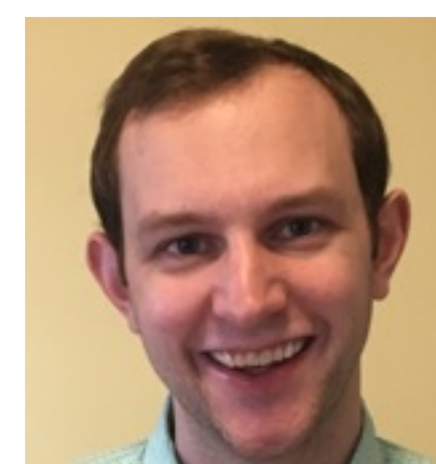
Automated Scalable Bayesian Inference via Data Summarization

Tamara Broderick
ITT Career Development
Assistant Professor,
MIT

With: Trevor Campbell, Jonathan H. Huggins



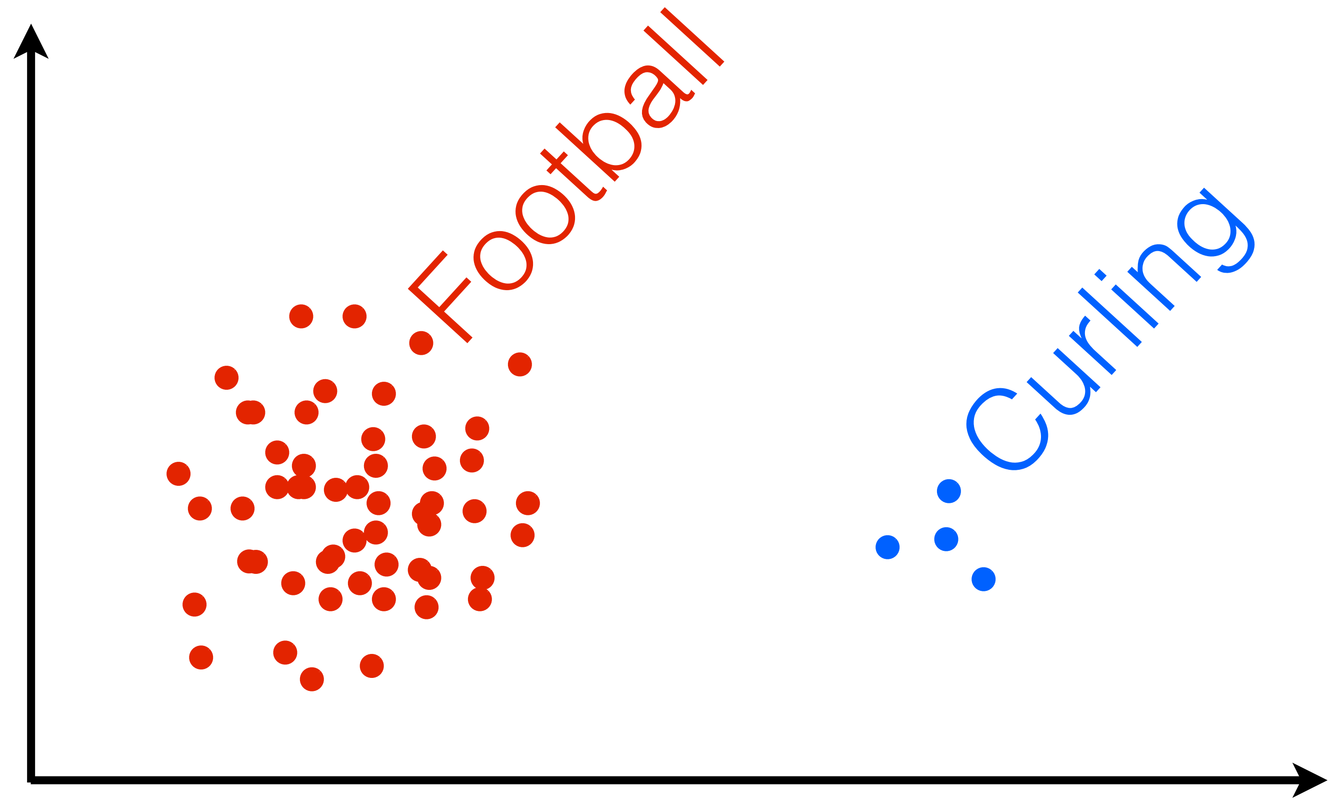
“Core” of the data set

“Core” of the data set

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

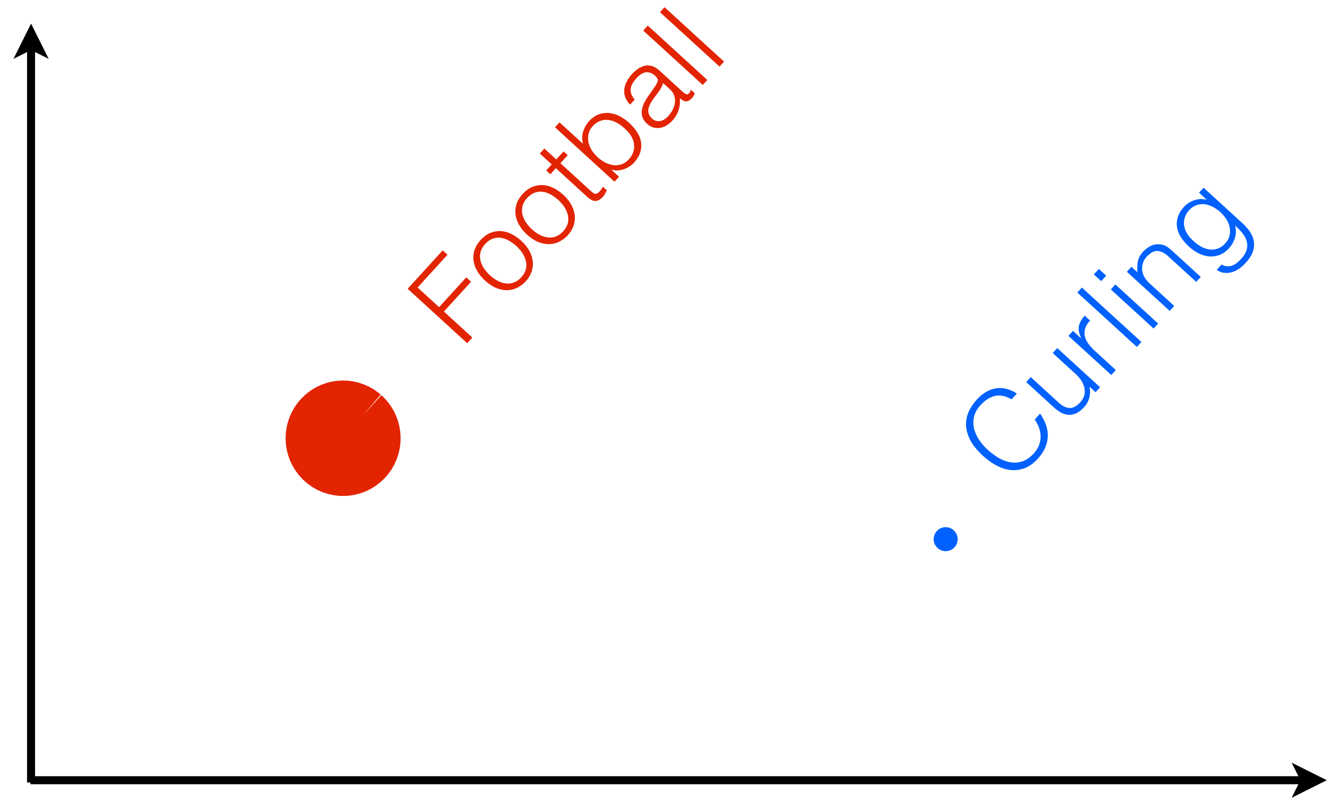
“Core” of the data set

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



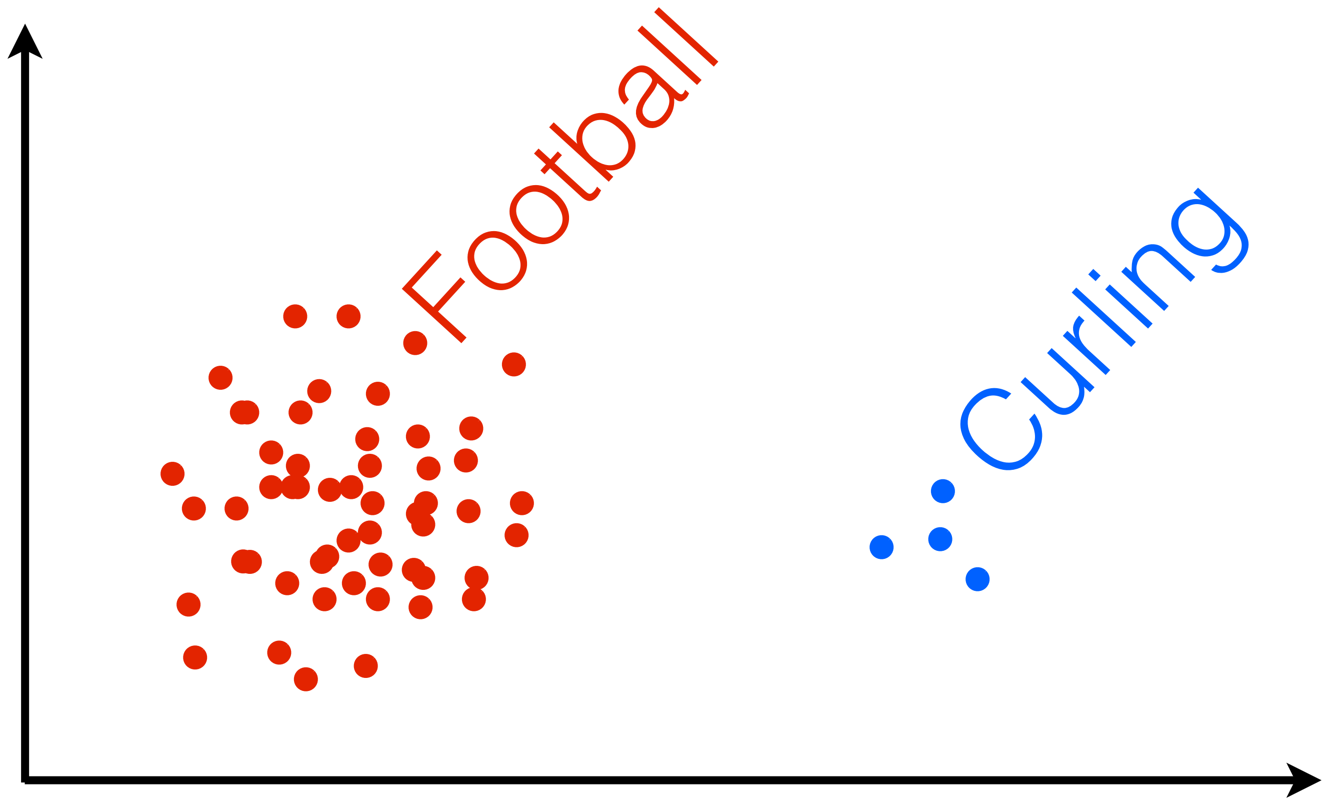
“Core” of the data set

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



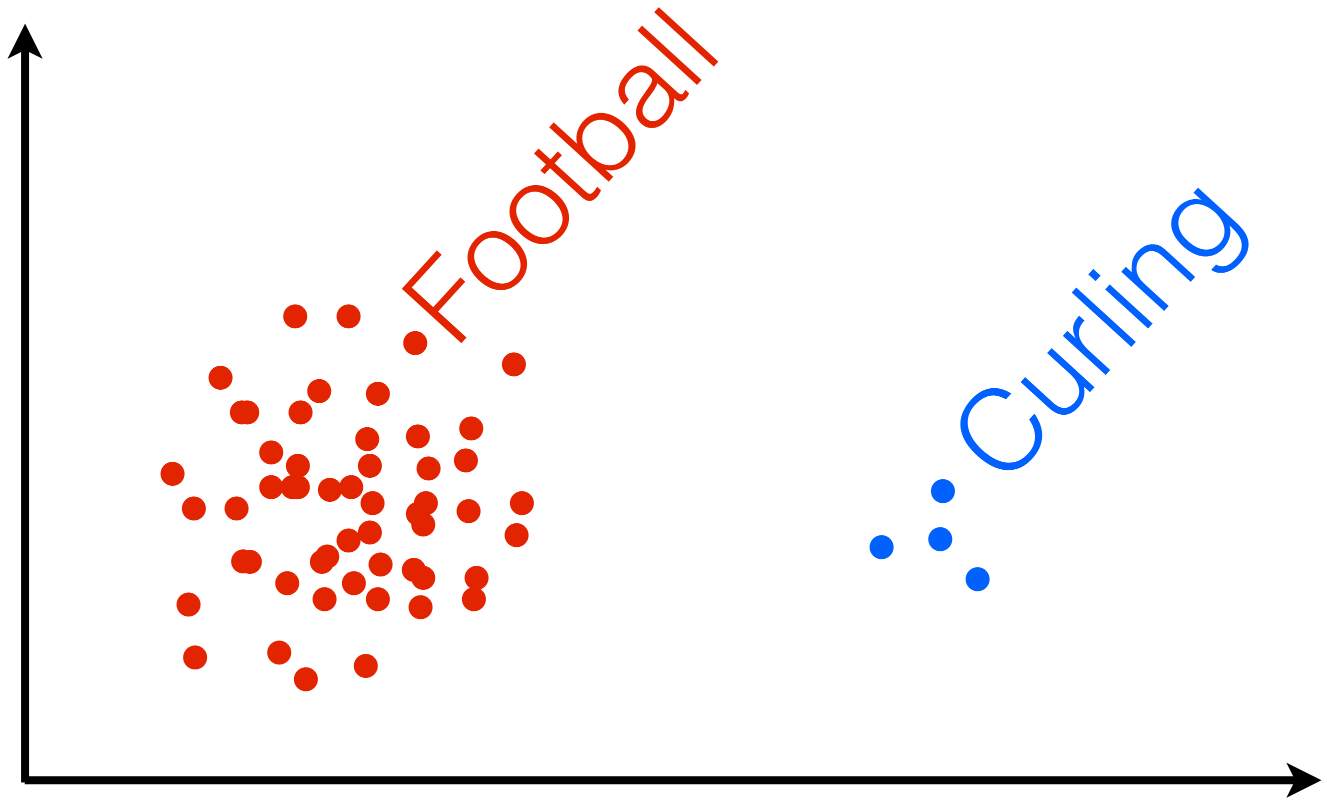
“Core” of the data set

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



“Core” of the data set

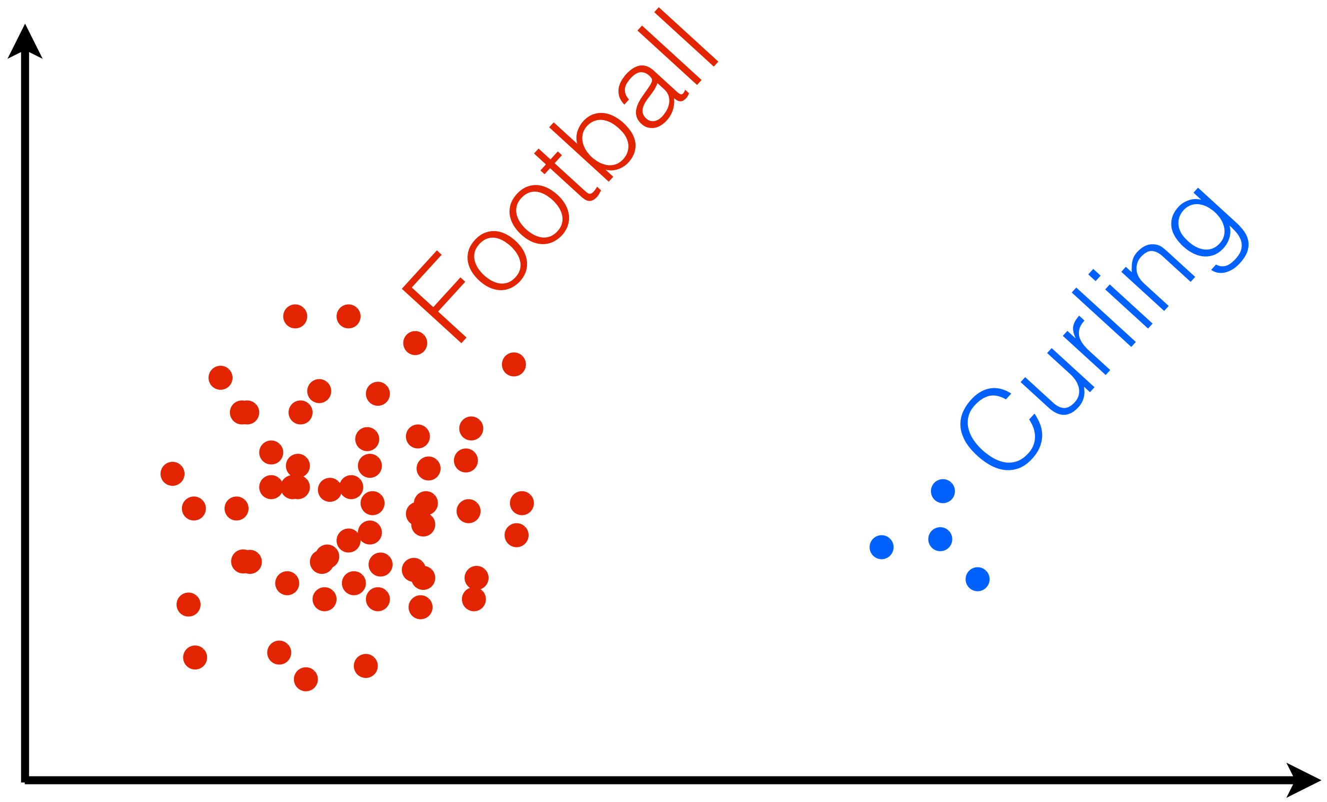
- Observe: redundancies can exist even if data isn’t “tall”
- Coresets: pre-process data to get a smaller, weighted data set



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

“Core” of the data set

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

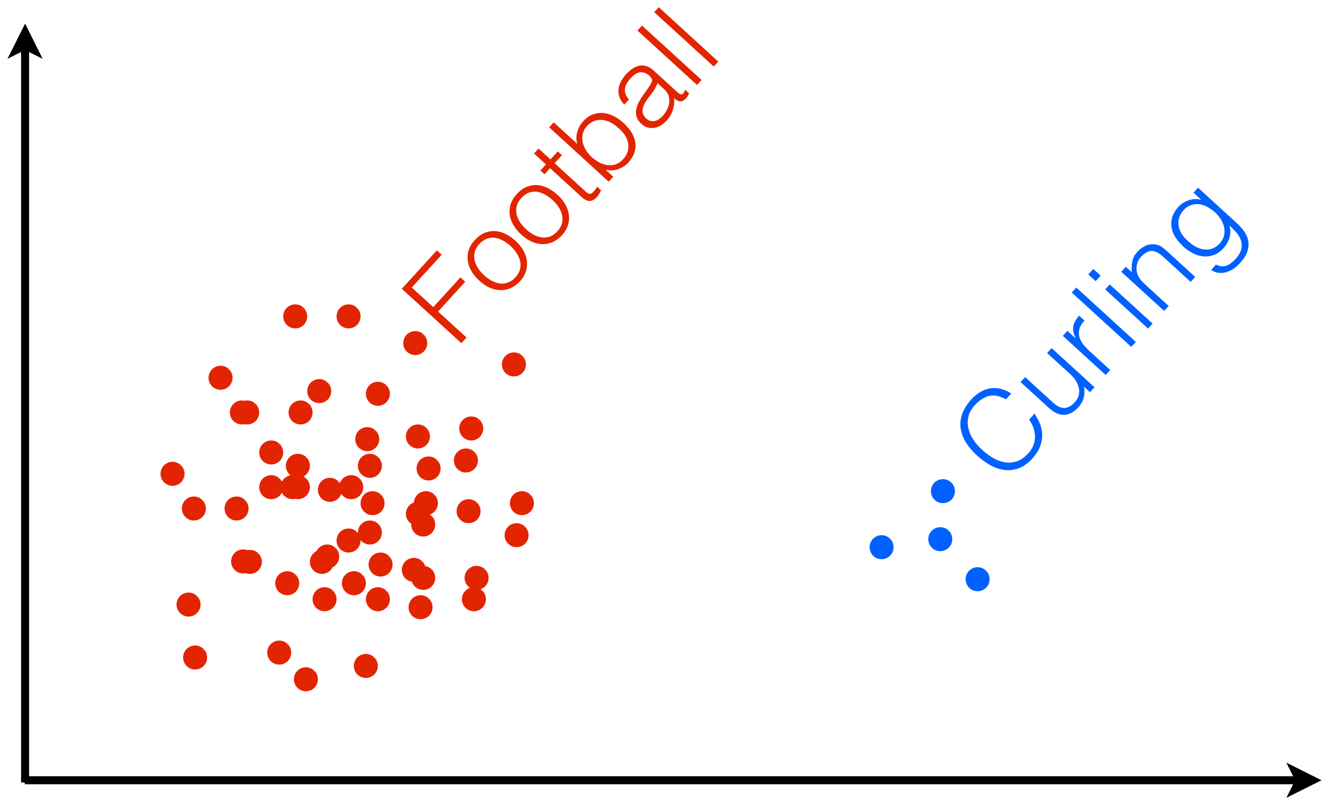


- Theoretical guarantees on quality

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

“Core” of the data set

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

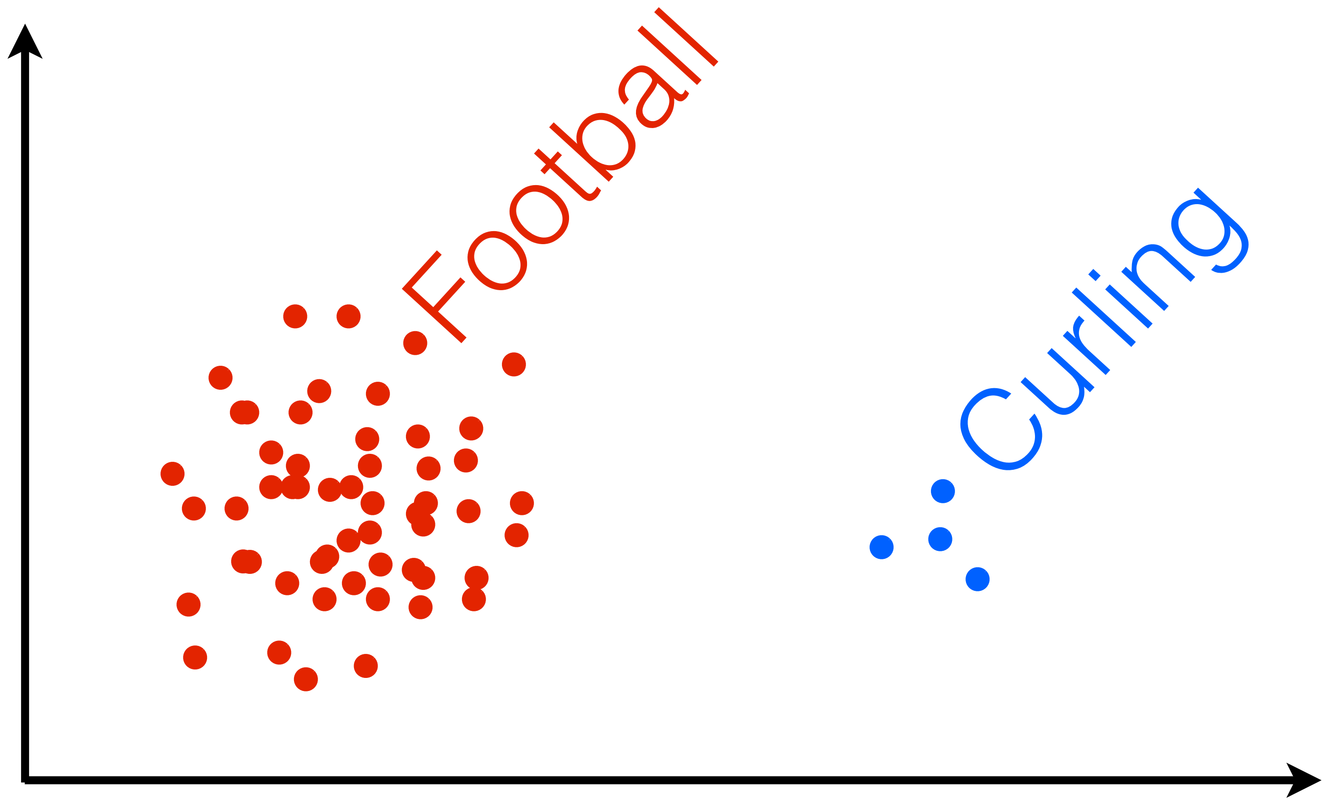


- Theoretical guarantees on quality
- How to develop **coresets for diverse tasks/geometries**?

[Bădoiu, Har-Peled, Indyk 2002; Agarwal et al 2005; Feldman & Langberg 2011; Huggins, Campbell, Broderick 2016; Campbell, Broderick 2017; Campbell, Broderick 2018]

“Core” of the data set

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

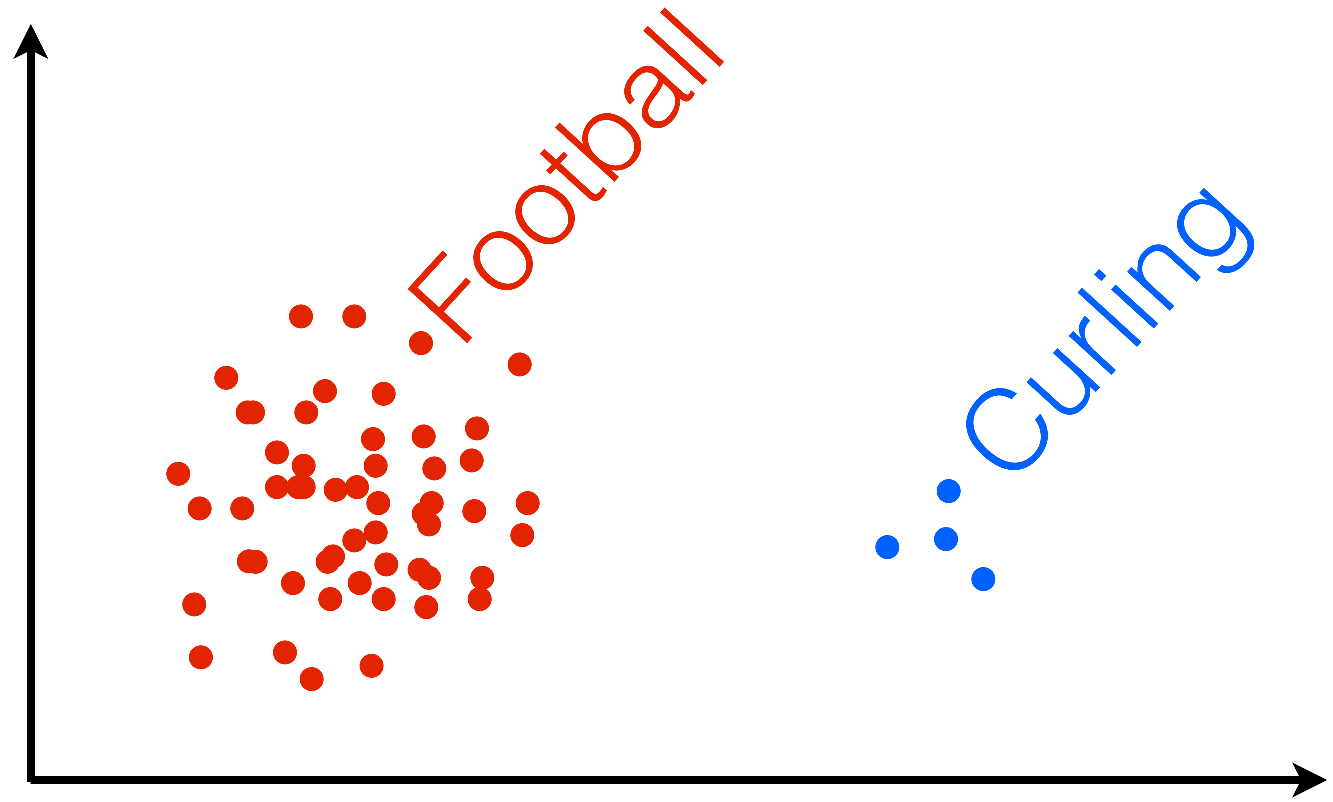


- Theoretical guarantees on quality
- How to develop **coresets for diverse tasks/geometries**?
- Previous heuristics: data squashing, big data GPs

[Bădoiu, Har-Peled, Indyk 2002; Agarwal et al 2005; Feldman & Langberg 2011; DuMouchel et al 1999; Madigan et al 2002; Huggins, Campbell, Broderick 2016; Campbell, Broderick 2017; Campbell, Broderick 2018]

“Core” of the data set

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

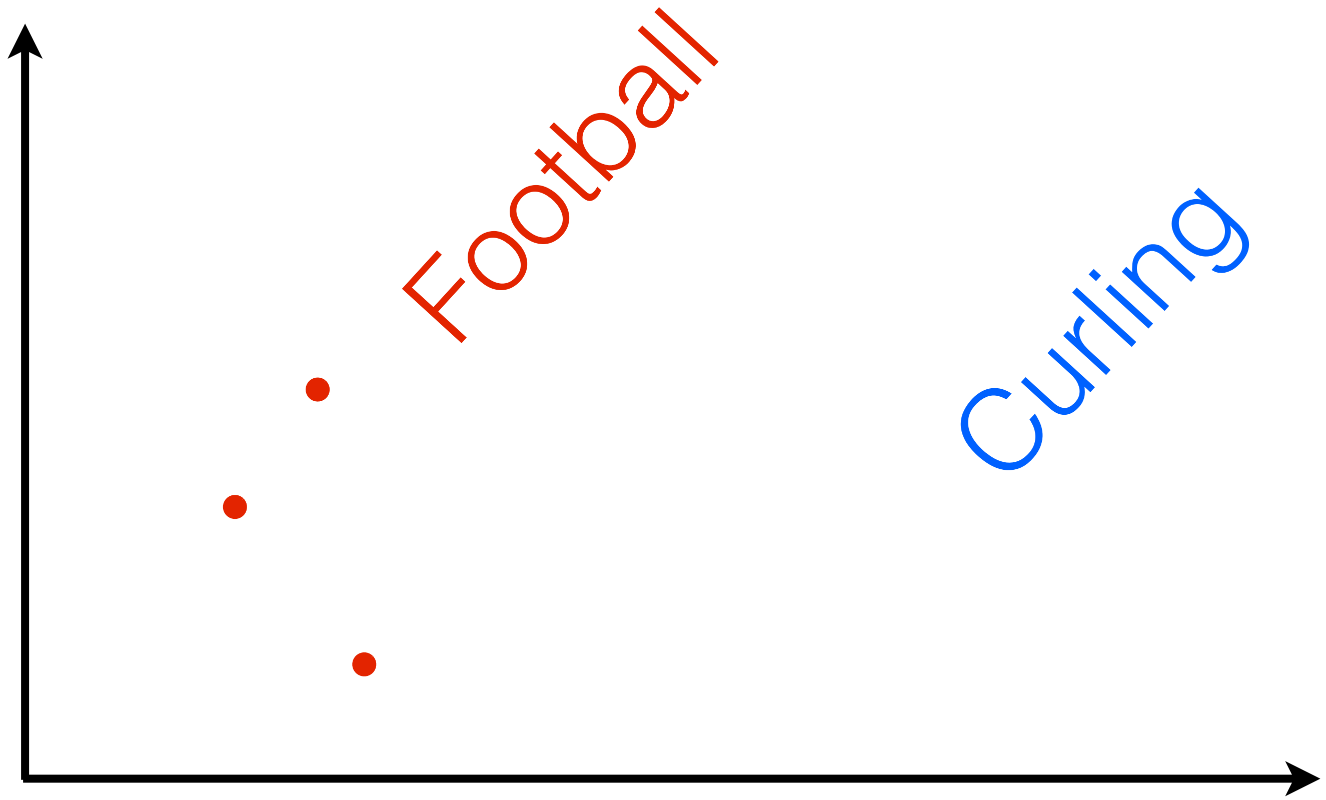


- Theoretical guarantees on quality
- How to develop **coresets for diverse tasks/geometries**?
- Previous heuristics: data squashing, big data GPs
- Cf. subsampling

[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 2017; Campbell, Broderick 2018]

“Core” of the data set

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

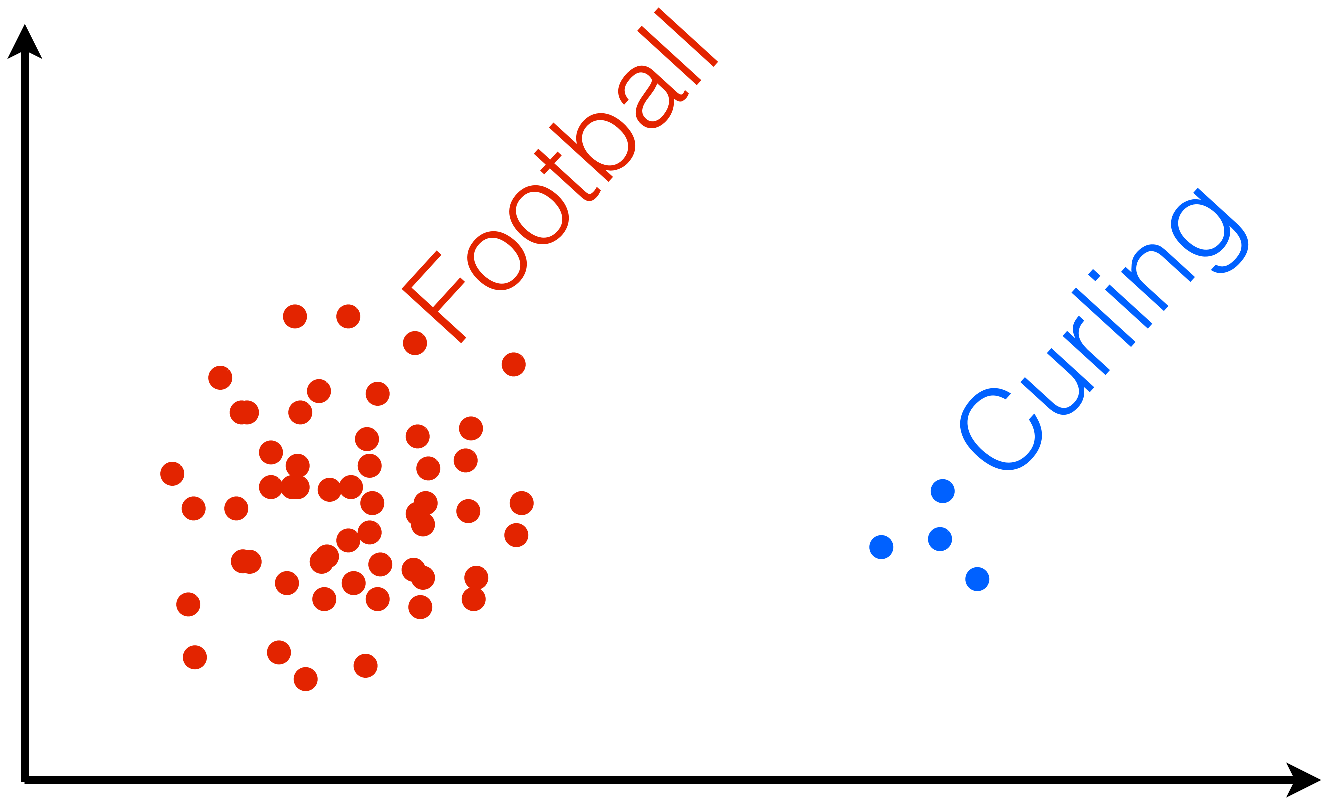


- Theoretical guarantees on quality
- How to develop **coresets for diverse tasks/geometries**?
- Previous heuristics: data squashing, big data GPs
- Cf. subsampling

[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 2017; Campbell, Broderick 2018]

“Core” of the data set

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



- Theoretical guarantees on quality
- How to develop **coresets for diverse tasks/geometries**?
- Previous heuristics: data squashing, big data GPs
- Cf. 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 2017; Campbell, Broderick 2018]

Roadmap

- The “core” of the data set

Roadmap

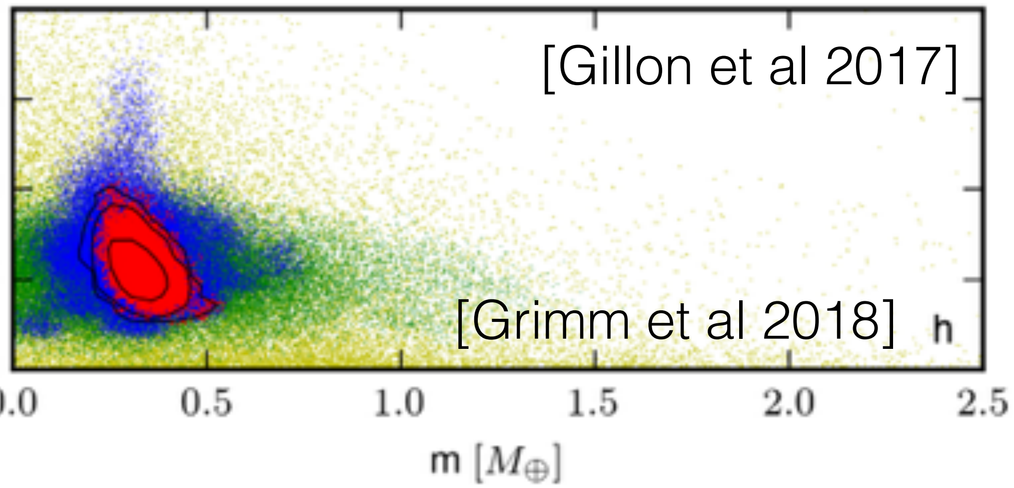
- The “core” of the data set
- Bayes setup
- Uniform data subsampling isn’t enough
- Importance sampling for “coresets”
- Optimization for “coresets”

Roadmap

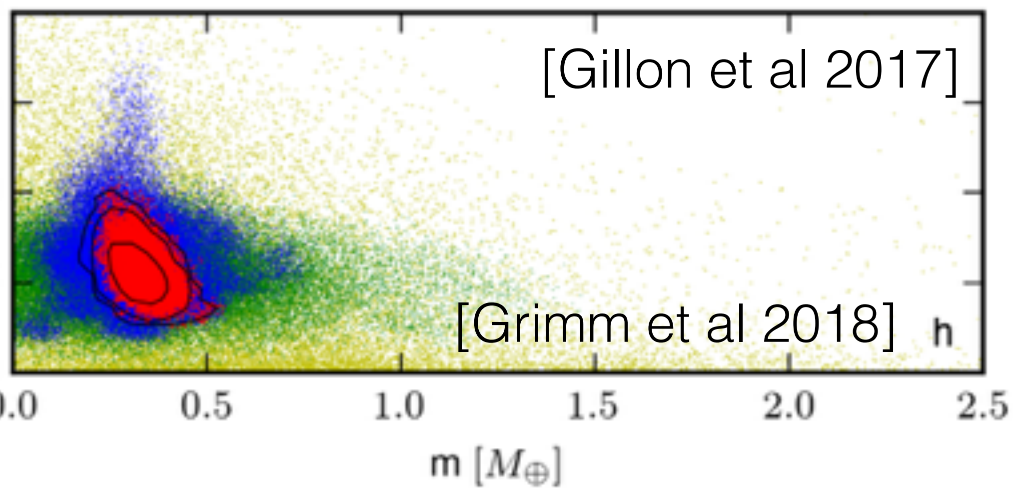
- The “core” of the data set
- Bayes setup
- Uniform data subsampling isn’t enough
- Importance sampling for “coresets”
- Optimization for “coresets”

Bayesian inference

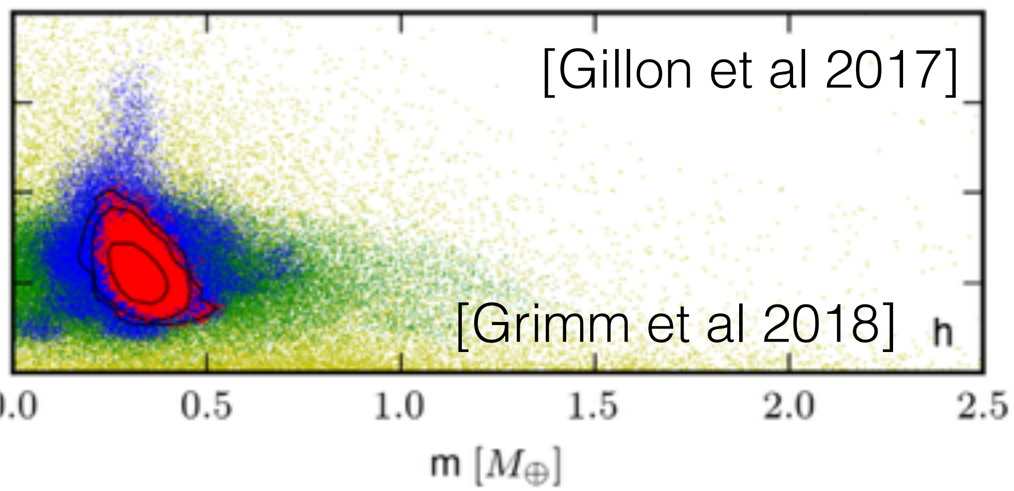
Bayesian inference

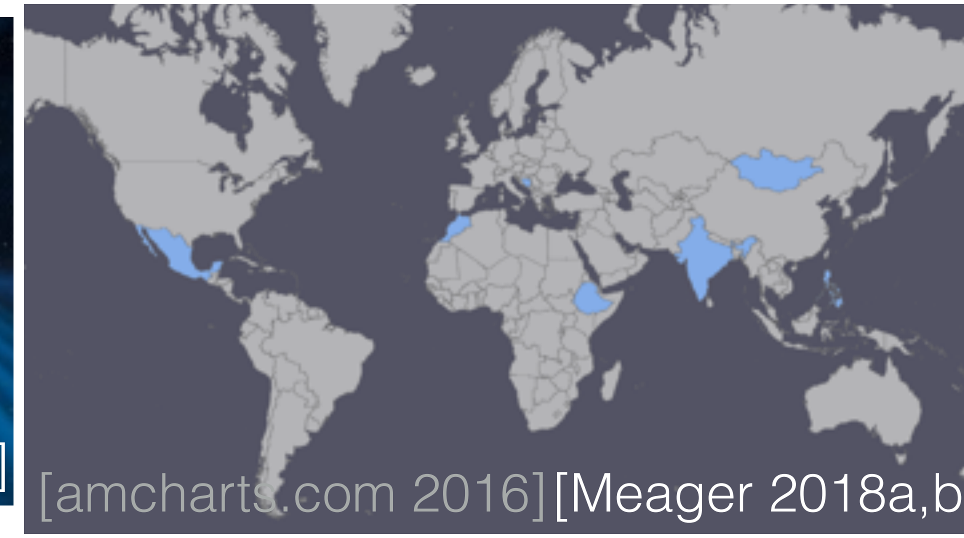


Bayesian inference

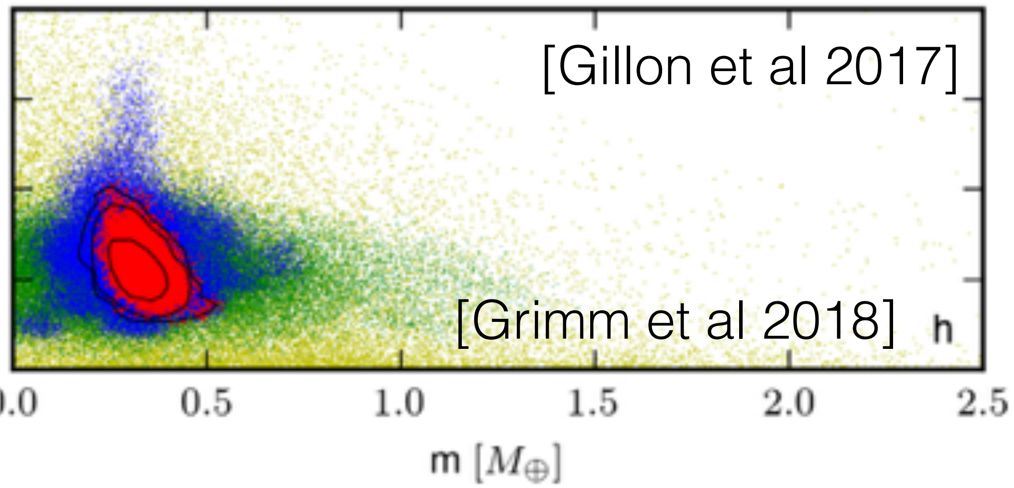


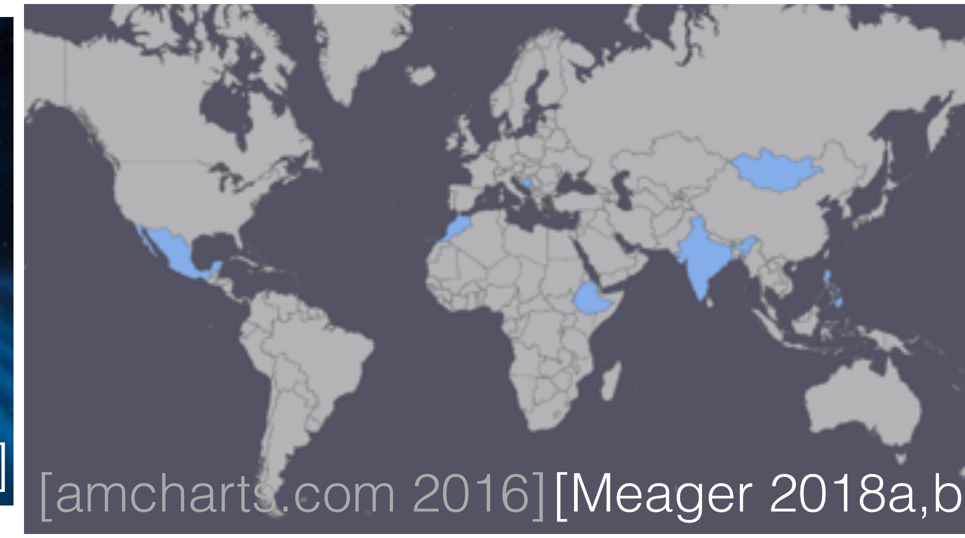
Bayesian inference

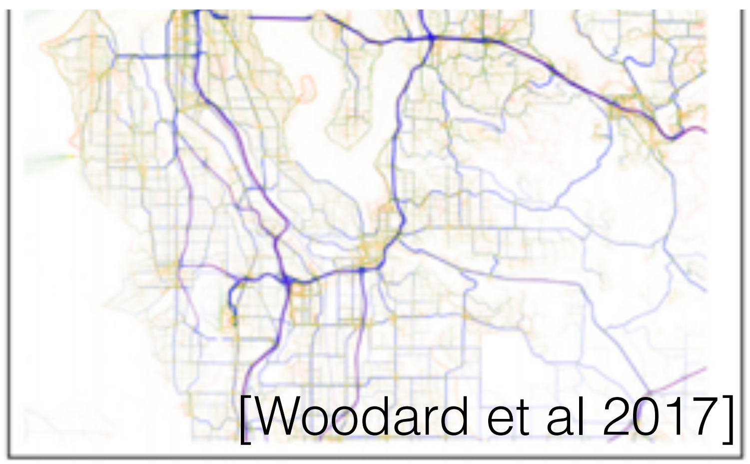




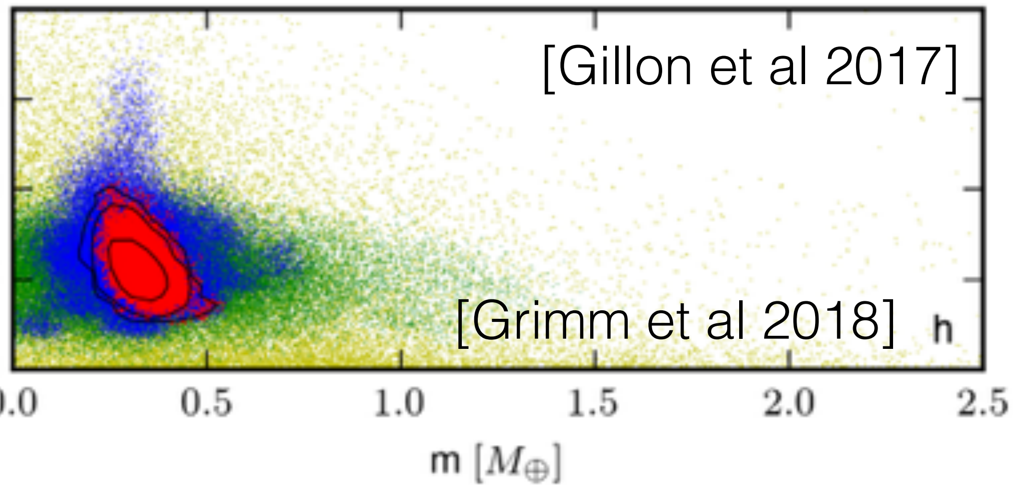
Bayesian inference

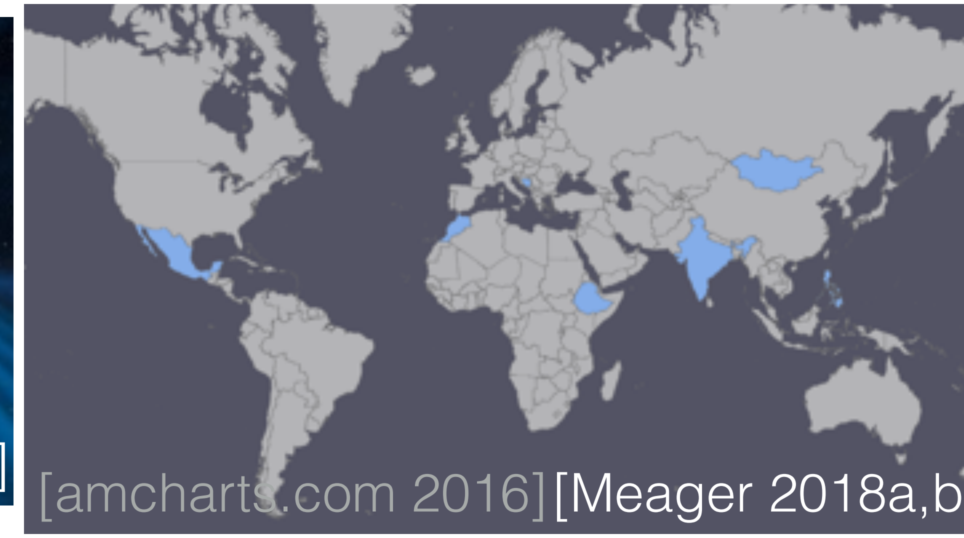


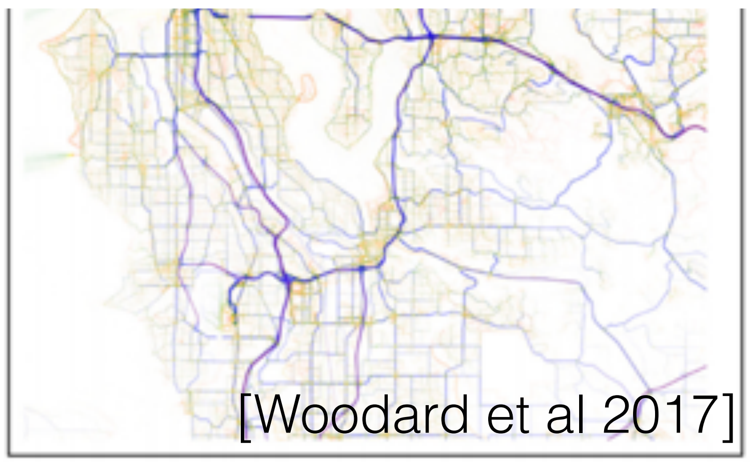




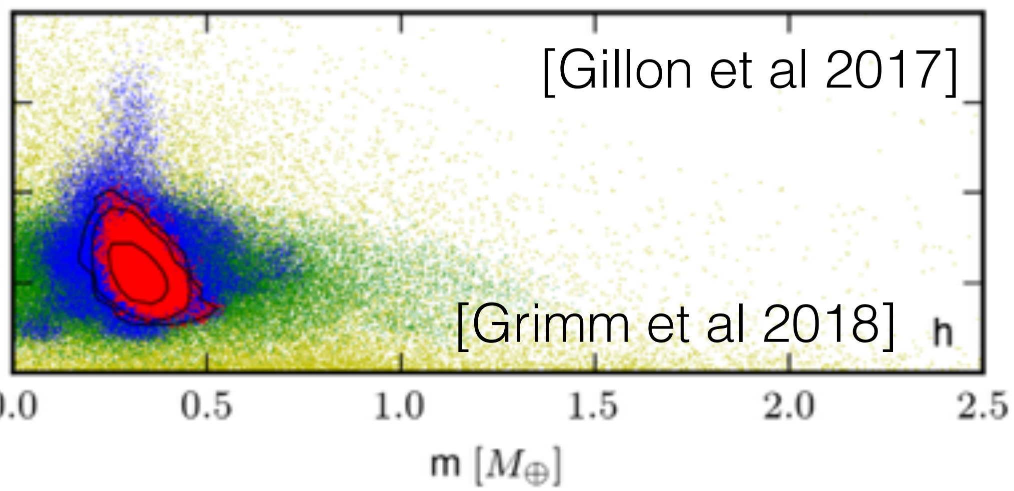
Bayesian inference

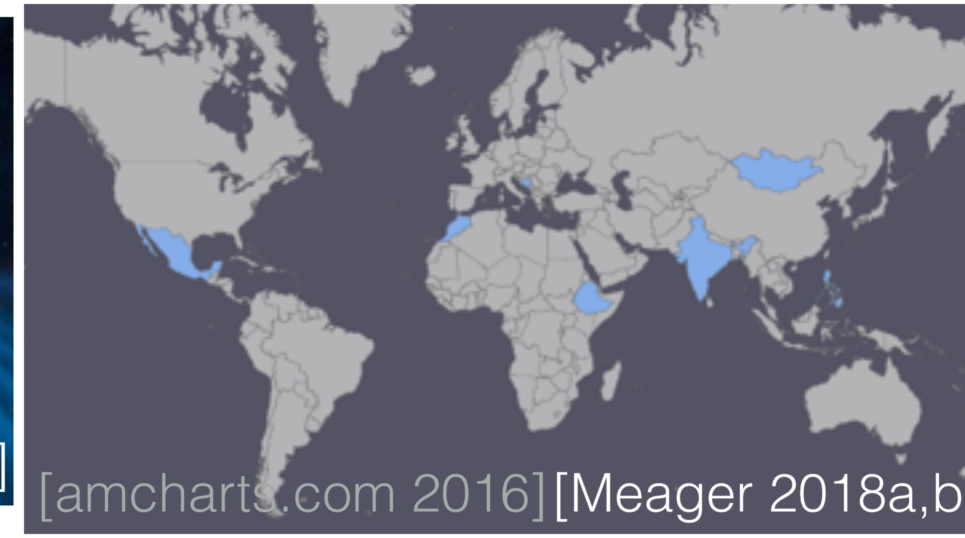


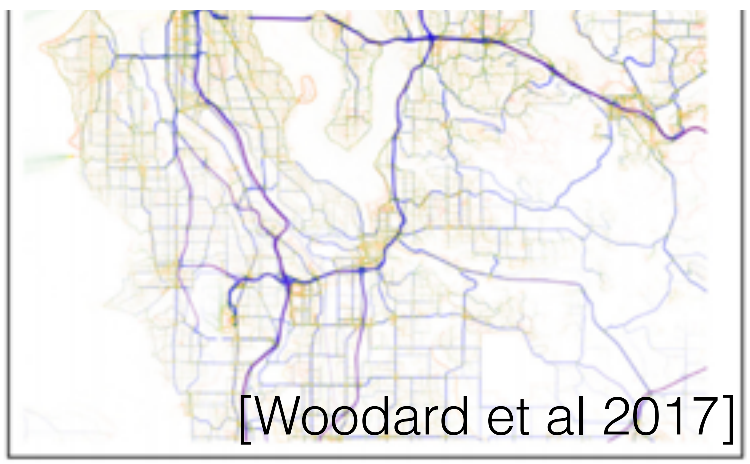




Bayesian inference

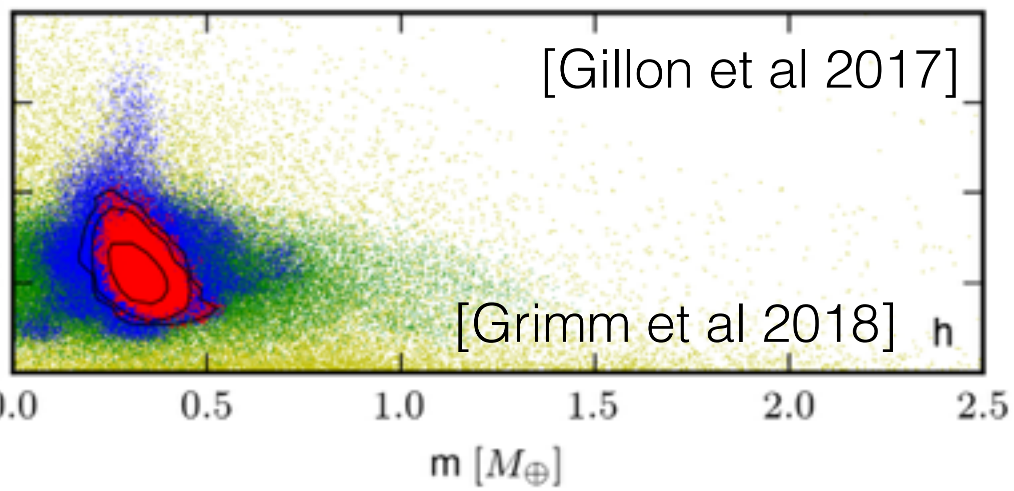


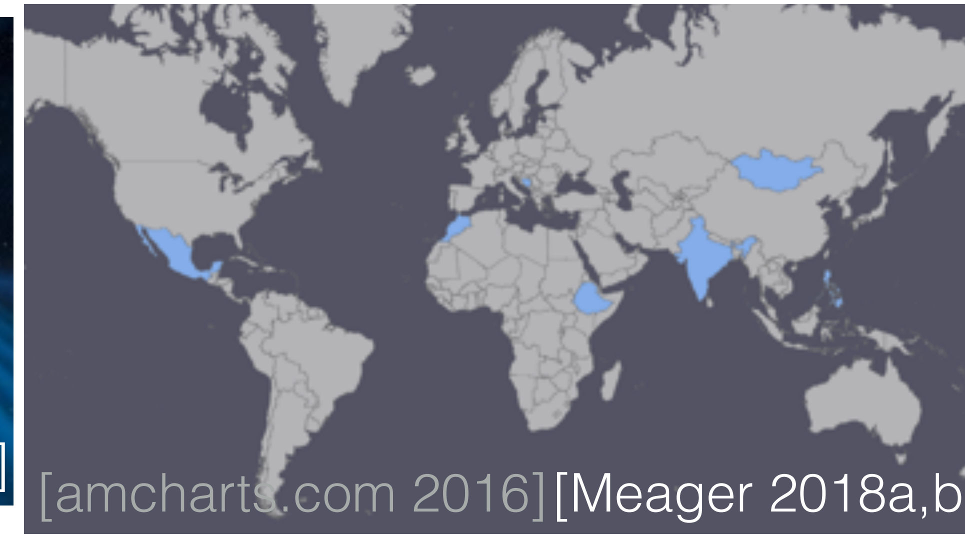


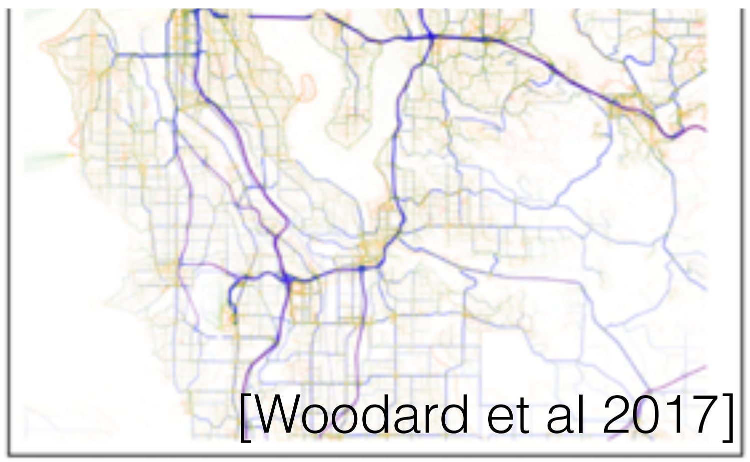


- Goal: Report point estimates, coherent uncertainties

Bayesian inference

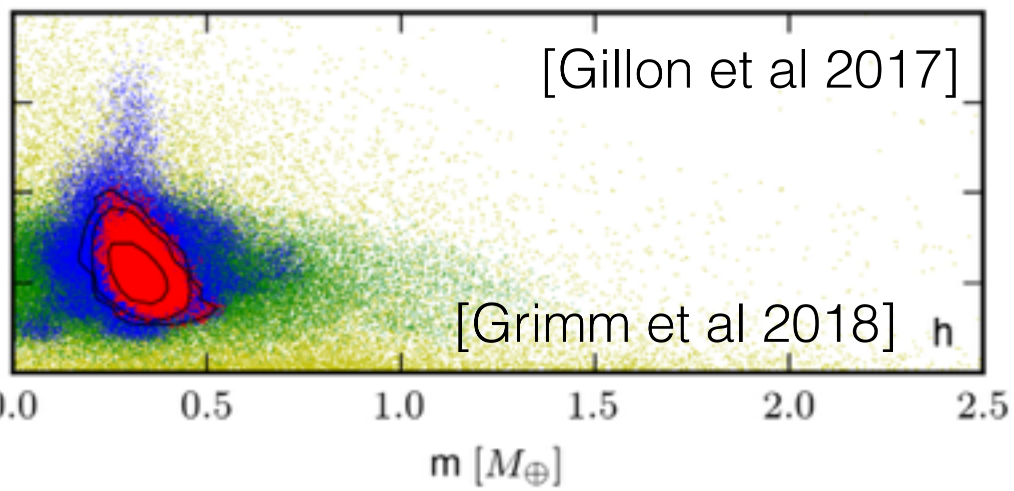


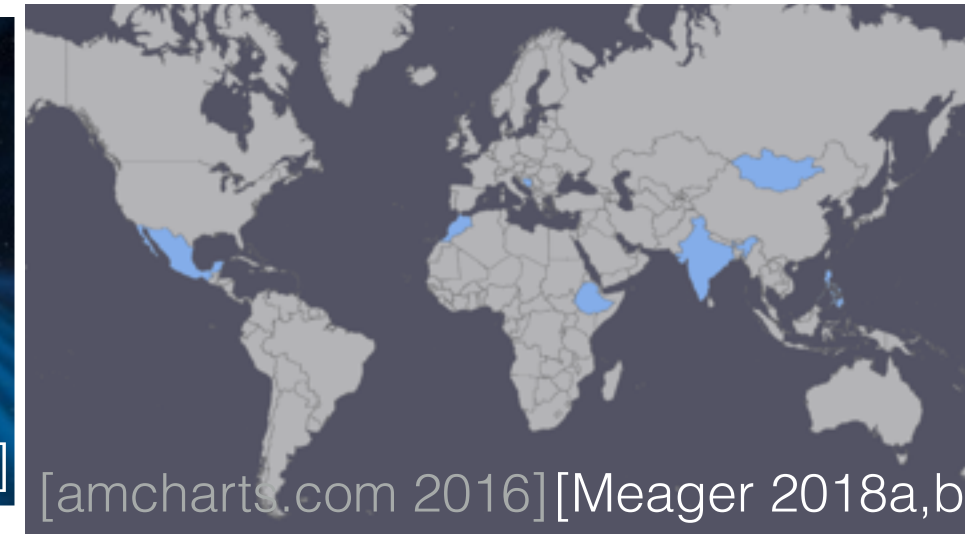


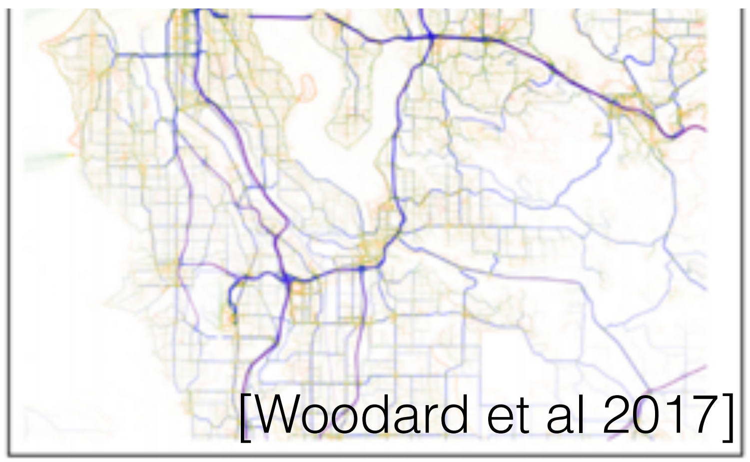


- Goal: Report point estimates, coherent uncertainties
- Challenge: existing methods can be slow, tedious, unreliable

Bayesian inference







- Goal: Report point estimates, coherent uncertainties
- Challenge: existing methods can be slow, tedious, unreliable
- Our proposal: use *efficient data summaries* for **scalable**, **automated** algorithms with **error bounds for finite data**

Bayesian inference

Bayesian inference

$p(\theta)$

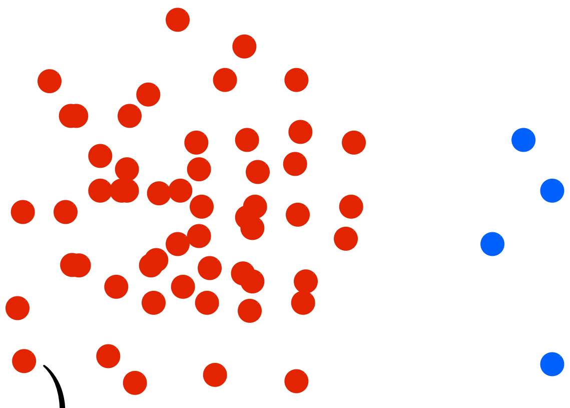
Bayesian inference

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

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

Bayesian inference

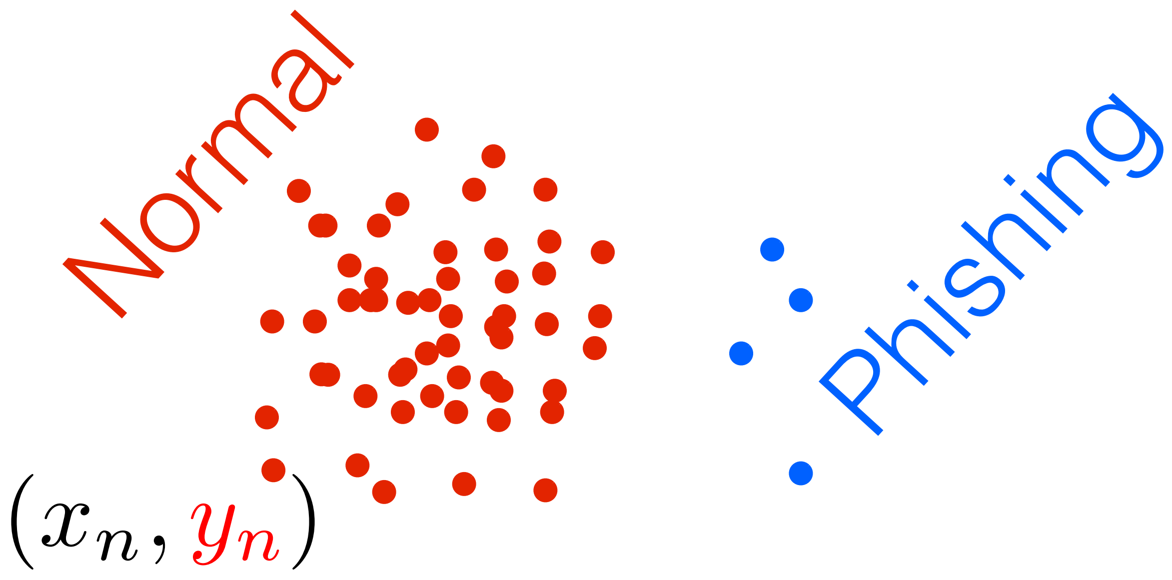
$$p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$$



(x_n, y_n)

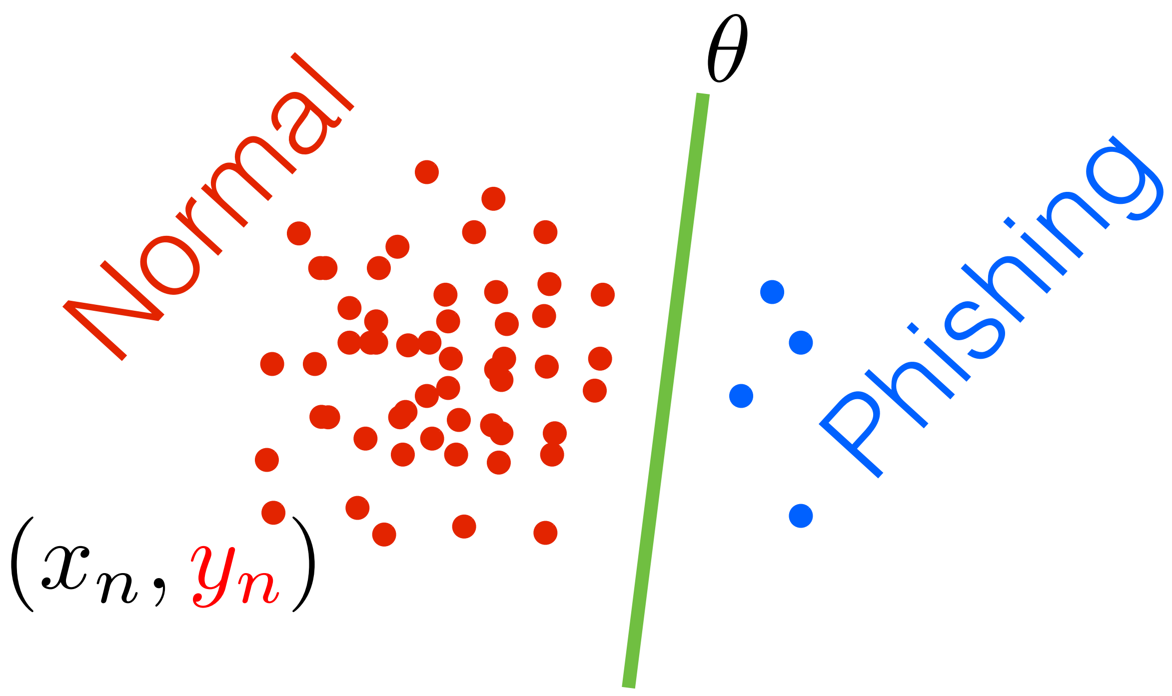
Bayesian inference

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



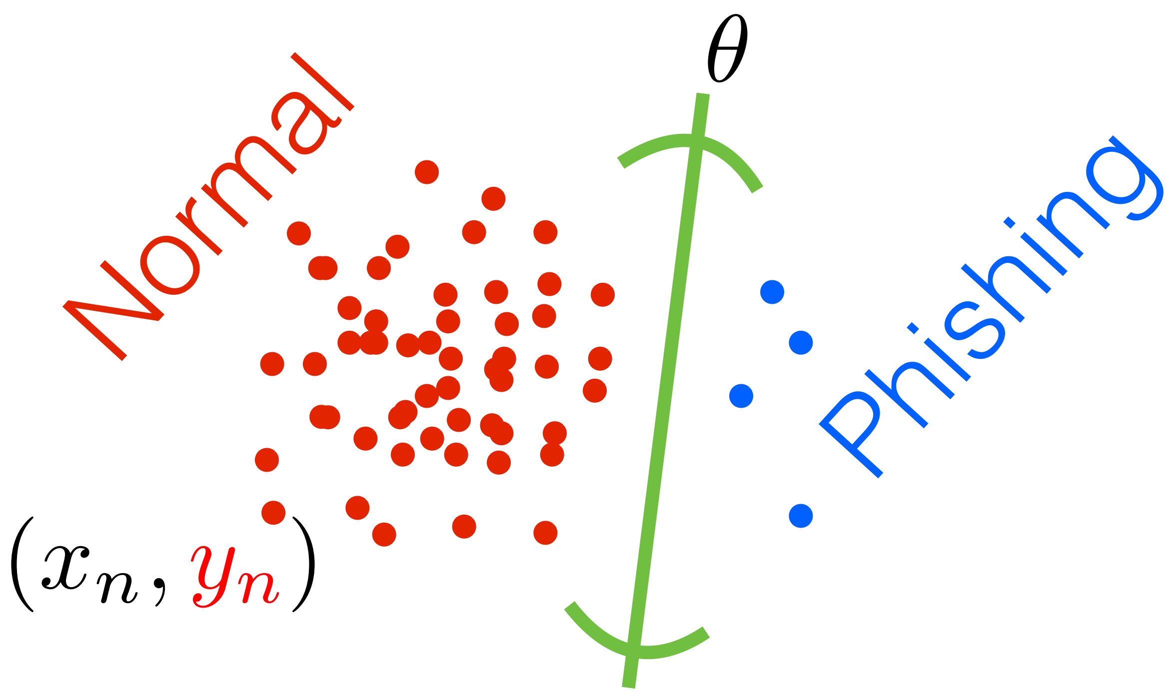
Bayesian inference

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



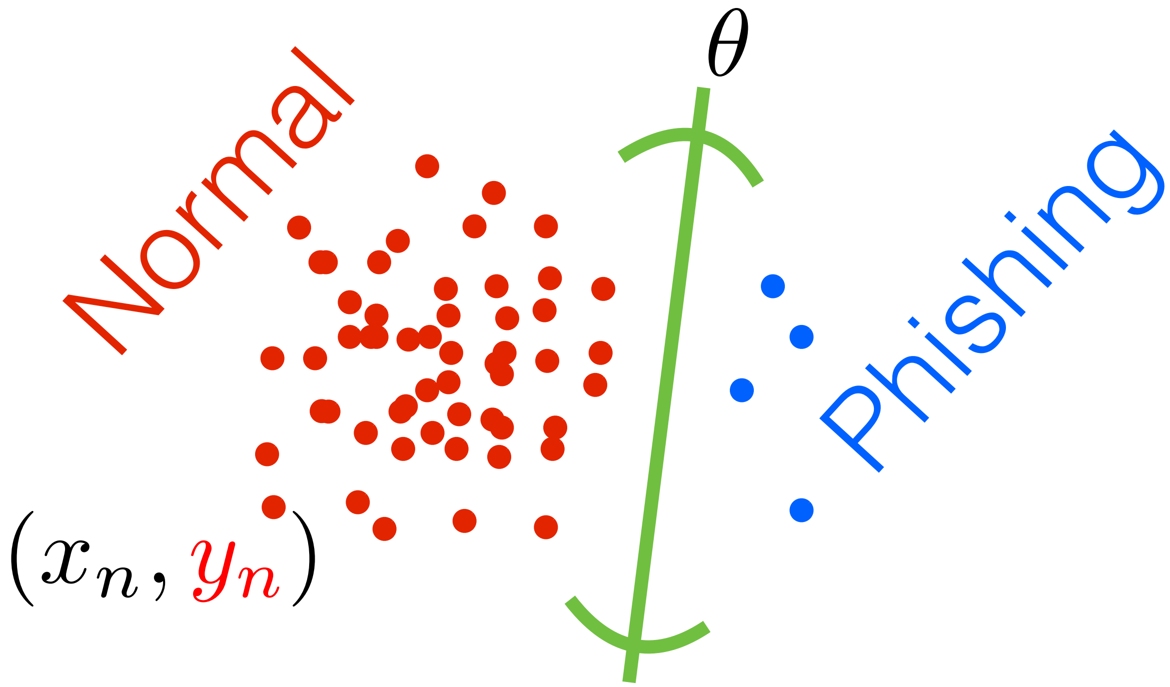
Bayesian inference

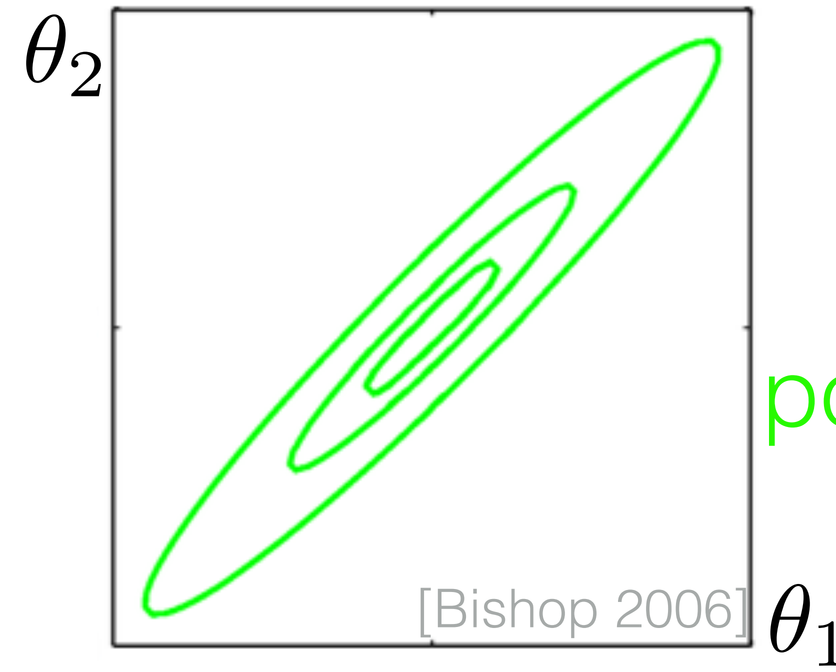
$$p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$$



Bayesian inference

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

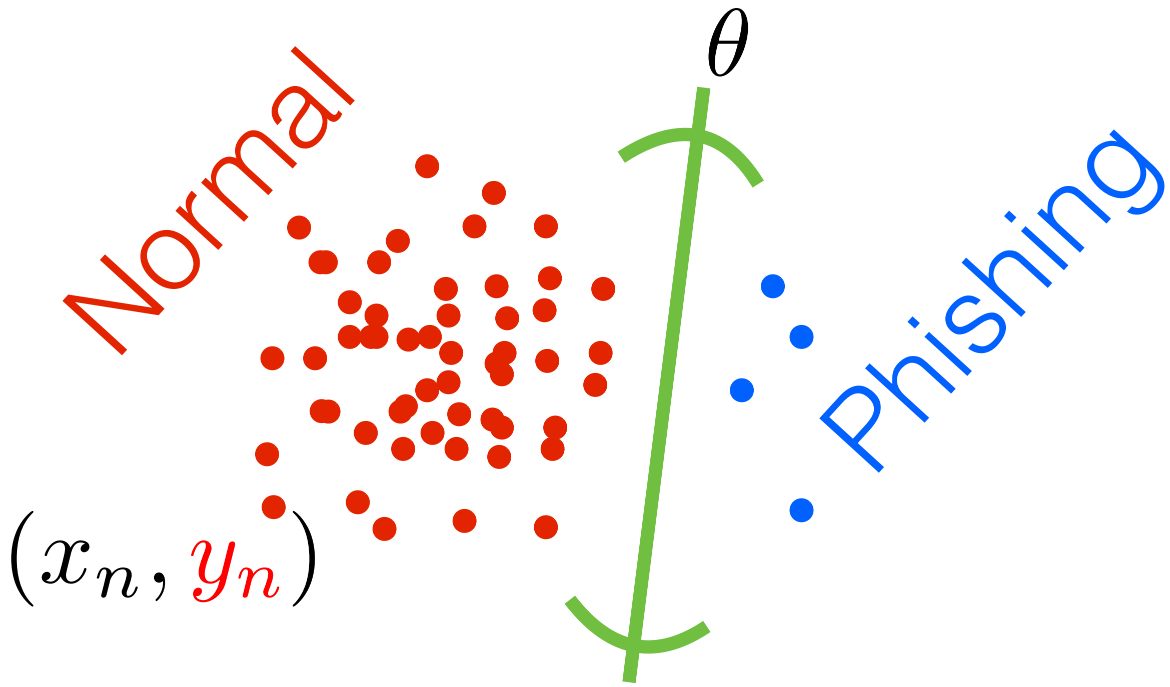


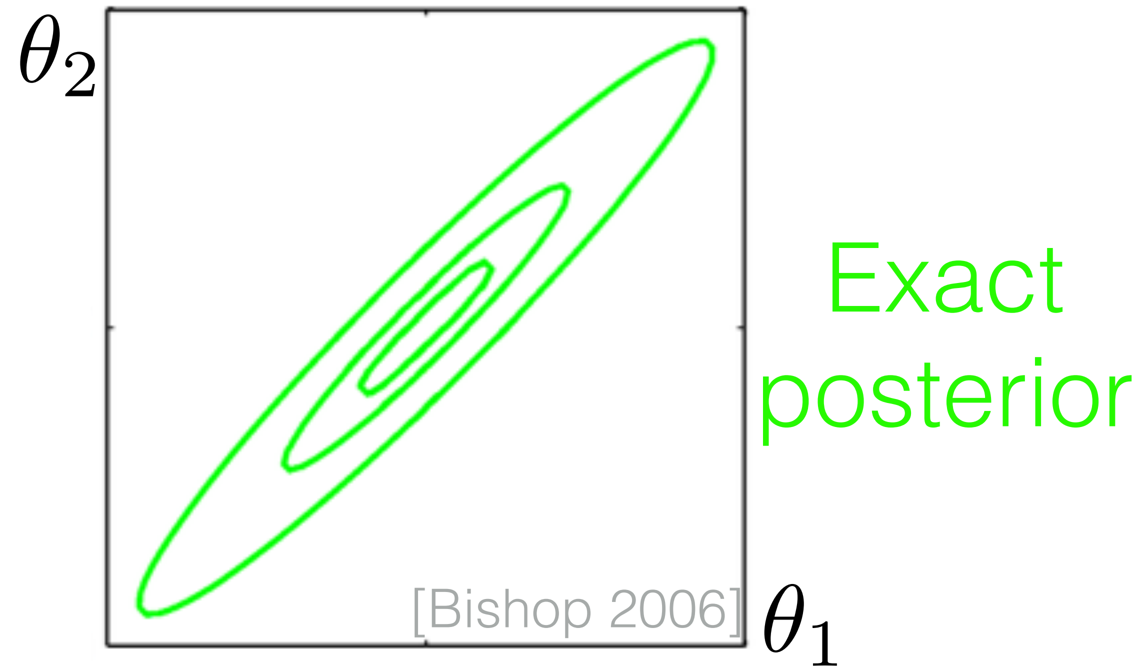


Exact posterior

Bayesian inference

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

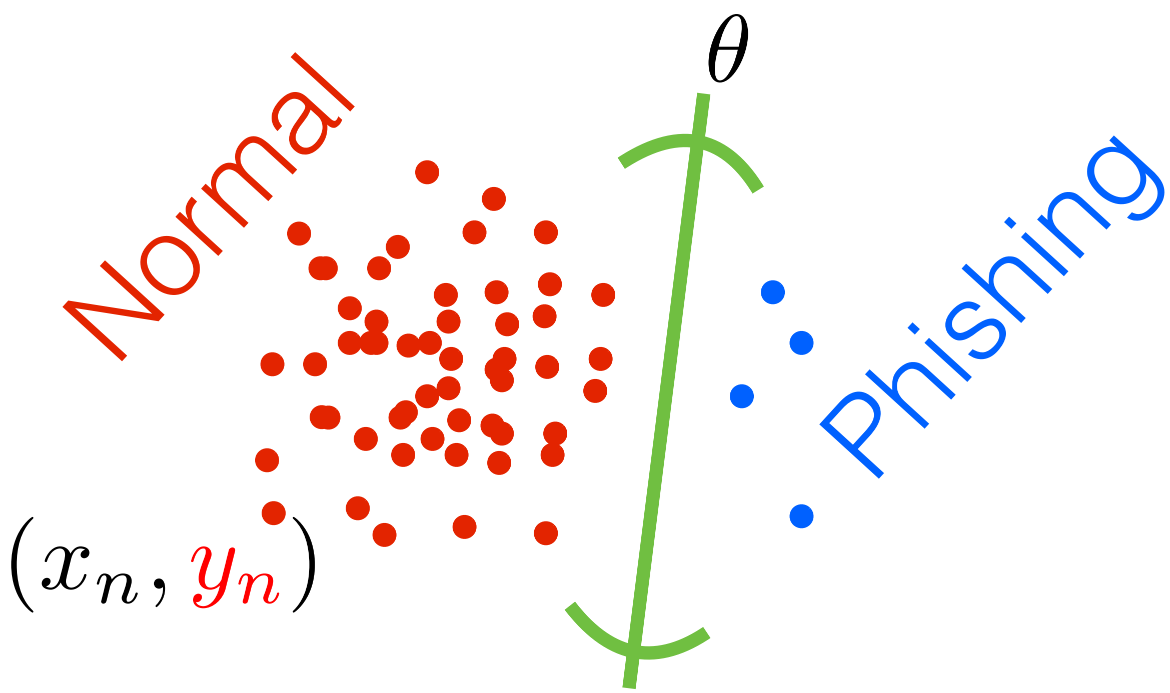


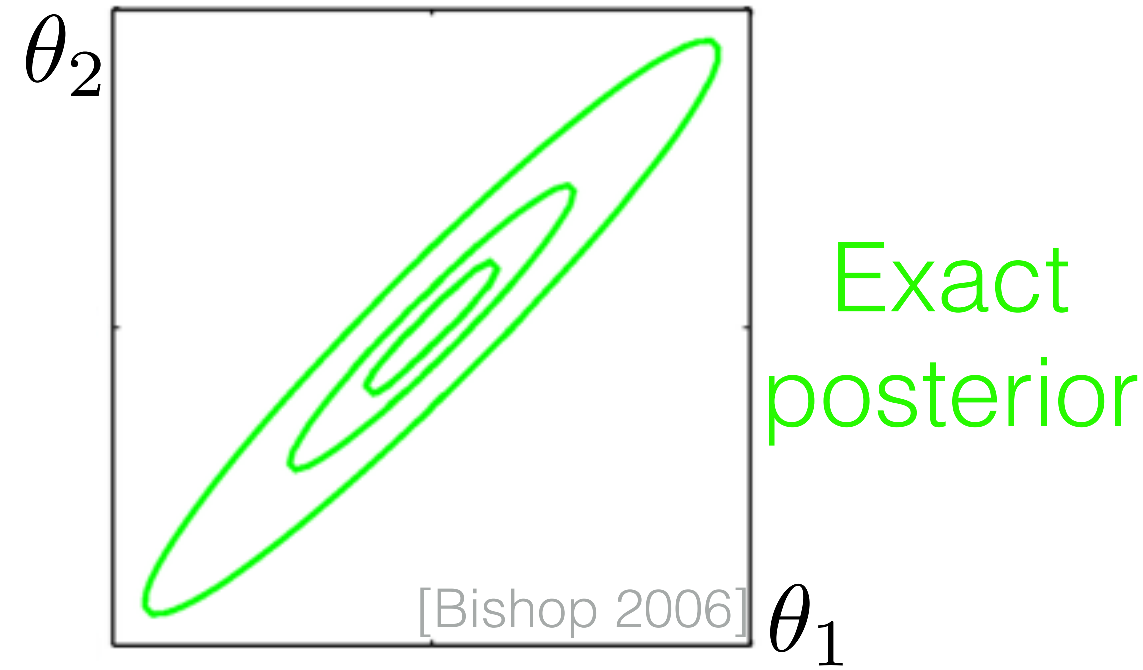


- MCMC: Eventually accurate but can be slow [Bardenet, Doucet, Holmes 2017]

Bayesian inference

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

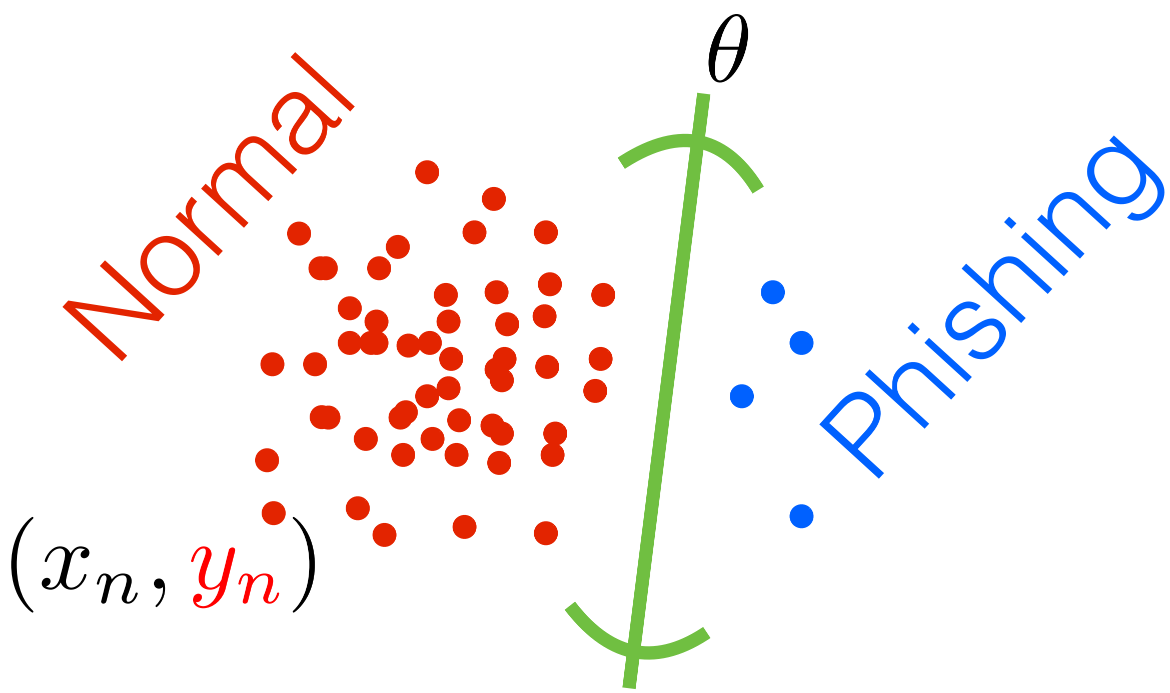


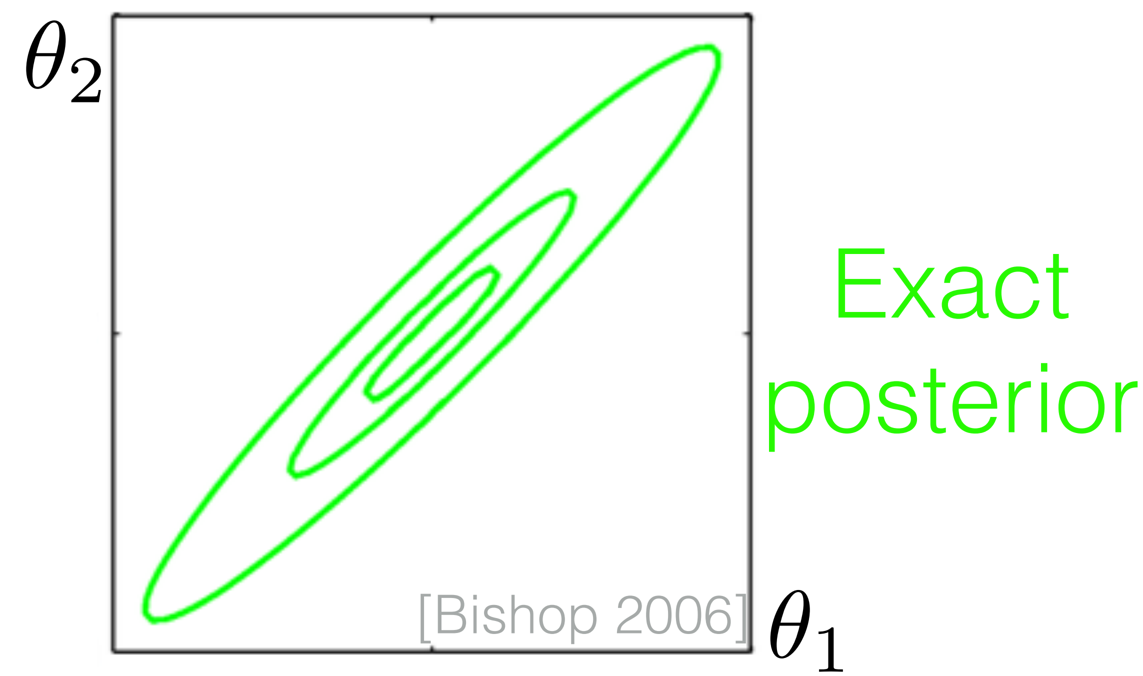


- MCMC: Eventually accurate but can be slow [Bardenet, Doucet, Holmes 2017]
- (Mean-field) variational Bayes: (MF)VB

Bayesian inference

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

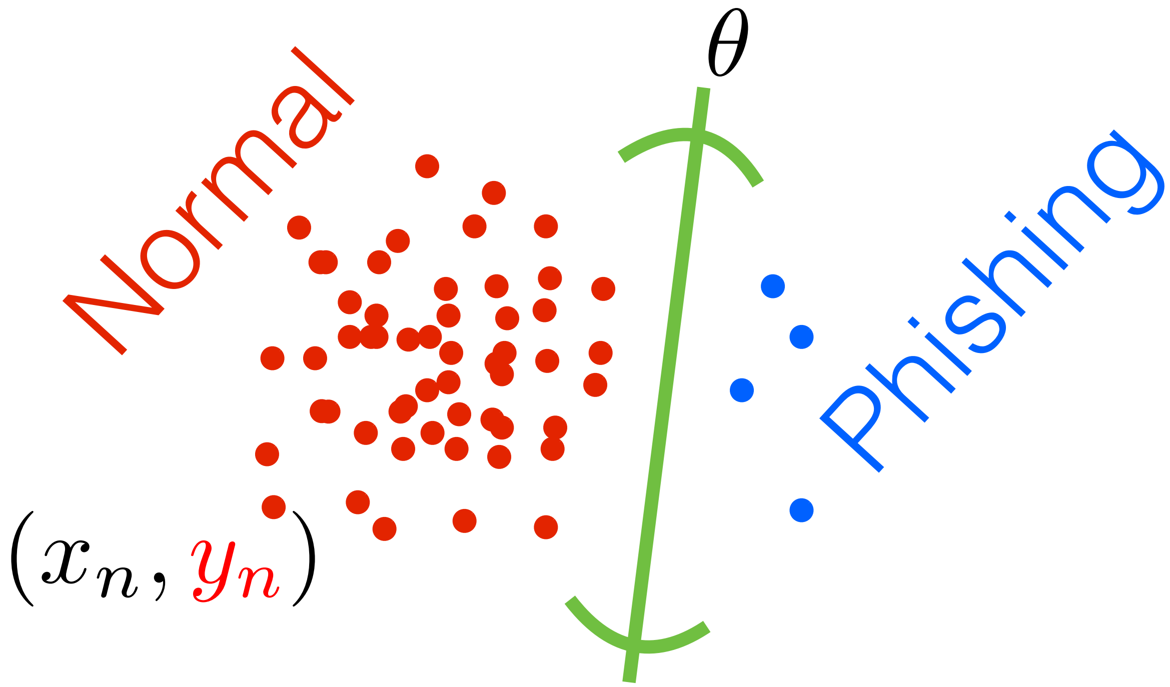


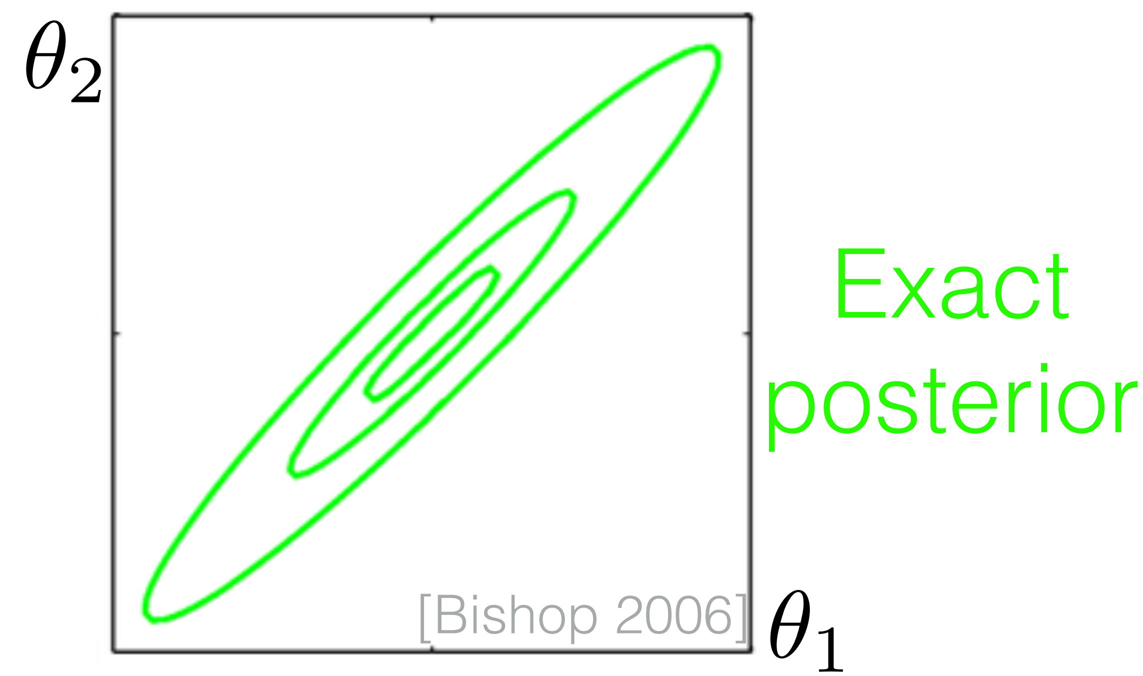


- MCMC: Eventually accurate but can be slow [Bardenet, Doucet, Holmes 2017]
- (Mean-field) variational Bayes: (MF)VB
  - Fast

Bayesian inference

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

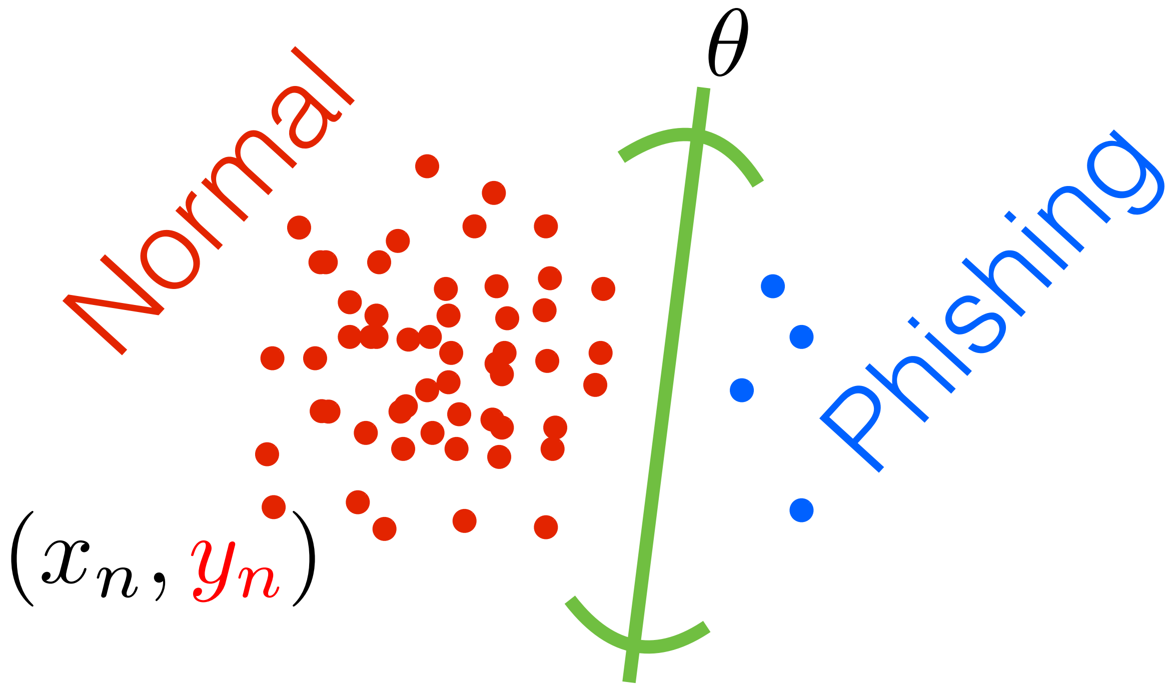


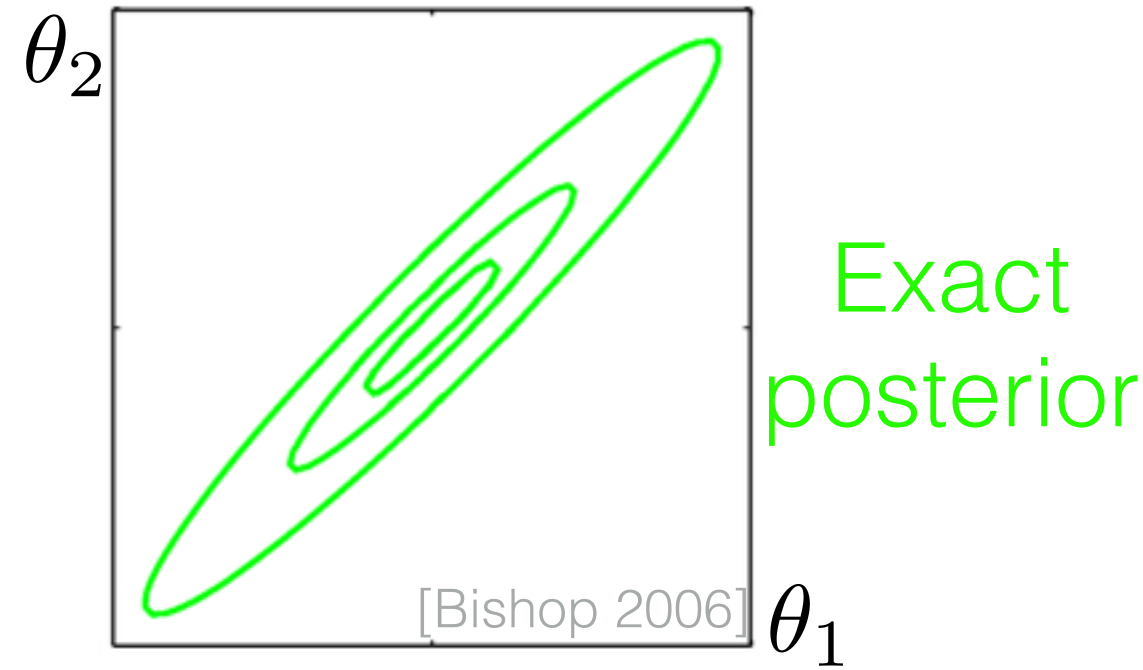


- MCMC: Eventually accurate but can be slow [Bardenet, Doucet, Holmes 2017]
- (Mean-field) variational Bayes: (MF)VB
  - Fast, streaming, distributed [Broderick, Boyd, Wibisono, Wilson, Jordan 2013]

Bayesian inference

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

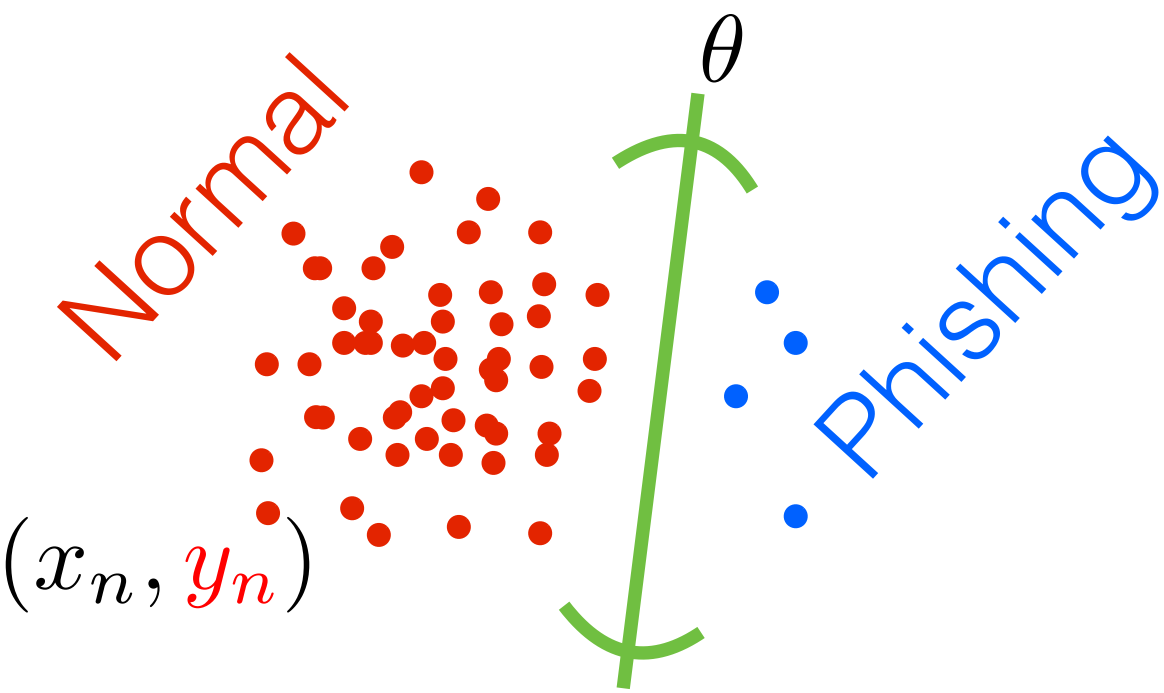


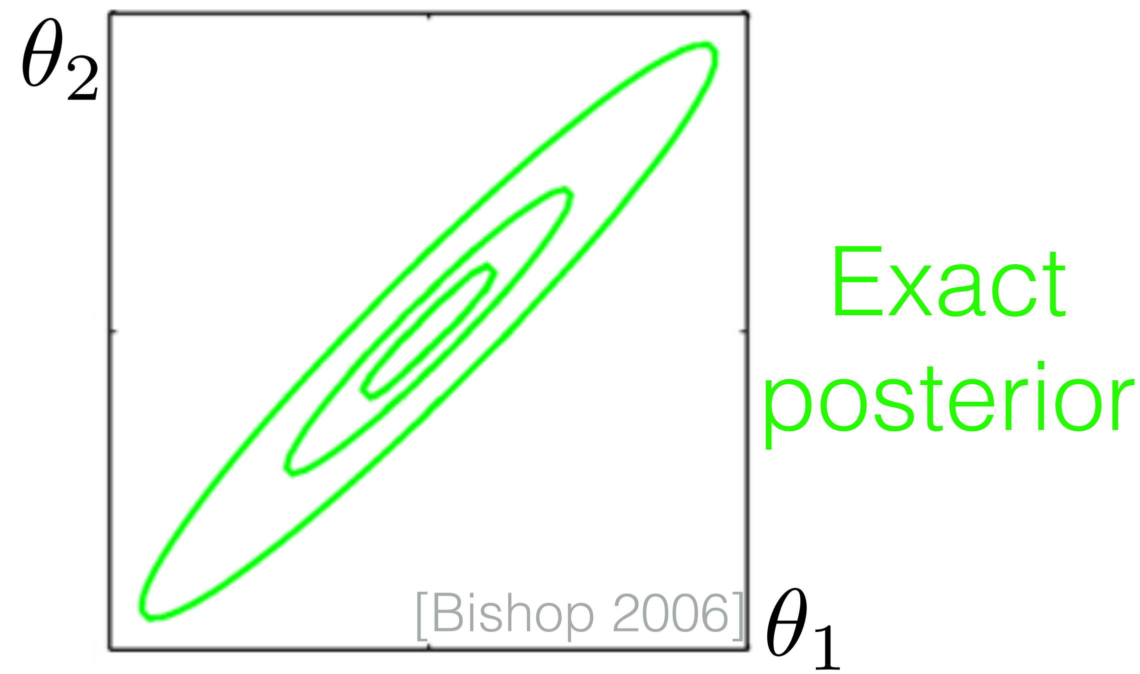


- MCMC: Eventually accurate but can be slow [Bardenet, Doucet, Holmes 2017]
- (Mean-field) variational Bayes: (MF)VB
  - Fast, streaming, distributed [Broderick, Boyd, Wibisono, Wilson, Jordan 2013] (3.6M Wikipedia, 32 cores, ~hour)

Bayesian inference

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



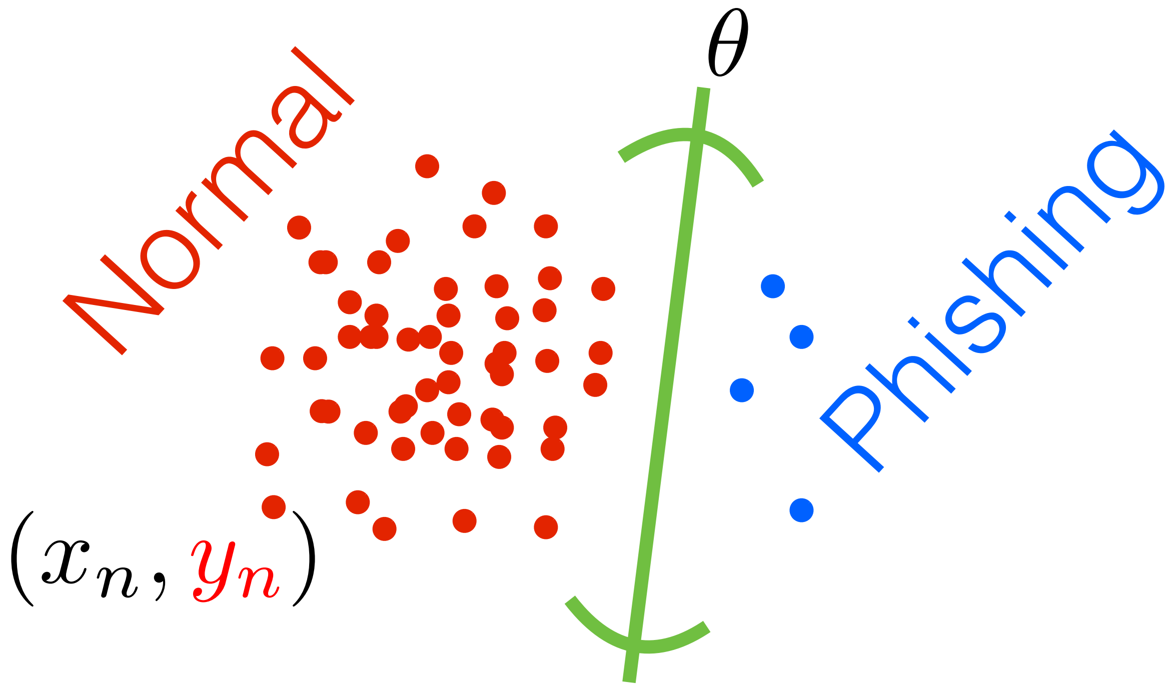


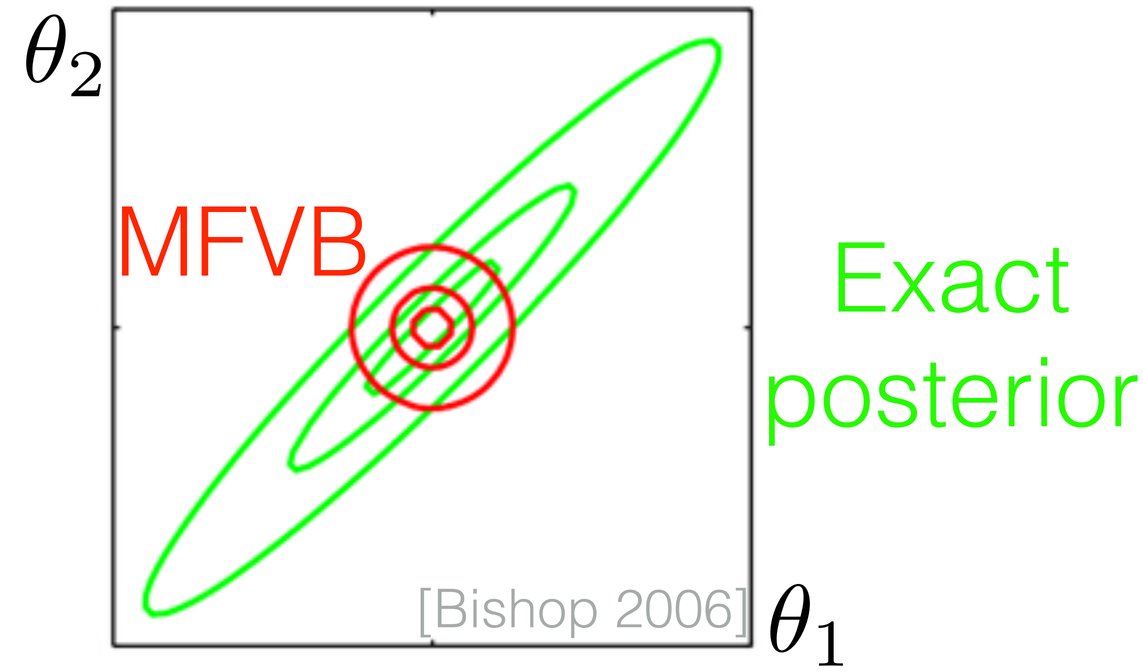
- MCMC: Eventually accurate but can be slow [Bardenet, Doucet, Holmes 2017]
- (Mean-field) variational Bayes: (MF)VB
  - Fast, streaming, distributed [Broderick, Boyd, Wibisono, Wilson, Jordan 2013] (3.6M Wikipedia, 32 cores, ~hour)
  - Misestimation & lack of quality guarantees

[MacKay 2003; Bishop 2006; Wang, Titterington 2004; Turner, Sahani 2011; Fosdick 2013; Dunson 2014; Bardenet, Doucet, Holmes 2017; Opper, Winther 2003; Giordano, Broderick, Jordan 2015, 2017; Huggins, Campbell, Kasprzak, Broderick 2018]

Bayesian inference

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



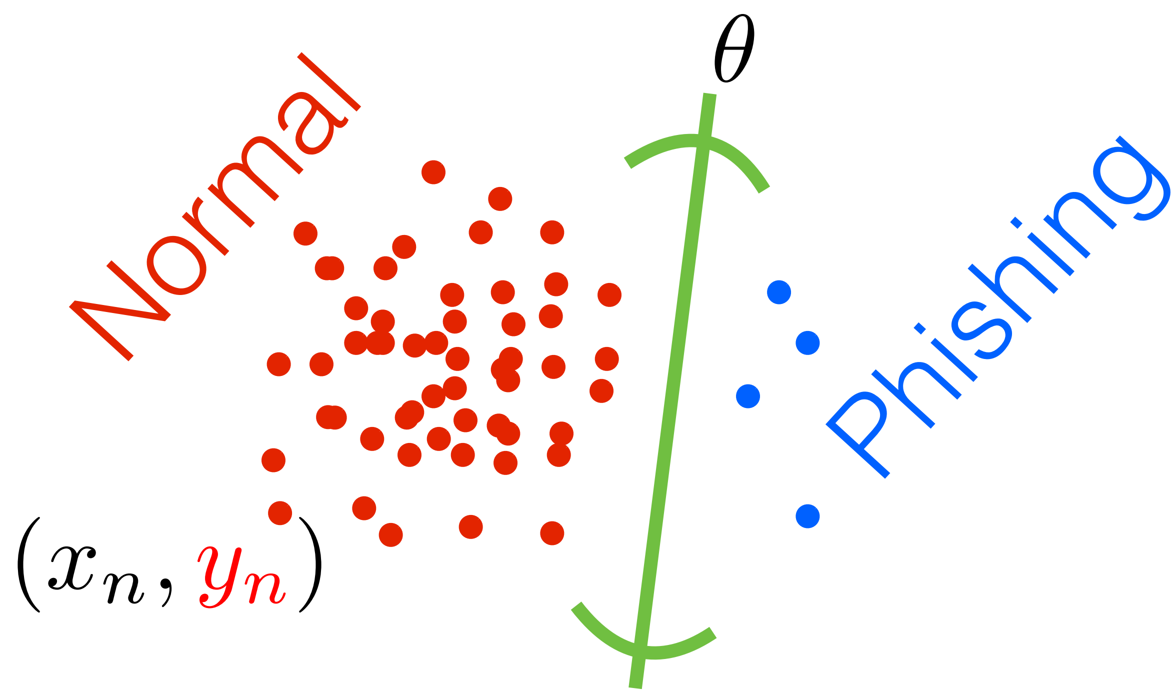


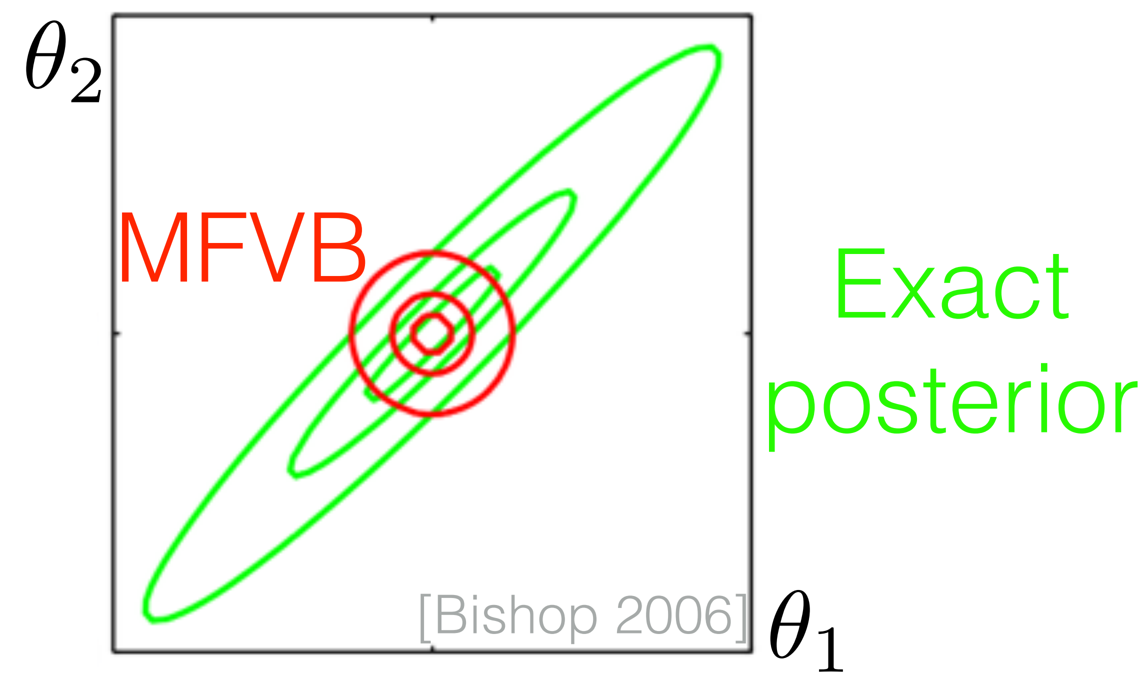
- MCMC: Eventually accurate but can be slow [Bardenet, Doucet, Holmes 2017]
- (Mean-field) variational Bayes: (MF)VB
  - Fast, streaming, distributed [Broderick, Boyd, Wibisono, Wilson, Jordan 2013] (3.6M Wikipedia, 32 cores, ~hour)
  - Misestimation & lack of quality guarantees

[MacKay 2003; Bishop 2006; Wang, Titterington 2004; Turner, Sahani 2011; Fosdick 2013; Dunson 2014; Bardenet, Doucet, Holmes 2017; Opper, Winther 2003; Giordano, Broderick, Jordan 2015, 2017; Huggins, Campbell, Kasprzak, Broderick 2018]

Bayesian inference

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





- MCMC: Eventually accurate but can be slow [Bardenet, Doucet, Holmes 2017]
- (Mean-field) variational Bayes: (MF)VB
  - Fast, streaming, distributed [Broderick, Boyd, Wibisono, Wilson, Jordan 2013] (3.6M Wikipedia, 32 cores, ~hour)
  - Misestimation & lack of quality guarantees

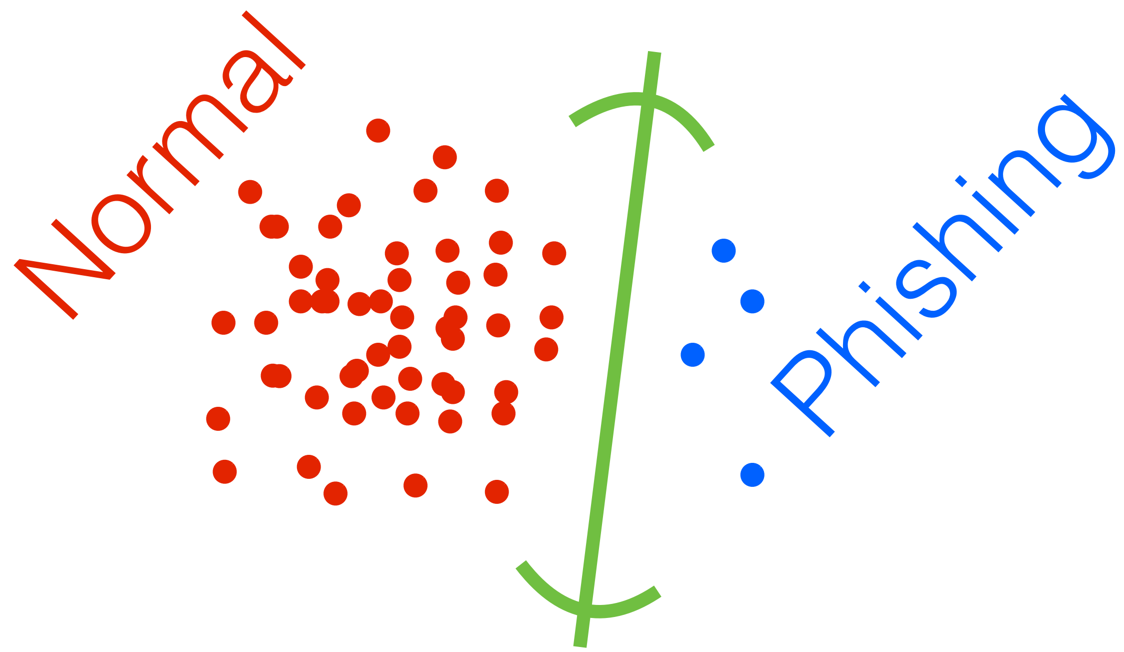
[MacKay 2003; Bishop 2006; Wang, Titterington 2004; Turner, Sahani 2011; Fosdick 2013; Dunson 2014; Bardenet, Doucet, Holmes 2017; Opper, Winther 2003; Giordano, Broderick, Jordan 2015, 2017; Huggins, Campbell, Kasprzak, Broderick 2018]

- Automation: e.g. Stan, NUTS, ADVI

[http://mc-stan.org/ ; Hoffman, Gelman 2014; Kucukelbir, Tran, Ranganath, Gelman, Blei 2017]

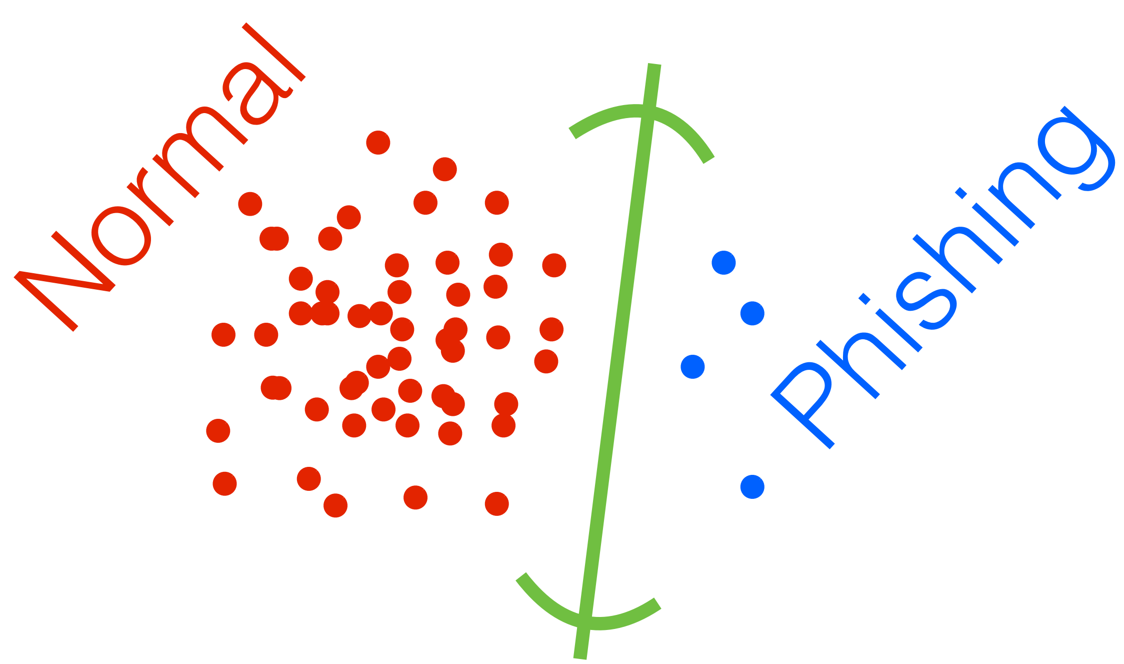
Bayesian coresets

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



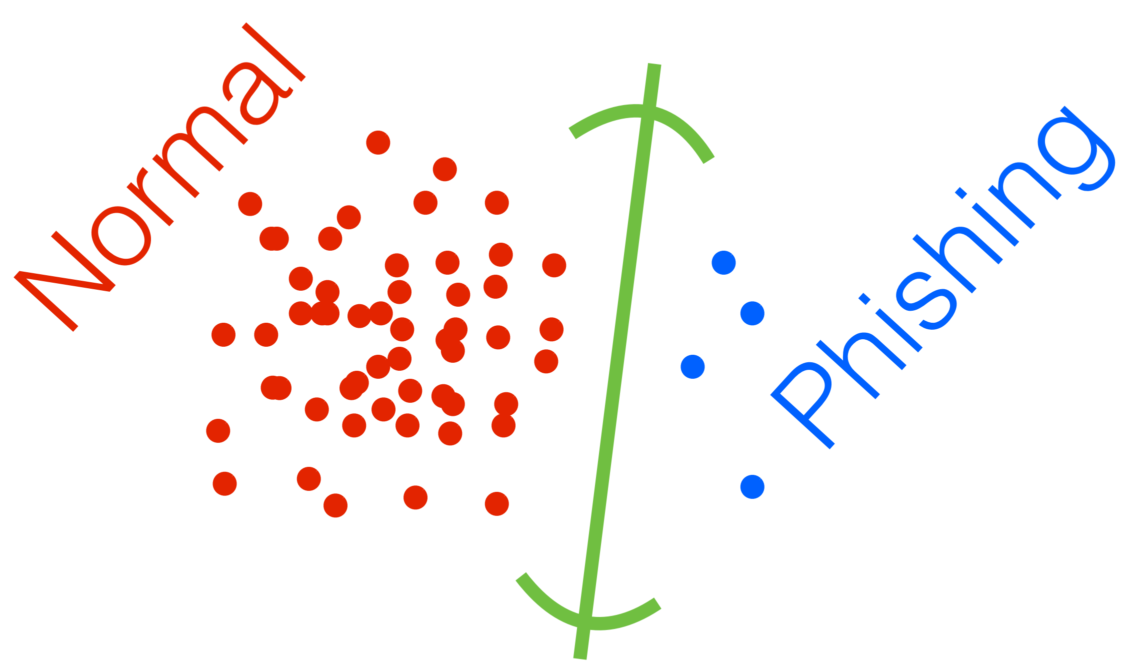
Bayesian coresets

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



Bayesian coresets

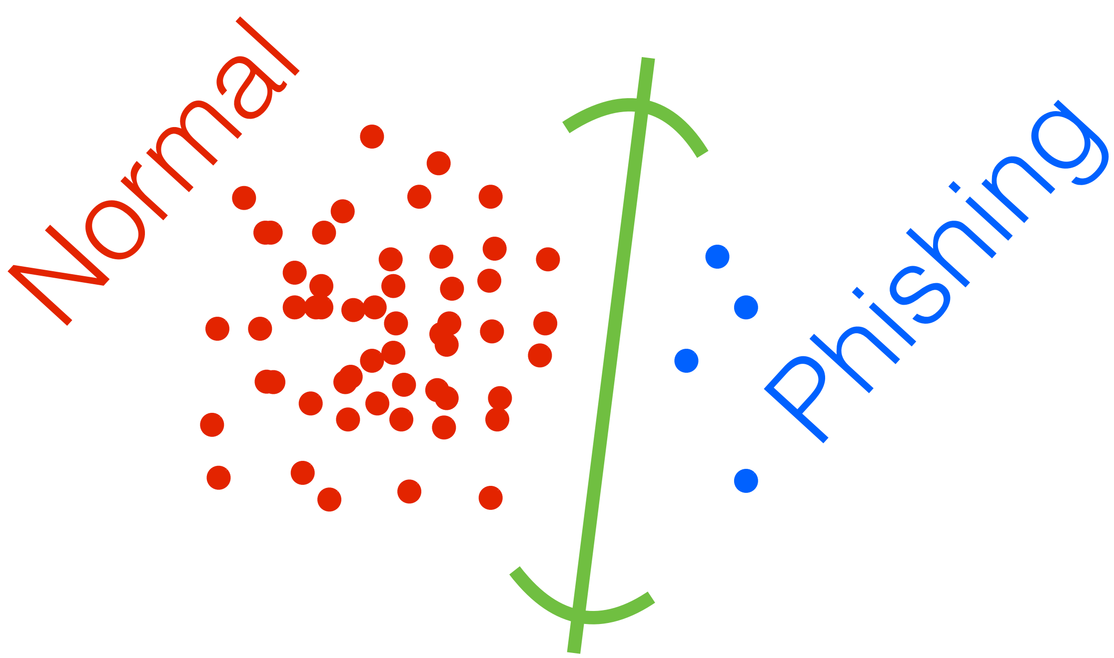
- Posterior $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$
- Log likelihood $\mathcal{L}_n(\theta) := \log p(y_n|\theta)\,, \;\; \mathcal{L}(\theta) := \sum_{n=1}^{N} \mathcal{L}_n(\theta)$
- Coreset log likelihood



Bayesian coresets

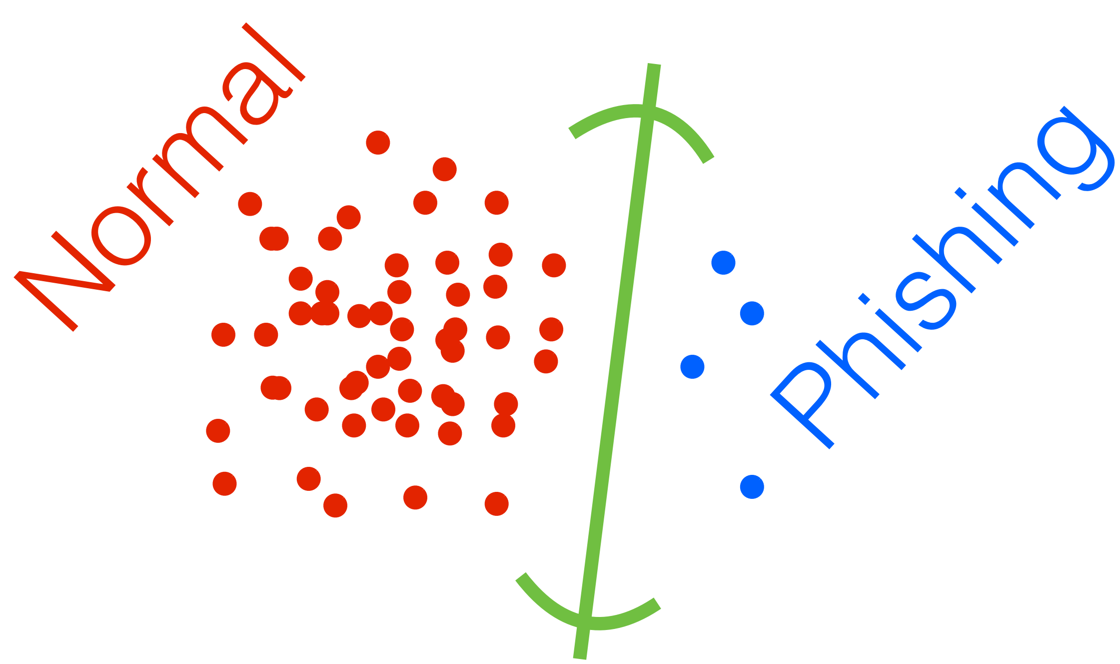
- Posterior $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$
- Log likelihood $\mathcal{L}_n(\theta) := \log p(y_n|\theta)\,, \;\; \mathcal{L}(\theta) := \sum_{n=1}^{N} \mathcal{L}_n(\theta)$
- Coreset log likelihood

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



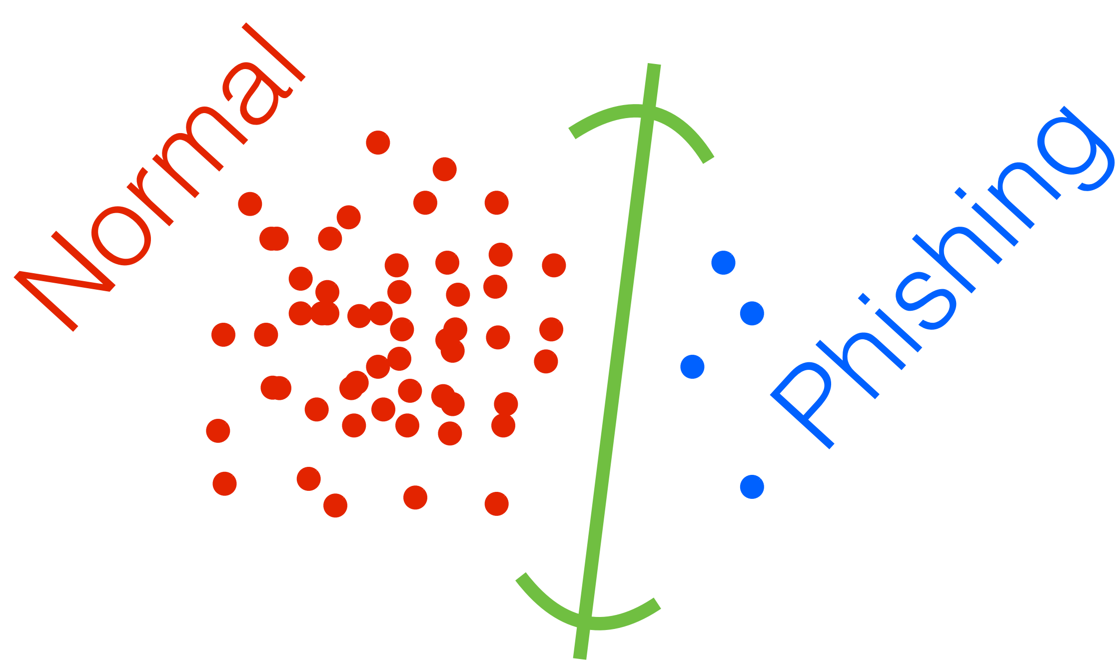
Bayesian coresets

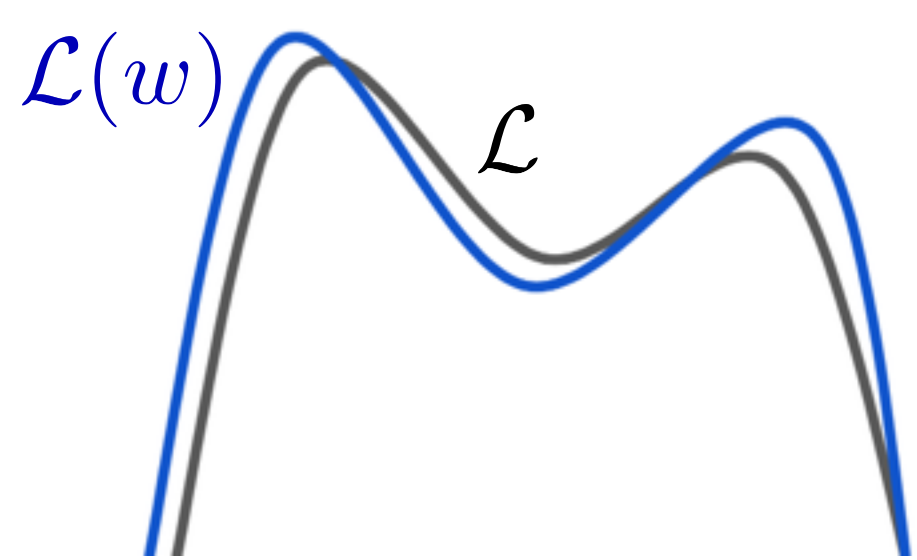
- Posterior $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$
- Log likelihood $\mathcal{L}_n(\theta) := \log p(y_n|\theta)\,,\ \ \mathcal{L}(\theta) := \sum_{n=1}^{N} \mathcal{L}_n(\theta)$
- Coreset log likelihood $\mathcal{L}(w,\theta) := \sum_{n=1}^{N} w_n \mathcal{L}_n(\theta)$ s.t. $\|w\|_0 \ll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



Bayesian coresets

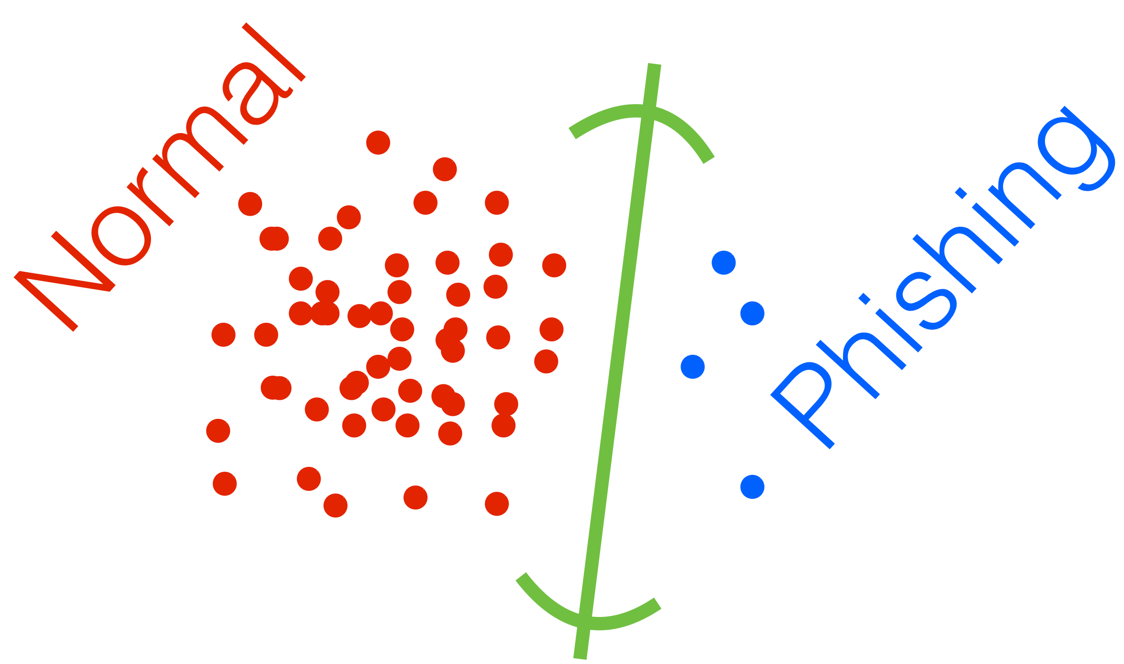
- Posterior $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$
- Log likelihood $\mathcal{L}_n(\theta) := \log p(y_n|\theta)\,,\;\; \mathcal{L}(\theta) := \sum_{n=1}^{N} \mathcal{L}_n(\theta)$
- Coreset log likelihood $\mathcal{L}(w,\theta) := \sum_{n=1}^{N} w_n \mathcal{L}_n(\theta)$ s.t. $\|w\|_0 \ll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

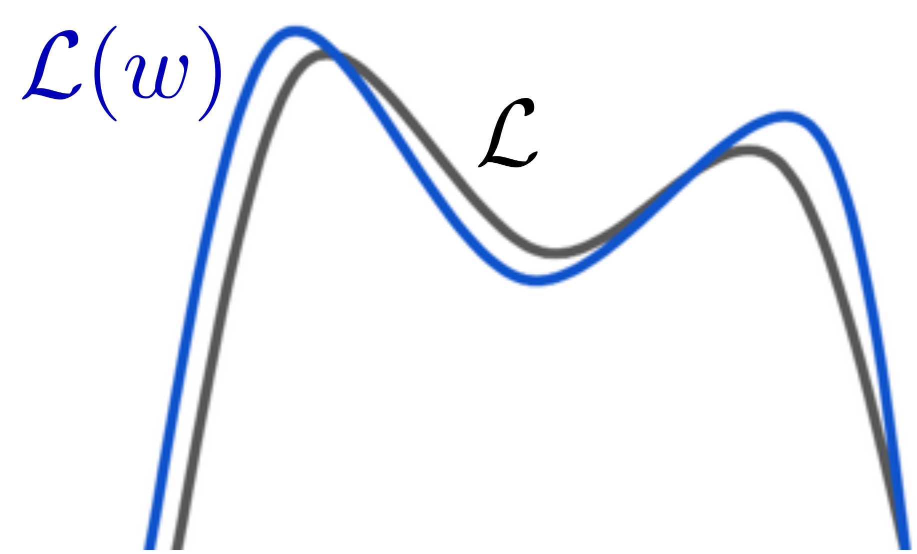




Bayesian coresets

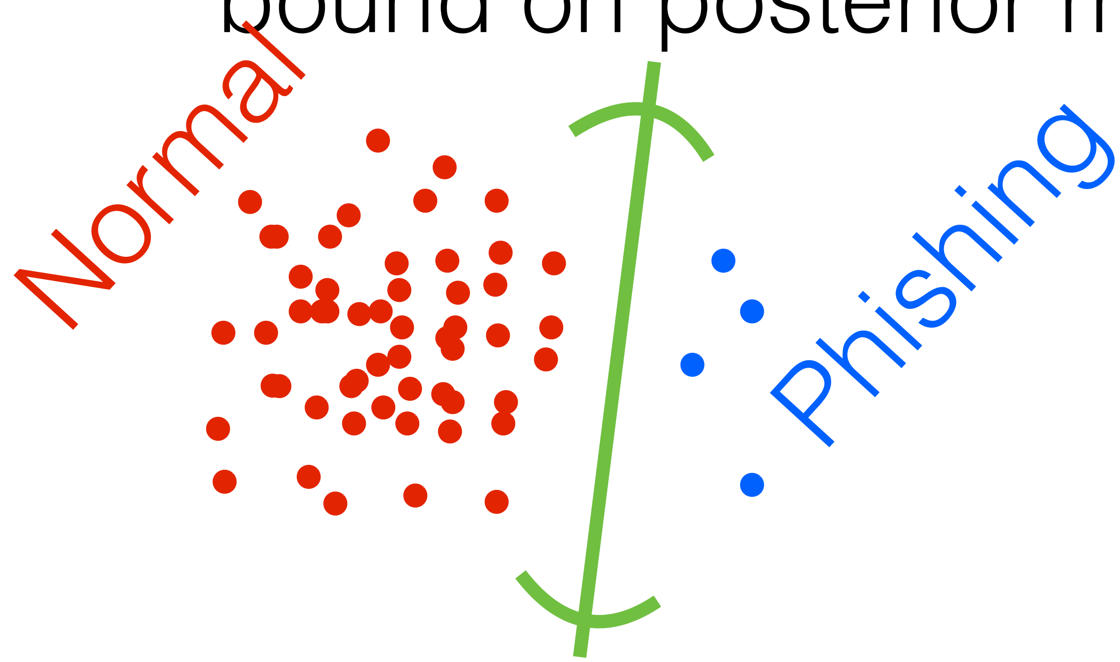
- Posterior $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$
- Log likelihood $\mathcal{L}_n(\theta) := \log p(y_n|\theta)\,, \;\; \mathcal{L}(\theta) := \sum_{n=1}^{N} \mathcal{L}_n(\theta)$
- Coreset log likelihood $\mathcal{L}(w,\theta) := \sum_{n=1}^{N} w_n \mathcal{L}_n(\theta)$ s.t. $\|w\|_0 \ll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

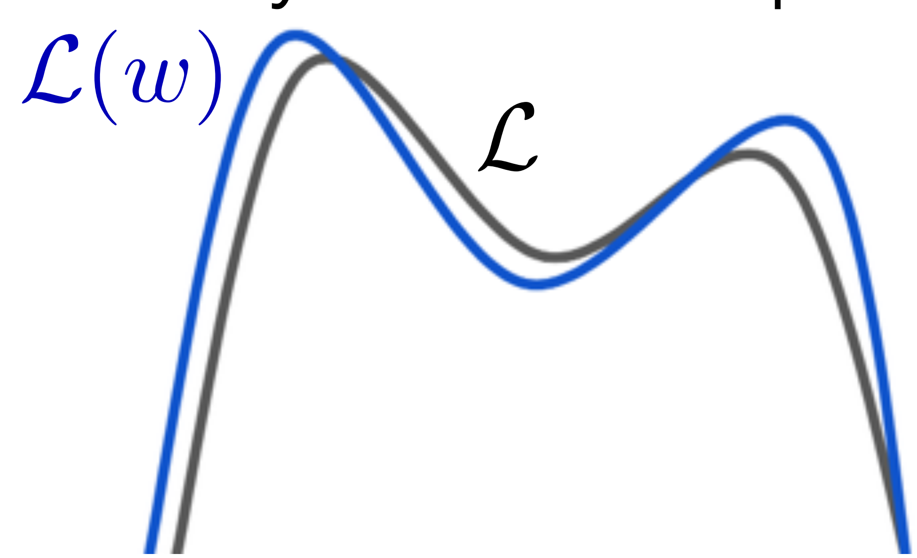




Bayesian coresets

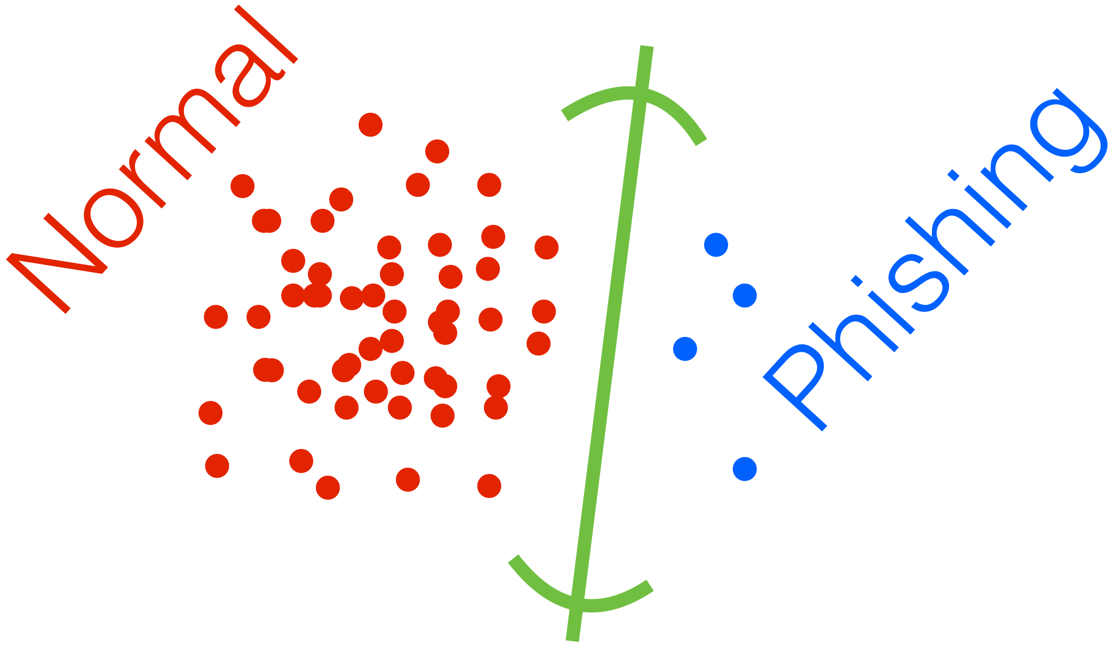
- Posterior $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$
- Log likelihood $\mathcal{L}_n(\theta) := \log p(y_n|\theta)\,, \;\; \mathcal{L}(\theta) := \sum_{n=1}^{N} \mathcal{L}_n(\theta)$
- Coreset log likelihood $\mathcal{L}(w,\theta) := \sum_{n=1}^{N} w_n \mathcal{L}_n(\theta)$ s.t. $\|w\|_0 \ll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



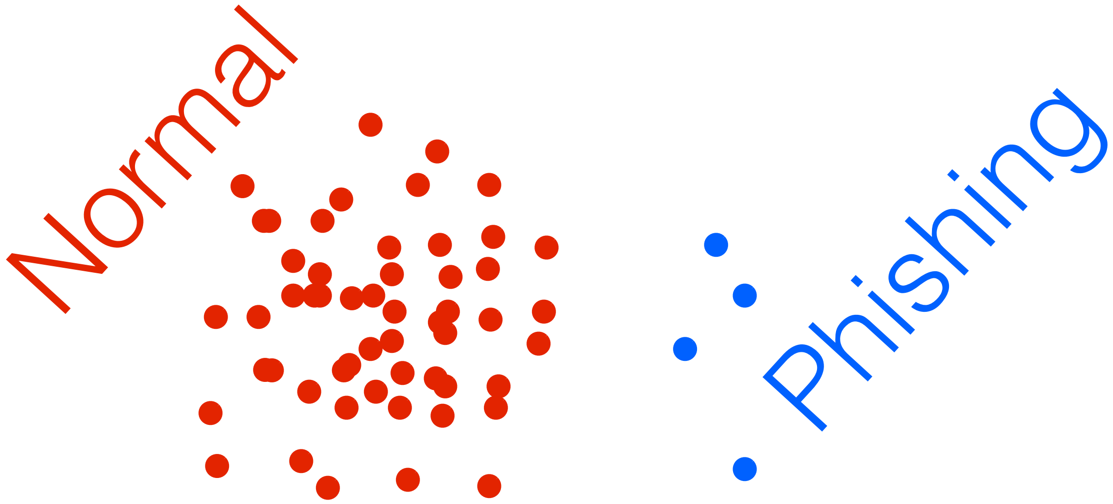


[Huggins, Campbell, Kasprzak, Broderick 2018; Broderick 2018]

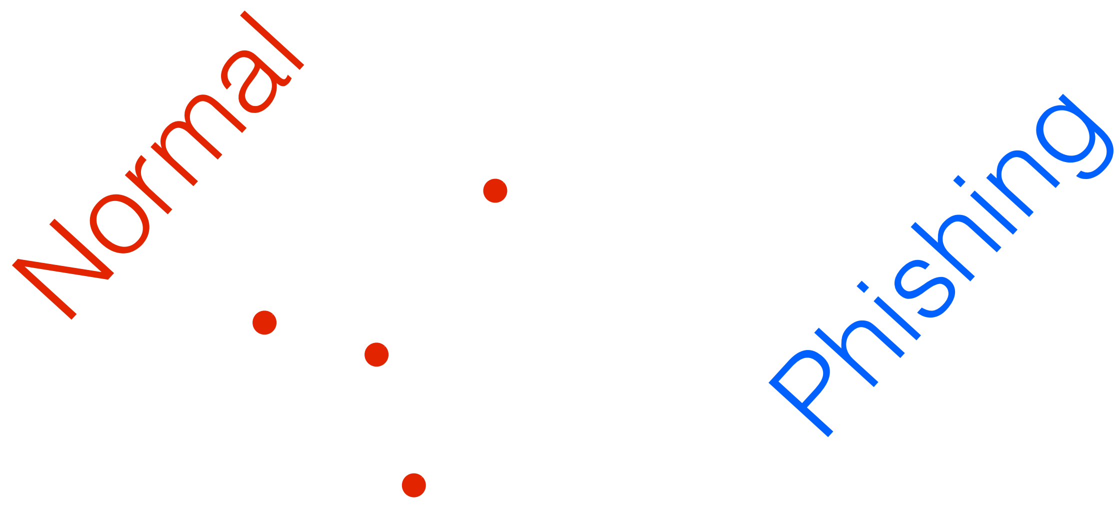
Uniform subsampling



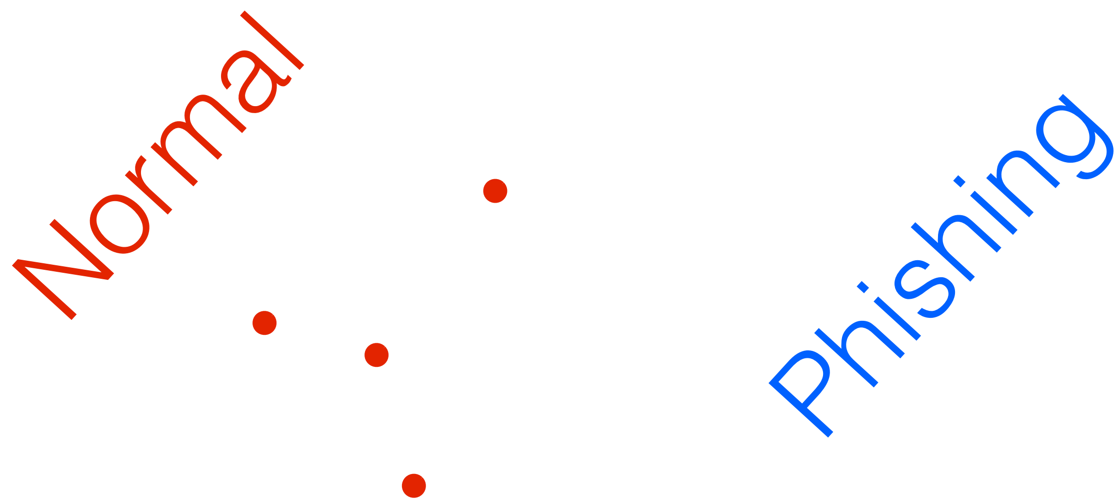
Uniform subsampling



Uniform subsampling

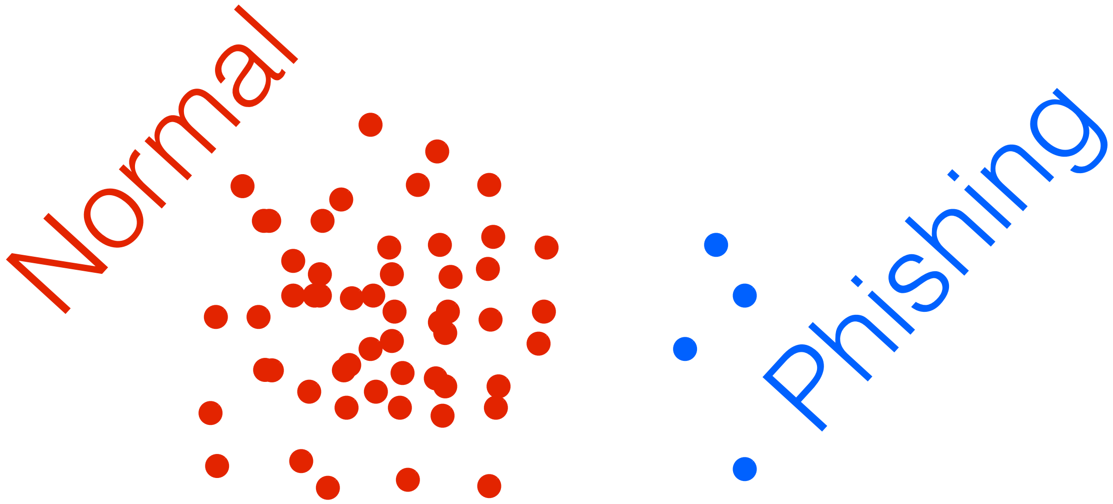


Uniform subsampling



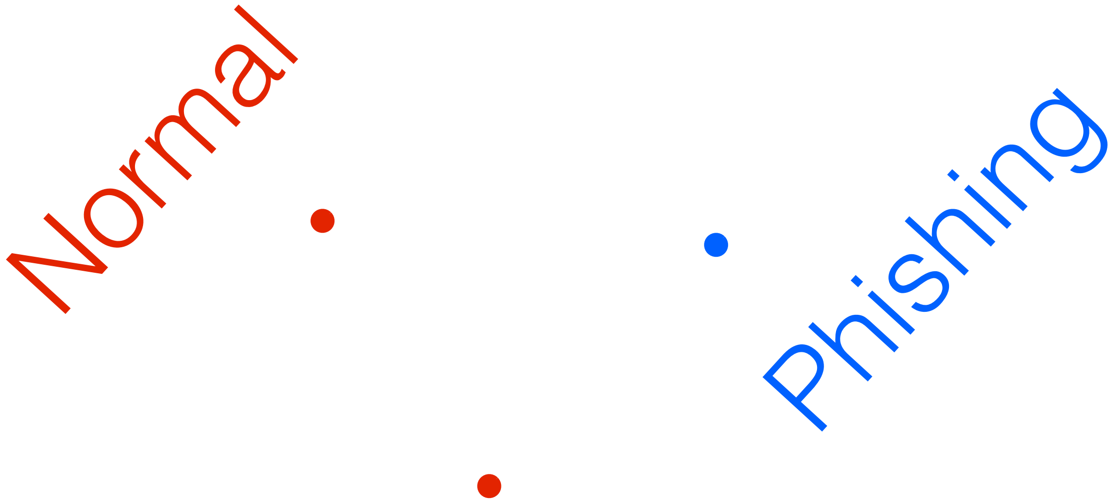
- Might miss important data

Uniform subsampling



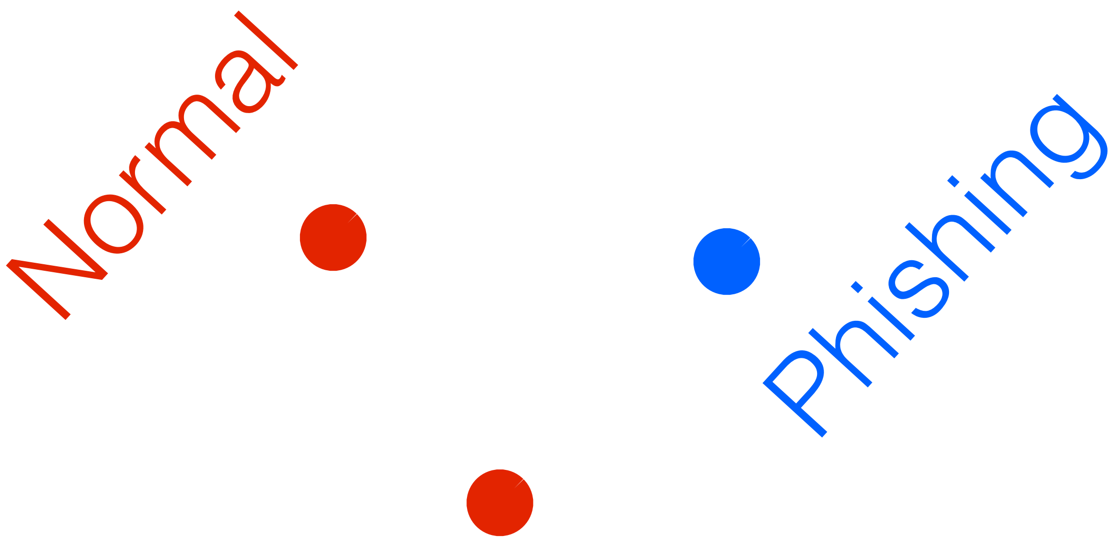
- Might miss important data

Uniform subsampling



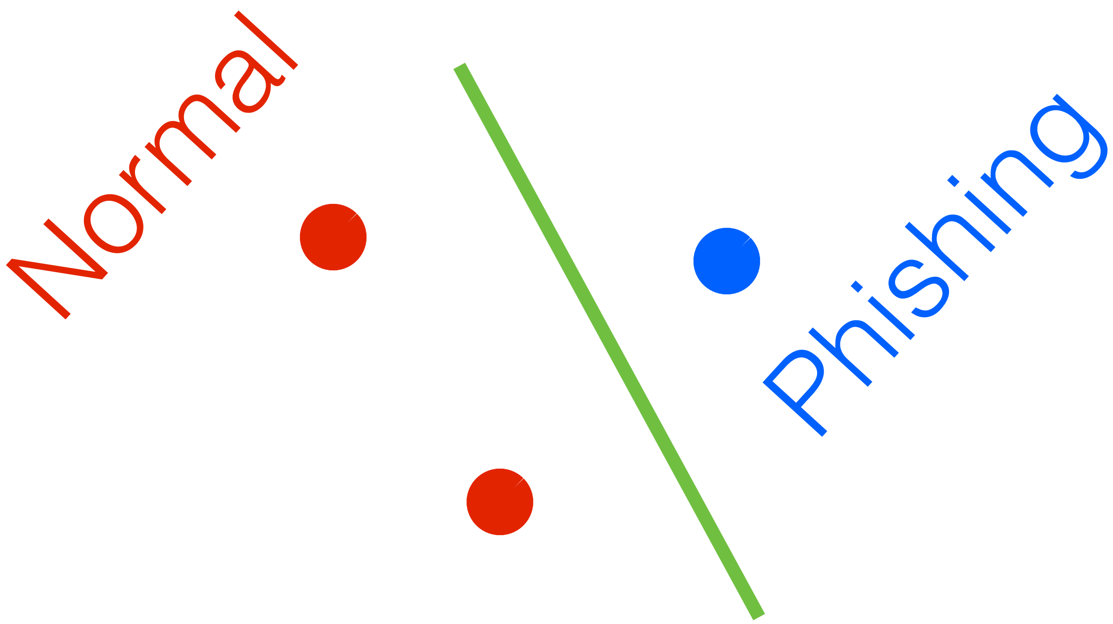
- Might miss important data

Uniform subsampling



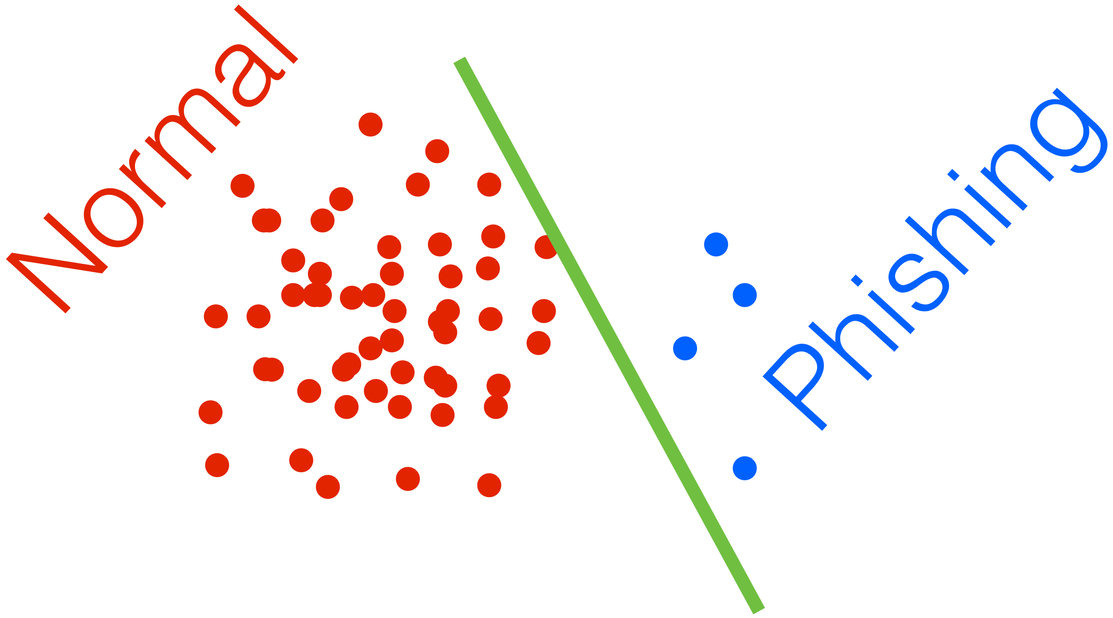
- Might miss important data

Uniform subsampling



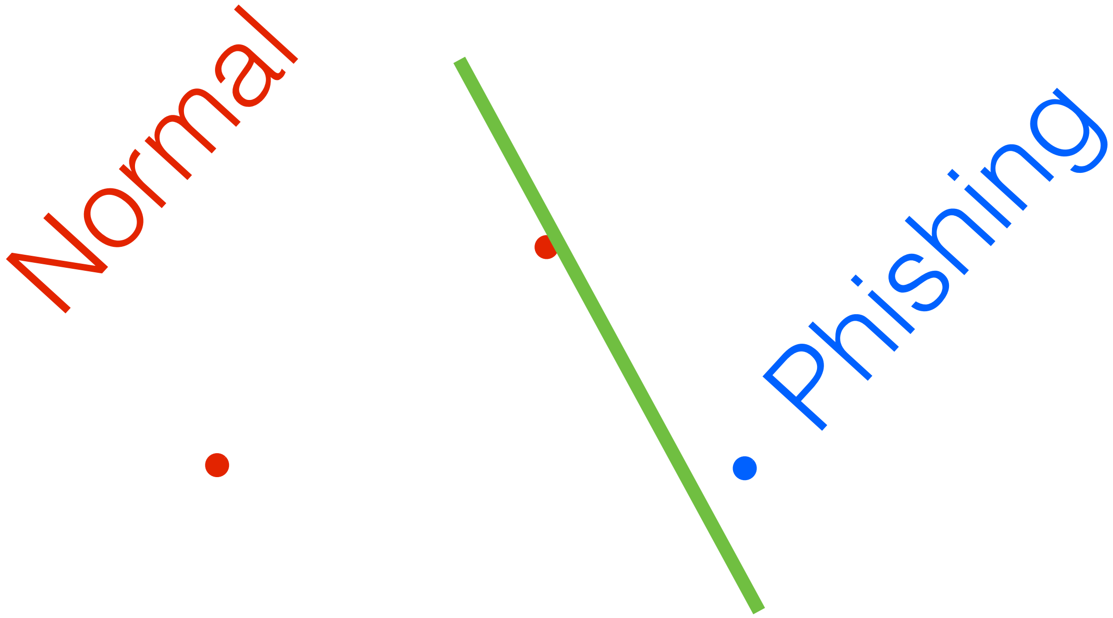
- Might miss important data

Uniform subsampling



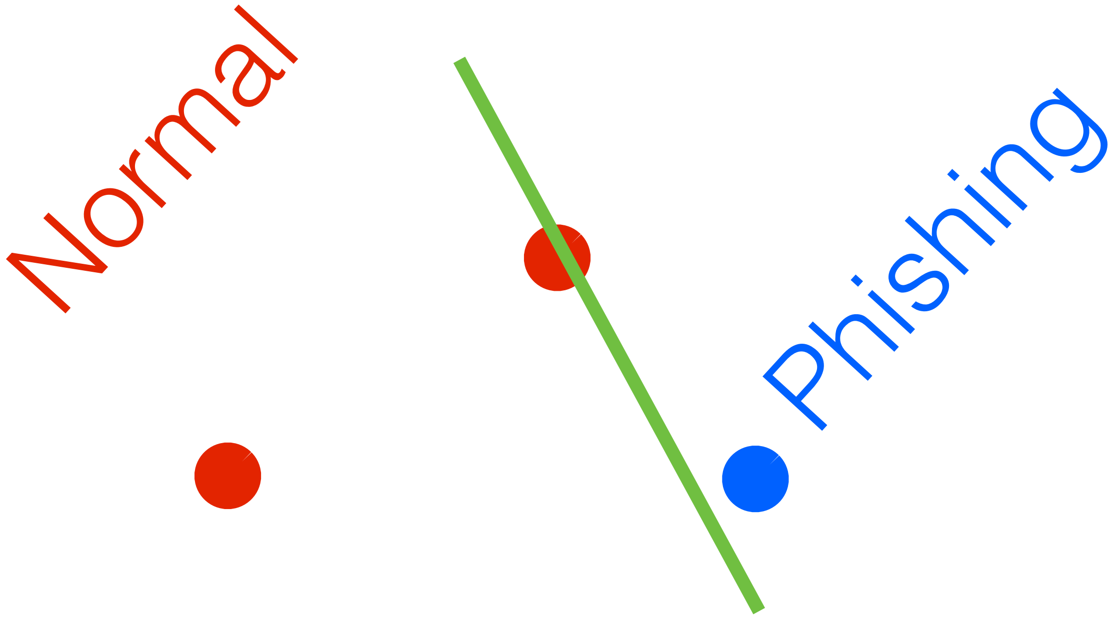
- Might miss important data

Uniform subsampling



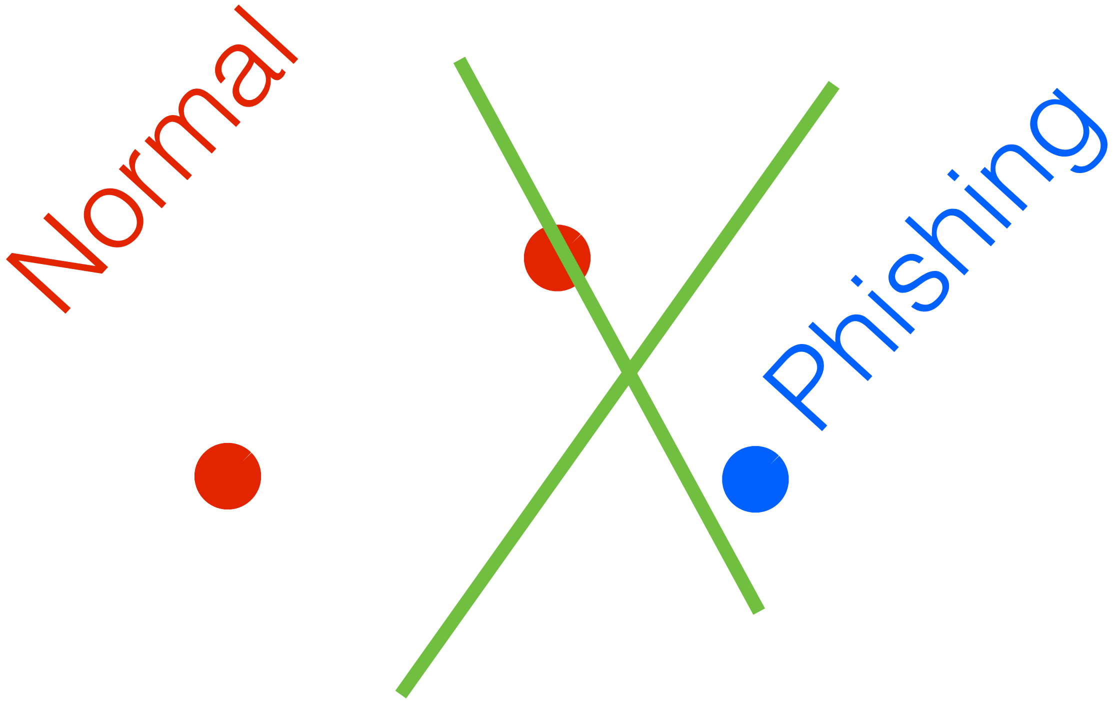
- Might miss important data

Uniform subsampling



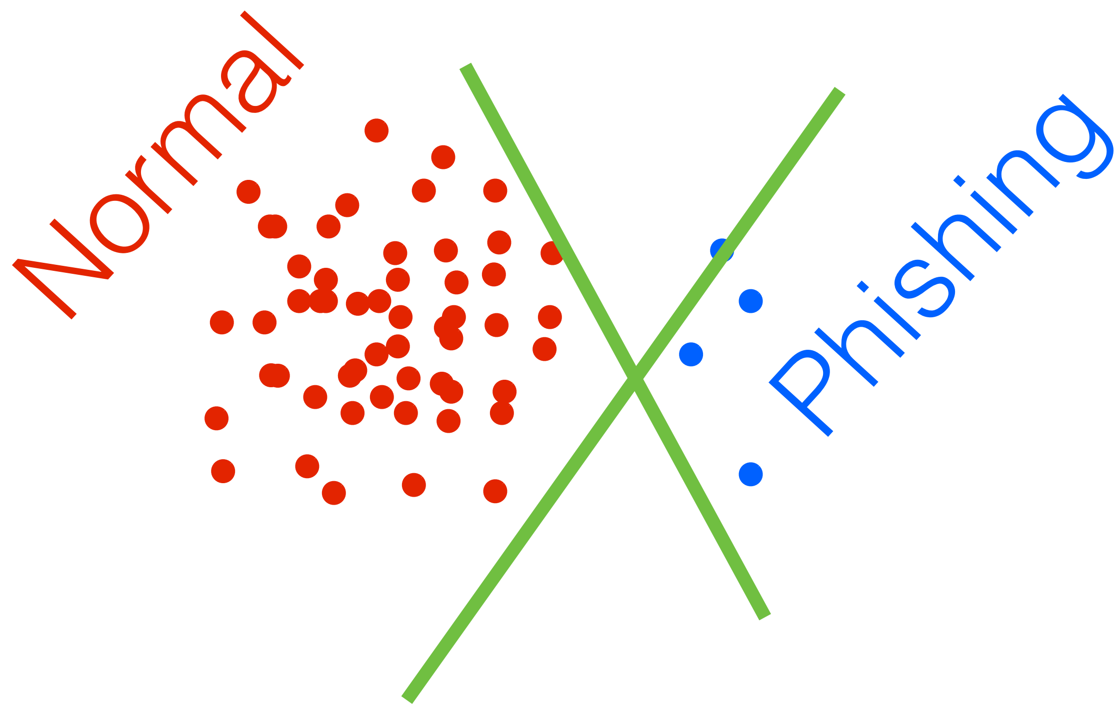
- Might miss important data

Uniform subsampling



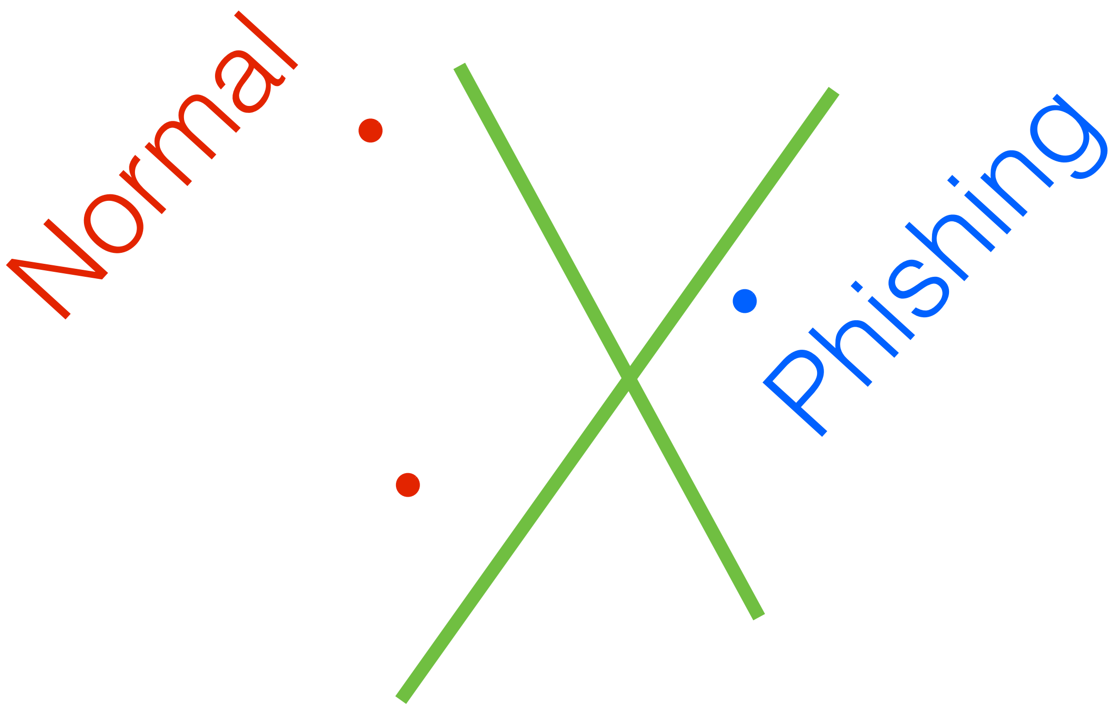
- Might miss important data

Uniform subsampling



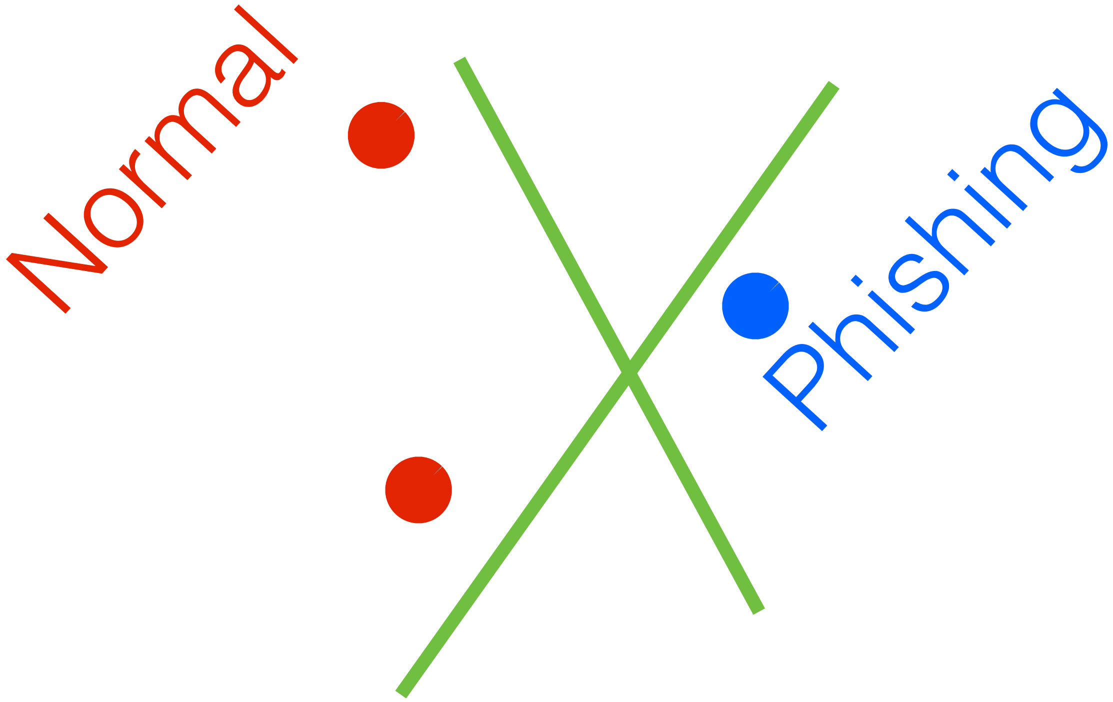
- Might miss important data

Uniform subsampling



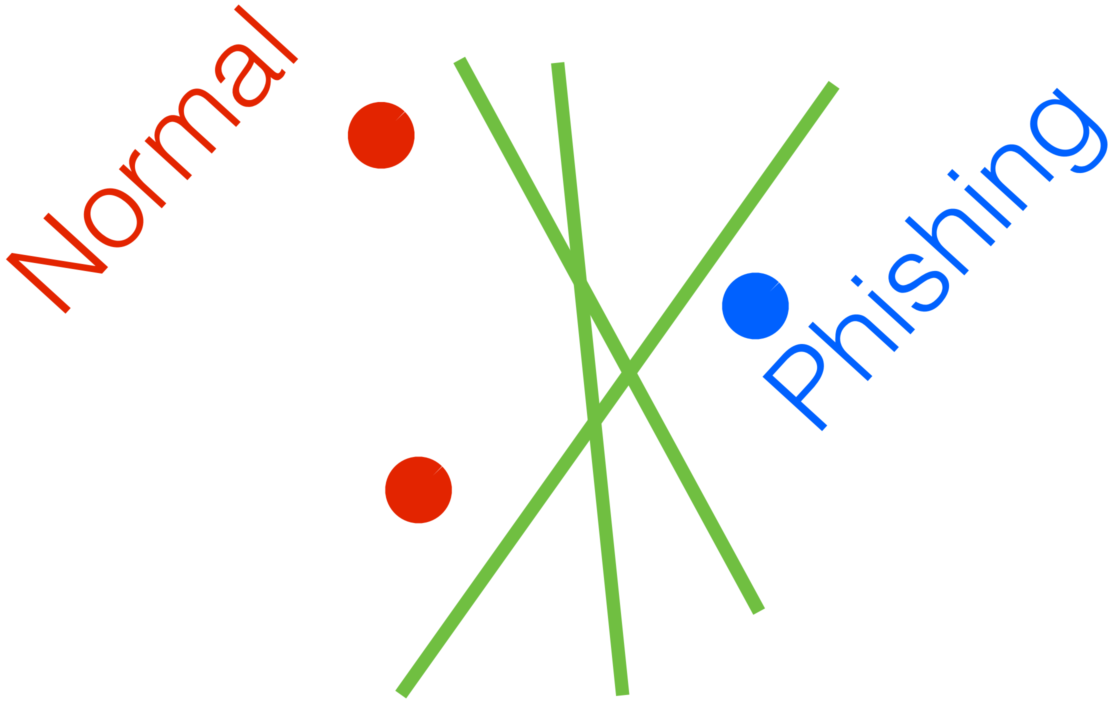
- Might miss important data

Uniform subsampling



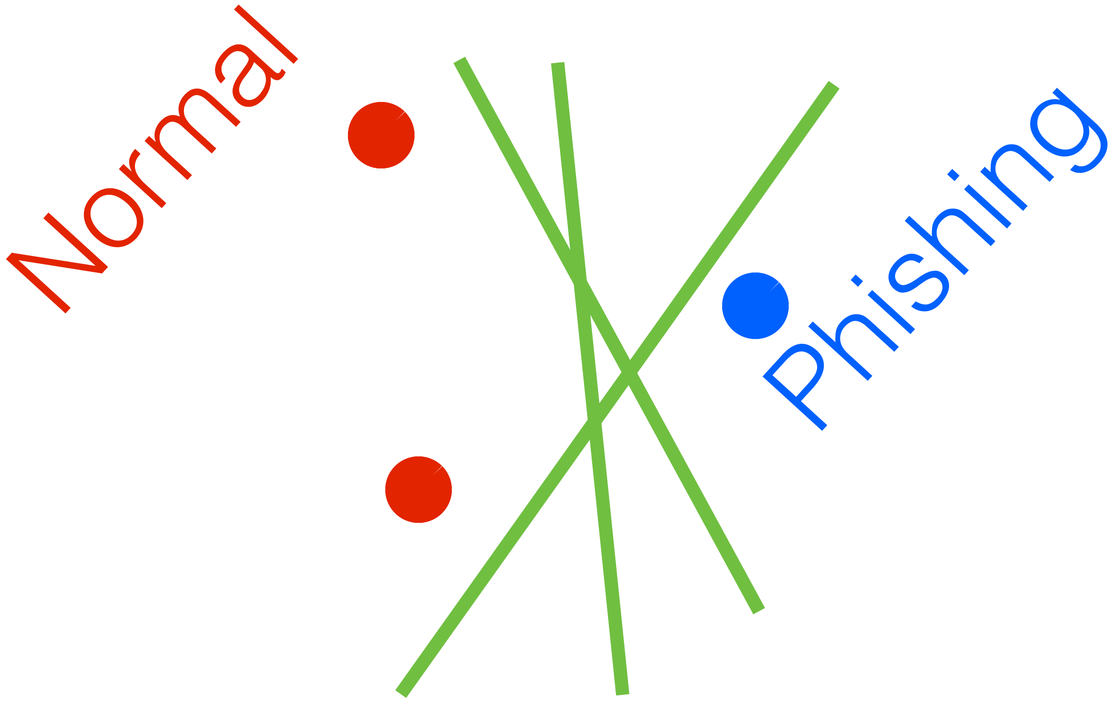
- Might miss important data

Uniform subsampling



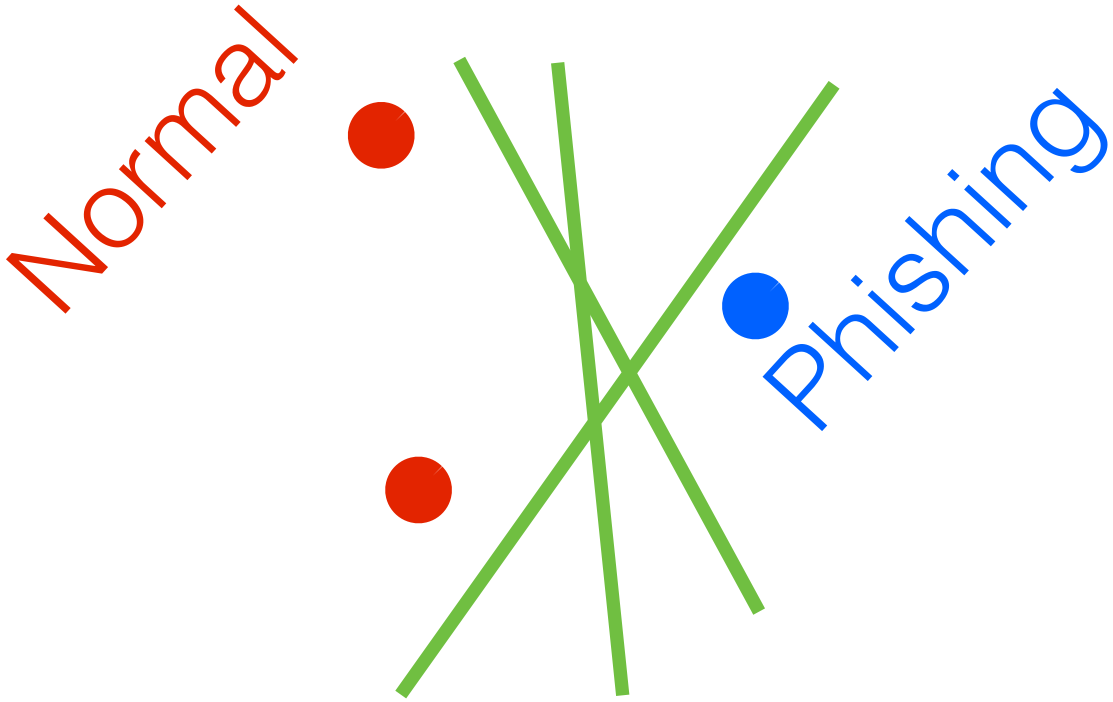
- Might miss important data

Uniform subsampling

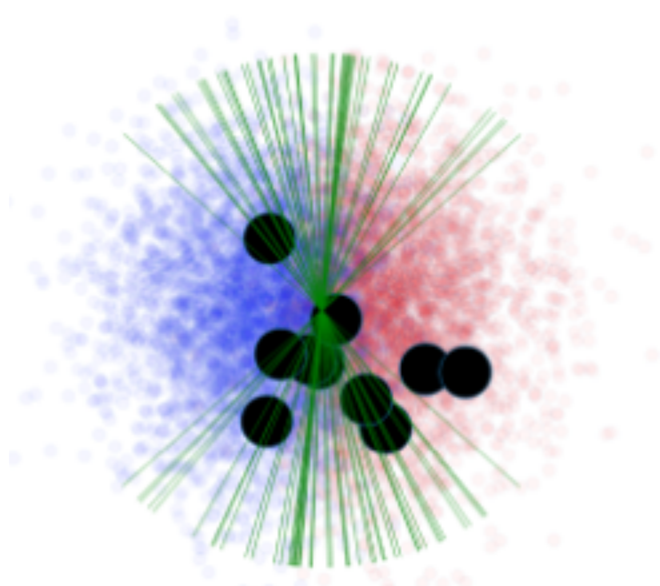


- Might miss important data
- Noisy estimates

Uniform subsampling

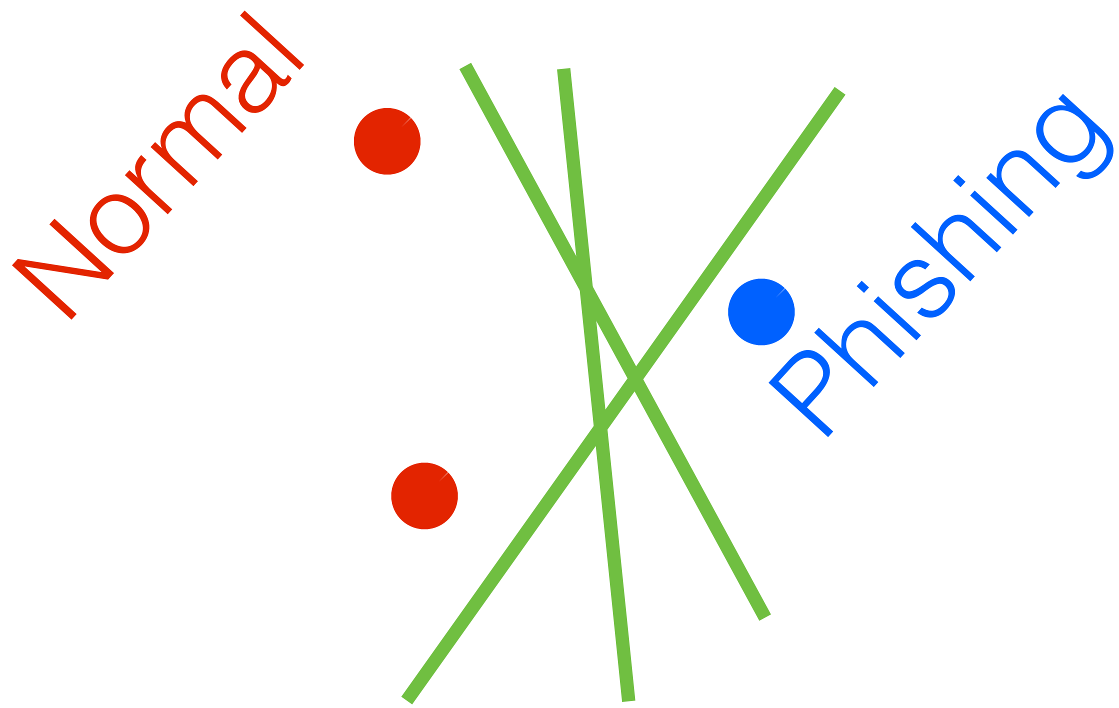


- Might miss important data
- Noisy estimates

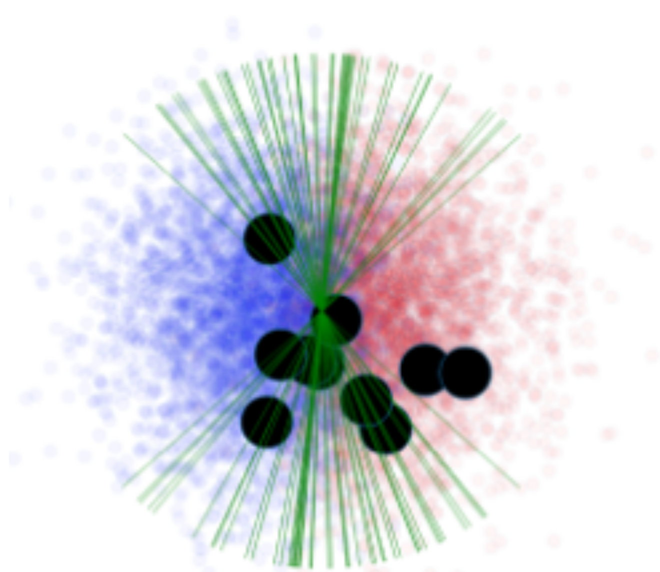


$M = 10$

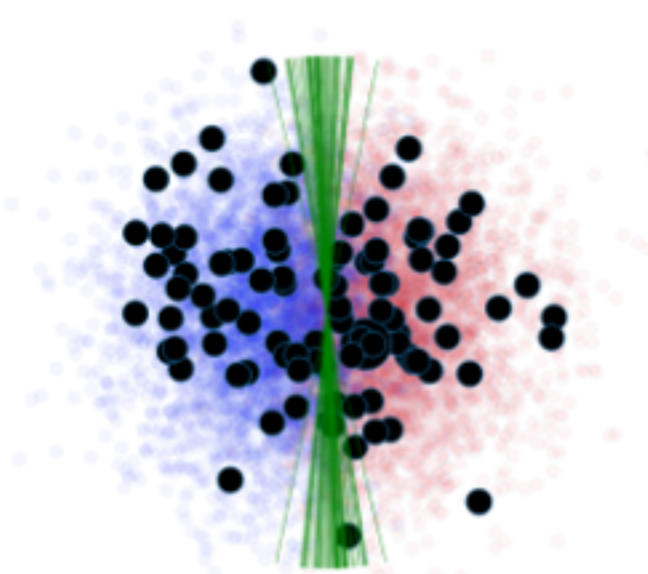
Uniform subsampling



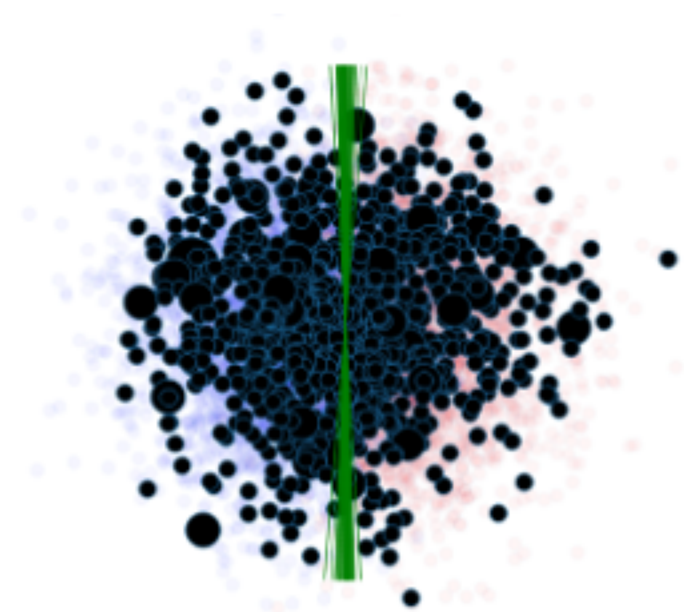
- Might miss important data
- Noisy estimates



$M = 10$



$M = 100$



$M = 1000$

Importance sampling

- "Optimal" importance weights

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

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

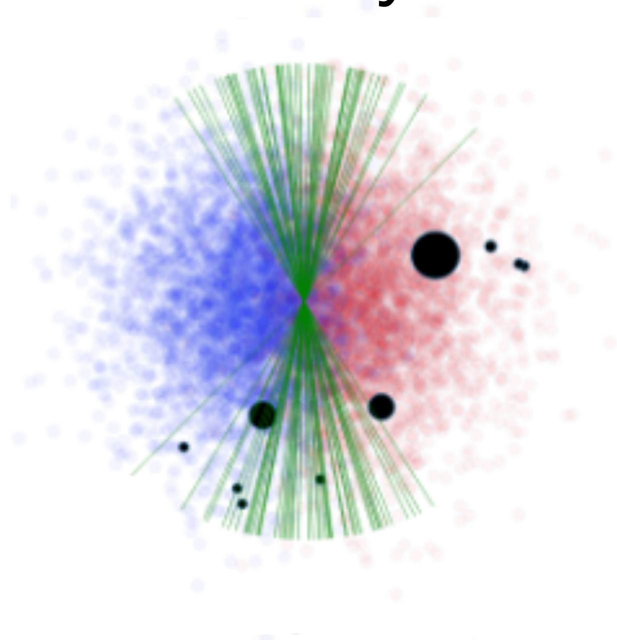
Importance sampling

- "Optimal" importance weights

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



$M = 10$

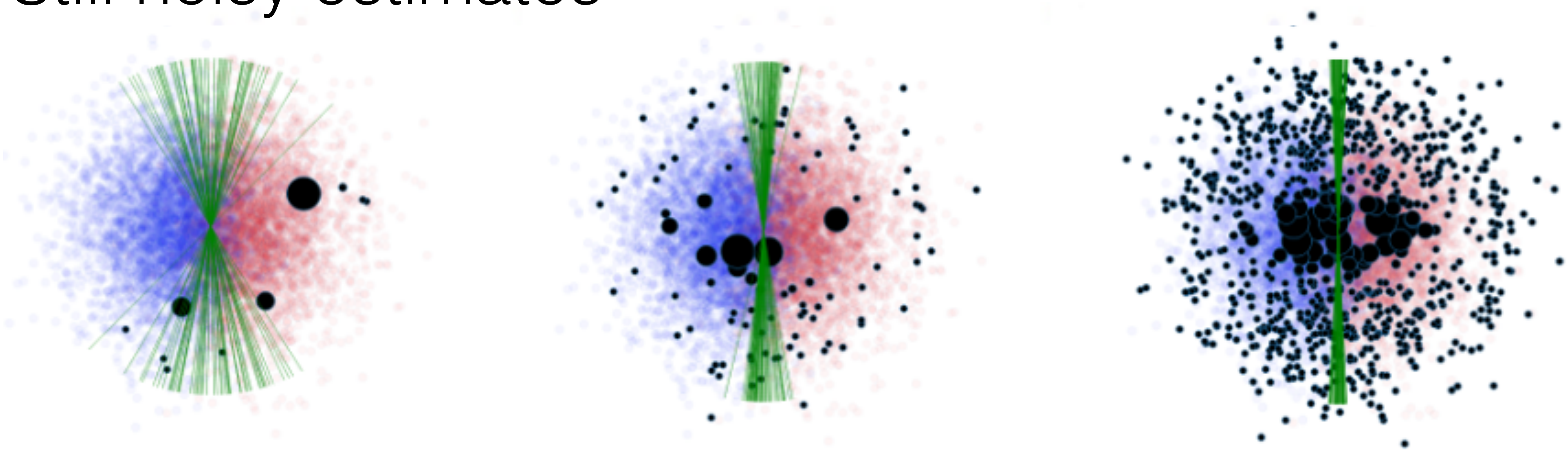
Importance sampling

- "Optimal" importance weights

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

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

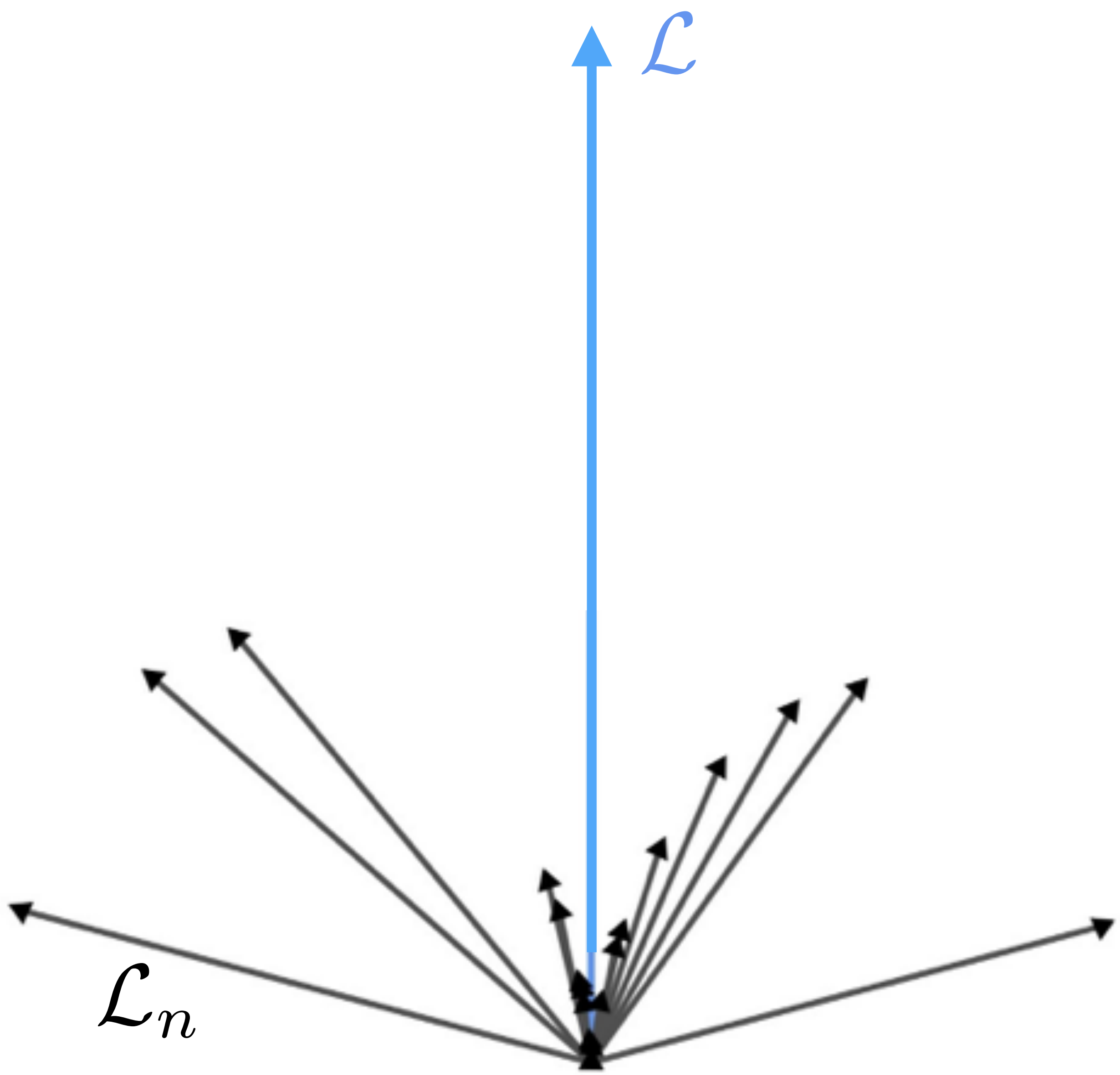
- Still noisy estimates



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

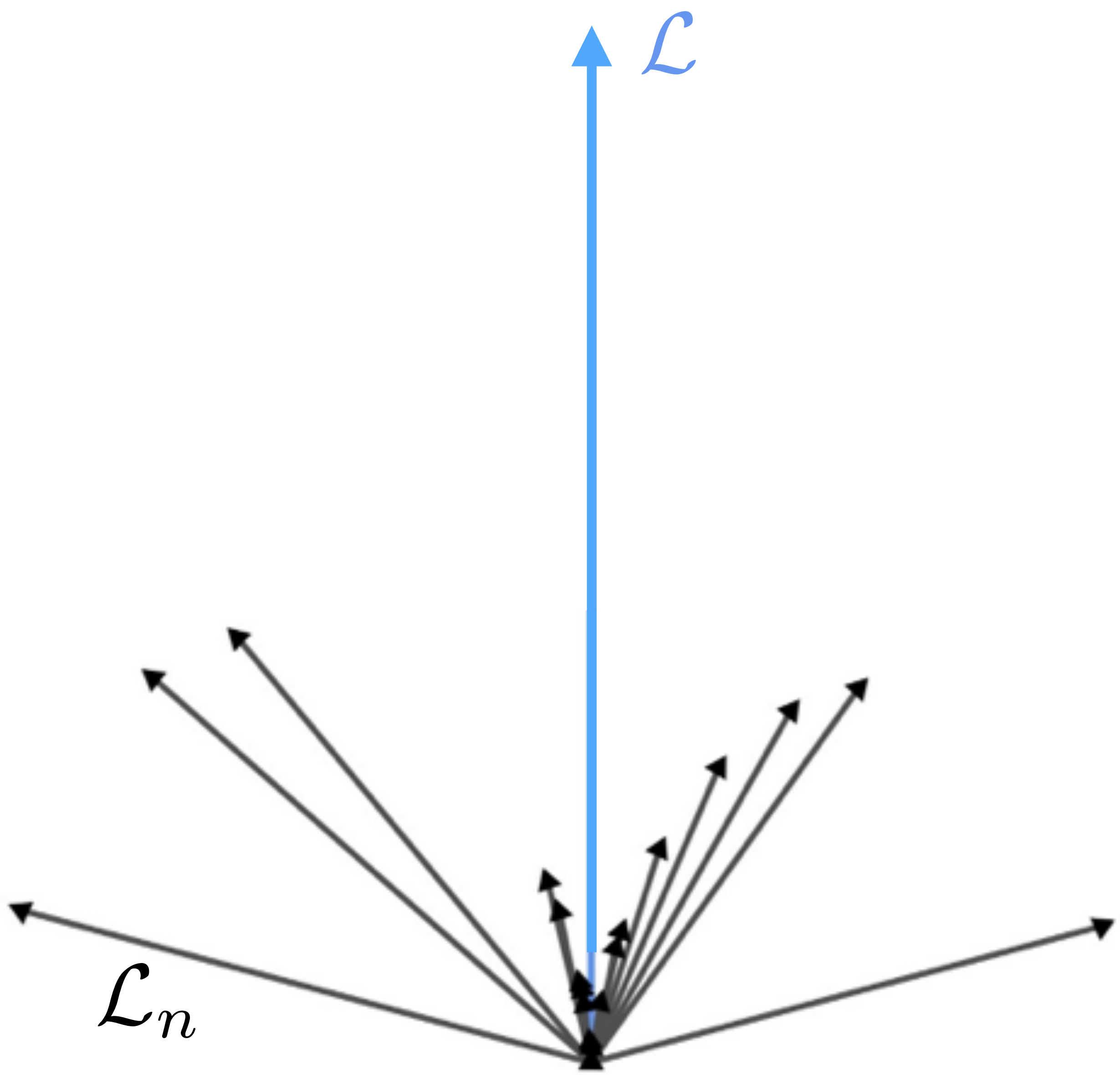
How to get a good Bayesian coreset?

- Want: Small error with few coreset points



How to get a good Bayesian coreset?

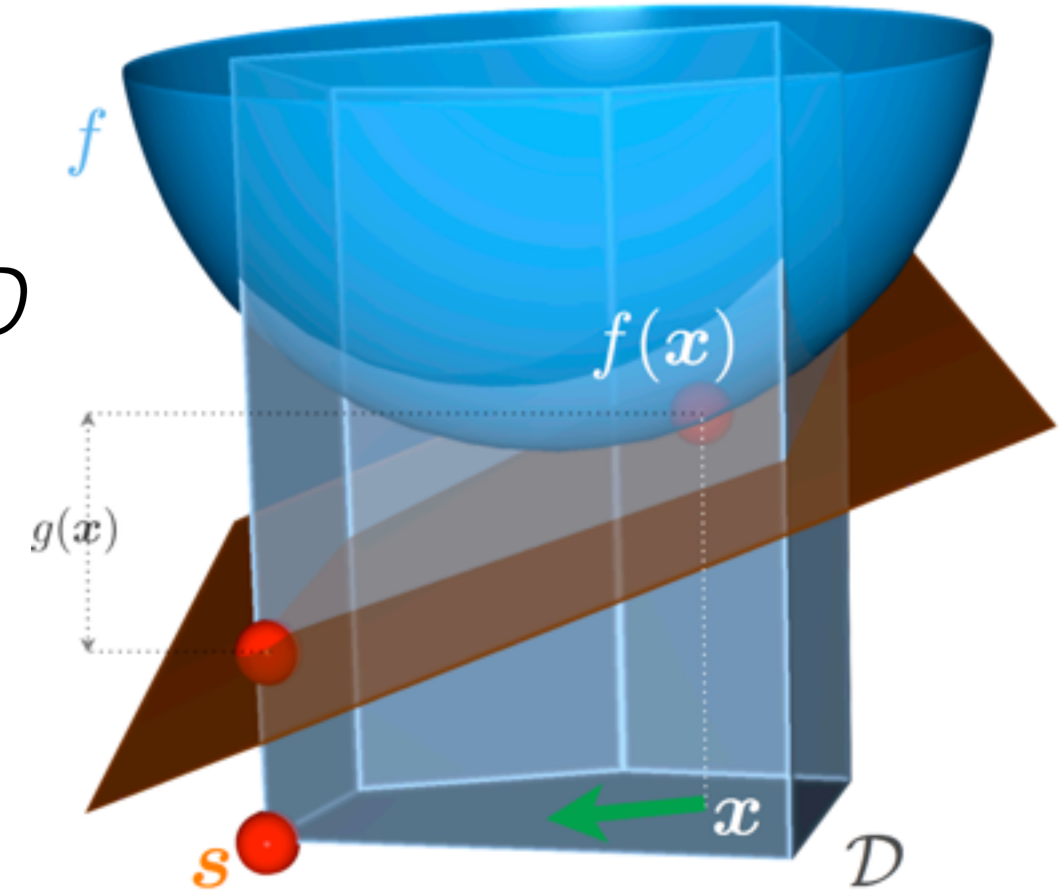
- Want: Small error with few coreset points



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

Frank-Wolfe

Convex optimization on a polytope *D*

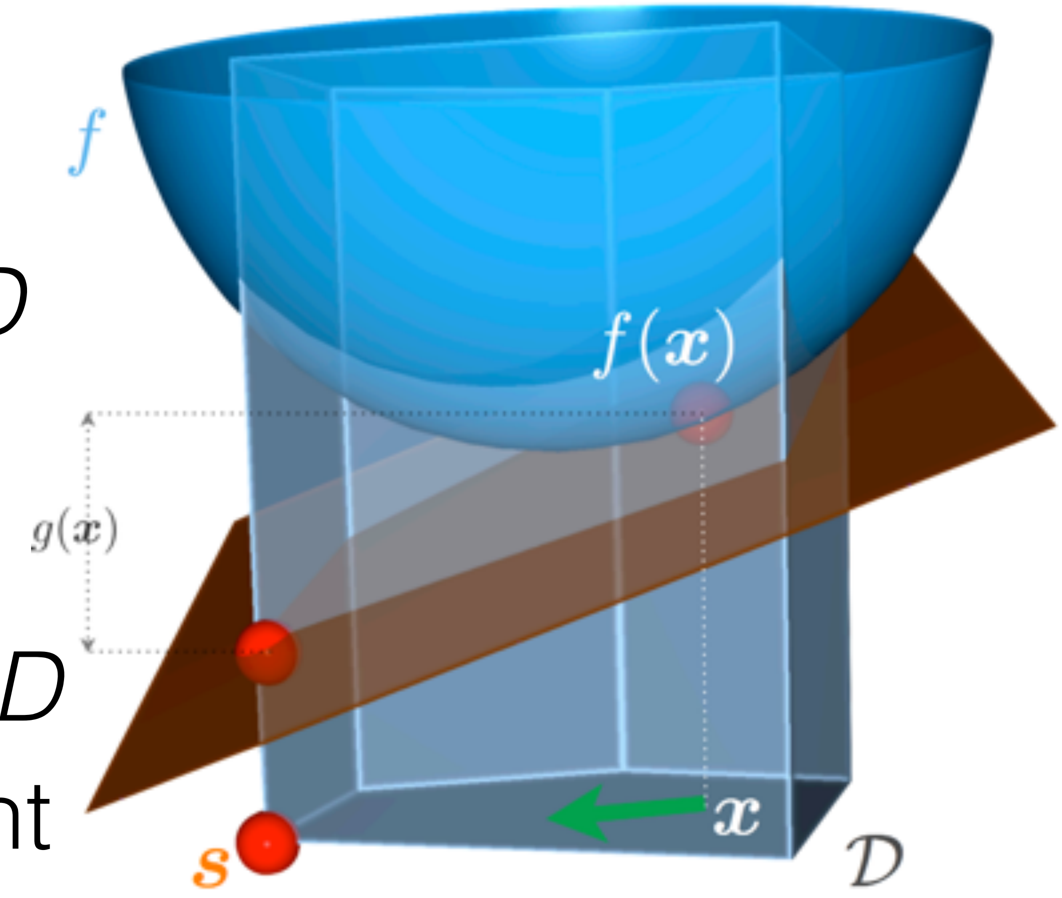


[Jaggi 2013]

Frank-Wolfe

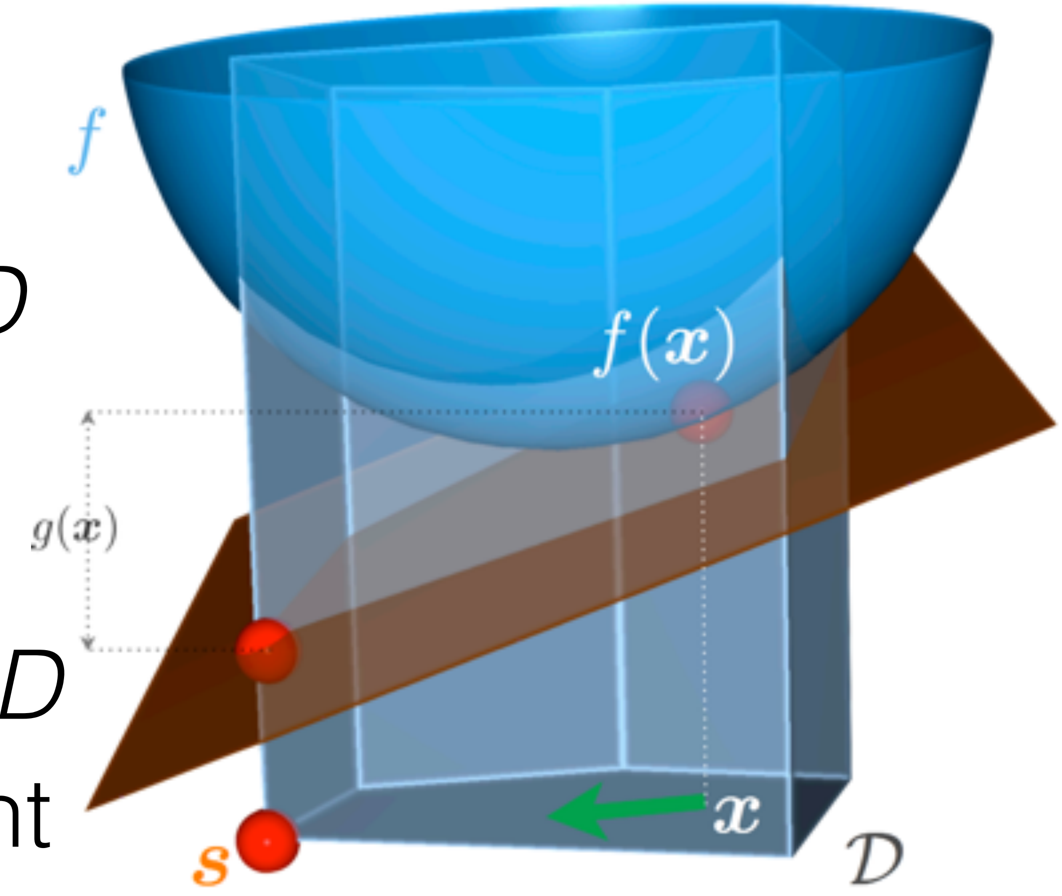
Convex optimization on a polytope *D*

- 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

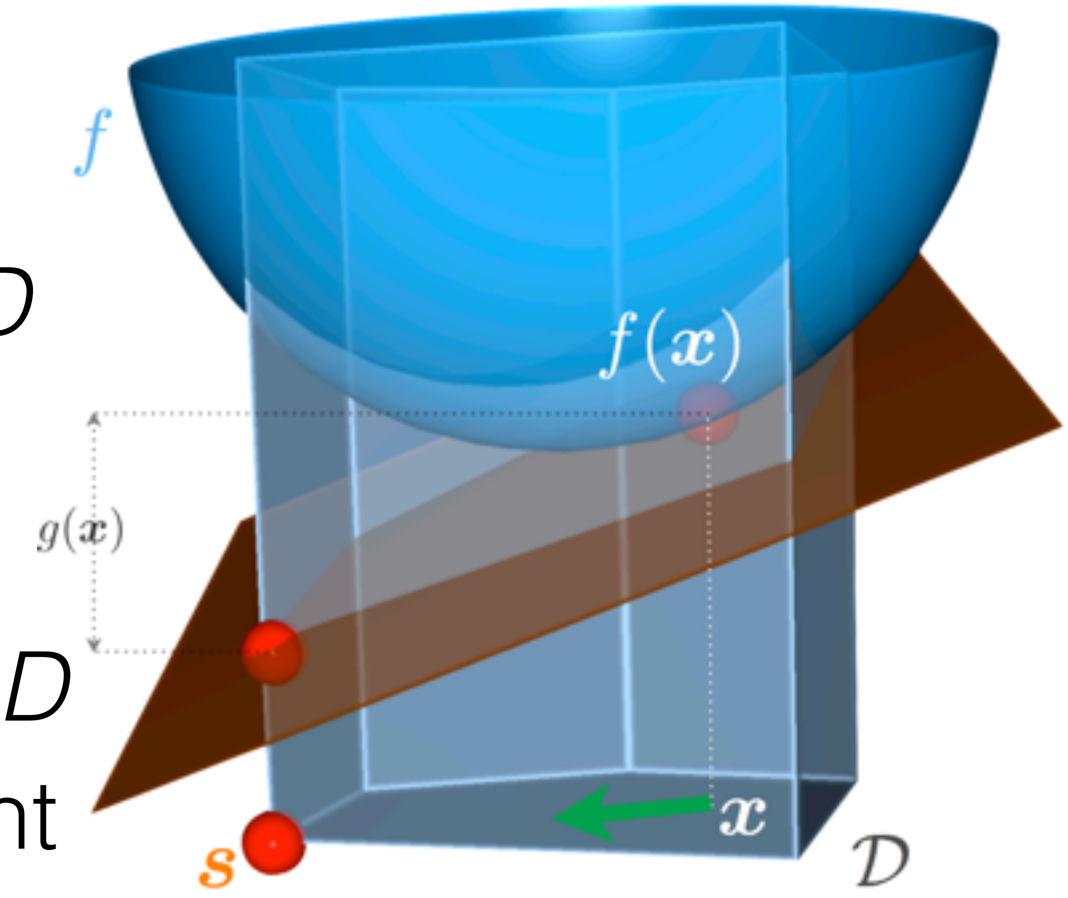


[Jaggi 2013]

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

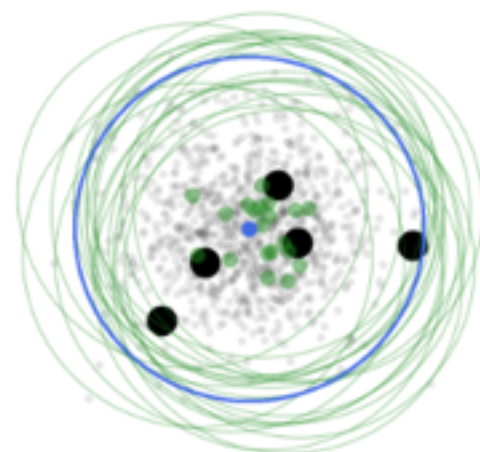
Thm (Campbell, B). After *M* iterations,

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

Gaussian model (simulated)

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

Uniform subsampling

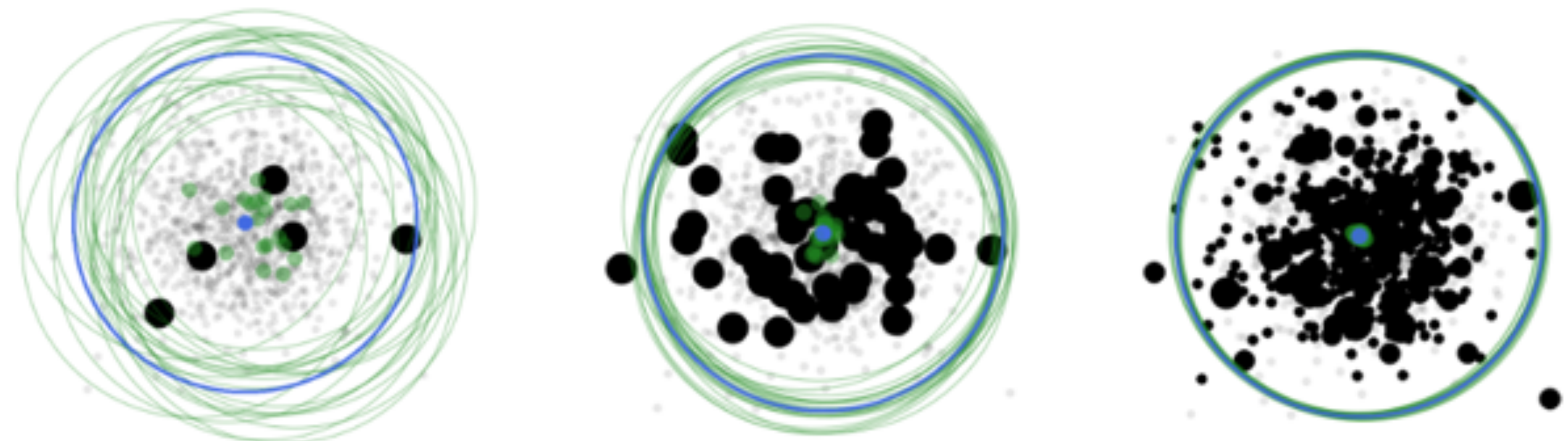


$M = 5$

Gaussian model (simulated)

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

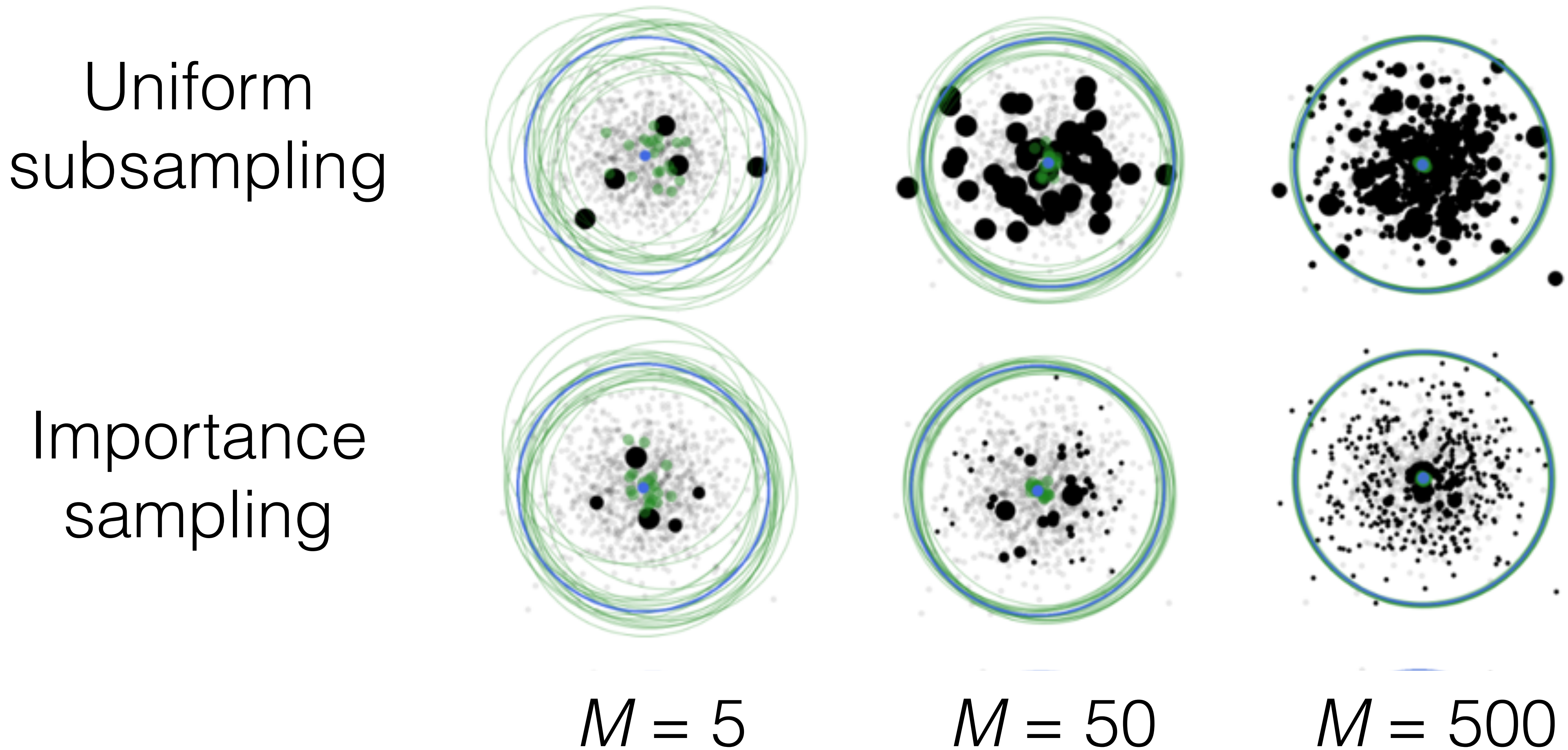
Uniform subsampling



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

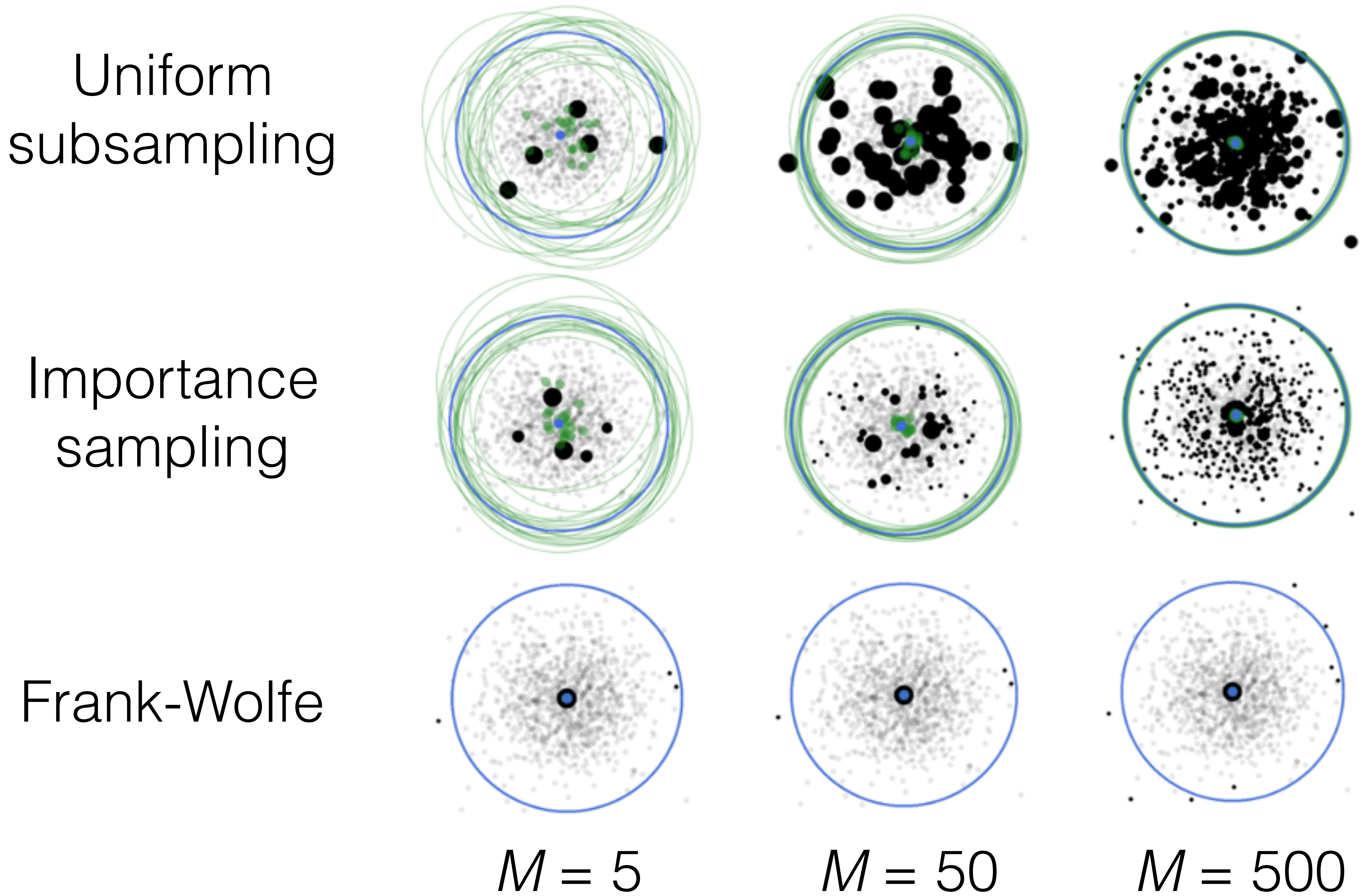
Gaussian model (simulated)

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



Gaussian model (simulated)

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



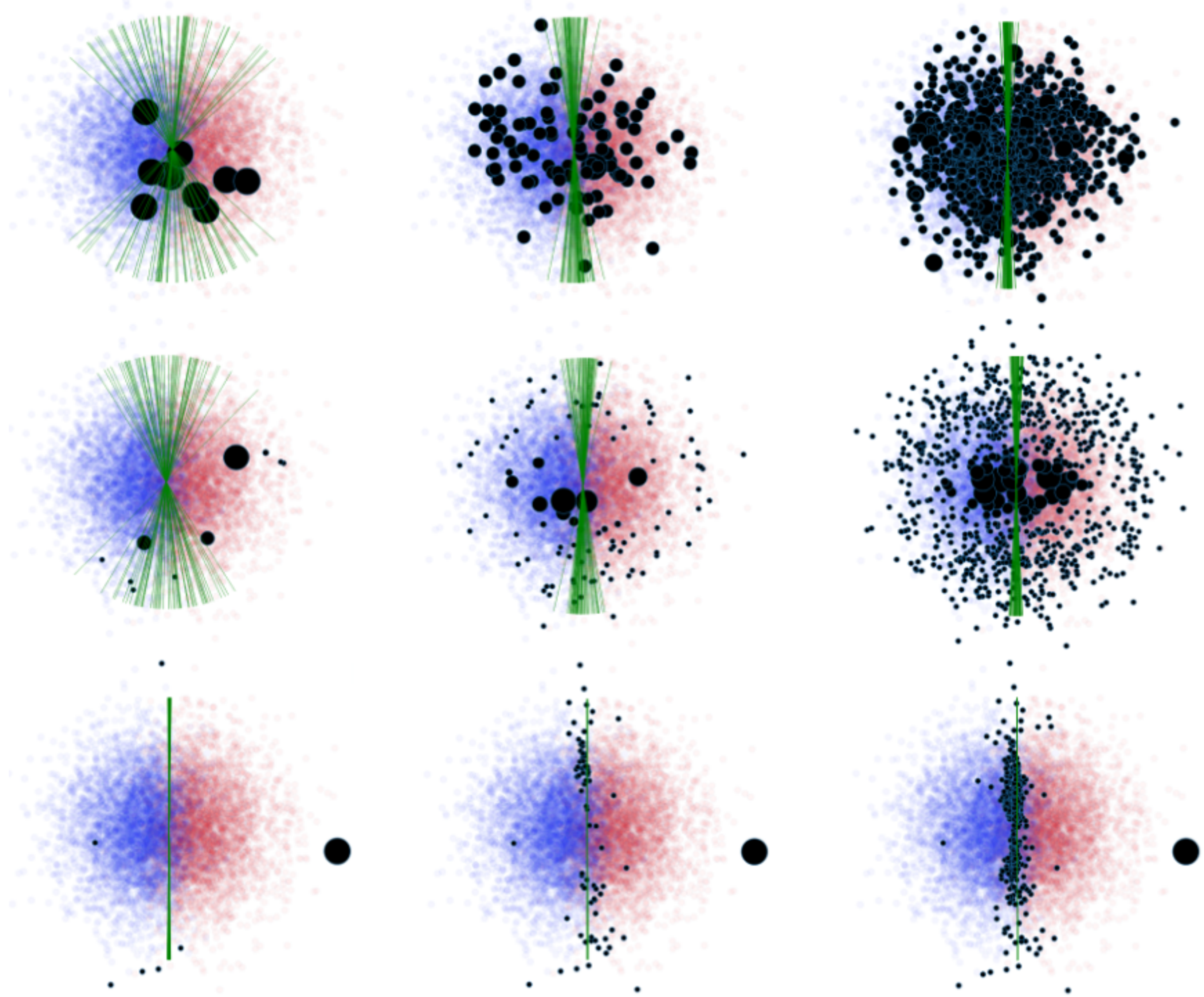
Logistic regression (simulated)

- 10K data points

Uniform subsampling

Importance sampling

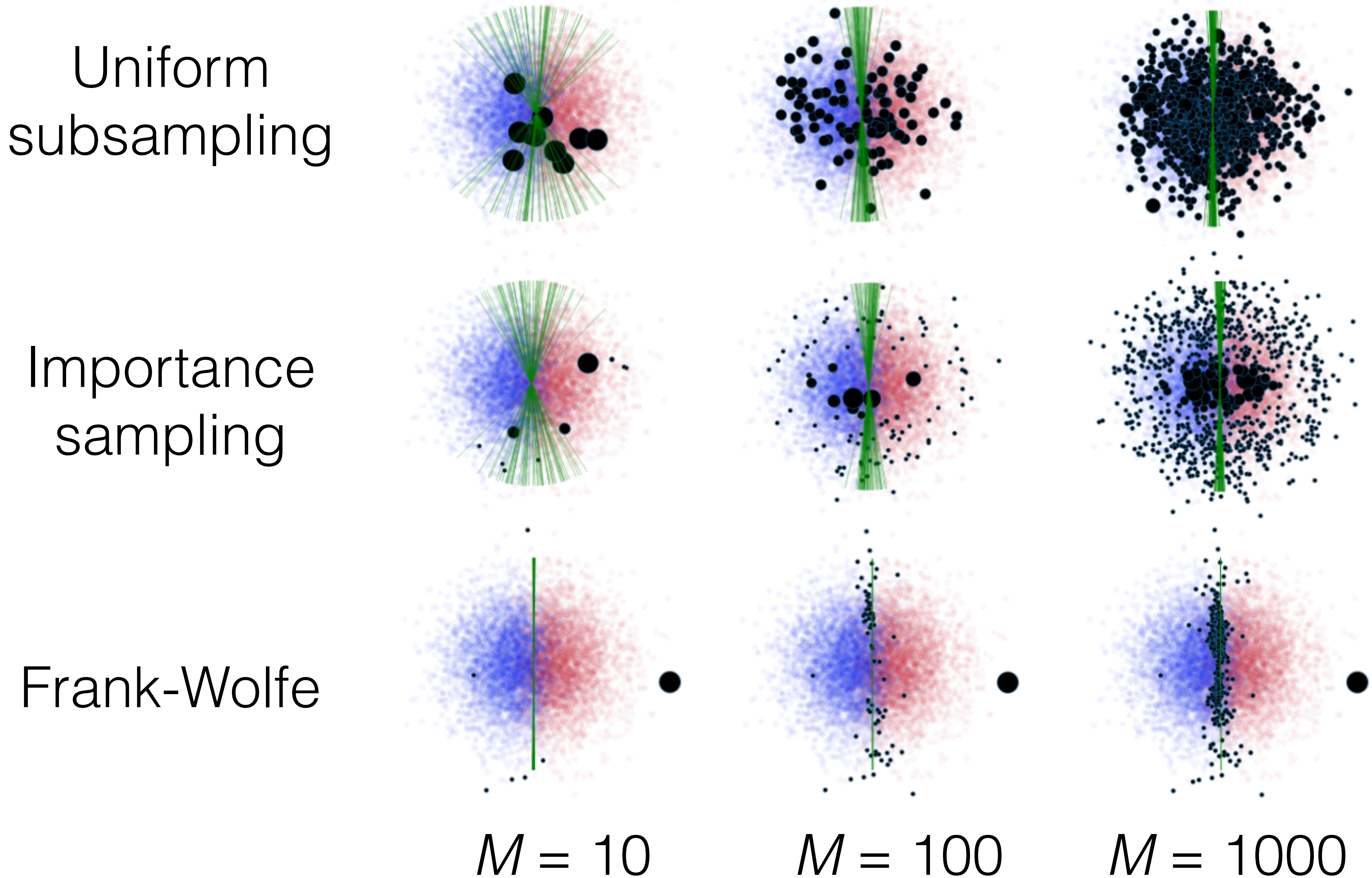
Frank-Wolfe



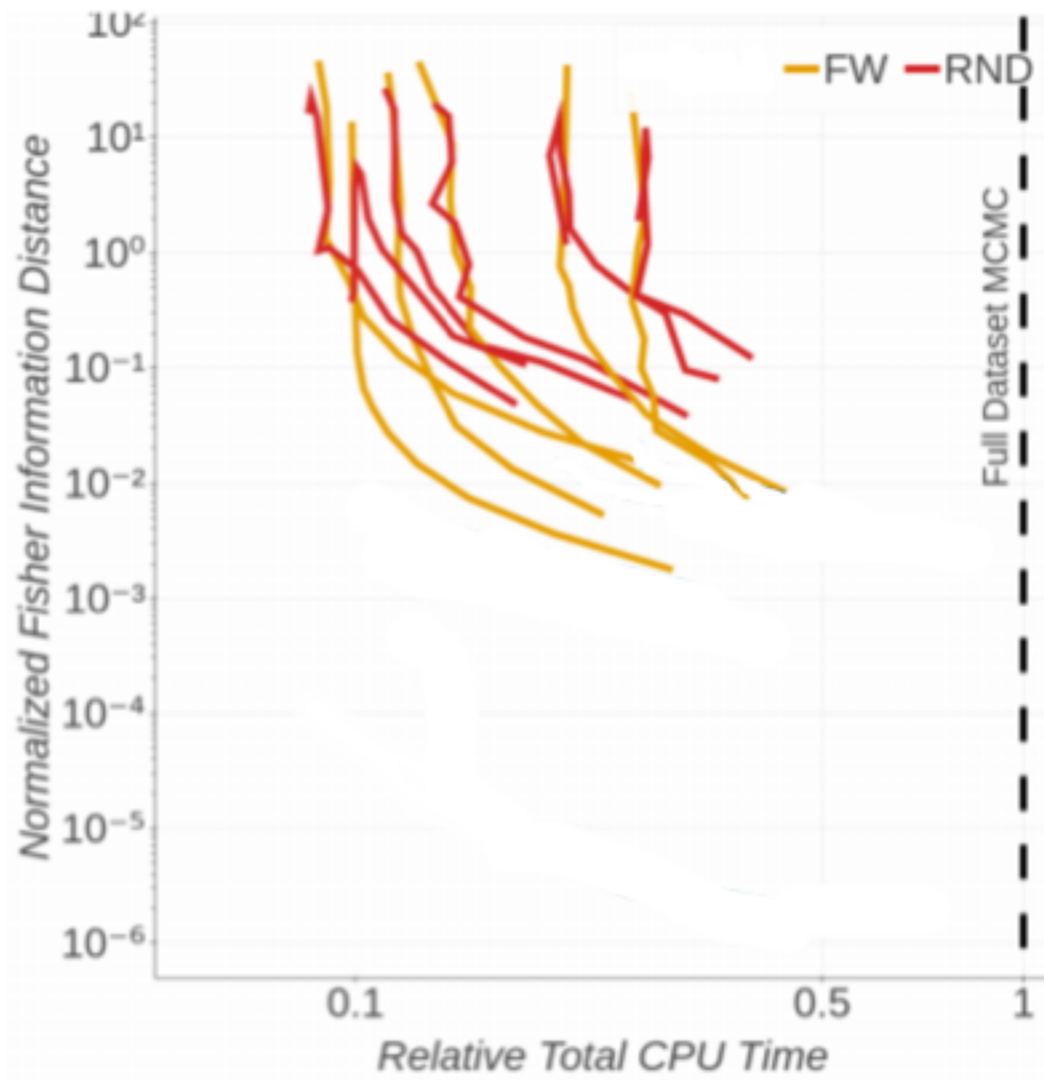
$M = 10$ $M = 100$ $M = 1000$

Logistic regression (simulated)

- 10K data points
- similar for Poisson regression, spherical clustering



Real data experiments



lower error

less total time

Uniform subsampling

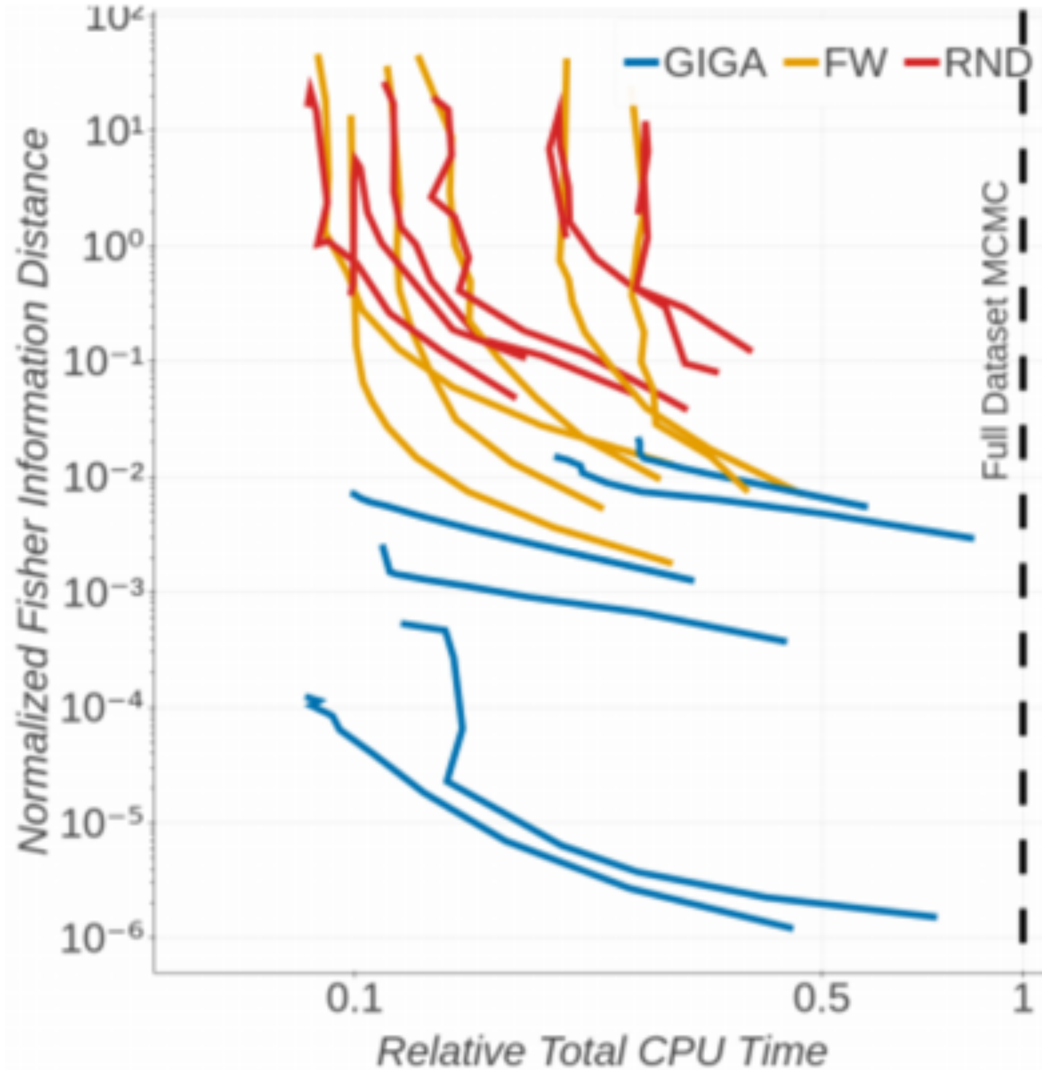
Frank Wolfe coresets

Data sets include:

- Phishing
- Chemical reactivity
- Bicycle trips
- Airport delays

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

Conclusions

- *Data summarization* for **scalable**, **automated** approx. Bayes algorithms with **error bounds on output quality (for finite data)**
  - Also: PASS-GLM: 6M pts, 1K features, 22 cores ➔16 s

Conclusions

- *Data summarization* for **scalable**, **automated** approx. Bayes algorithms with **error bounds on output quality (for finite data)**
  - Also: PASS-GLM: 6M pts, 1K features, 22 cores ➔16 s

R Agrawal, T Campbell, JH Huggins, and T Broderick. Data-dependent compression of random features for large-scale kernel approximation. ArXiv:1810.04249

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

T Campbell and T Broderick. Bayesian coreset construction via Greedy Iterative Geodesic Ascent. *ICML* 2018.

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.

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

References (2/6)

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. *NeurIPS* 2016.

T Broderick. “Variational Bayes and beyond: Bayesian inference for big data” ICML Tutorial, 2018. http://www.tamarabroderick.com/tutorial_2018_icml.html

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.

Y Chen, M Welling, and A Smola. Super-samples from kernel herding. *UAI* 2010.

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.

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.

References (3/6)

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. *NeurIPS* 2015.

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

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.

JH Huggins, T Campbell, M Kasprzak, and T Broderick. Scalable Gaussian Process inference with finite-data mean and variance guarantees. Under review. ArXiv:1806.10234.

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

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. *The Journal of Machine Learning Research* 18.1 (2017): 430-474.

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

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.

References (4/6)

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

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

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.

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

Application References (5/6)

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.

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.

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.

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.

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.

Additional image references (6/6)

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

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)