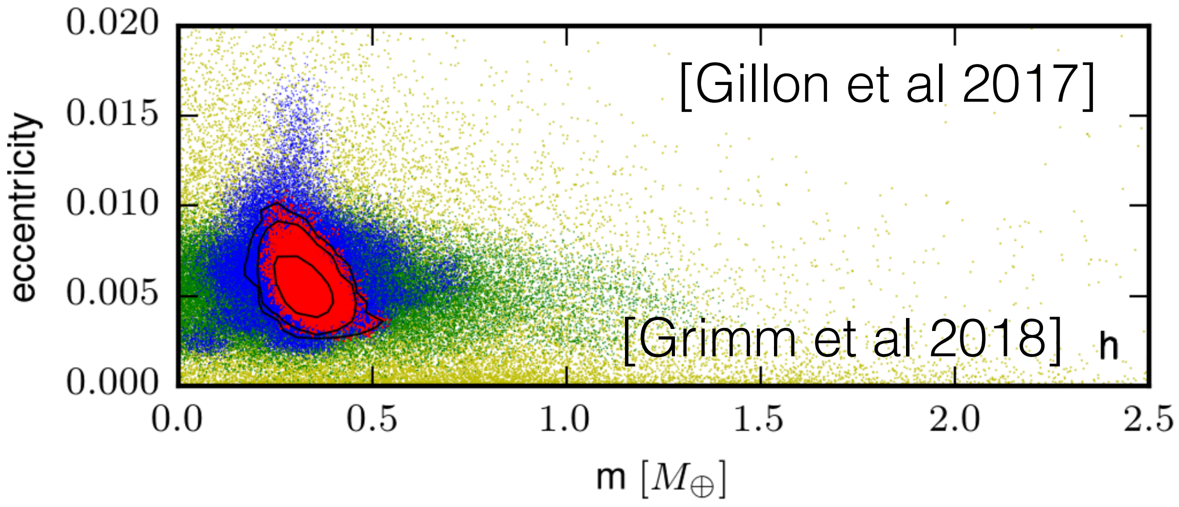
Variational Bayes and beyond: Bayesian inference for big data

Tamara Broderick
Associate Professor,
Electrical Engineering & Computer Science
MIT

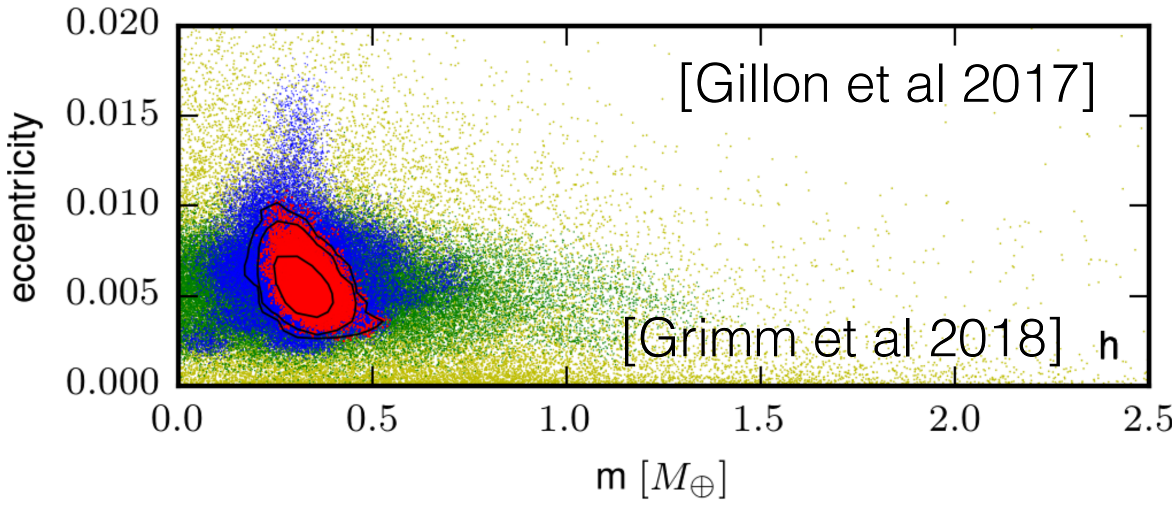
http://www.tamarabroderick.com/tutorials.html

Bayesian inference

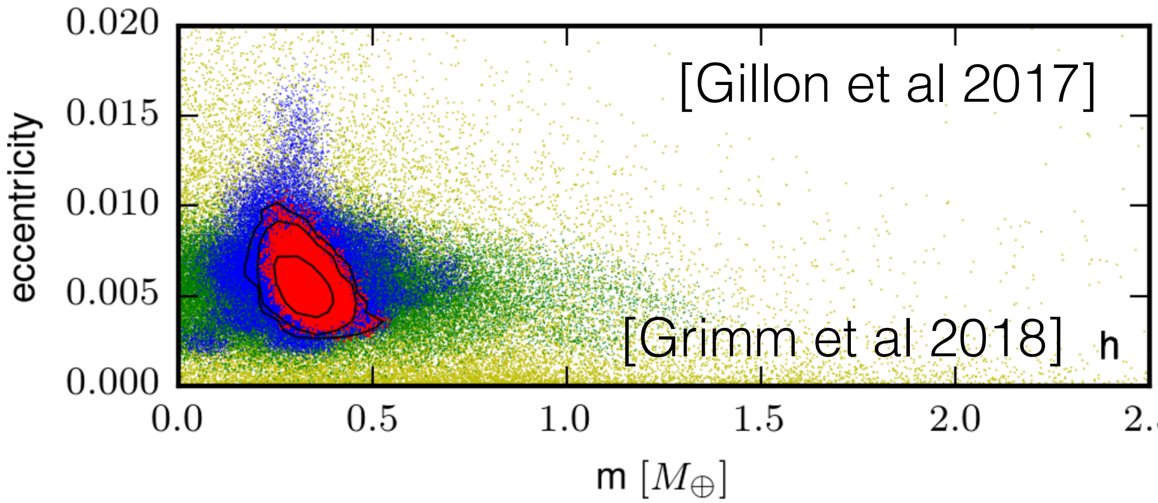
Bayesian inference

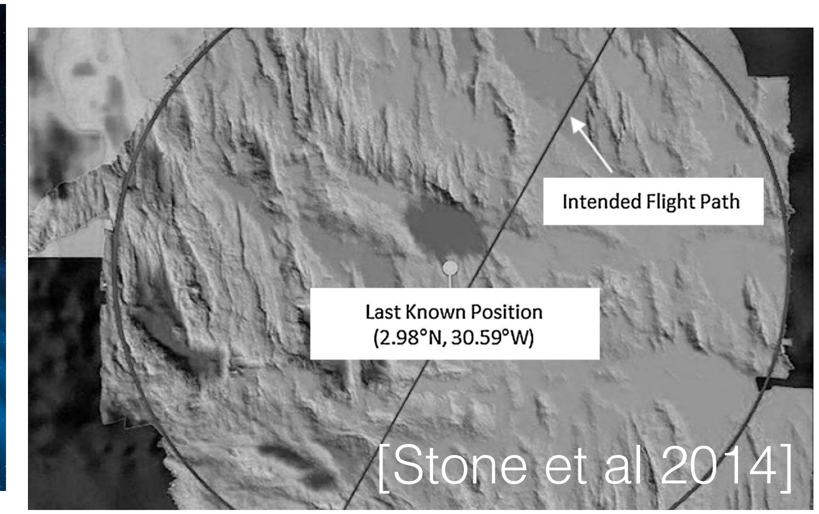


Bayesian inference

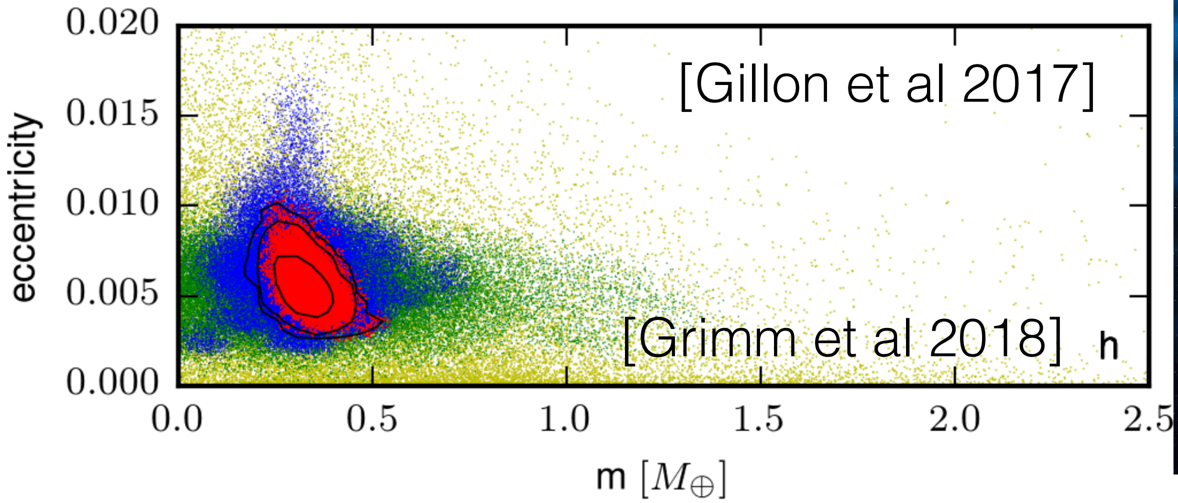


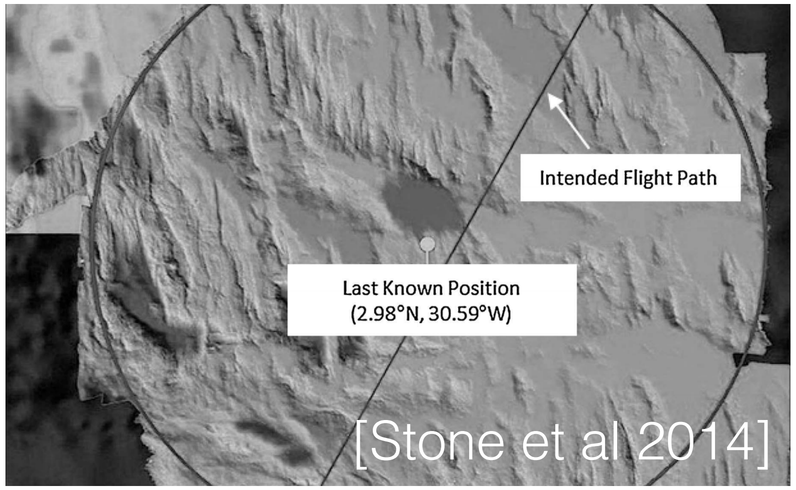
Bayesian inference

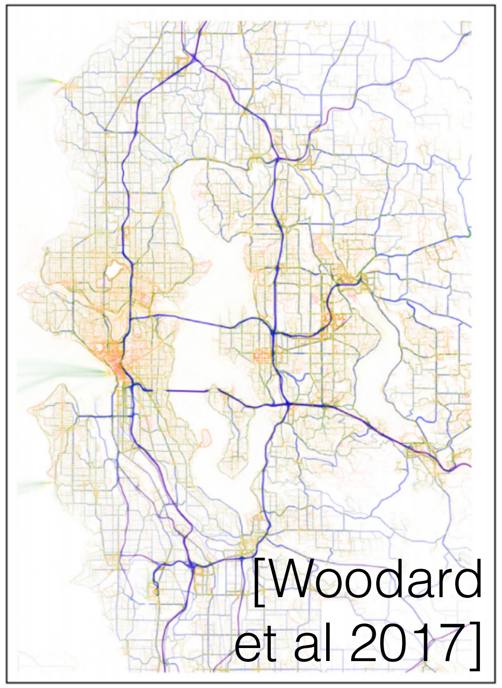




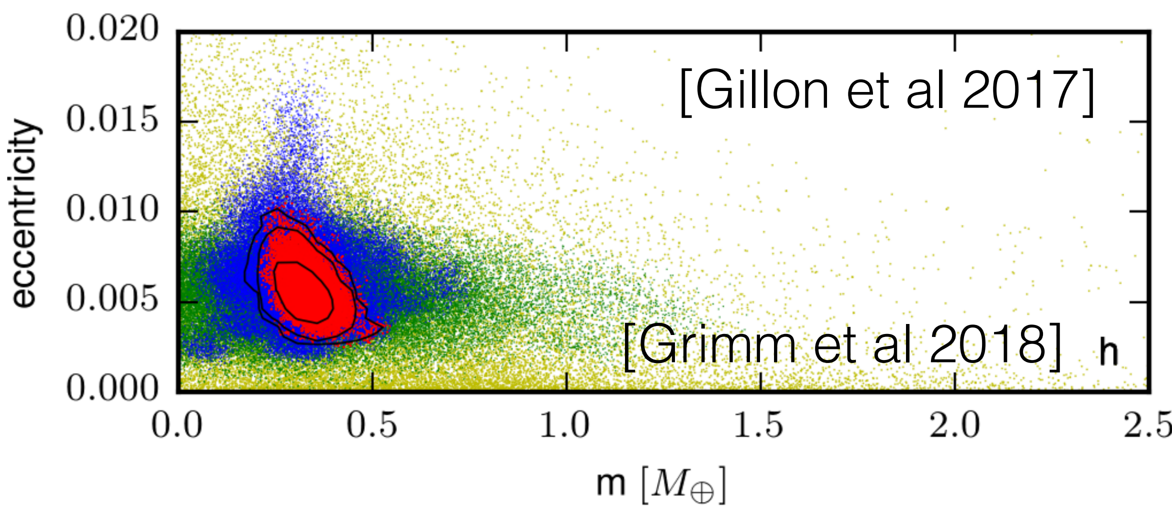
Bayesian inference

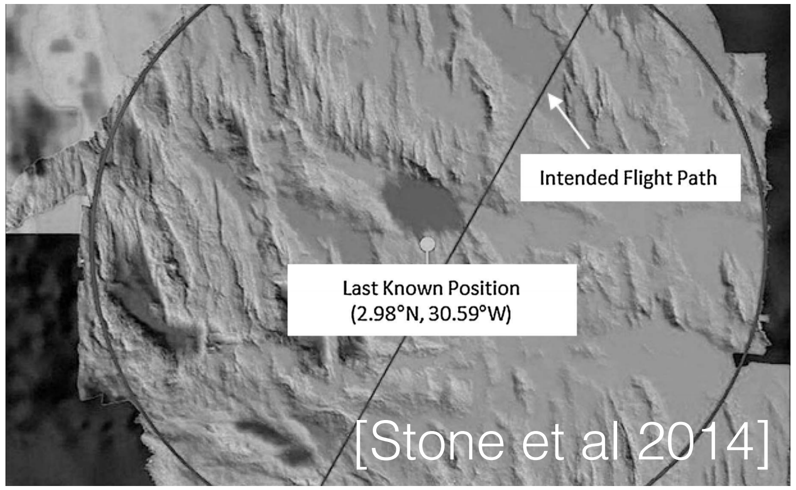


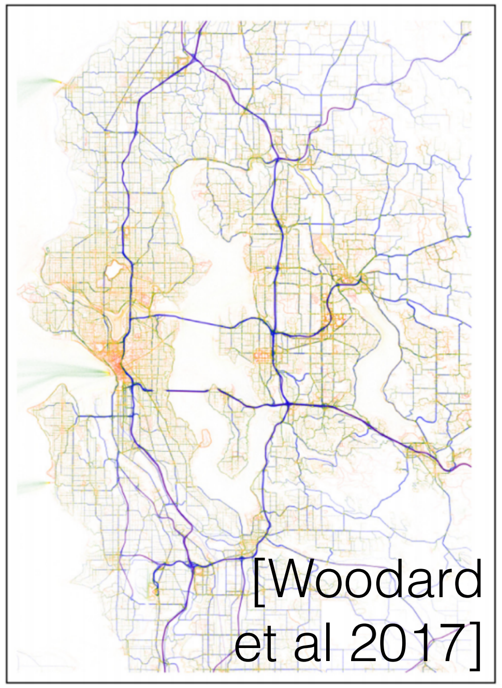


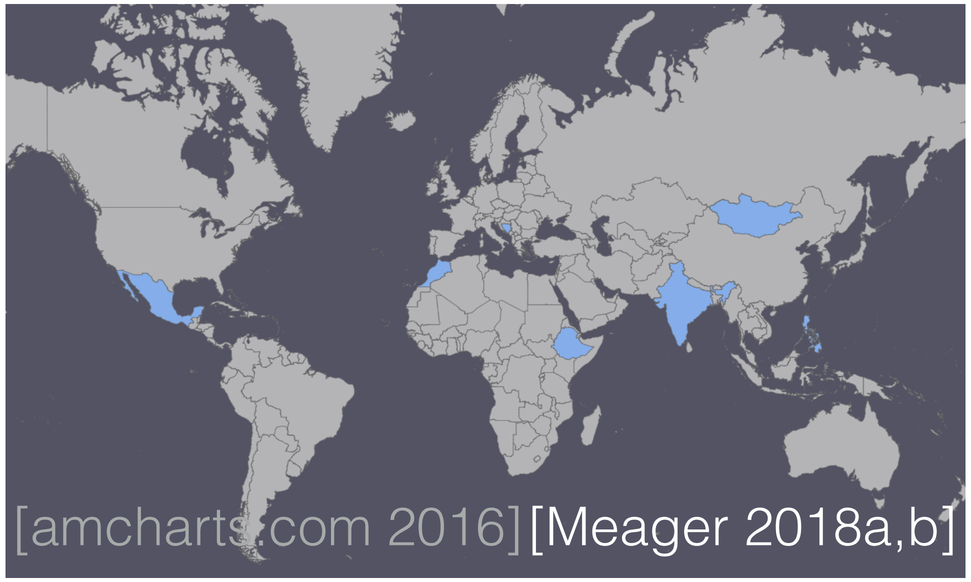


Bayesian inference

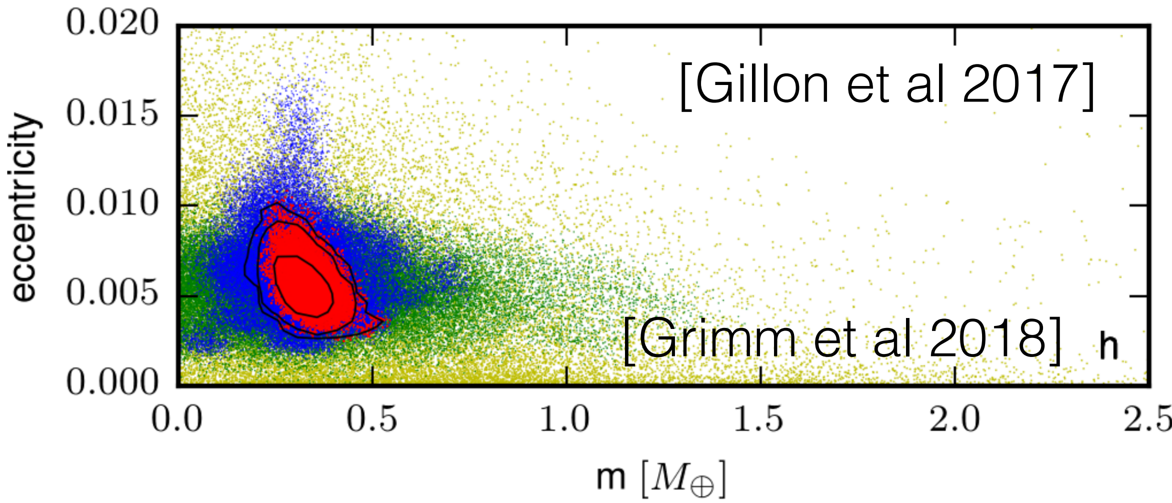


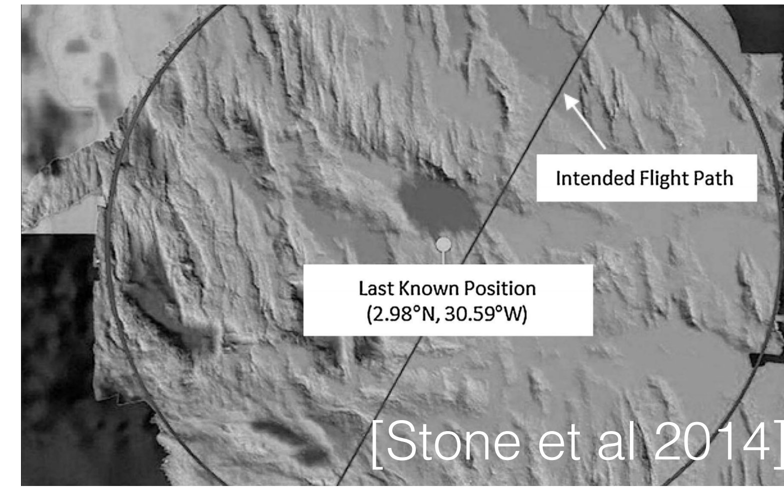


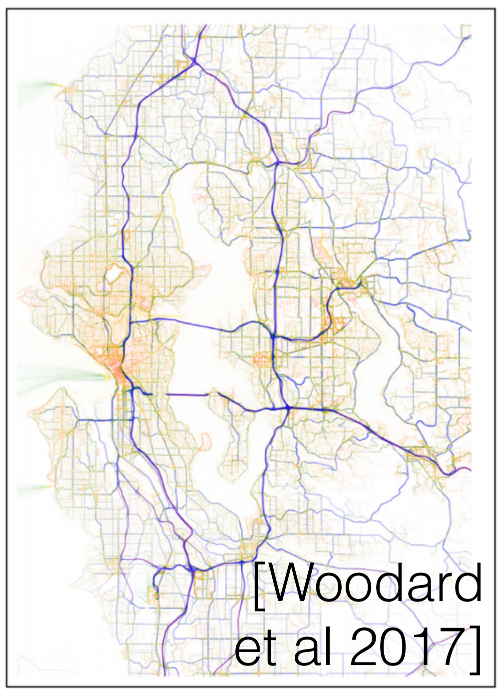


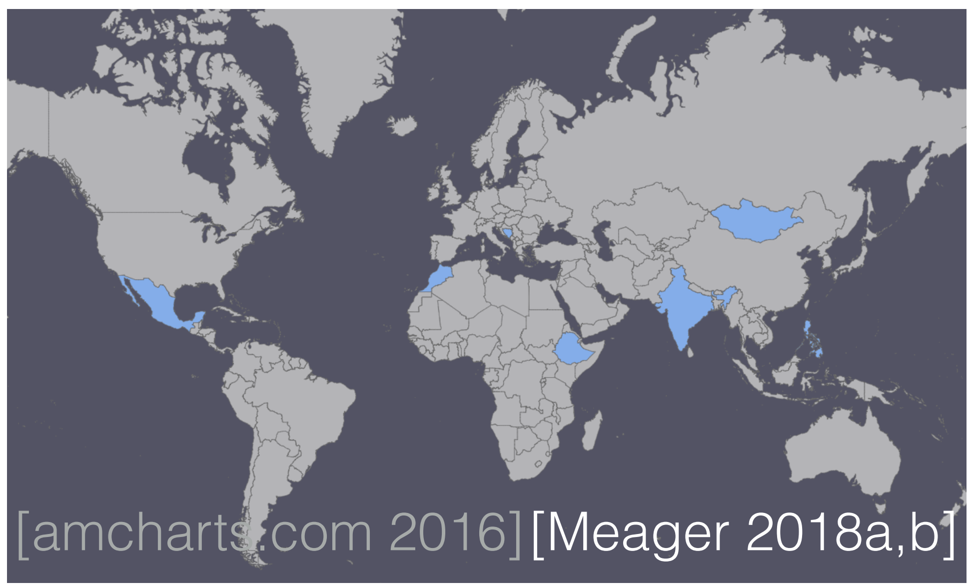


Bayesian inference

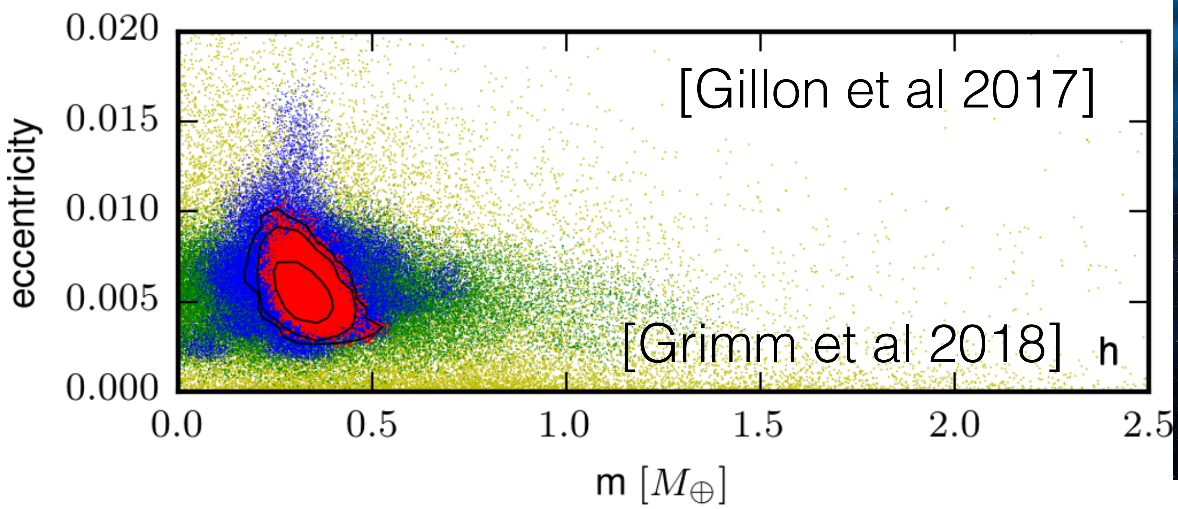


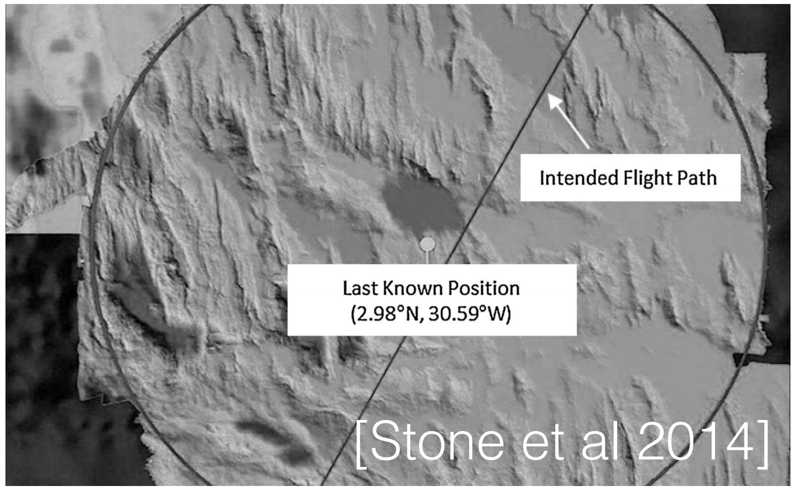


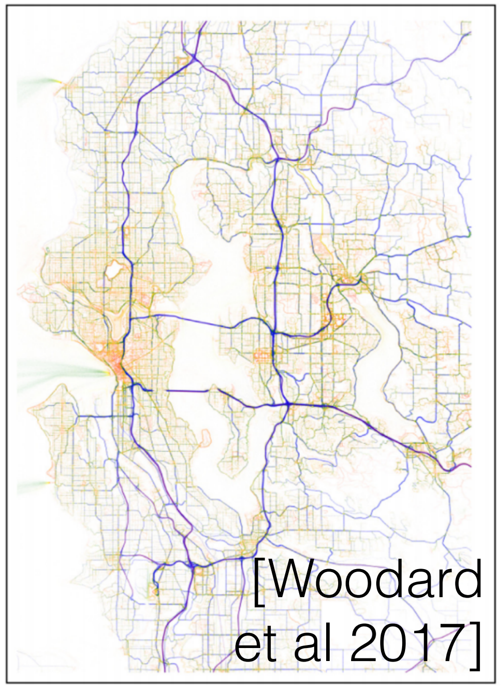


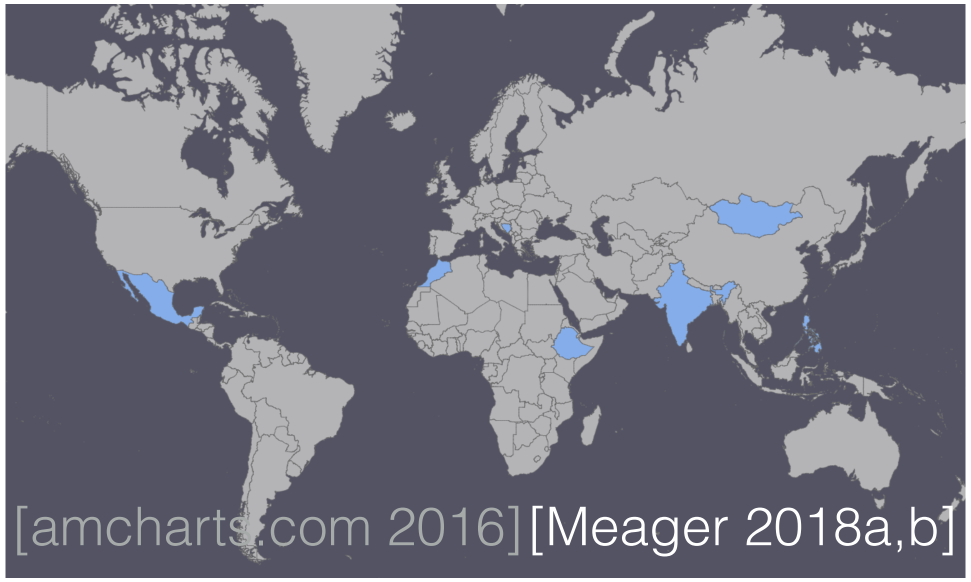


Bayesian inference

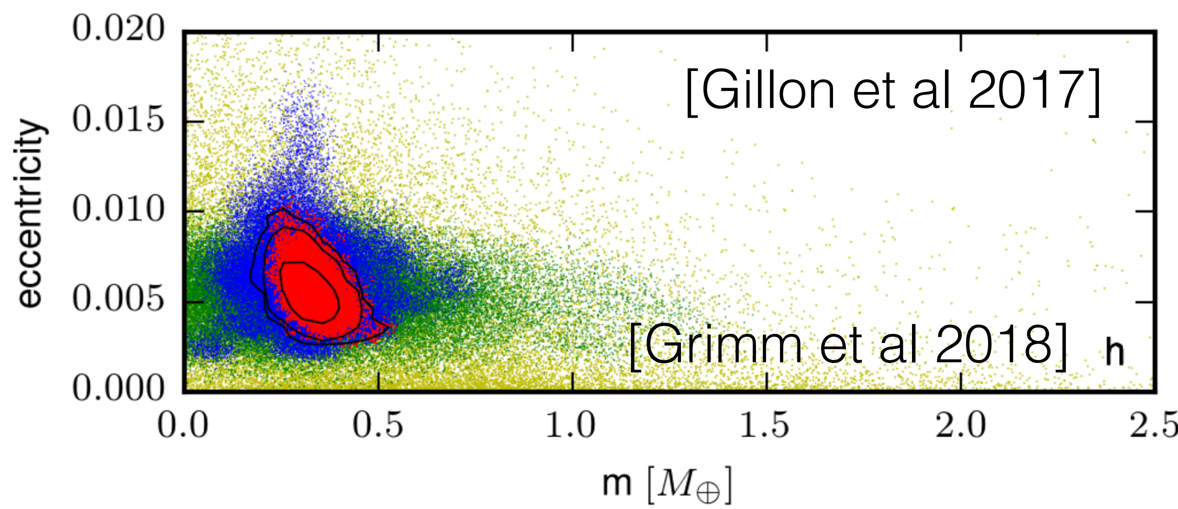


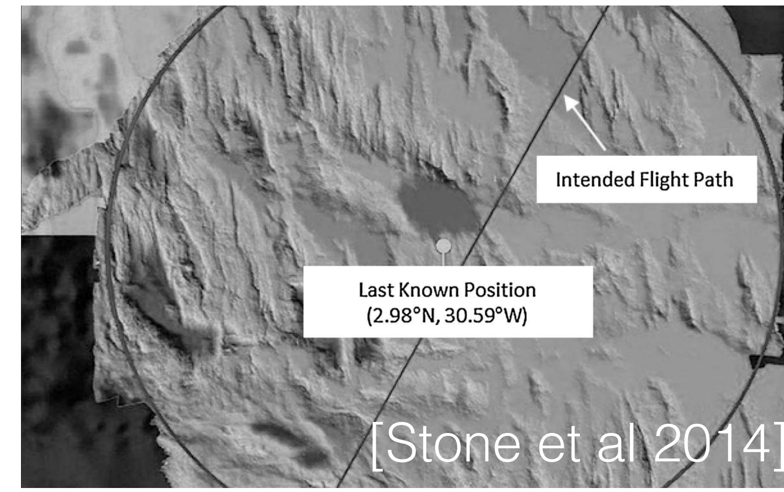


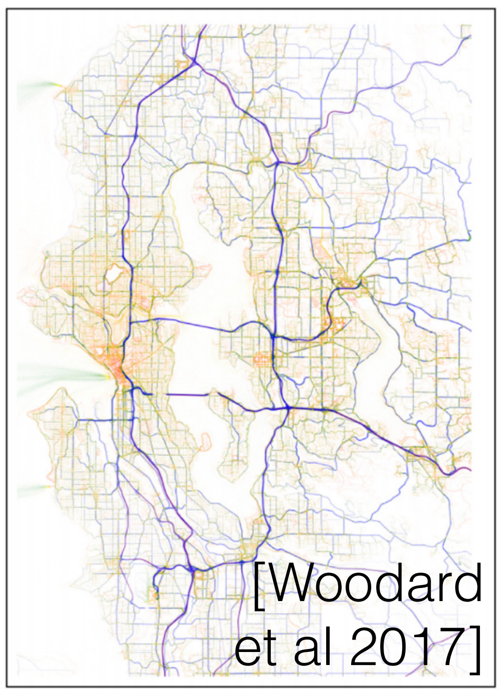


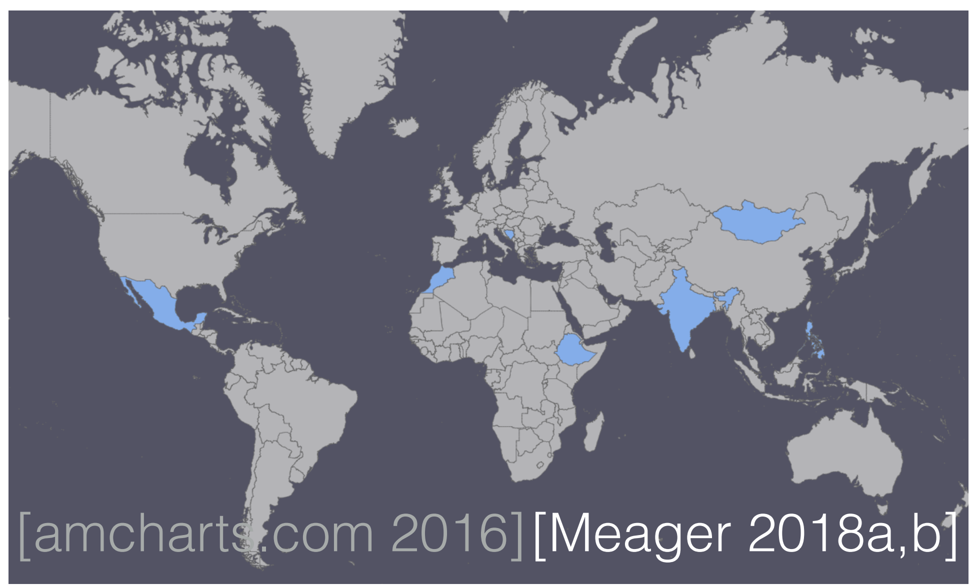


Bayesian inference



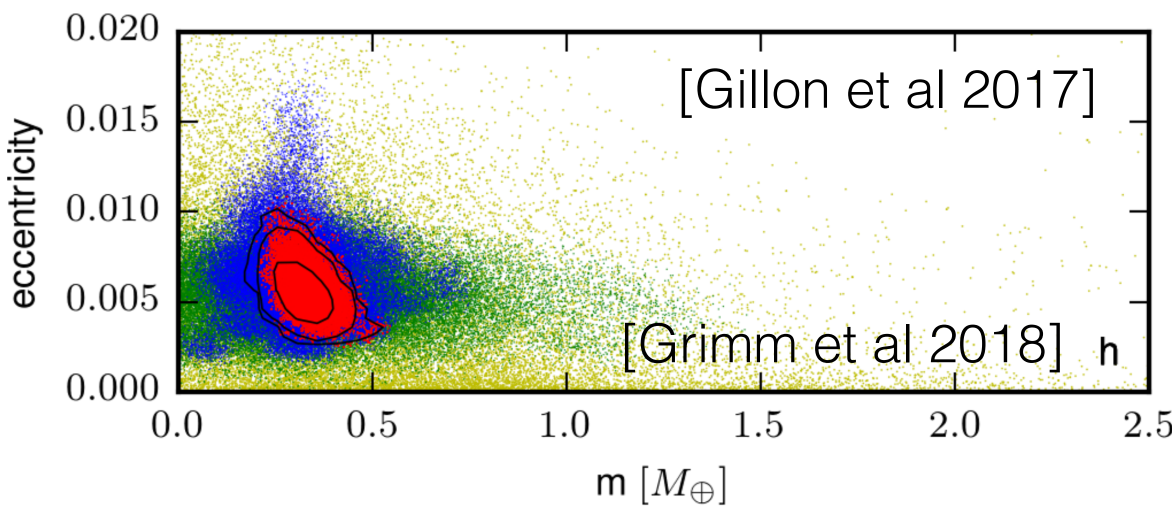


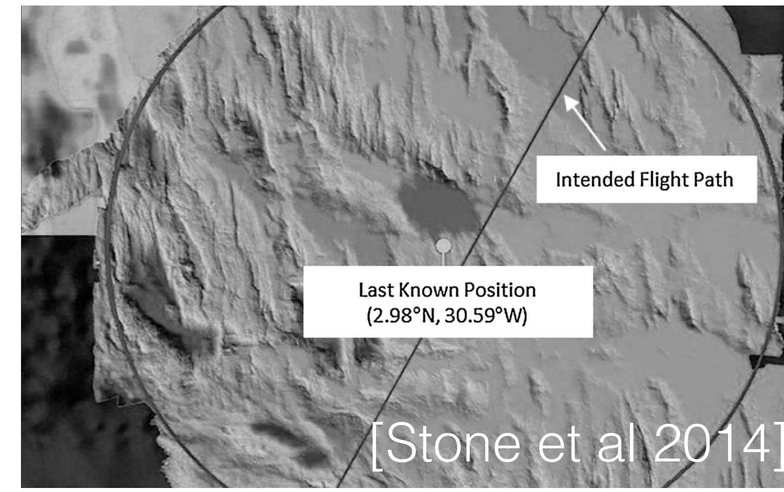


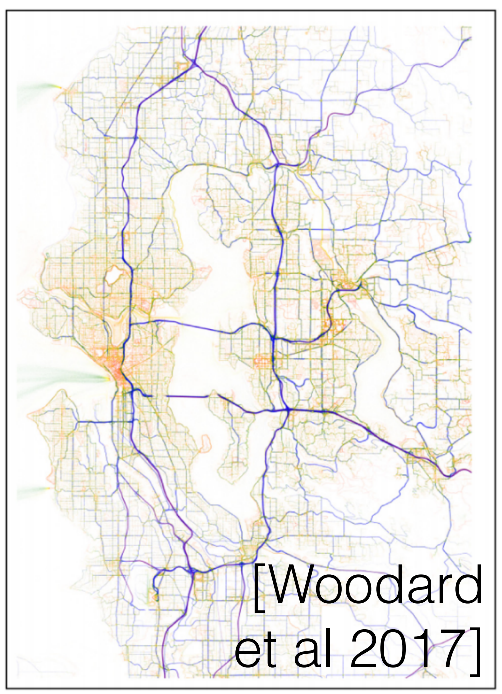


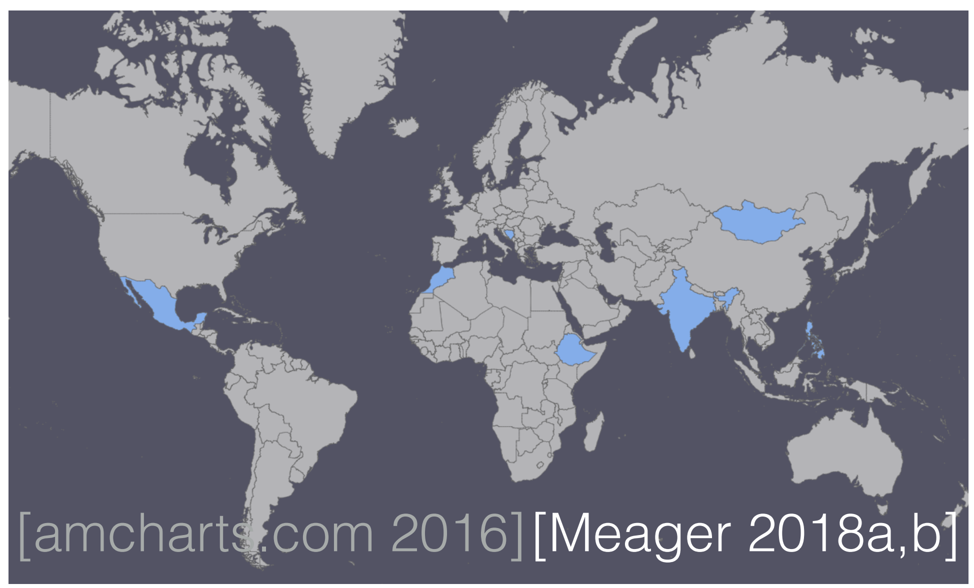
- Goals: good point estimates, uncertainty estimates

Bayesian inference



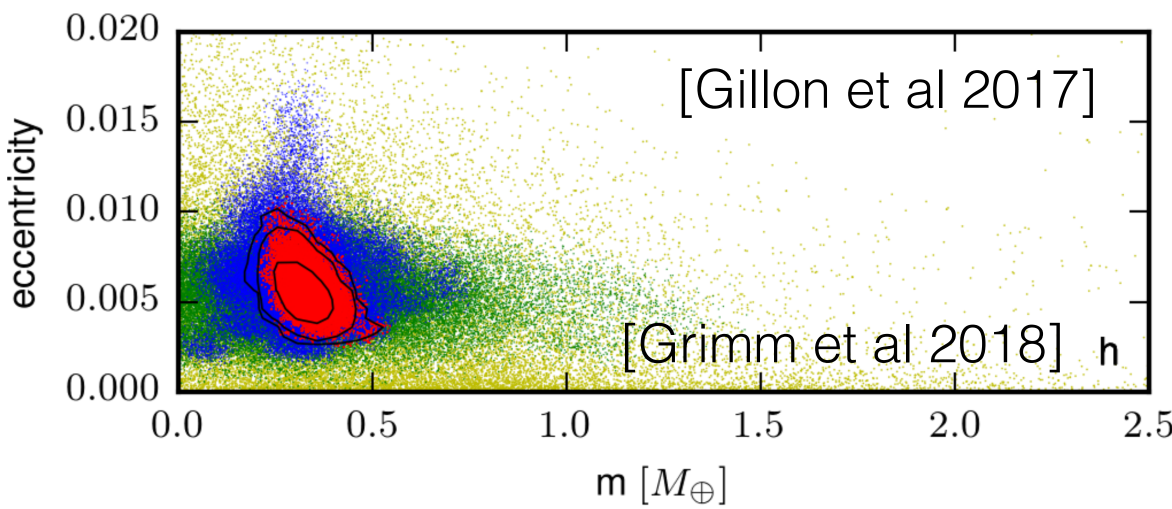


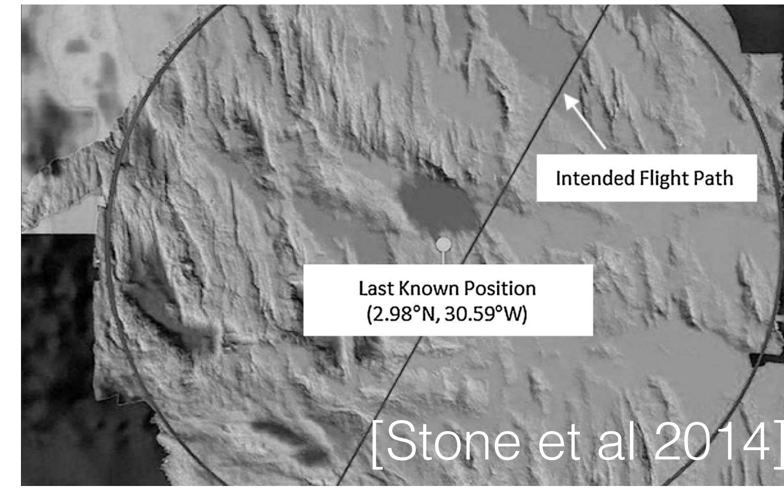


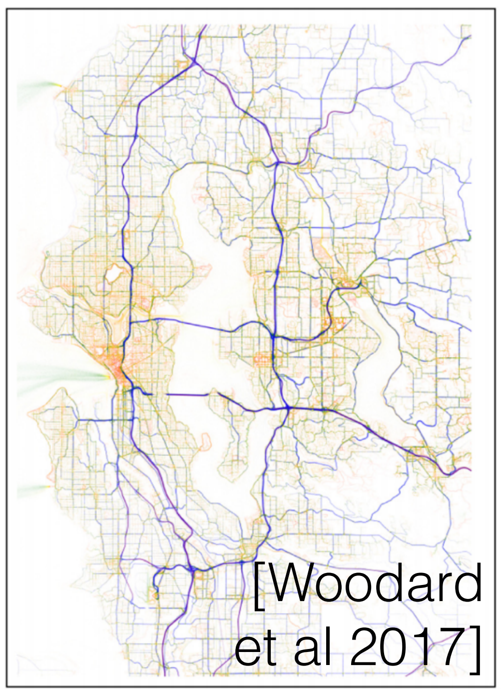


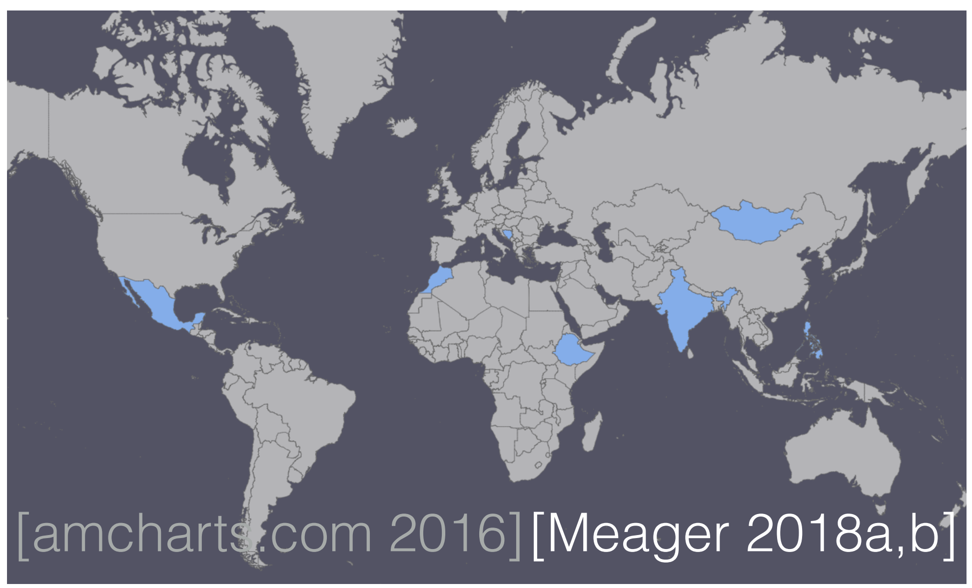
- Goals: good point estimates, uncertainty estimates
  - More: interpretable, flexible, modular, expert info

Bayesian inference





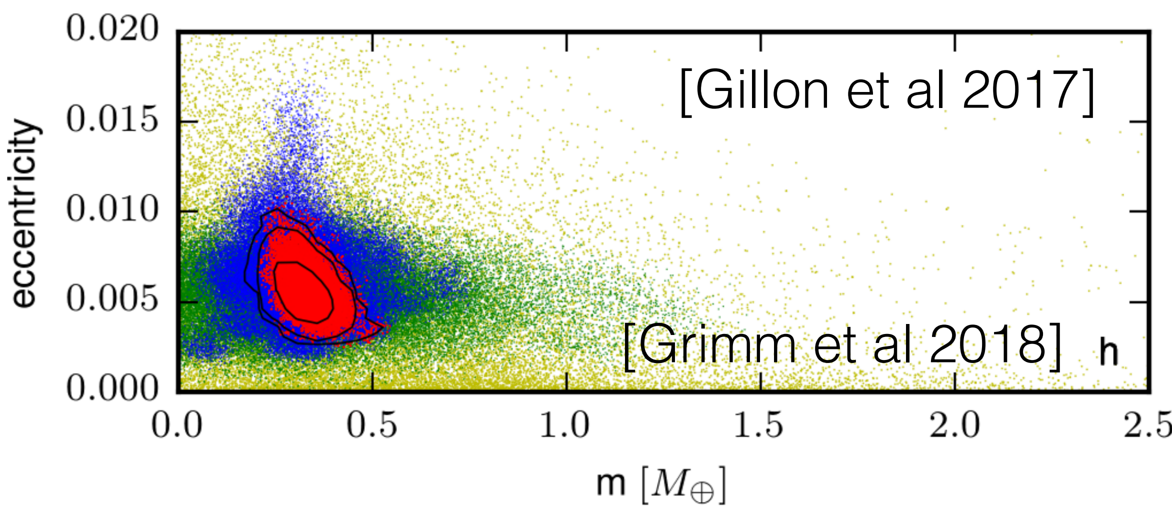


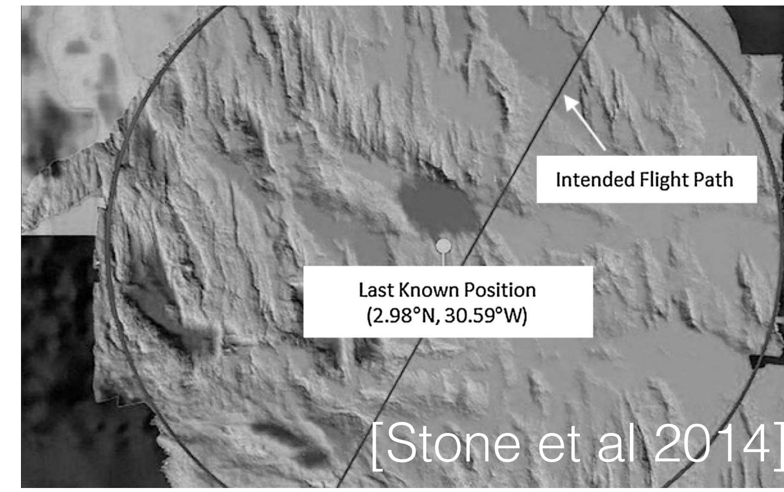


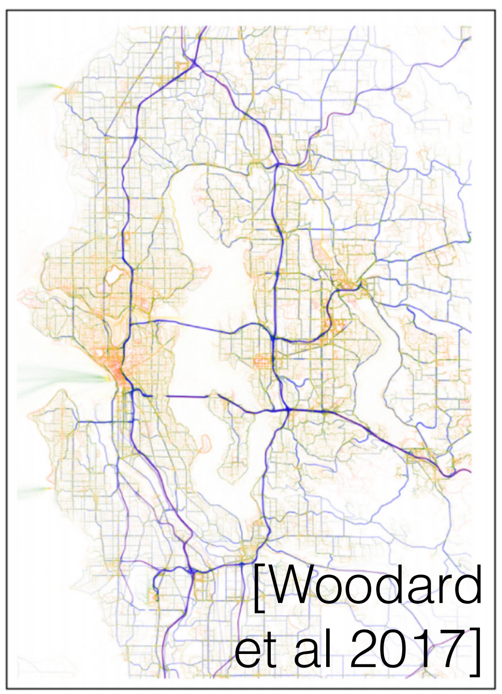
[mc-stan.org]

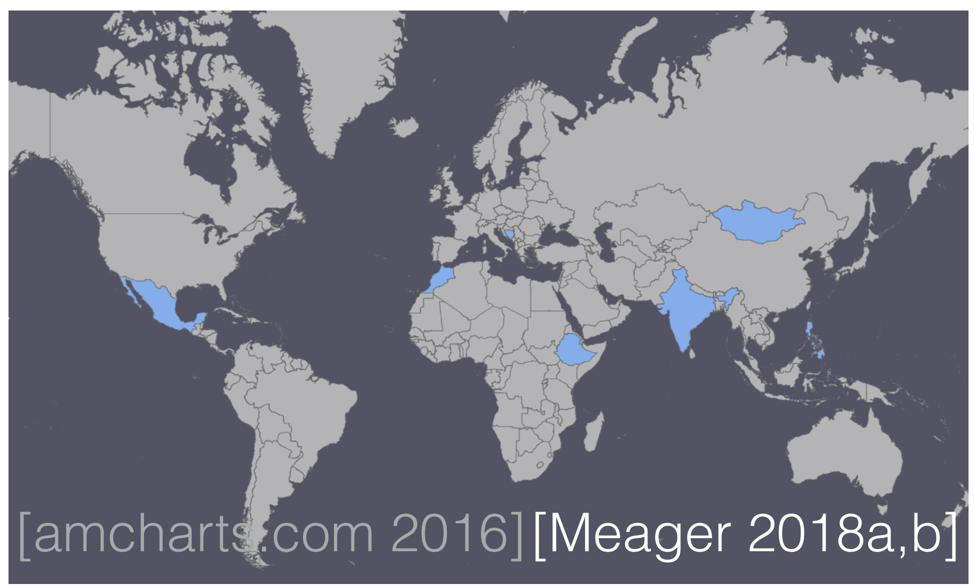
- Goals: good point estimates, uncertainty estimates
  - More: interpretable, flexible, modular, expert info

Bayesian inference





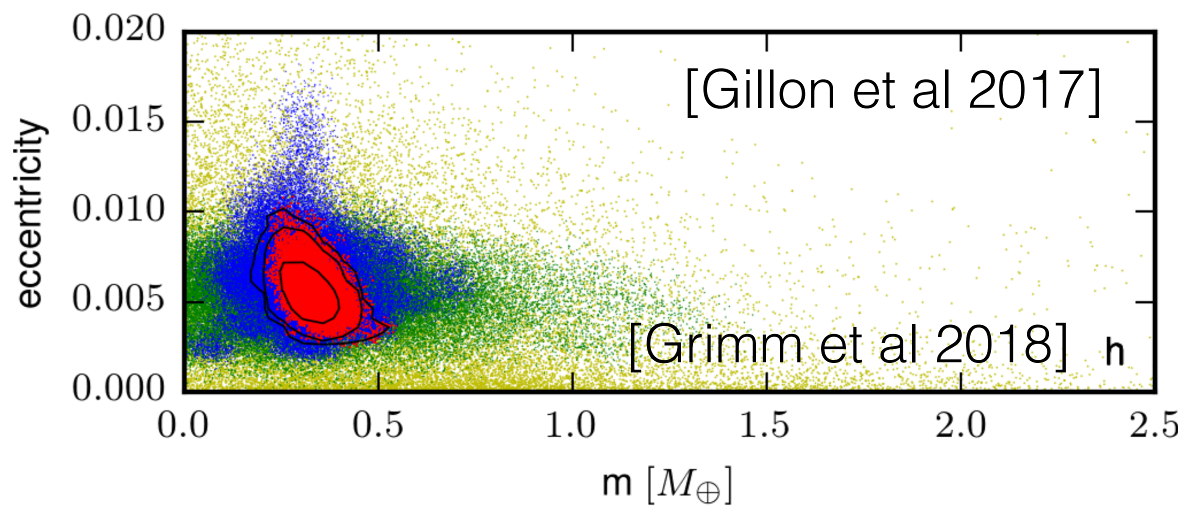


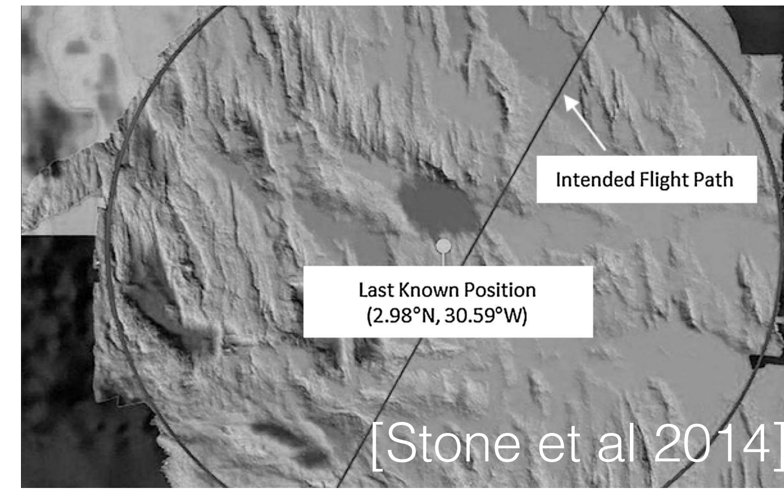


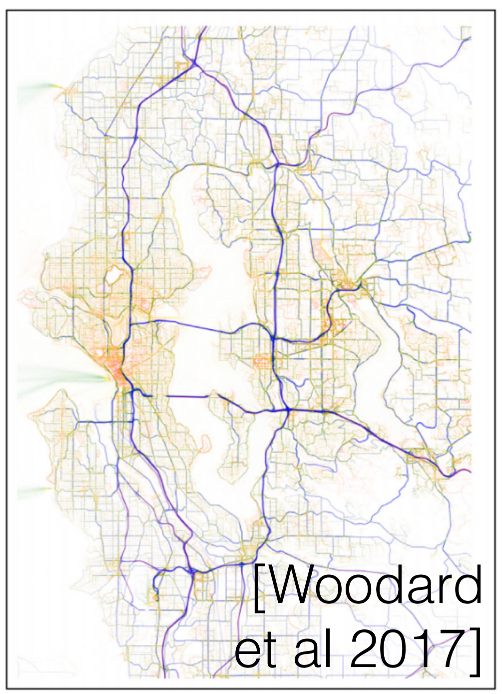
[mc-stan.org]

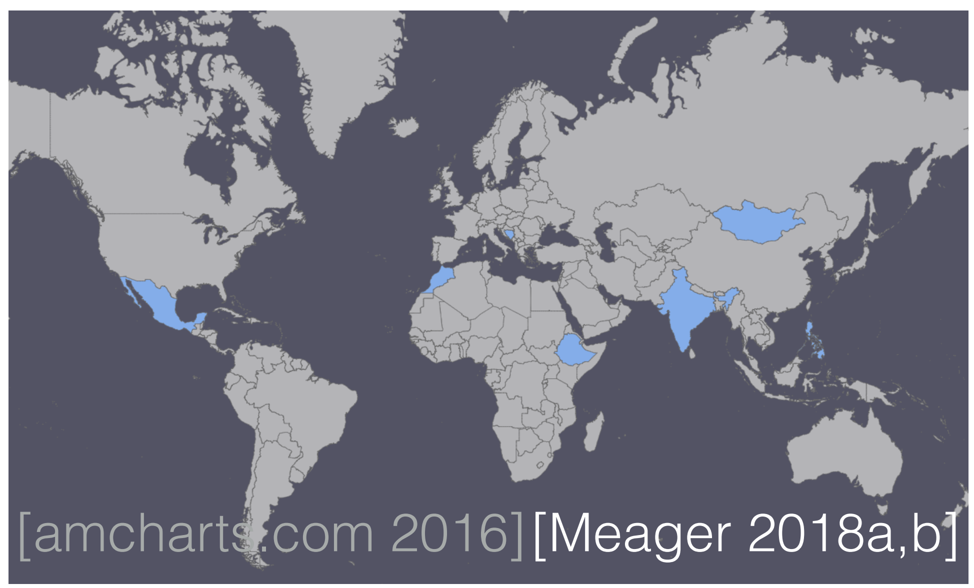
- Goals: good point estimates, uncertainty estimates
  - More: interpretable, flexible, modular, expert info
- Challenge: speed (compute, user), reliable inference

Bayesian inference









[mc-stan.org]

- Goals: good point estimates, uncertainty estimates
  - More: interpretable, flexible, modular, expert info
- Challenge: speed (compute, user), reliable inference
- Uncertainty doesn’t have to disappear in large data sets

Variational Bayes

Variational Bayes

- Modern problems: often large data, large dimensions

Variational Bayes

- Modern problems: often large data, large dimensions
- Variational Bayes can be very fast

Variational Bayes

- Modern problems: often large data, large dimensions
- Variational Bayes can be very fast

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

[Blei et al 2003]

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

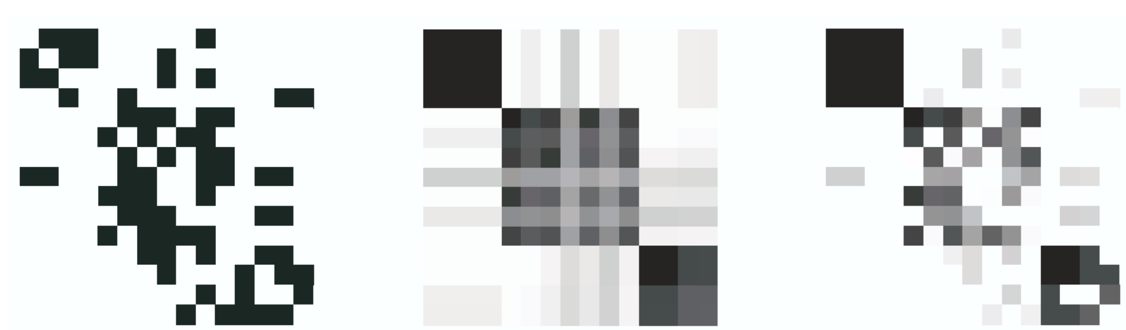
Variational Bayes

- Modern problems: often large data, large dimensions
- Variational Bayes can be very fast

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

[Blei et al 2003]

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.



[Airoldi et al 2008]

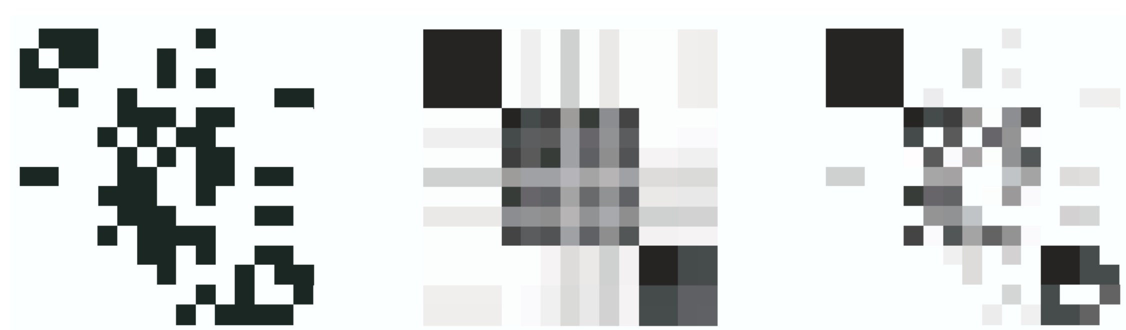
Variational Bayes

- Modern problems: often large data, large dimensions
- Variational Bayes can be very fast

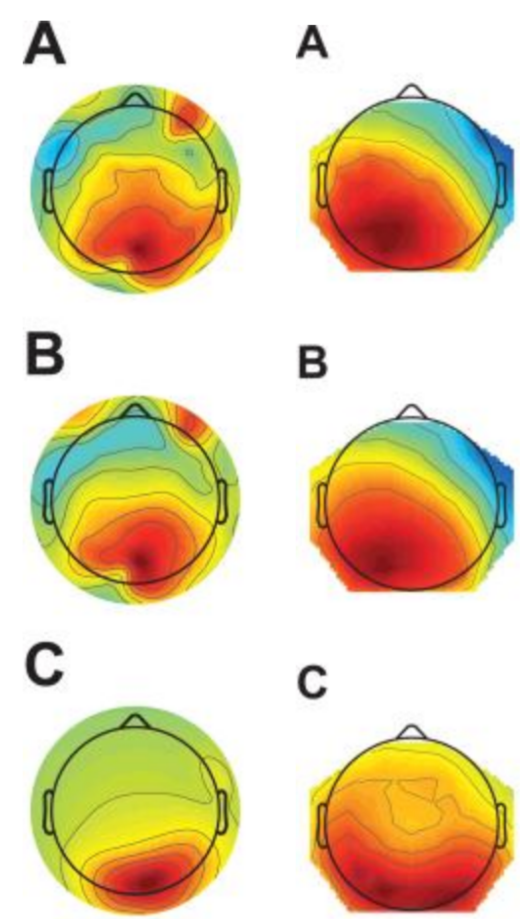
"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

[Blei et al 2003]

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.



[Airoldi et al 2008]



[Gershman et al 2014]

[Blei et al 2018]

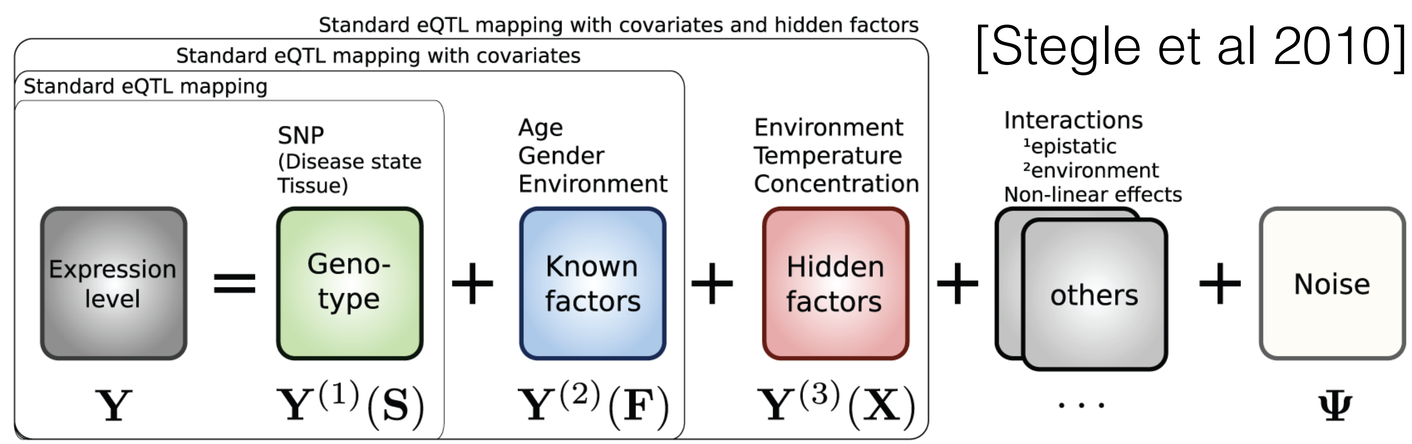
Variational Bayes

- Modern problems: often large data, large dimensions
- Variational Bayes can be very fast

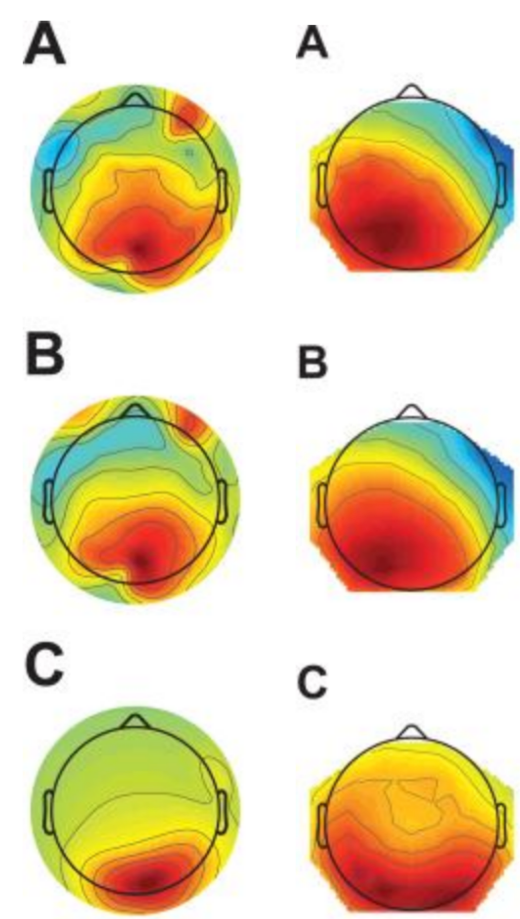
"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

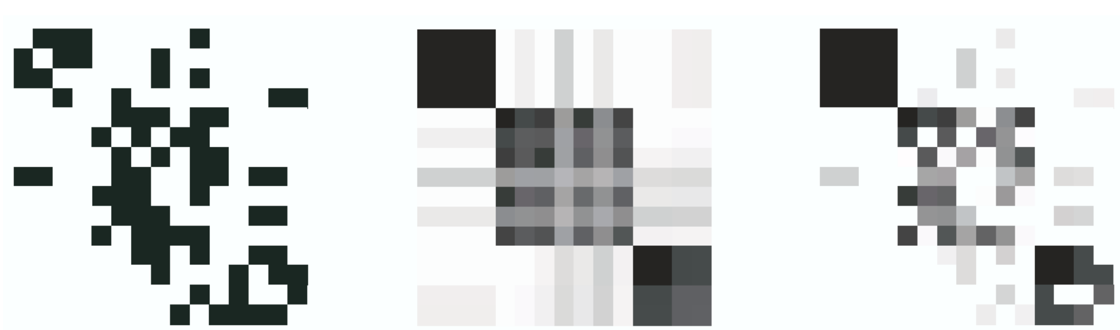
[Blei et al 2003]

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.



[Stegle et al 2010]





[Airoldi et al 2008]

[Gershman et al 2014]

[Blei et al 2018]

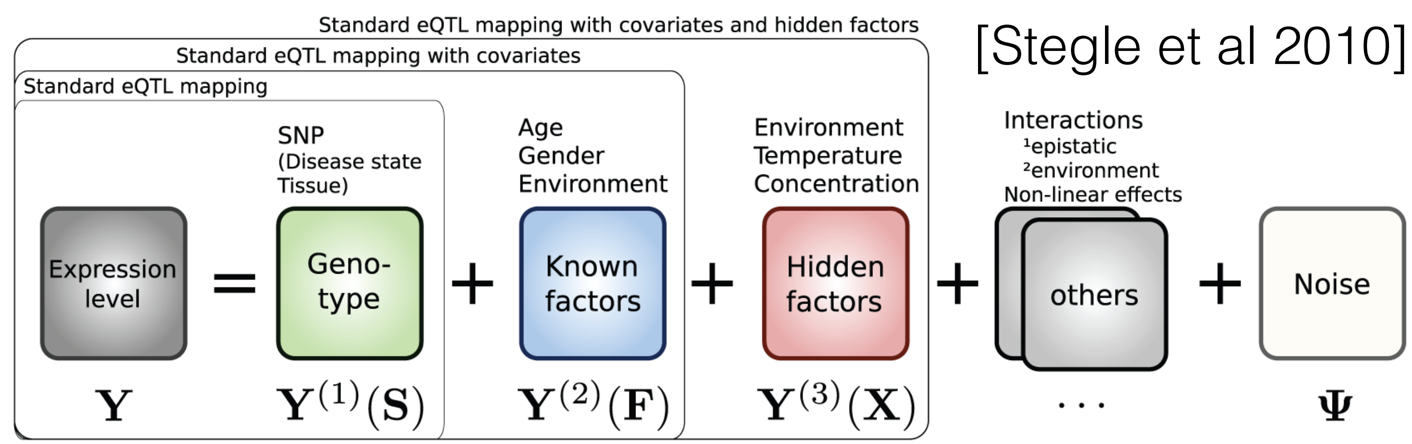
Variational Bayes

- Modern problems: often large data, large dimensions
- Variational Bayes can be very fast

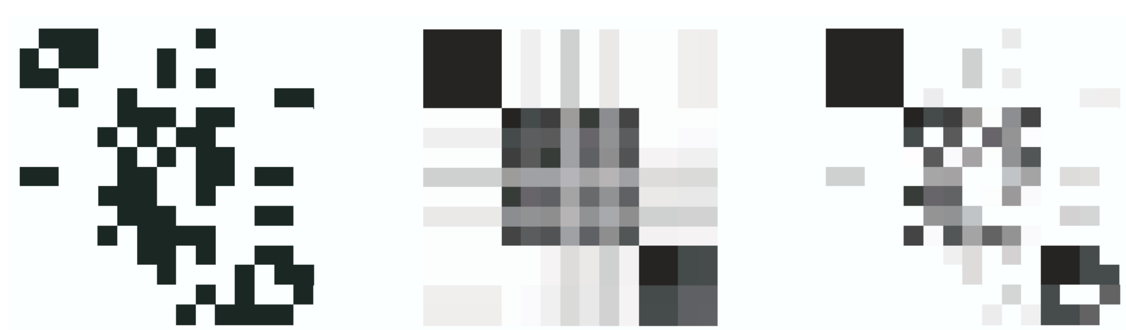
"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

[Blei et al 2003]

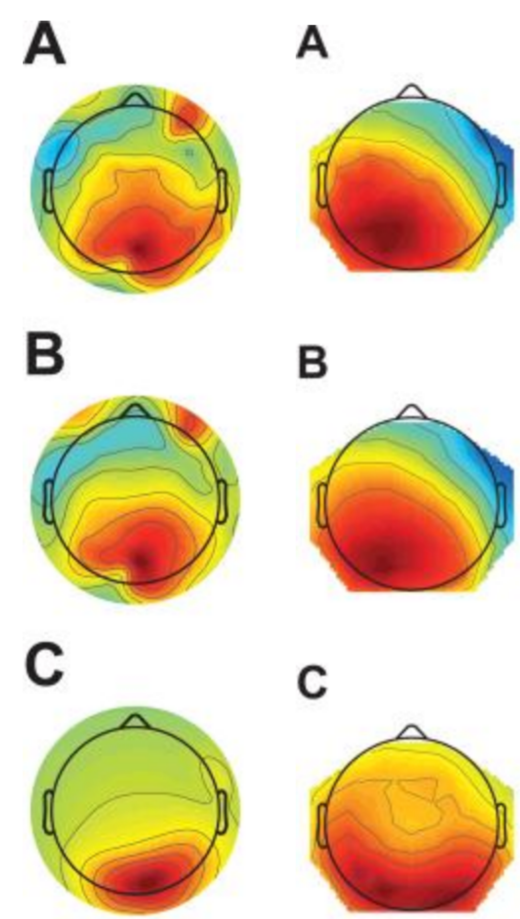
The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.



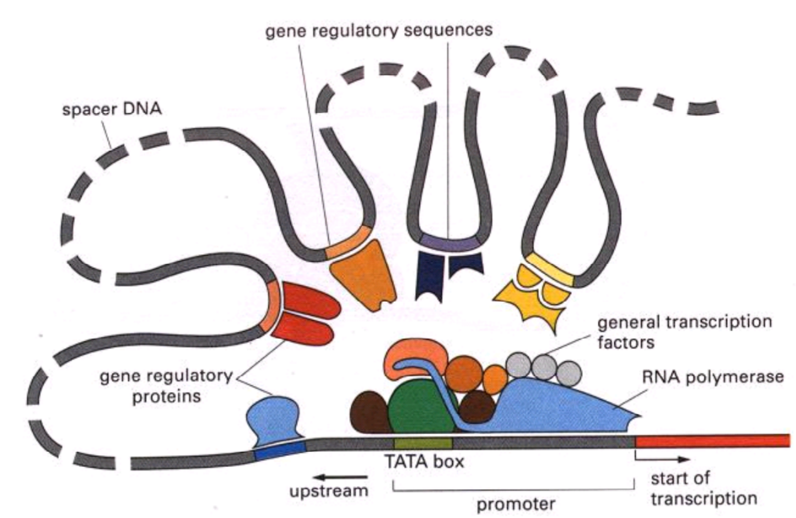
[Stegle et al 2010]



[Airoldi et al 2008]



[Gershman et al 2014]



[Xing et al 2004]
[Xing 2003]

[Blei et al 2018]

Roadmap

- Bayes & Approximate Bayes review

Roadmap

- Bayes & Approximate Bayes review
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)

Roadmap

- Bayes & Approximate Bayes review
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use MFVB?

Roadmap

- Bayes & Approximate Bayes review
- What is:
    - Variational Bayes (VB)
    - Mean-field variational Bayes (MFVB)
- Why use MFVB?
- When can we trust MFVB?

Roadmap

- Bayes & Approximate Bayes review
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use MFVB?
- When can we trust MFVB?
- Where do we go from here?

Roadmap

- Bayes & Approximate Bayes review
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use MFVB?
- When can we trust MFVB?
- Where do we go from here?

Bayesian inference

Bayesian inference

parameters

θ

Bayesian inference

parameters

$$p(\theta)$$

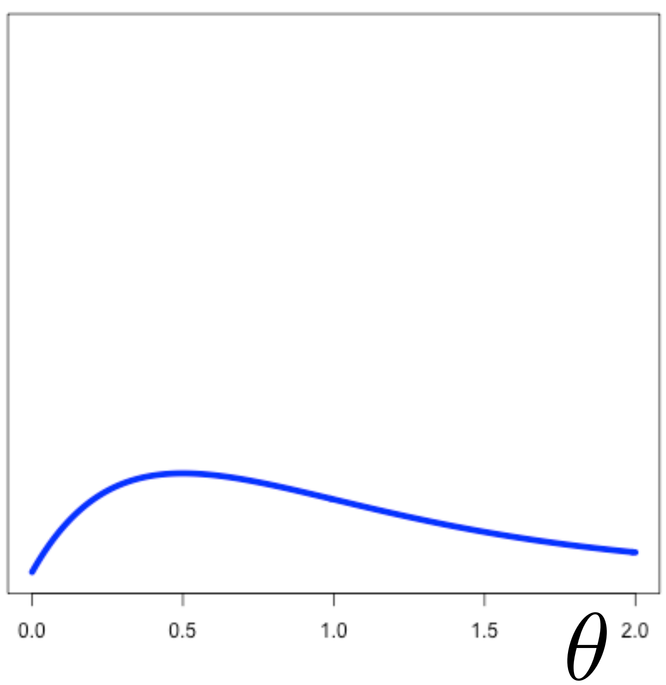
prior

Bayesian inference

parameters

$p(\theta)$

prior

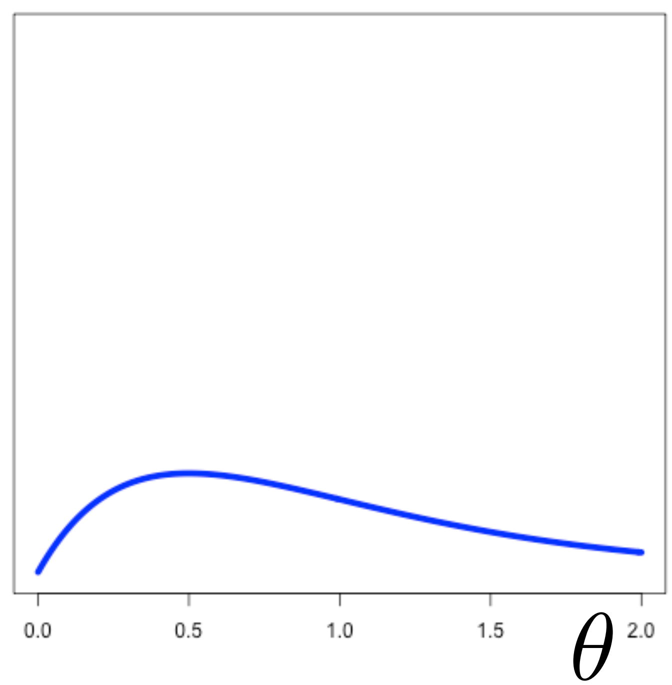


Bayesian inference

parameters

$$p(y_{1:N}|\theta)p(\theta)$$

likelihood prior

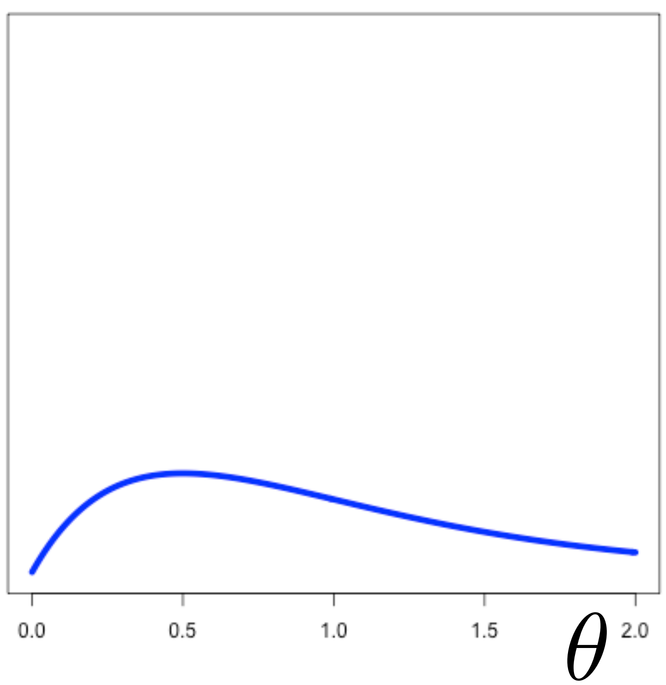


Bayesian inference

data　　parameters

$$p(y_{1:N}|\theta)p(\theta)$$

likelihood　prior

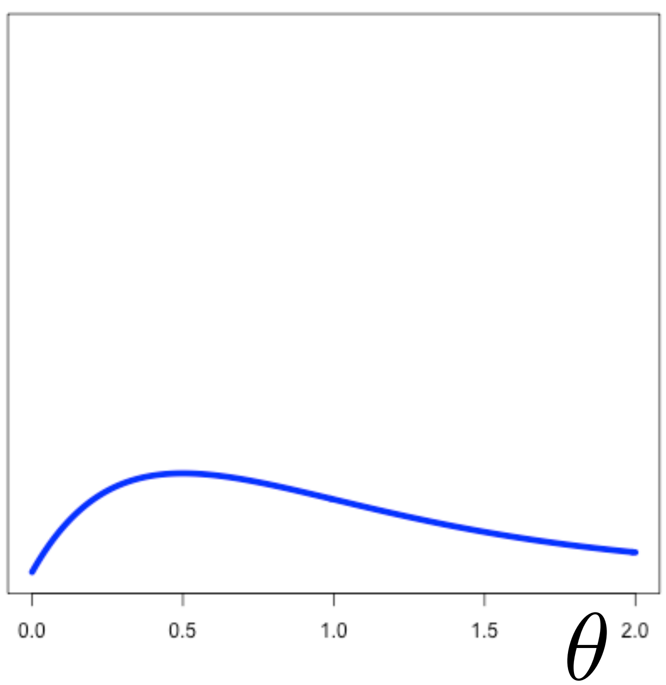


Bayesian inference

data parameters

$$p(\theta|y_{1:N}) \propto_\theta p(y_{1:N}|\theta)p(\theta)$$

posterior likelihood prior

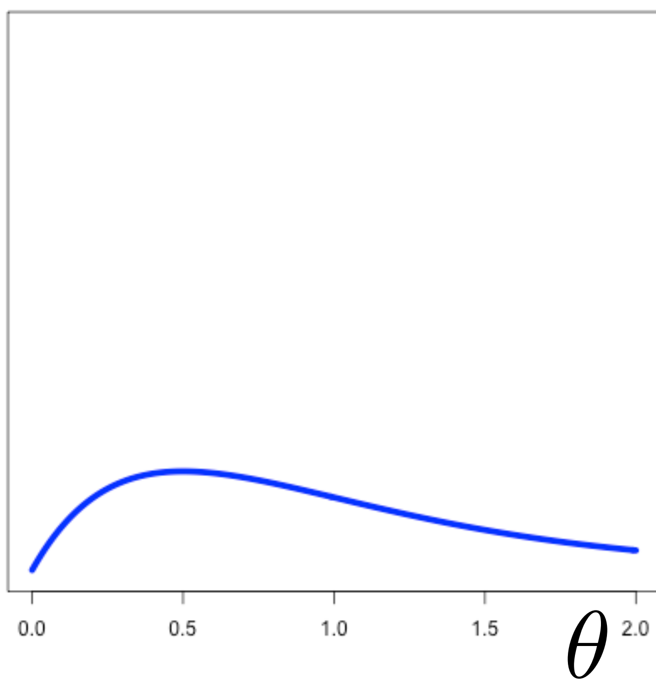


Bayesian inference

data → $y_{1:N}$, parameters → θ

$$p(\theta|y_{1:N}) \propto_\theta p(y_{1:N}|\theta)p(\theta)$$

posterior likelihood prior



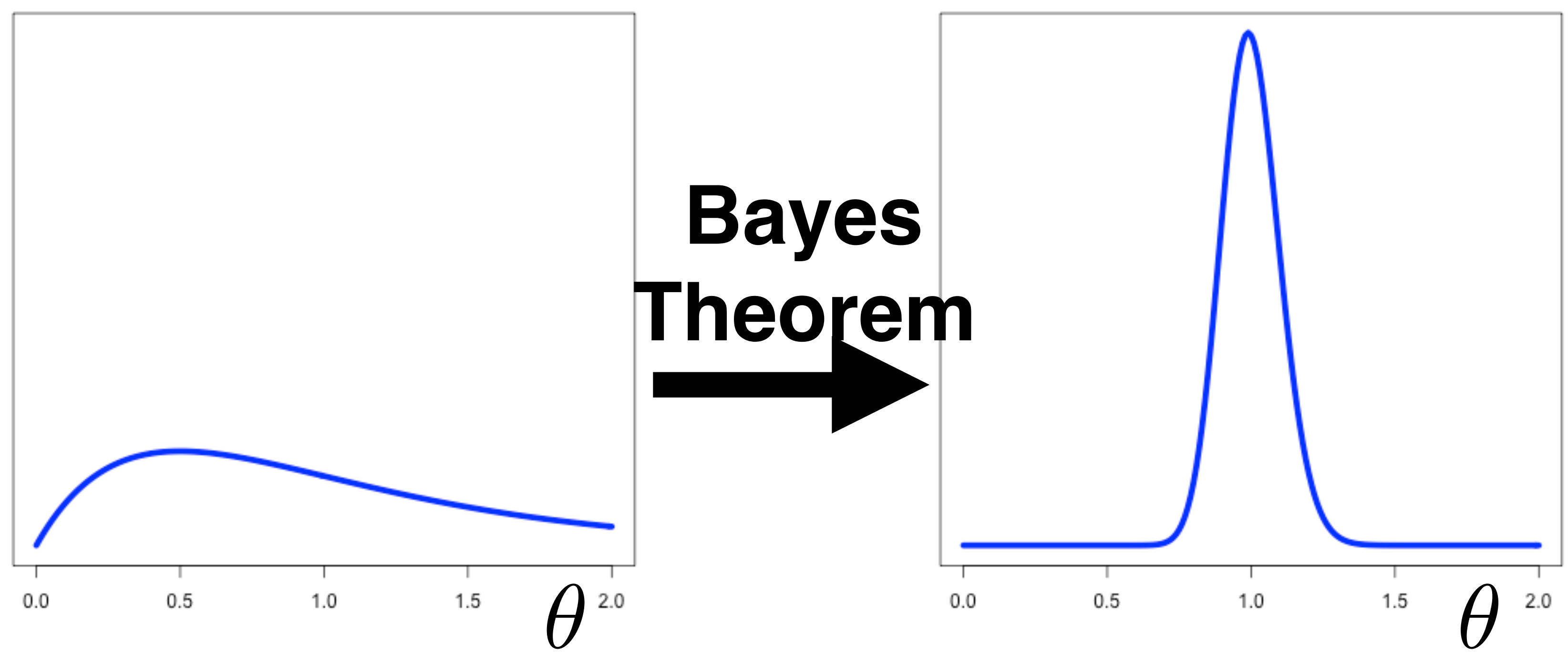
Bayes Theorem →

Bayesian inference

data parameters

$$p(\theta|y_{1:N}) \propto_\theta p(y_{1:N}|\theta)p(\theta)$$

posterior likelihood prior

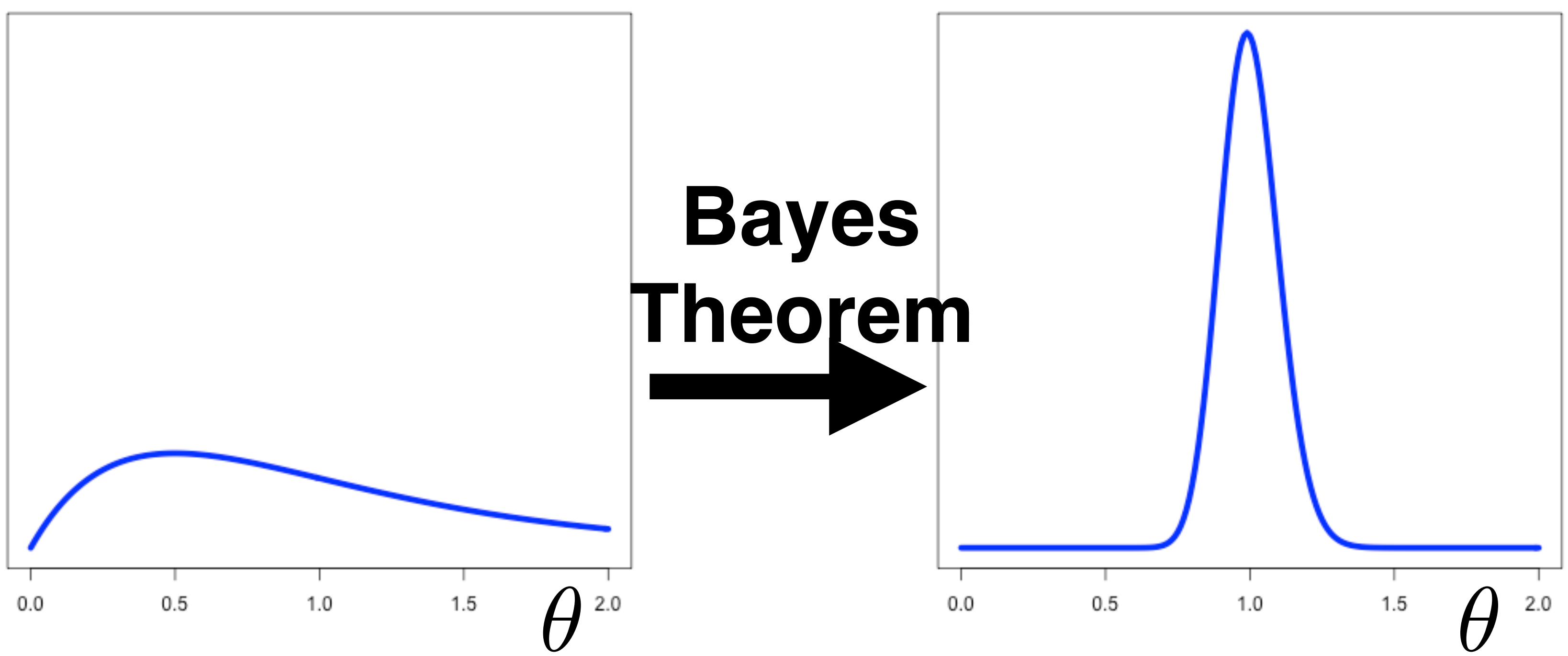


Bayesian inference

data parameters

$$p(\theta|y_{1:N}) \propto_{\theta} p(y_{1:N}|\theta)p(\theta)$$

posterior likelihood prior



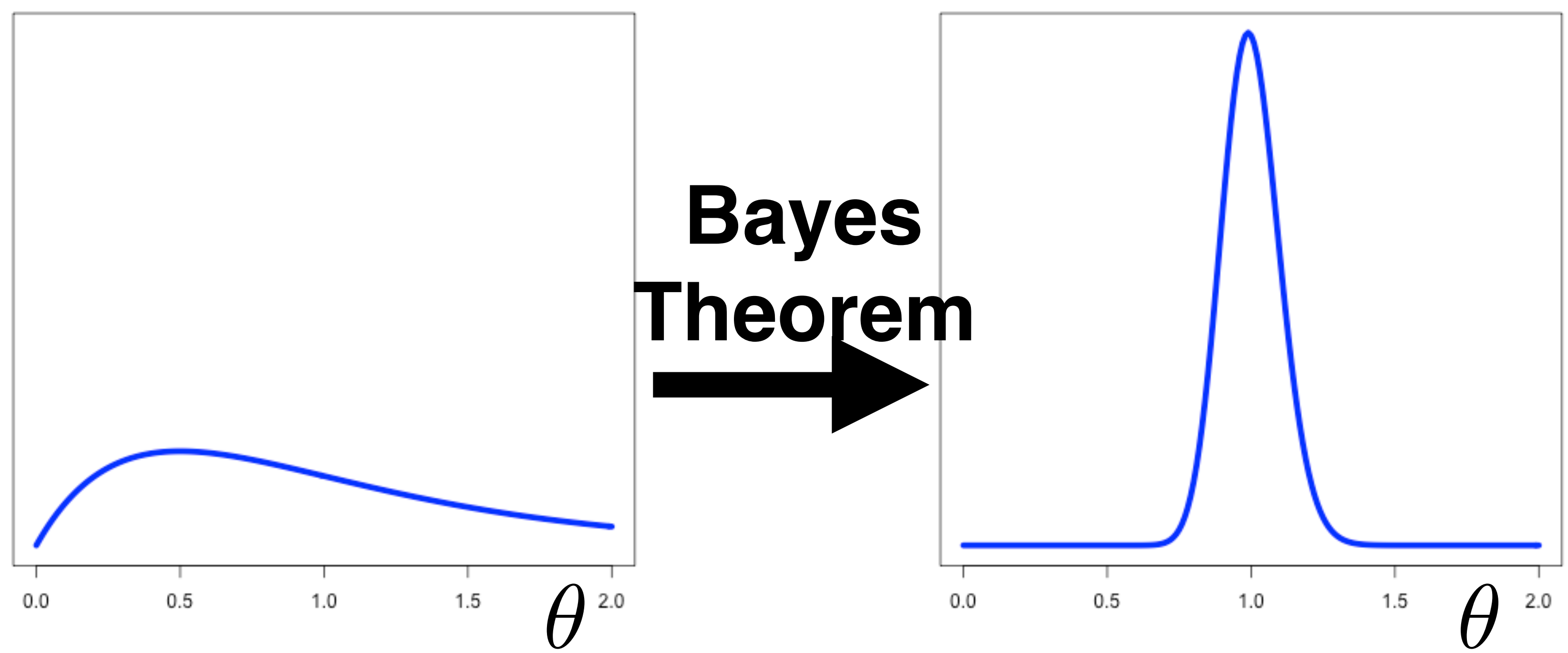
1. Build a model: choose prior & choose likelihood

Bayesian inference

data parameters

$$p(\theta|y_{1:N}) \propto_\theta p(y_{1:N}|\theta)p(\theta)$$

posterior likelihood prior



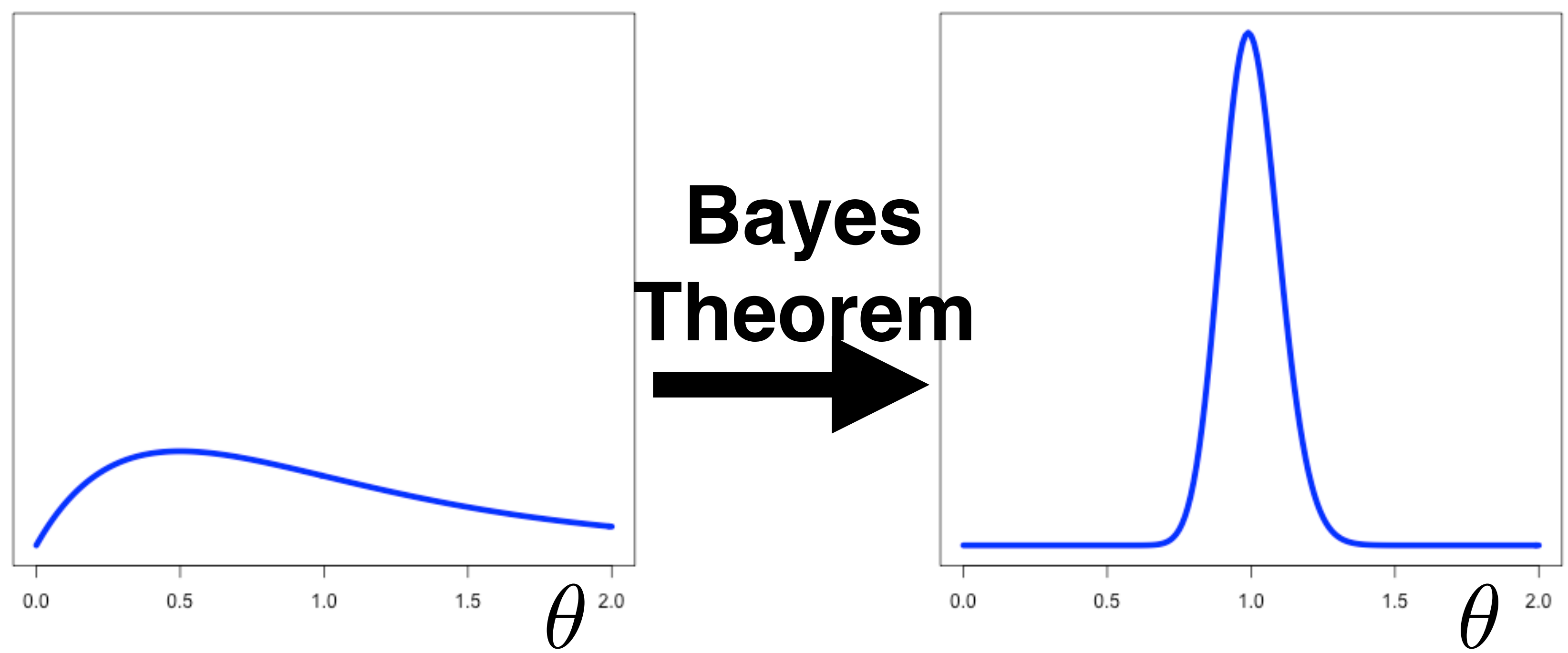
1. Build a model: choose prior & choose likelihood
2. Compute the posterior

Bayesian inference

data → likelihood; parameters → prior

$$p(\theta|y_{1:N}) \propto_\theta p(y_{1:N}|\theta)p(\theta)$$

posterior likelihood prior



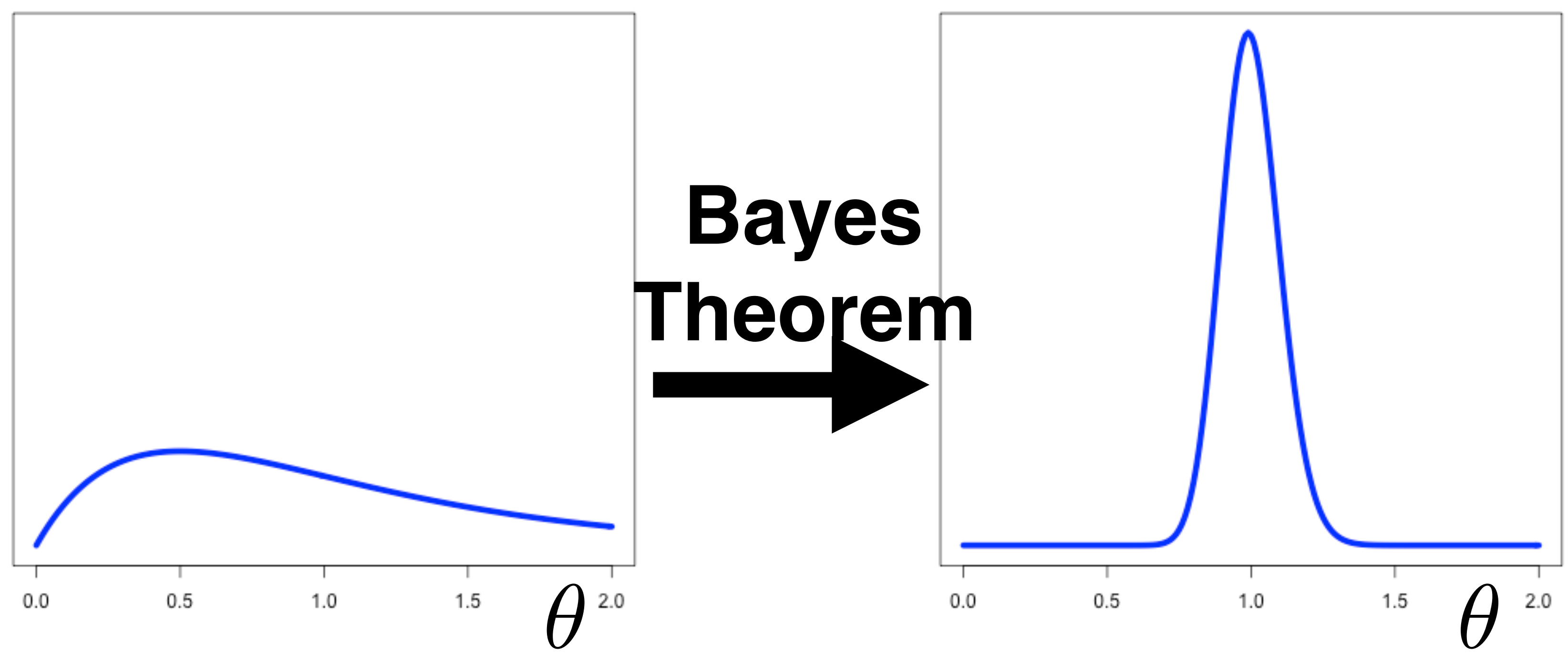
1. Build a model: choose prior & choose likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances

Bayesian inference

data → parameters

$$p(\theta|y_{1:N}) \propto_\theta p(y_{1:N}|\theta)p(\theta)$$

posterior likelihood prior



1. Build a model: choose prior & choose likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances

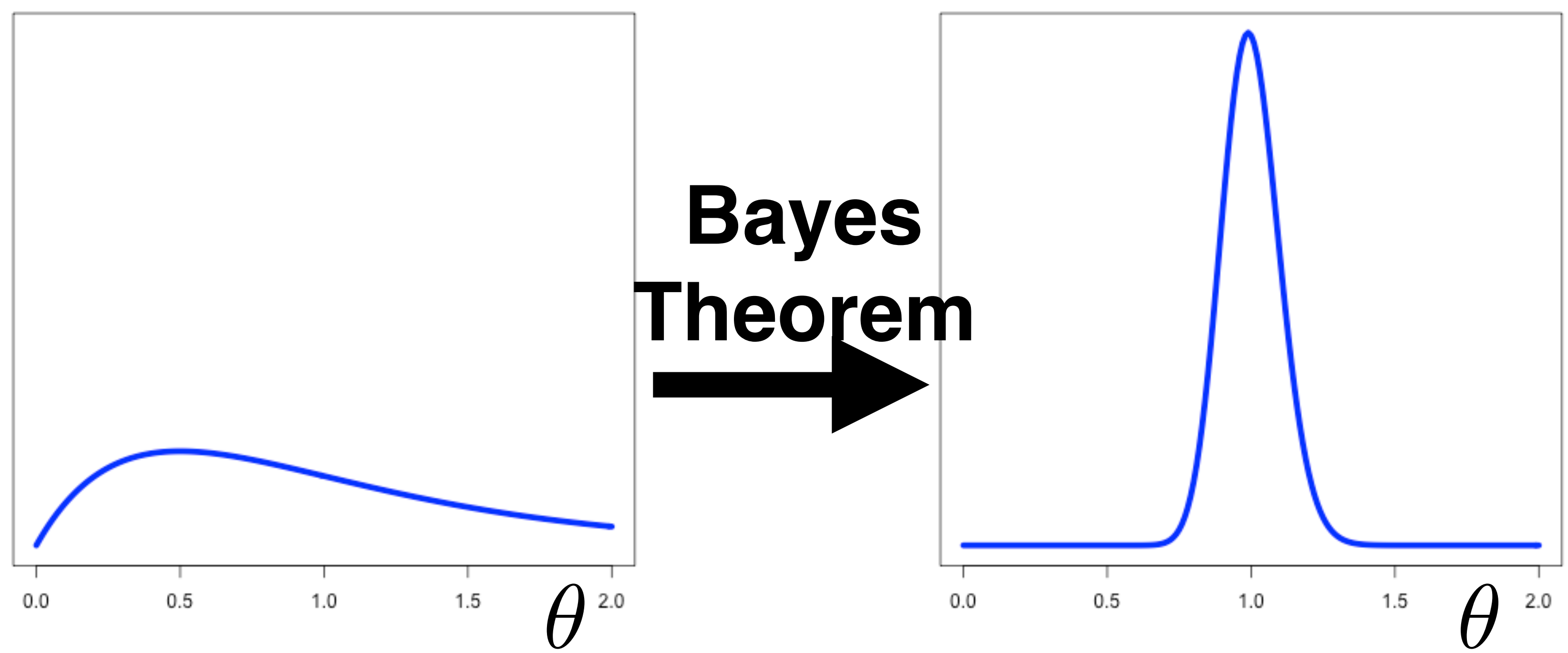
- Why are steps 2 and 3 hard?

Bayesian inference

data parameters

$$p(\theta|y_{1:N}) \propto_\theta p(y_{1:N}|\theta)p(\theta)$$

posterior likelihood prior



1. Build a model: choose prior & choose likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances

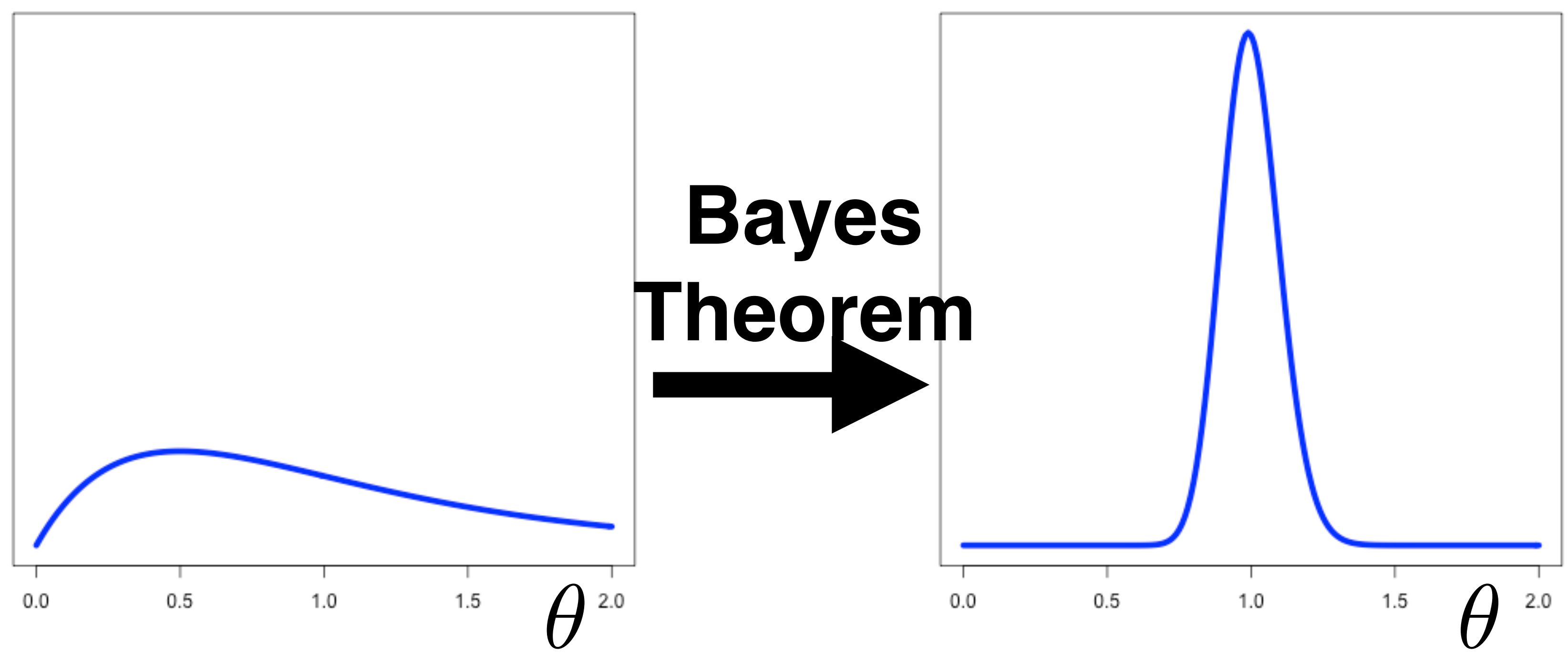
- Why are steps 2 and 3 hard?
  - Typically no closed form

Bayesian inference

data parameters

$$p(\theta|y_{1:N}) \propto_\theta p(y_{1:N}|\theta)p(\theta)$$

posterior likelihood prior



1. Build a model: choose prior & choose likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances

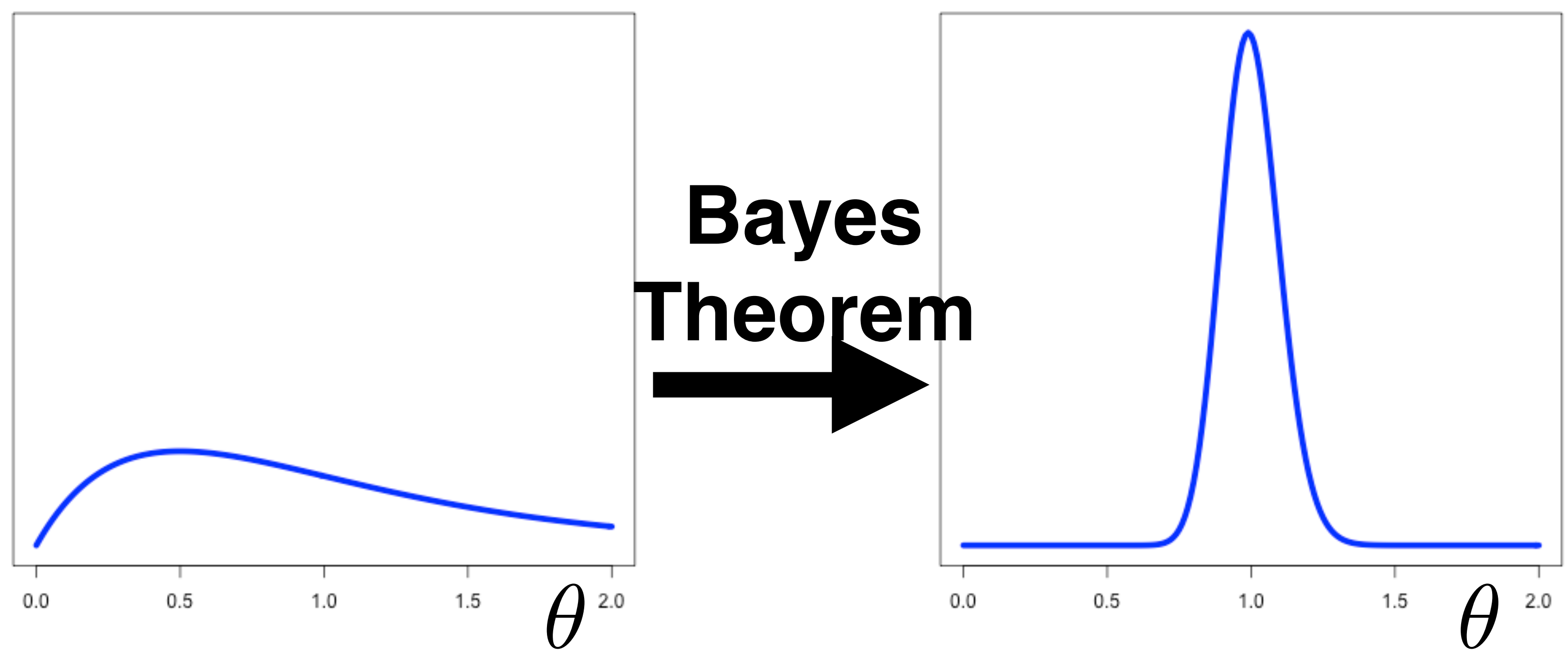
- Why are steps 2 and 3 hard?
  - Typically no closed form, high-dimensional integration

Bayesian inference

data parameters

$$p(\theta|y_{1:N}) = p(y_{1:N}|\theta)p(\theta)/p(y_{1:N})$$

posterior likelihood prior



1. Build a model: choose prior & choose likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances

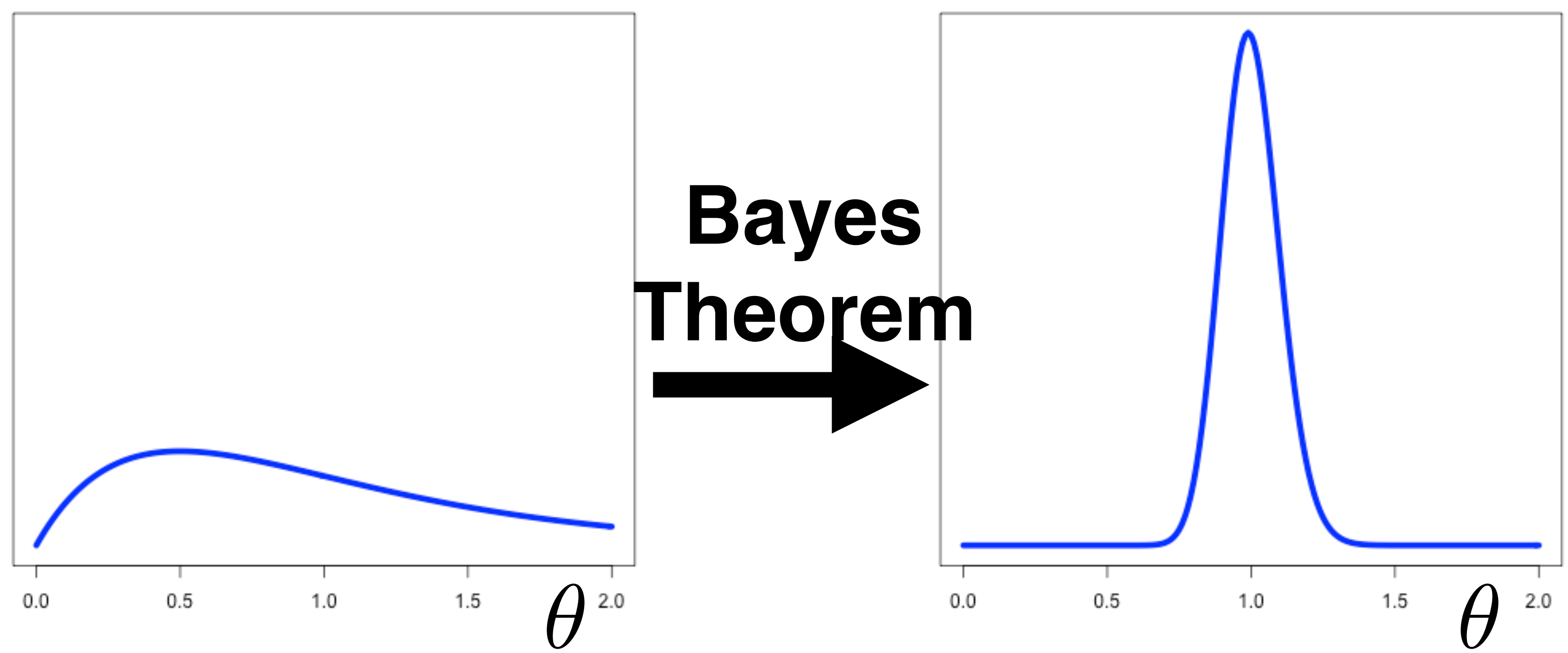
- Why are steps 2 and 3 hard?
  - Typically no closed form, high-dimensional integration

Bayesian inference

data parameters

$$p(\theta|y_{1:N}) = p(y_{1:N}|\theta)p(\theta)/p(y_{1:N})$$

posterior likelihood prior evidence



1. Build a model: choose prior & choose likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances

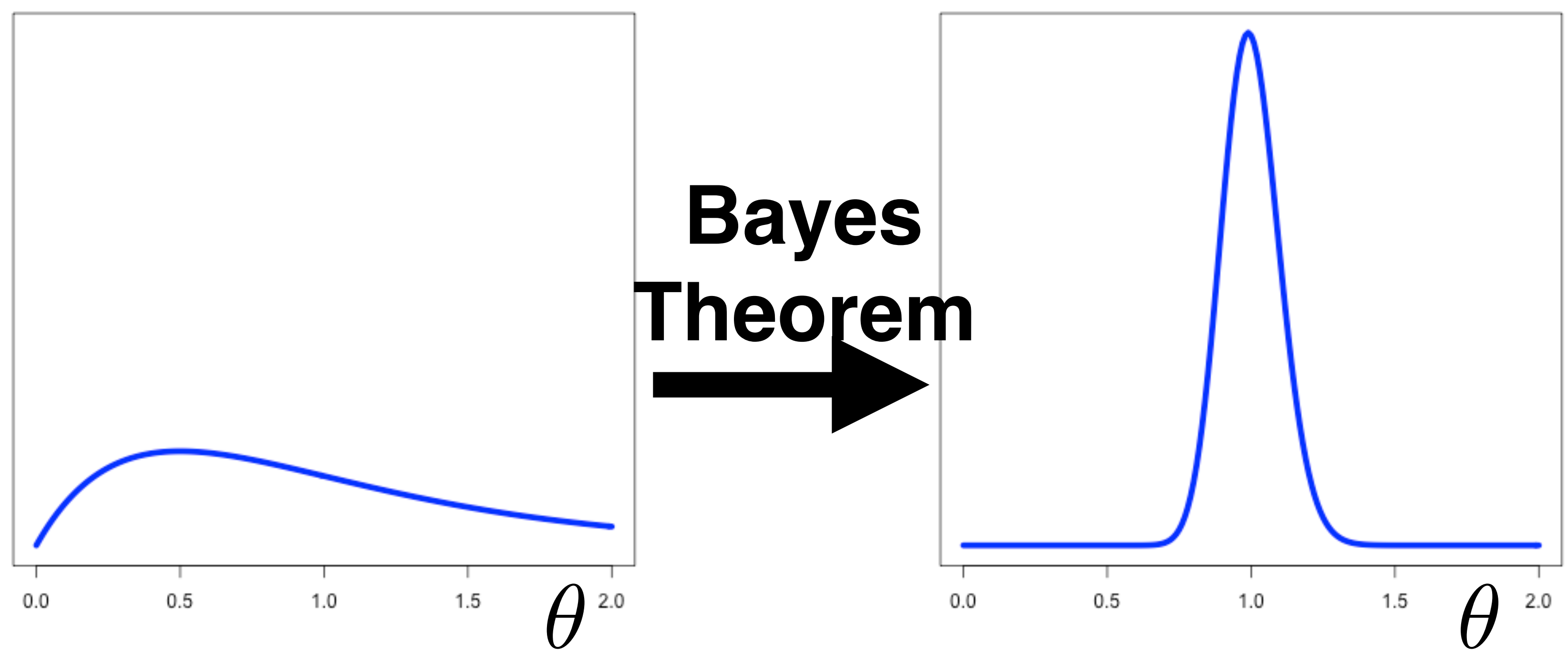
- Why are steps 2 and 3 hard?
  - Typically no closed form, high-dimensional integration

Bayesian inference

data parameters

$$p(\theta|y_{1:N}) = p(y_{1:N}|\theta)p(\theta) / \int p(y_{1:N}, \theta) d\theta$$

posterior likelihood prior evidence



1. Build a model: choose prior & choose likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances

- Why are steps 2 and 3 hard?
  - Typically no closed form, high-dimensional integration

Approximate Bayesian Inference

Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)

[Bardenet, Doucet, Holmes 2017]

Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

[Bardenet, Doucet, Holmes 2017]

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

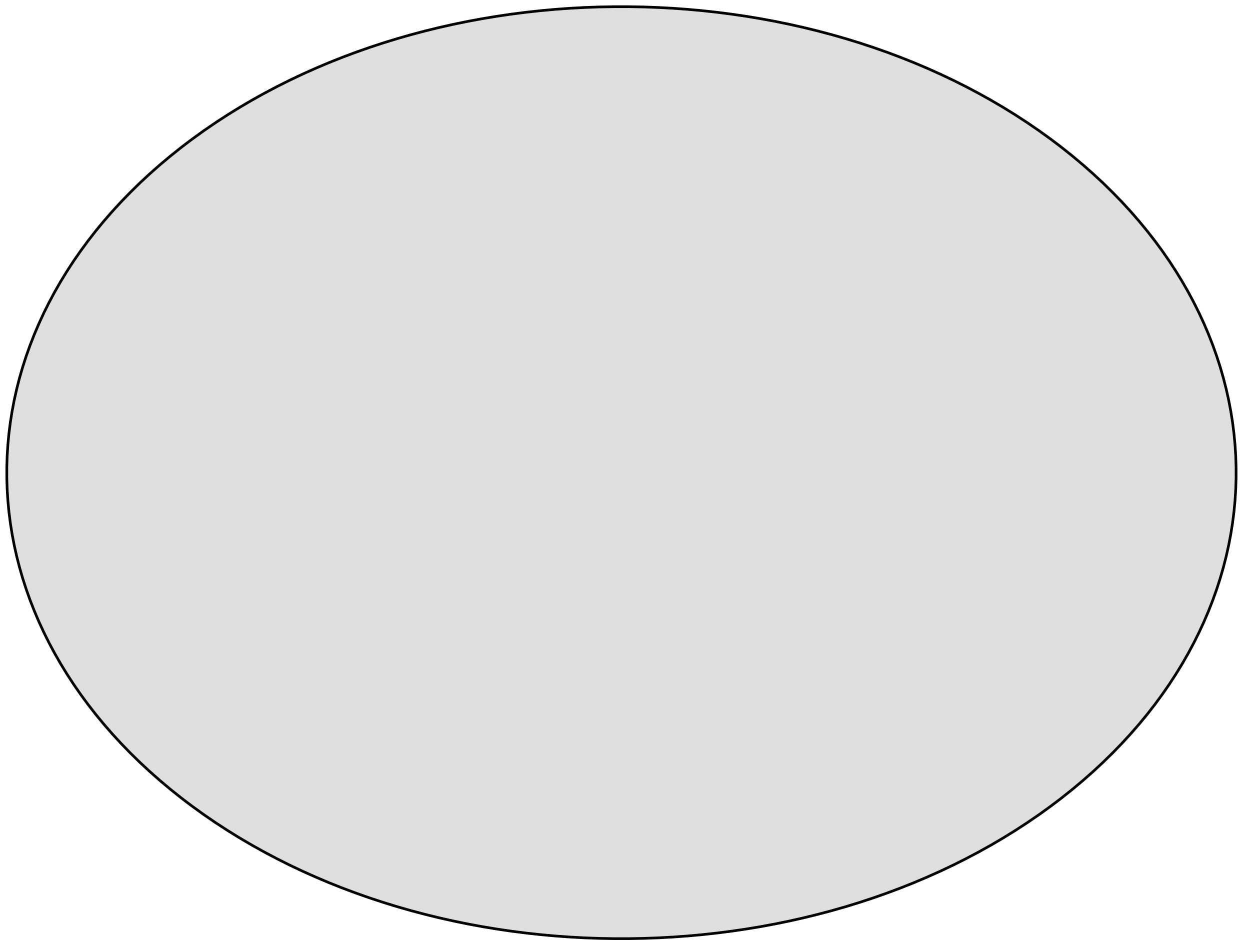
Instead: an optimization approach

- Approximate posterior with q^*

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow



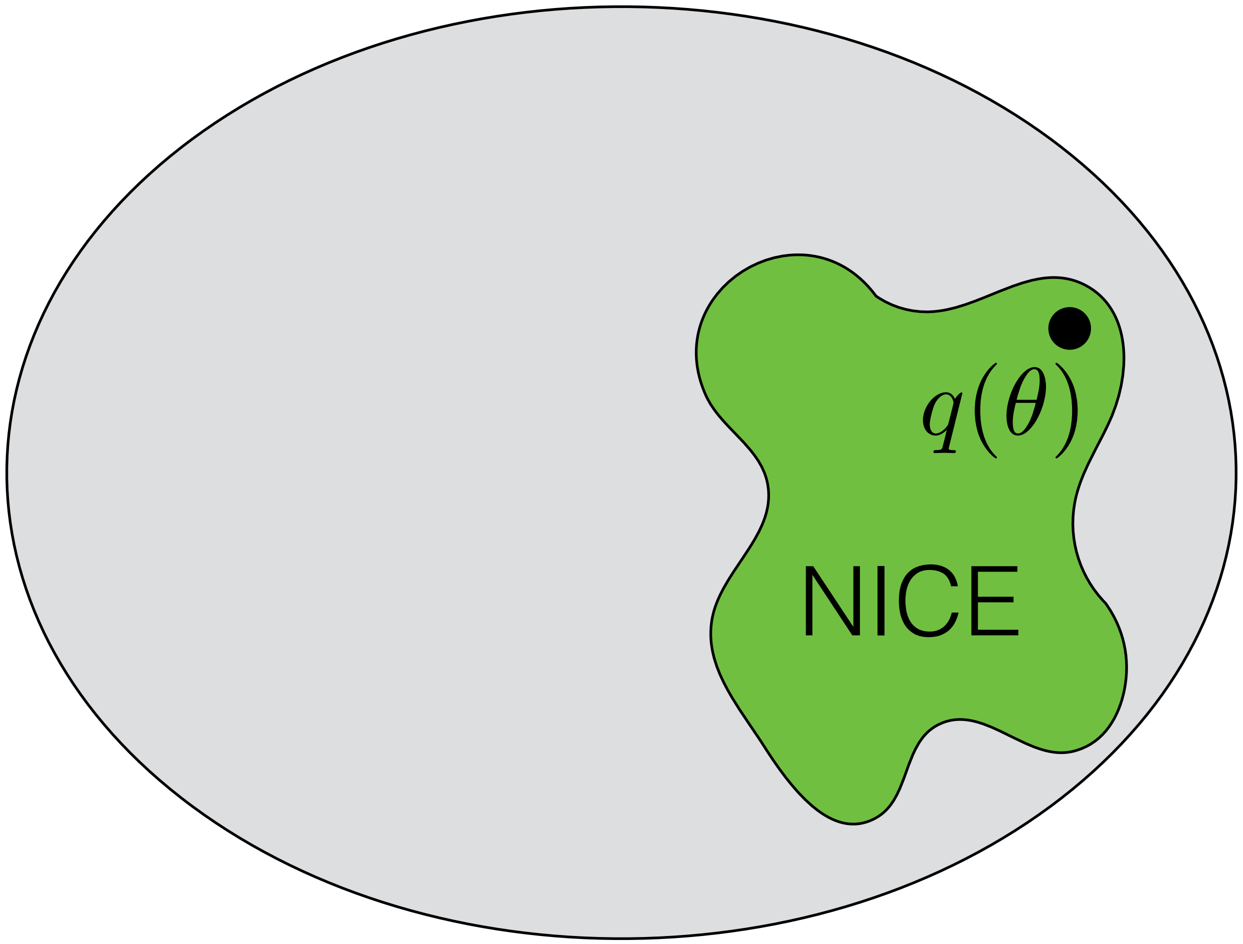
Instead: an optimization approach

- Approximate posterior with q^*

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow



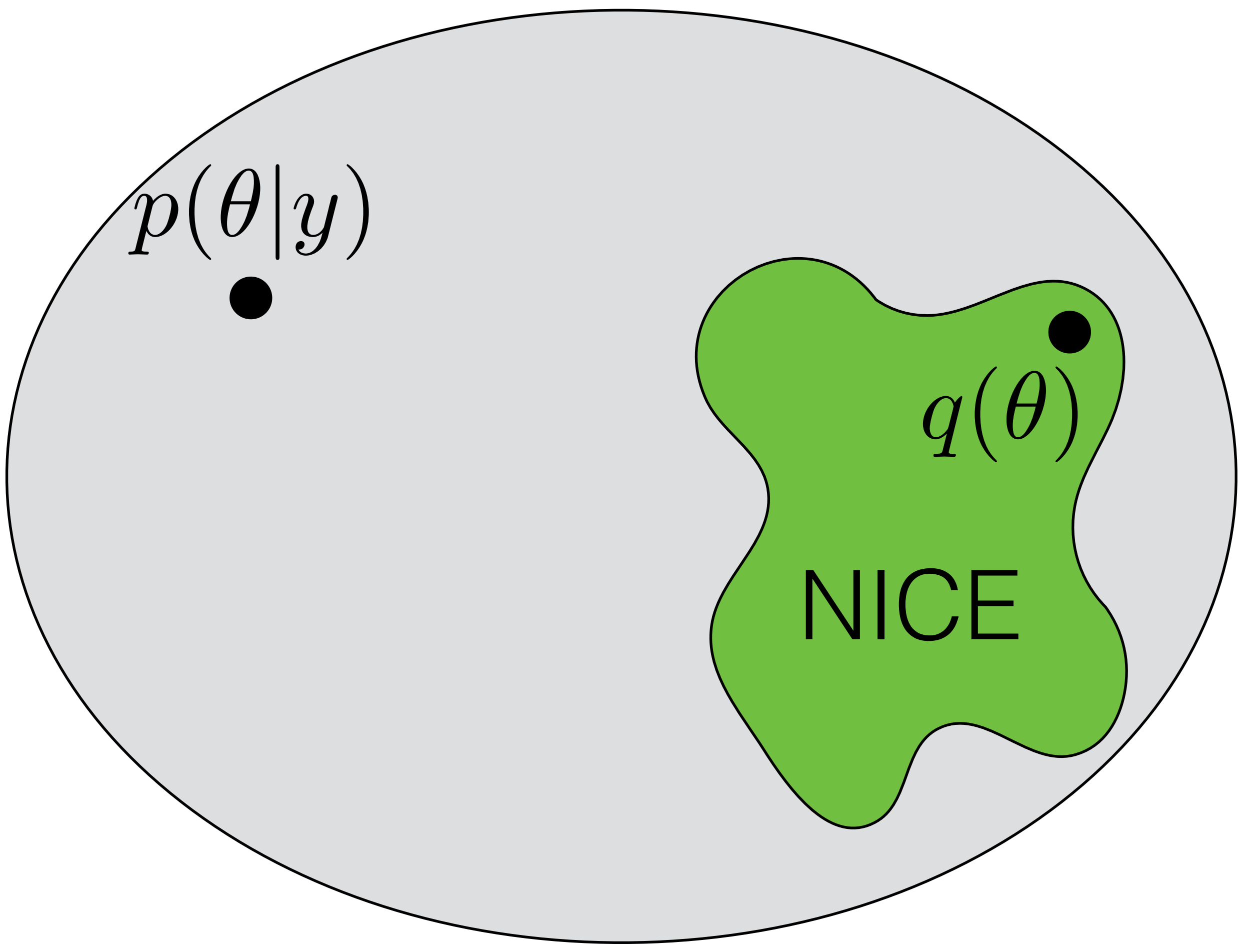
Instead: an optimization approach

- Approximate posterior with q^*

Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

[Bardenet, Doucet, Holmes 2017]



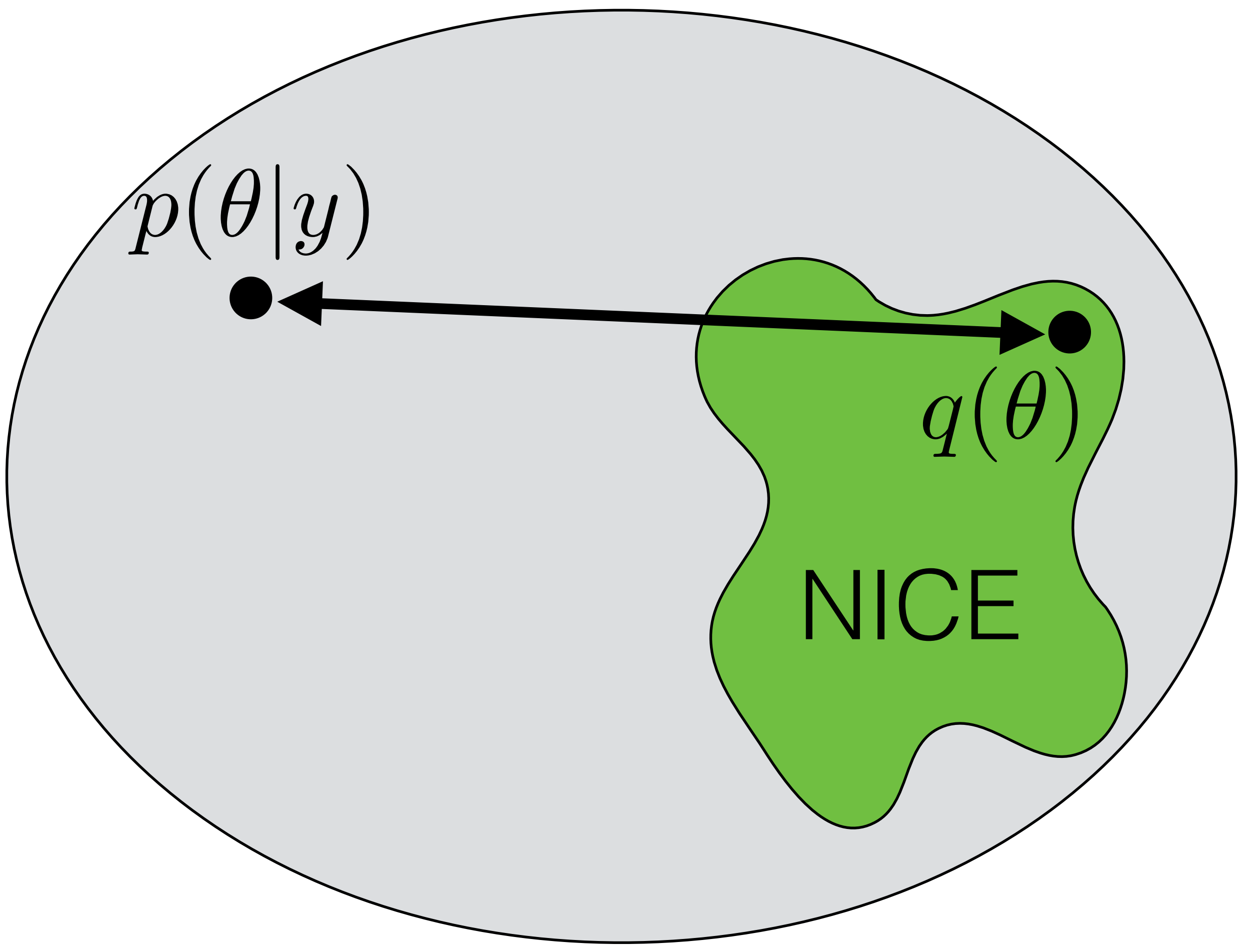
Instead: an optimization approach

- Approximate posterior with q^*

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow



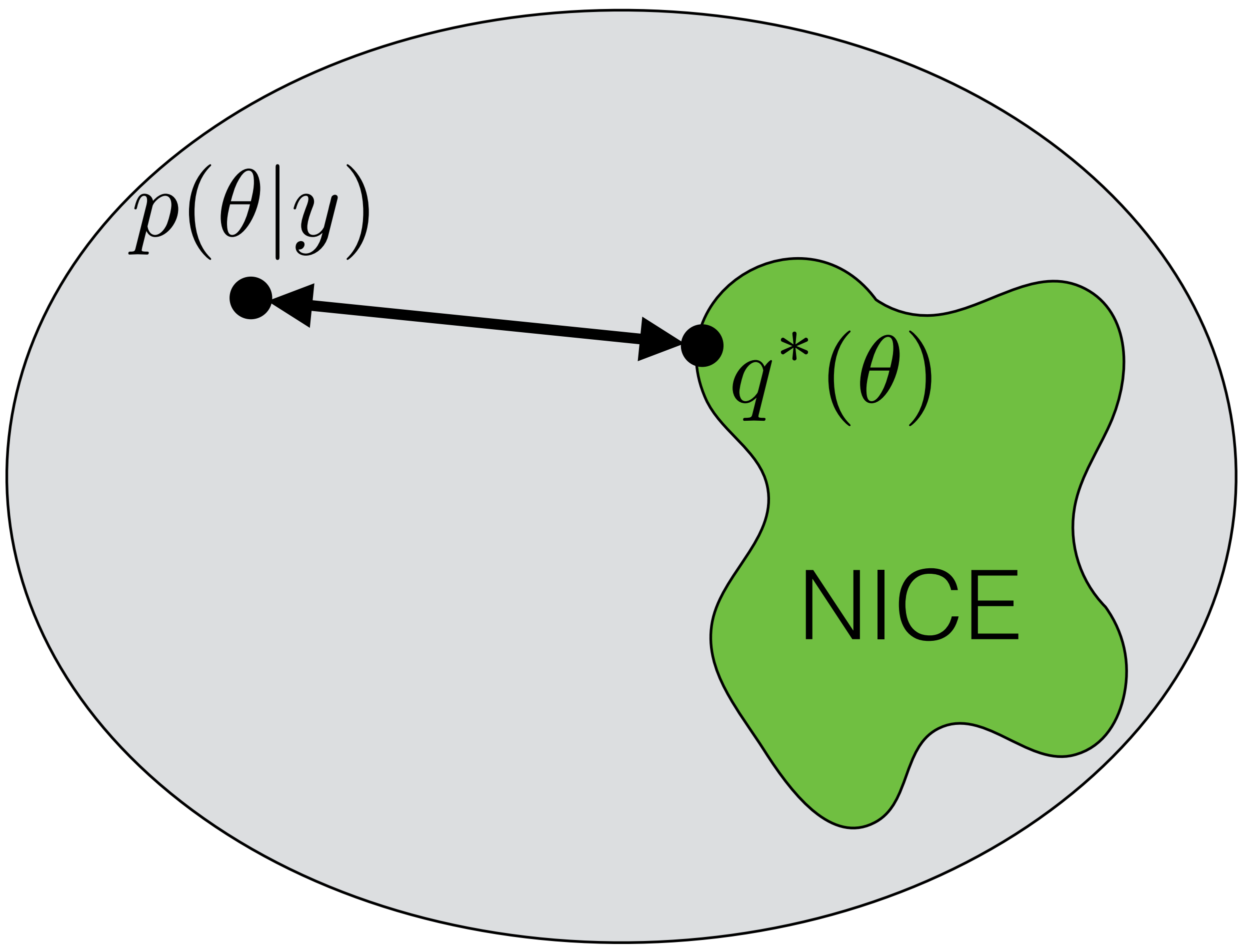
Instead: an optimization approach

- Approximate posterior with q^*

Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

[Bardenet, Doucet, Holmes 2017]



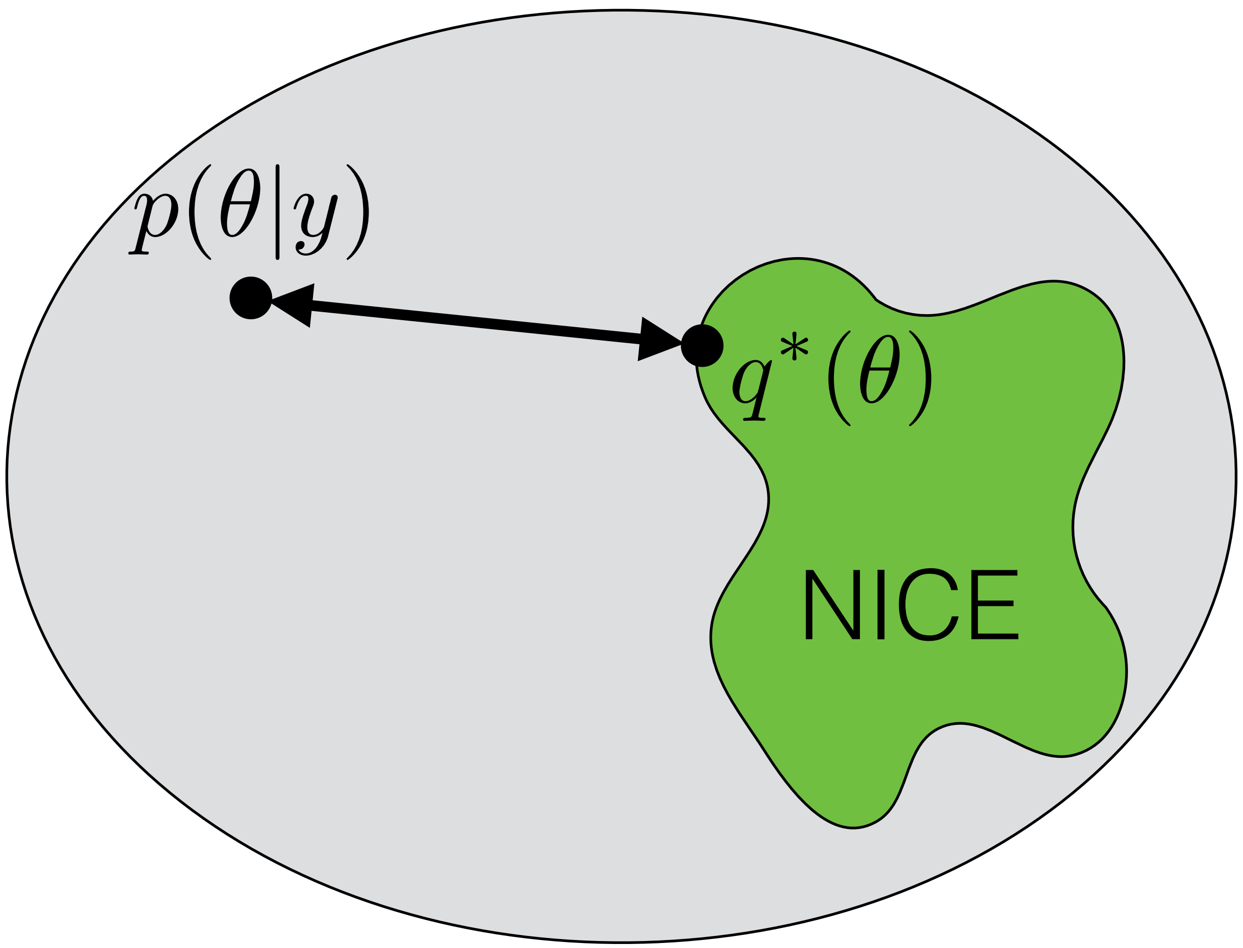
Instead: an optimization approach

- Approximate posterior with q^*

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow



Instead: an optimization approach

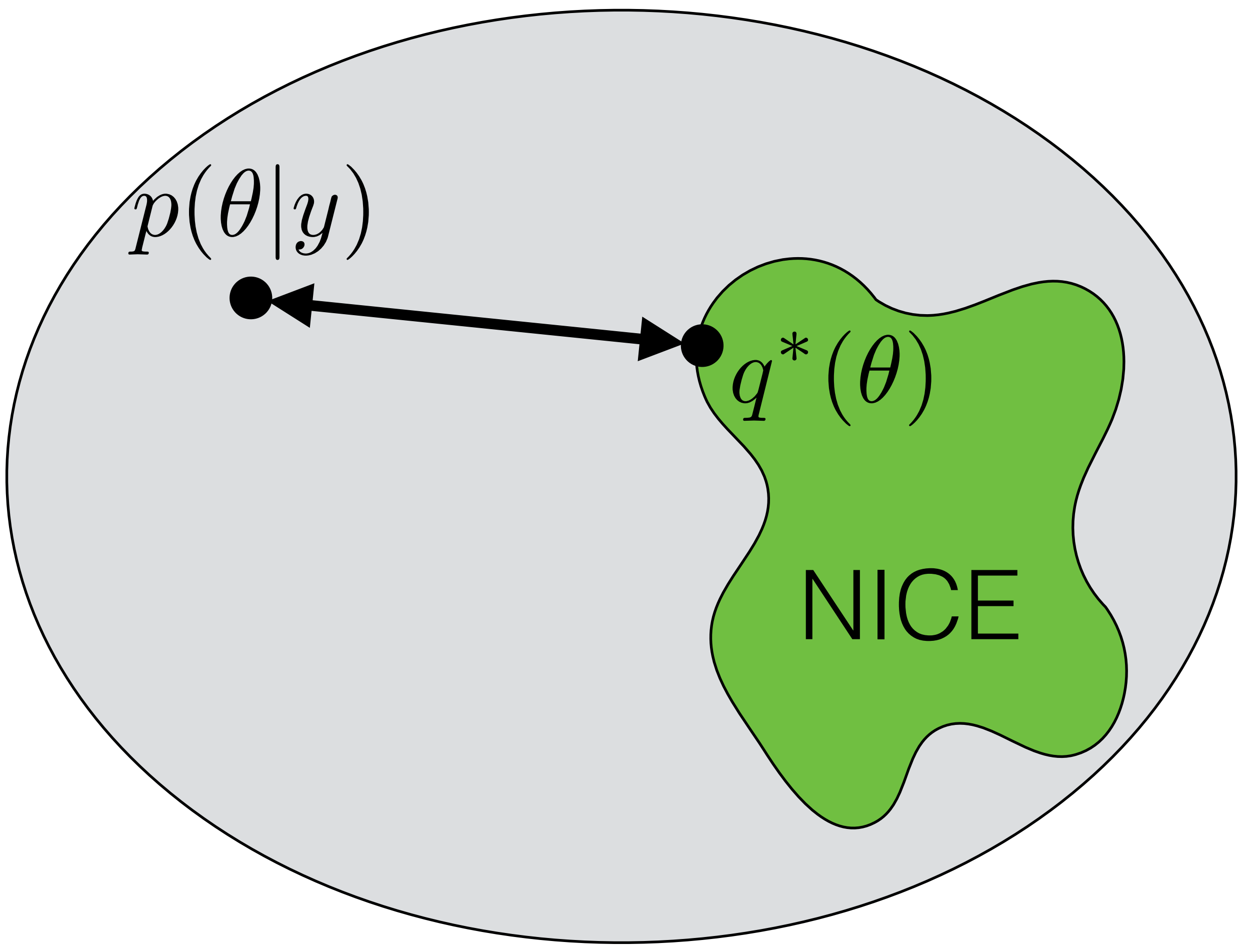
- Approximate posterior with q^*

$$q^* = \text{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

[Bardenet, Doucet, Holmes 2017]



Instead: an optimization approach

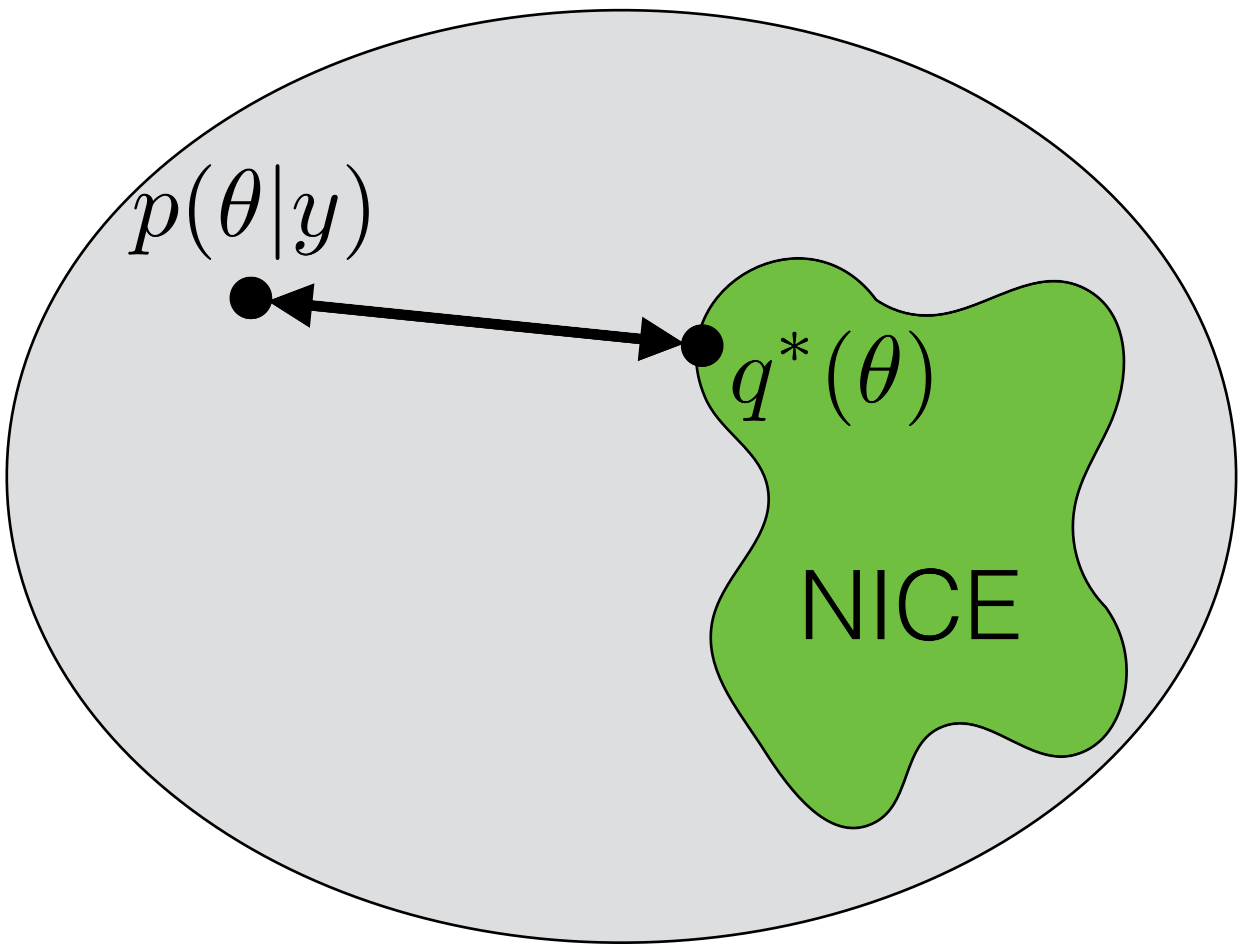
- Approximate posterior with q^*

$$q^* = \mathrm{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

[Bardenet, Doucet, Holmes 2017]



Instead: an optimization approach

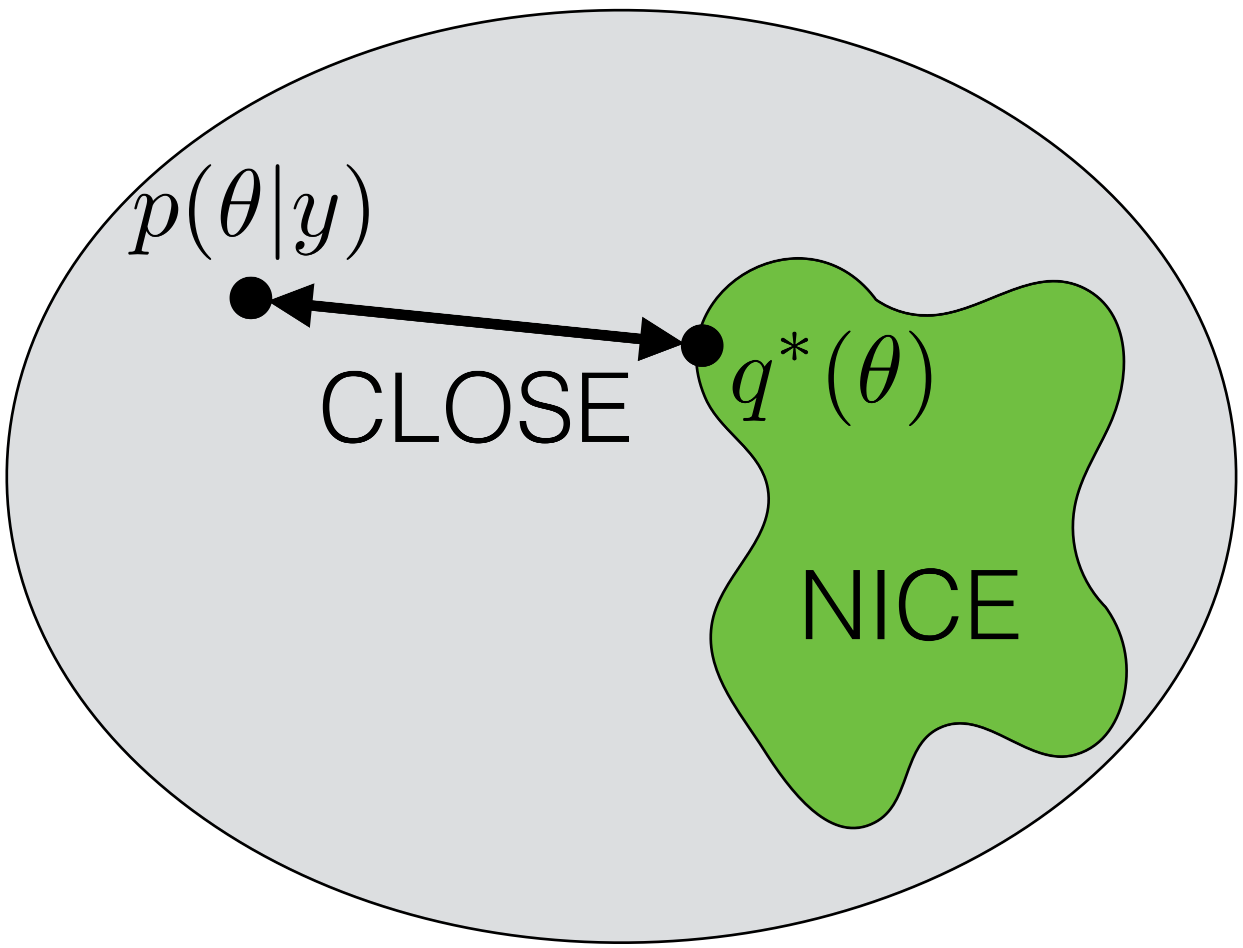
- Approximate posterior with q^*

$$q^* = \text{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow



Instead: an optimization approach

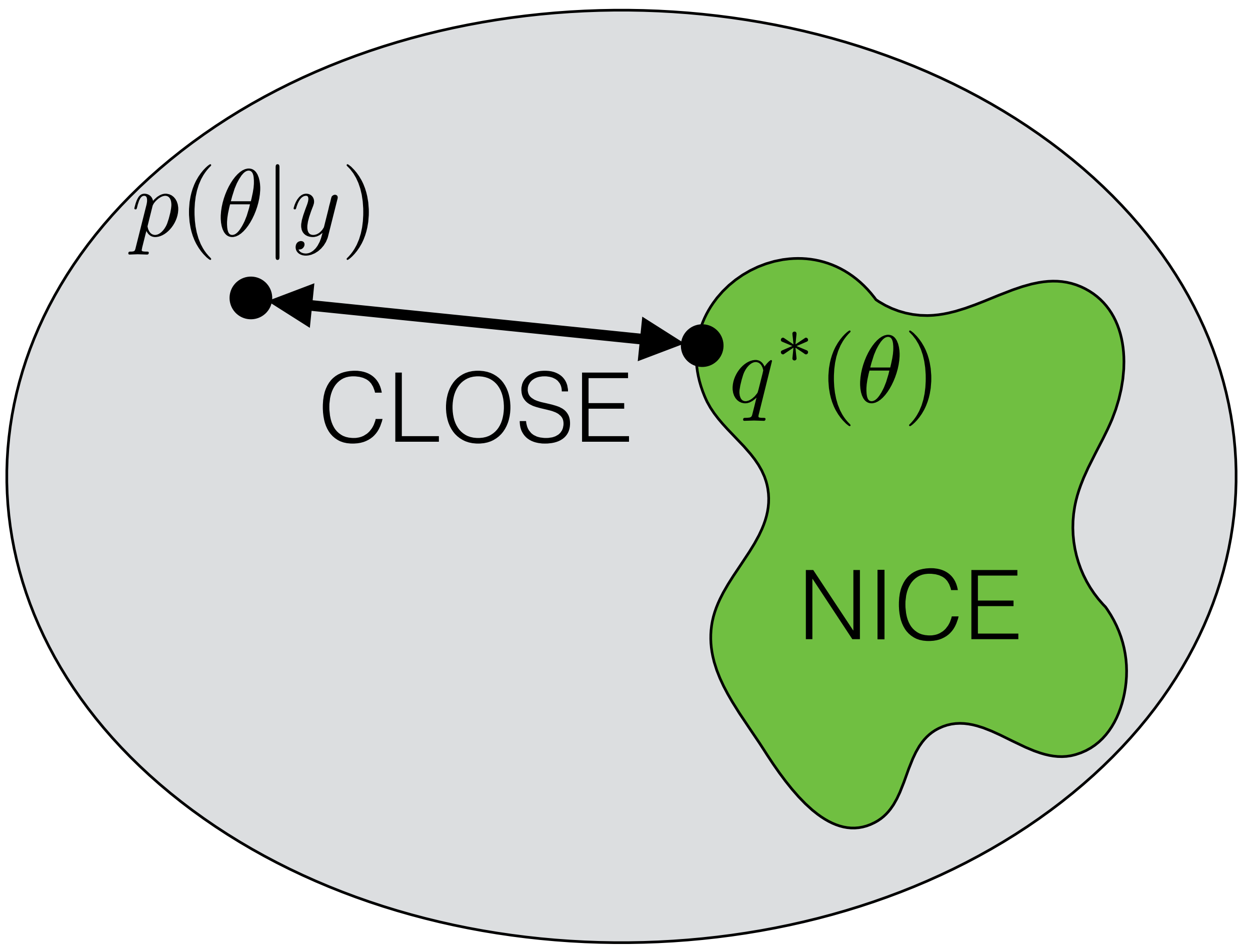
- Approximate posterior with q^*

$$q^* = \text{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow



Instead: an optimization approach

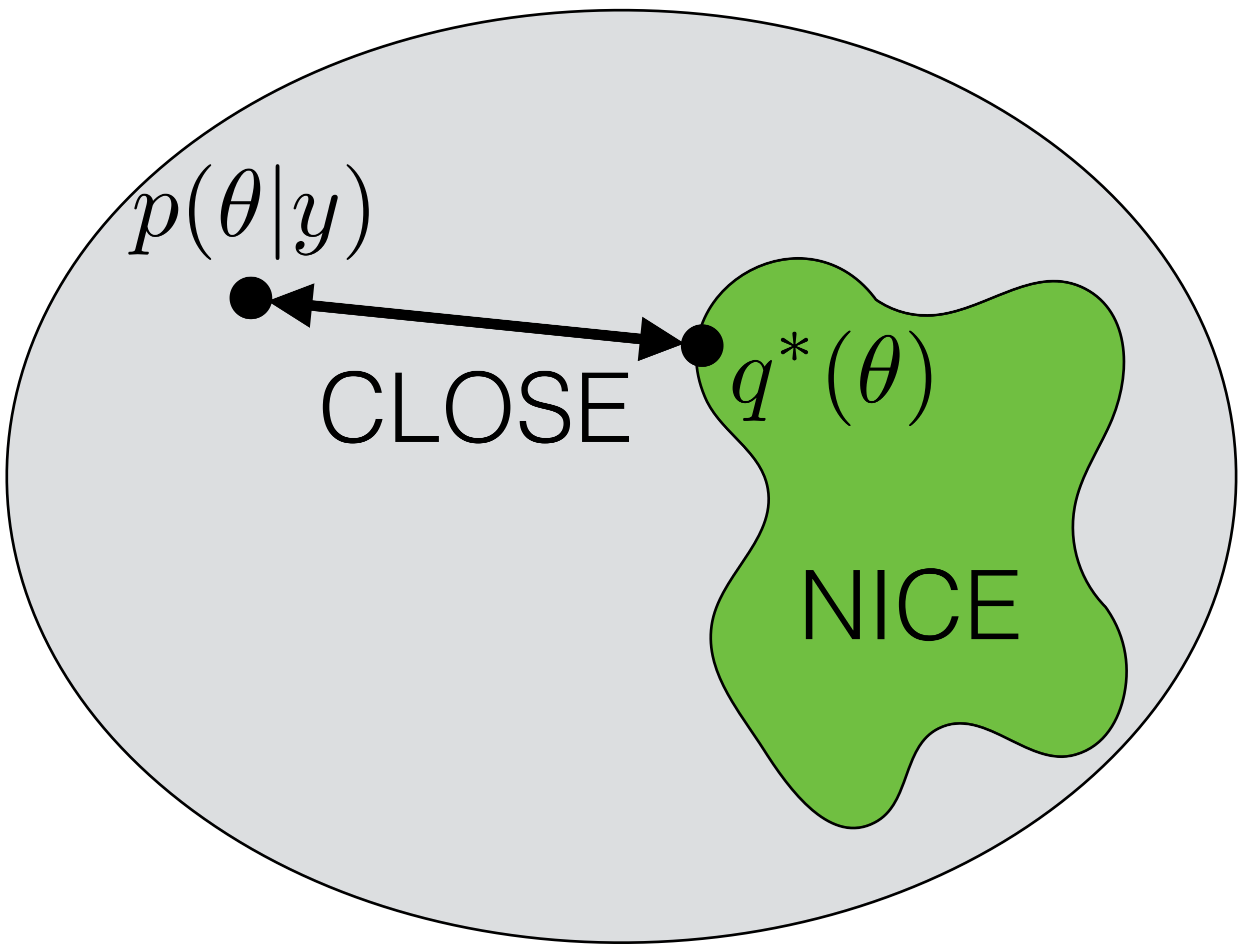
- Approximate posterior with q^*

$$q^* = \text{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow



Instead: an optimization approach

- Approximate posterior with q^*

$$q^* = \text{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

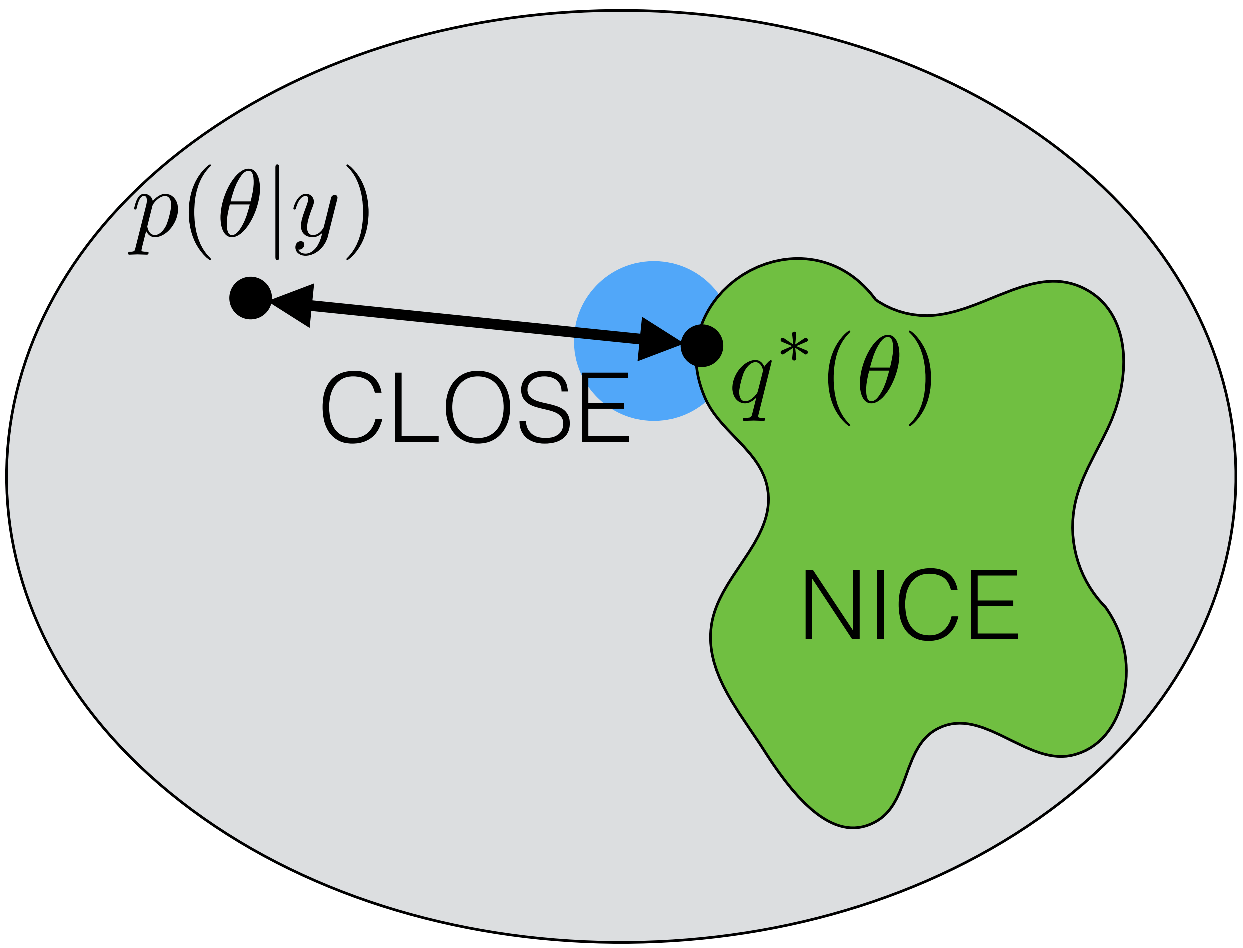
- Variational Bayes (VB): *f* is Kullback-Leibler divergence

$$KL(q(\cdot)||p(\cdot|y))$$

Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

[Bardenet, Doucet, Holmes 2017]



Instead: an optimization approach

- Approximate posterior with q^*

$$q^* = \text{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

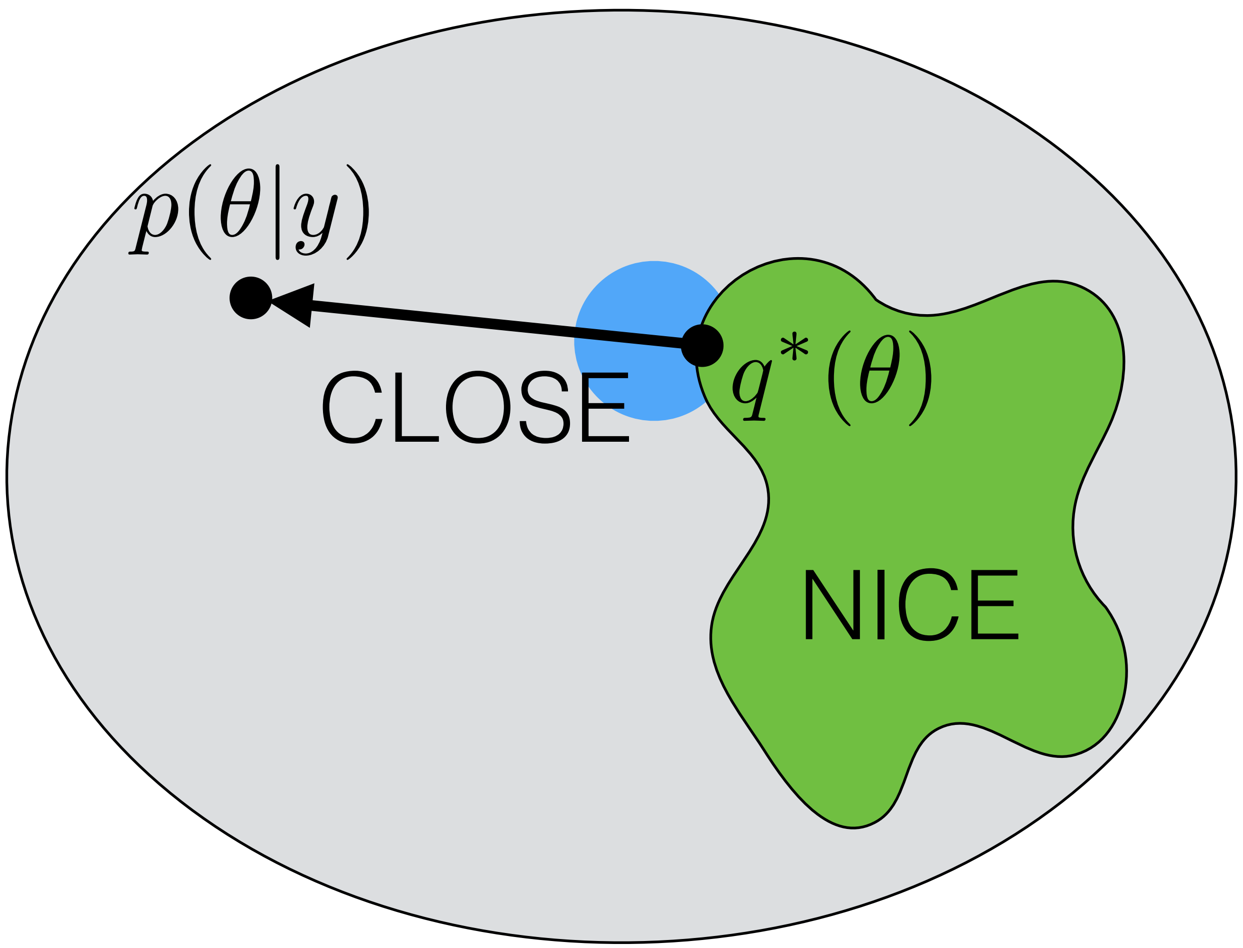
- Variational Bayes (VB): *f* is Kullback-Leibler divergence

$$KL(q(\cdot)||p(\cdot|y))$$

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow



Instead: an optimization approach

- Approximate posterior with q^*

$$q^* = \text{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

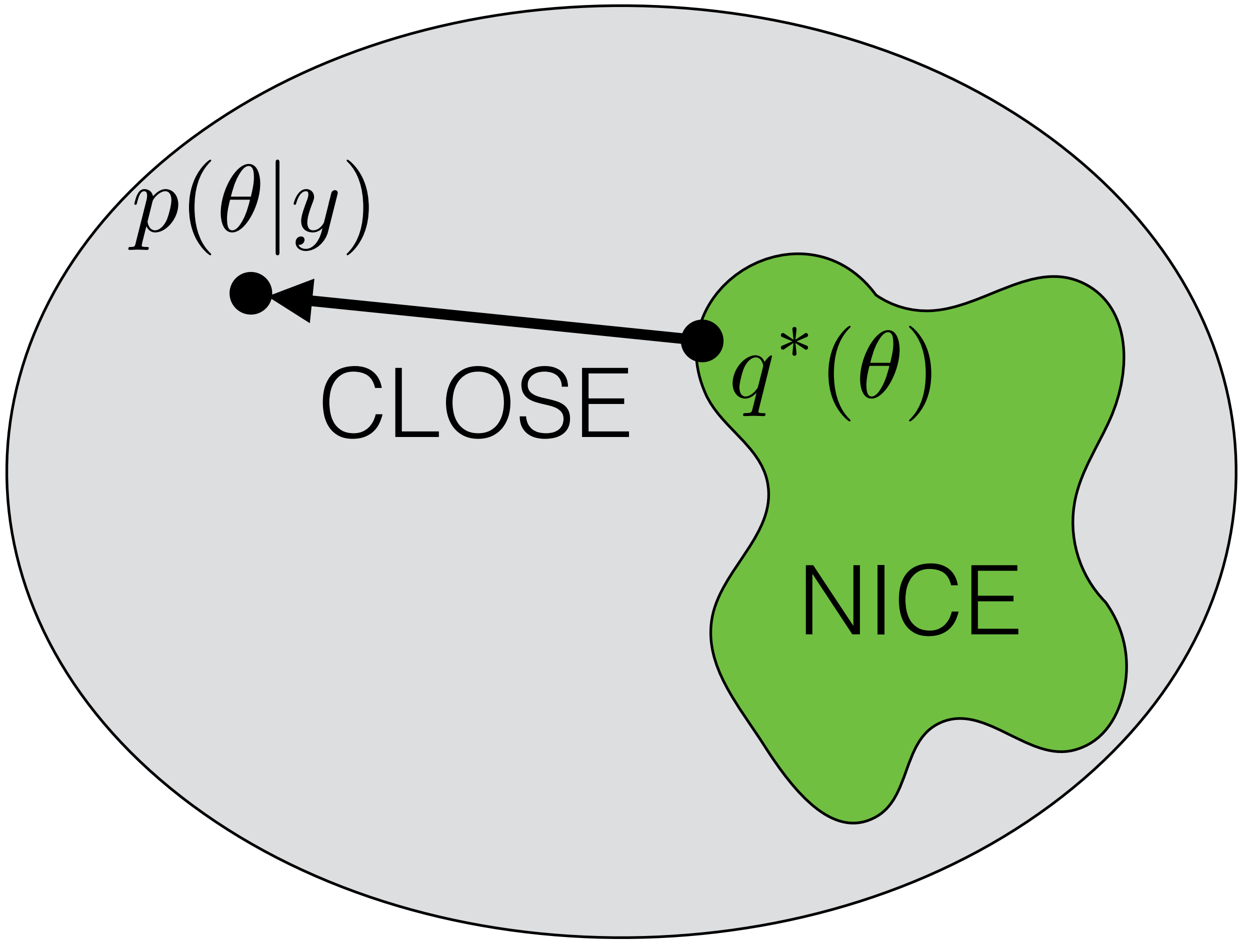
- Variational Bayes (VB): f is Kullback-Leibler divergence

$$KL(q(\cdot)||p(\cdot|y))$$

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow



Instead: an optimization approach

- Approximate posterior with *q**

$$q^* = \text{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

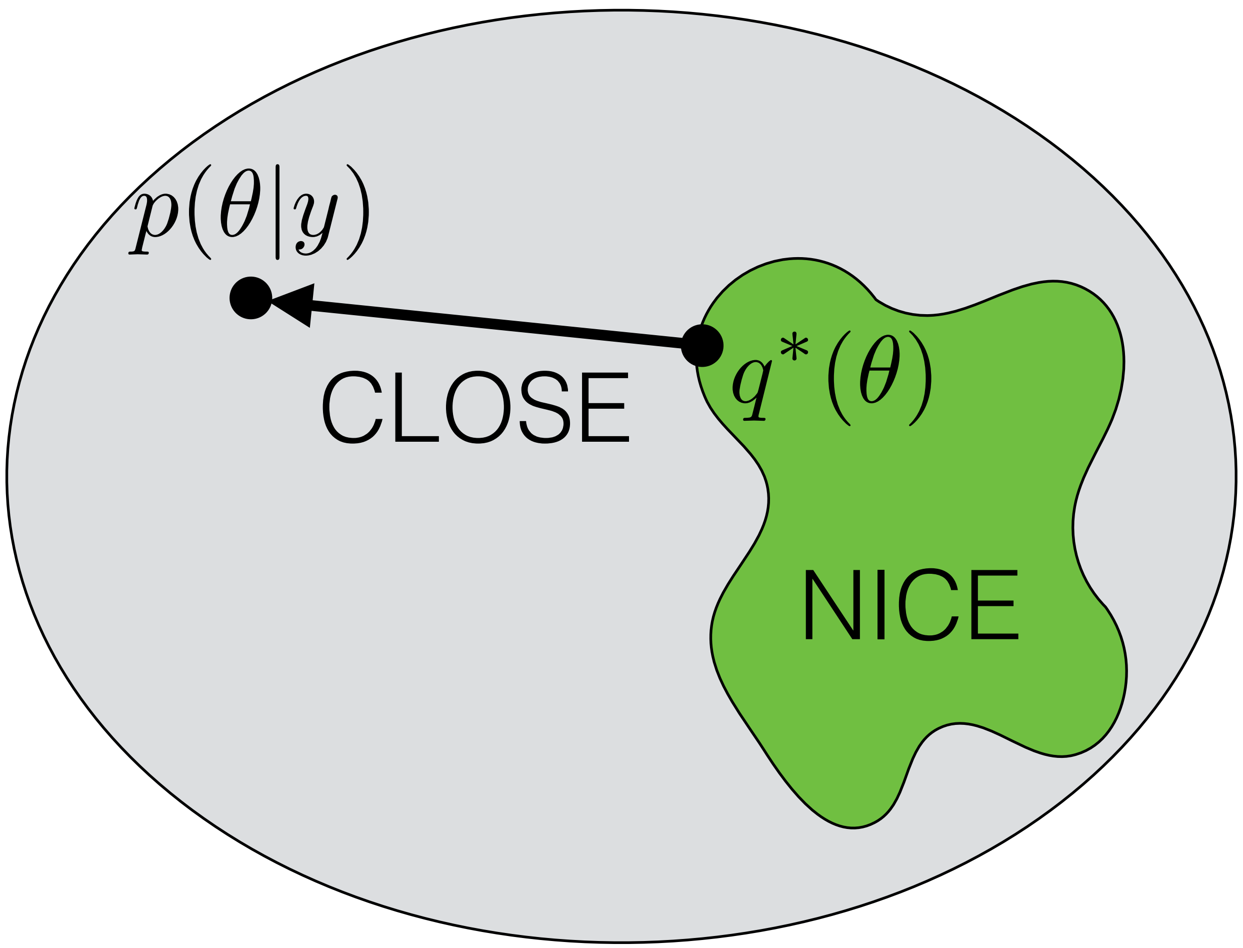
- Variational Bayes (VB): *f* is Kullback-Leibler divergence

$$KL(q(\cdot)||p(\cdot|y))$$

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow



Instead: an optimization approach

- Approximate posterior with q^*

$$q^* = \mathrm{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

- Variational Bayes (VB): *f* is Kullback-Leibler divergence

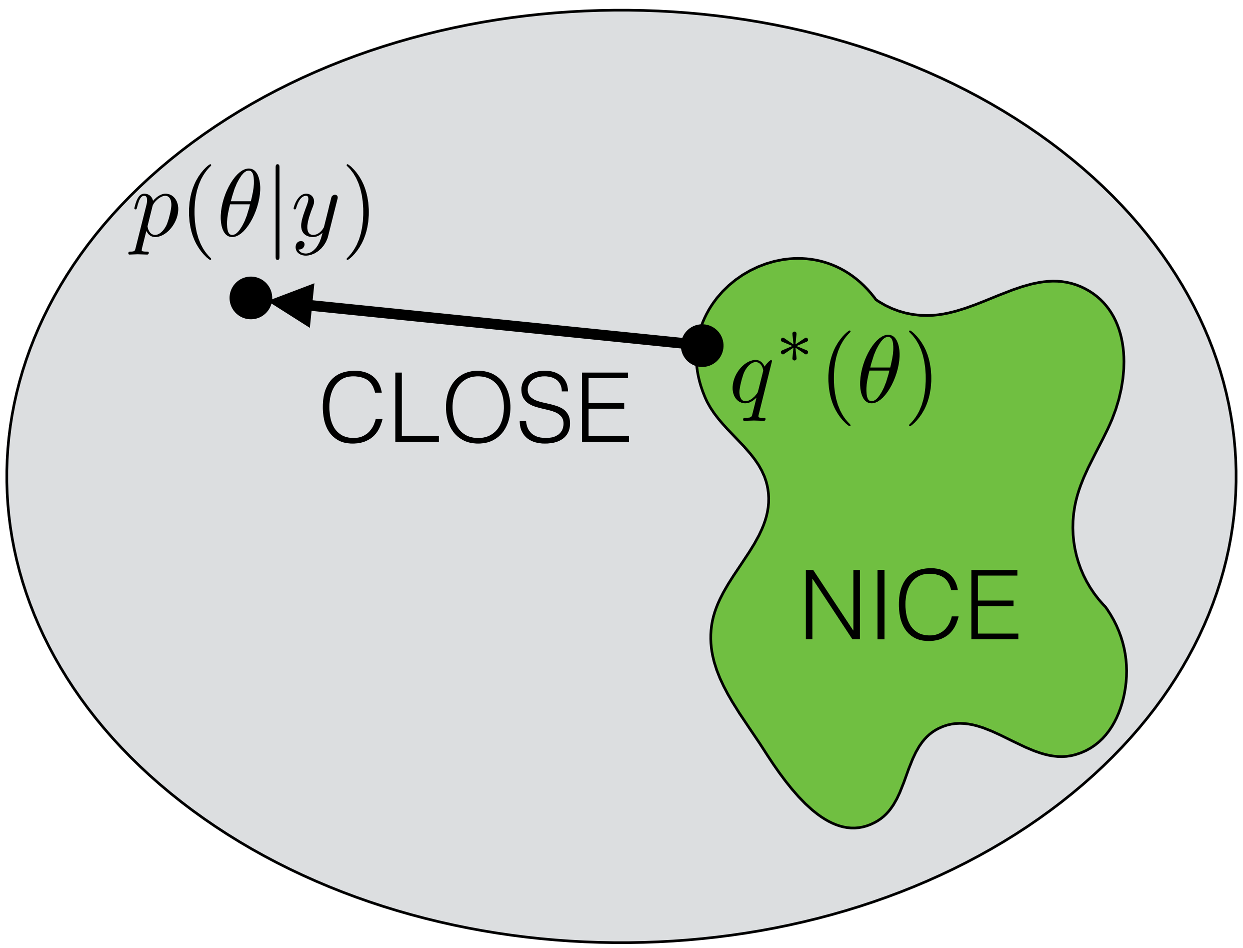
$$KL(q(\cdot)||p(\cdot|y))$$

- VB practical success

Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

[Bardenet, Doucet, Holmes 2017]



Instead: an optimization approach

- Approximate posterior with q^*

$$q^* = \text{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

- Variational Bayes (VB): *f* is Kullback-Leibler divergence

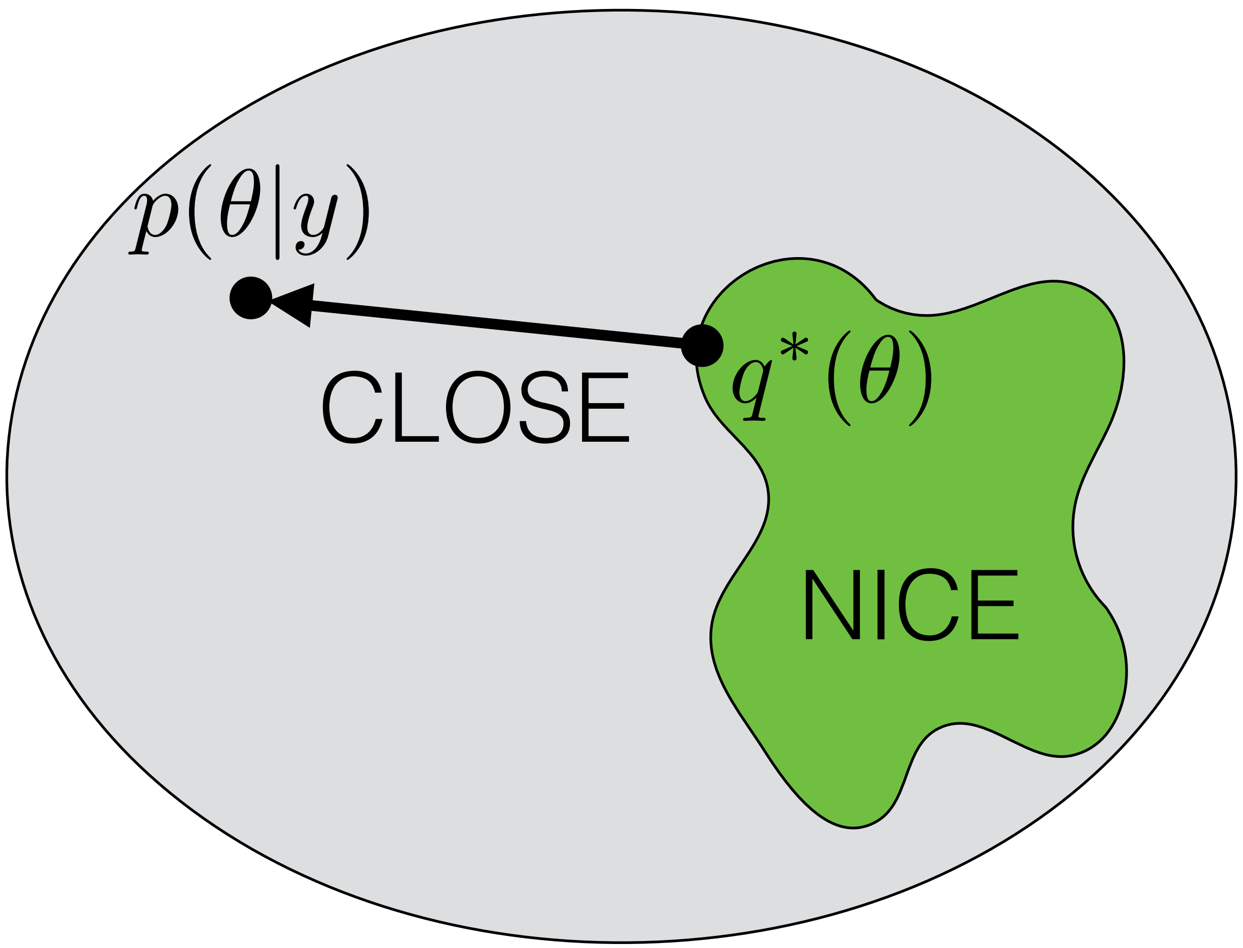
$$KL(q(\cdot)||p(\cdot|y))$$

- VB practical success: point estimates and prediction

Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

[Bardenet, Doucet, Holmes 2017]



Instead: an optimization approach

- Approximate posterior with q^*

$$q^* = \text{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

- Variational Bayes (VB): *f* is Kullback-Leibler divergence

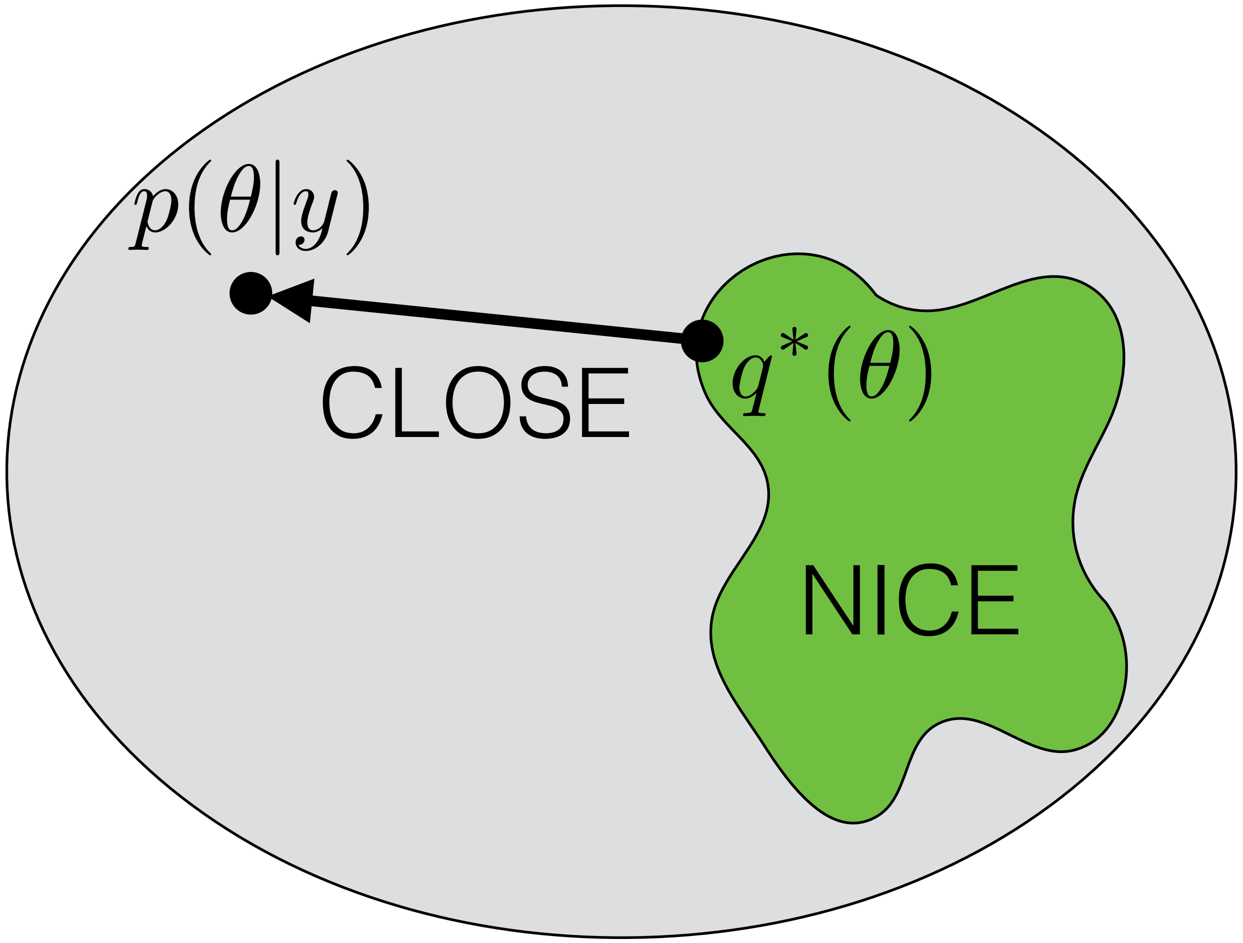
$$KL(q(\cdot)||p(\cdot|y))$$

- VB practical success: point estimates and prediction, fast

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow



Instead: an optimization approach

- Approximate posterior with q^*

$$q^* = \text{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

- Variational Bayes (VB): *f* is Kullback-Leibler divergence

$$KL(q(\cdot)||p(\cdot|y))$$

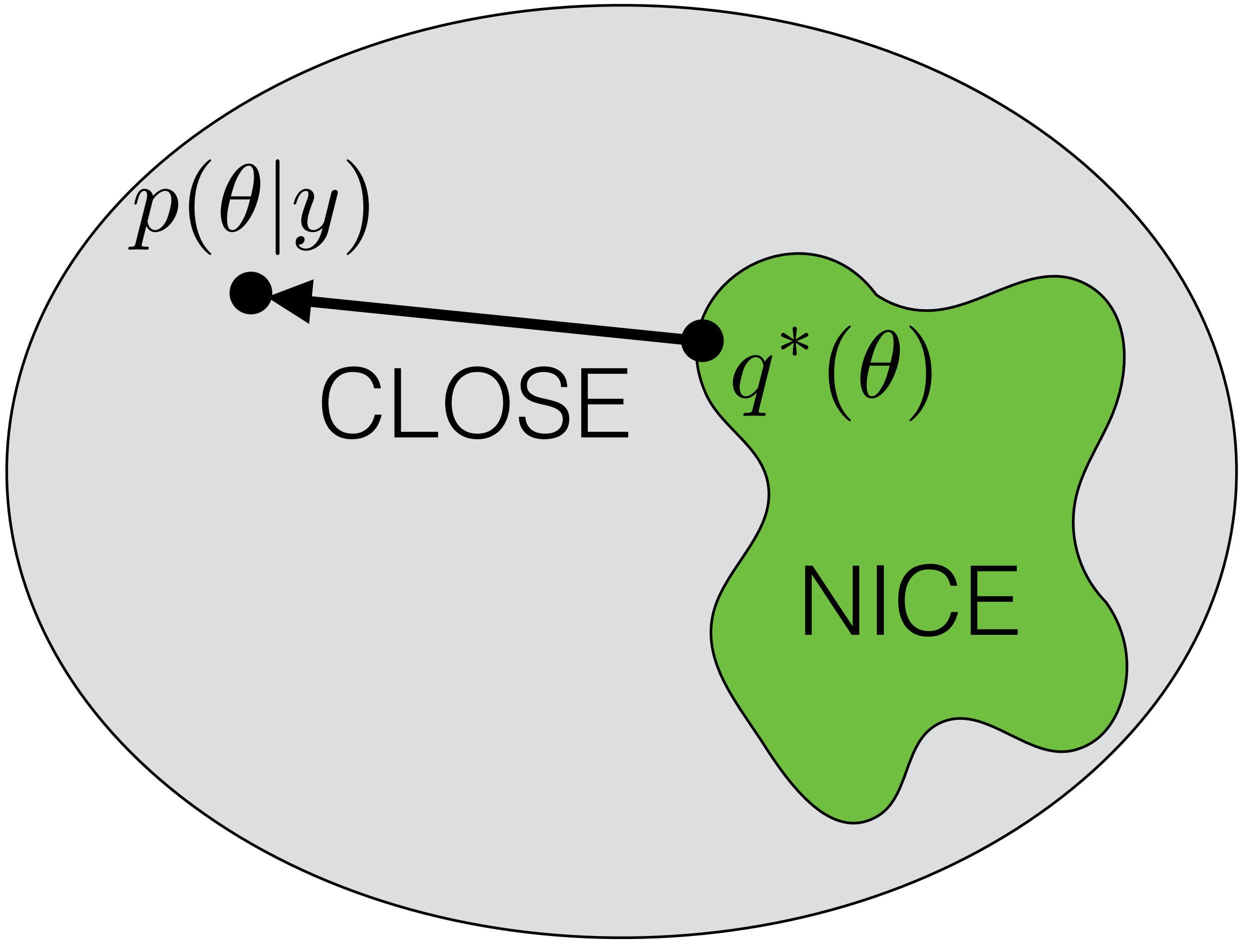
- VB practical success: point estimates and prediction, fast, streaming, distributed (3.6M Wikipedia, 350K Nature)

[Broderick, Boyd, Wibisono, Wilson, Jordan 2013]

Why KL?

- Variational Bayes

$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$

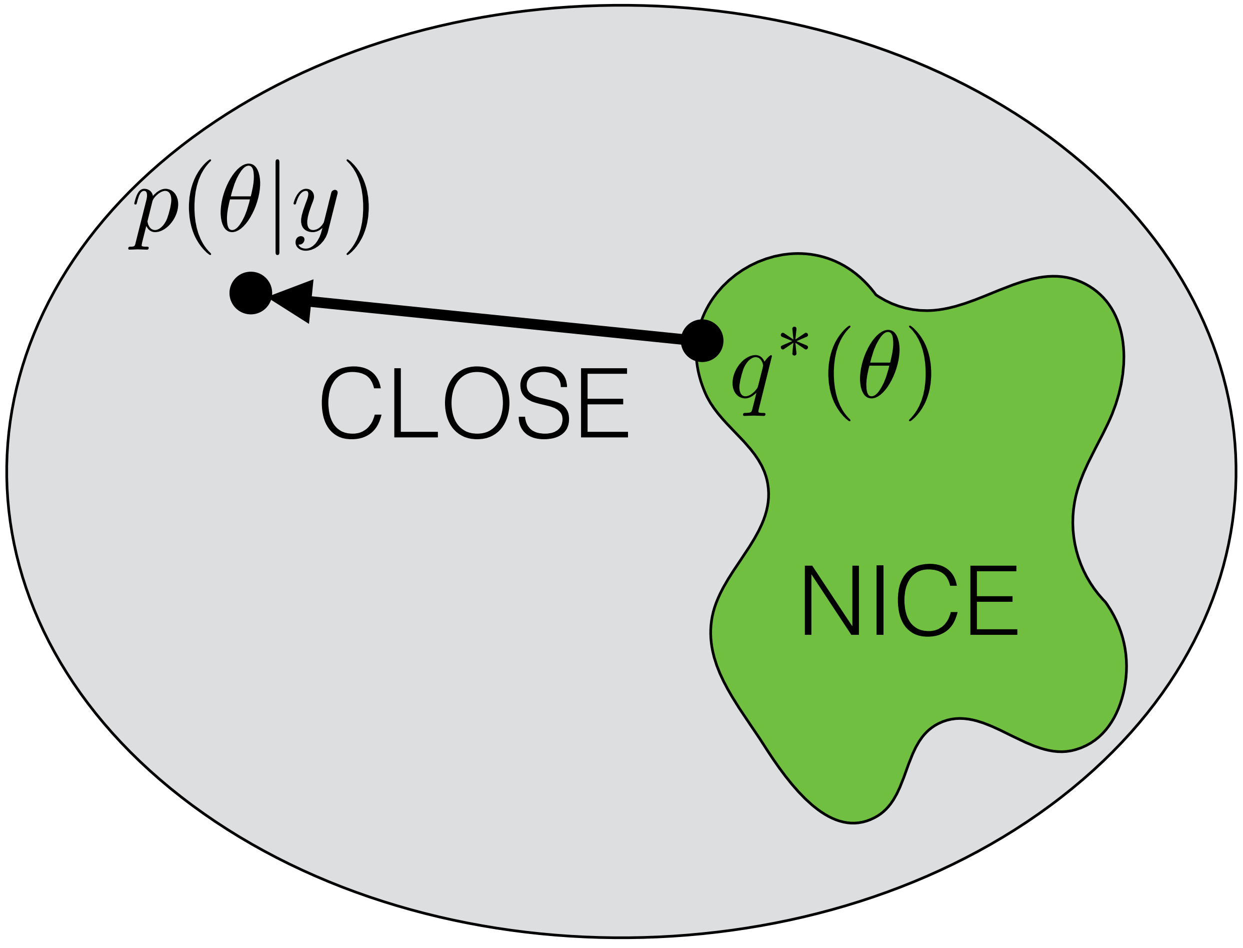


Why KL?

- Variational Bayes

$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot | y)\right)$$

$$\text{KL}\left(q(\cdot) || p(\cdot | y)\right)$$
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$



Why KL?

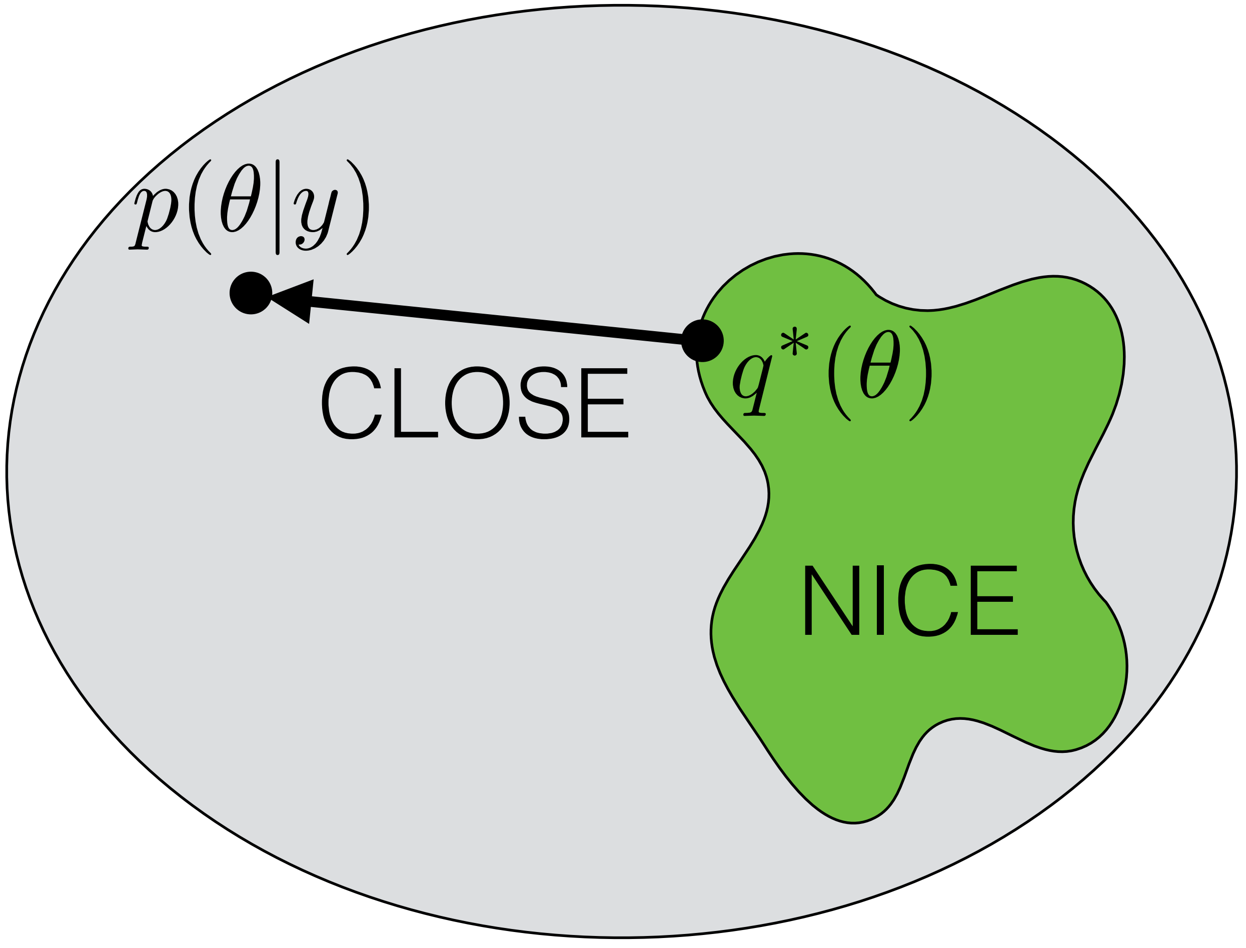
- Variational Bayes

$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$

$$\text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$

$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

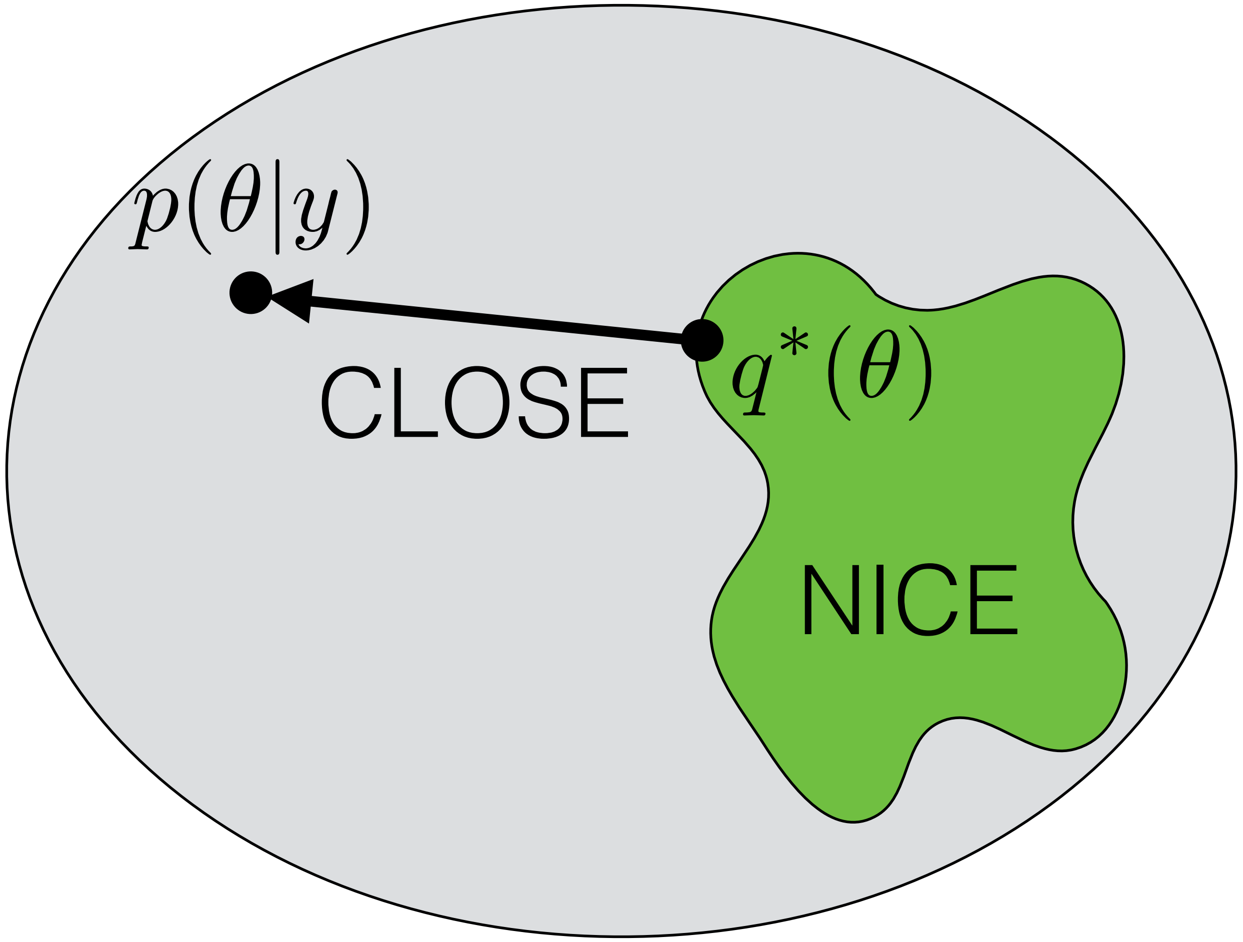
$$= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta$$



Why KL?

- Variational Bayes

$$q^* = \mathrm{argmin}_{q \in Q} \mathrm{KL}\left(q(\cdot) || p(\cdot | y)\right)$$

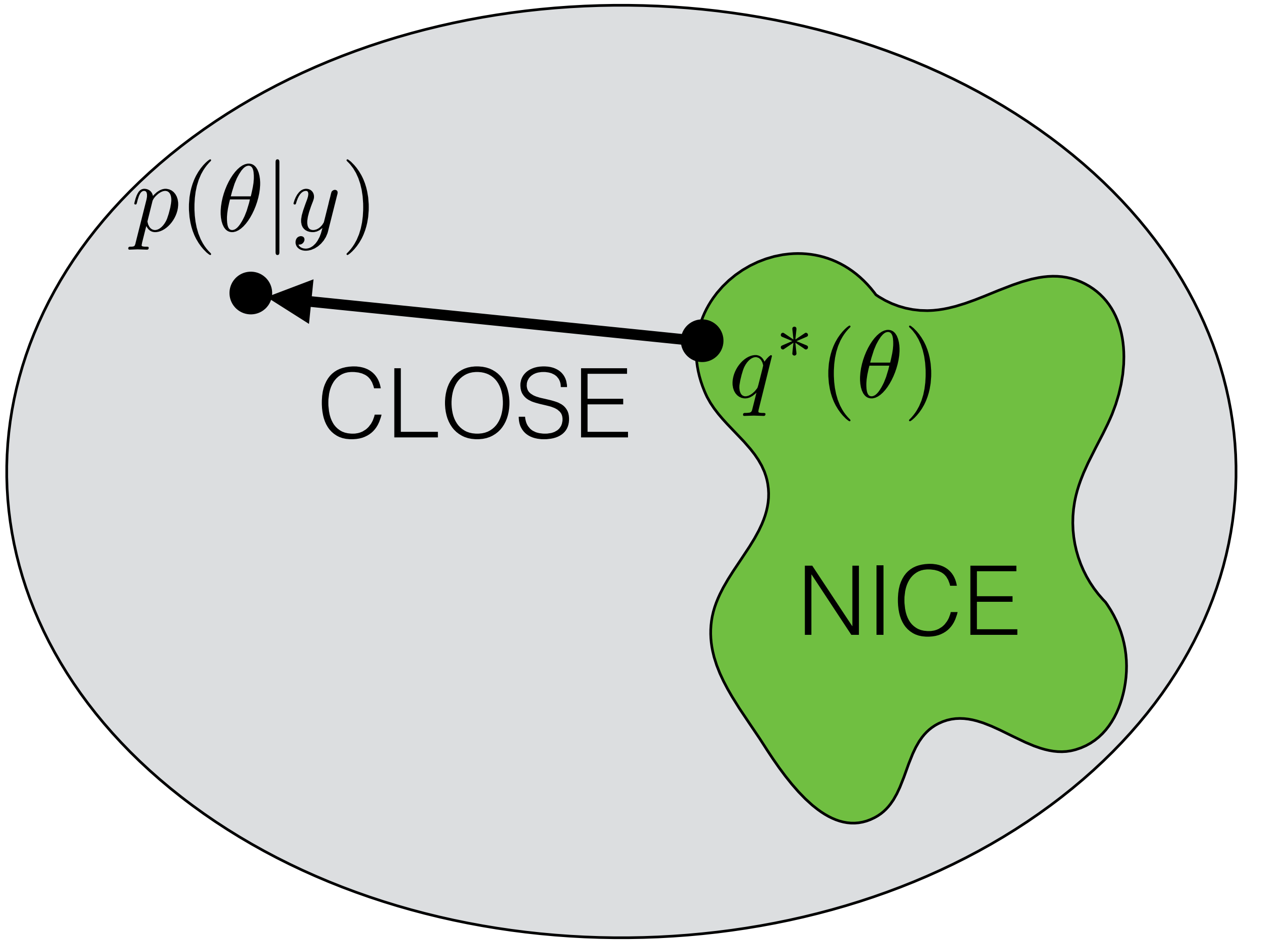


$$\mathrm{KL}\left(q(\cdot) || p(\cdot | y)\right)$$

$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

$$= \int q(\theta) \log \frac{q(\theta) p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta$$

Why KL?



- Variational Bayes

$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$

$$\text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$

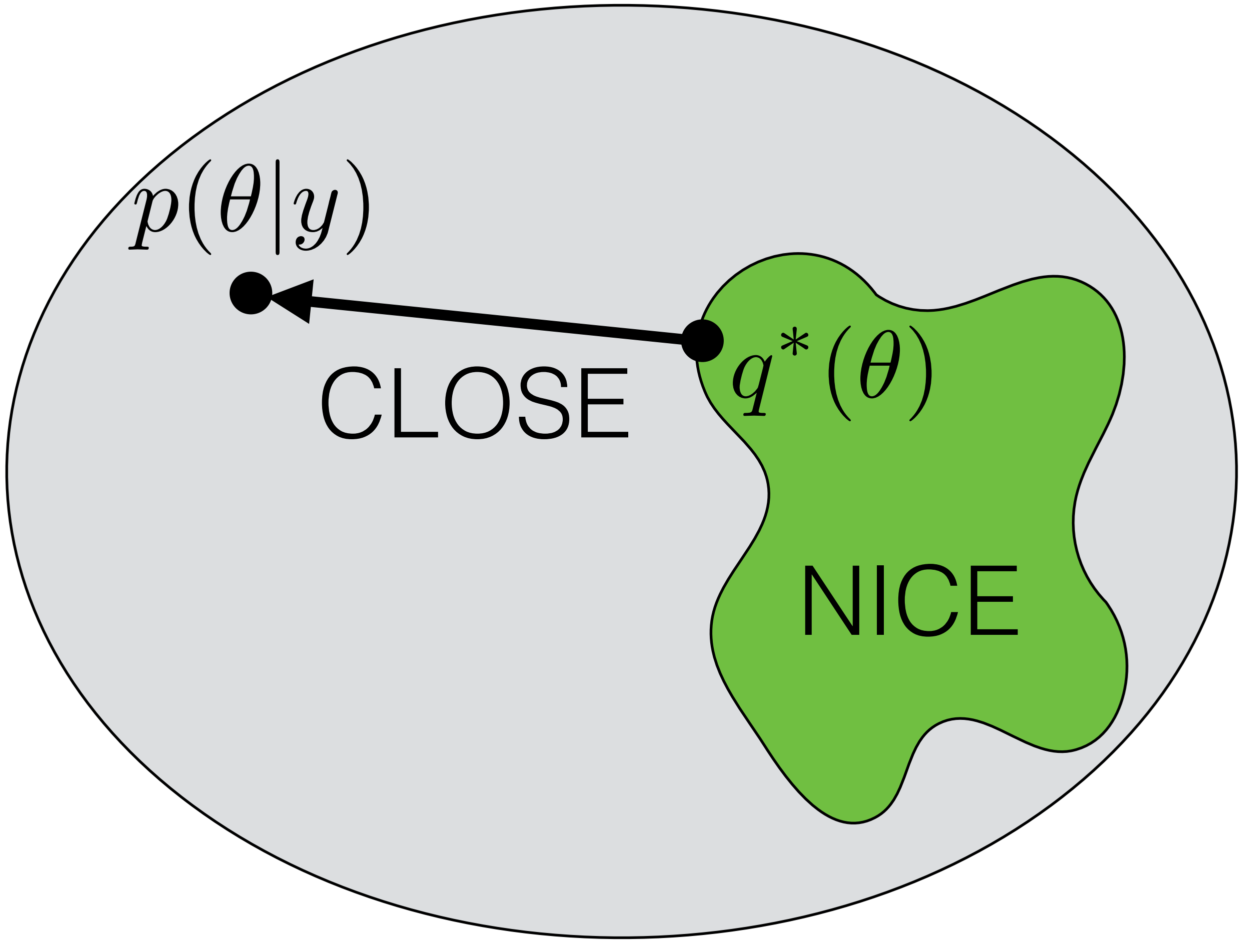
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

$$= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta,y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta,y)}{q(\theta)} d\theta$$

Why KL?

- Variational Bayes

$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$



$$\text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$

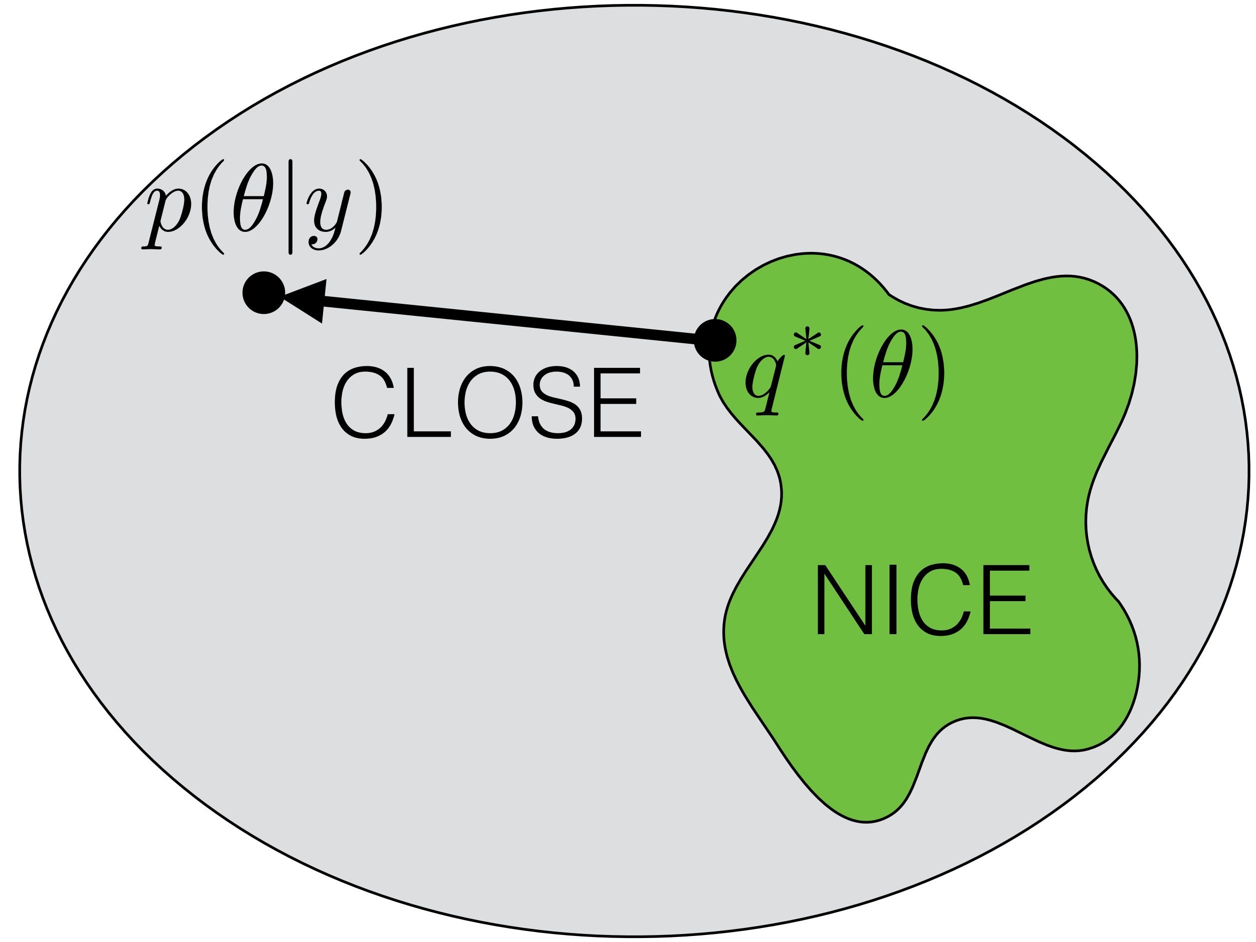
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

$$= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta$$

Why KL?

- Variational Bayes

$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot | y)\right)$$



$$\text{KL}\left(q(\cdot) || p(\cdot | y)\right)$$

$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

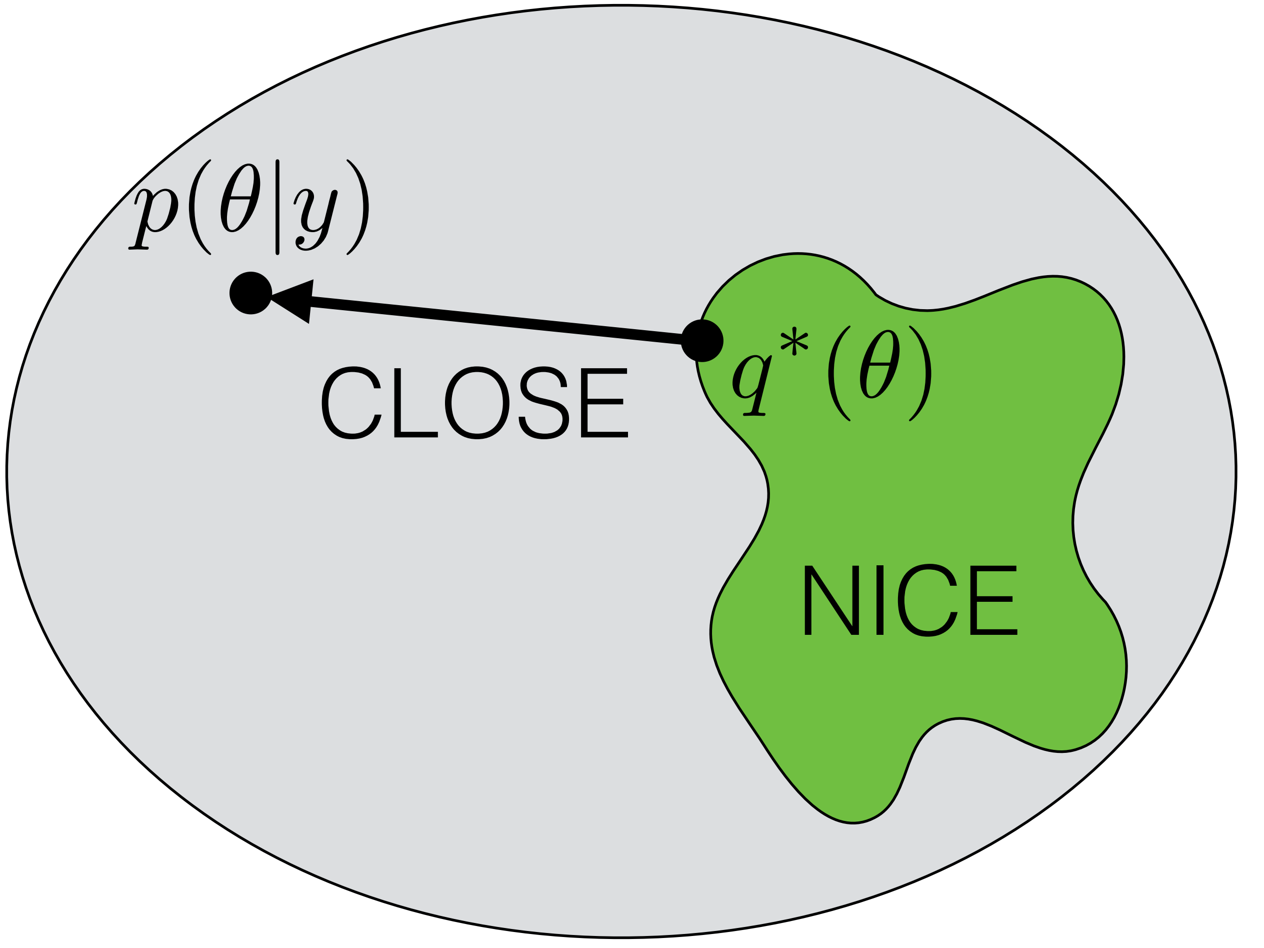
$$= \int q(\theta) \log \frac{q(\theta) p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta$$

"Evidence lower bound" (ELBO)

Why KL?

- Variational Bayes

$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$



$$\text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$

$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

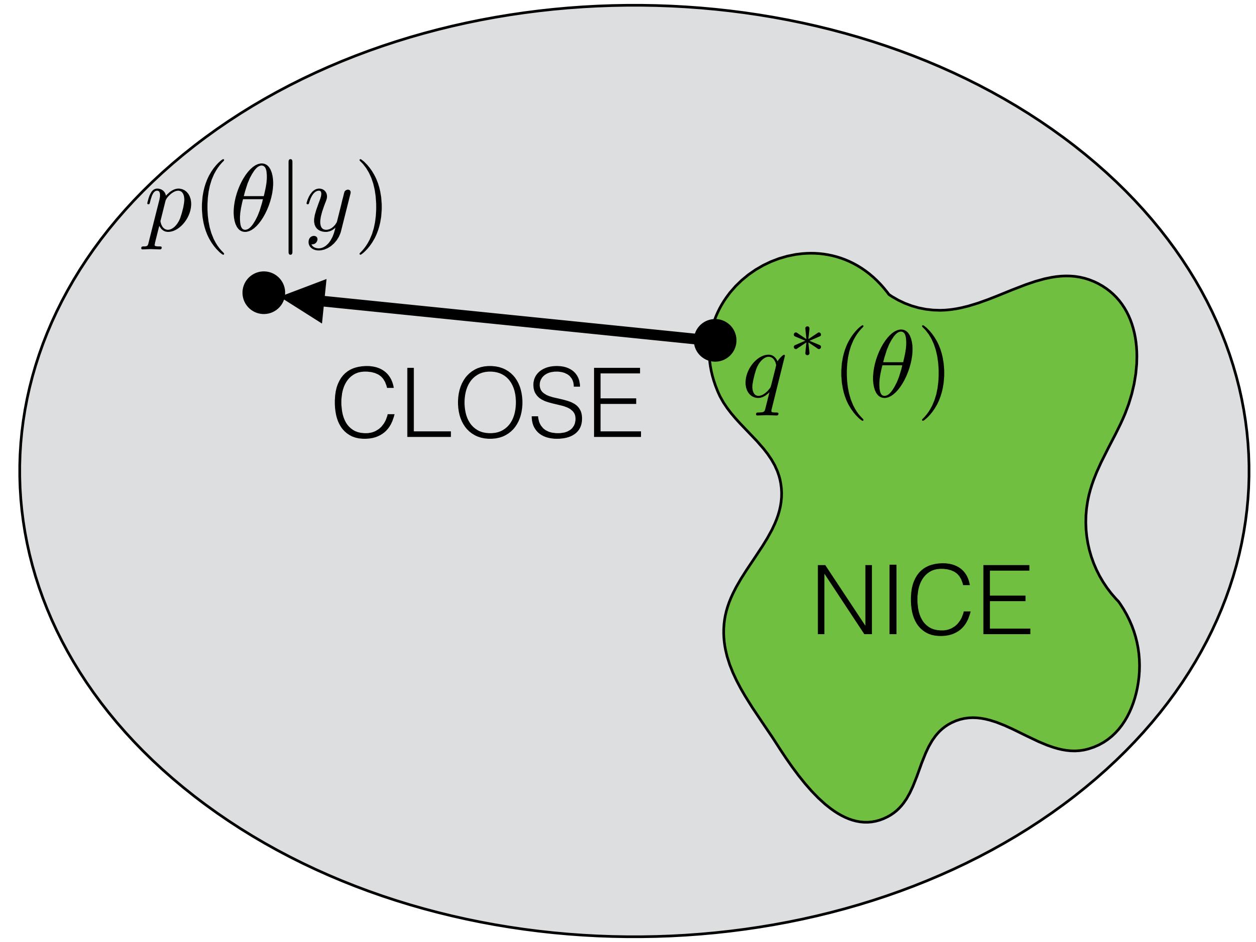
$$= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta$$

"Evidence lower bound" (ELBO)

Why KL?

- Variational Bayes

$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot | y)\right)$$



$$\text{KL}\left(q(\cdot) || p(\cdot | y)\right)$$

$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

$$= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta$$

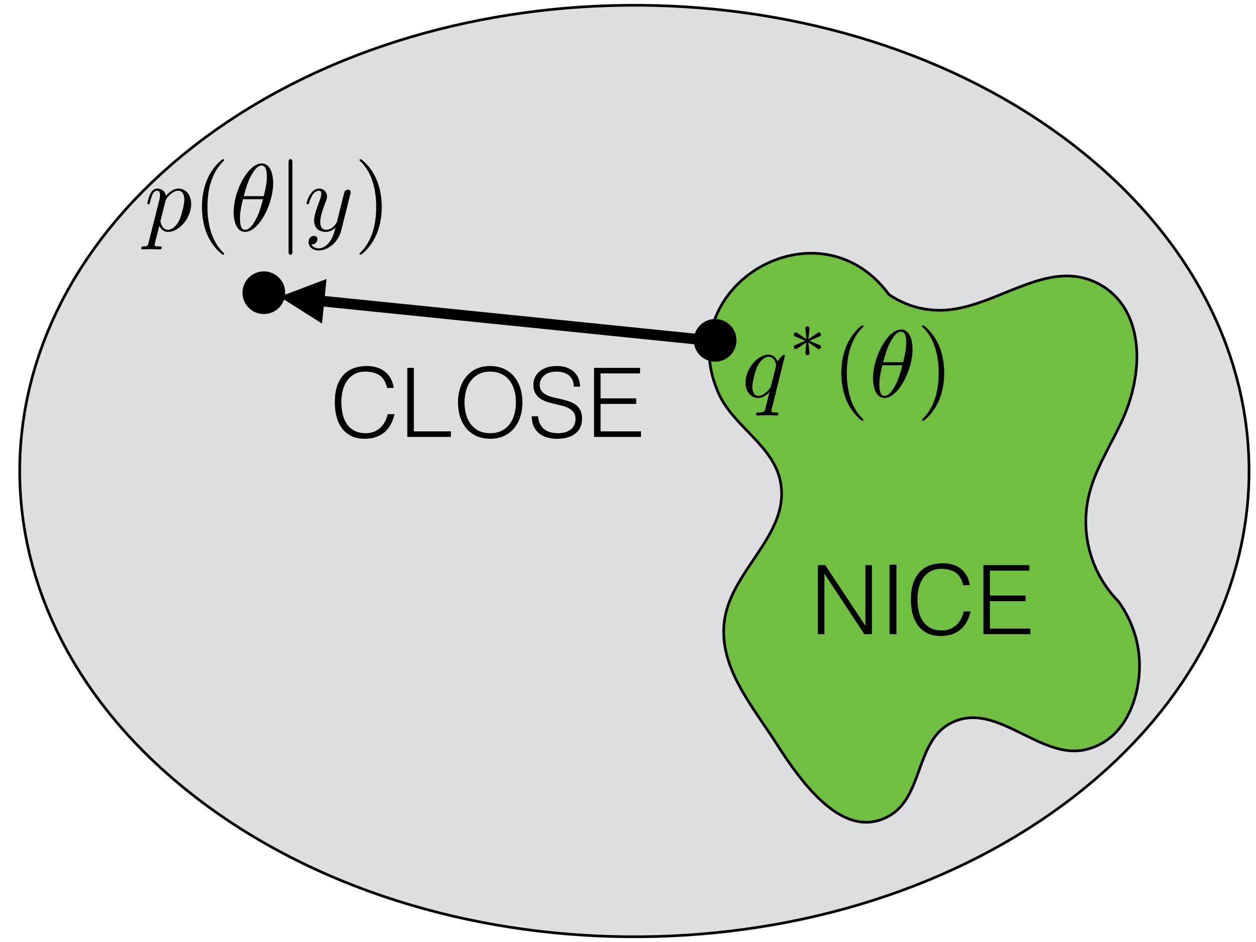
"Evidence lower bound" (ELBO)

- Exercise: Show $\text{KL} \geq 0$ [Bishop 2006, Sec 1.6.1]

Why KL?

- Variational Bayes

$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot | y)\right)$$



$$\text{KL}\left(q(\cdot) || p(\cdot | y)\right)$$

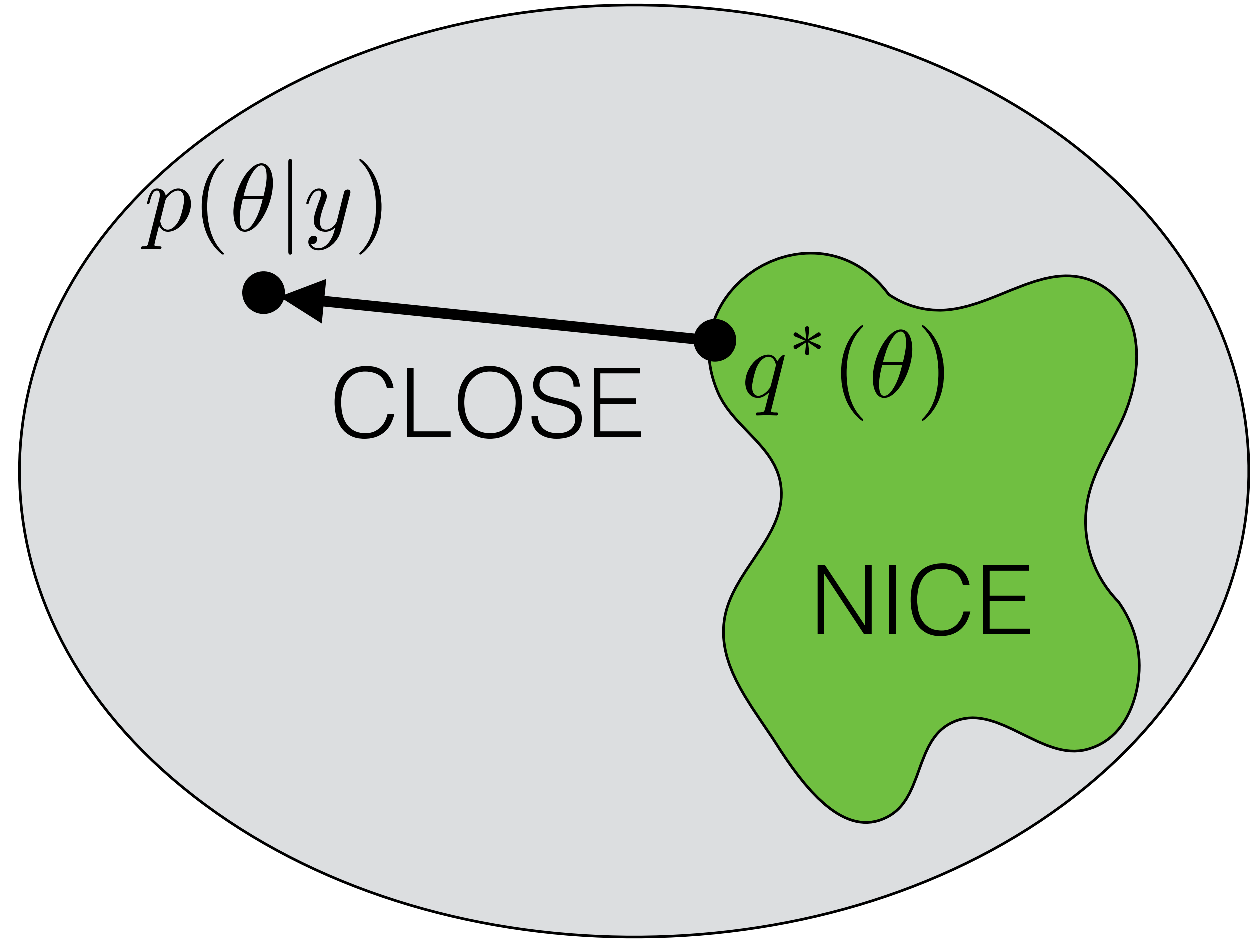
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

$$= \int q(\theta) \log \frac{q(\theta) p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta$$

"Evidence lower bound" (ELBO)

- Exercise: Show $\text{KL} \geq 0$ [Bishop 2006, Sec 1.6.1]
- $\text{KL} \geq 0 \Rightarrow \log p(y) \geq \text{ELBO}$

Why KL?



- Variational Bayes

$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot | y)\right)$$

$$\text{KL}\left(q(\cdot) || p(\cdot | y)\right)$$

$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta | y)} d\theta$$

$$= \int q(\theta) \log \frac{q(\theta) p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta$$

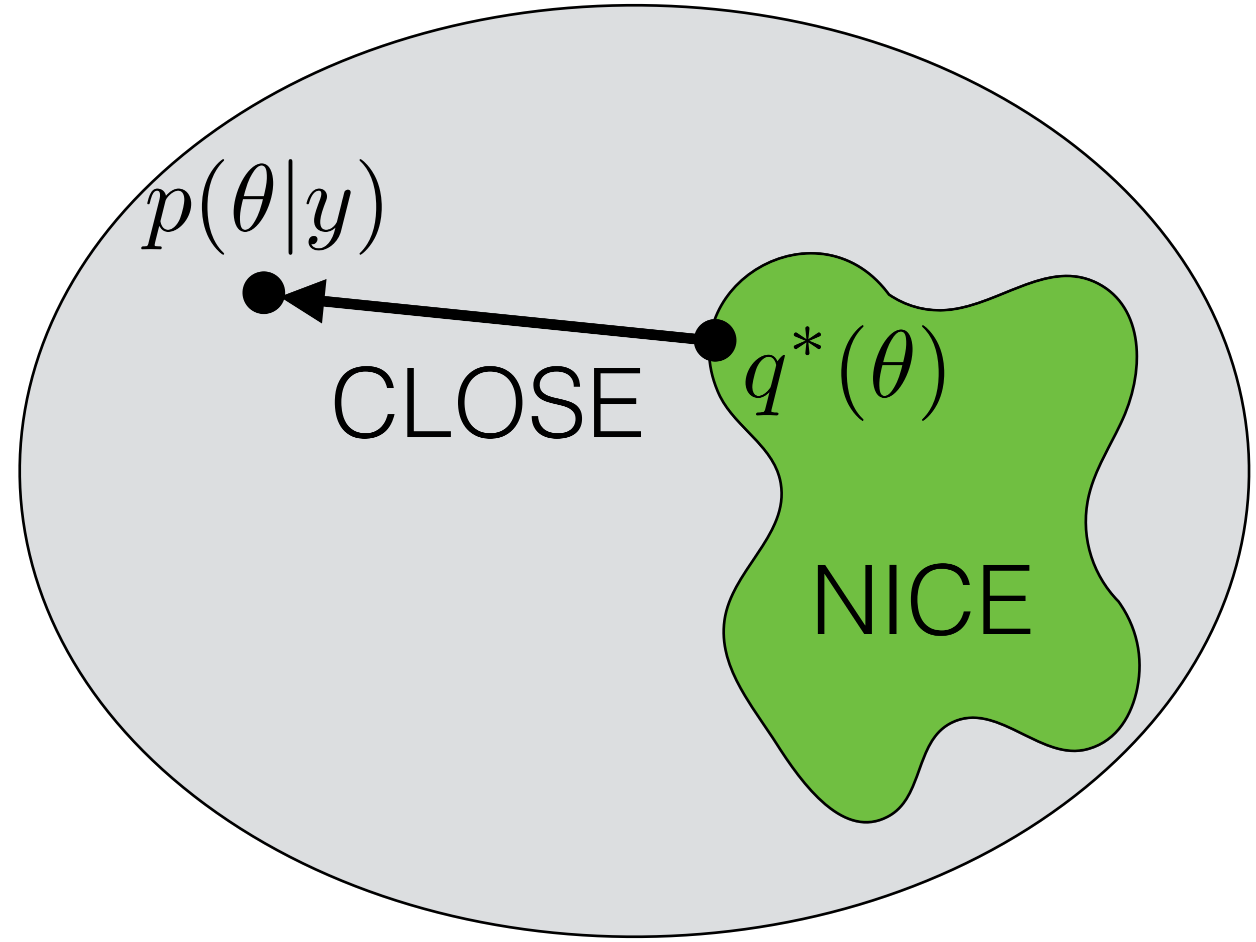
"Evidence lower bound" (ELBO)

- Exercise: Show $\text{KL} \geq 0$ [Bishop 2006, Sec 1.6.1]
- $\text{KL} \geq 0 \Rightarrow \log p(y) \geq \text{ELBO}$
- $q^* = \text{argmax}_{q \in Q} \text{ELBO}(q)$

Why KL?

- Variational Bayes

$$q^* = \operatorname{argmin}_{q \in Q} \mathrm{KL}\left(q(\cdot) || p(\cdot | y)\right)$$



$$\mathrm{KL}\left(q(\cdot) || p(\cdot | y)\right)$$

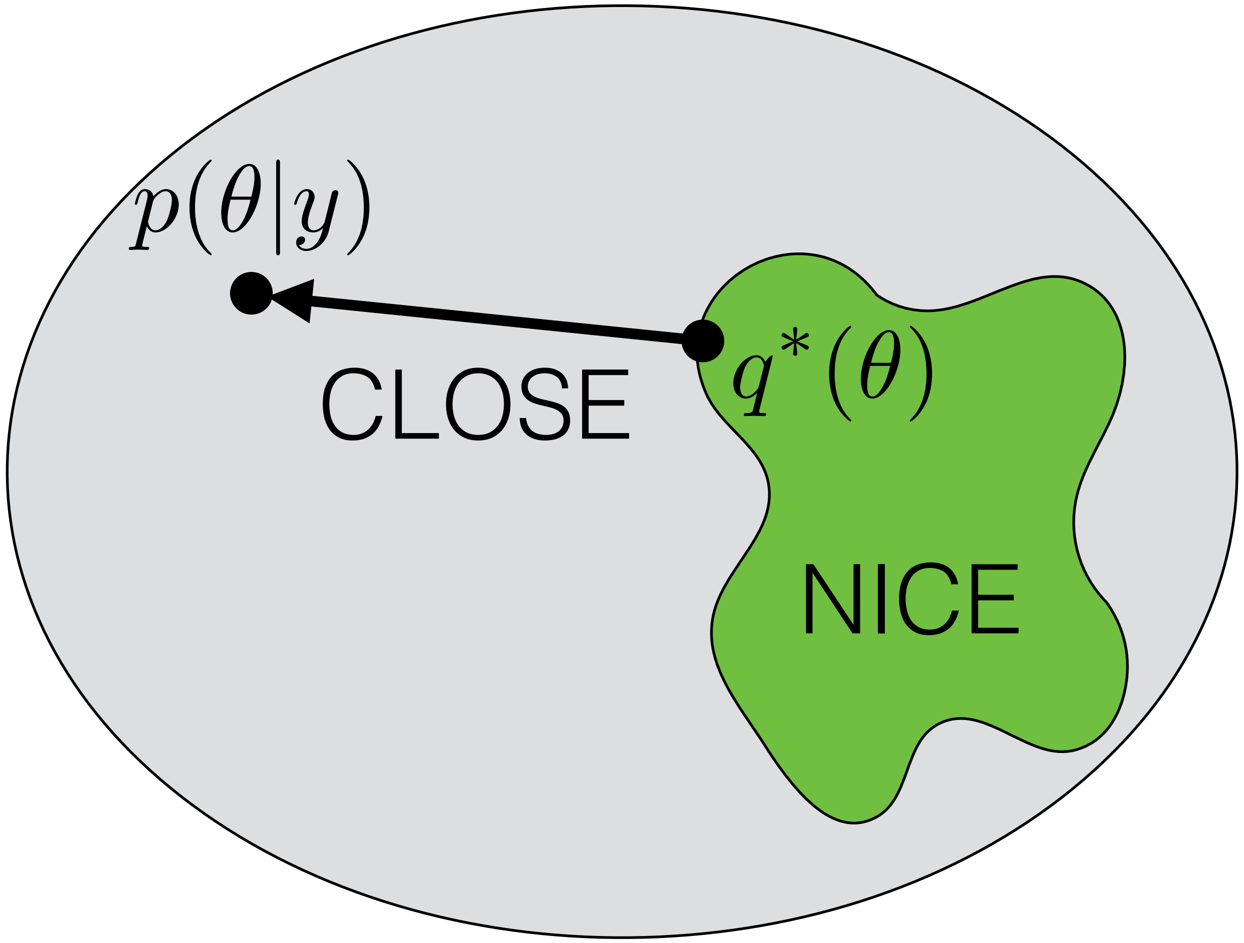
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

$$= \int q(\theta) \log \frac{q(\theta) p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta$$

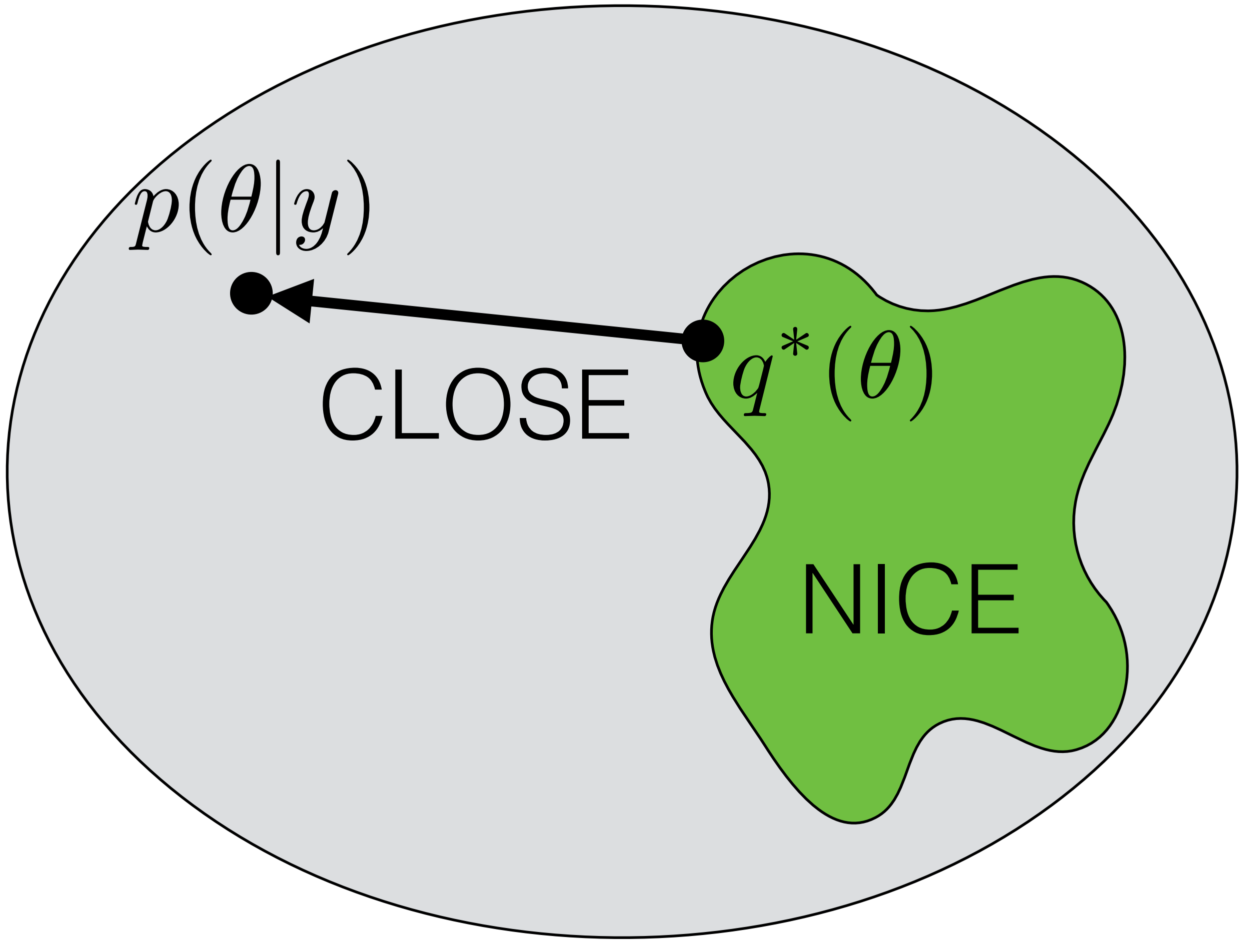
"Evidence lower bound" (ELBO)

- Exercise: Show $\mathrm{KL} \geq 0$ [Bishop 2006, Sec 1.6.1]
- $\mathrm{KL} \geq 0 \Rightarrow \log p(y) \geq \mathrm{ELBO}$
- $q^* = \operatorname{argmax}_{q \in Q} \mathrm{ELBO}(q)$
- Why KL (in this direction)?

Variational Bayes $q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot|y)\right)$



Variational Bayes $q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot|y)\right)$

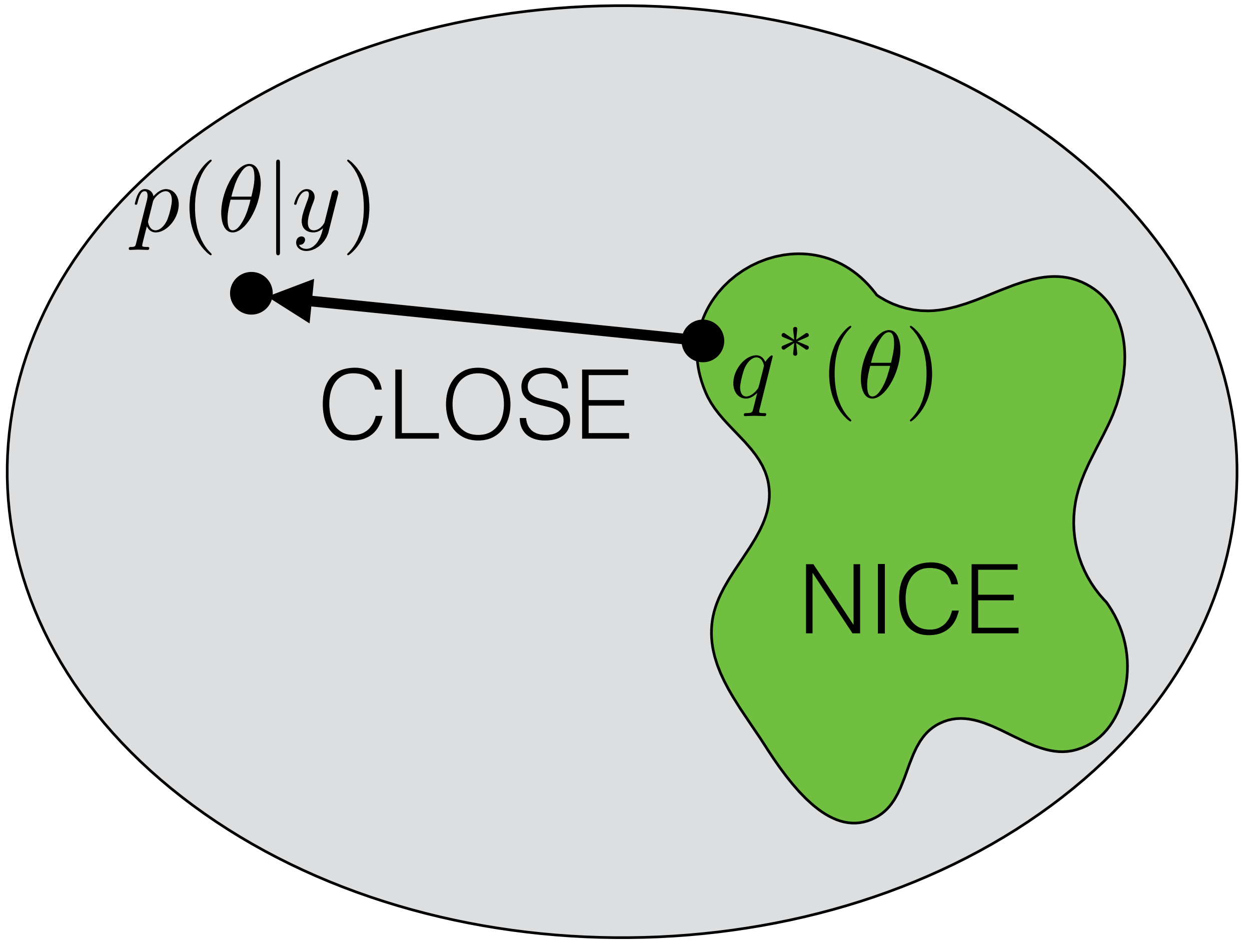


Choose "NICE" distributions

Variational Bayes

$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot)||p(\cdot|y)\right)$$

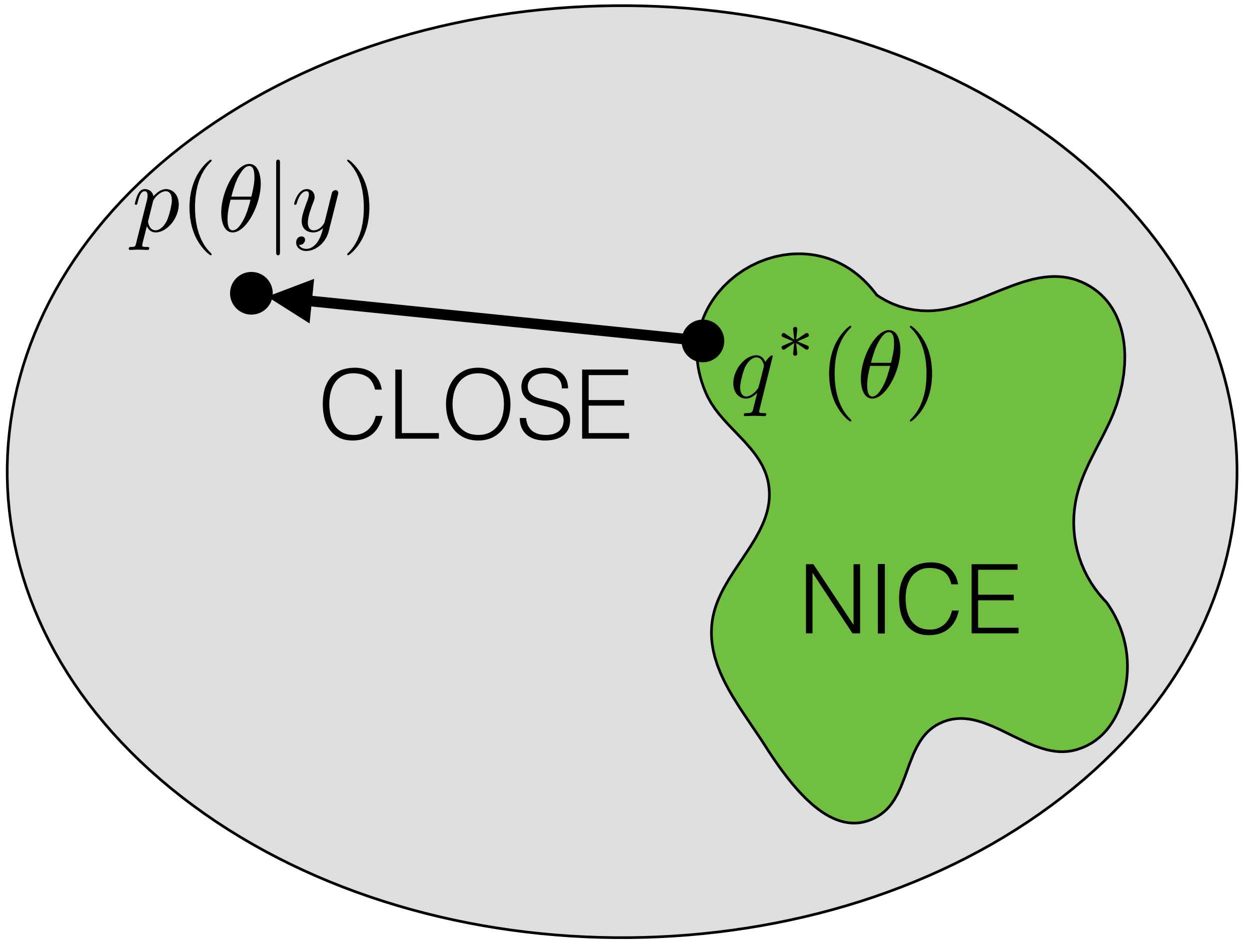
Choose "NICE" distributions



Variational Bayes

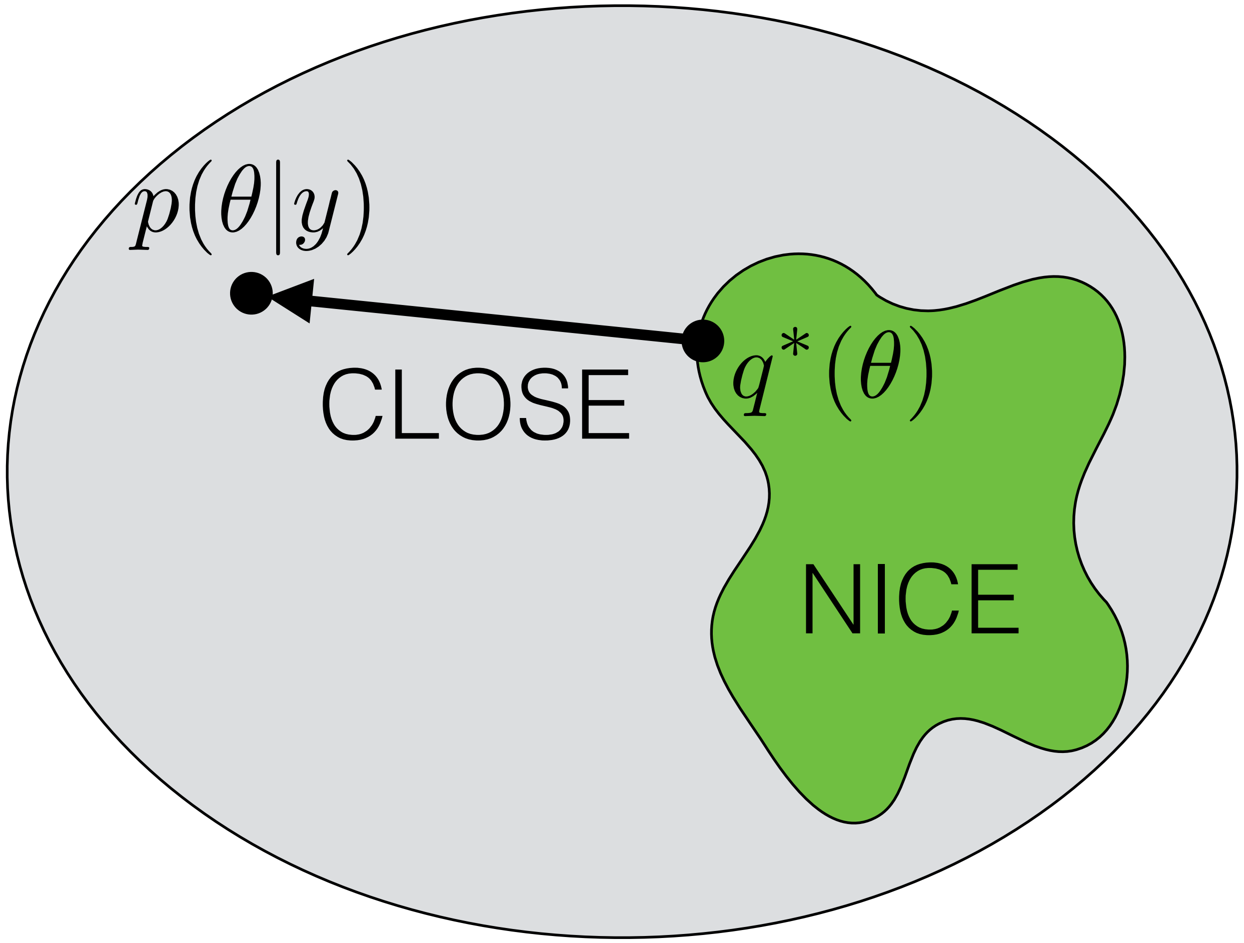
$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$

Choose "NICE" distributions



Variational Bayes

$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$



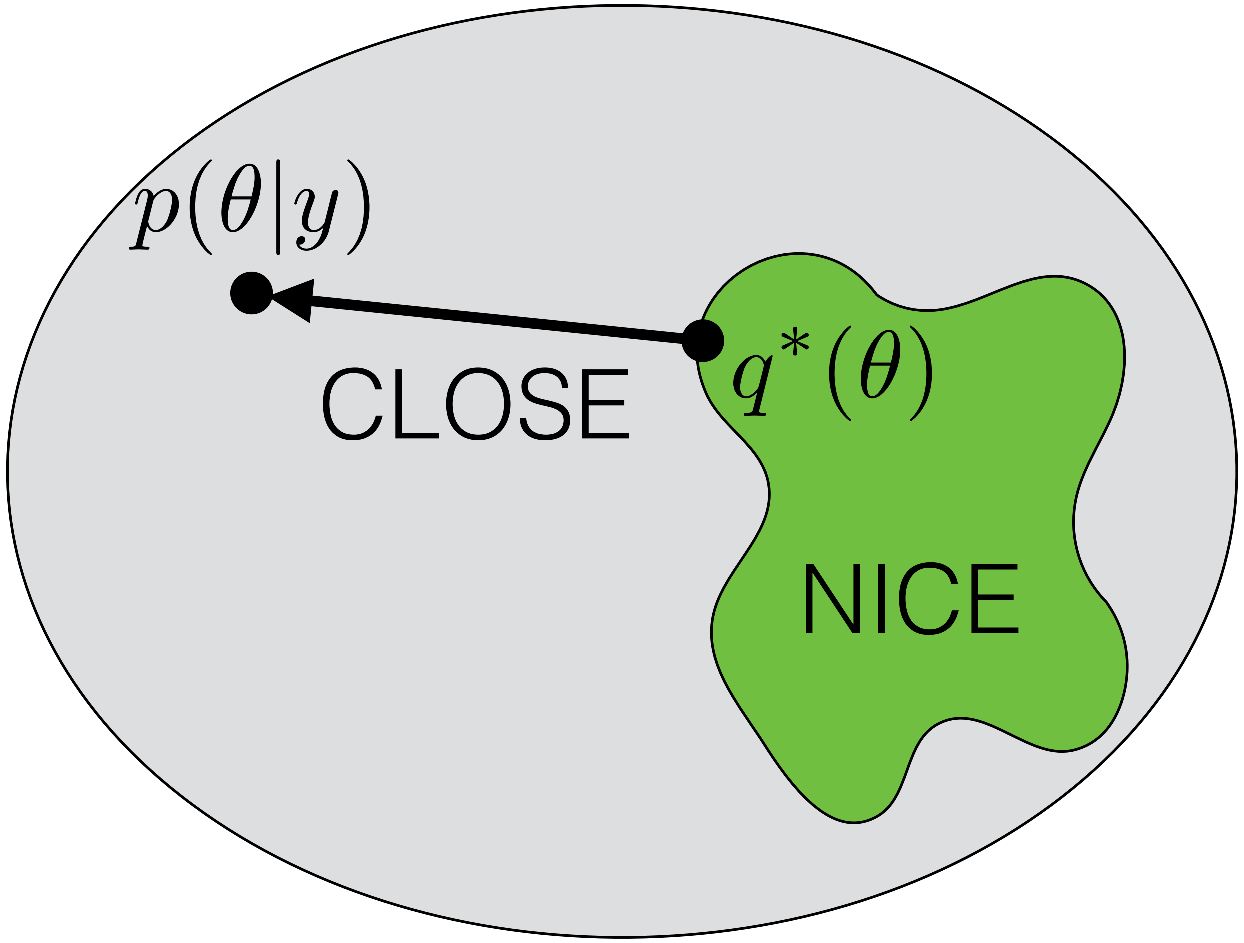
Choose "NICE" distributions

- Mean-field variational Bayes (MFVB)

$$Q_{MFVB} := \left\{ q : q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \right\}$$

Variational Bayes

$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$



Choose "NICE" distributions

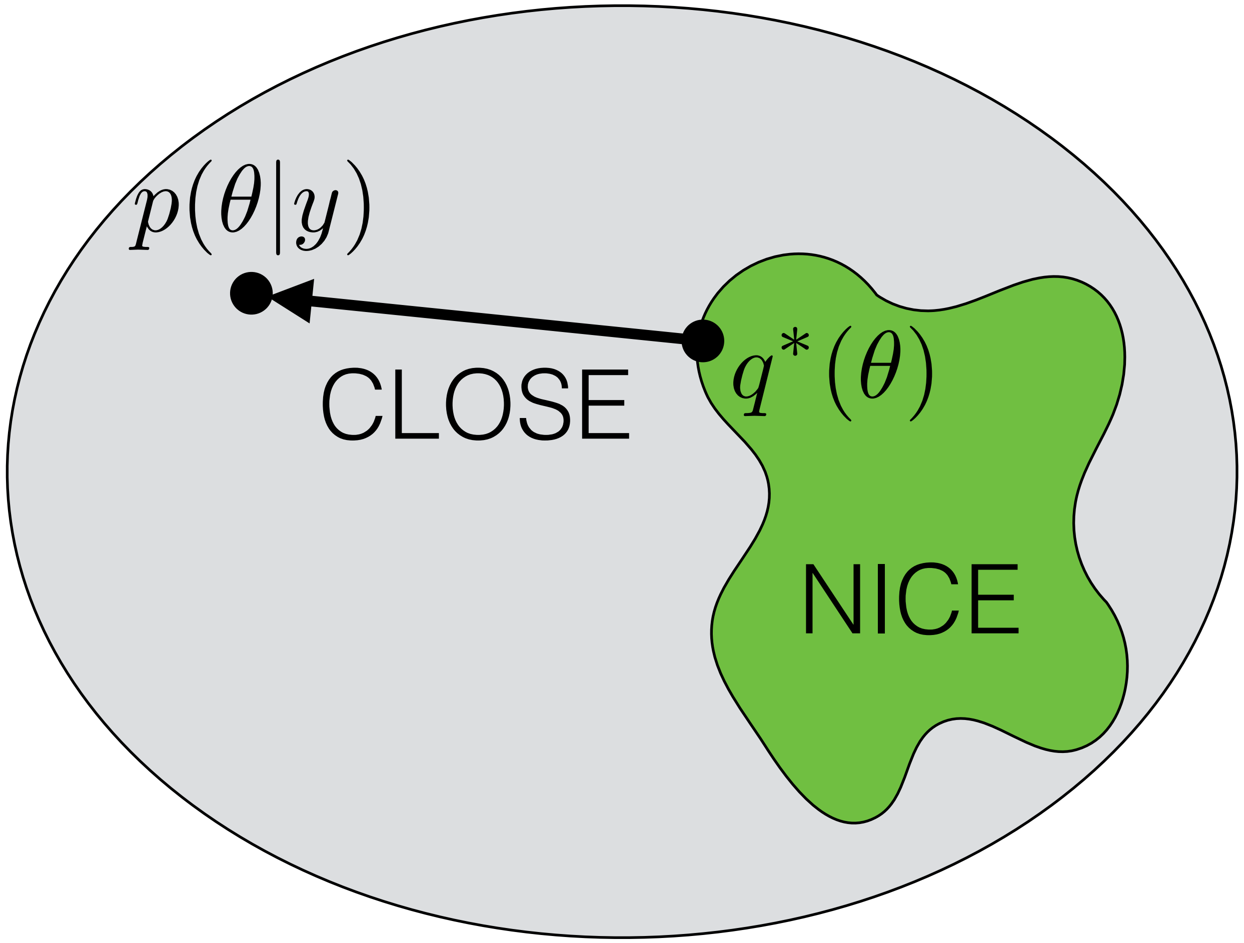
- Mean-field variational Bayes (MFVB)

$$Q_{MFVB} := \left\{ q : q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \right\}$$

- Often also exponential family

Variational Bayes

$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$



Choose "NICE" distributions

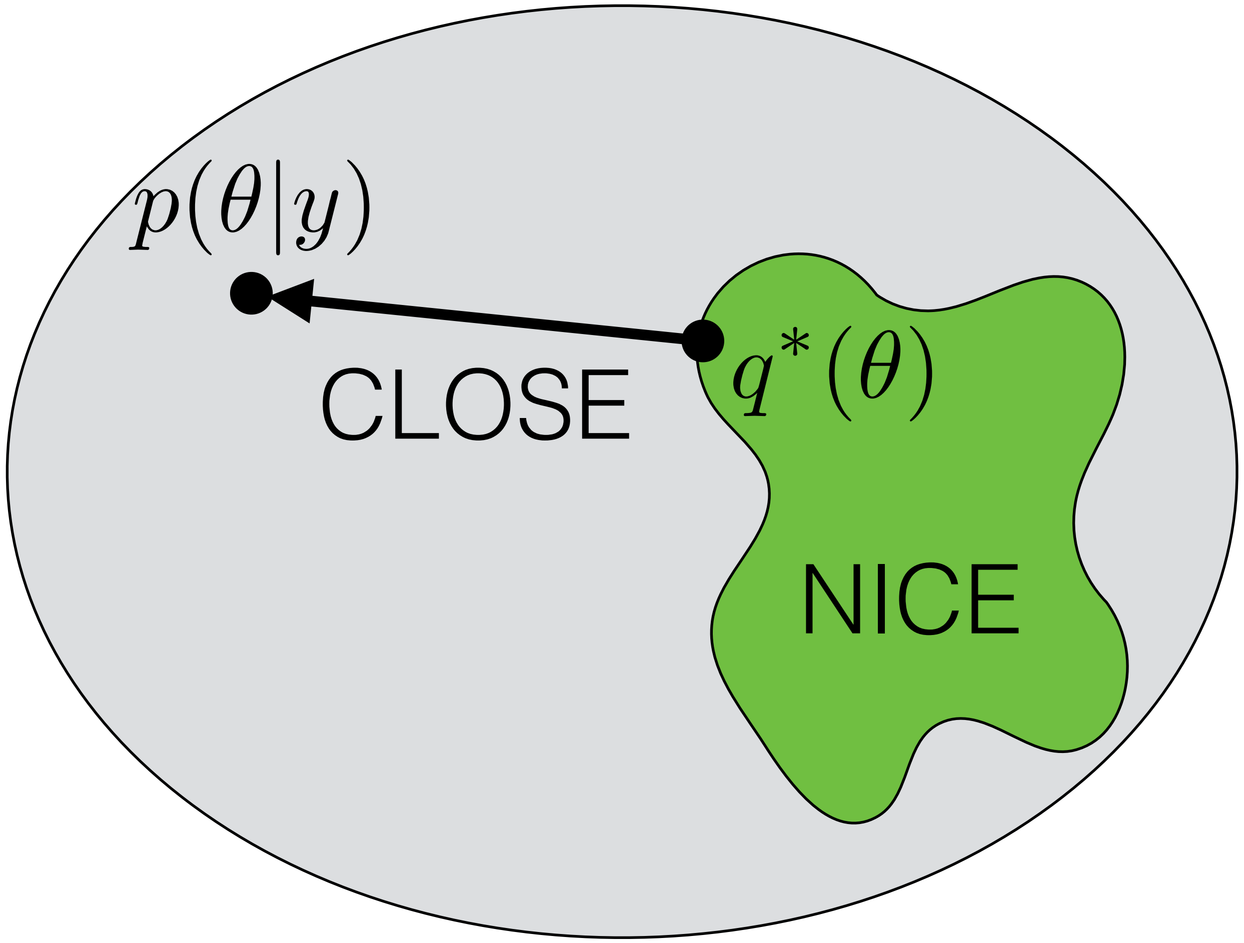
- Mean-field variational Bayes (MFVB)

$$Q_{MFVB} := \left\{ q : q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \right\}$$

- Often also exponential family
- *Not* a modeling assumption

Variational Bayes

$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$

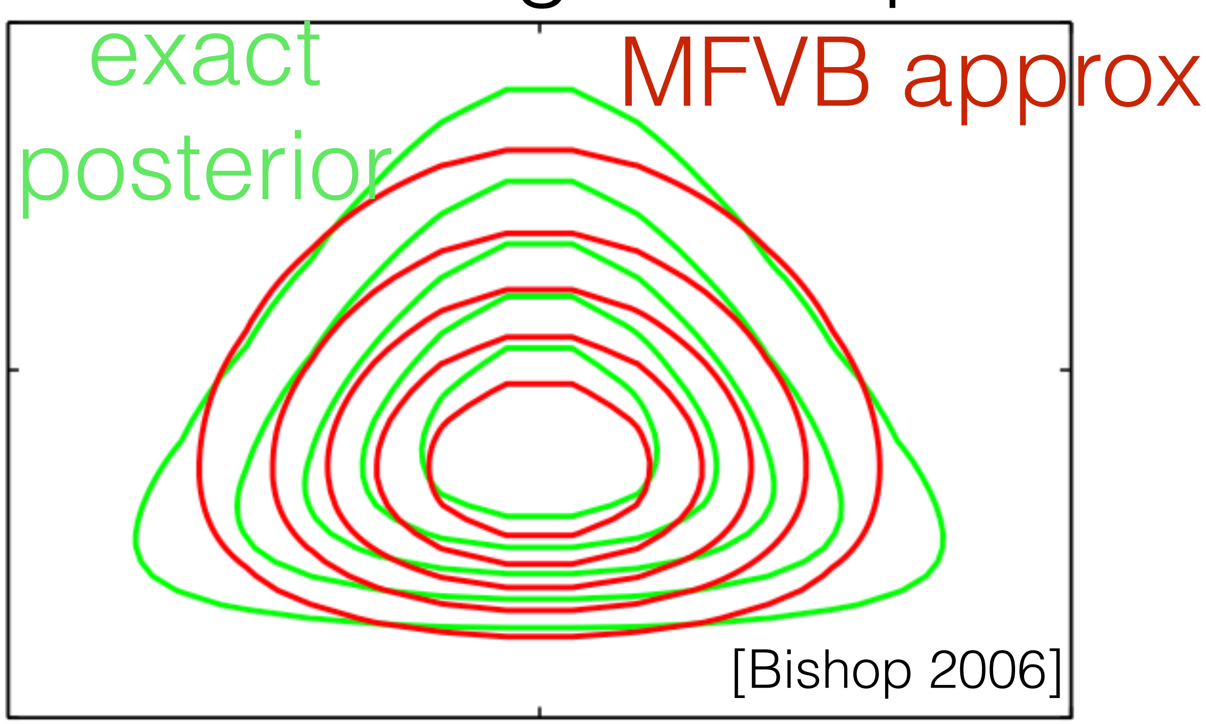


Choose "NICE" distributions

- Mean-field variational Bayes (MFVB)

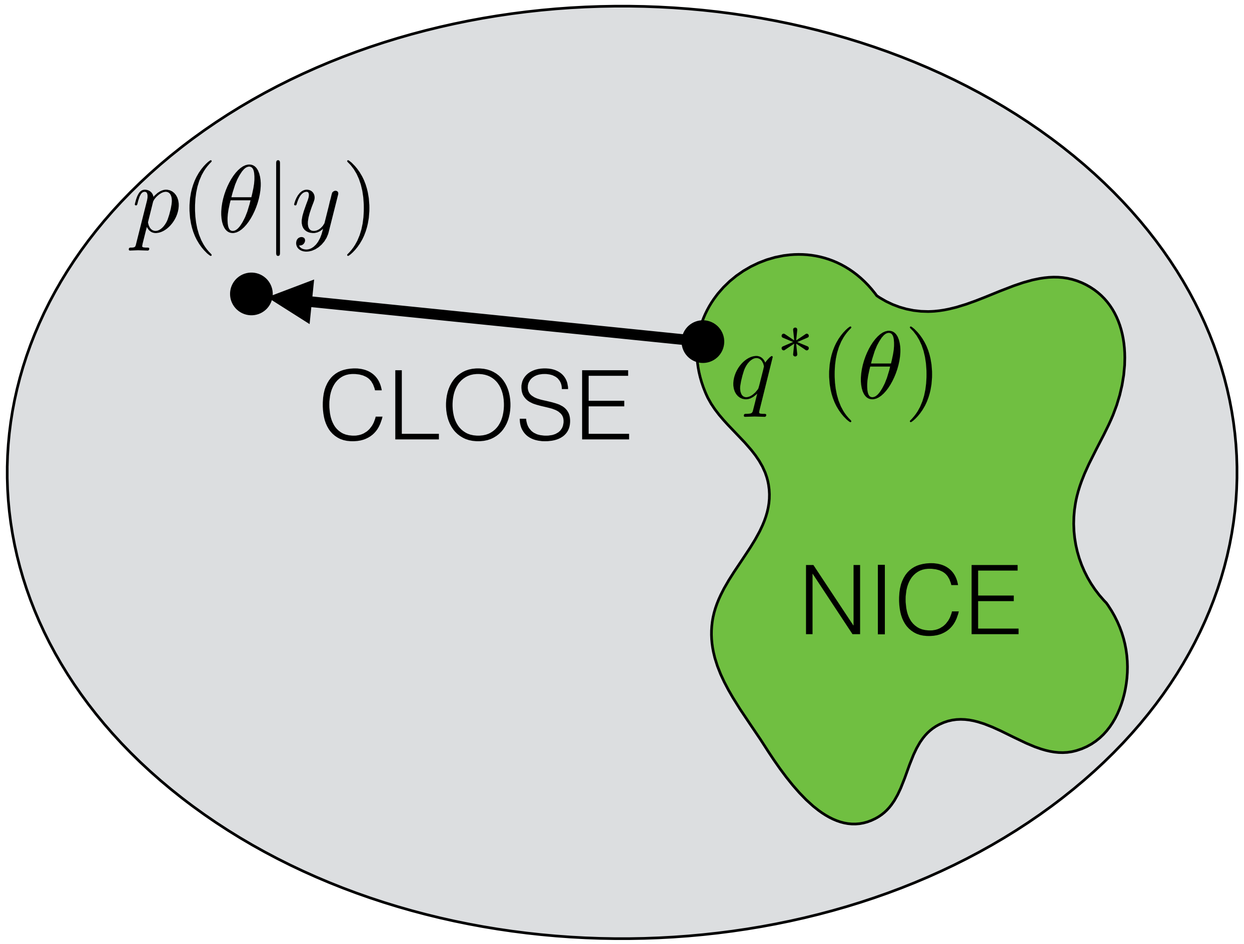
$$Q_{MFVB} := \left\{ q : q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \right\}$$

- Often also exponential family
- *Not* a modeling assumption



Variational Bayes

$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$



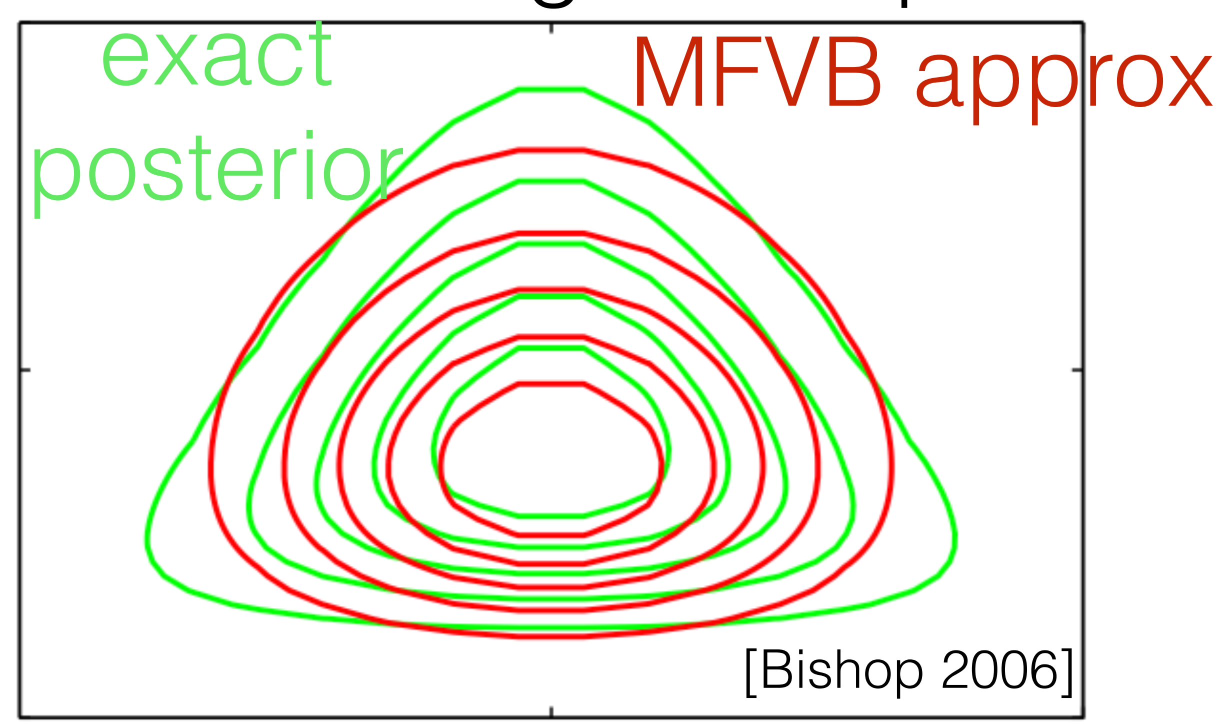
Choose "NICE" distributions

- Mean-field variational Bayes (MFVB)

$$Q_{MFVB} := \left\{ q : q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \right\}$$

- Often also exponential family
- *Not* a modeling assumption

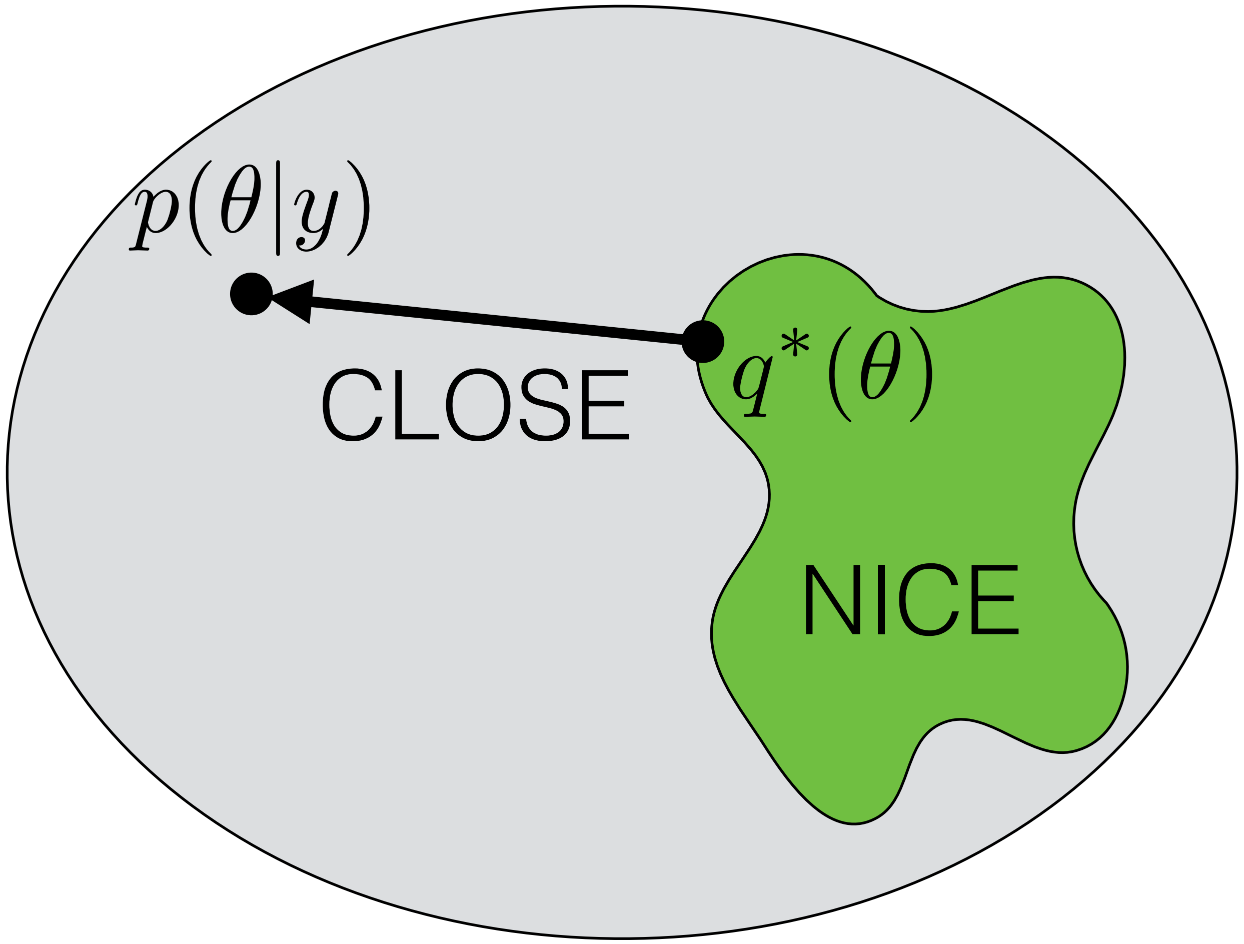
Now we have an optimization problem; how to solve it?



[Bishop 2006]

Variational Bayes

$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot | y)\right)$$

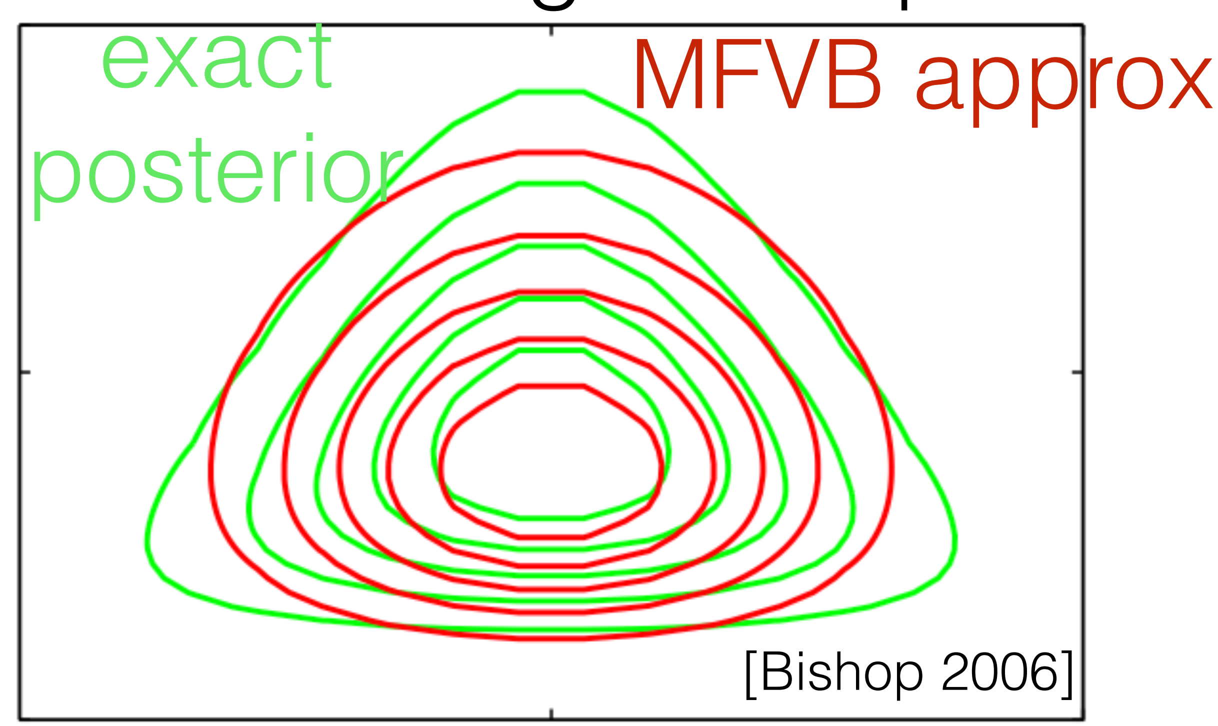


Choose "NICE" distributions

- Mean-field variational Bayes (MFVB)

$$Q_{MFVB} := \left\{ q : q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \right\}$$

- Often also exponential family
- *Not* a modeling assumption



Now we have an optimization problem; how to solve it?

- One option: Coordinate descent in $q_1, \ldots, q_J$

Approximate Bayesian inference

Approximate Bayesian inference

Use q^* to approximate $p(\cdot|y)$

Approximate Bayesian inference

Use q^* to approximate $p(\cdot|y)$

Optimization

$$q^* = \text{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

Approximate Bayesian inference

Use q^* to approximate $p(\cdot|y)$

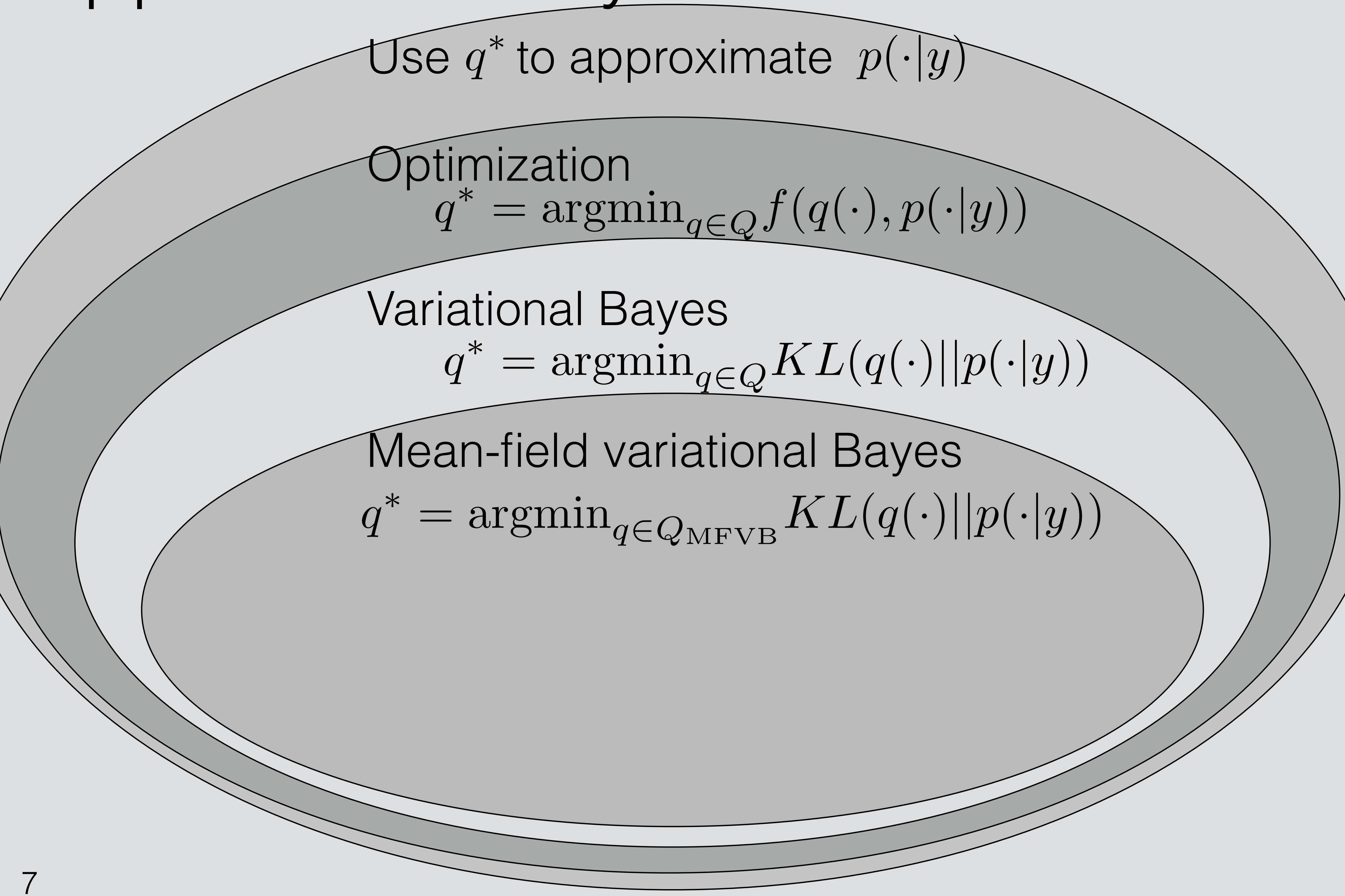
Optimization

$$q^* = \text{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

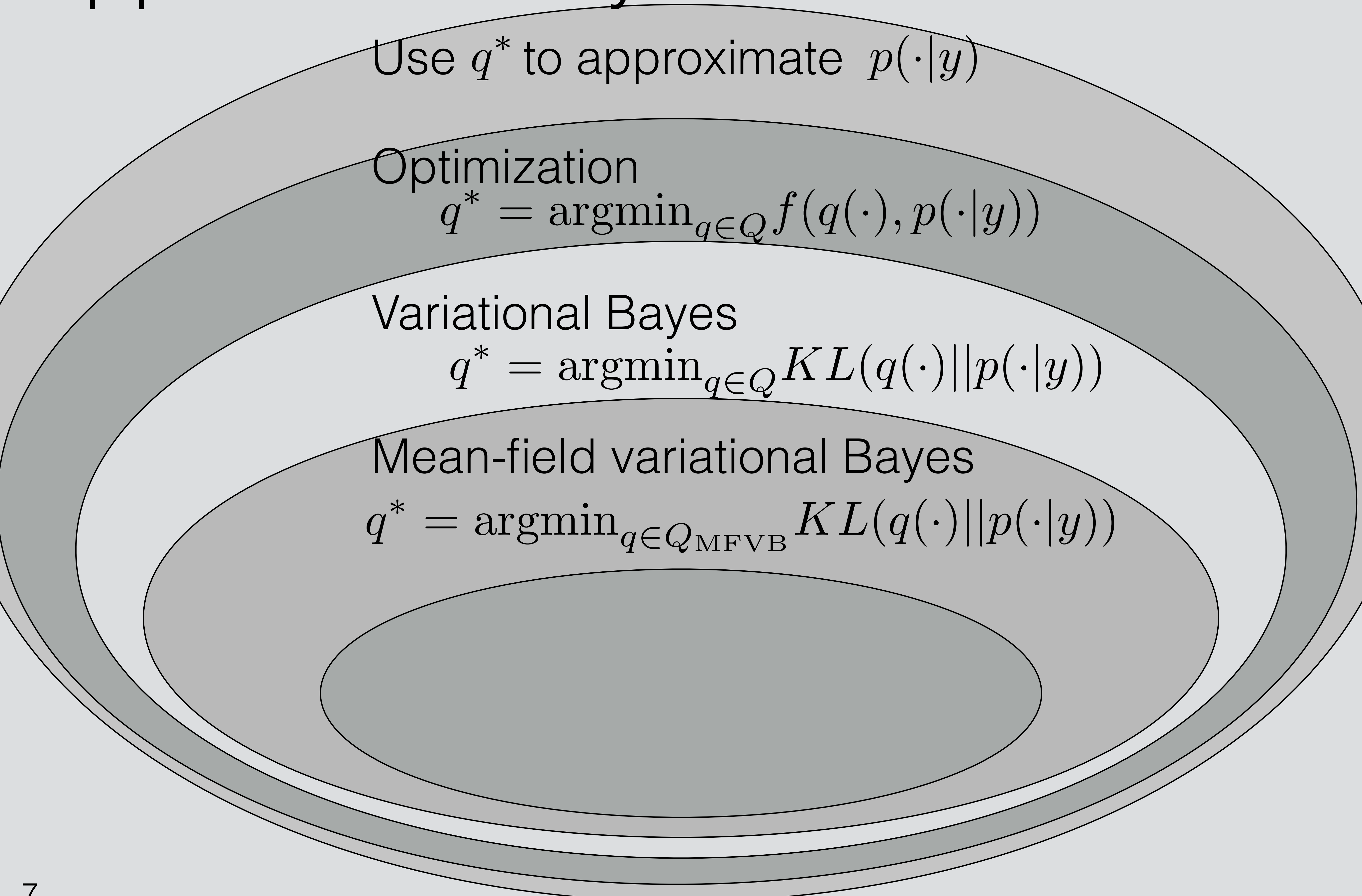
Variational Bayes

$$q^* = \text{argmin}_{q \in Q} KL(q(\cdot) || p(\cdot|y))$$

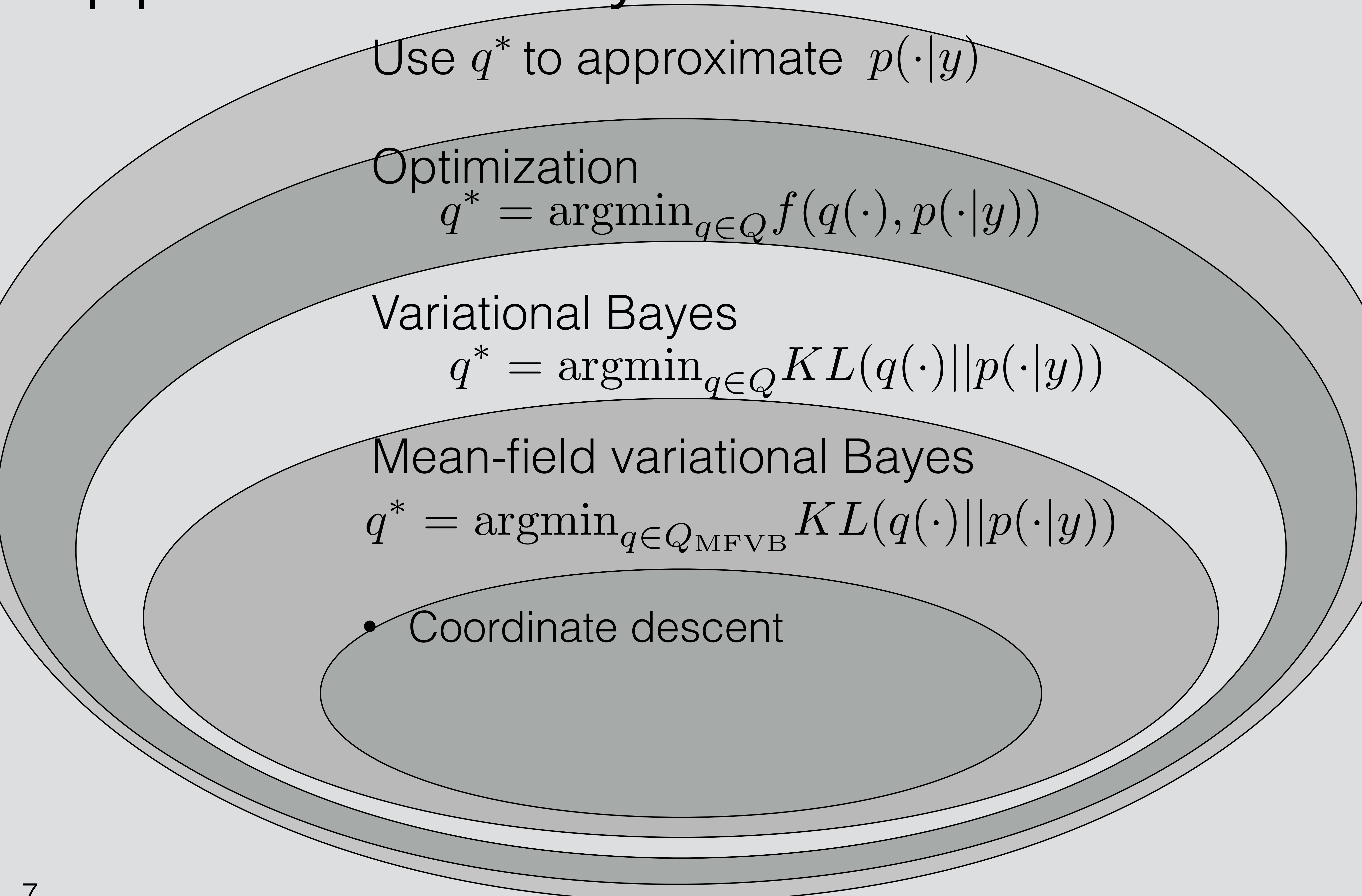
Approximate Bayesian inference



Approximate Bayesian inference



Approximate Bayesian inference



Approximate Bayesian inference

Use q^* to approximate $p(\cdot|y)$

Optimization

$$q^* = \text{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

Variational Bayes

$$q^* = \text{argmin}_{q \in Q} KL(q(\cdot)||p(\cdot|y))$$

Mean-field variational Bayes

$$q^* = \text{argmin}_{q \in Q_{\text{MFVB}}} KL(q(\cdot)||p(\cdot|y))$$

- Coordinate descent
- Stochastic variational inference (SVI) [Hoffman et al 2013]

Approximate Bayesian inference

Use q^* to approximate $p(\cdot|y)$

Optimization

$$q^* = \mathrm{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

Variational Bayes

$$q^* = \mathrm{argmin}_{q \in Q} KL(q(\cdot)||p(\cdot|y))$$

Mean-field variational Bayes

$$q^* = \mathrm{argmin}_{q \in Q_{\mathrm{MFVB}}} KL(q(\cdot)||p(\cdot|y))$$

- Coordinate descent
- Stochastic variational inference (SVI) [Hoffman et al 2013]
- Automatic differentiation variational inference (ADVI) [Kucukelbir et al 2015, 2017]

Roadmap

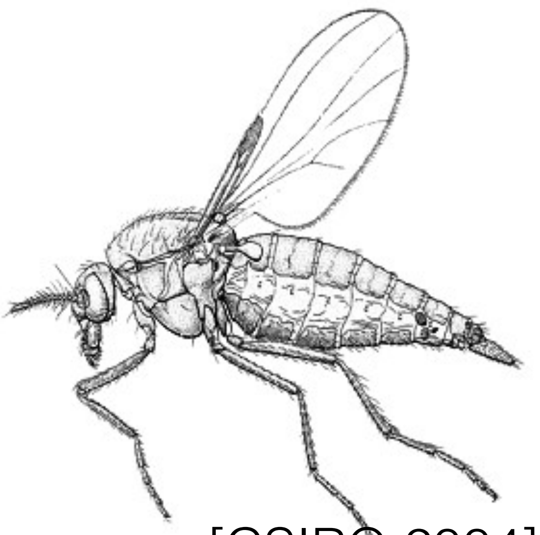
- Bayes & Approximate Bayes review
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use MFVB?
- When can we trust MFVB?
- Where do we go from here?

Roadmap

- Bayes & Approximate Bayes review
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use MFVB?
- When can we trust MFVB?
- Where do we go from here?

Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$



[CSIRO 2004]

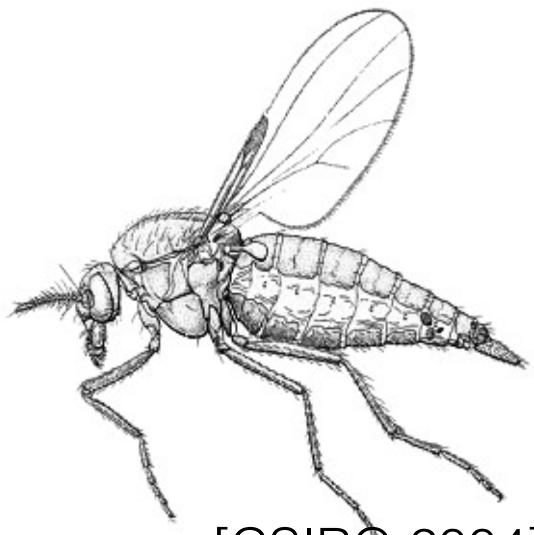
[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$

- Model:

$$p(y|\theta): \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \quad n = 1, \ldots, N$$



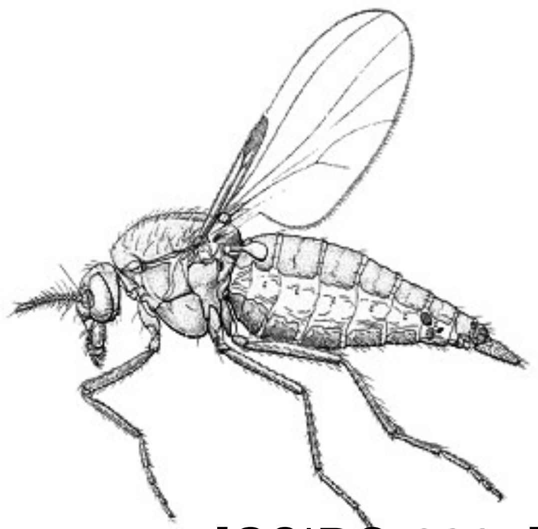
[CSIRO 2004]

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and variance $\theta = (\mu, \sigma^2)$
- Model:

$$p(y|\theta): \quad y_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \quad n = 1, \ldots, N$$



[CSIRO 2004]

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

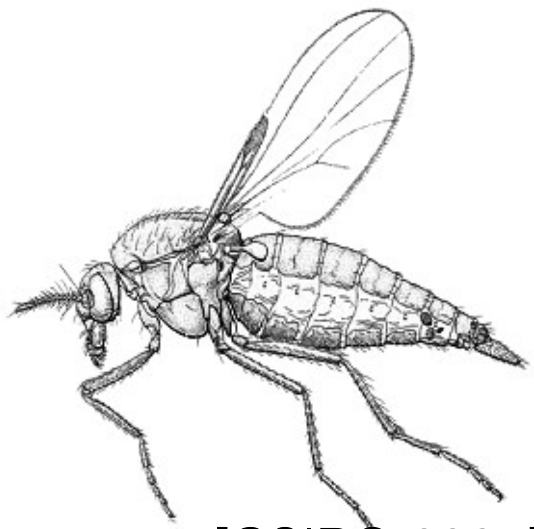
Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and variance $\theta = (\mu, \sigma^2)$
- Model:

$$p(y|\theta): \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \quad n = 1, \ldots, N$$

$$p(\theta): \quad (\sigma^2)^{-1} \sim \text{Gamma}(a_0, b_0)$$

$$\mu|\sigma^2 \sim \mathcal{N}(\mu_0, \lambda_0 \sigma^2)$$



[CSIRO 2004]

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

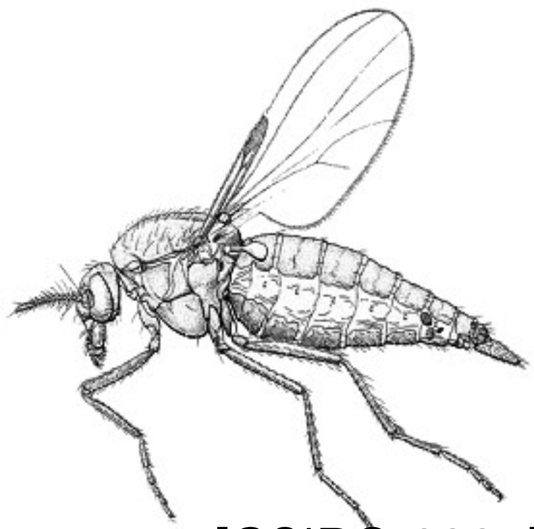
Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and variance $\theta = (\mu, \sigma^2)$
- Model (conjugate prior):

$$p(y|\theta): \quad y_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \quad n = 1, \ldots, N$$

$$p(\theta): \quad (\sigma^2)^{-1} \sim \text{Gamma}(a_0, b_0)$$

$$\mu|\sigma^2 \sim \mathcal{N}(\mu_0, \lambda_0 \sigma^2)$$



[CSIRO 2004]

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

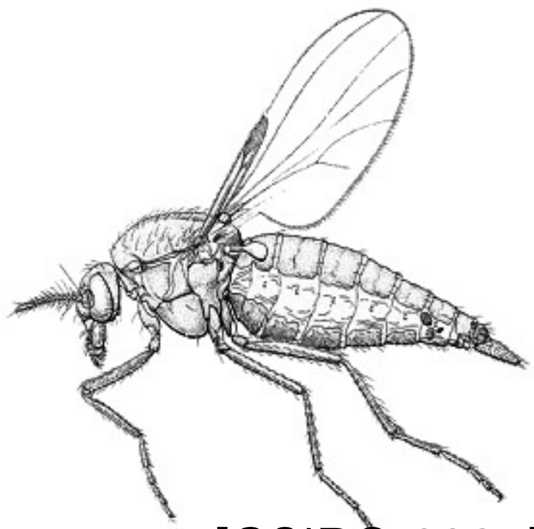
Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and variance $\theta = (\mu, \sigma^2)$
- Model (conjugate prior): [Exercise: find the posterior]

$$p(y|\theta): \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \quad n = 1, \ldots, N$$

$$p(\theta): \quad (\sigma^2)^{-1} \sim \text{Gamma}(a_0, b_0)$$

$$\mu | \sigma^2 \sim \mathcal{N}(\mu_0, \lambda_0 \sigma^2)$$



[CSIRO 2004]

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

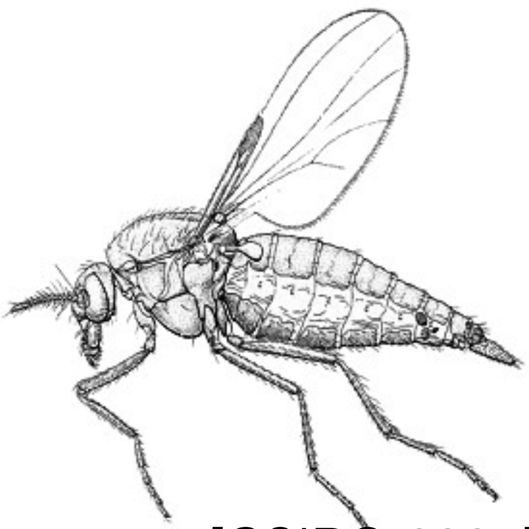
Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and variance $\theta = (\mu, \sigma^2)$
- Model (conjugate prior): [Exercise: find the posterior]

$$p(y|\theta): \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \quad n = 1, \ldots, N$$

$$p(\theta): \quad (\sigma^2)^{-1} \sim \text{Gamma}(a_0, b_0)$$

$$\mu|\sigma^2 \sim \mathcal{N}(\mu_0, \lambda_0 \sigma^2)$$



[CSIRO 2004]

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

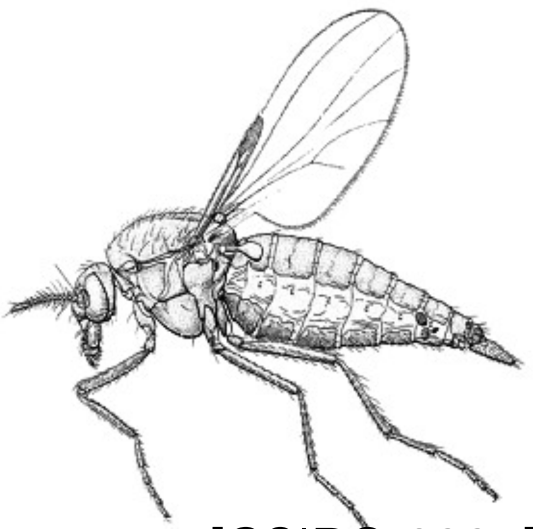
Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and precision $\theta = (\mu, \tau)$
- Model (conjugate prior): [Exercise: find the posterior]

$$p(y|\theta): \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \quad n = 1, \ldots, N$$

$$p(\theta): \quad (\sigma^2)^{-1} \sim \text{Gamma}(a_0, b_0)$$

$$\mu|\sigma^2 \sim \mathcal{N}(\mu_0, \lambda_0 \sigma^2)$$



[CSIRO 2004]

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

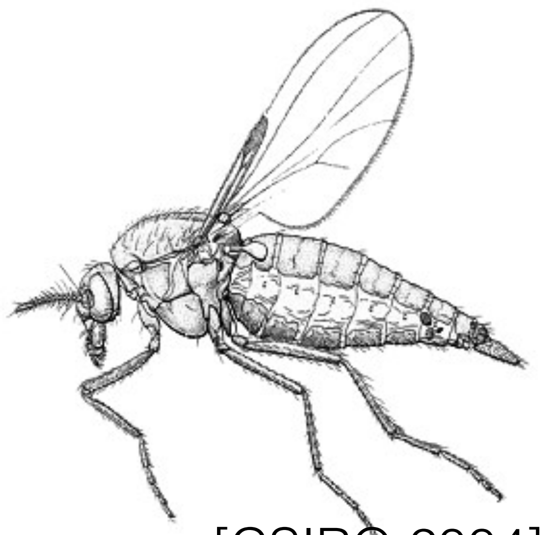
Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and precision $\theta = (\mu, \tau)$
- Model (conjugate prior): [Exercise: find the posterior]

$$p(y|\theta): \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \quad n = 1, \ldots, N$$

$$p(\theta): \quad (\sigma^2)^{-1} \sim \text{Gamma}(a_0, b_0)$$

$$\mu|\sigma^2 \sim \mathcal{N}(\mu_0, \lambda_0\sigma^2)$$



[CSIRO 2004]

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

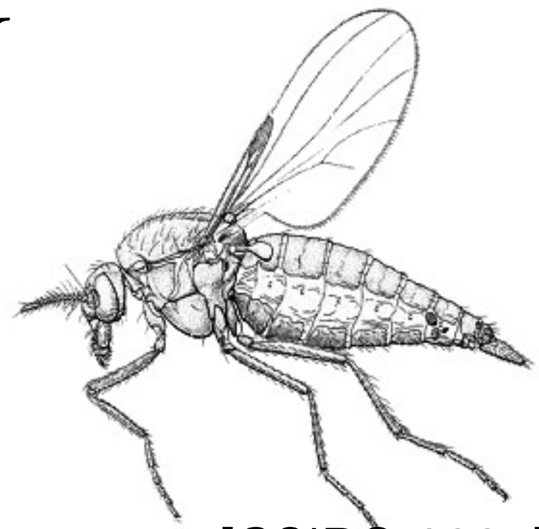
Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and precision $\theta = (\mu, \tau)$
- Model (conjugate prior): [Exercise: find the posterior]

$$p(y|\theta): \quad y_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \tau^{-1}), \quad n = 1, \ldots, N$$

$$p(\theta): \quad \tau \sim \text{Gamma}(a_0, b_0)$$

$$\mu|\tau \sim \mathcal{N}(\mu_0, (\rho_0 \tau)^{-1})$$



[CSIRO 2004]

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

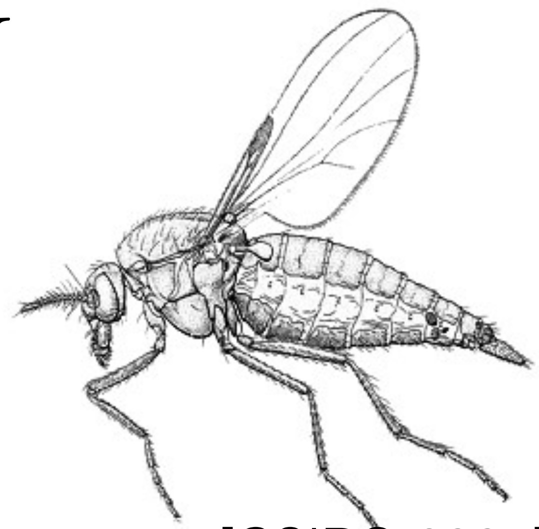
Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and precision $\theta = (\mu, \tau)$
- Model (conjugate prior): [Exercise: find the posterior]

$$p(y|\theta): \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \tau^{-1}), \quad n = 1, \ldots, N$$

$$p(\theta): \quad \tau \sim \text{Gamma}(a_0, b_0)$$

$$\mu|\tau \sim \mathcal{N}(\mu_0, (\rho_0\tau)^{-1})$$



[CSIRO 2004]

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

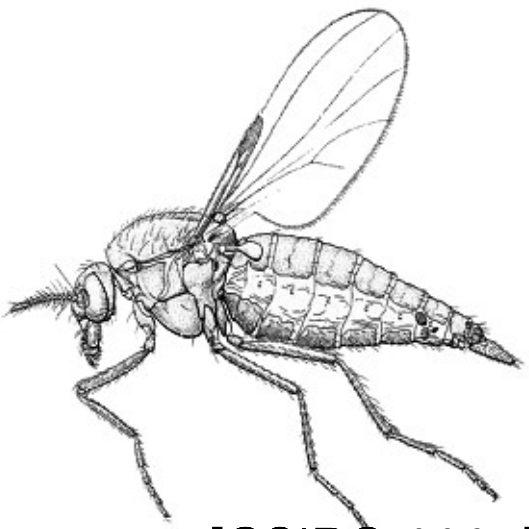
Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and precision $\theta = (\mu, \tau)$
- Model (conjugate prior): [Exercise: find the posterior]

$$p(y|\theta): \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \tau^{-1}), \quad n = 1, \ldots, N$$

$$p(\theta): \quad \tau \sim \text{Gamma}(a_0, b_0)$$

$$\mu|\tau \sim \mathcal{N}(\mu_0, (\rho_0 \tau)^{-1})$$



[CSIRO 2004]

- Exercise: check $p(\mu, \tau | y) \neq f_1(\mu, y) f_2(\tau, y)$

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

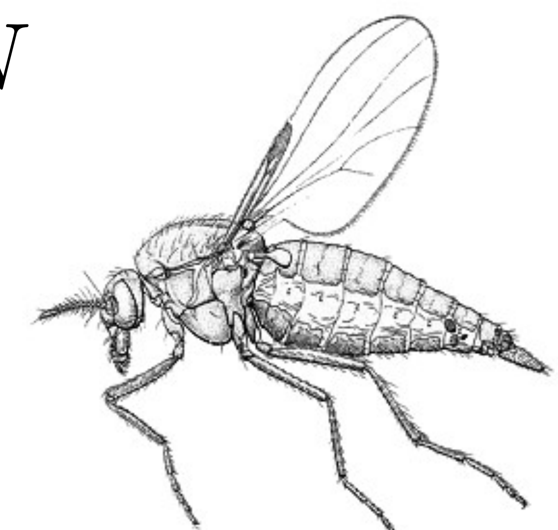
Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and precision $\theta = (\mu, \tau)$
- Model (conjugate prior): [Exercise: find the posterior]

$$p(y|\theta): \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \tau^{-1}), \quad n = 1, \ldots, N$$

$$p(\theta): \quad \tau \sim \text{Gamma}(a_0, b_0)$$

$$\mu|\tau \sim \mathcal{N}(\mu_0, (\rho_0 \tau)^{-1})$$



[CSIRO 2004]

- Exercise: check $p(\mu, \tau|y) \neq f_1(\mu, y) f_2(\tau, y)$
- MFVB approximation:

$$q^*(\mu, \tau) = q^*_\mu(\mu) q^*_\tau(\tau) = \text{argmin}_{q \in Q_{\text{MFVB}}} KL(q(\cdot) || p(\cdot|y))$$

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

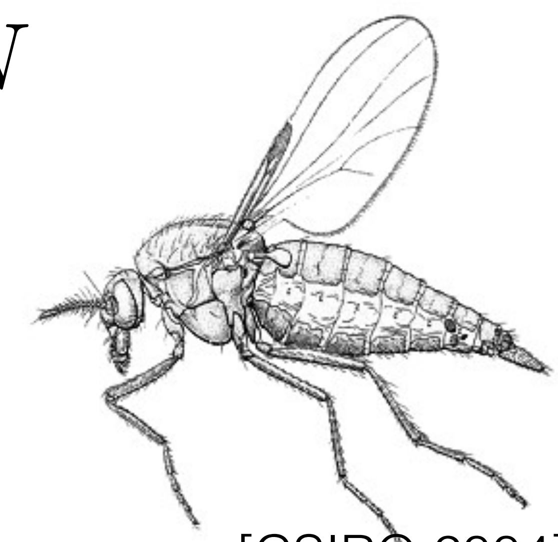
Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and precision $\theta = (\mu, \tau)$
- Model (conjugate prior): [Exercise: find the posterior]

$$p(y|\theta): \quad y_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \tau^{-1}), \quad n = 1, \ldots, N$$

$$p(\theta): \quad \tau \sim \text{Gamma}(a_0, b_0)$$

$$\mu|\tau \sim \mathcal{N}(\mu_0, (\rho_0\tau)^{-1})$$



[CSIRO 2004]

- Exercise: check $p(\mu, \tau|y) \neq f_1(\mu, y) f_2(\tau, y)$
- MFVB approximation:

$$q^*(\mu, \tau) = q^*_\mu(\mu) q^*_\tau(\tau) = \text{argmin}_{q \in Q_{\text{MFVB}}} KL(q(\cdot)||p(\cdot|y))$$

- Coordinate descent [Exercise: derive this] [Bishop 2006, Sec 10.1.3]

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

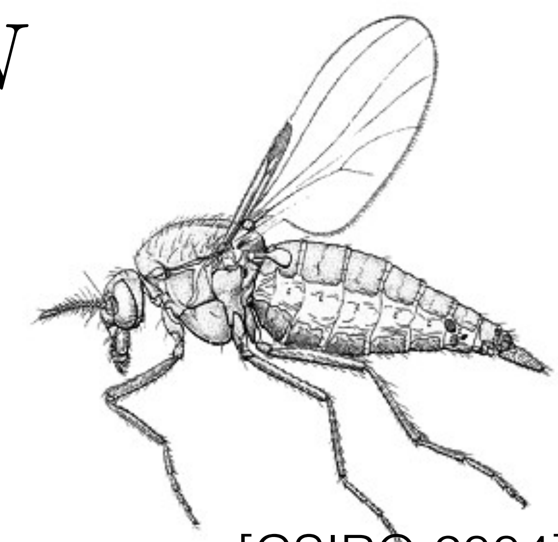
Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and precision $\theta = (\mu, \tau)$
- Model (conjugate prior): [Exercise: find the posterior]

$$p(y|\theta): \quad y_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \tau^{-1}), \quad n = 1, \ldots, N$$

$$p(\theta): \quad \tau \sim \text{Gamma}(a_0, b_0)$$

$$\mu|\tau \sim \mathcal{N}(\mu_0, (\rho_0 \tau)^{-1})$$



[CSIRO 2004]

- Exercise: check $p(\mu, \tau|y) \neq f_1(\mu, y) f_2(\tau, y)$
- MFVB approximation:

$$q^*(\mu, \tau) = q^*_\mu(\mu) q^*_\tau(\tau) = \text{argmin}_{q \in Q_{\text{MFVB}}} KL(q(\cdot)||p(\cdot|y))$$

- Coordinate descent [Exercise: derive this] [Bishop 2006, Sec 10.1.3]

$$q^*_\mu(\mu) = \mathcal{N}(\mu|\mu_N, \rho_N^{-1}) \qquad q^*_\tau(\tau) = \text{Gamma}(\tau|a_N, b_N)$$

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

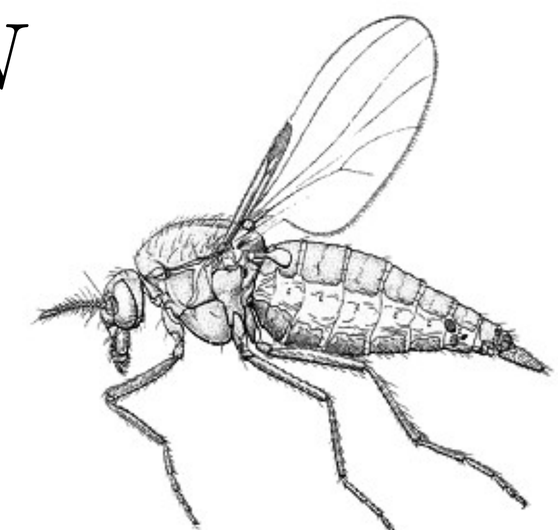
Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and precision $\theta = (\mu, \tau)$
- Model (conjugate prior): [Exercise: find the posterior]

$$p(y|\theta): \quad y_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \tau^{-1}), \quad n = 1, \ldots, N$$

$$p(\theta): \quad \tau \sim \text{Gamma}(a_0, b_0)$$

$$\mu|\tau \sim \mathcal{N}(\mu_0, (\rho_0 \tau)^{-1})$$



[CSIRO 2004]

- Exercise: check $p(\mu, \tau | y) \neq f_1(\mu, y) f_2(\tau, y)$
- MFVB approximation:

$$q^*(\mu, \tau) = q^*_\mu(\mu) q^*_\tau(\tau) = \text{argmin}_{q \in Q_{\text{MFVB}}} KL(q(\cdot) || p(\cdot|y))$$

- Coordinate descent [Exercise: derive this] [Bishop 2006, Sec 10.1.3]

$$q^*_\mu(\mu) = \mathcal{N}(\mu | \mu_N, \rho_N^{-1}) \qquad q^*_\tau(\tau) = \text{Gamma}(\tau | a_N, b_N)$$

"variational parameters"

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

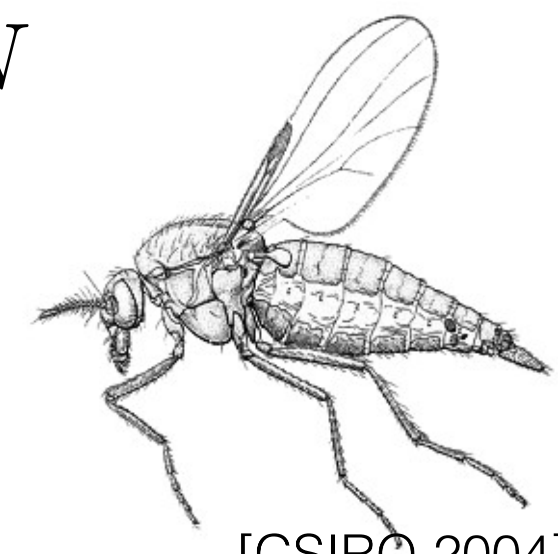
Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and precision $\theta = (\mu, \tau)$
- Model (conjugate prior): [Exercise: find the posterior]

$$p(y|\theta): \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \tau^{-1}), \quad n = 1, \ldots, N$$

$$p(\theta): \quad \tau \sim \text{Gamma}(a_0, b_0)$$

$$\mu|\tau \sim \mathcal{N}(\mu_0, (\rho_0\tau)^{-1})$$



[CSIRO 2004]

- Exercise: check $p(\mu, \tau|y) \neq f_1(\mu, y) f_2(\tau, y)$
- MFVB approximation:

$$q^*(\mu, \tau) = q^*_\mu(\mu) q^*_\tau(\tau) = \text{argmin}_{q \in Q_{\text{MFVB}}} KL(q(\cdot)||p(\cdot|y))$$

- Coordinate descent [Exercise: derive this] [Bishop 2006, Sec 10.1.3]

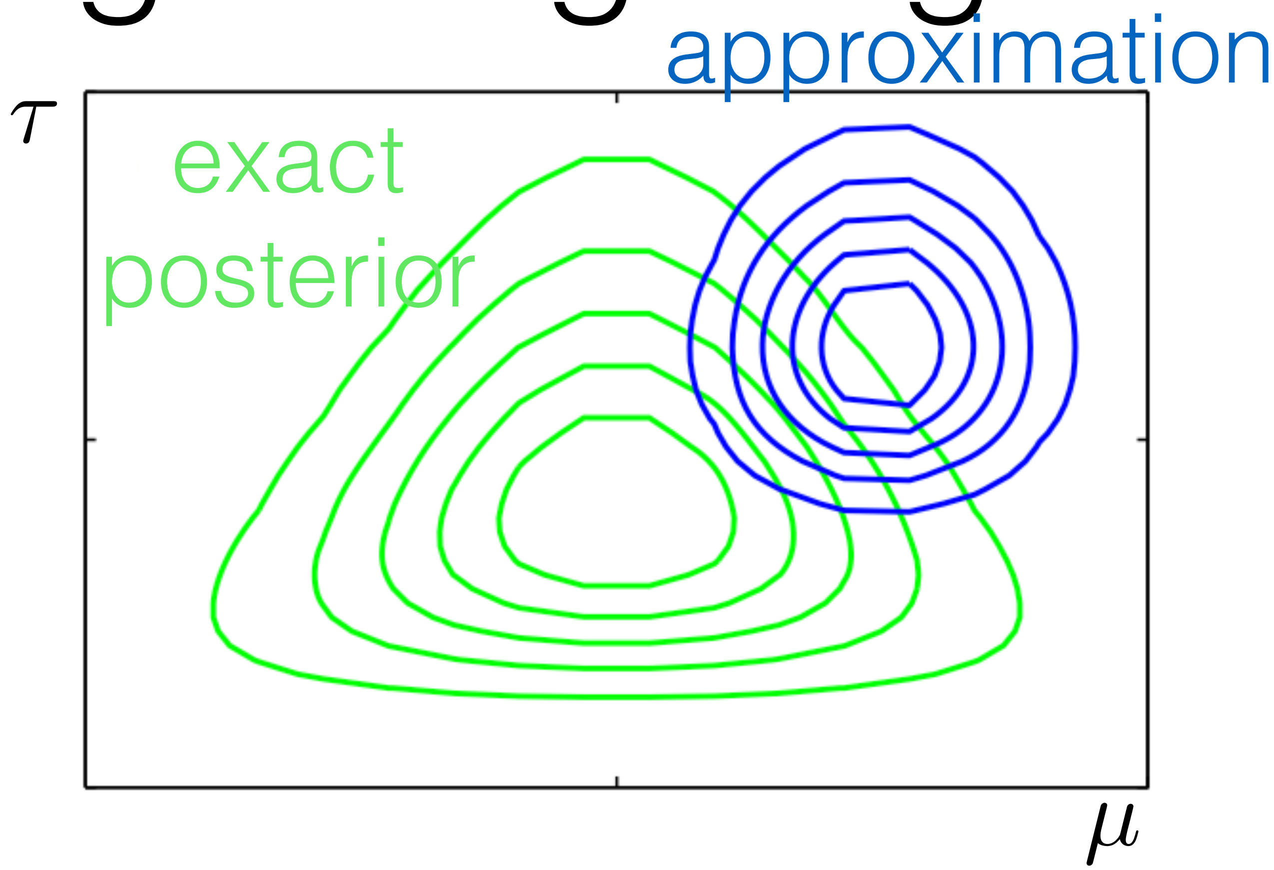
$$q^*_\mu(\mu) = \mathcal{N}(\mu|\mu_N, \rho_N^{-1}) \qquad q^*_\tau(\tau) = \text{Gamma}(\tau|a_N, b_N)$$

  - Iterate: $(\mu_N, \rho_N) = f(a_N, b_N)$
    $(a_N, b_N) = g(\mu_N, \rho_N)$

"variational parameters"

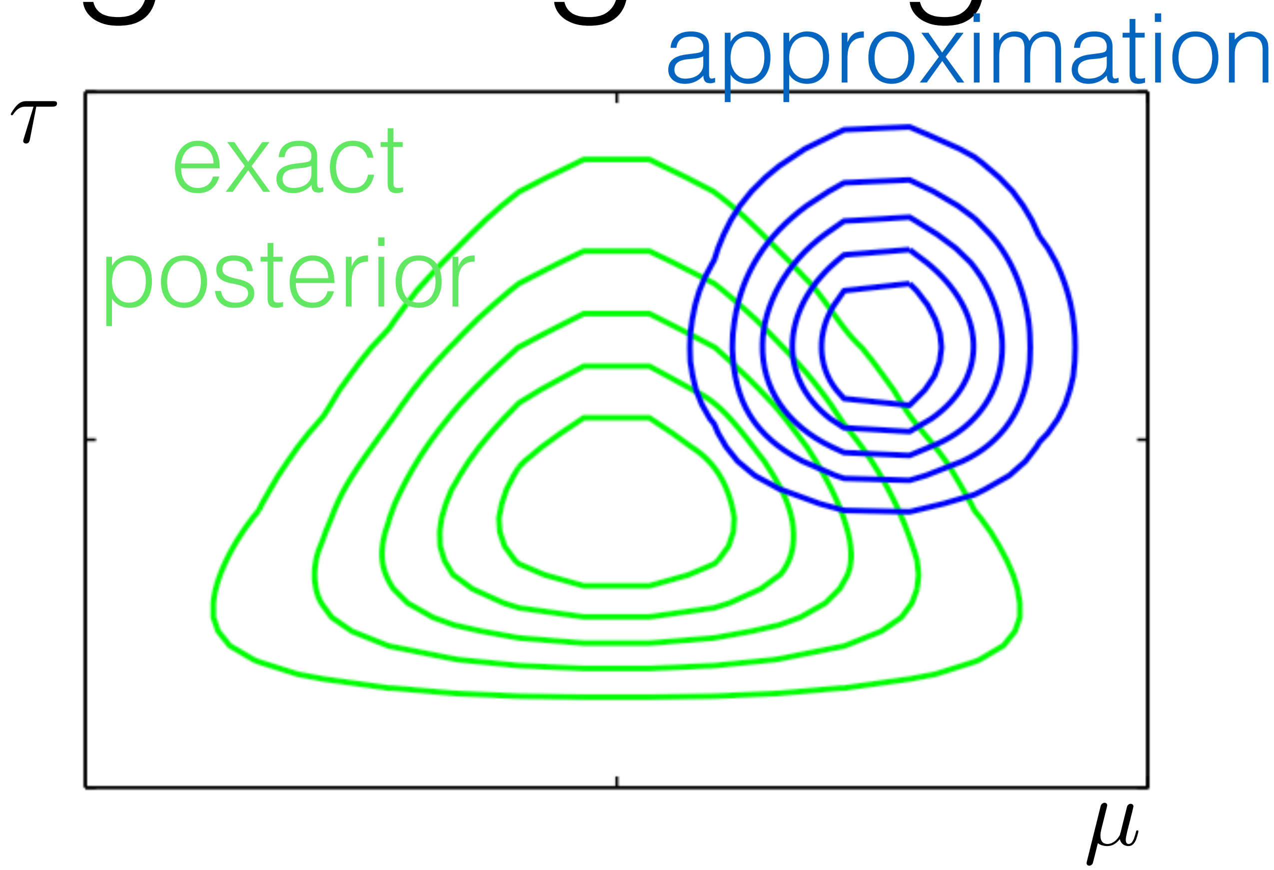
[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

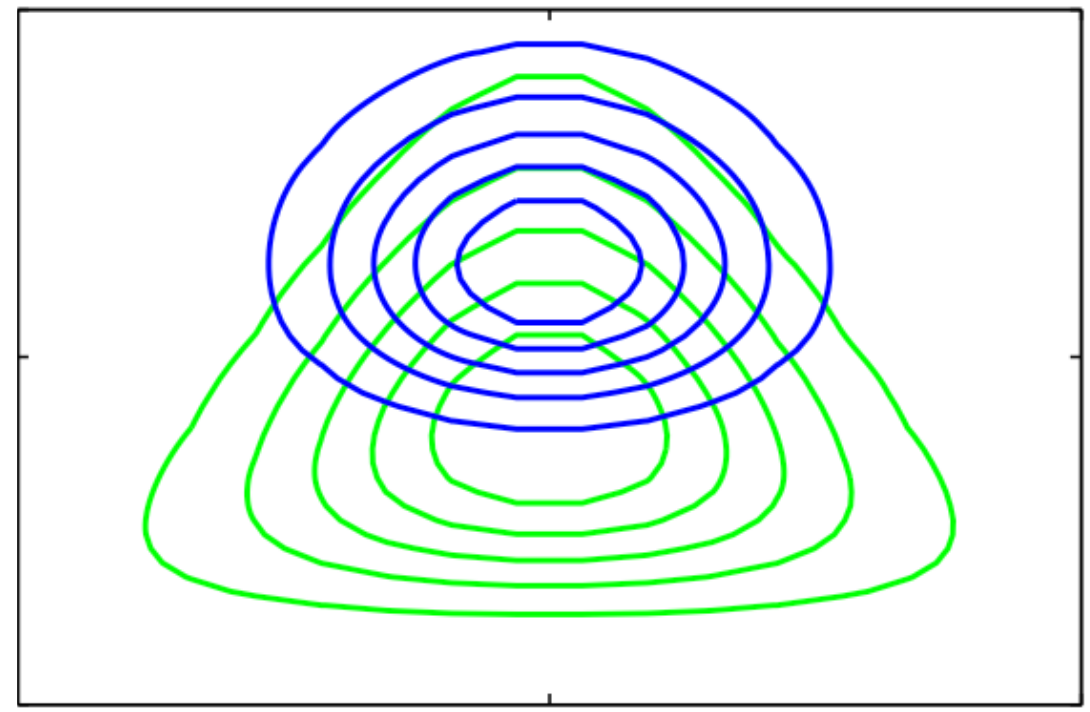
Midge wing length



[Bishop 2006]

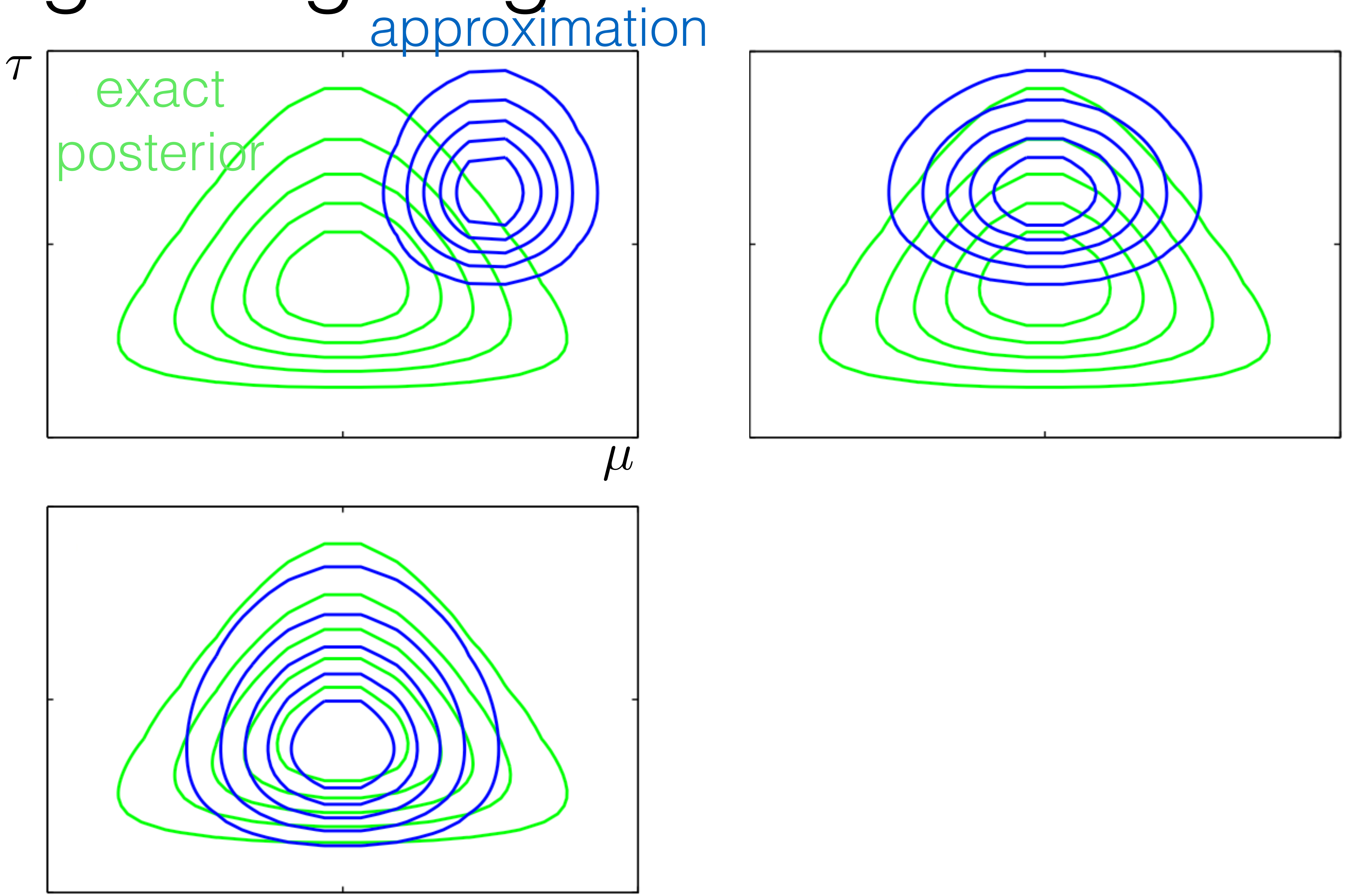
Midge wing length





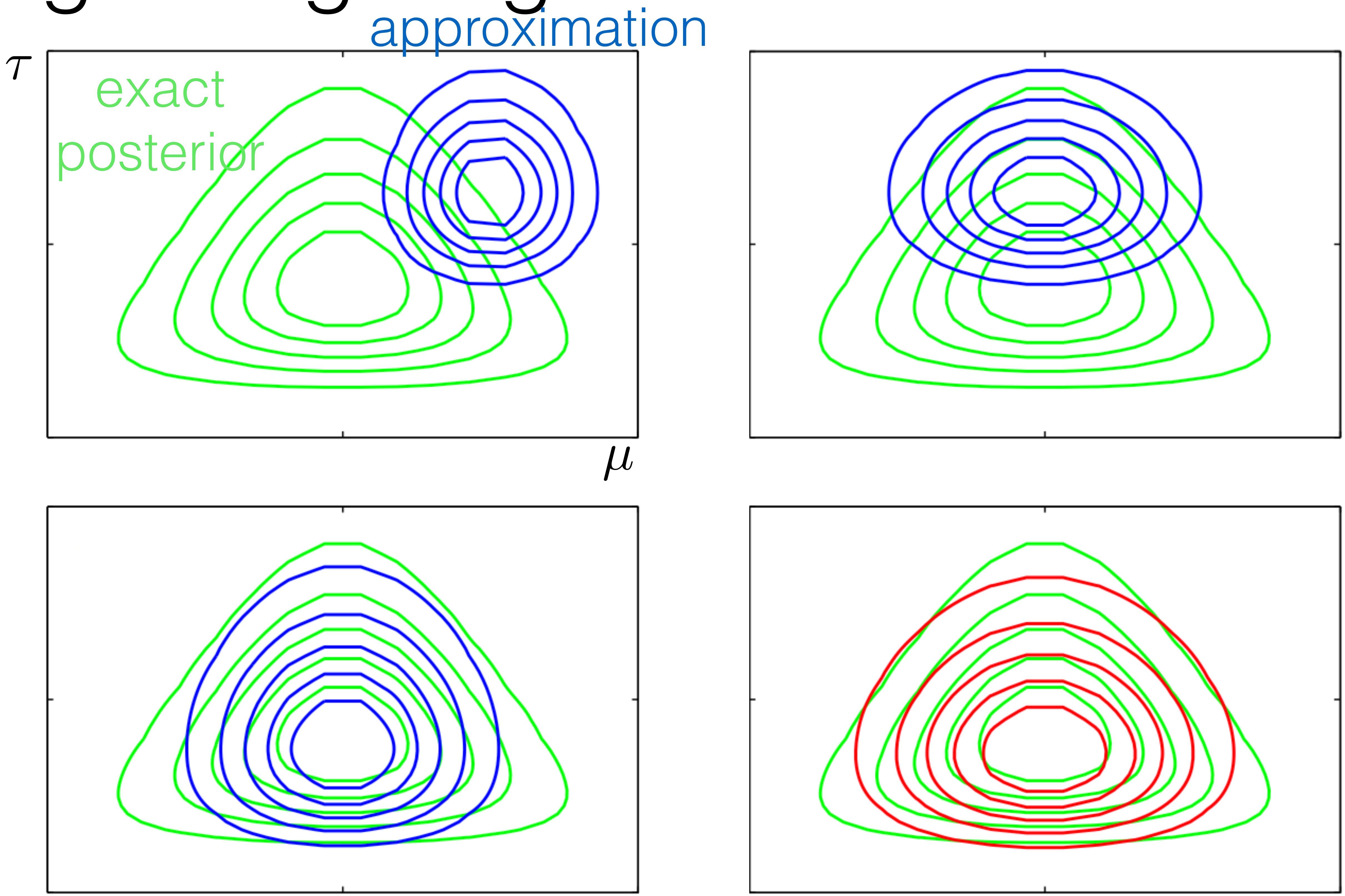
[Bishop 2006]

Midge wing length



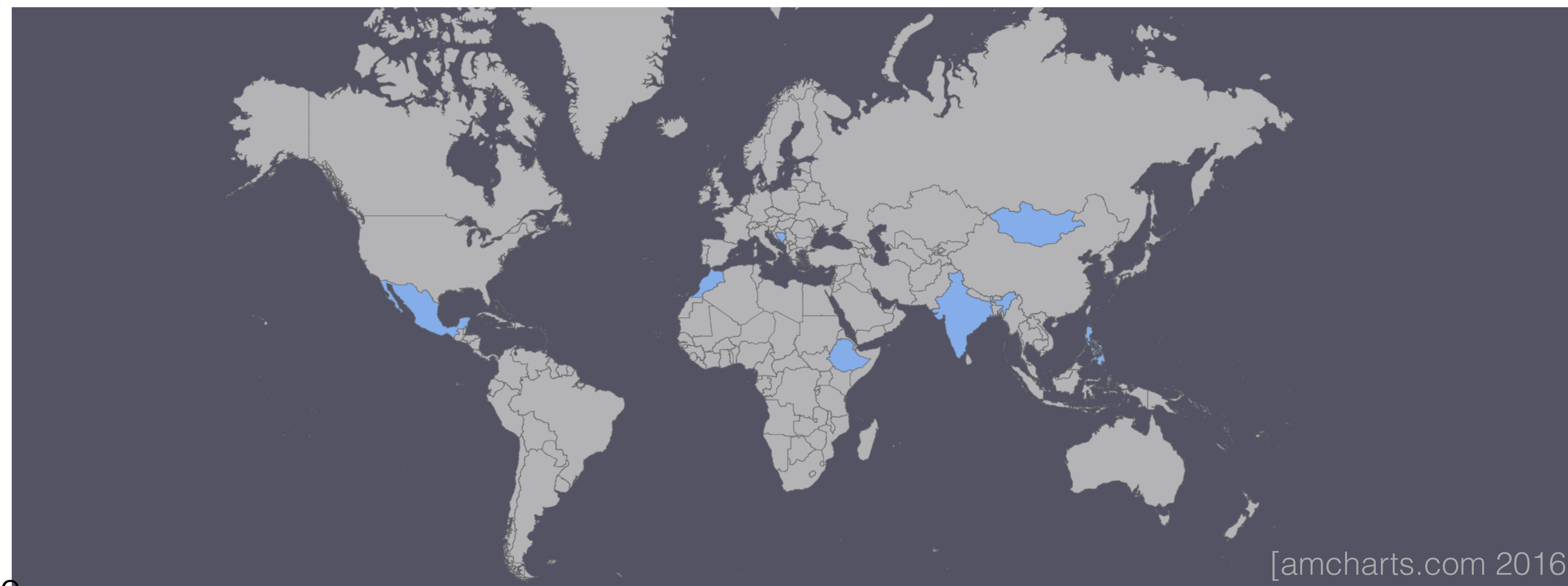
[Bishop 2006]

Midge wing length



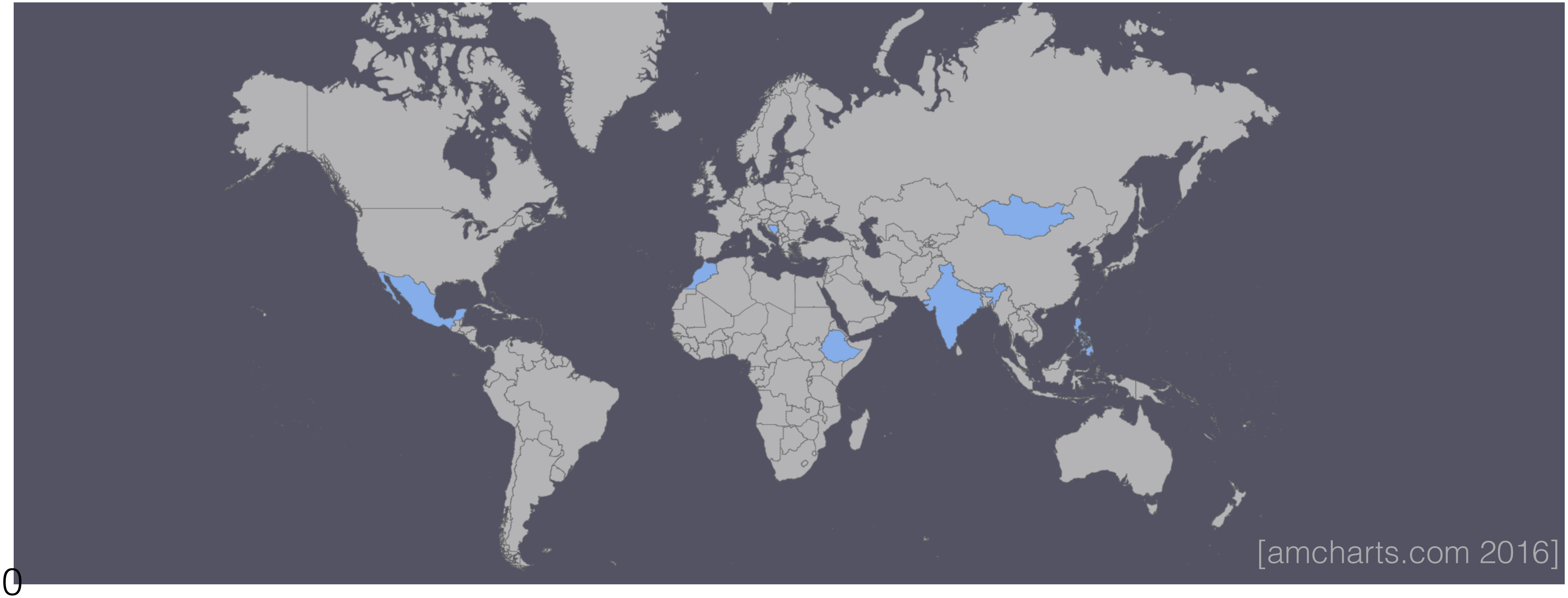
[Bishop 2006]

Microcredit Experiment



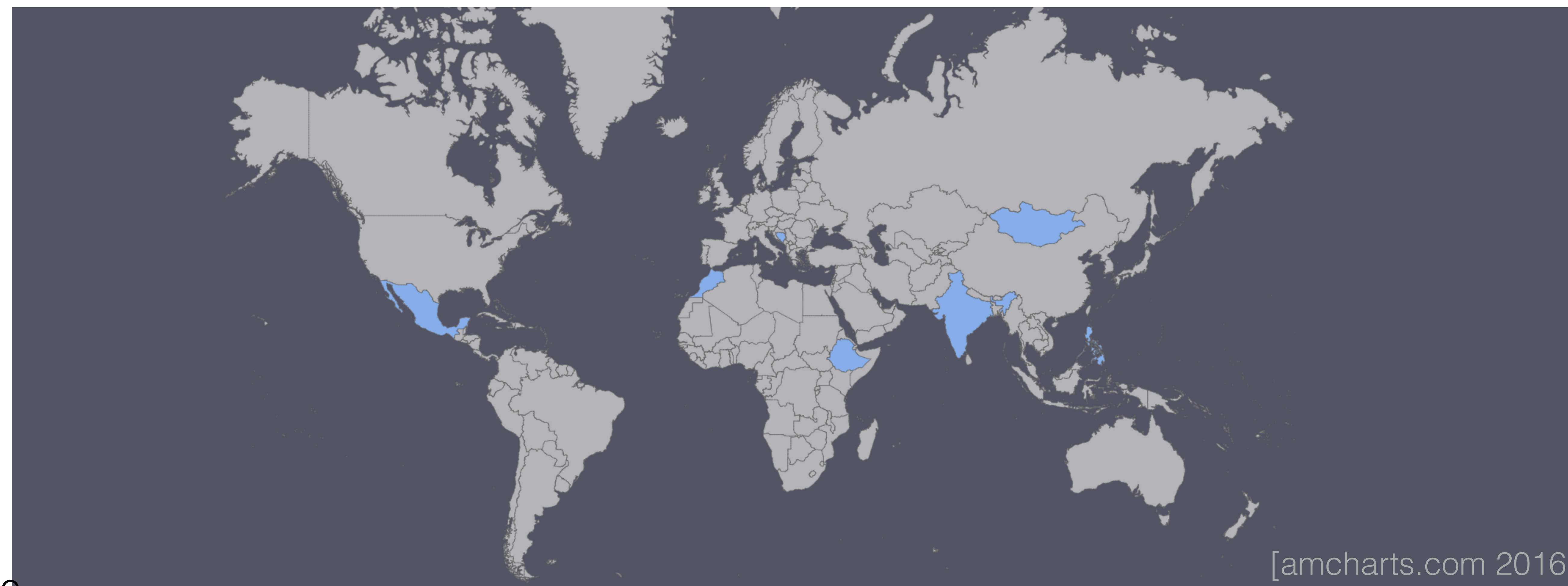
Microcredit Experiment

- Simplified from Meager (2018a)



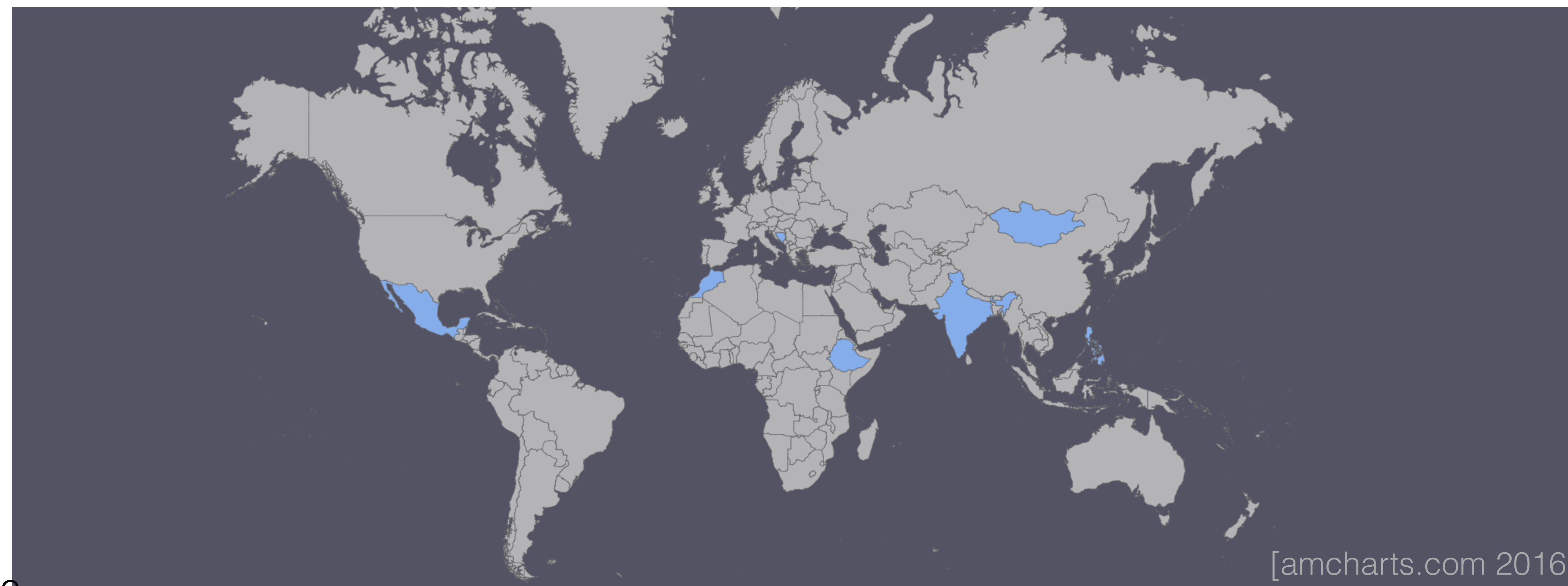
Microcredit Experiment

- Simplified from Meager (2018a)
- K = 7 microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)



Microcredit Experiment

- Simplified from Meager (2018a)
- K = 7 microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in kth site (~900 to ~17K)



Microcredit Experiment

- Simplified from Meager (2018a)
- $K = 7$ microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in kth site (~900 to ~17K)

Microcredit Experiment

- Simplified from Meager (2018a)
- $K = 7$ microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in kth site (~900 to ~17K)
- Profit of nth business at kth site:

Microcredit Experiment

- Simplified from Meager (2018a)
- K = 7 microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in kth site (~900 to ~17K)
- Profit of nth business at kth site:

profit → y_{kn}

Microcredit Experiment

- Simplified from Meager (2018a)
- K = 7 microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in kth site (~900 to ~17K)
- Profit of nth business at kth site:

profit

$$y_{kn} \stackrel{indep}{\sim} \mathcal{N}(\qquad\qquad , \quad)$$

Microcredit Experiment

- Simplified from Meager (2018a)
- K = 7 microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in kth site (~900 to ~17K)
- Profit of nth business at kth site:

profit

$$y_{kn} \overset{indep}{\sim} \mathcal{N}(\mu_k \qquad , \quad)$$

Microcredit Experiment

- Simplified from Meager (2018a)
- K = 7 microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in kth site (~900 to ~17K)
- Profit of nth business at kth site:

profit

$$y_{kn} \overset{indep}{\sim} \mathcal{N}(\mu_k + T_{kn}\tau_k, \quad)$$

Microcredit Experiment

- Simplified from Meager (2018a)
- K = 7 microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in kth site (~900 to ~17K)
- Profit of nth business at kth site:

profit

1 if microcredit

$$y_{kn} \overset{indep}{\sim} \mathcal{N}(\mu_k + T_{kn}\tau_k, \quad)$$

Microcredit Experiment

- Simplified from Meager (2018a)
- K = 7 microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in kth site (~900 to ~17K)
- Profit of nth business at kth site:

1 if microcredit

profit

$$y_{kn} \overset{indep}{\sim} \mathcal{N}(\mu_k + T_{kn}\tau_k, \quad)$$

Microcredit Experiment

- Simplified from Meager (2018a)
- K = 7 microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in kth site (~900 to ~17K)
- Profit of nth business at kth site:

profit → 1 if microcredit →

$$y_{kn} \overset{indep}{\sim} \mathcal{N}(\mu_k + T_{kn}\tau_k, \sigma_k^2)$$

Microcredit Experiment

- Simplified from Meager (2018a)
- $K = 7$ microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in kth site (~900 to ~17K)
- Profit of nth business at kth site:

1 if microcredit

profit

$$y_{kn} \overset{indep}{\sim} \mathcal{N}(\mu_k + T_{kn}\tau_k, \sigma_k^2)$$

Microcredit Experiment

- Simplified from Meager (2018a)
- K = 7 microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in kth site (~900 to ~17K)
- Profit of nth business at kth site:

1 if microcredit

profit

$$y_{kn} \overset{indep}{\sim} \mathcal{N}(\mu_k + T_{kn}\tau_k, \sigma_k^2)$$

- Priors and hyperpriors:

Microcredit Experiment

- Simplified from Meager (2018a)
- K = 7 microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in kth site (~900 to ~17K)
- Profit of nth business at kth site:

1 if microcredit

profit

$$y_{kn} \overset{indep}{\sim} \mathcal{N}(\mu_k + T_{kn}\tau_k, \sigma_k^2)$$

- Priors and hyperpriors:

$$\begin{pmatrix} \mu_k \\ \tau_k \end{pmatrix} \overset{iid}{\sim} \mathcal{N}\left(\begin{pmatrix} \mu \\ \tau \end{pmatrix}, C\right)$$

Microcredit Experiment

- Simplified from Meager (2018a)
- K = 7 microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in kth site (~900 to ~17K)
- Profit of nth business at kth site:

1 if microcredit

profit

$$y_{kn} \stackrel{indep}{\sim} \mathcal{N}(\mu_k + T_{kn}\tau_k, \sigma_k^2)$$

- Priors and hyperpriors:

$$\begin{pmatrix} \mu_k \\ \tau_k \end{pmatrix} \stackrel{iid}{\sim} \mathcal{N}\left(\begin{pmatrix} \mu \\ \tau \end{pmatrix}, C\right)$$

$$\sigma_k^{-2} \stackrel{iid}{\sim} \Gamma(a, b)$$

Microcredit Experiment

- Simplified from Meager (2018a)
- K = 7 microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- N_k businesses in kth site (~900 to ~17K)
- Profit of nth business at kth site:

profit → y_{kn}; 1 if microcredit → T_{kn}

$$y_{kn} \overset{indep}{\sim} \mathcal{N}(\mu_k + T_{kn}\tau_k, \sigma_k^2)$$

- Priors and hyperpriors:

$$\begin{pmatrix} \mu_k \\ \tau_k \end{pmatrix} \overset{iid}{\sim} \mathcal{N}\left(\begin{pmatrix} \mu \\ \tau \end{pmatrix}, C\right) \qquad \begin{pmatrix} \mu \\ \tau \end{pmatrix} \overset{iid}{\sim} \mathcal{N}\left(\begin{pmatrix} \mu_0 \\ \tau_0 \end{pmatrix}, \Lambda^{-1}\right)$$

$$\sigma_k^{-2} \overset{iid}{\sim} \Gamma(a, b) \qquad C \sim \mathrm{Sep\&LKJ}(\eta, c, d)$$

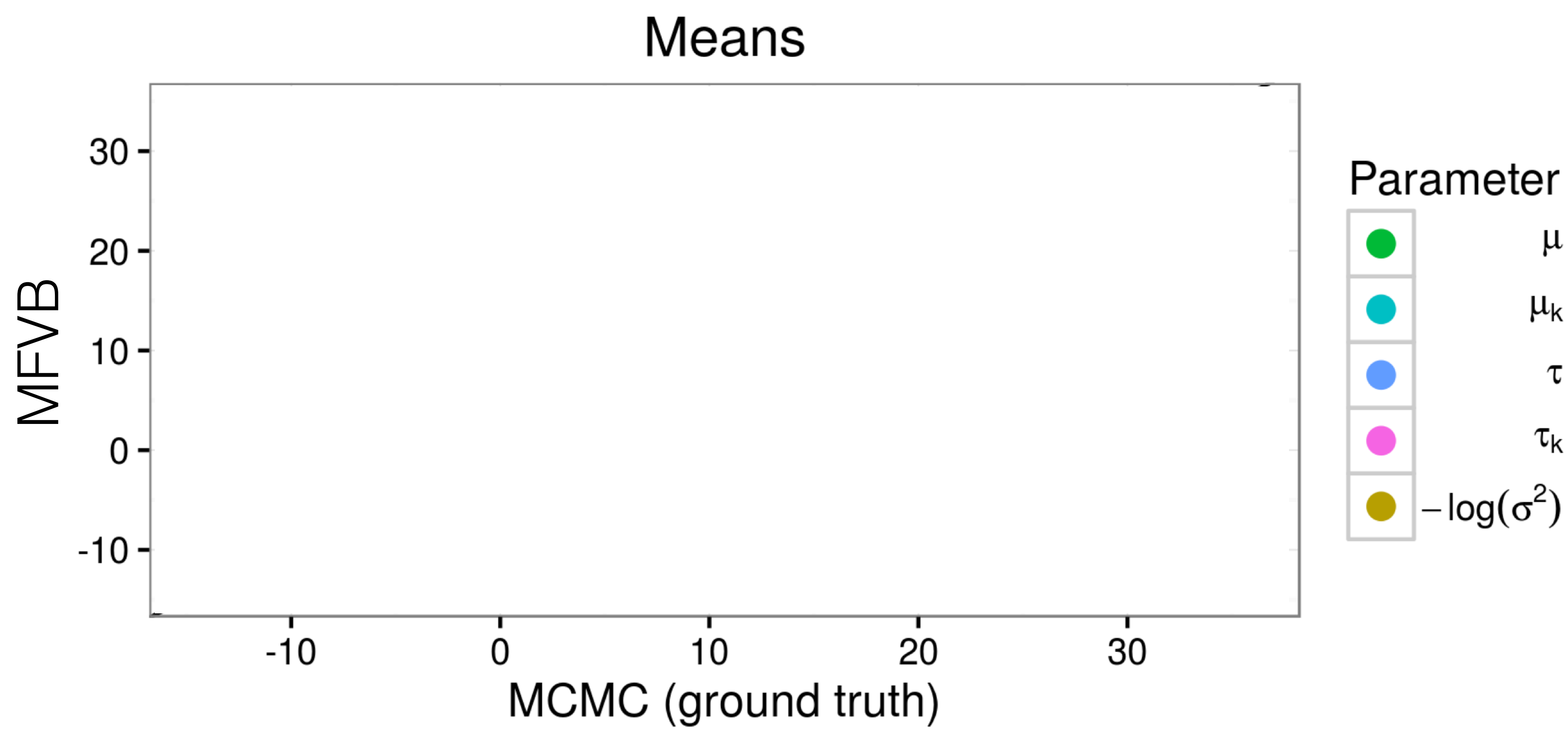
Microcredit

MFVB: Do we need to check the output?

Microcredit

MFVB: How will we know if it's working?

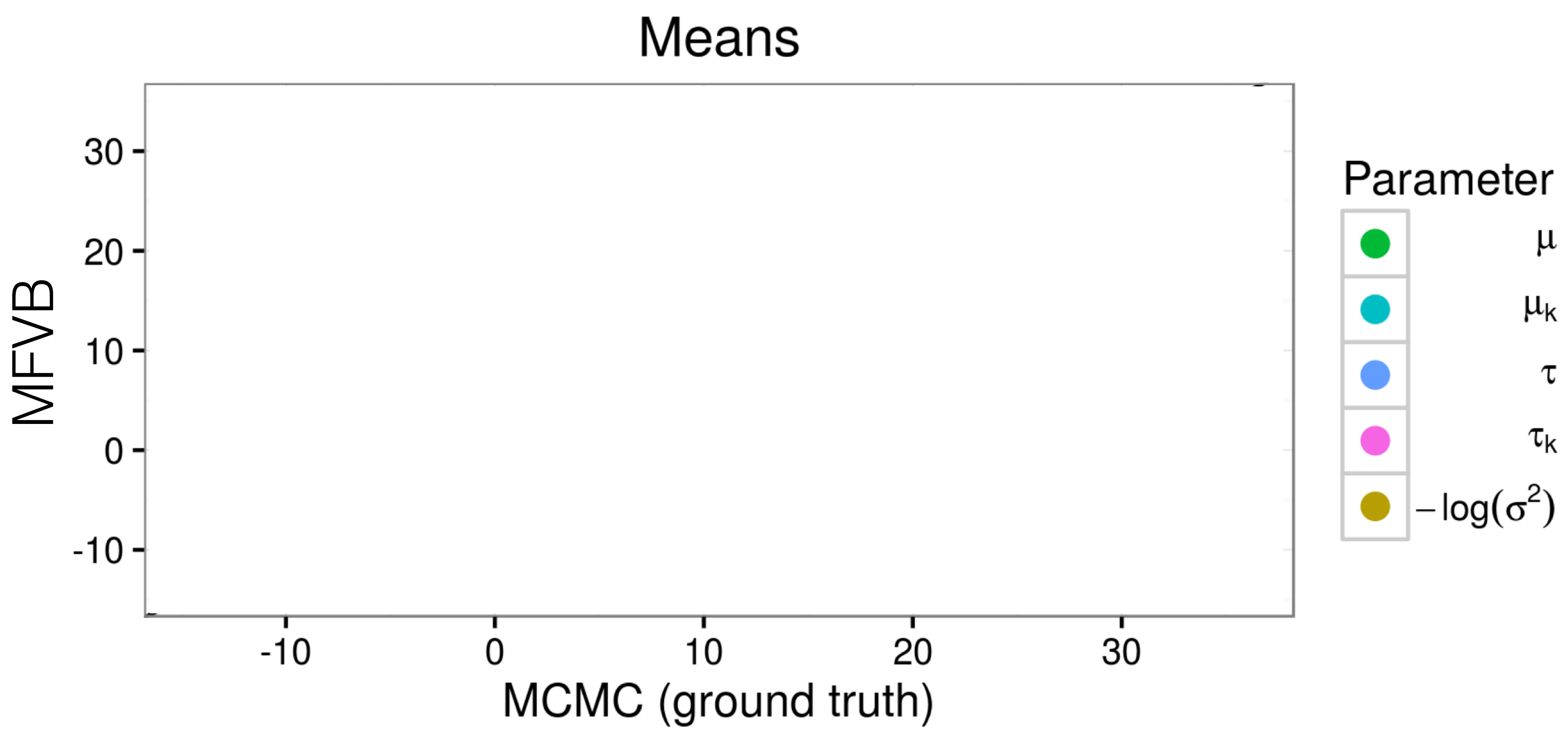
Microcredit



[Giordano, Broderick, Meager, Huggins, Jordan 2016]

Microcredit

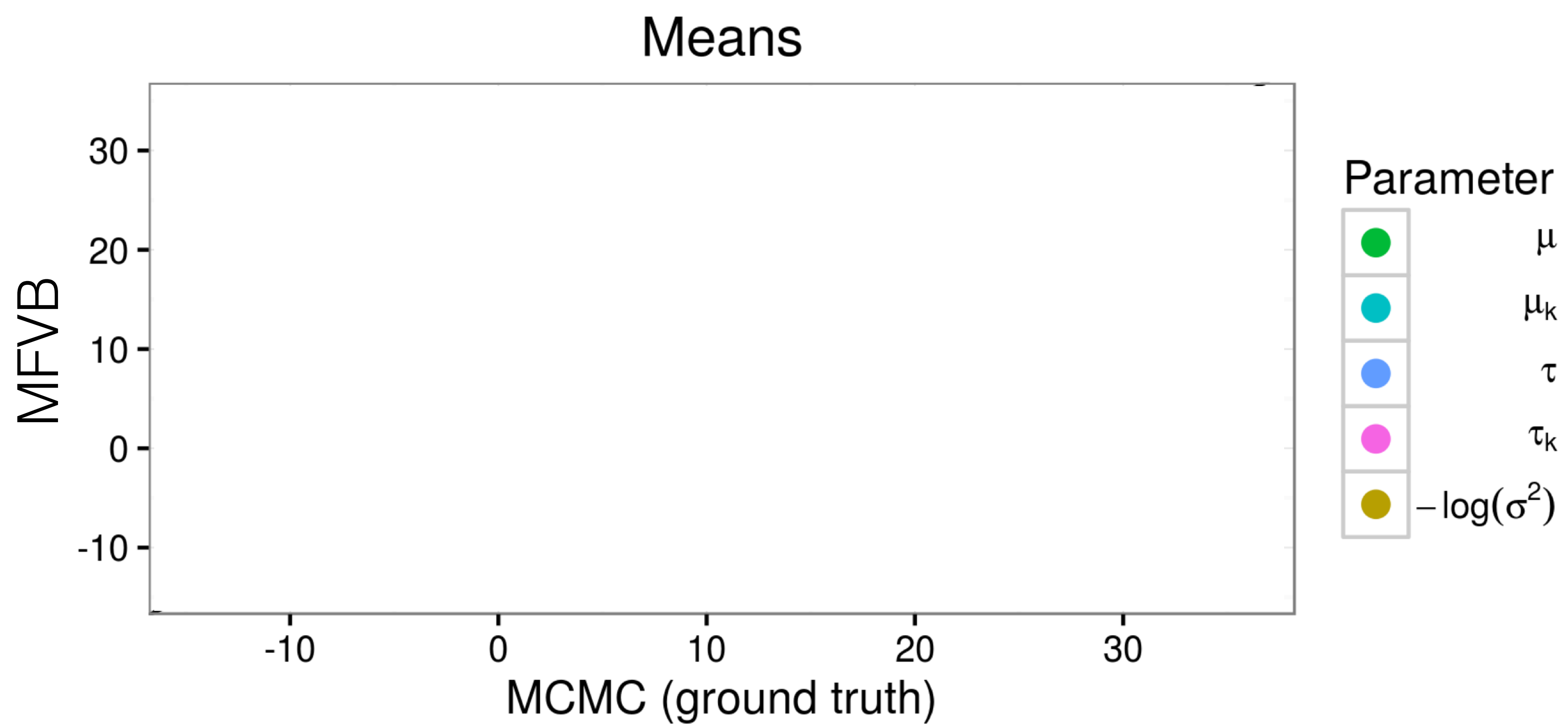
- *One set* of 2500 MCMC draws: **45 minutes**



[Giordano, Broderick, Meager, Huggins, Jordan 2016]

Microcredit

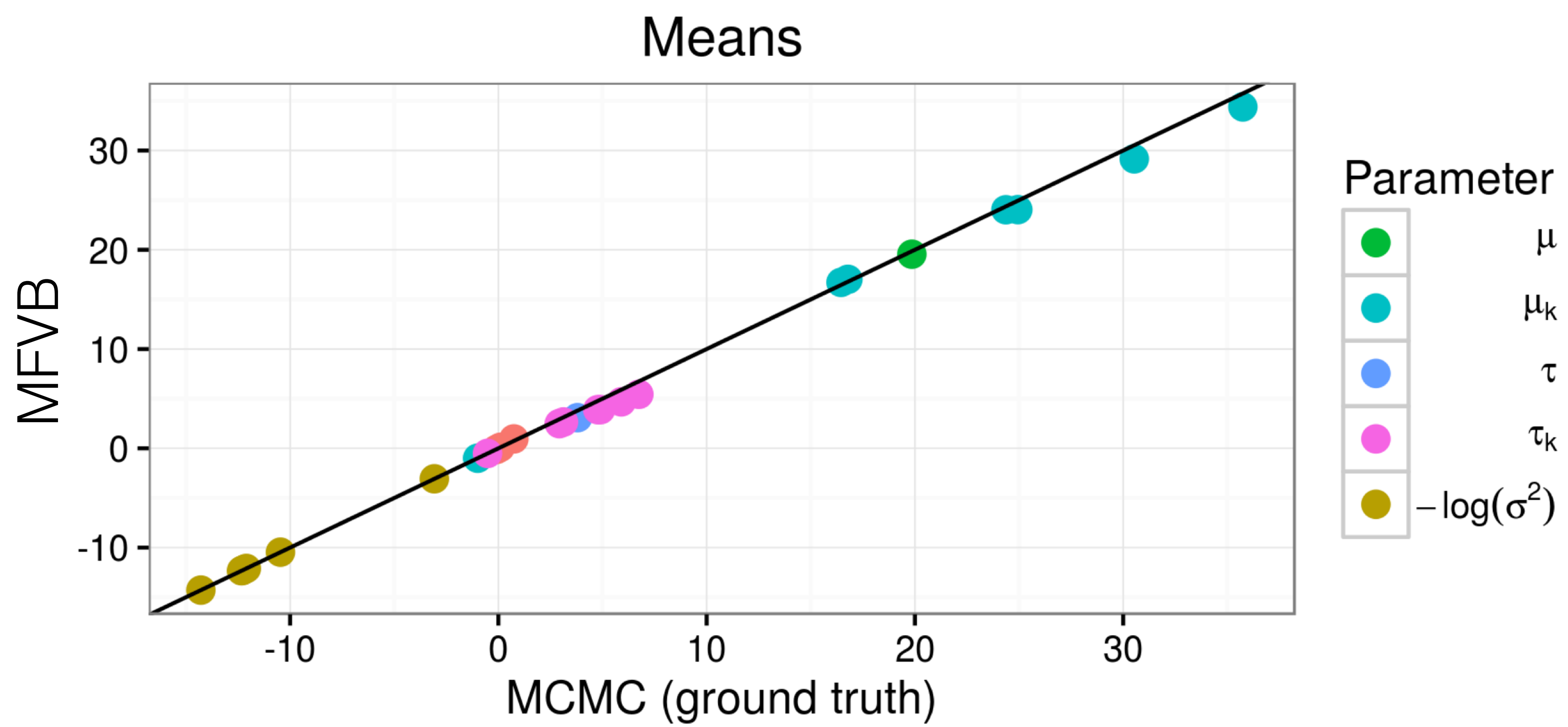
- *One set* of 2500 MCMC draws: **45 minutes**
- MFVB optimization: **<1 min**



[Giordano, Broderick, Meager, Huggins, Jordan 2016]

Microcredit

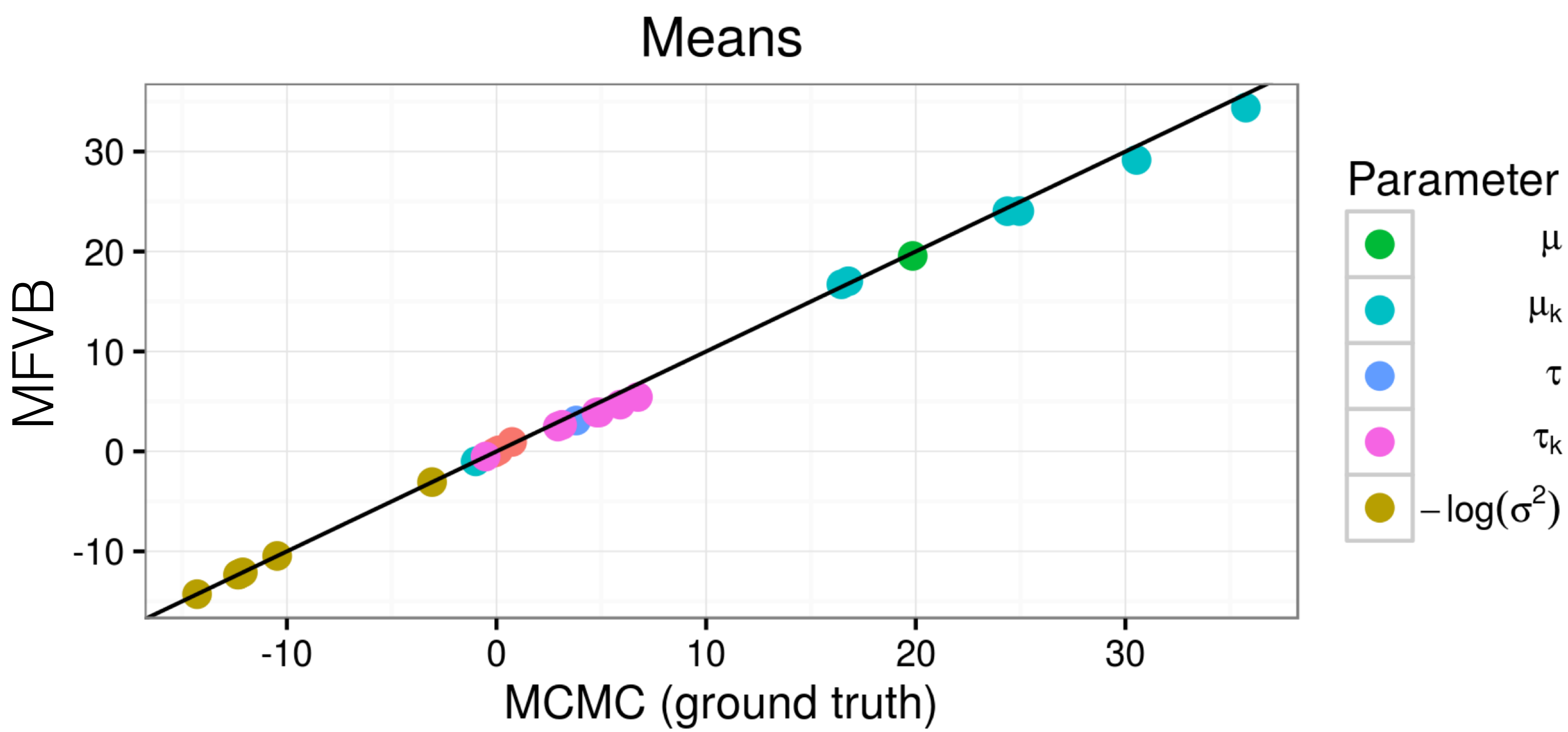
- *One set* of 2500 MCMC draws: **45 minutes**
- MFVB optimization: **<1 min**



[Giordano, Broderick, Meager, Huggins, Jordan 2016]

Microcredit

- *One set* of 2500 MCMC draws: **45 minutes**
- MFVB optimization: **<1 min**



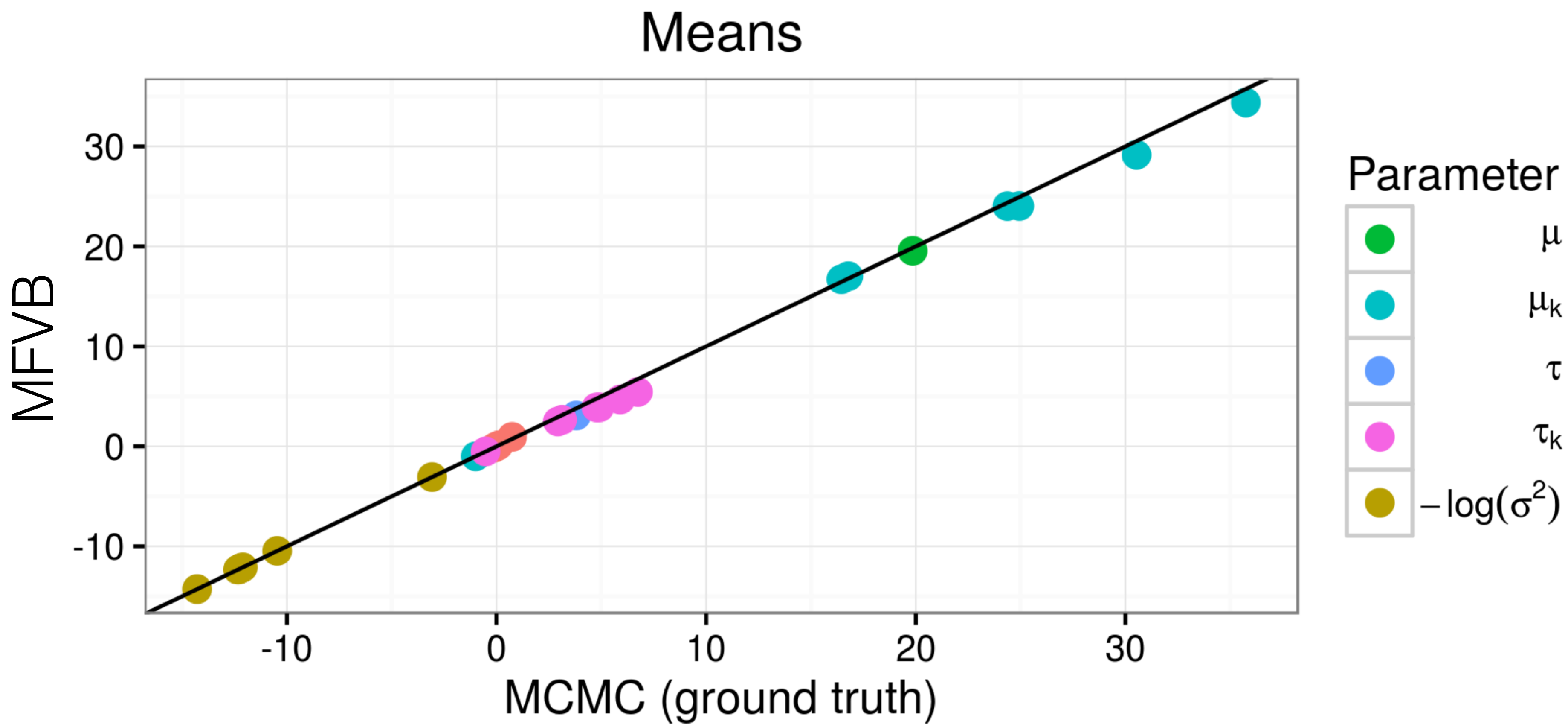
Criteo Online Ads Experiment

- Click-through conversion prediction
- Q: Will a customer (e.g.) buy a product after clicking?

[Giordano, Broderick, Meager, Huggins, Jordan 2016]

Microcredit

- *One set* of 2500 MCMC draws: **45 minutes**
- MFVB optimization: **<1 min**



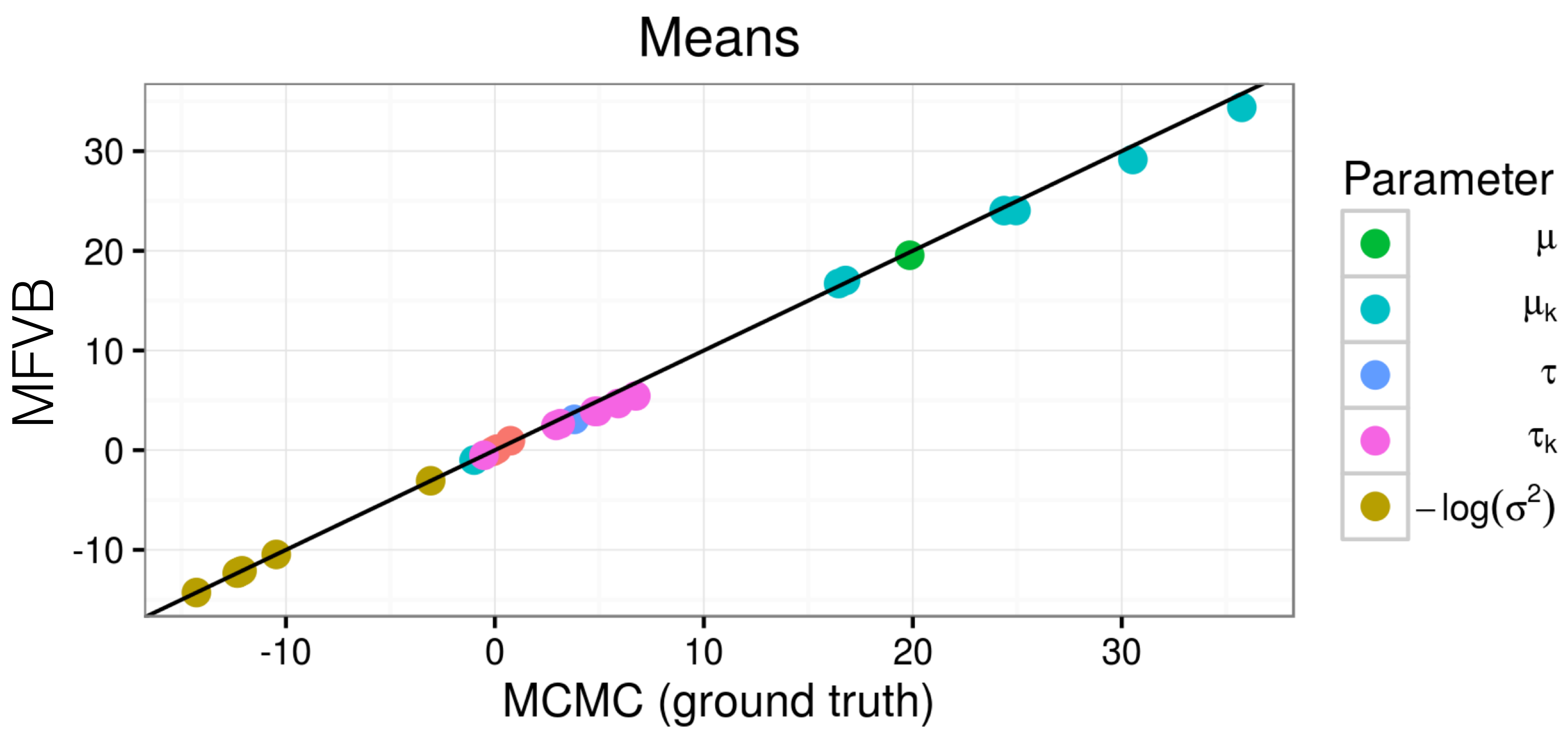
Criteo Online Ads Experiment

- Click-through conversion prediction
- Q: Will a customer (e.g.) buy a product after clicking?
- Q: How predictive of conversion are different features?

[Giordano, Broderick, Meager, Huggins, Jordan 2016]

Microcredit

- *One set* of 2500 MCMC draws: **45 minutes**
- MFVB optimization: **<1 min**



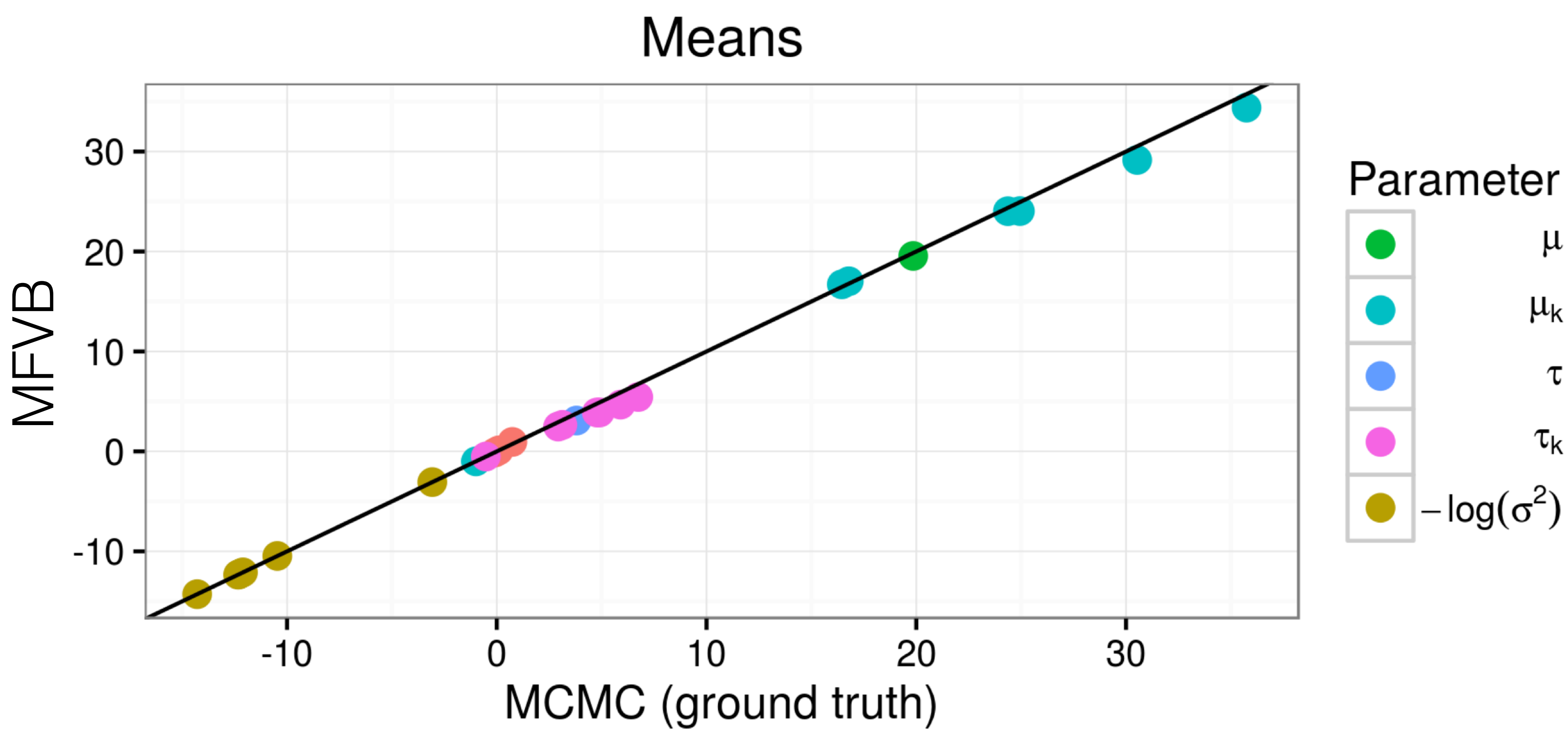
Criteo Online Ads Experiment

- Click-through conversion prediction
- Q: Will a customer (e.g.) buy a product after clicking?
- Q: How predictive of conversion are different features?
- Logistic GLMM

[Giordano, Broderick, Meager, Huggins, Jordan 2016; Giordano, Broderick, Jordan 2018]

Microcredit

- *One set* of 2500 MCMC draws: **45 minutes**
- MFVB optimization: **<1 min**



Criteo Online Ads Experiment

- Click-through conversion prediction
- Q: Will a customer (e.g.) buy a product after clicking?
- Q: How predictive of conversion are different features?
- Logistic GLMM; N = 61,895 subset to compare to MCMC

[Giordano, Broderick, Meager, Huggins, Jordan 2016; Giordano, Broderick, Jordan 2018]

Criteo Online Ads Experiment

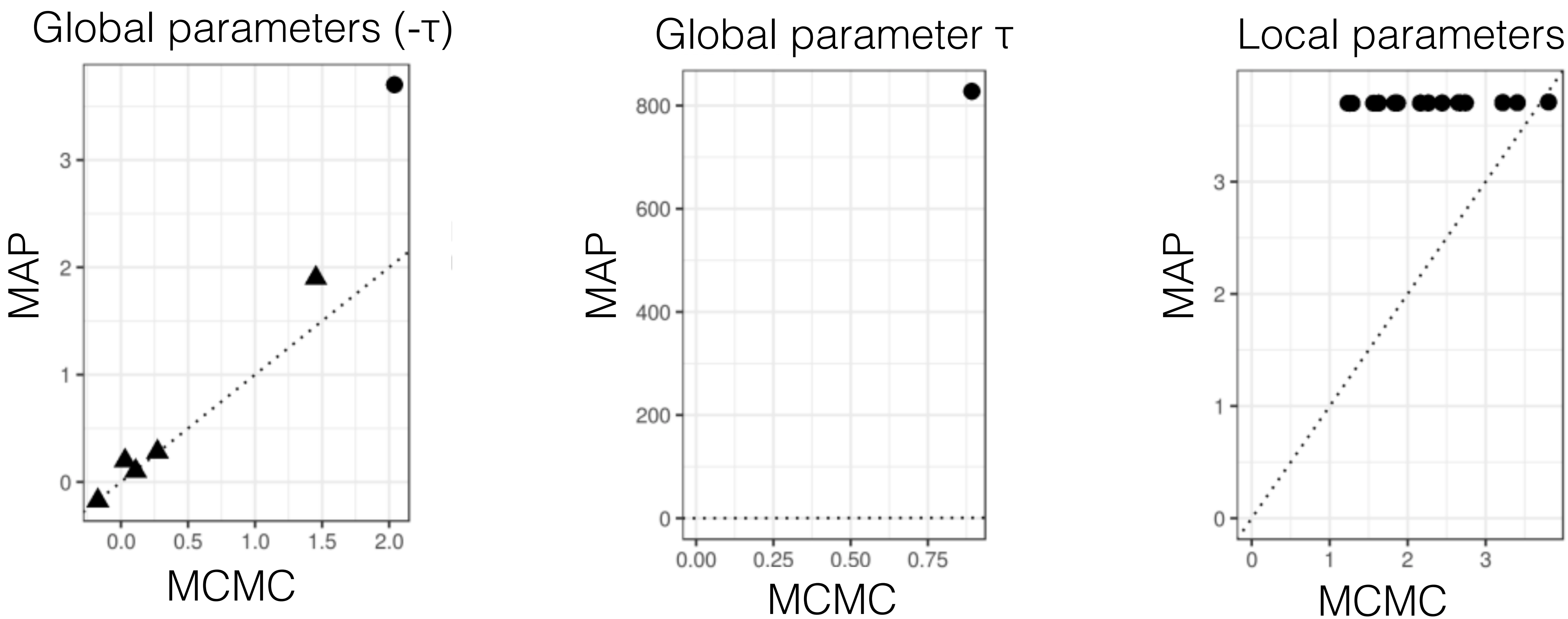
[Giordano, Broderick, Jordan 2018]

Criteo Online Ads Experiment

- MAP: **12 s**

[Giordano, Broderick, Jordan 2018]

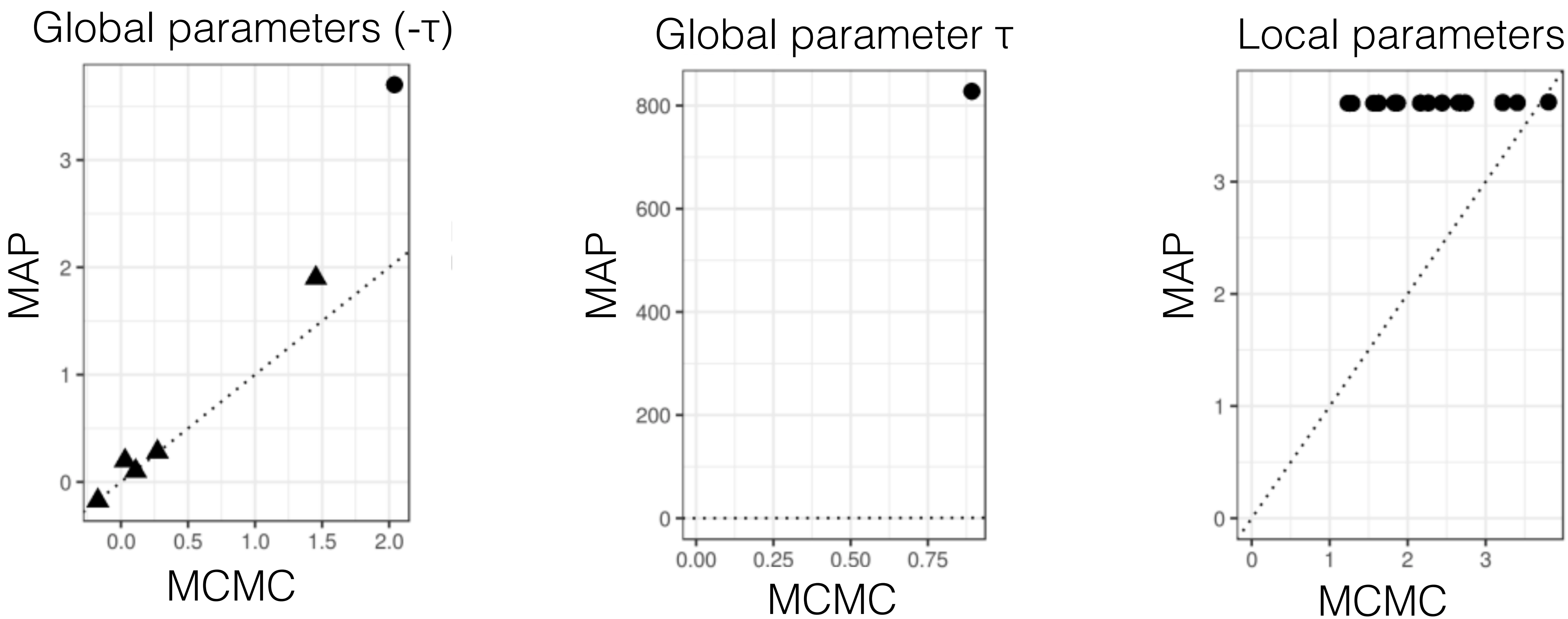
Criteo Online Ads Experiment



- MAP: **12 s**

[Giordano, Broderick, Jordan 2018]

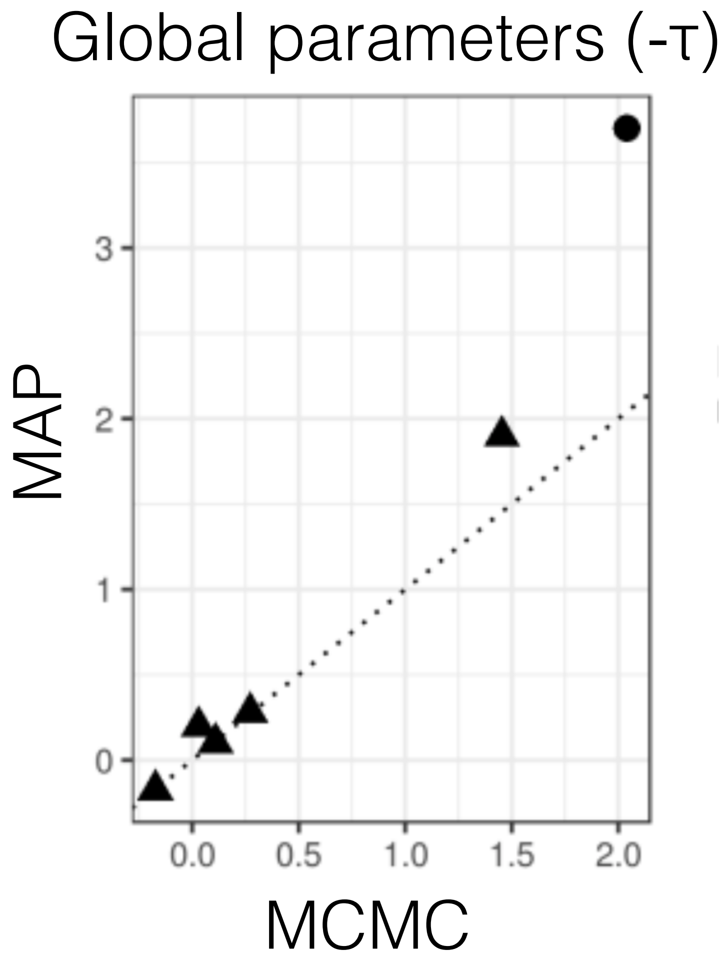
Criteo Online Ads Experiment

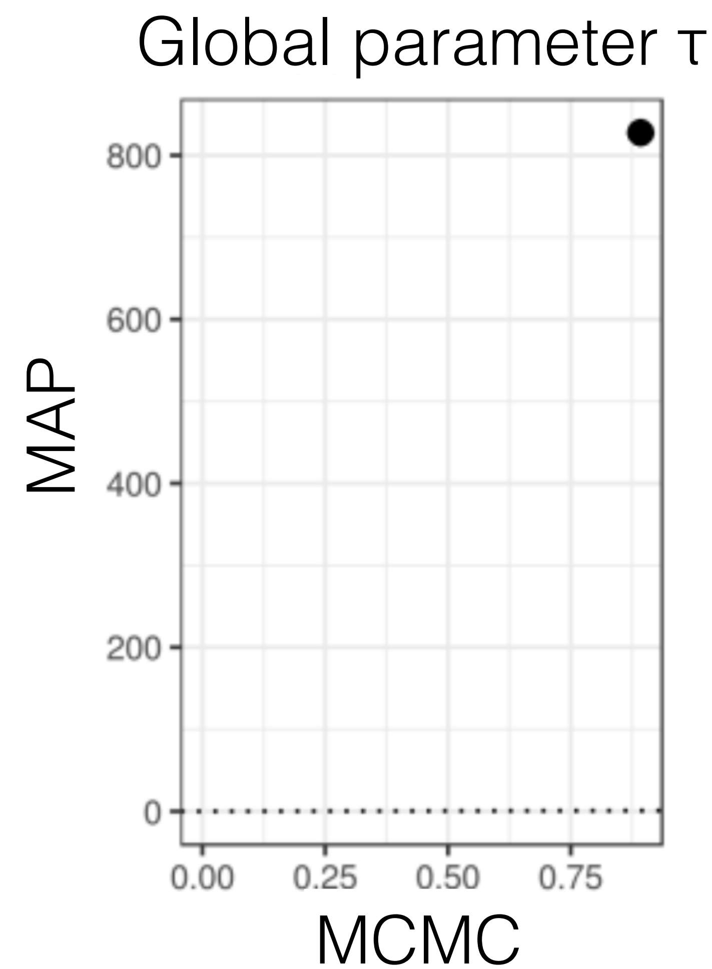


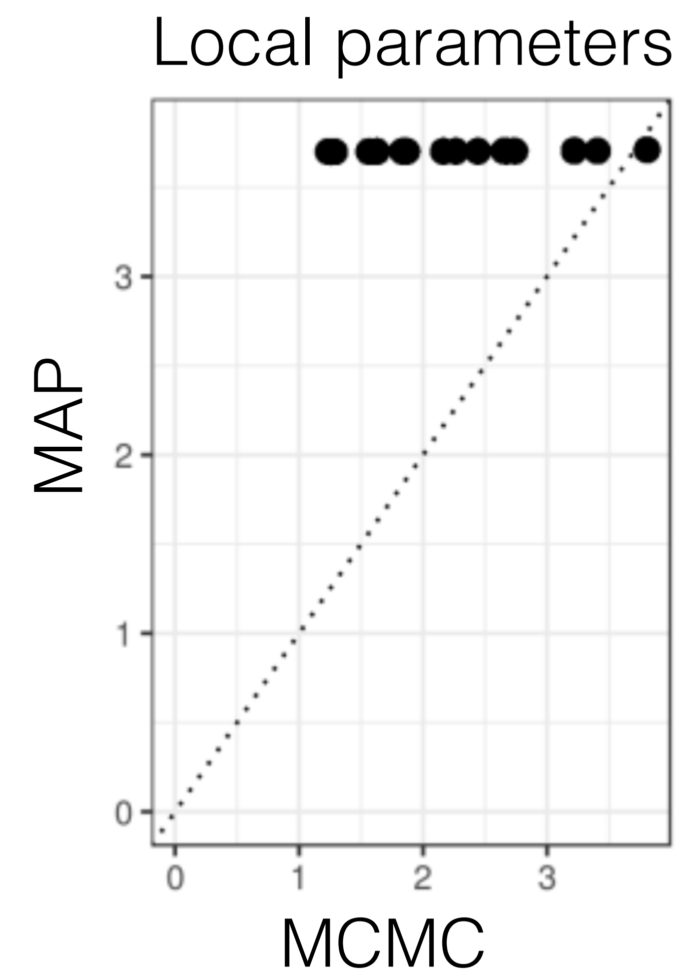
- MAP: **12 s**
- MFVB: **57 s**

[Giordano, Broderick, Jordan 2018]

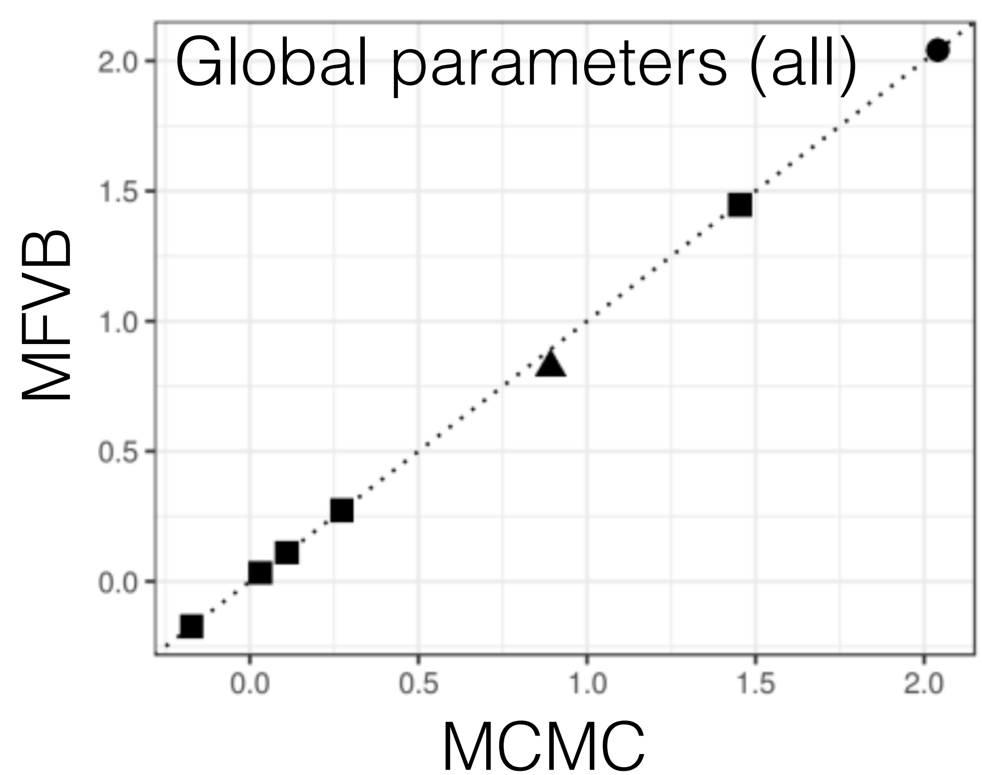
Criteo Online Ads Experiment

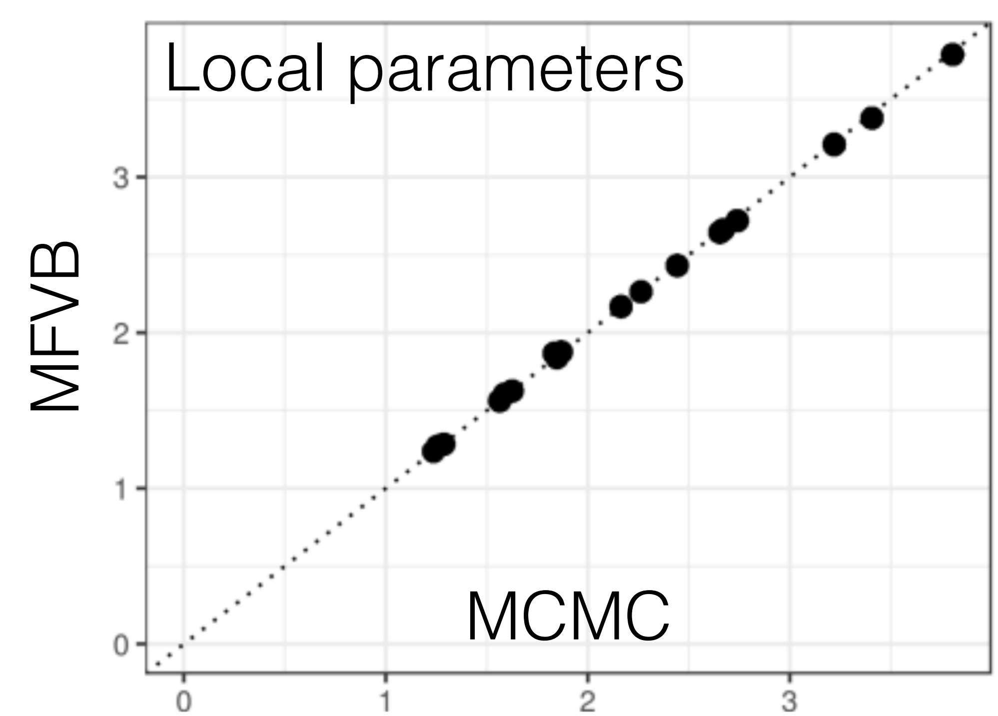






- MAP: **12 s**
- MFVB: **57 s**

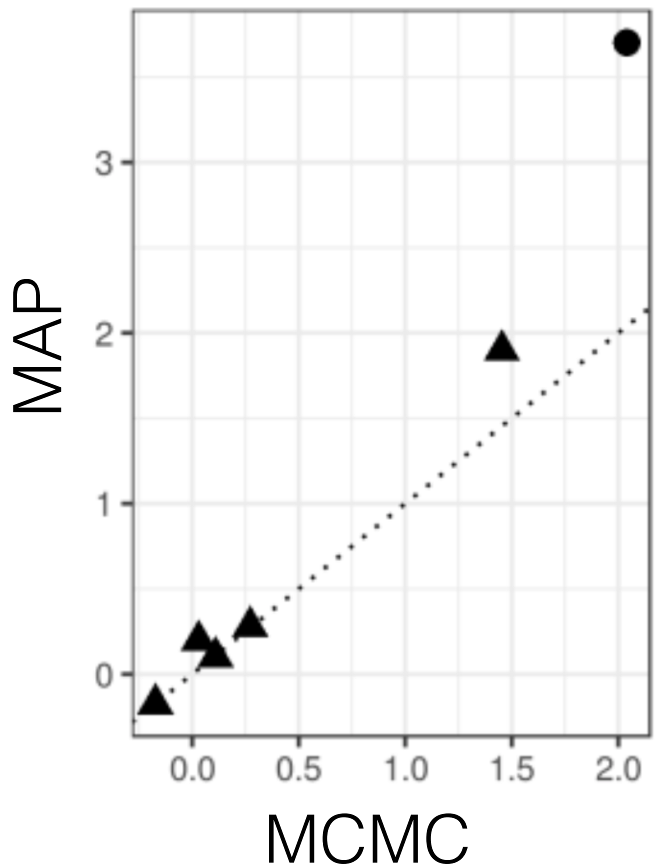




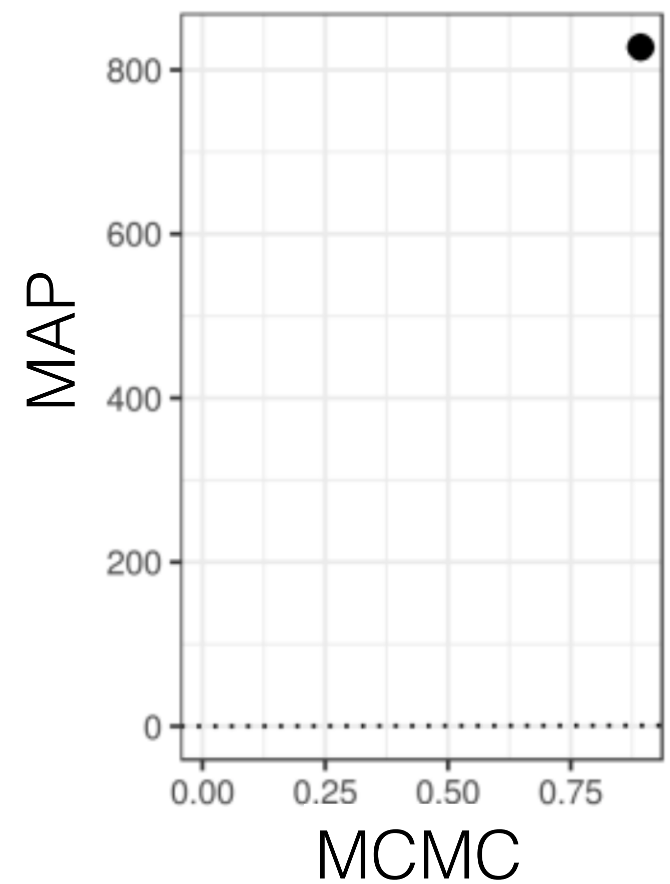
[Giordano, Broderick, Jordan 2018]

Criteo Online Ads Experiment

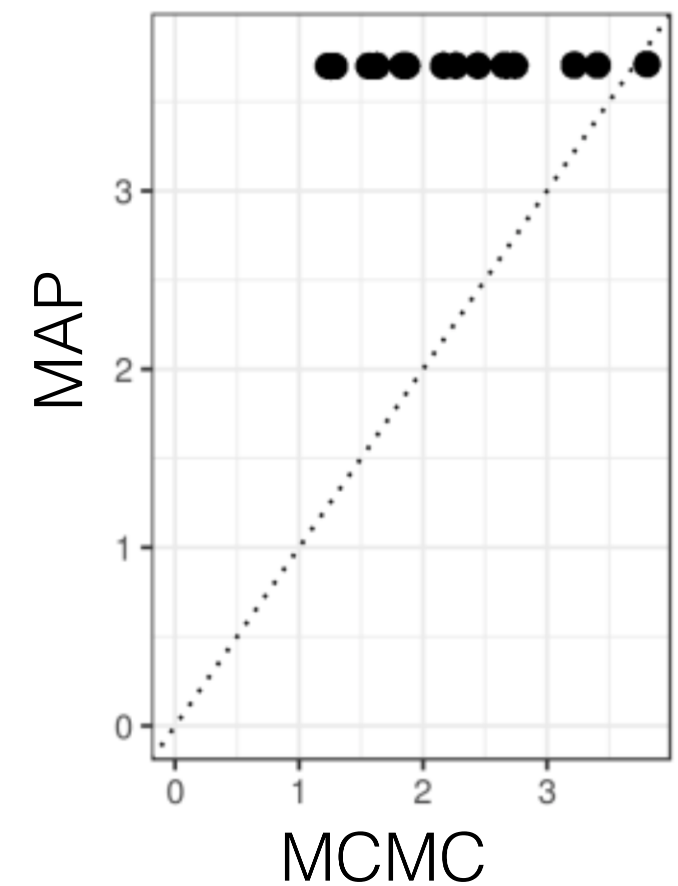
Global parameters (-τ)



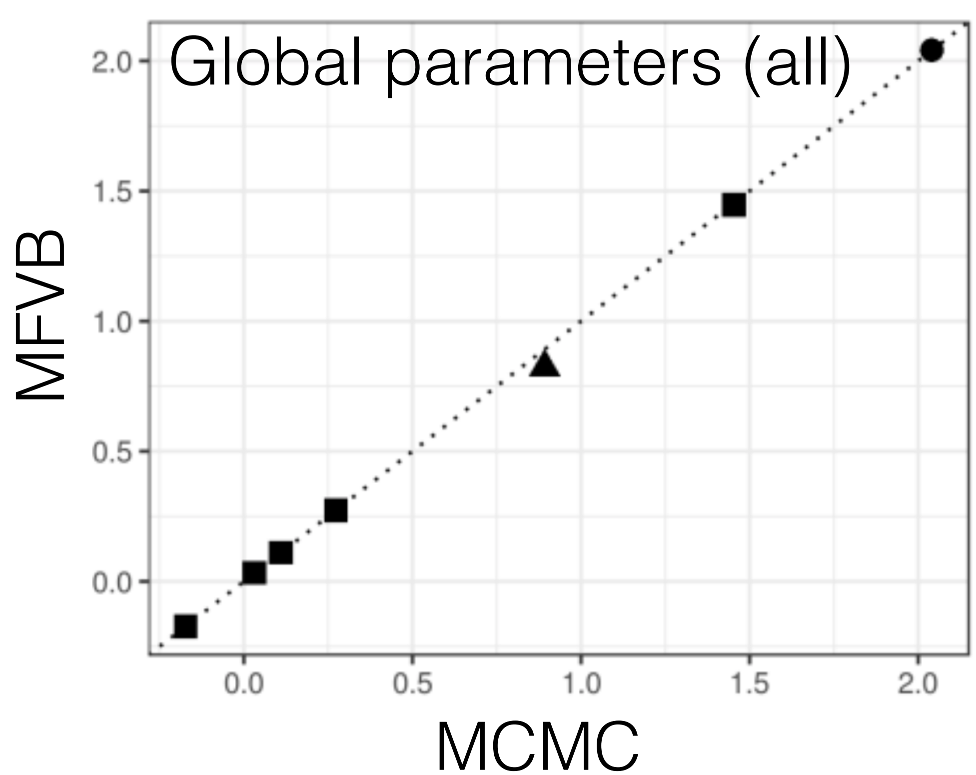
Global parameter τ

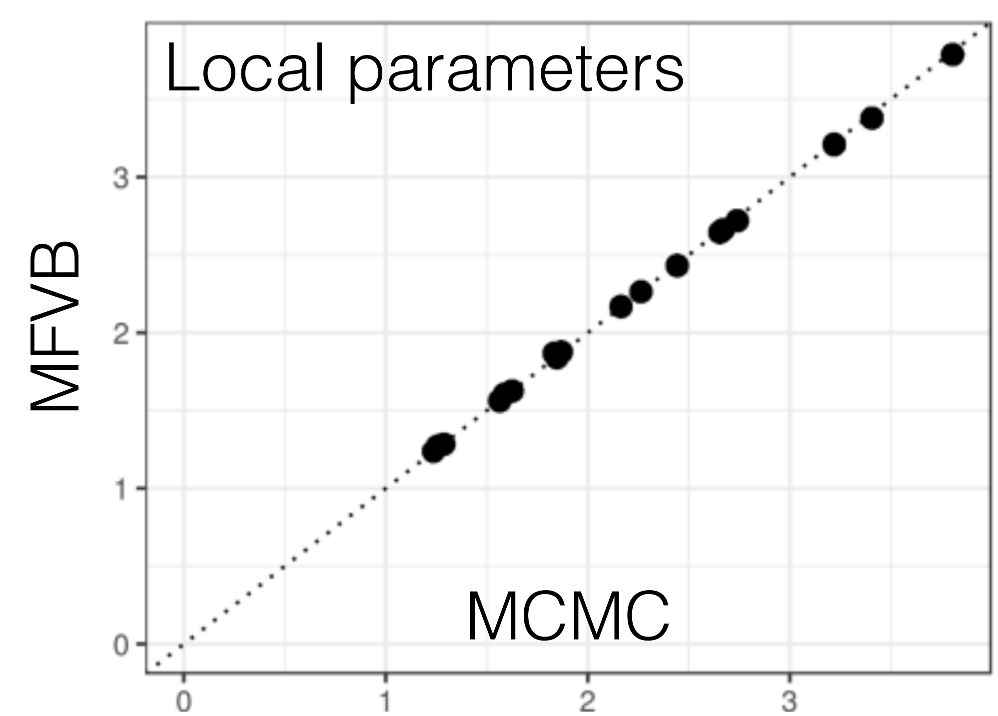


Local parameters



- MAP: **12 s**
- MFVB: **57 s**
- MCMC (5K samples): 21,066 s (**5.85 h**)





[Giordano, Broderick, Jordan 2018]

Why use MFVB?

- Topic discovery

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

[Blei et al 2003]

Why use MFVB?

- Topic discovery
  - Latent Dirichlet allocation (LDA)

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

[Blei et al 2003]

Why use MFVB?

- Topic discovery
  - Latent Dirichlet allocation (LDA): 27,700+ citations

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

[Blei et al 2003]

Roadmap

- Bayes & Approximate Bayes review
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use MFVB?
- When can we trust MFVB?
- Where do we go from here?

Roadmap

- Bayes & Approximate Bayes review
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use MFVB?
- When can we trust MFVB?
- Where do we go from here?

What about uncertainty?

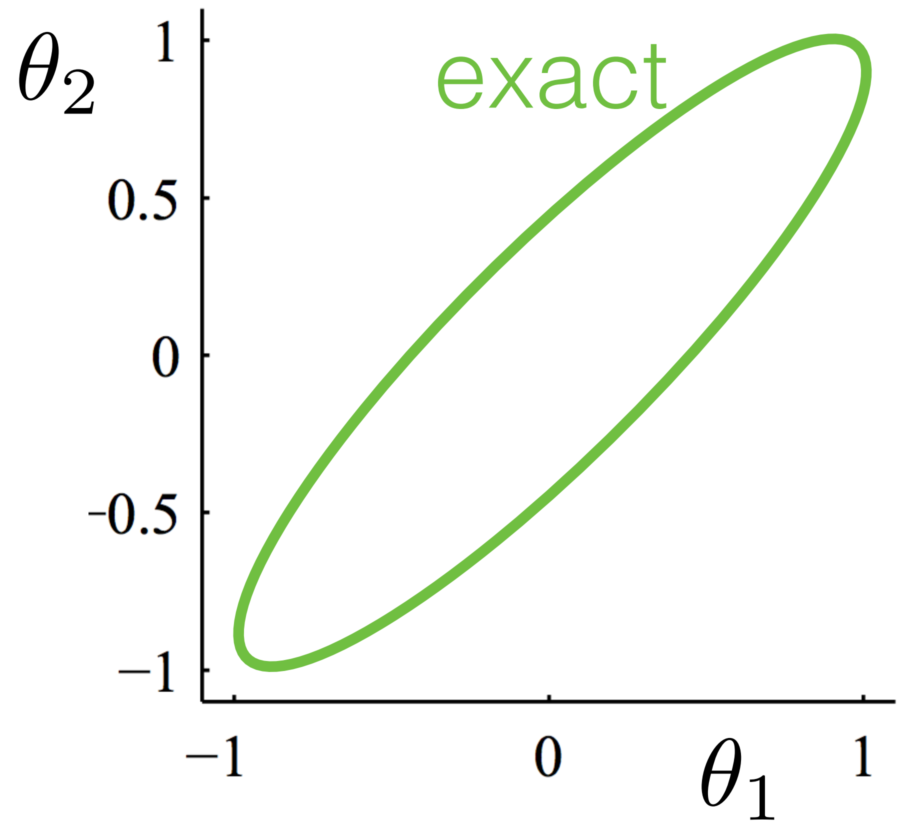
$$KL(q||p(\cdot|y)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

$$q(\theta) = \prod_{j=1}^{J} q_j(\theta_j)$$

What about uncertainty?

$$KL(q||p(\cdot|y)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

$$q(\theta) = \prod_{j=1}^{J} q_j(\theta_j)$$

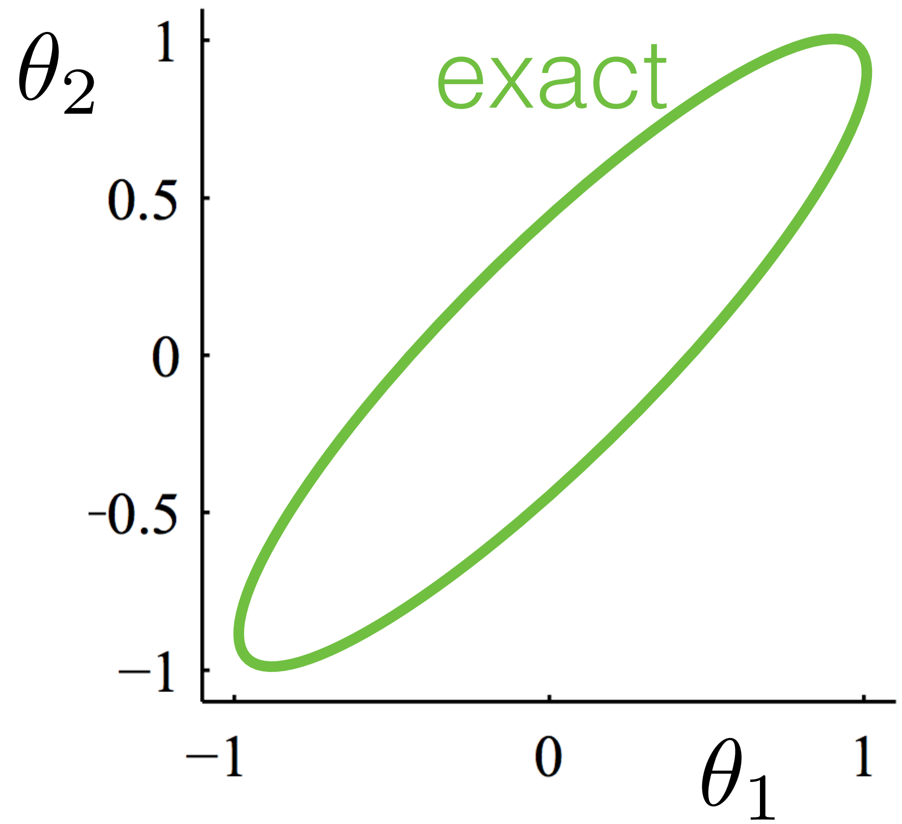


[Turner & Sahani 2011; MacKay 2003; Bishop 2006; Wang, Titterington 2004]

What about uncertainty?

$$KL(q||p(\cdot|y)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

$$q(\theta) = \prod_{j=1}^{J} q_j(\theta_j)$$



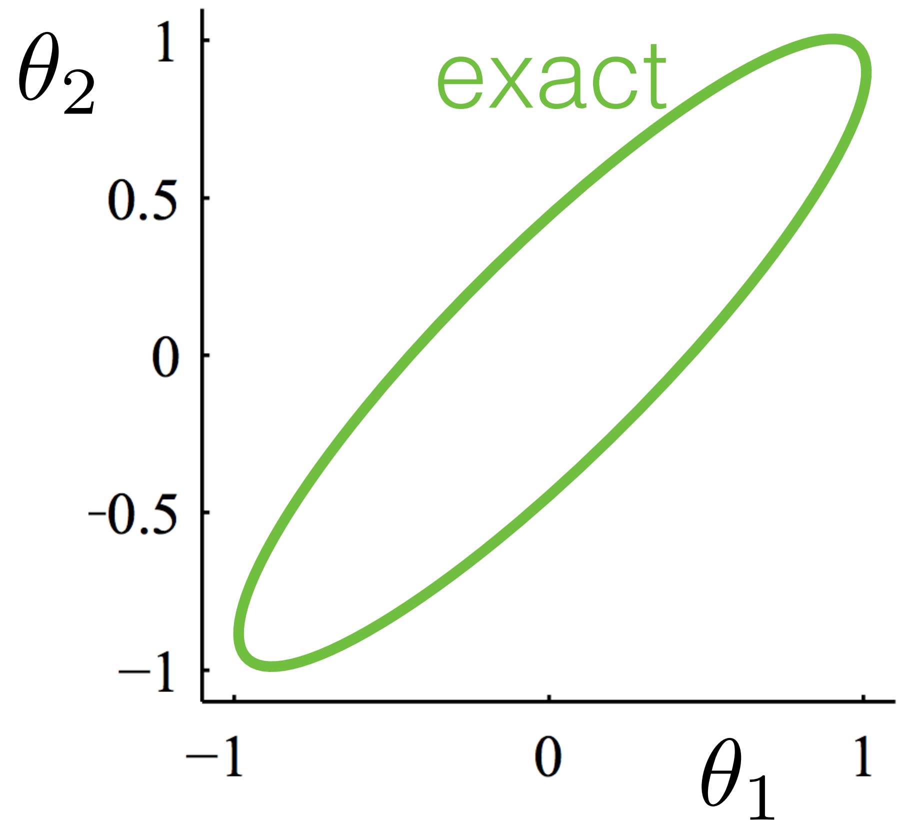
[Turner & Sahani 2011; MacKay 2003; Bishop 2006; Wang, Titterington 2004]

- Conjugate linear regression

What about uncertainty?

$$KL(q||p(\cdot|y)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

$$q(\theta) = \prod_{j=1}^{J} q_j(\theta_j)$$



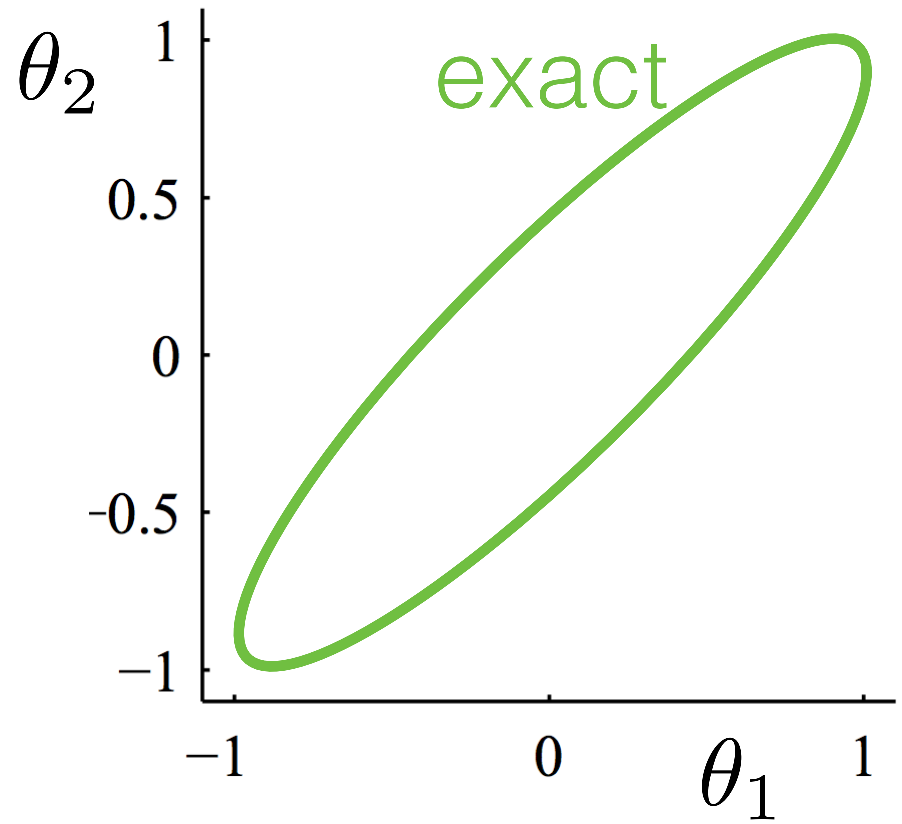
[Turner & Sahani 2011; MacKay 2003; Bishop 2006; Wang, Titterington 2004]

- Conjugate linear regression
- Bayesian central limit theorem

What about uncertainty?

$$KL(q||p(\cdot|y)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

$$q(\theta) = \prod_{j=1}^{J} q_j(\theta_j)$$



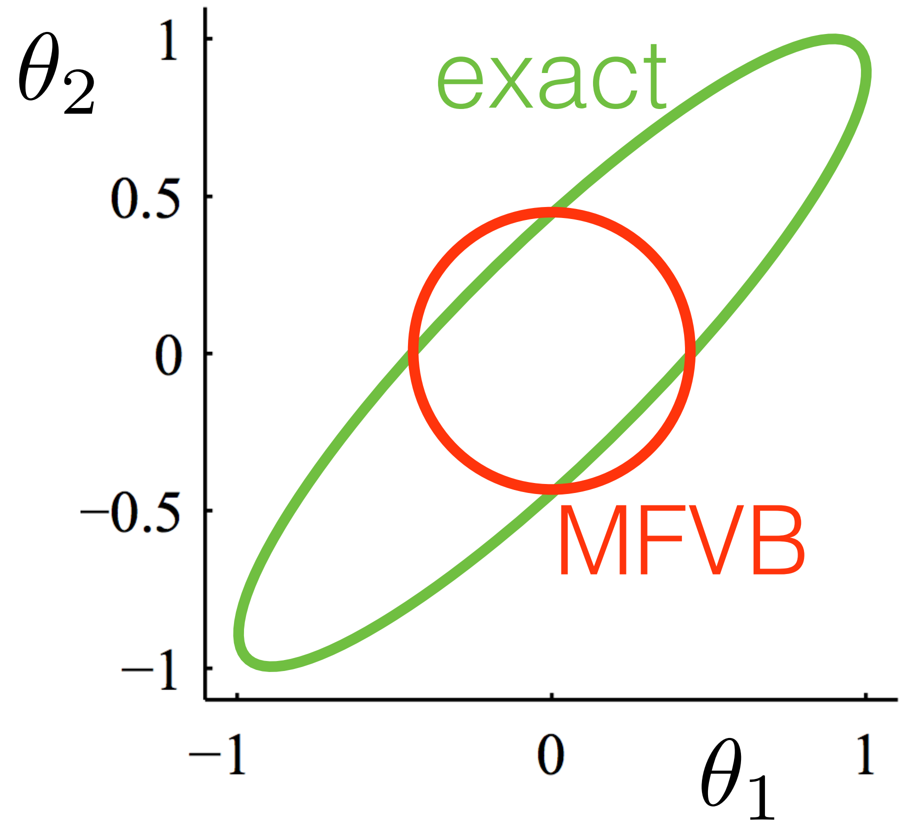
[Turner & Sahani 2011; MacKay 2003; Bishop 2006; Wang, Titterington 2004]

- Conjugate linear regression
- Bayesian central limit theorem

[Exercise: derive the MFVB-CA steps. Hint: use precision matrix.]

What about uncertainty?

$$KL(q||p(\cdot|y)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta \qquad q(\theta) = \prod_{j=1}^{J} q_j(\theta_j)$$



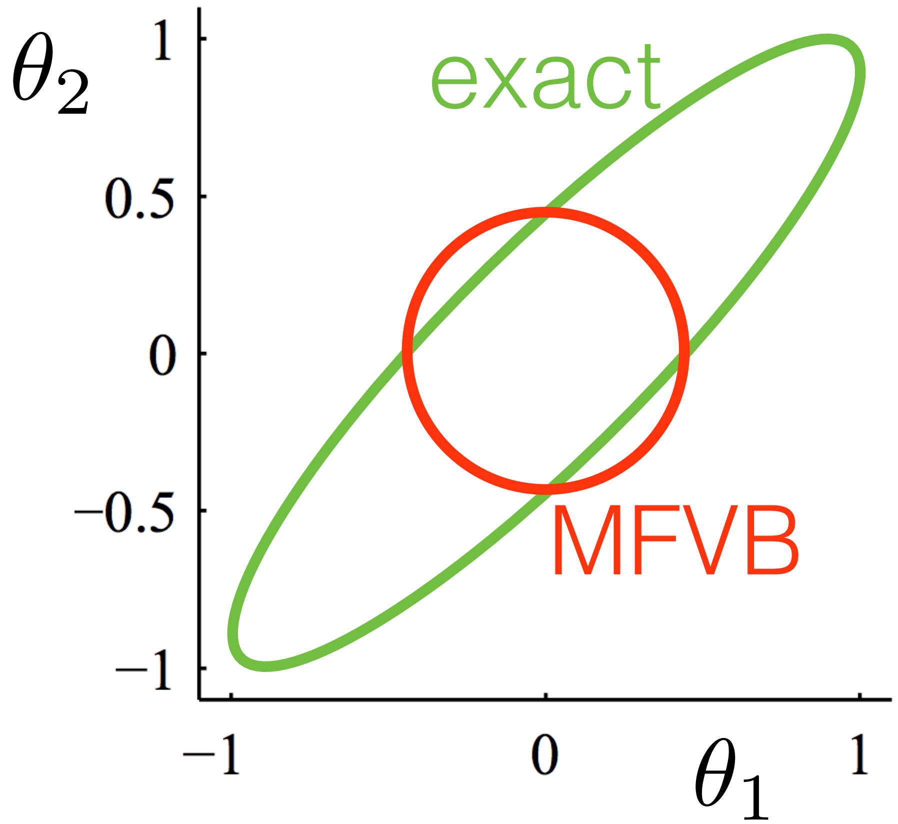
[Turner & Sahani 2011; MacKay 2003; Bishop 2006; Wang, Titterington 2004]

- Conjugate linear regression
- Bayesian central limit theorem

[Exercise: derive the MFVB-CA steps. Hint: use precision matrix.]

What about uncertainty?

$$KL(q||p(\cdot|y)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta \qquad q(\theta) = \prod_{j=1}^{J} q_j(\theta_j)$$



[Turner & Sahani 2011; MacKay 2003; Bishop 2006; Wang, Titterington 2004]

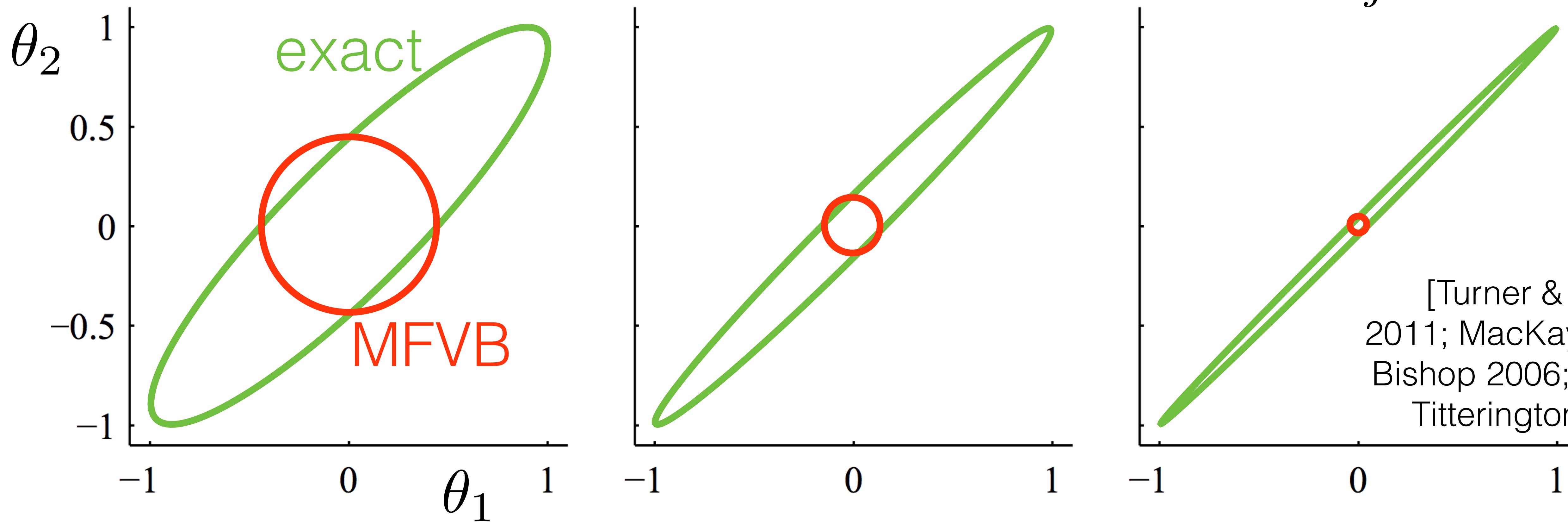
- Underestimates variance (sometimes severely)

- Conjugate linear regression
- Bayesian central limit theorem

[Exercise: derive the MFVB-CA steps. Hint: use precision matrix.]

What about uncertainty?

$$KL(q||p(\cdot|y)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta \qquad q(\theta) = \prod_{j=1}^{J} q_j(\theta_j)$$



[Turner & Sahani 2011; MacKay 2003; Bishop 2006; Wang, Titterington 2004]

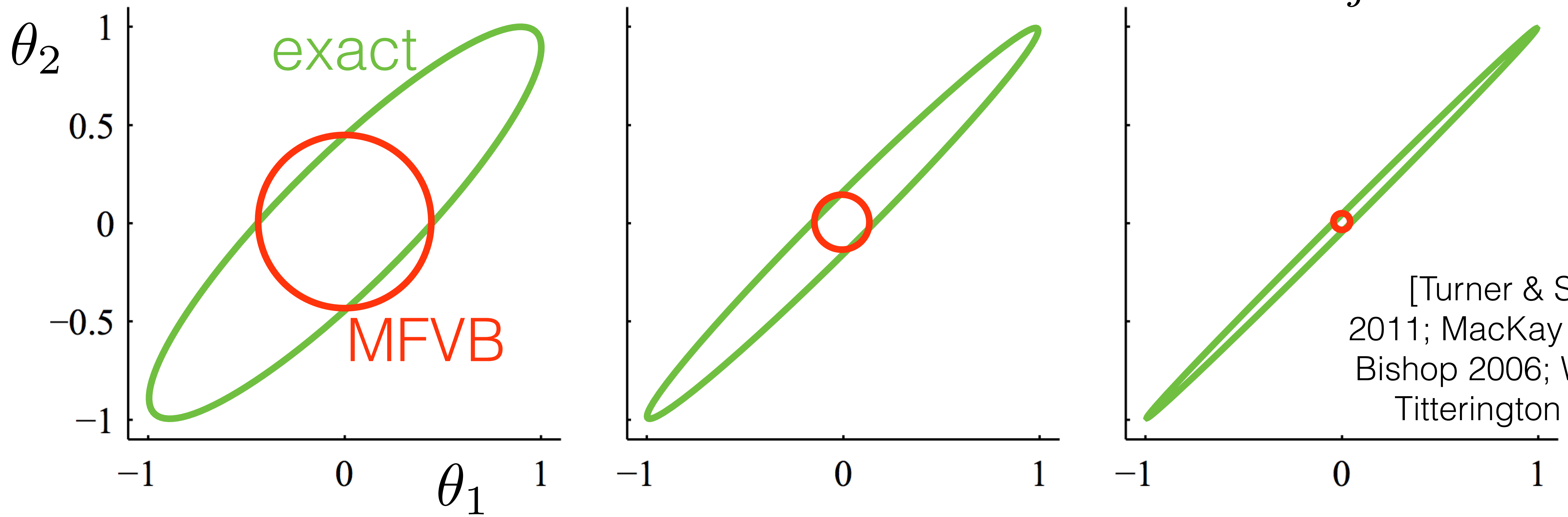
- Underestimates variance (sometimes severely)

- Conjugate linear regression
- Bayesian central limit theorem

[Exercise: derive the MFVB-CA steps. Hint: use precision matrix.]

What about uncertainty?

$$KL(q||p(\cdot|y)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta \qquad q(\theta) = \prod_{j=1}^{J} q_j(\theta_j)$$



[Turner & Sahani 2011; MacKay 2003; Bishop 2006; Wang, Titterington 2004]

- Underestimates variance (sometimes severely)
- No covariance estimates

- Conjugate linear regression
- Bayesian central limit theorem

[Exercise: derive the MFVB-CA steps. Hint: use precision matrix.]

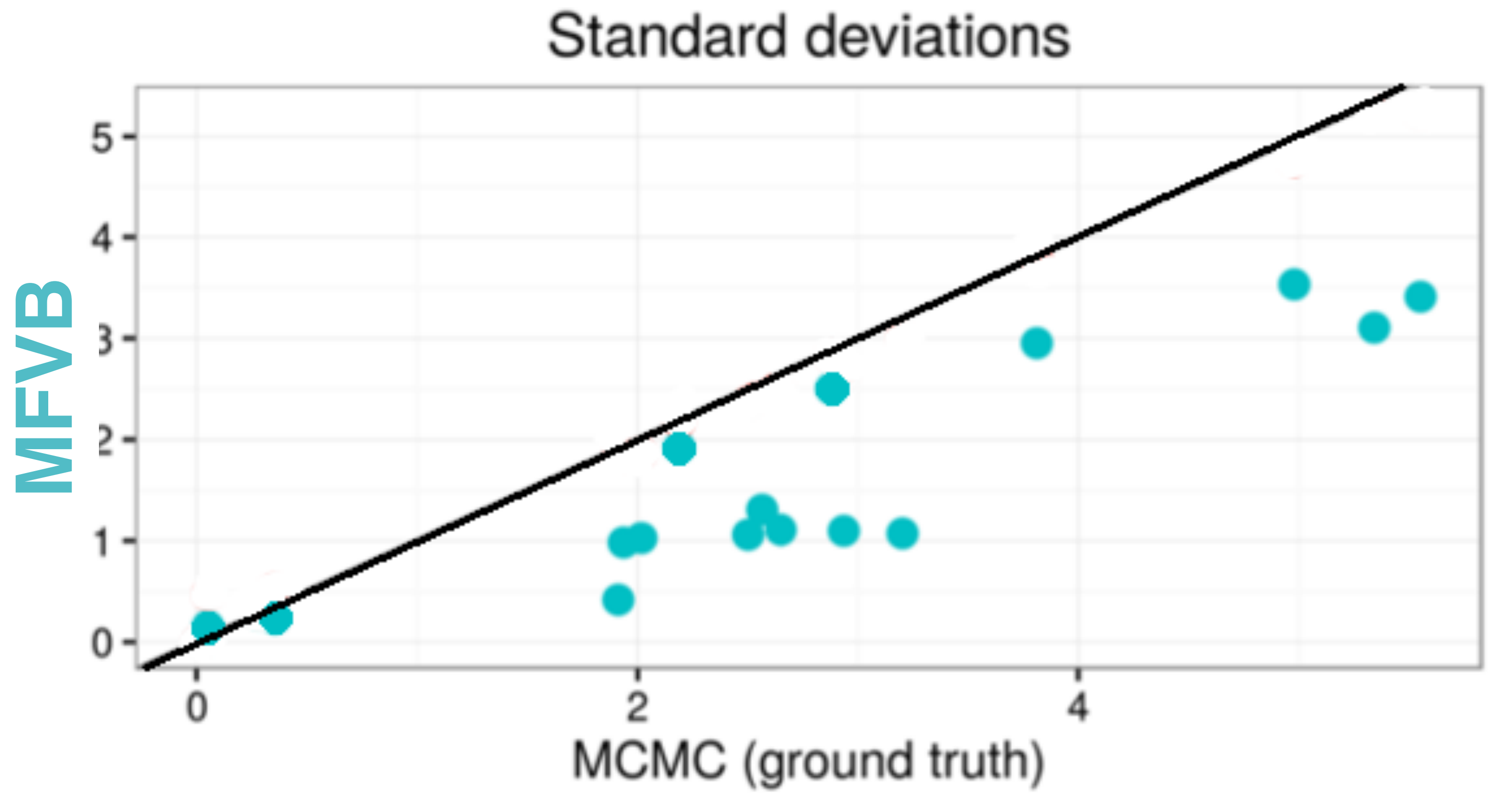
What about uncertainty?

- Microcredit

[Giordano, Broderick, Meager, Huggins, Jordan 2016]

What about uncertainty?

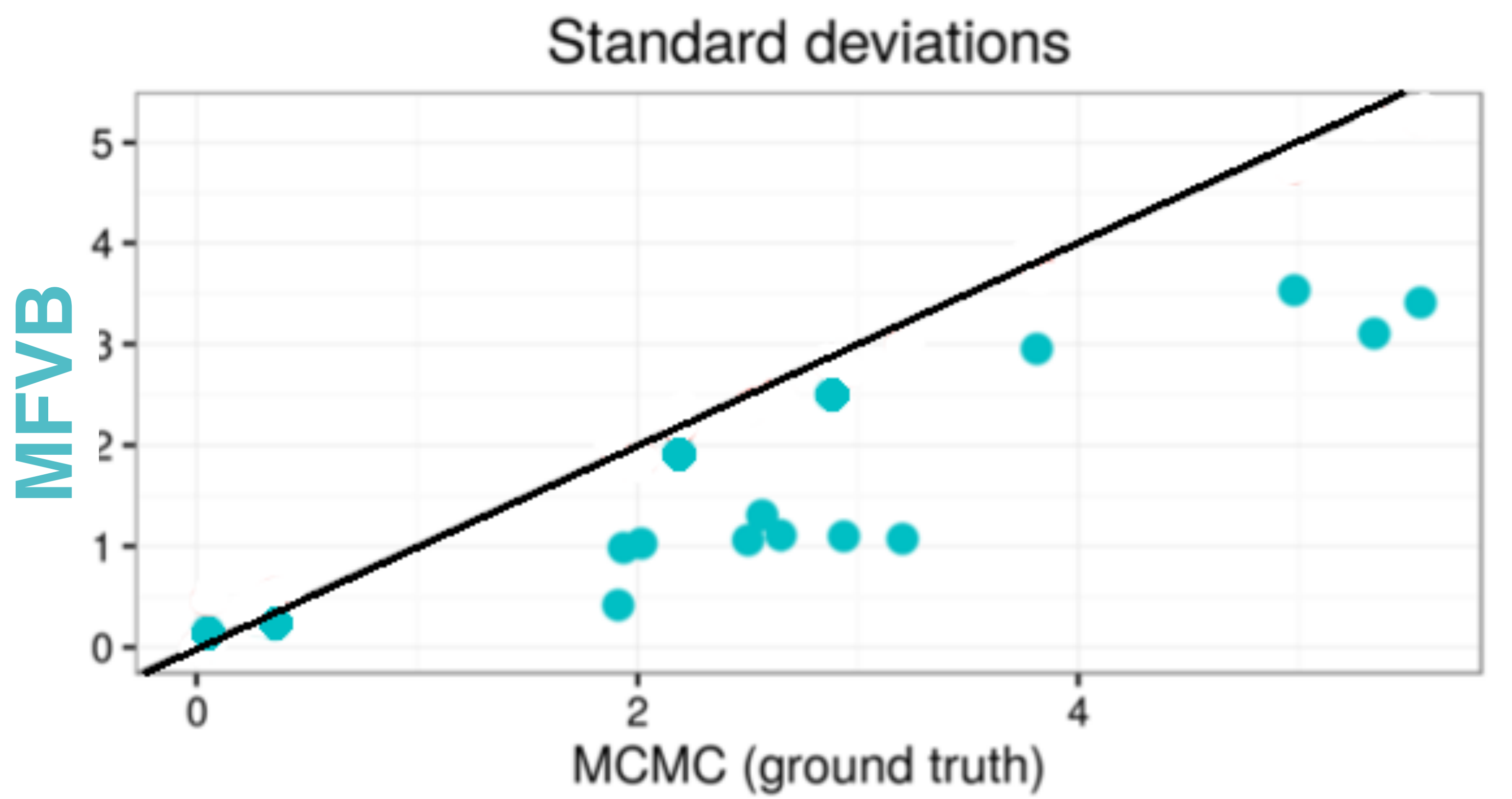
- Microcredit



[Giordano, Broderick, Meager, Huggins, Jordan 2016]

What about uncertainty?

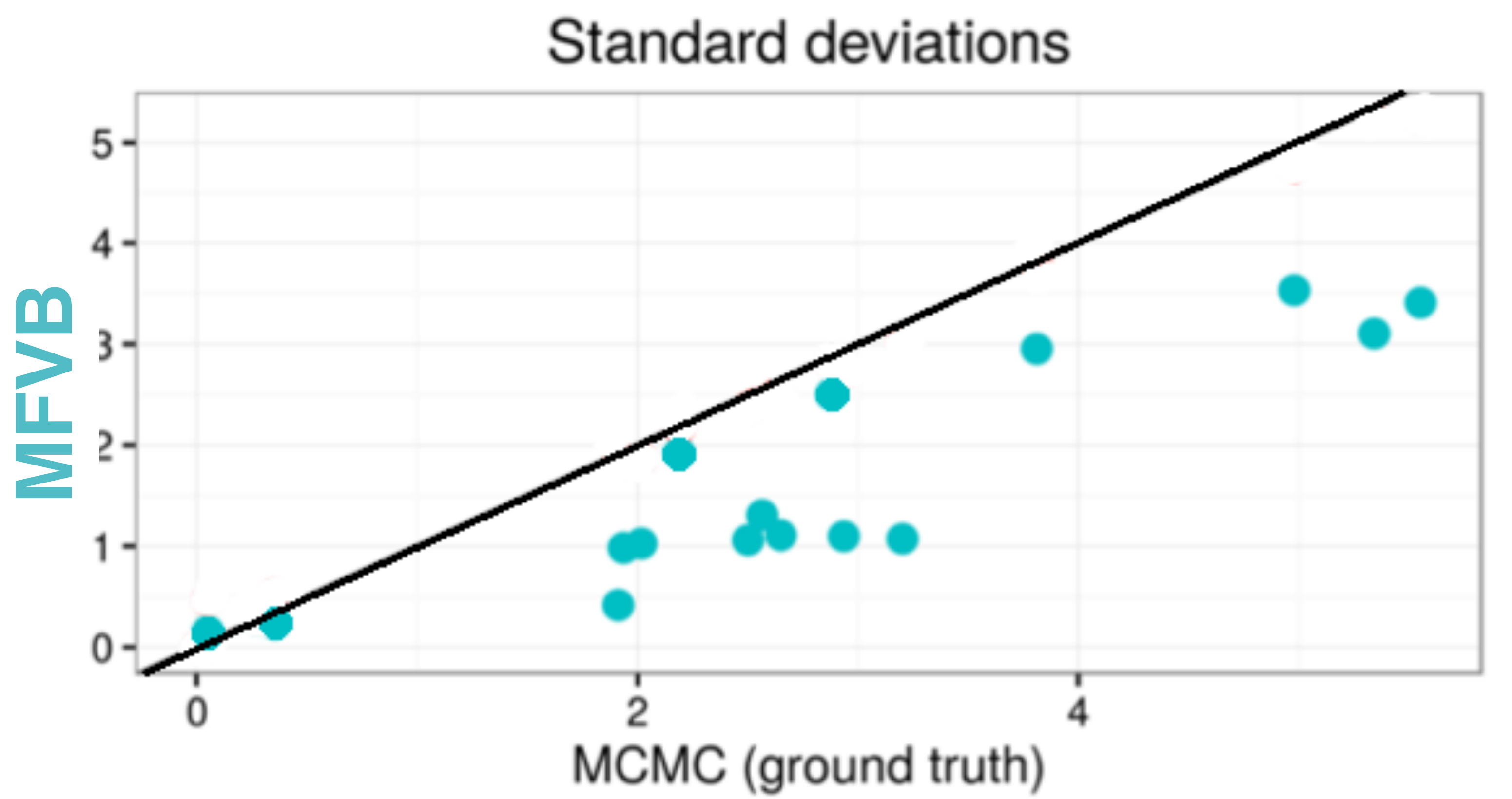
- Microcredit effect
- τ mean: 3.08 USD PPP



[Giordano, Broderick, Meager, Huggins, Jordan 2016]

What about uncertainty?

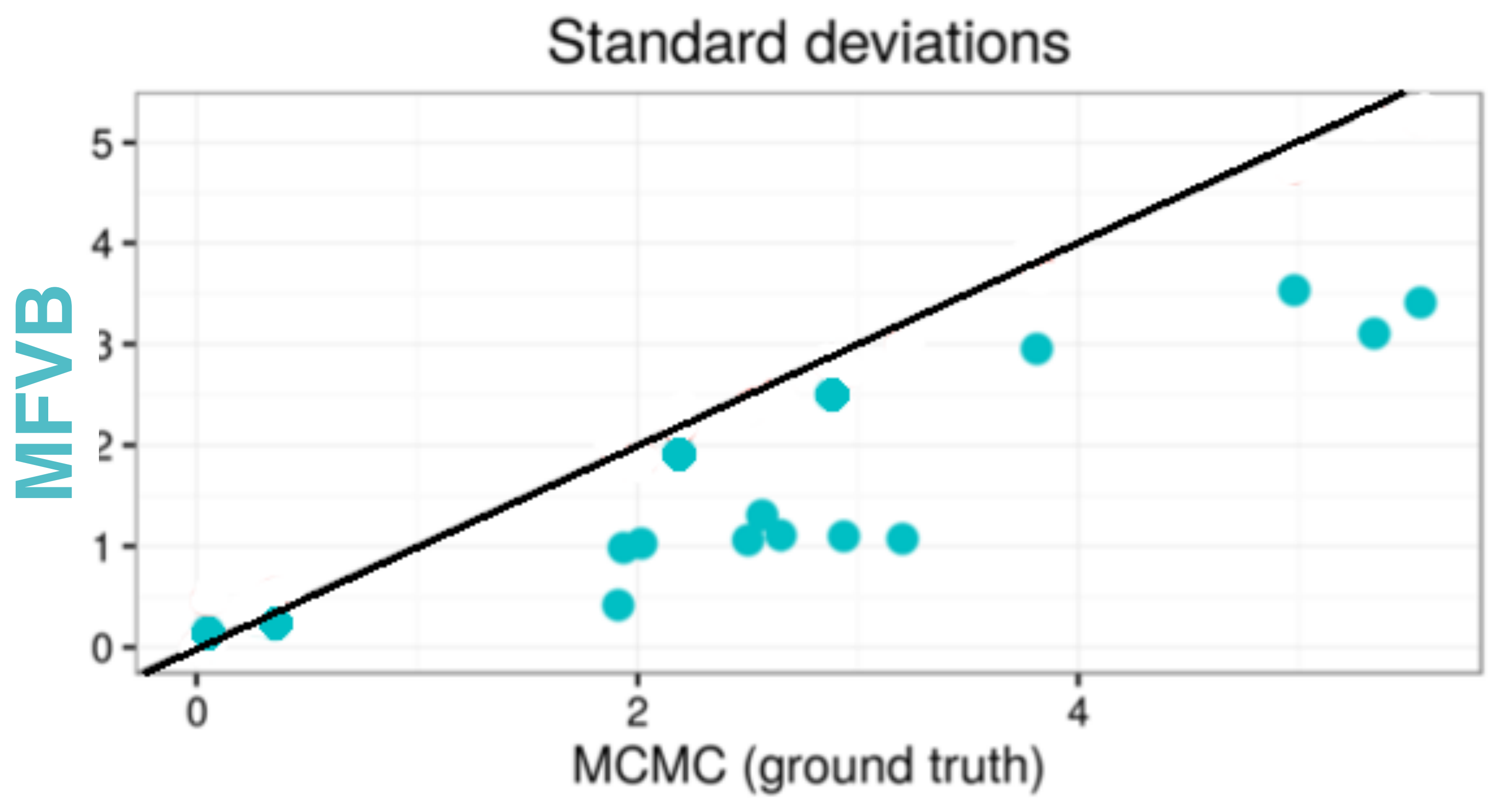
- Microcredit effect
- τ mean: 3.08 USD PPP
- τ std dev: 1.83 USD PPP



[Giordano, Broderick, Meager, Huggins, Jordan 2016]

What about uncertainty?

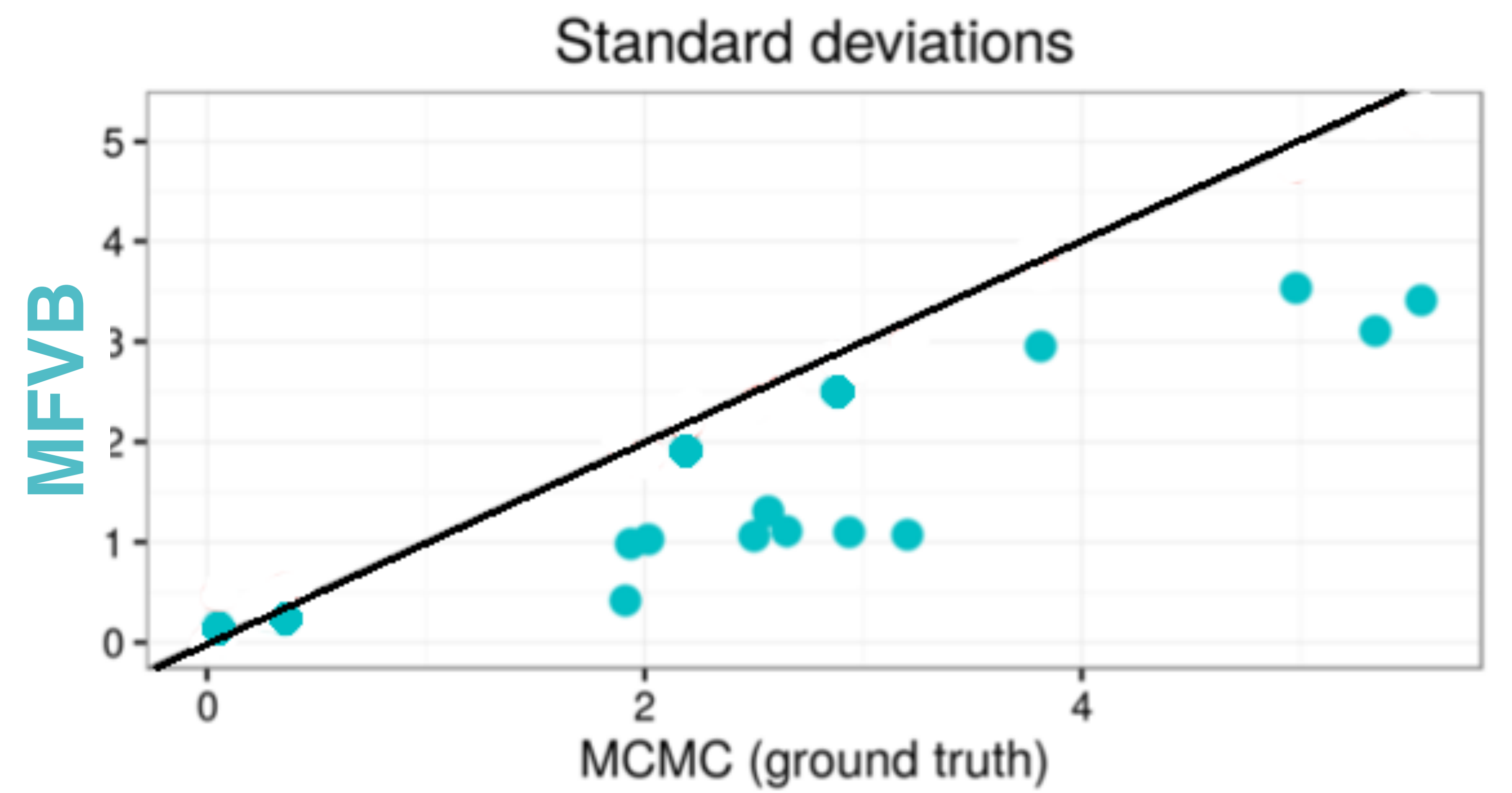
- Microcredit effect
- τ mean: 3.08 USD PPP
- τ std dev: 1.83 USD PPP
- Mean is 1.68 std dev from 0



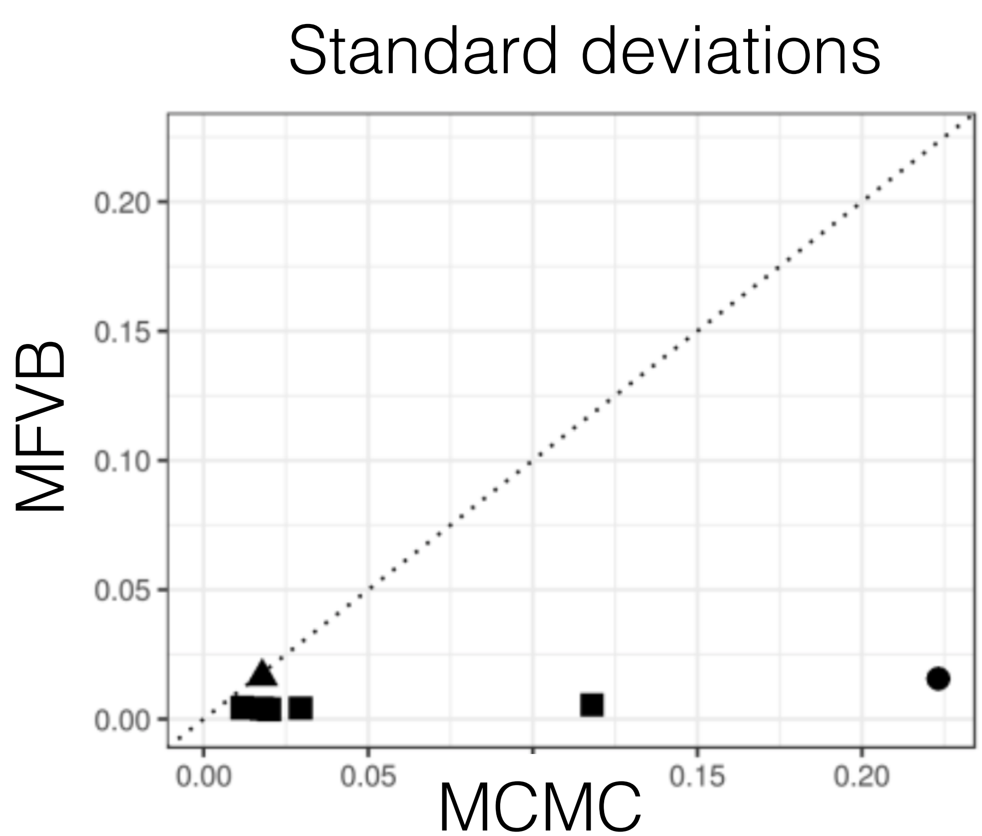
[Giordano, Broderick, Meager, Huggins, Jordan 2016]

What about uncertainty?

- Microcredit effect
- τ mean: 3.08 USD PPP
- τ std dev: 1.83 USD PPP
- Mean is 1.68 std dev from 0



- Criteo online ads experiment



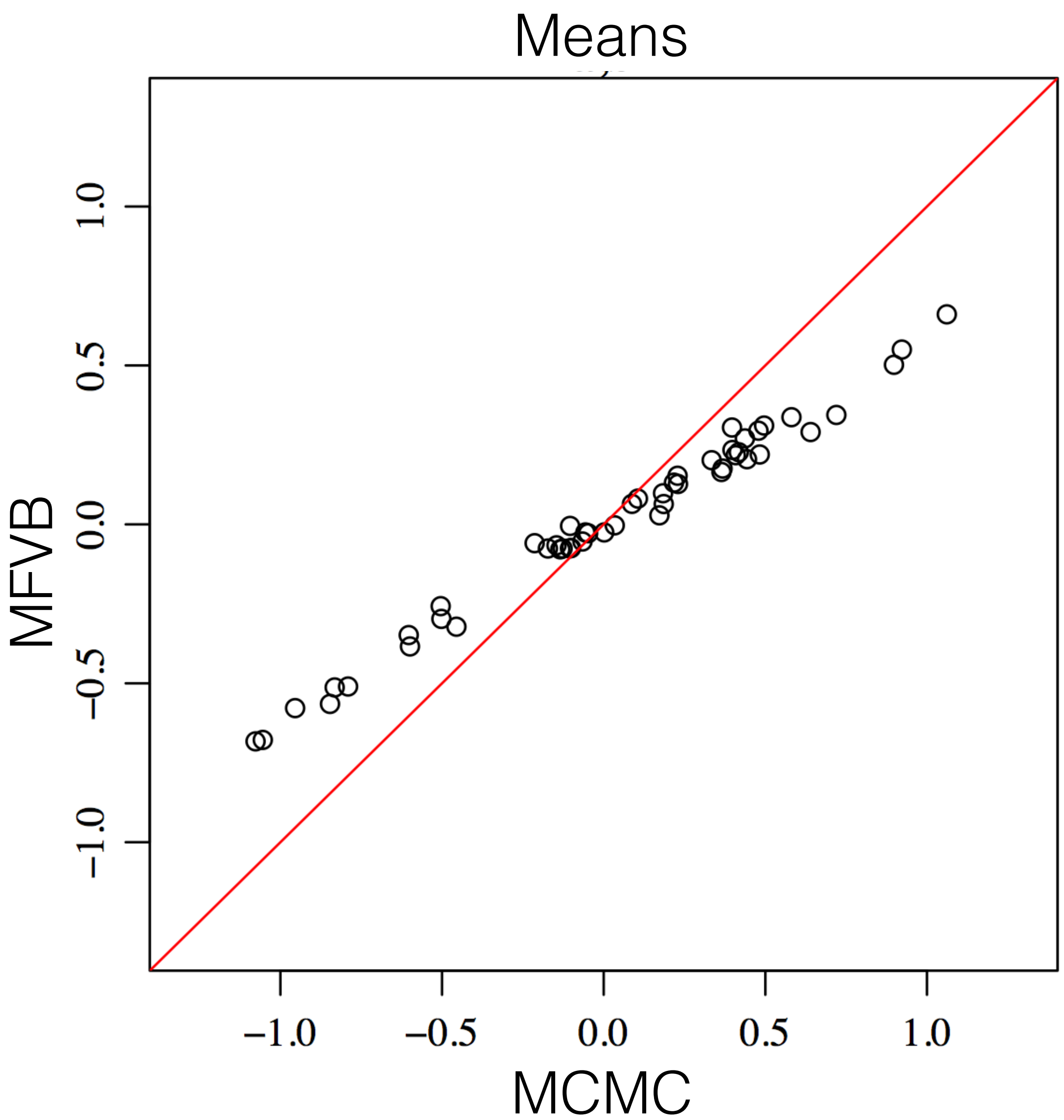
[Giordano, Broderick, Meager, Huggins, Jordan 2016; Giordano, Broderick, Jordan 2018]

What about means?

- Model for relational data with covariates
- When 1000+ nodes, MCMC > 1 day [Fosdick 2013, Ch 4]

What about means?

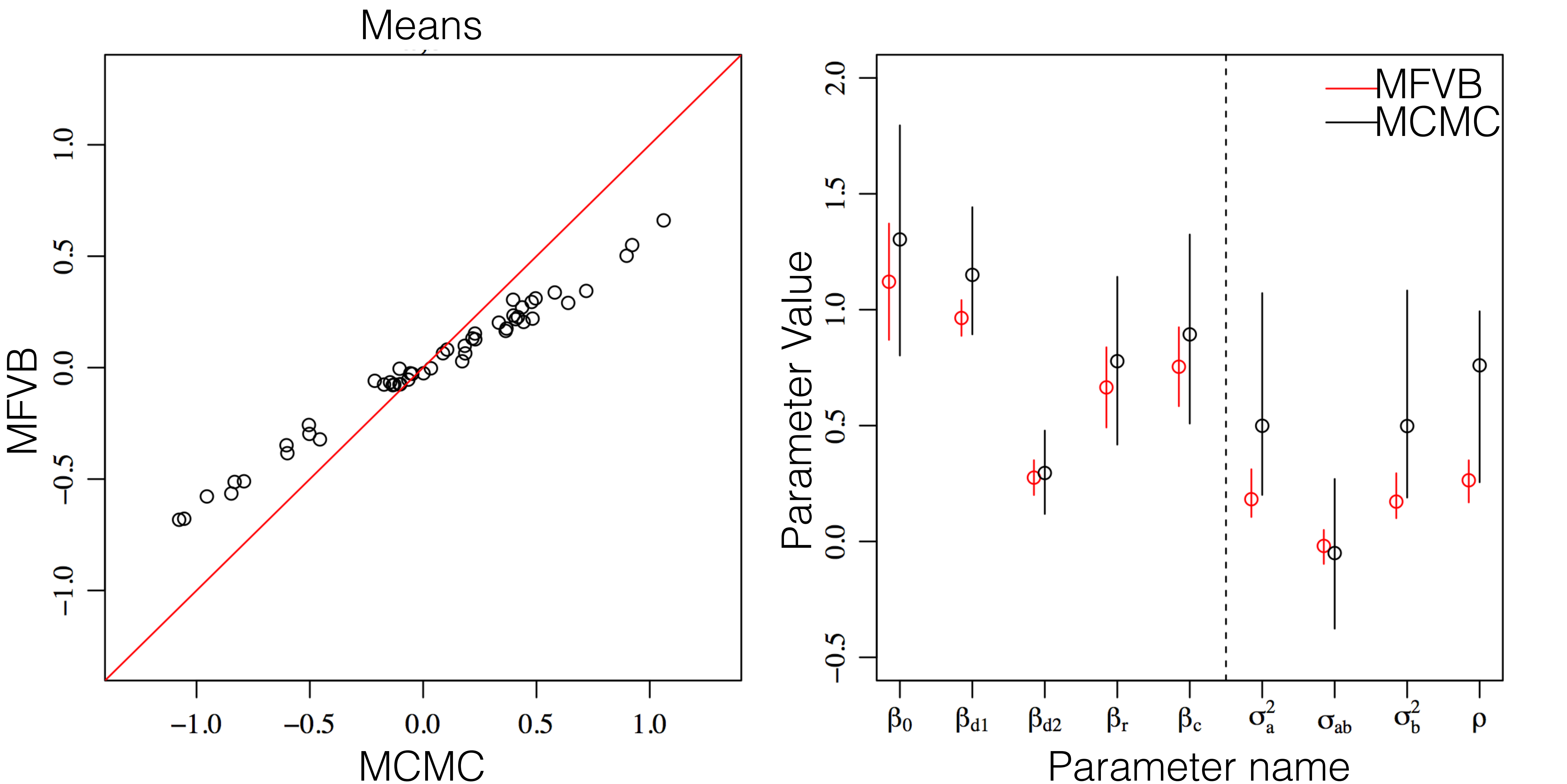
- Model for relational data with covariates
- When 1000+ nodes, MCMC > 1 day [Fosdick 2013, Ch 4]



[Fosdick 2013, Ch 4, Fig 4.3]

What about means?

- Model for relational data with covariates
- When 1000+ nodes, MCMC > 1 day [Fosdick 2013, Ch 4]



[Fosdick 2013, Ch 4, Fig 4.3]

Posterior means: revisited

- Want to predict college GPA y_n

[Giordano, Broderick, Jordan 2015]

Posterior means: revisited

- Want to predict college GPA y_n
- Collect: standardized test scores (e.g., SAT, ACT) x_n

[Giordano, Broderick, Jordan 2015]

Posterior means: revisited

- Want to predict college GPA y_n
- Collect: standardized test scores (e.g., SAT, ACT) x_n
- Collect: regional test scores r_n

[Giordano, Broderick, Jordan 2015]

Posterior means: revisited

- Want to predict college GPA y_n
- Collect: standardized test scores (e.g., SAT, ACT) x_n
- Collect: regional test scores r_n
- Model: $y_n | \beta, z, \sigma^2 \overset{indep}{\sim} \mathcal{N}(\beta^T x_n + z_{k(n)} r_n, \sigma^2)$

[Giordano, Broderick, Jordan 2015]

Posterior means: revisited

- Want to predict college GPA y_n
- Collect: standardized test scores (e.g., SAT, ACT) x_n
- Collect: regional test scores r_n
- Model: $y_n|\beta, z, \sigma^2 \overset{indep}{\sim} \mathcal{N}(\beta^T x_n + z_{k(n)} r_n, \sigma^2)$

$$z_k|\rho^2 \overset{iid}{\sim} \mathcal{N}(0, \rho^2) \qquad (\sigma^2)^{-1} \sim \mathrm{Gamma}(a_{\sigma^2}, b_{\sigma^2})$$

$$\beta \sim \mathcal{N}(0, \Sigma) \qquad (\rho^2)^{-1} \sim \mathrm{Gamma}(a_{\rho^2}, b_{\rho^2})$$

[Giordano, Broderick, Jordan 2015]

Posterior means: revisited

- Want to predict college GPA y_n
- Collect: standardized test scores (e.g., SAT, ACT) x_n
- Collect: regional test scores r_n
- Model:

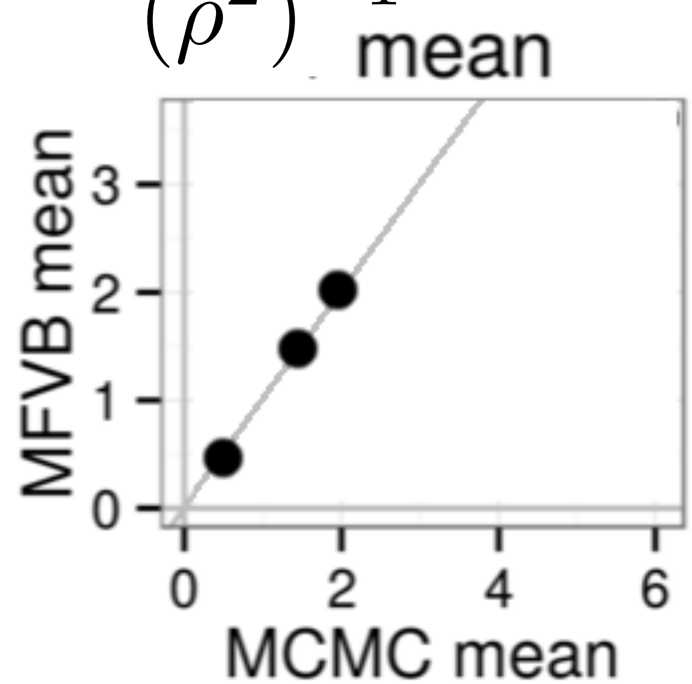
$$y_n|\beta, z, \sigma^2 \overset{indep}{\sim} \mathcal{N}(\beta^T x_n + z_{k(n)} r_n, \sigma^2)$$

$$z_k|\rho^2 \overset{iid}{\sim} \mathcal{N}(0, \rho^2) \qquad (\sigma^2)^{-1} \sim \text{Gamma}(a_{\sigma^2}, b_{\sigma^2})$$

$$\beta \sim \mathcal{N}(0, \Sigma) \qquad (\rho^2)^{-1} \sim \text{Gamma}(a_{\rho^2}, b_{\rho^2})$$

- Data simulated from model (3 data sets, 300 data points):

$(\rho^2)^{-1}$



[Giordano, Broderick, Jordan 2015]

Posterior means: revisited

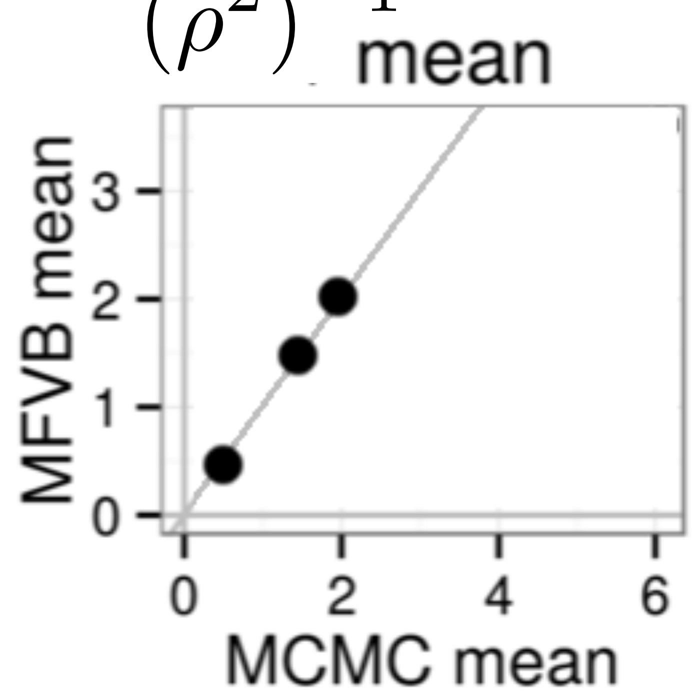
- Want to predict college GPA y_n
- Collect: standardized test scores (e.g., SAT, ACT) x_n
- Collect: regional test scores r_n
- Model: $y_n | \beta, z, \sigma^2 \overset{indep}{\sim} \mathcal{N}(\beta^T x_n + z_{k(n)} r_n, \sigma^2)$

$$z_k | \rho^2 \overset{iid}{\sim} \mathcal{N}(0, \rho^2) \qquad (\sigma^2)^{-1} \sim \text{Gamma}(a_{\sigma^2}, b_{\sigma^2})$$

$$\beta \sim \mathcal{N}(0, \Sigma) \qquad (\rho^2)^{-1} \sim \text{Gamma}(a_{\rho^2}, b_{\rho^2})$$

- Data simulated from model (3 data sets, 300 data points):

$(\rho^2)^{-1}$



[Giordano, Broderick, Jordan 2015]

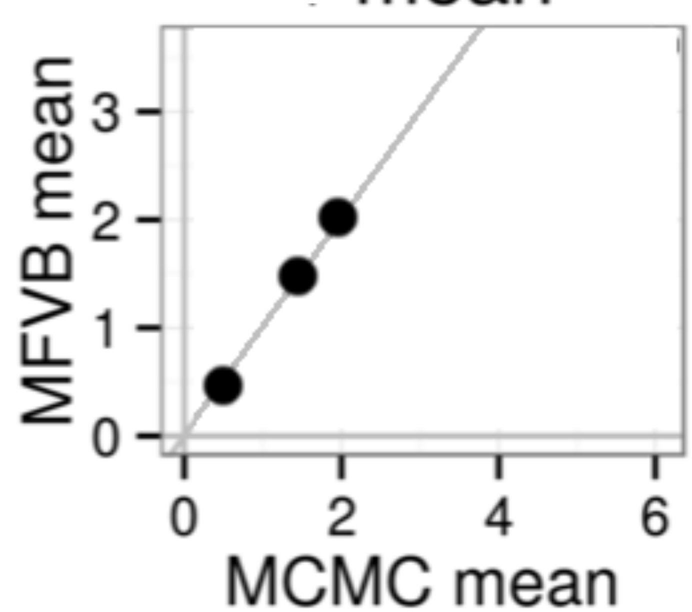
Posterior means: revisited

- Want to predict college GPA y_n
- Collect: standardized test scores (e.g., SAT, ACT) x_n
- Collect: regional test scores r_n
- Model: $y_n|\beta, z, \sigma^2 \overset{indep}{\sim} \mathcal{N}(\beta^T x_n + z_{k(n)} r_n, \sigma^2)$

$$z_k|\rho^2 \overset{iid}{\sim} \mathcal{N}(0, \rho^2) \qquad (\sigma^2)^{-1} \sim \text{Gamma}(a_{\sigma^2}, b_{\sigma^2})$$

$$\beta \sim \mathcal{N}(0, \Sigma) \qquad (\rho^2)^{-1} \sim \text{Gamma}(a_{\rho^2}, b_{\rho^2})$$

- Data simulated from model (100 data sets, 300 data points):



[Giordano, Broderick, Jordan 2015]

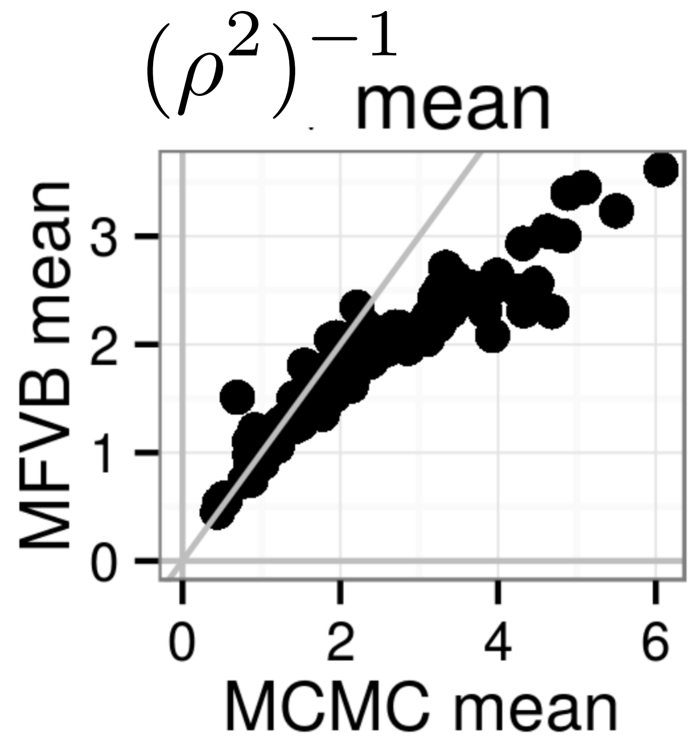
Posterior means: revisited

- Want to predict college GPA y_n
- Collect: standardized test scores (e.g., SAT, ACT) x_n
- Collect: regional test scores r_n
- Model: $y_n | \beta, z, \sigma^2 \overset{indep}{\sim} \mathcal{N}(\beta^T x_n + z_{k(n)} r_n, \sigma^2)$

$$z_k | \rho^2 \overset{iid}{\sim} \mathcal{N}(0, \rho^2) \qquad (\sigma^2)^{-1} \sim \text{Gamma}(a_{\sigma^2}, b_{\sigma^2})$$

$$\beta \sim \mathcal{N}(0, \Sigma) \qquad (\rho^2)^{-1} \sim \text{Gamma}(a_{\rho^2}, b_{\rho^2})$$

- Data simulated from model (100 data sets, 300 data points):



[Giordano, Broderick, Jordan 2015]

Approximate Bayesian inference

Use q^* to approximate $p(\cdot|y)$

Optimization

$$q^* = \text{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

Variational Bayes

$$q^* = \text{argmin}_{q \in Q} KL(q(\cdot)||p(\cdot|y))$$

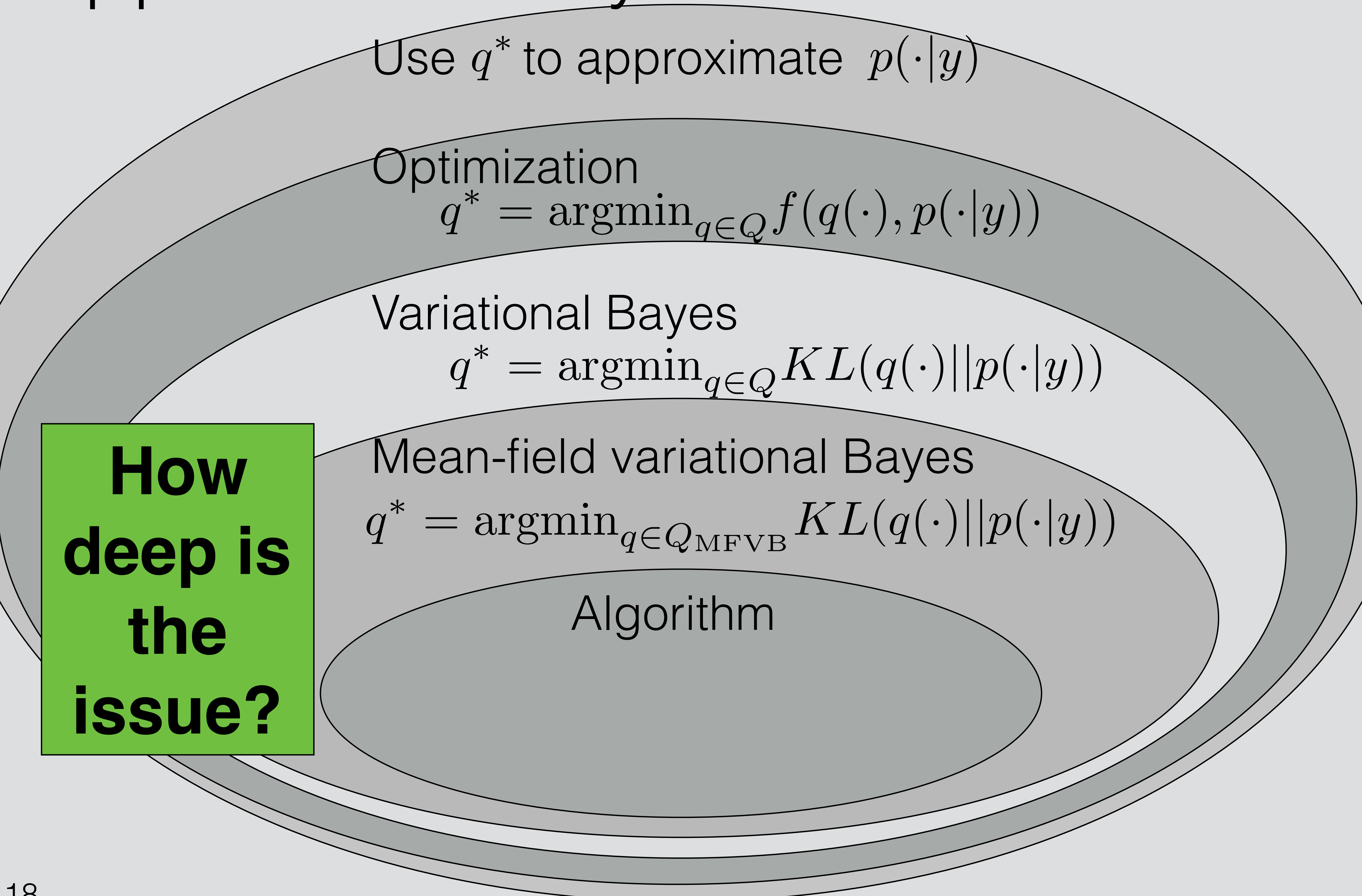
Mean-field variational Bayes

$$q^* = \text{argmin}_{q \in Q_{\text{MFVB}}} KL(q(\cdot)||p(\cdot|y))$$

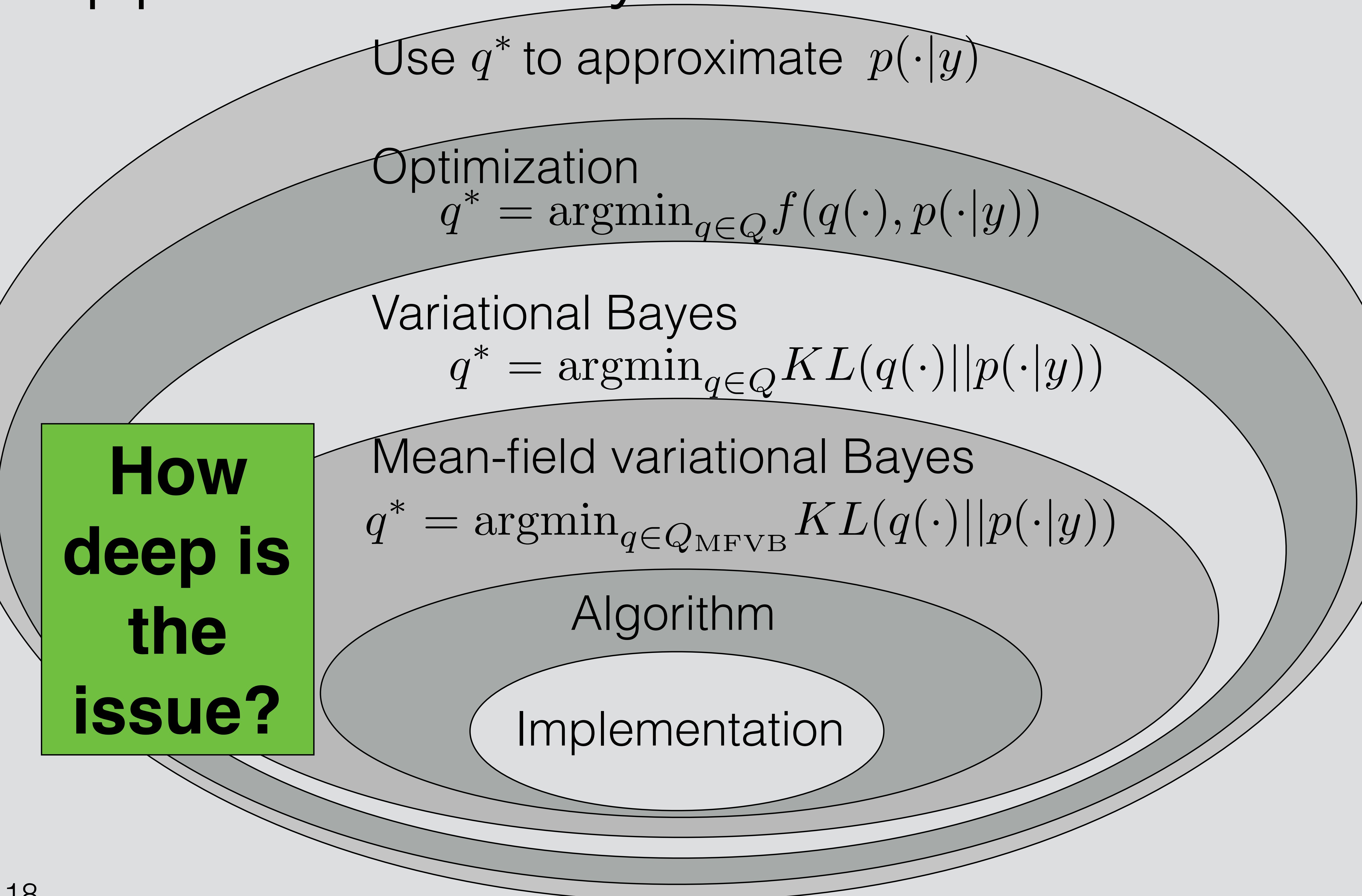
- Coordinate descent
- Stochastic variational inference (SVI) [Hoffman et al 2013]
- Automatic differentiation variational inference (ADVI) [Kucukelbir et al 2015, 2017]

How deep is the issue?

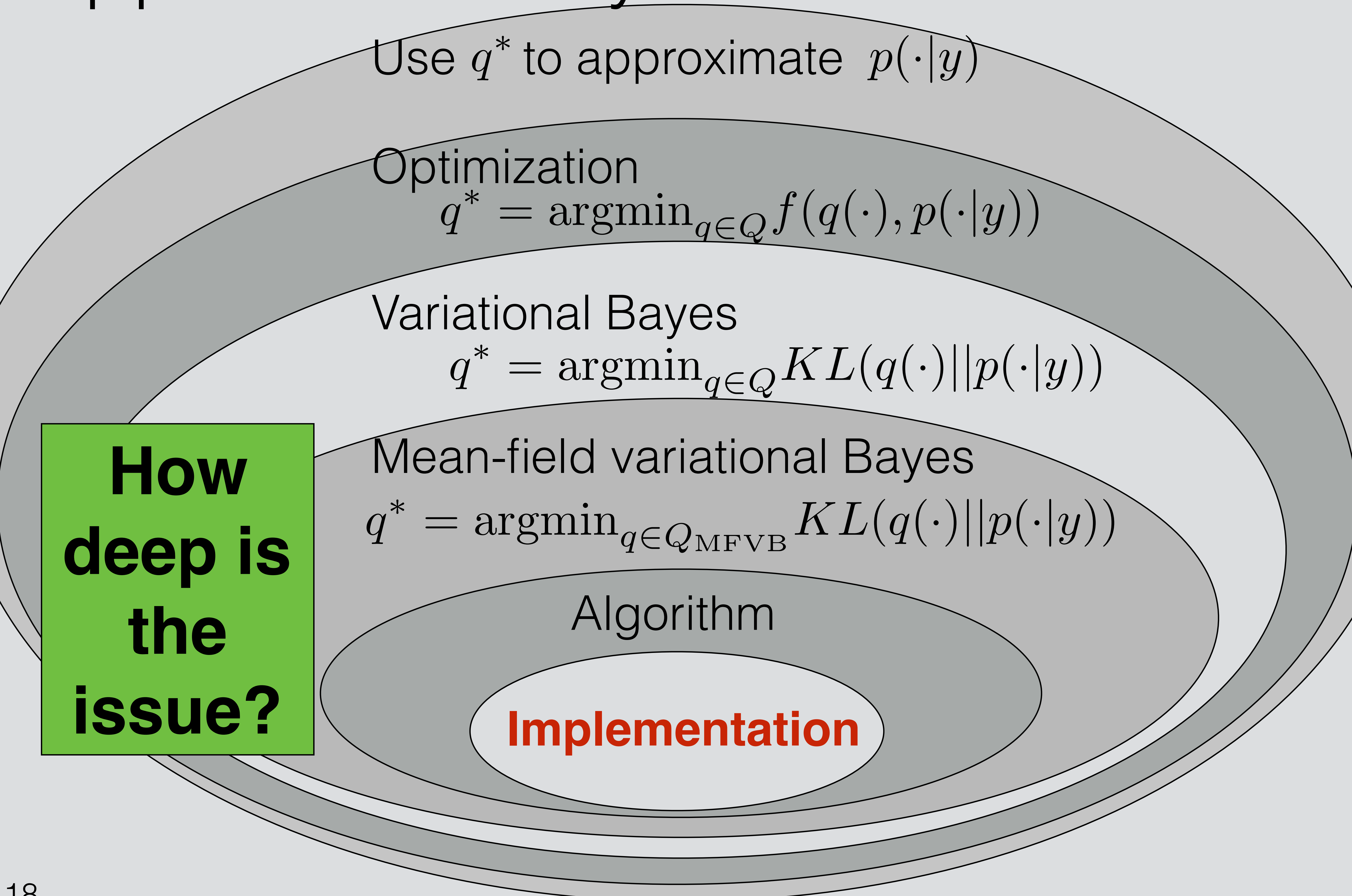
Approximate Bayesian inference



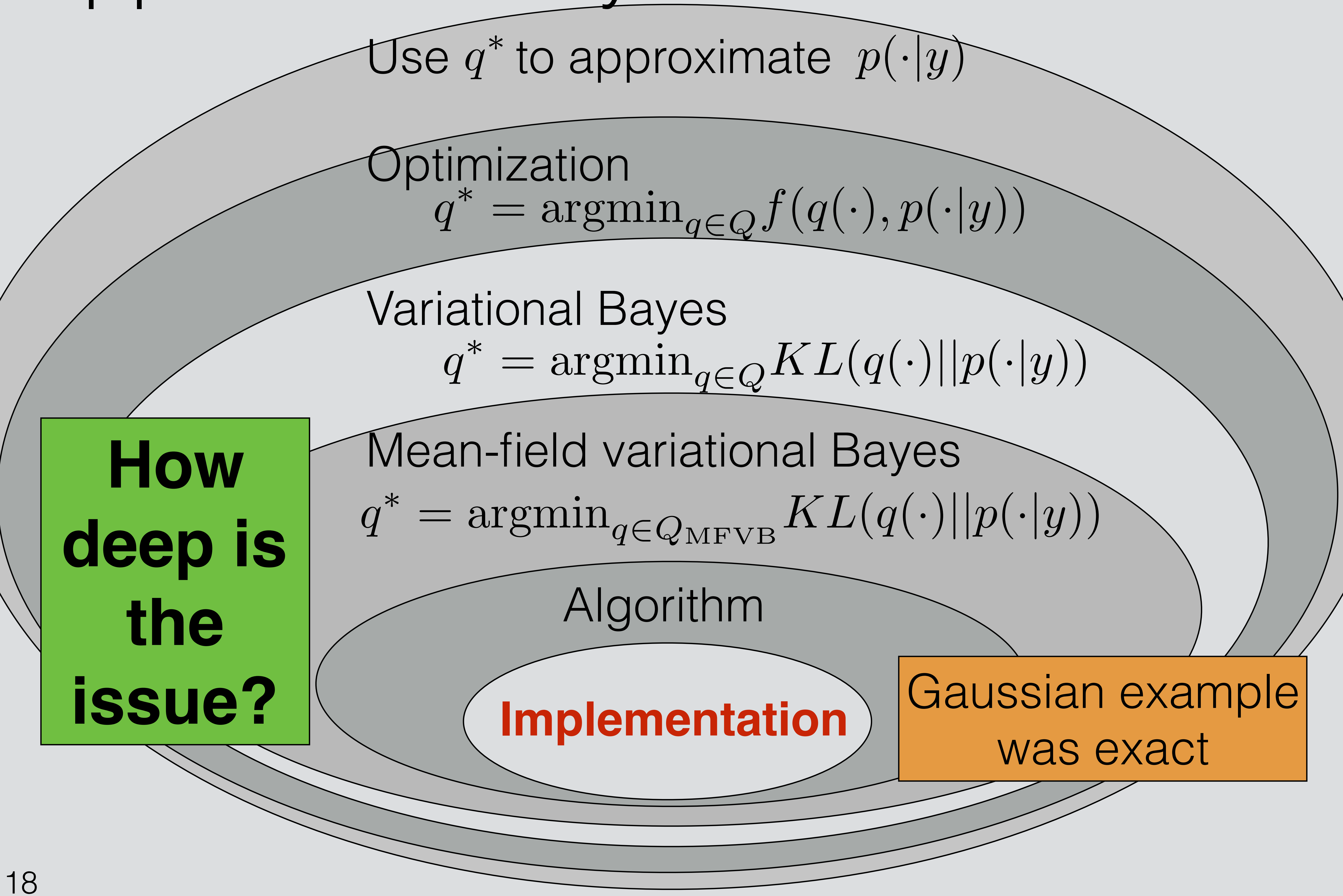
Approximate Bayesian inference



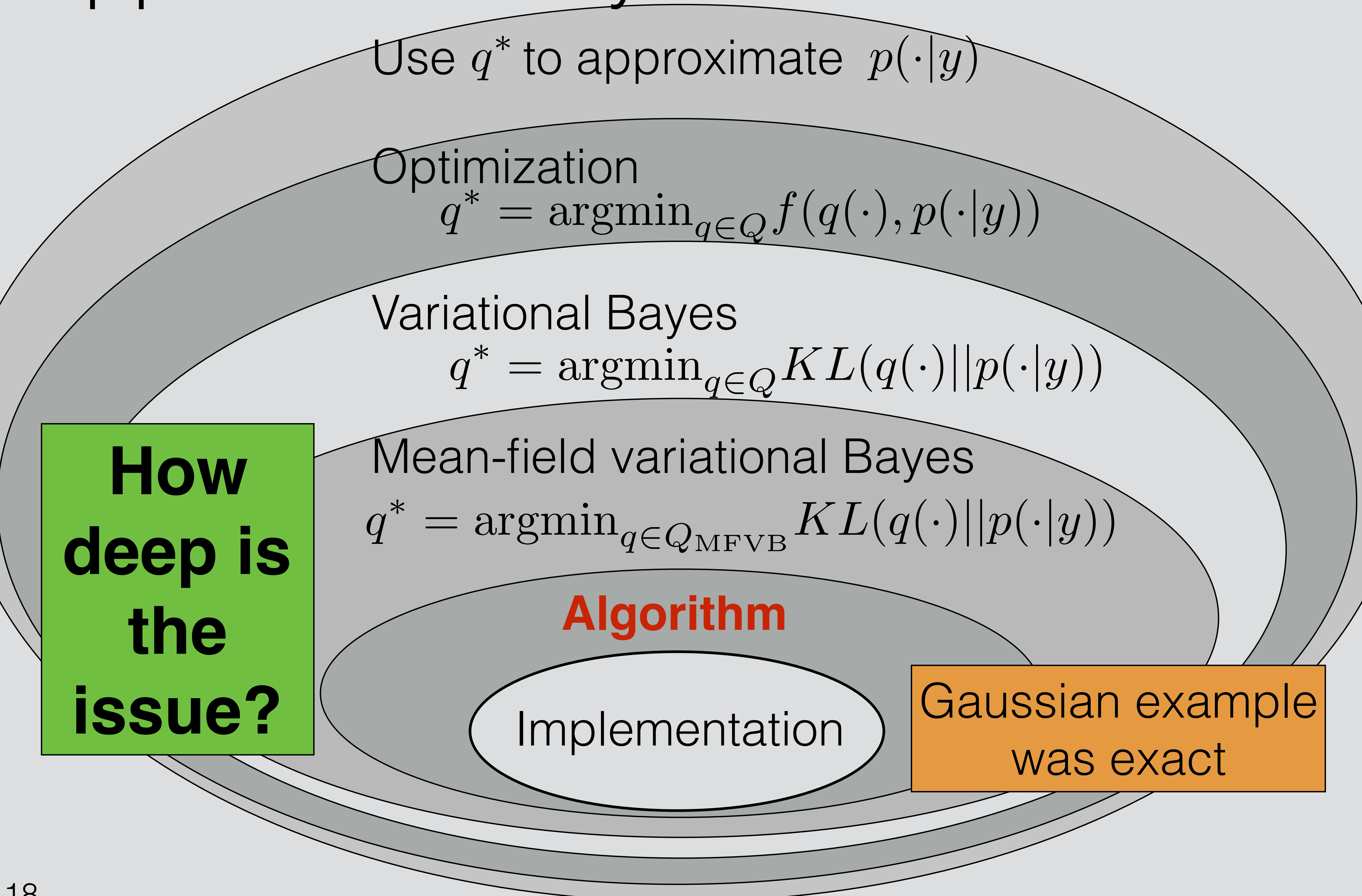
Approximate Bayesian inference



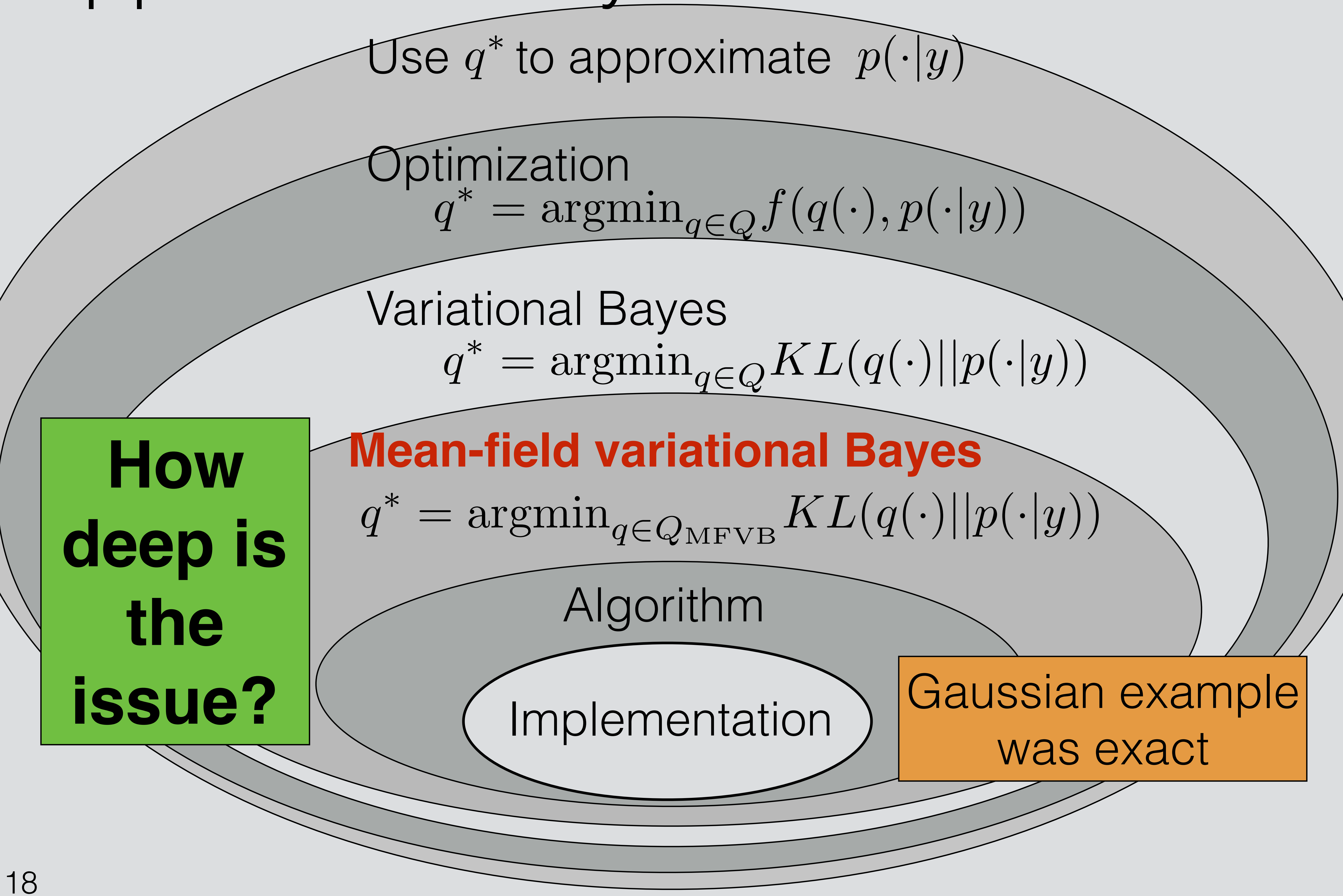
Approximate Bayesian inference



Approximate Bayesian inference



Approximate Bayesian inference

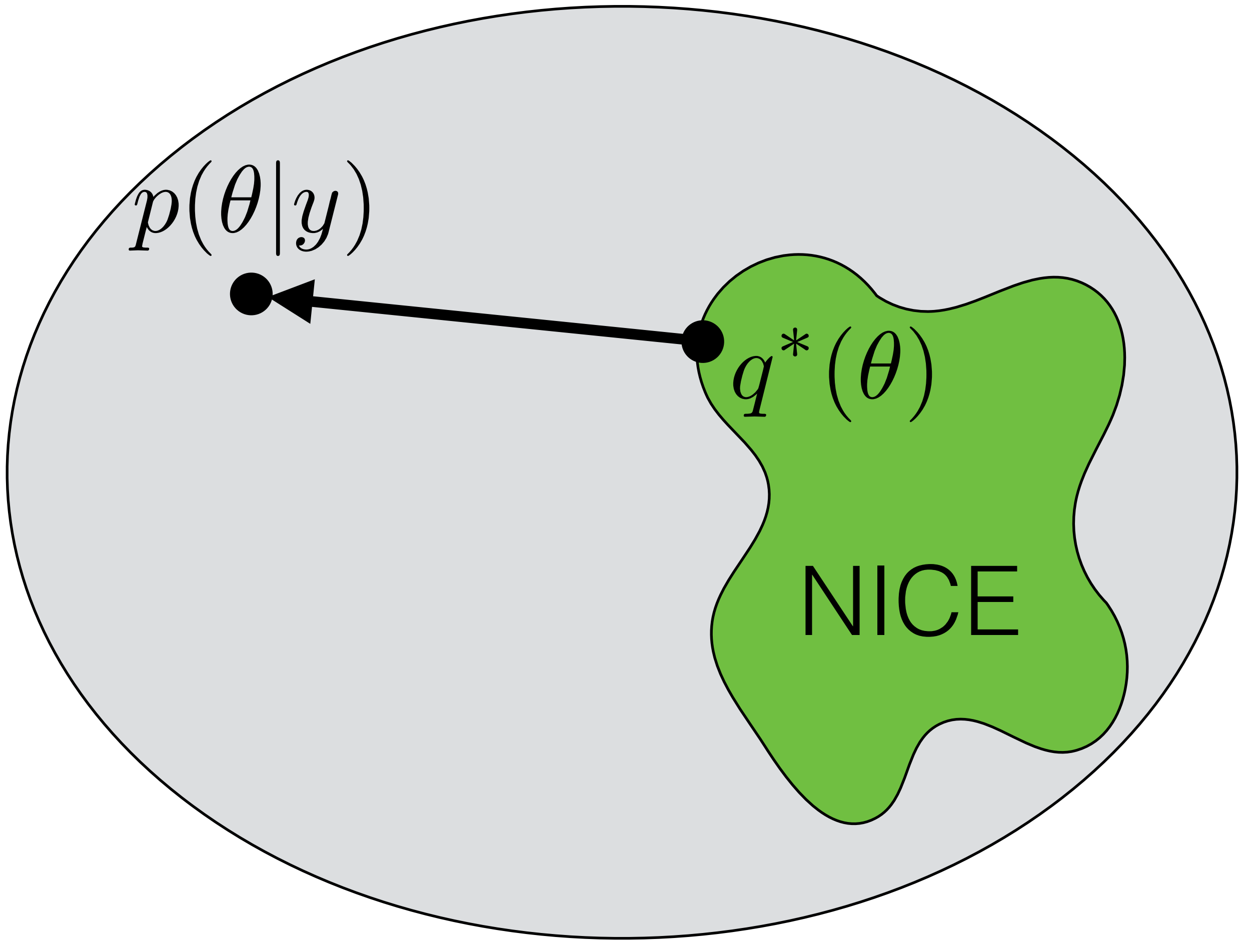


Is it just MFVB?

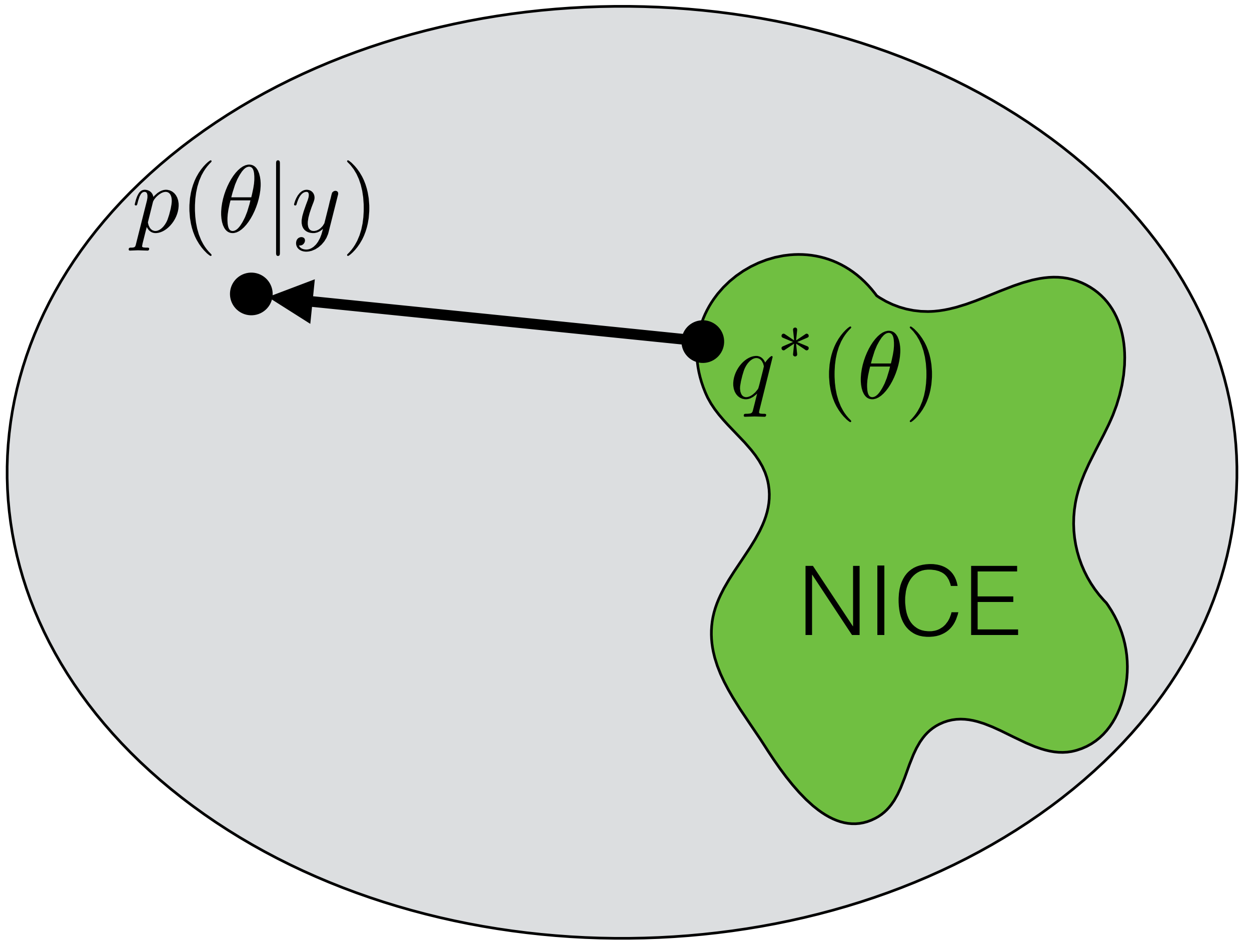
Is it just **MF**VB?

Is it just MFVB?

Is it just MFVB?

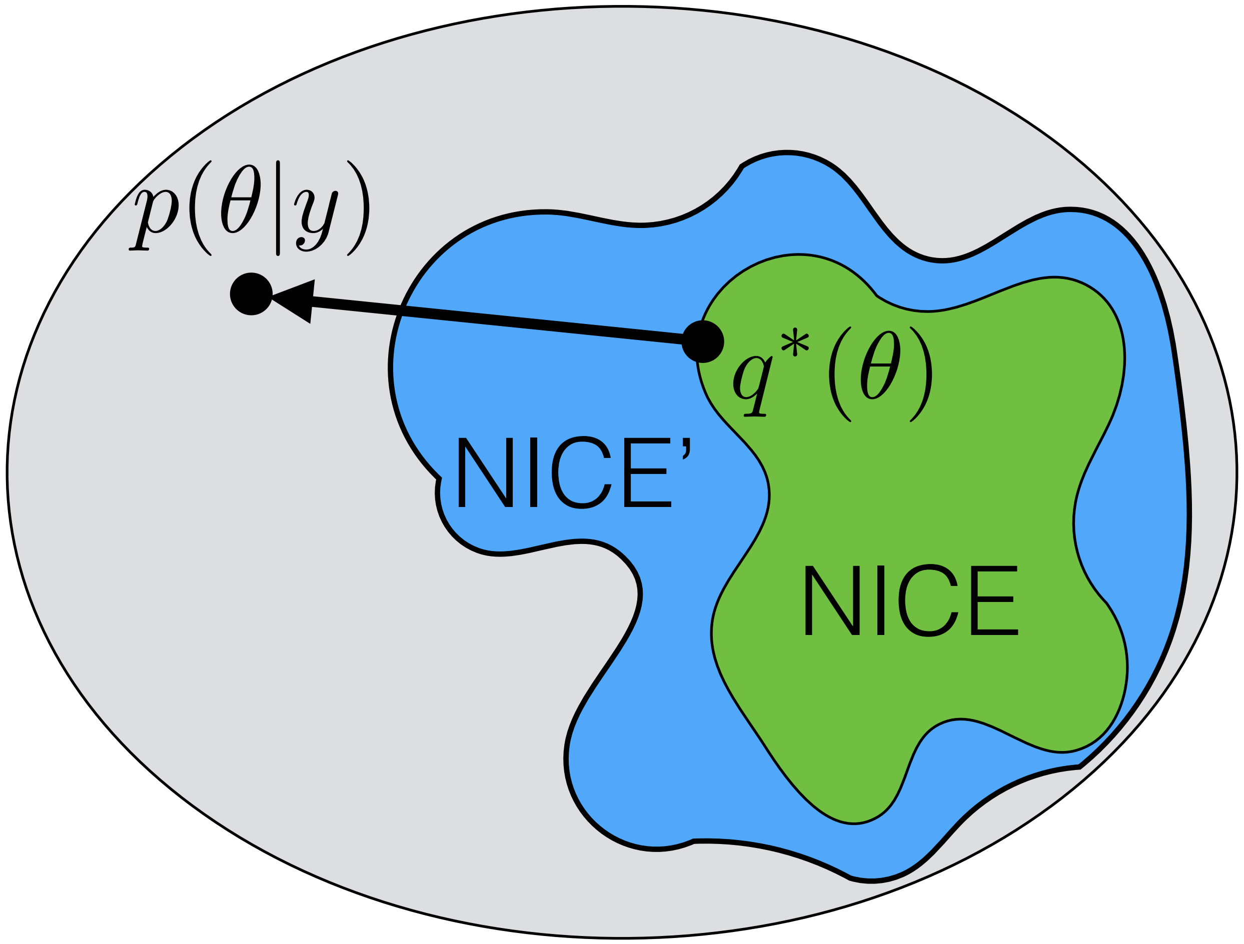


Is it just MFVB?



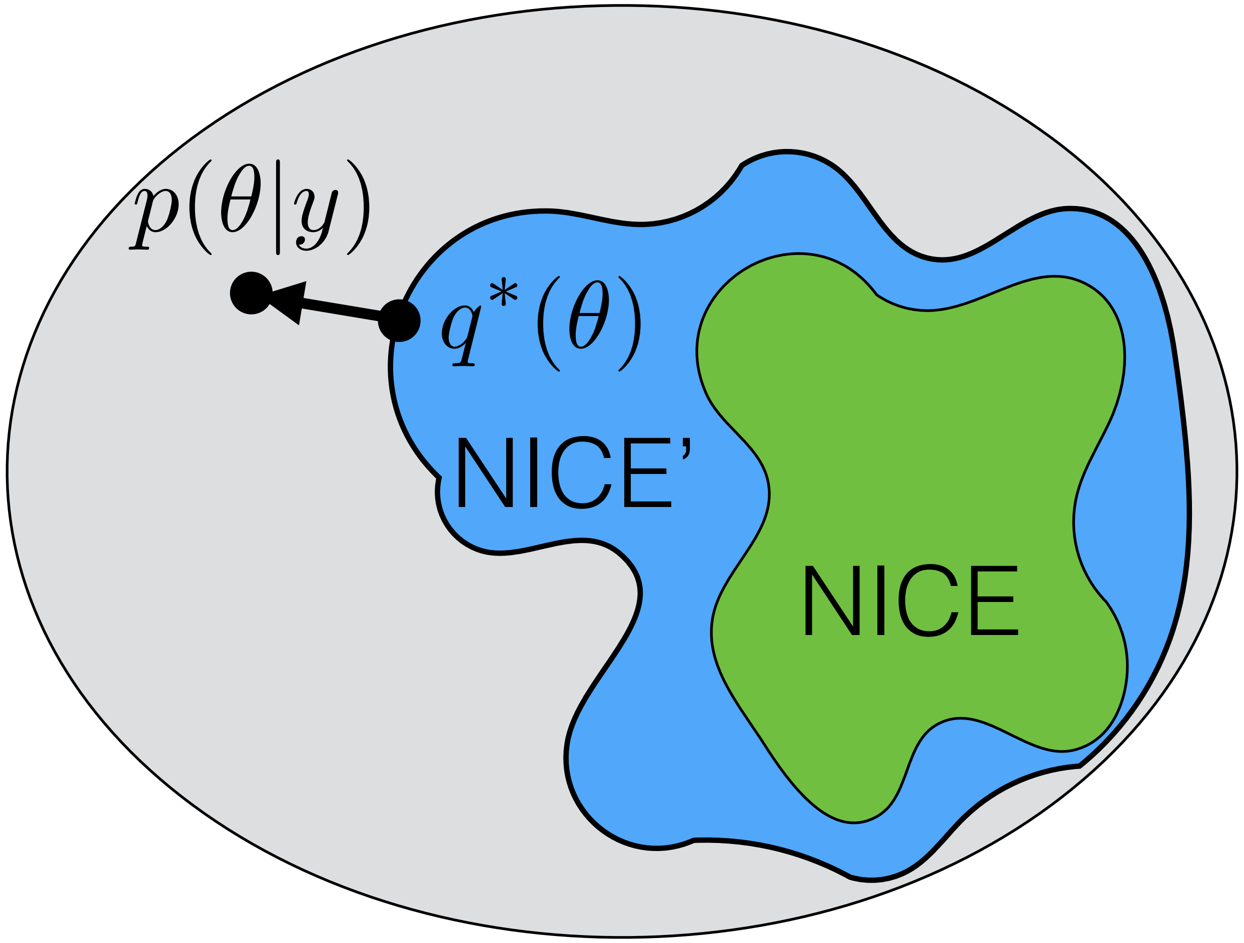
- Turner, Sahani (2011) showed (empirically) can have strictly larger NICE set but worse mean & variance estimates

Is it just MFVB?



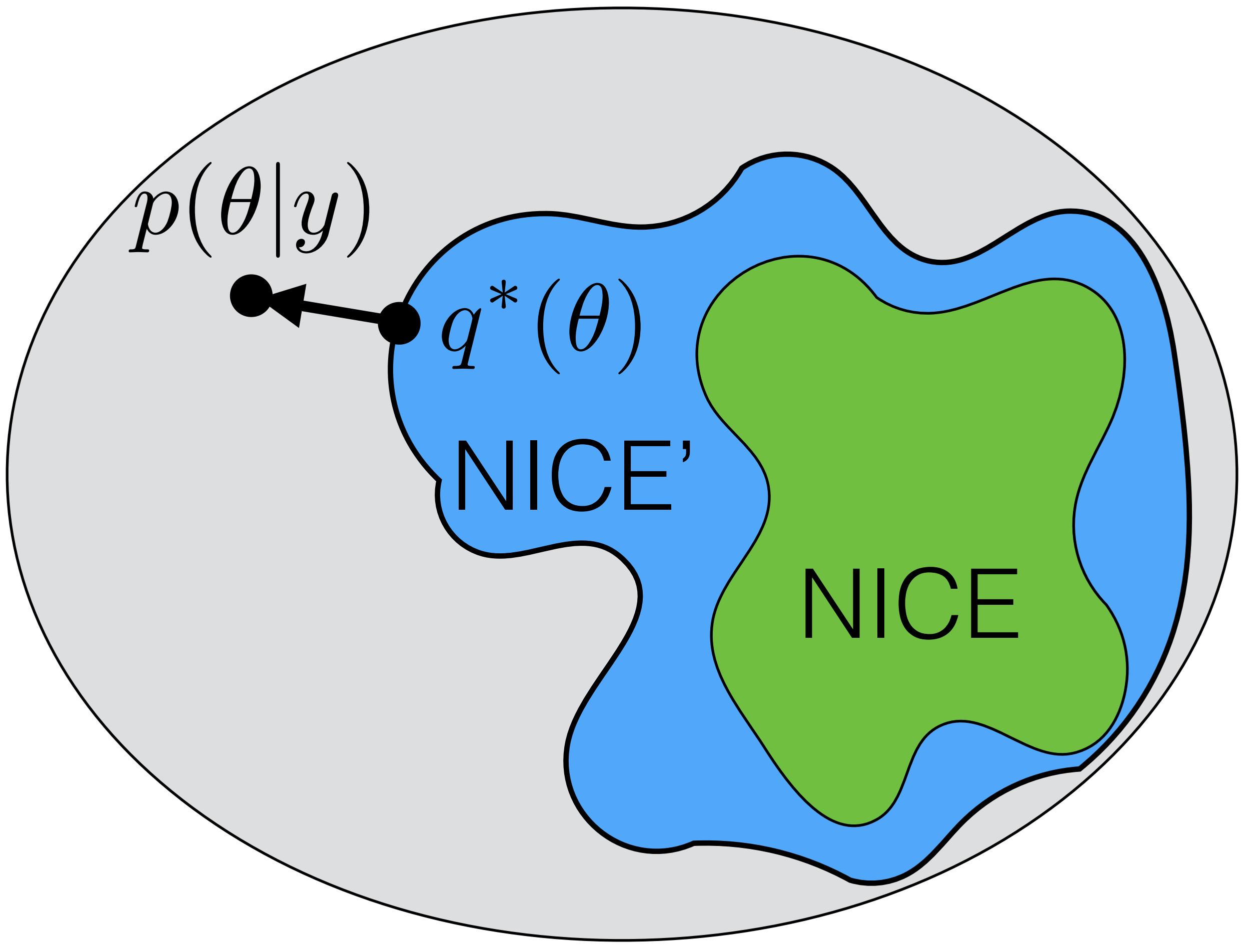
- Turner, Sahani (2011) showed (empirically) can have strictly larger NICE set but worse mean & variance estimates

Is it just MFVB?



- Turner, Sahani (2011) showed (empirically) can have strictly larger NICE set but worse mean & variance estimates

Is it just MFVB?



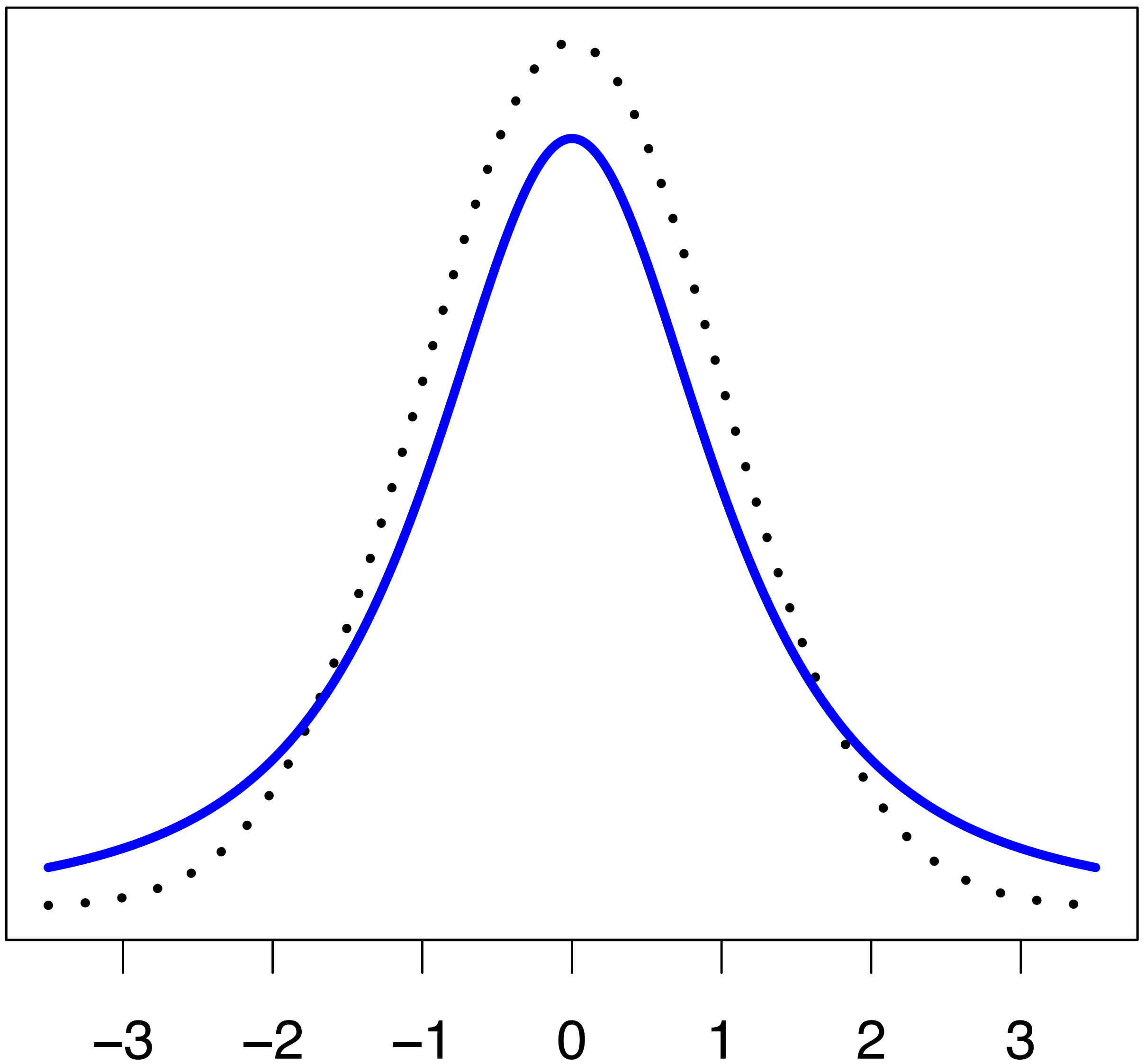
- Turner, Sahani (2011) showed (empirically) can have strictly larger NICE set but worse mean & variance estimates

- Takeaway: A smaller KL does not imply better mean and variance estimates
- Exercise: show this

Is it just MFVB?

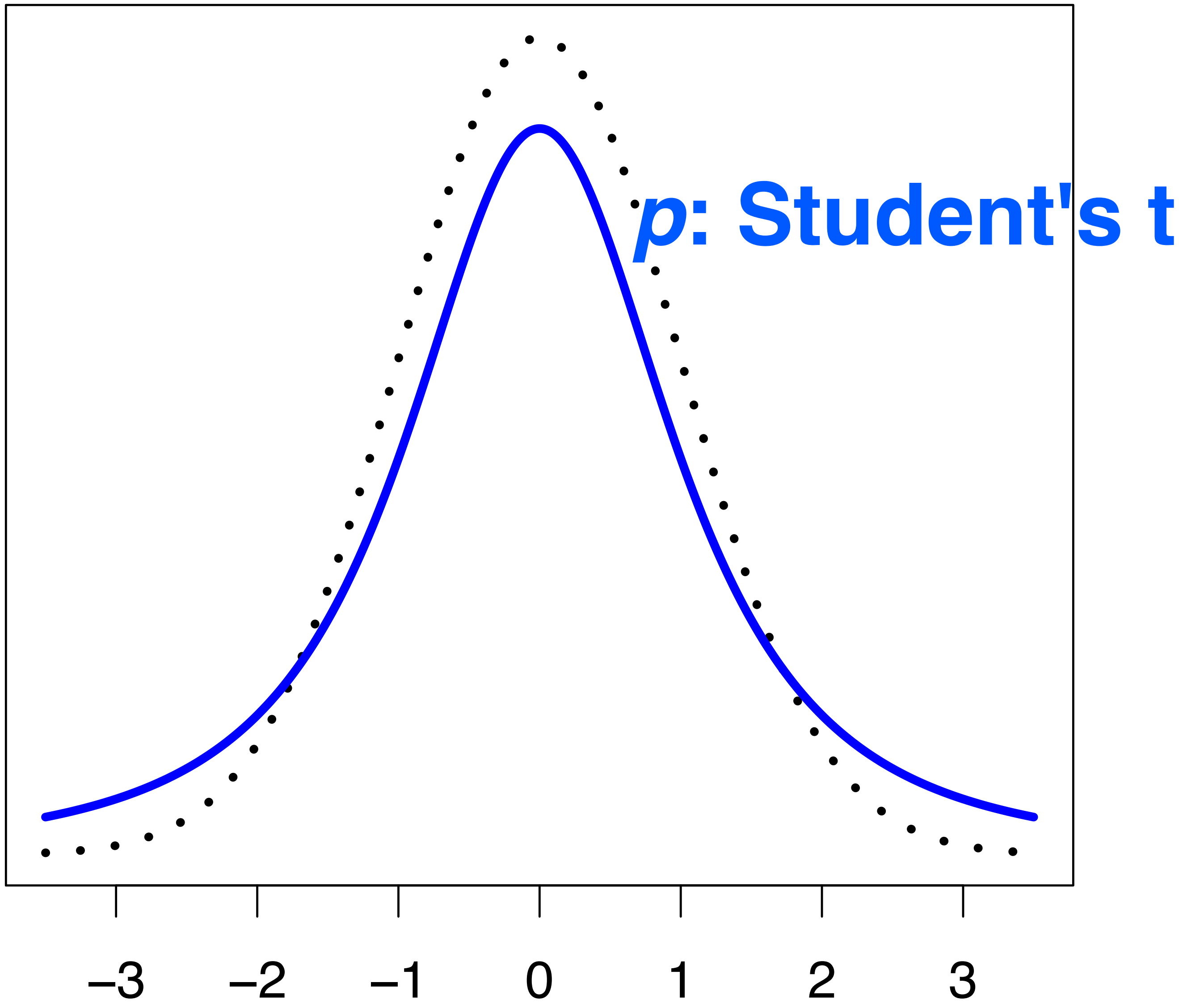
[Huggins, Karsprzak, Campbell, Broderick 2018]

Is it just MFVB?



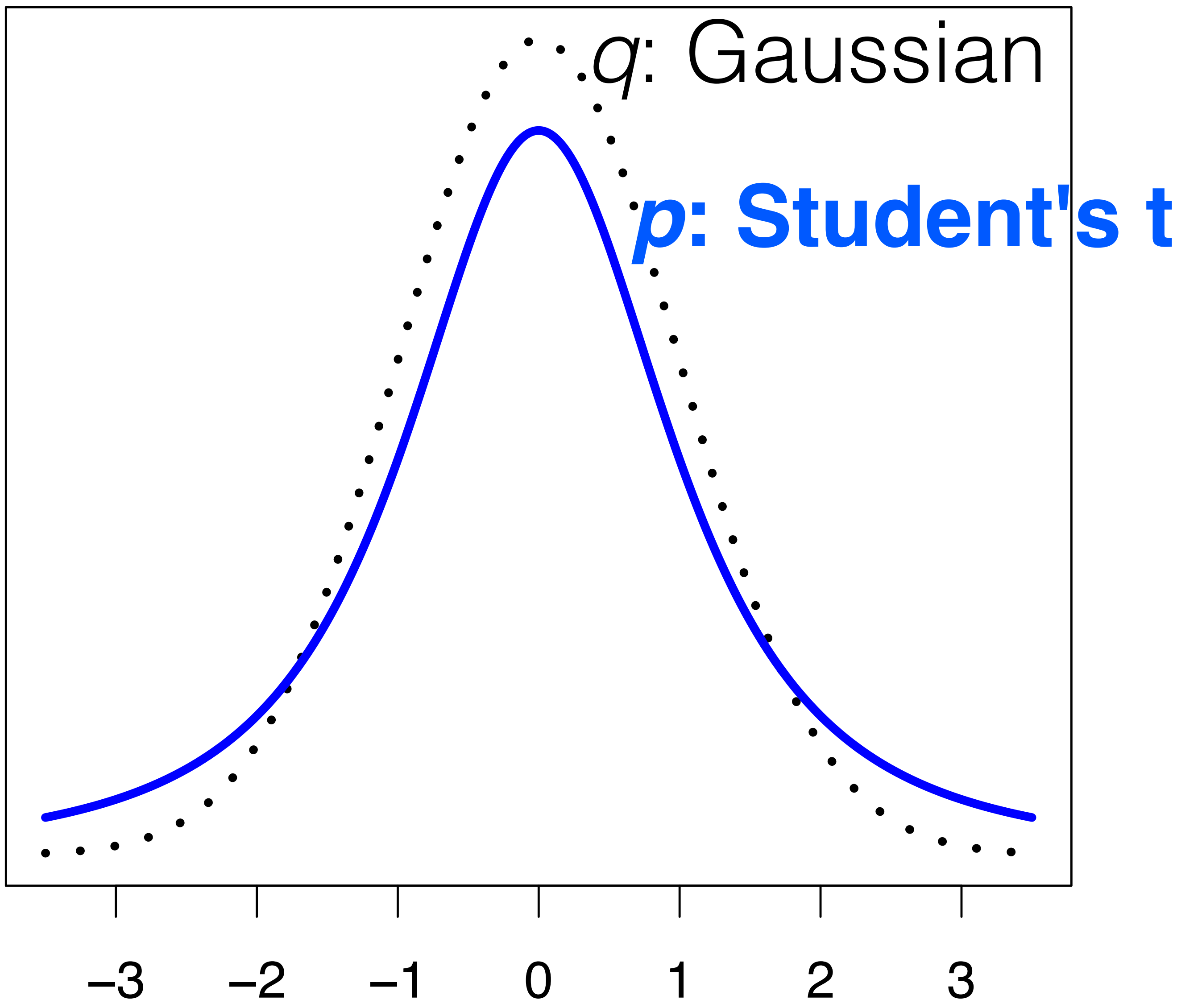
[Huggins, Karsprzak, Campbell, Broderick 2018]

Is it just MFVB?



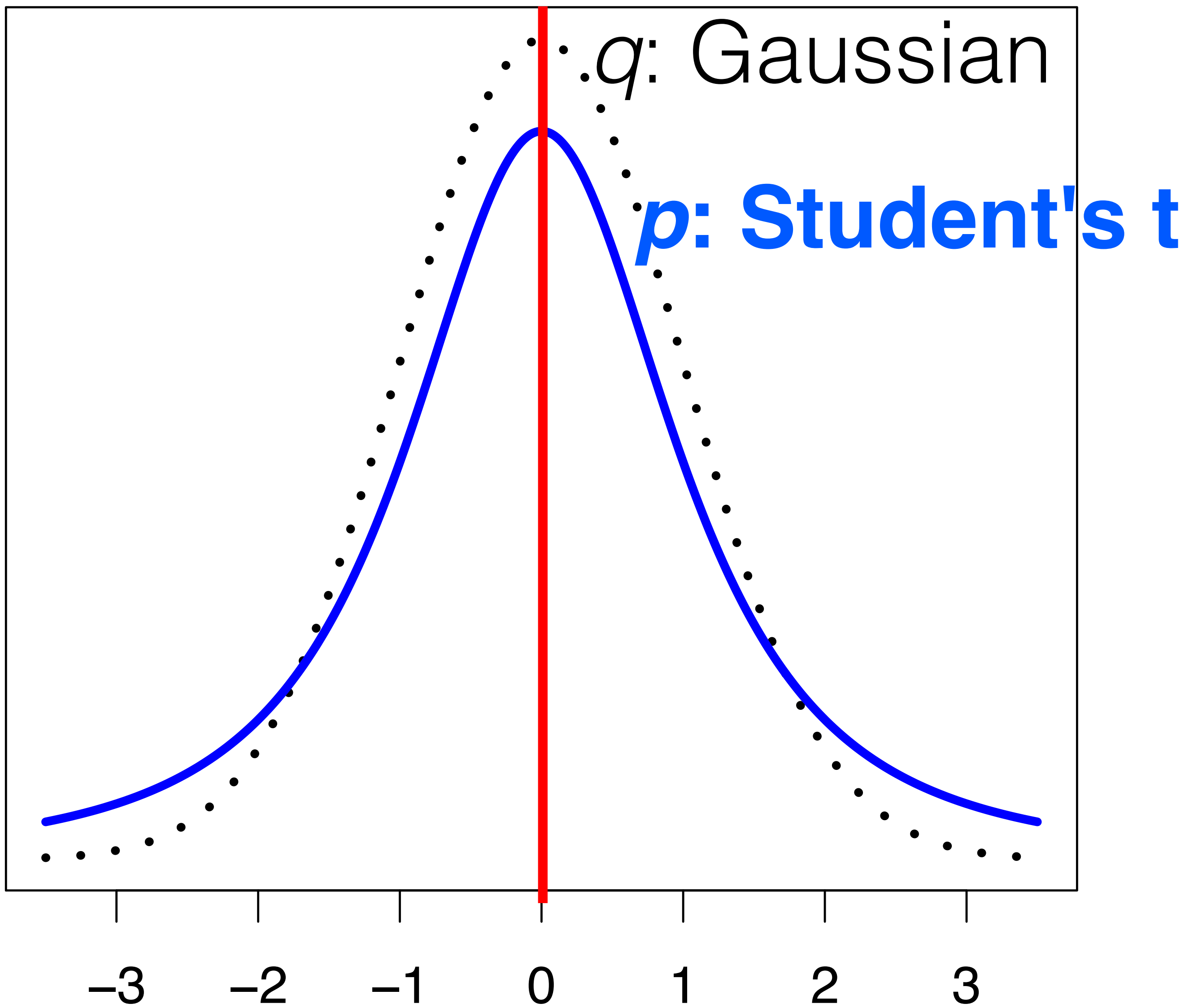
[Huggins, Karsprzak, Campbell, Broderick 2018]

Is it just MFVB?



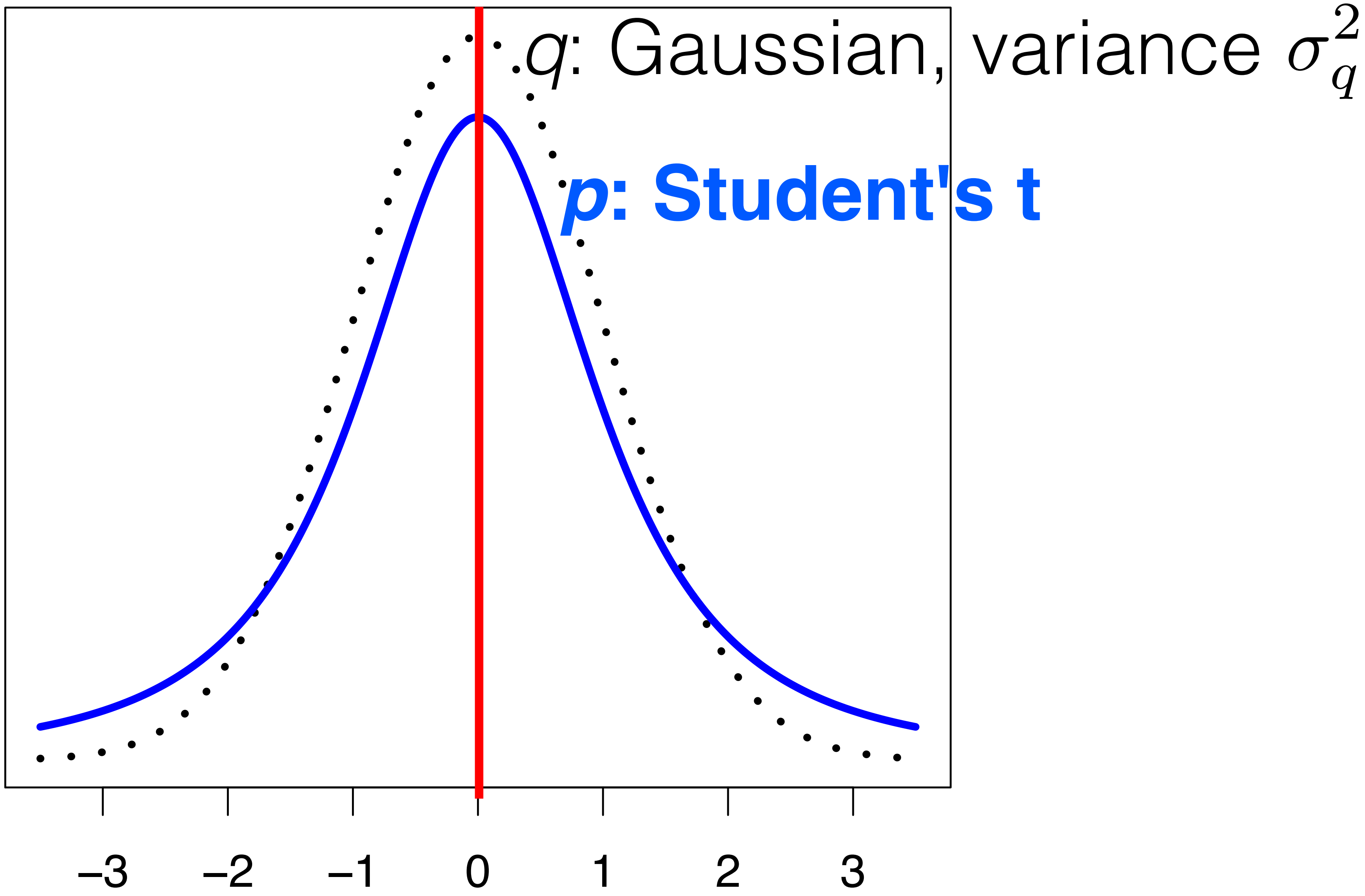
[Huggins, Karsprzak, Campbell, Broderick 2018]

Is it just MFVB?



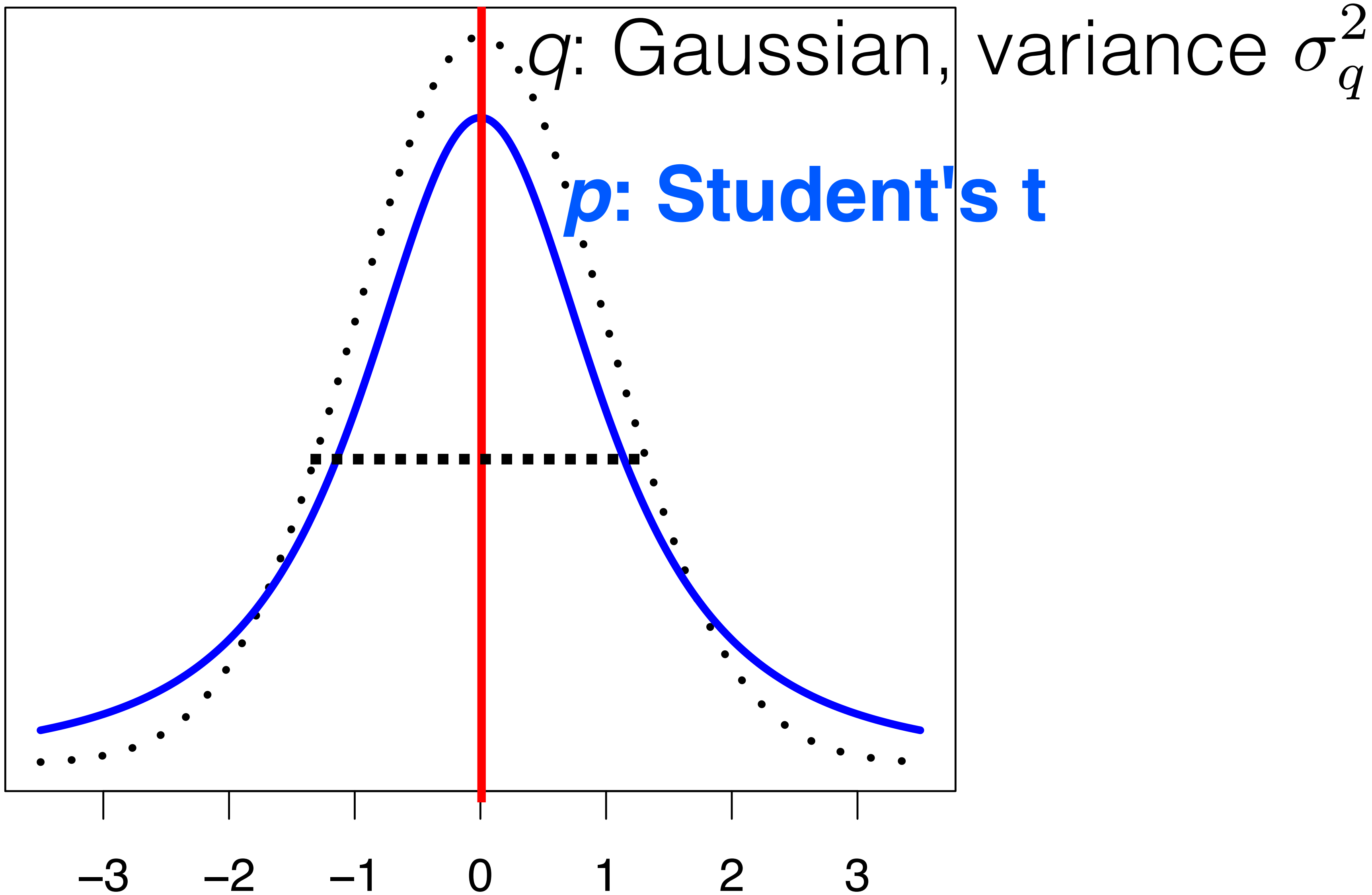
[Huggins, Karsprzak, Campbell, Broderick 2018]

Is it just MFVB?



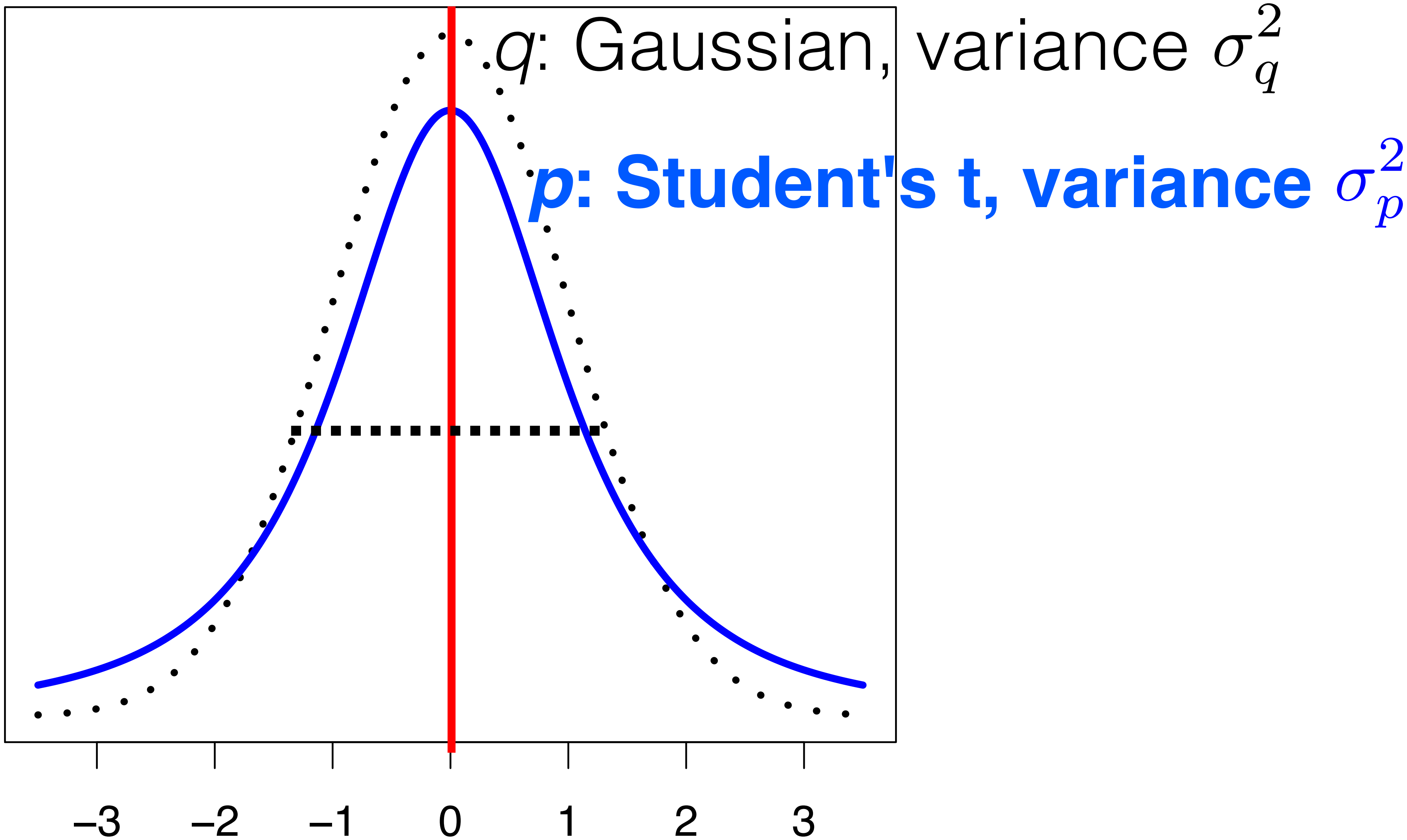
[Huggins, Karsprzak, Campbell, Broderick 2018]

Is it just MFVB?



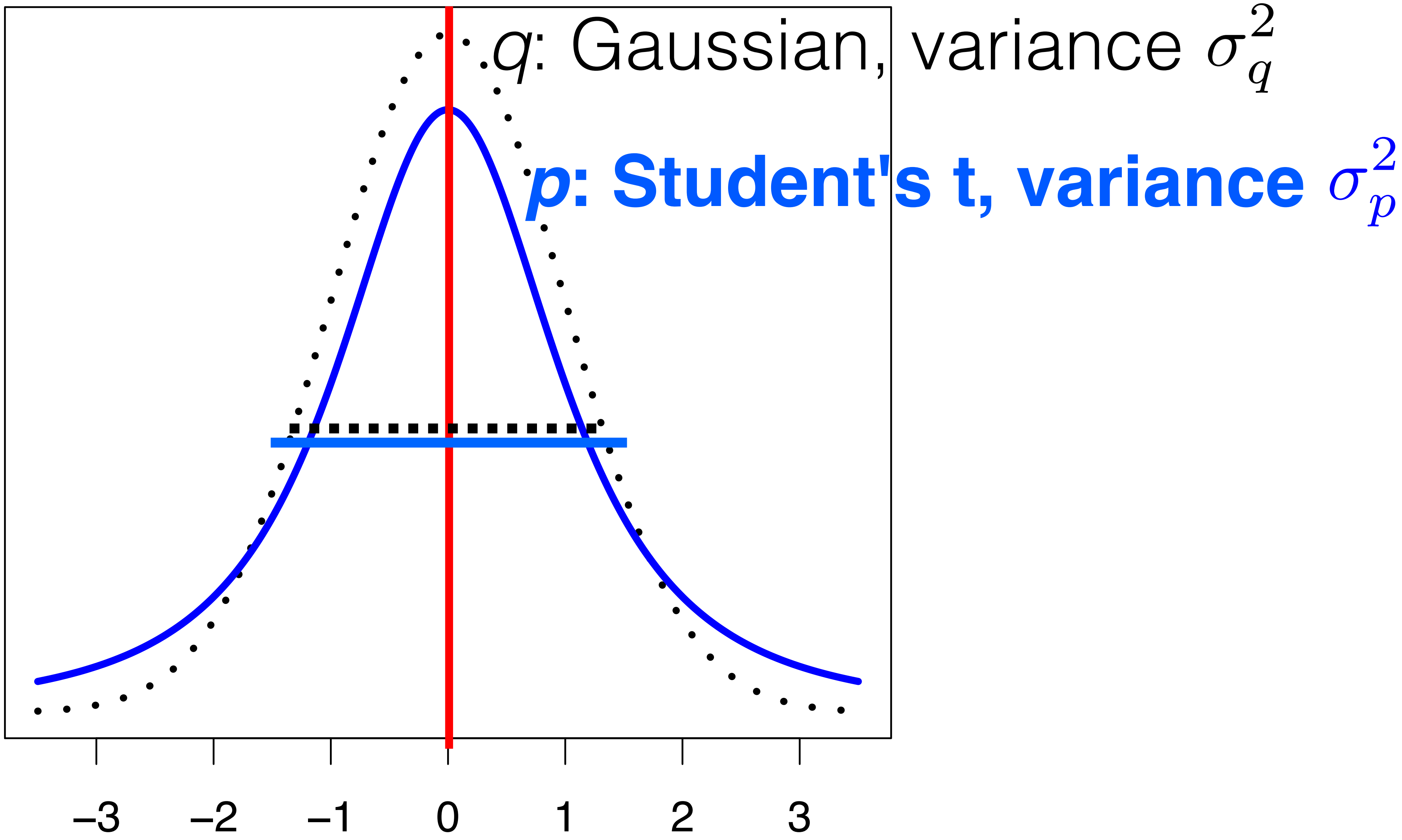
[Huggins, Karsprzak, Campbell, Broderick 2018]

Is it just MFVB?



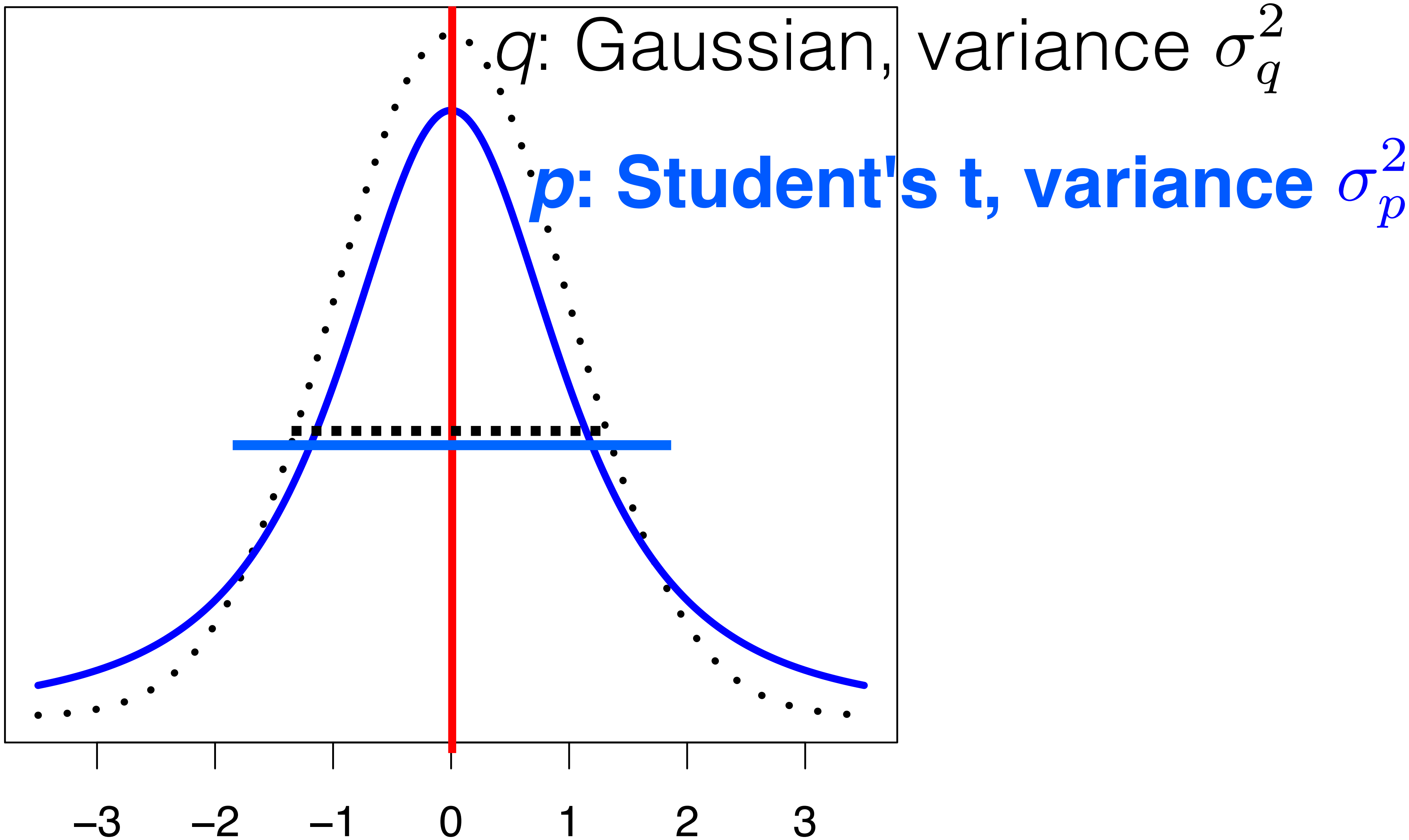
[Huggins, Karsprzak, Campbell, Broderick 2018]

Is it just MFVB?



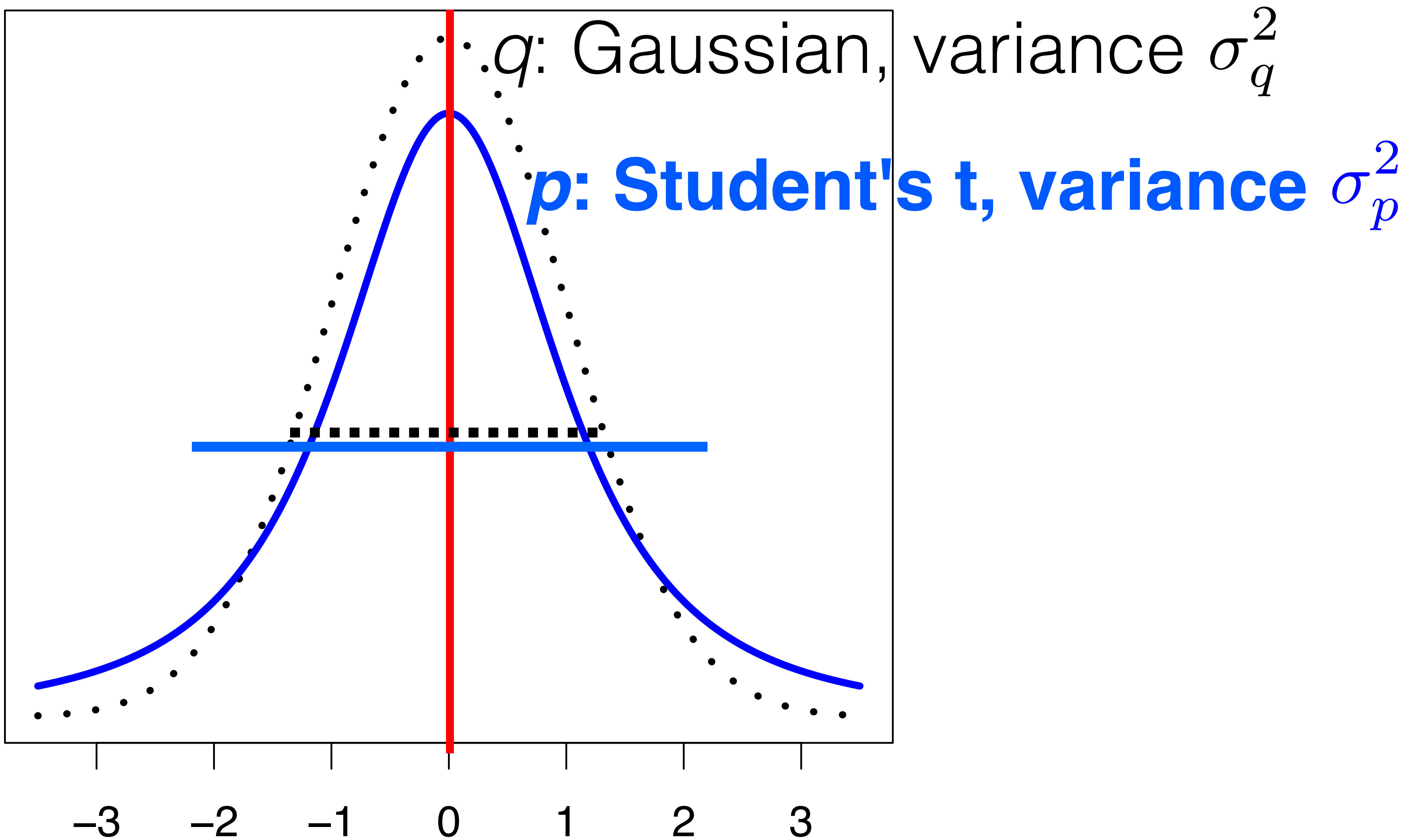
[Huggins, Karsprzak, Campbell, Broderick 2018]

Is it just MFVB?



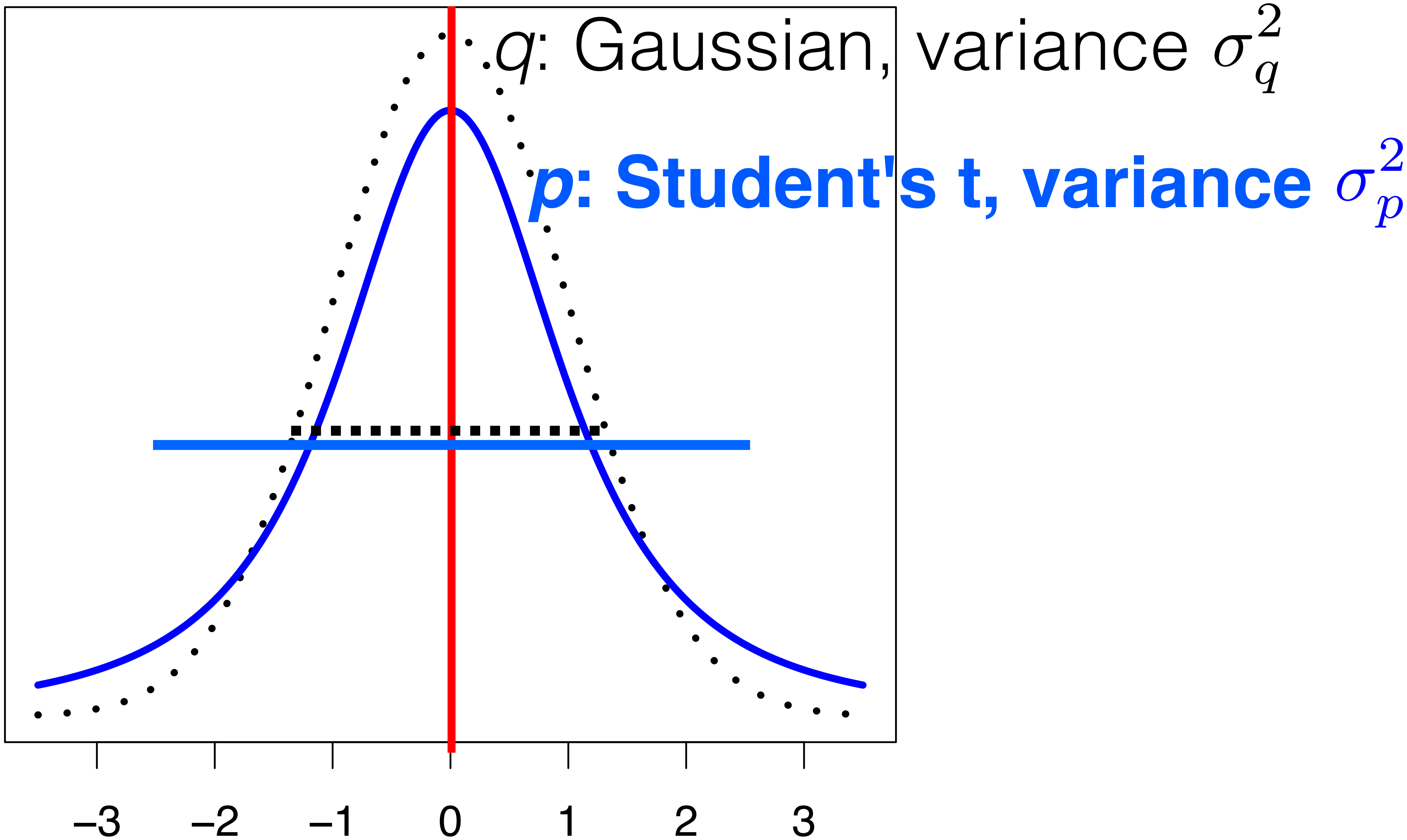
[Huggins, Karsprzak, Campbell, Broderick 2018]

Is it just MFVB?



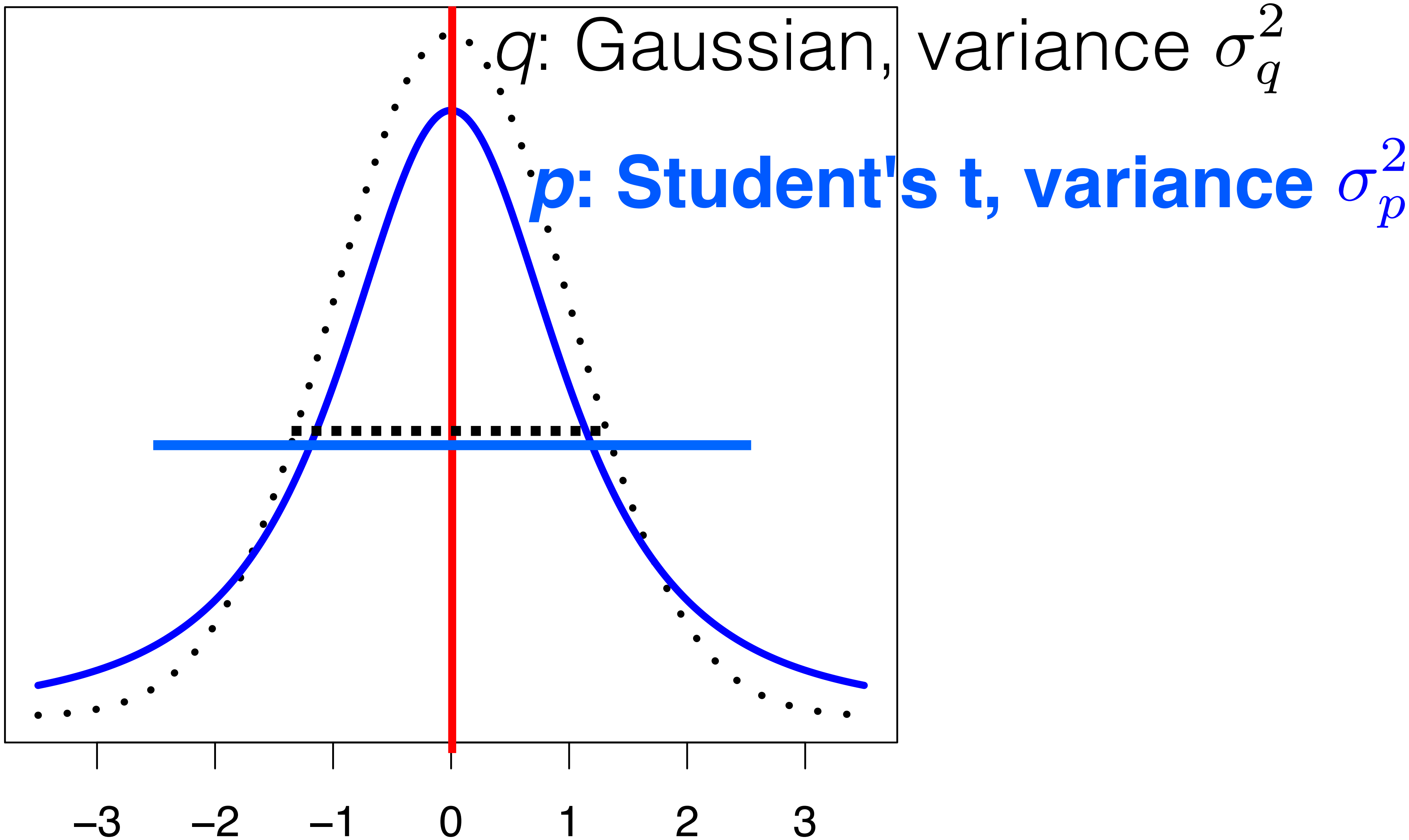
[Huggins, Karsprzak, Campbell, Broderick 2018]

Is it just MFVB?



[Huggins, Karsprzak, Campbell, Broderick 2018]

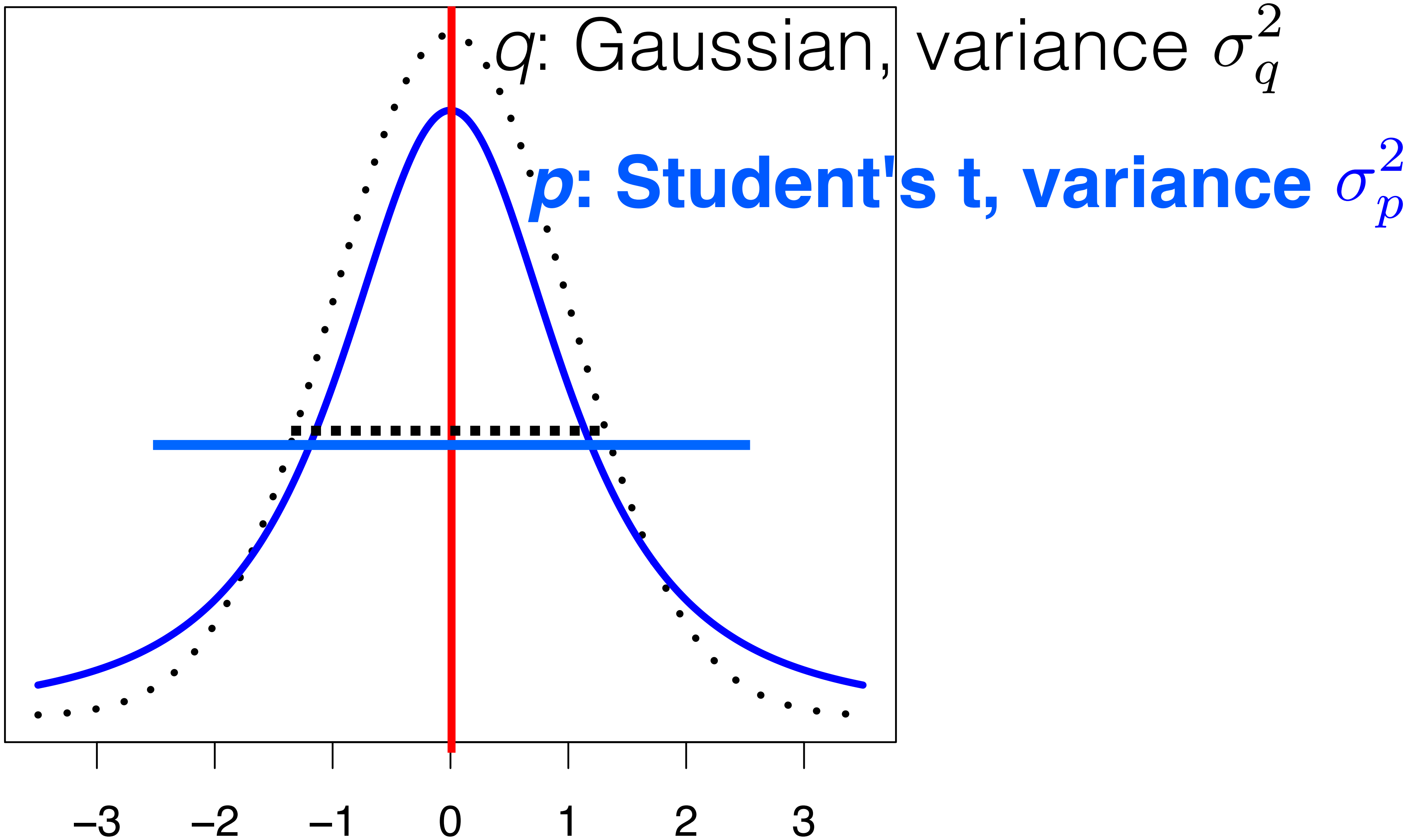
Is it just MFVB?



$$\sigma_p^2 \geq c\sigma_q^2$$

[Huggins, Karsprzak, Campbell, Broderick 2018]

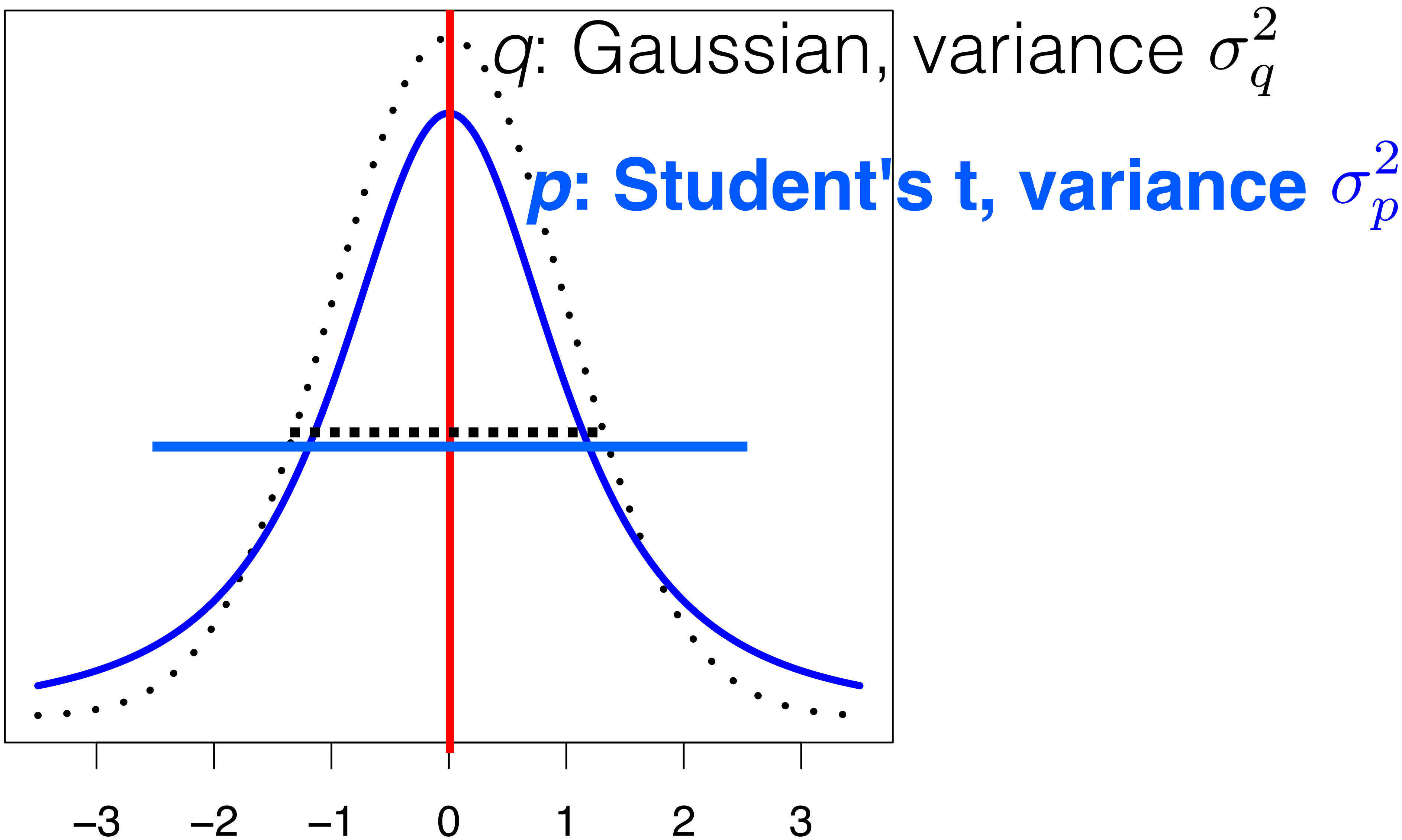
Is it just MFVB?



$$KL(q||p) < 0.802 \text{ but also } \sigma_p^2 \geq c\sigma_q^2$$

[Huggins, Karsprzak, Campbell, Broderick 2018]

Is it just MFVB?

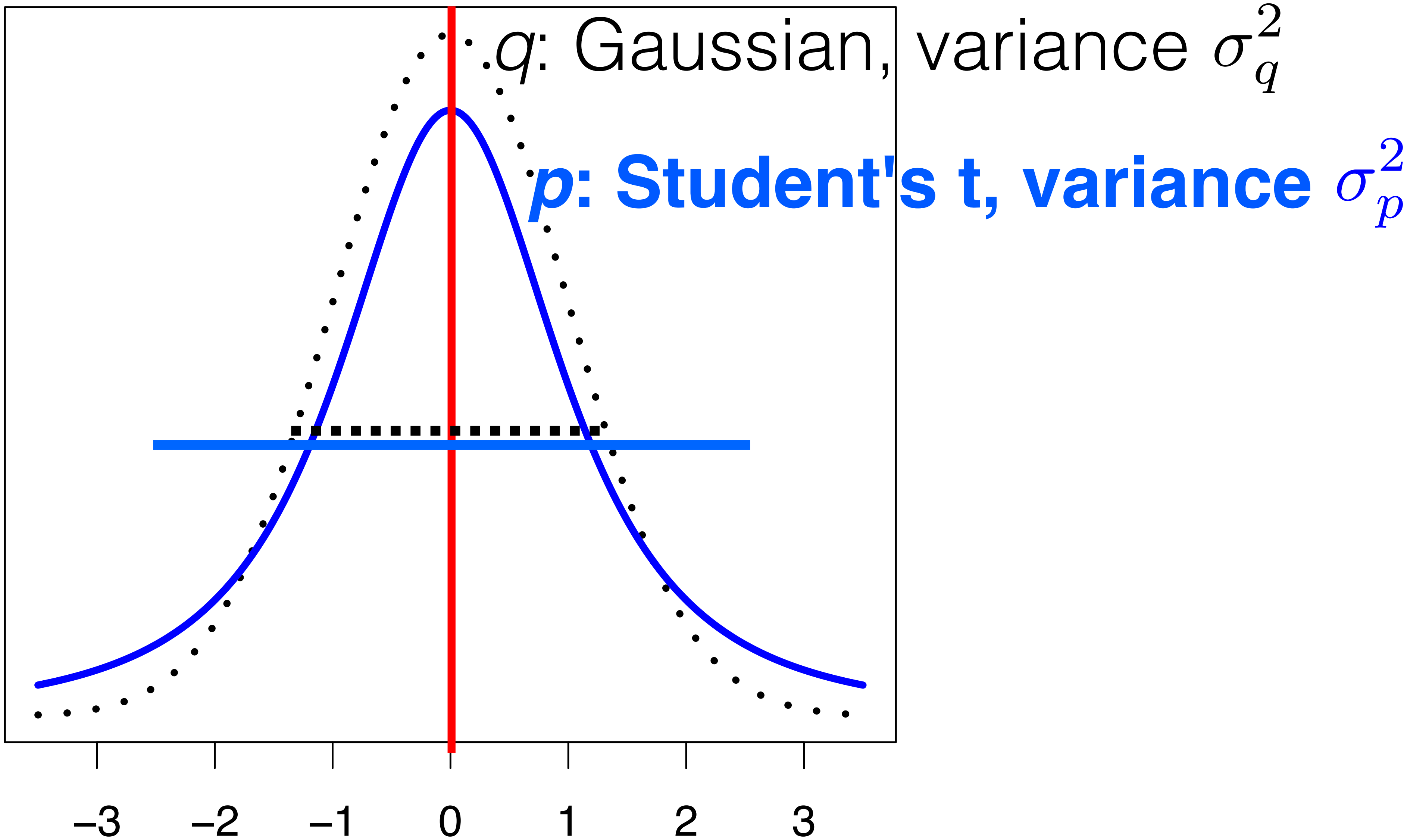


Proposition (HKCB). For any $c > 1$, there exist zero-mean, unimodal distributions q and p such that

$$KL(q||p) < 0.802 \text{ but also } \sigma_p^2 \geq c\sigma_q^2$$

[Huggins, Karsprzak, Campbell, Broderick 2018]

Is it just MFVB?



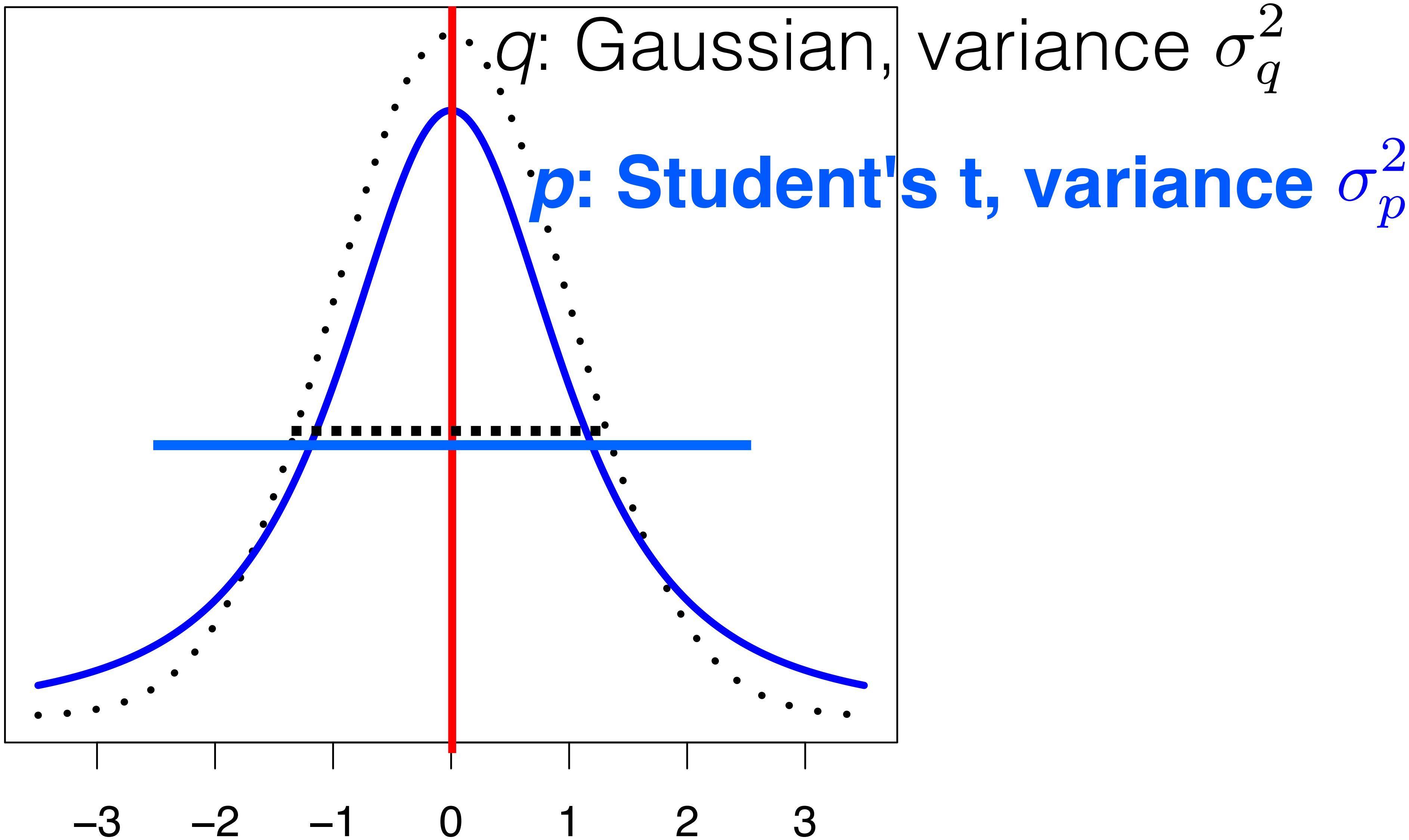
Proposition (HKCB). For any $c > 1$, there exist zero-mean, unimodal distributions q and p such that

$$KL(q||p) < 0.802 \text{ but also } \sigma_p^2 \geq c\sigma_q^2$$

Can have small KL and arbitrarily bad variance estimate

[Huggins, Karsprzak, Campbell, Broderick 2018]

Is it just MFVB?



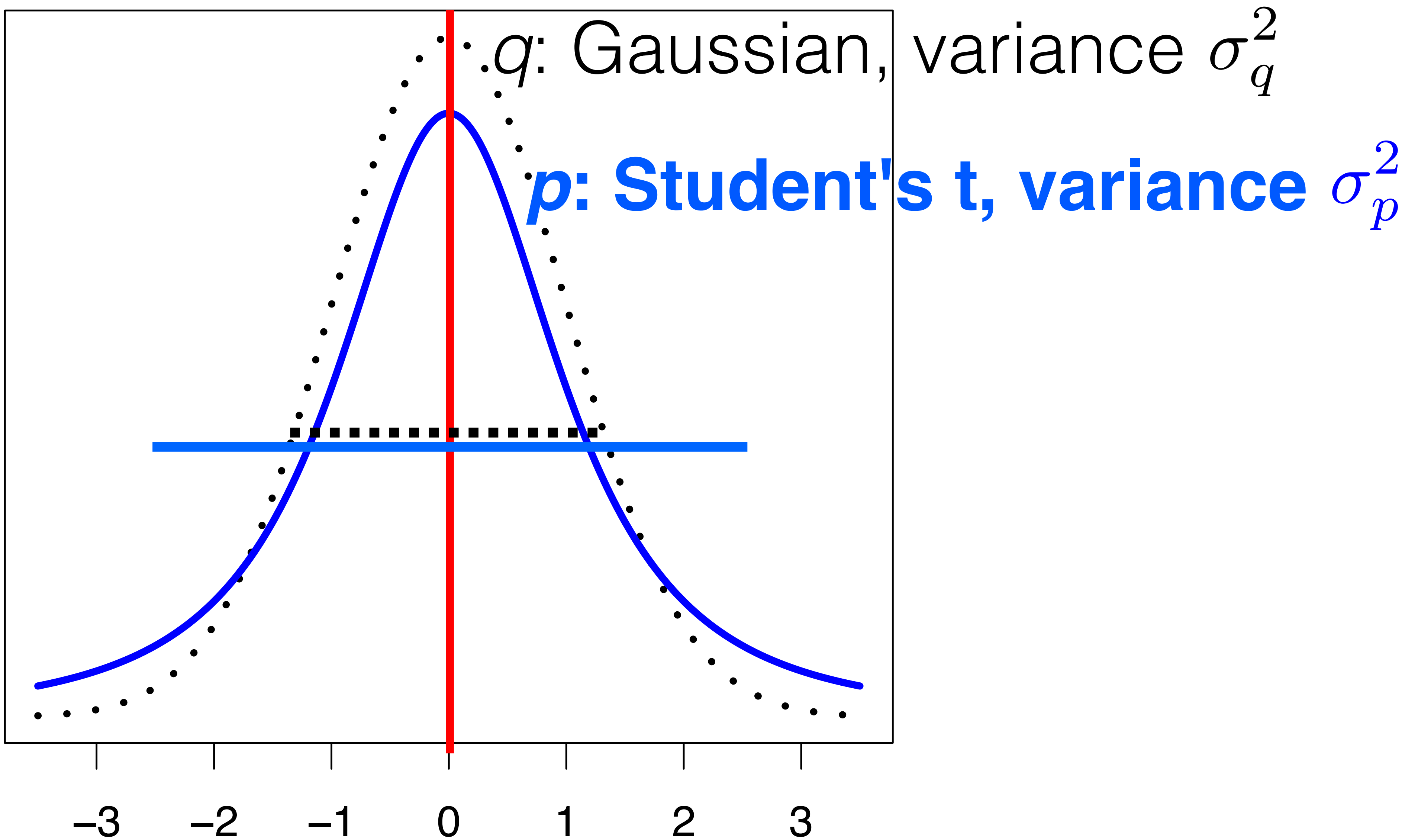
Proposition (HKCB). For any $c > 1$, there exist zero-mean, unimodal distributions q and p such that

$$KL(q||p) < 0.802 \text{ but also } \sigma_p^2 \geq c\sigma_q^2$$

Can have small KL and arbitrarily bad variance estimate

[Huggins, Karsprzak, Campbell, Broderick 2018]

Is it just MFVB?



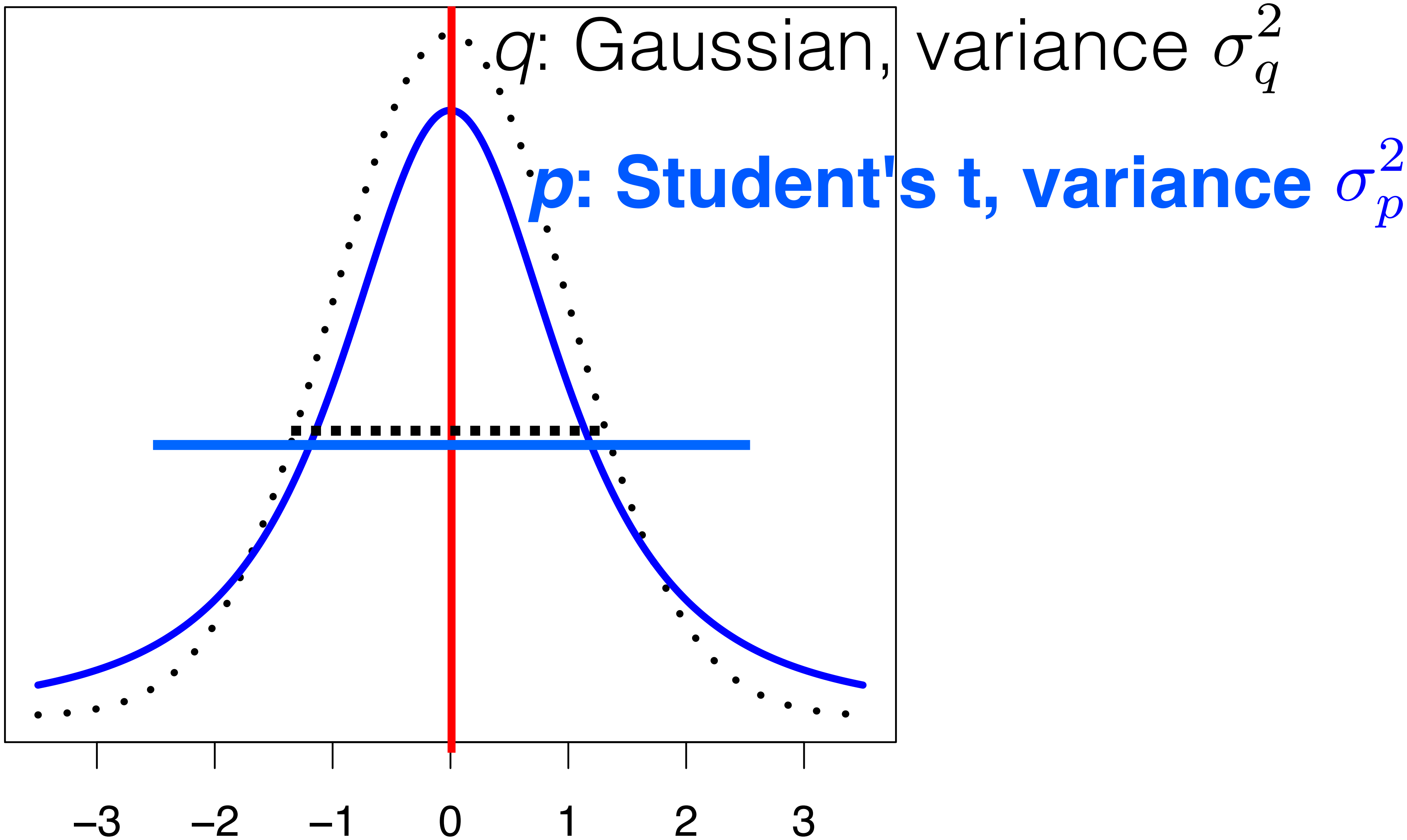
Conjecture (HKCB). For any $c > 1$, there exist zero-mean, unimodal distributions q and p such that

$$KL(q||p) < 0.802 \text{ but also } \sigma_p^2 \geq c\sigma_q^2$$

Can have small KL and arbitrarily bad variance estimate

[Huggins, Karsprzak, Campbell, Broderick 2018]

Is it just MFVB?



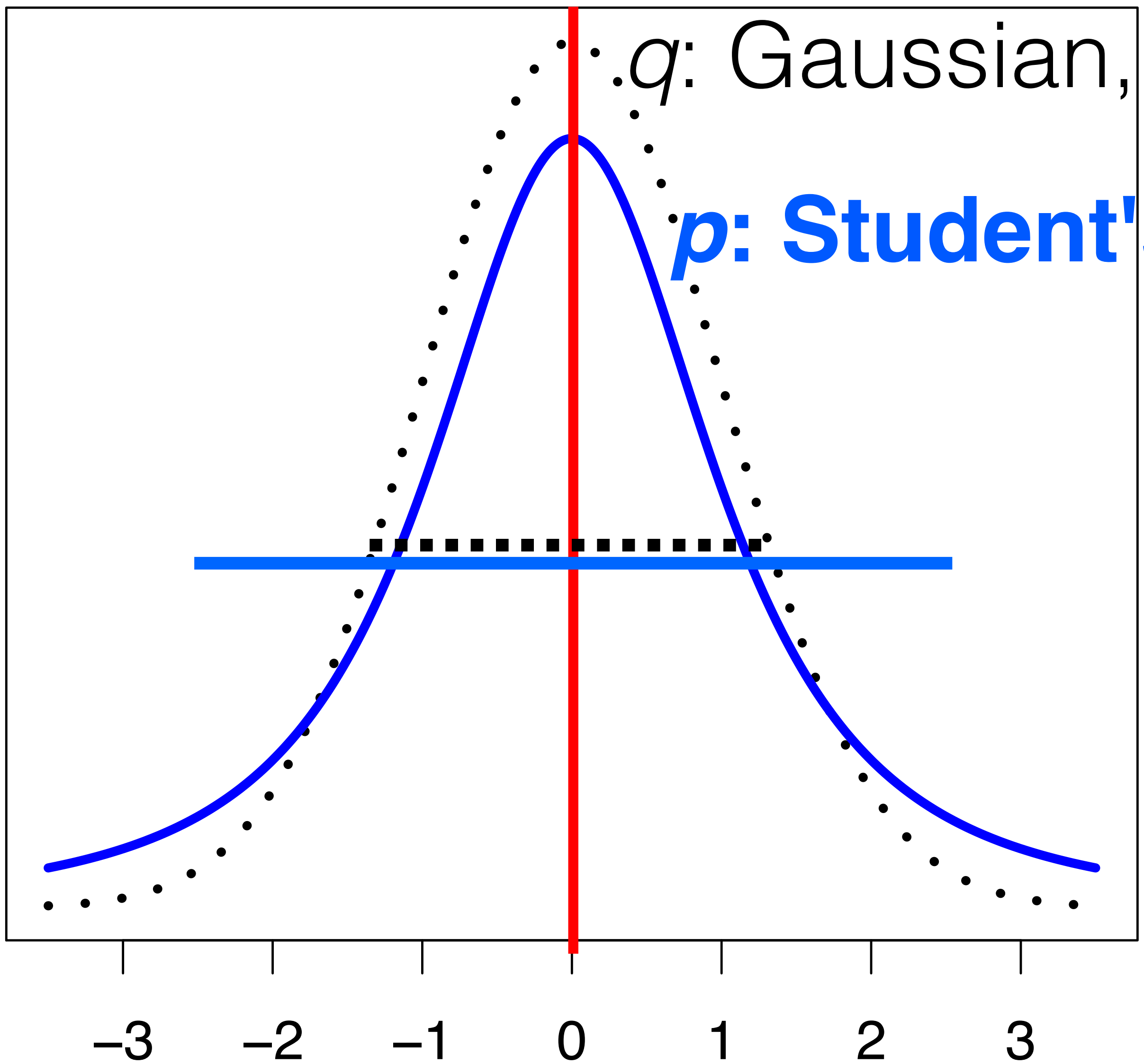
Conjecture (HKCB). For any $c > 1$, there exist zero-mean, unimodal distributions q and p such that

$$KL(q||p) < 0.802 \text{ but also } \sigma_p^2 \geq c\sigma_q^2$$

Can have small KL and arbitrarily bad variance estimate

[Huggins, Karsprzak, Campbell, Broderick 2018]

Is it just MFVB?



Conjecture (HKCB). For any $c > 1$, there exist zero-mean, unimodal distributions q and p such that

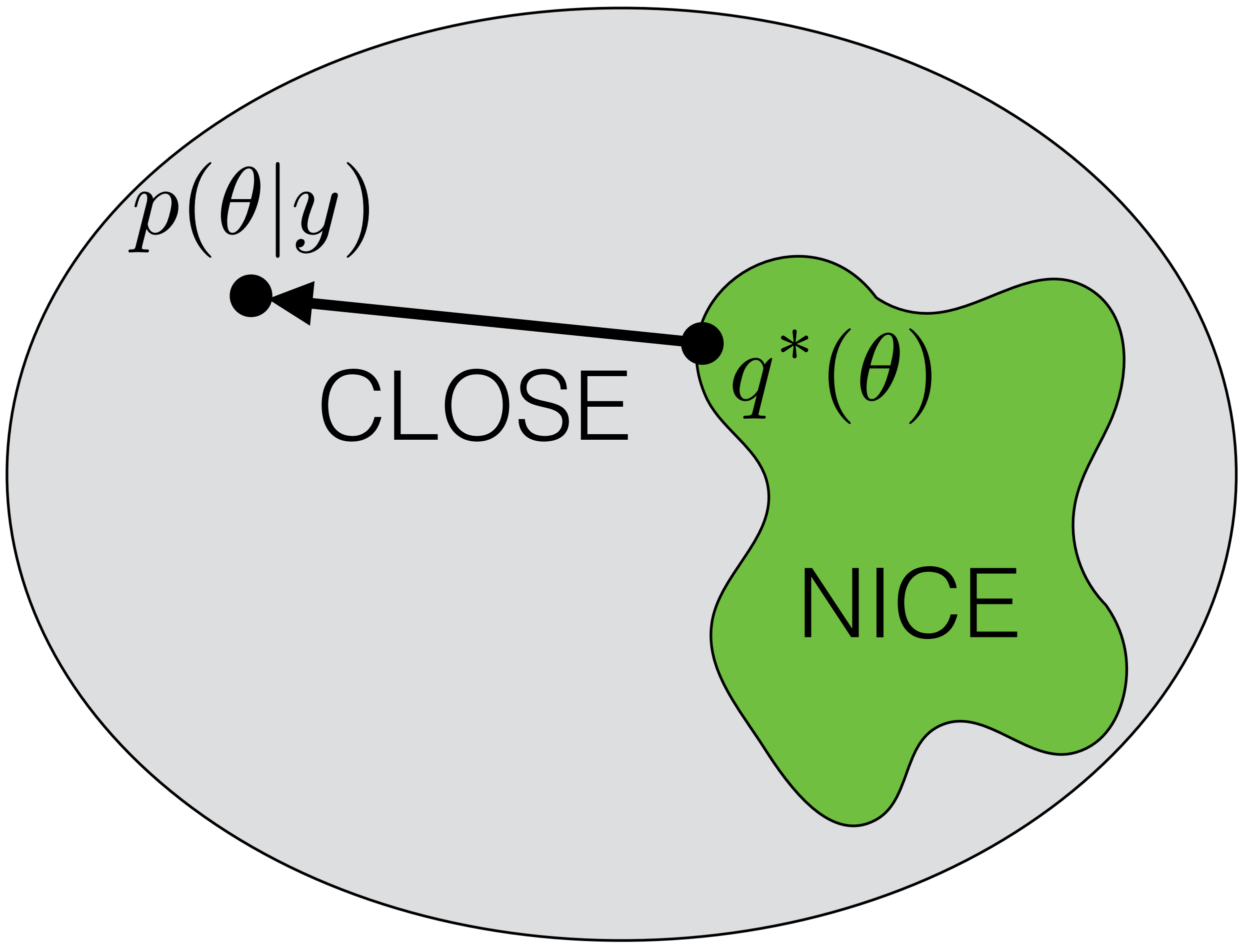
$$KL(q||p) < 0.12 \text{ but also } \sigma_p^2 \geq c\sigma_q^2$$

Can have small KL and arbitrarily bad variance estimate

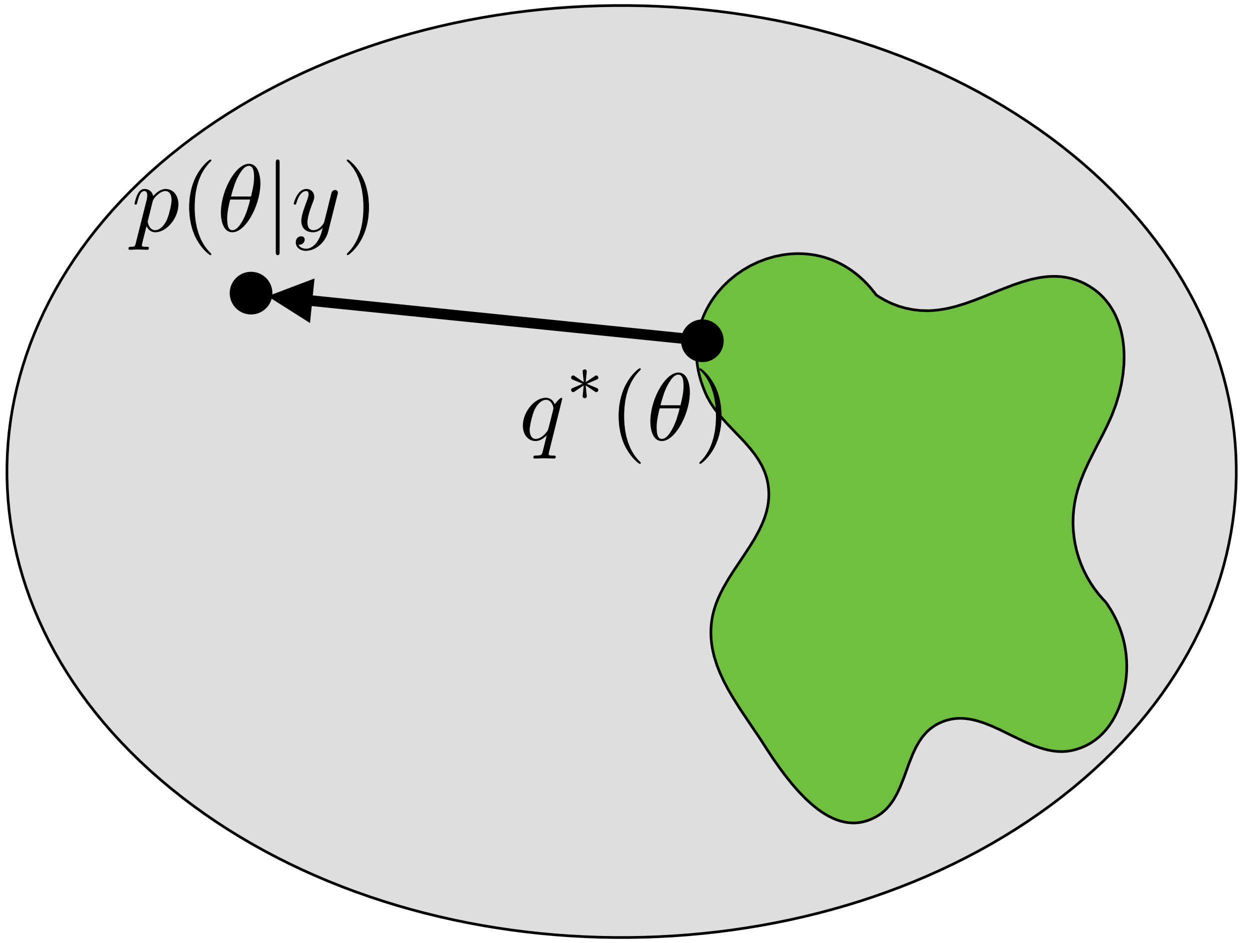
[Huggins, Karsprzak, Campbell, Broderick 2018]

How small is KL in practice?

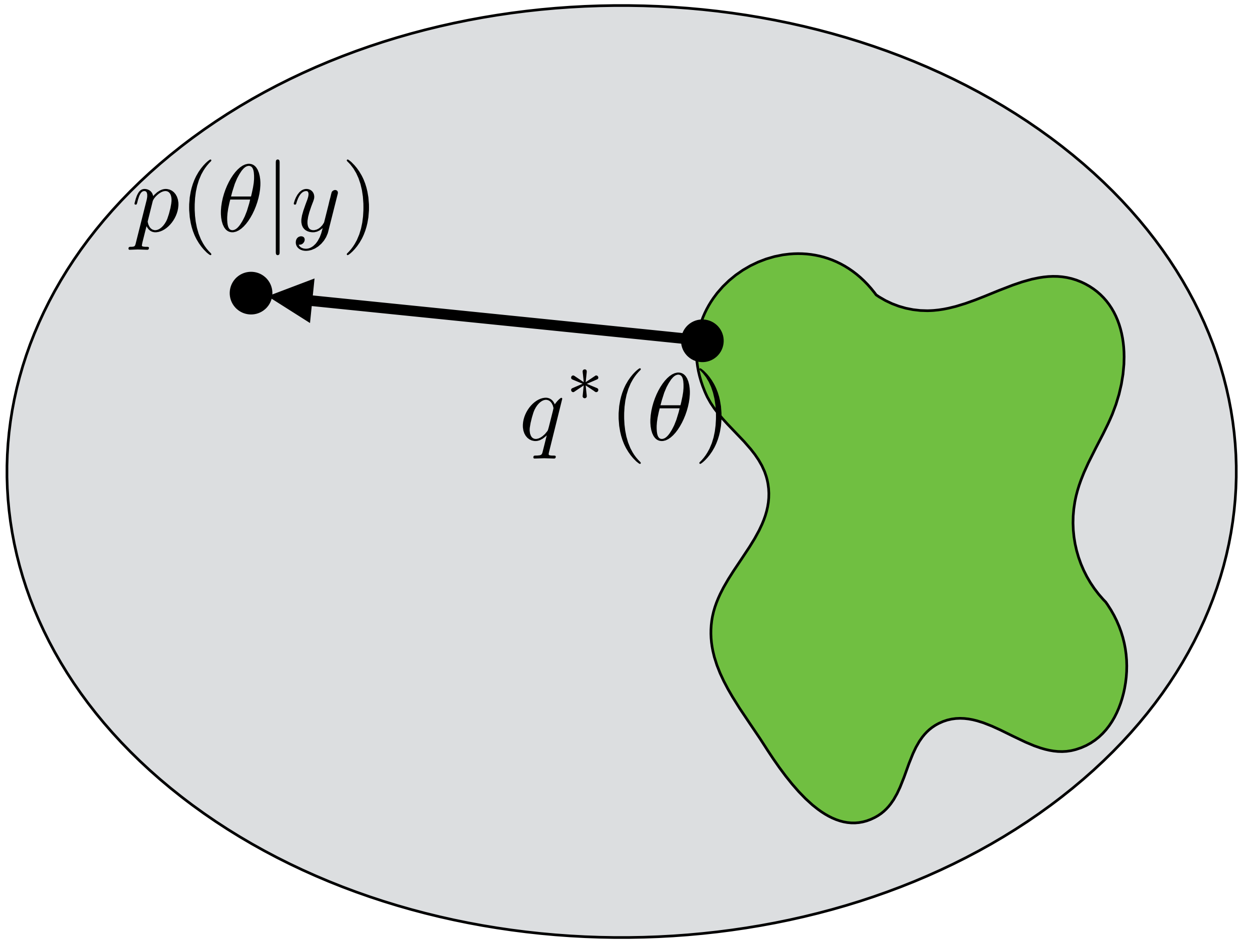
How small is KL in practice?



How small is KL in practice?

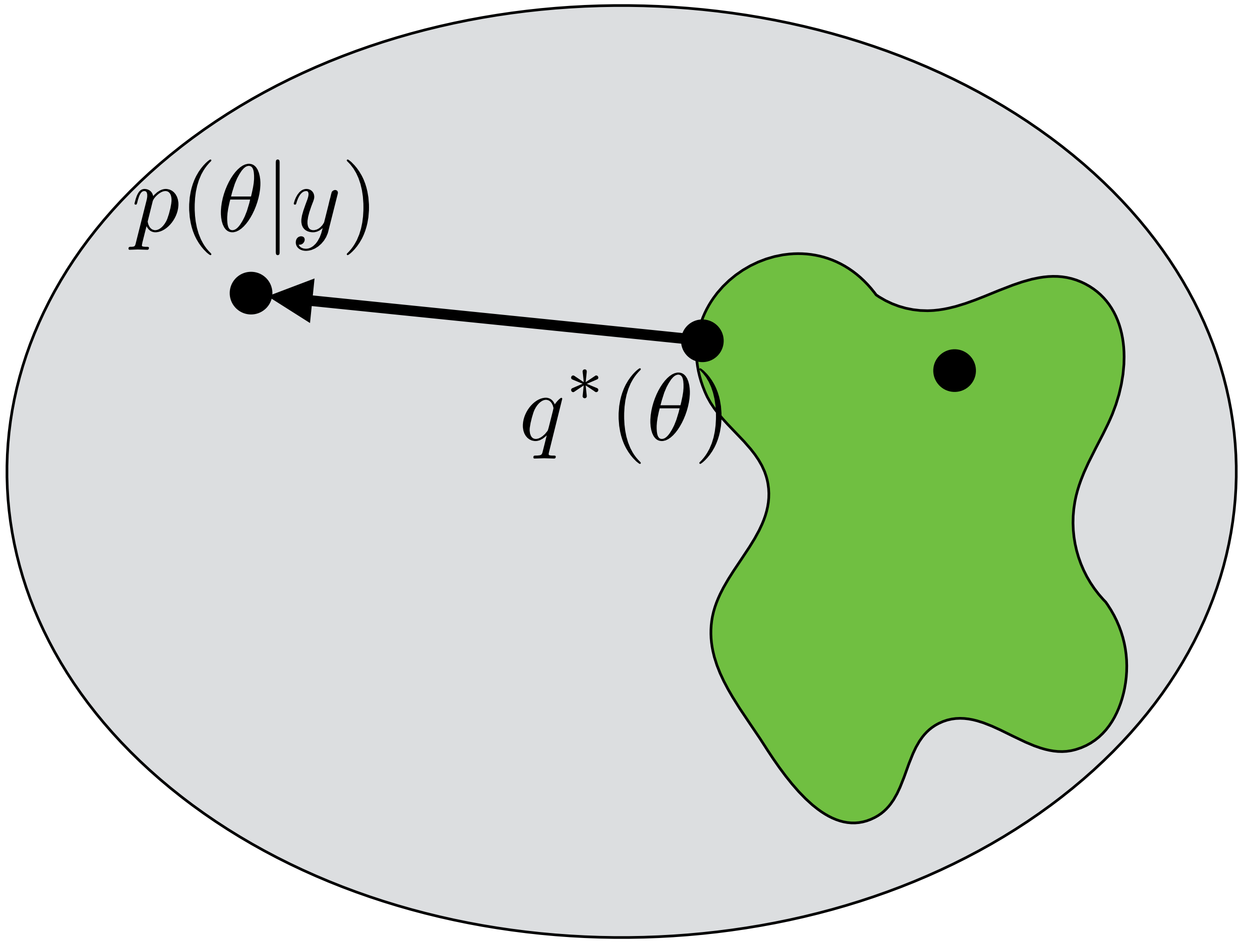


How small is KL in practice?



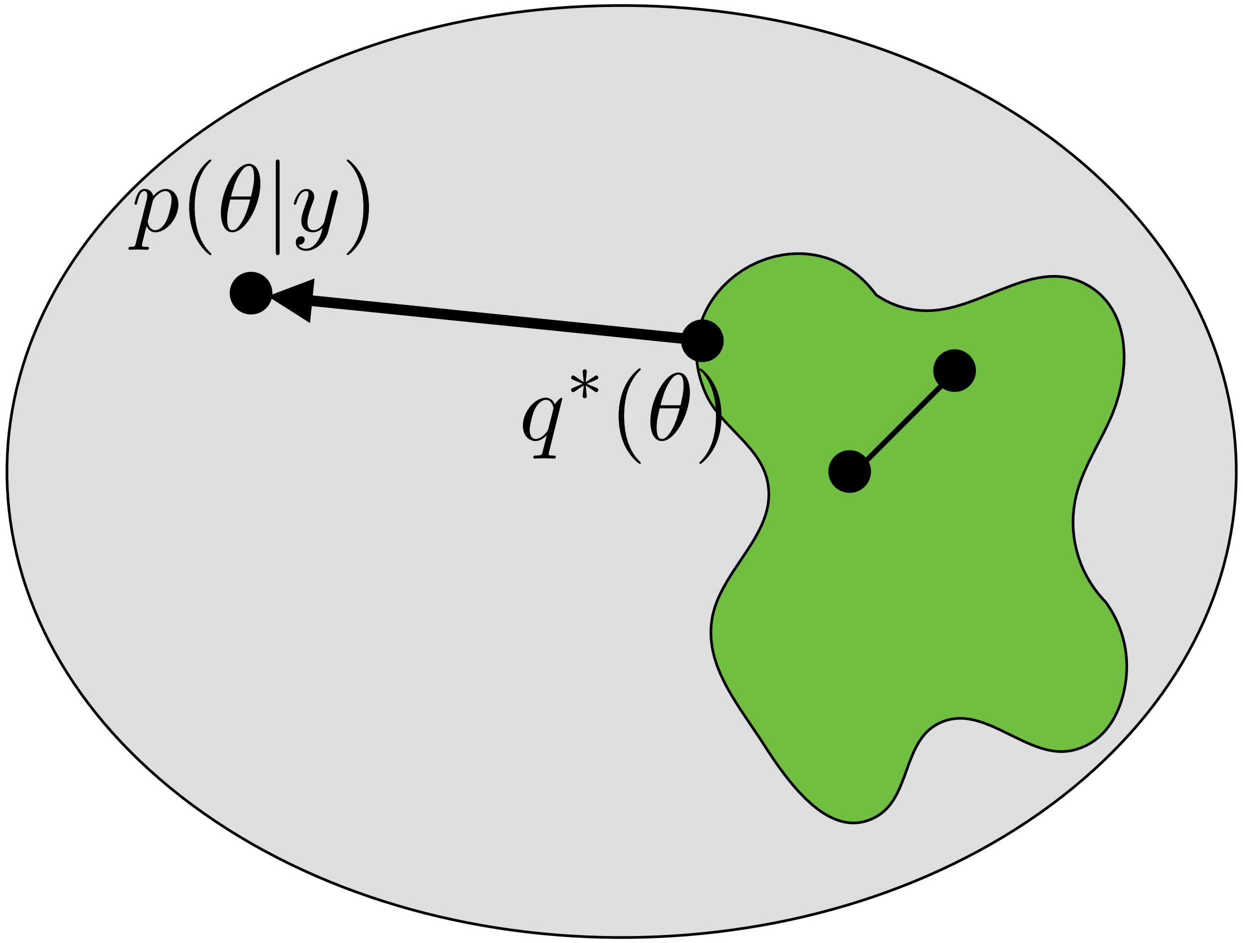
- Often optimum has non-zero KL (MFVB, Gaussian VB)

How small is KL in practice?



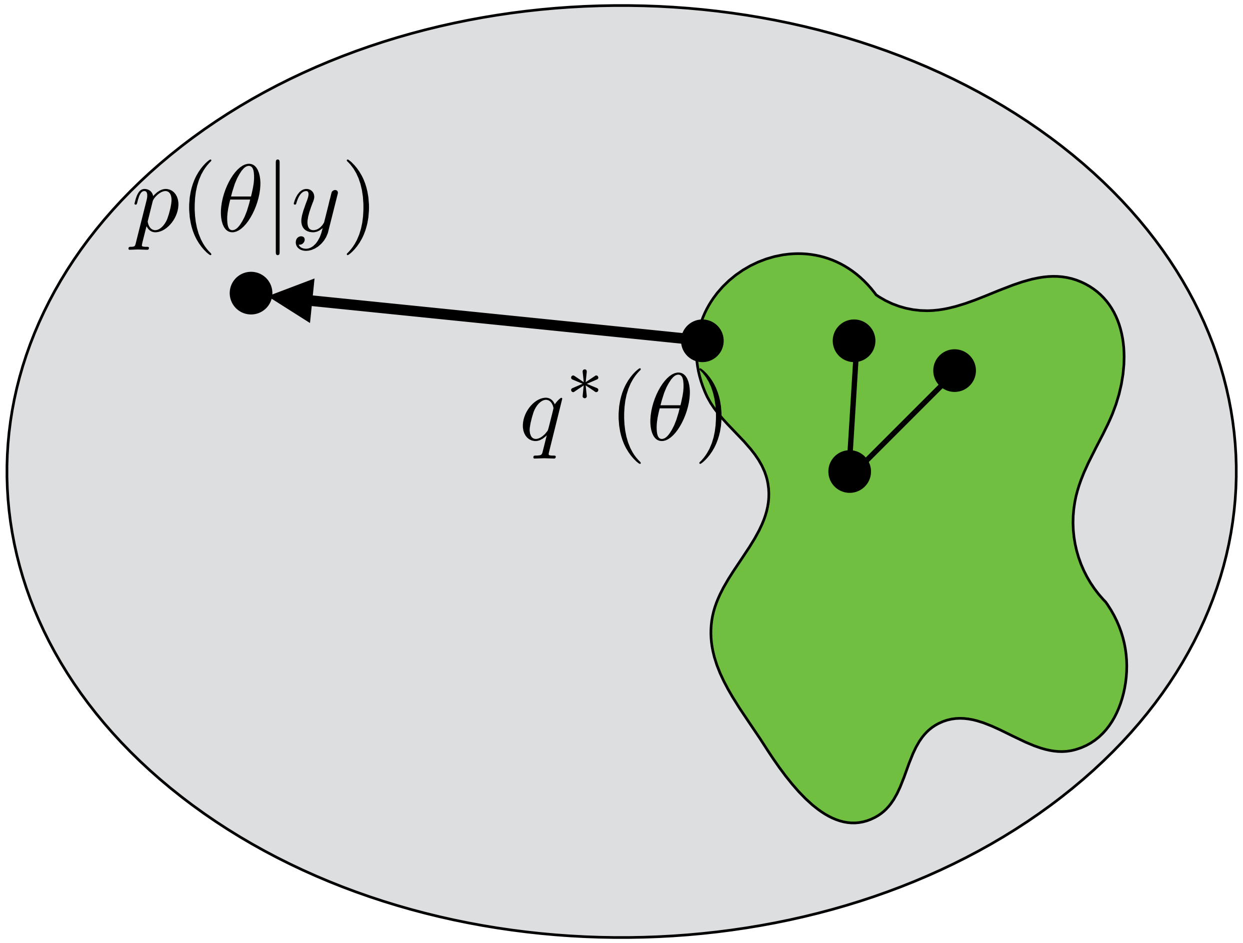
- Often optimum has non-zero KL (MFVB, Gaussian VB)

How small is KL in practice?



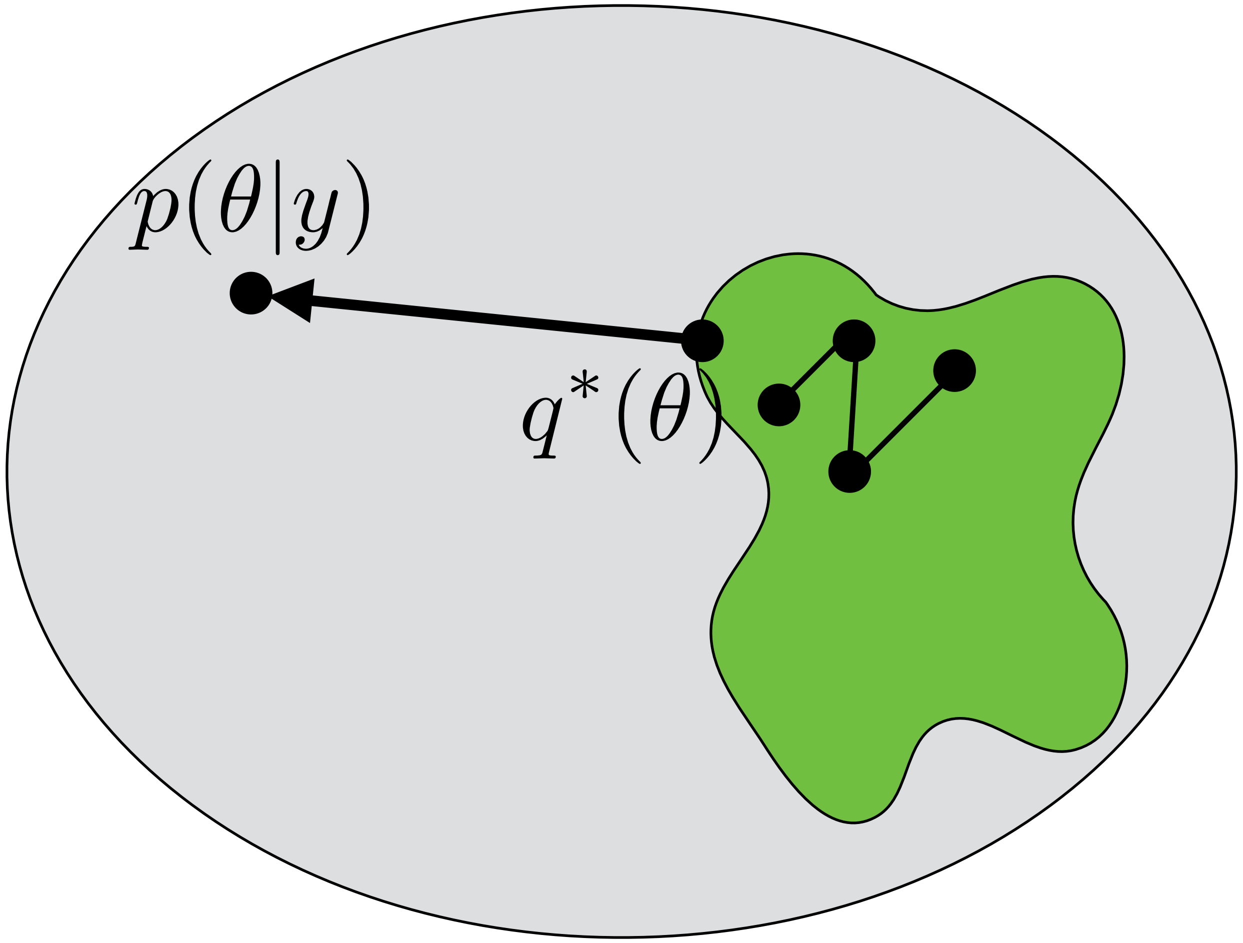
- Often optimum has non-zero KL (MFVB, Gaussian VB)

How small is KL in practice?



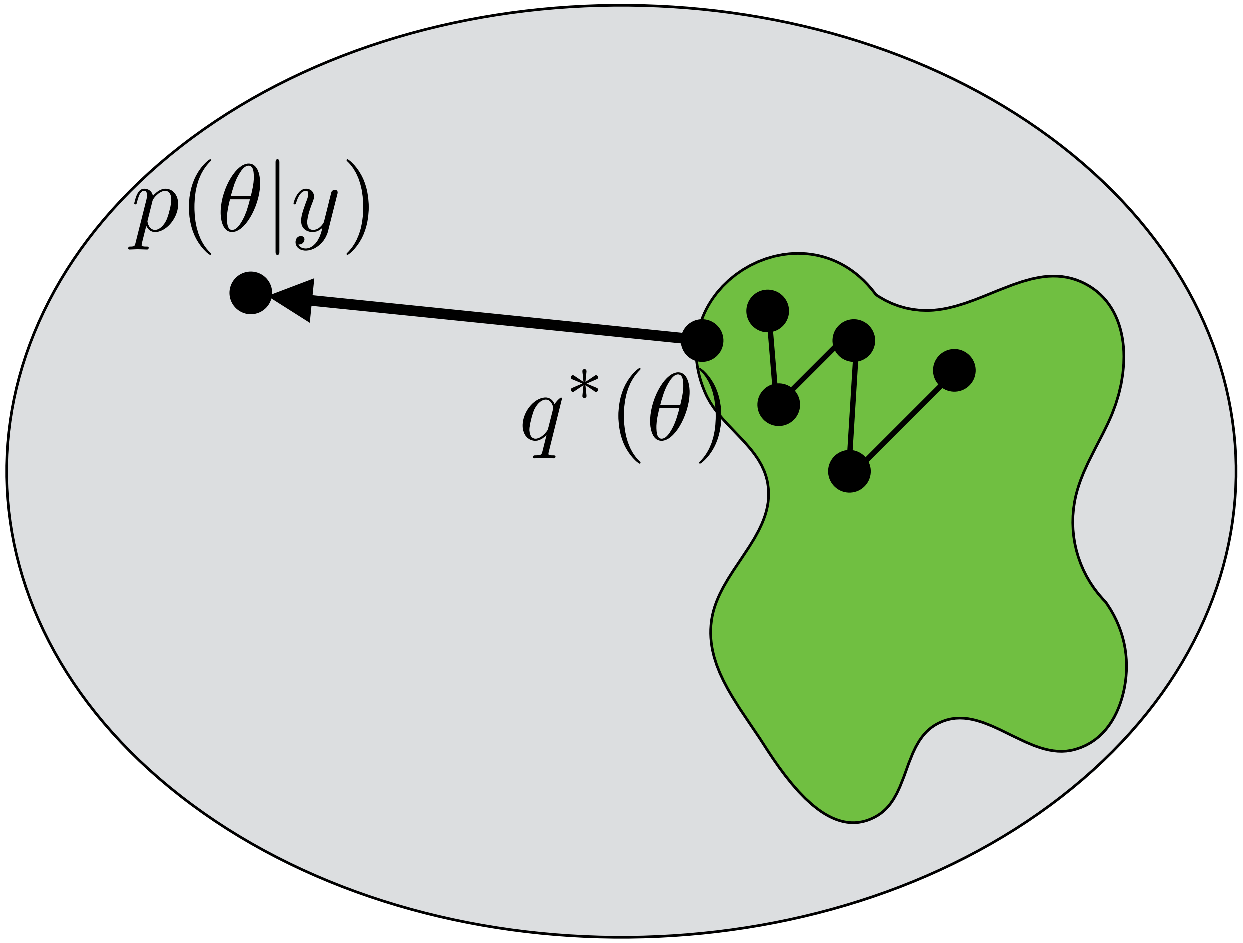
- Often optimum has non-zero KL (MFVB, Gaussian VB)

How small is KL in practice?



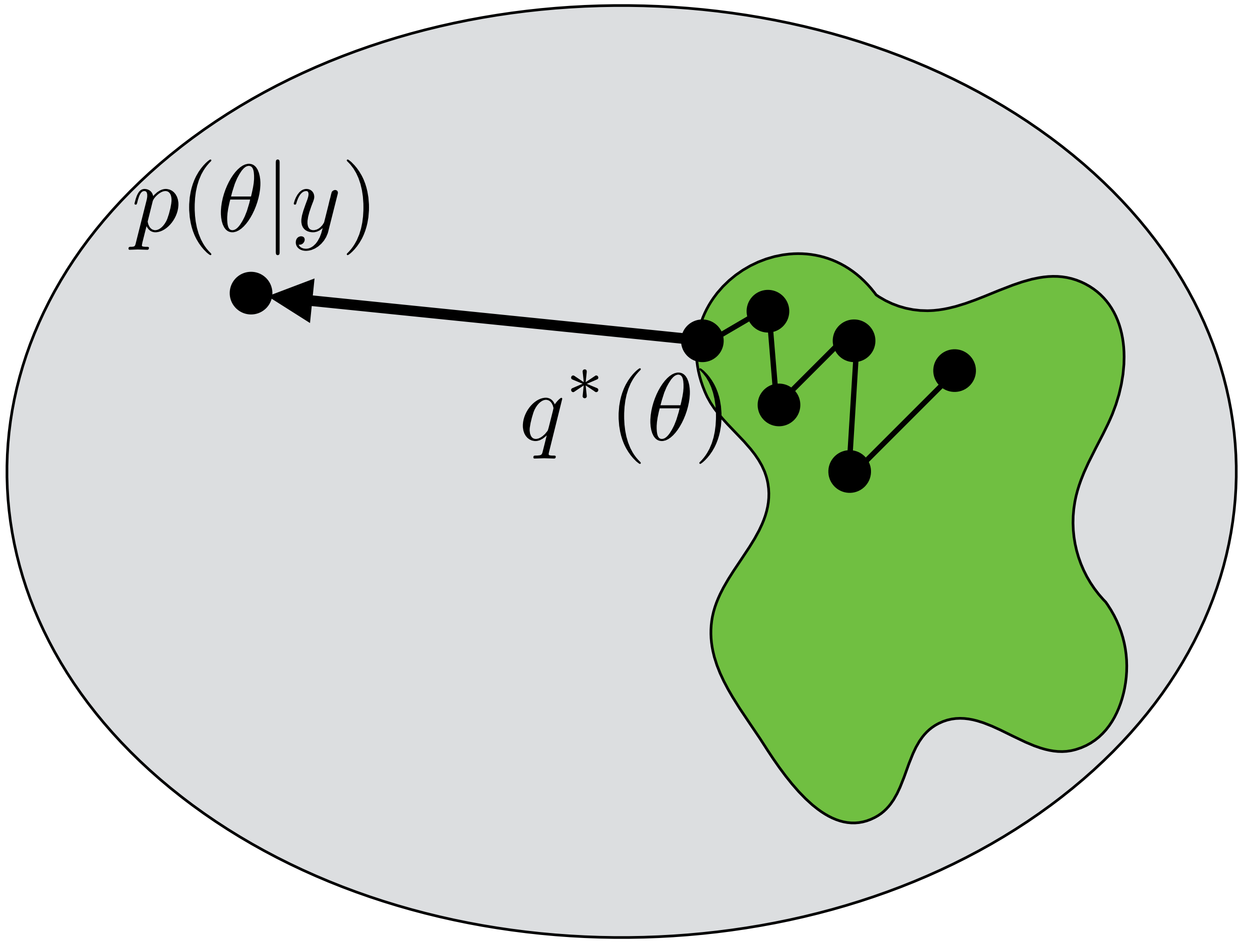
- Often optimum has non-zero KL (MFVB, Gaussian VB)

How small is KL in practice?



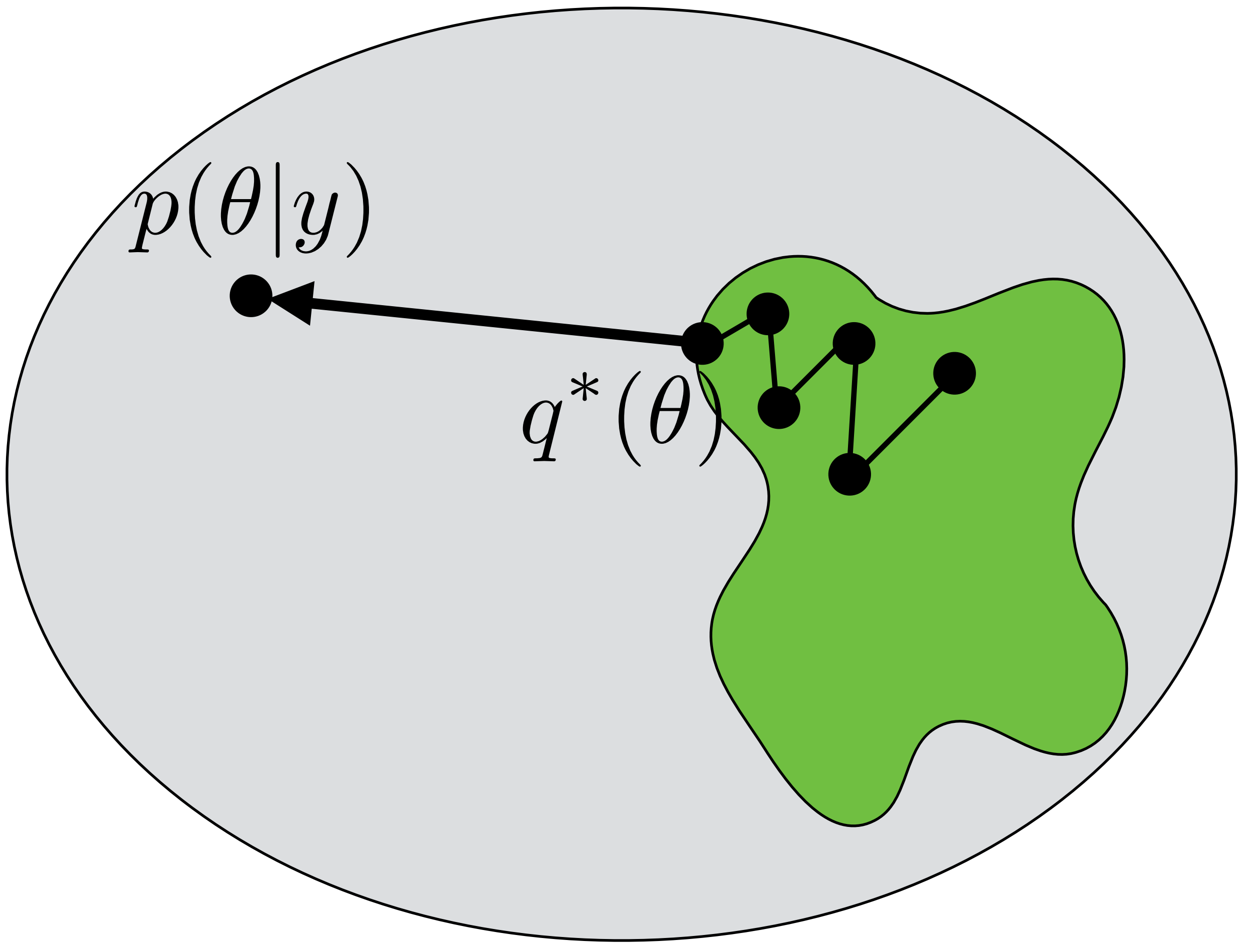
- Often optimum has non-zero KL (MFVB, Gaussian VB)

How small is KL in practice?



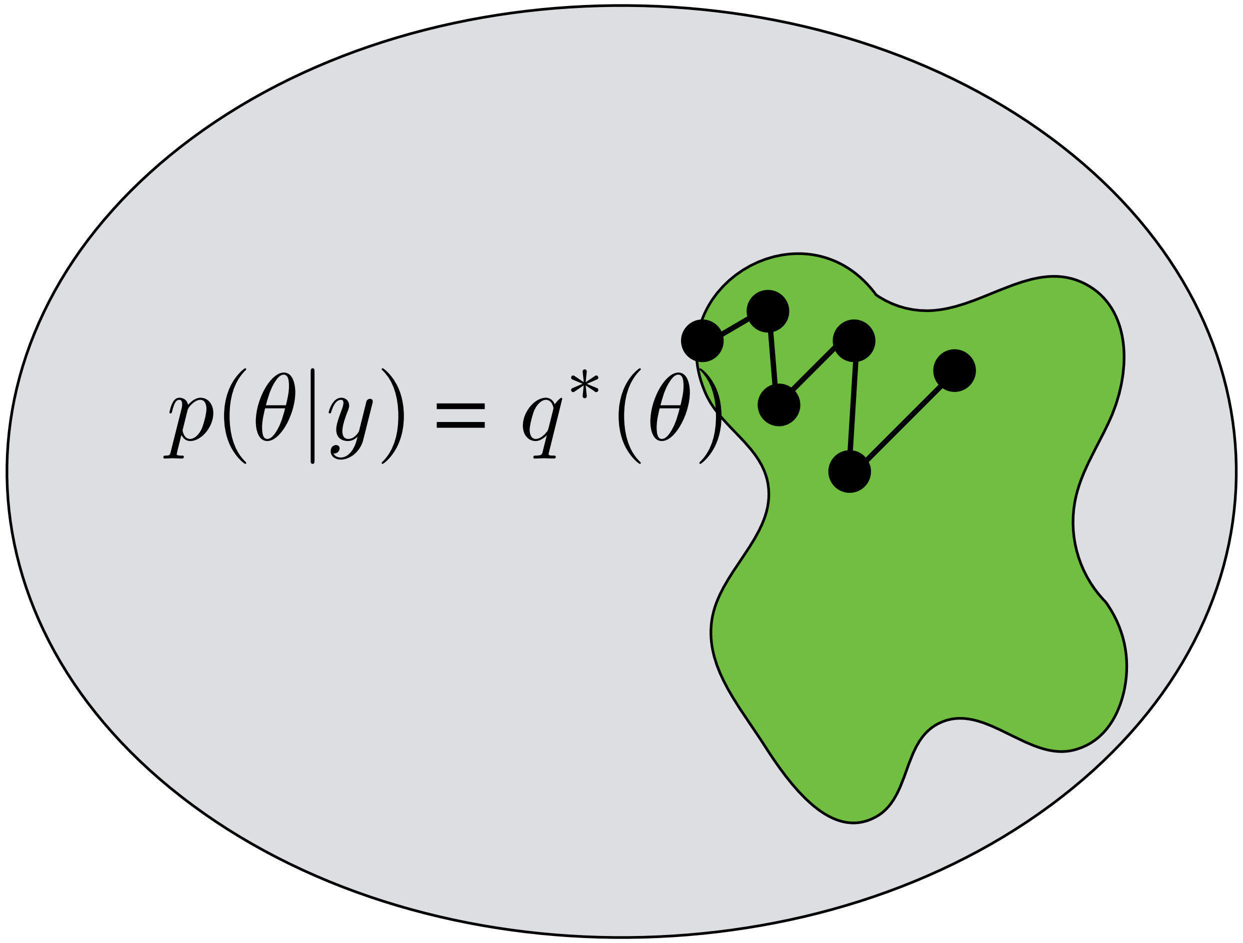
- Often optimum has non-zero KL (MFVB, Gaussian VB)

How small is KL in practice?



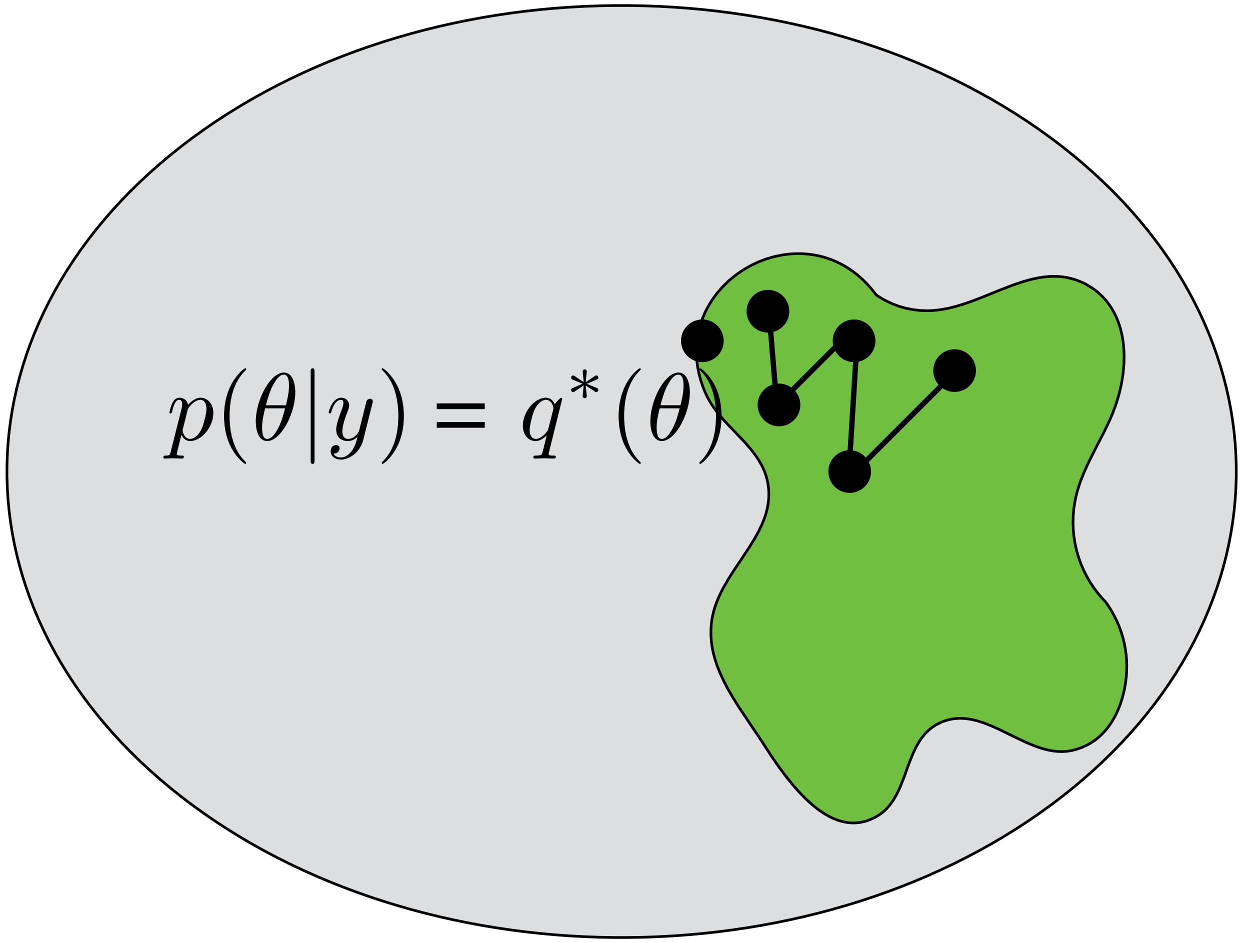
- Often optimum has non-zero KL (MFVB, Gaussian VB)
- Even if optimum has zero KL, algorithm might not reach zero

How small is KL in practice?



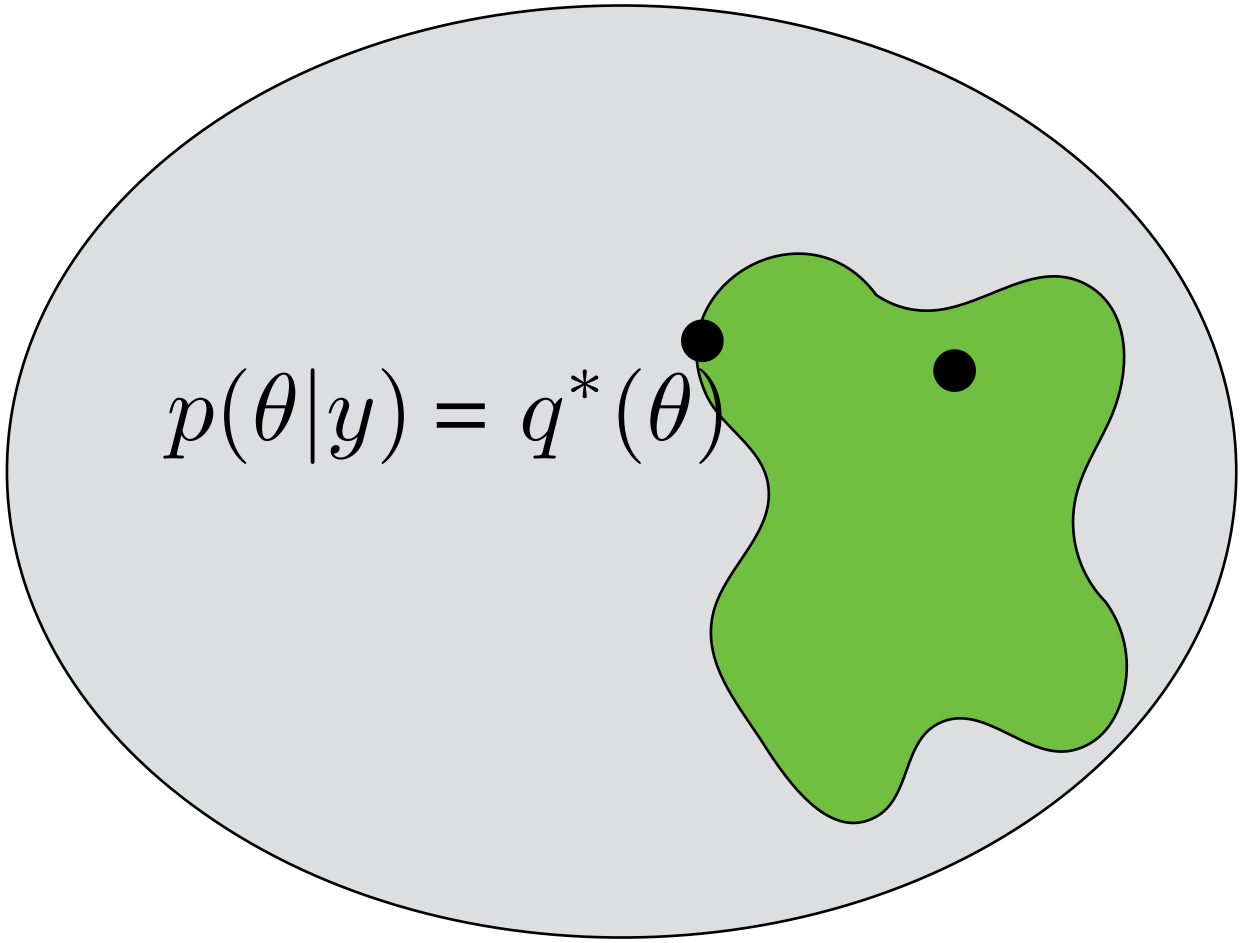
- Often optimum has non-zero KL (MFVB, Gaussian VB)
- Even if optimum has zero KL, algorithm might not reach zero

How small is KL in practice?



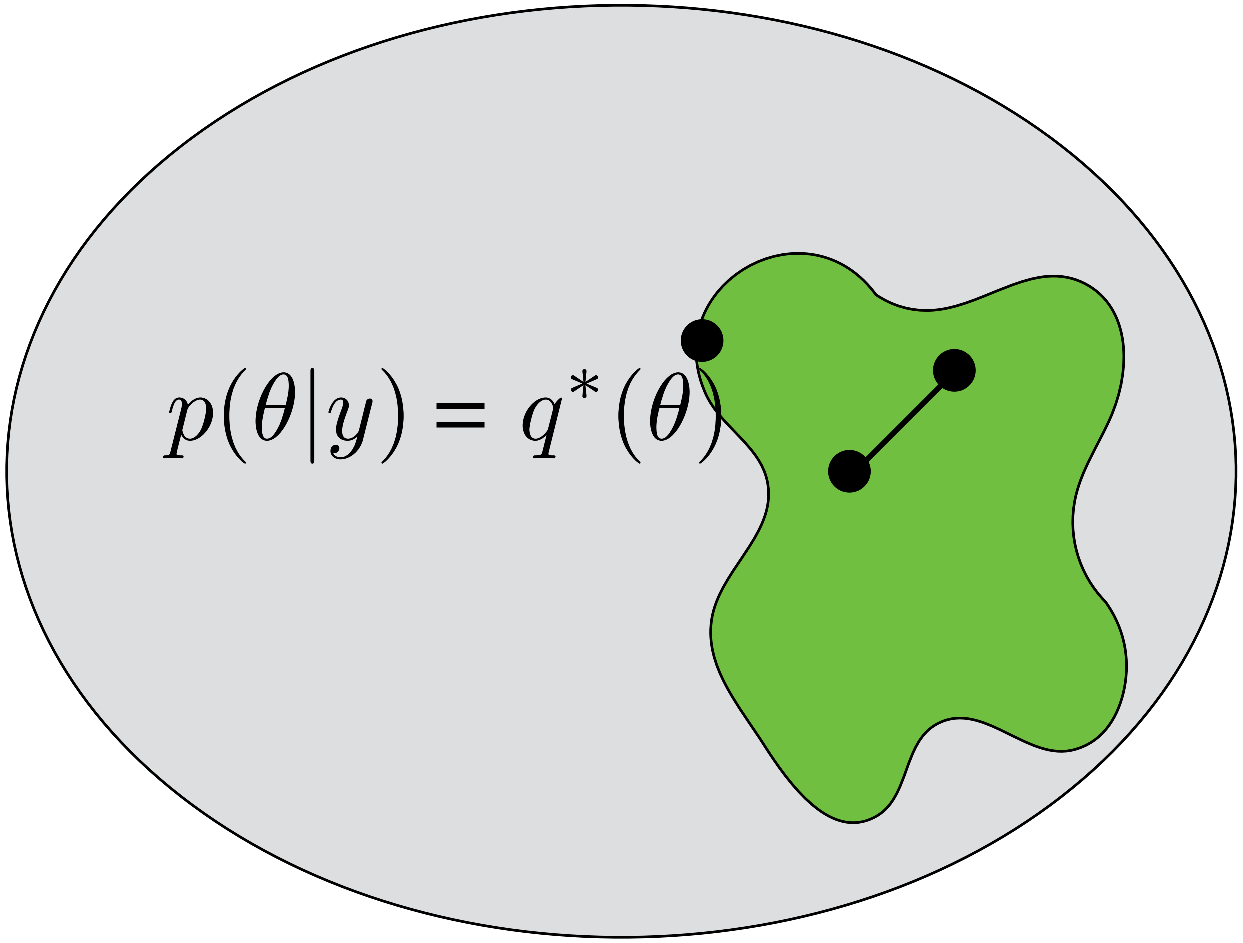
- Often optimum has non-zero KL (MFVB, Gaussian VB)
- Even if optimum has zero KL, algorithm might not reach zero

How small is KL in practice?



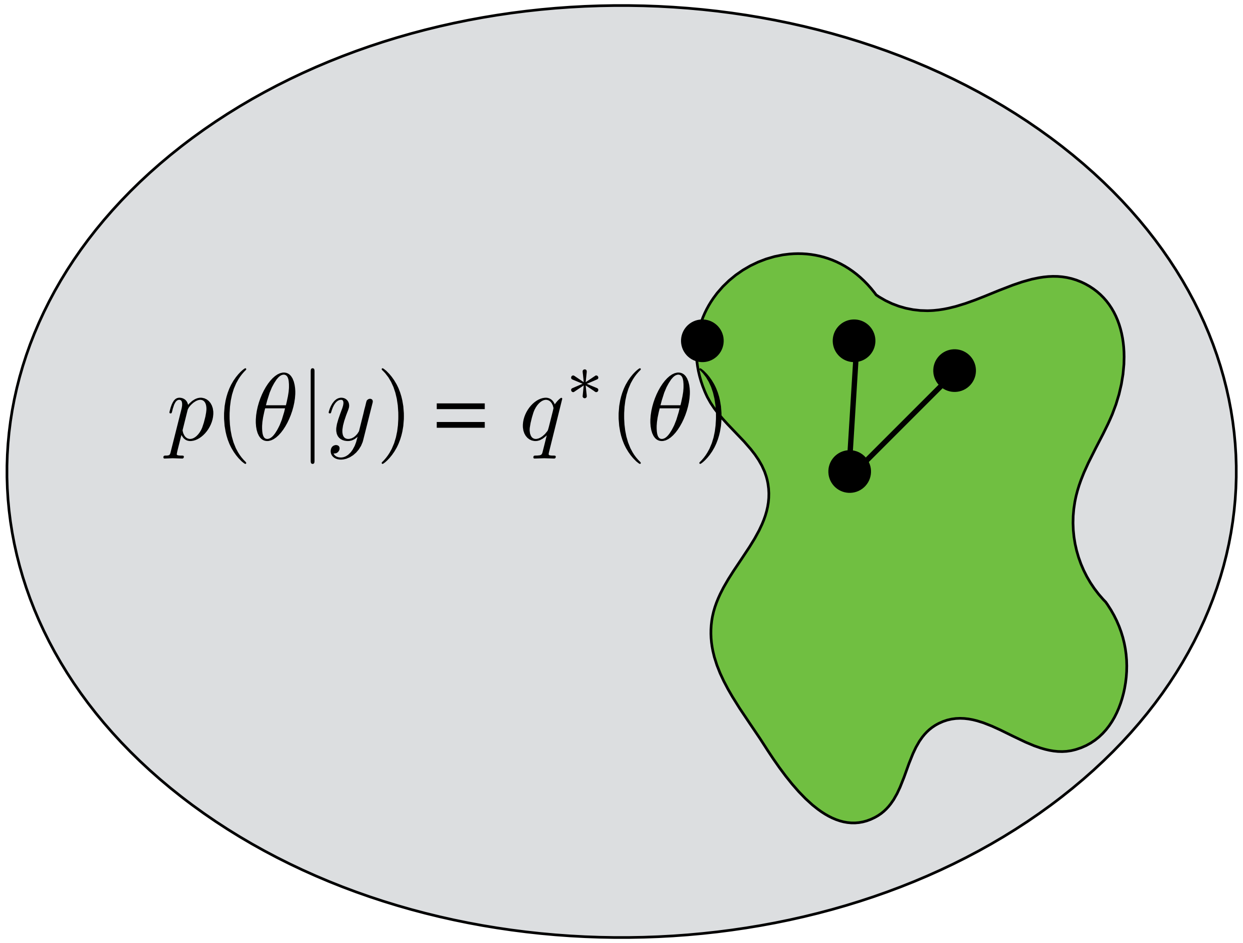
- Often optimum has non-zero KL (MFVB, Gaussian VB)
- Even if optimum has zero KL, algorithm might not reach zero

How small is KL in practice?



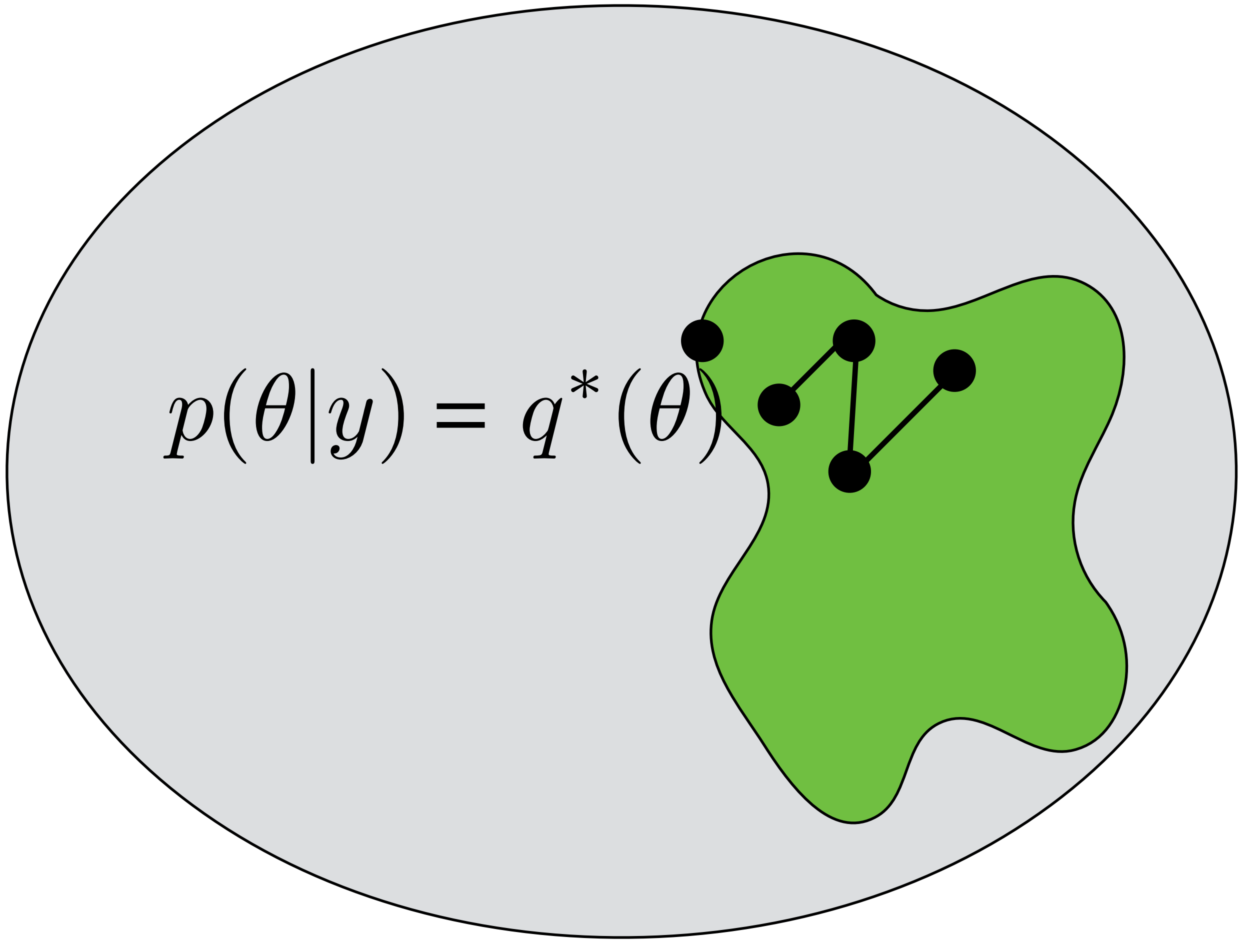
- Often optimum has non-zero KL (MFVB, Gaussian VB)
- Even if optimum has zero KL, algorithm might not reach zero

How small is KL in practice?



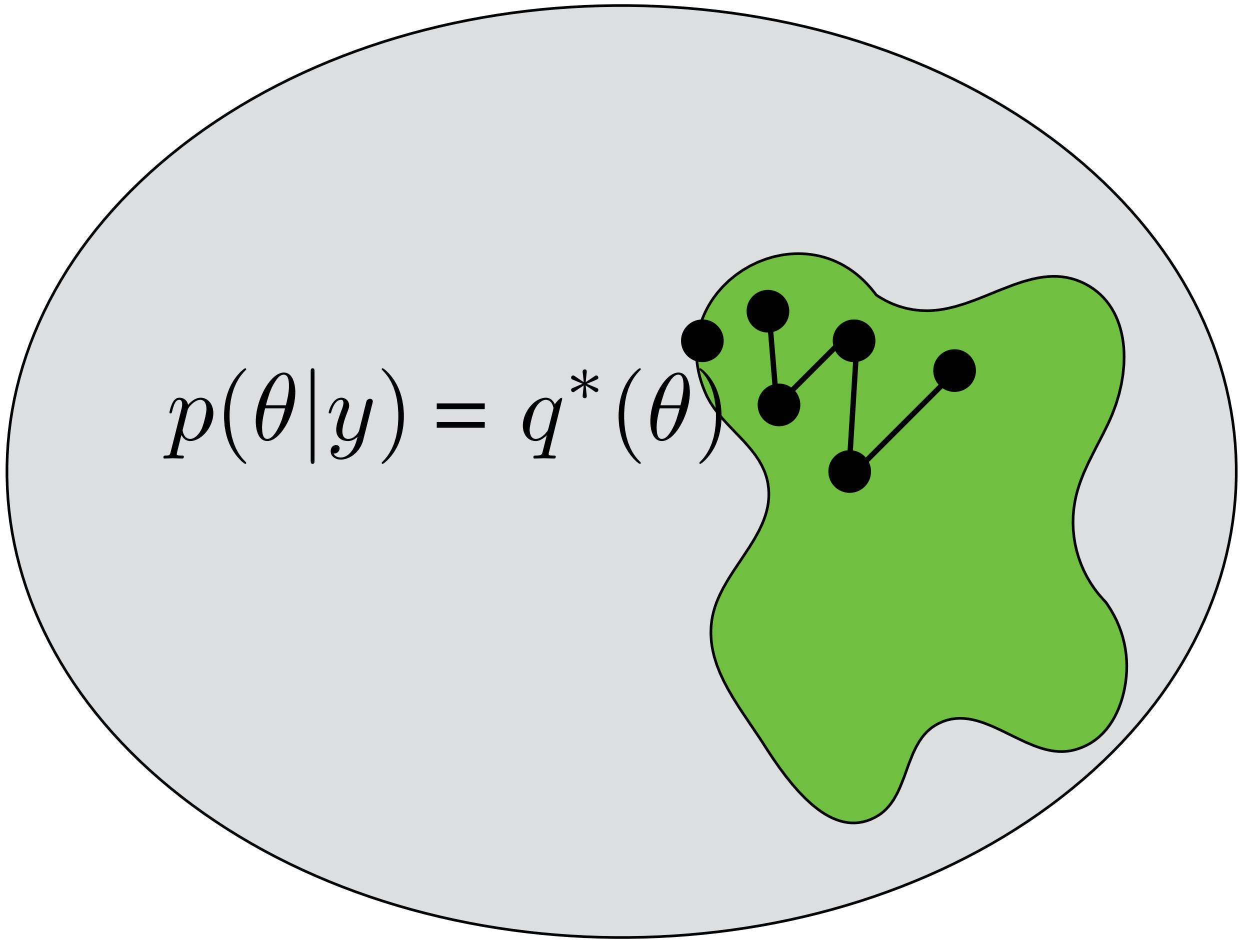
- Often optimum has non-zero KL (MFVB, Gaussian VB)
- Even if optimum has zero KL, algorithm might not reach zero

How small is KL in practice?



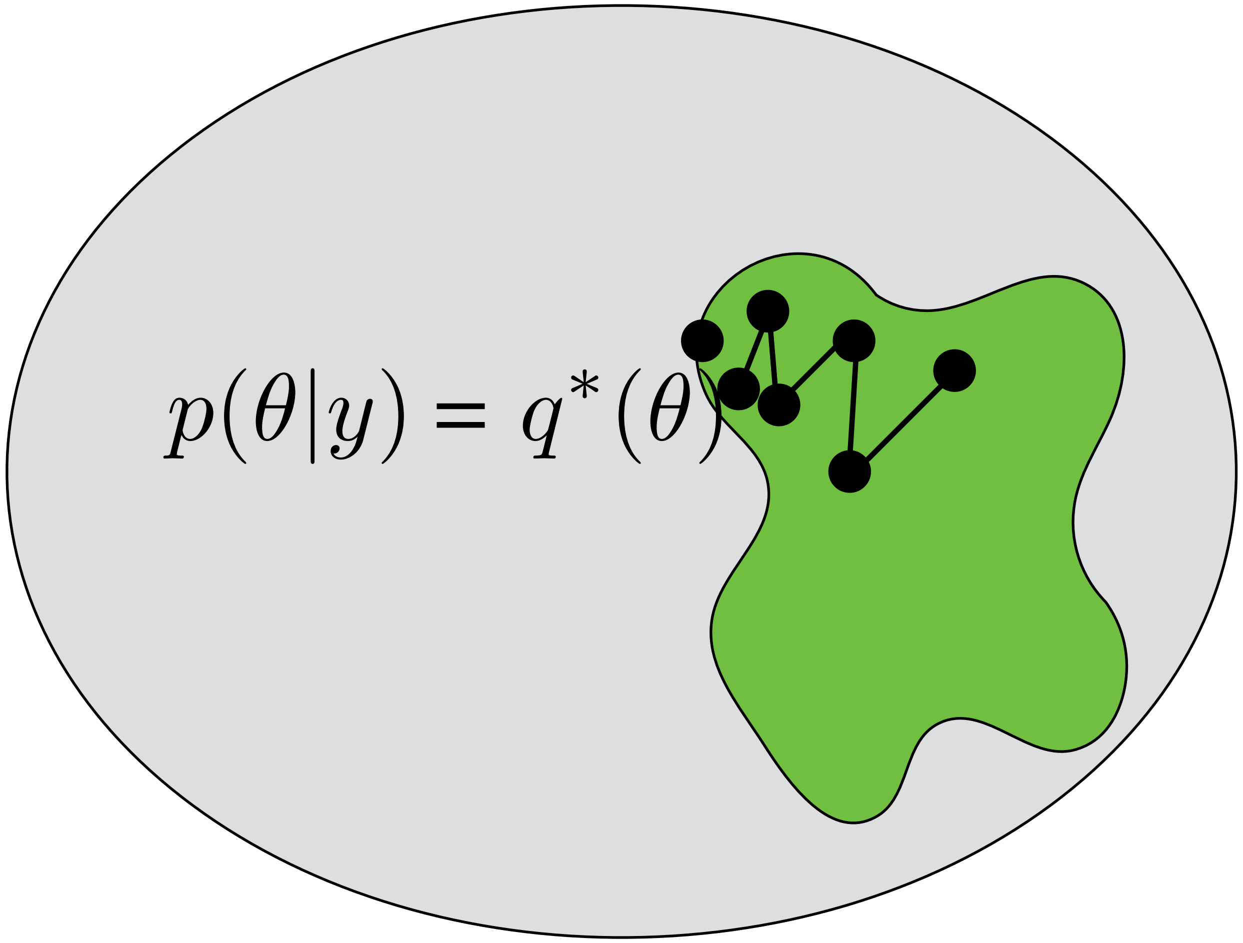
- Often optimum has non-zero KL (MFVB, Gaussian VB)
- Even if optimum has zero KL, algorithm might not reach zero

How small is KL in practice?



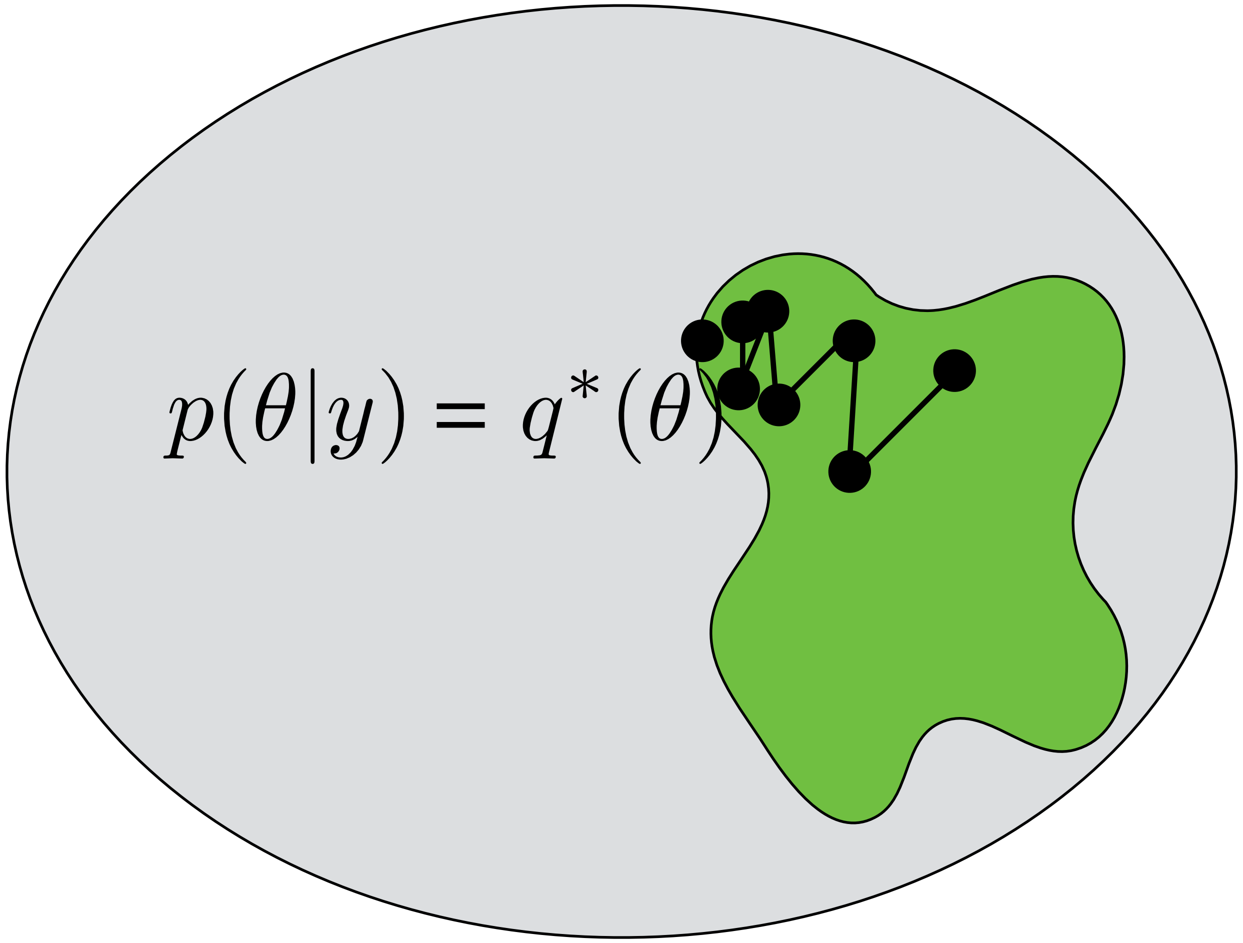
- Often optimum has non-zero KL (MFVB, Gaussian VB)
- Even if optimum has zero KL, algorithm might not reach zero

How small is KL in practice?



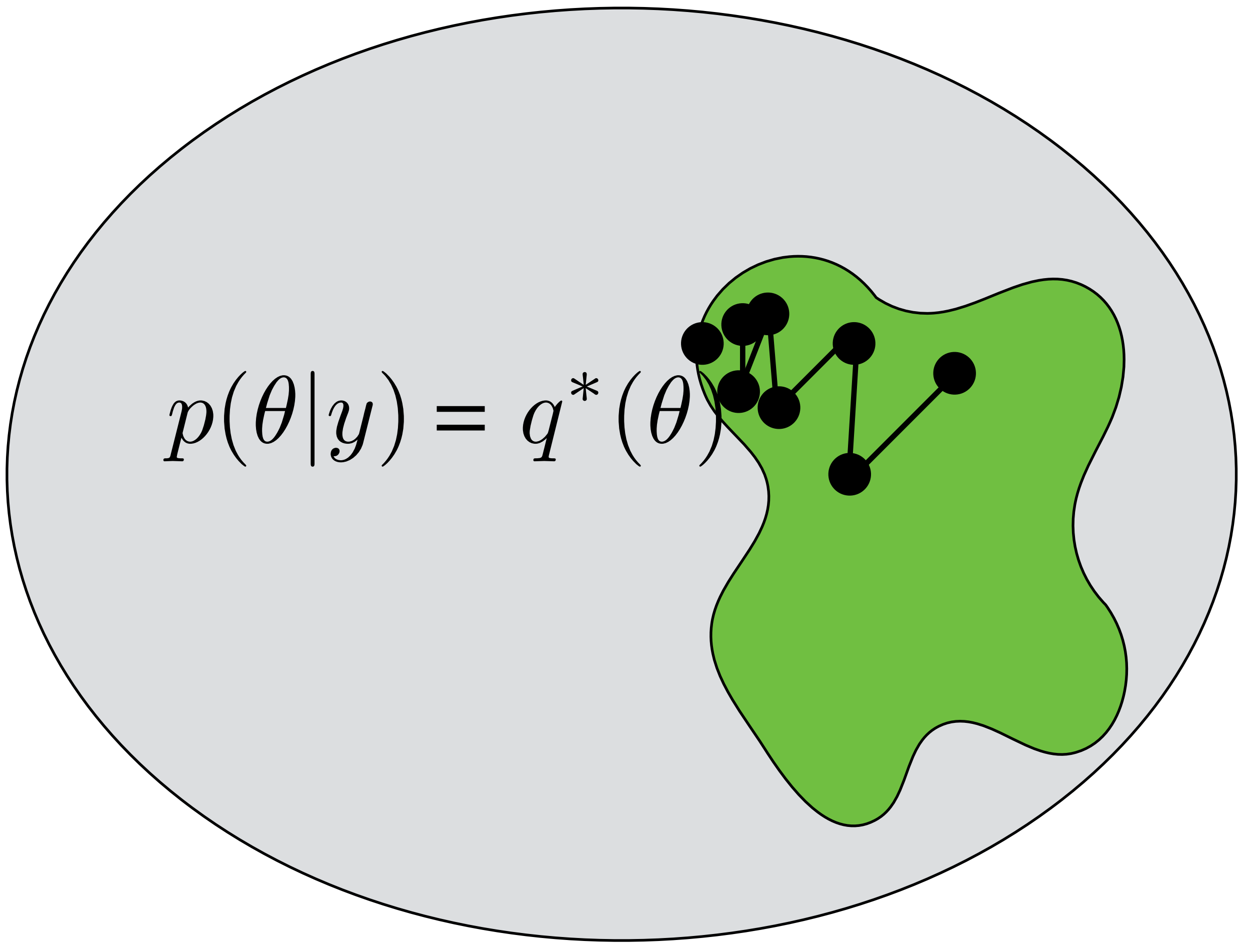
- Often optimum has non-zero KL (MFVB, Gaussian VB)
- Even if optimum has zero KL, algorithm might not reach zero

How small is KL in practice?



- Often optimum has non-zero KL (MFVB, Gaussian VB)
- Even if optimum has zero KL, algorithm might not reach zero

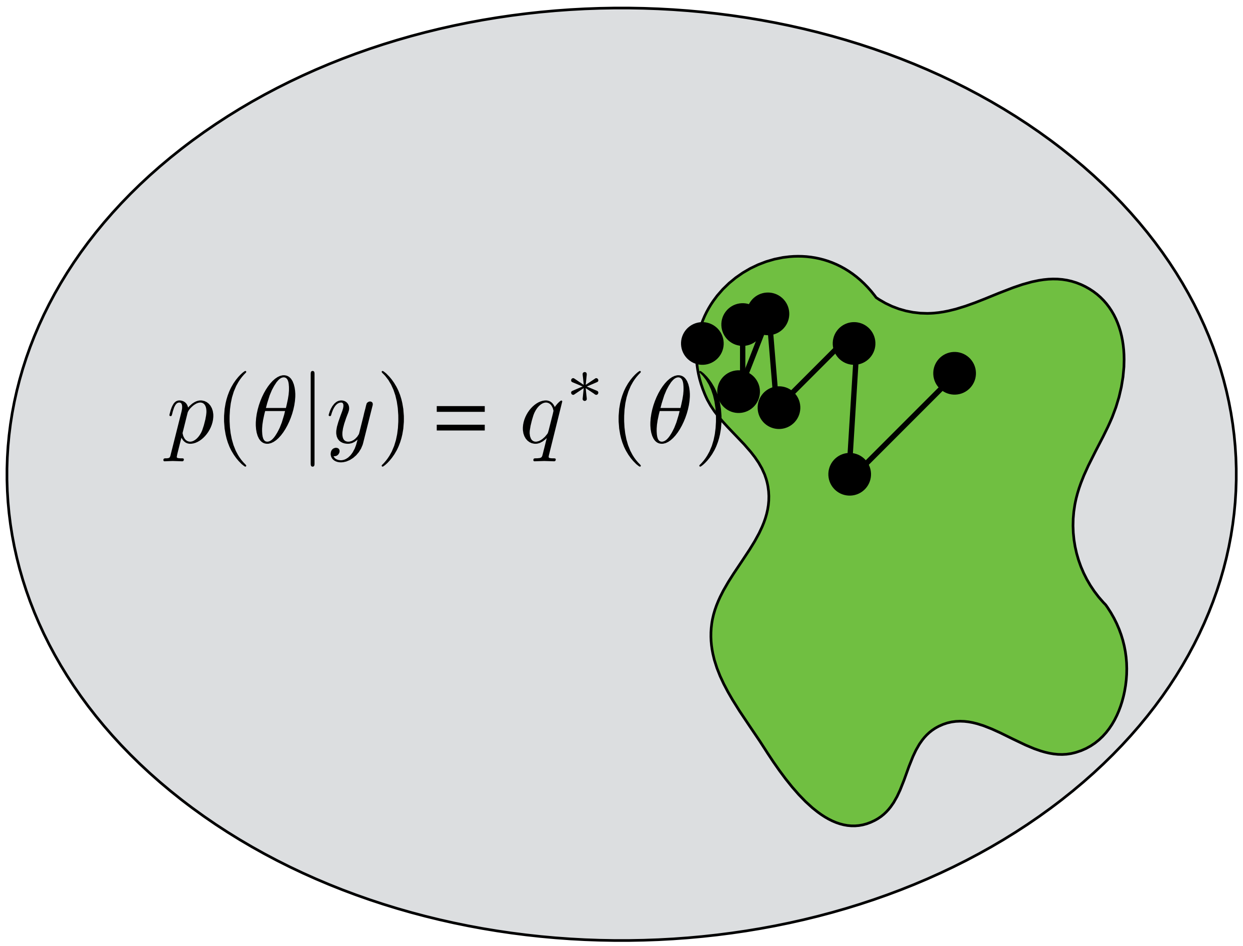
How small is KL in practice?



- Often optimum has non-zero KL (MFVB, Gaussian VB)
- Even if optimum has zero KL, algorithm might not reach zero

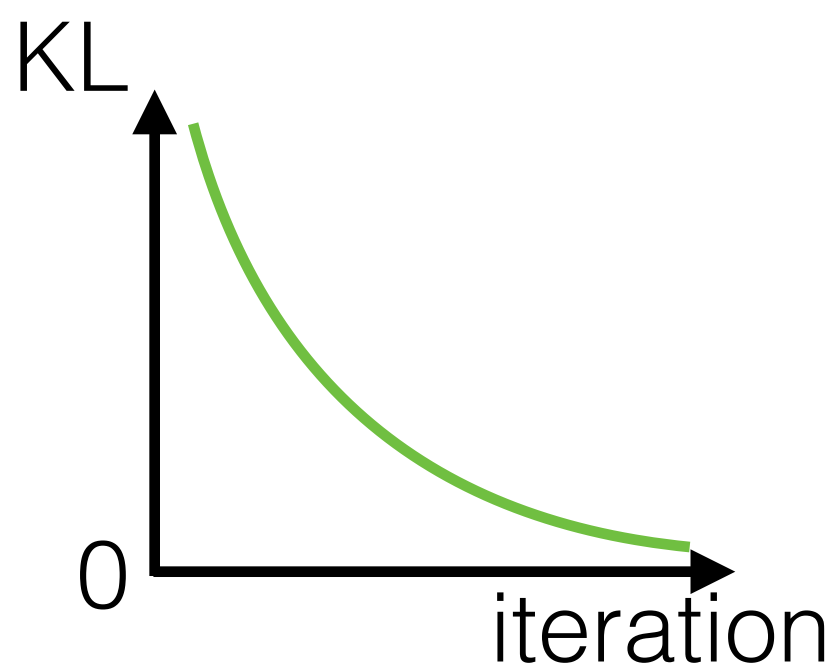
- Recall: the KL value isn't free

How small is KL in practice?

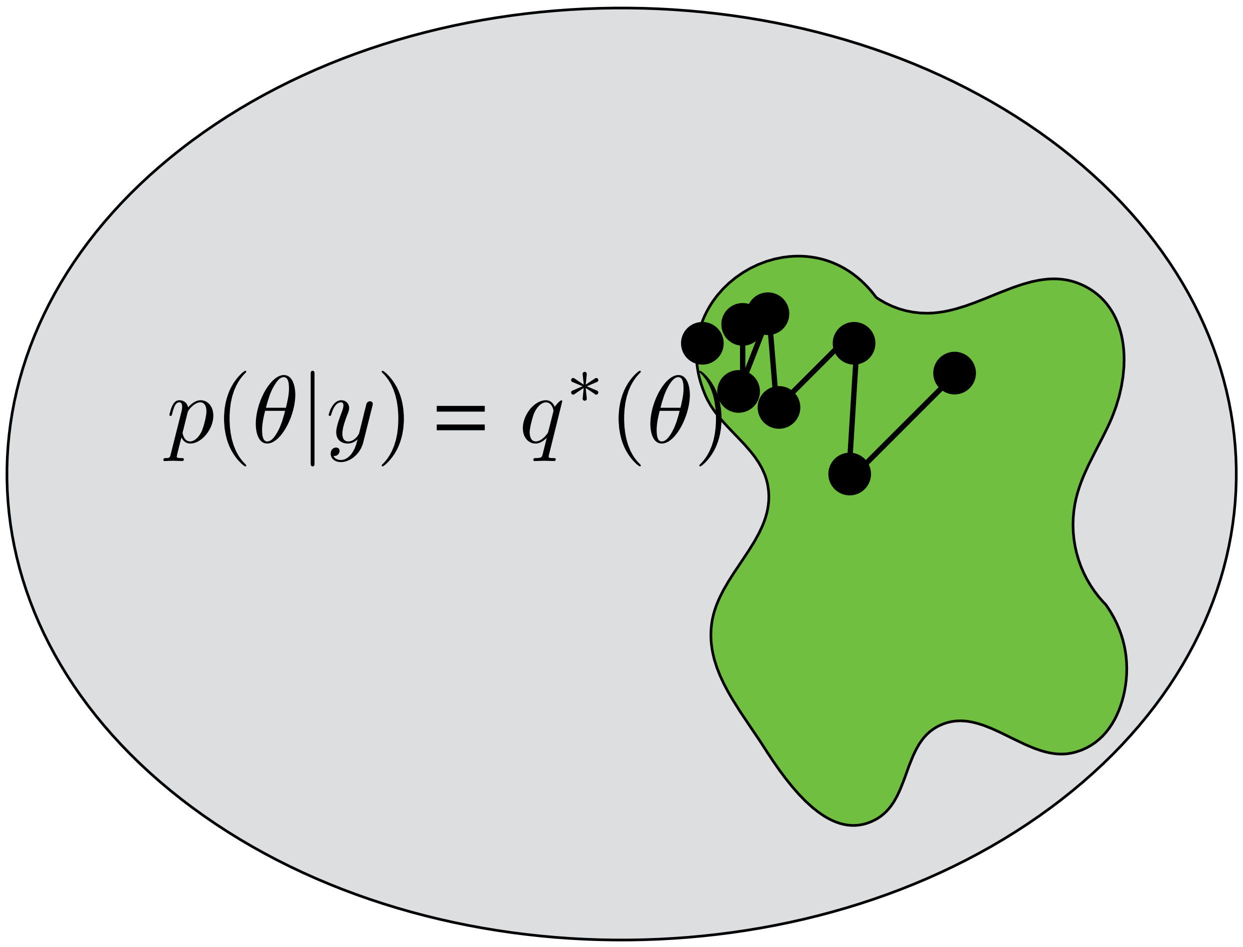


- Often optimum has non-zero KL (MFVB, Gaussian VB)
- Even if optimum has zero KL, algorithm might not reach zero

- Recall: the KL value isn't free

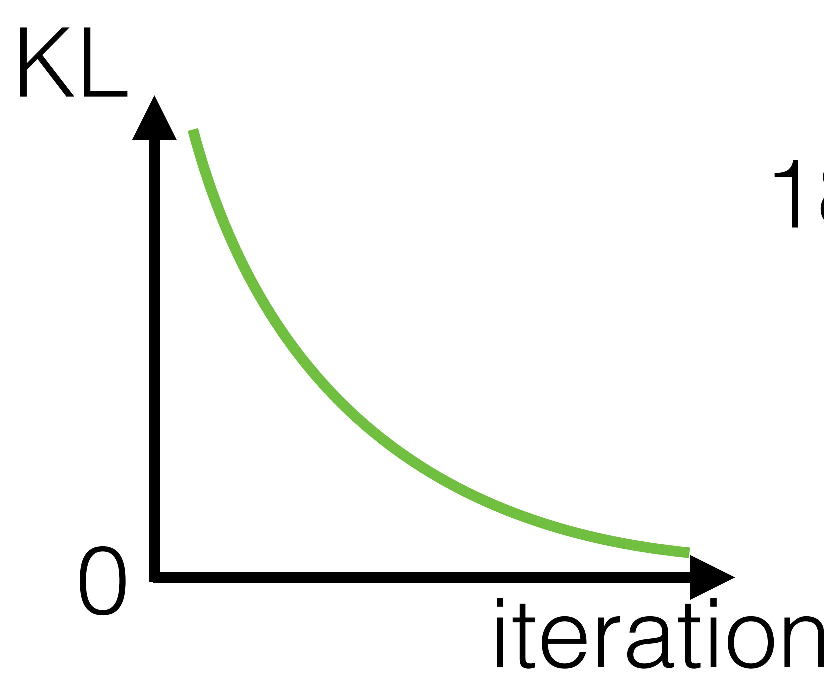


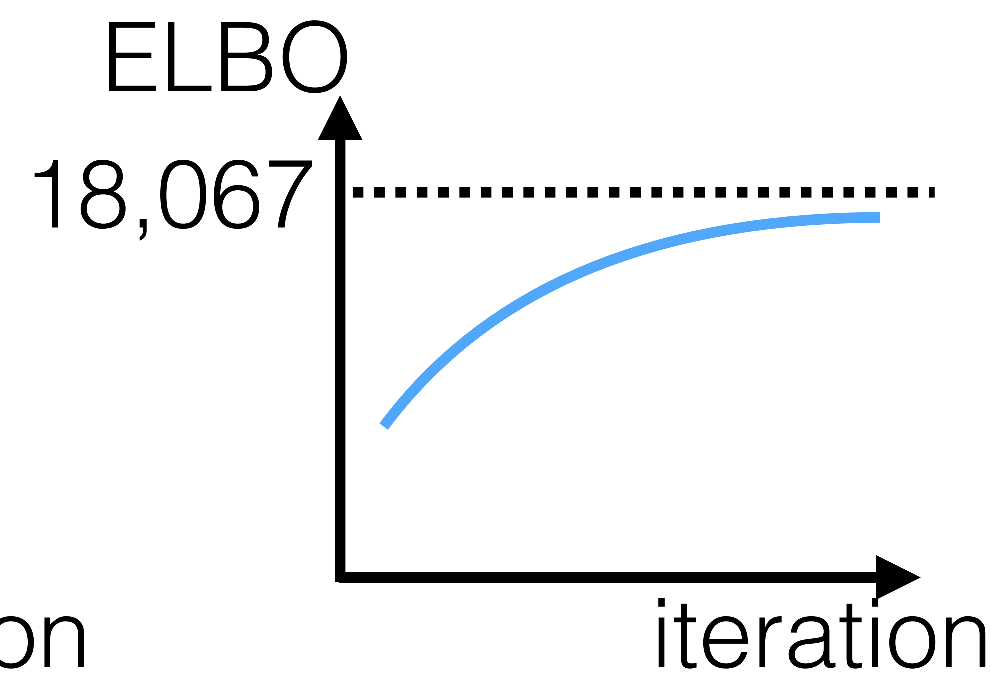
How small is KL in practice?



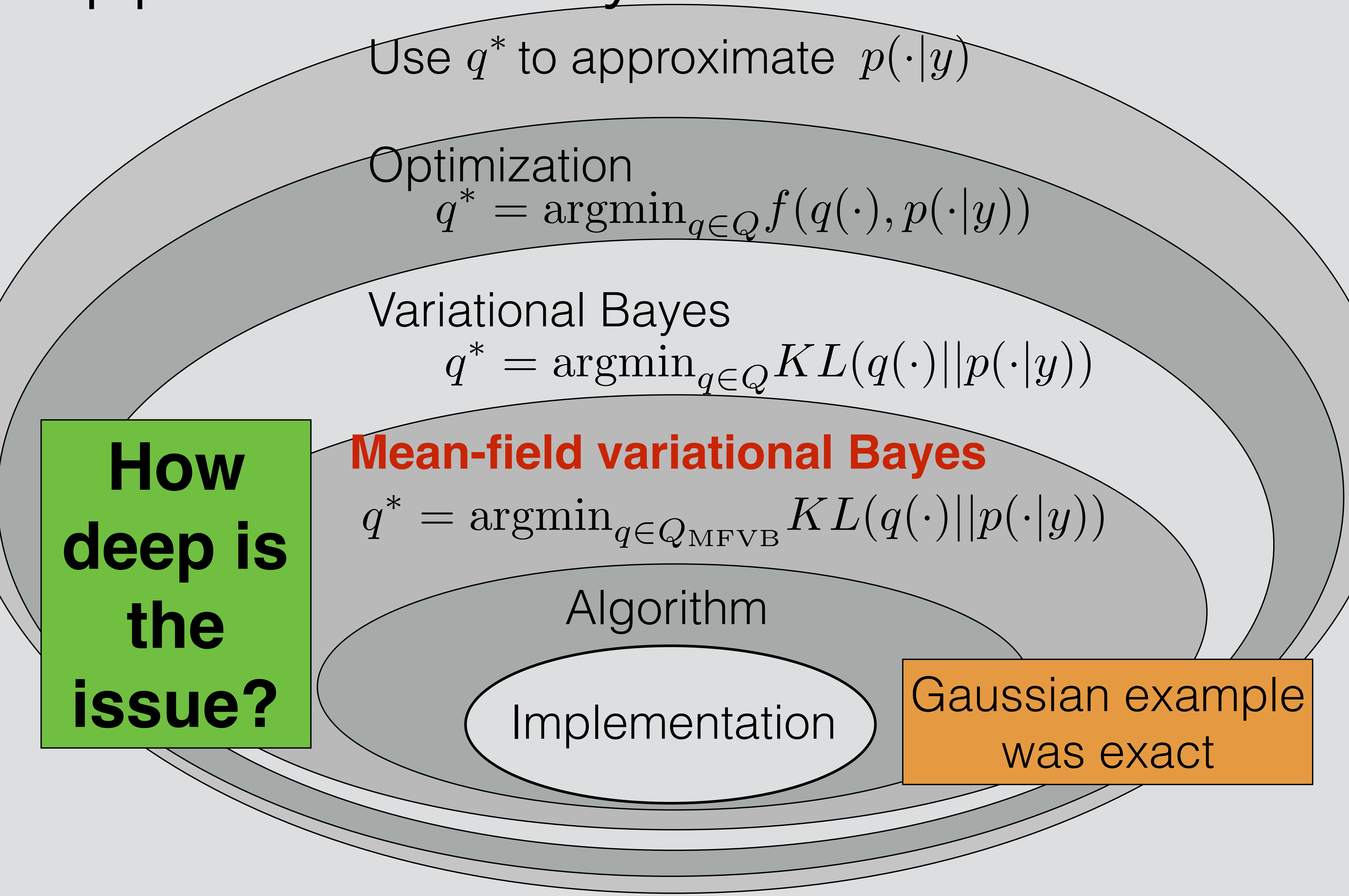
- Often optimum has non-zero KL (MFVB, Gaussian VB)
- Even if optimum has zero KL, algorithm might not reach zero

- Recall: the KL value isn't free

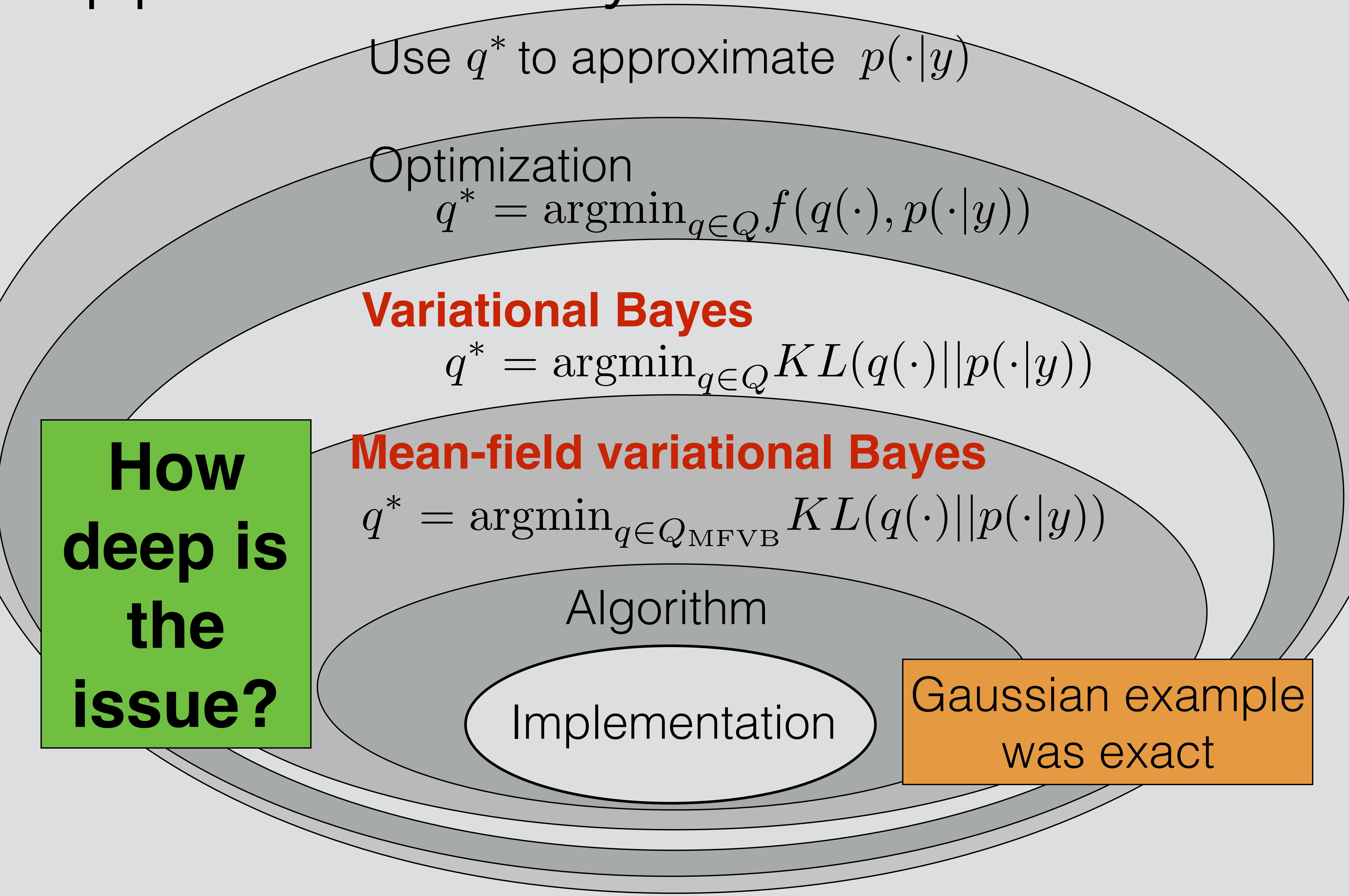




Approximate Bayesian inference



Approximate Bayesian inference



Roadmap

- Bayes & Approximate Bayes review
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use MFVB?
- When can we trust MFVB?
- Where do we go from here?

Roadmap

- Bayes & Approximate Bayes review
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use MFVB?
- When can we trust MFVB?
- Where do we go from here?

Roadmap

- Bayes & Approximate Bayes review
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use MFVB?
- When can we trust VB?
- Where do we go from here?

Roadmap

- Bayes & Approximate Bayes review
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use MFVB?
- When can we trust VB?
- Where do we go from here?

Roadmap

- Bayes & Approximate Bayes review
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use MFVB?
- When can we trust VB?
- Where do we go from here?

Latent Dirichlet Allocation (LDA)

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

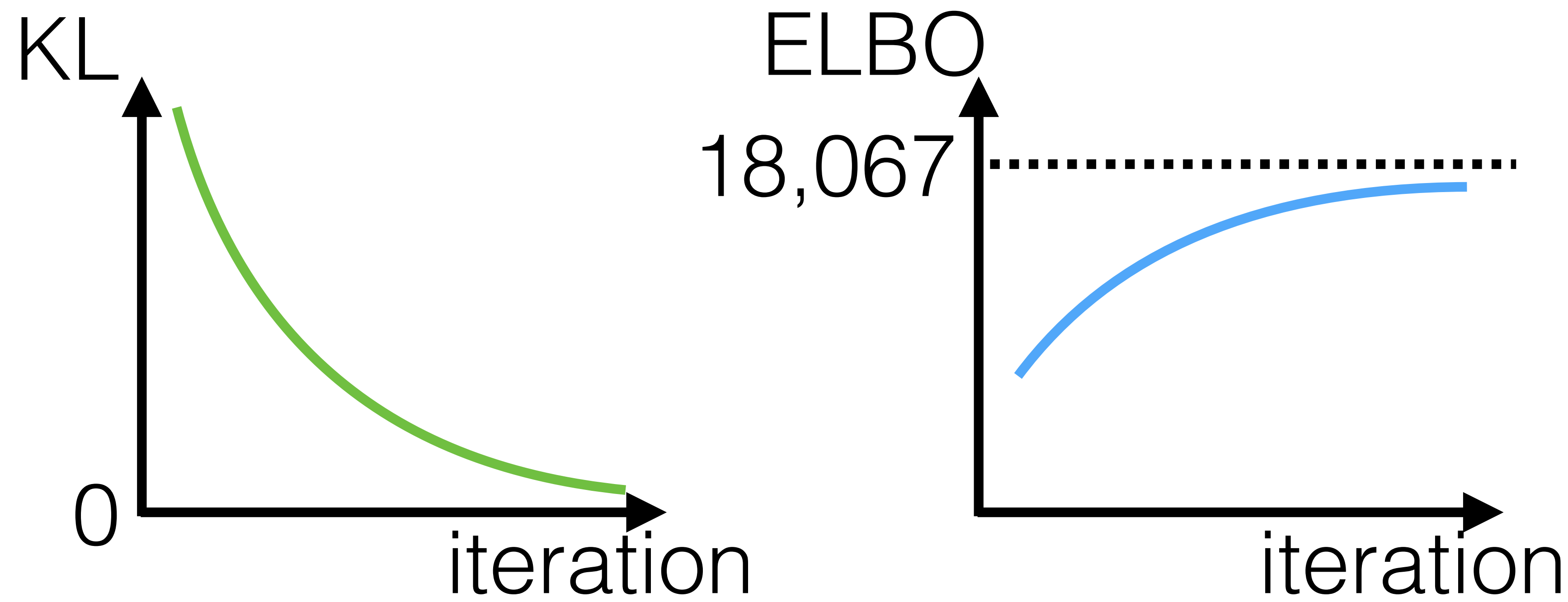
[Blei et al 2003]

What can we do?

- Reliable diagnostics

What can we do?

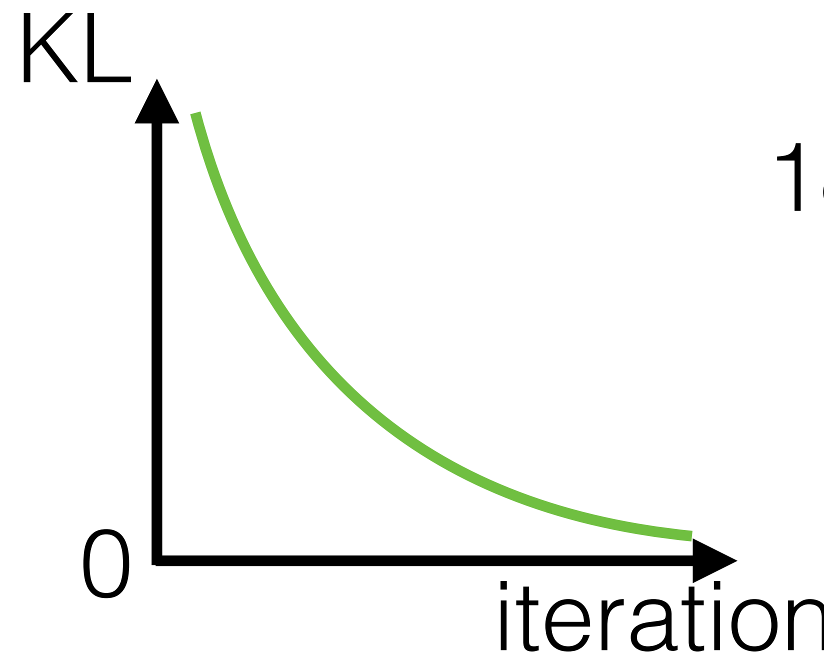
- Reliable diagnostics
    - cf. KL, ELBO

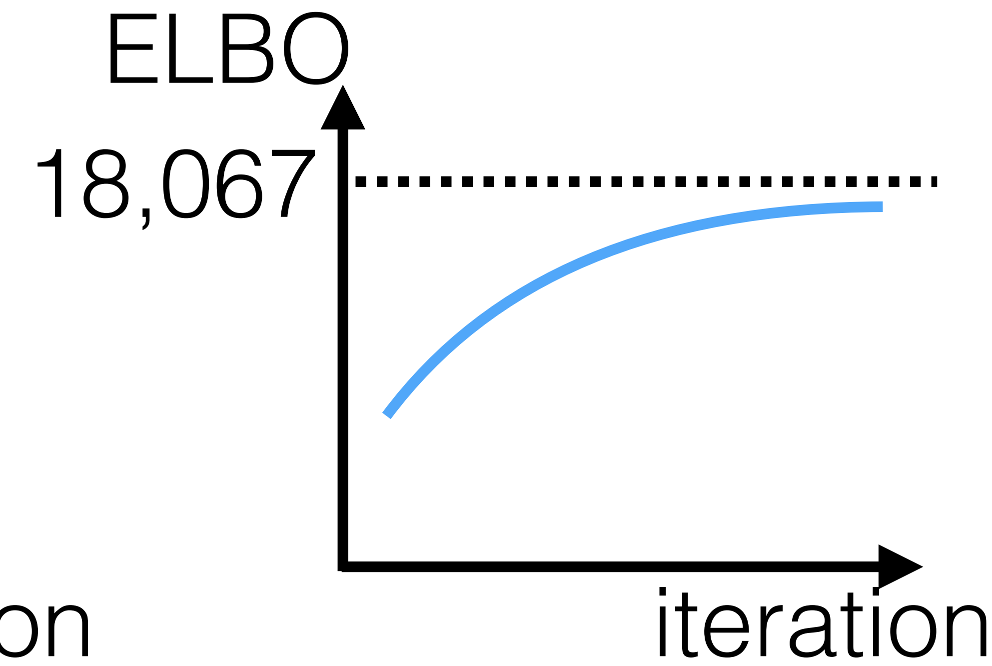


What can we do?

- Reliable diagnostics
  - cf. KL, ELBO

[Gorham, Mackey 2015, 2017; Chwialkowski, Strathmann, Gretton 2016; Jitkrittum et al 2017; Talts et al 2018; Yao et al 2018, etc.]



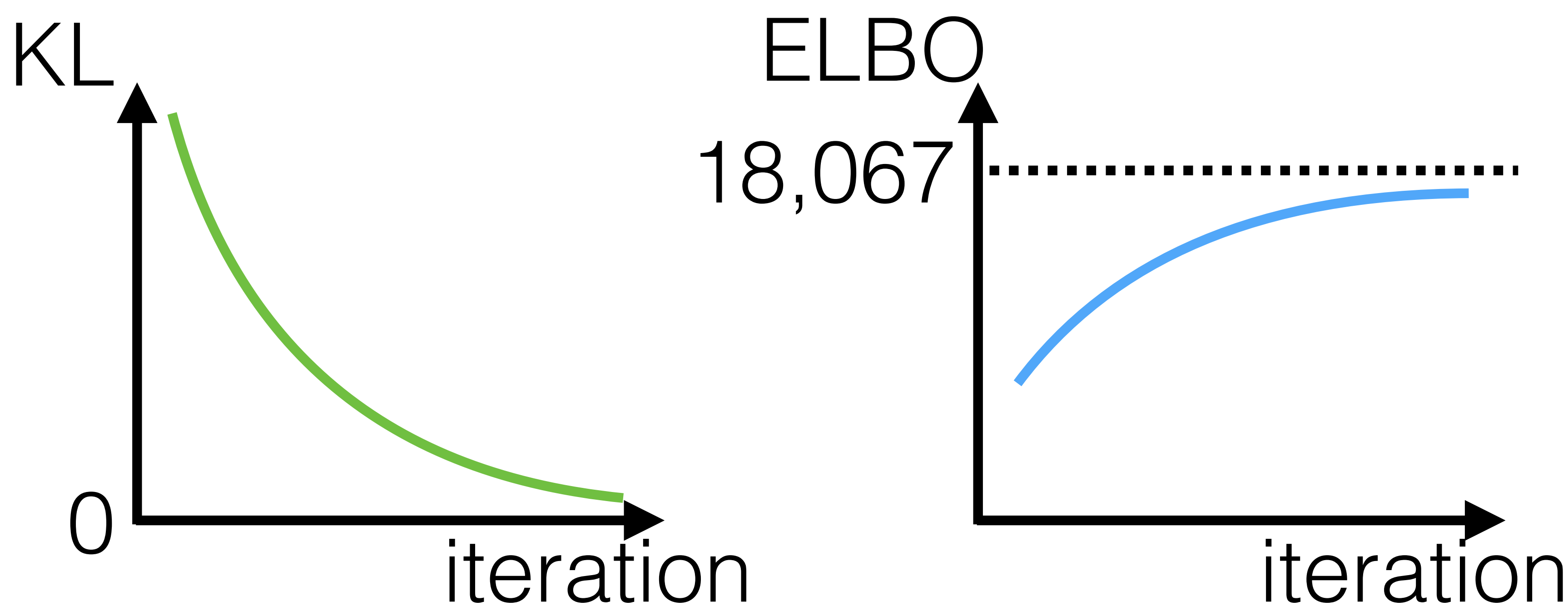


→ "Yes, but did it work? Evaluating variational inference" ICML 2018

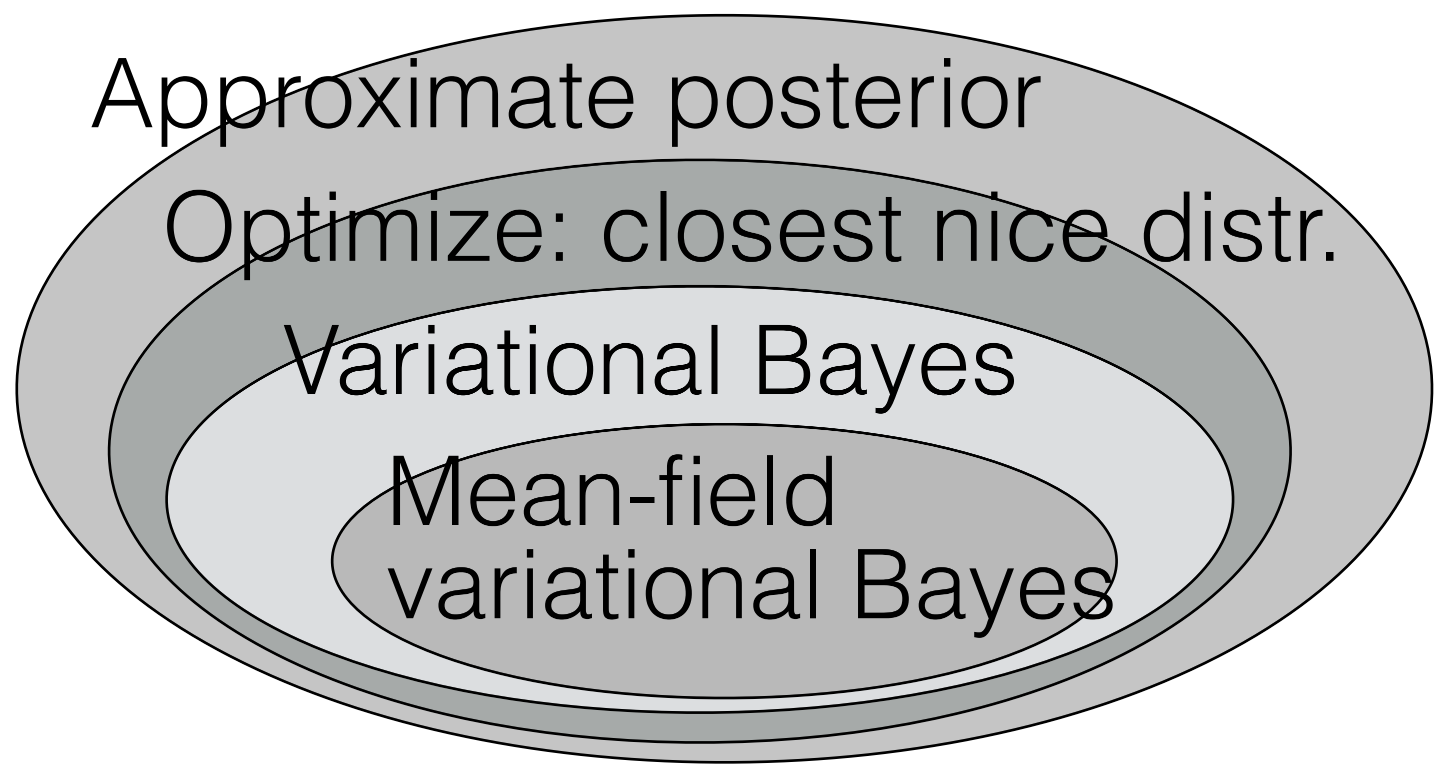
What can we do?

- Reliable diagnostics
  - cf. KL, ELBO

[Gorham, Mackey 2015, 2017; Chwialkowski, Strathmann, Gretton 2016; Jitkrittum et al 2017; Talts et al 2018; Yao et al 2018, etc.]



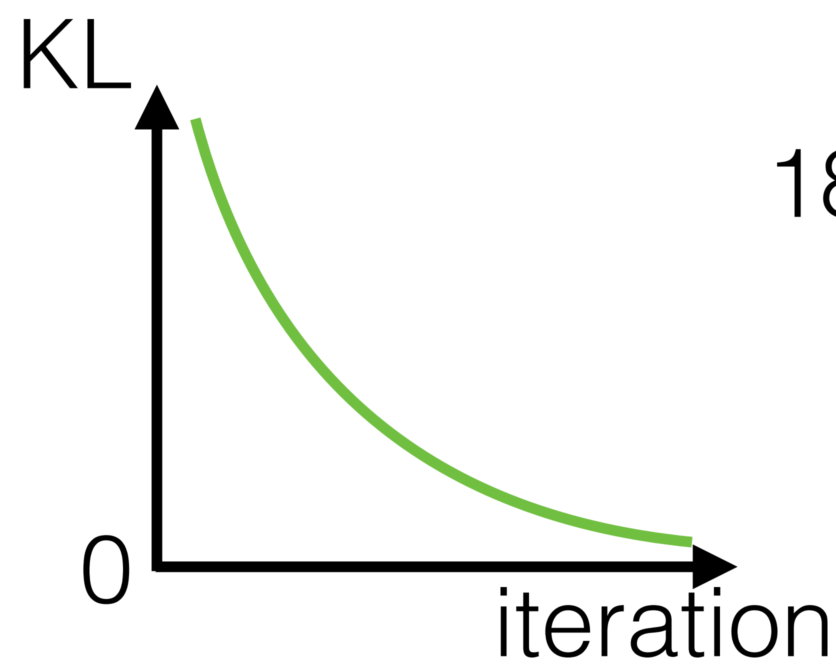
→ “Yes, but did it work? Evaluating variational inference” ICML 2018

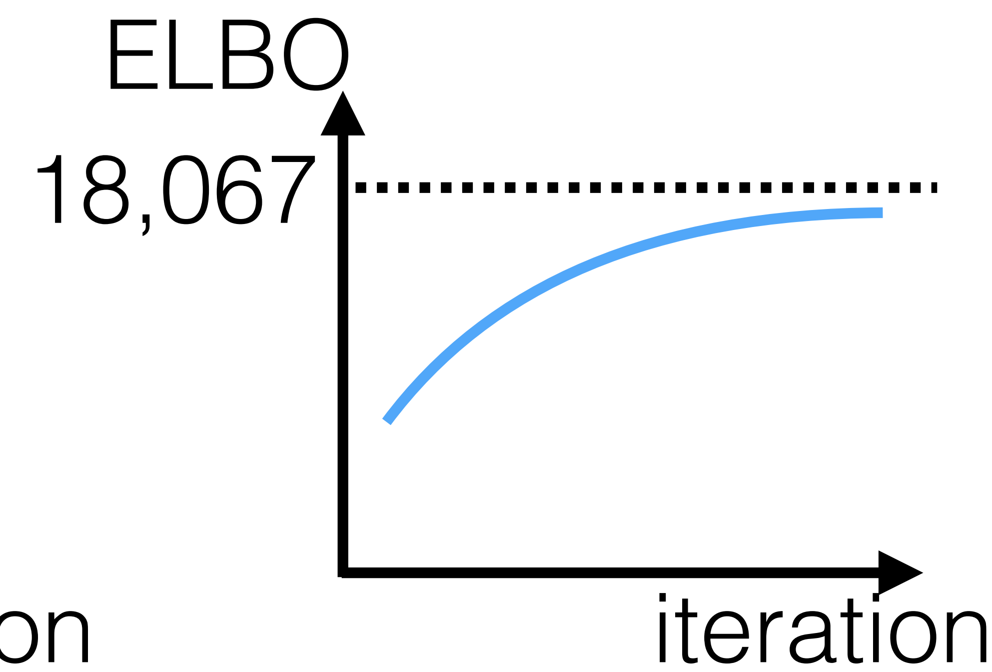


What can we do?

- Reliable diagnostics
  - cf. KL, ELBO

[Gorham, Mackey 2015, 2017; Chwialkowski, Strathmann, Gretton 2016; Jitkrittum et al 2017; Talts et al 2018; Yao et al 2018, etc.]

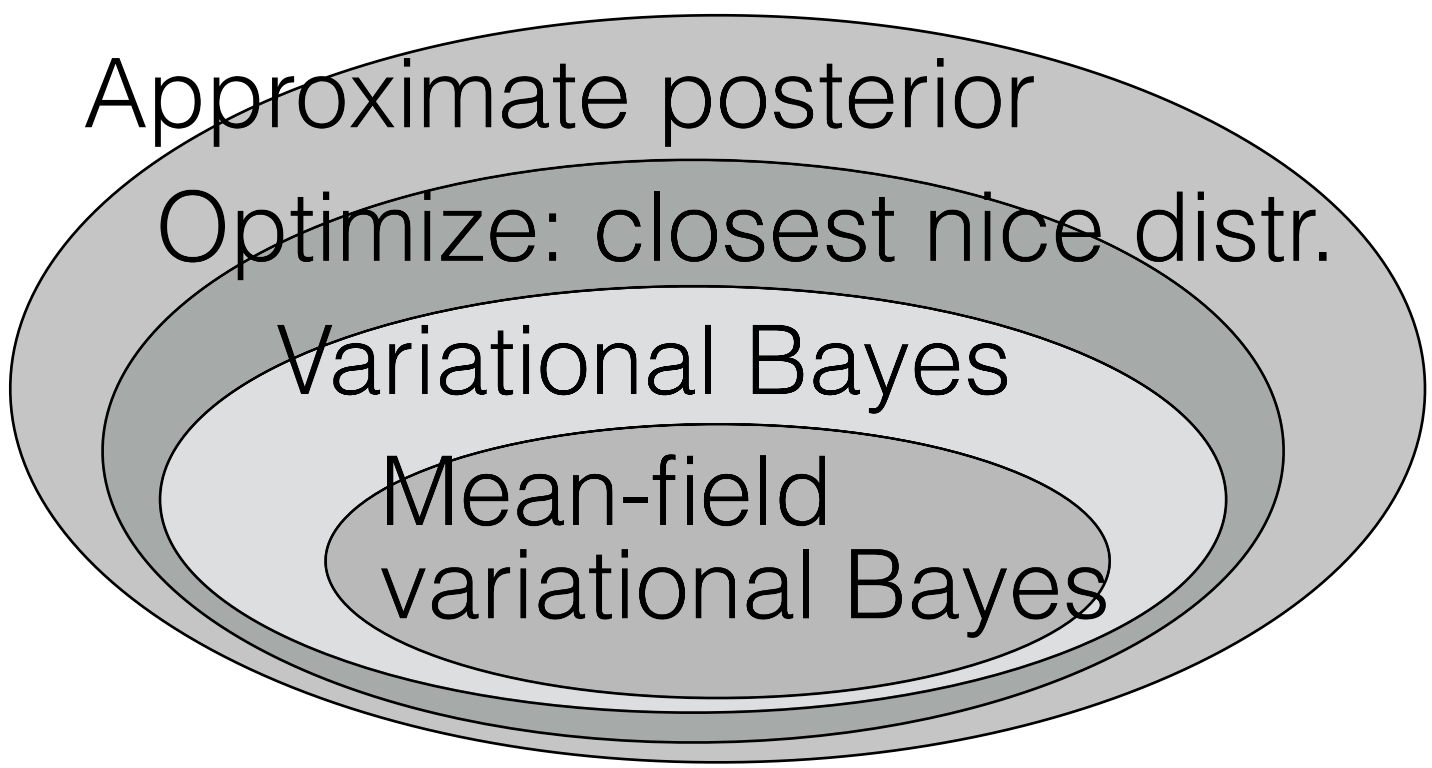




→ "Yes, but did it work? Evaluating variational inference" ICML 2018

- Alternative divergences: Time & accuracy

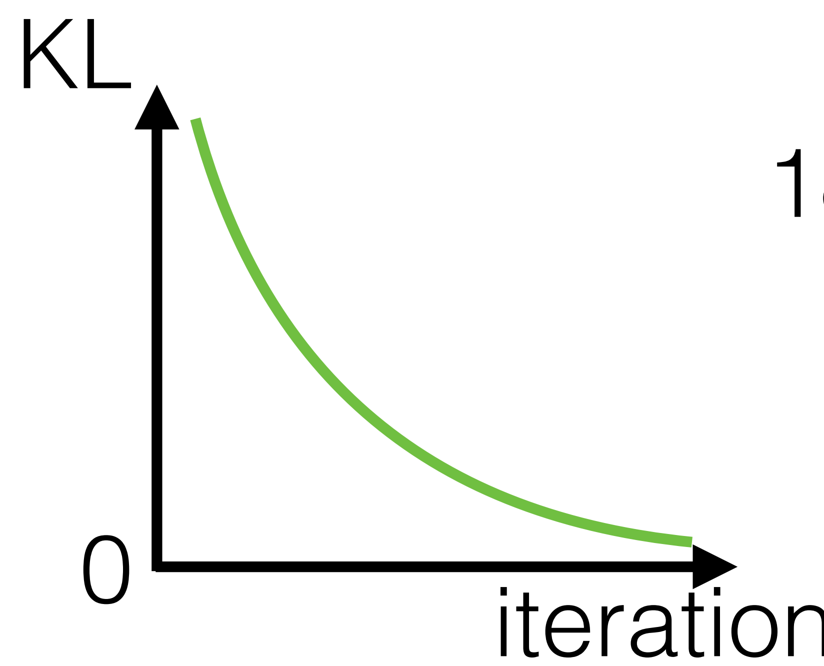
[Huggins, Kasprzak, Campbell, Broderick, 2018]

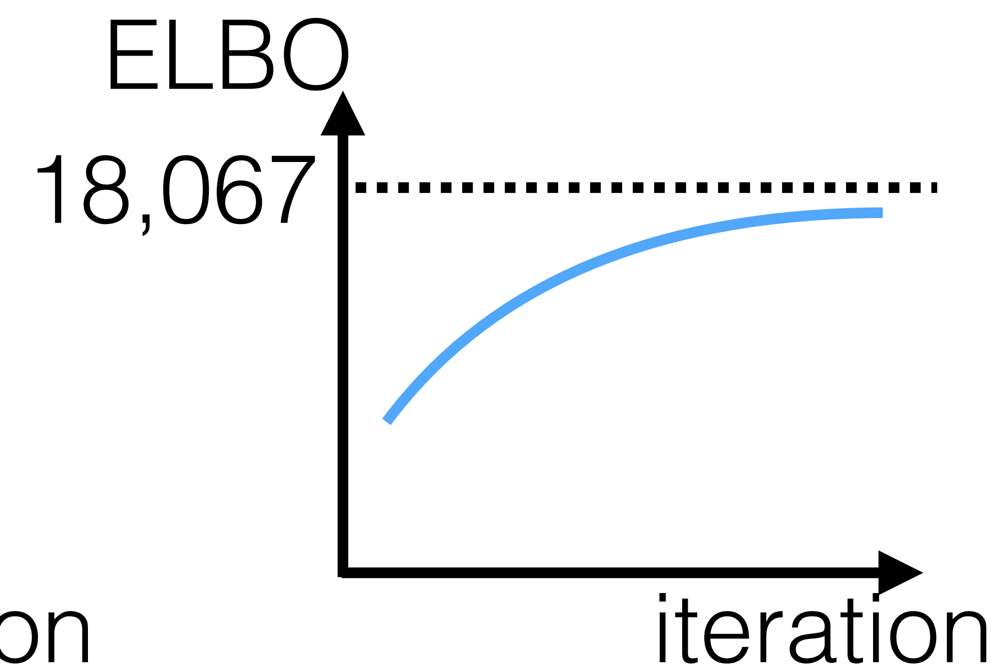


What can we do?

- Reliable diagnostics
    - cf. KL, ELBO

[Gorham, Mackey 2015, 2017; Chwialkowski, Strathmann, Gretton 2016; Jitkrittum et al 2017; Talts et al 2018; Yao et al 2018, etc.]





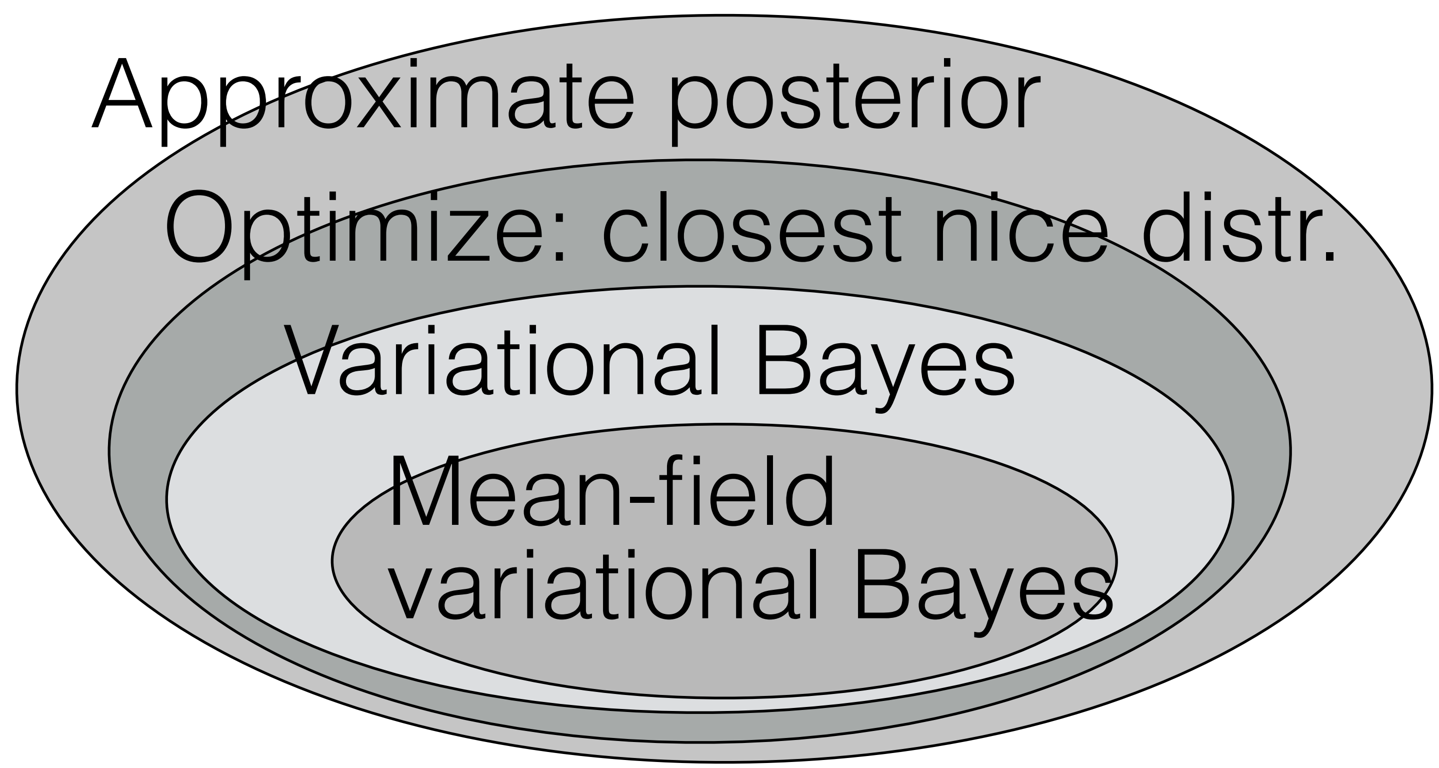
→ “Yes, but did it work? Evaluating variational inference” ICML 2018

- Alternative divergences: Time & accuracy

[Huggins, Kasprzak, Campbell, Broderick, 2018]

- Corrections

[Giordano, Broderick, Jordan 2018]

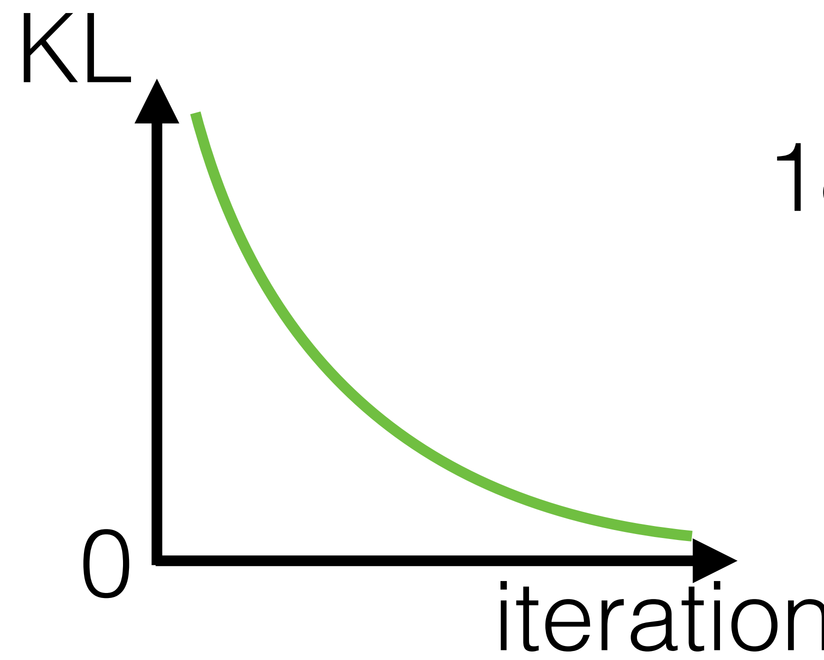


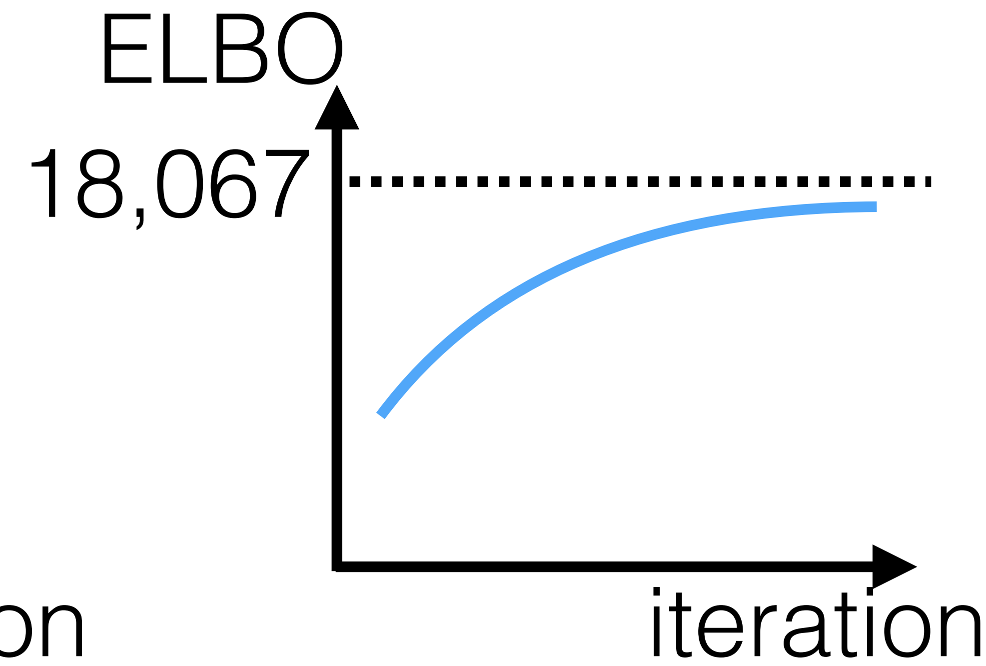
What can we do?

- Reliable diagnostics
  - cf. KL, ELBO

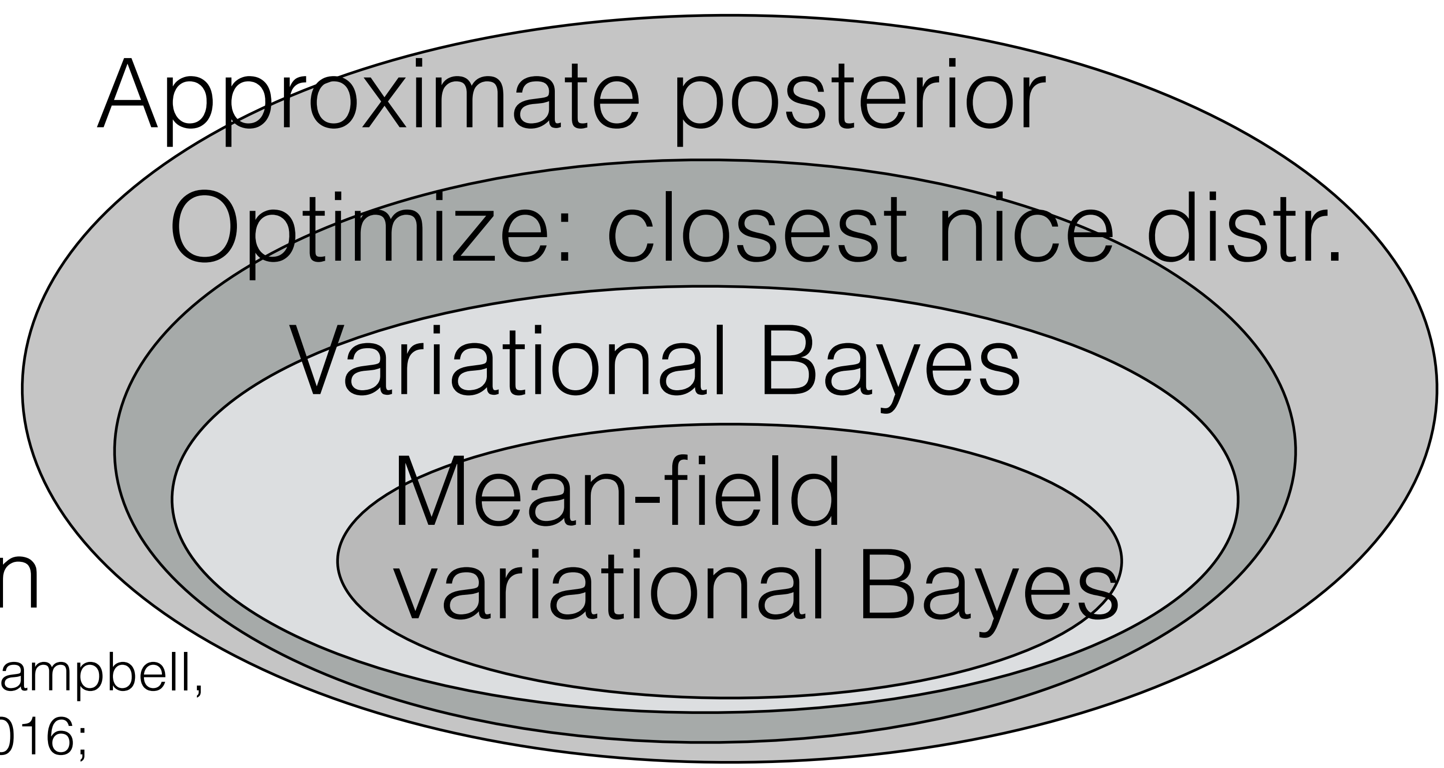
[Gorham, Mackey 2015, 2017; Chwialkowski, Strathmann, Gretton 2016; Jitkrittum et al 2017; Talts et al 2018; Yao et al 2018, etc.]

→ "Yes, but did it work? Evaluating variational inference" ICML 2018





- Alternative divergences: Time & accuracy [Huggins, Kasprzak, Campbell, Broderick, 2018]
- Corrections [Giordano, Broderick, Jordan 2018]
- Theoretical guarantees on finite-data quality [Huggins, Campbell, Broderick 2016; Campbell, Broderick 2018, 2019]



What to read next

Textbooks and Reviews

- Bishop. *Pattern Recognition and Machine Learning*, Ch 10. 2006.
- Blei, Kucukelbir, McAuliffe. Variational inference: A review for statisticians, *JASA* 2016.
- MacKay. *Information Theory, Inference, and Learning Algorithms*, Ch 33. 2003.
- Murphy. *Machine Learning: A Probabilistic Perspective*, Ch 21. 2012.
- Ormerod, Wand. Explaining Variational Approximations. *Amer Stat* 2010.
- Turner, Sahani. Two problems with variational expectation maximisation for time-series models. In *Bayesian Time Series Models*, 2011.
- Wainwright, Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 2008.

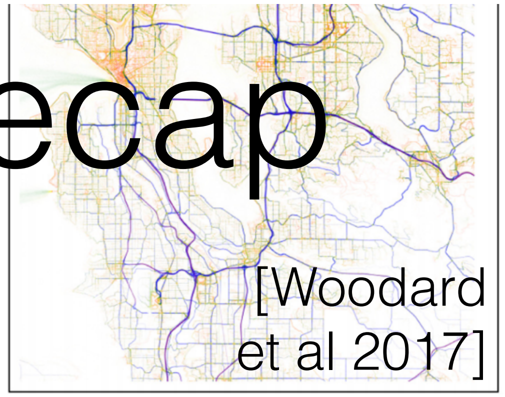
Our Experiments

- RJ Giordano, T Broderick, and MI Jordan. Linear response methods for accurate covariance estimates from mean field variational Bayes. *NeurIPS* 2015.
- RJ Giordano, T Broderick, R Meager, J Huggins, and MI Jordan. Fast robustness quantification with variational Bayes. *ICML Data4Good Workshop* 2016.
- RJ Giordano, T Broderick, and MI Jordan. Covariances, robustness, and variational Bayes. *Journal of Machine Learning Research*, 2018.
- J Huggins, M Kasprzak, T Campbell, T Broderick. Practical bounds on the error of Bayesian posterior approximations: A nonasymptotic approach, 2018. ArXiv: 1809.09505.

Automated, Scalable Bayesian Inference via Data Summarization

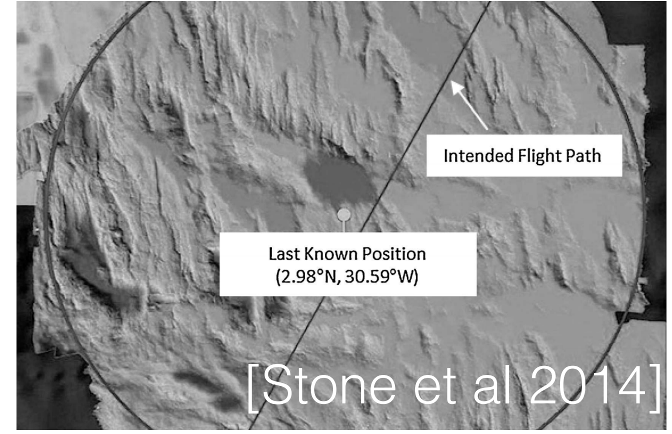
http://www.tamarabroderick.com/tutorials.html

Recap



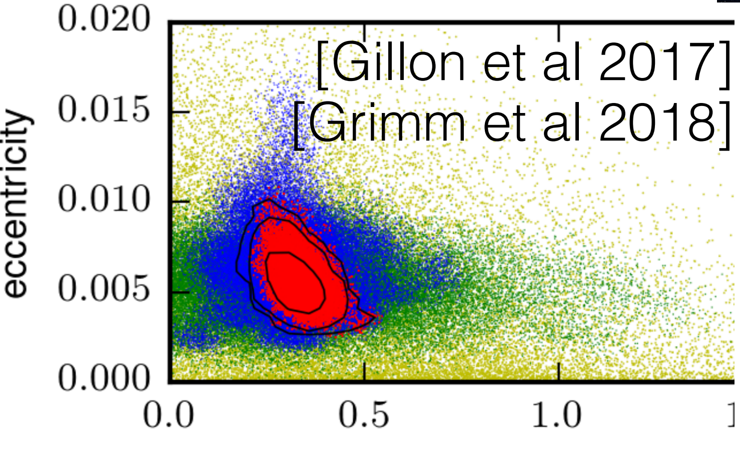
[Woodard et al 2017]

[ESO/ L. Calçada/ M. Kornmesser 2017] [Abbott et al 2016a,b]

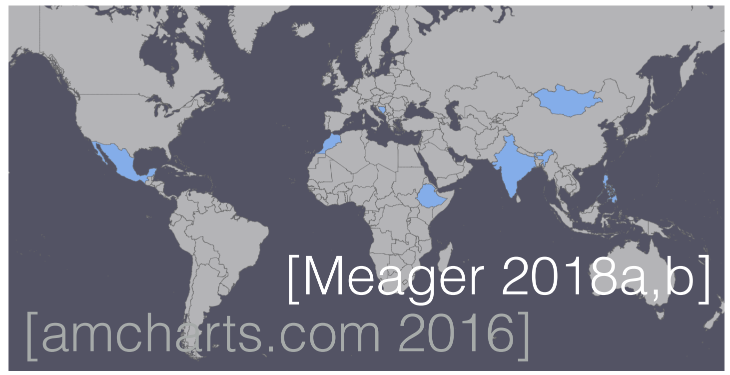


[Stone et al 2014]

[mc-stan.org]



[Gillon et al 2017]
[Grimm et al 2018]

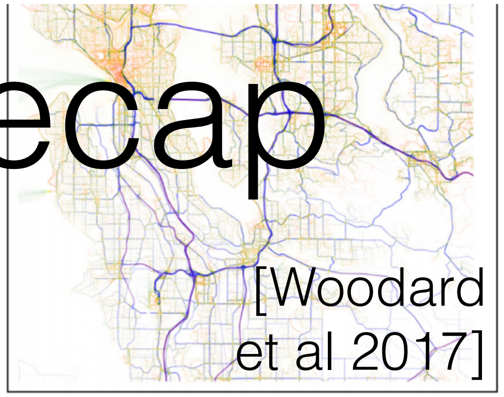


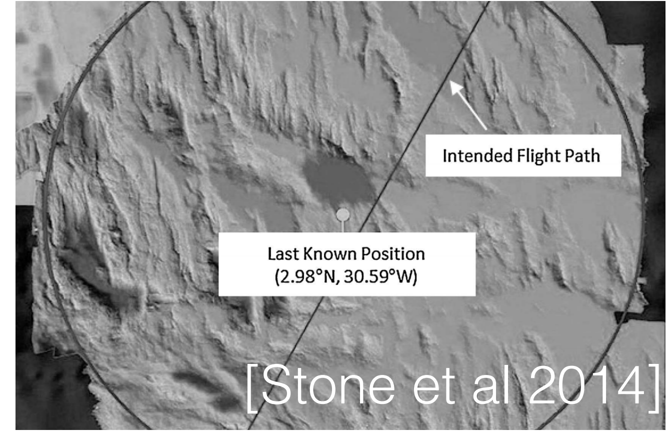
[Meager 2018a,b]
[amcharts.com 2016]

[Chati, Balakrishnan 2017]
[Julian Hertzog 2016]

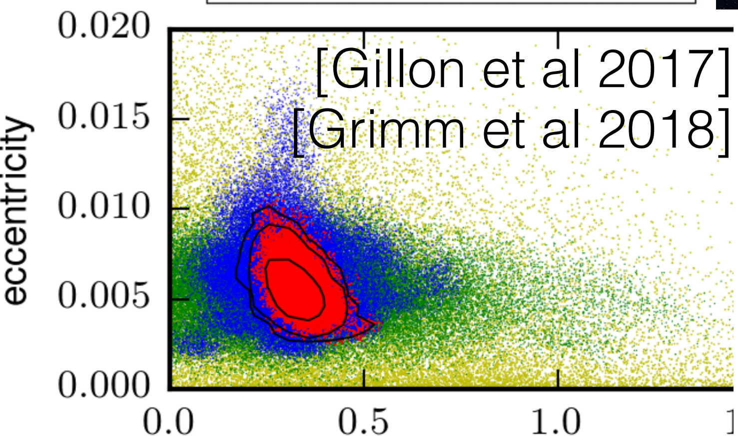
[Kuikka et al 2014]
[Baltic Salmon Fund]

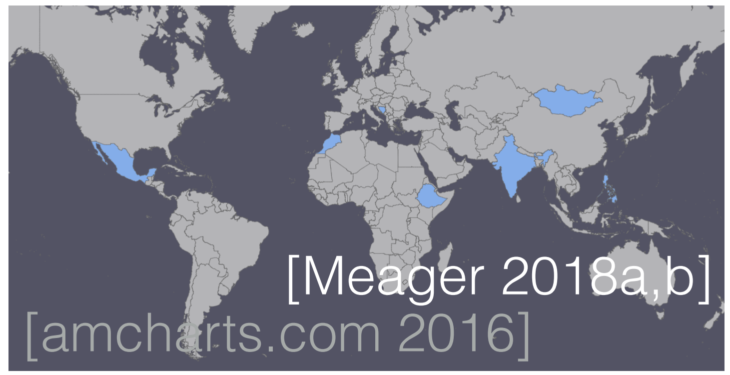
Recap





[mc-stan.org]

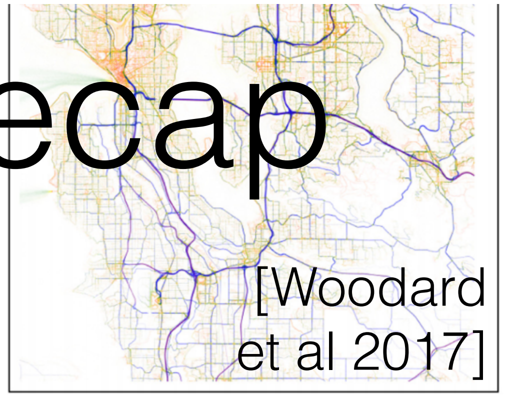


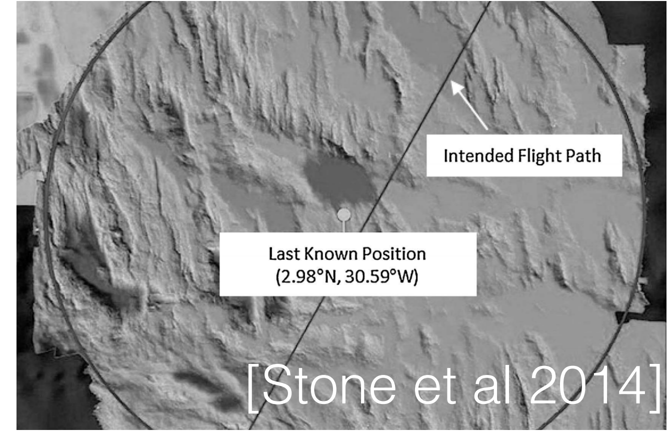


posterior likelihood prior

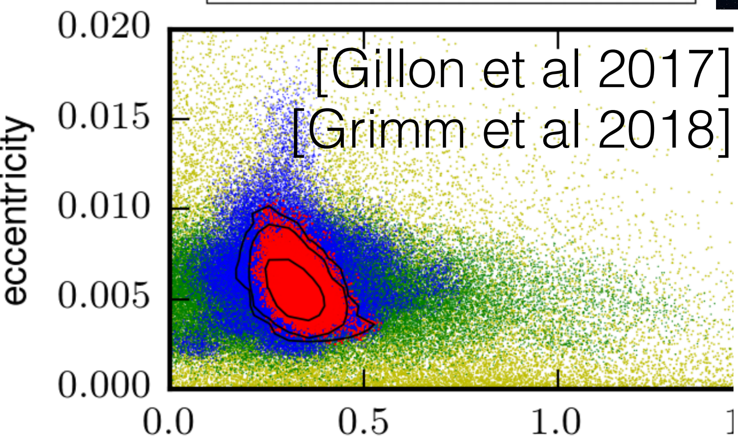
$$p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$$

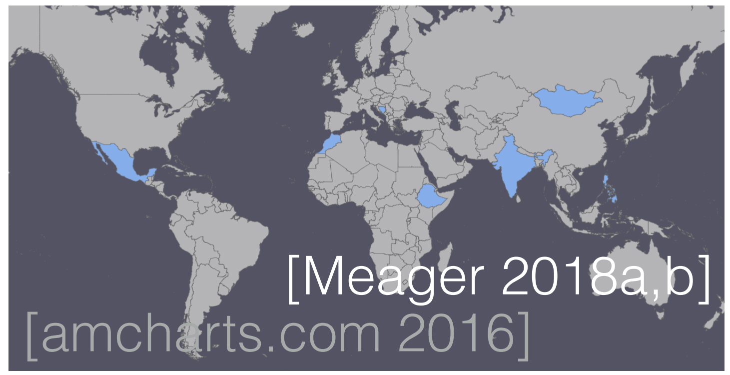
Recap





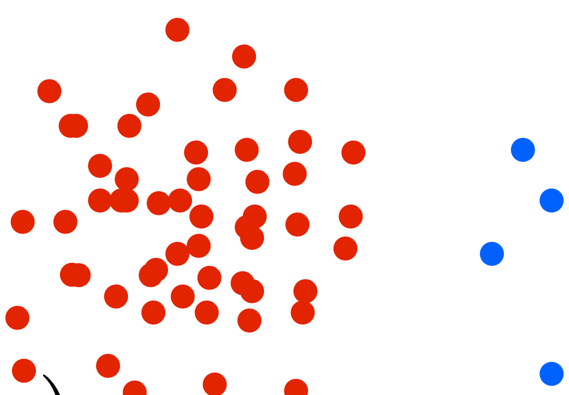
[mc-stan.org]





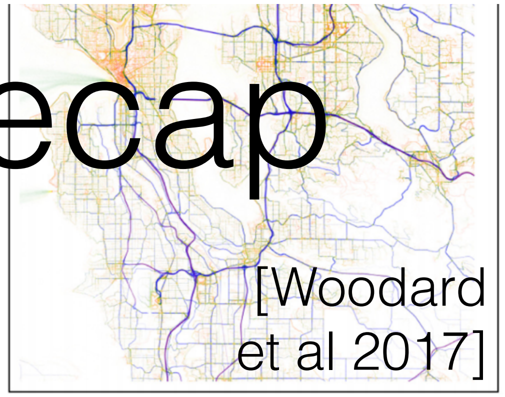
posterior likelihood prior

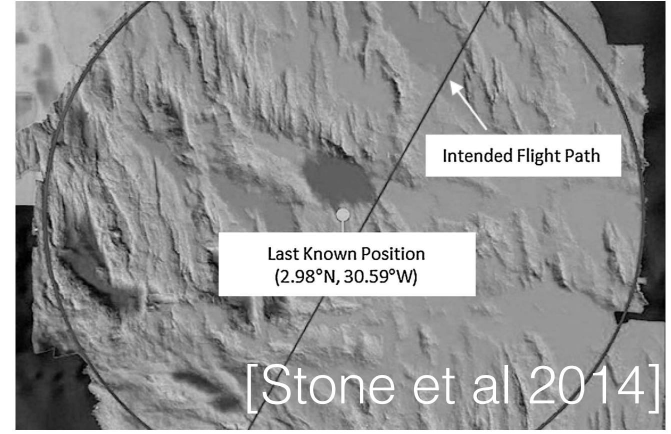
$$p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$$



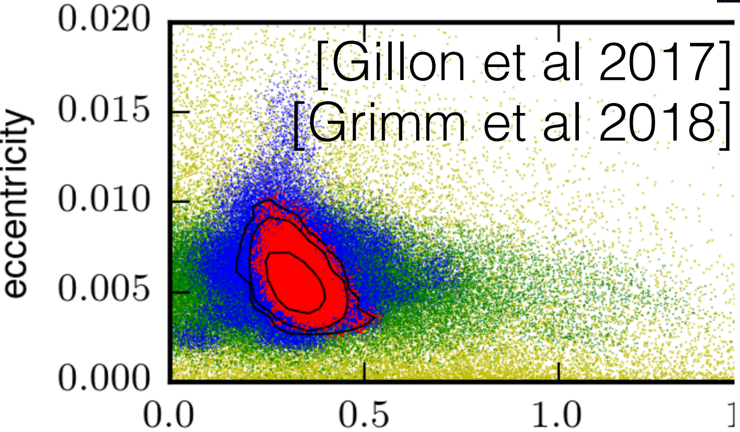
(x_n, y_n)

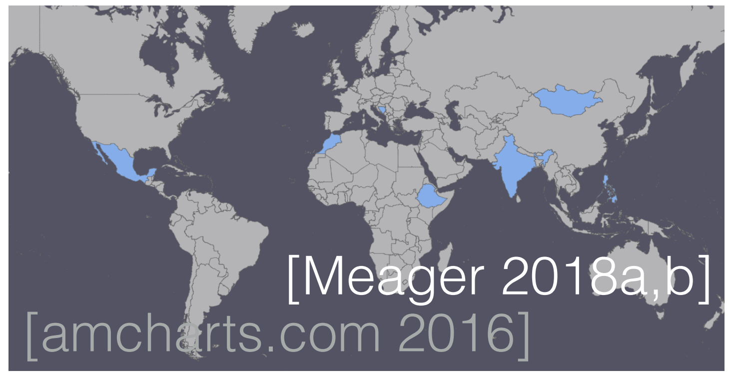
Recap





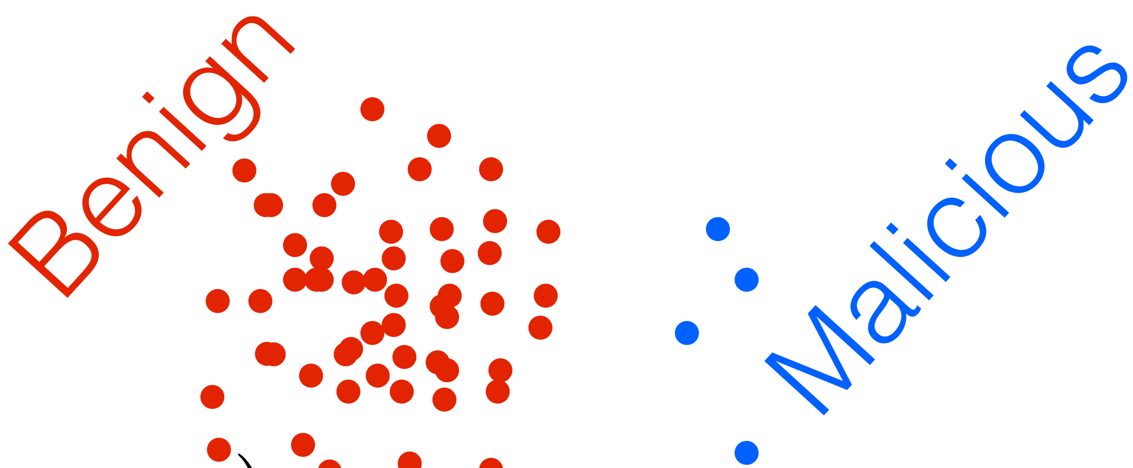
[mc-stan.org]





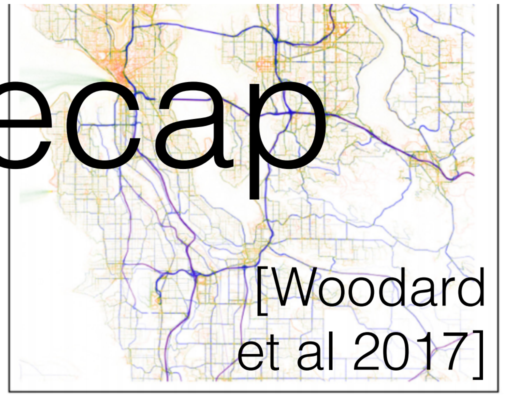
posterior likelihood prior

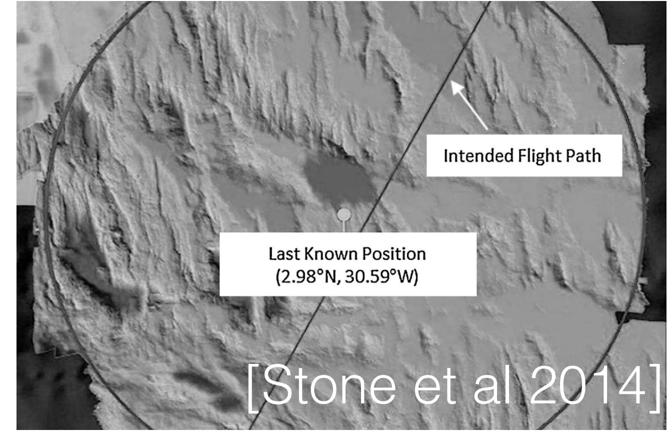
$$p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$$



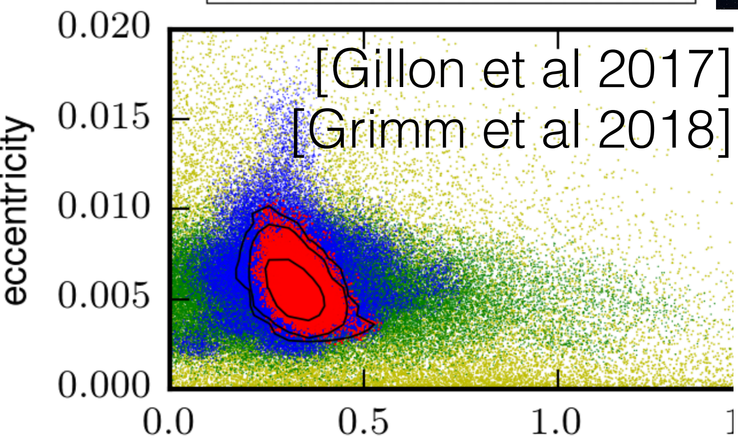
(x_n, y_n)

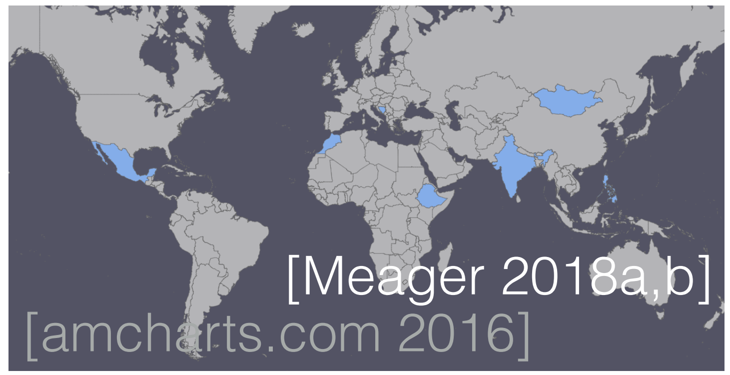
Recap





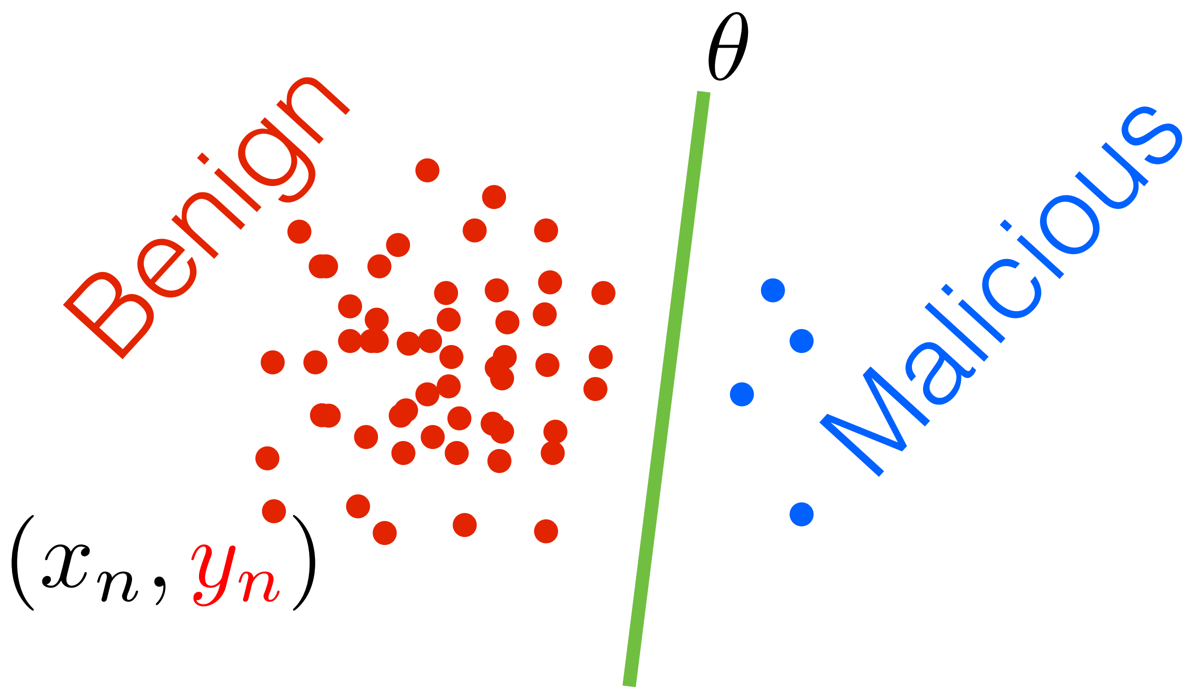
[mc-stan.org]



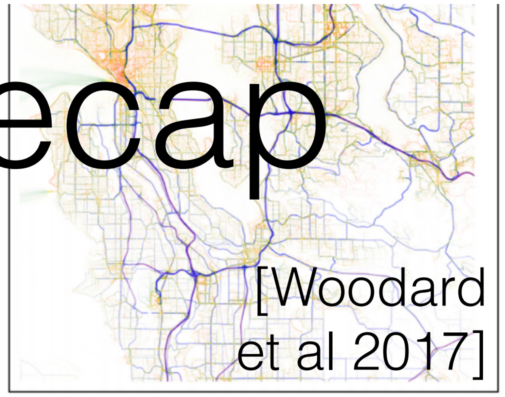


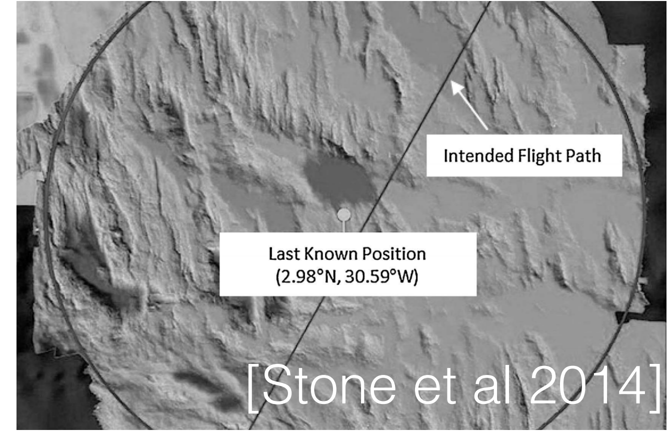
posterior likelihood prior

$$p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$$

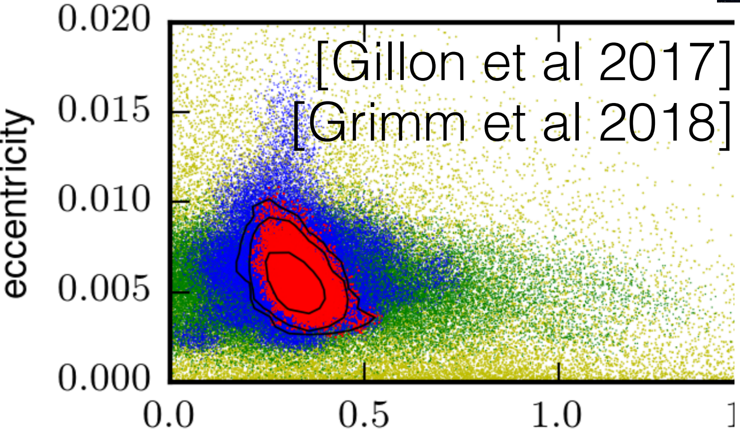


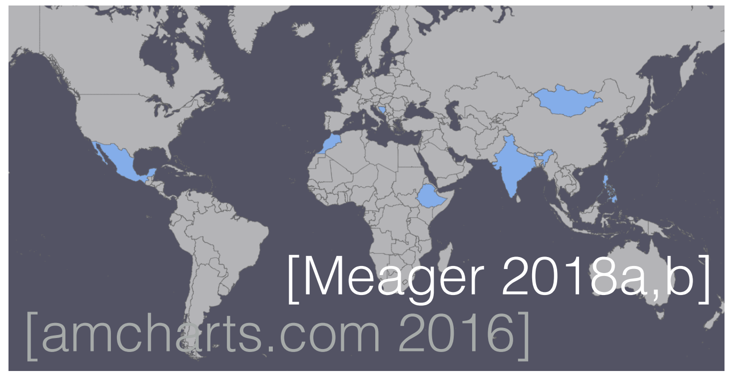
Recap





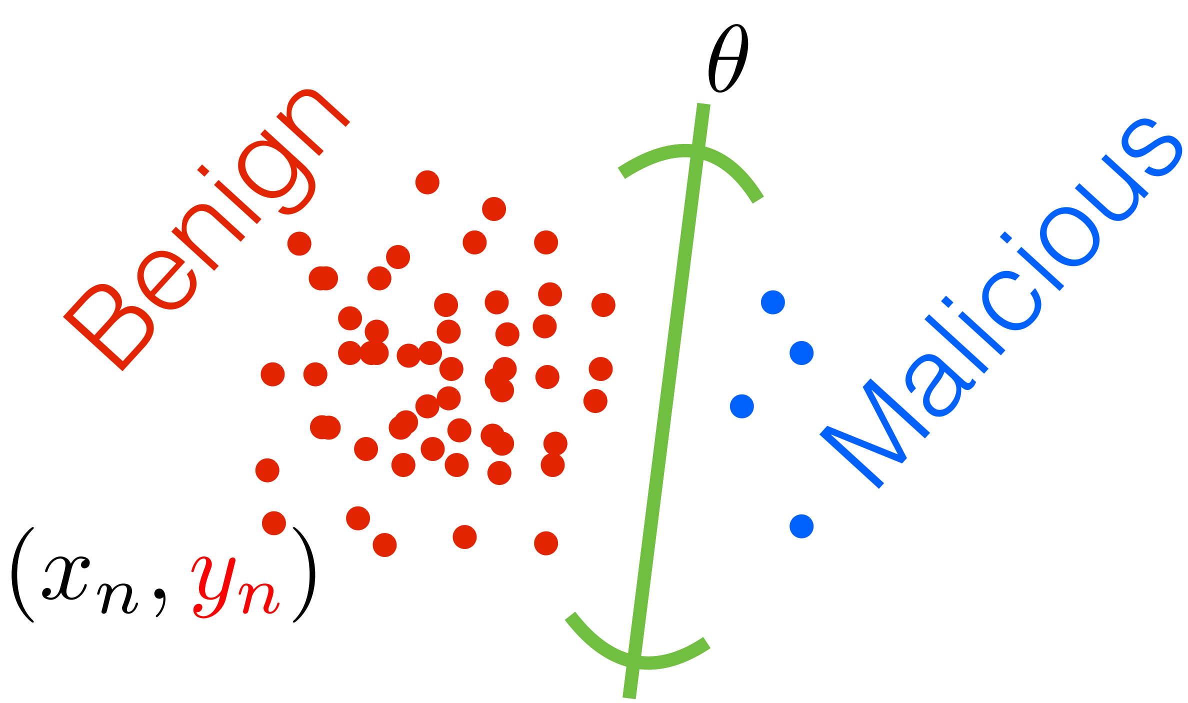
[mc-stan.org]



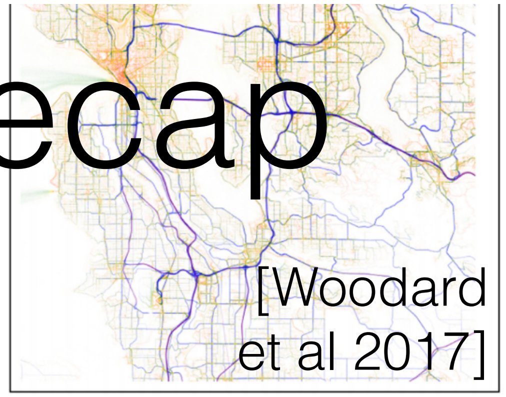


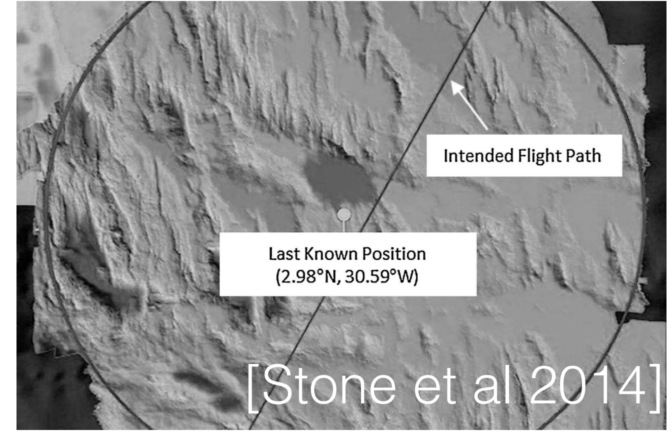
posterior likelihood prior

$$p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$$

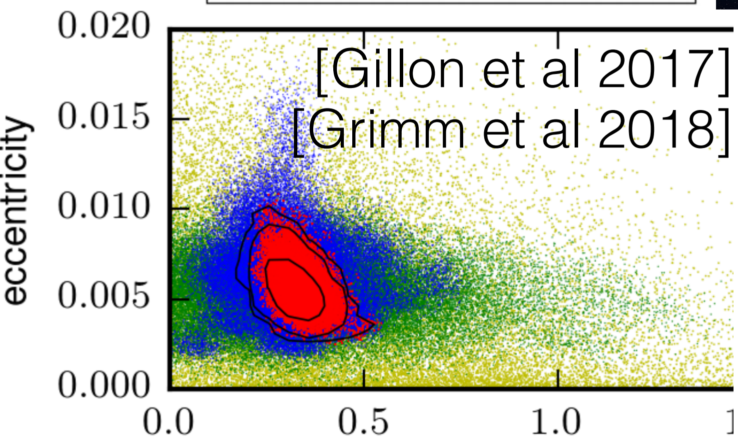


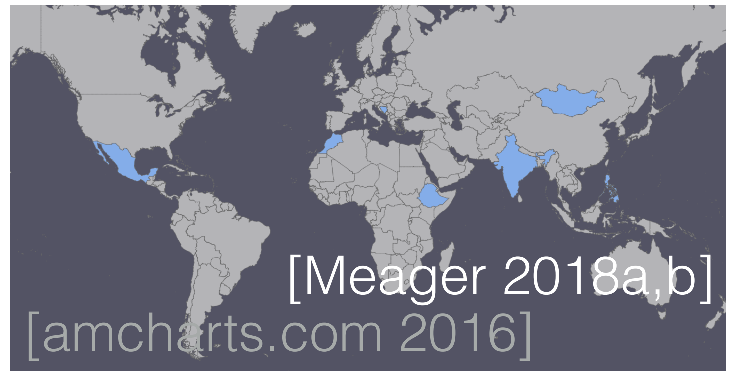
Recap





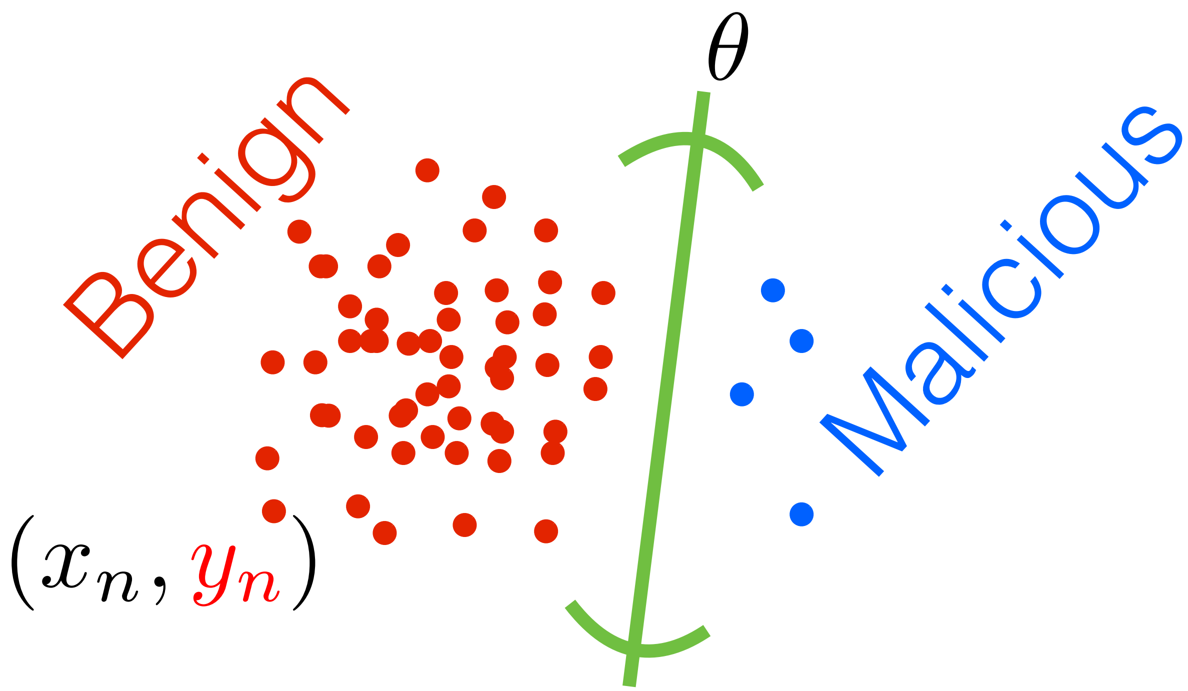
[mc-stan.org]

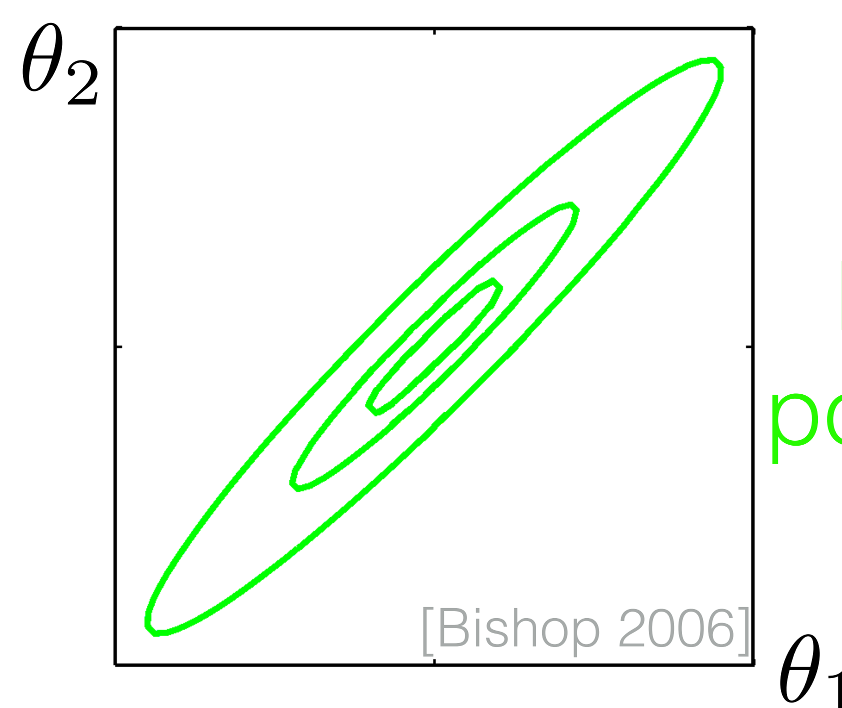




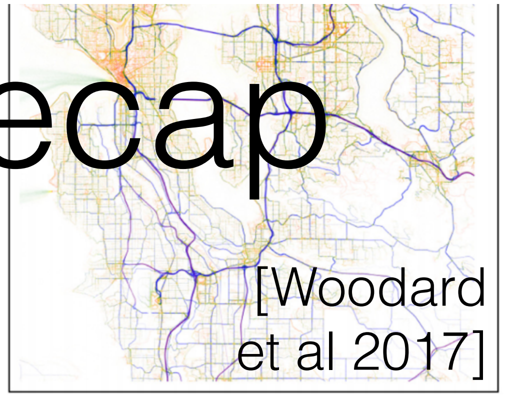
posterior likelihood prior

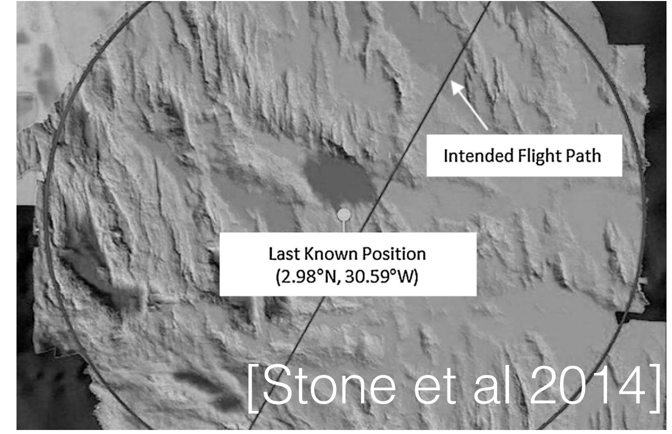
$$p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$$



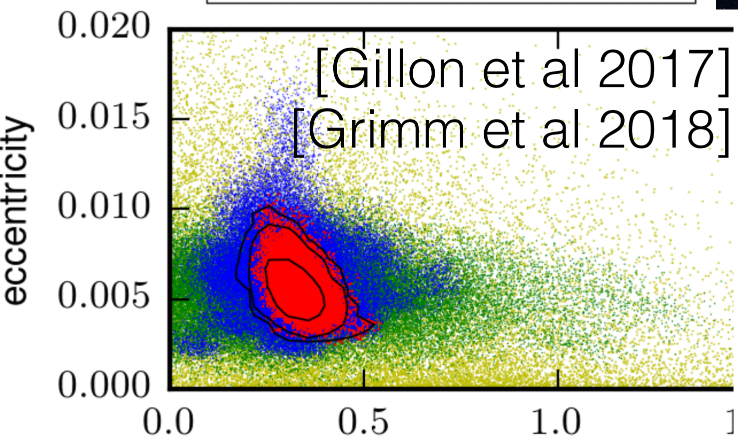


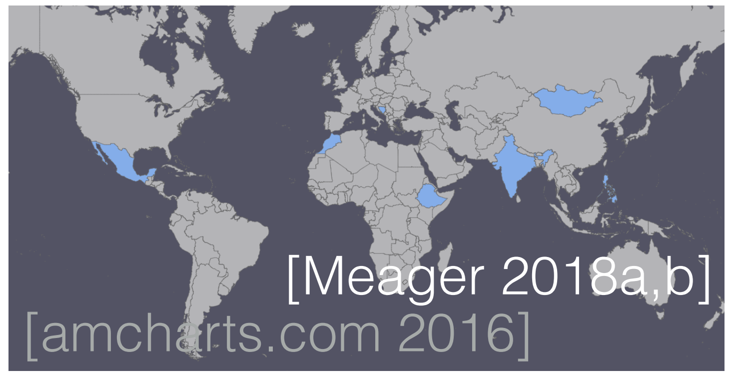
Recap





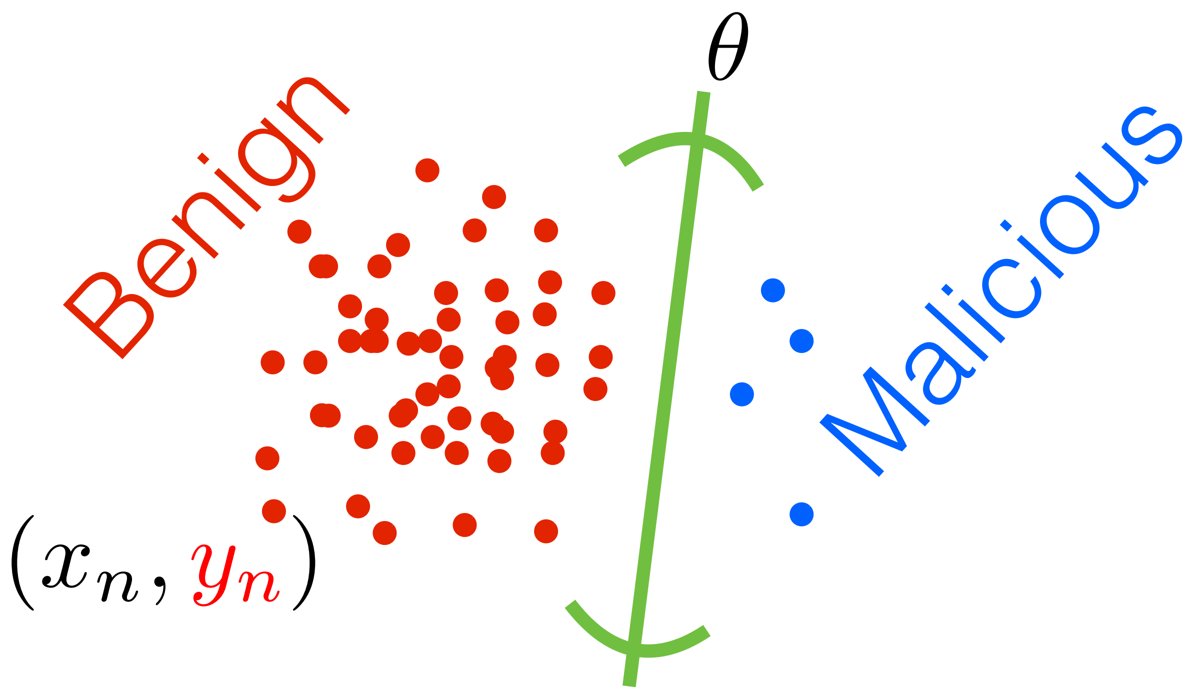
[mc-stan.org]

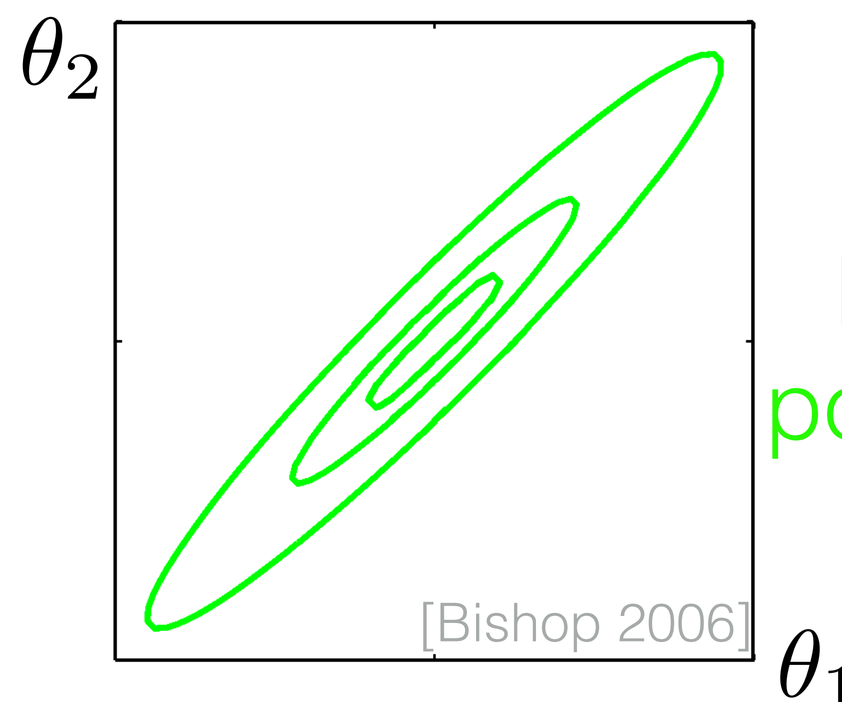




posterior likelihood prior

$$p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$$





Exact posterior

- Proposal: *efficient data summaries* for **fast**, **automated**, approximations with **error bounds for finite data**

Roadmap

Roadmap

- The “core” of the data set

Roadmap

- The “core” of the data set
- Uniform data subsampling isn’t enough

Roadmap

- The “core” of the data set
- Uniform data subsampling isn’t enough
- Importance sampling for “coresets”

Roadmap

- The “core” of the data set
- Uniform data subsampling isn’t enough
- Importance sampling for “coresets”
- Optimization for “coresets”

Roadmap

- The “core” of the data set
- Uniform data subsampling isn’t enough
- Importance sampling for “coresets”
- Optimization for “coresets”
- Approximate sufficient statistics

Roadmap

- The “core” of the data set
- Uniform data subsampling isn’t enough
- Importance sampling for “coresets”
- Optimization for “coresets”
- Approximate sufficient statistics

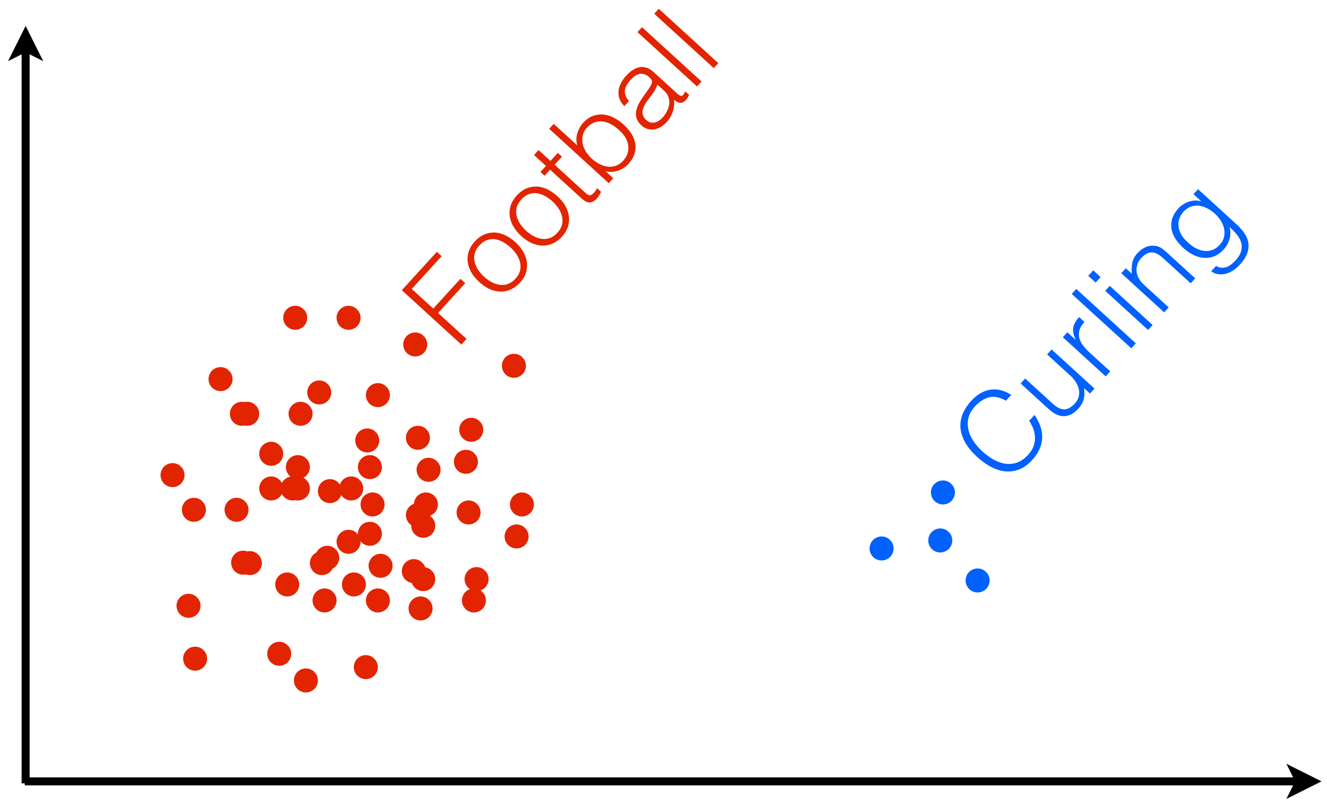
“Core” of the data set

“Core” of the data set

- Observe: redundancies can exist even if data isn’t “tall”

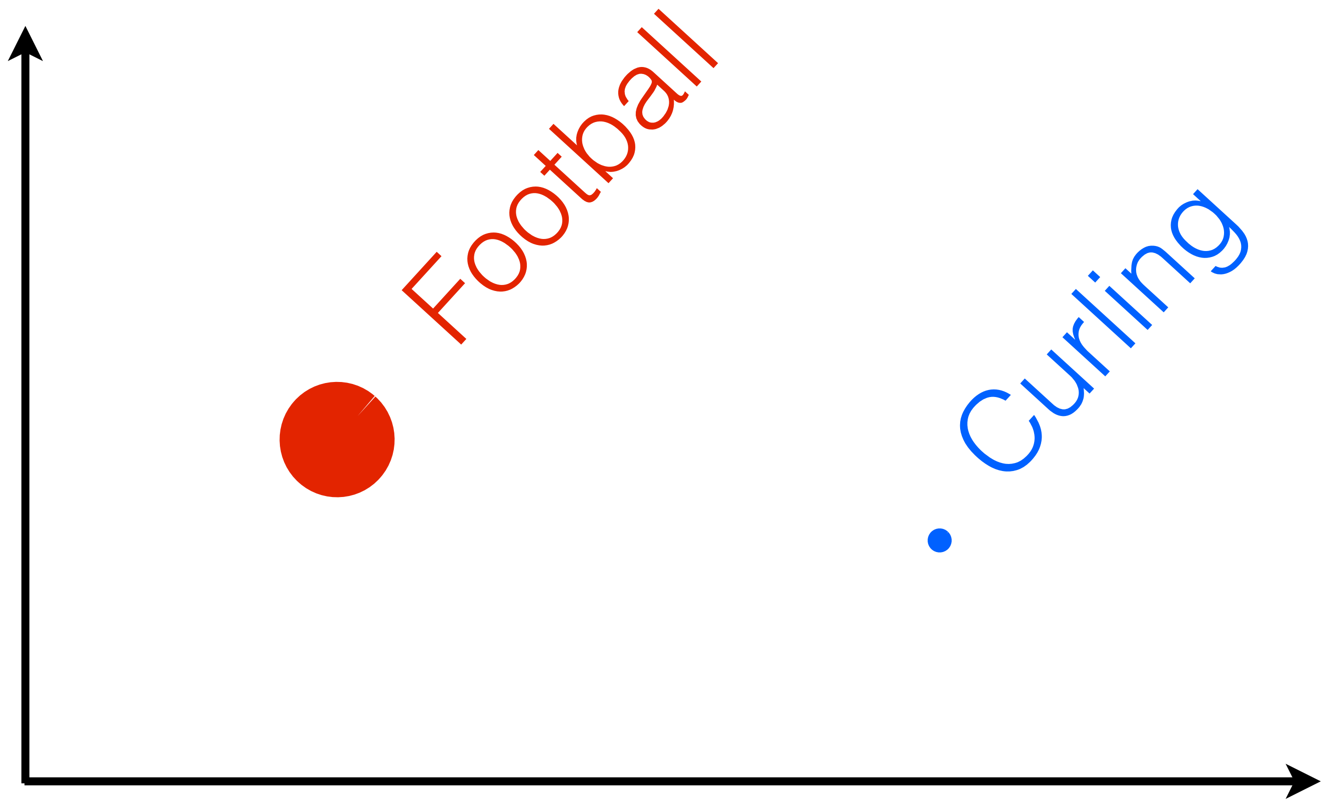
“Core” of the data set

- Observe: redundancies can exist even if data isn’t “tall”



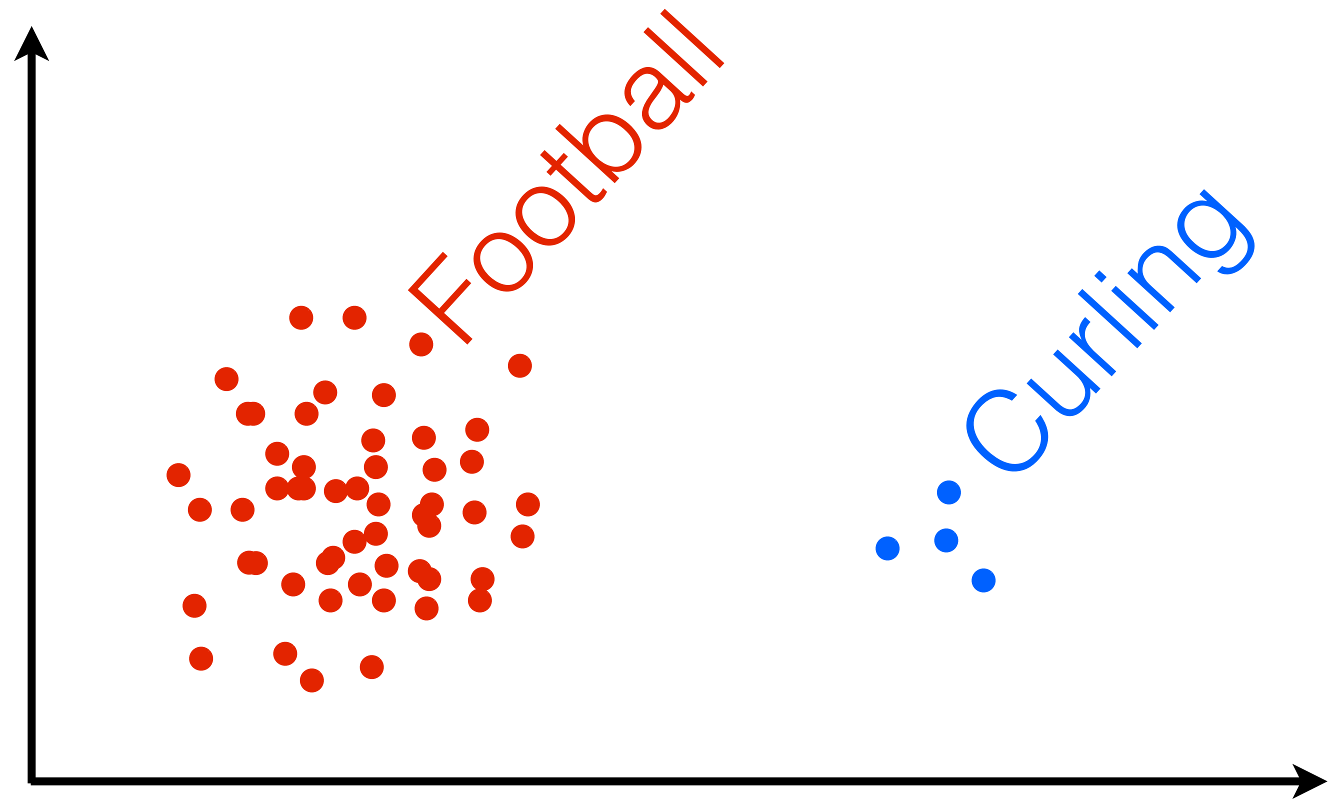
“Core” of the data set

- Observe: redundancies can exist even if data isn’t “tall”



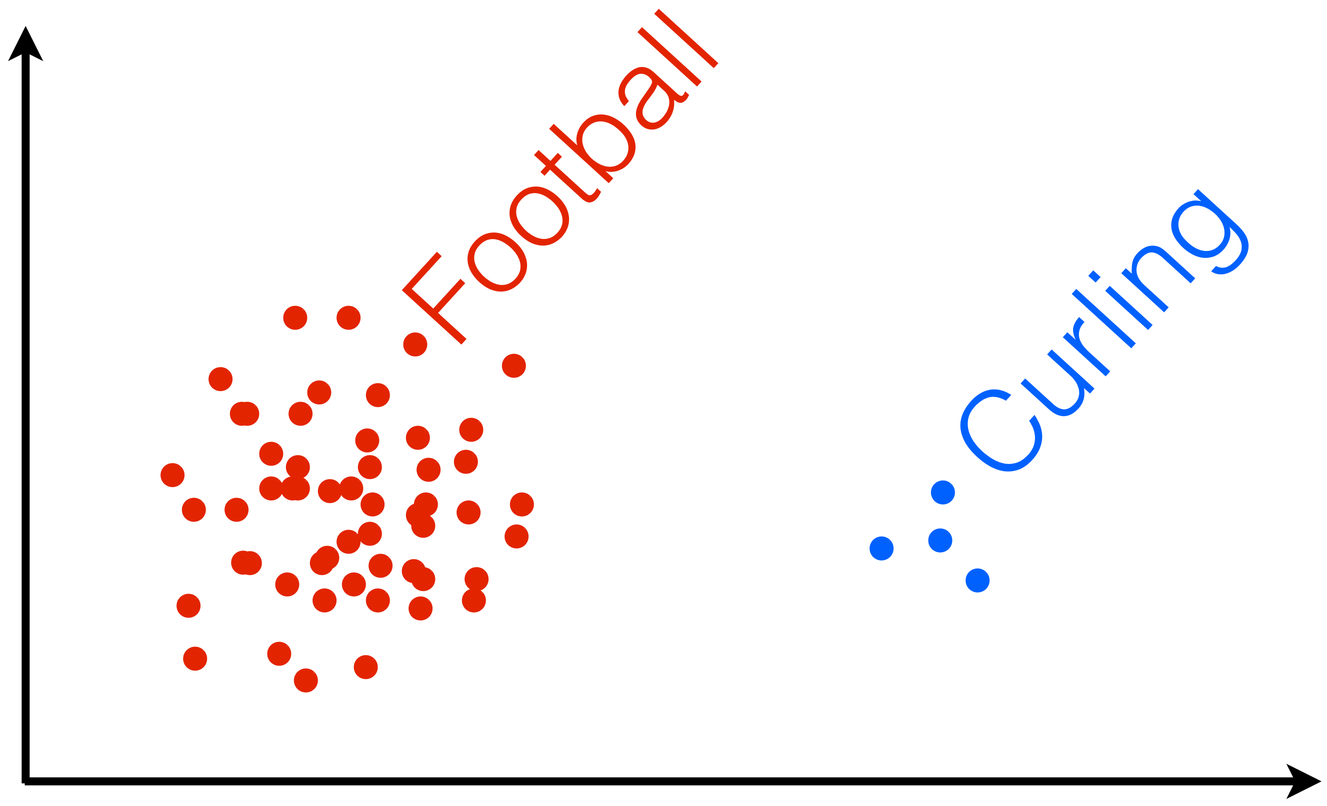
“Core” of the data set

- Observe: redundancies can exist even if data isn’t “tall”



"Core" of the data set

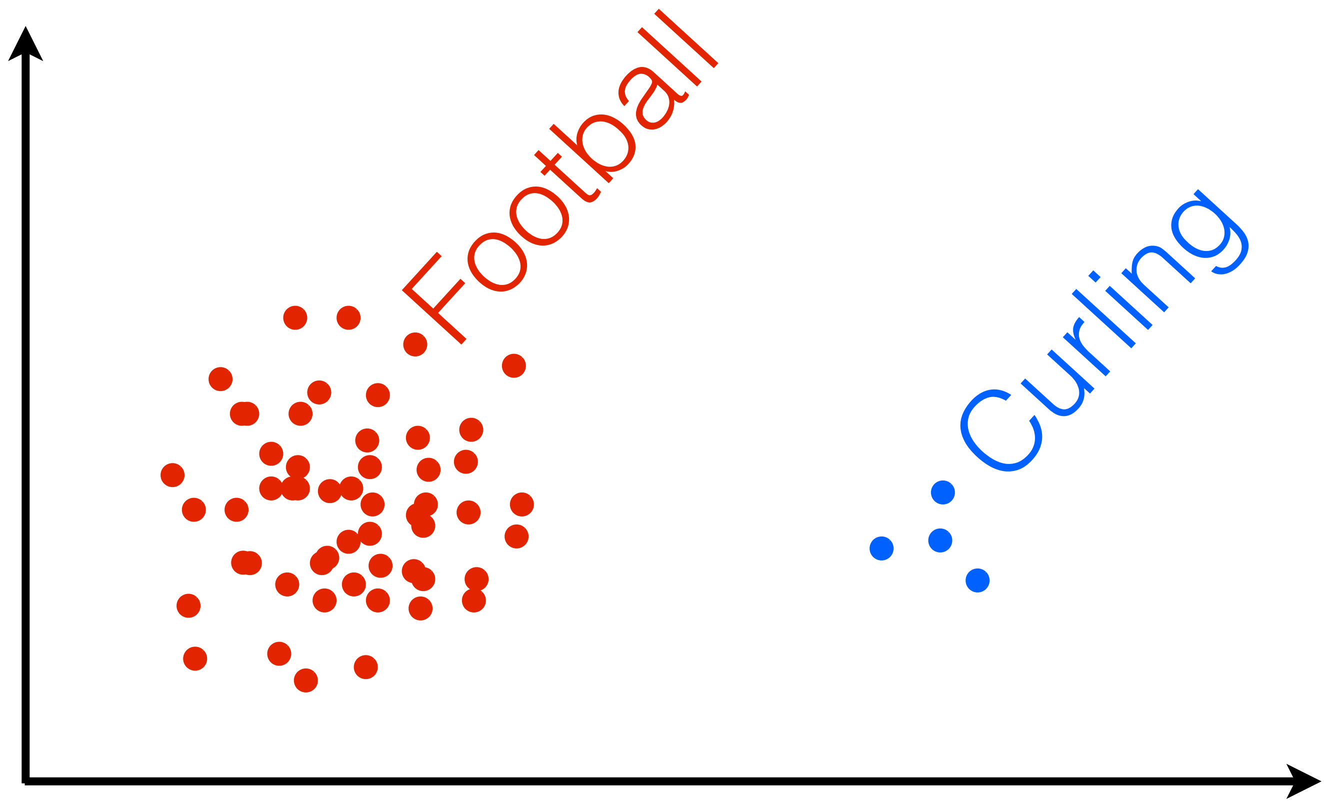
- Observe: redundancies can exist even if data isn't "tall"
- Coresets: pre-process data to get a smaller, weighted data set



[Bădoiu, Har-Peled, Indyk 2002; Agarwal et al 2005; Feldman & Langberg 2011]

"Core" of the data set

- Observe: redundancies can exist even if data isn't "tall"
- Coresets: pre-process data to get a smaller, weighted data set

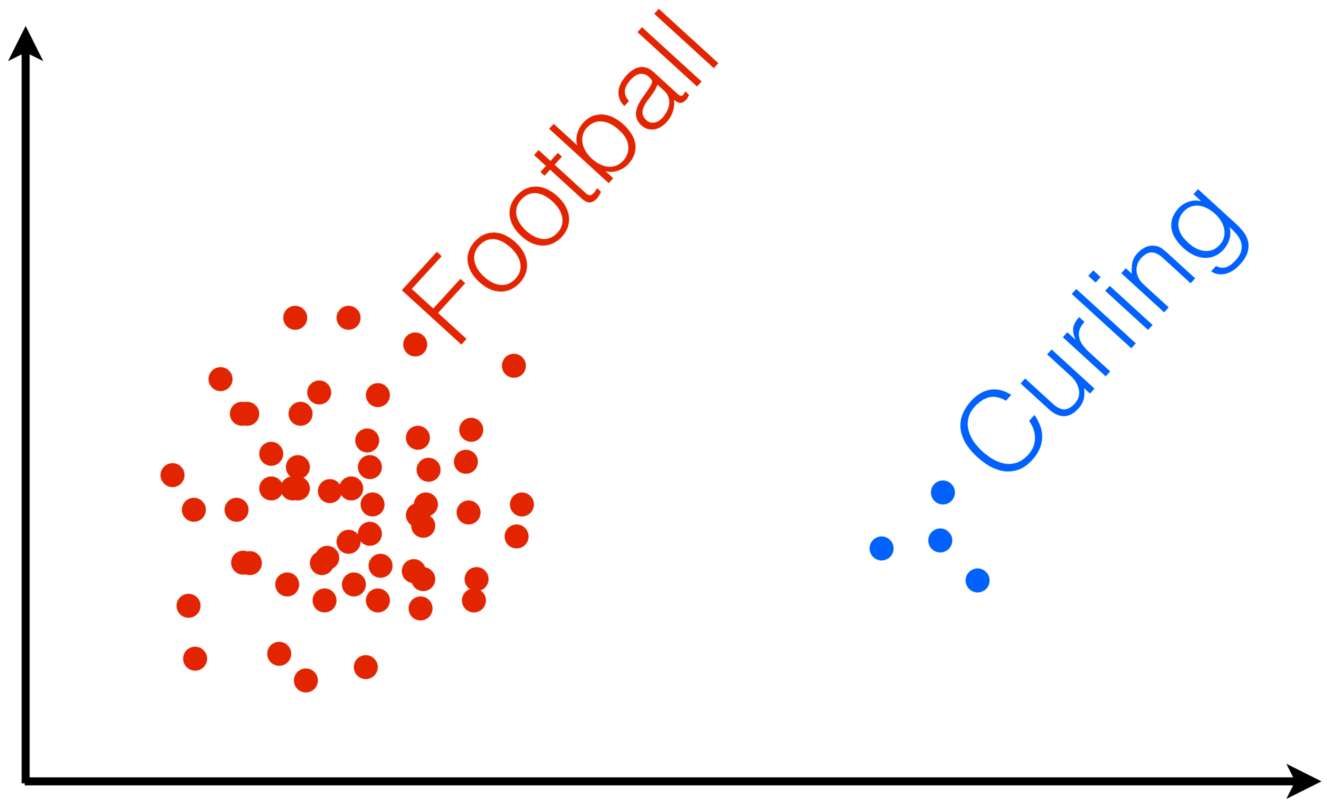


- Theoretical guarantees on quality

[Bădoiu, Har-Peled, Indyk 2002; Agarwal et al 2005; Feldman & Langberg 2011]

“Core” of the data set

- Observe: redundancies can exist even if data isn’t “tall”
- Coresets: pre-process data to get a smaller, weighted data set

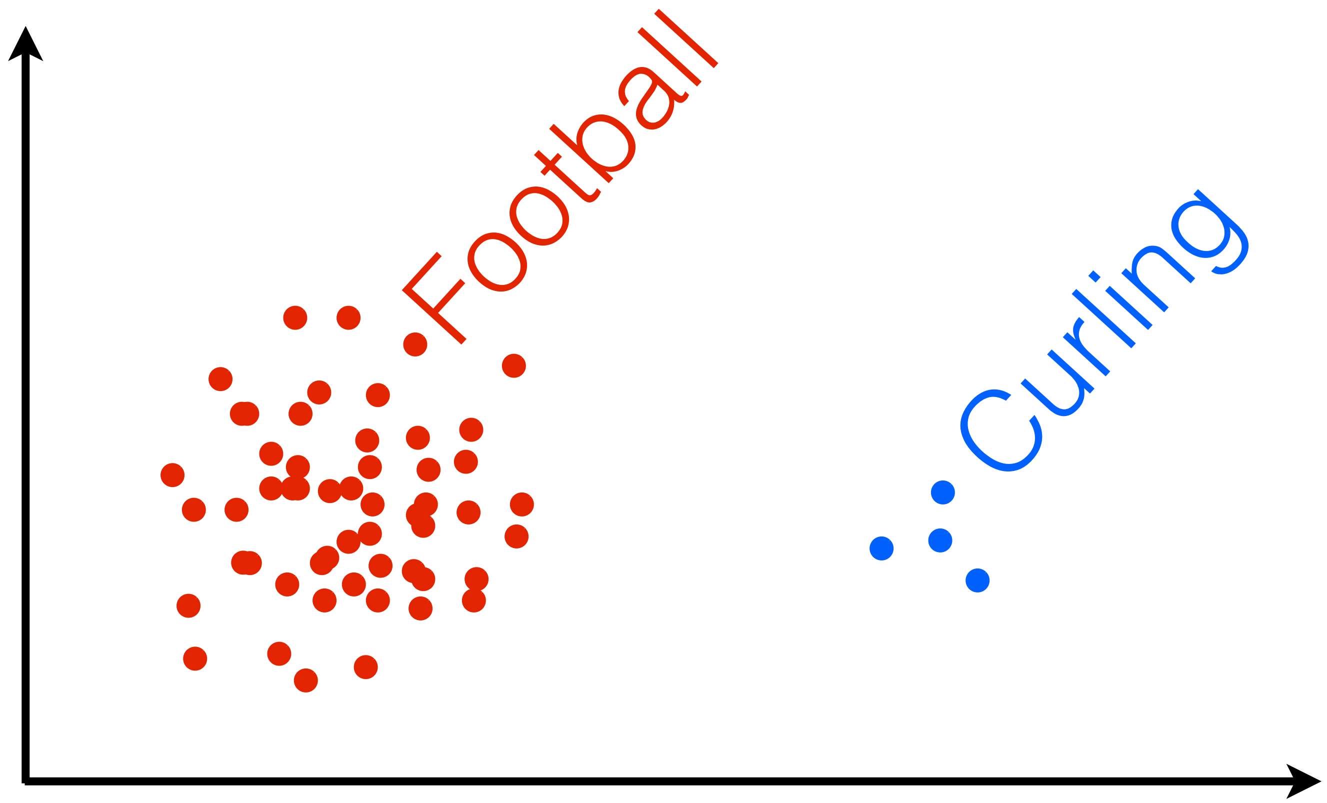


- Theoretical guarantees on quality
- How to develop **coresets for Bayes**?

[Bădoiu, Har-Peled, Indyk 2002; Agarwal et al 2005; Feldman & Langberg 2011; Huggins, Campbell, Broderick 2016; Campbell, Broderick 2019; Campbell, Broderick 2018; Agrawal, Campbell, Huggins, Broderick 2019]

"Core" of the data set

- Observe: redundancies can exist even if data isn't "tall"
- Coresets: pre-process data to get a smaller, weighted data set

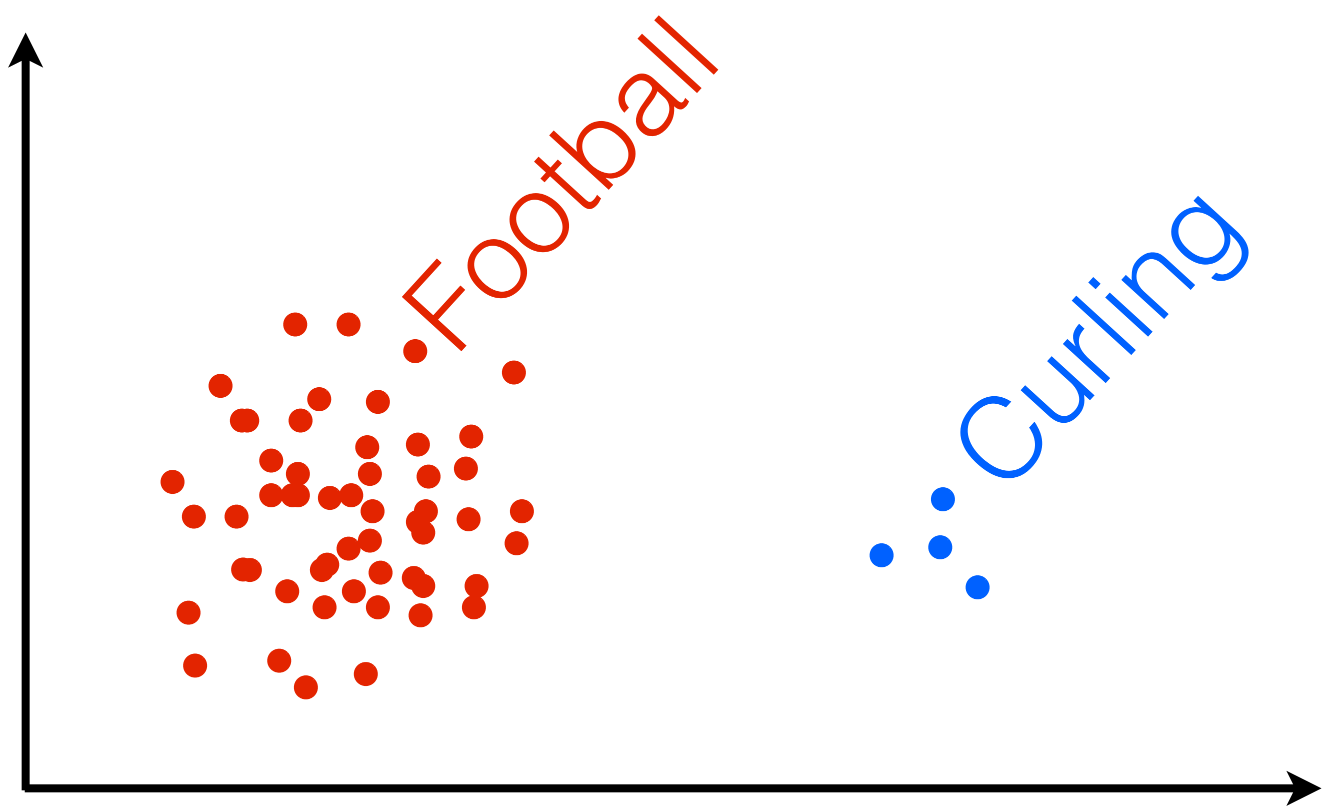


- Theoretical guarantees on quality
- How to develop **coresets for diverse tasks/geometries**?

[Bădoiu, Har-Peled, Indyk 2002; Agarwal et al 2005; Feldman & Langberg 2011; Huggins, Campbell, Broderick 2016; Campbell, Broderick 2019; Campbell, Broderick 2018; Agrawal, Campbell, Huggins, Broderick 2019]

“Core” of the data set

- Observe: redundancies can exist even if data isn’t “tall”
- Coresets: pre-process data to get a smaller, weighted data set

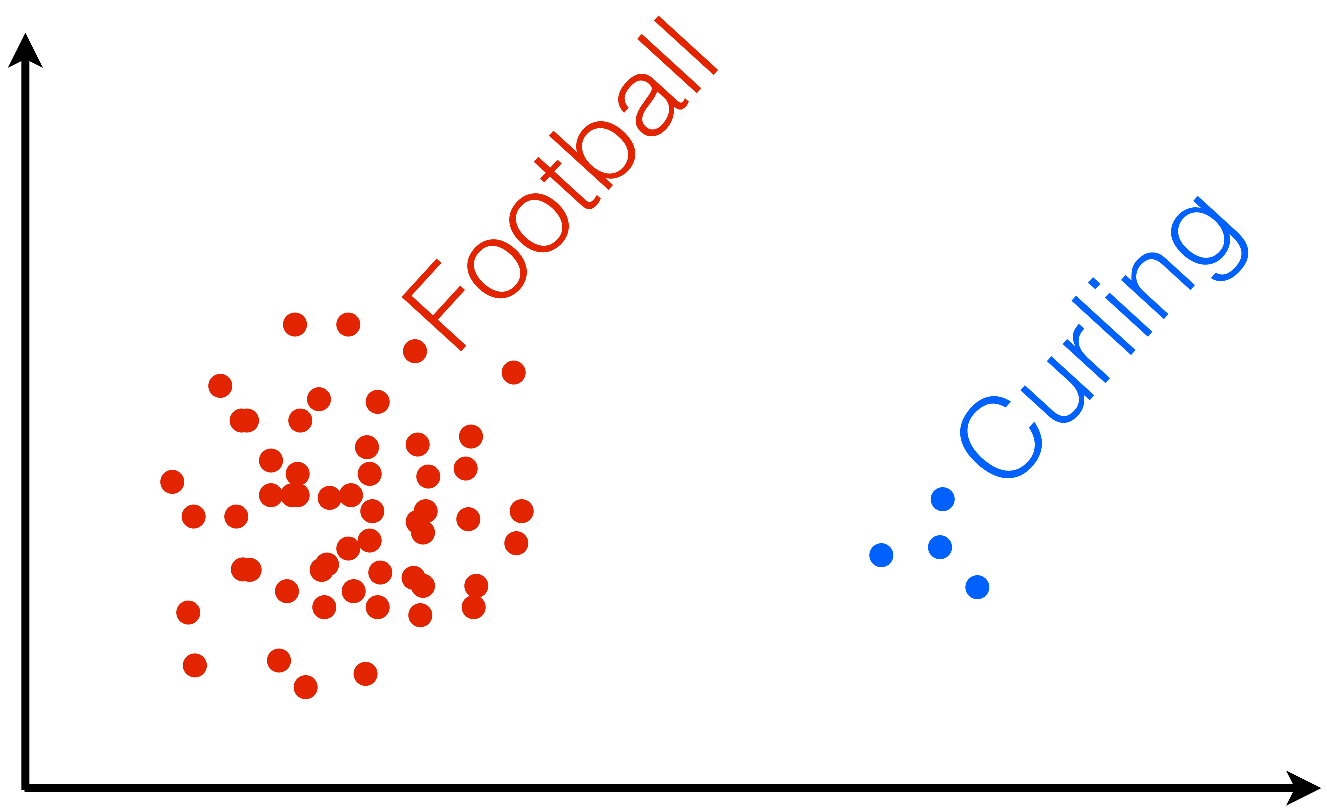


- Theoretical guarantees on quality
- How to develop **coresets for diverse tasks/geometries**?
- Previous heuristics: data squashing, big data GPs

[Bădoiu, Har-Peled, Indyk 2002; Agarwal et al 2005; Feldman & Langberg 2011; DuMouchel et al 1999; Madigan et al 2002; Huggins, Campbell, Broderick 2016; Campbell, Broderick 2019; Campbell, Broderick 2018; Agrawal, Campbell, Huggins, Broderick 2019]

"Core" of the data set

- Observe: redundancies can exist even if data isn't "tall"
- Coresets: pre-process data to get a smaller, weighted data set

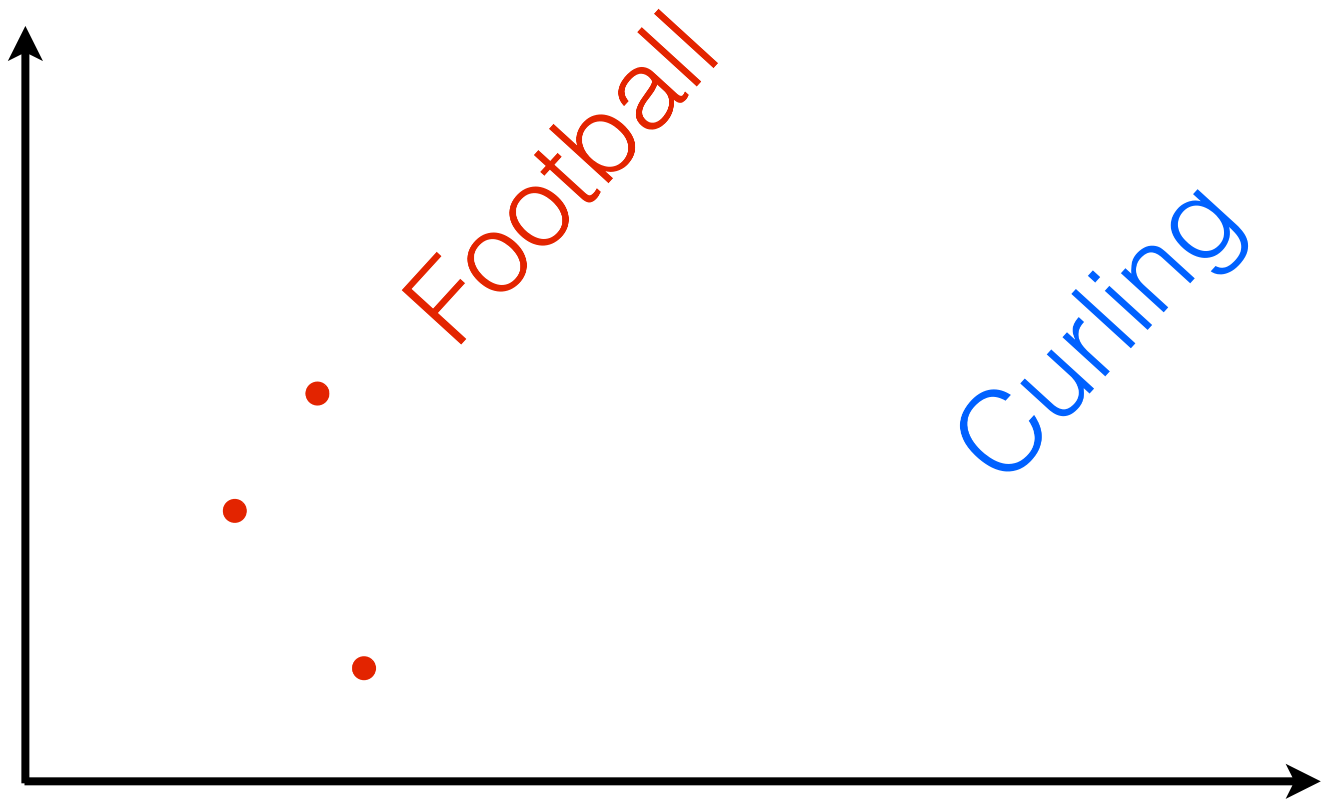


- Theoretical guarantees on quality
- How to develop **coresets for diverse tasks/geometries**?
- Previous heuristics: data squashing, big data GPs
- Compare to subsampling

[Bădoiu, Har-Peled, Indyk 2002; Agarwal et al 2005; Feldman & Langberg 2011; DuMouchel et al 1999; Madigan et al 2002; Huggins, Campbell, Broderick 2016; Campbell, Broderick 2019; Campbell, Broderick 2018; Agrawal, Campbell, Huggins, Broderick 2019]

“Core” of the data set

- Observe: redundancies can exist even if data isn’t “tall”
- Coresets: pre-process data to get a smaller, weighted data set

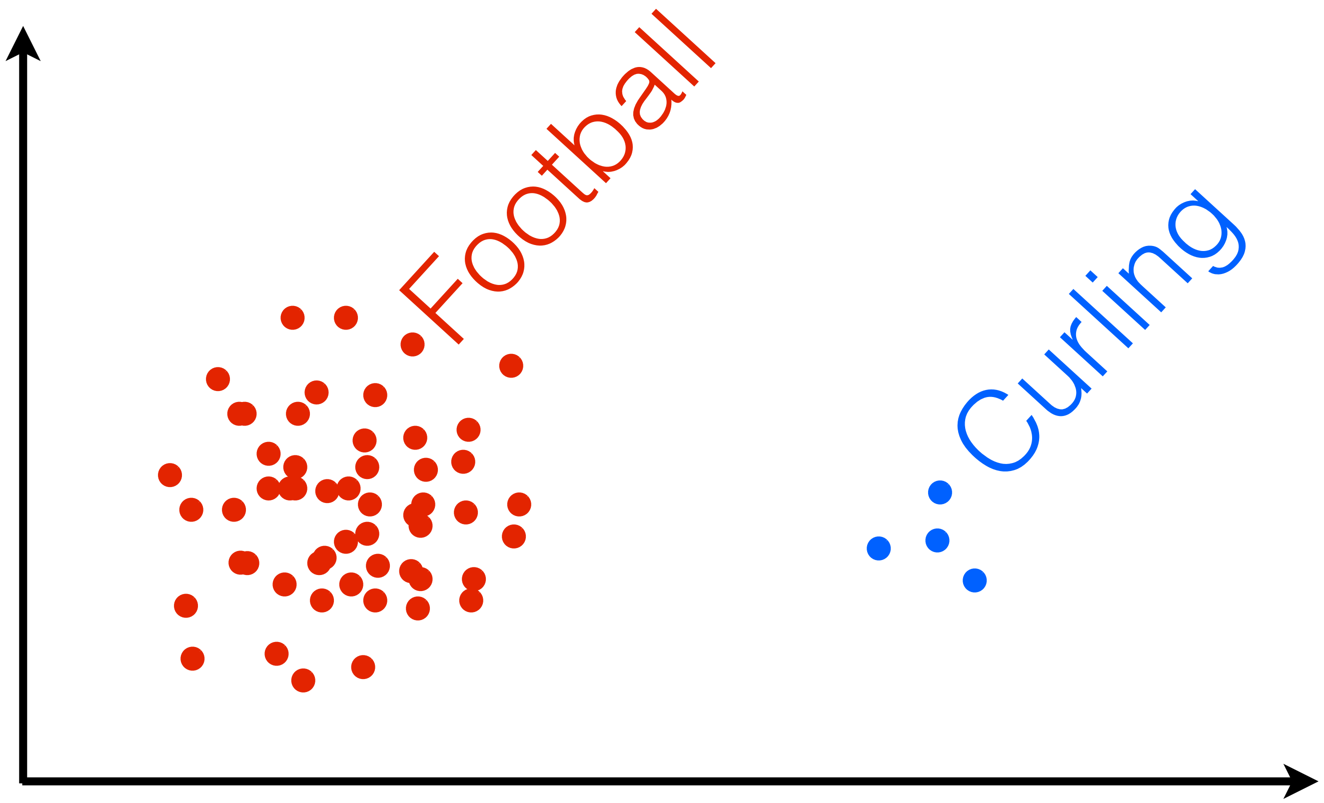


- Theoretical guarantees on quality
- How to develop **coresets for diverse tasks/geometries**?
- Previous heuristics: data squashing, big data GPs
- Compare to subsampling

[Bădoiu, Har-Peled, Indyk 2002; Agarwal et al 2005; Feldman & Langberg 2011; DuMouchel et al 1999; Madigan et al 2002; Huggins, Campbell, Broderick 2016; Campbell, Broderick 2019; Campbell, Broderick 2018; Agrawal, Campbell, Huggins, Broderick 2019]

“Core” of the data set

- Observe: redundancies can exist even if data isn’t “tall”
- Coresets: pre-process data to get a smaller, weighted data set

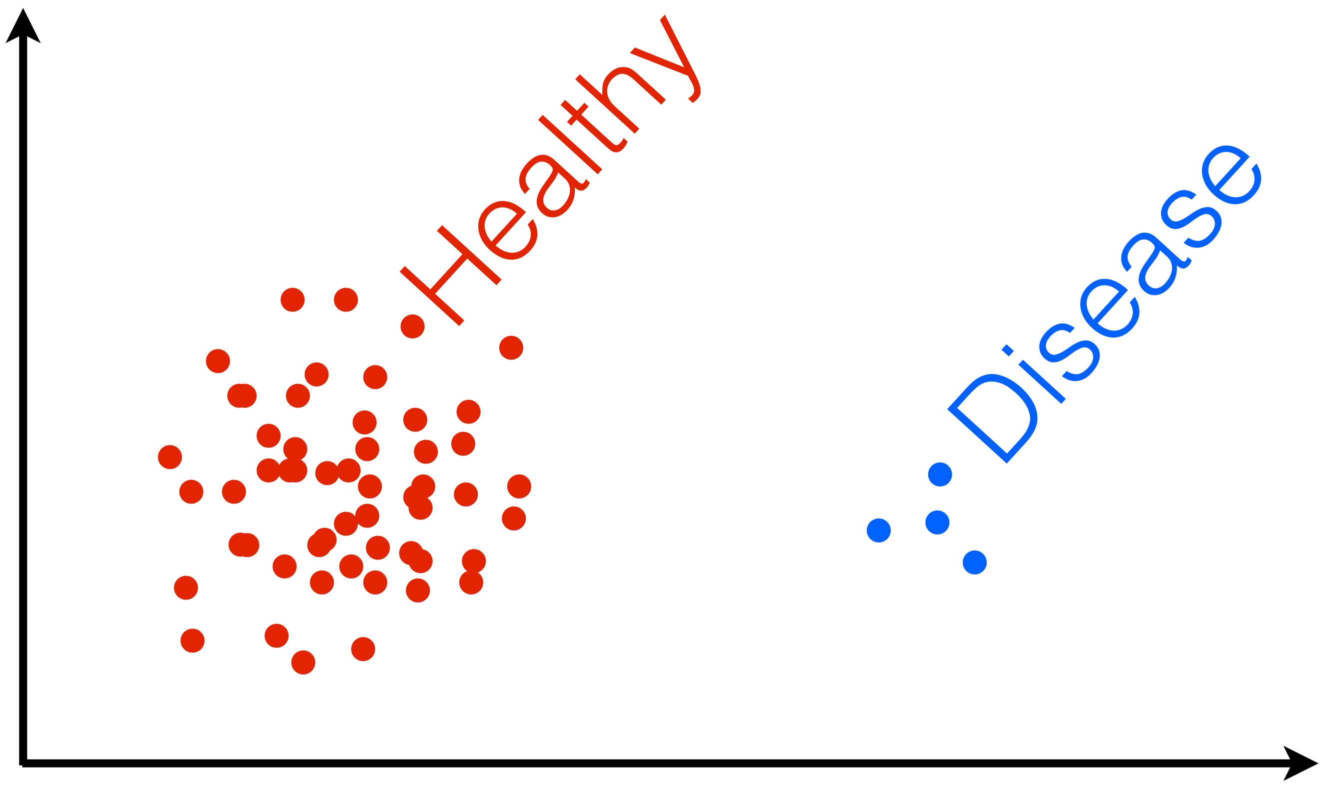


- Theoretical guarantees on quality
- How to develop **coresets for diverse tasks/geometries**?
- Previous heuristics: data squashing, big data GPs
- Compare to subsampling

[Bădoiu, Har-Peled, Indyk 2002; Agarwal et al 2005; Feldman & Langberg 2011; DuMouchel et al 1999; Madigan et al 2002; Huggins, Campbell, Broderick 2016; Campbell, Broderick 2019; Campbell, Broderick 2018; Agrawal, Campbell, Huggins, Broderick 2019]

“Core” of the data set

- Observe: redundancies can exist even if data isn’t “tall”
- Coresets: pre-process data to get a smaller, weighted data set

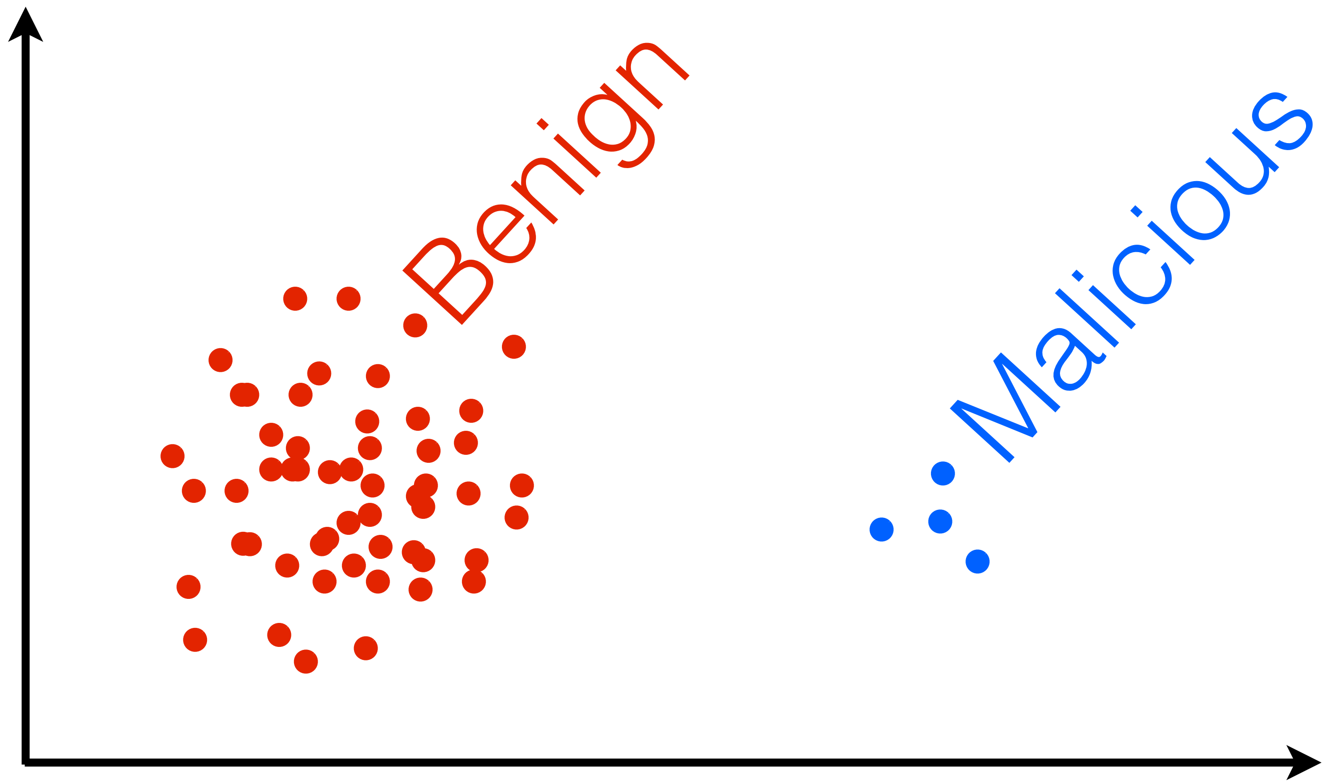


- Theoretical guarantees on quality
- How to develop **coresets for diverse tasks/geometries**?
- Previous heuristics: data squashing, big data GPs
- Compare to subsampling

[Bădoiu, Har-Peled, Indyk 2002; Agarwal et al 2005; Feldman & Langberg 2011; DuMouchel et al 1999; Madigan et al 2002; Huggins, Campbell, Broderick 2016; Campbell, Broderick 2019; Campbell, Broderick 2018; Agrawal, Campbell, Huggins, Broderick 2019]

“Core” of the data set

- Observe: redundancies can exist even if data isn’t “tall”
- Coresets: pre-process data to get a smaller, weighted data set

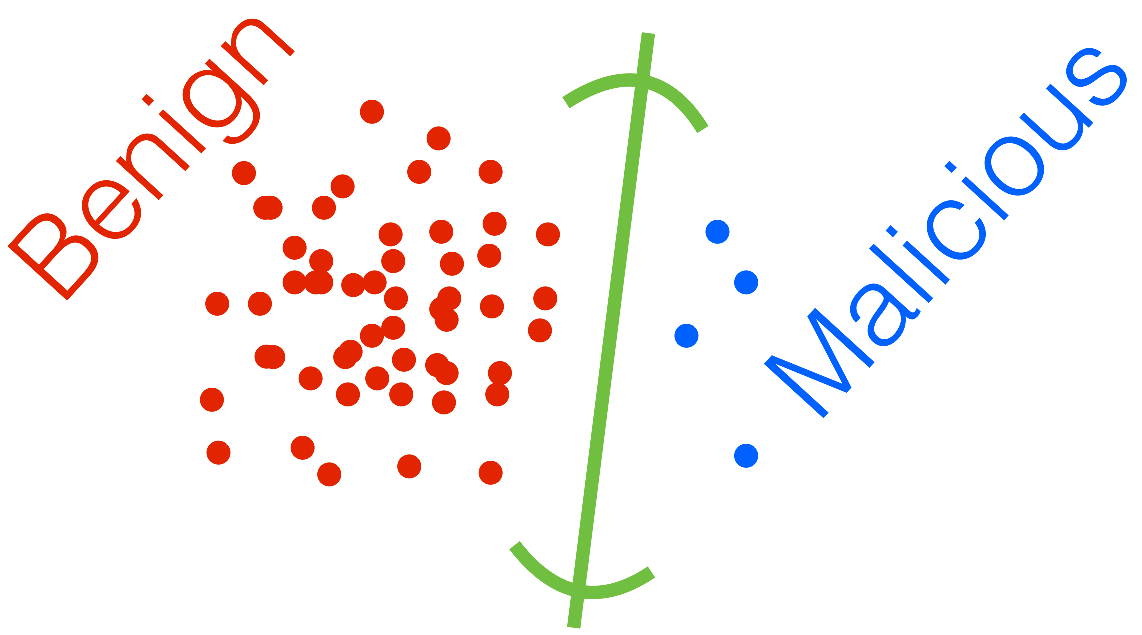


- Theoretical guarantees on quality
- How to develop **coresets for diverse tasks/geometries**?
- Previous heuristics: data squashing, big data GPs
- Compare to subsampling

[Bădoiu, Har-Peled, Indyk 2002; Agarwal et al 2005; Feldman & Langberg 2011; DuMouchel et al 1999; Madigan et al 2002; Huggins, Campbell, Broderick 2016; Campbell, Broderick 2019; Campbell, Broderick 2018; Agrawal, Campbell, Huggins, Broderick 2019]

Bayesian coresets

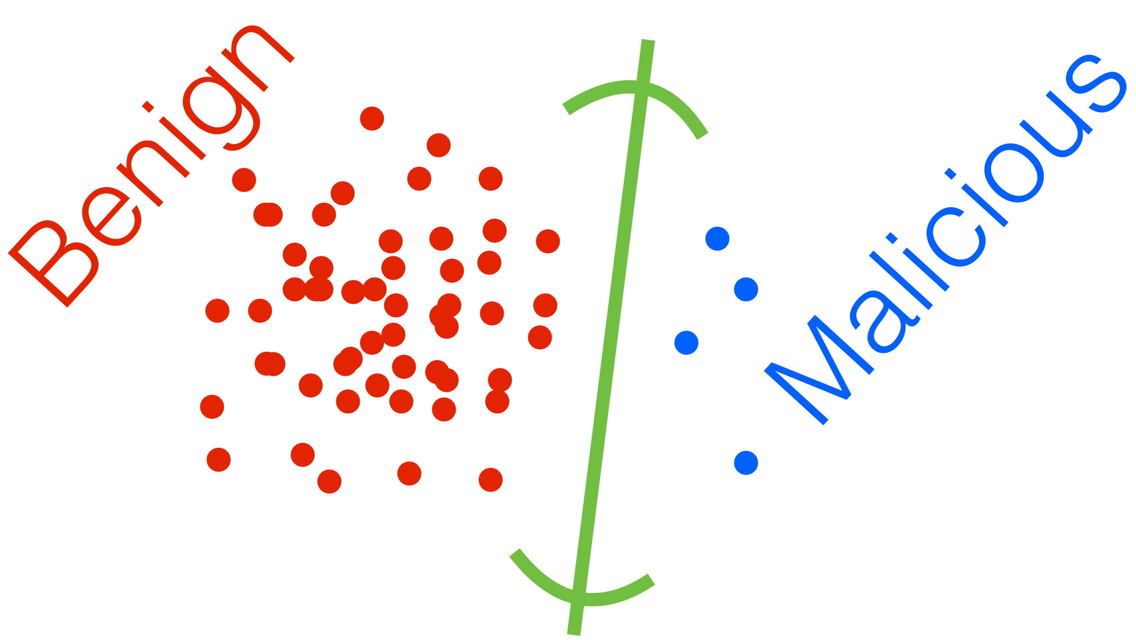
- Posterior $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$



[Huggins, Campbell, Broderick 2016]

Bayesian coresets

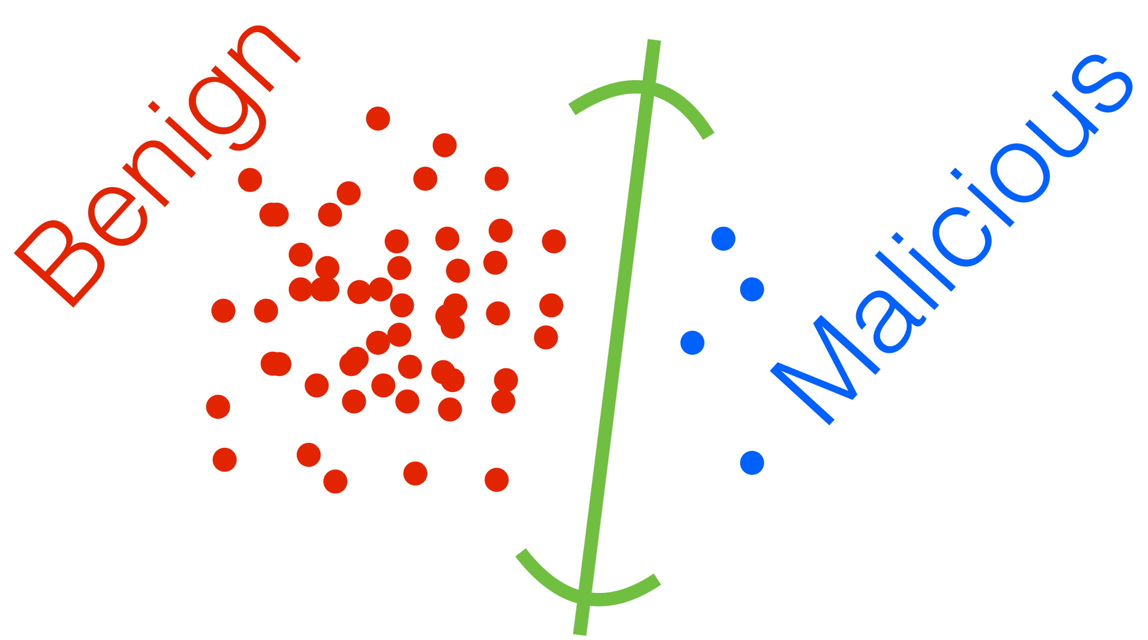
- Posterior $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$
- Log likelihood $\mathcal{L}_n(\theta) := \log p(y_n|\theta)\,, \;\; \mathcal{L}(\theta) := \sum_{n=1}^{N} \mathcal{L}_n(\theta)$



[Huggins, Campbell, Broderick 2016]

Bayesian coresets

- Posterior $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$
- Log likelihood $\mathcal{L}_n(\theta) := \log p(y_n|\theta)\,, \;\; \mathcal{L}(\theta) := \sum_{n=1}^{N} \mathcal{L}_n(\theta)$
- Coreset log likelihood

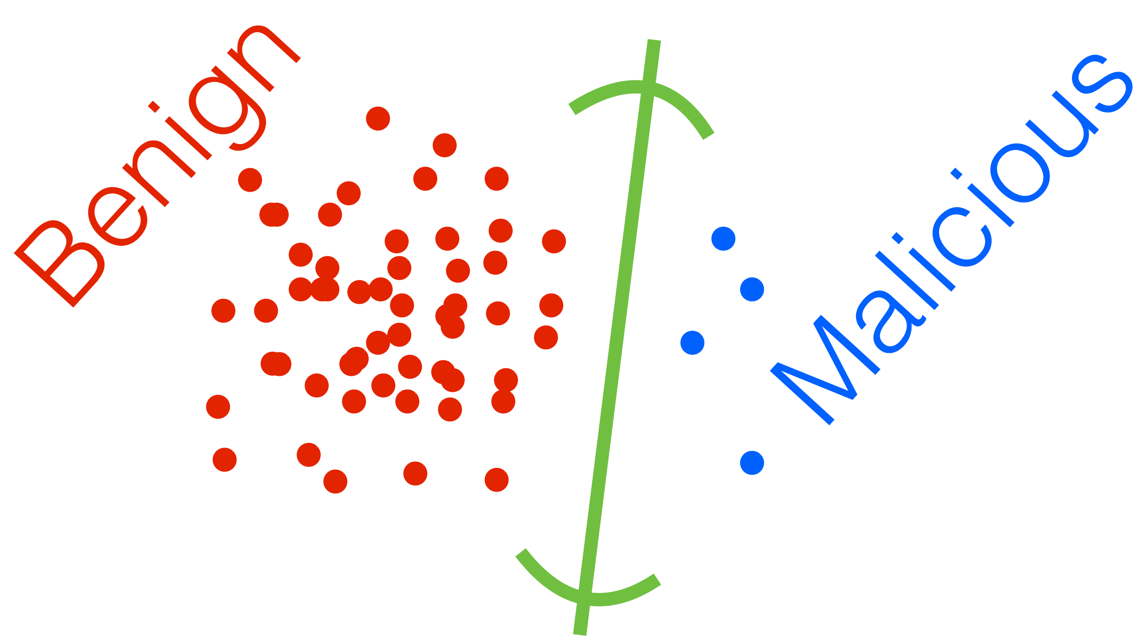


[Huggins, Campbell, Broderick 2016]

Bayesian coresets

- Posterior $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$
- Log likelihood $\mathcal{L}_n(\theta) := \log p(y_n|\theta)\,,\ \ \mathcal{L}(\theta) := \sum_{n=1}^{N} \mathcal{L}_n(\theta)$
- Coreset log likelihood

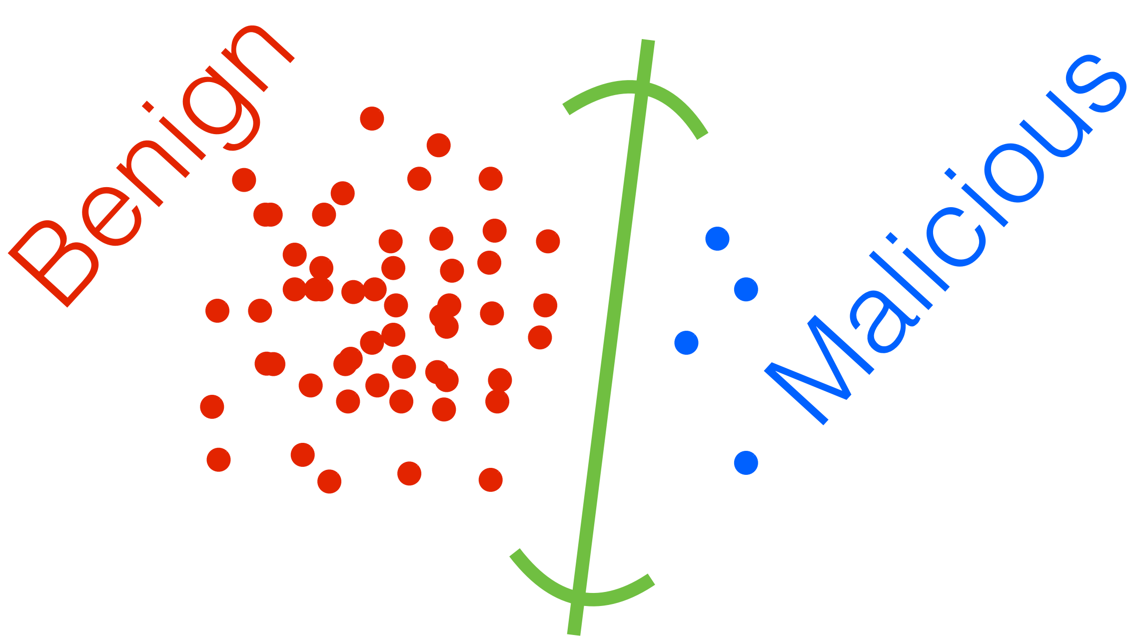
$$\|w\|_0 \ll N$$



[Huggins, Campbell, Broderick 2016]

Bayesian coresets

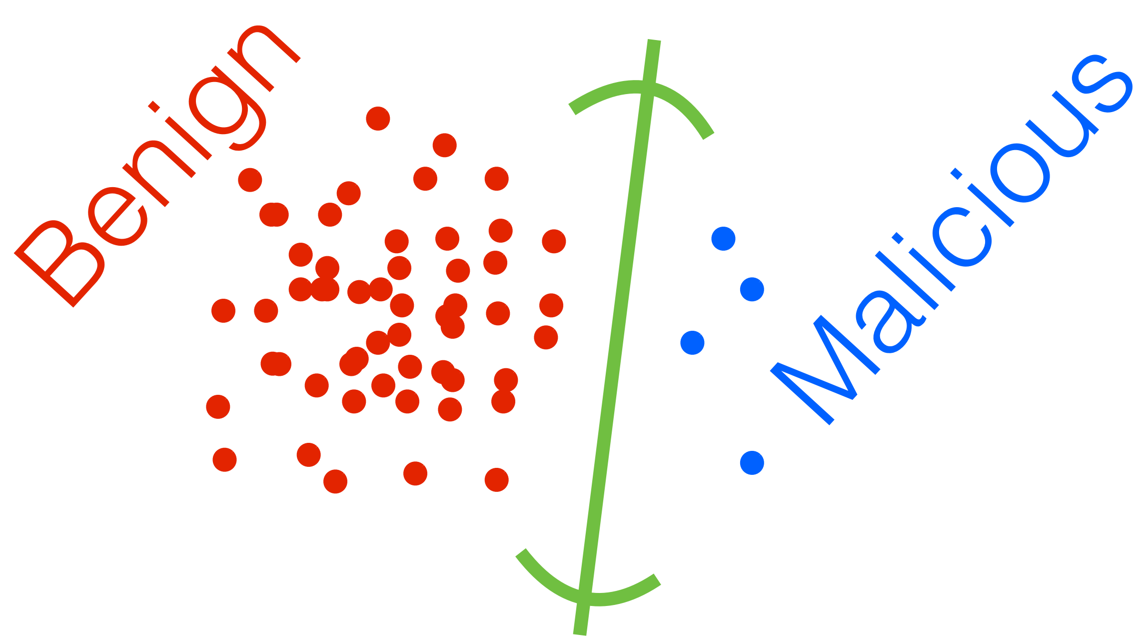
- Posterior $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$
- Log likelihood $\mathcal{L}_n(\theta) := \log p(y_n|\theta)\,, \;\; \mathcal{L}(\theta) := \sum_{n=1}^{N} \mathcal{L}_n(\theta)$
- Coreset log likelihood $\mathcal{L}(w,\theta) := \sum_{n=1}^{N} w_n \mathcal{L}_n(\theta)$ s.t. $\|w\|_0 \ll N$

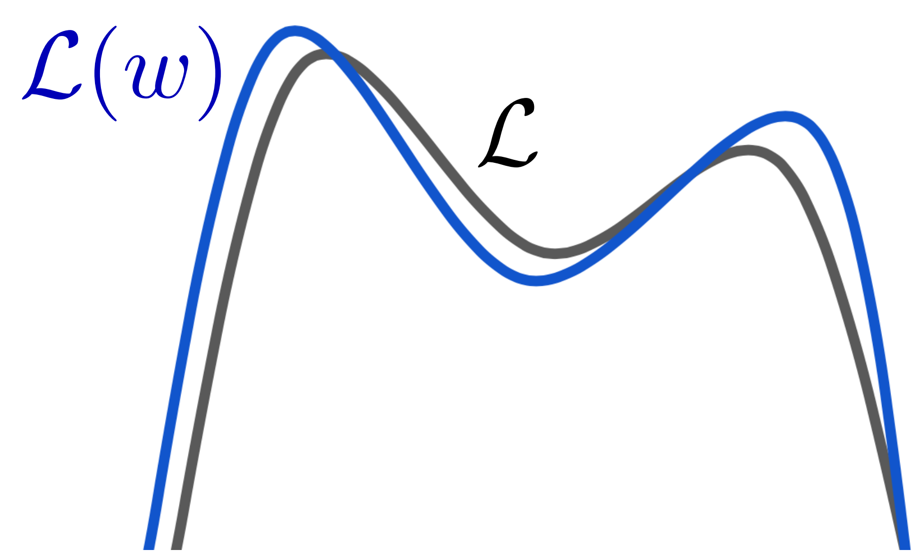


[Huggins, Campbell, Broderick 2016]

Bayesian coresets

- Posterior $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$
- Log likelihood $\mathcal{L}_n(\theta) := \log p(y_n|\theta)\,, \;\; \mathcal{L}(\theta) := \sum_{n=1}^{N} \mathcal{L}_n(\theta)$
- Coreset log likelihood $\mathcal{L}(w, \theta) := \sum_{n=1}^{N} w_n \mathcal{L}_n(\theta)$ s.t. $\|w\|_0 \ll N$

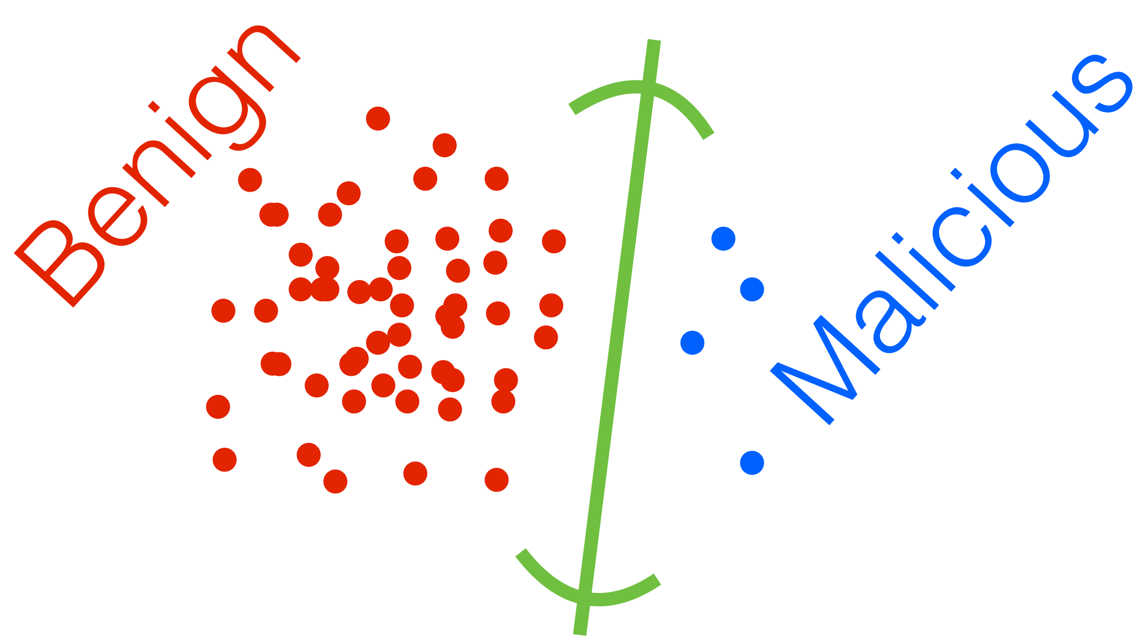


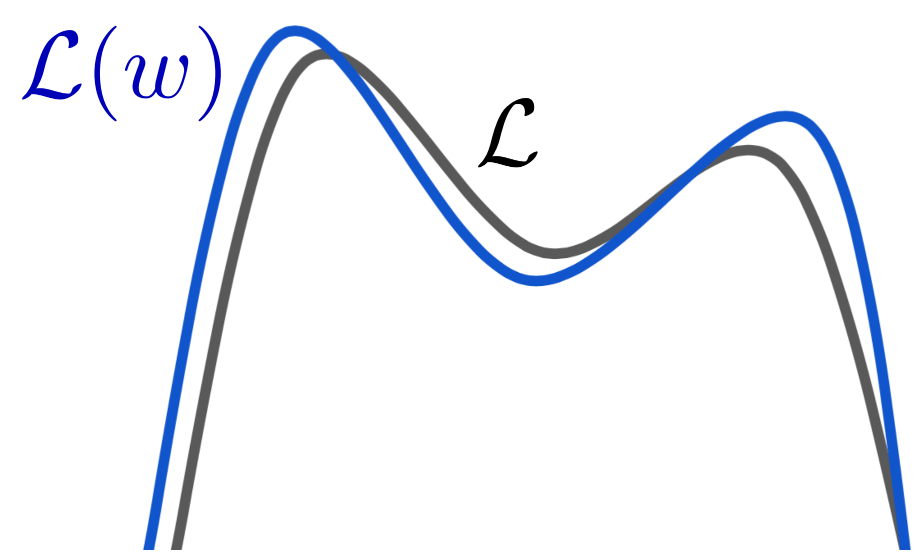


[Huggins, Campbell, Broderick 2016]

Bayesian coresets

- Posterior $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$
- Log likelihood $\mathcal{L}_n(\theta) := \log p(y_n|\theta)\,, \;\; \mathcal{L}(\theta) := \sum_{n=1}^{N} \mathcal{L}_n(\theta)$
- Coreset log likelihood $\mathcal{L}(w,\theta) := \sum_{n=1}^{N} w_n \mathcal{L}_n(\theta)$ s.t. $\|w\|_0 \ll N$
- ε-coreset: $\|\mathcal{L}(w) - \mathcal{L}\| \leq \epsilon$

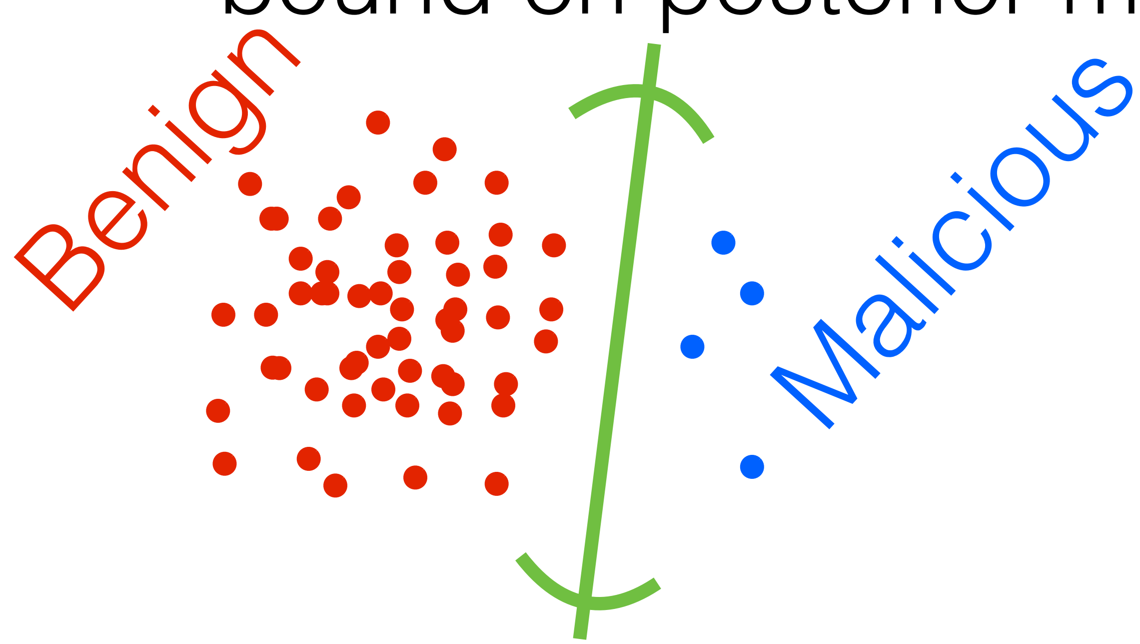


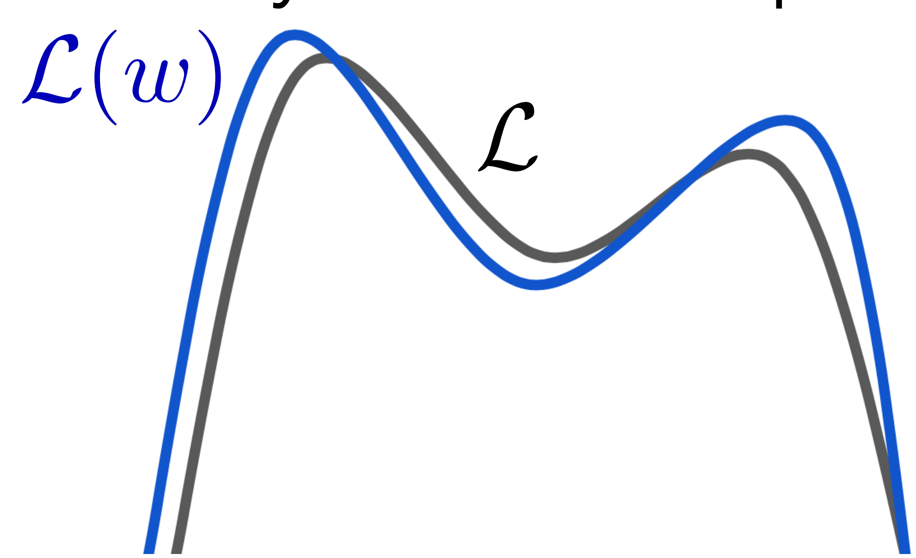


[Huggins, Campbell, Broderick 2016; Campbell, Broderick 2019]

Bayesian coresets

- Posterior $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$
- Log likelihood $\mathcal{L}_n(\theta) := \log p(y_n|\theta)$, $\mathcal{L}(\theta) := \sum_{n=1}^{N} \mathcal{L}_n(\theta)$
- Coreset log likelihood $\mathcal{L}(w, \theta) := \sum_{n=1}^{N} w_n \mathcal{L}_n(\theta)$ s.t. $\|w\|_0 \ll N$
- ε-coreset: $\|\mathcal{L}(w) - \mathcal{L}\| \leq \epsilon$
  - Bound on Wasserstein distance to exact posterior ➔ bound on posterior mean/uncertainty estimate quality





[Huggins, Campbell, Broderick 2016; Campbell, Broderick 2019; Huggins, Campbell, Kasprzak, Broderick 2018]

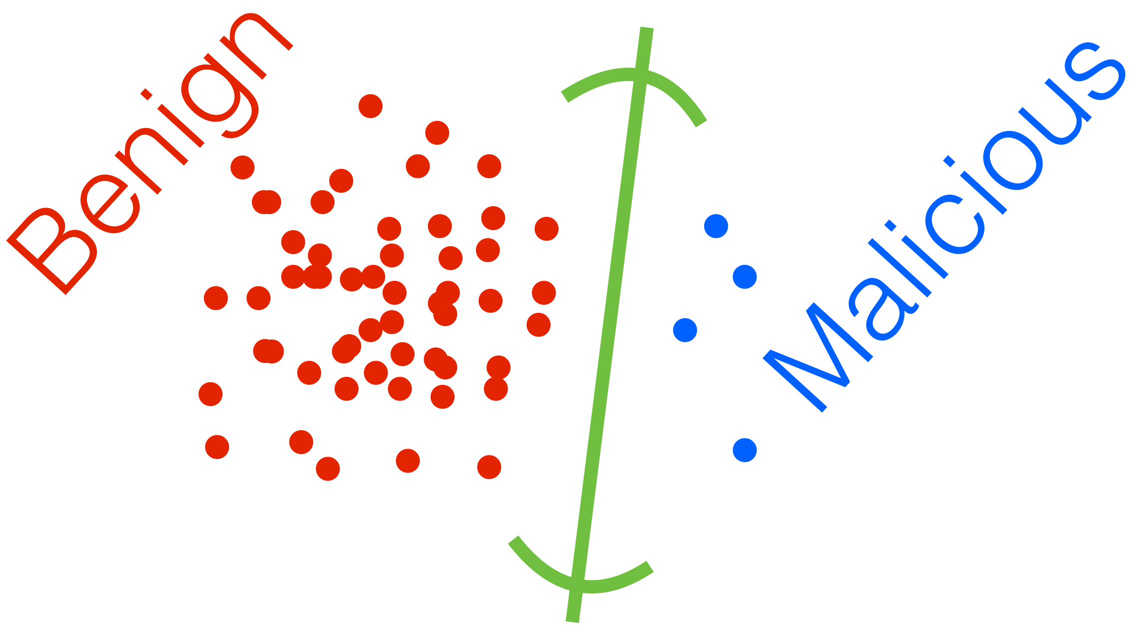
Roadmap

- The "core" of the data set
- Uniform data subsampling isn't enough
- Importance sampling for "coresets"
- Optimization for "coresets"
- Approximate sufficient statistics

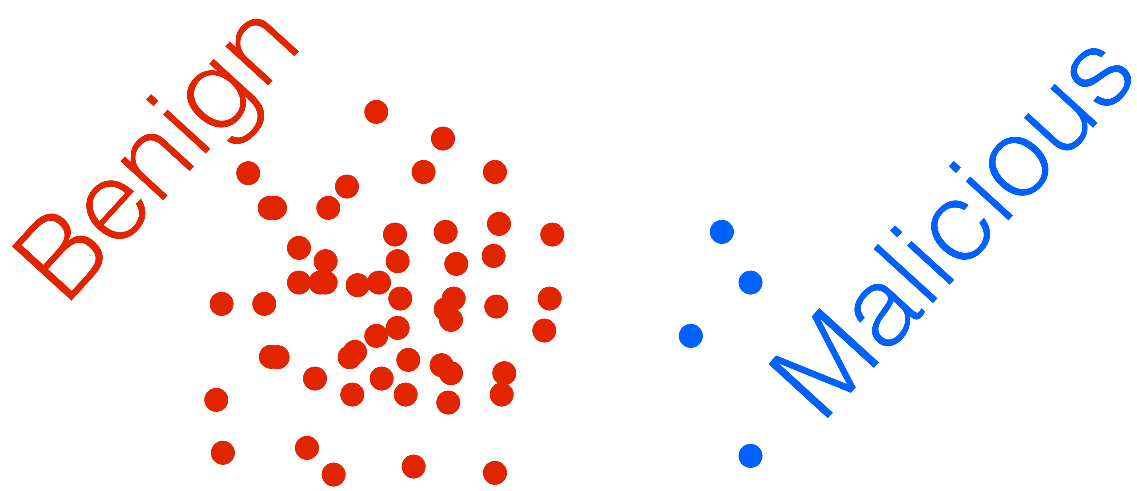
Roadmap

- The “core” of the data set
- Uniform data subsampling isn’t enough
- Importance sampling for “coresets”
- Optimization for “coresets”
- Approximate sufficient statistics

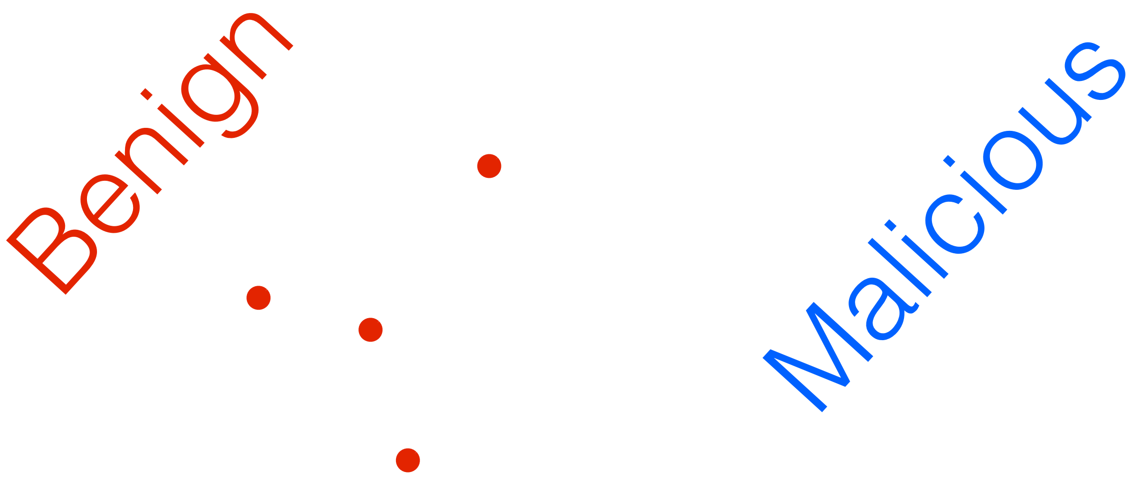
Uniform subsampling



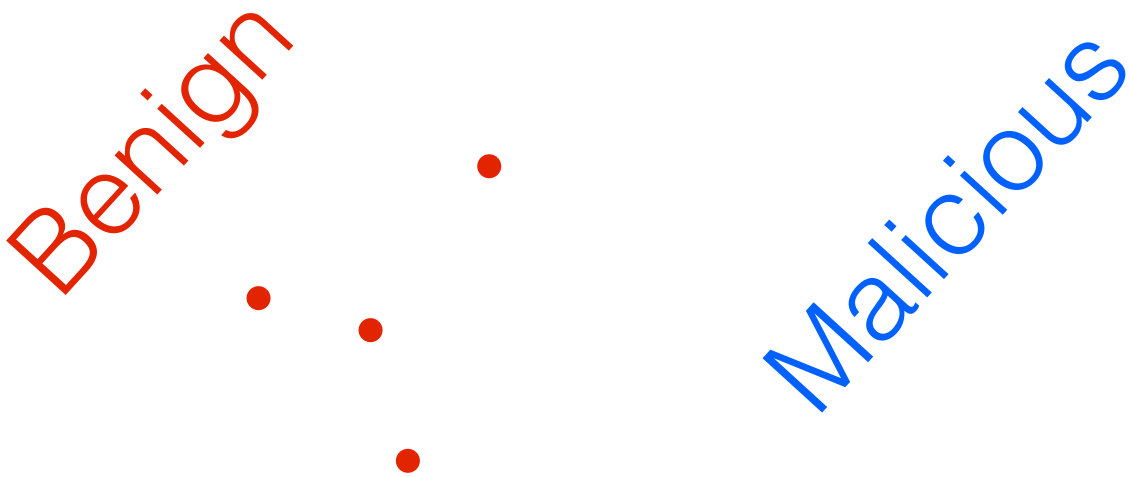
Uniform subsampling



Uniform subsampling

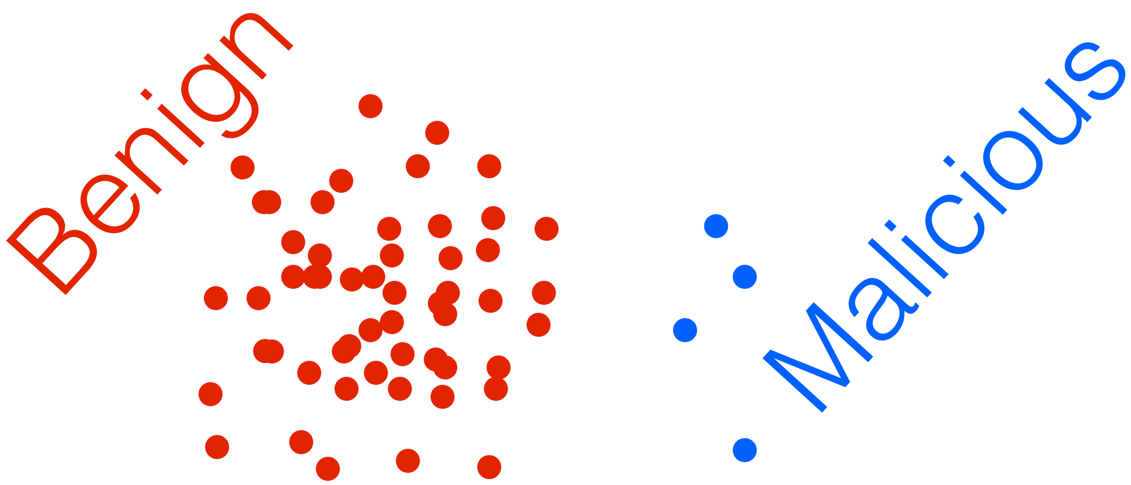


Uniform subsampling



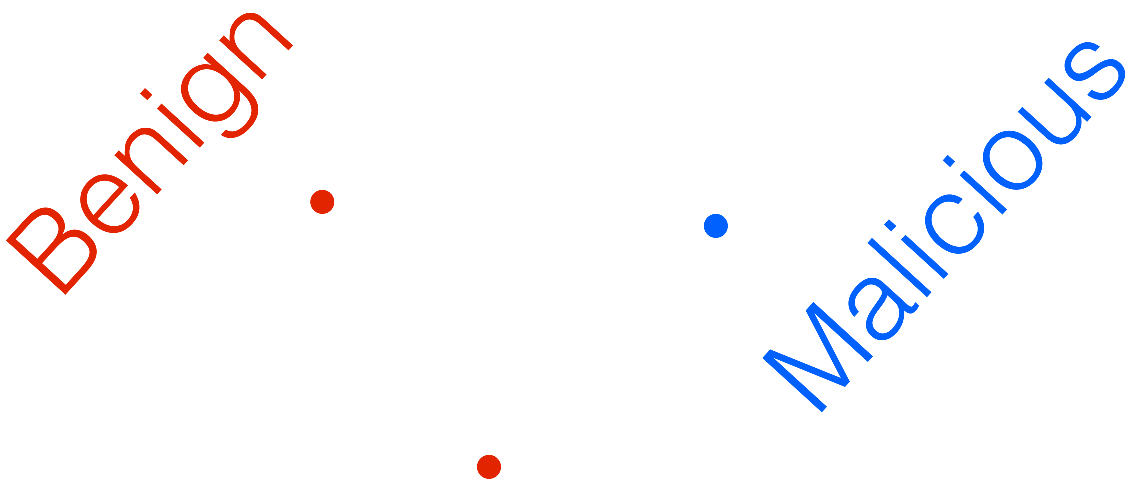
- Might miss important data

Uniform subsampling



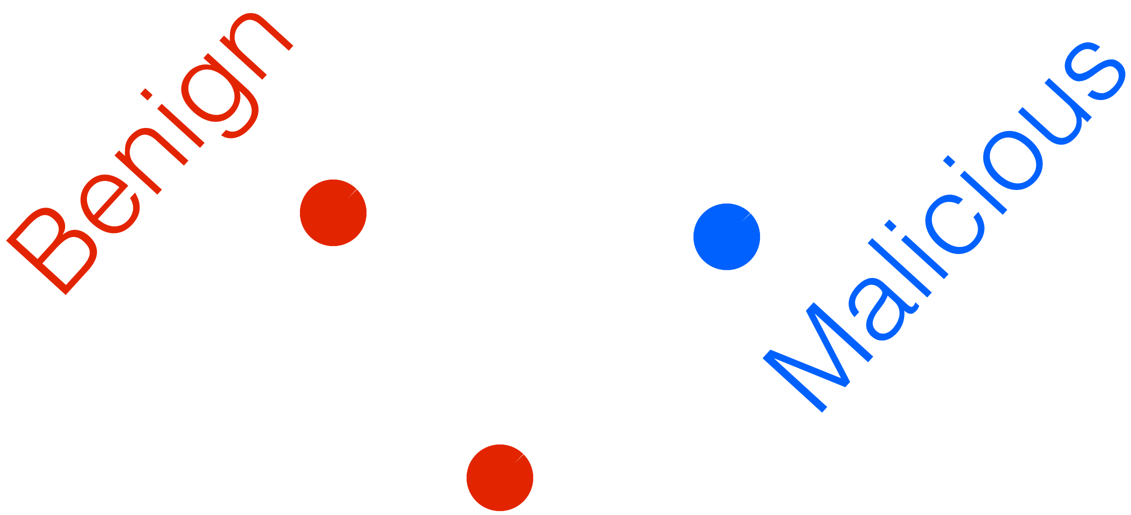
- Might miss important data

Uniform subsampling



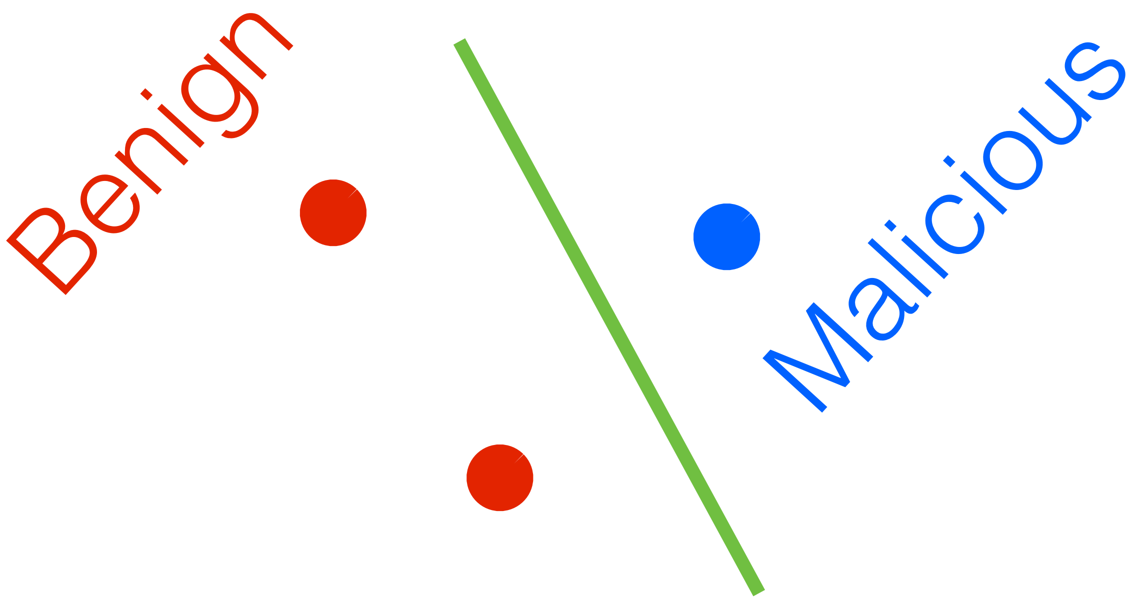
- Might miss important data

Uniform subsampling



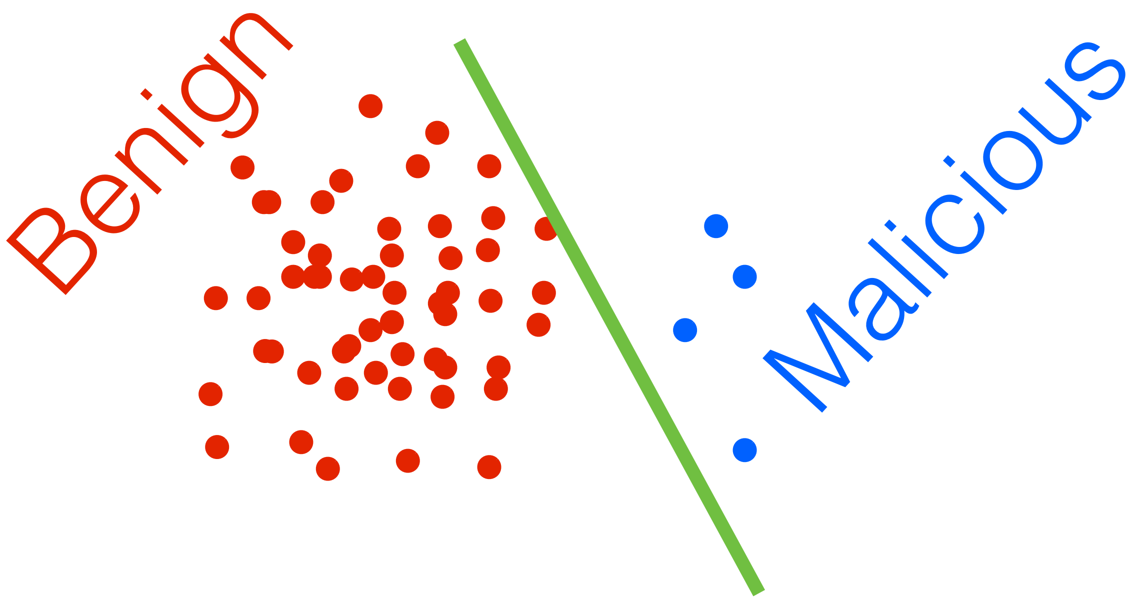
- Might miss important data

Uniform subsampling



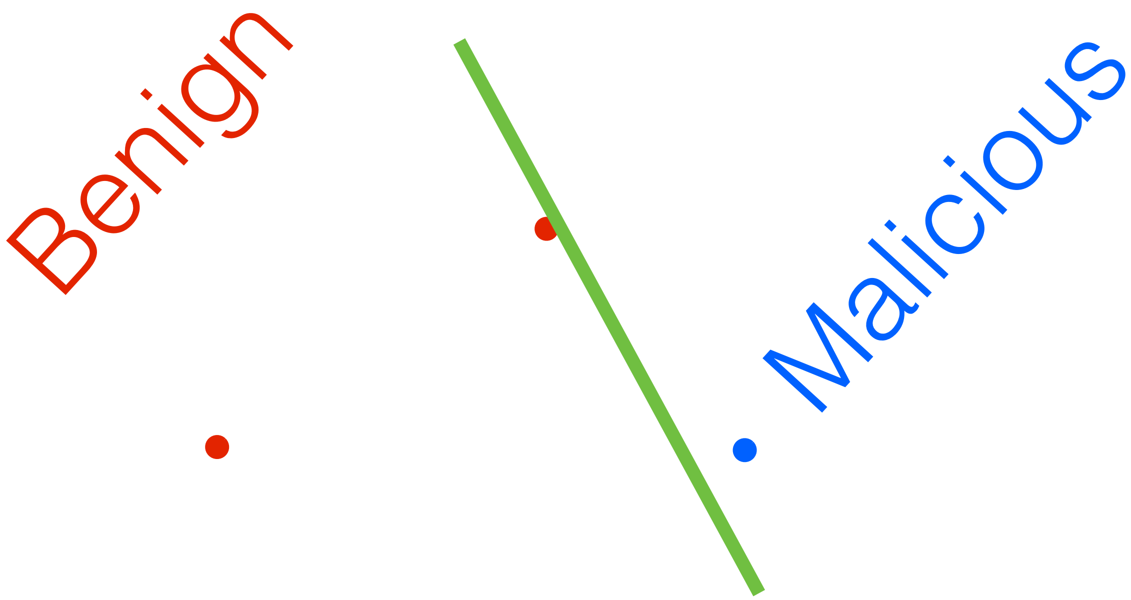
- Might miss important data

Uniform subsampling



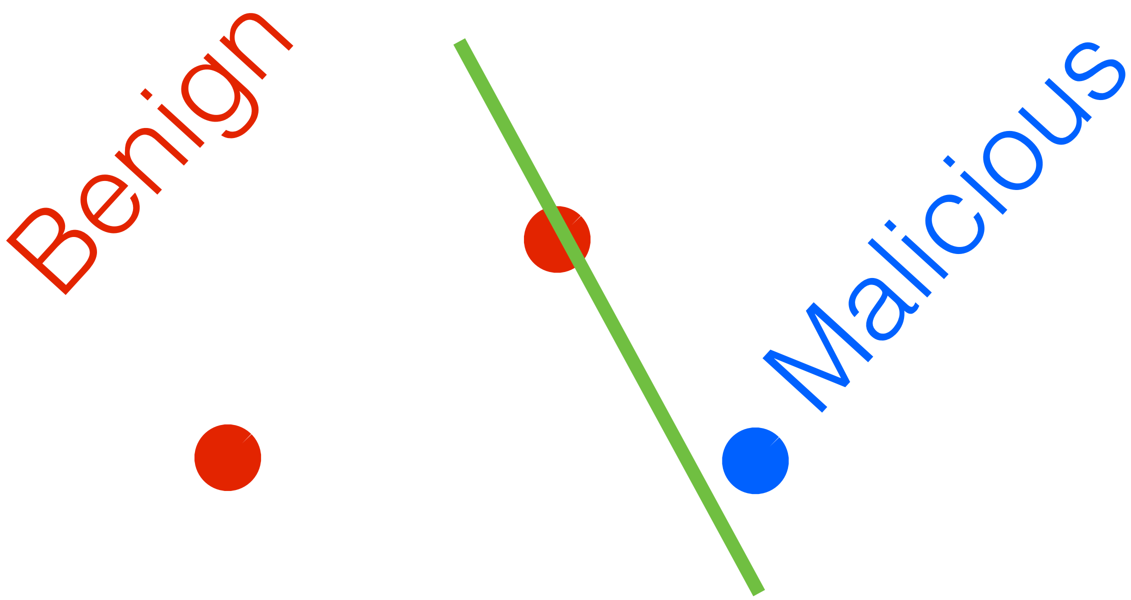
- Might miss important data

Uniform subsampling



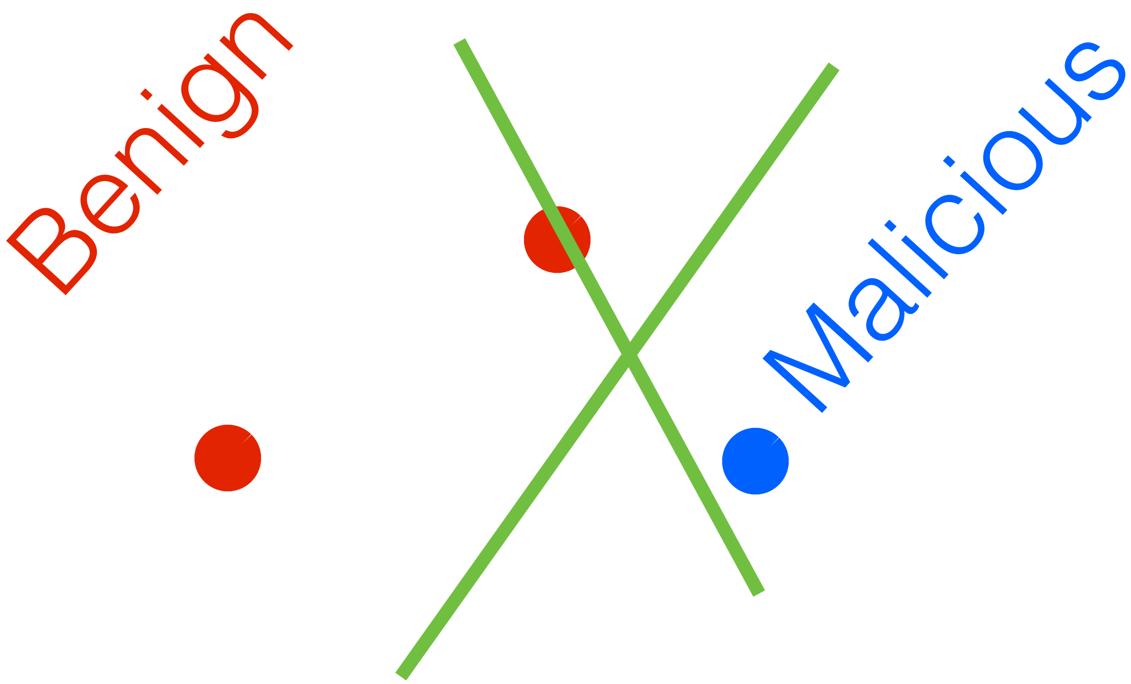
- Might miss important data

Uniform subsampling



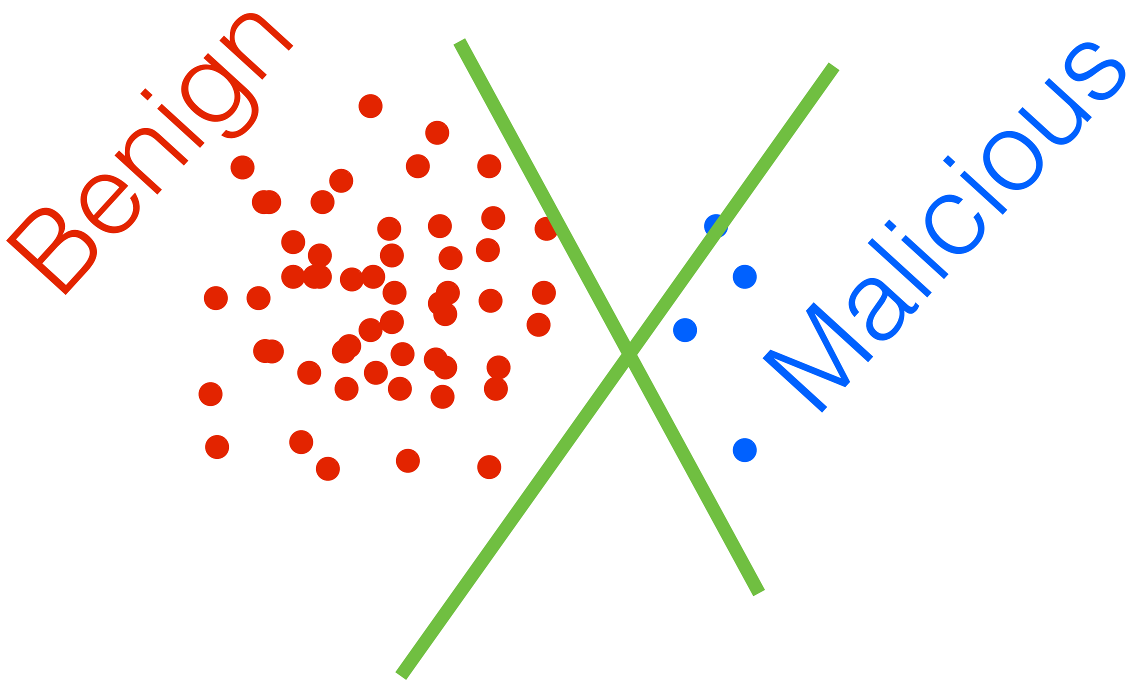
- Might miss important data

Uniform subsampling



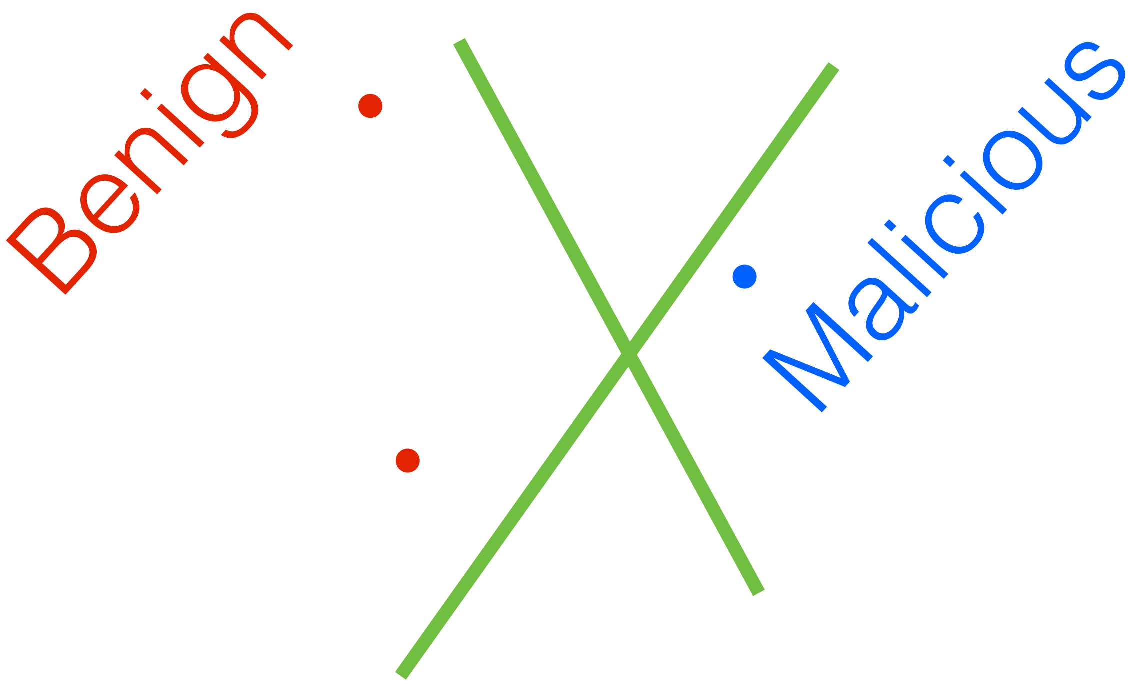
- Might miss important data

Uniform subsampling



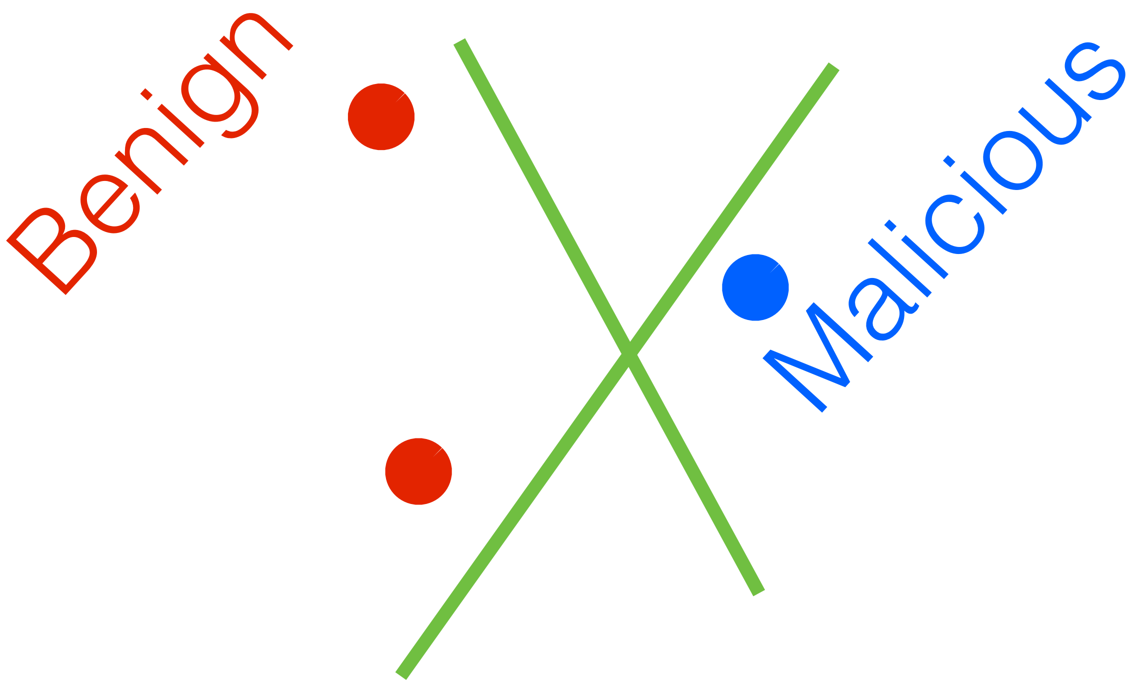
- Might miss important data

Uniform subsampling



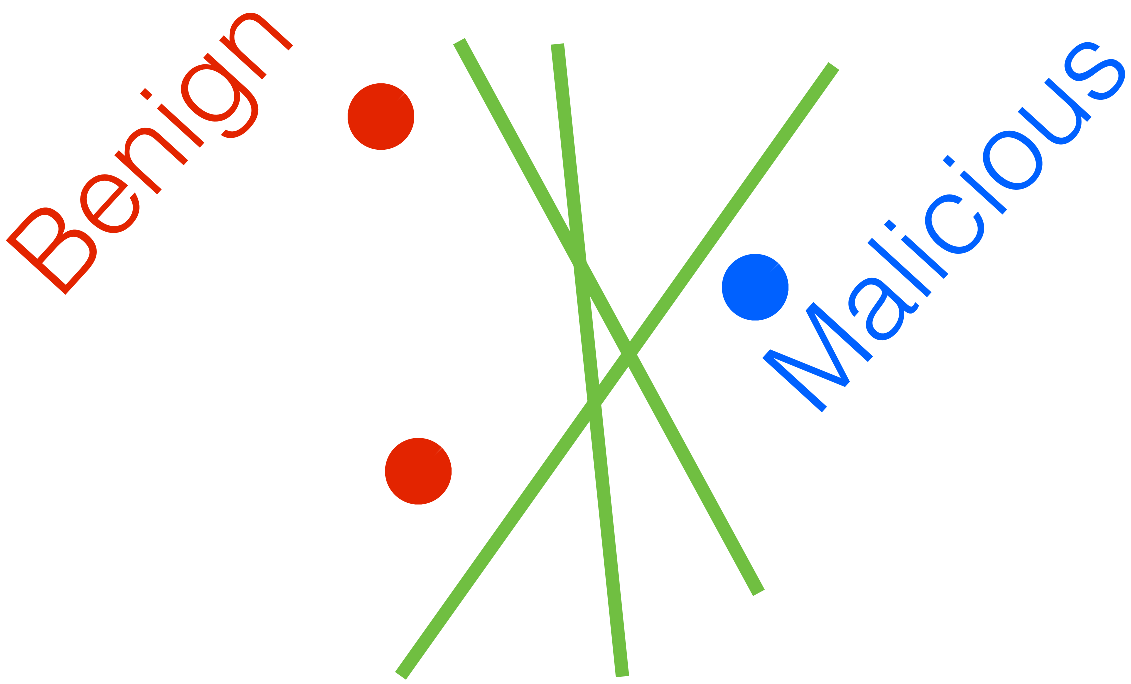
- Might miss important data

Uniform subsampling



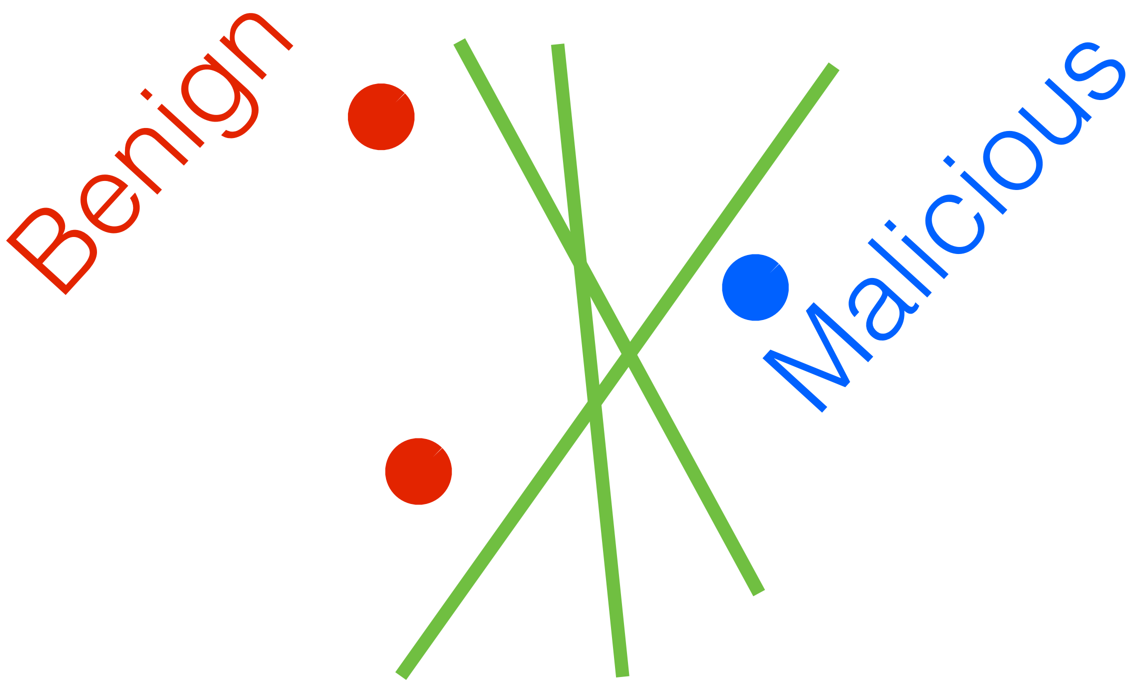
- Might miss important data

Uniform subsampling



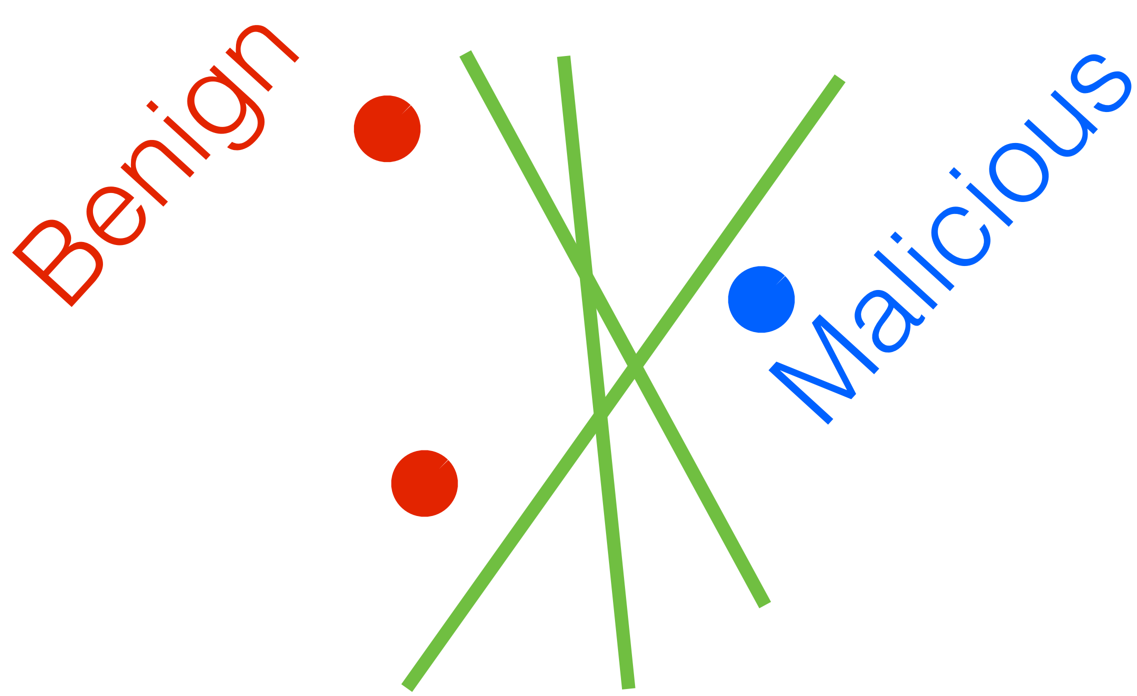
- Might miss important data

Uniform subsampling

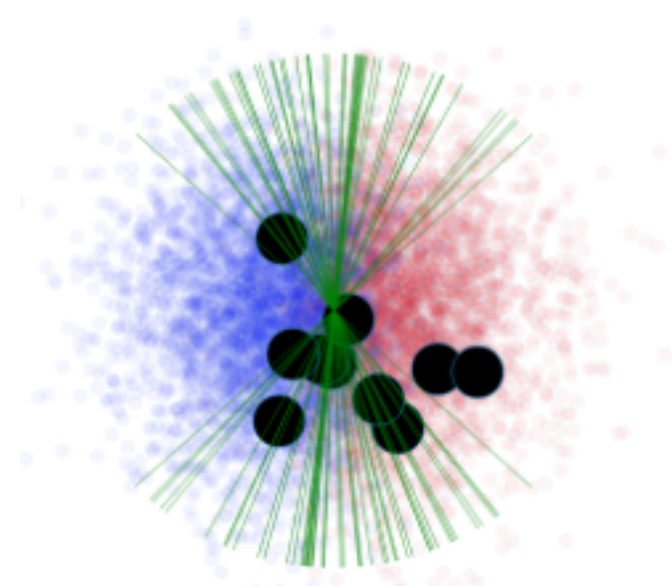


- Might miss important data
- Noisy estimates

Uniform subsampling

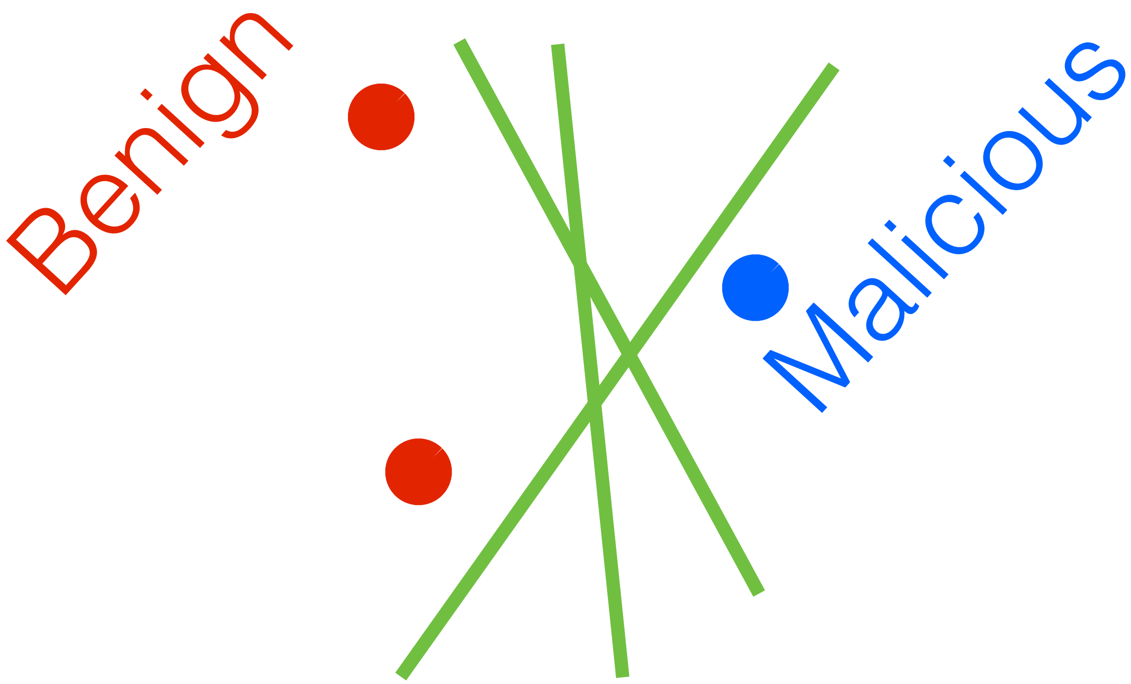


- Might miss important data
- Noisy estimates

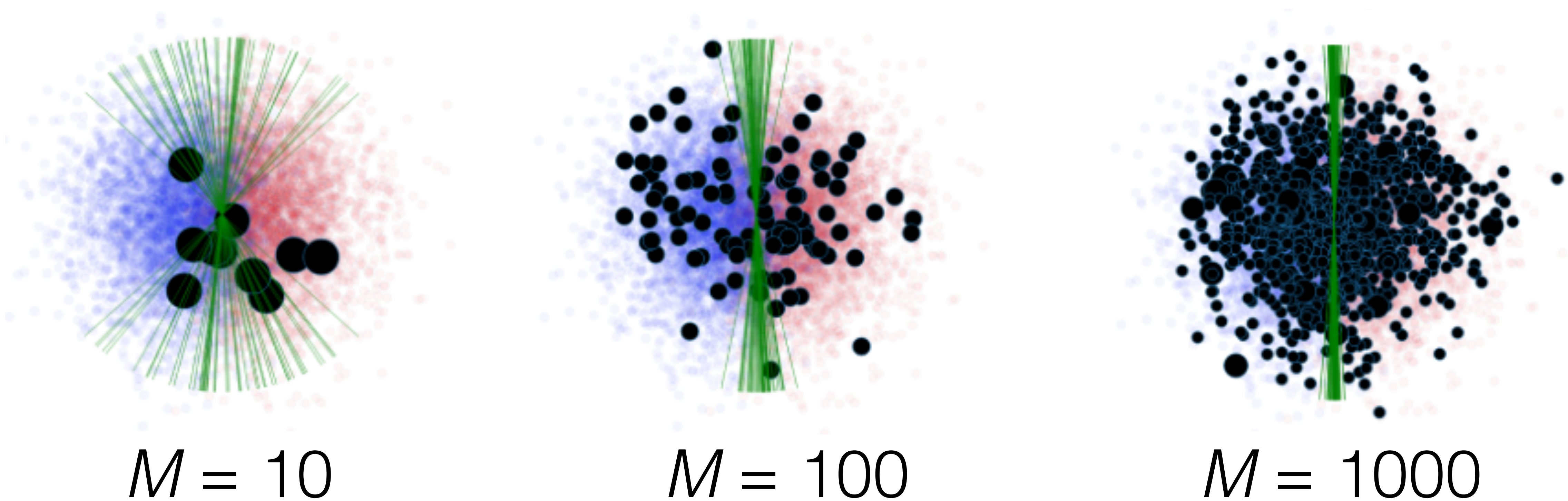


$M = 10$

Uniform subsampling



- Might miss important data
- Noisy estimates



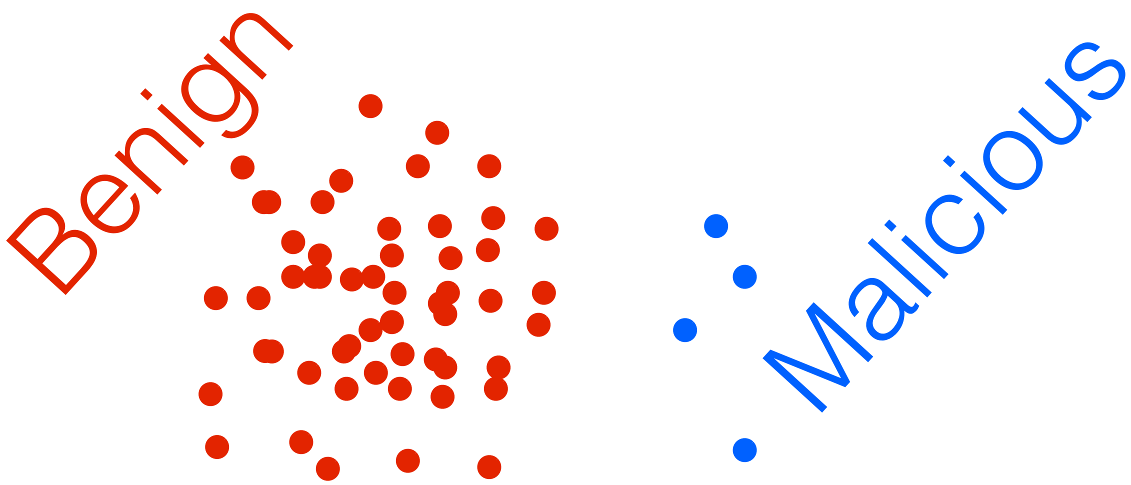
Roadmap

- The “core” of the data set
- Uniform data subsampling isn’t enough
- Importance sampling for “coresets”
- Optimization for “coresets”
- Approximate sufficient statistics

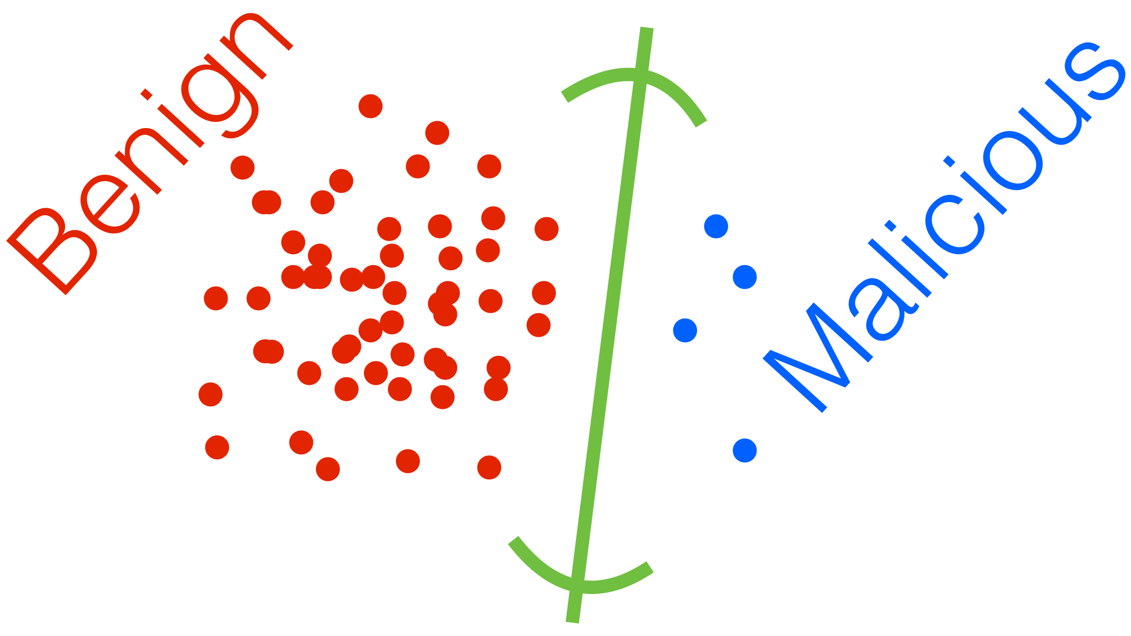
Roadmap

- The "core" of the data set
- Uniform data subsampling isn't enough
- Importance sampling for "coresets"
- Optimization for "coresets"
- Approximate sufficient statistics

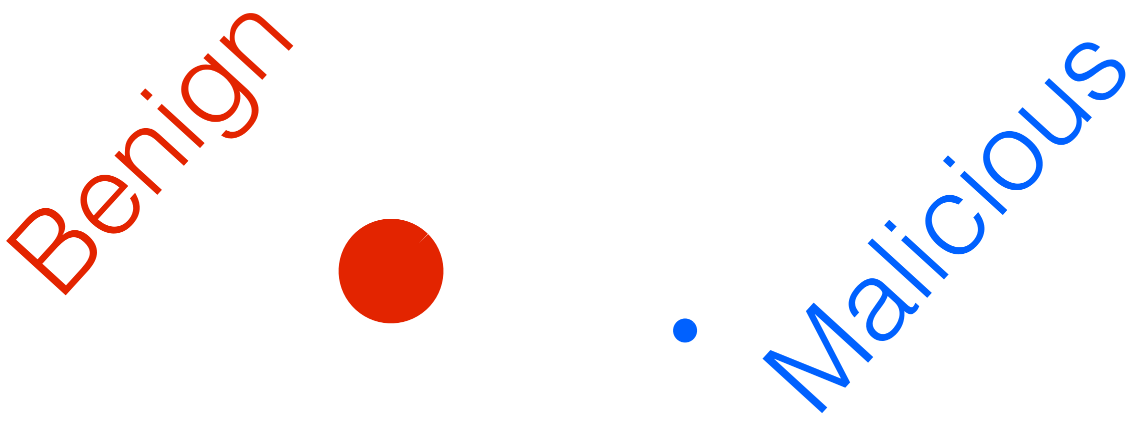
Importance sampling



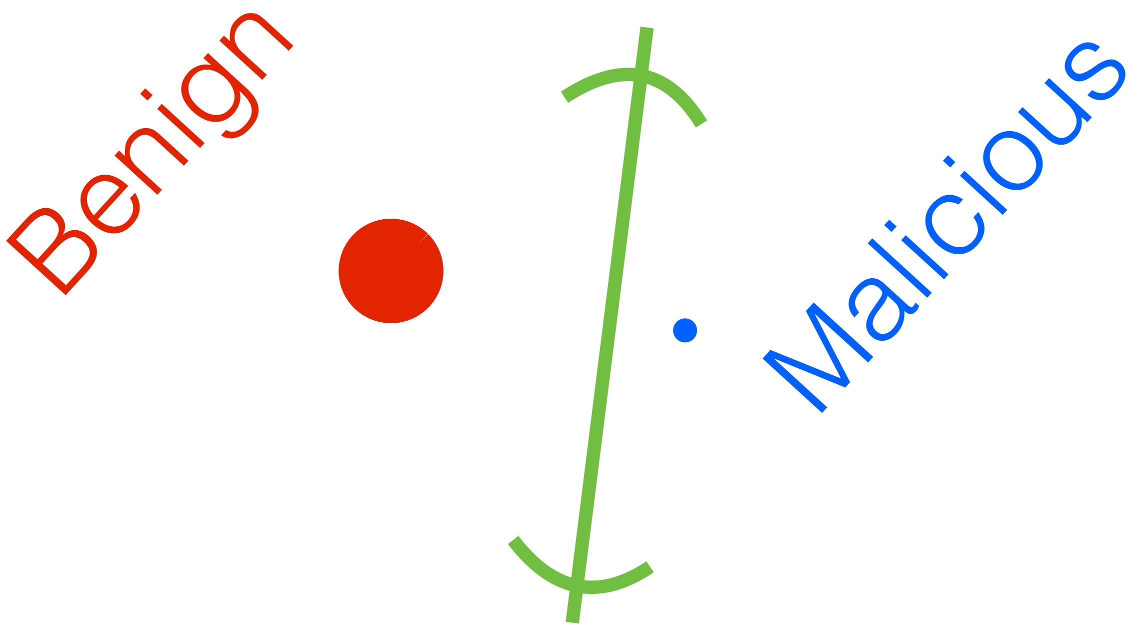
Importance sampling



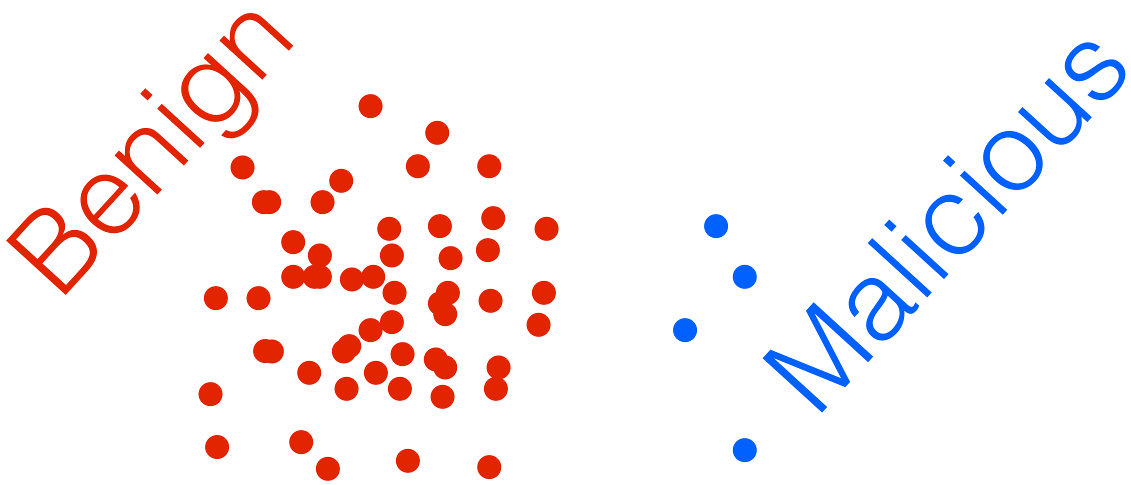
Importance sampling



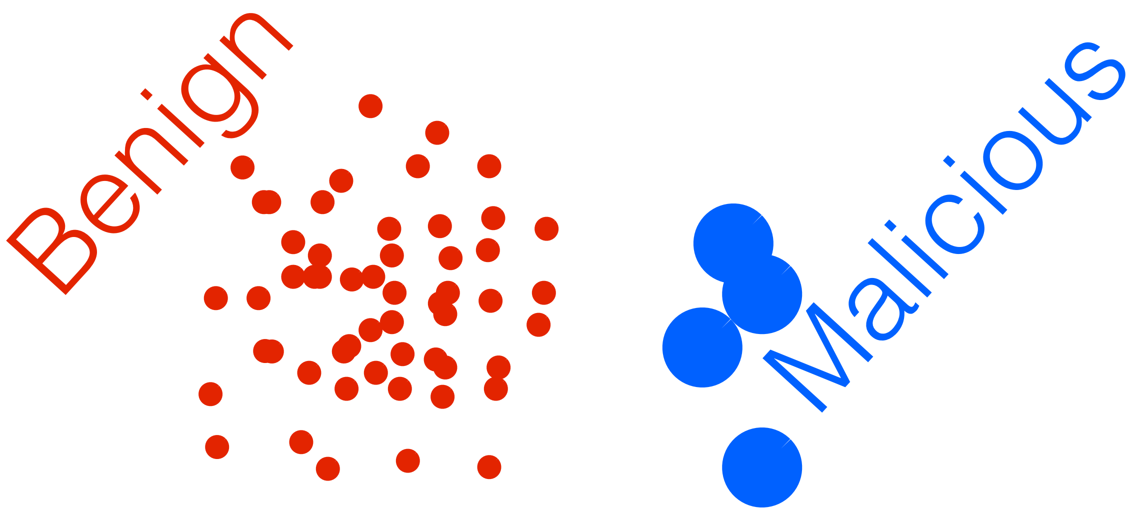
Importance sampling



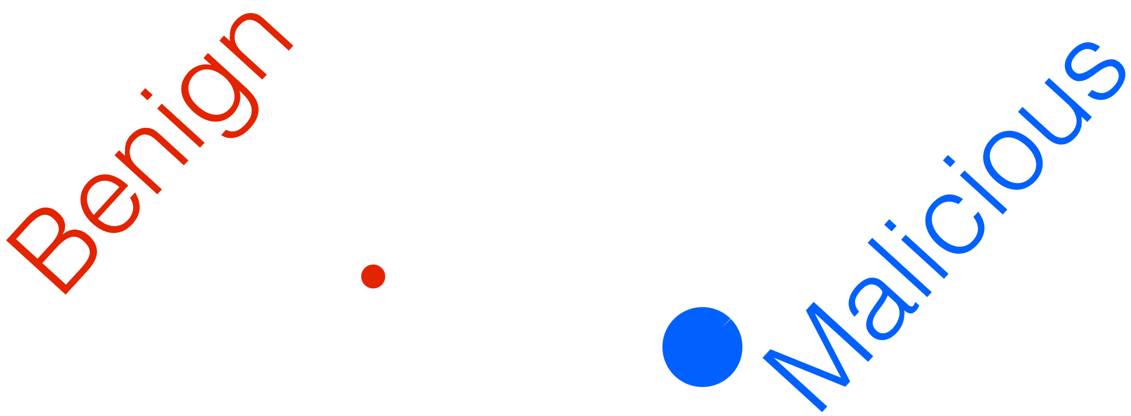
Importance sampling



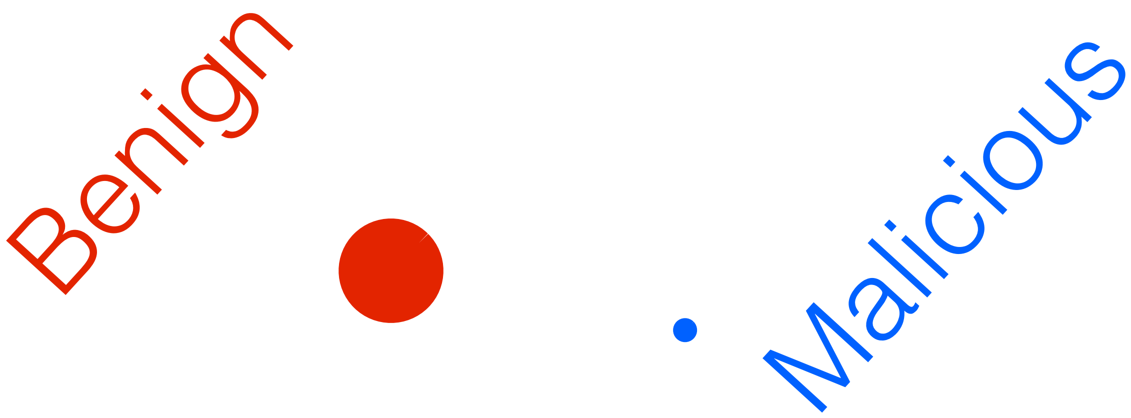
Importance sampling



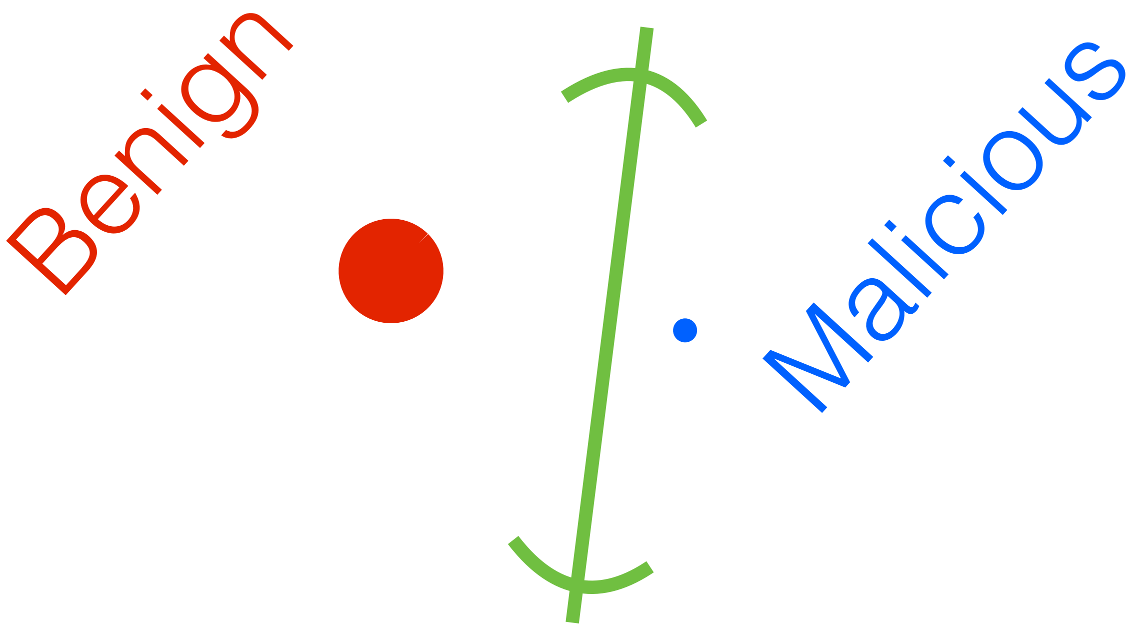
Importance sampling



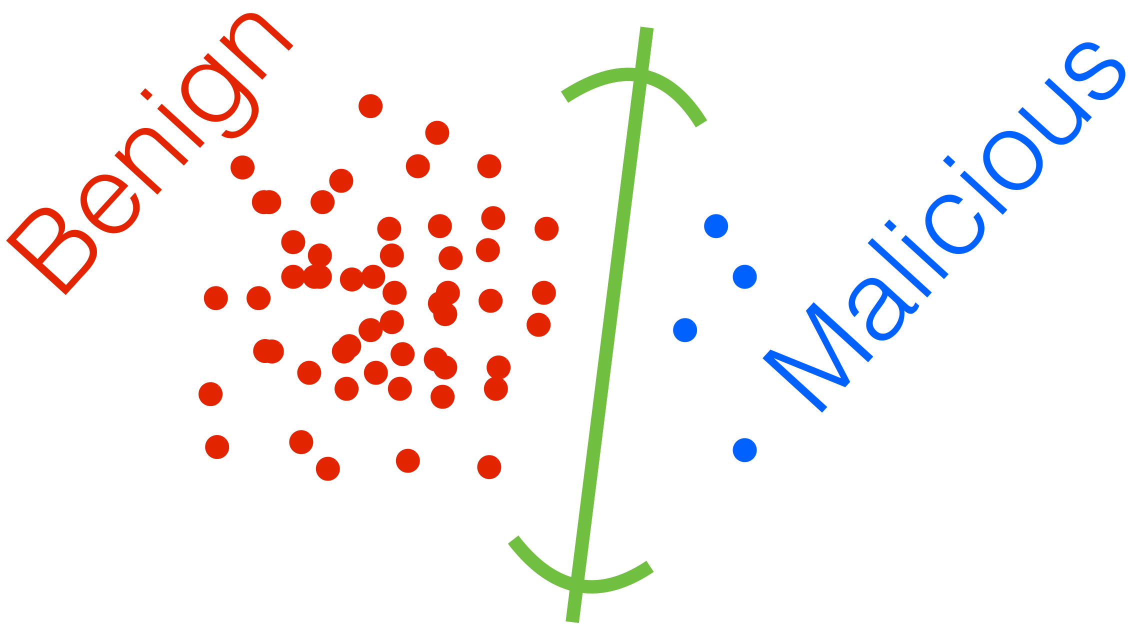
Importance sampling



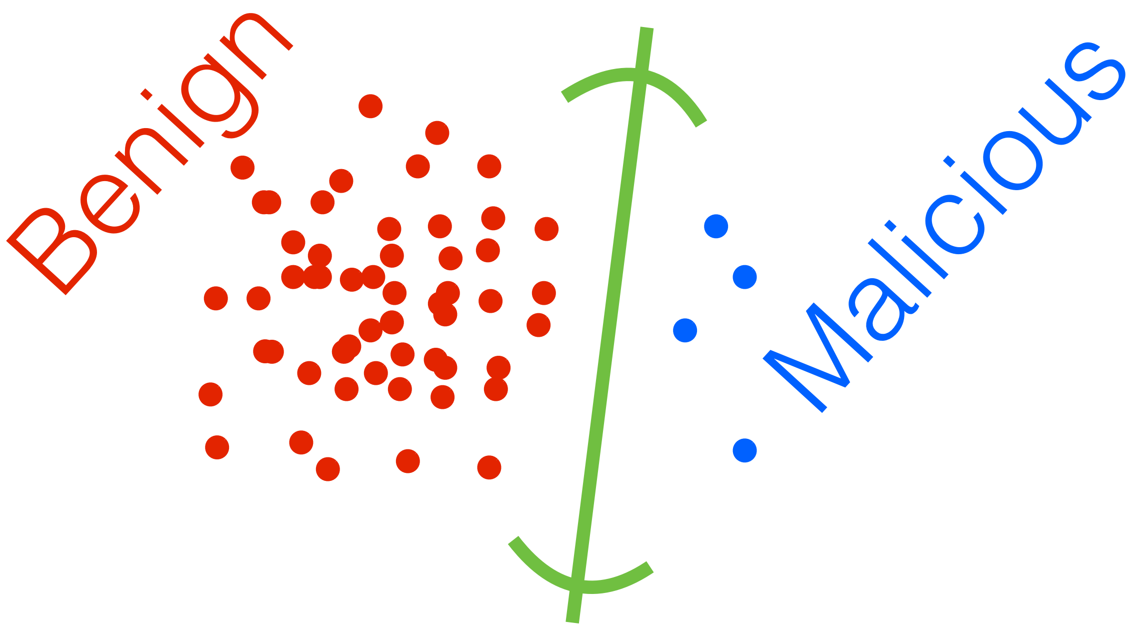
Importance sampling



Importance sampling

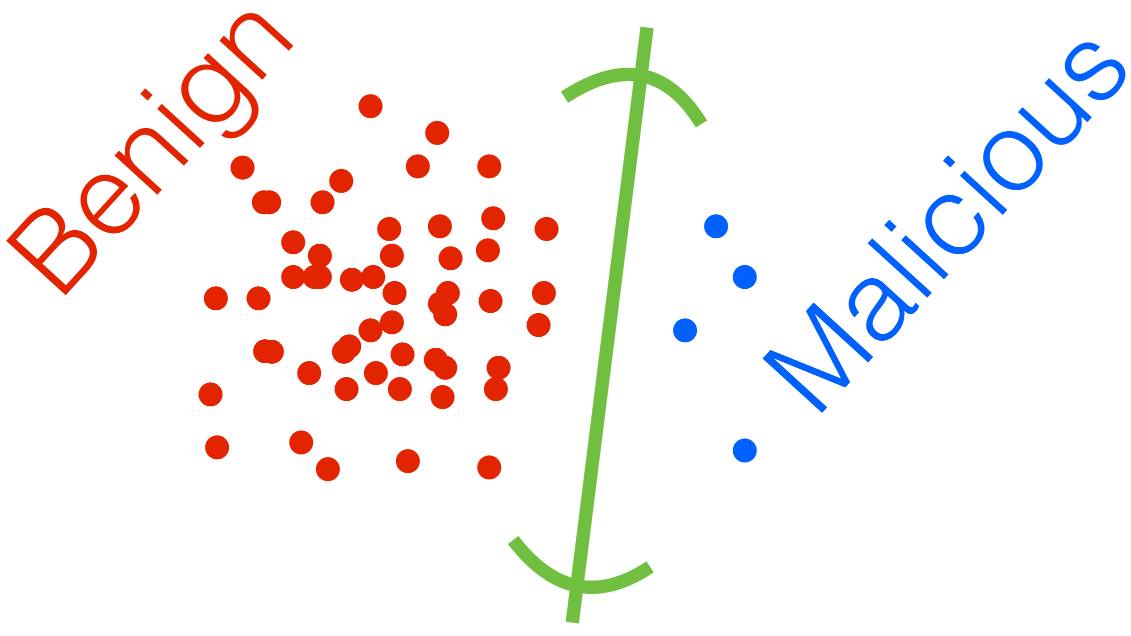


Importance sampling



$$\sigma_n \propto \|\mathcal{L}_n\|$$

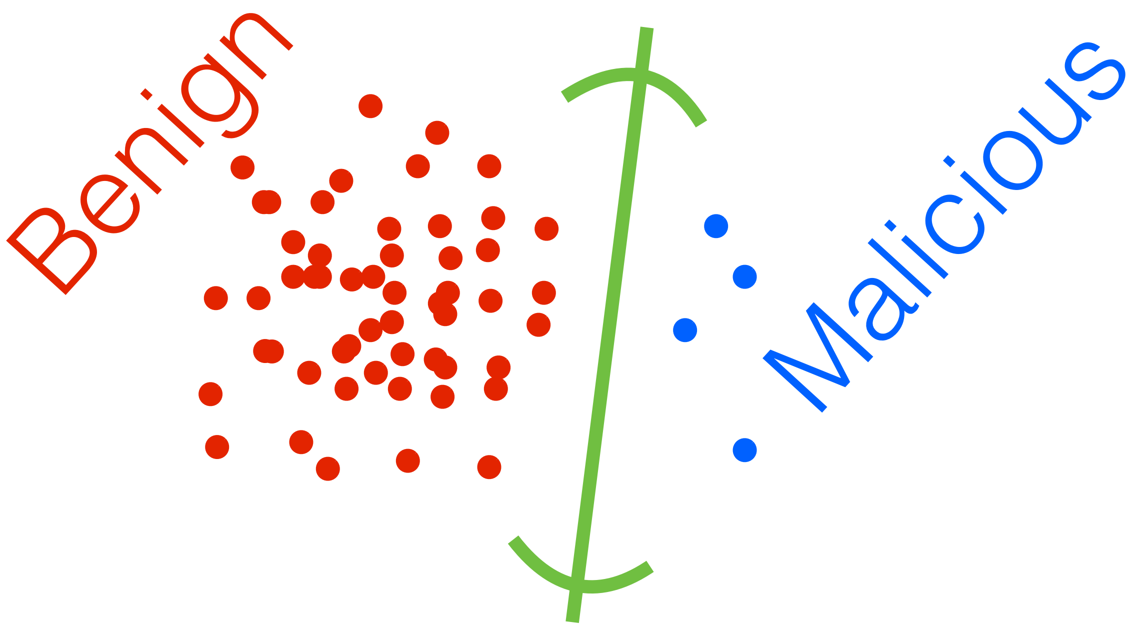
Importance sampling



$$\sigma := \sum_{n=1}^{N} \|\mathcal{L}_n\|$$

$$\sigma_n := \|\mathcal{L}_n\|/\sigma$$

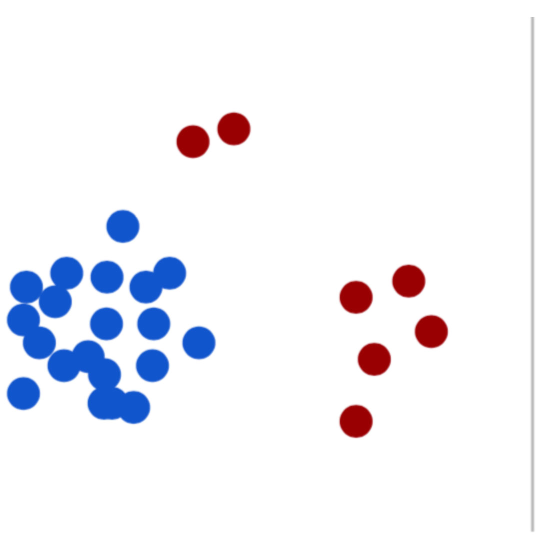
Importance sampling



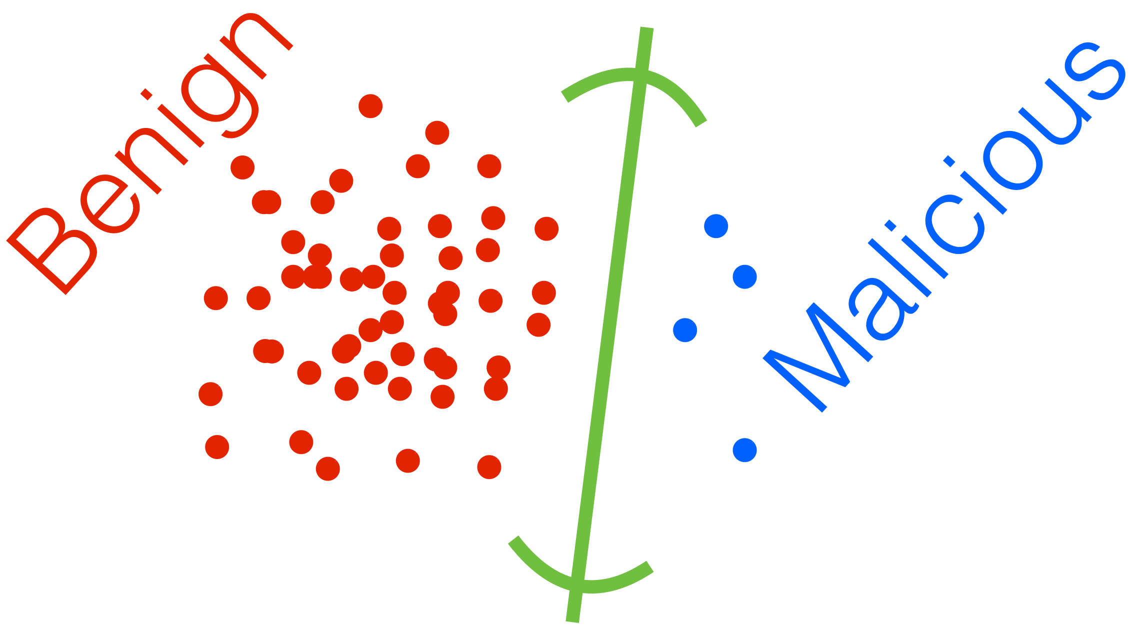
$$\sigma := \sum_{n=1}^{N} \|\mathcal{L}_n\|$$

$$\sigma_n := \|\mathcal{L}_n\| / \sigma$$

1. data

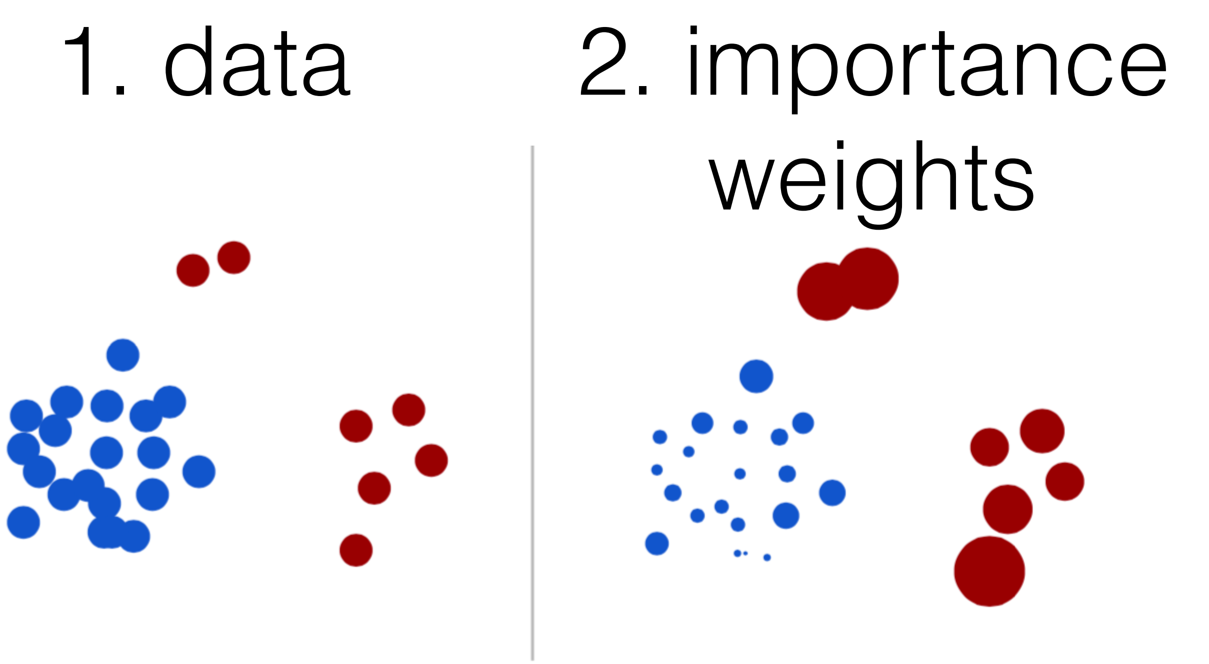


Importance sampling

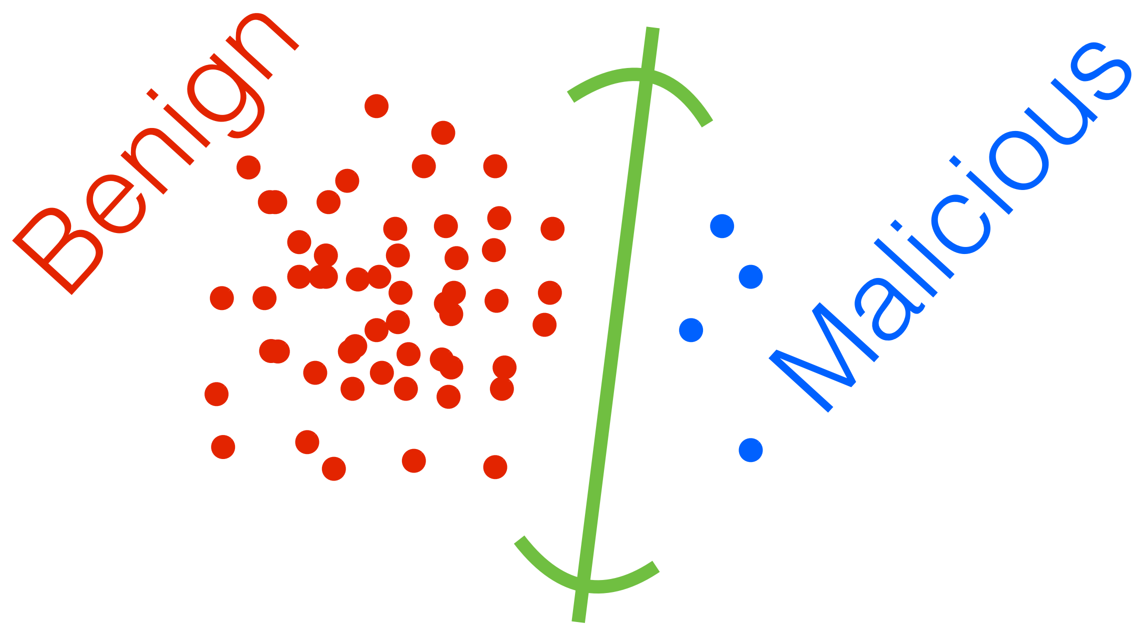


$$\sigma := \sum_{n=1}^{N} \|\mathcal{L}_n\|$$

$$\sigma_n := \|\mathcal{L}_n\|/\sigma$$

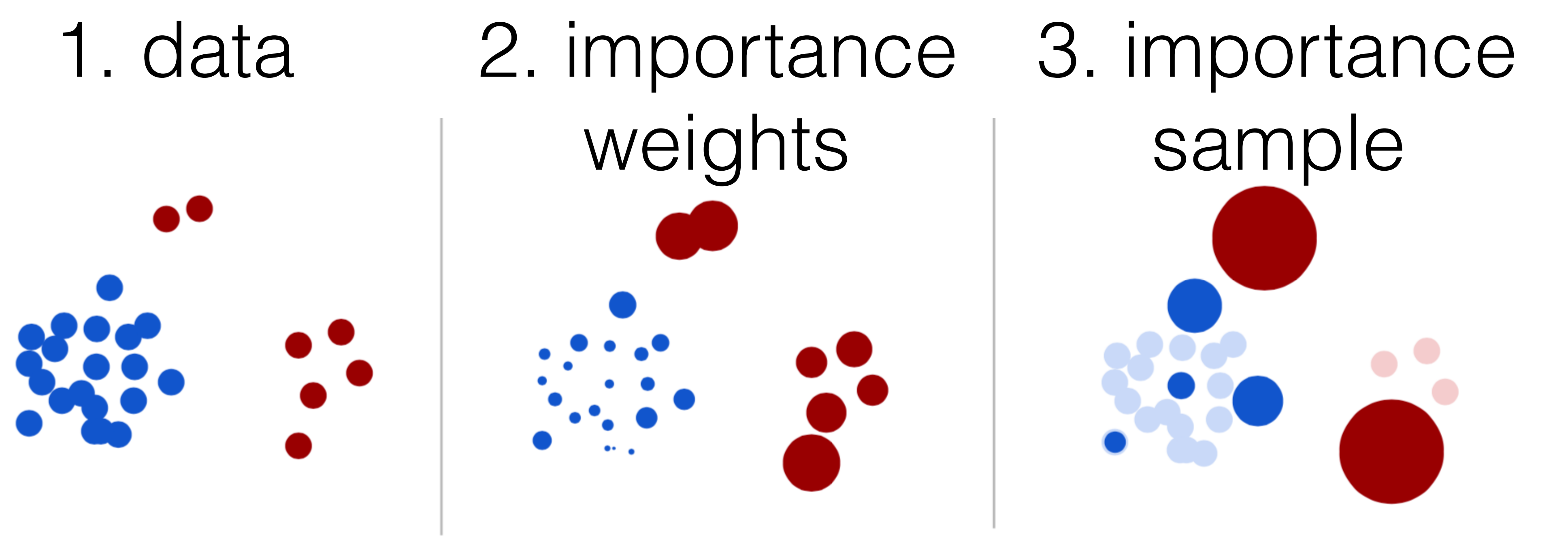


Importance sampling

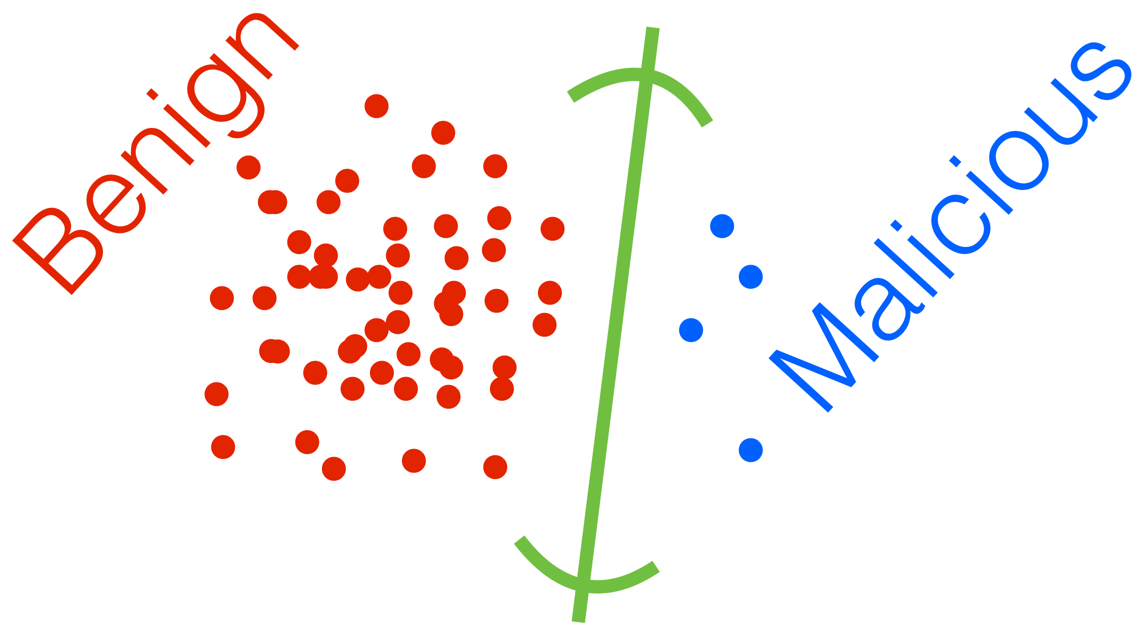


$$\sigma := \sum_{n=1}^{N} \|\mathcal{L}_n\|$$

$$\sigma_n := \|\mathcal{L}_n\| / \sigma$$

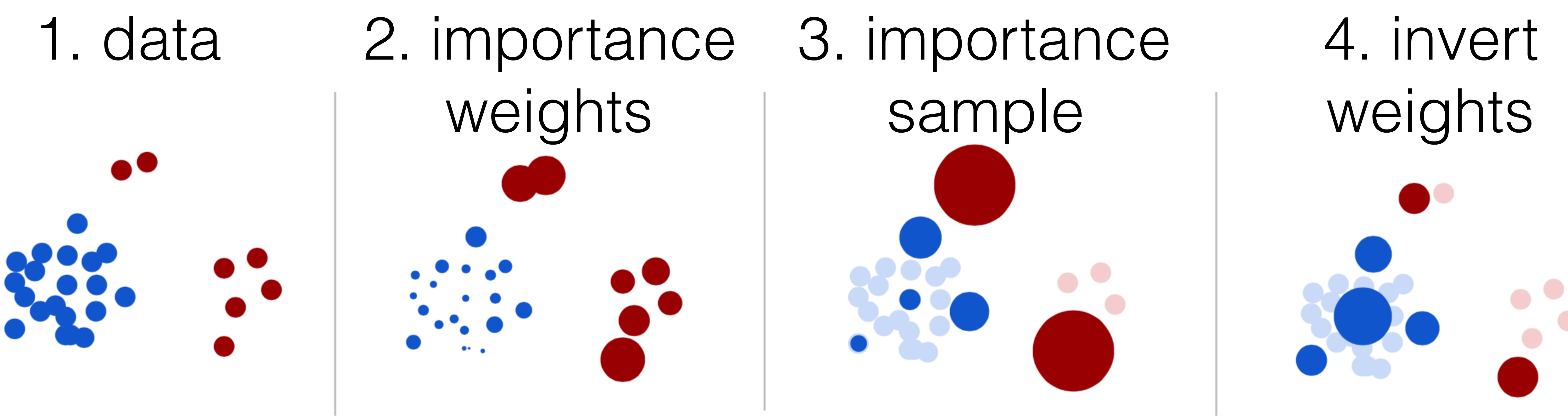


Importance sampling



$$\sigma := \sum_{n=1}^{N} \|\mathcal{L}_n\|$$

$$\sigma_n := \|\mathcal{L}_n\| / \sigma$$



Importance sampling

Thm (CB). δ ∈ (0,1). With probability ≥1 - δ, after *M* iterations,

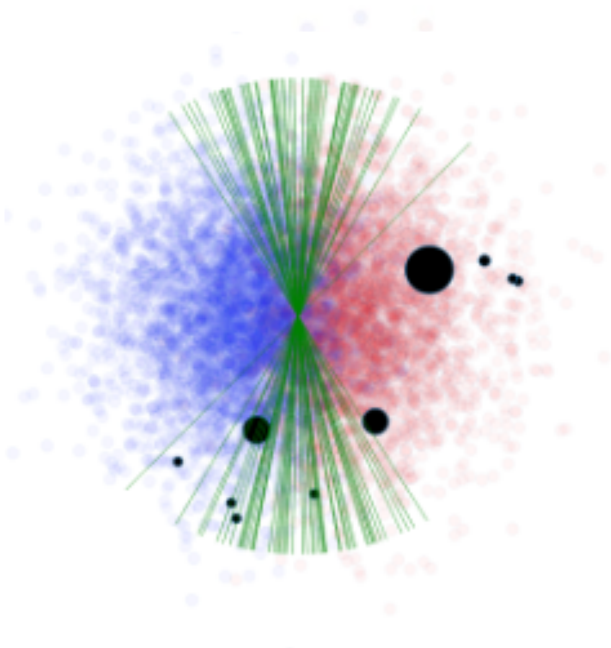
$$\|\mathcal{L}(w) - \mathcal{L}\| \leq \frac{\sigma\bar{\eta}}{\sqrt{M}}\left(1 + \sqrt{2\log\frac{1}{\delta}}\right)$$

Importance sampling

Thm (CB). δ ∈ (0,1). With probability ≥1 - δ, after *M* iterations,

$$\|\mathcal{L}(w) - \mathcal{L}\| \leq \frac{\sigma\bar{\eta}}{\sqrt{M}}\left(1 + \sqrt{2\log\frac{1}{\delta}}\right)$$

- Still noisy estimates



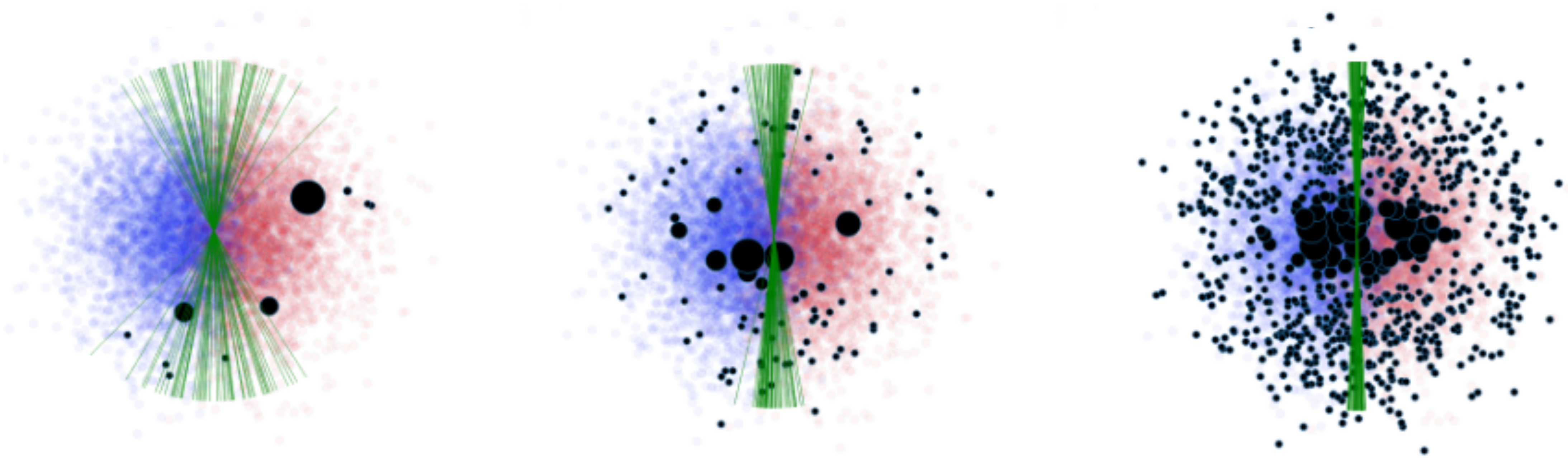
M = 10

Importance sampling

Thm (CB). δ ∈ (0,1). With probability ≥1 - δ, after *M* iterations,

$$\|\mathcal{L}(w) - \mathcal{L}\| \leq \frac{\sigma\bar{\eta}}{\sqrt{M}}\left(1 + \sqrt{2\log\frac{1}{\delta}}\right)$$

- Still noisy estimates



M = 10 *M* = 100 *M* = 1000

Hilbert coresets

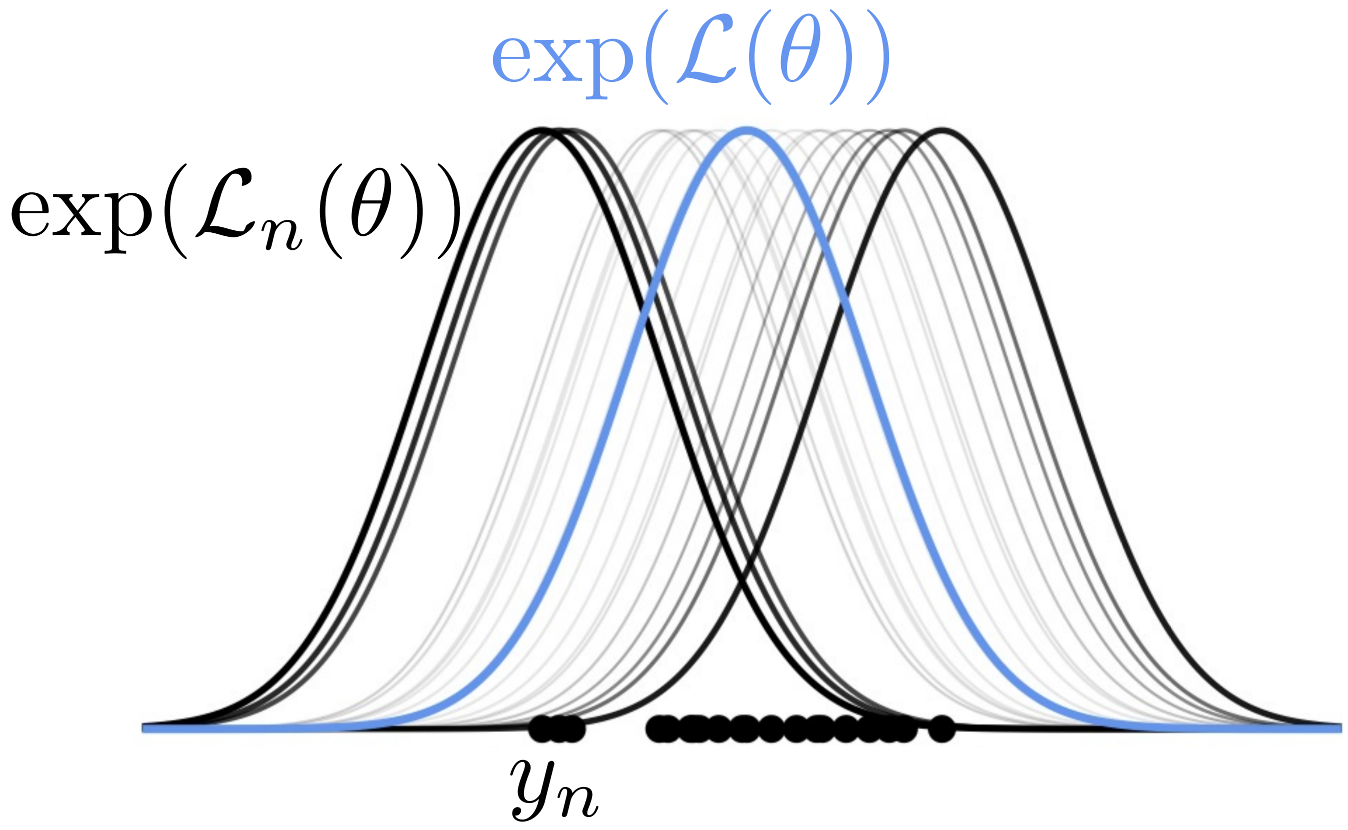
- Want a good coreset:

$$\min_{w \in \mathbb{R}^N} \|\mathcal{L}(w) - \mathcal{L}\|^2$$
$$\text{s.t. } w \geq 0, \|w\|_0 \leq M$$

Hilbert coresets

- Want a good coreset:

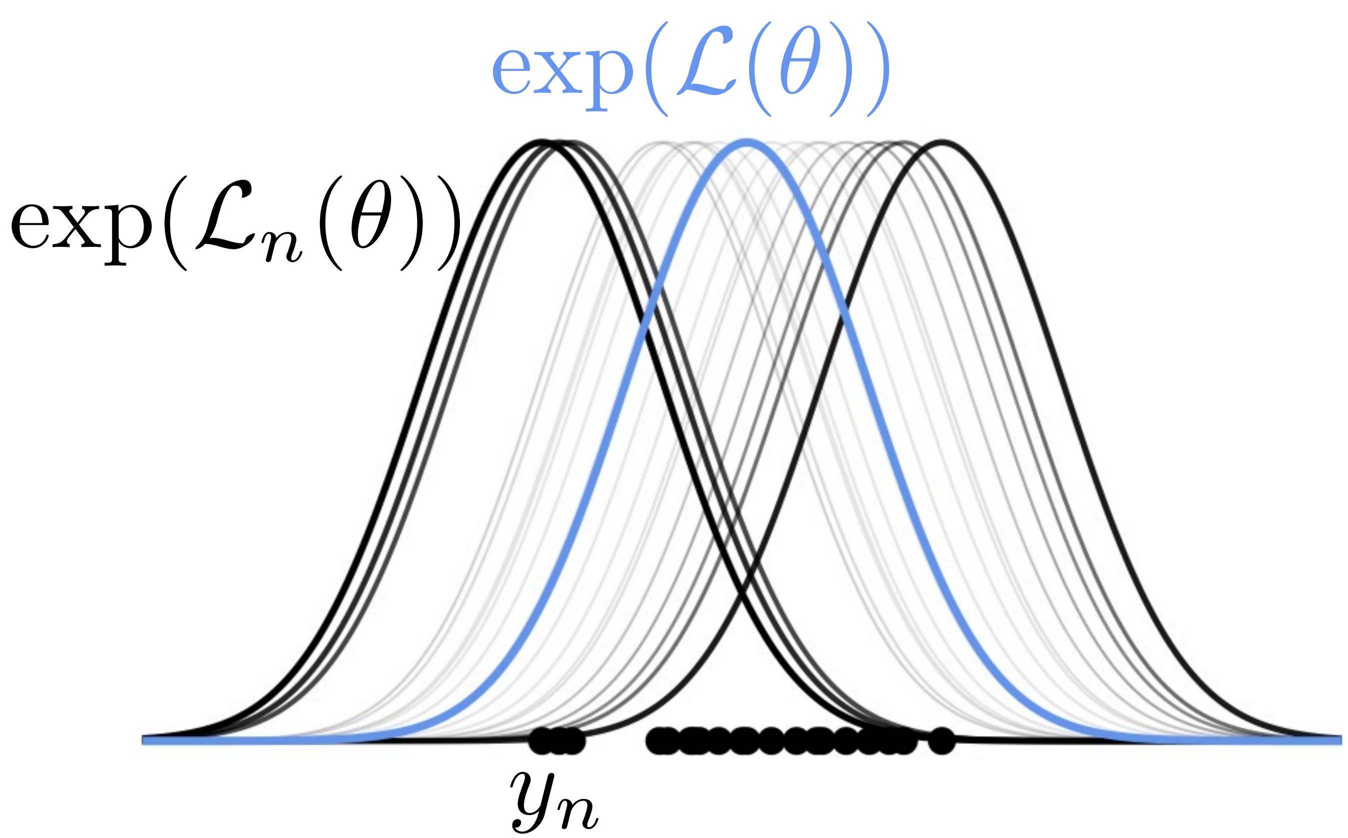
$$\min_{w \in \mathbb{R}^N} \|\mathcal{L}(w) - \mathcal{L}\|^2$$
$$\text{s.t. } w \geq 0, \|w\|_0 \leq M$$

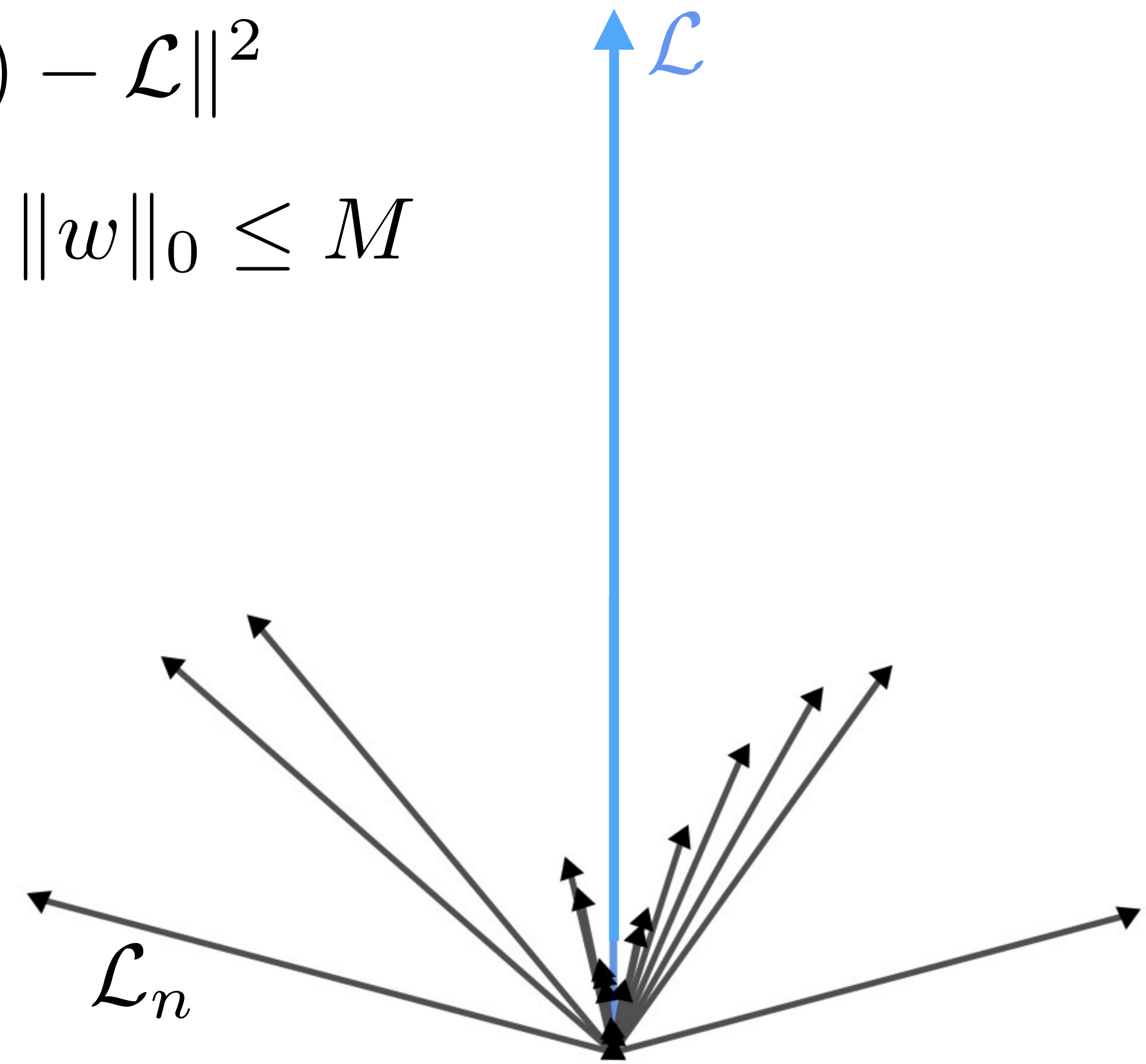


Hilbert coresets

- Want a good coreset:

$$\min_{w \in \mathbb{R}^N} \|\mathcal{L}(w) - \mathcal{L}\|^2$$
$$\text{s.t. } w \geq 0, \|w\|_0 \leq M$$

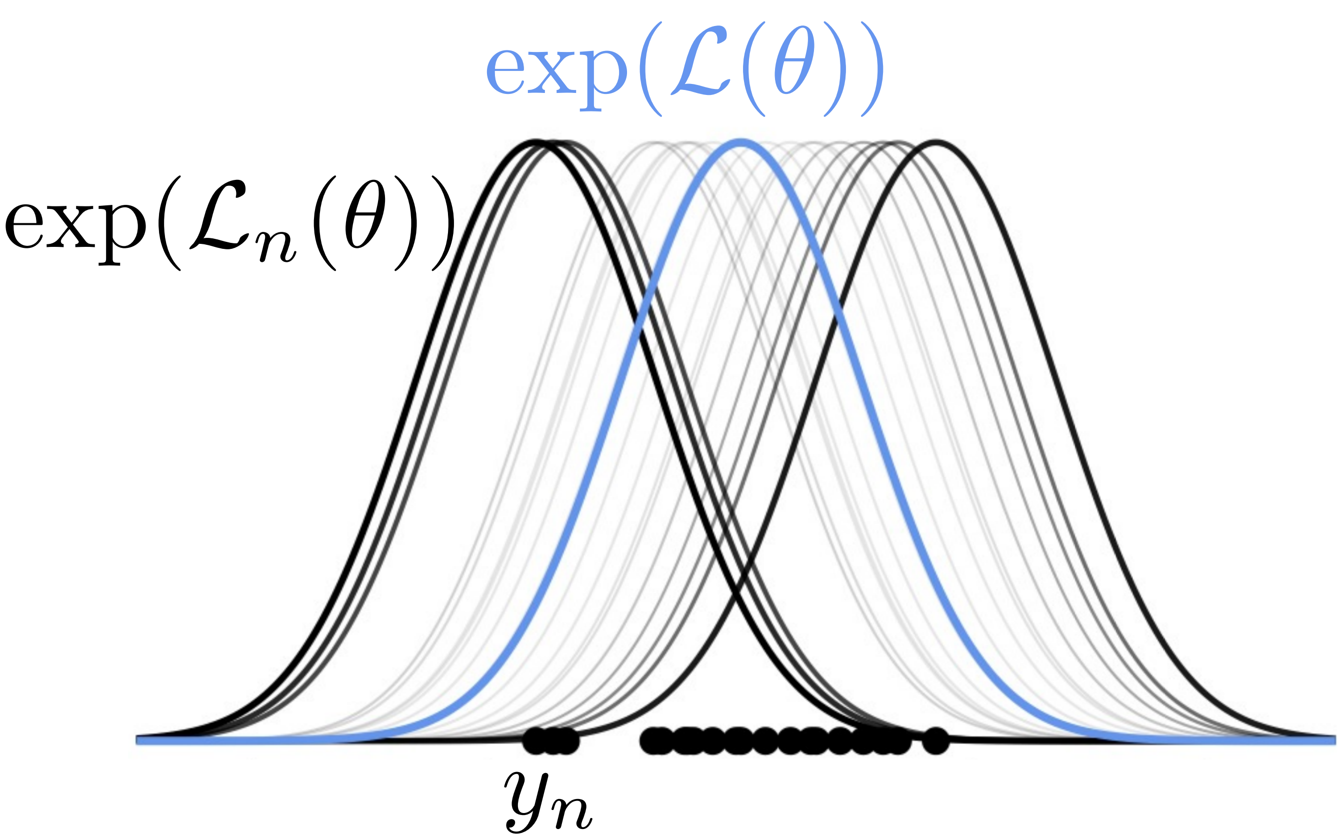


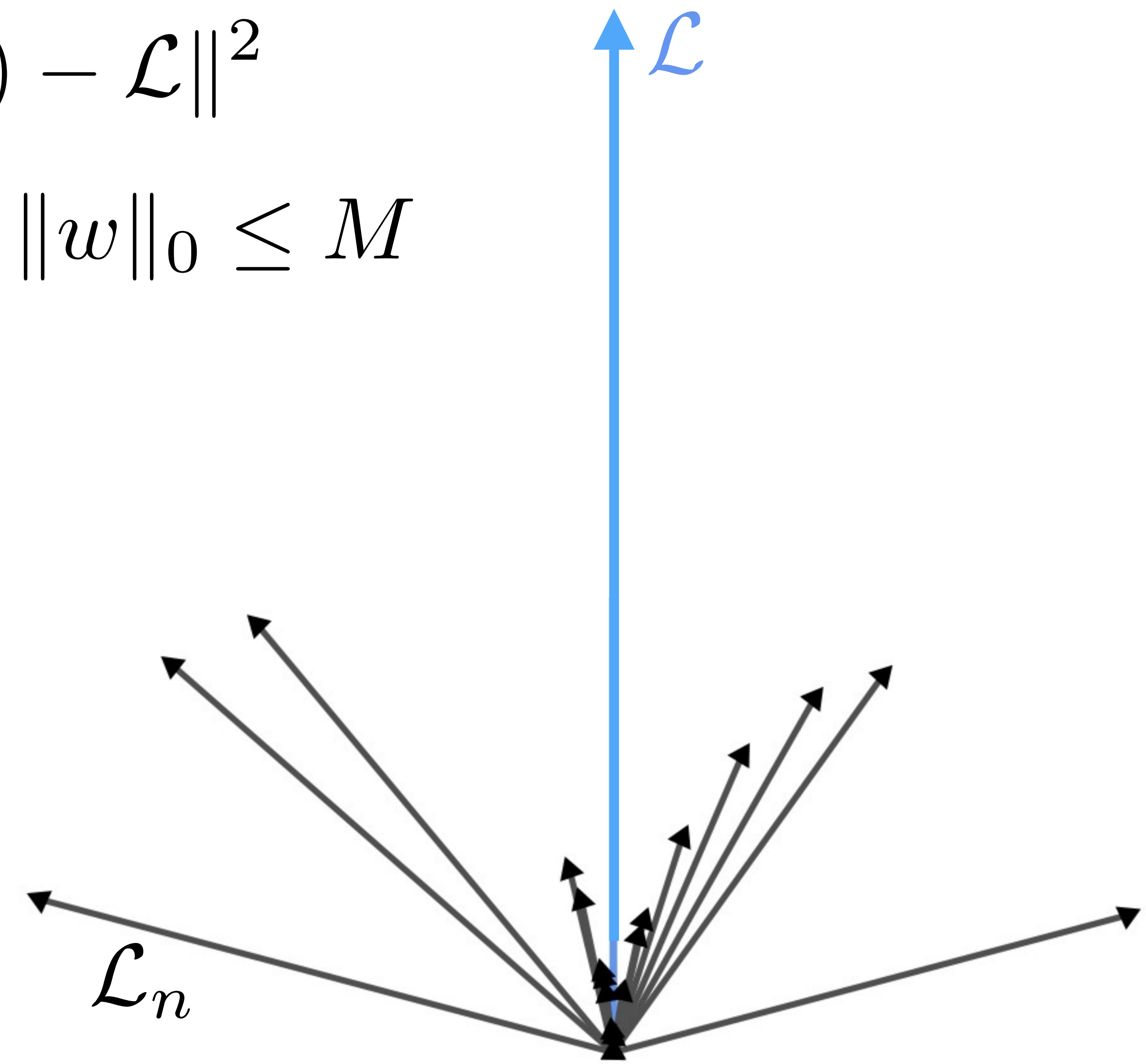


Hilbert coresets

- Want a good coreset:

$$\min_{w \in \mathbb{R}^N} \|\mathcal{L}(w) - \mathcal{L}\|^2$$
$$\text{s.t. } w \geq 0, \|w\|_0 \leq M$$



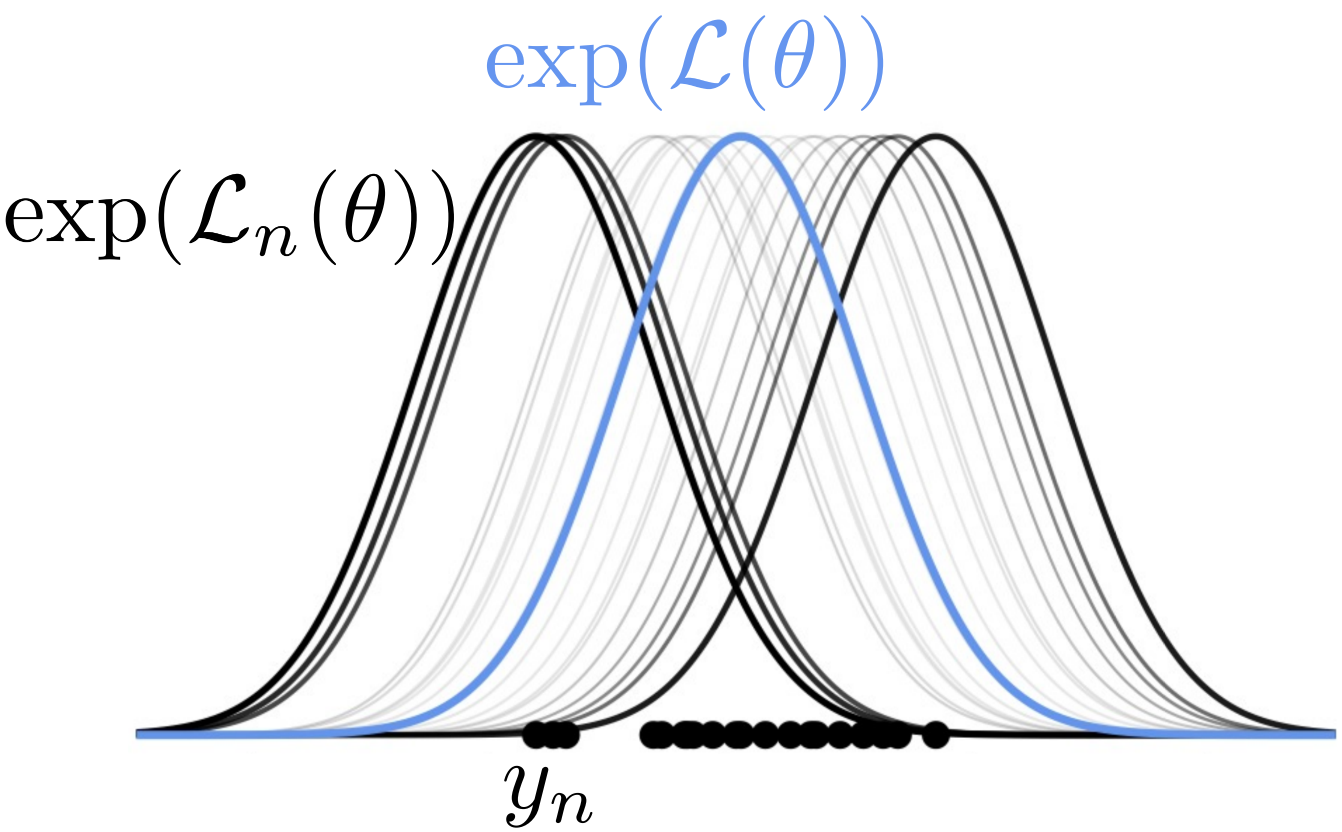


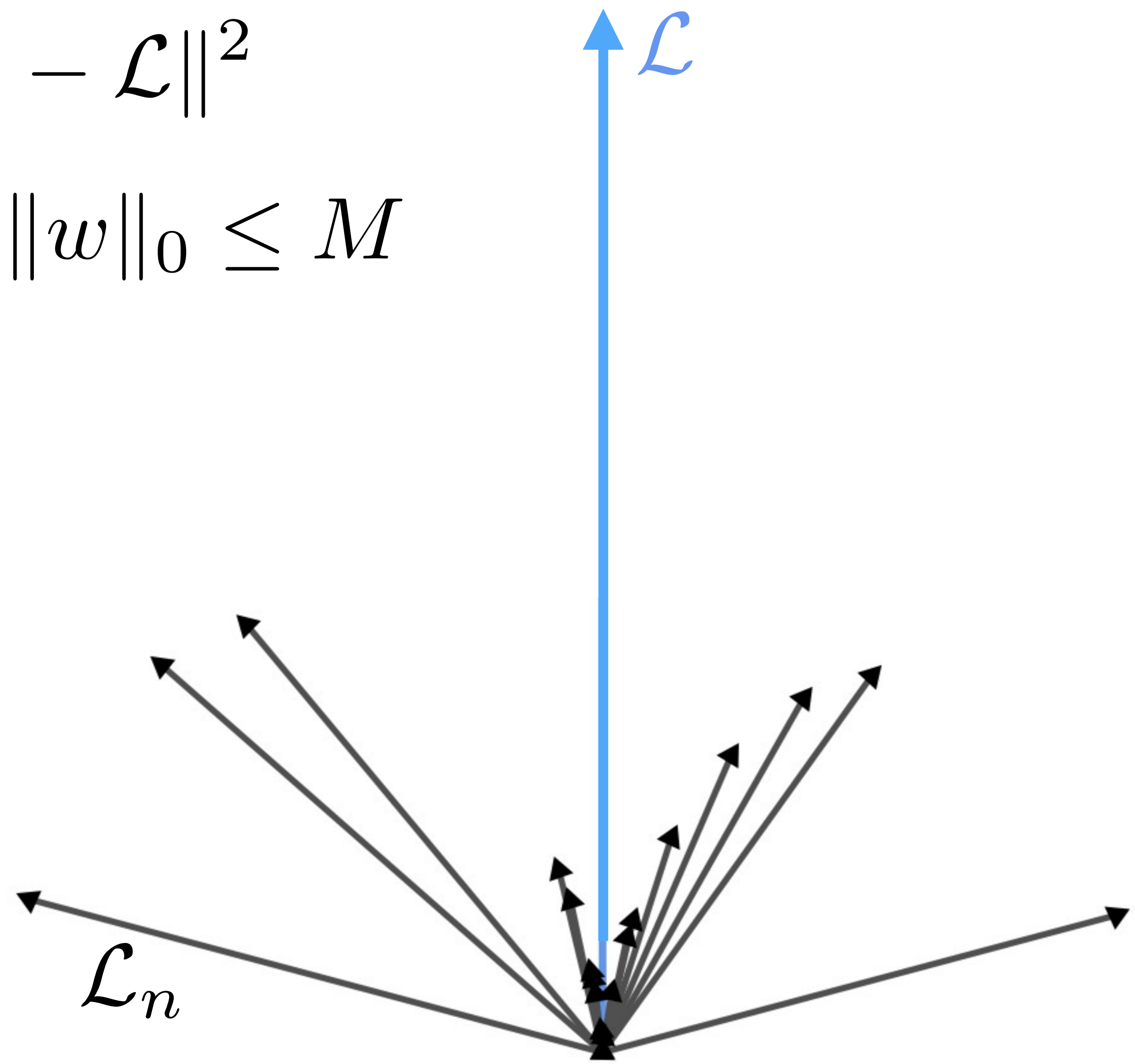
- need to consider (residual) error direction

Hilbert coresets

- Want a good coreset:

$$\min_{w \in \mathbb{R}^N} \|\mathcal{L}(w) - \mathcal{L}\|^2$$
$$\text{s.t. } w \geq 0, \|w\|_0 \leq M$$





- need to consider (residual) error direction
- sparse optimization

Roadmap

- The “core” of the data set
- Uniform data subsampling isn’t enough
- Importance sampling for “coresets”
- Optimization for “coresets”
- Approximate sufficient statistics

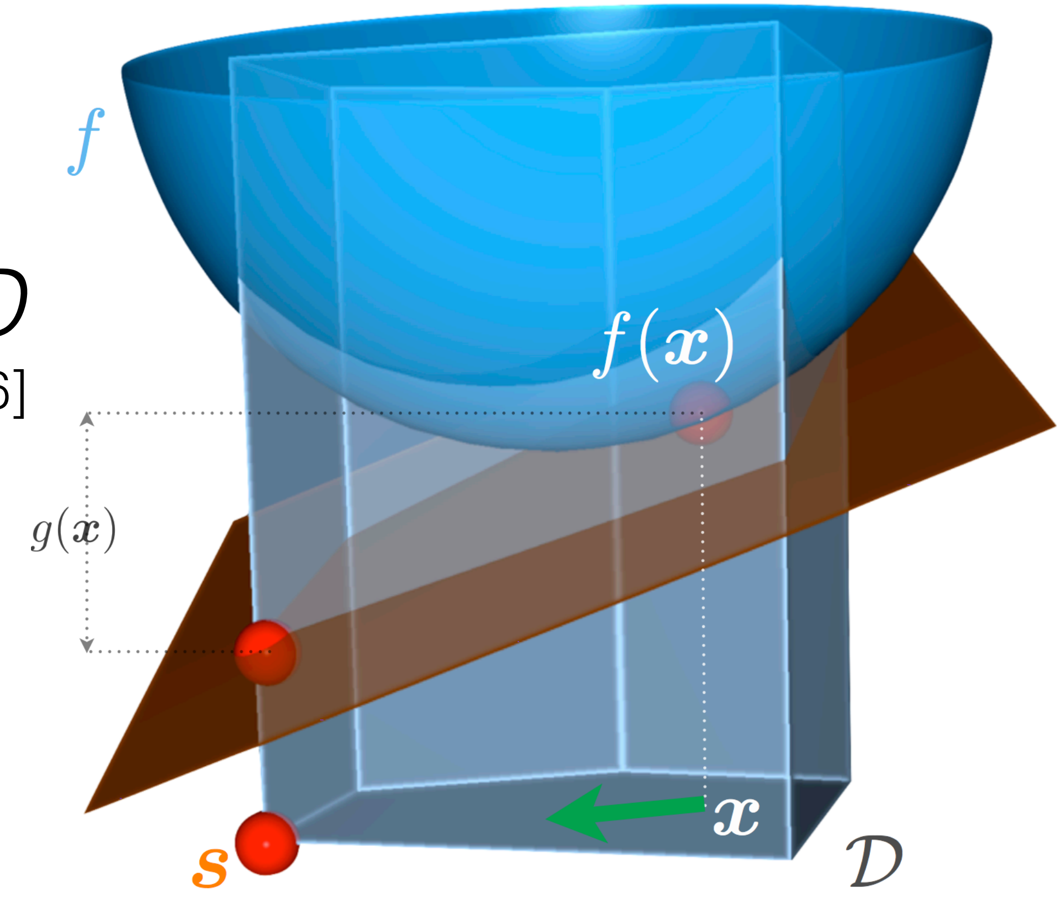
Roadmap

- The “core” of the data set
- Uniform data subsampling isn’t enough
- Importance sampling for “coresets”
- Optimization for “coresets”
- Approximate sufficient statistics

Frank-Wolfe

Convex optimization on a polytope *D*

[Frank, Wolfe 1956]



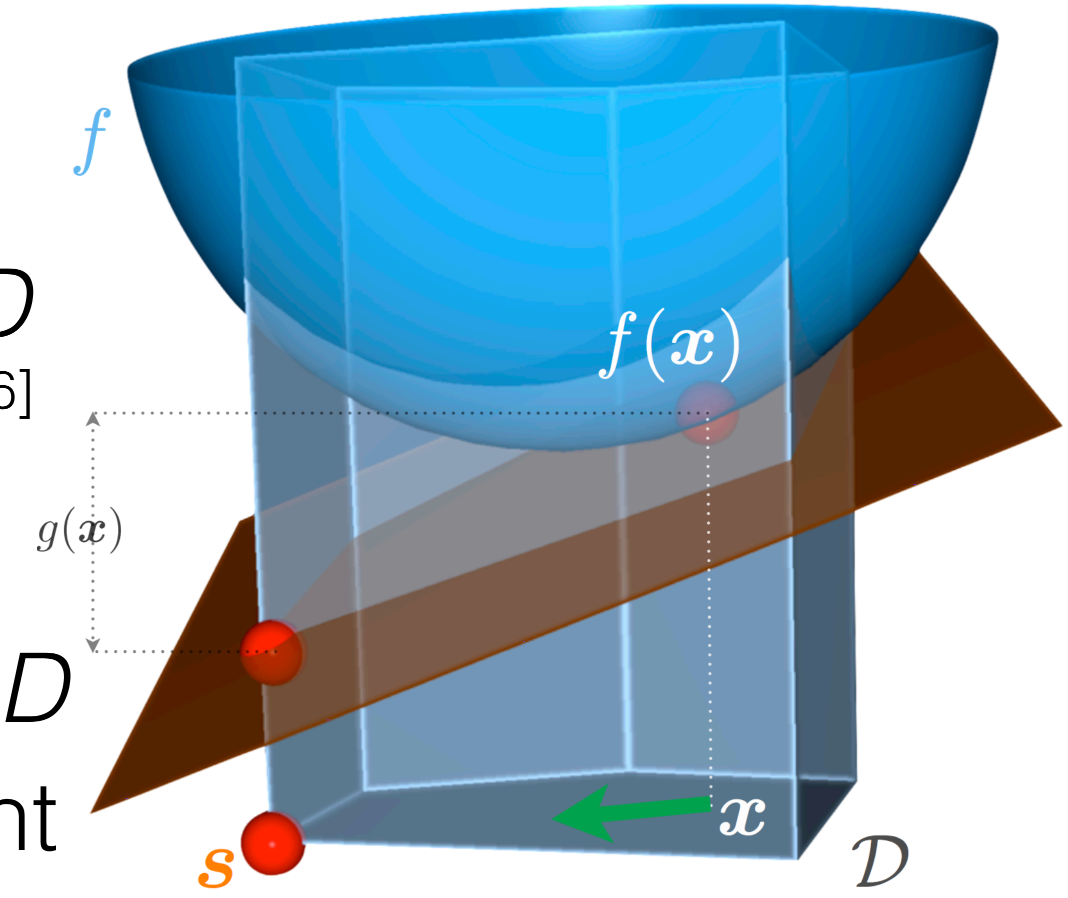
[Jaggi 2013]

Frank-Wolfe

Convex optimization on a polytope D

[Frank, Wolfe 1956]

- Repeat:
    1. Find gradient
    2. Find argmin point on plane in D
    3. Do line search between current point and argmin point



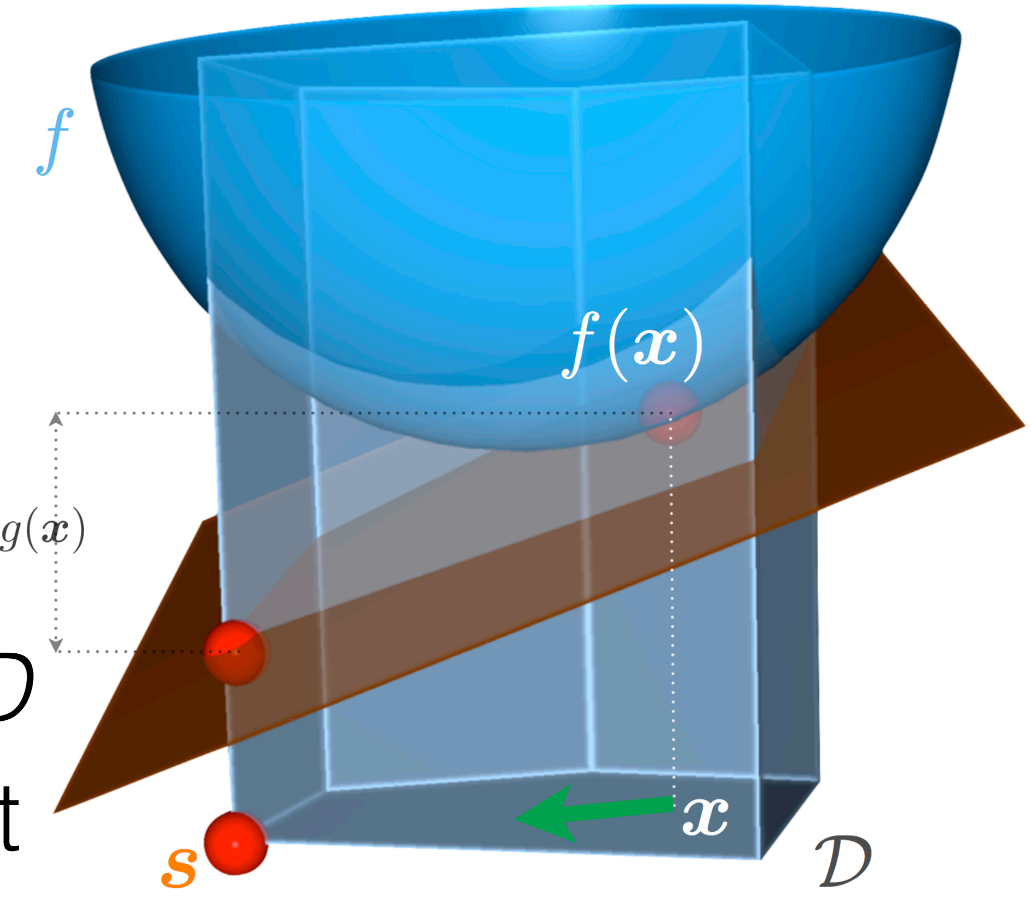
[Jaggi 2013]

Frank-Wolfe

Convex optimization on a polytope *D*

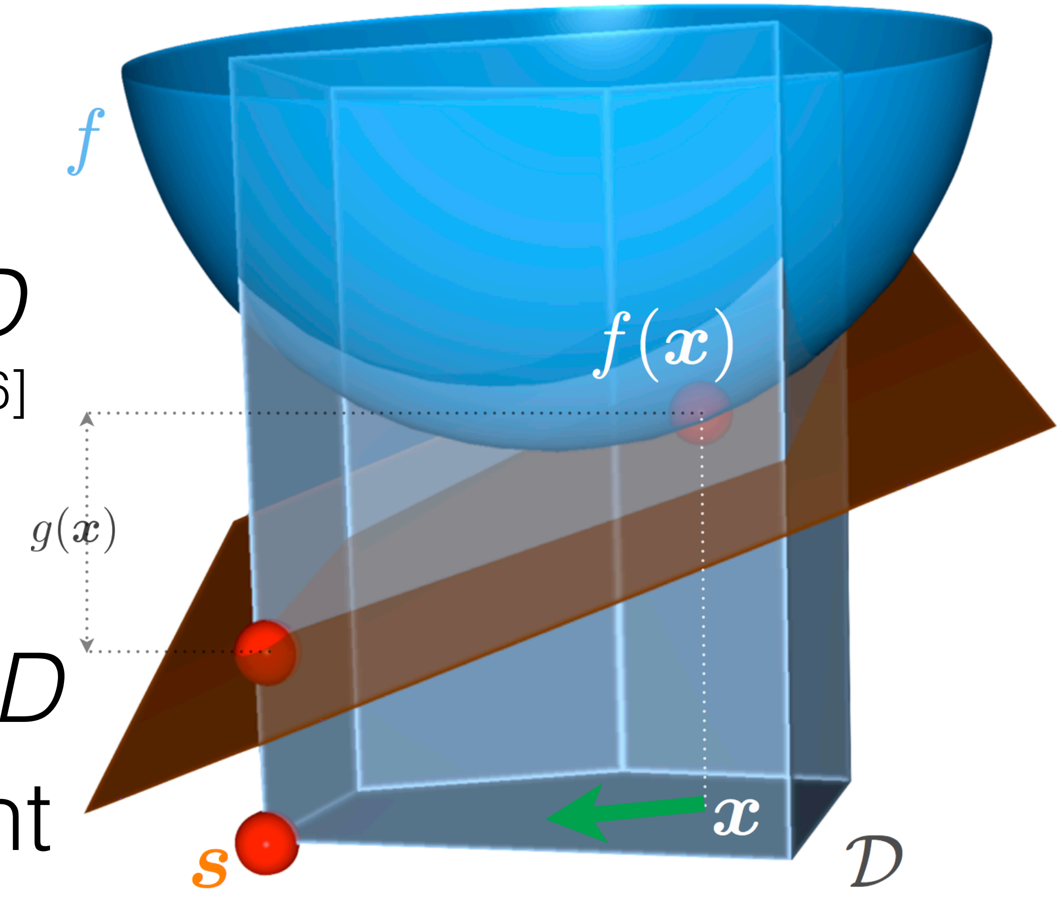
[Frank, Wolfe 1956]

- Repeat:
  1. Find gradient
  2. Find argmin point on plane in *D*
  3. Do line search between current point and argmin point
- Convex combination of *M* vertices after *M* -1 steps



[Jaggi 2013]

Frank-Wolfe



[Jaggi 2013]

Convex optimization on a polytope *D*

[Frank, Wolfe 1956]

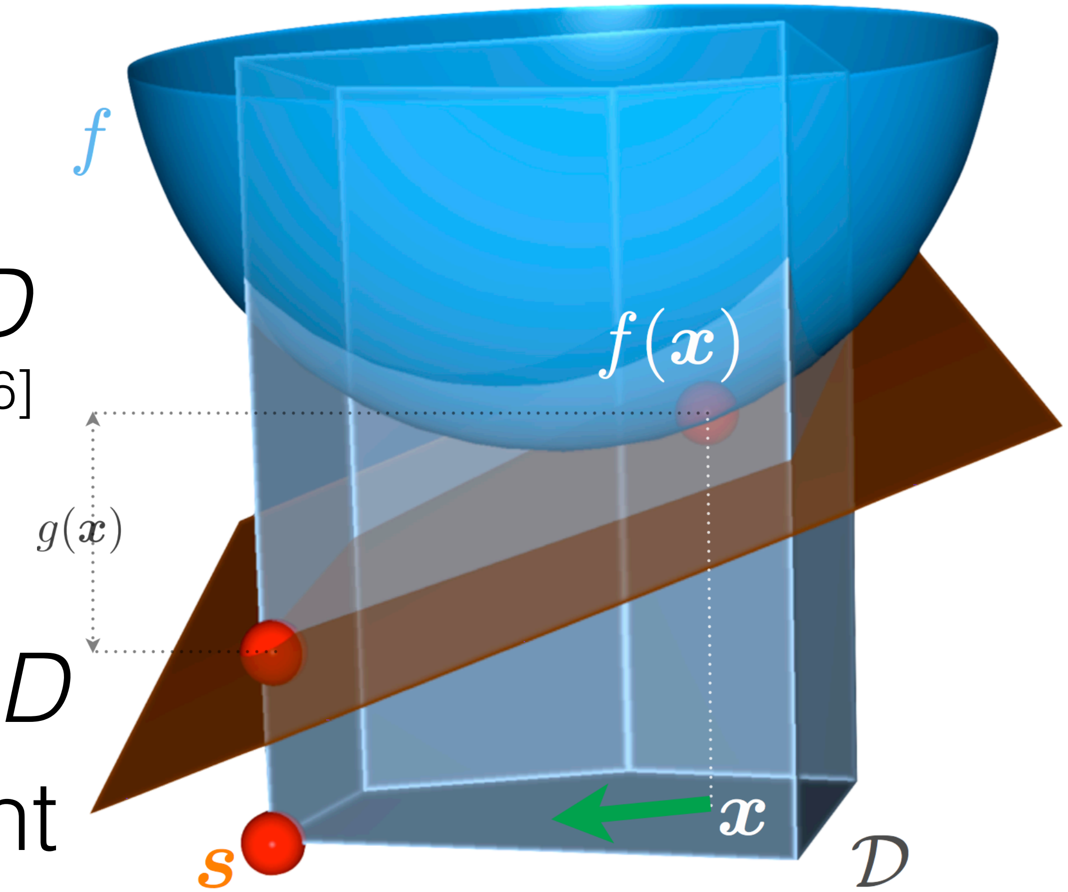
- Repeat:
  1. Find gradient
  2. Find argmin point on plane in *D*
  3. Do line search between current point and argmin point
- Convex combination of *M* vertices after *M* -1 steps
- Our problem: $\min_{w \in \mathbb{R}^N} \|\mathcal{L}(w) - \mathcal{L}\|$

Frank-Wolfe

Convex optimization on a polytope *D*

[Frank, Wolfe 1956]

- Repeat:
  1. Find gradient
  2. Find argmin point on plane in *D*
  3. Do line search between current point and argmin point



[Jaggi 2013]

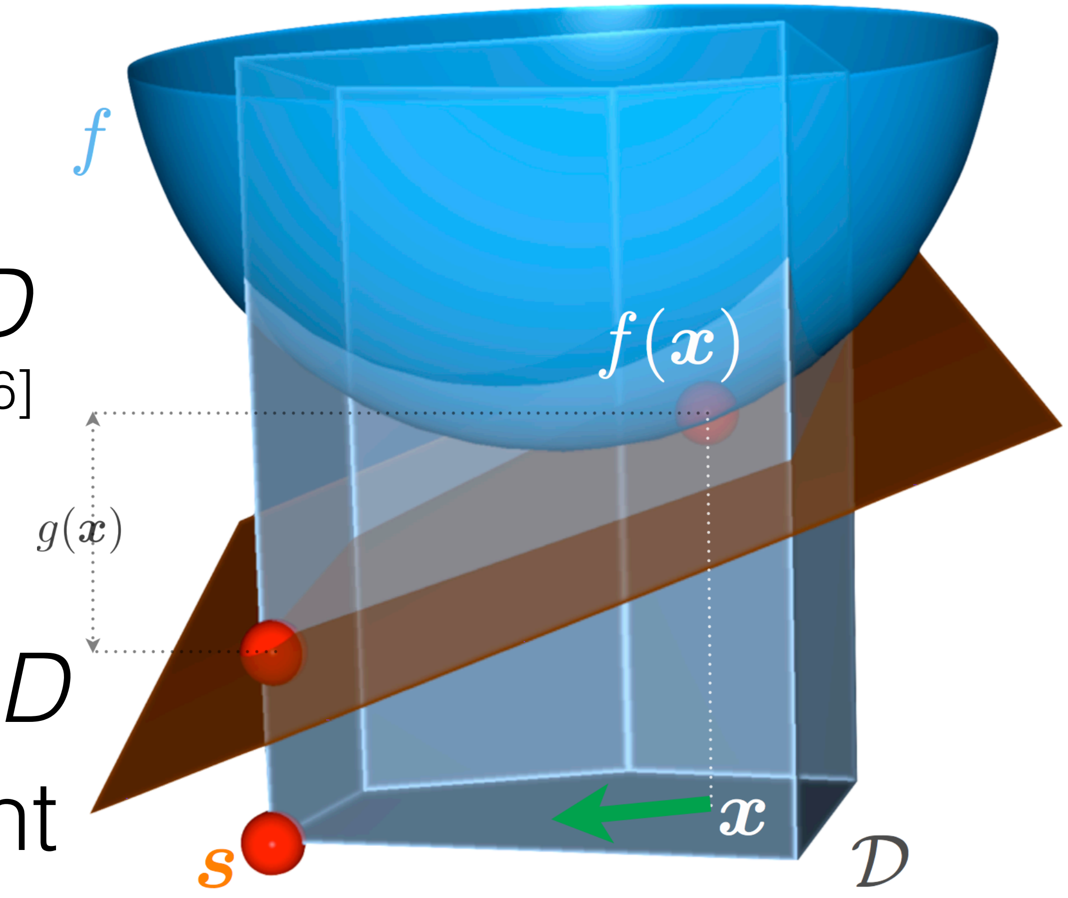
- Convex combination of *M* vertices after *M* -1 steps
- Our problem: $\min_{w \in \mathbb{R}^N} \|\mathcal{L}(w) - \mathcal{L}\|^2$

Frank-Wolfe

Convex optimization on a polytope *D*

[Frank, Wolfe 1956]

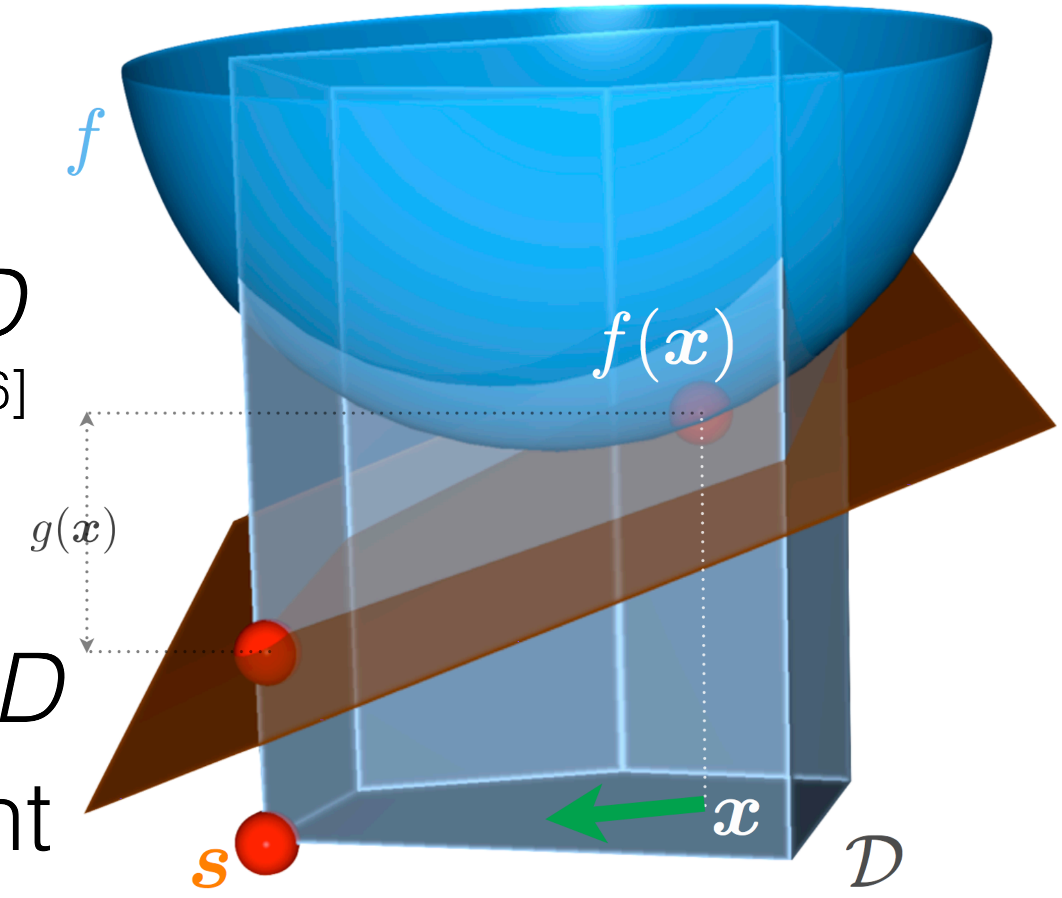
- Repeat:
  1. Find gradient
  2. Find argmin point on plane in *D*
  3. Do line search between current point and argmin point



[Jaggi 2013]

- Convex combination of *M* vertices after *M* -1 steps
- Our problem: $$\min_{w \in \mathbb{R}^N} \|\mathcal{L}(w) - \mathcal{L}\|^2$$ $$\text{s.t. } w \geq 0, \|w\|_0 \leq M$$

Frank-Wolfe



[Jaggi 2013]

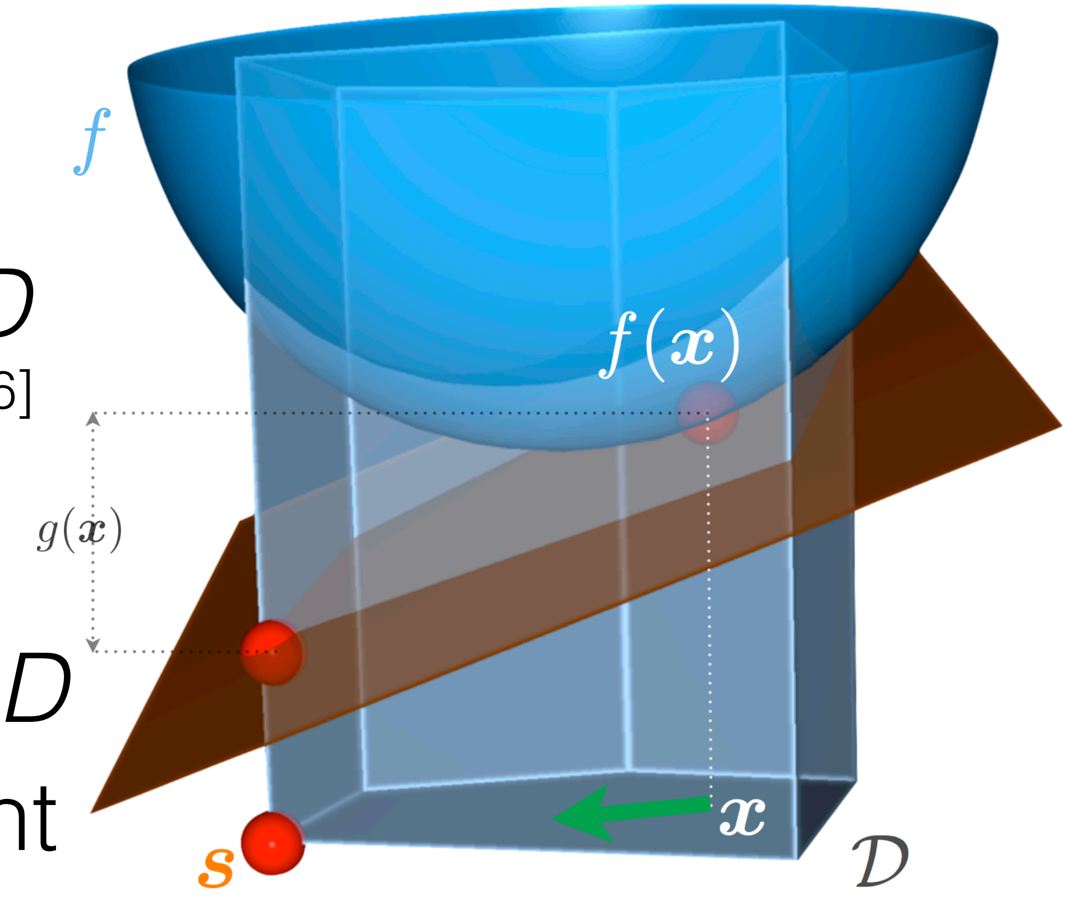
Convex optimization on a polytope *D*

[Frank, Wolfe 1956]

- Repeat:
    1. Find gradient
    2. Find argmin point on plane in *D*
    3. Do line search between current point and argmin point
- Convex combination of *M* vertices after *M* -1 steps
- Our problem: $\min_{w \in \mathbb{R}^N} \|\mathcal{L}(w) - \mathcal{L}\|^2$

$$\Delta^{N-1} := \left\{ w \in \mathbb{R}^N : \sum_{n=1}^{N} \sigma_n w_n = \sigma, w \geq 0 \right\}$$

Frank-Wolfe



[Jaggi 2013]

Convex optimization on a polytope *D*

[Frank, Wolfe 1956]

- Repeat:
  1. Find gradient
  2. Find argmin point on plane in *D*
  3. Do line search between current point and argmin point
- Convex combination of *M* vertices after *M* -1 steps
- Our problem: $\min_{w \in \mathbb{R}^N} \|\mathcal{L}(w) - \mathcal{L}\|^2$

$$\Delta^{N-1} := \left\{ w \in \mathbb{R}^N : \sum_{n=1}^N \sigma_n w_n = \sigma, w \geq 0 \right\}$$

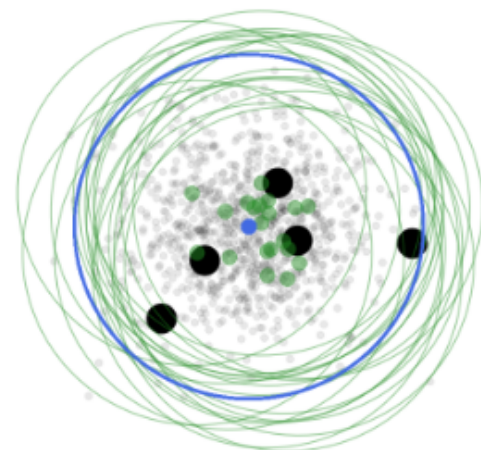
Thm (CB). After *M* iterations,

$$\|\mathcal{L}(w) - \mathcal{L}\| \leq \frac{c}{\sqrt{\alpha^{2M} + c'M}}$$

Gaussian model (simulated)

- 1K pts; norms, inference: closed-form

Uniform subsampling

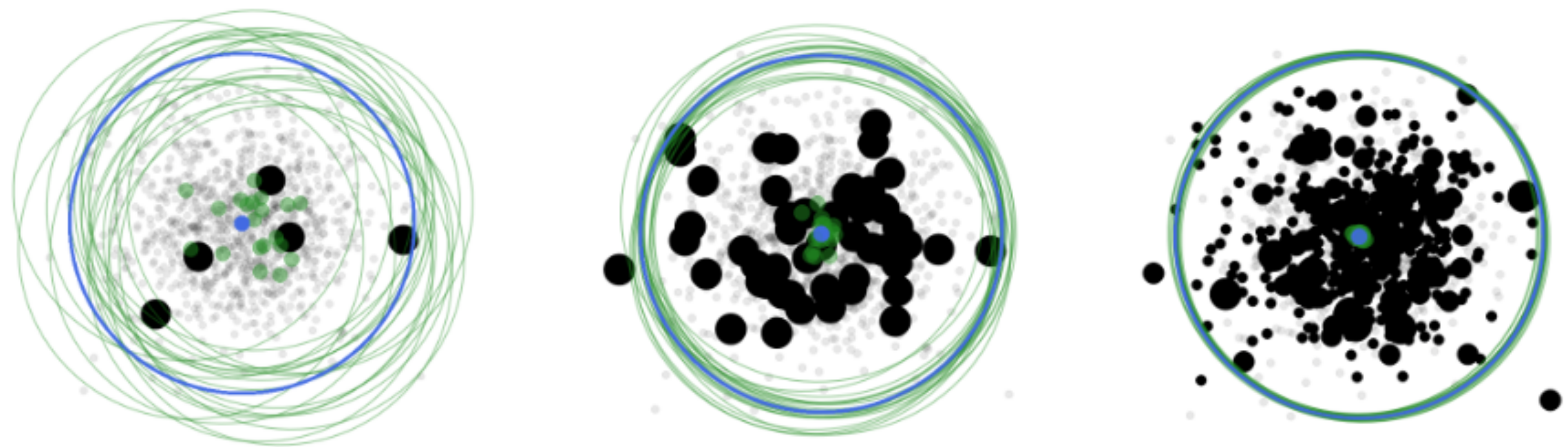


$M = 5$

Gaussian model (simulated)

- 1K pts; norms, inference: closed-form

Uniform subsampling



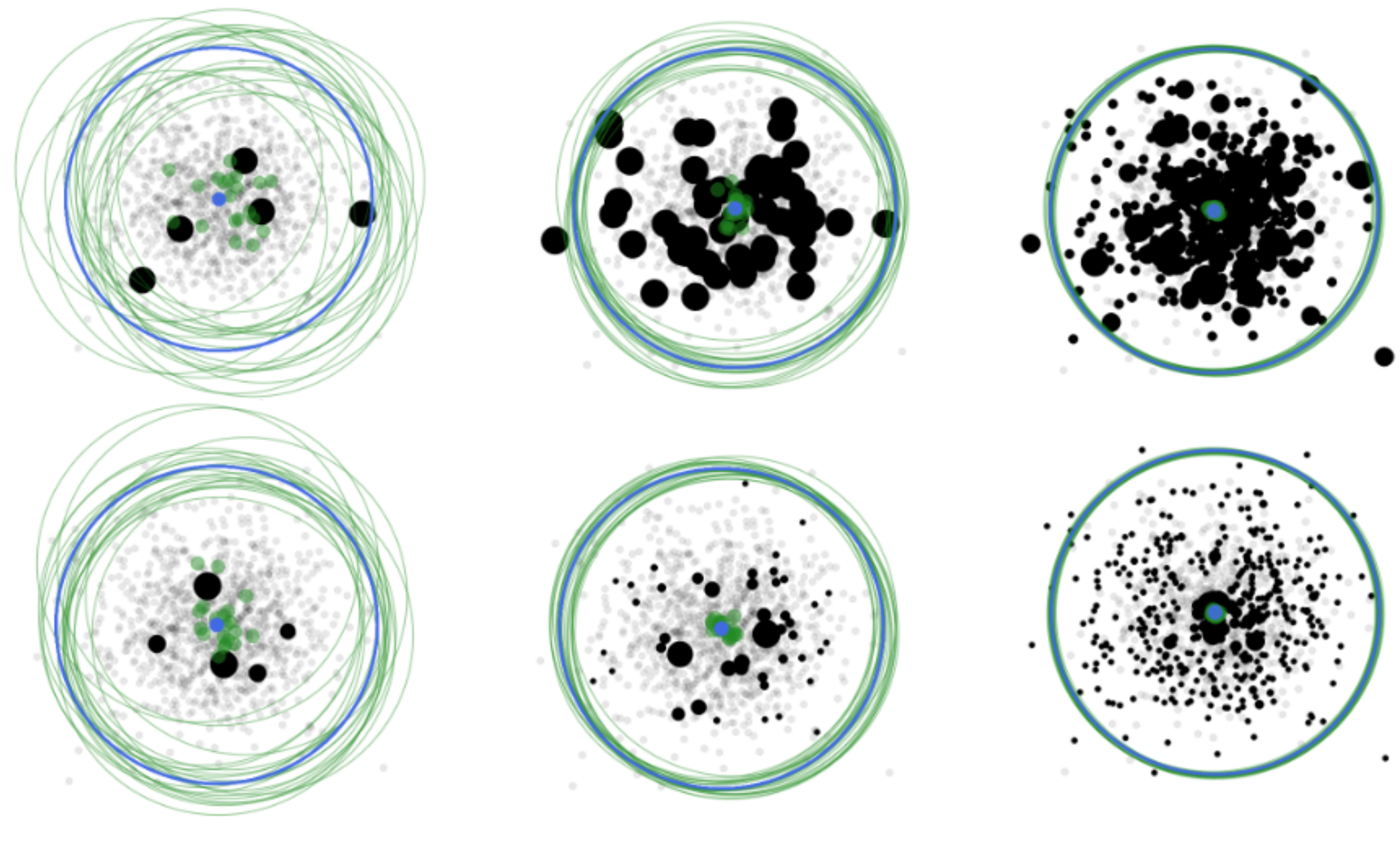
$M = 5$ $M = 50$ $M = 500$

Gaussian model (simulated)

- 1K pts; norms, inference: closed-form

Uniform subsampling

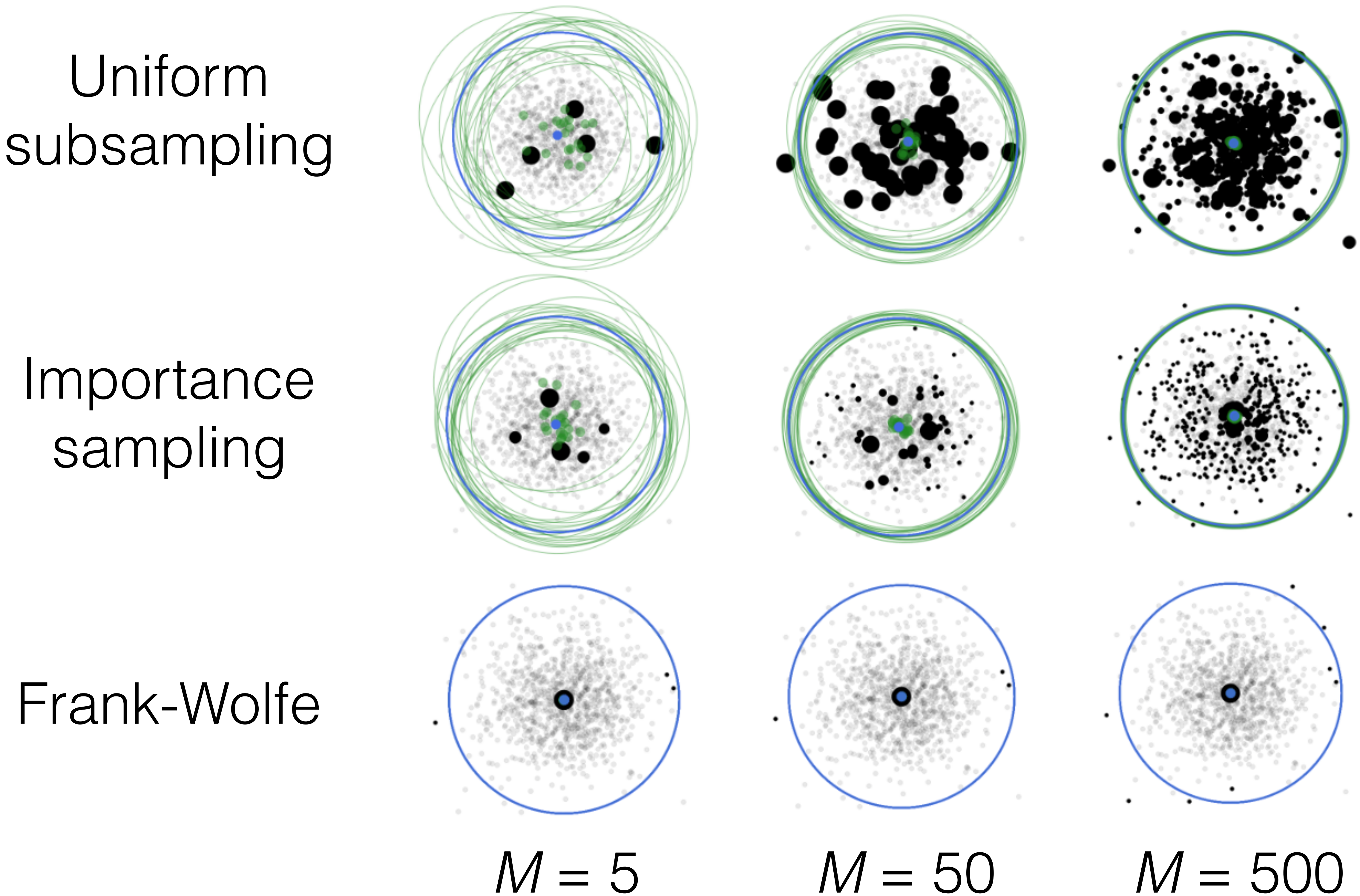
Importance sampling



$M = 5$ $M = 50$ $M = 500$

Gaussian model (simulated)

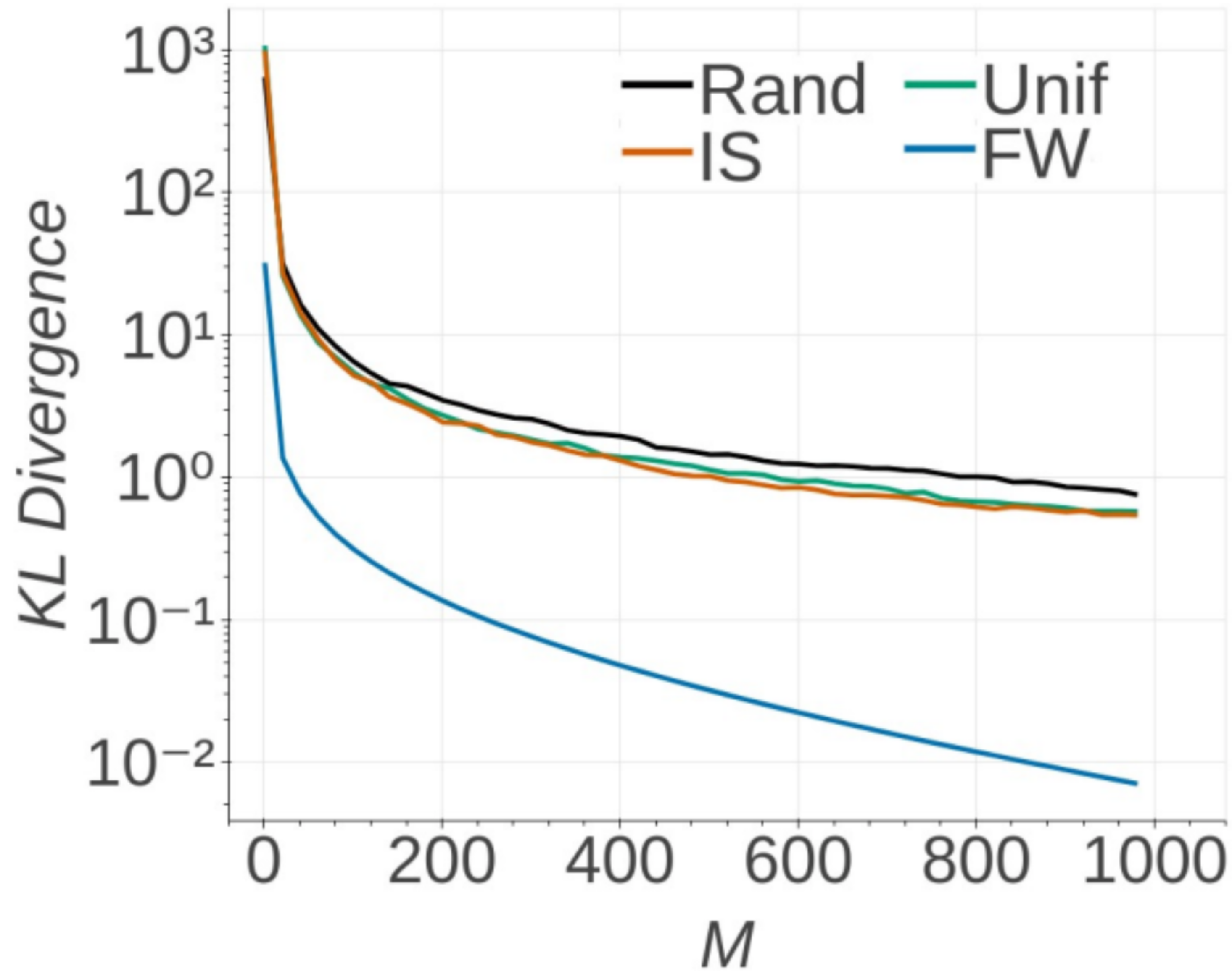
- 1K pts; norms, inference: closed-form



Gaussian model (simulated)

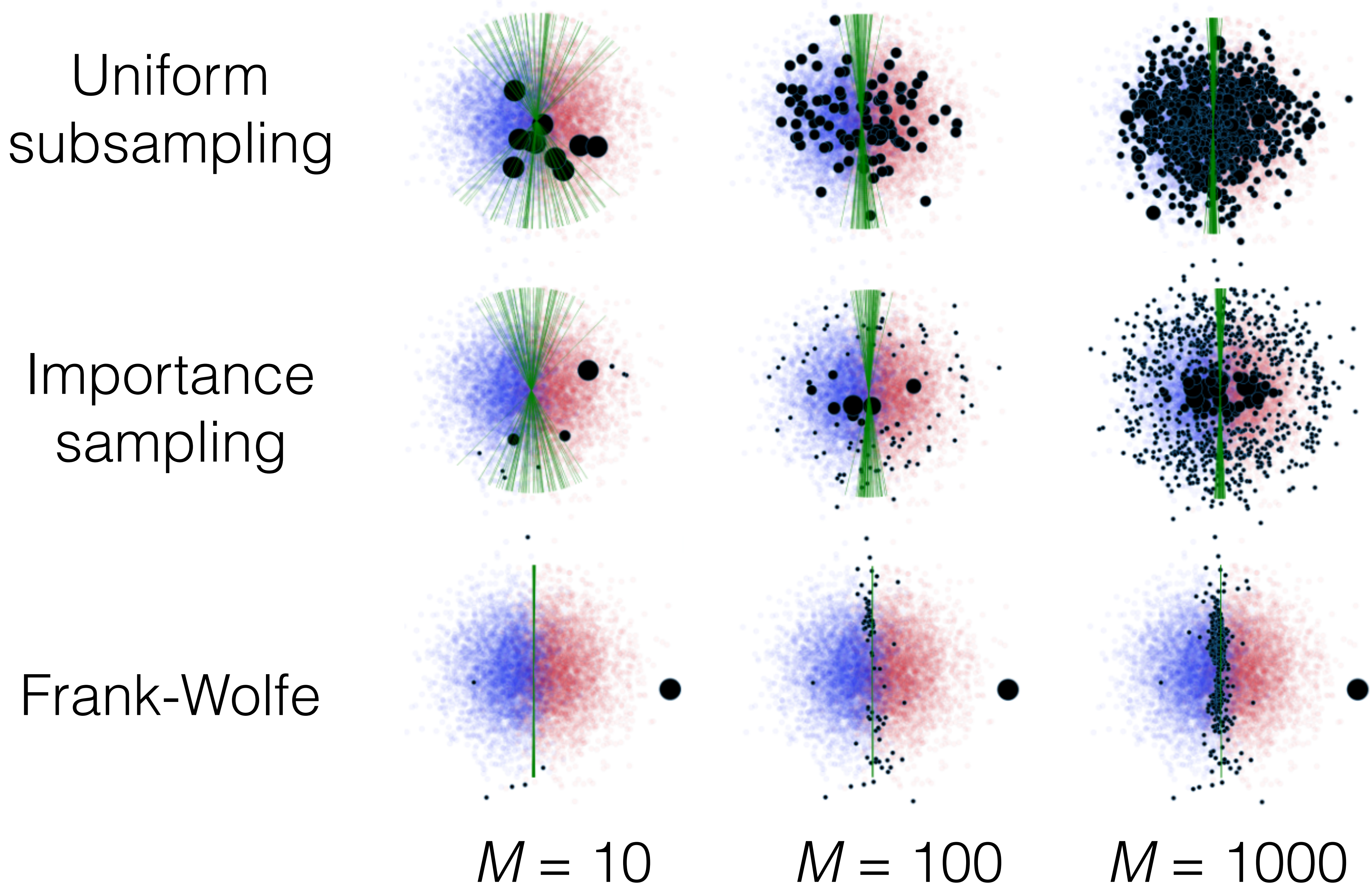
- 1K pts; norms, inference: closed-form

lower error



Logistic regression (simulated)

- 10K pts; general inference



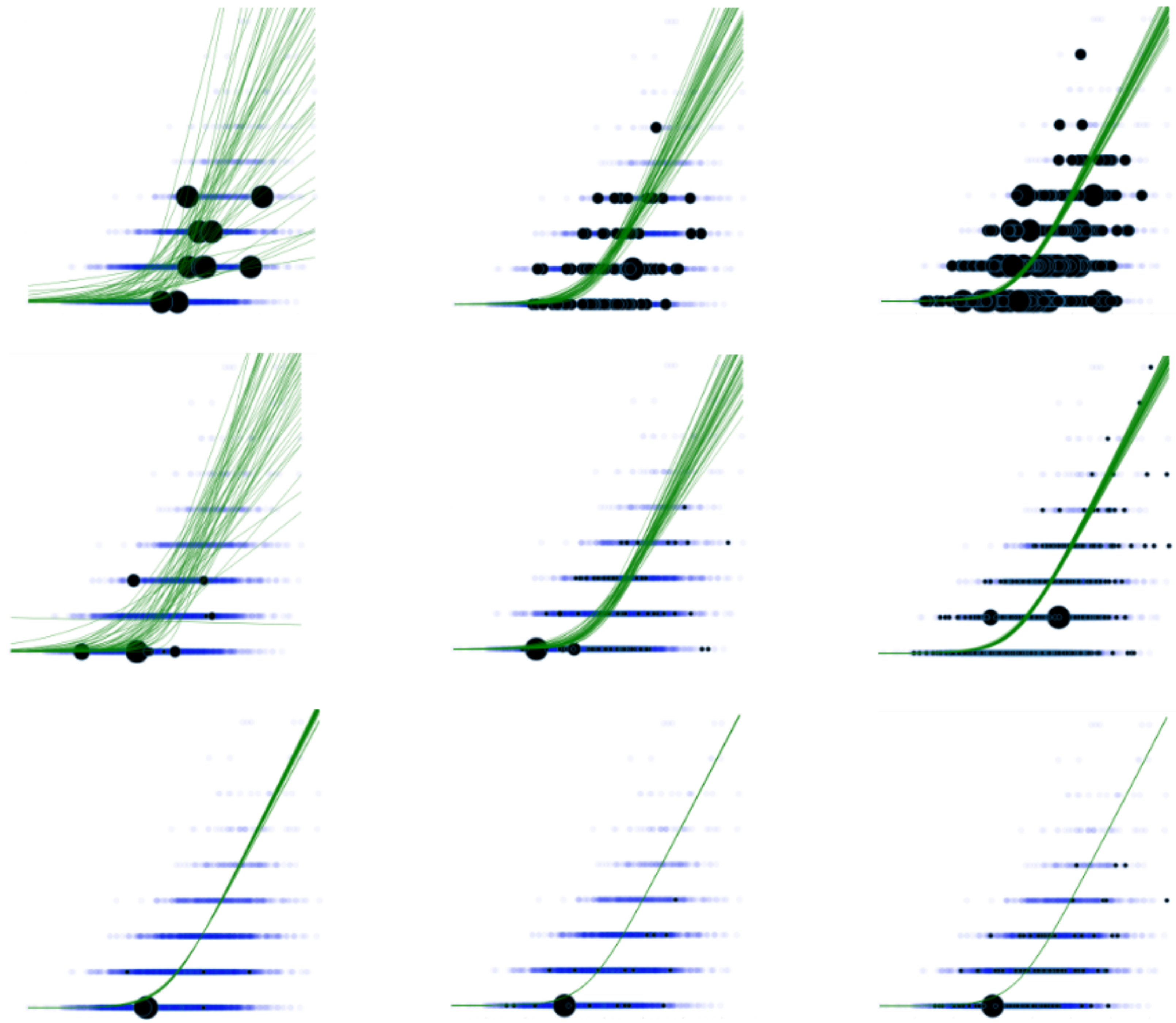
Poisson regression (simulated)

- 10K pts; general inference

Uniform subsampling

Importance sampling

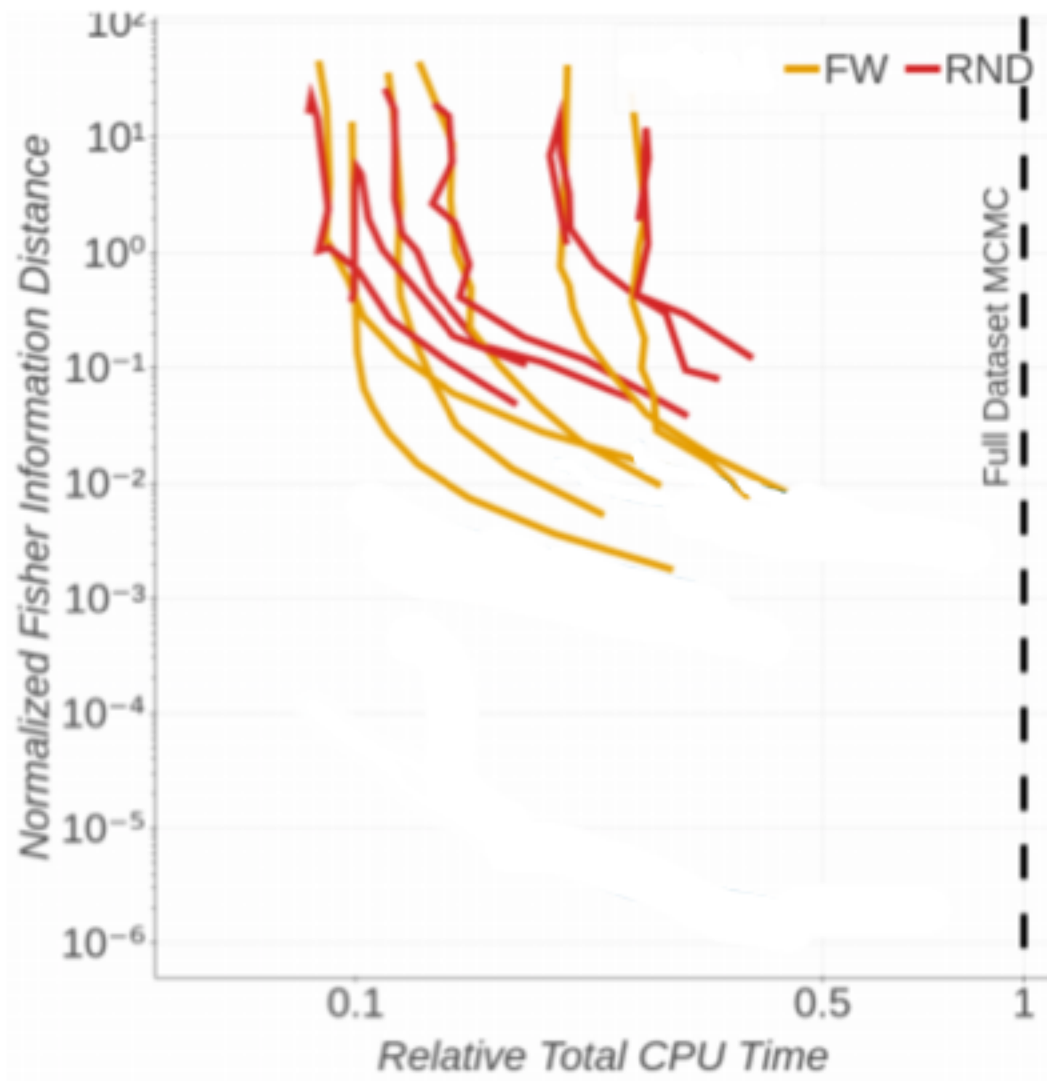
Frank-Wolfe



$M = 10$ $M = 100$ $M = 1000$

Real data experiments

lower error



less total time

Uniform subsampling

Frank Wolfe coresets

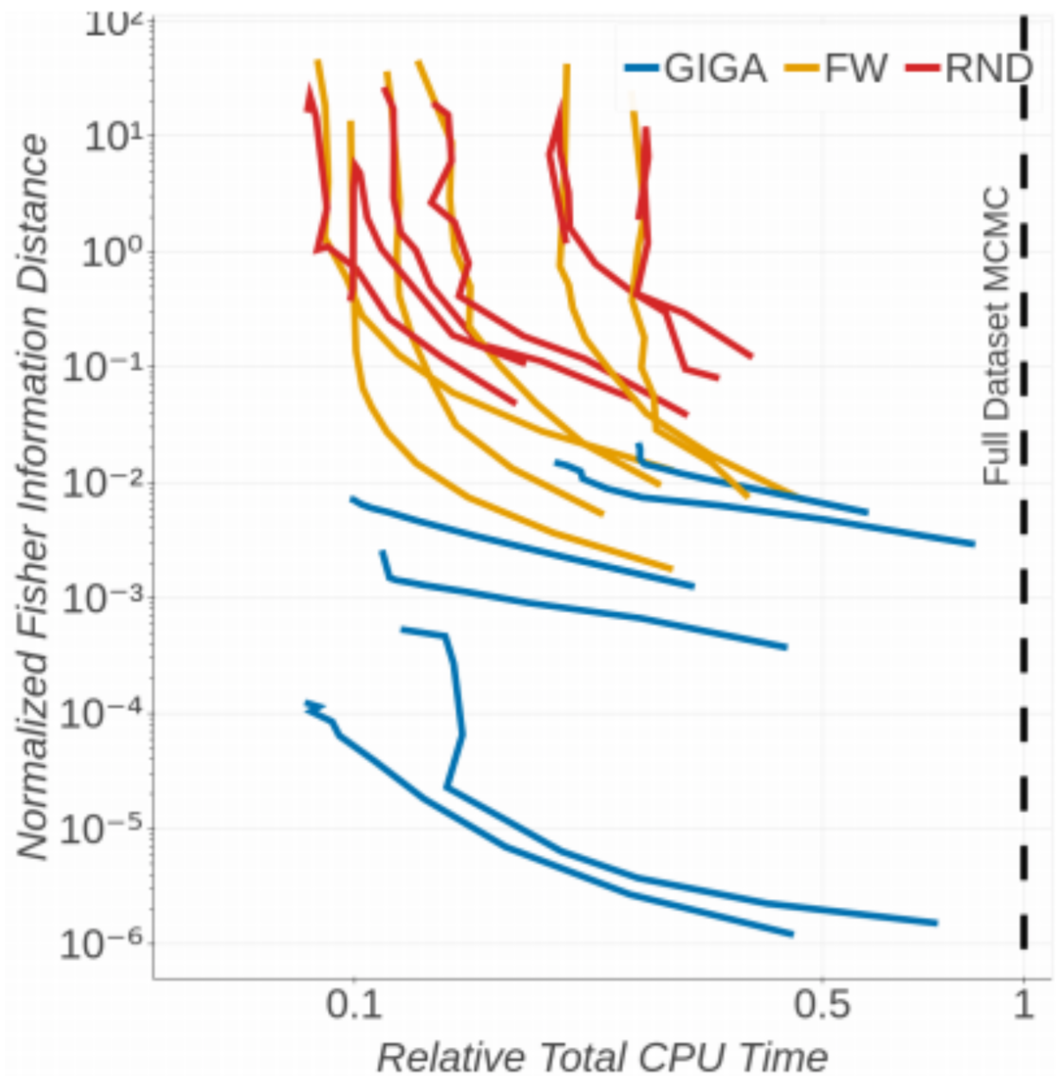
Data sets include:

- Phishing
- Chemical reactivity
- Bicycle trips
- Airport delays

[Campbell, Broderick 2019]

Real data experiments

lower error



less total time

Uniform subsampling

Frank Wolfe coresets

GIGA coresets

Data sets include:

- Phishing
- Chemical reactivity
- Bicycle trips
- Airport delays

[Campbell, Broderick 2019, 2018]

Roadmap

- The “core” of the data set
- Uniform data subsampling isn’t enough
- Importance sampling for “coresets”
- Optimization for “coresets”
- Approximate sufficient statistics

Roadmap

- The “core” of the data set
- Uniform data subsampling isn’t enough
- Importance sampling for “coresets”
- Optimization for “coresets”
- Approximate sufficient statistics

Data summarization

Data summarization

- Exponential family likelihood

Data summarization

- Exponential family likelihood

Sufficient statistics

$$p(y_{1:N}|x_{1:N},\theta) = \prod_{n=1}^{N} \exp\left[T(y_n,x_n)\cdot\eta(\theta)\right]$$

Data summarization

- Exponential family likelihood

Sufficient statistics

$$p(y_{1:N}|x_{1:N},\theta) = \prod_{n=1}^{N} \exp\left[T(y_n, x_n) \cdot \eta(\theta)\right]$$

$$= \exp\left[\left\{\sum_{n=1}^{N} T(y_n, x_n)\right\} \cdot \eta(\theta)\right]$$

Data summarization

- Exponential family likelihood

Sufficient statistics

$$p(y_{1:N}|x_{1:N}, \theta) = \prod_{n=1}^{N} \exp\left[T(y_n, x_n) \cdot \eta(\theta)\right]$$

$$= \exp\left[\left\{\sum_{n=1}^{N} T(y_n, x_n)\right\} \cdot \eta(\theta)\right]$$

- Scalable, single-pass, streaming, distributed, complementary to MCMC

Data summarization

- Exponential family likelihood

Sufficient statistics

$$p(y_{1:N}|x_{1:N}, \theta) = \prod_{n=1}^{N} \exp\left[T(y_n, x_n) \cdot \eta(\theta)\right]$$

$$= \exp\left[\left\{\sum_{n=1}^{N} T(y_n, x_n)\right\} \cdot \eta(\theta)\right]$$

- Scalable, single-pass, streaming, distributed, complementary to MCMC
- *But*: Often no simple sufficient statistics

Data summarization

- Exponential family likelihood

Sufficient statistics

$$p(y_{1:N}|x_{1:N},\theta) = \prod_{n=1}^{N} \exp\left[T(y_n, x_n) \cdot \eta(\theta)\right]$$

$$= \exp\left[\left\{\sum_{n=1}^{N} T(y_n, x_n)\right\} \cdot \eta(\theta)\right]$$

- Scalable, single-pass, streaming, distributed, complementary to MCMC
- *But*: Often no simple sufficient statistics
  - E.g. Bayesian logistic regression; GLMs; "deeper" models
    - Likelihood $p(y_{1:N}|x_{1:N},\theta) = \prod_{n=1}^{N} \frac{1}{1 + \exp(-y_n x_n \cdot \theta)}$

Data summarization

- Exponential family likelihood

Sufficient statistics

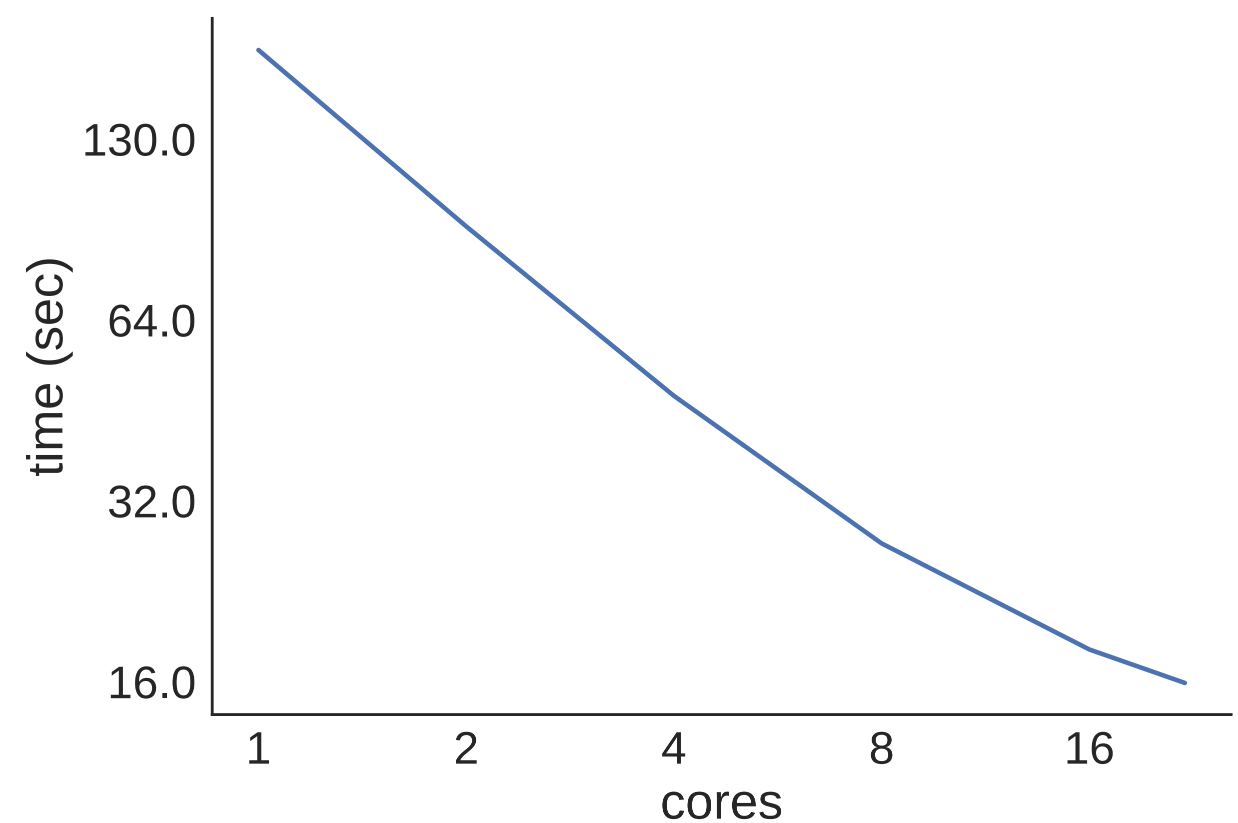
$$p(y_{1:N}|x_{1:N},\theta) = \prod_{n=1}^{N} \exp\left[T(y_n, x_n) \cdot \eta(\theta)\right]$$

$$= \exp\left[\left\{\sum_{n=1}^{N} T(y_n, x_n)\right\} \cdot \eta(\theta)\right]$$

- Scalable, single-pass, streaming, distributed, complementary to MCMC
- *But*: Often no simple sufficient statistics
  - E.g. Bayesian logistic regression; GLMs; “deeper” models
    - Likelihood $p(y_{1:N}|x_{1:N},\theta) = \prod_{n=1}^{N} \frac{1}{1+\exp(-y_n x_n \cdot \theta)}$
- Our proposal: (polynomial) *approximate* sufficient statistics

Data summarization

- 6M data points, 1000 features
- Streaming, distributed; minimal communication
- 22 cores, 16 sec
- Finite-data guarantees on Wasserstein distance to exact posterior



[Huggins, Adams, Broderick 2017]

Conclusions

- *Data summarization* for **scalable**, **automated** approximate Bayes algorithms with **error bounds on quality for finite data**

[Campbell, Broderick 2017, 2018; Huggins, Campbell, Broderick 2016; Huggins, Adams, Broderick 2017]

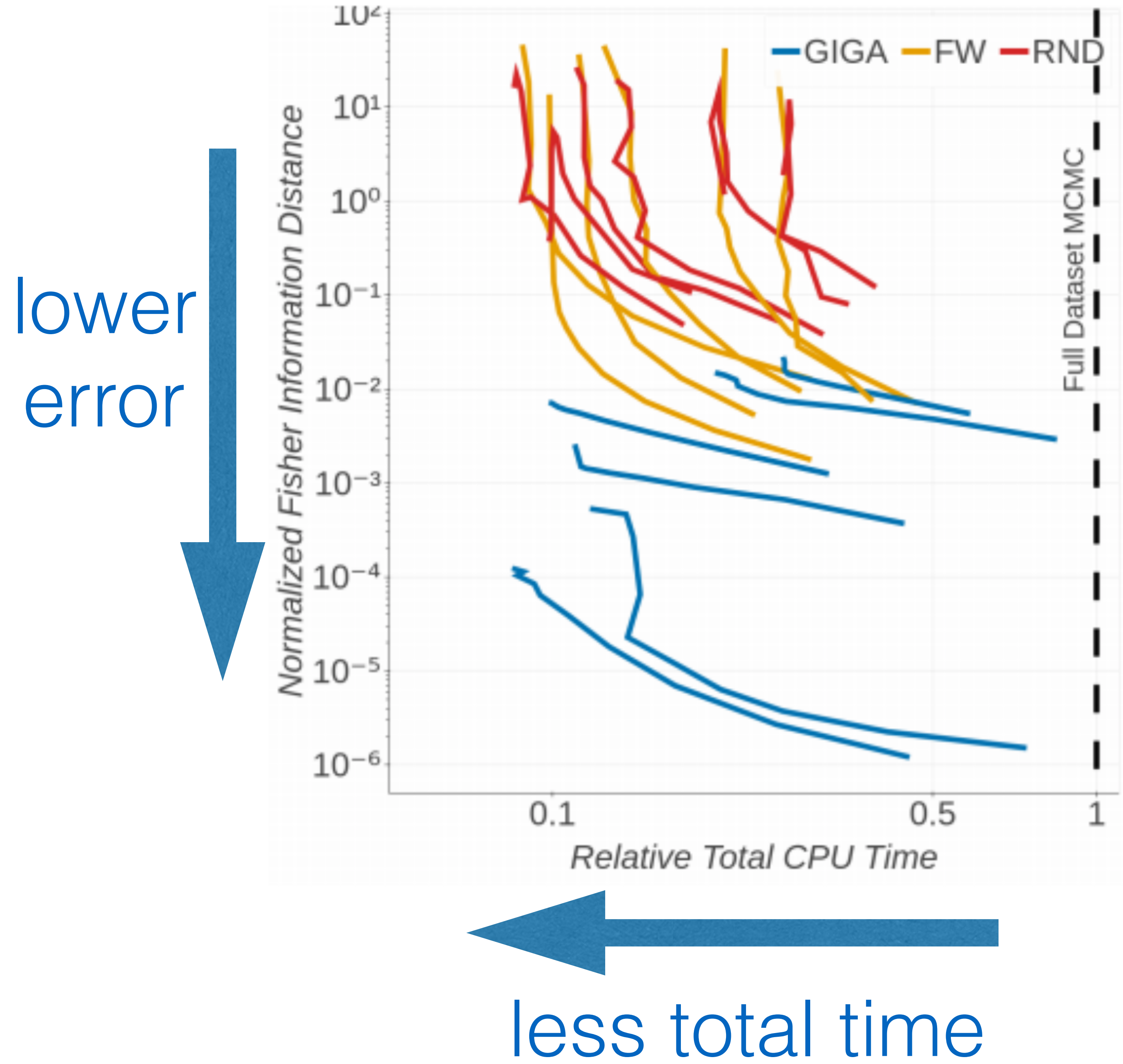
Conclusions

- *Data summarization* for **scalable**, **automated** approximate Bayes algorithms with **error bounds on quality for finite data**
  - Coresets
  - Approx. suff. statistics

[Campbell, Broderick 2017, 2018; Huggins, Campbell, Broderick 2016; Huggins, Adams, Broderick 2017]

Conclusions

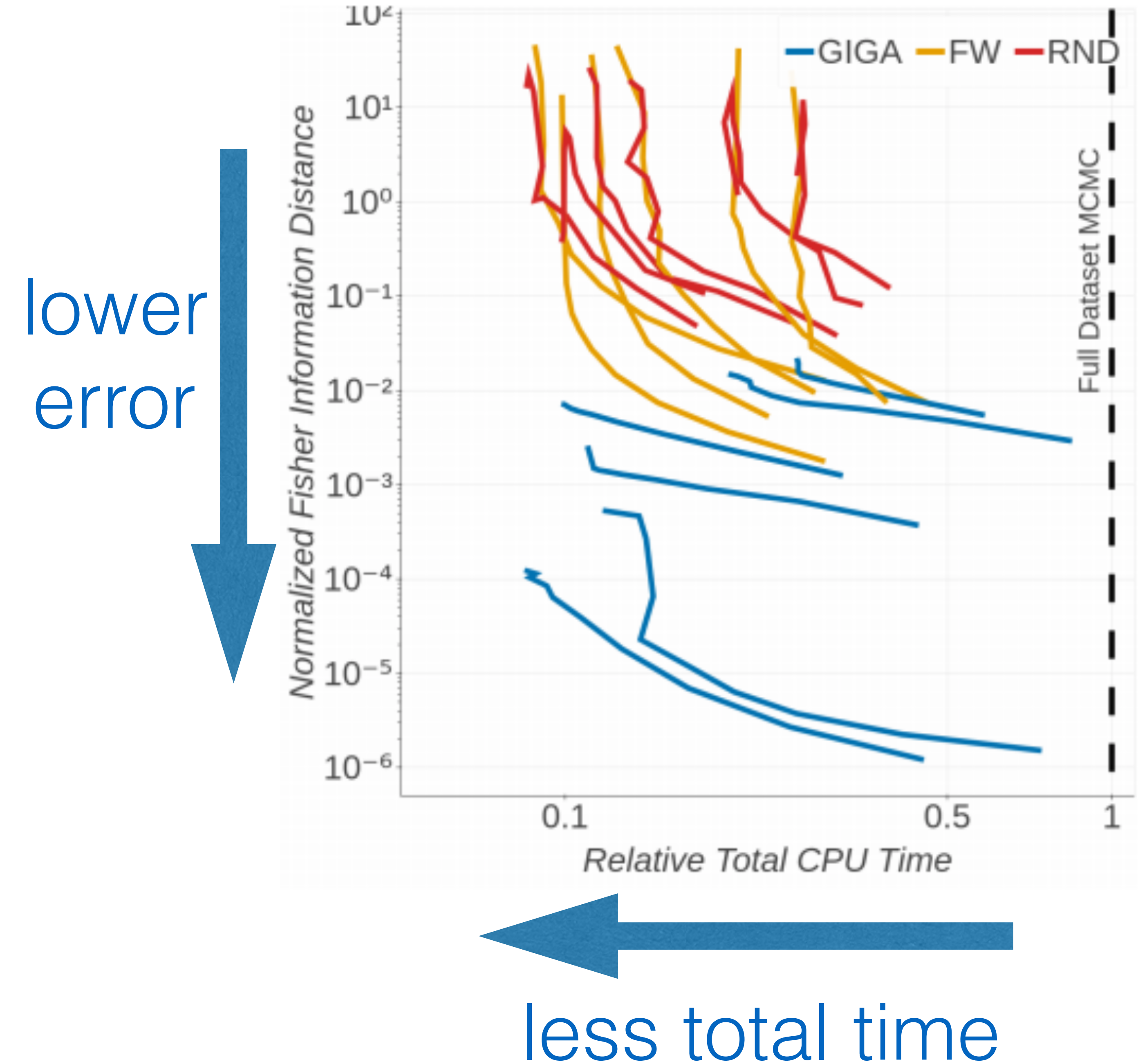
- *Data summarization* for **scalable**, **automated** approximate Bayes algorithms with **error bounds on quality for finite data**
  - Coresets
  - Approx. suff. statistics



[Campbell, Broderick 2017, 2018; Huggins, Campbell, Broderick 2016; Huggins, Adams, Broderick 2017]

Conclusions

- *Data summarization* for **scalable**, **automated** approximate Bayes algorithms with **error bounds on quality for finite data**
  - Coresets
  - Approx. suff. statistics
  - More accurate with more computation investment

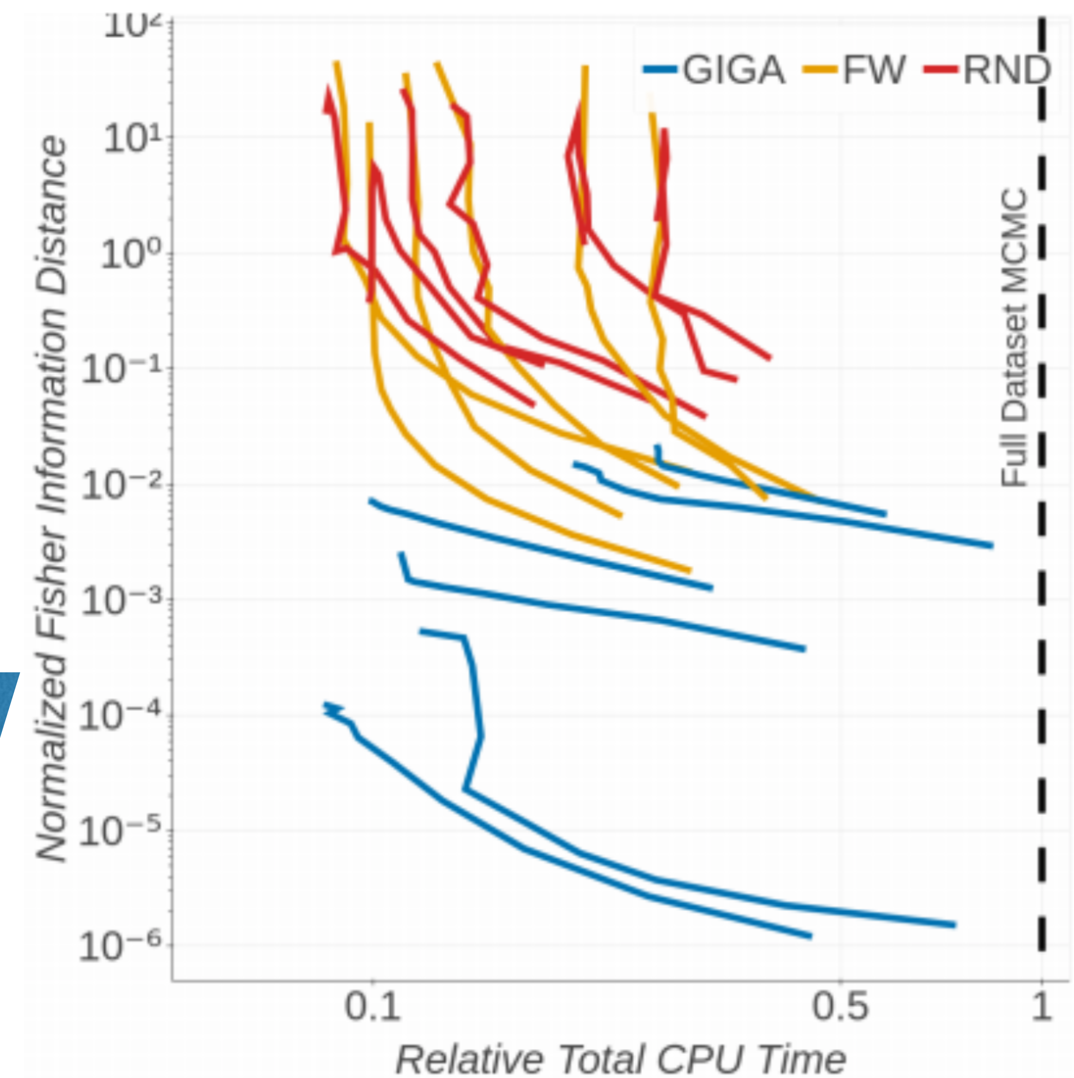


[Campbell, Broderick 2017, 2018; Huggins, Campbell, Broderick 2016; Huggins, Adams, Broderick 2017]

Conclusions

- *Data summarization* for **scalable**, **automated** approximate Bayes algorithms with **error bounds on quality for finite data**
  - Coresets
  - Approx. suff. statistics
  - More accurate with more computation investment
- A start
  - Lots of potential improvements/ directions

lower error



less total time

[Campbell, Broderick 2017, 2018; Huggins, Campbell, Broderick 2016; Huggins, Adams, Broderick 2017]

References

T Campbell and T Broderick. Automated scalable Bayesian inference via Hilbert coresets. *Journal of Machine Learning Research*, 2019.

T Campbell and T Broderick. Bayesian coreset construction via greedy iterative geodesic ascent. *ICML* 2018.

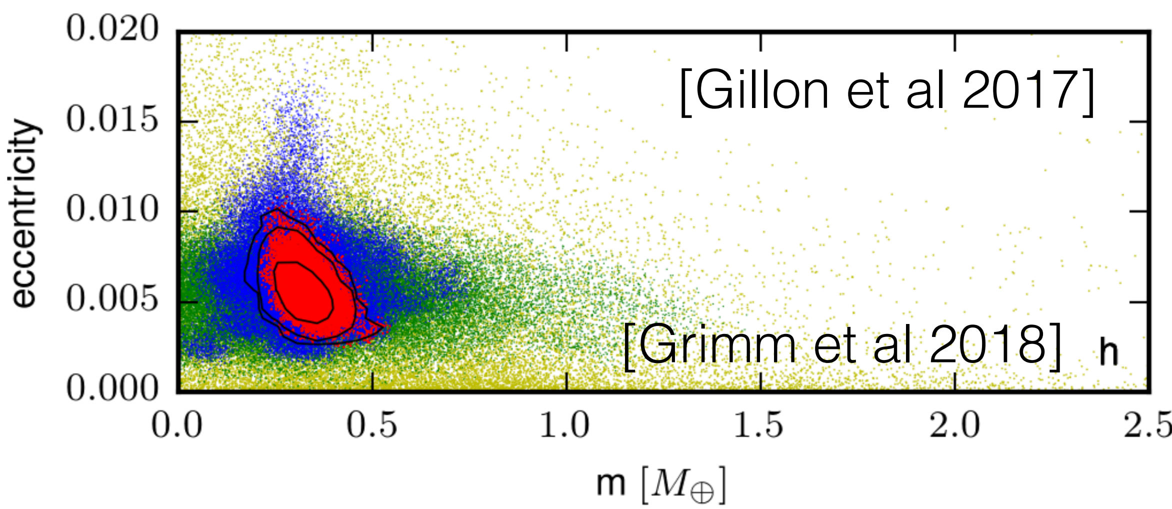
JH Huggins, T Campbell, M Kasprzak, and T Broderick. Practical bounds on the error of Bayesian posterior approximations: A nonasymptotic approach. ArXiv:1809.09505.

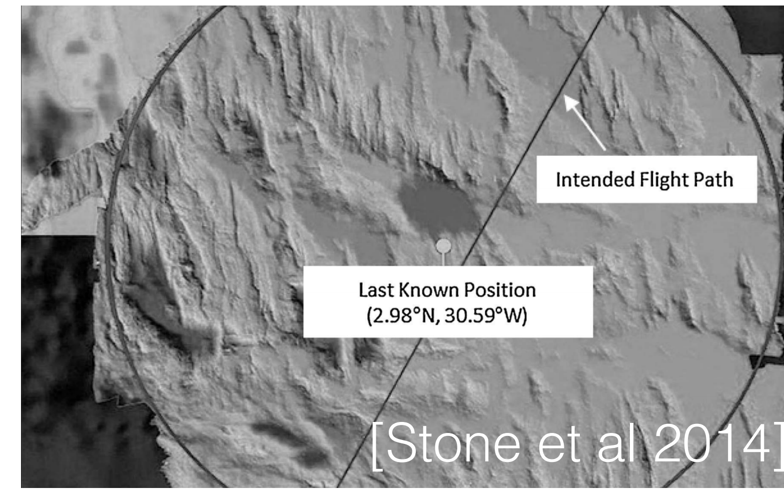
JH Huggins, T Campbell, and T Broderick. Coresets for scalable Bayesian logistic regression. *NeurIPS* 2016.

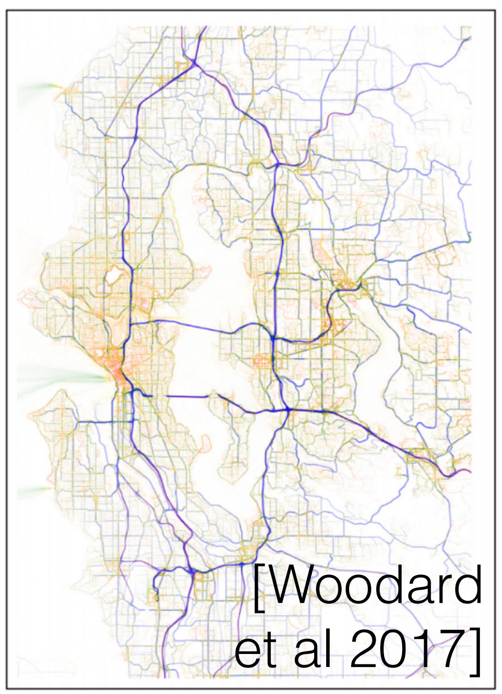
JH Huggins, RP Adams, and T Broderick. PASS-GLM: Polynomial approximate sufficient statistics for scalable Bayesian GLM inference. *NeurIPS* 2017.

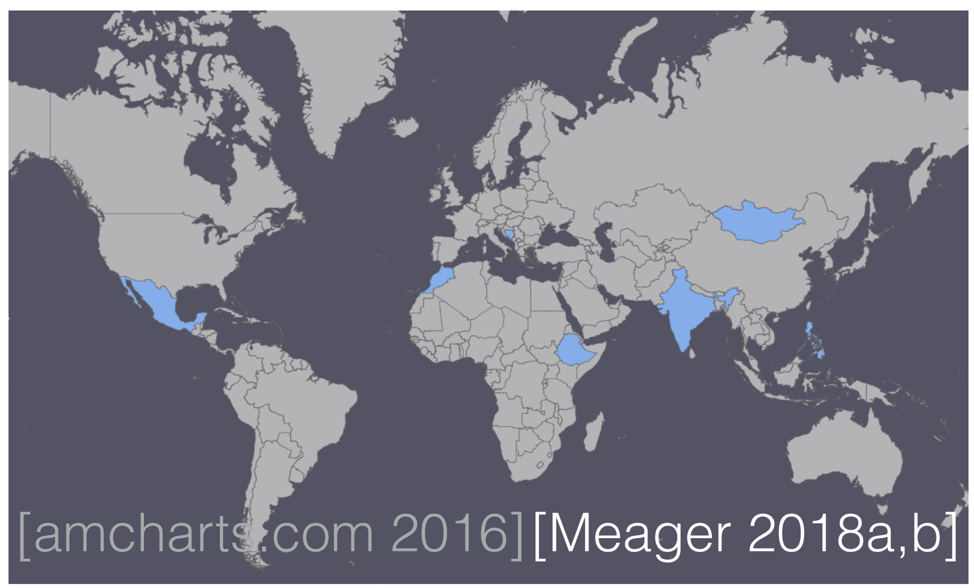
R Agrawal, T Campbell, JH Huggins, and T Broderick. Data-dependent compression of random features for large-scale kernel approximation. *AISTATS* 2019.

Bayesian inference



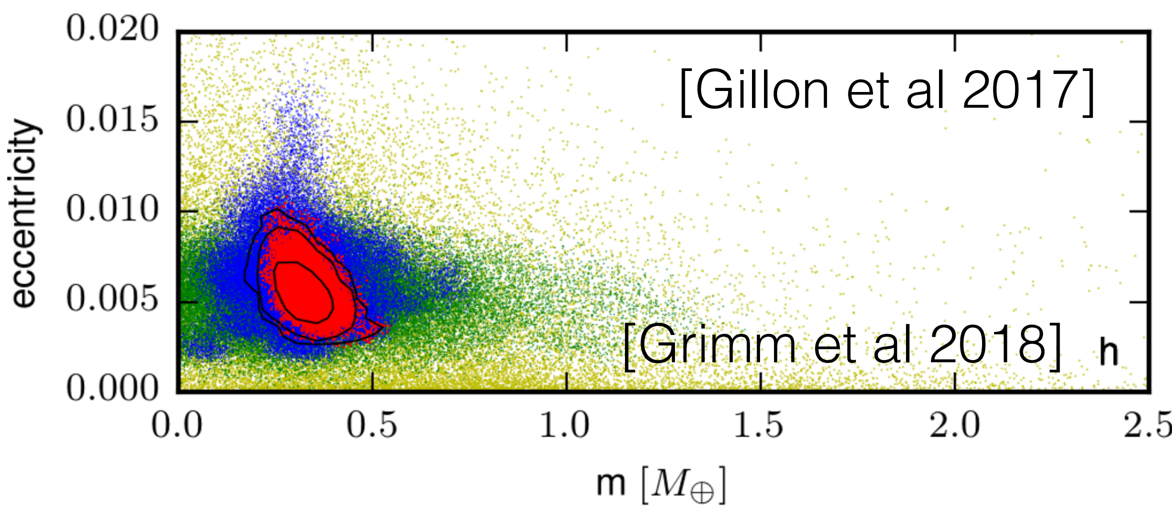


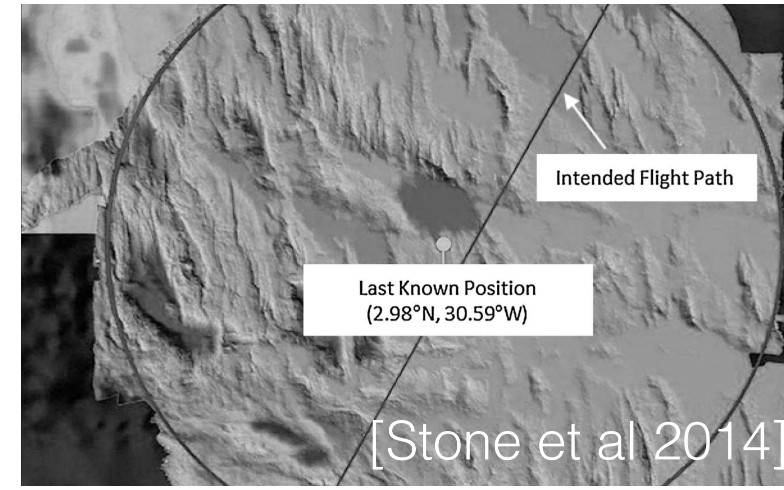


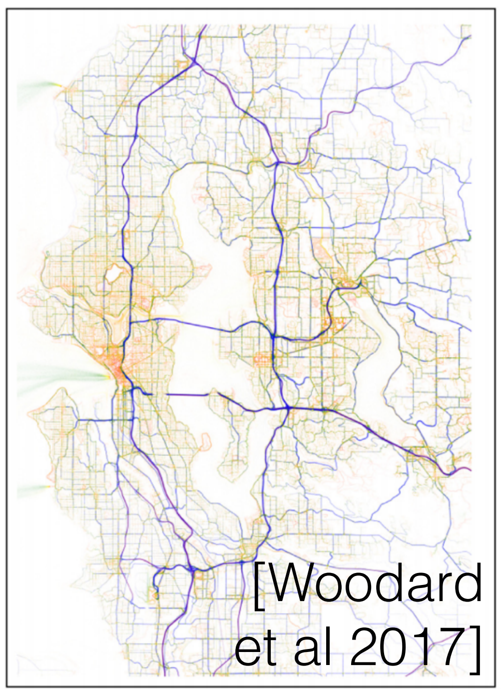


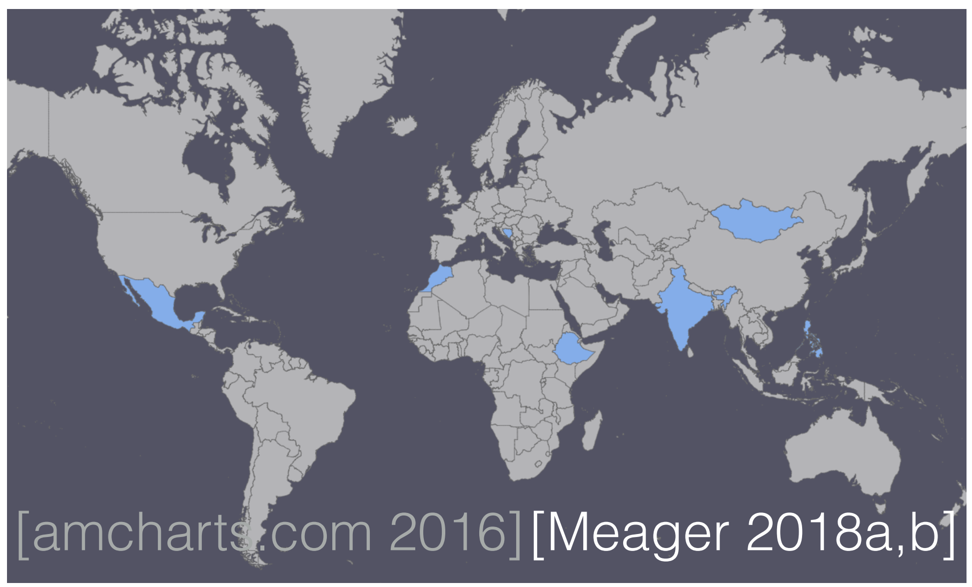
[mc-stan.org]

Bayesian inference





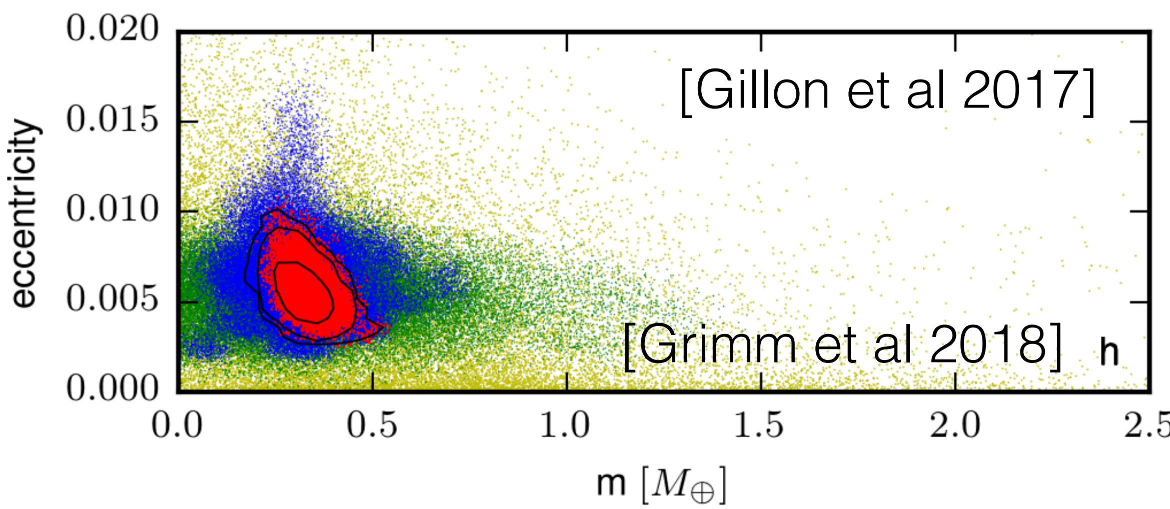


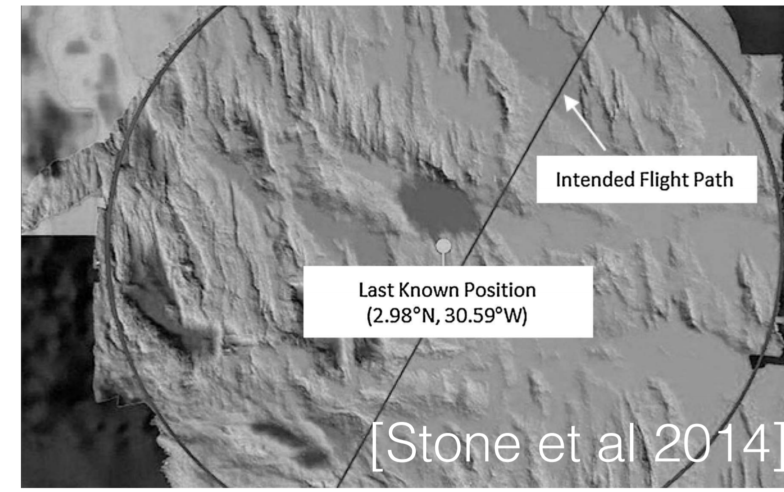


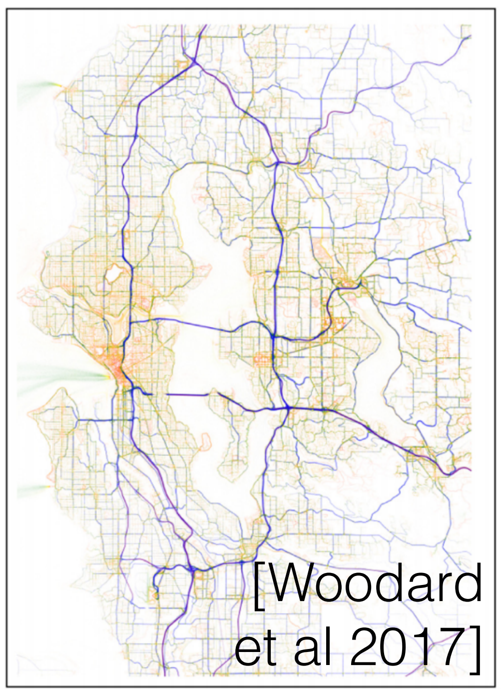
[mc-stan.org]

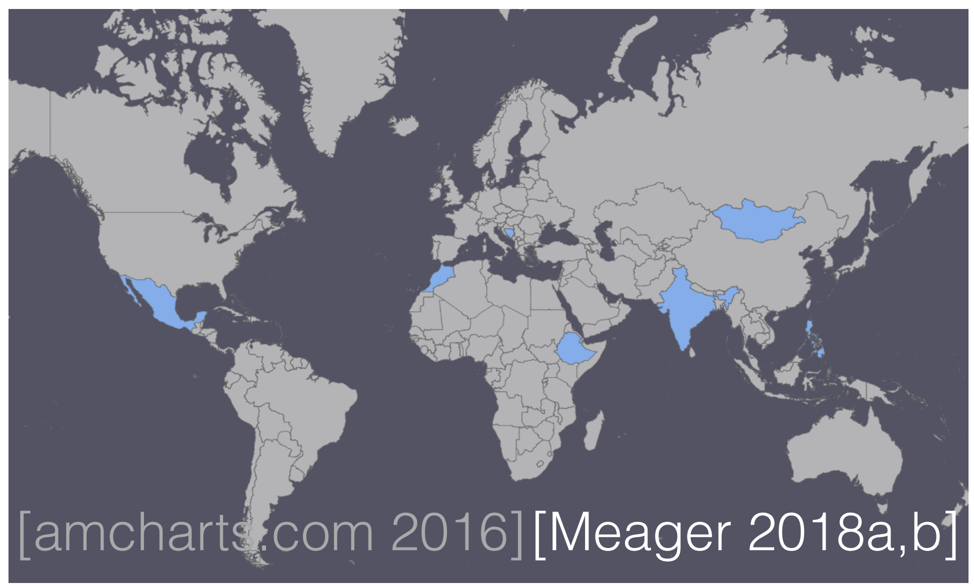
- Challenge: fast (compute, user), reliable inference

Bayesian inference





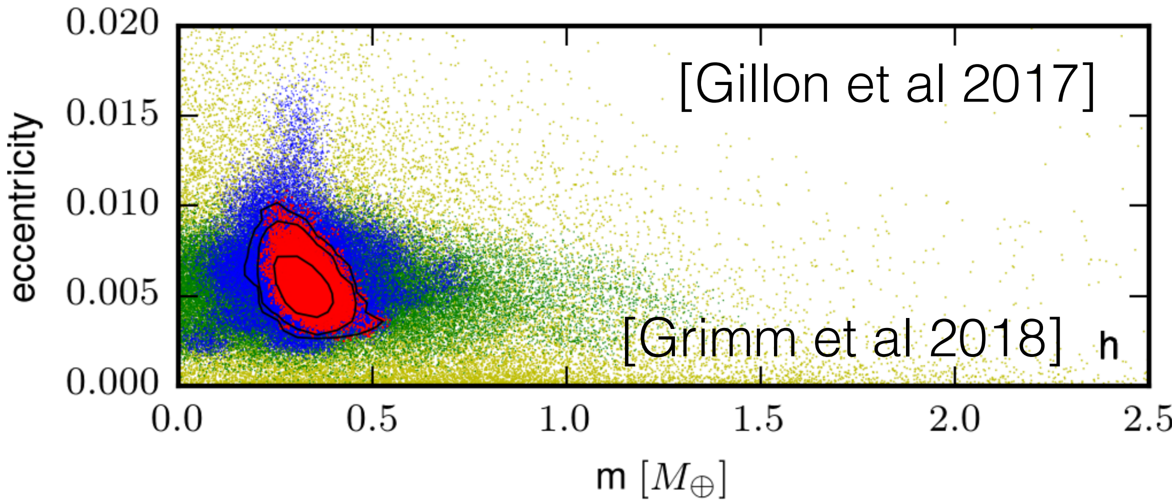


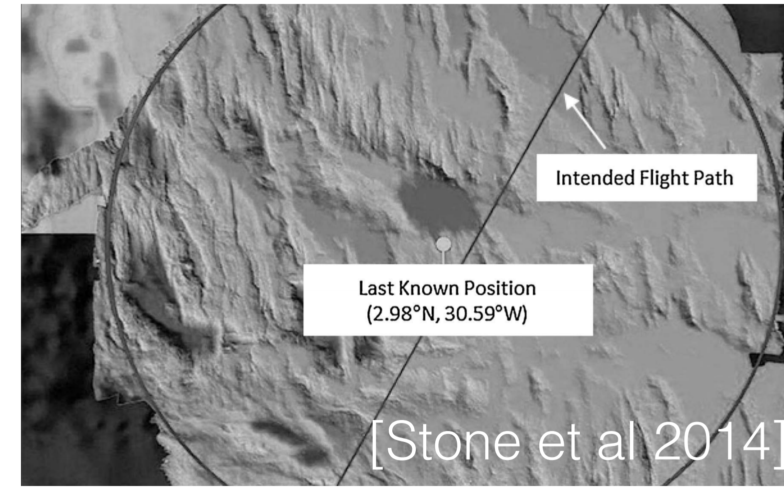


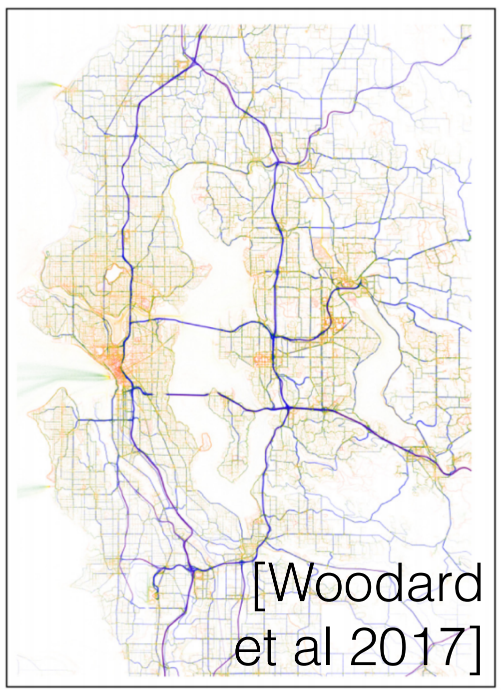
[mc-stan.org]

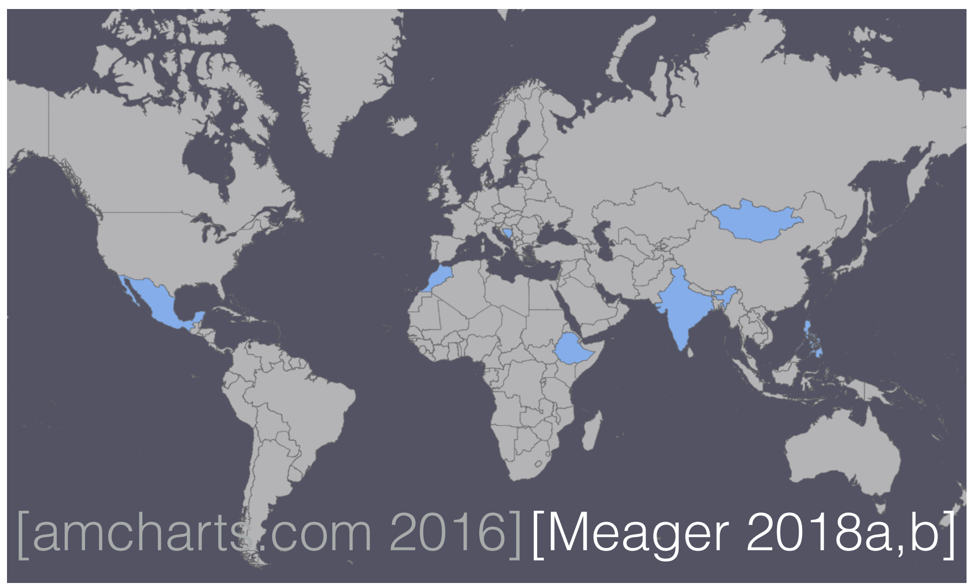
- Challenge: fast (compute, user), reliable inference
- Today: variational Bayes and beyond

Bayesian inference









[mc-stan.org]

- Challenge: fast (compute, user), reliable inference
- Today: variational Bayes and beyond
- **Fundamental questions**
  - **What is achievable in speed and accuracy?**

References (1/6)

R Bardenet, A Doucet, and C Holmes. "On Markov chain Monte Carlo methods for tall data." *Journal of Machine Learning Research* 18.1 (2017): 1515-1557.

AG Baydin, BA Pearlmutter, AA Radul, and JM Siskind. "Automatic differentiation in machine learning: a survey." *Journal of Machine Learning Research*, 2018.

DM Blei, A Kucukelbir, and JD McAuliffe. "Variational inference: A review for statisticians." *Journal of the American Statistical Association* 112.518 (2017): 859-877.

T Broderick, N Boyd, A Wibisono, AC Wilson, and MI Jordan. Streaming variational Bayes. *NeurIPS* 2013.

CM Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 2006.

T Campbell and T Broderick. Automated scalable Bayesian inference via Hilbert coresets. *Journal of Machine Learning Research*, 2019.

T Campbell and T Broderick. Bayesian Coreset Construction via Greedy Iterative Geodesic Ascent. *ICML* 2018.

RJ Giordano, T Broderick, and MI Jordan. "Linear response methods for accurate covariance estimates from mean field variational Bayes." *NeurIPS* 2015.

R Giordano, T Broderick, R Meager, J Huggins, and MI Jordan. "Fast robustness quantification with variational Bayes." *ICML 2016 Workshop on #Data4Good: Machine Learning in Social Good Applications*, 2016.

RJ Giordano, T Broderick, and MI Jordan. Covariances, robustness, and variational Bayes. *Journal of Machine Learning Research*, 2018.

References (2/6)

J Gorham and L Mackey. "Measuring sample quality with Stein's method." *NeurIPS* 2015.

J Gorham, and L Mackey. "Measuring sample quality with kernels." ArXiv:1703.01717 (2017).

PD Hoff. *A first course in Bayesian statistical methods*. Springer Science & Business Media, 2009.

MD Hoffman, DM Blei, C Wang, and J Paisley. "Stochastic variational inference." *The Journal of Machine Learning Research* 14.1 (2013): 1303-1347.

JH Huggins, T Campbell, and T Broderick. Coresets for scalable Bayesian logistic regression. *NeurIPS* 2016.

JH Huggins, RP Adams, and T Broderick. PASS-GLM: Polynomial approximate sufficient statistics for scalable Bayesian GLM inference. *NeurIPS* 2017.

J Huggins, T Campbell, M Kasprzak, T Broderick. Practical bounds on the error of Bayesian posterior approximations: A nonasymptotic approach, 2018. ArXiv:1809.09505.

A Kucukelbir, R Ranganath, A Gelman, and D Blei. Automatic variational inference in Stan. *NeurIPS* 2015.

A Kucukelbir, D Tran, R Ranganath, A Gelman, and DM Blei. Automatic differentiation variational inference. *Journal of Machine Learning Research* 18.1 (2017): 430-474.

DJC MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

Stan (open source software). http://mc-stan.org/ Accessed: 2018.

References (3/6)

S Talts, M Betancourt, D Simpson, A Vehtari, and A Gelman. Validating Bayesian Inference Algorithms with Simulation-Based Calibration. ArXiv:1804.06788 (2018).

RE Turner and M Sahani. Two problems with variational expectation maximisation for time-series models. In D Barber, AT Cemgil, and S Chiappa, editors, *Bayesian Time Series Models*, 2011.

Y Yao, A Vehtari, D Simpson, and A Gelman. Yes, but Did It Work?: Evaluating Variational Inference. *ICML* 2018.

Application References (4/6)

Abbott, Benjamin P., et al. "Observation of gravitational waves from a binary black hole merger." *Physical Review Letters* 116.6 (2016): 061102.

Abbott, Benjamin P., et al. "The rate of binary black hole mergers inferred from advanced LIGO observations surrounding GW150914." *The Astrophysical Journal Letters* 833.1 (2016): L1.

Airoldi, Edoardo M., David M. Blei, Stephen E. Fienberg, and Eric P. Xing. "Mixed membership stochastic blockmodels." *Journal of Machine Learning Research* 9.Sep (2008): 1981-2014.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet allocation." *Journal of Machine Learning Research* 3.Jan (2003): 993-1022.

Chati, Yashovardhan Sushil, and Hamsa Balakrishnan. "A Gaussian process regression approach to model aircraft engine fuel flow rate." *Cyber-Physical Systems (ICCPS), 2017 ACM/IEEE 8th International Conference on*. IEEE, 2017.

Gershman, Samuel J., David M. Blei, Kenneth A. Norman, and Per B. Sederberg. "Decomposing spatiotemporal brain patterns into topographic latent sources." NeuroImage 98 (2014): 91-102.

Gillon, Michaël, et al. "Seven temperate terrestrial planets around the nearby ultracool dwarf star TRAPPIST-1." *Nature* 542.7642 (2017): 456.

Grimm, Simon L., et al. "The nature of the TRAPPIST-1 exoplanets." *Astronomy & Astrophysics* 613 (2018): A68.

Application References (5/6)

Grogan Jr, William L., and Willis W. Wirth. "A new American genus of predaceous midges related to Palpomyia and Bezzia (Diptera: Ceratopogonidae). Un nuevo género Americano de purrujas depredadoras relacionadas con Palpomyia y Bezzia (Diptera: Ceratopogonidae)." *Proceedings of the Biological Society of Washington*. 94.4 (1981): 1279-1305.

Kuikka, Sakari, Jarno Vanhatalo, Henni Pulkkinen, Samu Mäntyniemi, and Jukka Corander. "Experiences in Bayesian inference in Baltic salmon management." *Statistical Science* 29.1 (2014): 42-49.

Meager, Rachael. "Understanding the impact of microcredit expansions: A Bayesian hierarchical analysis of 7 randomized experiments." *AEJ: Applied*, to appear, 2018a.

Meager, Rachael. "Aggregating Distributional Treatment Effects: A Bayesian Hierarchical Analysis of the Microcredit Literature." Working paper, 2018b.

Stegle, Oliver, Leopold Parts, Richard Durbin, and John Winn. "A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies." PLoS computational biology 6.5 (2010): e1000770.

Stone, Lawrence D., Colleen M. Keller, Thomas M. Kratzke, and Johan P. Strumpfer. "Search for the wreckage of Air France Flight AF 447." *Statistical Science* (2014): 69-80.

Woodard, Dawn, Galina Nogin, Paul Koch, David Racz, Moises Goldszmidt, and Eric Horvitz. "Predicting travel time reliability using mobile phone GPS data." *Transportation Research Part C: Emerging Technologies* 75 (2017): 30-44.

Xing, Eric P., Wei Wu, Michael I. Jordan, and Richard M. Karp. "LOGOS: a modular Bayesian model for de novo motif detection." Journal of Bioinformatics and Computational Biology 2.01 (2004): 127-154.

Additional image references (6/6)

amCharts. Visited Countries Map. https://www.amcharts.com/visited_countries/ Accessed: 2016.

Baltic Salmon Fund. https://www.en.balticsalmonfund.org/about_us Accessed: 2018.

ESO/L. Calçada/M. Kornmesser. 16 October 2017, 16:00:00. Obtained from: https://commons.wikimedia.org/wiki/File:Artist%E2%80%99s_impression_of_merging_neutron_stars.jpg || Source: https://www.eso.org/public/images/eso1733a/ (Creative Commons Attribution 4.0 International License)

J. Herzog. 3 June 2016, 17:17:30. Obtained from: https://commons.wikimedia.org/wiki/File:Airbus_A350-941_F-WWCF_MSN002_ILA_Berlin_2016_17.jpg (Creative Commons Attribution 4.0 International License)

E. Xing. 2003. Slides "LOGOS: a modular Bayesian model for de novo motif detection." Obtained from: https://www.cs.cmu.edu/~epxing/papers/Old_papers/slide_CSB03/CSB1.pdf Accessed: 2018.