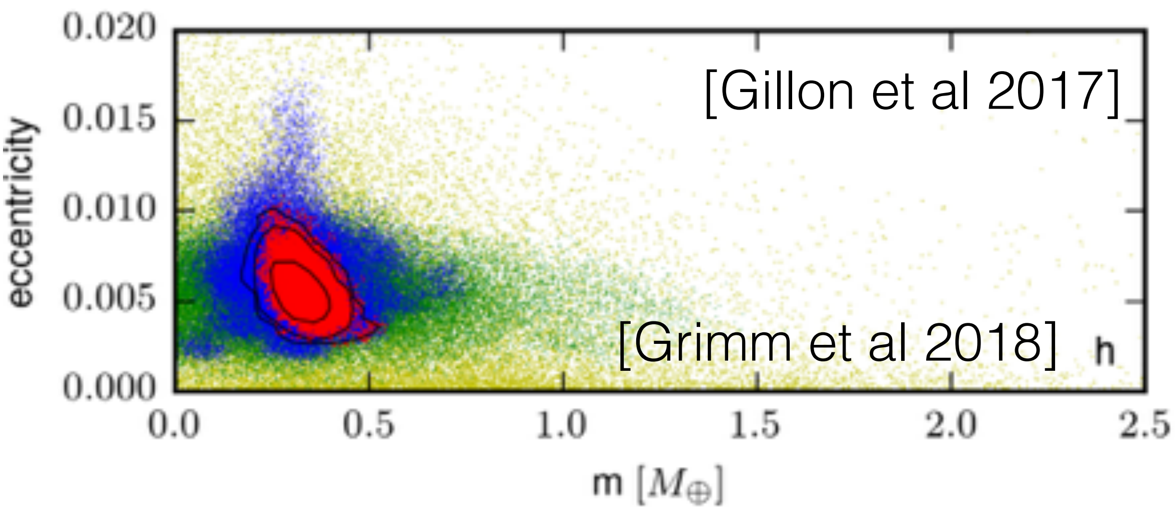
Variational Bayes and beyond: Bayesian inference for big data

Tamara Broderick
ITT Career Development
Assistant Professor,
MIT

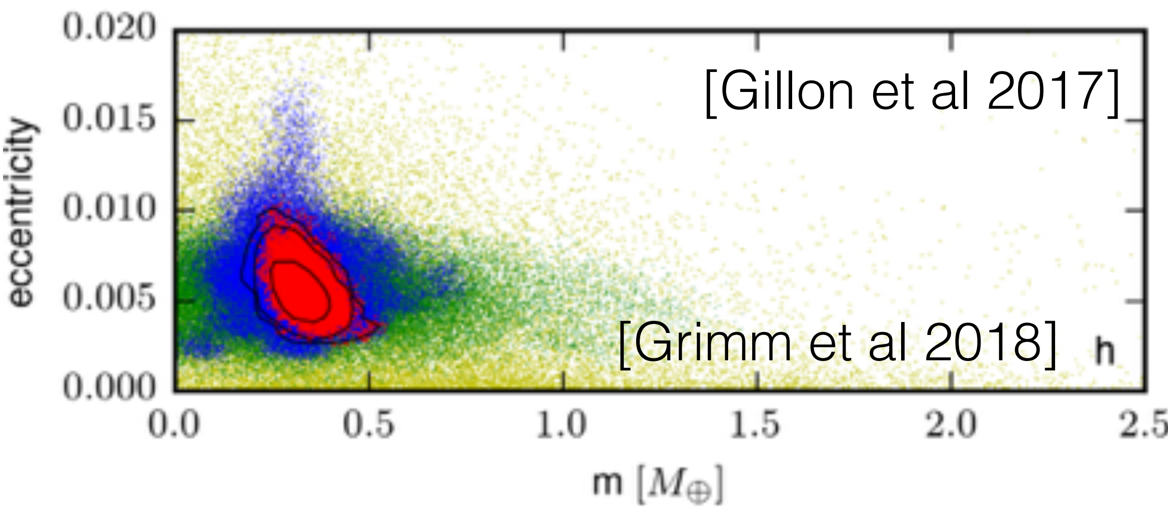
http://www.tamarabroderick.com/tutorial_2018_lugano.html

Bayesian inference

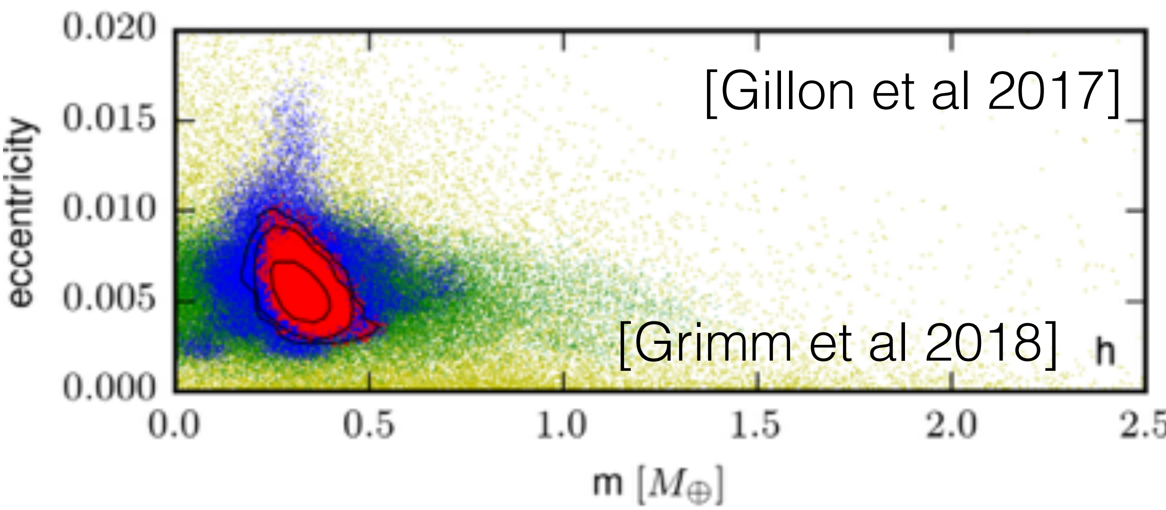
Bayesian inference

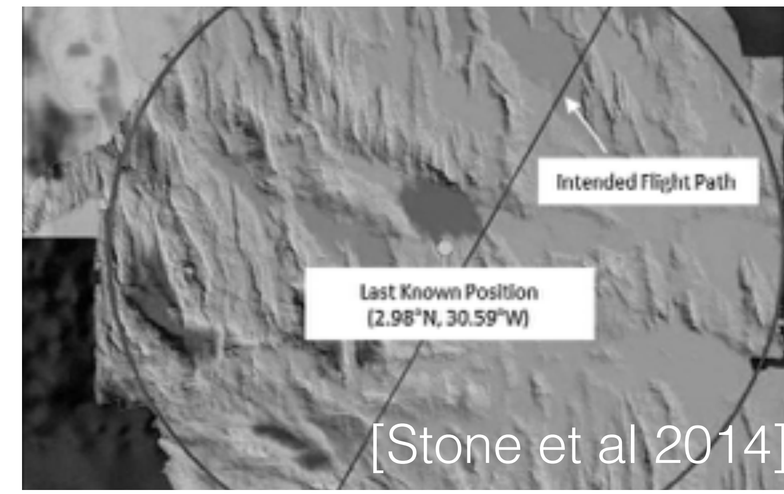


Bayesian inference

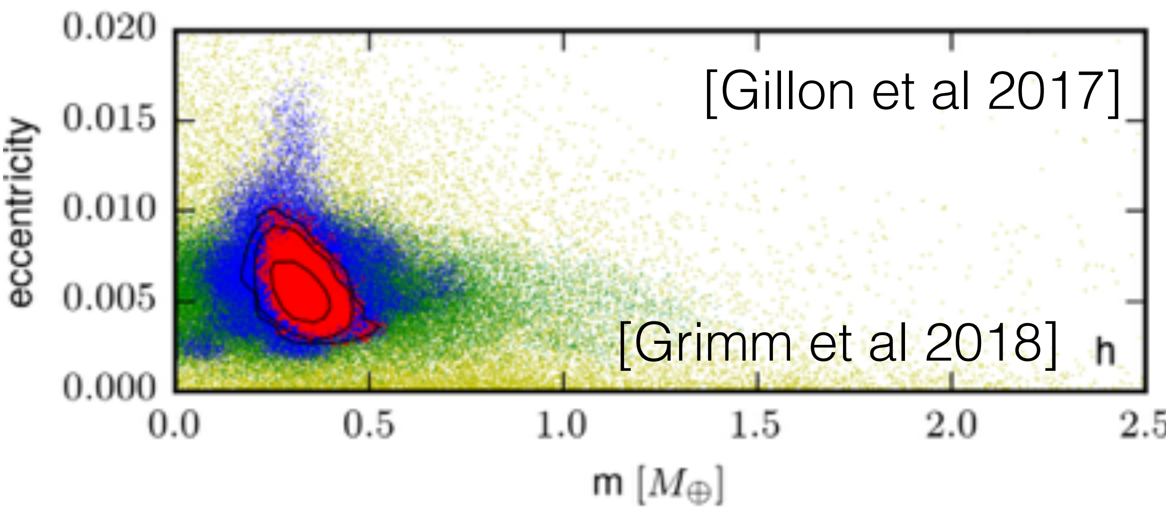


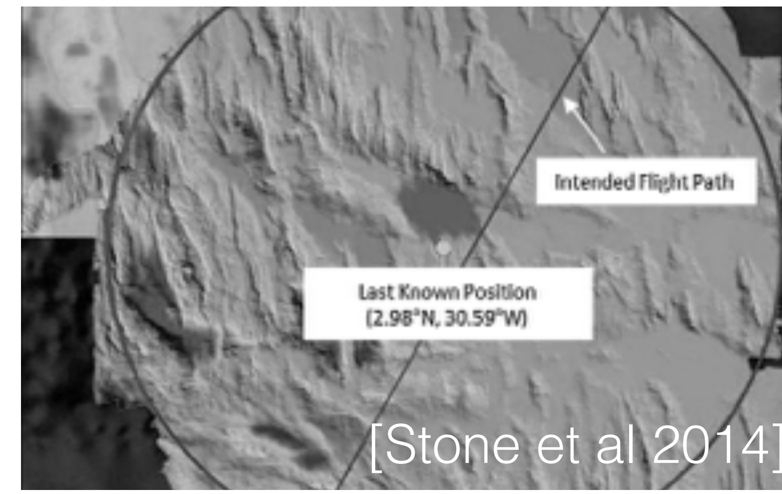
Bayesian inference

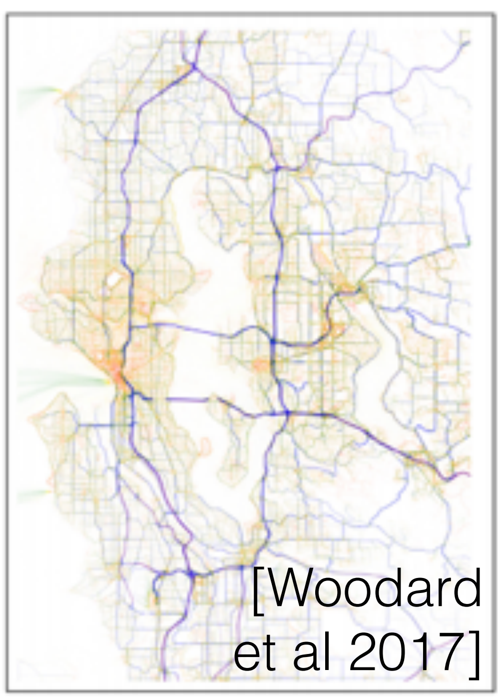




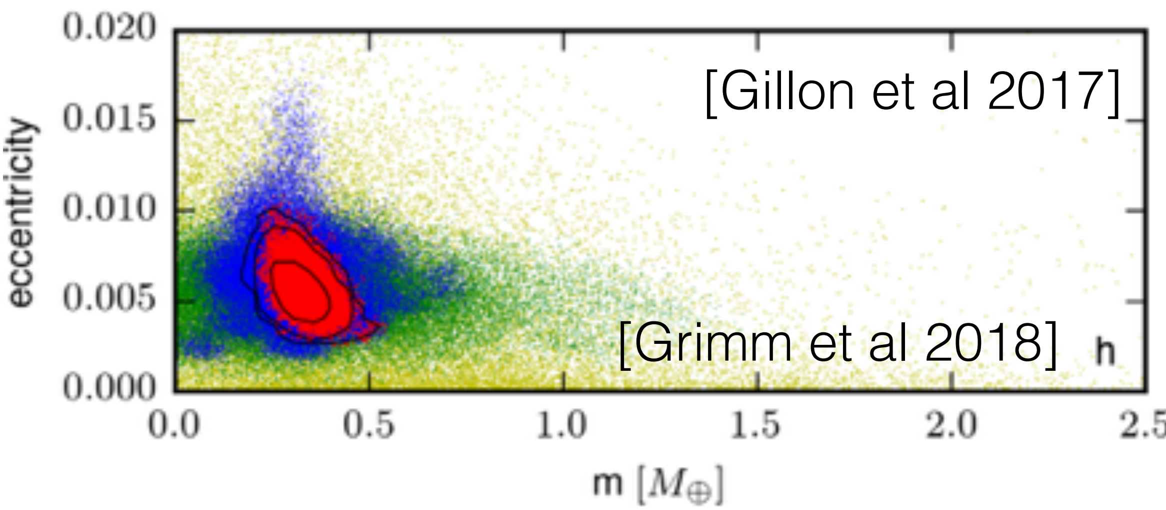
Bayesian inference

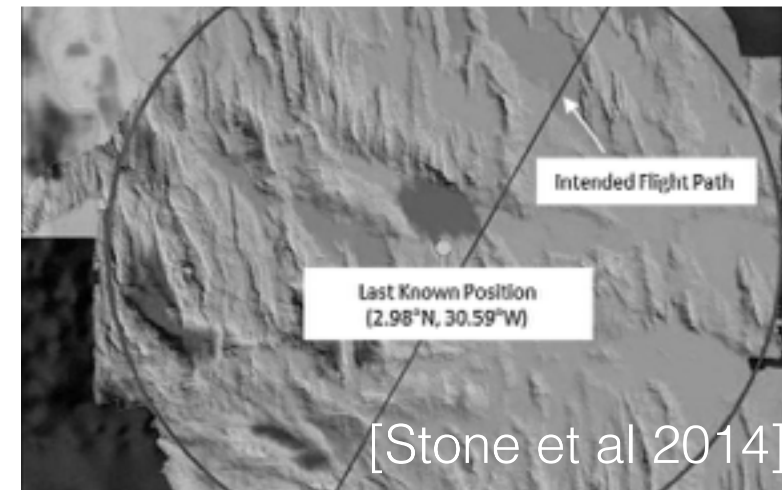


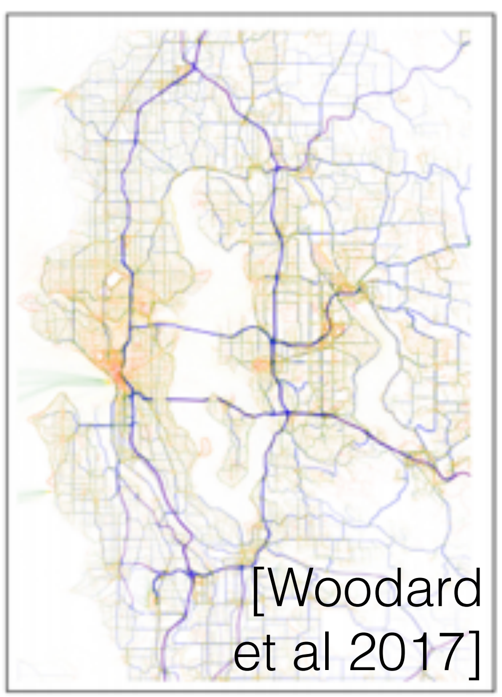


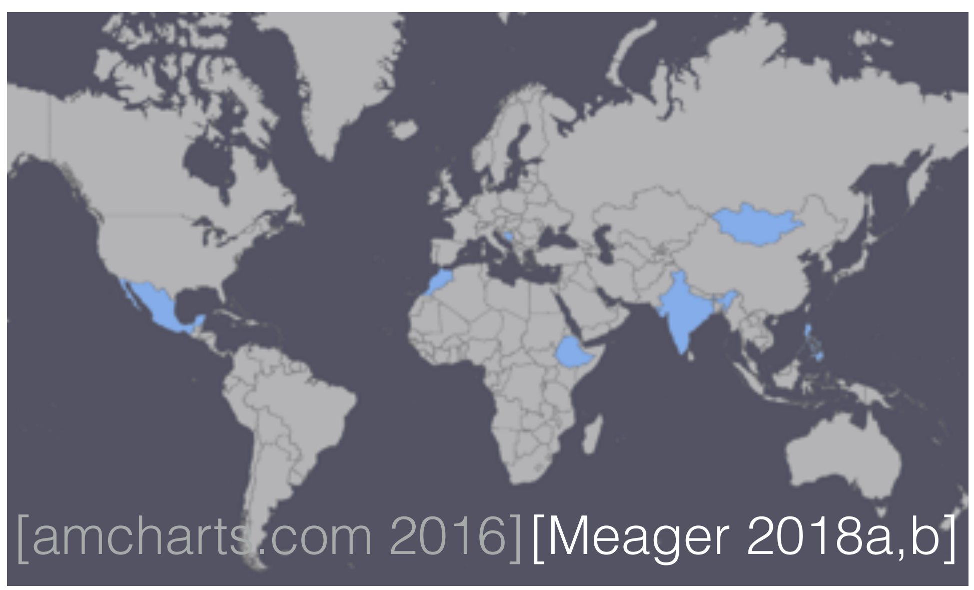


Bayesian inference

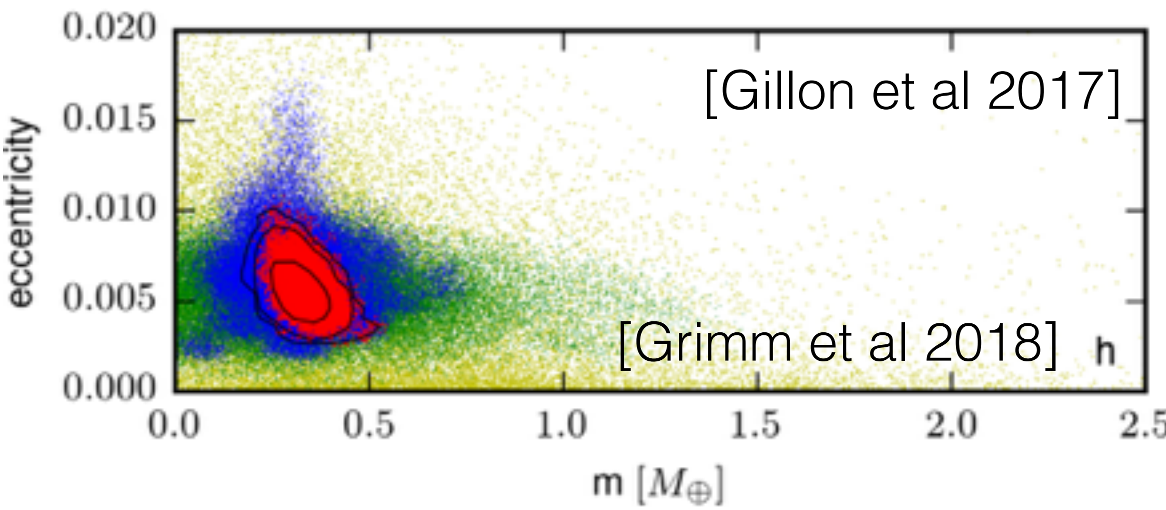


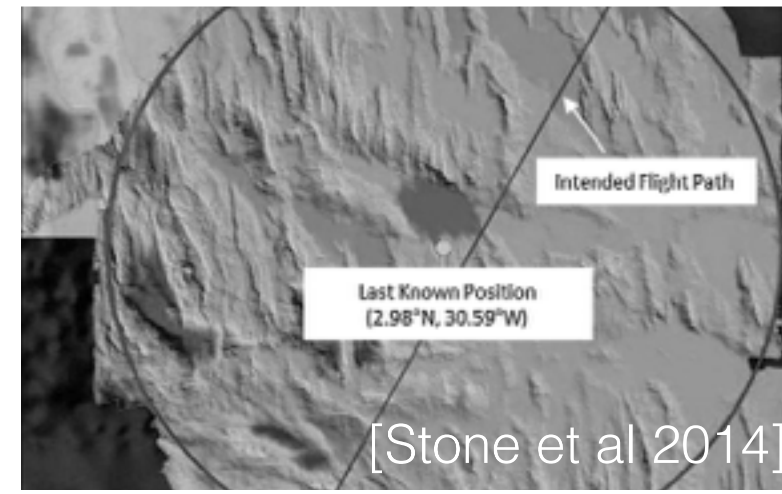


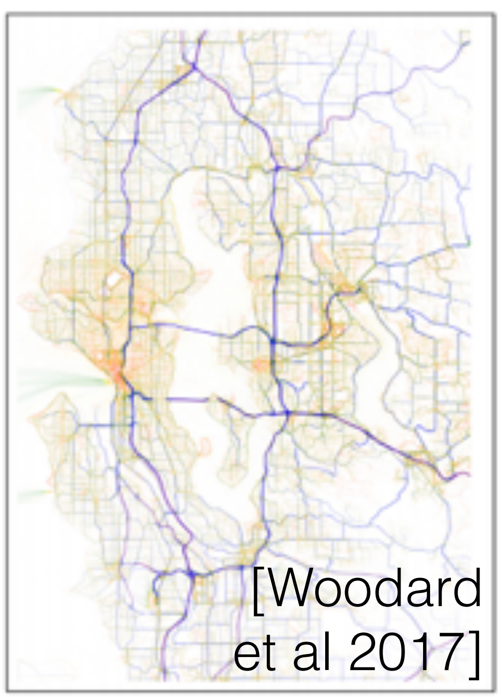


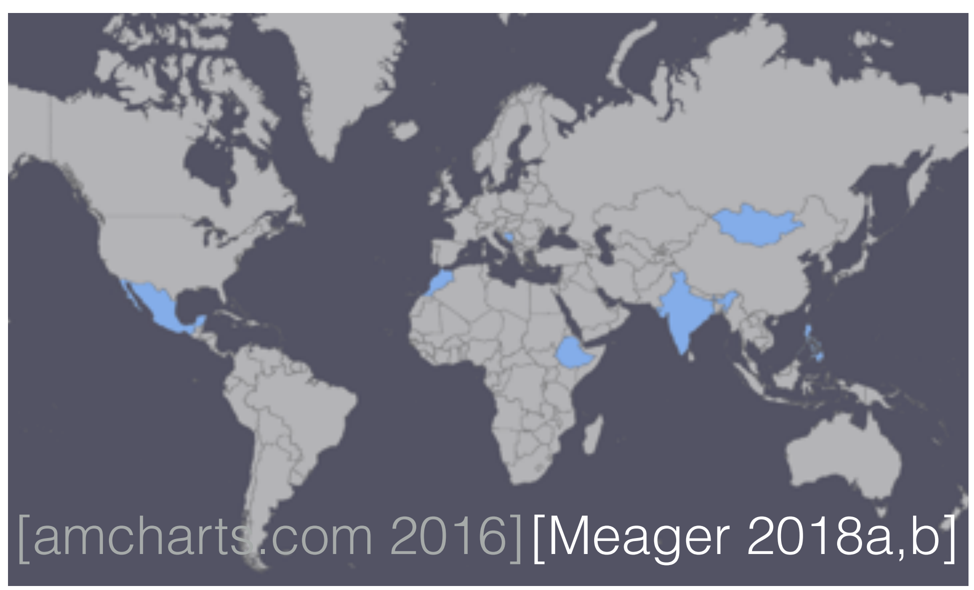


Bayesian inference

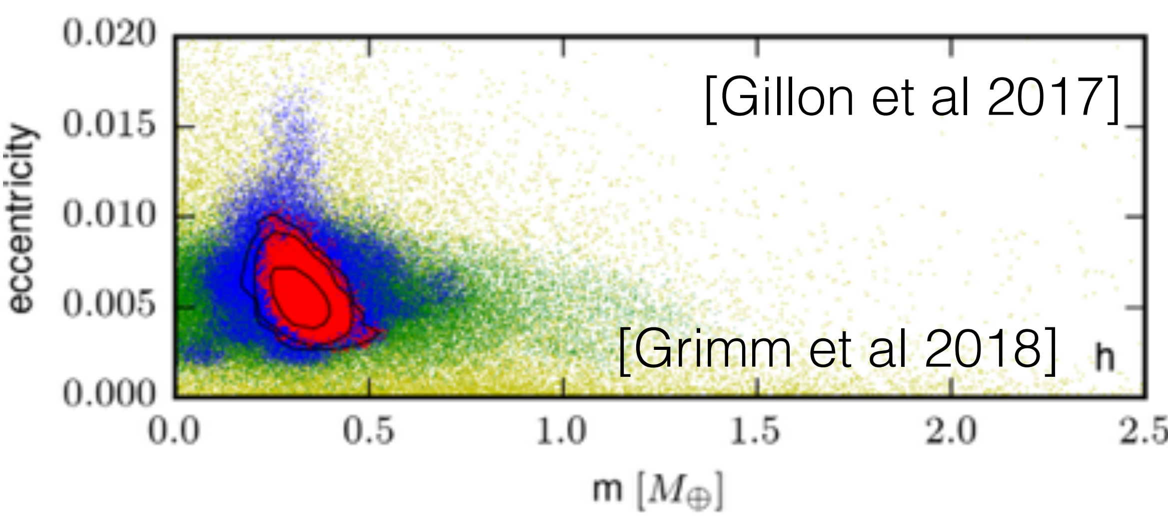


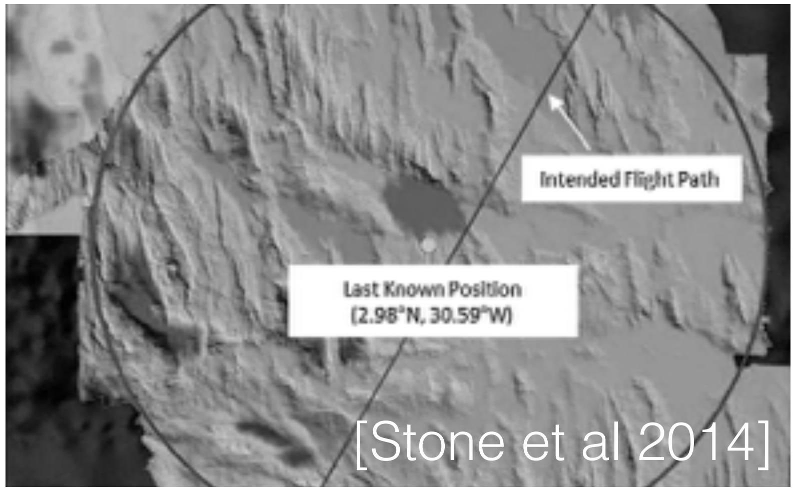


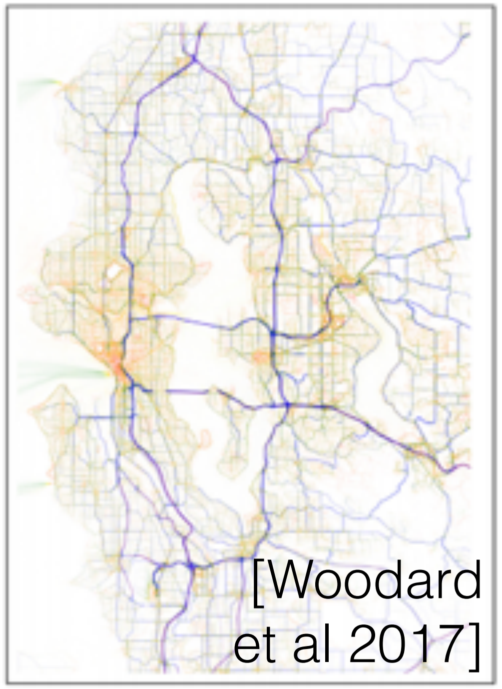


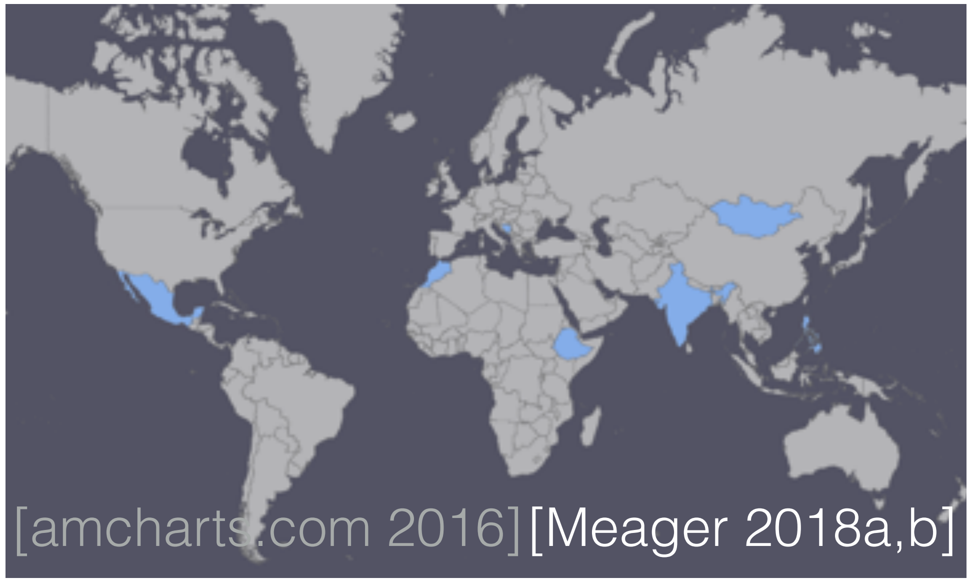


Bayesian inference



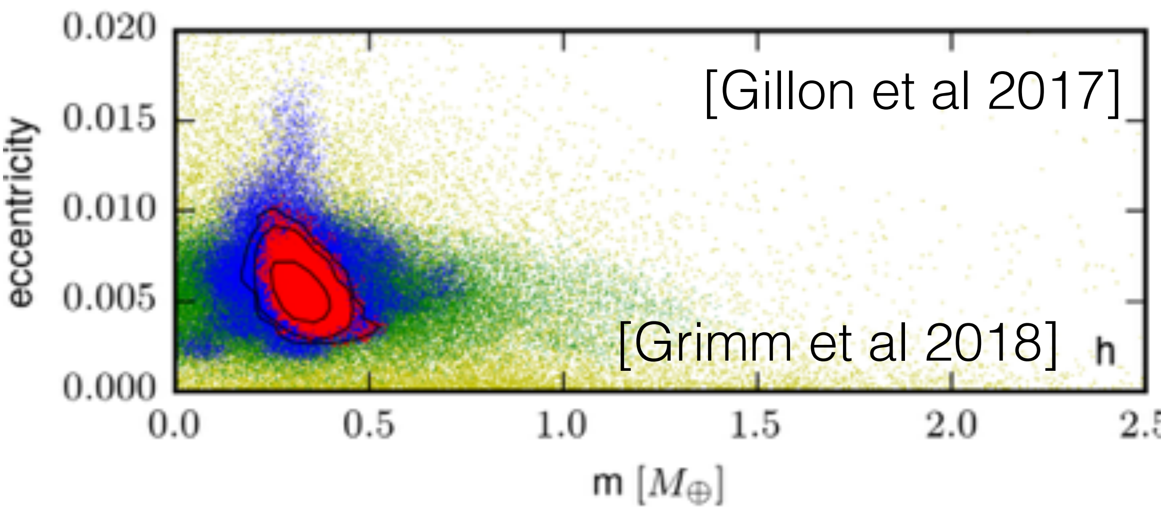


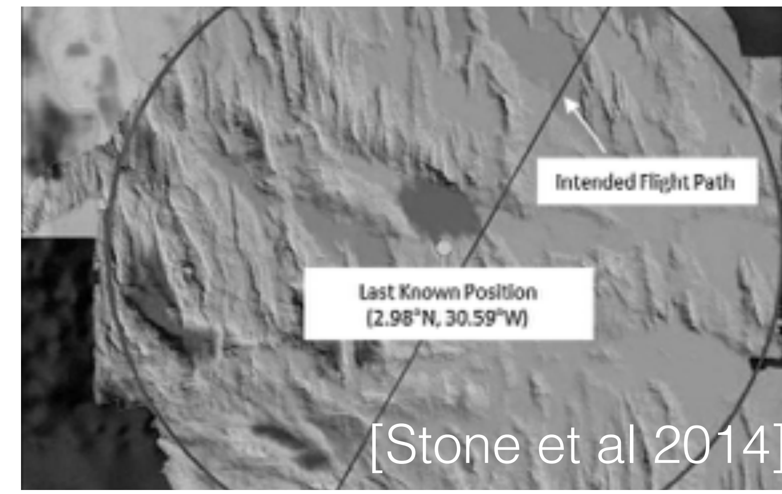


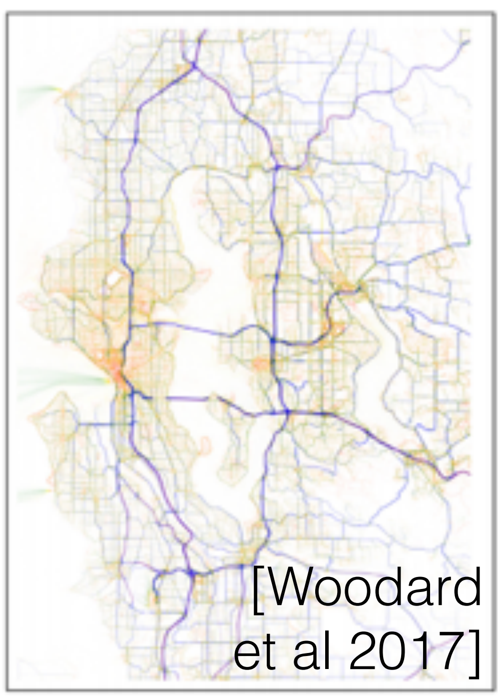


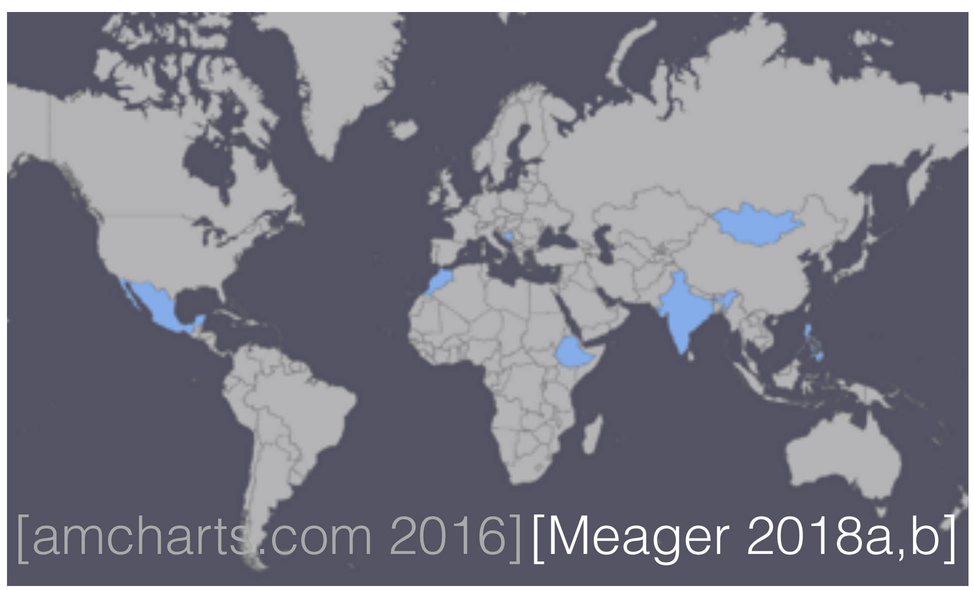
Bayesian inference

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



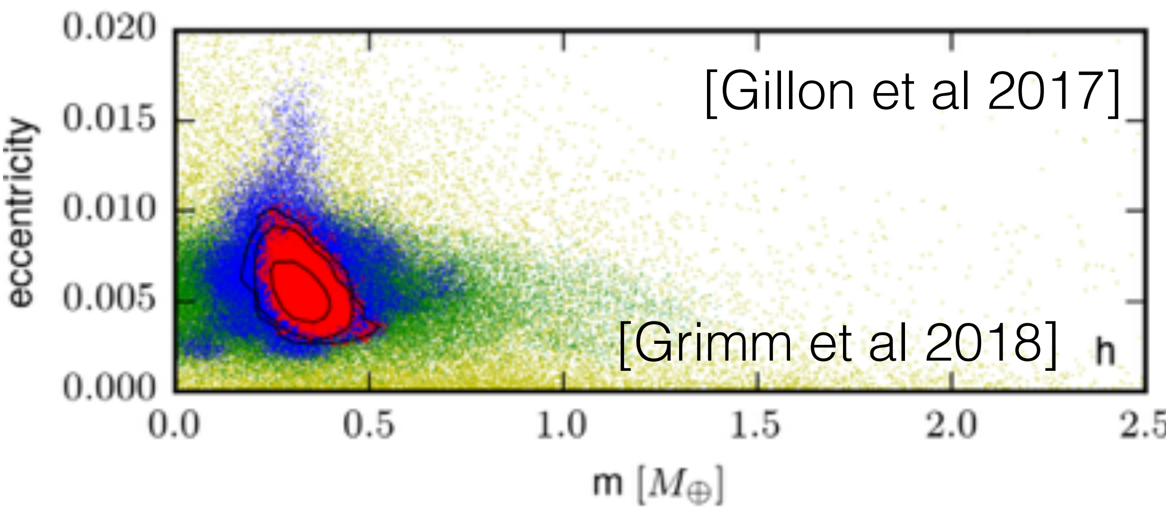


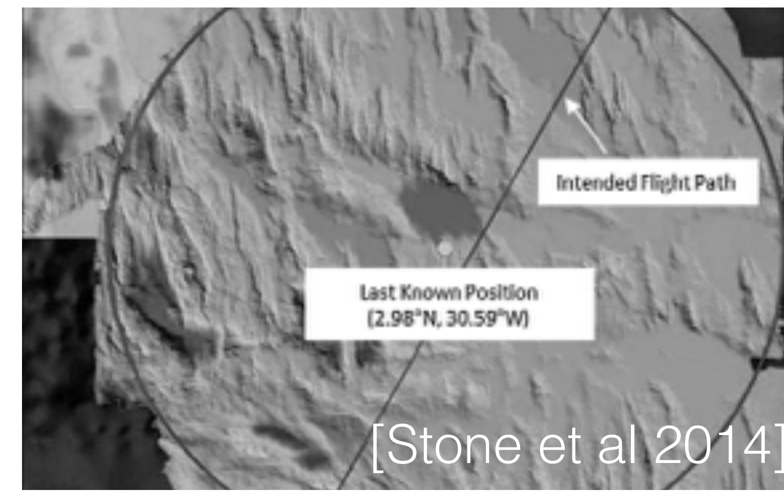


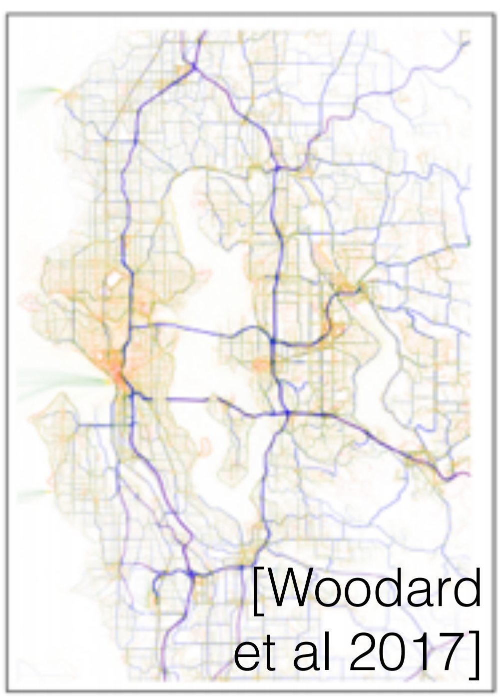


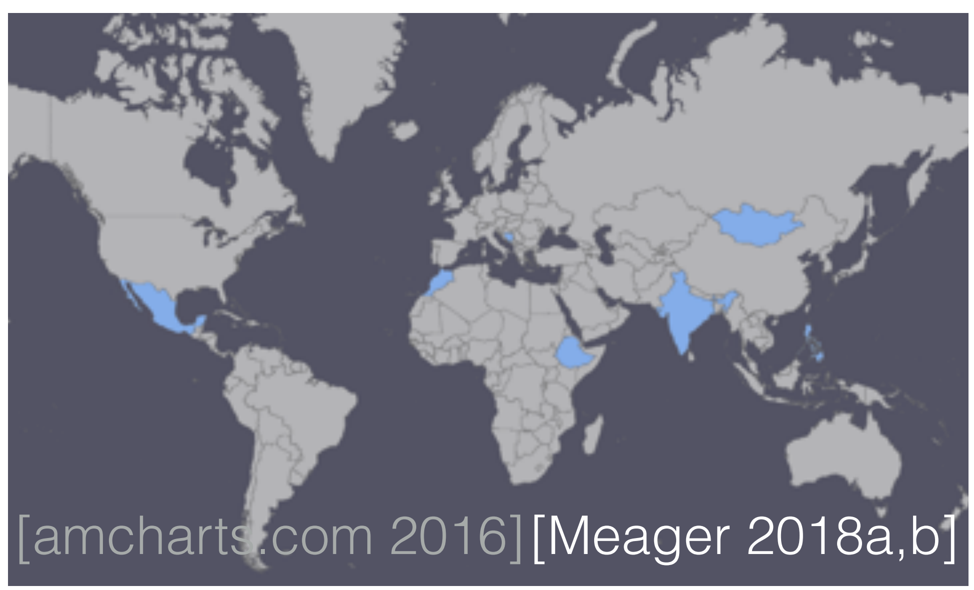
Bayesian inference

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





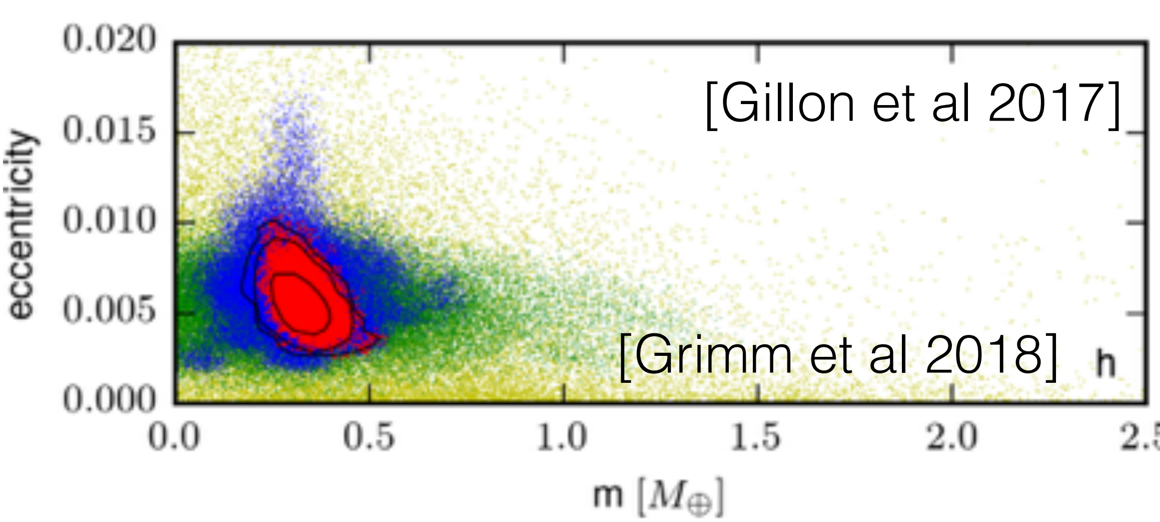


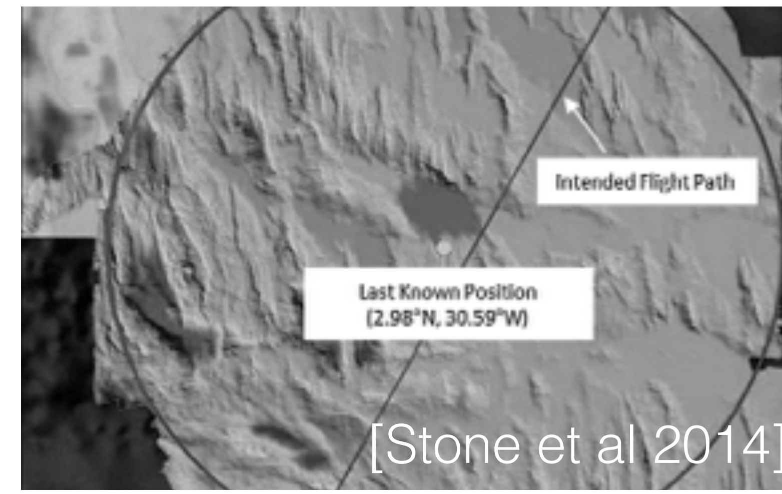


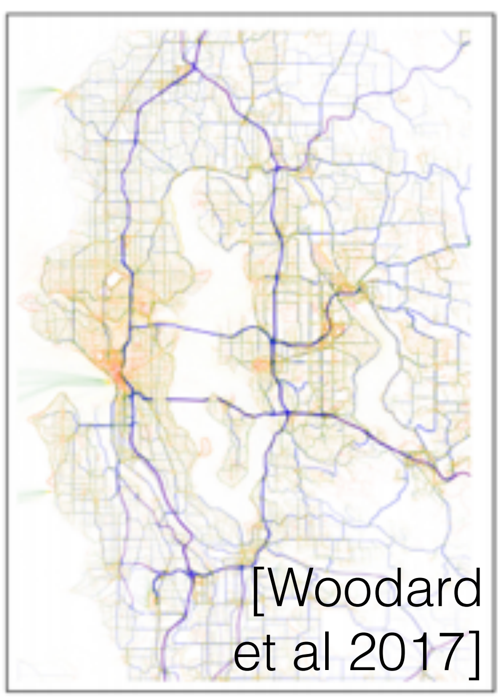
[mc-stan.org]

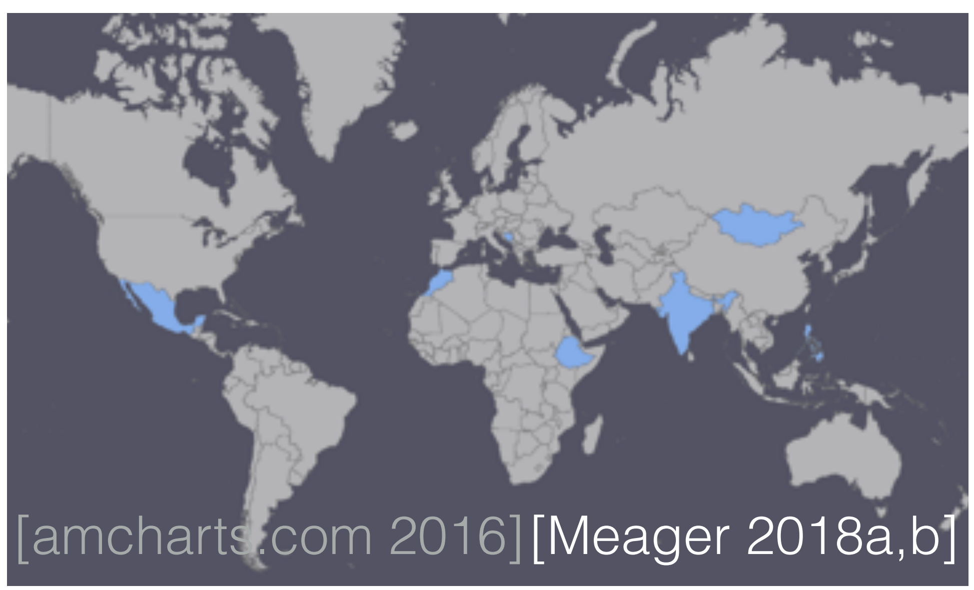
Bayesian inference

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







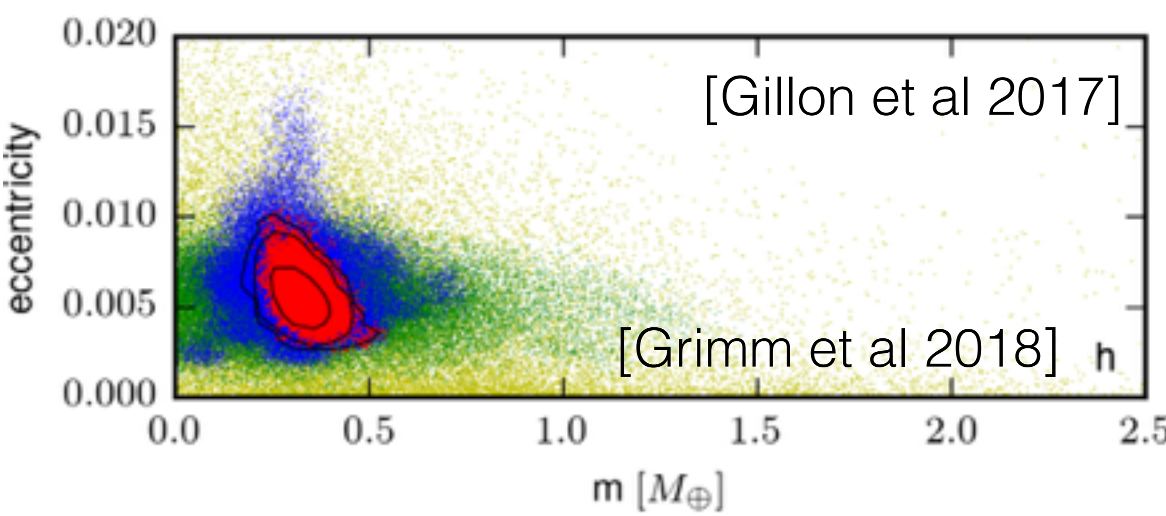


[mc-stan.org]

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

Bayesian inference

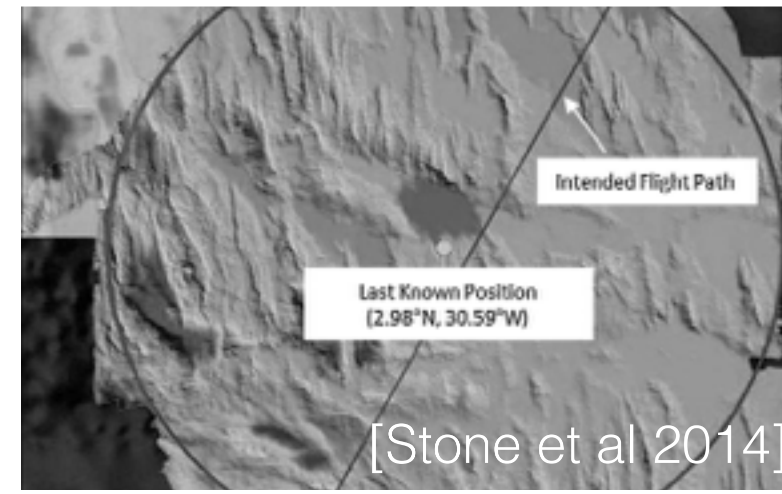
- Analysis goals: Point estimates, coherent uncertainties
    - Interpretable, complex, modular; expert information



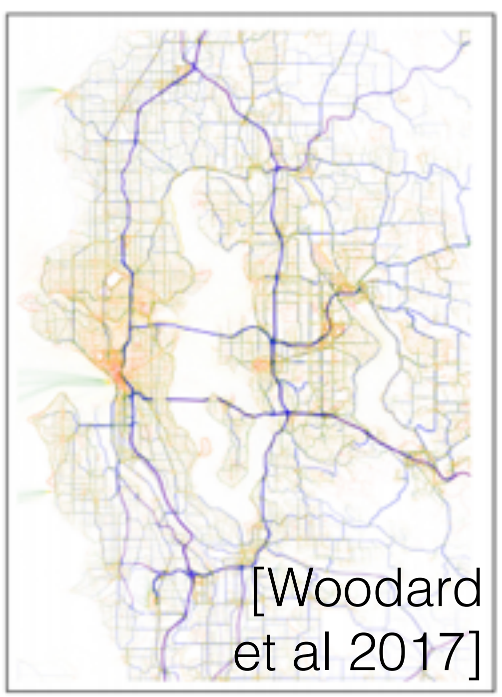
[Gillon et al 2017]

[Grimm et al 2018]

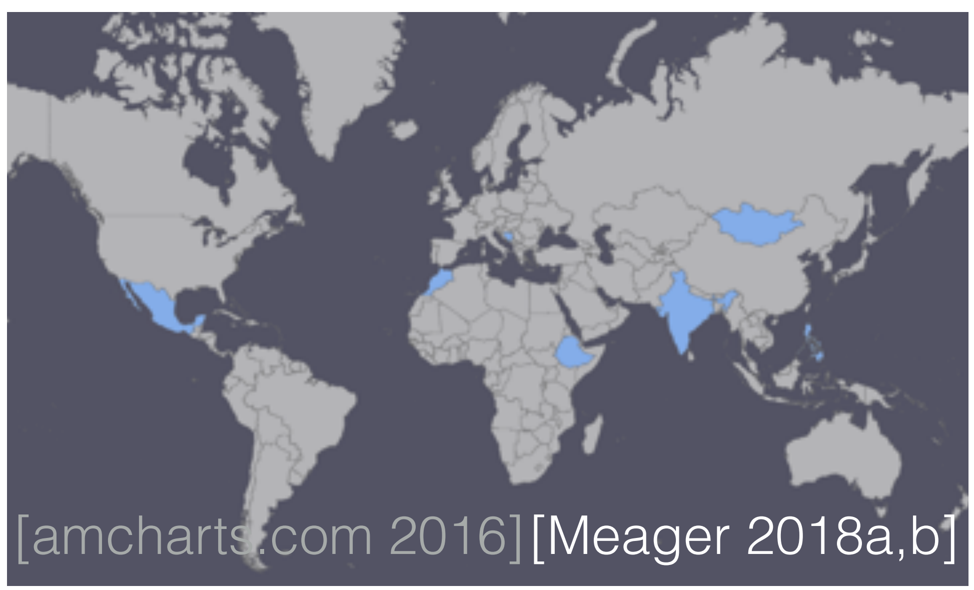
[ESO/ L. Calçada/ M. Kornmesser 2017] [Abbott et al 2016a,b]



[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.

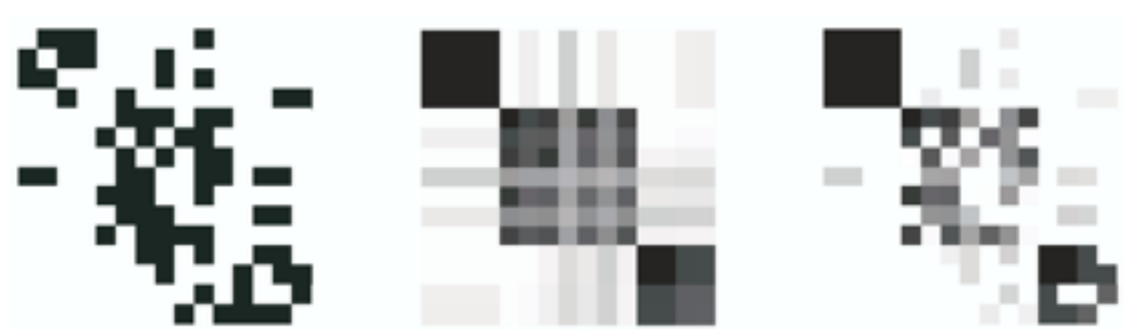
Variational Bayes

- 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]

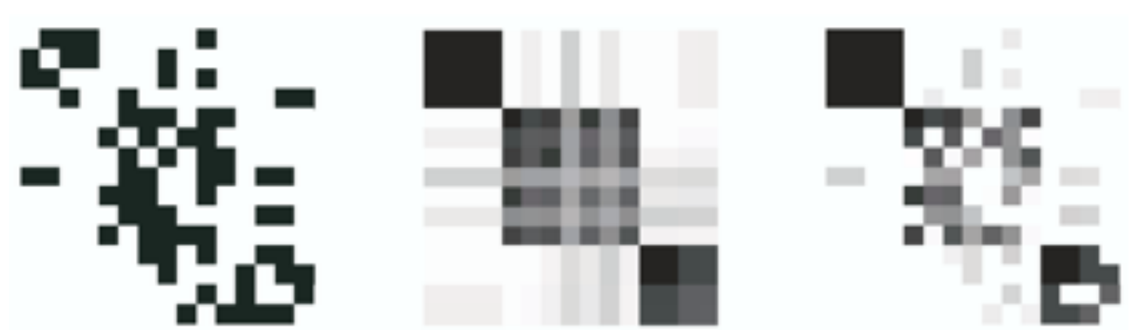
Variational Bayes

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

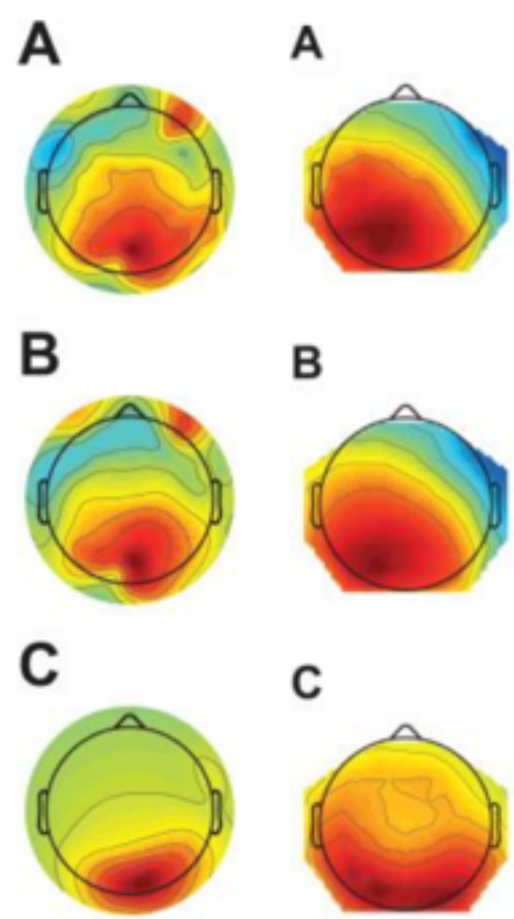
"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY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

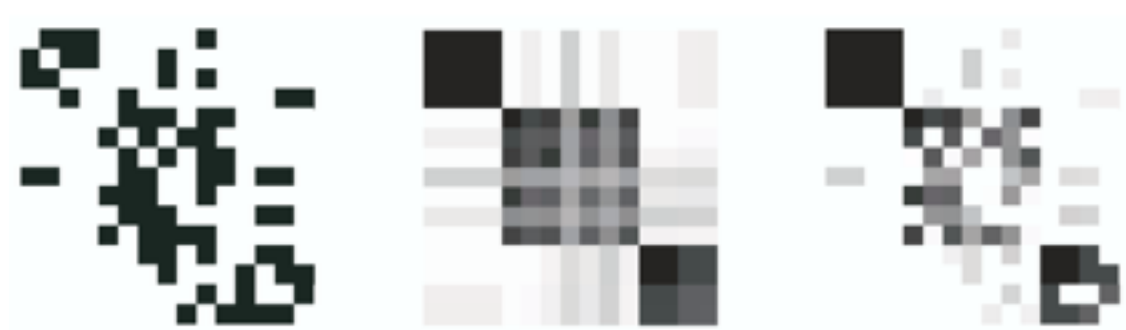
Variational Bayes

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

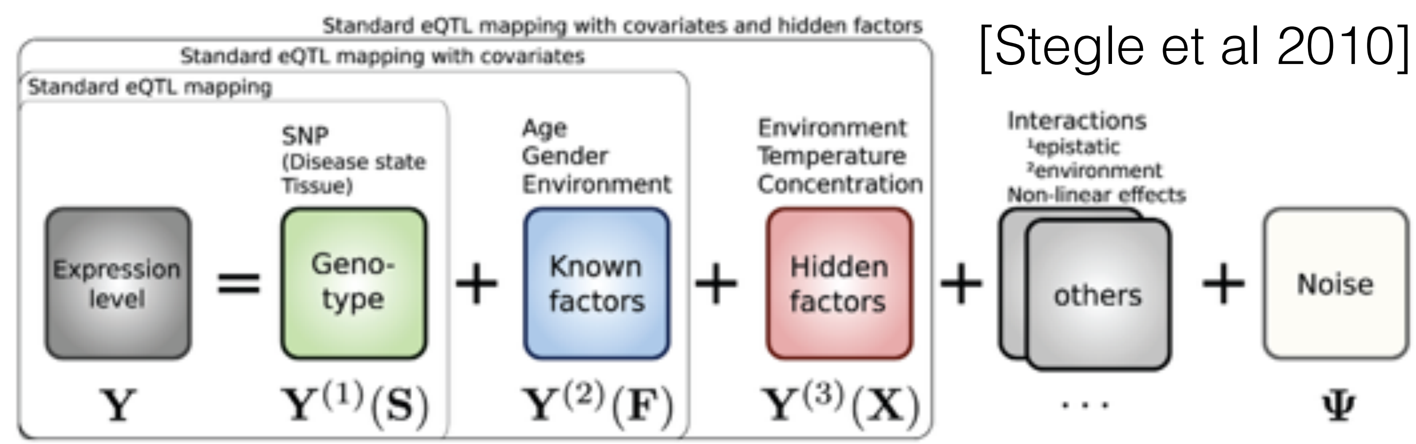
"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY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

[Blei et al 2003]

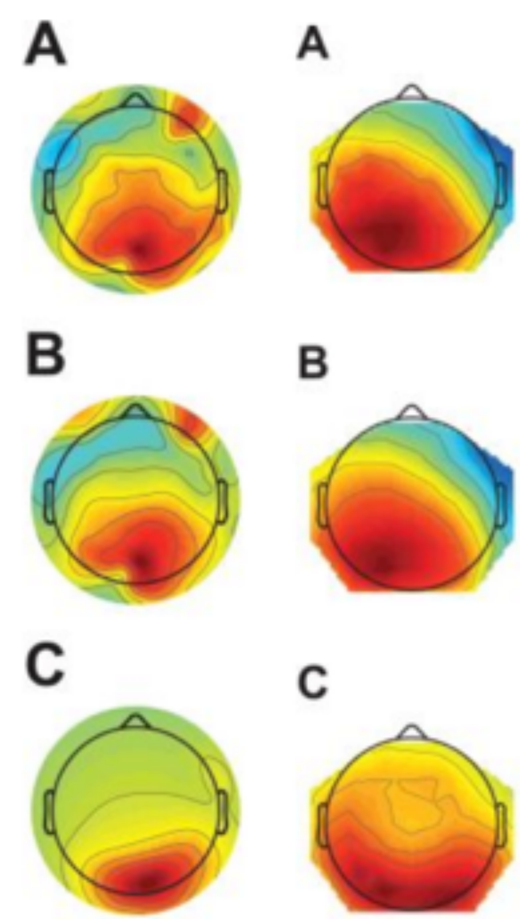
The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our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

[Blei et al 2018]

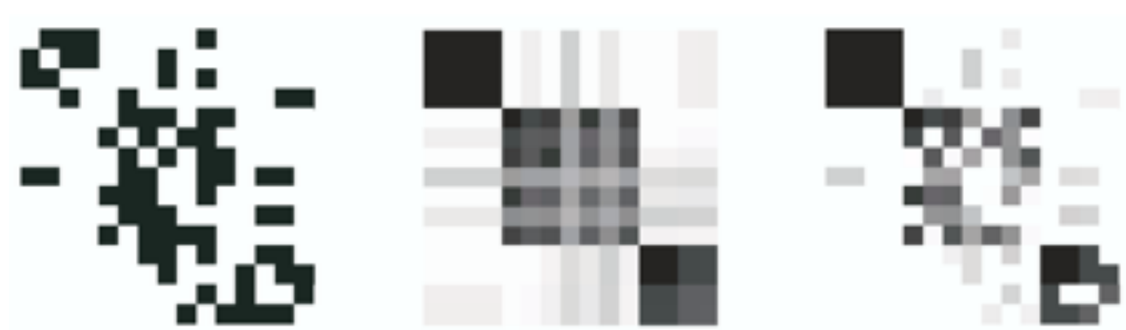
Variational Bayes

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

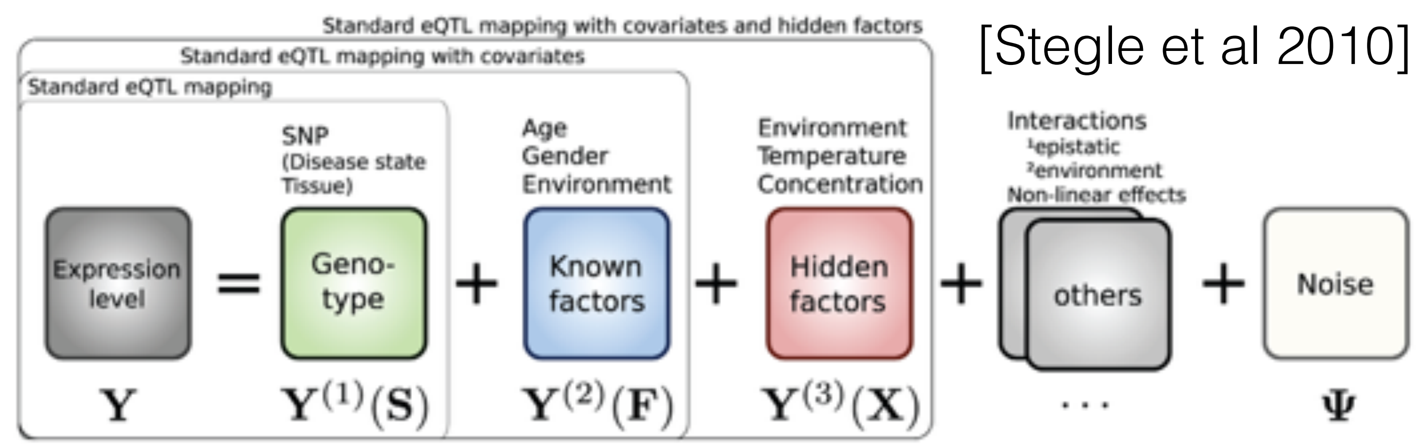
"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY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

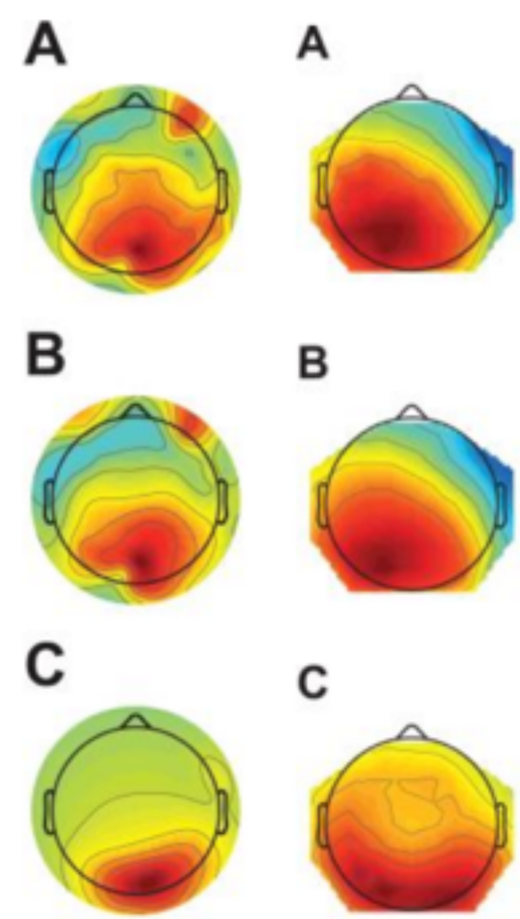
[Blei et al 2003]



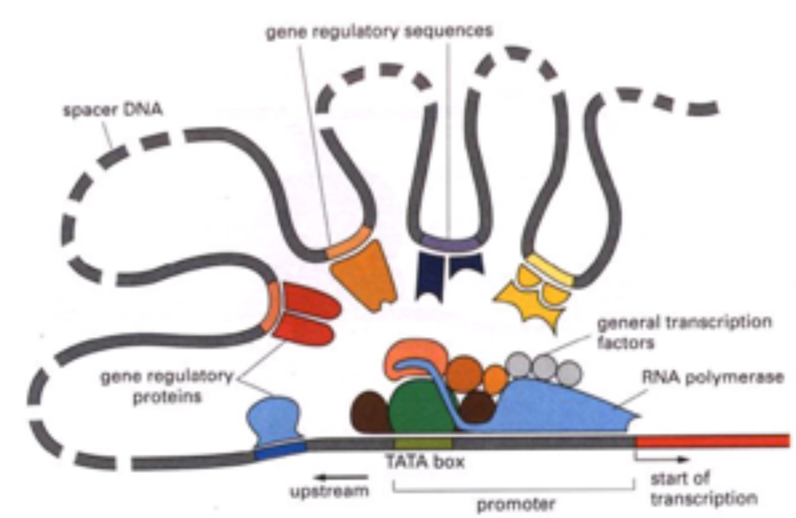
[Airoldi et al 2008]



[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

θ

Bayesian inference

parameters

$p(\theta)$

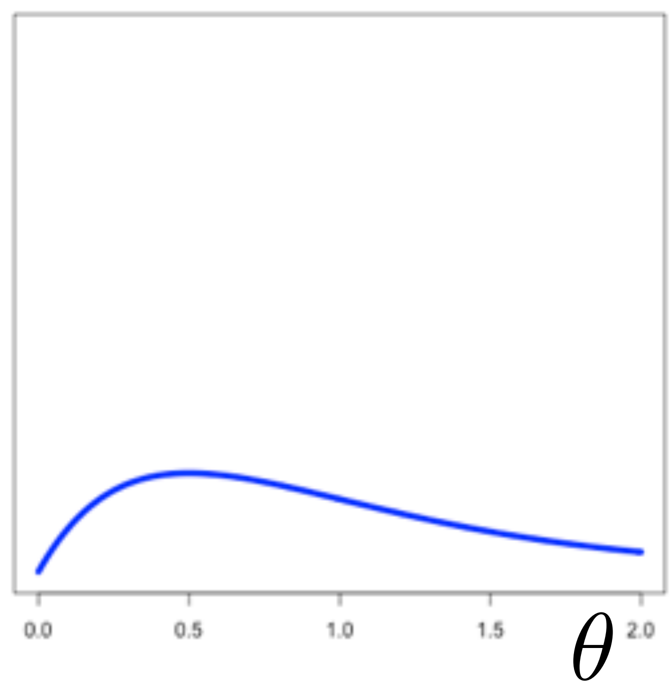
prior

Bayesian inference

parameters

$p(\theta)$

prior

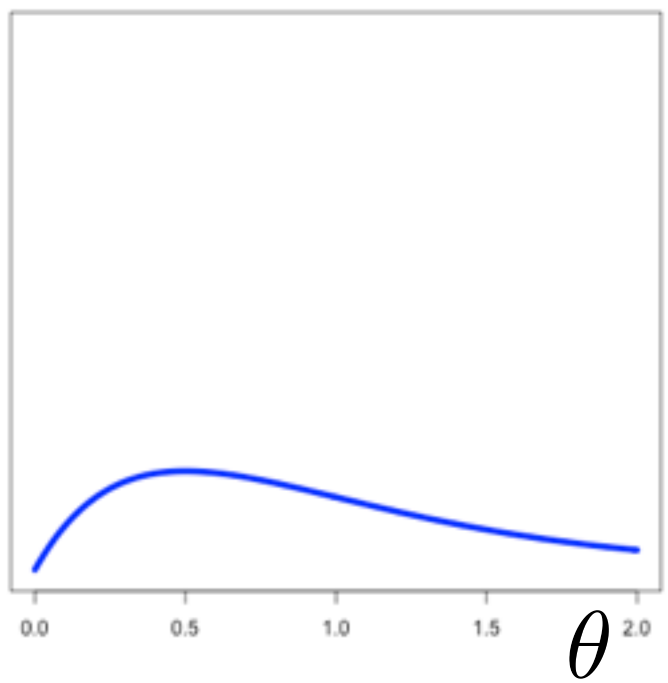


Bayesian inference

parameters

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

likelihood prior

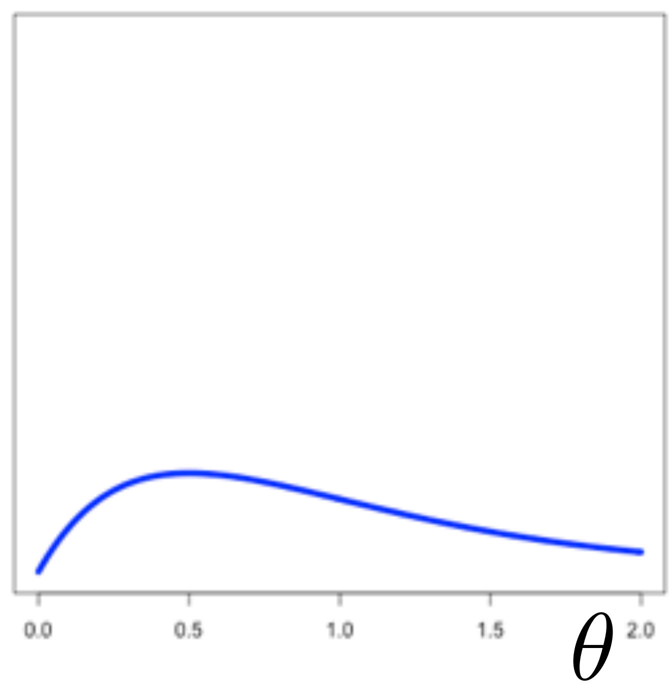


Bayesian inference

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

data → $y_{1:N}$; parameters → θ

likelihood prior

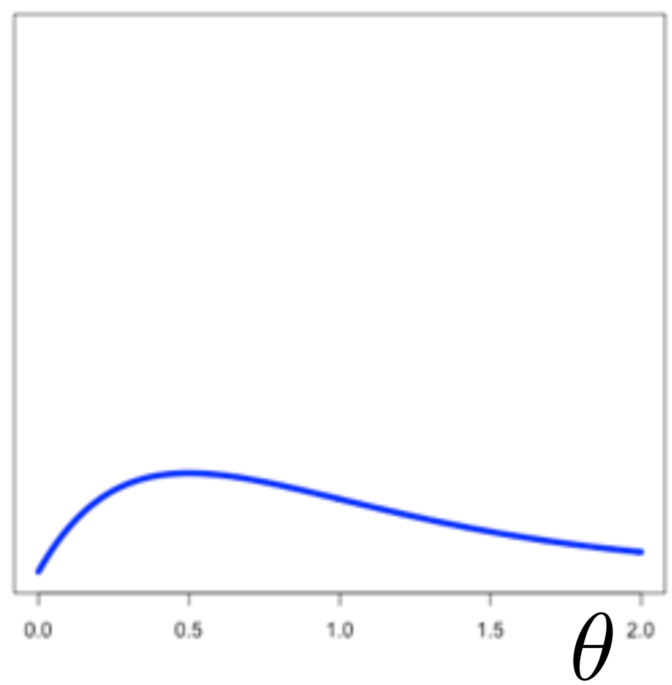


Bayesian inference

data → parameters

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

posterior likelihood prior

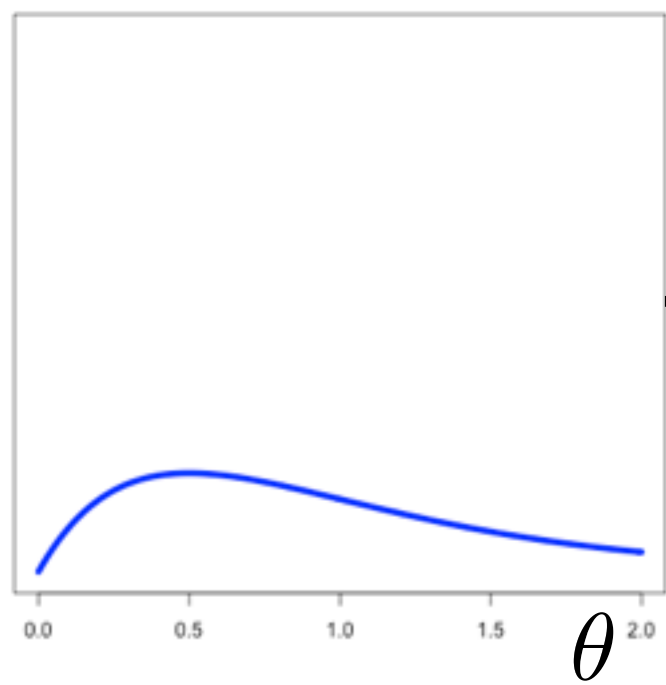


Bayesian inference

data parameters

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

posterior likelihood prior



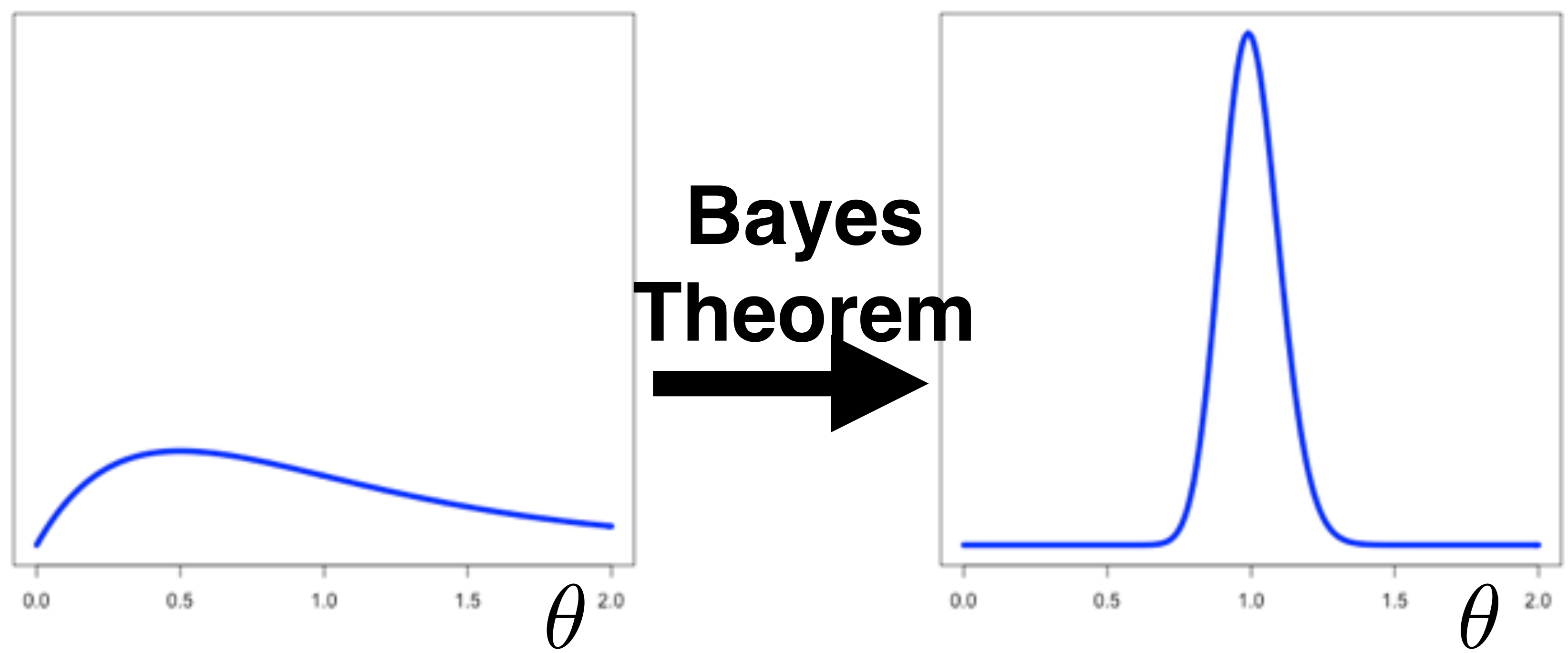
Bayes Theorem

Bayesian inference

data parameters

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

posterior likelihood prior

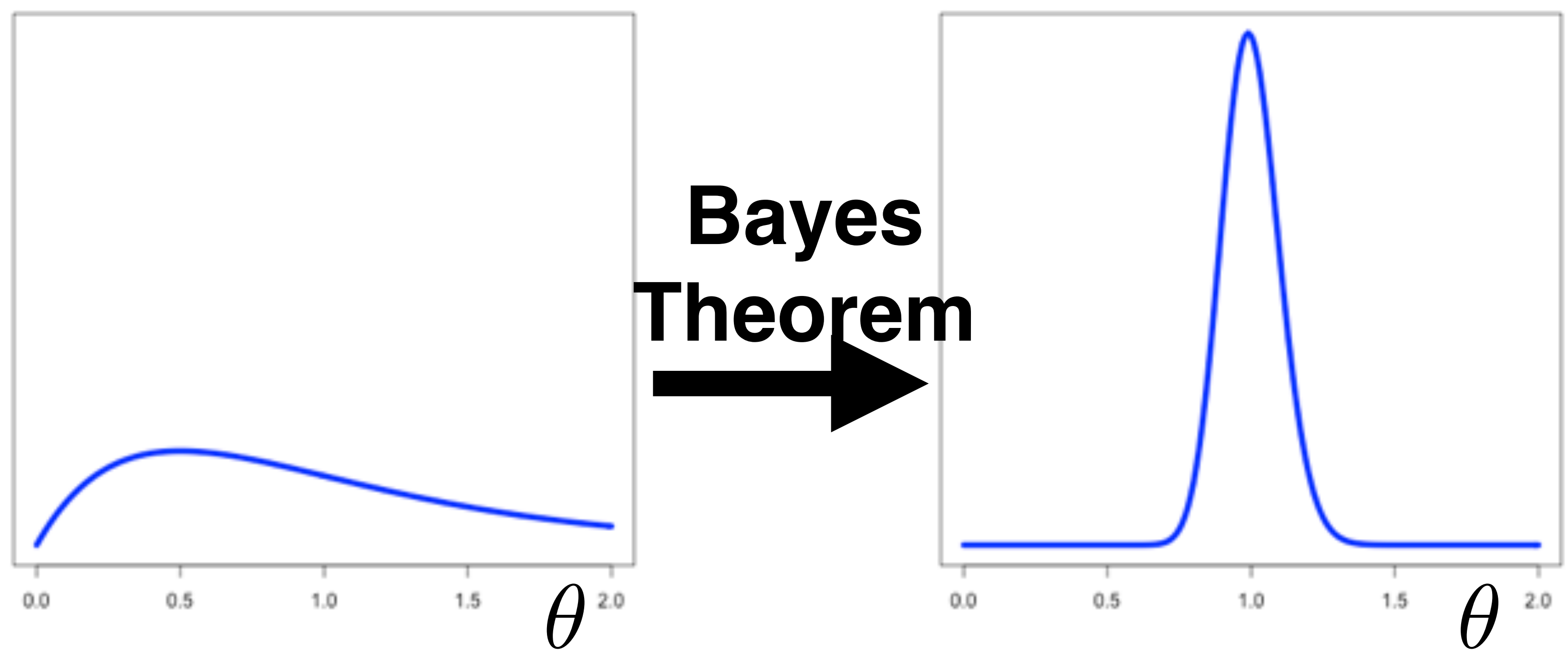


Bayesian inference

data parameters

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

posterior likelihood prior



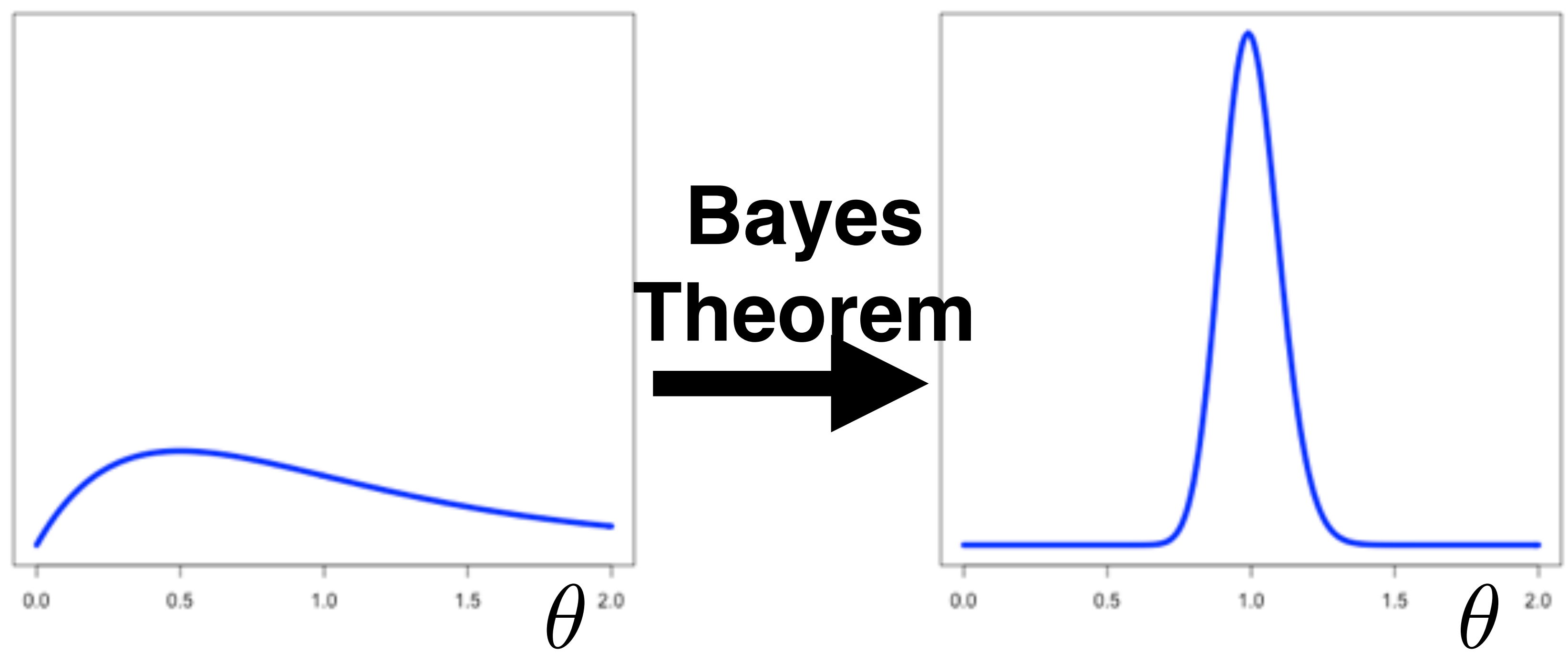
1. Build a model: choose prior & choose likelihood

Bayesian inference

data → $p(y_{1:N}|\theta)$; parameters → $p(\theta)$

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

posterior likelihood prior



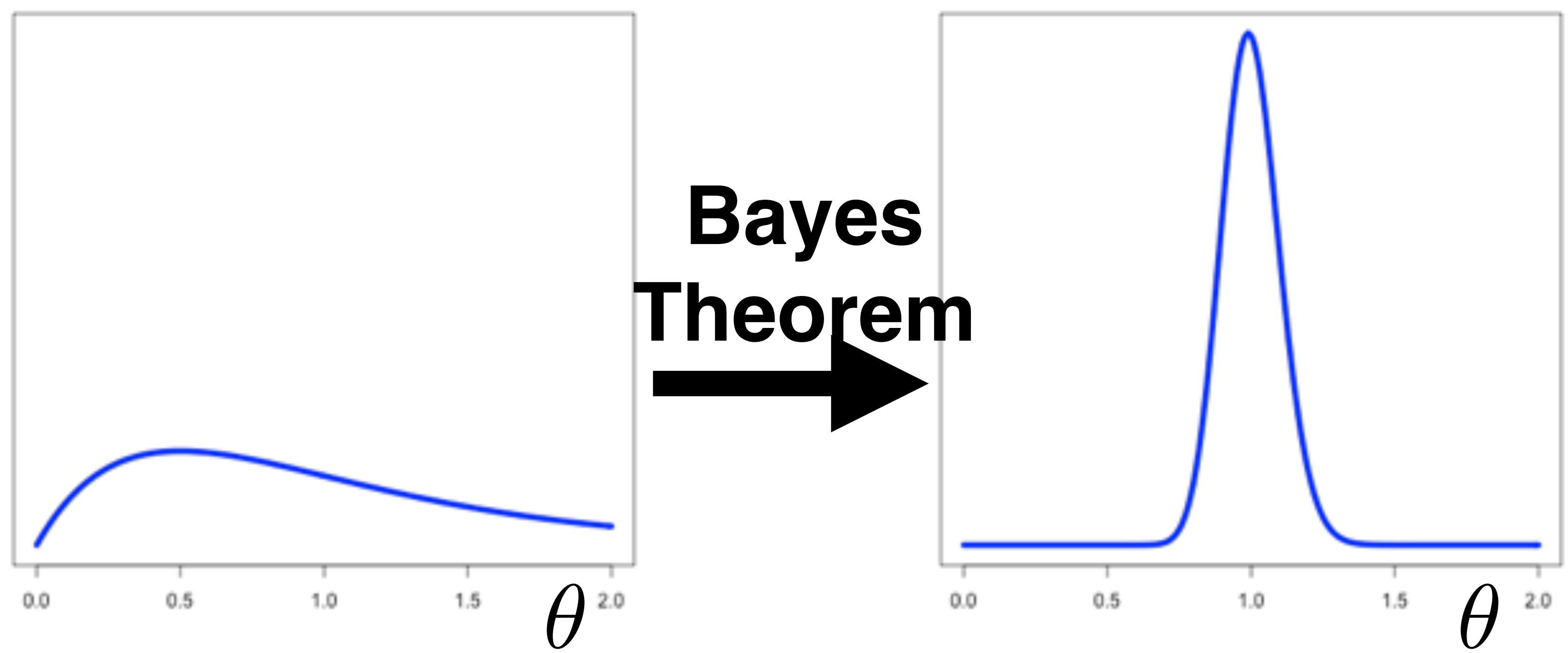
1. Build a model: choose prior & choose likelihood
2. Compute the posterior

Bayesian inference

data parameters

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

posterior likelihood prior



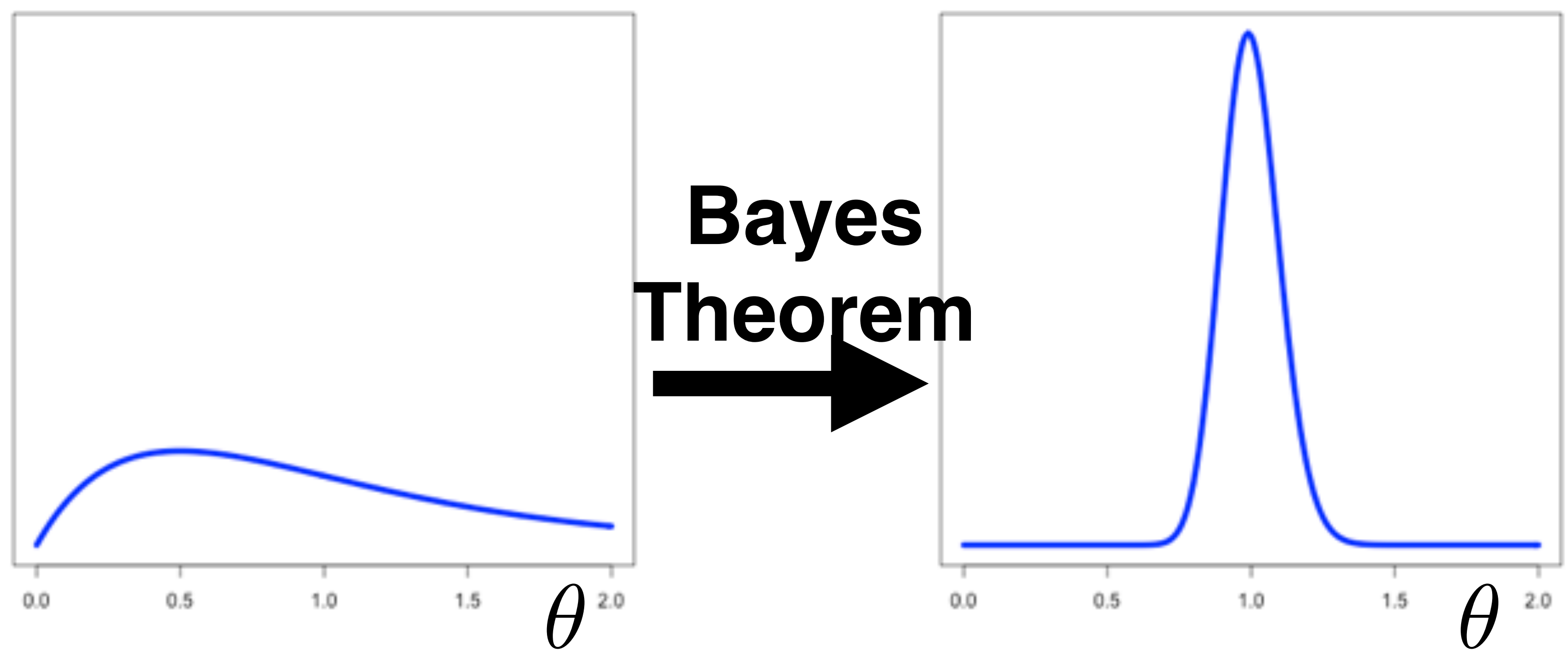
1. Build a model: choose prior & choose likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances

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

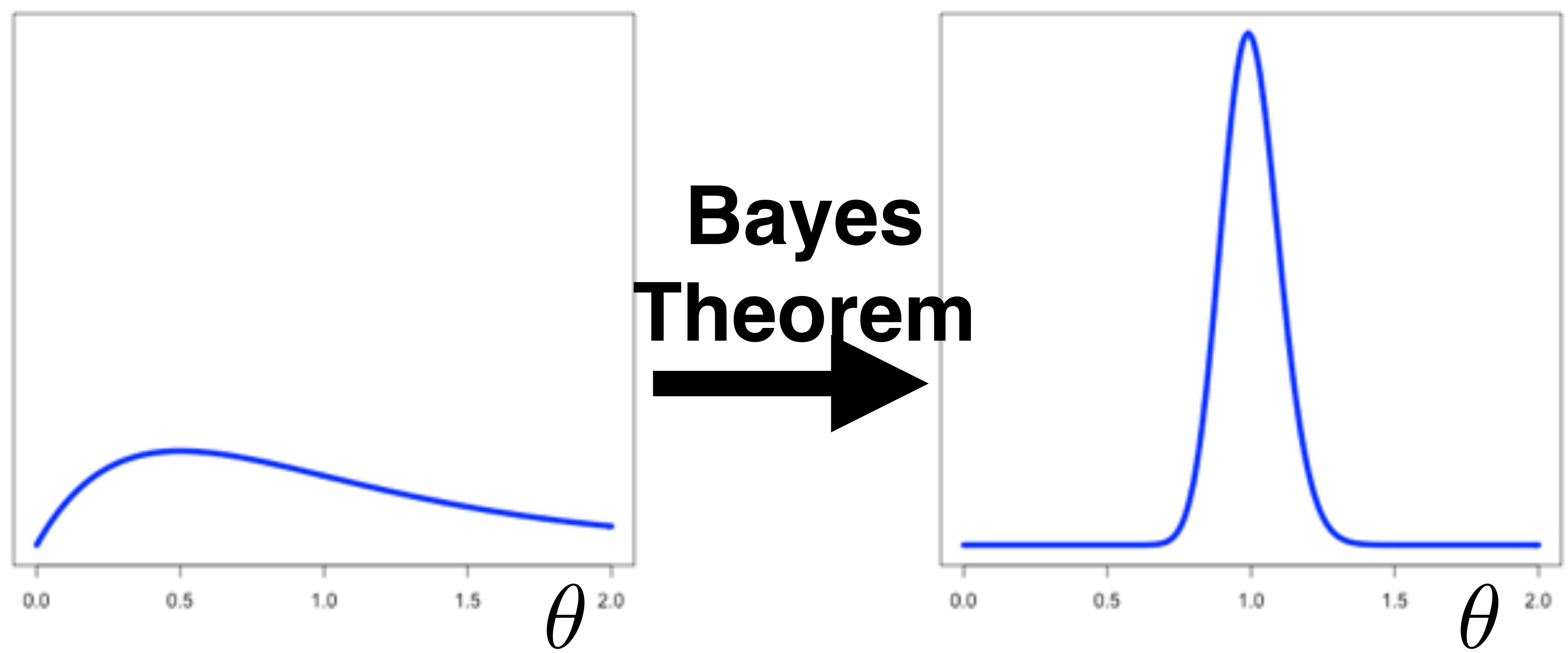
- Why are steps 2 and 3 hard?

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

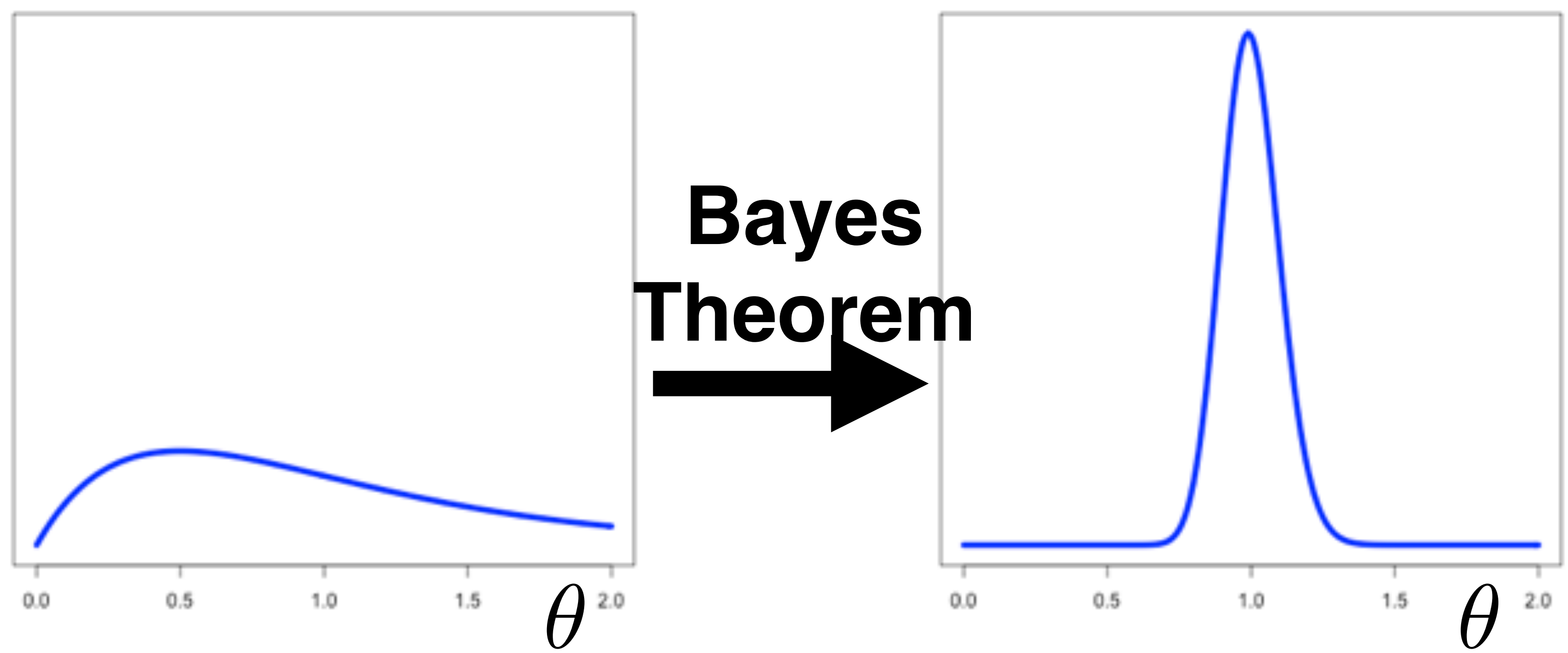
- Why are steps 2 and 3 hard?
  - Typically no closed form

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

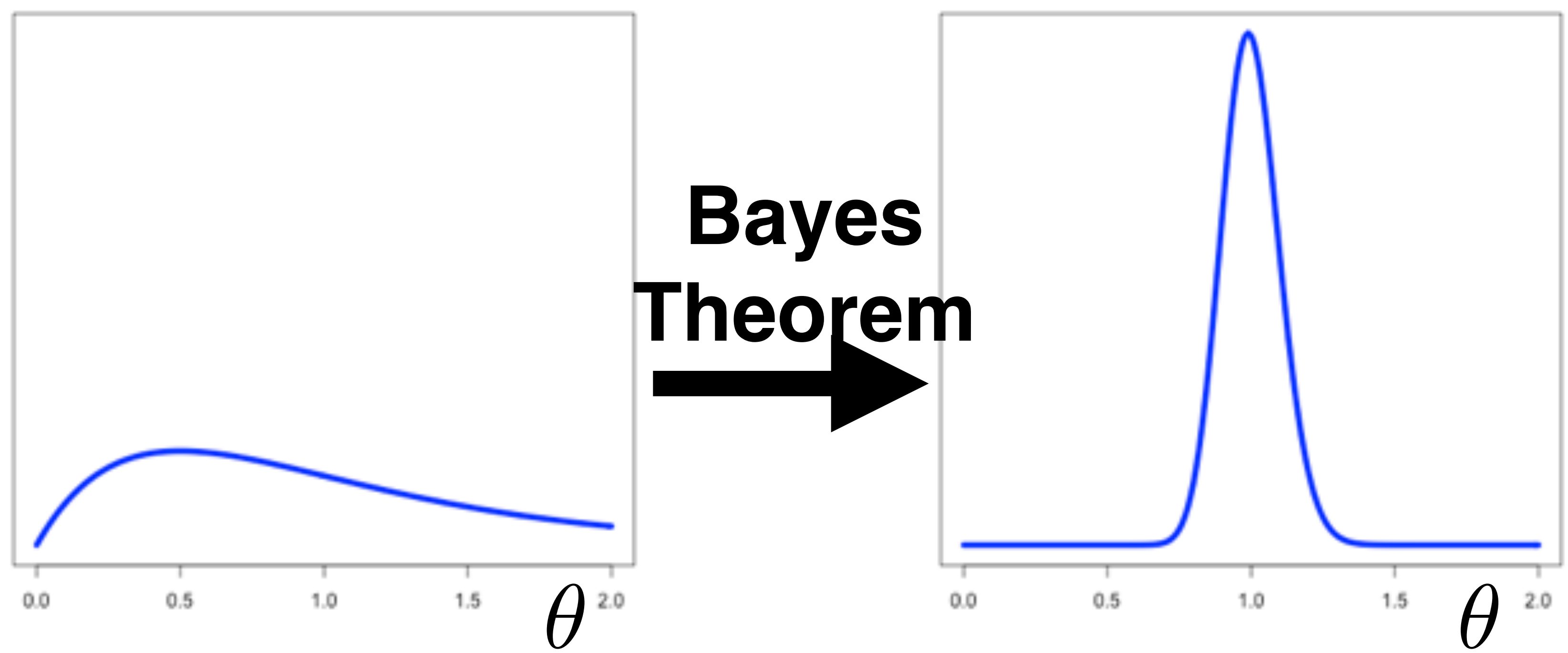
- Why are steps 2 and 3 hard?
    - Typically no closed form, high-dimensional integration

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

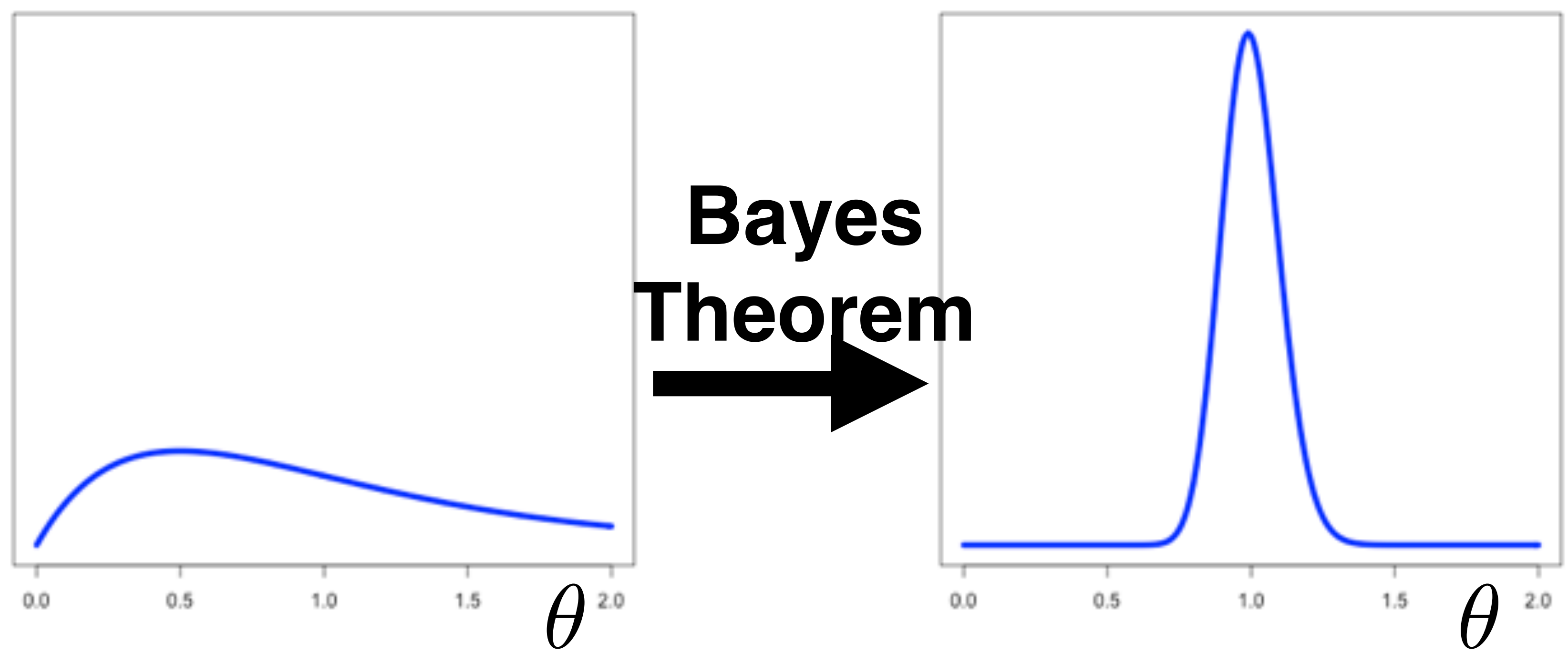
- Why are steps 2 and 3 hard?
  - Typically no closed form, high-dimensional integration

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

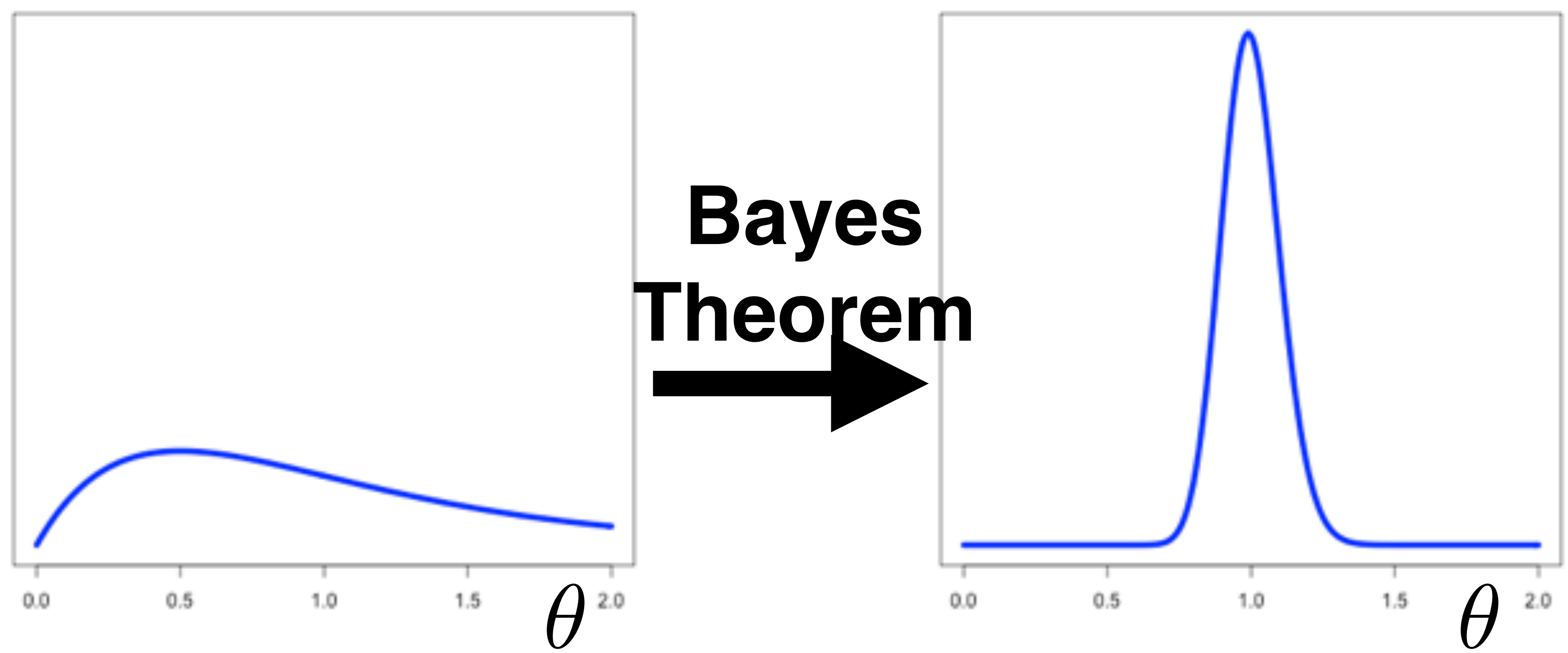
- Why are steps 2 and 3 hard?
    - Typically no closed form, high-dimensional integration

Bayesian inference

data → $y_{1:N}$, parameters → θ

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

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

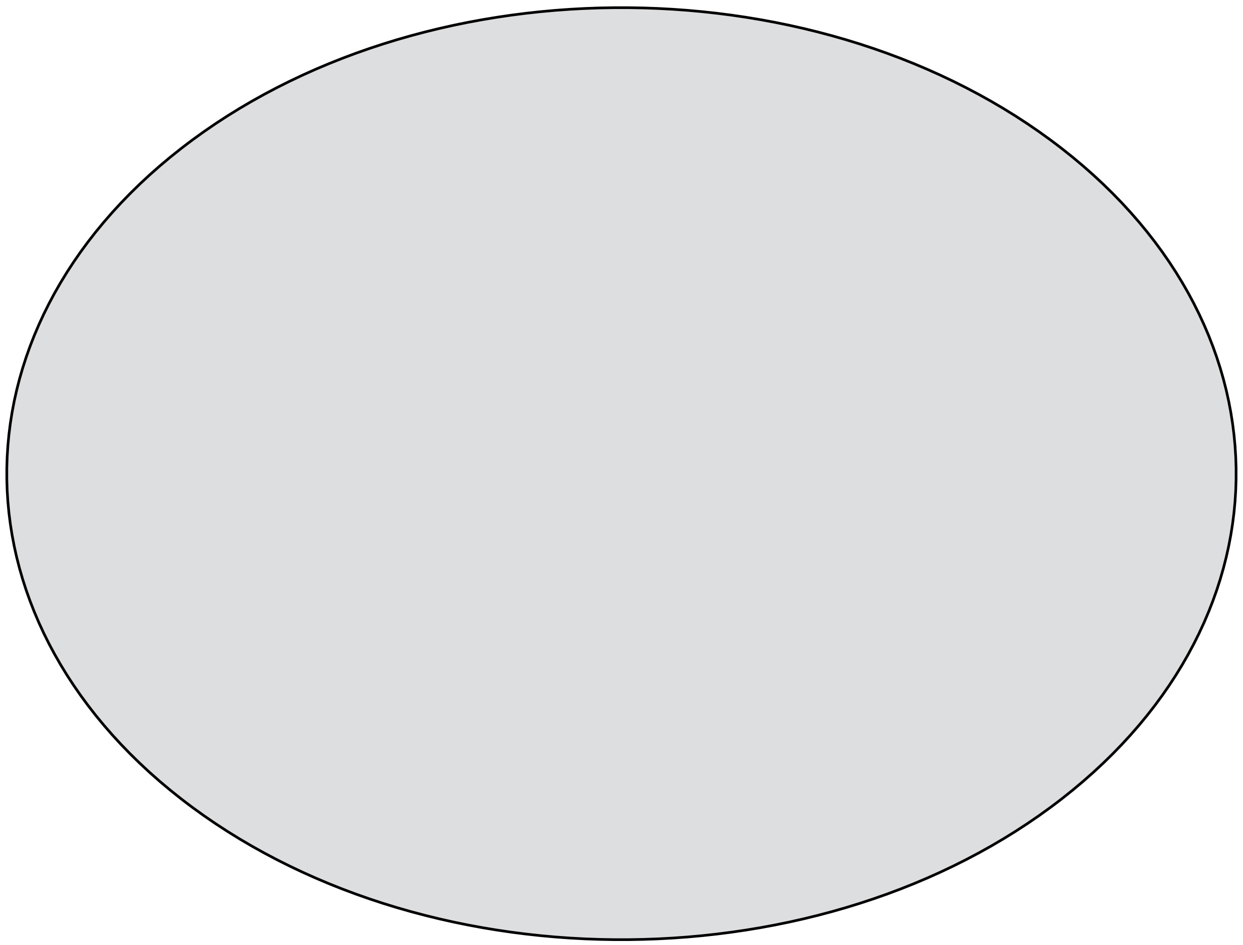
Instead: an optimization approach

- Approximate posterior with q^*

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

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



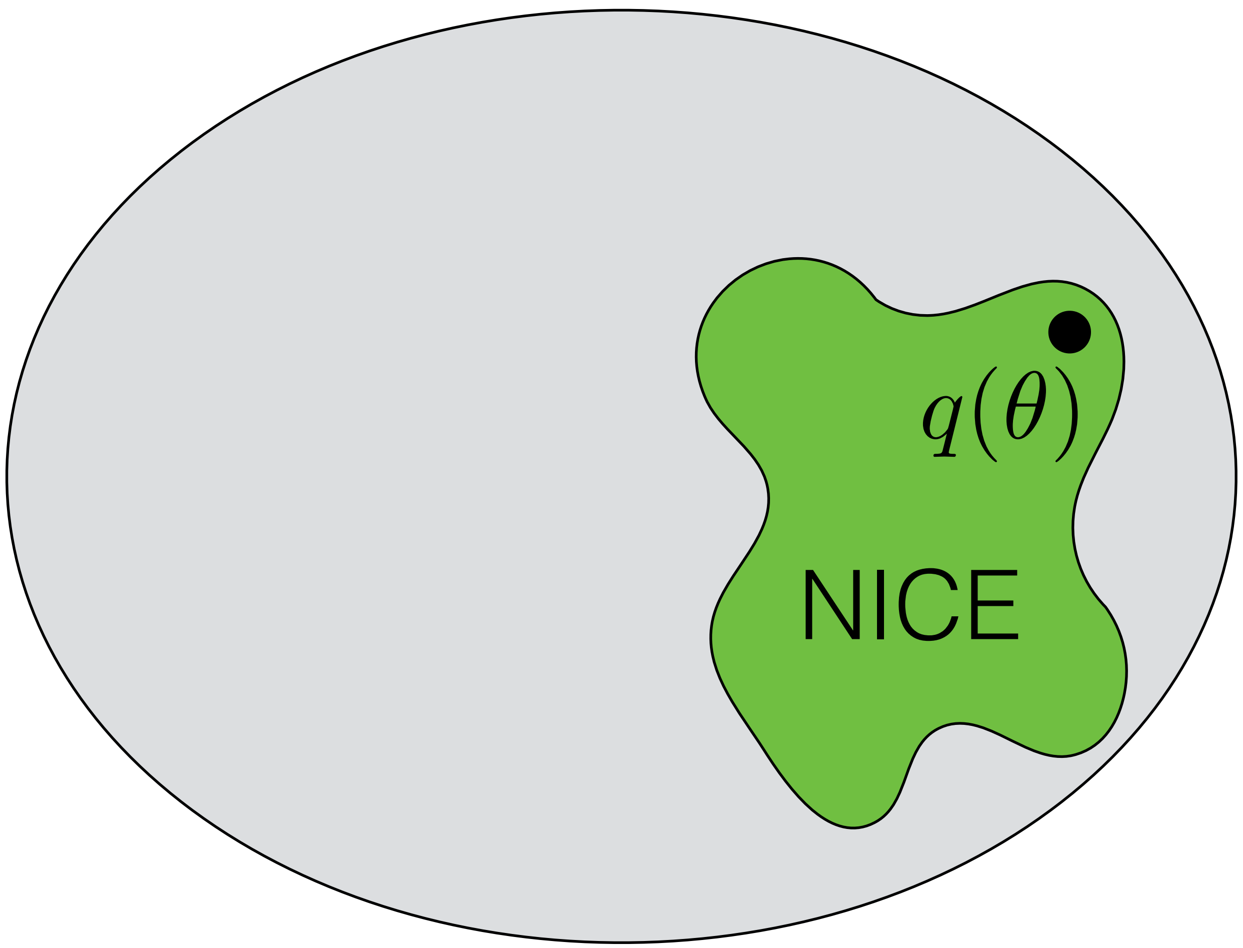
Instead: an optimization approach

- Approximate posterior with q^*

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

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



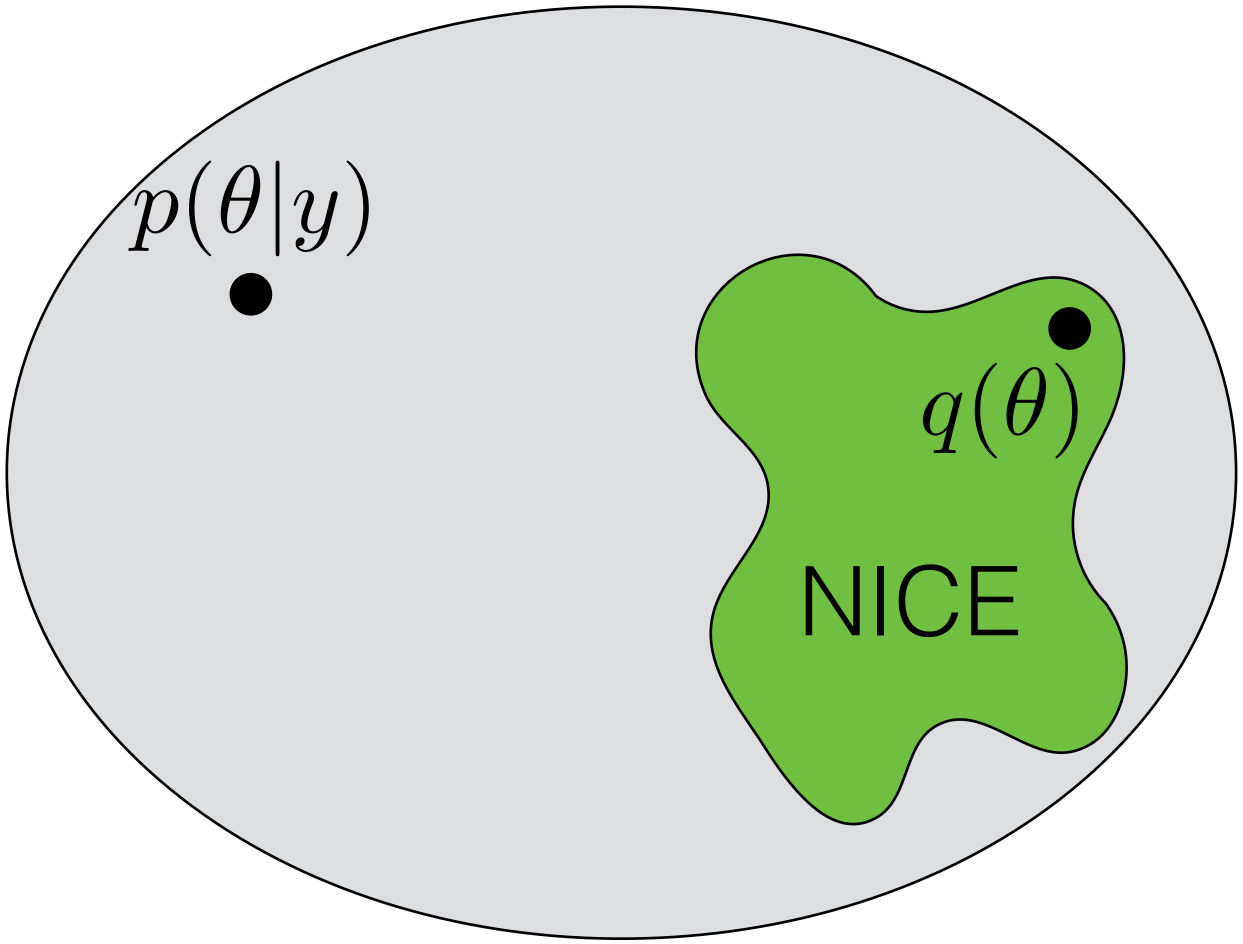
Instead: an optimization approach

- Approximate posterior with q^*

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

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



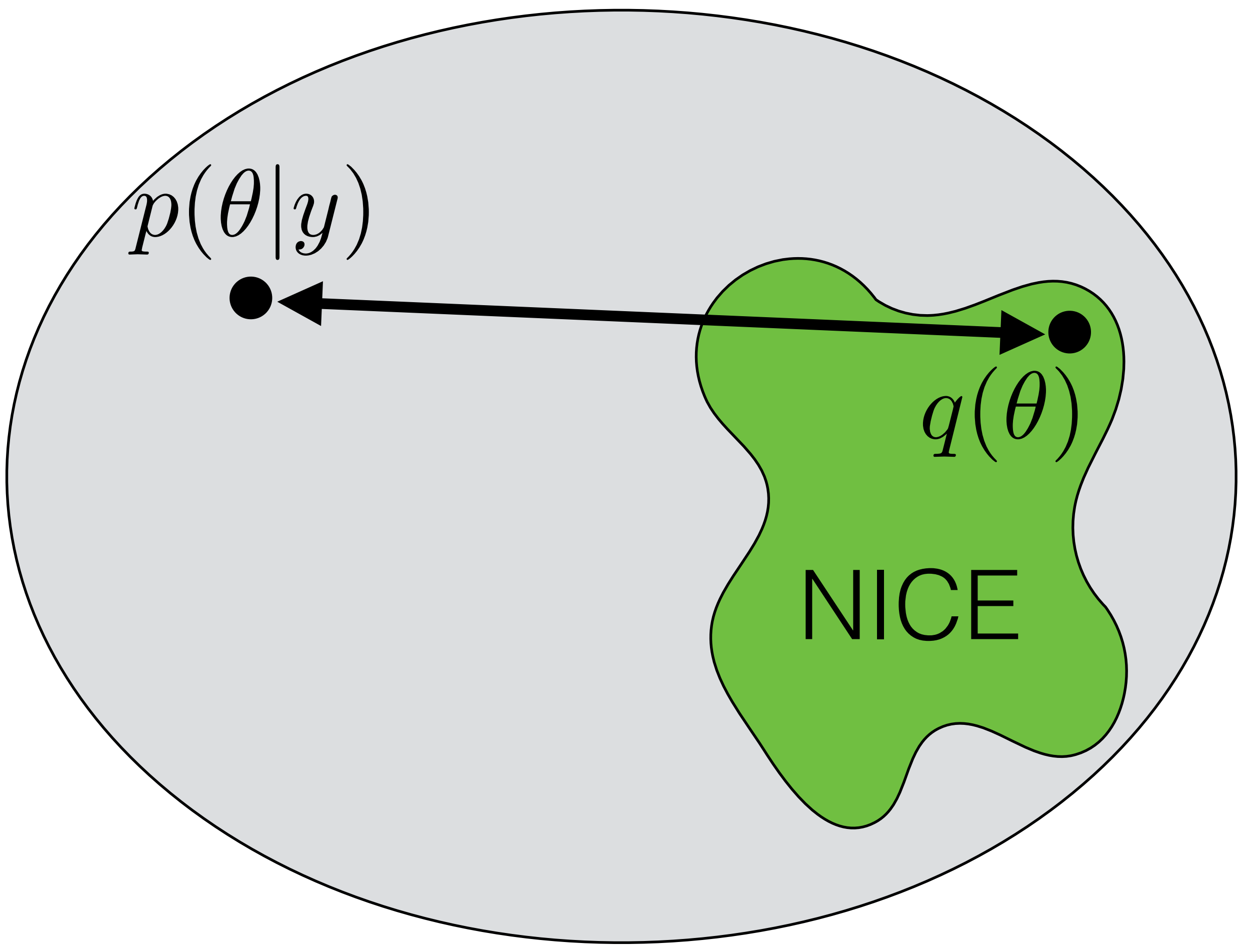
Instead: an optimization approach

- Approximate posterior with q^*

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

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



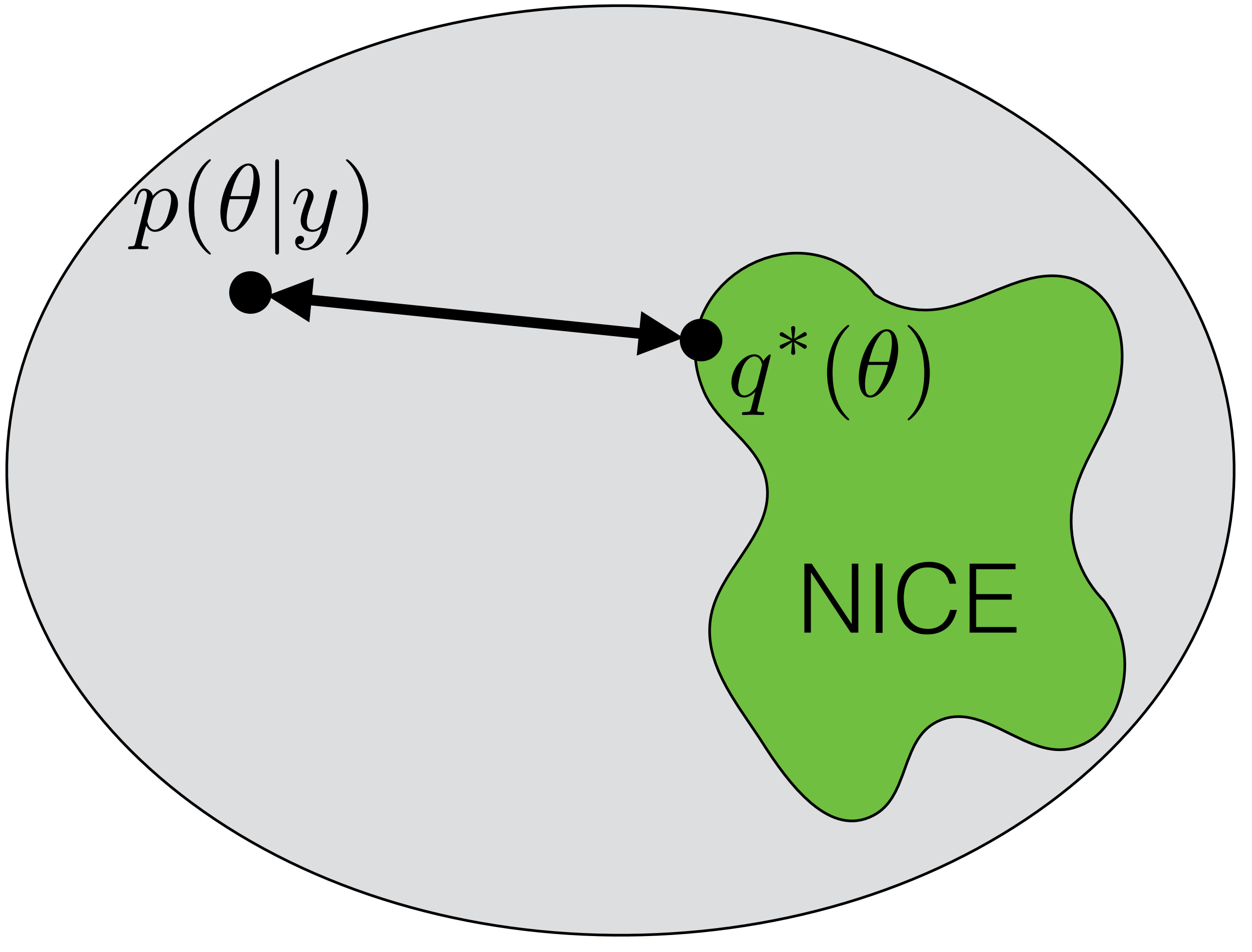
Instead: an optimization approach

- Approximate posterior with q^*

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

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



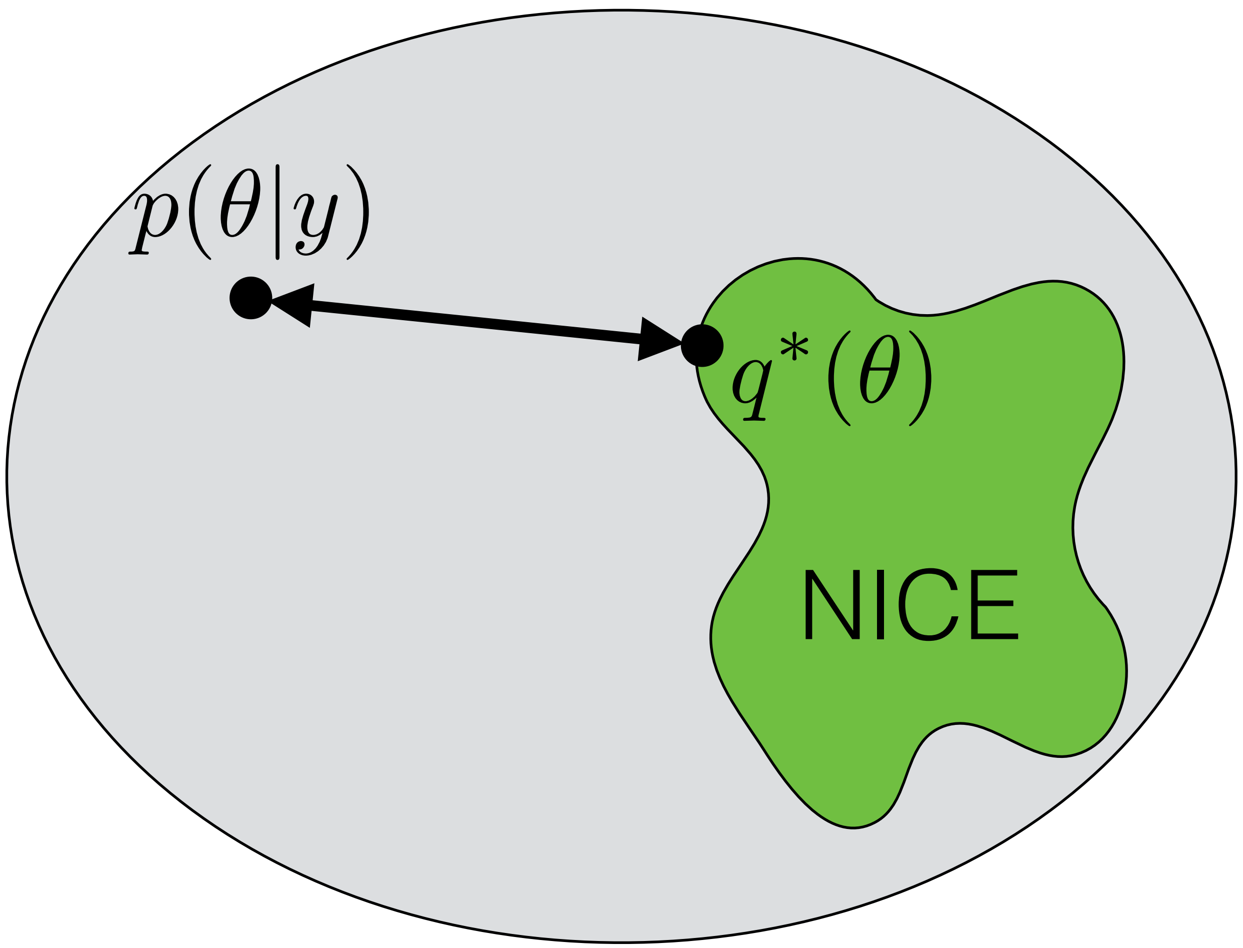
Instead: an optimization approach

- 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

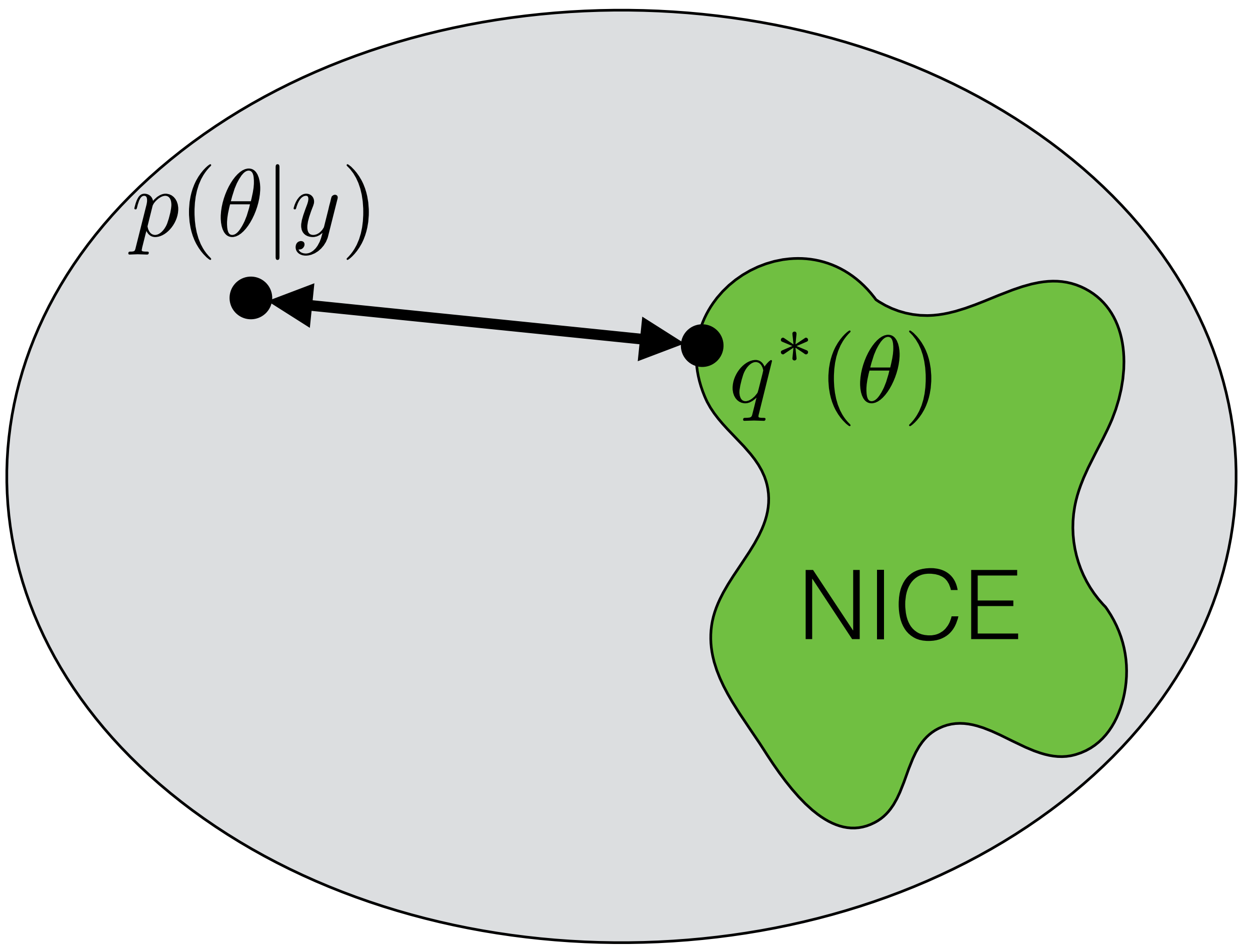
- Approximate posterior with q^*

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

Approximate Bayesian Inference

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

[Bardenet, Doucet, Holmes 2017]



Instead: an optimization approach

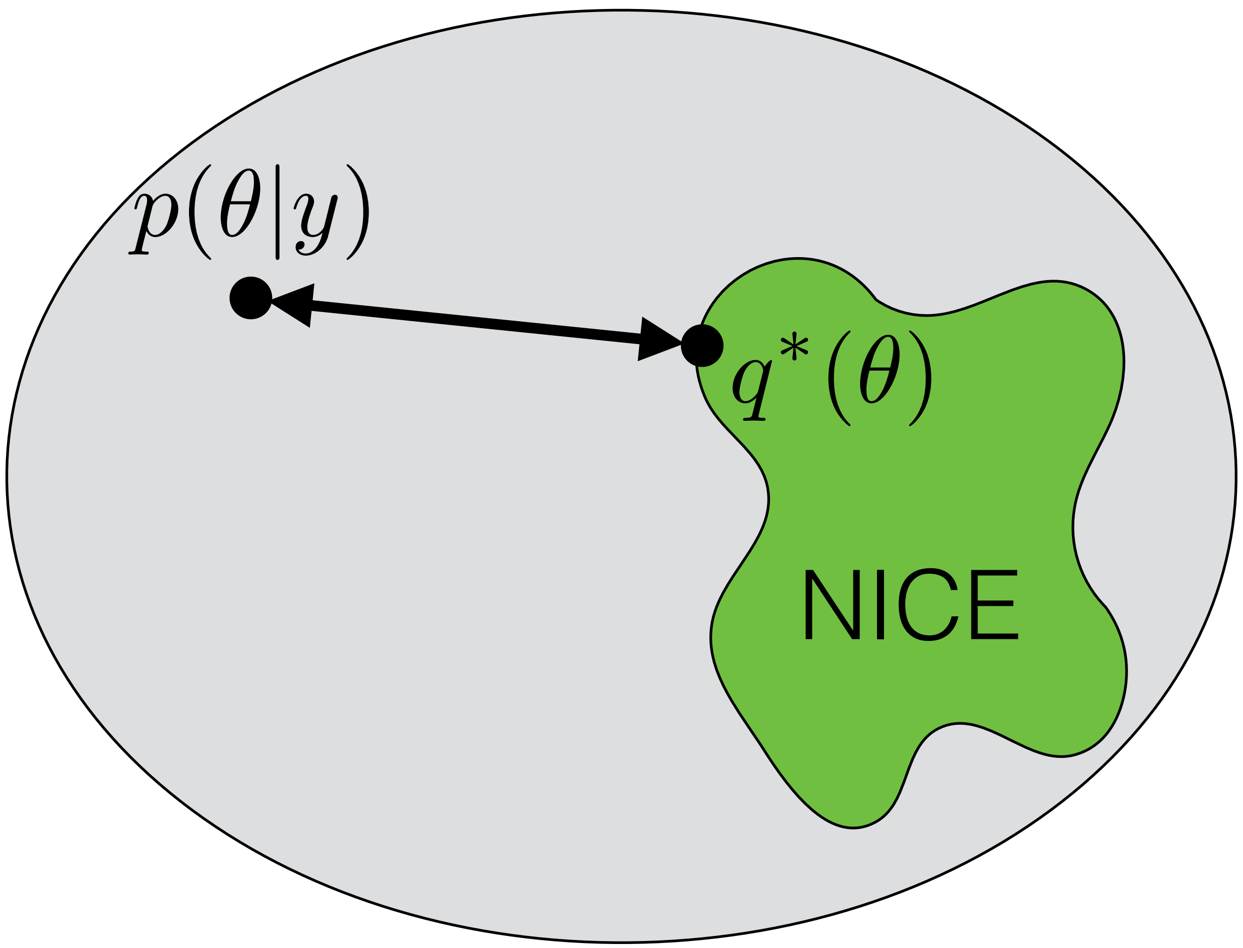
- Approximate posterior with q^*

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

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

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



Instead: an optimization approach

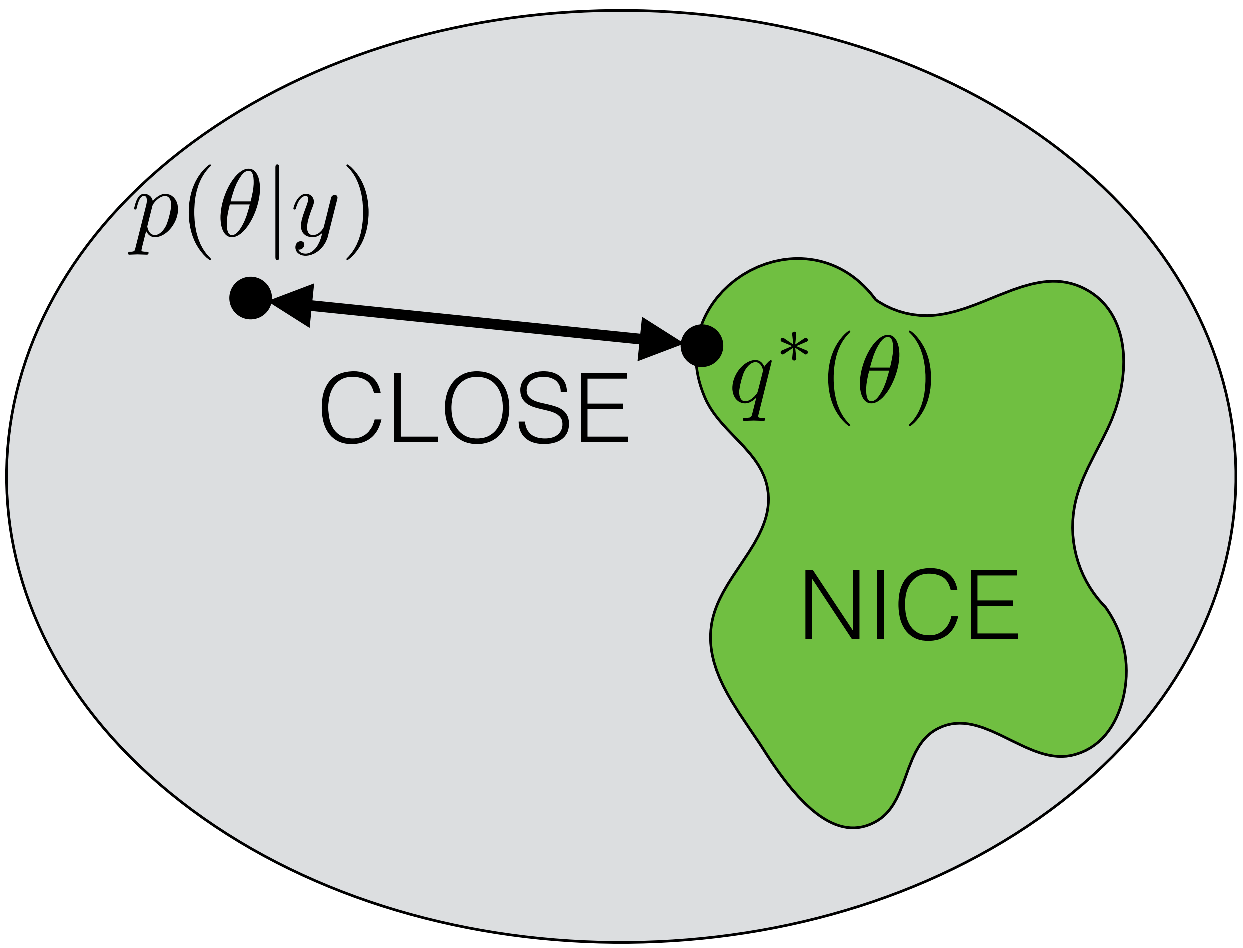
- Approximate posterior with q^*

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

Approximate Bayesian Inference

[Bardenet, Doucet, Holmes 2017]

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



Instead: an optimization approach

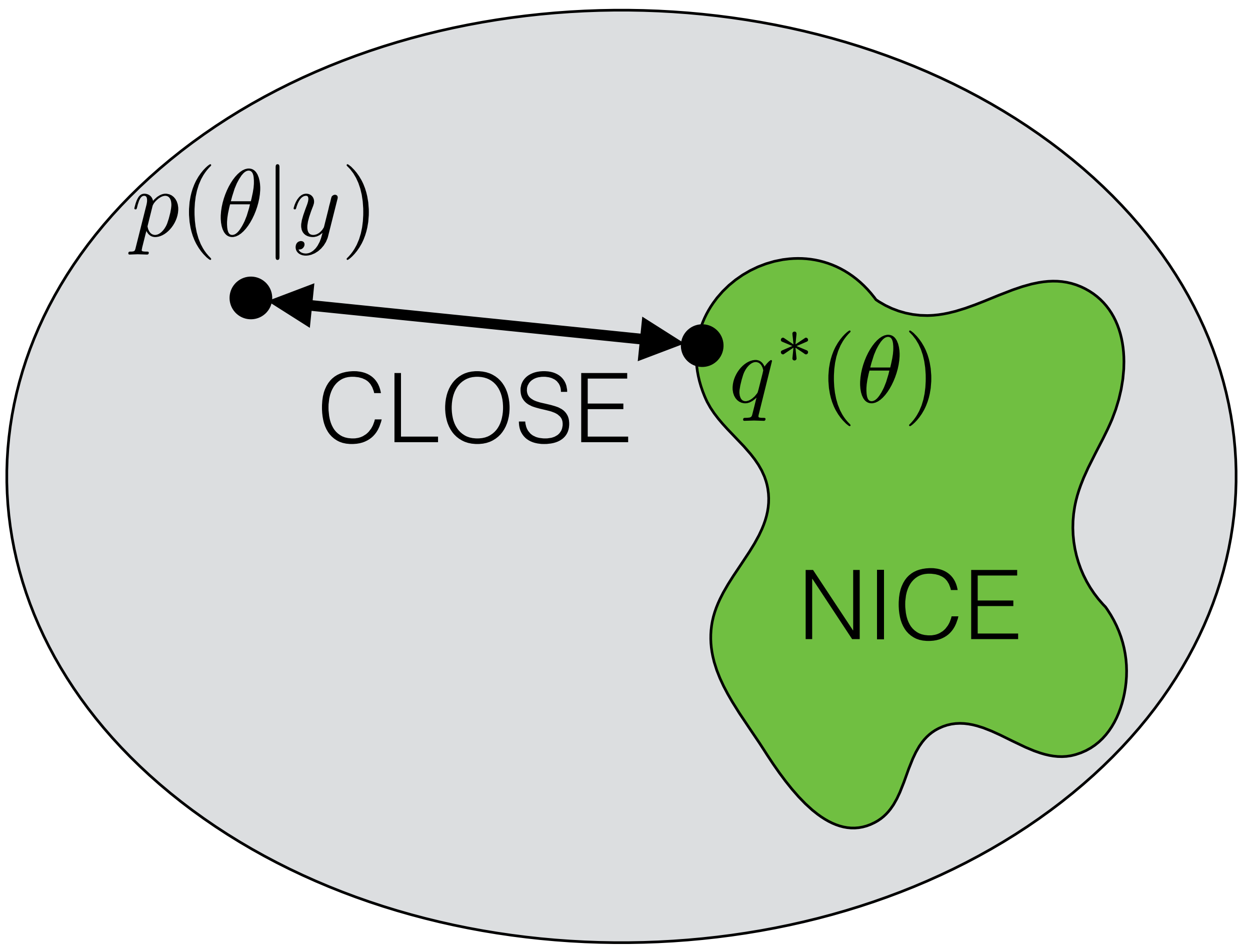
- Approximate posterior with q^*

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

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^*

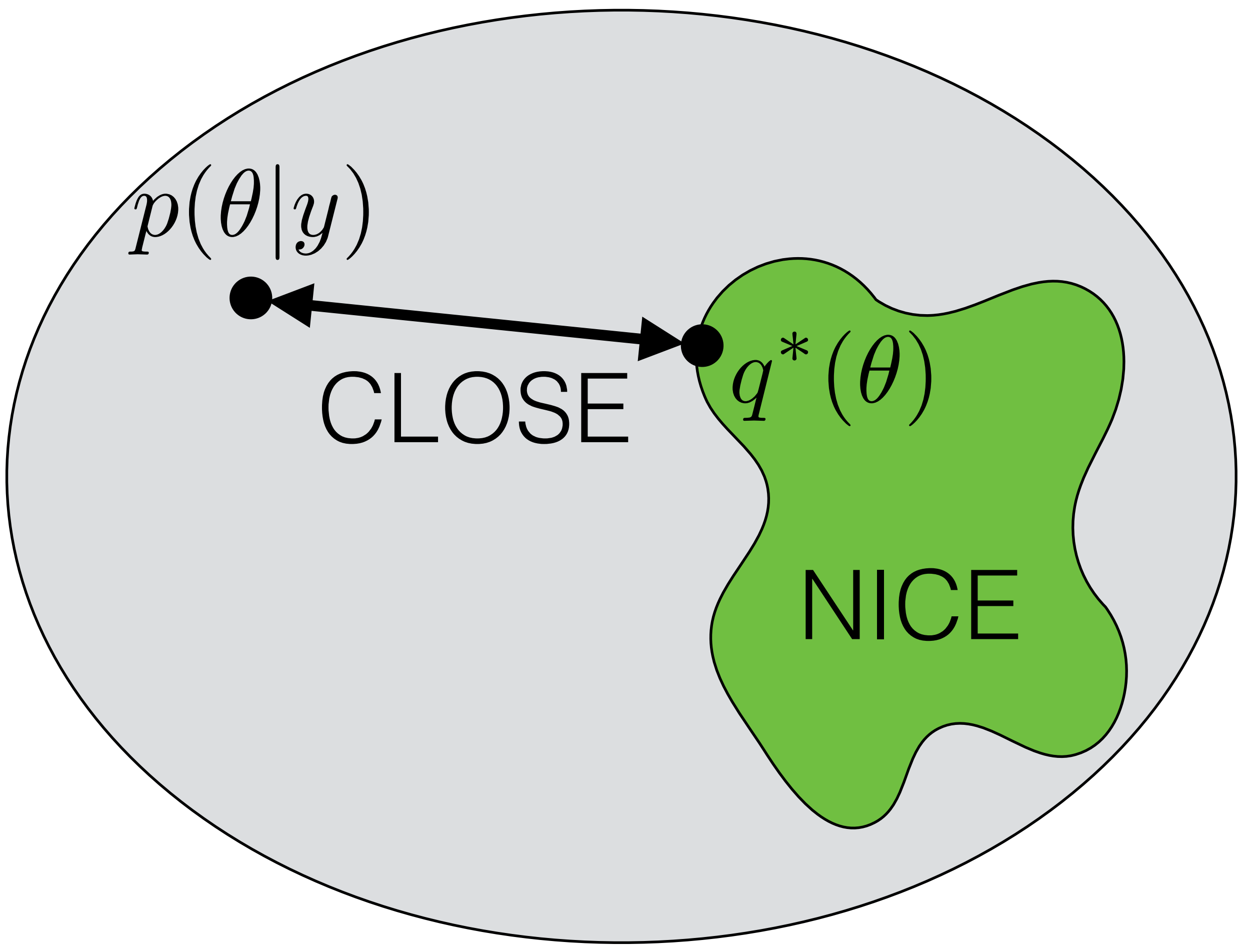
$$q^* = \mathrm{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

[board]

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^*

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

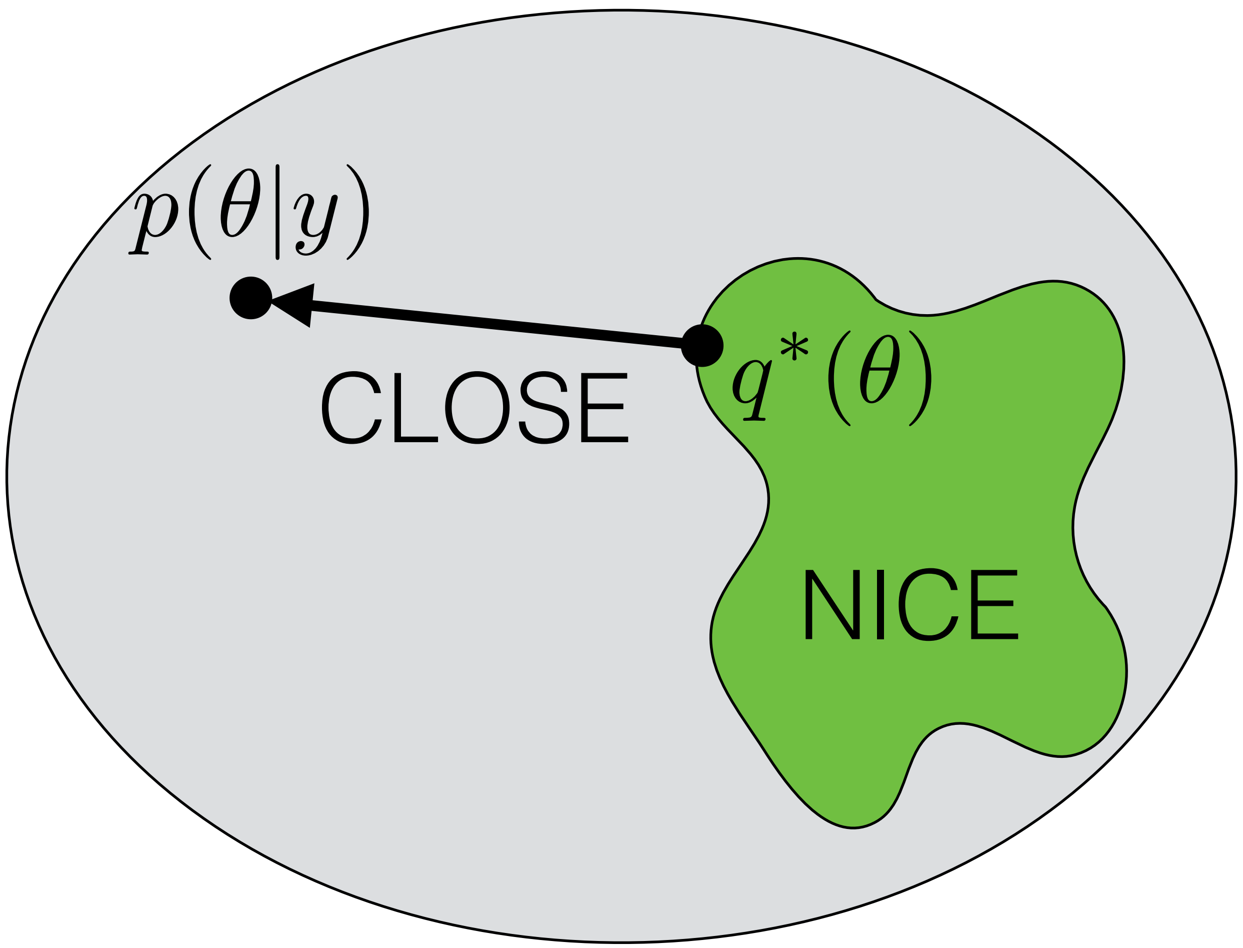
- Variational Bayes (VB): *f* is Kullback-Leibler divergence

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

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^*

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

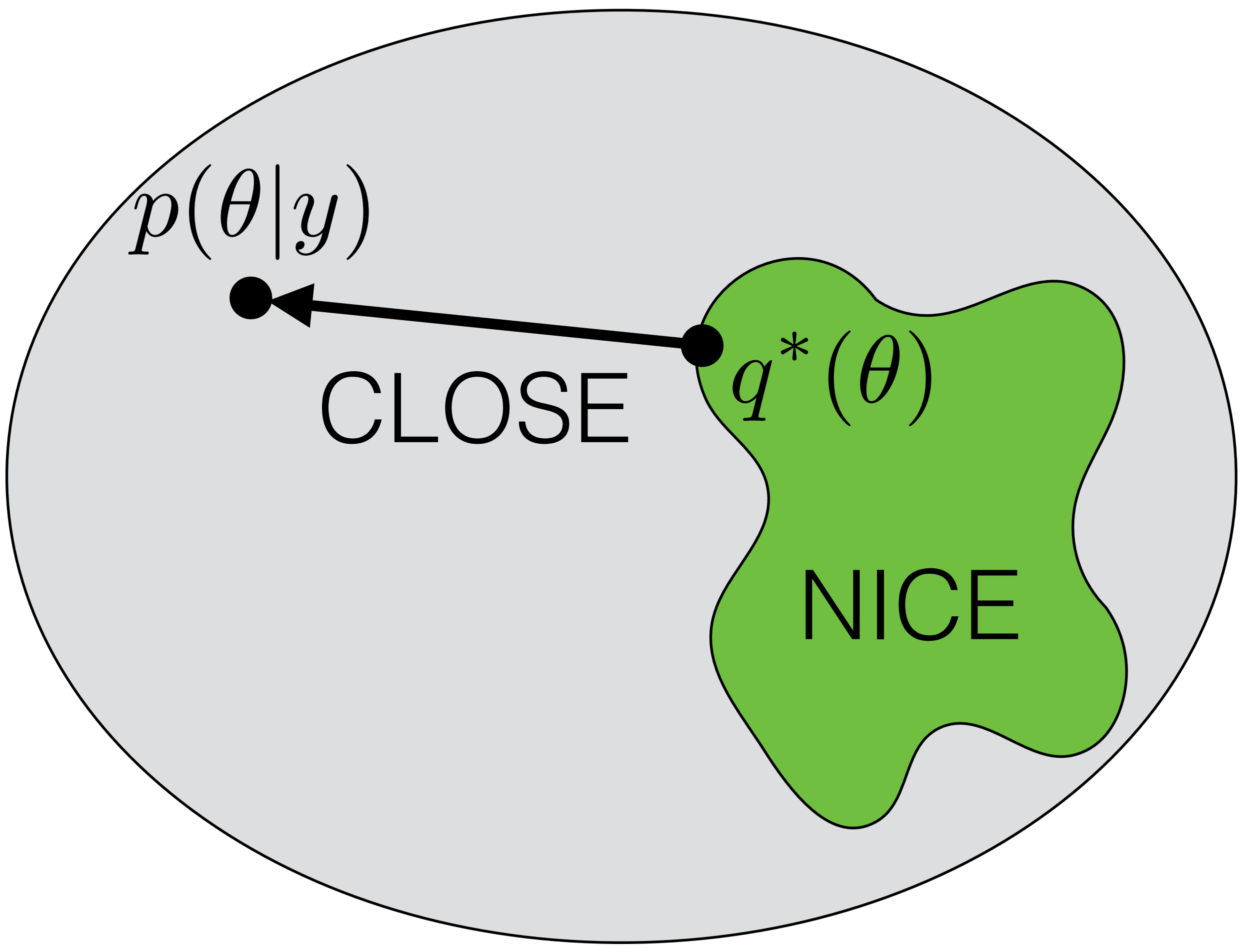
- Variational Bayes (VB): *f* is Kullback-Leibler divergence

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

Approximate Bayesian Inference

- 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^*

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

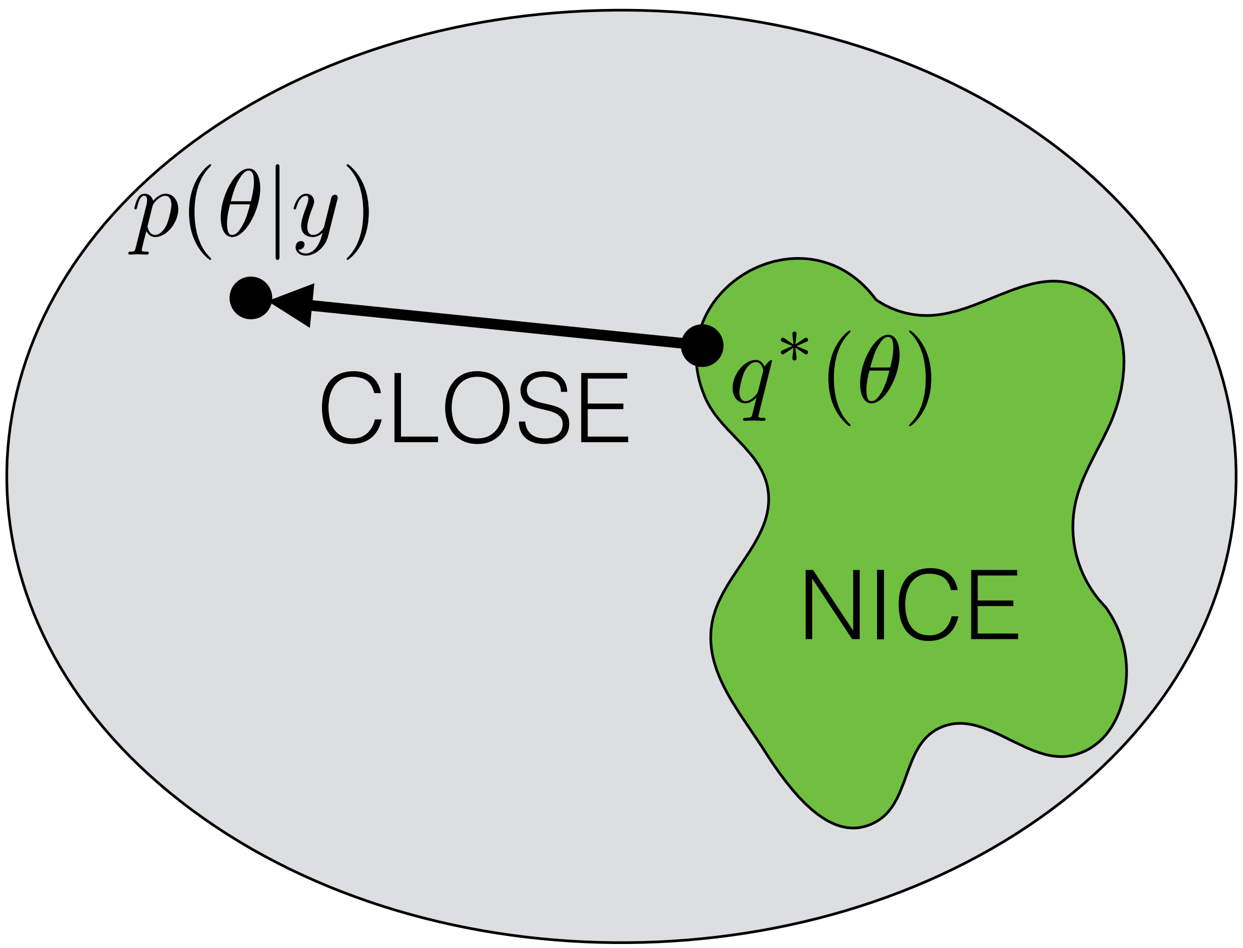
- Variational Bayes (VB): *f* is Kullback-Leibler divergence

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

- VB practical success

Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC) [Bardenet, Doucet, Holmes 2017]
  - 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

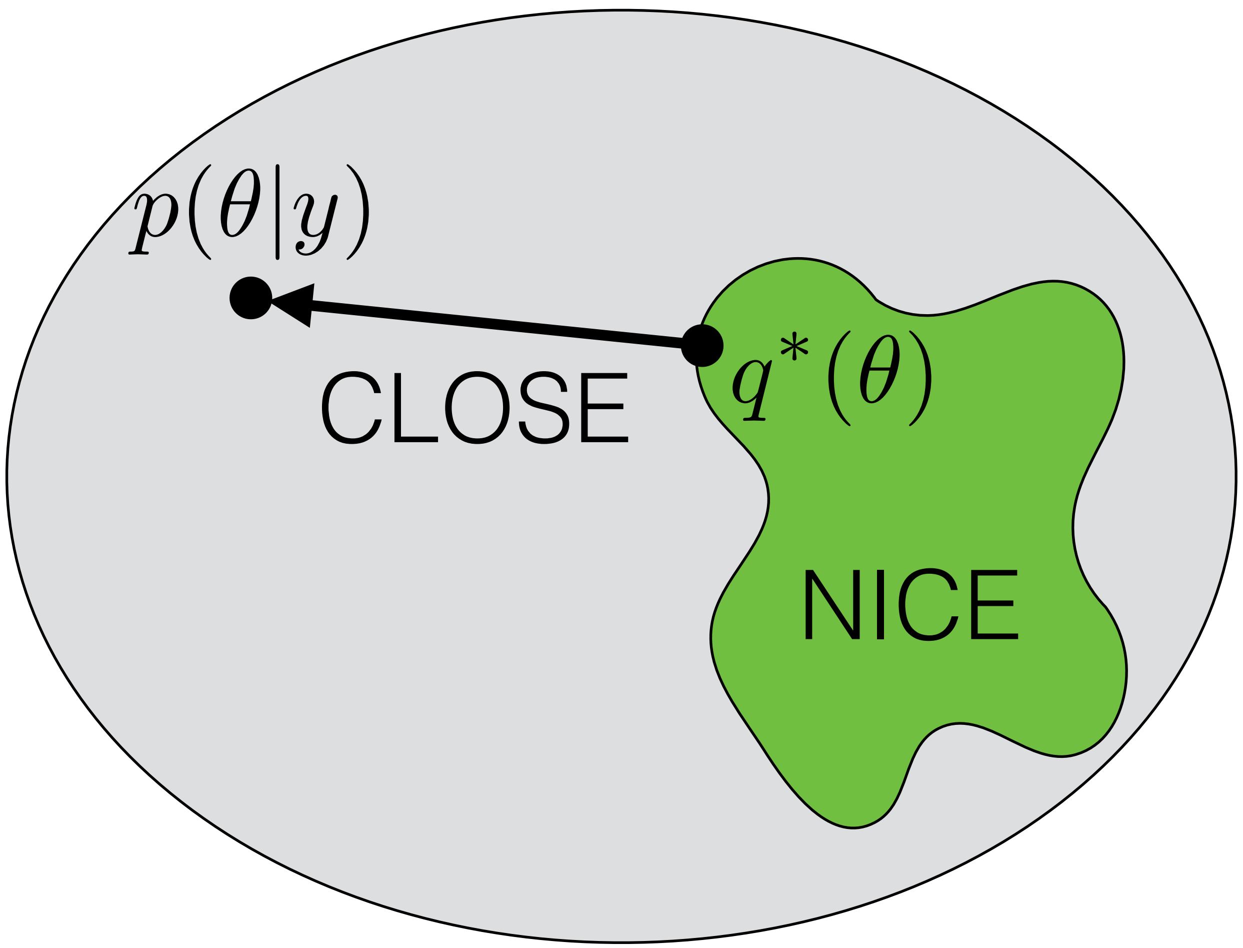
$$KL(q(\cdot)||p(\cdot|y))$$

- VB practical success: point estimates and prediction

Approximate Bayesian Inference

[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

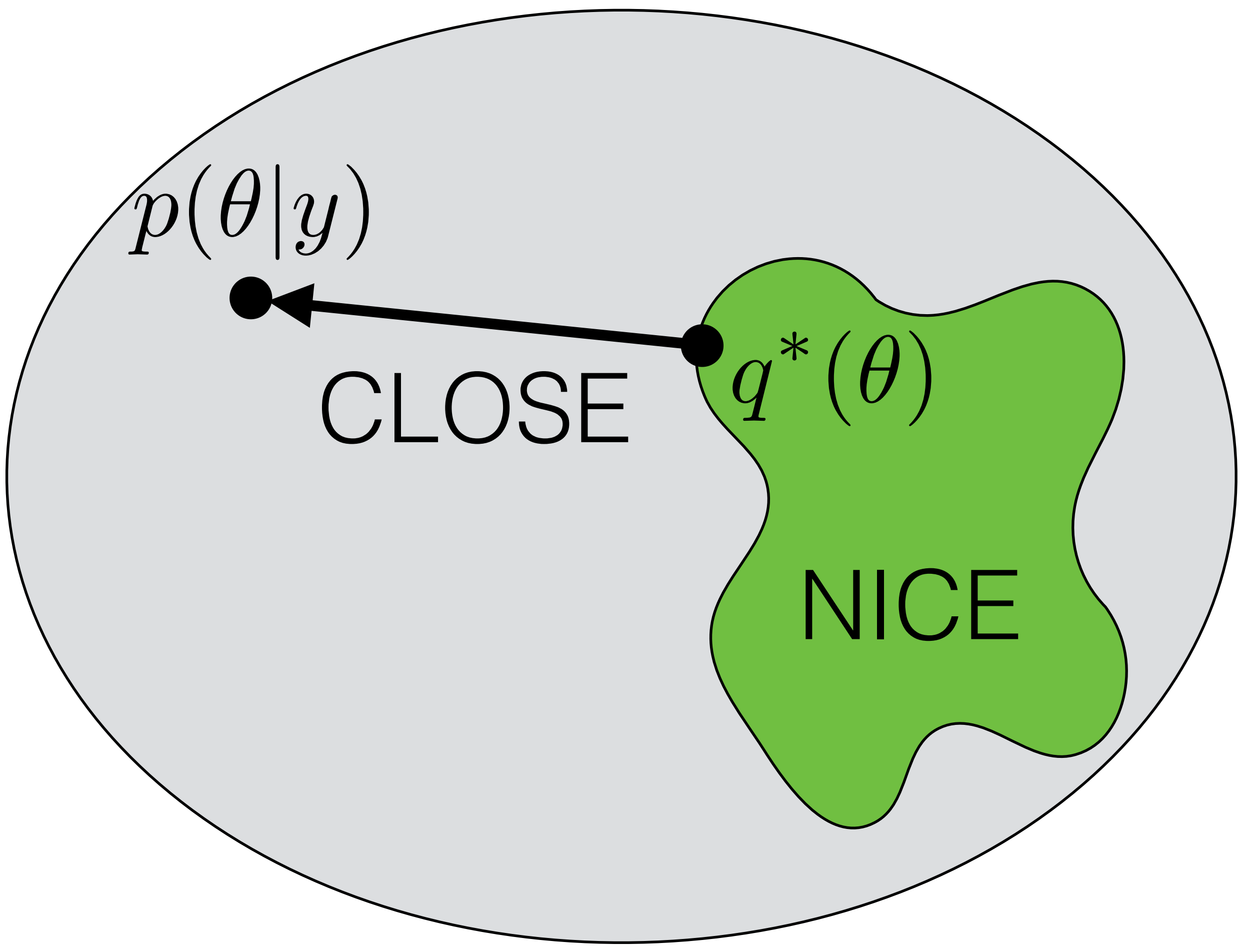
$$KL(q(\cdot)||p(\cdot|y))$$

- VB practical success: point estimates and prediction, fast

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

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

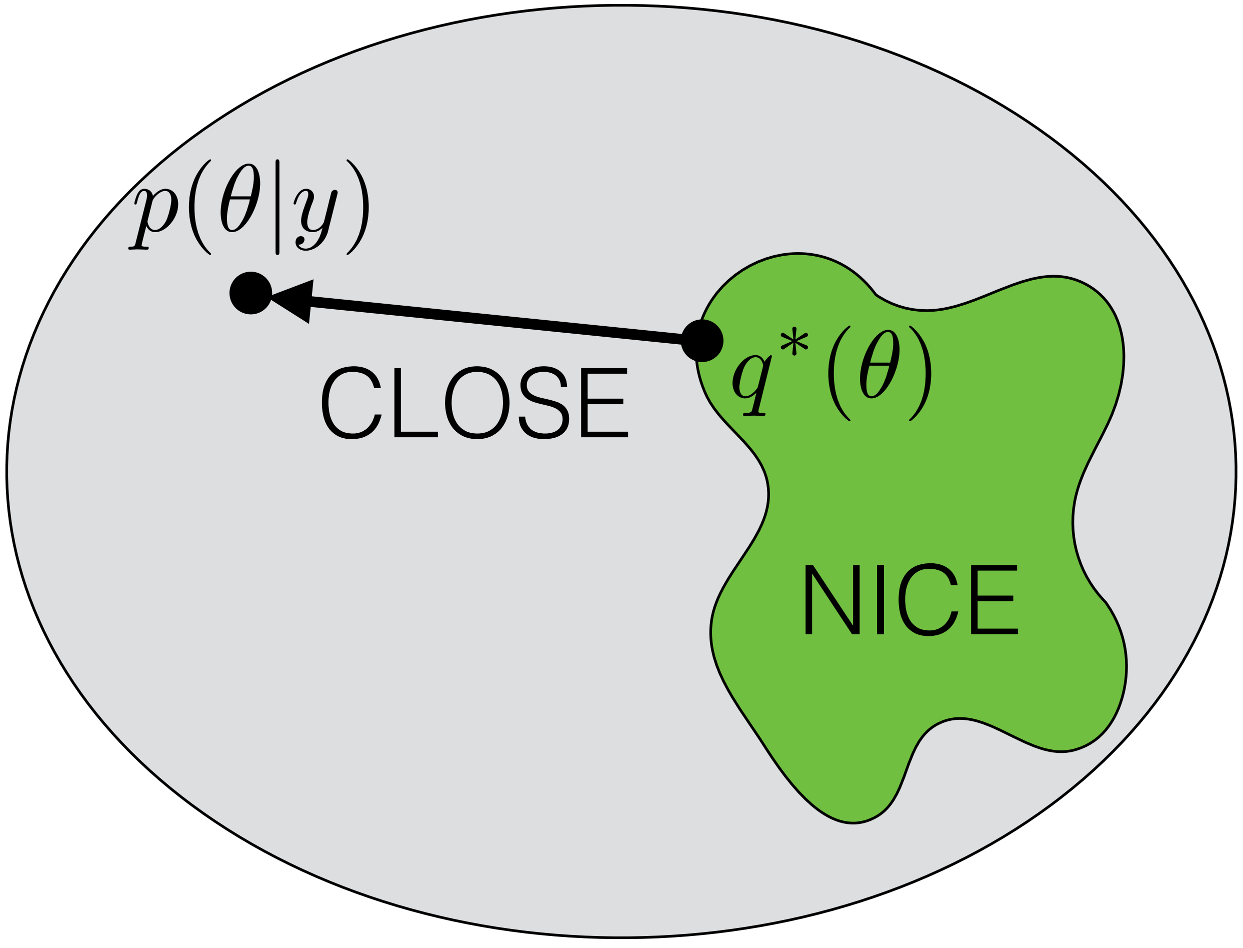
- VB practical success: point estimates and prediction, fast, streaming, distributed (3.6M Wikipedia, 350K Nature)

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

Why KL?

- Variational Bayes

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

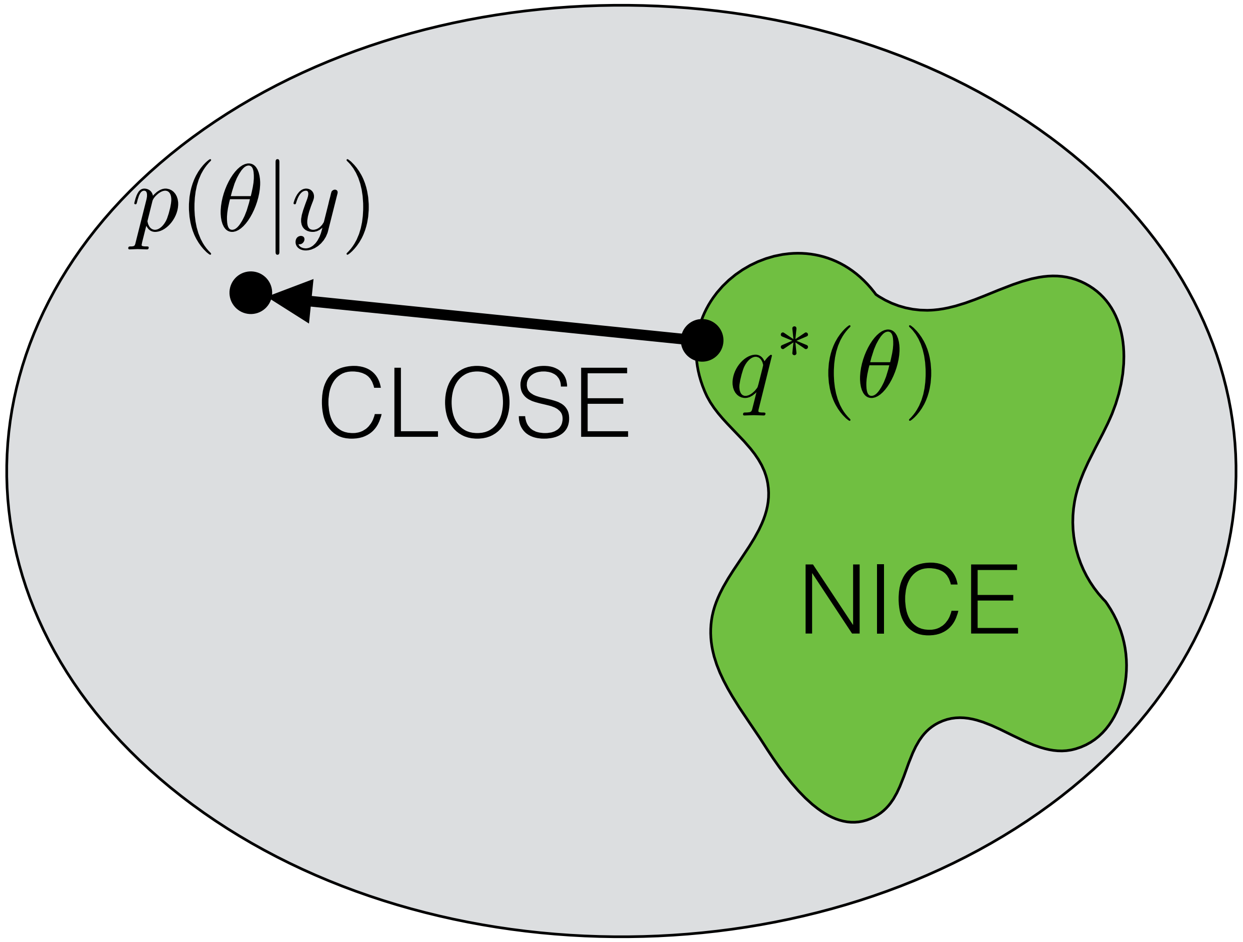


Why KL?

- Variational Bayes

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

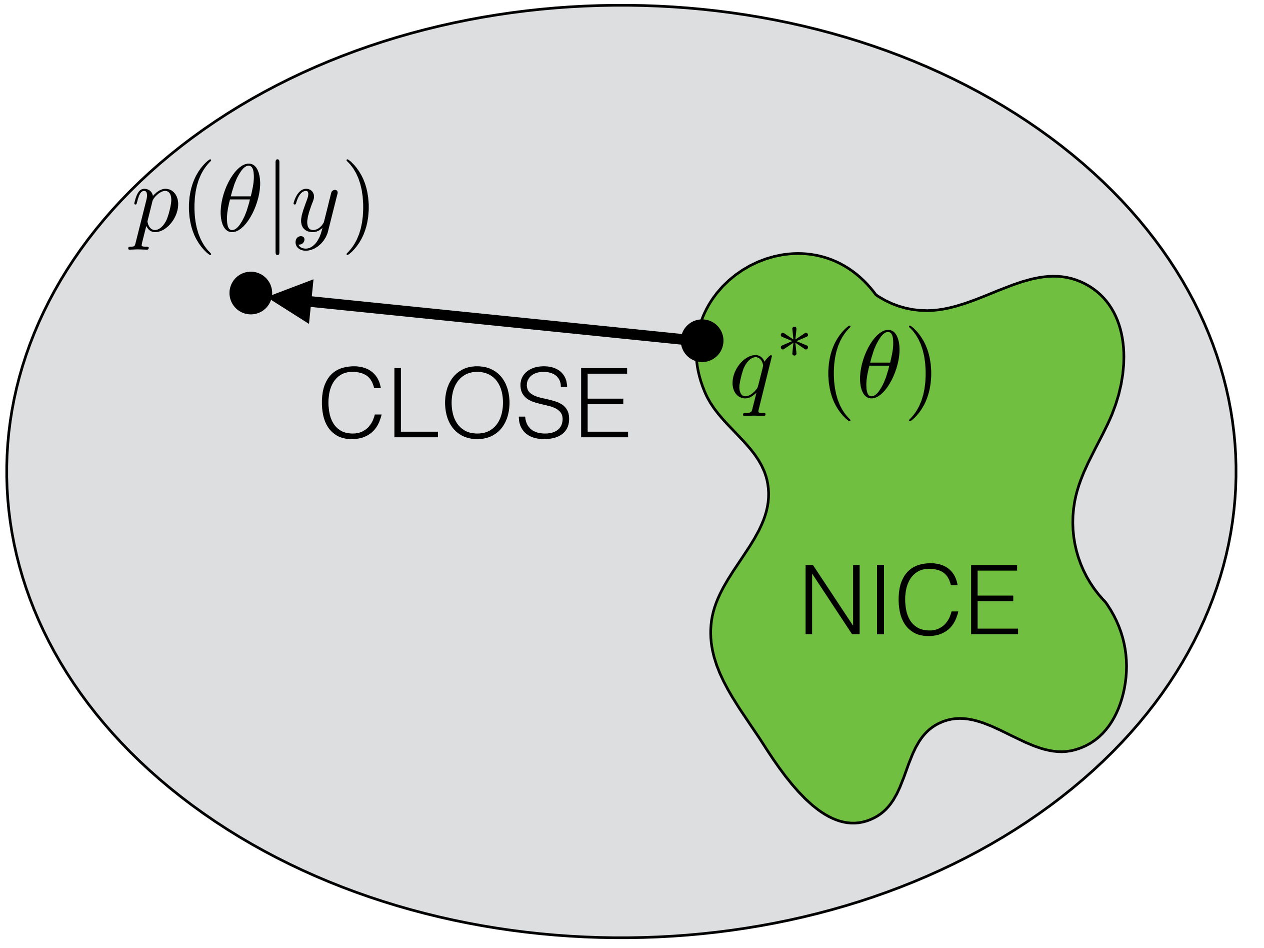
$$\text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$



Why KL?

- Variational Bayes

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



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

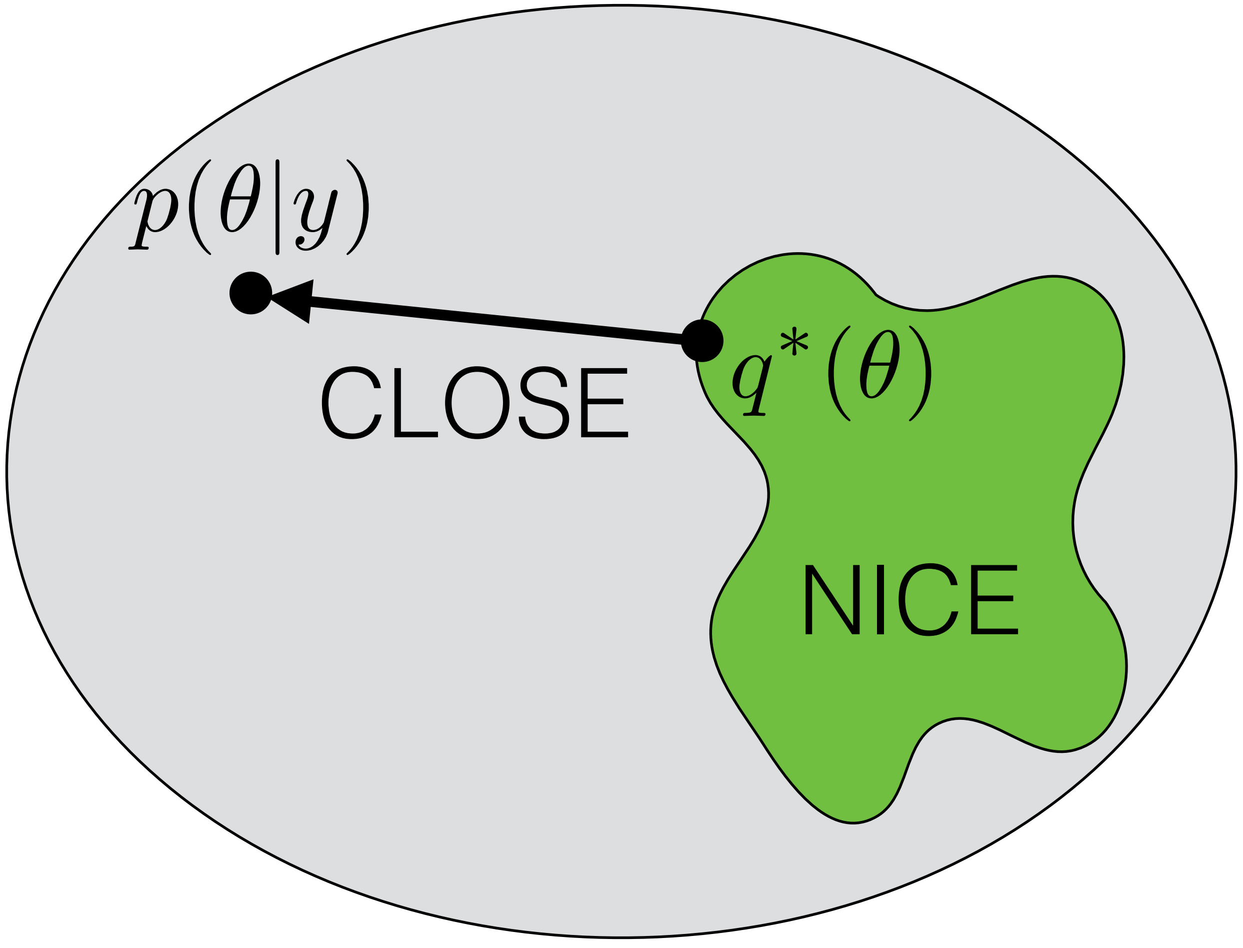
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

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

Why KL?

- Variational Bayes

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



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

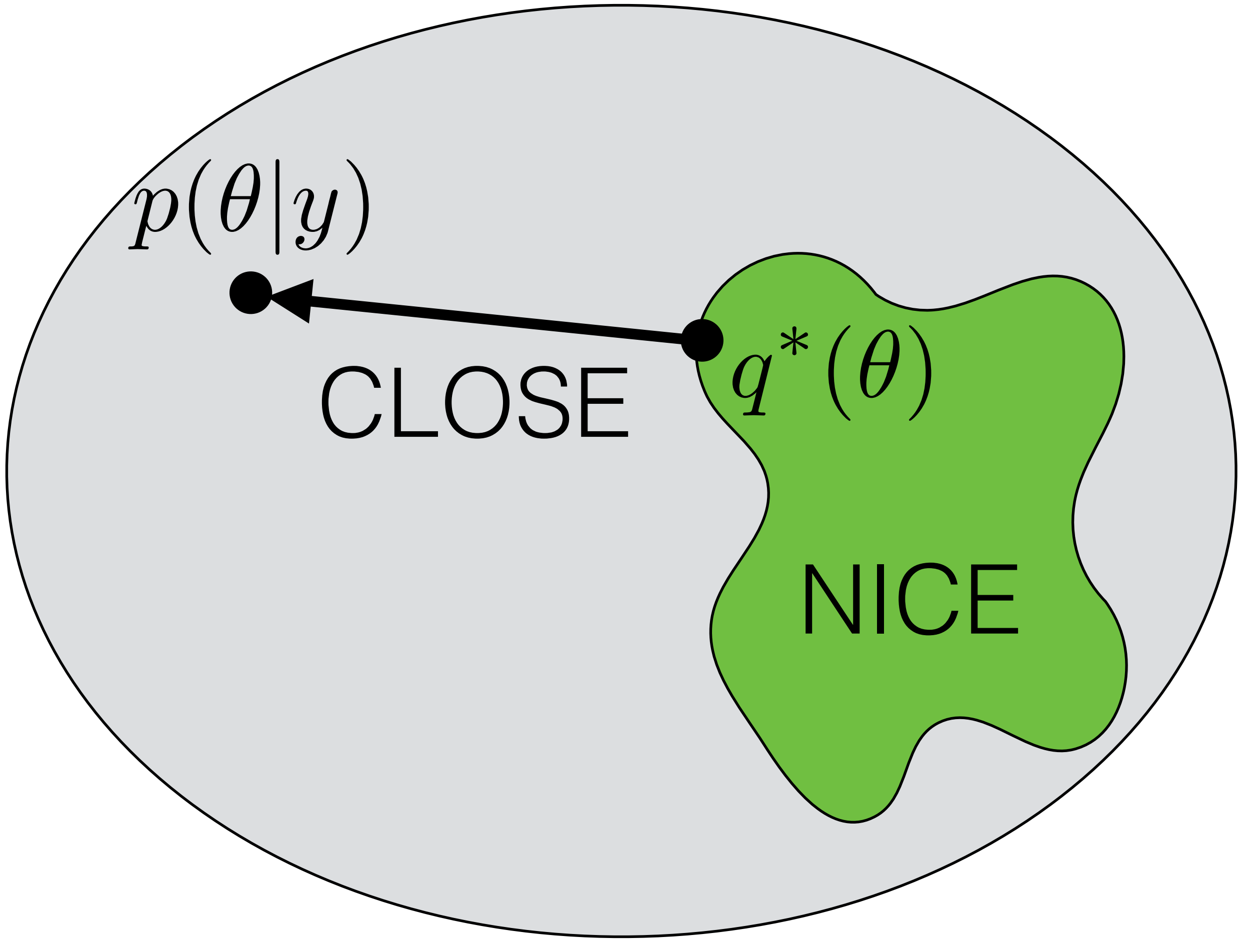
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

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

Why KL?

- Variational Bayes

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



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

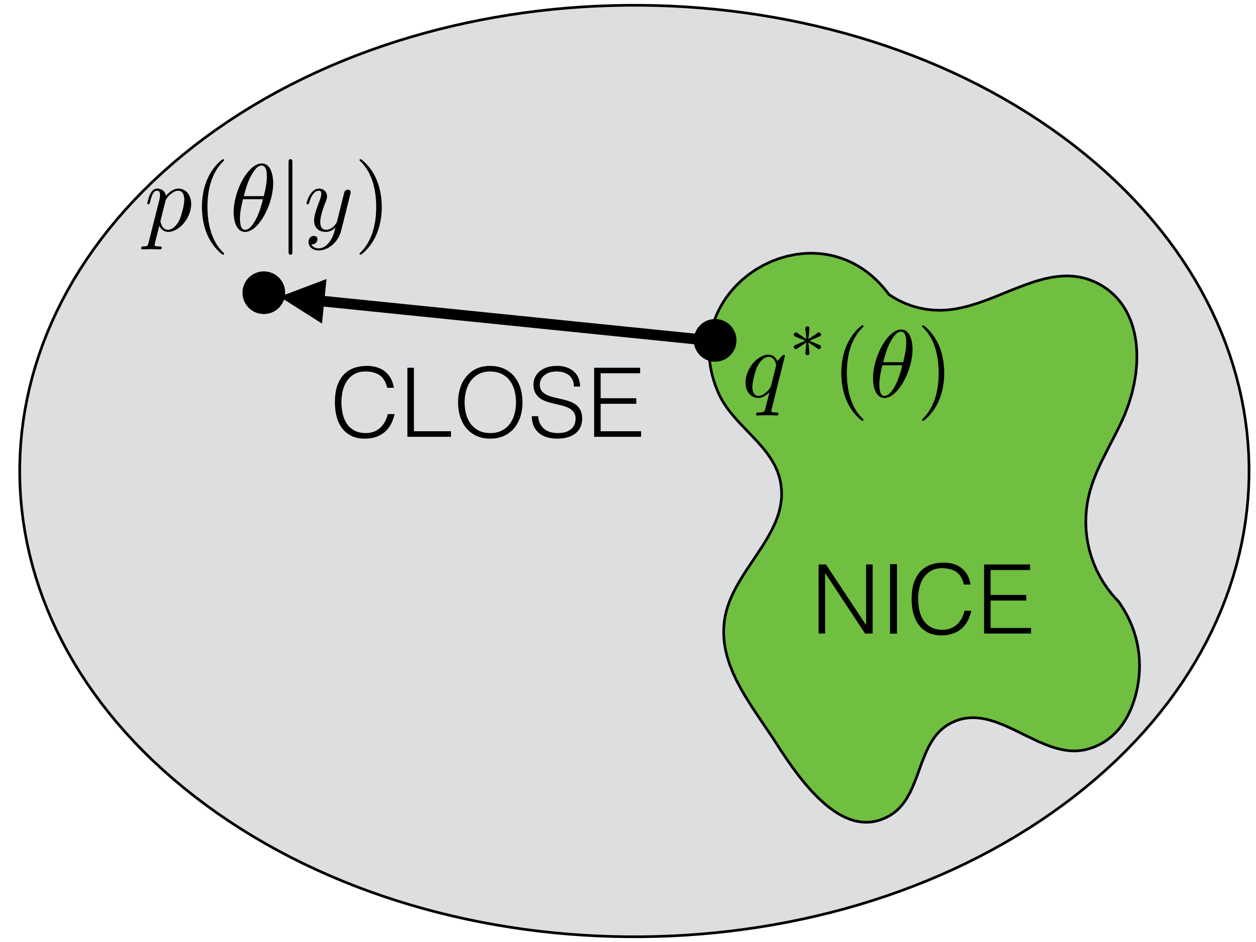
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

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

Why KL?

- Variational Bayes

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



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

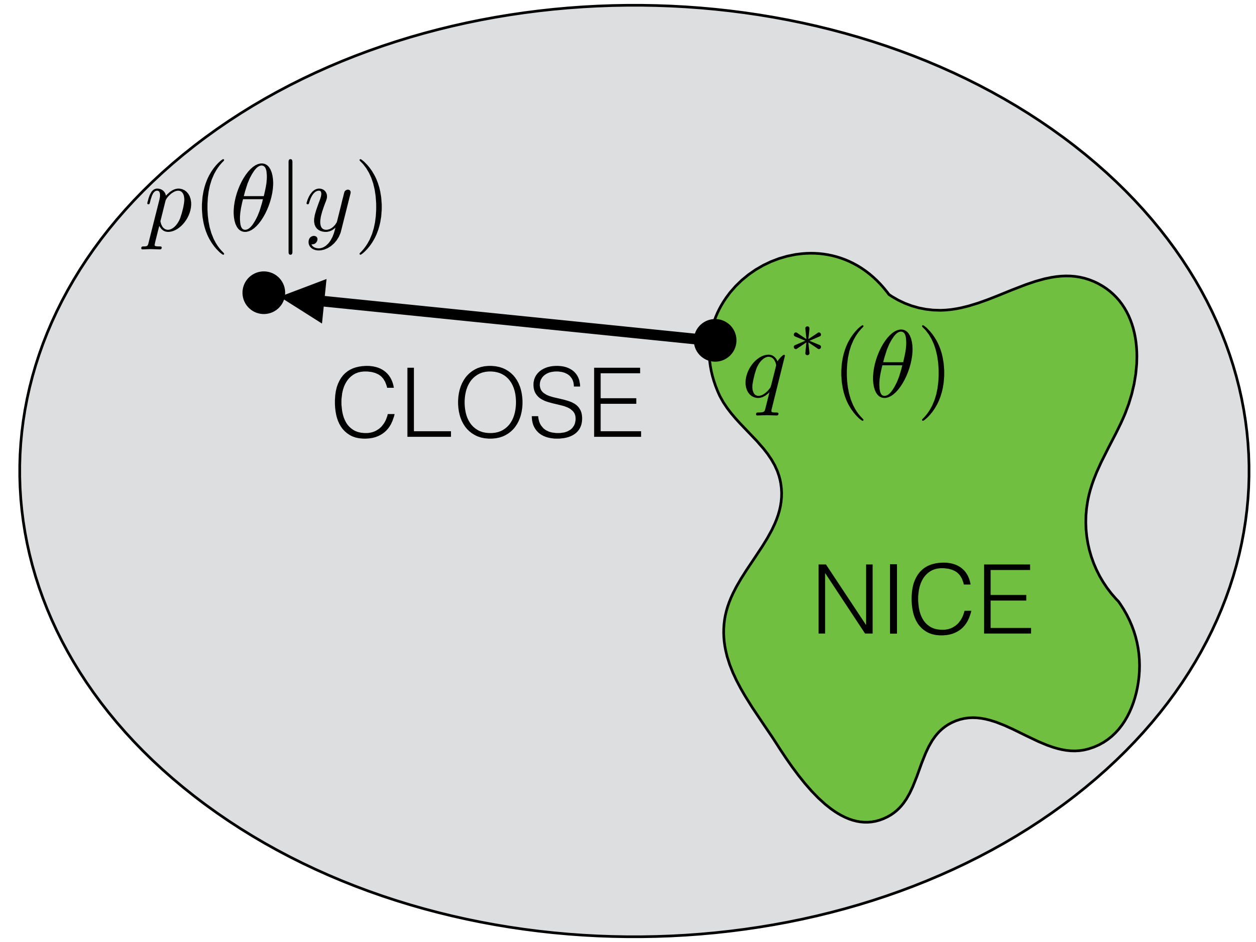
$$:= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta$$

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

Why KL?

- Variational Bayes

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



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

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

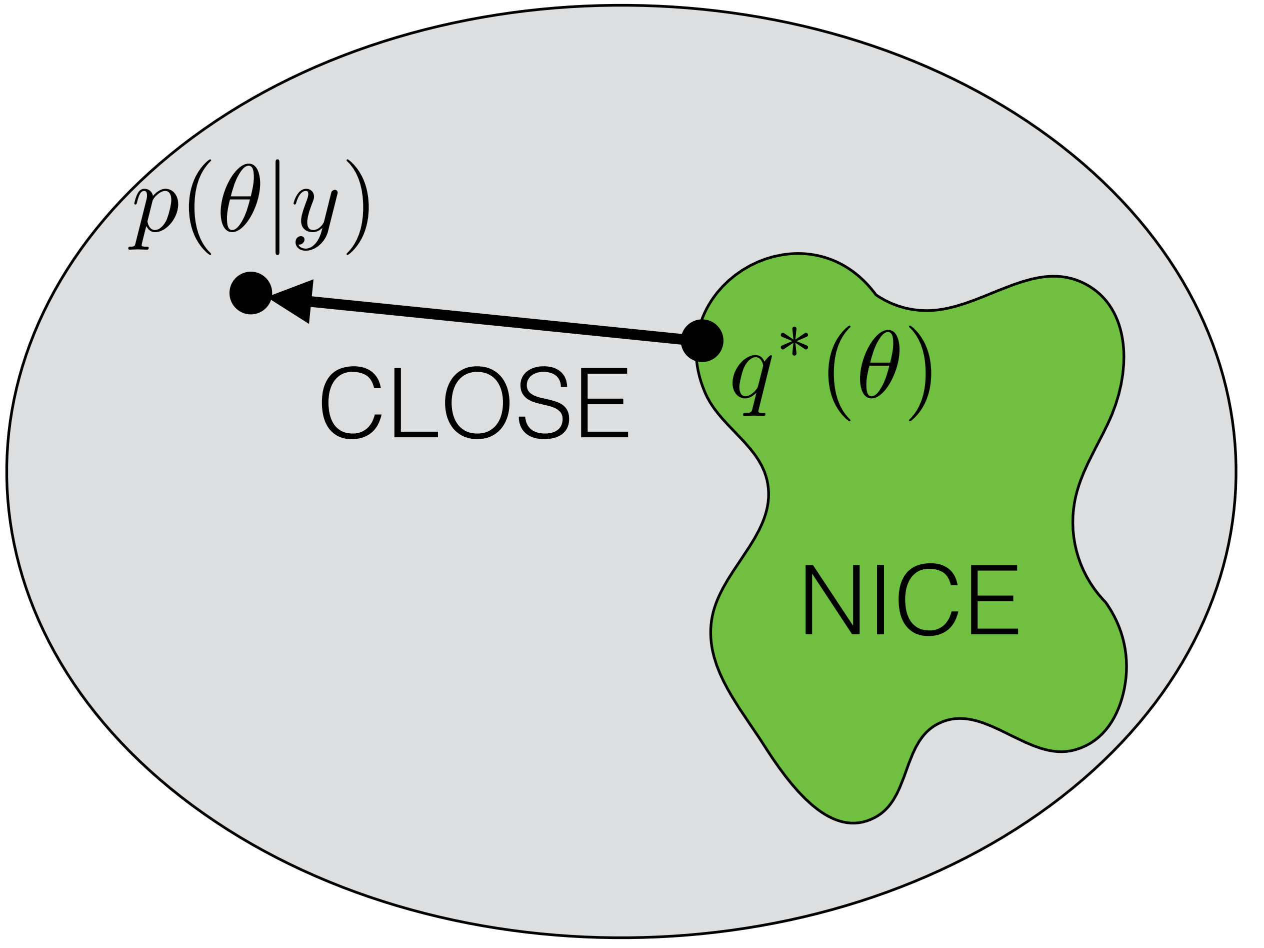
$$= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta$$

"Evidence lower bound" (ELBO)

Why KL?

- Variational Bayes

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



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

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

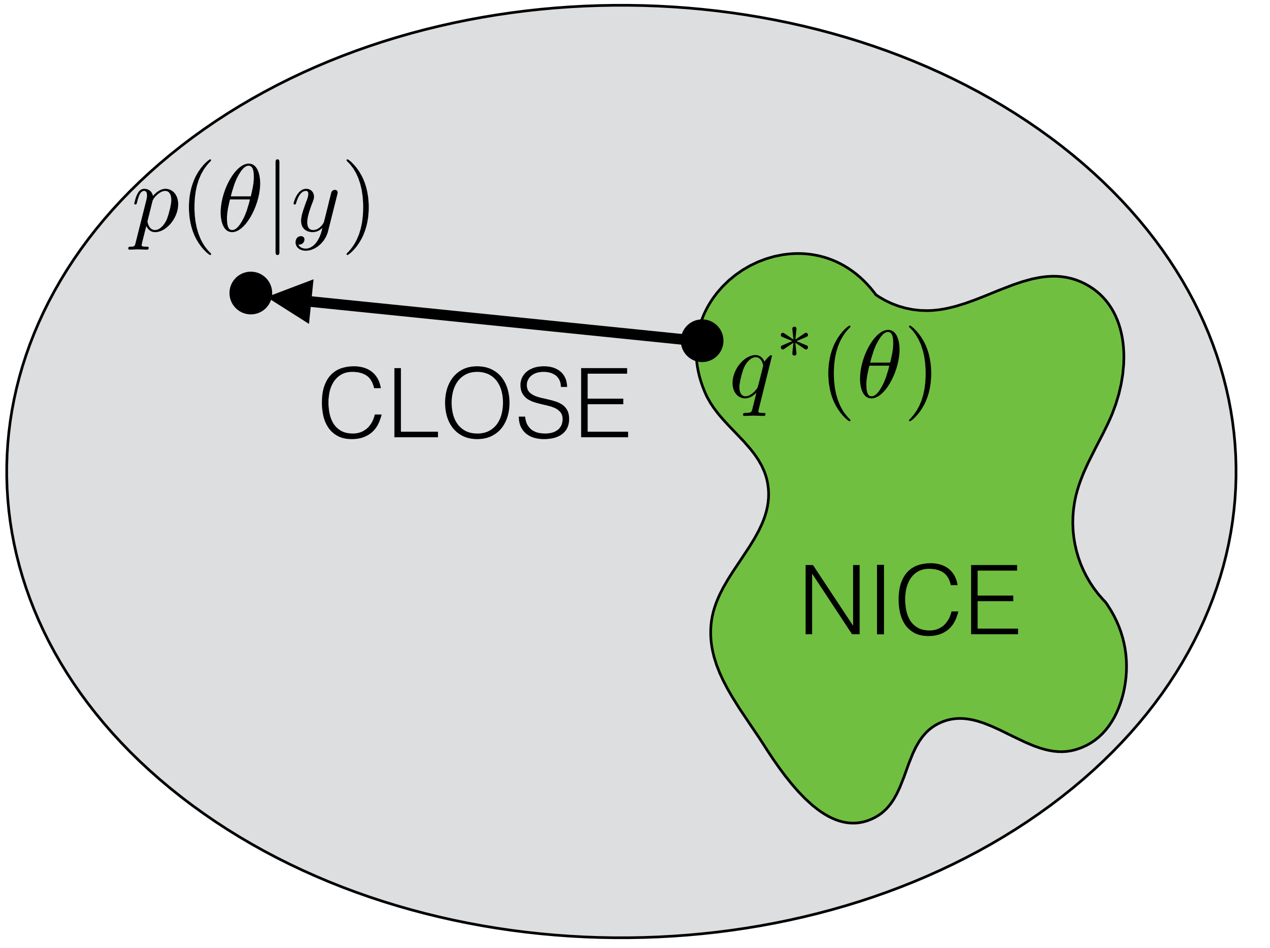
$$= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta$$

"Evidence lower bound" (ELBO)

Why KL?

- Variational Bayes

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



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

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

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

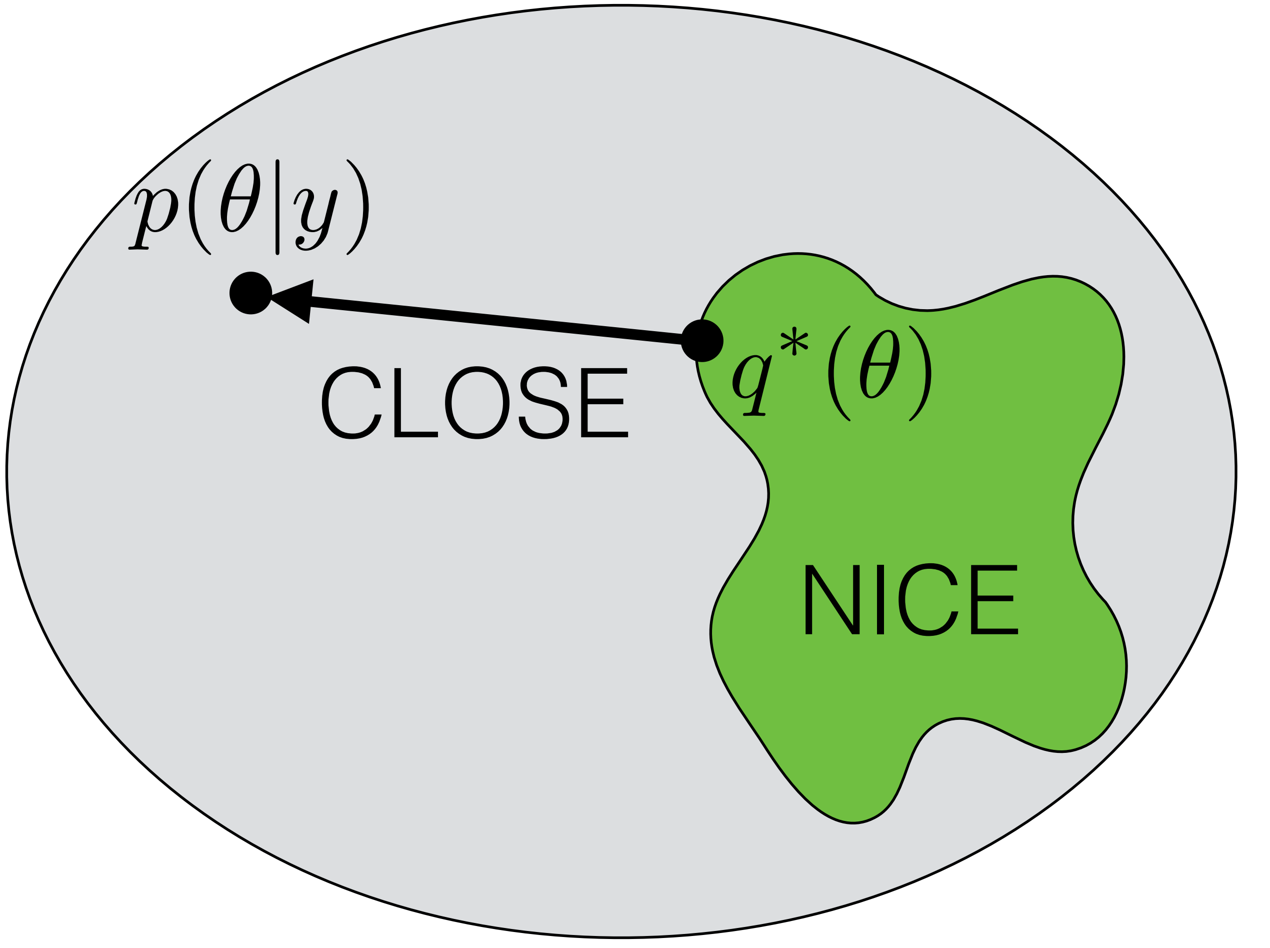
"Evidence lower bound" (ELBO)

- $q^* = \text{argmax}_{q \in Q} \text{ELBO}(q)$

Why KL?

- Variational Bayes

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



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

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

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

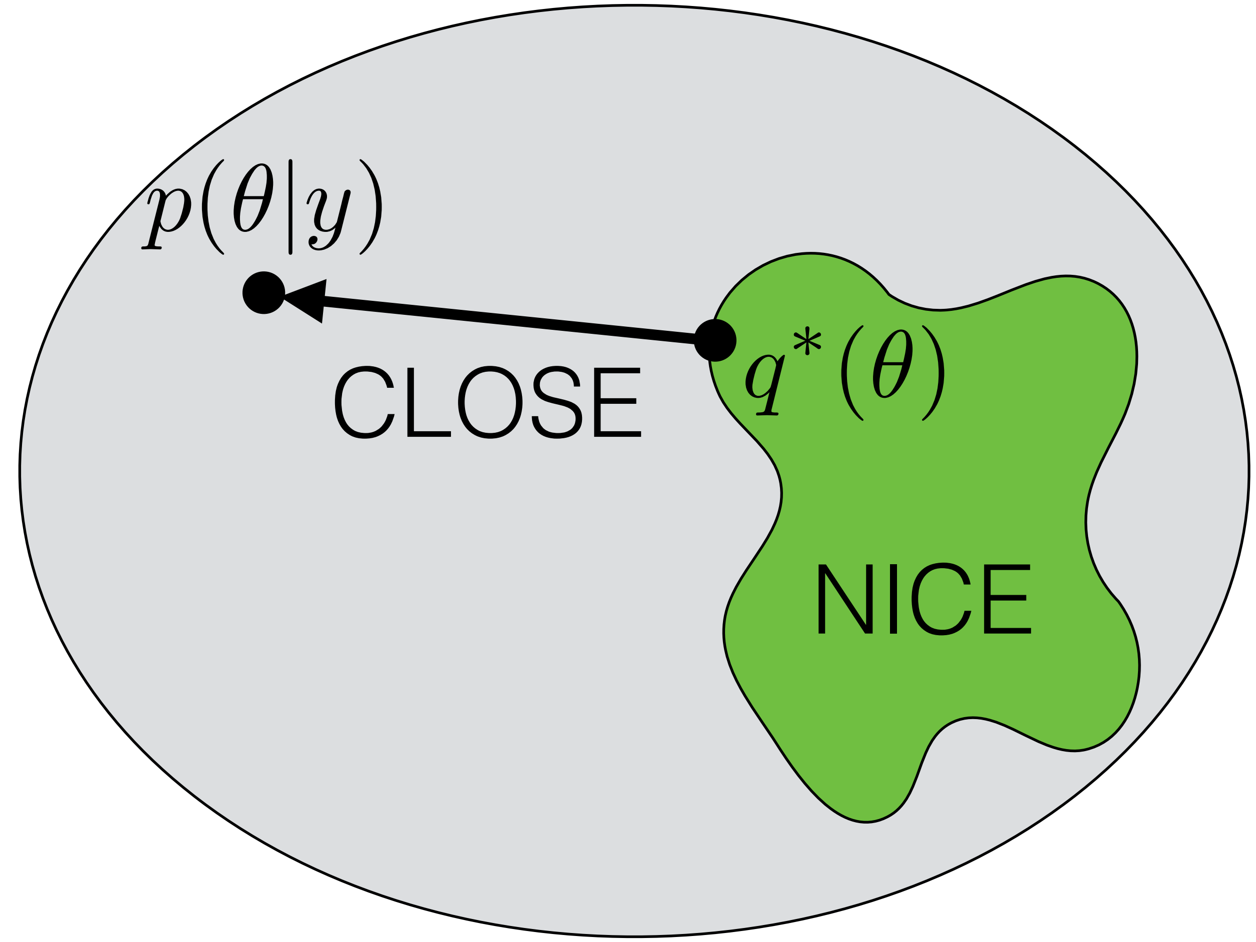
"Evidence lower bound" (ELBO)

- $q^* = \text{argmax}_{q \in Q} \text{ELBO}(q)$
- KL is positive definite [Board; Bishop 2006, Sec 1.6.1]

Why KL?

- Variational Bayes

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



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

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

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

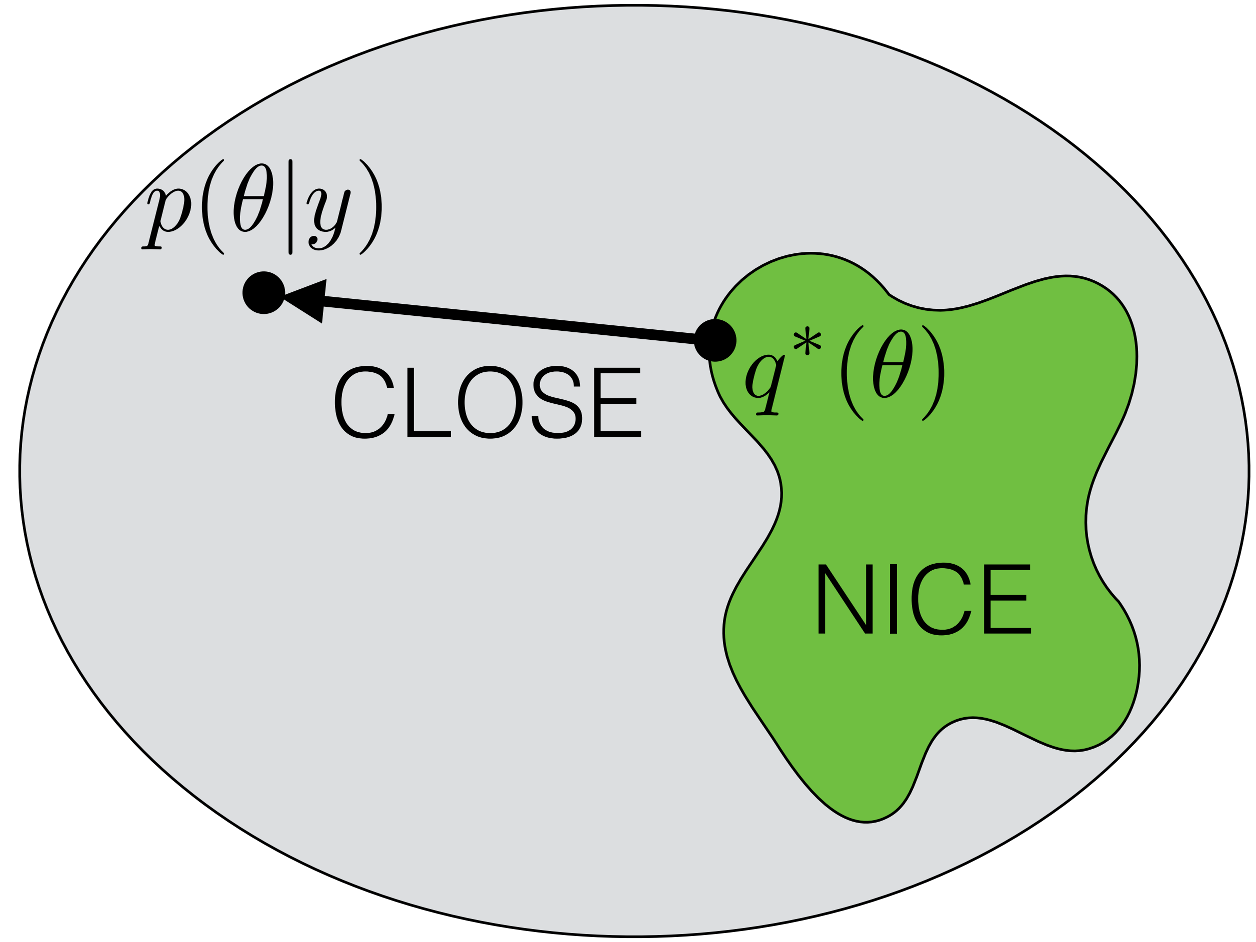
"Evidence lower bound" (ELBO)

- $q^* = \text{argmax}_{q \in Q} \text{ELBO}(q)$
- KL is positive definite [Board; Bishop 2006, Sec 1.6.1]
- $\text{KL} \geq 0 \Rightarrow \log p(y) \geq \text{ELBO}$

Why KL?

- Variational Bayes

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



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

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

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

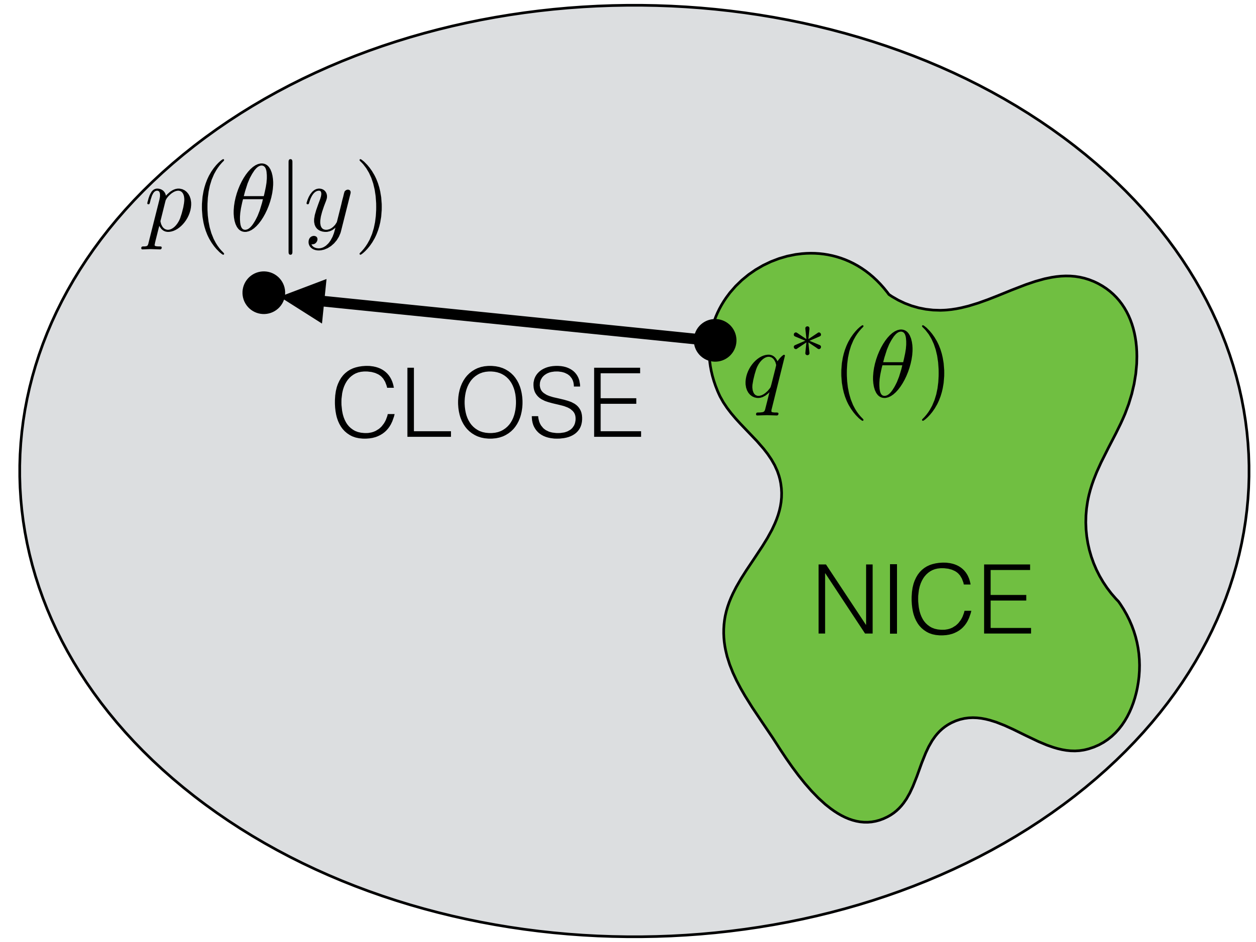
"Evidence lower bound" (ELBO)

- $q^* = \text{argmax}_{q \in Q} \text{ELBO}(q)$
- KL is positive definite [Board; Bishop 2006, Sec 1.6.1]
- $\text{KL} \geq 0 \Rightarrow \log p(y) \geq \text{ELBO}$
- Why KL?

Why KL?

- Variational Bayes

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



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

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

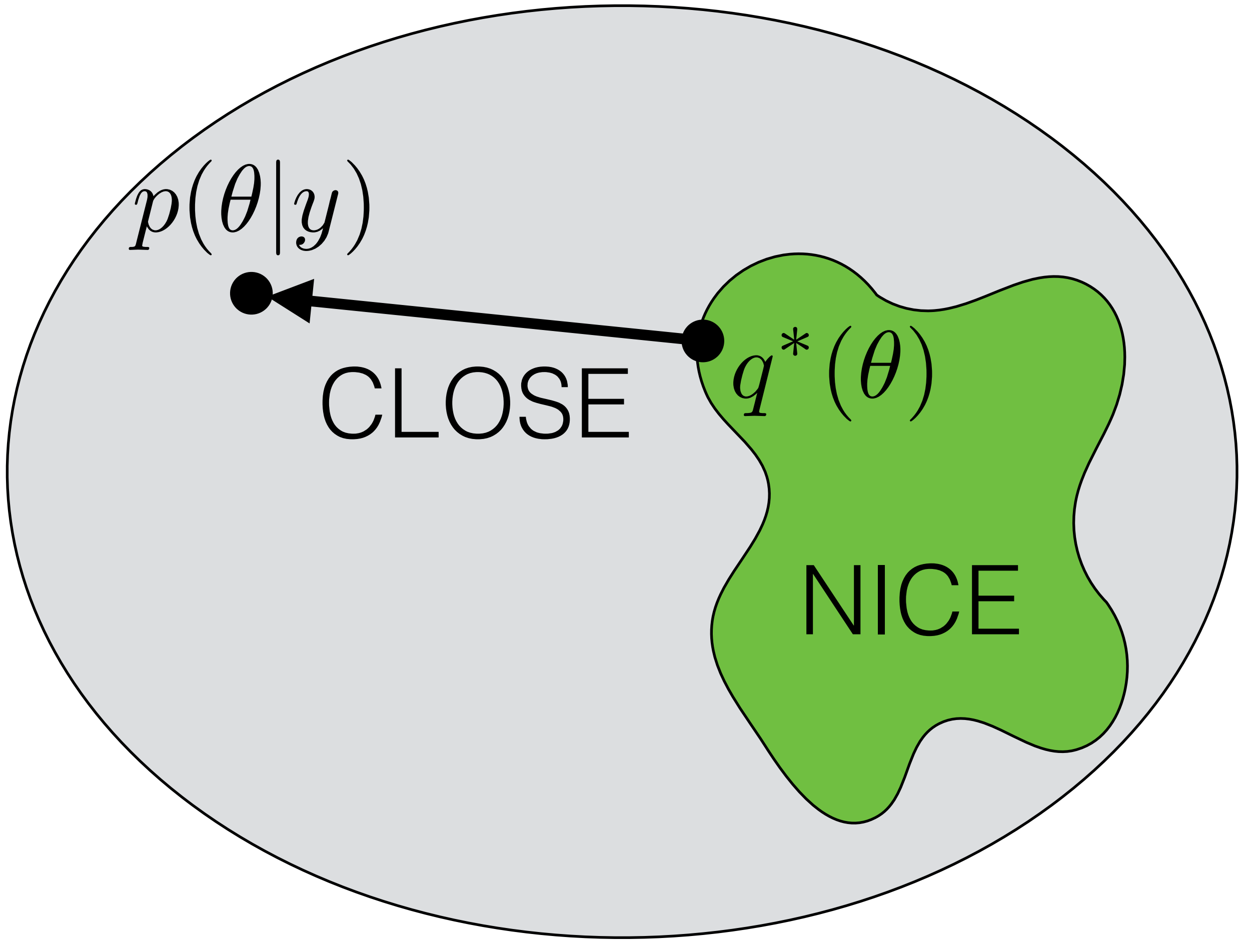
$$= \int q(\theta) \log \frac{q(\theta) p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta$$

"Evidence lower bound" (ELBO)

- $q^* = \text{argmax}_{q \in Q} \text{ELBO}(q)$
- KL is positive definite [Board; Bishop 2006, Sec 1.6.1]
- $\text{KL} \geq 0 \Rightarrow \log p(y) \geq \text{ELBO}$
- Why KL (in this direction)?

Variational Bayes

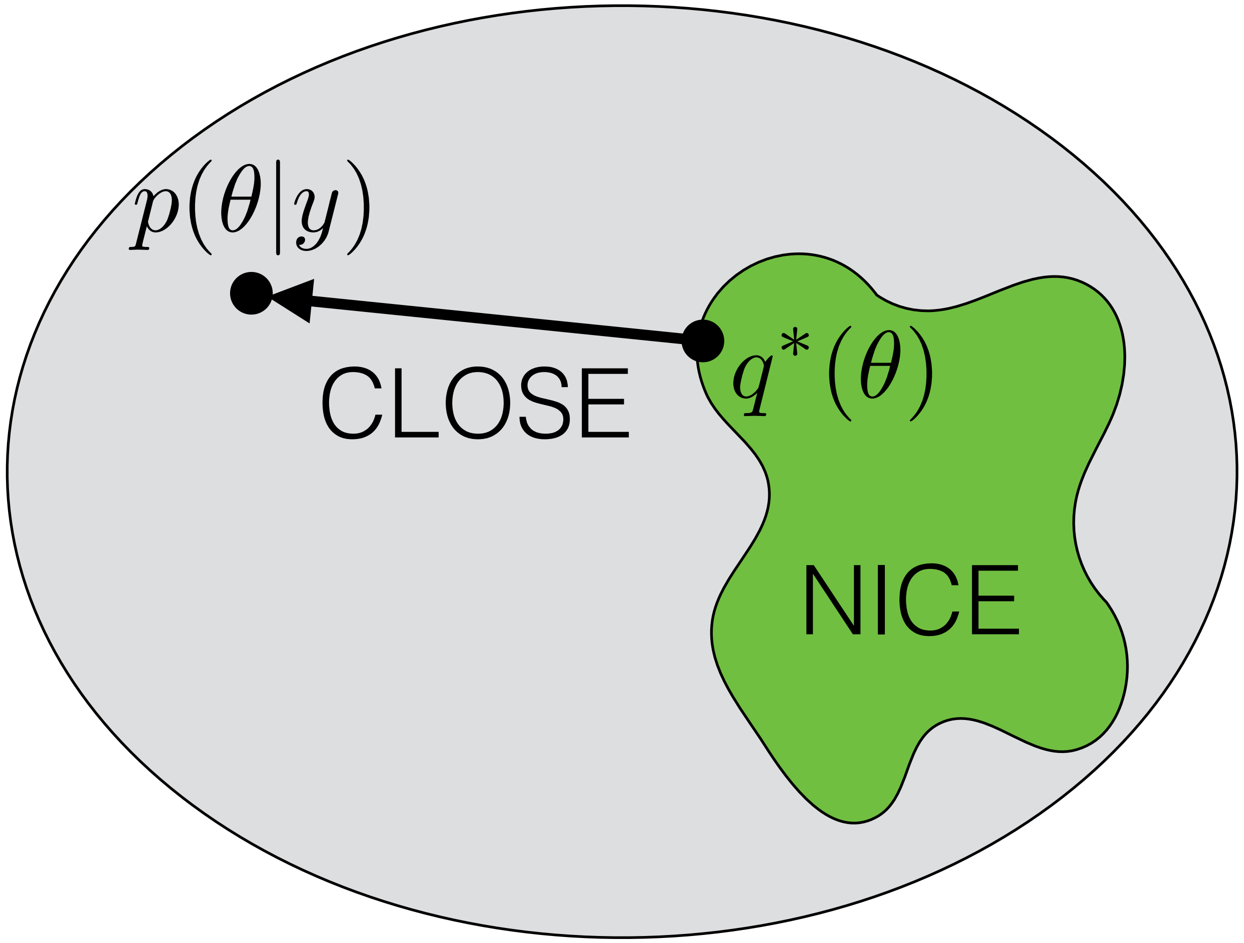
$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot|y)\right)$



Variational Bayes

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

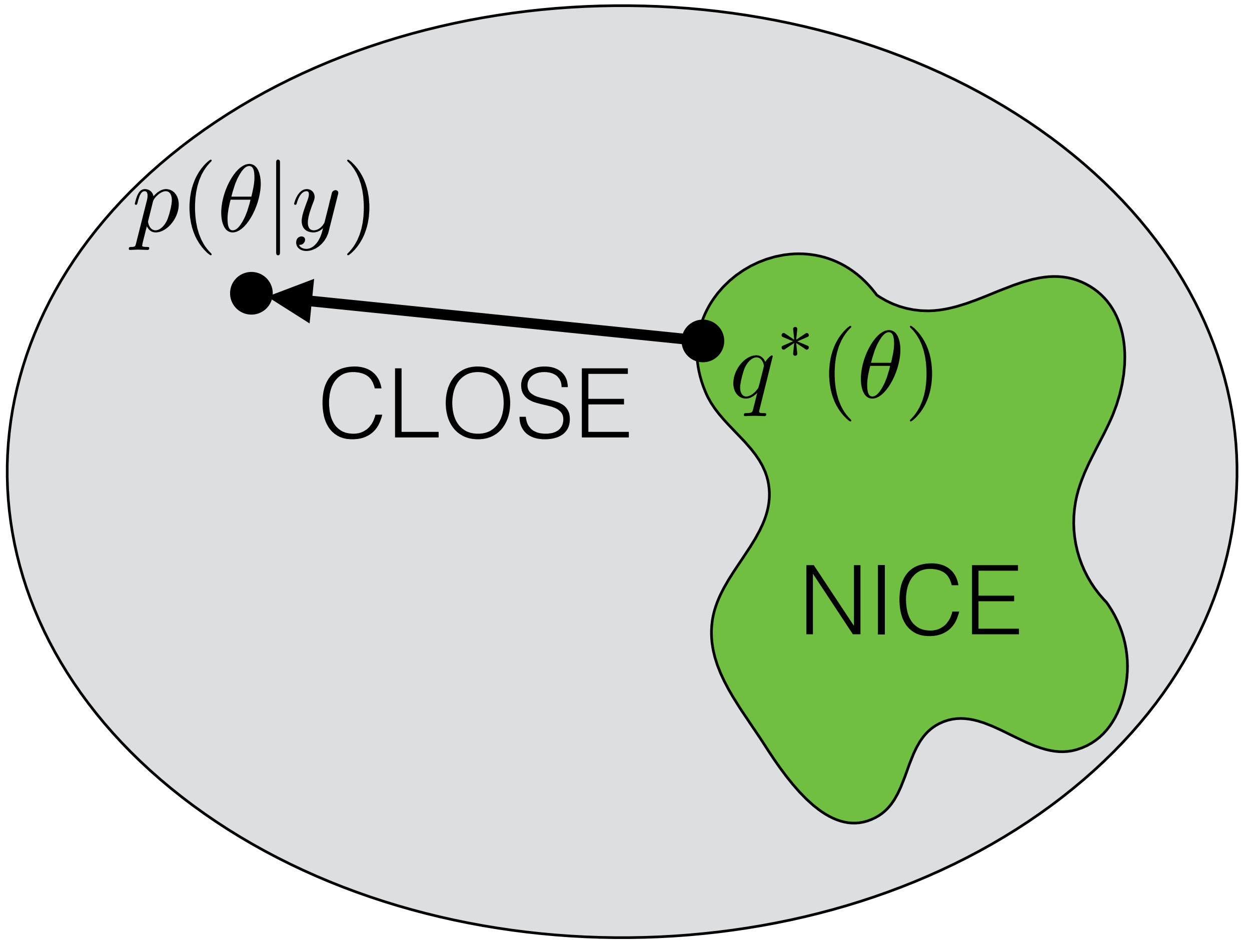
Choose "NICE" distributions



Variational Bayes

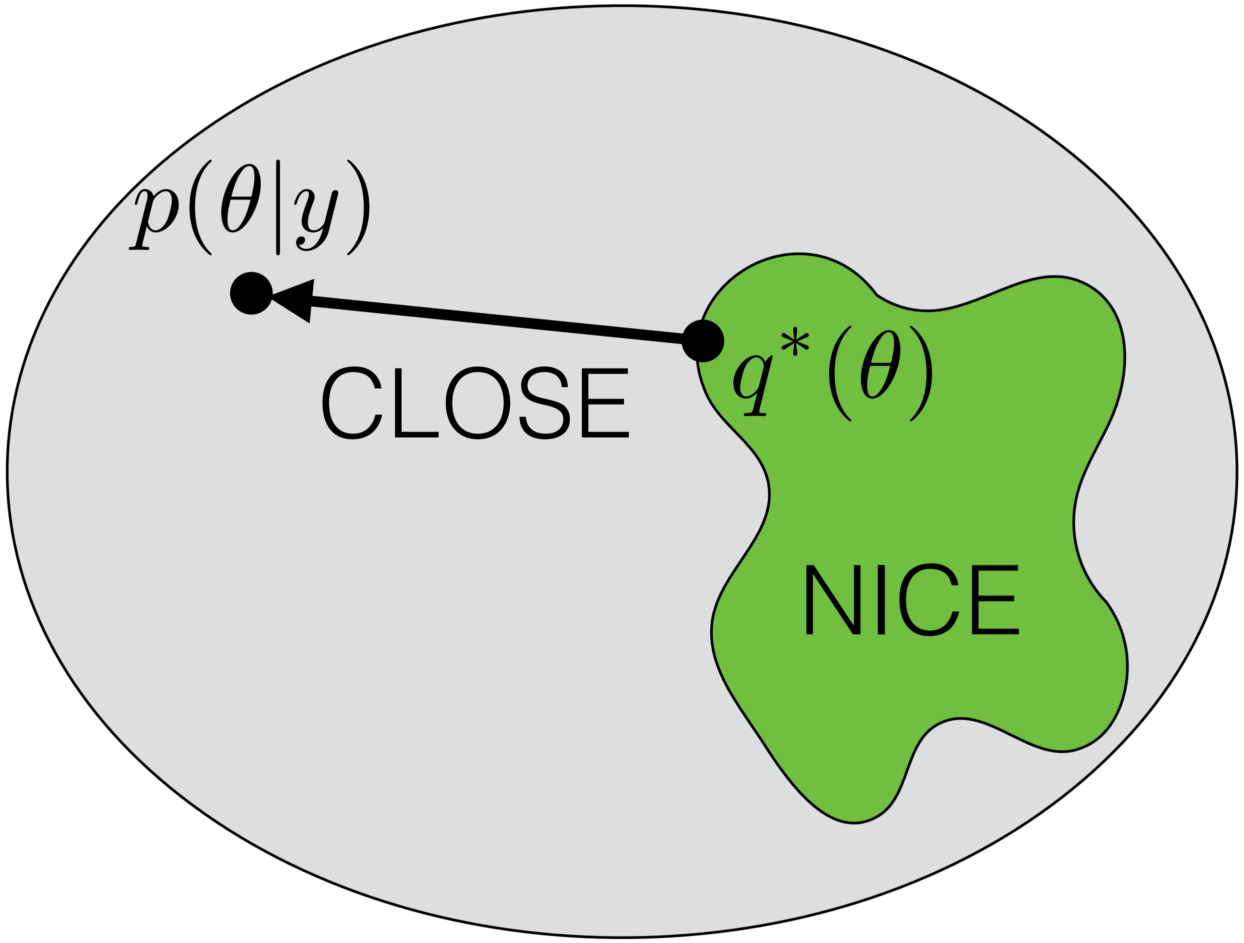
$$q^* = \operatorname{argmin}_{q \in Q} \mathrm{KL}\left(q(\cdot) || p(\cdot|y)\right)$$

Choose "NICE" distributions



Variational Bayes

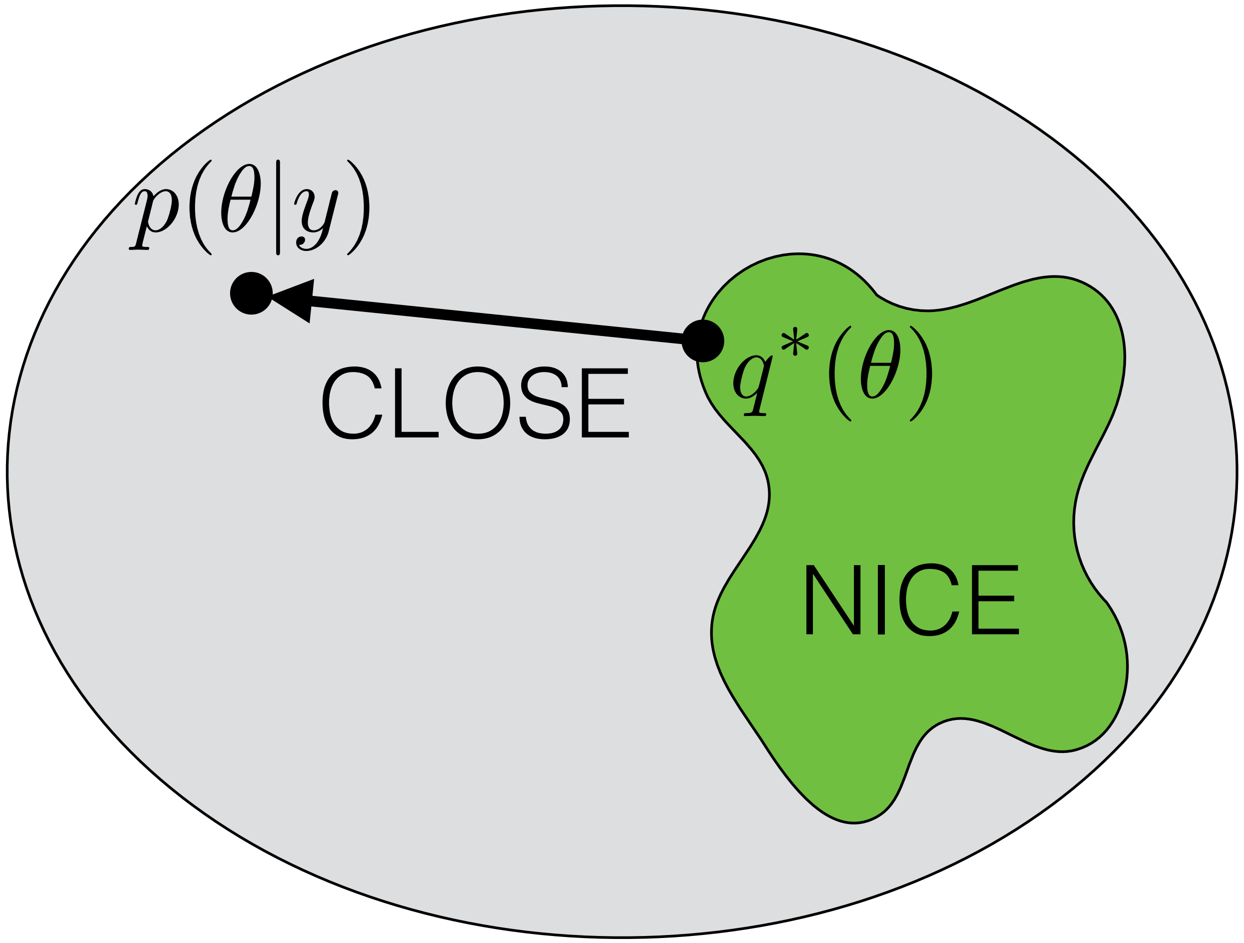
$$q^* = \text{argmin}_{q \in Q} \text{KL}\left(q(\cdot) || p(\cdot|y)\right)$$



Choose "NICE" distributions

Variational Bayes

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



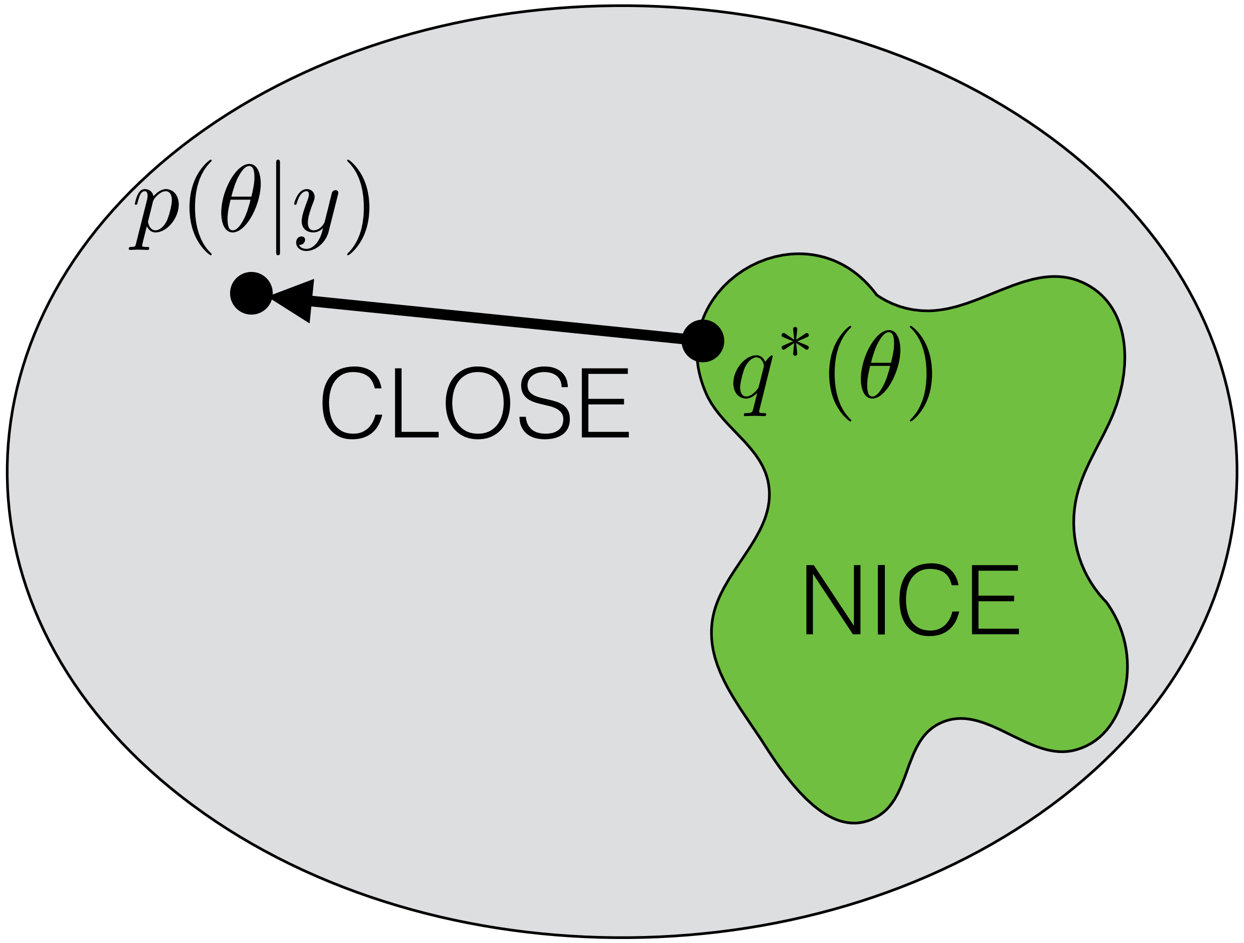
Choose "NICE" distributions

- Mean-field variational Bayes (MFVB)

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

Variational Bayes

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



Choose "NICE" distributions

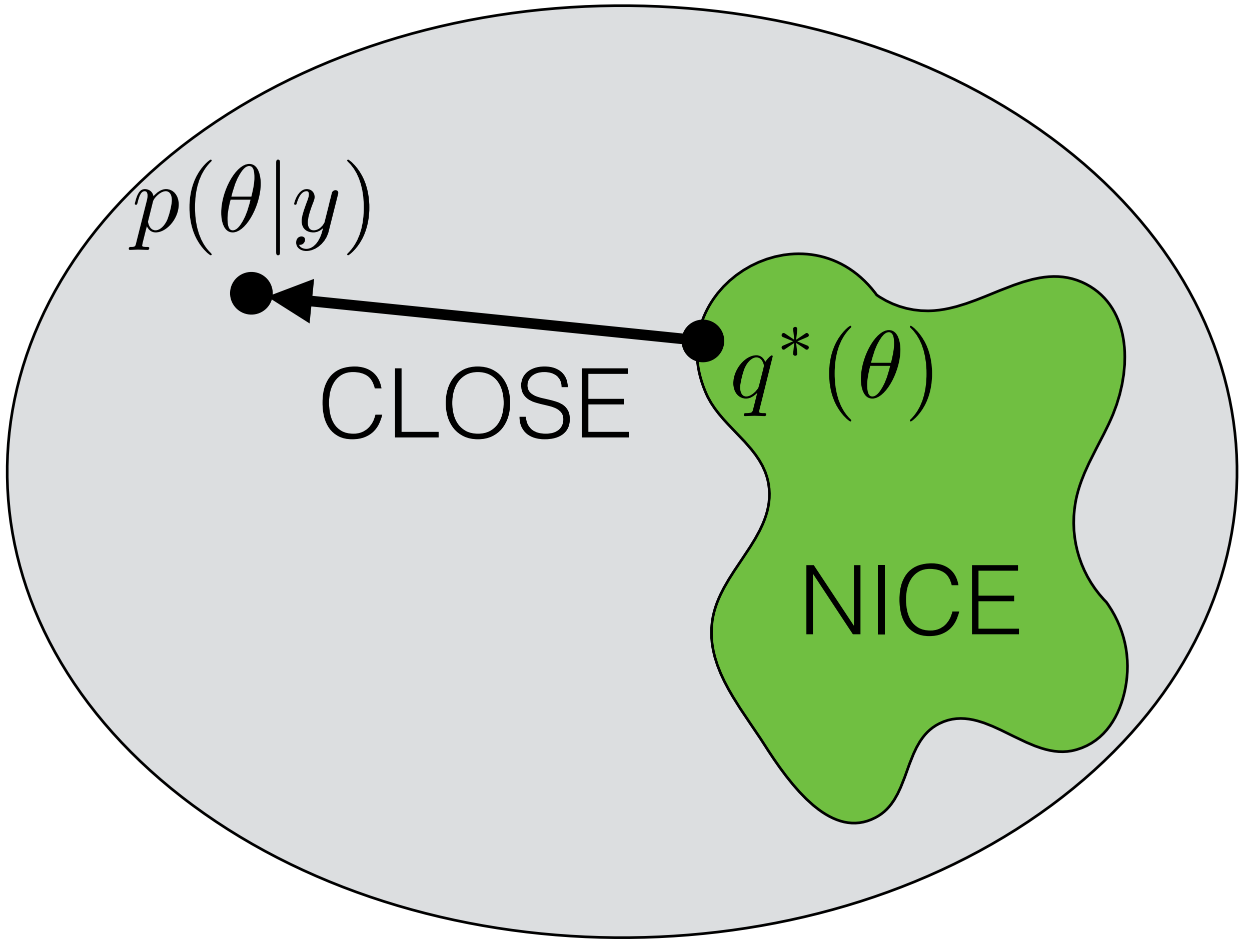
- Mean-field variational Bayes (MFVB)

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

- Often also exponential family

Variational Bayes

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



Choose "NICE" distributions

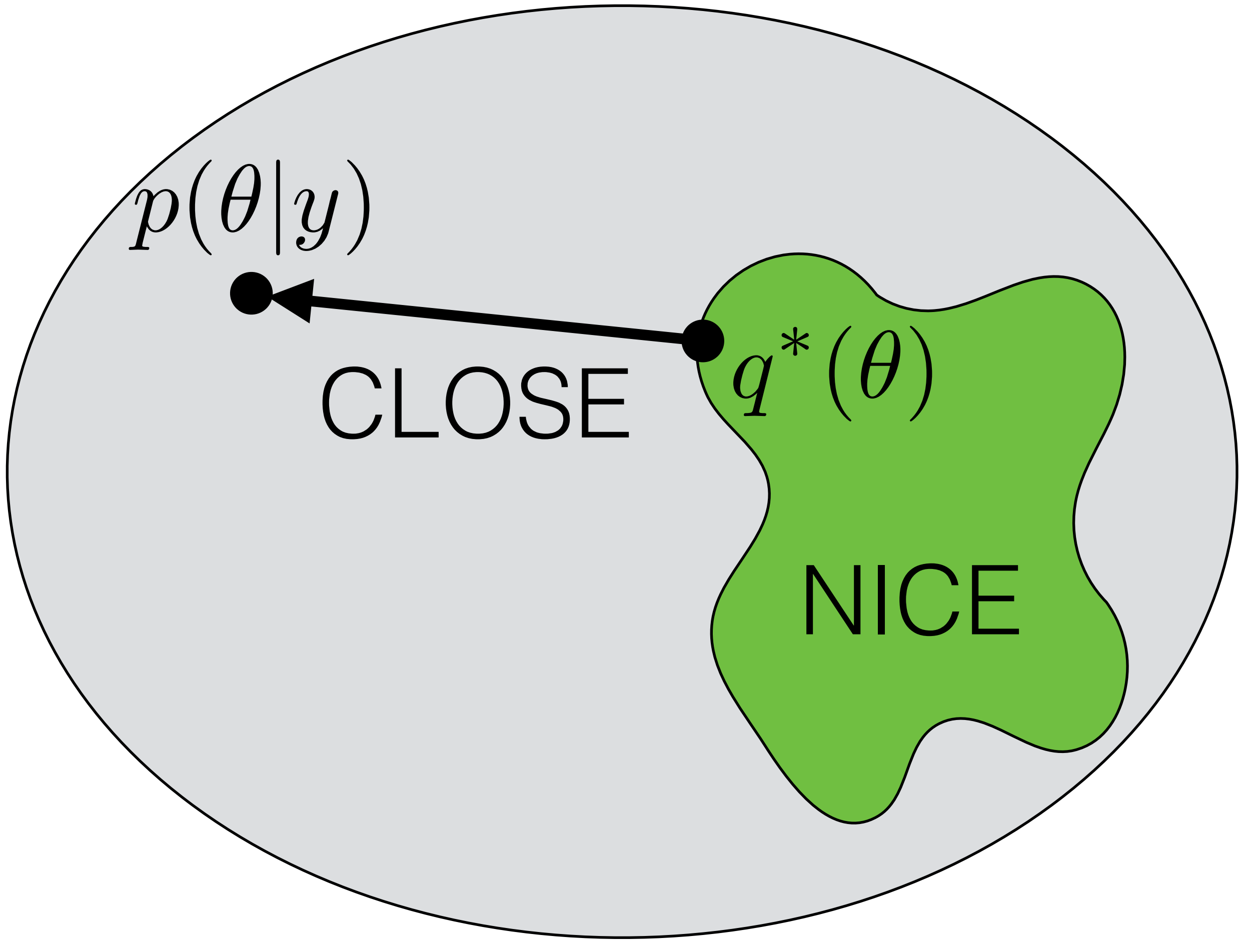
- Mean-field variational Bayes (MFVB)

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

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

Variational Bayes

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

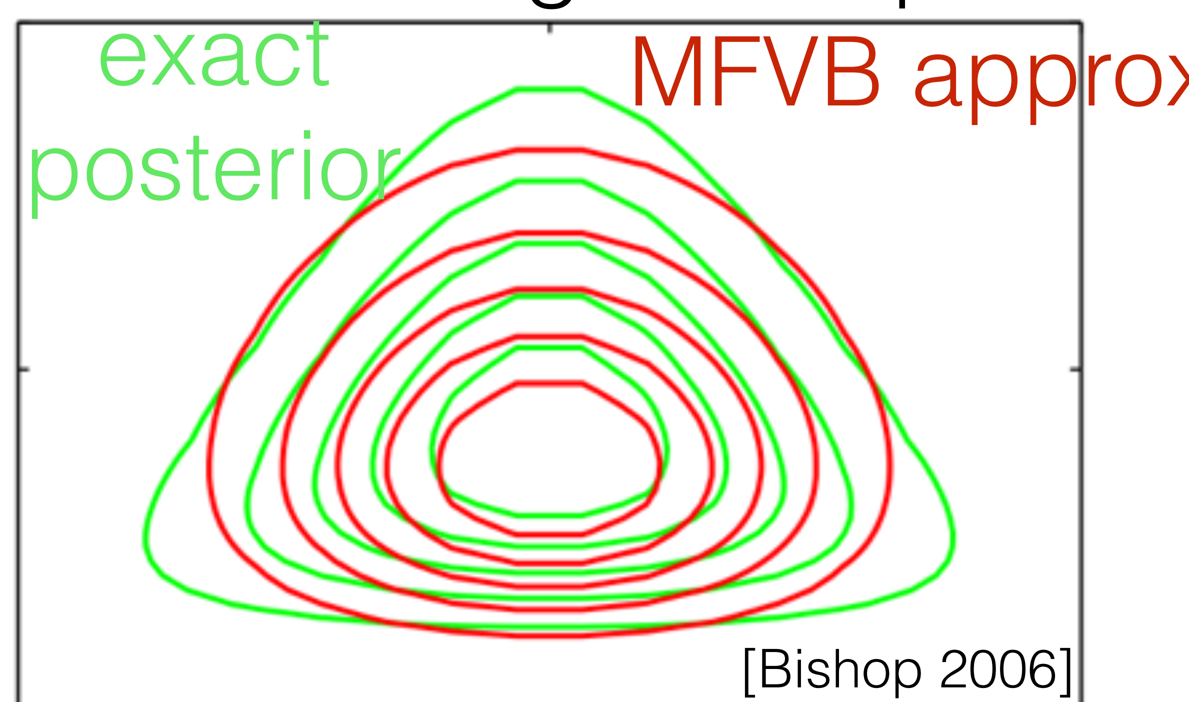


Choose "NICE" distributions

- Mean-field variational Bayes (MFVB)

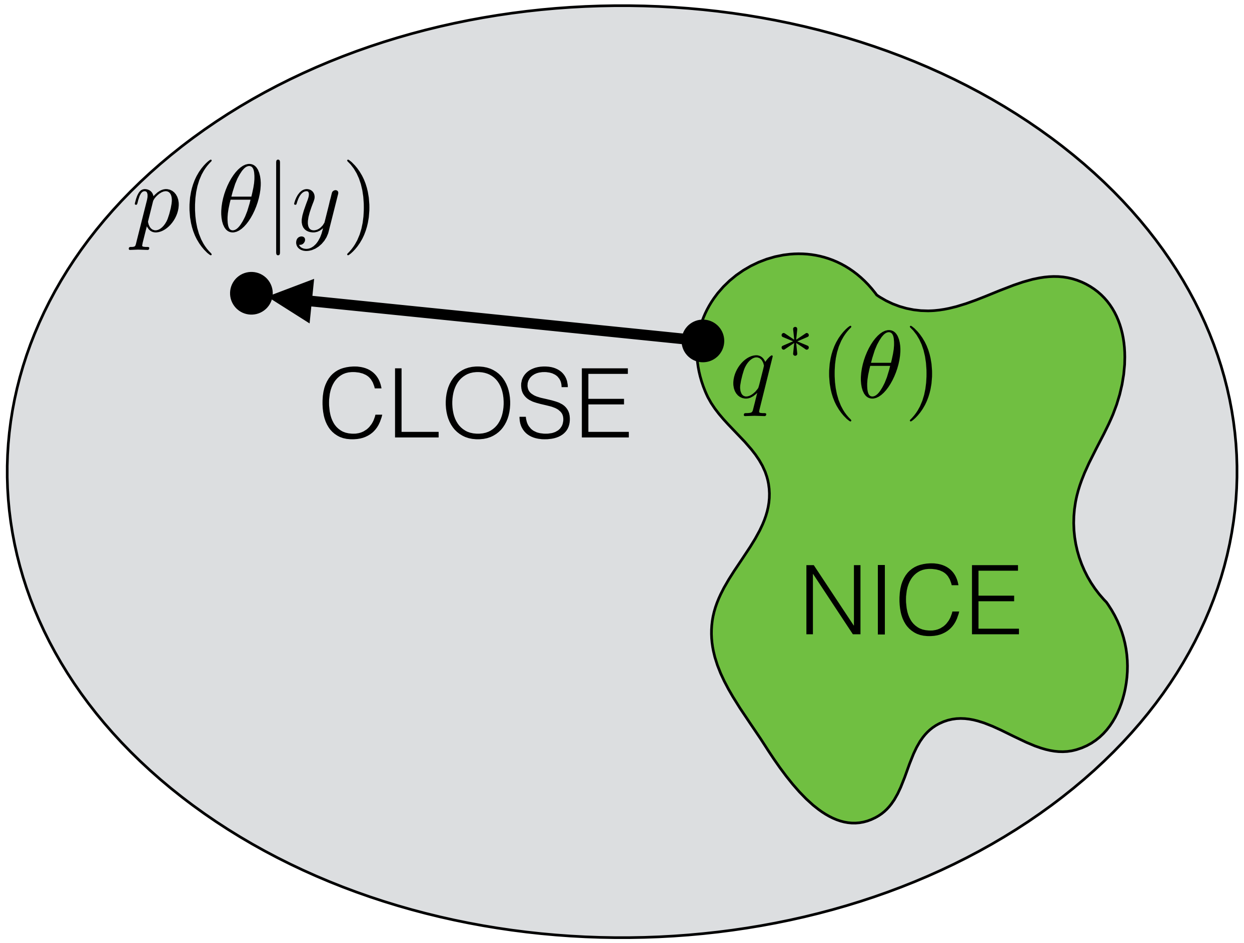
$$Q_{MFVB} := \left\{ q : q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \right\}$$

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



Variational Bayes

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

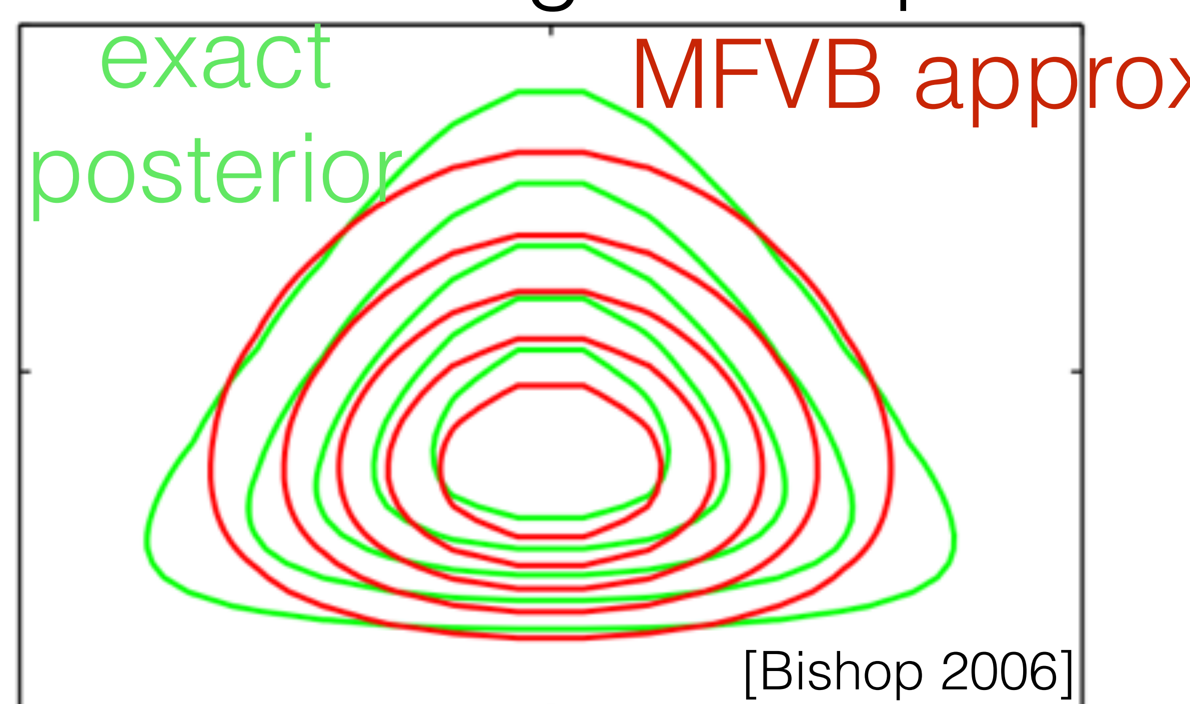


Choose “NICE” distributions

- Mean-field variational Bayes (MFVB)

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

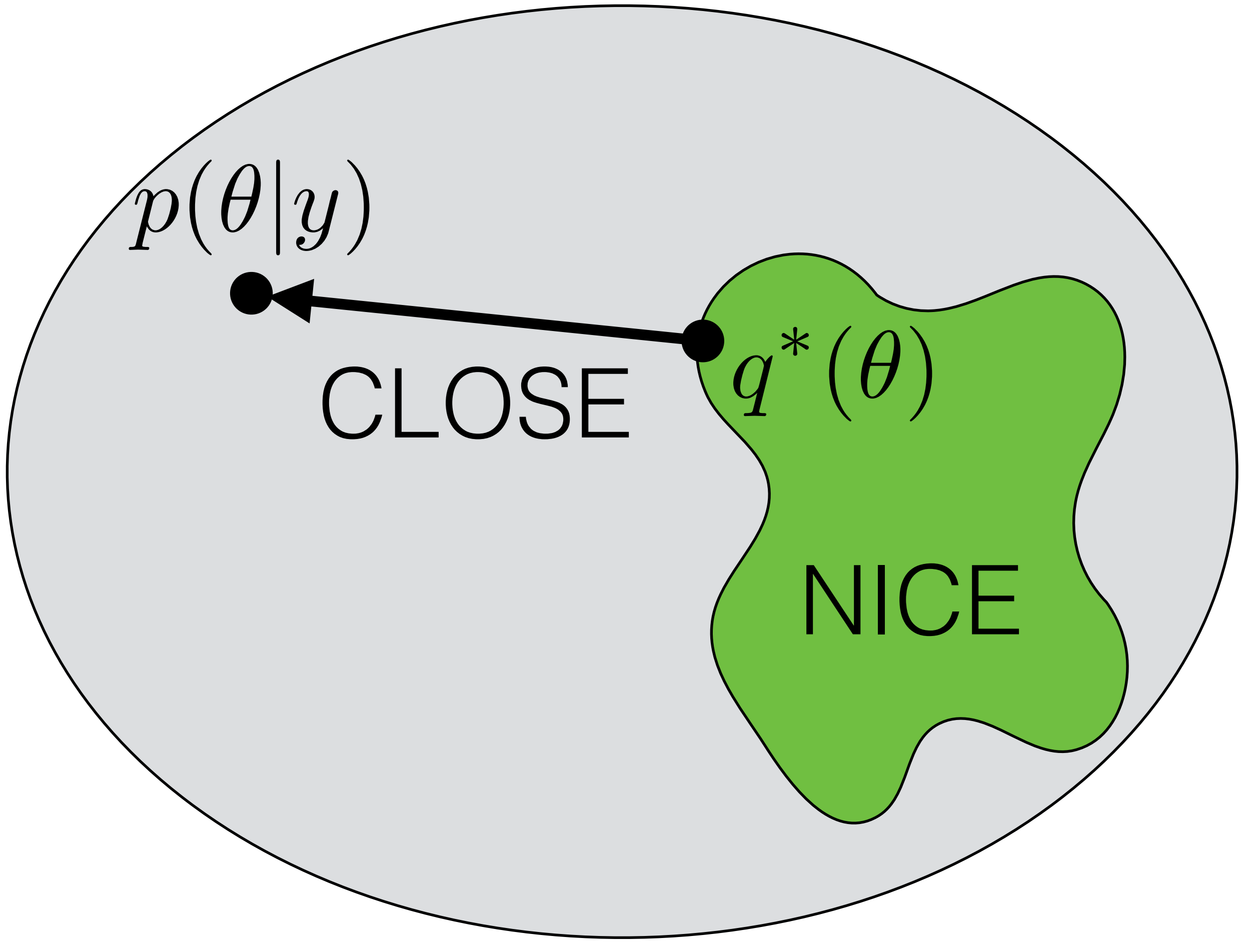
- Often also exponential family
- *Not* a modeling assumption



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

Variational Bayes

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

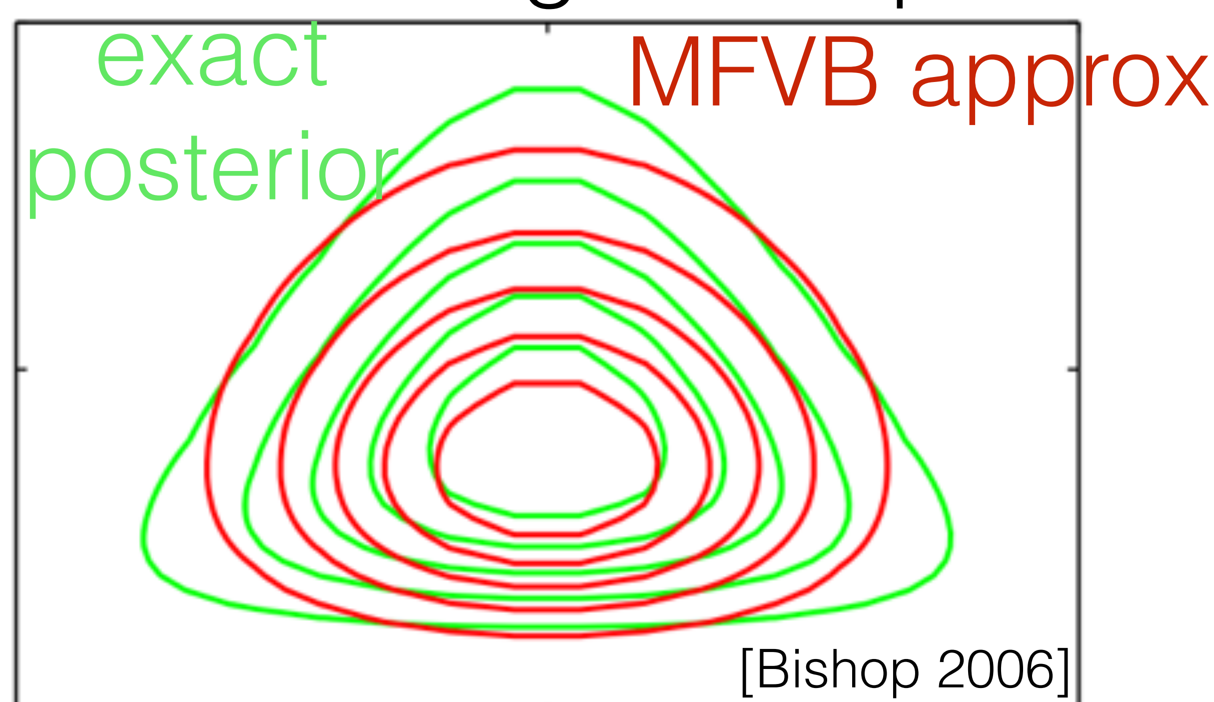


Choose "NICE" distributions

- 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



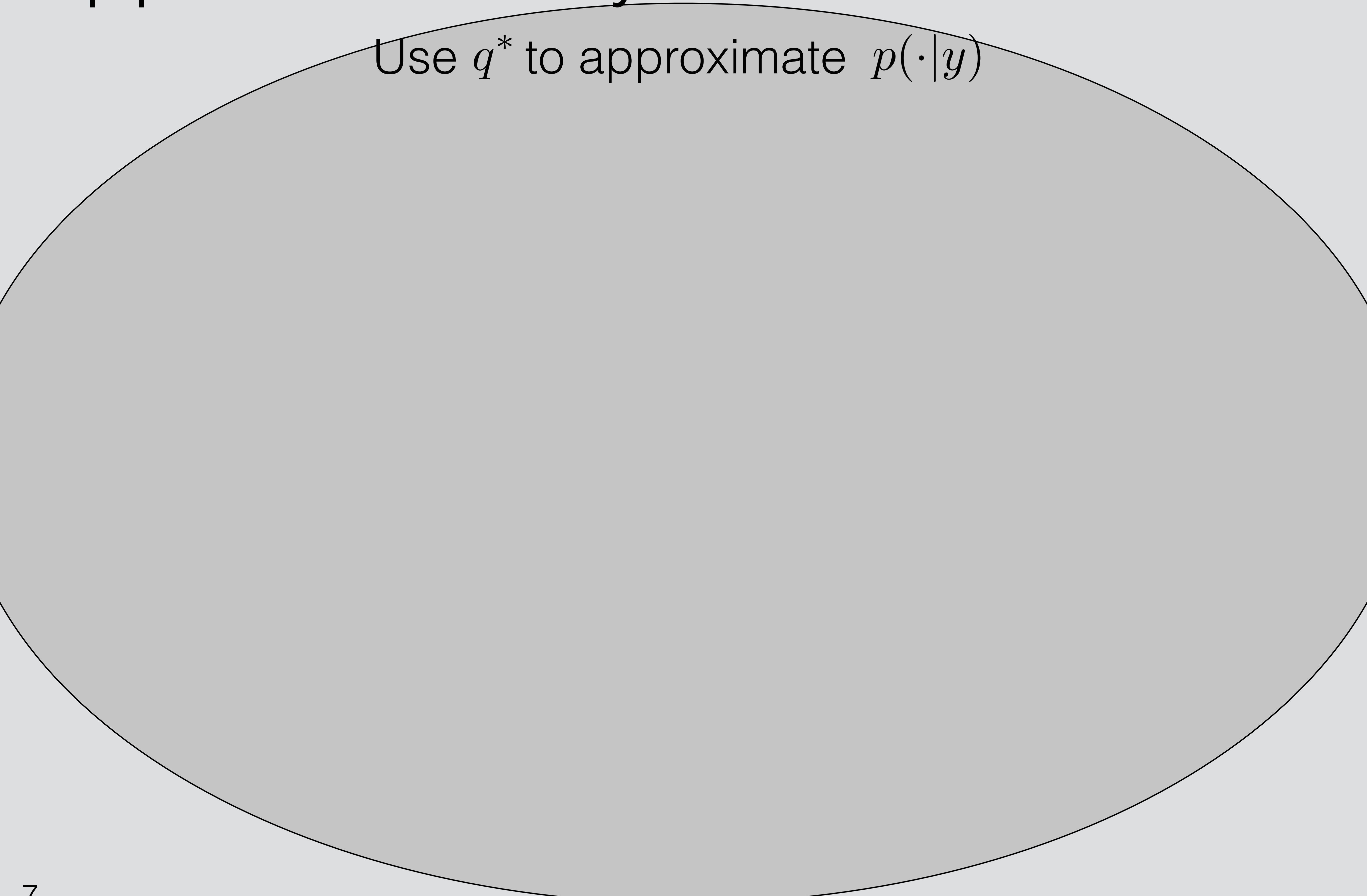
Now we have an optimization problem; how to solve it?

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

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

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

Variational Bayes

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

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?

References

- See the end of Part II for reference list