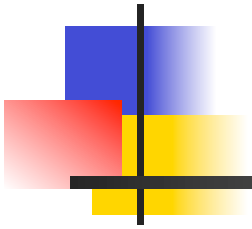


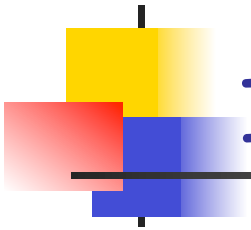
BIG Biomedicine and the Foundations of **BIG** Data Analysis

Michael W. Mahoney

ICSI and Dept of Statistics, UC Berkeley

May 2014

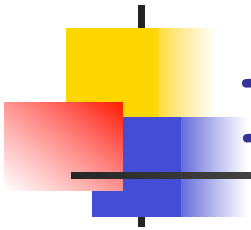
(For more info, see: http://www.stat.berkeley.edu/~mmahoney)



Insider's vs outsider's views (1 of 2)

Ques: Genetics vs molecular biology vs biochemistry vs biophysics:

- What's the difference?



Insider's vs outsider's views (1 of 2)

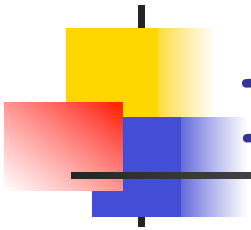
Ques: Genetics vs molecular biology vs biochemistry vs biophysics:

- What's the difference?

Answer: *Not much*, (if you are a "methods" person*)

- they are all biology
- you get data from any of those areas, ignoring important domain details, and evaluate your method qua method
- your reviewers evaluate the methods and don't care about the science
- ...

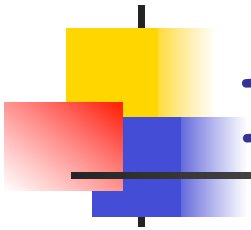
*E.g., one who self-identifies as doing data analysis or machine learning or statistics or theory of algorithms or artificial intelligence or ...



Insider's vs outsider's views (2 of 2)

Ques: Data analysis vs machine learning vs statistics vs theory of algorithms vs artificial intelligence (vs scientific computing vs computational mathematics vs databases ...):

- What's the difference?



Insider's vs outsider's views (2 of 2)

Ques: Data analysis vs machine learning vs statistics vs theory of algorithms vs artificial intelligence (vs scientific computing vs computational mathematics vs databases ...):

- What's the difference?

Answer: *Not much*, (if you are a "science" person*)

- they are all just tools
- you get a tool from any of those areas and bury details in a methods section
- your reviewers evaluate the science and don't care about the methods
- ...

*E.g., one who self identifies as doing genetics or molecular biology or biochemistry or biophysics or ...

BIG data??? MASSIVE data????

NYT, Feb 11, 2012: "The Age of Big Data"

- "What is Big Data? A meme and a marketing term, for sure, but also shorthand for advancing trends in technology that open the door to a new approach to understanding the world and making decisions. …"

Why are big data big?

- Generate data at different places/times and different resolutions
- Factor of 10 more data is not just more data, but *different data*

Thinking about large-scale data

Data generation is modern version of microscope/telescope:

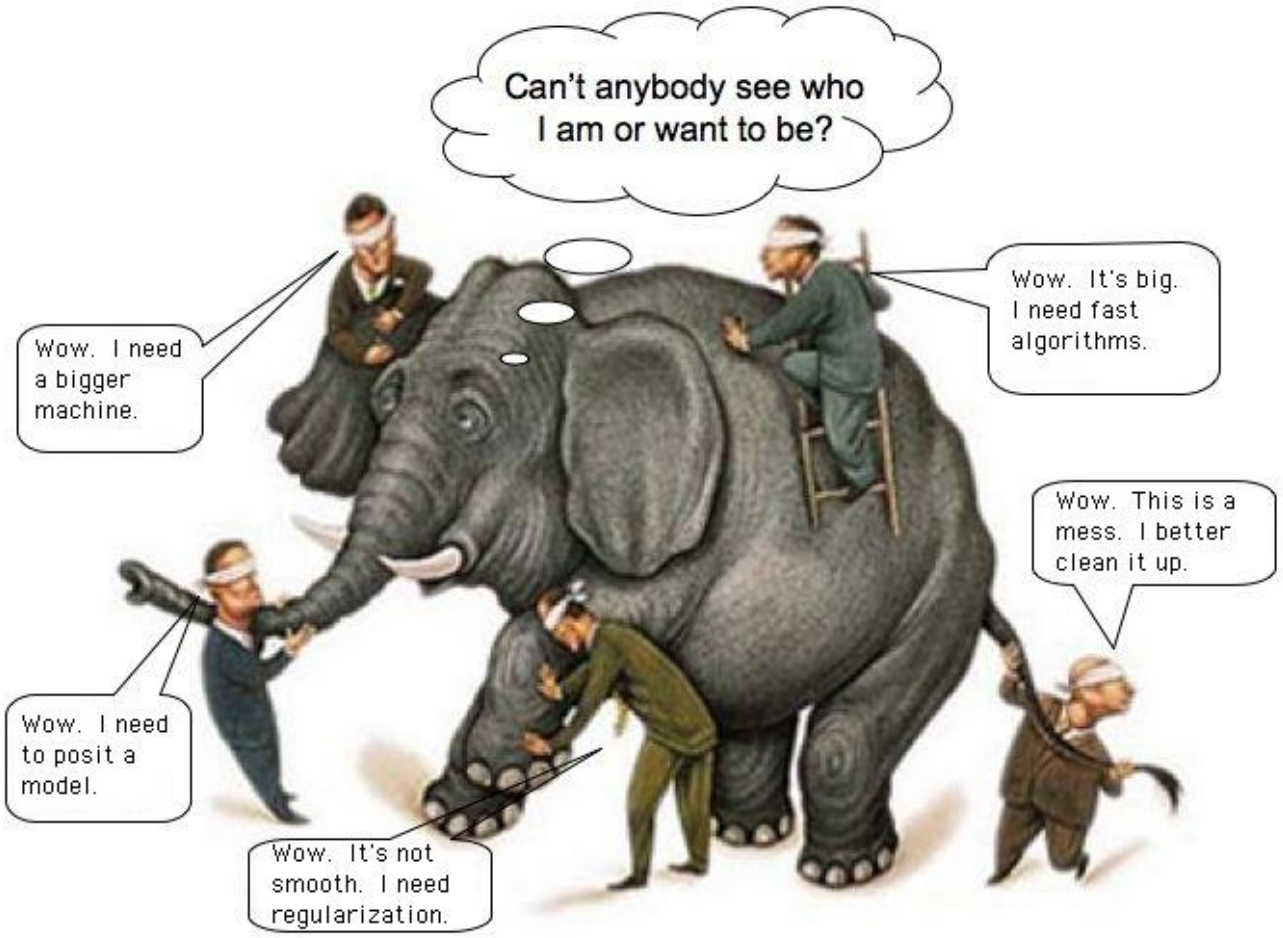
- See things couldn't see before: e.g., fine-scale movement of people, fine-scale clicks and interests; fine-scale tracking of packages; fine-scale measurements of temperature, chemicals, etc.
- Those inventions ushered new scientific eras and new understanding of the world and new technologies to do stuff

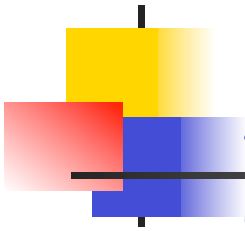
Easy things become hard and hard things become easy:

- Easier to see the other side of universe than bottom of ocean
- Means, sums, medians, correlations is easy with small data

Our ability to generate data far exceeds our ability to extract insight from data.

How do we view BIG data?





Algorithmic vs. Statistical Perspectives

Lambert (2000), Mahoney (2010)

Computer Scientists

- *Data*: are a record of everything that happened.
- *Goal*: process the data to find interesting patterns and associations.
- *Methodology*: Develop approximation algorithms under different models of data access since the goal is typically computationally hard.

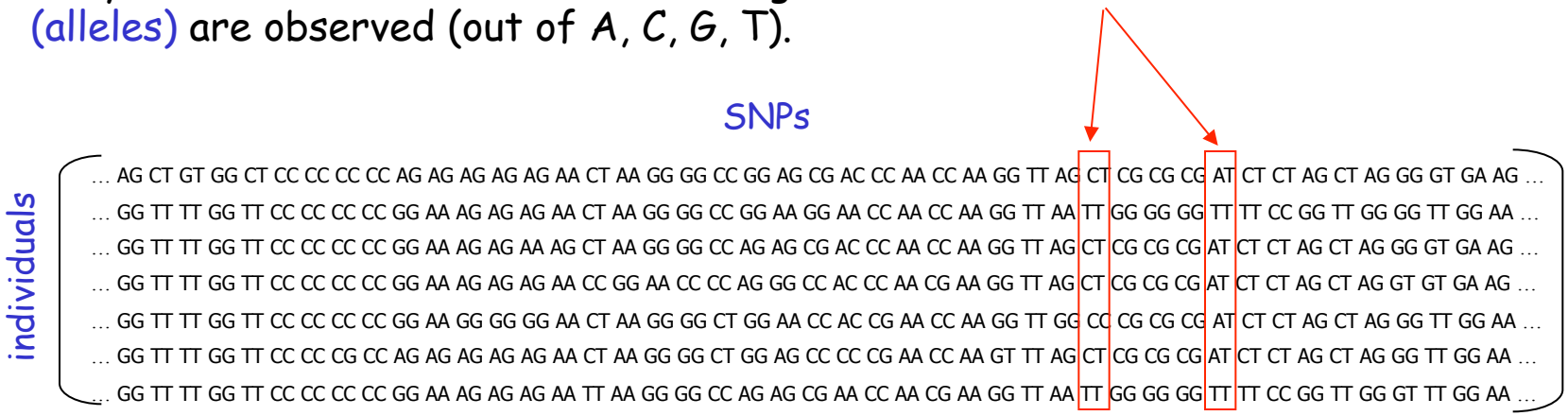
Statisticians (and Natural Scientists)

- *Data*: are a particular random instantiation of an underlying process describing unobserved patterns in the world.
- *Goal*: is to extract information about the world from noisy data.
- *Methodology*: Make inferences (perhaps about unseen events) by positing a model that describes the random variability of the data around the deterministic model.

Applications in: Human Genetics

Single Nucleotide Polymorphisms: the most common type of genetic variation in the genome across different individuals.

They are known locations at the human genome where two alternate nucleotide bases (alleles) are observed (out of A, C, G, T).

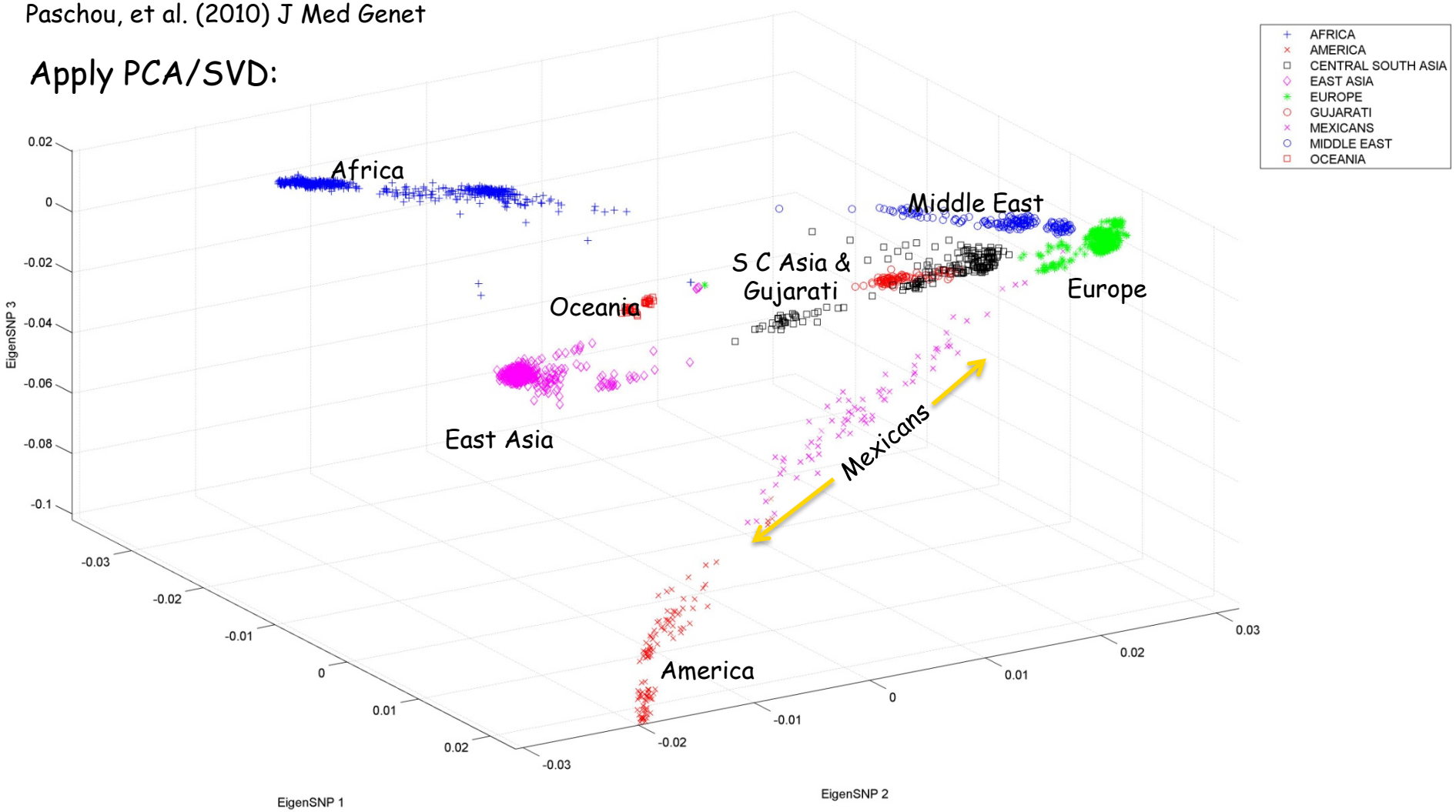


Matrices including thousands of individuals and hundreds of thousands if SNPs are available, and more/bigger/better are coming soon.

This can be written as a "matrix," assume it's been preprocessed properly, so let's call black box matrix algorithms.

Paschou, et al. (2010) J Med Genet

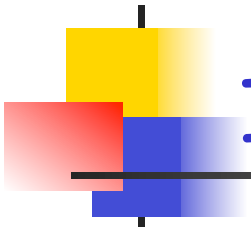
Apply PCA/SVD:



Not altogether satisfactory: the principal components are linear combinations of all SNPs, and – of course – can not be assayed!

Can we find **actual SNPs** that capture the information in the singular vectors?

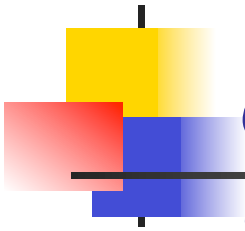
Formally: spanning the same subspace, optimizing variance, computationally efficient.



Issues with eigen-analysis

- **Computing large SVDs: computational time**
  - In commodity hardware (e.g., a 4GB RAM, dual-core laptop), using MatLab 7.0 (R14), the computation of the SVD of the dense 2,240-by-447,143 matrix A takes about 20 minutes.
  - Computing this SVD is not a one-liner, since we can not load the whole matrix in RAM (runs out-of-memory in MatLab). Instead, compute the SVD of AA^T.
  - In a similar" experiment," compute **1,200 SVDs** on matrices of dimensions (approx.) 1,200-by-450,000 (roughly, a full leave-one-out cross-validation experiment).

- **Selecting *actual columns* that "capture the structure" of the top PCs**
  - Combinatorial optimization problem; hard even for small matrices.
  - Often called the Column Subset Selection Problem (CSSP).
  - Not clear that such "good" columns even exist.
  - Avoid "reification" problem of "interpreting" singular vectors!



CUR matrix decompositions

Mahoney and Drineas "CUR Matrix Decompositions for Improved Data Analysis" (PNAS, 2009)

Goal. Solve the following problem:

"While very efficient basis vectors, the (singular) vectors themselves are completely artificial and do not correspond to actual (DNA expression) profiles. . . . Thus, it would be interesting to try to find basis vectors for all experiment vectors, using actual experiment vectors and not artificial bases that offer little insight." Kuruvilla et al. (2002)

Theorem:

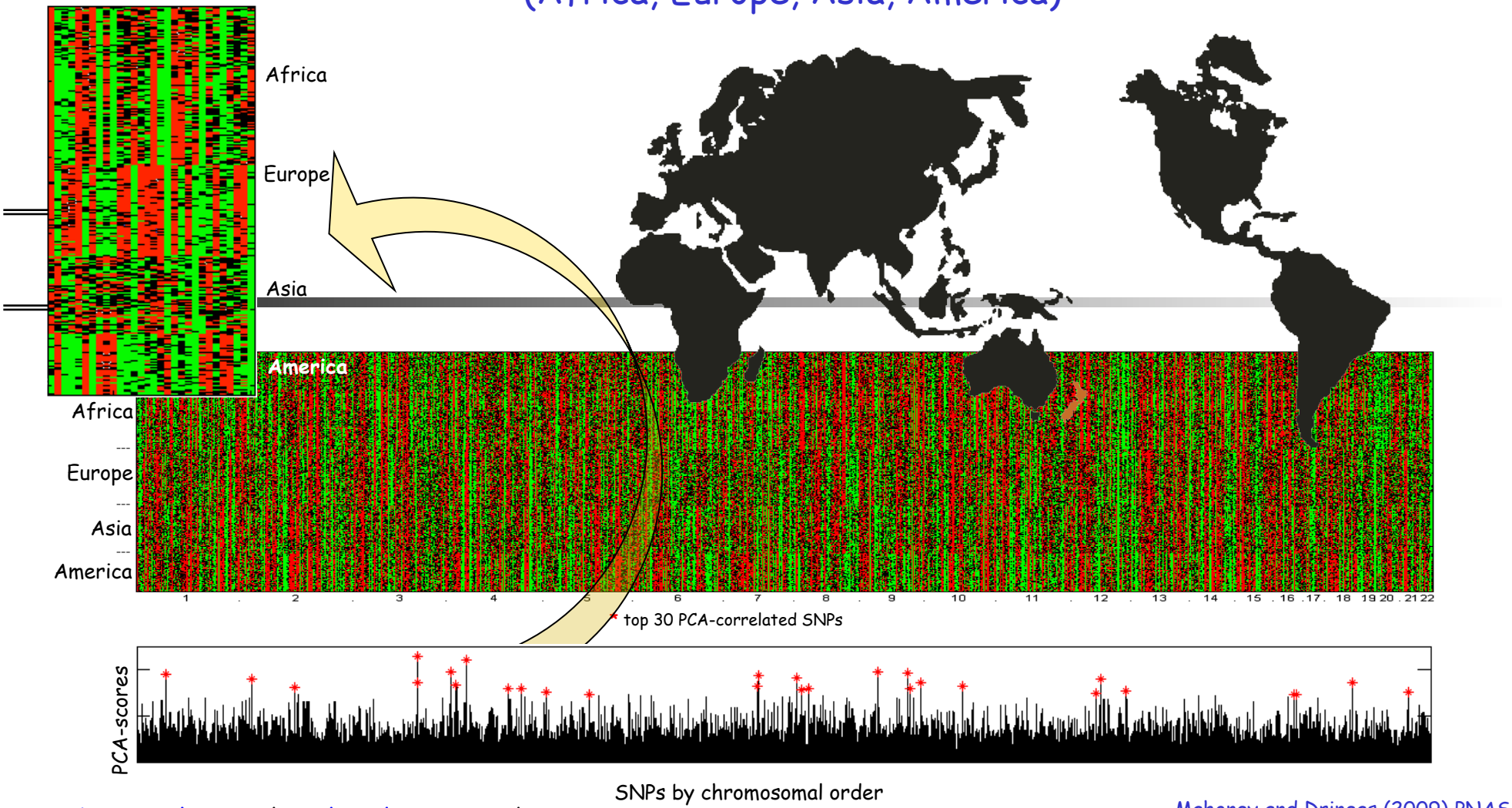
Given an arbitrary matrix, call a black box that I won't describe.

- You get a small number of actual columns/rows that are only marginally worse than the truncated PCA/SVD.
- The black box runs faster than computing a truncated PCA/SVD for arbitrary input.
- It's very robust to heuristic modifications.

Corollary:

We can use the same methods to approximate the PCA/SVD.

Selecting PCA SNPs for individual assignment to four continents (Africa, Europe, Asia, America)



- Data analysis and machine learning and statistics and theory of algorithms and scientific computing ... and genetics and astronomy and mass spectrometry and ... likes this---but each for different reasons!
- Good "hydrogen atom" for methods development!

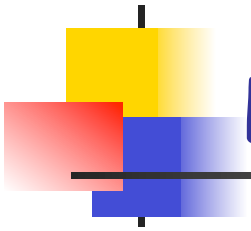
Mahoney and Drineas (2009) PNAS

Paschou et al (2007; 2008) PLoS Genetics

Paschou et al (2010) J Med Genet

Drineas et al (2010) PLoS One

Javed et al (2011) Annals Hum Genet



Bioinformatics: a cautionary tale?

• How did/does bioinformatics relate to computer science, statistics, and applied mathematics, "technically" and "sociologically"?

• How did NIH choose to fund graduate students and postdocs in the budget expansion of the 90s?

• What effect did this have on the number of American/foreign going into biomedical research?

• How will the pay structure of biomedical researchers effect which cs/stats "data scientists" engage you in your efforts?

• What effect does med schools deciding not to do joint faculty hires with cs departments have on bioinformatics and big biomedical data?

• How is this Big Biomedical Data phenomenon similar to and different than the Bioinformatics experience?

Big changes in the past ... and future

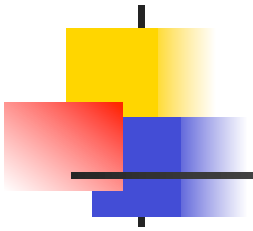
Consider the creation of:

- Modern Physics
- Computer Science
- Molecular Biology
- OR and Management Science
- Transistors and Microelectronics
- Biotechnology

These were driven by *new measurement techniques* and *technological advances*, but they led to:

- big new (academic and applied) questions
- new perspectives on the world
- lots of downstream applications

We are in the middle of a similarly big shift!



MMDS Workshop on "Algorithms for Modern Massive Data Sets"

(http://mmds-data.org)

at UC Berkeley, June 17-20, 2014

Objectives:

- Address algorithmic, statistical, and mathematical challenges in modern statistical data analysis.

- Explore novel techniques for modeling and analyzing massive, high-dimensional, and nonlinearly-structured data.

- Bring together computer scientists, statisticians, mathematicians, and data analysis practitioners to promote cross-fertilization of ideas.

Organizers: M. W. Mahoney, A. Shkolnik, P. Drineas, R. Zadeh, and F. Perez

Registration is available now!