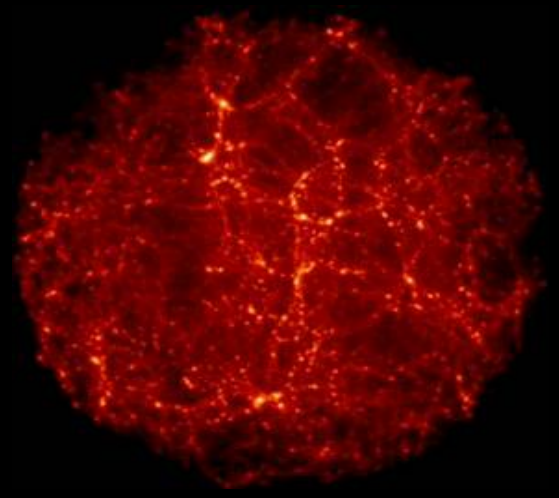
Computational Mathematics for Large-Scale Data Analysis

Enabling Scalable Astrostatistics



Alexander Gray

Georgia Institute of Technology

College of Computing

FASTlab: Fundamental Algorithmic and Statistical Tools

Is science in 2008 different from science in 1908?

Instruments

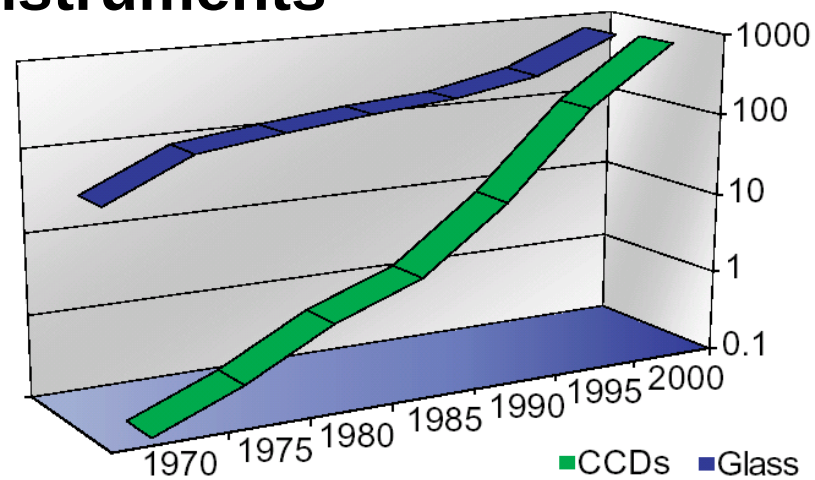


Fig. 1. Telescope area doubles every 25 years, whereas telescope CCD pixels double every 2 years. This rate seems to be accelerating. It

[*Science*, Szalay & J. Gray, 2001]

Is science in 2008 different from science in 1908?

Instruments

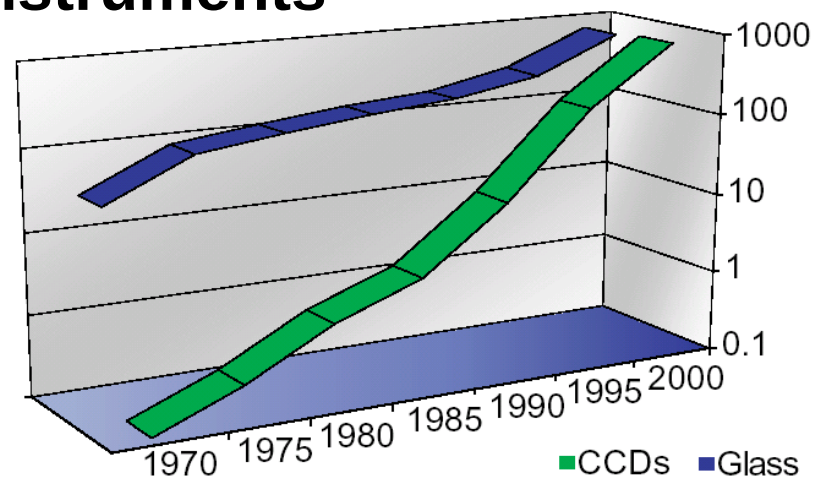
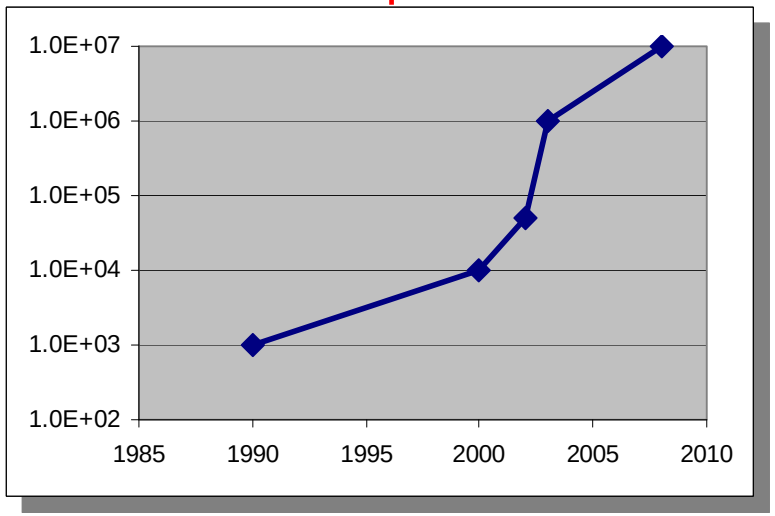


Fig. 1. Telescope area doubles every 25 years, whereas telescope CCD pixels double every 2 years. This rate seems to be accelerating. It

[*Science*, Szalay & J. Gray, 2001]

Data: CMB Maps



1990 COBE	1,000
2000 Boomerang	10,000
2002 CBI	50,000
2003 WMAP	1 Million
2008 Planck	10 Million

Data: Local Redshift Surveys

1986 CfA	3,500
1996 LCRS	23,000
2003 2dF	250,000
2005 SDSS	800,000

Data: Angular Surveys

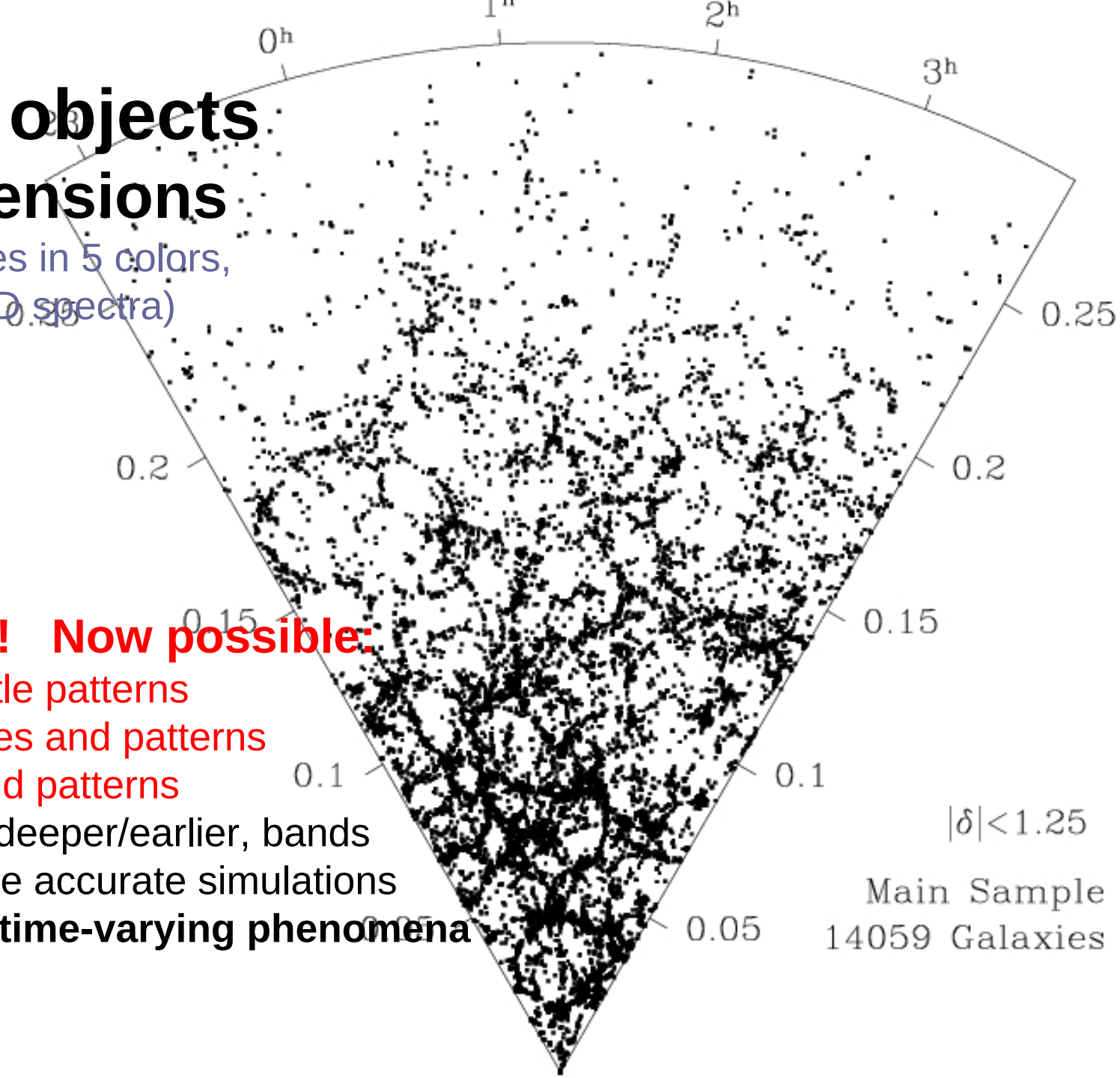
1970 Lick	1M
1990 APM	2M
2005 SDSS	200M
2008 LSST	2B

Sloan Digital Sky Survey (SDSS)

1 billion objects

144 dimensions

(~250M galaxies in 5 colors,
~1M 2000-D spectra)



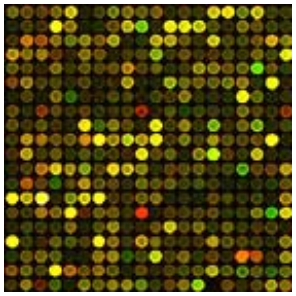
Size matters! Now possible:

- low noise: subtle patterns
- global properties and patterns
- rare objects and patterns
- more info: 3d, deeper/earlier, bands
- in parallel: more accurate simulations
- **2008: LSST – time-varying phenomena**

Happening everywhere!

microarray chips

Molecular biology

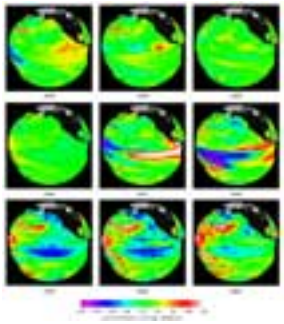


nuclear mag. resonance

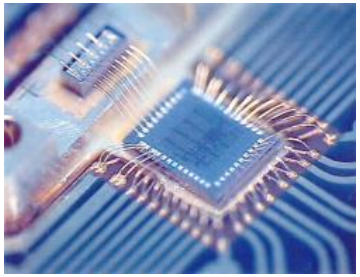
Drug discovery

satellite topography

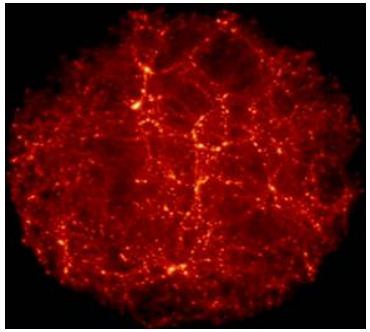
Earth sciences



microprocessor

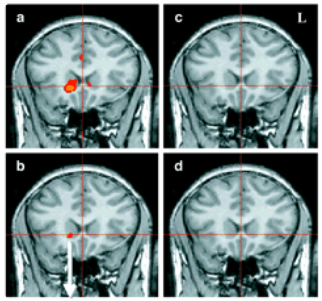


Physical simulation

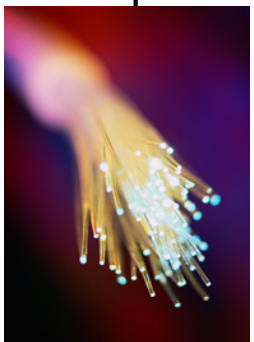


functional MRI

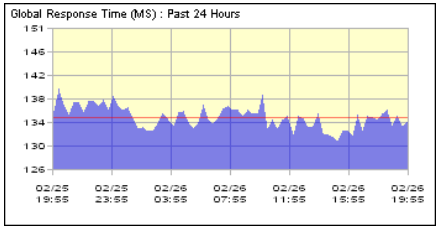
Neuroscience



fiber optics



Internet



Astrophysicist

Robert Nichol

1. How did galaxies evolve?
2. What was the early universe like?
3. Does dark energy exist?
4. Is our model (GR+inflation) right?

R. Nichol, Inst. Cosmol. Gravitation
A. Connolly, U. Pitt Physics
C. Miller, NOAO
R. Brunner, NCSA
G. Kulkarni, Inst. Cosmol. Gravitation
D. Wake, Inst. Cosmol. Gravitation
R. Scranton, U. Pitt Physics
M. Balogh, U. Waterloo Physics
I. Szapudi, U. Hawaii Inst. Astronomy
G. Richards, Princeton Physics
A. Szalay, Johns Hopkins Physics

Machine learning/
statistics guy

Astrophysicist

Robert Nichol

1. How did galaxies evolve?
2. What was the early universe like?
3. Does dark energy exist?
4. Is our model (GR+inflation) right?

R. Nichol, Inst.
A. Connolly, U
C. Miller, NOAC
R. Brunner, NC
G. Kulkarni, In
D. Wake, Inst.
R. Scranton, U.
M. Balogh, U. Waterloo Physics
I. Szapudi, U. Hawaii Inst. Astro.
G. Richards, Princeton Physics
A. Szalay, Johns Hopkins Physics

- Kernel density estimator
- n-point spatial statistics
- Nonparametric Bayes classifier
- Support vector machine
- Nearest-neighbor statistics
- Gaussian process regression
- Hierarchical clustering

Machine learning/
statistics guy

Astrophysicist

Robert Nichol

1. How did galaxies evolve?
2. What was the early universe like?
3. Does dark energy exist?
4. Is our model (GR+inflation) right?

R. Nichol, Inst.
A. Connolly, U
C. Miller, NOAC
R. Brunner, N
G. Kulkarni, In
D. Wake, Inst.
R. Scranton, U.
M. Balogh, U. Waterloo Physics
I. Szapudi, U. Hawaii Inst. Astro.
G. Richards, Princeton Physics
A. Szalay, Johns Hopkins Physics

- Kernel density estimator $O(N^2)$
- n-point spatial statistics $O(N^n)$
- Nonparametric Bayes classifier $O(N^2)$
- Support vector machine $O(N^2)$
- Nearest-neighbor statistics $O(N^2)$
- Gaussian process regression $O(N^3)$
- Hierarchical clustering $O(N^3)$

Machine learning/
statistics guy

Astrophysicist

Robert Nichol

1. How did galaxies evolve?
2. What was the early universe like?
3. Does dark energy exist?
4. Is our model (GR+inflation) right?

R. Nichol, Inst.
A. Connolly, U
C. Mi r, NOAO
R. Bru r, NO
G. Ku In
D. Wa
R. Scra
M. Bal rloo Physics
L. Szal
A. Szalay, Johns Hopkins Physics

- Kernel density estimator $O(N^2)$
- n-point spatial statistics $O(N^n)$
- Nonparametric Bayes classifier $O(N^2)$
- Support vector machine $O(N^2)$
- Nearest-neighbor statistics $O(N^2)$
- Gaussian process regression $O(N^3)$
- Hierarchical clustering $O(N^3)$

But I have 1 million points

Machine learning/ statistics guy

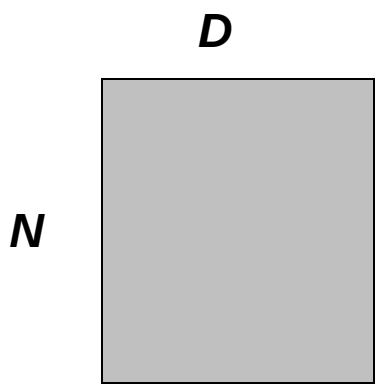
Statistics/learning challenges

Statistical (modeling, validation)**:**

- *Best performance with fewest assumptions*

Computational:

- *Large **N** (#data), **D** (#features)*



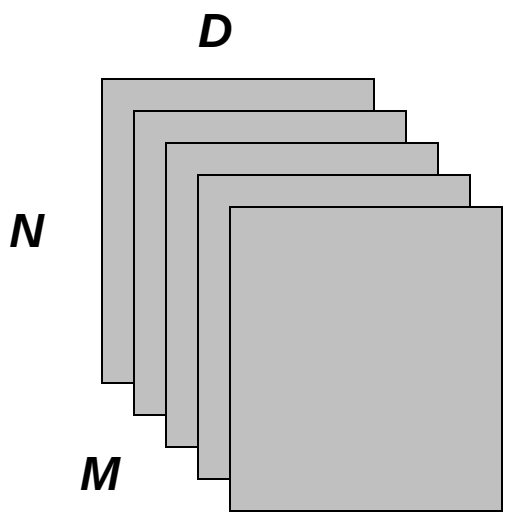
Statistics/learning challenges

Statistical (modeling, validation)**:**

- *Best performance with fewest assumptions*

Computational:

- *Large **N** (#data), **D** (#features), **M** (#models)*



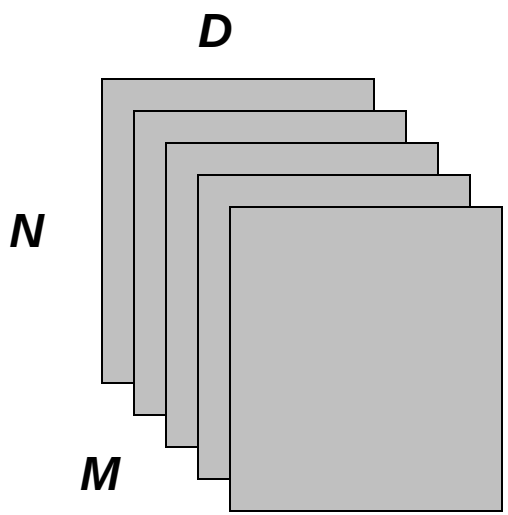
Statistics/learning challenges

Statistical (modeling, validation)**:**

- *Best performance with fewest assumptions*

Computational:

- *Large **N** (#data), **D** (#features), **M** (#models)*



Reduce? Simplify? **Poor modeling**

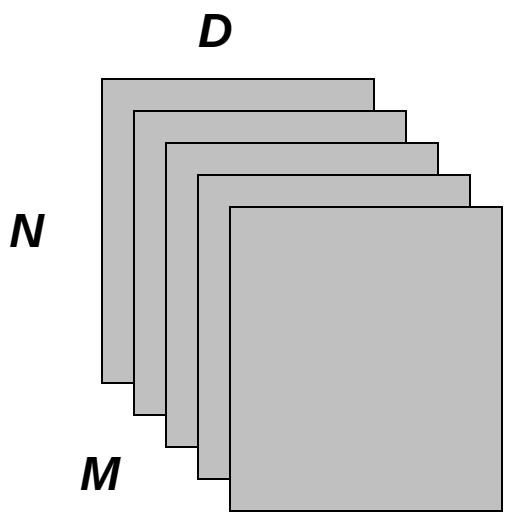
Statistics/learning challenges

Statistical (modeling, validation)**:**

- *Best performance with fewest assumptions*

Computational:

- *Large **N** (#data), **D** (#features), **M** (#models)*



Reduce? Simplify? **Poor modeling**

Avoid hard problems? **Poor funding**

My motivating datasets

- 1993-1999: POSS-II
- 1999-2008: SDSS
- Coming: Pan-STARRS, LSST
- Also:
  - Millennium simulation data
  - Large Hadron Collider data
  - network traffic (email) data
  - Inbio ecology data

What I like to think about…

- The **statistical problems and methods** needed for answering scientific questions
- The **computational problems and methods** involved in scaling them up to big datasets
- MLPACK: **software** for large-scale machine learning (later in 2008)

OUTLINE

1. What are some of the **statistical problems and methods** to consider?
2. What are some of the **computational problems and methods** to consider?
3. What might the **software** which implements all this look like?

OUTLINE

1. What are some of the **statistical problems and methods** to consider?
2. What are some of the **computational problems and methods** to consider?
3. What might the **software** which implements all this look like?

10 data analysis problems, and scalable tools we'd like for them

- **Querying** *(e.g. characterizing a region of space, defining a trigger)***:** nearest-neighbor, spherical range-search, orthogonal range-search
- **Density estimation** *(e.g. comparing galaxy types)***:** kernel density estimation, mixture of Gaussians
- **Regression** *(e.g. optical redshifts)**:*** linear regression, kernel regression, Gaussian process regression

10 data analysis problems, and scalable tools we'd like for them

- **Classification** *(e.g. quasar detection, star-galaxy separation)***:** nearest-neighbor classifier, nonparametric Bayes classifier, support vector machine
- **Dimension reduction** *(e.g. galaxy characterization)***:** principal component analysis, kernel PCA, maximum variance unfolding
- **Outlier detection** *(e.g. new object types, data cleaning)***:** by robust L_2 estimation, by density estimation, by dimension reduction

10 data analysis problems, and scalable tools we'd like for them

- **Clustering** *(e.g. automatic Hubble sequence)***:** k-means, hierarchical clustering ("friends-of-friends"), by dimension reduction
- **Time series analysis** *(e.g. asteroid tracking, variable objects)***:** Kalman filter, hidden Markov model, trajectory tracking
- **2-sample testing** *(e.g. cosmological validation)***:** n-point correlation
- **Cross-match** *(e.g. multiple databases)**:*** bipartite matching

OUTLINE

1. What are some of the **statistical problems and methods** to consider?
2. What are some of the **computational problems and methods** to consider?
3. What might the **software** which implements all this look like?

Core computational problems

What are the basic mathematical operations, or bottleneck subroutines, can we focus on developing fast algorithms for?

Core computational problems

- **Aggregations**
- **Generalized N-body problems**
- **Graphical model inference**
- **Linear algebra**
- **Optimization**

Core computational problems

Aggregations, GNPs, graphical models, linear algebra, optimization

- **Querying:** nearest-neighbor, sph range-search, ortho range-search
- **Density estimation:** kernel density estimation, mixture of Gaussians
- **Regression:** linear regression, kernel regression, Gaussian process regression
- **Classification:** nearest-neighbor classifier, nonparametric Bayes classifier, support vector machine
- **Dimension reduction:** principal component analysis, kernel PCA, maximum variance unfolding
- **Outlier detection:** by robust L_2 estimation, by density estimation, by dimension reduction
- **Clustering:** k-means, hierarchical clustering ("friends-of-friends"), by dimension reduction
- **Time series analysis:** Kalman filter, hidden Markov model, trajectory tracking
- **2-sample testing:** n-point correlation
- **Cross-match:** bipartite matching

Aggregations

- **How it appears:** nearest-neighbor, sph range-search, ortho range-search
- **Common methods:** locality sensitive hashing, kd-trees, metric trees, disk-based trees
- **Mathematical challenges:** high dimensions, provable runtime
- **Mathematical topics:** computational geometry, randomized algorithms

Aggregations

- **Interesting method:** *Cover-trees [Beygelzimer et al 2004]*
  - Provable runtime
  - Consistently good performance, even in higher dimensions
- **Interesting method:** *Learning trees [Cayton et al 2007]*
  - Learning data-optimal data structures
  - Improves performance over kd-trees

Generalized N-body Problems

- **How it appears:** kernel density estimation, mixture of Gaussians, kernel regression, Gaussian process regression, nearest-neighbor classifier, nonparametric Bayes classifier, support vector machine, kernel PCA, hierarchical clustering, trajectory tracking, n-point correlation
- **Common methods:** FFT, Fast Gauss Transform, Well-Separated Pair Decomposition
- **Mathematical challenges:** high dimensions, strong error guarantee
- **Mathematical topics:** approximation theory, computational physics

Generalized N-body Problems

- **Interesting method:** *Generalized Fast Multipole Method, aka multi-tree methods [Gray et al. 2000-2008]*
  - Fastest practical algorithms for most of the problems to which it has been applied
  - Hard relative error bounds
  - Automatic parallelization (THOR: Tree-based Higher-Order Reduce)
  - Big astrophysics results (dark energy evidence *Science* 2003, cosmic magnification verification *Nature* 2005, 1M quasars 2008)

Graphical model inference

- **How it appears:** hidden Markov models, bipartite matching
- **Common methods:** belief propagation, expectation propagation
- **Mathematical challenges:** large cliques, upper and lower bounds, graphs with loops
- **Mathematical topics:** variational methods, statistical physics, turbo codes

Graphical model inference

- **Interesting method:** *Survey propagation [Mezard et al 2002]*
  - *Good results for combinatorial optimization*
  - *Based on statistical physics ideas*

Linear algebra

- **How it appears:** linear regression, Gaussian process regression, PCA, kernel PCA, Kalman filter
- **Common methods:** QR, Krylov
- **Mathematical challenges:** numerical stability, sparsity preservation
- **Mathematical topics:** linear algebra

Linear algebra

- **Interesting method:** *Monte Carlo SVD [Frieze, Drineas, et al. 1998-2008]*
  - *Sample either columns or rows, from squared length distribution*
  - *Relative error bounds*

Optimization

- **How it appears:** support vector machine, maximum variance unfolding, robust L_2 estimation
- **Common methods:** interior point, Newton's method
- **Mathematical challenges:** large number of variables / constraints
- **Mathematical topics:** optimization theory, linear algebra, convex analysis

Optimization

- **Interesting method:** *Sequential minimization optimization (SMO) [Platt 1999]*
  - Much more efficient than interior-point, for SVM QPs
- **Interesting method:** *Stochastic quasi-Newton [Schraudolf 2007]*
  - Does not require scan of entire data

Interaction between statistics and computation

- **Explicitly trade off** between statistical accuracy and runtime
- **Monte Carlo:** a statistical idea for computational purposes
- **Active learning**, aka design of experiments: choose the important points

OUTLINE

1. What are some of the **statistical problems and methods** to consider?
2. What are some of the **computational problems and methods** to consider?
3. What might the **software** which implements all this look like?

Keep in mind the machine

- **Memory hierarchy:** cache, RAM, out-of-core
- Dataset bigger than one machine: **parallel/distributed**
- Everything is becoming **multicore**

Keep in mind the overall system

- **Databases** can be more useful than ASCII files
- **Workflows** can be more useful than brittle perl scripts
- **Visual analytics** connects visualization/HCI with data analysis

Keep in mind the software complexity

- Automatic **code generation** (e.g. MapReduce)
- Automatic **tuning** (e.g. OSKI)
- Automatic **algorithm derivation** (e.g. AutoBayes, SPIRAL)

The end

Always looking for hard problems!

Alexander Gray
agray@cc.gatech.edu