

Biomedical Data Science: An Introduction

M Gerstein

Slides freely downloadable from
Lectures.GersteinLab.org

Twitter meeting hashtag:

#CBDS7Feb

#3 - Simulation

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

Emphasis:
Supercomputers

#4 - Data Mining

Classifying information & discovering unexpected relationships

Emphasis: DB emphasis

Gray died in '07.
Book about came out in '09...

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:

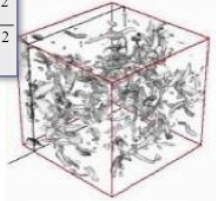
data exploration (eScience)

unify theory, experiment, and simulation

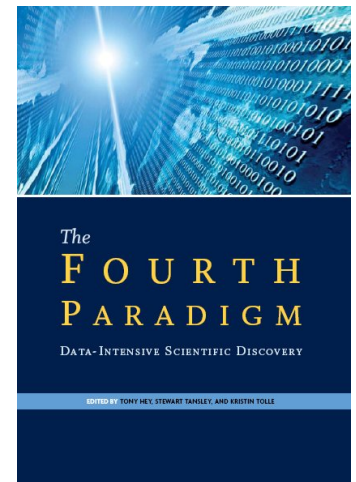
- Data captured by instruments
Or generated by simulator
- Processed by software
- Information/Knowledge stored in computer
- Scientist analyzes database / files
using data management and statistics



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



Historical Perspective on Data Sci. – Jim Gray



What is Data Science, Today?

- Data Science encompasses the study of the entire **lifecycle of data**
 - Understanding of how data are **gathered**
 - Knowledge of what data sources are available & how they may be synthesized
 - The **storage**, access, annotation, management, & transformation of data
- Data Science encompasses many aspects of primary **data analysis**
 - Statistical inference, machine learning, & the design of algorithms and computing systems that enable **data mining**
 - Connecting this mining where possible with analytic **modeling**
 - The presentation & intuitive **visualization** of analysis results
 - The use of data analysis to make **practical decisions**
- Consideration of secondary aspects of data (eg the **data exhaust**)
 - **Creative secondary uses**, not related to intent at collection
 - EG Science of science
 - The appropriate protection of **privacy**

- Commercial Data
 - Social activity generated
 - Ads, supply optimization, &c
 - Integral to success of GOOG, AMZN, FB, WMT...

- Scientific Data
 - Pre-dated commercial mining
 - Large sets generated by special instruments
 - Often created by large teams - not to address single hypothesis but to be mined broadly



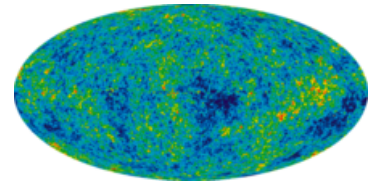
**Data
Science in
Traditional
Science
v
business
world**



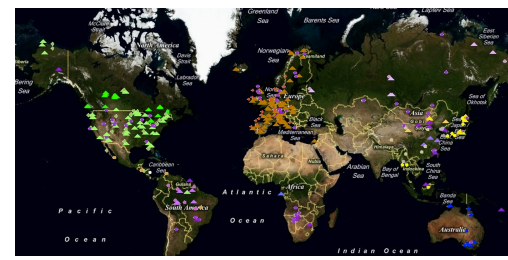
High energy physics - LHC



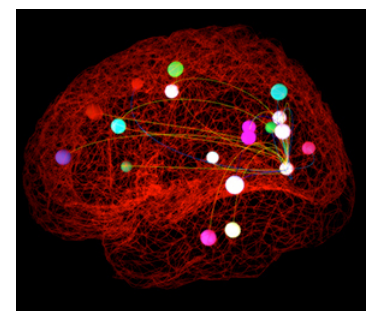
Genomics - DNA sequencer



Astronomy - Sloan Digital Sky survey



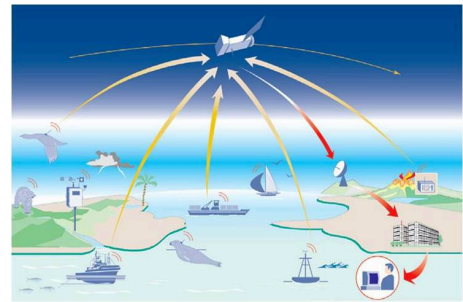
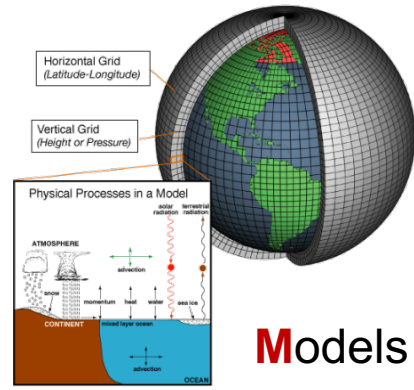
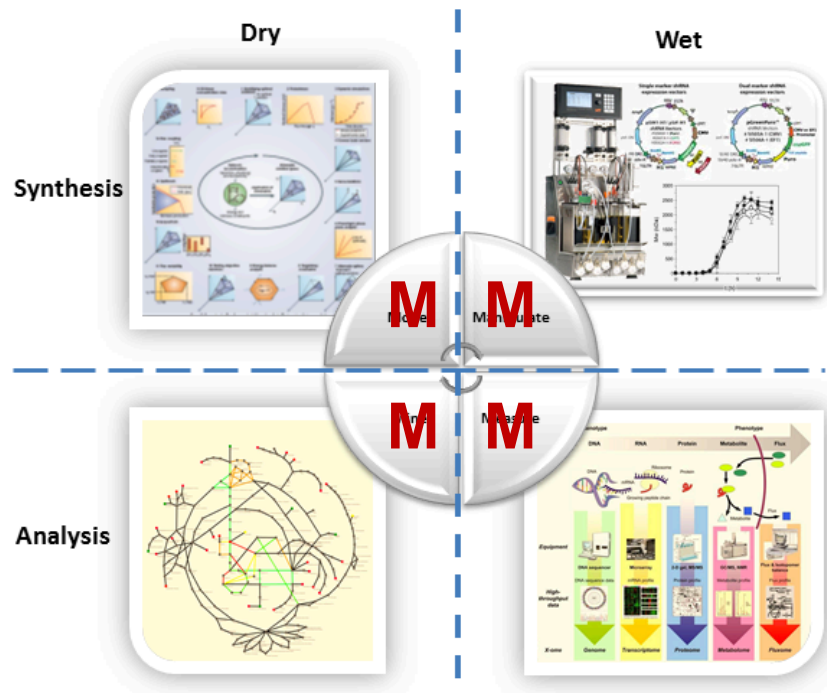
Ecology & Earth Sci. - Fluxnet



Neuroscience - Connectome Project

Coupling of Scientific Data to Models & Experiments

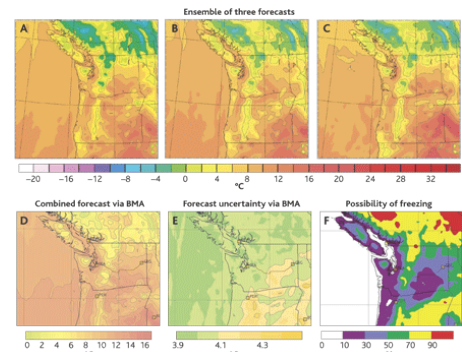
- Scientific data often coupled to a physical/biological model
- Lauffenburger's Sys. Biol. **4Ms**:
Measurement, **M**ining, **M**odeling & **M**anipulation
 (Ideker et al.'06. Annals of Biomed. Eng.)
- Weather forecasting as an exemplar
 - Physical models & simulation useful but not sufficient ("butterfly"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



Models + **D**ata **M**ining

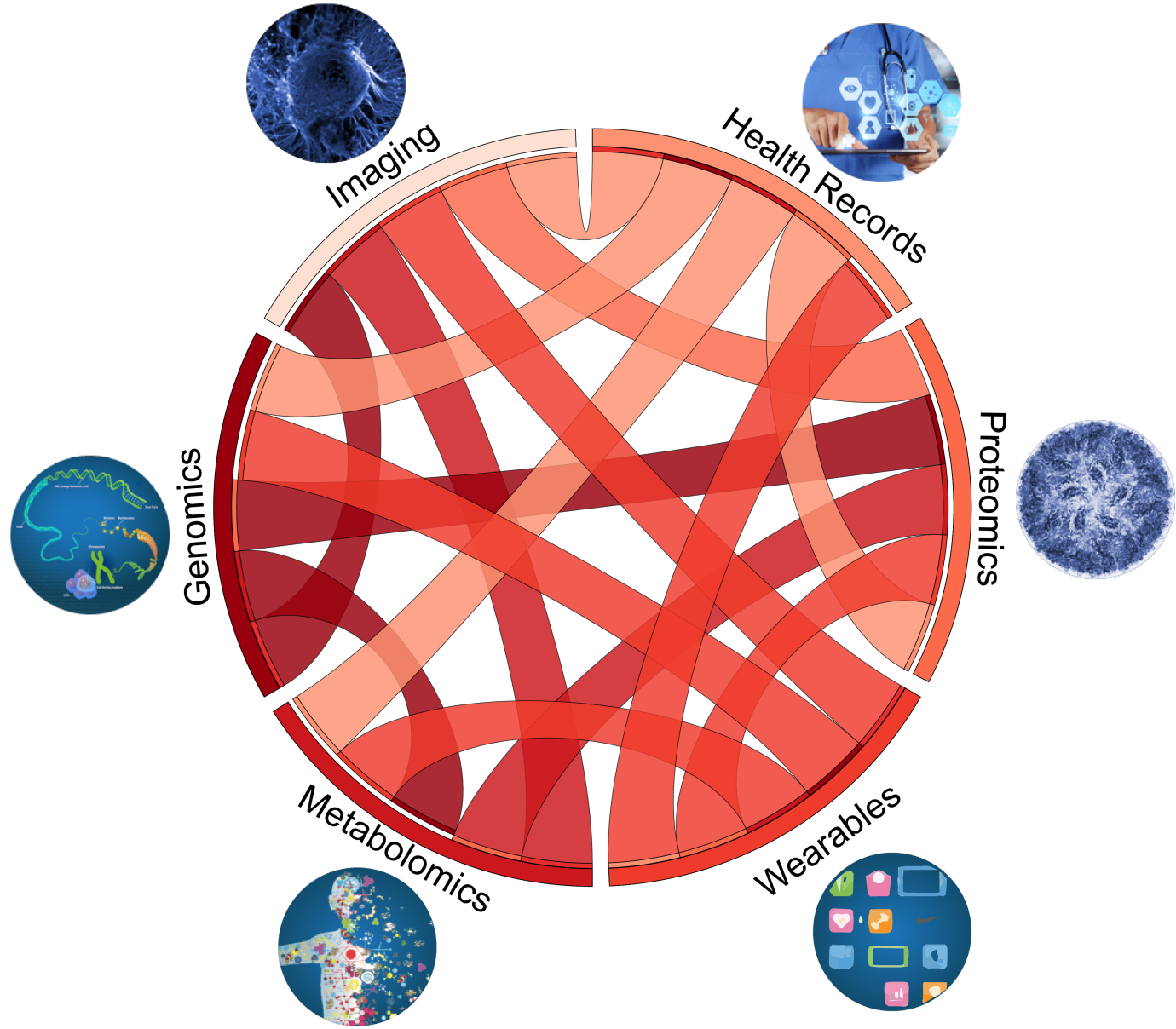


Forecasts

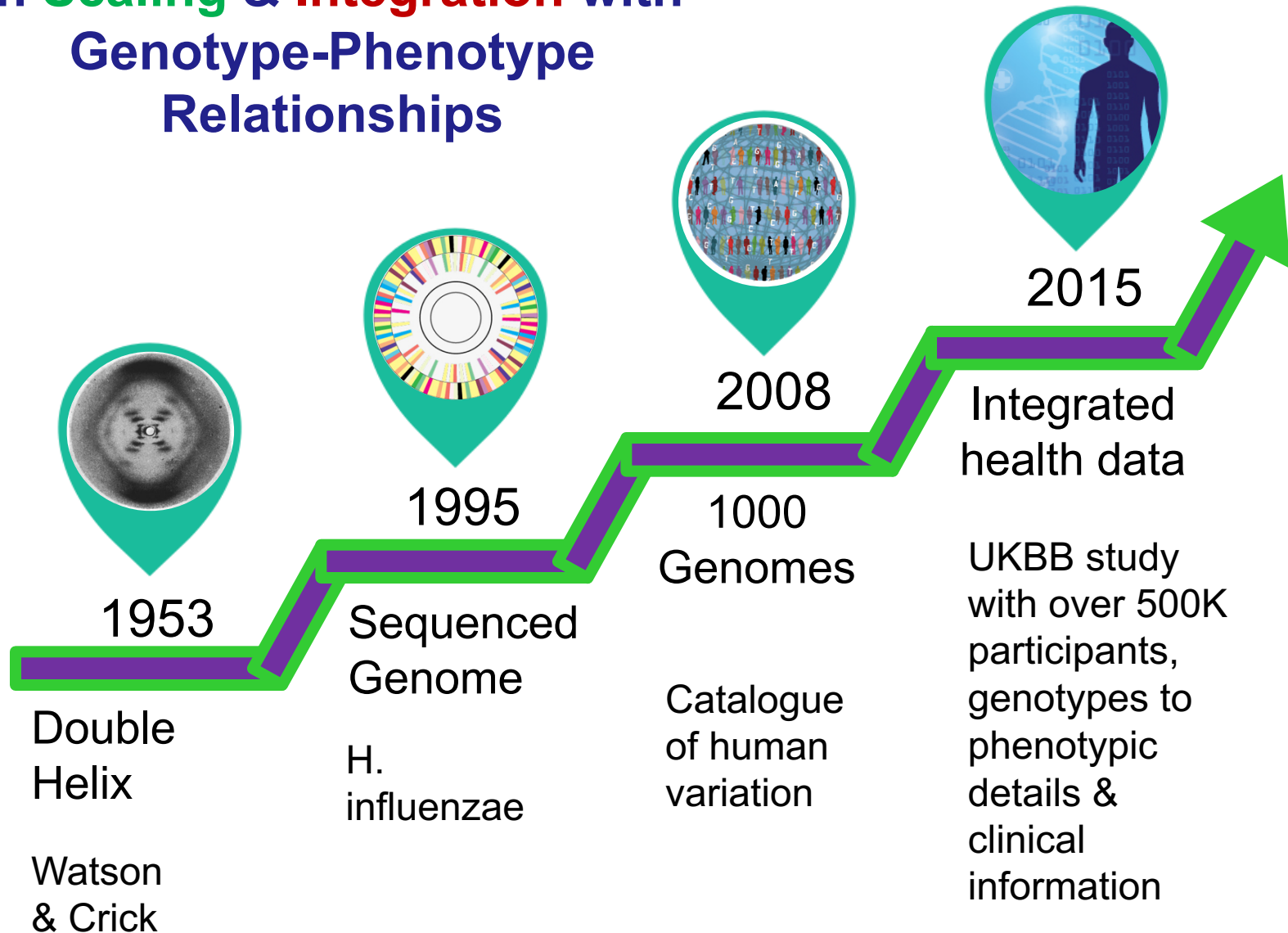


Drivers of Biomedical Data Science

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



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

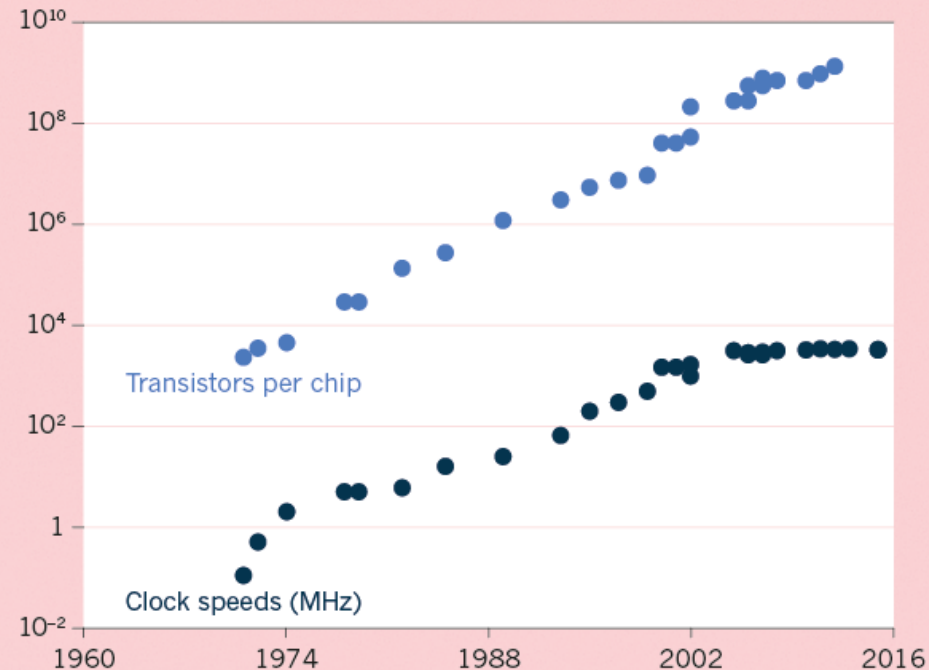
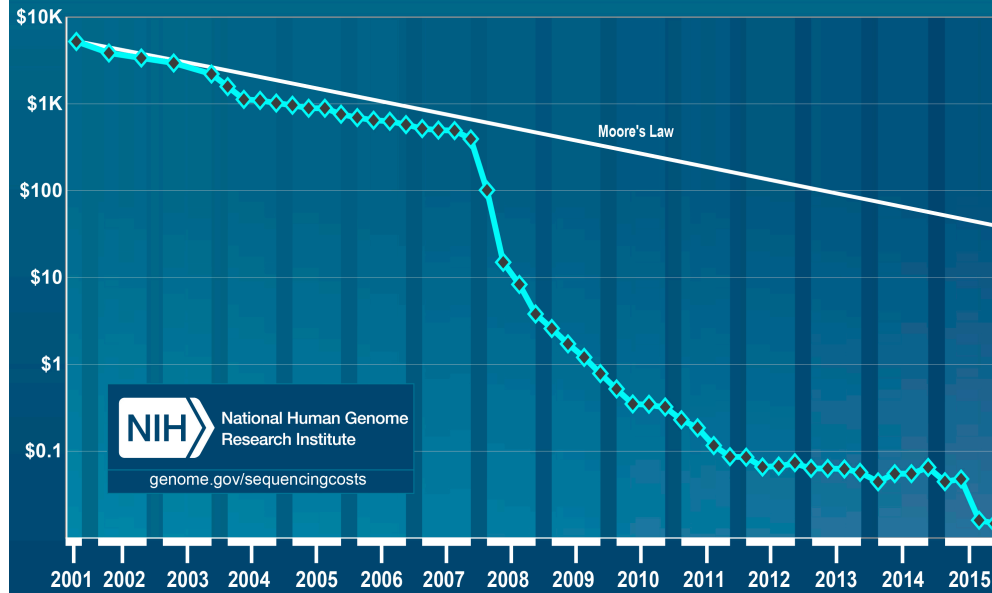


The **Scaling** of Genomic Data Science:

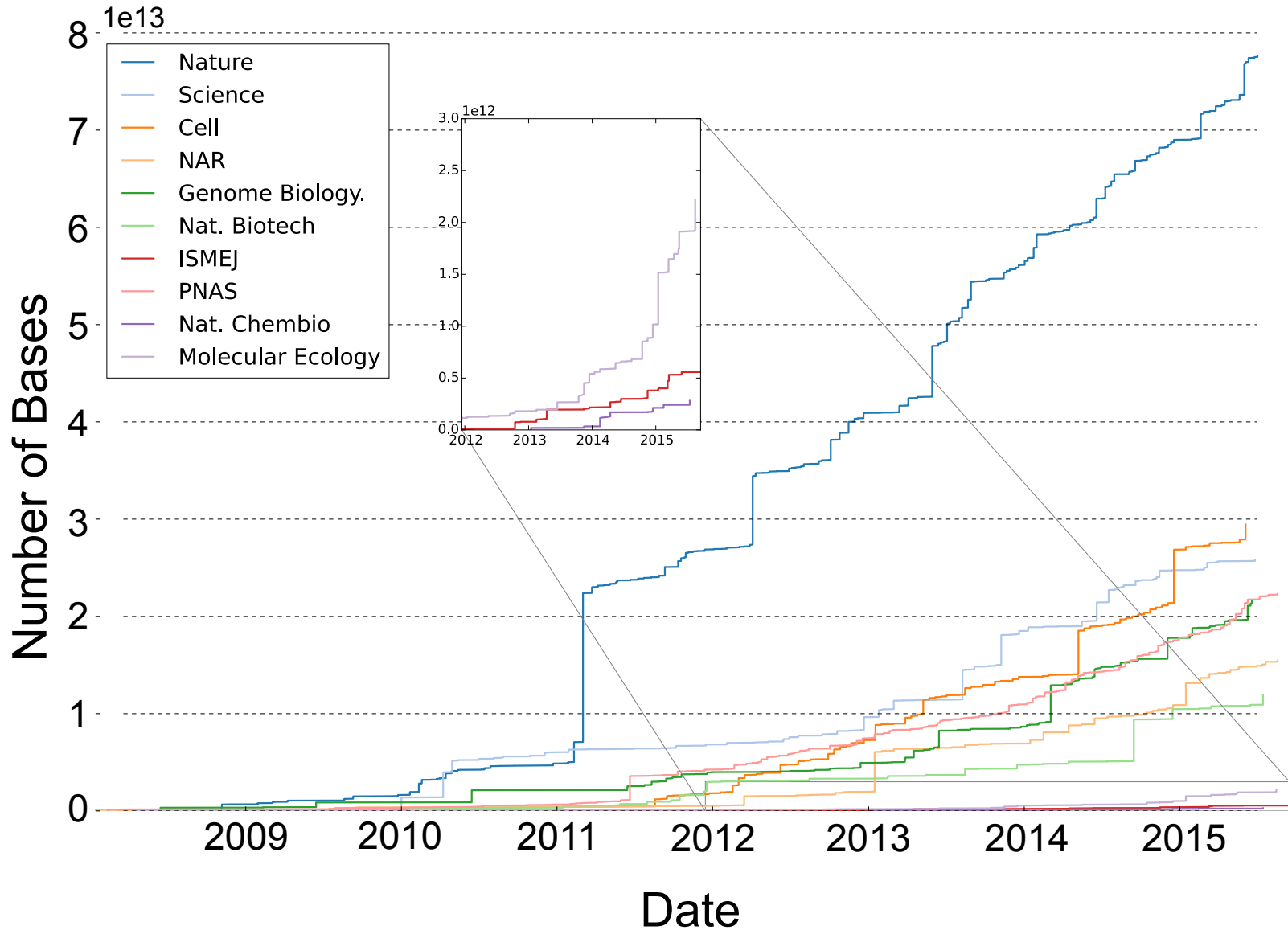
Powered by exponential increases in data & computing

(**Moore's Law**)

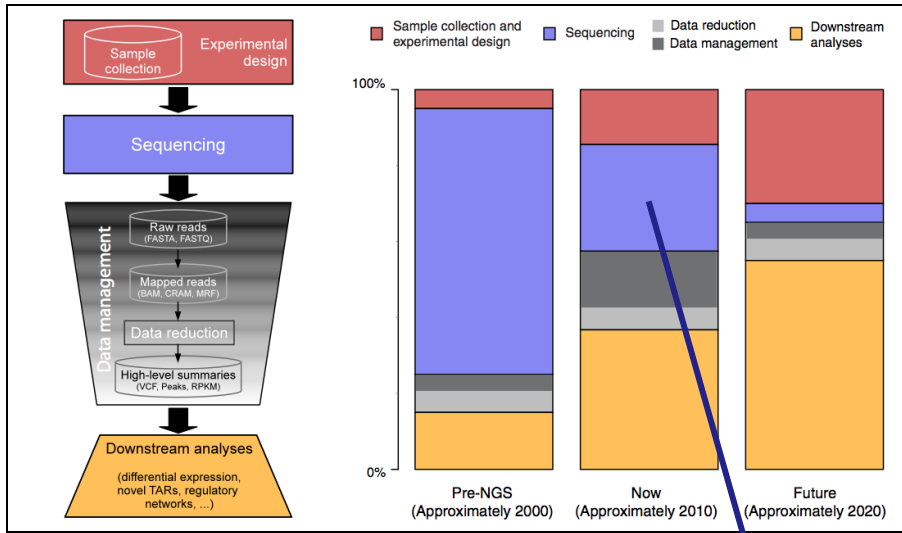
Cost per Raw Megabase of DNA Sequence



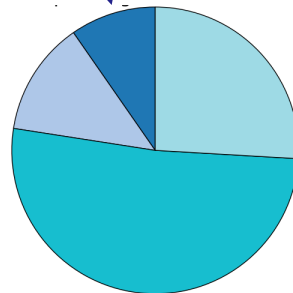
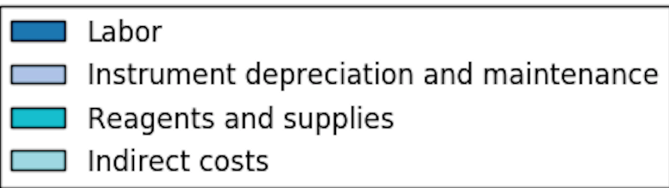
Exponential **Scaling** Changes Fields Using Genomic Data



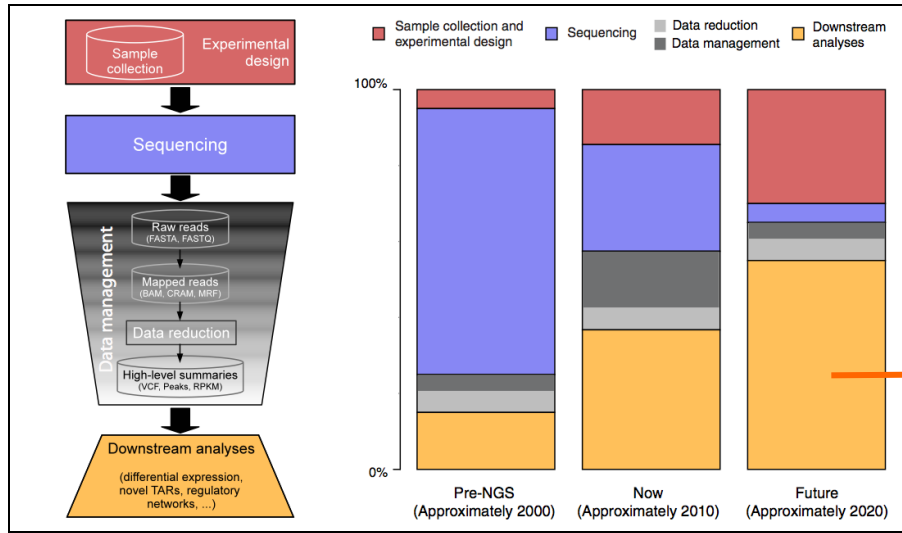
Exponential **Scaling** changes Genomics Itself



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

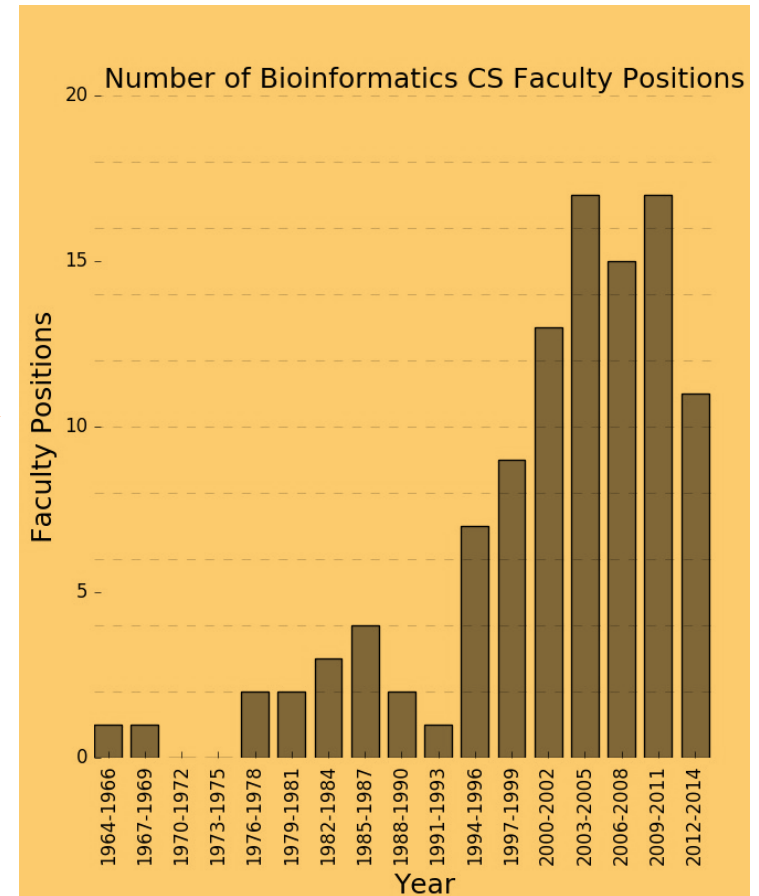


Exponential **Scaling** changes Genomics Itself

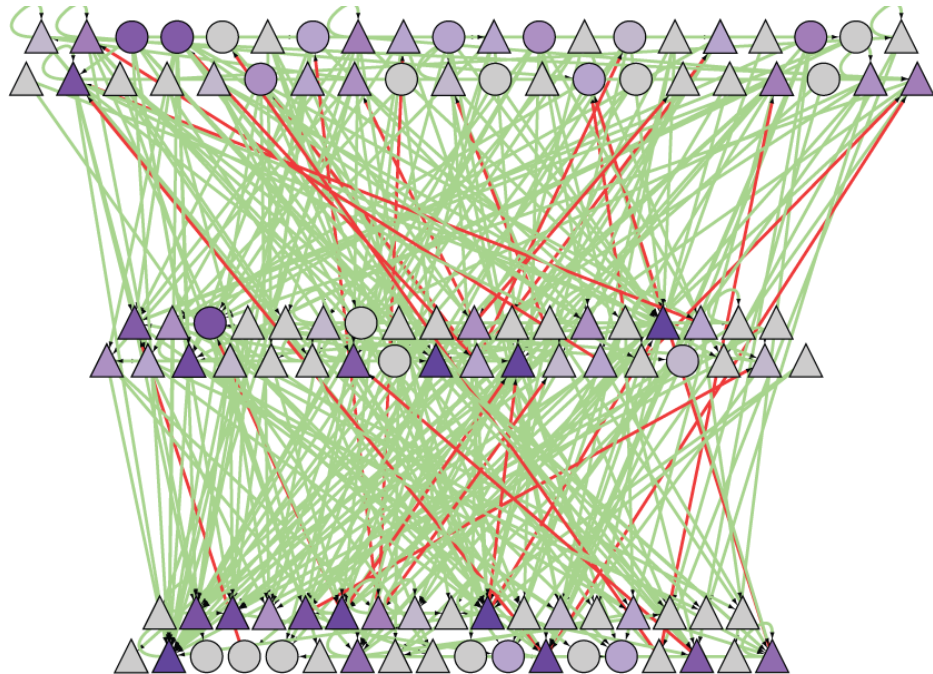


From '00 to ~' 20,
cost of DNA sequencing expt. shifts from
the actual seq. to sample
collection & analysis

Comparison w/ photography



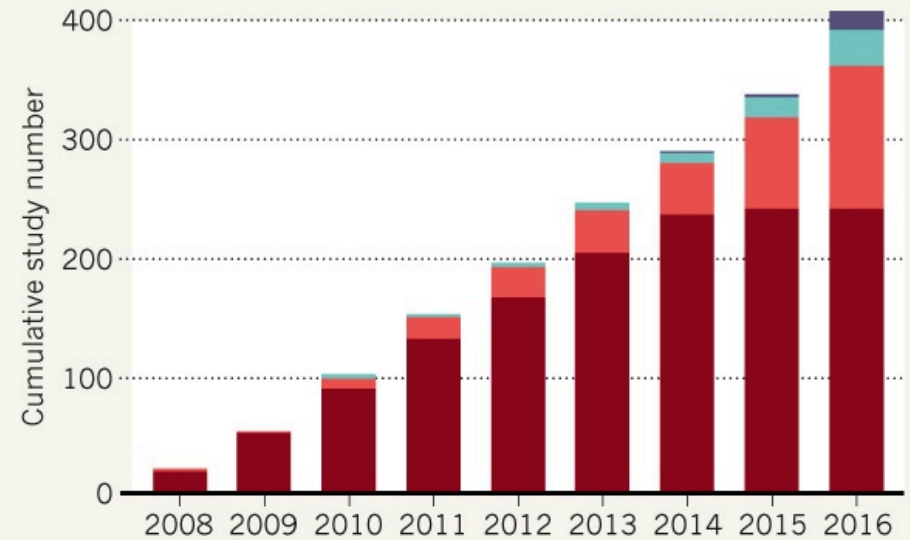
**A Success of
Scale & Integration:
Many GWAS
variants found,
most not in genes,
but affecting
regulatory network**



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.

Sample sizes: ■ More than 200,000 ■ 100,000–199,999
■ 50,000–99,999 ■ 10,000–49,999



©nature

- 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

EX of 'omics research on focused patient cohorts: Yale Research in Human Neurogenomics

- **Representative Nat'l Initiatives:**

CMG, BrainSpan,
psychENCODE, BSMN,
SFARI, MVP/PTSD, PGC,
NIDA Neuroproteomics

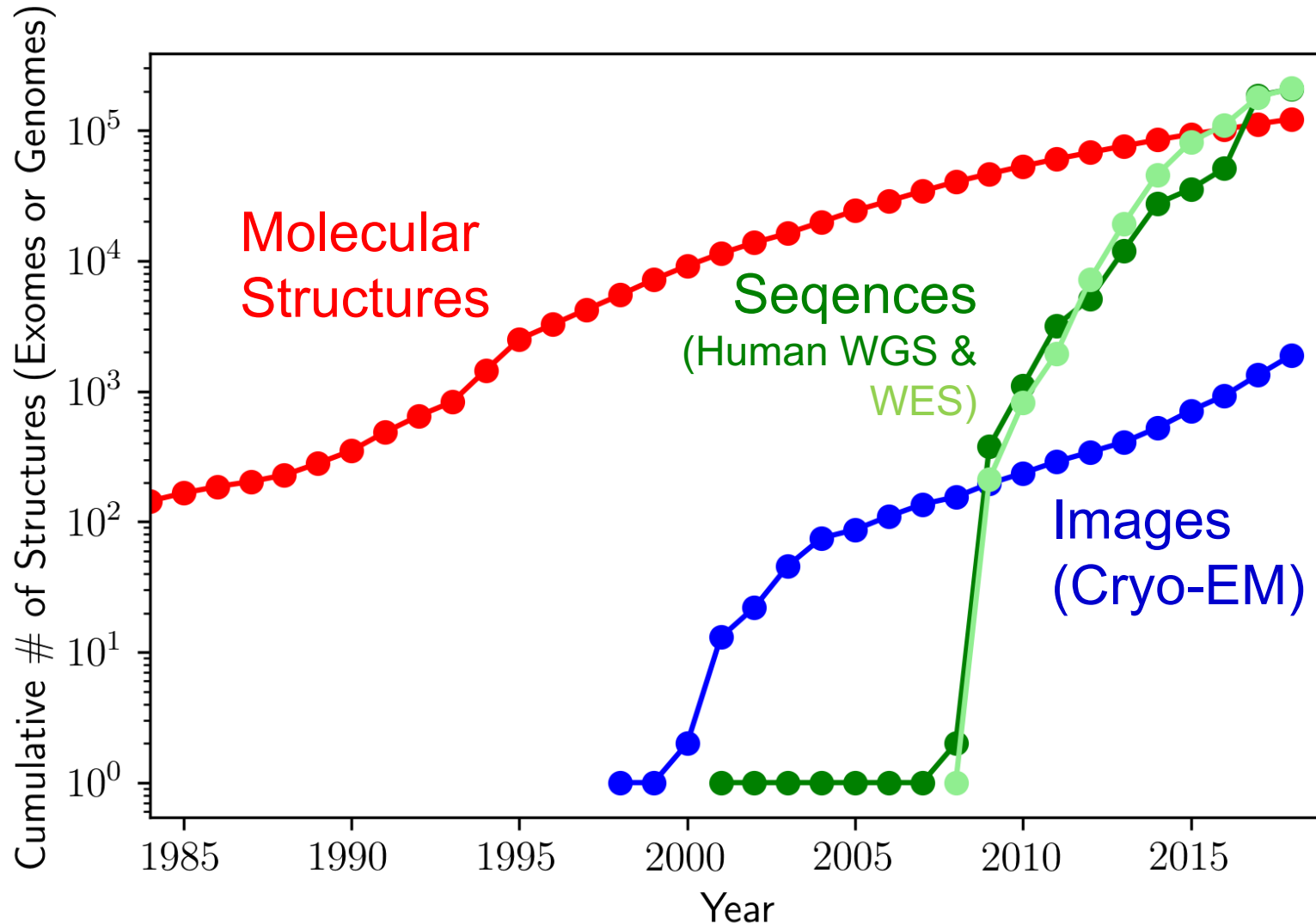


- **Yale investigators involved in these Initiatives:**

M Gunel, S Mane, N Sestan, F Vaccarino, J Noonan, P Rakic,
J Gelernter, A Nairn, K Williams, M Levine, R Polimanti

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

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



Center for Biomedical Data Science

Yale SCHOOL OF MEDICINE



Yale School of Medicine
Chairs' Lunch
November 4, 2016

Carolyn
Slayman



Center for Biomedical Data Science

