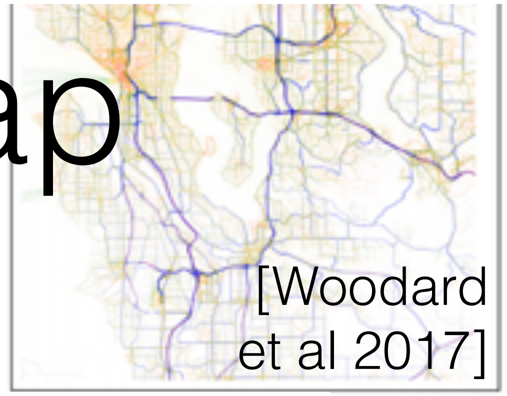
Part II: Automated, Scalable Bayesian Inference via Data Summarization

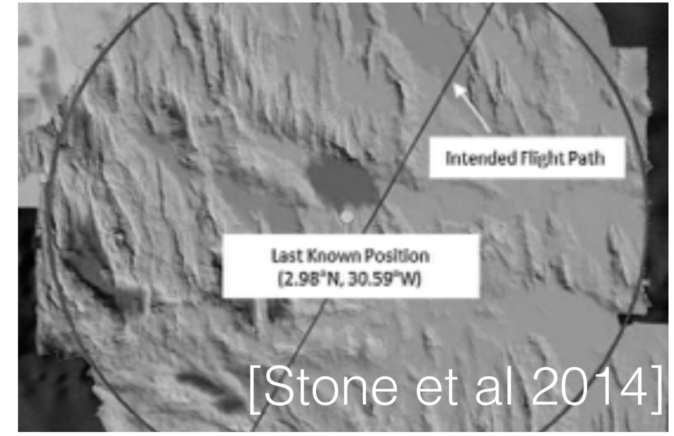
http://www.tamarabroderick.com/tutorial_2018_icml.html

Recap



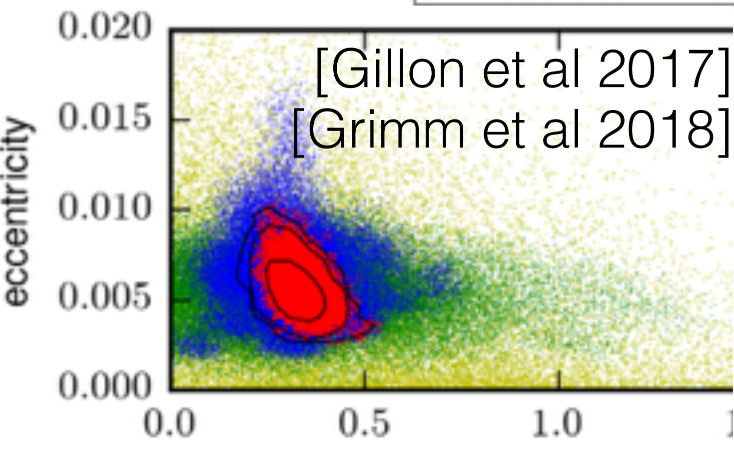
[Woodard et al 2017]

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

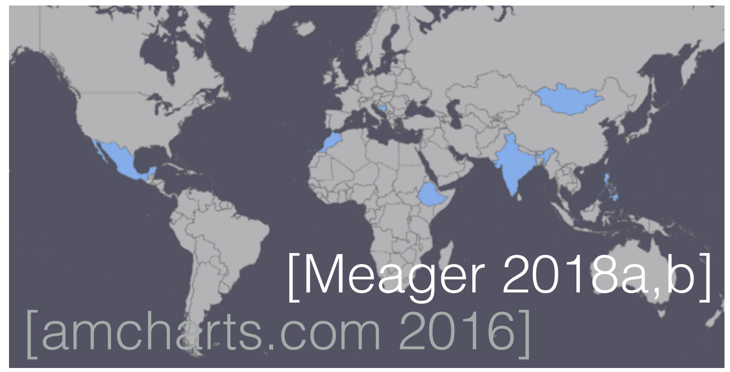


[Stone et al 2014]

[mc-stan.org]



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

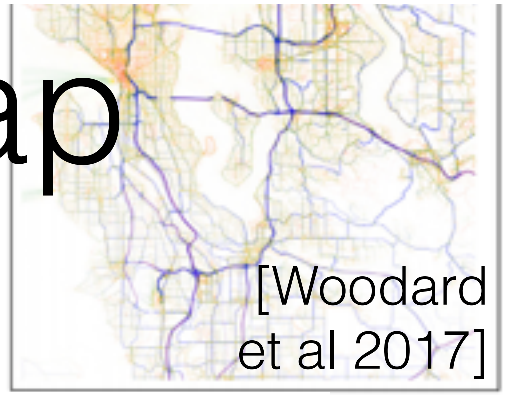


[Meager 2018a,b]
[amcharts.com 2016]

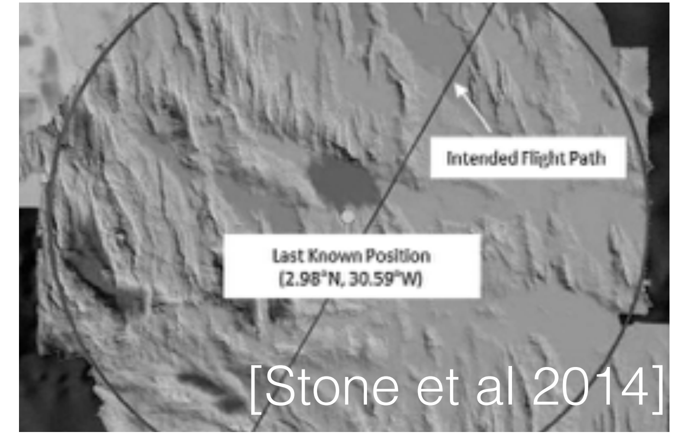
[Chati, Balakrishnan 2017]
[Julian Hertzog 2016]

[Kuikka et al 2014]
[Baltic Salmon Fund]

Recap

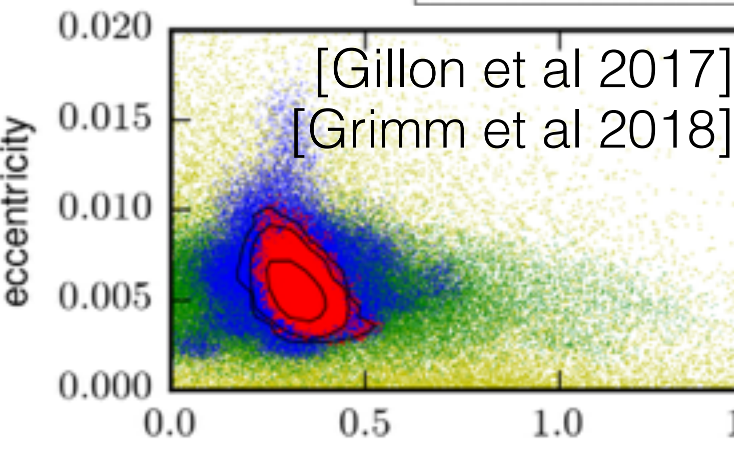

[Woodard et al 2017]


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

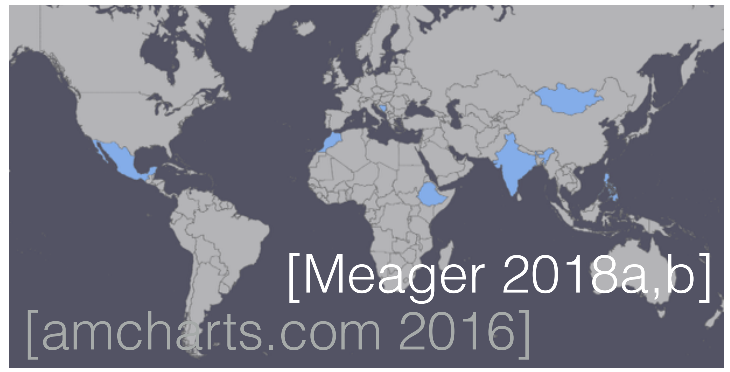


[Stone et al 2014]

[mc-stan.org]



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

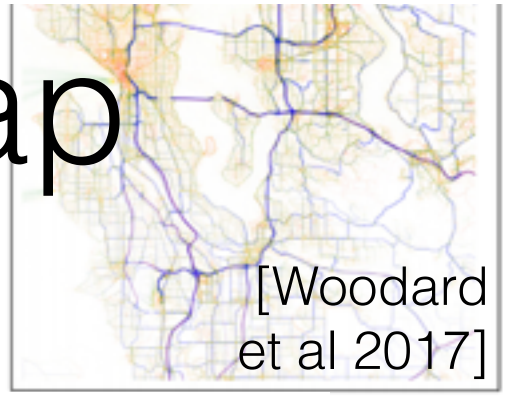

[Meager 2018a,b]
[amcharts.com 2016]

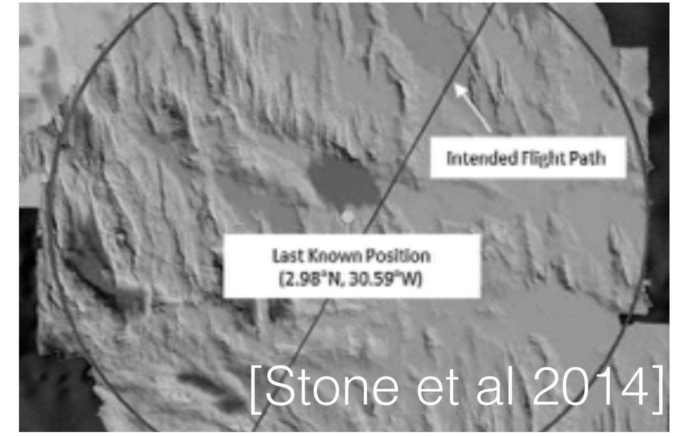

[Chati, Balakrishnan 2017]
[Julian Hertzog 2016]


[Kuikka et al 2014]
[Baltic Salmon Fund]

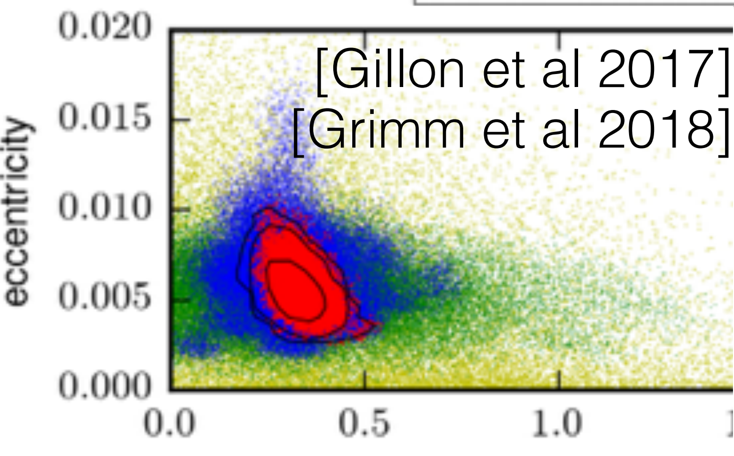
$$p(\theta)$$

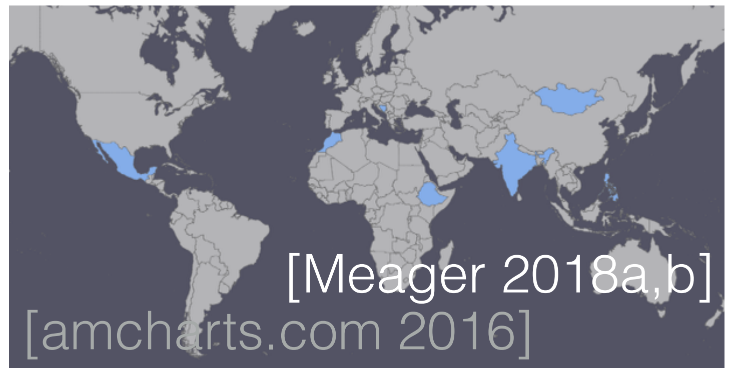
Recap





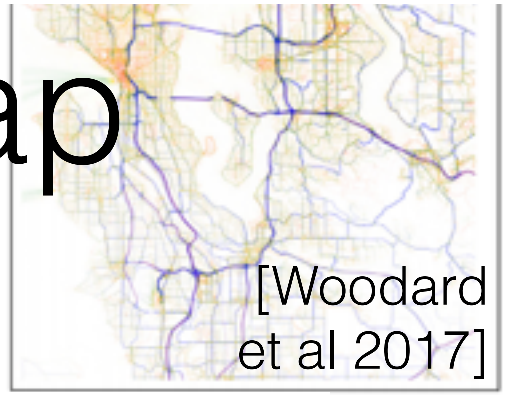
[mc-stan.org]

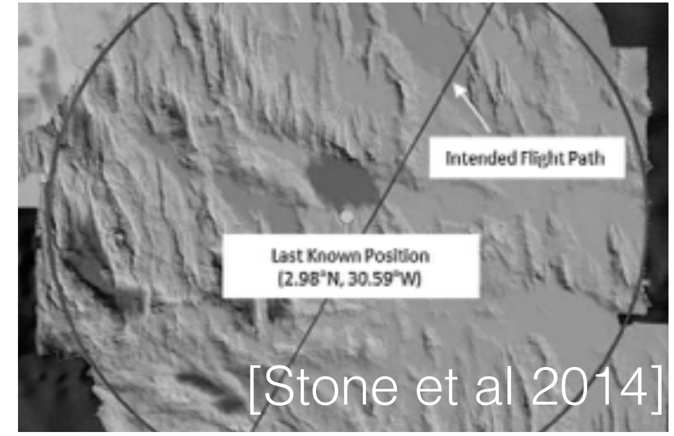




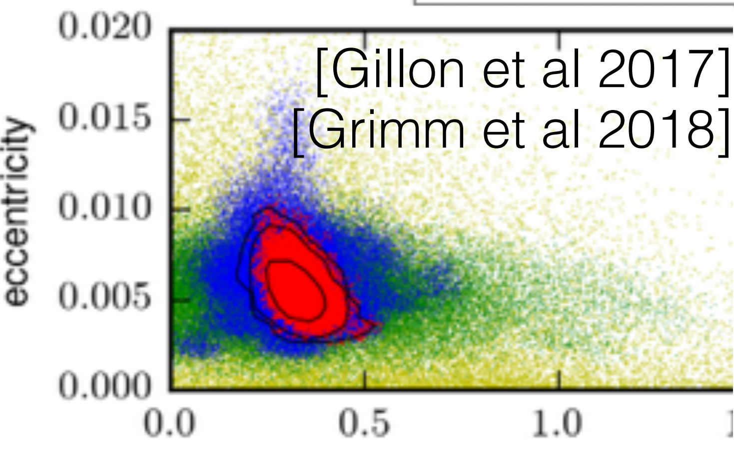
$$p(y|\theta)p(\theta)$$

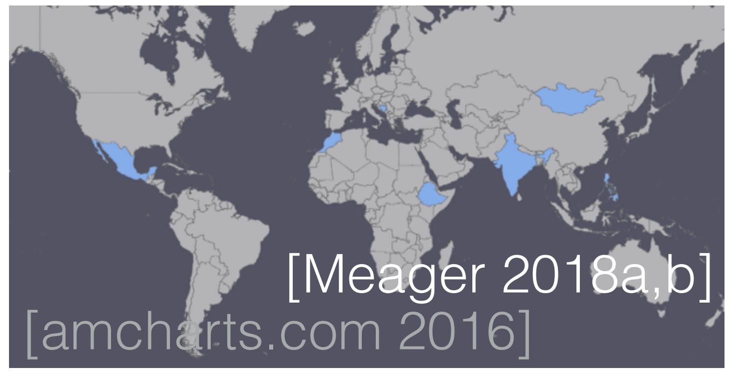
Recap





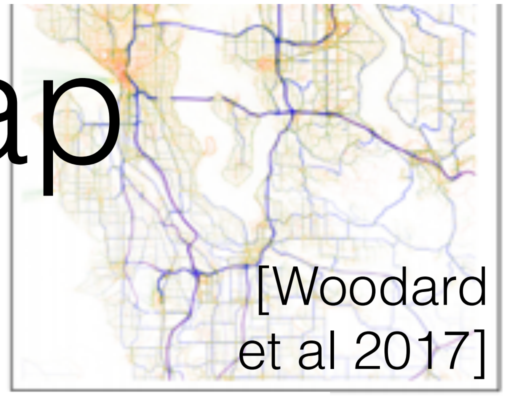
[mc-stan.org]

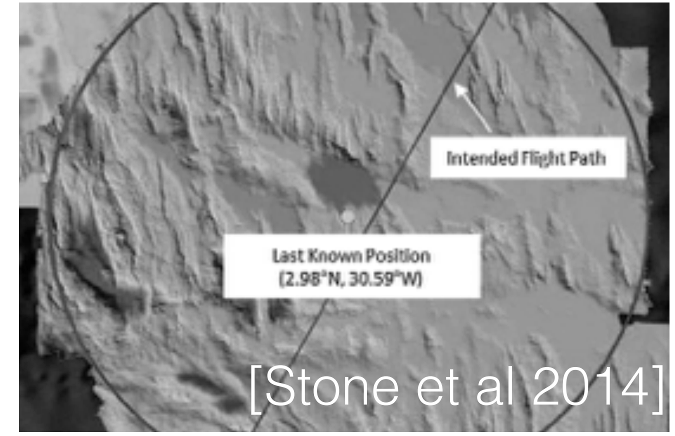




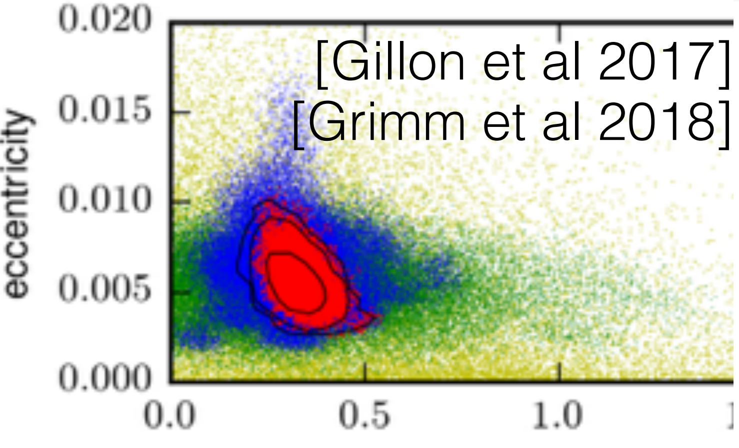
$$p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$$

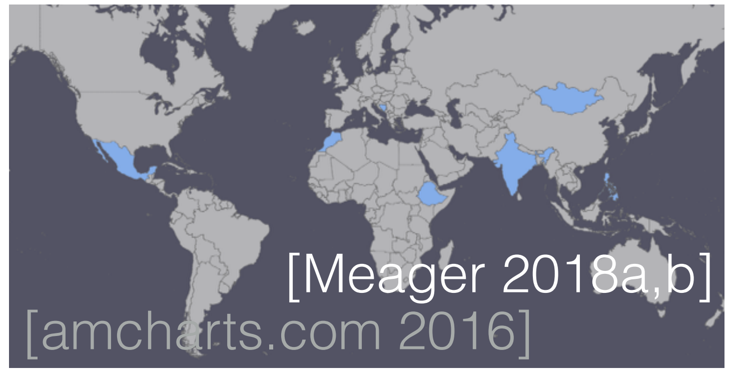
Recap



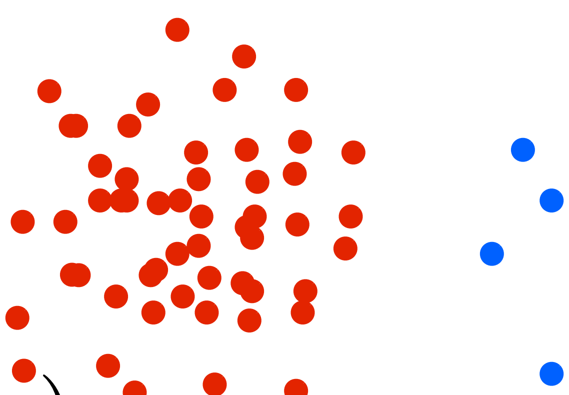


[mc-stan.org]



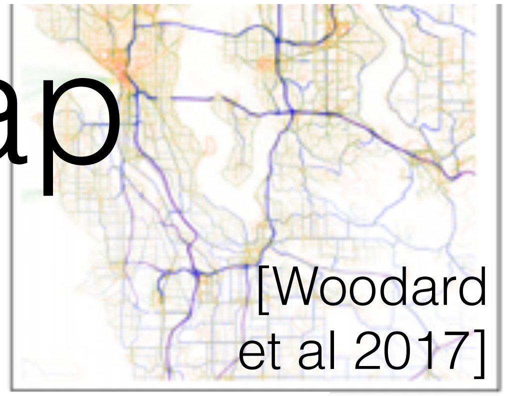


$$p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$$

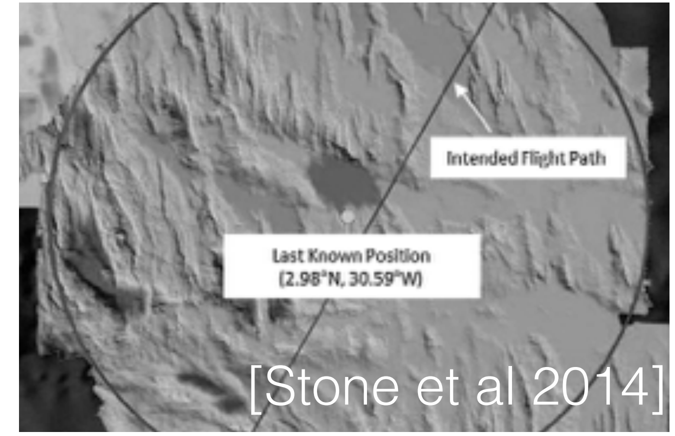


$(x_n, {\color{red} y_n})$

Recap

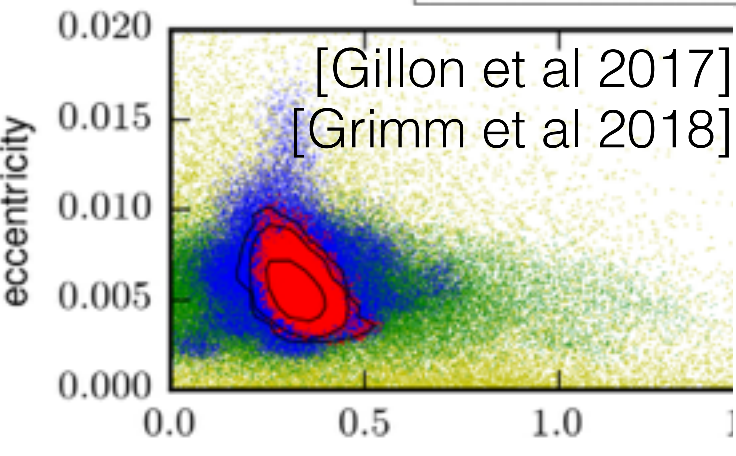

[Woodard et al 2017]


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

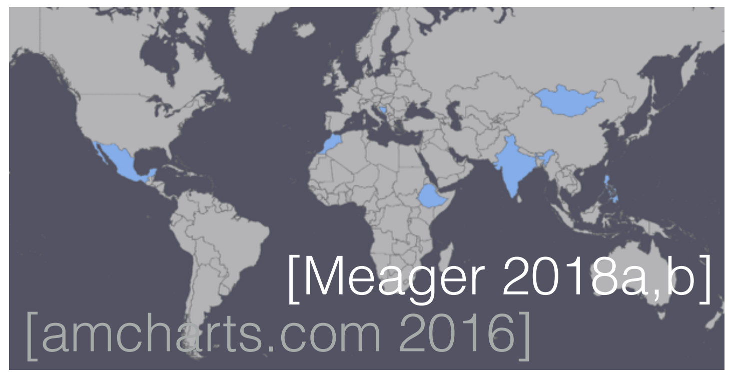


[Stone et al 2014]

[mc-stan.org]



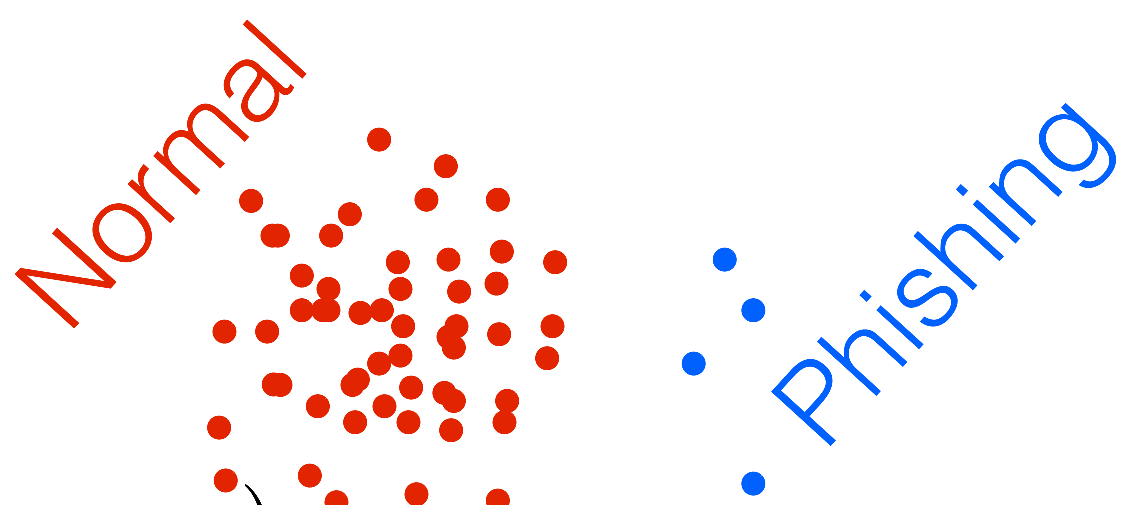
[Gillon et al 2017] [Grimm et al 2018]


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


[Julian Hertzog 2016] [Chati, Balakrishnan 2017]

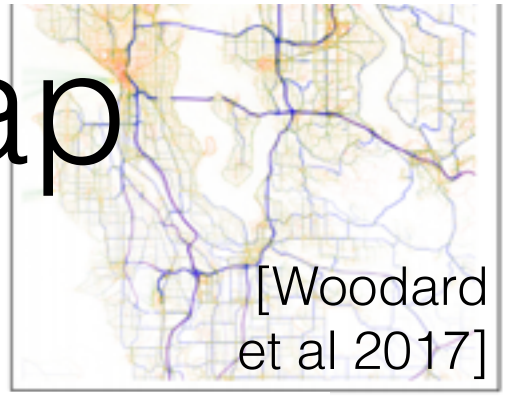

[Kuikka et al 2014] [Baltic Salmon Fund]

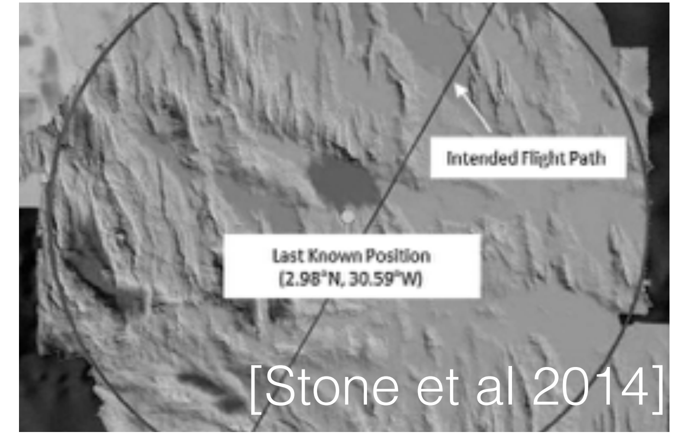
$$p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$$



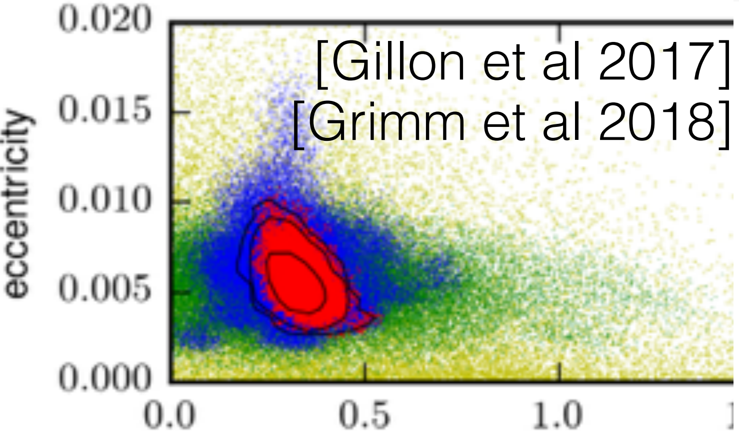
(x_n, y_n)

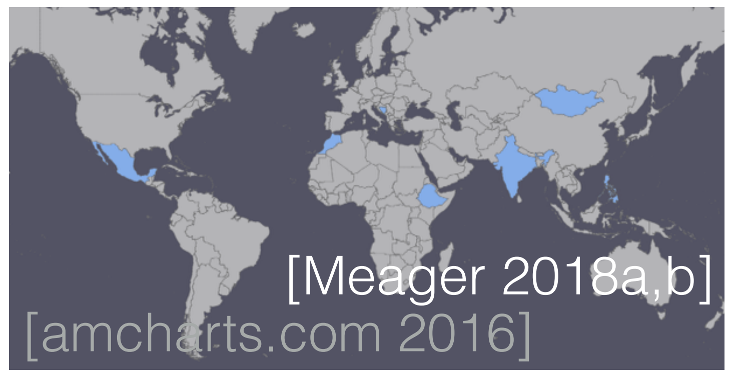
Recap



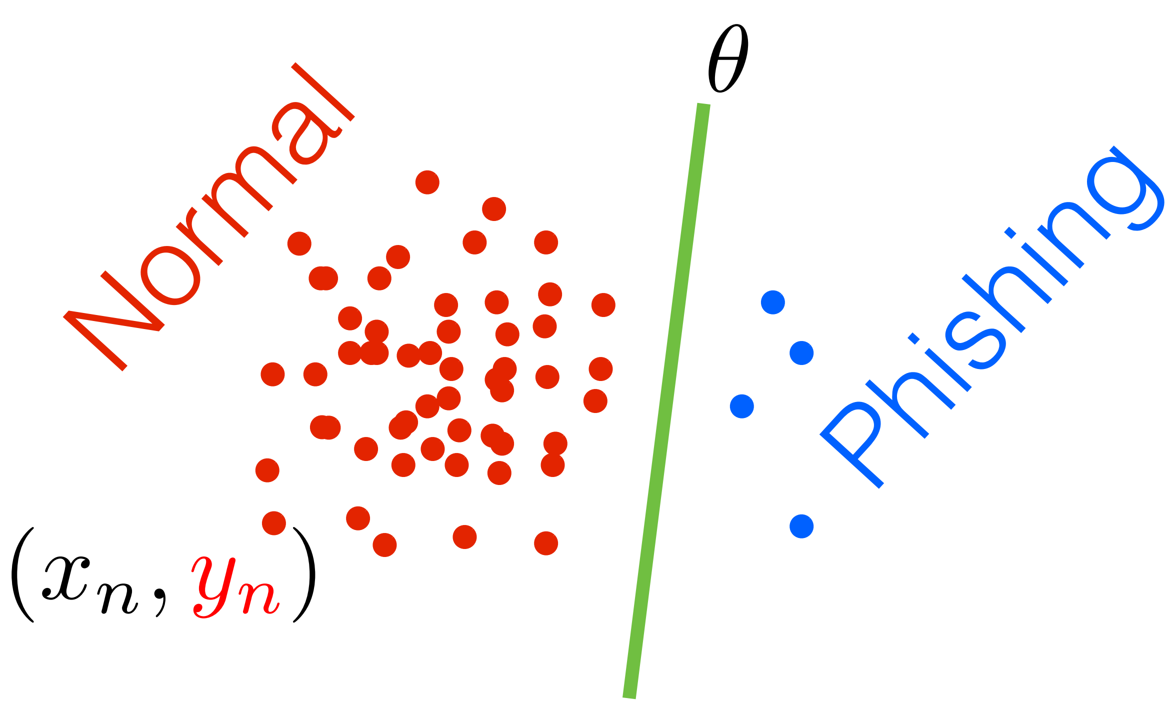


[mc-stan.org]

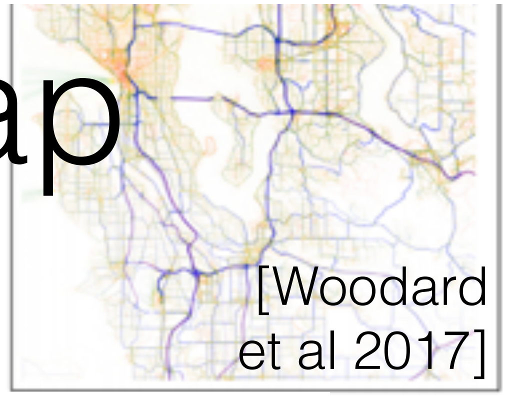




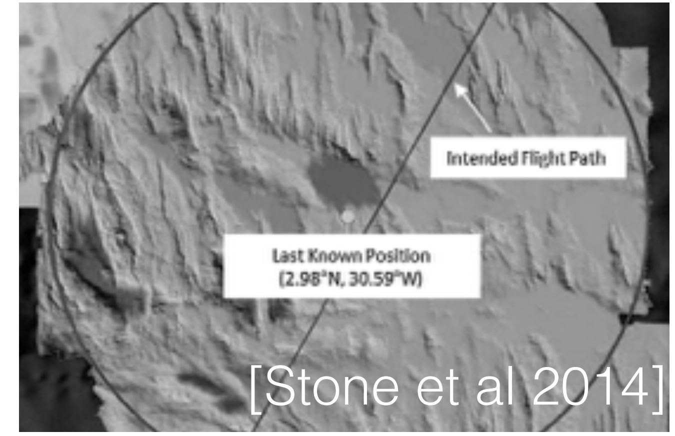
$$p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$$



Recap

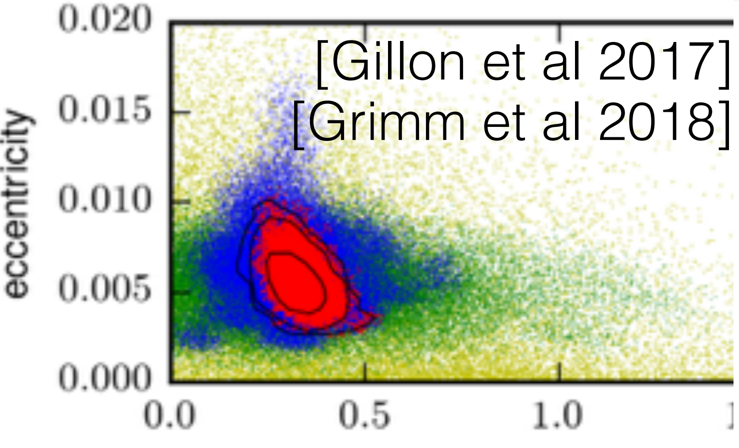

[Woodard et al 2017]


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

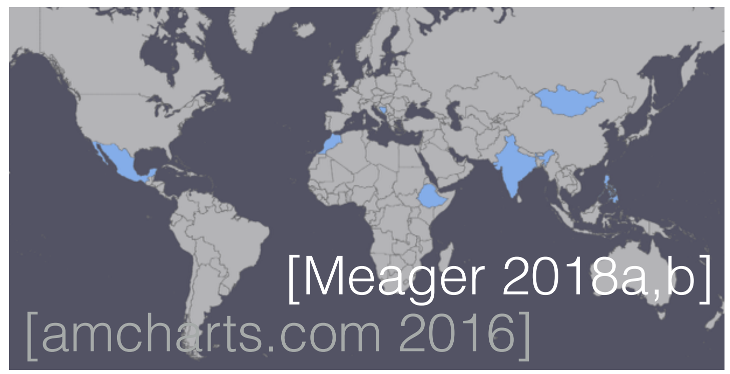


[Stone et al 2014]

[mc-stan.org]



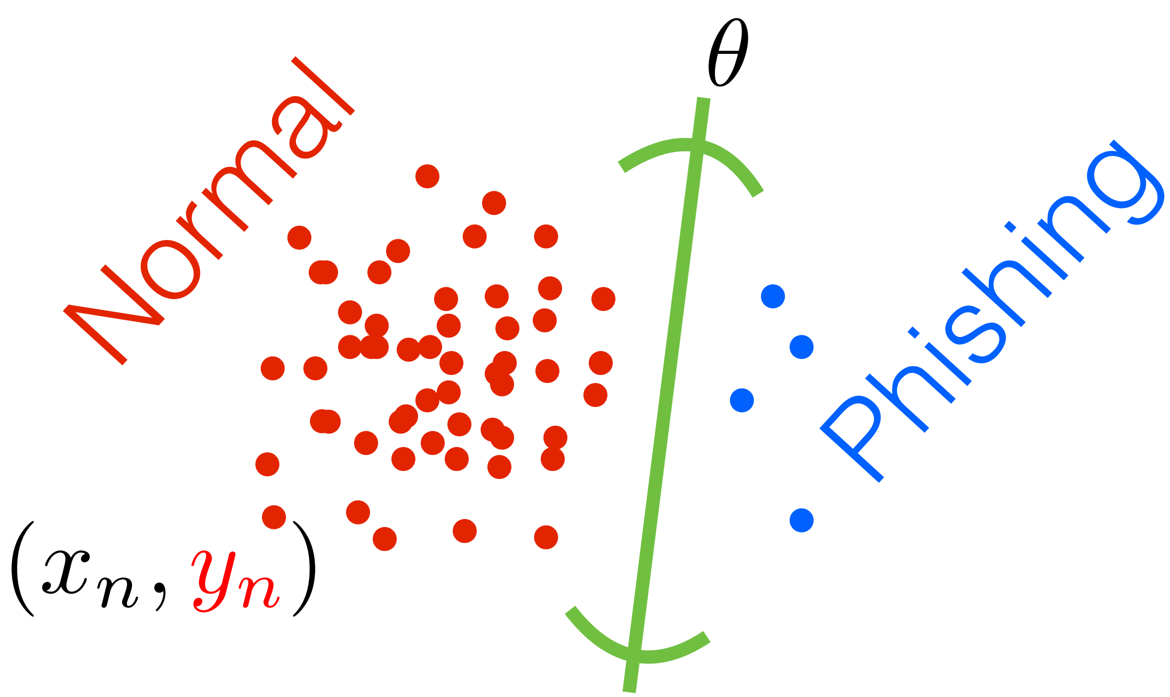
[Gillon et al 2017] [Grimm et al 2018]


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

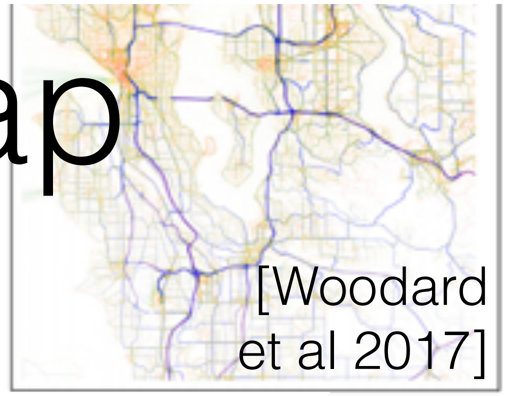

[Chati, Balakrishnan 2017] [Julian Hertzog 2016]

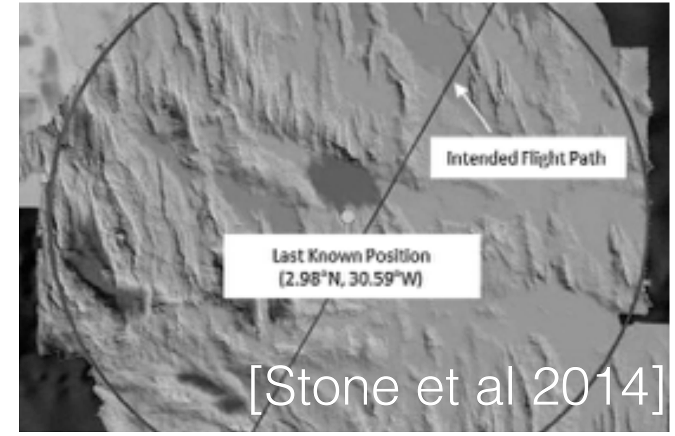

[Kuikka et al 2014] [Baltic Salmon Fund]

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

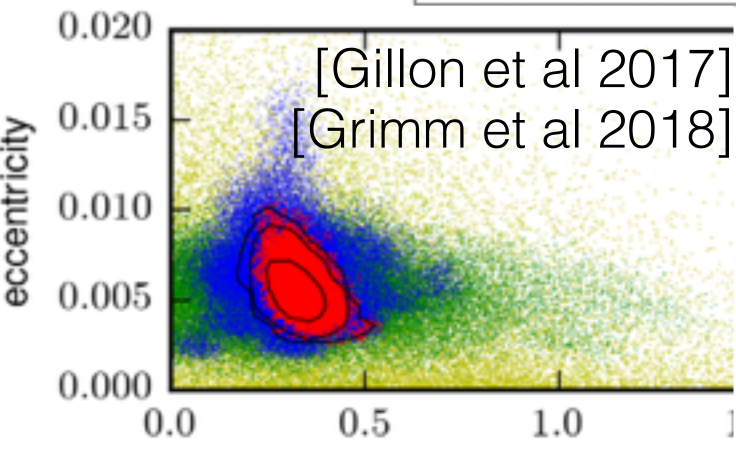


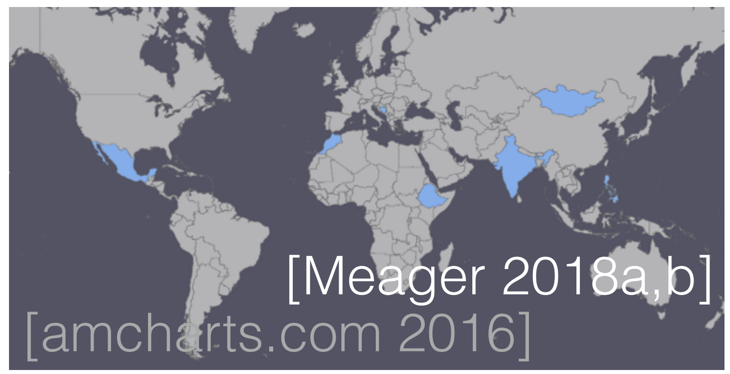
Recap



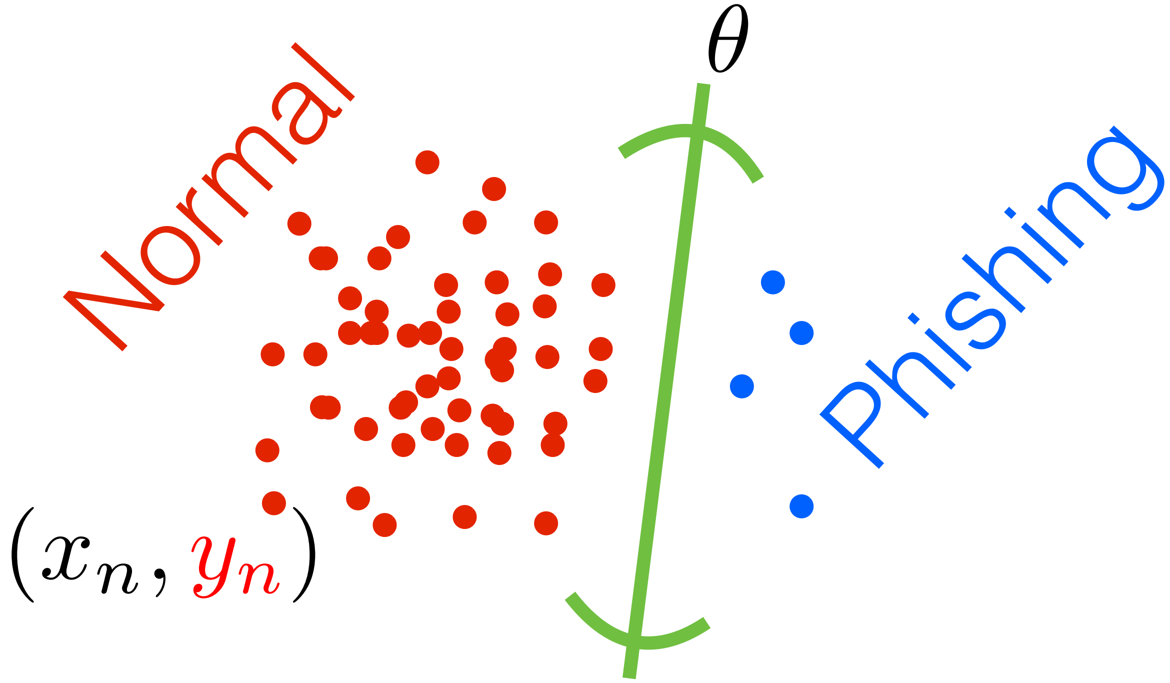


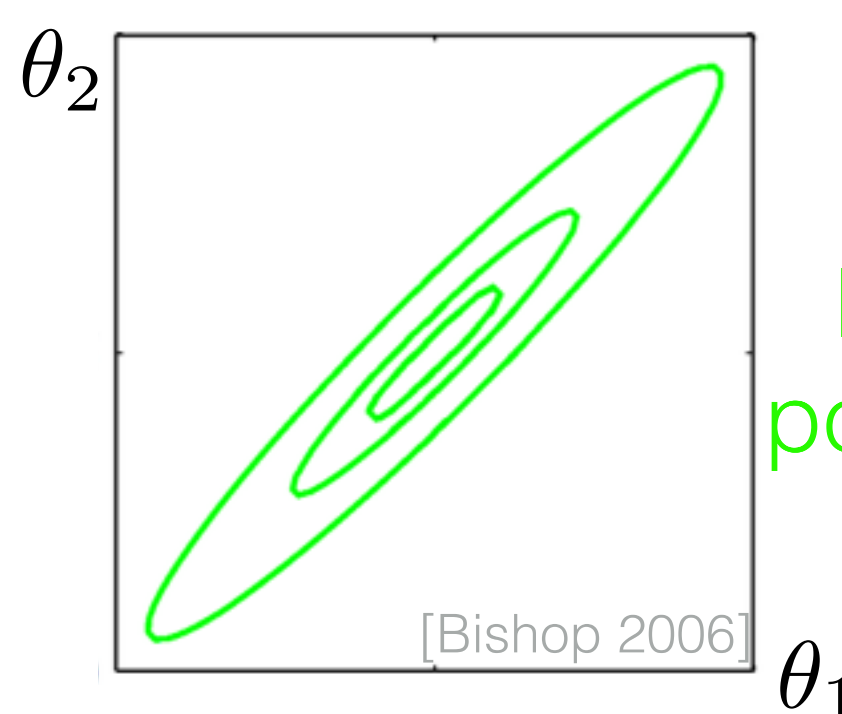
[mc-stan.org]



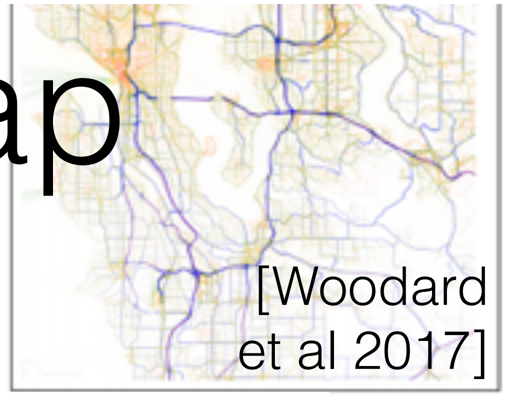


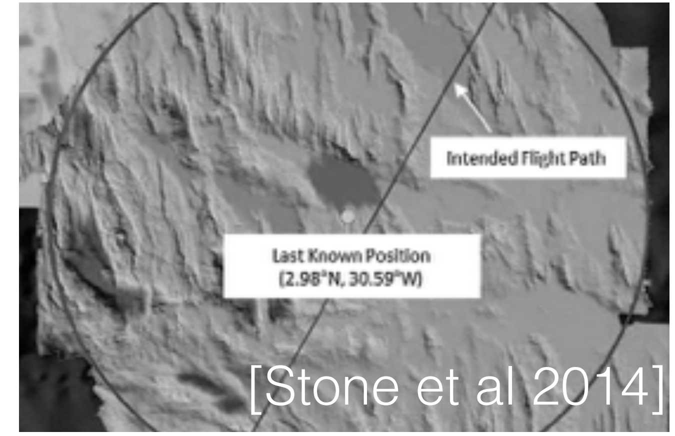
$$p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$$



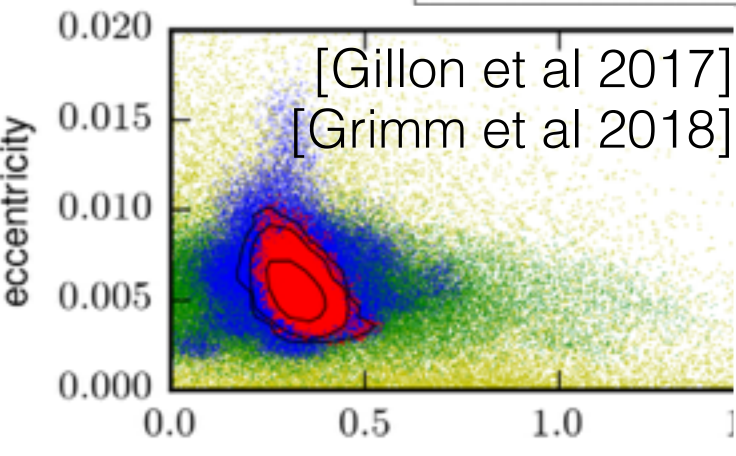


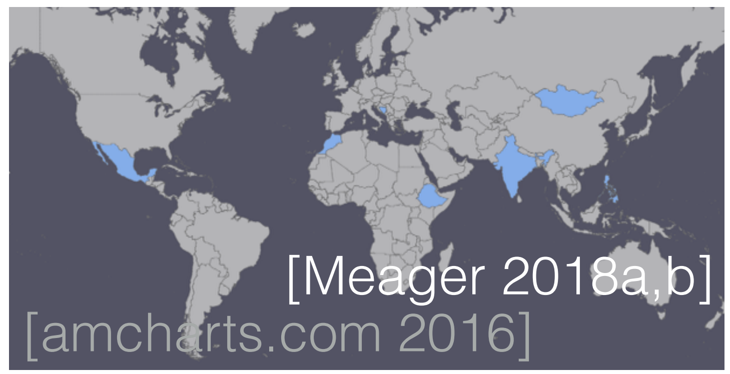
Recap



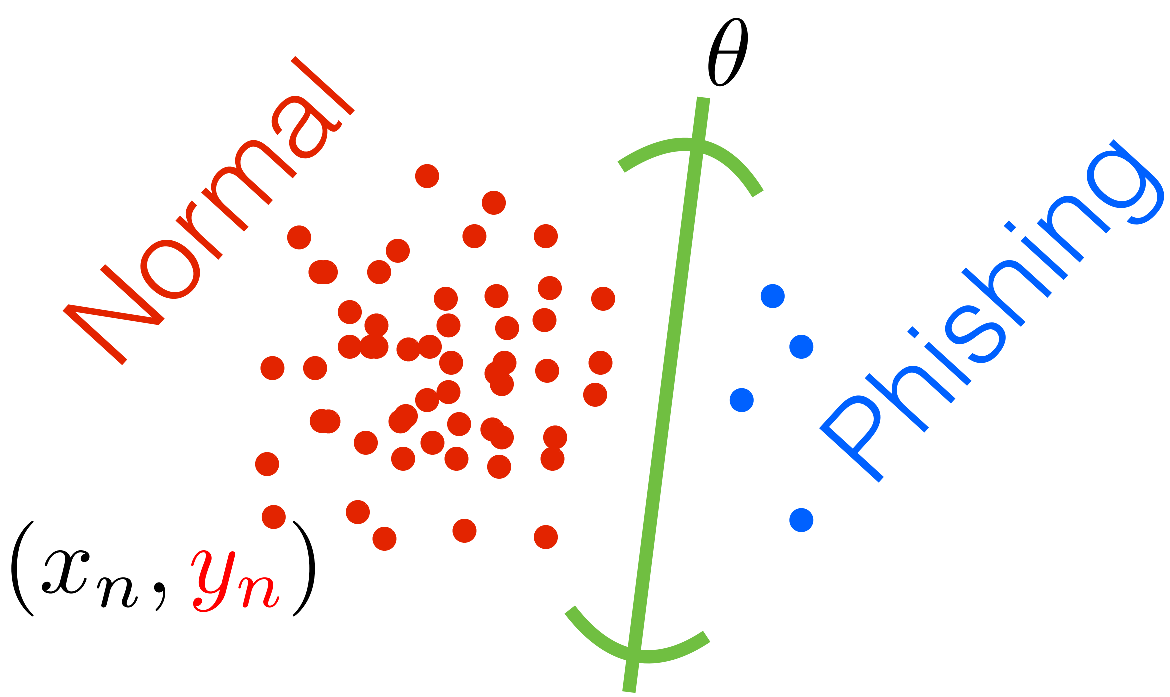


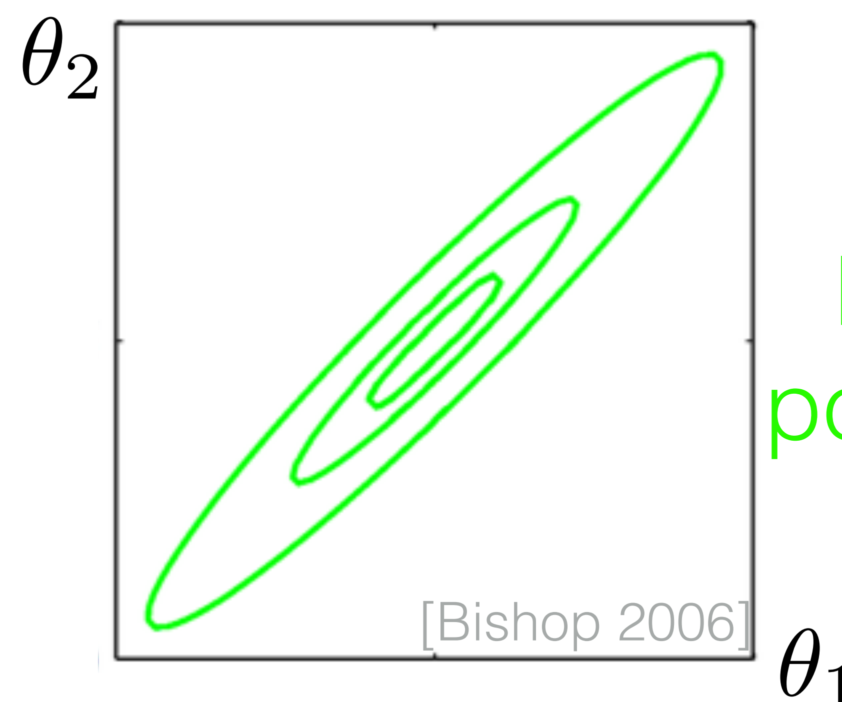
[mc-stan.org]





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





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

Roadmap

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

Roadmap

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

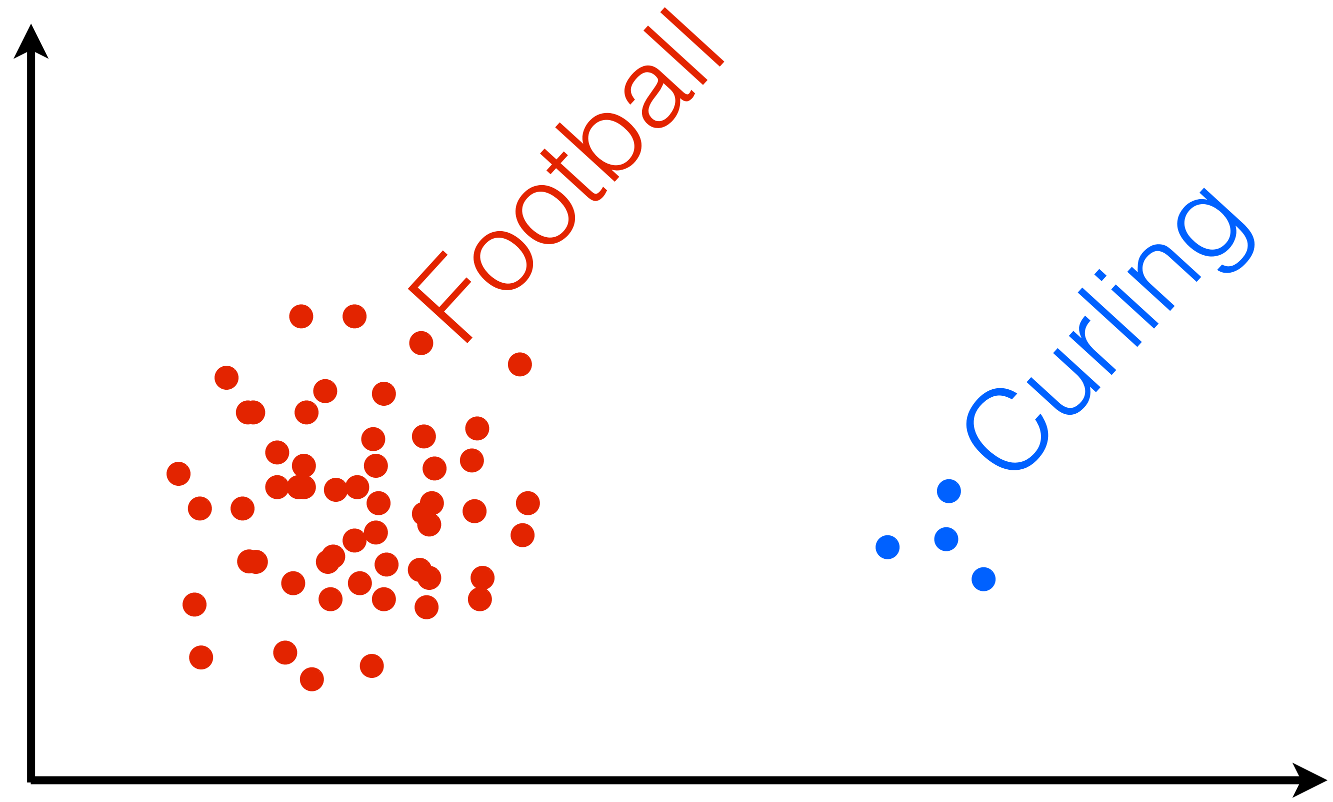
Bayesian coresets

Bayesian coresets

- Observe: redundancies can exist even if data isn't "tall"

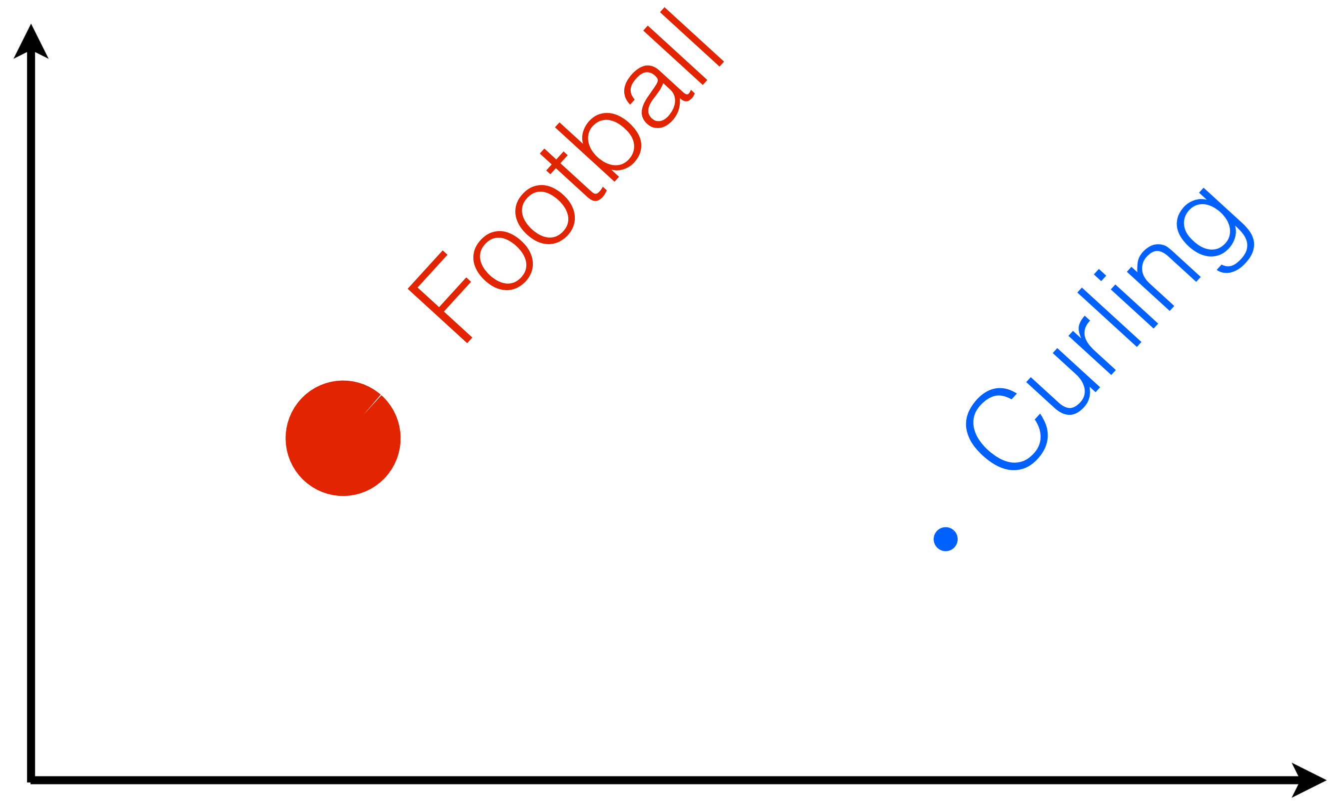
Bayesian coresets

- Observe: redundancies can exist even if data isn't "tall"



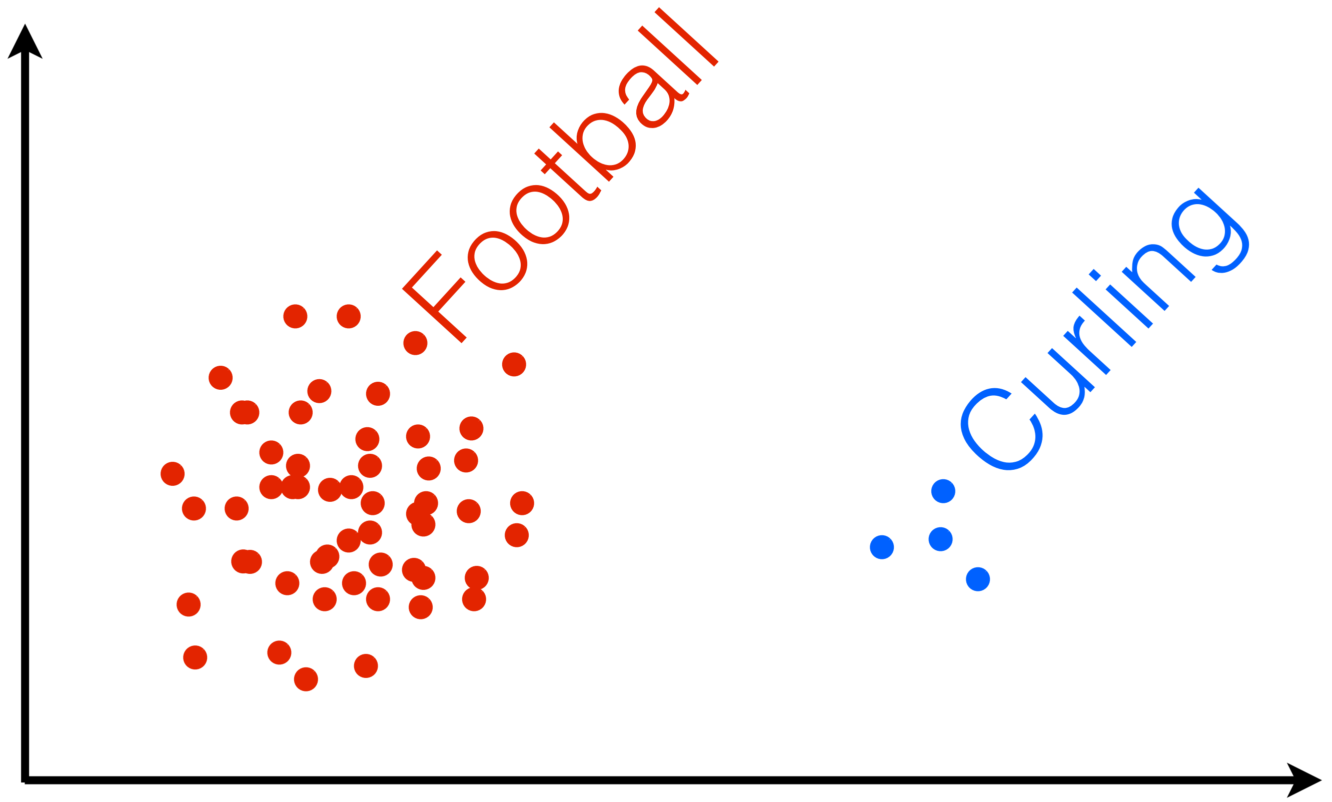
Bayesian coresets

- Observe: redundancies can exist even if data isn't "tall"



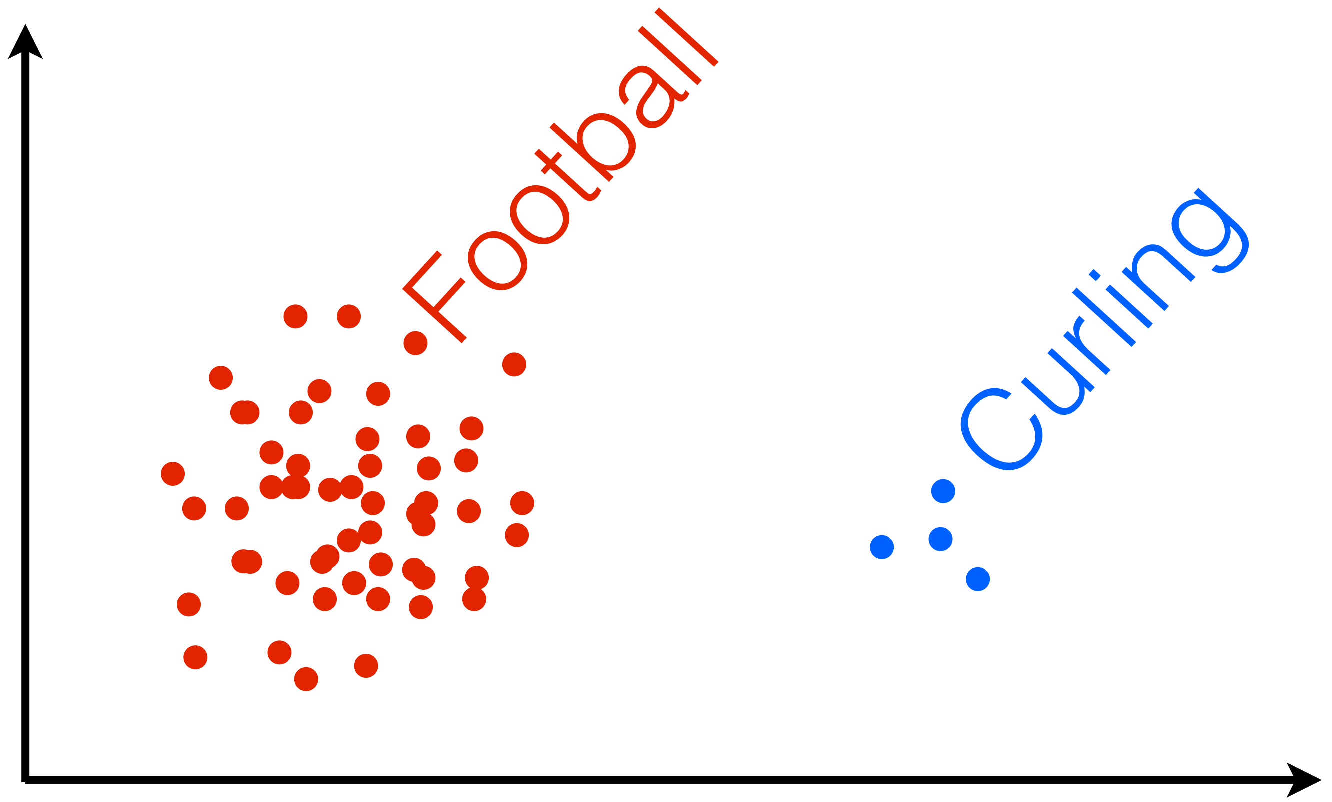
Bayesian coresets

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



Bayesian coresets

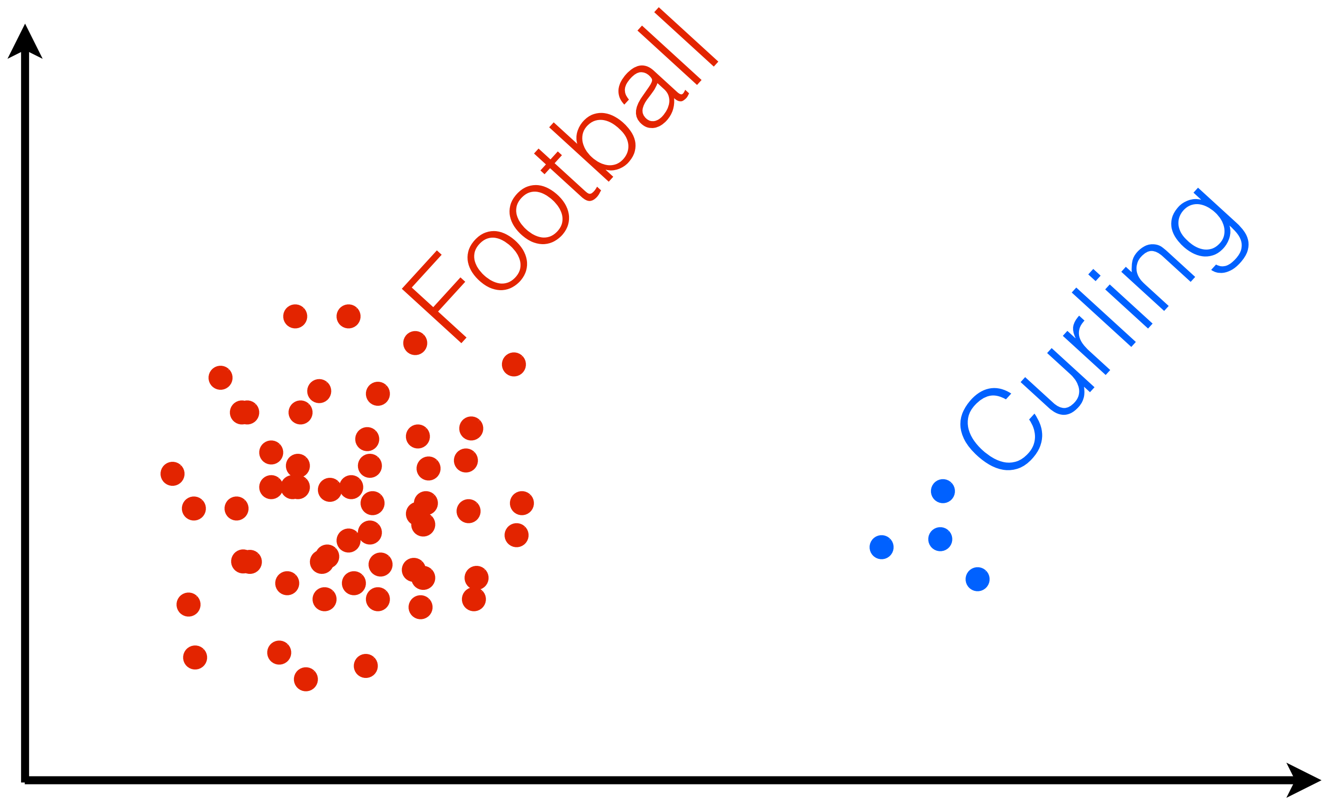
- Observe: redundancies can exist even if data isn't "tall"
- Coresets: pre-process data to get a smaller, weighted data set



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

Bayesian coresets

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

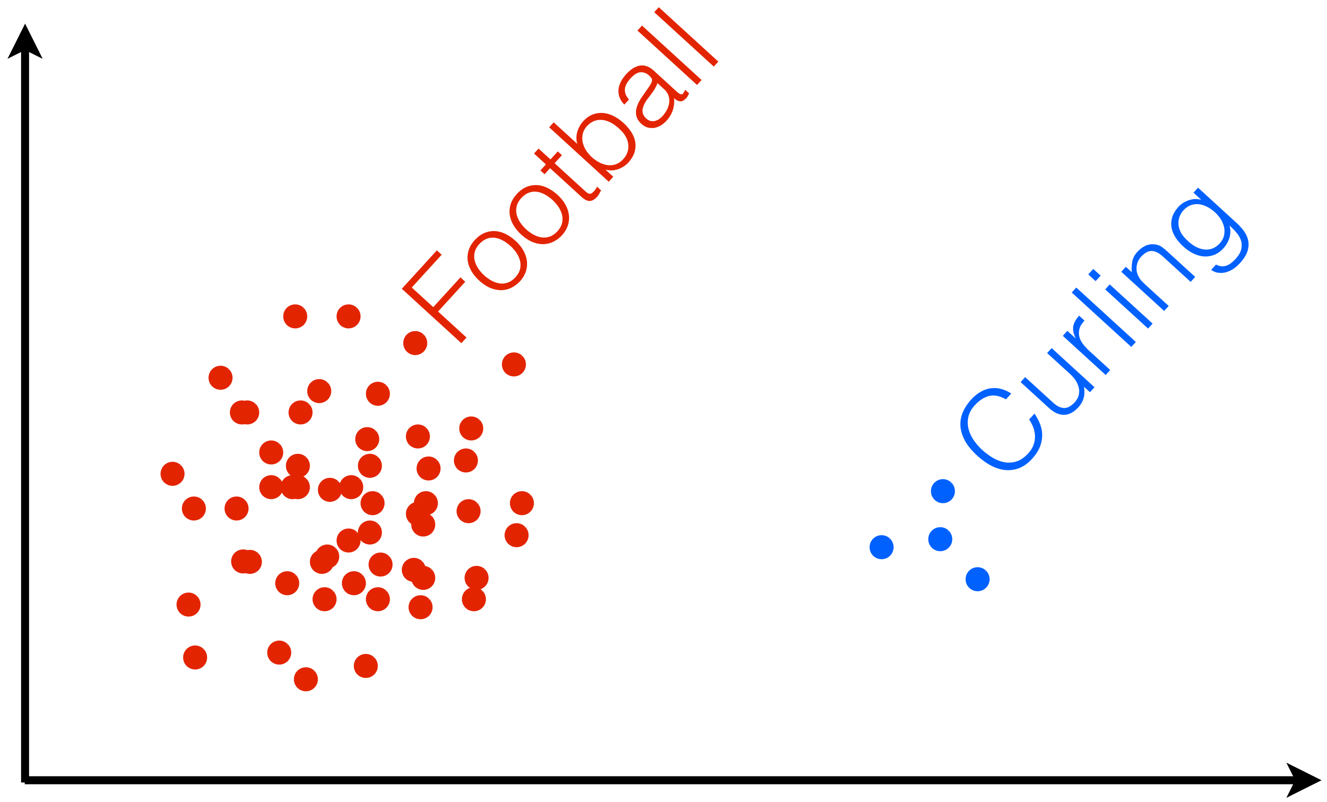


- Theoretical guarantees on quality

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

Bayesian coresets

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

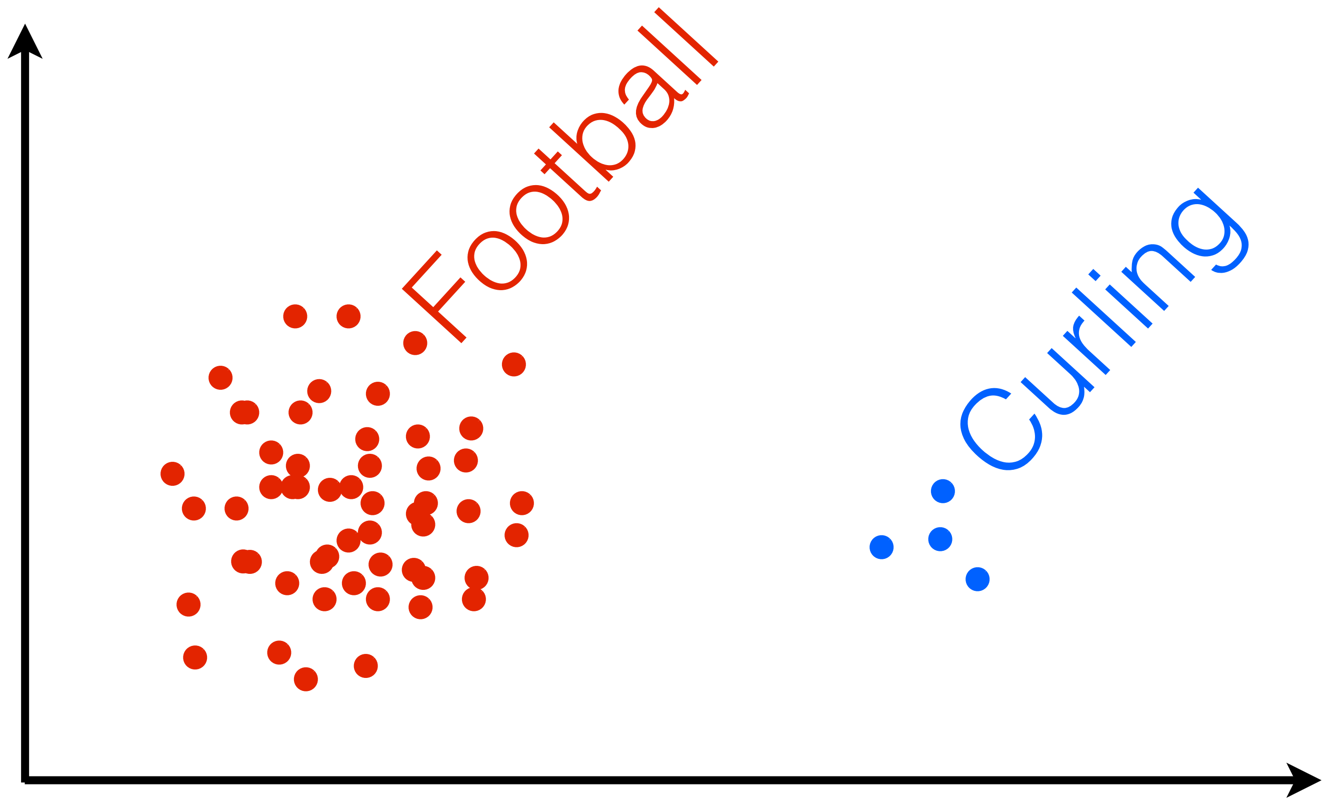


- Theoretical guarantees on quality
- Previous heuristics: data squashing, big data GPs

[Bădoiu, Har-Peled, Indyk 2002; Agarwal et al 2005; Feldman & Langberg 2011; DuMouchel et al 1999; Madigan et al 2002]

Bayesian coresets

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

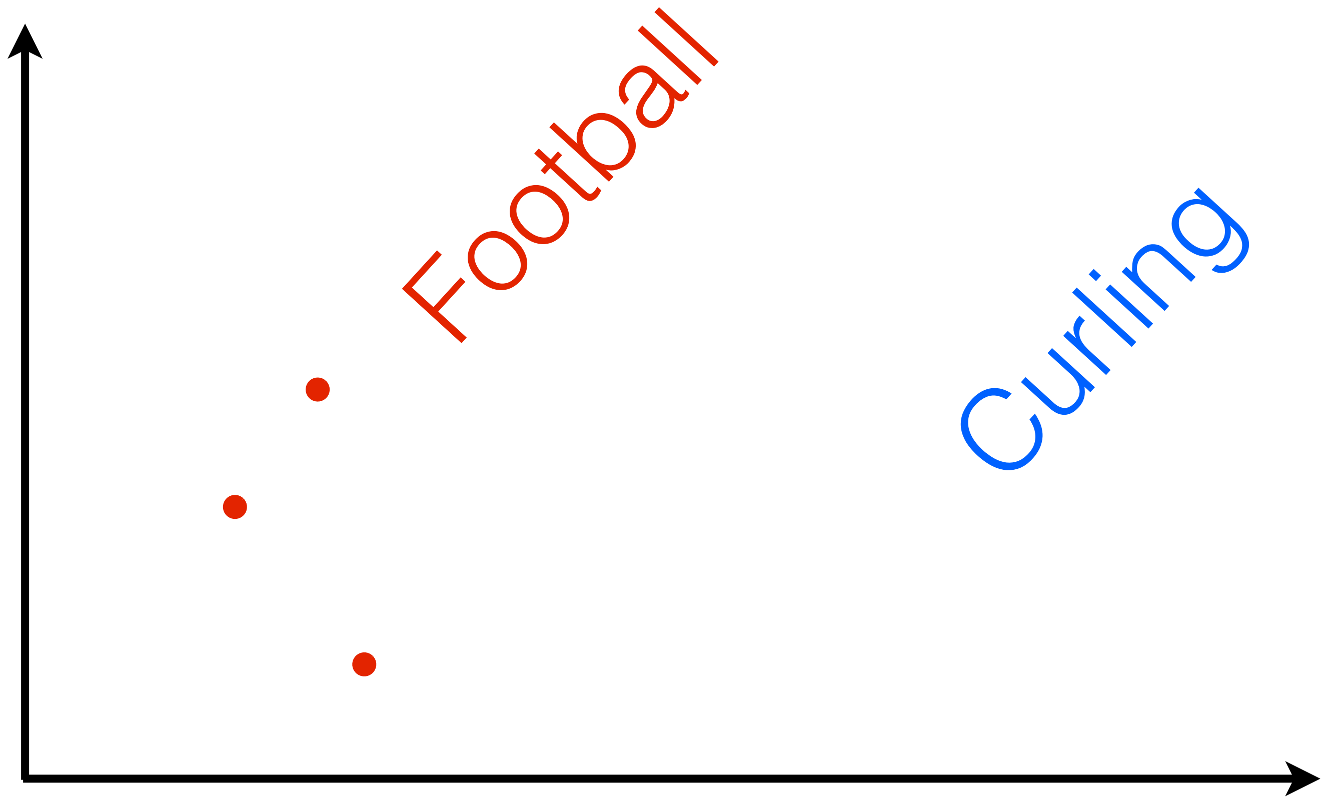


- Theoretical guarantees on quality
- Previous heuristics: data squashing, big data GPs
- Cf. subsampling

[Bădoiu, Har-Peled, Indyk 2002; Agarwal et al 2005; Feldman & Langberg 2011; DuMouchel et al 1999; Madigan et al 2002]

Bayesian coresets

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

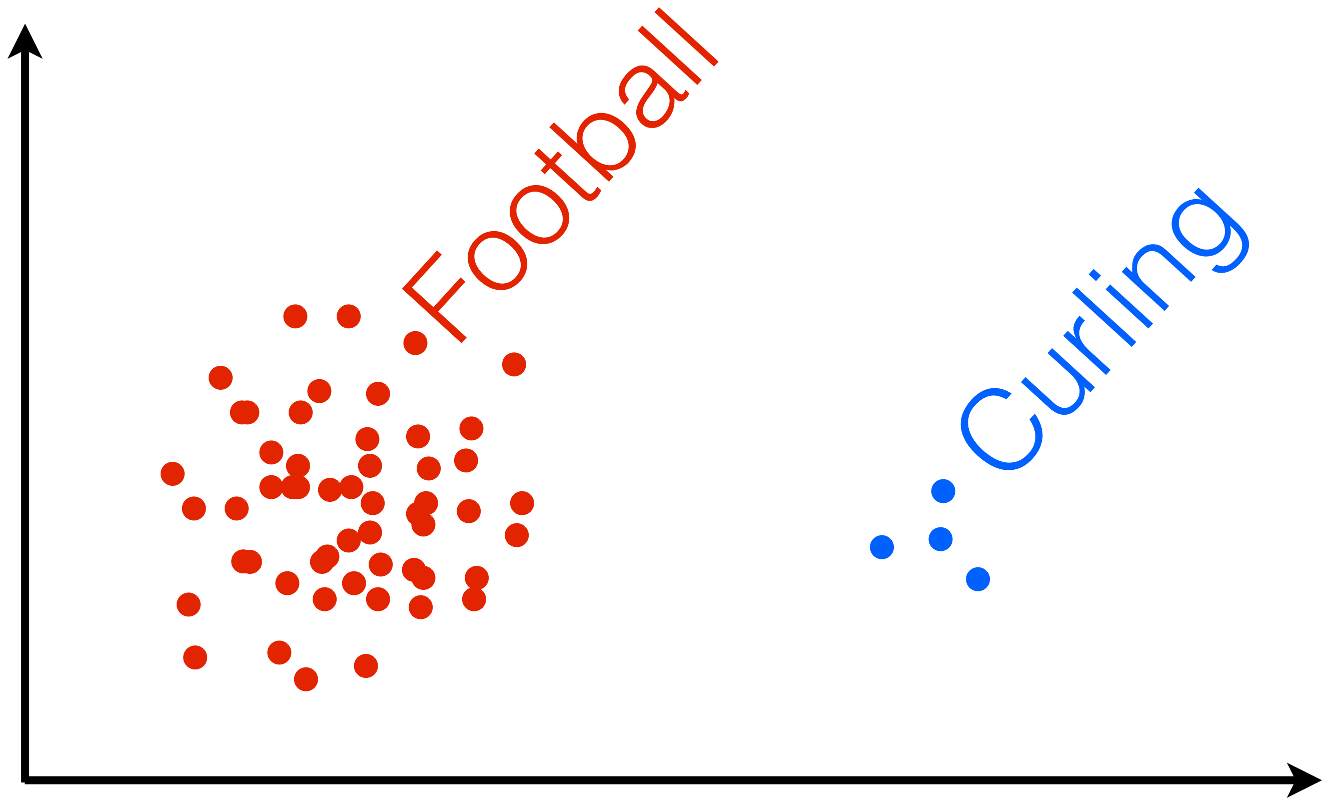


- Theoretical guarantees on quality
- Previous heuristics: data squashing, big data GPs
- Cf. subsampling

[Bădoiu, Har-Peled, Indyk 2002; Agarwal et al 2005; Feldman & Langberg 2011; DuMouchel et al 1999; Madigan et al 2002]

Bayesian coresets

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

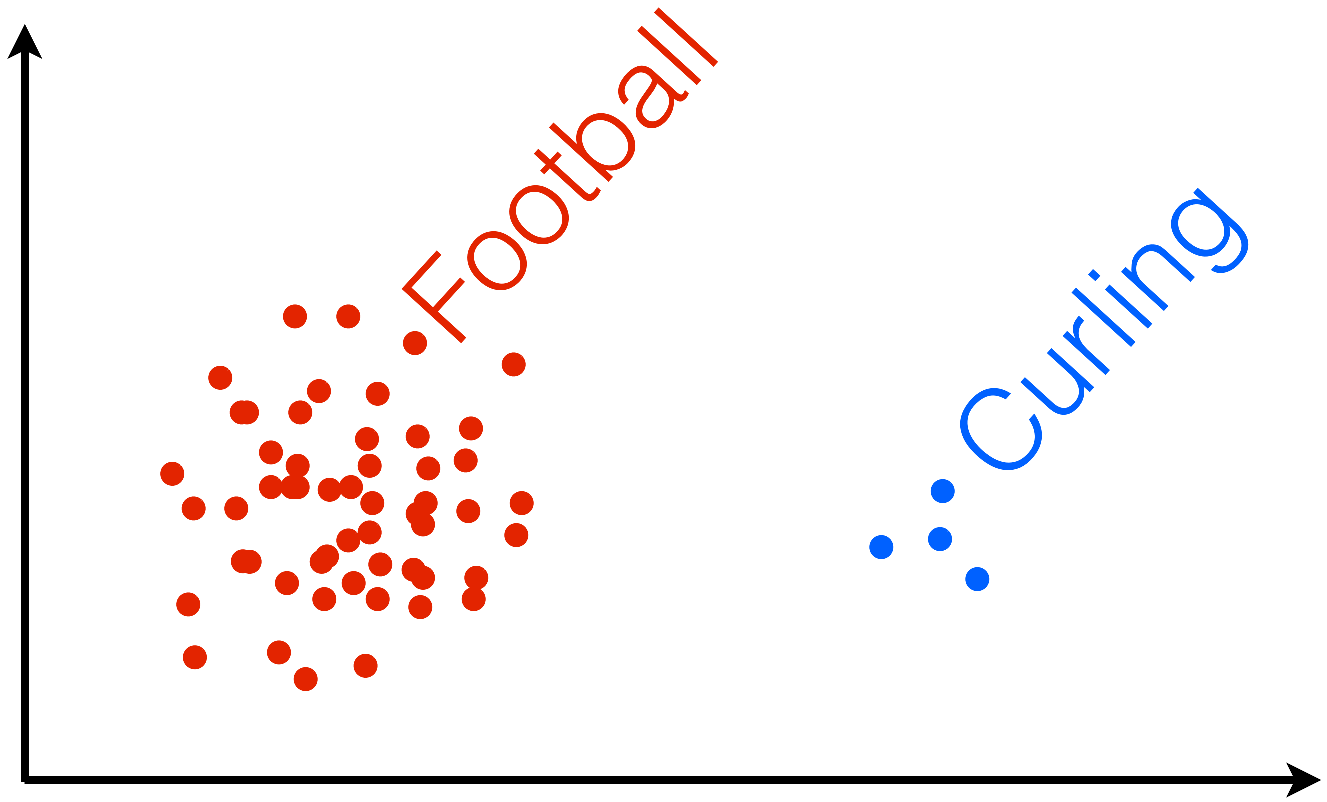


- Theoretical guarantees on quality
- Previous heuristics: data squashing, big data GPs
- Cf. subsampling

[Bădoiu, Har-Peled, Indyk 2002; Agarwal et al 2005; Feldman & Langberg 2011; DuMouchel et al 1999; Madigan et al 2002]

Bayesian coresets

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

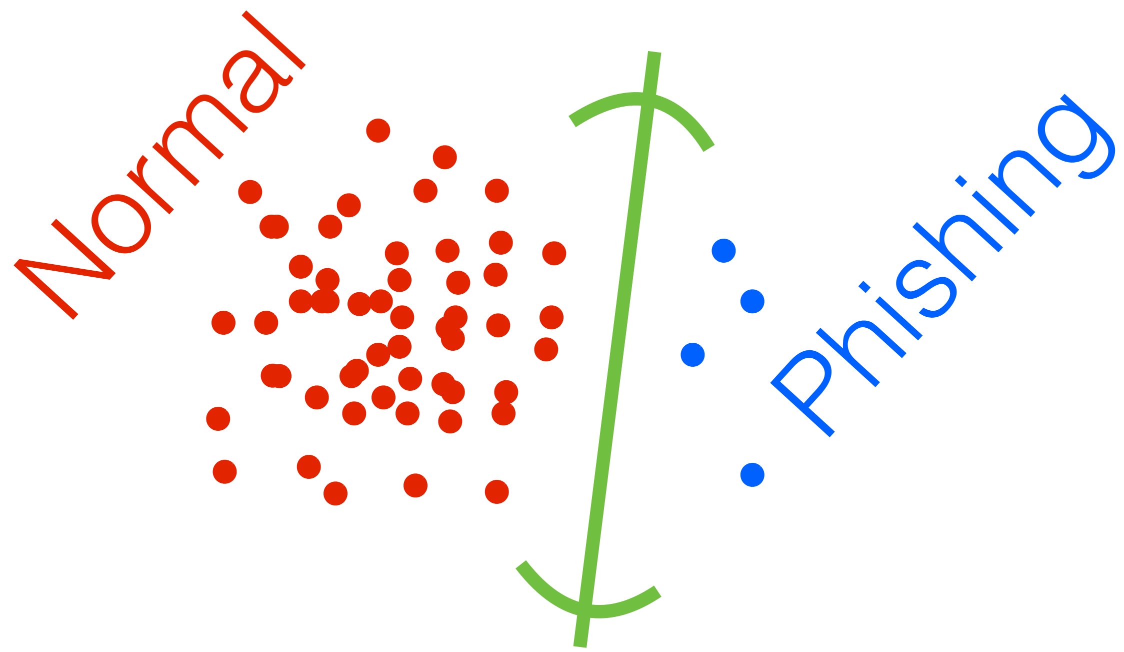


- Theoretical guarantees on quality
- Previous heuristics: data squashing, big data GPs
- Cf. subsampling
- How to develop **coresets for Bayes**?

[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 2017; Campbell, Broderick 2018]

Bayesian coresets

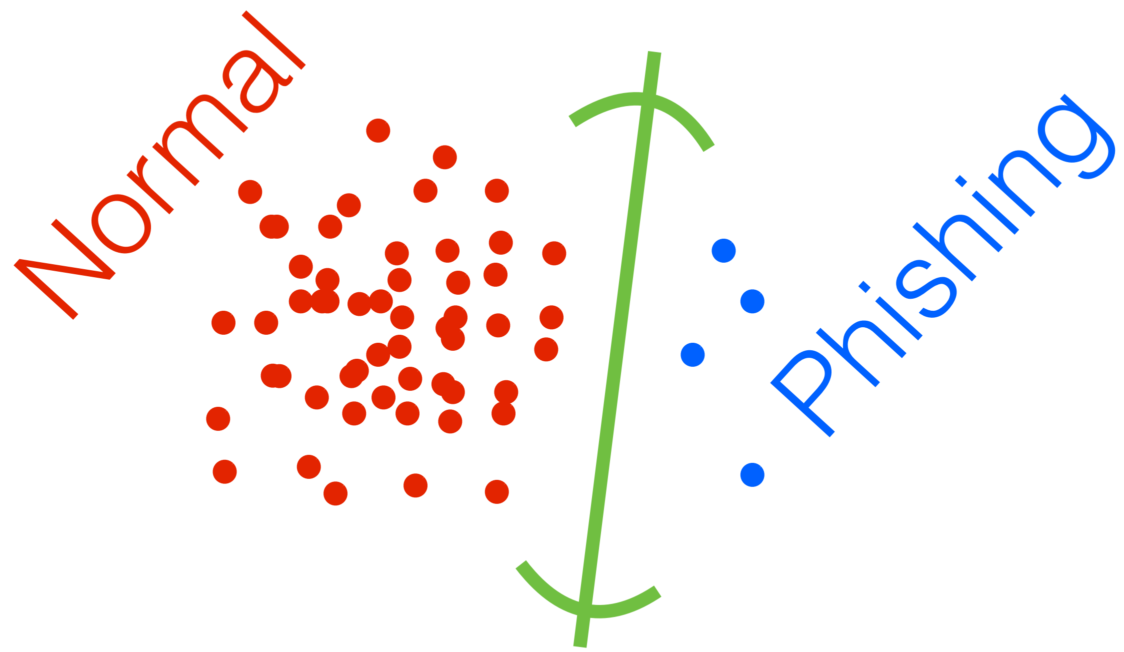
- Posterior $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$



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

Bayesian coresets

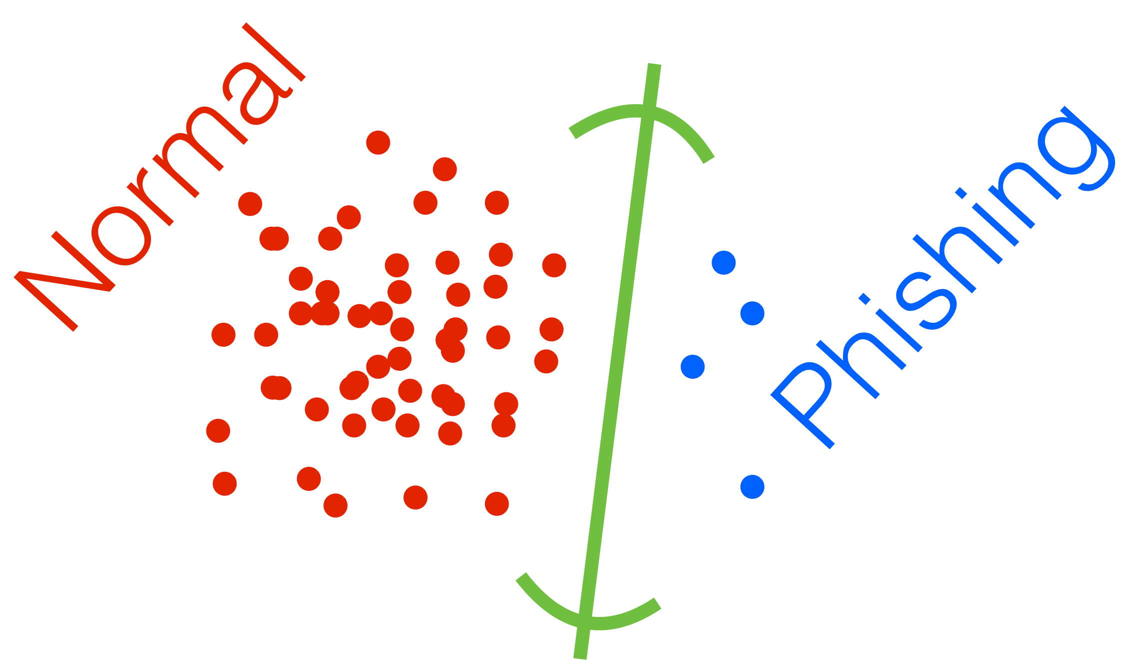
- Posterior $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$
- Log likelihood $\mathcal{L}_n(\theta) := \log p(y_n|\theta)\,, \;\; \mathcal{L}(\theta) := \sum_{n=1}^{N} \mathcal{L}_n(\theta)$



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

Bayesian coresets

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

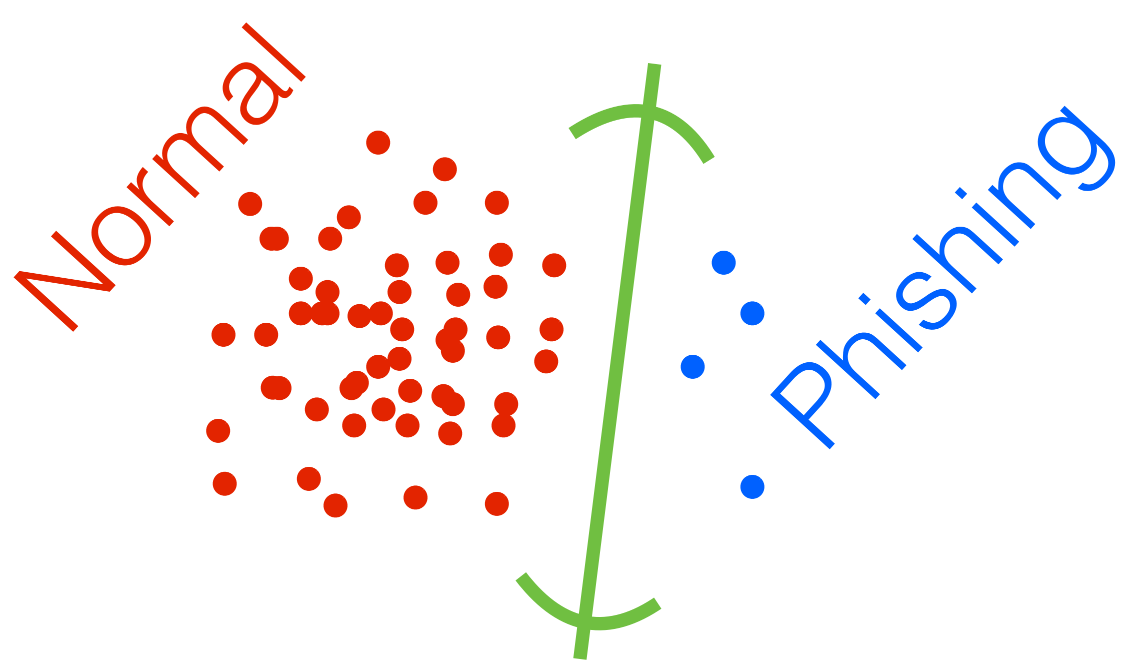


[Huggins, Campbell, Broderick 2016; Campbell, Broderick 2017, 2018]

Bayesian coresets

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

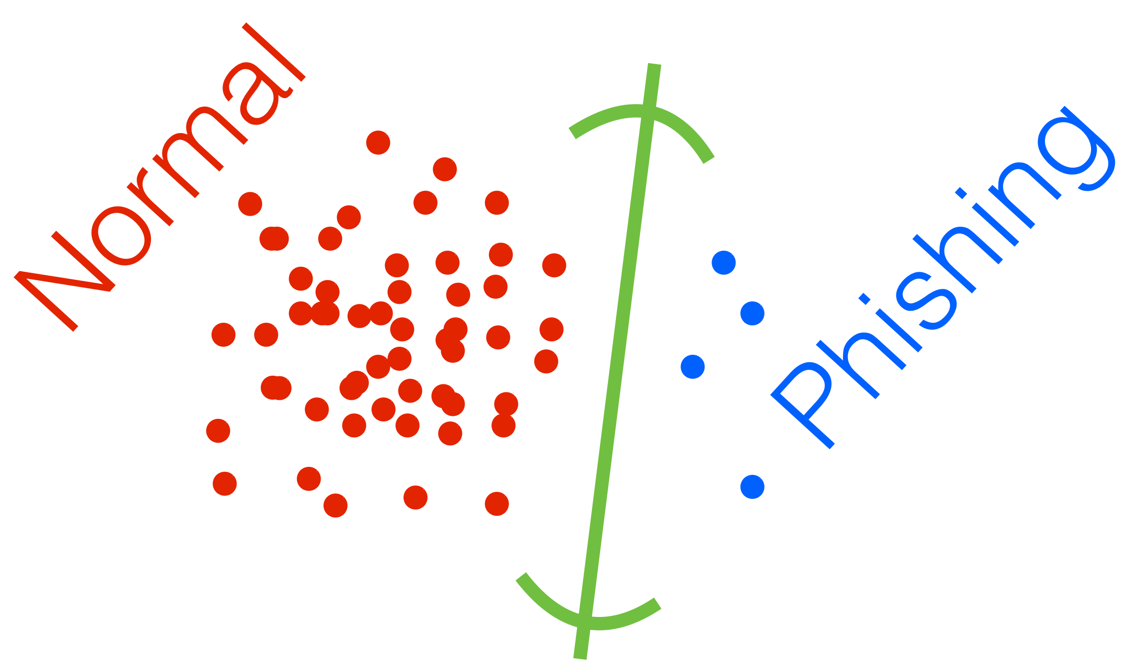
$$\|w\|_0 \ll N$$



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

Bayesian coresets

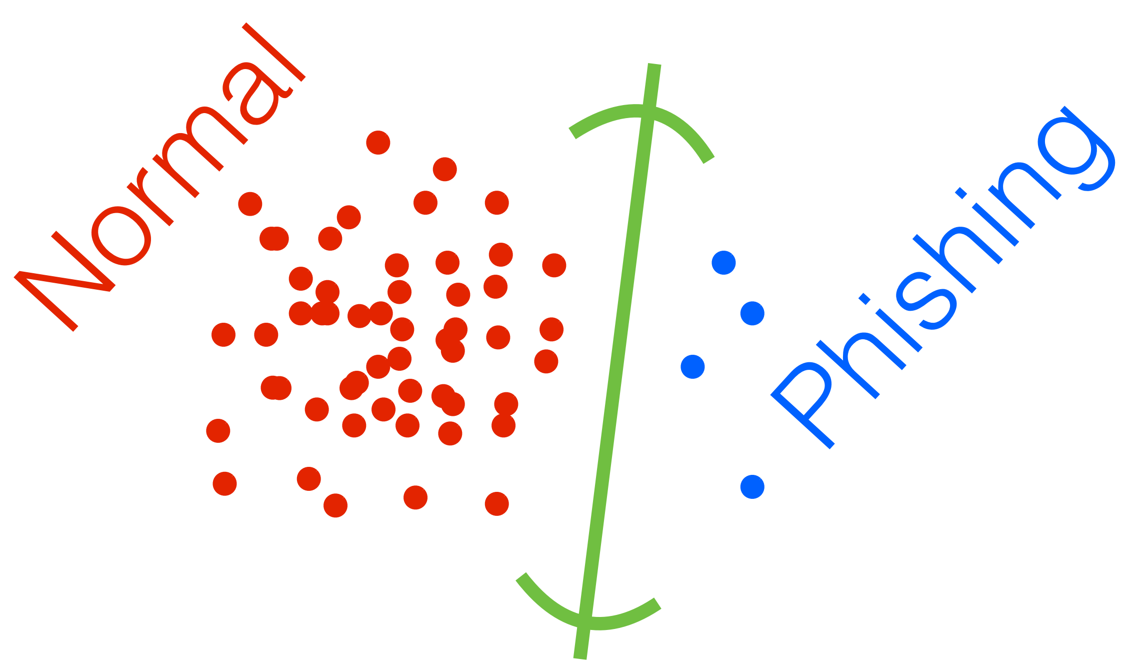
- Posterior $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$
- Log likelihood $\mathcal{L}_n(\theta) := \log p(y_n|\theta)\,, \;\; \mathcal{L}(\theta) := \sum_{n=1}^{N} \mathcal{L}_n(\theta)$
- Coreset log likelihood $\mathcal{L}(w,\theta) := \sum_{n=1}^{N} w_n \mathcal{L}_n(\theta)$ s.t. $\|w\|_0 \ll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

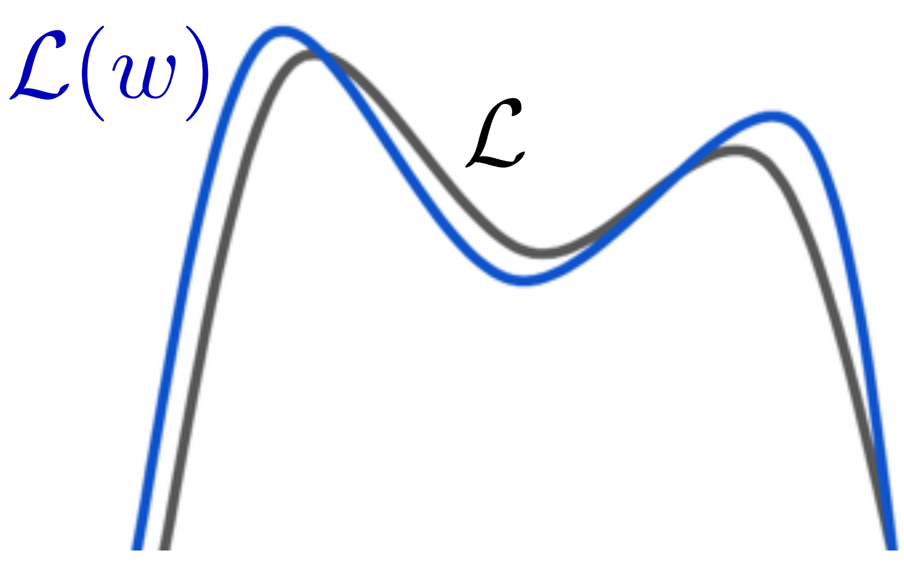


[Huggins, Campbell, Broderick 2016; Campbell, Broderick 2017, 2018]

Bayesian coresets

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

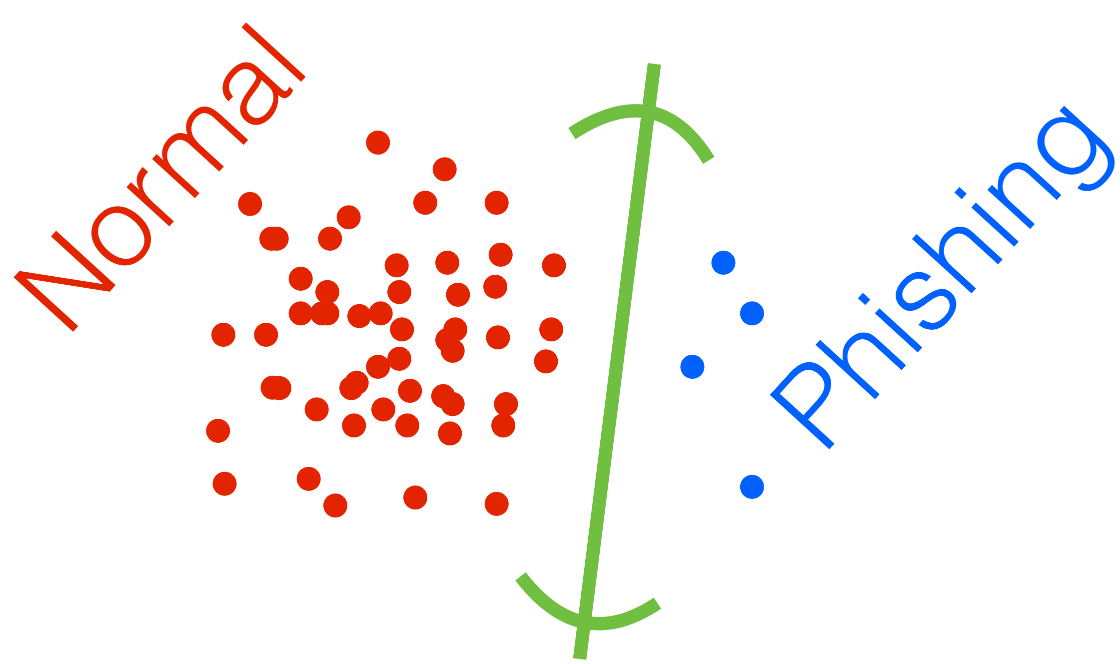


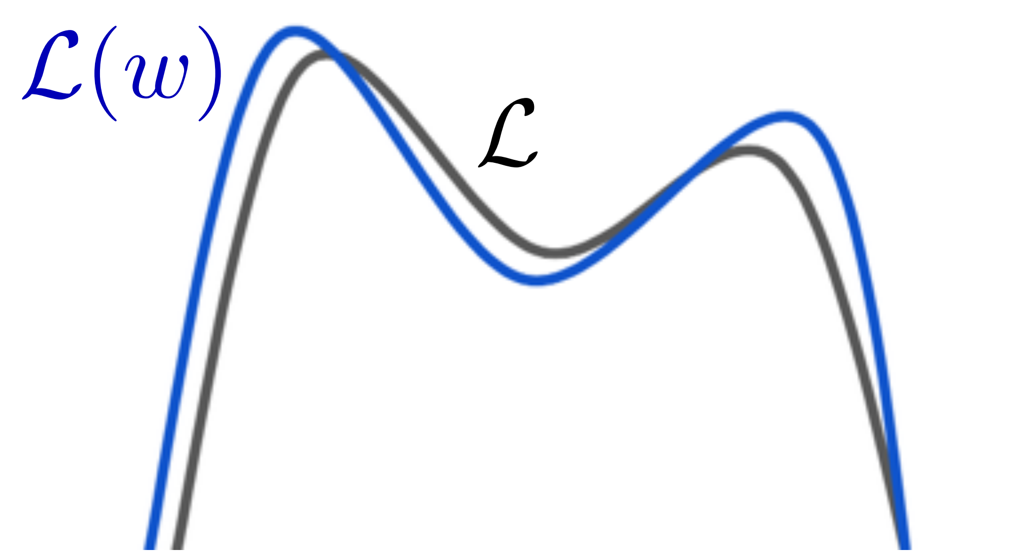


[Huggins, Campbell, Broderick 2016; Campbell, Broderick 2017, 2018]

Bayesian coresets

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

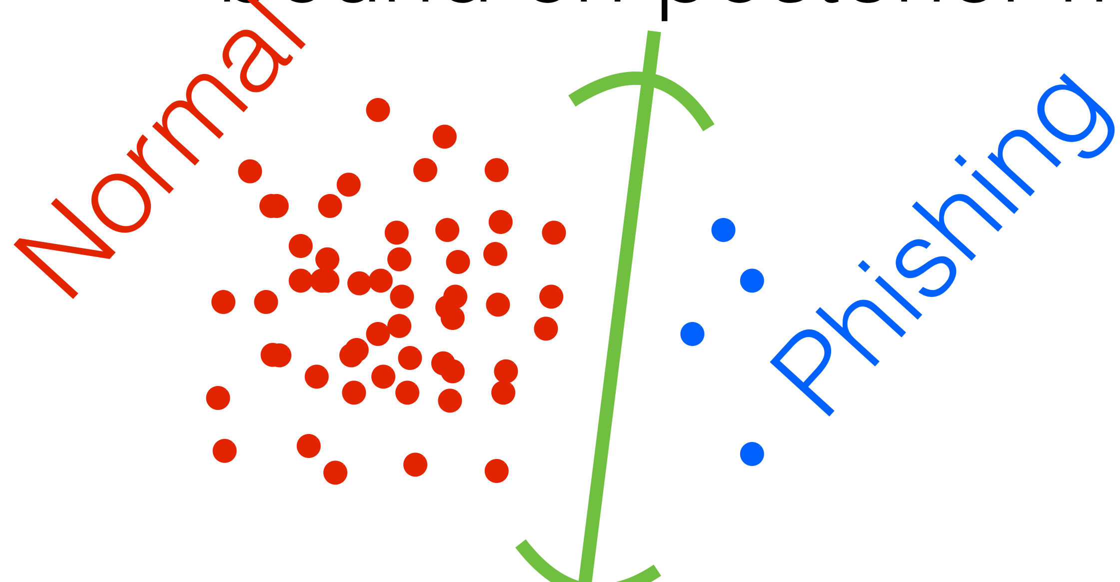


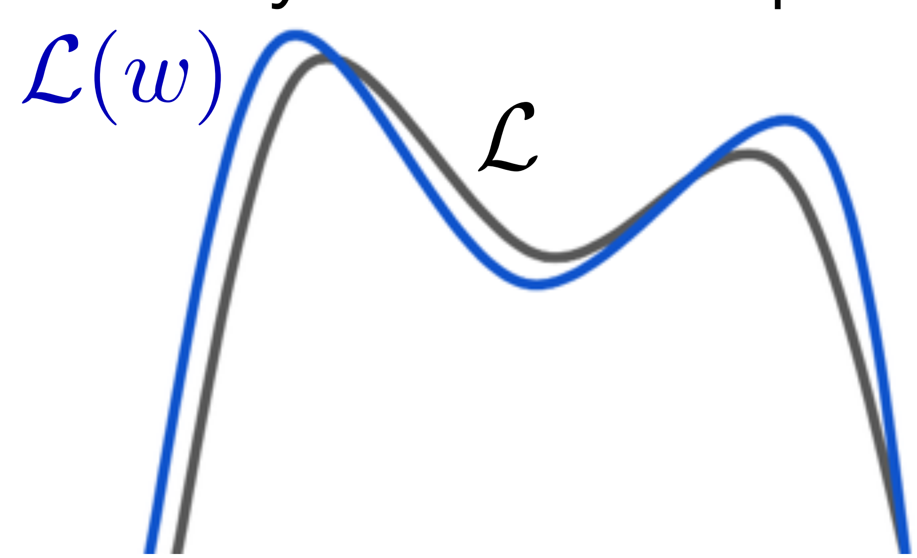


[Huggins, Campbell, Broderick 2016; Campbell, Broderick 2017, 2018]

Bayesian coresets

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





[Huggins, Campbell, Broderick 2016; Campbell, Broderick 2017, 2018; Huggins, Kasprzak, Campbell, Broderick, in preparation]

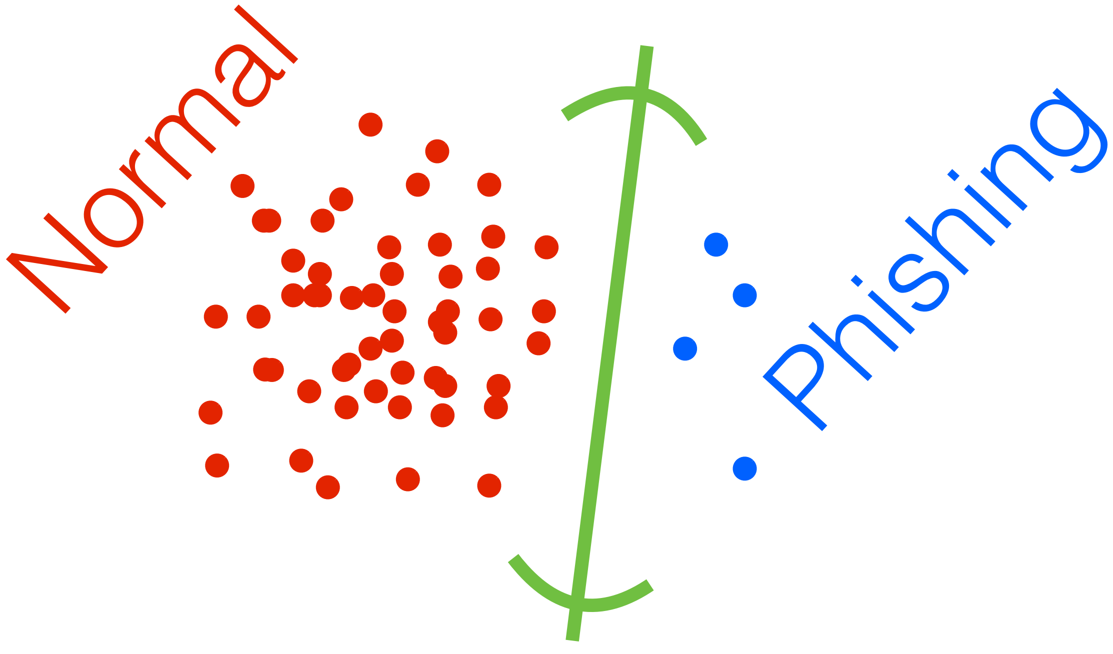
Roadmap

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

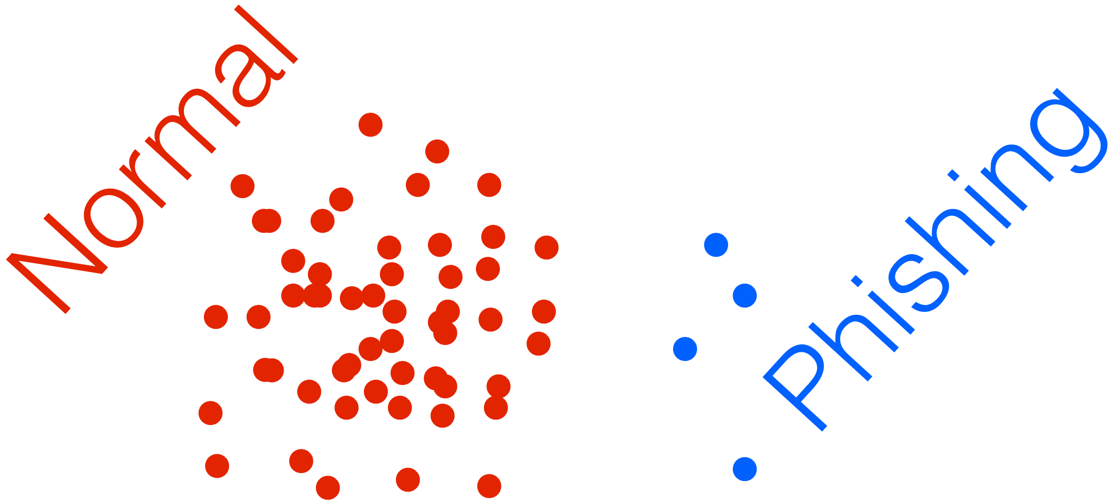
Roadmap

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

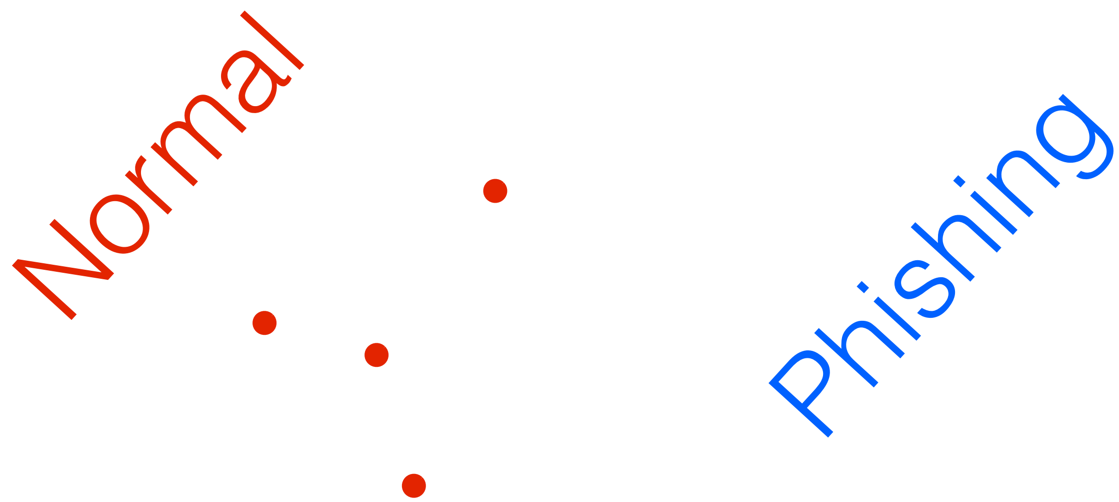
Uniform subsampling revisited



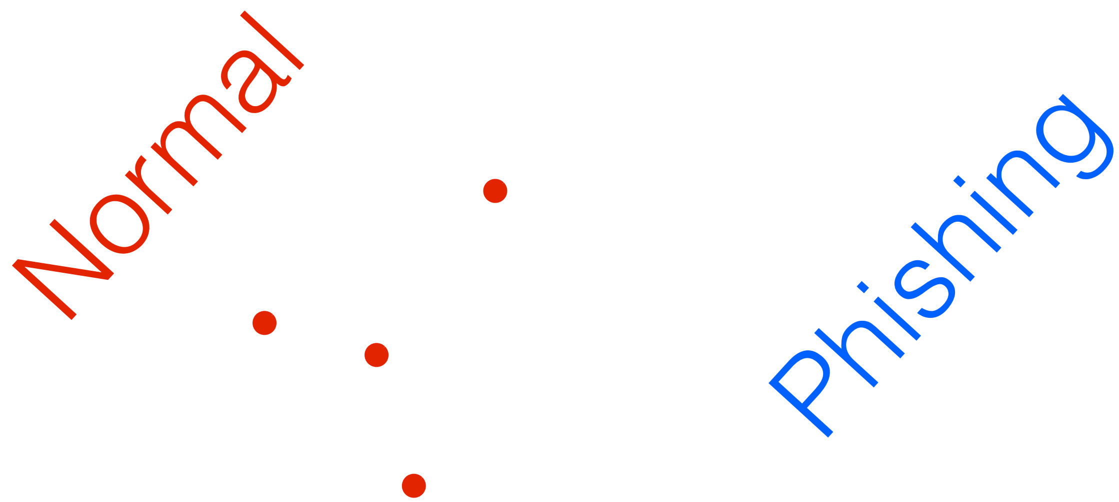
Uniform subsampling revisited



Uniform subsampling revisited

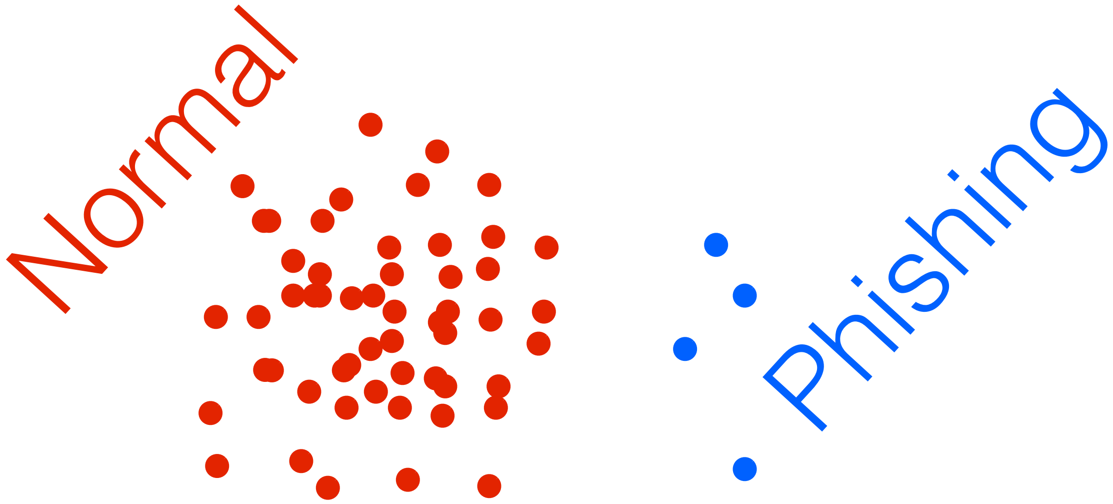


Uniform subsampling revisited



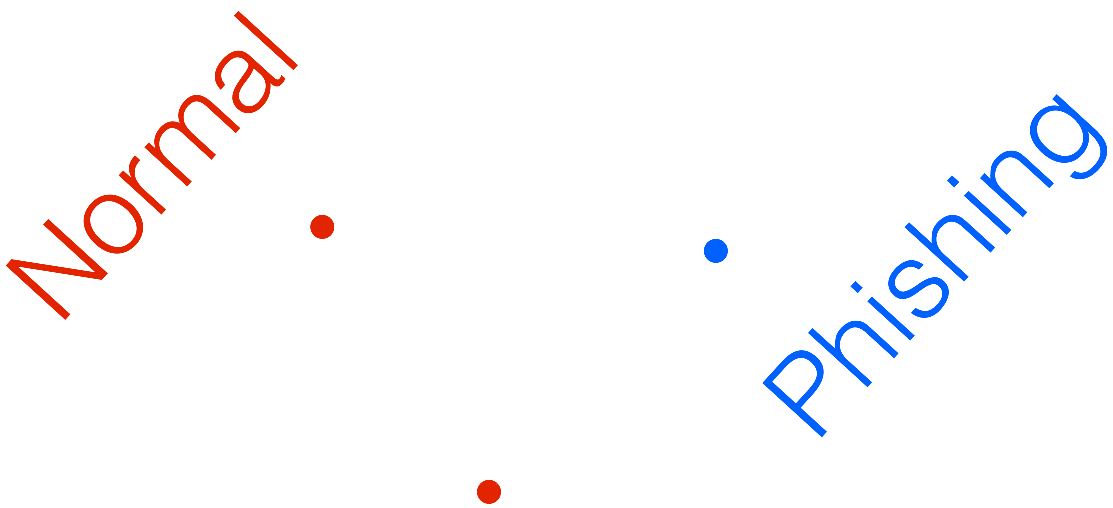
- Might miss important data

Uniform subsampling revisited



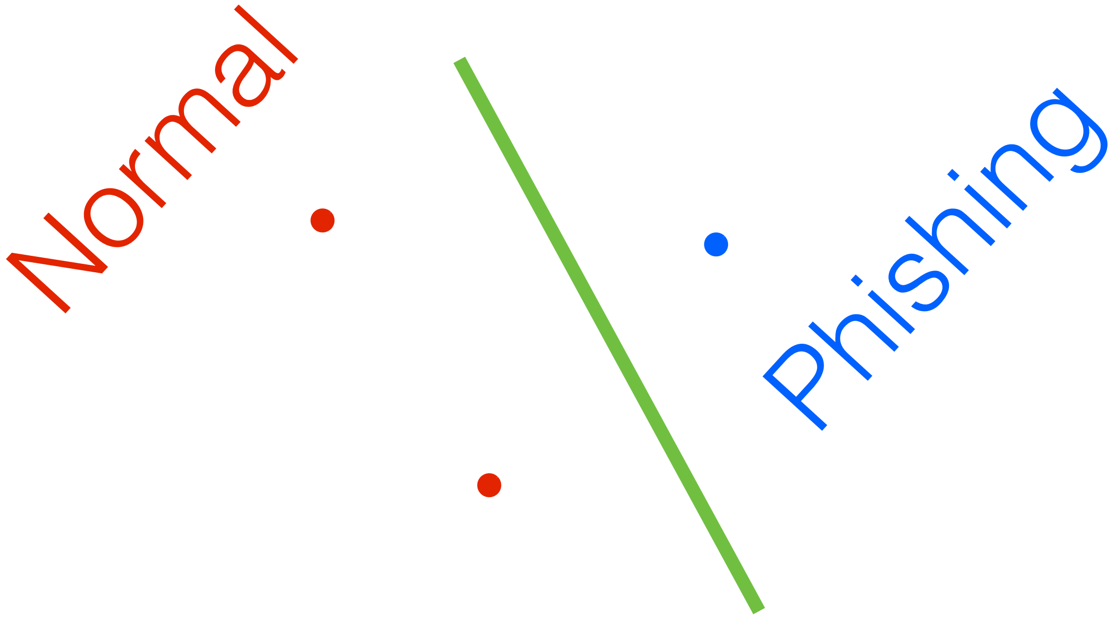
- Might miss important data

Uniform subsampling revisited



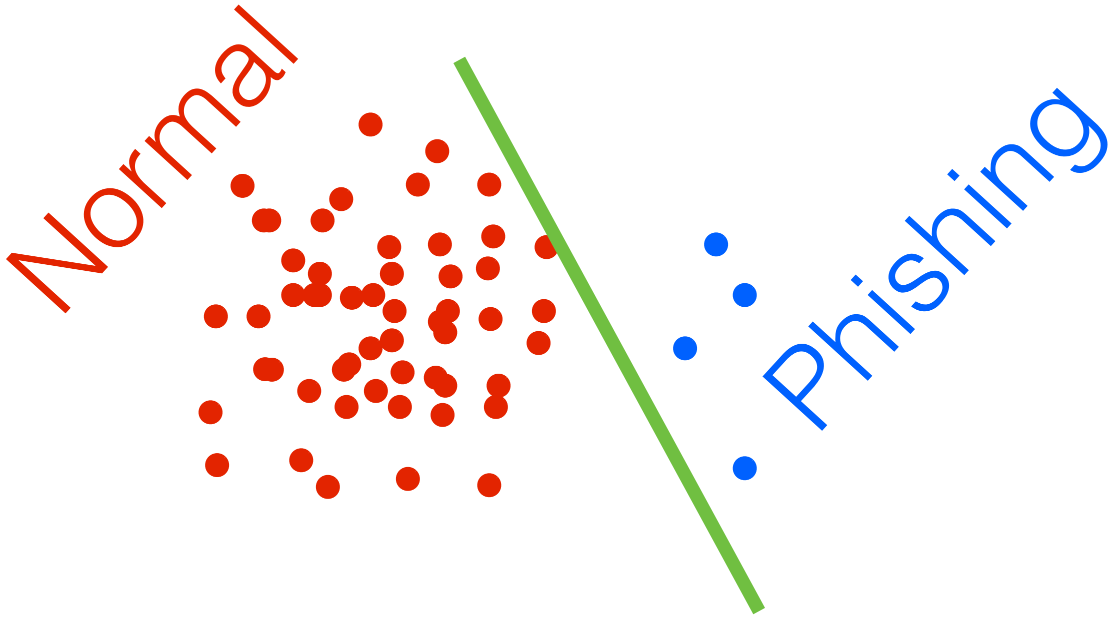
- Might miss important data

Uniform subsampling revisited



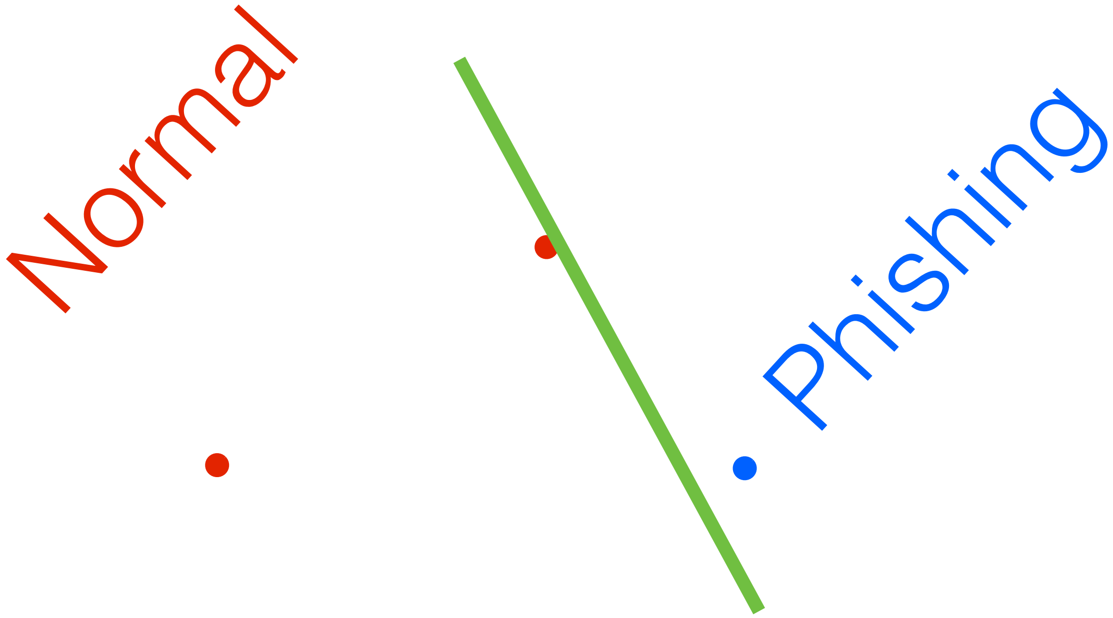
- Might miss important data

Uniform subsampling revisited



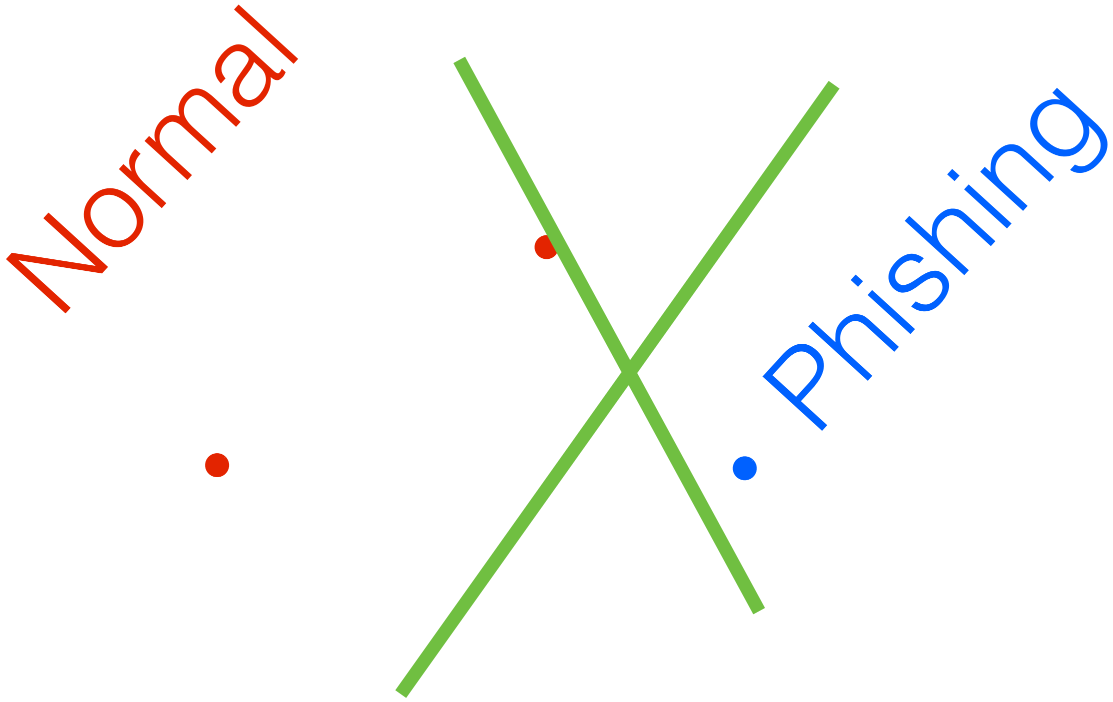
- Might miss important data

Uniform subsampling revisited



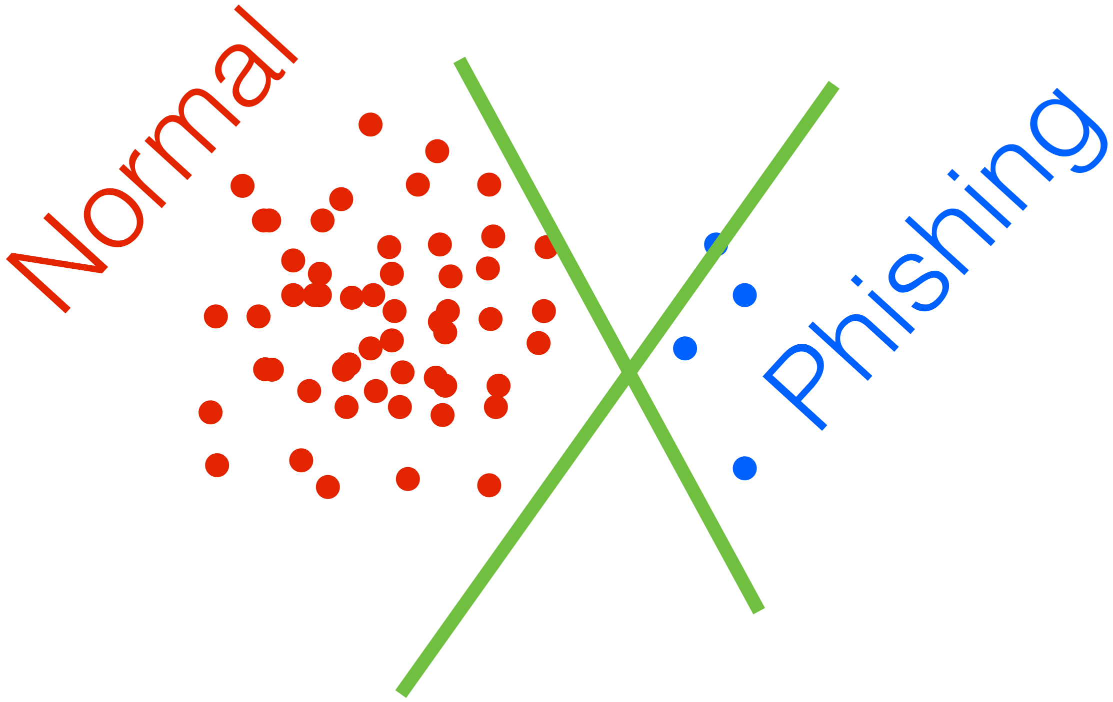
- Might miss important data

Uniform subsampling revisited



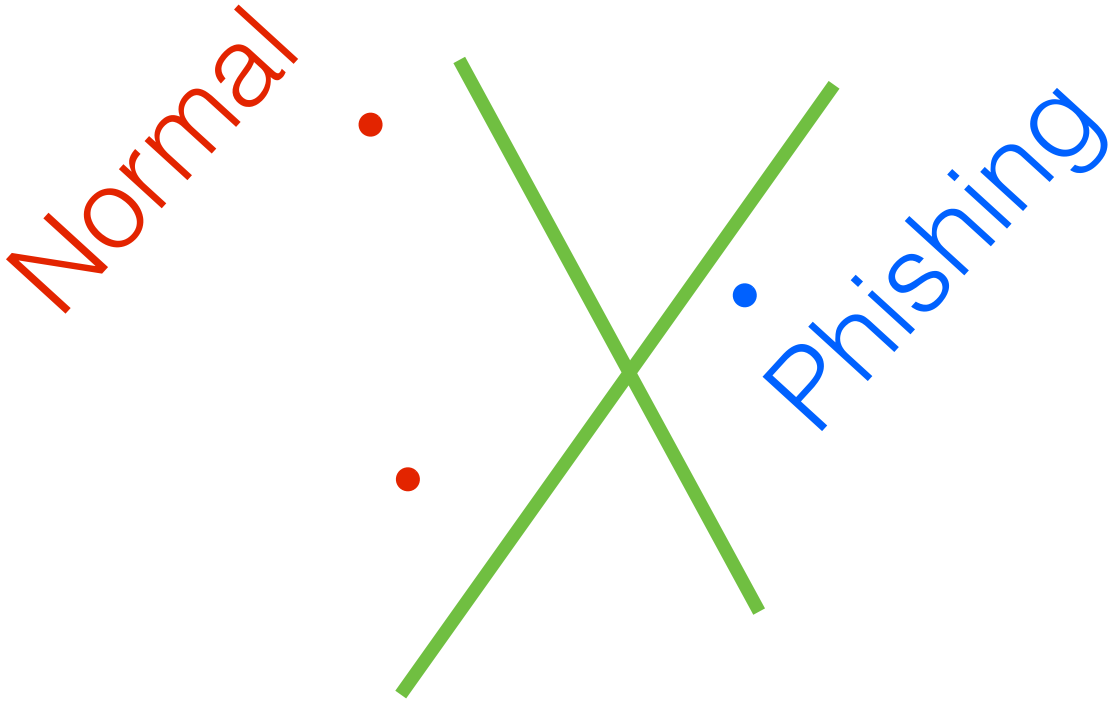
- Might miss important data

Uniform subsampling revisited



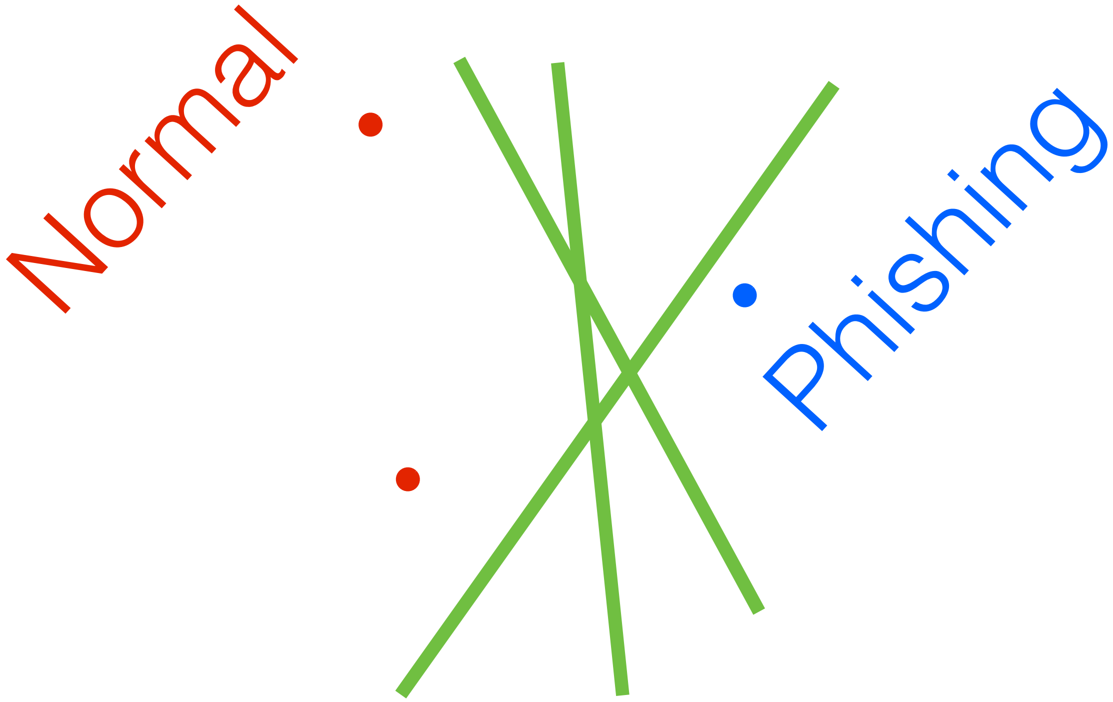
- Might miss important data

Uniform subsampling revisited



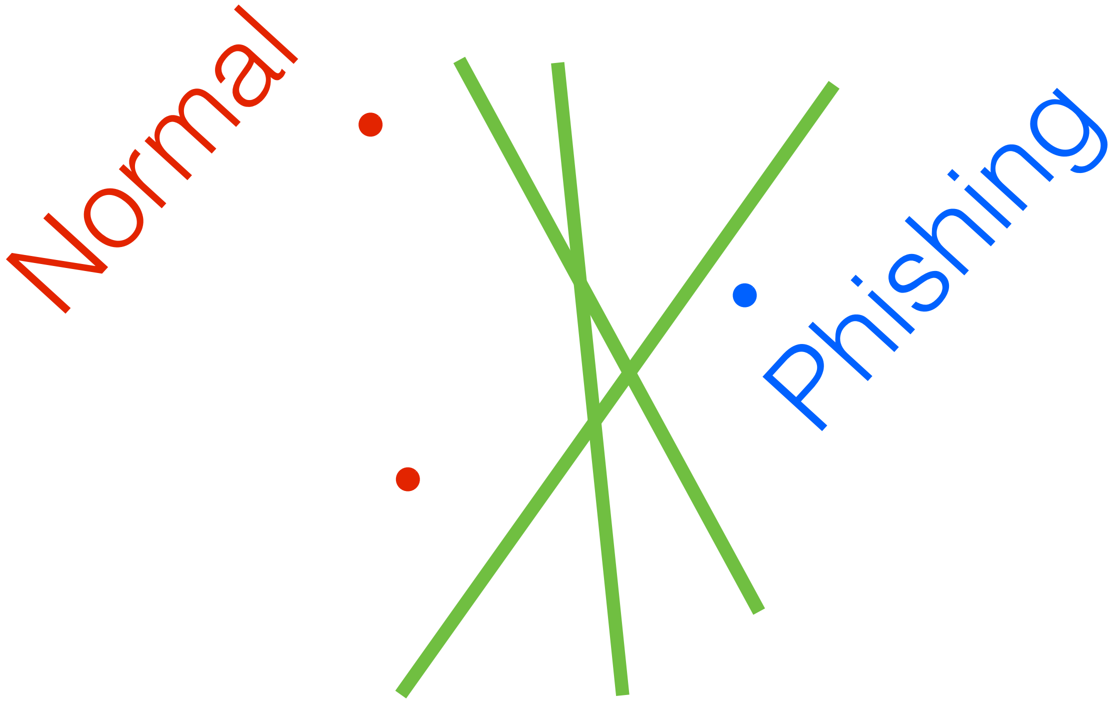
- Might miss important data

Uniform subsampling revisited



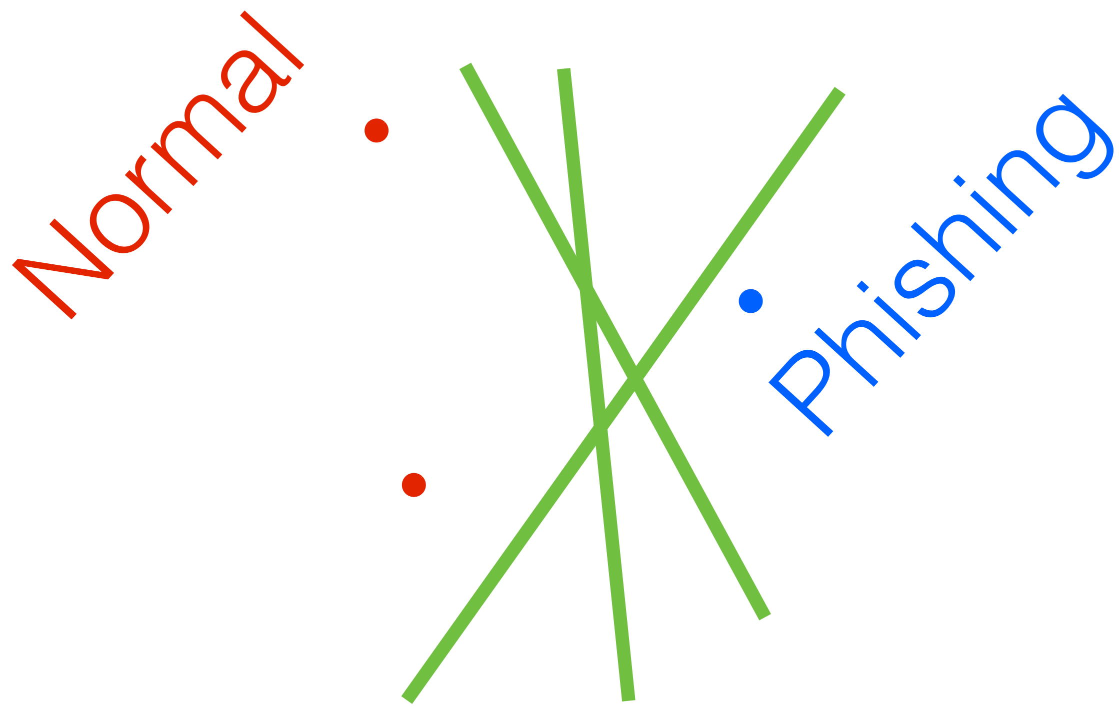
- Might miss important data

Uniform subsampling revisited

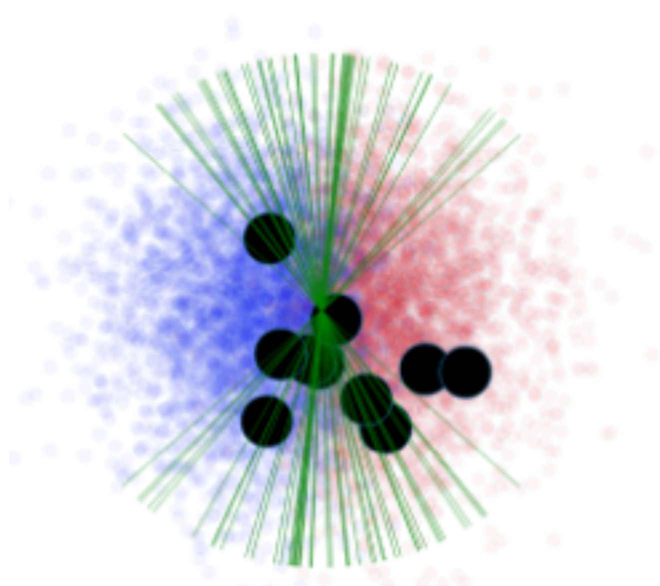


- Might miss important data
- Noisy estimates

Uniform subsampling revisited

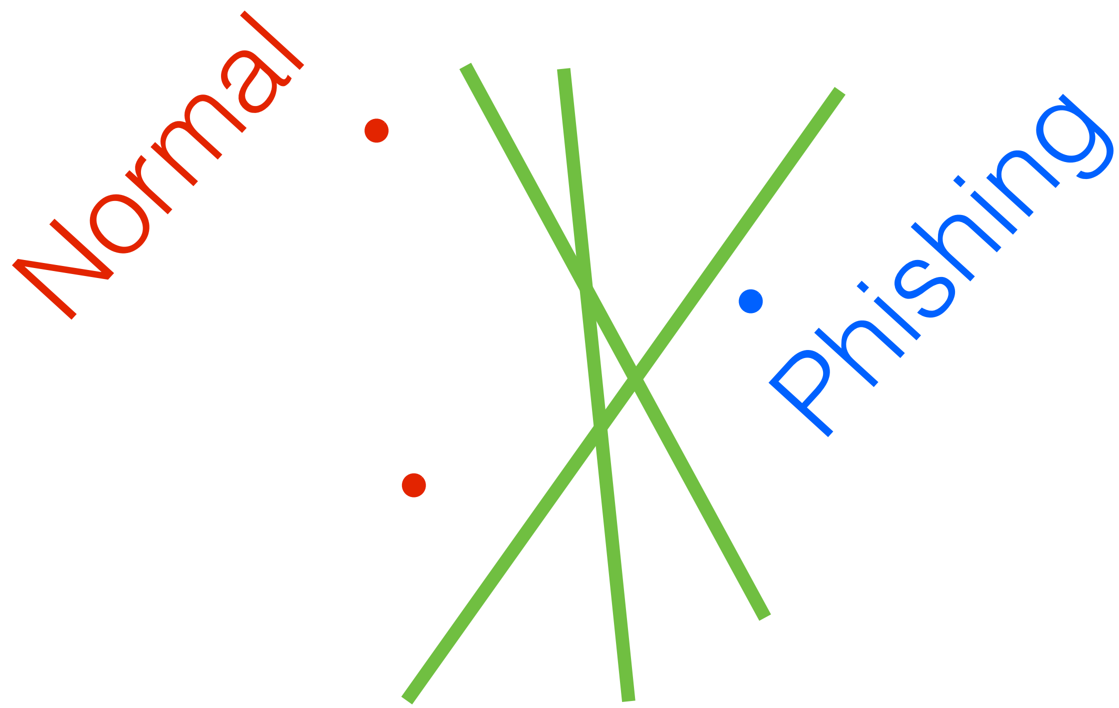


- Might miss important data
- Noisy estimates

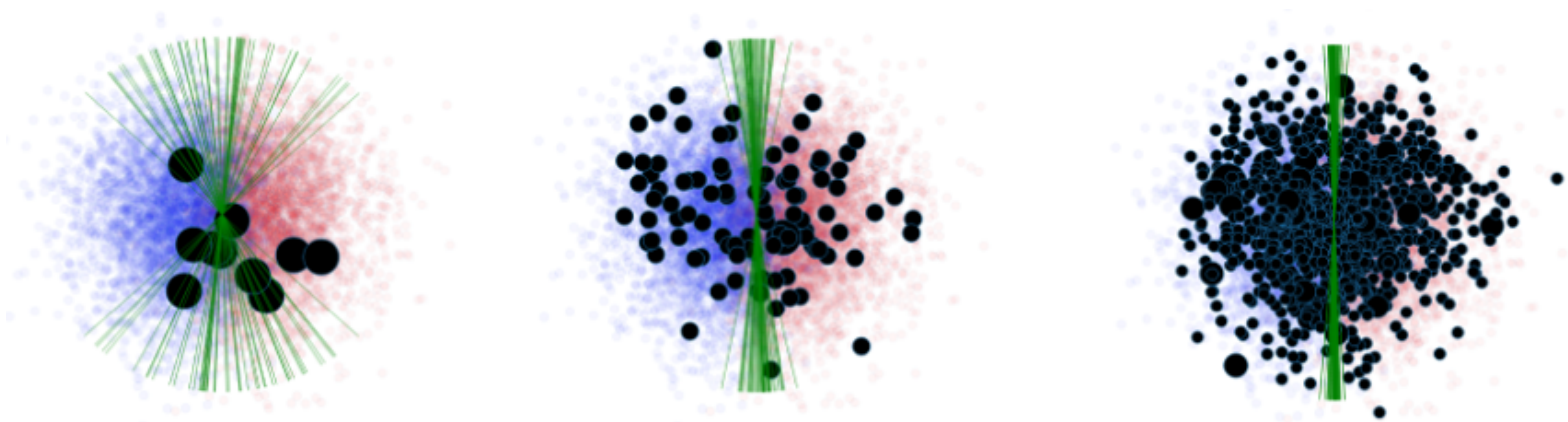


$M = 10$

Uniform subsampling revisited



- Might miss important data
- Noisy estimates



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

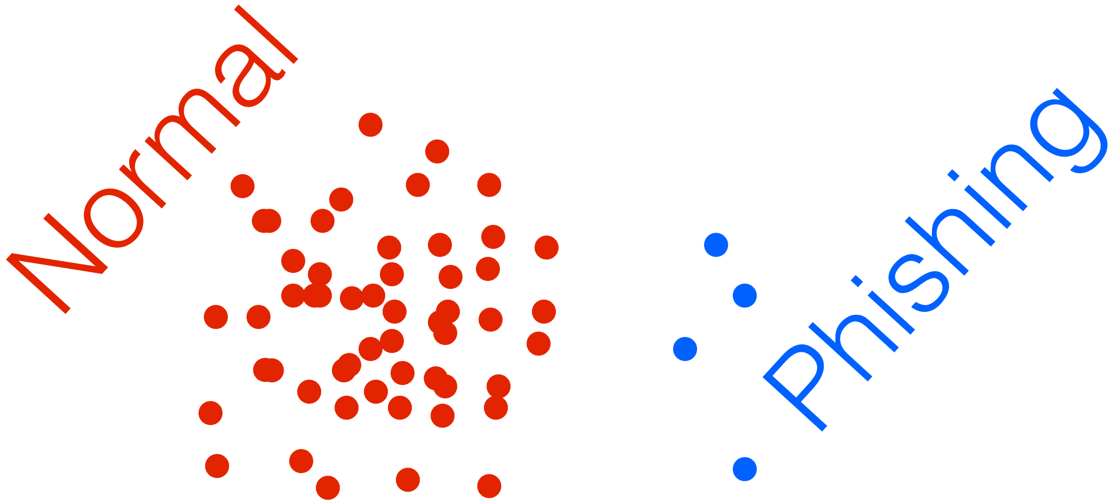
Roadmap

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

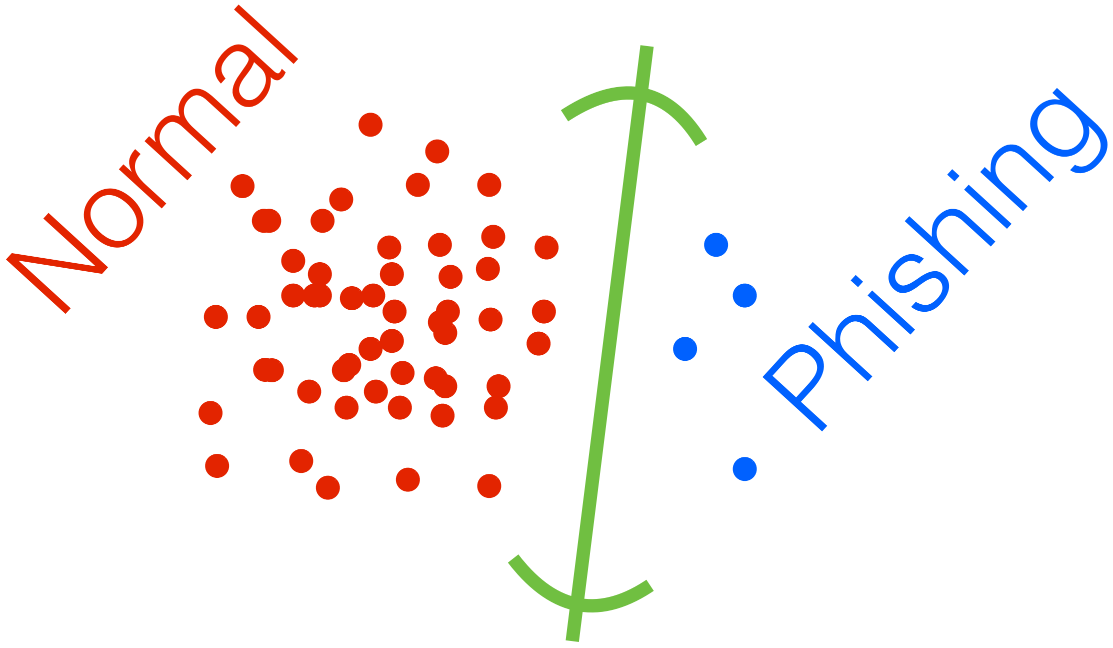
Roadmap

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

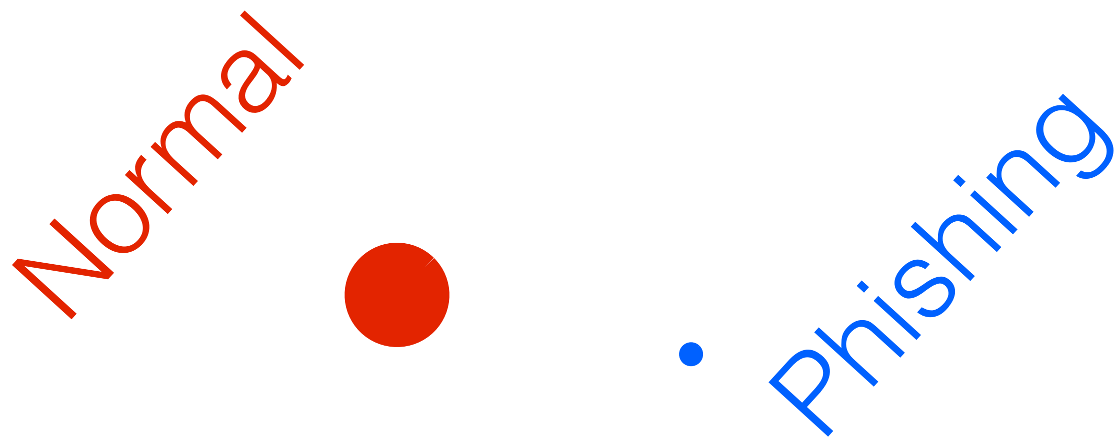
Importance sampling



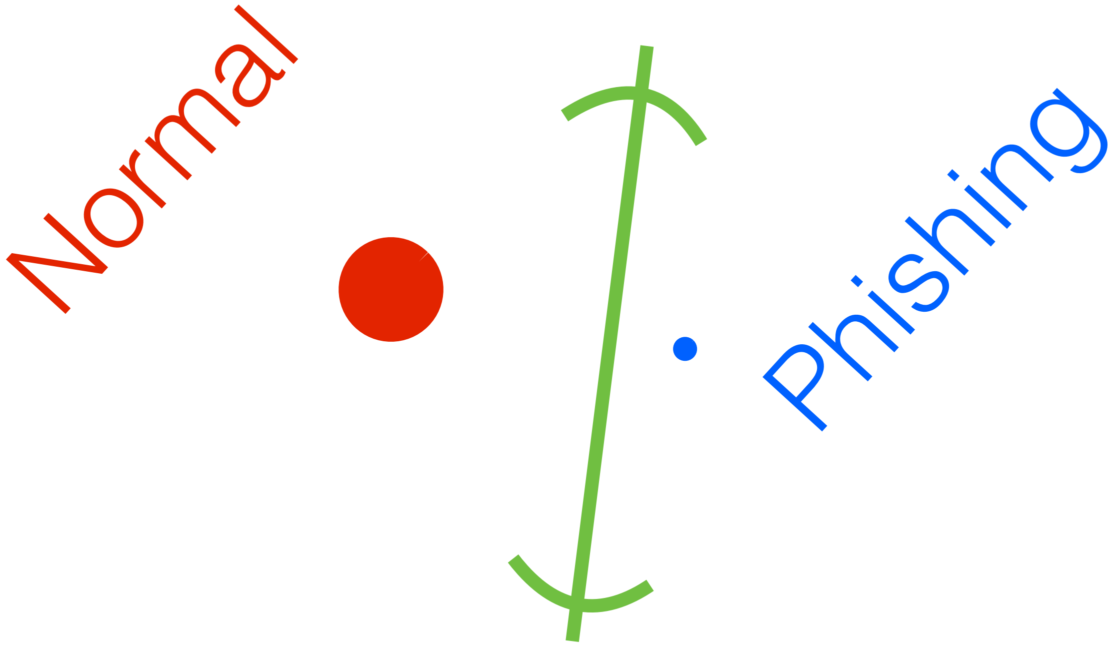
Importance sampling



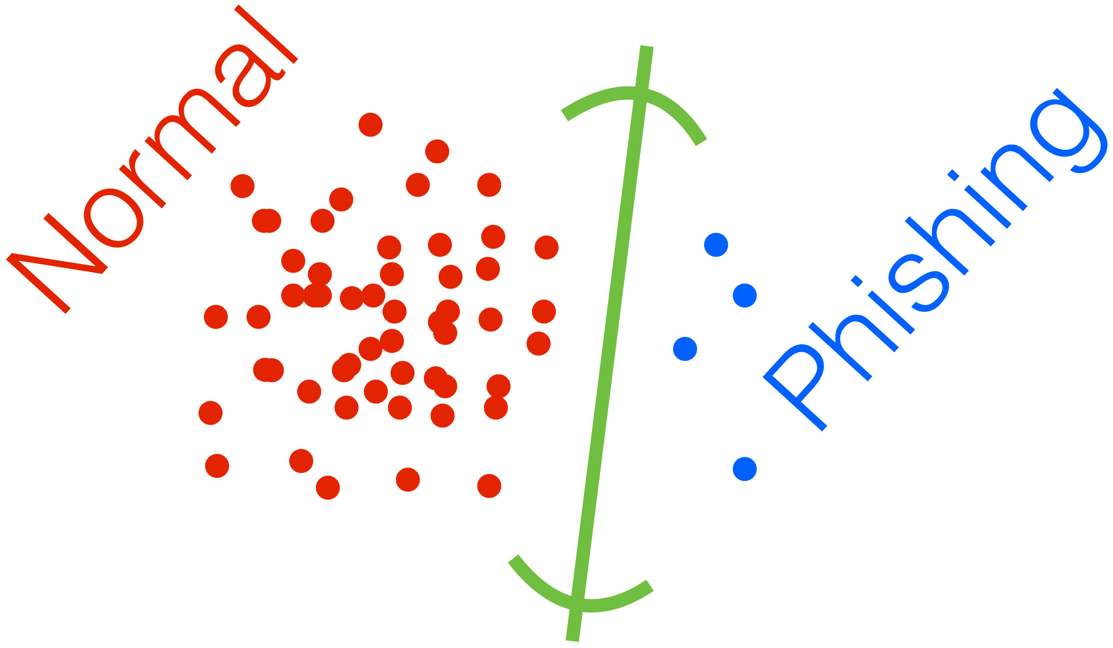
Importance sampling



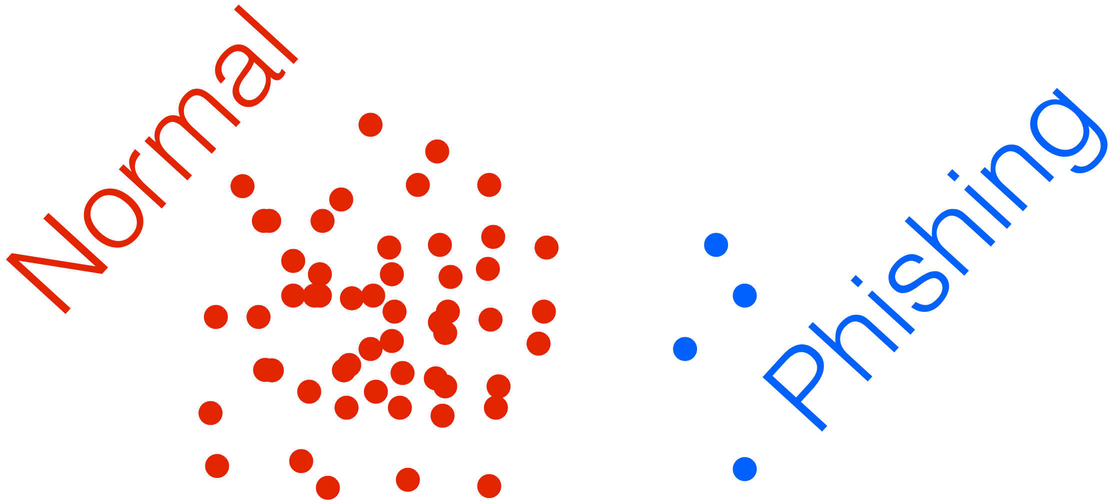
Importance sampling



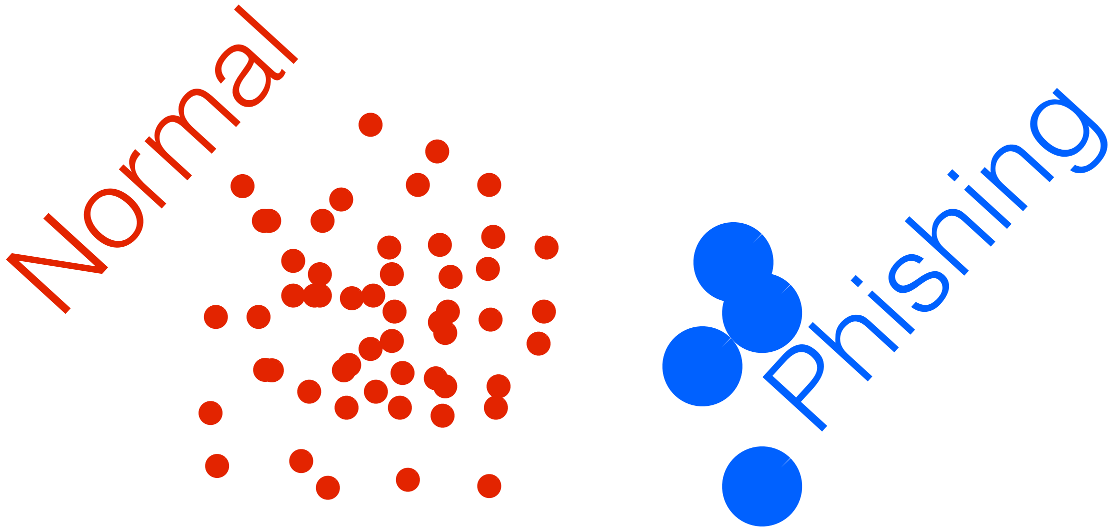
Importance sampling



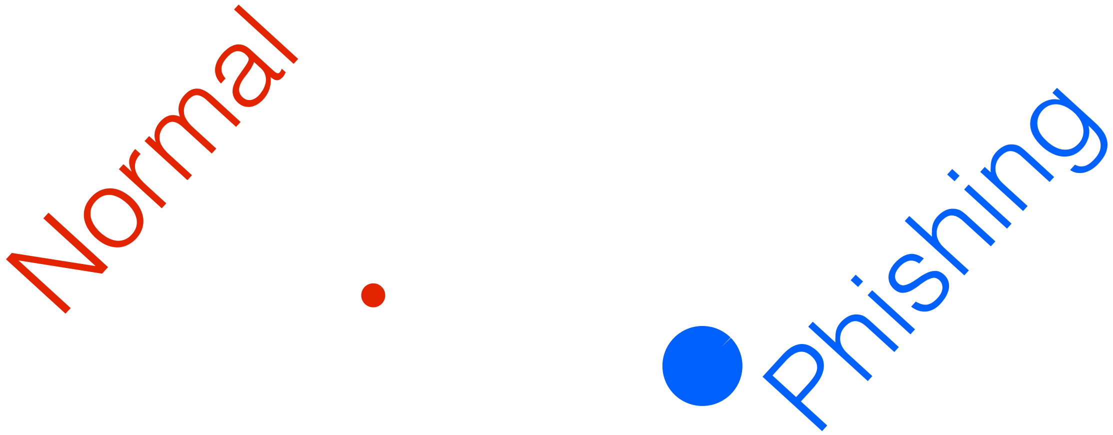
Importance sampling



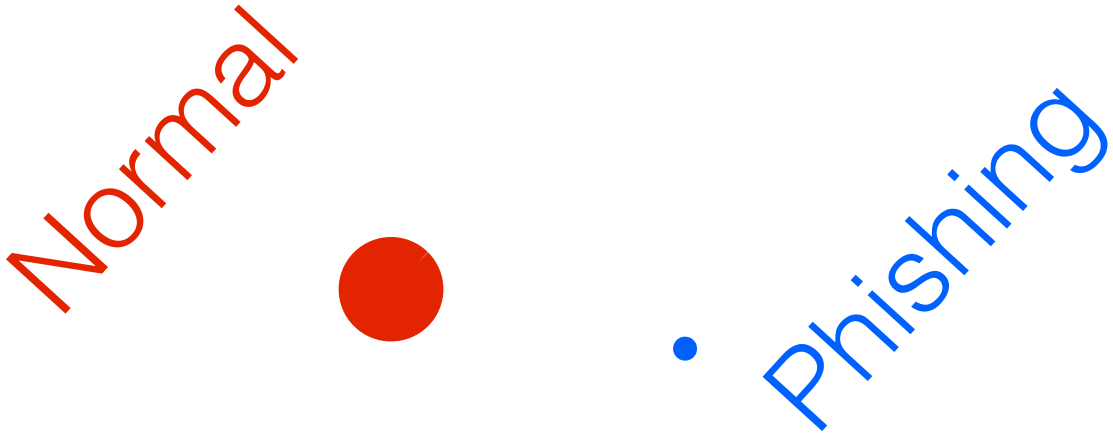
Importance sampling



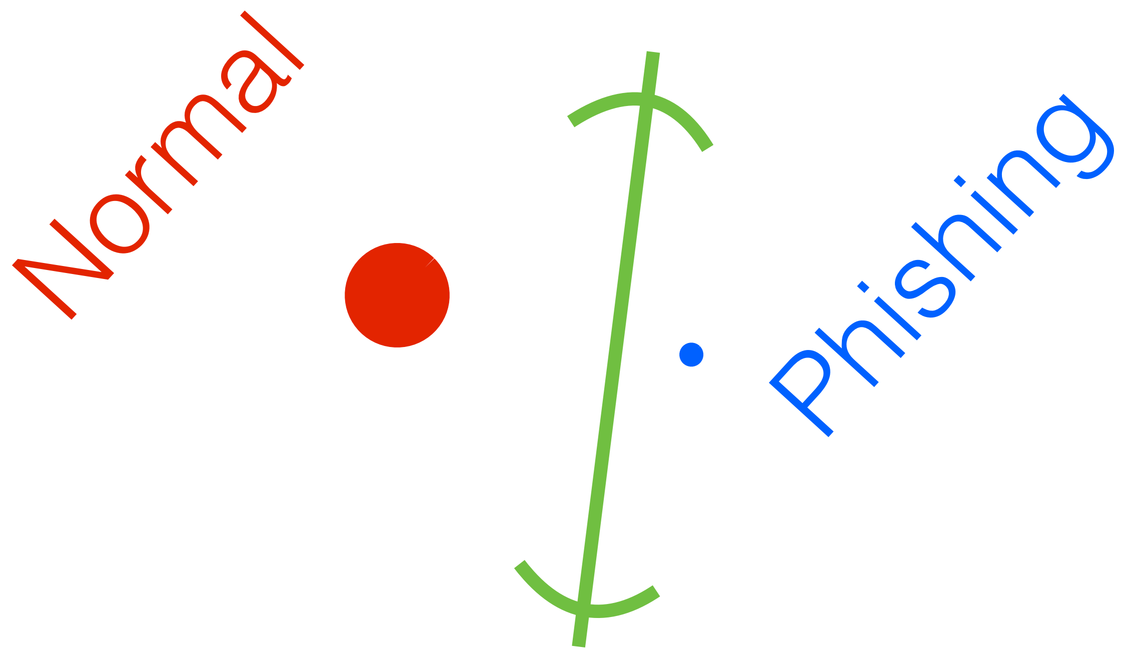
Importance sampling



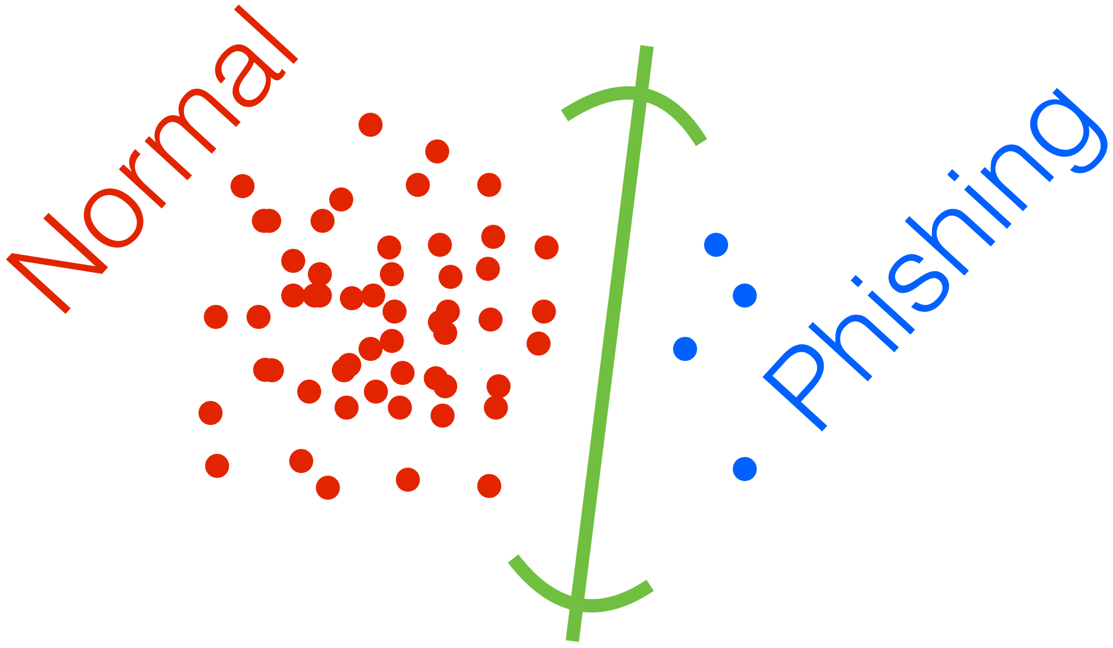
Importance sampling



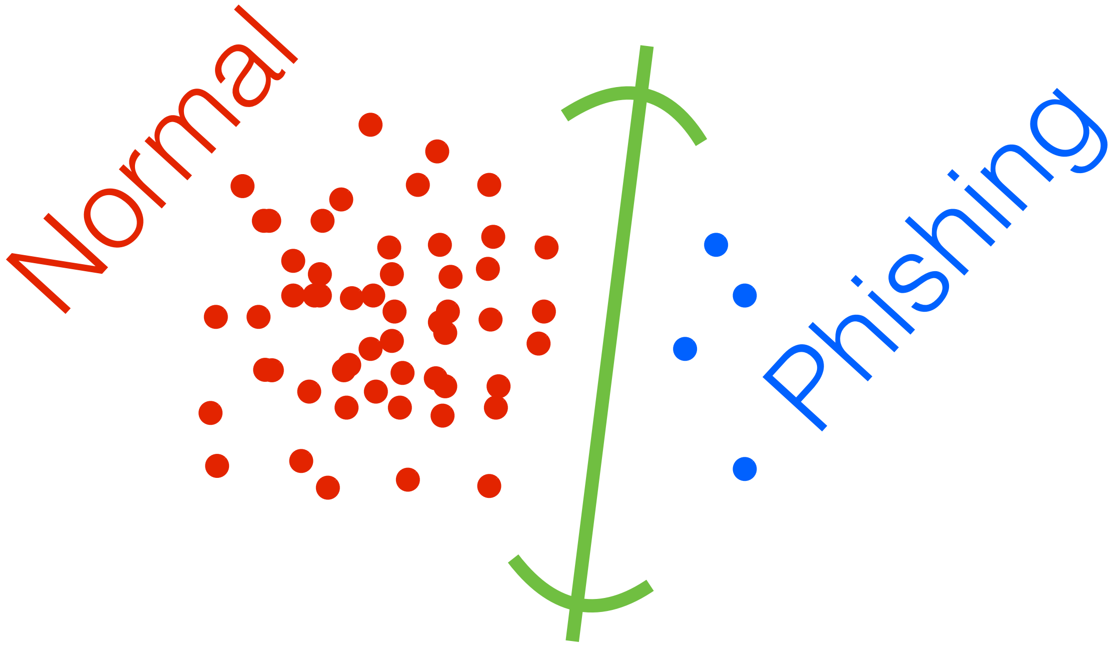
Importance sampling



Importance sampling

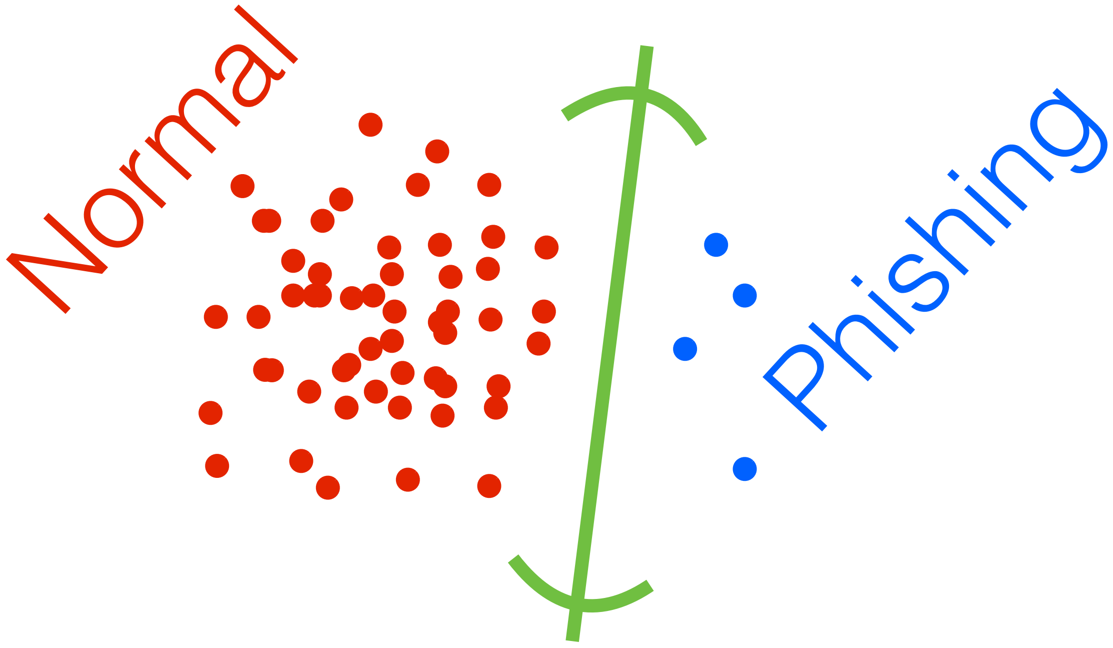


Importance sampling



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

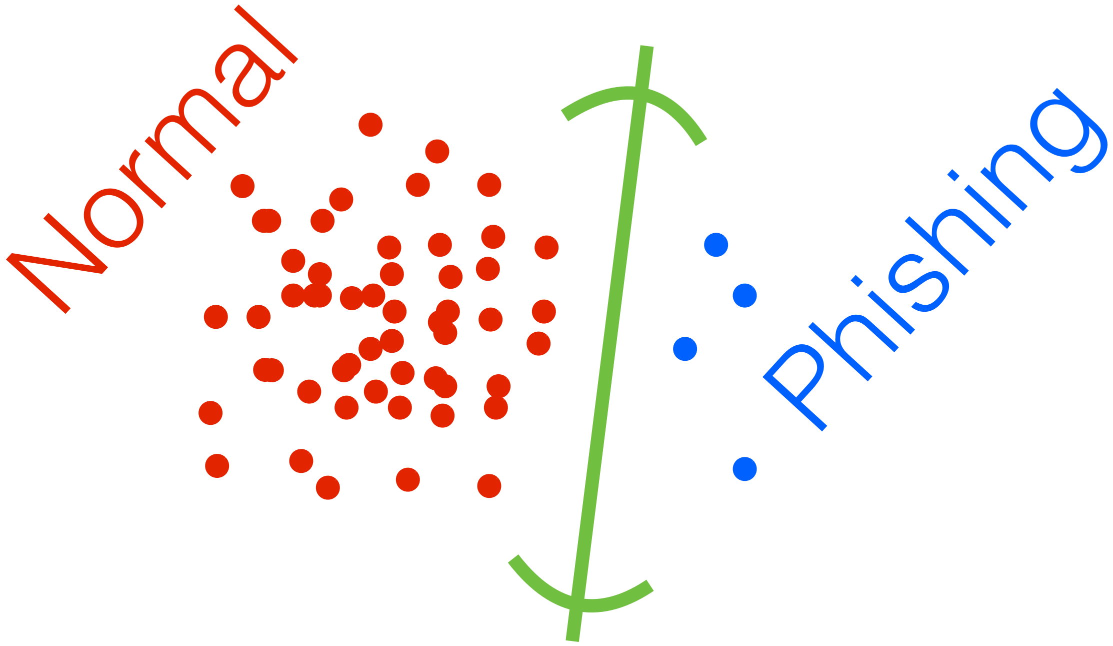
Importance sampling



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

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

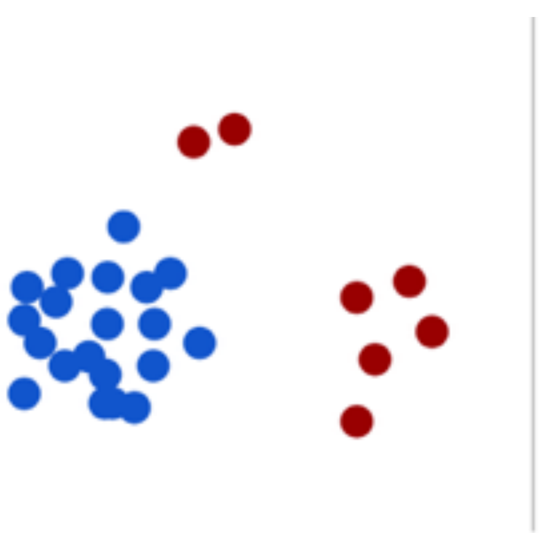
Importance sampling



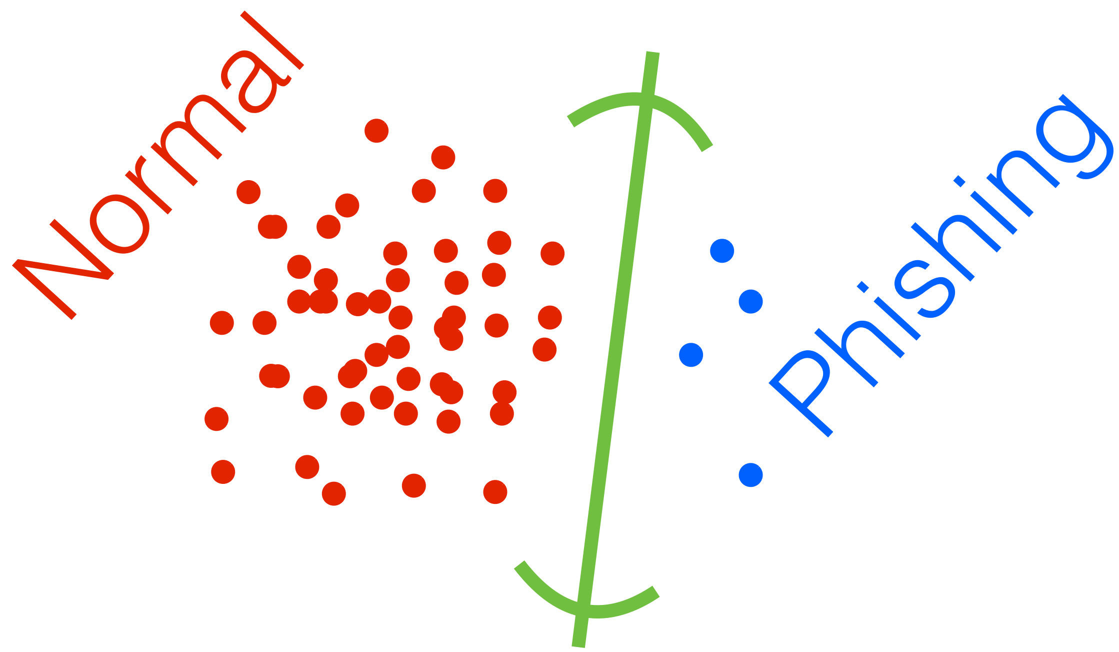
$$\sigma := \sum_{n=1}^{N} \|\mathcal{L}_n\|$$

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

1. data



Importance sampling

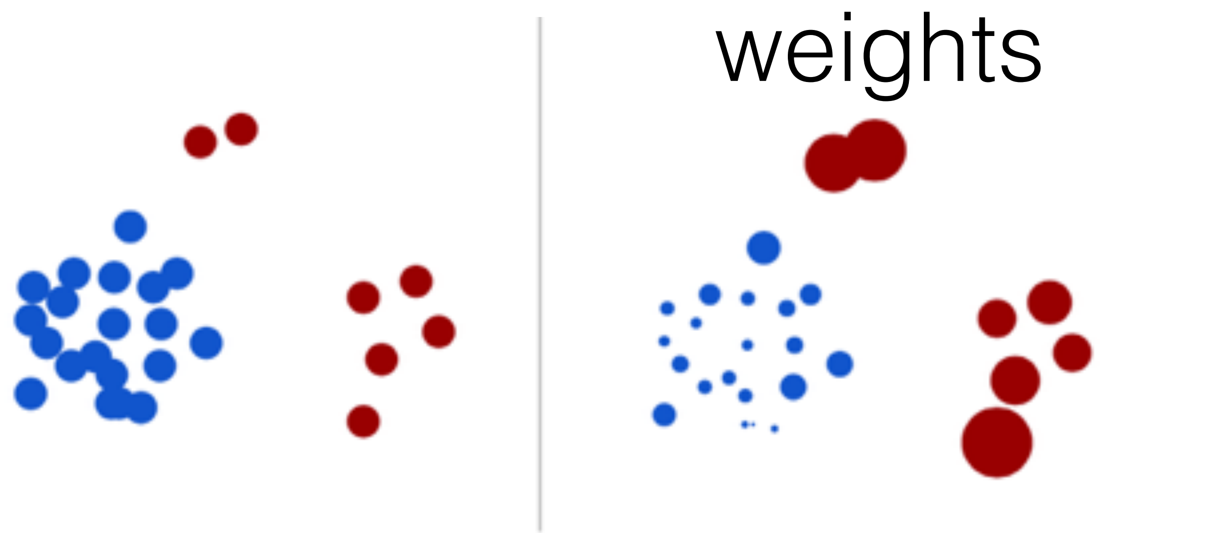


$$\sigma := \sum_{n=1}^{N} \|\mathcal{L}_n\|$$

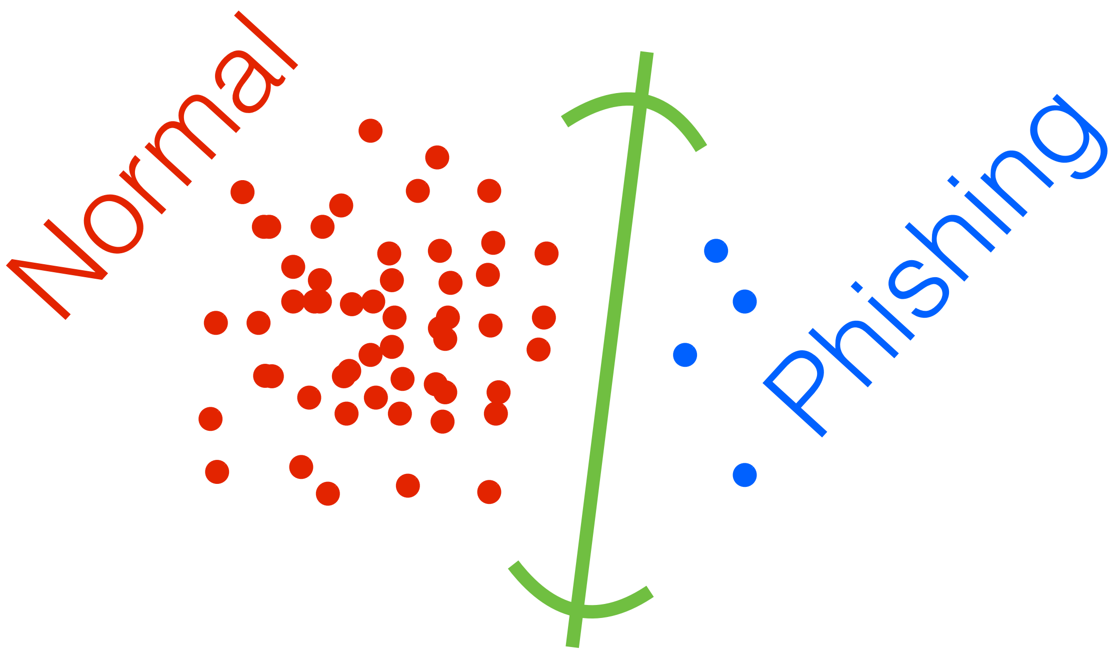
$$\sigma_n := \|\mathcal{L}_n\|/\sigma$$

1. data

2. importance weights

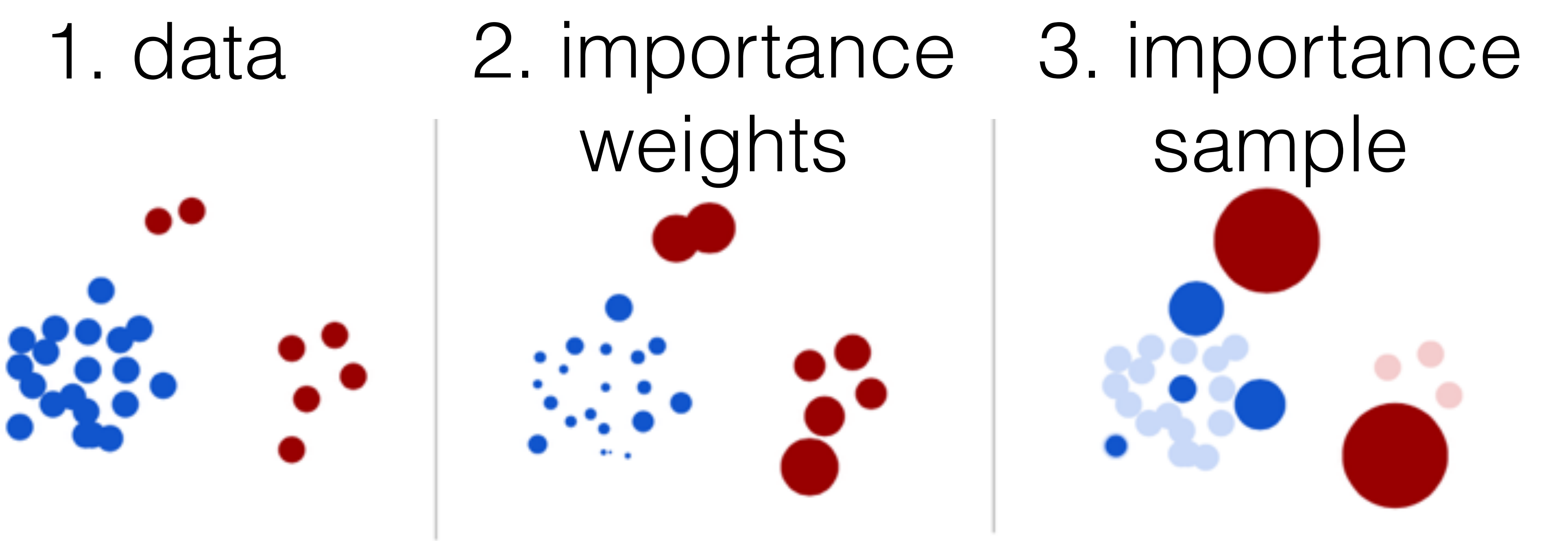


Importance sampling

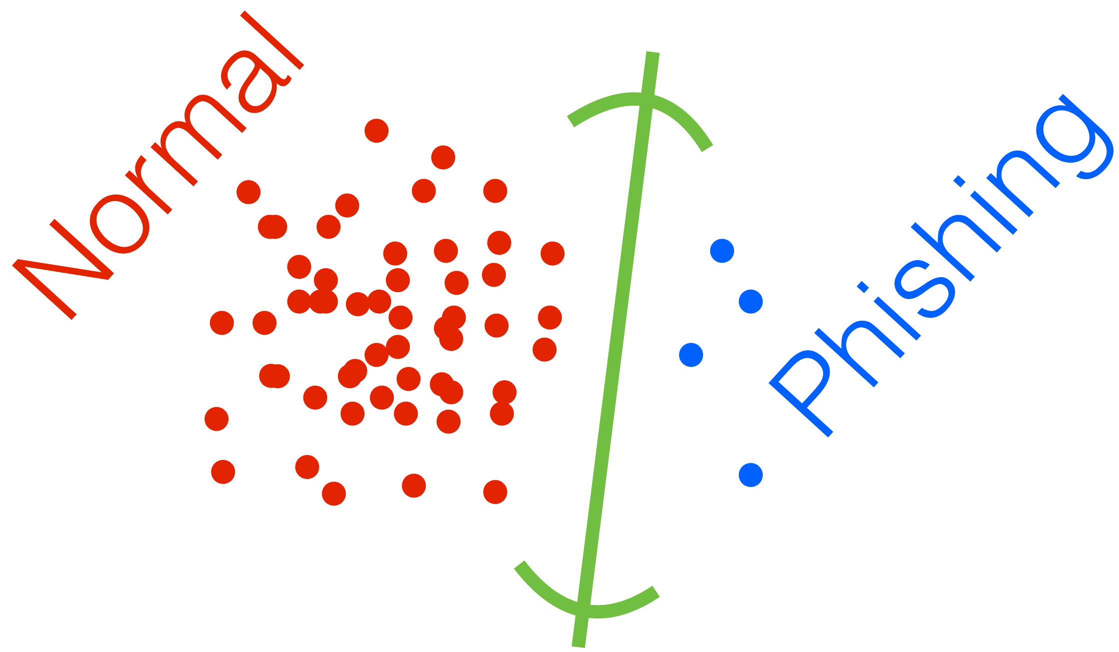


$$\sigma := \sum_{n=1}^{N} \|\mathcal{L}_n\|$$

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

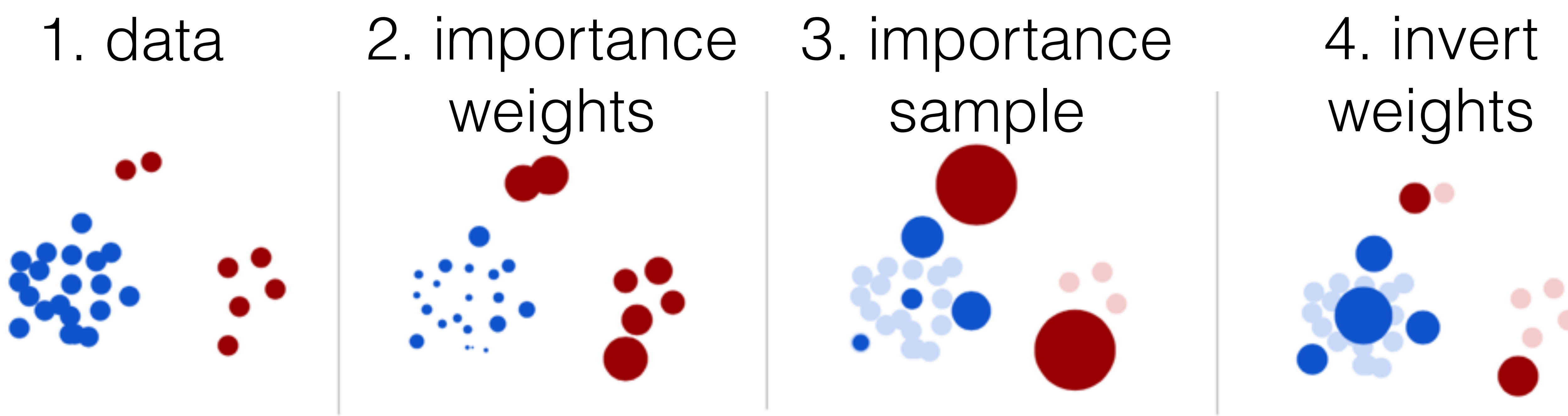


Importance sampling



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

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



Importance sampling

Thm (Campbell, B). δ ∈ (0,1). W.p. ≥1 - δ, after *M* iterations,

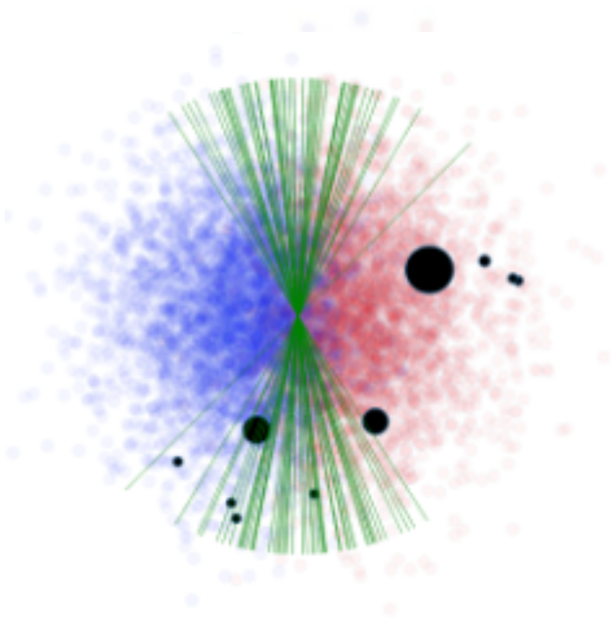
$$\|\mathcal{L}(w) - \mathcal{L}\| \leq \frac{\sigma\bar{\eta}}{\sqrt{M}}\left(1 + \sqrt{2\log\frac{1}{\delta}}\right)$$

Importance sampling

Thm (Campbell, B). δ ∈ (0,1). W.p. ≥1 - δ, after *M* iterations,

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

- Still noisy estimates



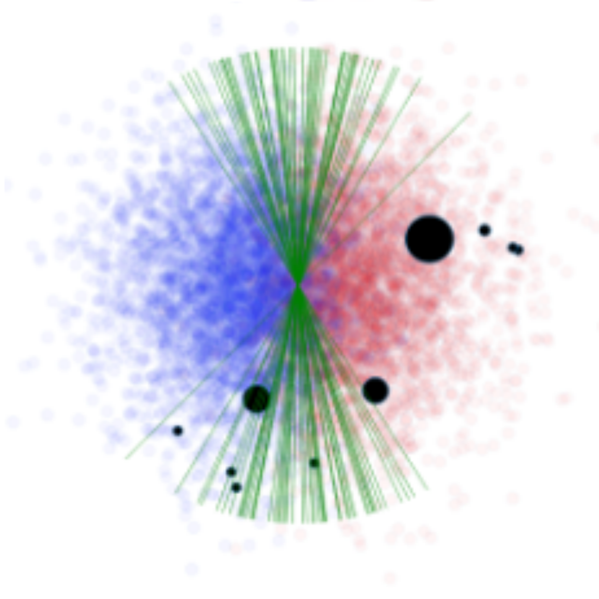
M = 10

Importance sampling

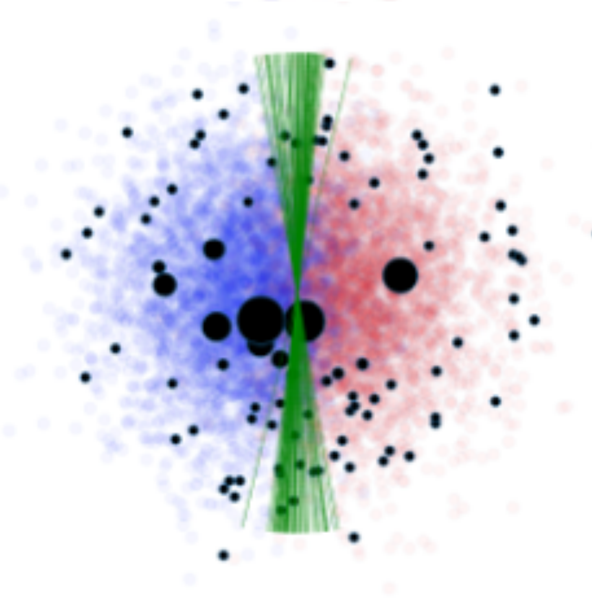
Thm (Campbell, B). δ ∈ (0,1). W.p. ≥1 - δ, after *M* iterations,

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

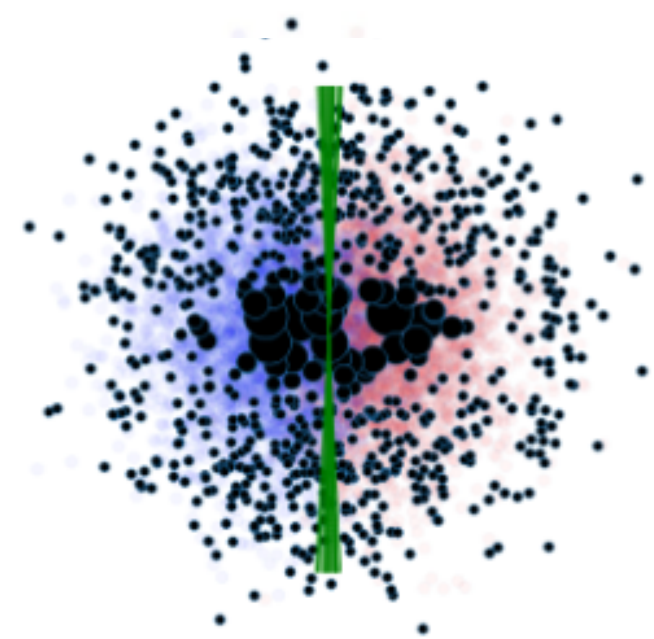
- Still noisy estimates



M = 10



M = 100



M = 1000

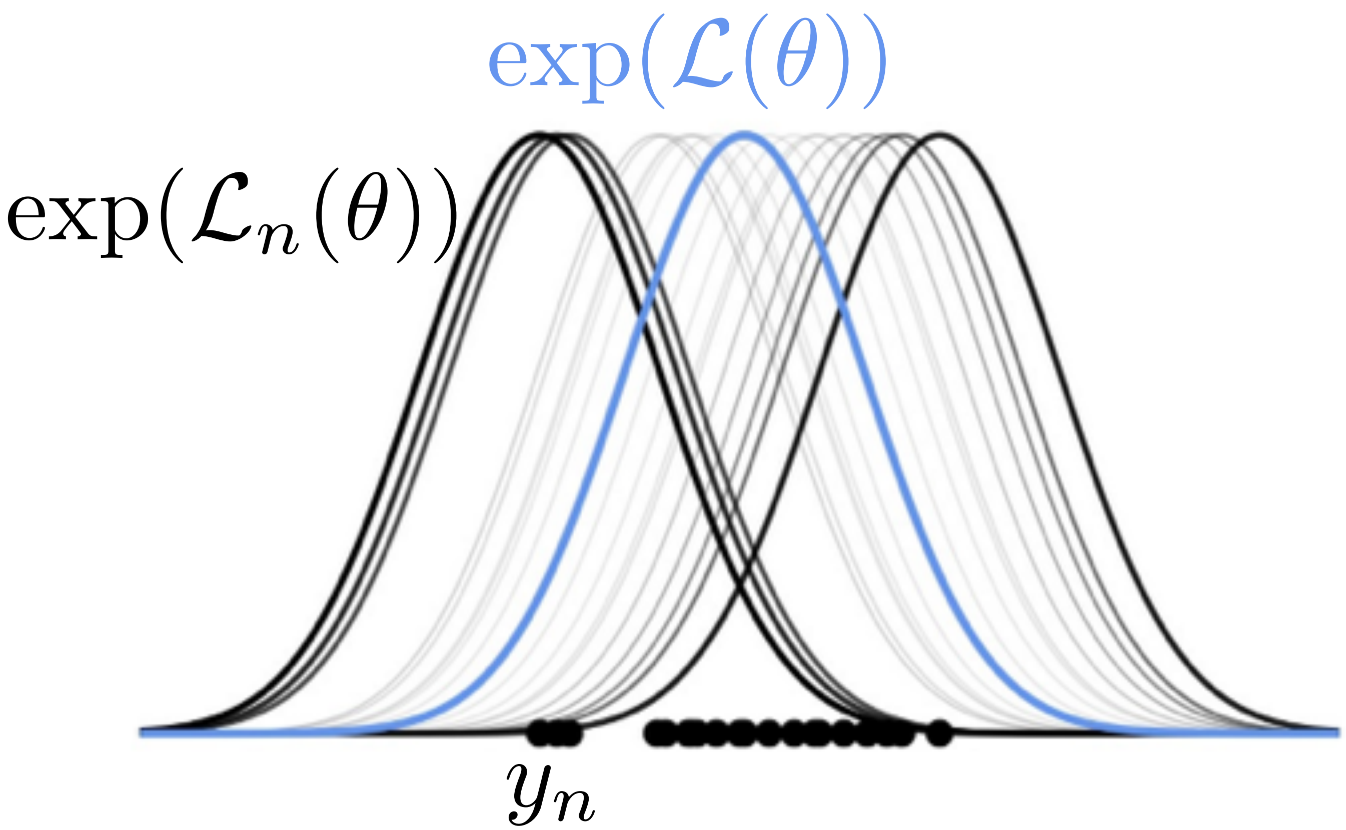
Hilbert coresets

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

Hilbert coresets

- Want a good coreset: $\min_{w \in \mathbb{R}^N} \|\mathcal{L}(w) - \mathcal{L}\|$

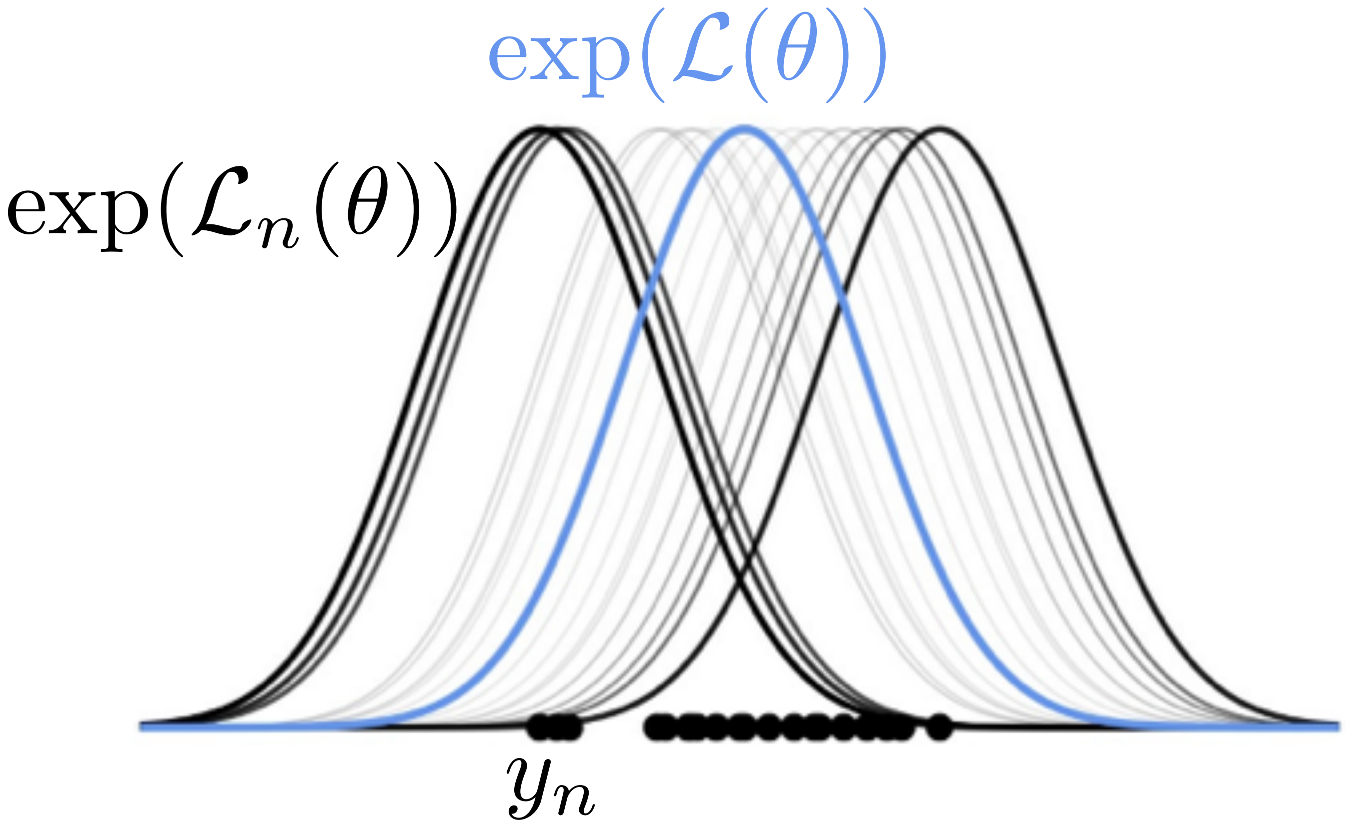
$$\text{s.t. } w \geq 0, \|w\|_0 \leq M$$

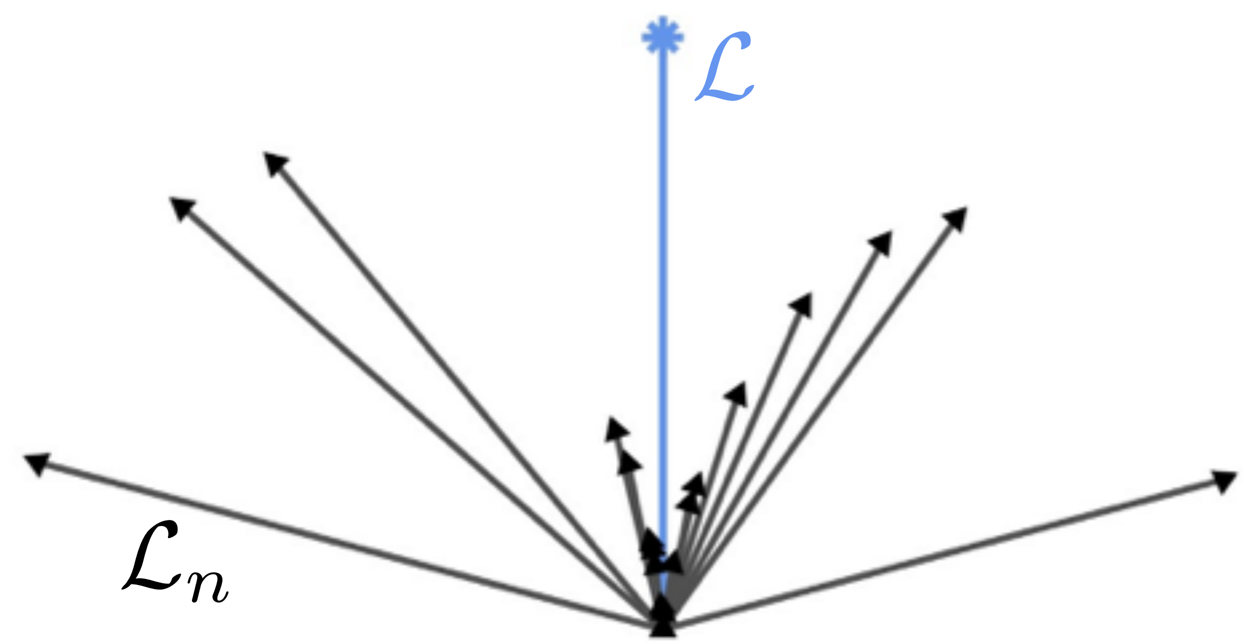


Hilbert coresets

- Want a good coreset: $\min_{w \in \mathbb{R}^N} \|\mathcal{L}(w) - \mathcal{L}\|$

$$\text{s.t. } w \geq 0, \|w\|_0 \leq M$$

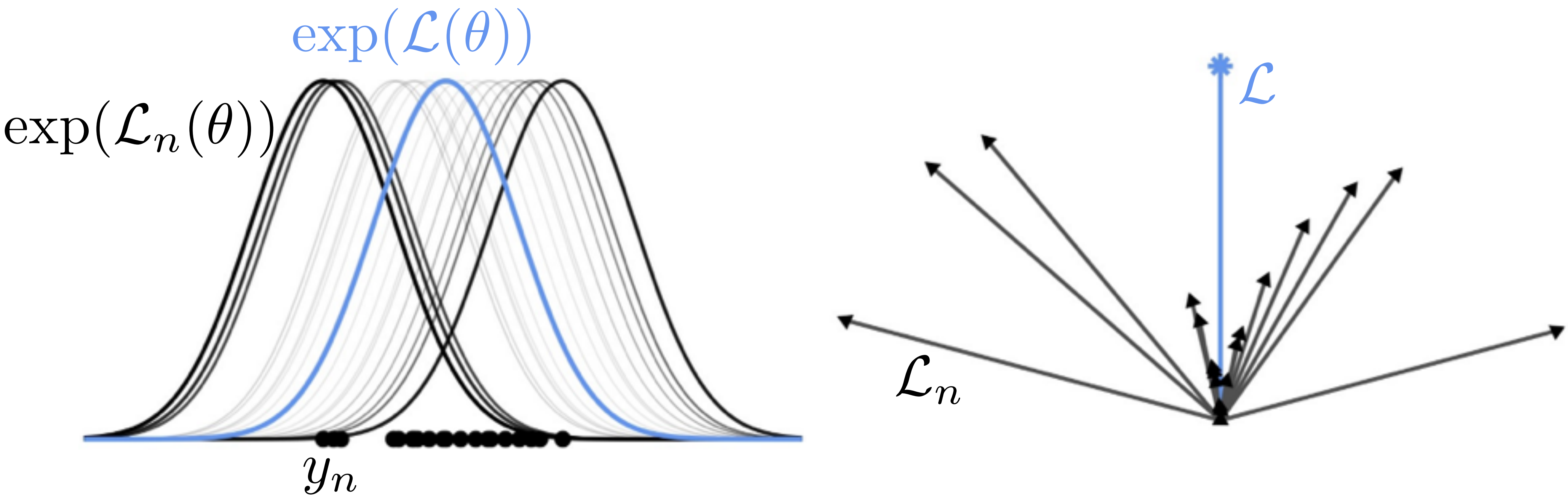




Hilbert coresets

- Want a good coreset: $\min_{w \in \mathbb{R}^N} \|\mathcal{L}(w) - \mathcal{L}\|$

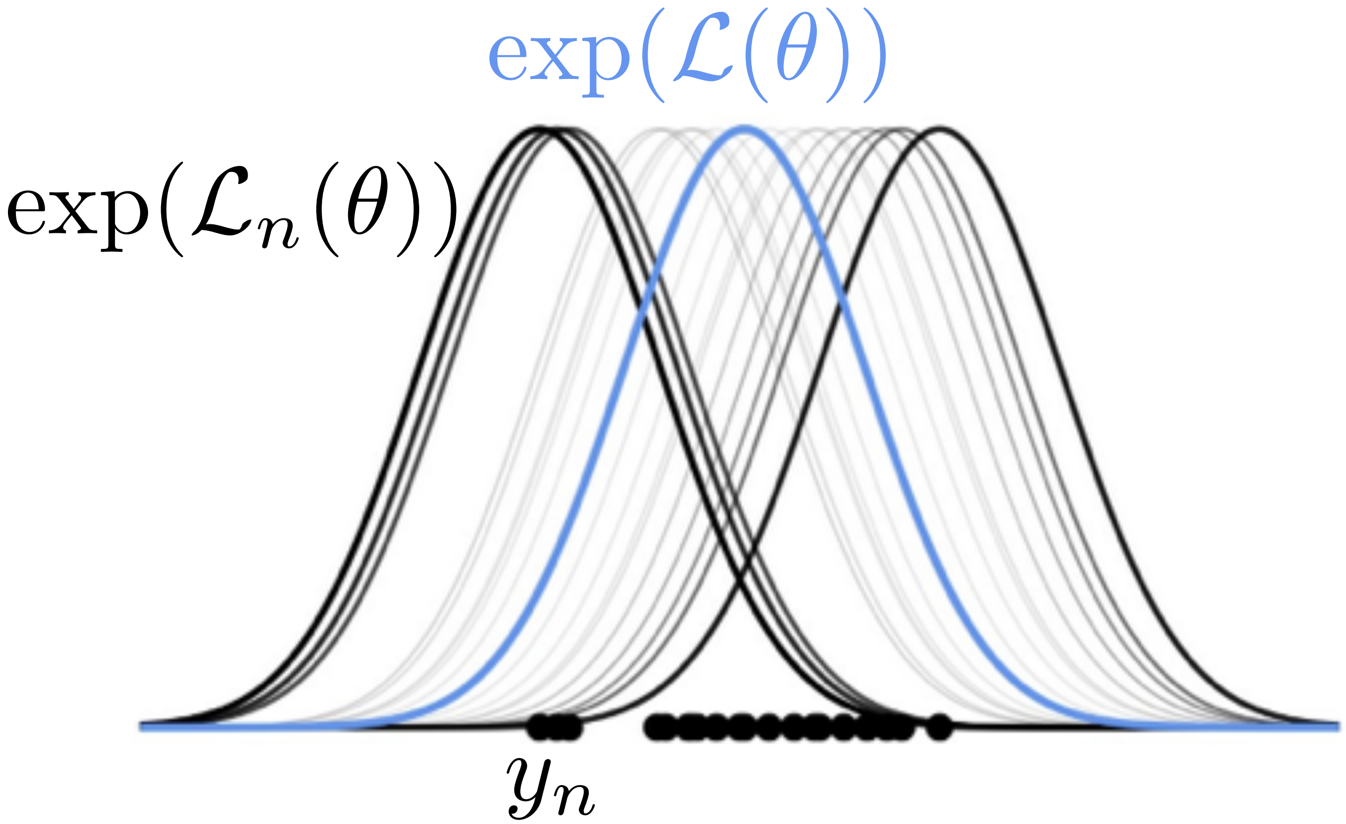
$$\text{s.t. } w \geq 0, \|w\|_0 \leq M$$

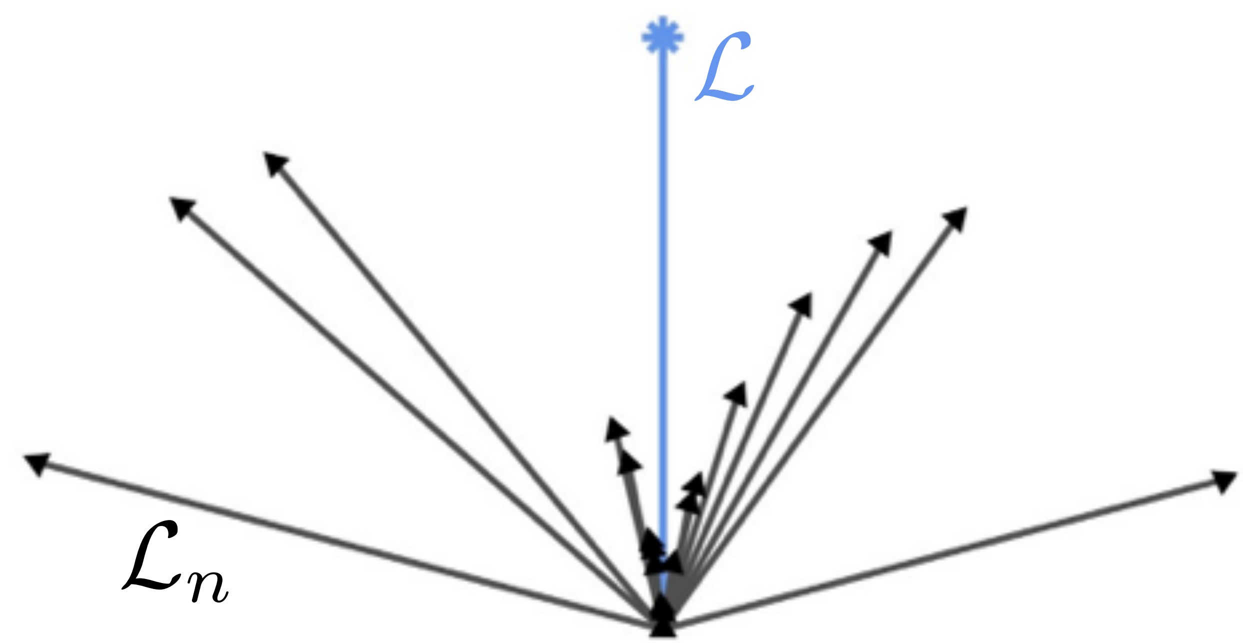


- need to consider (residual) error direction

Hilbert coresets

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





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

Roadmap

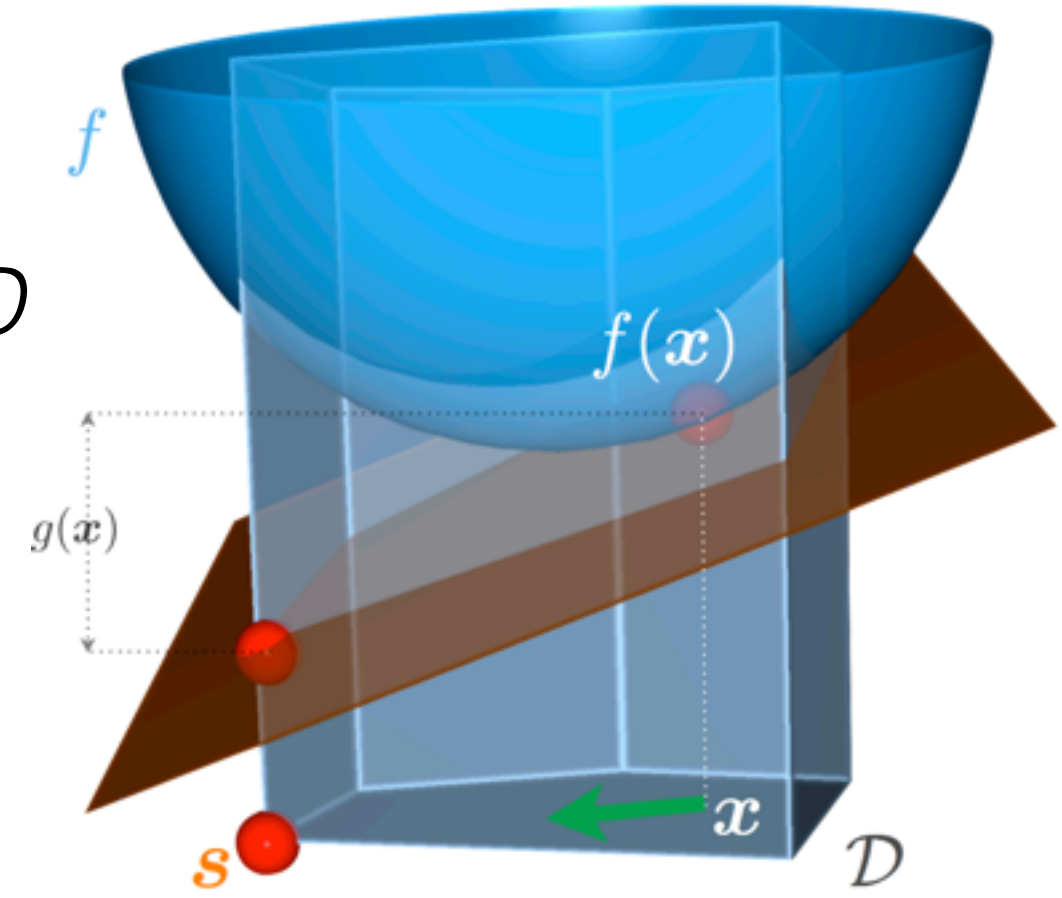
- The “core” of the data set
- Uniform data subsampling isn’t enough
- Importance sampling for “coresets”
- Optimization for “coresets”
- Approximate sufficient statistics

Roadmap

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

Frank-Wolfe

Convex optimization on a polytope *D*

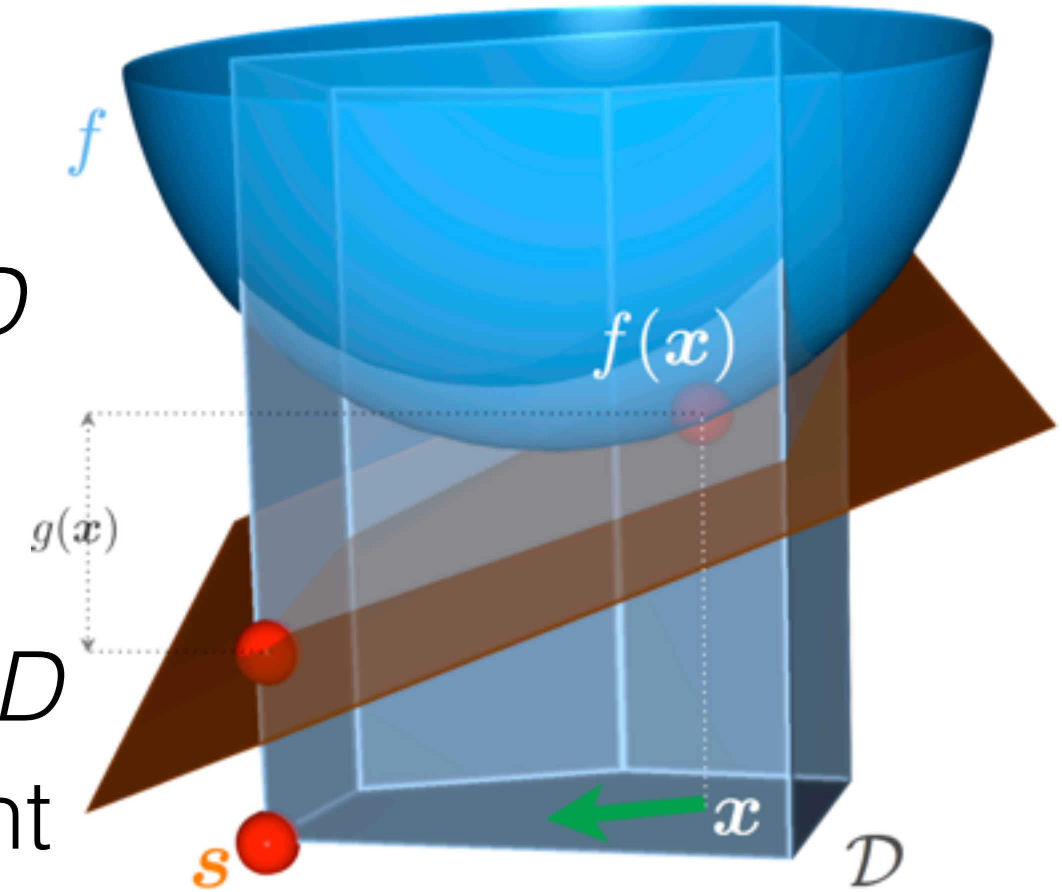


[Jaggi 2013]

Frank-Wolfe

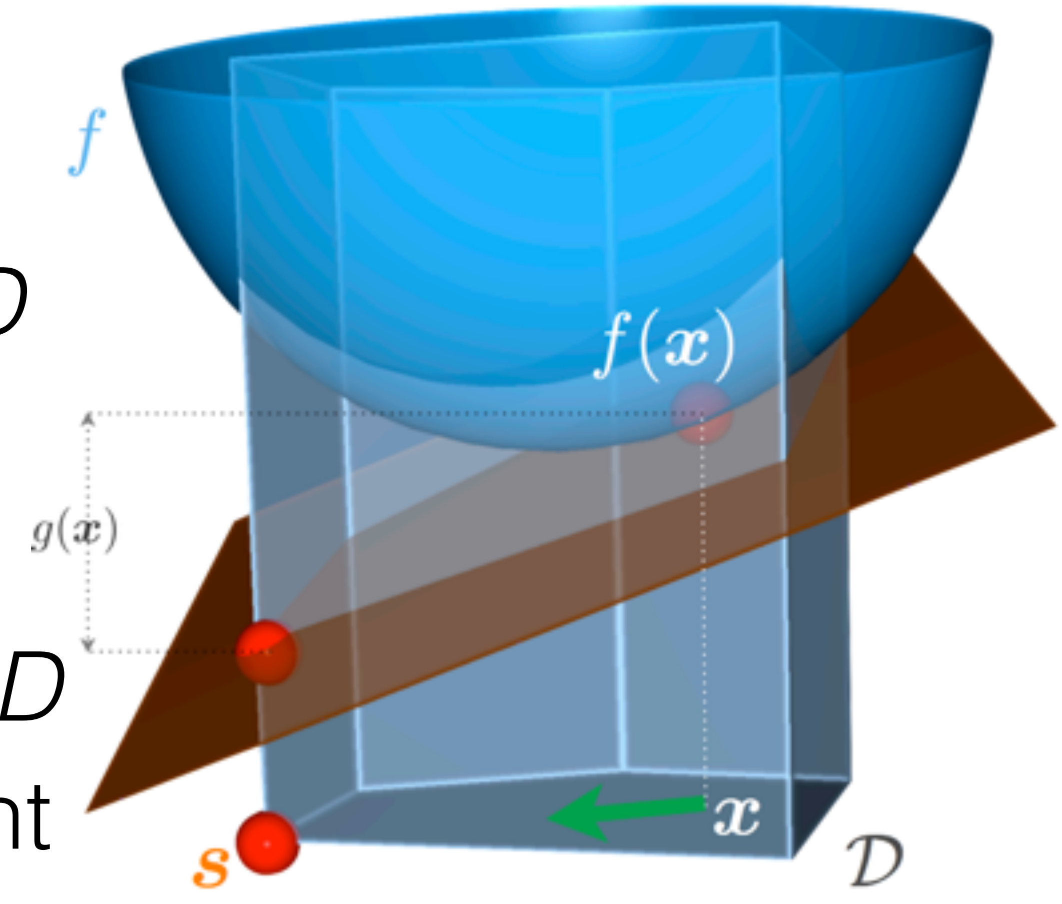
Convex optimization on a polytope *D*

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



[Jaggi 2013]

Frank-Wolfe

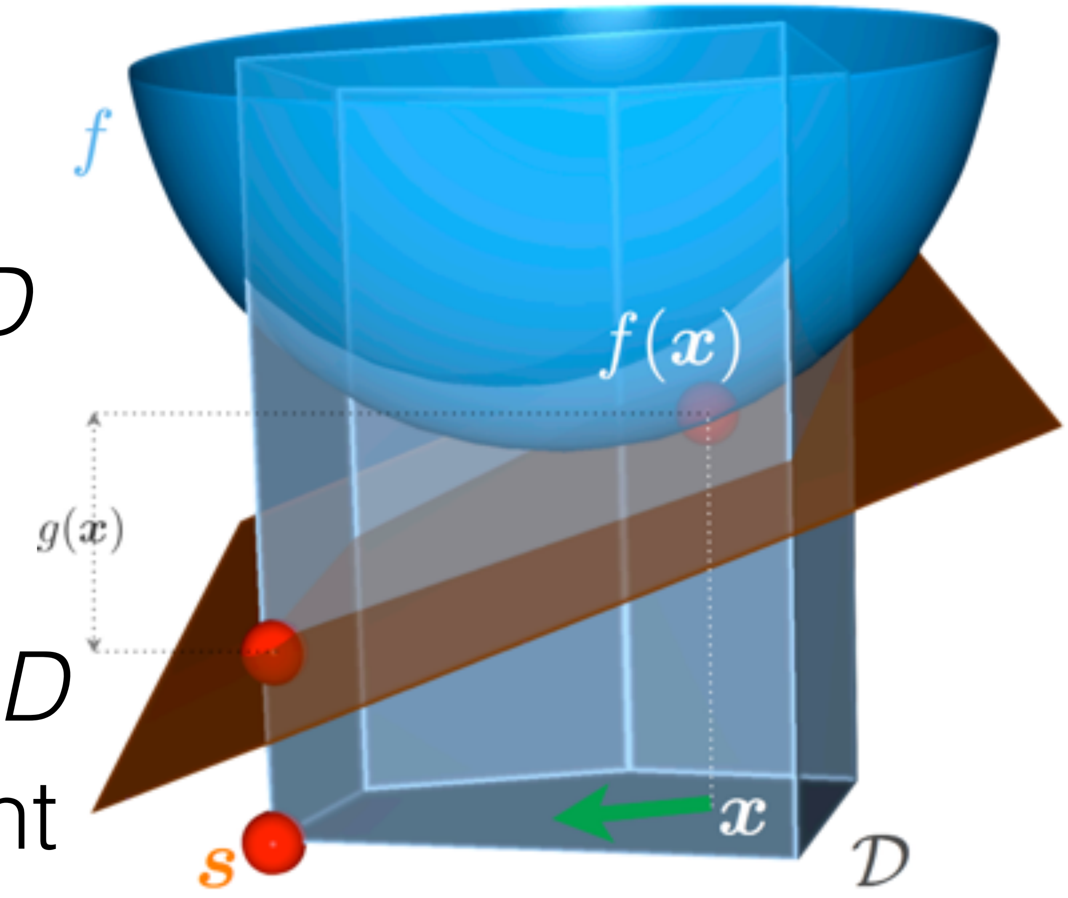


[Jaggi 2013]

Convex optimization on a polytope *D*

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

Frank-Wolfe

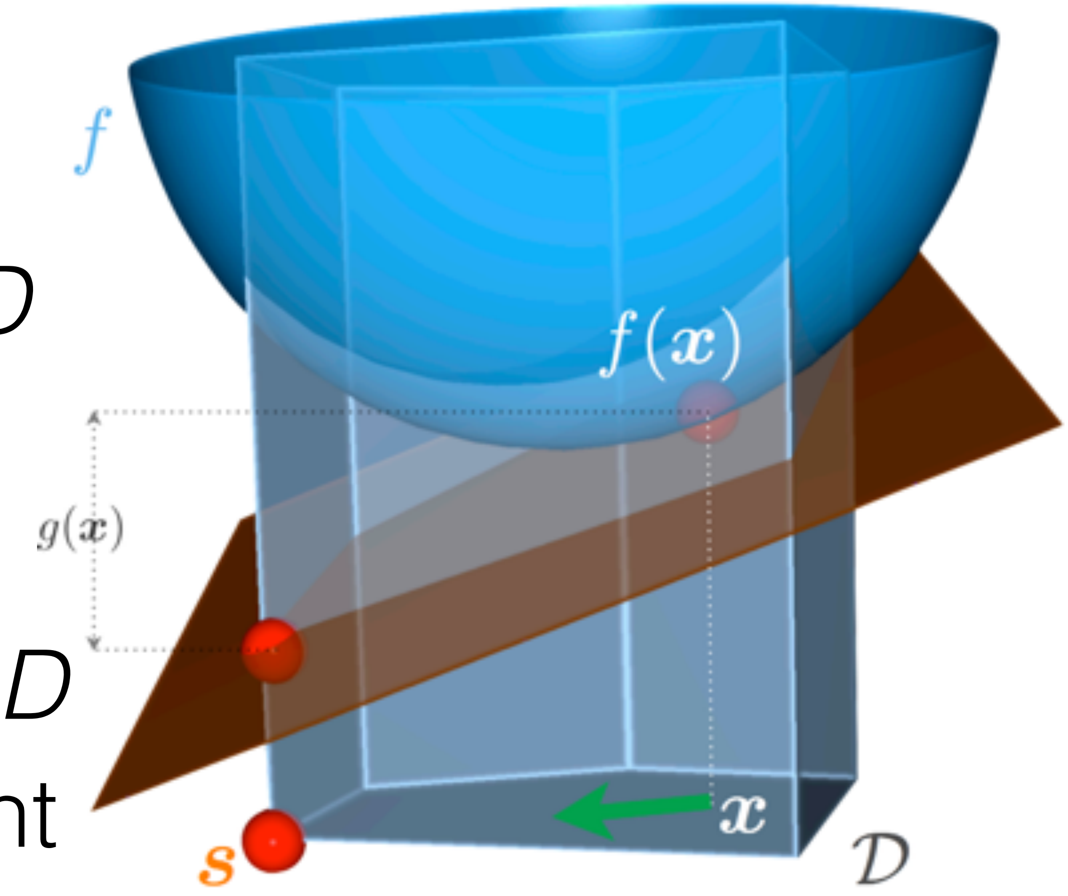


[Jaggi 2013]

Convex optimization on a polytope *D*

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

Frank-Wolfe

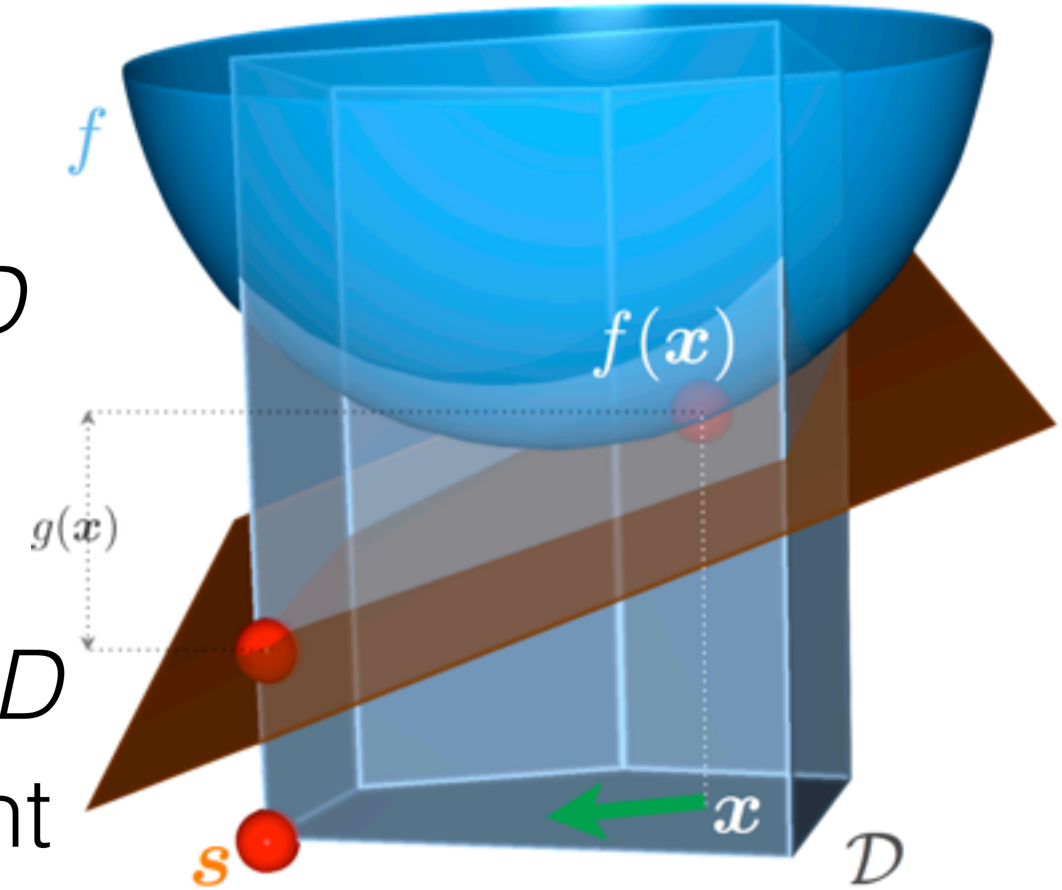


[Jaggi 2013]

Convex optimization on a polytope *D*

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

Frank-Wolfe



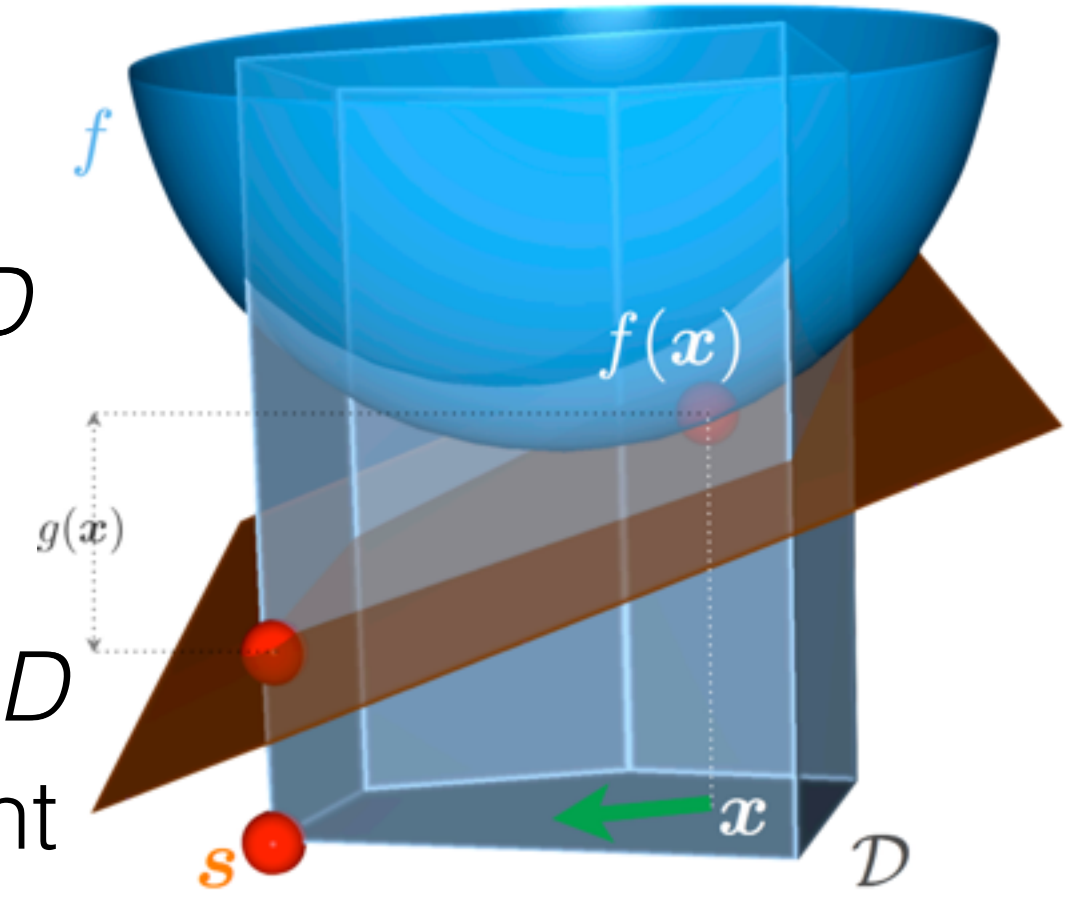
[Jaggi 2013]

Convex optimization on a polytope *D*

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

$$\text{s.t. } w \geq 0, \|w\|_0 \leq M$$

Frank-Wolfe



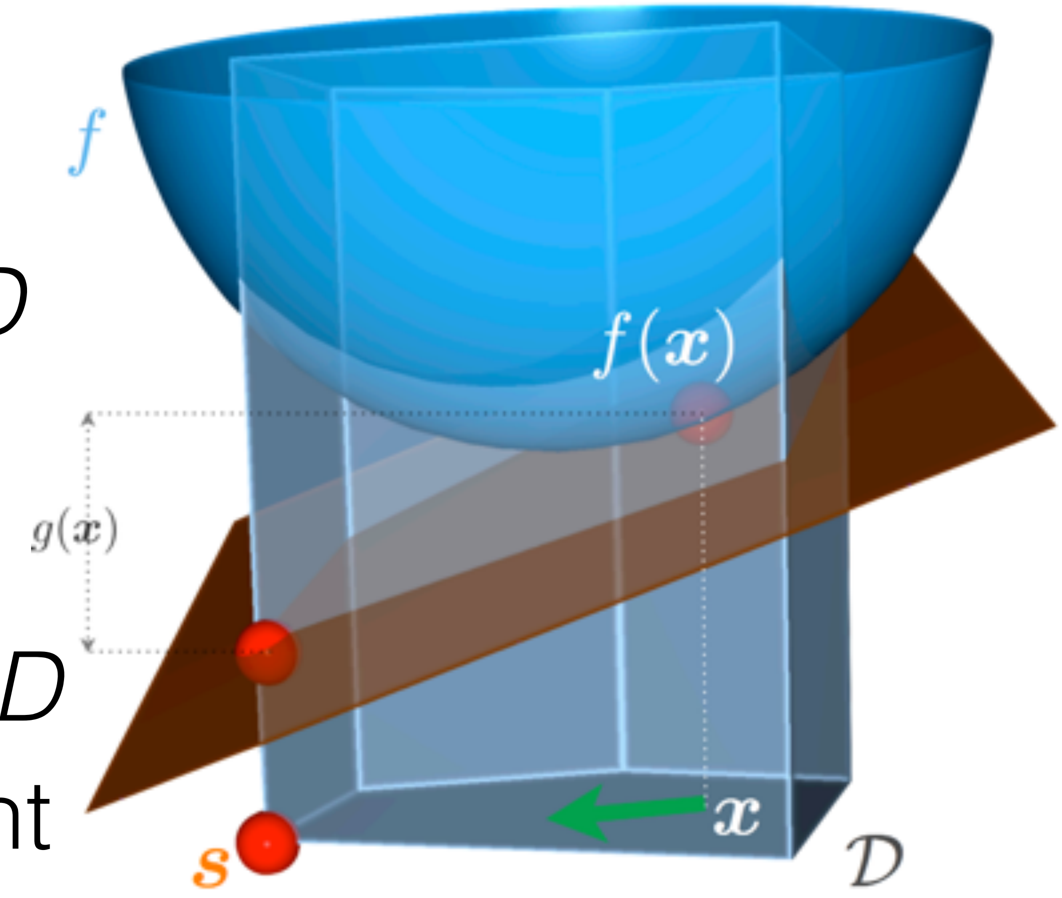
[Jaggi 2013]

Convex optimization on a polytope *D*

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

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

Frank-Wolfe



[Jaggi 2013]

Convex optimization on a polytope D

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

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

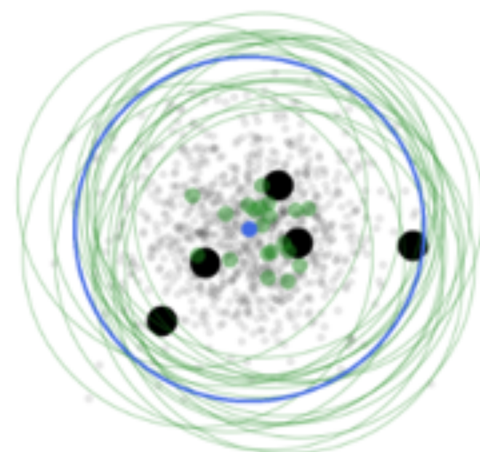
Thm (Campbell, B). After M iterations,

$$\|\mathcal{L}(w) - \mathcal{L}\| \leq \frac{\sigma \bar{\eta}}{\sqrt{\alpha^{2M} + M}}$$

Gaussian model (simulated)

- 10K pts; norms, inference: closed-form

Uniform subsampling

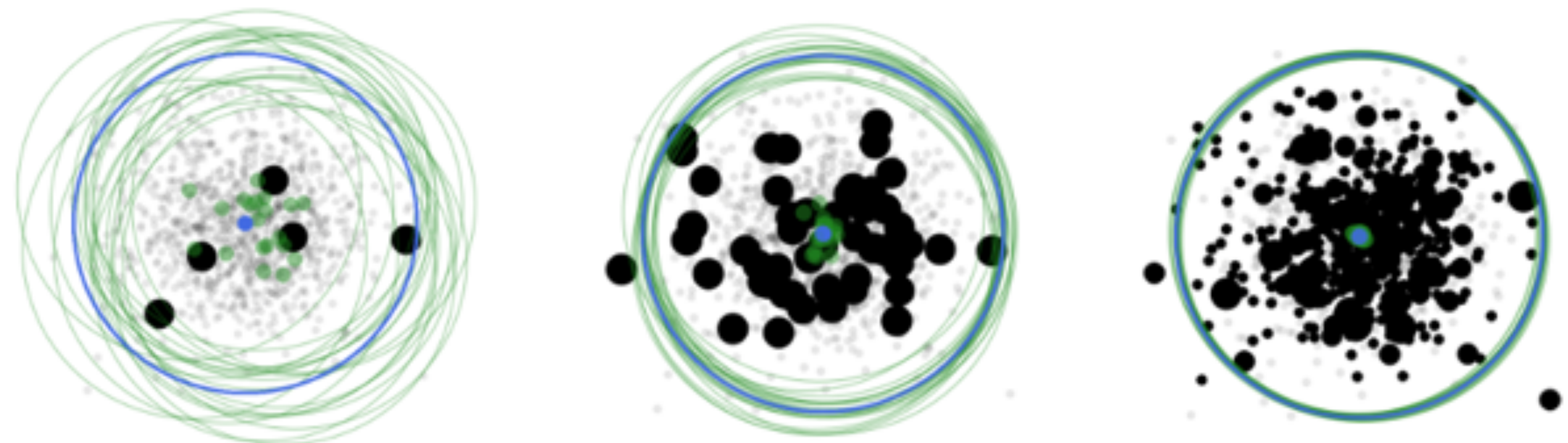


$M = 5$

Gaussian model (simulated)

- 10K pts; norms, inference: closed-form

Uniform subsampling



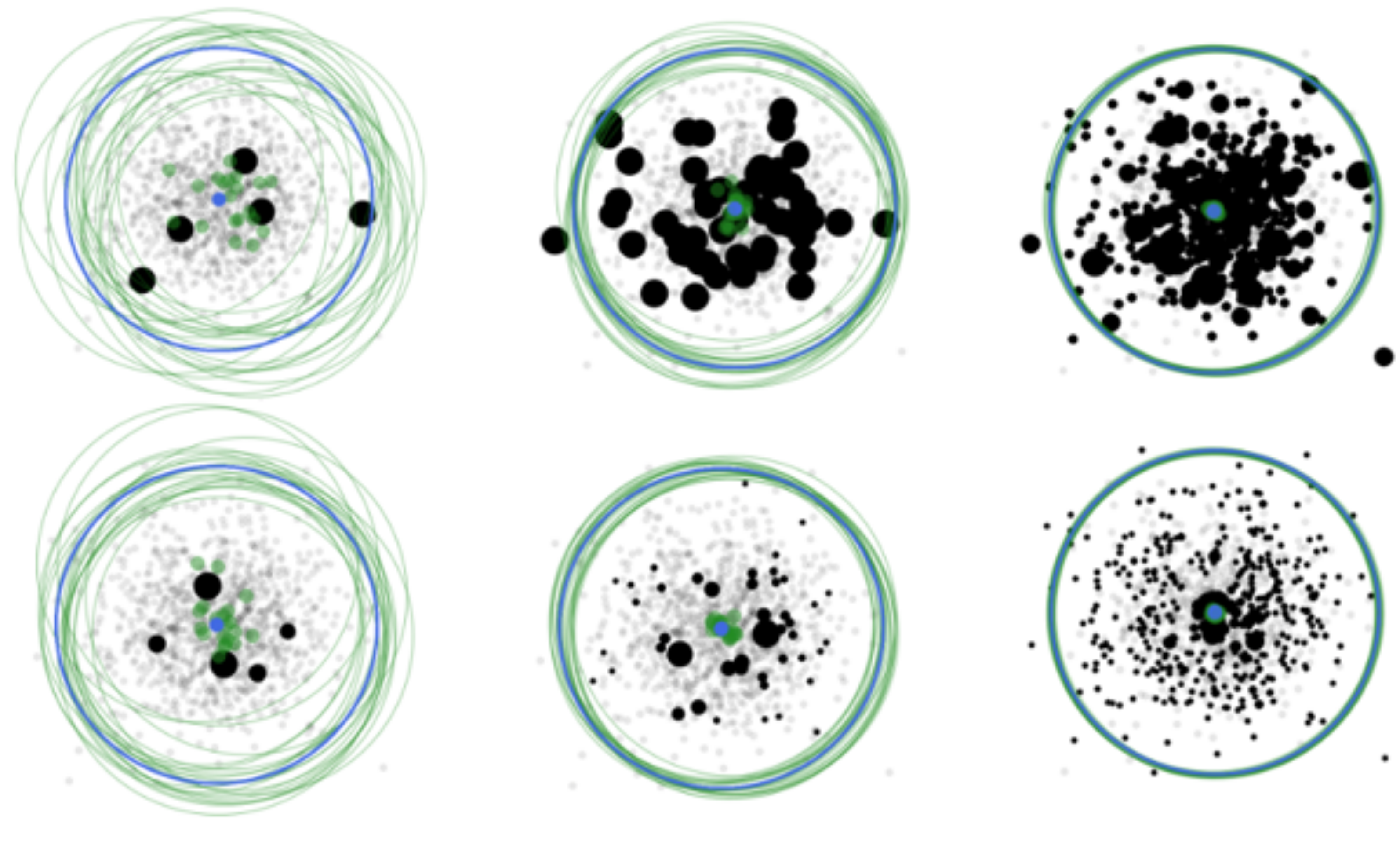
$M = 5$ $M = 50$ $M = 500$

Gaussian model (simulated)

- 10K pts; norms, inference: closed-form

Uniform subsampling

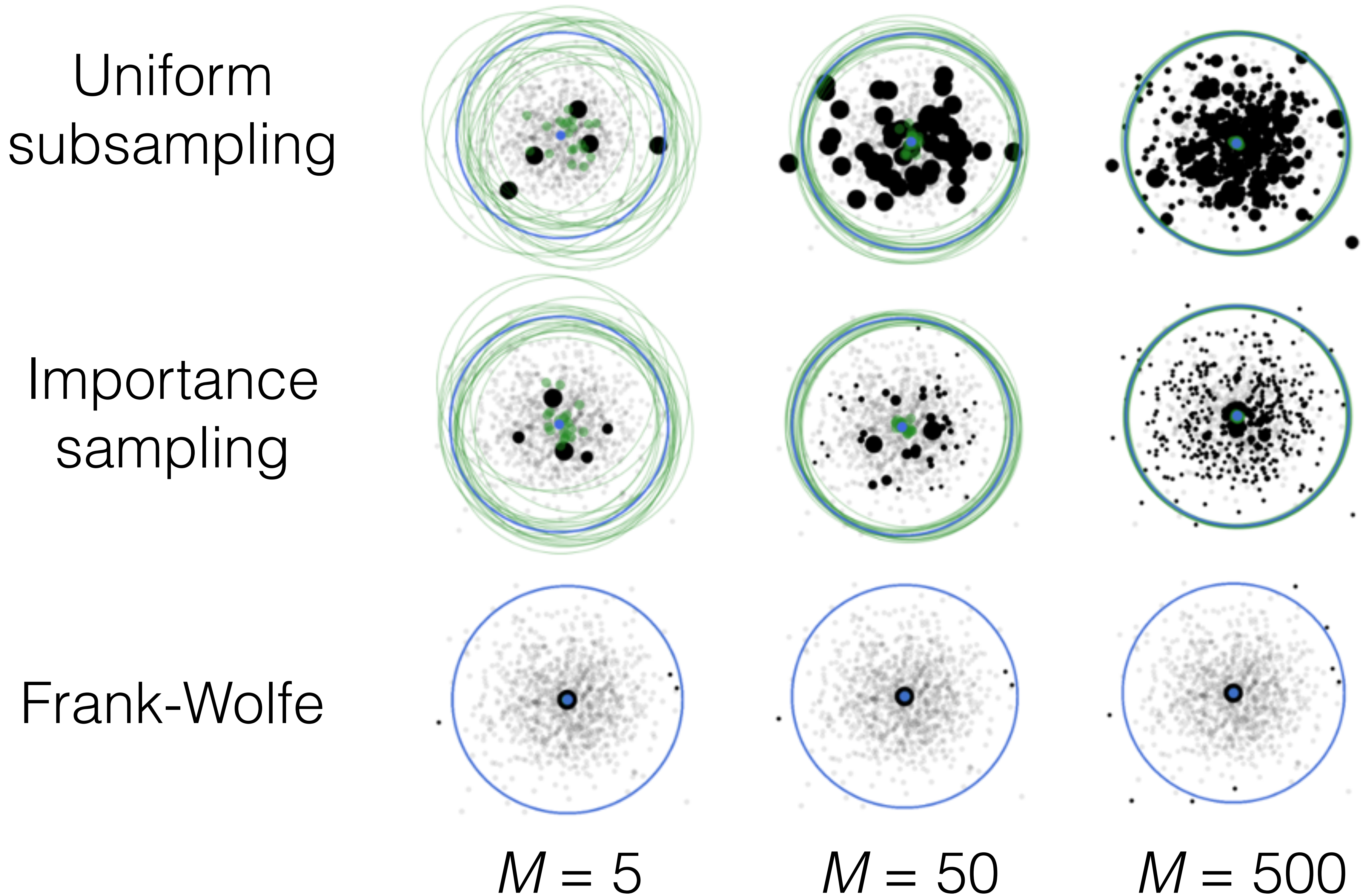
Importance sampling



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

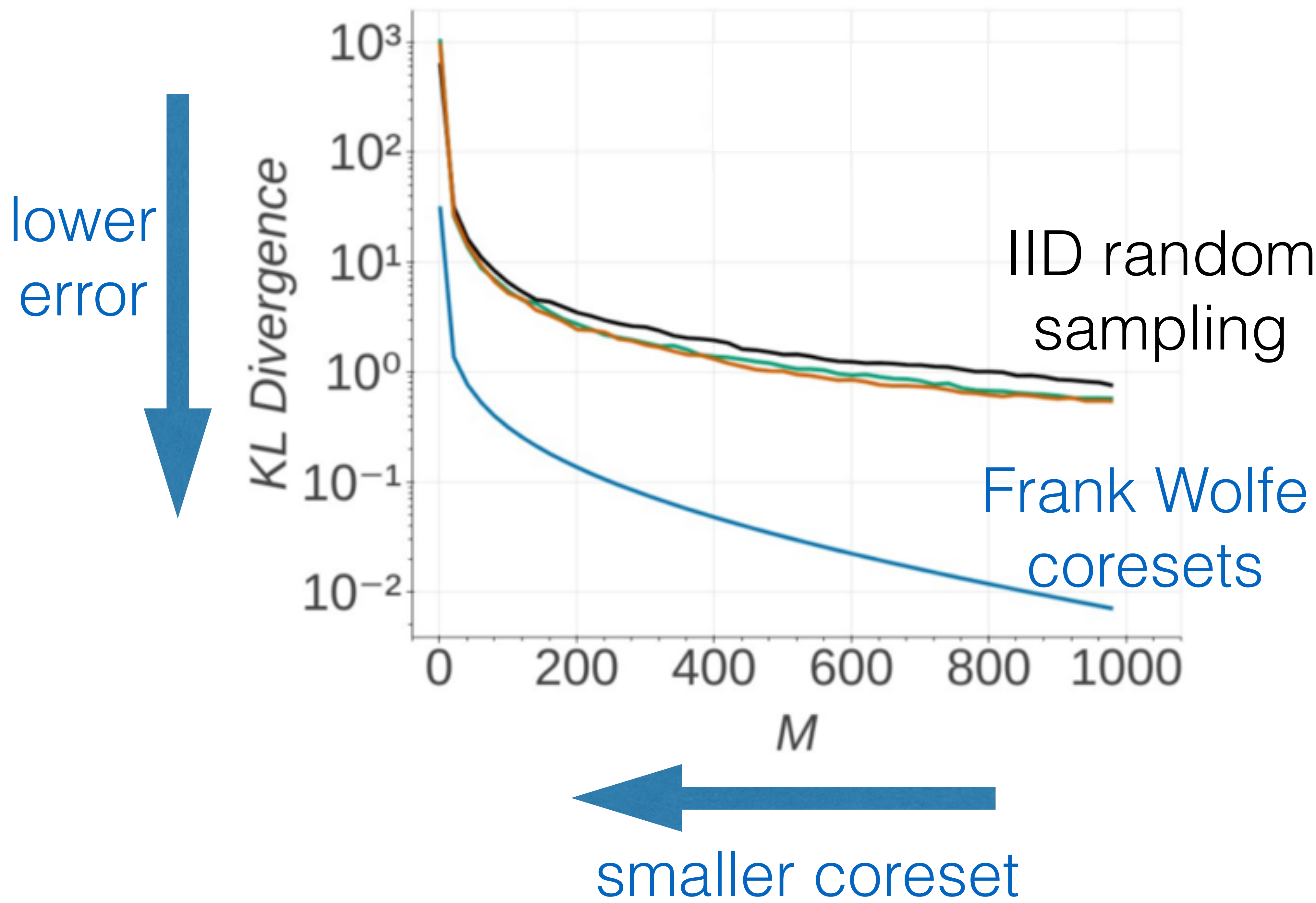
Gaussian model (simulated)

- 10K pts; norms, inference: closed-form



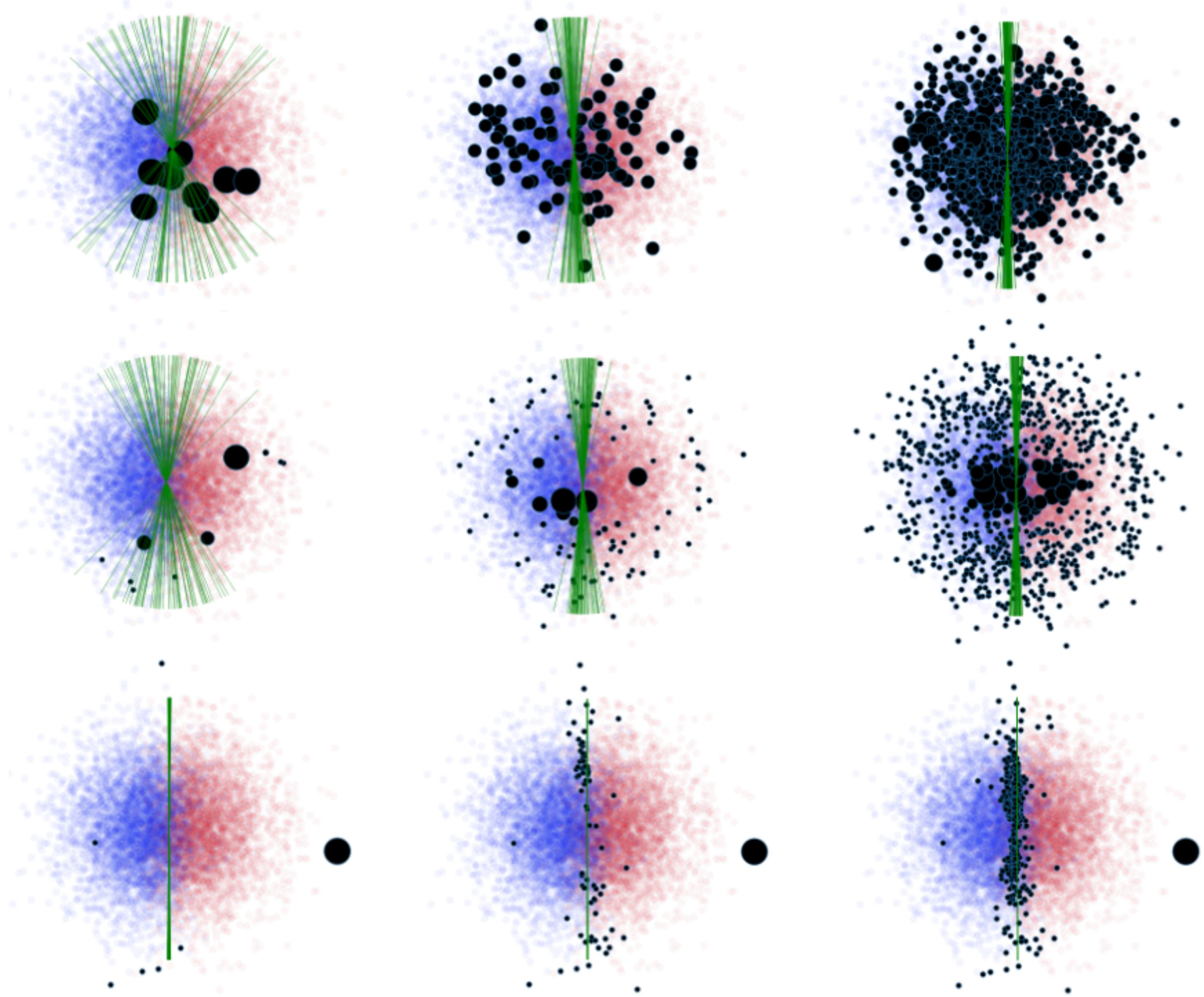
Gaussian model (simulated)

- 10K pts; norms, inference: closed-form



Logistic regression (simulated)

- 10K pts; general inference



Uniform subsampling

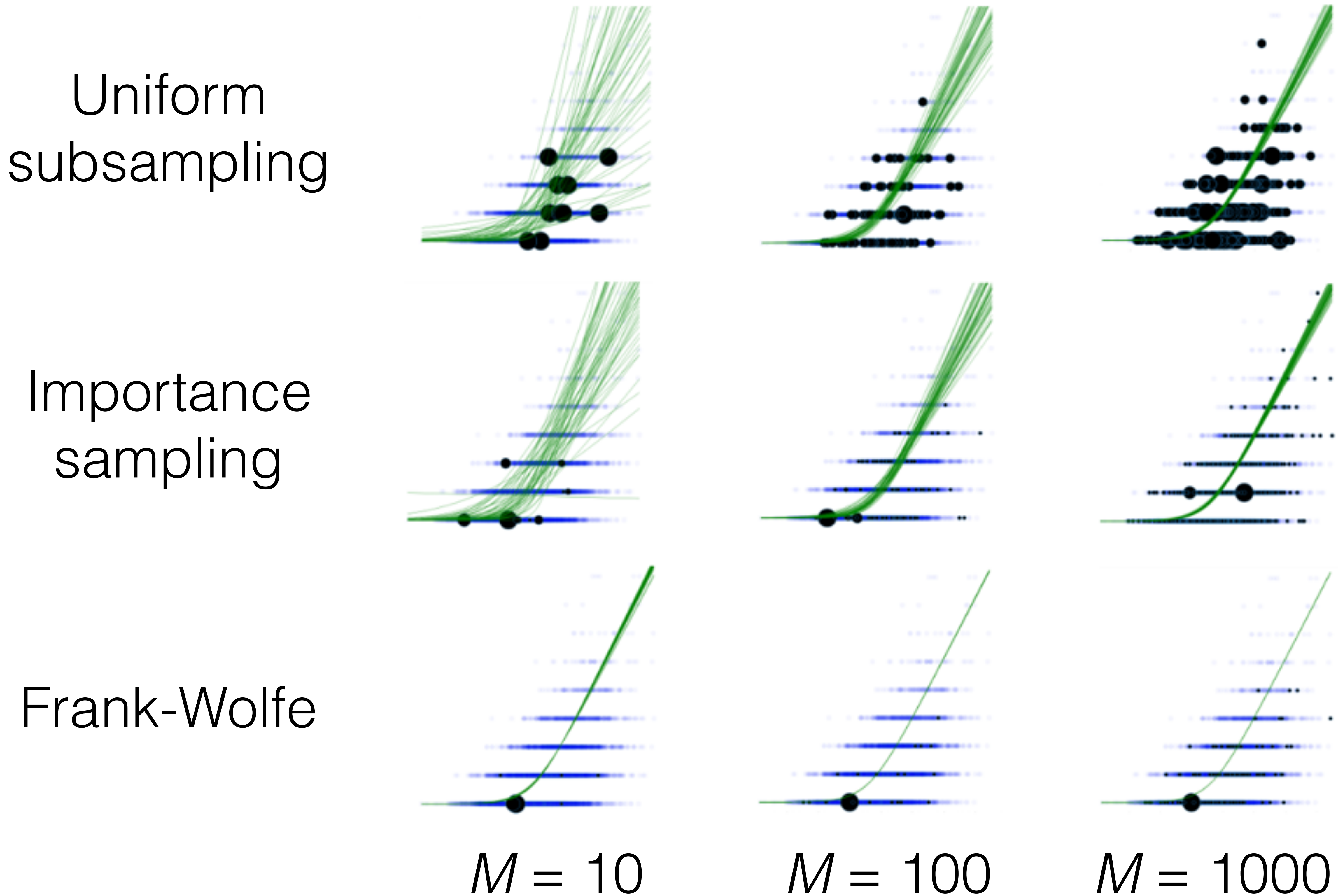
Importance sampling

Frank-Wolfe

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

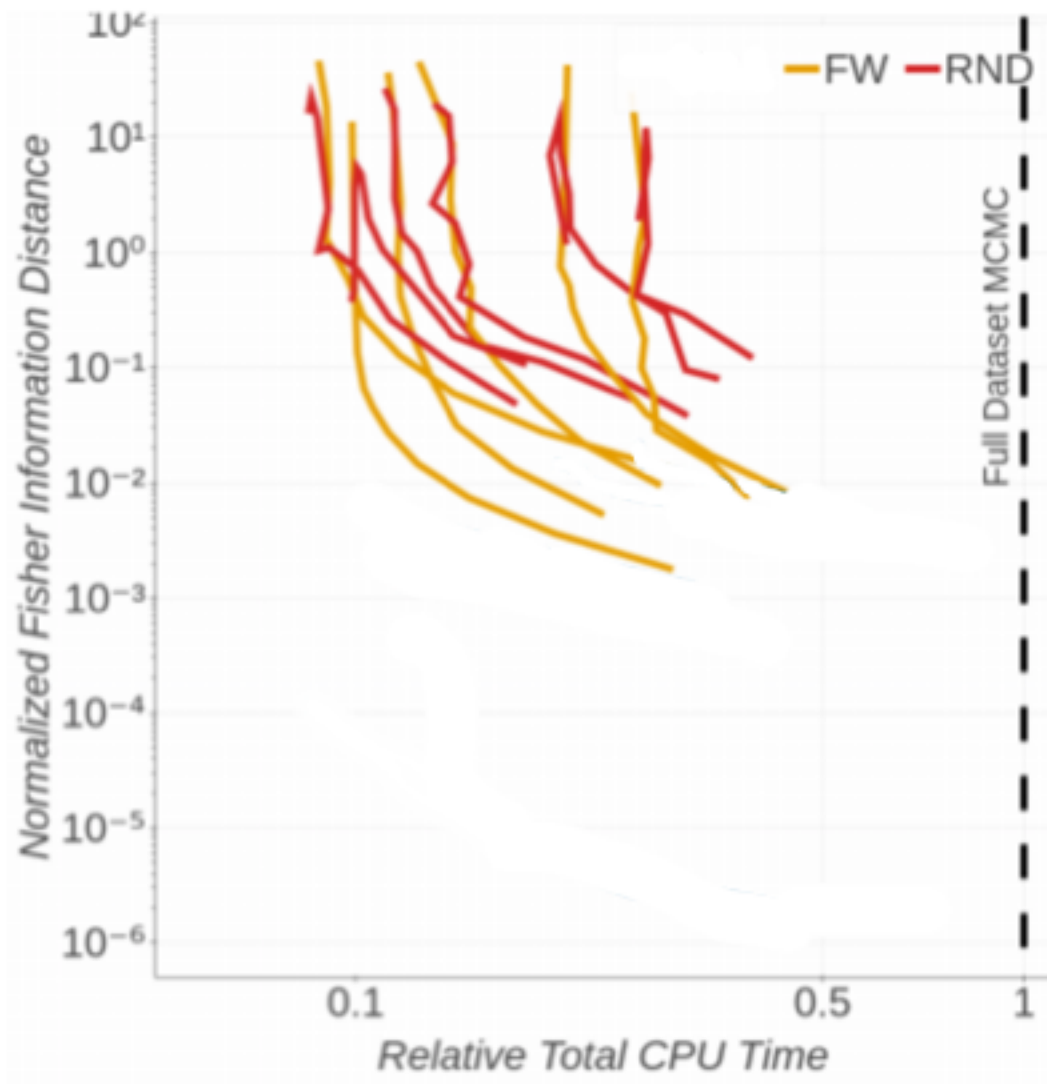
Poisson regression (simulated)

- 10K pts; general inference



Real data experiments

lower error



less total time

Uniform subsampling

Frank Wolfe coresets

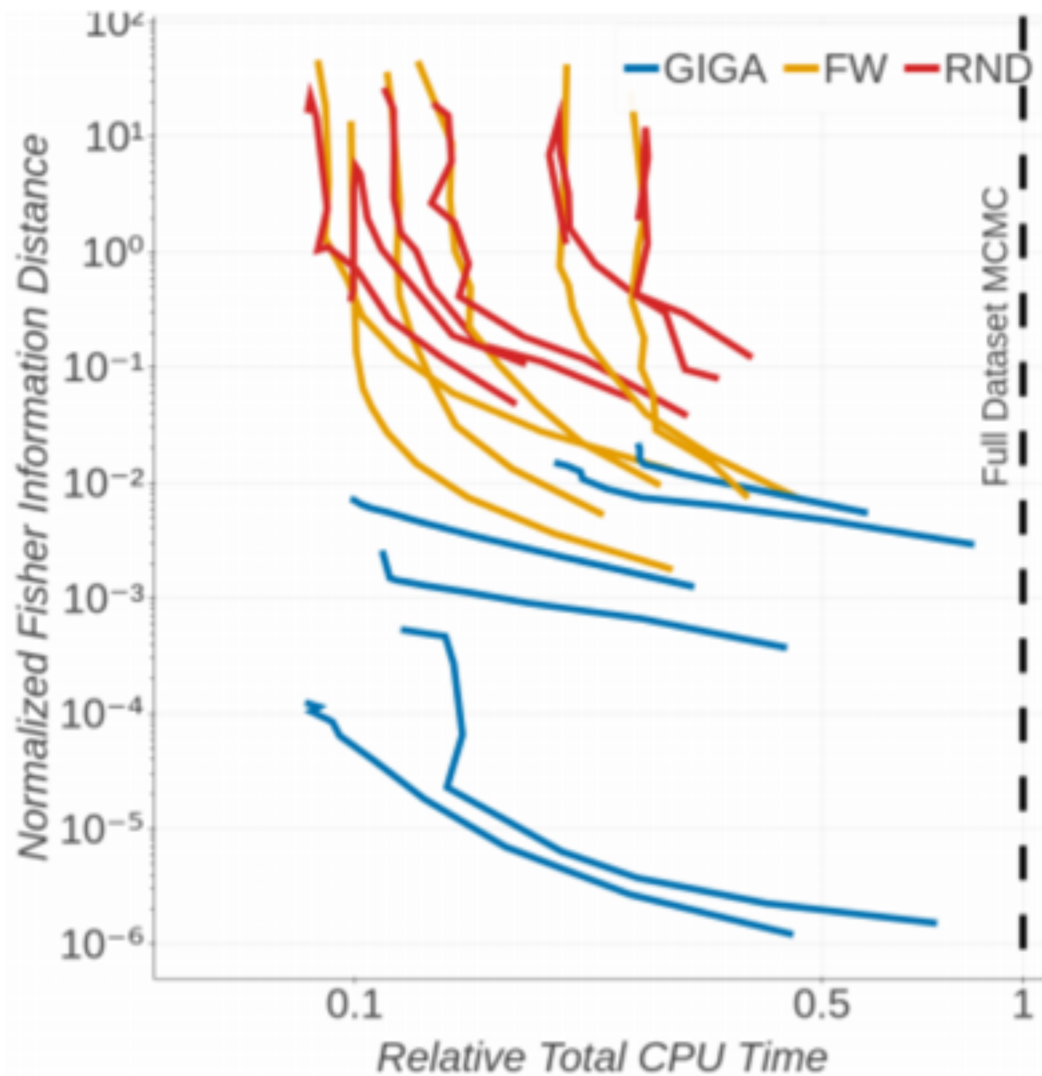
Data sets include:

- Phishing
- Chemical reactivity
- Bicycle trips
- Airport delays

[Campbell, Broderick 2017]

Real data experiments

lower error



less total time

Uniform subsampling

Frank Wolfe coresets

GIGA coresets

Data sets include:

- Phishing
- Chemical reactivity
- Bicycle trips
- Airport delays

[Campbell, Broderick 2017, 2018]

Roadmap

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

Roadmap

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

Data summarization

Data summarization

- Exponential family likelihood

Data summarization

- Exponential family likelihood

Sufficient statistics

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

Data summarization

- Exponential family likelihood

Sufficient statistics

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

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

Data summarization

- Exponential family likelihood

Sufficient statistics

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

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

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

Data summarization

- Exponential family likelihood

Sufficient statistics

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

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

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

Data summarization

- Exponential family likelihood

Sufficient statistics

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

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

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

Data summarization

- Exponential family likelihood

Sufficient statistics

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

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

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

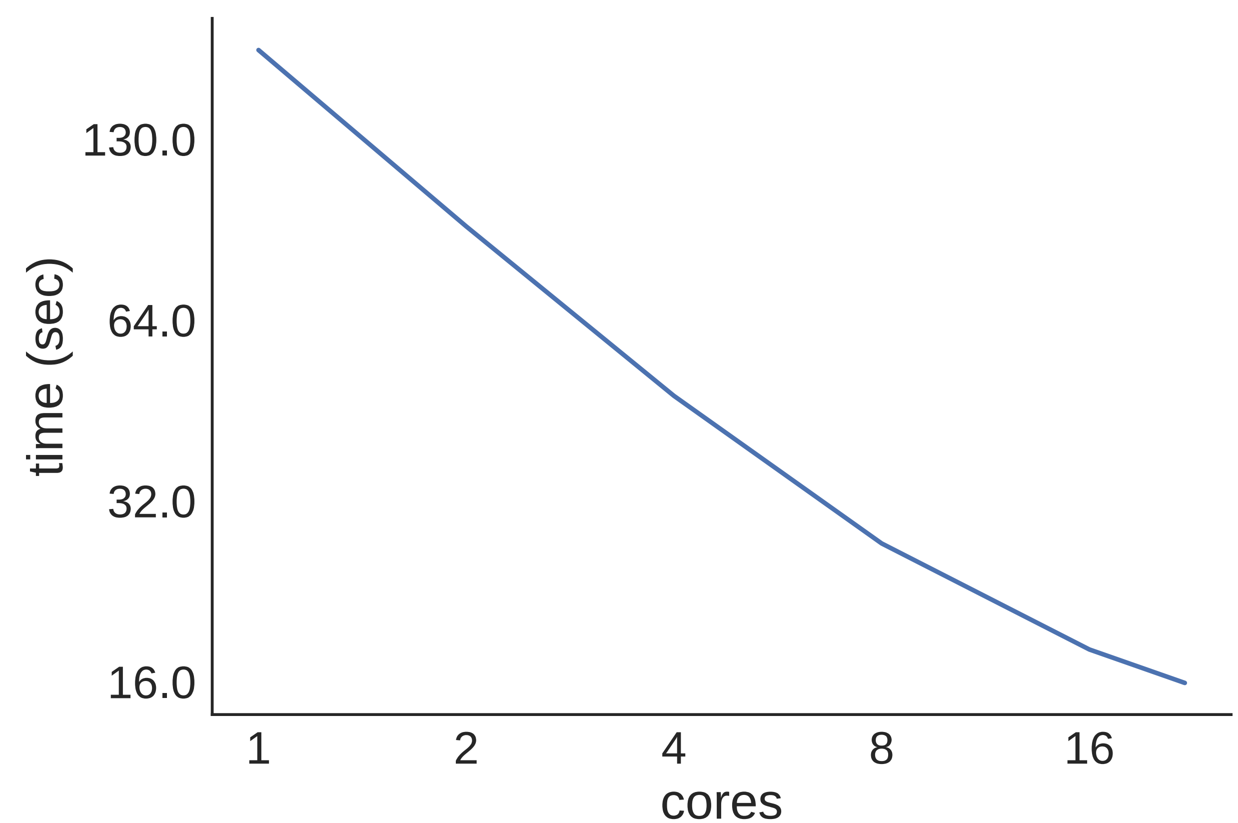
Data summarization

Criteo Labs > Algorithms > Criteo Releases its New Dataset

Criteo Releases its New Dataset

By: CriteoLabs / 31 Mar 2015

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



[Huggins, Adams, Broderick 2017]

Conclusions

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

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

Conclusions

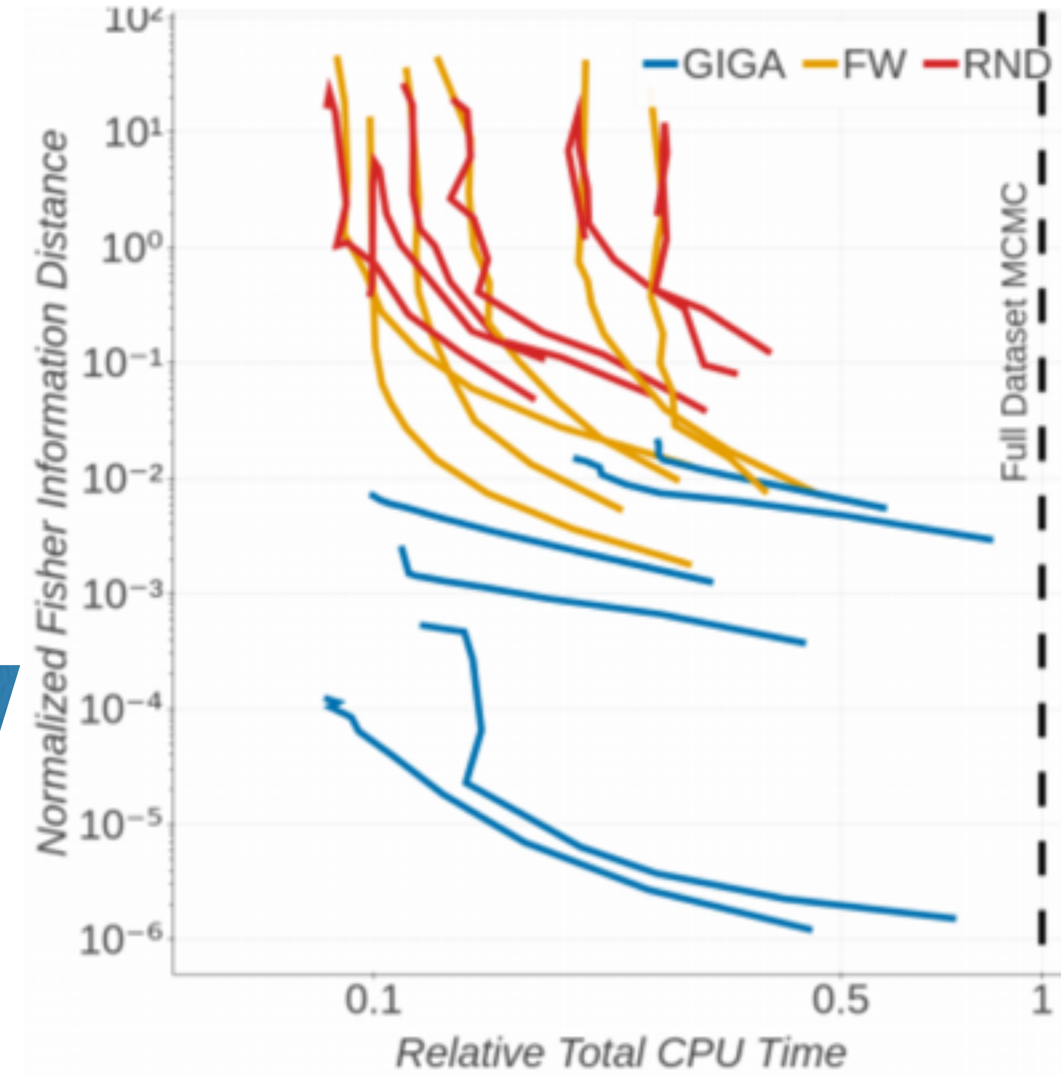
- *Data summarization* for **scalable**, **automated** approx. Bayes algorithms with **error bounds on quality for finite data**
  - Coresets
  - Approx. suff. stats

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

Conclusions

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

lower error



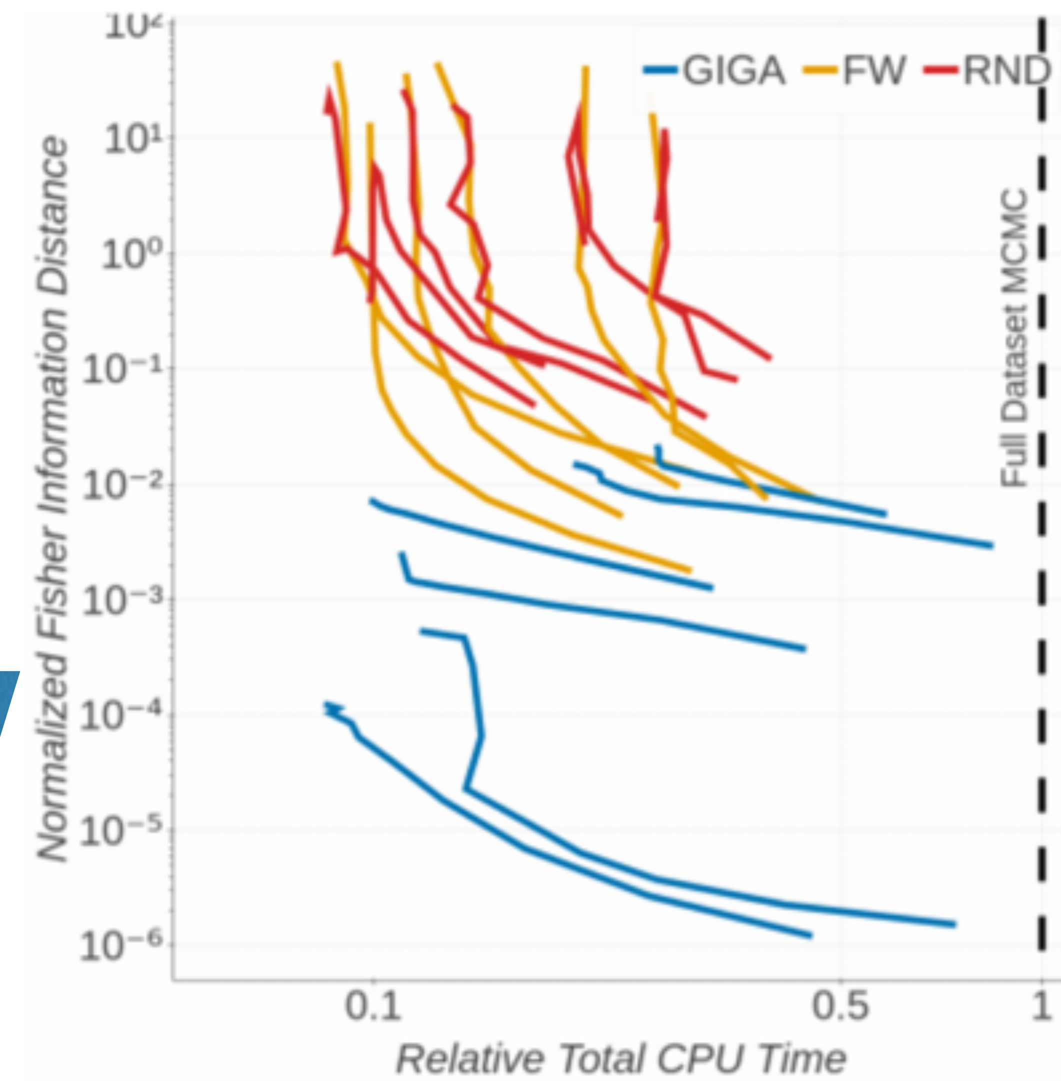
less total time

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

Conclusions

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

lower error

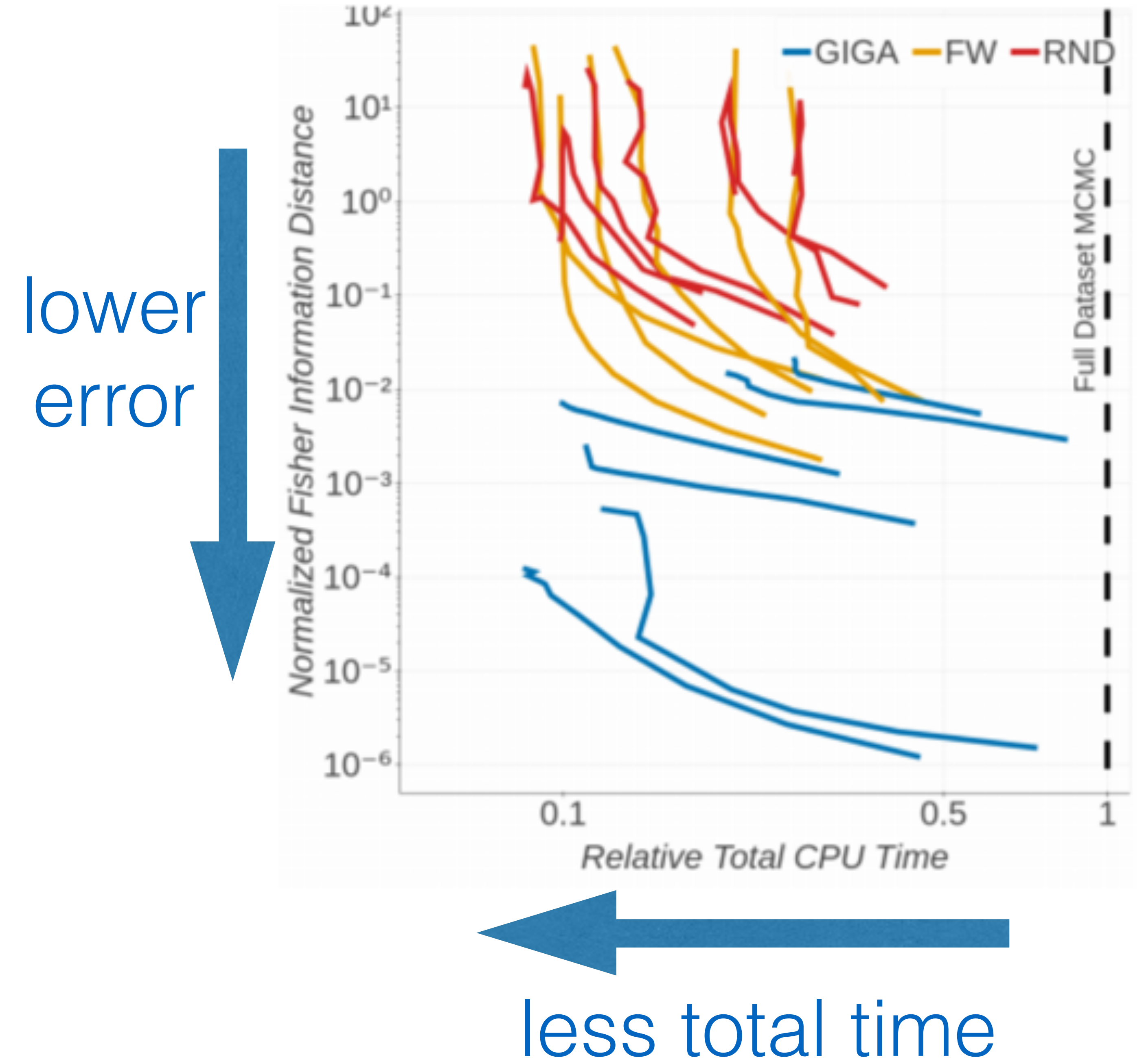


less total time

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

Conclusions

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



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

R Agrawal, C Uhler, and T Broderick. Minimal I-MAP MCMC for Scalable Structure Discovery in Causal DAG Models. *ICML* 2018: Fri 5:20--5:40PM @A5

T Campbell and T Broderick. Automated scalable Bayesian inference via Hilbert coresets. Under review. ArXiv:1710.05053.

* Code: https://github.com/trevorcampbell/bayesian-coresets

T Campbell and T Broderick. Bayesian Coreset Construction via Greedy Iterative Geodesic Ascent. *ICML* 2018: Fri 5:00--5:20PM @A4

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

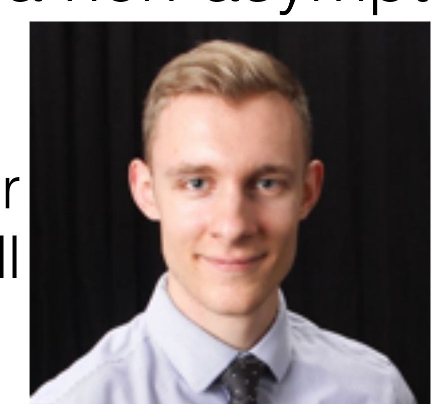
JH Huggins, RP Adams, and T Broderick. PASS-GLM: Polynomial approximate sufficient statistics for scalable Bayesian GLM inference. *NIPS* 2017.

JH Huggins, T Campbell, M Kasprzak, and T Broderick. Scalable Gaussian Process inference with finite-data mean and variance guarantees. https://arxiv.org/abs/1806.10234

JH Huggins, M Kasprzak, T Campbell, and T Broderick. Bayesian posterior mean and uncertainty estimates: a non-asymptotic approach. In preparation.

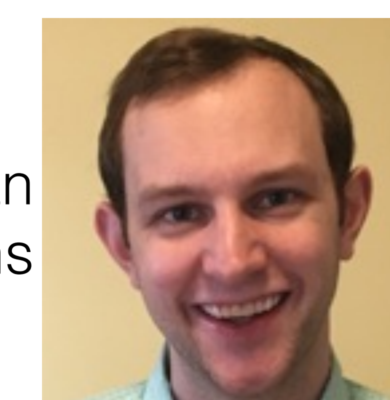
Raj Agrawal

Trevor Campbell



Ryan Giordano

Jonathan Huggins



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

R Agrawal, C Uhler, and T Broderick. Minimal I-MAP MCMC for Scalable Structure Discovery in Causal DAG Models. *ICML* 2018: Fri 5:20--5:40PM @A5

T Campbell and T Broderick. Automated scalable Bayesian inference via Hilbert coresets. Under review. ArXiv:1710.05053.
 * Code: https://github.com/trevorcampbell/bayesian-coresets

T Campbell and T Broderick. Bayesian Coreset Construction via Greedy Iterative Geodesic Ascent. *ICML* 2018: Fri 5:00--5:20PM @A4

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

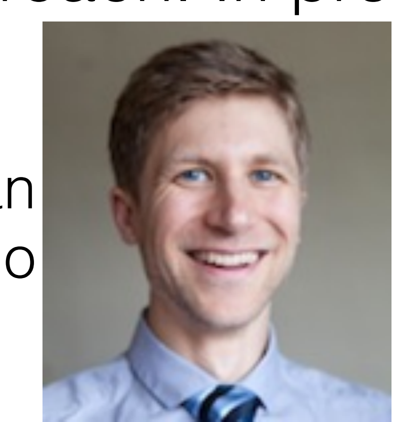
JH Huggins, RP Adams, and T Broderick. PASS-GLM: Polynomial approximate sufficient statistics for scalable Bayesian GLM inference. *NIPS* 2017.

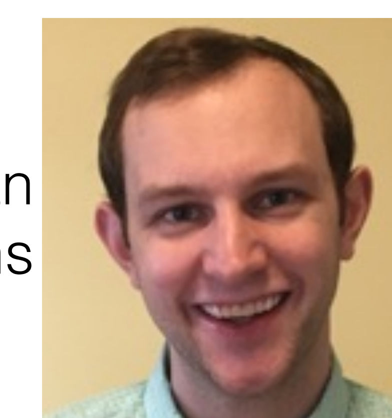
JH Huggins, T Campbell, M Kasprzak, and T Broderick. Scalable Gaussian Process inference with finite-data mean and variance guarantees. https://arxiv.org/abs/1806.10234

JH Huggins, M Kasprzak, T Campbell, and T Broderick. Bayesian posterior mean and uncertainty estimates: a non-asymptotic approach. In preparation.

Raj Agrawal


Trevor Campbell


Ryan Giordano


Jonathan Huggins


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

R Agrawal, C Uhler, and T Broderick. Minimal I-MAP MCMC for Scalable Structure Discovery in Causal DAG Models. *ICML* 2018: Fri 5:20--5:40PM @A5

T Campbell and T Broderick. Automated scalable Bayesian inference via Hilbert coresets. Under review. ArXiv:1710.05053.

* Code: https://github.com/trevorcampbell/bayesian-coresets

T Campbell and T Broderick. Bayesian Coreset Construction via Greedy Iterative Geodesic Ascent. *ICML* 2018: Fri 5:00--5:20PM @A4

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

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

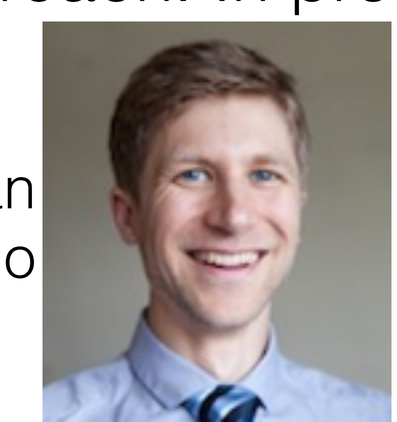
JH Huggins, T Campbell, M Kasprzak, and T Broderick. Scalable Gaussian Process inference with finite-data mean and variance guarantees. https://arxiv.org/abs/1806.10234

JH Huggins, M Kasprzak, T Campbell, and T Broderick. Bayesian posterior mean and uncertainty estimates: a non-asymptotic approach. In preparation.

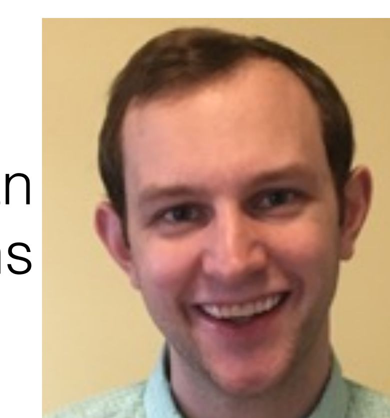
Raj Agrawal

Trevor Campbell

Ryan Giordano



Jonathan Huggins



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

R Agrawal, C Uhler, and T Broderick. Minimal I-MAP MCMC for Scalable Structure Discovery in Causal DAG Models. ***ICML* 2018: Fri 5:20--5:40PM @A5**

T Campbell and T Broderick. Automated scalable Bayesian inference via Hilbert coresets. Under review. ArXiv:1710.05053.

* Code: https://github.com/trevorcampbell/bayesian-coresets

T Campbell and T Broderick. Bayesian Coreset Construction via Greedy Iterative Geodesic Ascent. ***ICML* 2018: Fri 5:00--5:20PM @A4**

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

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

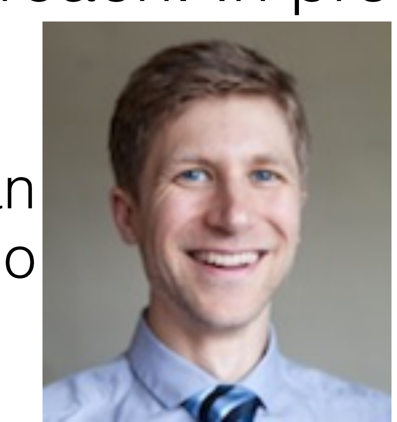
JH Huggins, T Campbell, M Kasprzak, and T Broderick. Scalable Gaussian Process inference with finite-data mean and variance guarantees. https://arxiv.org/abs/1806.10234

JH Huggins, M Kasprzak, T Campbell, and T Broderick. Bayesian posterior mean and uncertainty estimates: a non-asymptotic approach. In preparation.

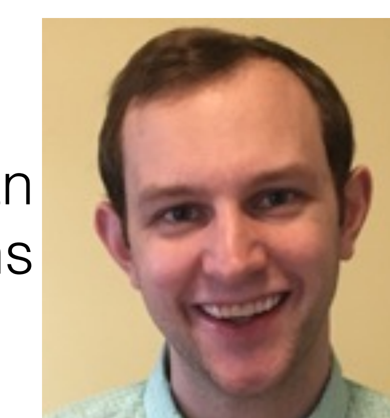
Raj Agrawal

Trevor Campbell

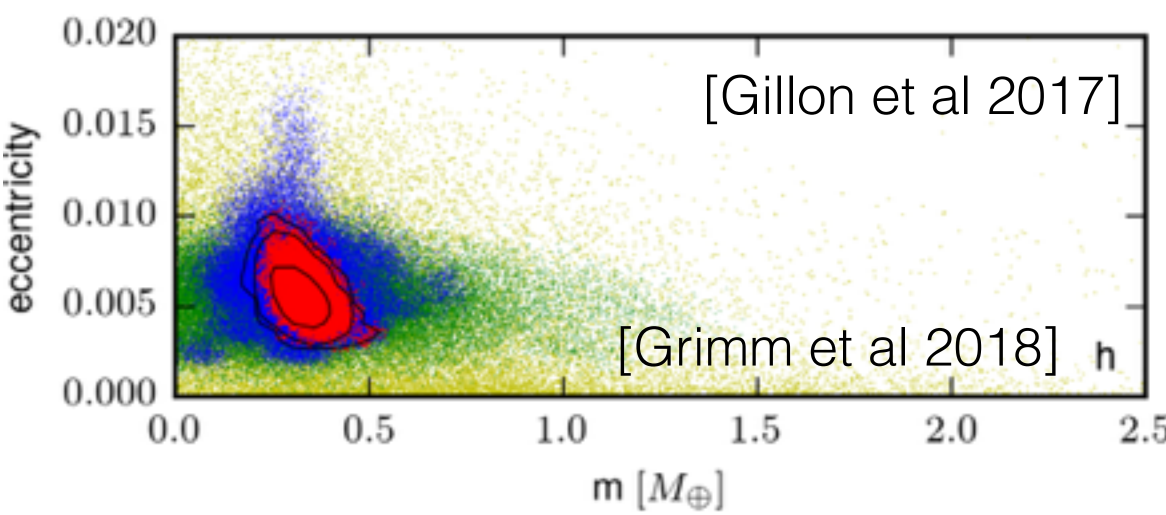
Ryan Giordano

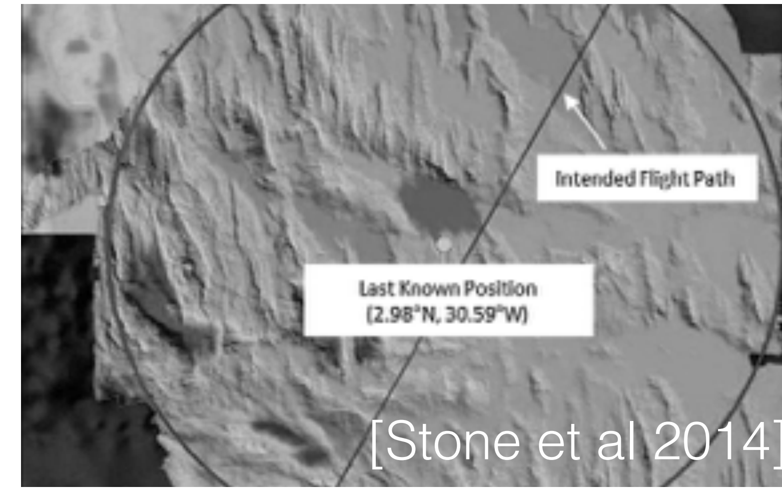


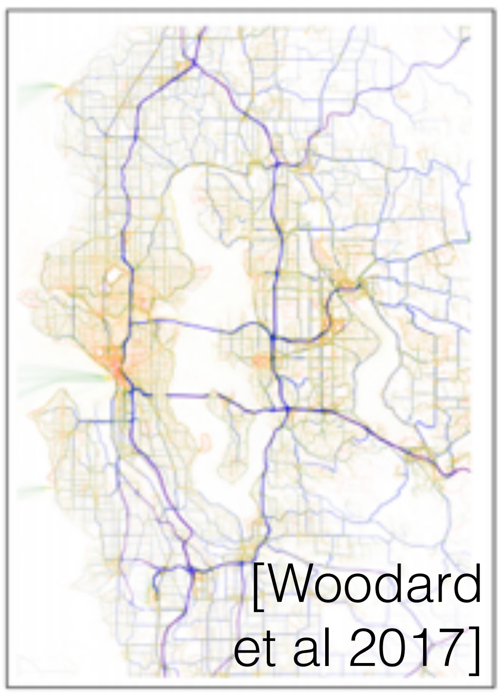
Jonathan Huggins

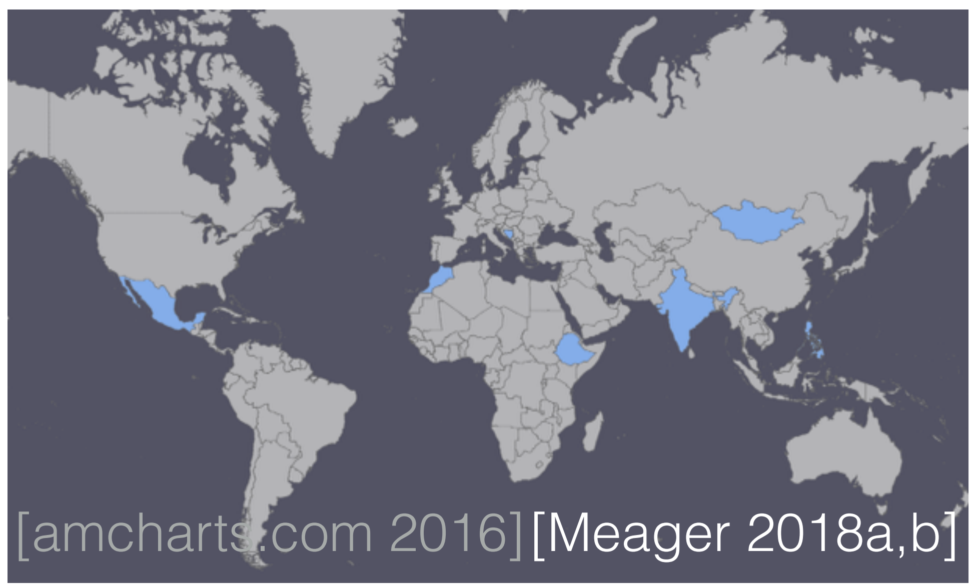


Bayesian inference



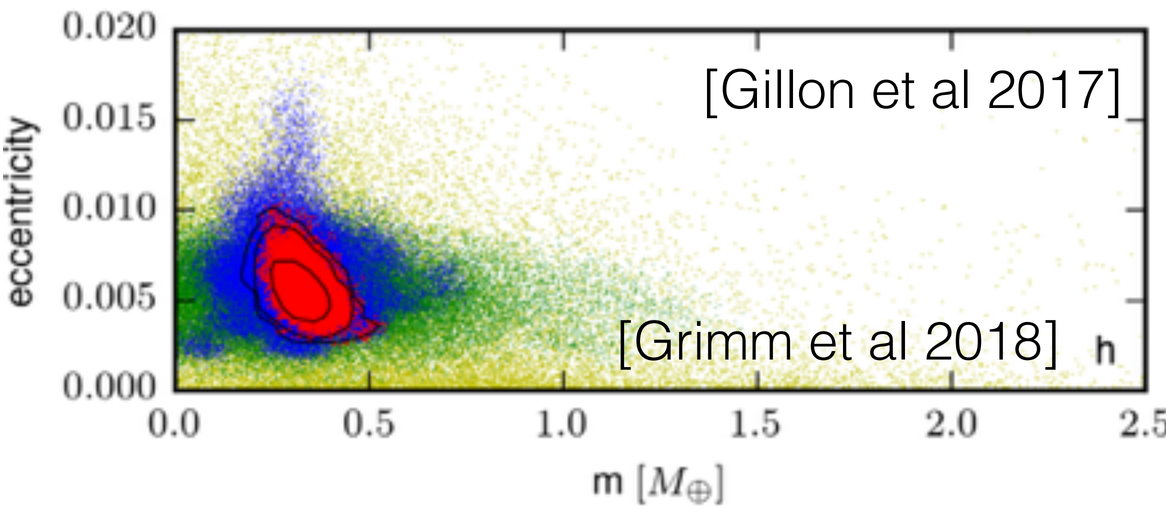


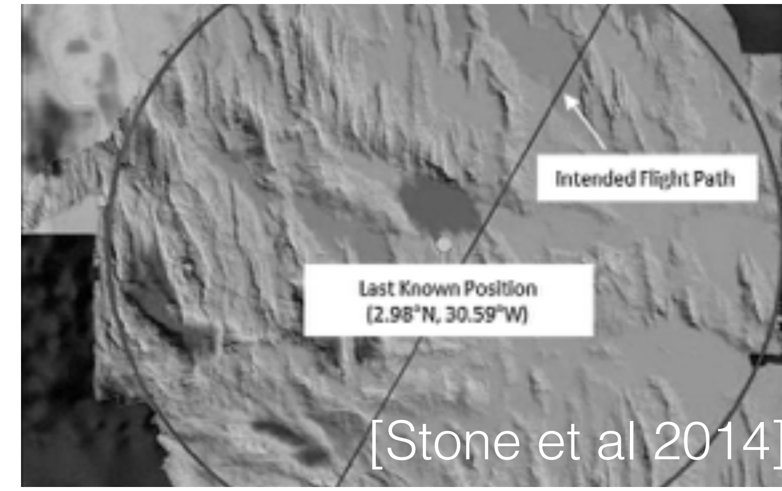


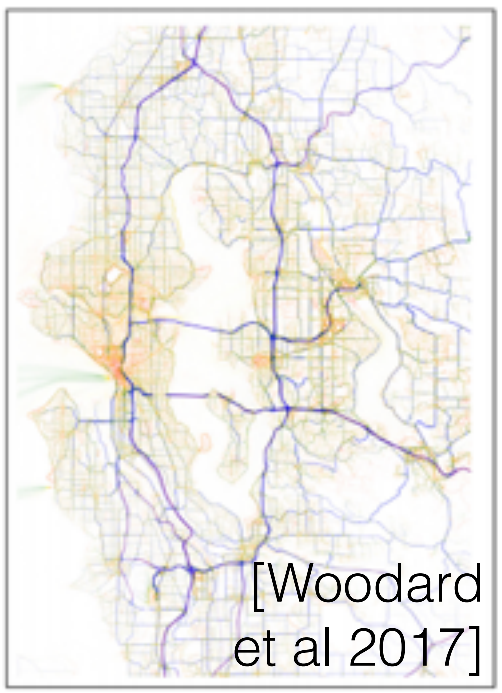


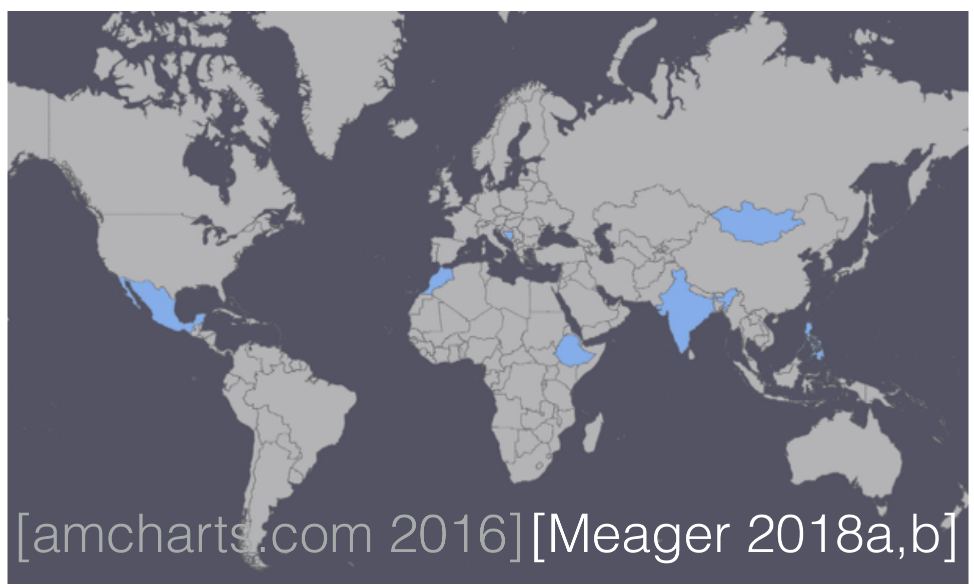
[mc-stan.org]

Bayesian inference





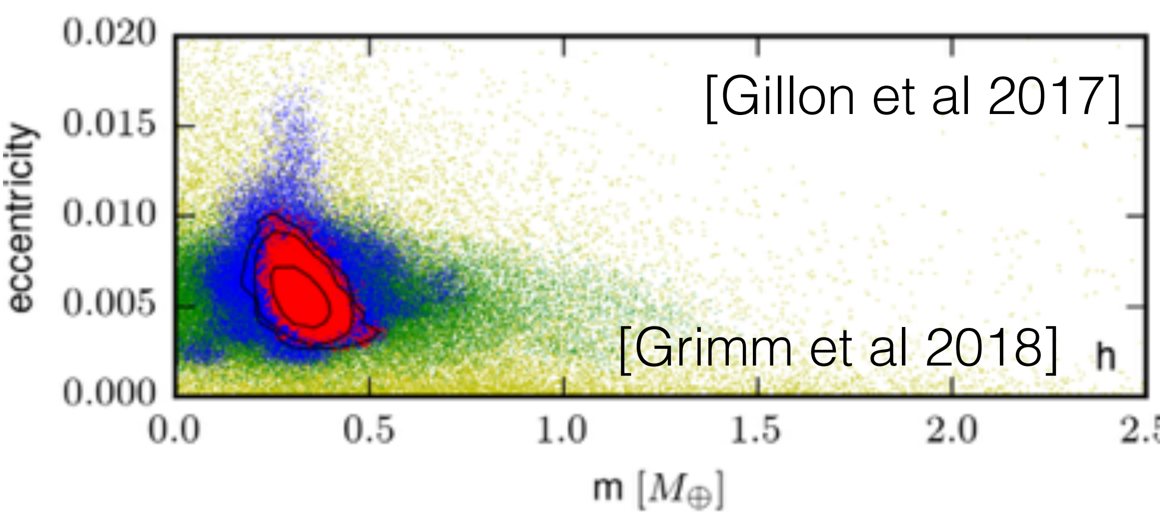


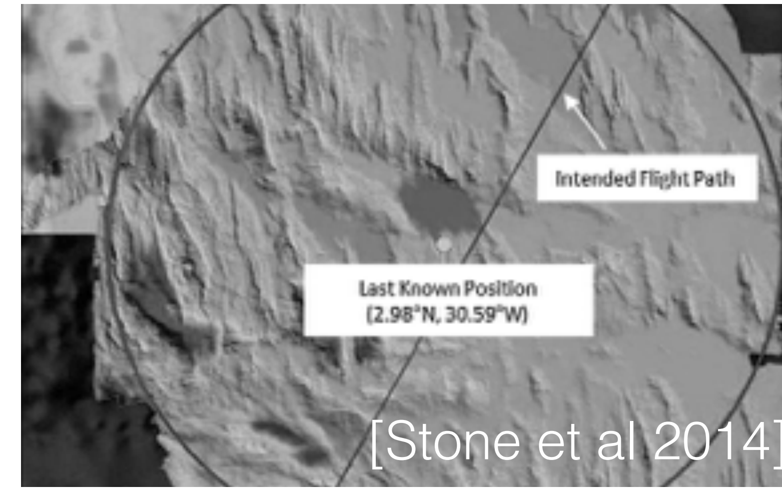


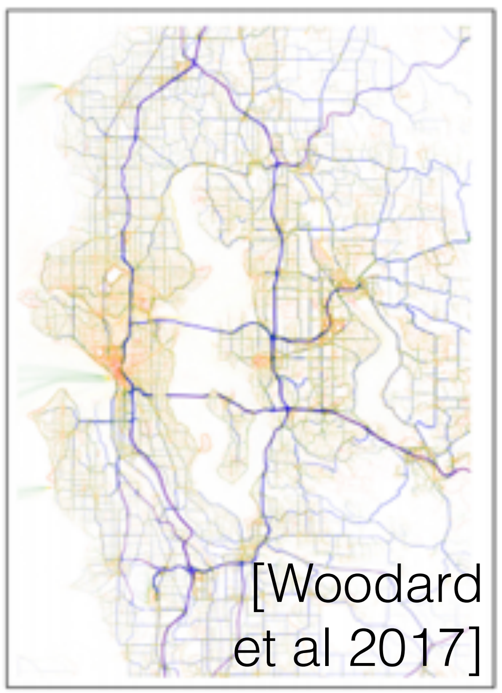
[mc-stan.org]

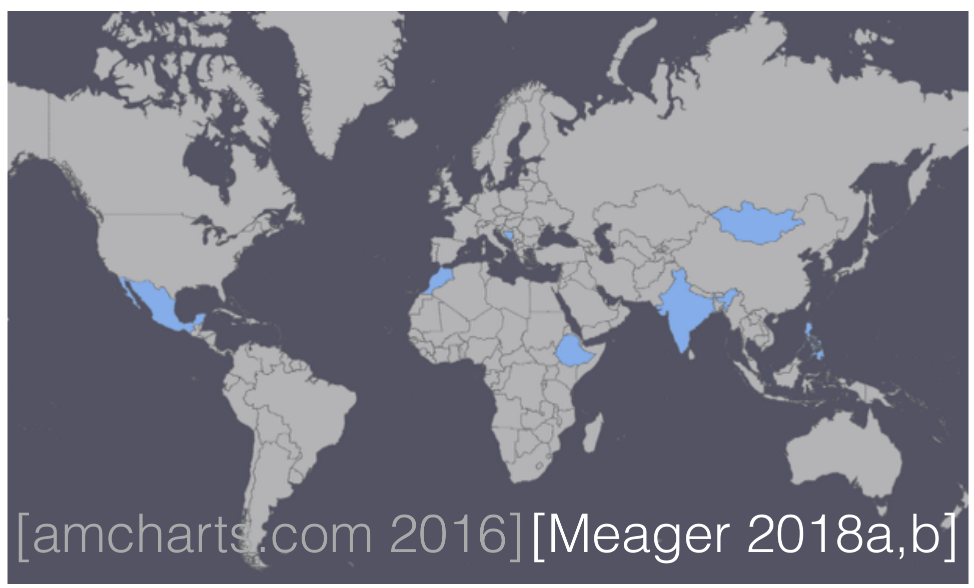
- Challenge: fast (compute, user), reliable inference

Bayesian inference









[mc-stan.org]

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

Further References (3/8)

[See earlier slides for first two pages of references]

PK Agarwal, S Har-Peled, and KR Varadarajan. Geometric approximation via coresets. *Combinatorial and Computational Geometry* 52 (2005): 1-30.

M Bădoiu, S Har-Peled, and P Indyk. Approximate clustering via core-sets. *Proceedings of the 34th Annual ACM Symposium on Theory of Computing*, 2002.

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

M Bauer, M van der Wilk, and CE Rasmussen. Understanding probabilistic sparse Gaussian process approximations. *NIPS* 2016.

AG Baydin, BA Pearlmutter, AA Radul, and JM Siskind. Automatic differentiation in machine learning: a survey. ArXiv:1502.05767v4 (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. *NIPS* 2013.

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

K Chwialkowski, H Strathmann, and A Gretton. A kernel test of goodness of fit. *ICML* 2016.

W DuMouchel, C Volinsky, T Johnson, C Cortes, and D Pregibon. Squashing flat files flatter. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 6-15. ACM, 1999.

D Dunson. Robust and scalable approach to Bayesian inference. Talk at *ISBA* 2014.

References (4/8)

D Feldman, and M Langberg. A unified framework for approximating and clustering data. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pp. 569-578. ACM, 2011.

B Fosdick. *Modeling Heterogeneity within and between Matrices and Arrays*, Chapter 4.7. PhD Thesis, University of Washington, 2013.

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

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

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

MD Hoffman, and A Gelman. The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* 15, no. 1 (2014): 1593-1623.

M Jaggi. Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization. *ICML* 2013.

W Jitkrittum, W Xu, Z Szabó, K Fukumizu, and A Gretton. A linear-time kernel goodness-of-fit test. *NIPS* 2017.

References (5/8)

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

A Kucukelbir, D Tran, R Ranganath, A Gelman, and DM Blei. Automatic differentiation variational inference. *The 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.

D Madigan, N Raghavan, W Dumouchel, M Nason, C Posse, and G Ridgeway. Likelihood-based data squashing: A modeling approach to instance construction. *Data Mining and Knowledge Discovery* 6, no. 2 (2002): 173-190.

M Opper and O Winther. Variational linear response. *NIPS* 2003.

G Rosman, M Volkov, D Feldman, JW Fisher III, D Rus. Coresets for k-segmentation of streaming data. *NIPS* 2014.

S Talts, M Betancourt, D Simpson, A Vehtari, and A Gelman. "Validating Bayesian Inference Algorithms with Simulation-Based Calibration." aArXiv: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.

B Wang and M Titterington. Inadequacy of interval estimates corresponding to variational Bayesian approximations. *AISTATS*, 2004.

Application References (6/8)

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.

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

Application References (7/8)

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

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

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

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

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

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

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

Additional image references (8/8)

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

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

CSIRO. 11 November 2004. Obtained from: https://commons.wikimedia.org/wiki/File:Leptoconops_spp._from_CSIRO.jpg & http://www.ces.csiro.au/aicn/system/c_1153.htm Accessed: 2018.

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

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

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