MIT EECS
ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

MIT STATISTICS + DATA SCIENCE CENTER

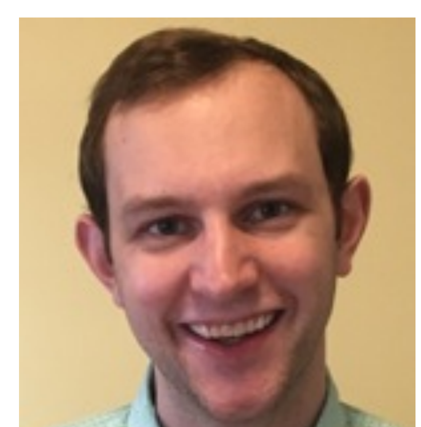
CSAIL

Coresets for Bayesian Logistic Regression

Tamara Broderick

ITT Career Development
Assistant Professor,
MIT

With: Jonathan H. Huggins, Trevor Campbell



Bayesian inference

Bayesian inference

- Complex, modular

Bayesian inference

- Complex, modular; coherent uncertainties

Bayesian inference

- Complex, modular; coherent uncertainties; prior info

Bayesian inference

- Complex, modular; coherent uncertainties; prior info

$$p(\theta)$$

Bayesian inference

- Complex, modular; coherent uncertainties; prior info

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

Bayesian inference

- Complex, modular; coherent uncertainties; prior info

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

Bayesian inference

- Complex, modular; coherent uncertainties; prior info

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

- MCMC

Bayesian inference

- Complex, modular; coherent uncertainties; prior info

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

- MCMC: Accurate but can be slow [Bardenet, Doucet, Holmes 2015]

Bayesian inference

- Complex, modular; coherent uncertainties; prior info

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

- MCMC: Accurate but can be slow [Bardenet, Doucet, Holmes 2015]
- (Mean-field) variational Bayes: (MF)VB

Bayesian inference

- Complex, modular; coherent uncertainties; prior info

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

- MCMC: Accurate but can be slow [Bardenet, Doucet, Holmes 2015]
- (Mean-field) variational Bayes: (MF)VB
  - Fast

Bayesian inference

- Complex, modular; coherent uncertainties; prior info

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

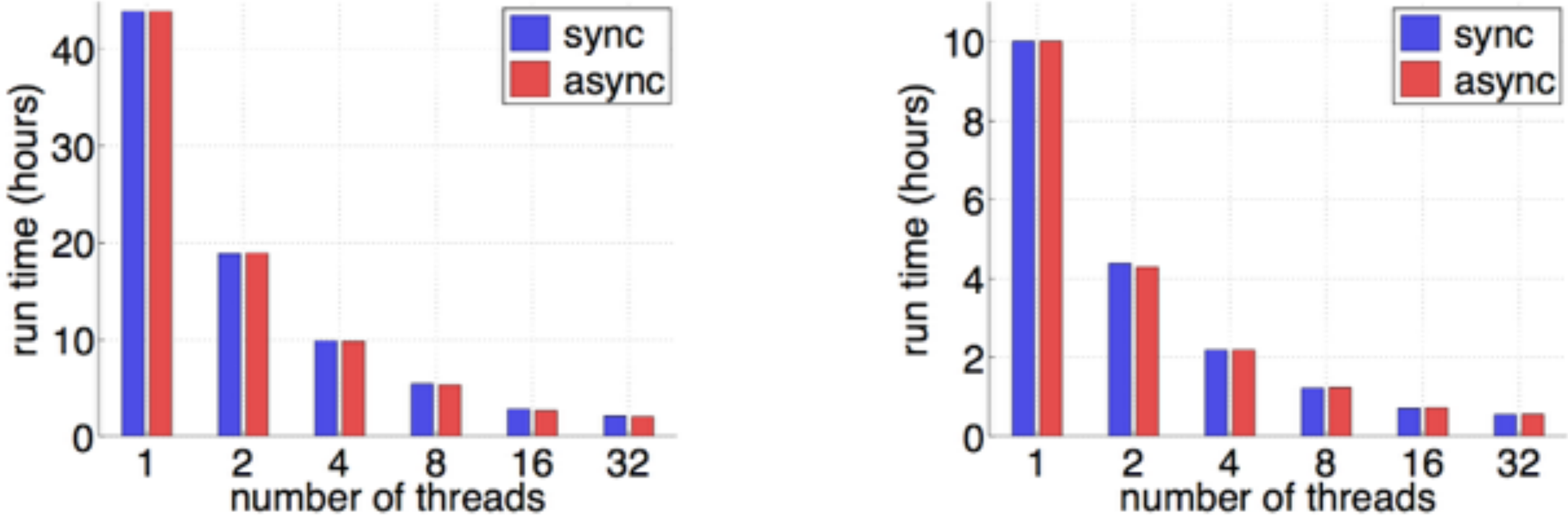
- MCMC: Accurate but can be slow [Bardenet, Doucet, Holmes 2015]
- (Mean-field) variational Bayes: (MF)VB
  - Fast, streaming, distributed [Broderick, Boyd, Wibisono, Wilson, Jordan 2013]

Bayesian inference

- Complex, modular; coherent uncertainties; prior info

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

- MCMC: Accurate but can be slow [Bardenet, Doucet, Holmes 2015]
- (Mean-field) variational Bayes: (MF)VB
  - Fast, streaming, distributed [Broderick, Boyd, Wibisono, Wilson, Jordan 2013]



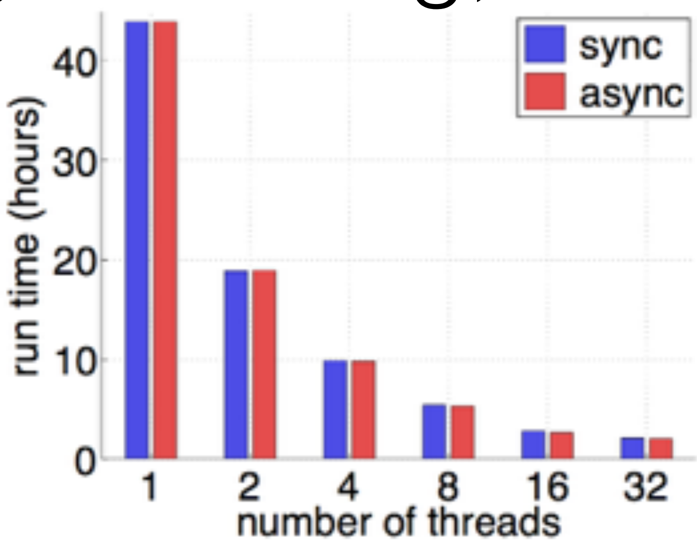
(c) Wikipedia (3.6M) (d) Nature (350K)

Bayesian inference

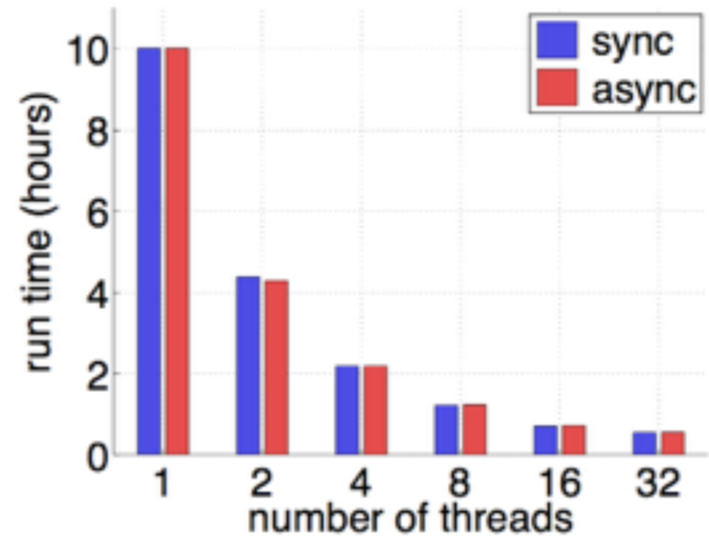
- Complex, modular; coherent uncertainties; prior info

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

- MCMC: Accurate but can be slow [Bardenet, Doucet, Holmes 2015]
- (Mean-field) variational Bayes: (MF)VB
  - Fast, streaming, distributed [Broderick, Boyd, Wibisono, Wilson, Jordan 2013]



(c) Wikipedia (3.6M)



(d) Nature (350K)

- Misestimation & lack of quality guarantees

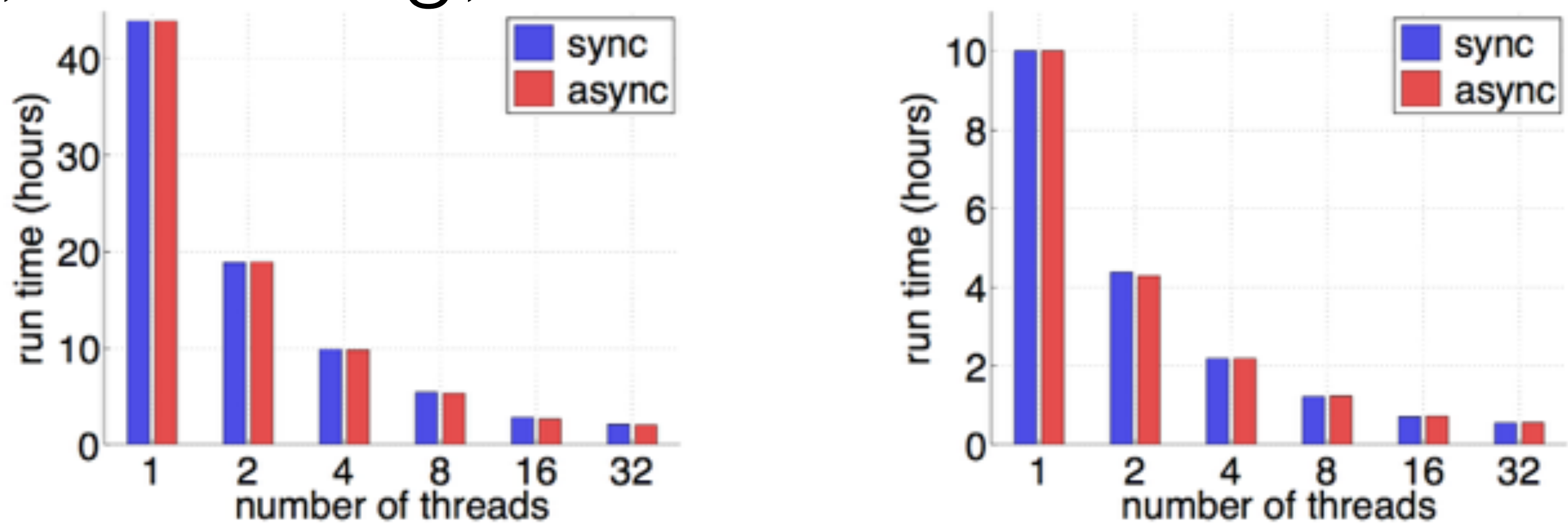
[MacKay 2003; Bishop 2006; Wang, Titterington 2004; Turner, Sahani 2011]

Bayesian inference

- Complex, modular; coherent uncertainties; prior info

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

- MCMC: Accurate but can be slow [Bardenet, Doucet, Holmes 2015]
- (Mean-field) variational Bayes: (MF)VB
  - Fast, streaming, distributed [Broderick, Boyd, Wibisono, Wilson, Jordan 2013]



(c) Wikipedia (3.6M) (d) Nature (350K)

- Misestimation & lack of quality guarantees

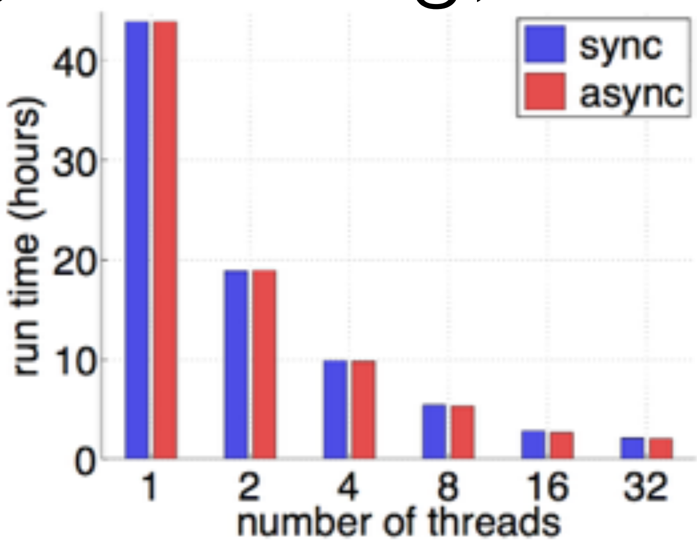
[MacKay 2003; Bishop 2006; Wang, Titterington 2004; Turner, Sahani 2011; Fosdick 2013; Dunson 2014; Bardenet, Doucet, Holmes 2015]

Bayesian inference

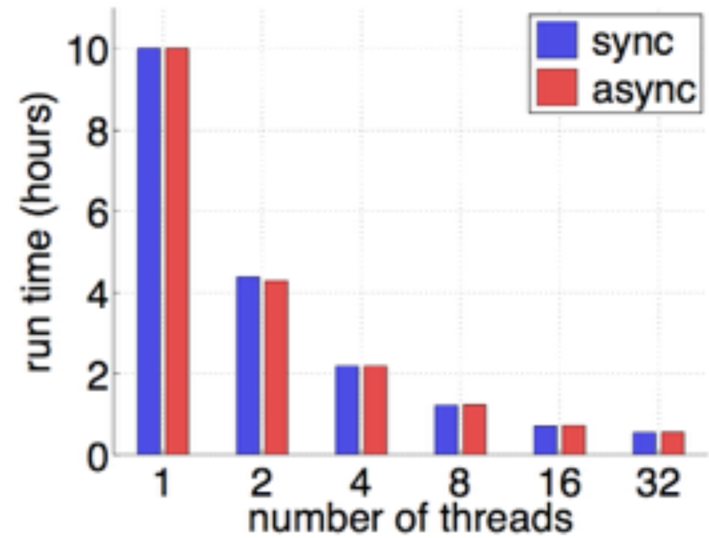
- Complex, modular; coherent uncertainties; prior info

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

- MCMC: Accurate but can be slow [Bardenet, Doucet, Holmes 2015]
- (Mean-field) variational Bayes: (MF)VB
  - Fast, streaming, distributed [Broderick, Boyd, Wibisono, Wilson, Jordan 2013]



(c) Wikipedia (3.6M)



(d) Nature (350K)

- Misestimation & lack of quality guarantees

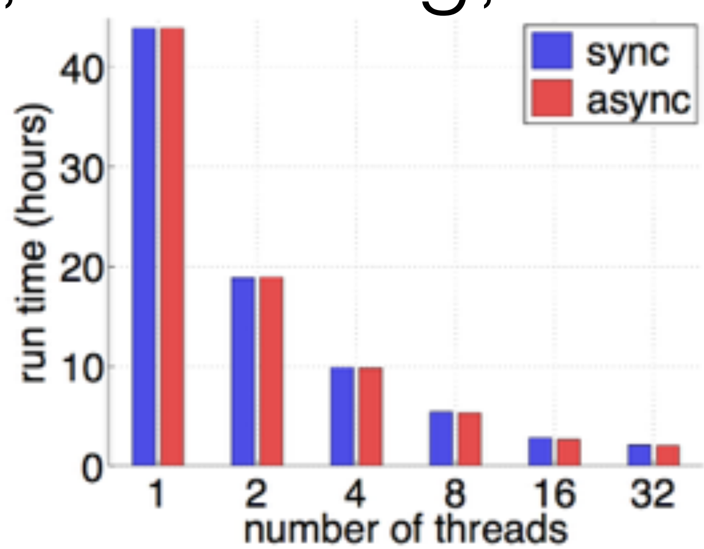
[MacKay 2003; Bishop 2006; Wang, Titterington 2004; Turner, Sahani 2011; Fosdick 2013; Dunson 2014; Bardenet, Doucet, Holmes 2015; Opper, Winther 2003; Giordano, Broderick, Jordan 2015]

Bayesian inference

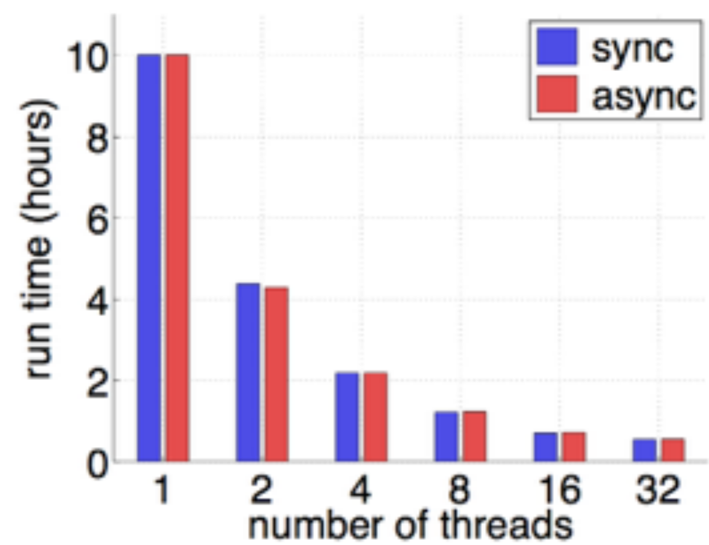
- Complex, modular; coherent uncertainties; prior info

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

- MCMC: Accurate but can be slow [Bardenet, Doucet, Holmes 2015]
- (Mean-field) variational Bayes: (MF)VB
  - Fast, streaming, distributed [Broderick, Boyd, Wibisono, Wilson, Jordan 2013]



(c) Wikipedia (3.6M)



(d) Nature (350K)

- Misestimation & lack of quality guarantees

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

- Our proposal: use data summarization for fast, streaming, distributed algs. with theoretical guarantees

Data summarization

Data summarization

- Exponential family likelihood

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

Data summarization

- Exponential family likelihood

Sufficient statistics

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

Data summarization

- Exponential family likelihood

Sufficient statistics

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

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

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

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 $p(y_{1:N}|x_{1:N},\theta) = \prod_{n=1}^{N} \frac{1}{1+\exp(-y_n x_n \cdot \theta)}$

Data summarization

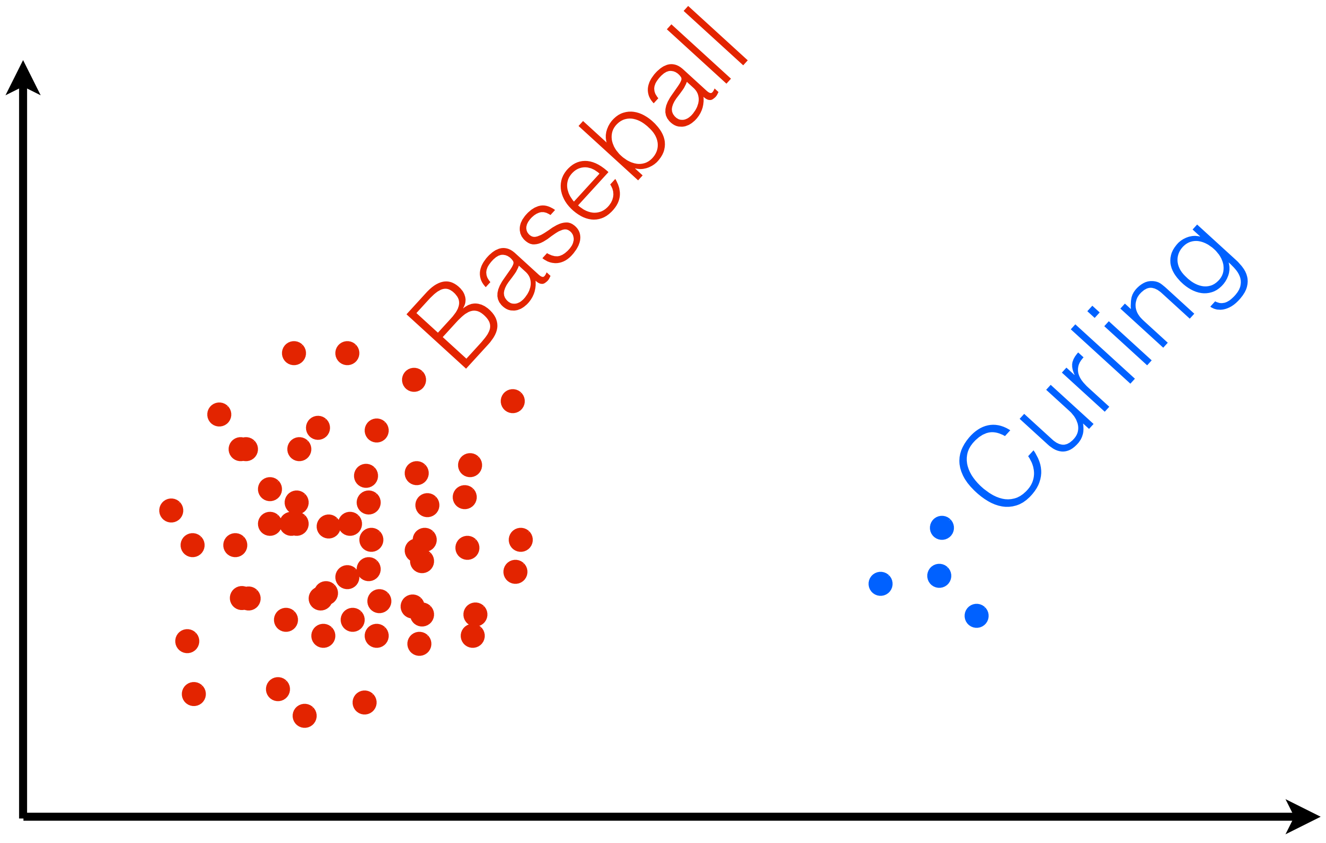
- Exponential family likelihood

Sufficient statistics

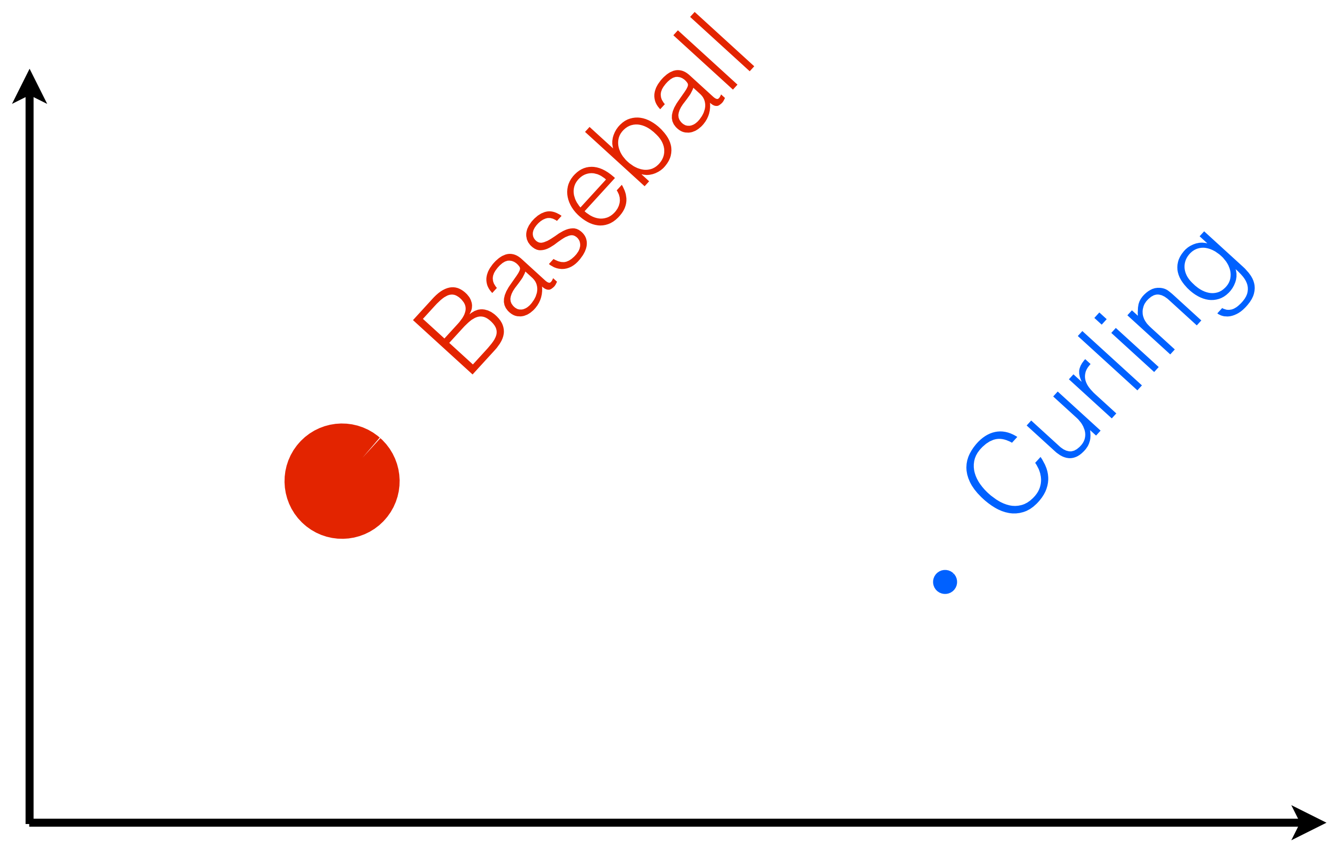
$$p(y_{1:N}|x_{1:N},\theta) = \prod_{n=1}^{N} \exp\left[\boxed{T(y_n,x_n)} \cdot \eta(\theta)\right]$$

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

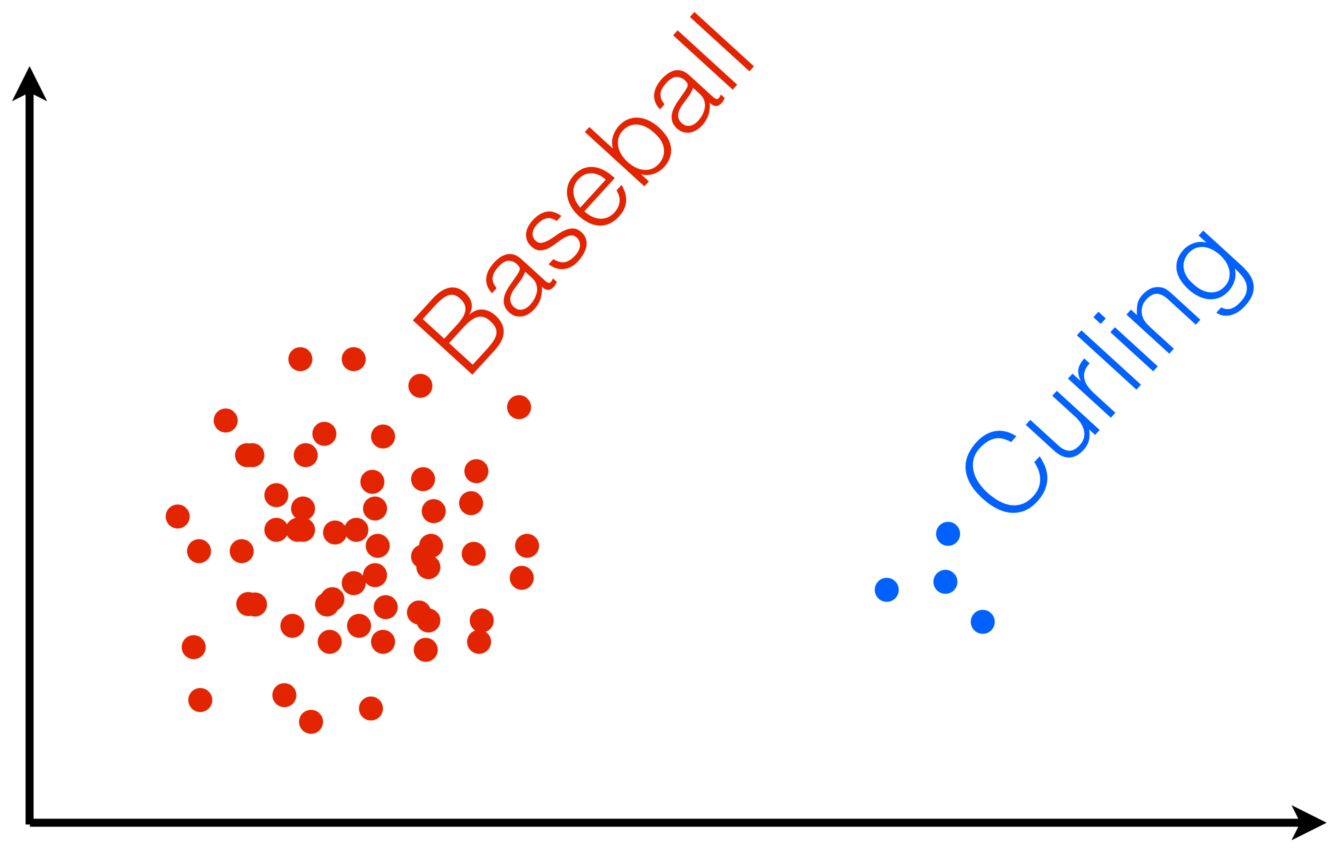
- Scalable, single-pass, streaming, distributed, complementary to MCMC
- *But*: Often no simple sufficient statistics
  - E.g. Bayesian logistic regression; GLMs; "deeper" models
    - Likelihood $p(y_{1:N}|x_{1:N},\theta) = \prod_{n=1}^{N} \frac{1}{1+\exp(-y_n x_n \cdot \theta)}$
- Our proposal: *approximate*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



[Agarwal et al 2005, Feldman & Langberg 2011]

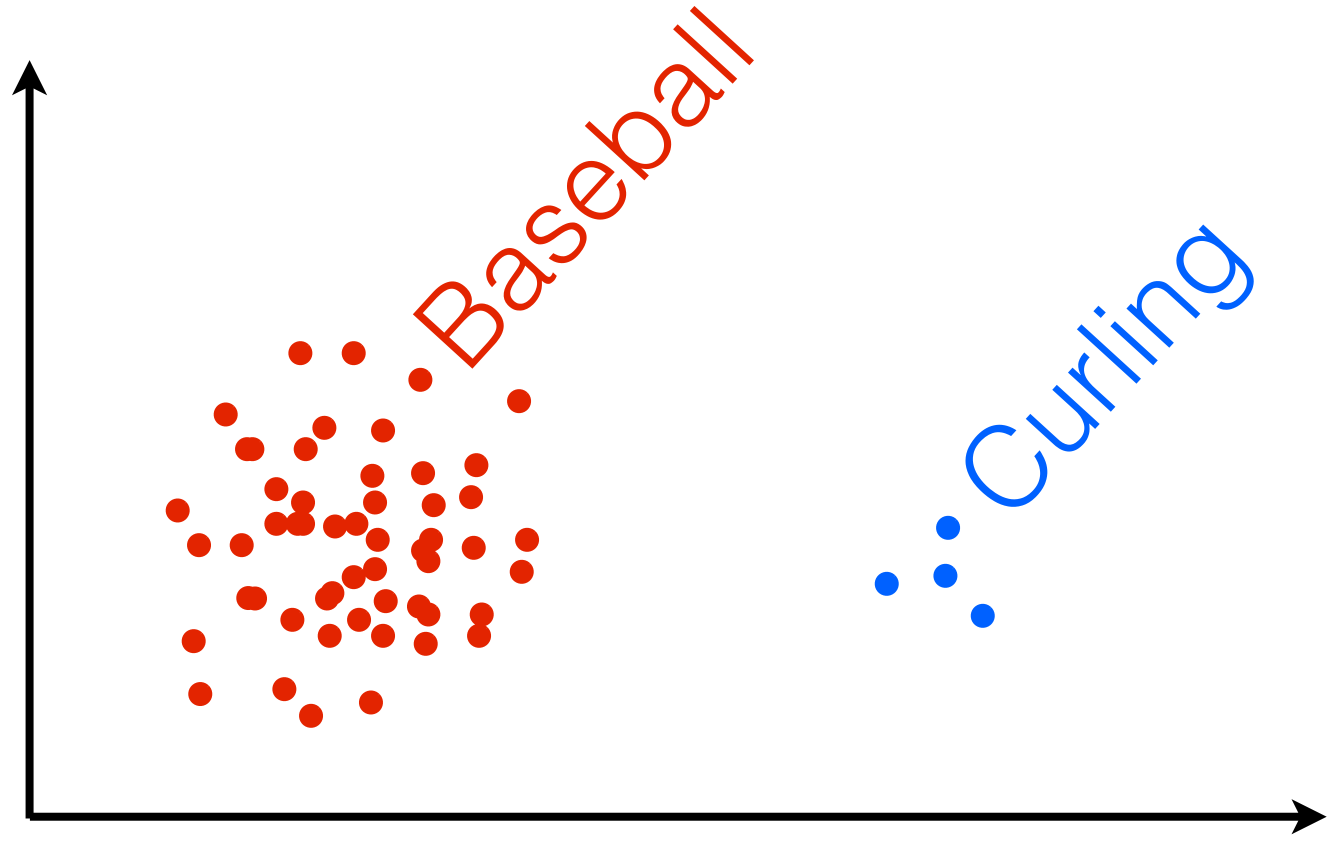


[Agarwal et al 2005, Feldman & Langberg 2011]



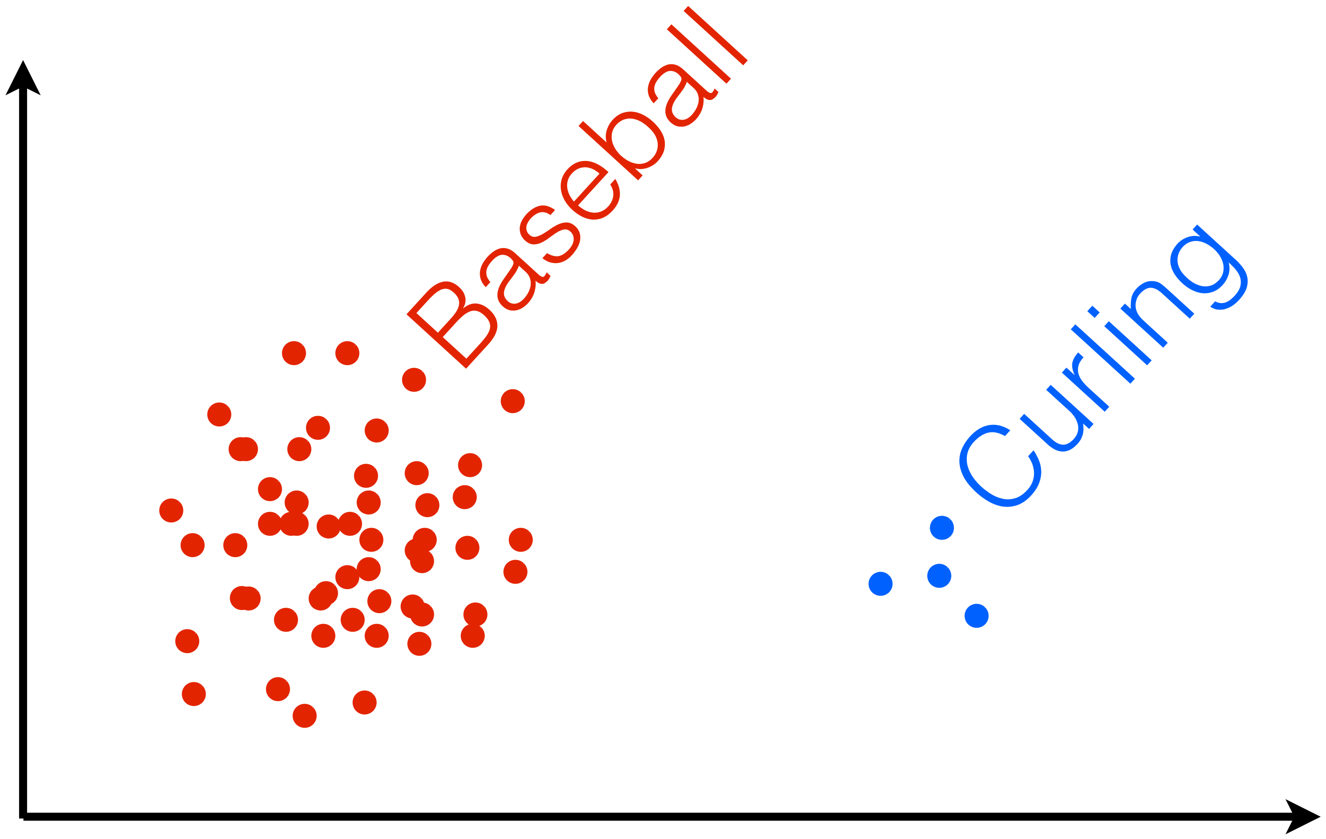
[Agarwal et al 2005, Feldman & Langberg 2011]

Coresets



[Agarwal et al 2005, Feldman & Langberg 2011]

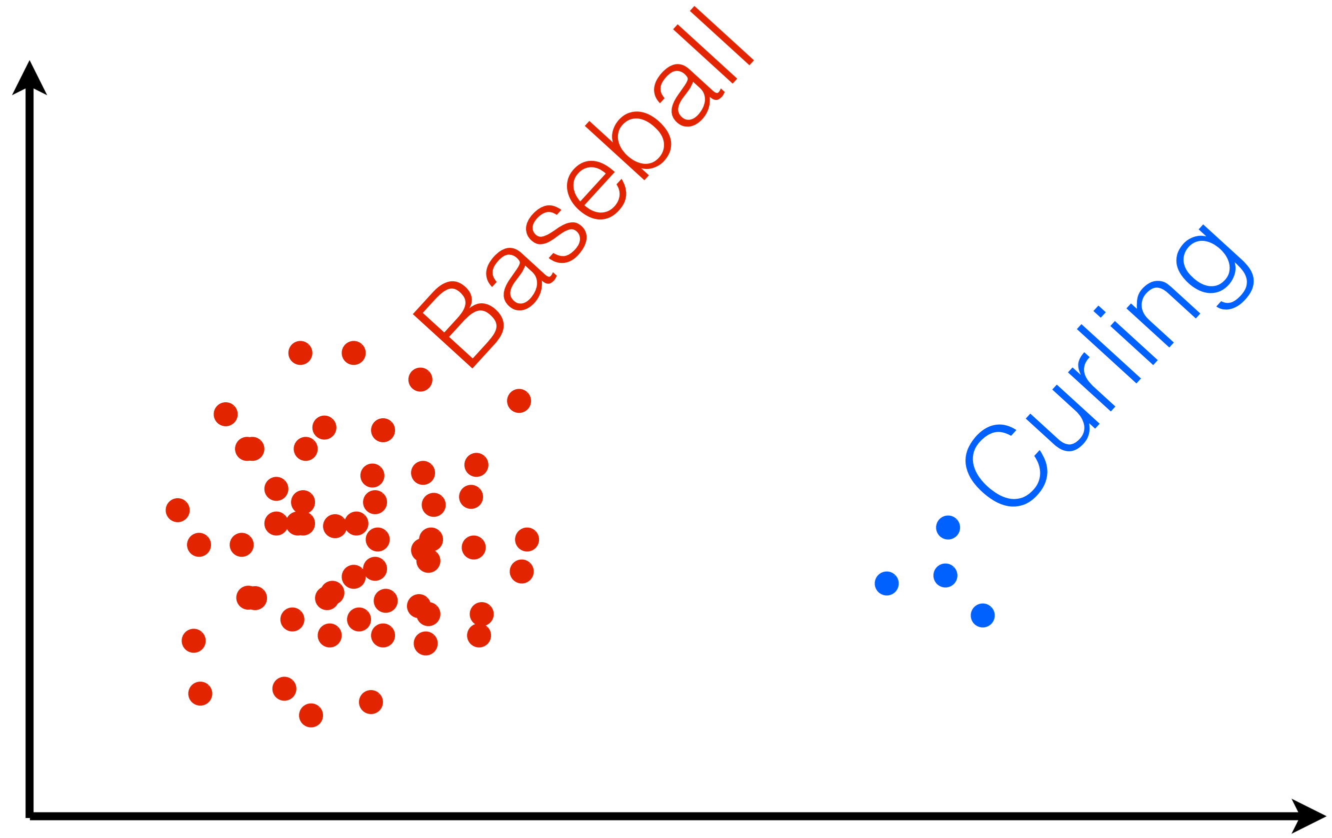
Coresets



- Pre-process data to get a smaller, weighted data set

[Agarwal et al 2005, Feldman & Langberg 2011]

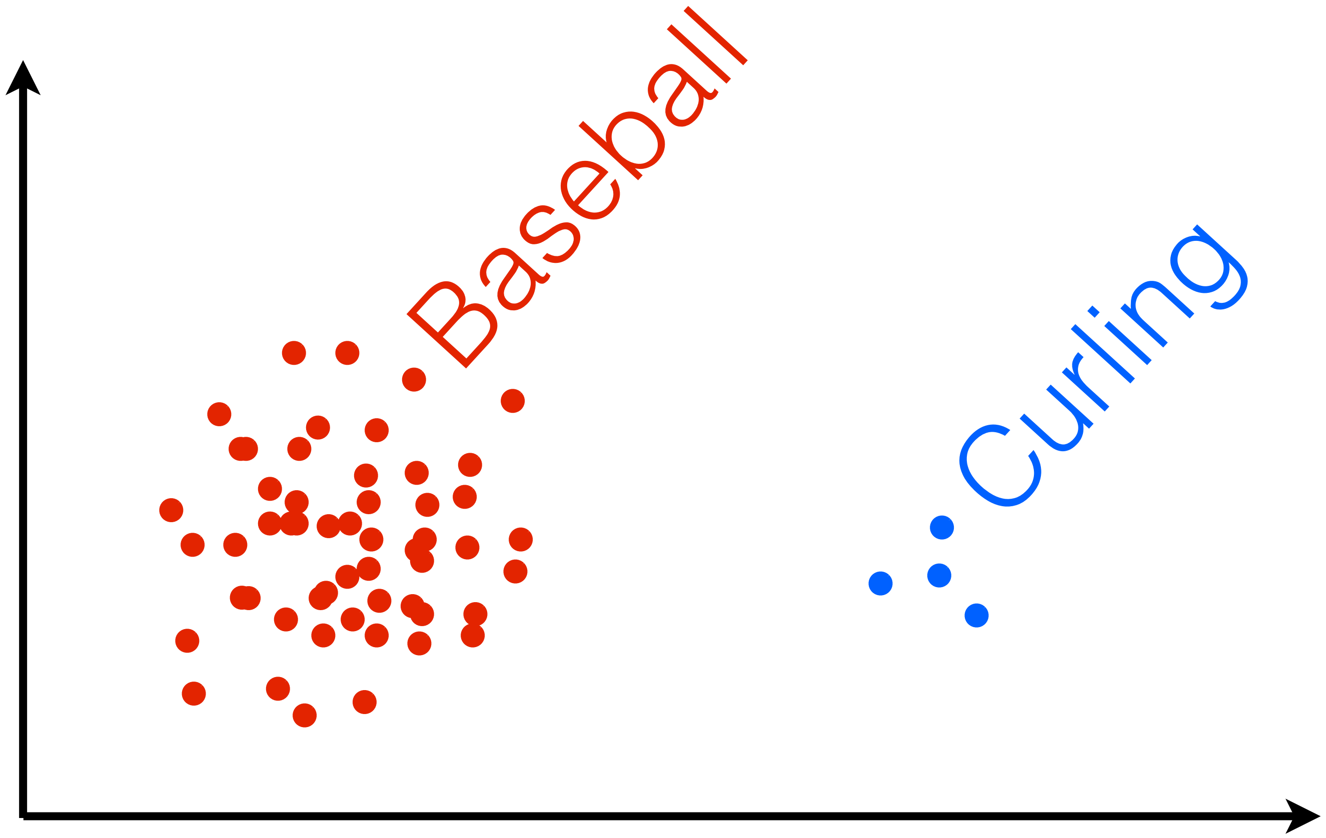
Coresets



- Pre-process data to get a smaller, weighted data set
- Theoretical guarantees on quality

[Agarwal et al 2005, Feldman & Langberg 2011]

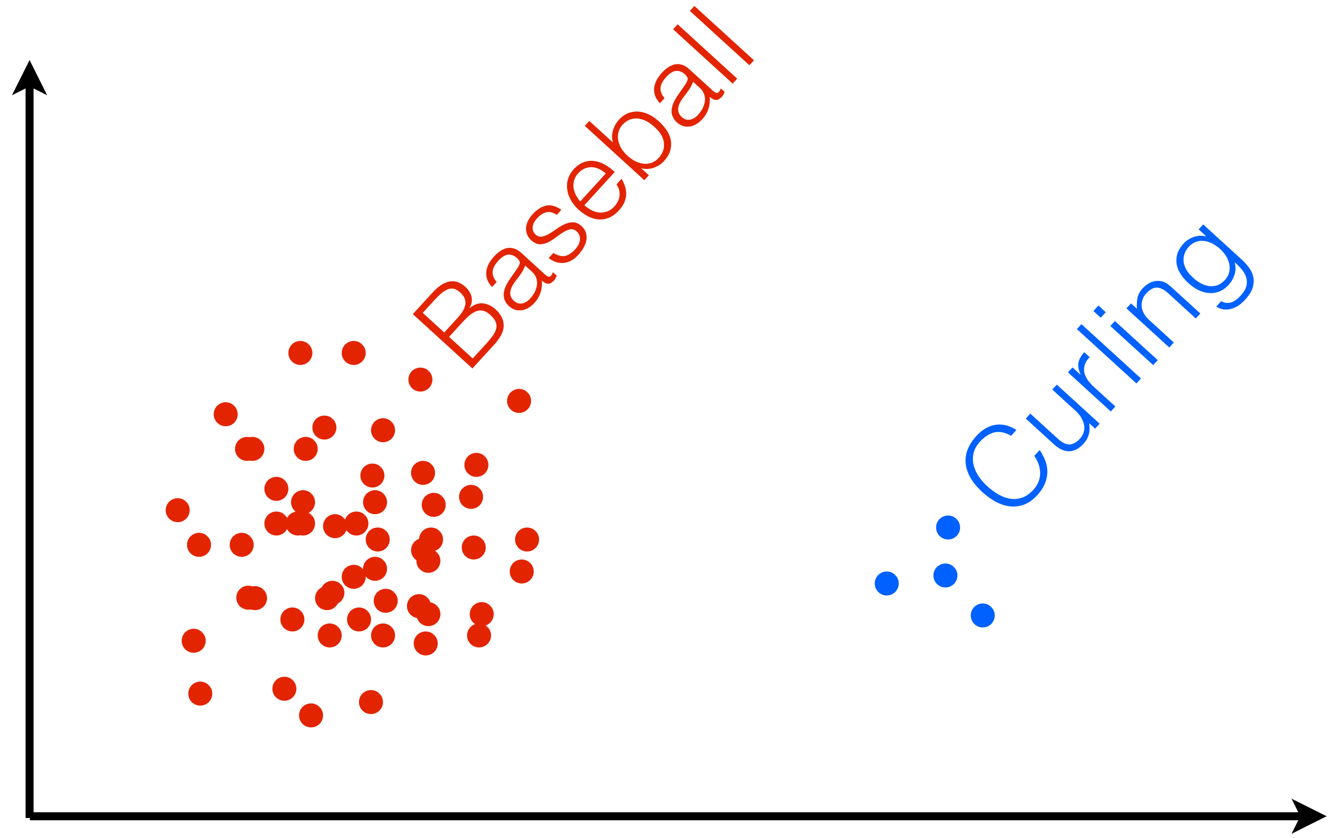
Coresets



- Pre-process data to get a smaller, weighted data set
- Theoretical guarantees on quality
- Fast algorithms; error bounds for streaming, distributed

[Agarwal et al 2005, Feldman & Langberg 2011]

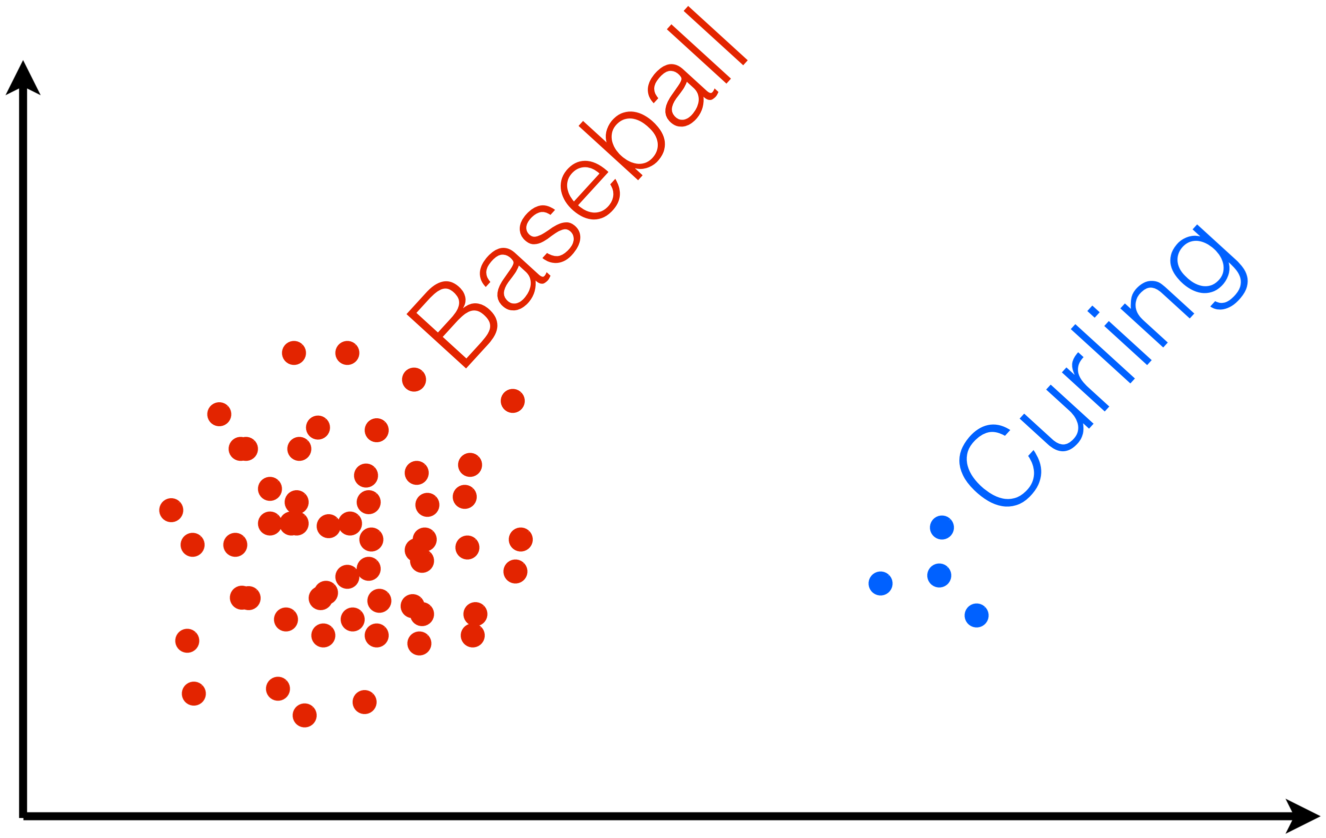
Coresets



- Pre-process data to get a smaller, weighted data set
- Theoretical guarantees on quality
- Fast algorithms; error bounds for streaming, distributed
- Cf. data squashing, big data GP ideas

[Agarwal et al 2005, Feldman & Langberg 2011, DuMouchel et al 1999, Madigan et al 1999]

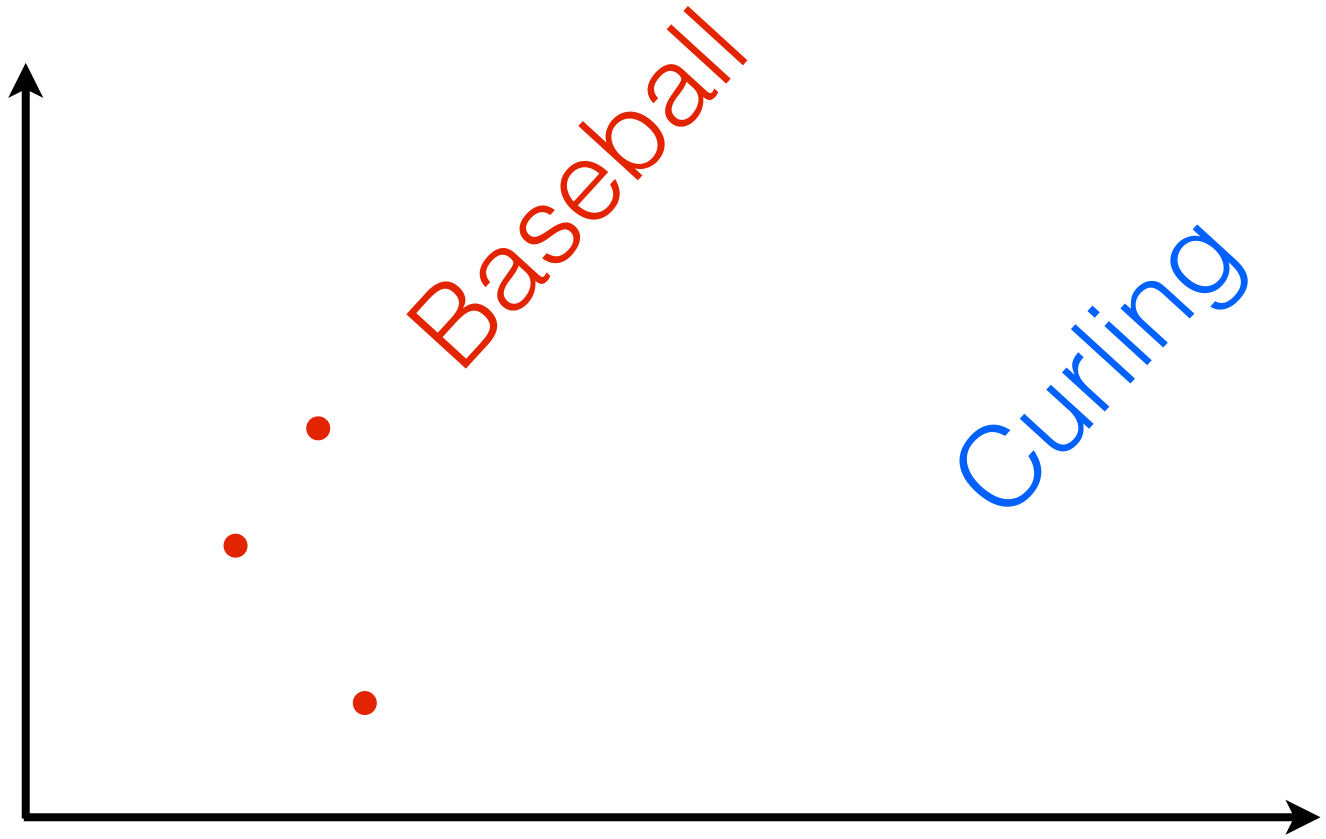
Coresets



- Pre-process data to get a smaller, weighted data set
- Theoretical guarantees on quality
- Fast algorithms; error bounds for streaming, distributed
- Cf. data squashing, big data GP ideas, subsampling

[Agarwal et al 2005, Feldman & Langberg 2011, DuMouchel et al 1999, Madigan et al 1999]

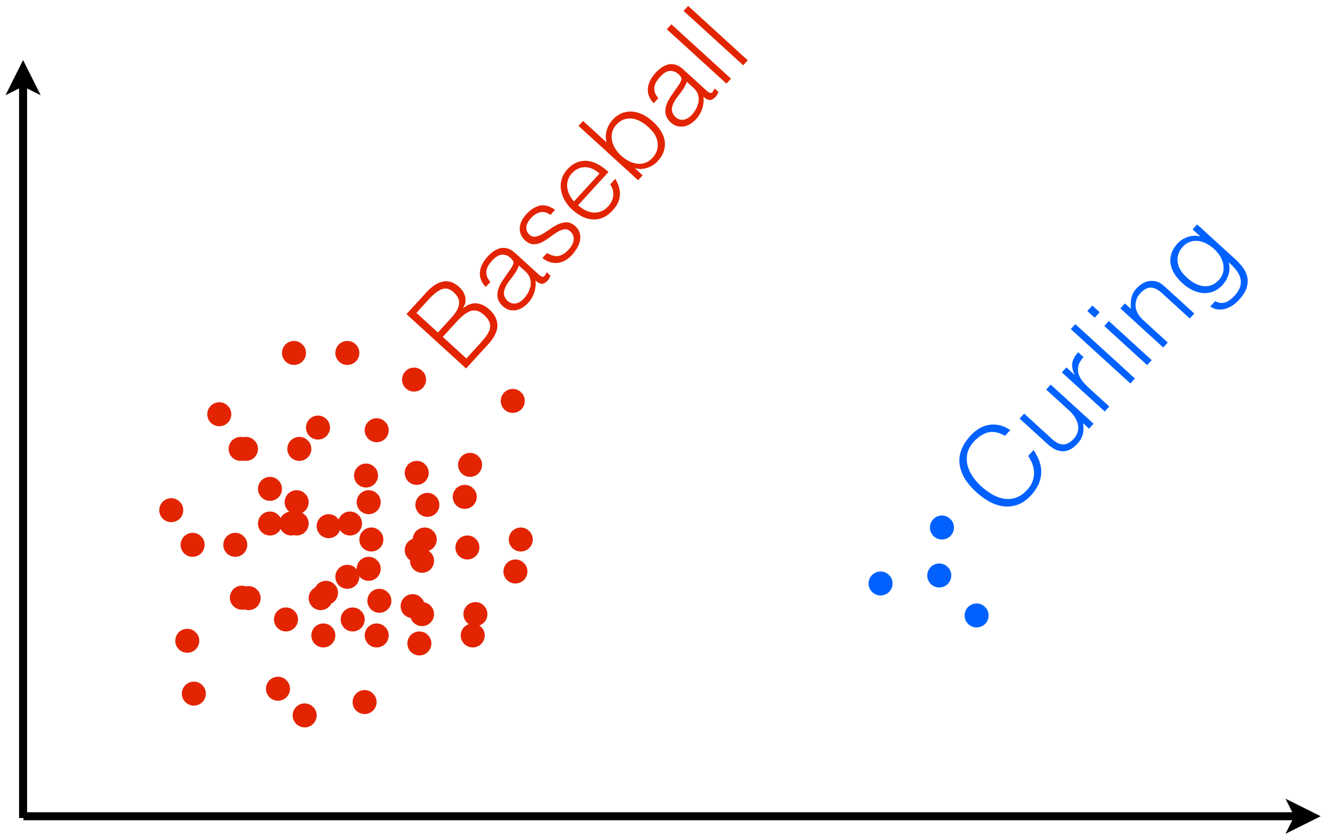
Coresets



- Pre-process data to get a smaller, weighted data set
- Theoretical guarantees on quality
- Fast algorithms; error bounds for streaming, distributed
- Cf. data squashing, big data GP ideas, subsampling

[Agarwal et al 2005, Feldman & Langberg 2011, DuMouchel et al 1999, Madigan et al 1999]

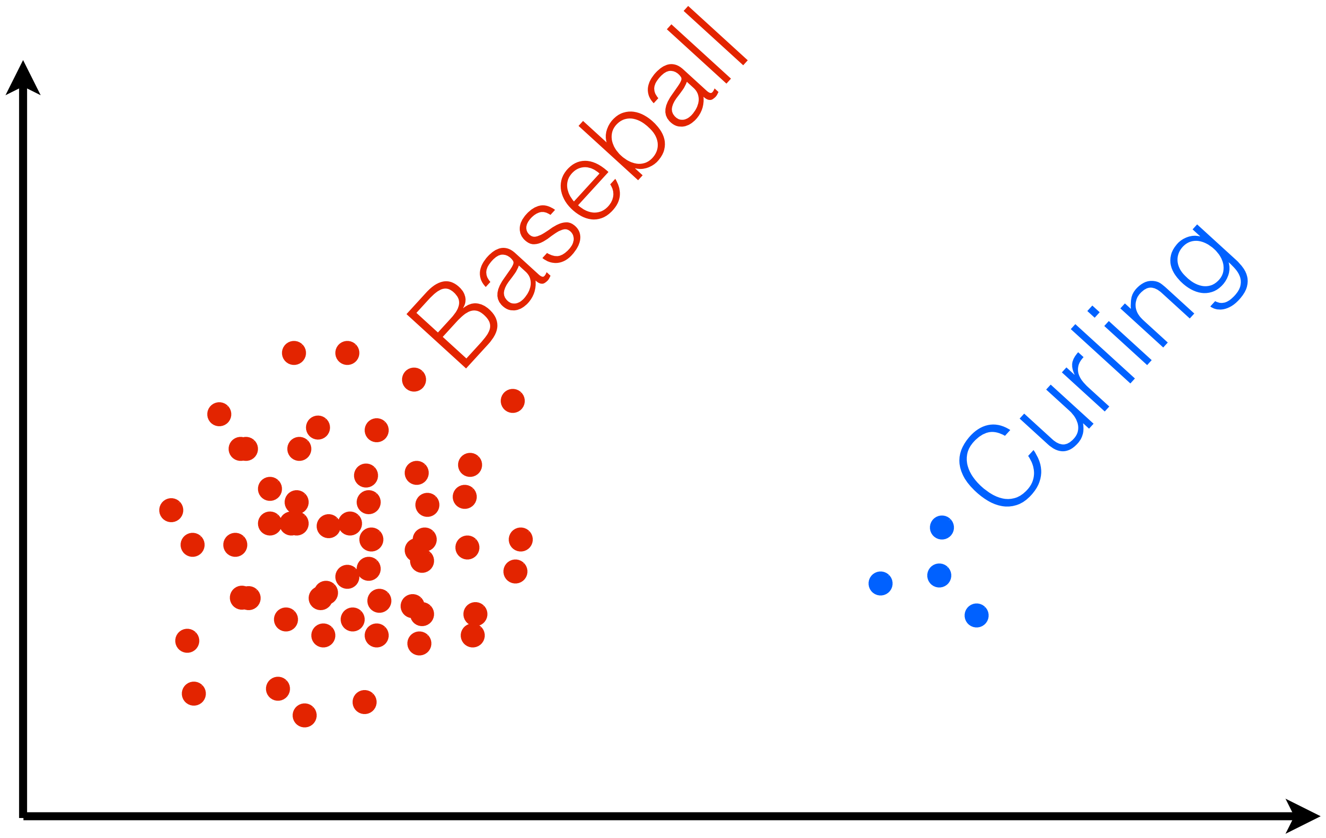
Coresets



- Pre-process data to get a smaller, weighted data set
- Theoretical guarantees on quality
- Fast algorithms; error bounds for streaming, distributed
- Cf. data squashing, big data GP ideas, subsampling

[Agarwal et al 2005, Feldman & Langberg 2011, DuMouchel et al 1999, Madigan et al 1999]

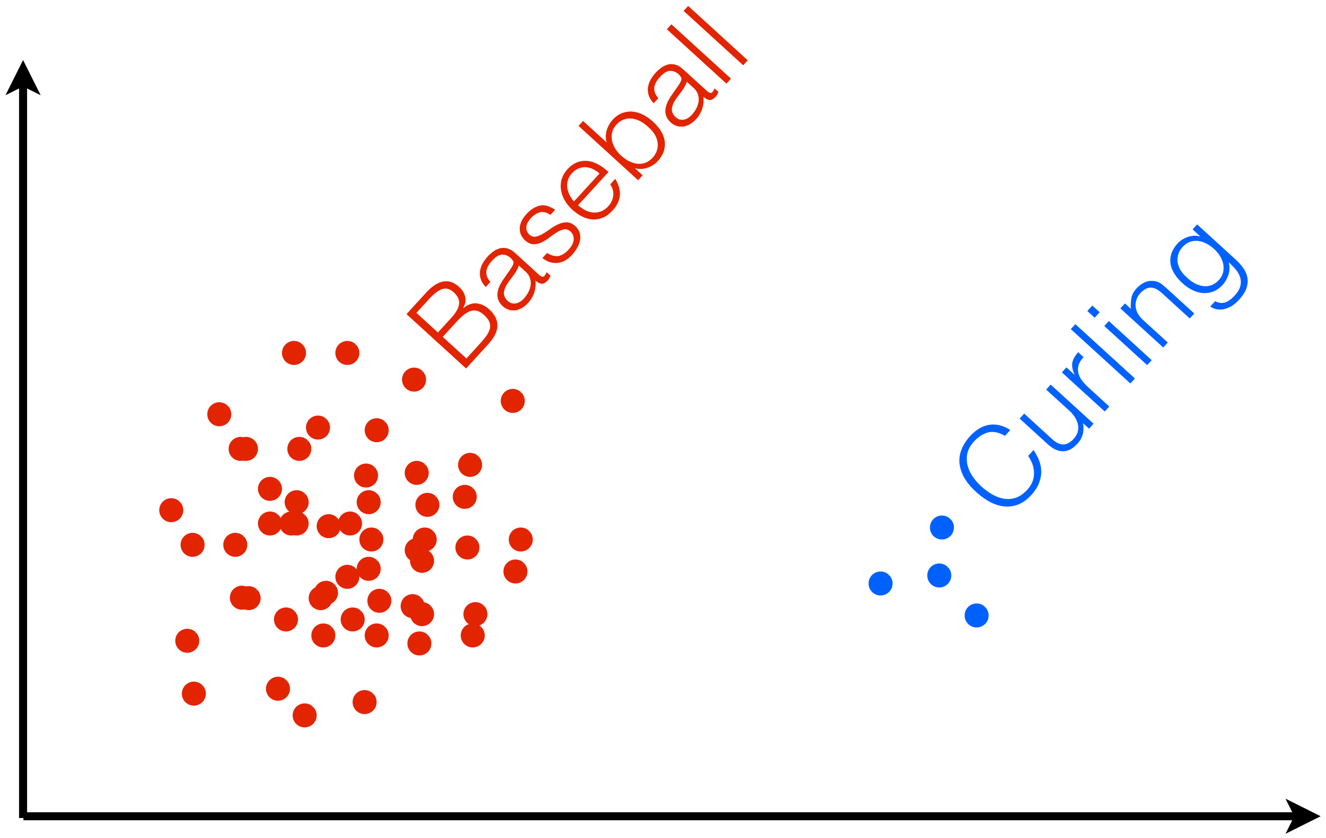
Coresets



- Pre-process data to get a smaller, weighted data set
- Theoretical guarantees on quality
- Fast algorithms; error bounds for streaming, distributed
- Cf. data squashing, big data GP ideas, subsampling
- We develop: coresets for Bayes

[Agarwal et al 2005, Feldman & Langberg 2011, DuMouchel et al 1999, Madigan et al 1999; Huggins, Campbell, Broderick 2016]

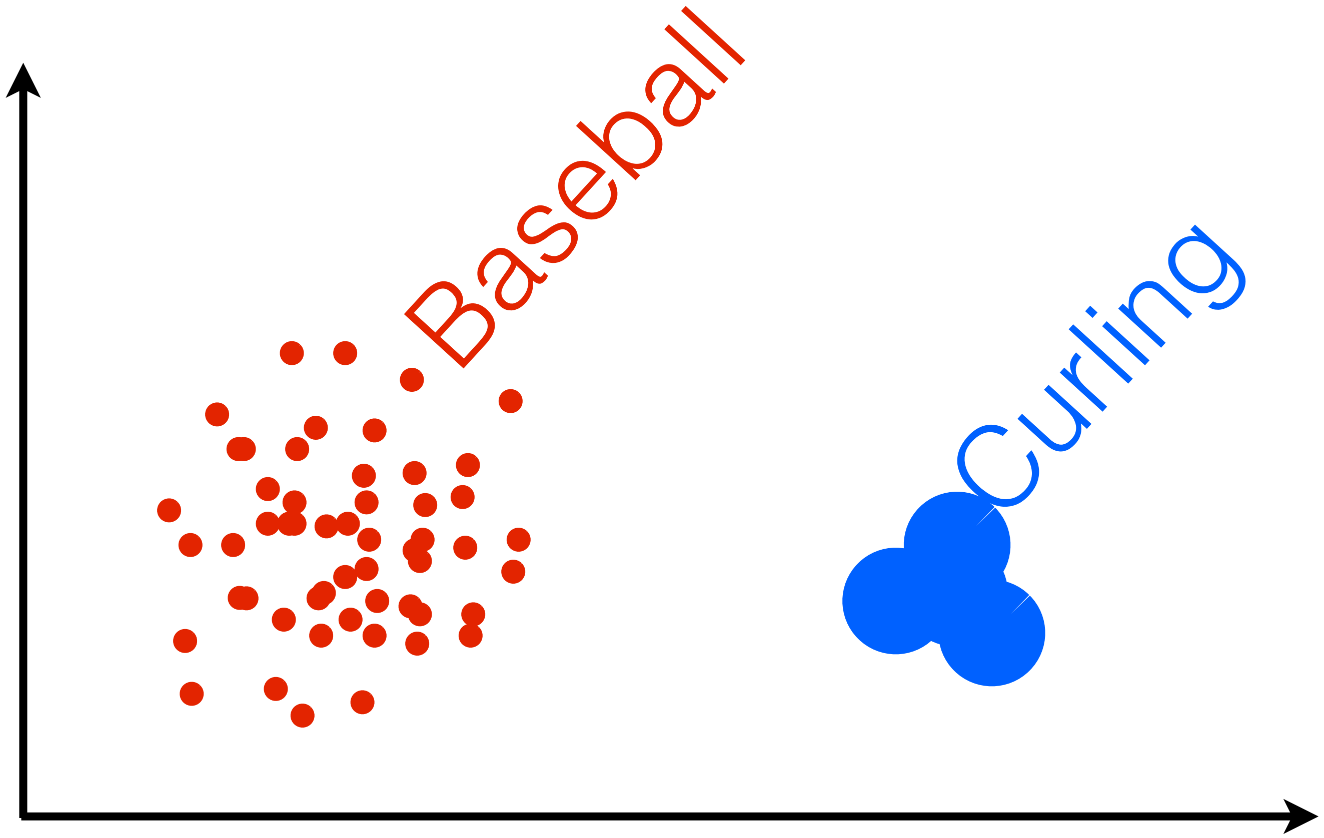
Coresets



- Pre-process data to get a smaller, weighted data set
- Theoretical guarantees on quality
- Fast algorithms; error bounds for streaming, distributed
- Cf. data squashing, big data GP ideas, subsampling
- We develop: coresets for Bayes
  - Focus on: Logistic regression

[Agarwal et al 2005, Feldman & Langberg 2011, DuMouchel et al 1999, Madigan et al 1999; Huggins, Campbell, Broderick 2016]

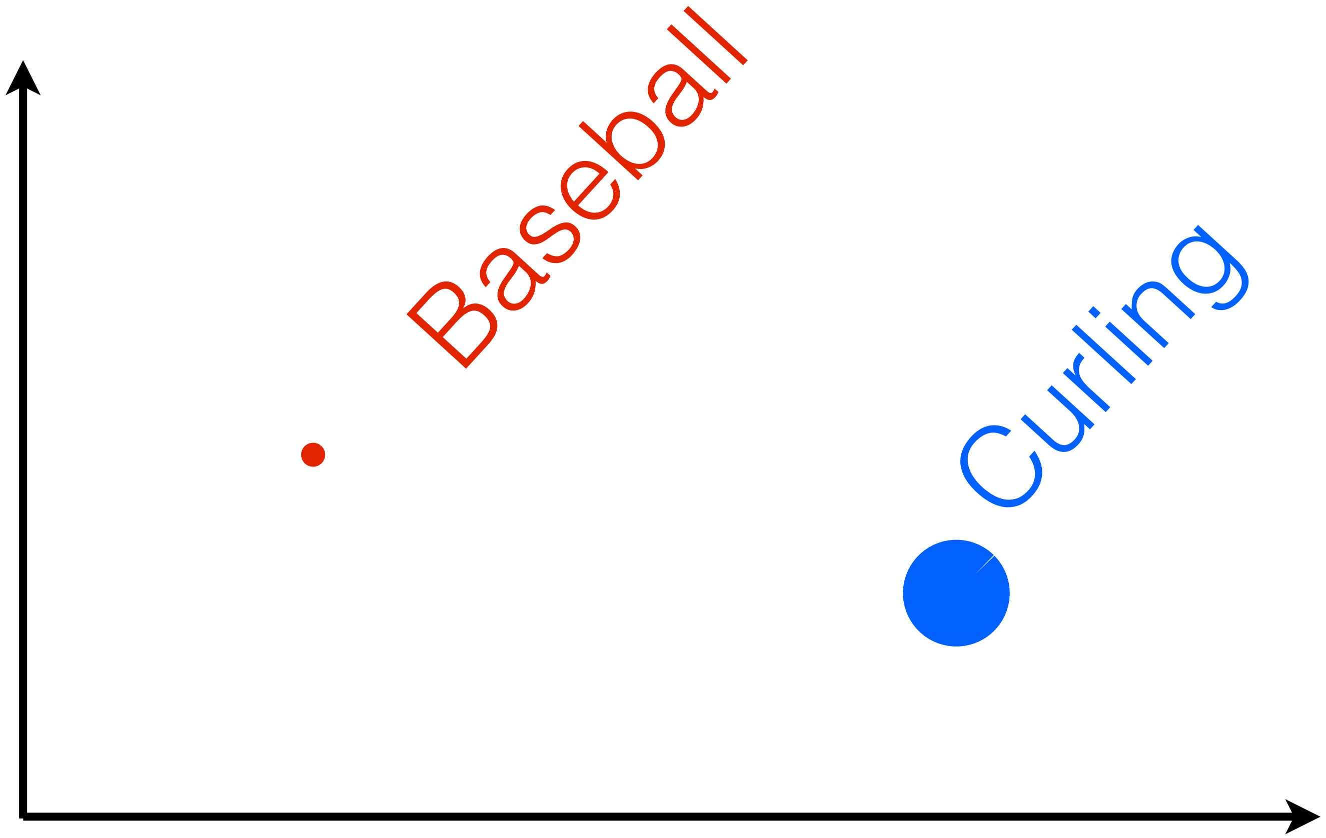
Coresets



- Pre-process data to get a smaller, weighted data set
- Theoretical guarantees on quality
- Fast algorithms; error bounds for streaming, distributed
- Cf. data squashing, big data GP ideas, subsampling
- We develop: coresets for Bayes
  - Focus on: Logistic regression

[Agarwal et al 2005, Feldman & Langberg 2011, DuMouchel et al 1999, Madigan et al 1999; Huggins, Campbell, Broderick 2016]

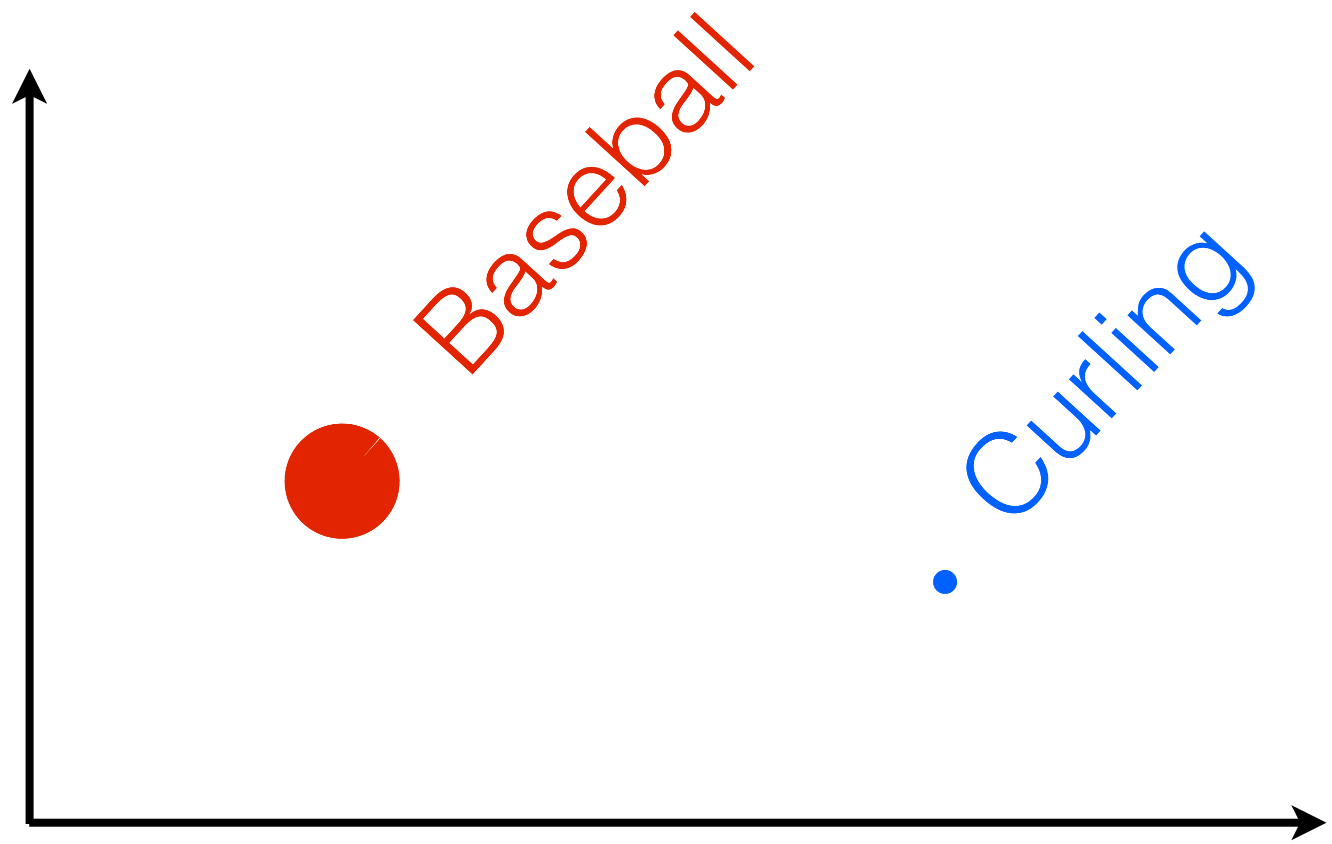
Coresets



- Pre-process data to get a smaller, weighted data set
- Theoretical guarantees on quality
- Fast algorithms; error bounds for streaming, distributed
- Cf. data squashing, big data GP ideas, subsampling
- We develop: coresets for Bayes
  - Focus on: Logistic regression

[Agarwal et al 2005, Feldman & Langberg 2011, DuMouchel et al 1999, Madigan et al 1999; Huggins, Campbell, Broderick 2016]

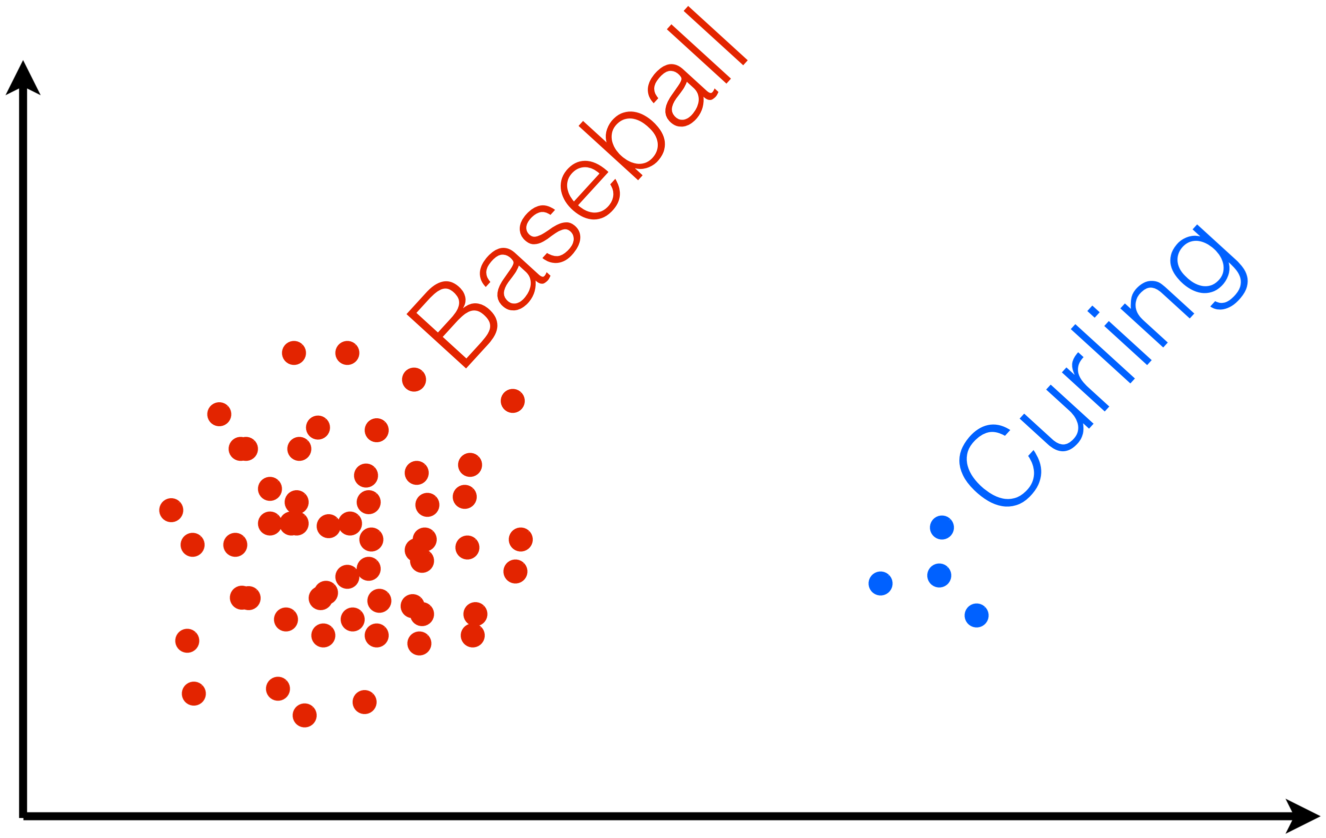
Coresets



- Pre-process data to get a smaller, weighted data set
- Theoretical guarantees on quality
- Fast algorithms; error bounds for streaming, distributed
- Cf. data squashing, big data GP ideas, subsampling
- We develop: coresets for Bayes
  - Focus on: Logistic regression

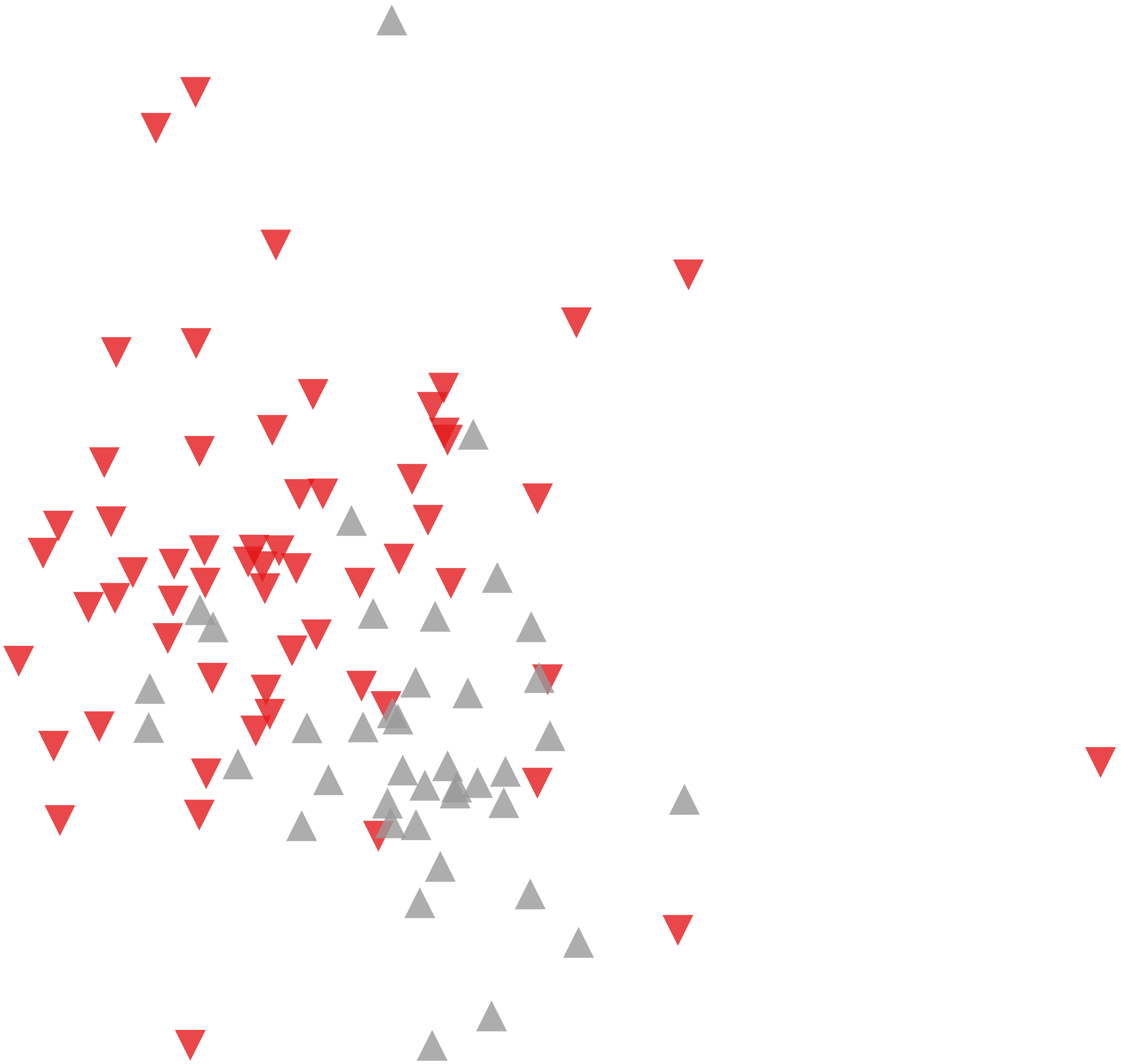
[Agarwal et al 2005, Feldman & Langberg 2011, DuMouchel et al 1999, Madigan et al 1999; Huggins, Campbell, Broderick 2016]

Coresets



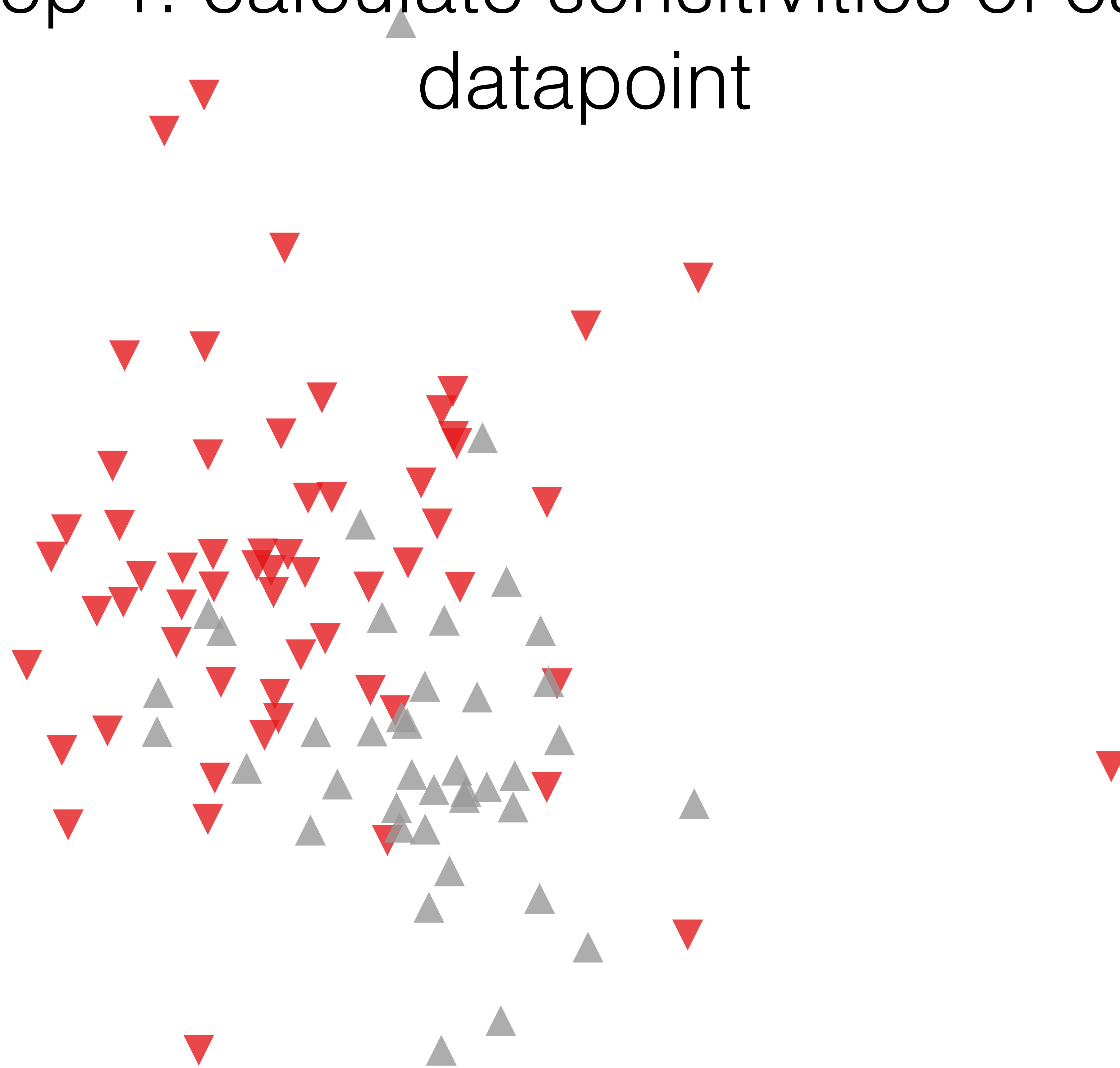
- Pre-process data to get a smaller, weighted data set
- Theoretical guarantees on quality
- Fast algorithms; error bounds for streaming, distributed
- Cf. data squashing, big data GP ideas, subsampling
- We develop: coresets for Bayes
  - Focus on: Logistic regression

[Agarwal et al 2005, Feldman & Langberg 2011, DuMouchel et al 1999, Madigan et al 1999; Huggins, Campbell, Broderick 2016]



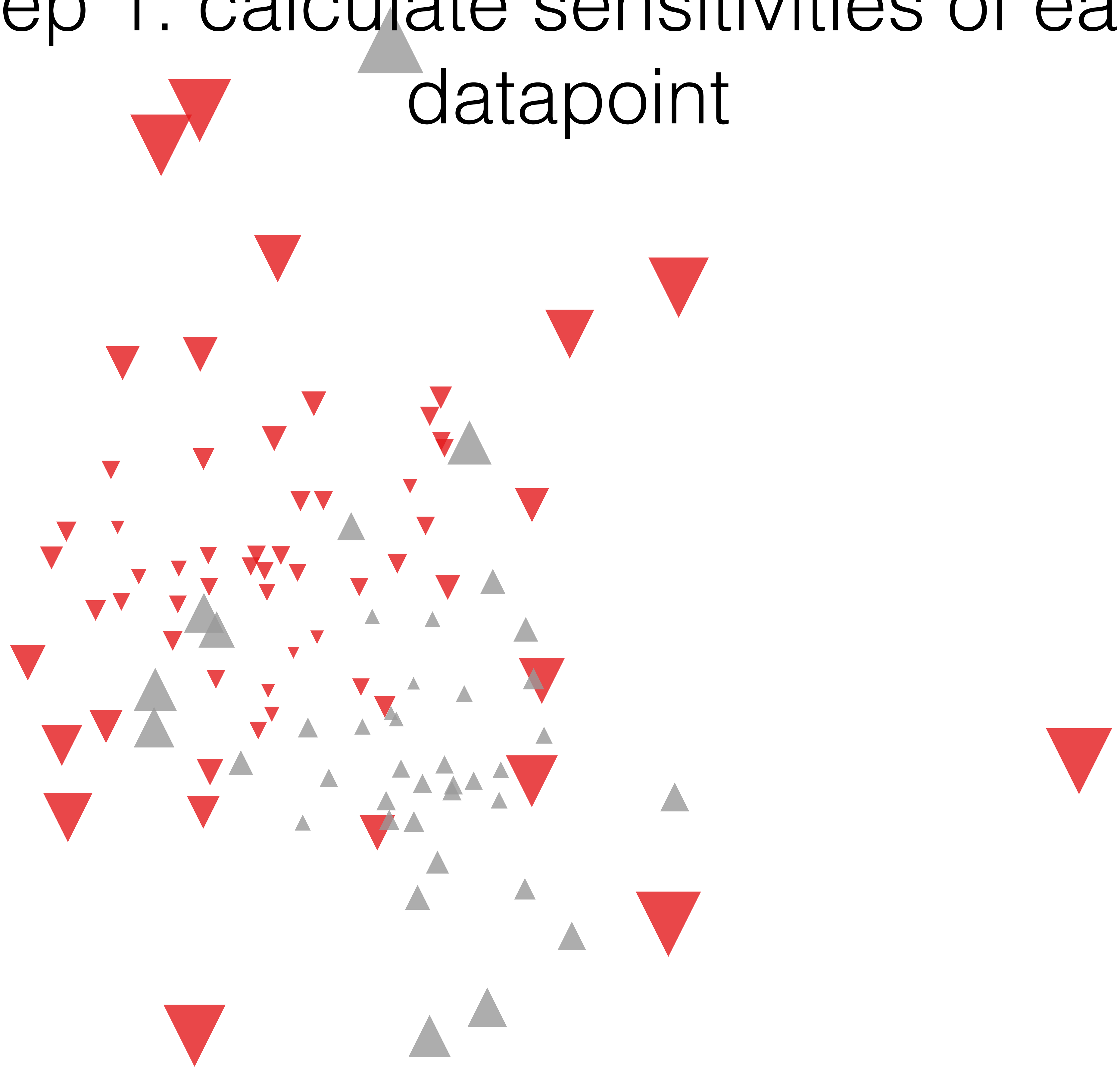
[Huggins, Campbell, Broderick 2016]

Step 1: calculate sensitivities of each datapoint



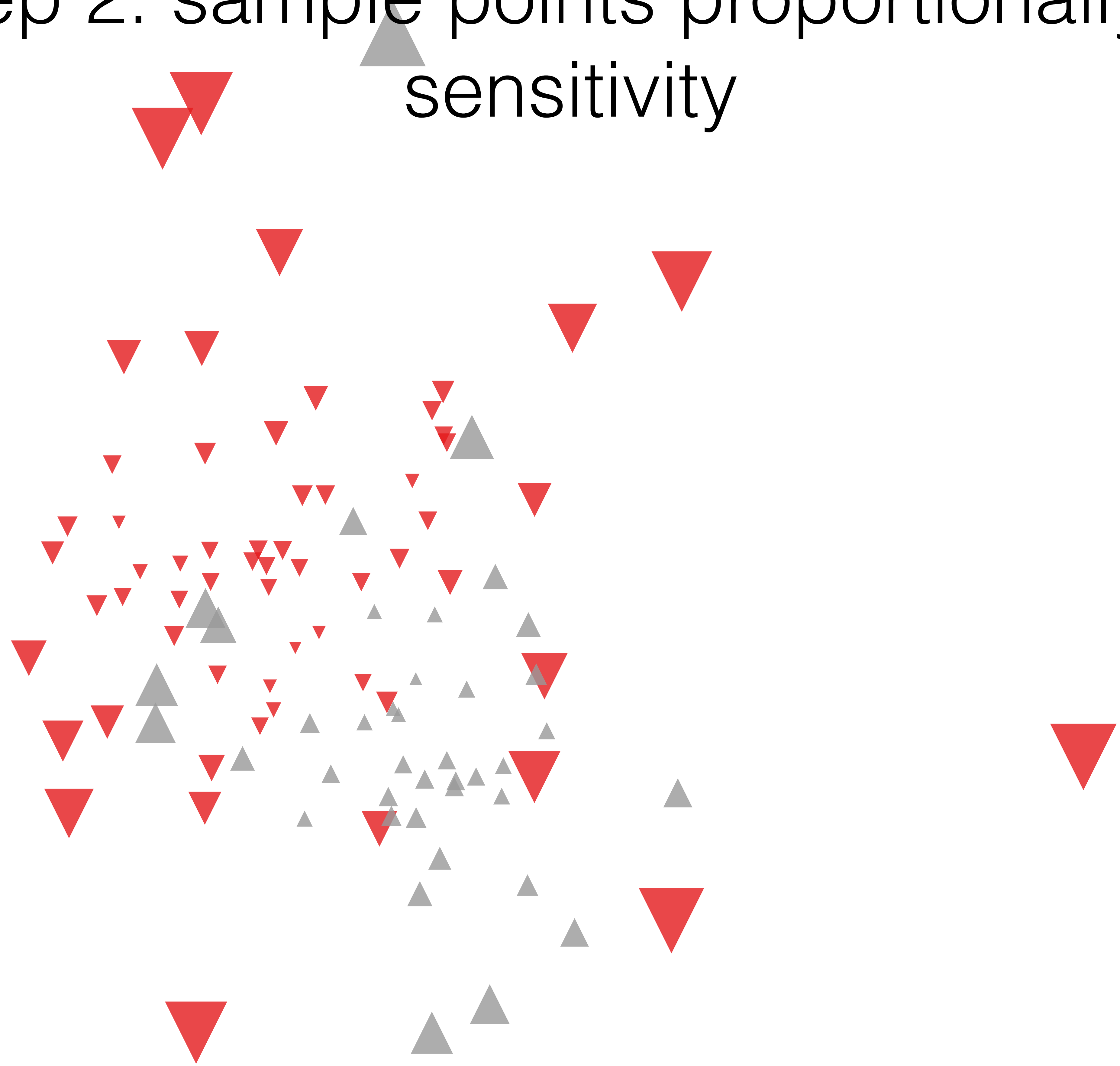
[Huggins, Campbell, Broderick 2016]

Step 1: calculate sensitivities of each datapoint



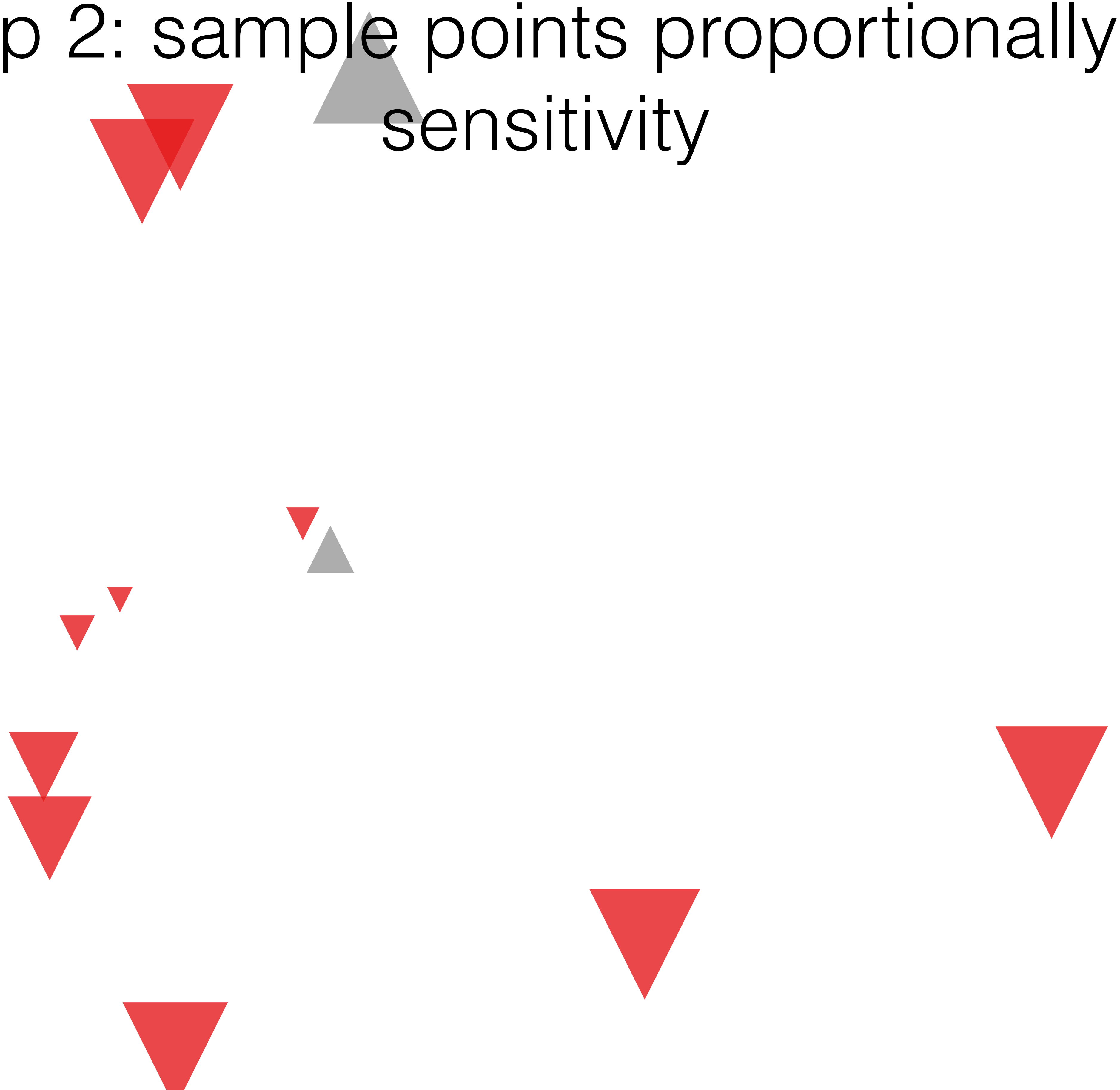
[Huggins, Campbell, Broderick 2016]

Step 2: sample points proportionally to sensitivity



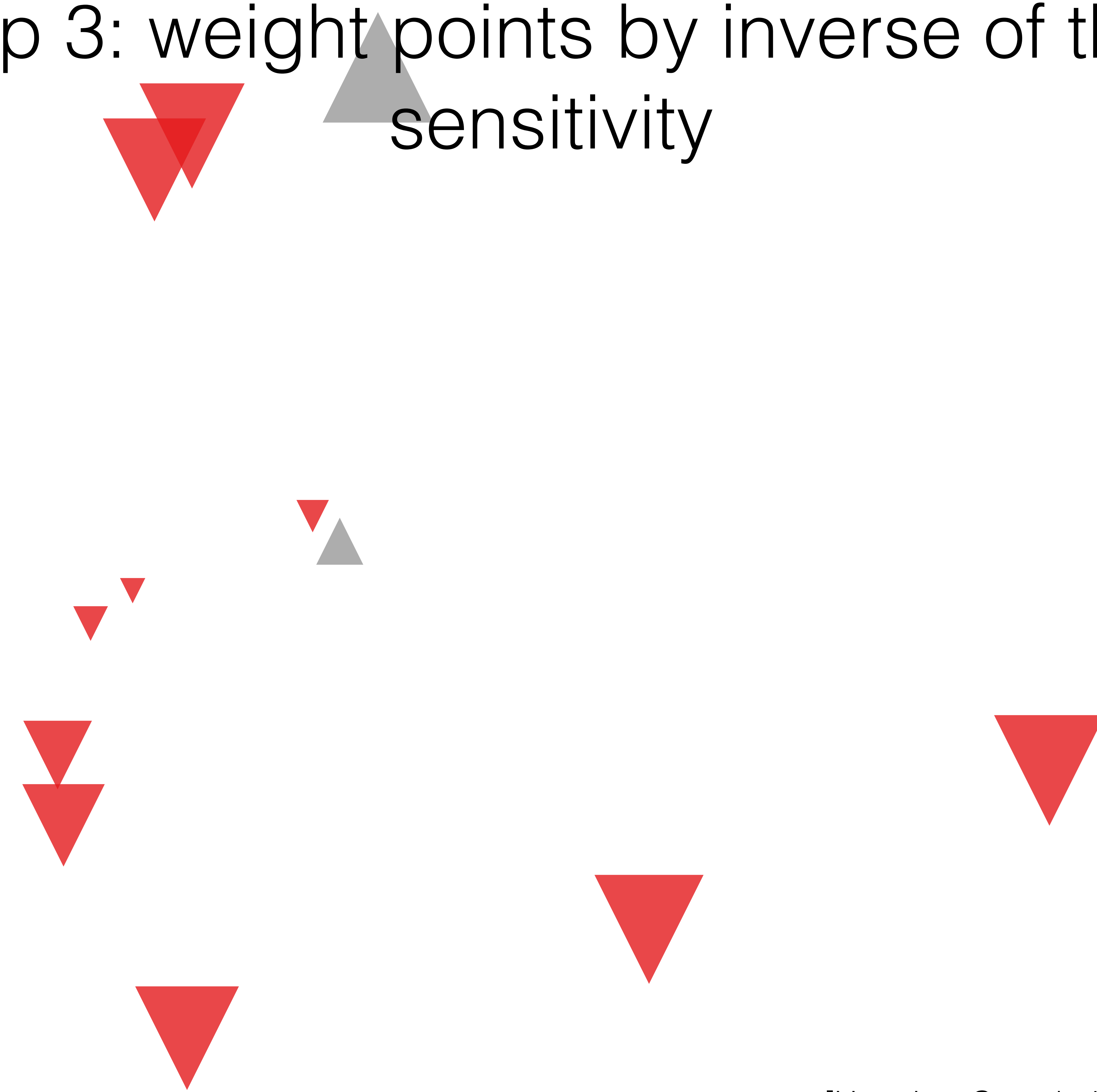
[Huggins, Campbell, Broderick 2016]

Step 2: sample points proportionally to sensitivity



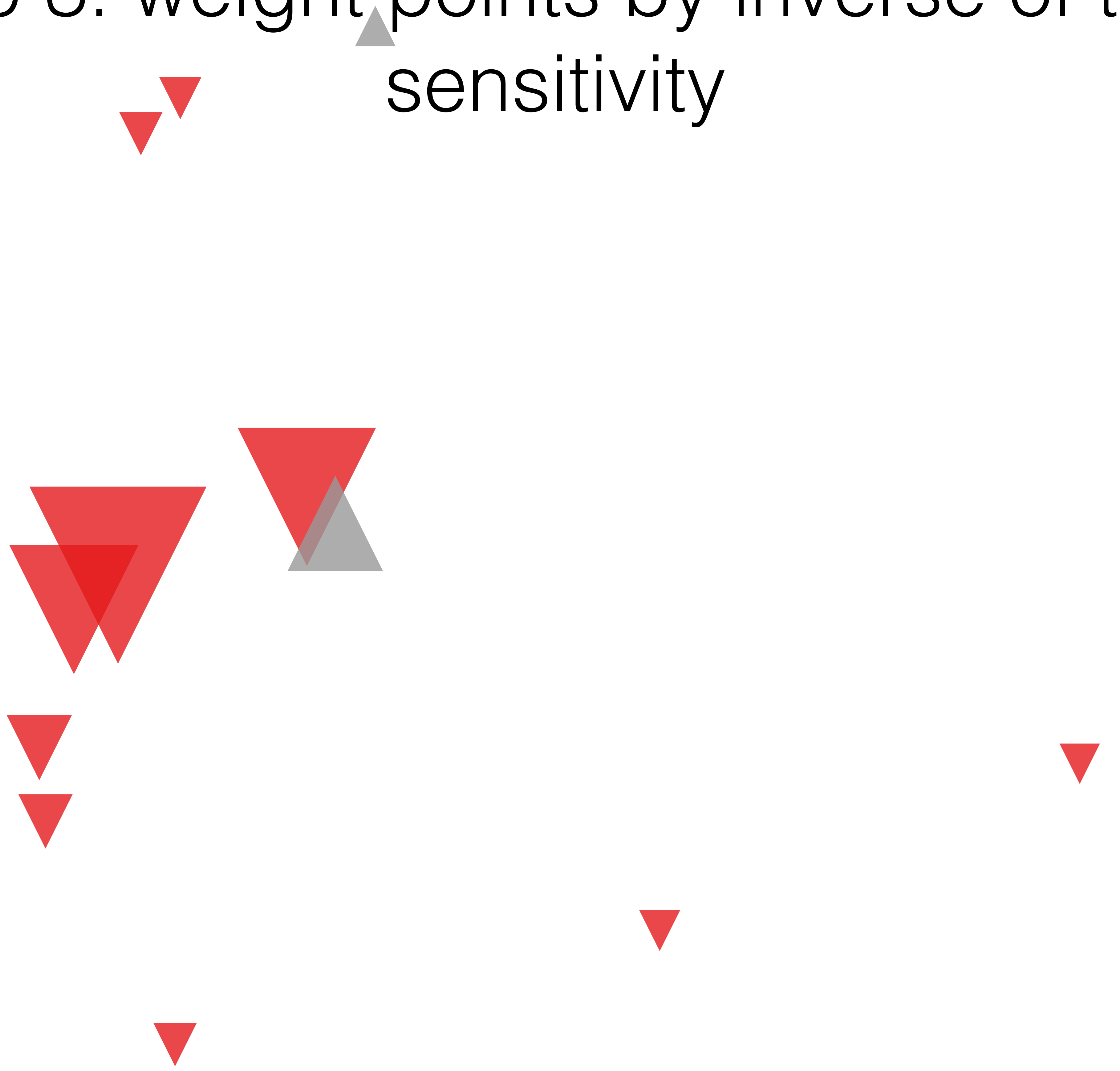
[Huggins, Campbell, Broderick 2016]

Step 3: weight points by inverse of their sensitivity



[Huggins, Campbell, Broderick 2016]

Step 3: weight points by inverse of their sensitivity

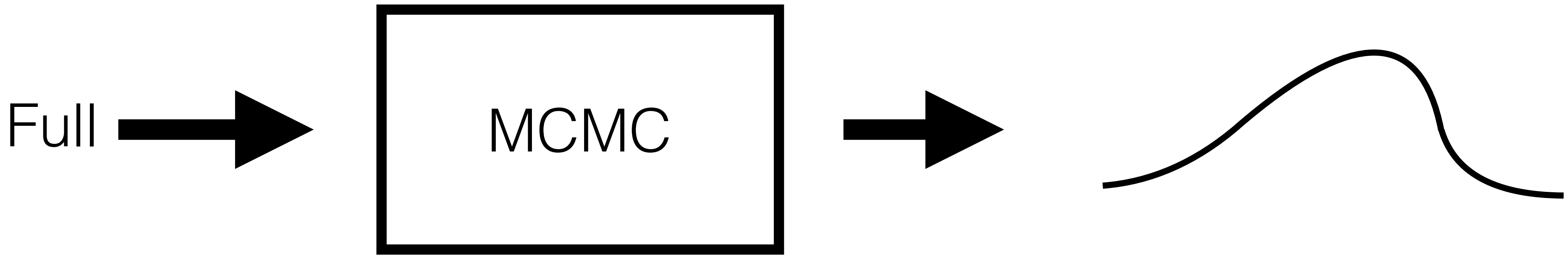


[Huggins, Campbell, Broderick 2016]

Step 4: input weighted points to existing approximate posterior algorithm

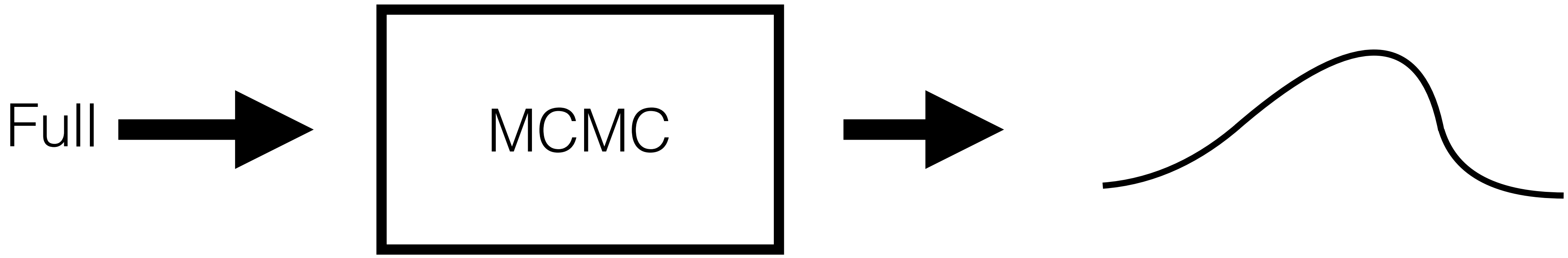
[Huggins, Campbell, Broderick 2016]

Step 4: input weighted points to existing approximate posterior algorithm



[Huggins, Campbell, Broderick 2016]

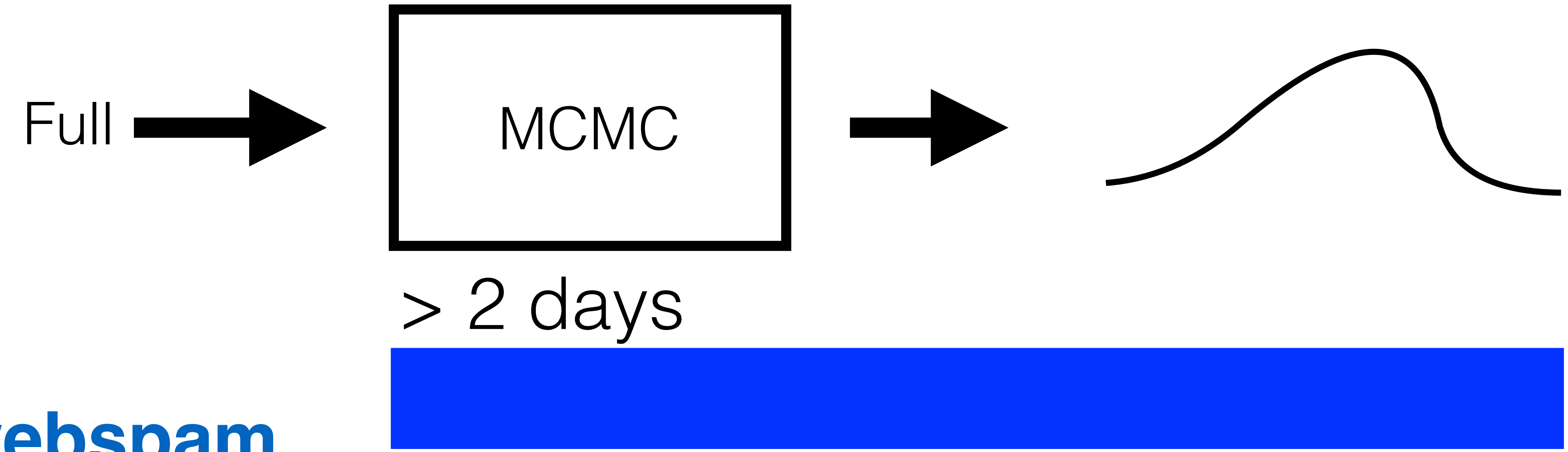
Step 4: input weighted points to existing approximate posterior algorithm



webspam
350K points
127 features

[Huggins, Campbell, Broderick 2016]

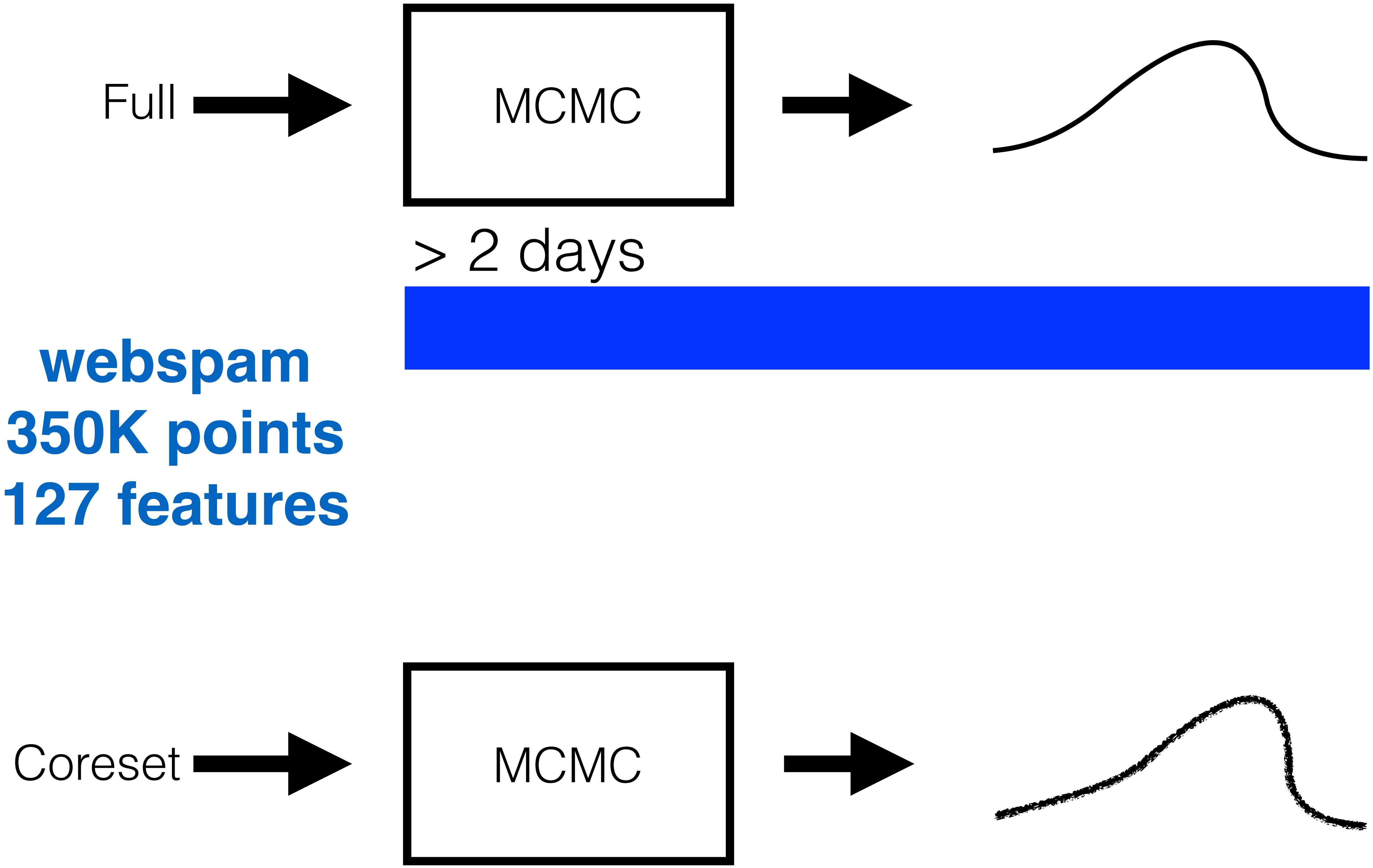
Step 4: input weighted points to existing approximate posterior algorithm



webspam
350K points
127 features

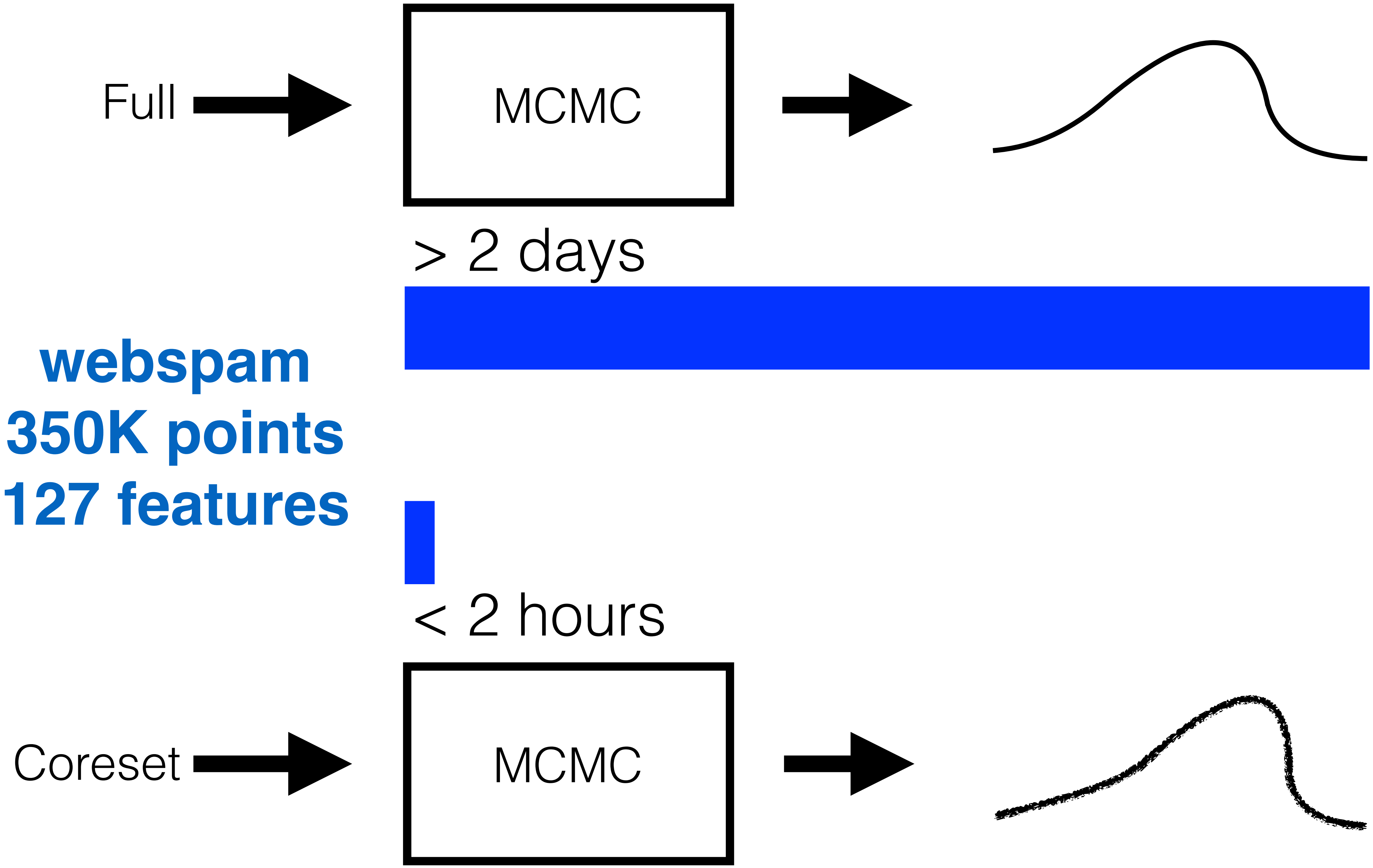
[Huggins, Campbell, Broderick 2016]

Step 4: input weighted points to existing approximate posterior algorithm



[Huggins, Campbell, Broderick 2016]

Step 4: input weighted points to existing approximate posterior algorithm



[Huggins, Campbell, Broderick 2016]

Theory

[Huggins, Campbell, Broderick 2016]

Theory

- Finite-data theoretical guarantee

Thm sketch (HCB). Choose $\varepsilon > 0$, $\delta \in (0,1)$. Our algorithm runs in $O(N)$ time and creates $\text{coreset-size} \sim \text{const} \cdot \epsilon^{-2} + \log(1/\delta)$

W.p. 1 - δ, it constructs a coreset with $\left|\ln \mathcal{E} - \ln \tilde{\mathcal{E}}\right| \leq \epsilon \left|\ln \mathcal{E}\right|$

[Huggins, Campbell, Broderick 2016]

Theory

- Finite-data theoretical guarantee
  - On the log evidence (vs. posterior mean, uncertainty, etc)

Thm sketch (HCB). Choose ε > 0, δ ∈ (0,1). Our algorithm runs in *O*(*N*) time and creates $\text{coreset-size} \sim \text{const} \cdot \epsilon^{-2} + \log(1/\delta)$

W.p. 1 - δ, it constructs a coreset with $\left|\ln \mathcal{E} - \ln \tilde{\mathcal{E}}\right| \leq \epsilon \left|\ln \mathcal{E}\right|$

[Huggins, Campbell, Broderick 2016]

Theory

- Finite-data theoretical guarantee
  - On the log evidence (vs. posterior mean, uncertainty, etc)

Thm sketch (HCB). Choose ε > 0, δ ∈ (0,1). Our algorithm runs in *O*(*N*) time and creates $\text{coreset-size} \sim \text{const} \cdot \epsilon^{-2} + \log(1/\delta)$

W.p. 1 - δ, it constructs a coreset with $\left|\ln \mathcal{E} - \ln \tilde{\mathcal{E}}\right| \leq \epsilon \left|\ln \mathcal{E}\right|$

- Can quantify the propagation of error in streaming and parallel settings

1. If D_i' is an ε-coreset for D_i, then $D_1' \cup D_2'$ is an ε-coreset for $D_1 \cup D_2$.
2. If D' is an ε-coreset for D and D'' is an ε'-coreset for D', then D'' is an ε''-coreset for D, where ε'' = (1 + ε)(1 + ε') − 1.

[Huggins, Campbell, Broderick 2016]

06/18/15

Criteo Releases Industry's Largest-Ever Dataset for Machine Learning to Academic Community

Over one terabyte of data released to help researchers benchmark distributed learning algorithms in critical research

[Huggins, Adams, Broderick, submitted]

06/18/15

Criteo Releases Industry's Largest-Ever Dataset for Machine Learning to Academic Community

Over one terabyte of data released to help researchers benchmark distributed learning algorithms in critical research

- Subset yields 6M data points, 1K features

[Huggins, Adams, Broderick, submitted]

Polynomial approximate sufficient statistics

06/18/15

Criteo Releases Industry's Largest-Ever Dataset for Machine Learning to Academic Community

Over one terabyte of data released to help researchers benchmark distributed learning algorithms in critical research

- Subset yields 6M data points, 1K features

[Huggins, Adams, Broderick, submitted]

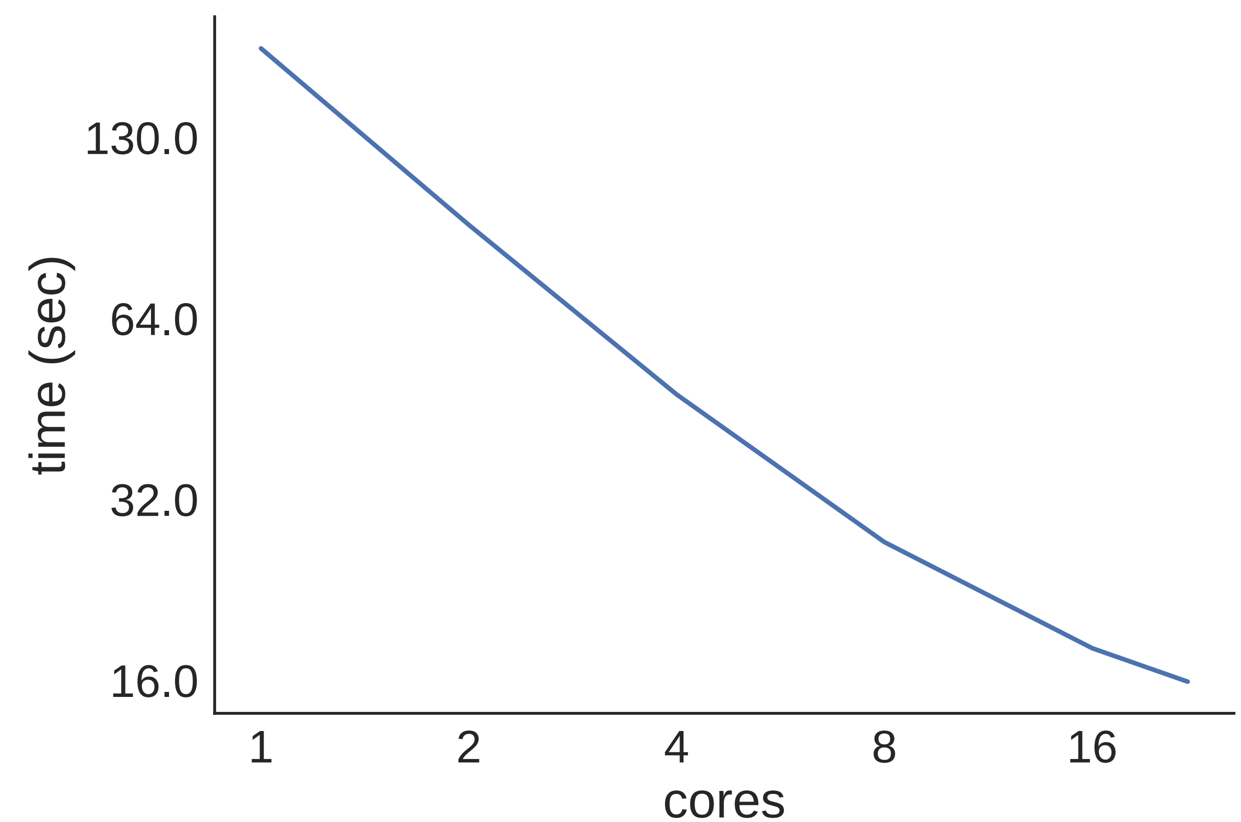
Polynomial approximate sufficient statistics

06/18/15

Criteo Releases Industry's Largest-Ever Dataset for Machine Learning to Academic Community

Over one terabyte of data released to help researchers benchmark distributed learning algorithms in critical research

- Subset yields 6M data points, 1K features



[Huggins, Adams, Broderick, submitted]

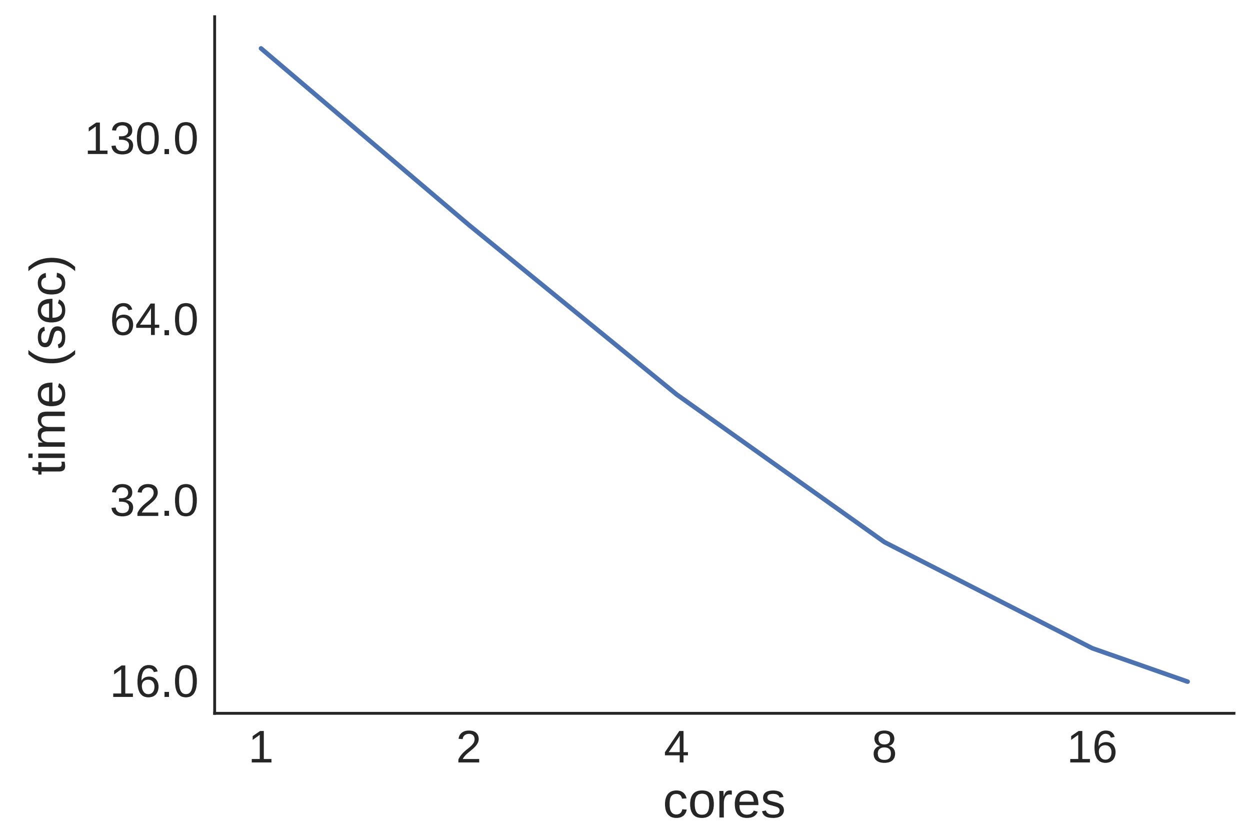
Polynomial approximate sufficient statistics


06/18/15

Criteo Releases Industry's Largest-Ever Dataset for Machine Learning to Academic Community

Over one terabyte of data released to help researchers benchmark distributed learning algorithms in critical research

- Subset yields 6M data points, 1K features
- Streaming, distributed; minimal communication



[Huggins, Adams, Broderick, submitted]

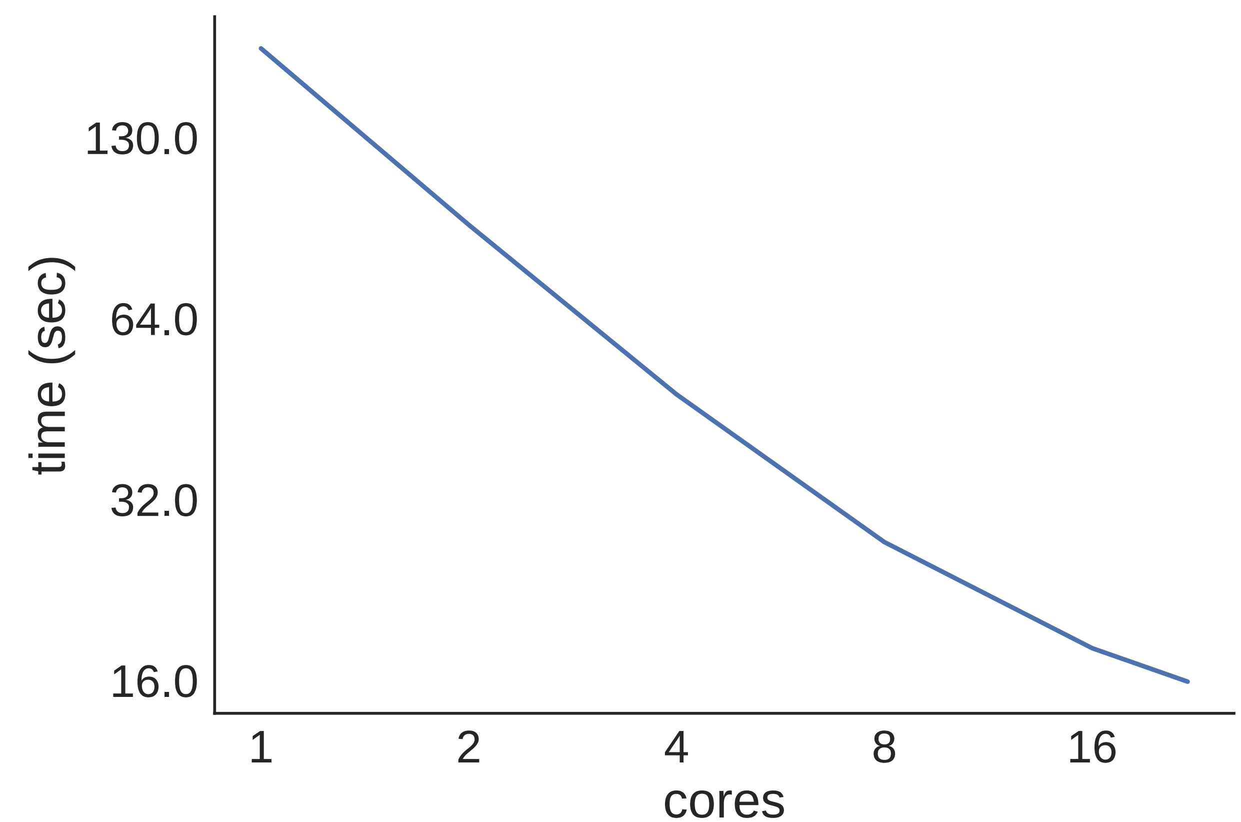
Polynomial approximate sufficient statistics

06/18/15

Criteo Releases Industry's Largest-Ever Dataset for Machine Learning to Academic Community

Over one terabyte of data released to help researchers benchmark distributed learning algorithms in critical research

- Subset yields 6M data points, 1K features
- Streaming, distributed; minimal communication
- 24 cores, <20 sec



[Huggins, Adams, Broderick, submitted]

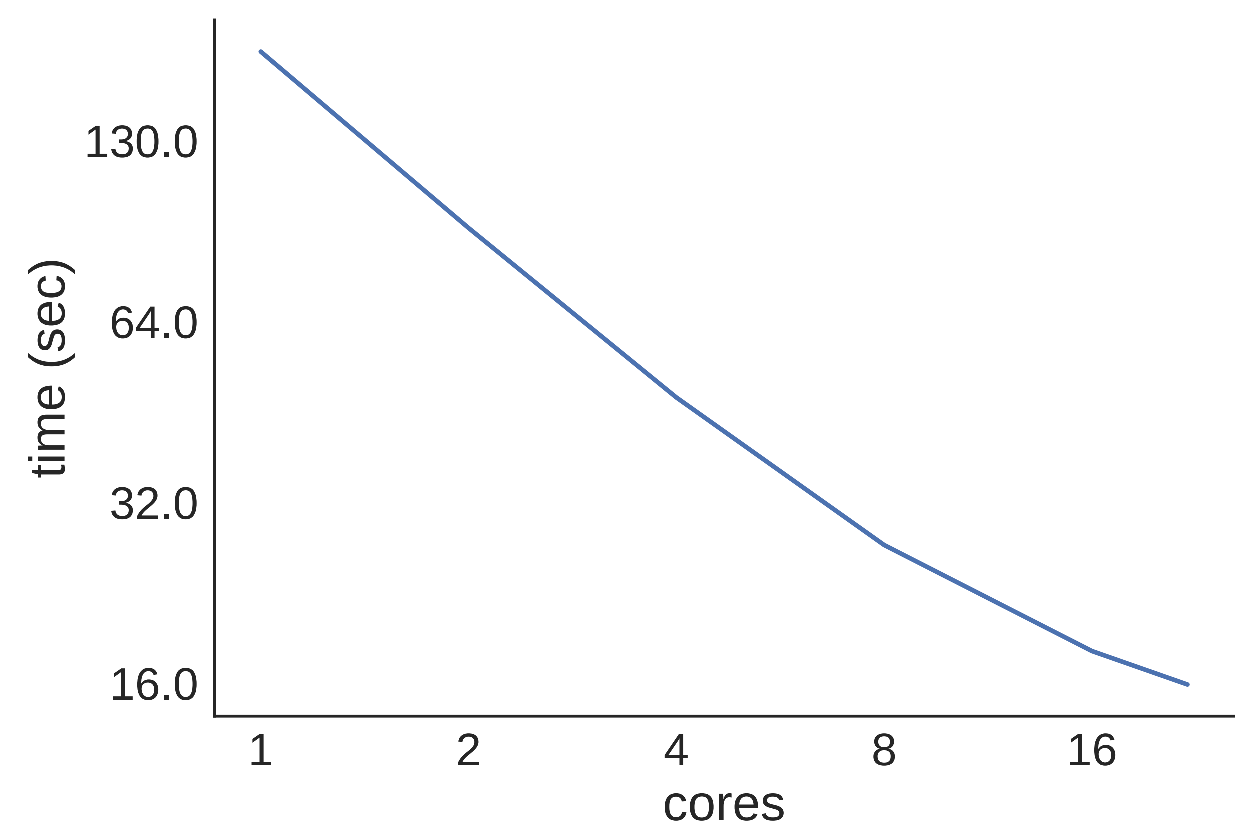
Polynomial approximate sufficient statistics

06/18/15

Criteo Releases Industry's Largest-Ever Dataset for Machine Learning to Academic Community

Over one terabyte of data released to help researchers benchmark distributed learning algorithms in critical research

- Subset yields 6M data points, 1K features
- Streaming, distributed; minimal communication
- 24 cores, <20 sec
- Bounds on Wasserstein



[Huggins, Adams, Broderick, submitted]

Conclusions

- **Reliable** Bayesian inference at scale via data summarization
  - Coresets, polynomial approximate sufficient statistics
  - Streaming, distributed
- Challenges and opportunities:
  - Beyond logistic regression
  - Generalized linear models; deep models; high-dimensional models

[Huggins, Campbell, Broderick 2016; Huggins, Adams, Broderick, submitted; Bardenet, Maillard 2015; Geppert, Ickstadt, Munteanu, Quedenfeld, Sohler 2017; Ahfock, Astle, Richardson 2017]

References

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

T Campbell*, JH Huggins*, J How, and T Broderick. Truncated random measures. Submitted. ArXiv:1603.00861.

R 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, JH Huggins, and MI Jordan. Fast robustness quantification with variational Bayes. *ICML Workshop on #Data4Good: Machine Learning in Social Good Applications*, 2016. ArXiv:1606.07153.

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

References

- PK Agarwal, S Har-Peled, and KR Varadarajan. Geometric approximation via coresets. *Combinatorial and computational geometry*, 2005.
- D Ahfock, WJ Astle, S Richardson. Statistical properties of sketching algorithms. arXiv 1706.03665.
- R Bardenet, A Doucet, and C Holmes. On Markov chain Monte Carlo methods for tall data. arXiv, 2015.
- R Bardenet, O-A Maillard, A note on replacing uniform subsampling by random projections in MCMC for linear regression of tall datasets. Preprint.
- CM Bishop. *Pattern Recognition and Machine Learning*, 2006.
- W DuMouchel, C Volinsky, T Johnson, C Cortes, D Pregibon. Squashing flat files flatter. *SIGKDD* 1999.
- D Dunson. Robust and scalable approach to Bayesian inference. Talk at *ISBA* 2014.
- D Feldman and . Langberg. A unified framework for approximating and clustering data. *Symposium on Theory of Computing*, 2011.
- B Fosdick. *Modeling Heterogeneity within and between Matrices and Arrays*, Chapter 4.7. PhD Thesis, University of Washington, 2013.
- LN Geppert, K Ickstadt, A Munteanu, J Quedenfeld, C Sohler. Random projections for Bayesian regression. *Statistics and Computing*, 2017.
- DJC MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.
- D Madigan, N Raghavan, W Dumouchel, M Nason, C Posse, G Ridgeway. Likelihood-based data squashing: A modeling approach to instance construction. *Data Mining and Knowledge Discovery*, 2002.
- M Opper and O Winther. Variational linear response. *NIPS* 2003.
- 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.