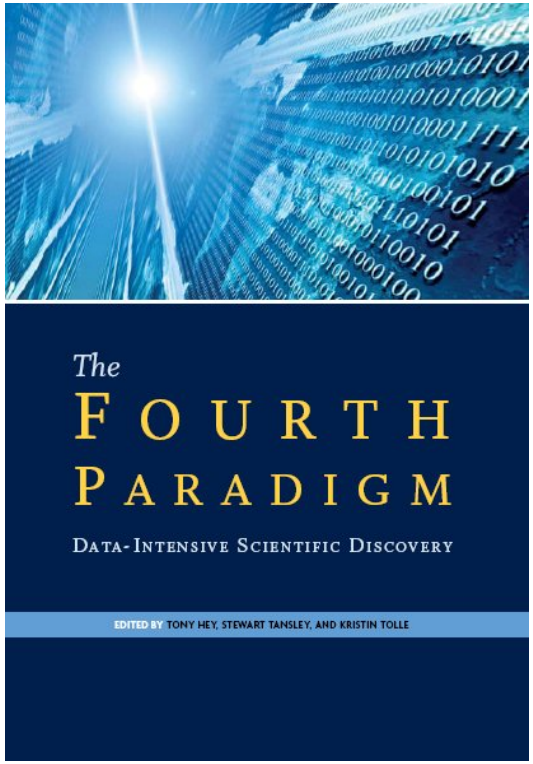
Overview:
what is
Biomed. Data science?
(Placing it into context)

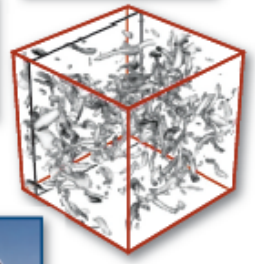
Jim Gray's 4th Paradigm



Science Paradigms

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

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



Jim Gray's 4th Paradigm

#3 - Simulation

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

Emphasis on: Supercomputers

#4 - Data Mining

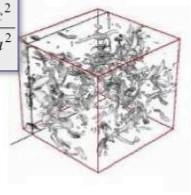
Classifying information & discovering unexpected relationships

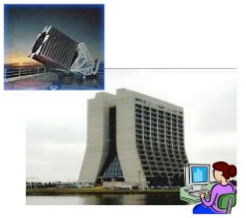
Emphasis: networks, "federated" DBs

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

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





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

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

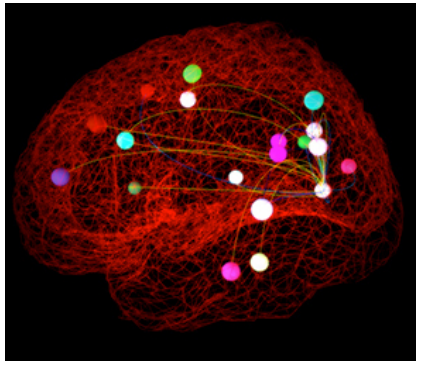
Data Science in the wider world: a buzz-word for successful Ads

[Oct. '12 issue]

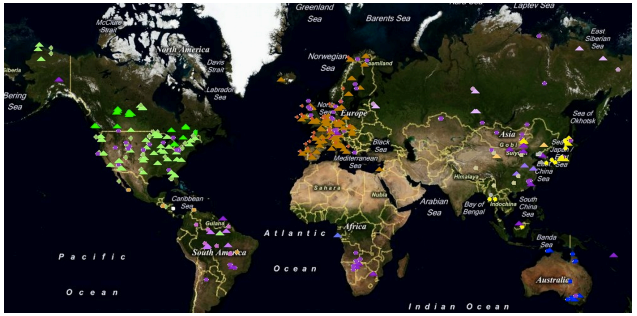
Data Science in Traditional Science

- 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

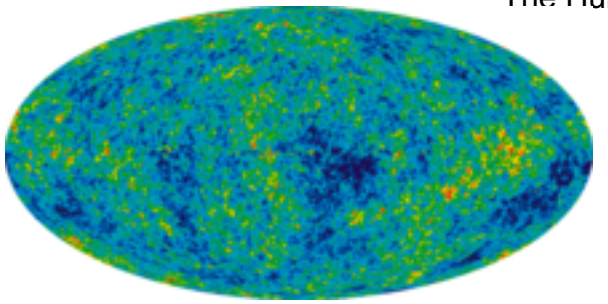
High energy physics -
Large Hadron Collider



Neuroscience -
The Human Connectome Project



Ecology
& Earth Sci.
- Fluxnet



Astronomy -
Sloan Digital Sky survey

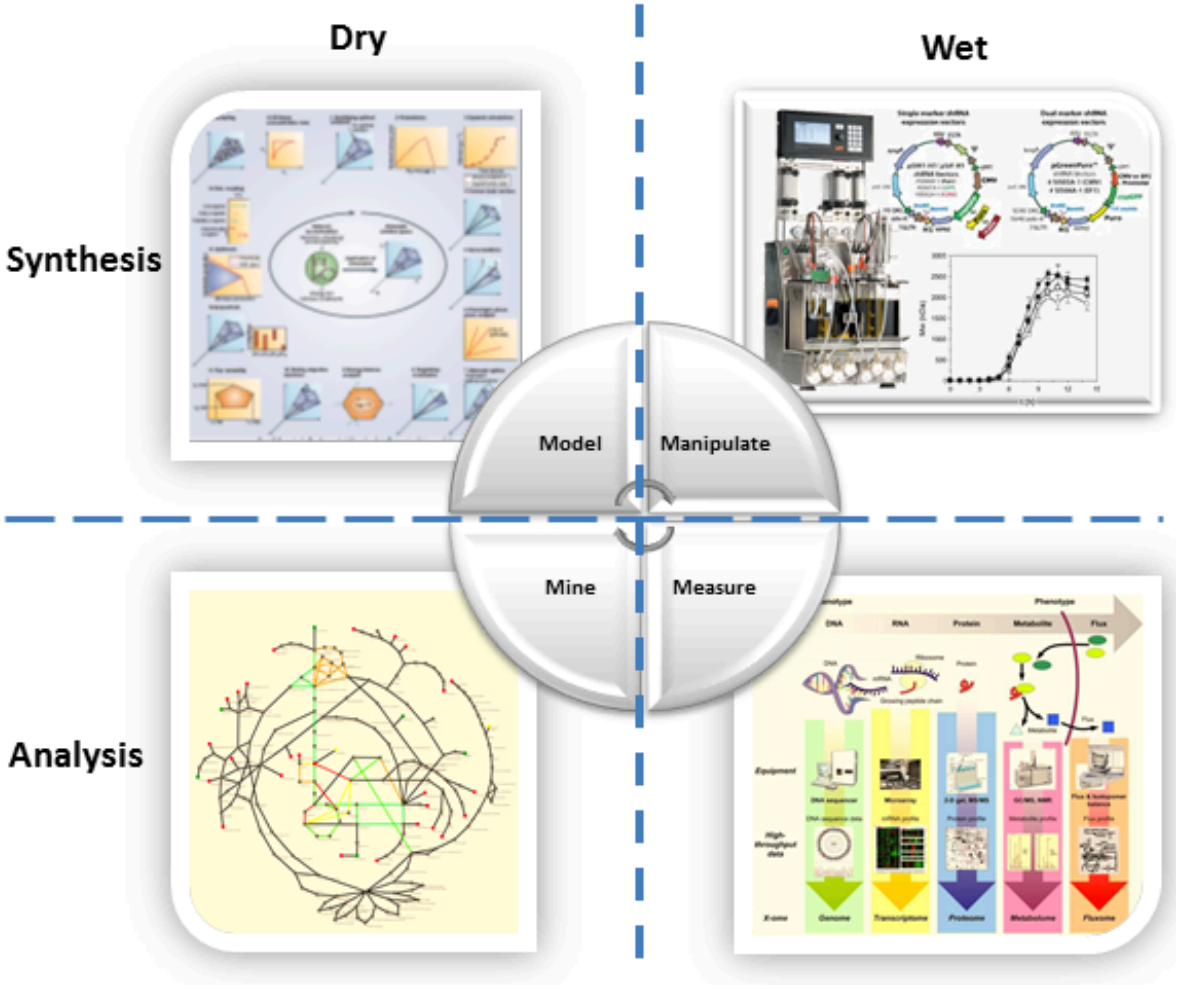
Genomics
DNA
sequencer

4Ms:

Measurement,
Mining,
Modeling
& Manipulation

TREY IDEKER, L. RAIMOND WINSLOW & A. DOUGLAS LAUFFENBURGER ('06). "Bioengineering and Systems Biology," Annals of Biomedical Engineering DOI: 10.1007/s10439-005-9047-7

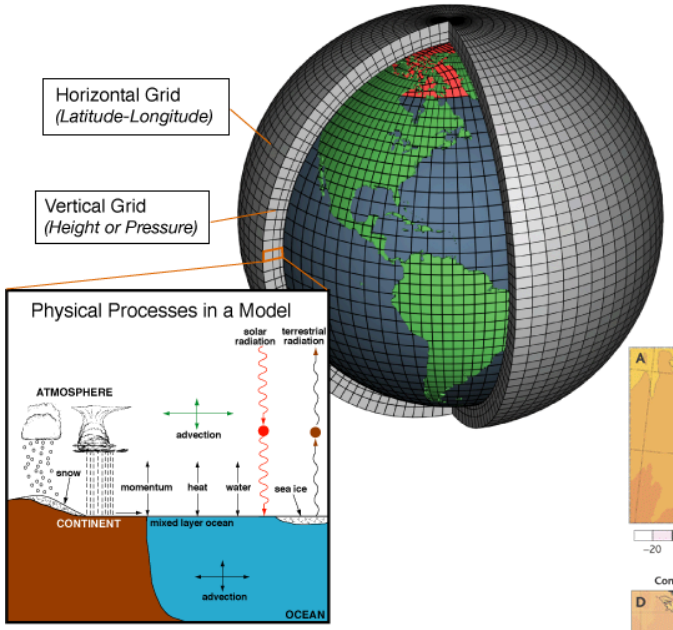
Image from http://web.aibn.uq.edu.au/cssb/ResearchProjects.html



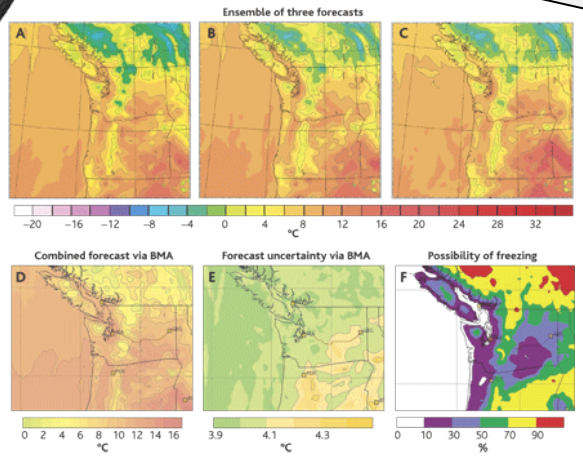
Weather forecasting

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

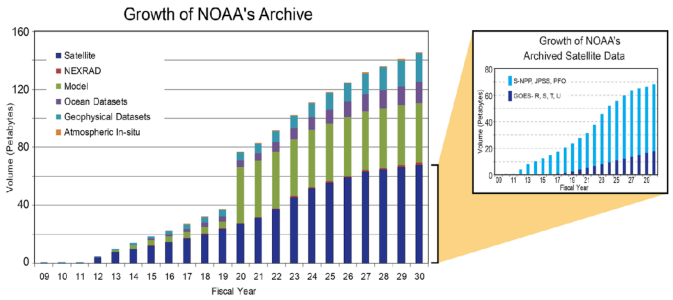
How do they do it?
Physical models & massive sim. useful
(but not sufficient - think “butterfly” effect.)
Large-scale data collection via sensors

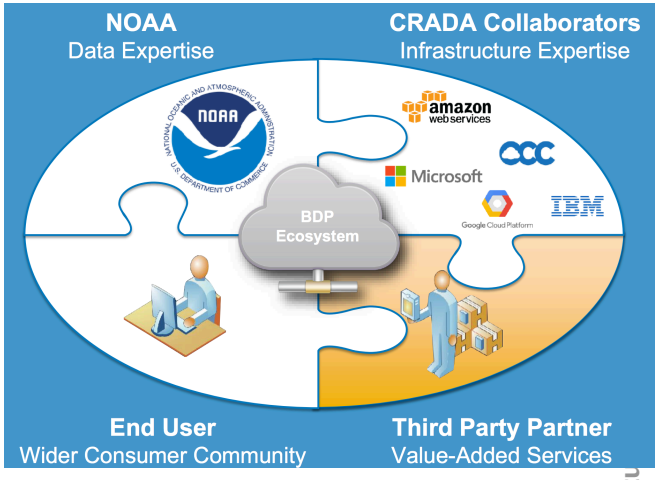


1964, first climate model



90s, ensemble methods





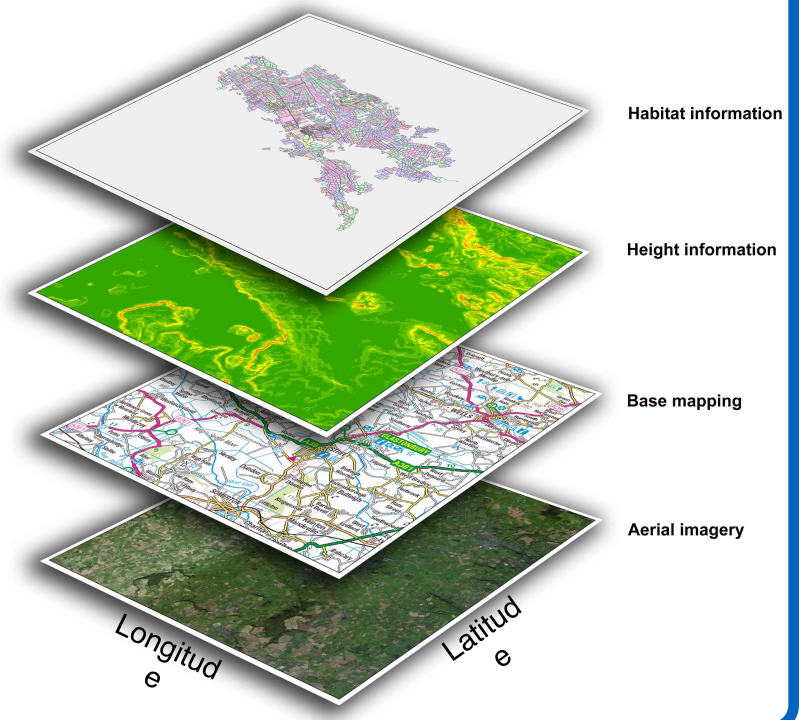
2010s, big data project

Biomedical Data Science

- The ambition of **data map** & eventually model of the genome, connectome, organs…
- The recent success of genomics (to highlight) but maybe a **changing landscape**
- How **scaling** is integral to the changing landscape
- Using large-scale data as an **anchor** for heterogeneous phenotype/medical data

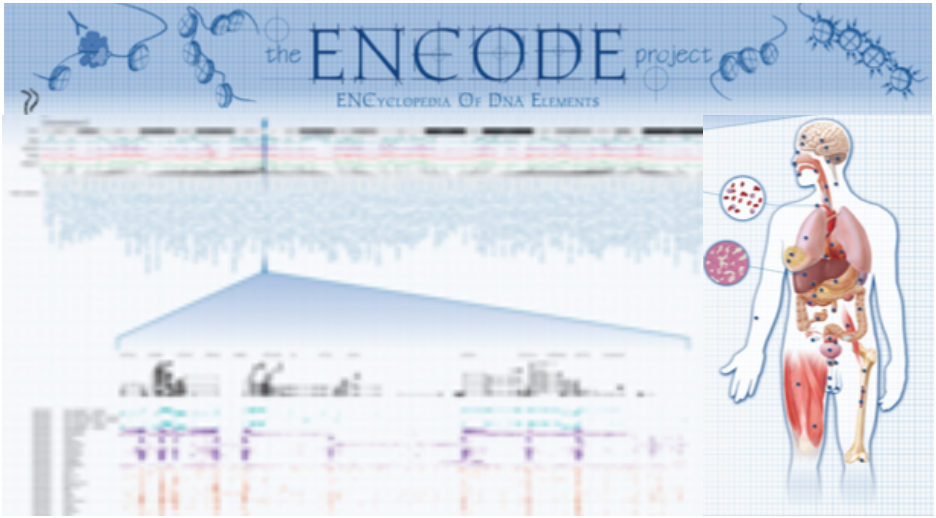
Human genome annotation — a non-intuitive map

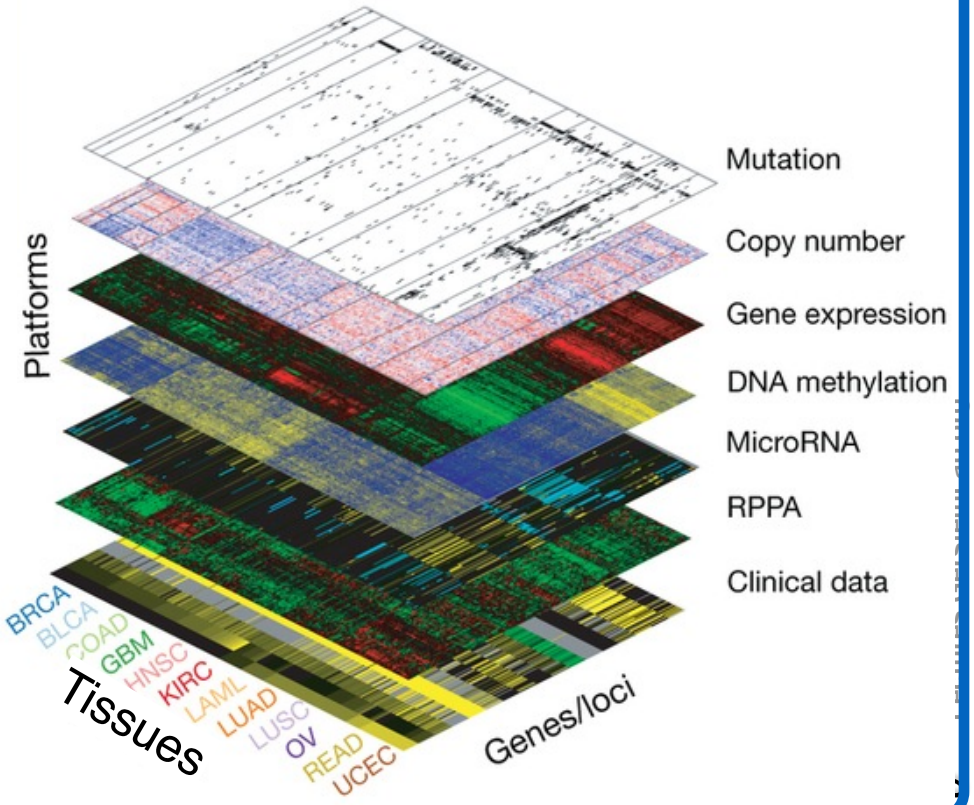
geographical information



- Large-scale organisation providing an overview of the genome
- Integration of heterogeneous data

genomic information



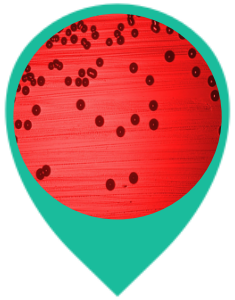


Biomed. Data Sci. via Example: Huge Success in Amassing Genotype-Phenotype Relationships

1953

Double Helix

The discovery of double helix by James Watson and Francis Crick



1995

Sequenced Genome

Haemophilus influenzae as the first organism's genome completely sequenced

2008

Thousand Genomes

By far the most detailed catalogue of human genetic variation

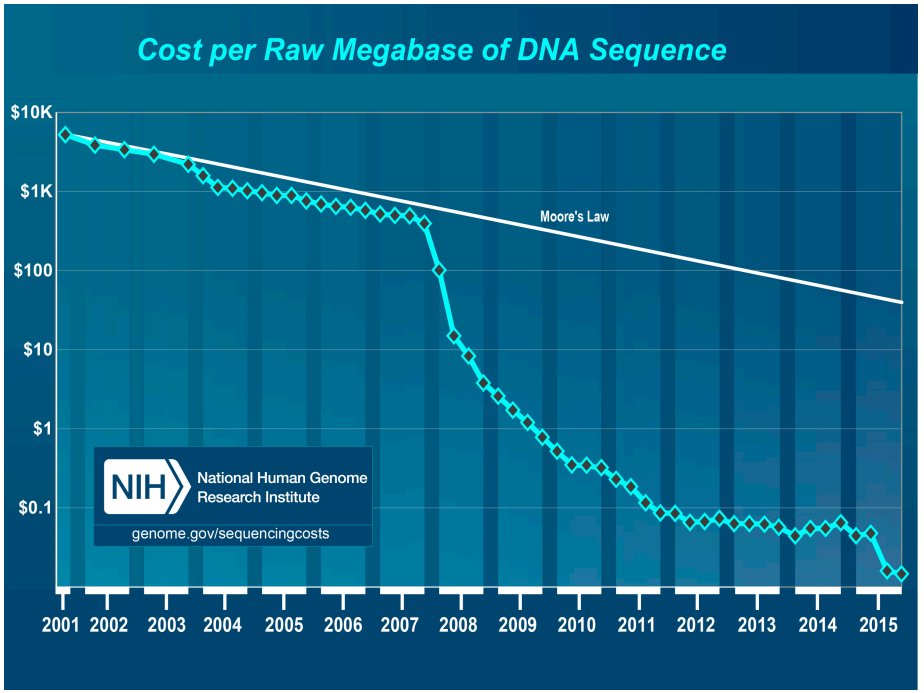
2015

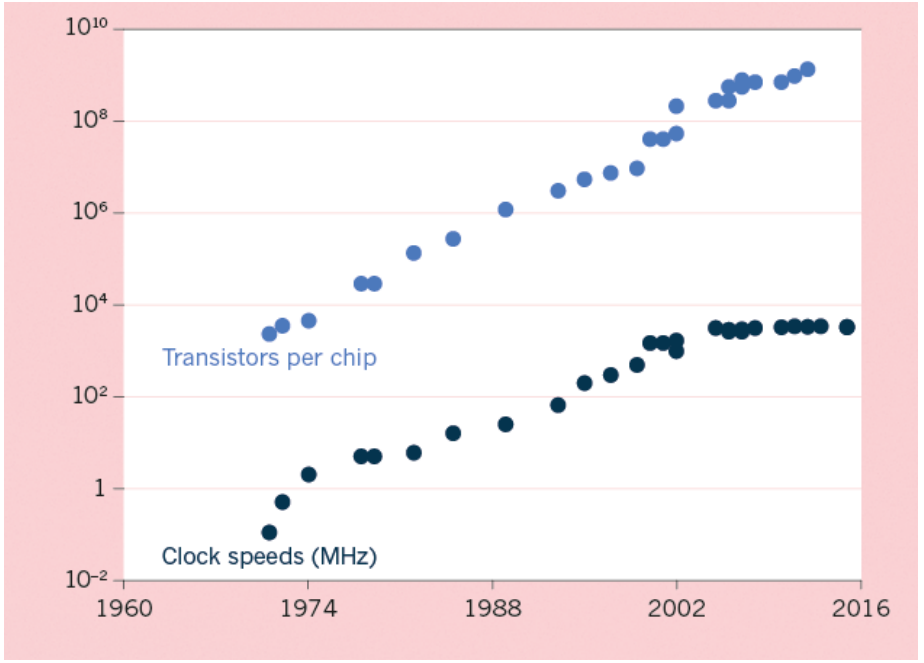
Integrated health data

Study with over 0.5M participants collecting integrated data from genotypes to phenotypic details and clinical information

Sequencing Data Explosion:

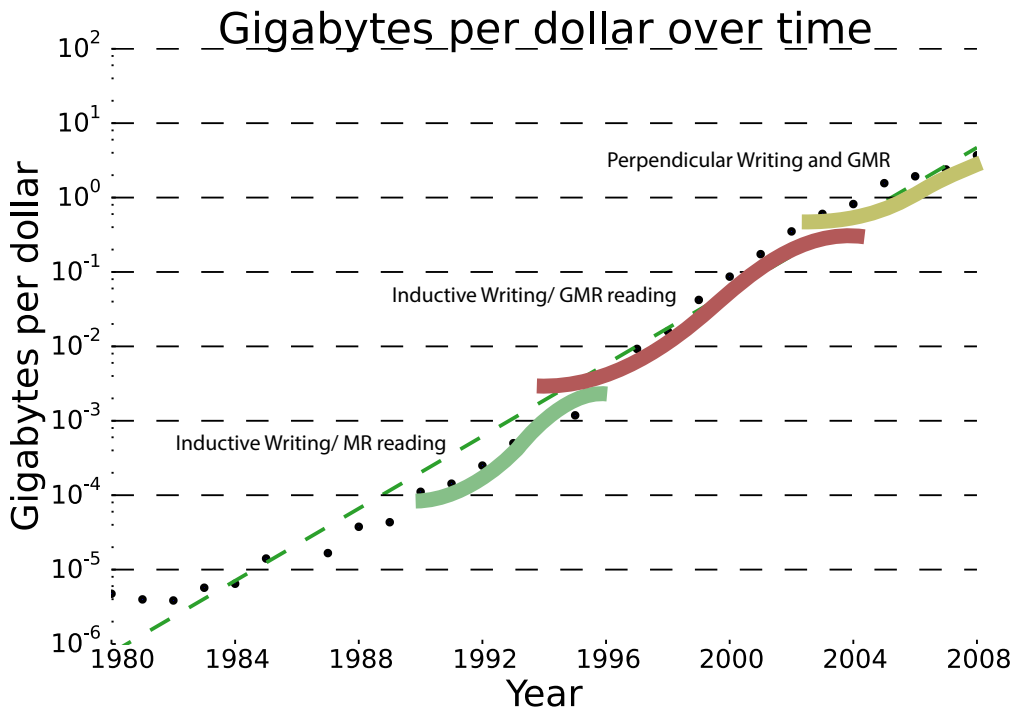
Powered by hyper-exponential incr. in data & exponential increase in computing (Moore's Law)

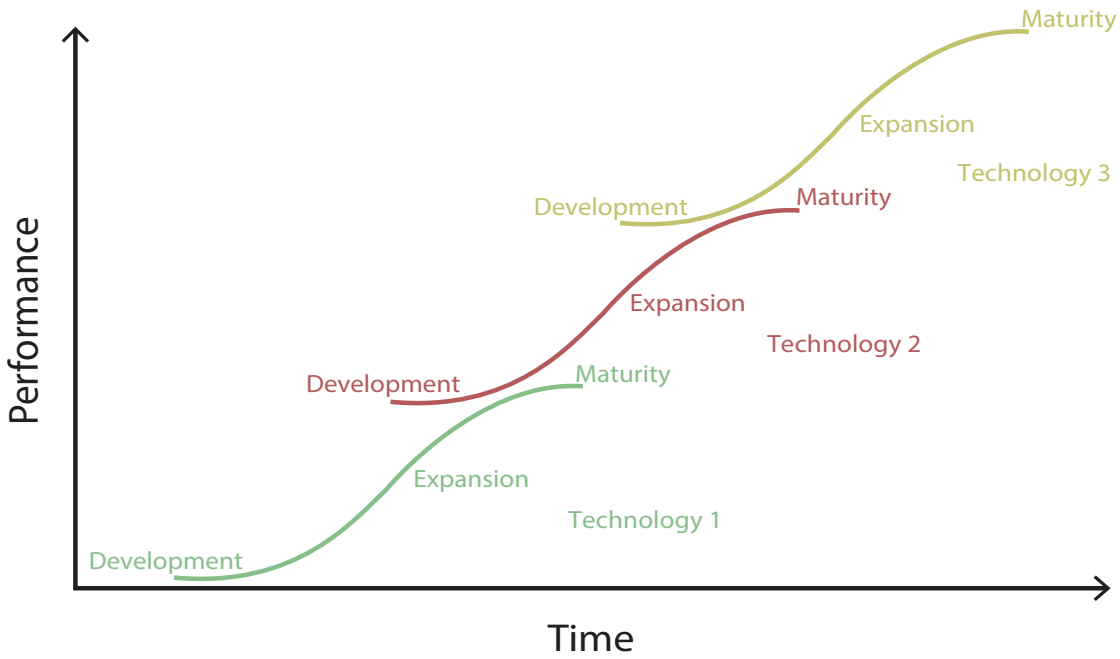




Kryder's Law and S-curves underlying exponential growth

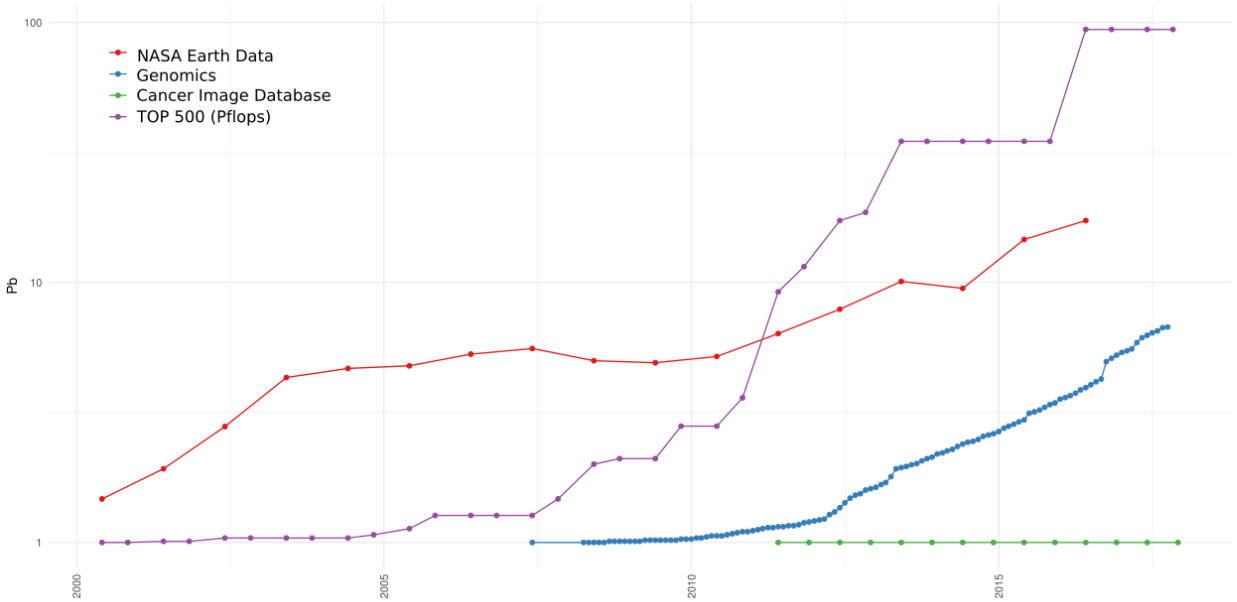
- Moore's & Kryder's Laws
  - As important as the increase in computer speed has been, the ability to store large amounts of information on computers is even more crucial
- Exponential increase seen in Kryder's law is a superposition of S-curves for different technologies

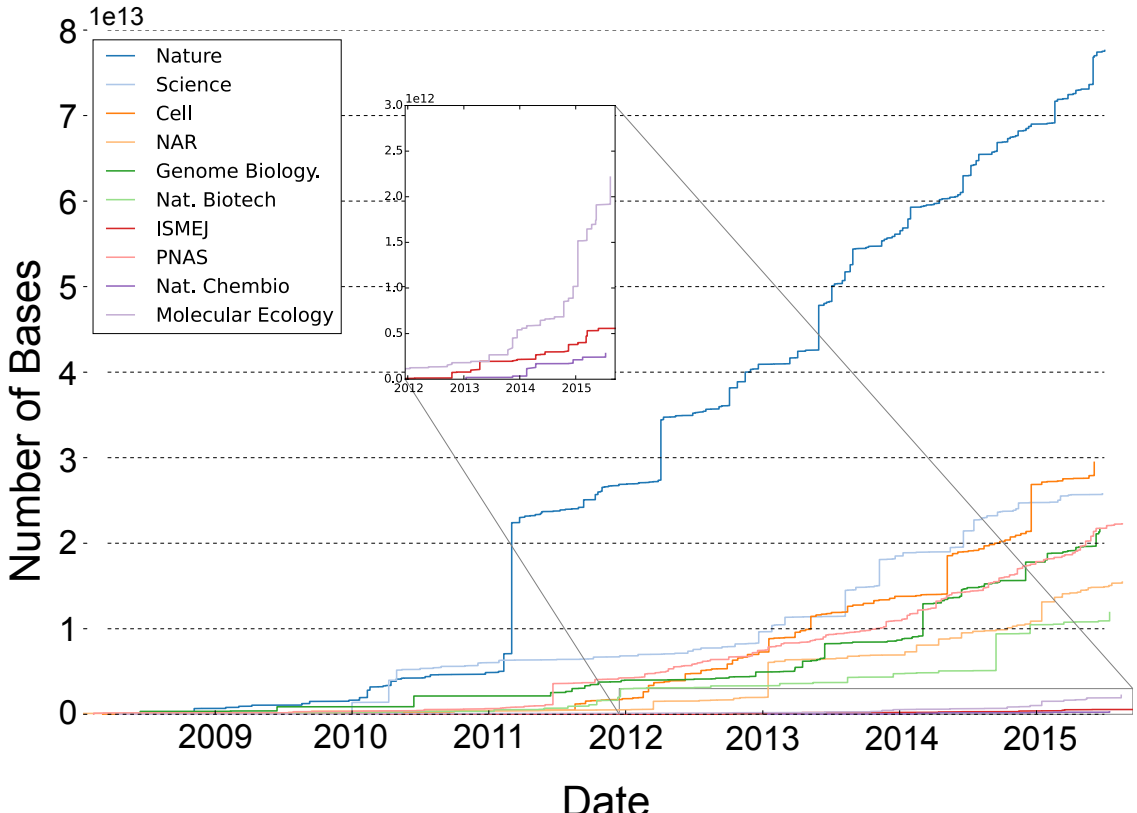




[Muir et al. ('15) GenomeBiol.]

Sequencing cost reductions have resulted in an explosion of data

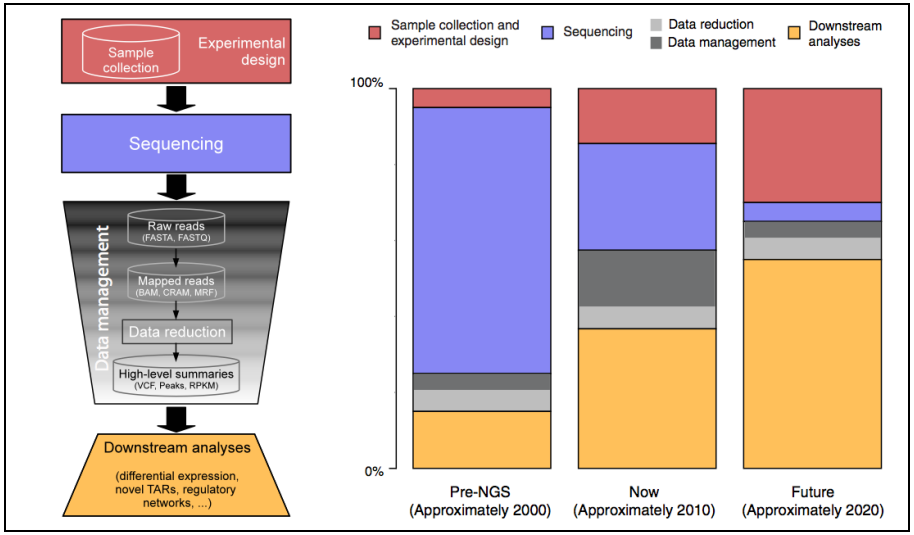




- The type of sequence data deposited has changed as well.

[Muir et al. (‘15) GenomeBiol.]

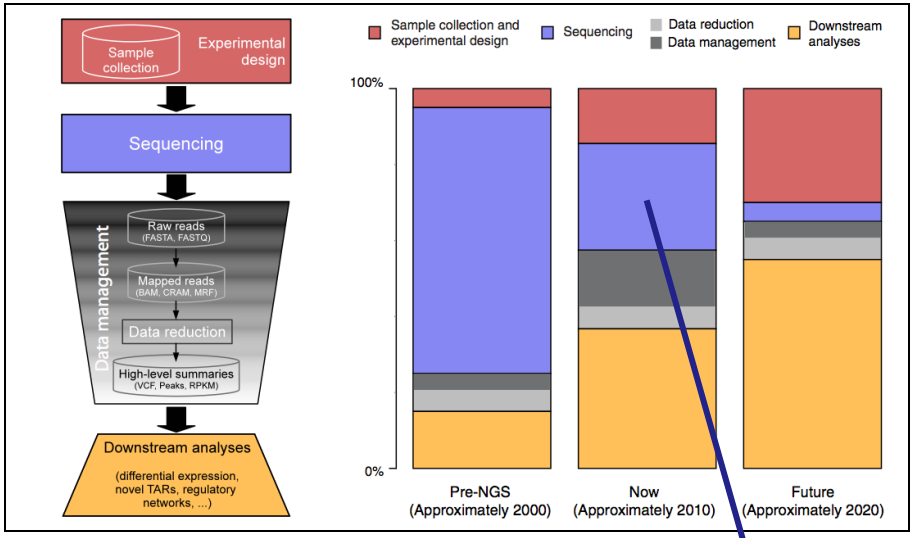
The changing costs of a sequencing pipeline



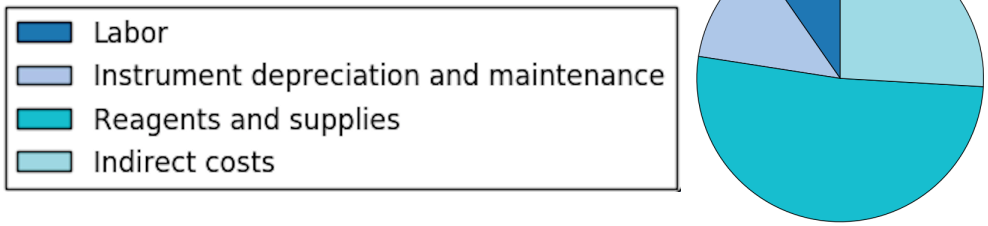
From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

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

The changing costs of a sequencing pipeline

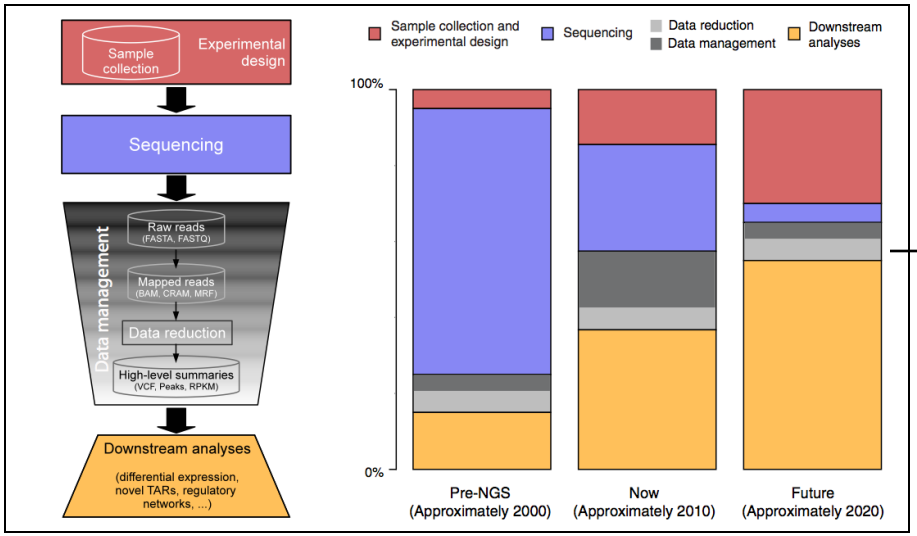


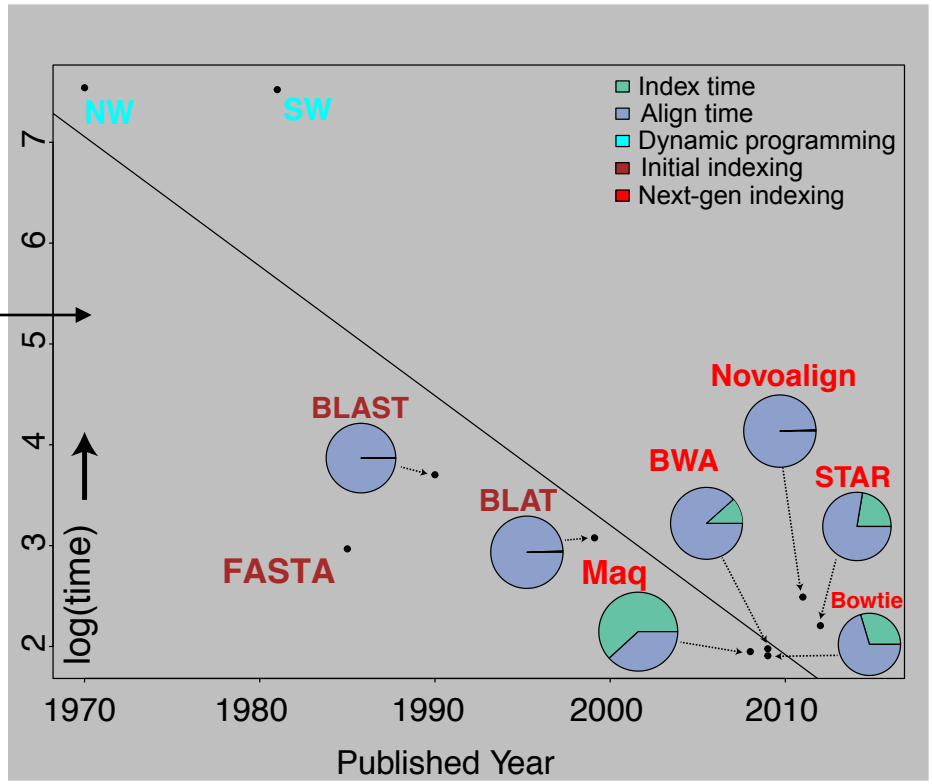
From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis



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

The changing costs of a sequencing pipeline



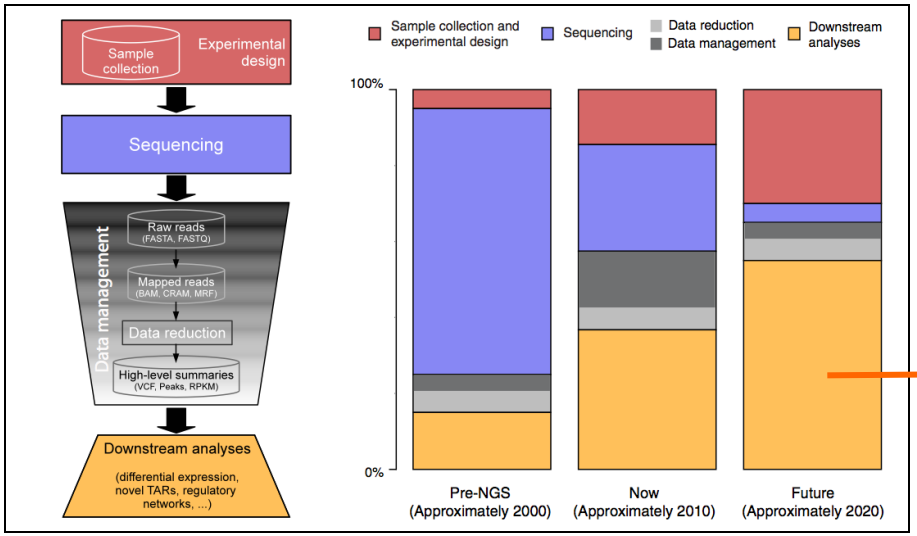


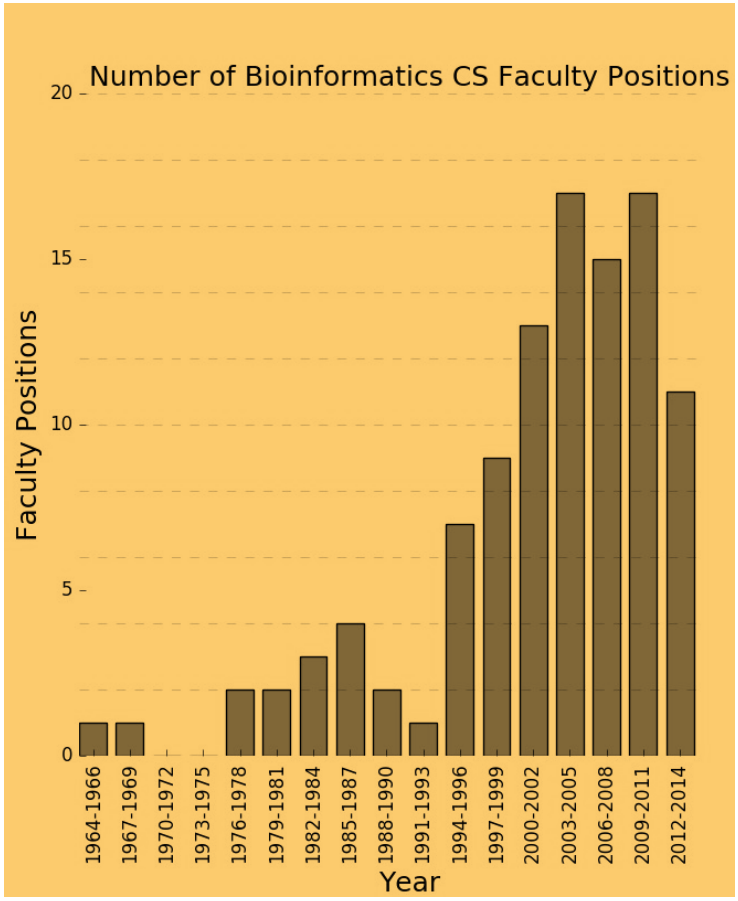
From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

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



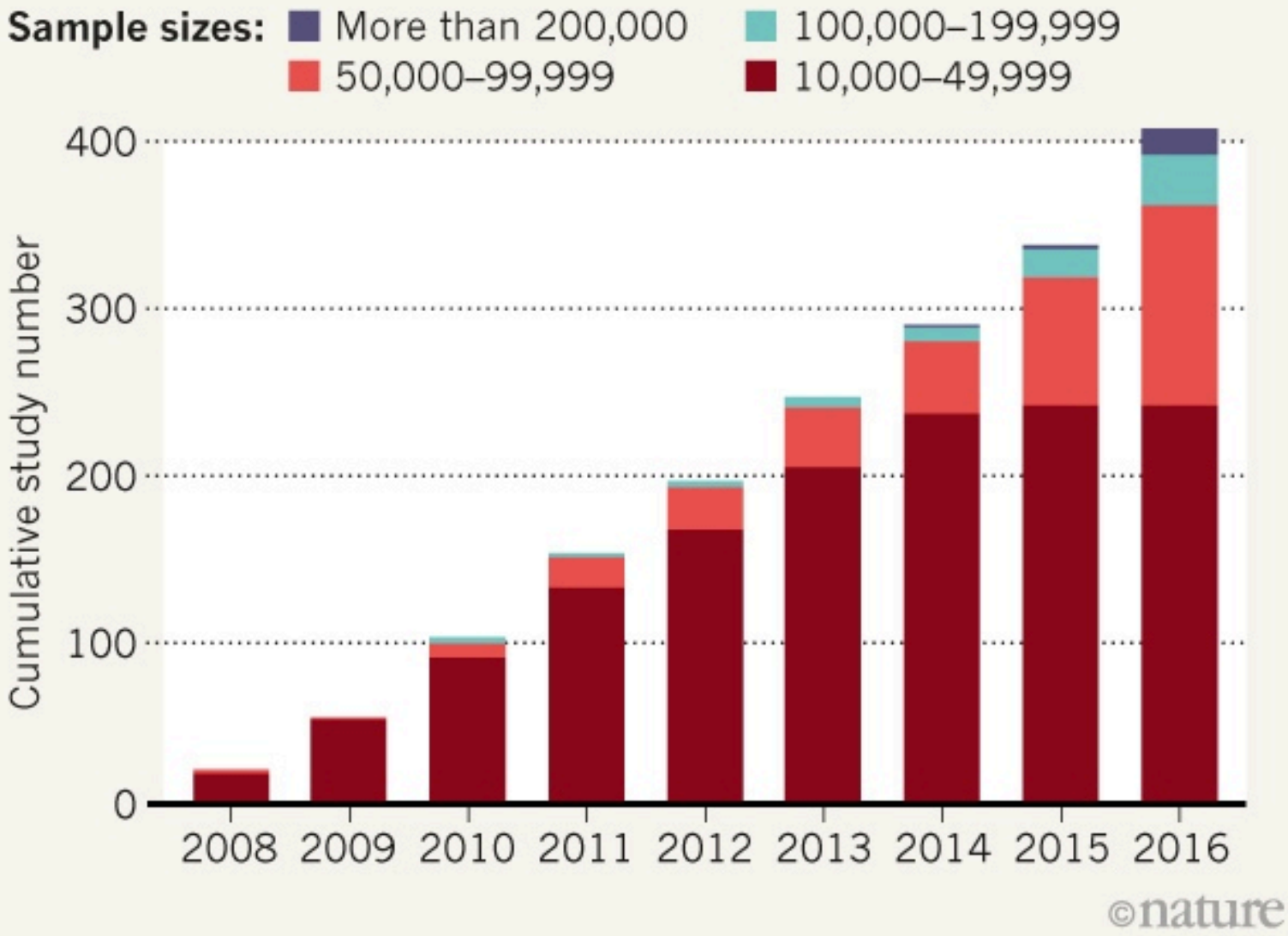


From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

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

THE GENOME-WIDE TIDE

Large genome-wide association studies that involve more than 10,000 people are growing in number every year — and their sample sizes are increasing.

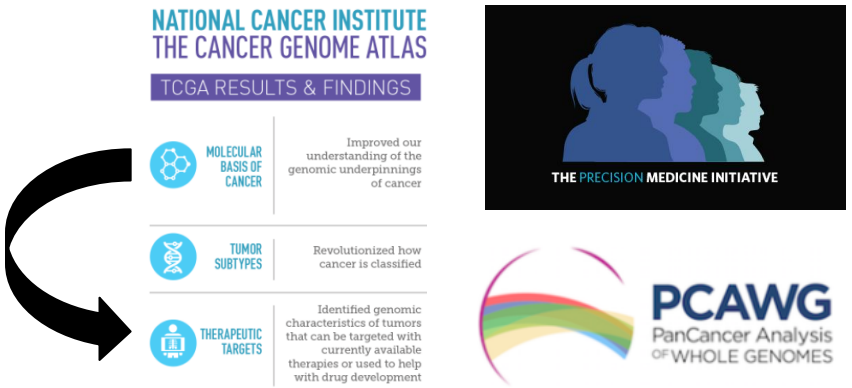


Basic Science to Medical Practice

Research Initiatives and Biomedical Startups

Large-scale genomics data as a anchor to organize large amounts of phenotype data – EMRs, wearables…

INITIATIVES



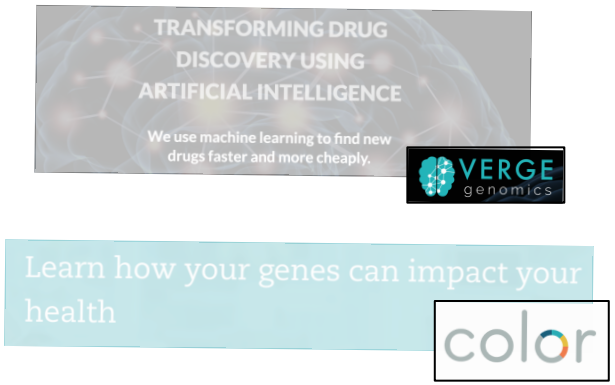
1.Genomics of disease-focused cohorts; GWAS [2002-present], TCGA, PCAWG [2006-present]

2.Integration of genomic data with rich clinical phenotypes; UKBiobank, All of Us [2016-present]

3.Integration of genomic data in EMRs for clinical decision support & wearables; [Near future]

4.Home-based routine sequencing of DNA and RNA in blood as part of preventive care [Speculative future]

STARTUPS



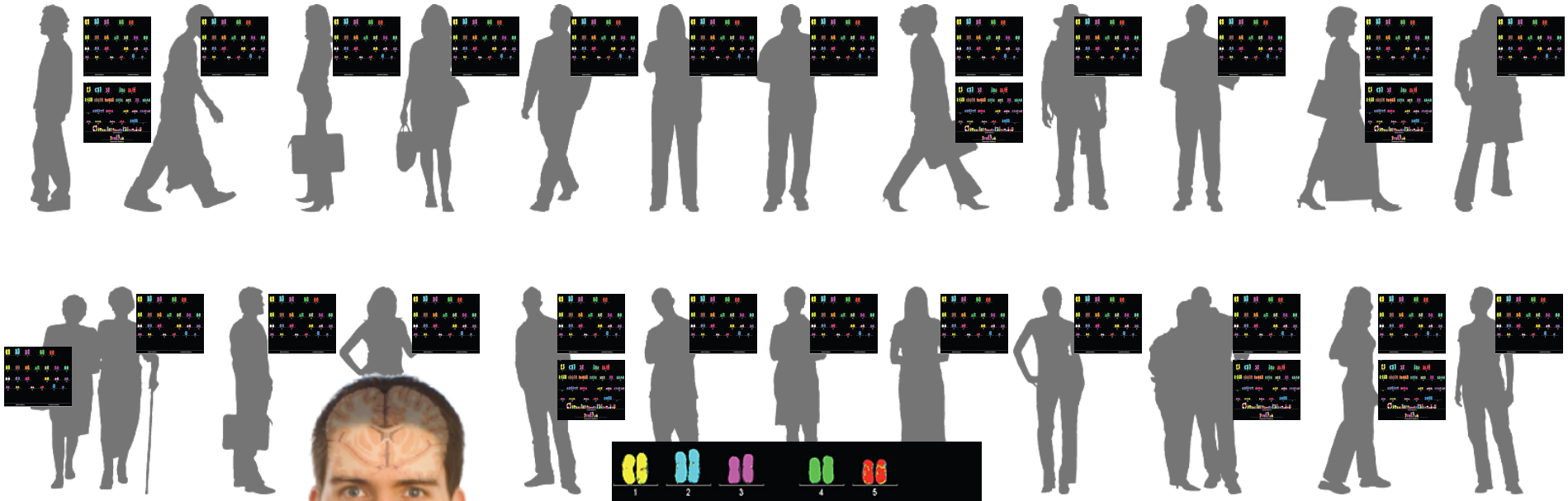
Medical Big Data: Promise and Challenges (Lee and Yoon , *Kidney Res. Clin. Pract.*, 2017)

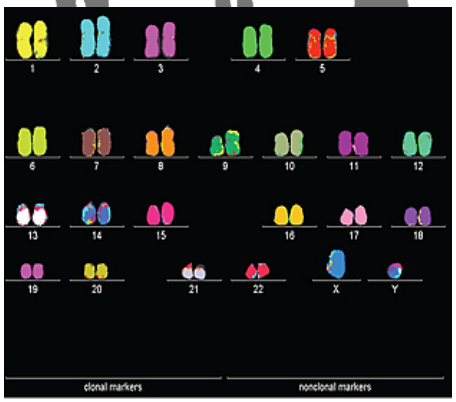
EX of 'omics research on focused patient cohorts: Many Yale Researchers Involved in Neurogenomics

- Involved national initiatives:
  psychENCODE, CMG,
  BrainSpan, BSMN,
  NIDA Neuroproteomics

- Yale investigators:
  M Gunel, N Sestan, F Vaccarino, J Noonan,
  J Gelernter, A Nairn

- DNA variants, altered protein & RNA levels in brains in development & various diseases (eg ASD, SCZ)

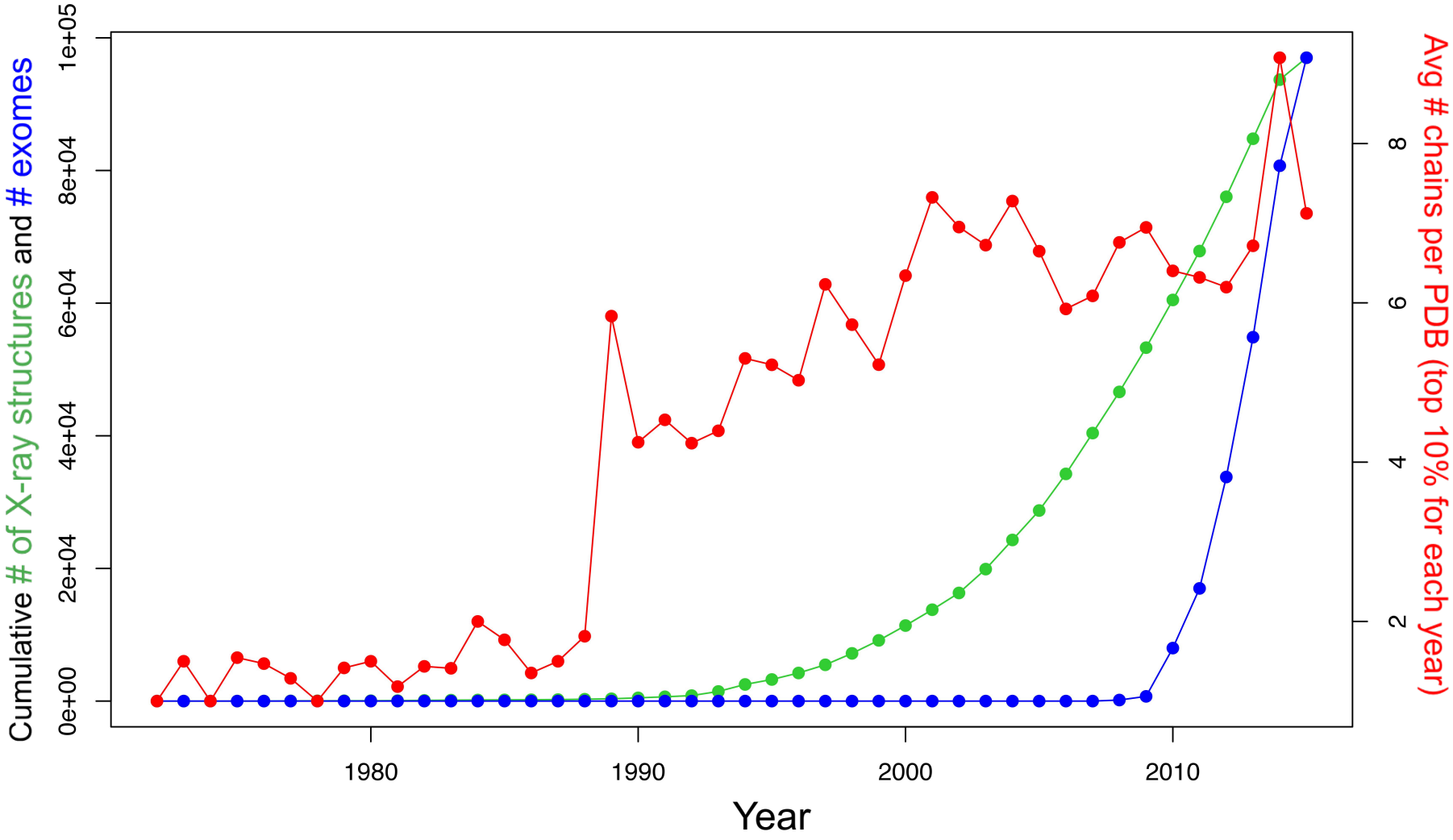




Placing the individual into the context of the population & using the population to build a interpretative model

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

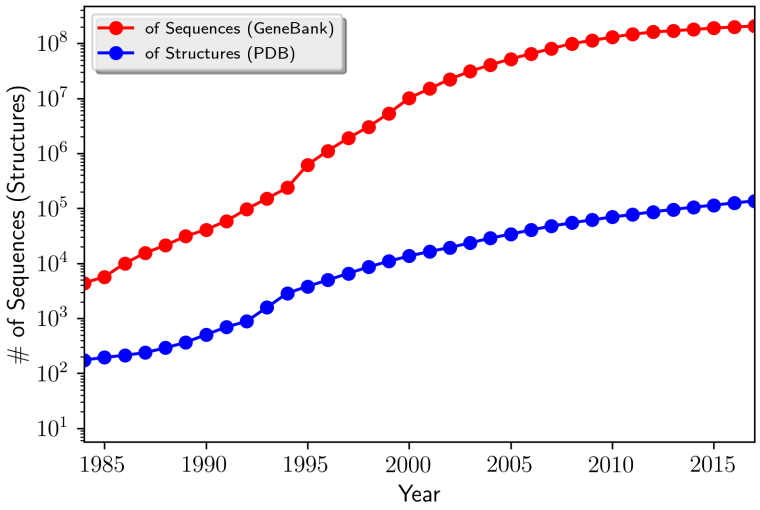
The volume of sequenced exomes is outpacing that of structures, while solved structures have become more complex in nature.

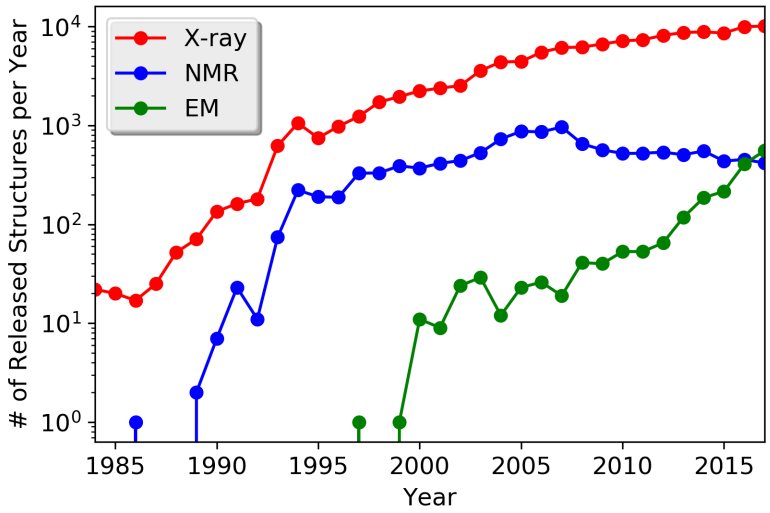


Exome data hosted on NCBI Sequence Read Archive (SRA)

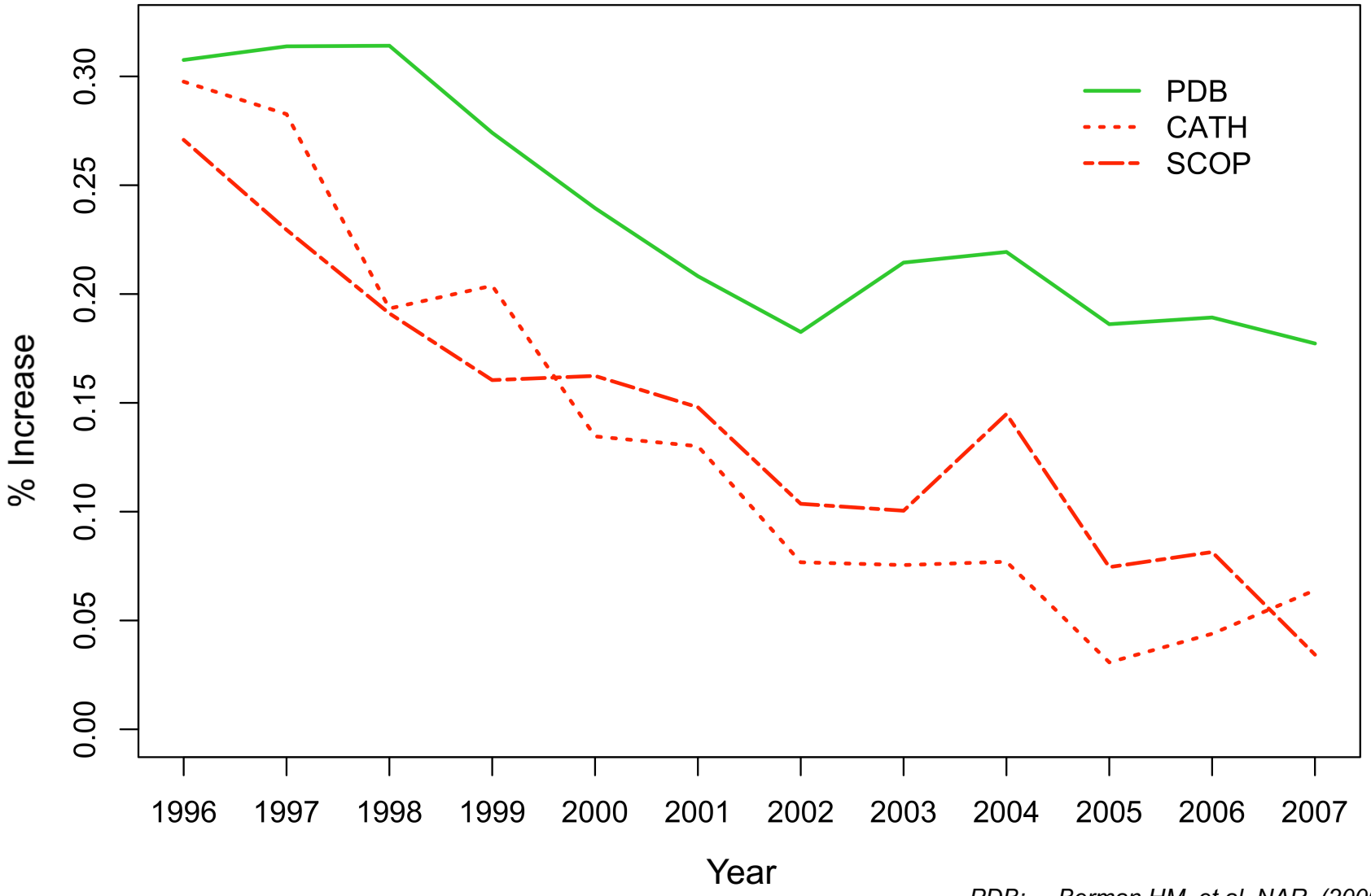
[Sethi et al. COSB ('15)]

Experimental determination of 3D structures can not keep up with the explosive growth of sequence information
The Electron Microscopy (EM) has emerged as a powerful tool in determining 3D structures





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

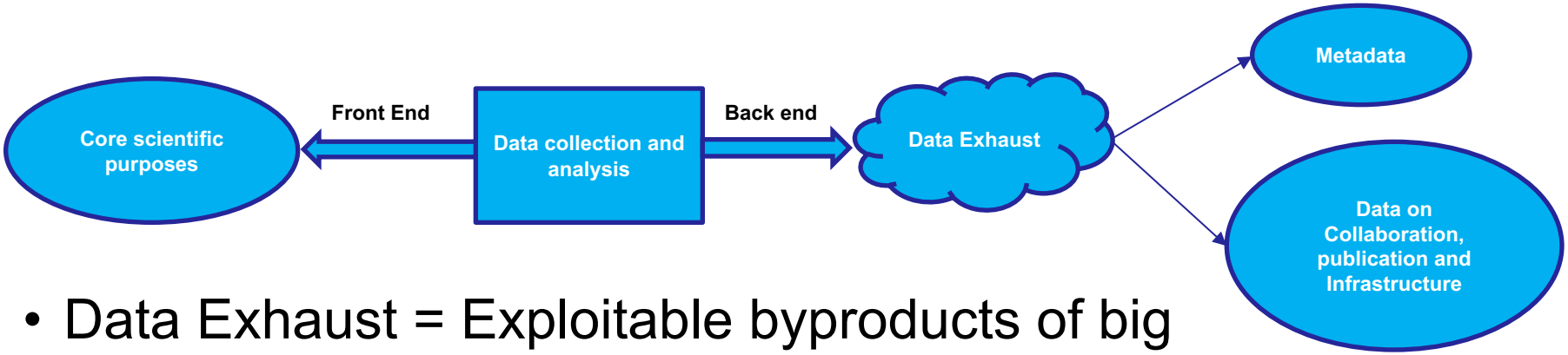


[Sethi et al. COSB ('15)]

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

Core Qs v Creative Use of the Data

Data Exhaust



- Data Exhaust = Exploitable byproducts of big data collection and analysis
- Creative use of Data is key to Data Science !
- Aspects of Privacy but also Science of Science

[PHOTO: RELAXNEWS; from http://www.lapresse.ca]

Genomics: as Data Science sub-discipline

- Developing ways of organizing & mining categorizing information on a large scale
  - Very fundamental & early form of "Big Data", feeding into other enterprises (classification approach, R)
  - Also importing tech. developed in other big data disciplines (Hadoop)