


Computational Challenges for Real-Time Marketing with Large Datasets

Alan Montgomery
Associate Professor
Carnegie Mellon University
Tepper School of Business

e-mail: alan.montgomery@cmu.edu
web: <http://www.andrew.cmu.edu/user/alm3>

*Seventh Invitational Choice Symposium,
Philadelphia, Pennsylvania
14 June 2007*



Outline

- Real-Time Marketing
- Current State of Computation in Industry
- Example Choice Problem
 - Computational Considerations
- Conclusions

2

Real-Time Marketing

Examples, Methods, and Problems

Real-Time Marketing Problems

- Web design
- Browsing
- Web search
- Promotion/Pricing

The screenshot displays a Microsoft Internet Explorer browser window with a search for "Real Time Marketing" on Google. The search results page features a "Great Offers" section for a Kodak Easyshare C743 camera, priced at \$149.99. The browser window also shows the Amazon.com website in the background.

Real-Time Marketing Problems

- These problems can be characterized as consumer choice problems. If we can predict likelihood of choices as a function of decision environment then we can determine best decision.
- Consider requirements for an e-Retailer:
 - 10,000 user-sessions per day @ 10 requests per session =
 - 100,000 page requests per day (~1 per second) =
 - 500 transactions per day (one every ~20 seconds)
 - User history could be thousands of pieces of information, megabytes of information per user
 - Must be able to respond to user in less than a second

5

Proposal

- Model of choice using Multinomial probit models
 - MCMC with data augmentation step proposed by Rossi, McCulloch, and Allenby (1996), "The value of purchase history data in target marketing"
- Heavy emphasis on Bayesian Models which generally required numerical integration through simulation methods to yield solutions, such as Monte-Carlo Markov Chain (MCMC)
- These methods are too slow and generally sequential
 - Computational times in hours for "small" problems with hundreds of users
- Goal: Modify existing approaches to work in a grid environment

6



Counterarguments

- Use rule based approaches which are fast, interpretable, and simple to implement
 - Not using data efficiency
- Use logit models or models where numerical integration is not necessary
 - Inadequate to model task
- Use frequentist methods over Bayes methods which emphasize point estimates instead of distribution
 - Want best decision theoretic solutions

7


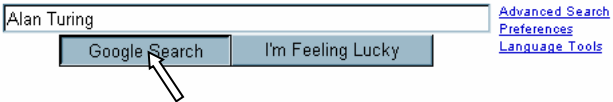


Current State of Computation in Industry

Google Example

Source: Randy Bryant,
SCS, Carnegie Mellon University

Motivation

- 200+ processors
- 200+ terabyte database
- 10^{10} total clock cycles
- 0.1 second response time
- 5¢ average advertising revenue

Source: Randy Bryant

9

Google's Computing Infrastructure

- System
 - ~ 3 million processors in clusters of ~2000 processors each
 - Commodity parts
 - x86 processors, IDE disks, Ethernet communications
 - Gain reliability through redundancy & software management
 - Partitioned workload
 - Data: Web pages, indices distributed across processors
 - Function: crawling, index generation, index search, document retrieval, Ad placement
- A Data-Intensive Super Computer (DISC)
 - Large-scale computer centered around data
 - Collecting, maintaining, indexing, computing

Source: Randy Bryant

Barroso, Dean, Hölzle, "Web Search for a Planet: The Google Cluster Architecture" IEEE Micro 2003 10

Google's Economics

- Making Money from Search
 - \$5B search advertising revenue in 2006
 - Est. 100 B search queries
 - → 5¢ / query average revenue
- That's a Lot of Money!
 - Only get revenue when someone clicks sponsored link
 - Some clicks go for \$10's
- That's Really Cheap!
 - Google + Yahoo + Microsoft: \$5B infrastructure investments in 2007

Sponsored Links

[Do you have mesothelioma?](#)
Let Our Law Firm Fight the Asbestos Companies for You!
[www.masolawsuit.com](#)

[Have Mesothelioma Cancer?](#)
We'll Fight To Win The Compensation You Deserve! Call (800) 946-9646
[www.MesotheliomaNews.com](#)

[Mesothelioma Lawyer](#)
Mesothelioma Cases is all we do. Get \$5 and protect your family.
[www.Legal-Mesothelioma-Help.com](#)

[Mesothelioma](#)
Mesothelioma medical & legal resource. Get legal assistance.
[www.MesotheliomaFYI.com](#)
Pennsylvania

[Mesothelioma Cancer](#)
Medical & Legal Resources for Those With Mesothelioma - Get Help Here.
[www.MesotheliomaCenter.org](#)

[Asbestos exposure kills](#)
Make a claim for your compensation. There is money available - just ask
[www.askaboutmesothelioma.com](#)

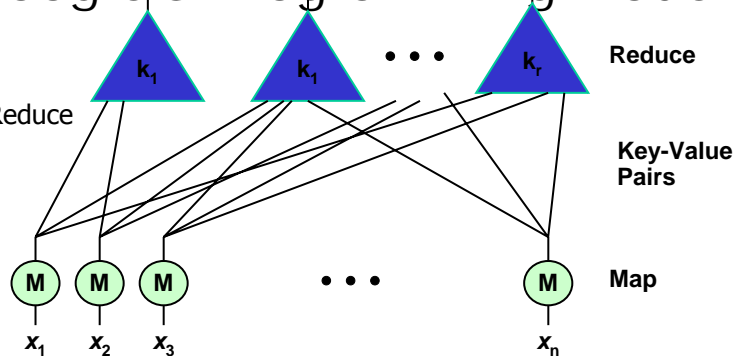
[Mesothelioma Treatment](#)
Mesothelioma Treatment Information
Mesothelioma Treatment Attorney
[MesotheliomaTreatmentHelpCenter.com](#)

[Mesothelioma Empowerment](#)
Patient profiles, medical help.
We only handle mesothelioma cases.
[www.mesothel.com](#)

Source: Randy Bryant

Google's Programming Model

- MapReduce



- Map computation across many objects
 - E.g., 10^{10} Internet web pages
- Aggregate results in many different ways
- System deals with issues of resource allocation & reliability

Source: Randy Bryant

Dean & Ghemawat: "MapReduce: Simplified Data Processing on Large Clusters", OSDI 2004

Other Grid Environments

- Sun Grid
 - \$1/CPU-hr, SunFire dual process Opteron-based servers with 4 Gb of RAM per CPU
- IBM Grid
- Amazon Elastic Compute Cloud (EC2)
- Potential for Sony to over commercial PS3 computing grid
 - Folding@Home: 31,761 PS3 CPUs churning out 416 teraflops versus 184,134 active Windows machines producing 175 teraflops

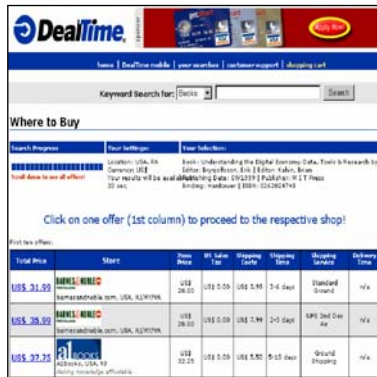
13

Example Choice Problem

"The Great Equalizer? An Empirical Study of Consumer Choice at a Shopbot",

Erik Brynjolfsson, Michael Smith, and Alan Montgomery (2007) Working Paper.

What is a shopbot?



Shopping Robot's automatically search a large number of stores for a specific product

Makes search quick and simple. Average range in prices is \$12, and Amazon is lowest only 5% of time (data from 2001).

Example:

John Grisham's *The Brethren*, list price \$27.95, prices range between \$13.49 (buy.com) and \$50.75 (totalinformation.com)

15

Shopbots as Choice Problems

Total Price	State	Item Price	US Sales Tax	Shipping Costs	Shipping Time	Shipping Service	Delivery Time
US\$ 43.98	amazon.com	US\$ 35.00	US\$ 0.00	US\$ 8.98	2 days	Second Day Air	5 days
US\$ 45.95	buy.com	US\$ 35.00	US\$ 0.00	US\$ 10.95	1 day	Next Day Air	n/a
US\$ 46.00	alBooks	US\$ 32.25	US\$ 0.00	US\$ 13.75	2 days	UPS 2nd Day (blue)	n/a
US\$ 48.95	alibris	US\$ 46.00	US\$ 0.00	US\$ 2.95	3-14 days	U.S. Postal Service	n/a
US\$ 48.98	amazon.com	US\$ 35.00	US\$ 0.00	US\$ 13.98	1 day	Next Day Air	4 days
US\$ 49.95	alibris	US\$ 46.00	US\$ 0.00	US\$ 3.95	1-6 days	UPS	n/a
US\$ 36.75	1 Booktree.com, USA, CA	US\$ 36.75	US\$ 0.00	US\$ 0.00	3-14 days	USPS Parcel Post	6-21 days
US\$ 38.88	AlphaCraze.com, USA	US\$ 35.00	US\$ 0.00	US\$ 3.88	4-14 days	USPS Special Rate	5-15 days
US\$ 40.48	AlphaCraze.com, USA	US\$ 35.00	US\$ 0.00	US\$ 5.48	2-3 days	Express Priority Mail	3-4 days
US\$ 40.75	Page 1 Book, USA, NM	US\$ 35.00	US\$ 0.00	US\$ 5.75	2-6 days	Federal Express Ground	n/a
US\$ 42.70	1 Booktree.com, USA, CA	US\$ 36.75	US\$ 0.00	US\$ 5.95	4-6 days	UPS Ground	7-13 days
US\$ 43.70	1 Booktree.com, USA, CA	US\$ 36.75	US\$ 0.00	US\$ 6.95	2-3 days	Priority Mail	5-10 days

16

Unique Choice Aspects

- Choice is many from many
 - Usual multinomial logit/probit models consider only the problem of one from N, here we observe M from N; Requires a multivariate probit model
- Search behavior is dynamic
 - The amount of search could depend upon the price of the book, when the search is made, the expertise of the user, ...
- Ordering is very important
 - Shelf Design information is frequently not known in scanner choice data, here we know the exact tabular format

17

Our Choice Problem

User selects two offers from 45 presented, we observe all attributes and selections

Many from many choice problem

Total Price	Store	Item Price	% Sales Tax	Shipping Costs	Shipping Time	Shipping Service	Delivery Time
US\$ 43.98	amazon.com	US\$ 35.00	US\$ 0.00	US\$ 8.98	2 days	Second Day Air	5 days
US\$ 45.95	buy.com	US\$ 35.00	US\$ 0.00	US\$ 10.95	1 day	Next Day Air	n/a
US\$ 46.00	1	US\$ 32.25	US\$ 0.00	US\$ 13.75	2 days	US 2nd Day (Blue)	n/a
US\$ 48.95	Walmart	US\$ 46.00	US\$ 0.00	US\$ 2.95	3-14 days	U.S. Postal Service	n/a
US\$ 48.98	amazon.com	US\$ 35.00	US\$ 0.00	US\$ 13.98	1 day	Next Day Air	4 days
US\$ 49.95	Walmart	US\$ 46.00	US\$ 0.00	US\$ 3.95	1-4 days	UPS	n/a
US\$ 36.75	1 Bookstree.com, USA, CA	US\$ 36.75	US\$ 0.00	US\$ 0.00	3-14 days	USPS Parcel Post	4-21 days
US\$ 38.88	Alphatrace.com, USA	US\$ 35.00	US\$ 0.00	US\$ 3.88	4-14 days	USPS Special Rate	5-15 days
US\$ 40.48	Alphatrace.com, USA	US\$ 35.00	US\$ 0.00	US\$ 5.48	2-3 days	Express Priority Mail	3-4 days
US\$ 40.75	Page 1 Book, USA, NH	US\$ 35.00	US\$ 0.00	US\$ 5.75	2-4 days	Federal Express Ground	n/a
US\$ 42.78	1 Bookstree.com, USA, CA	US\$ 36.75	US\$ 0.00	US\$ 5.95	4-4 days	UPS Ground	7-13 days
US\$ 43.78	1 Bookstree.com, USA, CA	US\$ 36.75	US\$ 0.00	US\$ 6.95	2-3 days	Priority Mail	5-10 days

Generalizing the Multivariate Probit Model

- User considers all items with latent utility greater than a threshold:

$$y_{ij} = \begin{cases} 1 & \text{if } u_{ij} \geq \lambda_{ij} \\ 0 & \text{otherwise} \end{cases}$$

- Threshold is equal to the P th order statistic of the latent utilities (where the user views N offers):

$$\lambda_{ij} = \mathbf{u}_{it}^{-j} \langle P_{it} \rangle$$

- Multivariate probit model occurs when $\lambda=0$. Also, multinomial probit when $p=1$ (or maximum):

$$\lambda_{ij} = \mathbf{u}_{it}^{-j} \langle 1 \rangle = \max(u_{it1}, \dots, u_{it,j-1}, u_{it,j+1}, \dots, u_{itK})$$

19

Modeling Choice Set Size

- Introduce a poisson-log normal regression model. (Note: the log normal error overcomes the common overdispersion problem of the poisson model)

$$P_{it} \sim \text{TruncatedPoisson}(\theta_{it}), \quad P_{it} = 0, 1, \dots, N_{it}$$

$$\ln(\theta_{it}) = \underbrace{\boldsymbol{\gamma}'_i \mathbf{w}_{it}} + \alpha_{it}, \quad \alpha_{it} \sim N(0, \tau^2)$$

Book price, book type, time of search, cumulative number of dealtime visits, ...

20

Consumer Utility Model

- Additive utility model for the k th product given N alternatives with A attributes shown in the set:

$$u_{ij} = \underbrace{\beta'_{it} \mathbf{x}_{ij}} + \underbrace{\varepsilon_{ij}}, \quad \varepsilon_{ij} \sim N(0, 1)$$

Observable attributes: store name, price, logo, order, etc.

Allow error to follow spatial autoregression: if an offer is next to an unexpectedly good one, it's more likely to be selected

21

Hierarchical Bayesian Model

- Session Coefficients follow a linear model:


$$\beta_{it} \sim N(\mathbf{K}\mathbf{w}_{it}, \mathbf{H})$$

$$\delta_{it} \sim N(\bar{\delta}, \mathbf{X}) \cdot \mathbf{I}(-1 \leq \delta_{it} \leq 1), \quad \delta_{it} = [\phi_{it} \quad \theta_{it}]'$$

- User Coefficients follow a linear model:

$$\gamma_i \sim N(\mathbf{\Gamma}\mathbf{z}_i, \mathbf{\Omega})$$


22



Some MCMC Estimation Notes

- The choice dimension (~ 40 offers) is quite large so we need to be very efficient. Use properties of partitioned matrices so only one 40×40 matrix inversion is necessary instead of 40 39×39 matrix inversions. (Dramatically speeds things up for updating latent utilities)
- Use a slice sampler to estimate ARMA(1,1) parameters, follow Tiao and Ali (1971) for an efficient scheme to invert covariance matrix and calculate covariance.
- Use a slice sampler to estimate truncated poisson-log normal distribution

23



Proposed estimation algorithm

- Generally we must estimate all users and coefficients at the same time, however here we can focus on estimating only the latent utilities.
- We must also simultaneously consider our decision problem, and evaluate the expected probability from various choices

24

Computational Considerations

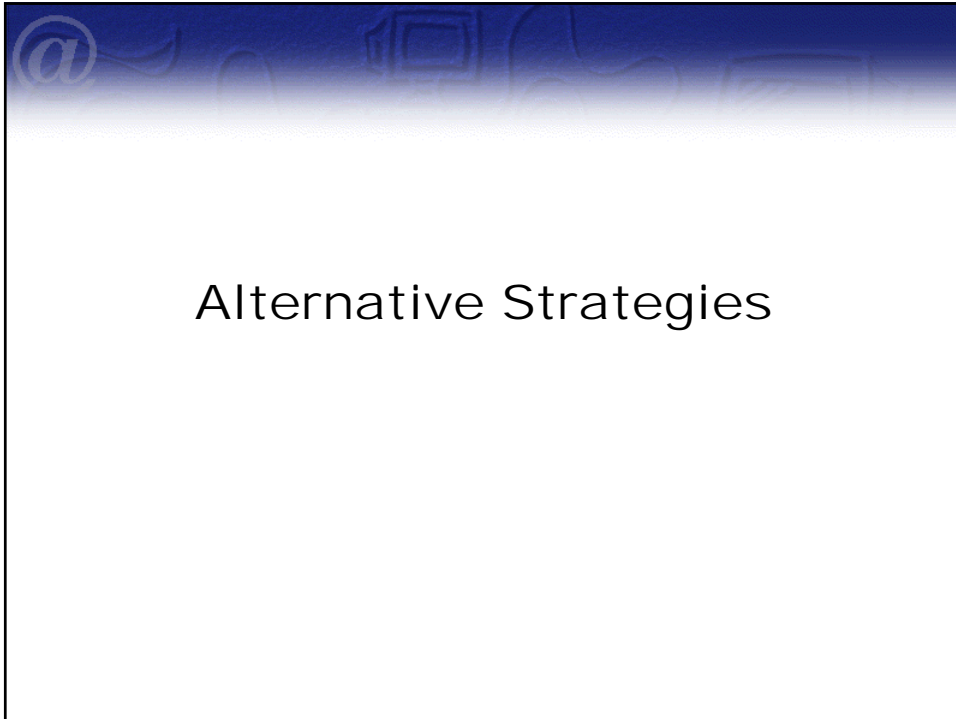
- Must be able to lookup and retrieve user history from memory. To aid in start-up time we could initialize the starting point.
 - 20 Sessions X 40 Observations X 10 Variables [80 bytes] = 64kB of User History
 - Can we sample the user's history? Which searches important?
 - Potentially keep user history online:
 - Consider that 10,000 users x 64kB = 64 mB
 - Consider that 1,000,000 users x 64kB = 6.4 gB
- Evaluate grid of points

25

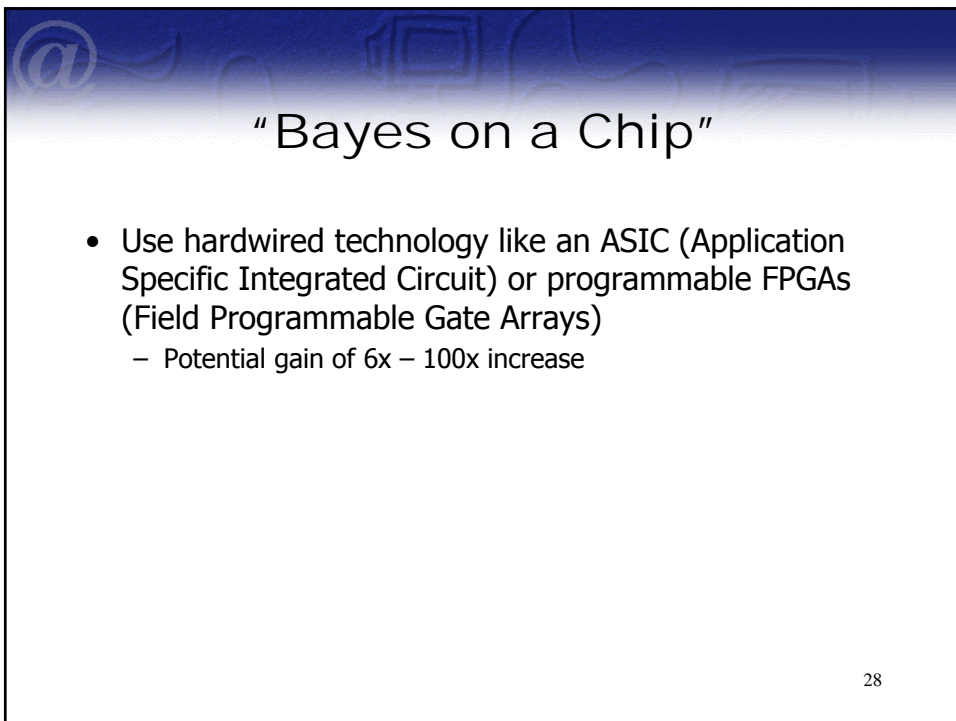
MCMC Approaches

- Griddy Gibbs
 - Discretize parameter space
 - Could we select reasonable 'grid' points offline
- Metropolis-Hastings or Slice Sampler
 - Could employ exponential cascading, in which one processor could suggest points that others should sample at
 - Randomly shuffle conditional draws

26

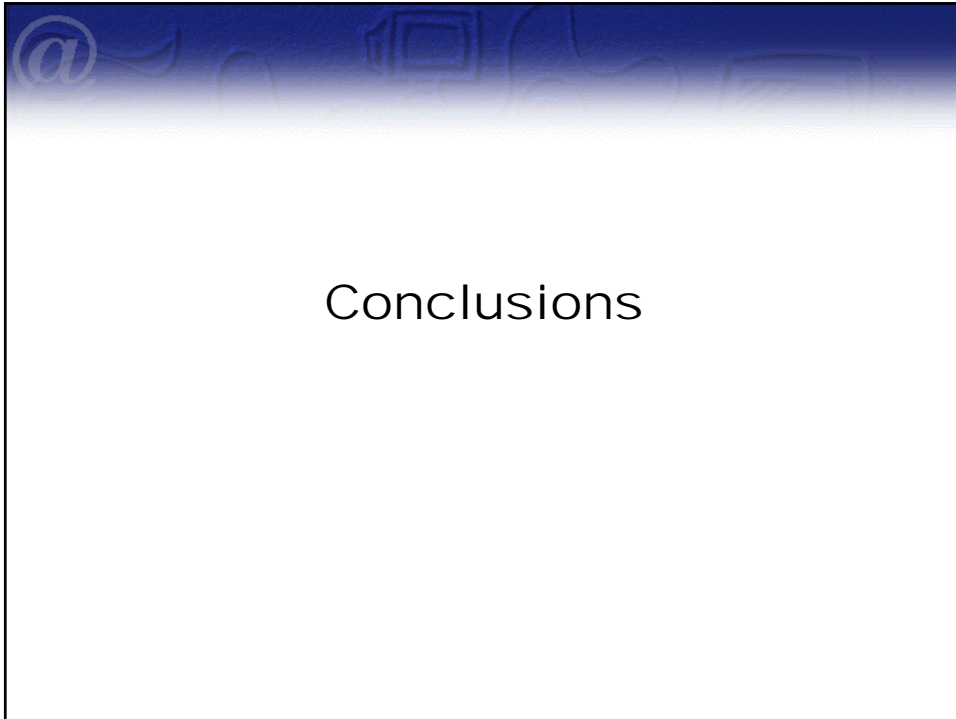


Alternative Strategies

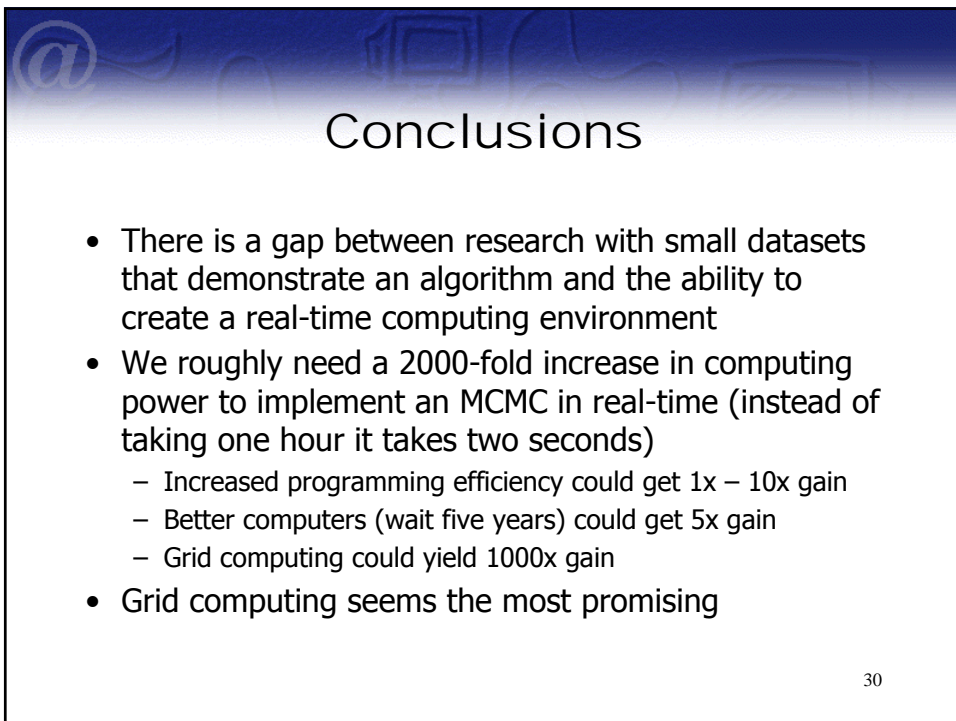


"Bayes on a Chip"

- Use hardwired technology like an ASIC (Application Specific Integrated Circuit) or programmable FPGAs (Field Programmable Gate Arrays)
 - Potential gain of 6x – 100x increase



Conclusions



Conclusions

- There is a gap between research with small datasets that demonstrate an algorithm and the ability to create a real-time computing environment
- We roughly need a 2000-fold increase in computing power to implement an MCMC in real-time (instead of taking one hour it takes two seconds)
 - Increased programming efficiency could get 1x – 10x gain
 - Better computers (wait five years) could get 5x gain
 - Grid computing could yield 1000x gain
- Grid computing seems the most promising

30