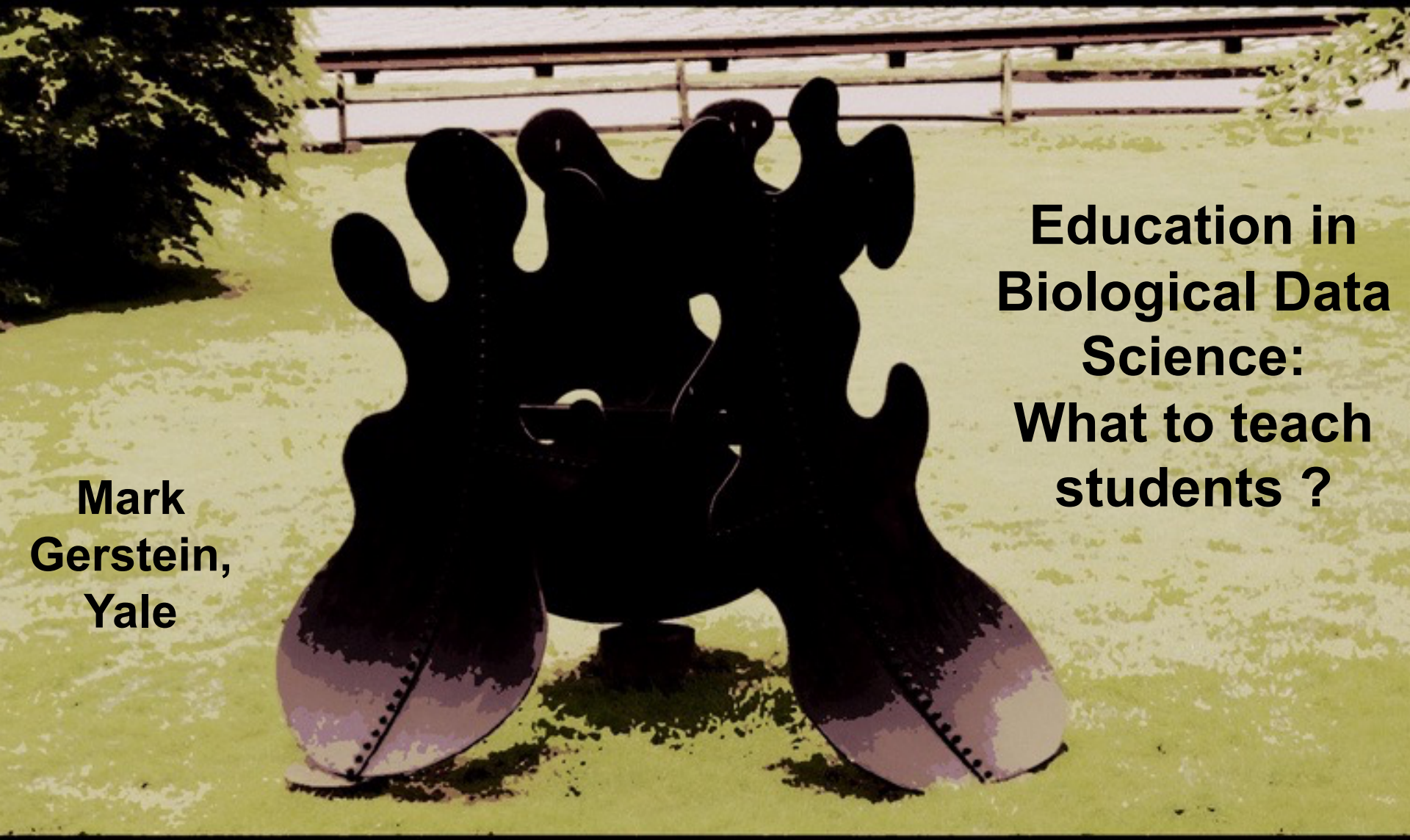


Education in Biological Data Science: What to teach students ?

Mark Gerstein, Yale

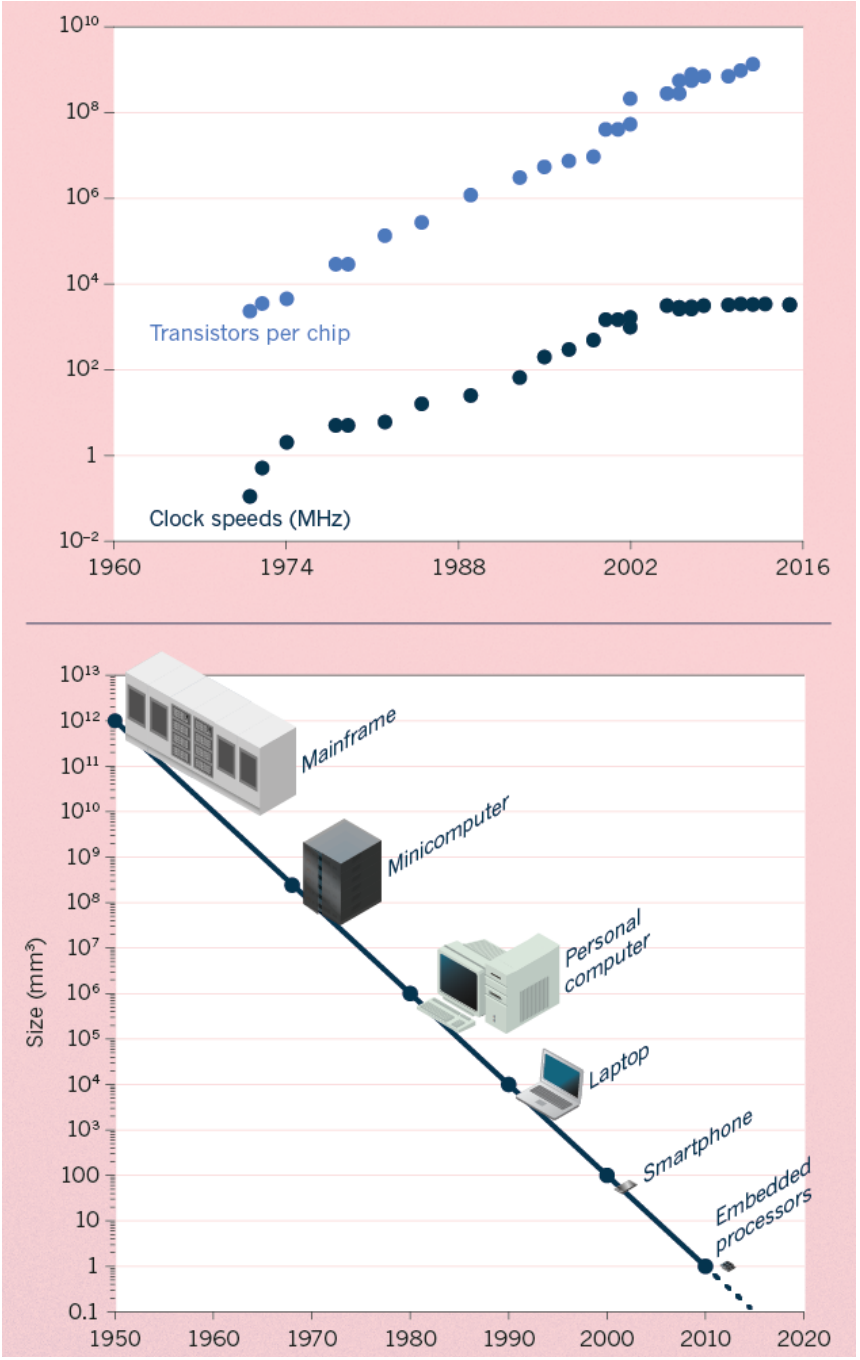
Slides freely downloadable from Lectures.GersteinLab.org
& “tweetable” (via @markgerstein). See last slide for more info.

Education in Biological Data Science: What to teach students ?

- Field Definers
  - **Q:** What is driving the development of biological data science as a field?
  - **Q:** How is biological data science related to other sub-disciplines of data science ?
- The Students
  - **Q:** What backgrounds do students of this discipline have?
  - **Q:** What careers are we preparing for?
- Curriculum
  - **Q:** Should we have standard curriculum ? What would we put into in?
  - **Q:** Does data science include physical modelling?
  - **Q:** What are good metaphors for the subject?

Moore's Law: Exponential Scaling of Computer Technology

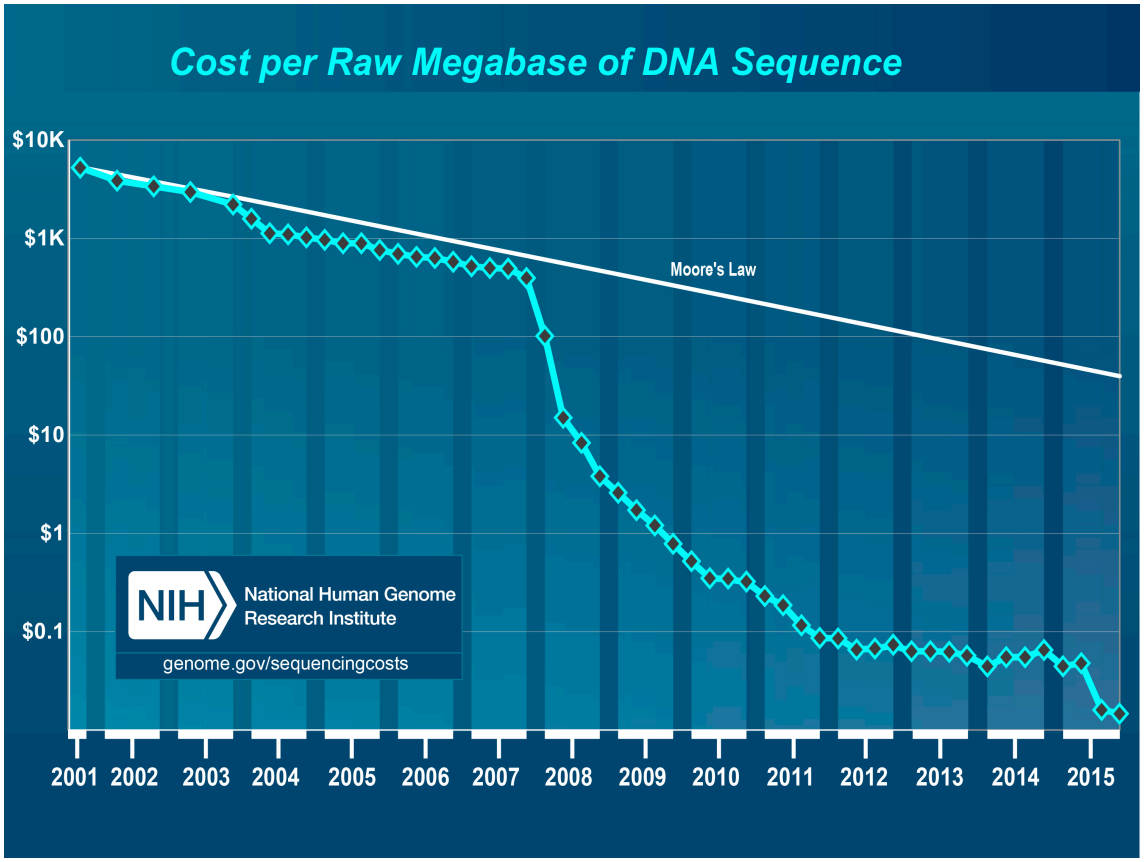
- Exponential increase in the number of transistors per chip.



[Waldrop ('15) Nature]

Sequencing Data Explosion: Faster than Moore's Law for a Time

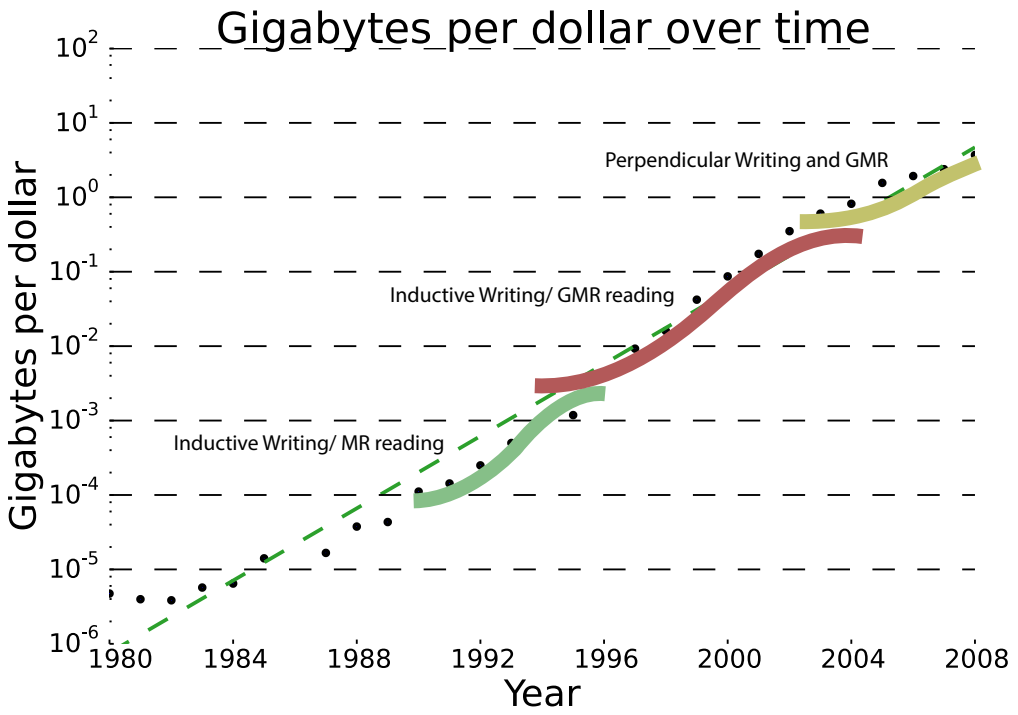
- In the early 2000's, improvements in Sanger sequencing produced a scaling pattern similar to Moore's law.
- The advent of NGS was a shift to a new technology with dramatic decrease in cost).

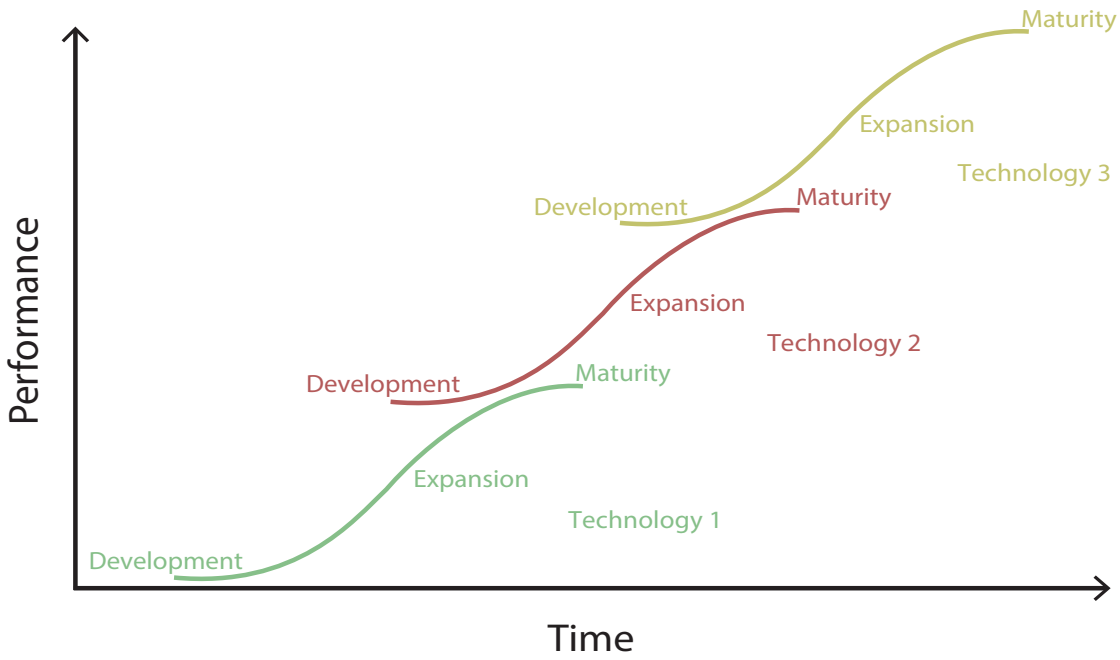


Kryder's Law and S-curves underlying exponential growth

- Moore's & Kryder's Laws
- Exponential increase seen in Kryder's law is a superposition of S-curves for different technologies

- Does this apply to Next-gen sequencing ?

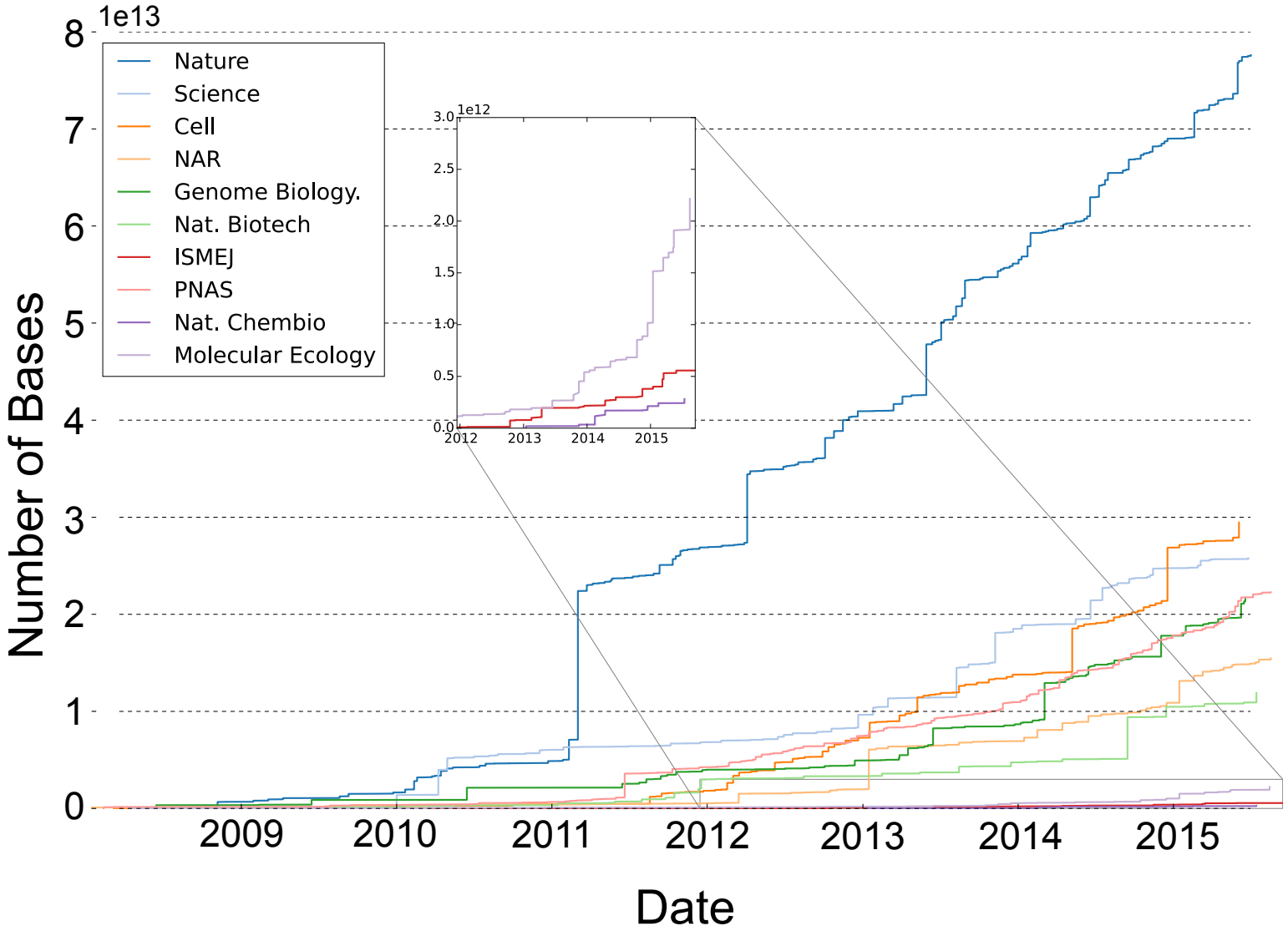




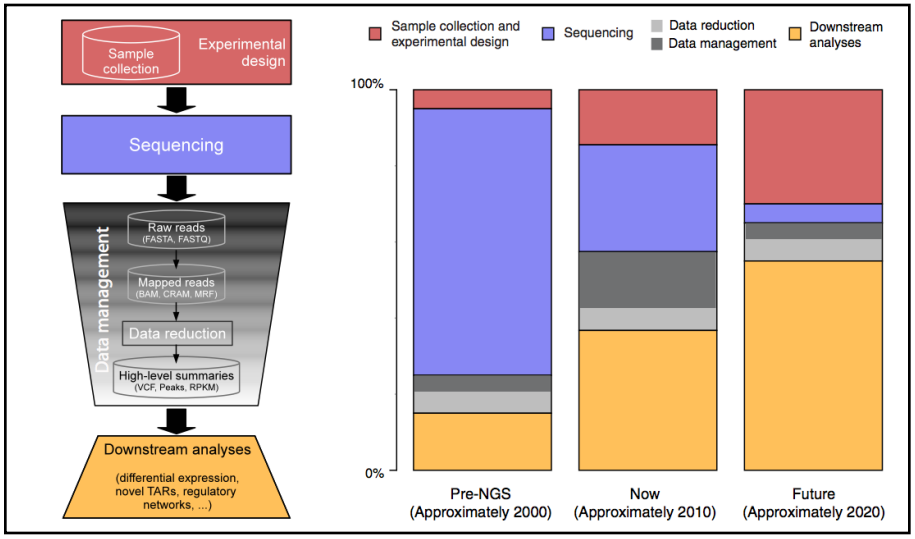
[Muir et al. ('15) GenomeBiol.]

The well known data explosion and not as appreciated increased diversity in sequence data sources

[Muir et al. ('15) GenomeBiol.]



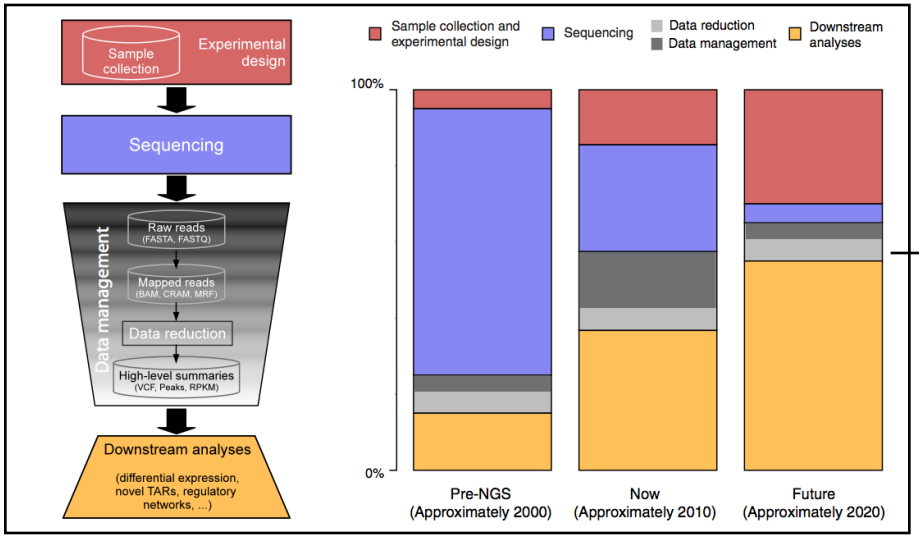
The changing costs of a sequencing pipeline

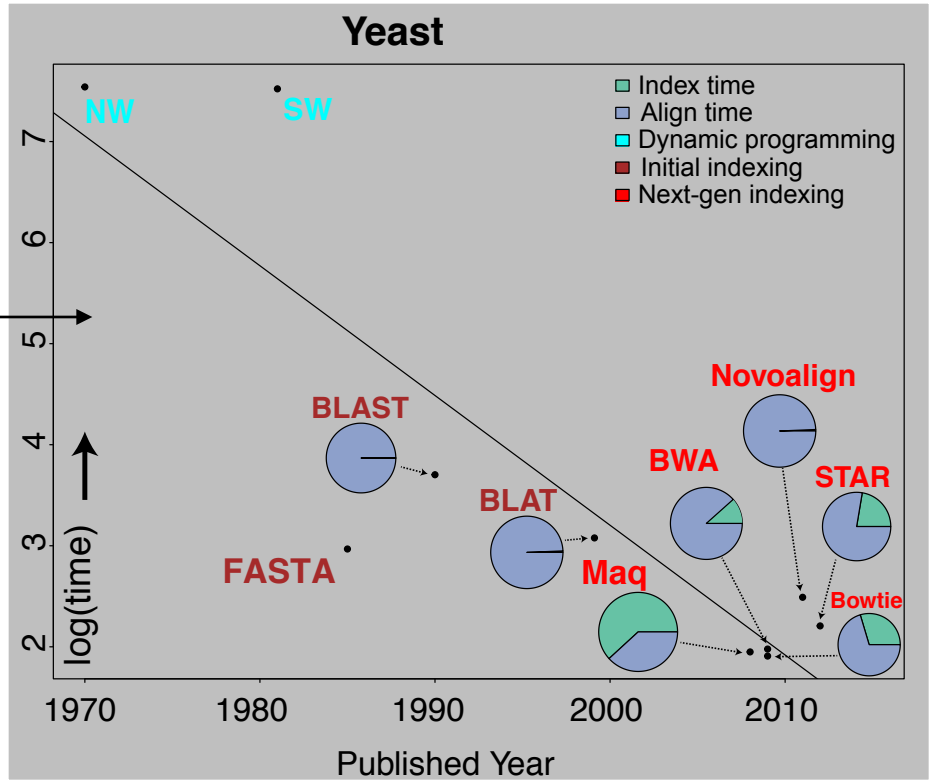


From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

[Sboner et al. (‘11), Muir et al. (‘15) Genome Biology]

The changing costs of a sequencing pipeline



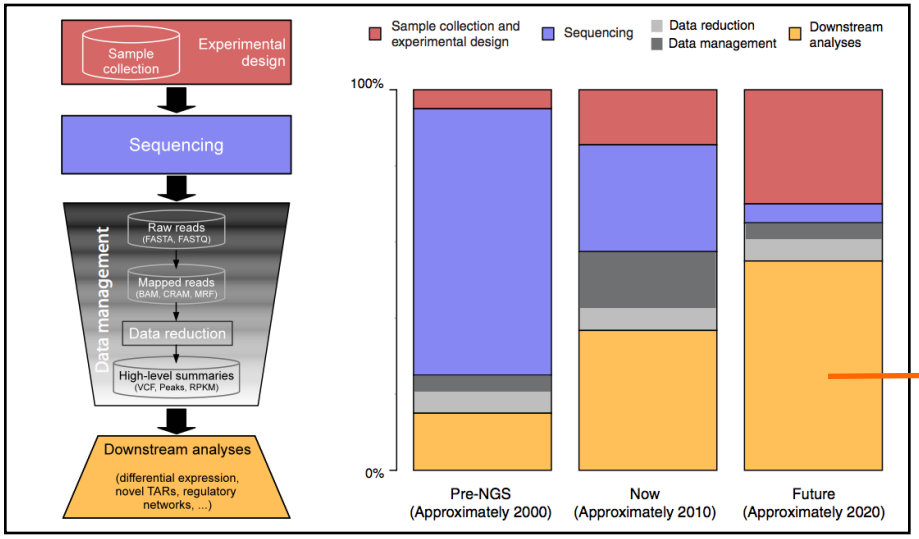


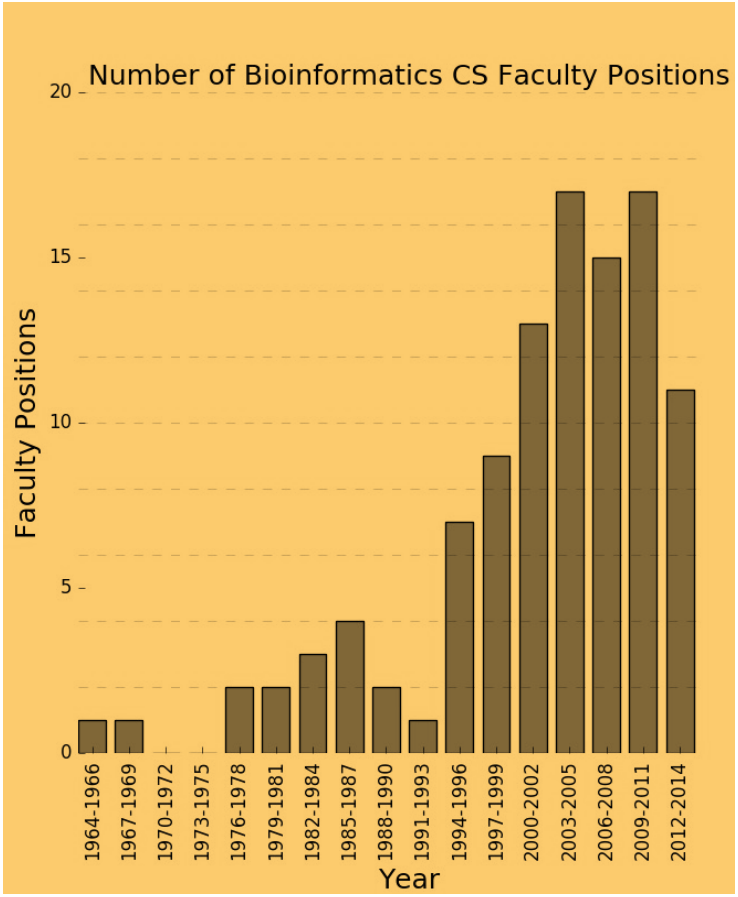
From '00 to ~' 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

Alignment algorithms scaling to keep pace with data generation

[Sboner et al. ('11), Muir et al. ('15) Genome Biology]

The changing costs of a sequencing pipeline

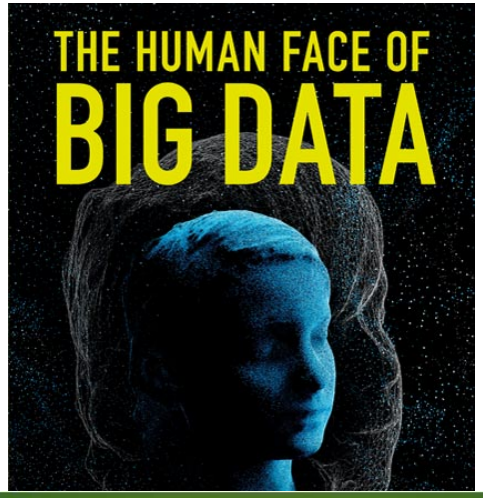


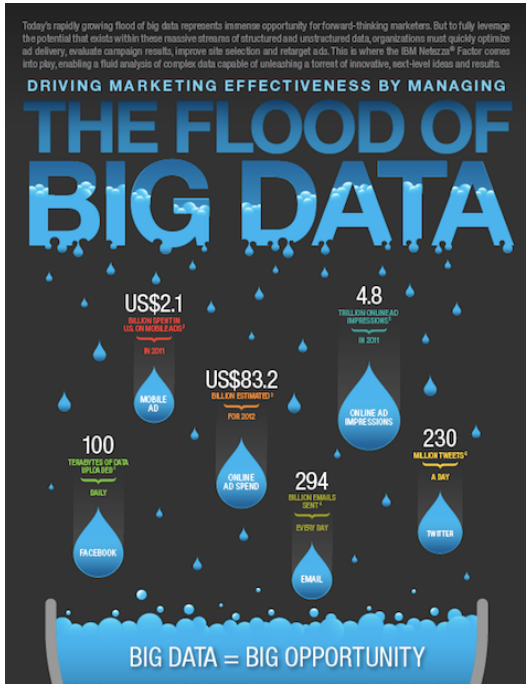


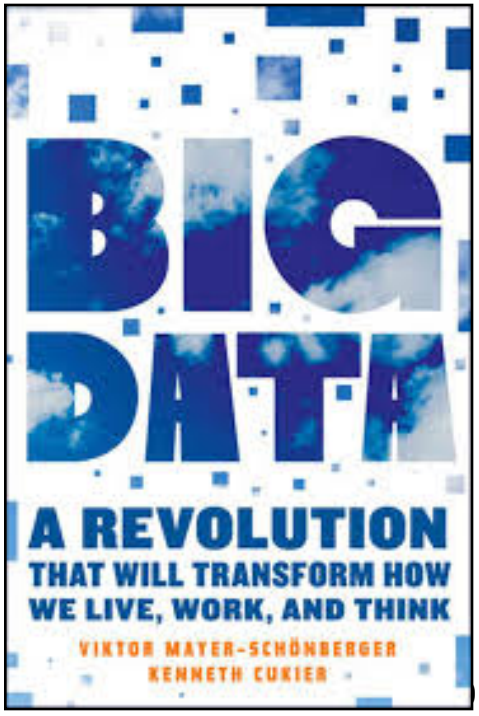
From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

[Sboner et al. (‘11), Muir et al. (‘15) Genome Biology]

Commercial World Data: Financial & Retail Data







Big Data:

a current buzz-word

Harvard Business Review

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

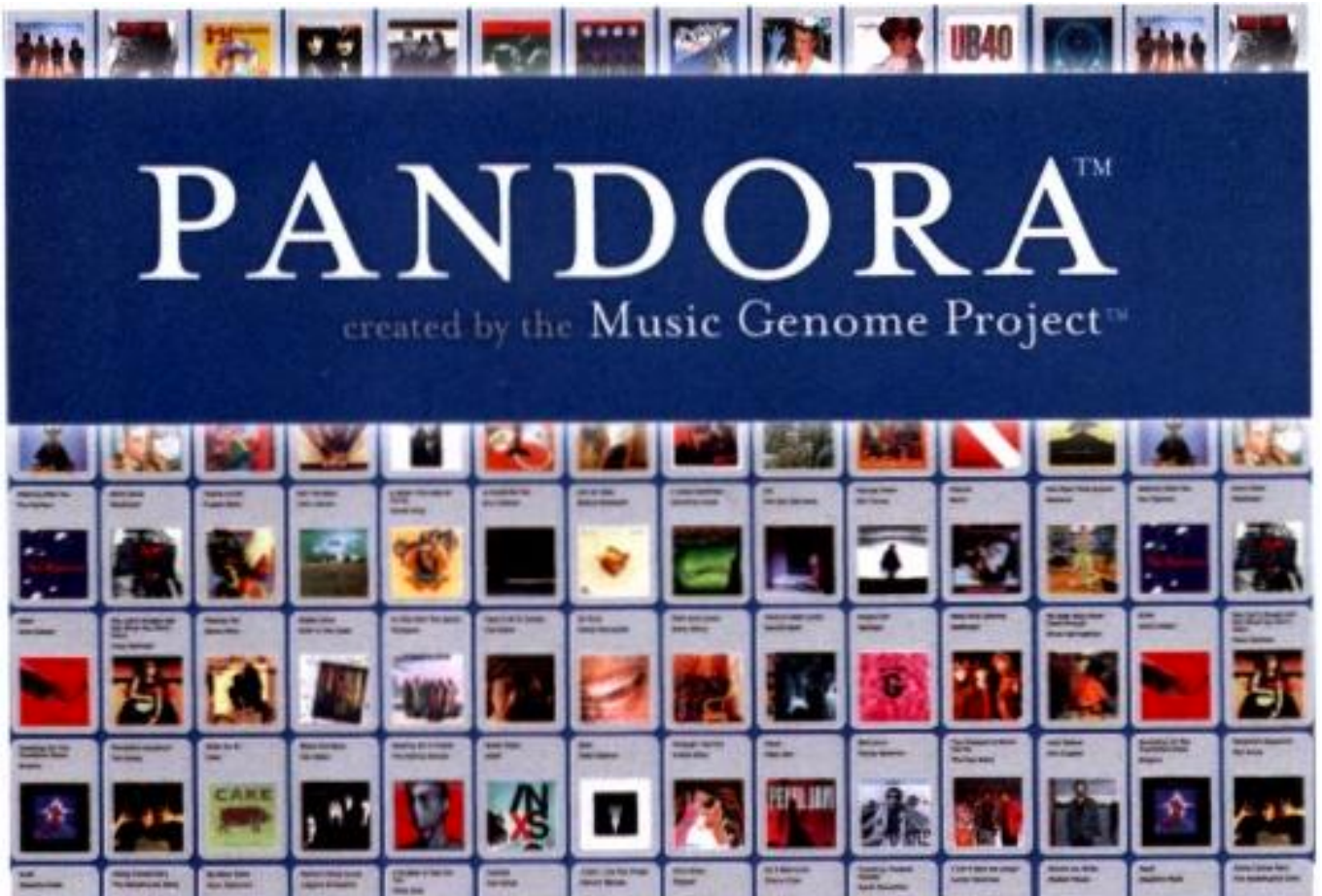
Artwork: **Tamar Cohen, *Andrew J Buboltz,*** 2011, silk screen on a page from a high

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business ne
up. The company had just under 8 million accounts, and the number was growing qu
friends and colleagues to join. But users weren't seeking out connections with the pe
rate executives had expected. Something was apparently missing in the social expe

[Oct. '12 issue]

Genomics: as an exemplar Data Science sub-discipline

- Developing ways of organizing & mining genomic information on a large scale
  - Very fundamental & early form of "Big Data"
- Perhaps we can learn from other data science disciplines
  &, in turn, teach them how to do this?

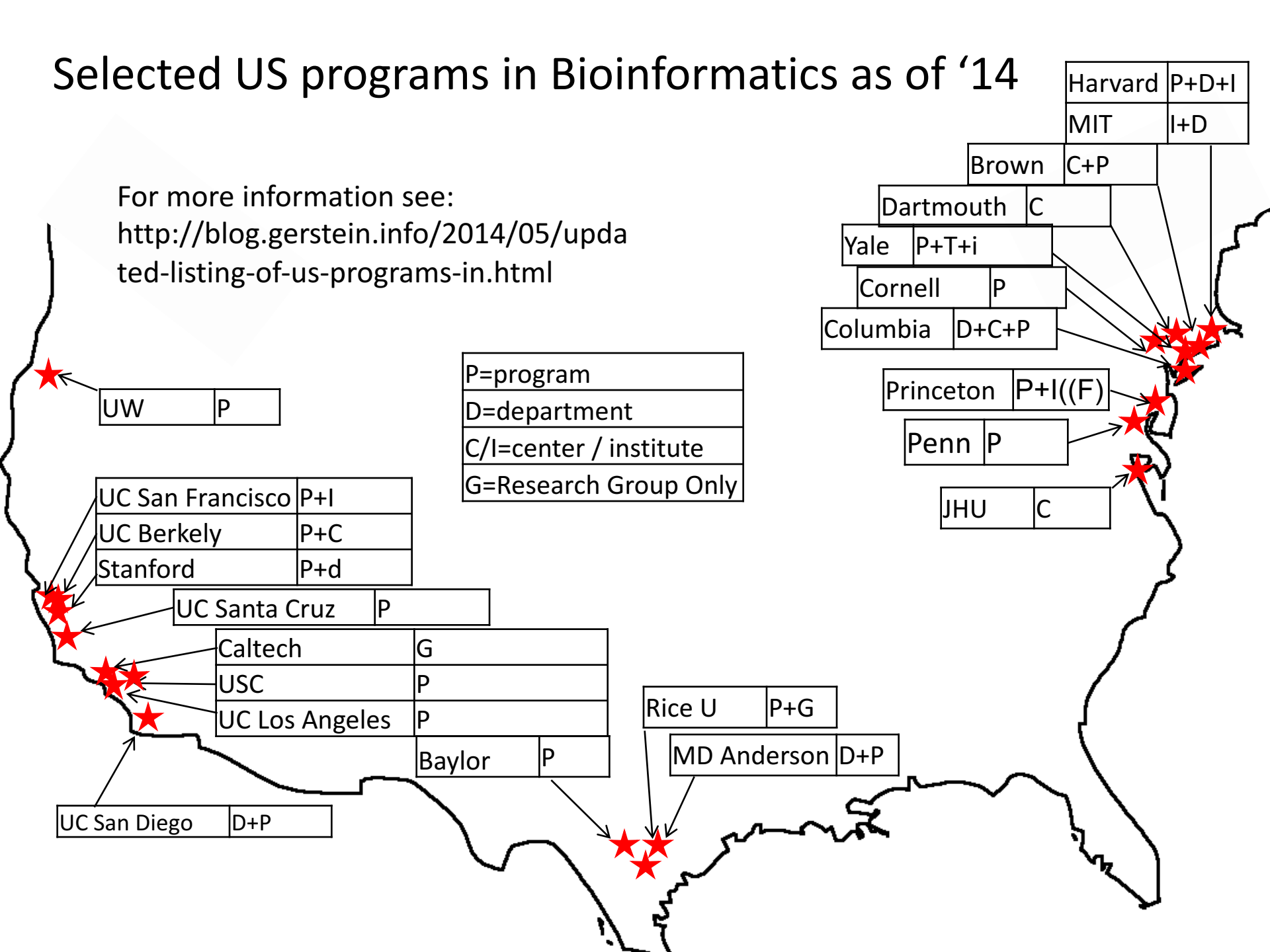


Education in Biological Data Science: What to teach students ?

- Field Definers
  - **Q:** What is driving the development of biological data science as a field?
  - **Q:** How is biological data science related to other sub-disciplines of data science ?
- The Students
  - **Q:** What backgrounds do students of this discipline have?
  - **Q:** What careers are we preparing for?
- Curriculum
  - **Q:** Should we have standard curriculum ? What would we put into in?
  - **Q:** Does data science include physical modelling?
  - **Q:** What are good metaphors for the subject?

Selected US programs in Bioinformatics as of '14

For more information see:
http://blog.gerstein.info/2014/05/updated-listing-of-us-programs-in.html



History & Current Structure of Yale Computational Biology PhD Program (CBB.yale.edu)

- History
  - Started in '02 1st as admissions track & in '03 then as a PhD granting program
- Key Numbers from 3 years ago
  - 77 matriculated, 34 graduated so far
  - ~7 students/yr (~40% non-US)

Inputs as of '14

- CBB Graduates – Undergrad Majors

Biology	Bioinformatics	Informatics	Other
19	3	15	5

- CBB Current Students – Undergrad Majors

Biology	Bioinformatics	Informatics	Other
18	8	8	1

Outputs

- Over last 7 yrs
- Some faculty, split betw. **more bio. oriented** & **bioinformatics** appointments
- Many in industry, split betw. **traditional bioinfo. route in biotech/pharma** & more general **"data-science" business** positions

Fac.

2003-2007	Assoc Professor, ASU
2002-2007	Asst Professor, UT
2005-2010	UCLA Lecturer
2009-2014	Asst Professor, UNC
2006-2012	Assoc Bioinformatics Scientist , Children's Hospital of Philadelphia

Postdoc

2002-2008	Postdoc, Stanford University
2002-2009	Postdoc, Dana Farber Institute
2004-2010	Resident in General Surgery, Yale
2007-2012	Computational Biologist, Broad Institute, MA
2007-2012	Postdoc, Stanford University
2008-2013	Postdoc, Stanford University
2006-2013	Programmer Anaylst II, Yale University

Industry

2002-2007	Sr. Bioinformatics Scientist, Illumina
2004-2009	Data Integration Officer, St. Jude, Memphis
2003-2010	Scientist, Celgene
2004-2010	Quantitative Trader, Laurion Capital Mgt
2005-2010	Director of Informatics, Bina Technologies Inc.
2005-2010	Investigator, Novartis Institutes for BioMedical Research
2004-2010	Sr. Developer, Schrodinger, Inc.
2006-2011	Assoc Principal Scientist, Merck Company
2005-2011	Product Manager & Bioinformatics Analyst, 5AM Solutions
2005-2011	Financial firm in Beijing
2006-2011	Quantitative Analyst, Google
2005-2011	Data Analyst/NLP Specialist, Elsevier
2007-2012	Lead Bioinformatics R&D Developer, Regeneron Pharmaceuticals Inc.
2006-2012	Software Developer, Berkeley Nat Lab
2009-2012	Information Technology and Services, Germany
2008-2013	Economic Modeling Senior, Freddie Mac
2007-2013	Analytics Consultant, SeqWise Next Generation Sequencing Consulting
2008-2014	Research Scientist, GE Global Research
2008-2014	Bioinformatics Scientist, Illumina
2009-2014	Senior Consulting Engineer, Attivio, Inc.

Education in Biological Data Science: What to teach students ?

- Field Definers
  - **Q:** What is driving the development of biological data science as a field?
  - **Q:** How is biological data science related to other sub-disciplines of data science ?
- The Students
  - **Q:** What backgrounds do students of this discipline have?
  - **Q:** What careers are we preparing for?
- Curriculum
  - **Q:** Should we have standard curriculum ? What would we put into in?
  - **Q:** Does data science include physical modelling?
  - **Q:** What are good metaphors for the subject?

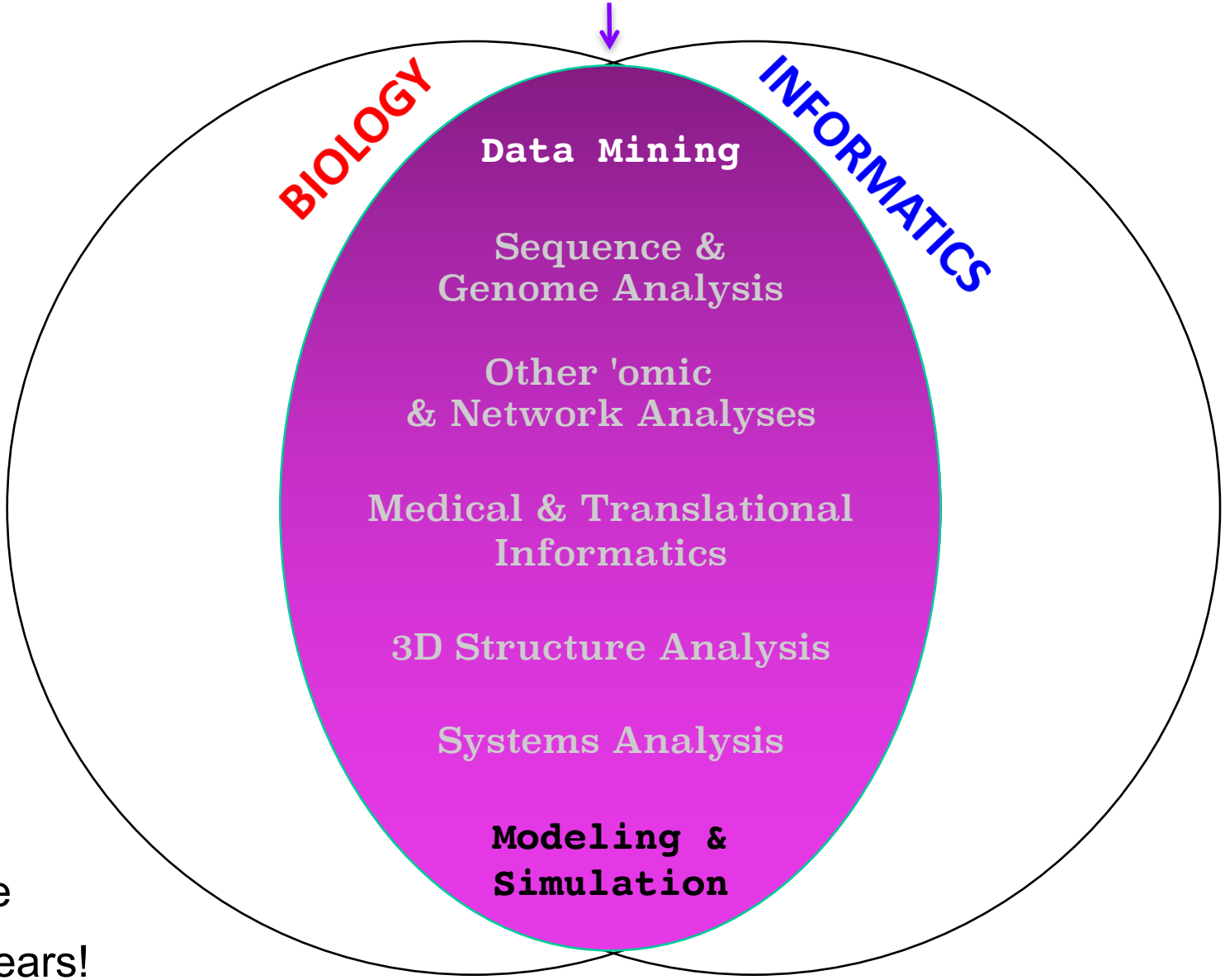
Curriculum: Courses & Competency in Core CBB, Biological Sciences & Informatics

- 10 Courses in Three Core Areas of Competency
  - Computational Biology & Bioinformatics (3 grad courses)
    - **CBB 752b Bioinformatics: Practical Applications of Simulation & Data Mining**
    - CBB 740a Clinical and Translational Informatics
    - CBB 562a Dynamical Systems in Biology
  - Biological sciences (2 grad courses)
  - Informatics - e.g., CS, stats, app. math (2 grad courses)
  - Electives (2 undergrad or grad courses, in any of the above)

- Competency of incoming students (need to take courses to get to this level)
  - Biology & Natural Science: introductory biology, biochemistry, chemistry
  - CS: introduction to CS, data structures & programming techniques
  - Math & Stat: introduction to probability and statistical inference, multivariate calculus and linear algebra

[More detail in Gerstein et al. ('07) *J Biomed. Inf.*]

Bioinformatics: Practical Applications of Simulation & Data Mining



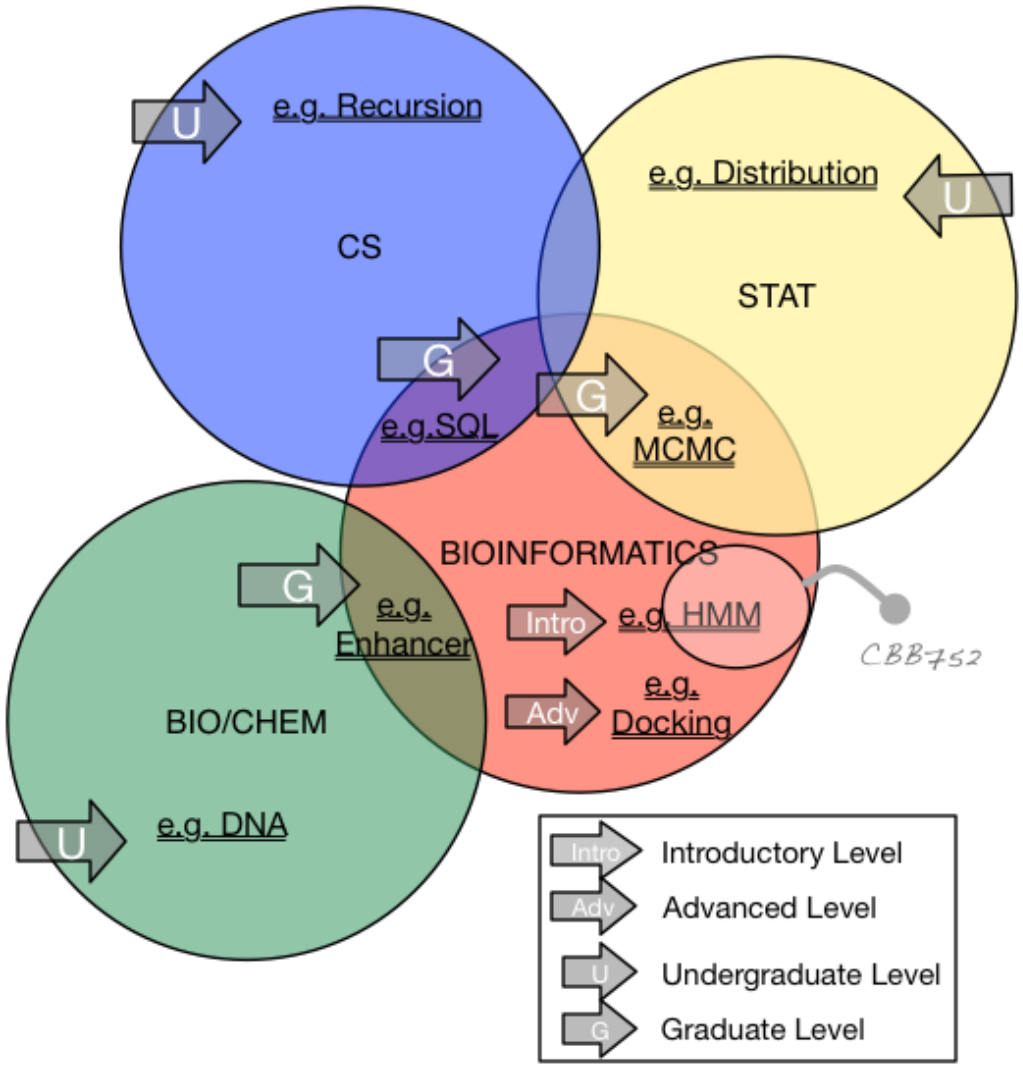
In existence for **18** years!

Gersteinlab.org/courses/452

Changing name in '17 to => Biomedical Data Science: Mining & Modelling

Defining Bioinformatics – by crowd-sourced judgement

- Bioinformatics
  - Related terms
    - Biological Data Science
    - Bioinformatics & / or / vs Computational Biology
    - Biocomputing
    - Systems Biology
    - Qbio
- What are its boundaries
  - Determining the "Support Vectors"



- Topic list & crowded-sourced comments at goo.gl/**303KXr**

Are They or Aren't They Comp. Bio.? (#1)

- (Digital Libraries & Medical Record Analysis
  - Automated Bibliographic Search and Textual Comparison
  - Knowledge bases for biological literature
- (Motif Discovery Using Gibb's Sampling
- (Methods for Structure Determination
  - Computational Crystallography
    - Refinement
  - NMR Structure Determination
    - (Distance Geometry
- (Metabolic Pathway Simulation
- (The DNA Computer

Are They or Aren't They Comp. Bio.? (#1, Answers)

- (YES?) Digital Libraries & Medical Record Analysis
  - Automated Bibliographic Search and Textual Comparison
  - Knowledge bases for biological literature
- (YES) Motif Discovery Using Gibb's Sampling
- (NO?) Methods for Structure Determination
  - Computational Crystallography
    - Refinement
  - NMR Structure Determination
    - (YES) Distance Geometry
- (YES) Metabolic Pathway Simulation
- (NO) The DNA Computer

Are They or Aren't They Comp. Bio.? (#2)

- (Gene identification by sequence characteristics
  - Prediction of splice sites
- (DNA methods in forensics
- (Modeling of Populations of Organisms
  - Ecological Modeling (predator & prey)
- (Modeling the nervous system
  - Computational neuroscience
  - Understanding how brains think & using this to make a better computer
- (Molecular phenotype discovery – looking for gene expression signatures of cancer
  - What if it included non-molecular data such as age ?

Are They or Aren't They Comp. Bio.? (#2, Answers)

- **(YES)** Gene identification by sequence characteristics
  - Prediction of splice sites
- **(YES)** DNA methods in forensics
- **(NO)** Modeling of Populations of Organisms
  - Ecological Modeling (predator & prey)
- **(NO?)** Modeling the nervous system
  - Computational neuroscience
  - Understanding how brains think & using this to make a better computer
- **(YES)** Molecular phenotype discovery – looking for gene expression signatures of cancer
  - What if it included non-molecular data such as age ?

Are They or Aren't They Comp. Bio.? (#3)

- (RNA structure prediction
- (Radiological Image Processing
  - Computational Representations for Human Anatomy (visible human)
- (Artificial Life Simulations
  - Artificial Immunology / Computer Security
  - (Genetic Algorithms in molecular biology
- (Homology Modeling & Drug Docking
- (Char. drugs & other small molecules (QSAR)
- (Computerized Diagnosis based on Pedigrees
- (Processing of NextGen sequencing image files
- (Module finding in protein networks

Are They or Aren't They Comp. Bio.? (#3, Answers)

- **(YES)** RNA structure prediction
- **(NO)** Radiological Image Processing
  - Computational Representations for Human Anatomy (visible human)
- **(NO)** Artificial Life Simulations
  - Artificial Immunology / Computer Security
  - **(NO?)** Genetic Algorithms in molecular biology
- **(YES)** Homology Modeling & Drug Docking
- **(YES)** Char. drugs & other small molecules (QSAR)
- **(NO)** Computerized Diagnosis based on Pedigrees
- **(NO)** Processing of NextGen sequencing image files
- **(YES)** Module finding in protein networks

General Types of "Informatics" techniques in Computational Biology – a mix between mining & modeling

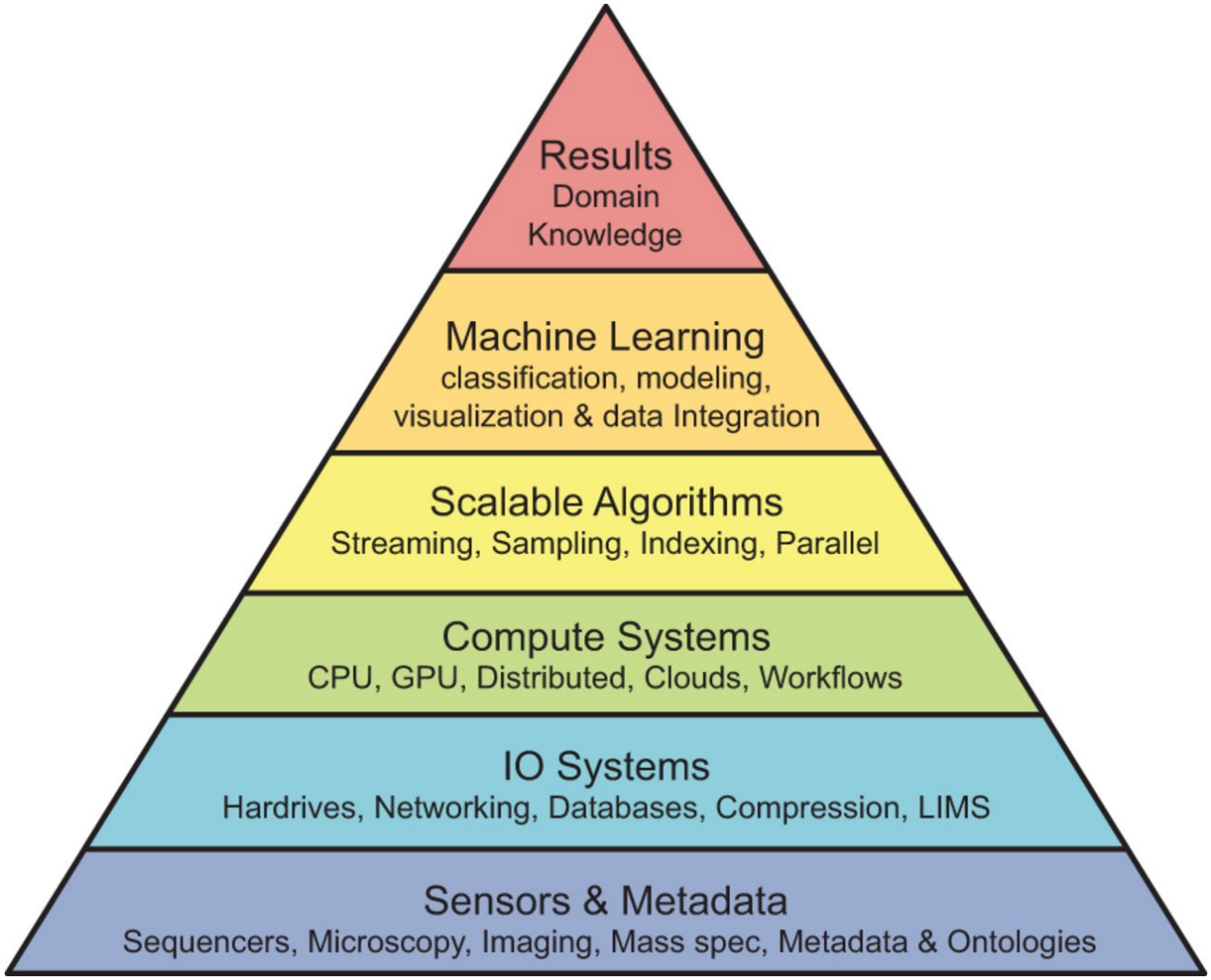
- **Databases**
  - Building, Querying
  - Representing Complex data
- **Data mining**
  - Machine Learning techniques
  - Clustering & Tree construction
  - Rapid Text String Comparison & textmining
  - Detailed statistics of significance & association
- **Network Analysis**
  - Analysis of Topology (eg Hubs)
  - Predicting Connectivity
- **Structure Analysis & Geometry**
  - Graphics (Surfaces, Volumes)
  - Comparison & 3D Matching (Vision, recognition, docking)
- **Physical Modeling**
  - Newtonian Mechanics
  - Minimization & Simulation
  - Modeling Chemical Reactions & Cellular Processes

Weather forecasing as a model for bioinformatics: successfully fusing large-scale data with physical models to create useful predictions

- Lampooned but actually very successful
  - No ability to predict a century ago, & now forecasts checked by billions every day
  - Interpretable & useful statistical predictions, informing everything from clothing choices to commerce

- How do they do it?
  - Physical models & massive simulation useful (but not sufficient - think “butterfly” effect.)
  - Large-scale data collection via sensors

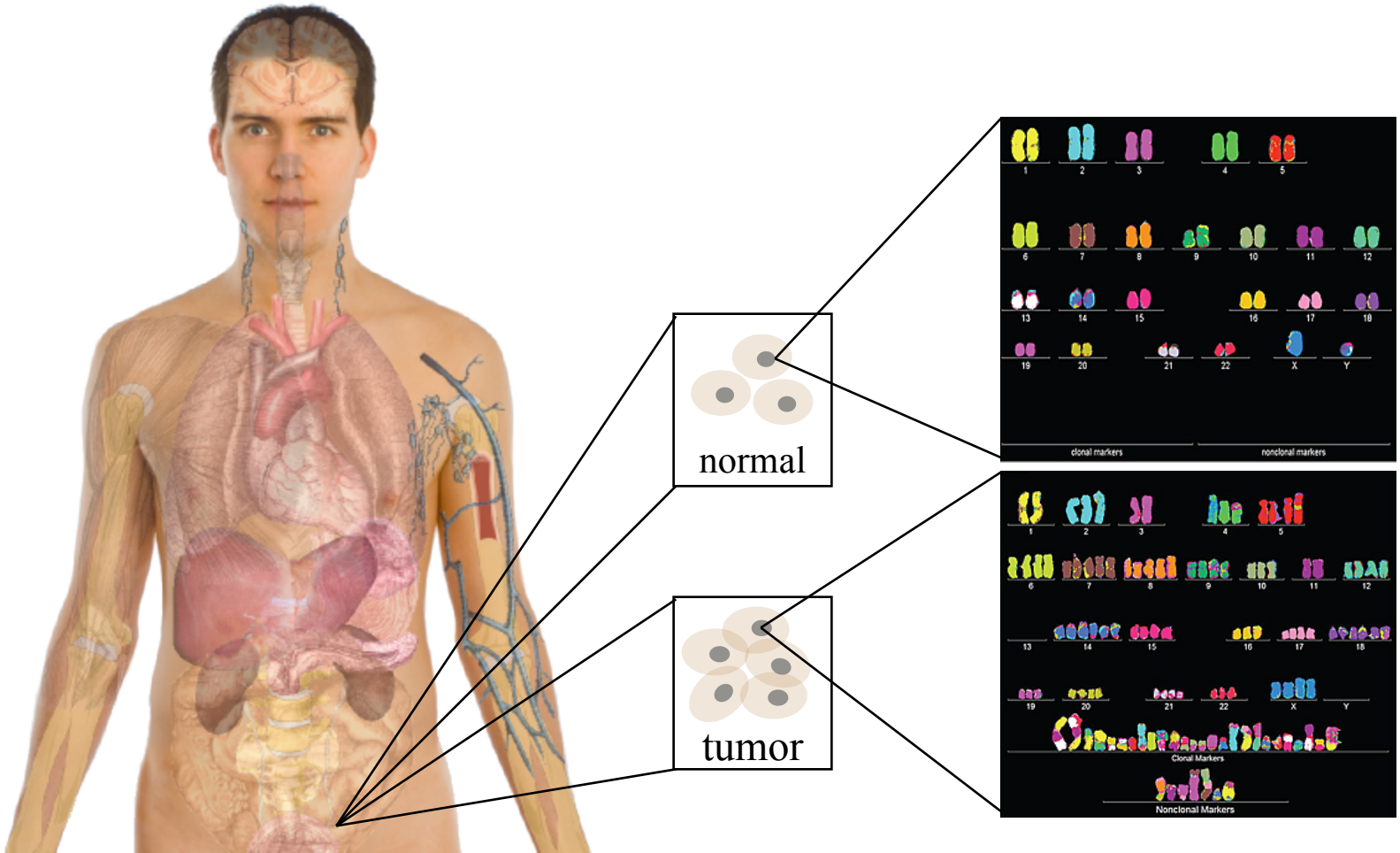
Data science analysis stack.



Michael C. Schatz Genome Res. 2015;25:1417-1422

Our field as future Gateway – Personal Genomics as a Gateway into Biology

Personal genomes soon will become a commonplace part of medical research & eventually treatment (esp. for cancer). They will provide a primary connection for biological science to the general public.



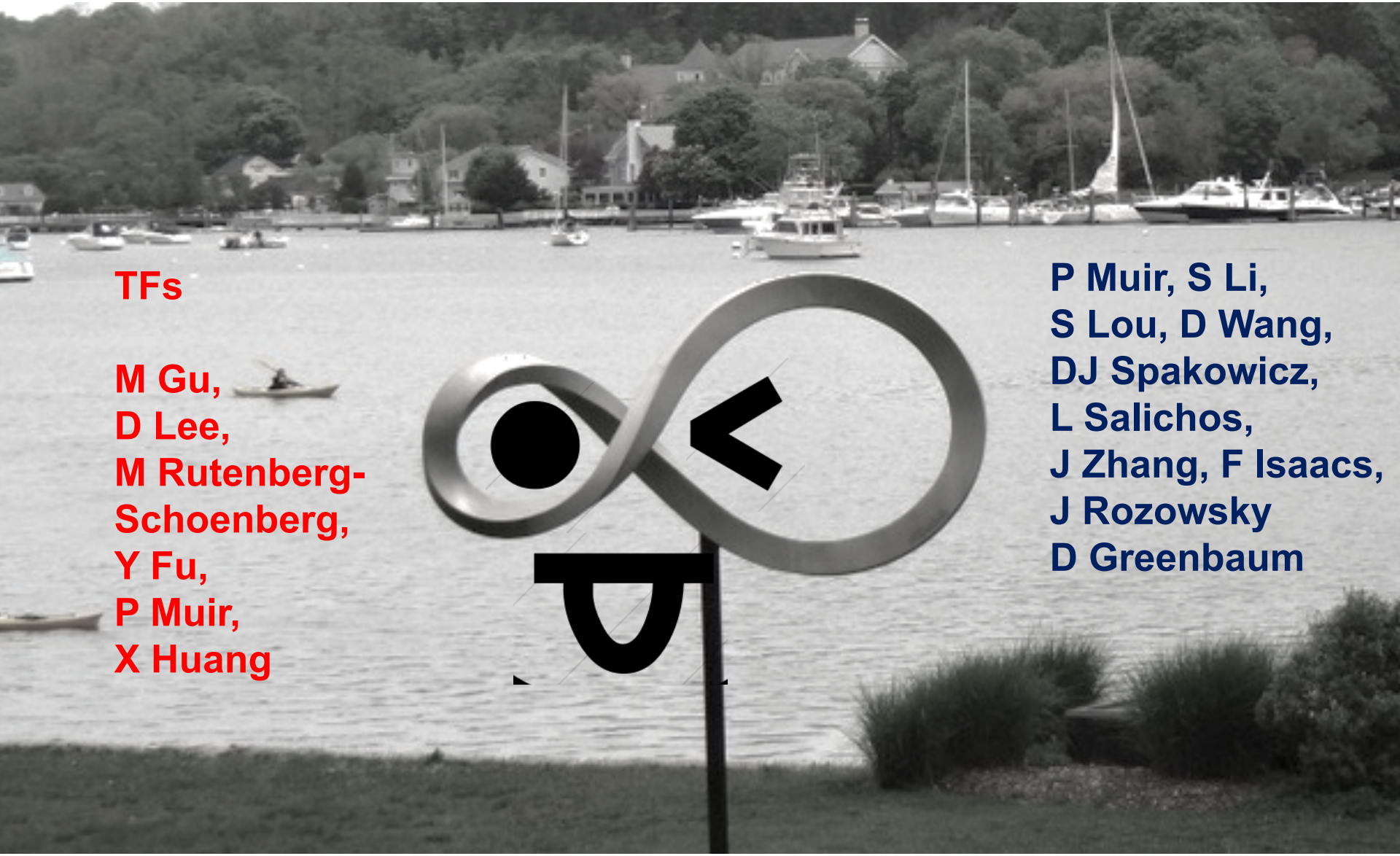
Education in Biological Data Science: What to teach students ?

- Field Definers
  - **Q:** What is driving the development of biological data science as a field?
  - **Q:** How is biological data science related to other sub-disciplines of data science ?
- The Students
  - **Q:** What backgrounds do students of this discipline have?
  - **Q:** What careers are we preparing for?
- Curriculum
  - **Q:** Should we have standard curriculum ? What would we put into in?
  - **Q:** Does data science include physical modelling?
  - **Q:** What are good metaphors for the subject?

Education in Biological Data Science: What to teach students ?

- Field Definers
  - **Q:** What is driving the development of biological data science as a field?
    A: Parallel Moore's-law scaling of computing & data generation
  - **Q:** How is biological data science related to other sub-disciplines of data science ?
    A: one of the 1st. Perhaps an exemplar for others
- The Students
  - **Q:** What backgrounds do students of this discipline have?
    A: Yale CBB as a case study: mixed betw. bio/chem & informatics
  - **Q:** What careers are we preparing for?
    A: Lots in "data industry", not nec. in biotech
- Curriculum
  - **Q:** Should we have standard curriculum ? What would we put into in?
    A: A crowded-sourced response, intersecting informatics & bio. subjects with some unique to bioinformatics
  - **Q:** Does data science include physical modelling?
    A: Yes
  - **Q:** What are good metaphors for the subject?
    A: Weather forecasting, knowledge stack & gateway

Acknowledgements



TFs

M Gu,
D Lee,
M Rutenberg-Schoenberg,
Y Fu,
P Muir,
X Huang

P Muir, S Li,
S Lou, D Wang,
DJ Spakowicz,
L Salichos,
J Zhang, F Isaacs,
J Rozowsky
D Greenbaum

Lectures.GersteinLab.org for class + Hiring Postdocs. See gersteinlab.org/jobs

Info about content in this slide pack

- General PERMISSIONS
  - This Presentation is copyright Mark Gerstein, Yale University, 2014.
  - Please read permissions statement at http://www.gersteinlab.org/misc/permissions.html .
  - Feel free to use slides & images in the talk with PROPER acknowledgement (via citation to relevant papers or link to gersteinlab.org).
  - Paper references in the talk were mostly from Papers.GersteinLab.org.

- For SeqUniverse slide, please contact Heidi Sofia, NHGRI

- PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see http://streams.gerstein.info .
  - In particular, many of the images have particular EXIF tags, such as kwpotppt , that can be easily queried from flickr, viz: http://www.flickr.com/photos/mbgmbg/tags/kwpotppt