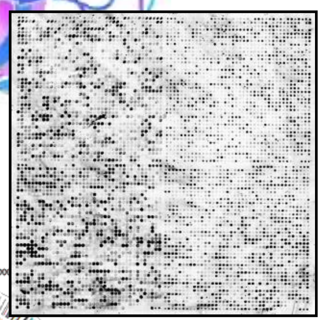
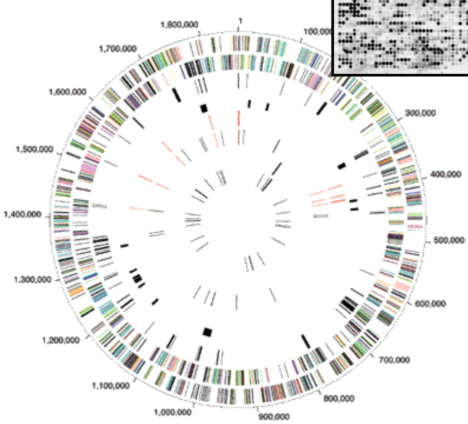
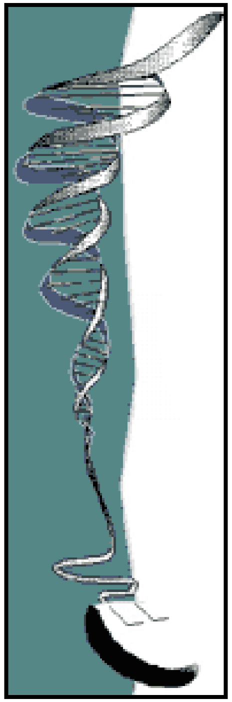
Biomedical Data Science: Introduction

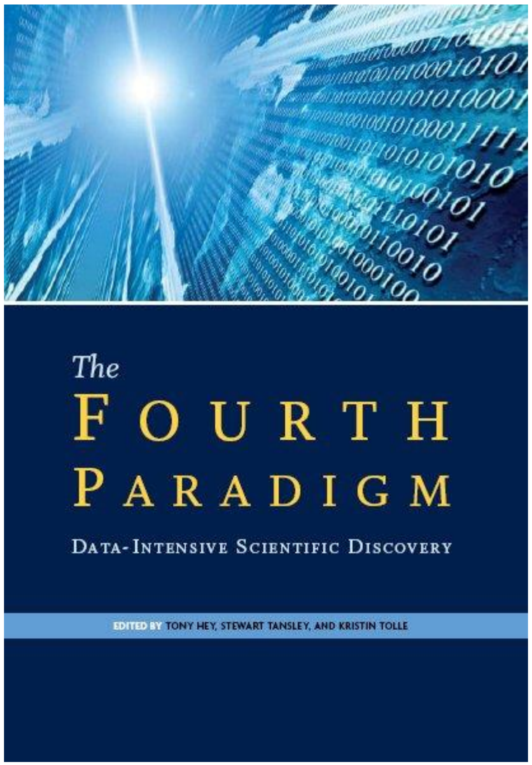






Mark Gerstein, Yale University
GersteinLab.org/courses/452
(last edit in spring ’22)

Overview: what is Biomed. Data science?

(Placing it into the context of Data Science, in general)

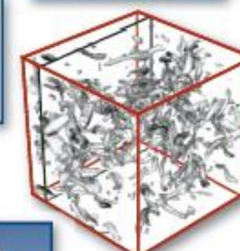
Jim Gray's 4th Paradigm



Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi Gp}{3} - K\frac{c^2}{a^2}$$



- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  - Data captured by instruments
    or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files
    using data management and statistics

Jim Gray's 4th Paradigm

#3 - Simulation

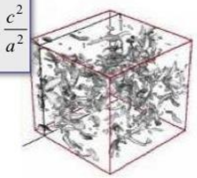
Prediction based on physical principles (eg Exact Determination of Rocket Trajectory)

Emphasis on: Supercomputers

Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today:
  data exploration (eScience)
  unify theory, experiment, and simulation
  - Data captured by instruments
    Or generated by simulator
  - Processed by software
  - Information/Knowledge stored in computer
  - Scientist analyzes database / files
    using data management and statistics

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$



#4 - Data Mining

Classifying information & discovering unexpected relationships

Emphasis: networks, "federated" DBs

Gray died in '07.
Book about his ideas came out in '09…..

[Slide from : http://research.microsoft.com/en-us/um/people/gray/talks/stanford%2520symbolic%2520systems%2520seminar.ppt]

What is Data Science? An overall, bland definition…

- Data Science encompasses the study of the entire **lifecycle of data**
  - Understanding of how data are **gathered** & the issues that arise in its collection
  - Knowledge of what data sources are available & how they may be synthesized to solve problems
  - The **storage**, access, annotation, management, & transformation of data
- Data Science encompasses many aspects of **data analysis**
  - Statistical inference, machine learning, & the design of algorithms and computing systems that enable **data mining**
  - Connecting this mining where possible with **physical modeling**
  - The presentation and **visualization** of data analysis
  - The use of data analysis to make **practical decisions** & policy
- Secondary aspects of data, not its intended use – eg the **data exhaust**
  - The appropriate protection of **privacy**
  - Creative **secondary uses** of data – eg for Science of science
  - The elimination of inappropriate bias in the entire process

Data Science in the wider world: a buzz-word for successful Ads

- Ads, media, product placement, supply optimization,
- Integral to success of GOOG, FB, AMZN, WMT…


Forbes
New Posts
+36 posts this hour
Most Popular
16-Year-Old Innovator
Lists
Promising Companies

CIO Network
INSIGHTS AND IDEAS FOR TECHNOLOGY LEADERS.
+ Follow (489)

TECH | 12/12/2012 @ 1:57AM | 3,289 views

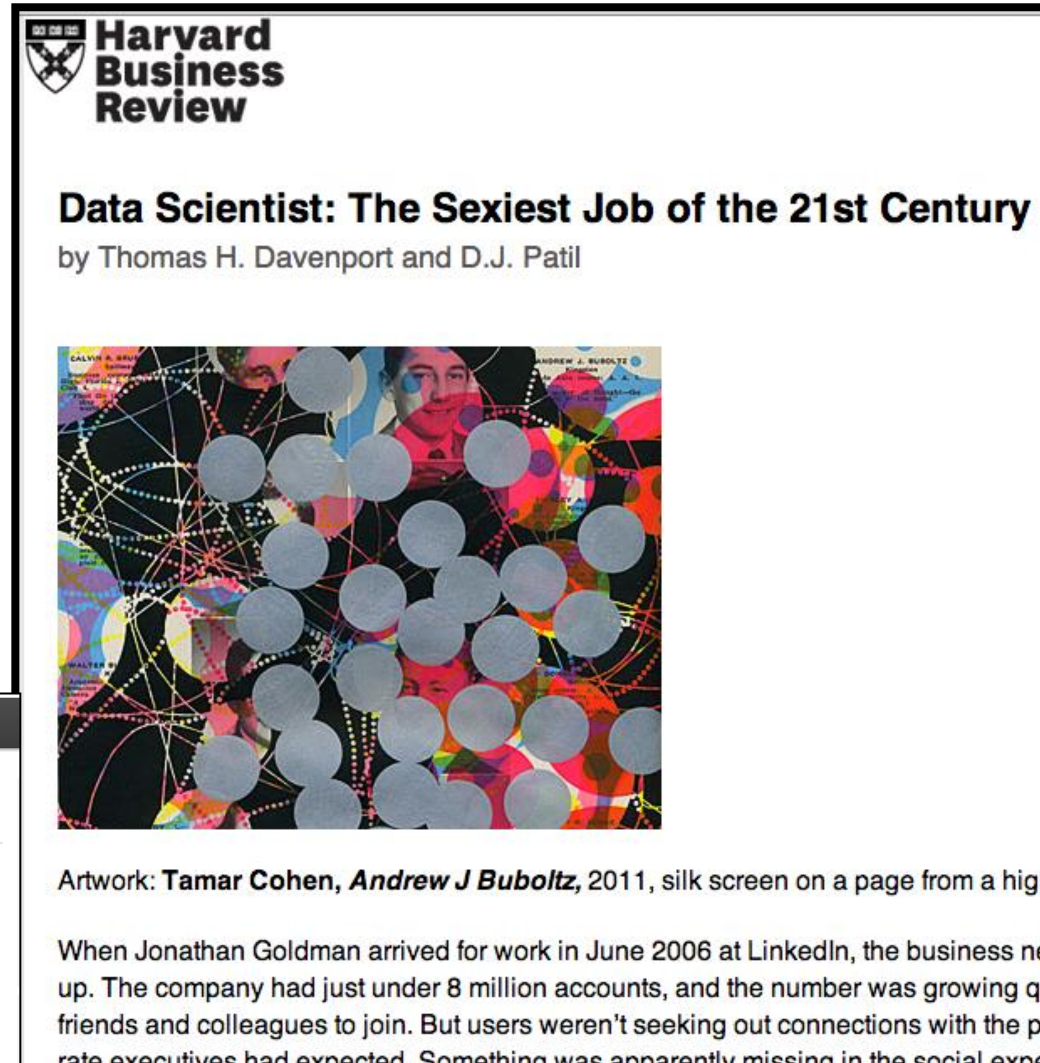
Why Big Data Is All Retailers Want for Christmas

Eric Savitz, Forbes Staff

+ Comment Now + Follow Comments

Guest post written by Quentin Gallivan

Quentin Gallivan is CEO of Pentaho Corp., an Orlando, Florida-based provider of business analytics software.


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

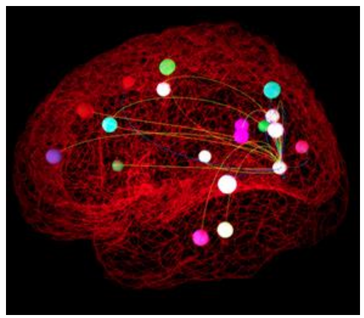
When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business ne
up. The company had just under 8 million accounts, and the number was growing qu
friends and colleagues to join. But users weren't seeking out connections with the pe
rate executives had expected. Something was apparently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

[Oct. '12 issue]

Data Science in Traditional Science

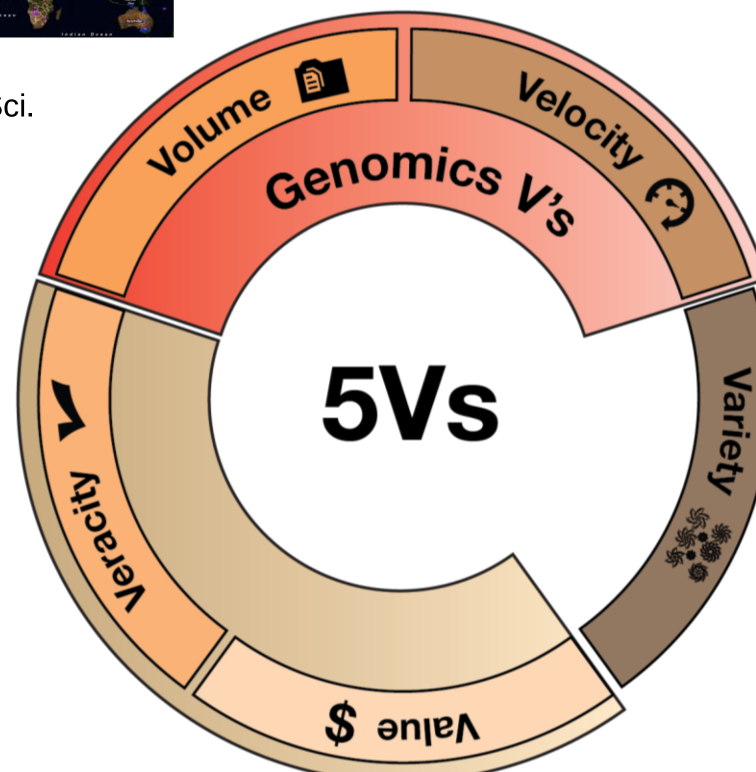
- Pre-dated commercial mining
- Instrument generated
- Large data sets often created by large teams not to answer one Q but to be mined broadly
- Often coupled to a physical/biological model
- Interplay w/ experiments

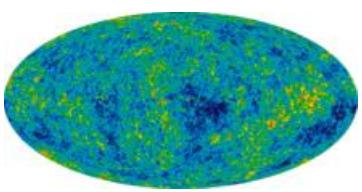
High energy physics - Large Hadron Collider



Neuroscience - The Human Connectome Project

Ecology & Earth Sci. - Fluxnet





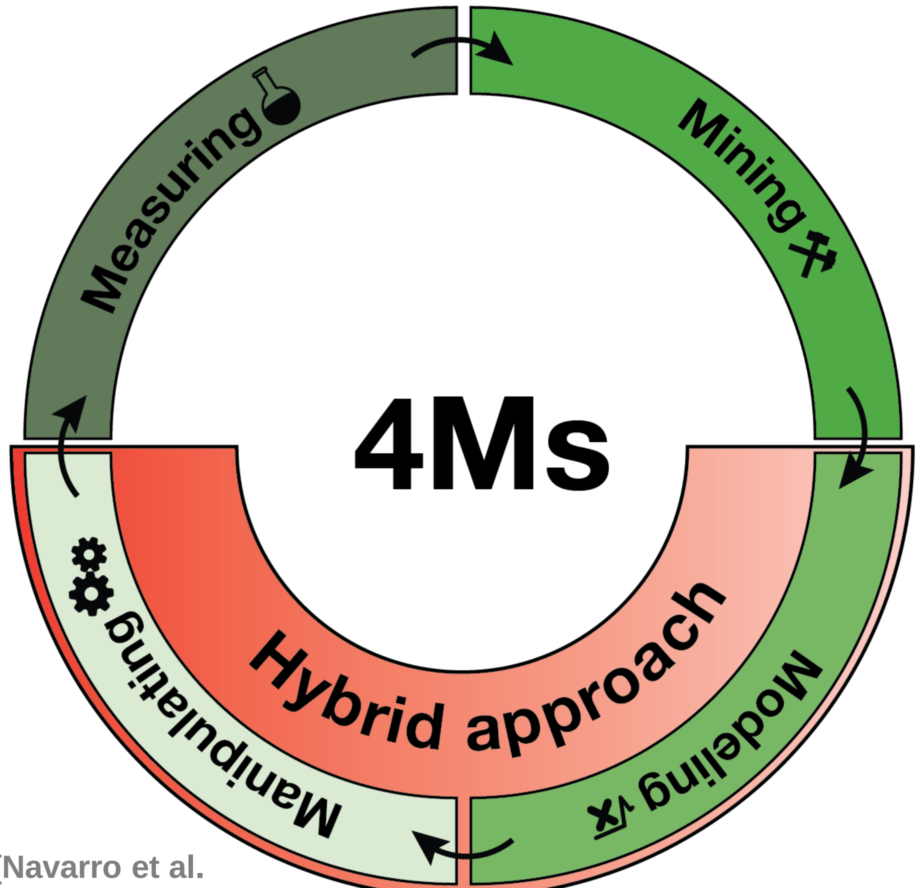
Astronomy - Sloan Digital Sky survey

Genomics DNA sequencer

[Navarro et al. GenomeBiol. ('19, in press)]

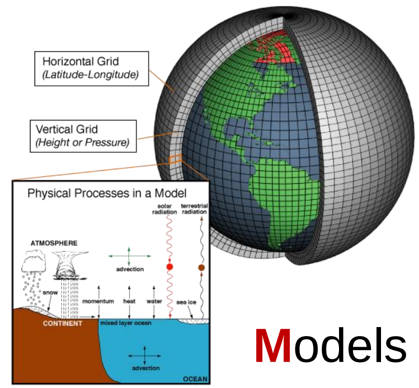
Coupling of Scientific Data to Models & Experiments

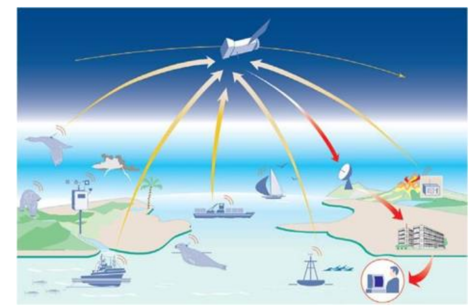
- Scientific data often coupled to a physical/biological model
- Lauffenburger's Sys. Biol. **4Ms:**
  Measurement, Mining, Modeling & Manipulation
  (Ideker et al.'06. Annals of Biomed. Eng.)
- Weather forecasting as an exemplar
  - Physical models & simulation useful but not sufficient ("butterfly" effect)
  - Success via coupling to large-scale sensor data collection



[Navarro et al. GenomeBiol. ('19, in press)]

Image from http://web.aibn.uq.edu.au/cssb/ResearchProjects.htm

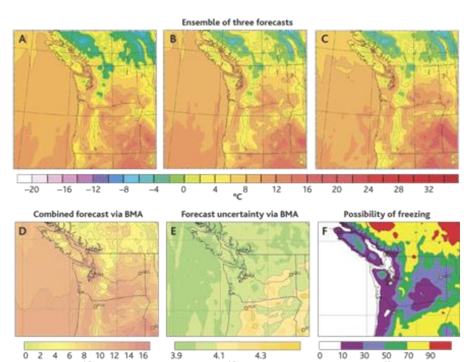




Models + Data Mining

↓

Forecasts



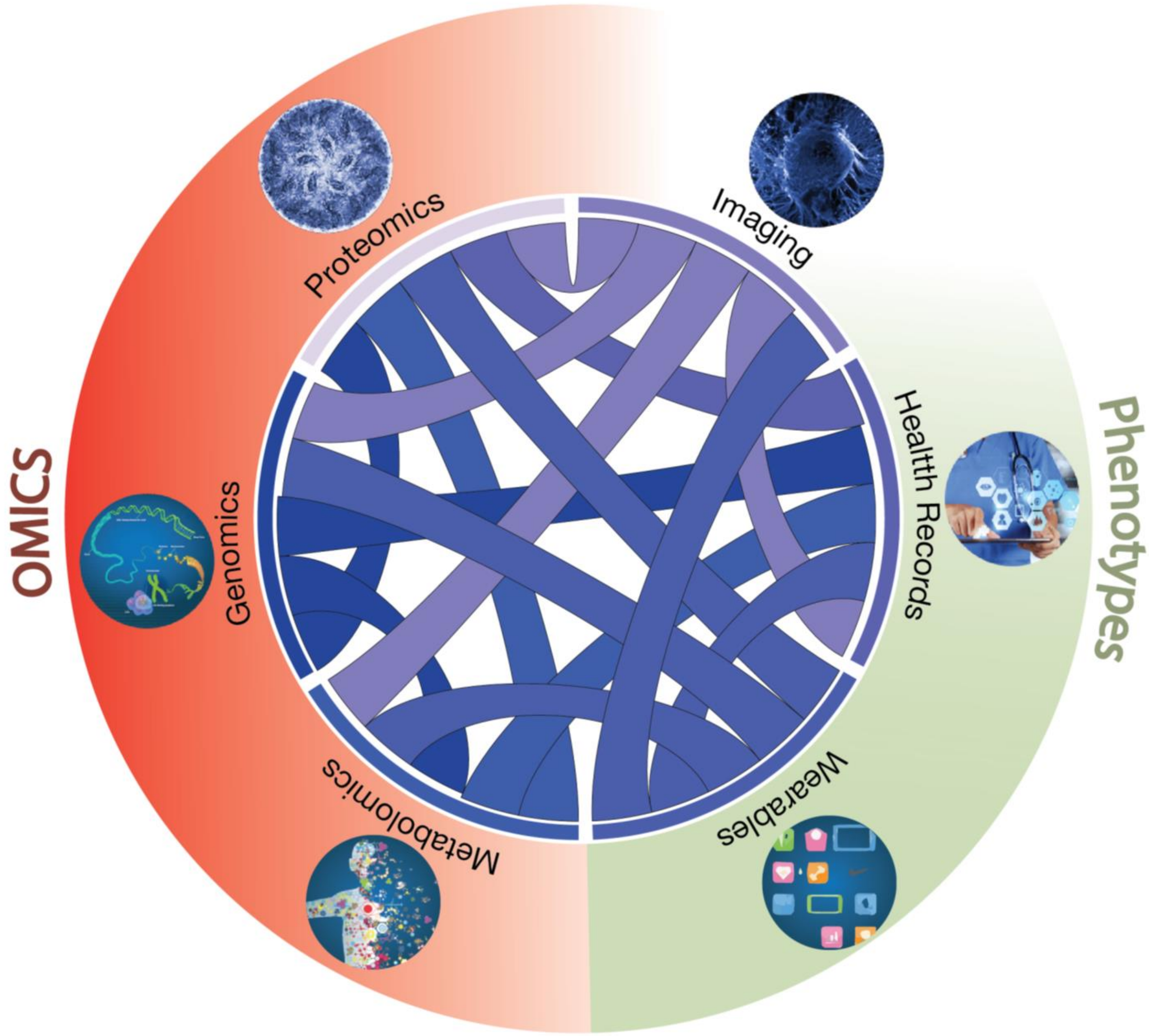
NOAA

Biomed. Data science:

Scaling & Integration

Drivers of Biomedical Data Science

- **Integration** across data types
- **Scaling** of individual data types



[Navarro et al. GenomeBiol. ('19, in press)]

Case Study: Amazing Progress in Scaling & Integration with Genotype-Phenotype Relationships

1953

Double Helix

Watson & Crick

1995

Sequenced Genome

H. influenzae

2008

1000 Genomes

Catalogue of human variation

2015

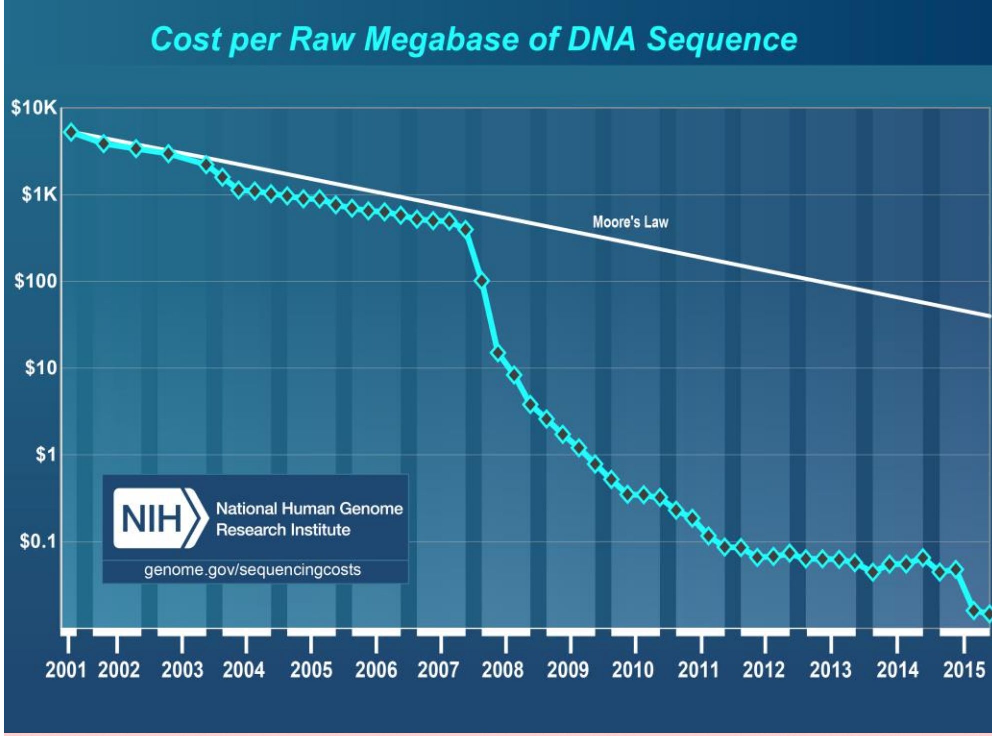
Integrated health data

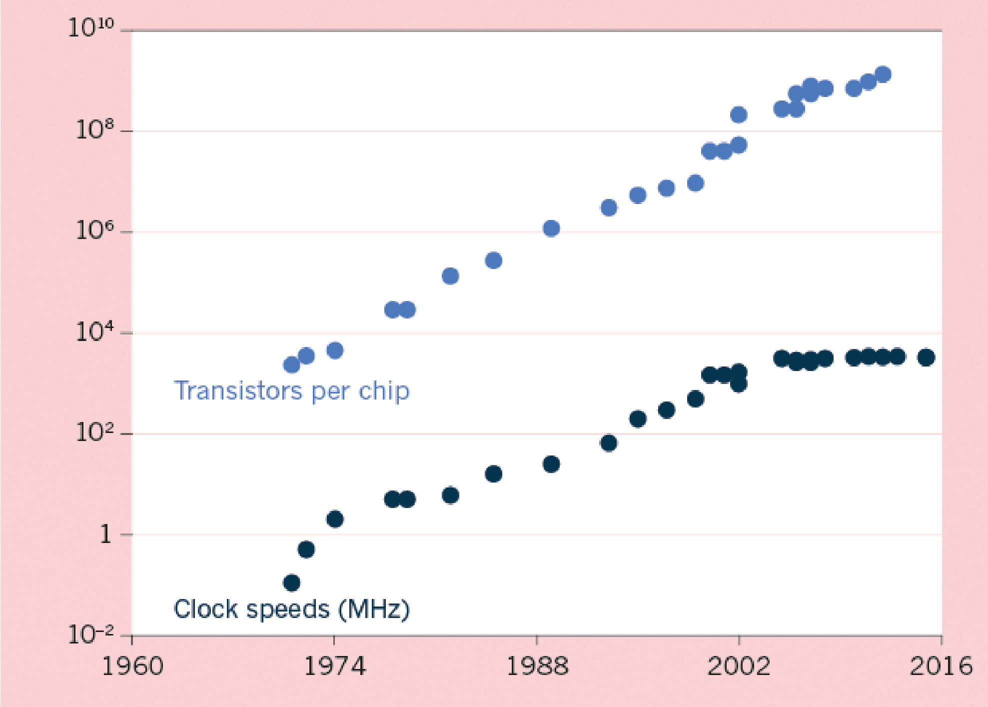
UKBB study with over 500K participants, genotypes to phenotypic details & clinical information

The Scaling of Genomic Data Science:

Powered by exponential increases in data & computing

(Moore’s Law)

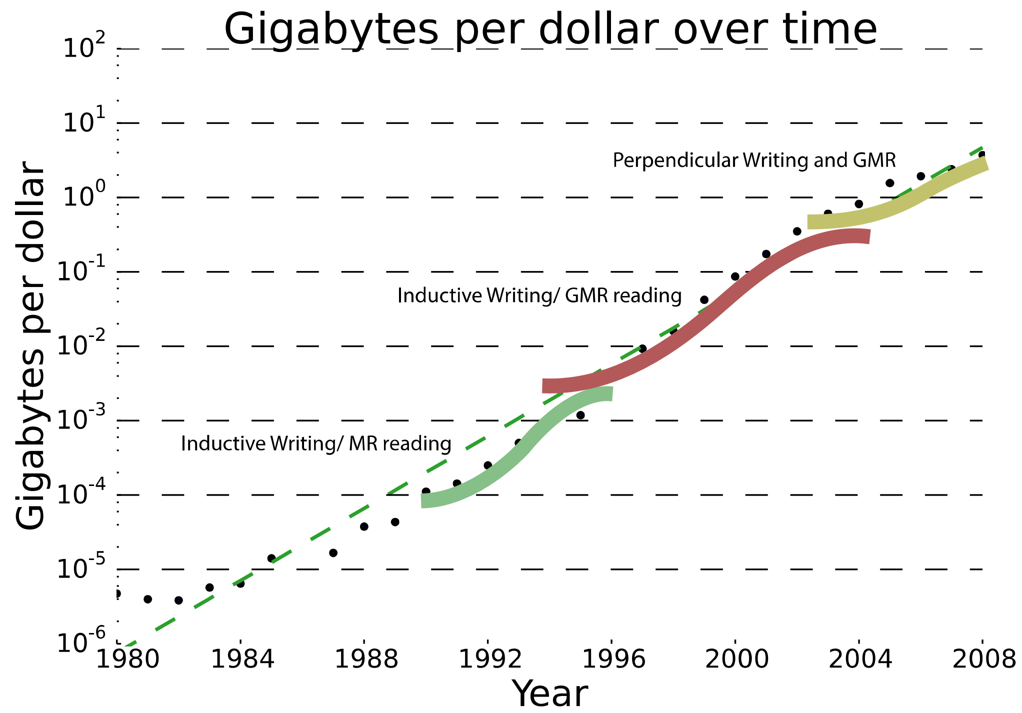


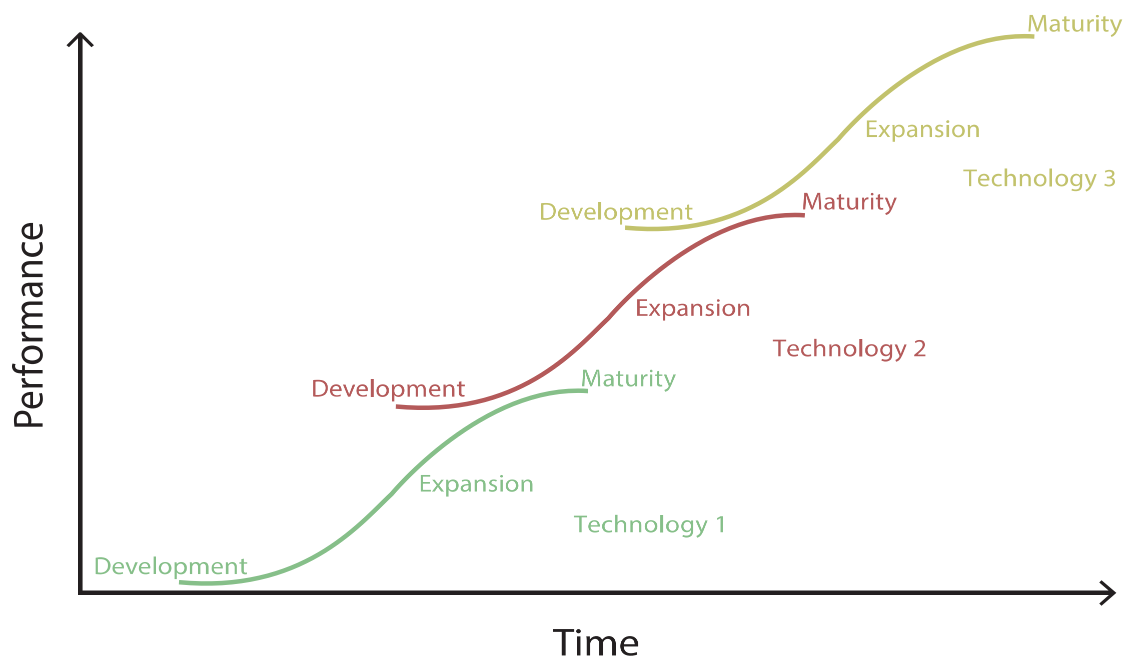


[NHGRI website + Waldrop (‘15) Nature]

Kryder's Law and S-curves underlying exponential growth

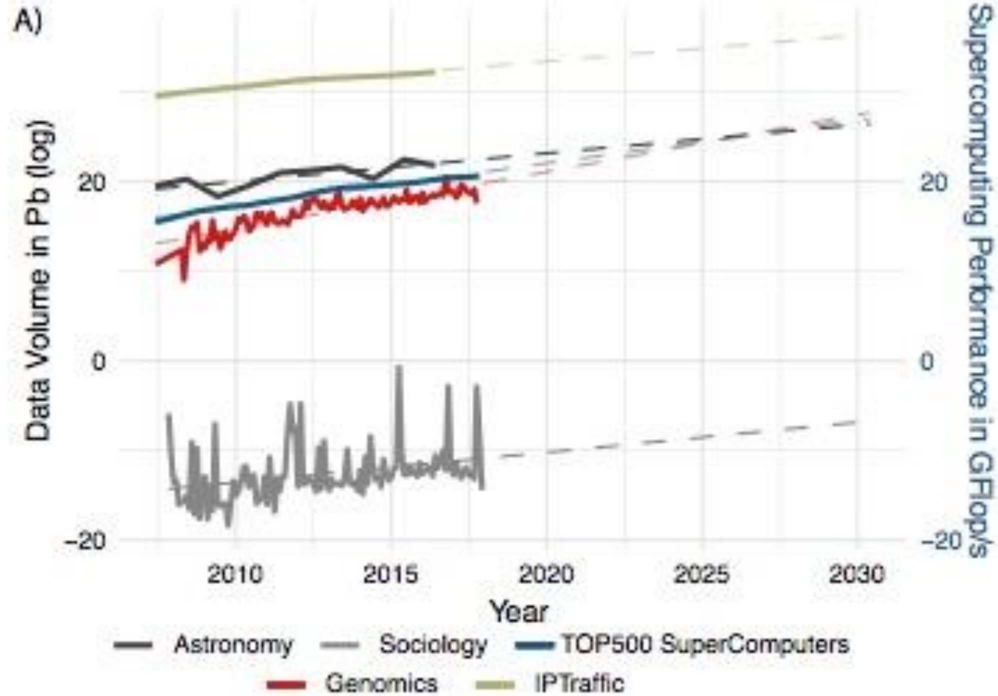
- Moore's & Kryder's Laws
    - As important as the increase in computer speed has been, the ability to store large amounts of information on computers is even more crucial
- Exponential increase seen in Kryder's law is a superposition of S-curves for different technologies

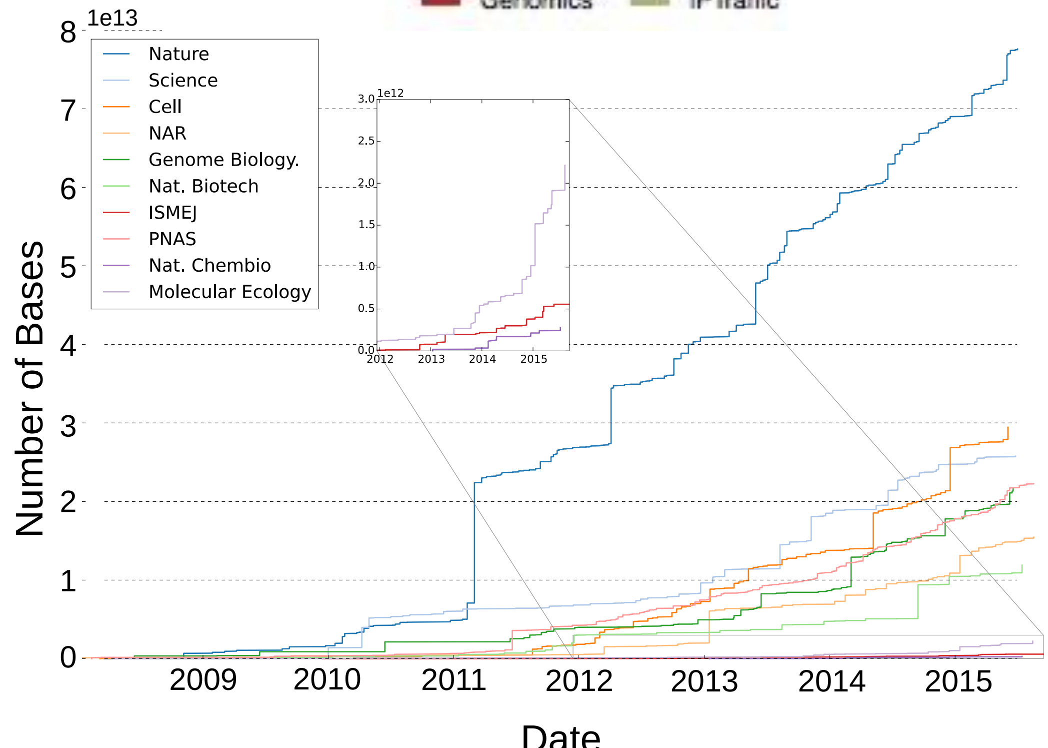




[Muir et al. ('15) GenomeBiol.]

Sequencing cost reductions have resulted in an explosion of data

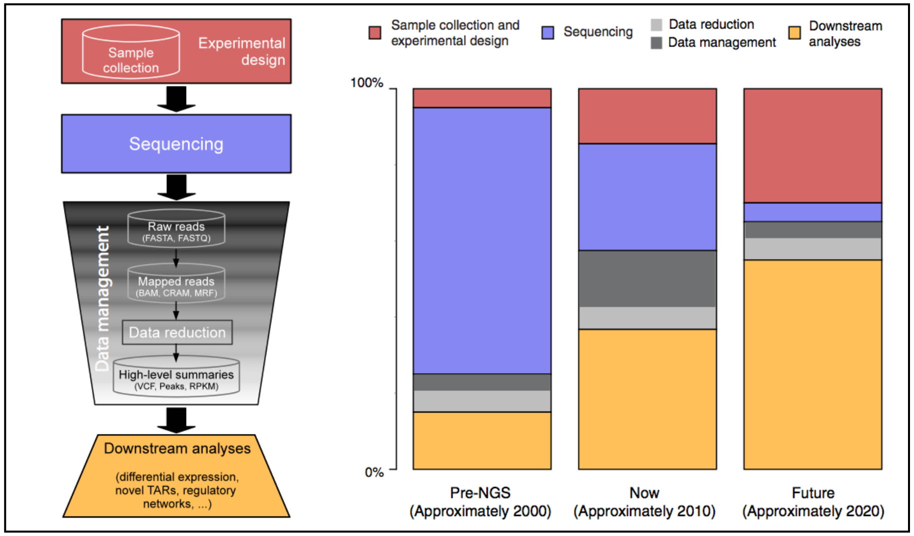




- The type of sequence data deposited has changed as well.

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

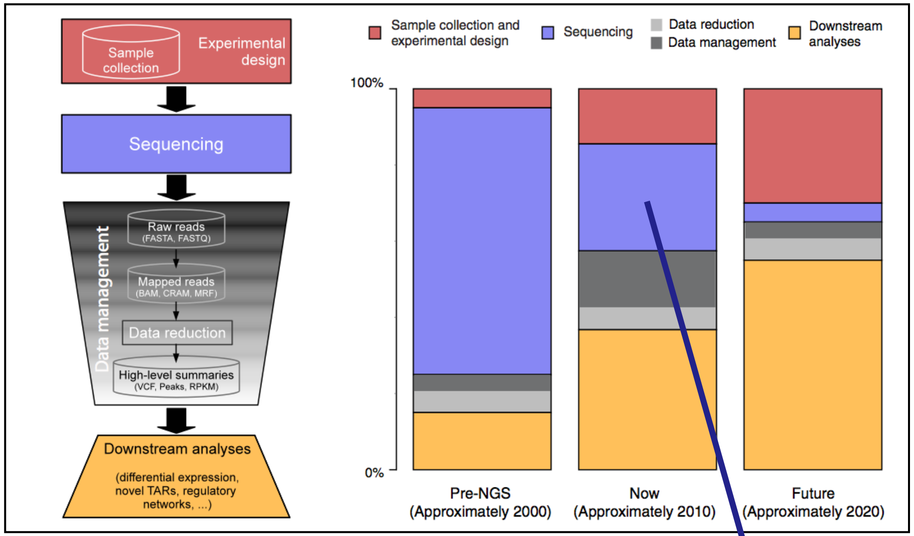
The changing costs of a sequencing pipeline



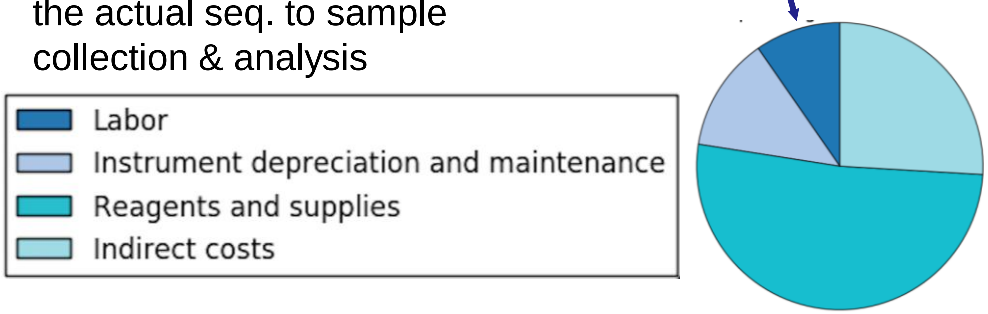
From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

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

The changing costs of a sequencing pipeline

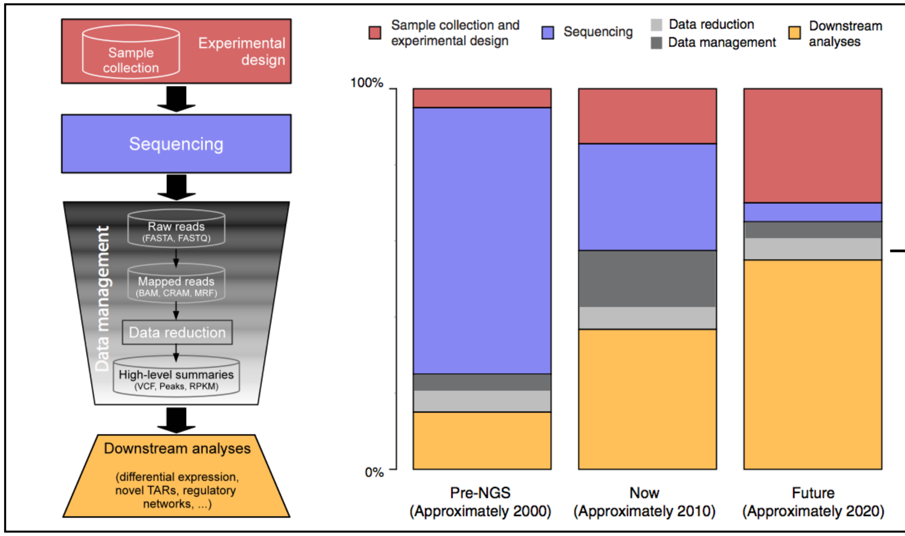


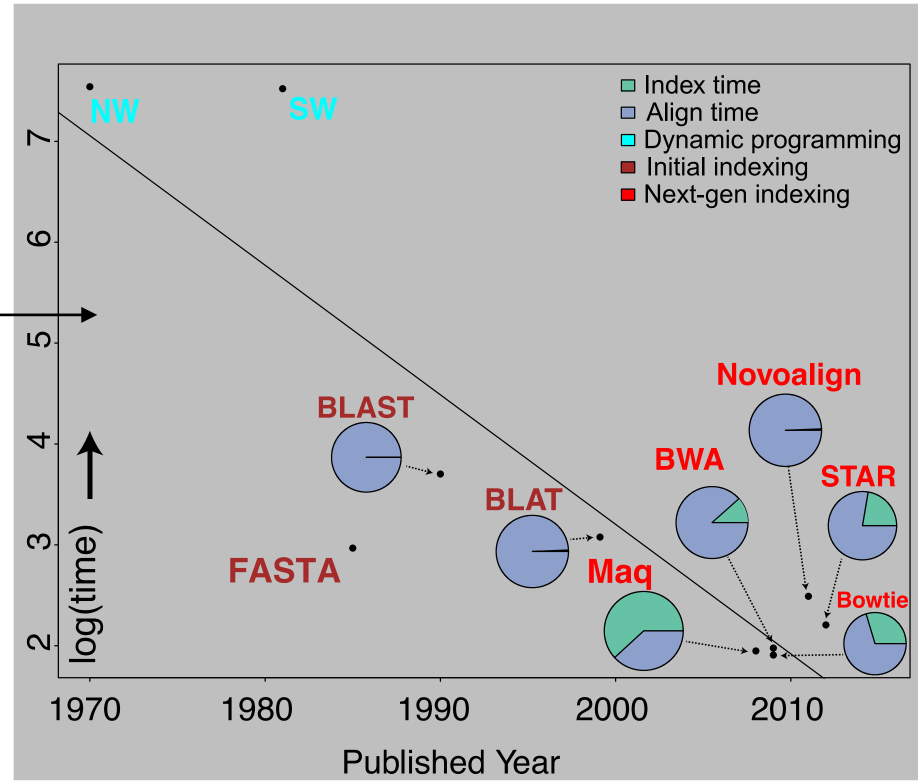
From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis



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

The changing costs of a sequencing pipeline



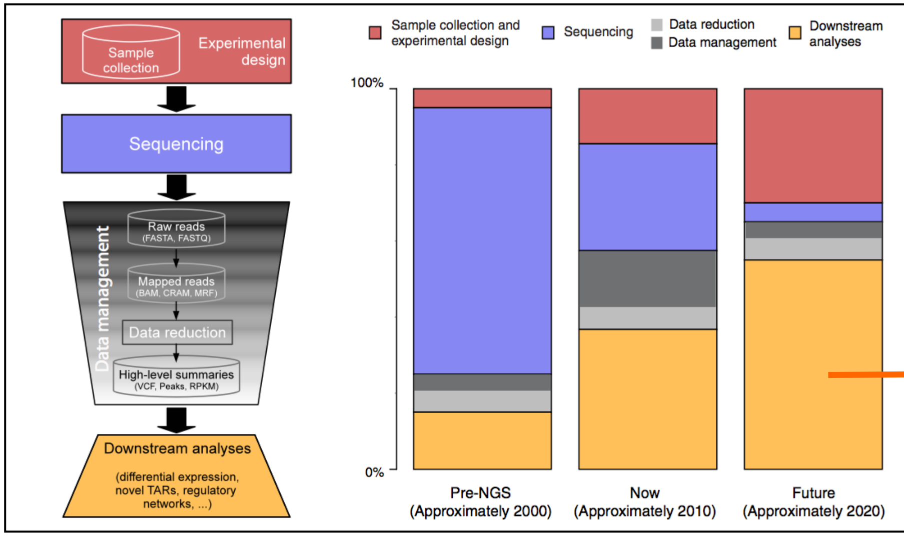


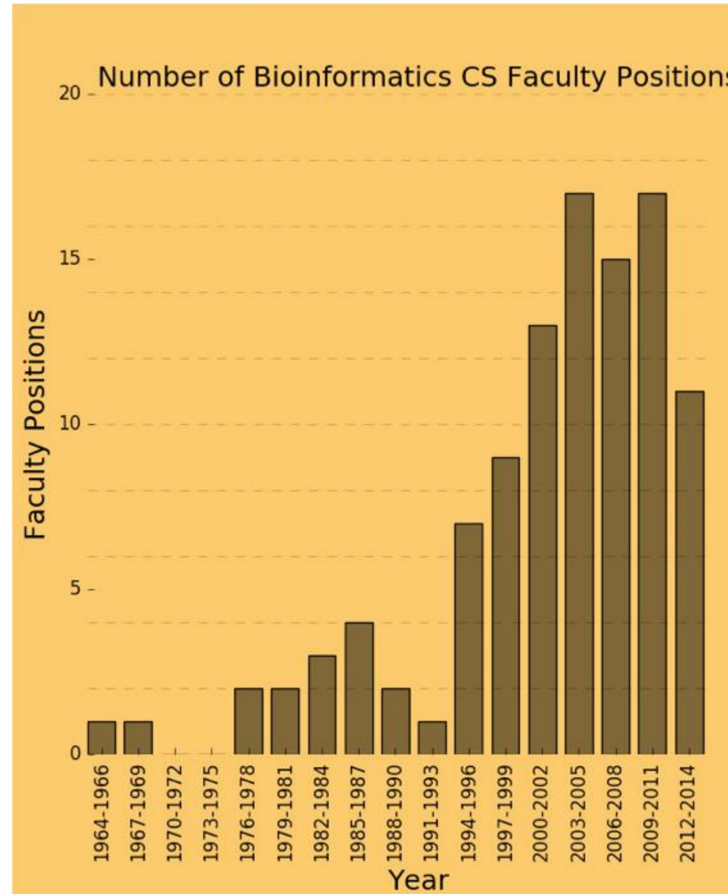
From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

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

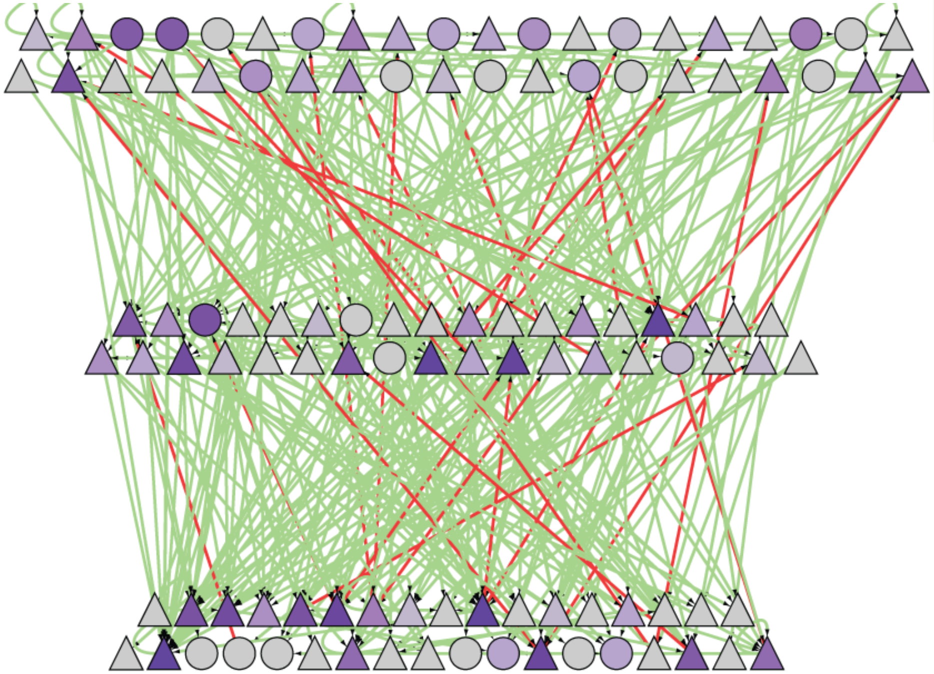


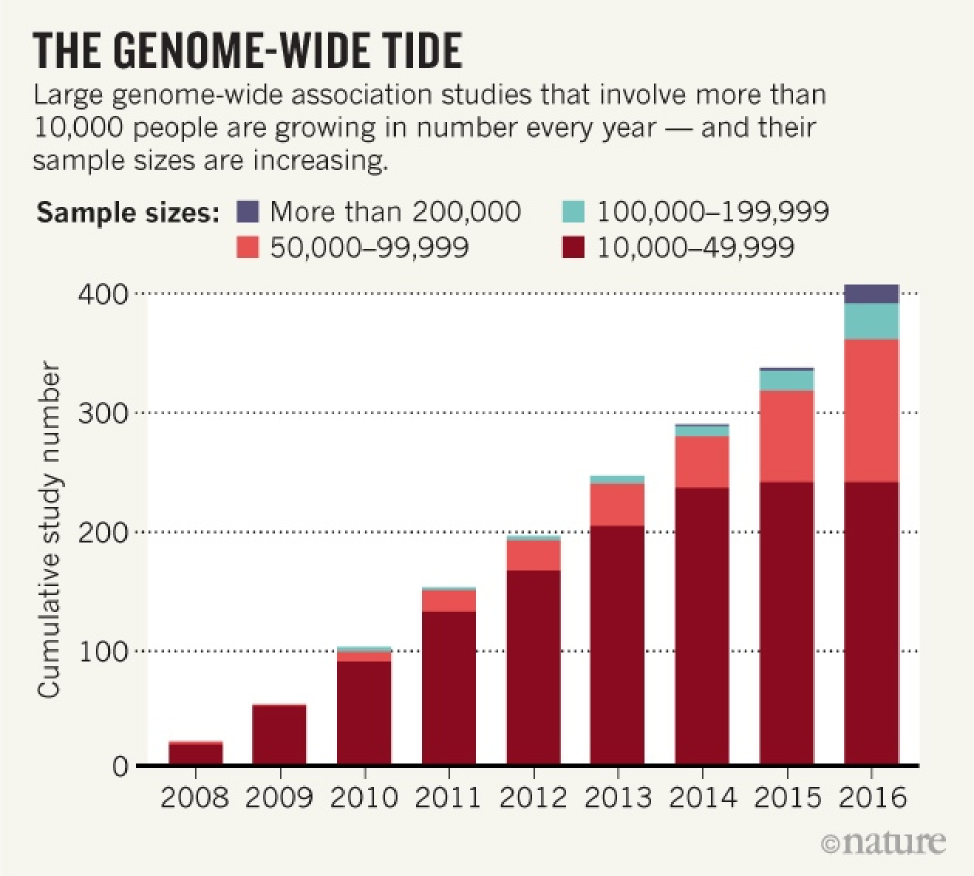


From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

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

A Success of Scale & Integration: Many GWAS variants found, most not in genes, but affecting regulatory network





- A 1st GWAS done at Yale, for AMD: (Klein et al. 05, Science)
- Many since then
- Most SNVs fall into non-coding regulatory regions
  (major contributions by Yale groups to this ENCODE annotation effort)

[*Nature* 489: 91]

Basic Science to Medicine

INITIATIVES

STARTUPS

- Large-scale 'omics data as an anchor to organize phenotypic data – EMRs, wearables…
- 1st ['05-]: Exomes & chips of disease-focused cohorts – init. GWAS, TCGA, PGC
- 2nd ['15-]: Integration of full WGS with rich & diverse phenotypes - UKBiobank, TopMed, Genomics England, PCAWG, All of Us

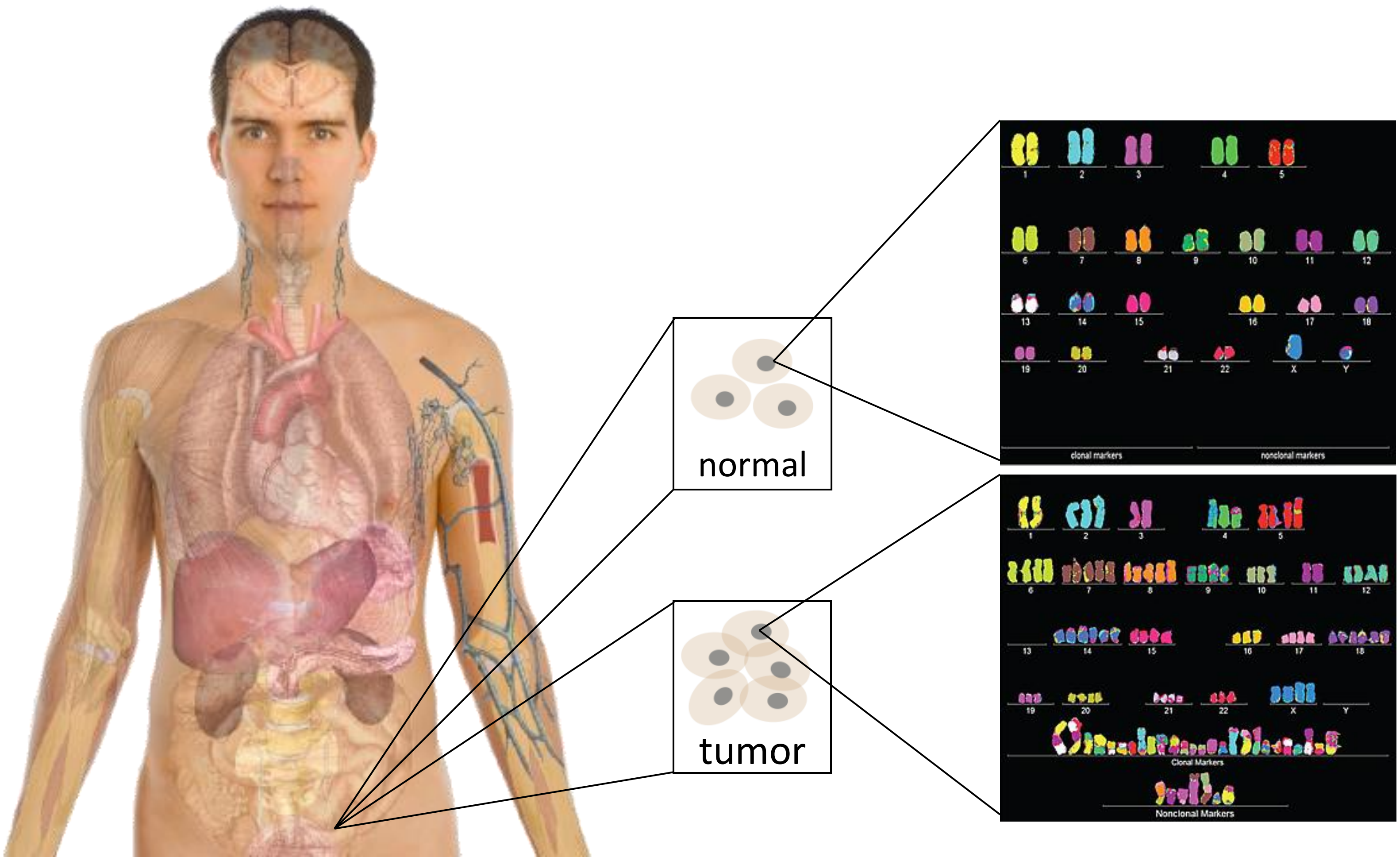
Medical Big Data: Promise and Challenges (Lee and Yoon , *Kidney Res. Clin. Pract.*, 2017)

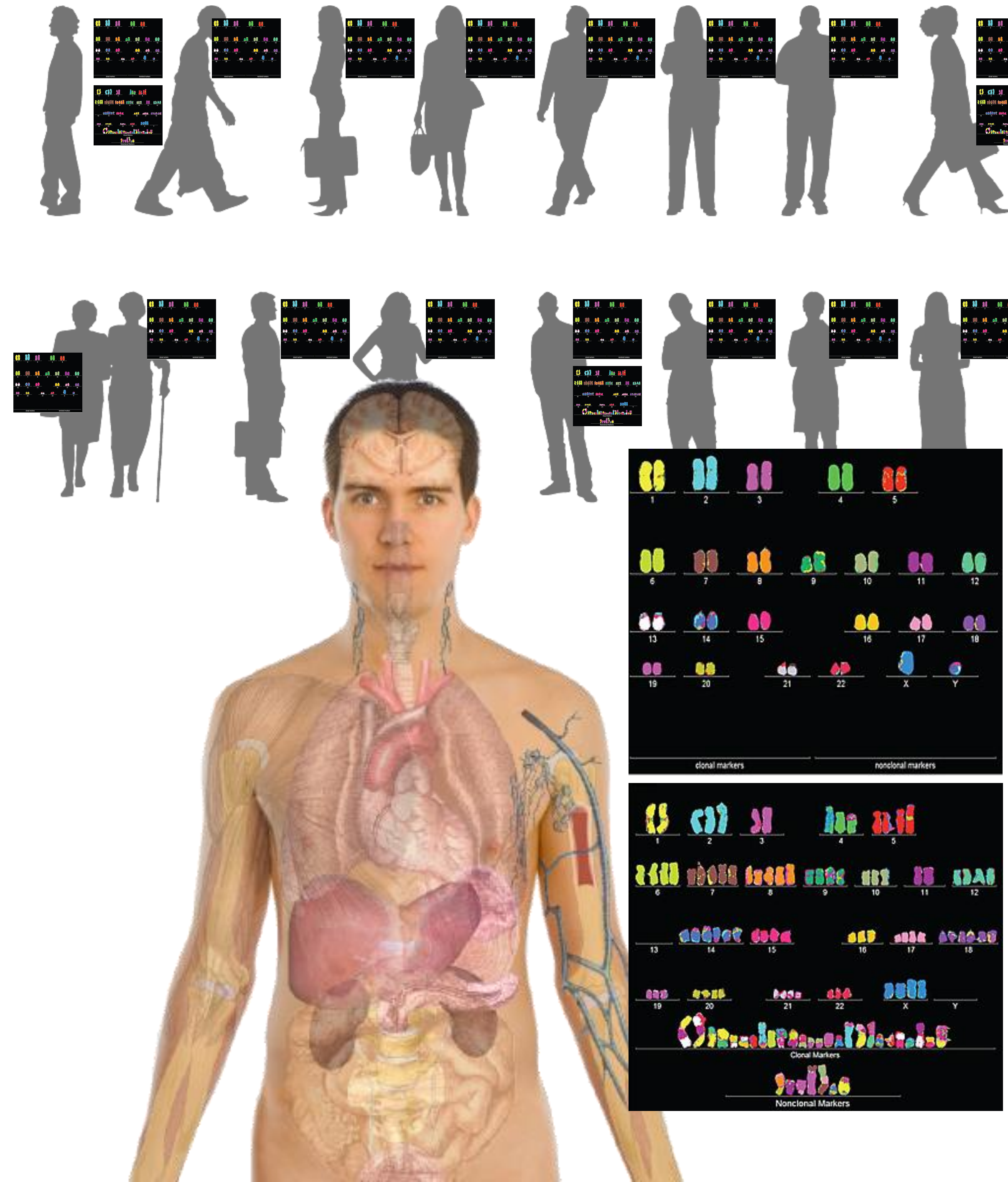
Biomed. Data science:

The Future

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

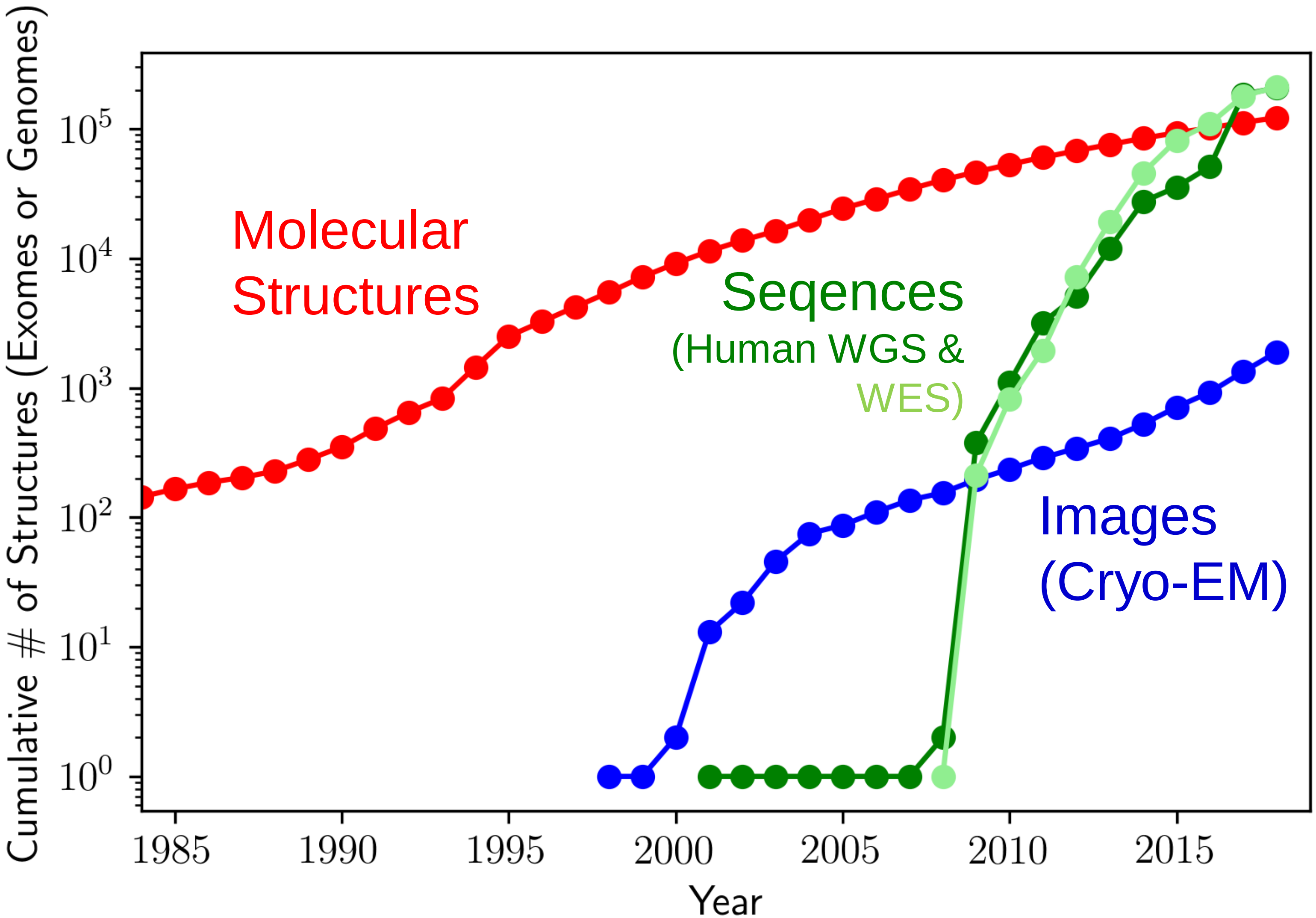
Personal genomes soon will become a commonplace part of medical research & eventually treatment (esp. for cancer). They will provide a primary connection for biological science to the general public.





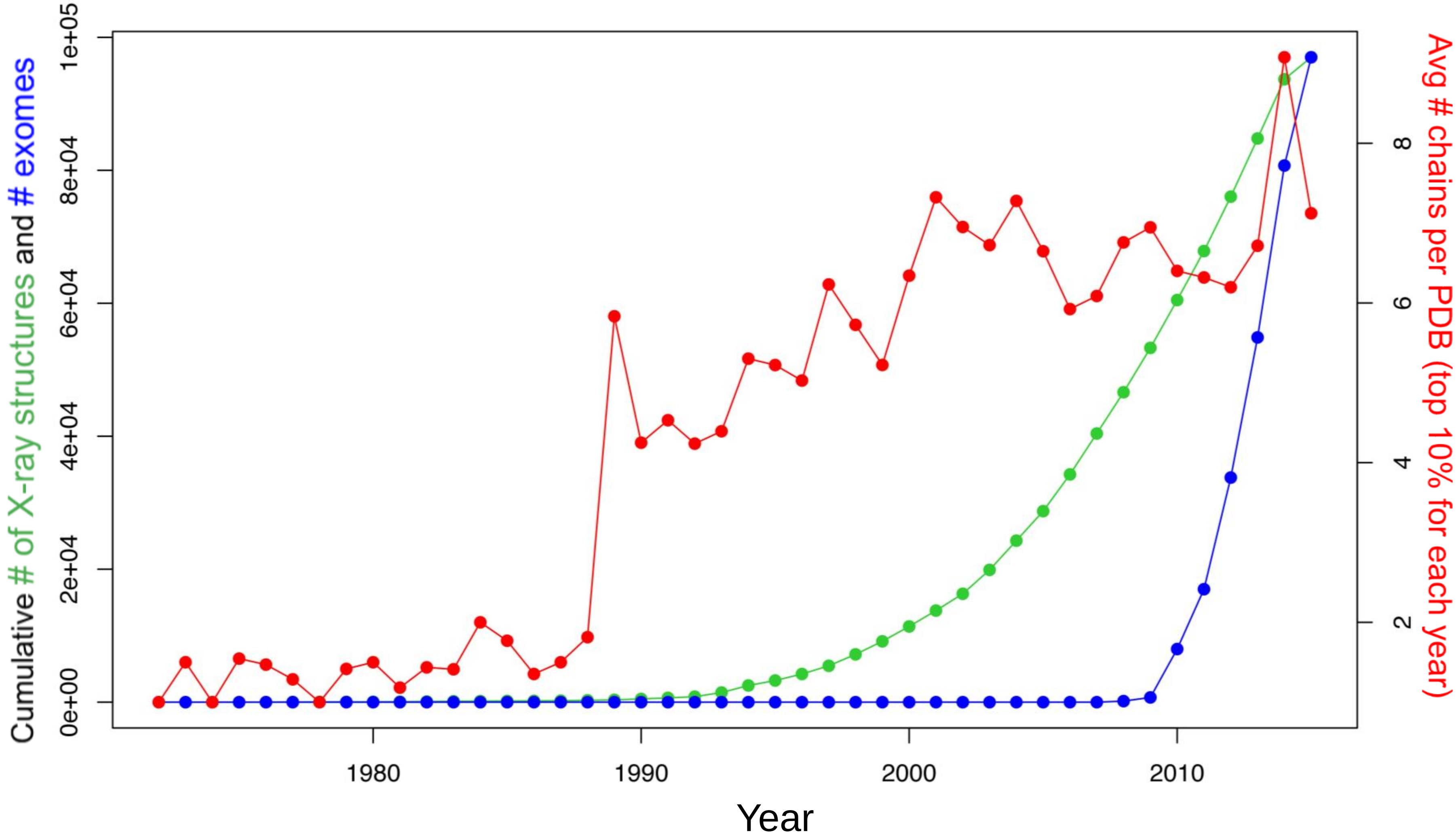
Placing the individual into the context of the population & using the population to build a interpretative model

How will the Data Scaling Continue? The Past, Present & Future Ecosystem of Large-scale Biomolecular Data



Trends in data generation point to growing opportunities for leveraging sequence variants to study structure (and vice versa)

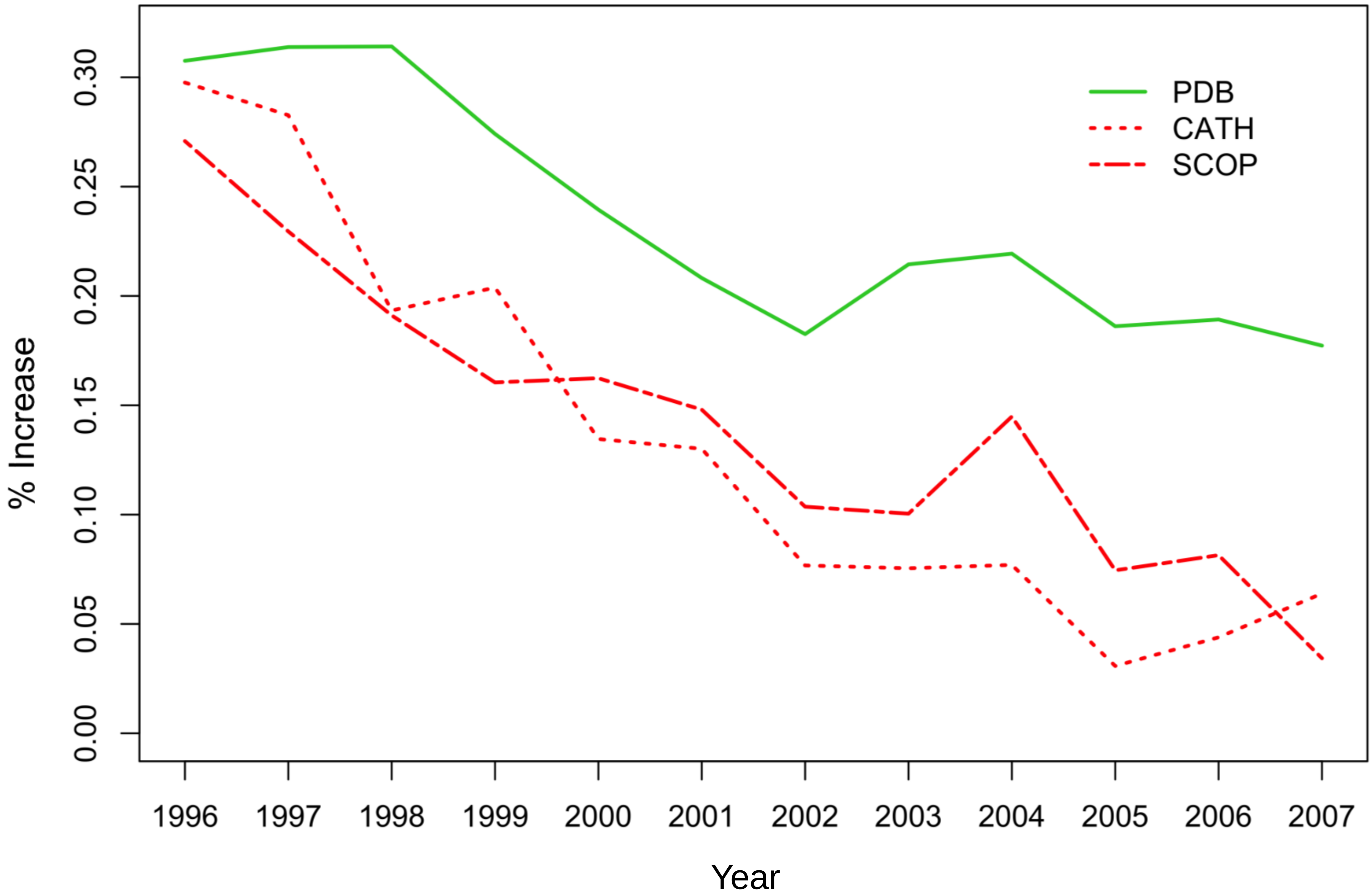
The volume of sequenced exomes is outpacing that of structures, while solved structures have become more complex in nature.



Exome data hosted on NCBI Sequence Read Archive (SRA)

[Sethi et al. COSB ('15)]

Growing sequence redundancy in the PDB (as evidenced by a reduced pace of novel fold discovery) offers a more comprehensive view of how such sequences occupy conformational landscapes – Gene & Struc. Families as main organizing principle



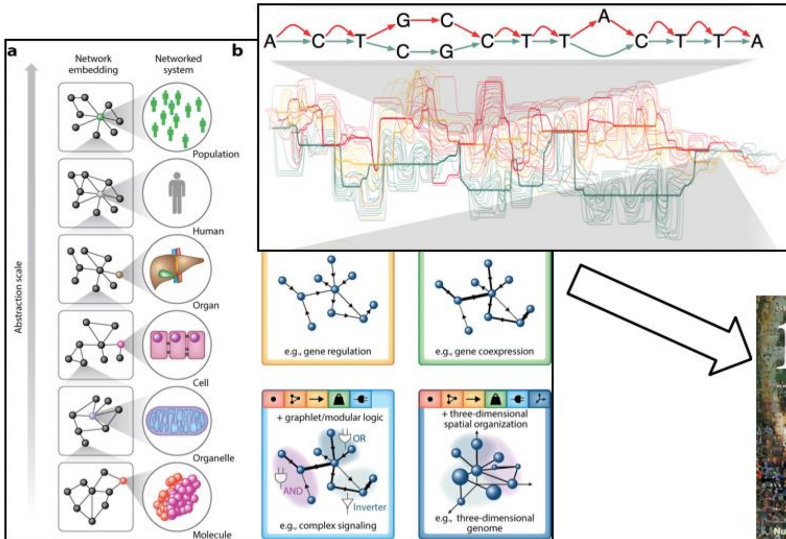
PDB: Berman HM, et al. NAR. (2000)
CATH: Sillitoe I, et al. NAR. (2015)
SCOP: Fox NK et al. NAR. (2014)

[Sethi et al. COSB ('15)]

Examples of Imports & Exports to/from Genomics & Other Data Science Application Areas

Technical Imports

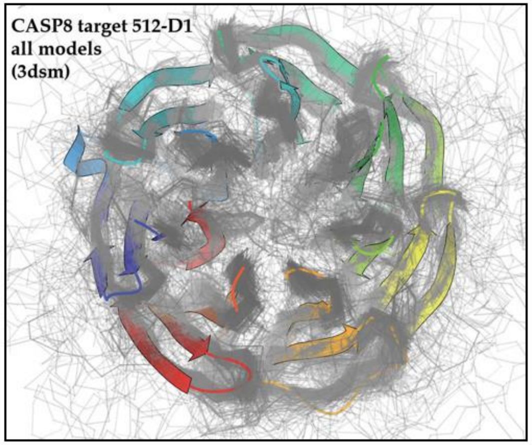
Networks and graphs



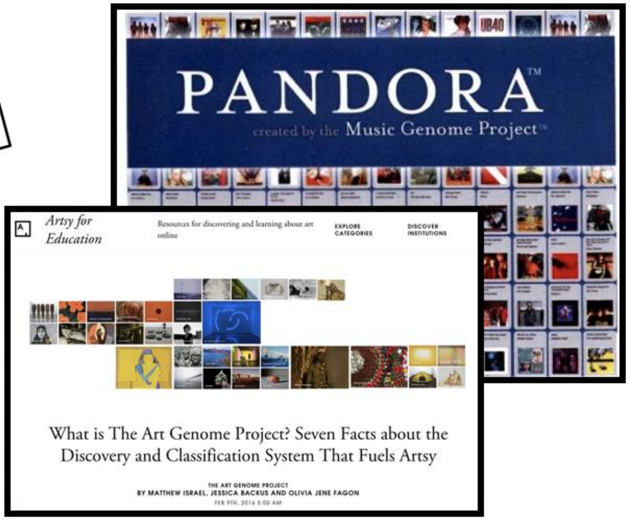
Importing tech. developed in other big data disciplines

Cultural Imports

CASP

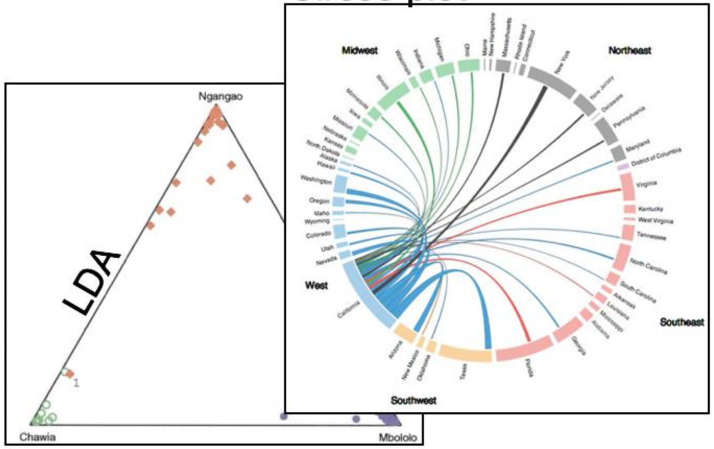


Cultural Exports

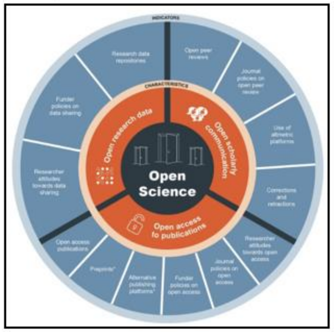


Technical Exports

Circos plot



Open Science



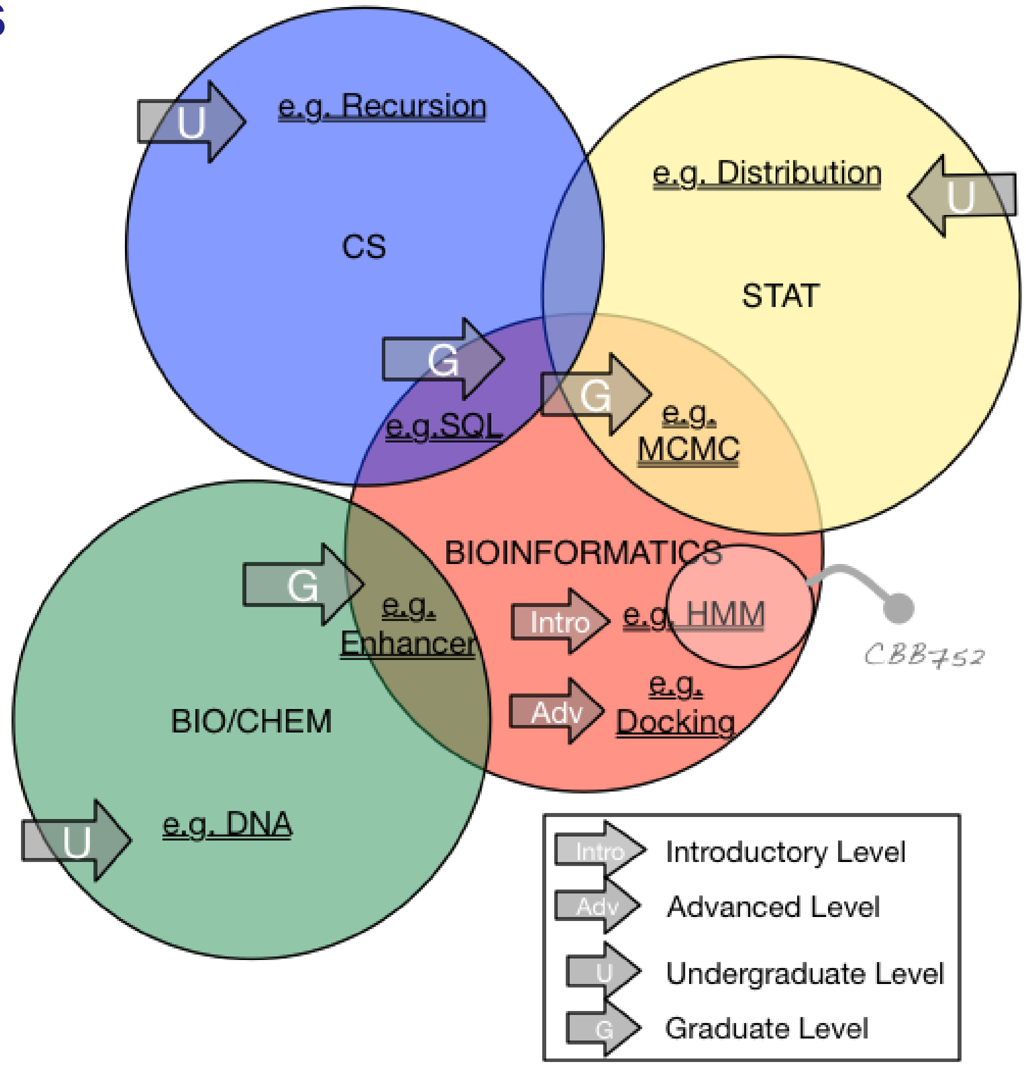
[Navarro et al. GenomeBiol. ('19, in press)]

Biomed. Data science:

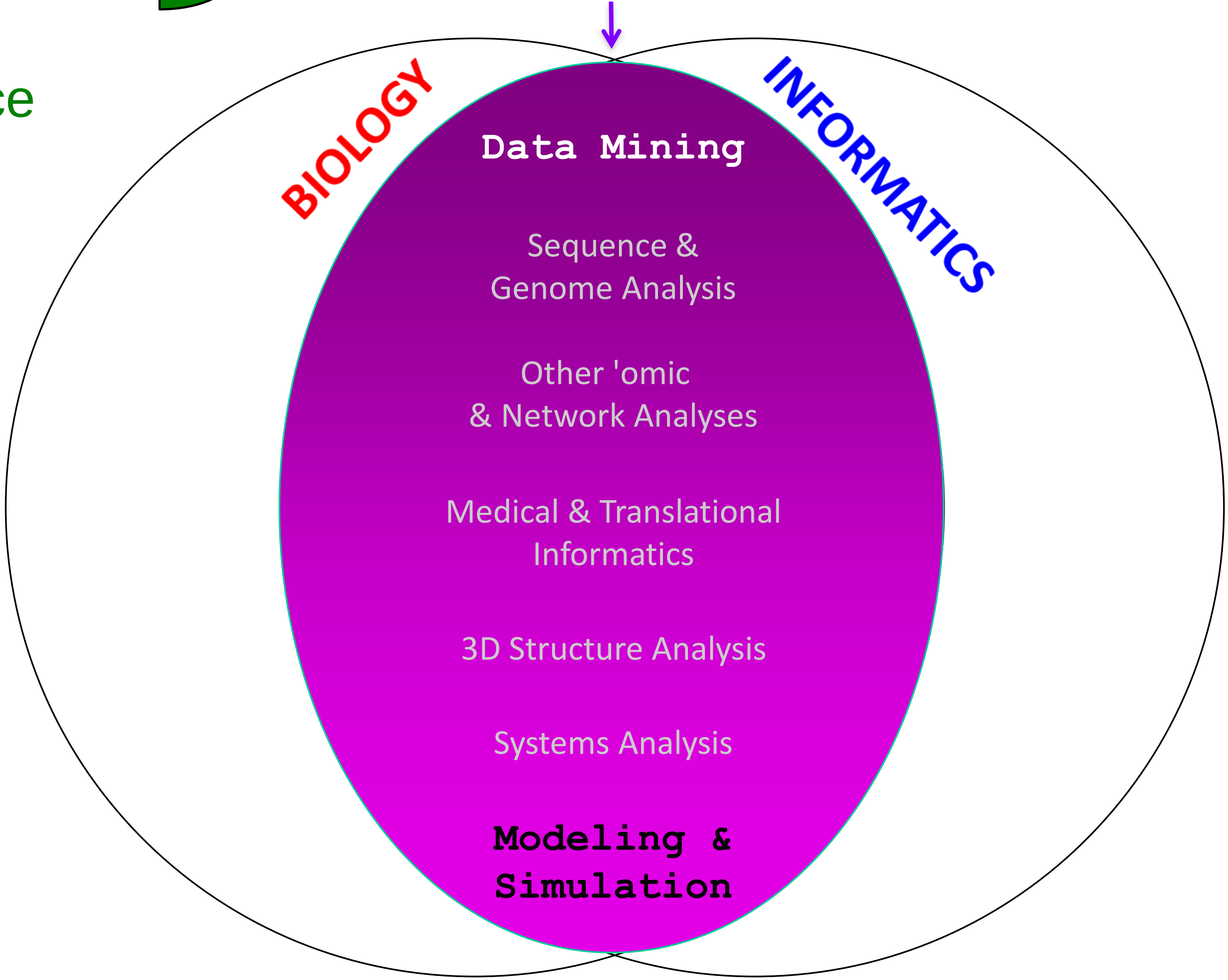
The Course

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"



Biomedical Data Science ⊃ (Molecular) BIOINFORMATICS



[Luscombe et al. ('01). *Methods Inf Med* 40: 346]

What is Bioinformatics?

- *(Molecular)* **Bio** - **`informatics`**
- One idea for a definition?
  Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **"informatics" techniques** (derived from disciplines such as applied math, CS, and statistics) to **organize, mine, model & understand the information associated** with these molecules, **on a large-scale.**
- Bioinformatics is a practical discipline with many **applications**.

[Luscombe et al. ('01). *Methods Inf Med* 40: 346]

Class Web Page

GersteinLab.org/courses/452

Short Office Hours

Today right after class.

After that email me!

(in Bass 432,

contact.gerstein.info)