Big Data, Statistics, and the Internet

Steven L. Scott

Google

April 1, 2014

Summary

- Big data live on more than one machine.
- Computing takes place in the MapReduce / Hadoop paradigm, where communicating is expensive.
- Methods that treat "the data" as a single object don't work. We need to treat MapReduce as a basic assumption.

Outline of the talk

Big data, Statistics, and the Internet

- "Big data" are real, and sometimes necessary
- Bayes is important
- Experiments in the Internet age
- MapReduce: A case study in how to do it wrong

Consensus Monte Carlo

- Examples
  - Binomial
  - Logistic regression
  - Hierarchical models
  - BART

Conclusion

Dan Ariely

January 2013 Facebook post

Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...

Big data, machine learning, and statistics

Statistics

Models summarize data so humans can better understand it.

- Does spending more money on ads make my business more profitable?
- Is treatment A better than treatment B?

Sampling/aggregation works just fine.
(Humans can't handle the complexity anyway.)

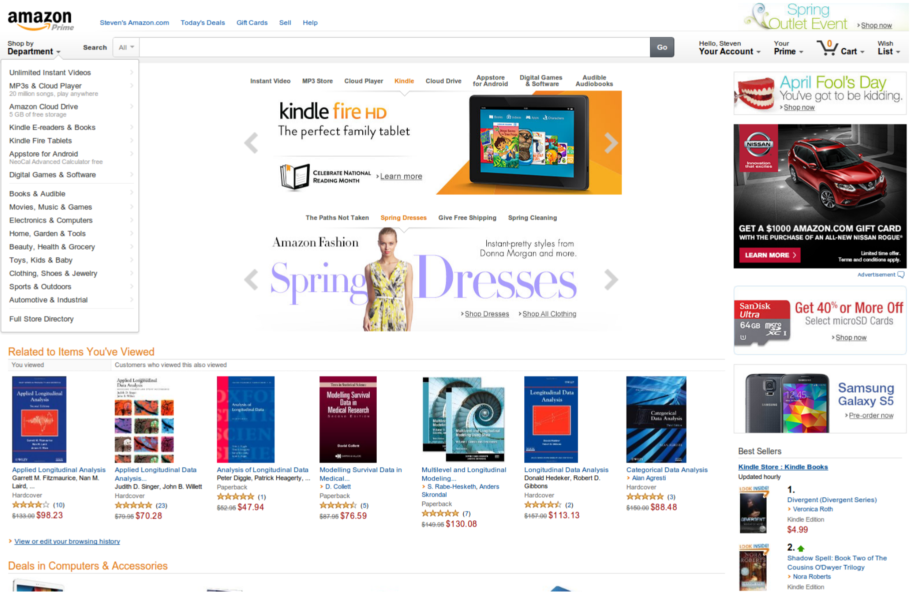
Machine learning

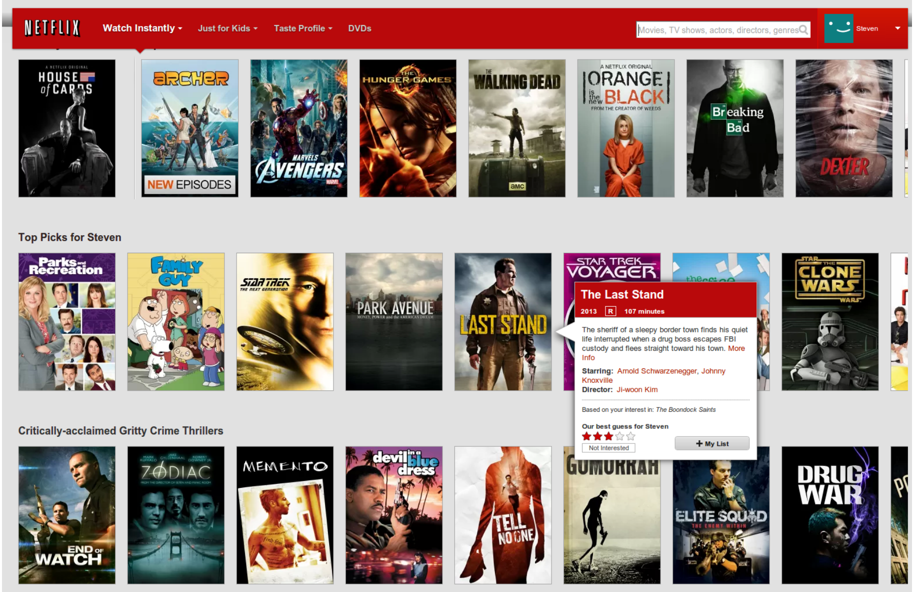
Models allow machines to make decisions.

- Google search, Amazon product recommendations, Facebook news feed, ...

Need big data to match diverse users to diverse content.

The canonical problem in Silicon Valley is personalization.


amazon Prime
Steven's Amazon.com
Today's Deals
Gift Cards
Sell
Help
Spring Outlet Event
Shop now
Shop by Department
Search
All
Go
Hello, Steven
Your Account
Your Prime
0 Cart
Wish List
Unlimited Instant Videos
MP3s & Cloud Player
20 million songs, play anywhere
Amazon Cloud Drive
5 GB of free storage
Kindle E-readers & Books
Kindle Fire Tablets
Appstore for Android
NeoCal Advanced Calculator free
Digital Games & Software
Books & Audible
Movies, Music & Games
Electronics & Computers
Home, Garden & Tools
Beauty, Health & Grocery
Toys, Kids & Baby
Clothing, Shoes & Jewelry
Sports & Outdoors
Automotive & Industrial
Full Store Directory
Instant Video
MP3 Store
Cloud Player
Kindle
Cloud Drive
Appstore for Android
Digital Games & Software
Audible Audiobooks
kindle fire HD
The perfect family tablet
Celebrate National Reading Month
Learn more
The Paths Not Taken
Spring Dresses
Give Free Shipping
Spring Cleaning
Amazon Fashion
Instant-pretty styles from Donna Morgan and more.
Spring Dresses
Shop Dresses
Shop All Clothing
April Fool's Day
You've got to be kidding.
Shop now
NISSAN
Innovation that excites
GET A $1000 AMAZON.COM GIFT CARD
WITH THE PURCHASE OF AN ALL-NEW NISSAN ROGUE®
LEARN MORE
Limited time offer. Terms and conditions apply.
Advertisement
SanDisk Ultra 64GB micro SD XC I
Get 40% or More Off
Select microSD Cards
Shop now
Samsung Galaxy S5
Pre-order now
Related to Items You've Viewed
You viewed
Customers who viewed this also viewed
Applied Longitudinal Analysis
Garrett M. Fitzmaurice, Nan M. Laird, ...
Hardcover
(10)
$133.00 $98.23
Applied Longitudinal Data Analysis...
Judith D. Singer, John B. Willett
Hardcover
(23)
$79.95 $70.28
Analysis of Longitudinal Data
Peter Diggle, Patrick Heagerty, ...
Paperback
(1)
$52.95 $47.94
Modelling Survival Data in Medical...
D. Collett
Paperback
(5)
$87.95 $76.59
Multilevel and Longitudinal Modeling...
S. Rabe-Hesketh, Anders Skrondal
Paperback
(7)
$149.95 $130.08
Longitudinal Data Analysis
Donald Hedeker, Robert D. Gibbons
Hardcover
(2)
$157.00 $113.13
Categorical Data Analysis
Alan Agresti
Hardcover
(3)
$150.00 $88.48
View or edit your browsing history
Deals in Computers & Accessories
Best Sellers
Kindle Store : Kindle Books
Updated hourly
1.
Divergent (Divergent Series)
Veronica Roth
Kindle Edition
$4.99
2.
Shadow Spell: Book Two of The Cousins O'Dwyer Trilogy
Nora Roberts
Kindle Edition


NETFLIX
Watch Instantly
Just for Kids
Taste Profile
DVDs
Movies, TV shows, actors, directors, genres
Steven
Top Picks for Steven
Critically-acclaimed Gritty Crime Thrillers
The Last Stand
2013 R 107 minutes
The sheriff of a sleepy border town finds his quiet life interrupted when a drug boss escapes FBI custody and flees straight toward his town. More Info
Starring: Arnold Schwarzenegger, Johnny Knoxville
Director: Ji-woon Kim
Based on your interest in: The Boondock Saints
Our best guess for Steven
Not Interested
My List

takls on big data

Web | News | Videos | Images | Shopping | More | Search tools

About 457,000,000 results (0.56 seconds)

Showing results for ***talks*** on big data

Search instead for takls on big data

Big Data Whitepaper - CenturyLinkTechnology.Com

Ad go.centurylinktechnology.com/ (855) 670-4801

Learn The Four Golden Goals Of **Big Data** Solutions. Free Whitepaper

Big Data Analytics Tool - Actuate.com

Ad www.actuate.com/Analytics-FreeTrial

Fastest **Data** Engine in the Market Turn **Big Data** into **Big** Value

Big Data Explorer Guide - emc.com

Ad www.emc.com/**Big-Data**-Explorer

Learn Everything **Big Data** You Need To Know. Free Interactive Guide.

EMC has 3,678 followers on Google+

The surprising seeds of a **big-data** revolution in healthcare - ...

www.ted.com/**talks**/joel_selanikio_the_surprisin... TED

Data geek Joel Selanikio **talks** through the sea change in collecting health **data** in the past decade ...

Making sense of too much **data** | Playlist | TED

www.ted.com/playlists/56/making_sense_of_too_much_**data** TED

Data mining innovator Shyam Sankar explains why solving **big** problems ... Stats and myths collide in this fascinating **talk** that ends with a remarkable insight. 8.

TED **Talks** about **Data** | TED.com

www.ted.com/topics/**data** TED

Ads

Big Data

cloud.google.com/**Big**Query

Sign-up For Real-time **Big Data** Analytics On Google Bigquery

Big Data Dashboards

www.qlik.com/**Big_Data**

Get **Big** Value From Your **Big Data**. Spot Trend's & Insights In Seconds!

100% Uptime for Hadoop

www.wandisco.com/hadoop

No Downtime No **Data** Loss No Latency 100% reliable realtime availability

Big Data

www.intel.com/**BigData**

Intel® Delivers Technology Built to Scale Your **Big Data** Challenges.

Big Data Analytics eBook

www.**data**meer.com/learn

Resources for **Big Data** Analytics Download your free copy today!

Cloudera **Big Data**

www.cloudera.com/hadoop-training

Big Data Thought Leadership. Software, Training, Certification.

Personalization is a "big logistic regression"

- Response = convert (yes / no)
- Data are structured...

Search queries

- Search query
- Search history (long term)
- Session history
- Demographics
- ...

Search results

- Individual pages
- "One boxes"
- Nested in sites
- Links between sites
- ...

Ads

- Ad creative
- Keywords
- Landing page quality
- Nested in campaigns
- ...

... but structure is often ignored

- Many billions of queries and potential results, billion+ users, many millions of ads (at any given time).
- There are content $\times$ user interactions.

Sp rsi y

Sparsity plays an important role in modeling internet data

- Models are “big” because of a small number of factors with many levels.
- Big data problems are often big collections of small data problems.

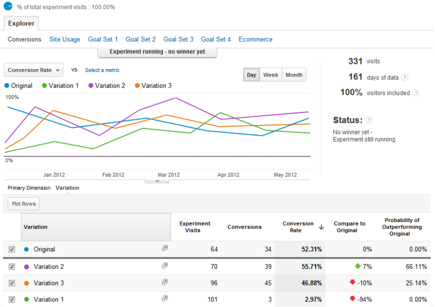
Bayesian ideas remain important in big data

Bayesian themes:

Prediction	Uncertainty	Combine information
Average over unknowns, don't maximize.	Probability coherently represents uncertainty.	Hierarchical models combine information from multiple sources.

- The importance of better predictions is obvious.
- Now for an example of the others...

Classical vs internet experiments



Classical experiments

- Results take a long time.
- Planning is important.
- Type I errors costly.

Internet experiments

- Results come quickly and sequentially.
- All cost is opportunity cost.
- Cost of Type I error is 0.

Multi-armed bandits

Entirely driven by parameter uncertainty

- Problem statement:
  - There are several potential actions available.
  - Rewards come from a distribution $f_a(y|\theta)$.
  - If you knew θ you could choose the best a.
  - Run a sequential experiment to find the best a by learning θ.
- Tradeoff:

Explore	Experiment in case your model is wrong.
Exploit	Do what your model says is best.

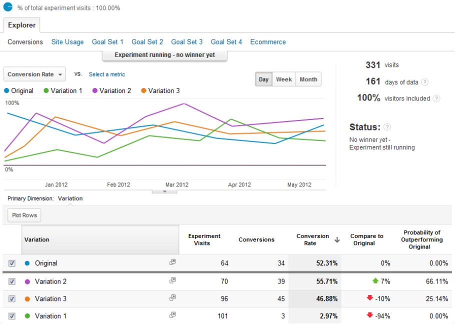
- Thompson sampling heuristic:
  - Given current data $\mathbf{y}$, compute

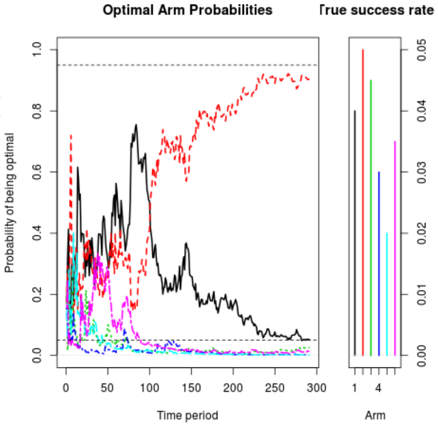
$$\begin{aligned} w_a &= p(\text{action } a \text{ is best}|\mathbf{y}) \\ &= \int p(\text{action } a \text{ is best}|\theta) p(\theta|\mathbf{y}) \, d\theta \end{aligned}$$

  - Assign action a with probability w_a.

Application: Website optimization

For details search Google for [Google analytics multi-armed bandit]

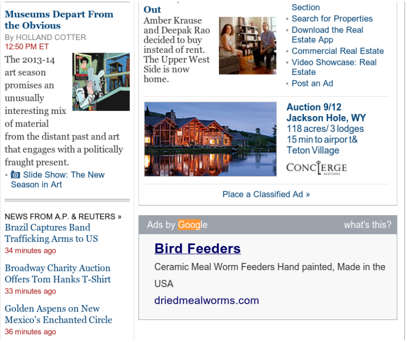


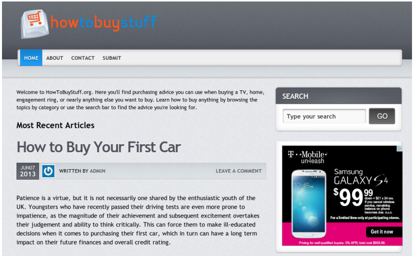


Each website is run as an independent experiment.

“Too easy” for a big data problem.

What about ad optimization?





- The font, text size, and background color of the ad should work well with the font, text size, and color scheme of the domain.
- There are $\sim 600,000$ domains in the (proof of concept) data. Many have too little traffic to support an independent experiment.
- Need uncertainty + information pooling.

Test case: hierarchical logistic regression

$$logitPr(y_{di} = 1|\beta_d) = \beta_d^T \mathbf{x}_{di}$$

$$\beta_d \sim \mathcal{N}(\mu, \Sigma)$$

$$p(\mu, \Sigma) = NIW$$

- d is one of $\sim$600,000 internet domains (blah.com)
- y_{di} indicates whether the ad on domain d was clicked at impression i.
- $\mathbf{x}_{di}$ contains details about ad characteristics: fonts, colors, etc.
  $\mathbf{x}_{di}$ has roughly 10 dimensions.

NOTE: The model is (somewhat) pedagogical.
More sophisticated shrinkage is needed.

The "obvious" algorithm

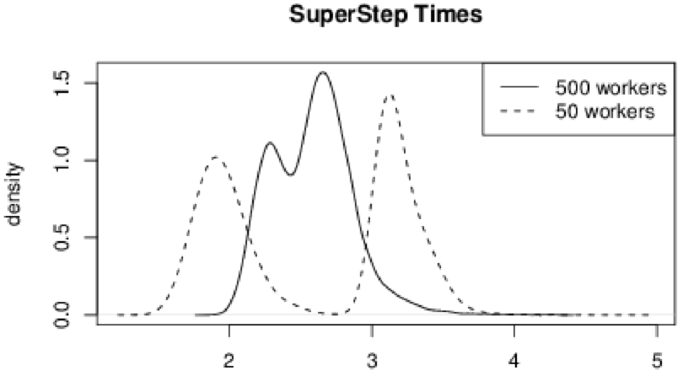
Algorithm

Alternate between the following steps many times.

Map	Each worker simulates $p(\beta_d \mid \mathbf{y}_d, \mu, \Sigma)$.
Reduce	Master node simulates $p(\mu, \Sigma \mid \boldsymbol{\beta})$.

- Theoretically identical to the standard "single machine" MCMC.
- Works great on HPC/GPU.
- It works terribly in MapReduce.

SuperStep Times



500 worker job has

- Less time per step drawing β_d
- Same amount of work drawing μ, Σ.

But the μ, Σ draw takes *more* time!
Extra time spent *coordinating* the extra machines.

Workers	Hours
50	$\sim$5.20
500	$\sim$2.75

Both too slow!

The MapReduce paradox

- We started off compute constrained.
- We parallelized.
- We wound up using essentially **0%** of the compute resources on the worker machines.

Conclusion

For Bayesian methods to work in a MapReduce / Hadoop environment, we need algorithms that require very little communication.

Consensus Monte Carlo

Embarassingly parallel: only one communication

- Partition the data $\mathbf{y}$ into "shards" $\mathbf{y}_1, \ldots, \mathbf{y}_S$.
- Run a separate Monte Carlo on each shard-level posterior $p(\theta|\mathbf{y}_s)$. (MCMC, SMC, QMC, MC, ...)
- Combine the draws to form a "consensus" posterior distribution. (Like doing a meta-analysis)

This should work great, because most Bayesian problems can be written

$$p(\theta|\mathbf{y}) = \prod_{s=1}^{S} p(\theta|\mathbf{y}_s)$$

Complication: You have draws from $p(\theta|\mathbf{y}_s)$, not the distribution itself.

Benefits: No communication overhead. You can use existing software.

Two issues

- What to do about the prior?
  - Fractionate: $p_s(\theta) = p(\theta)^{1/S}$
    (reasonable, but beware impropriety and consider shrinkage)
  - Adjust by multiplying / dividing "working priors."

- How to form the consensus?
  - Average [Scott *et. al* (2013)], [Wang and Dunson (2013)]
  - Other methods:
    - Importance resampling (SFS?) [Huang and Gelman (2005)]
    - Kernel density estimates [Neiswanger, *et. al* (2013)]
    - Parametric approximation [Huang and Gelman (2005)]
      [Machine Learning]

Averaging individual draws? The Gaussian case

- If p_1 and p_2 are normal, then $p_1 p_2 =$ "prior $\times$ likelihood".
- If $z \sim p_1 p_2$ then

$$z \sim \mathcal{N}\left(\frac{\frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}, \frac{1}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}\right)$$

- If $x \sim p_1$ and $y \sim p_2$, then $w_1 x + w_2 y \sim p_1 p_2$ (where $w_i \propto 1/\sigma_i^2$).

This requires knowing the weights to use in the average, but

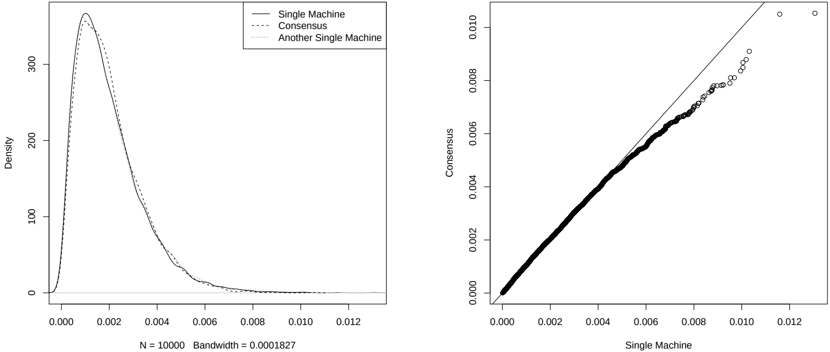
1. Sometimes you know them.
2. We've got a full MC for each shard, so we can easily estimate within-shard variances.

Averaging vs. other methods.

- Advertising is robust.
  - Not susceptible to infinite resampling weights or the curse of dimensionality.
  - Can't capture multi-modality or discrete parameter spaces.
- Kernel density estimates will break down in high dimensions
  - Despite claims of being "asymptotically exact".
  - Beware the constant in $\mathcal{O}(\cdot)$.
- Importance resampling is probably the long term answer
  - Details still need to be worked out.
  - What to use for importance weights?
  - Worker level posteriors can be widely separated, and not overlap the true posterior, in which case resampling would fail.
- Parametric models
  - Variational Bayes gets the marginal histograms right, but not the joint distribution.
  - Parametric assumptions are reasonable if they're reasonable.

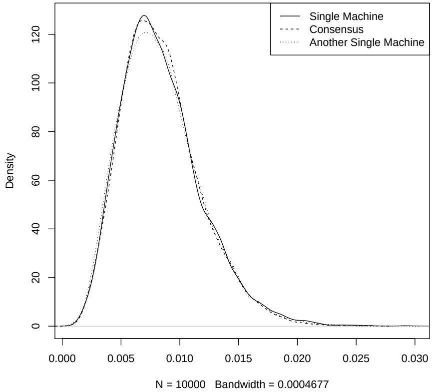
Binomial

1000 Bernoulli outcomes, with one success, 100 workers.



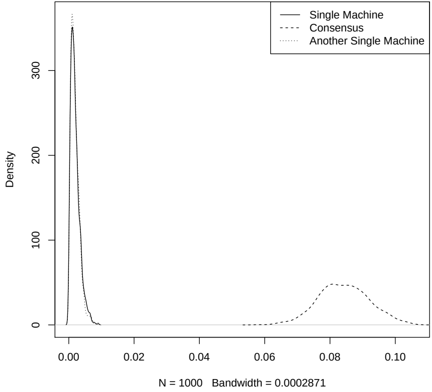
This beta distribution is highly skewed (non-Gaussian) but it still works.

Unbalanced is okay



- 5 workers (100, 20, 20, 70, 500) observations.
- Weights (proportional to n) handle information uncertainty correctly.

Be careful with the prior



- Beta prior
- Distribute information ("fake data") evenly across shards.
- Notion of a "natural scale" is still a mystery.

- If S copies of the prior contain meaningful information then you need to be careful.
- If not, then you can afford to be sloppy.

Logistic regression

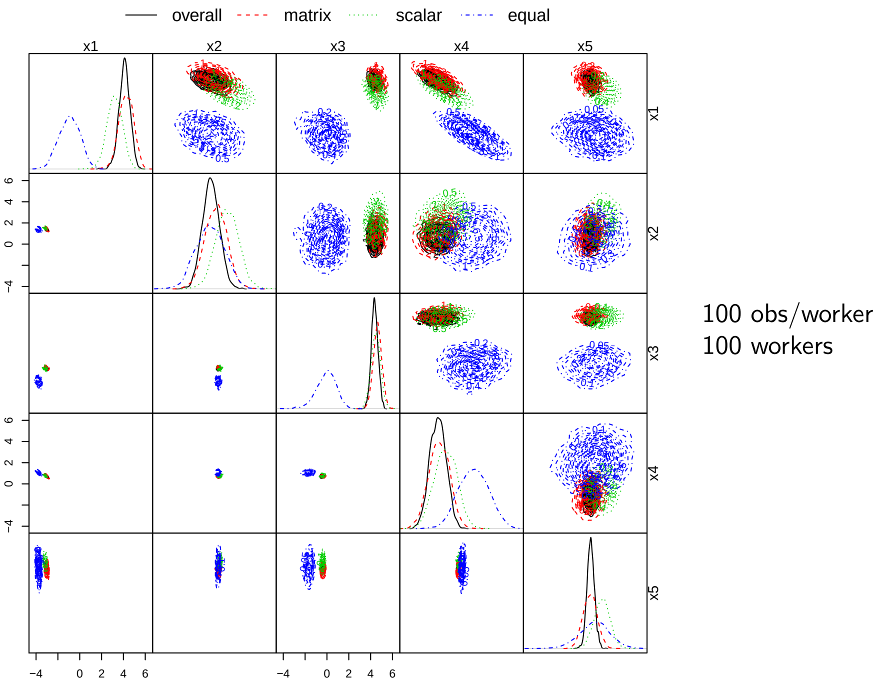
- Test data has 5 binary predictor variables

	x_1	x_2	x_3	x_4	x_5
frequency	1	.2	.3	.5	.01
coefficient	-3	1.2	-.5	.8	3

- Last one is rare, but highly predictive when it occurs.

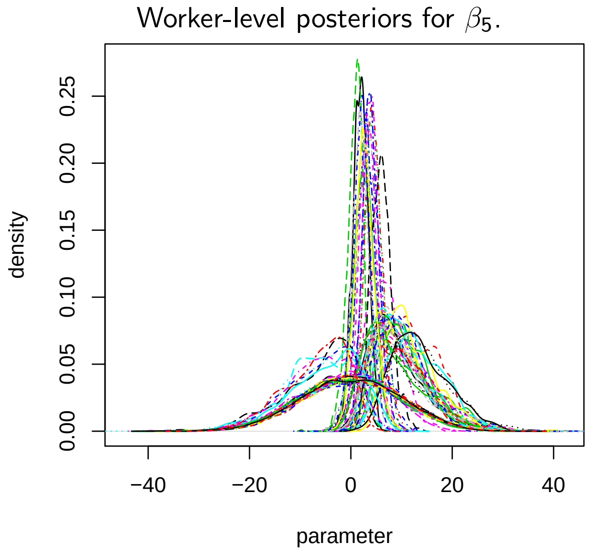
Data

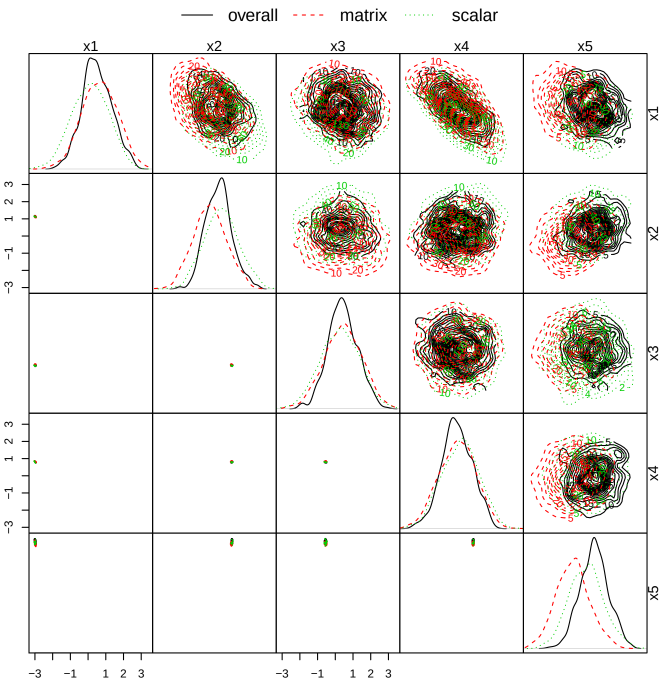
y	n	x_1	x_2	x_3	x_4	x_5
266	2755	1	0	0	1	0
116	2753	1	0	0	0	0
34	1186	1	0	1	0	0
190	717	1	1	0	1	0
61	1173	1	0	1	1	0
37	305	1	1	1	0	0
68	301	1	1	1	1	0
119	706	1	1	0	0	0
18	32	1	0	0	0	1
13	17	1	0	1	1	1
18	24	1	0	0	1	1
8	10	1	1	0	1	1
2	2	1	1	1	0	1
7	13	1	0	1	0	1
2	2	1	1	1	1	1
3	4	1	1	0	0	1


overall
matrix
scalar
equal
x1
x2
x3
x4
x5
100 obs/worker
100 workers

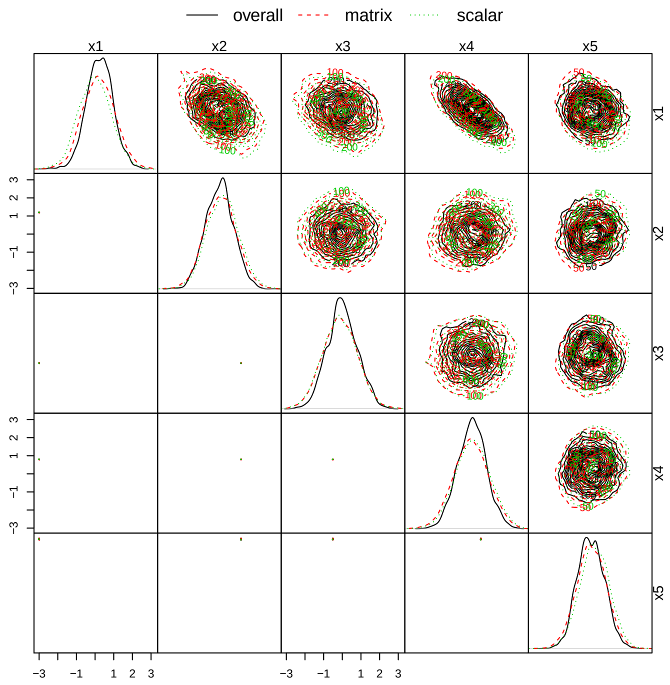
Why equal weighting fails

Even though workers have identical sampling distributions.





1000 obs / worker
100 workers



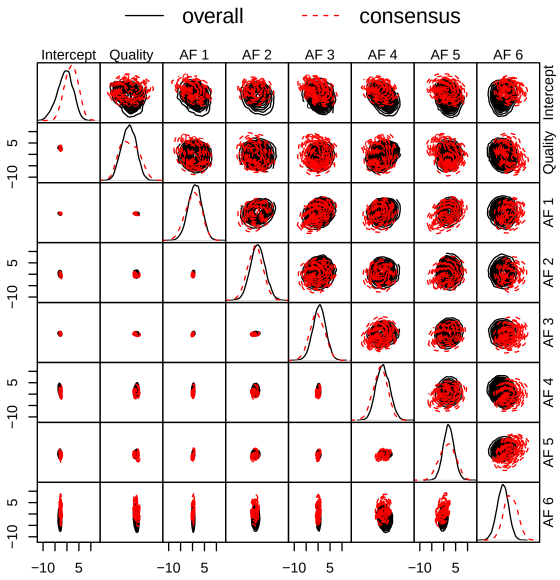
10,000 obs / worker
100 workers

Hierarchical Poisson regression

- 24 million observations (ad impressions)
- Predictors include a "quality score" and indicators for 6 different "ad formats".
- Coefficients vary by advertiser (roughly 11,000 advertisers here).
- Data sharded by advertiser. 867 shards. No shard has more than 50 thousand observations. Median shard size is 27,000 observations.

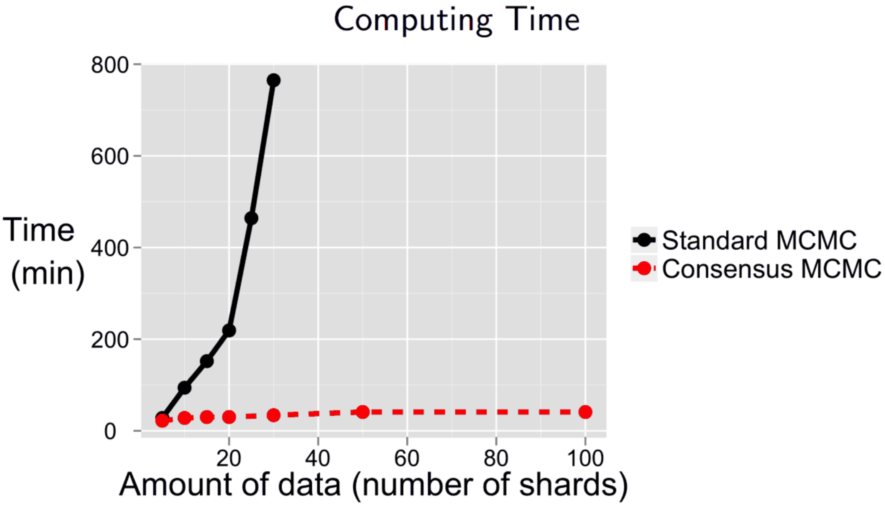
Hierarchical Poisson regression

Posterior for hyperparameters given 6 data shards



- Very close, even on joint distribution.
- Parameters are embarassingly parallel, given marginal for hyperparameters.
- Can get parameters in one more “Map” step.

Compute time for hierarchical models



BART

[Chipman, George, McCulloch (2010)]

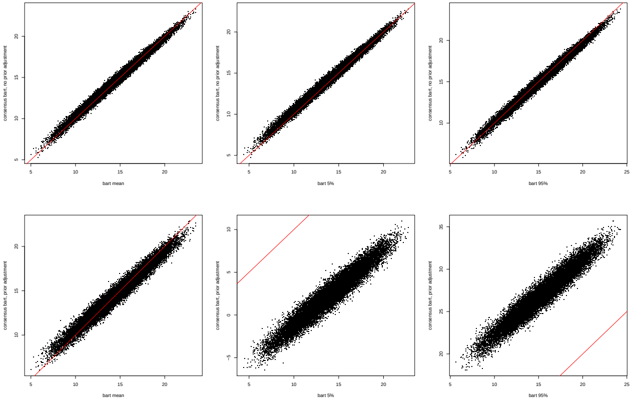
$$y_i = \sum_j f_j(\mathbf{x}_i) + \epsilon_i \qquad \epsilon_i \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right)$$

- Each f_i is a tree.
- Priors keep the trees small.
  "Sum of weak learners" idea akin to boosting.
- Trees have means in the leaves.

Can't average trees, but we can average $\hat{f}(\mathbf{x})$.

Consensus Bart vs Bart

Fits vs Friedman's test function. 30 shards, 20K total observations.



- Columns: mean, and (.025, .975) quantiles. Consensus vs. single machine.
- Top row: same prior (nearly perfect match)
- Bottom row: prior$^{1/30}$ (fractionated prior $\rightarrow$ overdispersed posterior)

Conclusions

- Bayesian ideas are important in big data problems for the same reasons they're important in small data problems.
- Naive parallelizations of single machine algorithms are not effective in a MapReduce environment (too much coordination).
- Consensus Monte Carlo

Good	- Requires only one communication. - Reasonably good approximation. - Can use existing code.
Bad	- No theoretical guarantees yet (in non-Gaussian case). - Averaging only works where averaging makes sense. - Not good for discrete parameters, spike-and-slab, etc. - Need to work out resampling theory.

References

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010).
Bart: Bayesian additive regression trees.
The Annals of Applied Statistics **4**, 266–298.

Dean, J. and Ghemawat, S. (2004).
Mapreduce: Simplied data processing on large clusters.
In *OSDI'04: Fifth Symposium on Operating System Design and Implementation*.

Huang, Z. and Gelman, A. (2005)
Sampling for Bayesian computation with large datasets
(unpublished)

Malewicz, G., Austern, Matthew H. Bik, A. J. C., Dehnert, J. C., Horn, I., Leiser, N., and Czajkowski, G. (2010).
Pregel: A system for large-scale graph processing.
In *SIGMOD'10*, 135–145.

Neiswanger, W, Wang, C, and Xing, E (2013).
Asymptotically Exact, Embarrassingly Parallel MCMC
arXiv preprint arXiv:1311.4780.

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I., and McCulloch, R. E. (2013)
Bayes and Big Data: The Consensus Monte Carlo Algorithm
http://research.google.com/pubs/pub41849.html.

Wang, X. and Dunson, D. B. (2013)
Parallel MCMC via Weierstrass Sampler
arXiv preprint arXiv:1312.4605.