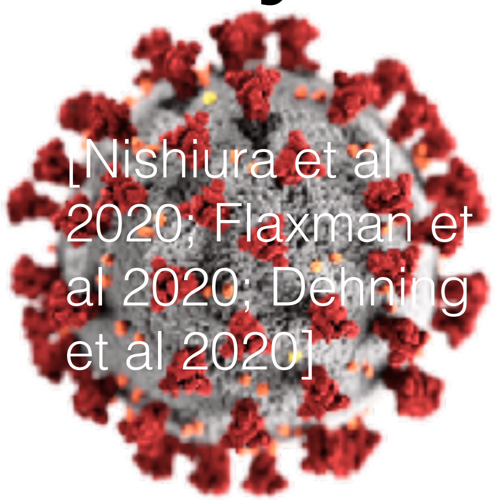
Variational Bayes and beyond: Foundations of scalable Bayesian inference (Part III)

Tamara Broderick
Associate Professor
MIT

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

Bayesian inference

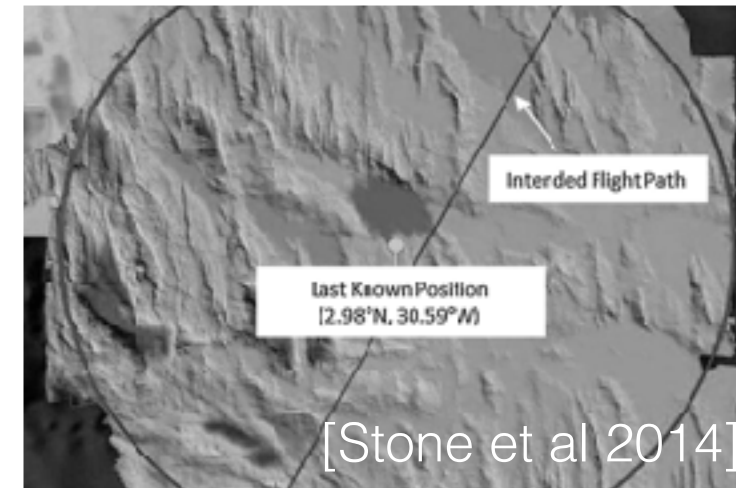


[Nishiura et al 2020; Flaxman et al 2020; Dehning et al 2020]

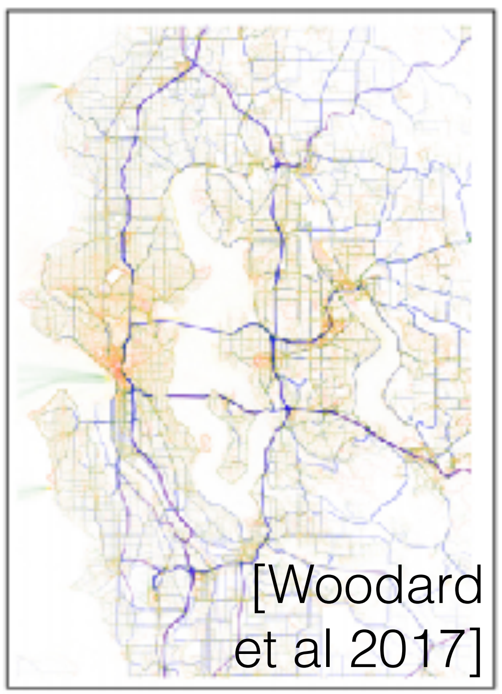
[Gillon et al 2017l Grimm et al 2018]

[Abbott et al 2016a,b]

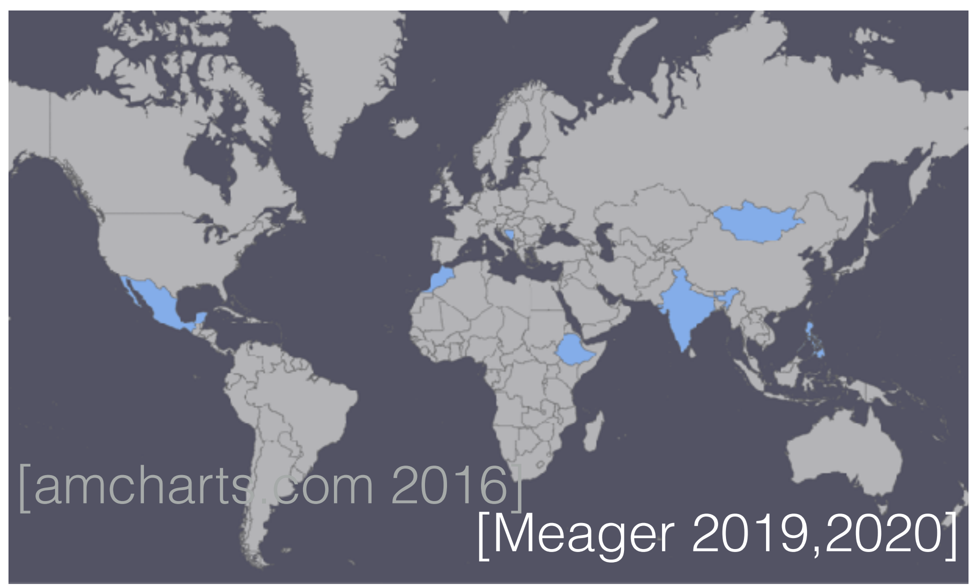
[ESO/ L. Calçada/ M. Kornmesser 2017]



[Stone et al 2014]



[Woodard et al 2017]



[amcharts.com 2016]

[Meager 2019,2020]

[Kuikka et al 2014]

[Baltic Salmon Fund]

[Chati, Balakrishnan 2017]

[Julian Hertzog 2016]

[mc-stan.org]

- Goals: good point estimates, uncertainty estimates
- Challenge: speed (compute, user), reliable inference

Criteo Online Ads Experiment

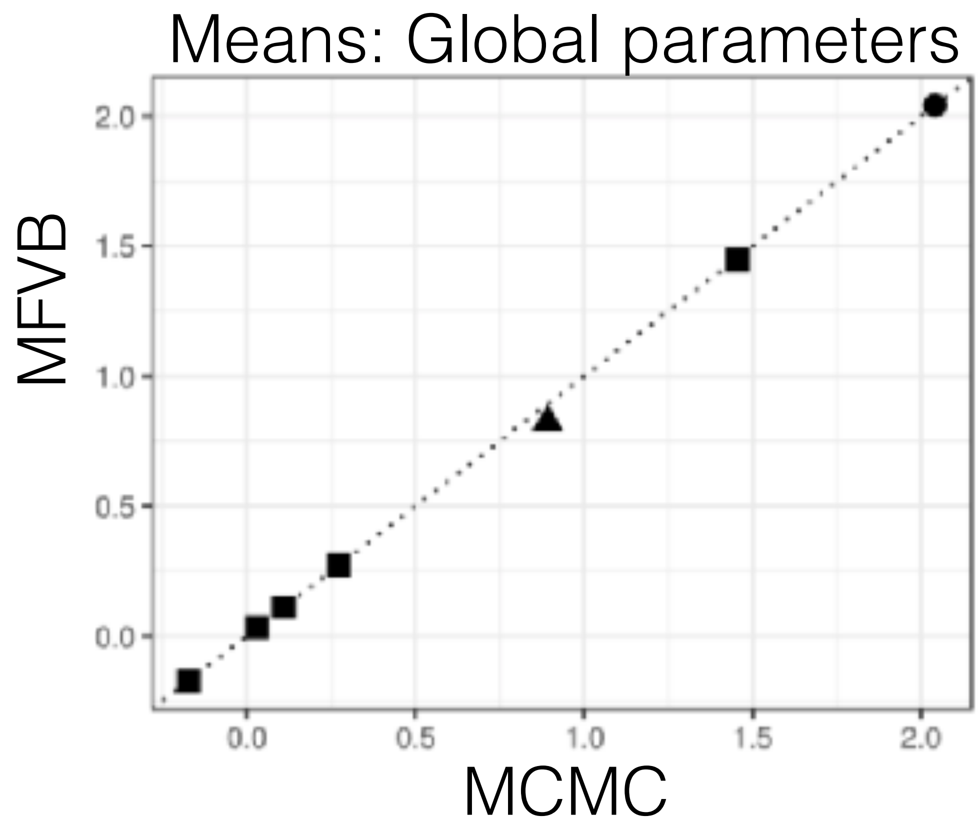
- MAP: **12 s**
- MFVB: **57 s**
- MCMC (5K samples): 21,066 s (**5.85 h**)

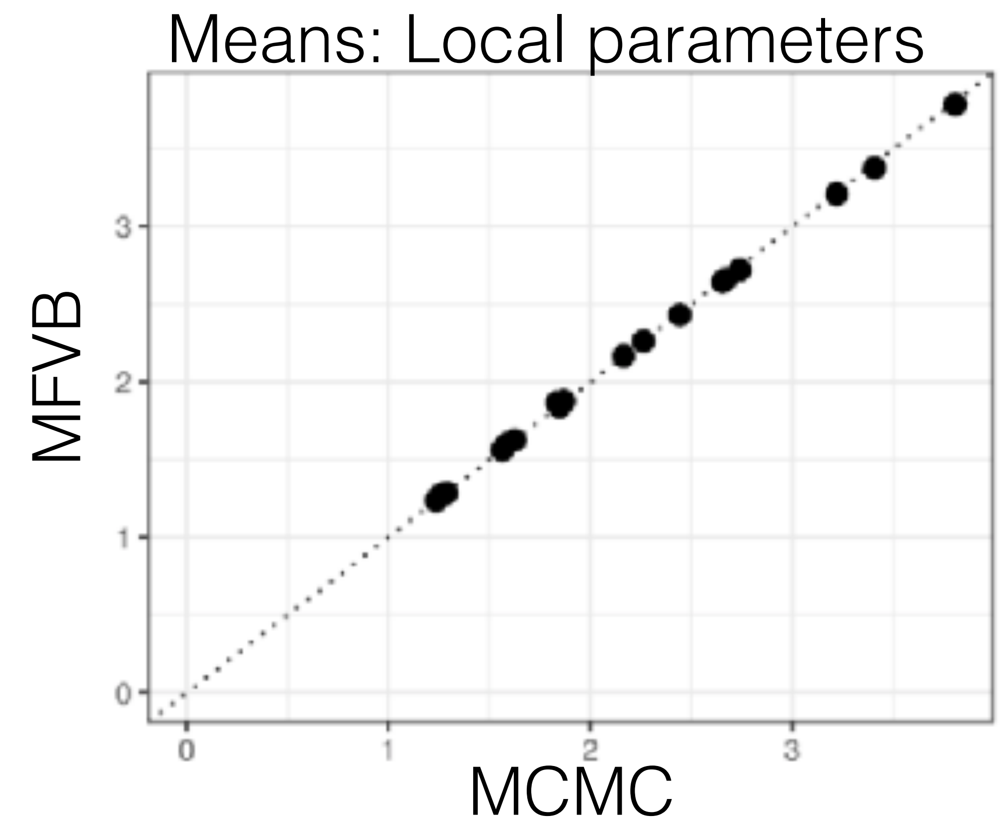
[Giordano, Broderick, Jordan 2018]

Criteo Online Ads Experiment

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

[Giordano, Broderick, Jordan 2018]

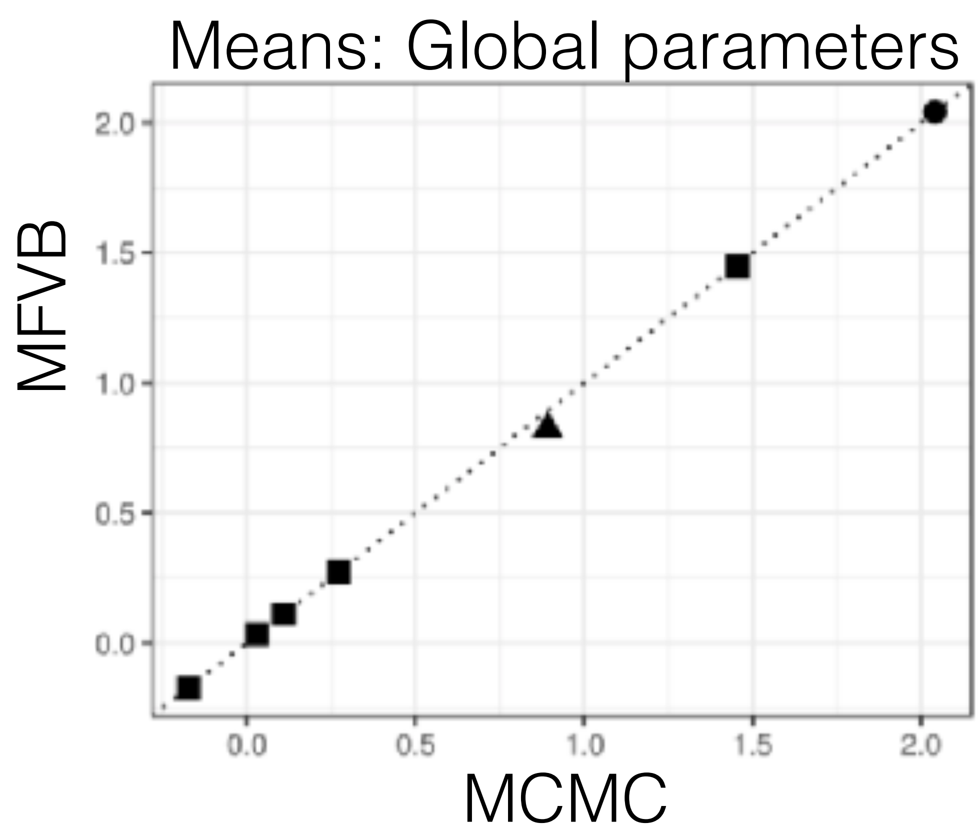


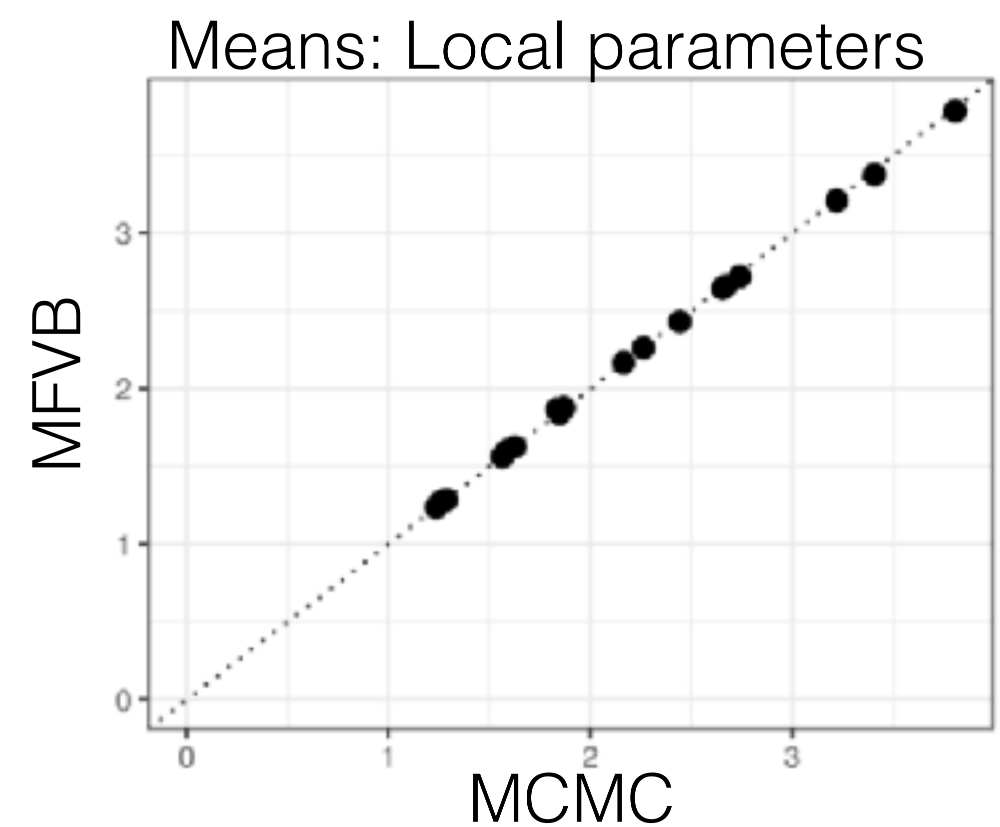


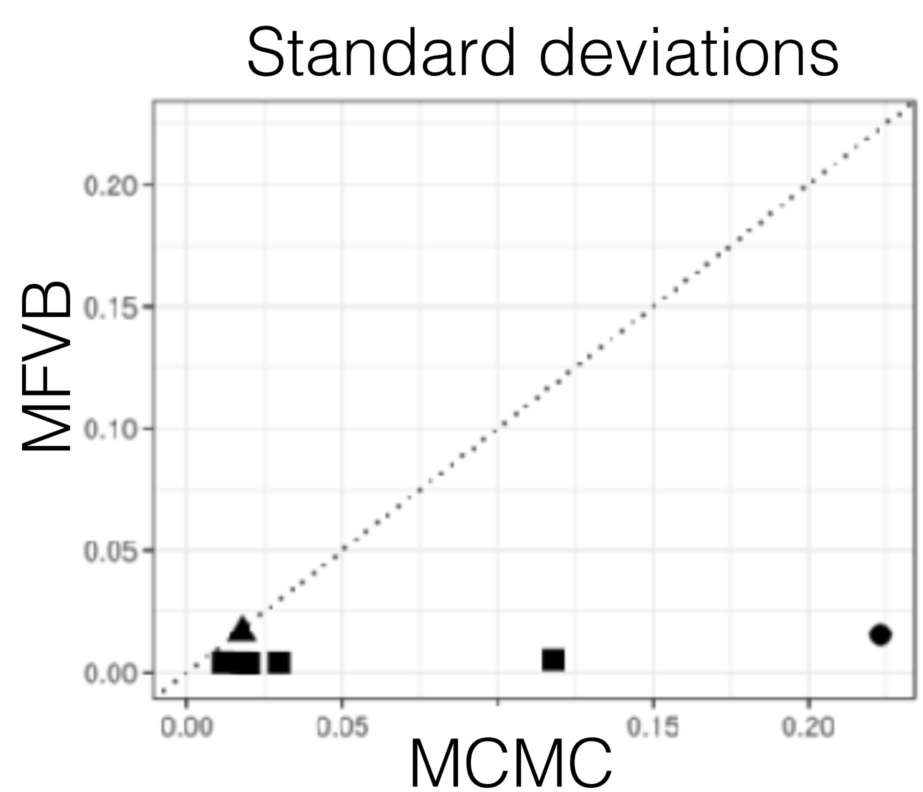
Criteo Online Ads Experiment

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

[Giordano, Broderick, Jordan 2018]



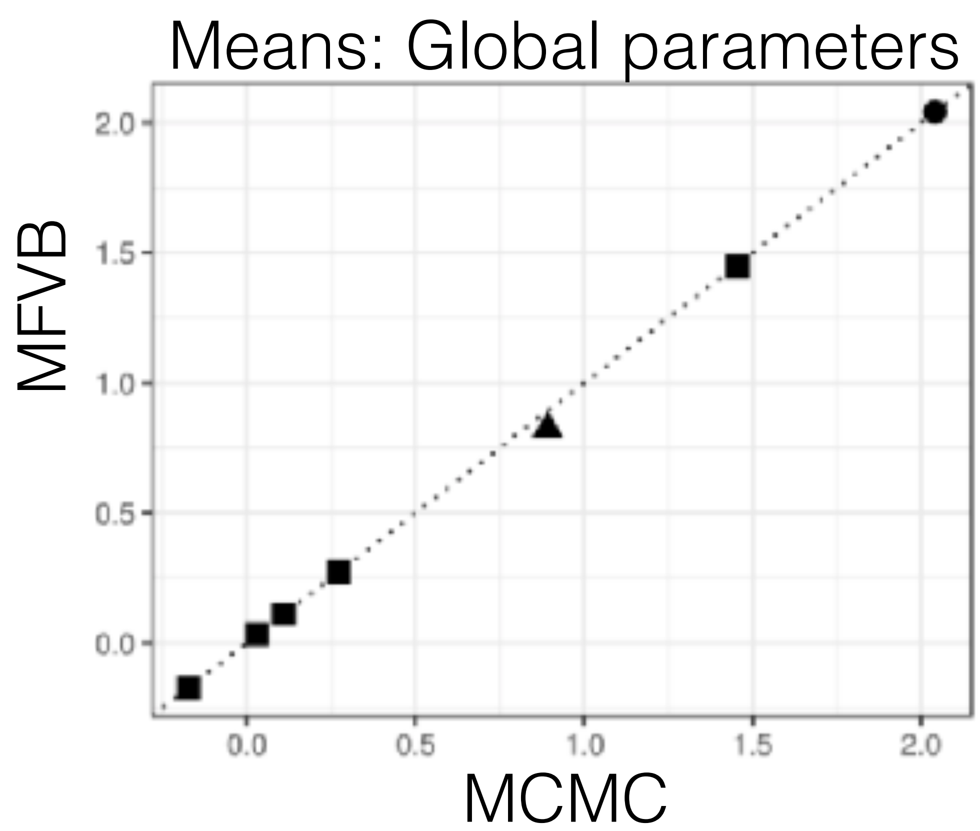


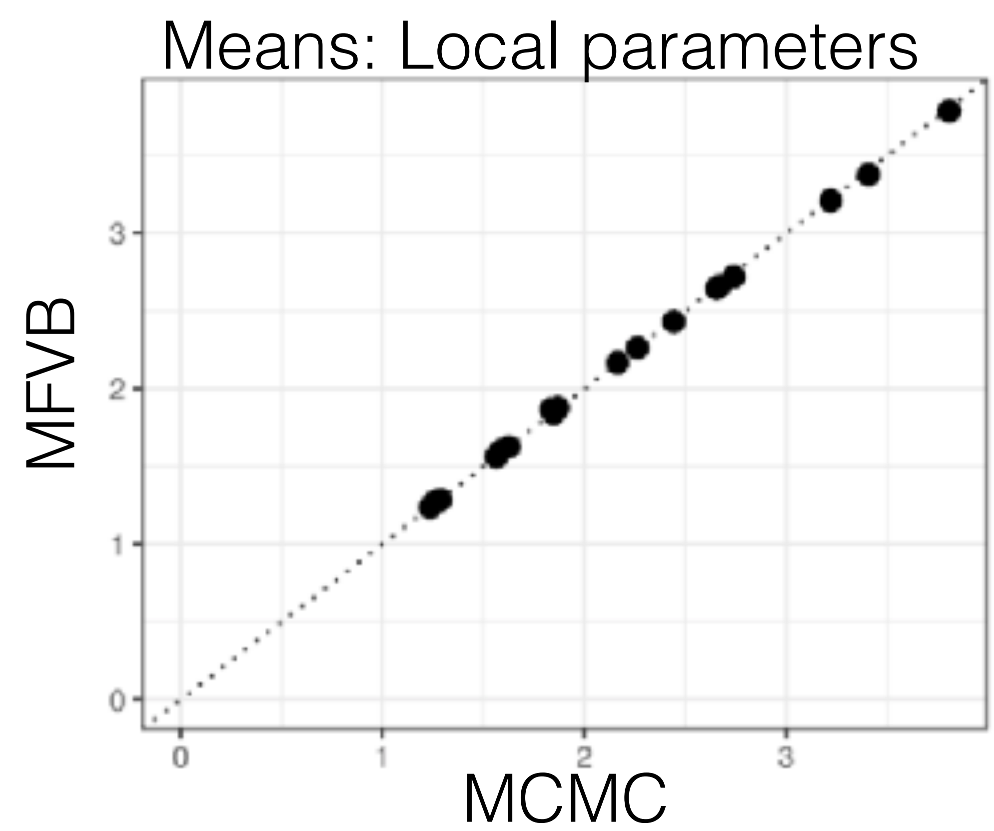


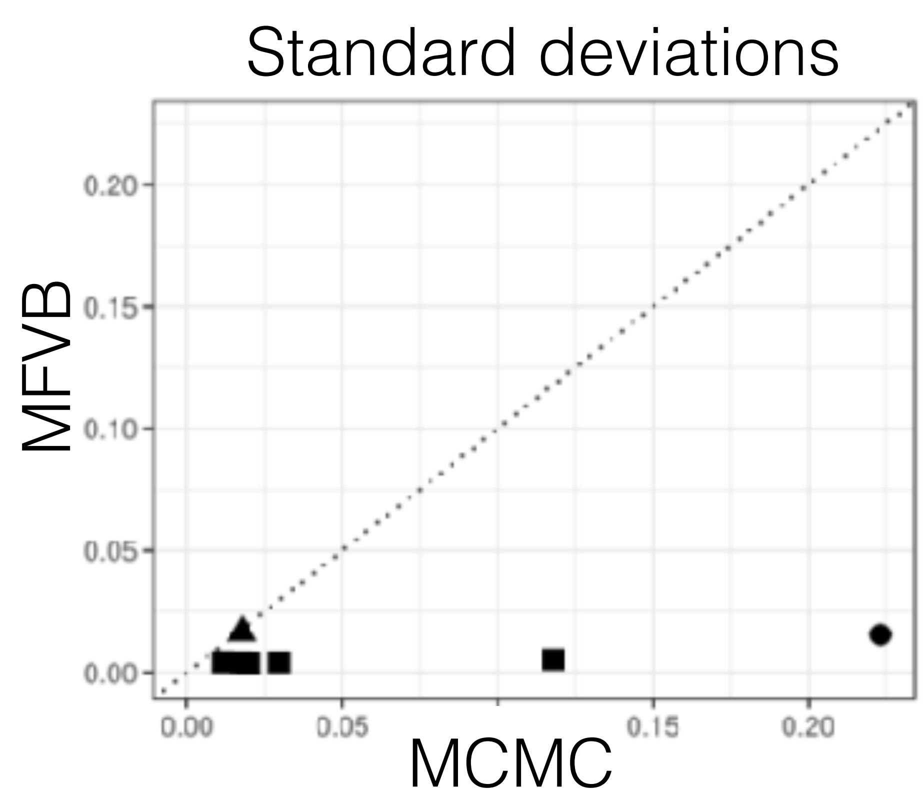
Criteo Online Ads Experiment

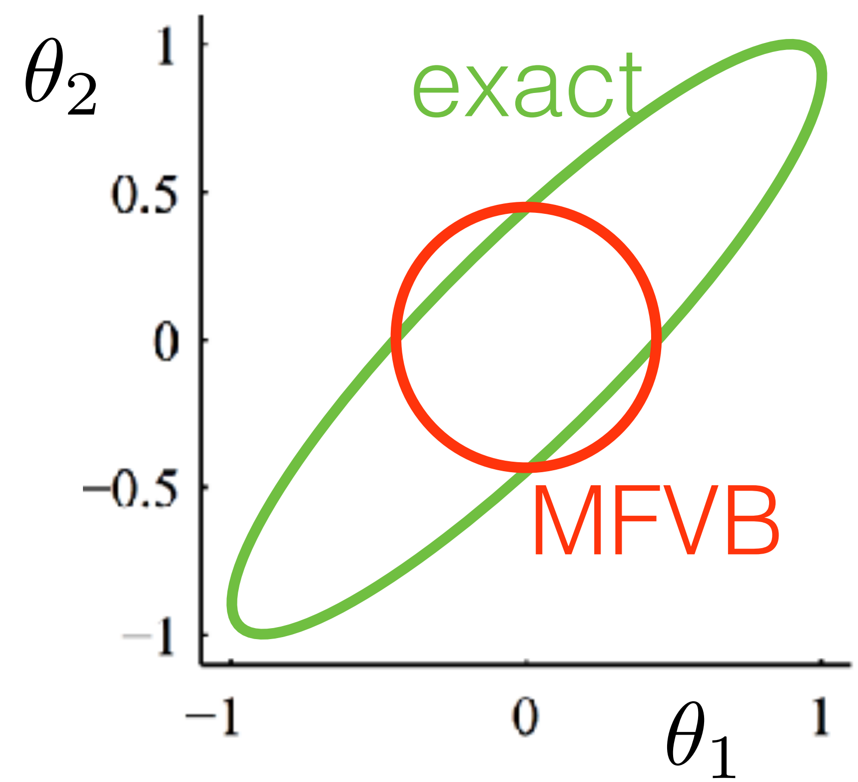
- MAP: **12 s**
- MFVB: **57 s**
- MCMC (5K samples): 21,066 s (**5.85 h**)

[Giordano, Broderick, Jordan 2018]









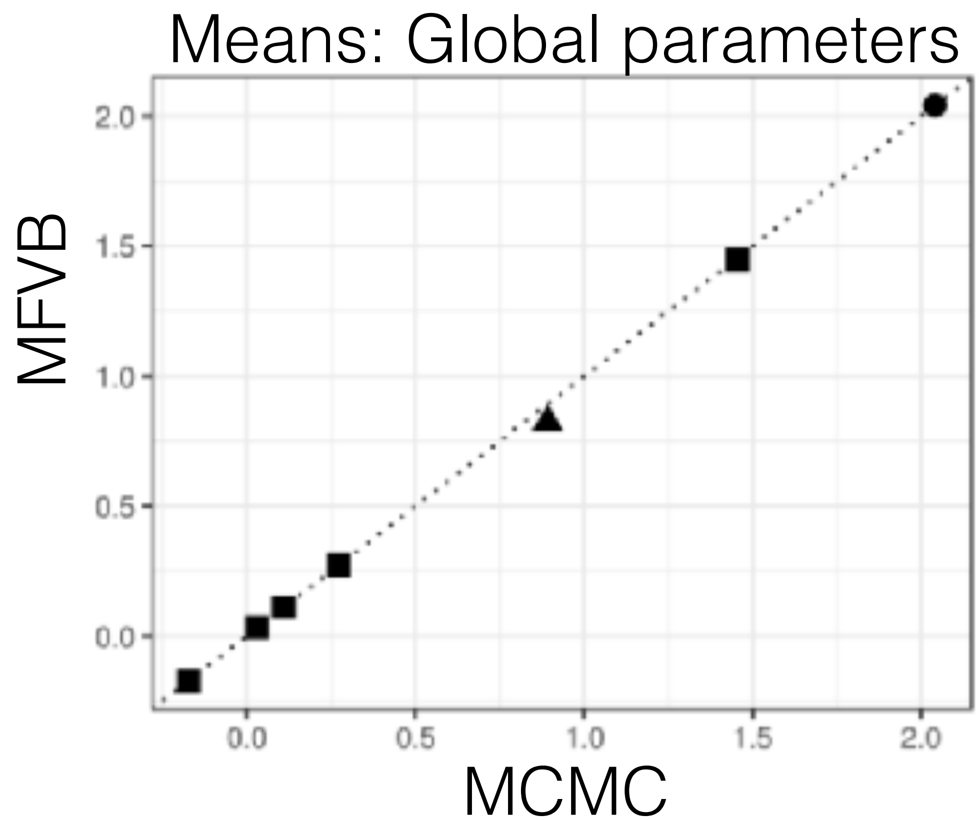
[Turner & Sahani 2011; MacKay 2003; Bishop 2006; Wang, Titterington 2004]

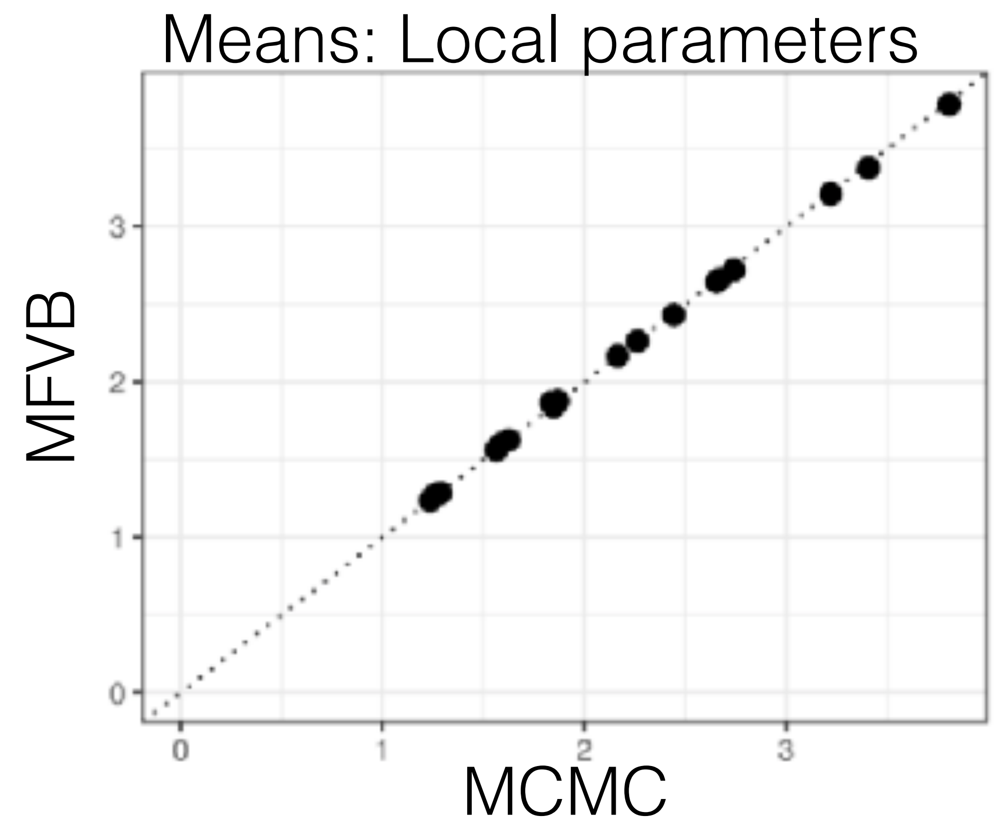
Criteo Online Ads Experiment

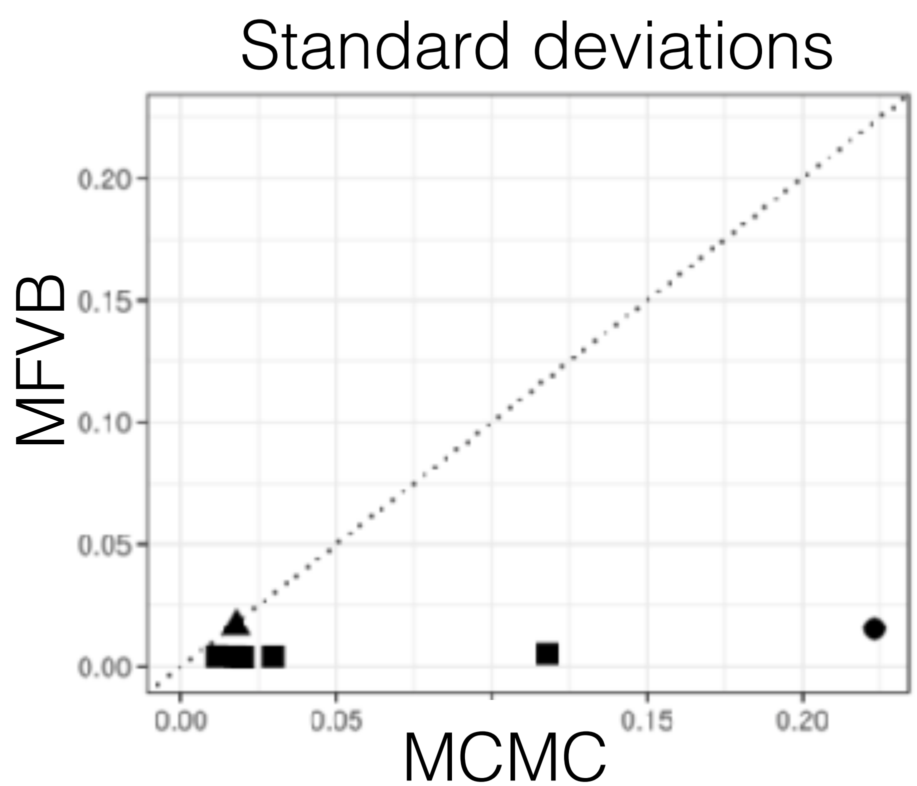
- MAP: **12 s**
- MFVB: **57 s**
- MCMC (5K samples): 21,066 s (**5.85 h**)

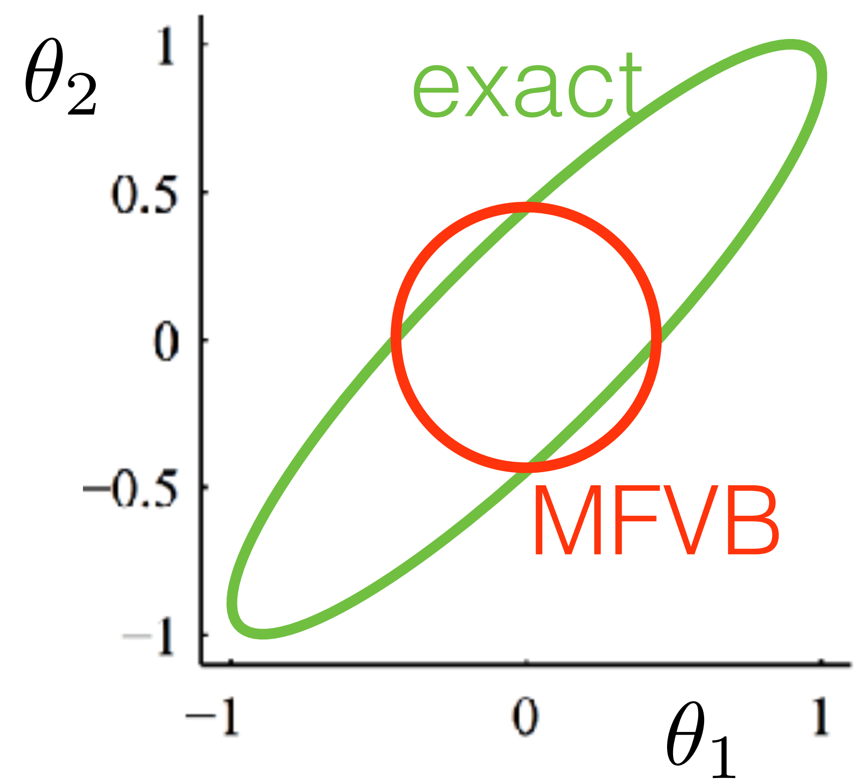
[Giordano, Broderick, Jordan 2018]

- Also: microcredit, graph/ network data, etc







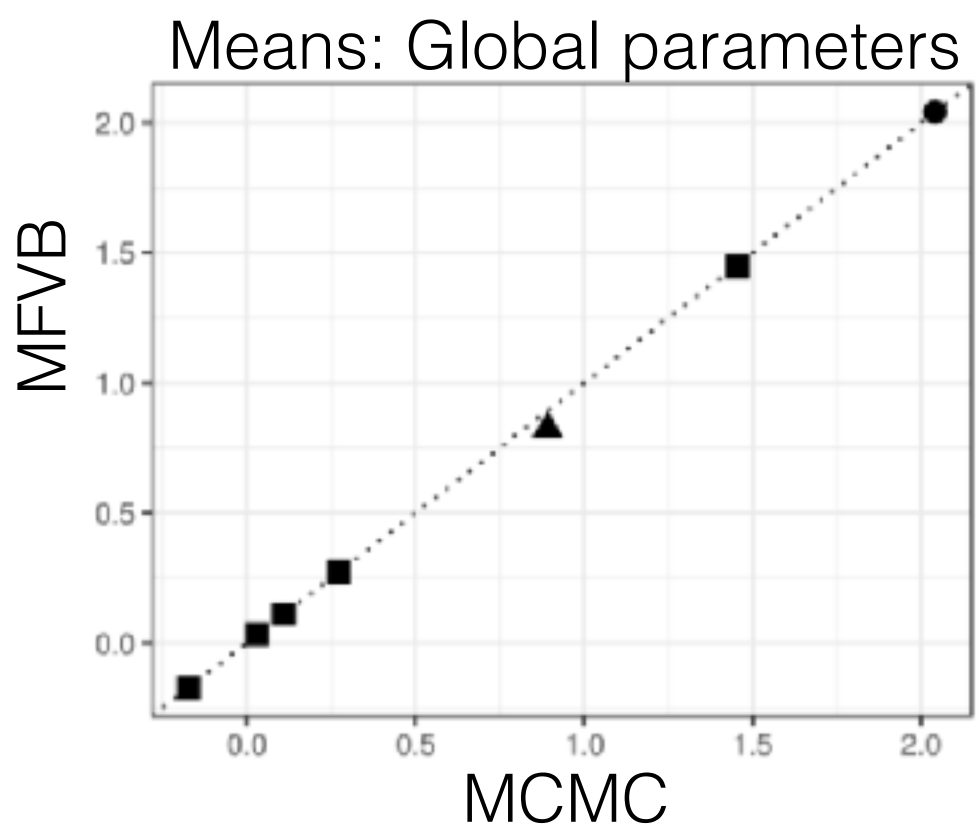


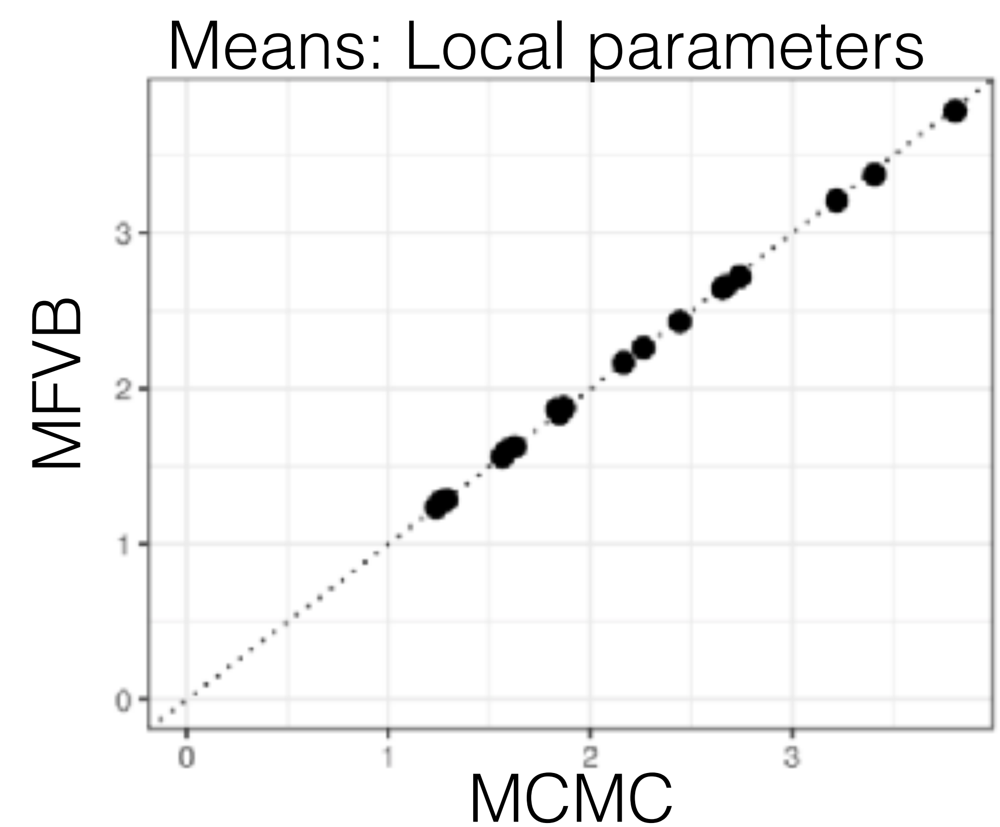
[Turner & Sahani 2011; MacKay 2003; Bishop 2006; Wang, Titterington 2004]

Criteo Online Ads Experiment

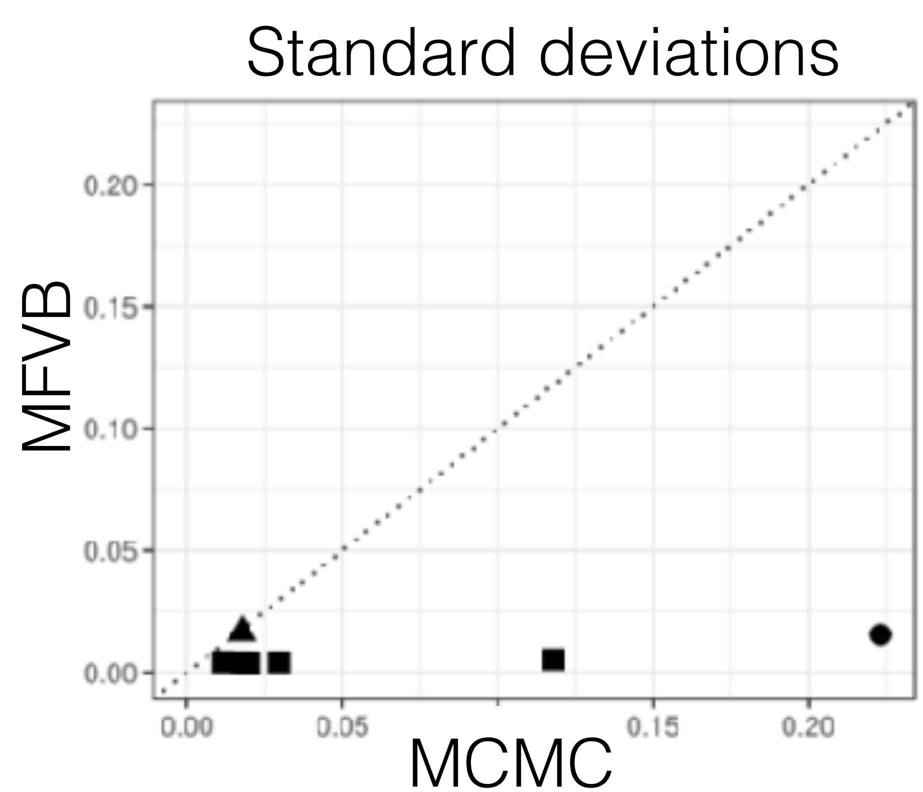
- MAP: **12 s**
- MFVB: **57 s**
- MCMC (5K samples): 21,066 s (**5.85 h**)

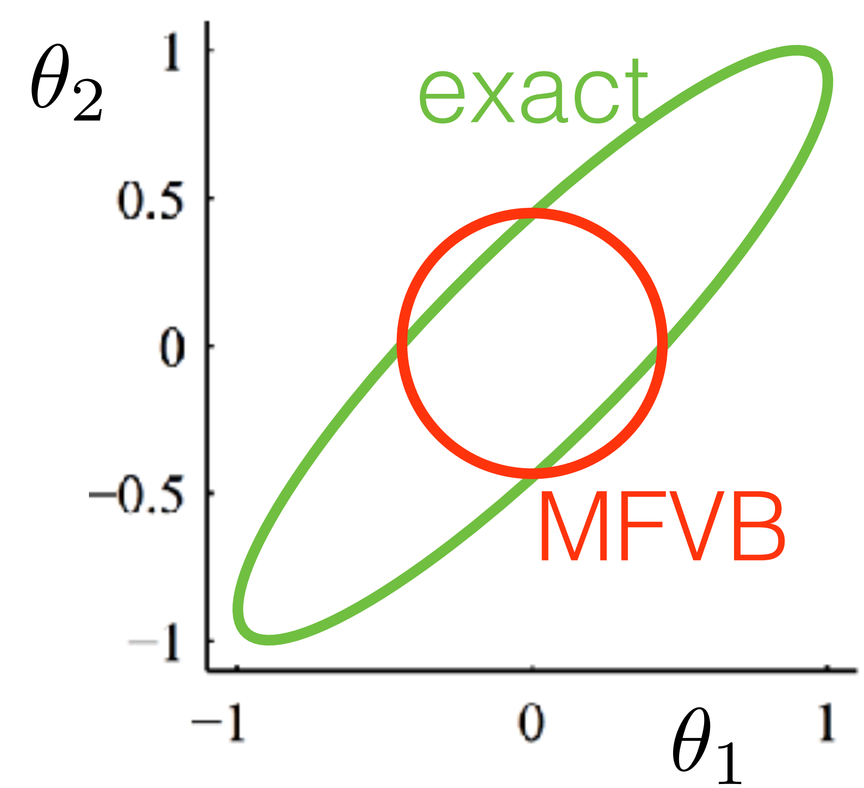
[Giordano, Broderick, Jordan 2018]





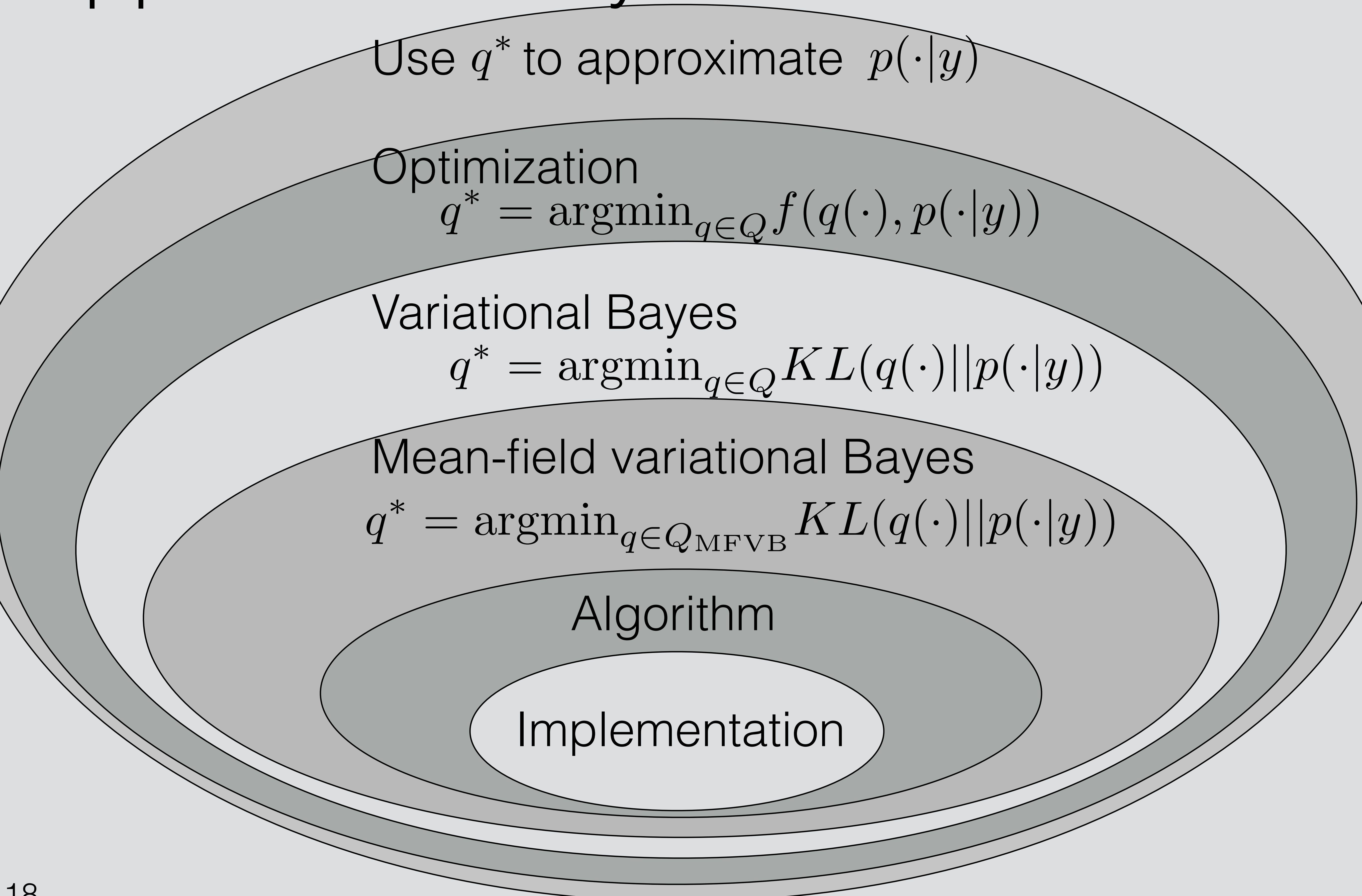
- Also: microcredit, graph/ network data, etc
- Posterior means





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

Approximate Bayesian inference



Approximate Bayesian inference

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

Optimization

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

Variational Bayes

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

Mean-field variational Bayes

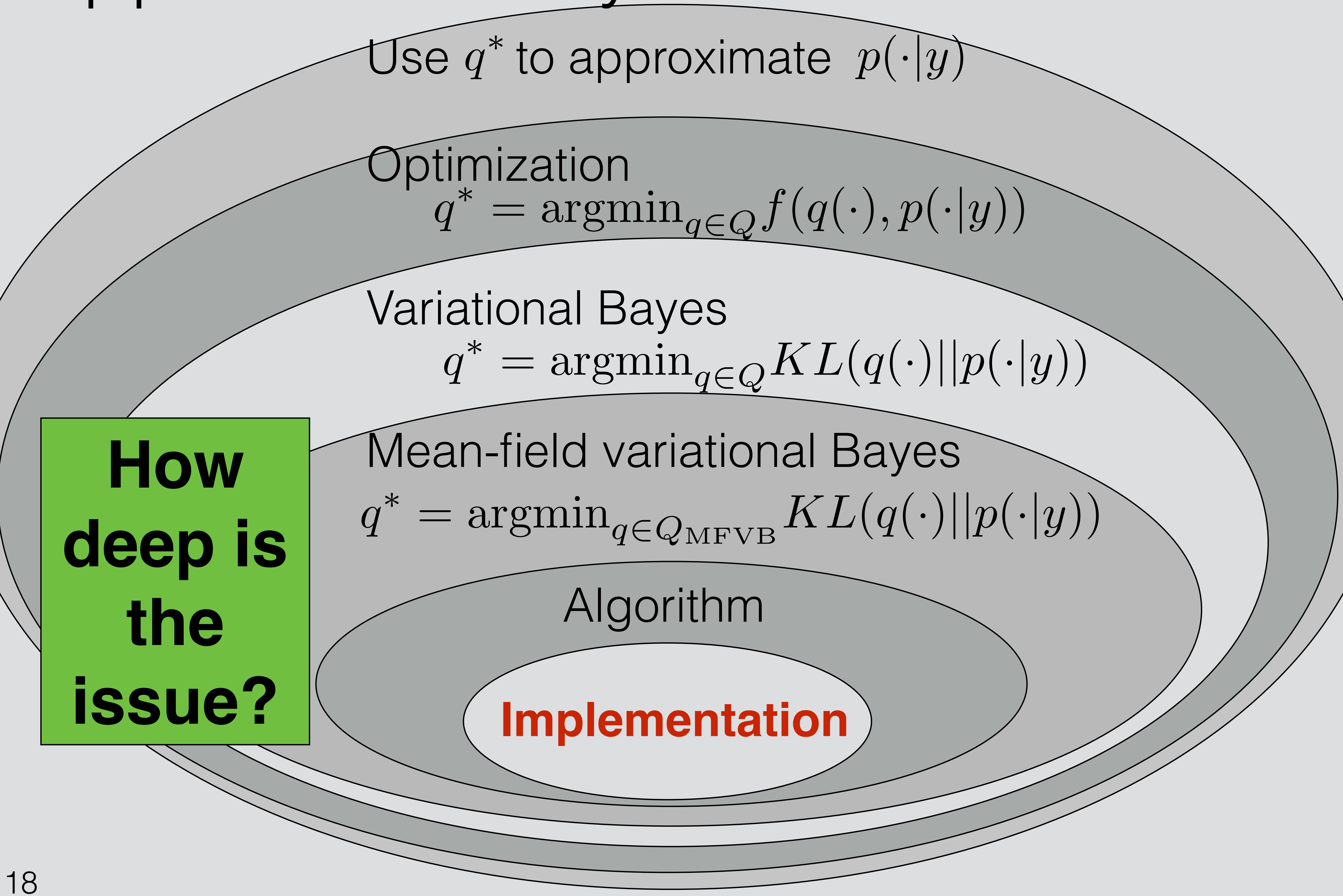
$$q^* = \mathrm{argmin}_{q \in Q_{\mathrm{MFVB}}} KL(q(\cdot)||p(\cdot|y))$$

Algorithm

Implementation

How deep is the issue?

Approximate Bayesian inference



Approximate Bayesian inference

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

Optimization

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

Variational Bayes

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

Mean-field variational Bayes

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

Algorithm

Implementation

How deep is the issue?

Gaussian example was exact

Approximate Bayesian inference

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

Optimization

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

Variational Bayes

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

Mean-field variational Bayes

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

Algorithm

Implementation

How deep is the issue?

Gaussian example was exact

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

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

Algorithm

Implementation

How deep is the issue?

Gaussian example was exact

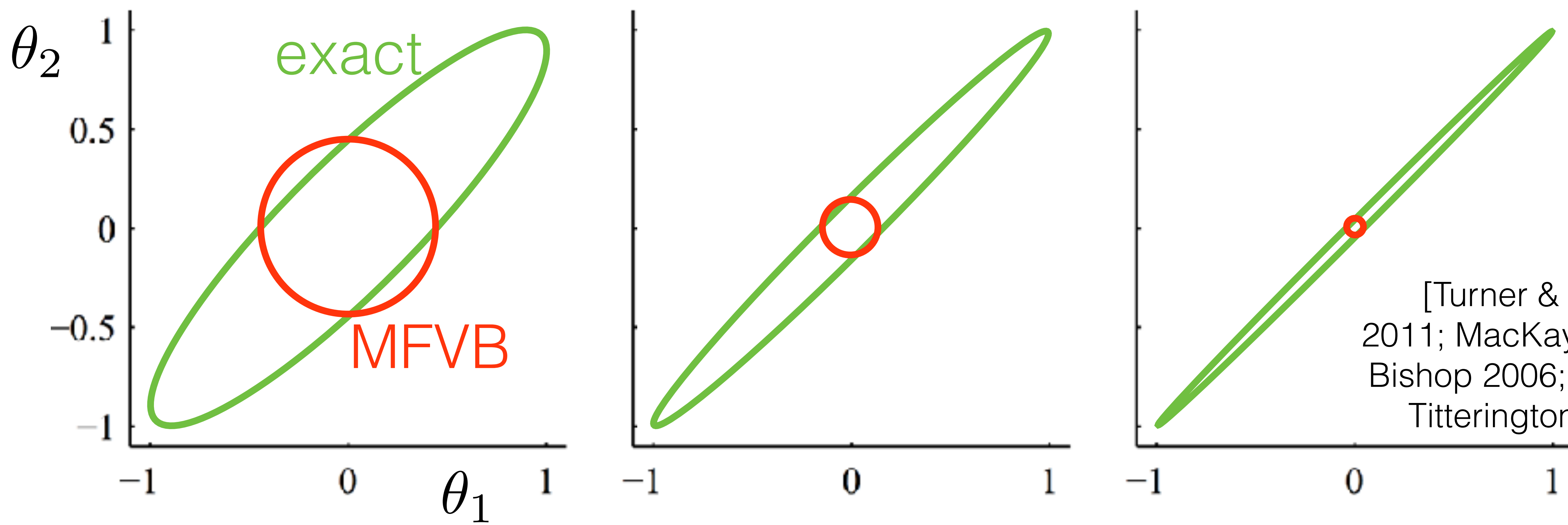
Is it just MFVB?

Is it just **MF**VB?

Is it just MFVB?

Is it just MFVB?

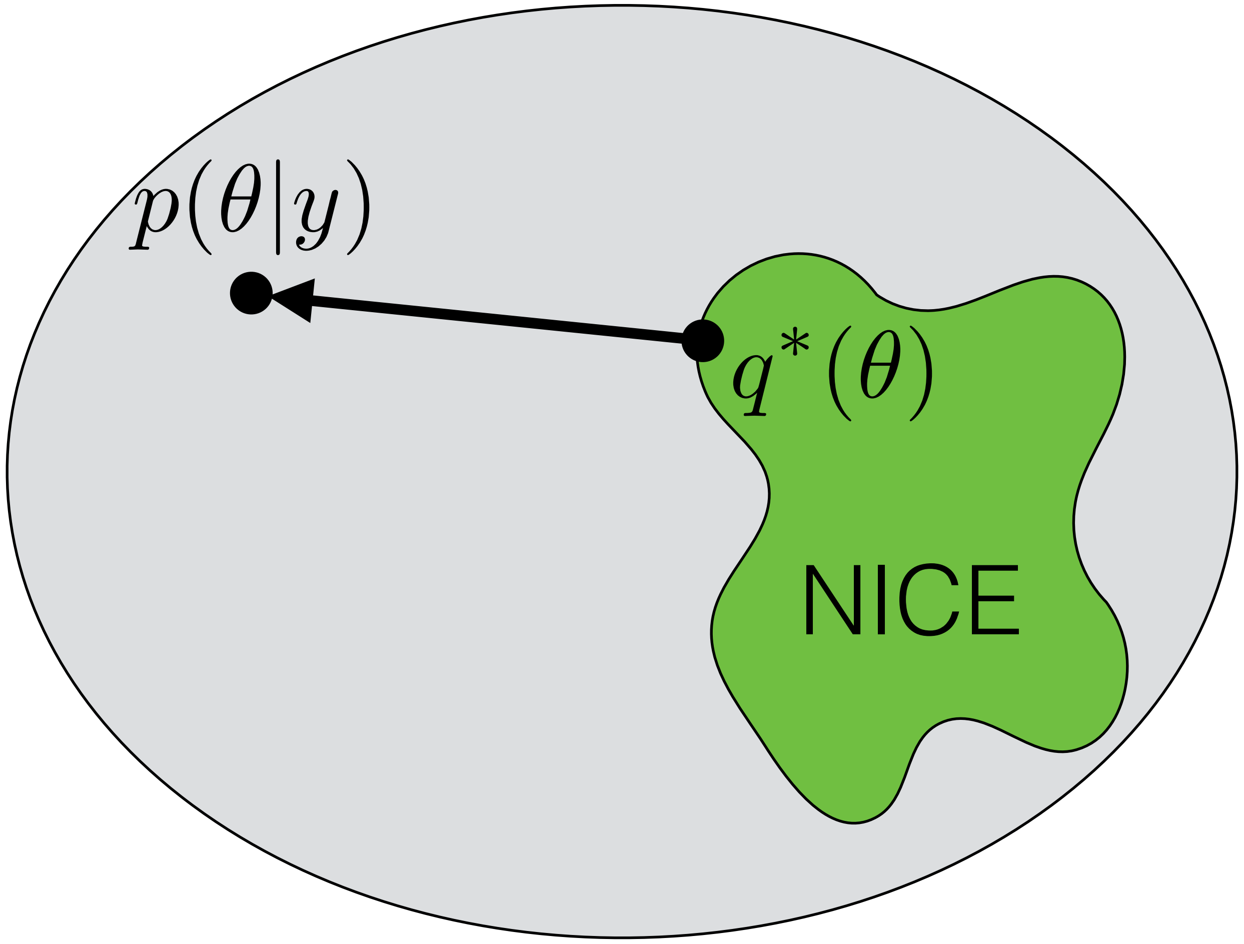
$$KL(q||p(\cdot|y)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta \qquad q(\theta) = \prod_{j=1}^{J} q_j(\theta_j)$$



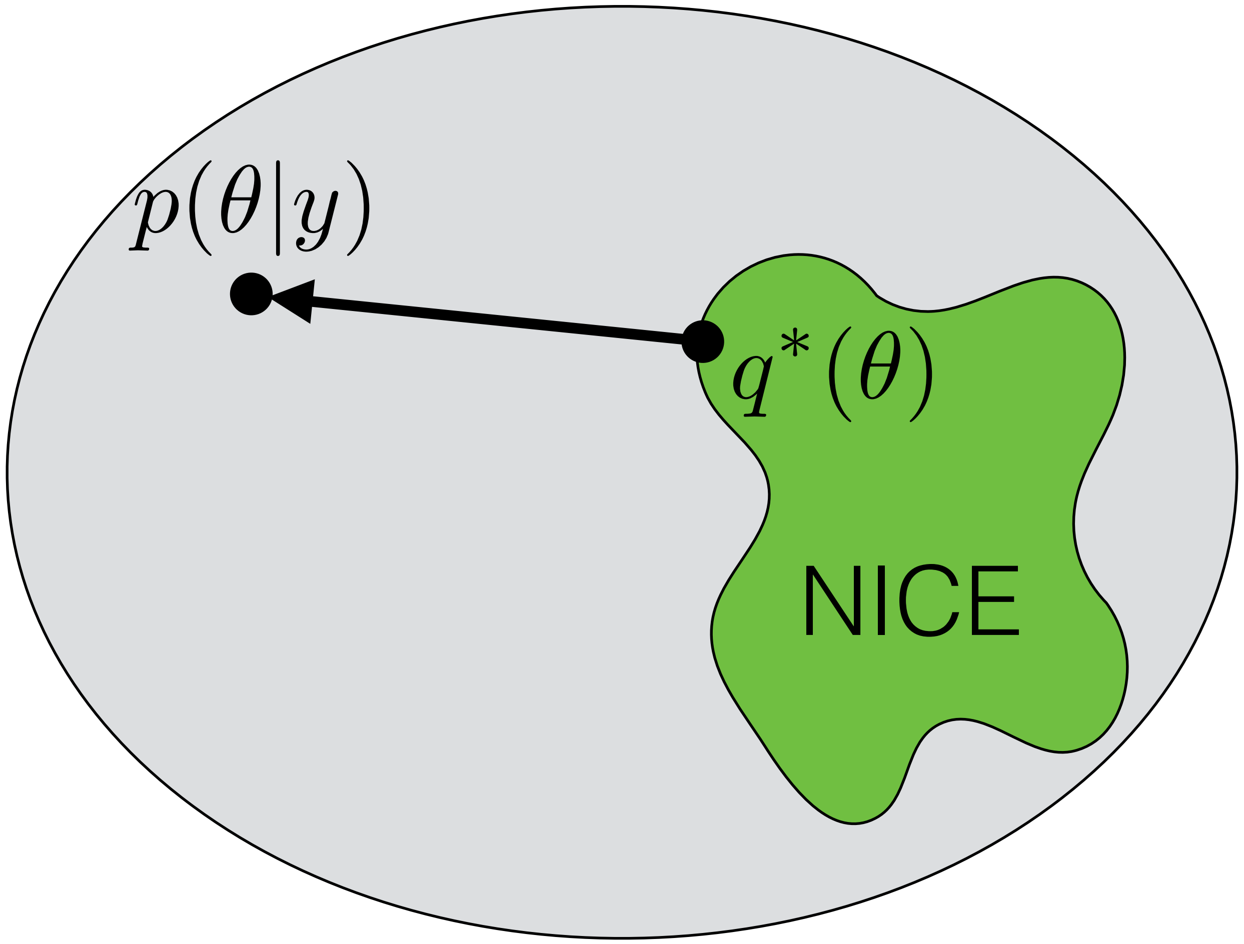
[Turner & Sahani 2011; MacKay 2003; Bishop 2006; Wang, Titterington 2004]

Is it just MFVB?

Is it just MFVB?

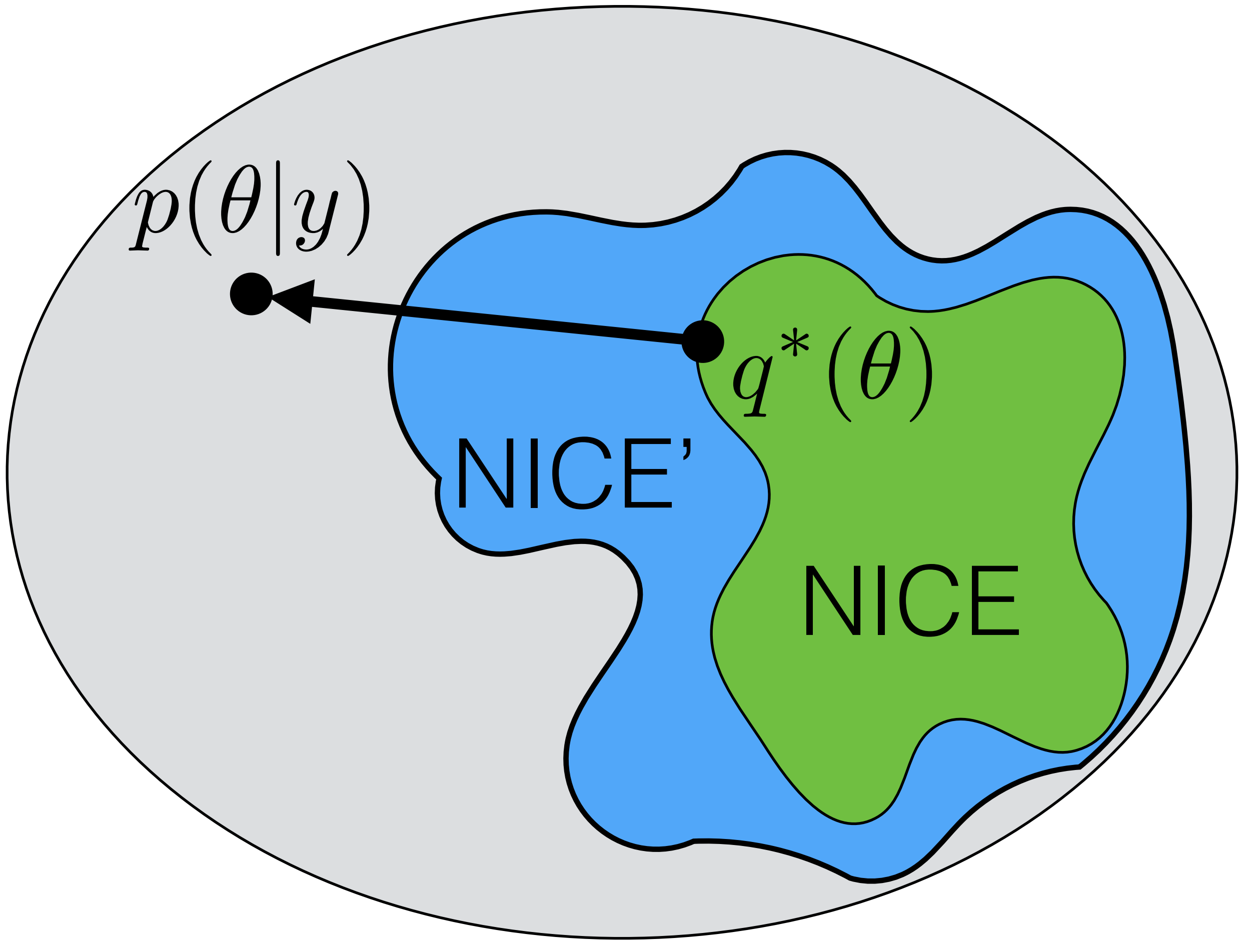


Is it just MFVB?



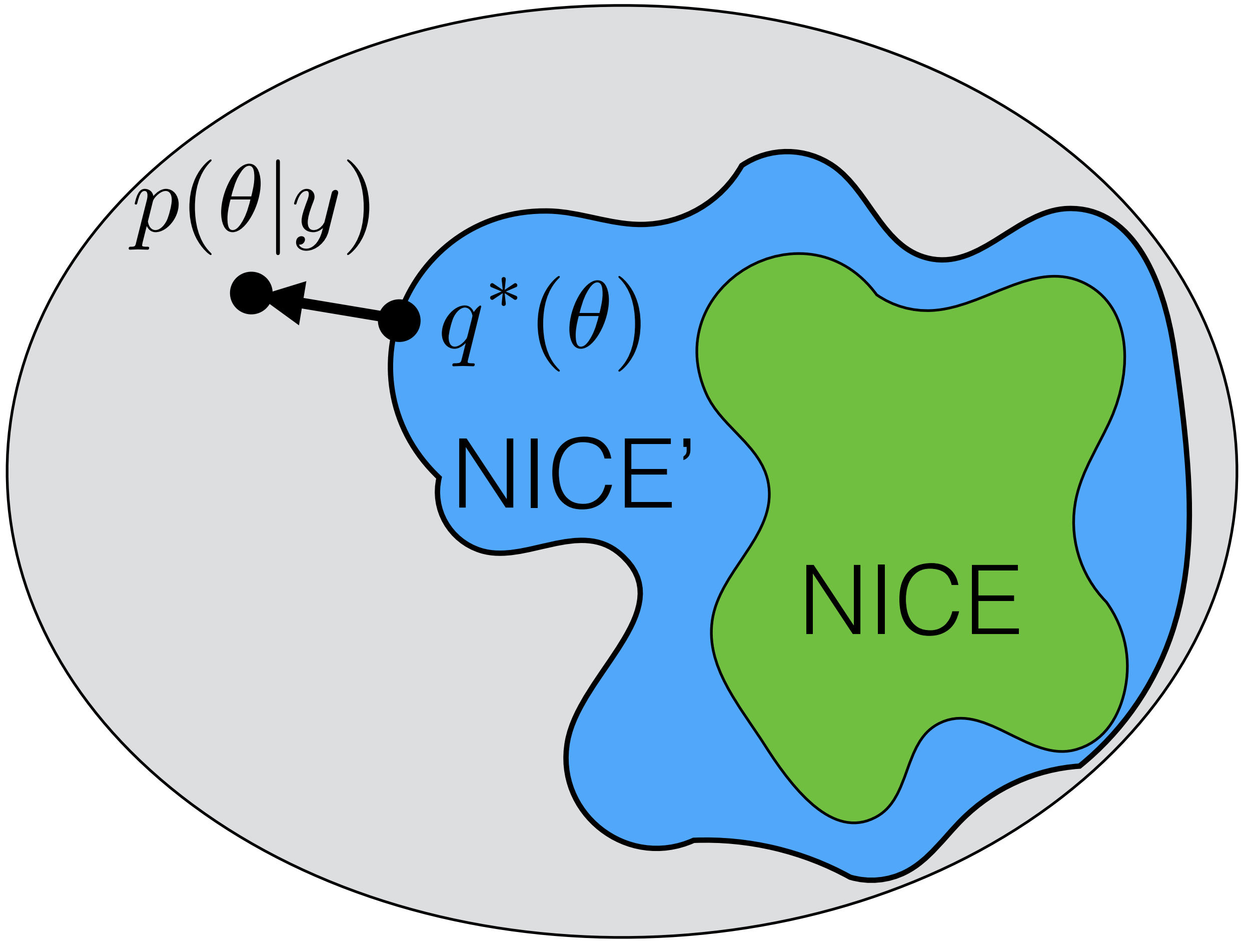
- Turner, Sahani (2011) showed (empirically) can have strictly larger NICE set but worse mean & variance estimates

Is it just MFVB?



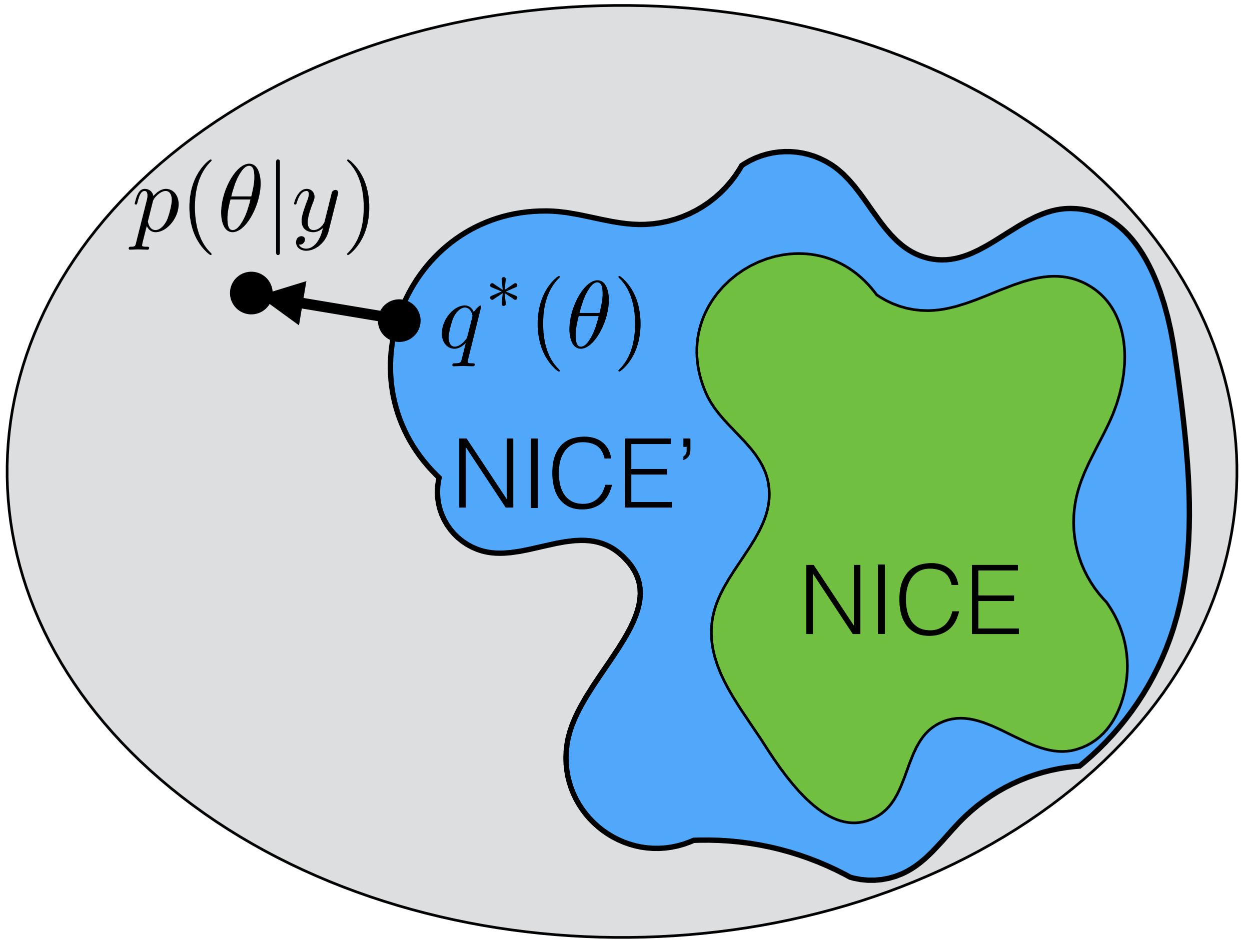
- Turner, Sahani (2011) showed (empirically) can have strictly larger NICE set but worse mean & variance estimates

Is it just MFVB?



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

Is it just MFVB?

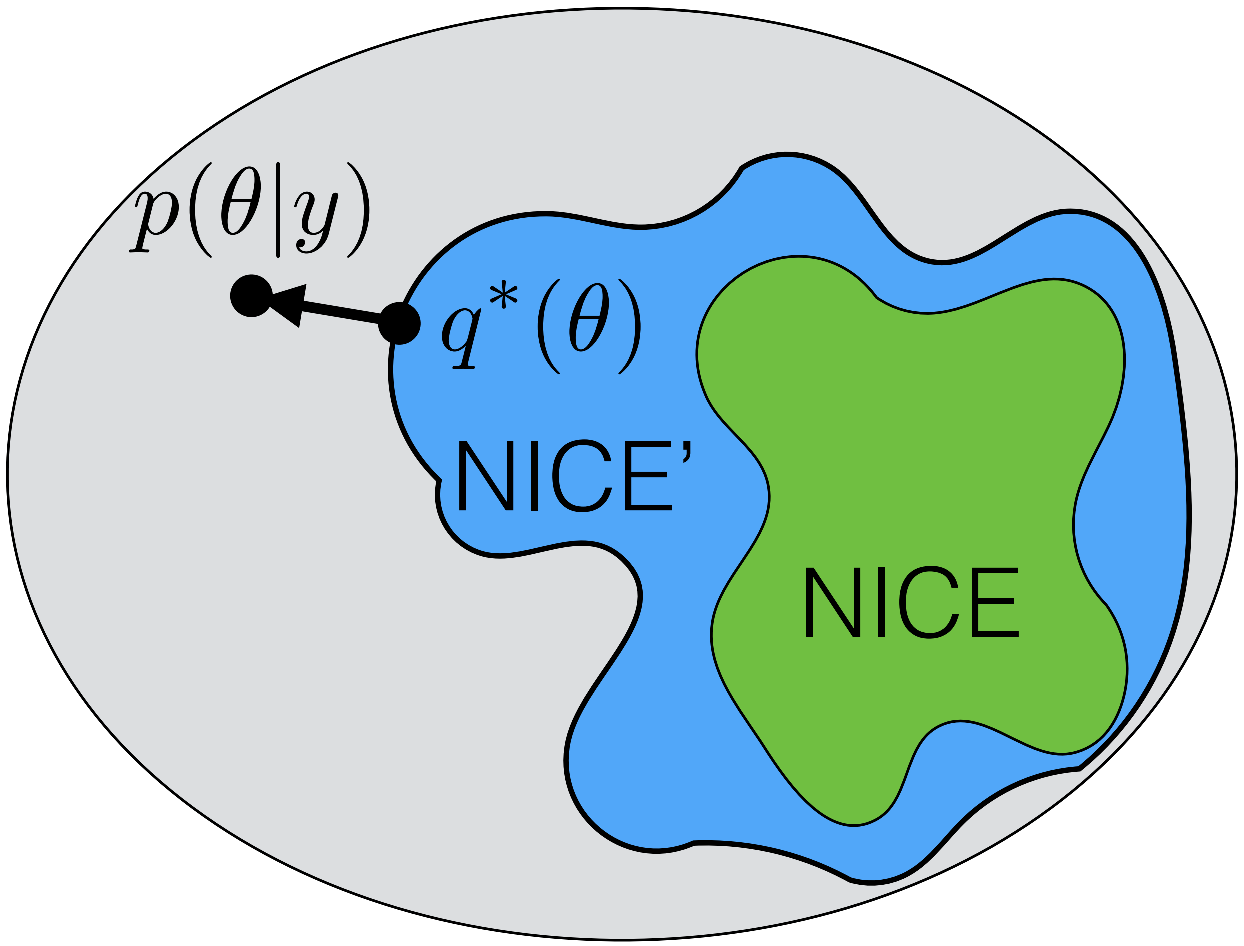


- Turner, Sahani (2011) showed (empirically) can have strictly larger NICE set but worse mean & variance estimates

- Exercise: Show, with a simple example, that a smaller KL does not imply better mean and variance estimates

[Baqué et al 2017; Huggins, Karsprzak, Campbell, Broderick 2020]

Is it just MFVB?



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

- Exercise: Show, with a simple example, that a smaller KL does not imply better mean and variance estimates
- But how much worse can the estimates be? And could it have just been the implementation?

[Baqué et al 2017; Huggins, Karsprzak, Campbell, Broderick 2020]

Is it just MFVB?

[Huggins, Karsprzak, Campbell, Broderick 2020]

Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3 [Baqué et al 2017; Huggins et al 2020]

[Huggins, Karsprzak, Campbell, Broderick 2020]

Is it just MFVB?

- Some KL values seen in practice:
  ~1 to ~70, 0.5 to 3 [Baqué et al 2017; Huggins et al 2020]
- Take any constant *c*

[Huggins, Karsprzak, Campbell, Broderick 2020]

Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3 [Baqué et al 2017; Huggins et al 2020]
- Take any constant *c*

Proposition. Can have small KL (<0.23) & arbitrarily bad variance estimate

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

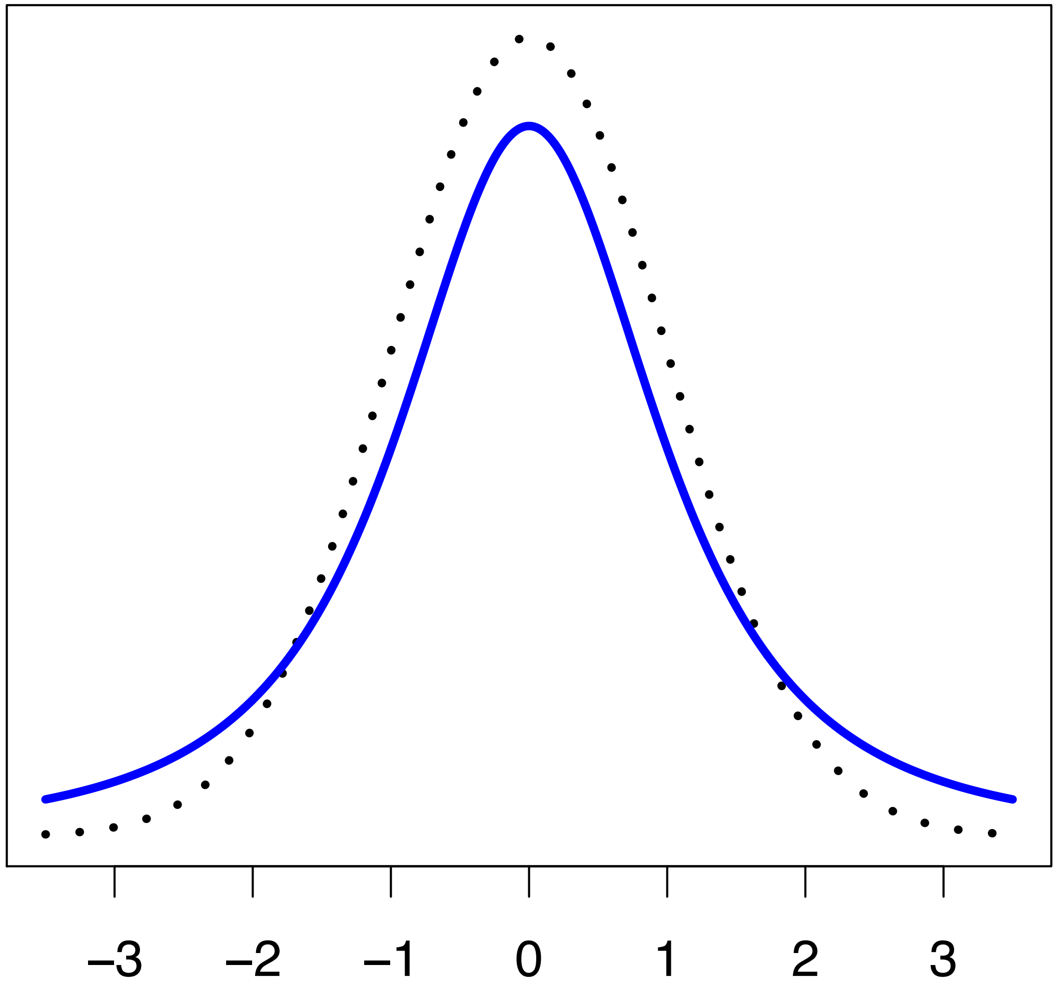
[Huggins, Karsprzak, Campbell, Broderick 2020]

Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3 [Baqué et al 2017; Huggins et al 2020]
- Take any constant *c*

Proposition. Can have small KL (<0.23) & arbitrarily bad variance estimate

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



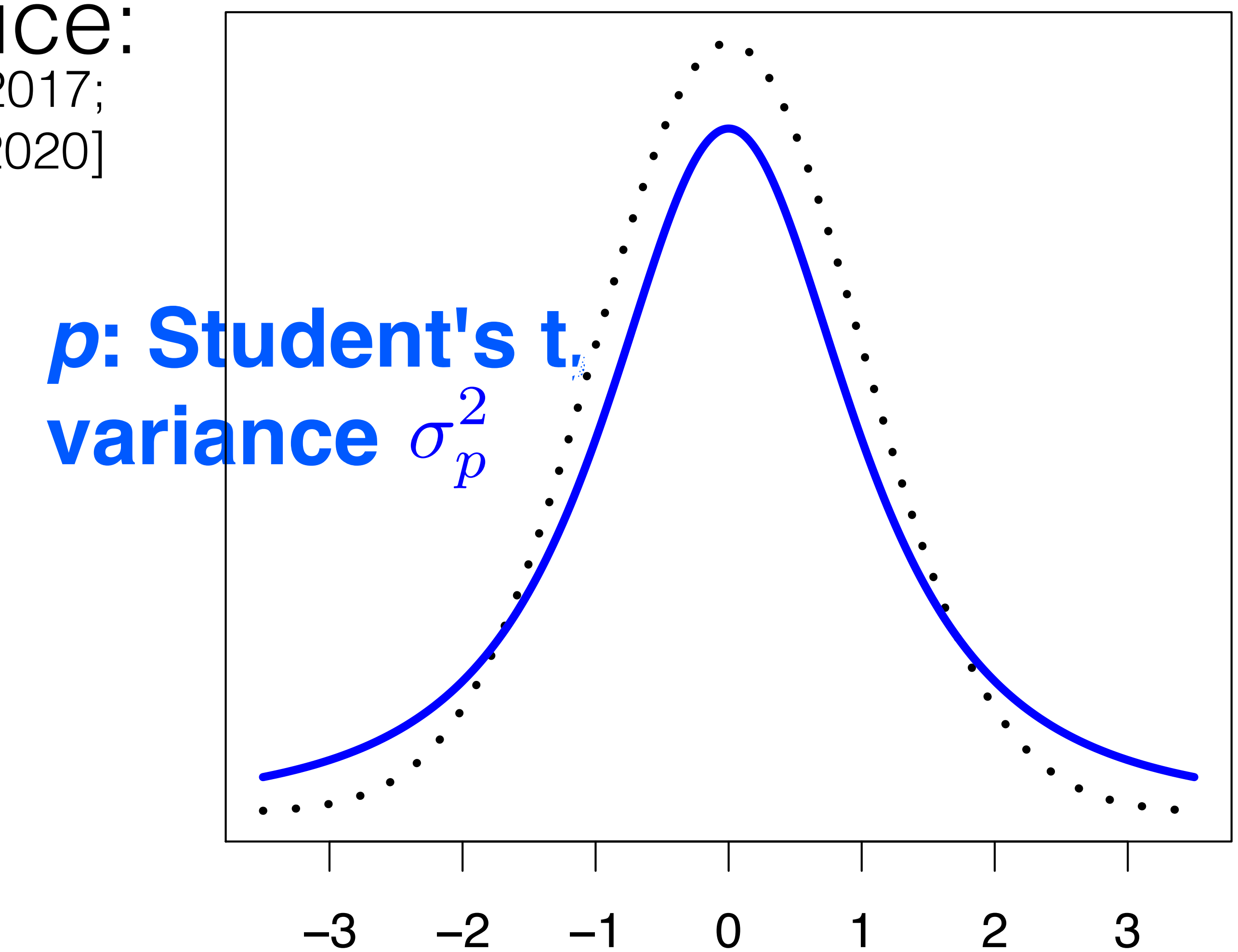
[Huggins, Karsprzak, Campbell, Broderick 2020]

Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3 [Baqué et al 2017; Huggins et al 2020]
- Take any constant *c*

Proposition. Can have small KL (<0.23) & arbitrarily bad variance estimate

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



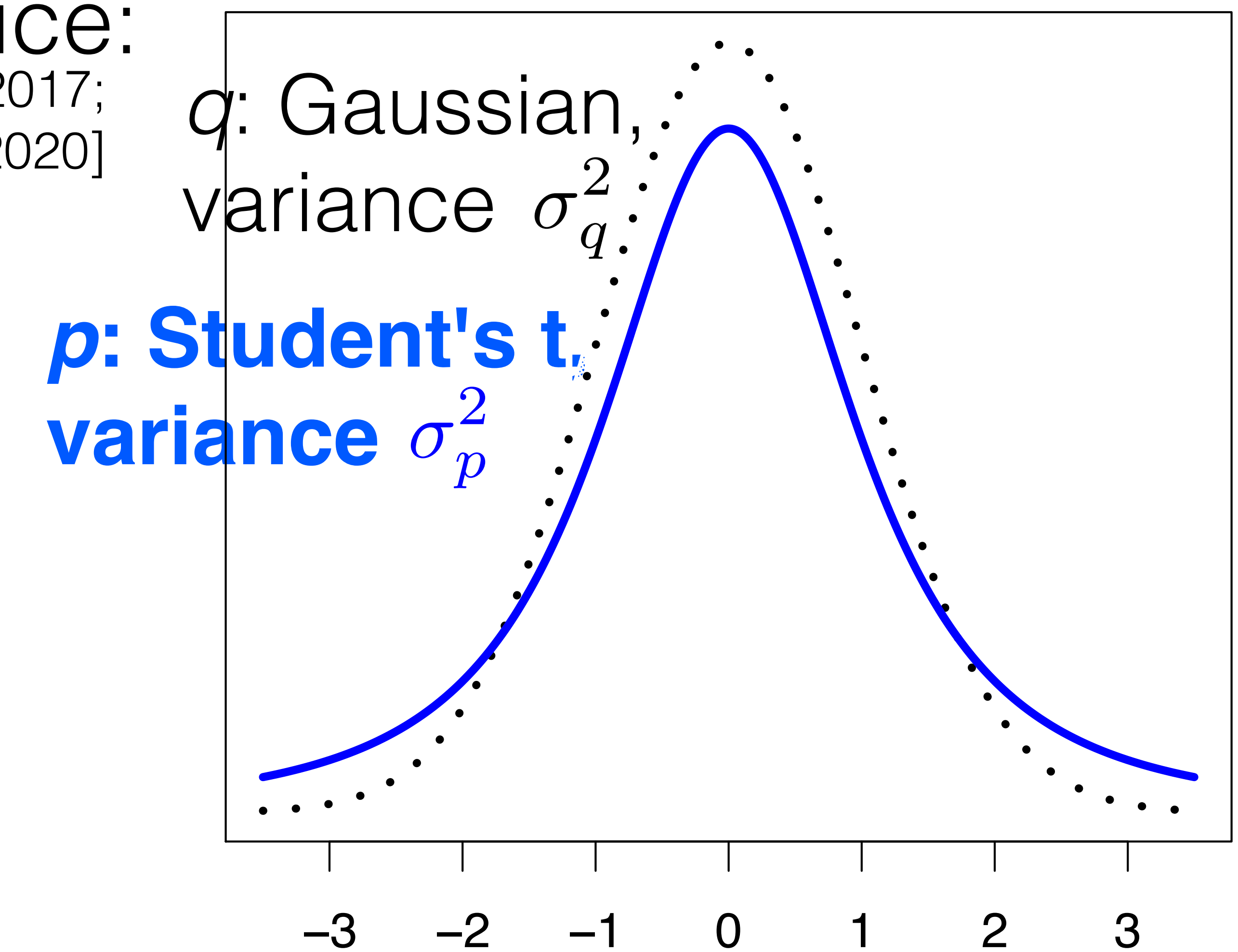
[Huggins, Karsprzak, Campbell, Broderick 2020]

Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3 [Baqué et al 2017; Huggins et al 2020]
- Take any constant *c*

Proposition. Can have small KL (<0.23) & arbitrarily bad variance estimate

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

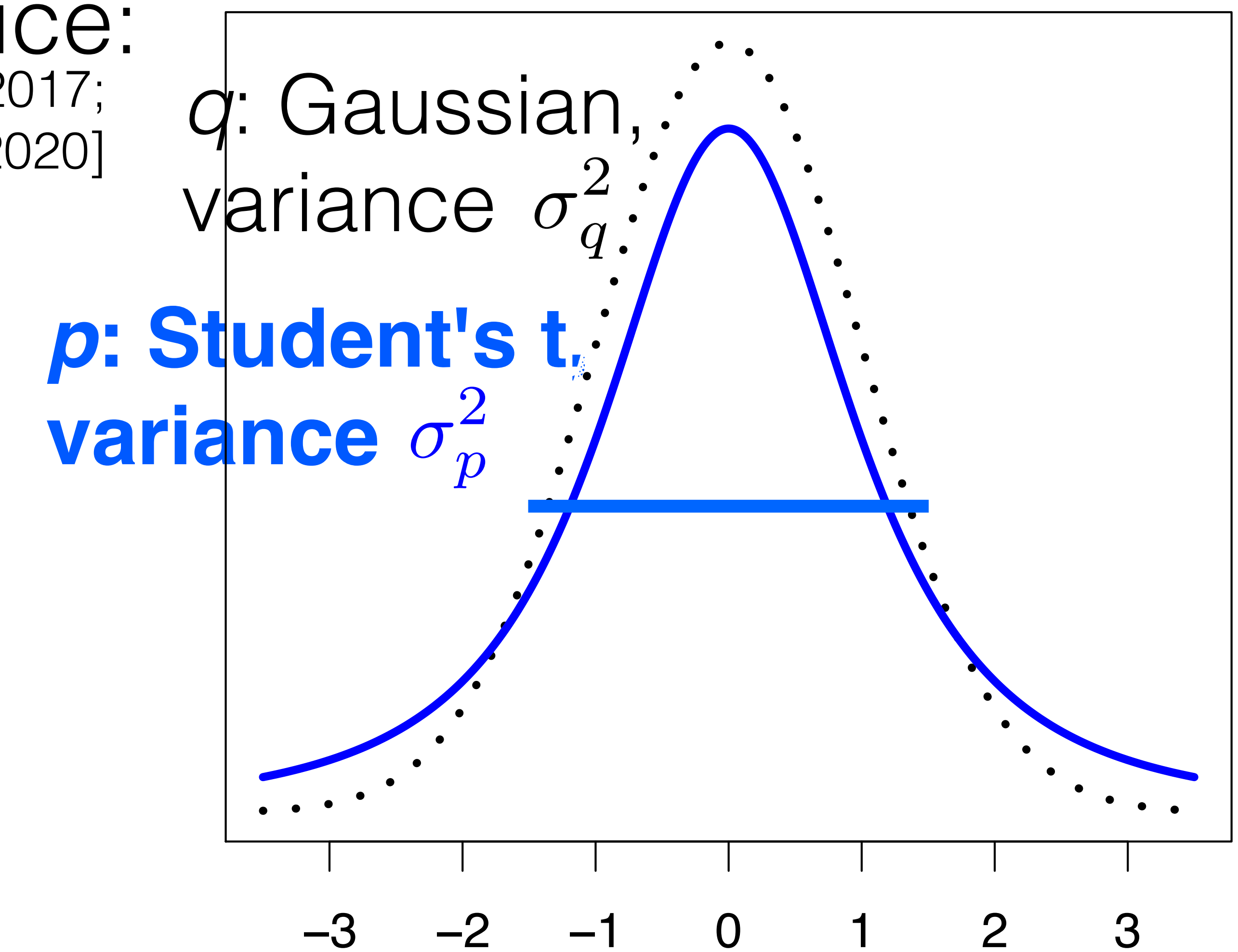


[Huggins, Karsprzak, Campbell, Broderick 2020]

Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3 [Baqué et al 2017; Huggins et al 2020]
- Take any constant *c*

Proposition. Can have small KL (<0.23) & arbitrarily bad variance estimate
$$\sigma_p^2 \geq c\sigma_q^2$$



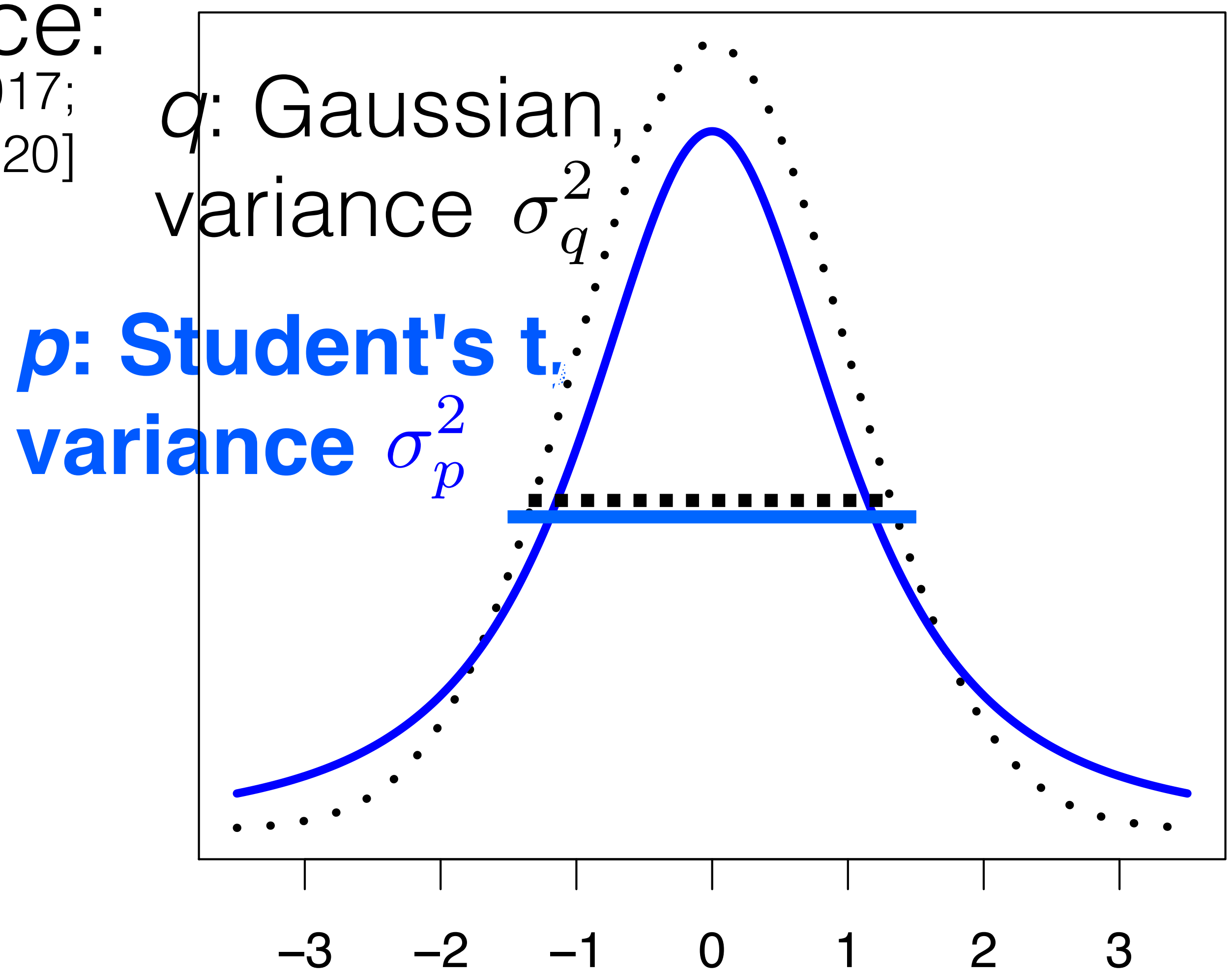
[Huggins, Karsprzak, Campbell, Broderick 2020]

Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3 [Baqué et al 2017; Huggins et al 2020]
- Take any constant c

Proposition. Can have small KL (<0.23) & arbitrarily bad variance estimate

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



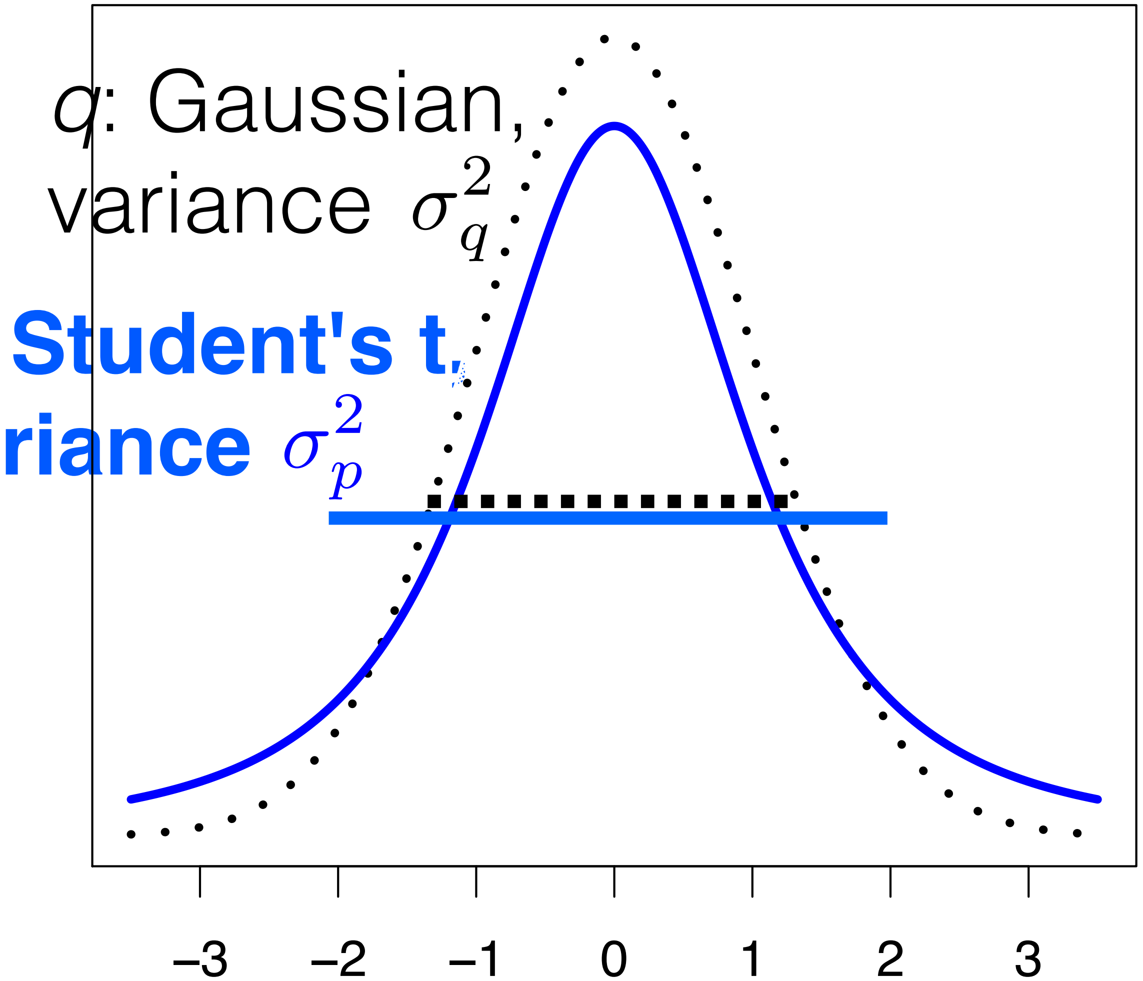
[Huggins, Karsprzak, Campbell, Broderick 2020]

Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3 [Baqué et al 2017; Huggins et al 2020]
- Take any constant *c*

Proposition. Can have small KL (<0.23) & arbitrarily bad variance estimate

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



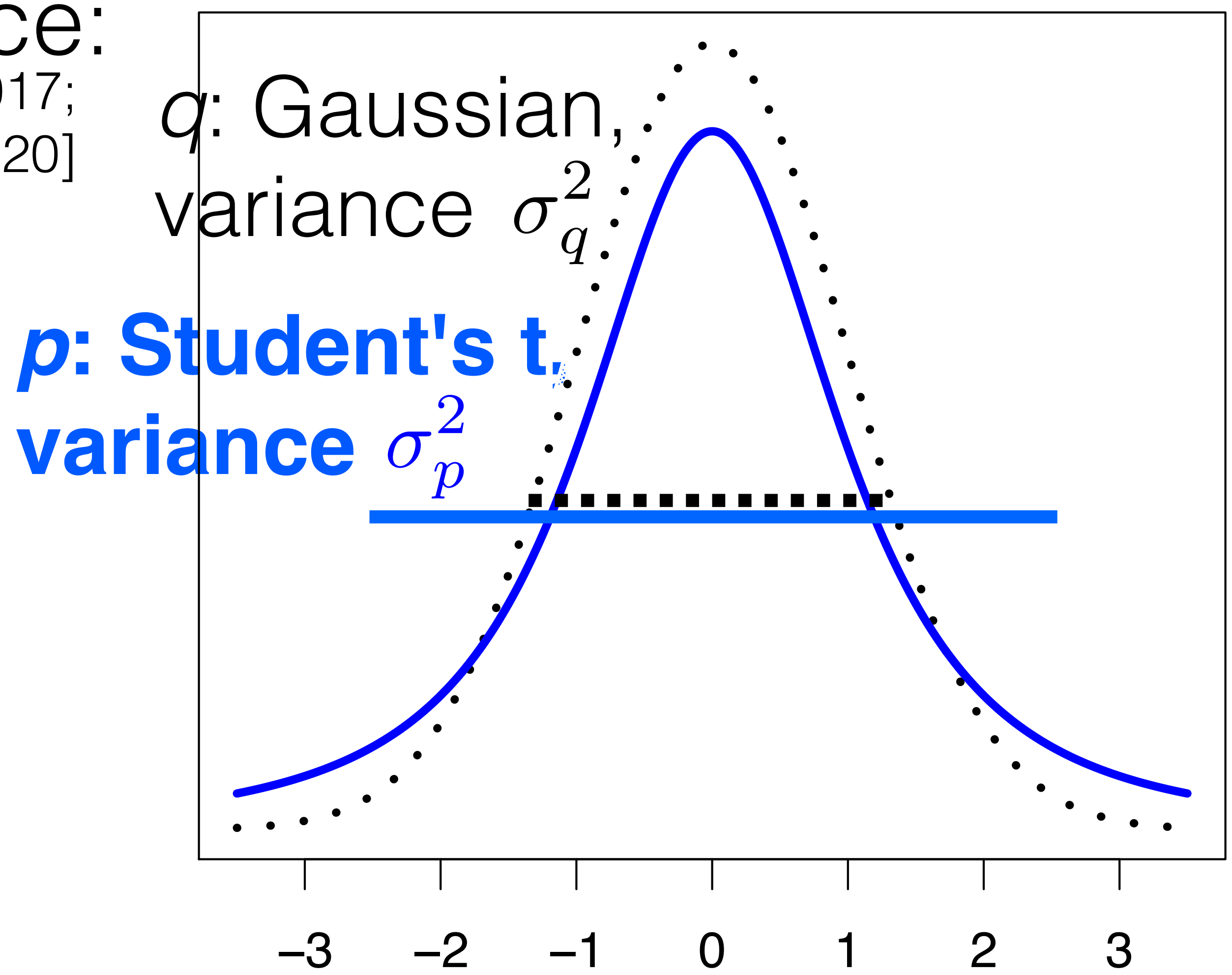
[Huggins, Karsprzak, Campbell, Broderick 2020]

Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3 [Baqué et al 2017; Huggins et al 2020]
- Take any constant *c*

Proposition. Can have small KL (<0.23) & arbitrarily bad variance estimate

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



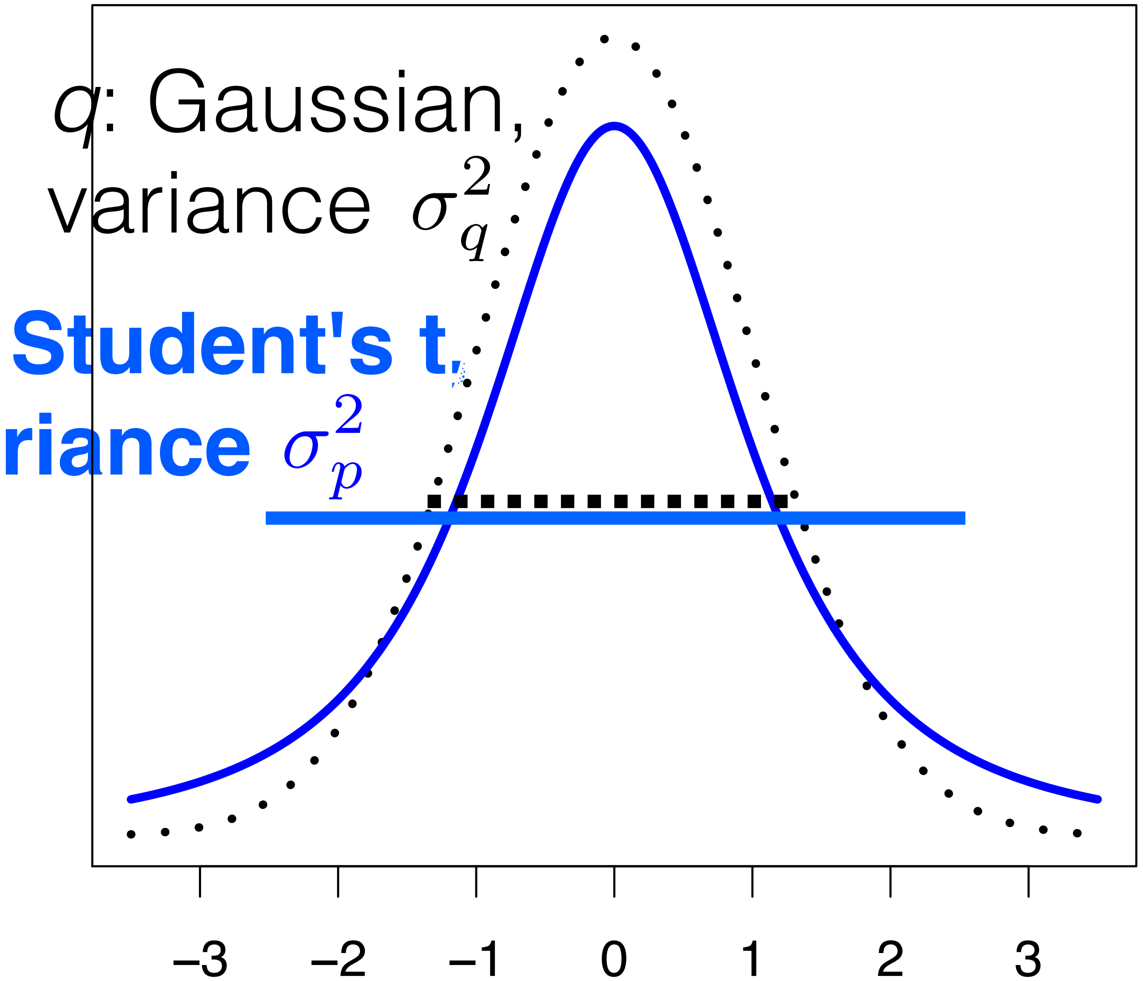
[Huggins, Karsprzak, Campbell, Broderick 2020]

Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3 [Baqué et al 2017; Huggins et al 2020]
- Take any constant *c*

Proposition. Can have small KL (<0.23) & arbitrarily bad variance estimate

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



Proposition. Can have small KL (<0.9) and arbitrarily bad mean estimate

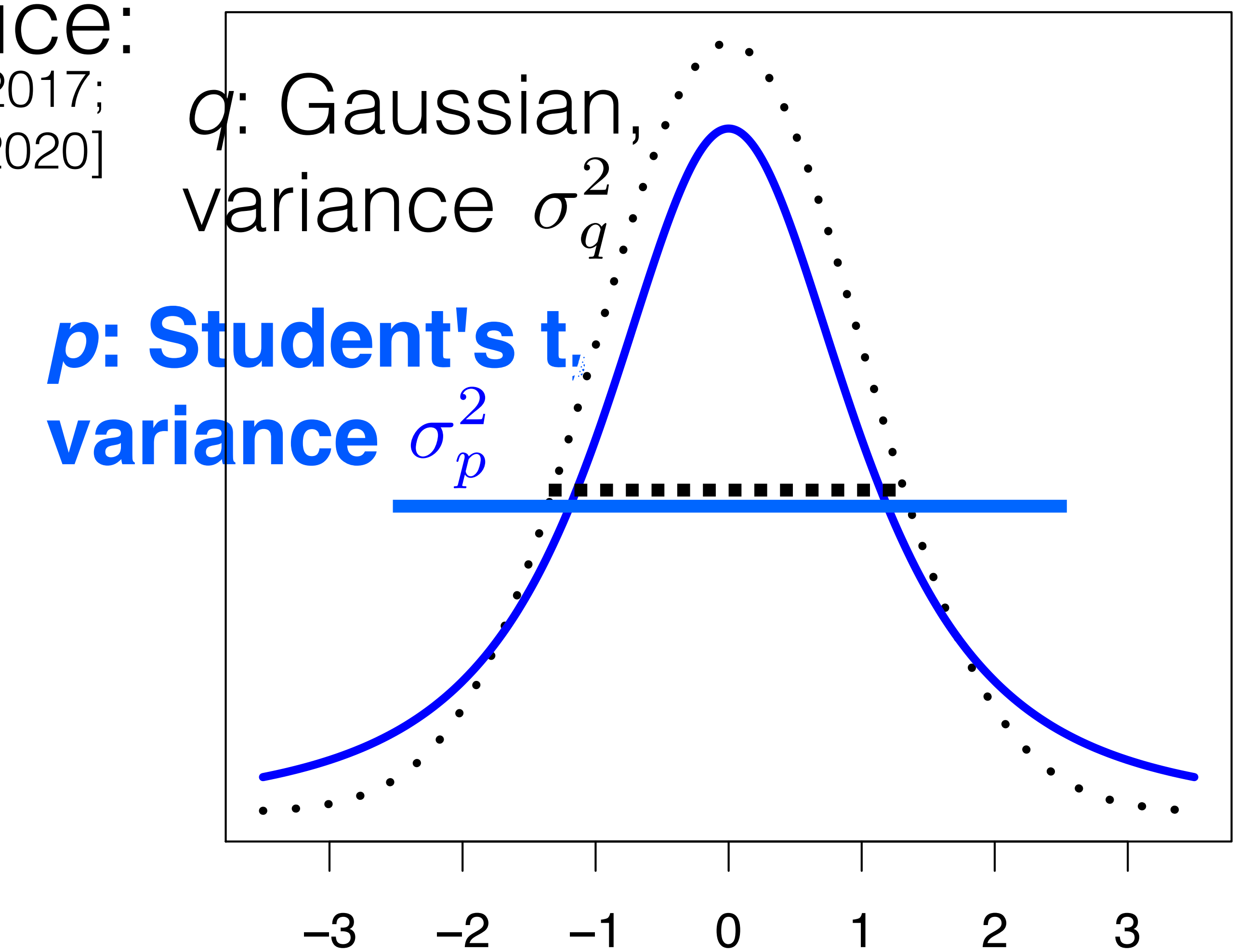
$$(m_p - m_q)^2 \geq c\sigma_p^2$$

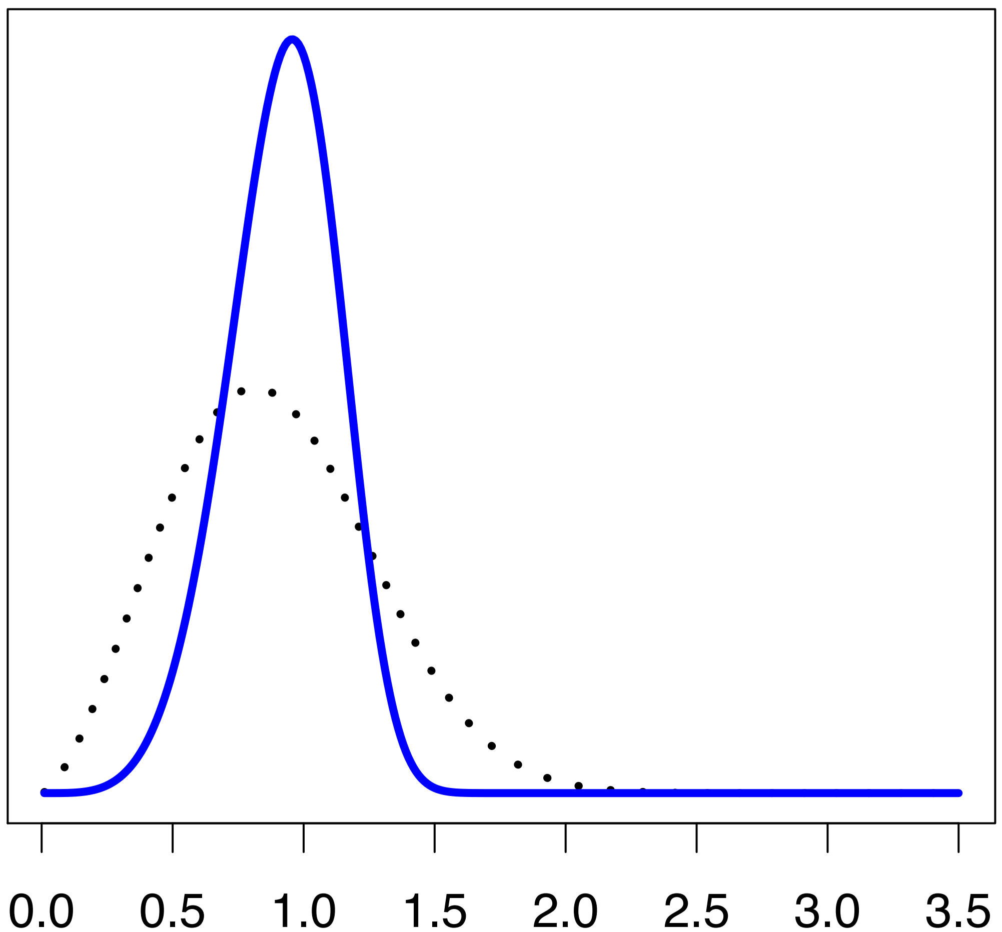
[Huggins, Karsprzak, Campbell, Broderick 2020]

Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3 [Baqué et al 2017; Huggins et al 2020]
- Take any constant *c*

Proposition. Can have small KL (<0.23) & arbitrarily bad variance estimate
$$\sigma_p^2 \geq c\sigma_q^2$$





Proposition. Can have small KL (<0.9) and arbitrarily bad mean estimate
$$(m_p - m_q)^2 \geq c\sigma_p^2$$

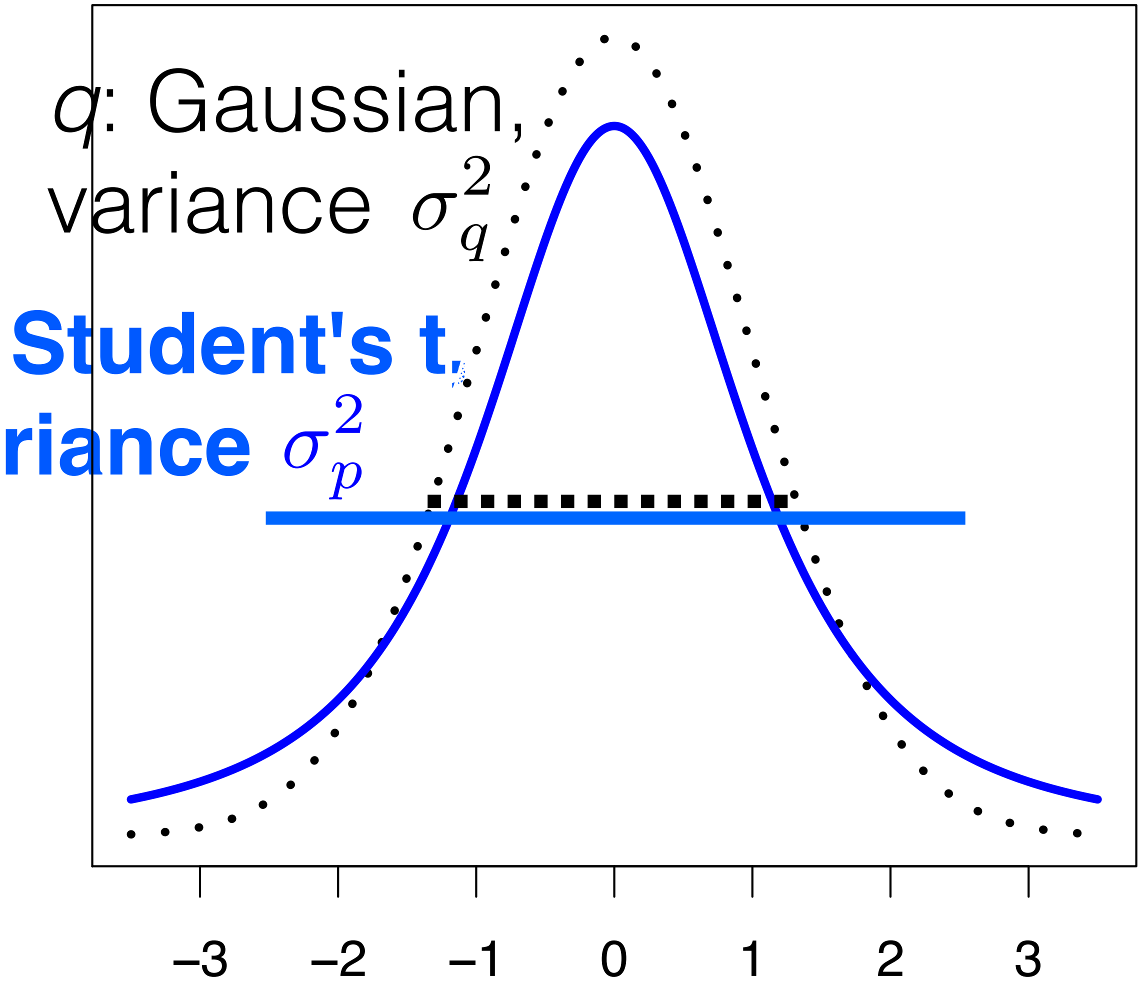
[Huggins, Karsprzak, Campbell, Broderick 2020]

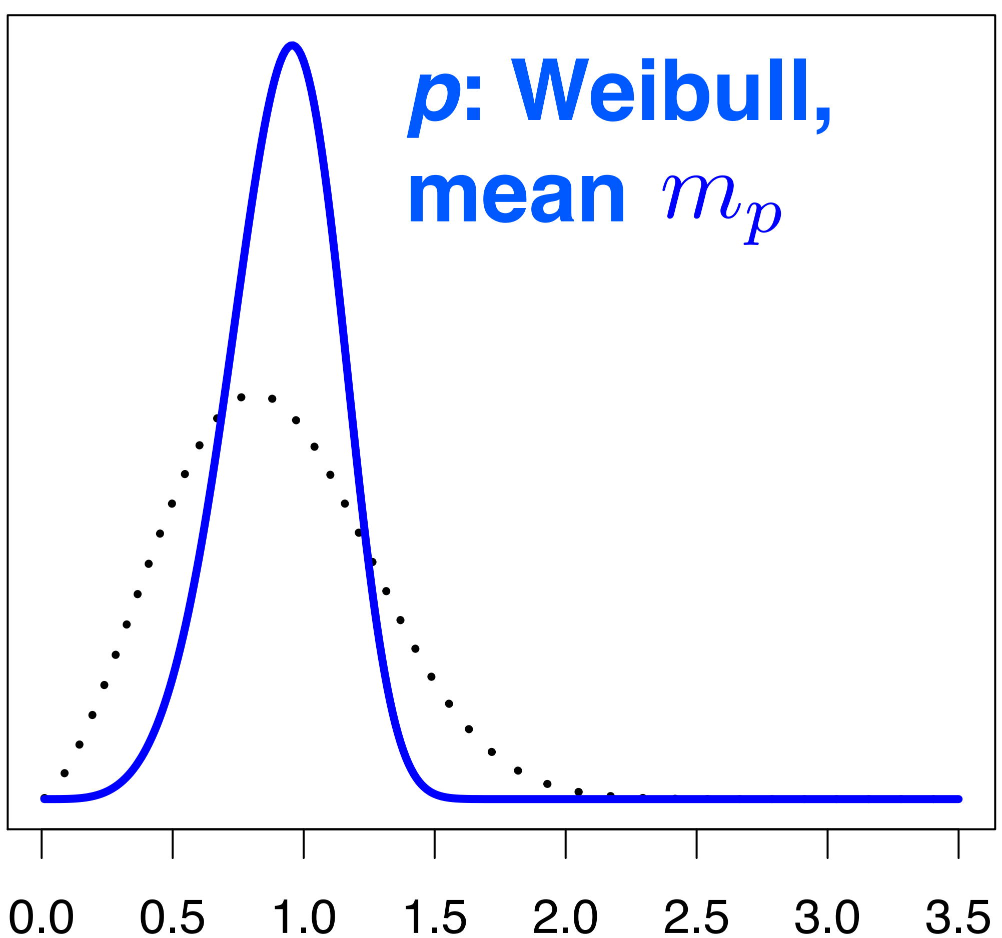
Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3 [Baqué et al 2017; Huggins et al 2020]
- Take any constant *c*

Proposition. Can have small KL (<0.23) & arbitrarily bad variance estimate

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





Proposition. Can have small KL (<0.9) and arbitrarily bad mean estimate

$$(m_p - m_q)^2 \geq c\sigma_p^2$$

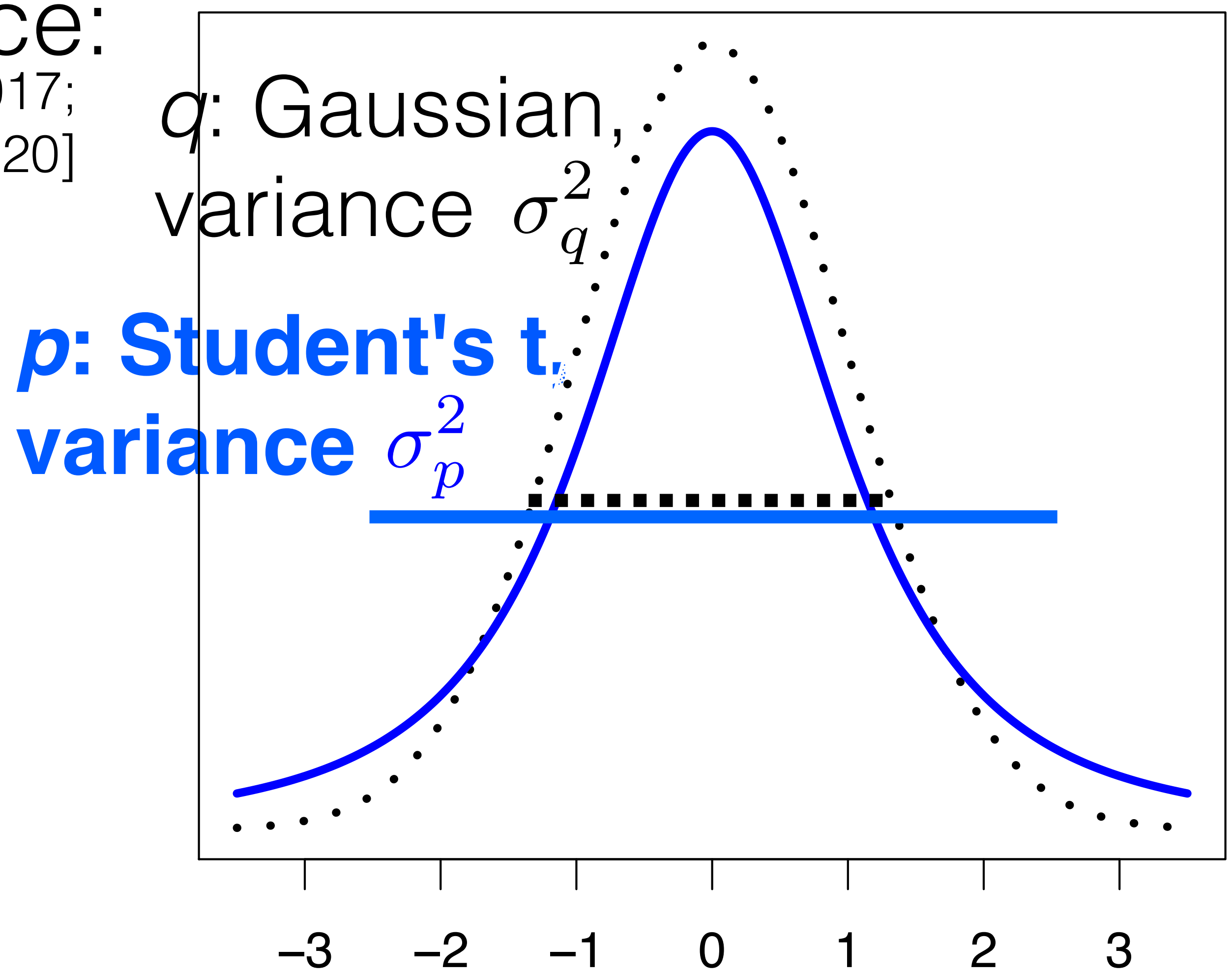
[Huggins, Karsprzak, Campbell, Broderick 2020]

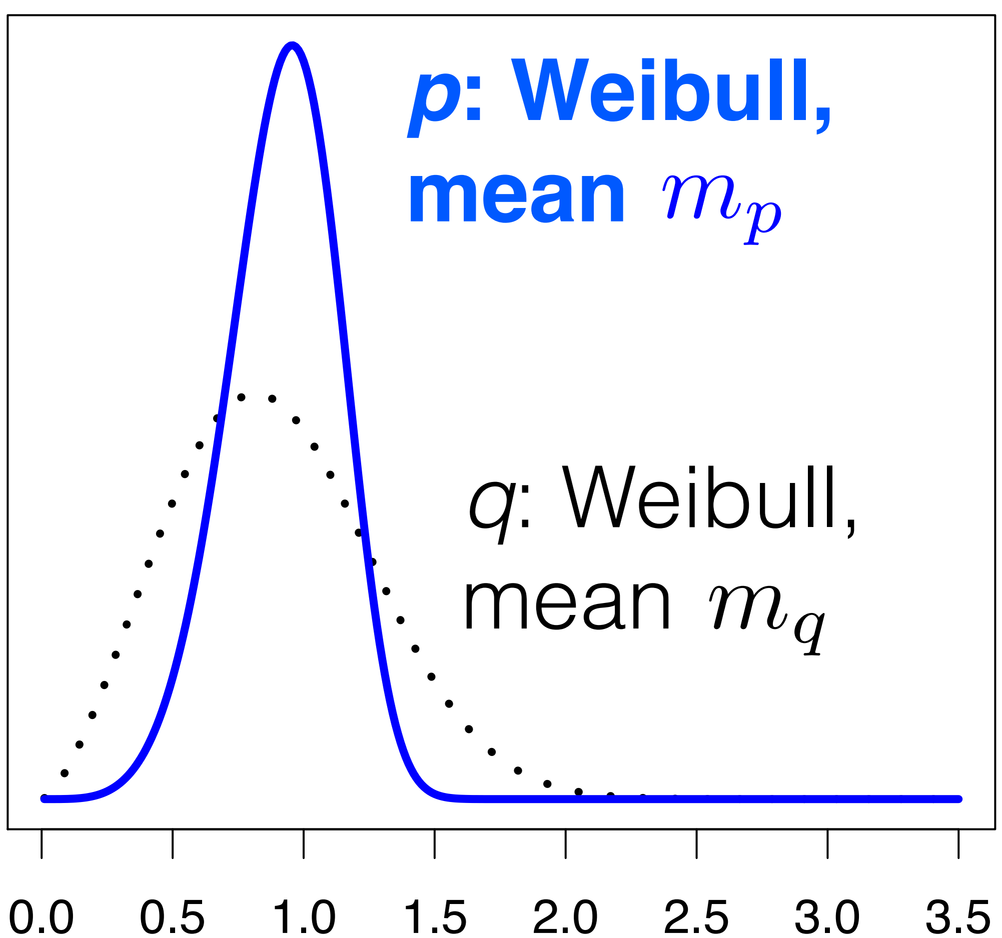
Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3 [Baqué et al 2017; Huggins et al 2020]
- Take any constant c

Proposition. Can have small KL (<0.23) & arbitrarily bad variance estimate

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





Proposition. Can have small KL (<0.9) and arbitrarily bad mean estimate

$$(m_p - m_q)^2 \geq c\sigma_p^2$$

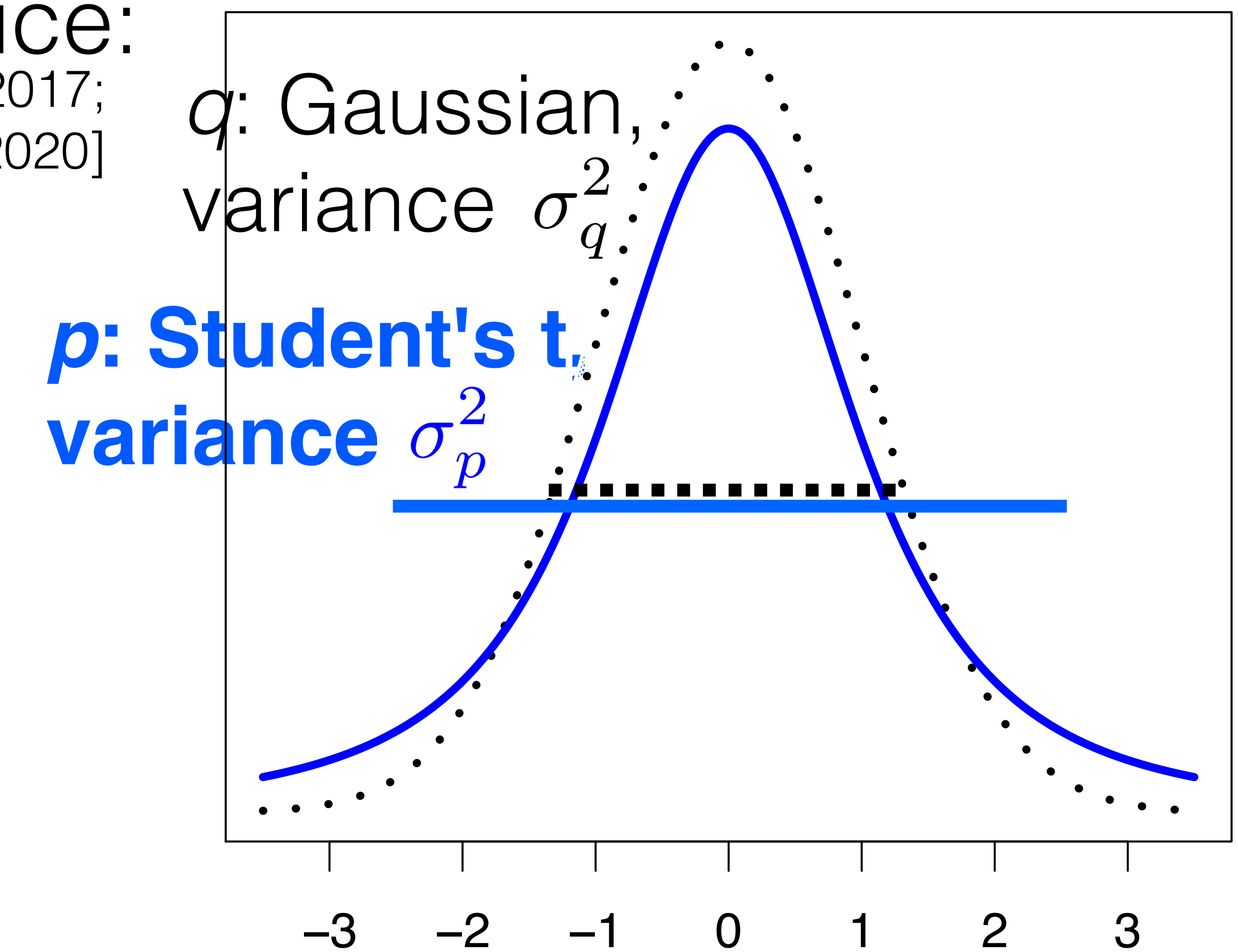
[Huggins, Karsprzak, Campbell, Broderick 2020]

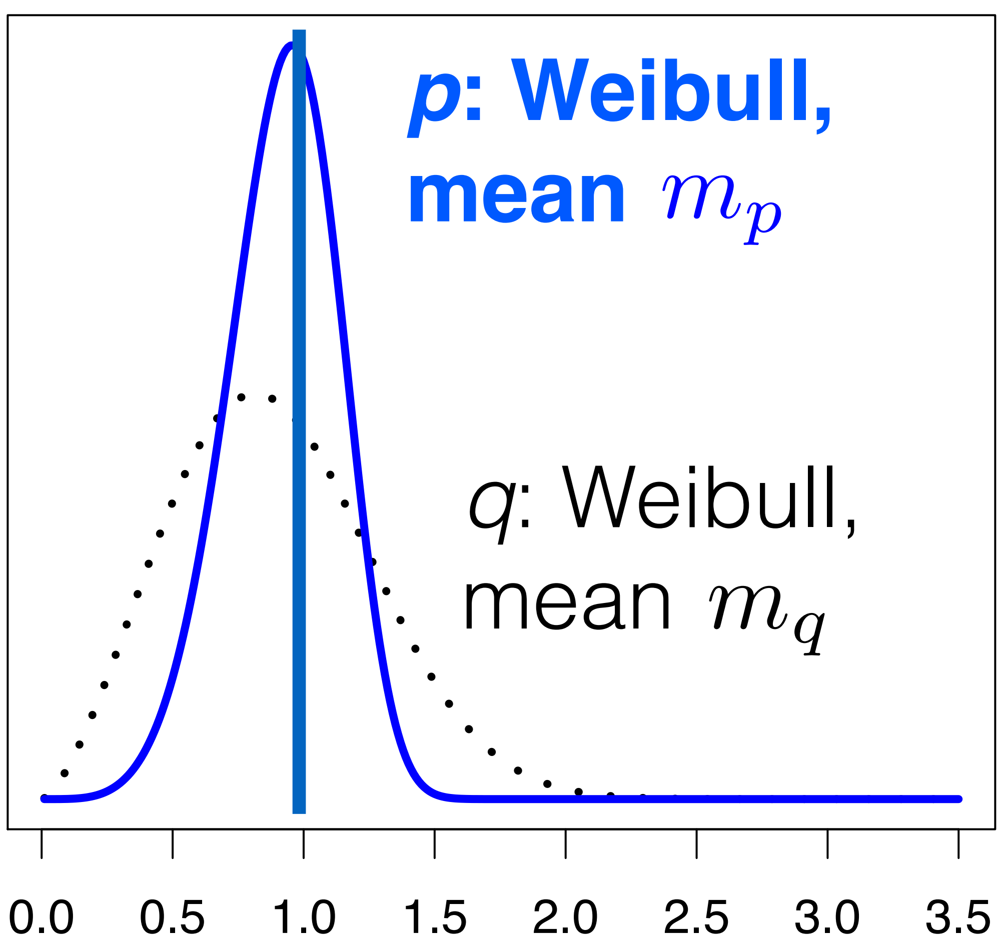
Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3 [Baqué et al 2017; Huggins et al 2020]
- Take any constant c

Proposition. Can have small KL (<0.23) & arbitrarily bad variance estimate

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





Proposition. Can have small KL (<0.9) and arbitrarily bad mean estimate

$$(m_p - m_q)^2 \geq c\sigma_p^2$$

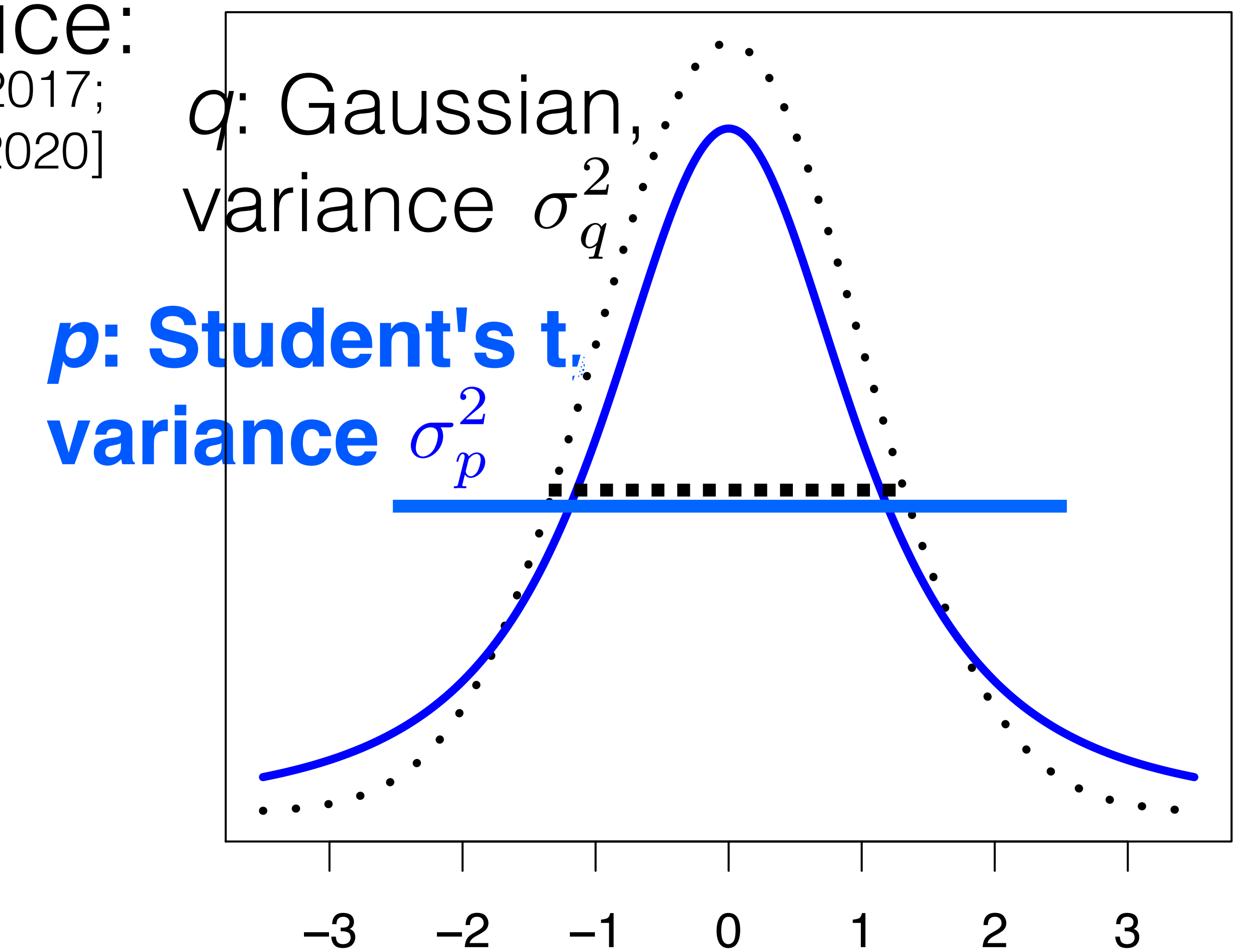
[Huggins, Karsprzak, Campbell, Broderick 2020]

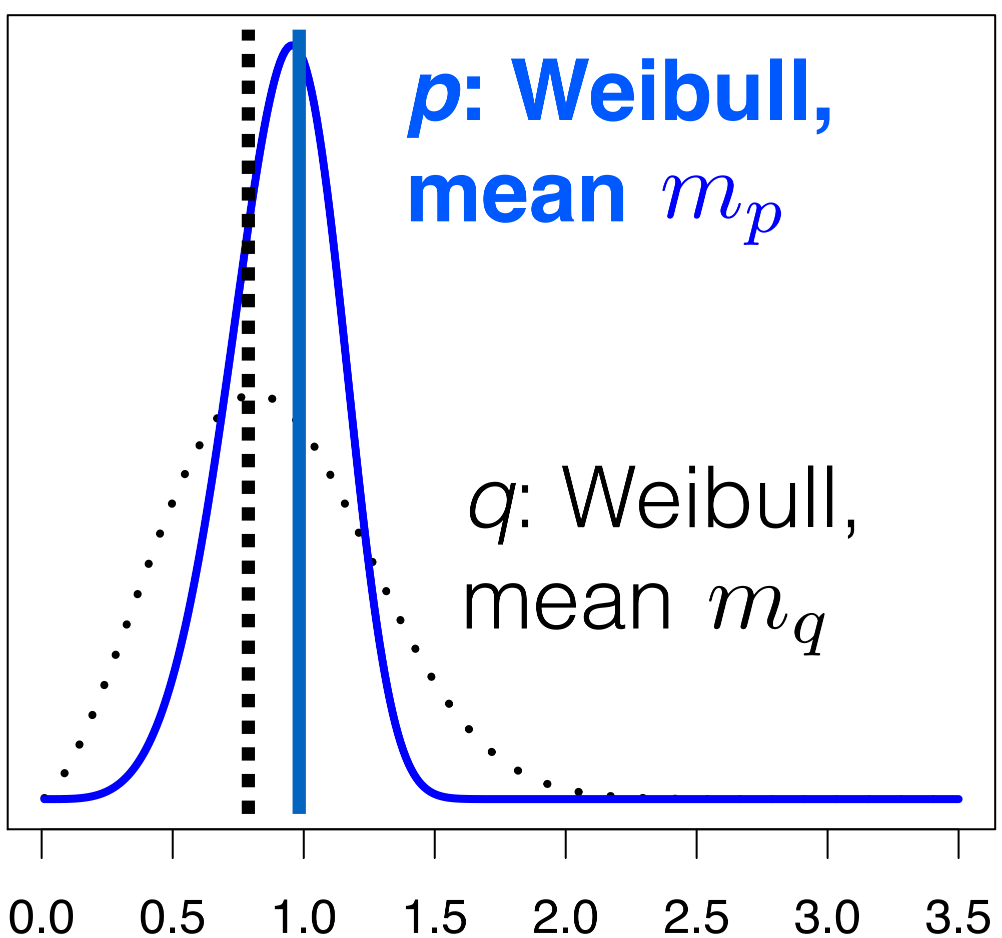
Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3 [Baqué et al 2017; Huggins et al 2020]
- Take any constant c

Proposition. Can have small KL (<0.23) & arbitrarily bad variance estimate

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



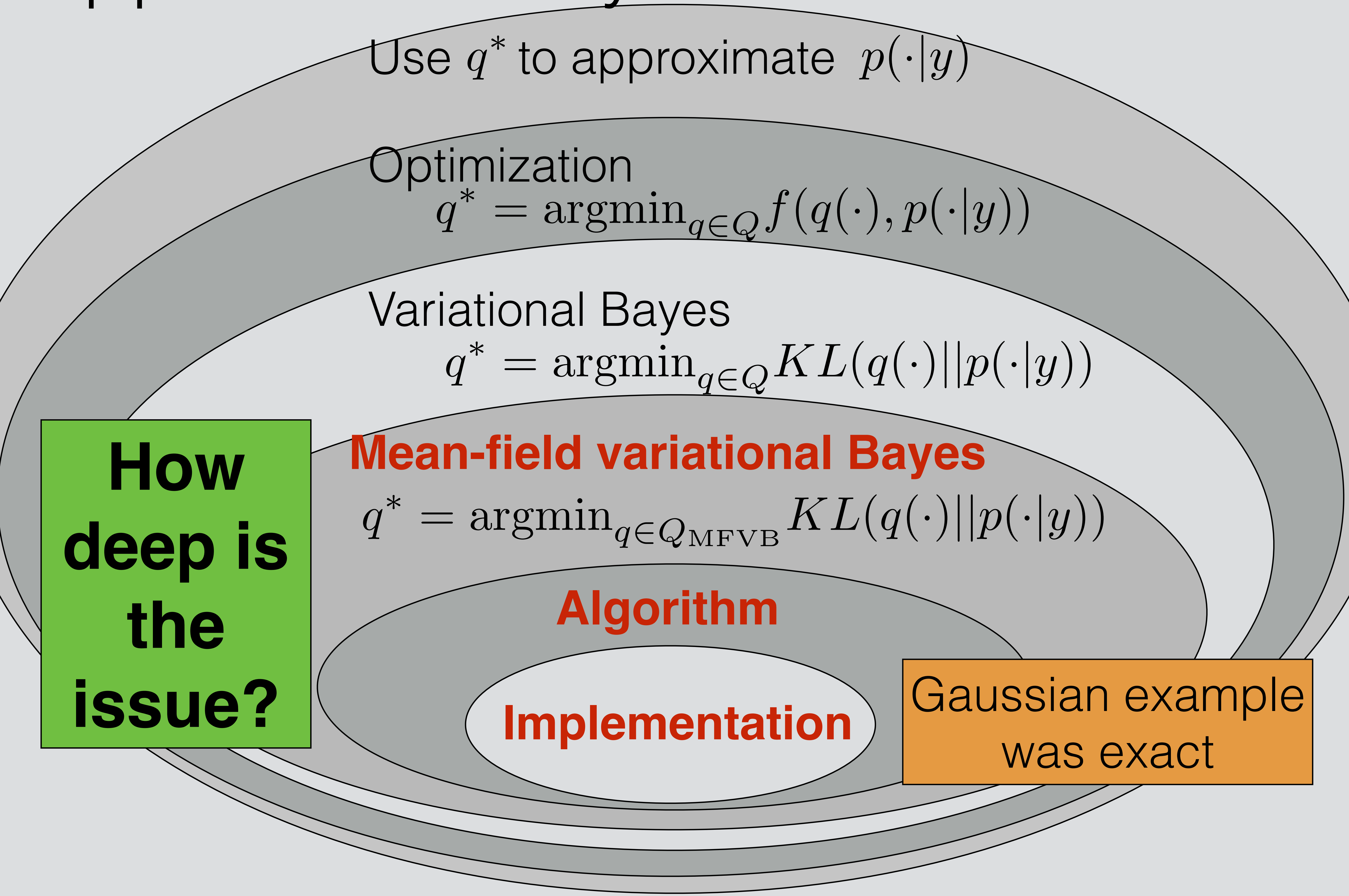


Proposition. Can have small KL (<0.9) and arbitrarily bad mean estimate

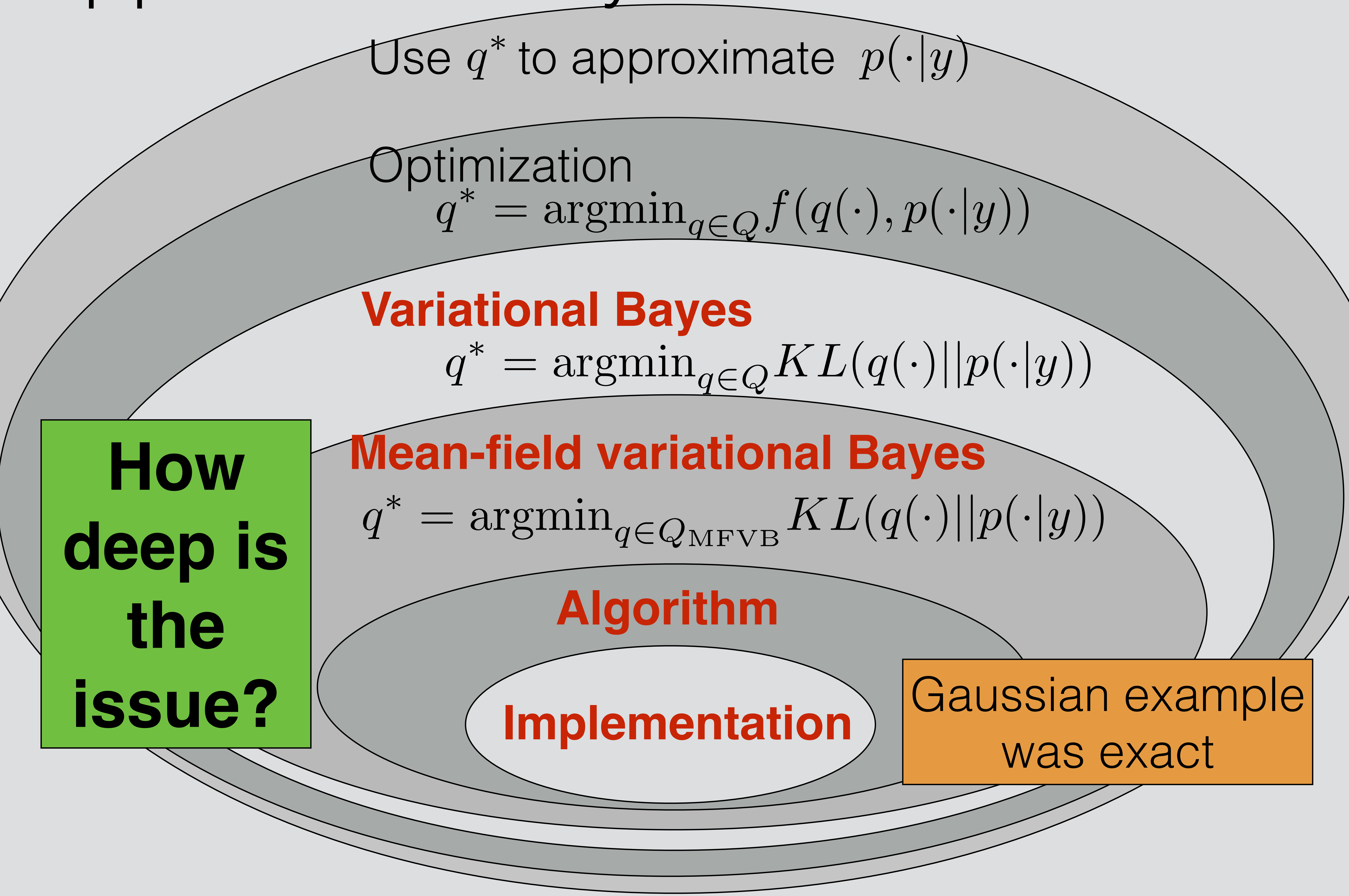
$$(m_p - m_q)^2 \geq c\sigma_p^2$$

[Huggins, Karsprzak, Campbell, Broderick 2020]

Approximate Bayesian inference



Approximate Bayesian inference



Roadmap

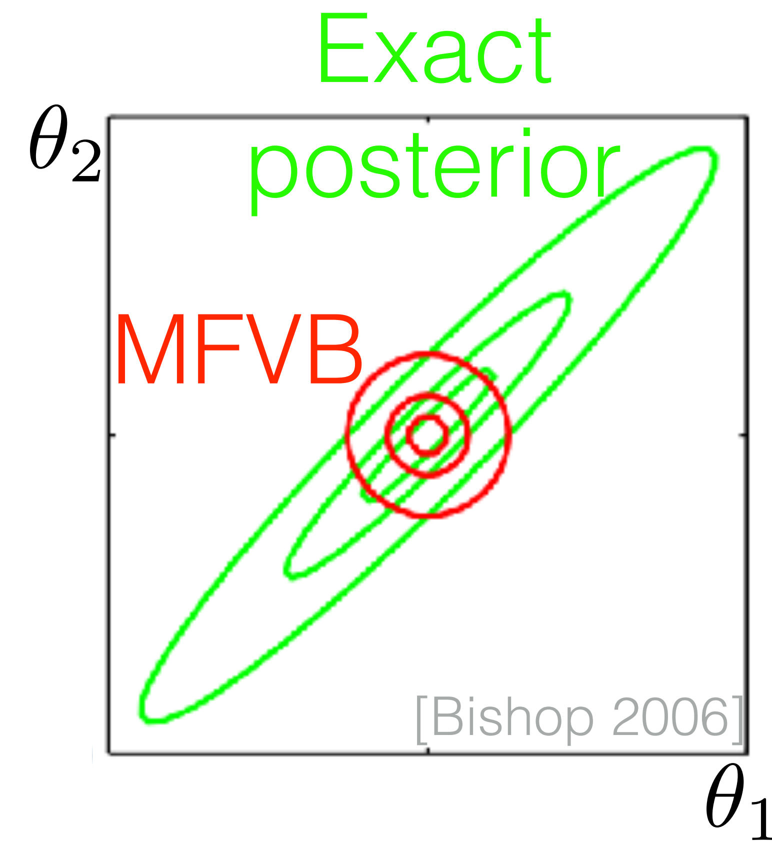
- Bayes & Approximate Bayes review
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use VB?
- When can we trust VB?
- Where do we go from here?

Roadmap

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

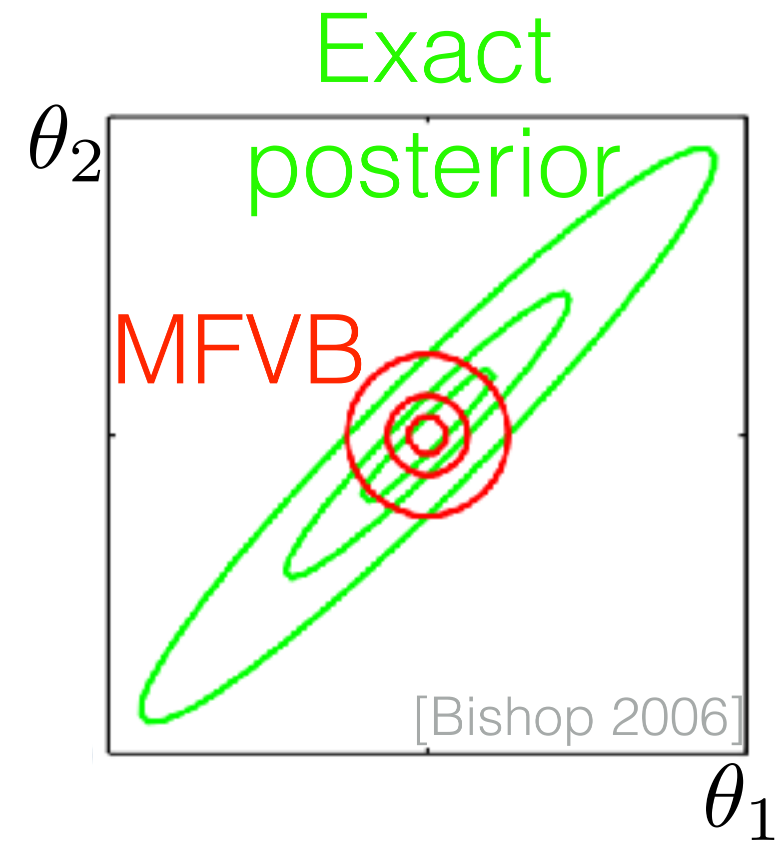
We can fix VB uncertainty

We can fix VB uncertainty



We can fix VB uncertainty

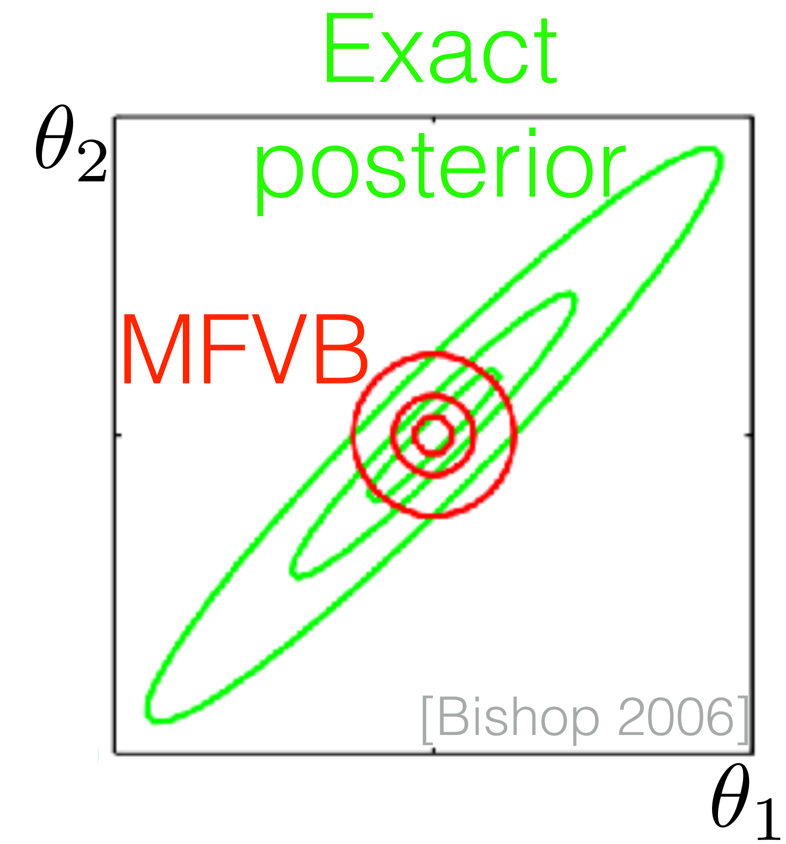
- We provide: linear response variational Bayes (LRVB) [see also Opper, Winther 2003]



[Giordano, Broderick, Jordan 2015, 2018]

We can fix VB uncertainty

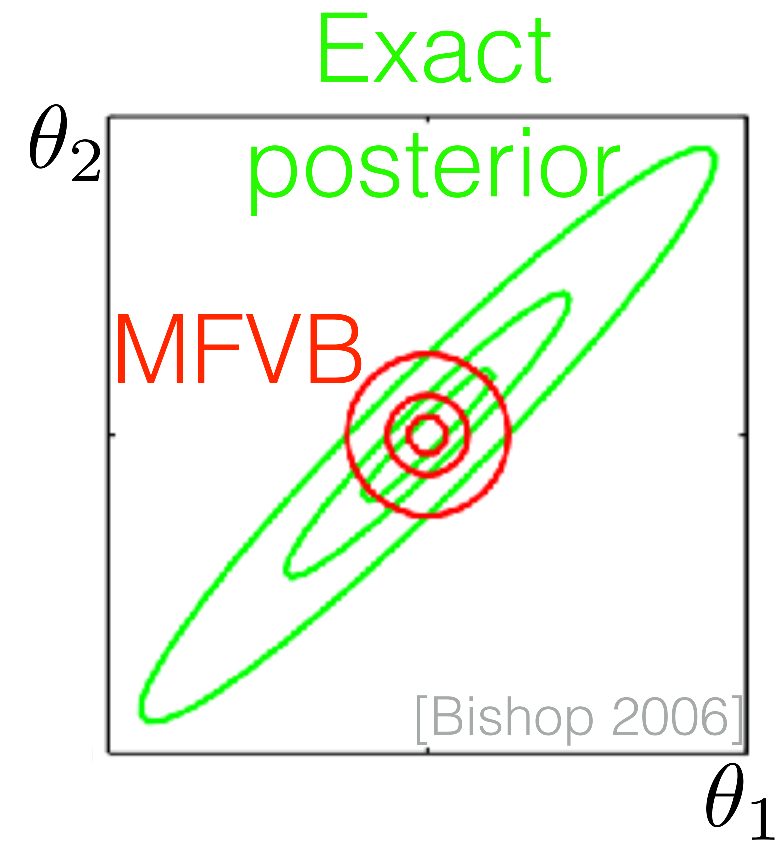
- We provide: linear response variational Bayes (LRVB) [see also Opper, Winther 2003]
- Procedure: VB (e.g. MFVB), then LRVB



[Giordano, Broderick, Jordan 2015, 2018]

We can fix VB uncertainty

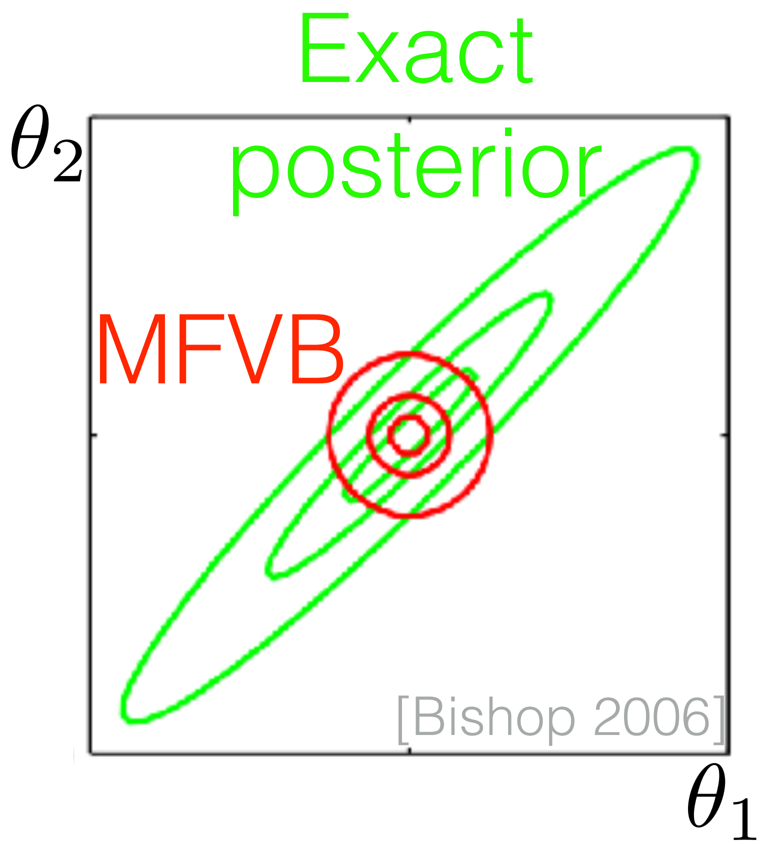
- We provide: linear response variational Bayes (LRVB) [see also Opper, Winther 2003]
- Procedure: VB (e.g. MFVB), then LRVB
  - Perturbation ideas, statistical physics



[Giordano, Broderick, Jordan 2015, 2018]

We can fix VB uncertainty

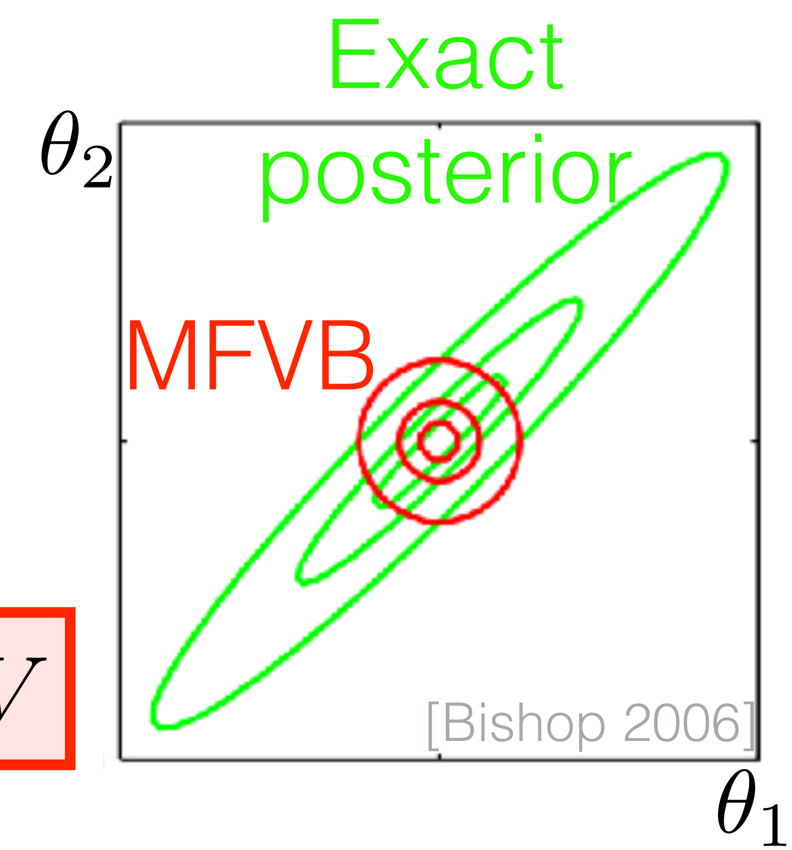
- We provide: linear response variational Bayes (LRVB) [see also Opper, Winther 2003]
- Procedure: VB (e.g. MFVB), then LRVB
  - Perturbation ideas, statistical physics
  - Correction to VB:



[Giordano, Broderick, Jordan 2015, 2018]

We can fix VB uncertainty

- We provide: linear response variational Bayes (LRVB) [see also Opper, Winther 2003]
- Procedure: VB (e.g. MFVB), then LRVB
    - Perturbation ideas, statistical physics
    - Correction to VB: $\boxed{\hat{\Sigma}} = (I - VH)^{-1}\boxed{V}$

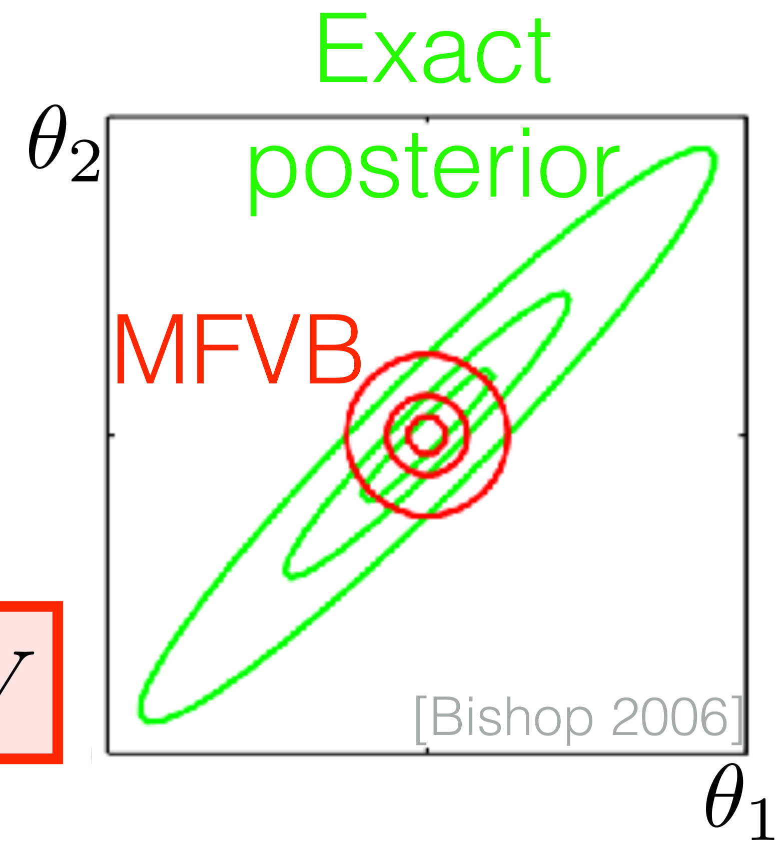


[Giordano, Broderick, Jordan 2015, 2018]

We can fix VB uncertainty

- We provide: linear response variational Bayes (LRVB) [see also Opper, Winther 2003]
- Procedure: VB (e.g. MFVB), then LRVB
  - Perturbation ideas, statistical physics
  - Correction to VB: $\boxed{\hat{\Sigma}} = (I - VH)^{-1} \boxed{V}$

computable from model with autodiff

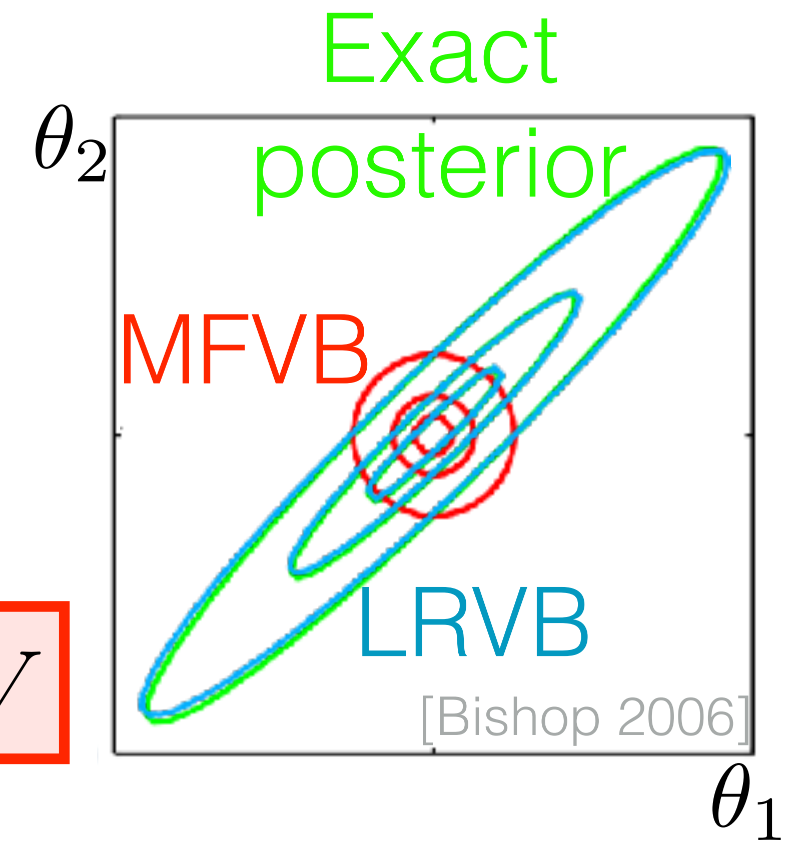


[Giordano, Broderick, Jordan 2015, 2018]

We can fix VB uncertainty

- We provide: linear response variational Bayes (LRVB) [see also Opper, Winther 2003]
- Procedure: VB (e.g. MFVB), then LRVB
  - Perturbation ideas, statistical physics
  - Correction to VB: $\boxed{\hat{\Sigma}} = (I - VH)^{-1} \boxed{V}$

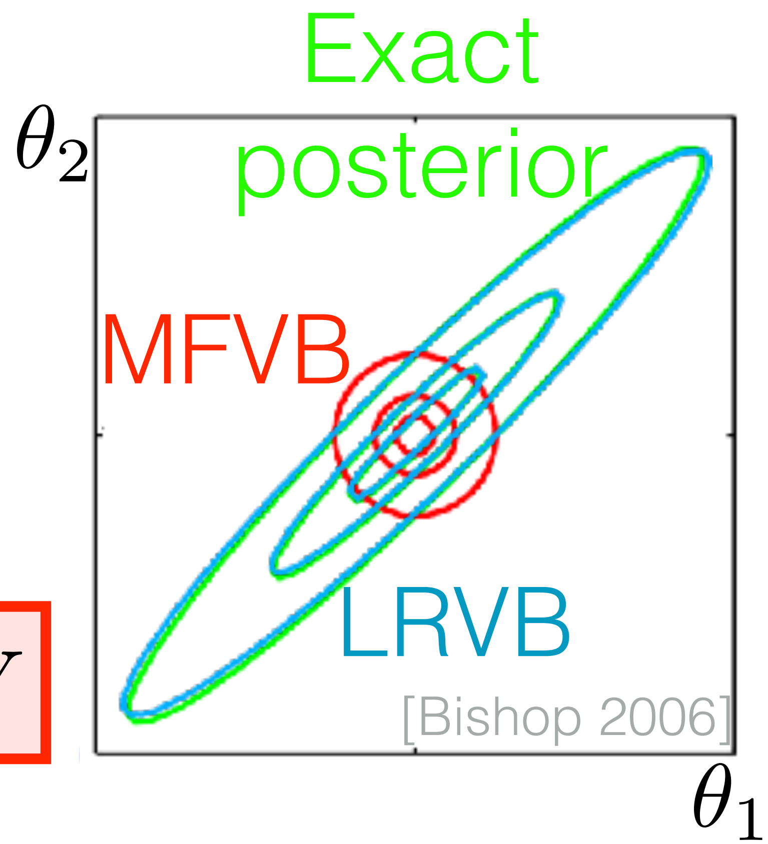
computable from model with autodiff



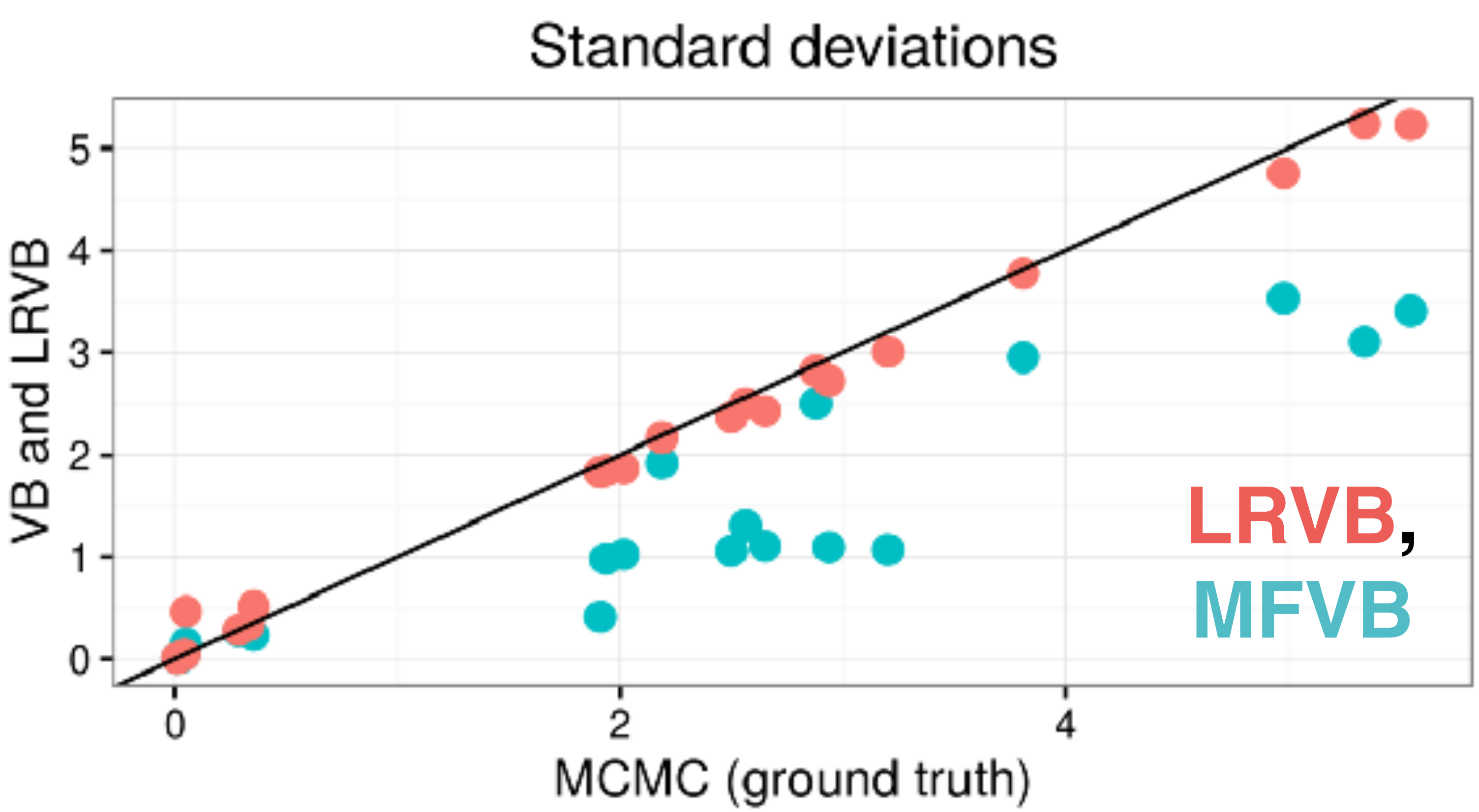
[Giordano, Broderick, Jordan 2015, 2018]

We can fix VB uncertainty

- We provide: linear response variational Bayes (LRVB) [see also Opper, Winther 2003]
- Procedure: VB (e.g. MFVB), then LRVB
  - Perturbation ideas, statistical physics
  - Correction to VB: $\boxed{\hat{\Sigma}} = (I - VH)^{-1} \boxed{V}$



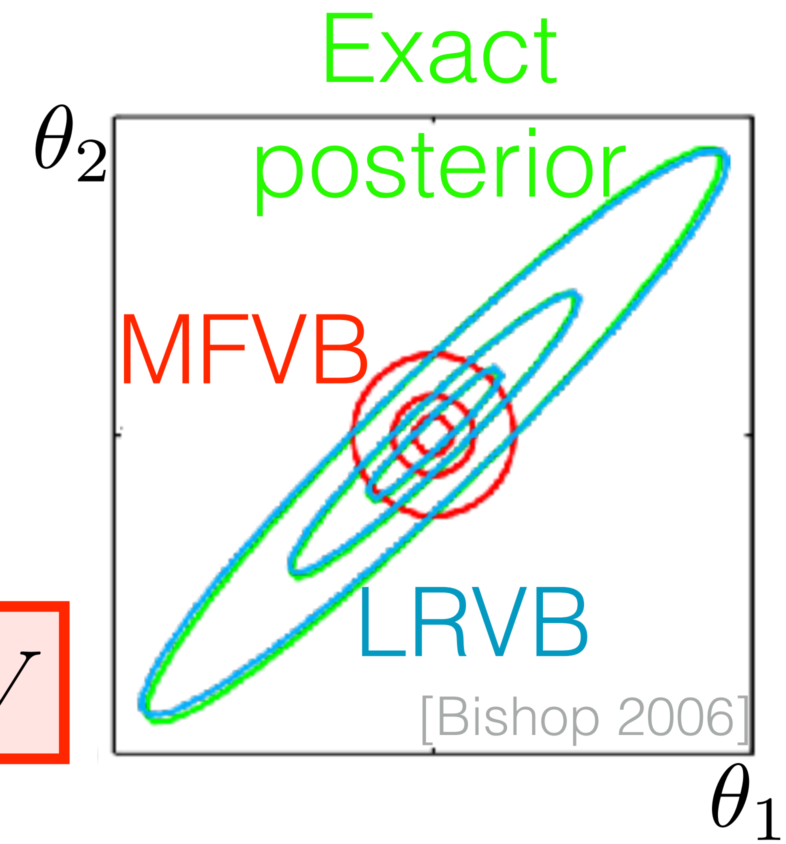
computable from model with autodiff



[Giordano, Broderick, Jordan 2015, 2018]

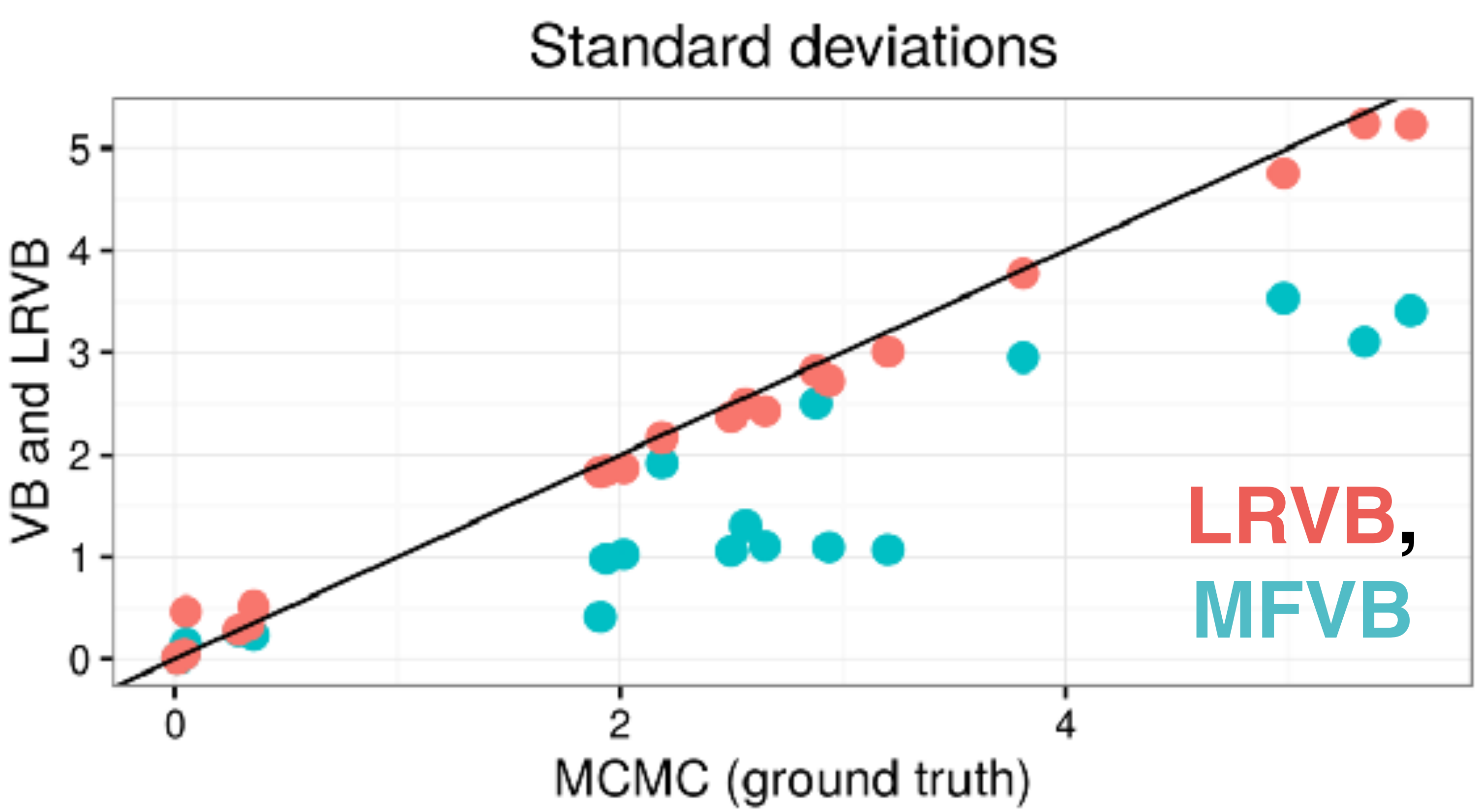
We can fix VB uncertainty

- We provide: linear response variational Bayes (LRVB) [see also Opper, Winther 2003]
- Procedure: VB (e.g. MFVB), then LRVB
  - Perturbation ideas, statistical physics
  - Correction to VB: $\boxed{\hat{\Sigma}} = (I - VH)^{-1} \boxed{V}$



computable from model with autodiff

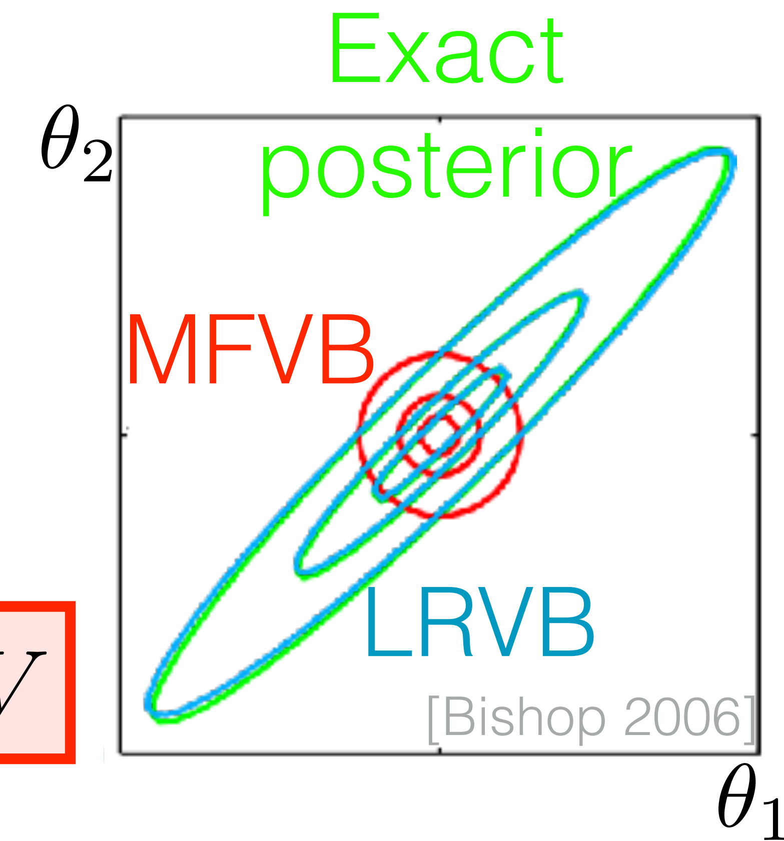
- Exact for Gaussians



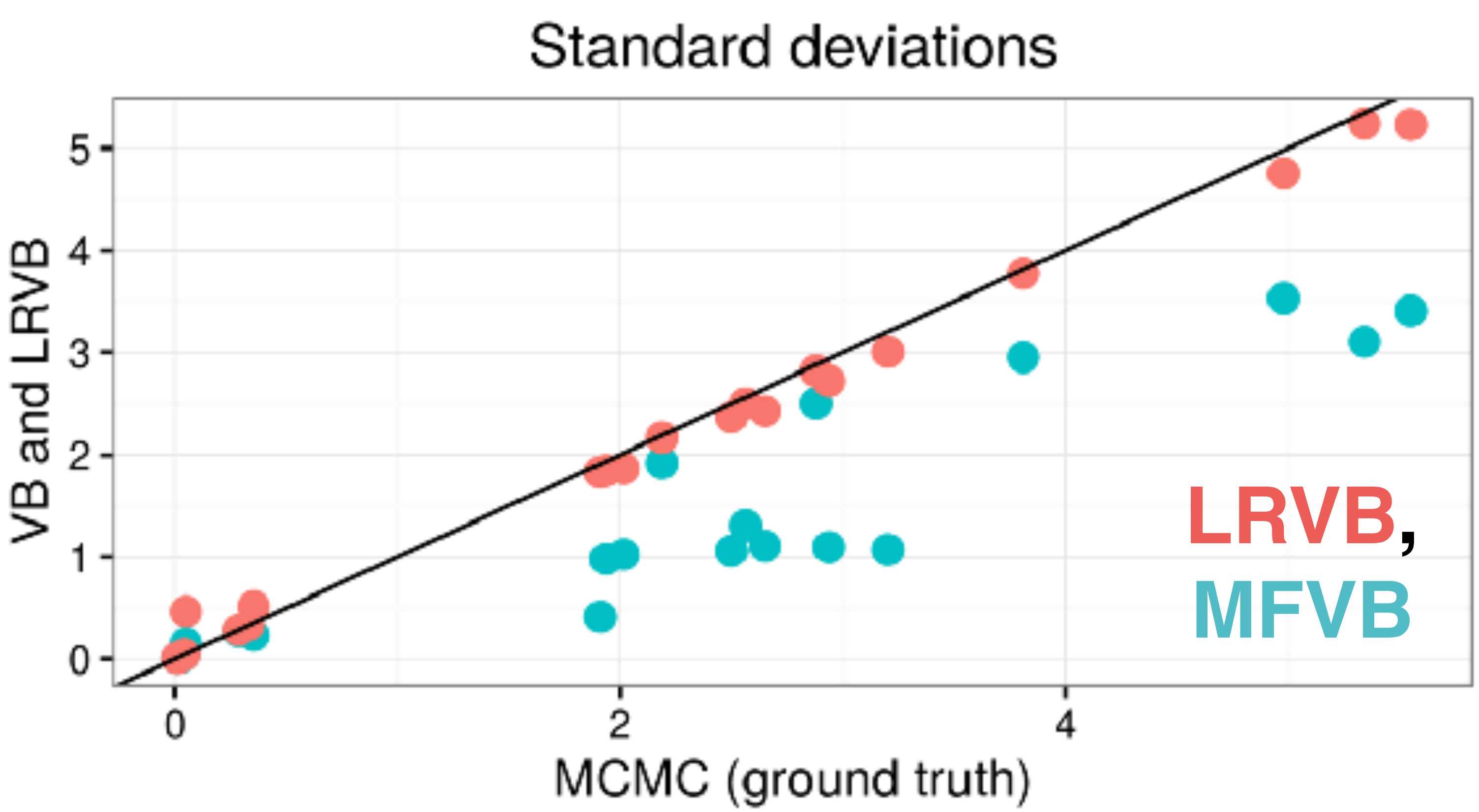
[Giordano, Broderick, Jordan 2015, 2018]

We can fix VB uncertainty

- We provide: linear response variational Bayes (LRVB) [see also Opper, Winther 2003]
- Procedure: VB (e.g. MFVB), then LRVB
  - Perturbation ideas, statistical physics
  - Correction to VB: $\boxed{\hat{\Sigma}} = (I - VH)^{-1}\boxed{V}$



computable from model with autodiff



- Exact for Gaussians
- Needs good posterior mean approximation in practice

[Giordano, Broderick, Jordan 2015, 2018]

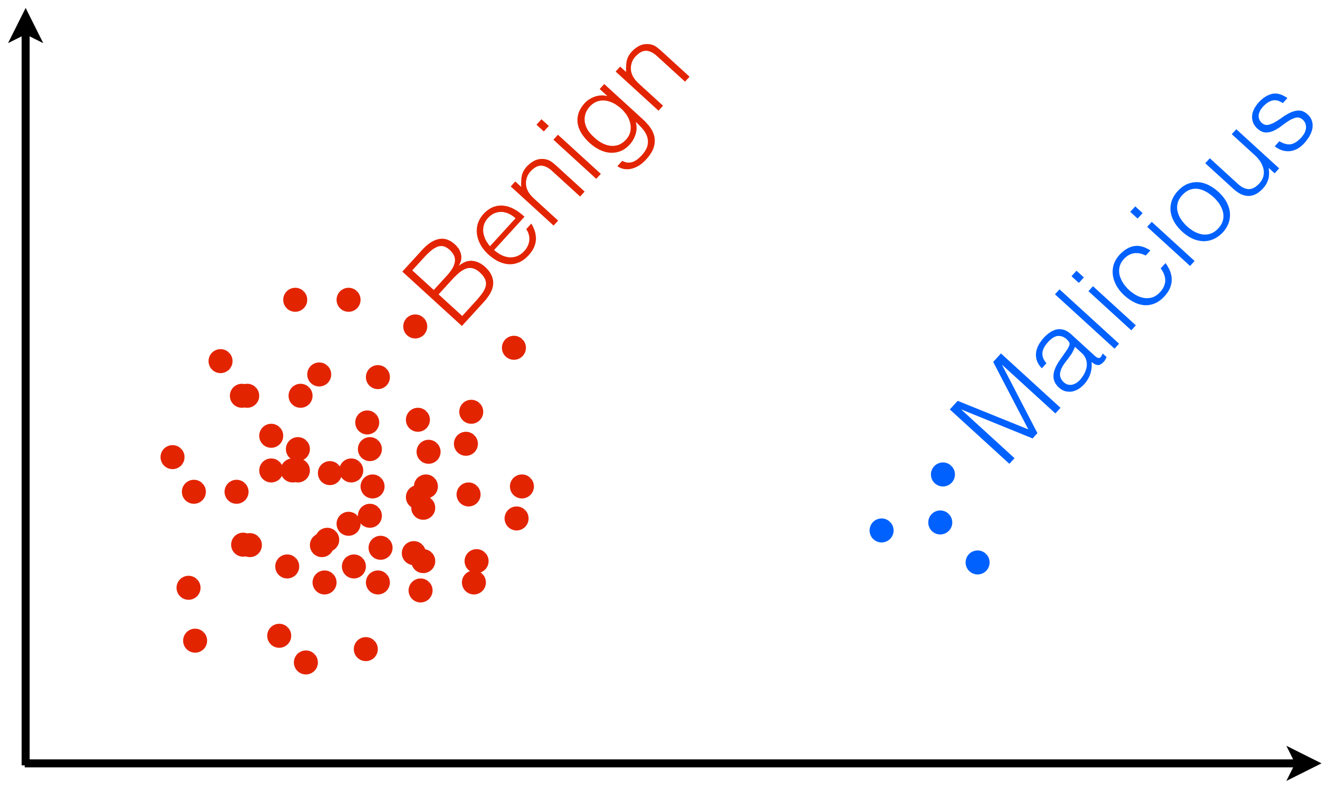
Focus on data “core” for guarantees

Focus on data “core” for guarantees

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

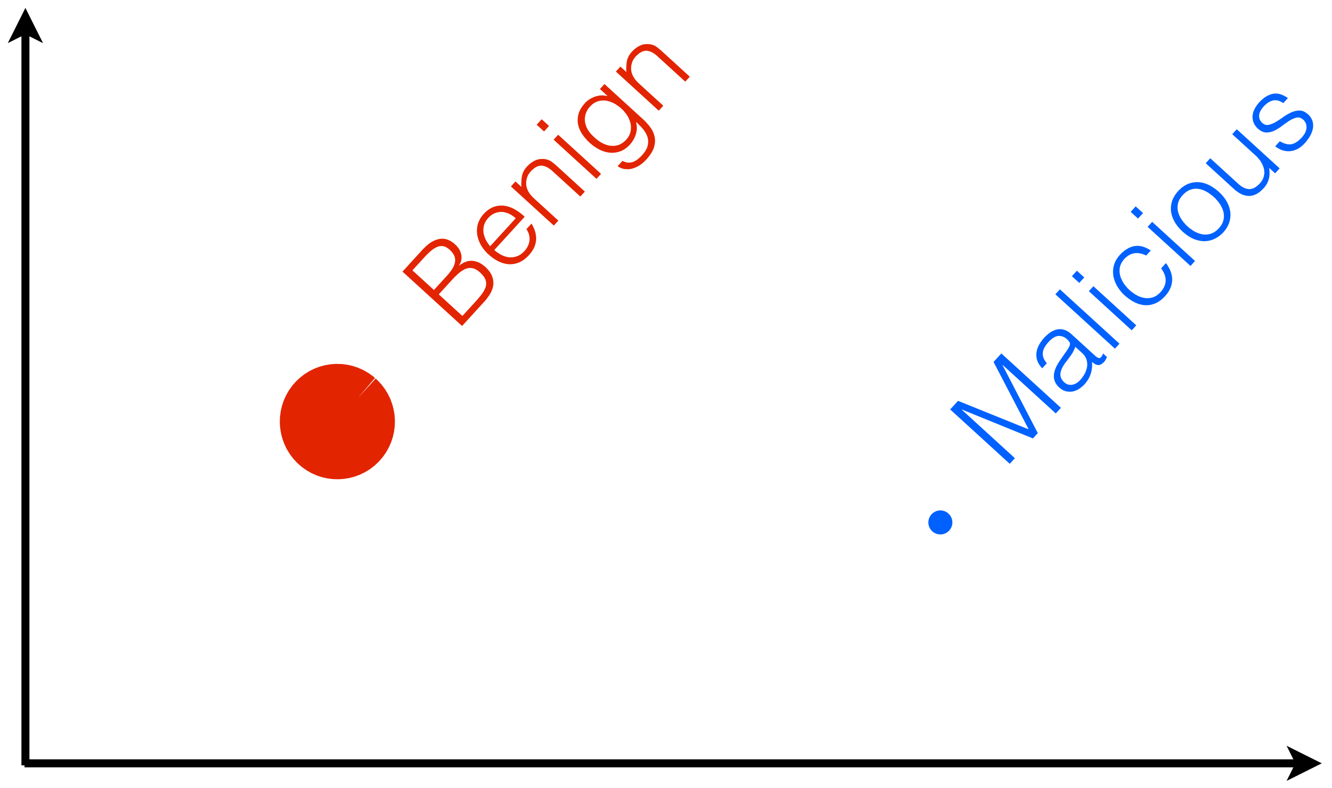
Focus on data “core” for guarantees

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



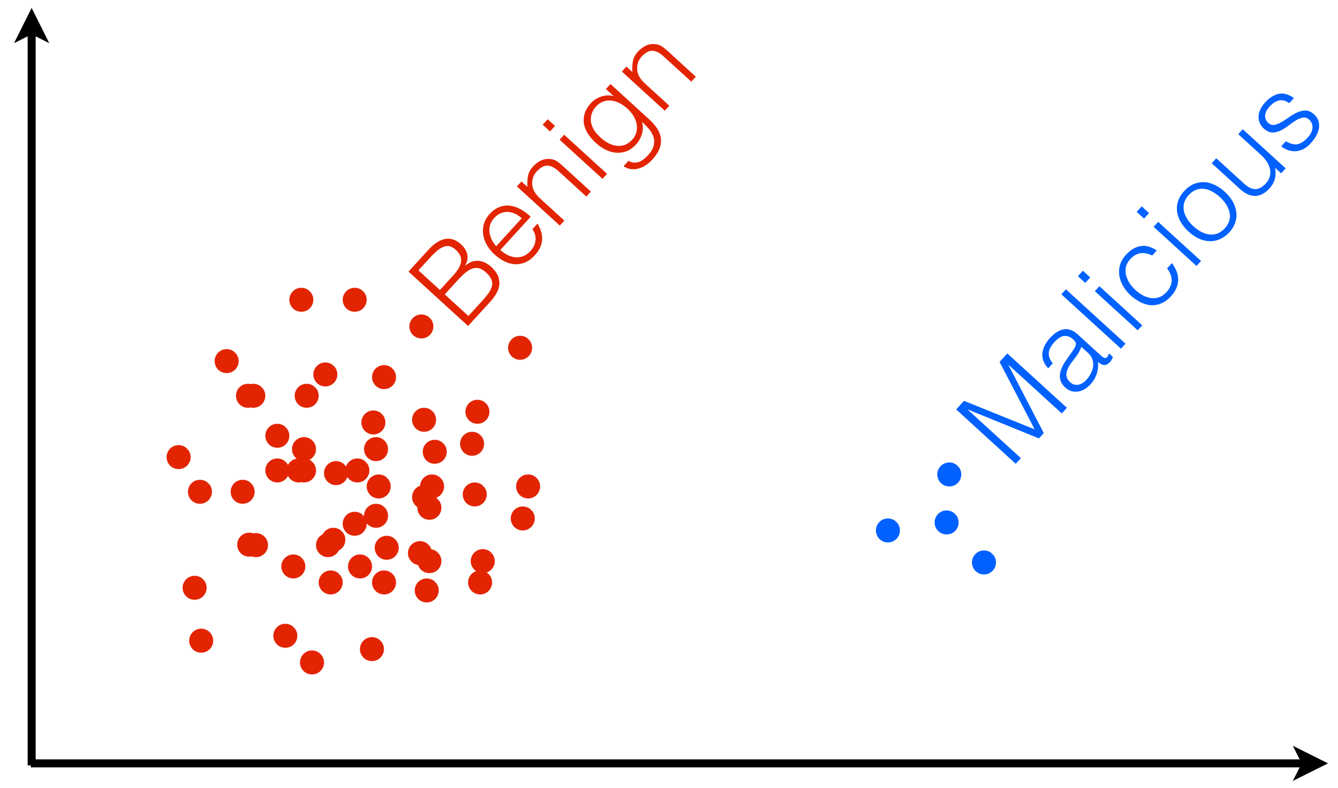
Focus on data “core” for guarantees

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



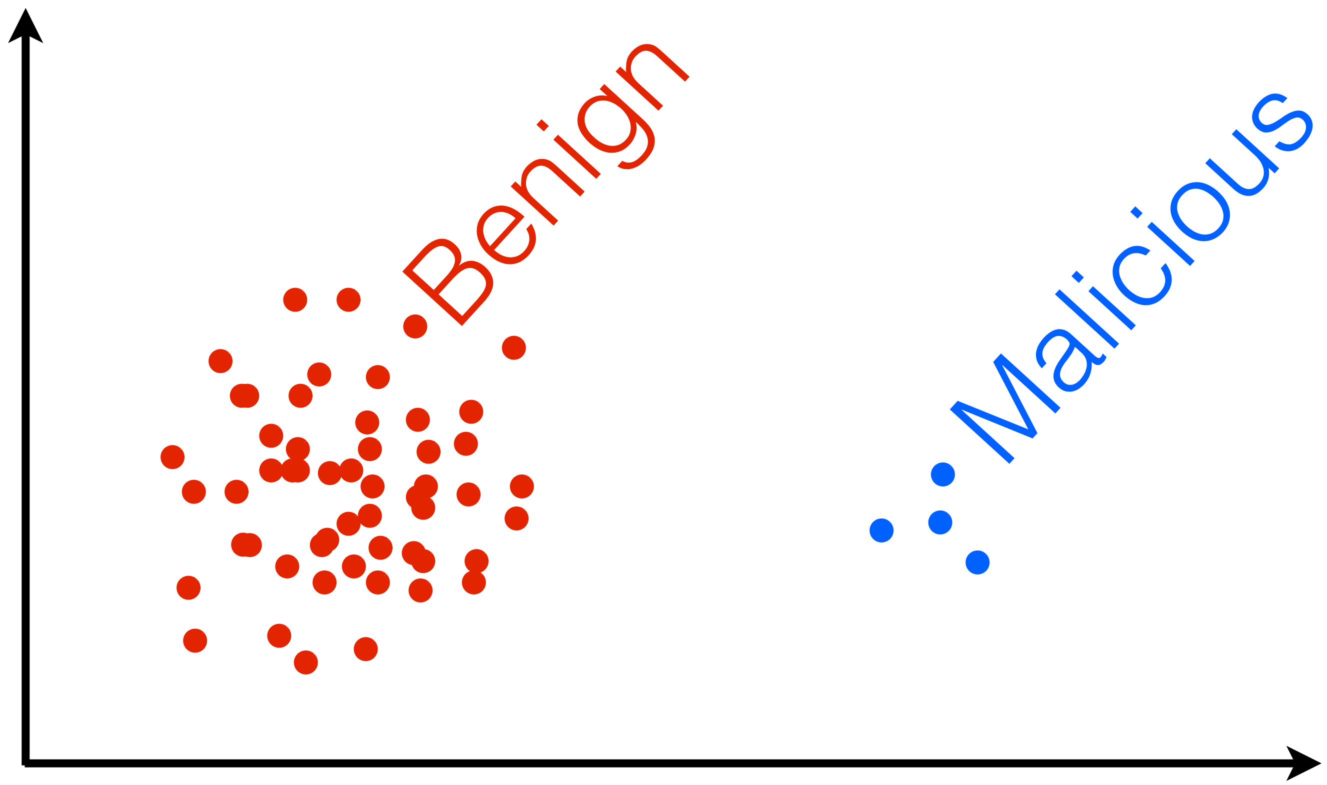
Focus on data “core” for guarantees

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



Focus on data "core" for guarantees

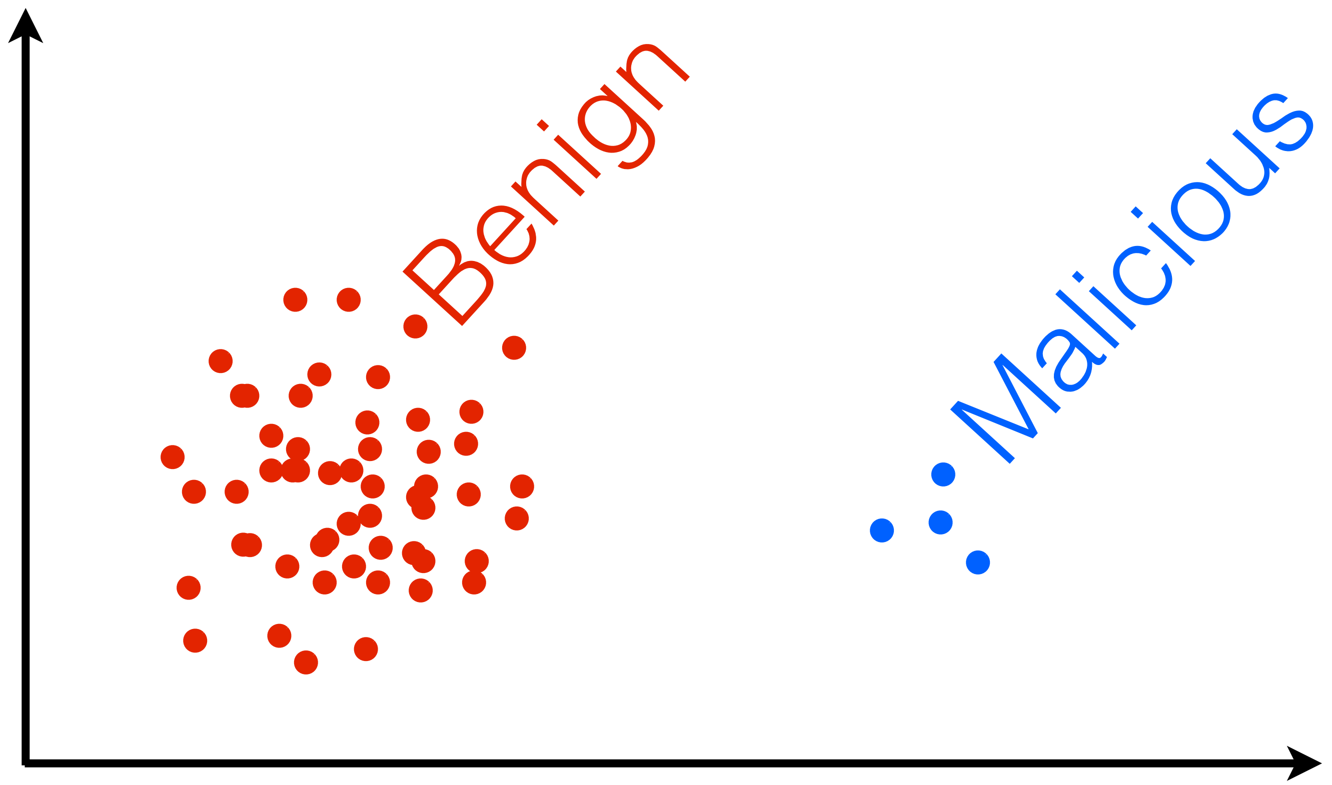
- Observe: redundancies can exist even if data isn't "tall"
- Coresets: pre-process data to get a smaller, weighted data set



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

Focus on data “core” for guarantees

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

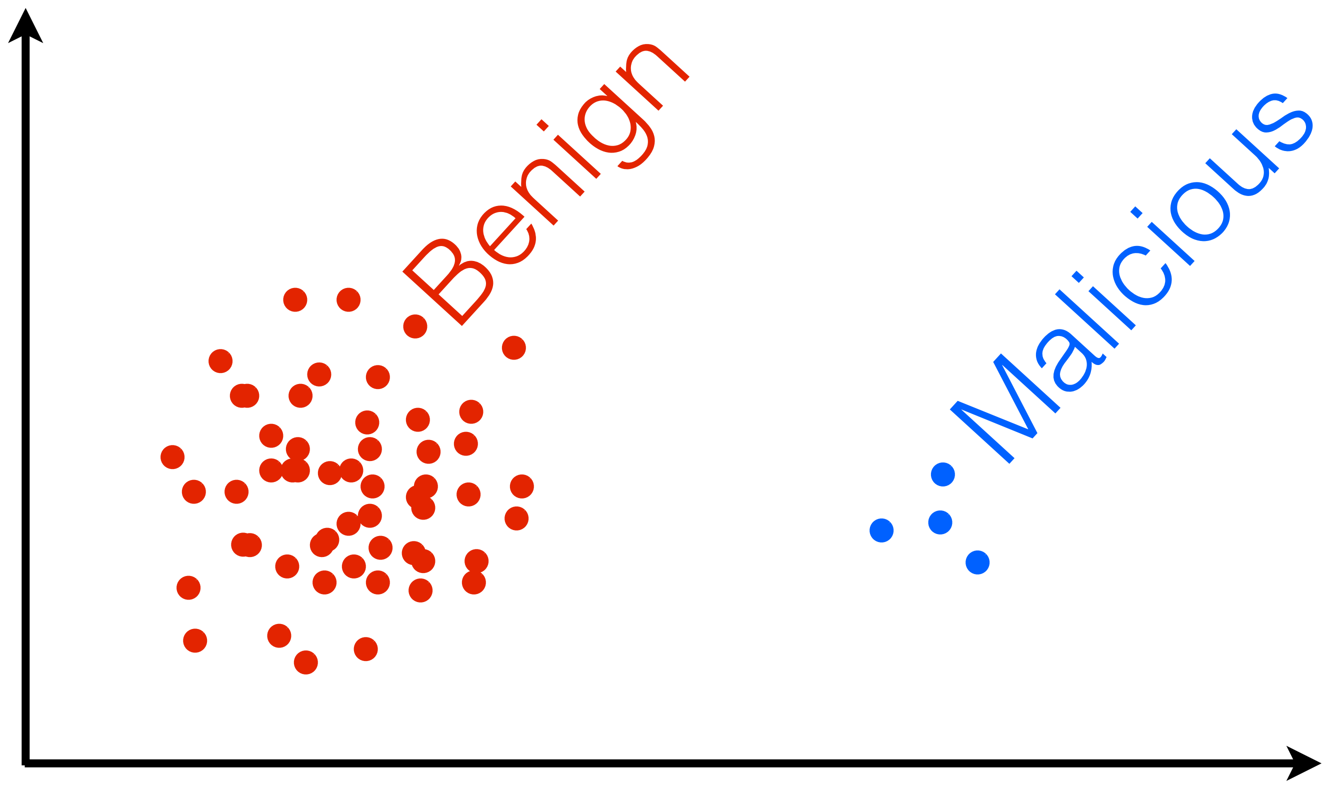


- Theoretical guarantees on quality

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

Focus on data “core” for guarantees

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

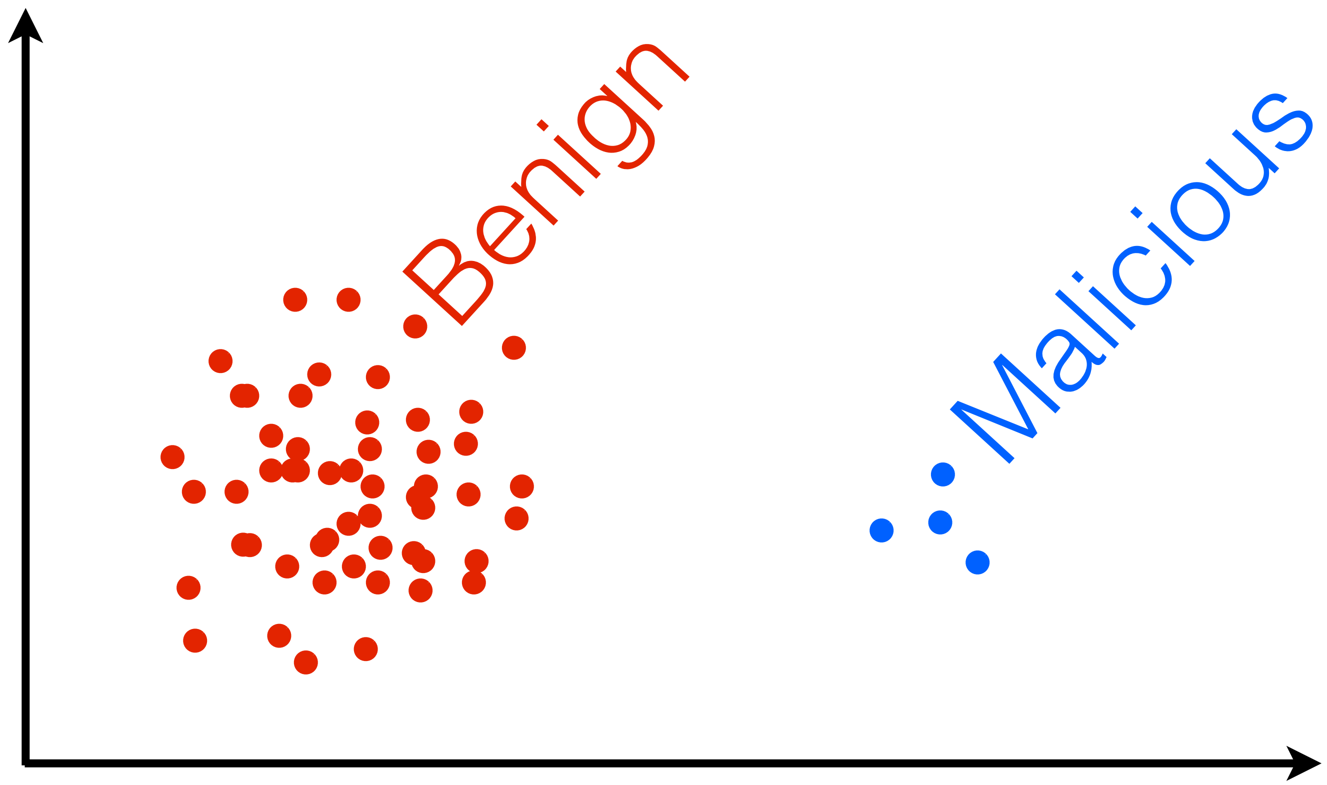


- Theoretical guarantees on quality
- How to develop **coresets for Bayes**?

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

Focus on data "core" for guarantees

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

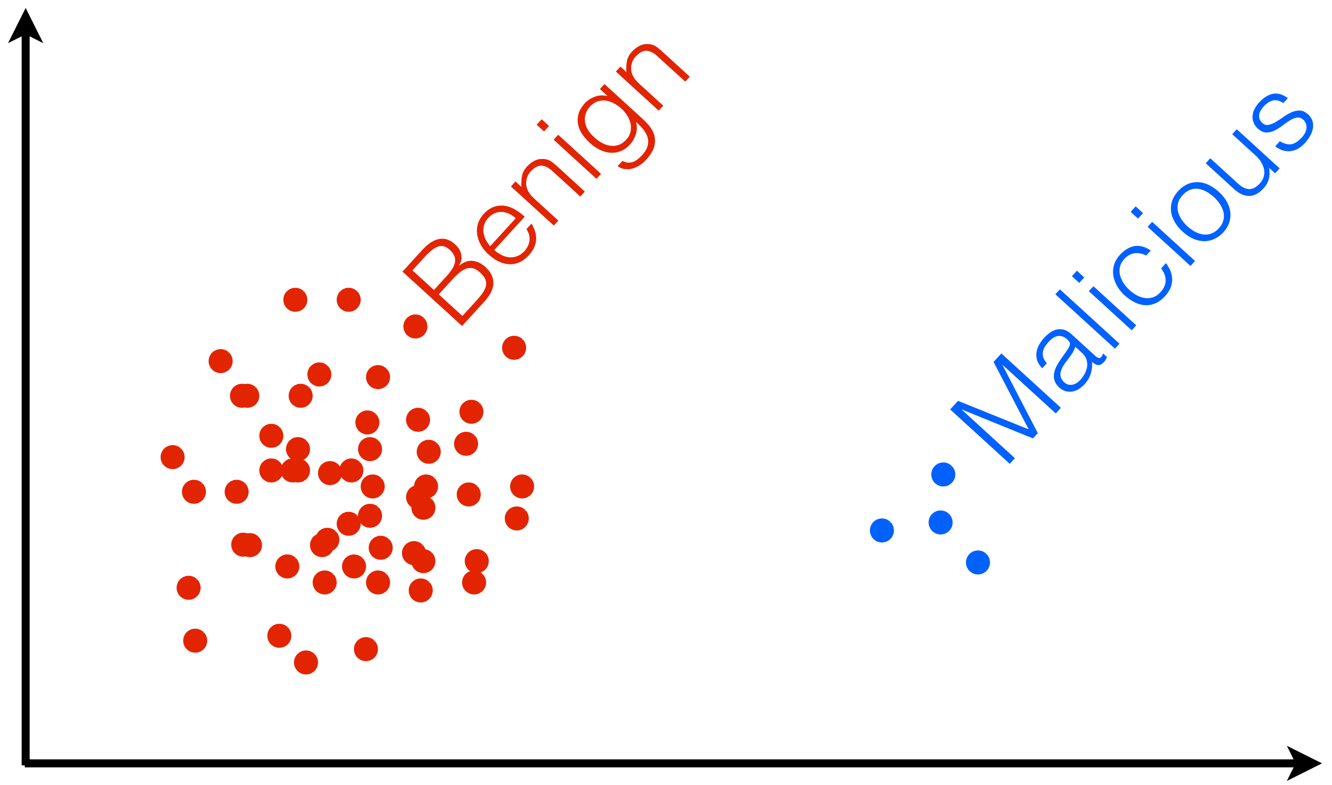


- Theoretical guarantees on quality
- How to develop **coresets for diverse tasks/geometries**?

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

Focus on data "core" for guarantees

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

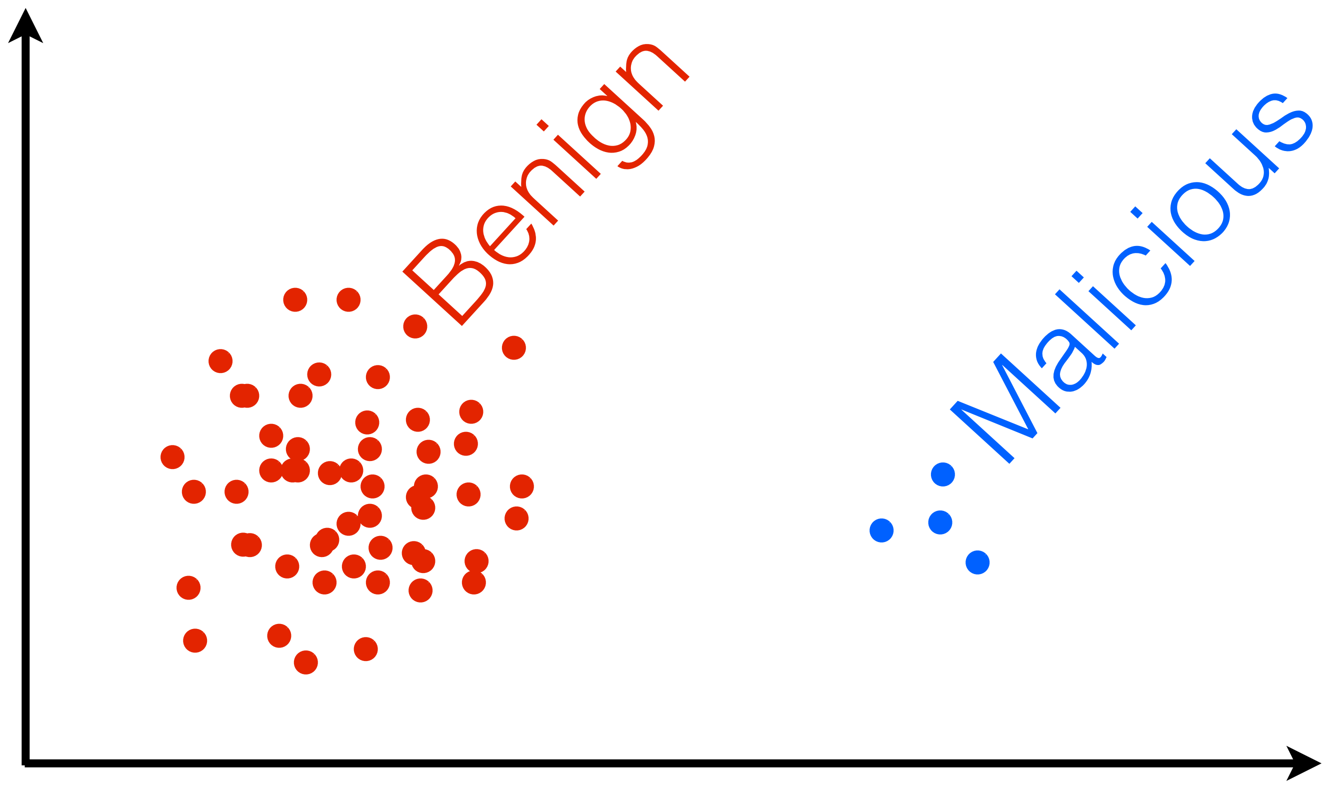


- Theoretical guarantees on quality
- How to develop **coresets for diverse tasks/geometries**?
- Previous heuristics: data squashing, big data GPs

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

Focus on data "core" for guarantees

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

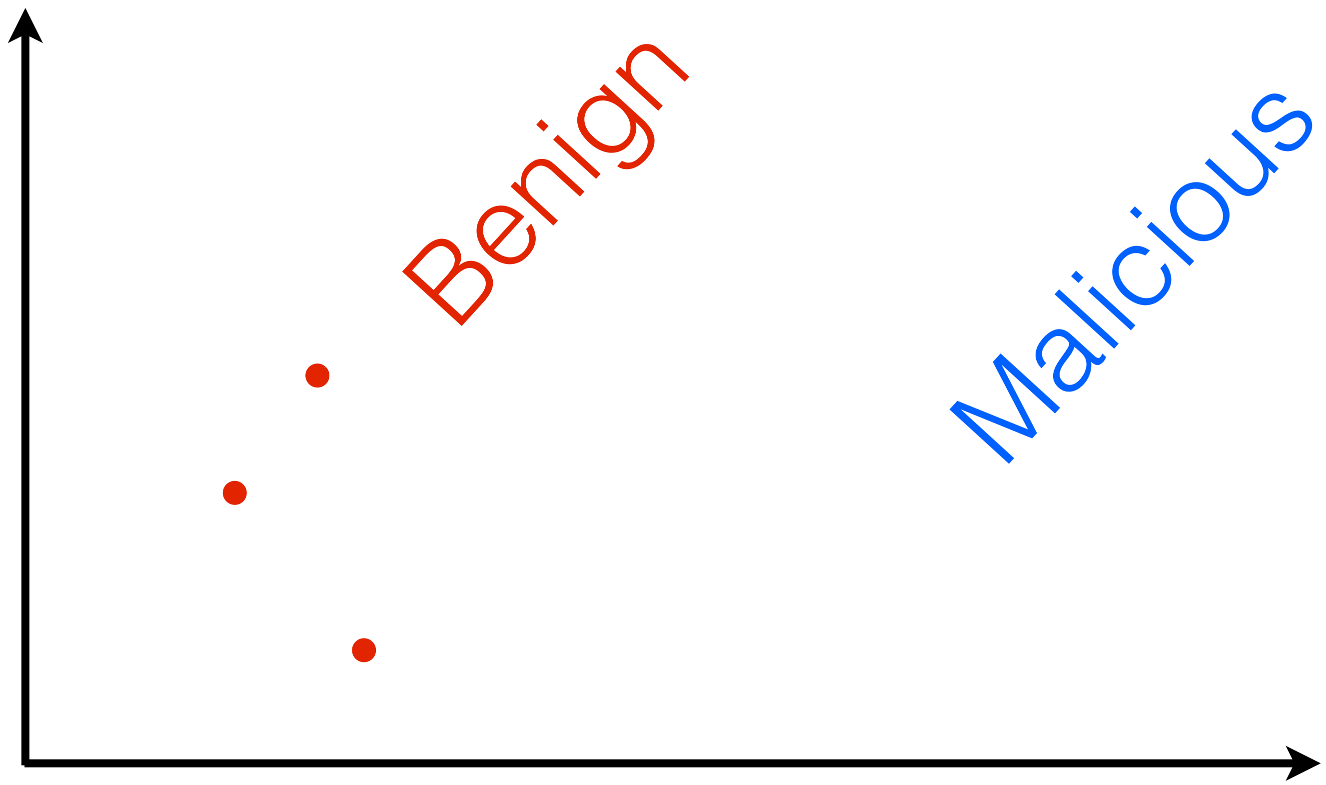


- Theoretical guarantees on quality
- How to develop **coresets for diverse tasks/geometries**?
- Previous heuristics: data squashing, big data GPs
- Compare to subsampling

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

Focus on data “core” for guarantees

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

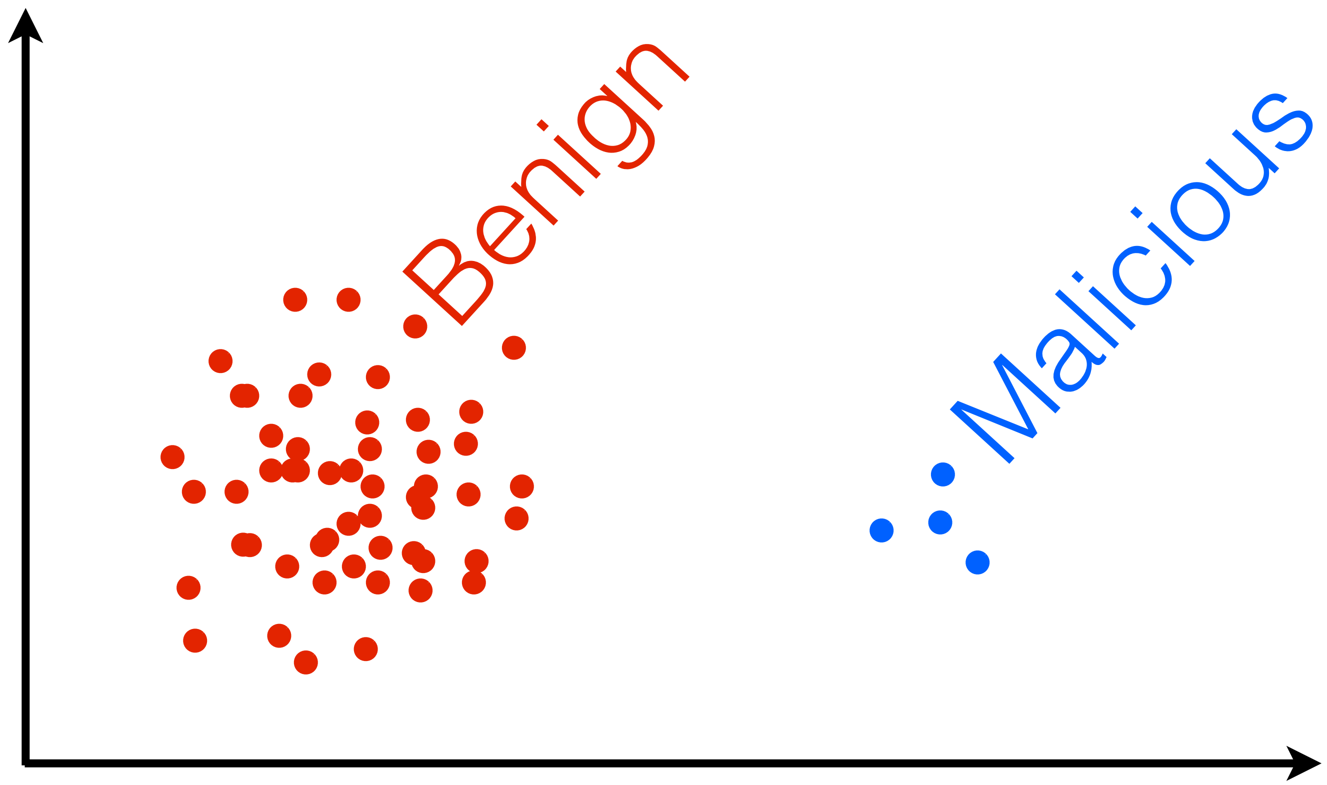


- Theoretical guarantees on quality
- How to develop **coresets for diverse tasks/geometries**?
- Previous heuristics: data squashing, big data GPs
- Compare to subsampling

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

Focus on data "core" for guarantees

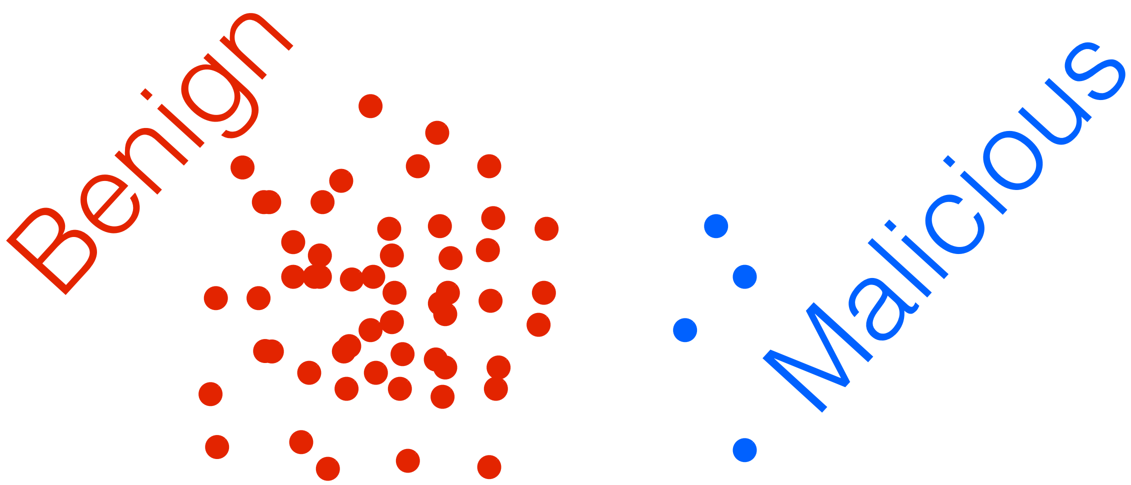
- Observe: redundancies can exist even if data isn't "tall"
- Coresets: pre-process data to get a smaller, weighted data set



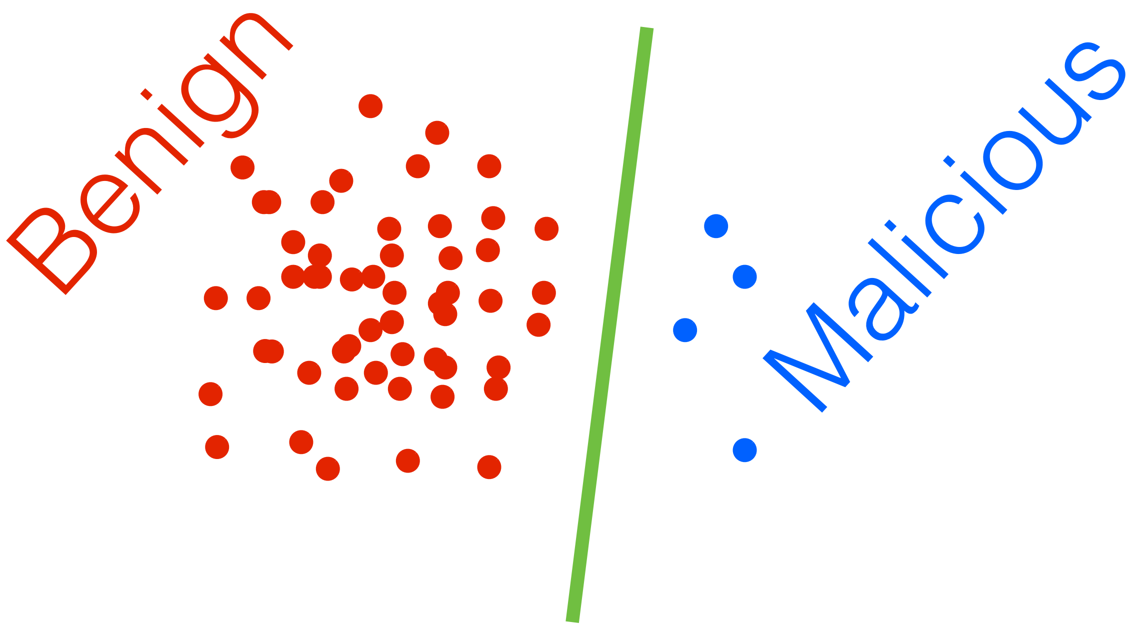
- Theoretical guarantees on quality
- How to develop **coresets for diverse tasks/geometries**?
- Previous heuristics: data squashing, big data GPs
- Compare to subsampling

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

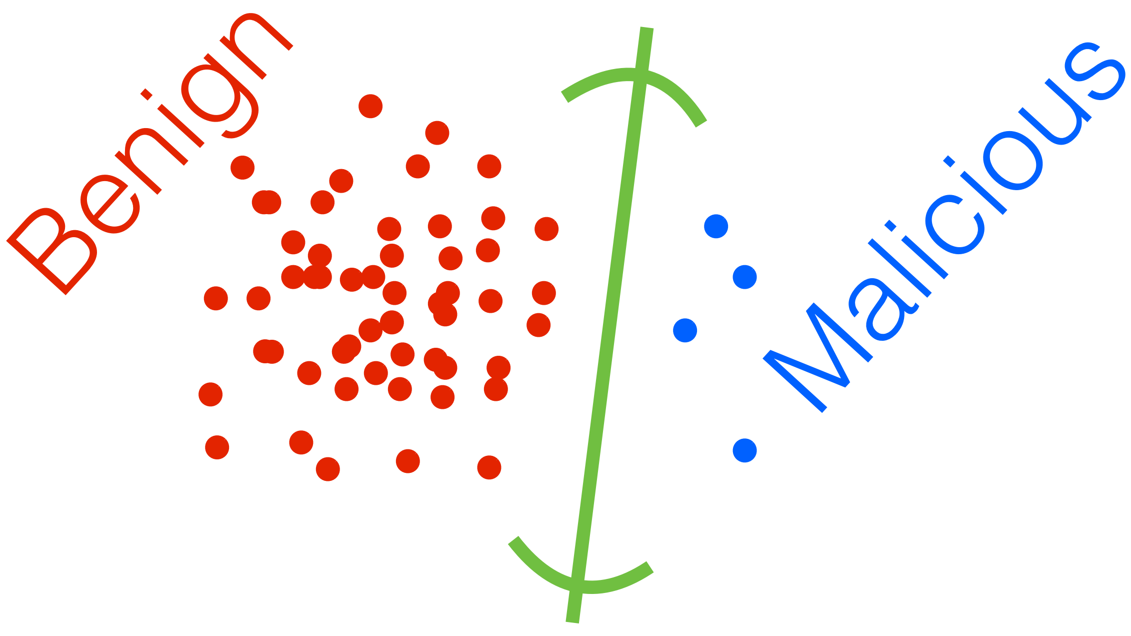
Uniform subsampling



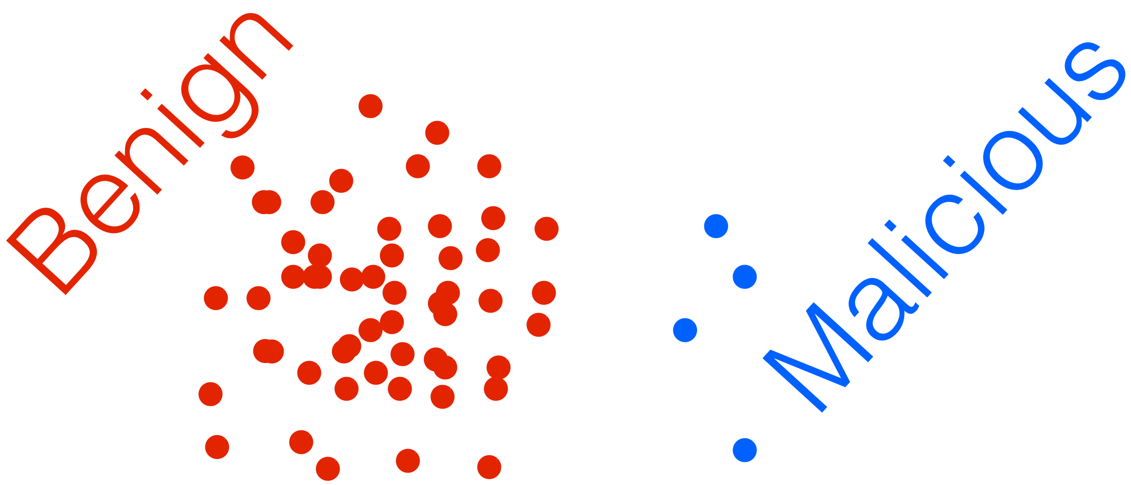
Uniform subsampling



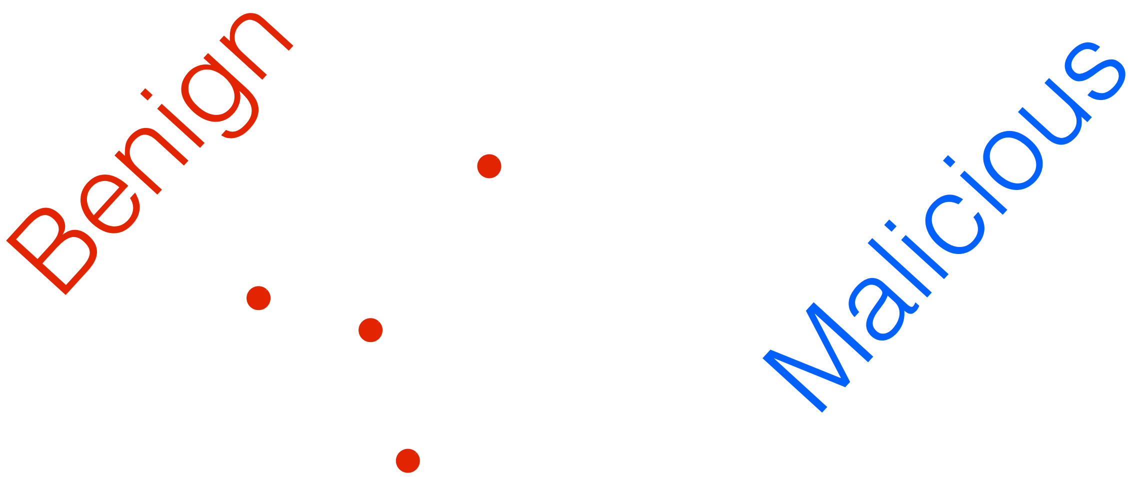
Uniform subsampling



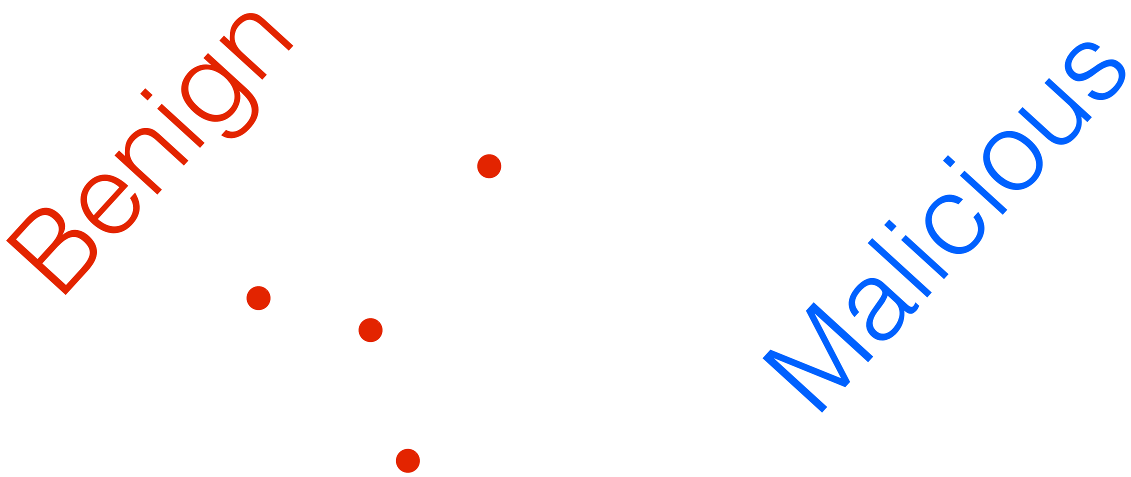
Uniform subsampling



Uniform subsampling

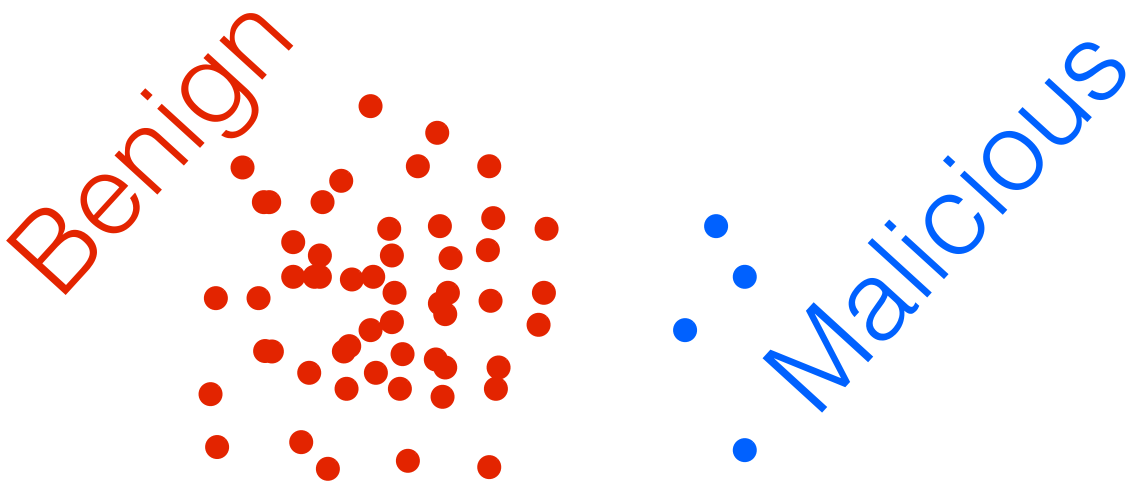


Uniform subsampling



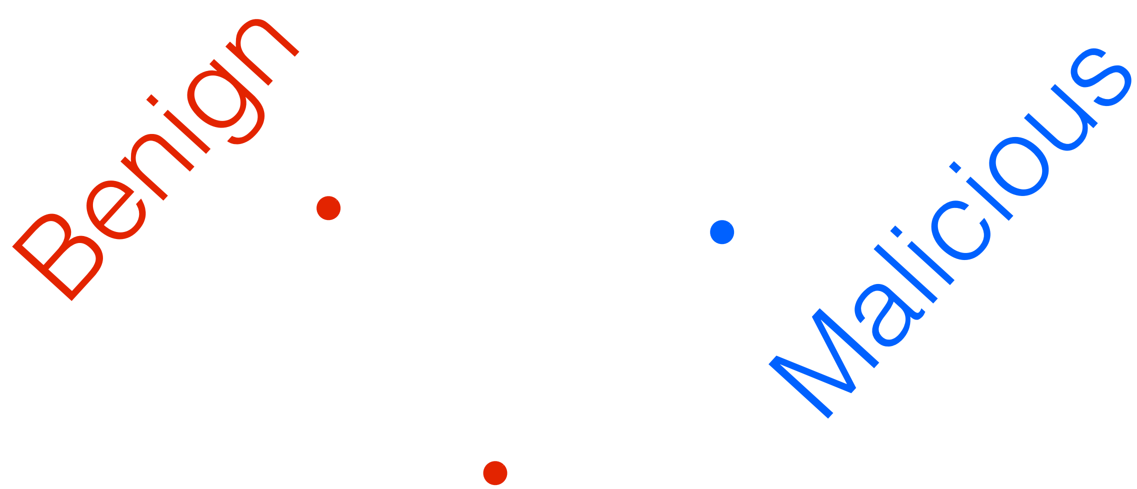
- Might miss important data

Uniform subsampling



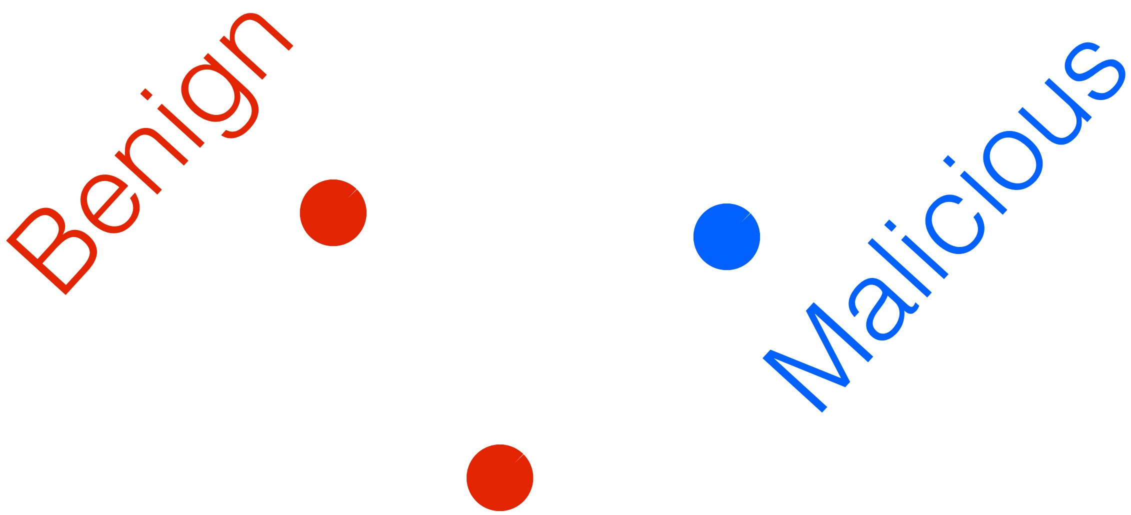
- Might miss important data

Uniform subsampling



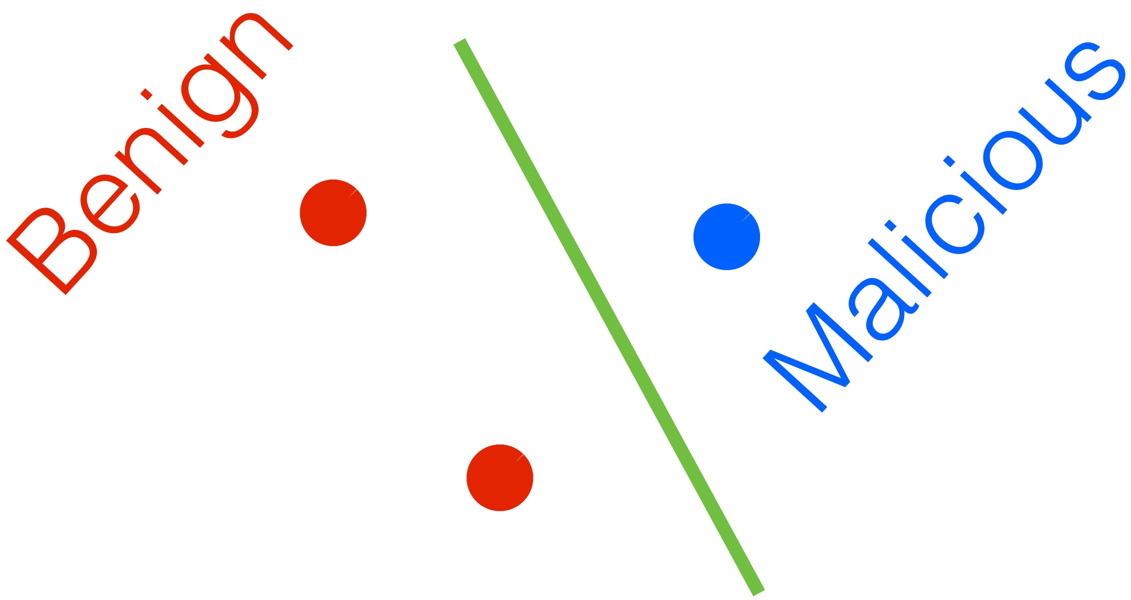
- Might miss important data

Uniform subsampling



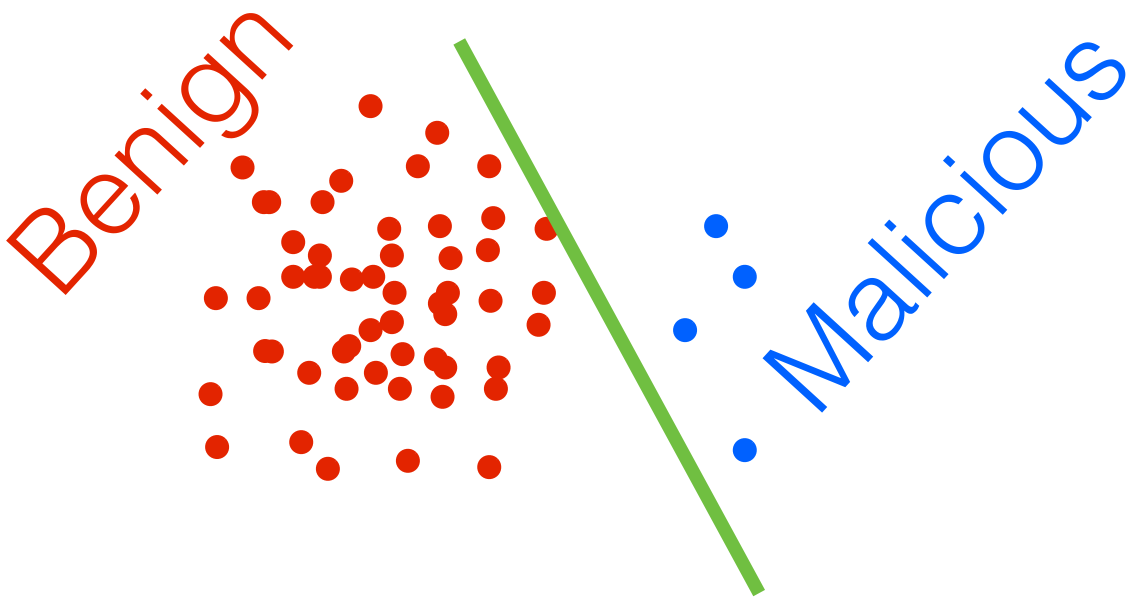
- Might miss important data

Uniform subsampling



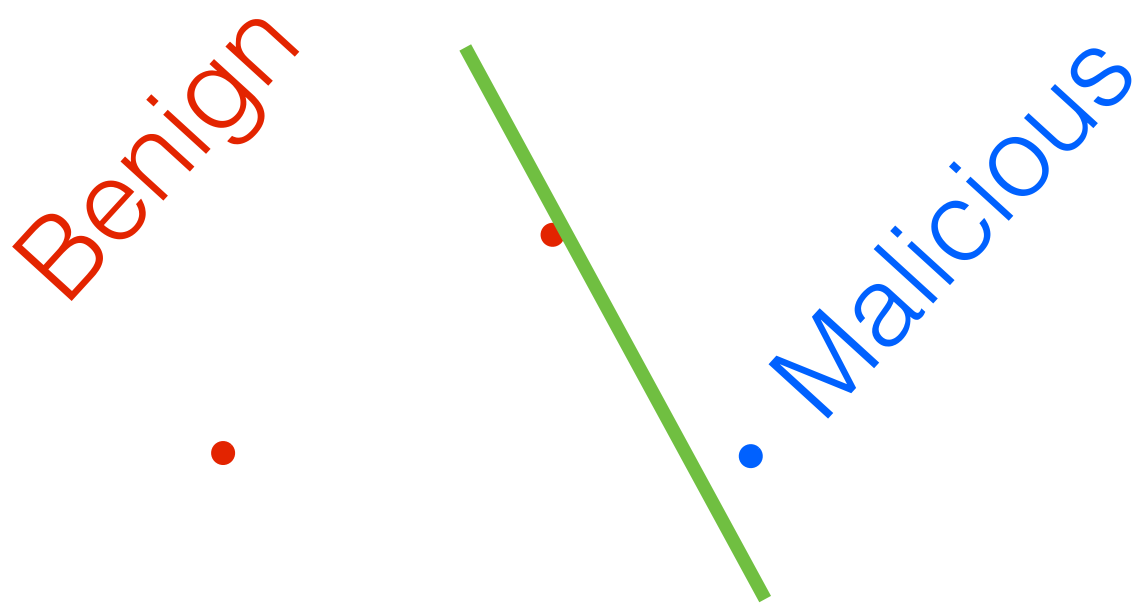
- Might miss important data

Uniform subsampling



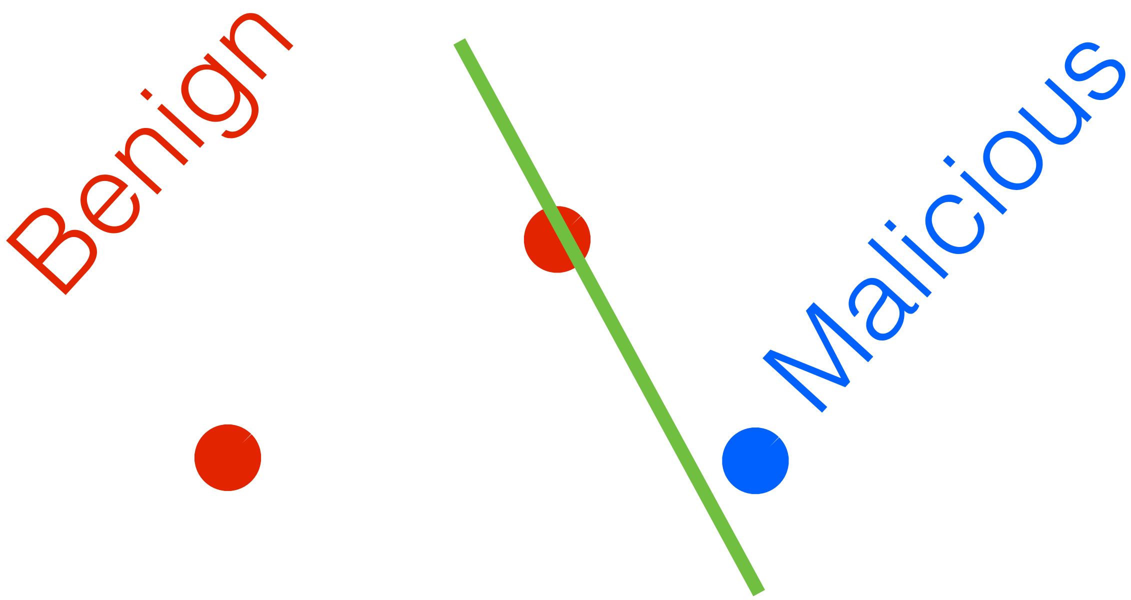
- Might miss important data

Uniform subsampling



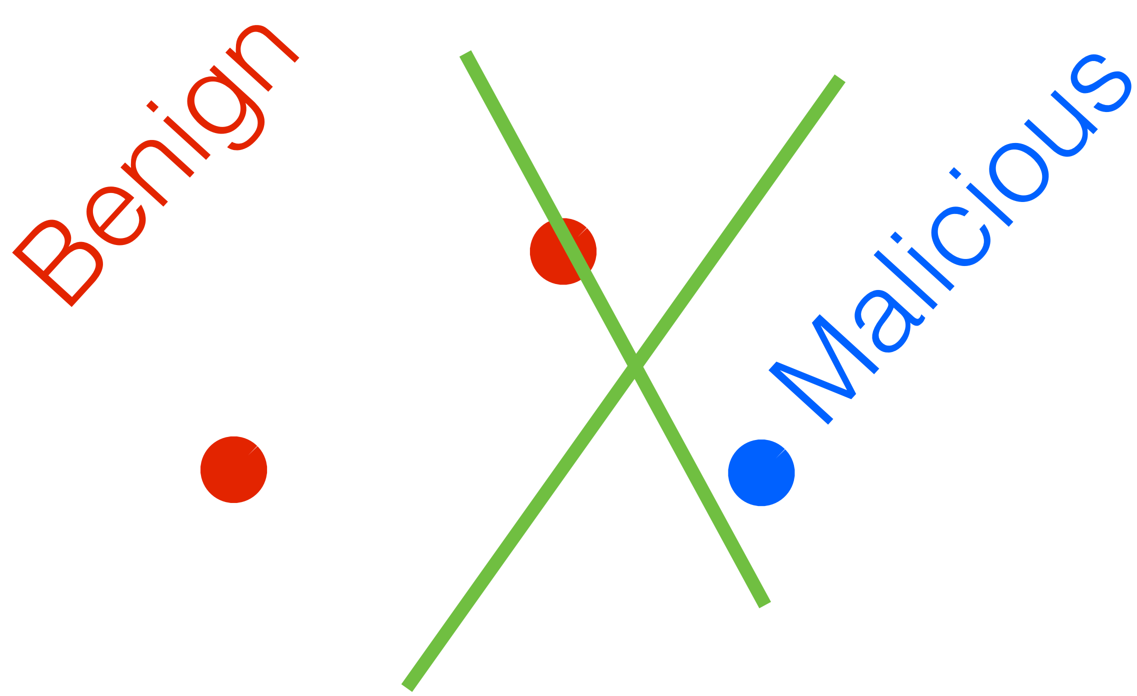
- Might miss important data

Uniform subsampling



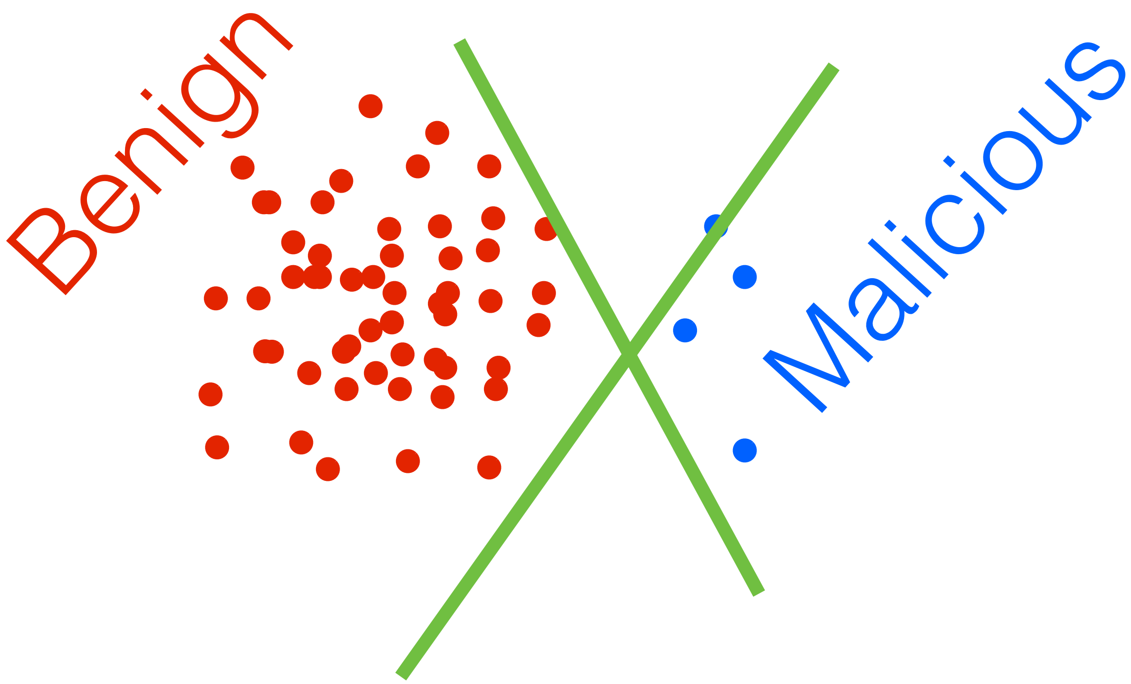
- Might miss important data

Uniform subsampling



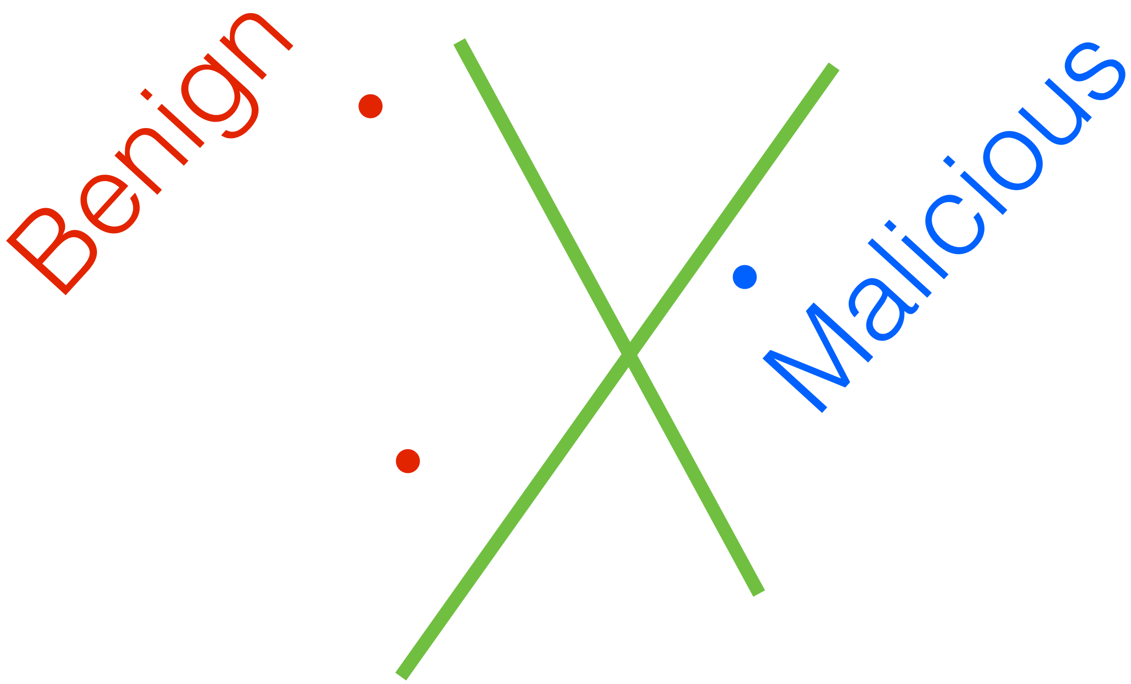
- Might miss important data

Uniform subsampling



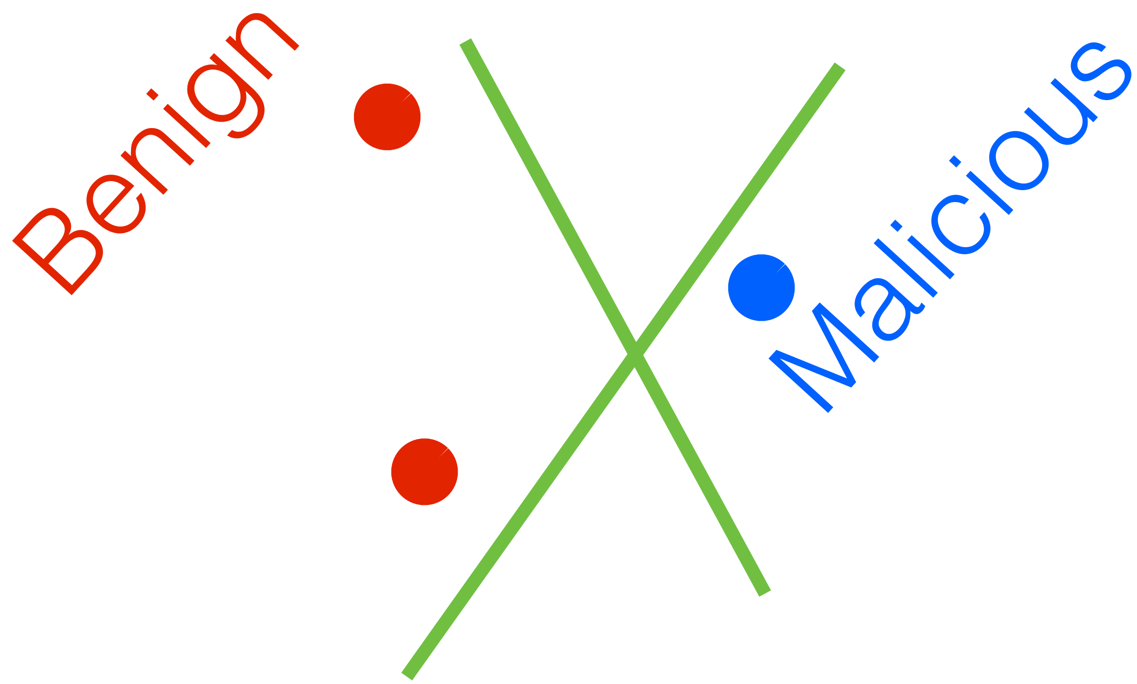
- Might miss important data

Uniform subsampling



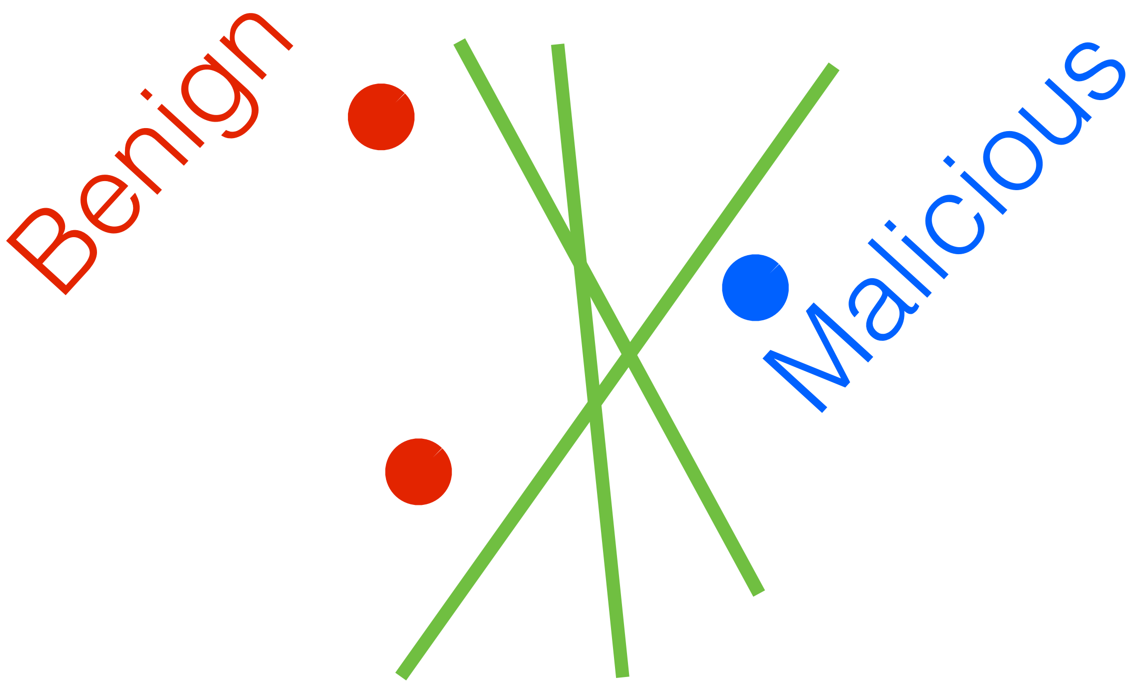
- Might miss important data

Uniform subsampling



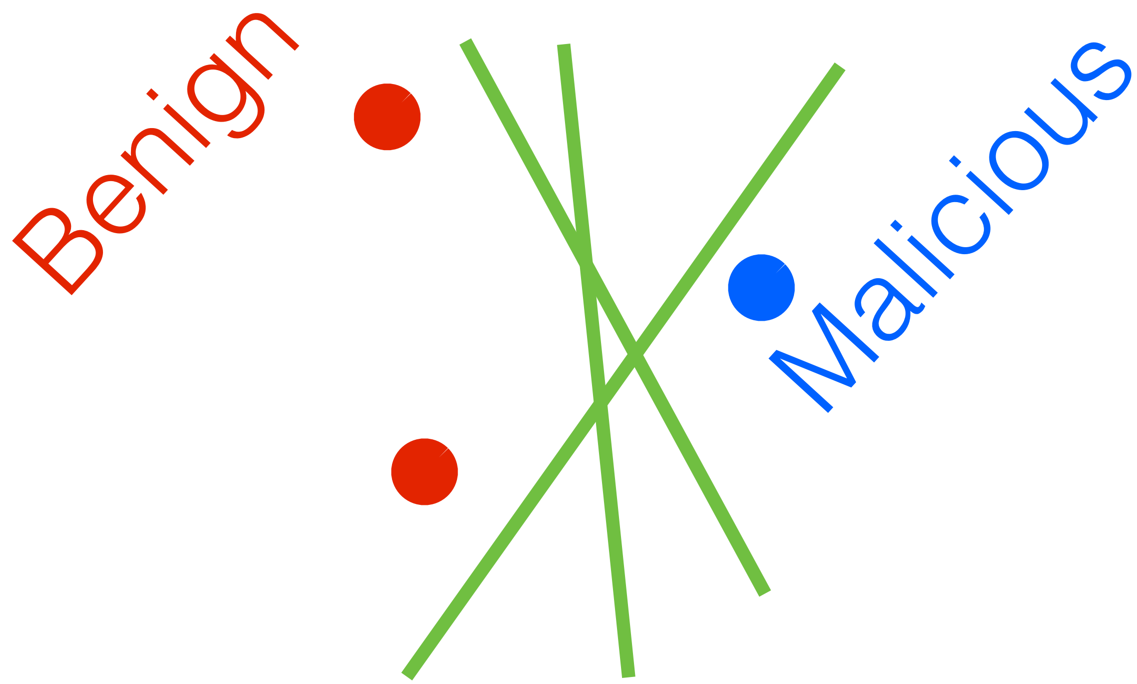
- Might miss important data

Uniform subsampling



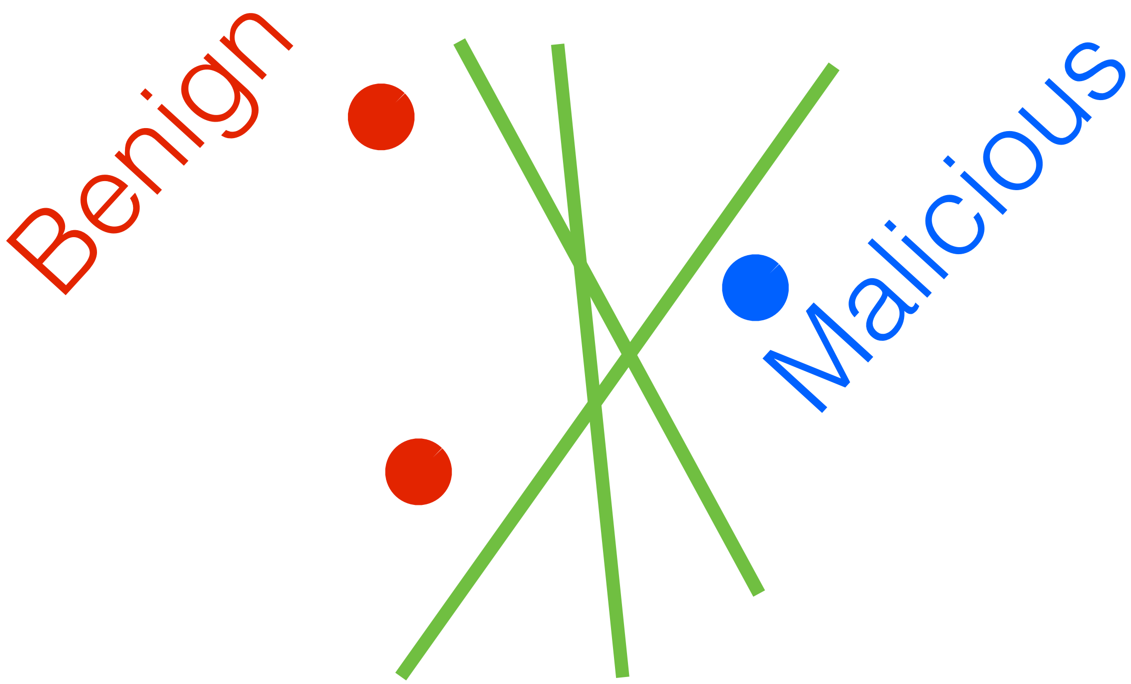
- Might miss important data

Uniform subsampling

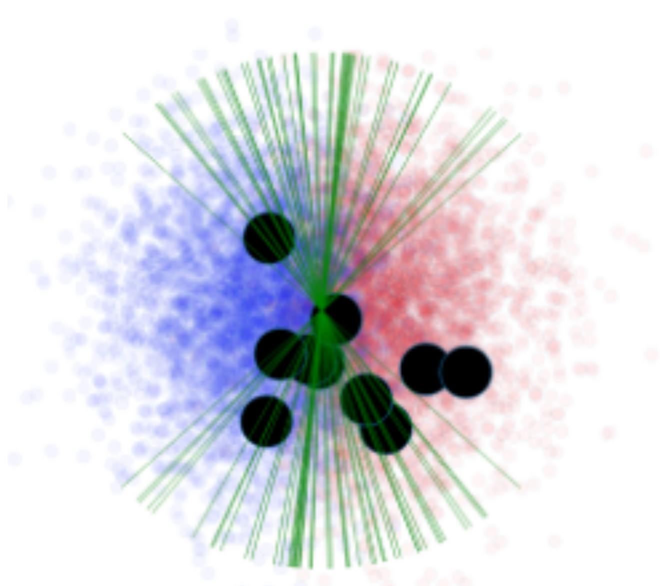


- Might miss important data
- Noisy estimates

Uniform subsampling



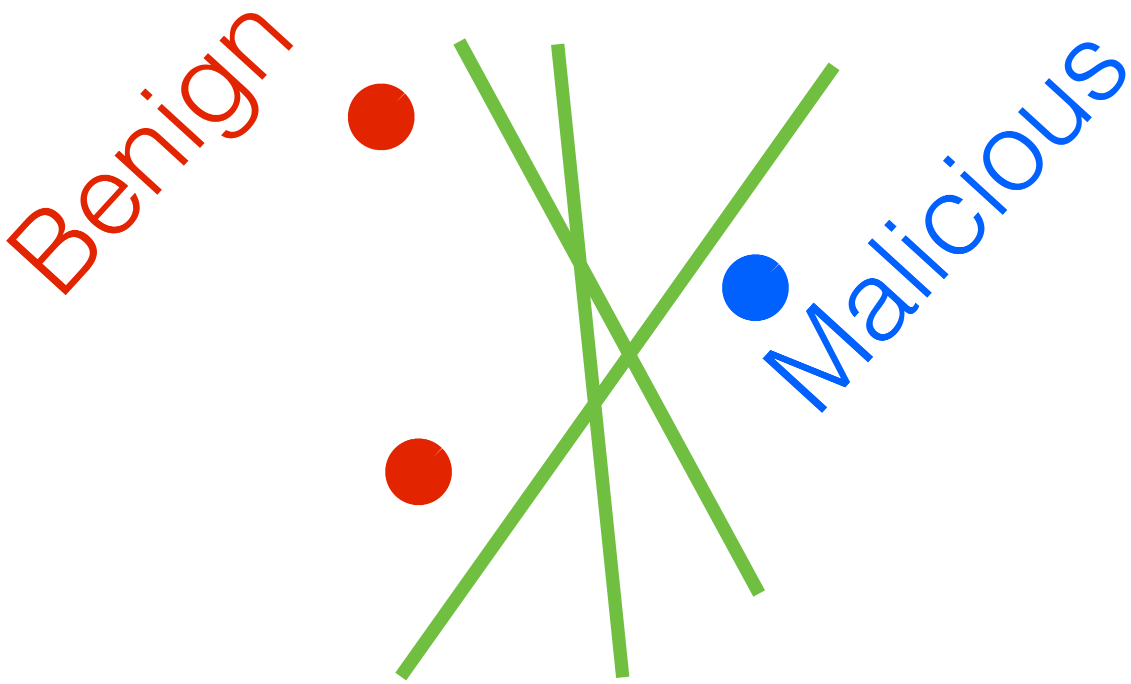
- Might miss important data
- Noisy estimates



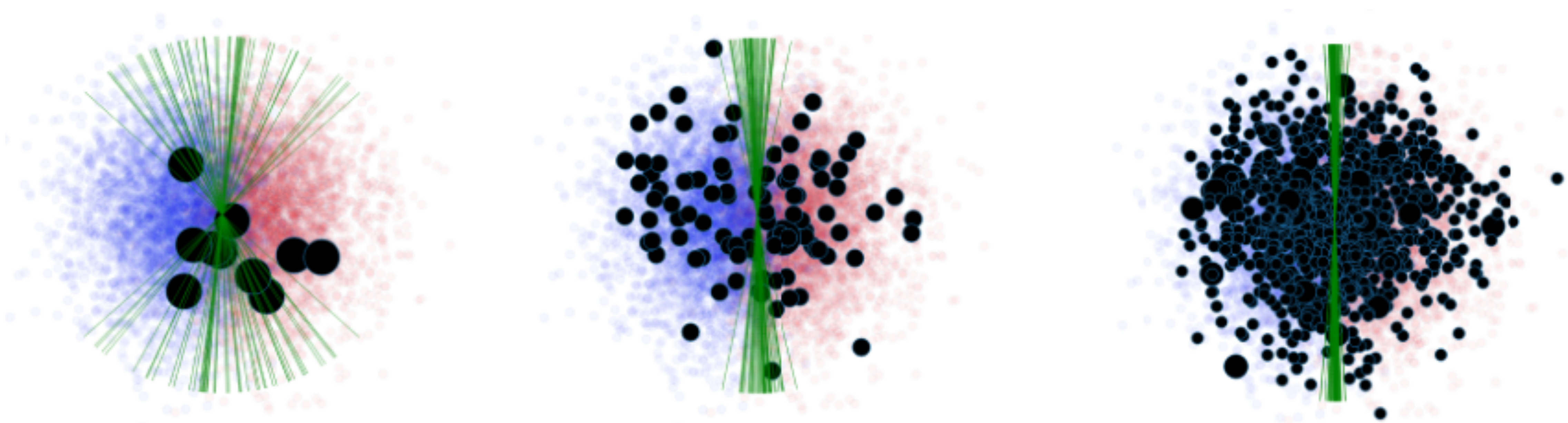
$M = 10$

[Campbell, Broderick 2018, 2019]

Uniform subsampling



- Might miss important data
- Noisy estimates

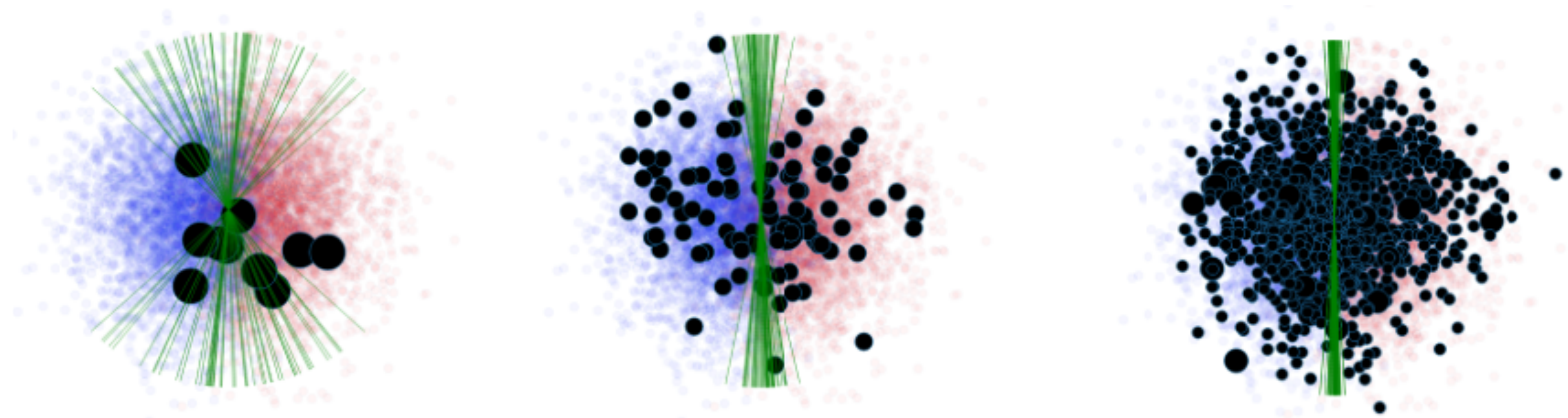


$M = 10$ $M = 100$ $M = 1000$

[Campbell, Broderick 2018, 2019]

Data summarization alternatives

Uniform subsampling



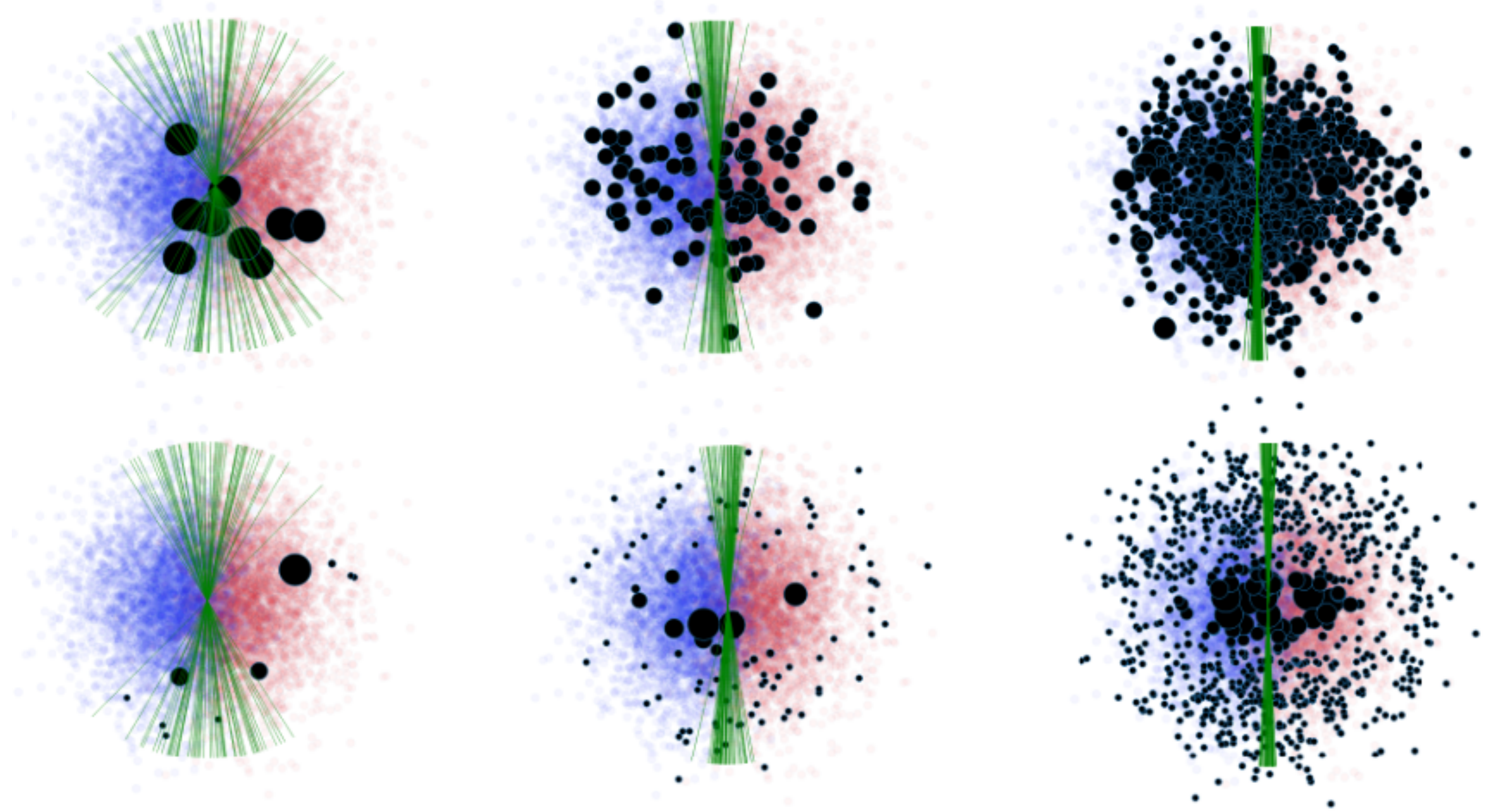
$M = 10$ $M = 100$ $M = 1000$

[Campbell, Broderick 2018, 2019]

Data summarization alternatives

Uniform subsampling

Importance sampling



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

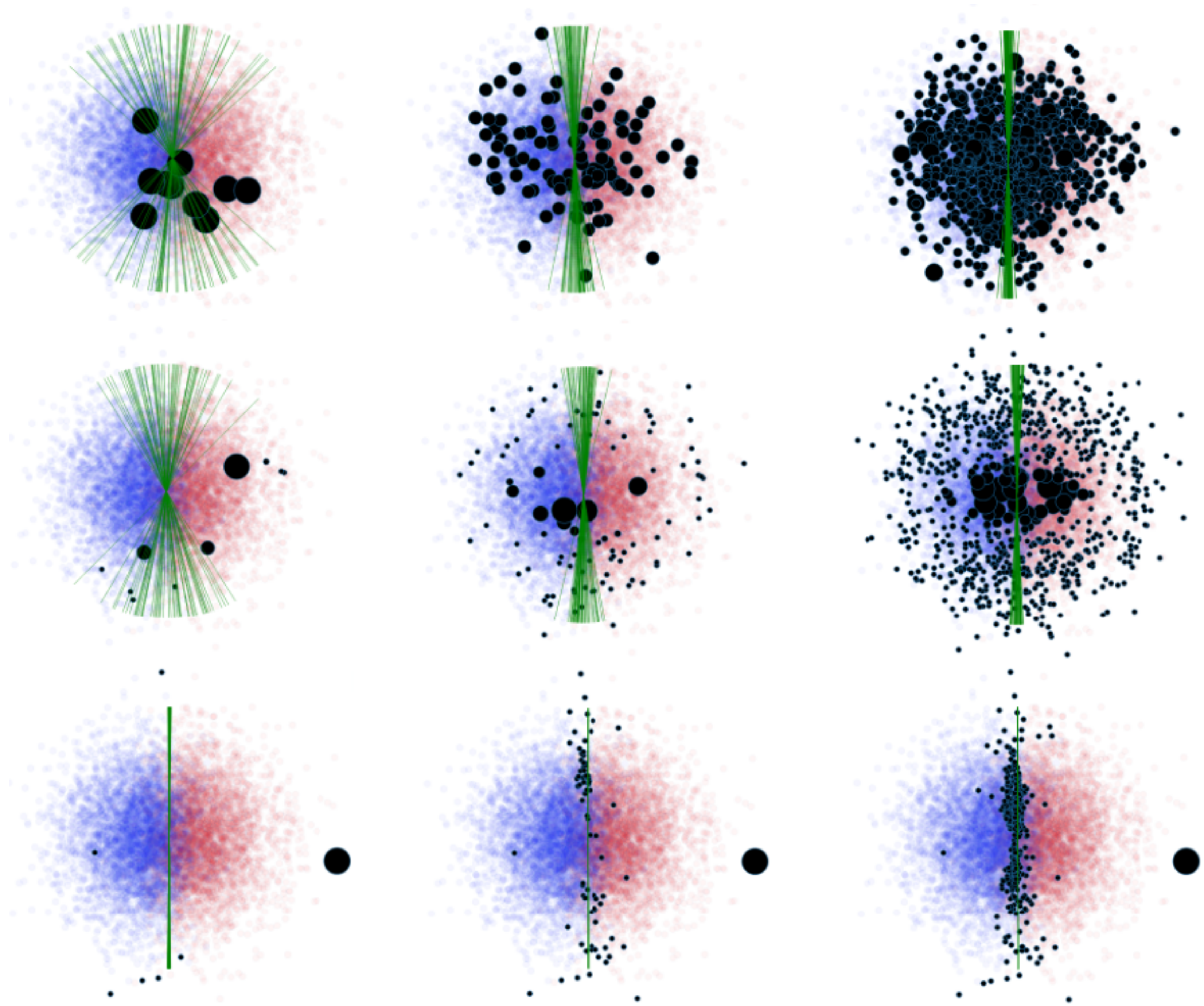
[Campbell, Broderick 2018, 2019]

Data summarization alternatives

Uniform subsampling

Importance sampling

Bayesian/Hilbert coresets



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

[Campbell, Broderick 2018, 2019]

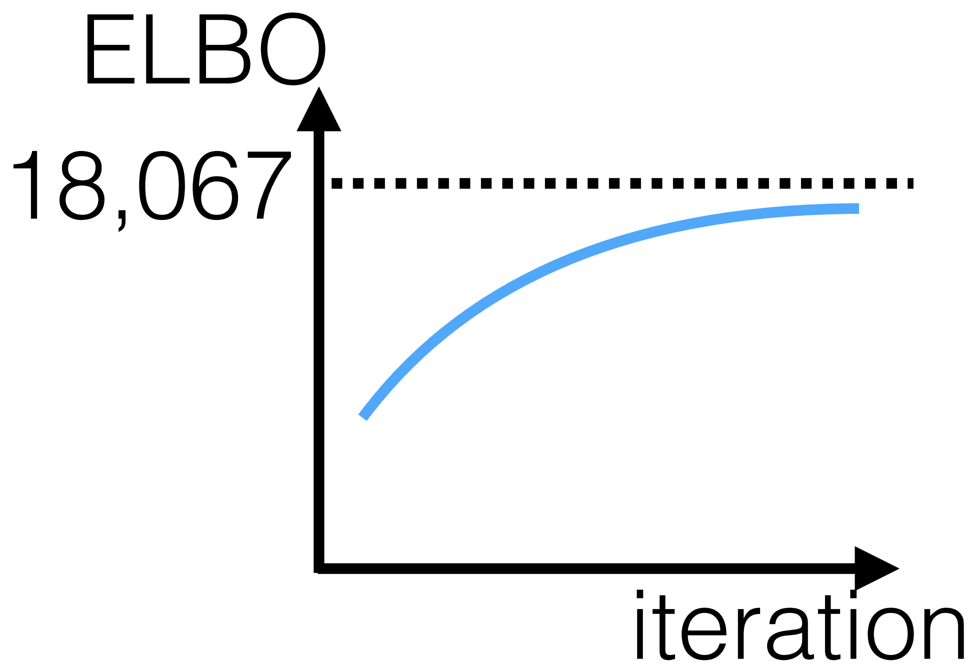
Reliable diagnostics

Reliable diagnostics

- ELBO or KL alone isn't enough

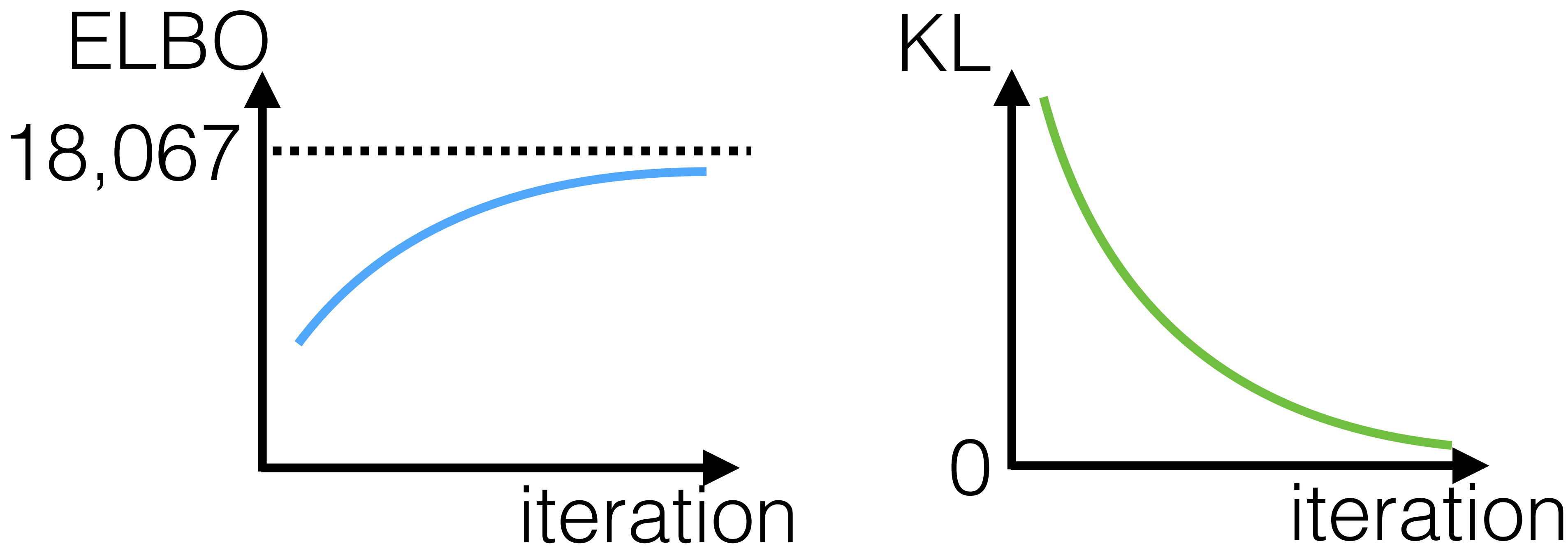
Reliable diagnostics

- ELBO or KL alone isn't enough



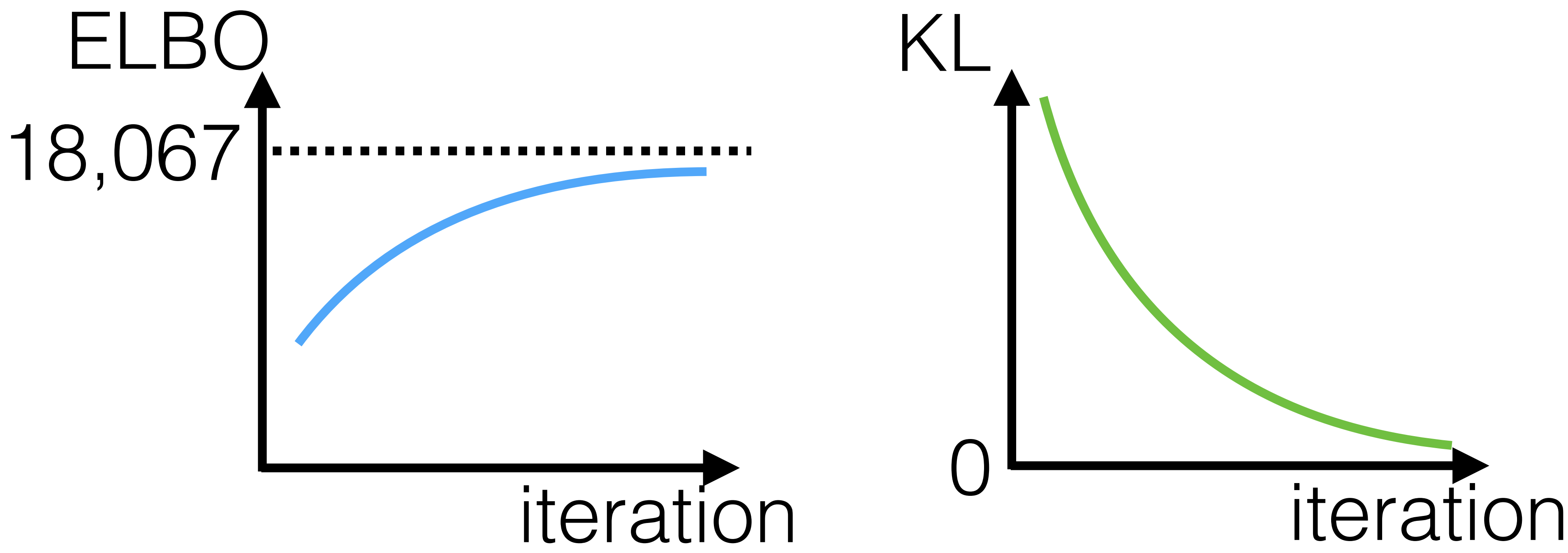
Reliable diagnostics

- ELBO or KL alone isn't enough



Reliable diagnostics

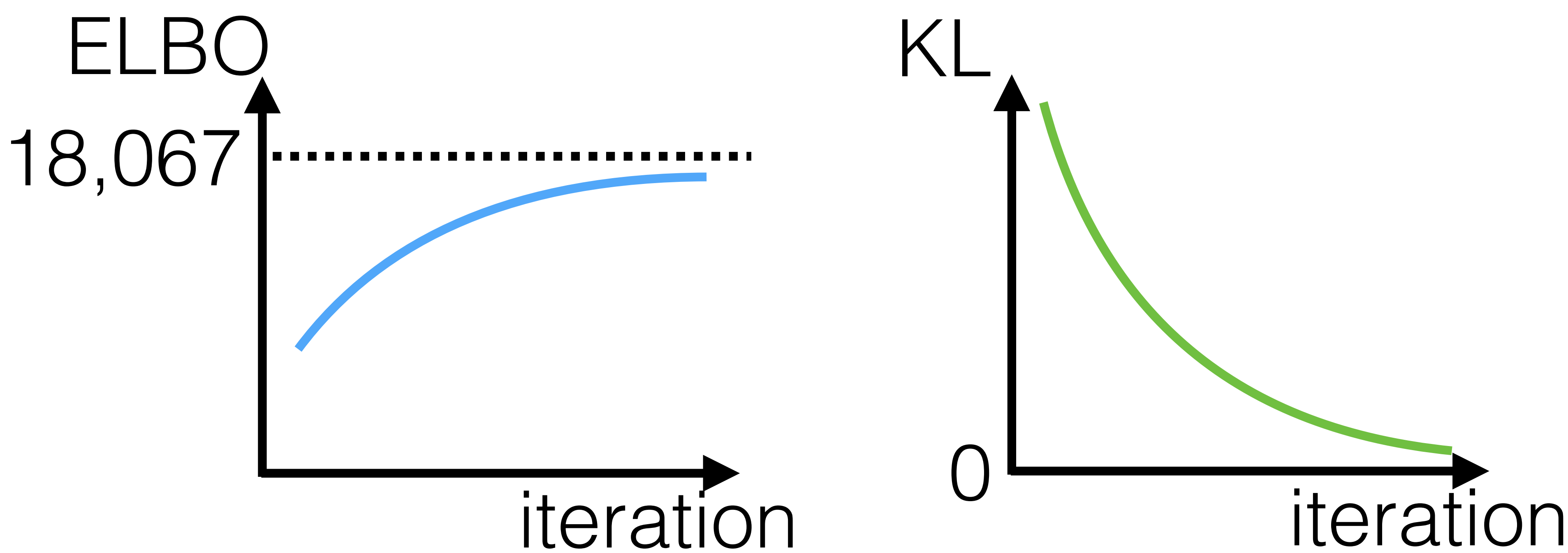
- ELBO or KL alone isn't enough



- Instead: easy-to-compute bound on Wasserstein [Huggins, Kasprzak, Campbell, Broderick, 2020]
  - Wasserstein bounds error in posterior mean and variance

Reliable diagnostics

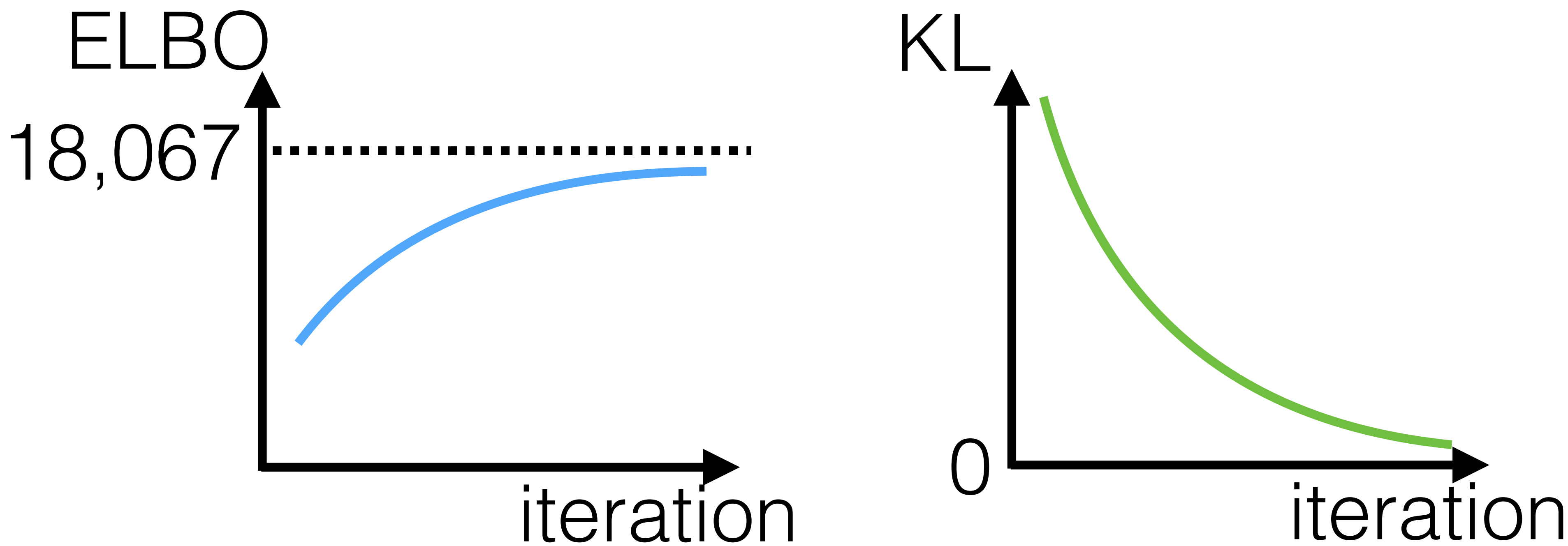
- ELBO or KL alone isn't enough



- Instead: easy-to-compute bound on Wasserstein [Huggins, Kasprzak, Campbell, Broderick, 2020]
  - Wasserstein bounds error in posterior mean and variance
- Part of a validated workflow for VB [Huggins, Kasprzak, Campbell, Broderick, 2020]

Reliable diagnostics

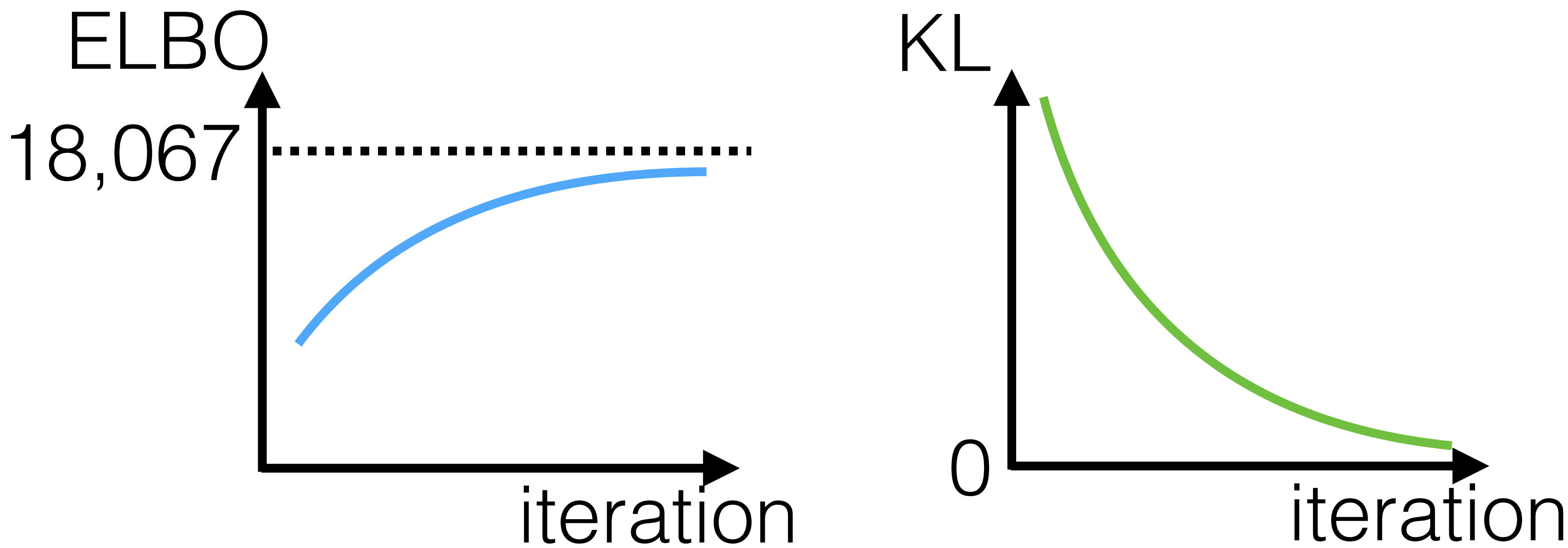
- ELBO or KL alone isn't enough



- Instead: easy-to-compute bound on Wasserstein [Huggins, Kasprzak, Campbell, Broderick, 2020]
  - Wasserstein bounds error in posterior mean and variance
- Part of a validated workflow for VB [Huggins, Kasprzak, Campbell, Broderick, 2020]
  - Builds on e.g. [Dieng et al 2017; Yao et al 2018]

Reliable diagnostics

- ELBO or KL alone isn't enough

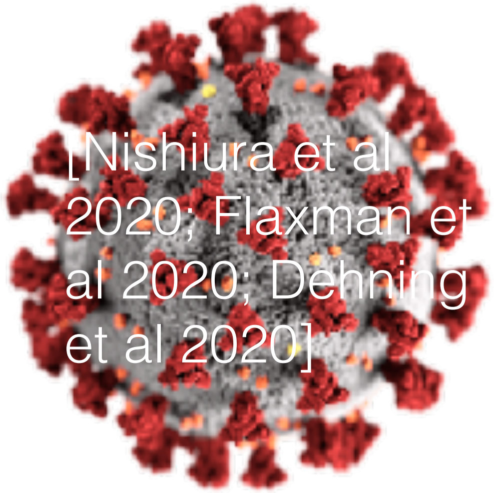


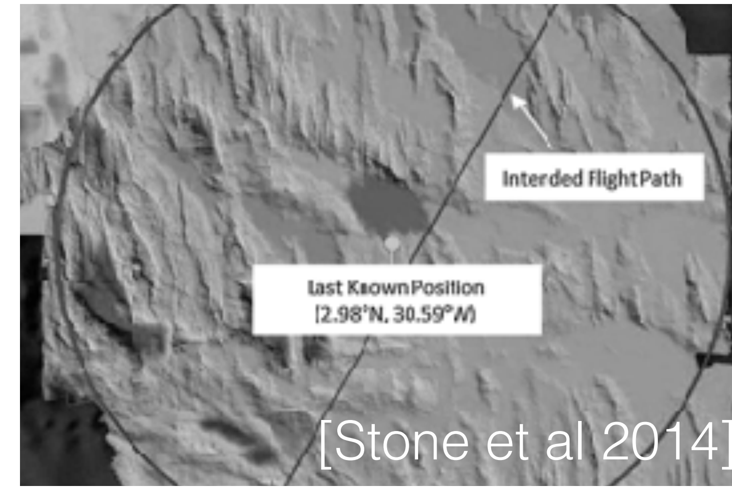
- Instead: easy-to-compute bound on Wasserstein [Huggins, Kasprzak, Campbell, Broderick, 2020]
  - Wasserstein bounds error in posterior mean and variance
- Part of a validated workflow for VB [Huggins, Kasprzak, Campbell, Broderick, 2020]
  - Builds on e.g. [Dieng et al 2017; Yao et al 2018]
- See also [Gorham, Mackey 2015, 2017; Chwialkowski, Strathmann, Gretton 2016; Jitkrittum et al 2017; Talts et al 2018, etc.]

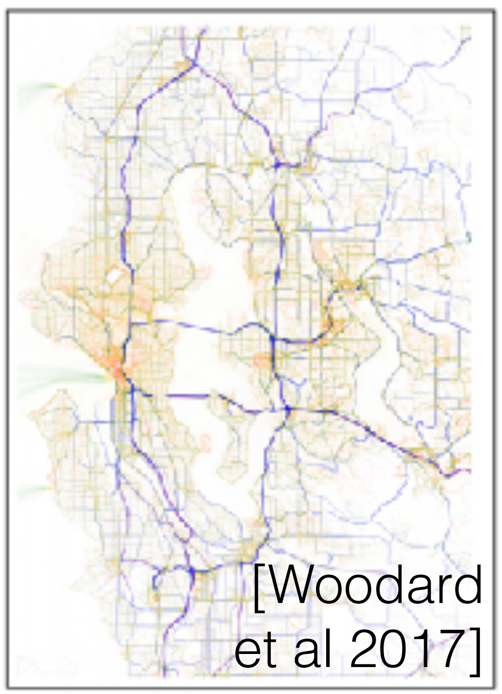
Roadmap

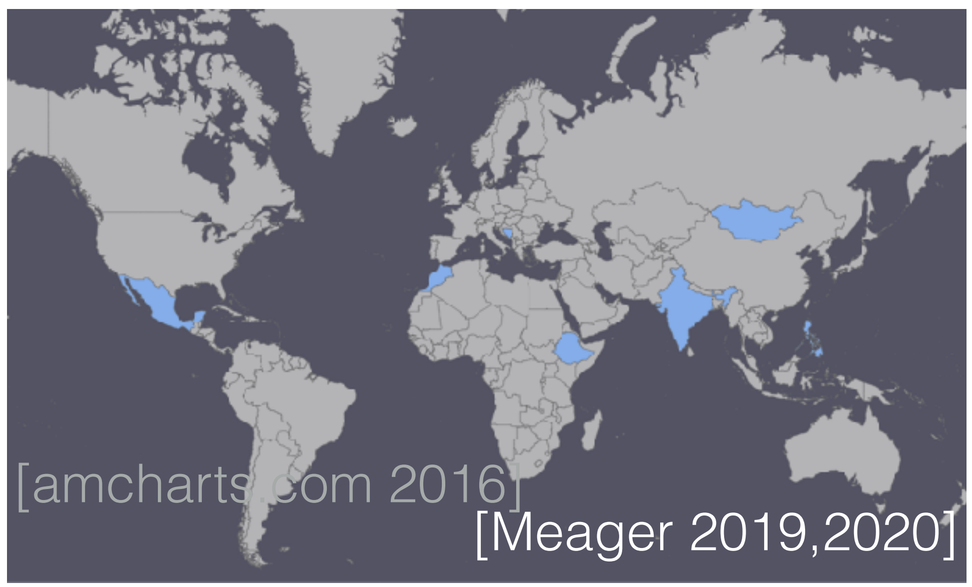
- Bayes & Approximate Bayes review
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use VB?
- When can we trust VB?
- Where do we go from here?

Bayesian inference









[mc-stan.org]

- Goals: good point estimates, uncertainty estimates
- Challenge: speed (compute, user), reliable inference

What to read next

Textbooks and Reviews

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

Our Experiments

- R Giordano, T Broderick, and MI Jordan. Linear response methods for accurate covariance estimates from mean field variational Bayes. *NeurIPS* 2015.
- R Giordano, T Broderick, R Meager, JH Huggins, and MI Jordan. Fast robustness quantification with variational Bayes. *ICML Data4Good Workshop* 2016.
- R Giordano, T Broderick, and MI Jordan. Covariances, robustness, and variational Bayes. *JMLR* 2018.
- J Huggins, M Kasprzak, T Campbell, T Broderick. Validated Variational Inference via Practical Posterior Error Bounds. ArXiv: 1910.04102. *AISTATS* 2020, to appear.
- T Campbell and T Broderick. Automated scalable Bayesian inference via Hilbert coresets. *JMLR* 2019.
- T Campbell and T Broderick. Bayesian Coreset Construction via Greedy Iterative Geodesic Ascent. *ICML* 2018.

References (1/6)

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

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

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

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

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

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

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

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

AB Dieng, D Tran, R Ranganath, J Paisley, & D Blei. Variational Inference via χ Upper Bound Minimization, *NeurIPS* 2017.

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

References (2/6)

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

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

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

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

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

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

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

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

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

J Huggins, M Kasprzak, T Campbell, T Broderick. Validated Variational Inference via Practical Posterior Error Bounds. ArXiv:1910.04102, *AISTATS* 2020, to appear.

References (3/6)

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

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

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

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

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

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

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

Application References (4/6)

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

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

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

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

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

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

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

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

Application References (5/6)

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

Meager, Rachael. “Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of 7 randomized experiments.” *AEJ: Applied*, 2019.

Meager, Rachael. “Aggregating Distributional Treatment Effects: A Bayesian Hierarchical Analysis of the Microcredit Literature.” Under review, 2020.

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

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

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

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

Additional image references (6/6)

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

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

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

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

A. Kongrut. 23 Jan 2020. Obtained from: https://www.bangkokpost.com/opinion/opinion/1841569/bungling-govt-is-losing-the-pm2-5-war

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