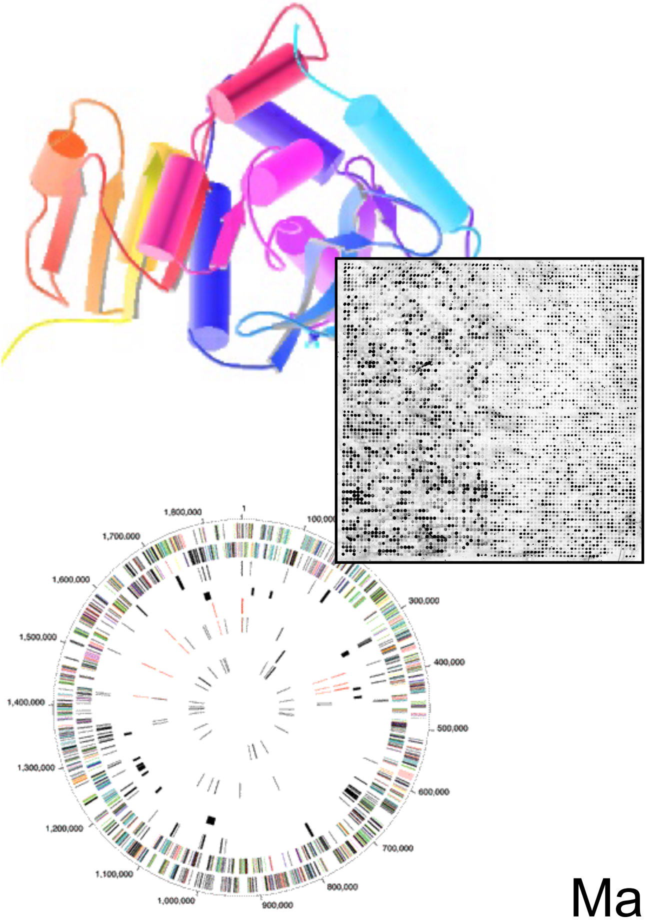
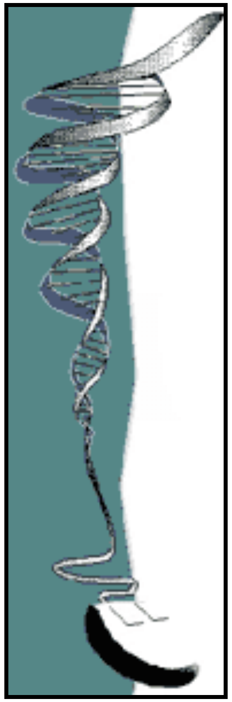
Biomedical Data Science: Introduction



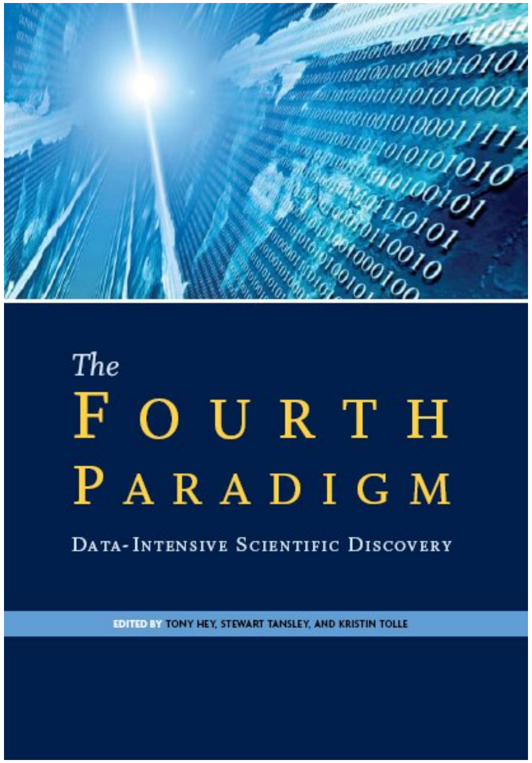


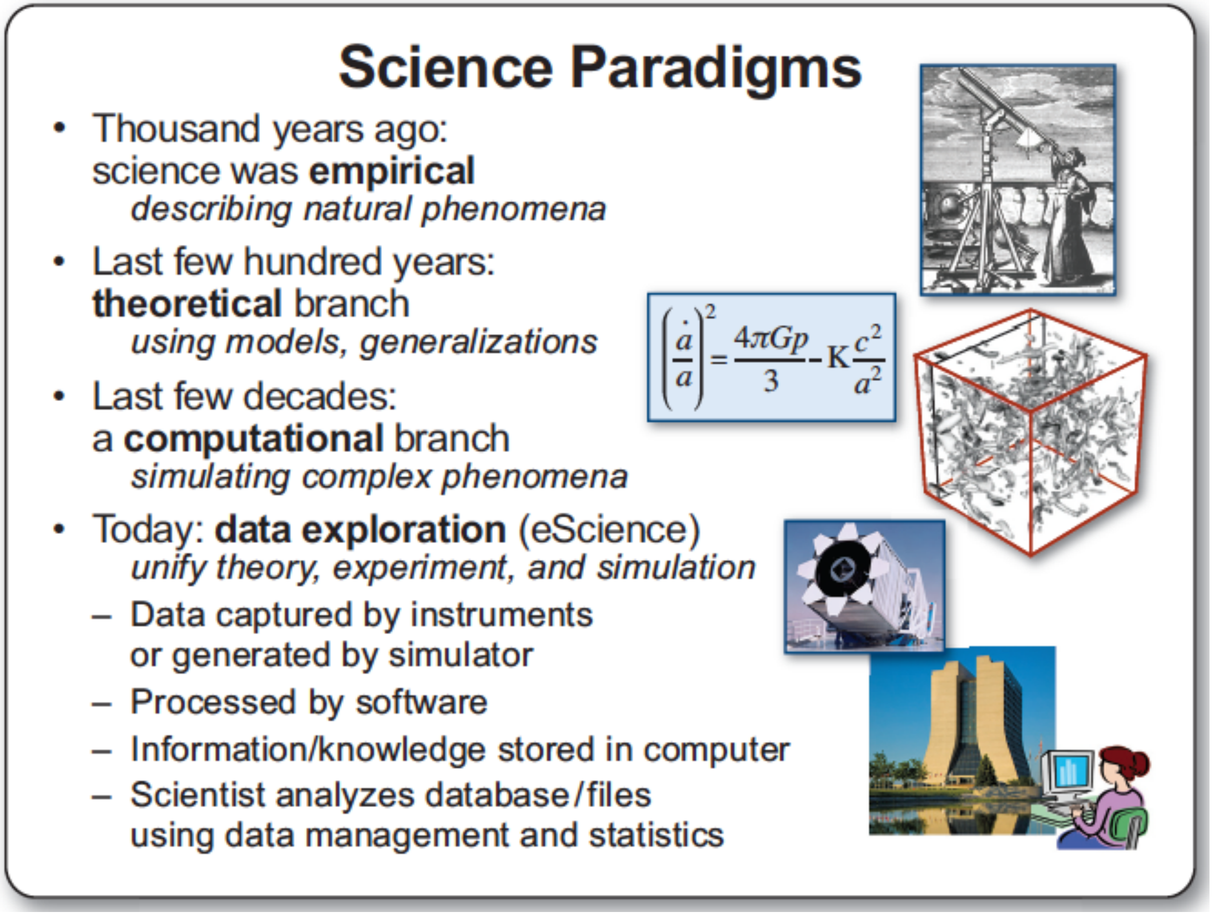
Mark Gerstein, Yale University
GersteinLab.org/courses/452
(last edit in spring ’20)

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
- 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

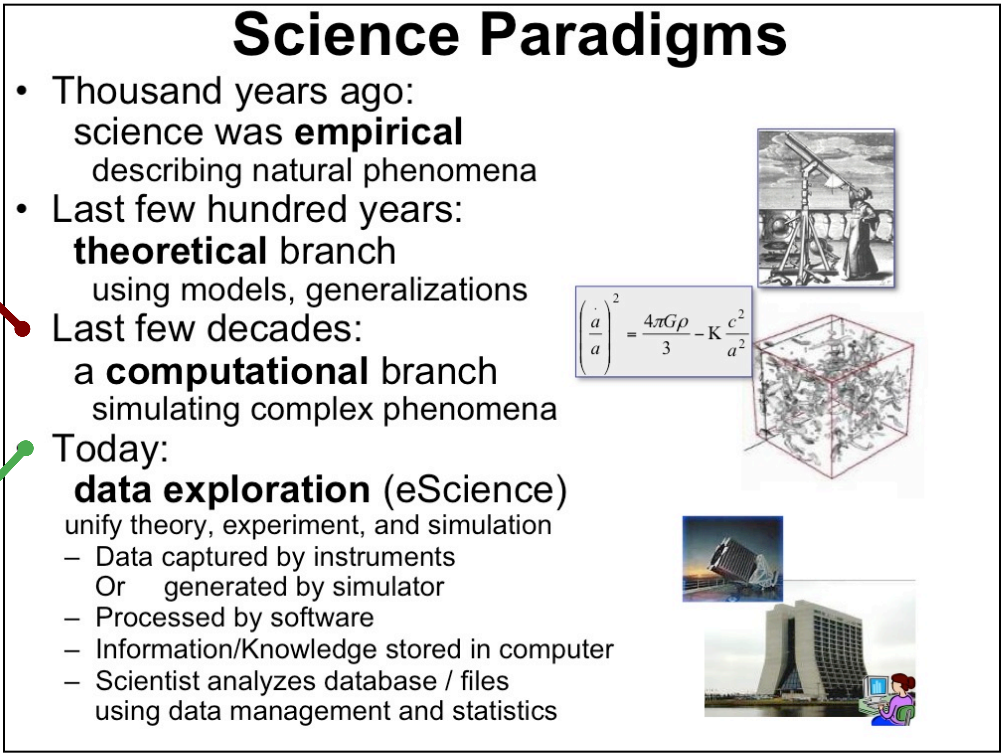
$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi Gp}{3} - K\frac{c^2}{a^2}$$

Jim Gray's 4th Paradigm

#3 - Simulation

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

Emphasis on: Supercomputers



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…

CIO Network
INSIGHTS AND IDEAS FOR TECHNOLOGY LEADERS.

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

Why Big Data Is All Retailers Want for Christmas

Eric Savitz, Forbes Staff

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

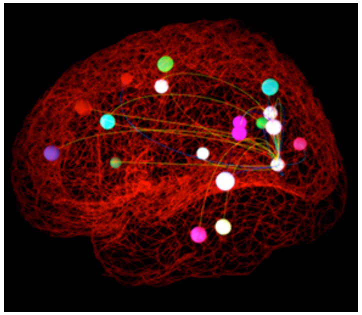
When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business ne
up. The company had just under 8 million accounts, and the number was growing qu
friends and colleagues to join. But users weren't seeking out connections with the pe
rate executives had expected. Something was apparently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

[Oct. '12 issue]

Data Science in Traditional Science

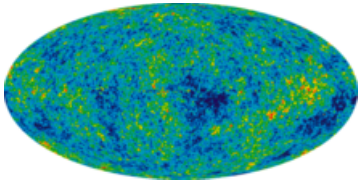
- Pre-dated commercial mining
- Instrument generated
- Large data sets often created by large teams not to answer one Q but to be mined broadly
- Often coupled to a physical/biological model
- Interplay w/ experiments

High energy physics -
Large Hadron Collider



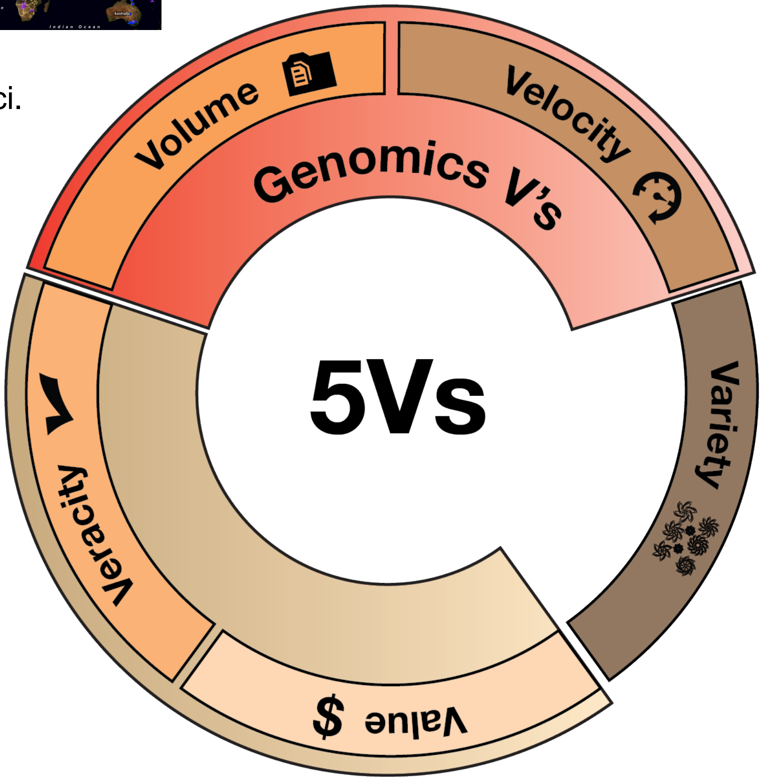
Neuroscience -
The Human
Connectome Project

Ecology
& Earth Sci.
- Fluxnet



Astronomy -
Sloan Digital Sky survey

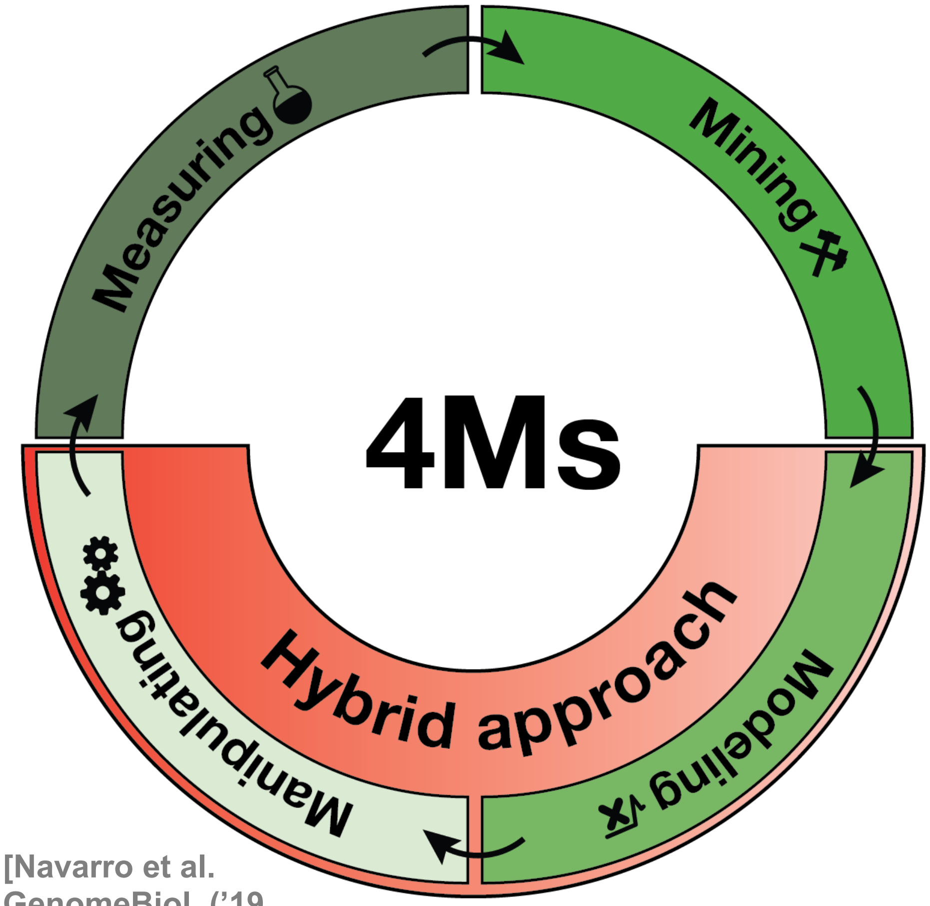
Genomics
DNA
sequencer



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

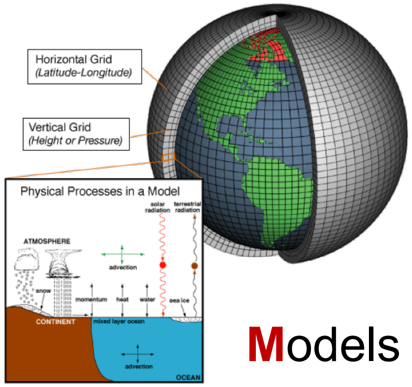
Coupling of Scientific Data to Models & Experiments

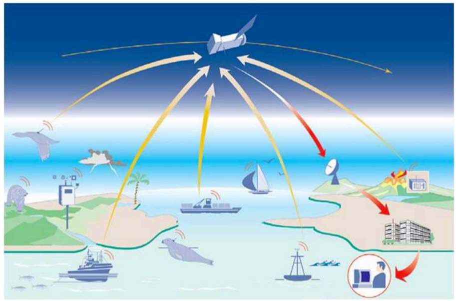
- Scientific data often coupled to a physical/biological model
- Lauffenburger’s Sys. Biol. **4Ms:**
  Measurement, Mining, Modeling & Manipulation
  (Ideker et al.’06. Annals of Biomed. Eng.)
- Weather forecasting as an exemplar
  - Physical models & simulation useful but not sufficient (“butterfly” 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

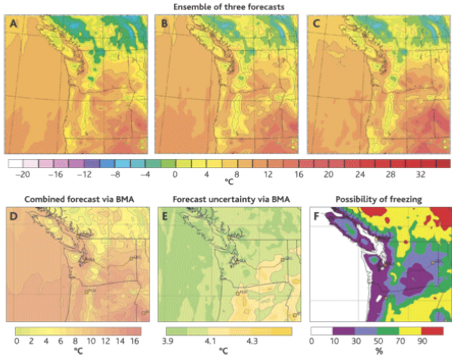




Models + Data **Mining**

↓

Forecasts



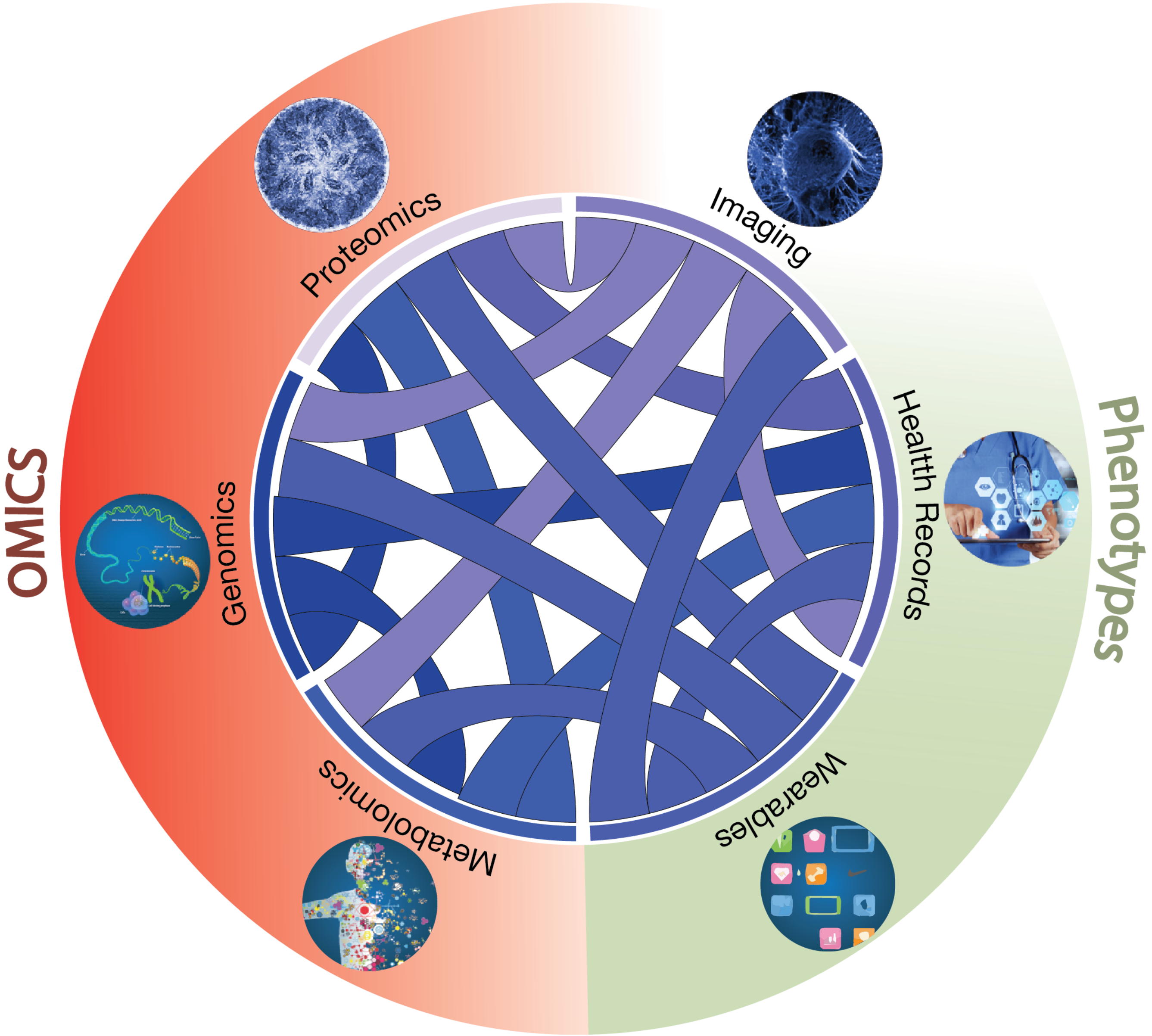
NOAA

Biomed. Data science:

Scaling & Integration

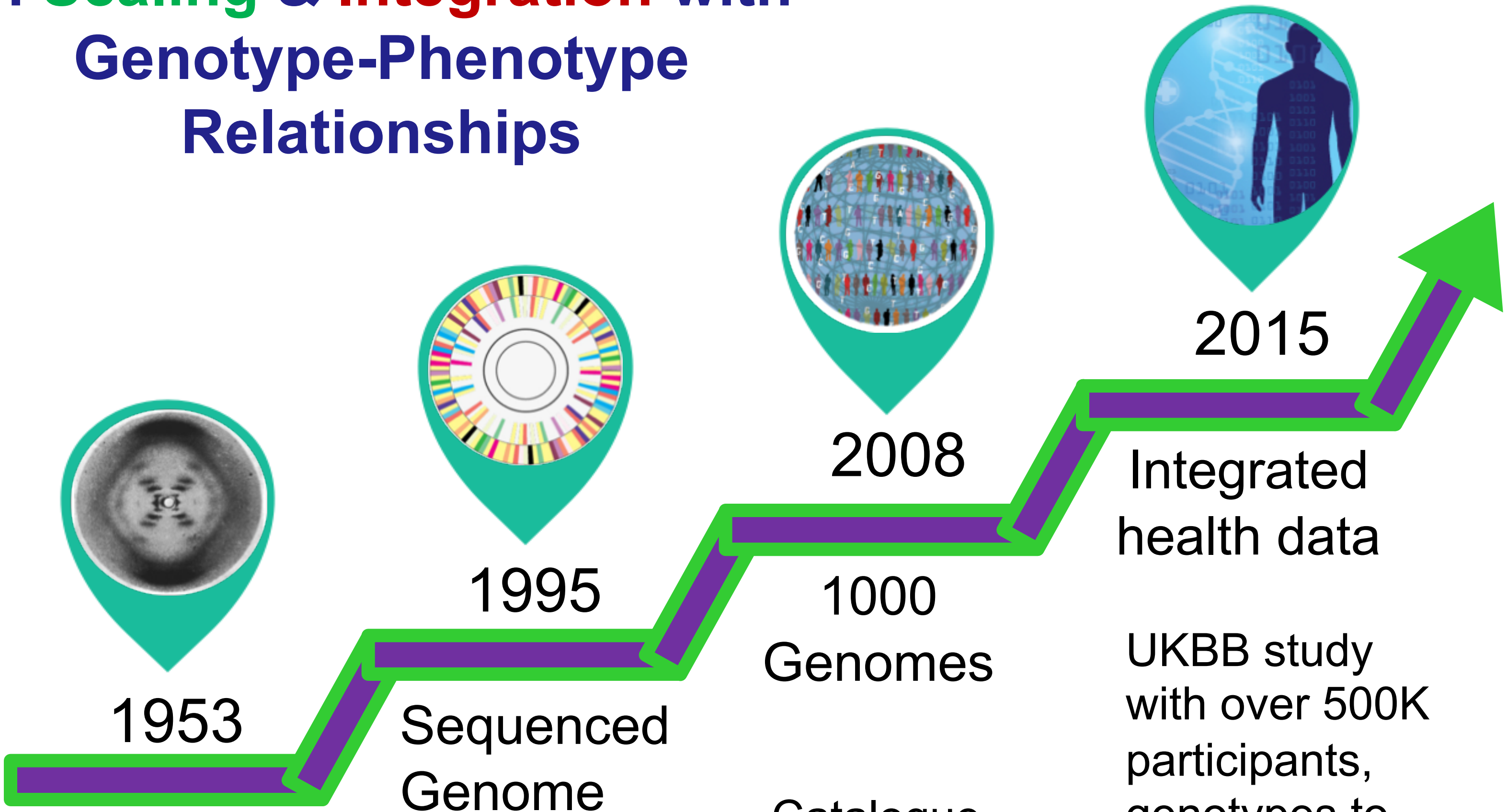
Drivers of Biomedical Data Science

- **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

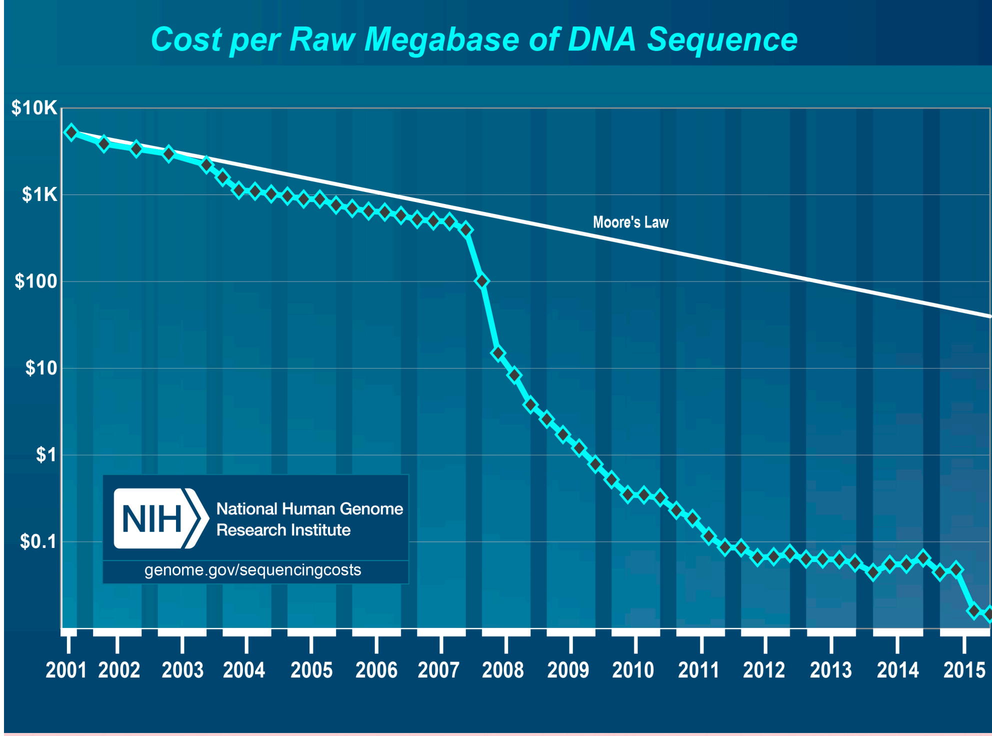
Integrated health data

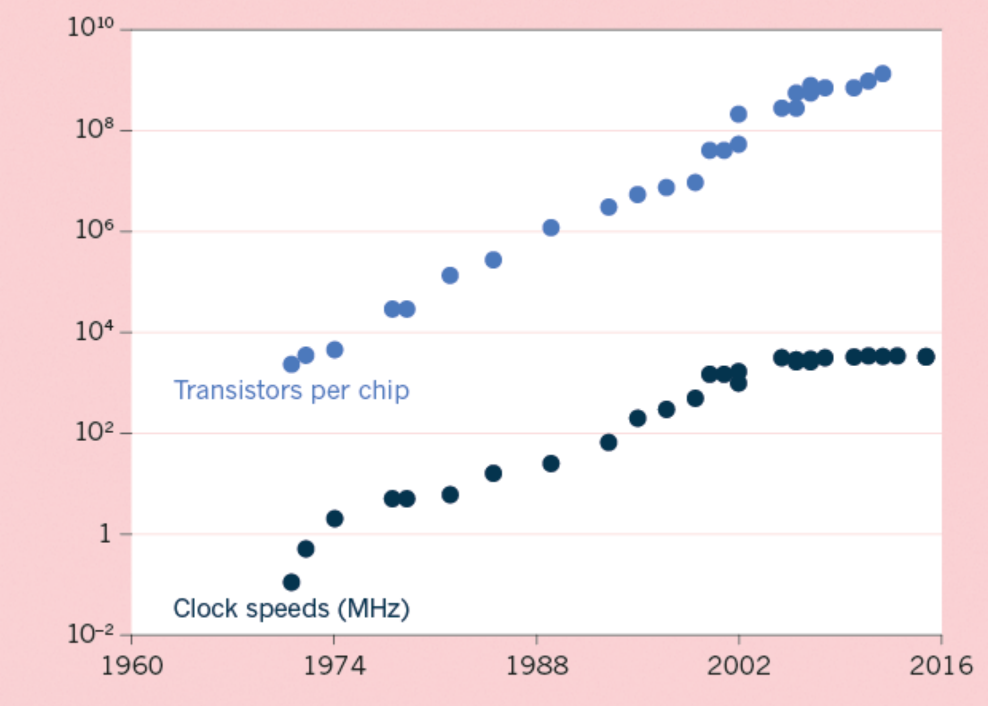
UKBB study with over 500K participants, genotypes to phenotypic details & clinical information

The Scaling of Genomic Data Science:

Powered by exponential increases in data & computing

(Moore's Law)

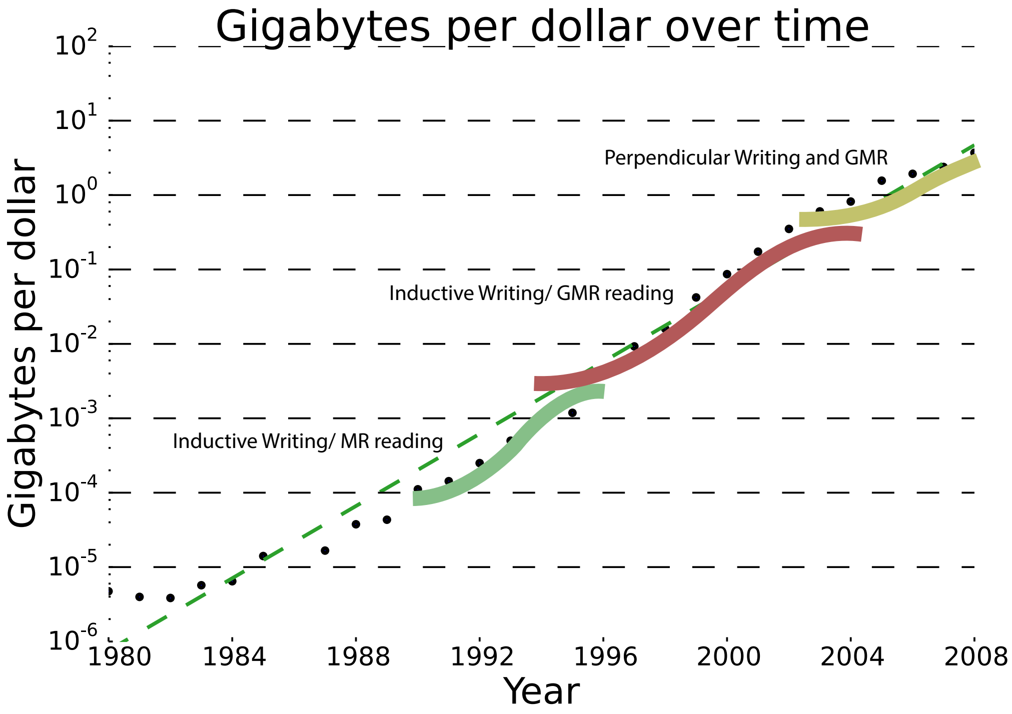


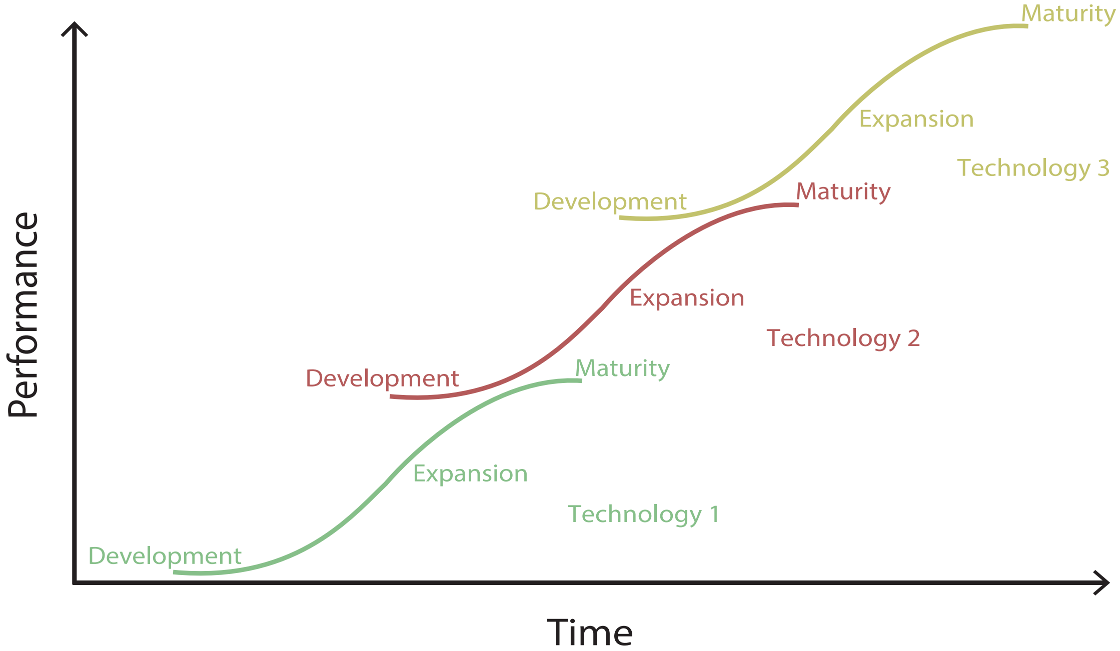


[NHGRI website + Waldrop ('15) Nature]

Kryder's Law and S-curves underlying exponential growth

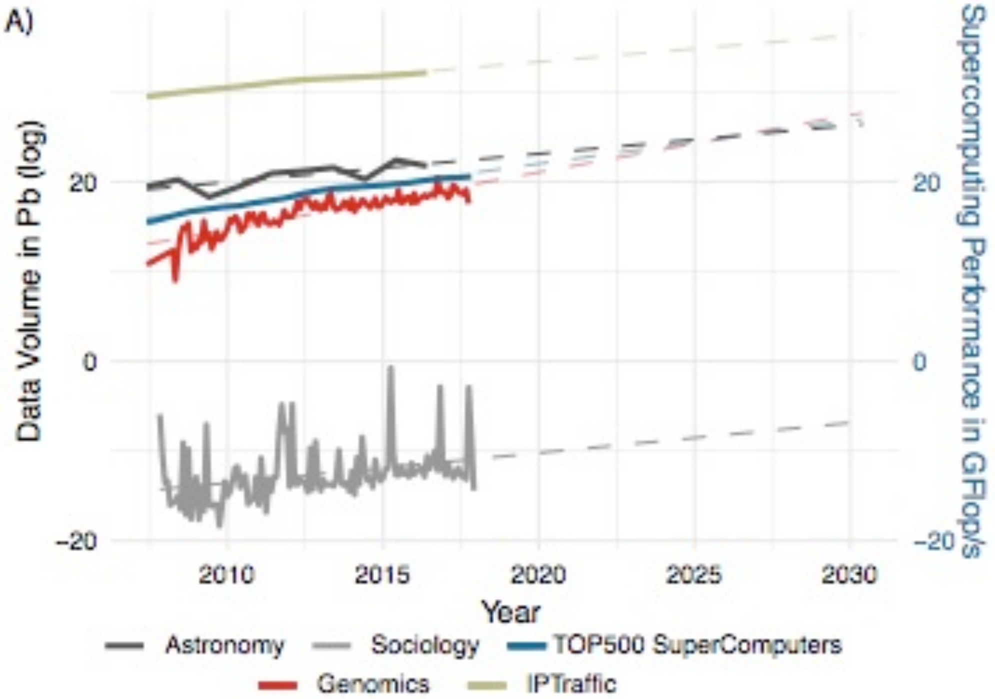
- Moore's & Kryder's Laws
  - As important as the increase in computer speed has been, the ability to store large amounts of information on computers is even more crucial
- Exponential increase seen in Kryder's law is a superposition of S-curves for different technologies

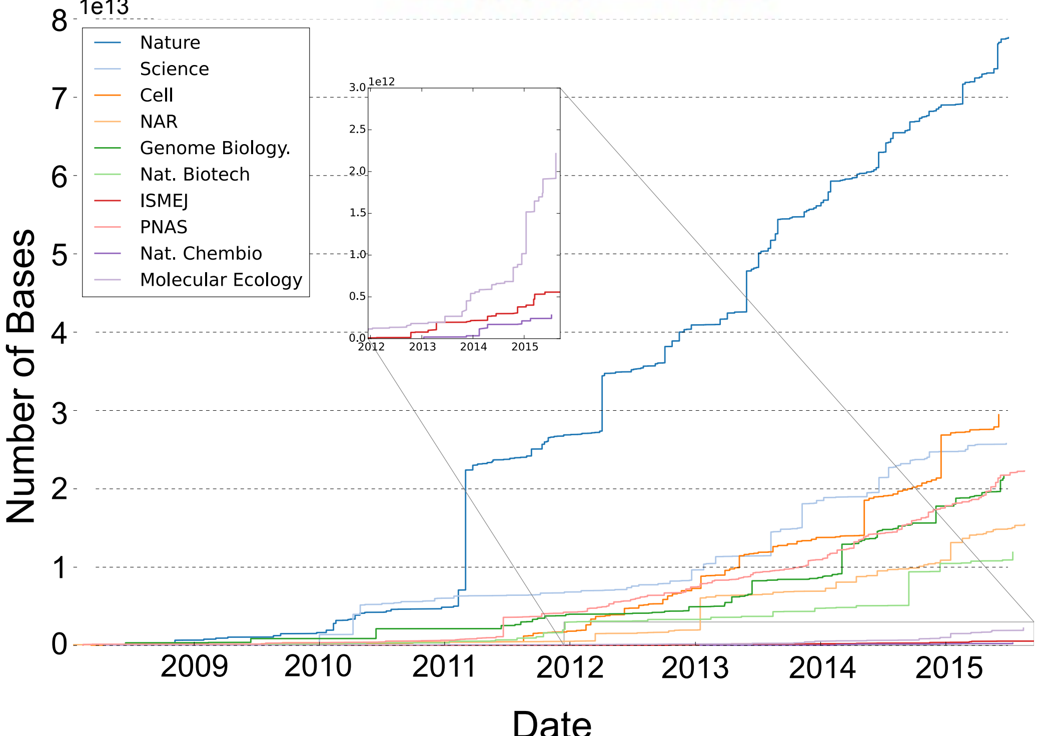

Gigabytes per dollar over time



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

Sequencing cost reductions have resulted in an explosion of data

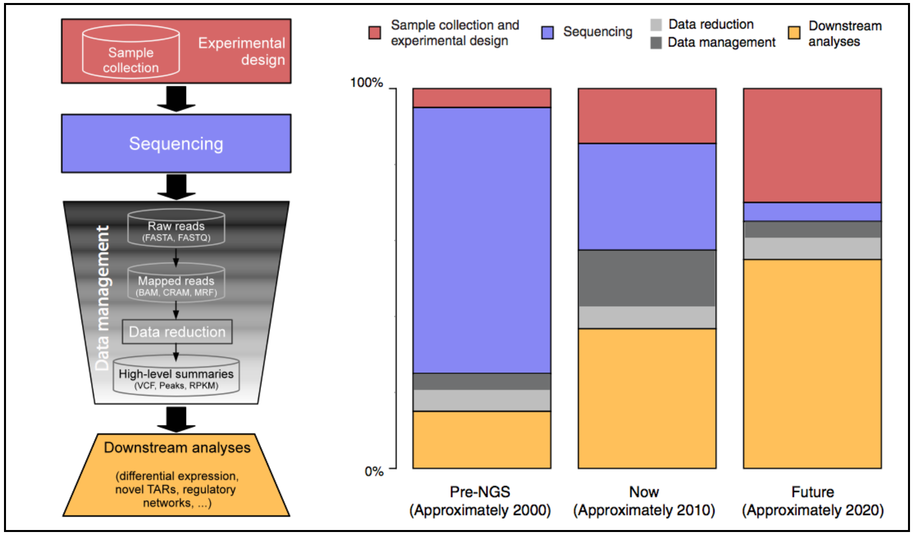




- The type of sequence data deposited has changed as well.

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

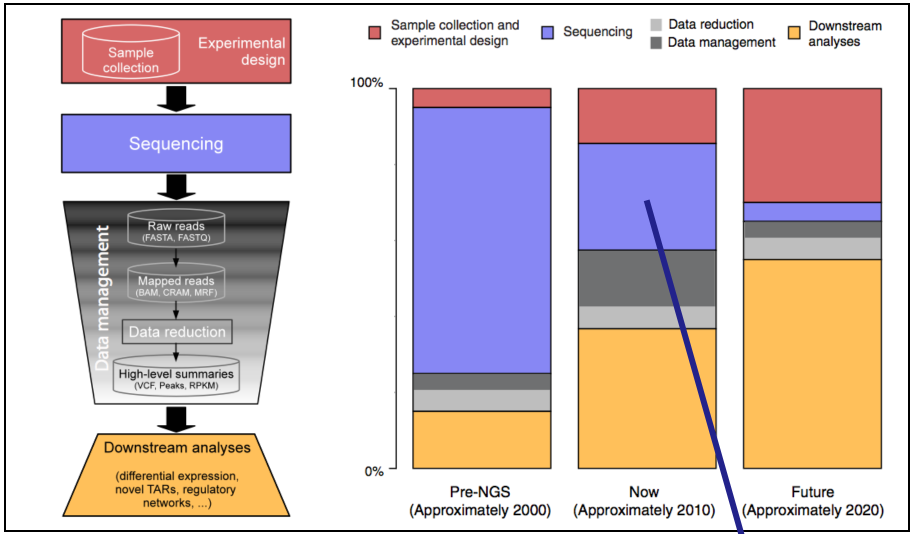
The changing costs of a sequencing pipeline



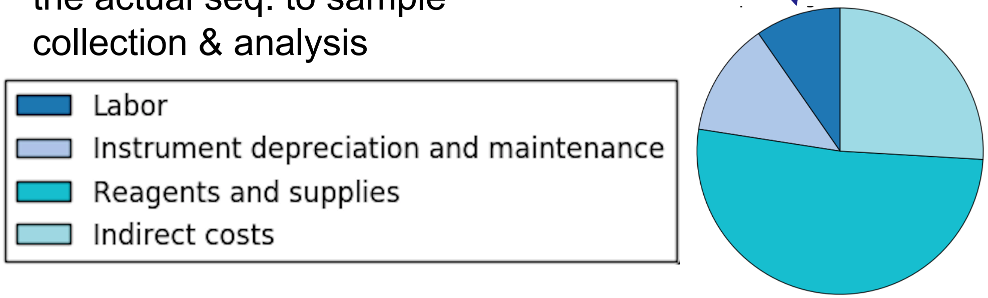
From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from
the actual seq. to sample
collection & analysis

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

The changing costs of a sequencing pipeline

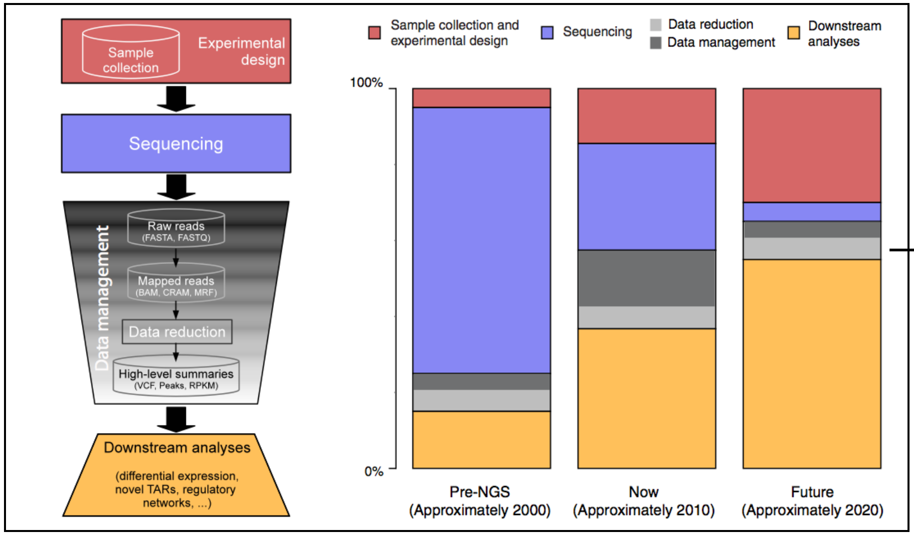


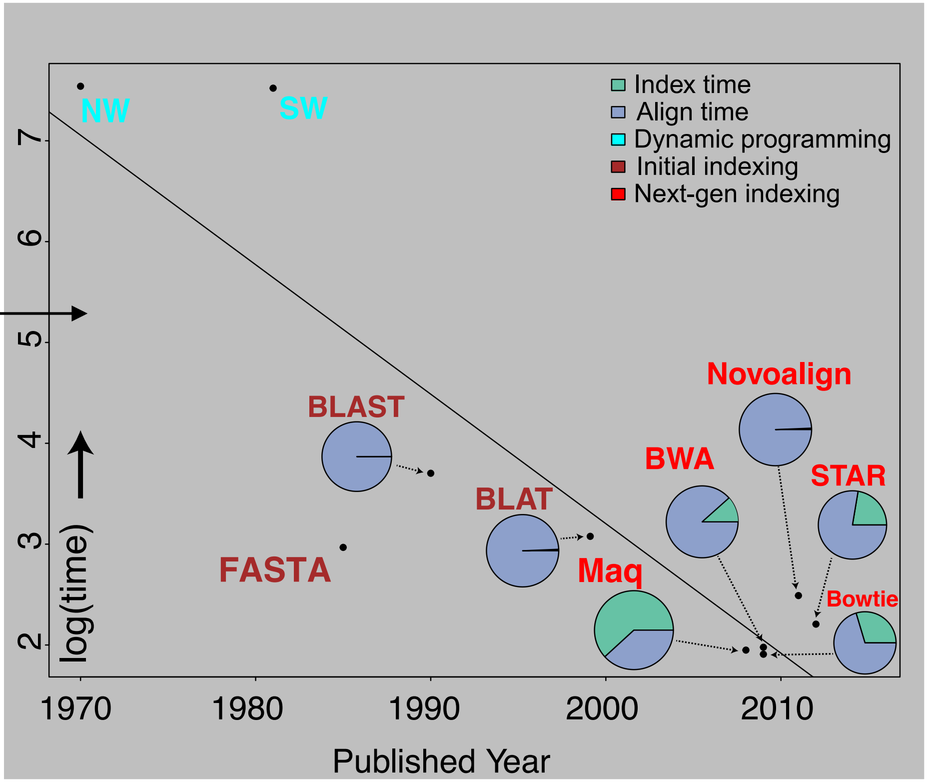
From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis



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

The changing costs of a sequencing pipeline



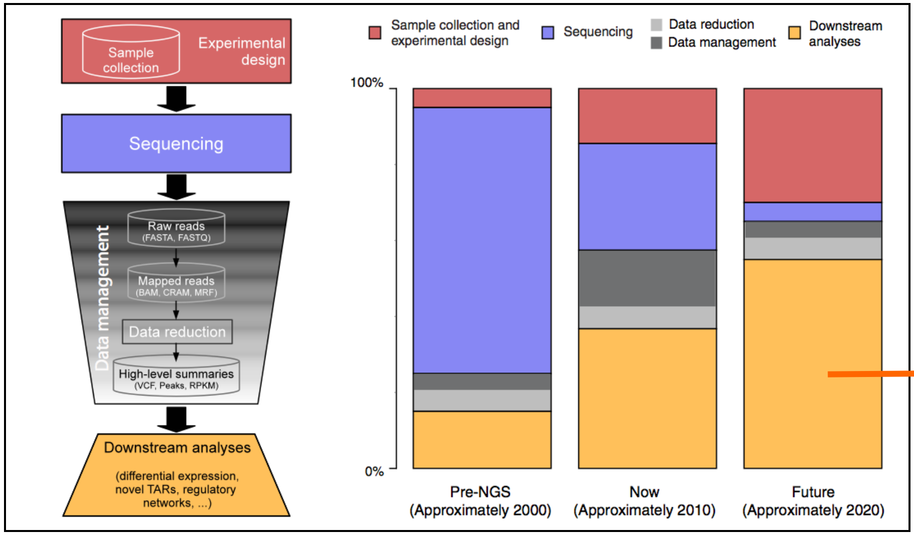


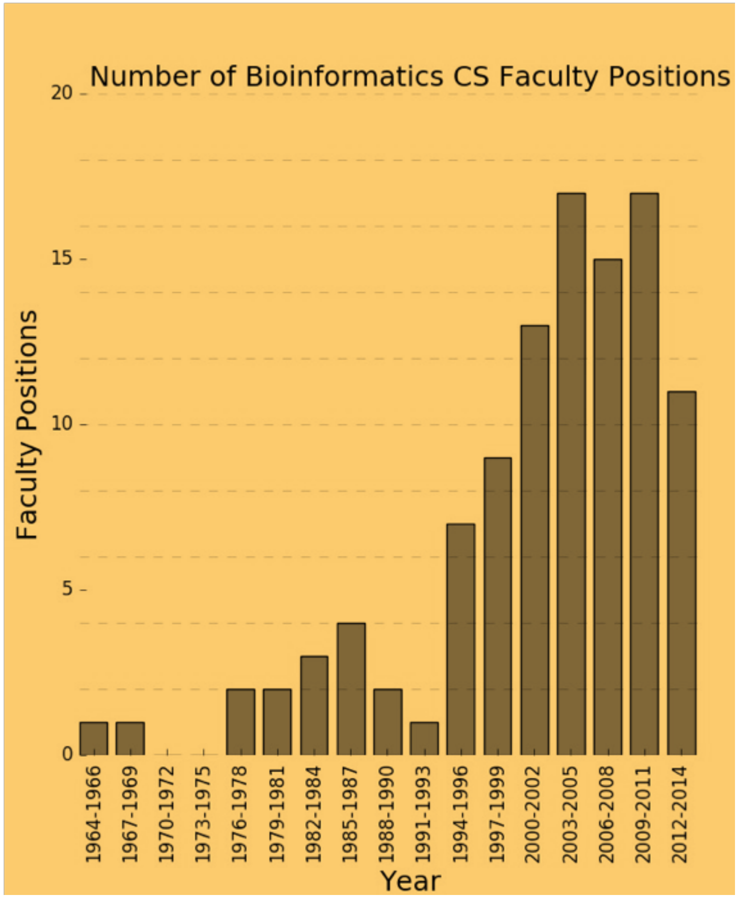
From '00 to ~'20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

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

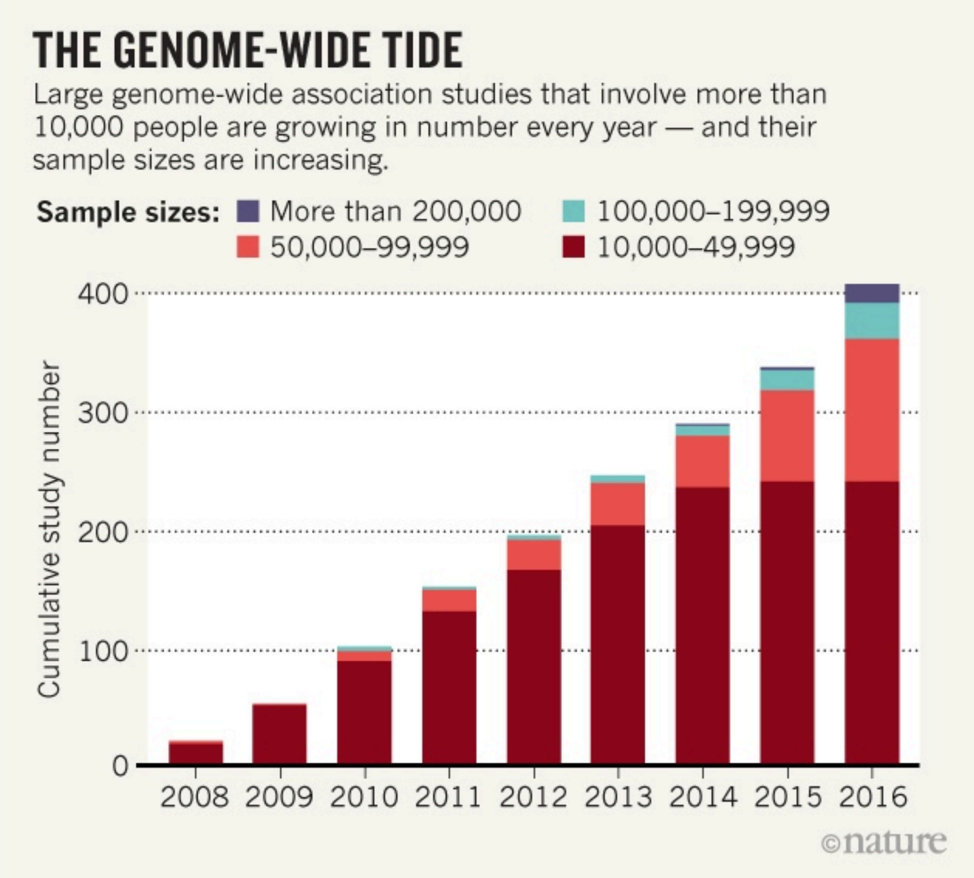




From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

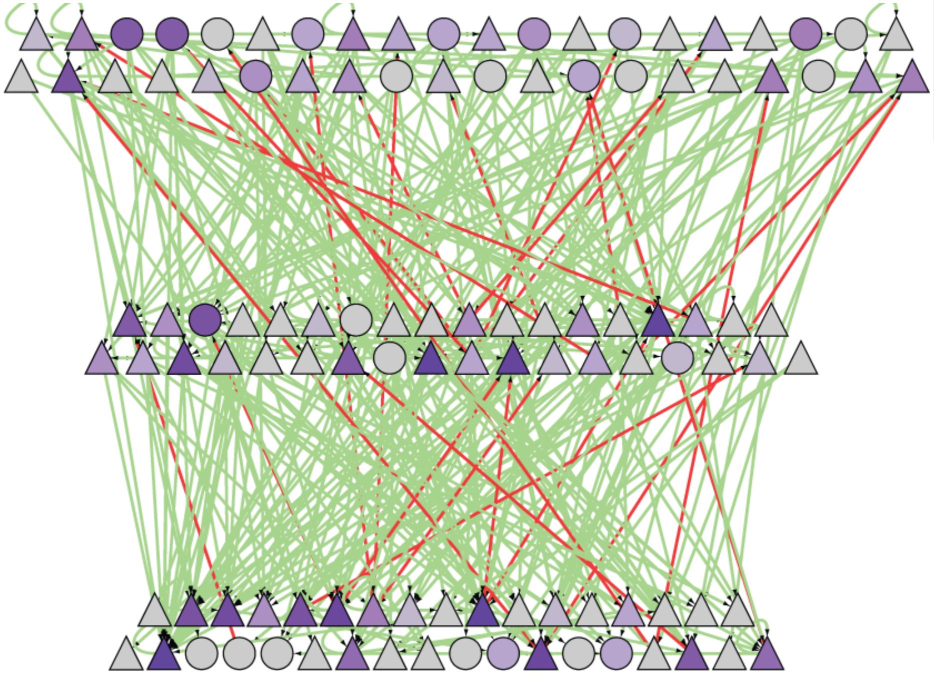
[Sboner et al. (‘11), Muir et al. (‘15) Genome Biology]

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



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.



- 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

AGTC

biobank uk

MVP

TCGA

Genomics england

PCAWG
PanCancer Analysis of WHOLE GENOMES

THE PRECISION MEDICINE INITIATIVE

STARTUPS

color

23andMe

Veritas
The Genome Company

FOUNDATION MEDICINE®

- 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

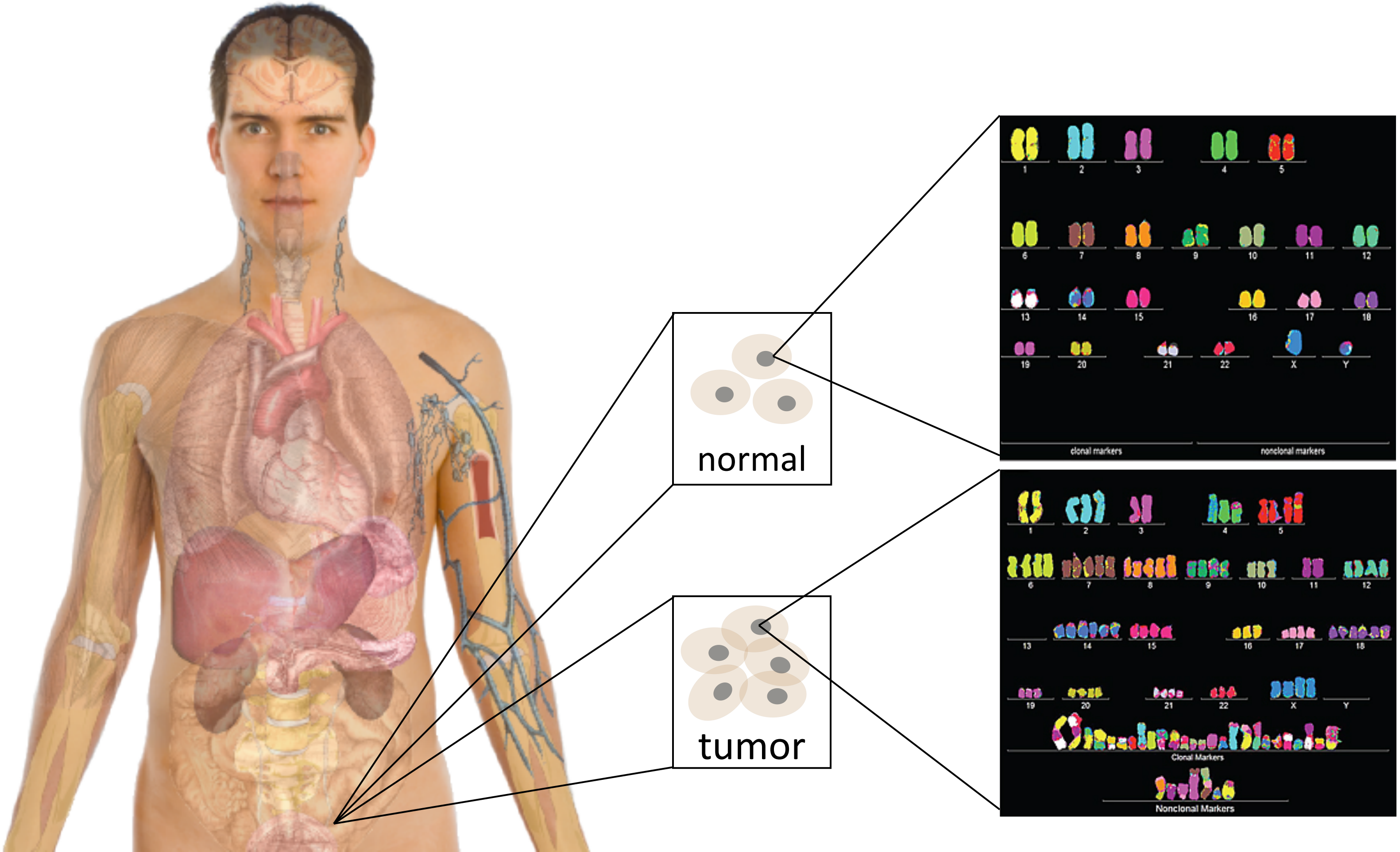
Medical Big Data: Promise and Challenges (Lee and Yoon , *Kidney Res. Clin. Pract.*, 2017)

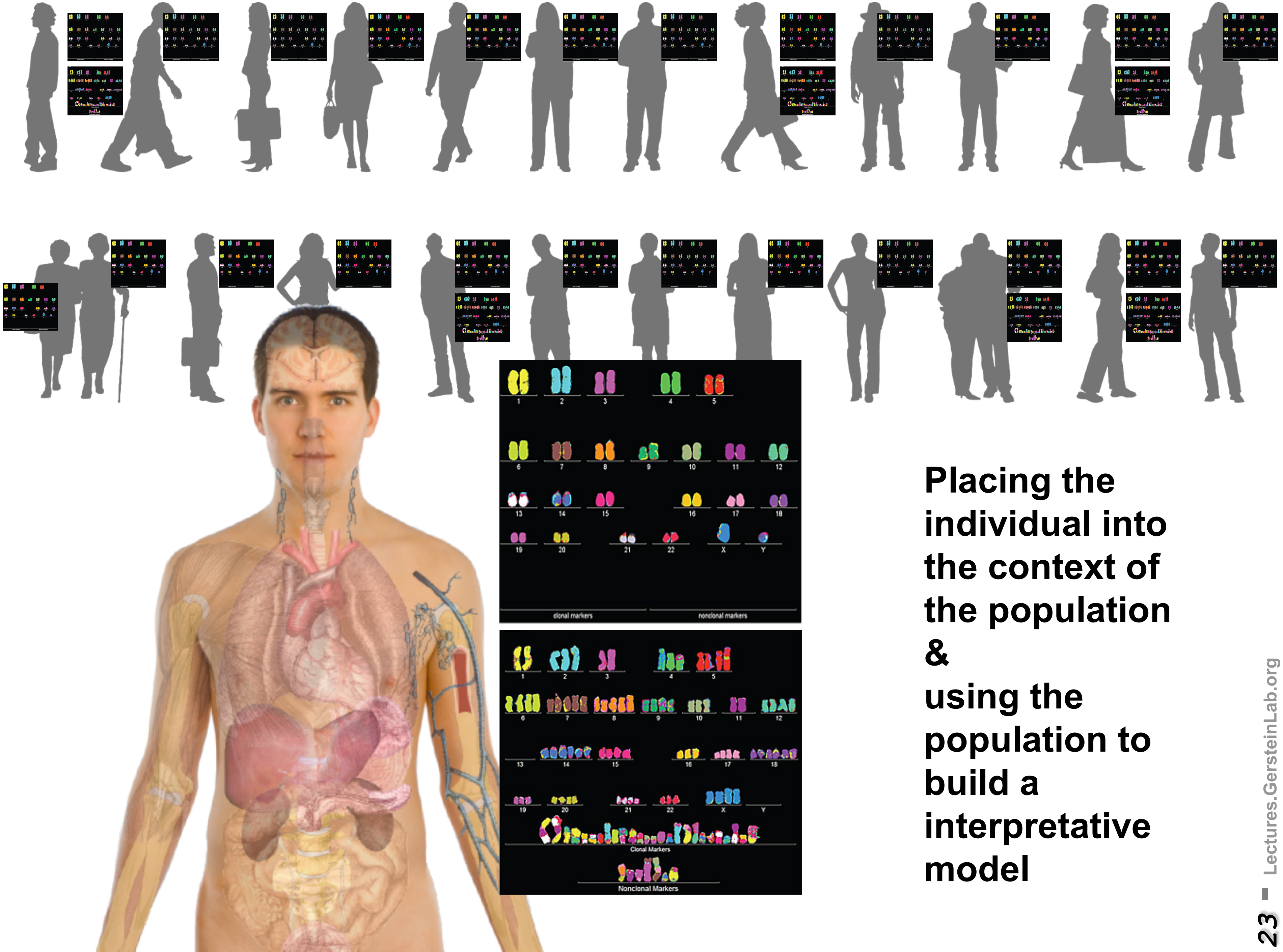
Biomed. Data science:

The Future

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

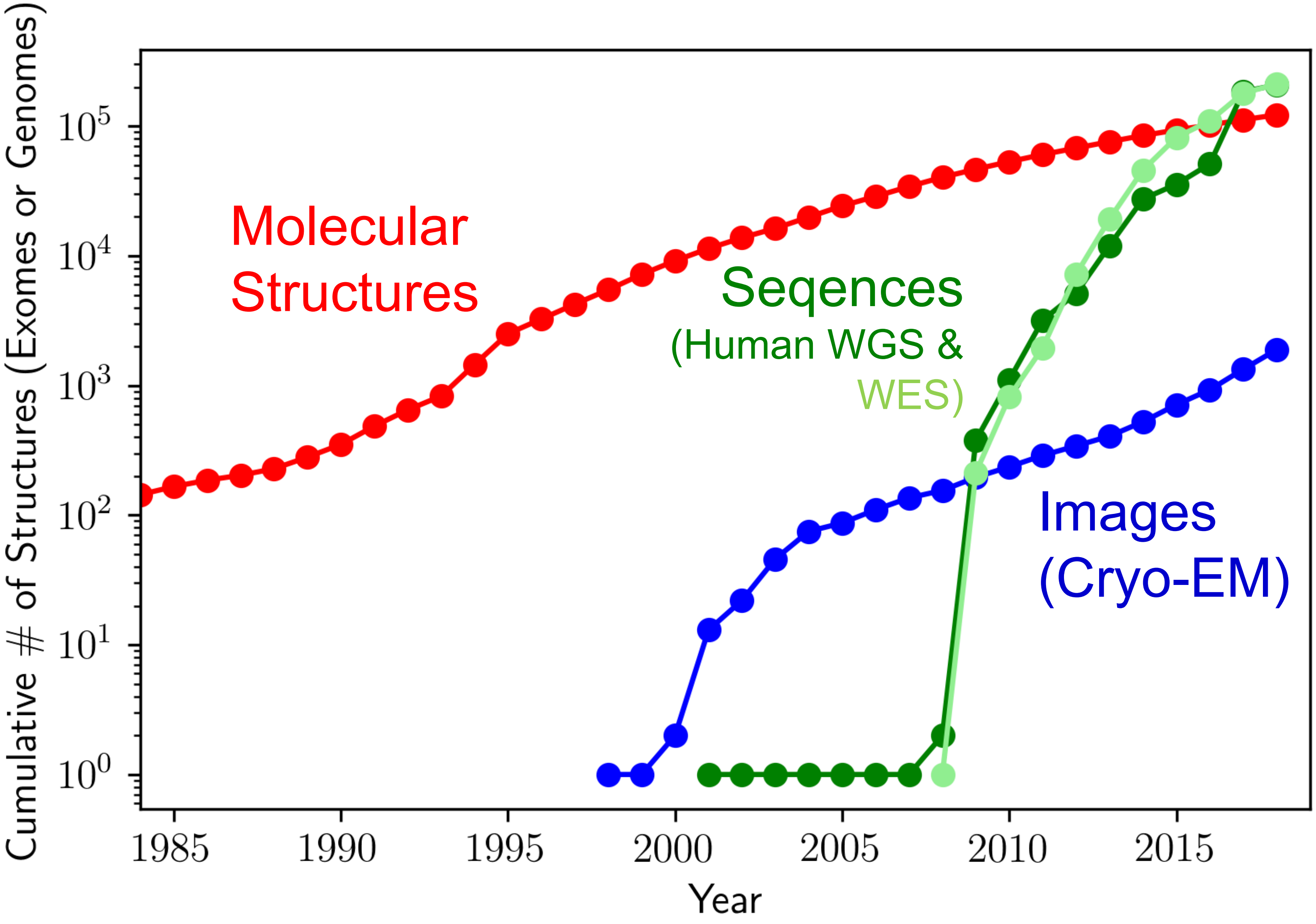
Personal genomes soon will become a commonplace part of medical research & eventually treatment (esp. for cancer). They will provide a primary connection for biological science to the general public.





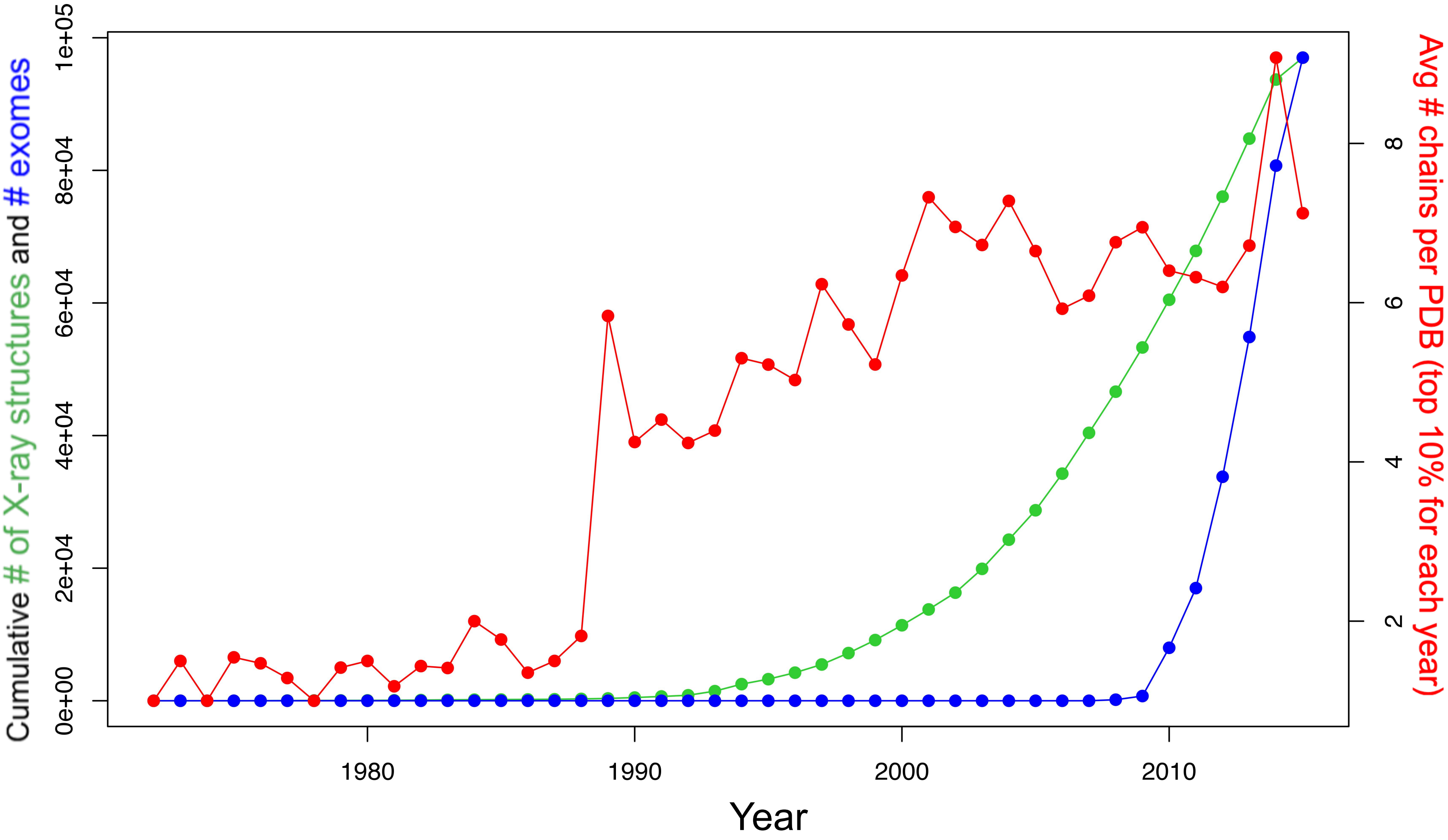
Placing the individual into the context of the population
&
using the population to build a interpretative model

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

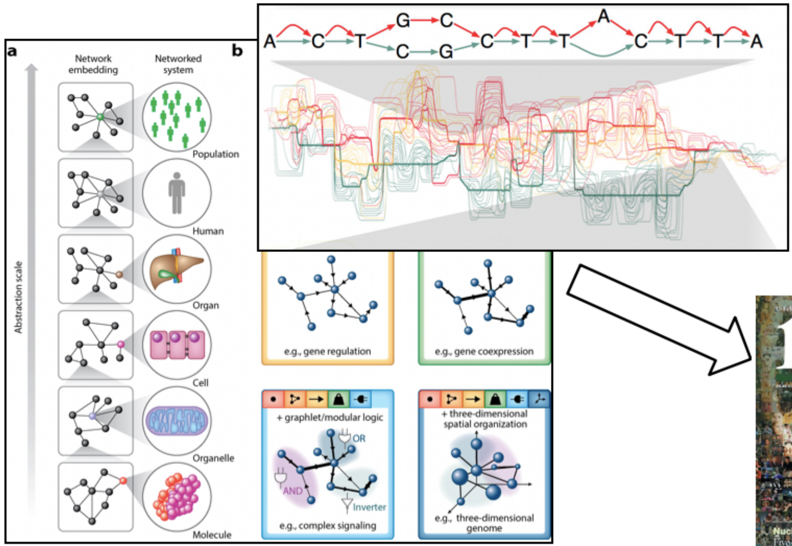
PDB: Berman HM, et al. NAR. (2000)
CATH: Sillitoe I, et al. NAR. (2015)
SCOP: Fox NK et al. NAR. (2014)

[Sethi et al. COSB ('15)]

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

Technical Imports

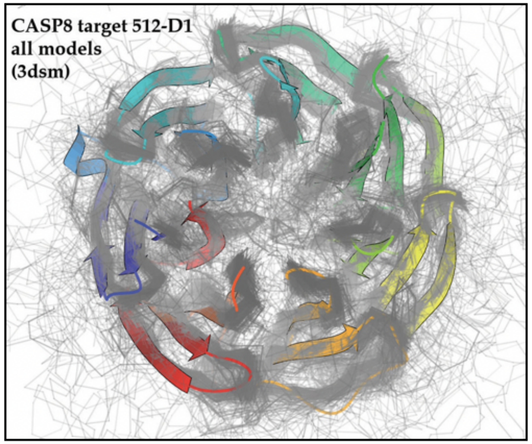
Networks and graphs



Importing tech. developed in other big data disciplines

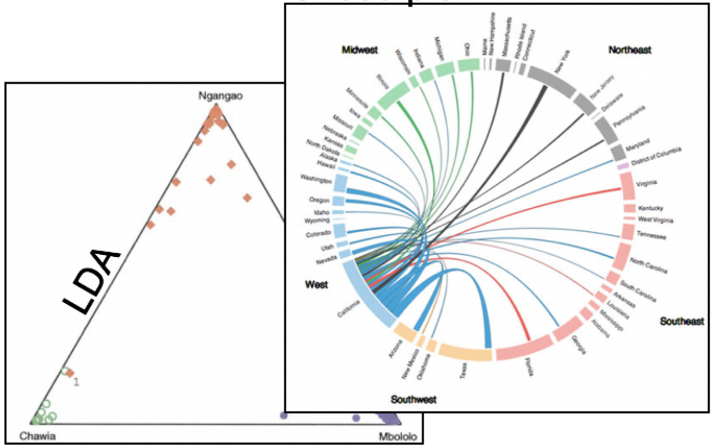
Cultural Imports

CASP

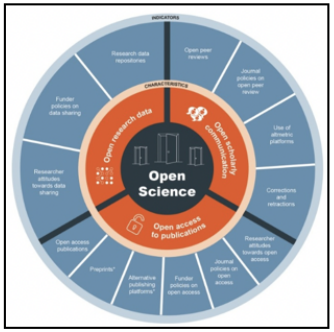


Technical Exports

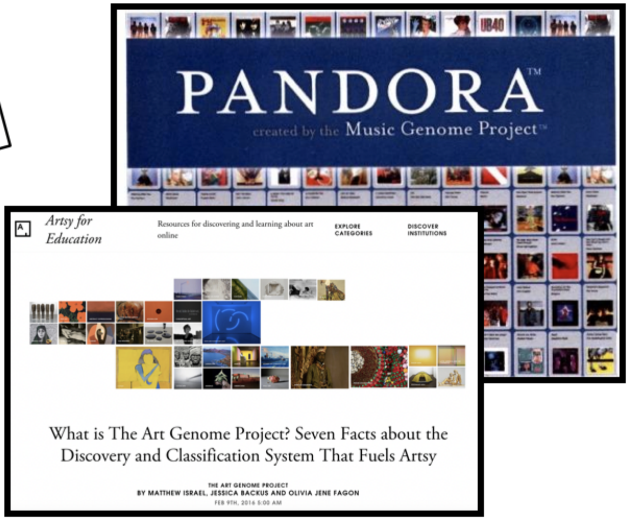
Circos plot



Open Science



Cultural Exports



Artsy for Education

Resources for discovering and learning about art online

EXPLORE CATEGORIES

DISCOVER INSTITUTIONS

What is The Art Genome Project? Seven Facts about the Discovery and Classification System That Fuels Artsy

THE ART GENOME PROJECT
BY MATTHEW ISRAEL, JESSICA BACKUS AND OLIVIA JENE FAGON
FEB 9TH, 2016 5:00 AM

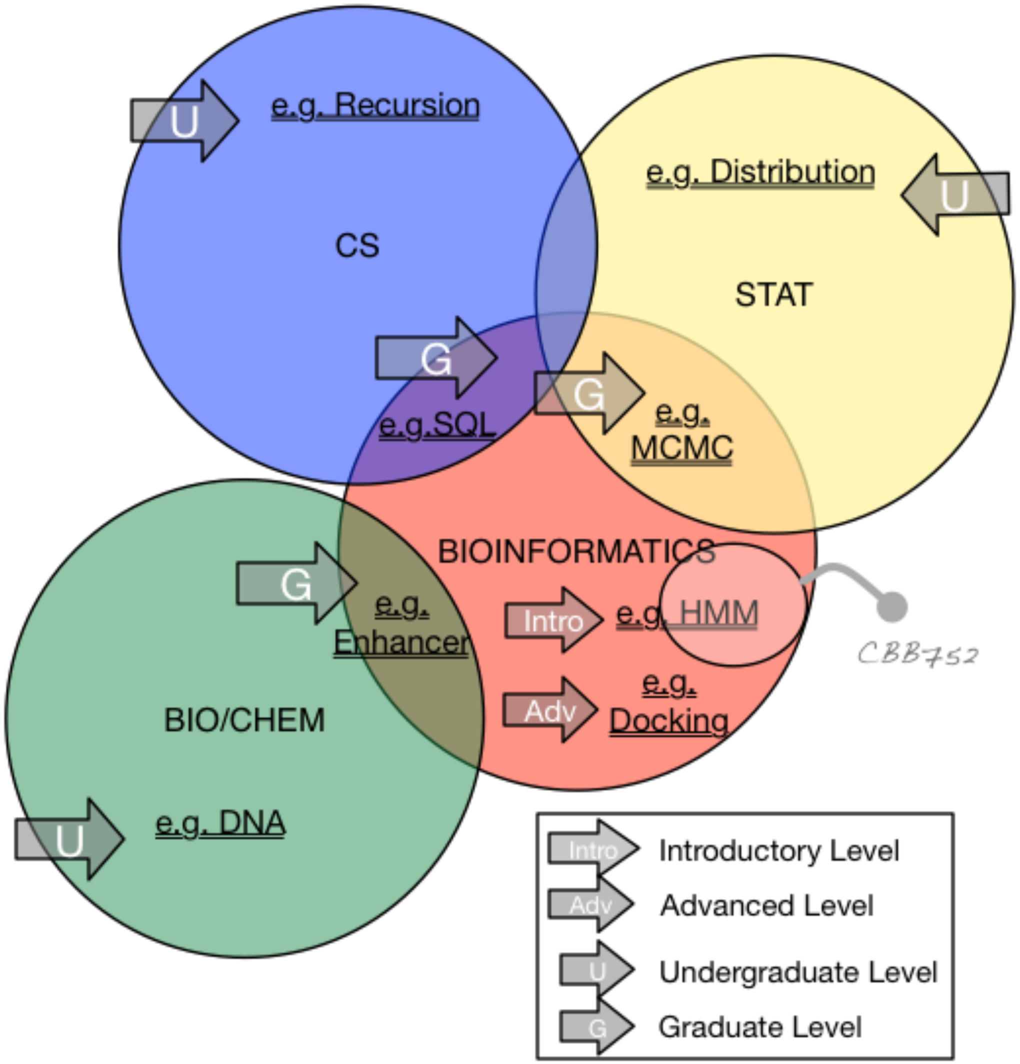
[Navarro et al. GenomeBiol. ('19, in press)]

Biomed. Data science:

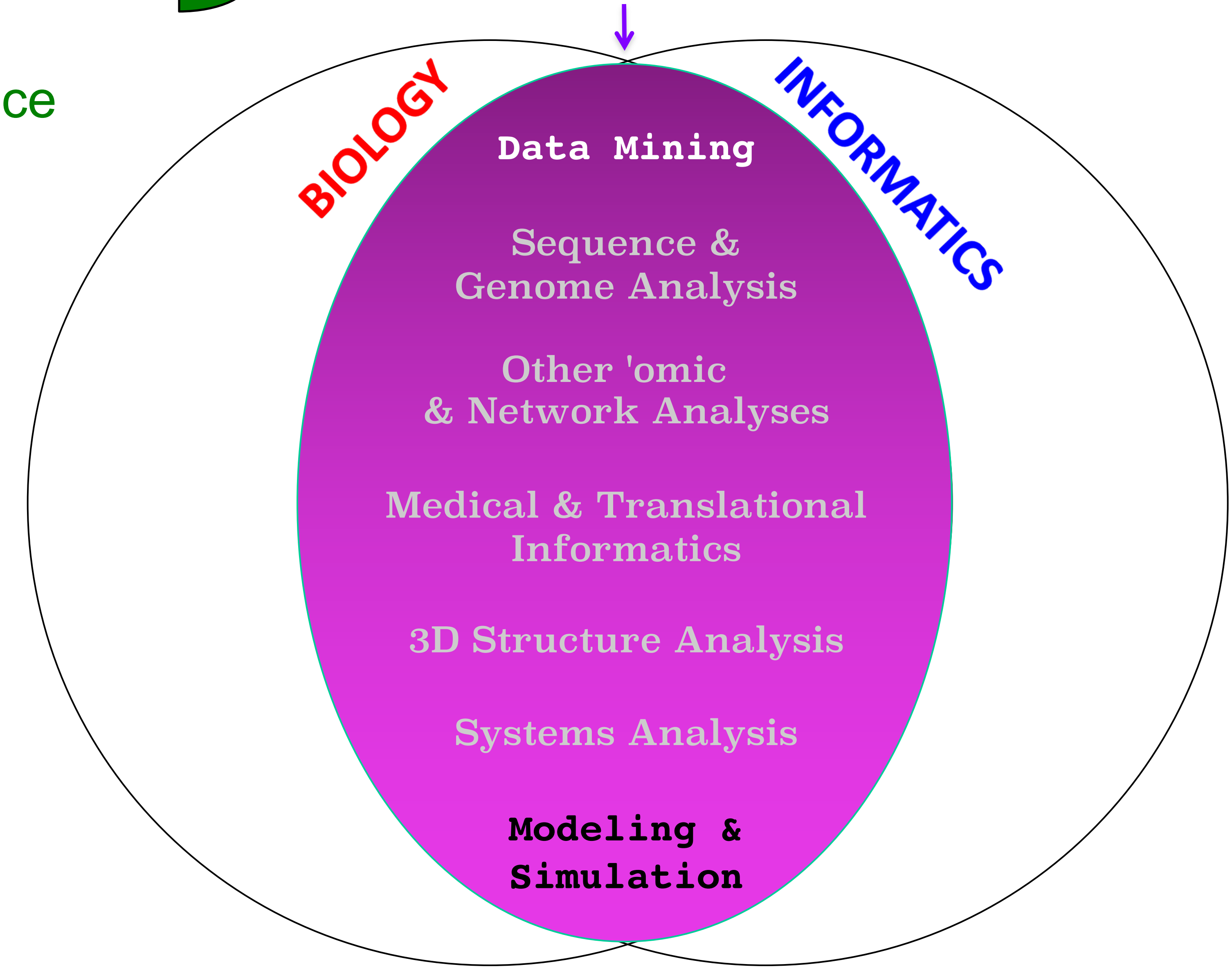
The Course

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



BIOLOGY

INFORMATICS

Data Mining

Sequence &
Genome Analysis

Other 'omic
& Network Analyses

Medical & Translational
Informatics

3D Structure Analysis

Systems Analysis

**Modeling &
Simulation**

[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

Assignment #0 Page

http://bit.ly/cbb752b20-hw0

Short Office Hours

Today right after class,

On Wed. 1/16:

Right before & after class,

After that email me!

(in Bass 432,

contact.gerstein.info)

Biomed. Data science:

More details on Bioinformatics as a sub-discipline

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]

What Information to Organize?

- **Sequences** (DNA & Protein)
- 3D Structures
- Network & Pathway Connectivity
- Phylogenetic tree relationships
- Large-scale gene expression & functional genomics data
- Phenotypic data & medical records....

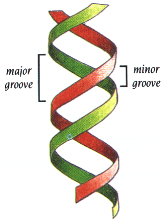
What is the Information?
Molecular Biology as an Information Science

- Central Dogma
  of Molecular Biology

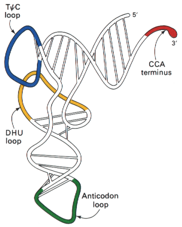
```
DNA
 -> RNA
  -> Protein
   -> Phenotype
    -> DNA
```

- Central Paradigm
  for Bioinformatics

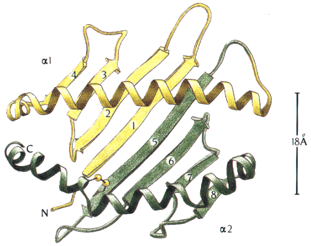
```
Genomic Sequence Information
 -> mRNA (level)
  -> Protein Sequence
   -> Protein Structure
    -> Biological Function
     -> Organismal Phenotype
```



•Genetic material



•Information transfer (mRNA)
•Protein synthesis (tRNA/mRNA)
•Some catalytic activity



Molecular Biology Information - DNA

- Raw DNA Sequence
  - 4 bases: AGCT
  - ~1 K in a gene, ~2 M in genome
  - ~3 Gb Human

```
atggcaattaaaattggtatcaatggttttggtcgtatcggccgtatcgtattccgtgca
gcacaacaccgtgatgacattgaagttgtaggtattaacgacttaatcgacgttgaatac
atggcttatatgttgaaatatgattcaactcacggtcgtttcgacggcactgttgaagtg
aaagatggtaacttagtggttaatggtaaaactatccgtgtaactgcagaacgtgatcca
gcaaacttaaactggggtgcaatcggtgttgatatcgctgttgaagcgactggtttattc
ttaactgatgaaactgctcgtaaacatatcactgcaggcgcaaaaaaagttgtattaact
ggcccatctaaagatgcaacccctatgttcgttcgtggtgtaaacttcaacgcatacgca
ggtcaagatatcgtttctaacgcatcttgtacaacaaactgtttagctcctttagcacgt
gttgttcatgaaactttcggtatcaaagatggtttaatgaccactgttcacgcaacgact
gcaactcaaaaaactgtggatggtccatcagctaaagactggcgcggcggccgcggtgca
tcacaaaacatcattccatcttcaacaggtgcagcgaaagcagtaggtaaagtattacct
gcattaaacggtaaattaactggtatggctttccgtgttccaacgccaaacgtatctgtt
gttgatttaacagttaatcttgaaaaaccagcttcttatgatgcaatcaaacaagcaatc
aaagatgcagcggaaggtaaaacgttcaatggcgaattaaaaggcgtattaggttacact
gaagatgctgttgtttctactgacttcaacggttgtgctttaacttctgtatttgatgca
gacgctggtatcgcattaactgattctttcgttaaattggtatc . . .


. . .   caaaaatagggttaatatgaatctcgatctccattttgttcatcgtattcaa
caacaagccaaaactcgtacaaatatgaccgcacttcgctataaagaacacggcttgtgg
cgagatatctcttggaaaaactttcaagagcaactcaatcaactttctcgagcattgctt
gctcacaatattgacgtacaagataaaatcgccatttttgcccataatatggaacgttgg
gttgttcatgaaactttcggtatcaaagatggtttaatgaccactgttcacgcaacgact
acaatcgttgacattgcgaccttacaaattcgagcaatcacagtgcctatttacgcaacc
aatacagcccagcaagcagaatttatcctaaatcacgccgatgtaaaaattctcttcgtc
ggcgatcaagagcaatacgatcaaacattggaaattgctcatcattgtccaaaattacaa
aaaattgtagcaatgaaatccaccattcaattacaacaagatcctctttcttgcacttgg
```

Molecular Biology Information: Protein Sequence

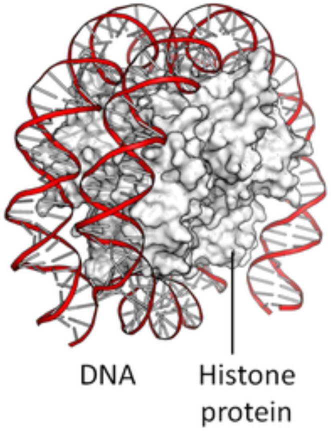
- 20 letter alphabet
  - `ACDEFGHIKLMNPQRSTVWY` but not `BJOUXZ`
- Strings of ~300 aa in an average protein (in bacteria), ~200 aa in a domain
- >12 M known protein sequences (uniprot, http://www.ebi.ac.uk/uniprot/TrEMBLstats/, 2011)

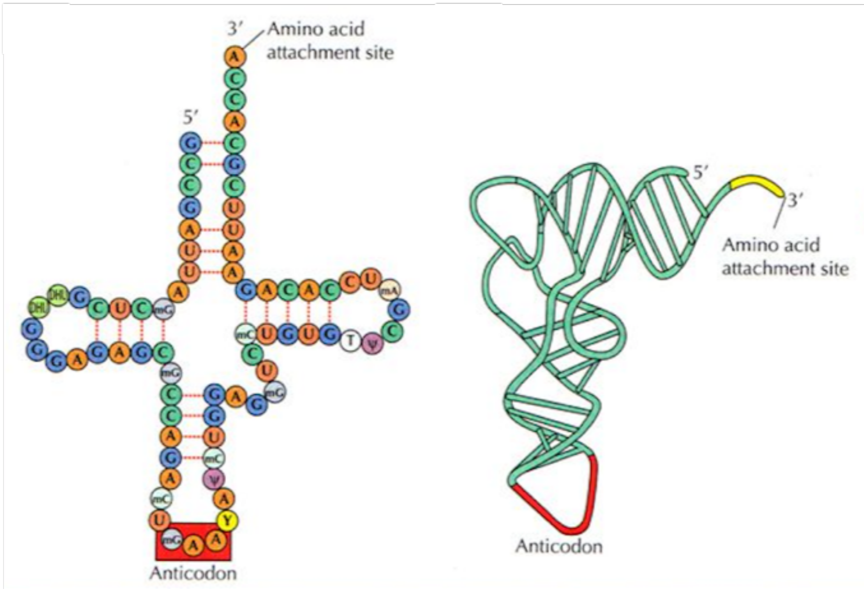
```
d1dhfa_ LNCIVAVSQNMGIGKNGDLPWPPLRNEFRYFQRMTTTSSVEGKQ-NLVIMGKKTWFSI
d8dfr__ LNSIVAVCQNMGIGKDGNLPWPPLRNEYKYFQRMTSTSHVEGKQ-NAVIMGKKTWFSI
d4dfra_ ISLIAALAVDRVIGMENAMPWN-LPADLAWFKRNTL--------NKPVIMGRHTWESI
d3dfr__ TAFLWAQDRDGLIGKDGHLPWH-LPDDLHYFRAQTV--------GKIMVVGRRTYESF

d1dhfa_ LNCIVAVSQNMGIGKNGDLPWPPLRNEFRYFQRMTTTSSVEGKQ-NLVIMGKKTWFSI
d8dfr__ LNSIVAVCQNMGIGKDGNLPWPPLRNEYKYFQRMTSTSHVEGKQ-NAVIMGKKTWFSI
d4dfra_ ISLIAALAVDRVIGMENAMPW-NLPADLAWFKRNTLD--------KPVIMGRHTWESI
d3dfr__ TAFLWAQDRNGLIGKDGHLPW-HLPDDLHYFRAQTVG--------KIMVVGRRTYESF

d1dhfa_ VPEKNRPLKGRINLVLSRELKEPPQGAHFLSRSLDDALKLTEQPELANKVDMVWIVGGSSVYKEAMNHP
d8dfr__ VPEKNRPLKDRINIVLSRELKEAPKGAHYLSKSLDDALALLDSPELKSKVDMVWIVGGTAVYKAAMEKP
d4dfra_ ---G-RPLPGRKNIILS-SQPGTDDRV-TWVKSVDEAIAACGDVP------EIMVIGGGRVYEQFLPKA
d3dfr__ ---PKRPLPERTNVVLTHQEDYQAQGA-VVVHDVAAVFAYAKQHLDQ----ELVIAGGAQIFTAFKDDV

d1dhfa_ -PEKNRPLKGRINLVLSRELKEPPQGAHFLSRSLDDALKLTEQPELANKVDMVWIVGGSSVYKEAMNHP
d8dfr__ -PEKNRPLKDRINIVLSRELKEAPKGAHYLSKSLDDALALLDSPELKSKVDMVWIVGGTAVYKAAMEKP
d4dfra_ -G---RPLPGRKNIILSSSQPGTDDRV-TWVKSVDEAIAACGDVPE----- IMVIGGGRVYEQFLPKA
d3dfr__ -P--KRPLPERTNVVLTHQEDYQAQGA-VVVHDVAAVFAYAKQHLD----QELVIAGGAQIFTAFKDDV
```

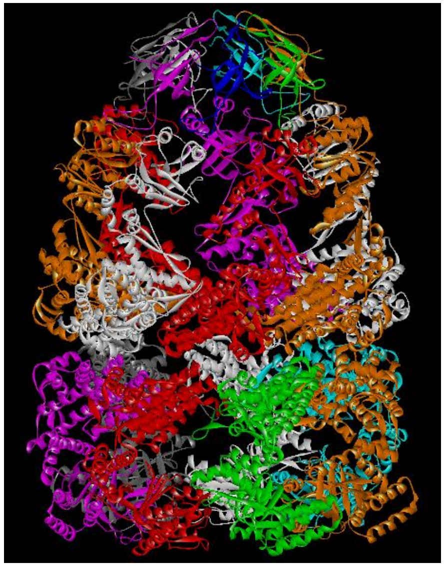
Molecular Biology Information: Macromolecular Structure

- DNA/RNA/Protein
  - Mostly protein

(RNA Adapted From D Soll Web Page,
Right Hand Top Protein from M Levitt web page)

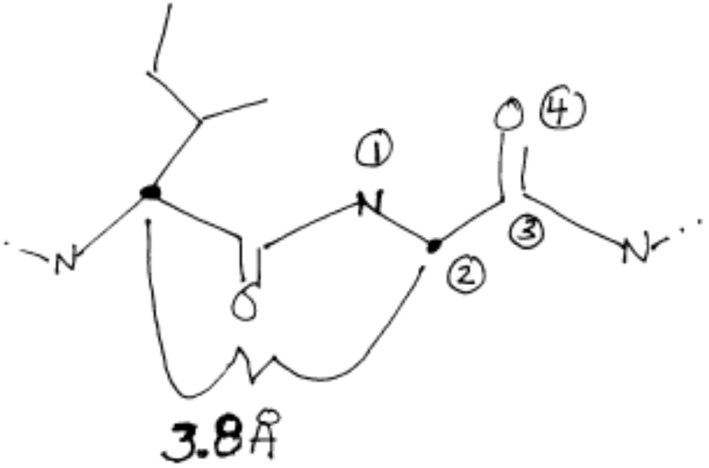






Molecular Biology Information: Protein Structure Details

- Statistics on Number of XYZ triplets
  - 200 residues/domain => 200 CA atoms, separated by 3.8 A
  - Avg. Residue is Leu: 4 backbone atoms + 4 sidechain atoms, 150 cubic A => ~1500 xyz triplets (=8x200) per p
  - >100K Domains, ~1200 folds (scop 1



```
ATOM      1  C   ACE     0       9.401  30.166  60.595  1.00 49.88      1GKY  67
ATOM      2  O   ACE     0      10.432  30.832  60.722  1.00 50.35      1GKY  68
ATOM      3  CH3 ACE     0       8.876  29.767  59.226  1.00 50.04      1GKY  69
ATOM      4  N   SER     1       8.753  29.755  61.685  1.00 49.13      1GKY  70
ATOM      5  CA  SER     1       9.242  30.200  62.974  1.00 46.62      1GKY  71
ATOM      6  C   SER     1      10.453  29.500  63.579  1.00 41.99      1GKY  72
ATOM      7  O   SER     1      10.593  29.607  64.814  1.00 43.24      1GKY  73
ATOM      8  CB  SER     1       8.052  30.189  63.974  1.00 53.00      1GKY  74
ATOM      9  OG  SER     1       7.294  31.409  63.930  1.00 57.79      1GKY  75
ATOM     10  N   ARG     2      11.360  28.819  62.827  1.00 36.48      1GKY  76
ATOM     11  CA  ARG     2      12.548  28.316  63.532  1.00 30.20      1GKY  77
ATOM     12  C   ARG     2      13.502  29.501  63.500  1.00 25.54      1GKY  78
. . .
ATOM   1444  CB  LYS   186      13.836  22.263  57.567  1.00 55.06      1GKY1510
ATOM   1445  CG  LYS   186      12.422  22.452  58.180  1.00 53.45      1GKY1511
ATOM   1446  CD  LYS   186      11.531  21.198  58.185  1.00 49.88      1GKY1512
ATOM   1447  CE  LYS   186      11.452  20.402  56.860  1.00 48.15      1GKY1513
ATOM   1448  NZ  LYS   186      10.735  21.104  55.811  1.00 48.41      1GKY1514
ATOM   1449  OXT LYS   186      16.887  23.841  56.647  1.00 62.94      1GKY1515
TER    1450      LYS   186                                              1GKY1516
```

Molecular Biology Information: Whole Genomes

- The Revolution Driving Everything

  Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., Fitzhugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L. I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O. & Venter, J. C. (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae rd." *Science* 269: 496-512.

  (Picture adapted from TIGR website, http://www.tigr.org)

- Timeline

  1995, HI (bacteria): 1.6 Mb & 1600 genes done
  1997, yeast: 13 Mb & ~6000 genes for yeast
  1998, worm: ~100Mb with 19 K genes
  1999: >30 completed genomes!
  2000, draft human
  2003, human: 3 Gb & 100 K genes...
  2010, 1000 human genomes!
  2017, 13K human genomes

1995

Bacteria,
1.6 Mb,
~1600 genes
[*Science* **269**: 496]

1997

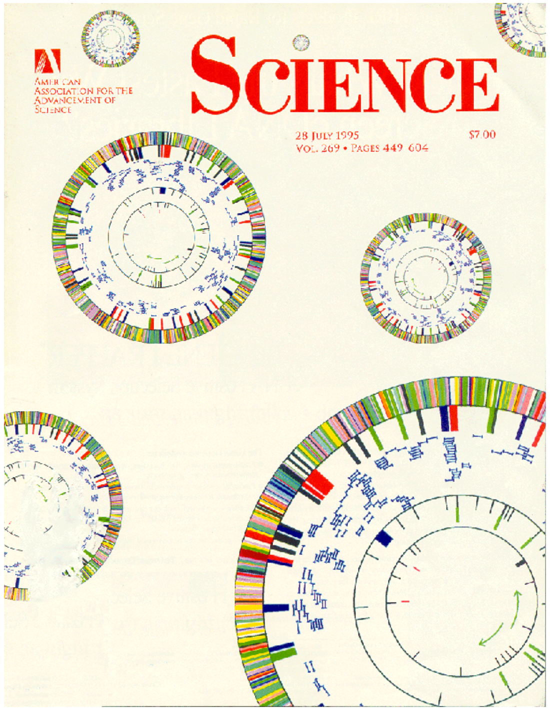
Eukaryote,
13 Mb,
~6K genes
[Nature **387**: 1]

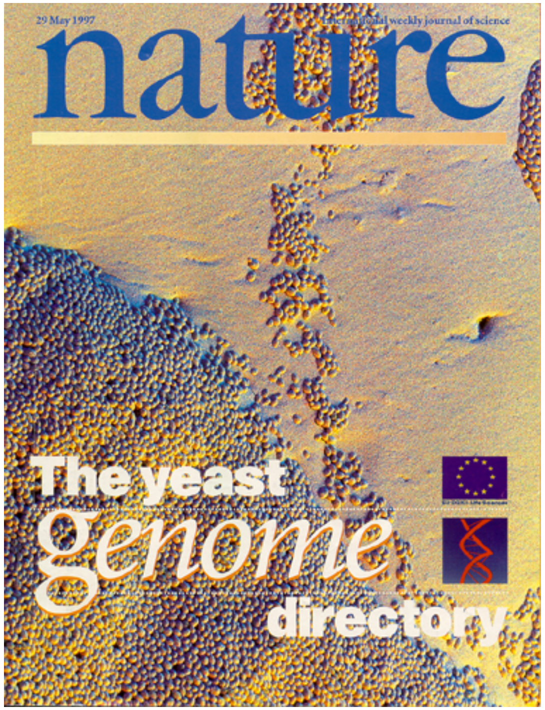
1998

Animal,
~100 Mb,
~20K genes
[*Science* **282**: 1945]

2000?

Human,
~3 Gb,
~20K genes





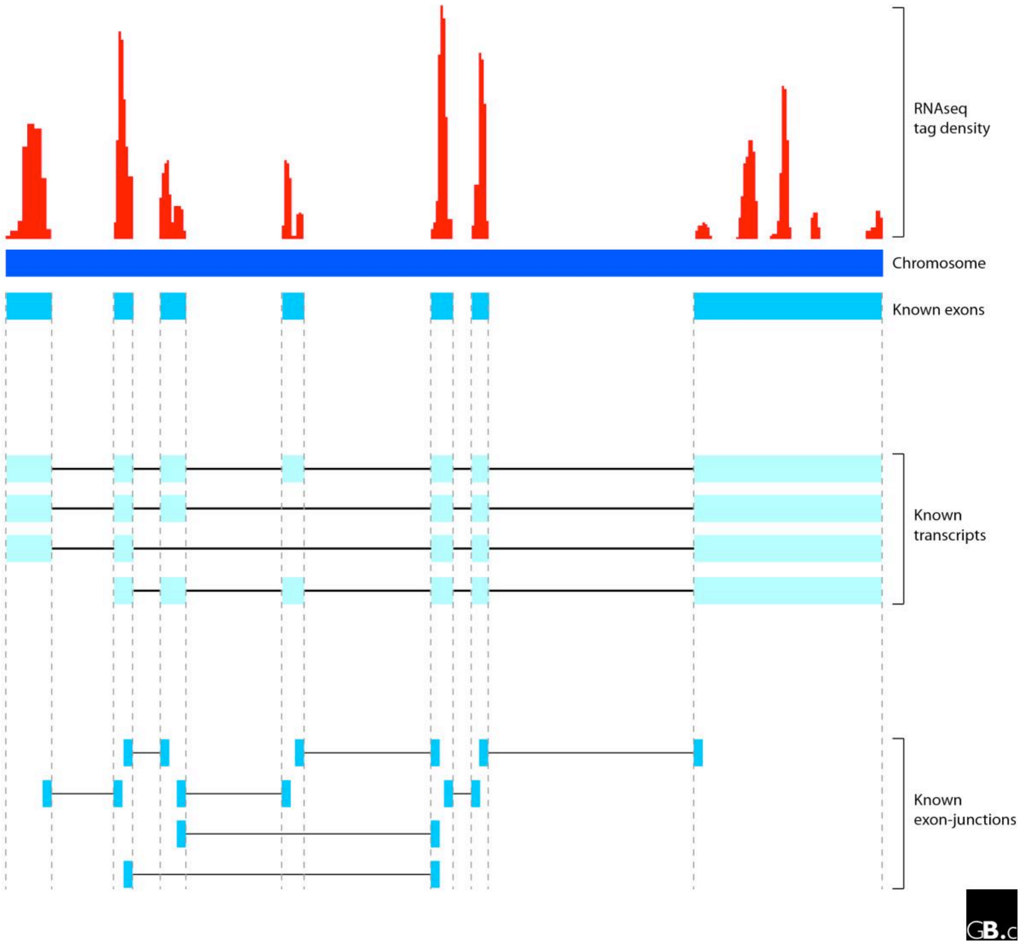
A Bioinformatics prediction that came true!

'98 spoof

⟷

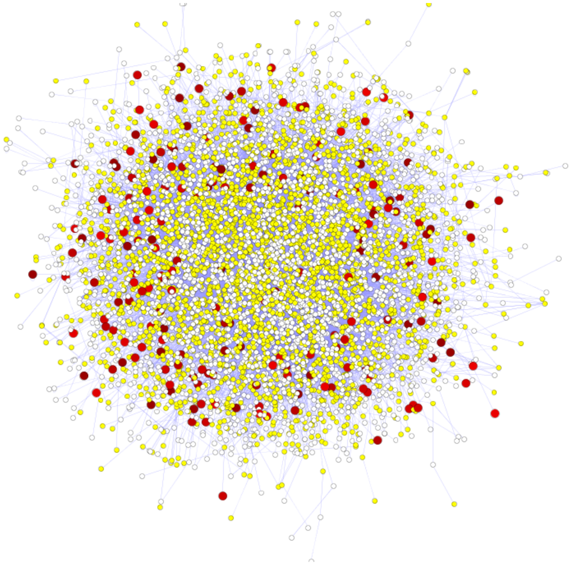
real thing, Apr '00

Gene Expression Data: On & Off

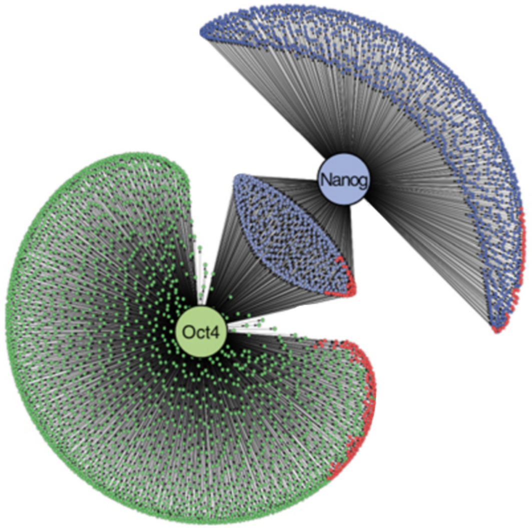


- Early experiments yeast
  - Complexity at 10 time points,
    6000 x 10 = 60K floats
- Then tiling array technology
  - 50 M data points to tile the human genome at ~50 bp res.
- Now Next-Gen Sequencing (RNAseq)
  - 10M+ reads on the human genome, counts
- Can only sequence genome once but can do an infinite variety of expression experiments

Molecular Networks: Connectivity



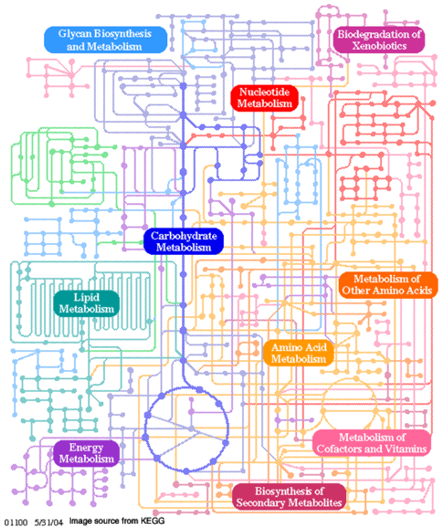
Protein-protein Interaction networks



TF-target-gene Regulatory networks

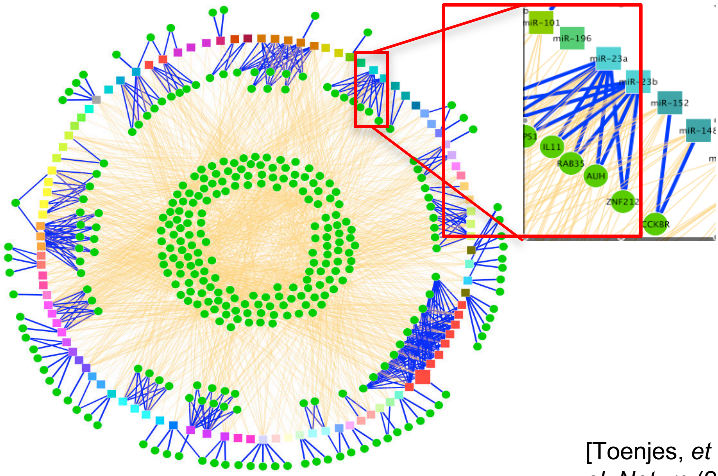
Regulatory Networks
- Get readouts of where proteins bind to DNA : Chip-chip then chip-seq

Protein Interaction Networks
- For yeast: 6000 x 6000 / 2 ~ 18M possible interactions (maybe ~30K real)



01100 5/31/04 Image source from KEGG

Metabolic pathway networks



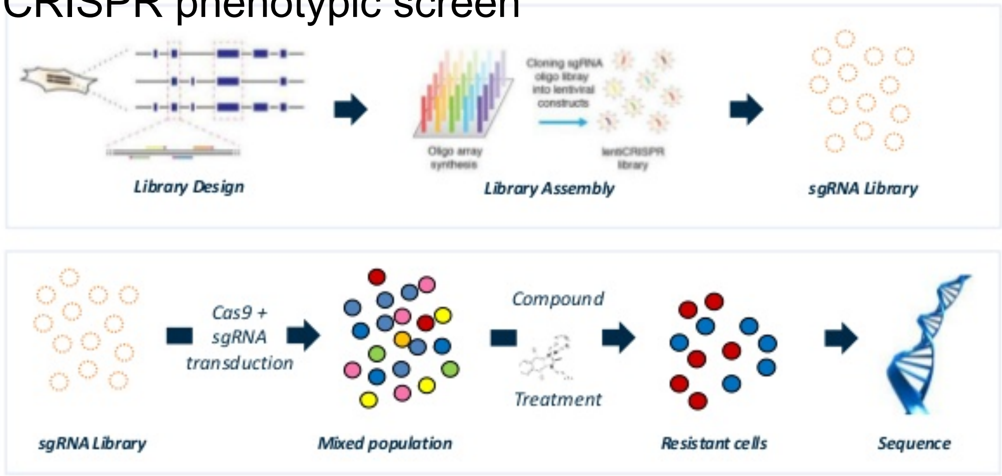
miRNA-target networks

[Toenjes, *et al*, *Mol. BioSyst.* (2008); Jeong *et al*, *Nature* (2001); [Horak, et al, Genes & Development, 16:3017-3033; DeRisi, Iyer, and Brown, Science, 278:680-686; Descalzo, et al, Mol Syst Biol, 9:694]

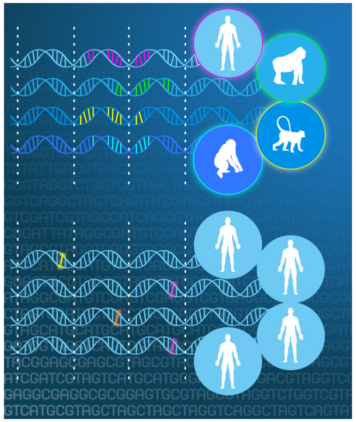
Molecular Biology Information: Other Integrative Data

- Information to understand genomes
  - Whole Organisms Phylogeny, traditional zoology
  - Environments, Habitats, ecology
  - Phenotype Experiments (large-scale KOs, transposons)
  - The Literature (MEDLINE)
- The Future....

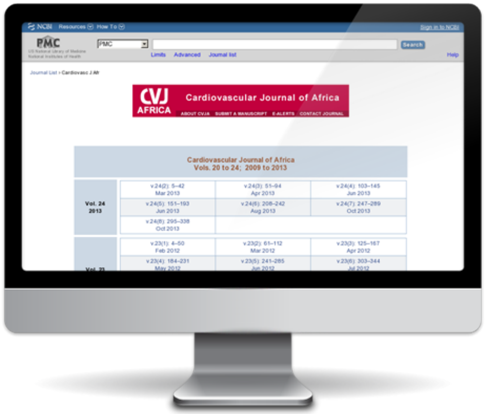
CRISPR phenotypic screen



Comparative genomics



Literature knowledgebase



Cardiovascular Journal of Africa
Vols. 20 to 24; 2009 to 2013

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**.

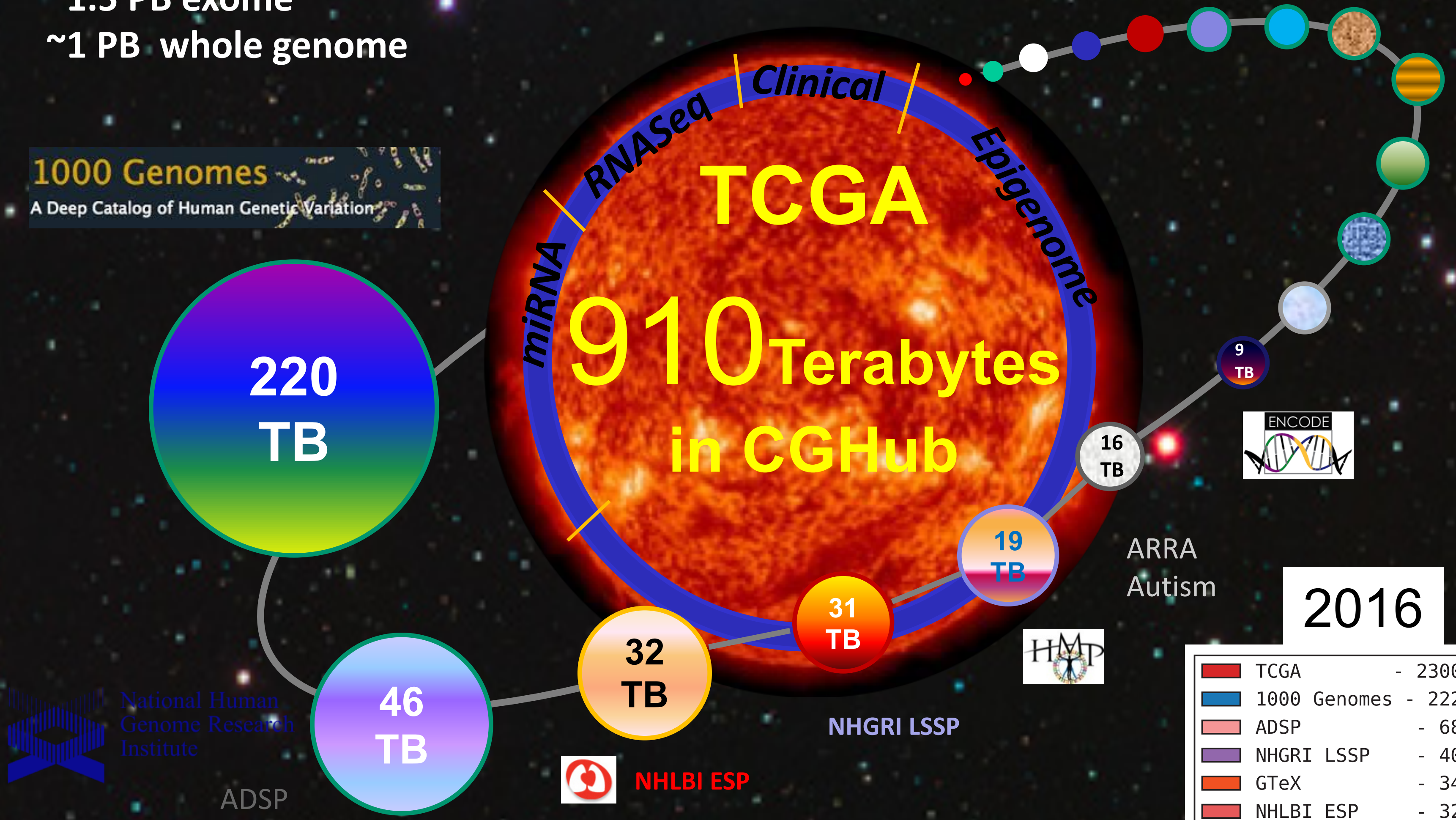
[Luscombe et al. ('01). *Methods Inf Med* 40: 346]

Seq Universe

[from Heidi Sofia, NHGRI]

SRA >1 petabyte

TCGA endpoint: ~2.5 Petabytes
~1.5 PB exome
~1 PB whole genome



2016

🟥	TCGA	- 2300 TB
🟦	1000 Genomes	- 222 TB
🟥	ADSP	- 68 TB
🟪	NHGRI LSSP	- 40 TB
🟧	GTeX	- 34 TB
🟥	NHLBI ESP	- 32 TB
🟦	HMP	- 29 TB
🟥	ARRA Autism	- 24 TB
🟦	ENCODE	- 9 TB

Sofia, 2-28-14

What is Bioinformatics?

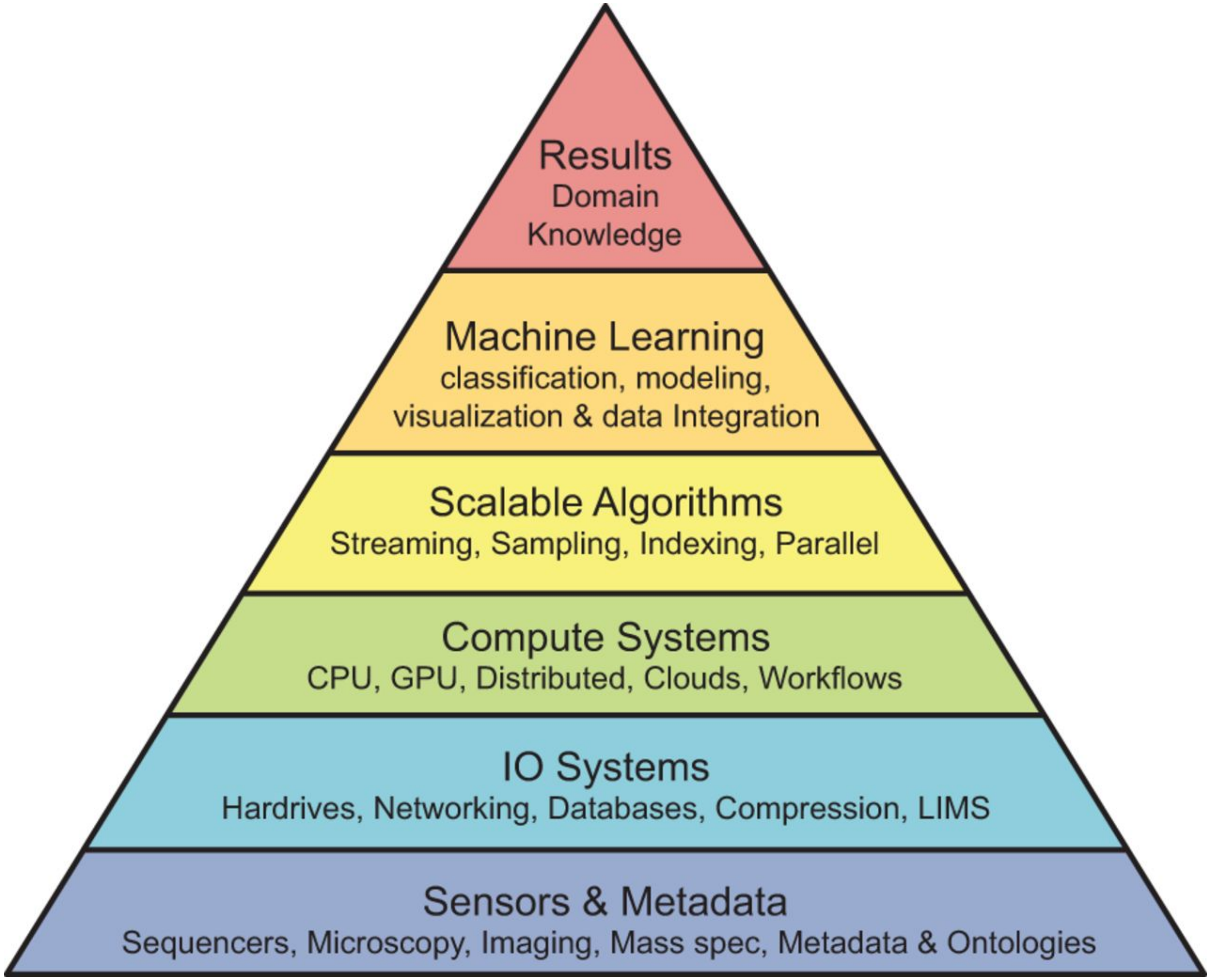
- *(Molecular)* **Bio** - **`informatics`**
- One idea for a definition?
  Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **"informatics" techniques** (derived from disciplines such as applied math, CS, and statistics) to **organize, 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]

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

Data science analysis stack.



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

Bioinformatics

Key Practical Applications

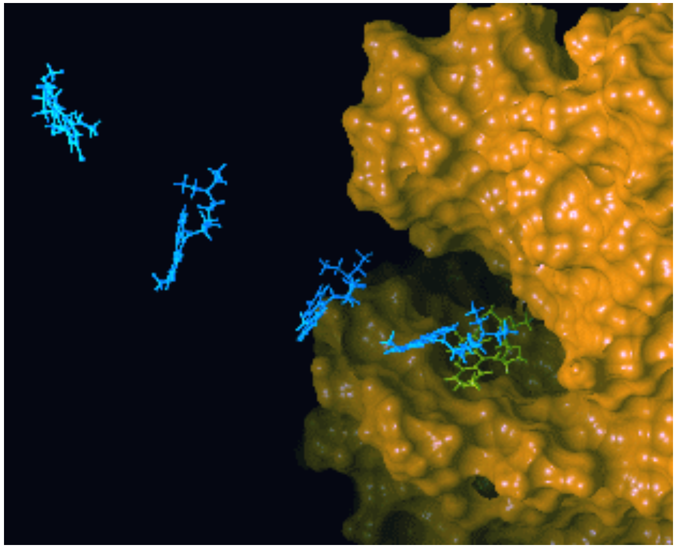
What is Bioinformatics?

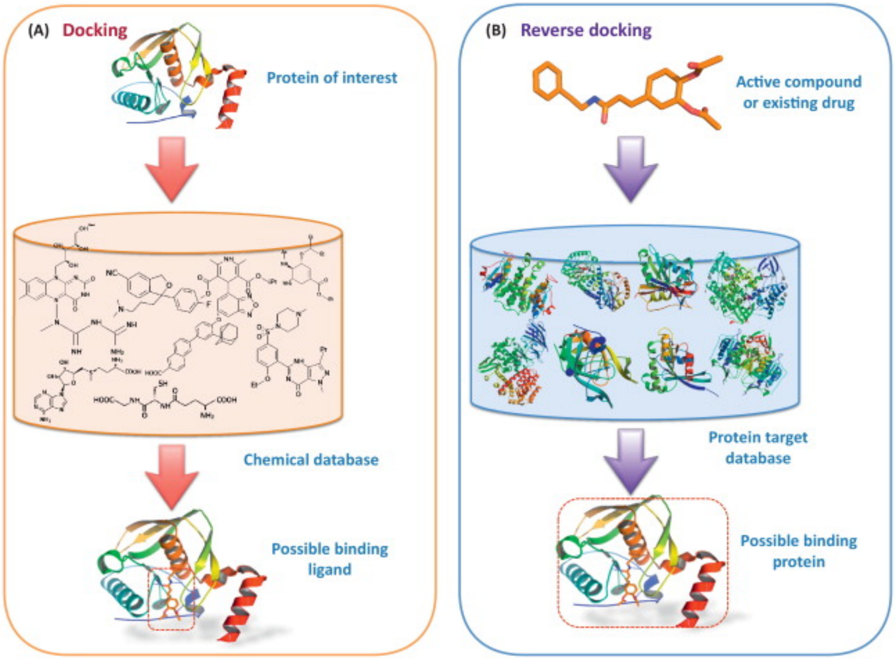
- *(Molecular)* **Bio** - `informatics`
- One idea for a definition?
  Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **"informatics" techniques** (derived from disciplines such as applied math, CS, and statistics) to **organize, 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]

Major Application I:
Designing Drugs from Structural Targets

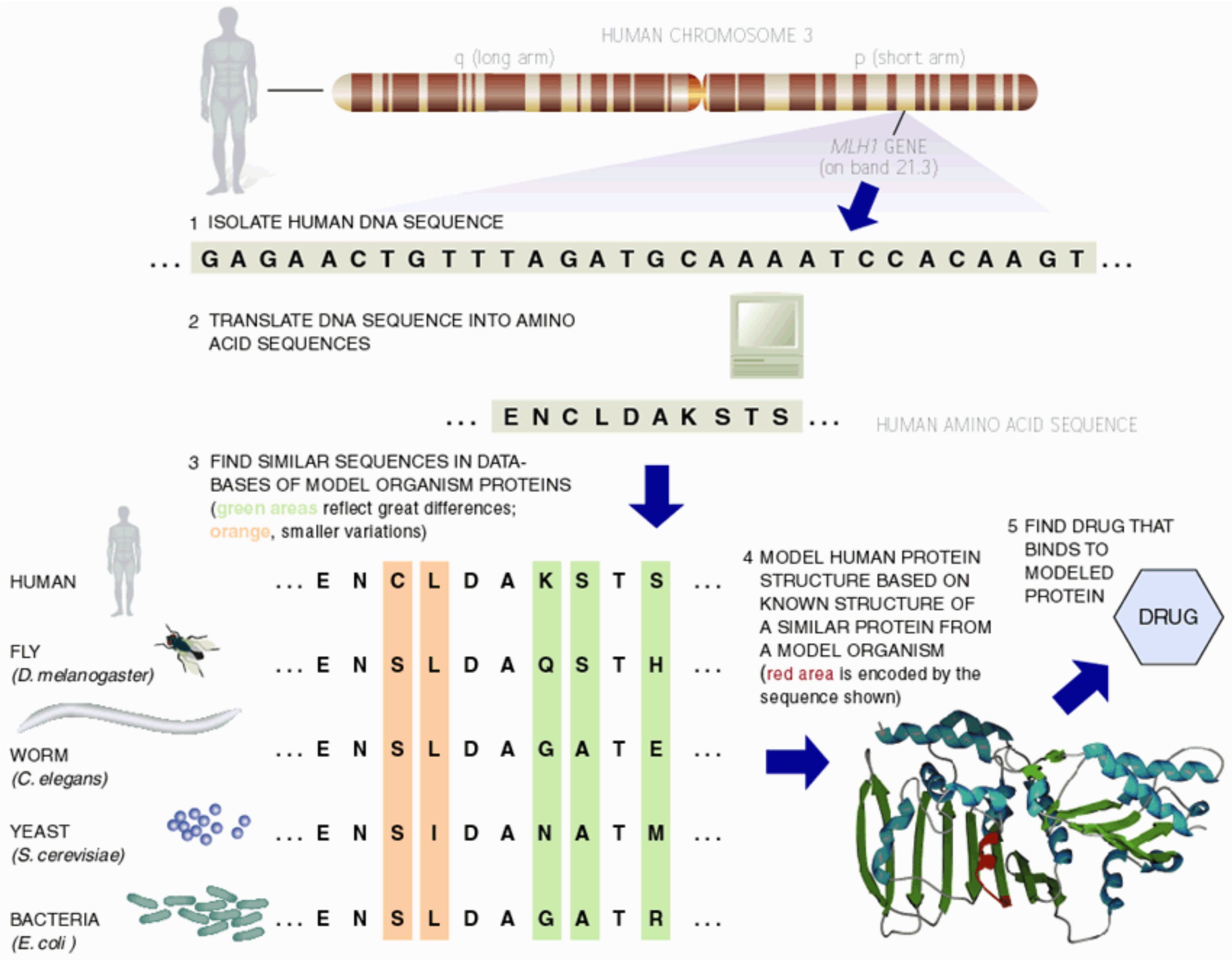
- Understanding how structures bind other molecules
- Designing inhibitors using docking, structure modeling
- *In silico* screens of chemical and protein databases





(From left to right, figures adapted from Olsen Group Docking Page at Scripps, Zheng et al. Trends in Pharmacological Sciences 2013)

Major Application II: Finding Homologs, to Find Experimentally Tractable Gene Targets



HUMAN CHROMOSOME 3

q (long arm)

p (short arm)

MLH1 GENE (on band 21.3)

1 ISOLATE HUMAN DNA SEQUENCE

... GAGAACTGTTTAGATGCAAAATCCACAAGT ...

2 TRANSLATE DNA SEQUENCE INTO AMINO ACID SEQUENCES

... ENCLDAKSTS ... HUMAN AMINO ACID SEQUENCE

3 FIND SIMILAR SEQUENCES IN DATABASES OF MODEL ORGANISM PROTEINS (green areas reflect great differences; orange, smaller variations)

HUMAN	... E	N	C	L	D	A	K	S	T	S	...
FLY (*D. melanogaster*)	... E	N	S	L	D	A	Q	S	T	H	...
WORM (*C. elegans*)	... E	N	S	L	D	A	G	A	T	E	...
YEAST (*S. cerevisiae*)	... E	N	S	I	D	A	N	A	T	M	...
BACTERIA (*E. coli*)	... E	N	S	L	D	A	G	A	T	R	...

4 MODEL HUMAN PROTEIN STRUCTURE BASED ON KNOWN STRUCTURE OF A SIMILAR PROTEIN FROM A MODEL ORGANISM (red area is encoded by the sequence shown)

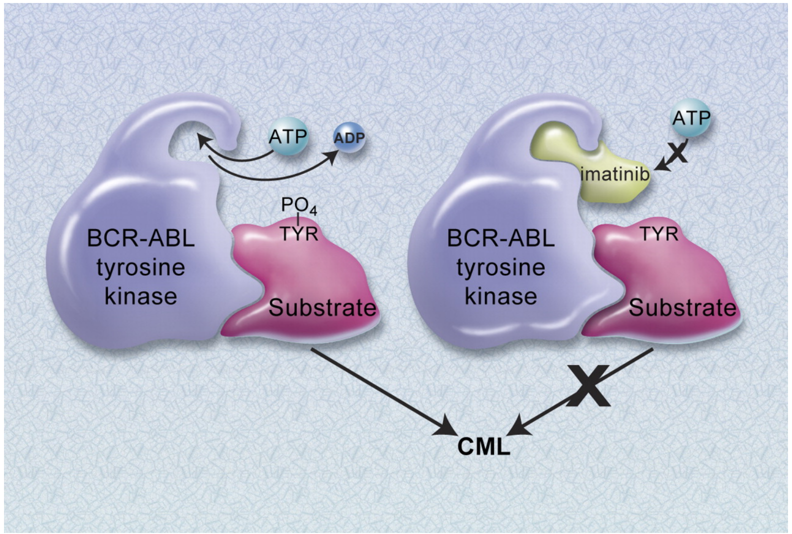
5 FIND DRUG THAT BINDS TO MODELED PROTEIN

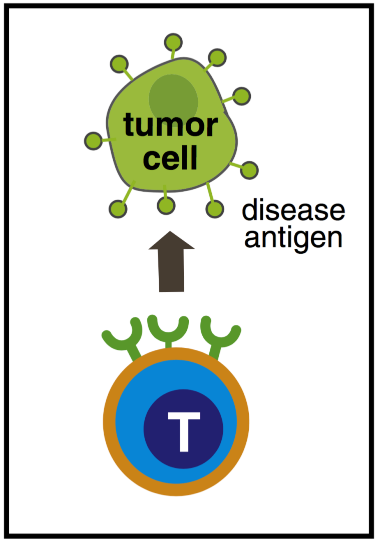
DRUG

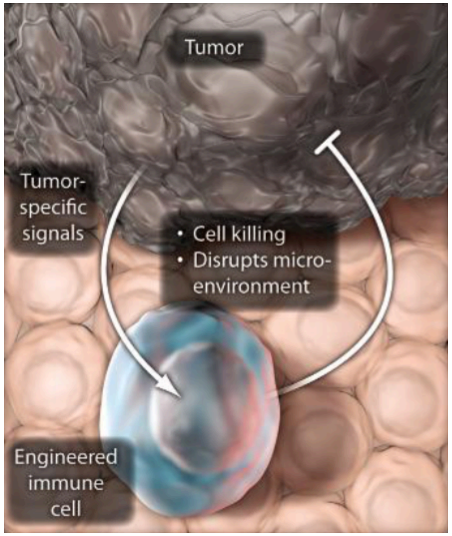
[Adapted from Sci. Am.]

Major Application III:
Customizing treatment in oncology

- Identifying disease causing mutations in individual patients
- Designing targeted therapeutics
  - e.g. BCR-abl and Gleevec
  - Cancer immunotherapies targeting neoantigens



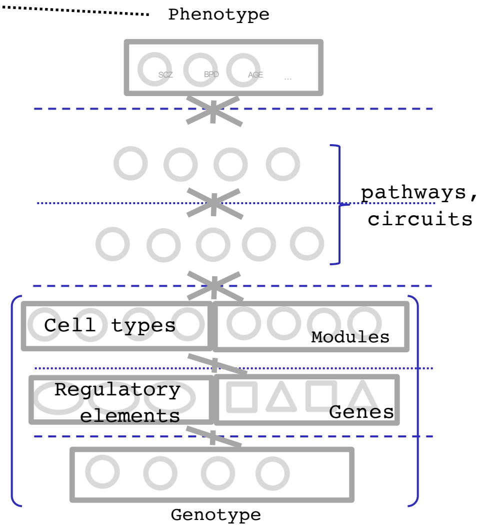




(From left to right, figures adapted from Druker BJ. Blood 2008 and the Lim Lab at UCSF)

Major Application IV:
Finding molecular mechanisms & drug targets for diseases we know little about (Neuro-psychiatic Diseases)

Disease	Heritability*	Molecular **Mechanisms**
Schizophrenia	**81%**	**-**
Bipolar disorder	70%	**-**
Alzheimer's disease	58 - 79%	Apolipoprotein E (APOE), Tau
Hypertension	30%	Renin–angiotensin–aldosterone
Heart disease	34-53%	Atherosclerosis, VCAM-1
Stroke	32%	Reactive oxygen species (ROS), Ischemia
Type-2 diabetes	26%	Insulin resistance
Breast Cancer	25-56%	BRCA, PTEN



*https://www.snpedia.com/index.php/Heritability

Many psychiatric conditions are highly heritable
- Schizophrenia: up to 80%

But we don't understand basic molecular mechanisms underpinning this association
- (in contrast to many other diseases such as cancer & heart disease)

Moreover, current models substantially underestimate heritability using genetic data
- Schizophrenia : ~25%

Thus, interested in developing predictive models of psychiatric traits which:
- Use observations at intermediate (molecular levels) levels to inform latent structure.
- Use the predictive features of these "molecular endo phenotypes" to begin to suggest actors involved in mechanism

Major Application IV: Finding molecular mechanisms & drug targets for diseases we know little about (Neuro-psychiatic Diseases)

Disease	Heritability*	Molecular **Mechanisms**
Schizophrenia	**81%**	-
Bipolar disorder	70%	-
Alzheimer's disease	58 - 79%	Apolipoprotein E (APOE), Tau
Hypertension	30%	Renin–angiotensin–aldosterone
Heart disease	34-53%	Atherosclerosis, VCAM-1
Stroke	32%	Reactive oxygen species (ROS), Ischemia
Type-2 diabetes	26%	Insulin resistance
Breast Cancer	25-56%	BRCA, PTEN

Many psychiatric conditions are highly heritable

Schizophrenia: up to 80%

But we don't understand basic molecular mechanisms underpinning this association

(in contrast to many e)

Moreover, current model jenetic data

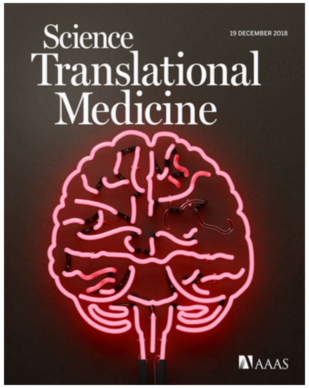
Schizophrenia : ~2

Thus, interested in devel sh:

Use observations a n latent structure.

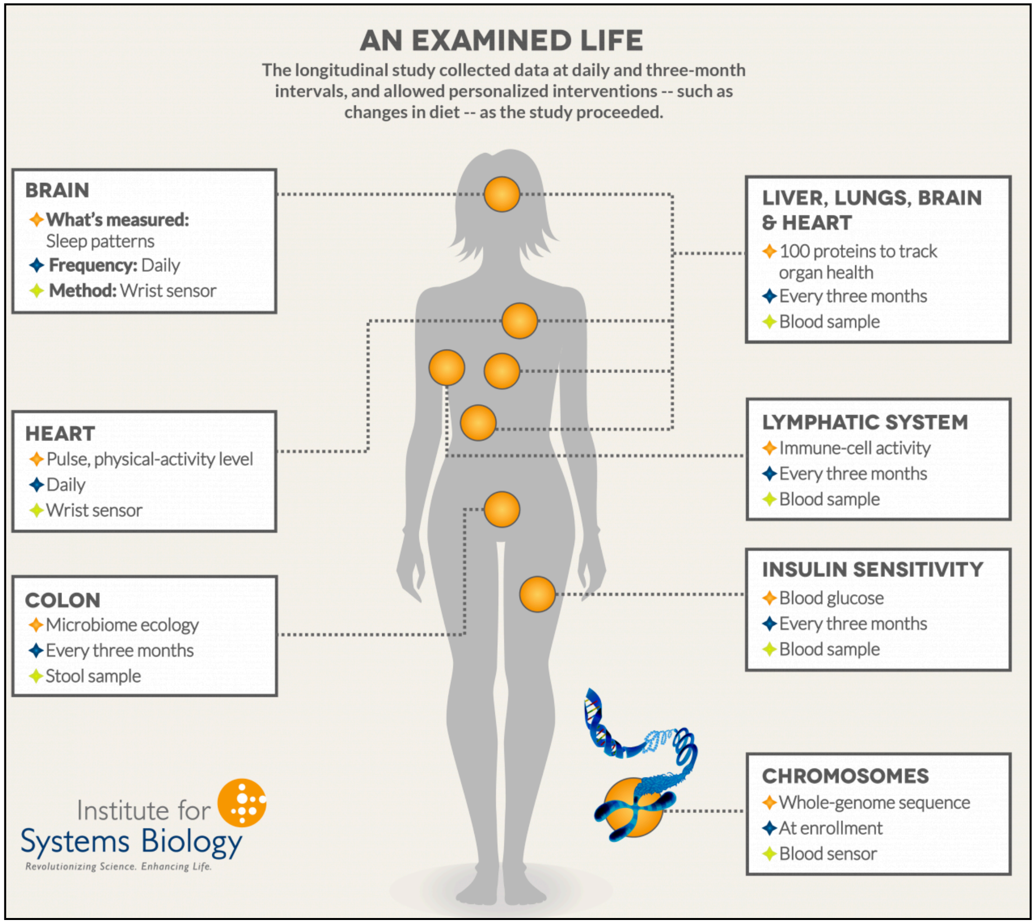
Use the predictive features of these “molecular endo phenotypes” to begin to suggest actors involved in mechanism

Recent Rollout in Science addressing this, involving many Yale Researchers



*https://www.snpedia.com/index.php/Heritability

Major Application V: Holistic Personal Genome Characterization, in Normal Individuals



(Figure from Institute for Systems Biology)

- Mental disease & cancer are two extremes with respect to genomics (CEN, 92: 26)
  - Many other conditions in between, often involving interaction with the environment
- Pers. Genome Characterization
  - Identify mutations in personal genomes (SNPs, SVs, &c)
  - Estimate phenotypic (deleterious or protective) impact of variants.
  - Compare one person to wider population.
- Track changes over time & consider interaction w/ environment
  - Transcriptome studies
  - Longitudinal health studies (e.g. 100K wellness project, Framingham Heart Study)

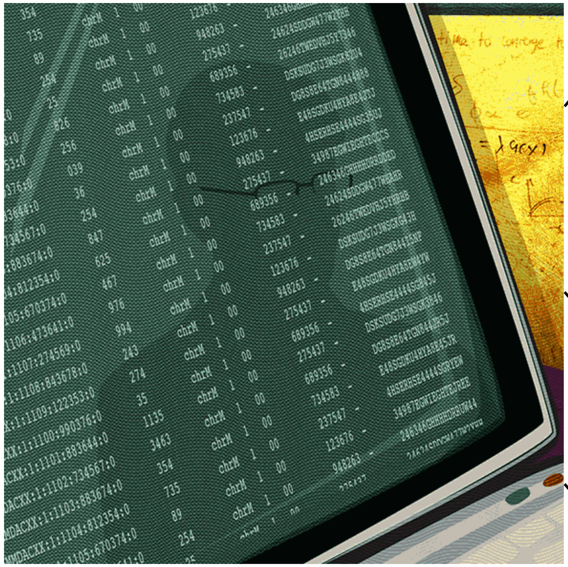
The Key Application in Personal Genomics

More Details

Personal Genomics as an an organizing theme for this class

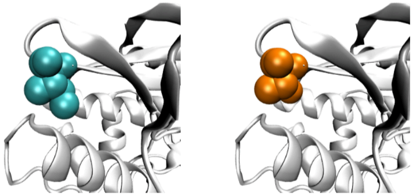
- A personal genome can reveal a lot about an individual.
  - Disease risks, ancestry, personal traits, etc.
- In this class, we will look at how to identify key genomic variants with the most impact.
- Next step: Personal genome annotation combined with multi-omic and longitudinal health data can inform new links between genotype and phenotype relevant to an individual and the larger population. It can act as a center for large-scale data integration.
- In preparation for this future, we will use analysis techniques including systems and network modeling as well as structural modeling to contextualize future multi-omics data sets and better interpret the mechanisms through which variants impact health
- Genomic privacy will become increasingly important as precision medicine becomes more common.

Analyzing Carl Zimmer’s genome



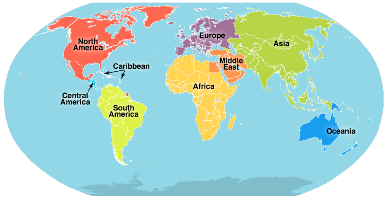
SNV AAGCT → ACGCT

Protein Structure

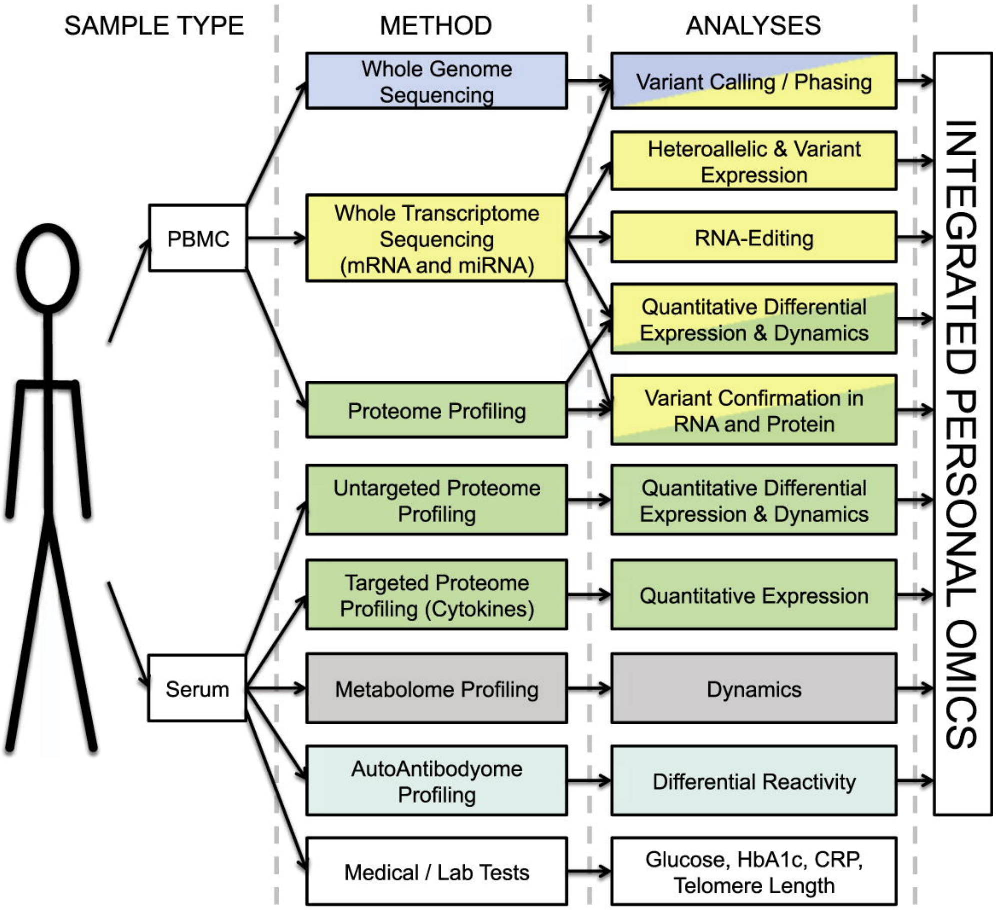


Wild-type *Mutated*

Ancestry



Personal Omics Profiling



SAMPLE TYPE

METHOD

ANALYSES

PBMC

Serum

Whole Genome Sequencing

Whole Transcriptome Sequencing (mRNA and miRNA)

Proteome Profiling

Untargeted Proteome Profiling

Targeted Proteome Profiling (Cytokines)

Metabolome Profiling

AutoAntibodyome Profiling

Medical / Lab Tests

Variant Calling / Phasing

Heteroallelic & Variant Expression

RNA-Editing

Quantitative Differential Expression & Dynamics

Variant Confirmation in RNA and Protein

Quantitative Differential Expression & Dynamics

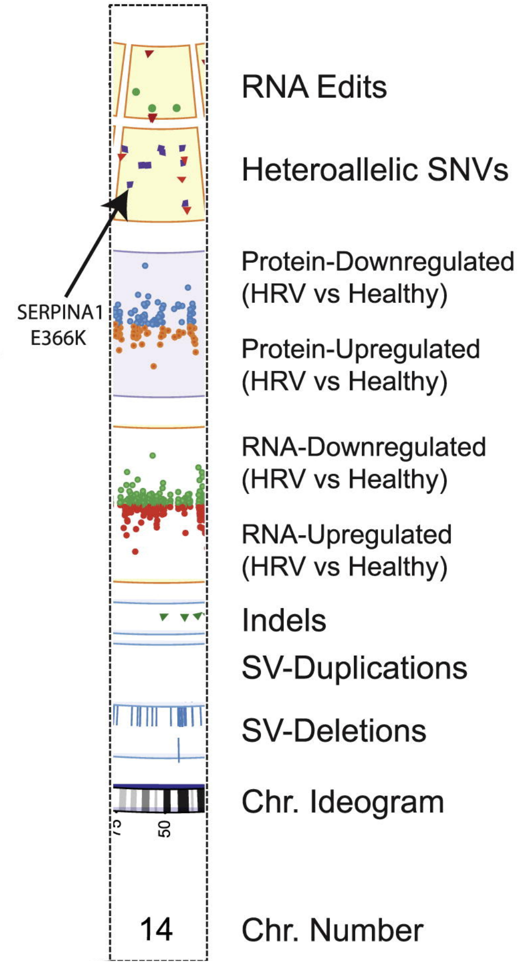
Quantitative Expression

Dynamics

Differential Reactivity

Glucose, HbA1c, CRP, Telomere Length

INTEGRATED PERSONAL OMICS



(Figure from Chen et al. Cell 2012)

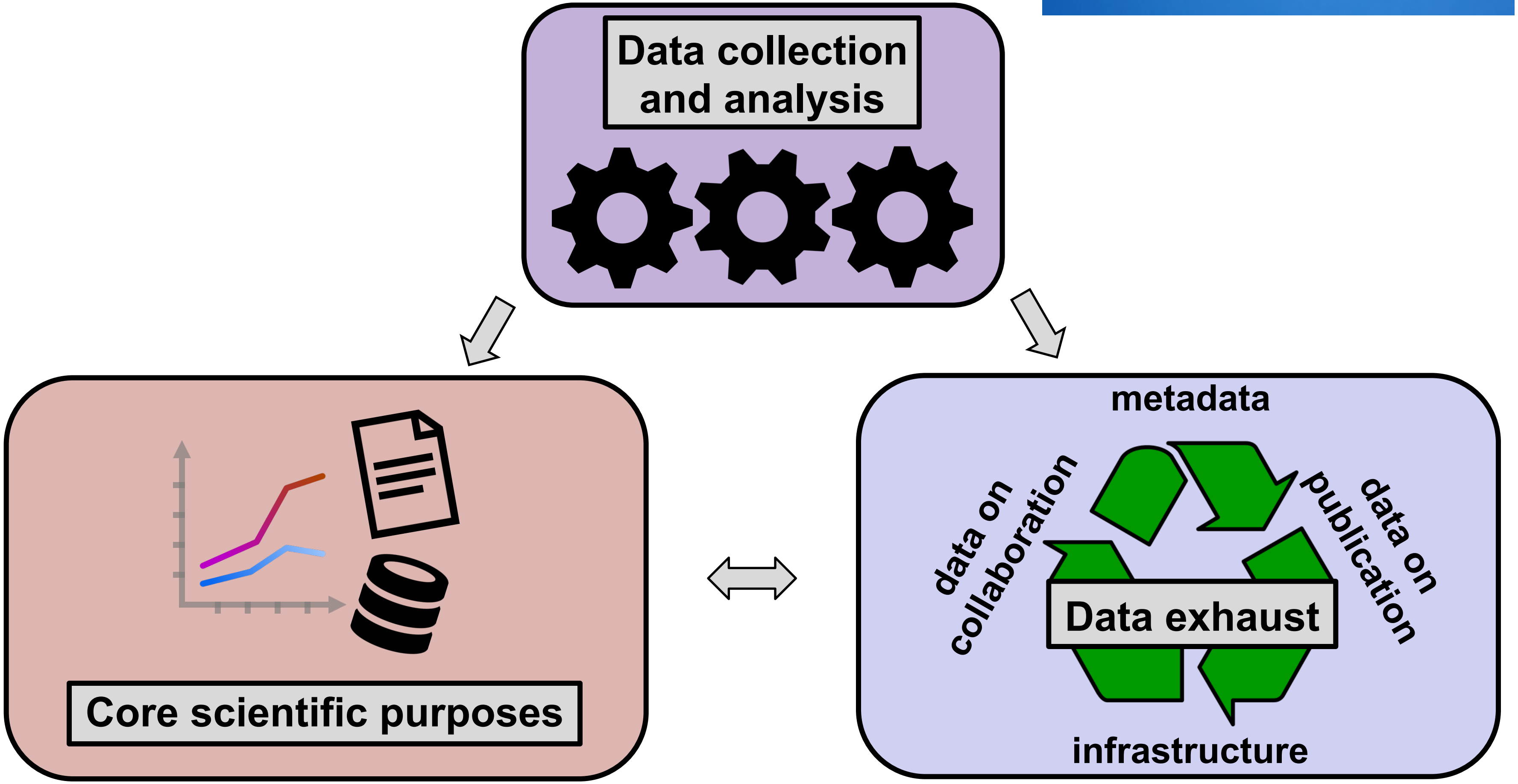
The Other Side of the Data Science Coin:

The Data Exhaust from Personal Genomics

(SOS)

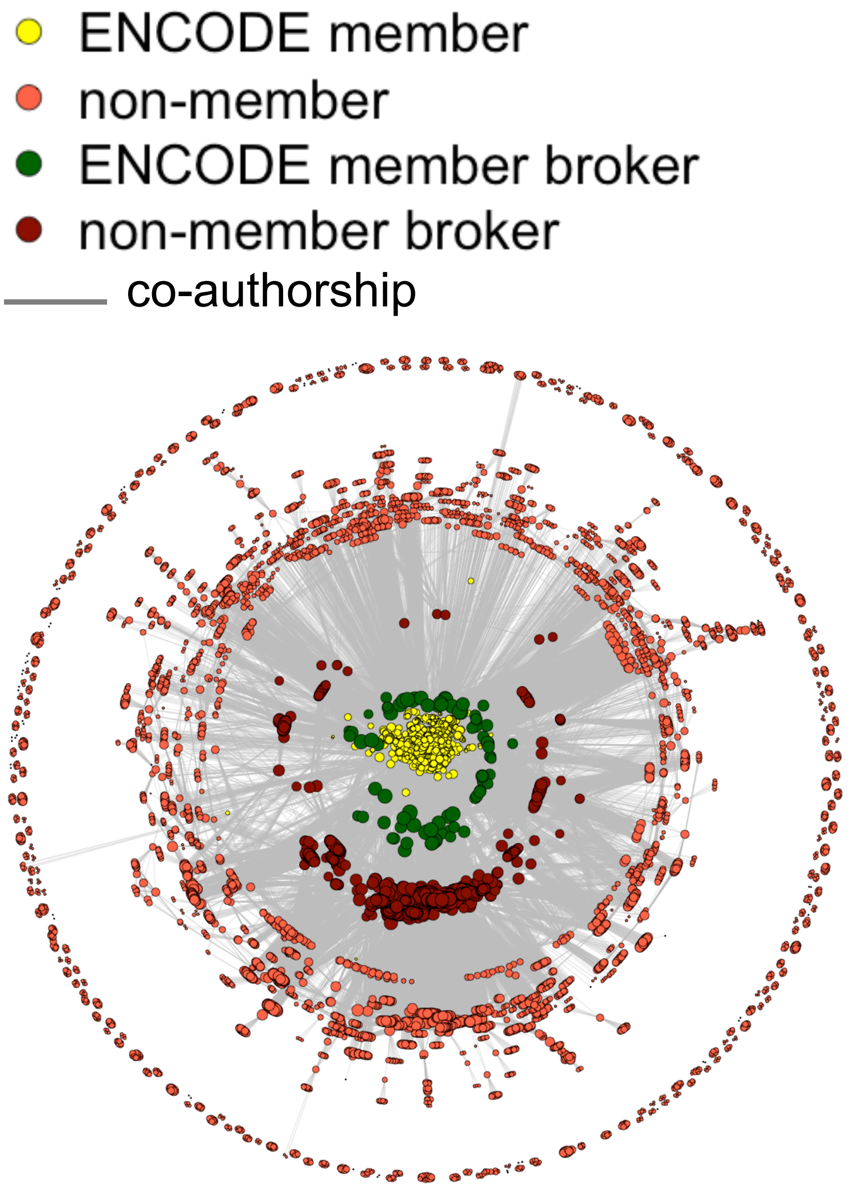
Data Exhaust

- Creative use of data is key to data science!
- Data exhaust = exploitable byproducts of big data collection and analysis



[photos: wikipedia/wikimedia]

Exhaust Mining Application: Using Science to Study Science (SOS)

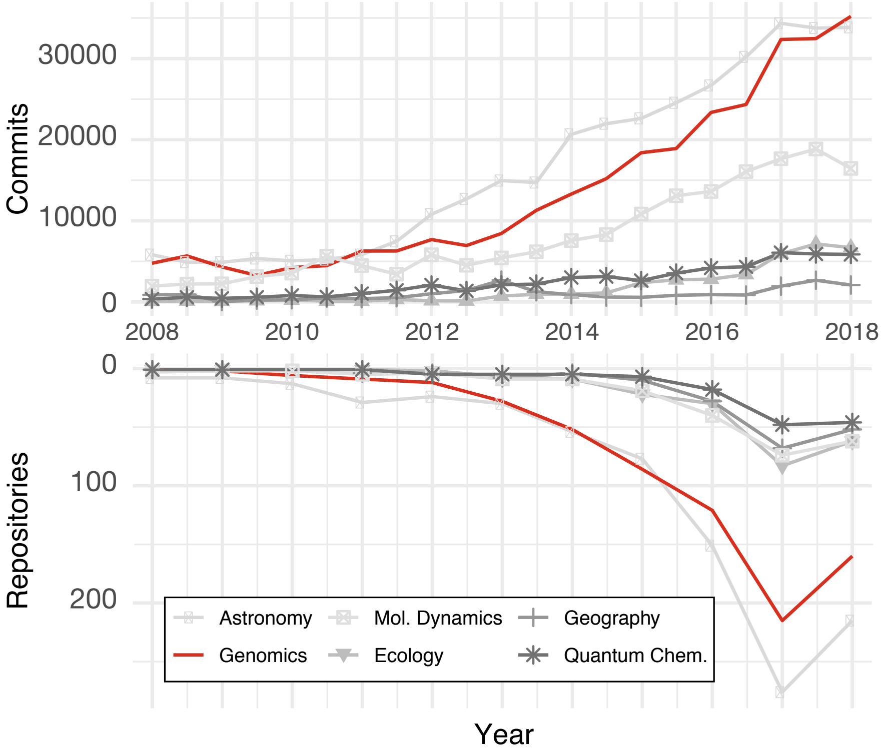


- Mining output of science (Scientific Publications) to understand how science works as a social enterprise

- Co-authorship network of members of the human genome annotation group (ENCODE) & users of this groups data

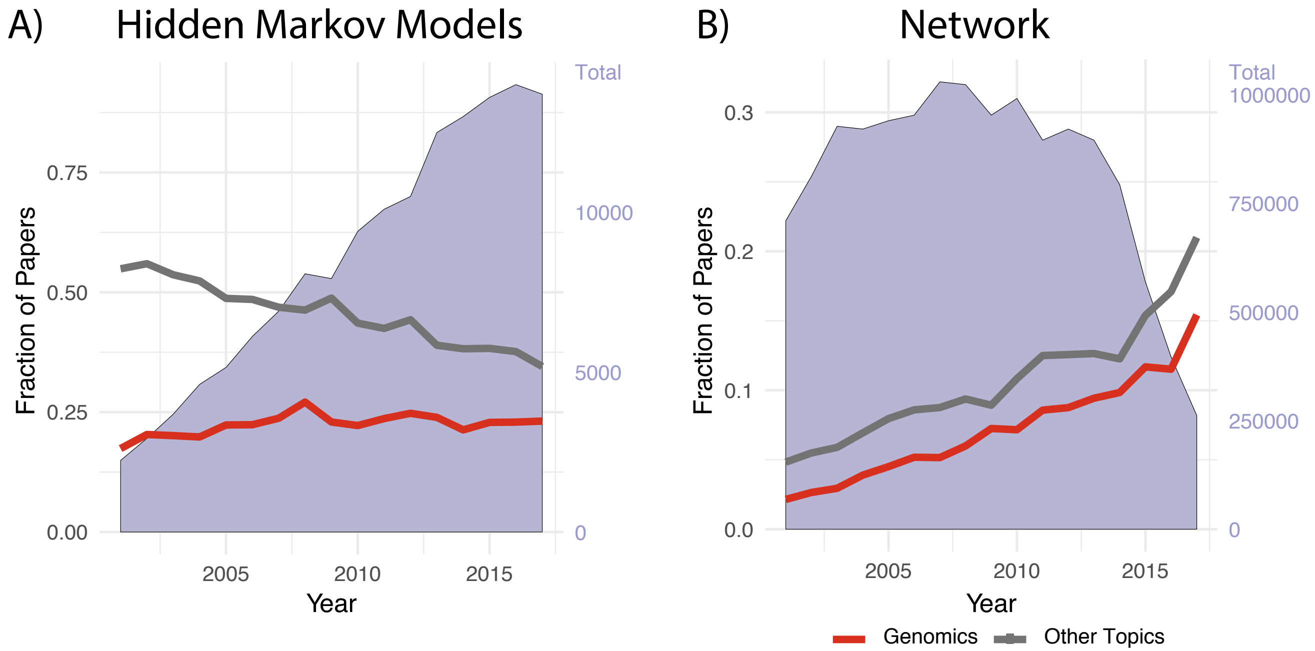
[Wang et al., TIG ('16)]

Quantifying Imports & exports from/to Genomics & other Data Science subfields



Genomics adoption of open source

Genomics contribution of HMM and Network papers



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

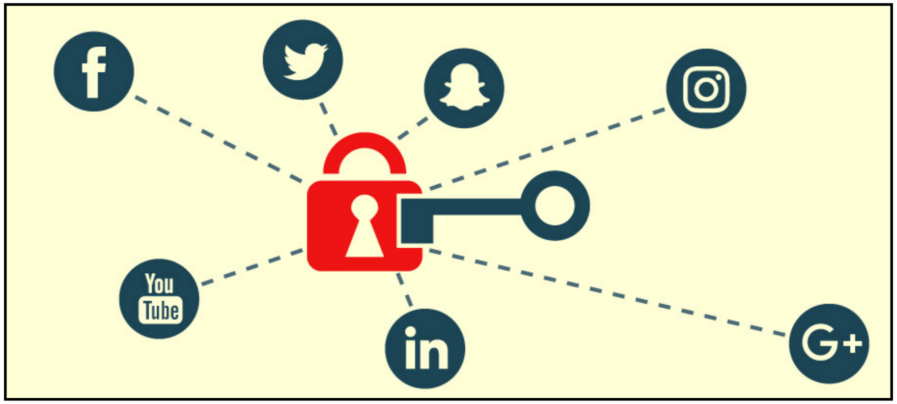
The Other Side of the Data Science Coin:

The Data Exhaust from Personal Genomics

(Privacy considerations)

Genomics has similar "Big Data" Dilemma as in the Rest of Society

- We confront privacy risks every day we access the internet (e.g., social media, e-commerce).
- Sharing & "peer-production" is central to success of many new ventures, with analogous risks to genomics
  - **EG web search**: Large-scale mining essential



Genetic Exceptionalism :

The Genome is very fundamental data, potentially very revealing about one's identity & characteristics

Personal Genomic info. essentially meaningless currently but will it be in 20 yrs? 50 yrs?

Genomic sequence very revealing about one's children. Is true consent possible?

Once put on the web it can't be taken back

Ethically challenged history of genetics

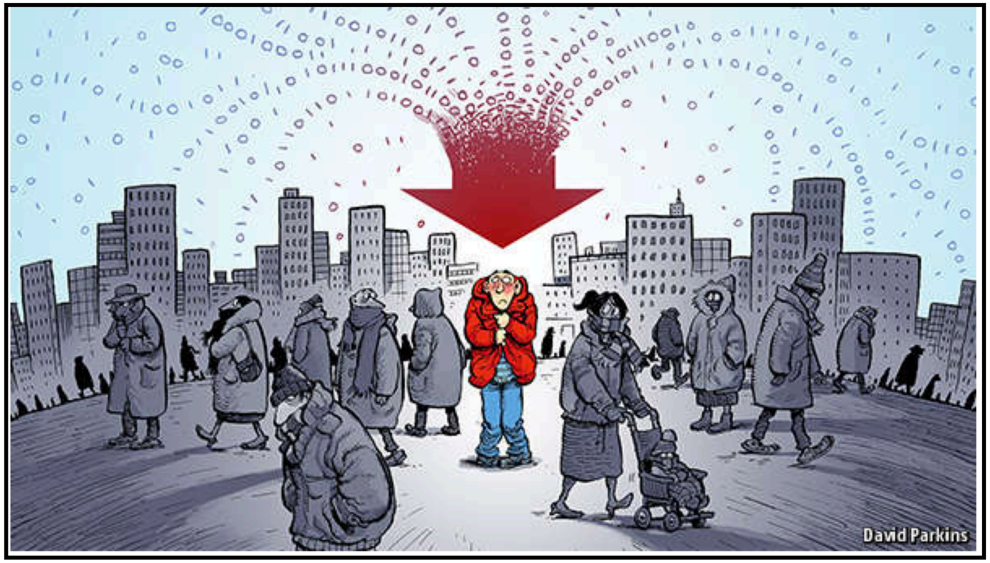
Ownership of the data & what consent means (Hela)

Could your genetic data give rise to a product line?

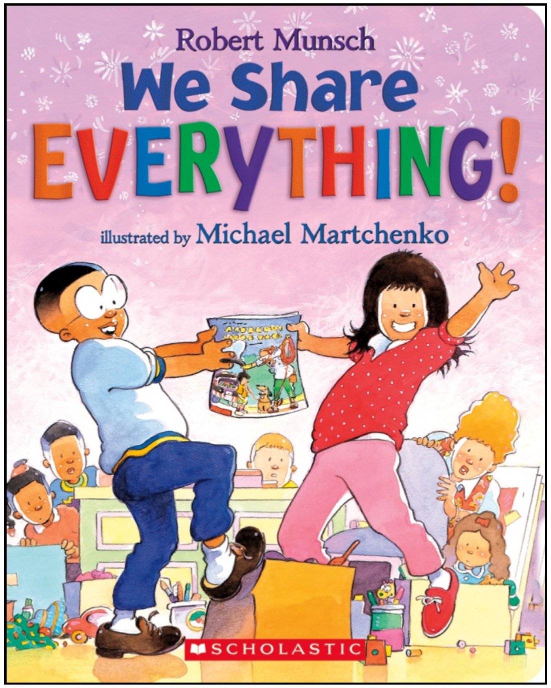
[Seringhaus & Gerstein ('09), *Hart. Courant* (Jun 5); Greenbaum & Gerstein ('11), *NY Times* (6 Oct), D Greenbaum & M Gerstein ('08). Am J. Bioethics; D Greenbaum & M Gerstein, Hartford Courant, 10 Jul. '08 ; SF Chronicle, 2 Nov. '08; Greenbaum et al. *PLOS CB* ('11) ; Greenbaum & Gerstein ('13), The Scientist; Photos from NY Times, it.wisc.edu]

The Other Side of the Coin: Why we should share

- Sharing helps **speed research**
  - Large-scale mining of this information is important for medical research
  - Statistical power
  - Privacy is cumbersome, particularly for big data



[Economist, 15 Aug ‘15]



The Dilemma

- The individual (harmed?) v the collective (benefits)
  - But do sick patients care about their privacy?
- How to balance risks v rewards
  - Quantification

[Yale Law Roundtable (‘10). Comp. in Sci. & Eng. 12:8; D Greenbaum & M Gerstein (‘09). Am. J. Bioethics; D Greenbaum & M Gerstein (‘10). SF Chronicle, May 2, Page E-4; Greenbaum et al. *PLOS CB* (‘11)]

Current Social & Technical Solutions: The quandary where are now

- Closed Data Approach
  - Consents
  - "Protected" distribution via dbGAP
  - Local computes on secure computer
- Issues with Closed Data
  - Non-uniformity of consents & paperwork
    - Different, confusing int'l norms
  - Computer security is burdensome
  - Many schemes get "hacked" .
  - **Tricky aspects of high-dimensional data** (ease of creating quasi-identifiers)

- Open Data
  - Genomic "test pilots" (ala PGP)?
    - Sports stars & celebrities?
  - Some public data & data donation is helpful but is this a realistic solution for an unbiased sample of ~1M

[Greenbuam et al ('04), Nat. Biotech; Greenbaum & Gerstein ('13), The Scientist; Photo: futurism.com]

Extra stuff

Class Web Page

GersteinLab.org/courses/452

Assignment #0 Page

http://bit.ly/cbb752b20-hw0

Plagiarism

(QUOTED(!) from Yale College Dean’s Office Memo of 9-Sept-2019)

- Please tell your students, in your syllabi and orally:
  - to cite all sources in papers, including drafts, every time they use a source,
  - to place quotation marks around any cited or cut-and-pasted materials, in addition to footnoting or otherwise marking the source,
  - to mark each paraphrased source, as though it were quoted,
  - to list all sources consulted for the draft or paper in the closing materials, such as a bibliography or roster of sources consulted,
  - to follow the above rules even in STEM courses that require a search of research literature,
  - not to collaborate unless explicitly allowed,
  - not to submit another person's solution to a problem, including a solution found online,
  - not to submit or reuse material from the same paper in multiple courses unless both instructors provide written permission to do so.

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