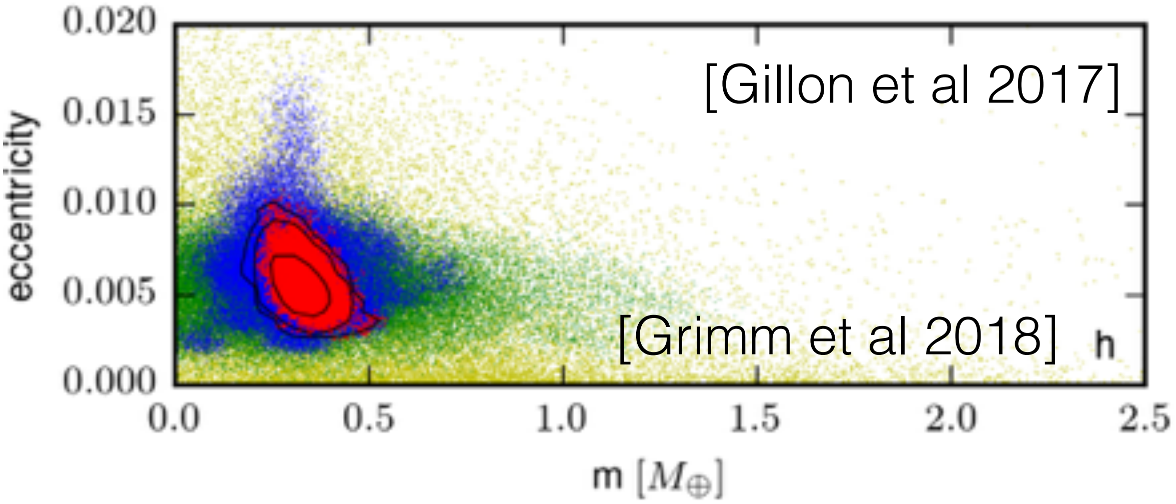
Variational Bayes and beyond: Bayesian inference for big data

Tamara Broderick
ITT Career Development
Assistant Professor,
MIT

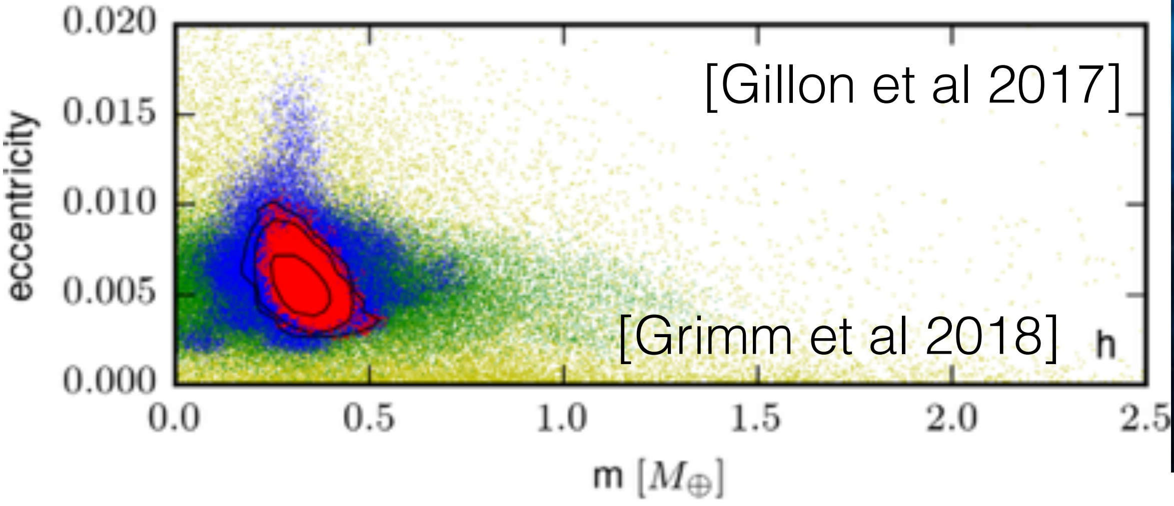
http://www.tamarabroderick.com/tutorial_2018_icml.html

Bayesian inference

Bayesian inference



Bayesian inference



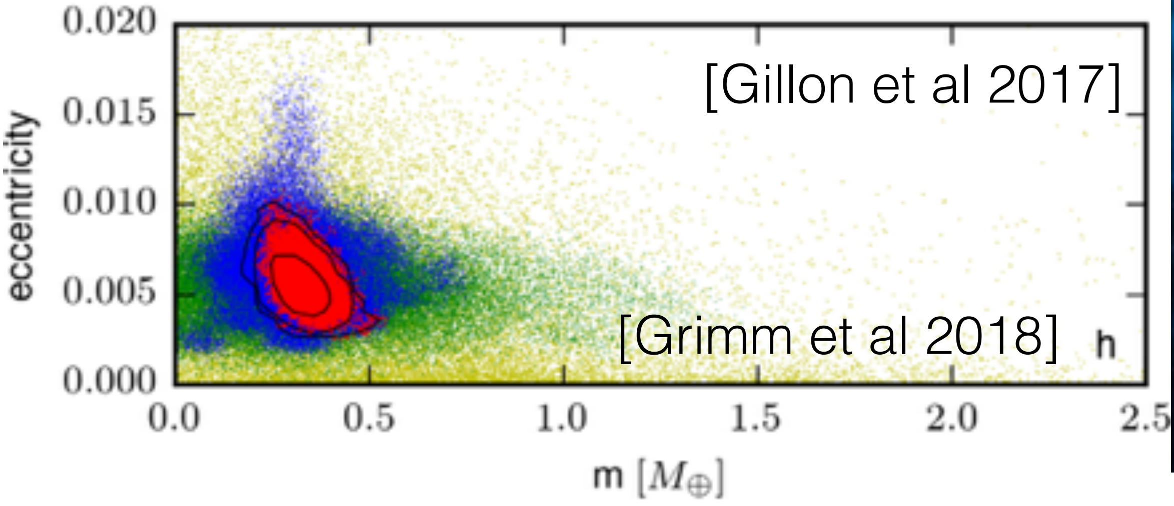
[Gillon et al 2017]

[Grimm et al 2018]

[ESO/
L. Calçada/
M. Kornmesser
2017]

[Abbott et al 2016a,b]

Bayesian inference

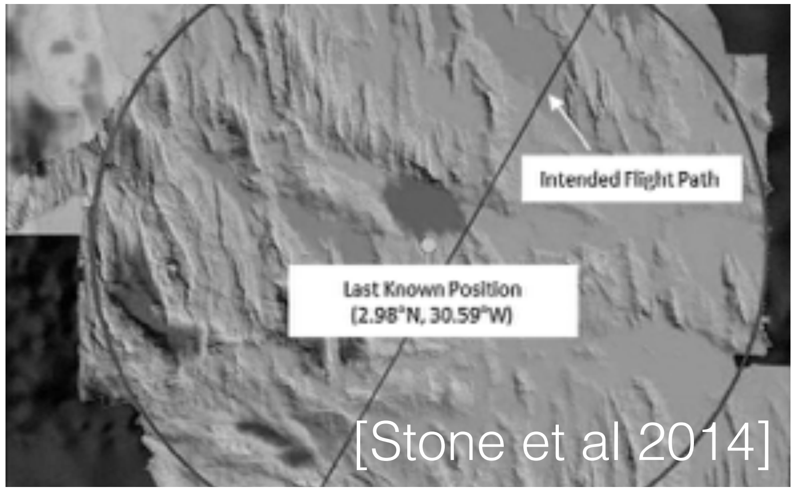


[Gillon et al 2017]

[Grimm et al 2018]

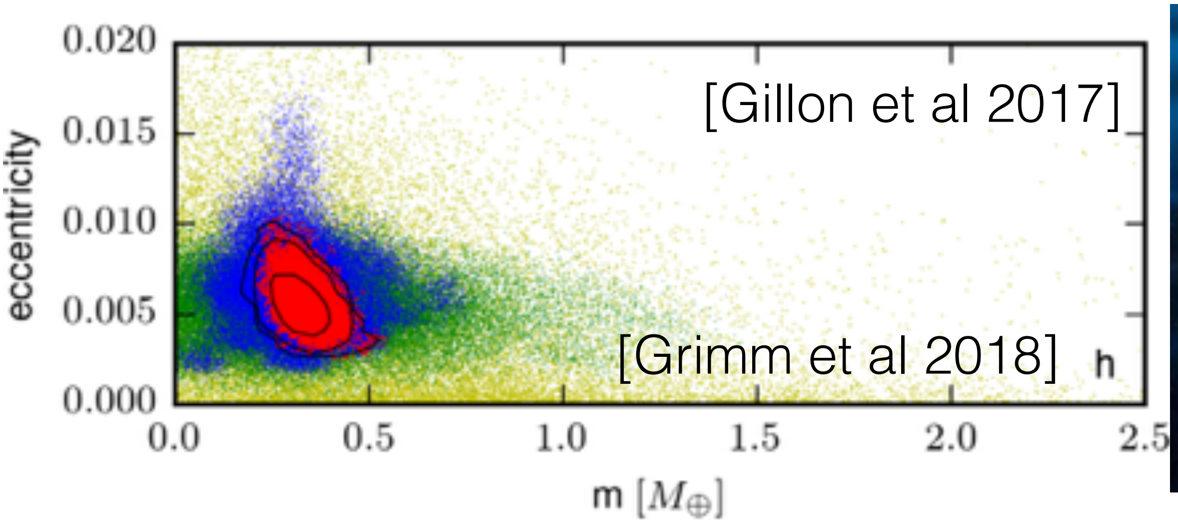
[ESO/
L. Calçada/
M. Kornmesser
2017]

[Abbott et al 2016a,b]



[Stone et al 2014]

Bayesian inference

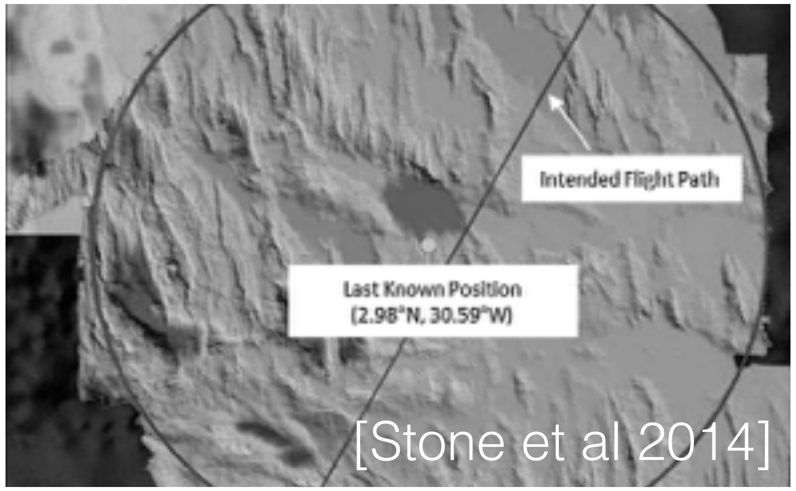


[Gillon et al 2017]

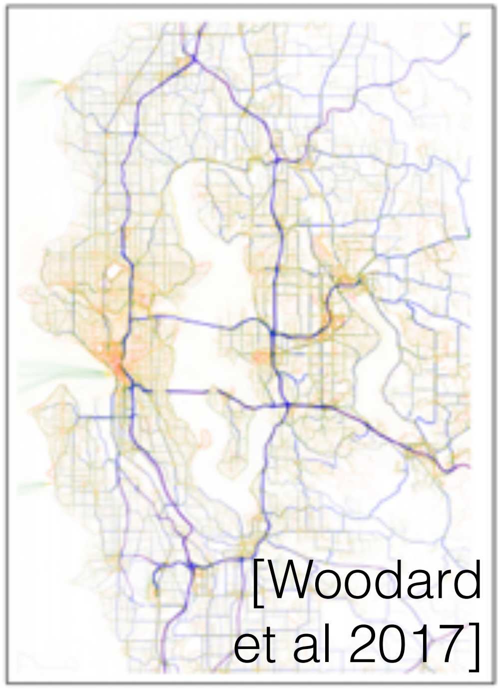
[Grimm et al 2018]

[ESO/
L. Calçada/
M. Kornmesser
2017]

[Abbott et al 2016a,b]

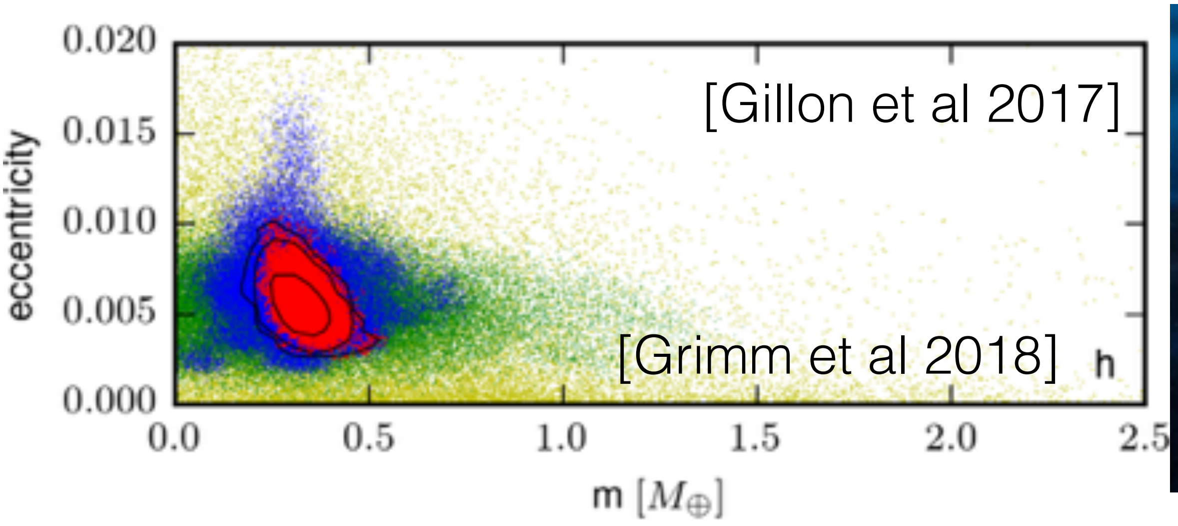


[Stone et al 2014]



[Woodard
et al 2017]

Bayesian inference

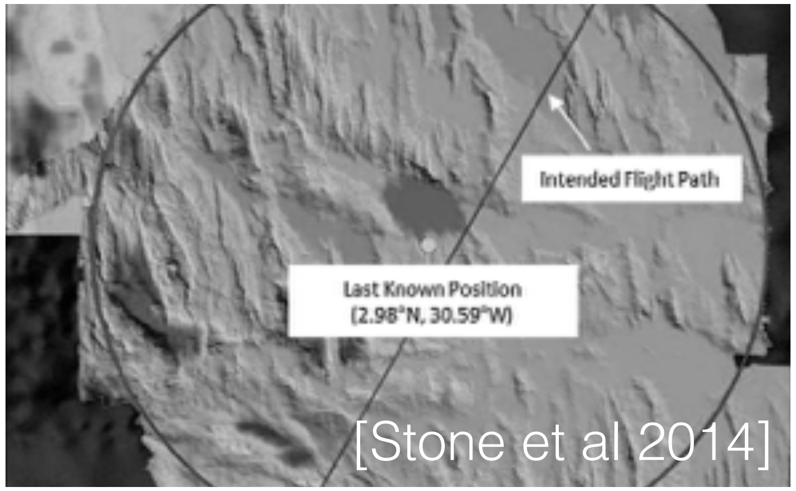


[Gillon et al 2017]

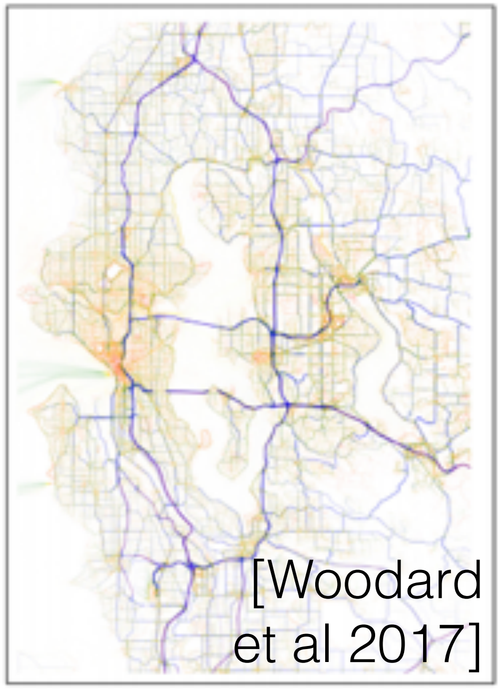
[Grimm et al 2018]

[ESO/
L. Calçada/
M. Kornmesser
2017]

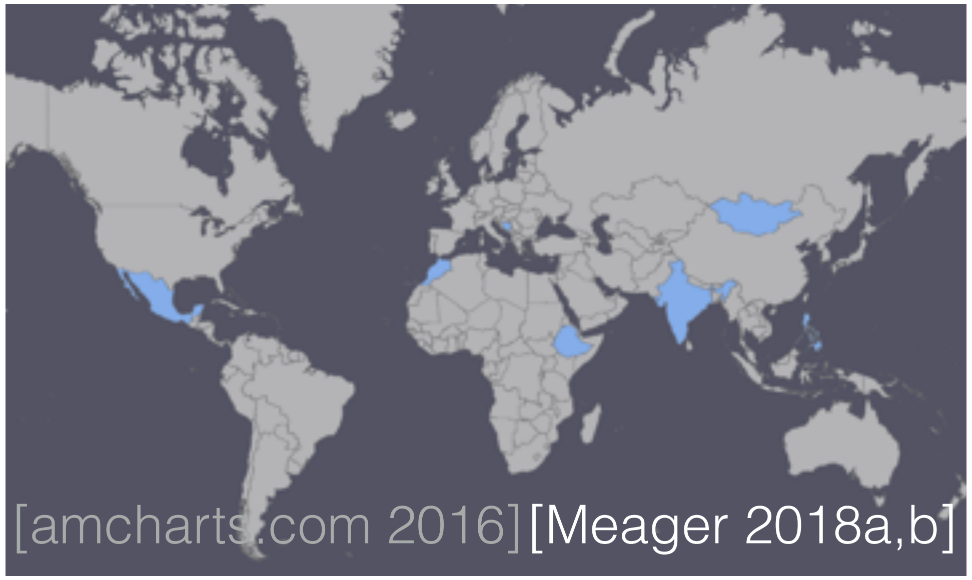
[Abbott et al 2016a,b]



[Stone et al 2014]



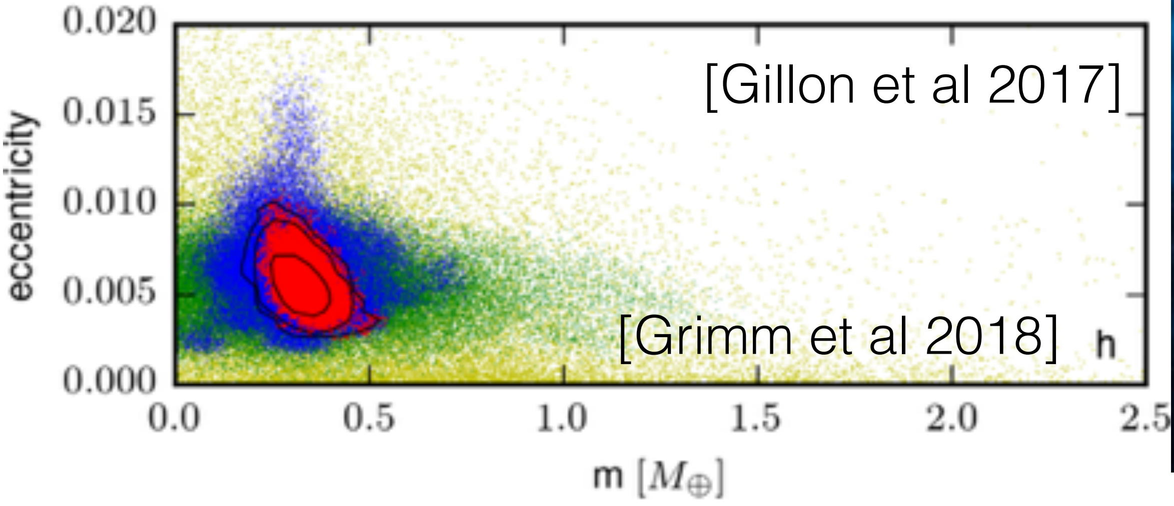
[Woodard
et al 2017]



[amcharts.com 2016]

[Meager 2018a,b]

Bayesian inference

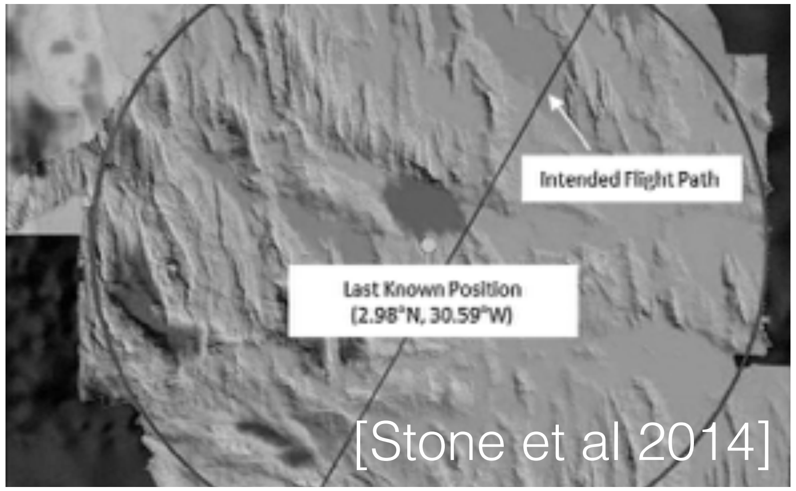


[Gillon et al 2017]

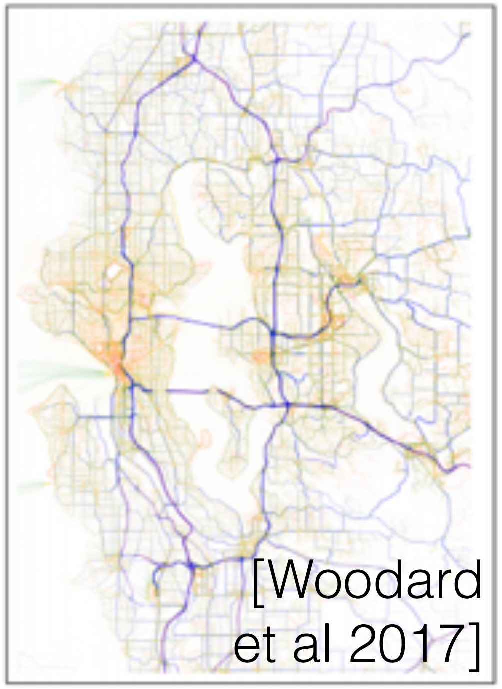
[Grimm et al 2018]

[ESO/
L. Calçada/
M. Kornmesser
2017]

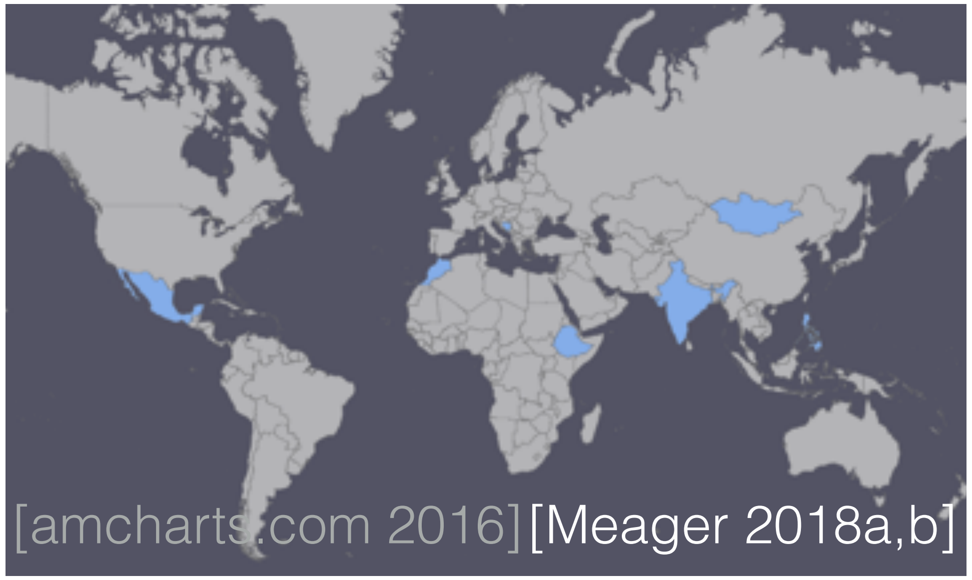
[Abbott et al 2016a,b]



[Stone et al 2014]



[Woodard
et al 2017]



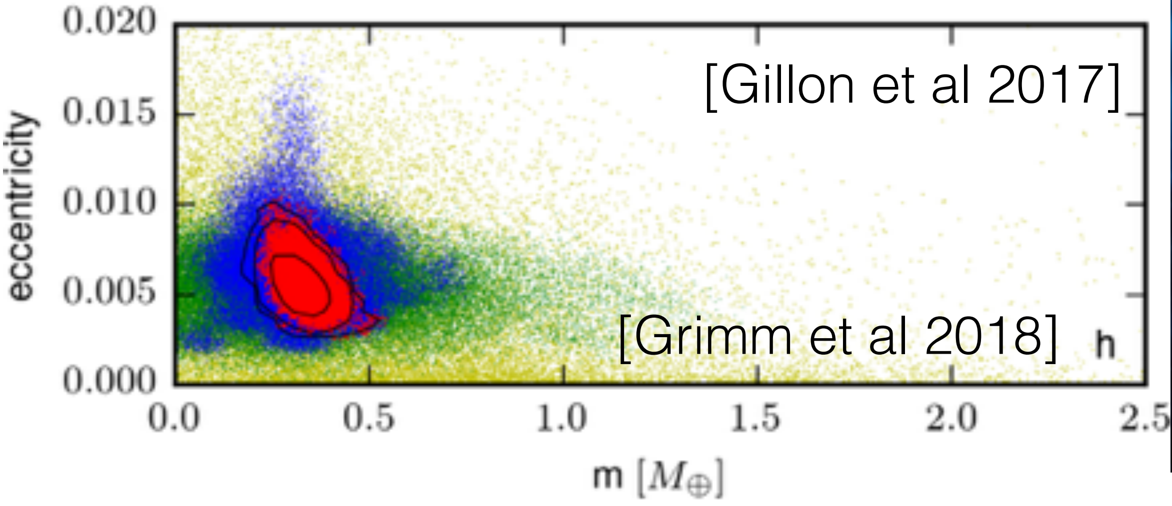
[amcharts.com 2016]

[Meager 2018a,b]

[Julian Hertzog 2016]

[Chati, Balakrishnan
2017]

Bayesian inference

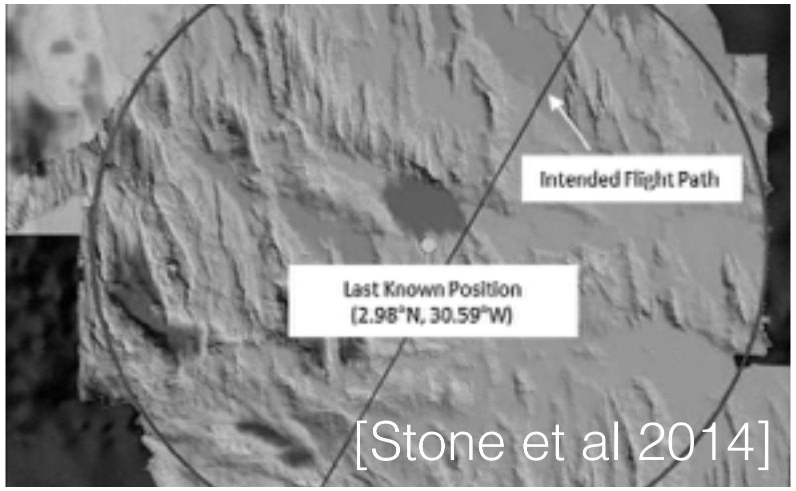


[Gillon et al 2017]

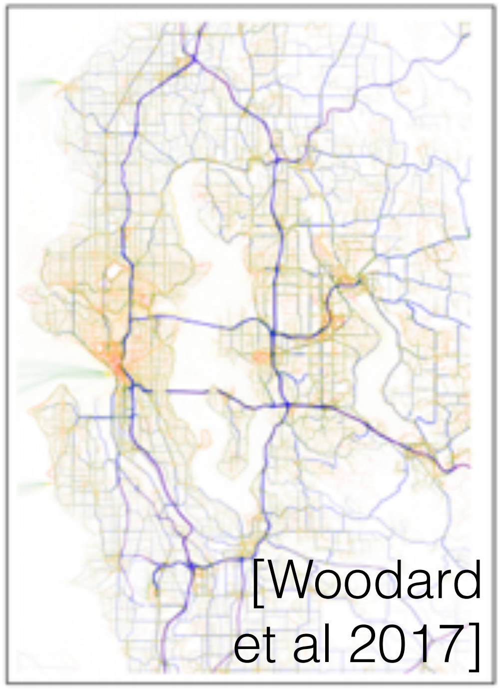
[Grimm et al 2018]

[ESO/
L. Calçada/
M. Kornmesser
2017]

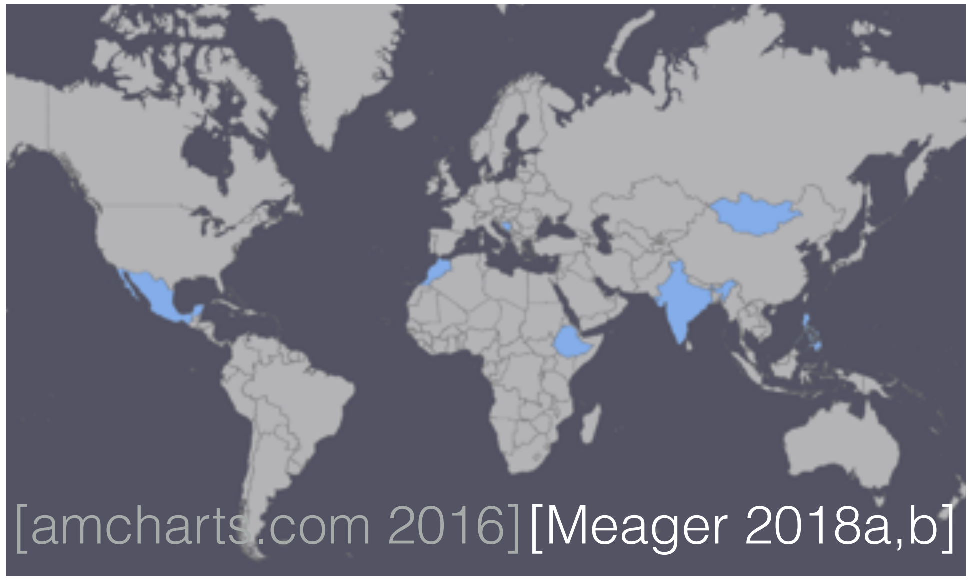
[Abbott et al 2016a,b]



[Stone et al 2014]



[Woodard
et al 2017]



[amcharts.com 2016]

[Meager 2018a,b]

[Kuikka et al 2014]

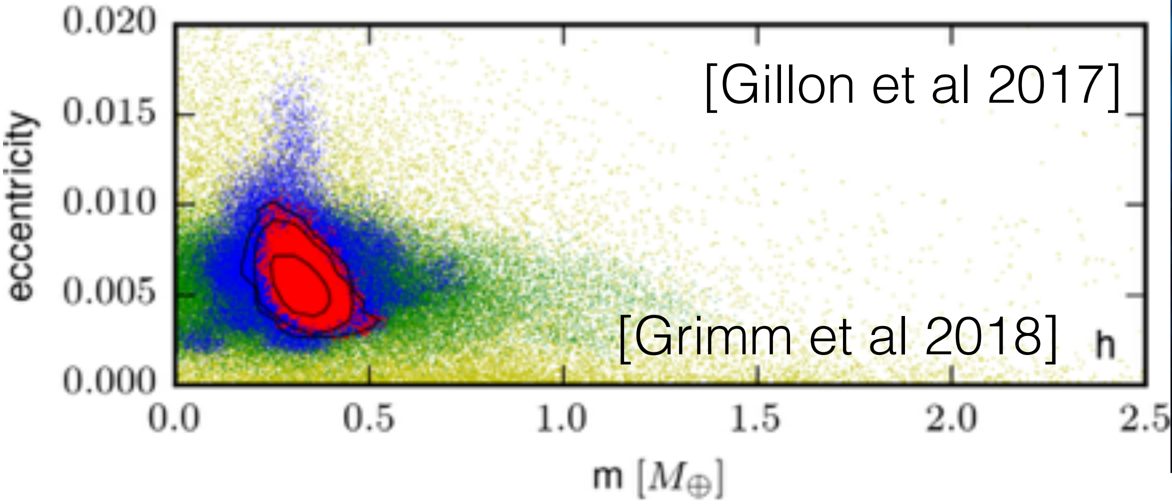
[Baltic Salmon Fund]

[Chati, Balakrishnan
2017]

[Julian Hertzog 2016]

Bayesian inference

- Analysis goals: Point estimates, coherent uncertainties
    - Interpretable, complex, modular; expert information

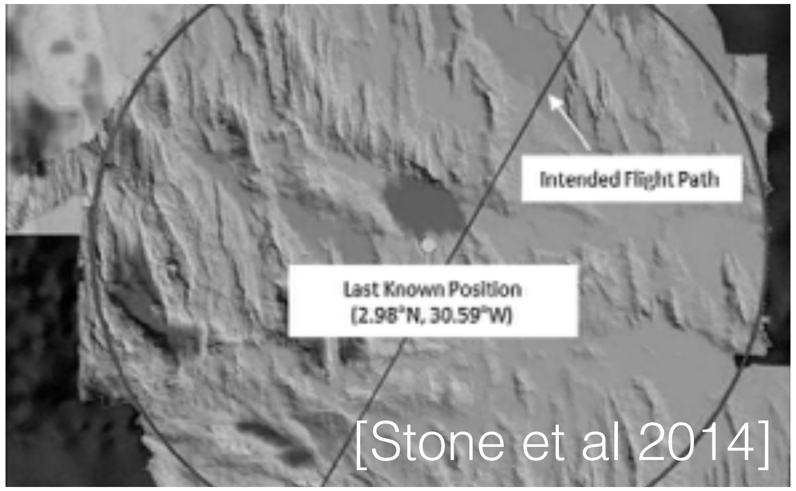


[Gillon et al 2017]

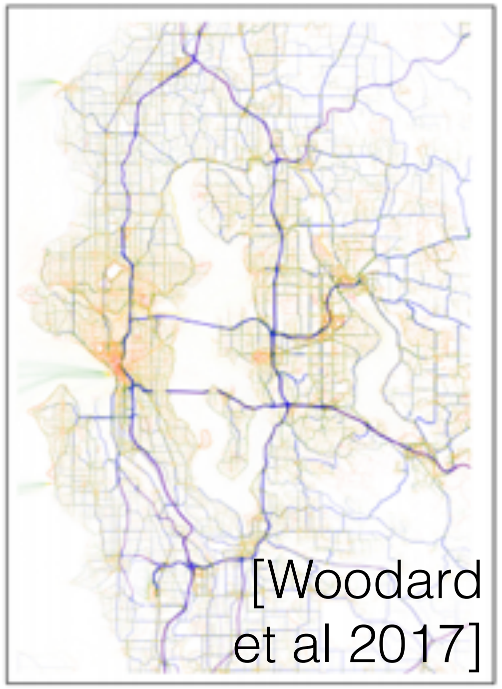
[Grimm et al 2018]

[ESO/
L. Calçada/
M. Kornmesser
2017]

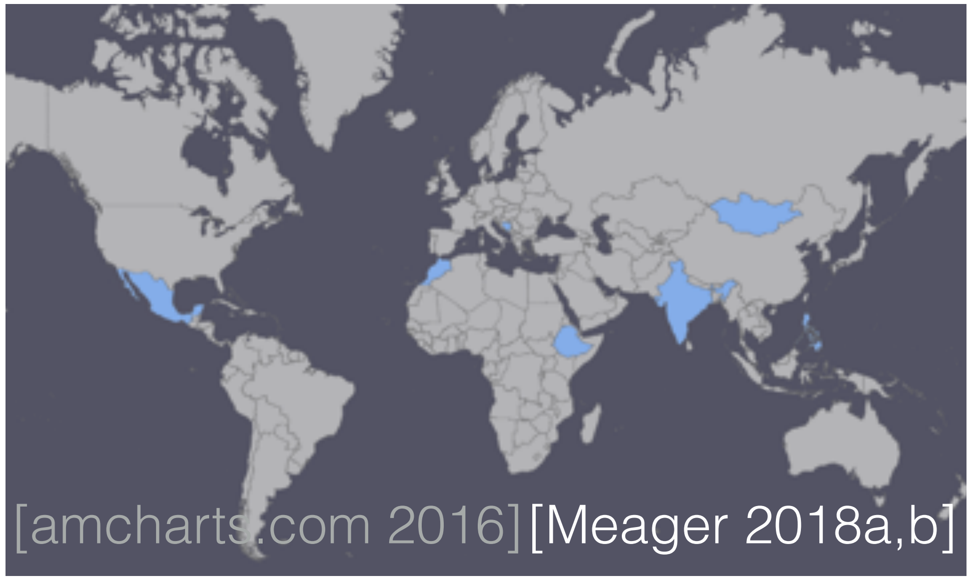
[Abbott et al 2016a,b]



[Stone et al 2014]



[Woodard
et al 2017]



[amcharts.com 2016]

[Meager 2018a,b]

[Kuikka et al 2014]

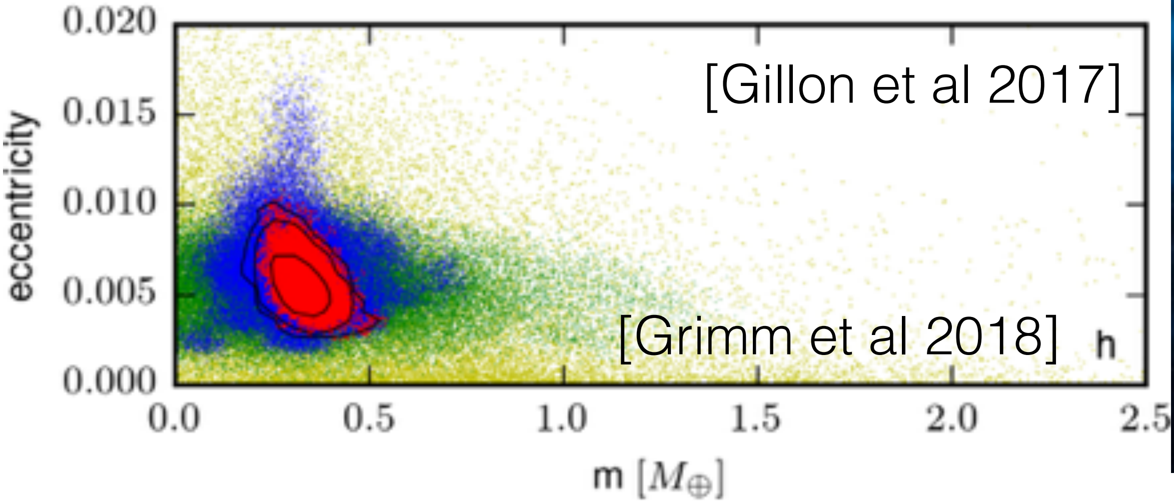
[Baltic Salmon Fund]

[Chati, Balakrishnan
2017]

[Julian Hertzog 2016]

Bayesian inference

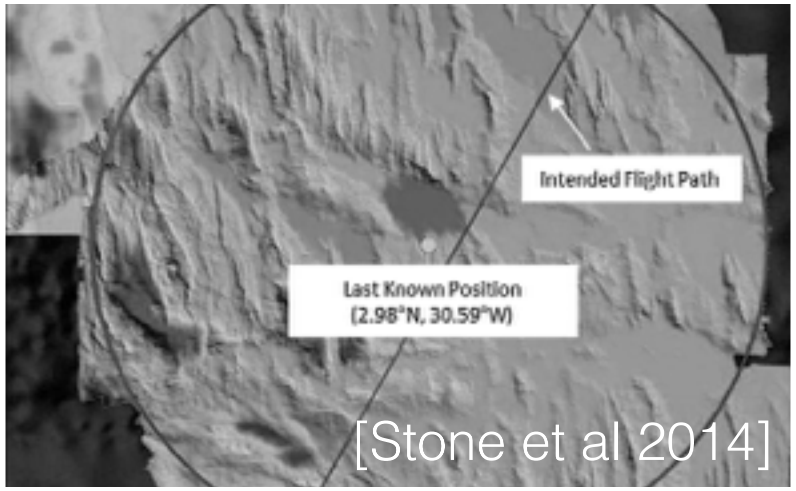
- Analysis goals: Point estimates, coherent uncertainties
  - Interpretable, complex, modular; expert information



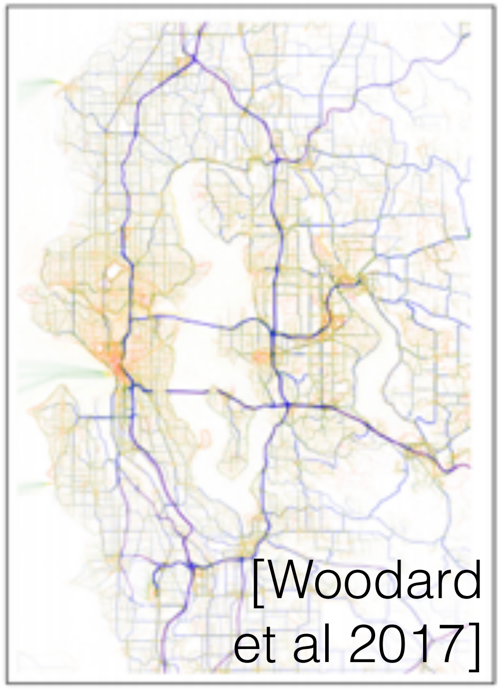
[Gillon et al 2017]

[Grimm et al 2018]

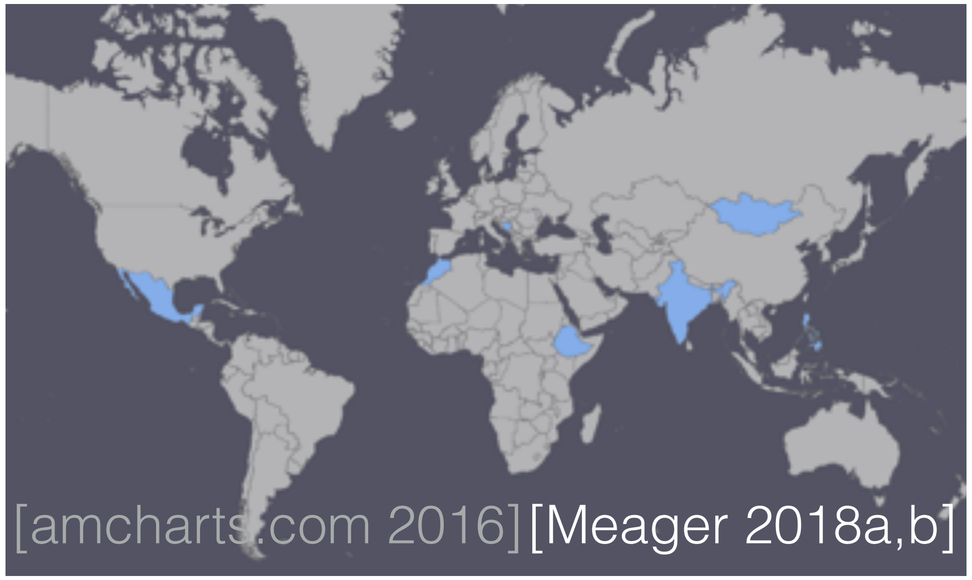
[ESO/ L. Calçada/ M. Kornmesser 2017] [Abbott et al 2016a,b]



[Stone et al 2014]



[Woodard et al 2017]



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

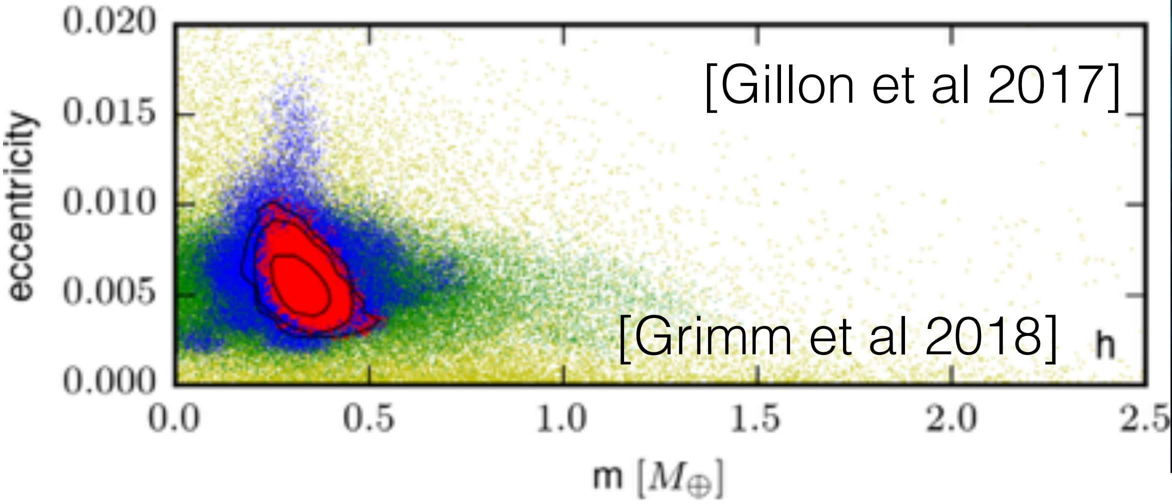
[Kuikka et al 2014]
[Baltic Salmon Fund]

[Chati, Balakrishnan 2017]
[Julian Hertzog 2016]

[mc-stan.org]

Bayesian inference

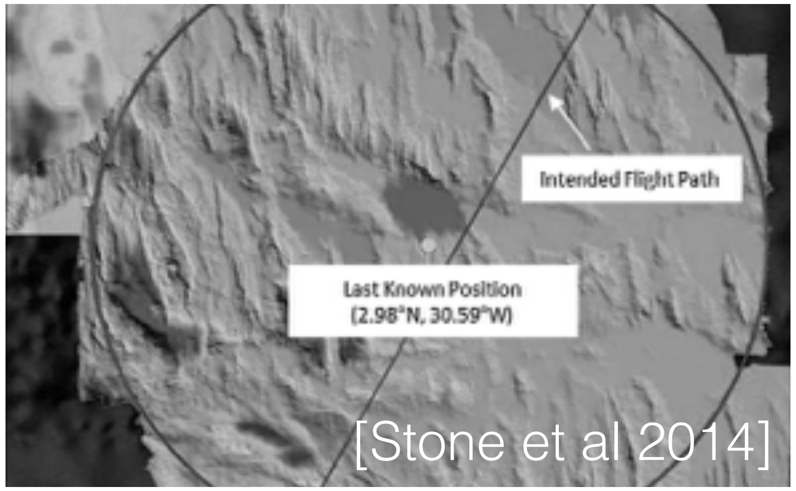
- Analysis goals: Point estimates, coherent uncertainties
  - Interpretable, complex, modular; expert information



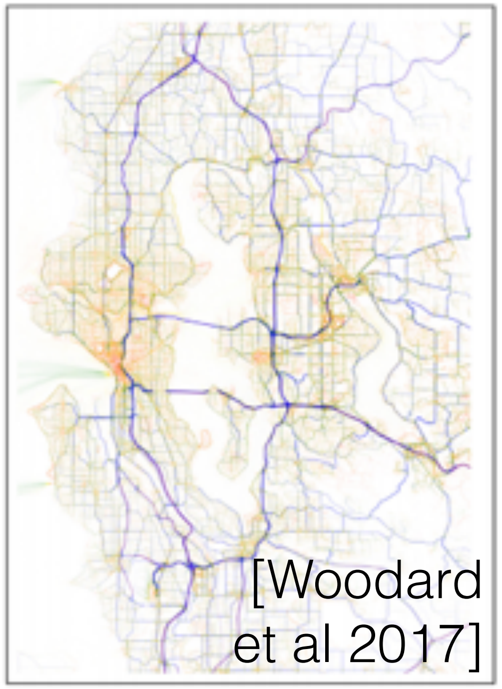
[Gillon et al 2017]

[Grimm et al 2018]

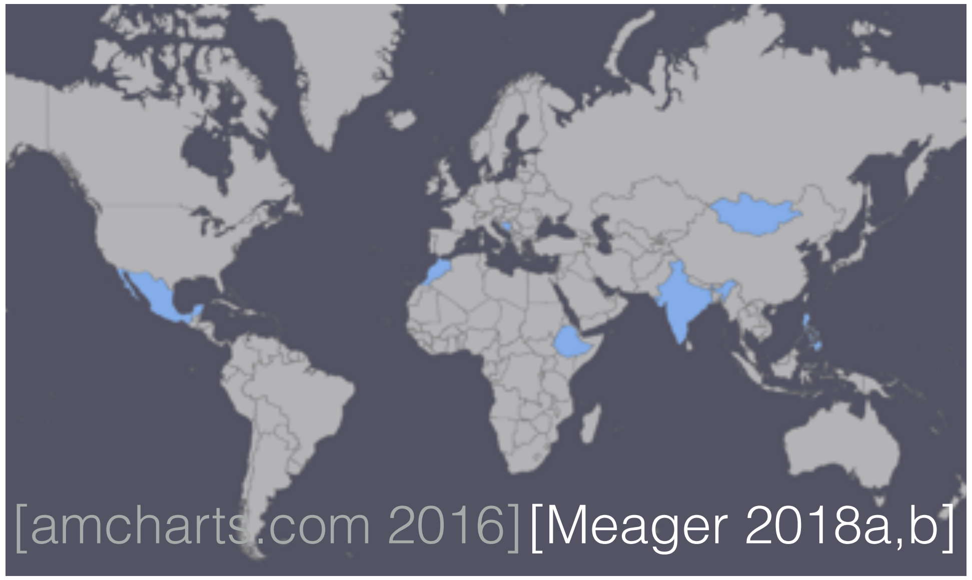
[ESO/ L. Calçada/ M. Kornmesser 2017] [Abbott et al 2016a,b]



[Stone et al 2014]



[Woodard et al 2017]



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

[Kuikka et al 2014] [Baltic Salmon Fund]

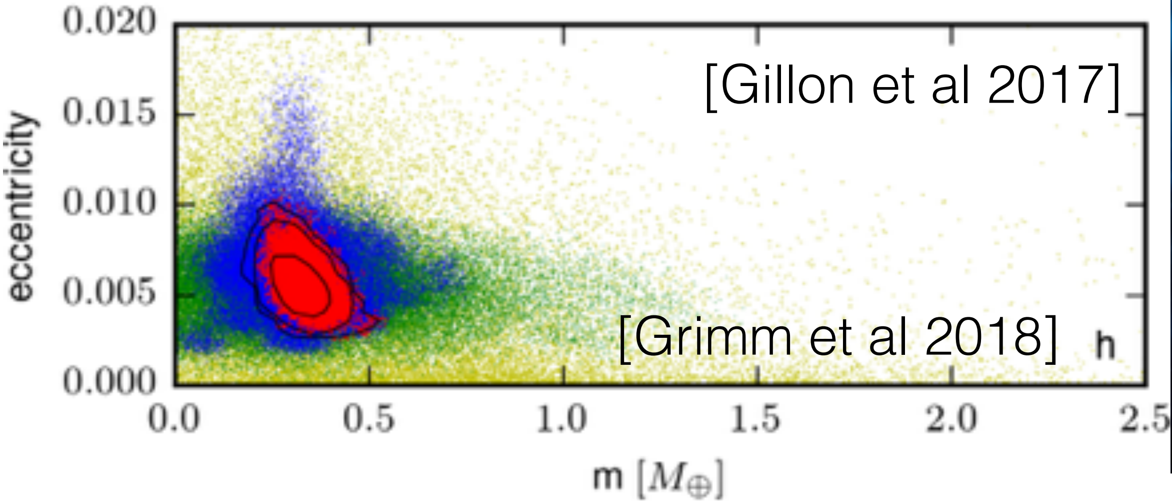
[Chati, Balakrishnan 2017] [Julian Hertzog 2016]

[mc-stan.org]

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

Bayesian inference

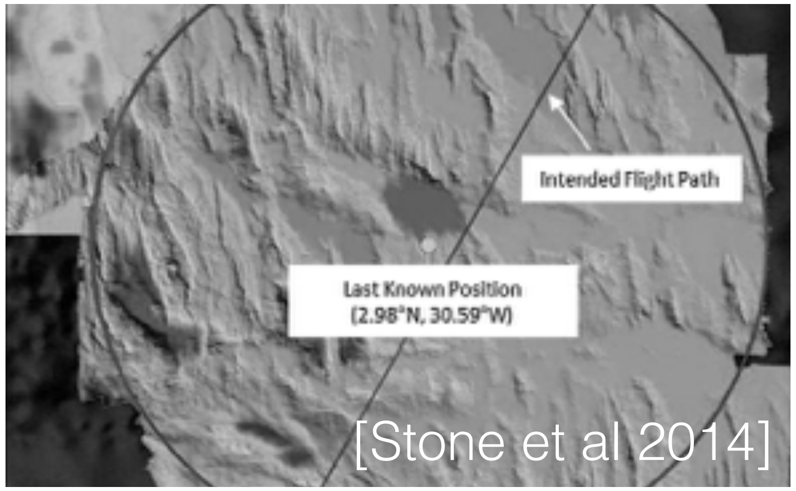
- Analysis goals: Point estimates, coherent uncertainties
  - Interpretable, complex, modular; expert information



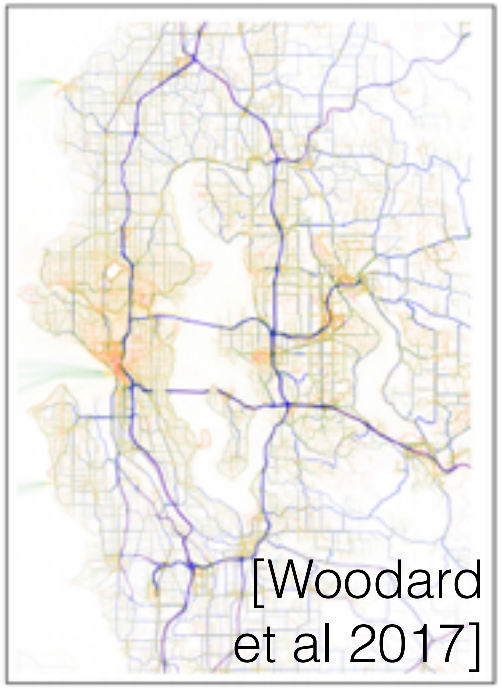
[Gillon et al 2017]

[Grimm et al 2018]

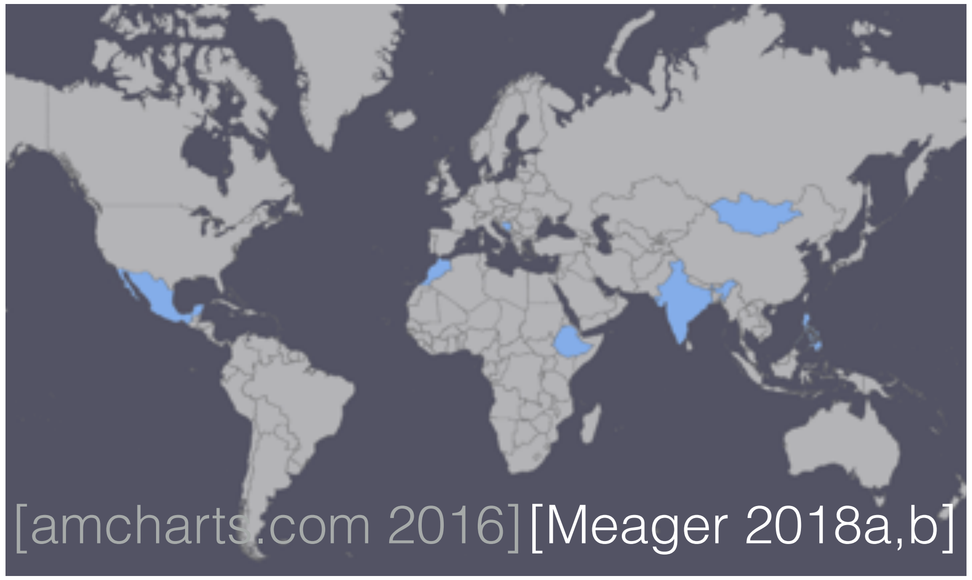
[ESO/ L. Calçada/ M. Kornmesser 2017] [Abbott et al 2016a,b]



[Stone et al 2014]



[Woodard et al 2017]



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

[Kuikka et al 2014] [Baltic Salmon Fund]

[Chati, Balakrishnan 2017] [Julian Hertzog 2016]

[mc-stan.org]

- Challenge: fast (compute, user), reliable inference
- Uncertainty doesn't have to disappear in large data sets

Variational Bayes

Variational Bayes

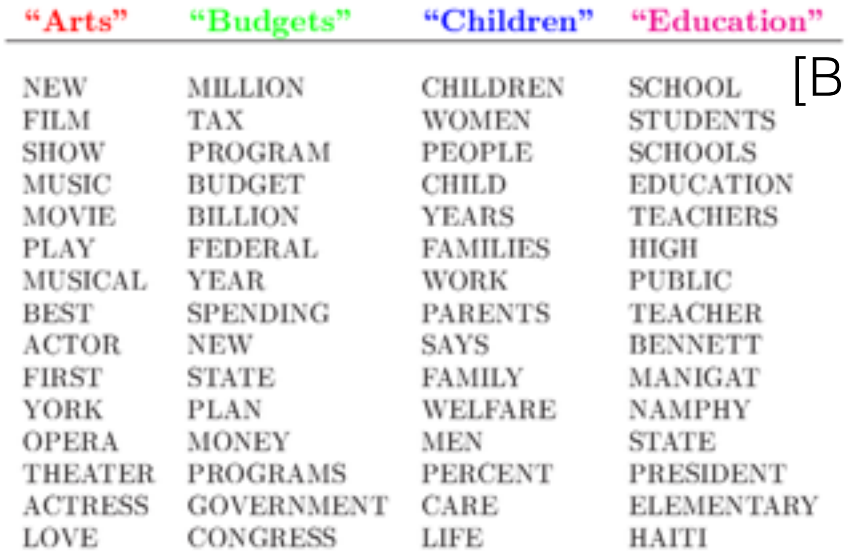
- Modern problems: often large data, large dimensions

Variational Bayes

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

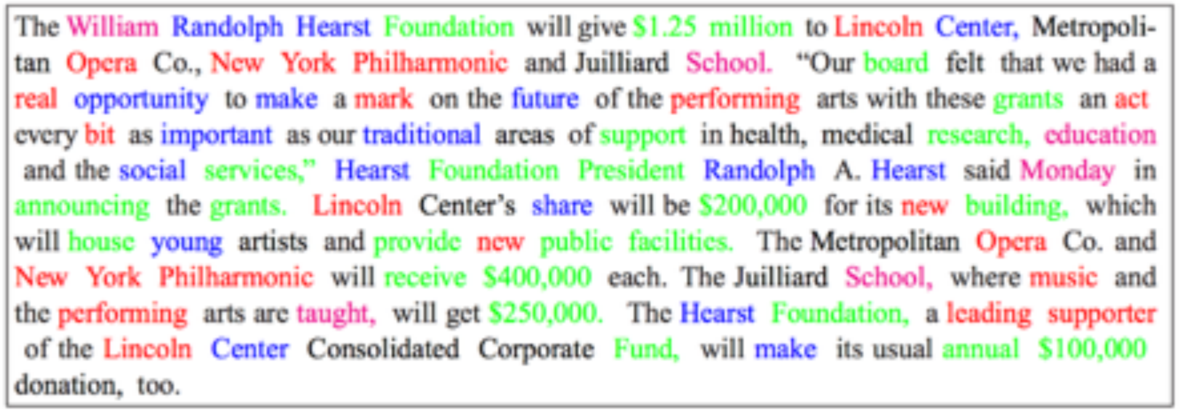
Variational Bayes

- 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

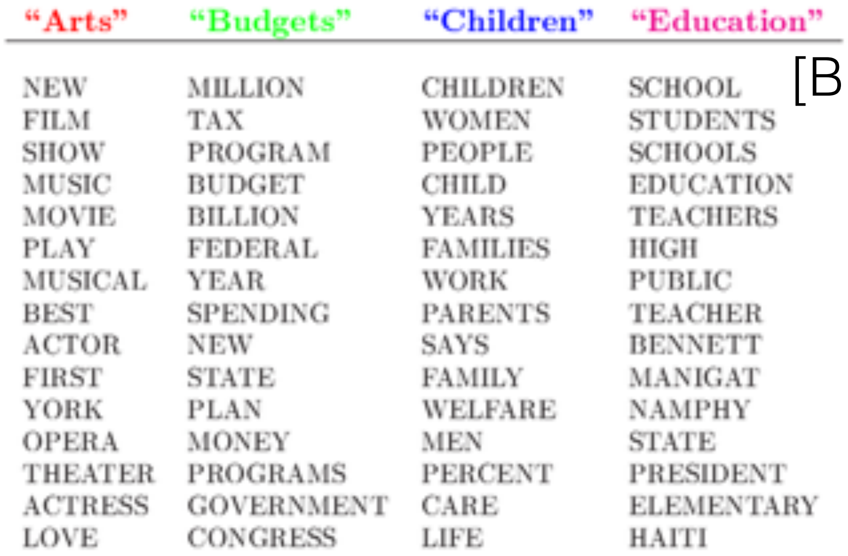
[Blei et al 2003]



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.

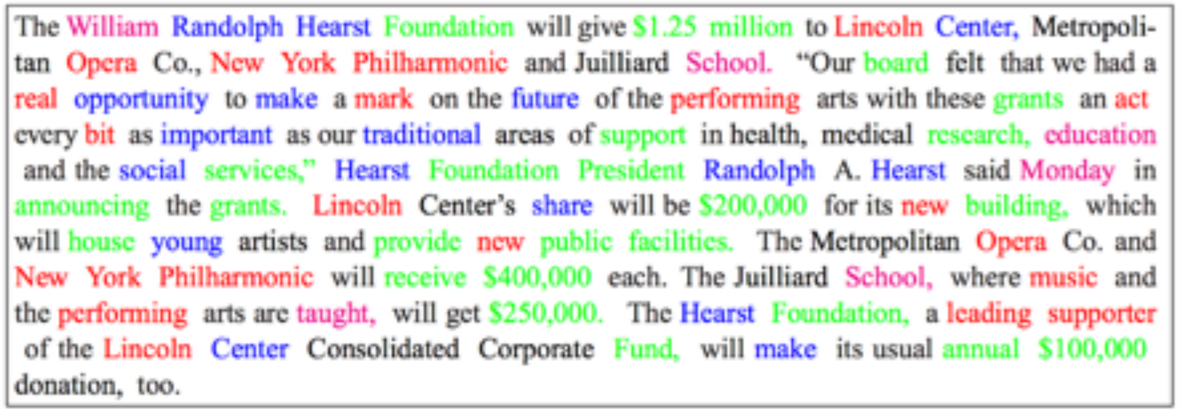
Variational Bayes

- 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

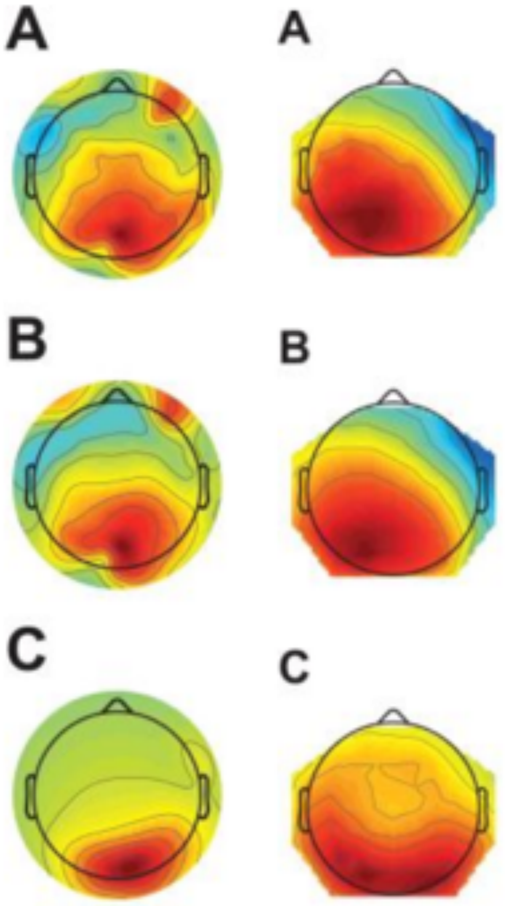
- Modern problems: often large data, large dimensions
- Variational Bayes can be very fast

"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]

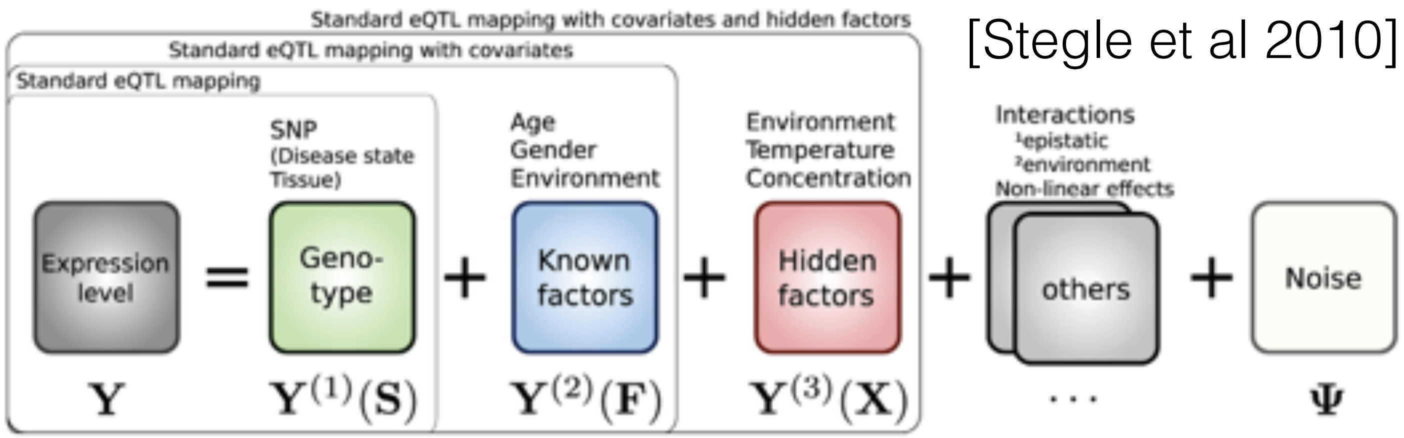
Variational Bayes

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

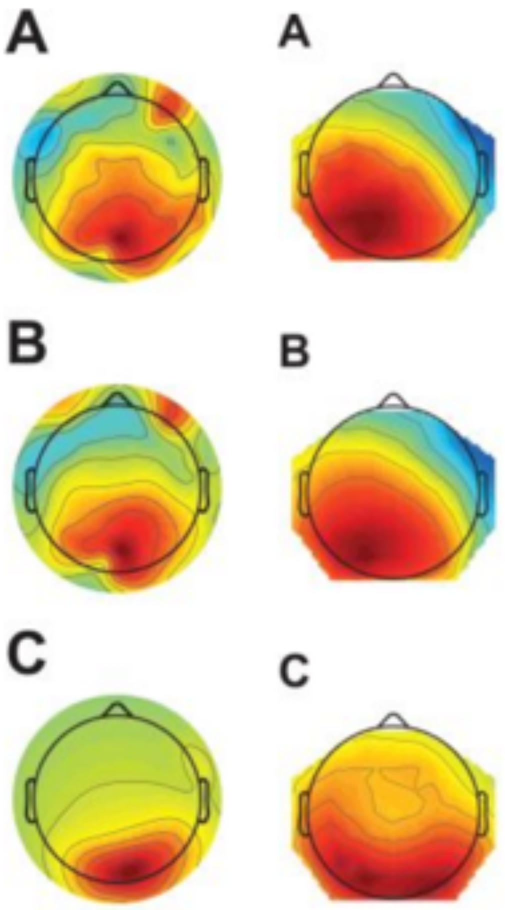
"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY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

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]



[Stegle et al 2010]



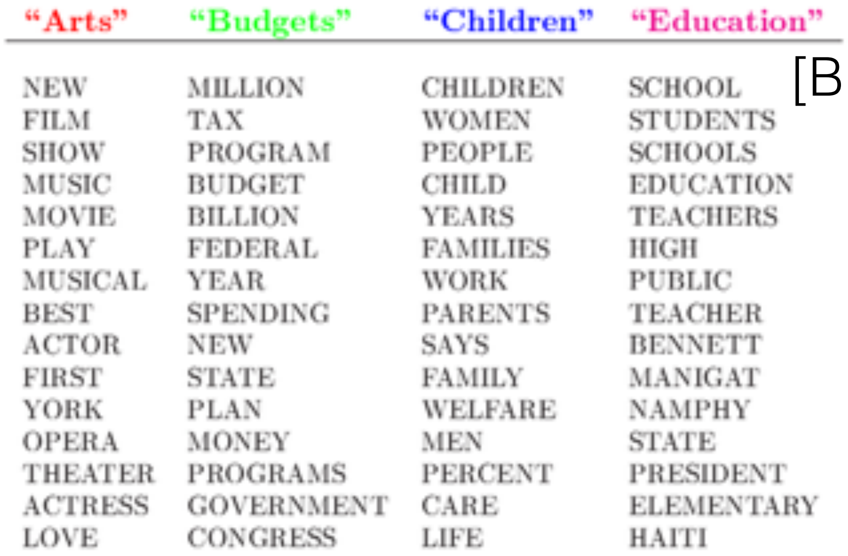
[Airoldi et al 2008]

[Gershman et al 2014]

[Blei et al 2018]

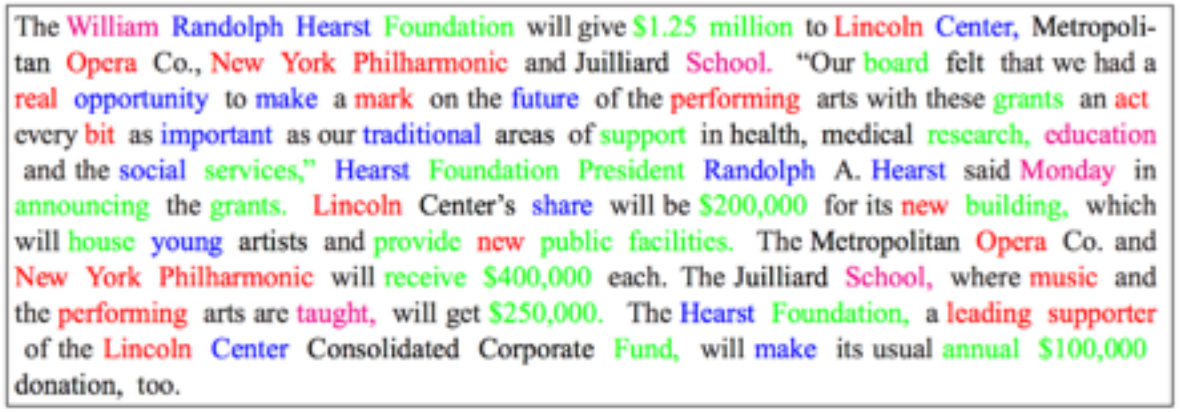
Variational Bayes

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



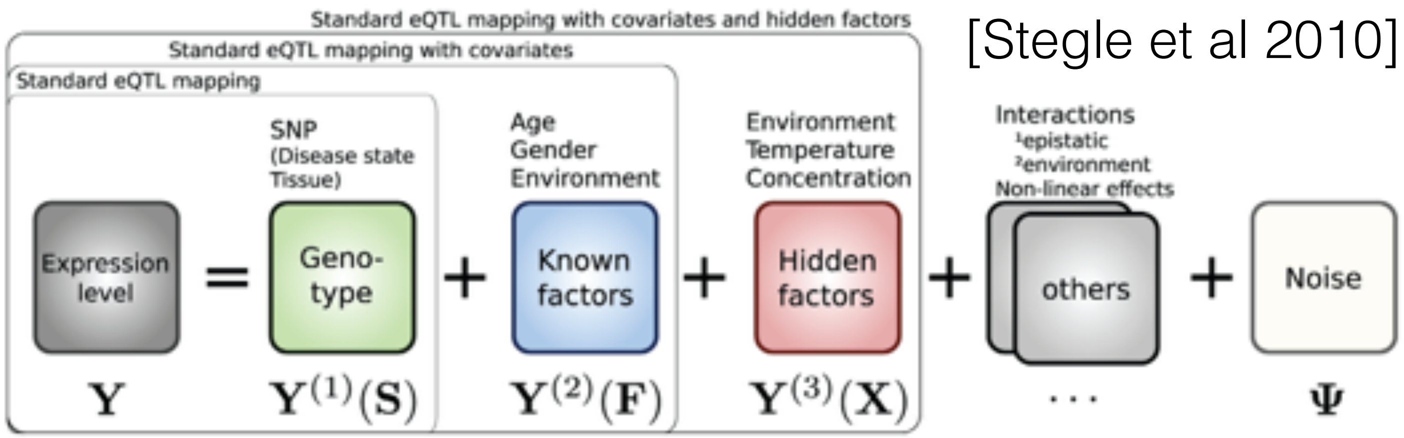
"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	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]

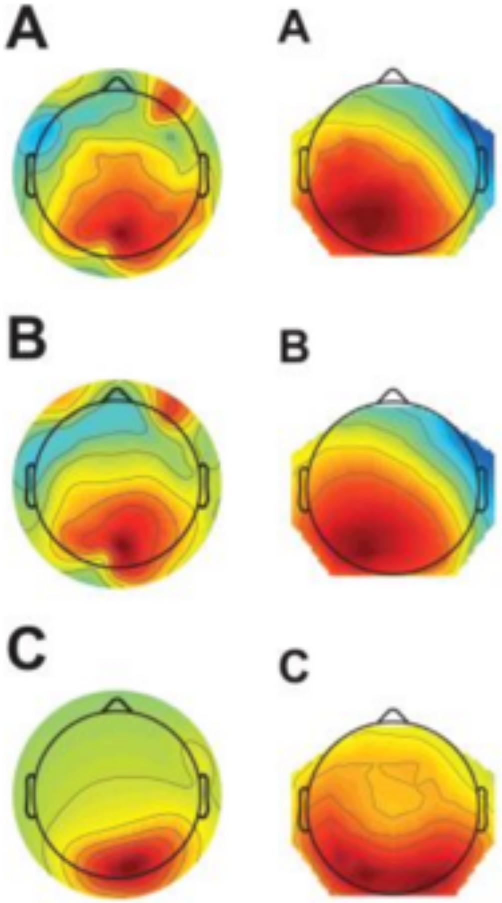


The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our 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.

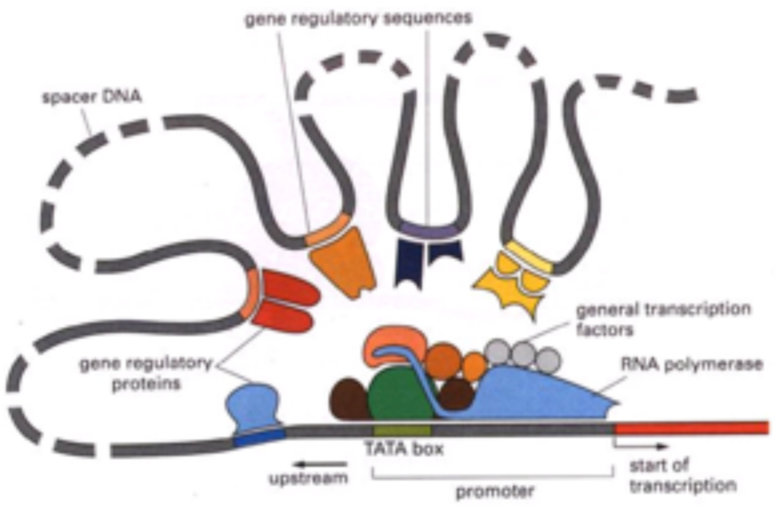
[Airoldi et al 2008]



[Stegle et al 2010]



[Gershman et al 2014]



[Xing et al 2004]
[Xing 2003]

[Blei et al 2018]

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

$$\theta$$

Bayesian inference

parameters

$$p(\theta)$$

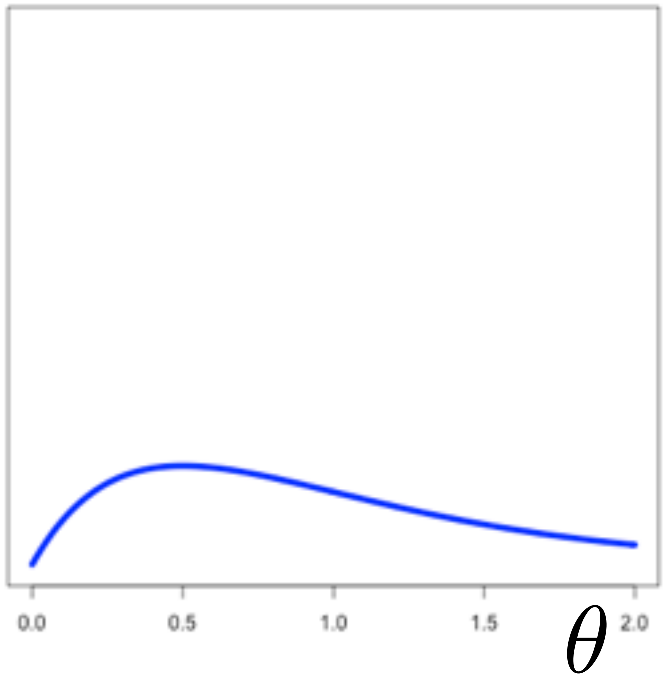
prior

Bayesian inference

parameters

$p(\theta)$

prior

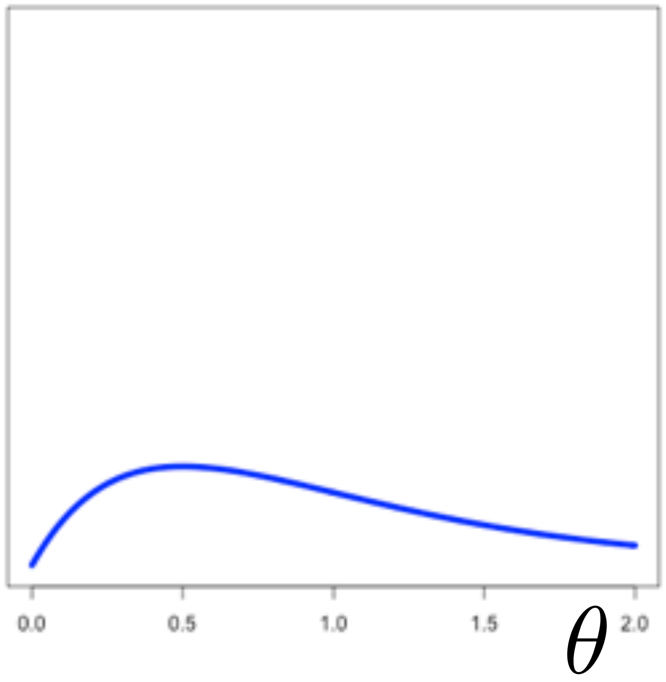


Bayesian inference

parameters

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

likelihood prior

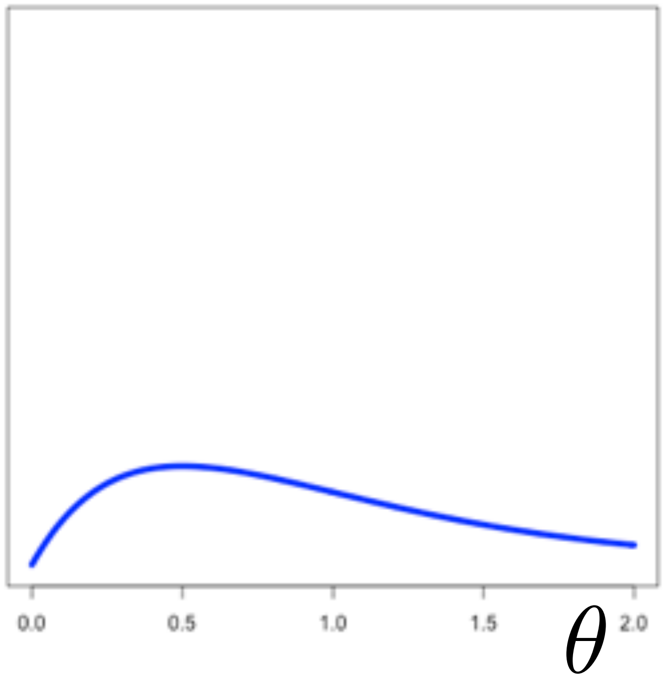


Bayesian inference

data parameters

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

likelihood prior

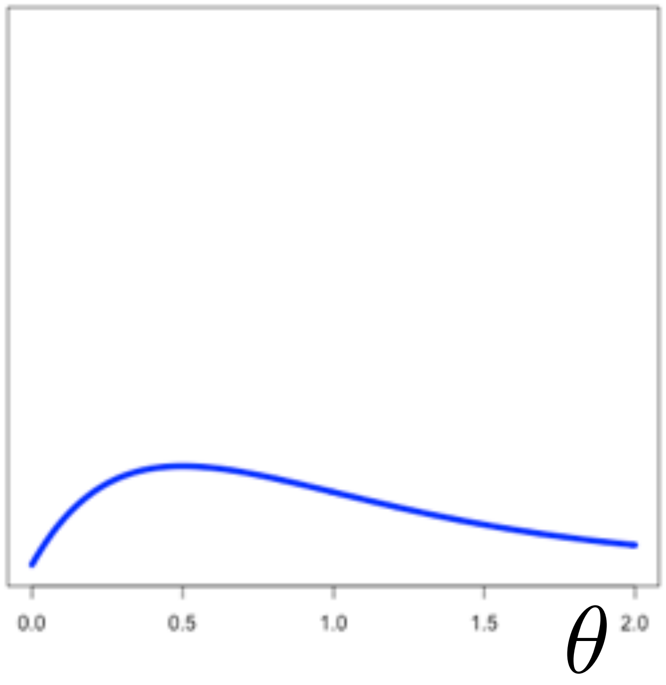


Bayesian inference

data parameters

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

posterior likelihood prior



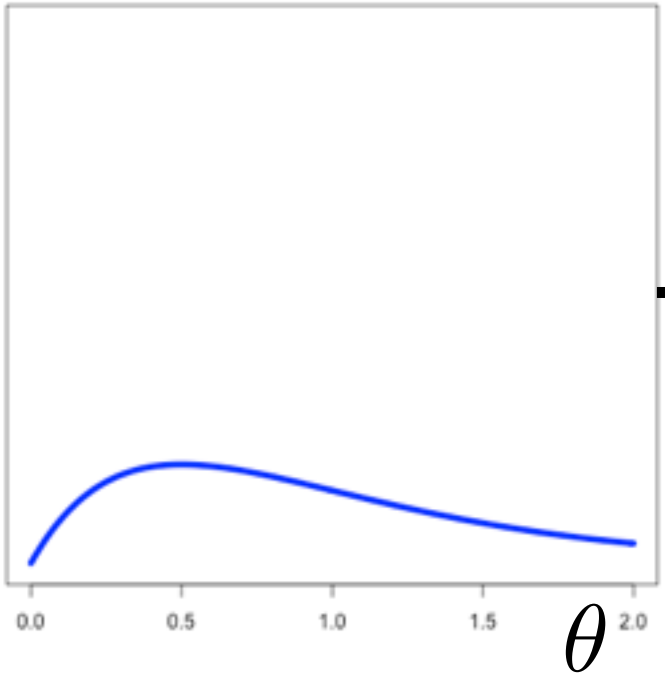
Bayesian inference

data

parameters

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

posterior likelihood prior



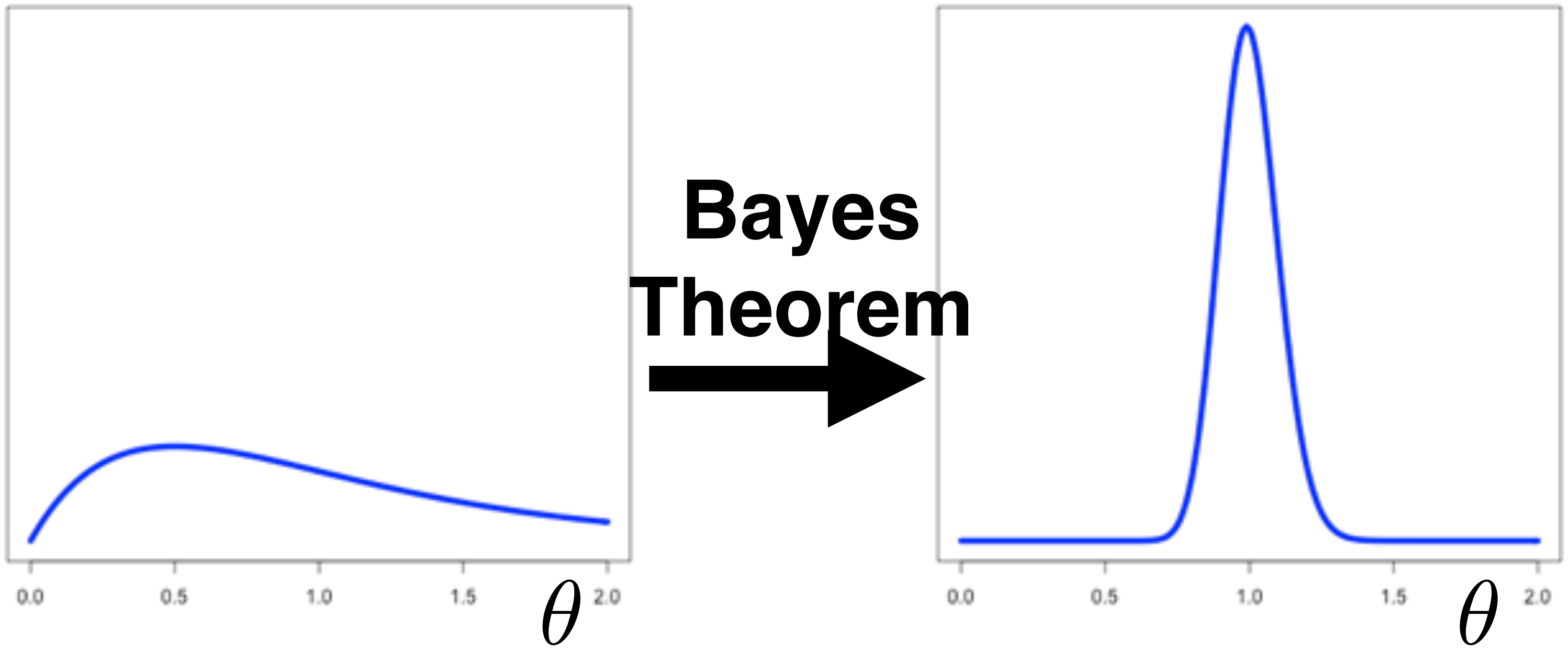
Bayes Theorem

Bayesian inference

data parameters

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

posterior likelihood prior

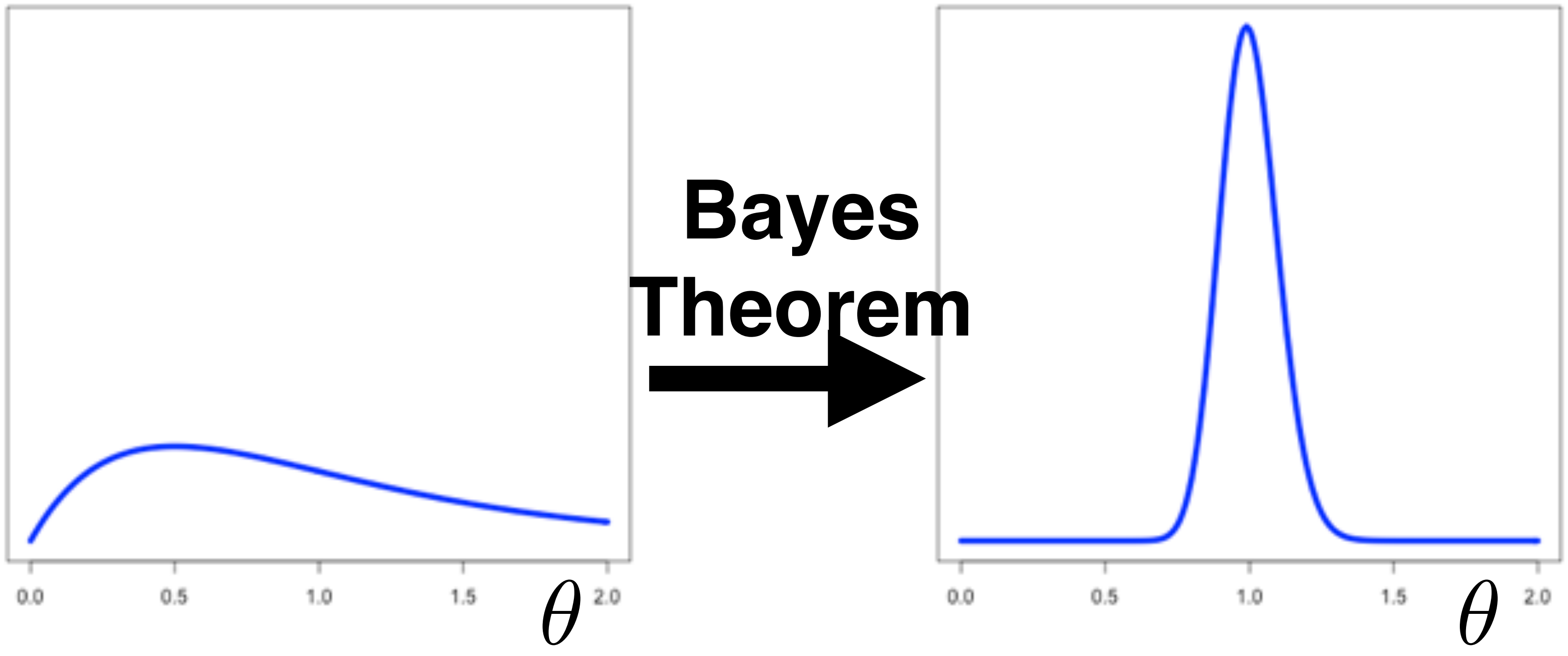


Bayesian inference

data parameters

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

posterior likelihood prior



Bayes Theorem

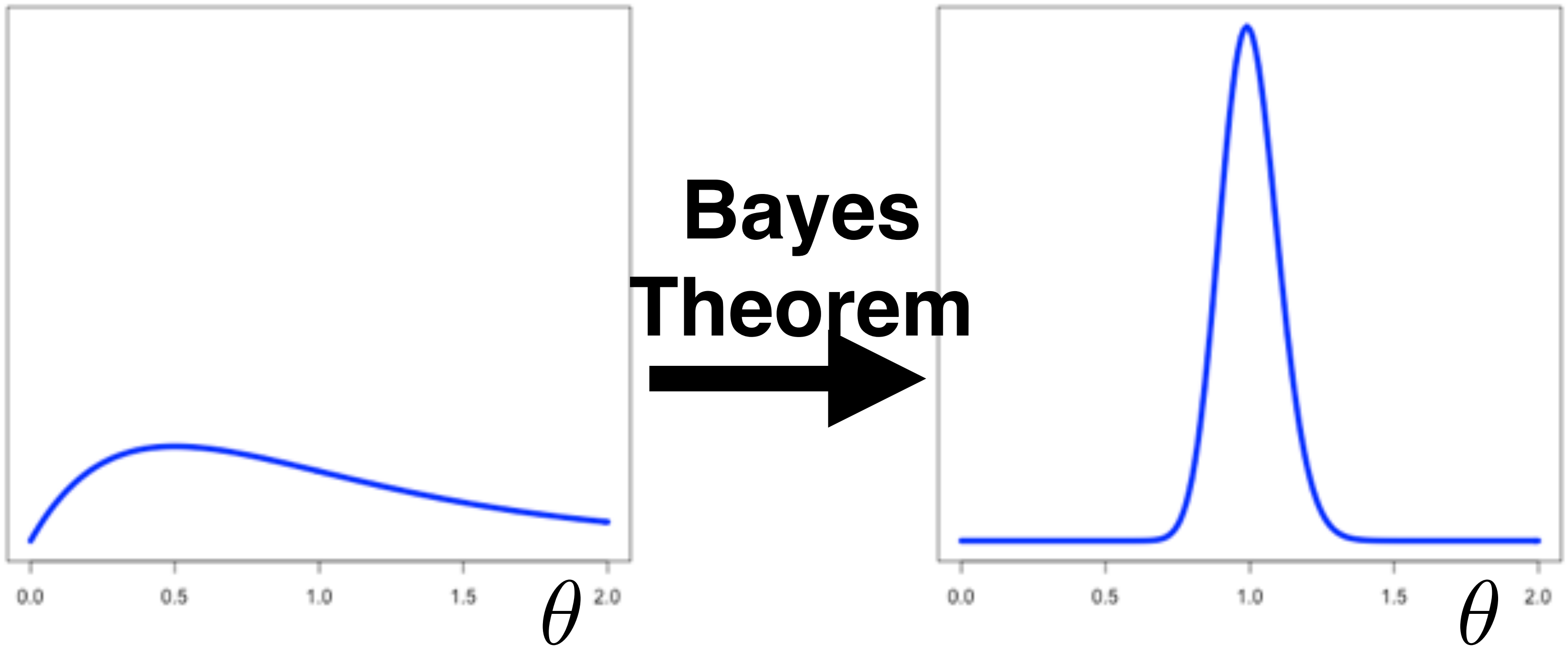
1. Build a model: choose prior & choose likelihood

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

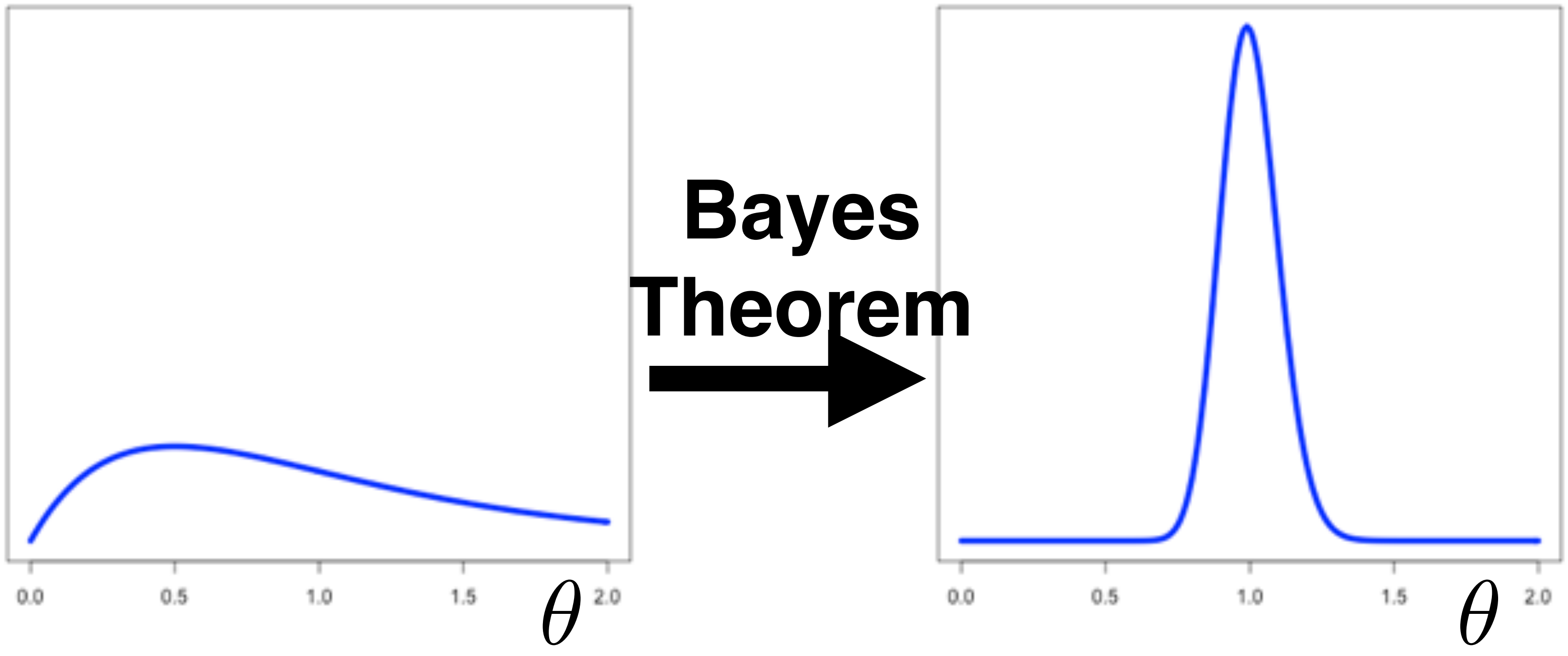
1. Build a model: choose prior & choose likelihood
2. Compute the posterior

Bayesian inference

data parameters

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

posterior likelihood prior



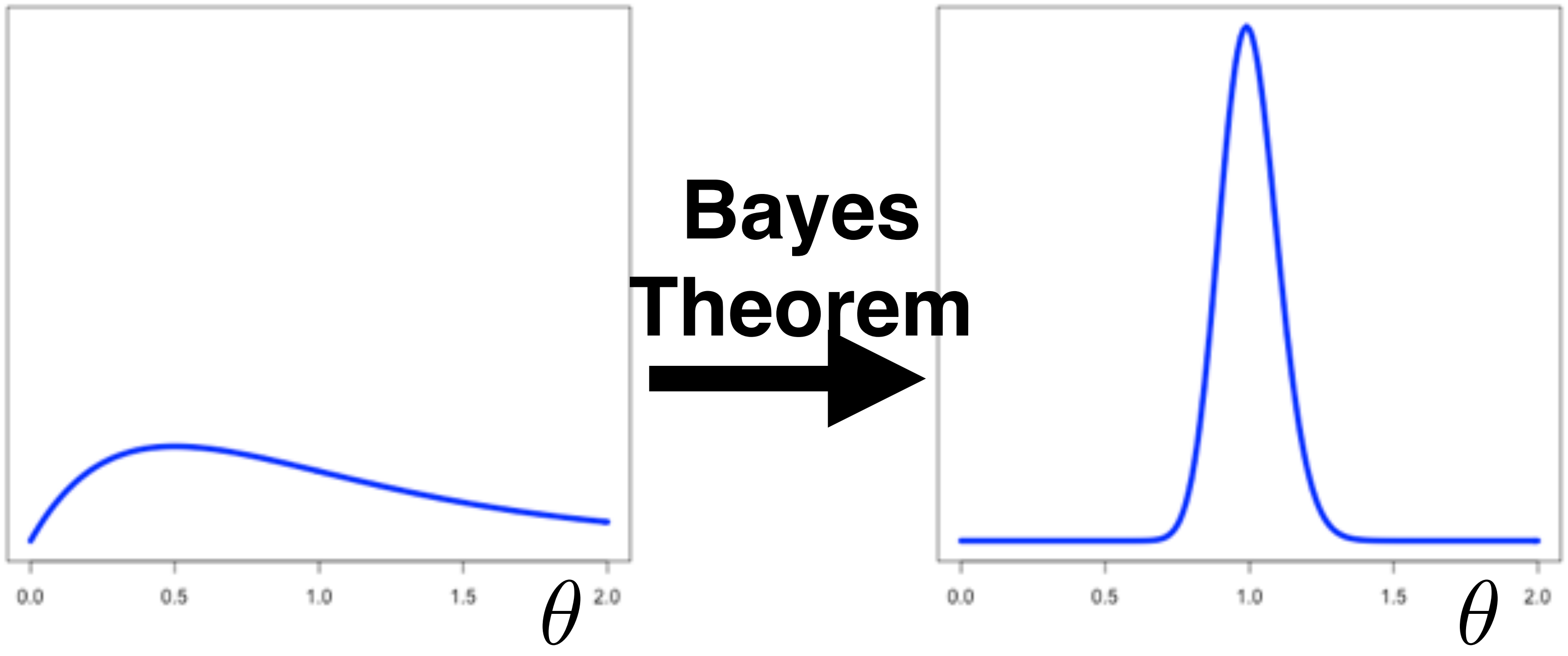
1. Build a model: choose prior & choose likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances

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

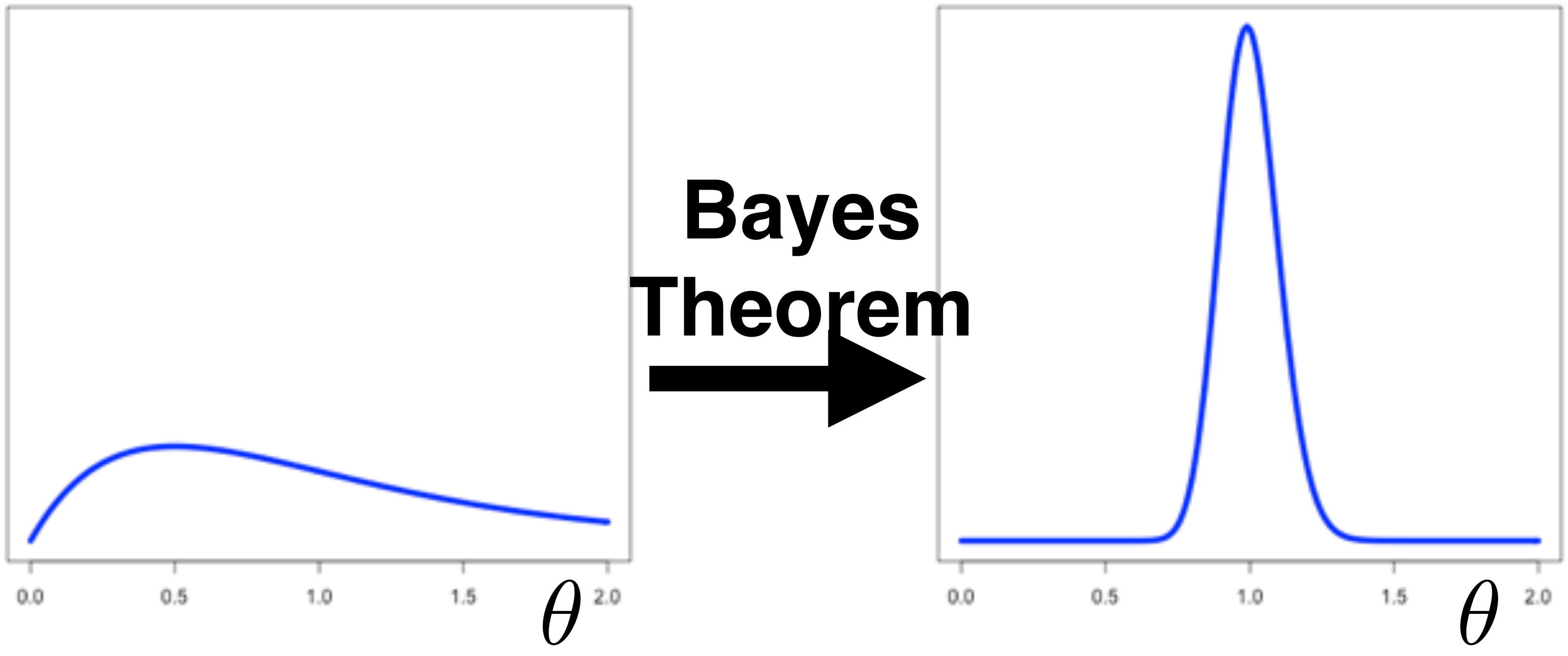
- Why are steps 2 and 3 hard?

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

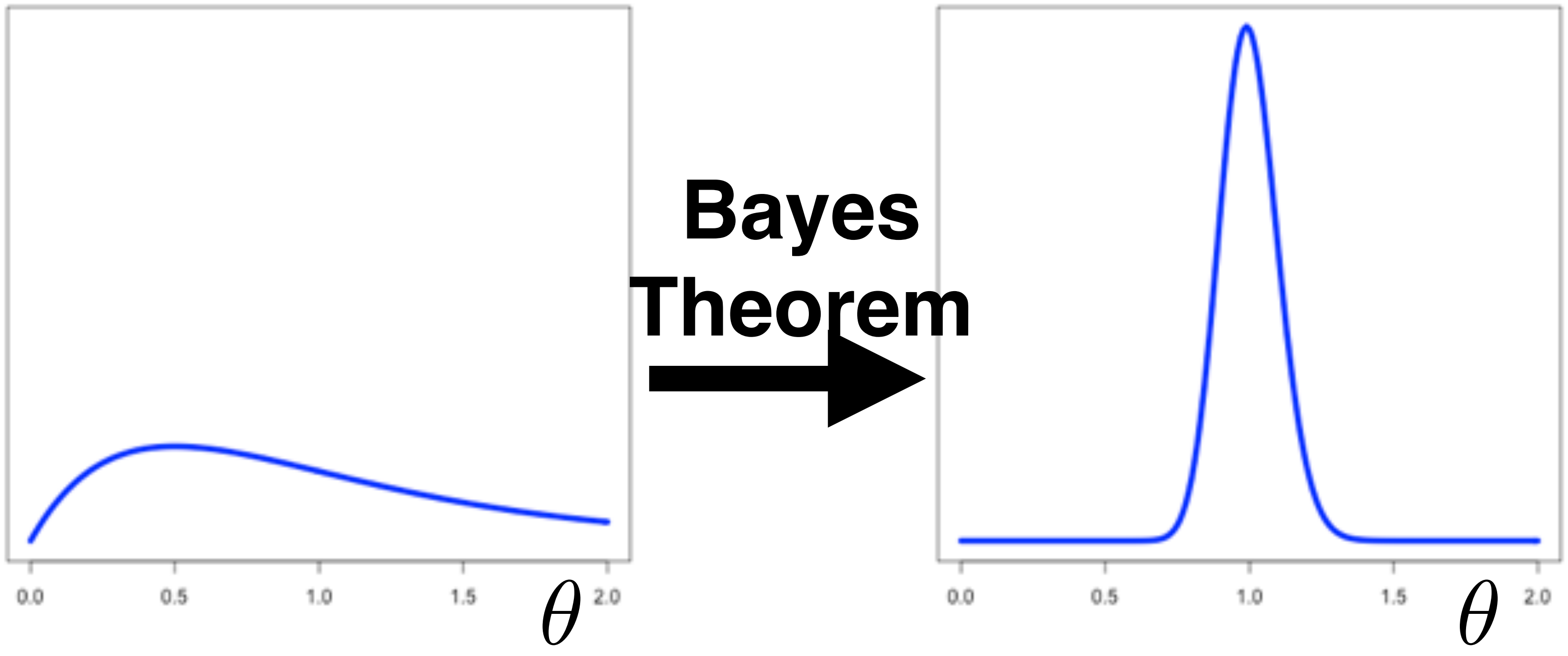
- Why are steps 2 and 3 hard?
  - Typically no closed form

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

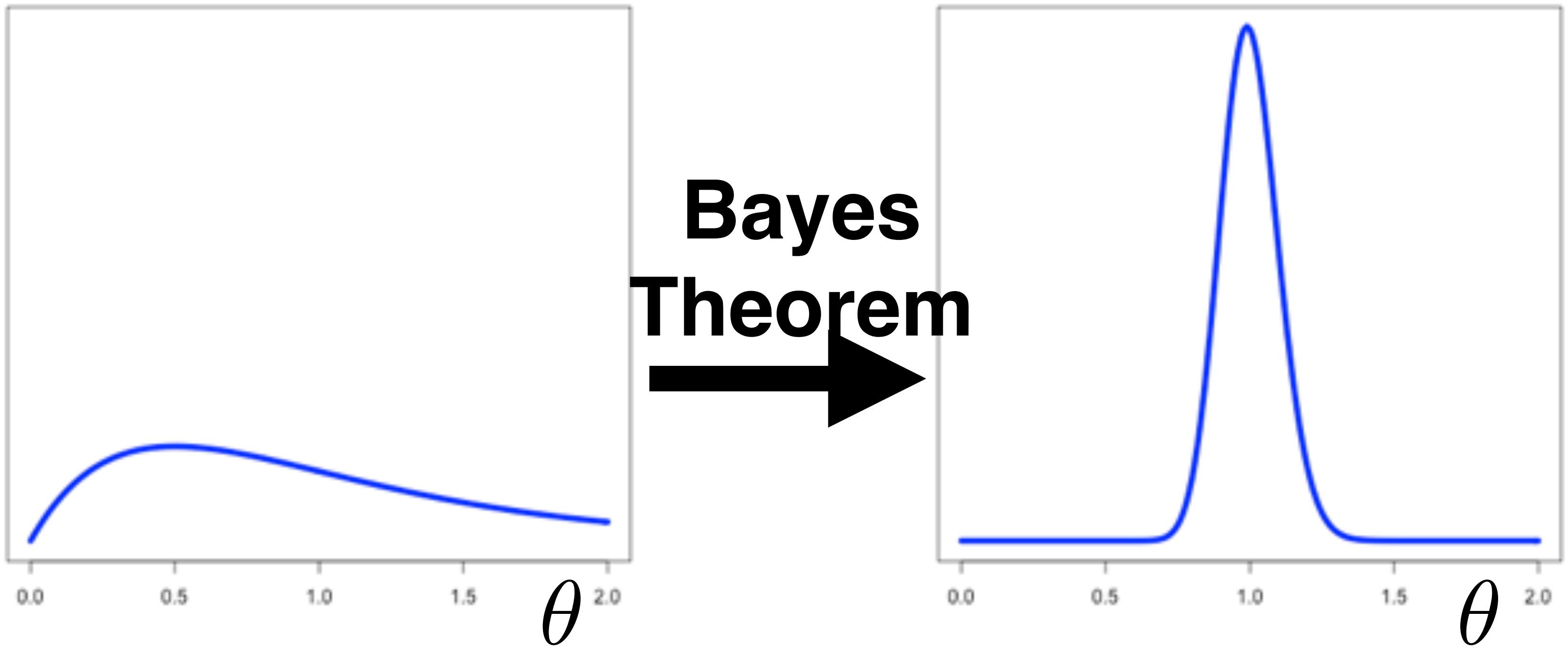
- Why are steps 2 and 3 hard?
    - Typically no closed form, high-dimensional integration

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

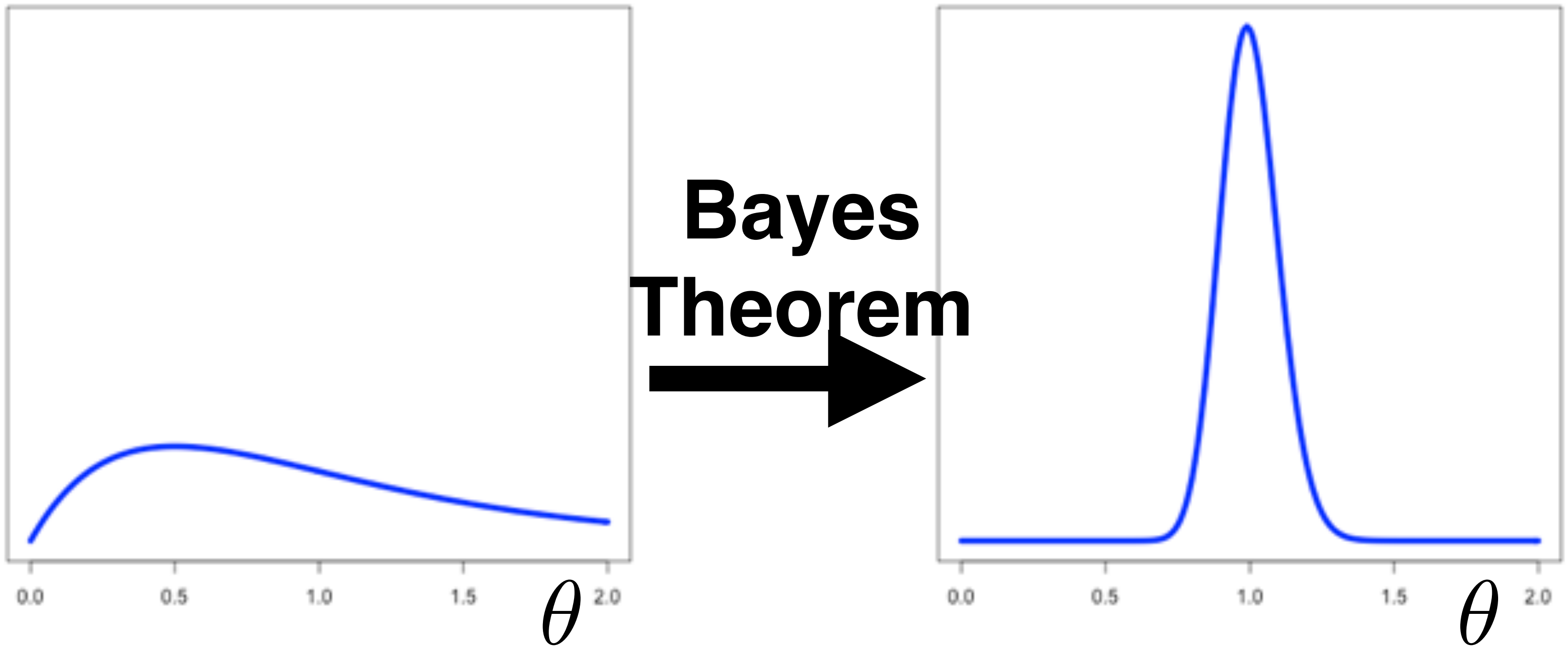
- Why are steps 2 and 3 hard?
  - Typically no closed form, high-dimensional integration

Bayesian inference

data parameters

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

posterior likelihood prior evidence

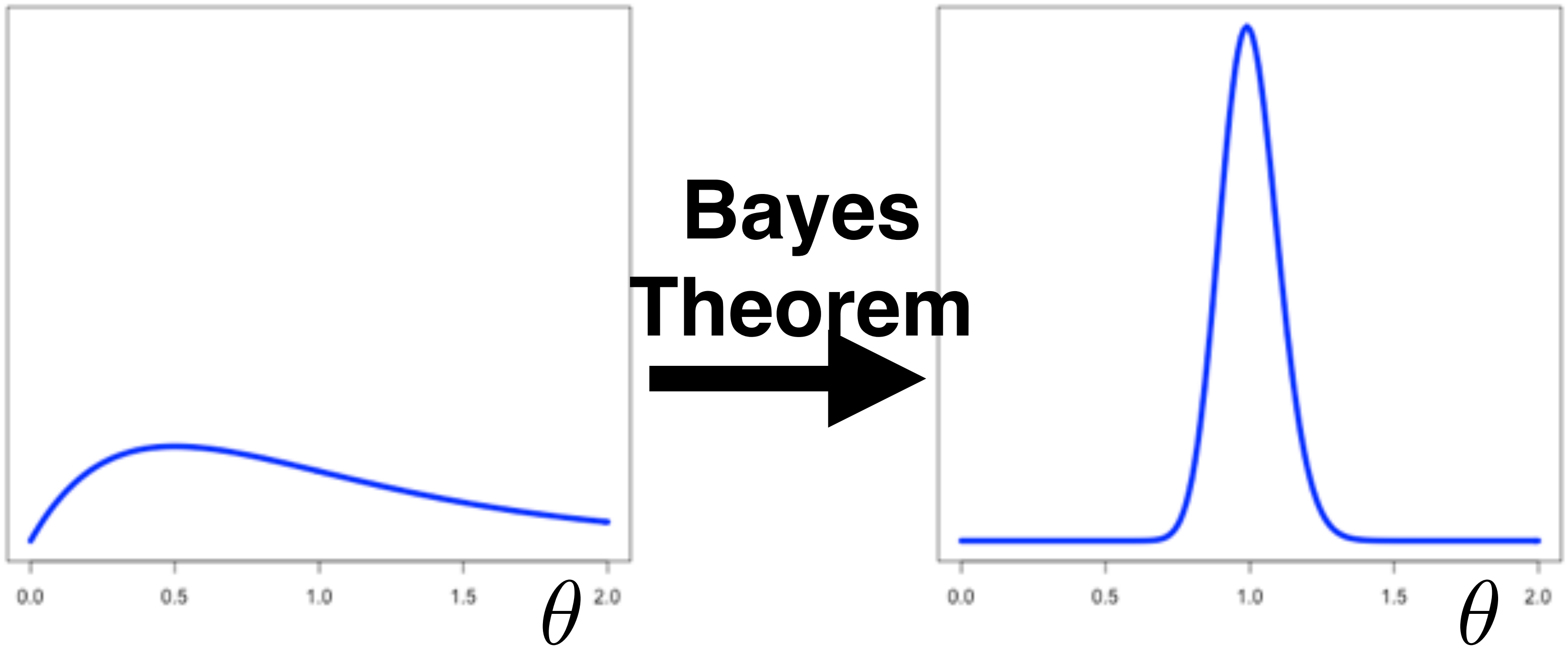


1. Build a model: choose prior & choose likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances

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

Bayesian inference

$$\underset{\text{posterior}}{p(\theta|y_{1:N})} = \underset{\text{likelihood}}{p(\overset{\text{data}}{y_{1:N}}|\theta)}\,\underset{\text{prior}}{\overset{\text{parameters}}{p(\theta)}}/\underset{\text{evidence}}{\int p(y_{1:N},\theta)d\theta}$$



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

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

[Bardenet, Doucet, Holmes 2017]

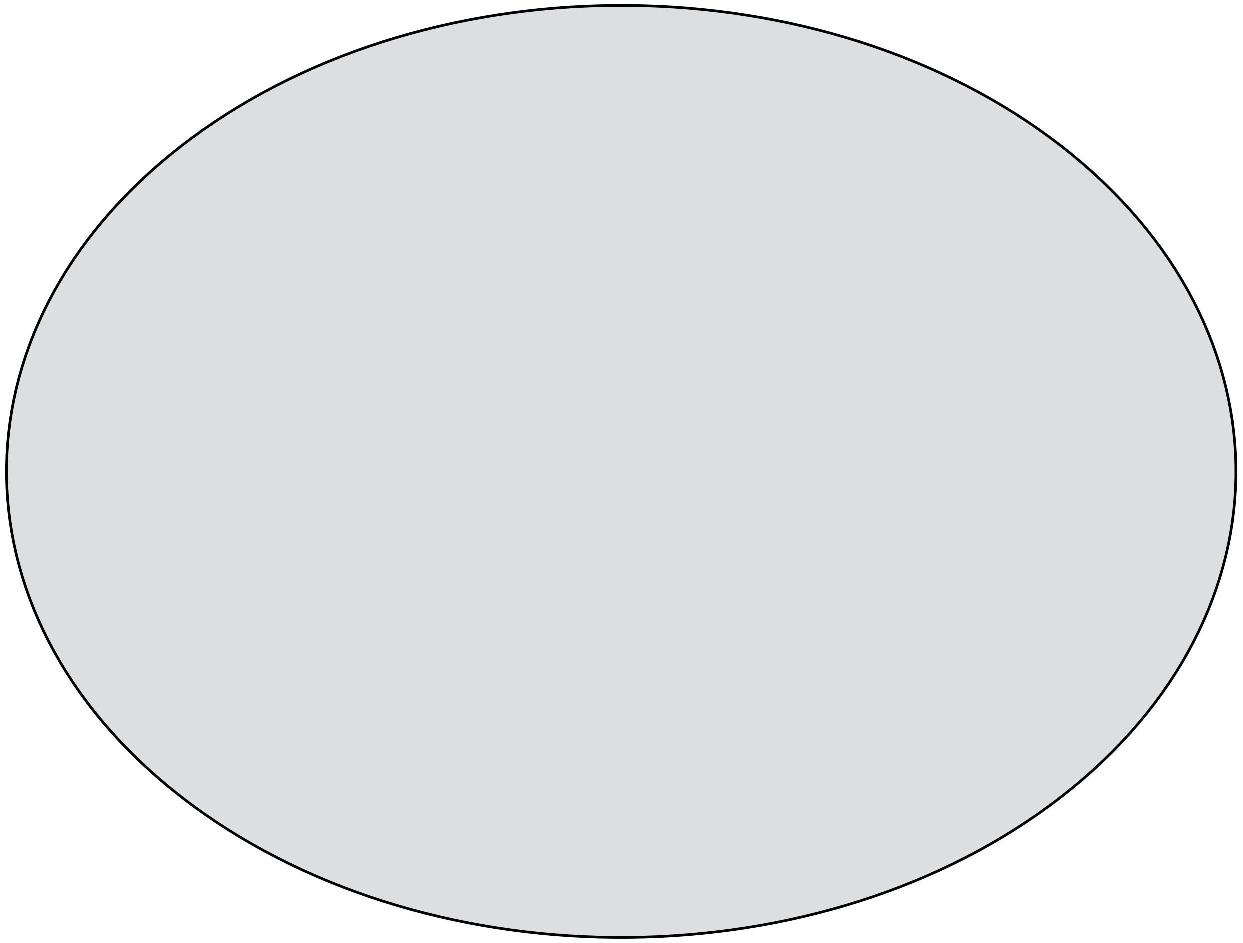
Instead: an optimization approach

- Approximate posterior with *q**

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

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



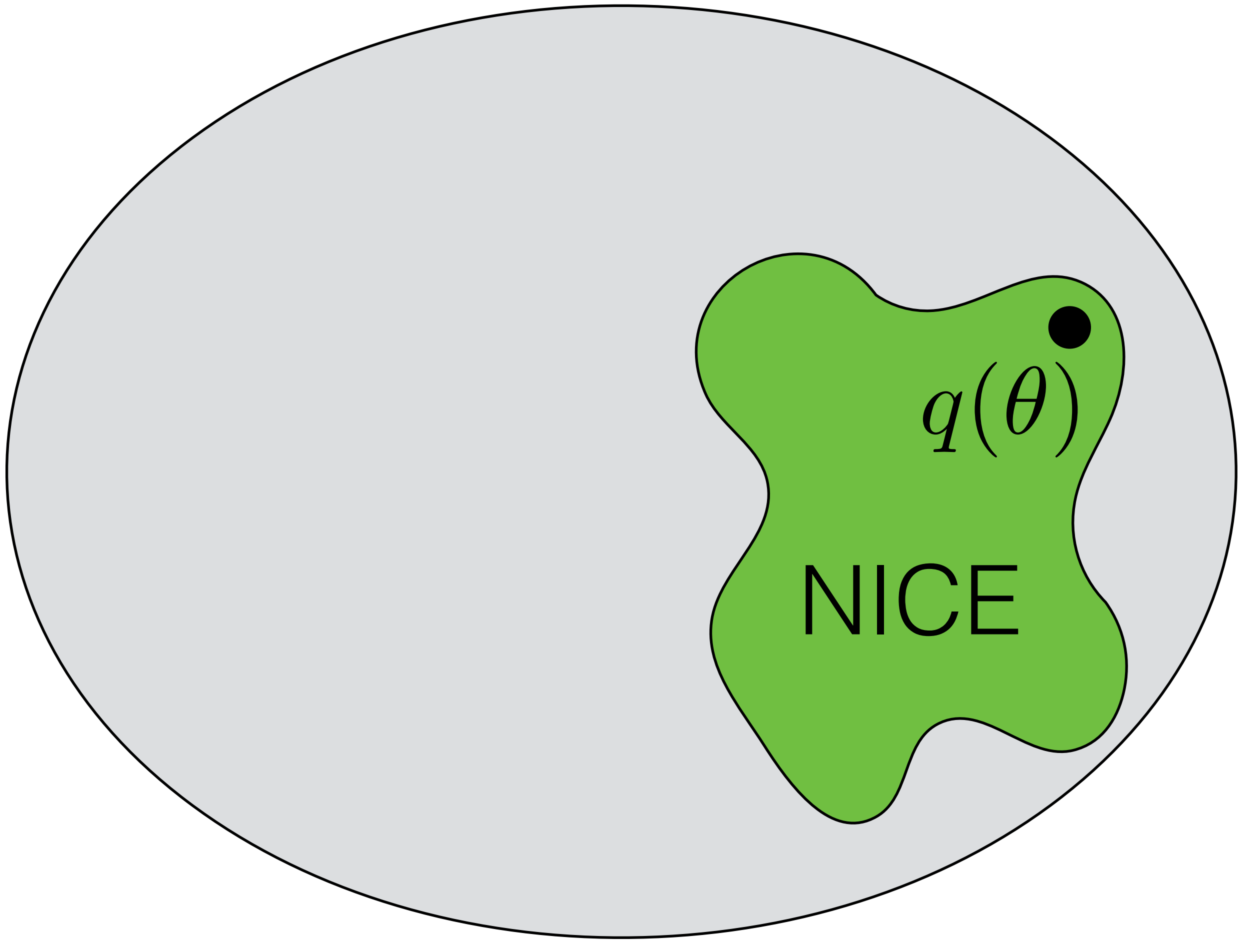
Instead: an optimization approach

- Approximate posterior with *q**

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

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



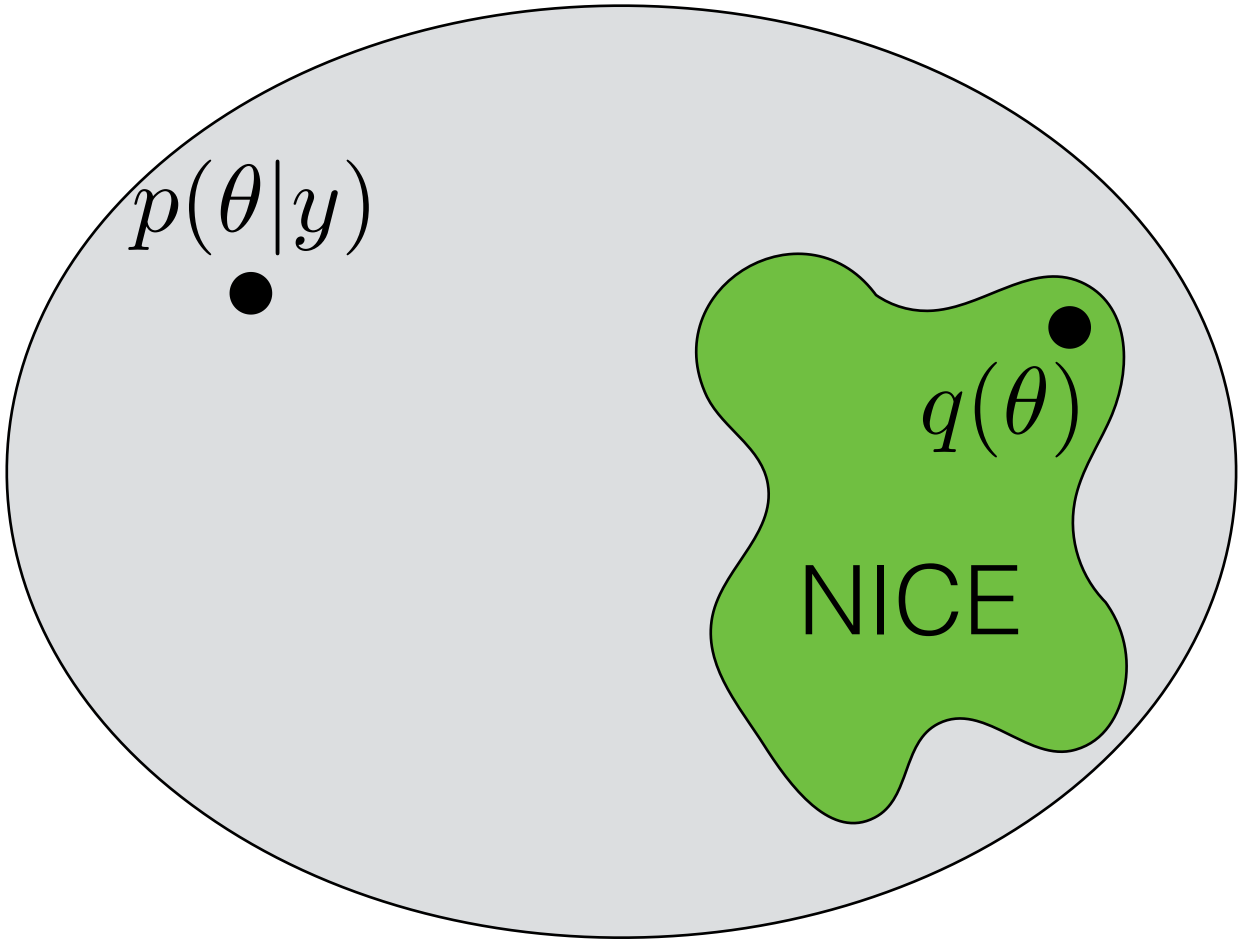
Instead: an optimization approach

- Approximate posterior with q^*

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

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



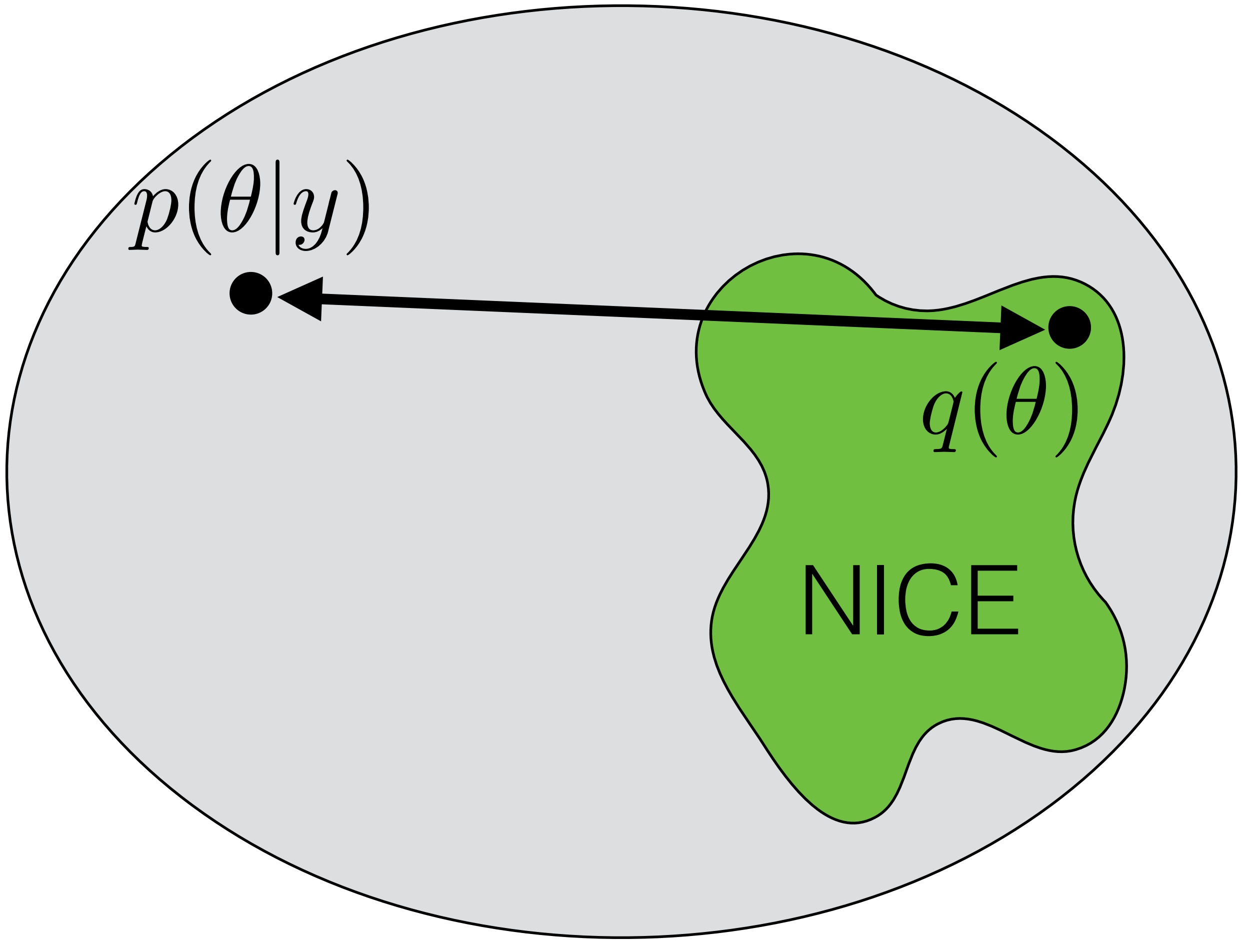
Instead: an optimization approach

- Approximate posterior with *q**

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

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



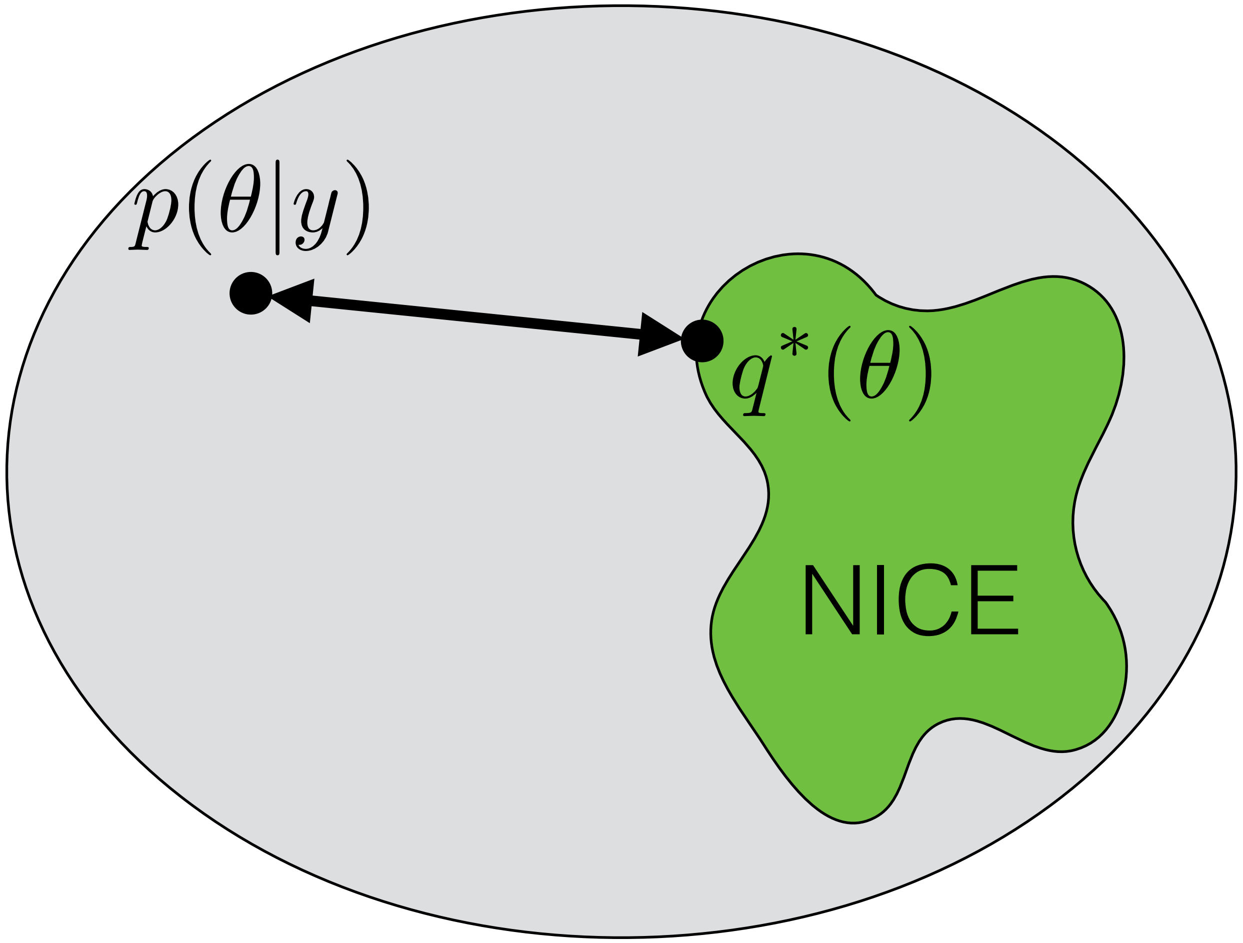
Instead: an optimization approach

- Approximate posterior with *q**

Approximate Bayesian Inference

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

[Bardenet, Doucet, Holmes 2017]



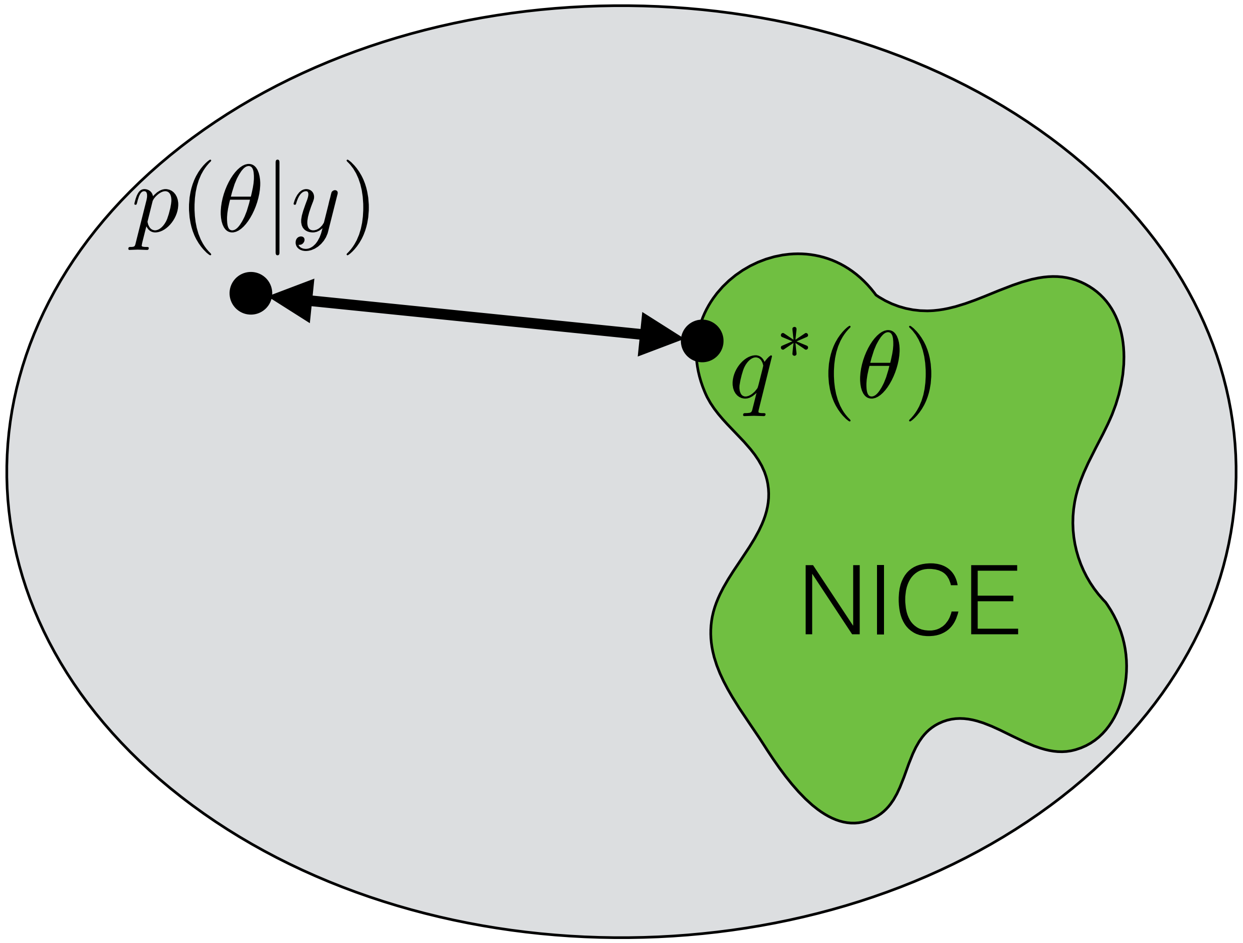
Instead: an optimization approach

- Approximate posterior with q^*

Approximate Bayesian Inference

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

[Bardenet, Doucet, Holmes 2017]



Instead: an optimization approach

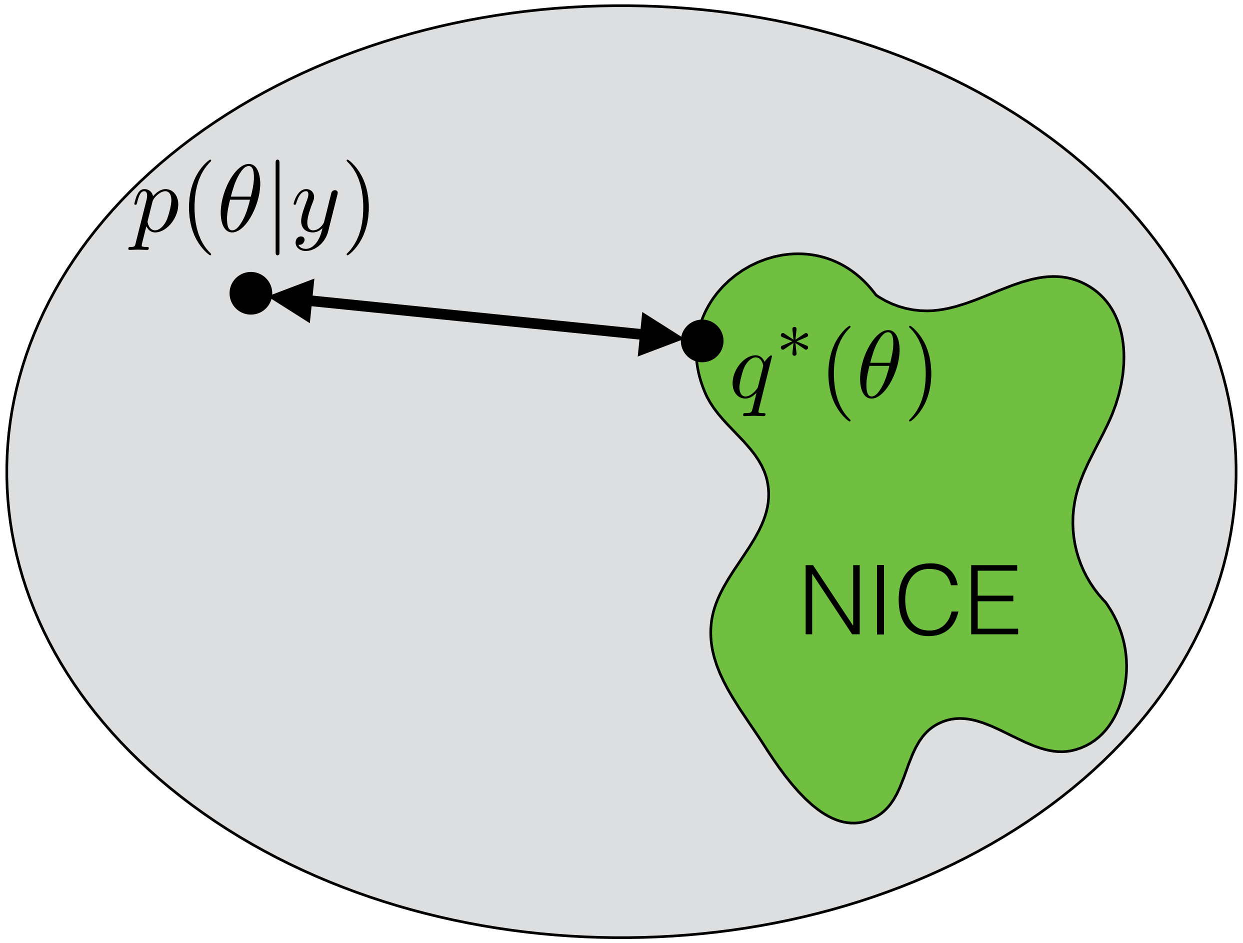
- Approximate posterior with q^*

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

Approximate Bayesian Inference

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

[Bardenet, Doucet, Holmes 2017]



Instead: an optimization approach

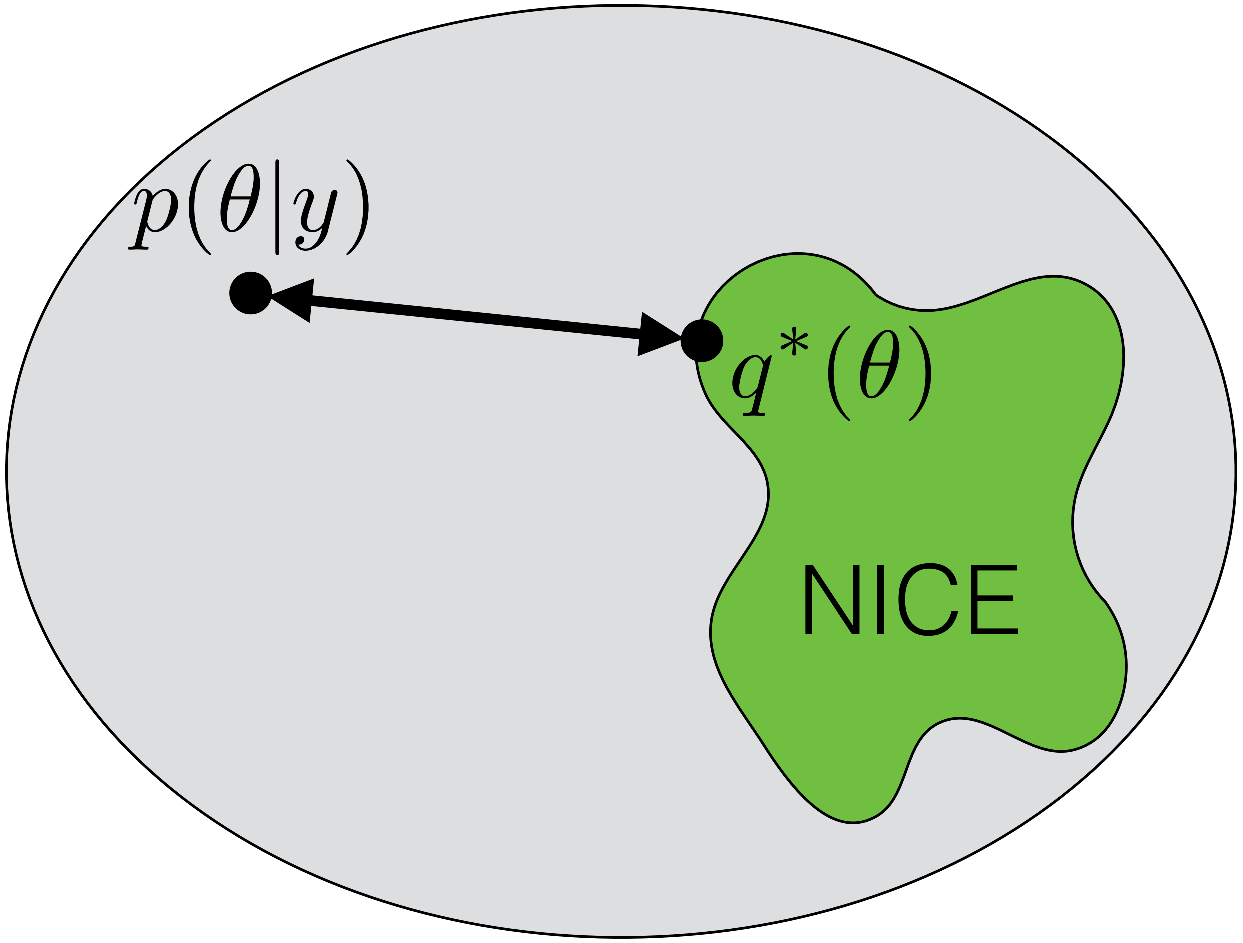
- Approximate posterior with q^*

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

Approximate Bayesian Inference

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

[Bardenet, Doucet, Holmes 2017]



Instead: an optimization approach

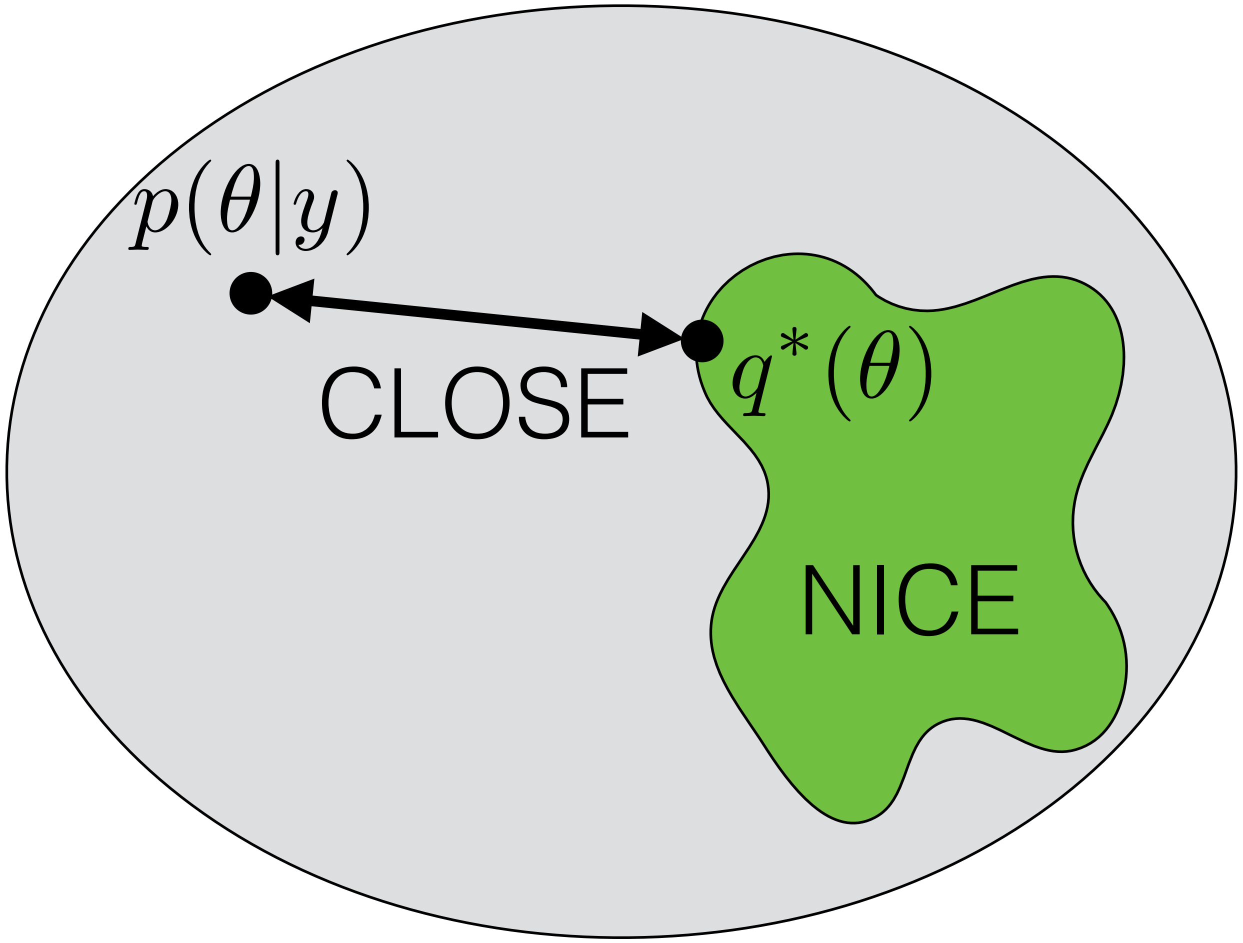
- Approximate posterior with q^*

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

Approximate Bayesian Inference

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

[Bardenet, Doucet, Holmes 2017]



Instead: an optimization approach

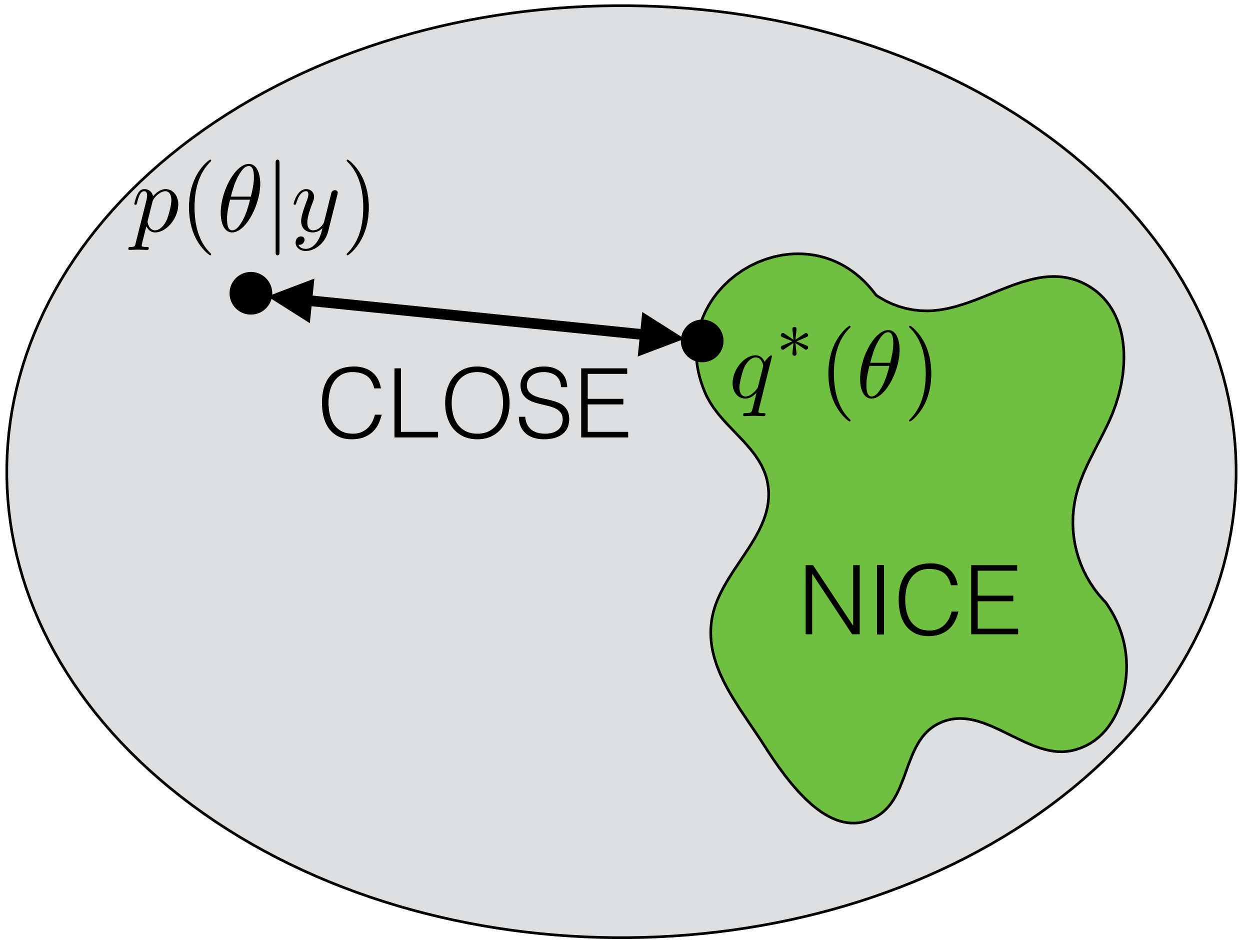
- Approximate posterior with q^*

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

Approximate Bayesian Inference

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

[Bardenet, Doucet, Holmes 2017]



Instead: an optimization approach

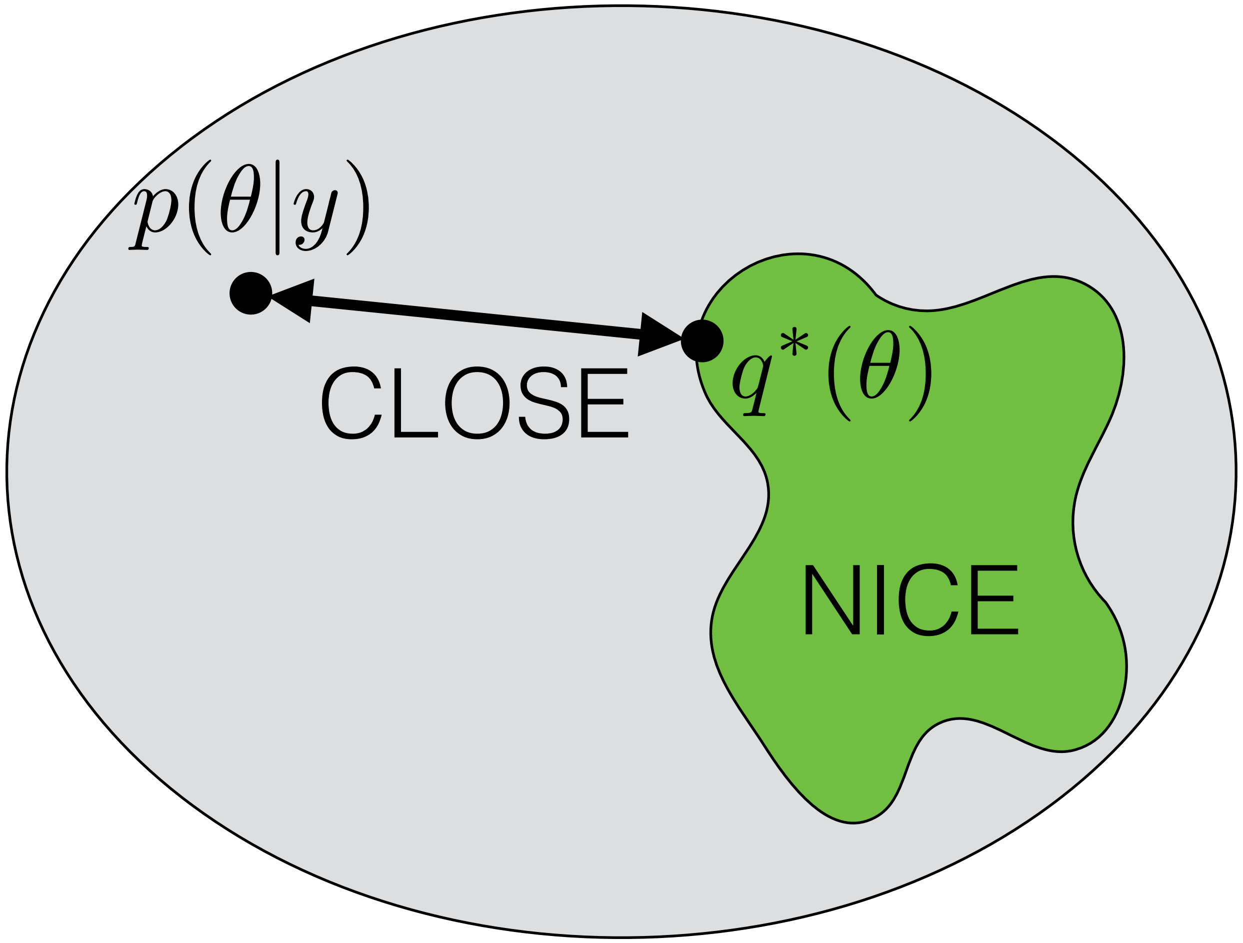
- Approximate posterior with q^*

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

Approximate Bayesian Inference

- 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^*

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

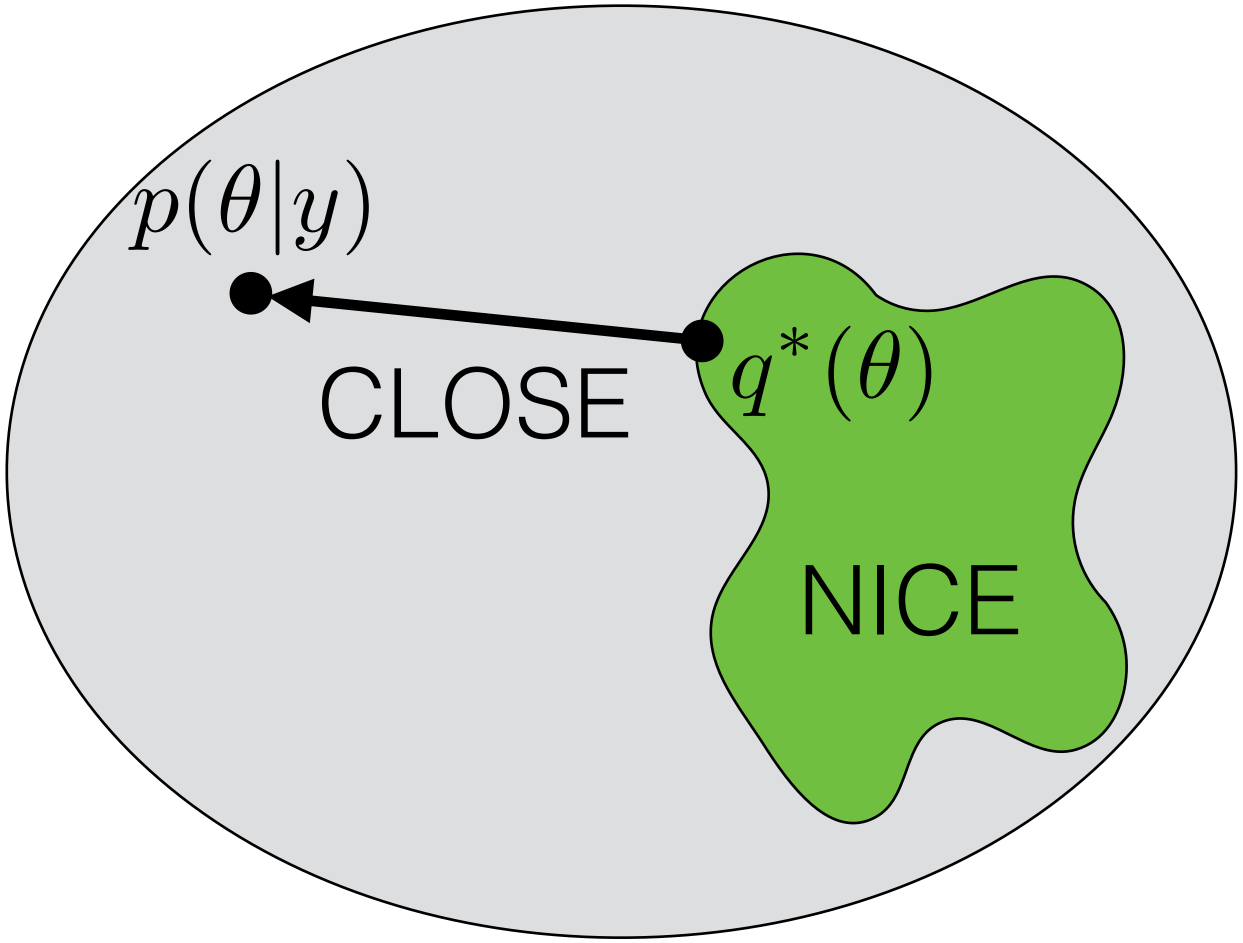
- Variational Bayes (VB): *f* is Kullback-Leibler divergence

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

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**

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

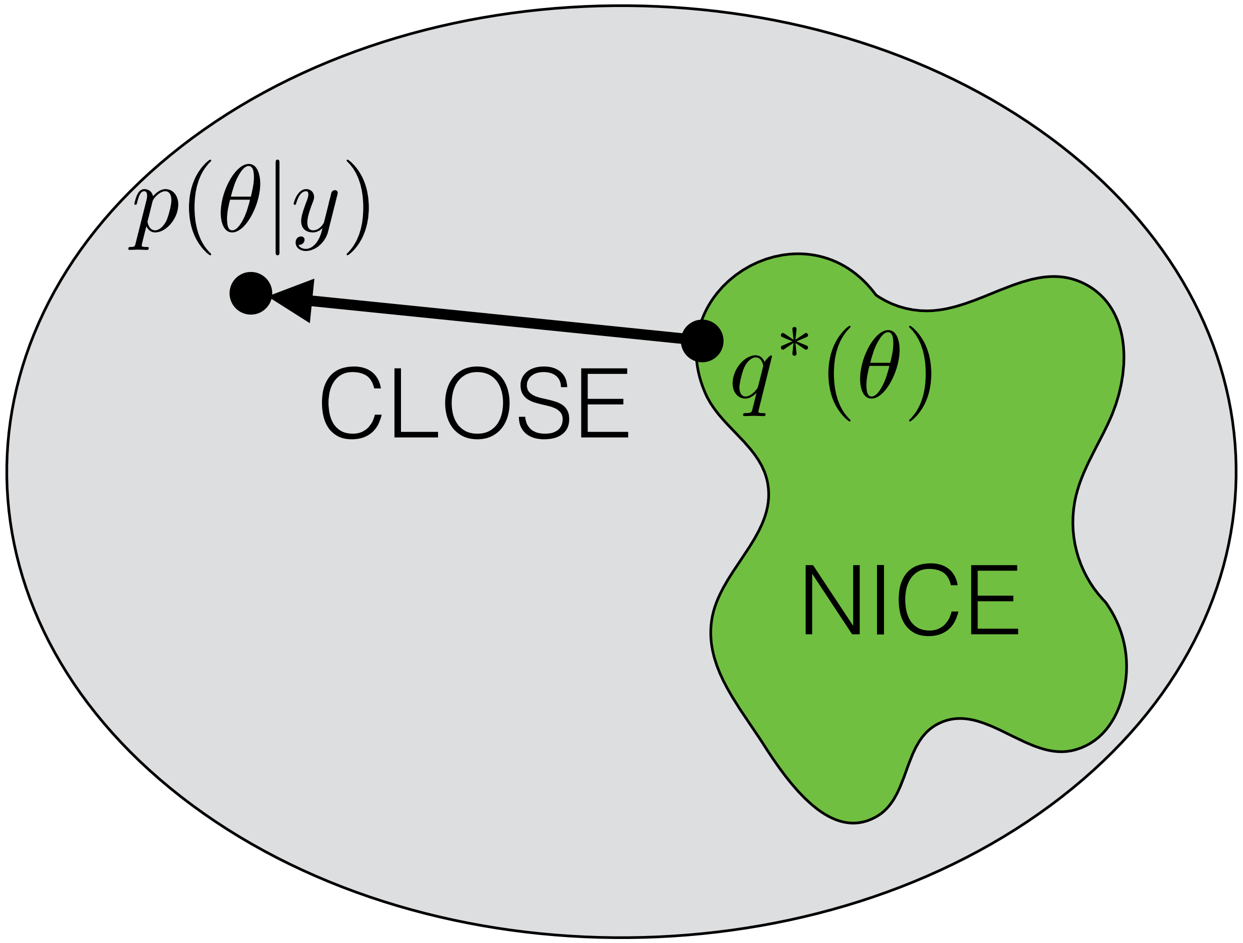
- Variational Bayes (VB): *f* is Kullback-Leibler divergence

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

Approximate Bayesian Inference

- 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

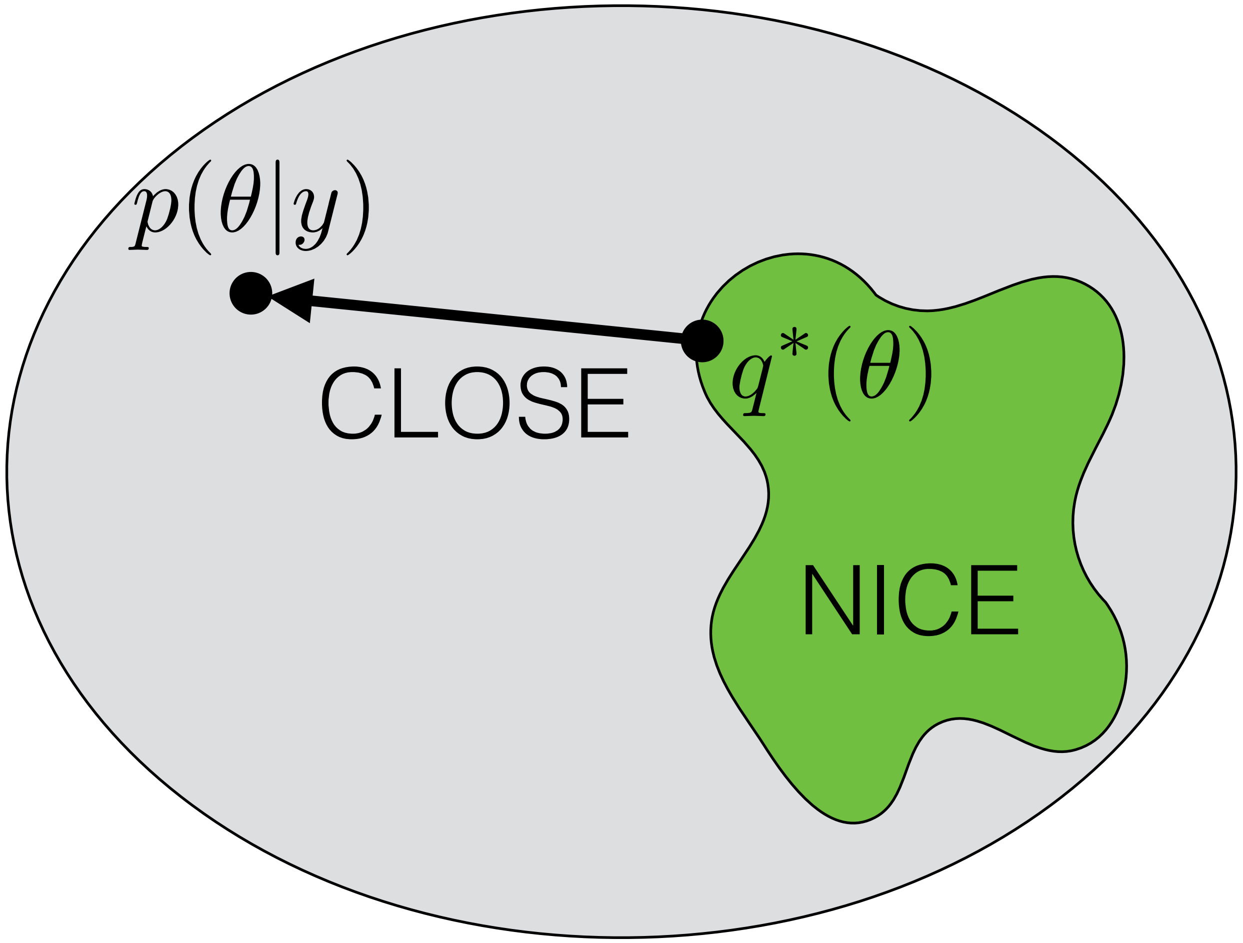
$$KL(q(\cdot)||p(\cdot|y))$$

- VB practical success

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

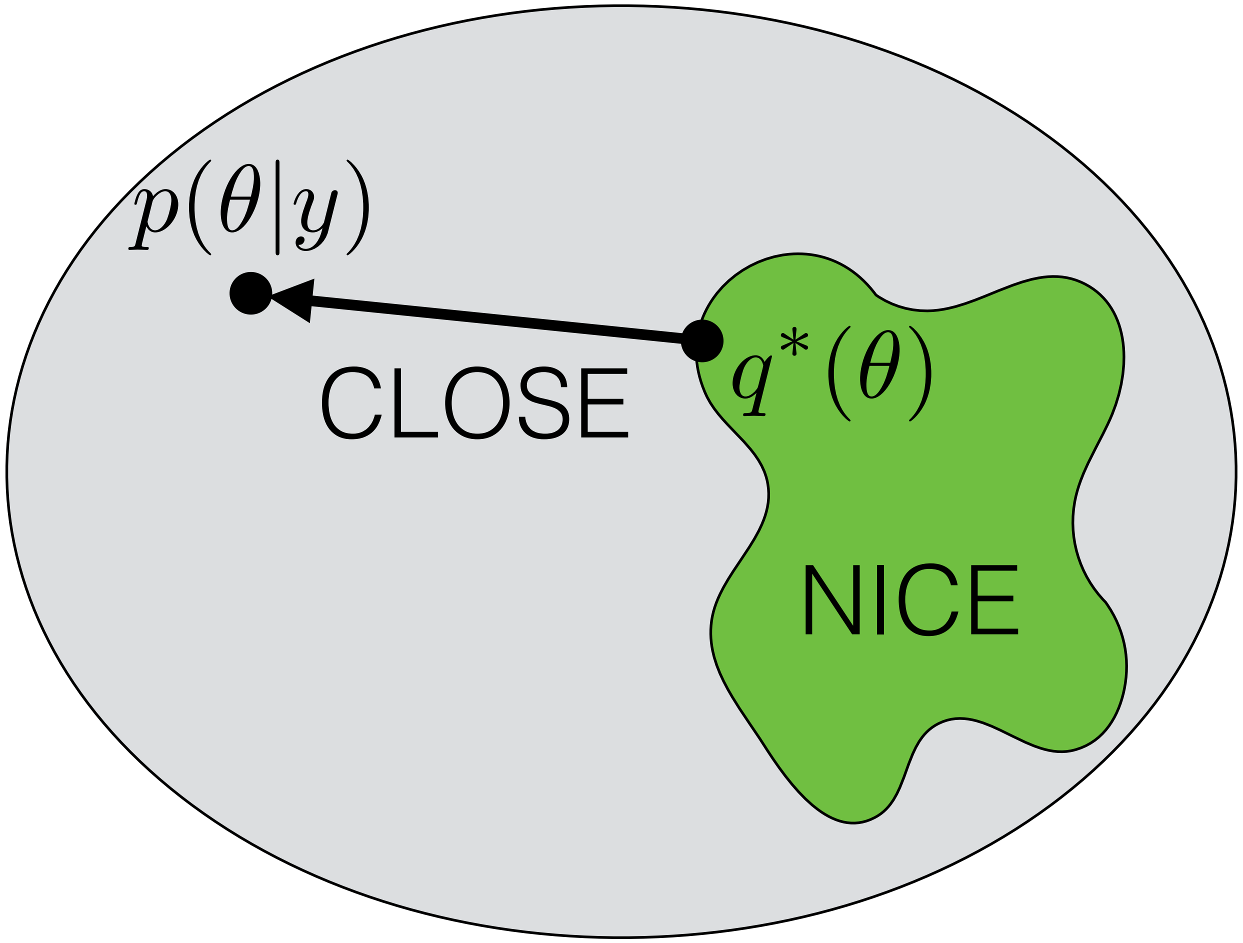
$$KL(q(\cdot)||p(\cdot|y))$$

- VB practical success: point estimates and prediction

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**

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

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

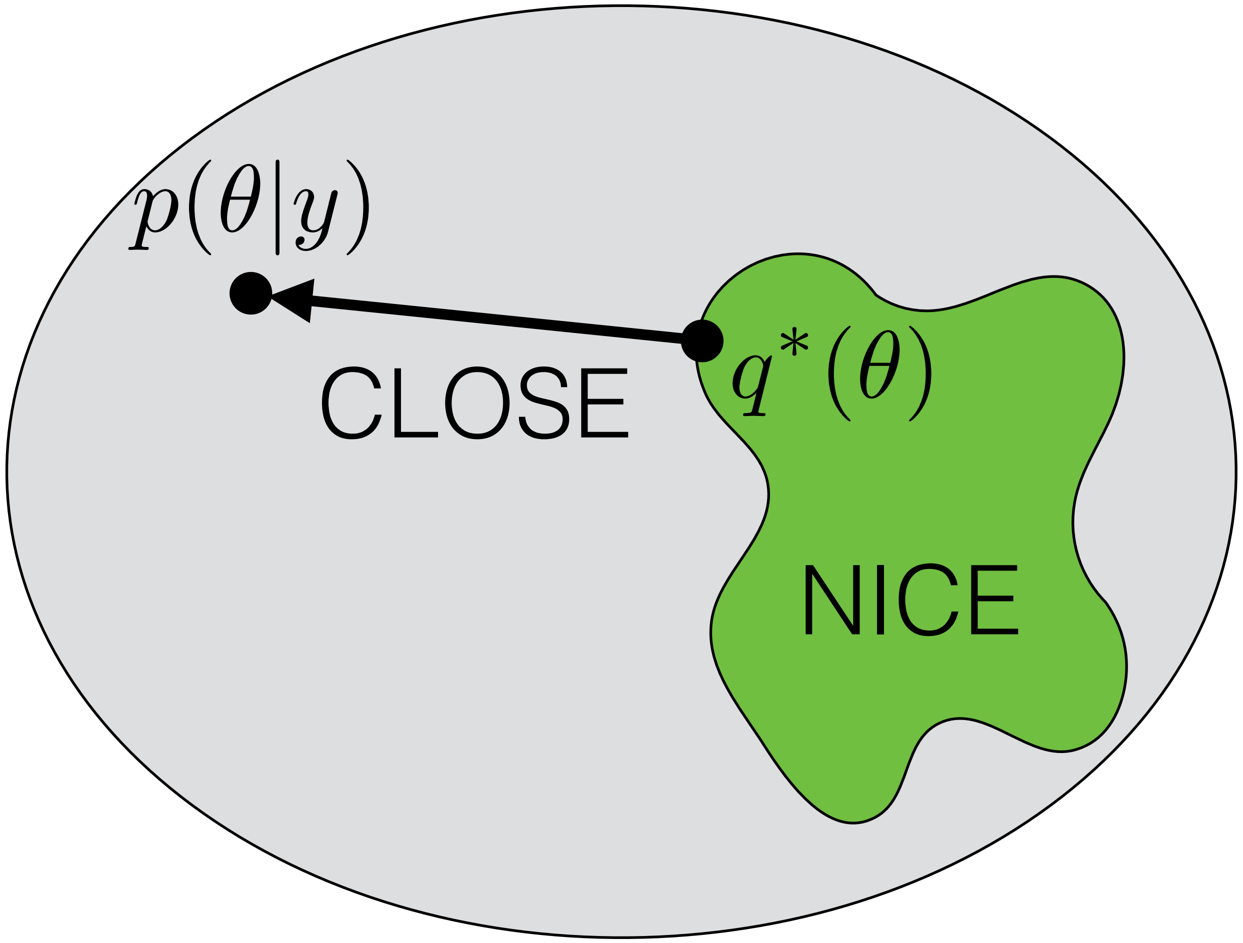
$$KL(q(\cdot)||p(\cdot|y))$$

- VB practical success: point estimates and prediction, fast

Approximate Bayesian Inference

- 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

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

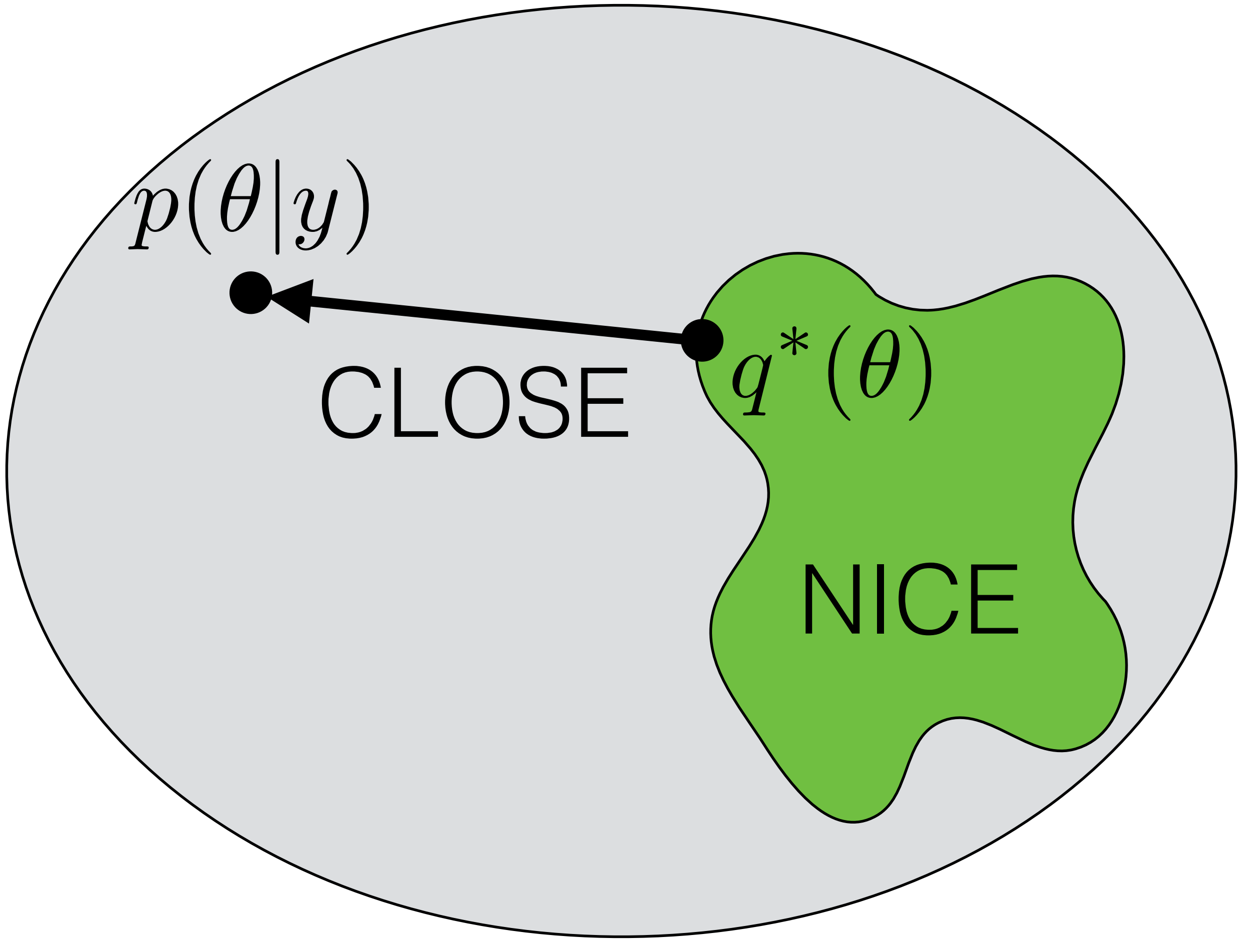
- VB practical success: point estimates and prediction, fast, streaming, distributed (3.6M Wikipedia, 350K Nature)

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

Why KL?

- Variational Bayes

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

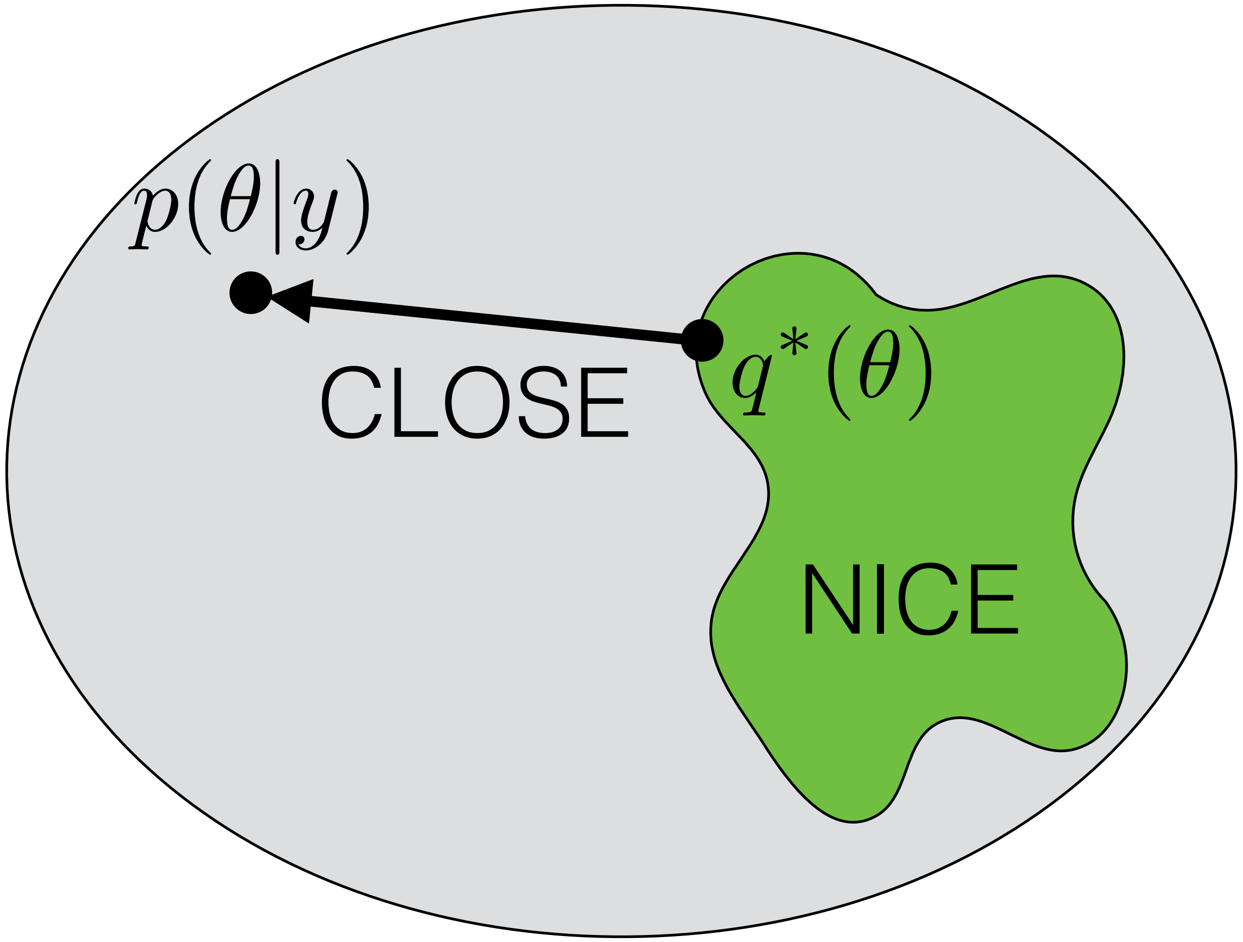


Why KL?

- Variational Bayes

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

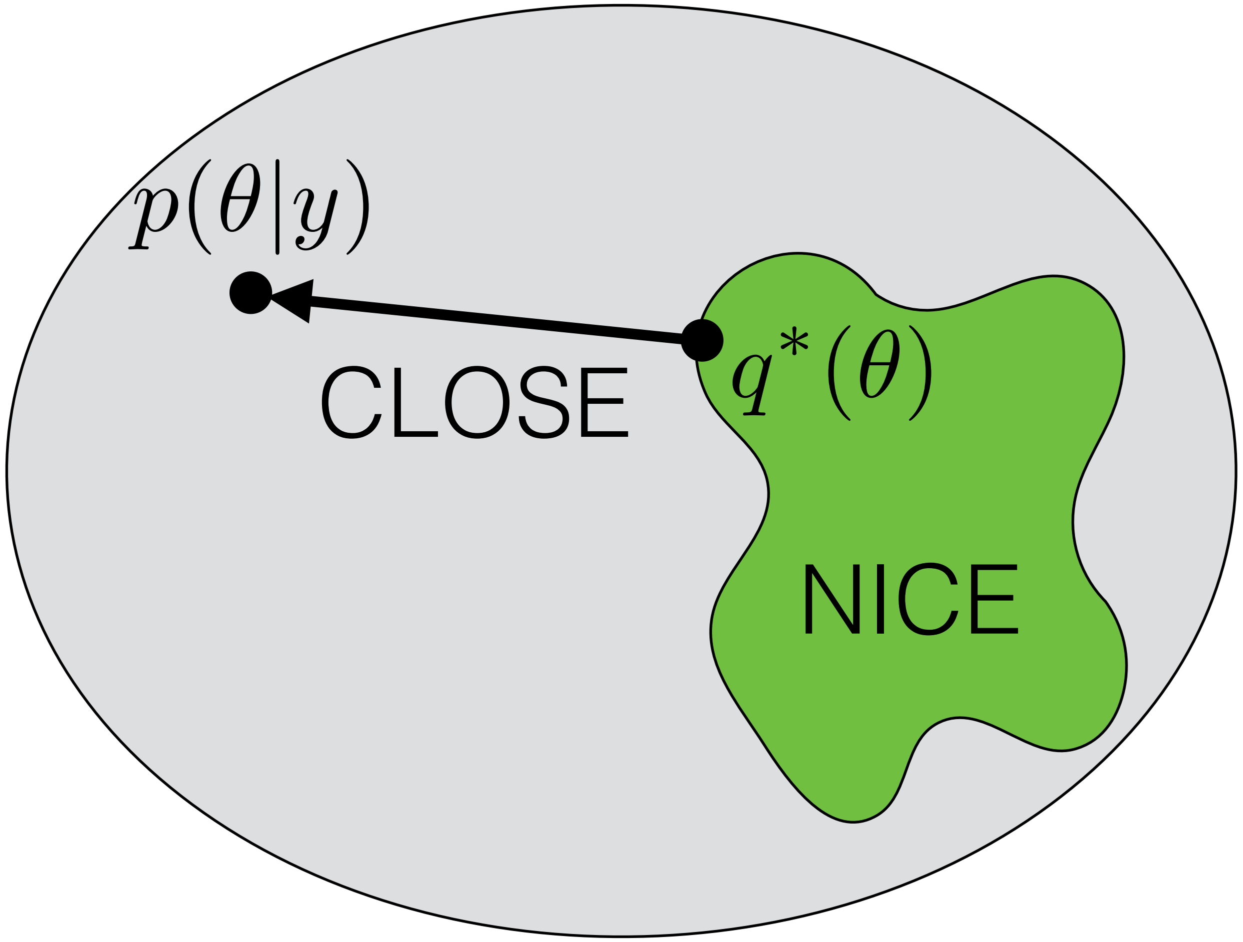
$$\mathrm{KL}\left(q(\cdot) || p(\cdot | y)\right)$$
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta | y)} d\theta$$



Why KL?

- Variational Bayes

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



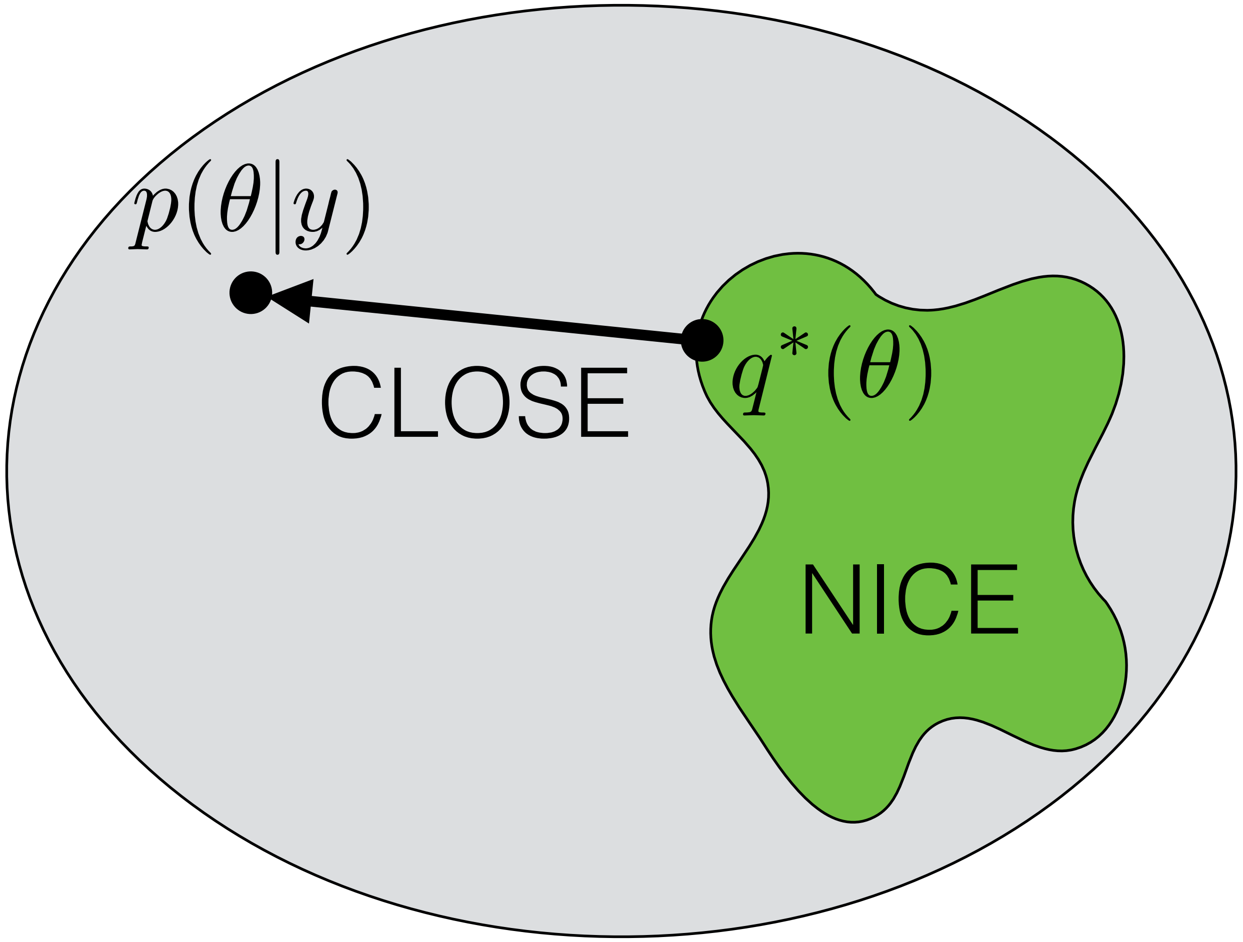
$$\mathrm{KL}\left(q(\cdot) || p(\cdot | y)\right)$$

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

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

Why KL?

- Variational Bayes



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

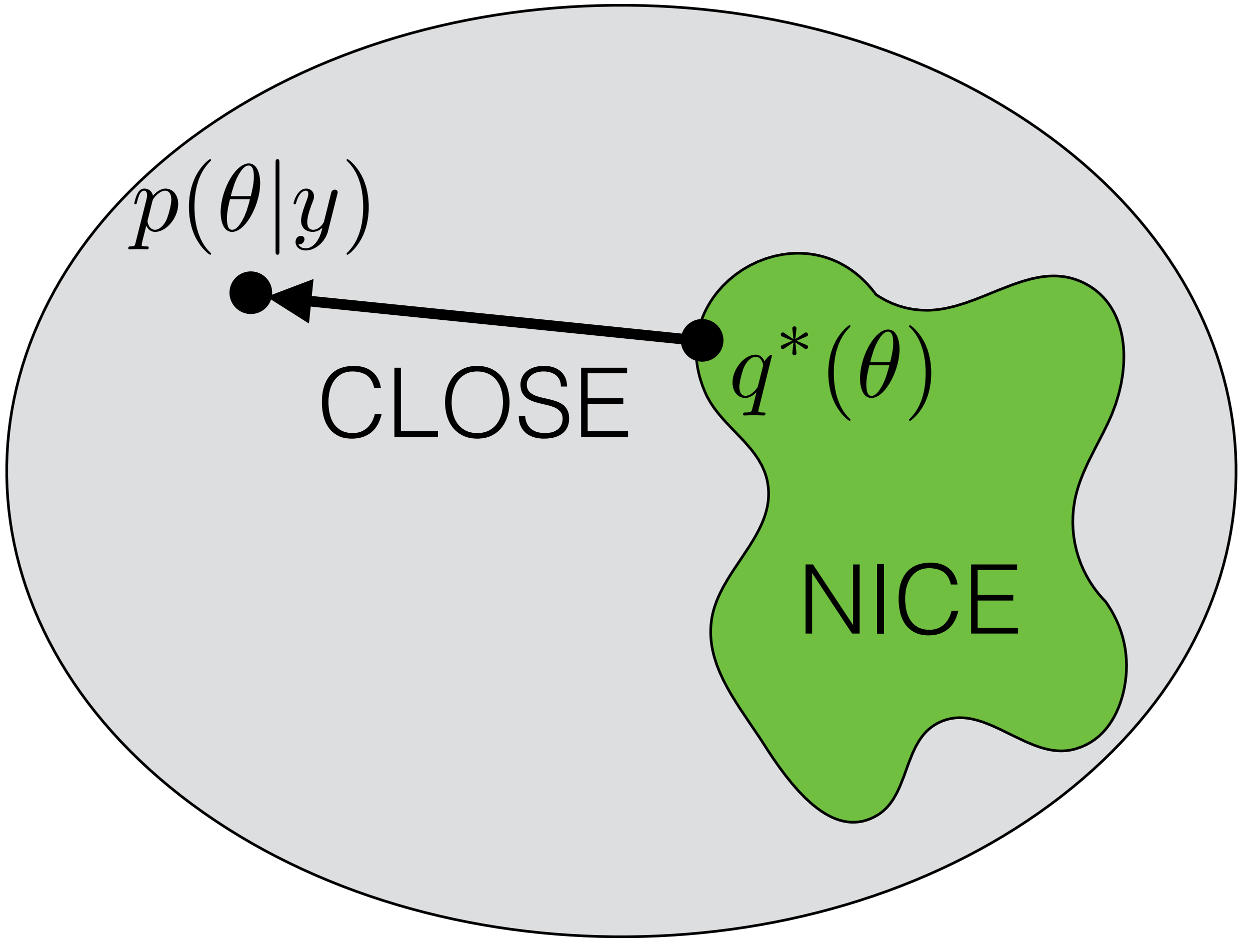
$$\text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$

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

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

Why KL?

- Variational Bayes



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

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

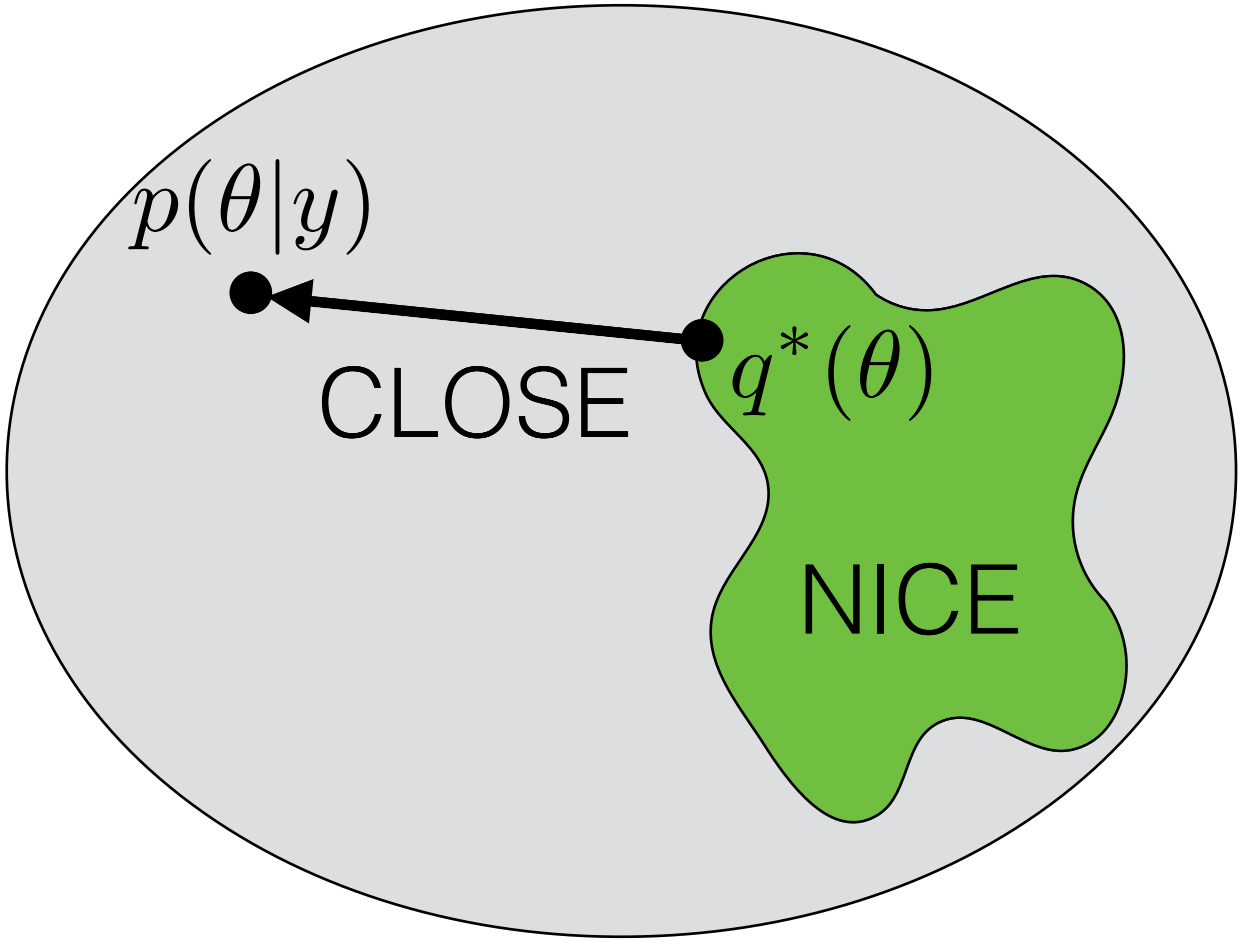
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

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

Why KL?

- Variational Bayes

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



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

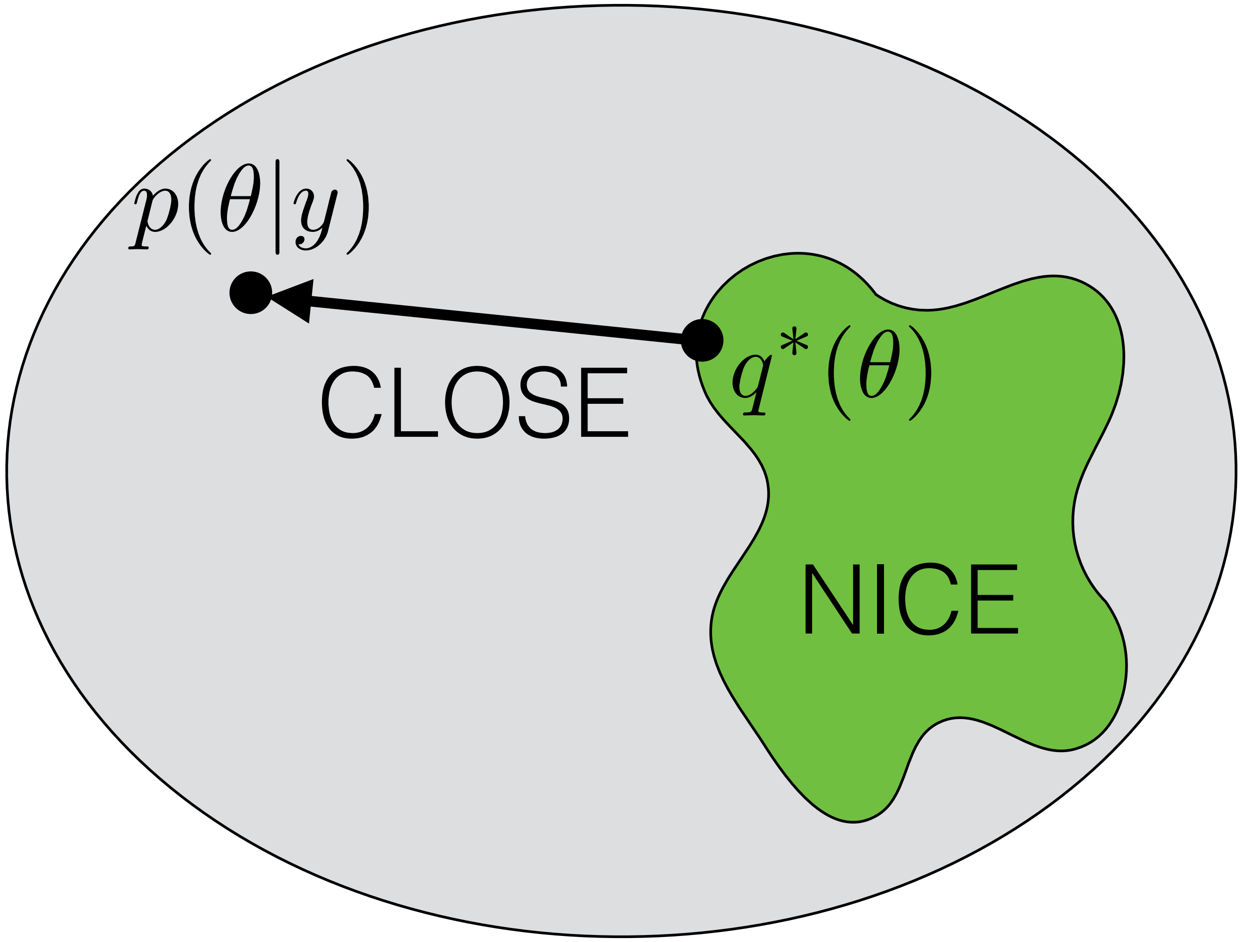
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

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

Why KL?

- Variational Bayes

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



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

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

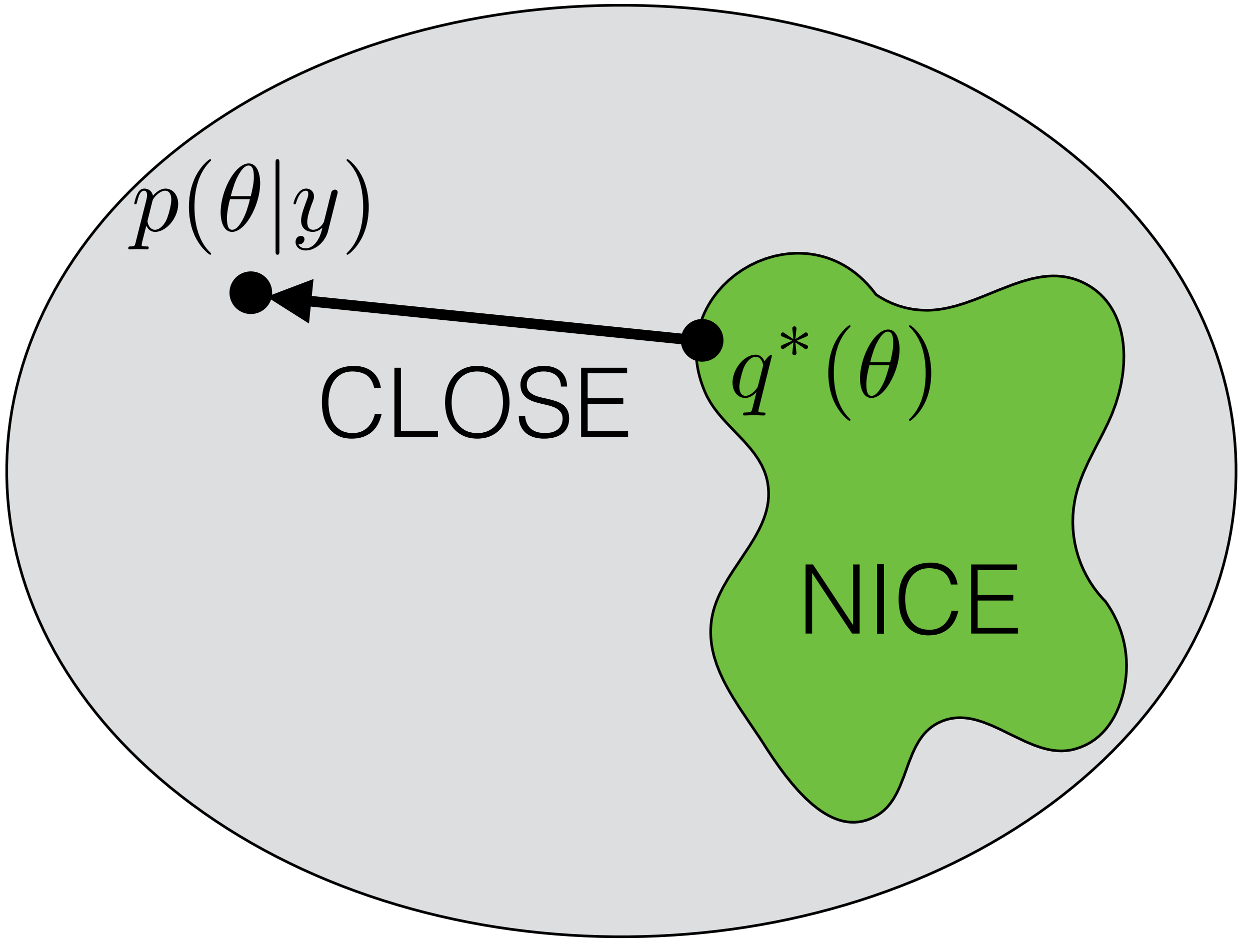
$$= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta$$

“Evidence lower bound” (ELBO)

Why KL?

- Variational Bayes

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



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

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

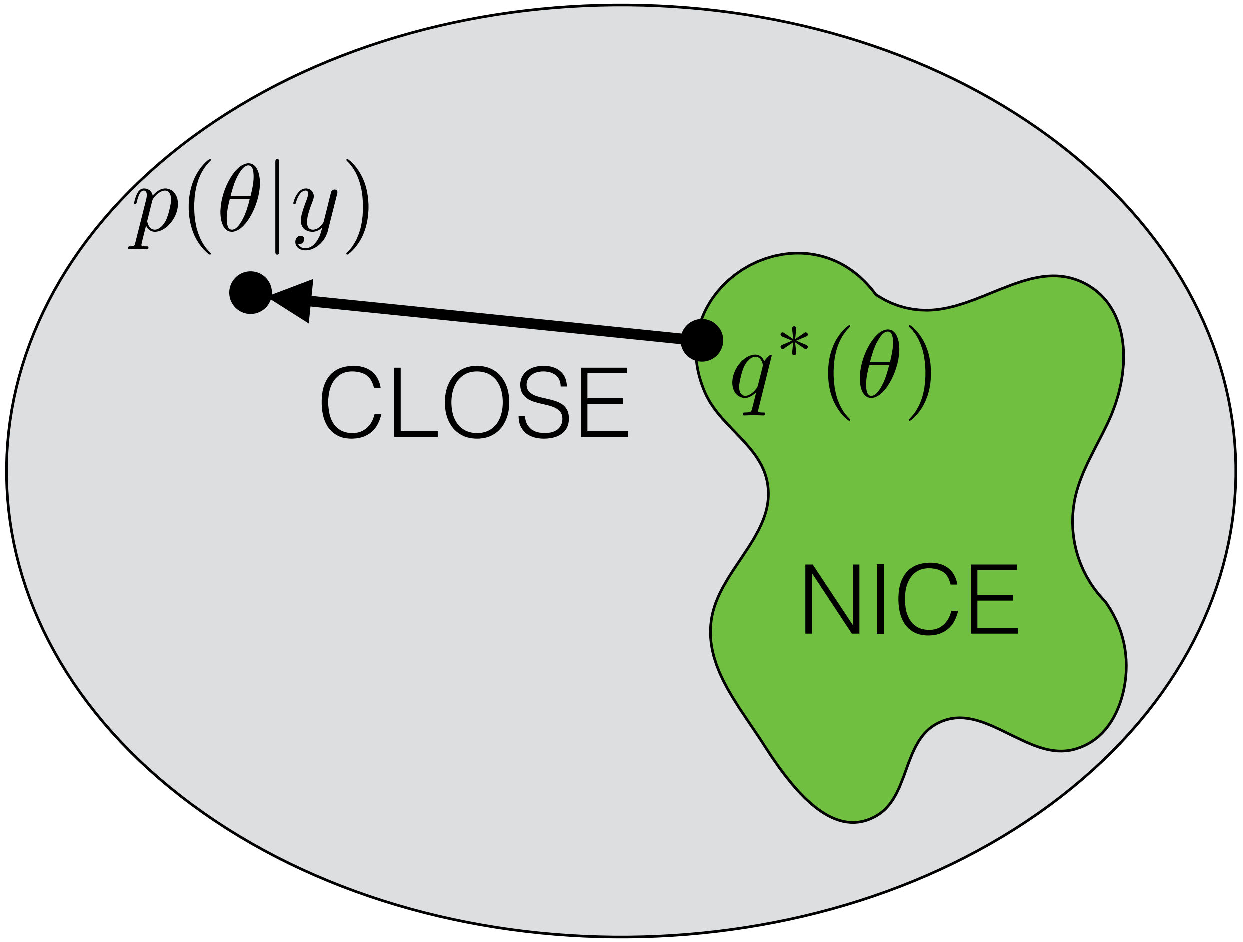
$$= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta,y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta,y)}{q(\theta)} d\theta$$

“Evidence lower bound” (ELBO)

Why KL?

- Variational Bayes

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



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

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

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

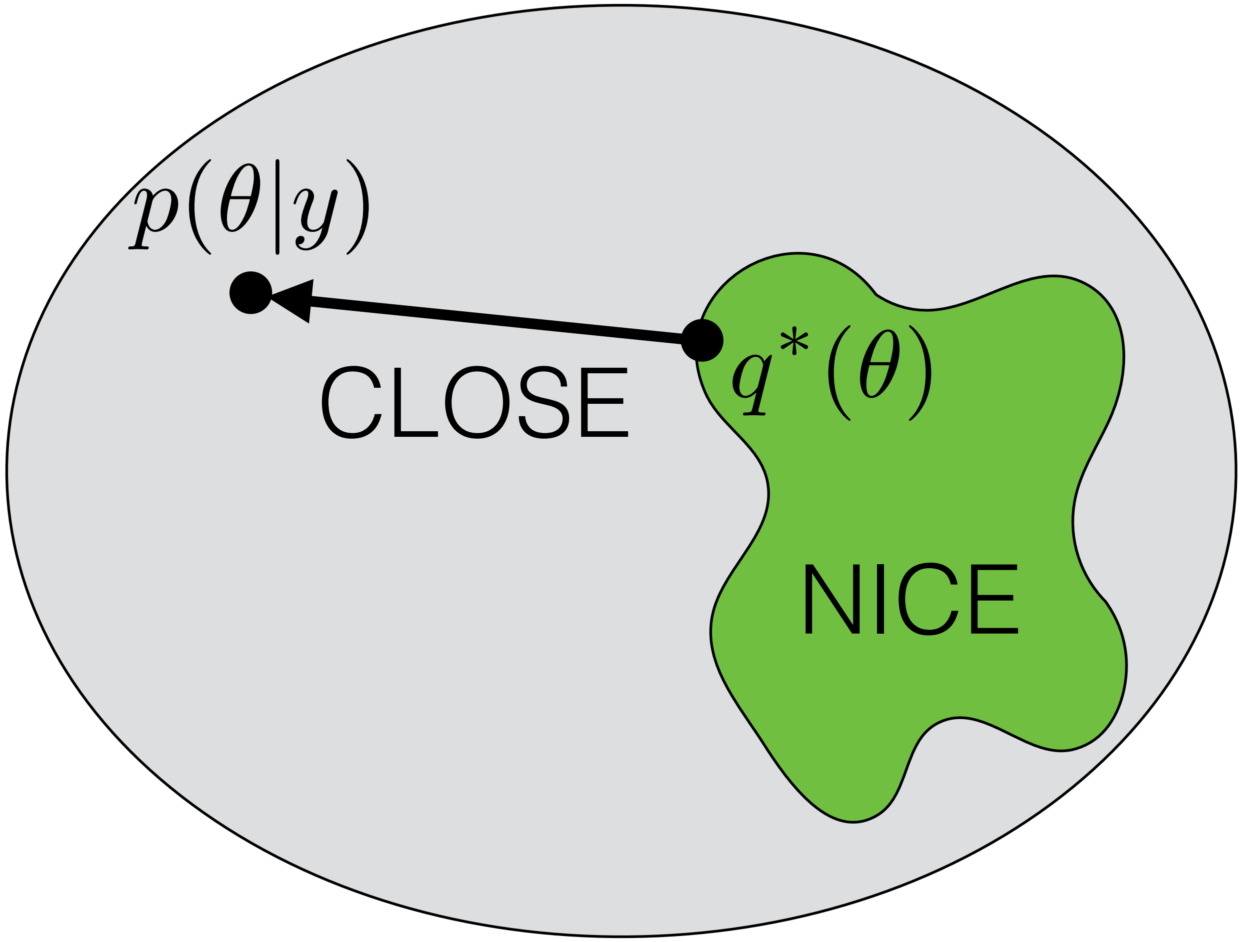
"Evidence lower bound" (ELBO)

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

Why KL?

- Variational Bayes

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



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

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

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

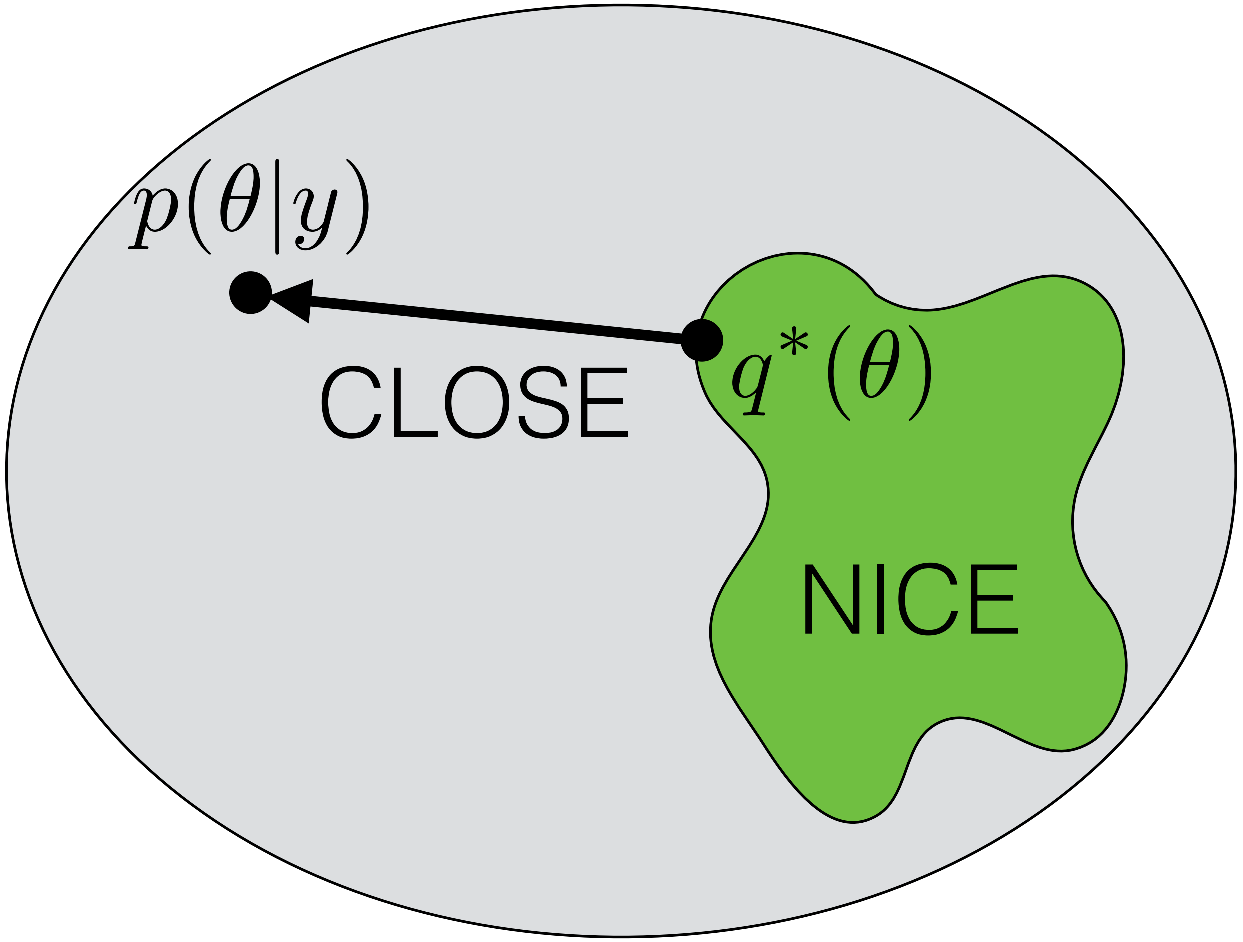
"Evidence lower bound" (ELBO)

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

Why KL?

- Variational Bayes

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



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

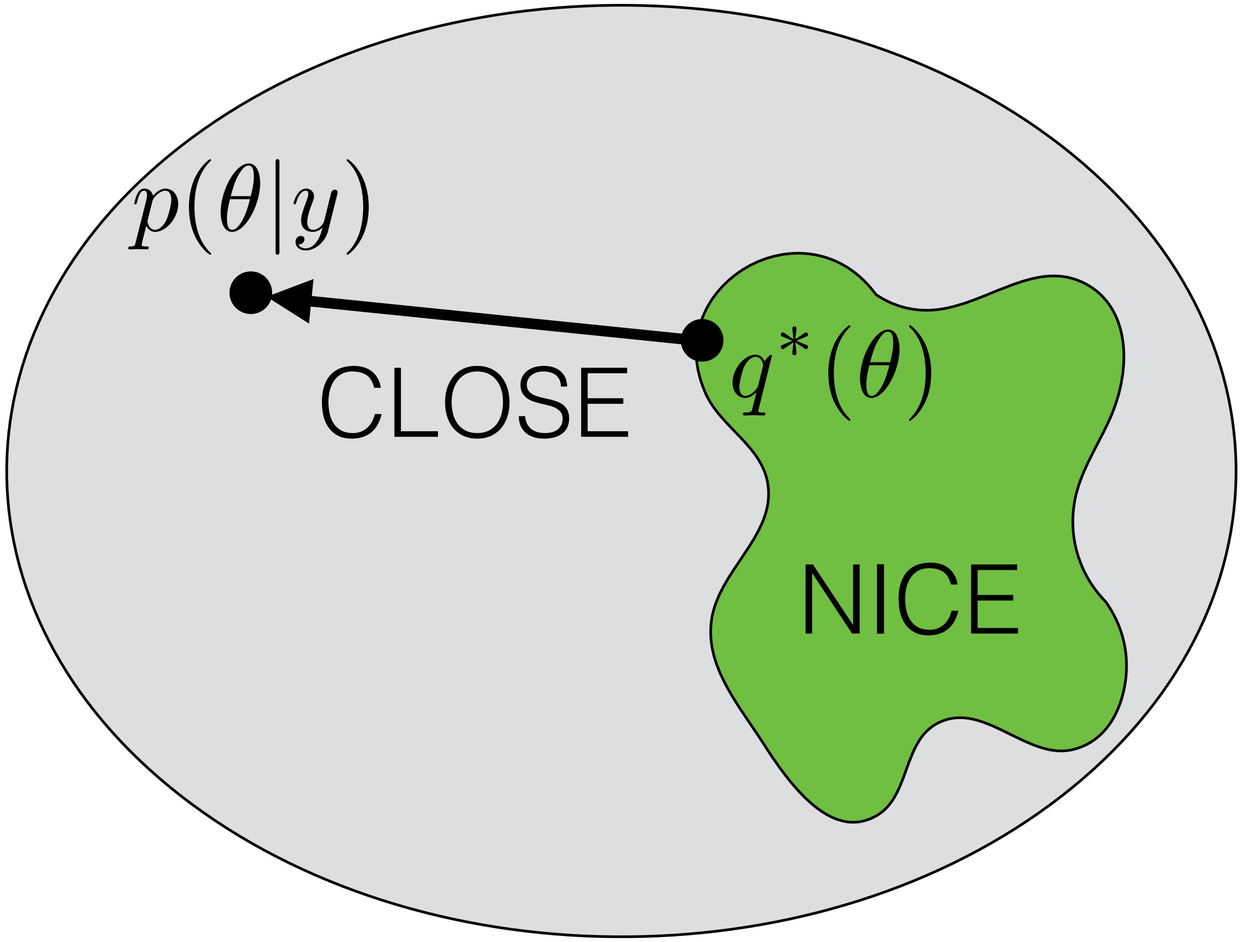
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

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

"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

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

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

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

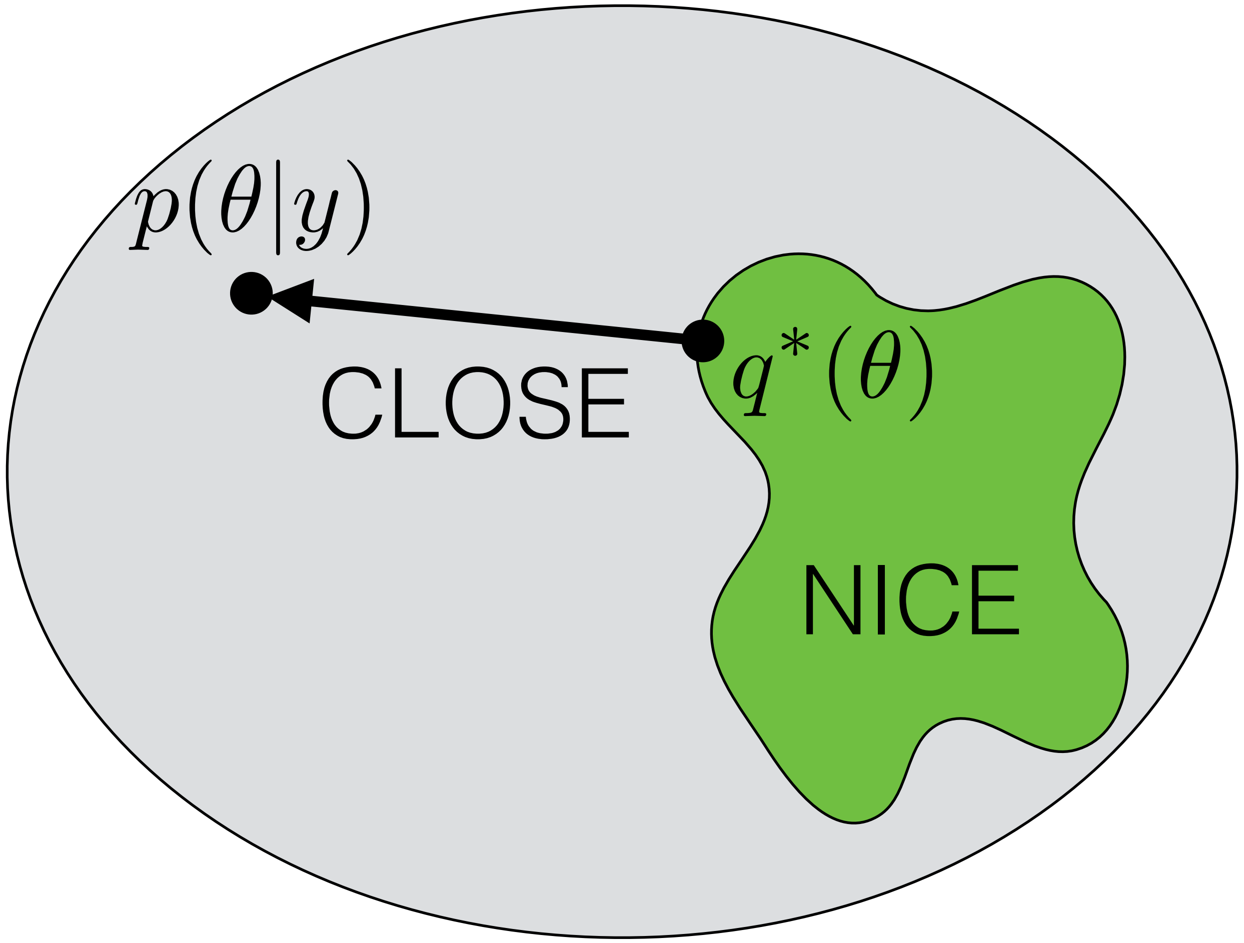
$$= \int q(\theta)\log\frac{q(\theta)p(y)}{p(\theta,y)}d\theta = \log p(y) - \int q(\theta)\log\frac{p(\theta,y)}{q(\theta)}d\theta$$

"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 (in this direction)?

Variational Bayes

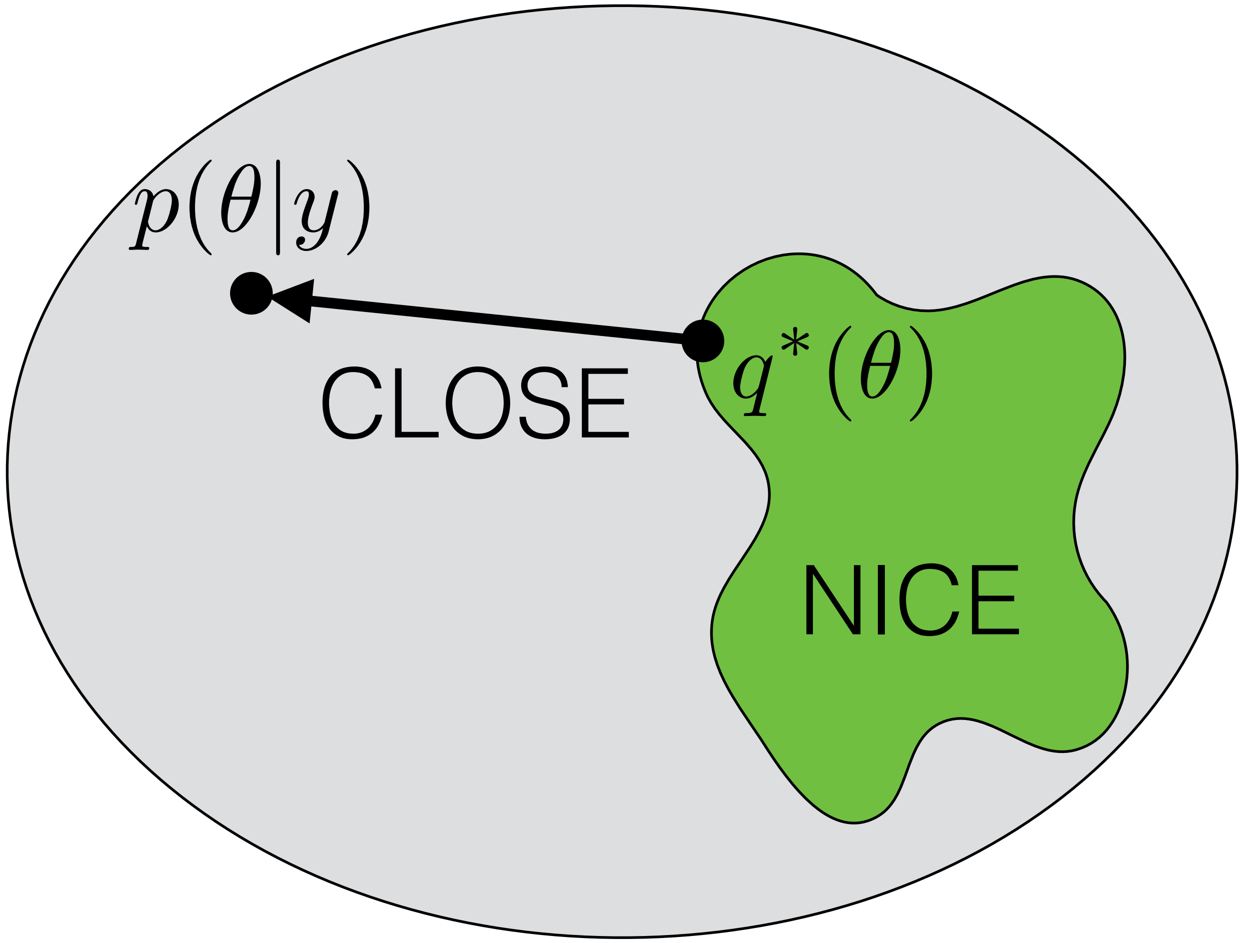
$$q^* = \operatorname{argmin}_{q \in Q} \mathrm{KL}\left(q(\cdot) || p(\cdot|y)\right)$$



Variational Bayes

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

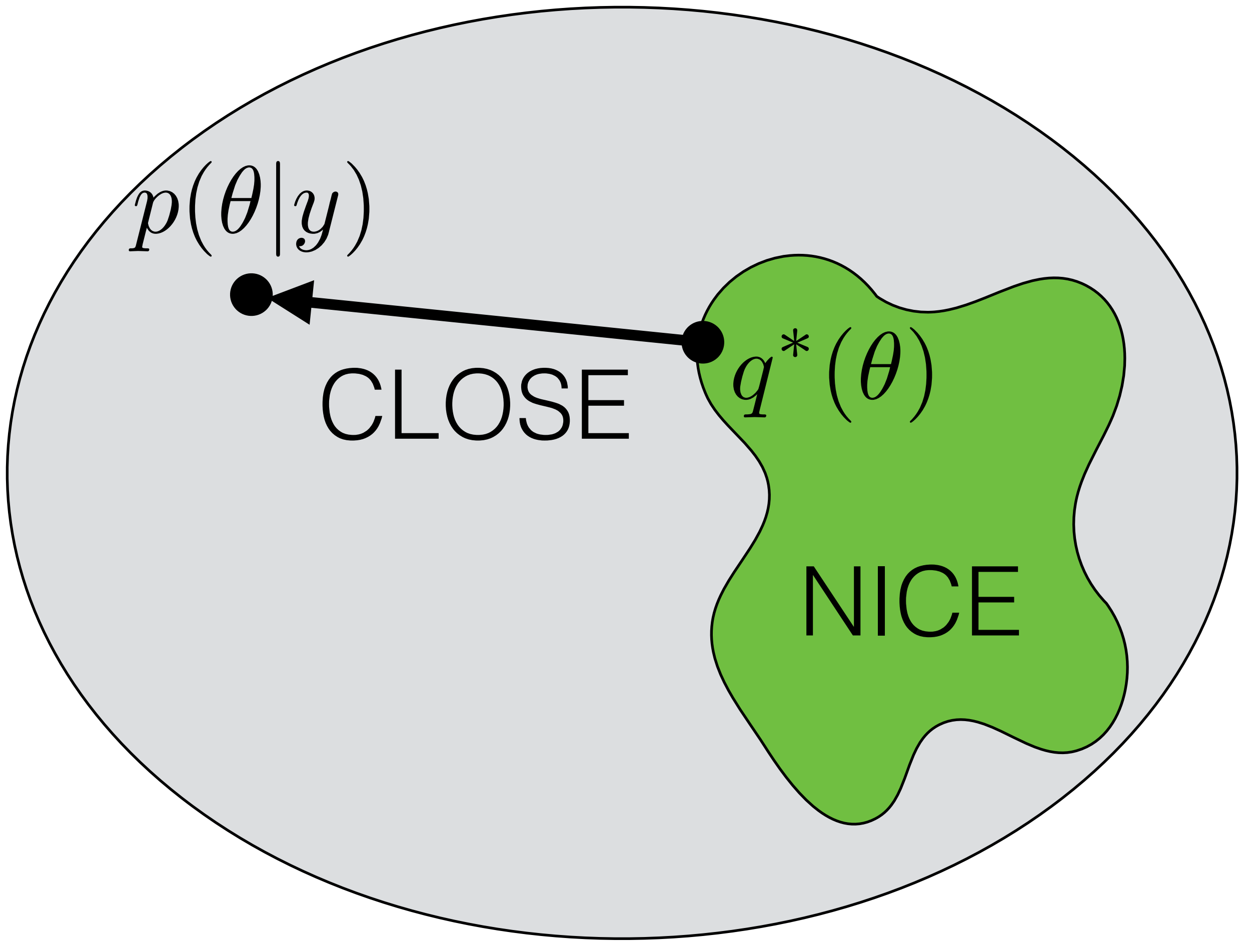
Choose “NICE” distributions



Variational Bayes

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

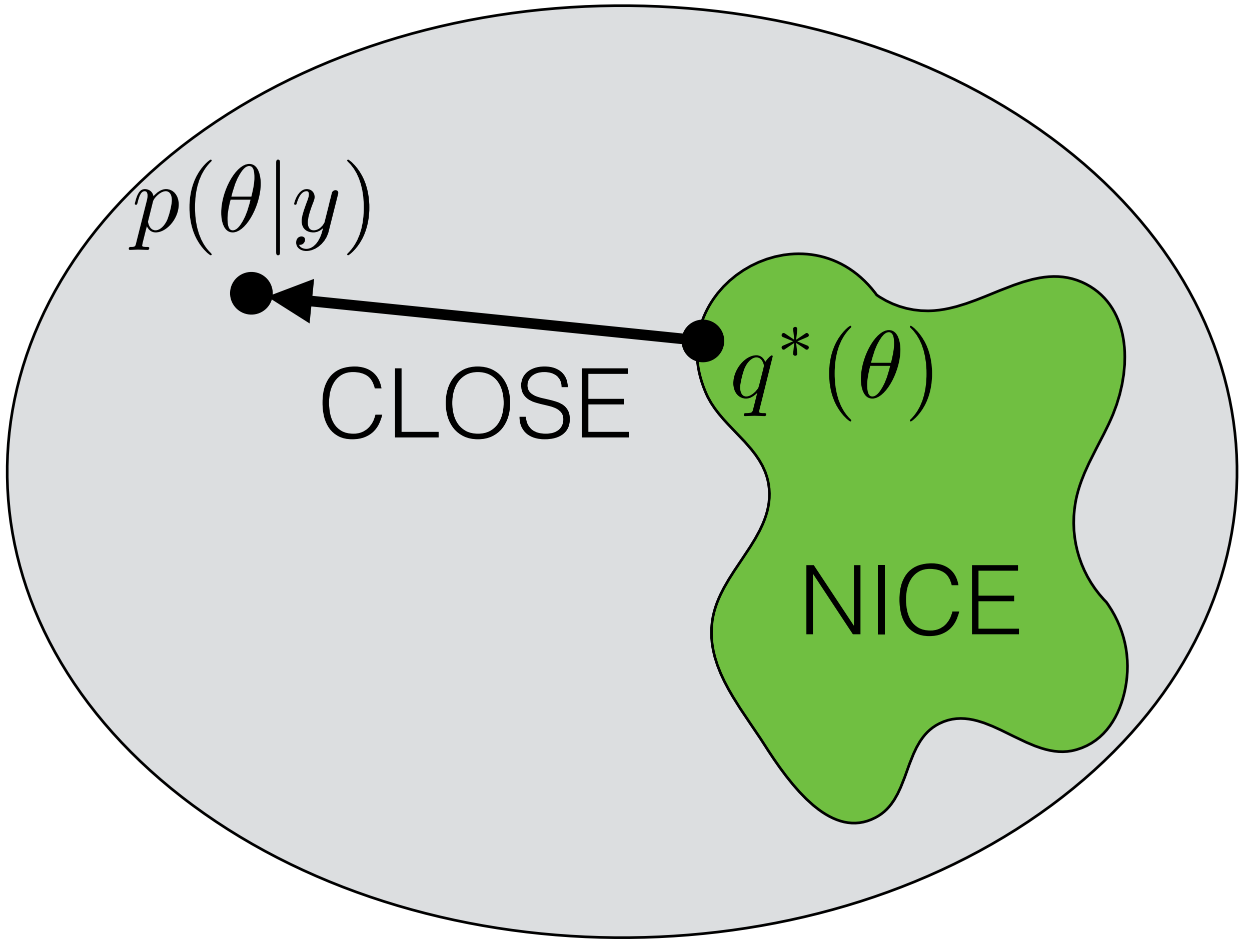
Choose “NICE” distributions



Variational Bayes

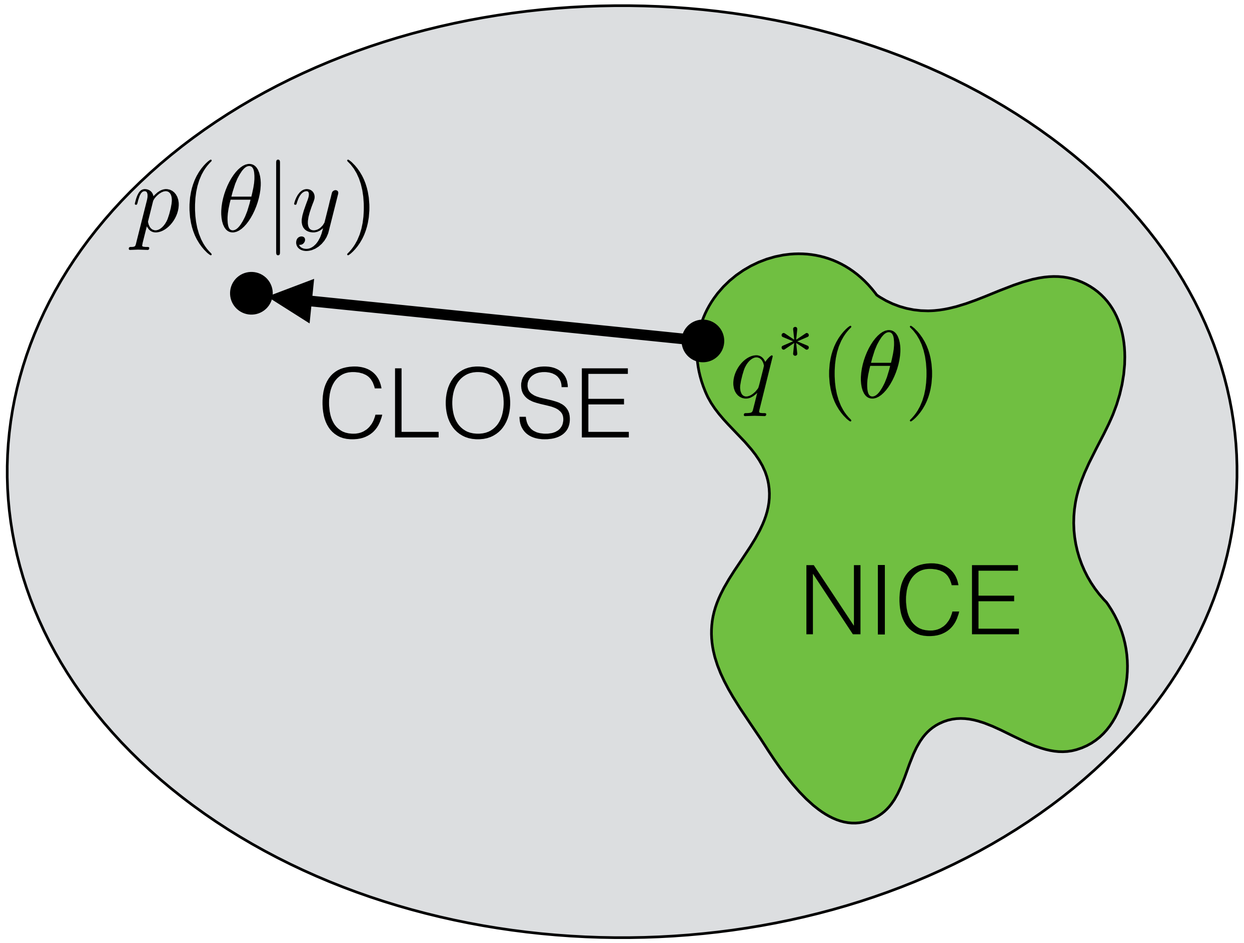
$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$

Choose “NICE” distributions



Variational Bayes

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



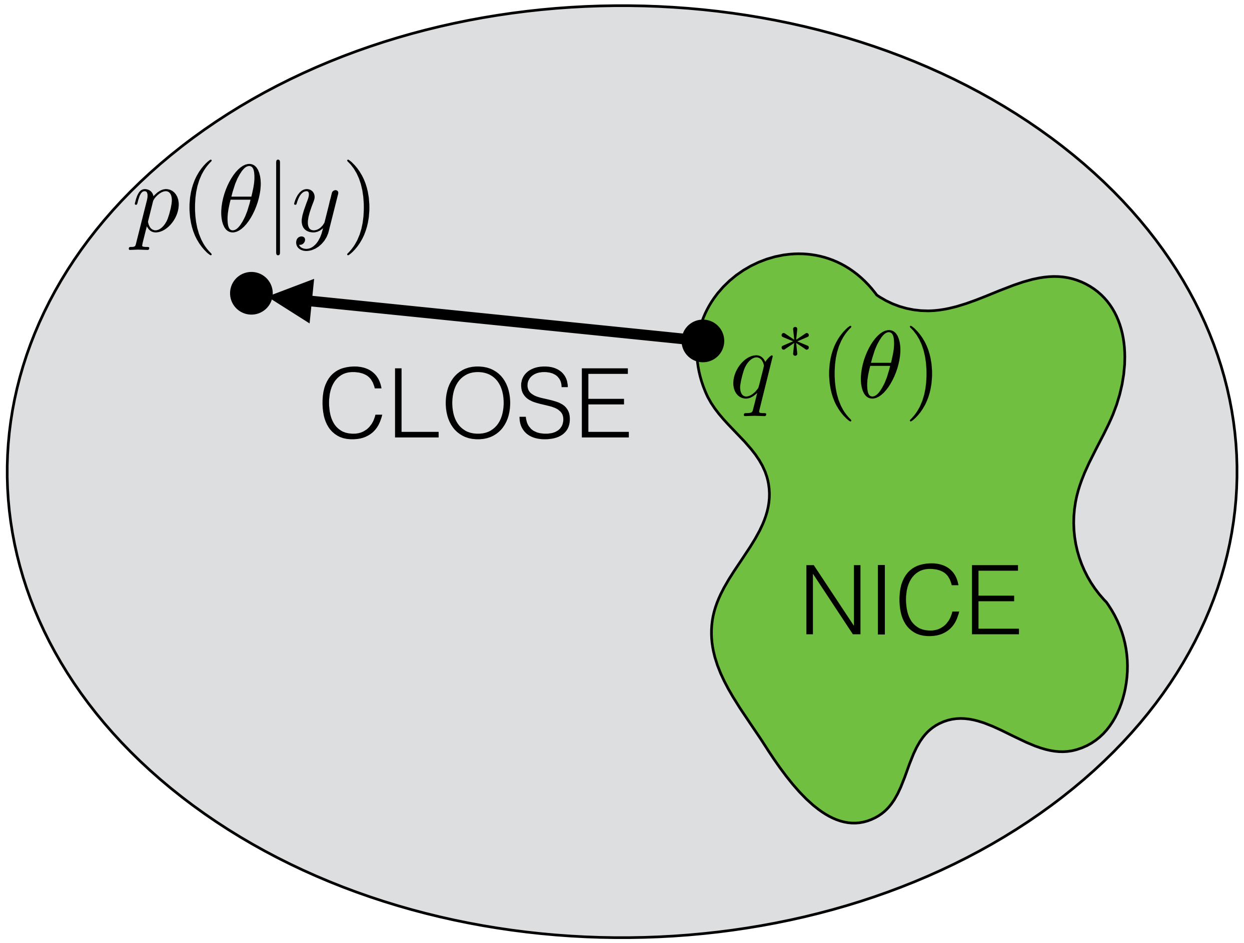
Choose “NICE” distributions

- Mean-field variational Bayes (MFVB)

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

Variational Bayes

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



Choose “NICE” distributions

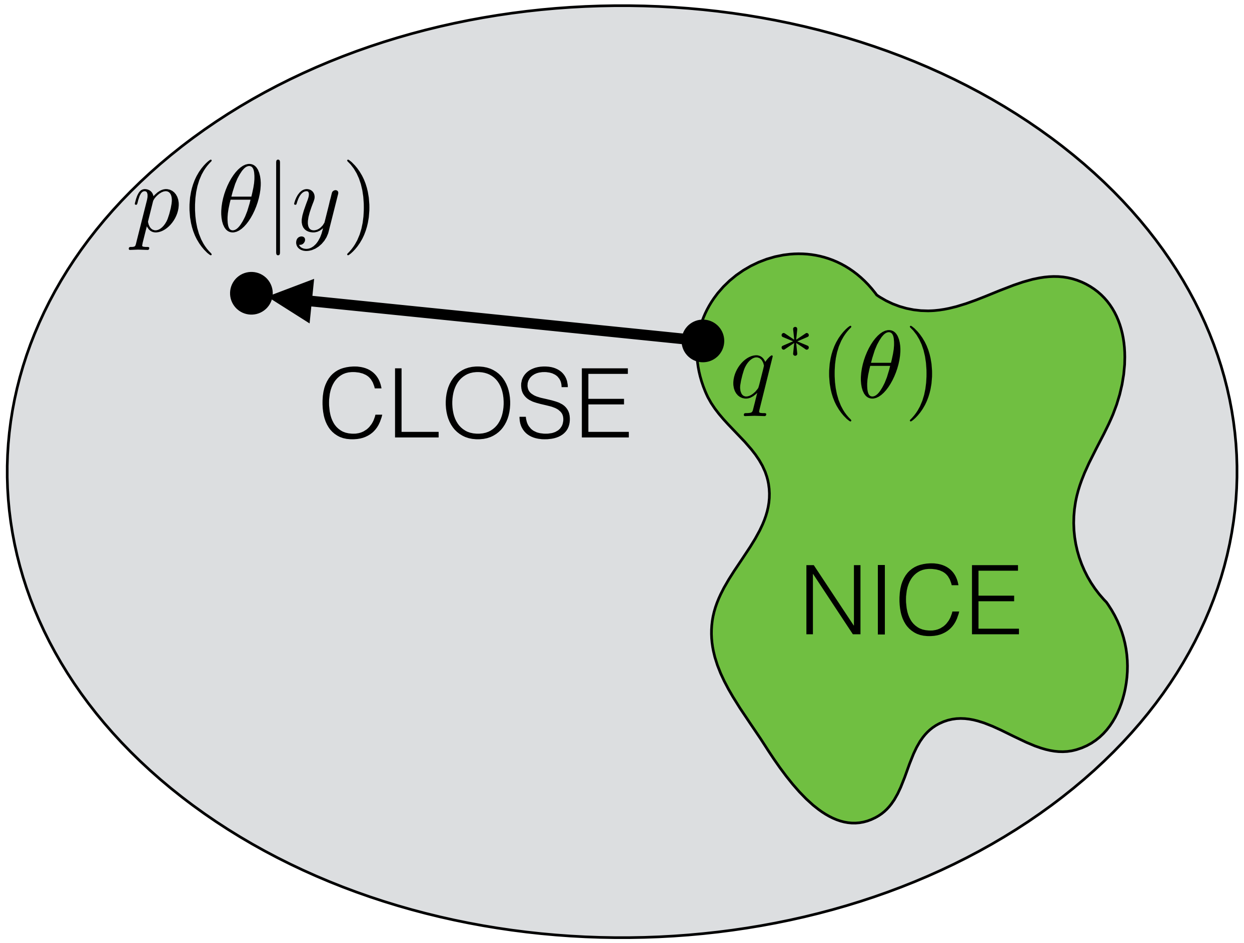
- Mean-field variational Bayes (MFVB)

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

- Often also exponential family

Variational Bayes

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



Choose "NICE" distributions

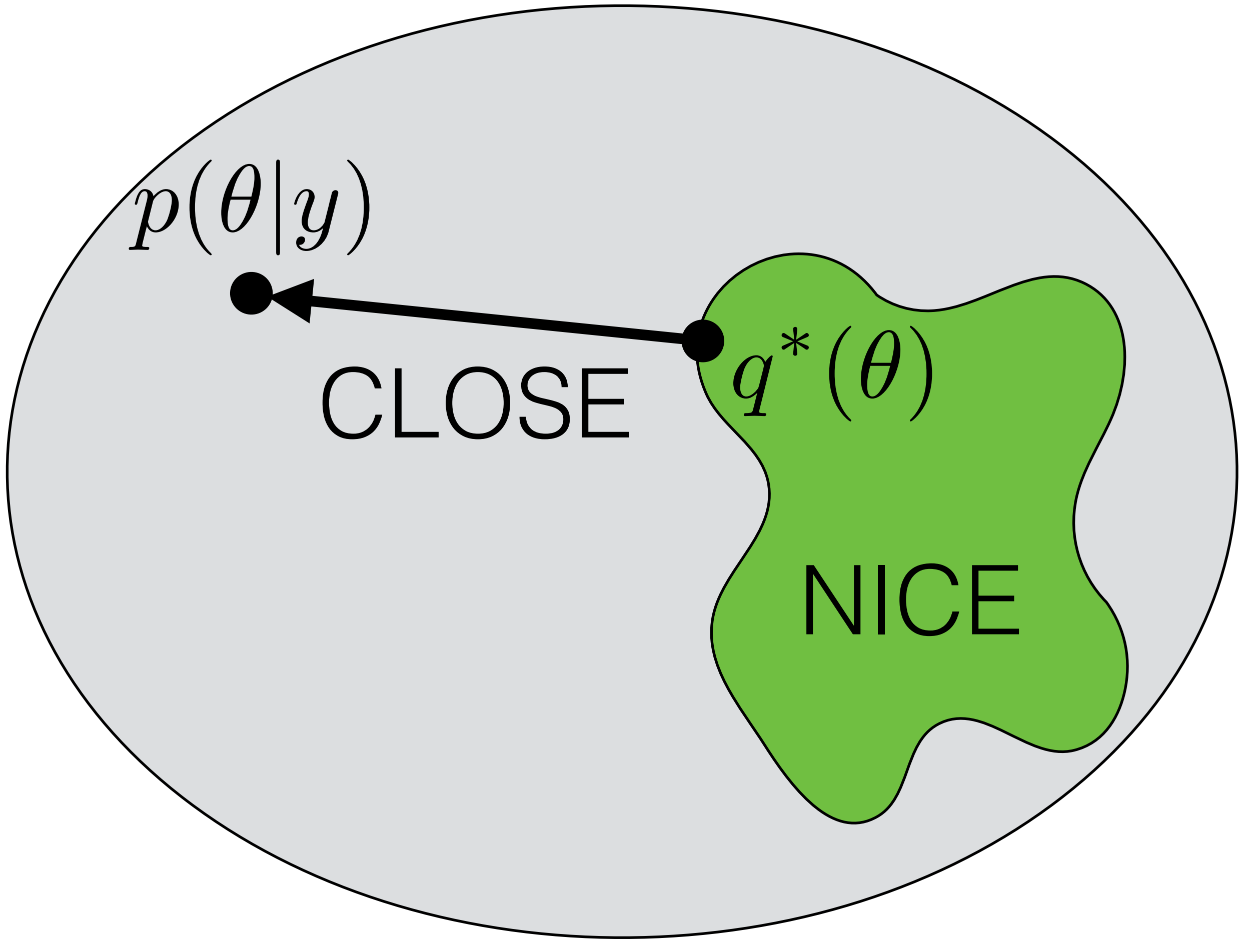
- Mean-field variational Bayes (MFVB)

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

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

Variational Bayes

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

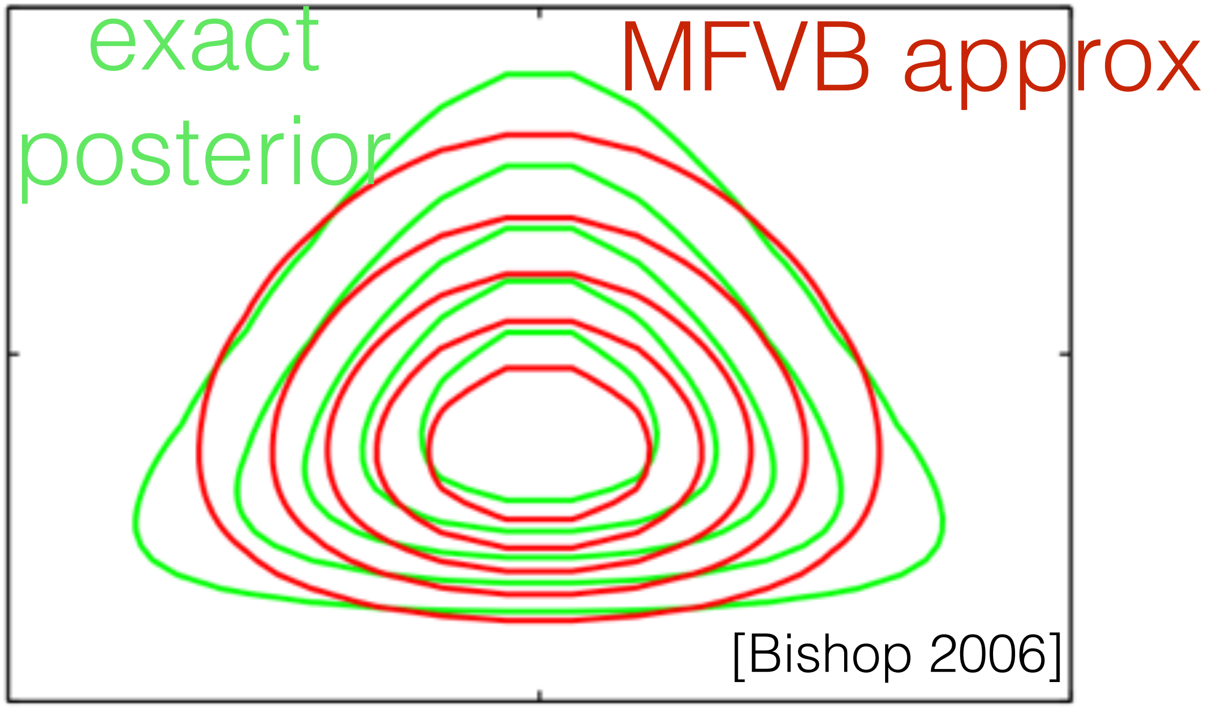


Choose "NICE" distributions

- Mean-field variational Bayes (MFVB)

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

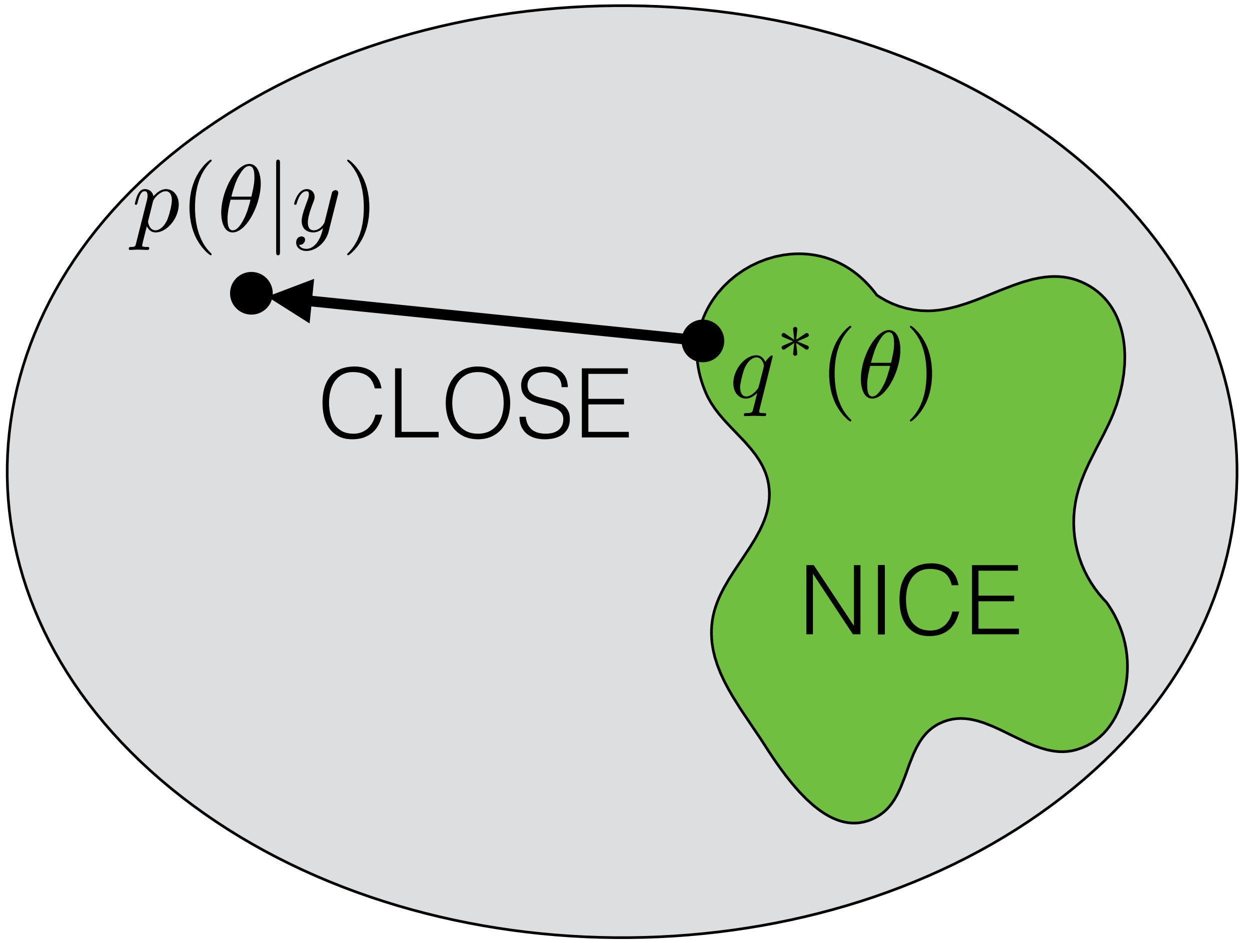
- Often also exponential family
- *Not* a modeling assumption



[Bishop 2006]

Variational Bayes

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

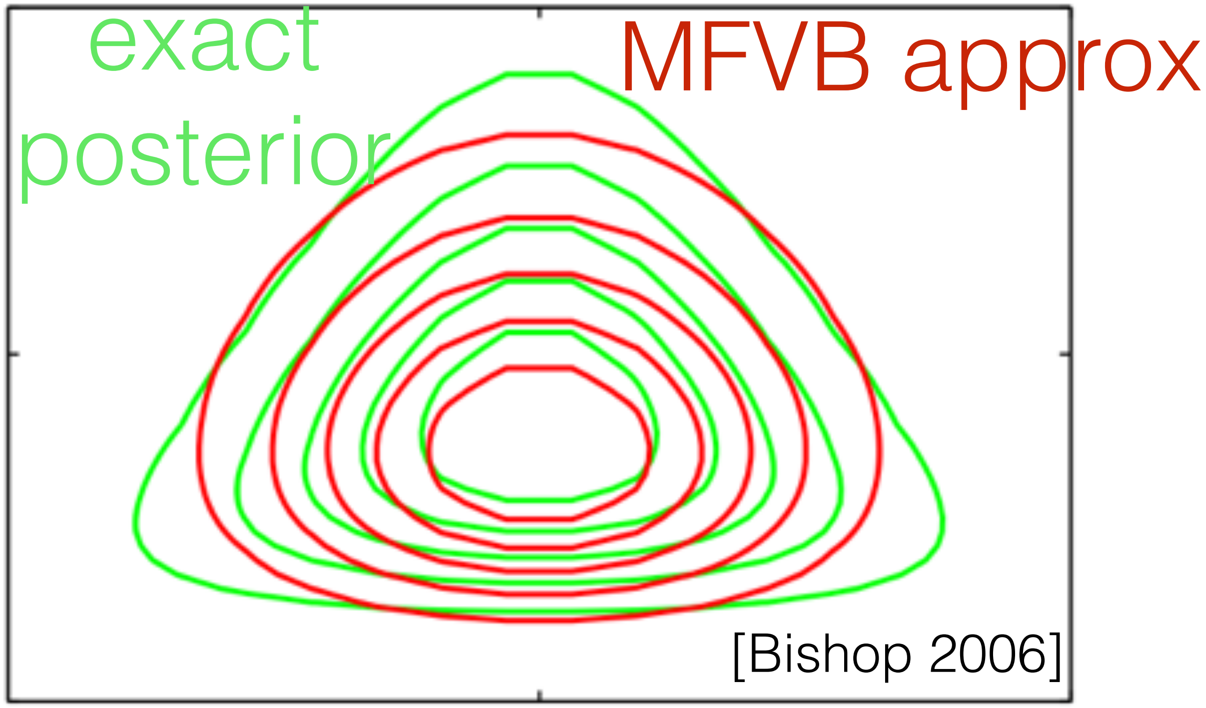


Choose “NICE” distributions

- 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

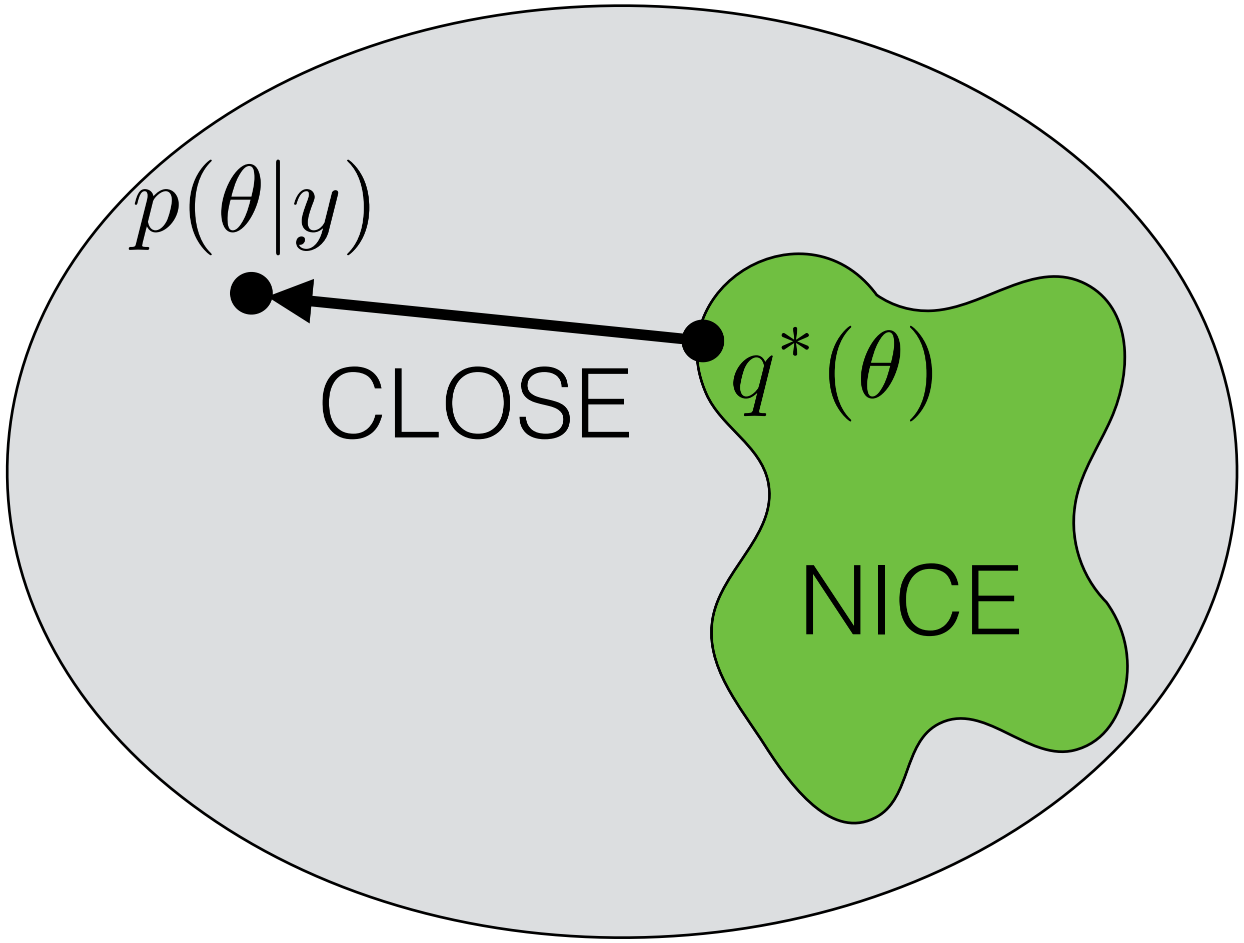


[Bishop 2006]

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

Variational Bayes

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



Choose “NICE” distributions

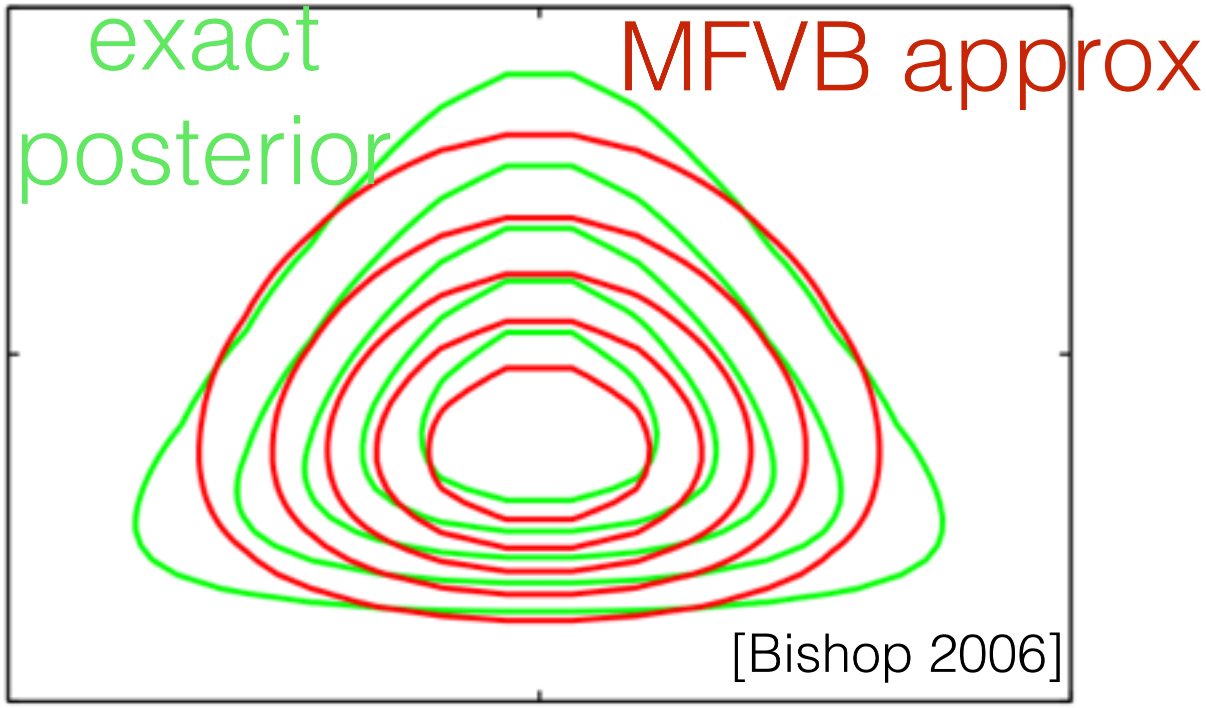
- Mean-field variational Bayes (MFVB)

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

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

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

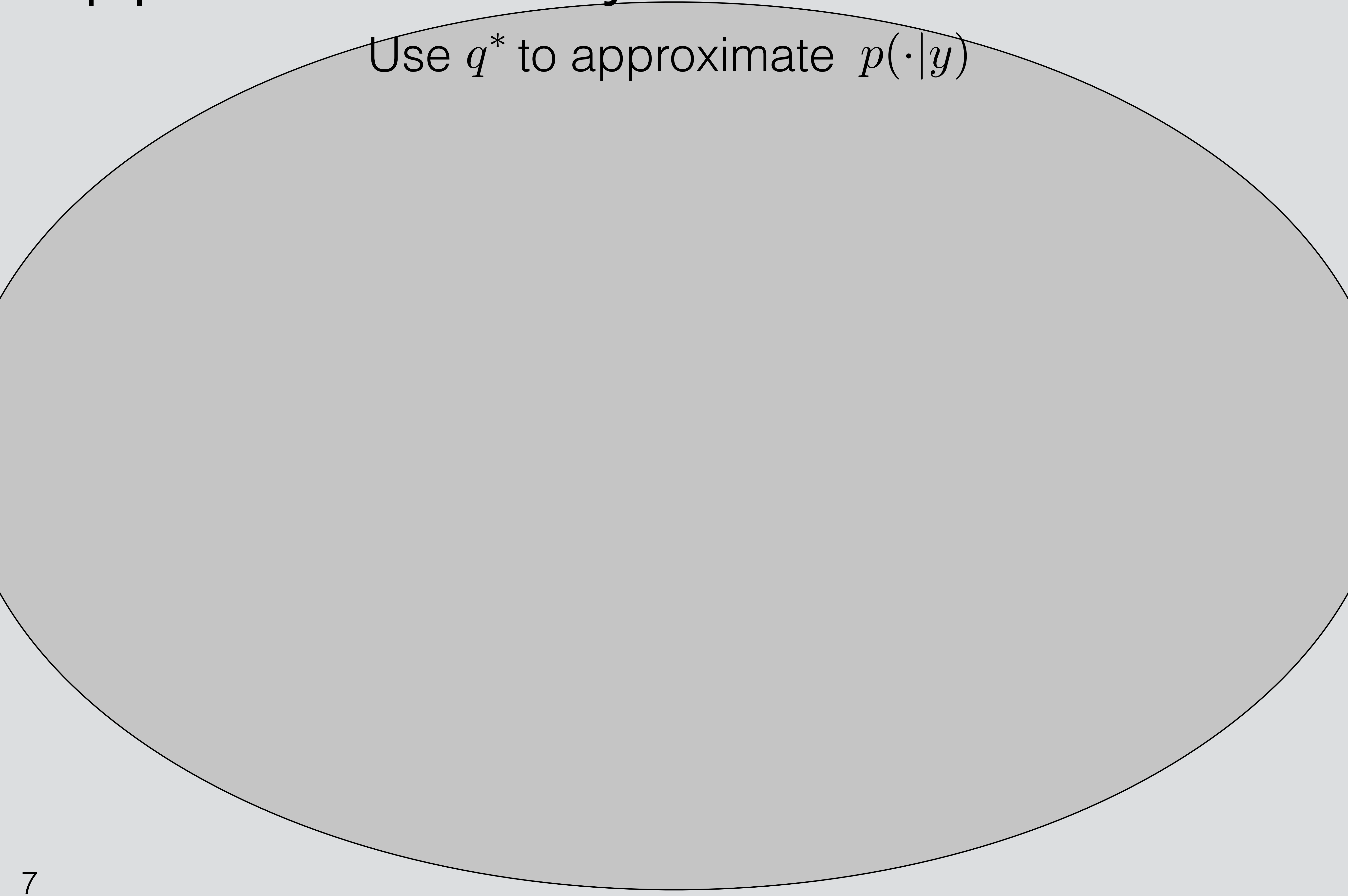
- One option: Coordinate descent in $q_1, \ldots, q_J$



[Bishop 2006]

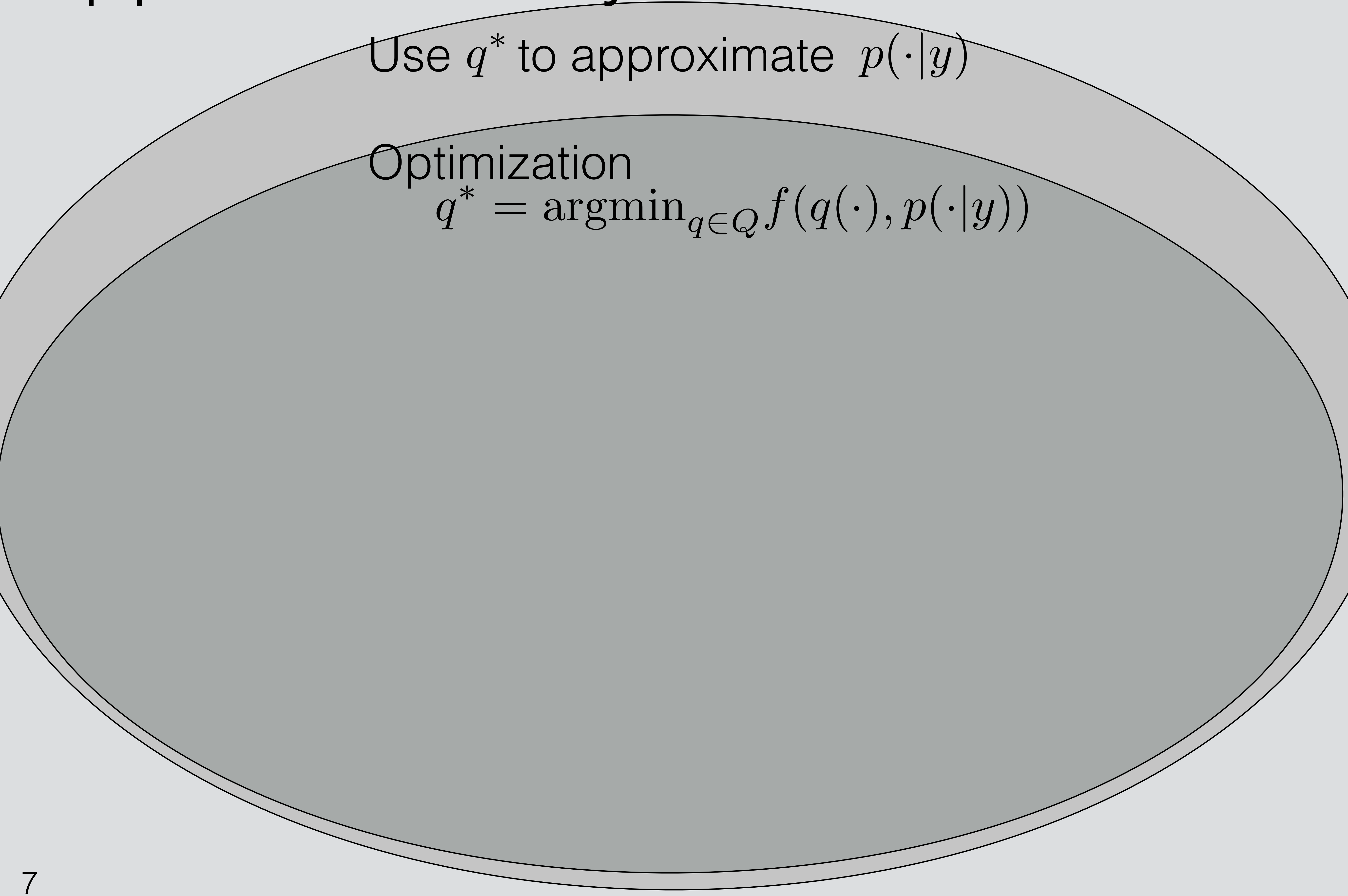
Approximate Bayesian inference

Approximate Bayesian inference



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

Approximate Bayesian inference

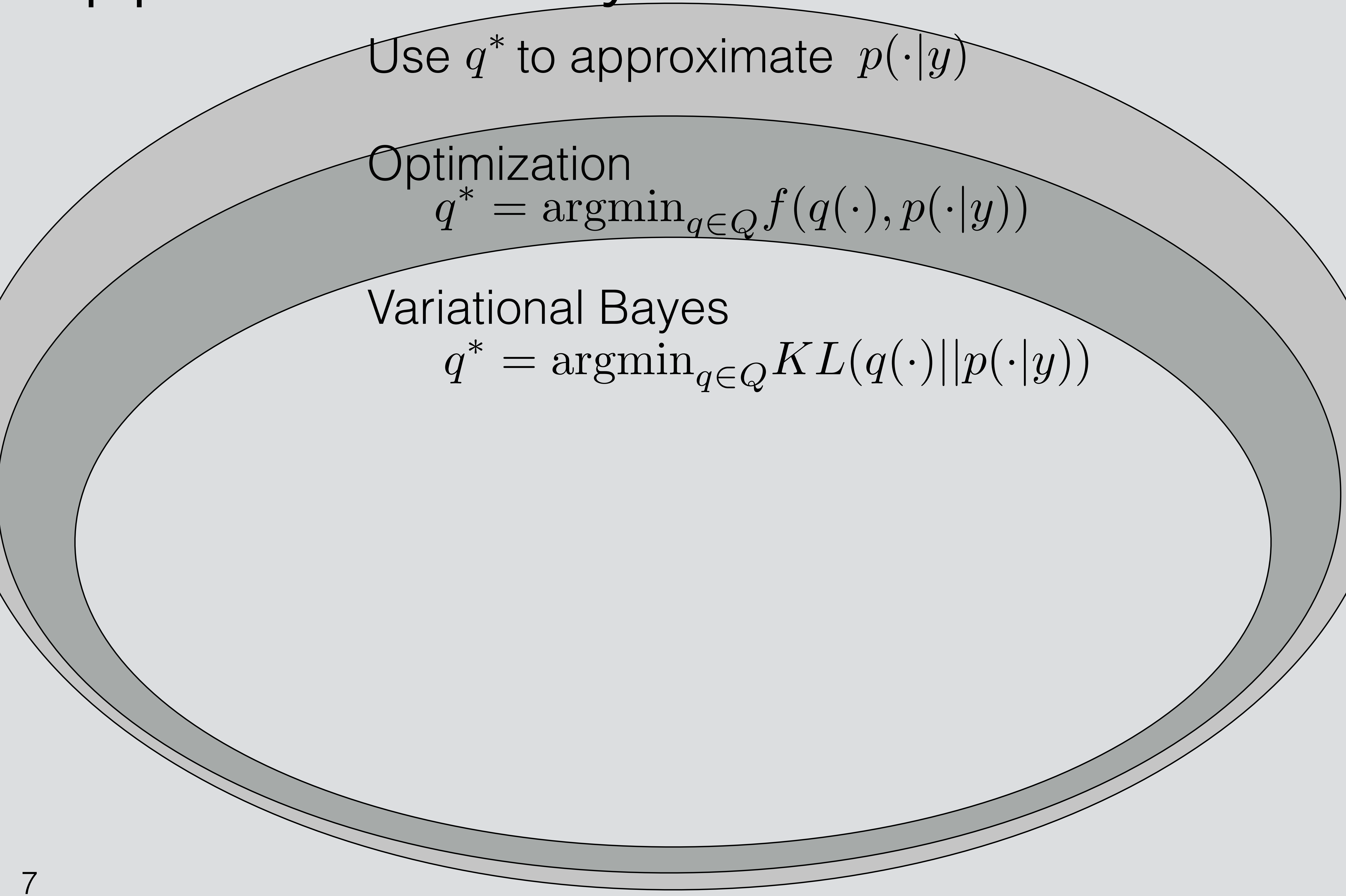


Use q^* to approximate $p(\cdot|y)$

Optimization

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

Approximate Bayesian inference



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

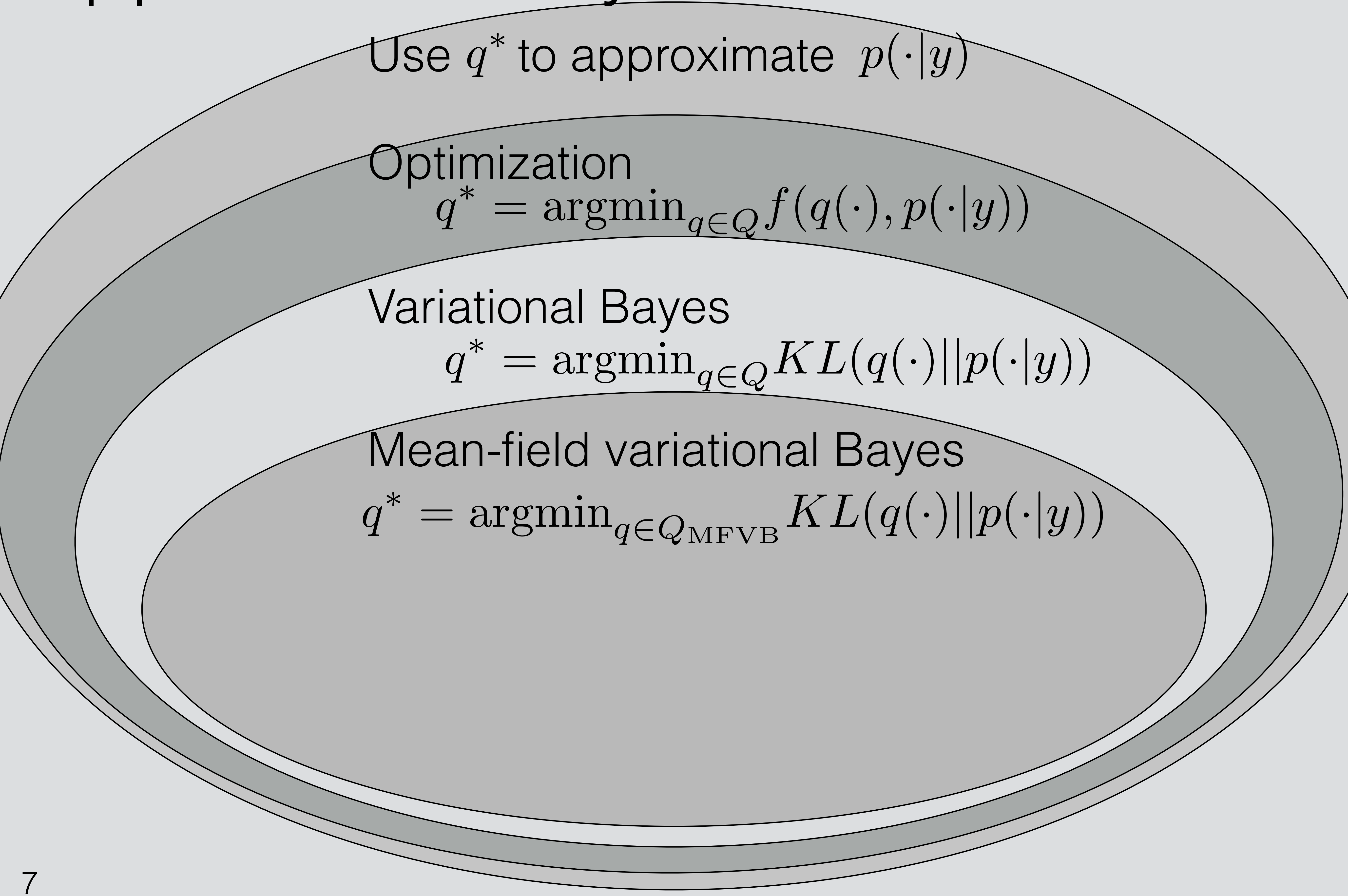
Optimization

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

Variational Bayes

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

Approximate Bayesian inference



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

Optimization

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

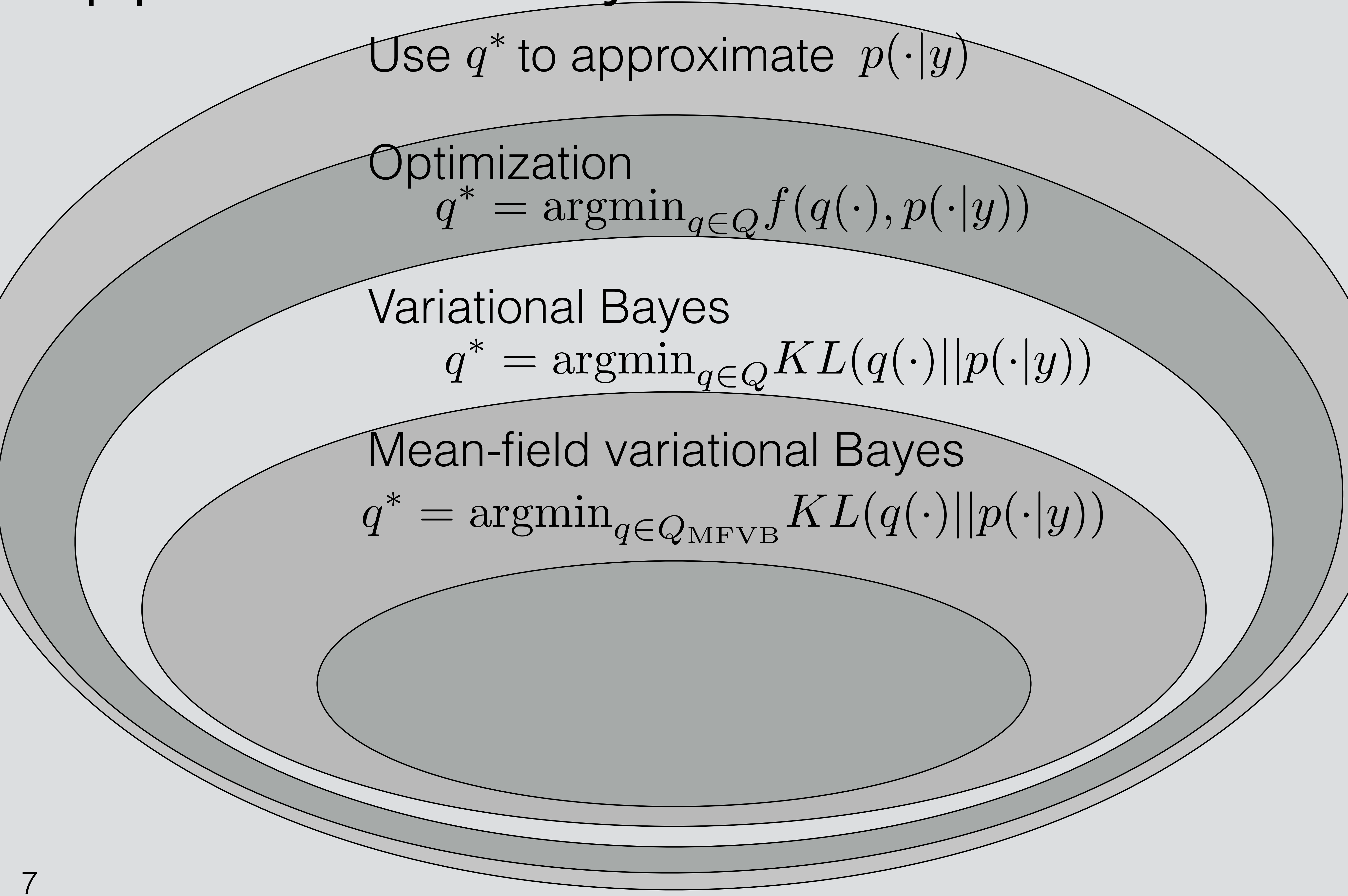
Variational Bayes

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

Mean-field variational Bayes

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

Approximate Bayesian inference



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

Optimization

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

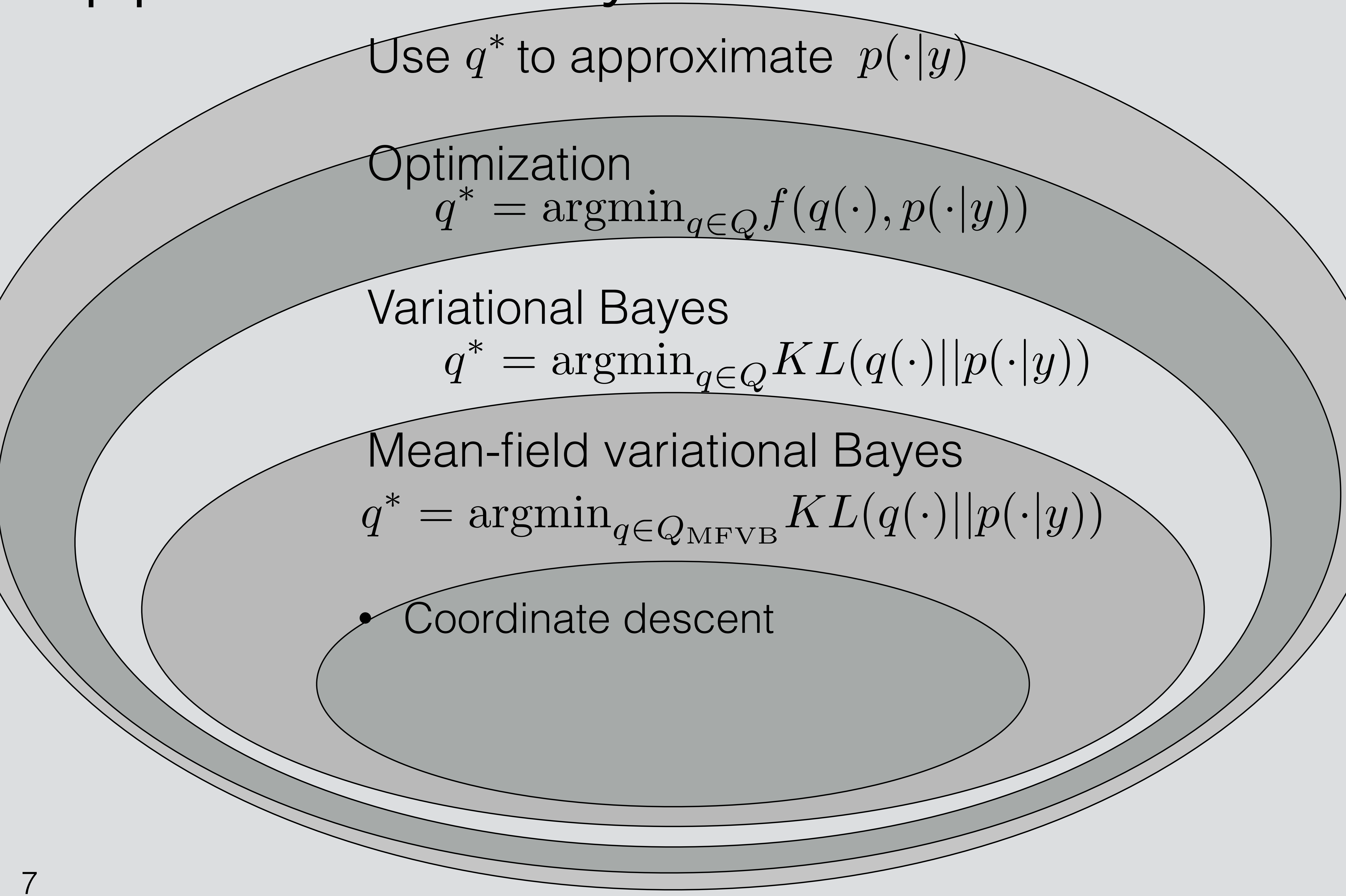
Variational Bayes

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

Mean-field variational Bayes

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

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

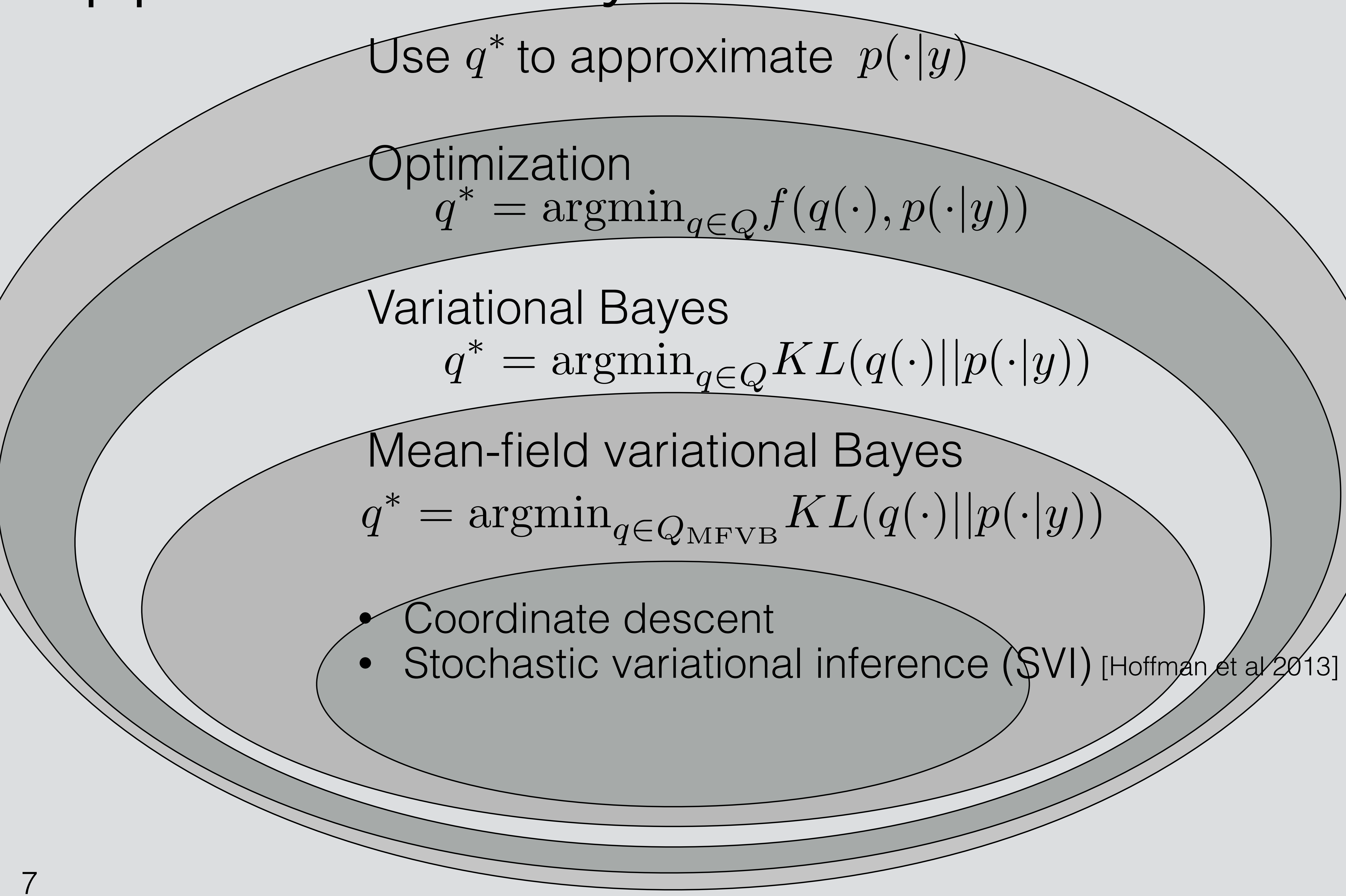
$$q^* = \text{argmin}_{q \in Q} KL(q(\cdot)||p(\cdot|y))$$

Mean-field variational Bayes

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

- Coordinate descent

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

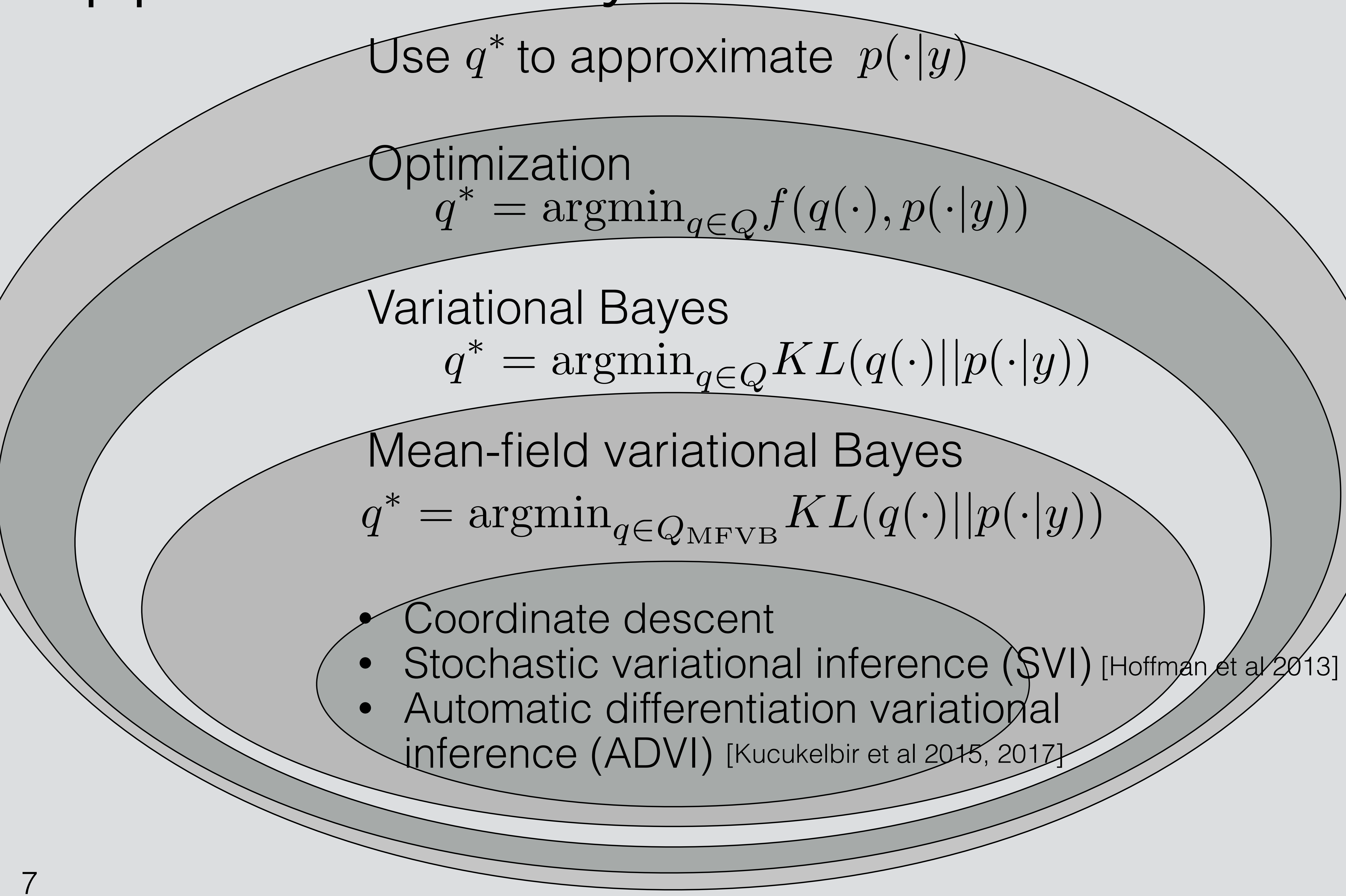
$$q^* = \text{argmin}_{q \in Q} KL(q(\cdot)||p(\cdot|y))$$

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

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

Variational Bayes

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

Mean-field variational Bayes

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

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

Roadmap

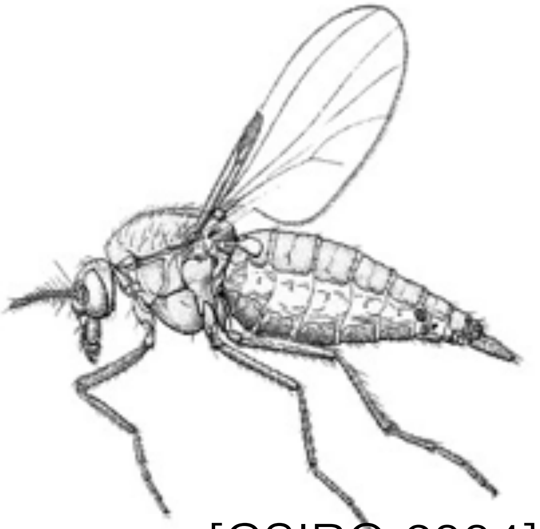
- Bayes & Approximate Bayes review
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use MFVB?
- When can we trust MFVB?
- Where do we go from here?

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:

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

[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:

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

[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:

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

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

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

[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:

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

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

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

- Exercise: check $p(\mu, \sigma^2|y) \neq f_1(\mu, y) f_2(\sigma^2, y)$

[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:

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

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

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

[CSIRO 2004]

- Exercise: check $p(\mu, \sigma^2|y) \neq f_1(\mu, y) f_2(\sigma^2, y)$
- MFVB approximation:

$$q^*(\mu, \sigma^2) = q^*_\mu(\mu) q^*_{\sigma^2}(\sigma^2) = \mathrm{argmin}_{q \in Q_{\mathrm{MFVB}}} KL(q(\cdot)||p(\cdot|y))$$

[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 variance $\theta = (\mu, \sigma^2)$
- Model:

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

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

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

[CSIRO 2004]

- Exercise: check $p(\mu, \sigma^2|y) \neq f_1(\mu, y) f_2(\sigma^2, y)$
- MFVB approximation:

$$q^*(\mu, \sigma^2) = q^*_\mu(\mu) q^*_{\sigma^2}(\sigma^2) = \text{argmin}_{q \in Q_{\text{MFVB}}} KL(q(\cdot)||p(\cdot|y))$$

- 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 variance $\theta = (\mu, \sigma^2)$
- Model:

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

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

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

[CSIRO 2004]

- Exercise: check $p(\mu, \sigma^2|y) \neq f_1(\mu, y) f_2(\sigma^2, y)$
- MFVB approximation:

$$q^*(\mu, \sigma^2) = q^*_\mu(\mu) q^*_{\sigma^2}(\sigma^2) = \text{argmin}_{q \in Q_{\text{MFVB}}} KL(q(\cdot)||p(\cdot|y))$$

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

$$q^*(\mu) = N(\mu|m_\mu, \rho^2_\mu) \qquad q^*((\sigma^2)^{-1}) = \text{Gamma}((\sigma^2)^{-1}|a_\sigma, b_\sigma)$$

[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:

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

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

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

[CSIRO 2004]

- Exercise: check $p(\mu, \sigma^2|y) \neq f_1(\mu, y) f_2(\sigma^2, y)$
- MFVB approximation:

$$q^*(\mu, \sigma^2) = q^*_\mu(\mu) q^*_{\sigma^2}(\sigma^2) = \mathrm{argmin}_{q \in Q_{\mathrm{MFVB}}} KL(q(\cdot)||p(\cdot|y))$$

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

$$q^*(\mu) = N(\mu|m_\mu, \rho^2_\mu) \qquad q^*((\sigma^2)^{-1}) = \mathrm{Gamma}((\sigma^2)^{-1}|a_\sigma, b_\sigma)$$

"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 variance $\theta = (\mu, \sigma^2)$
- Model:

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

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

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

[CSIRO 2004]

- Exercise: check $p(\mu, \sigma^2|y) \neq f_1(\mu, y) f_2(\sigma^2, y)$
- MFVB approximation:

$$q^*(\mu, \sigma^2) = q^*_\mu(\mu) q^*_{\sigma^2}(\sigma^2) = \text{argmin}_{q \in Q_{\text{MFVB}}} KL(q(\cdot)||p(\cdot|y))$$

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

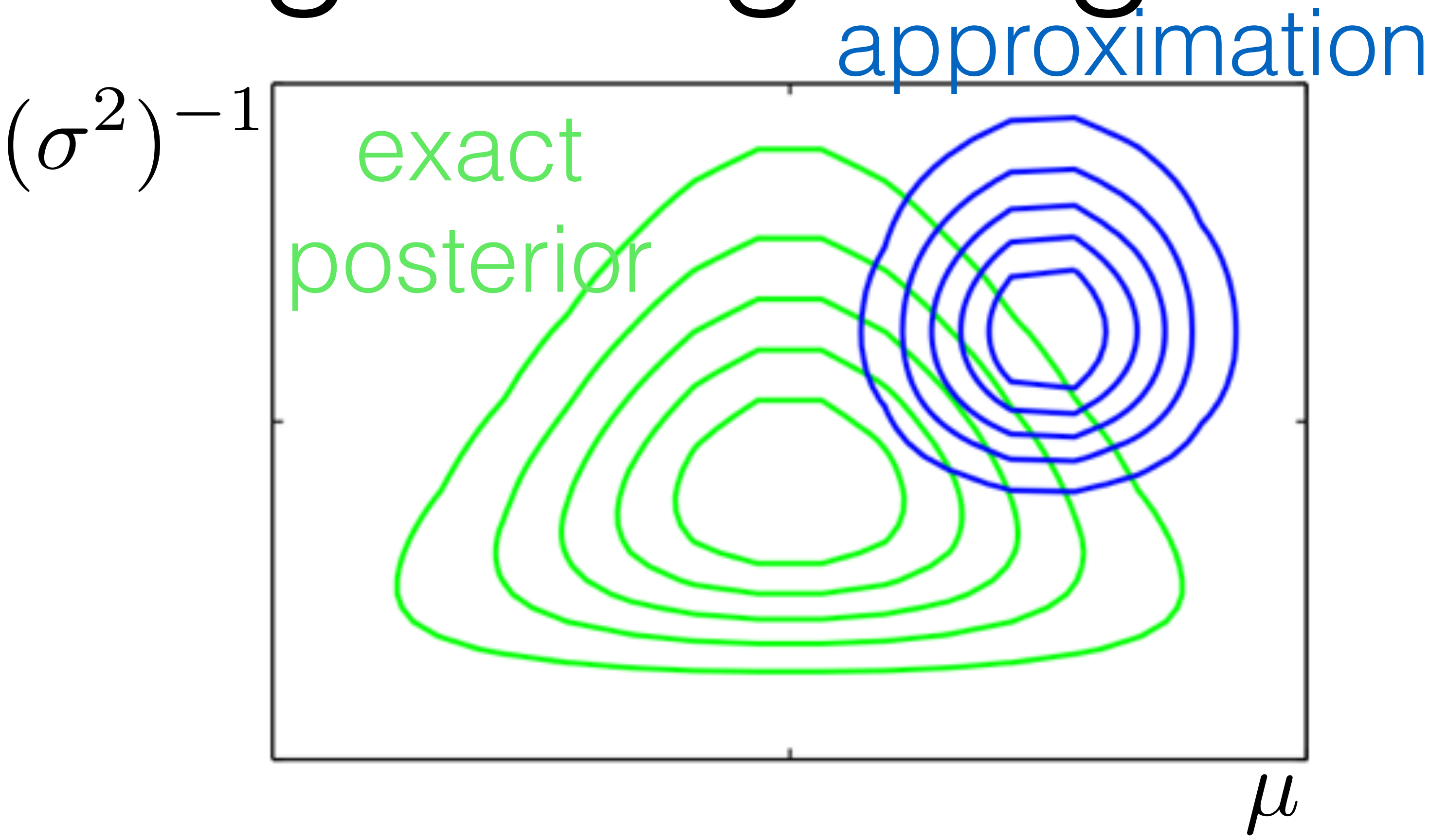
$$q^*(\mu) = N(\mu|m_\mu, \rho^2_\mu) \qquad q^*((\sigma^2)^{-1}) = \text{Gamma}((\sigma^2)^{-1}|a_\sigma, b_\sigma)$$

  - Iterate: $(m_\mu, \rho^2_\mu) = f(a_\sigma, b_\sigma)$
    $(a_\sigma, b_\sigma) = g(m_\mu, \rho^2_\mu)$

"variational parameters"

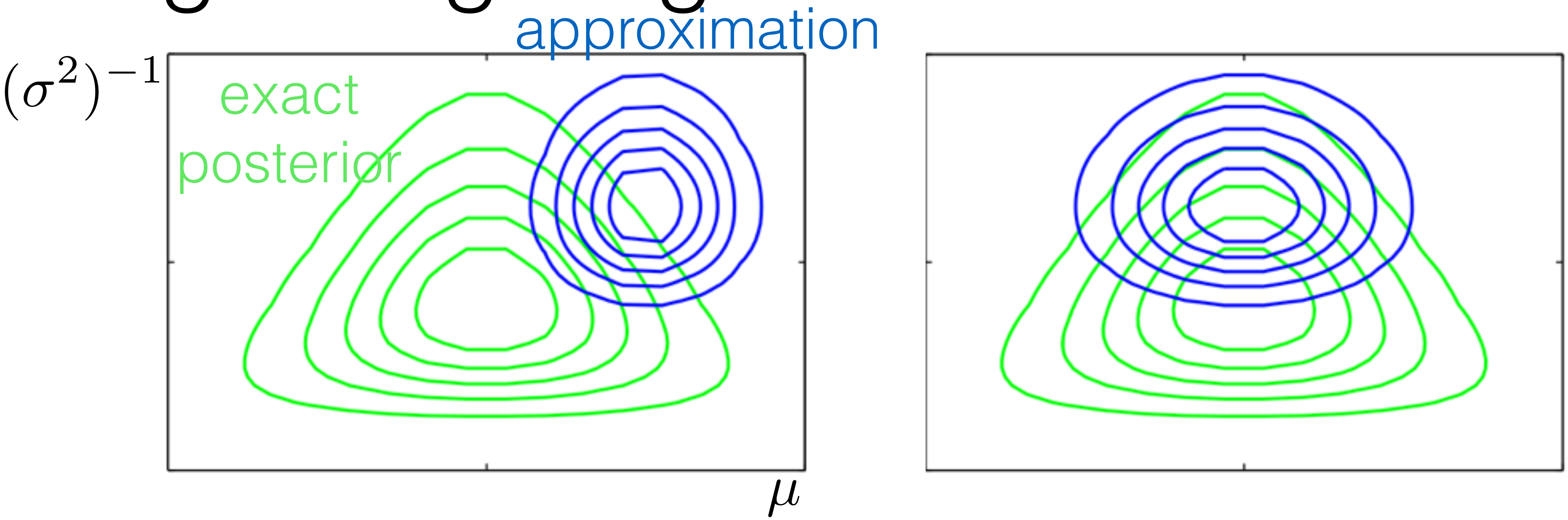
[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]

Midge wing length



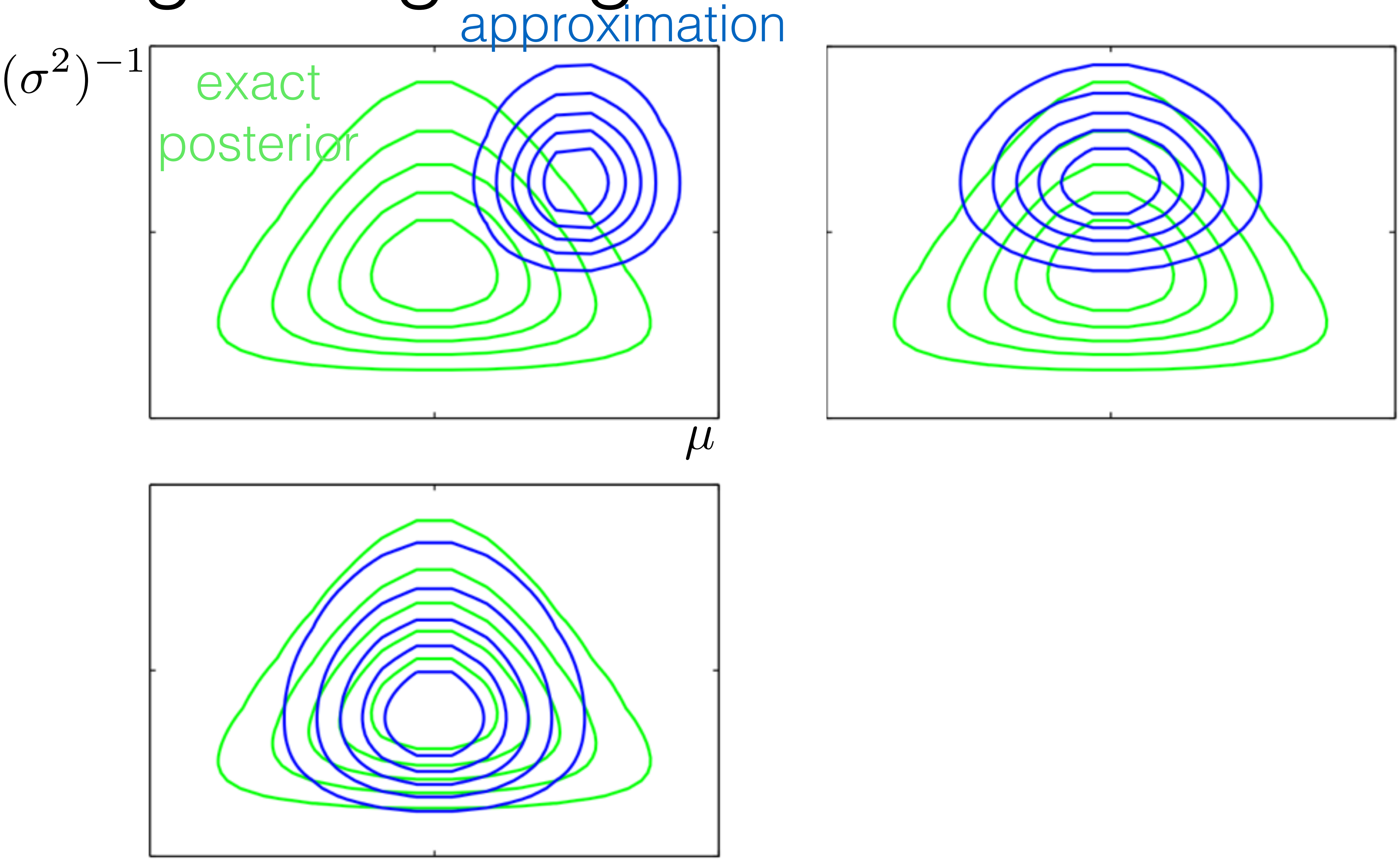
[Bishop 2006]

Midge wing length



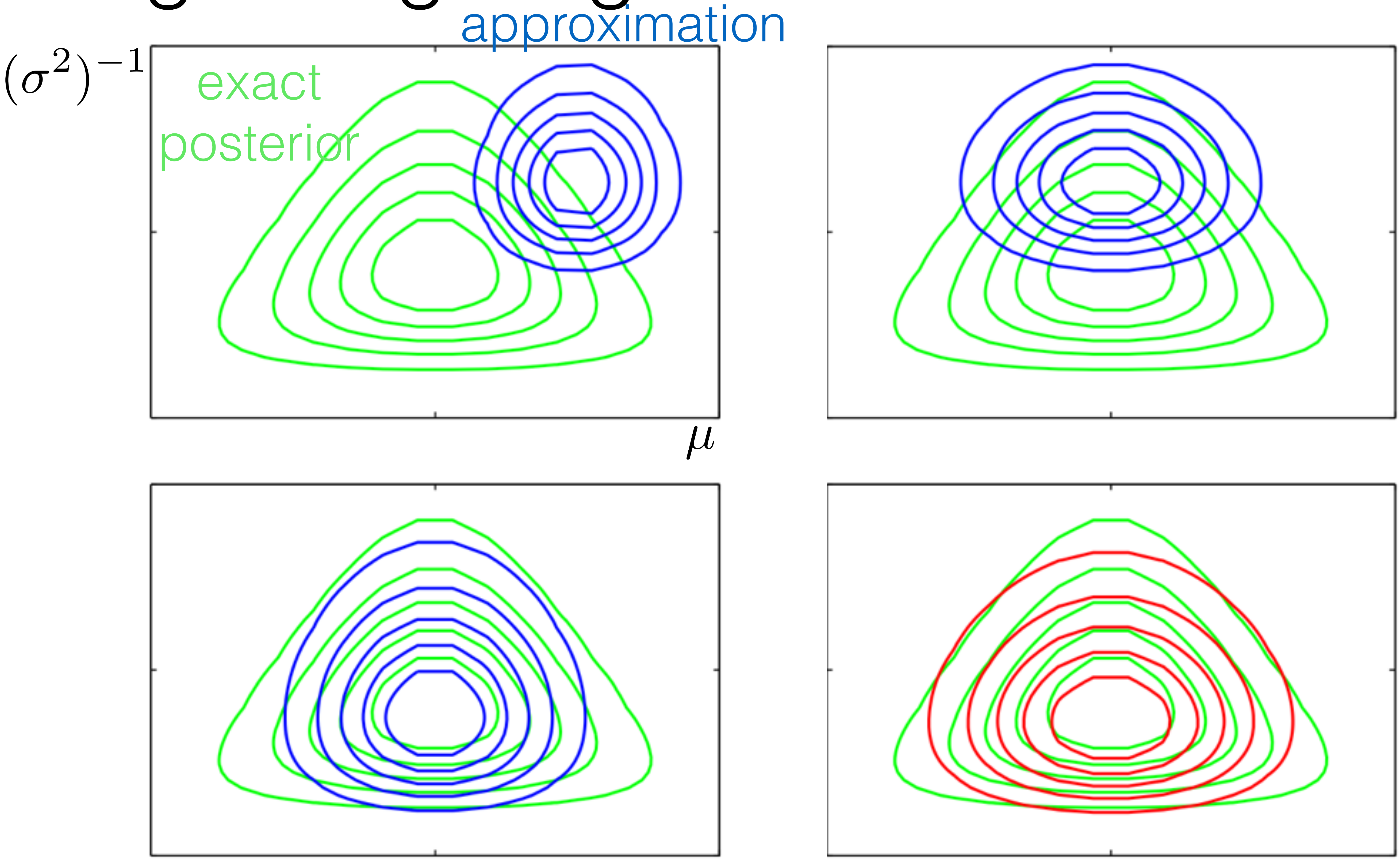
[Bishop 2006]

Midge wing length



[Bishop 2006]

Midge wing length

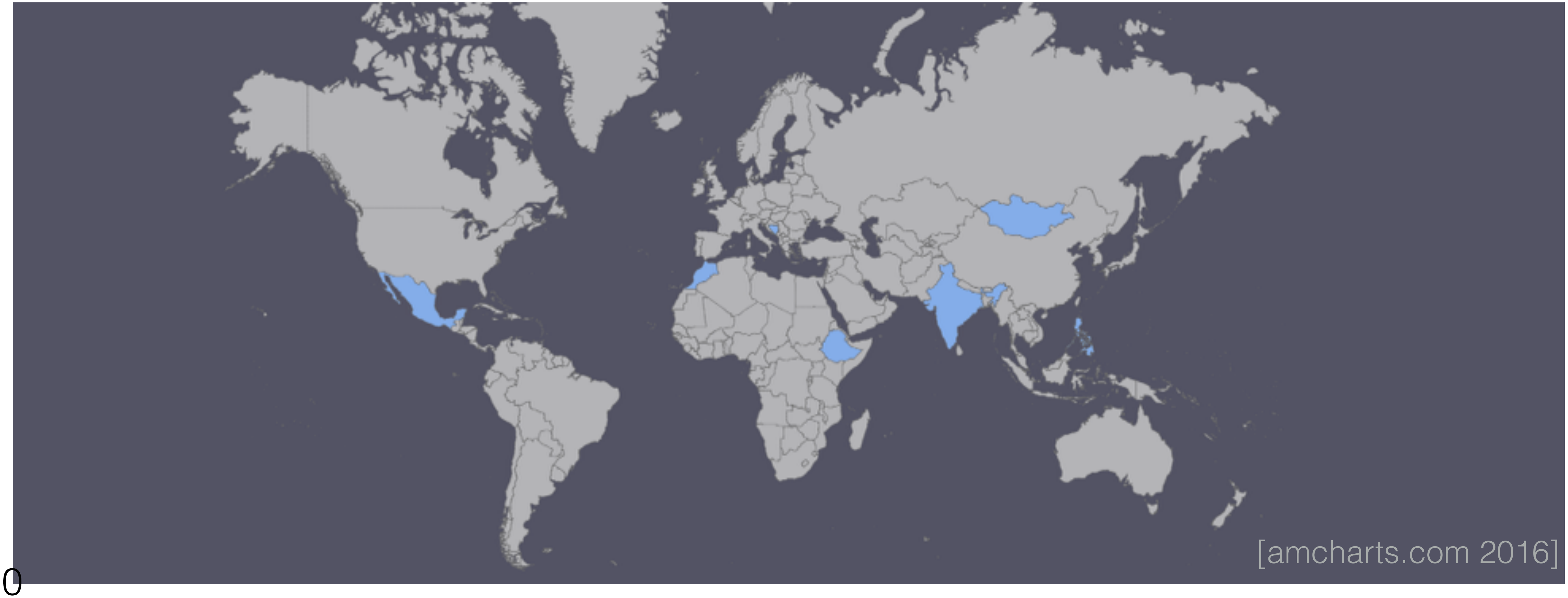


approximation

exact posterior

[Bishop 2006]

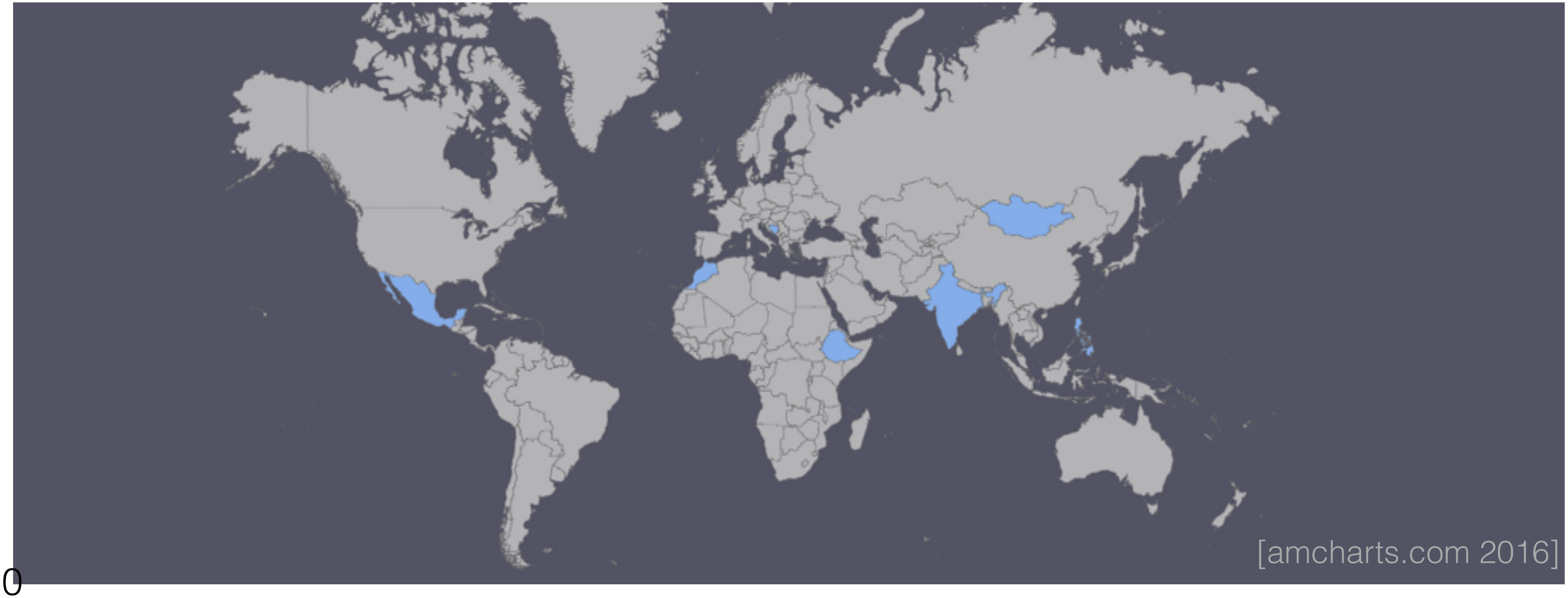
Microcredit Experiment



[amcharts.com 2016]

Microcredit Experiment

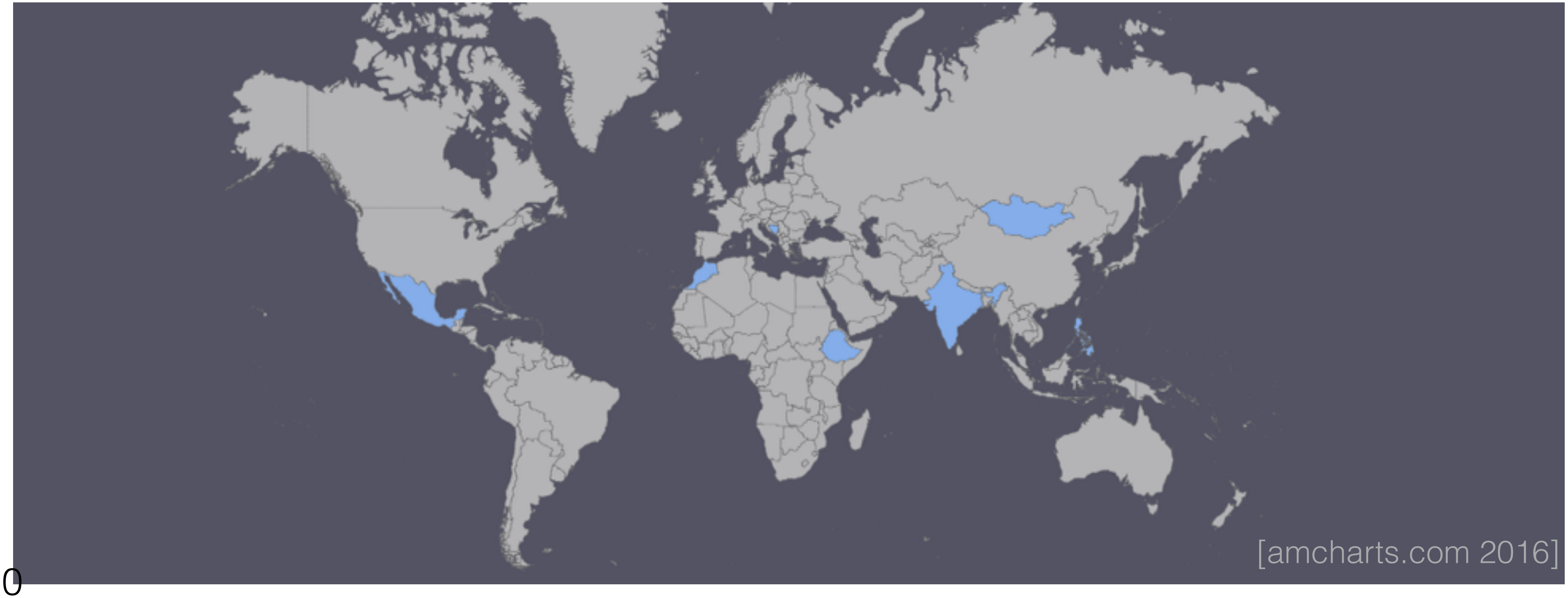
- Simplified from Meager (2018a)
- K = 7 microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)



[amcharts.com 2016]

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)



[amcharts.com 2016]

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

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

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:

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:

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 *k*th site (~900 to ~17K)
- Profit of *n*th business at *k*th 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

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

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 *k*th site (~900 to ~17K)
- Profit of *n*th business at *k*th site:

1 if microcredit

profit

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

- Priors and hyperpriors:

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

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

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:

1 if microcredit

profit

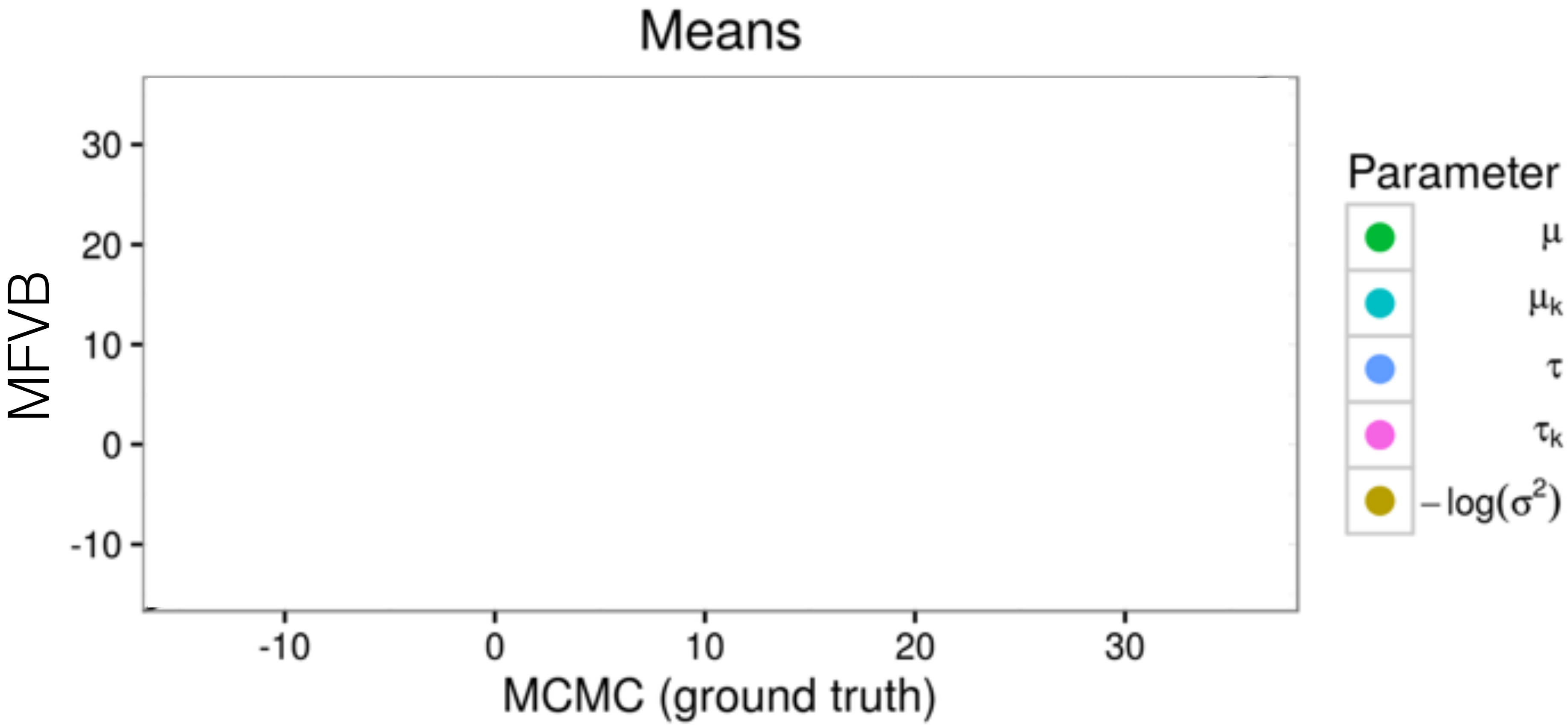
$$y_{kn} \overset{indep}{\sim} \mathcal{N}(\mu_k + T_{kn}\tau_k, \sigma_k^2)$$

- Priors and hyperpriors:

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

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

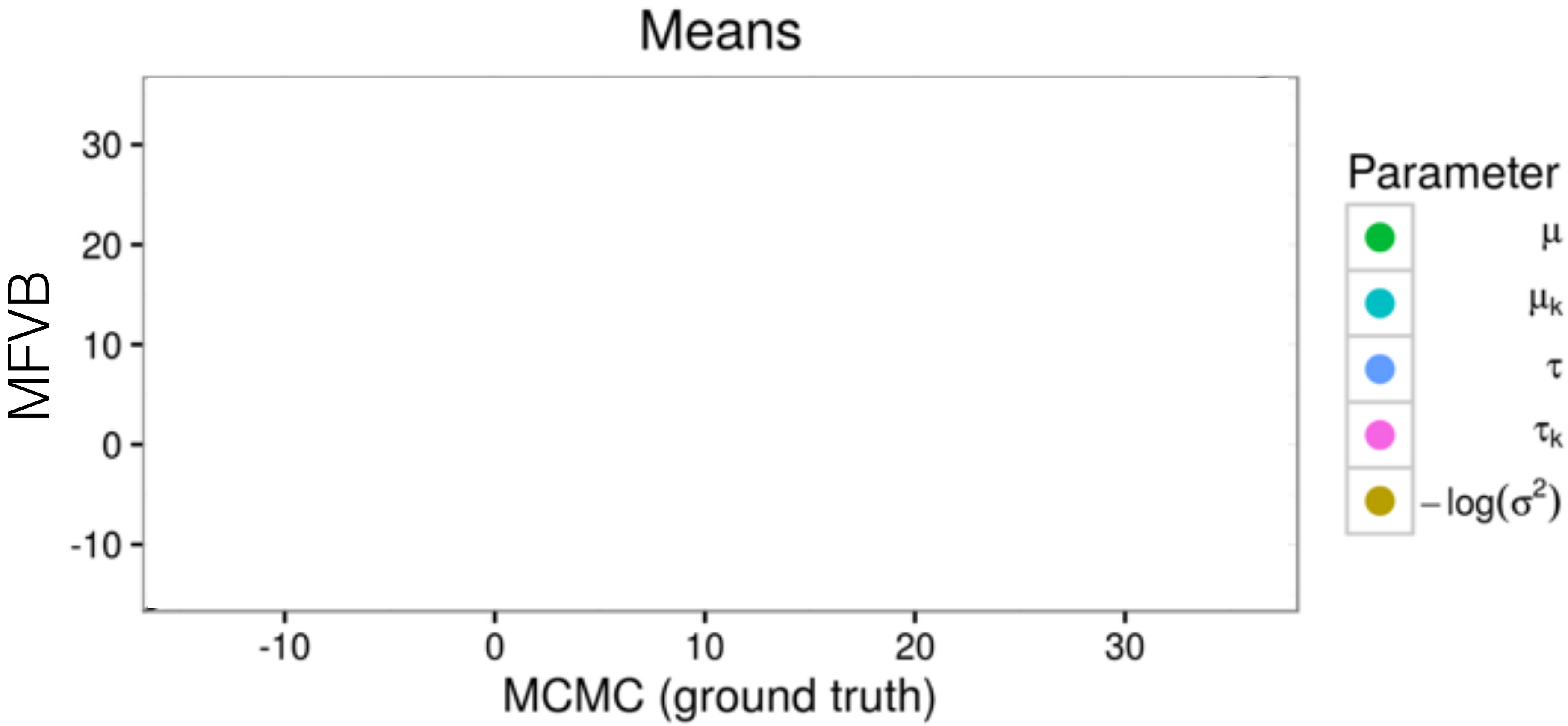
Microcredit



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

Microcredit

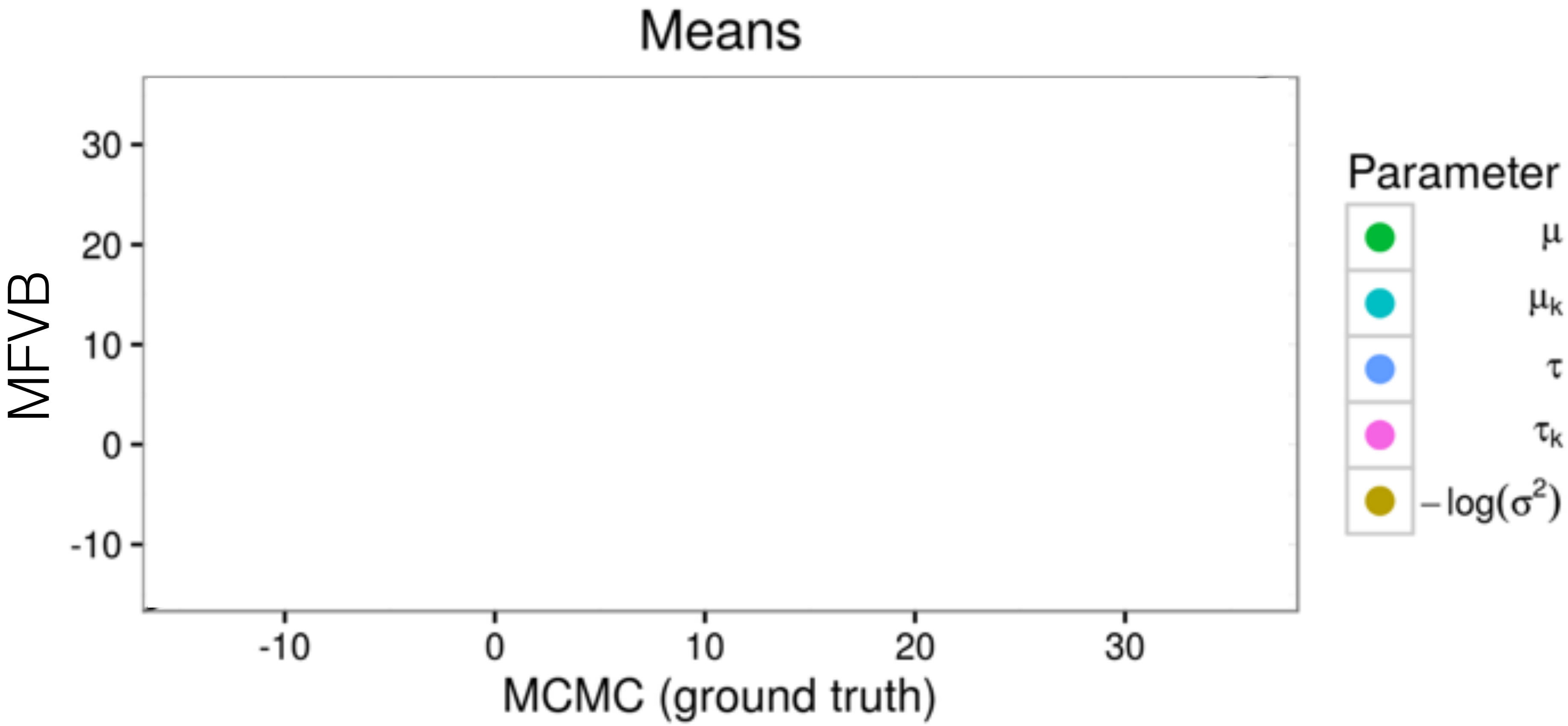
- *One set* of 2500 MCMC draws: **45 minutes**



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

Microcredit

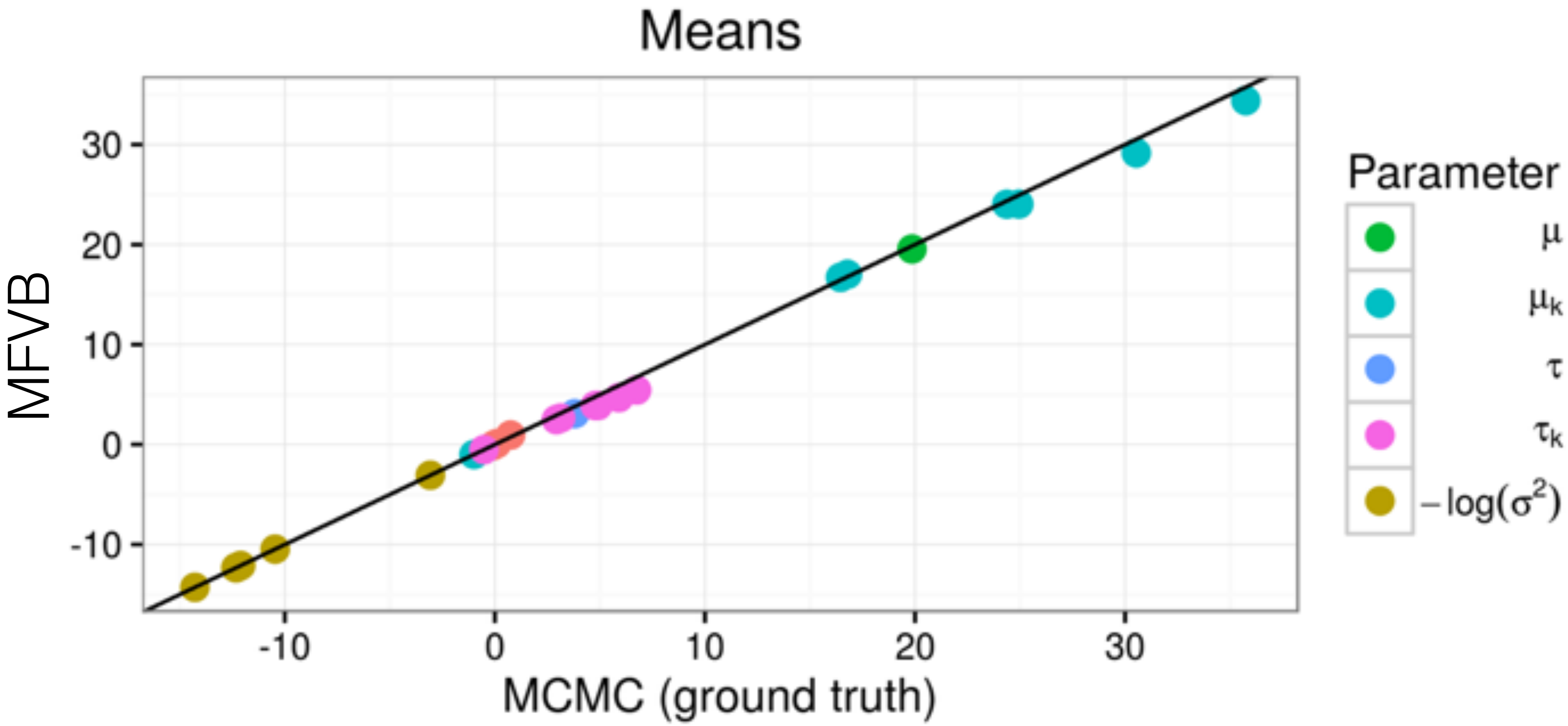
- *One set* of 2500 MCMC draws: **45 minutes**
- MFVB optimization: **<1 min**



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

Microcredit

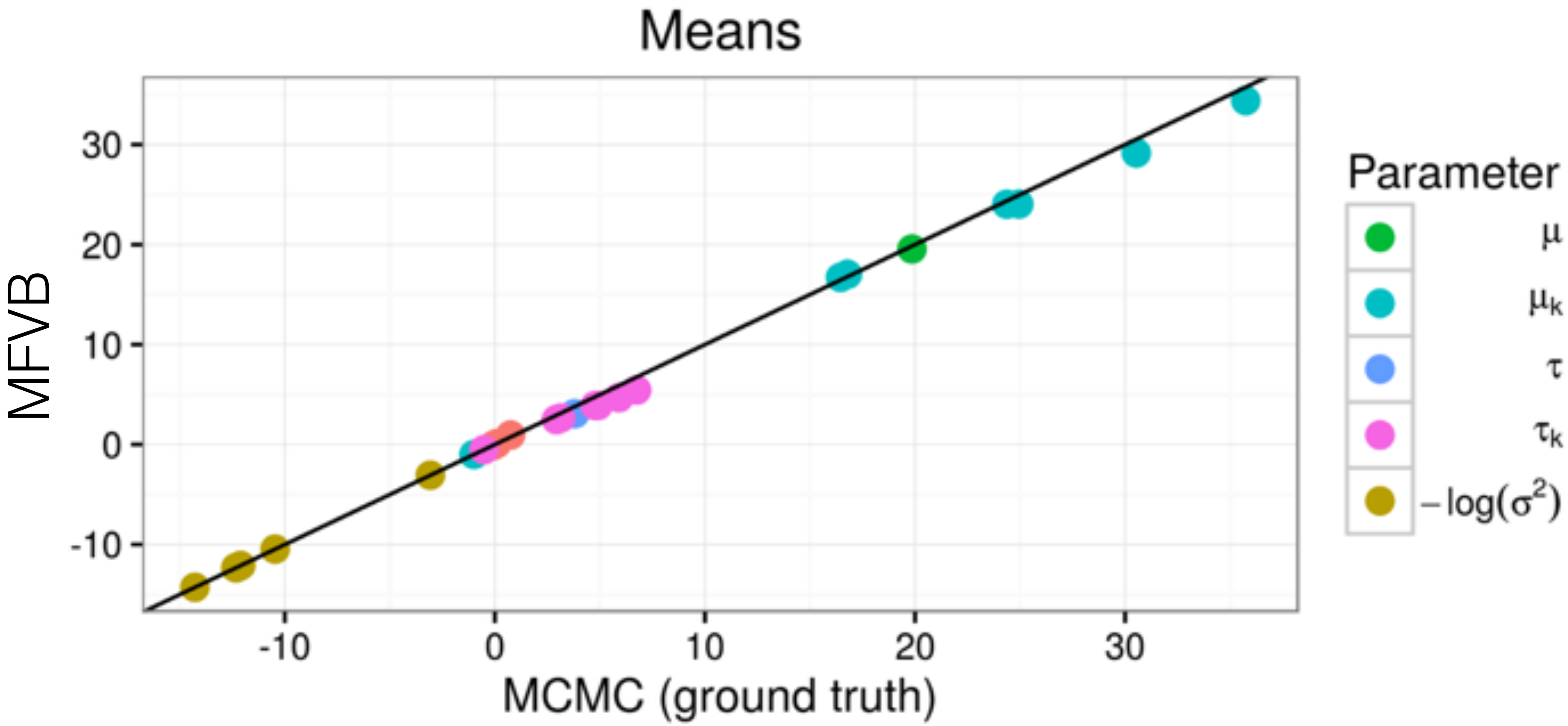
- *One set* of 2500 MCMC draws: **45 minutes**
- MFVB optimization: **<1 min**



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

Microcredit

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



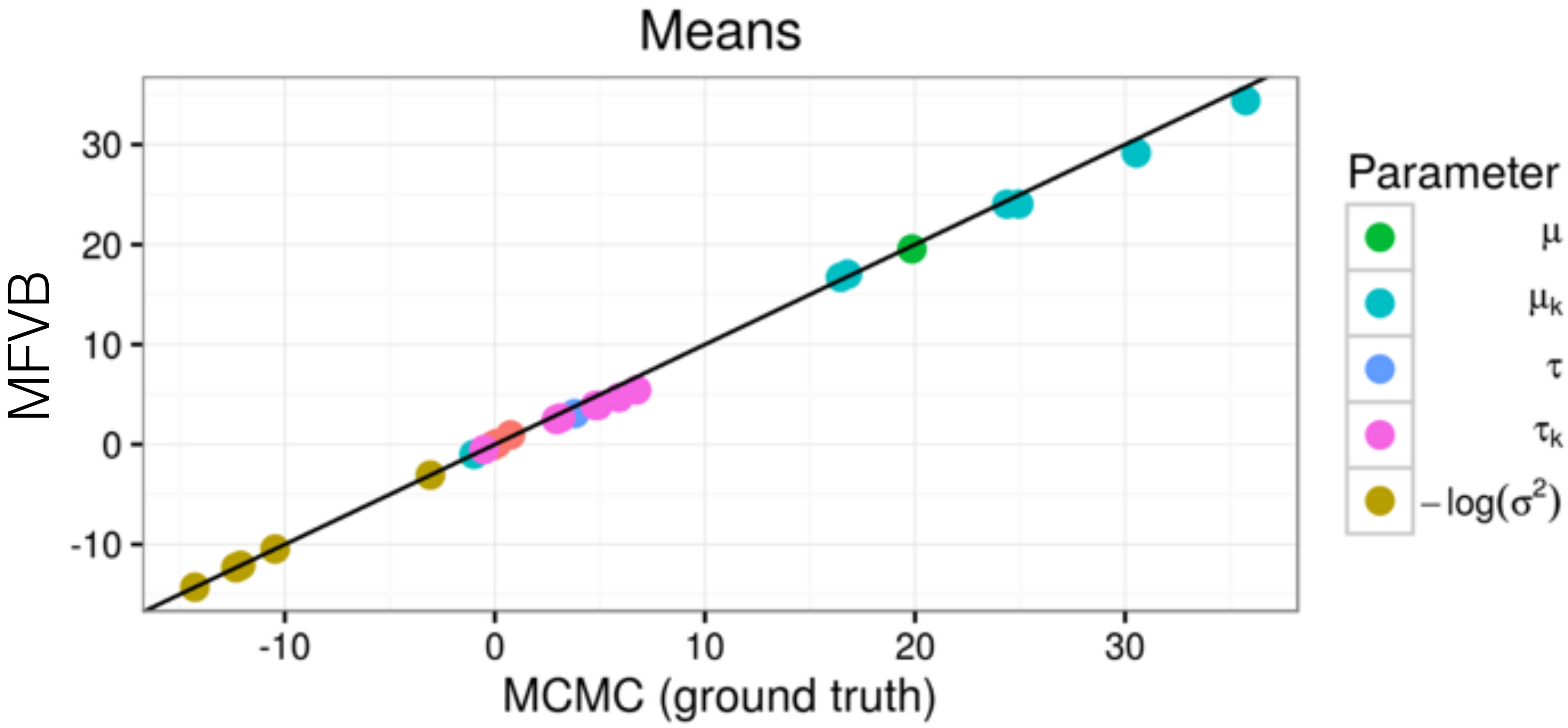
Criteo Online Ads Experiment

- 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**



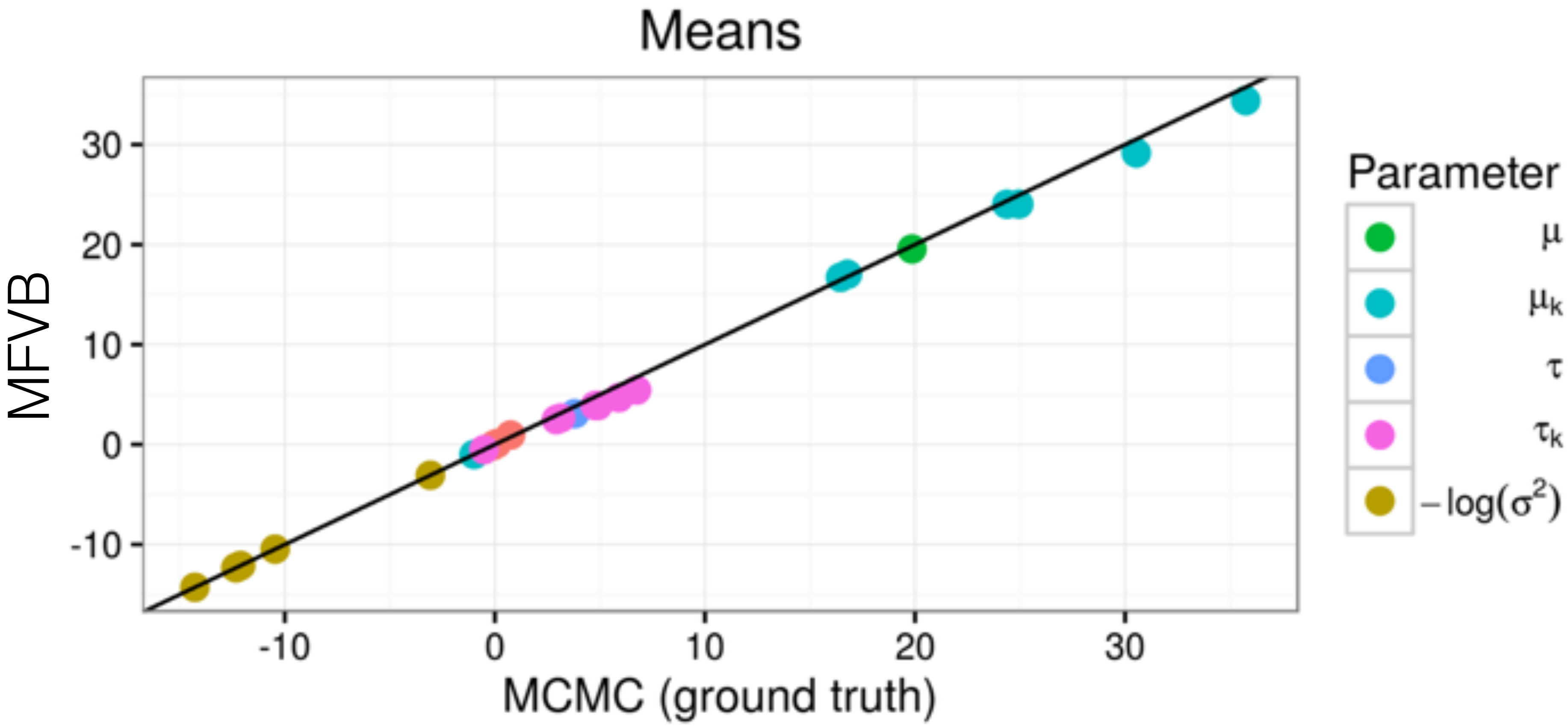
Criteo Online Ads Experiment

- 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; N = 61,895 subset to compare to MCMC

[Giordano, Broderick, Meager, Huggins, Jordan 2016; Giordano, Broderick, Jordan 2017]

Criteo Online Ads Experiment

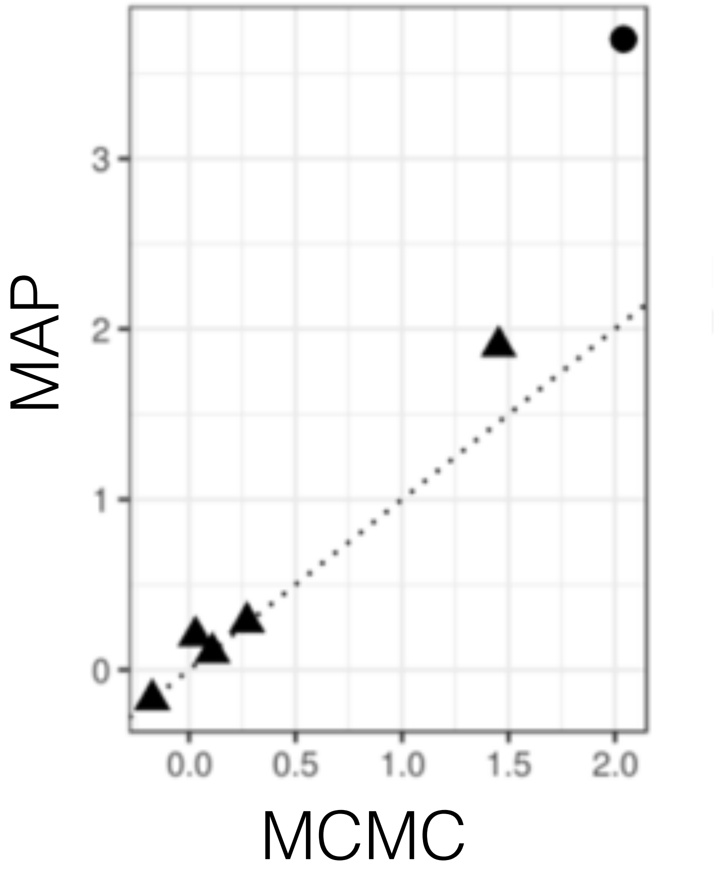
[Giordano, Broderick, Jordan 2017]

Criteo Online Ads Experiment

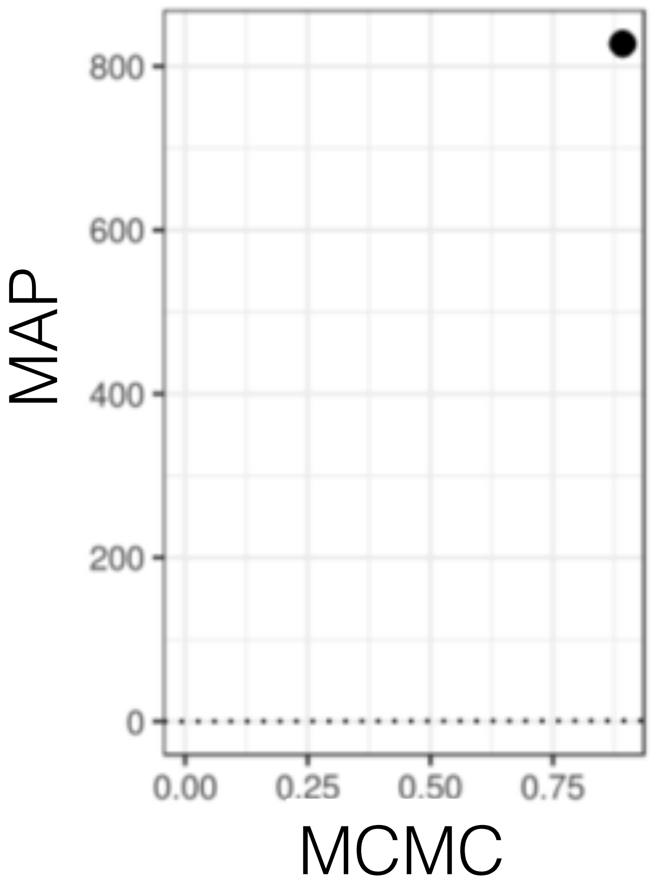
- MAP: **12 s**

[Giordano, Broderick, Jordan 2017]

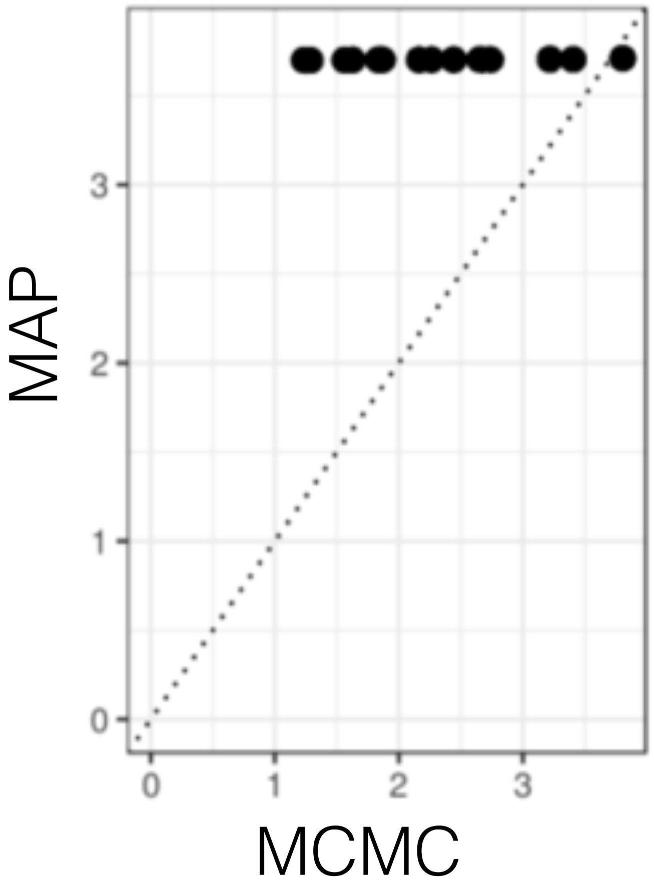
Criteo Online Ads Experiment



Global parameters (-τ)



Global parameter τ



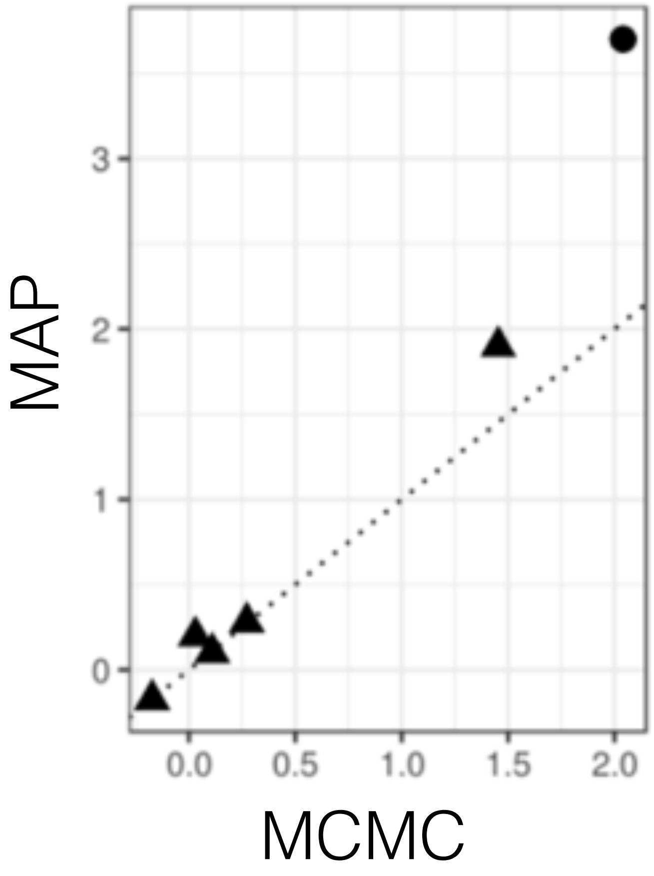
Local parameters

- MAP: **12 s**

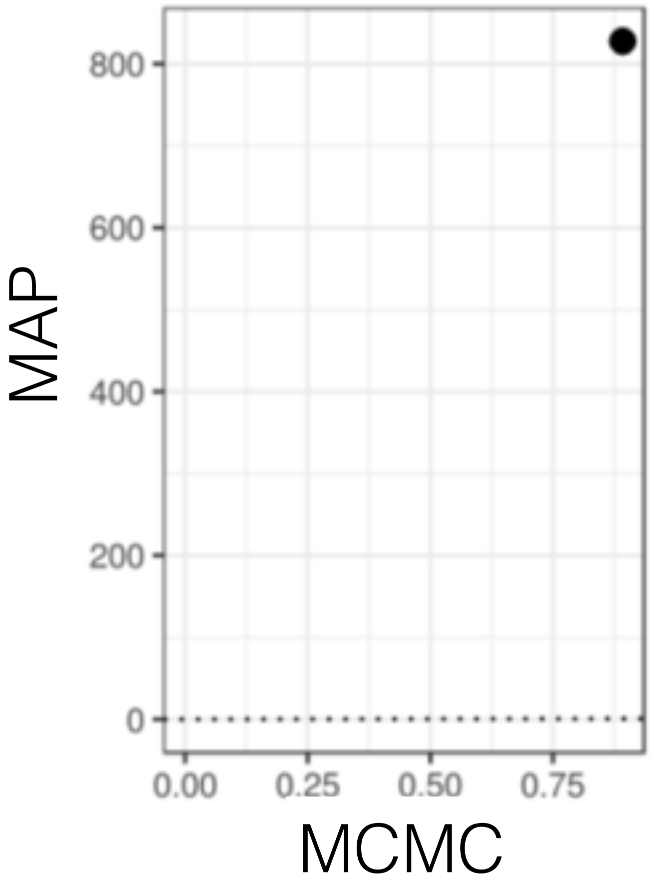
[Giordano, Broderick, Jordan 2017]

Criteo Online Ads Experiment

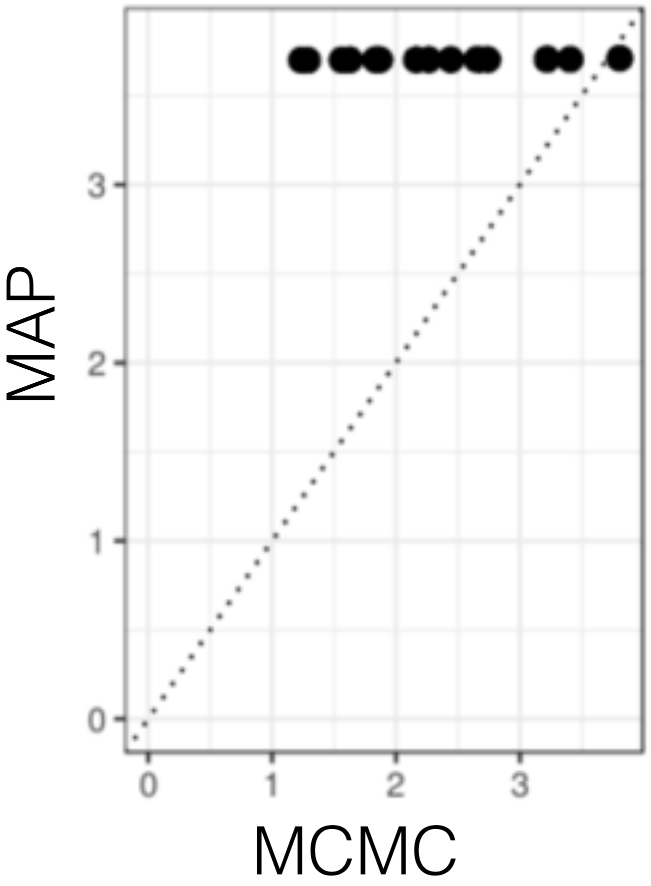
Global parameters (-τ)



Global parameter τ



Local parameters

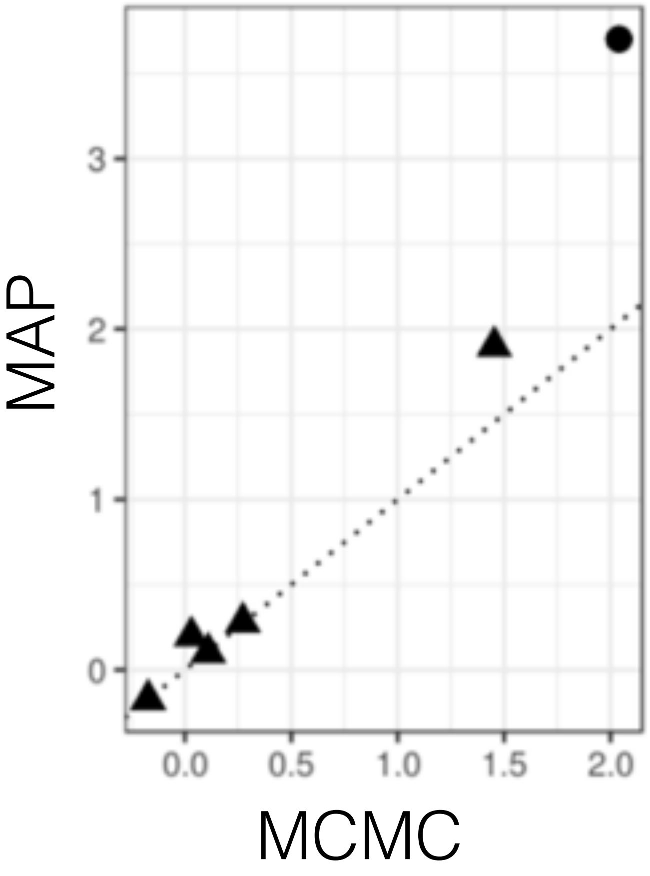


- MAP: **12 s**
- MFVB: **57 s**

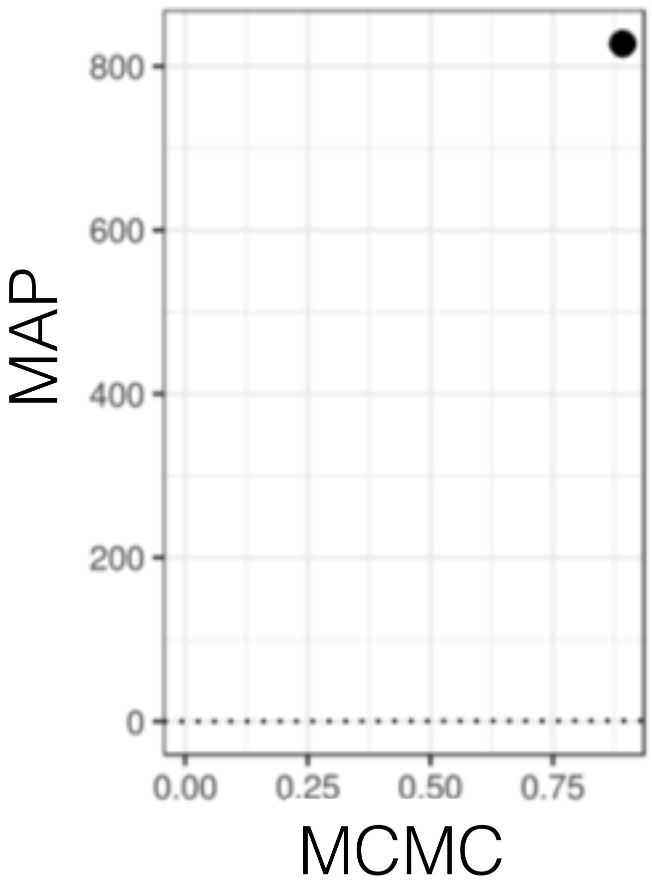
[Giordano, Broderick, Jordan 2017]

Criteo Online Ads Experiment

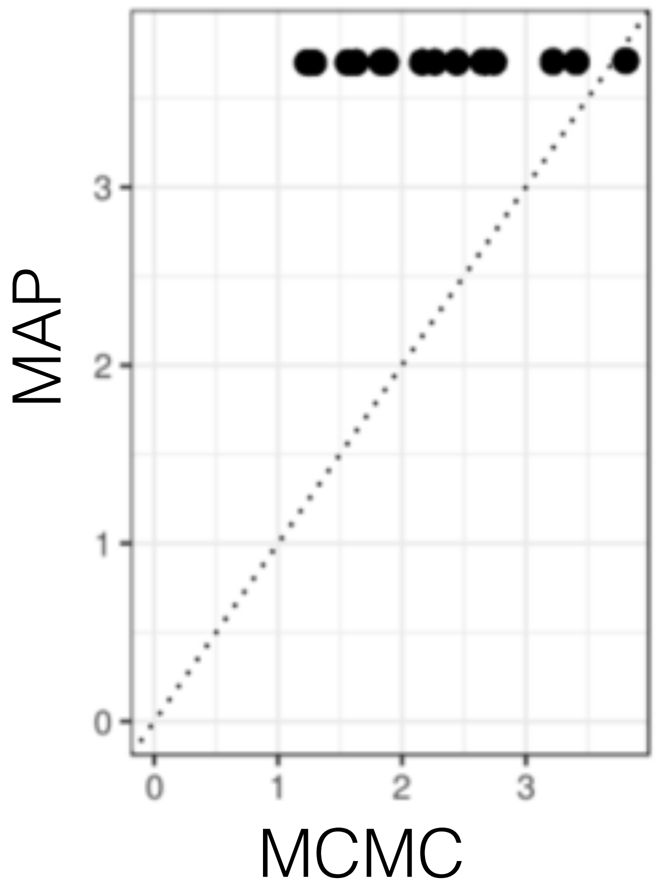
Global parameters (-τ)



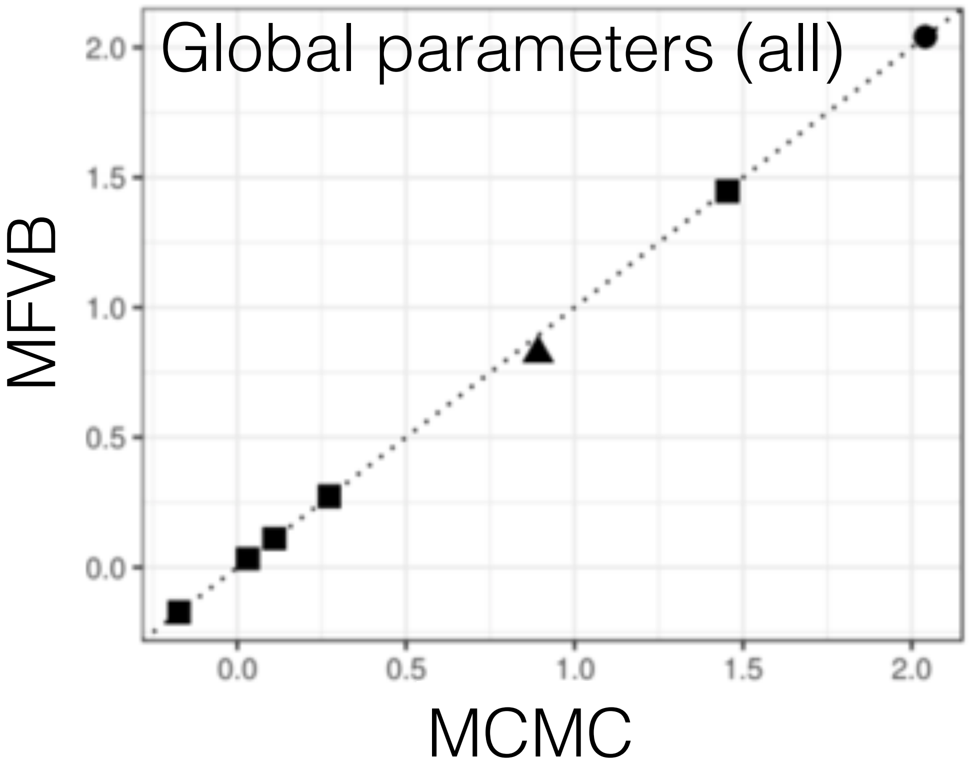
Global parameter τ

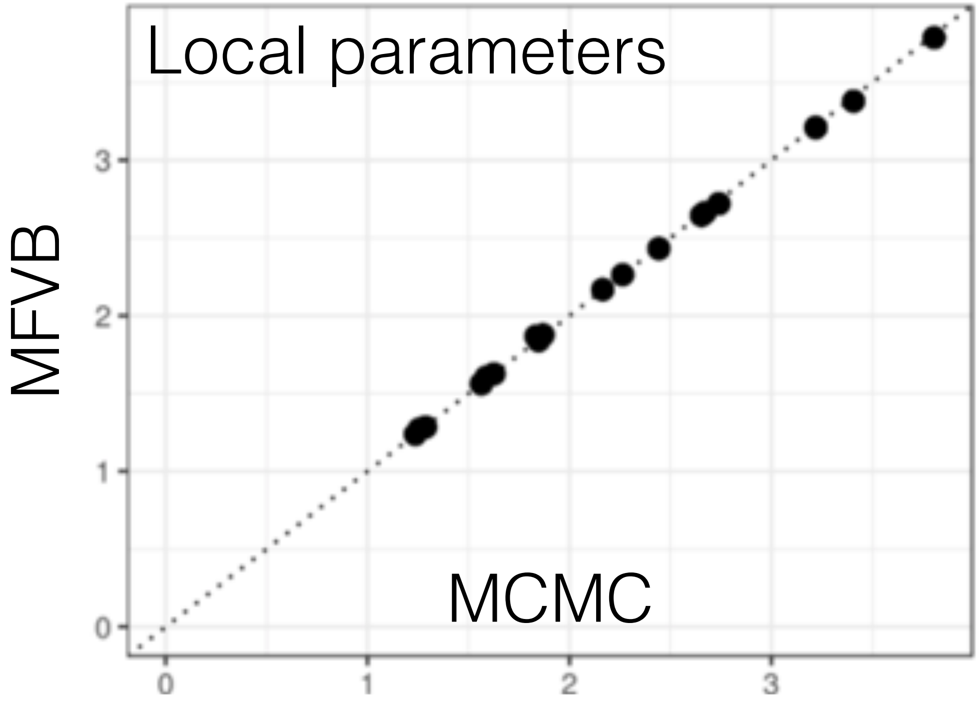


Local parameters



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

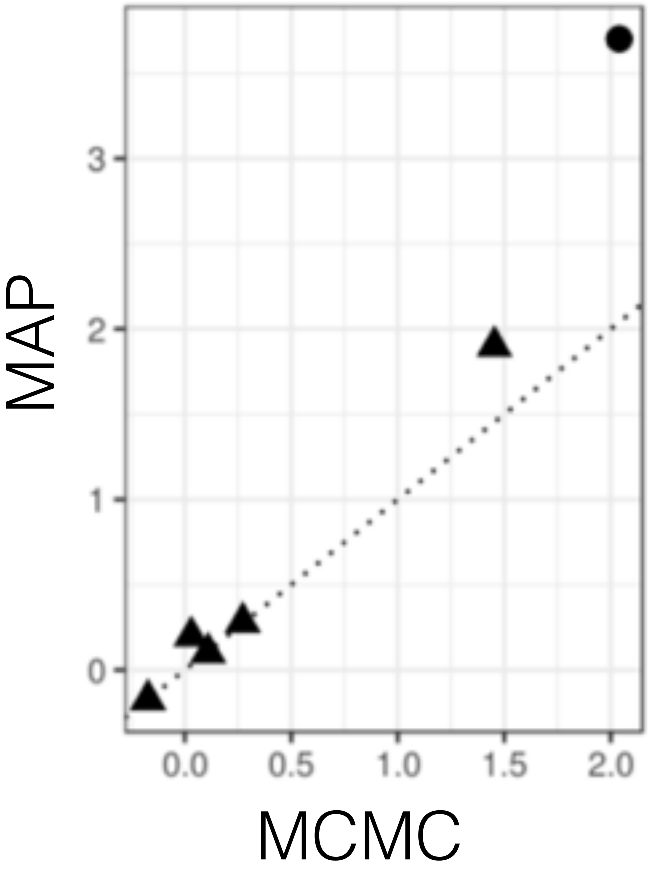

Global parameters (all)


Local parameters

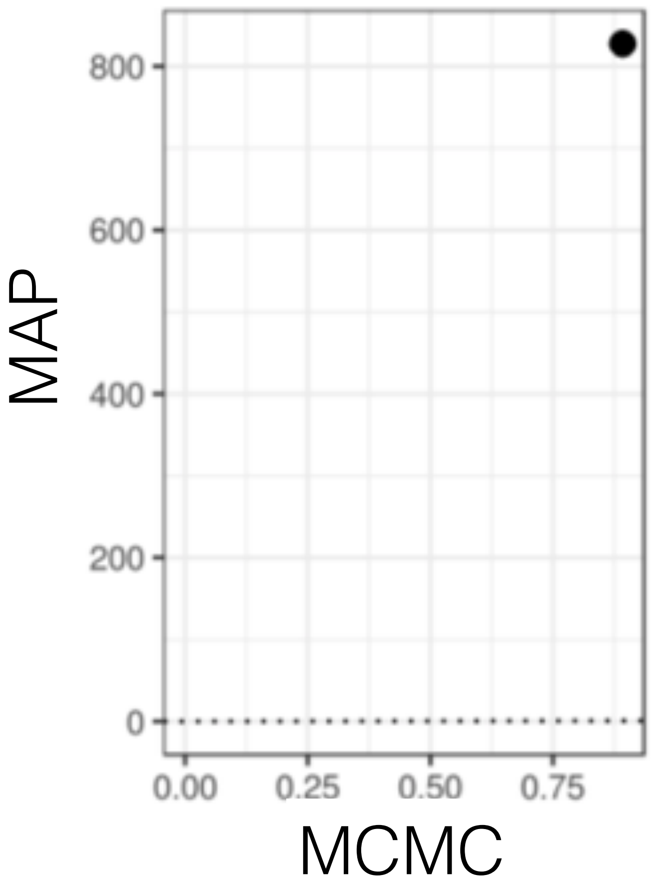
[Giordano, Broderick, Jordan 2017]

Criteo Online Ads Experiment

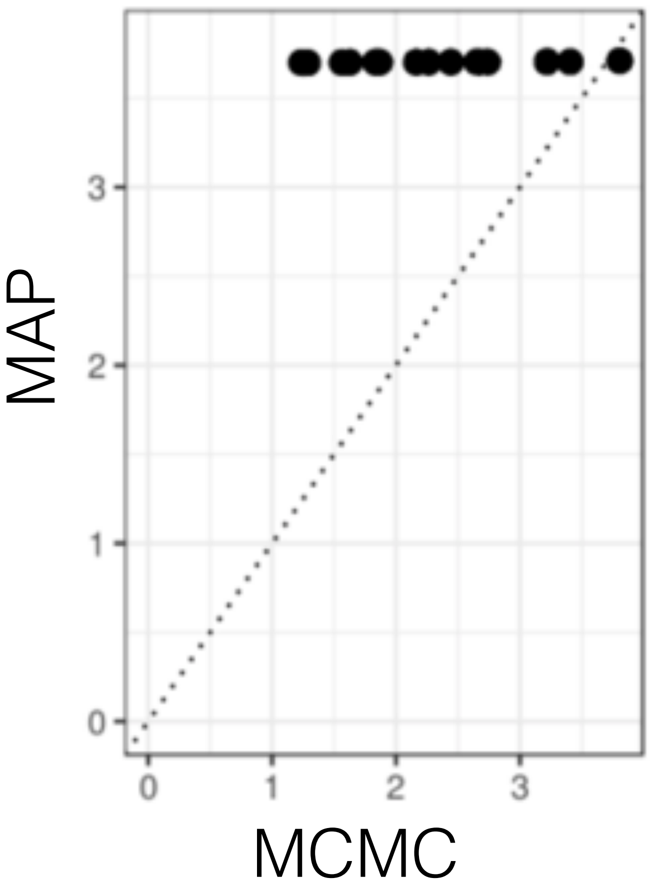
Global parameters (-τ)



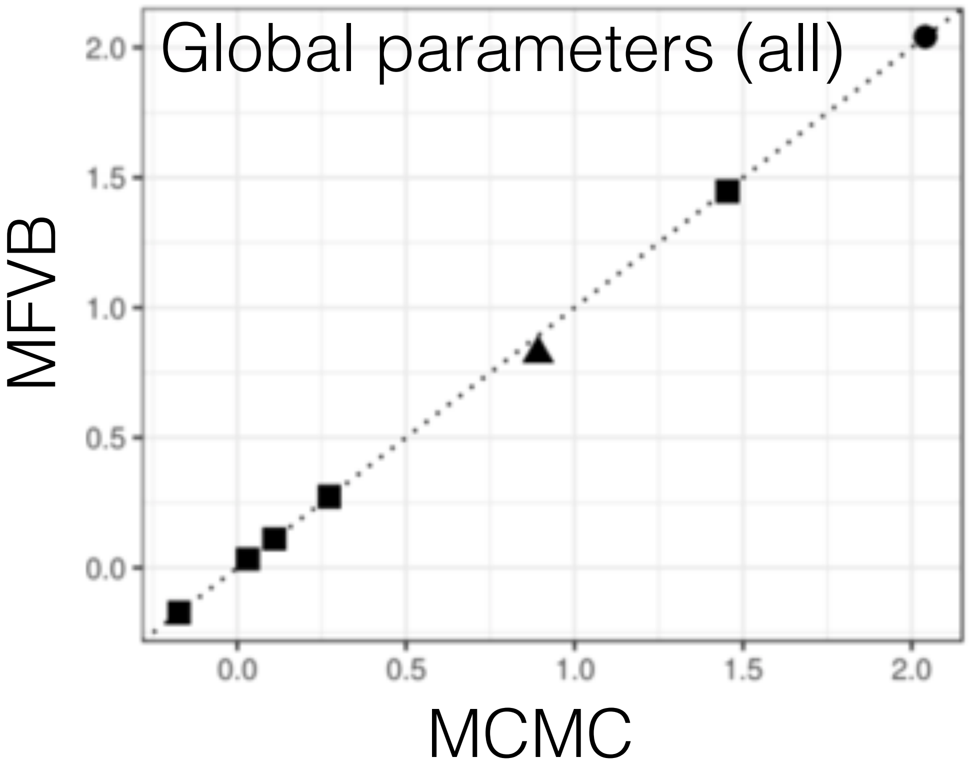
Global parameter τ

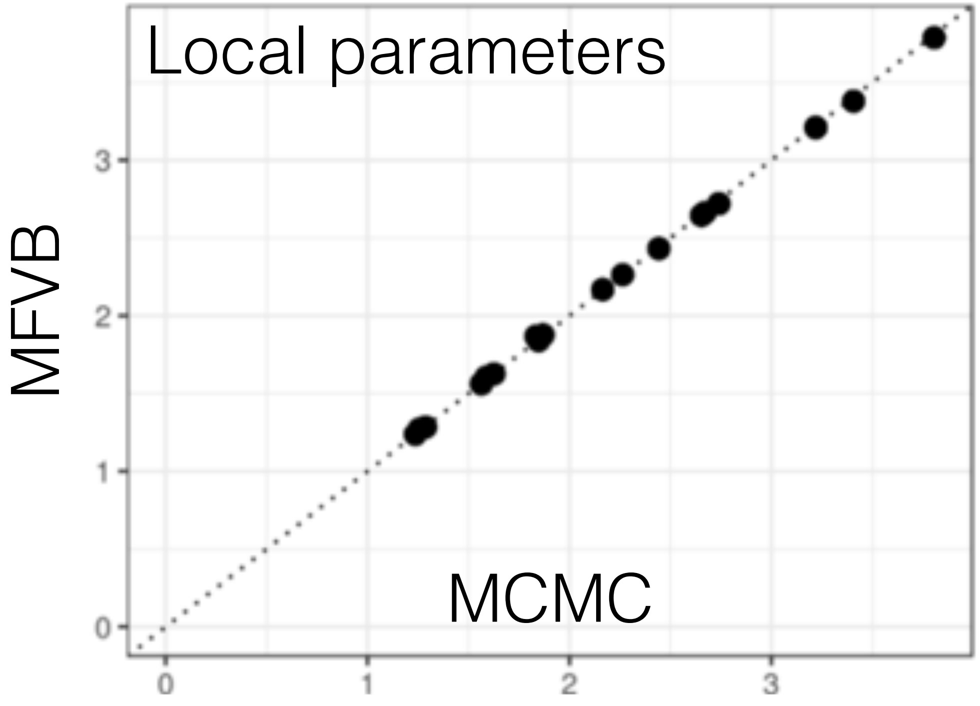


Local parameters



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


Global parameters (all)


Local parameters

[Giordano, Broderick, Jordan 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?

What about uncertainty?

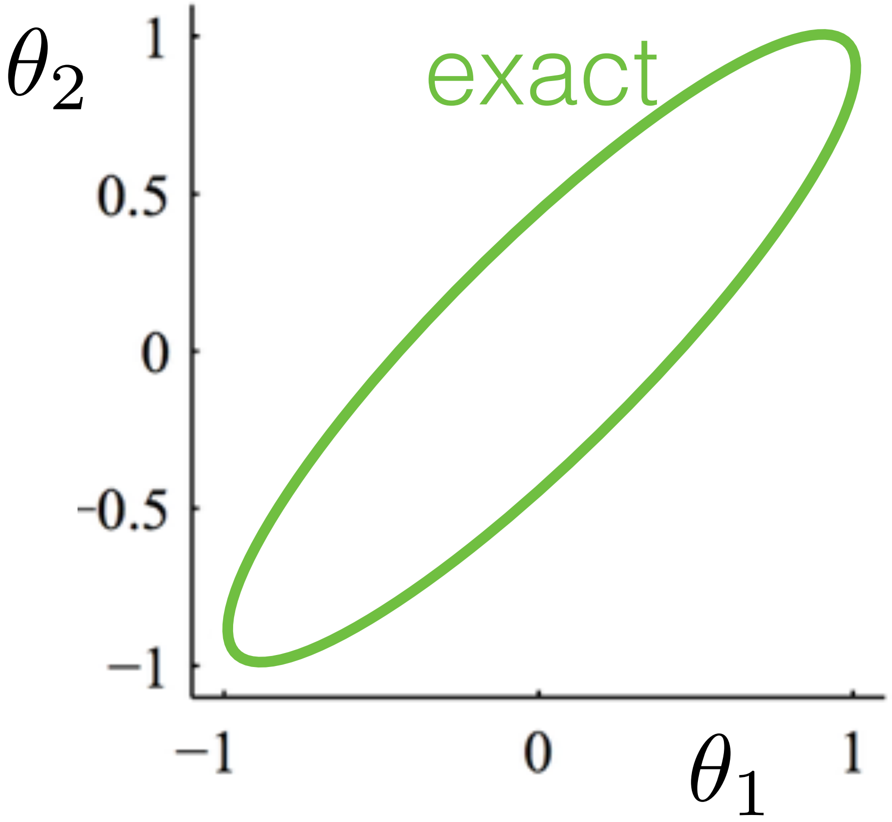
$$KL(q||p(\cdot|x)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|x)} d\theta$$

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

What about uncertainty?

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

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

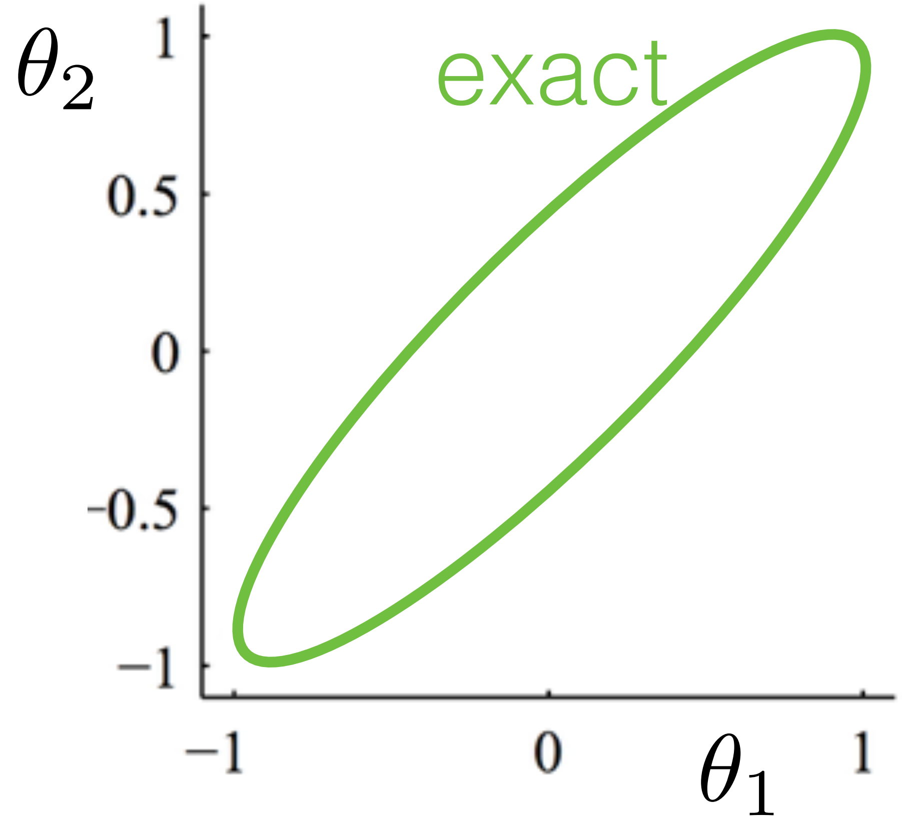


[Turner & Sahani 2011; MacKay 2003; Bishop 2006; Wang, Titterington 2004]

What about uncertainty?

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

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



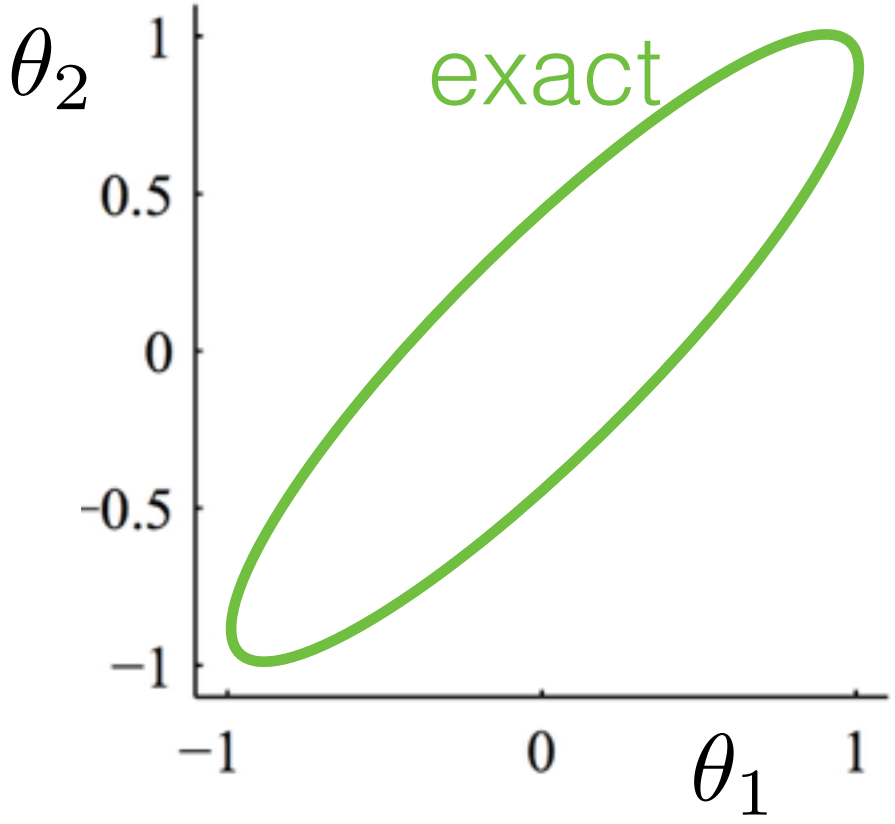
[Turner & Sahani 2011; MacKay 2003; Bishop 2006; Wang, Titterington 2004]

- Conjugate linear regression

What about uncertainty?

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

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



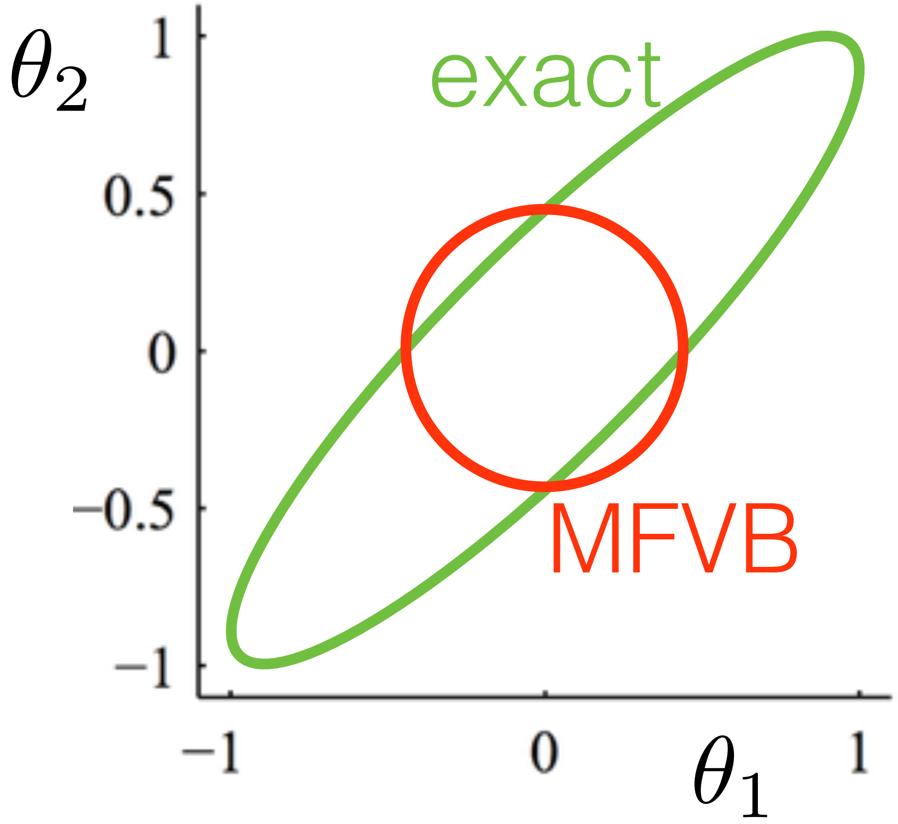
[Turner & Sahani
2011; MacKay 2003;
Bishop 2006; Wang,
Titterington 2004]

- Conjugate linear regression
- Bayesian central limit theorem

What about uncertainty?

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

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



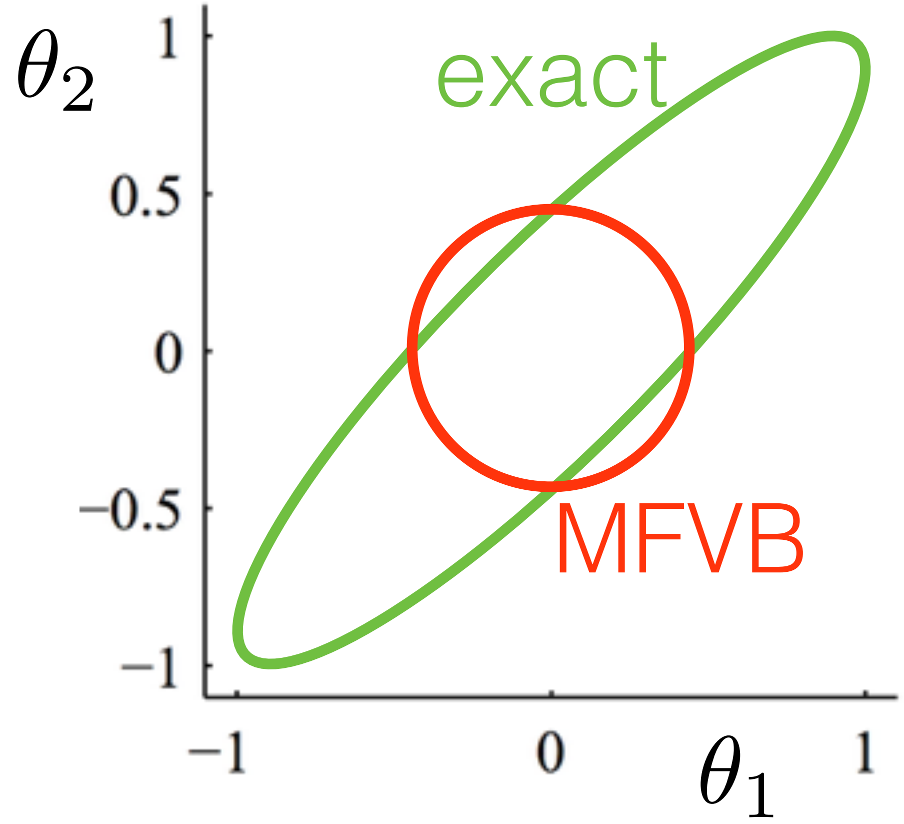
[Turner & Sahani 2011; MacKay 2003; Bishop 2006; Wang, Titterington 2004]

- Conjugate linear regression
- Bayesian central limit theorem

What about uncertainty?

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

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



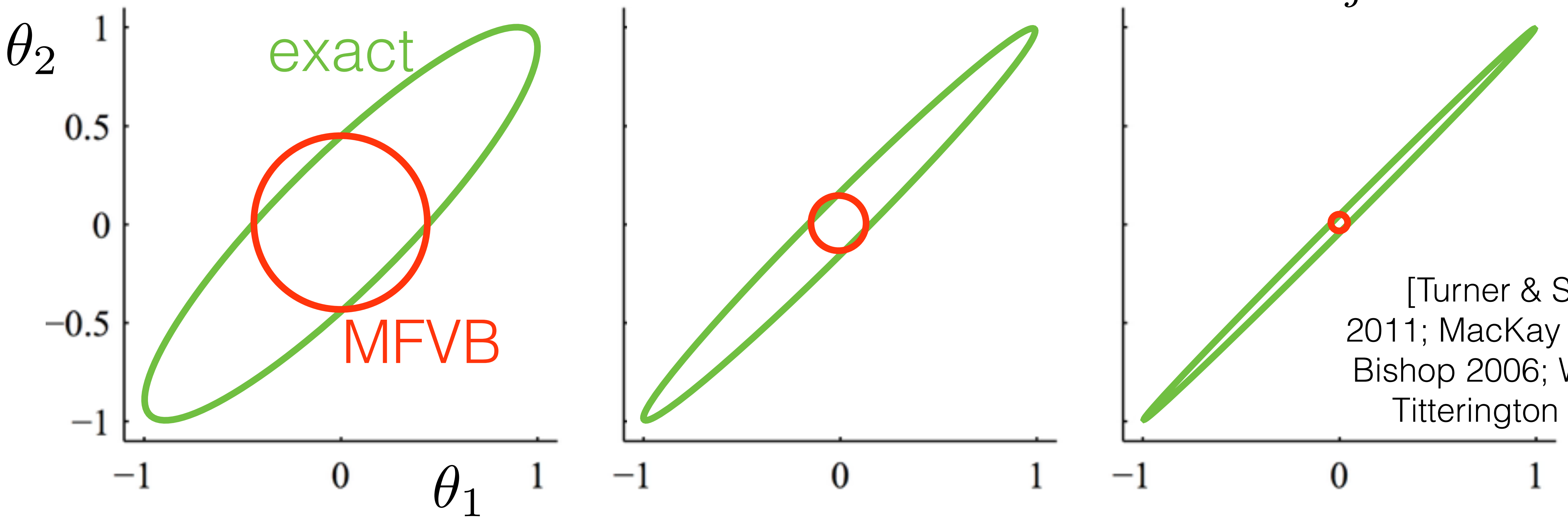
[Turner & Sahani 2011; MacKay 2003; Bishop 2006; Wang, Titterington 2004]

- Underestimates variance (sometimes severely)

- Conjugate linear regression
- Bayesian central limit theorem

What about uncertainty?

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



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

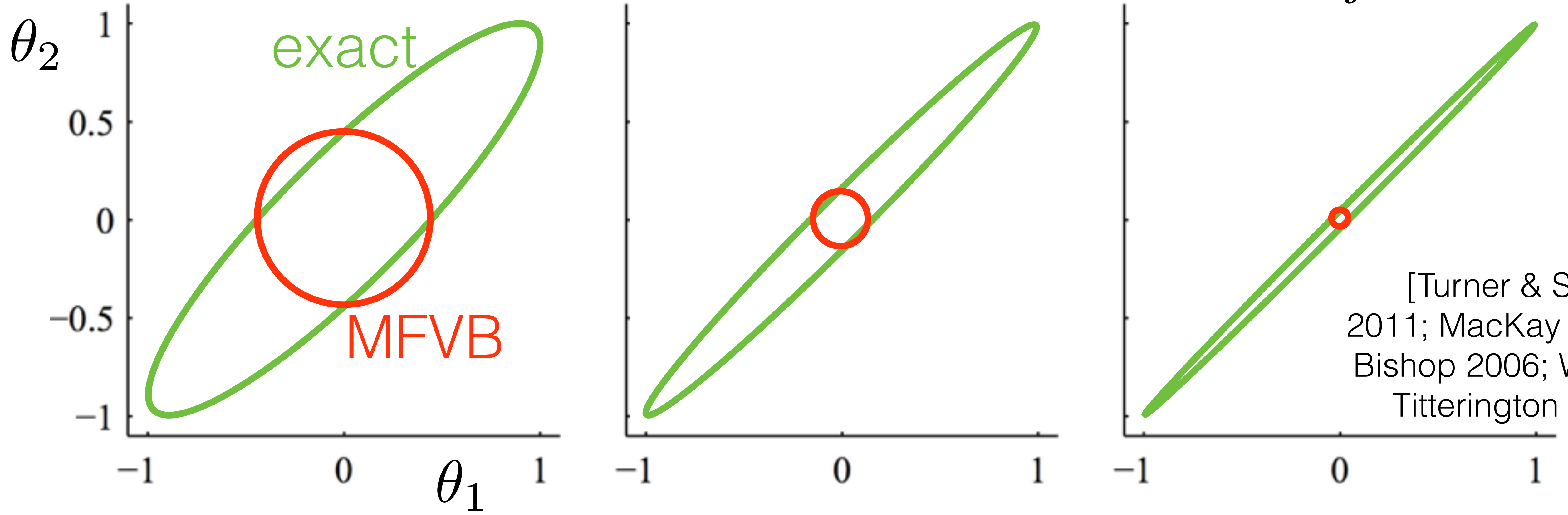
- Underestimates variance (sometimes severely)

- Conjugate linear regression
- Bayesian central limit theorem

What about uncertainty?

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

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



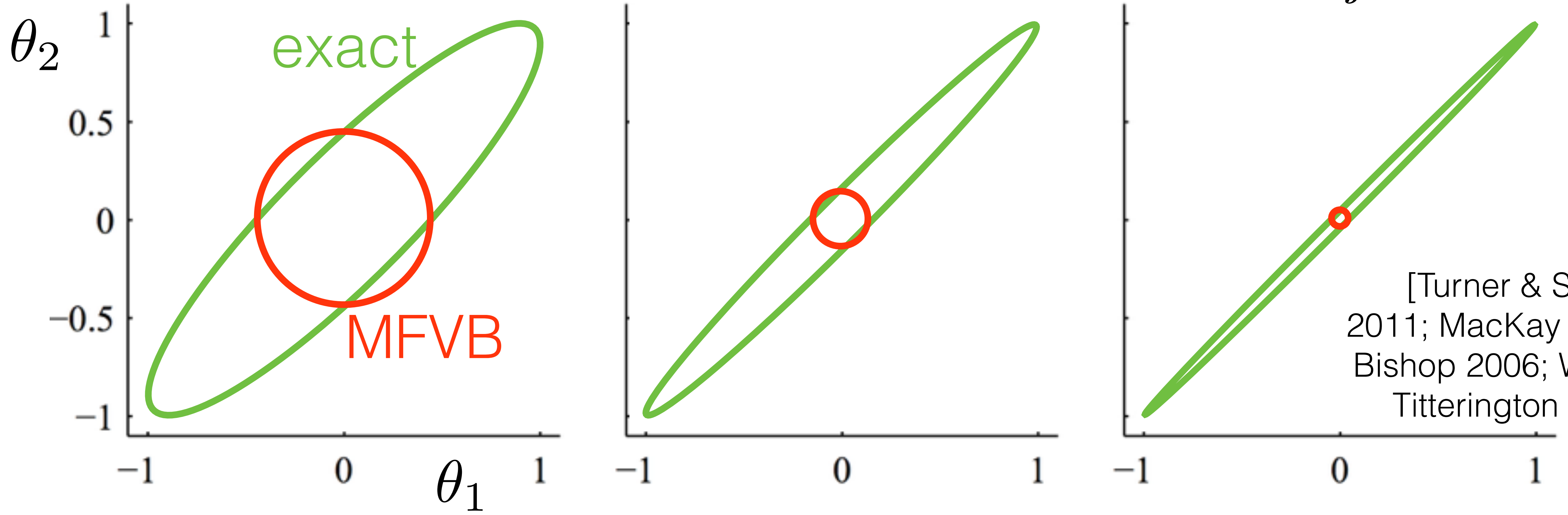
[Turner & Sahani 2011; MacKay 2003; Bishop 2006; Wang, Titterington 2004]

- Underestimates variance (sometimes severely)
- No covariance estimates

- Conjugate linear regression
- Bayesian central limit theorem

What about uncertainty?

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



[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 exact (closed) form of q^*

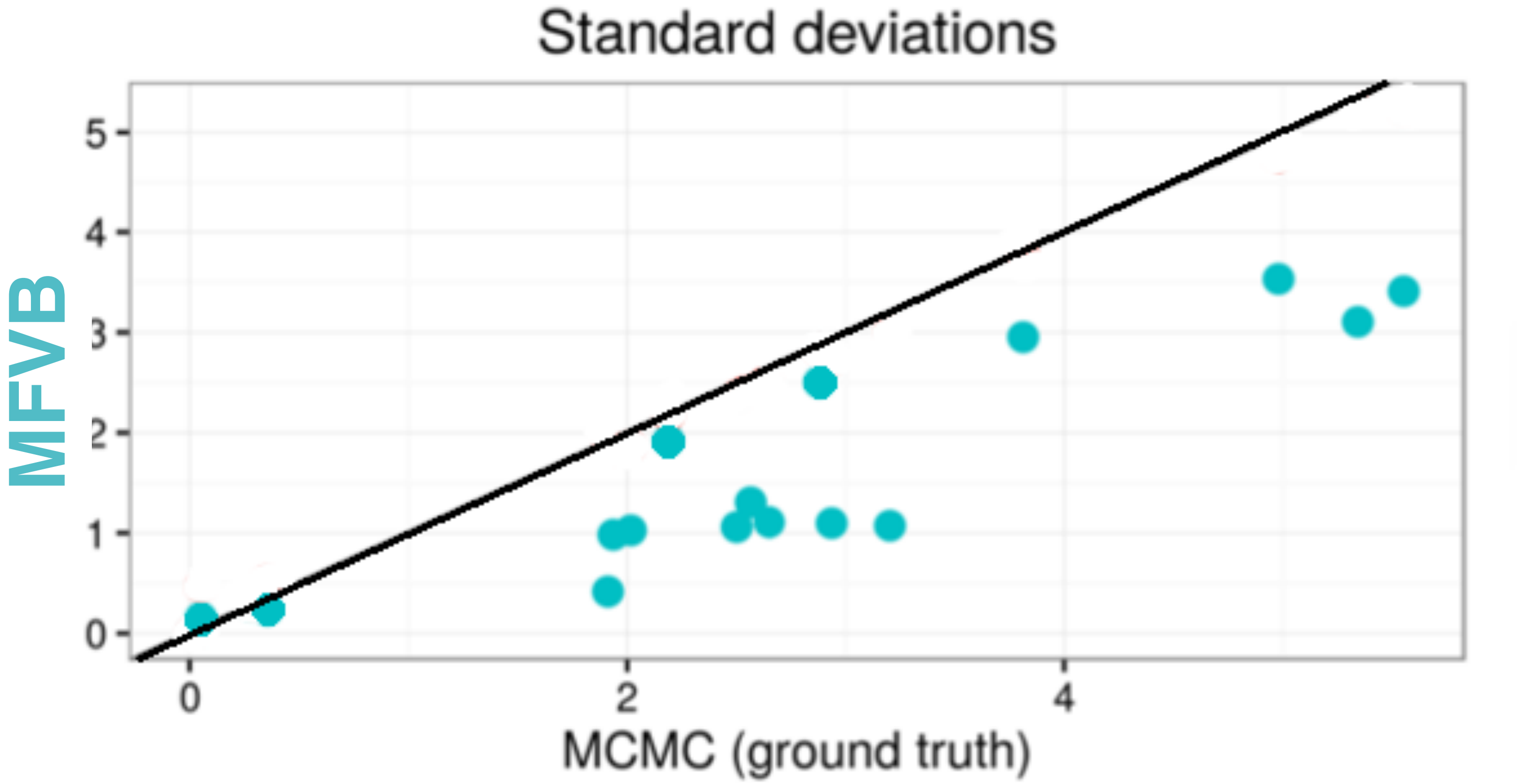
What about uncertainty?

- Microcredit

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

What about uncertainty?

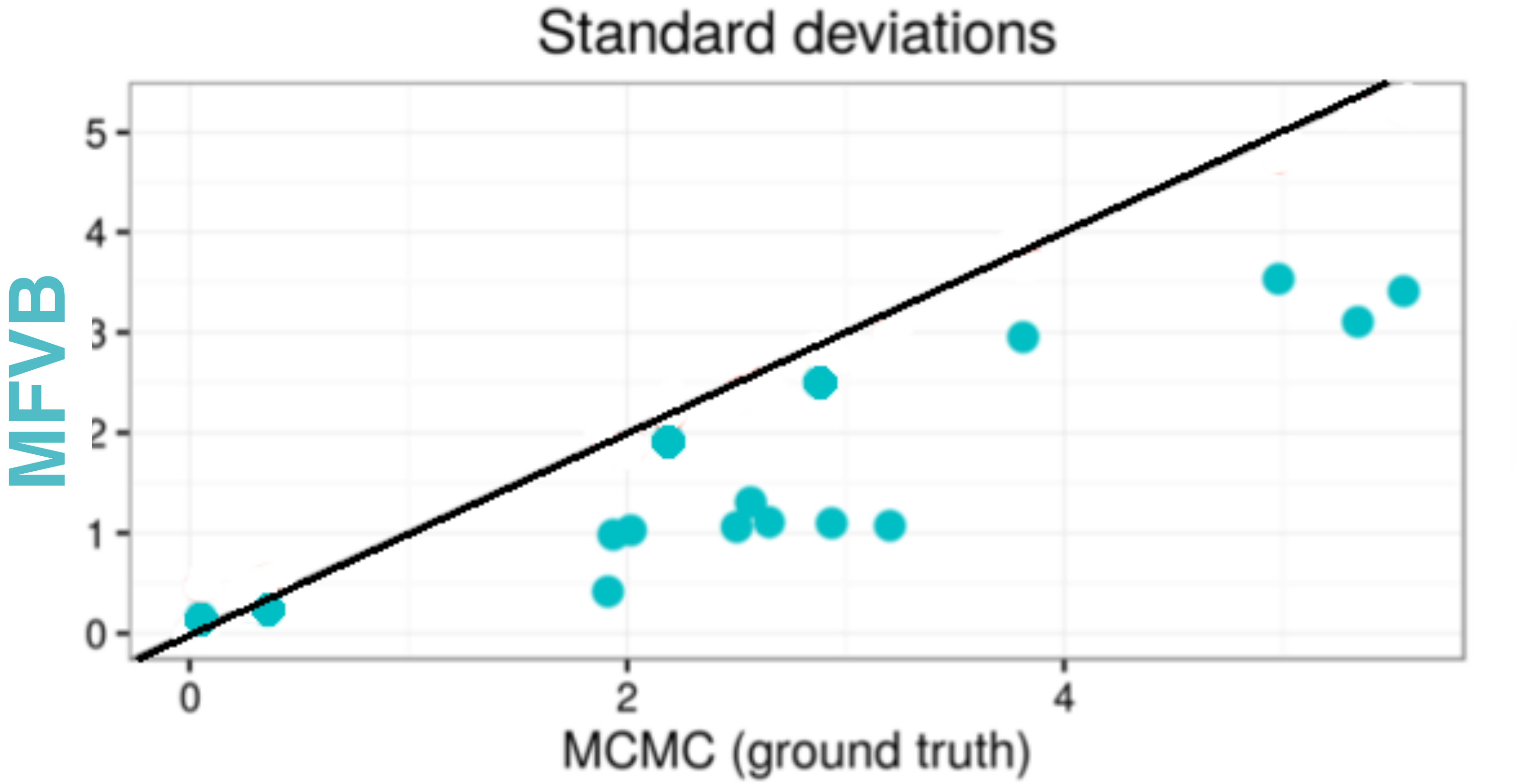
- Microcredit



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

What about uncertainty?

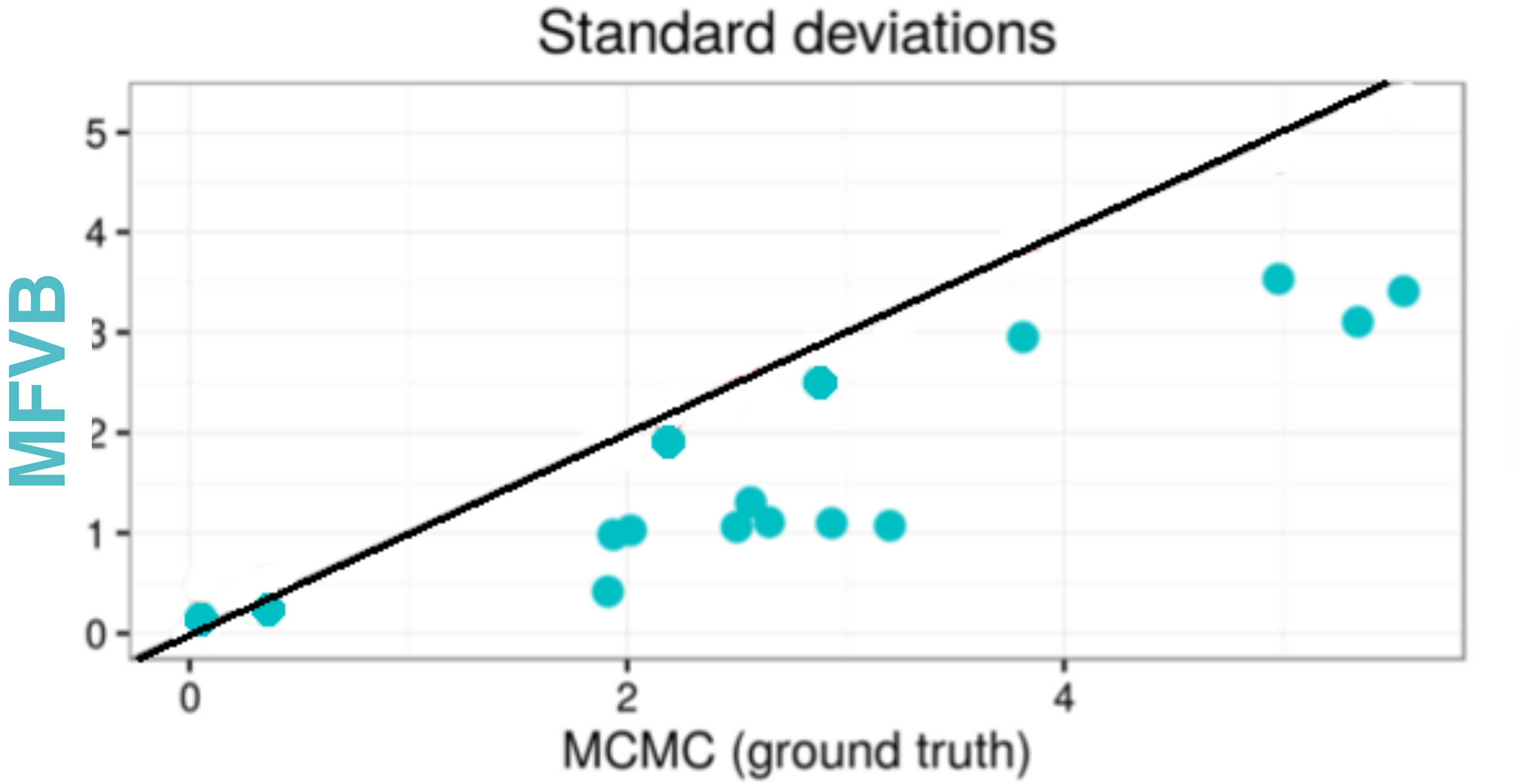
- Microcredit effect
- τ mean:
  3.08 USD PPP



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

What about uncertainty?

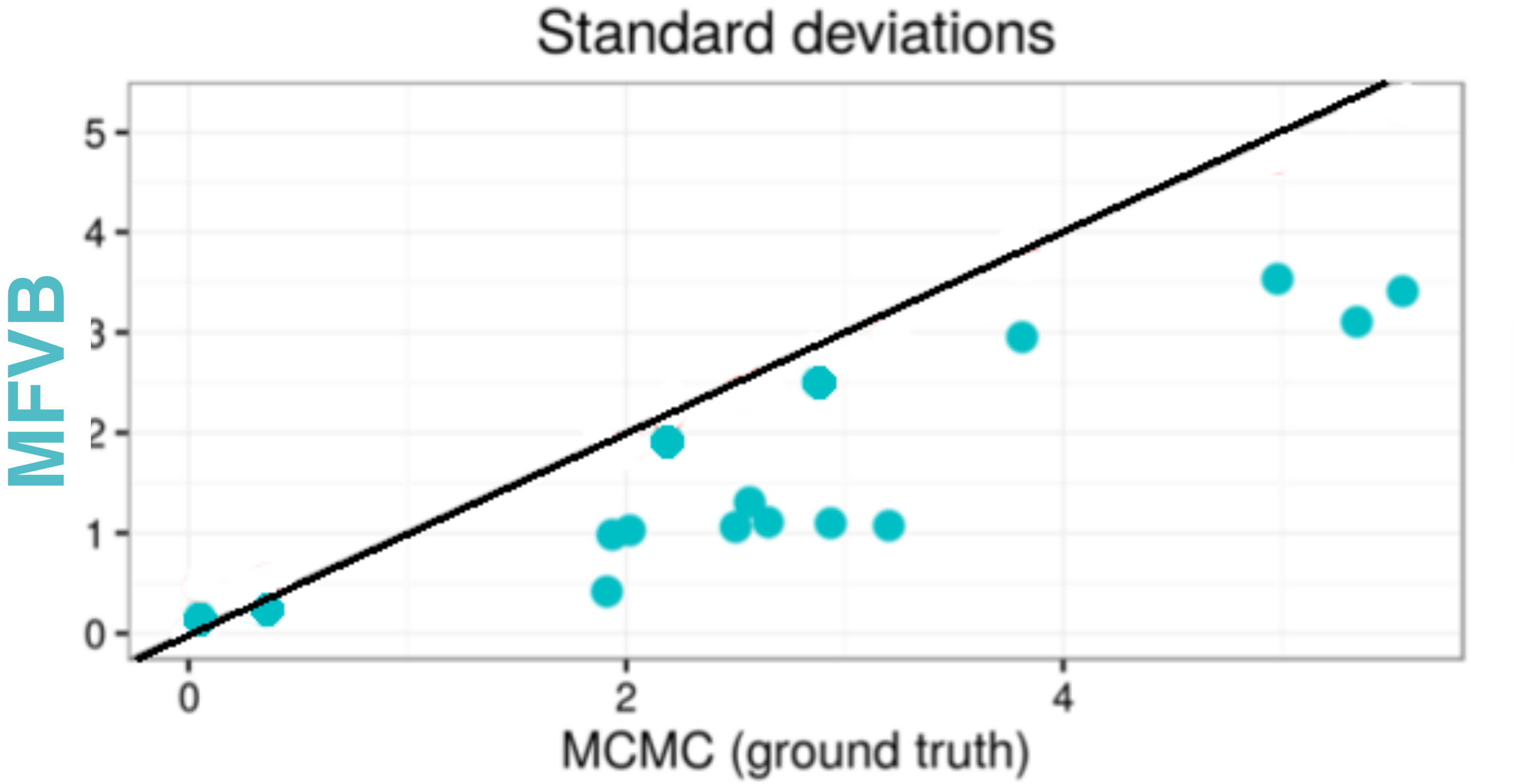
- Microcredit effect
- τ mean:
  3.08 USD PPP
- τ std dev:
  1.83 USD PPP



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

What about uncertainty?

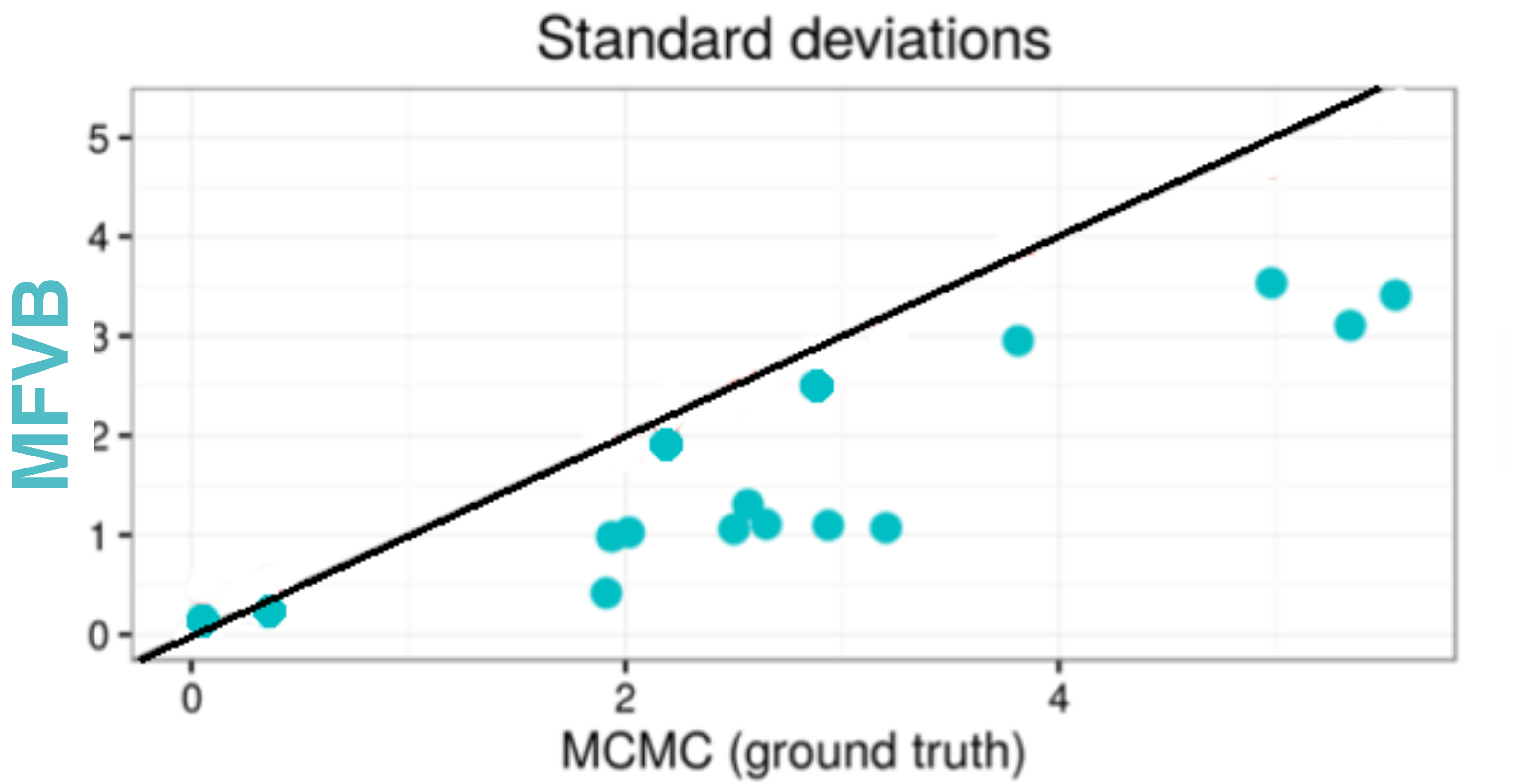
- Microcredit effect
- τ mean: 3.08 USD PPP
- τ std dev: 1.83 USD PPP
- Mean is 1.68 std dev from 0



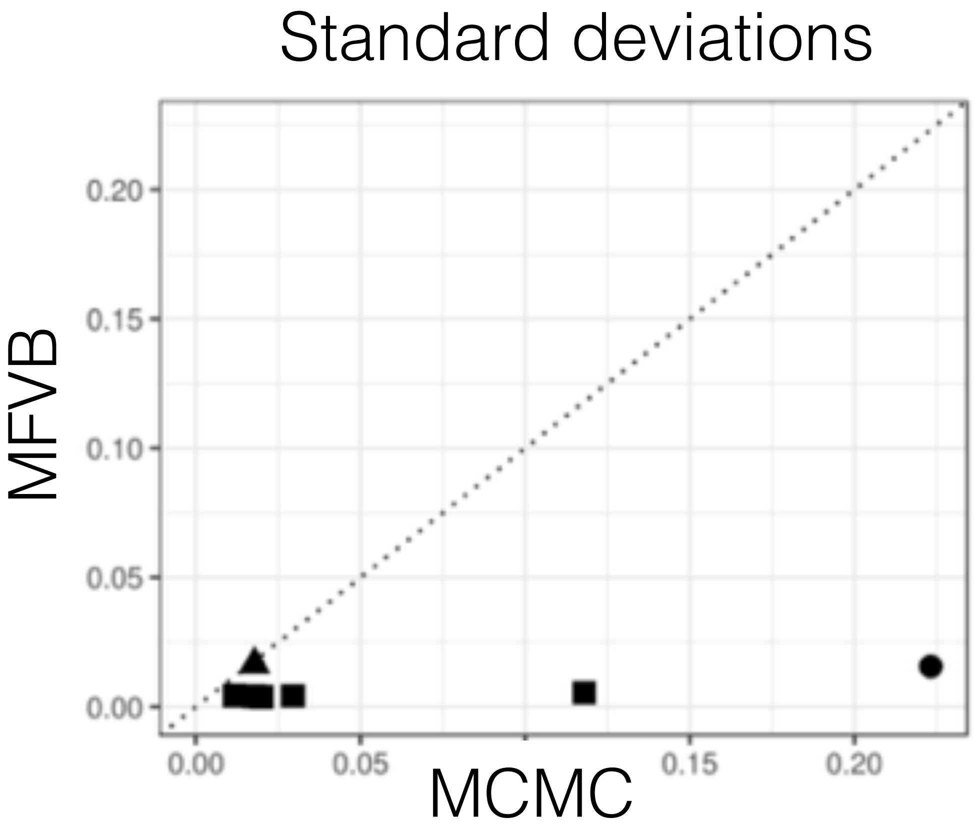
[Giordano, Broderick, Meager, Huggins, Jordan 2016]

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

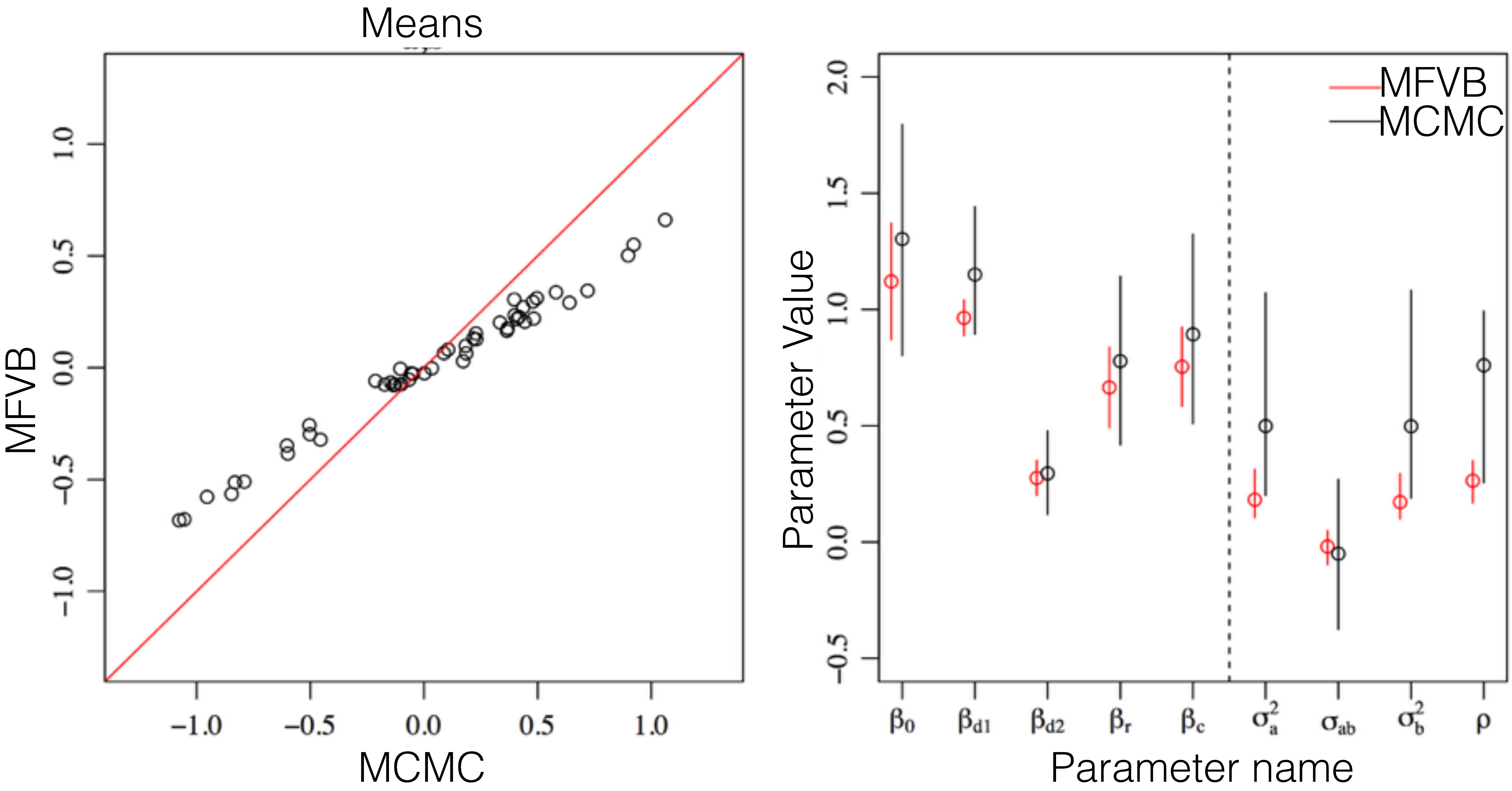


[Giordano, Broderick, Meager, Huggins, Jordan 2016; Giordano, Broderick, Jordan 2017]

What about means?

- Model for relational data with covariates
- 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: $y_n|\beta, z, \sigma^2 \overset{indep}{\sim} \mathcal{N}(\beta^T x_n + z_{k(n)} r_n, \sigma^2)$

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

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

[Giordano, Broderick, Jordan 2015]

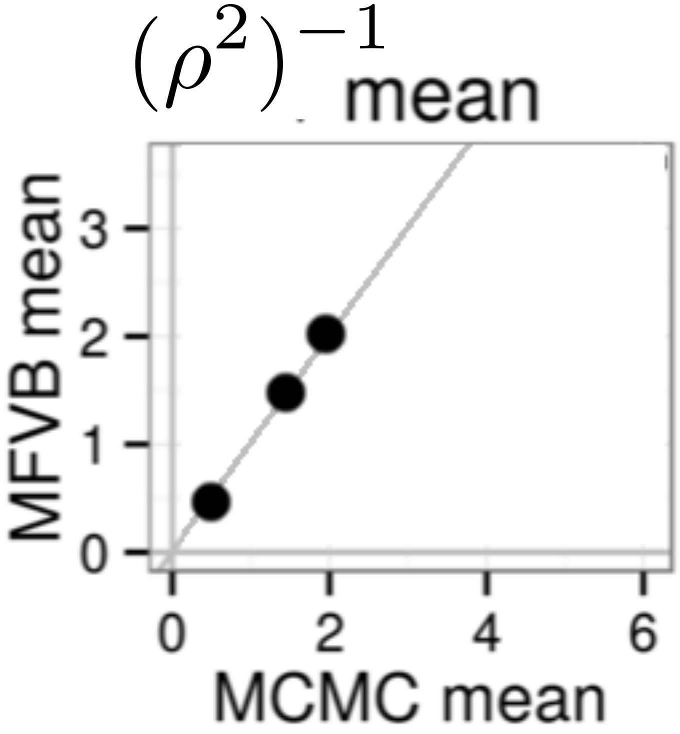
Posterior means: revisited

- 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):



[Giordano, Broderick, Jordan 2015]

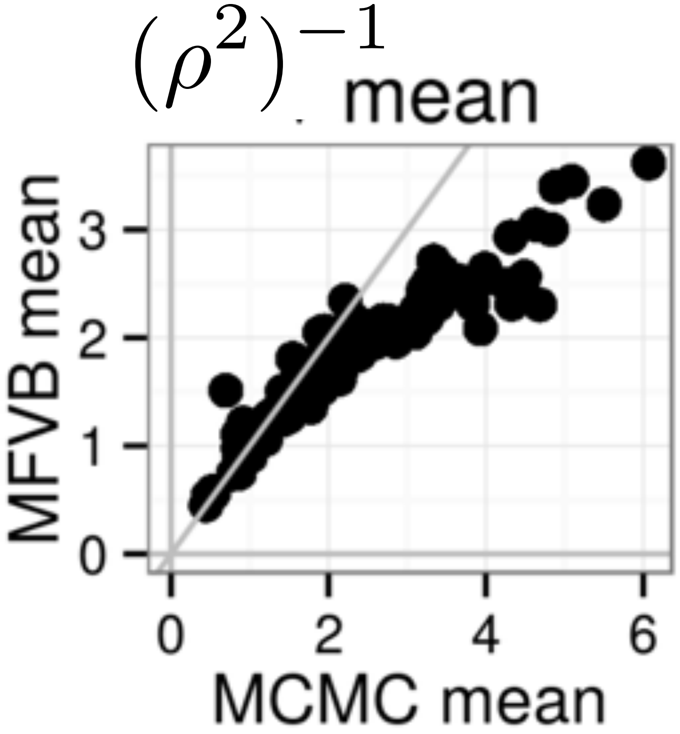
Posterior means: revisited

- 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]

What can we do?

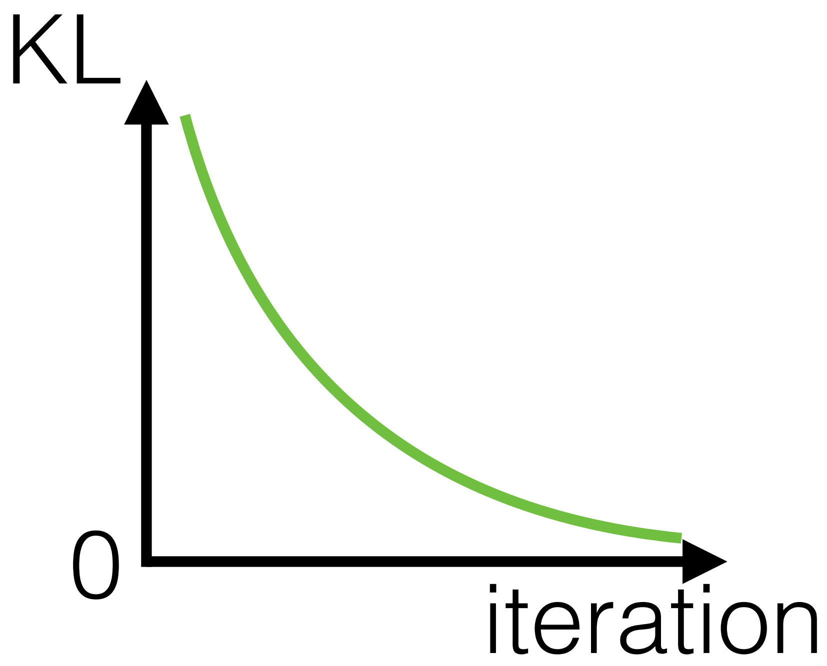
- Reliable diagnostics

What can we do?

- Reliable diagnostics
  - KL vs ELBO

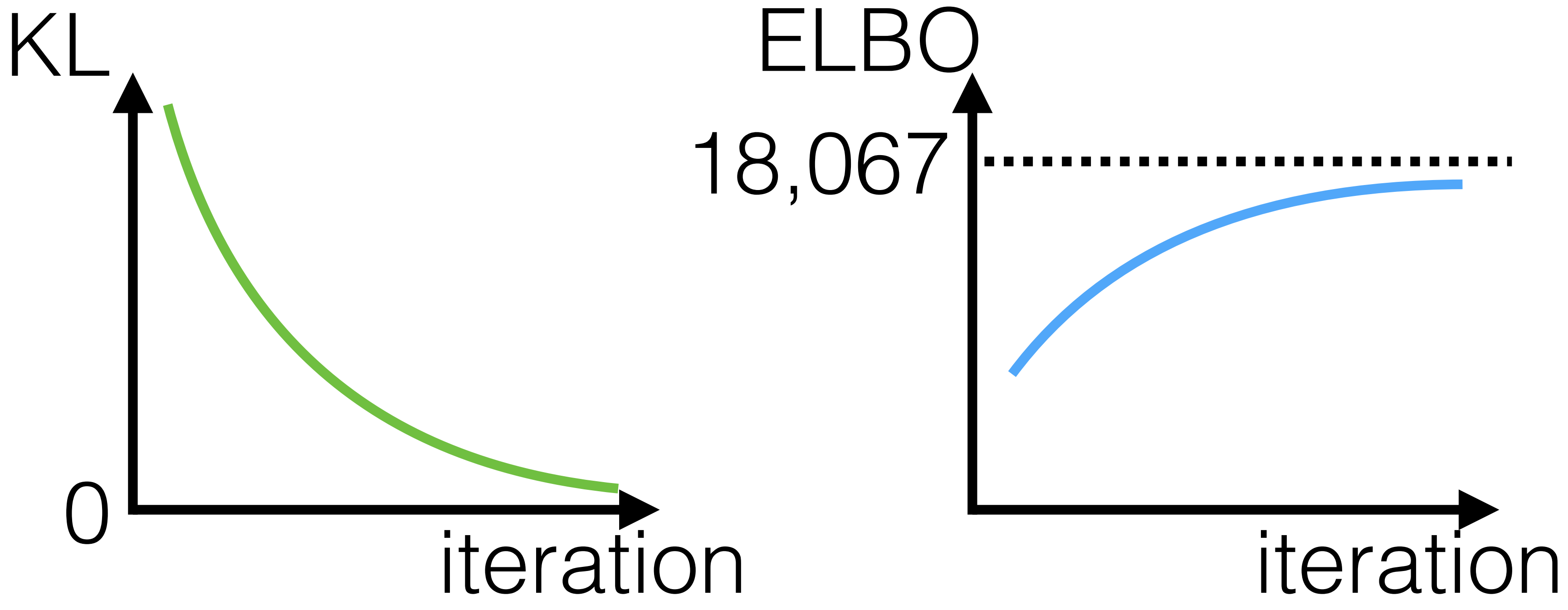
What can we do?

- Reliable diagnostics
    - KL vs ELBO



What can we do?

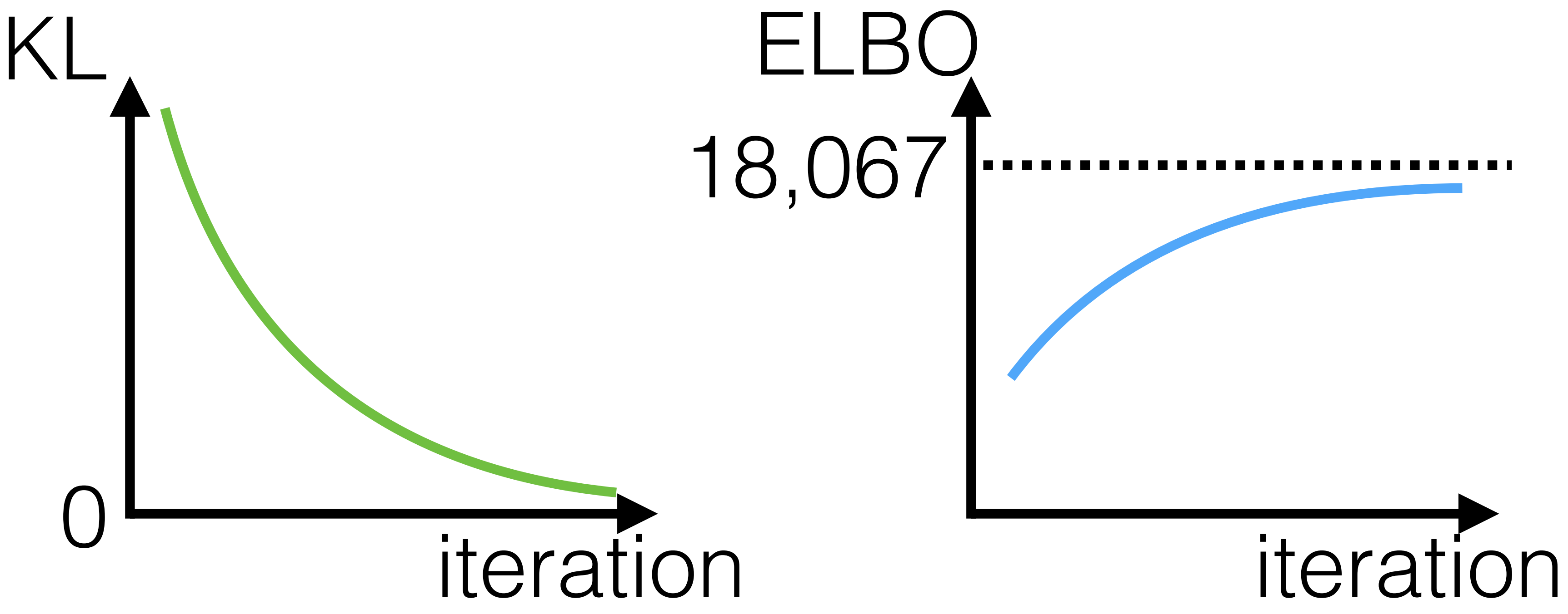
- Reliable diagnostics
    - KL vs ELBO



What can we do?

- Reliable diagnostics
  - KL vs ELBO

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

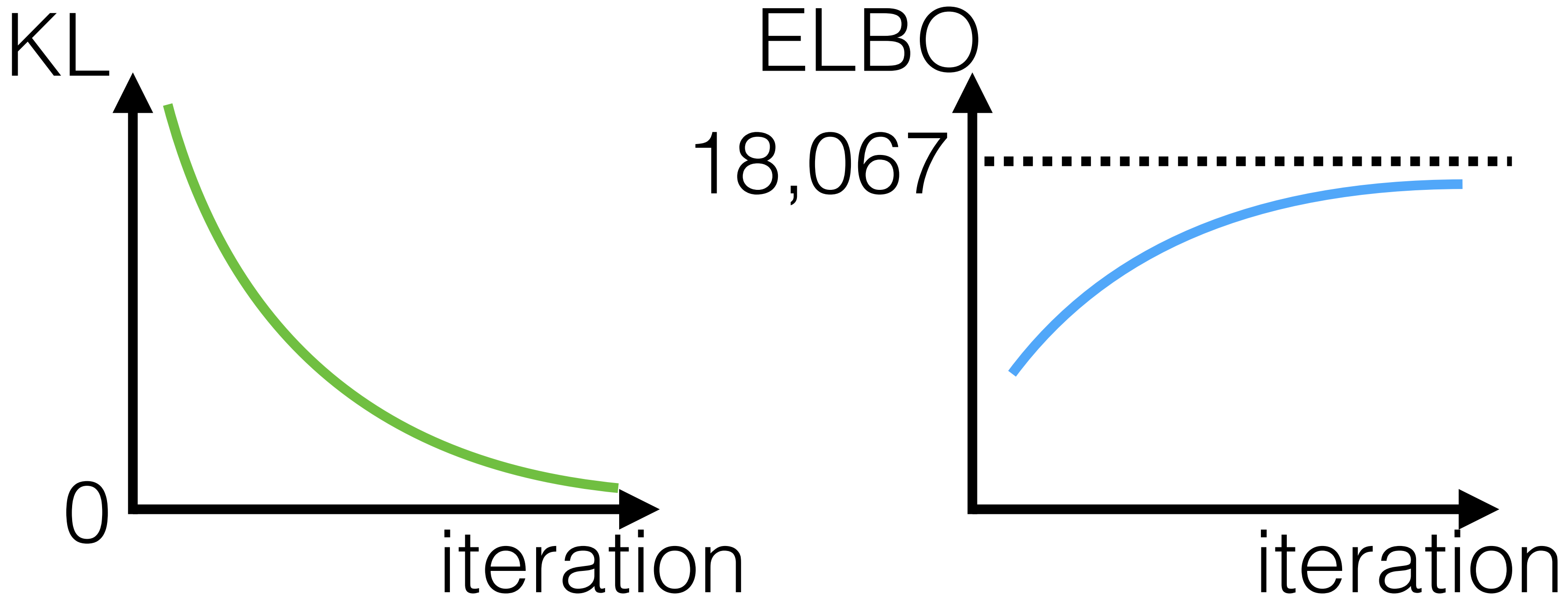


“Yes, but did it work? Evaluating variational inference” ICML Wedn 5pm

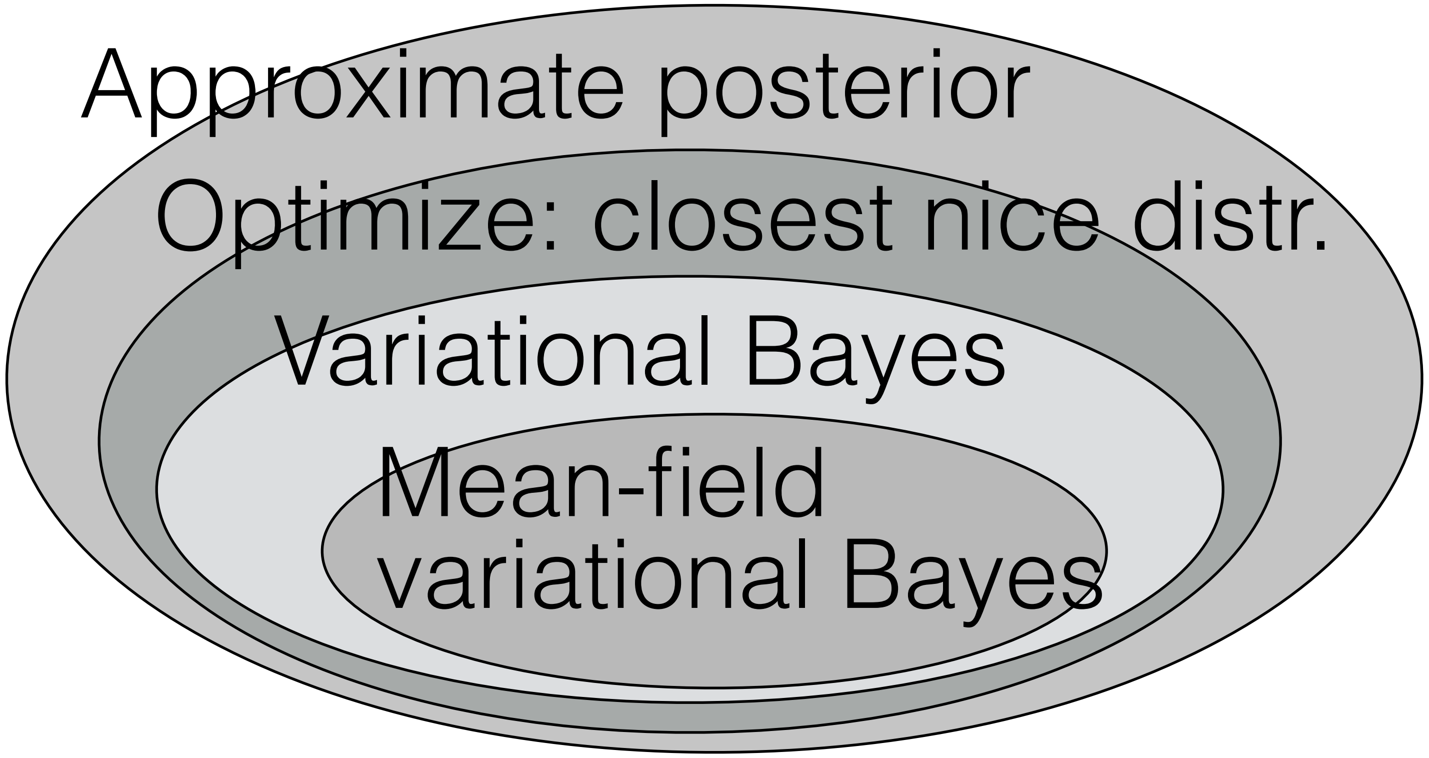
What can we do?

- Reliable diagnostics
  - KL vs 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 Wedn 5pm



Approximate posterior

Optimize: closest nice distr.

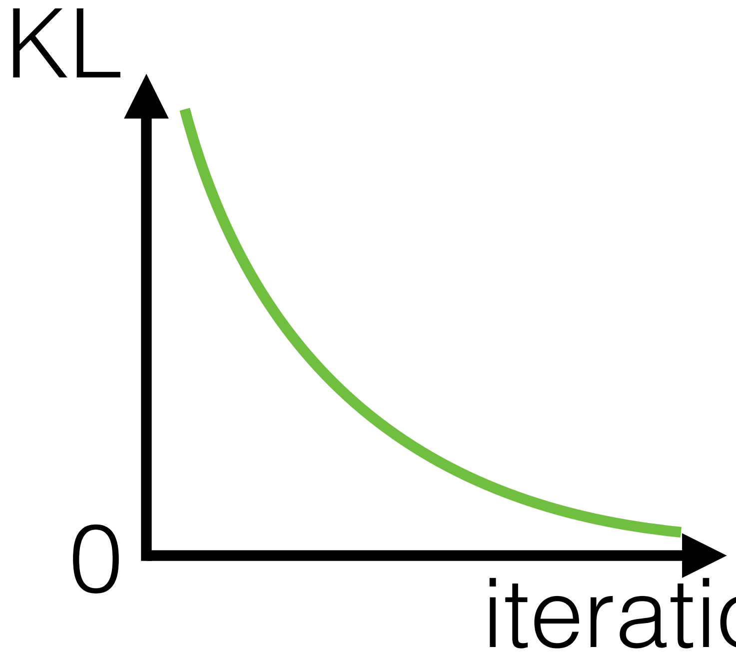
Variational Bayes

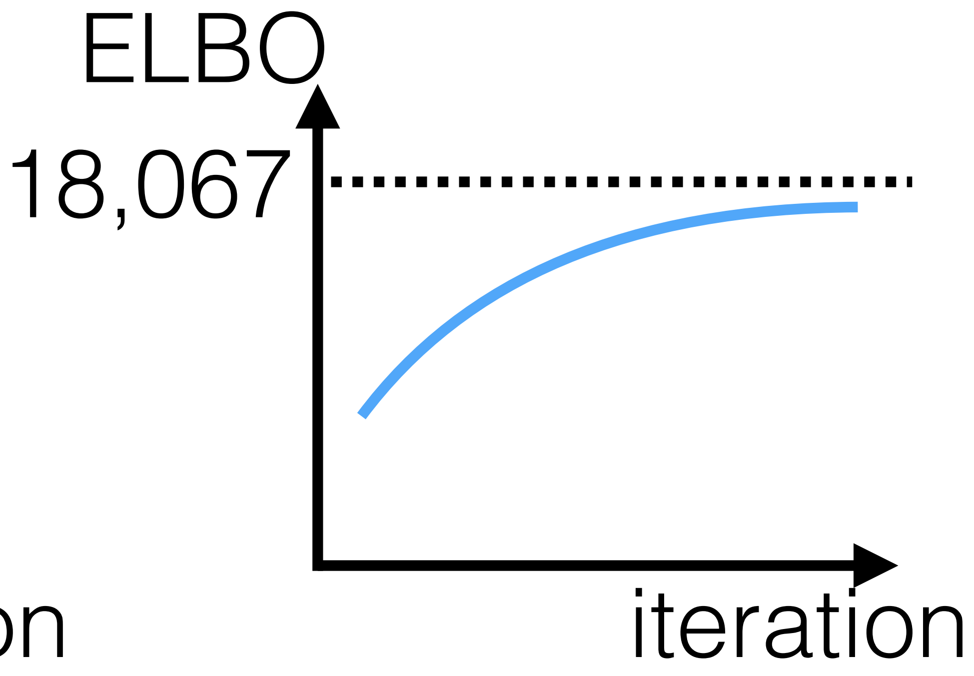
Mean-field variational Bayes

What can we do?

- Reliable diagnostics
  - KL vs ELBO

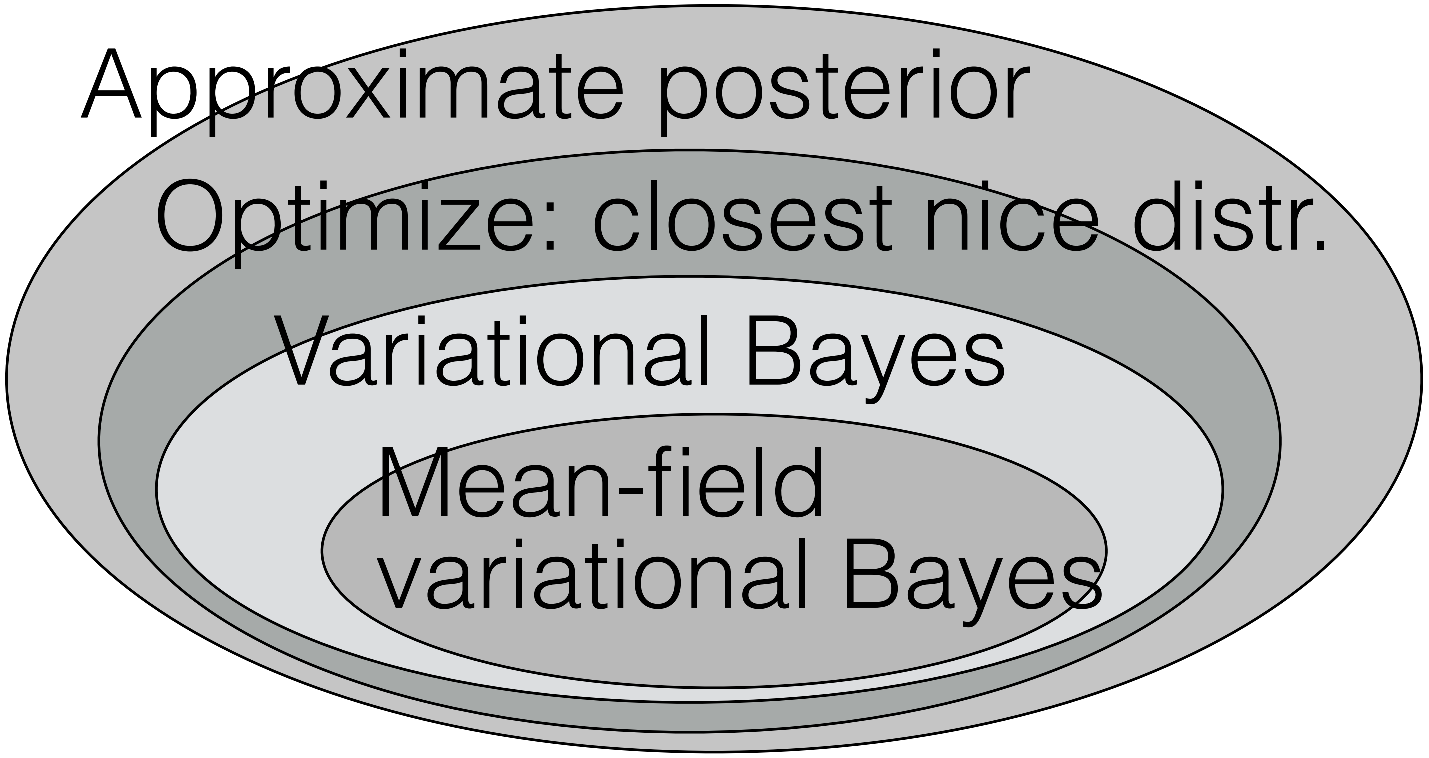
[Gorham, Mackey 2015, 2017; Chwialkowski, Strathmann, Gretton 2016; Jitkrittum et al 2017; Talts et al 2018; Yao et al 2018, etc.]





→ "Yes, but did it work? Evaluating variational inference" ICML Wedn 5pm

- Richer "nice" set; alternative divergences



Approximate posterior

Optimize: closest nice distr.

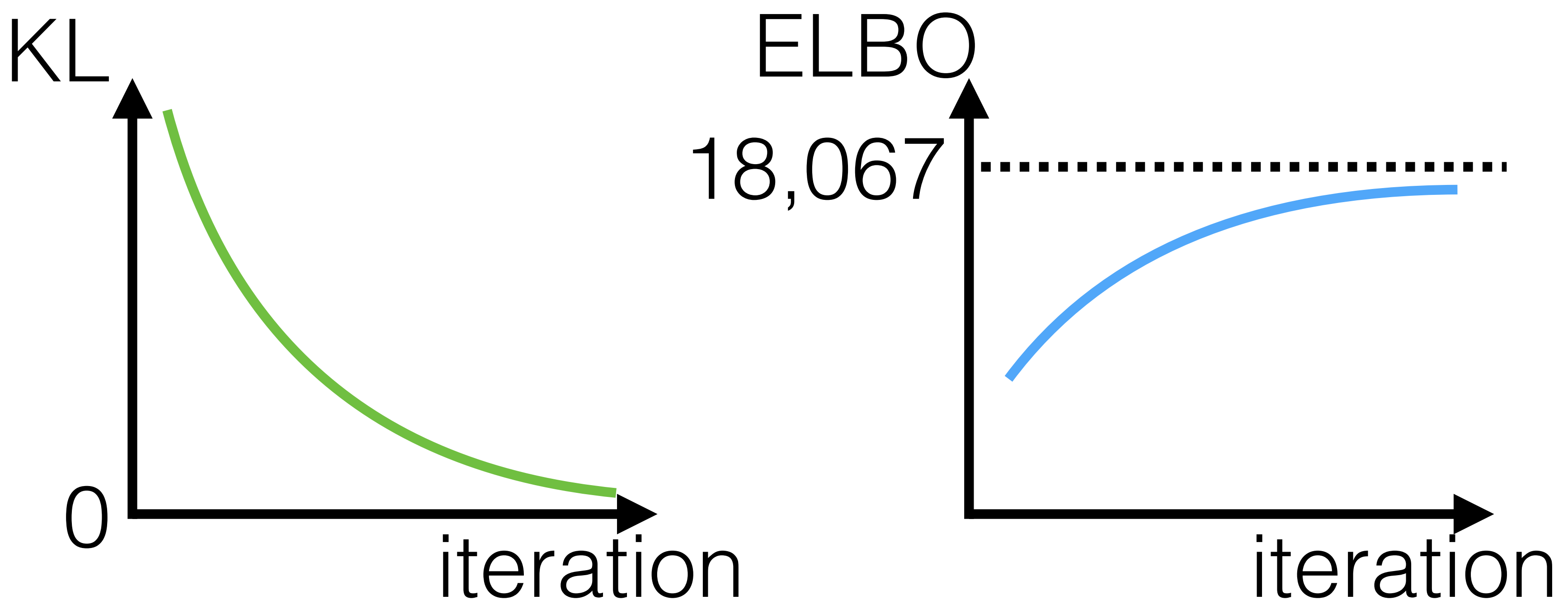
Variational Bayes

Mean-field variational Bayes

What can we do?

- Reliable diagnostics
  - KL vs ELBO

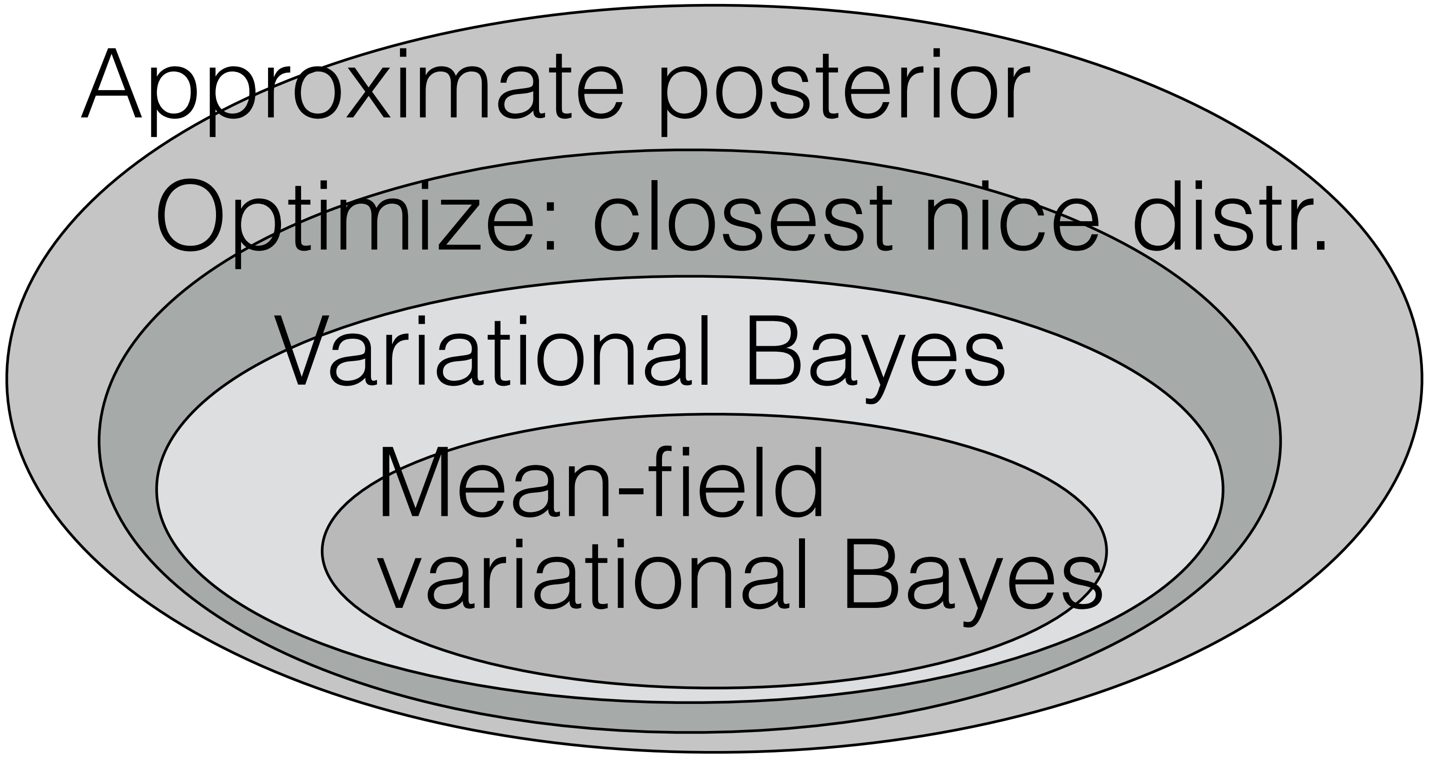
[Gorham, Mackey 2015, 2017; Chwialkowski, Strathmann, Gretton 2016; Jitkrittum et al 2017; Talts et al 2018; Yao et al 2018, etc.]



→ “Yes, but did it work? Evaluating variational inference” ICML Wedn 5pm

- Richer “nice” set; alternative divergences

[Turner, Sahani 2011]



Approximate posterior

Optimize: closest nice distr.

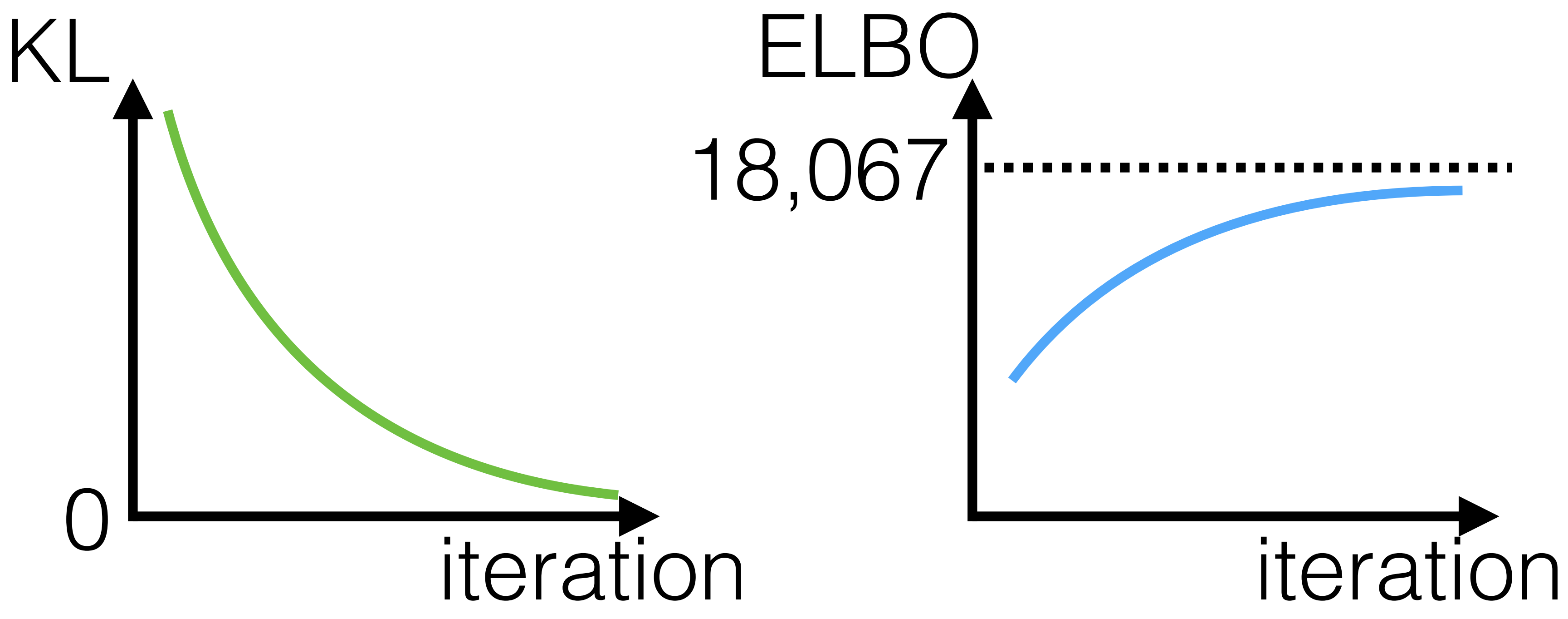
Variational Bayes

Mean-field variational Bayes

What can we do?

- Reliable diagnostics
  - KL vs ELBO

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

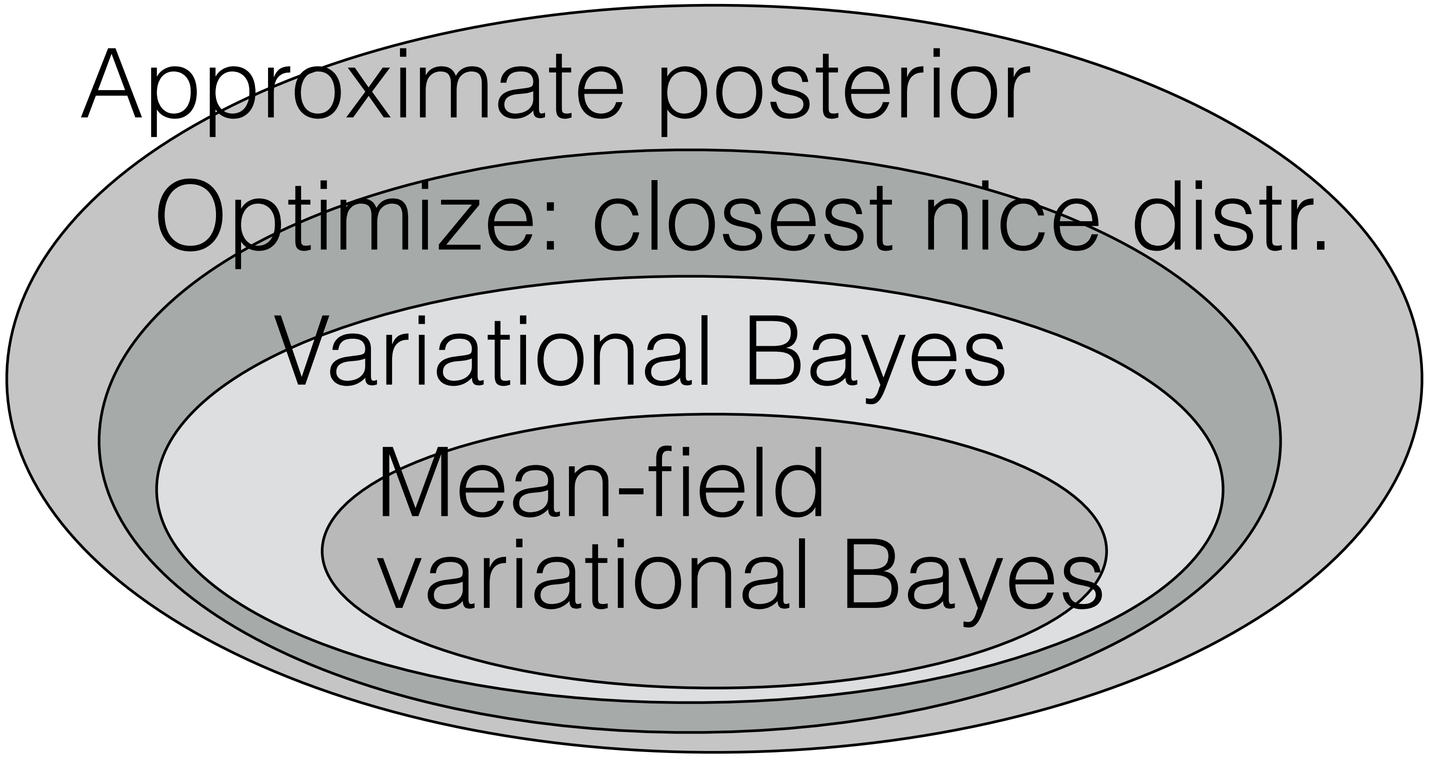


→ "Yes, but did it work? Evaluating variational inference" ICML Wedn 5pm

- Richer "nice" set; alternative divergences

[Turner, Sahani 2011]

- Theoretical guarantees on finite-data quality



Approximate posterior

Optimize: closest nice distr.

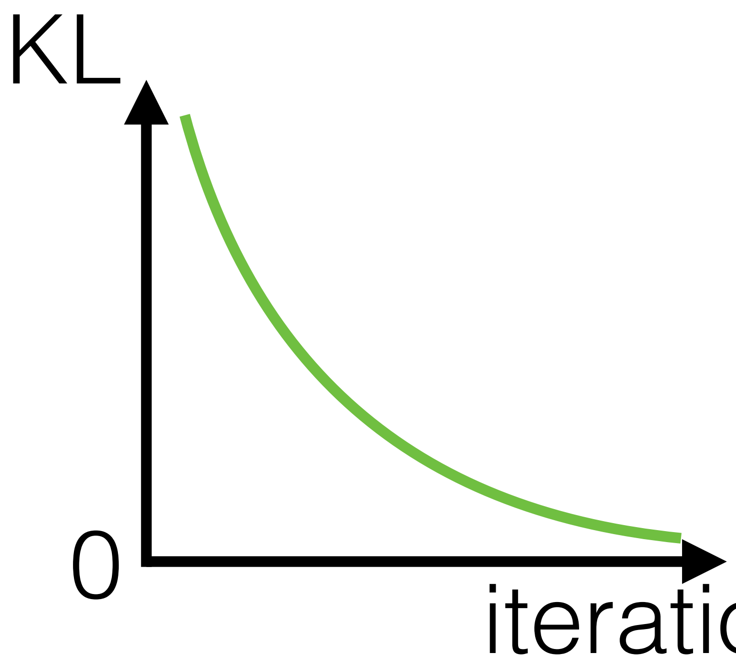
Variational Bayes

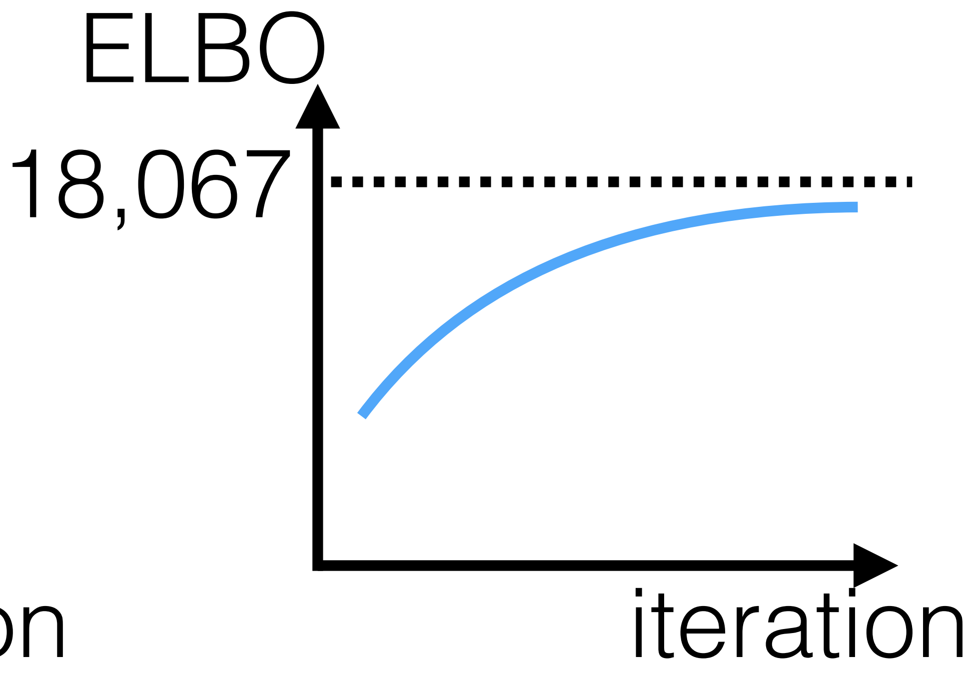
Mean-field variational Bayes

What can we do?

- Reliable diagnostics
  - KL vs ELBO

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

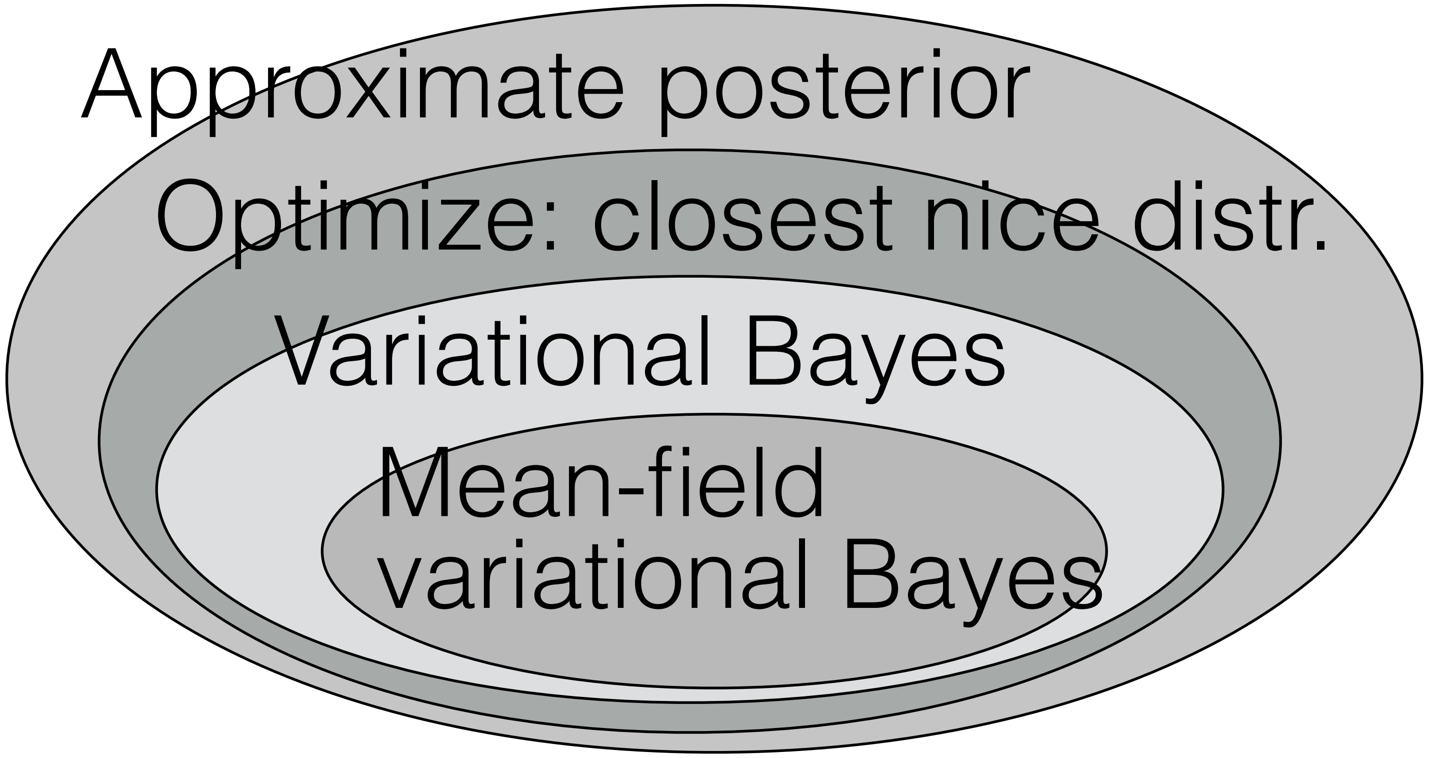




→ “Yes, but did it work? Evaluating variational inference” ICML Wedn 5pm

- Richer “nice” set; alternative divergences

[Turner, Sahani 2011]



Approximate posterior

Optimize: closest nice distr.

Variational Bayes

Mean-field variational Bayes

- Theoretical guarantees on finite-data quality
  - Data summarization

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.
- More Experiments
  - RJ Giordano, T Broderick, and MI Jordan. Linear response methods for accurate covariance estimates from mean field variational Bayes. *NIPS* 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, 2017. Under review. ArXiv:1709.02536.

See end of Part II slides for full reference list.