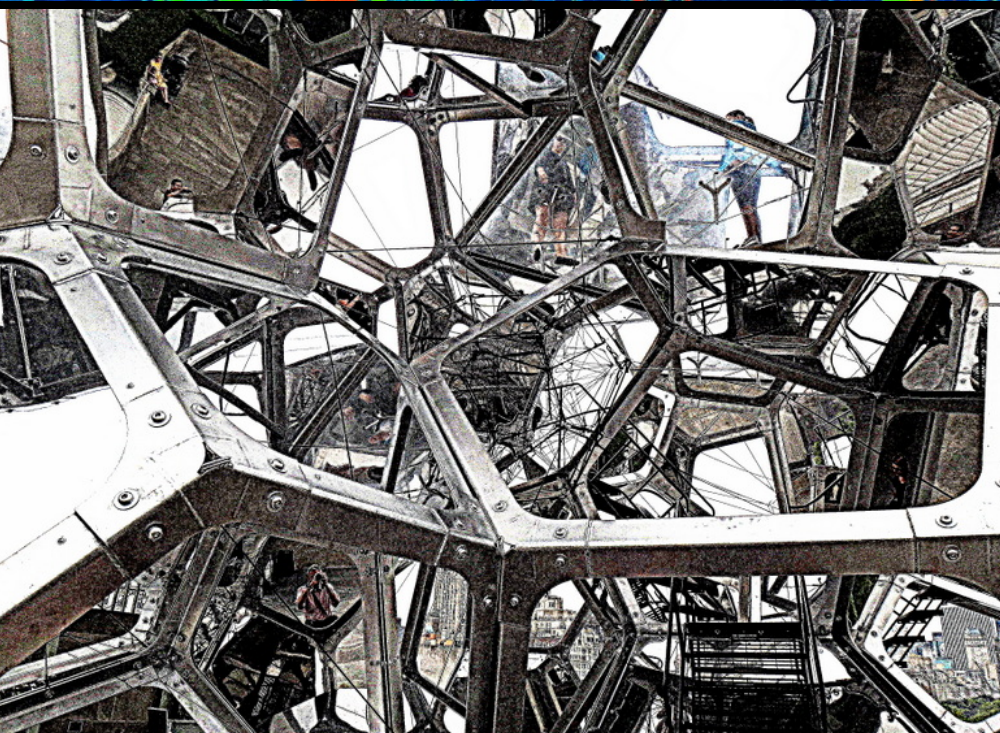


Biomedical Data Science: An Introduction



M Gerstein, Yale
(See last slide for more info.)
Credits & references on each slide

Slides freely downloadable
from **Lectures.GersteinLab.org**
& “tweetable” (via **@MarkGerstein**)

**Overview: what is
Biomed. Data science?**

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

Science Paradigms

#3 - Simulation

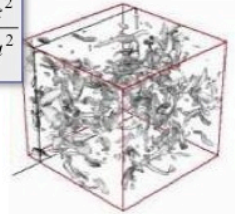
Prediction based on physical principles (eg Exact Determination of Rocket Trajectory)

Emphasis on: Supercomputers

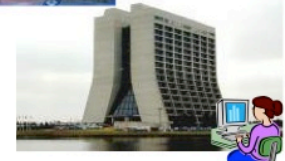
- Thousand years ago: science was **empirical** describing natural phenomena
- Last few hundred years: **theoretical** branch using models, generalizations
- Last few decades: a **computational** branch simulating complex phenomena



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



- 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



#4 - Data Science

Data gathering and storing

Data analysis including data mining, modeling, visualizing

Creative use of data exhaust and protection of privacy

Emphasis: networks, “federated” DBs

Gray died in '07.

Book about his ideas came out in '09.....



The
**FOURTH
PARADIGM**

DATA-INTENSIVE SCIENTIFIC DISCOVERY

EDITED BY TONY HILL, STEWART HANLEY, AND KRISTIN TOLL

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 “Big Tech”: a buzz-word for successful Ads



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 school yearbook.

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

Forbes · New Posts · Most Popular · Lists

108
349
193
353
12

CIO Network
INSIGHTS AND IDEAS FOR TECHNOLOGY LEADERS.
+ Follow (48)

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

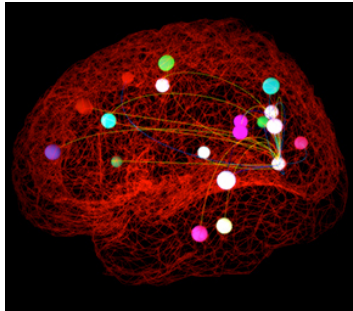
Guest post written by **Quentin Gallivan**
Quentin Gallivan is CEO of Pentaho Corp., an Orlando, Florida-based provider of business analytics software.

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



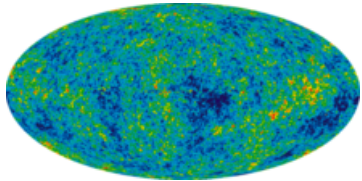
High energy physics - Large Hadron Collider



Neuroscience - The Human Connectome Project



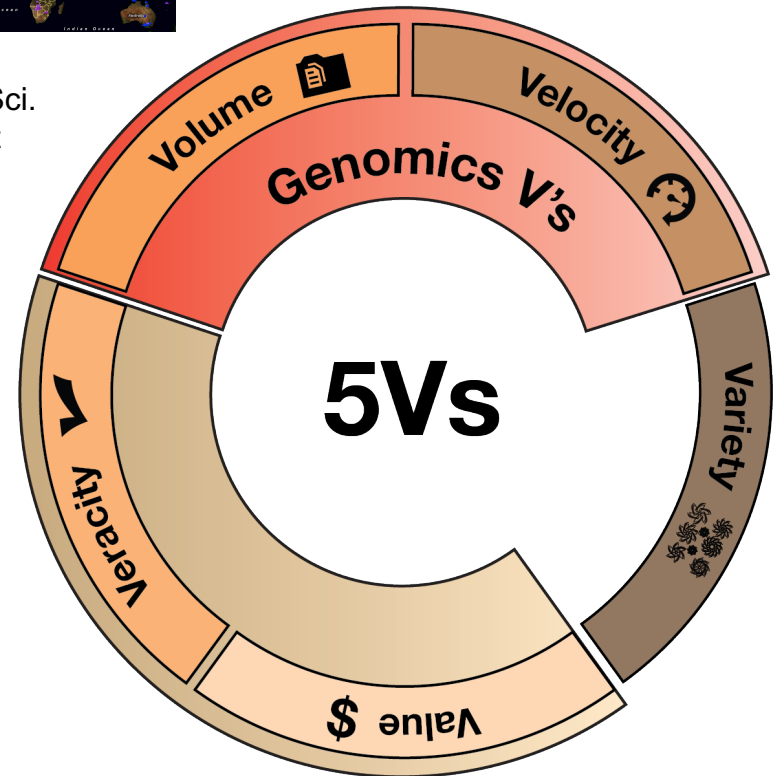
Ecology & Earth Sci. - Fluxnet



Astronomy - Sloan Digital Sky survey

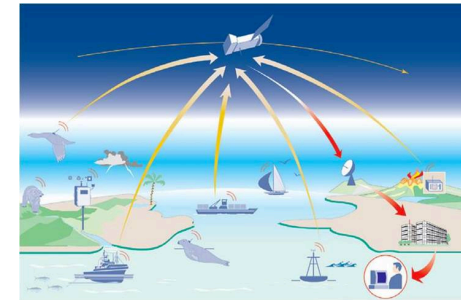
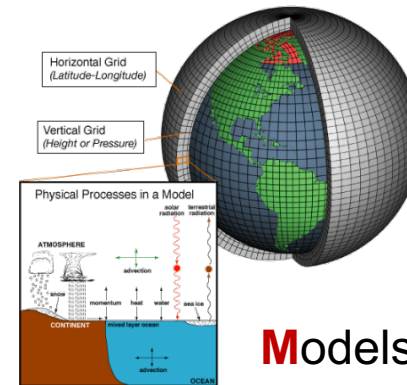
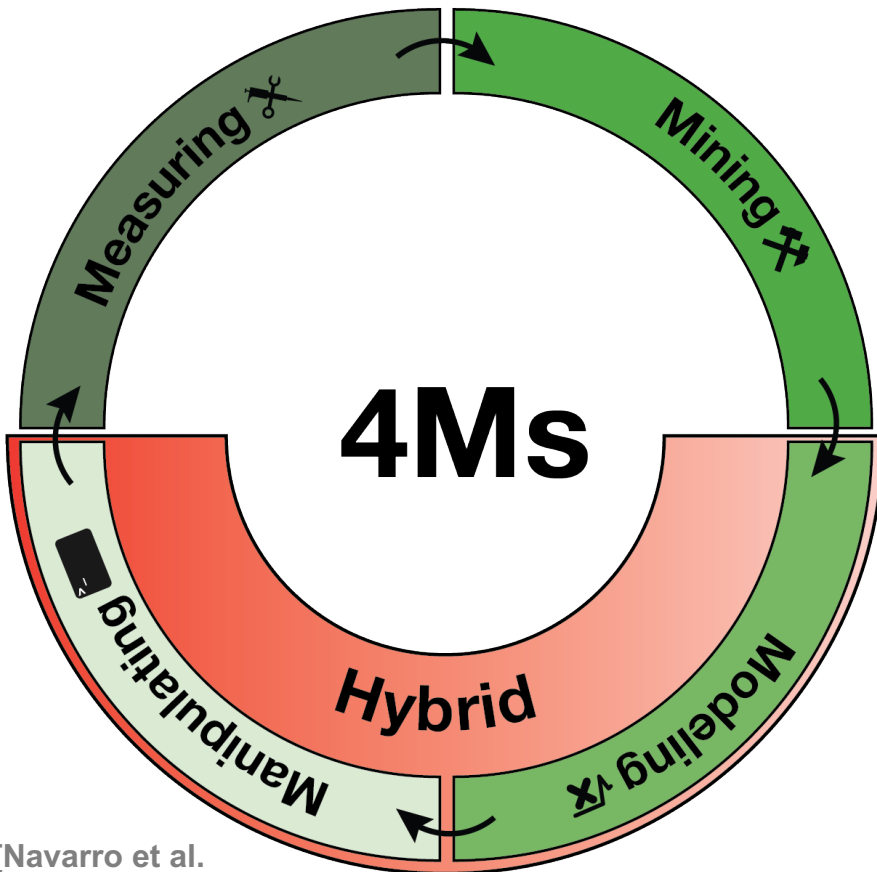


Genomics DNA sequencer



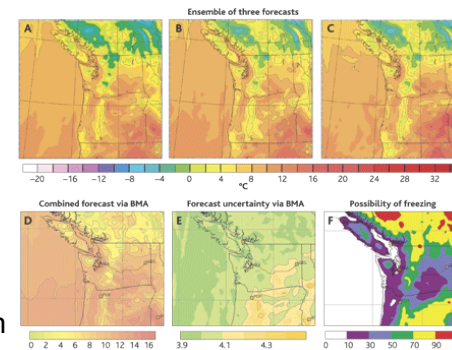
Coupling of Scientific Data to Models & Experiments

- 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



NOAA

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

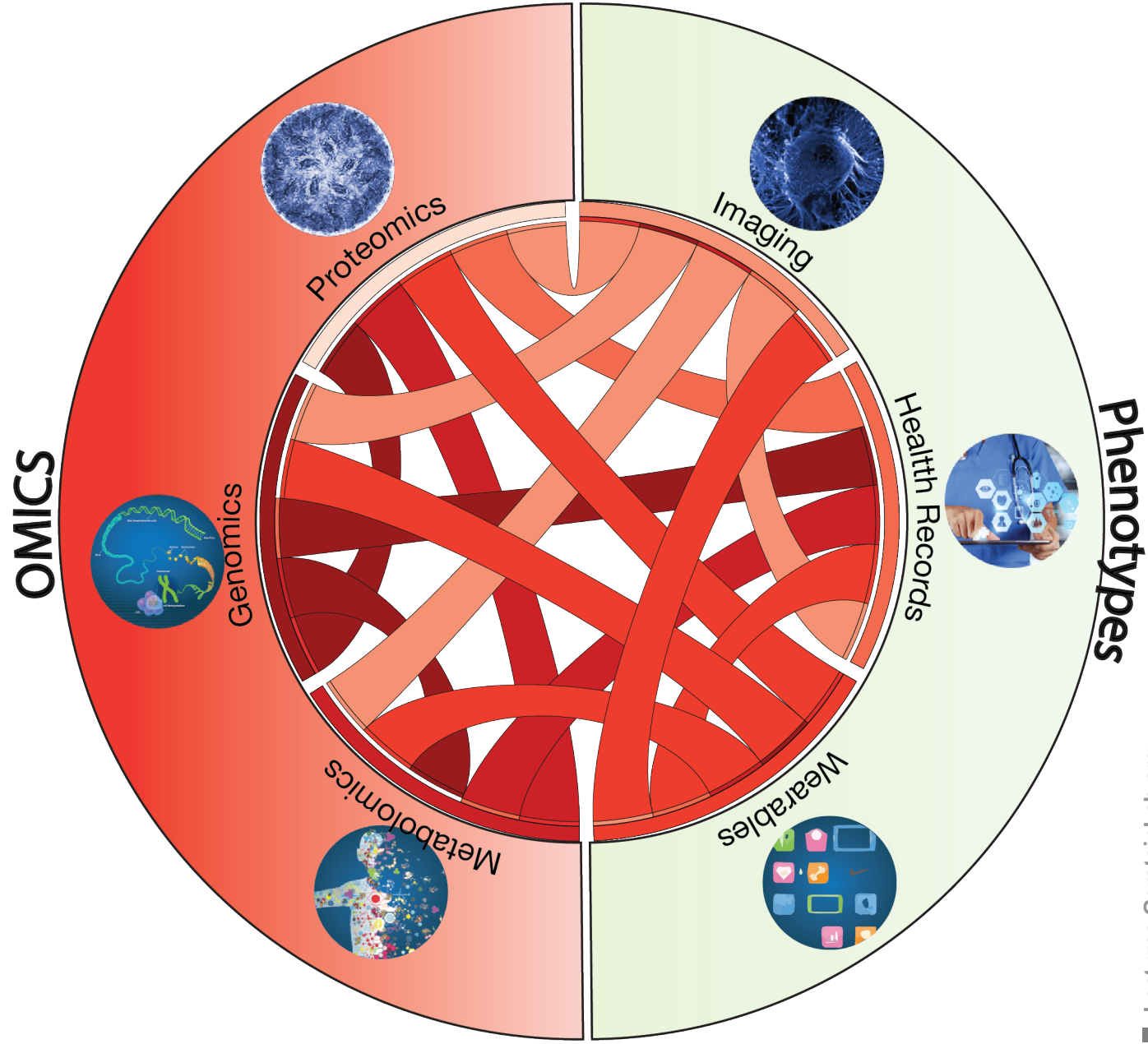
Image from <http://web.aibn.uq.edu.au/cssb/ResearchProjects.htm>

Biomed. Data science:

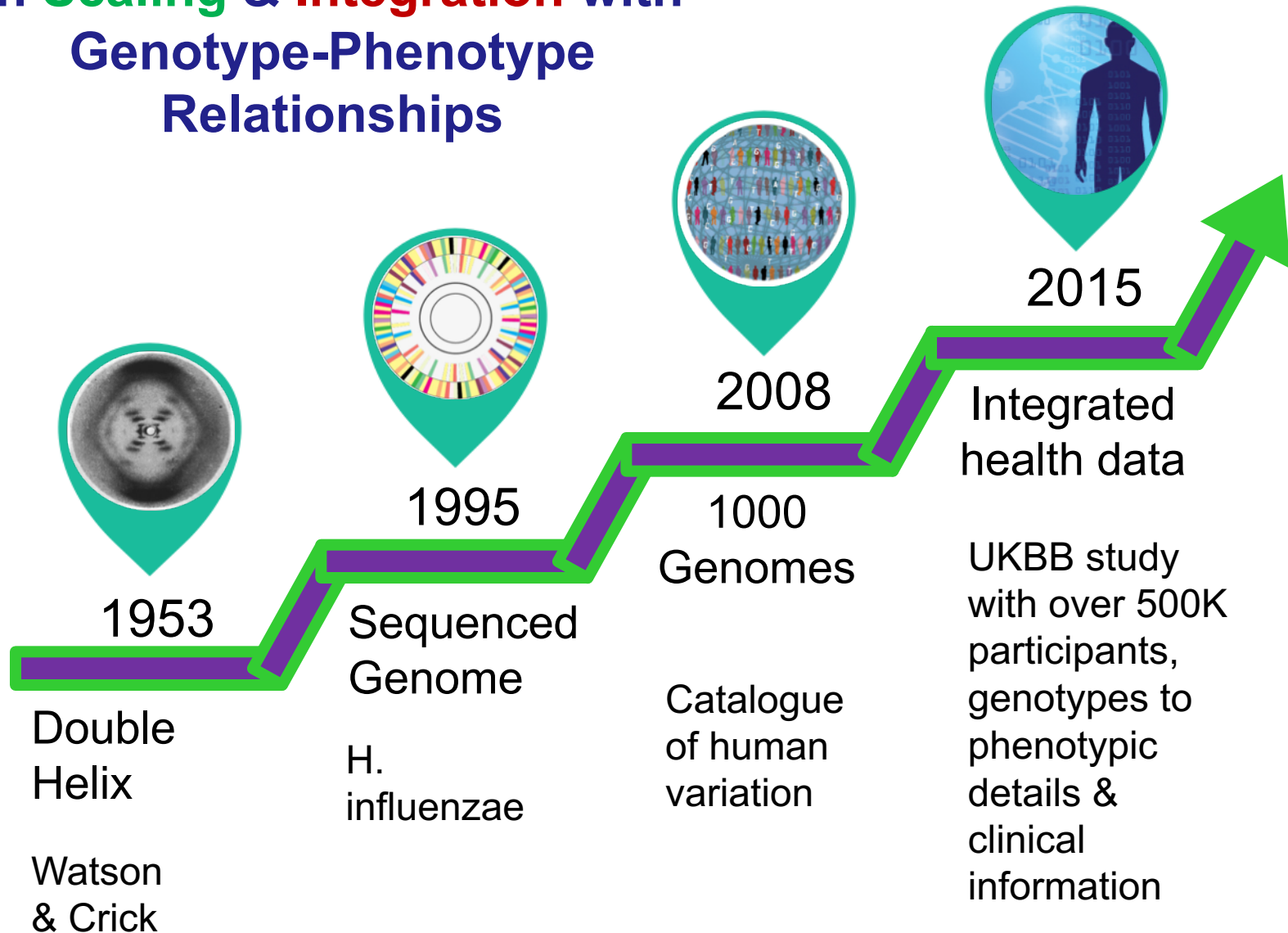
**Data Scaling
& Integration**

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

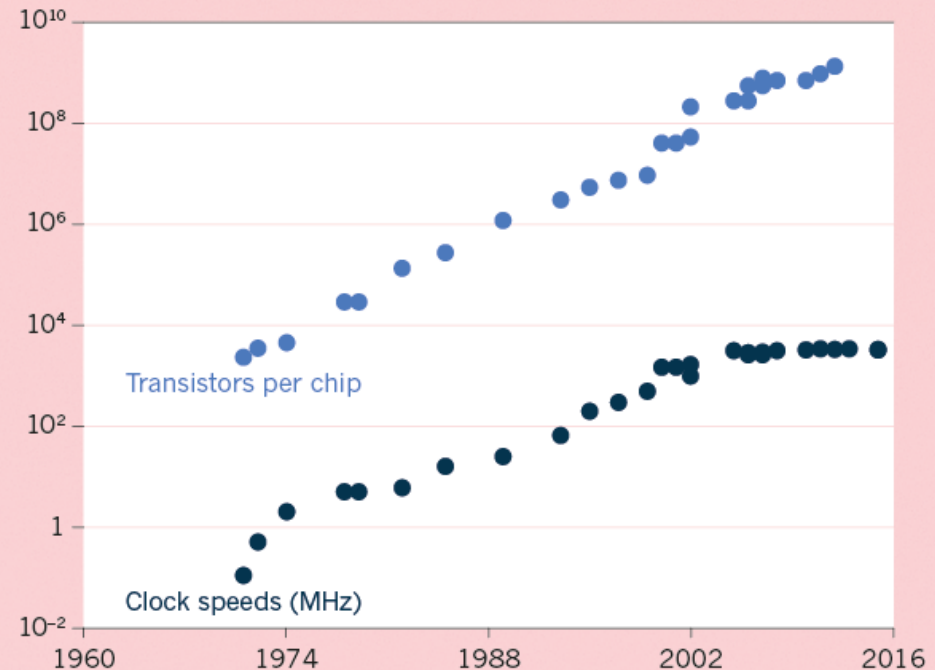
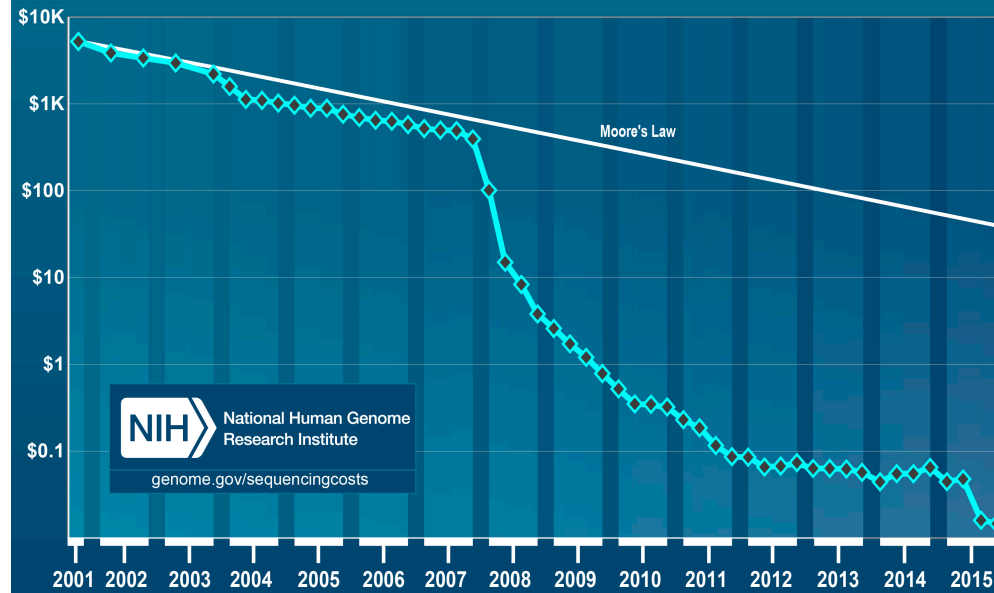


The **Scaling** of
Genomic Data
Science:

Powered by
exponential
increases in
data & computing

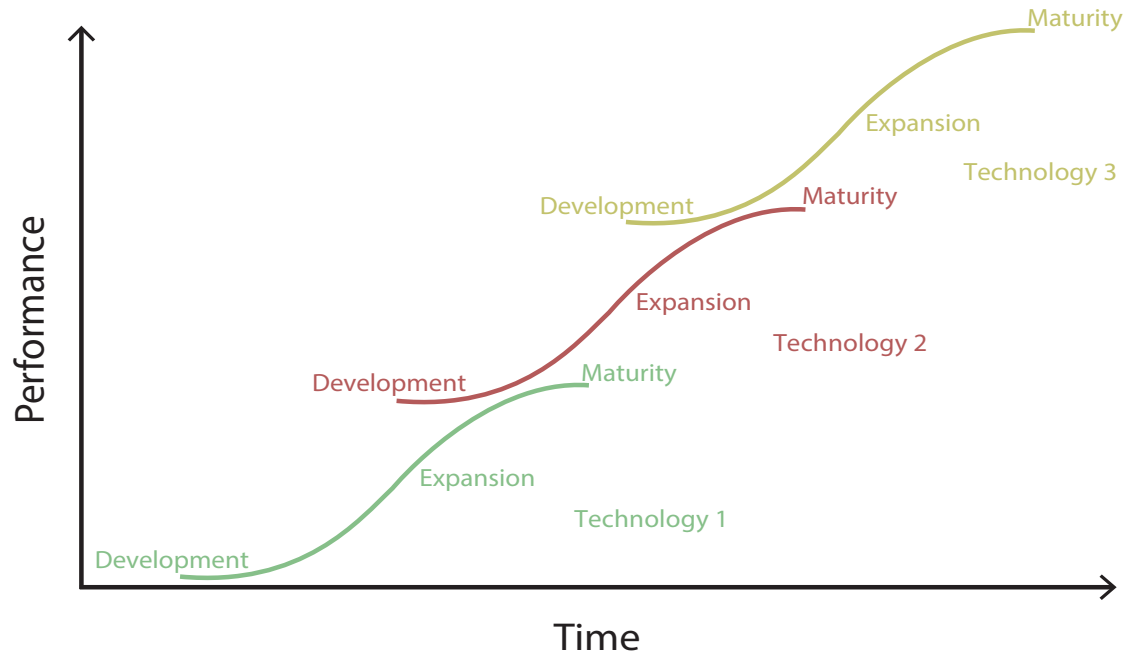
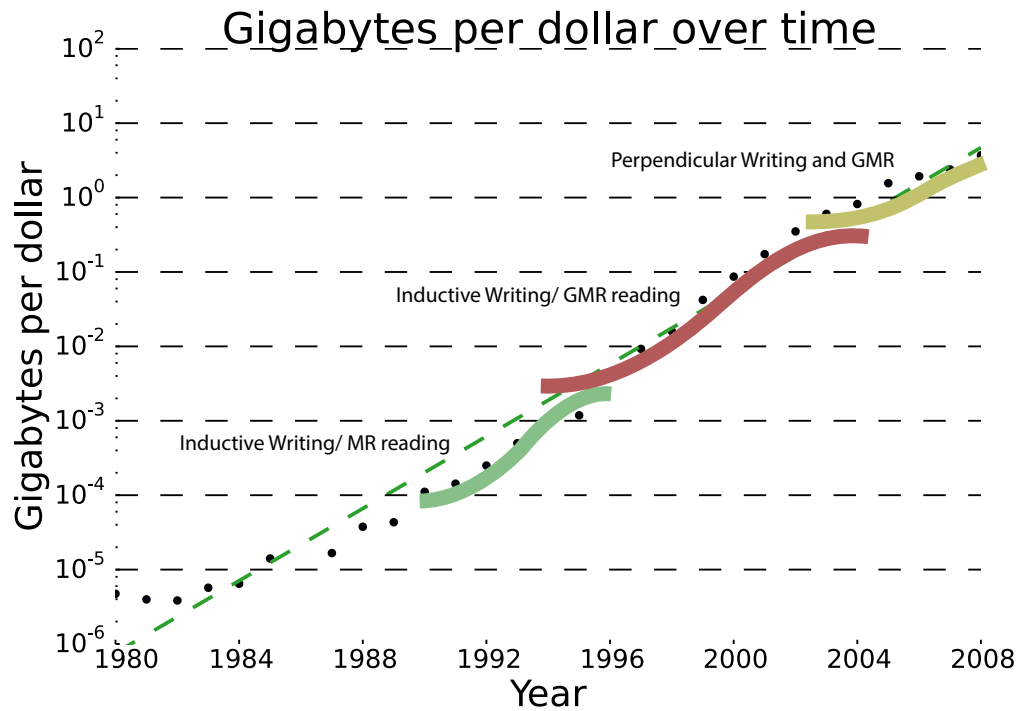
(**Moore's Law**)

Cost per Raw Megabase of DNA Sequence

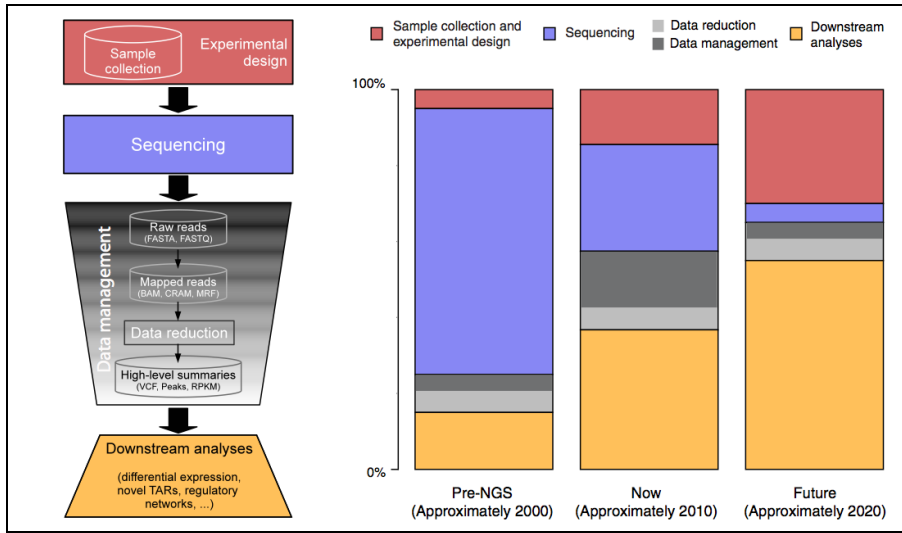


Kryder's Law and S-curves underlying exponential growth

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

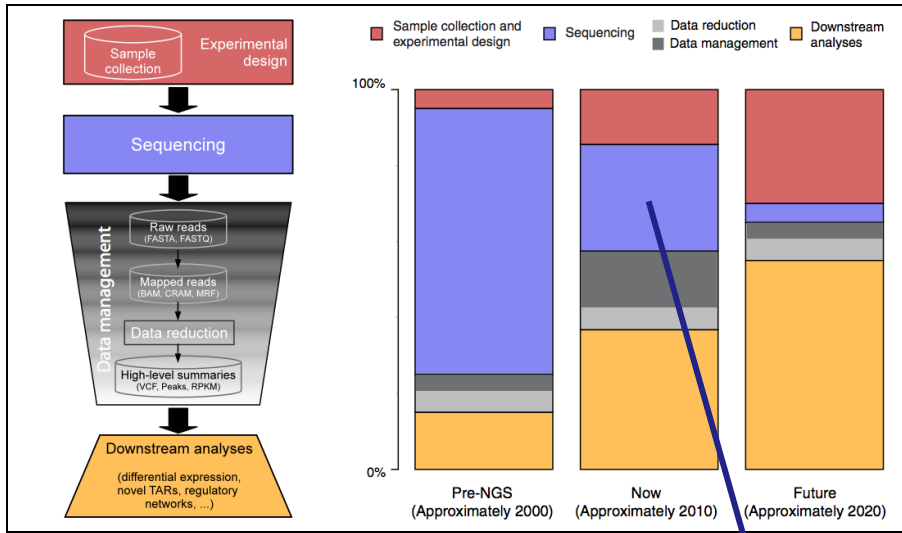


The changing costs of a sequencing pipeline

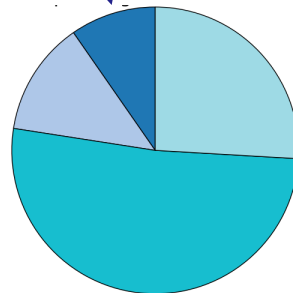
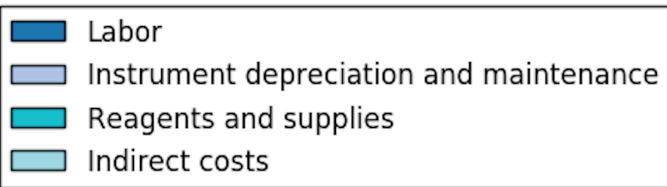


From '00 to ~' 20,
cost of DNA sequencing expt. shifts from
the actual seq. to sample
collection & analysis

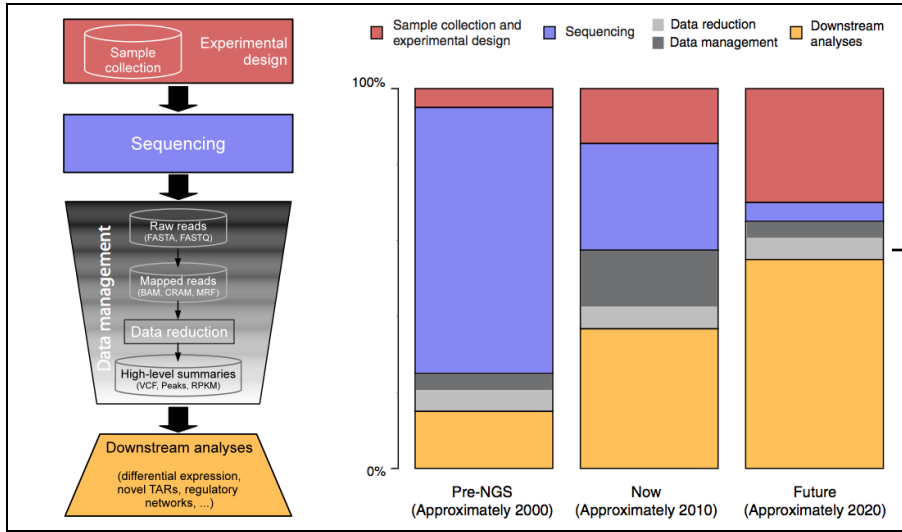
The changing costs of a sequencing pipeline



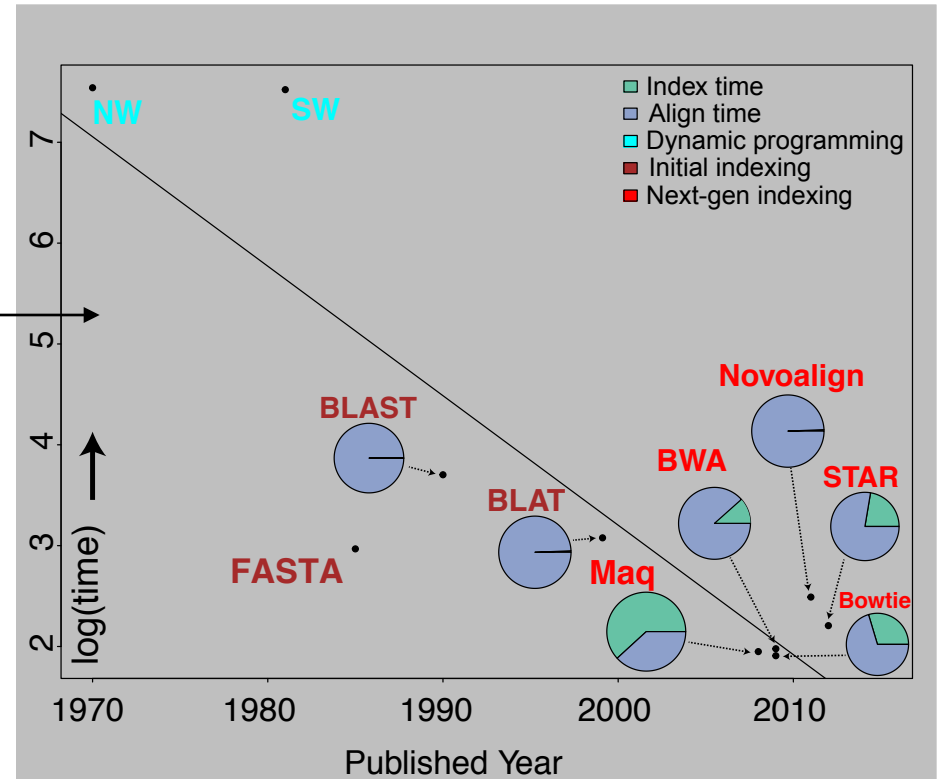
From '00 to ~' 20, cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis



The changing costs of a sequencing pipeline

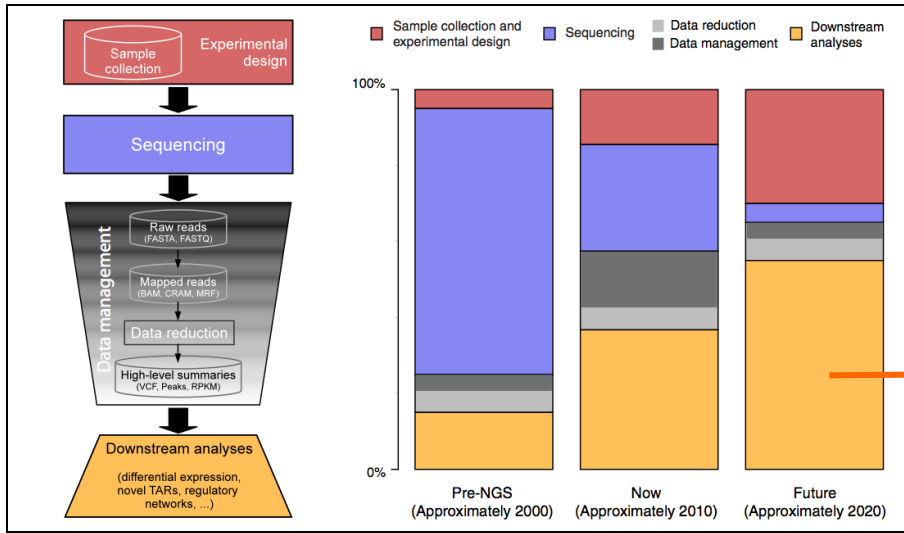


From '00 to ~' 20, cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

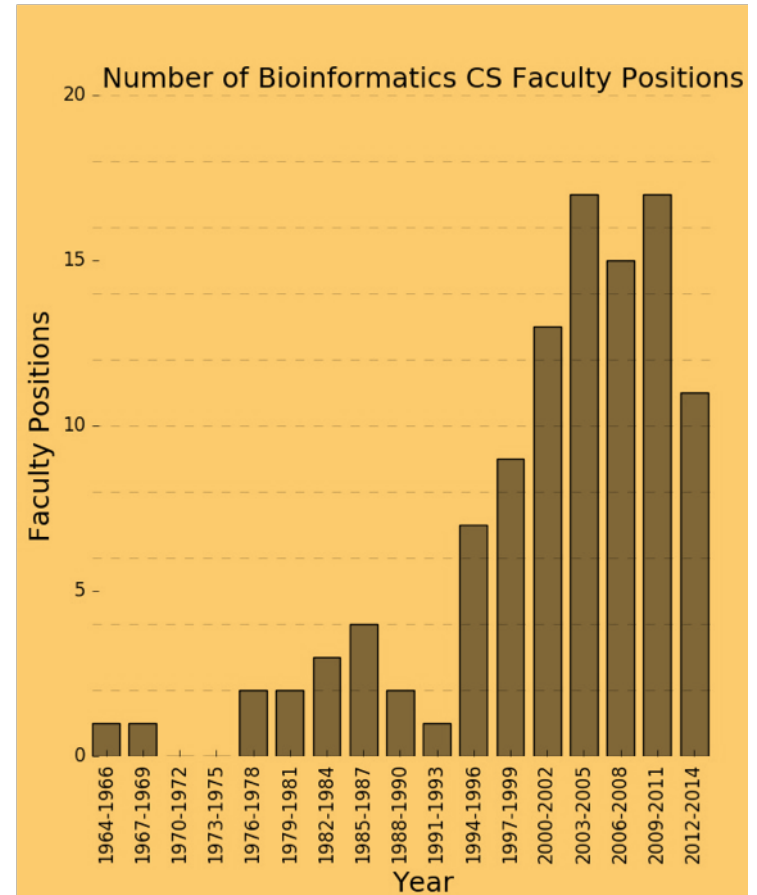


Alignment algorithms scaling to keep pace with data generation

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

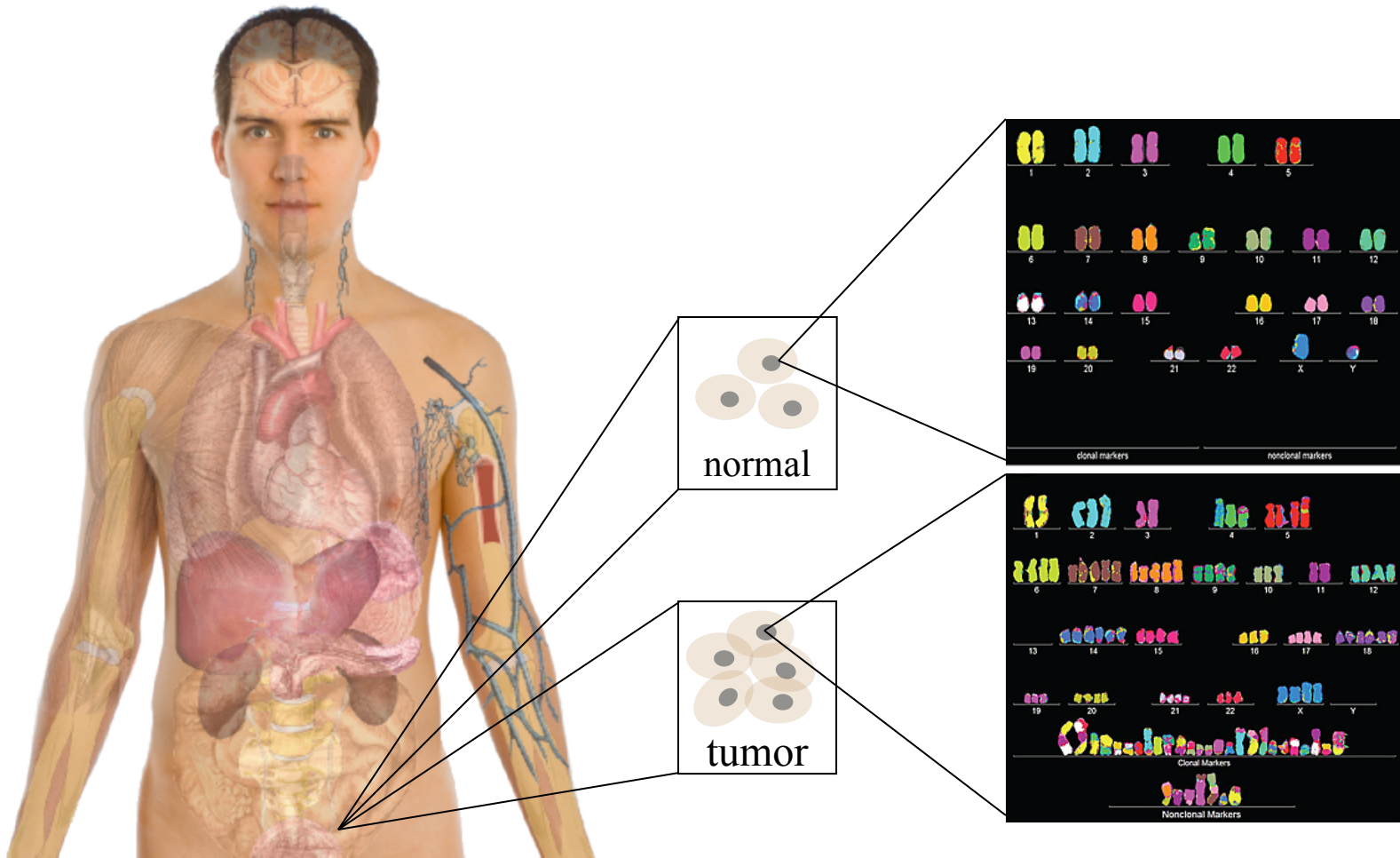


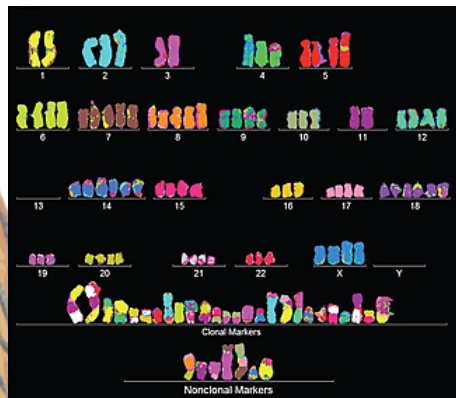
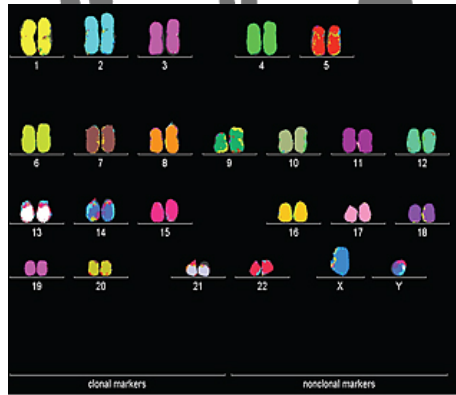
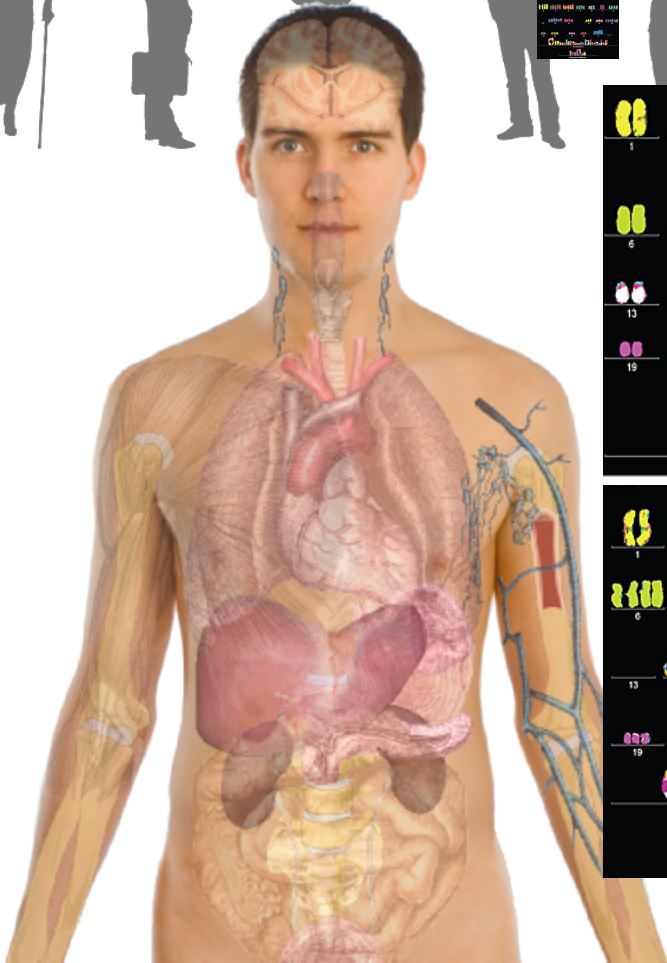
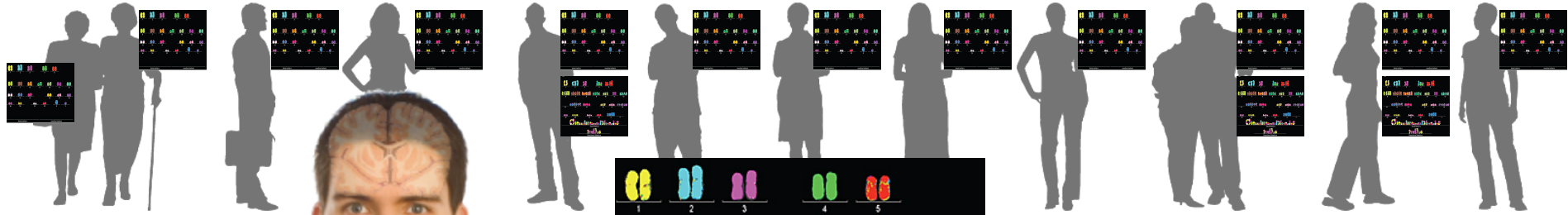
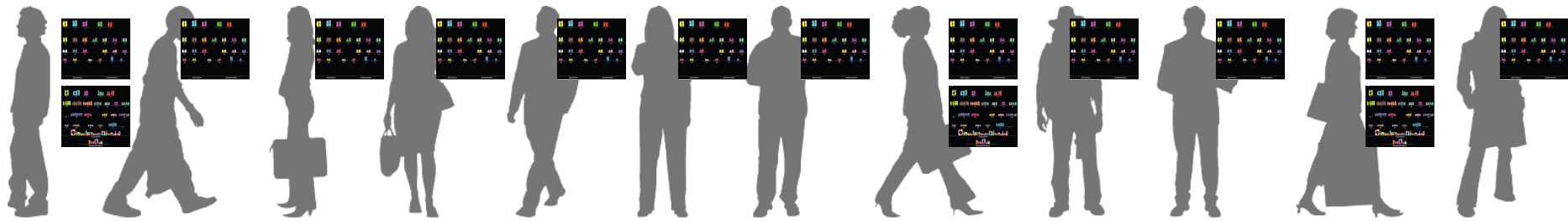
Biomed. Data science:

The Future

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

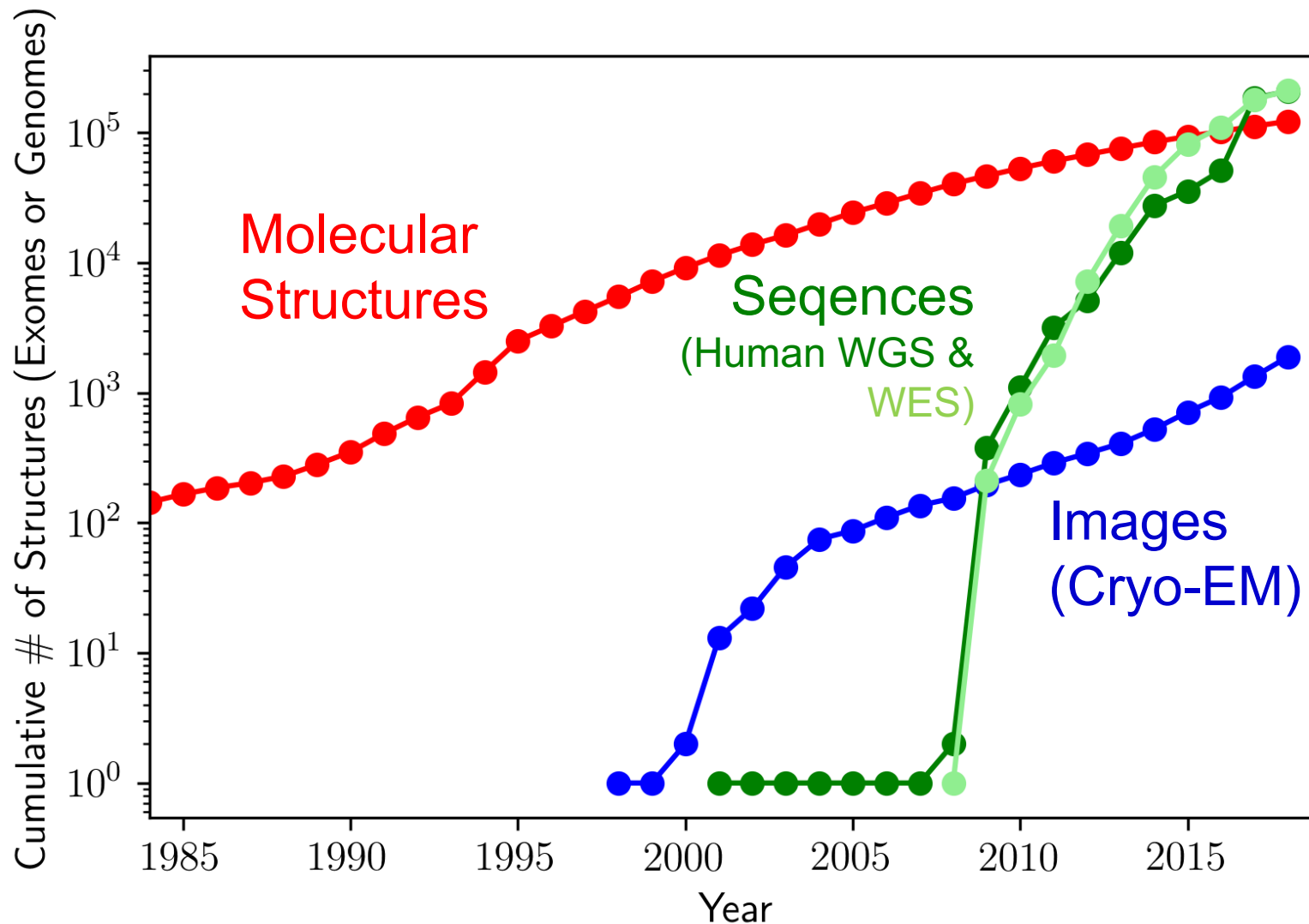
Personal genomes soon will become a commonplace part of medical research & eventually treatment (esp. for cancer). They will provide a primary connection for biological science to the general public.





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

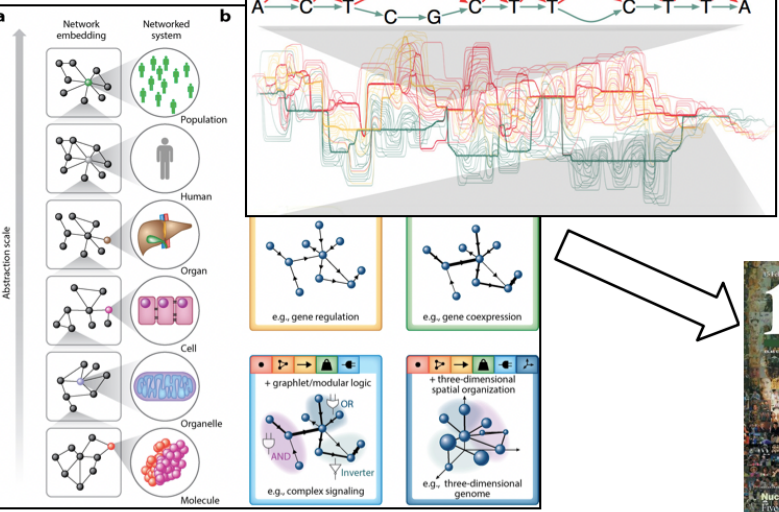


Examples of Imports & Exports to/from Genomics & Other Data

Science Application Areas

Technical Imports

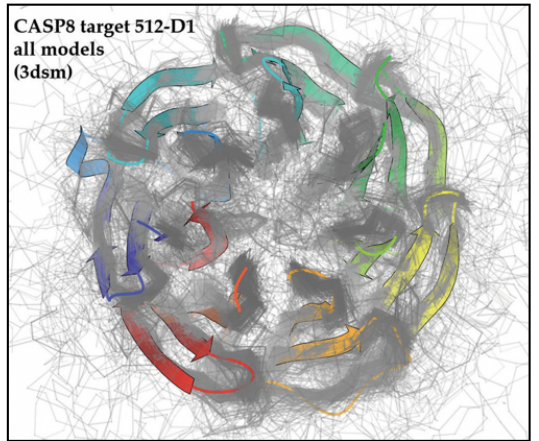
Networks and graphs



Importing tech. developed in other big data disciplines

Cultural Imports

CASP

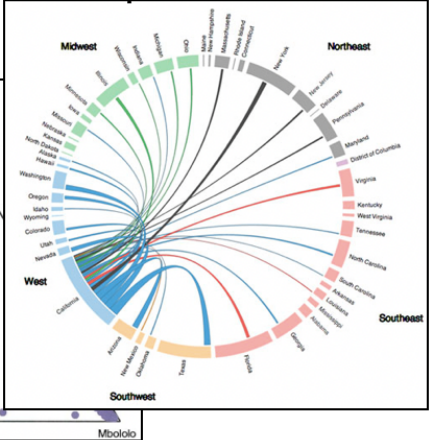


Cultural Exports

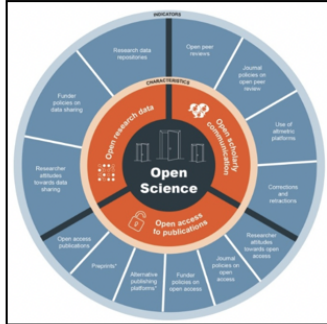


Technical Exports

Circos plot



Open Science



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

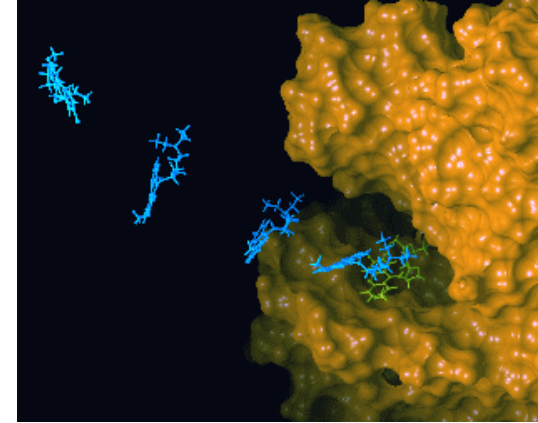
Biomedical Data Science

Key Practical Applications for Health

Major Bioinformatics Applications

I. Designing Drugs from Structural Targets

- Understanding how structures bind other molecules
- Designing inhibitors using docking, structure modeling



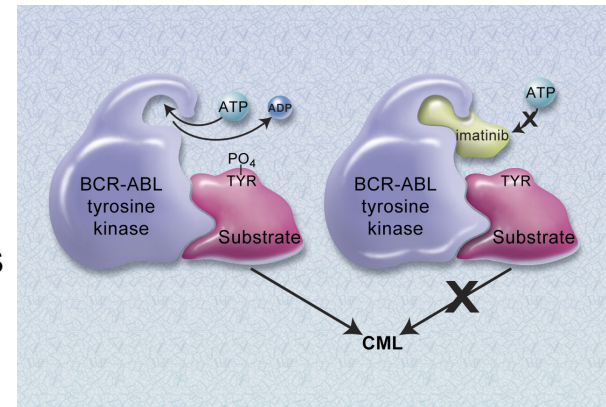
HUMAN	...	E	N	C	L	D	A	K	S	T	S	...
FLY (<i>D. melanogaster</i>)	...	E	N	S	L	D	A	Q	S	T	H	...
WORM (<i>C. elegans</i>)	...	E	N	S	L	D	A	G	A	T	E	...
YEAST (<i>S. cerevisiae</i>)	...	E	N	S	I	D	A	N	A	T	M	...
BACTERIA (<i>E. coli</i>)	...	E	N	S	L	D	A	G	A	T	R	...

II. Finding Homologs

- Model structures based on currently available structures
- Find experimentally tractable gene targets

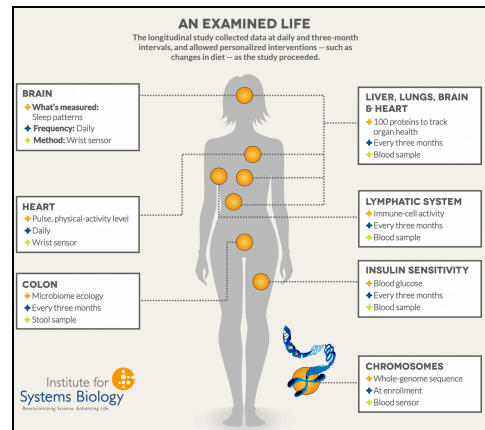
III. Customizing treatment in oncology

- Identifying key disease causing mutations
- Cancer immunotherapies targeting neo-antigens



IV. Personal Genome Characterization

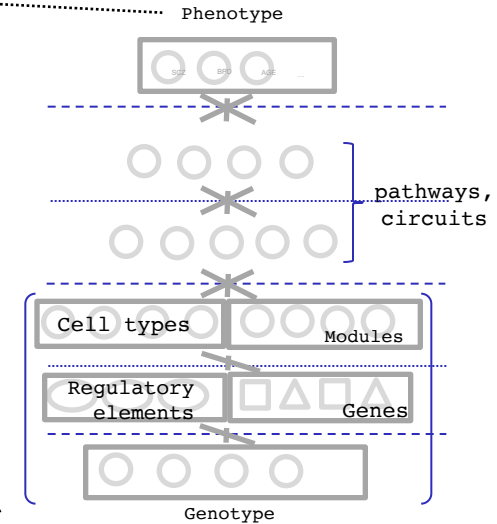
- Identify mutations in personal genomes (SNPs, SVs, &c)
- Integrate with digital phenotyping (eg wearables)



(From top to bottom: figures adapted from Olsen Group Docking Page at Scripps, *Sci. Am.*, Druker B.J. *Blood* 2008, Institute for Systems Biology)

Major Application V: Finding molecular mechanisms & drug targets for diseases we know little about (Neuropsychiatric Diseases)

Disease	Heritability*	Molecular Mechanisms
Schizophrenia	81%	C4A
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 other diseases such as cancer & heart disease)

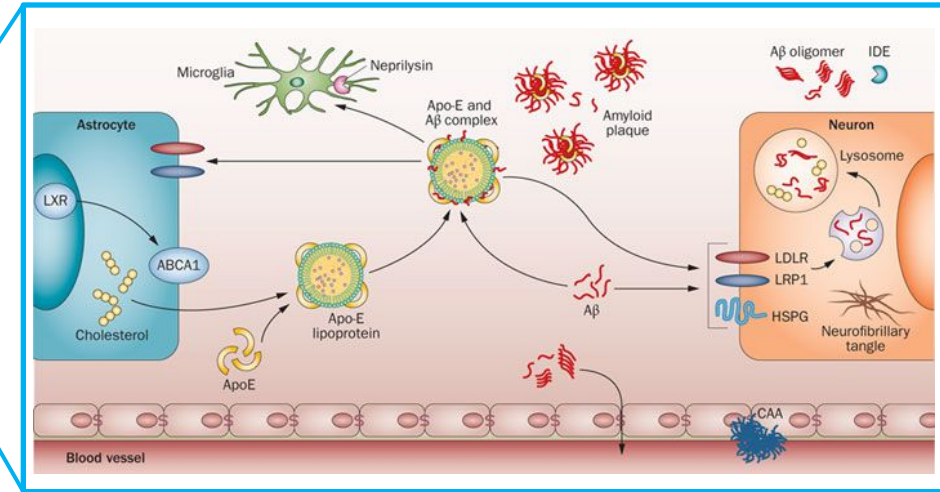
Thus, interested in developing predictive models of psychiatric traits which:

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

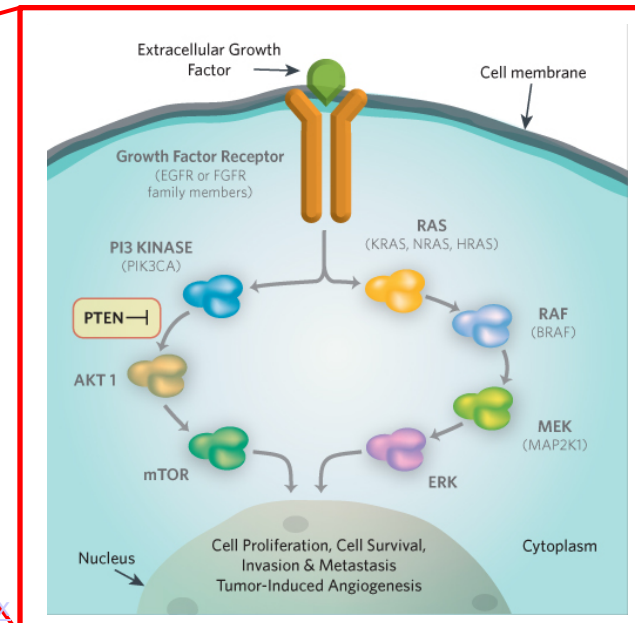
Major Application V: Finding molecular mechanisms & drug targets for diseases we know little about (Neuropsychiatric Diseases)

Disease	Heritability*	Molecular Mechanisms
Schizophrenia	81%	Complement Component 4A (C4A)
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



Major Application V: Finding molecular mechanisms & drug targets for diseases we know little about (Neuropsychiatric Diseases)

Disease	Heritability*	Molecular Mechanisms
Schizophrenia	81%	Complement Component 4A (C4A)
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



Major Application V: Finding molecular mechanisms & drug targets for diseases we know little about (Neuropsychiatric 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 other diseases)

Thus, interested in developing new treatments

Use observations of brain structure & function

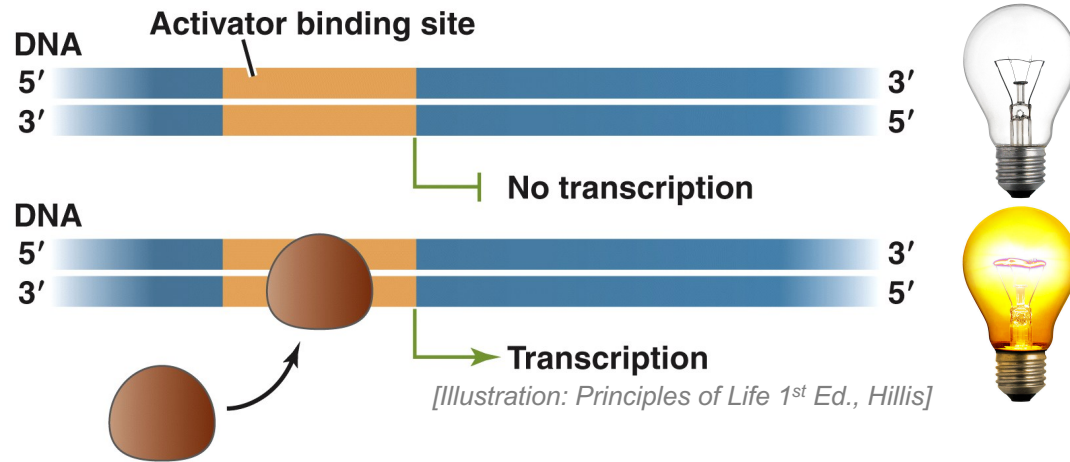
Use the predictive power of genetic data to suggest actors involved

Recent Rollout in Science addressing this, involving many Yale Researchers

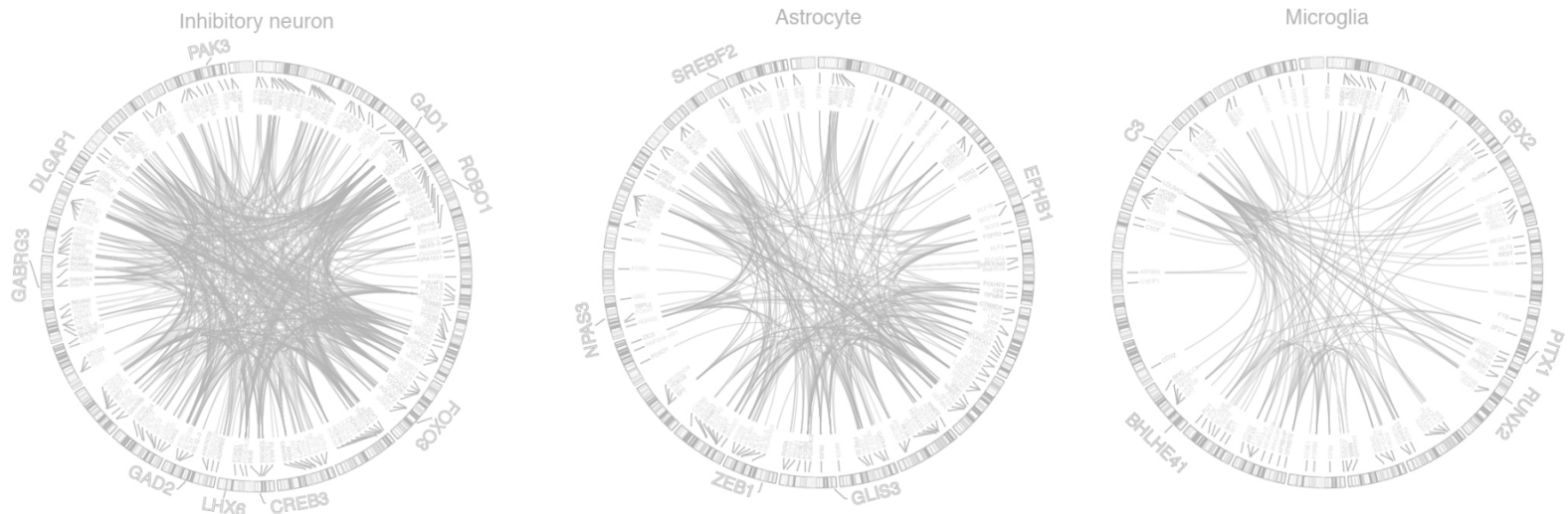
ch:
n latent
to begin to



Developing a gene regulatory network for the human brain

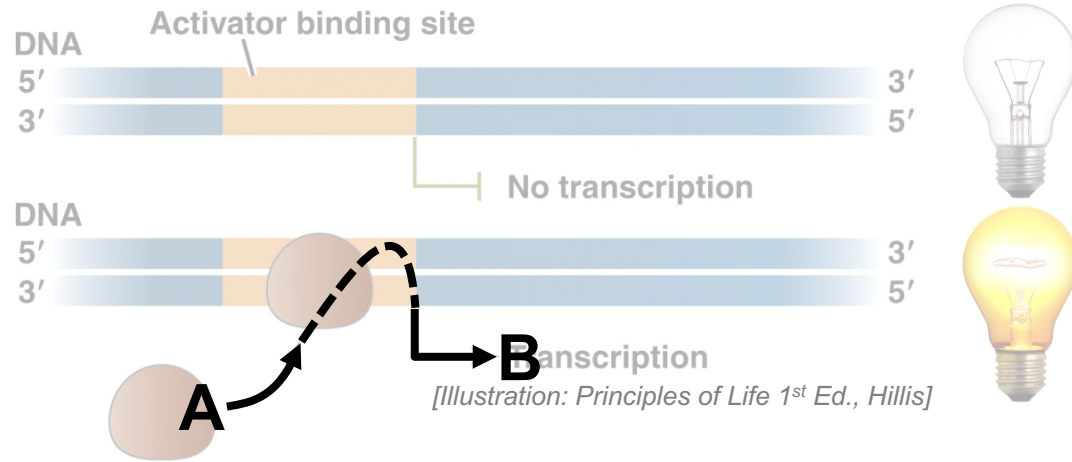


Many such gene regulatory relationships form a network

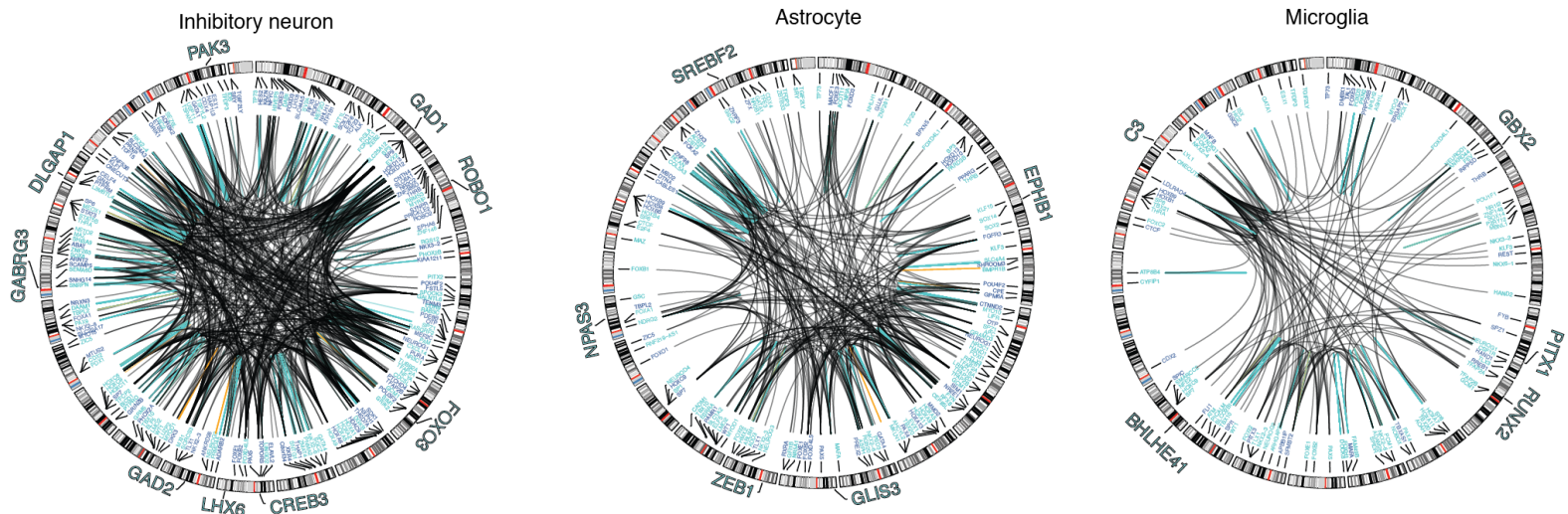


[Wang et al. ('18) Science]

Developing a gene regulatory network for the human brain



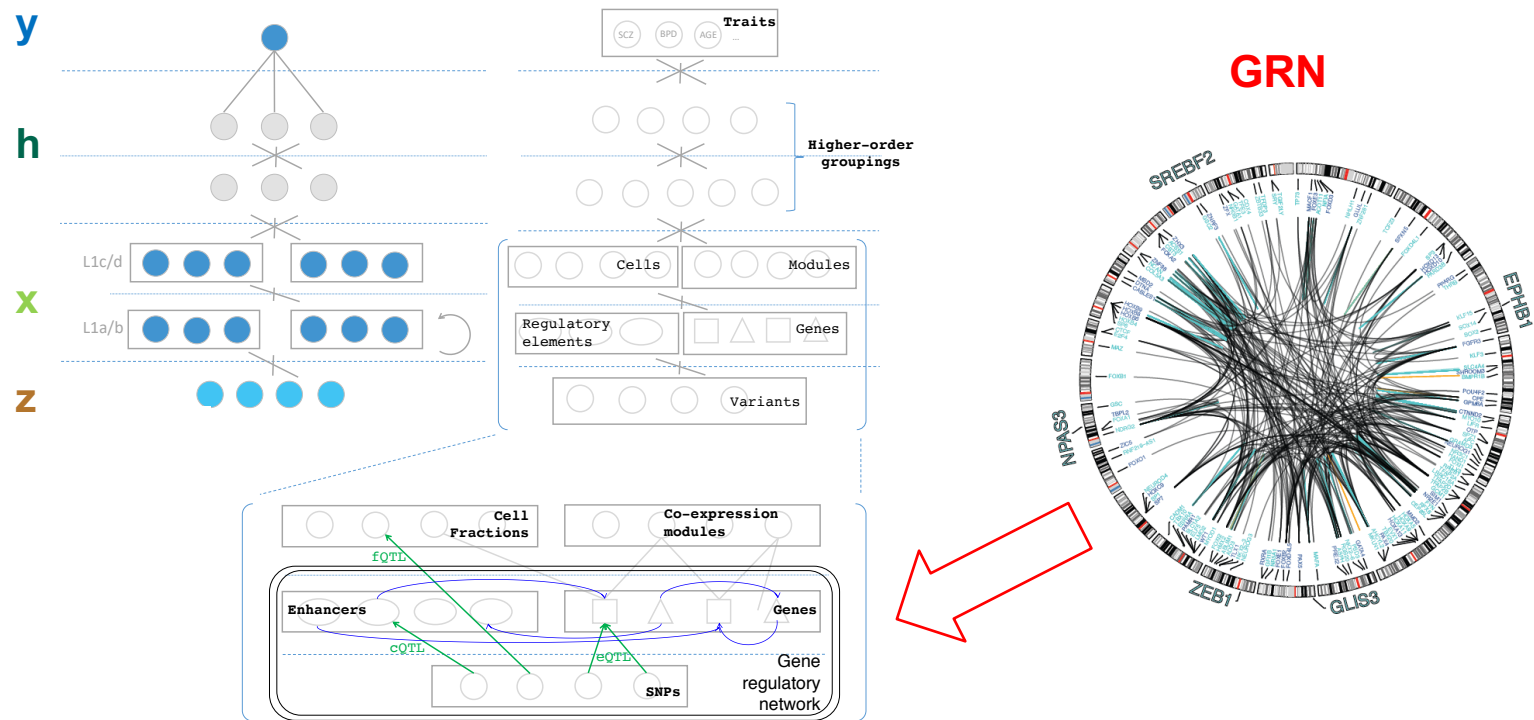
Many such gene regulatory relationships form a network



[Wang et al. ('18) Science]

Deep Structured Phenotype Network (DSPN)

- Embed **Gene Regulatory Network** in deep neural network
- Allows transcriptome (+other) imputation & trait prediction



y: phenotypes

x: intermediate phenotypes (e.g. expression, enhancers)

h: hidden units (e.g., circuits)

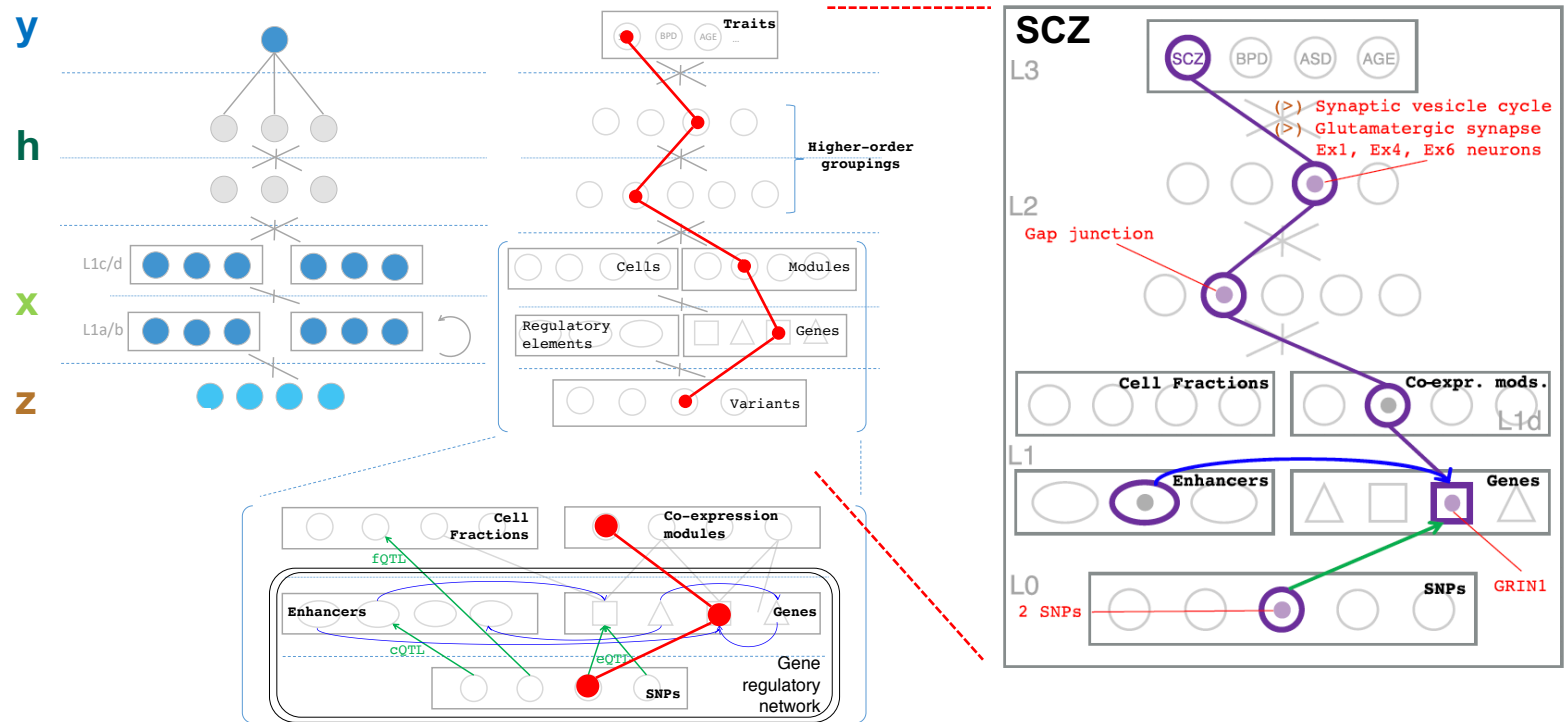
z: genotypes (e.g., SNPs)

Deep Boltzmann Machine Energy model:

$$p(\mathbf{x}, \mathbf{y}, \mathbf{h} | \mathbf{z}) \propto \exp(-E(\mathbf{x}, \mathbf{y}, \mathbf{h} | \mathbf{z}))$$

Deep Structured Phenotype Network (DSPN)

- Allows prioritization of genes / modules through network interpretation (using path tracing)



y: phenotypes

x: intermediate phenotypes (e.g. expression, enhancers)

h: hidden units (e.g., circuits)

z: genotypes (e.g., SNPs)

Deep Boltzmann Machine Energy model:

$$p(\mathbf{x}, \mathbf{y}, \mathbf{h} | \mathbf{z}) \propto \exp(-E(\mathbf{x}, \mathbf{y}, \mathbf{h} | \mathbf{z}))$$

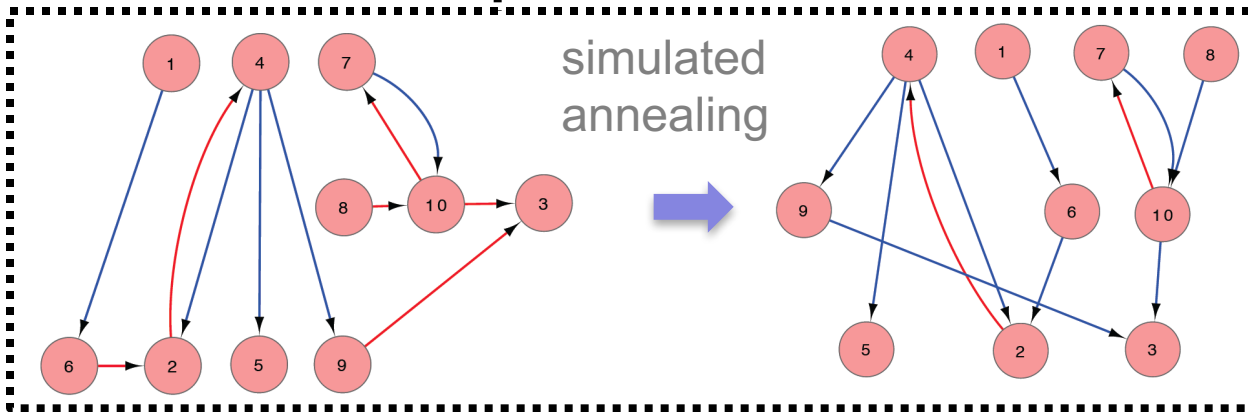
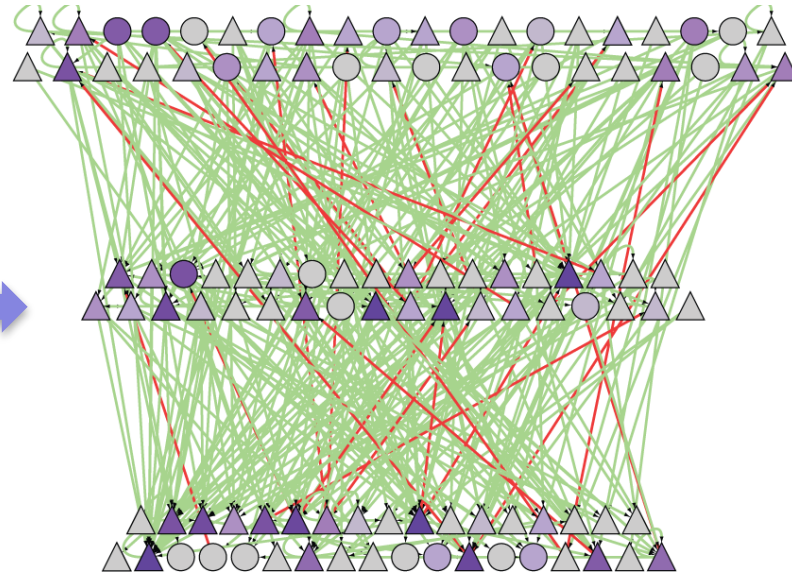
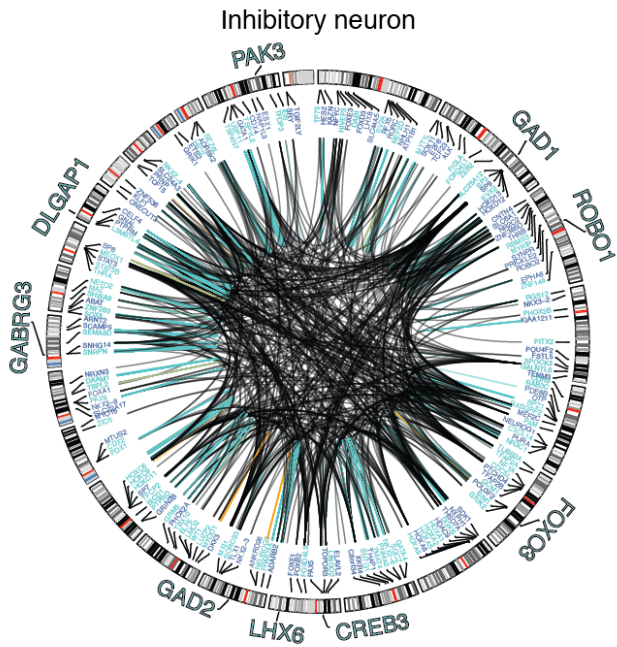
Biomedical Data Science

**General “Bioinspired”
Insights into Systems
Design**

Hierarchies in Networks

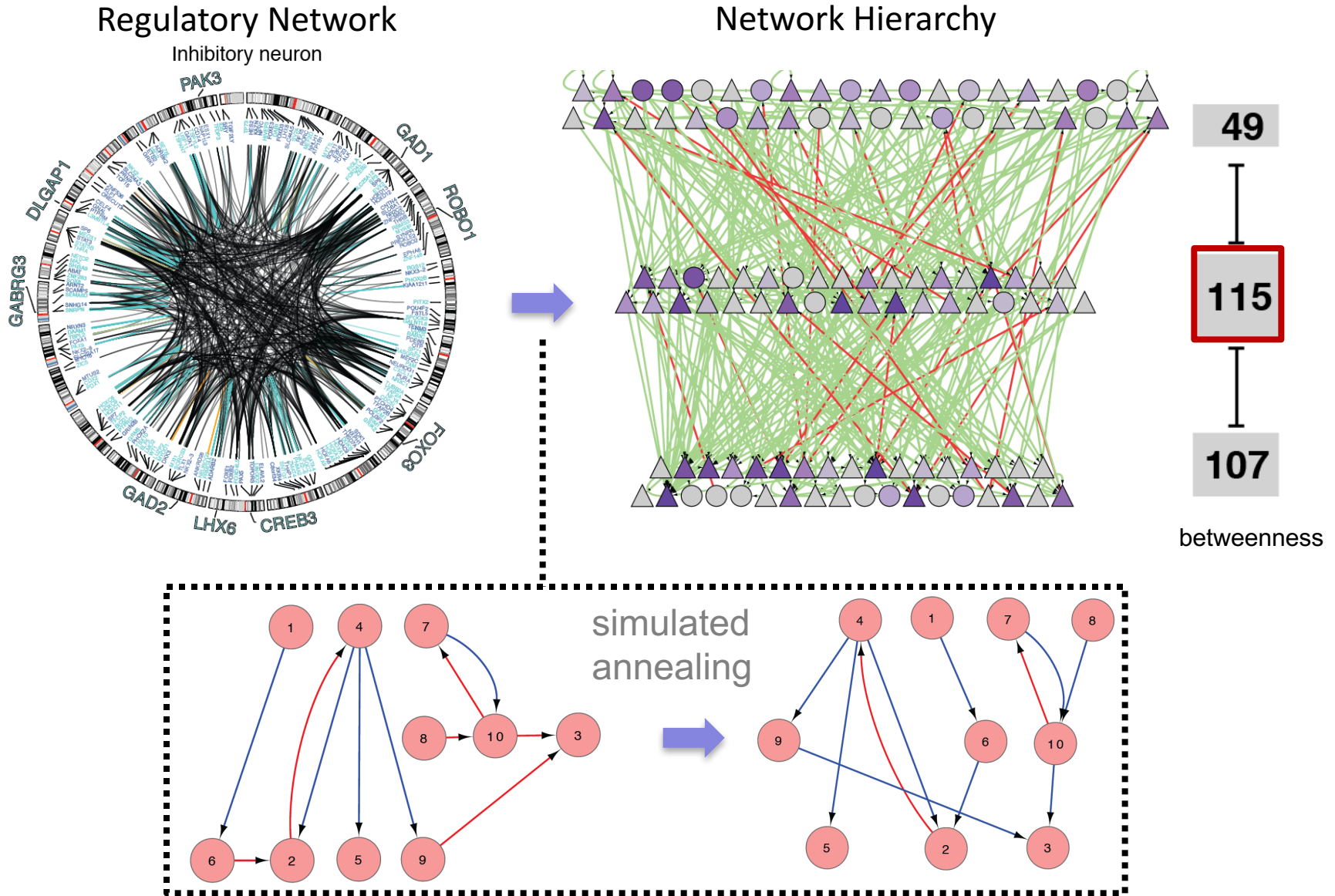
Regulatory Network

Network Hierarchy



[Wang et al. ('18) Science, Gerstein et al. Nature ('12)]

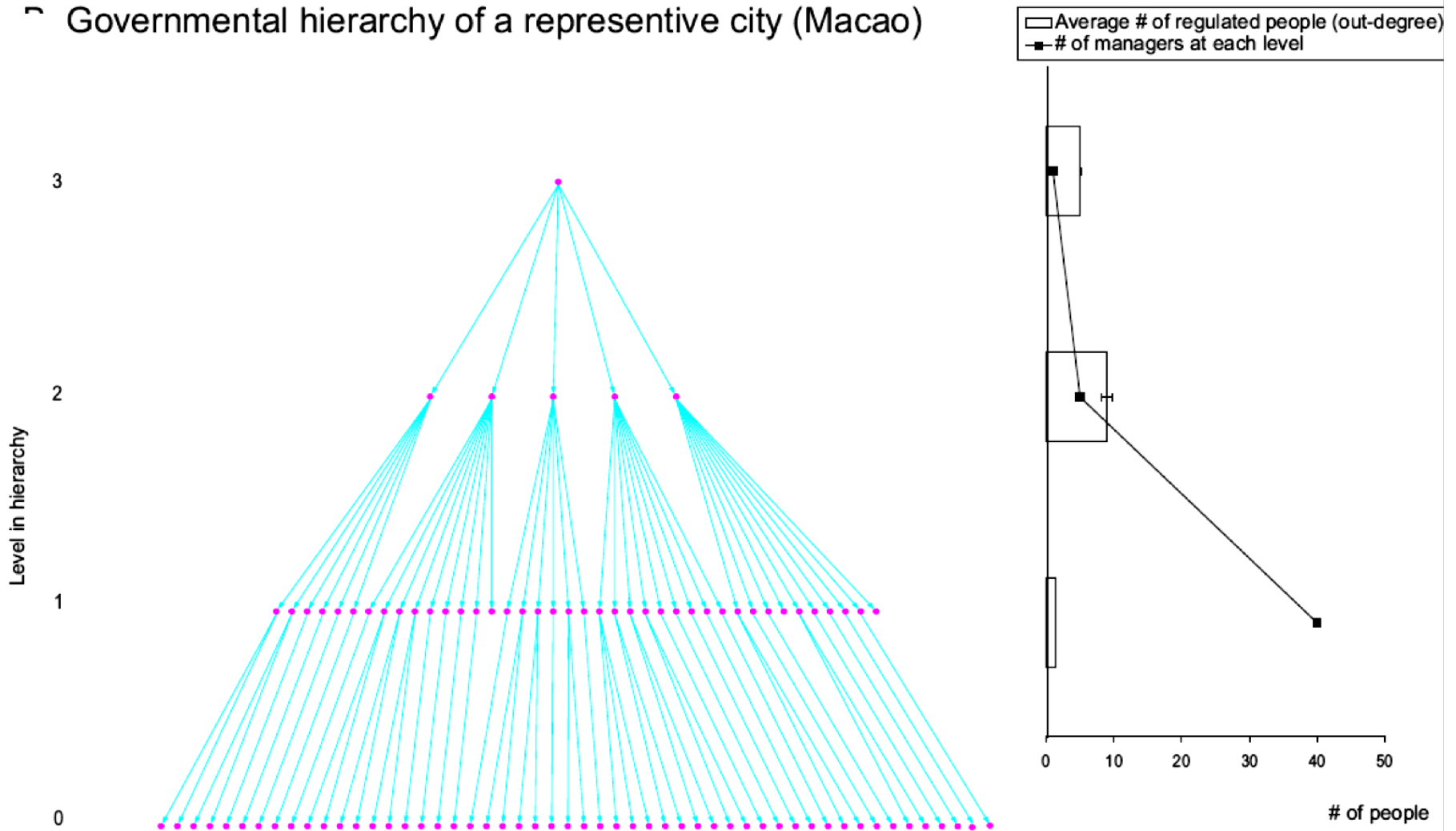
Hierarchies in Networks



[Wang et al. ('18) Science, Gerstein et al. Nature ('12)]

A government hierarchy and middle-managers

Governmental hierarchy of a representative city (Macao)

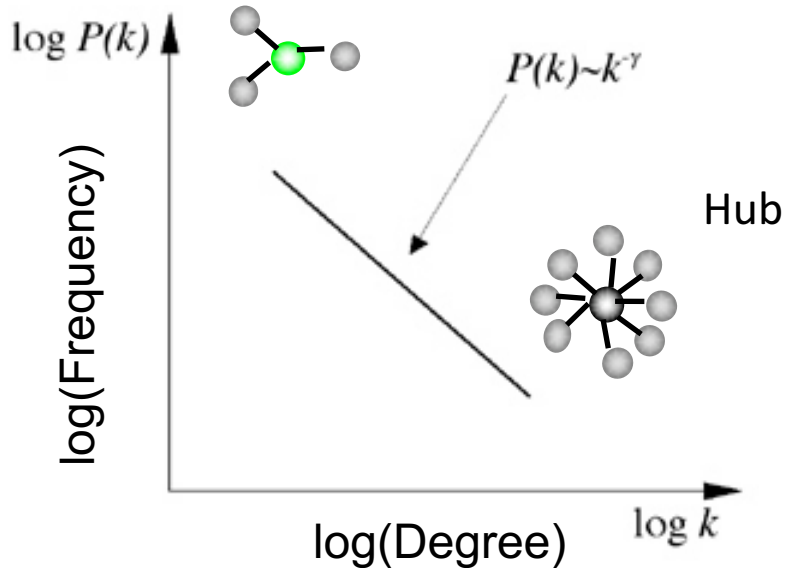


[Yu et al., PNAS ('06)]

[Yu et al., PNAS '06]

35 - Lectures.GersteinLab.org

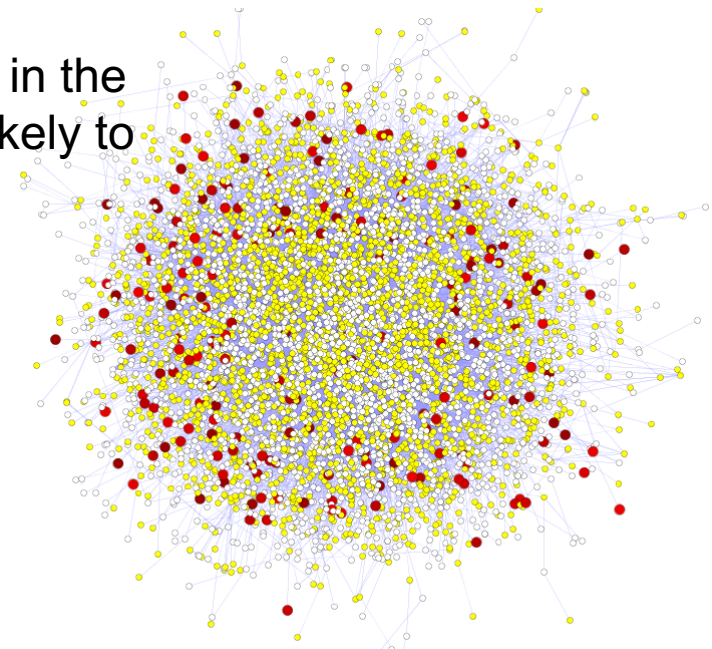
Power-law distribution



Hubs & Bottlenecks Under Constraint: A Finding from the Network Biology Community

- High likelihood of positive selection
- Lower likelihood of positive selection
- Not under positive selection
- No data about positive selection

[Nielsen et al. *PLoS Biol.* (2005), HPRD, Kim et al. *PNAS* (2007)]



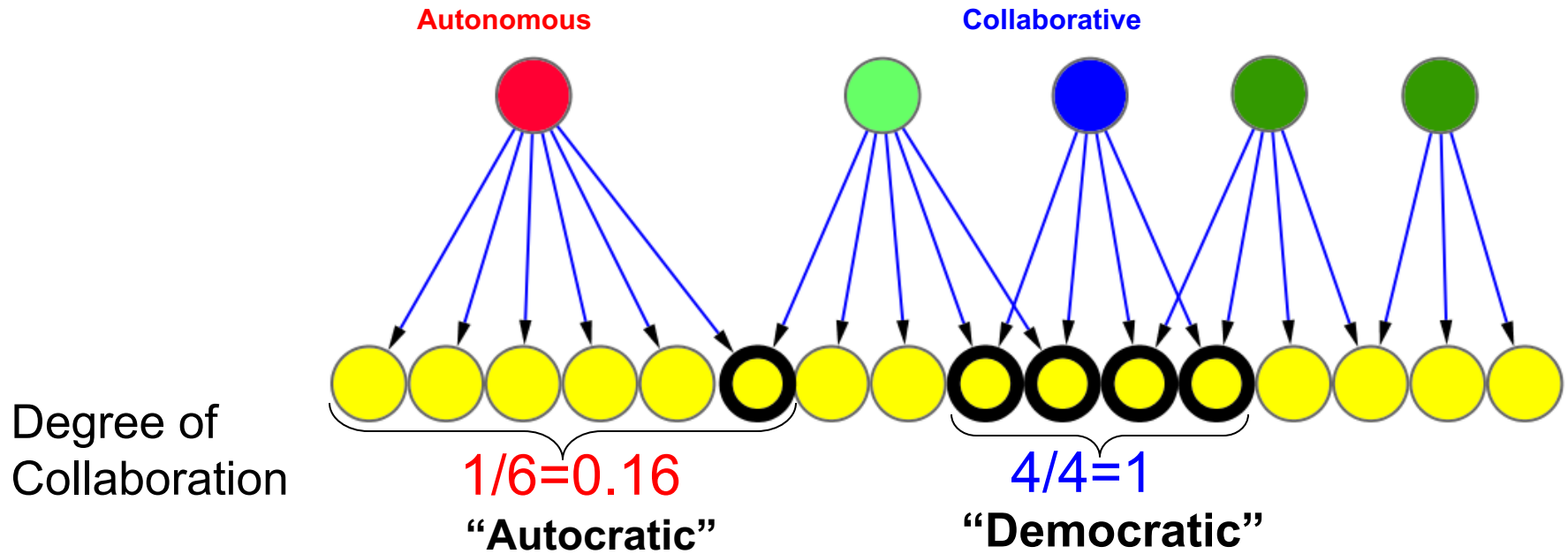
- **More Connectivity, More Constraint:**

Genes & proteins that have a more central position in the network tend to evolve more slowly and are more likely to be essential.

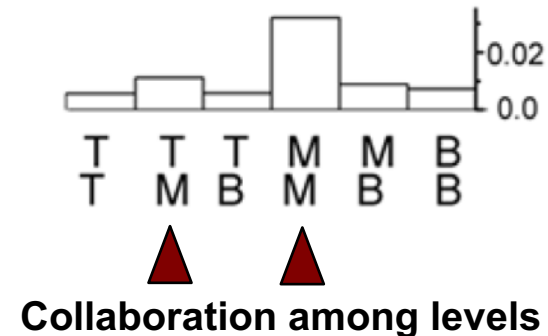
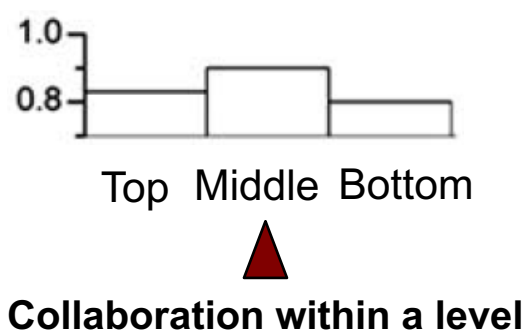
- This phenomenon is observed in **many organisms & different kinds of networks**

- Protein-Protein Interaction Networks – human, yeast, E coli Worm, fly
- miRNA Network

Collaborative Regulation Eases bottlenecks

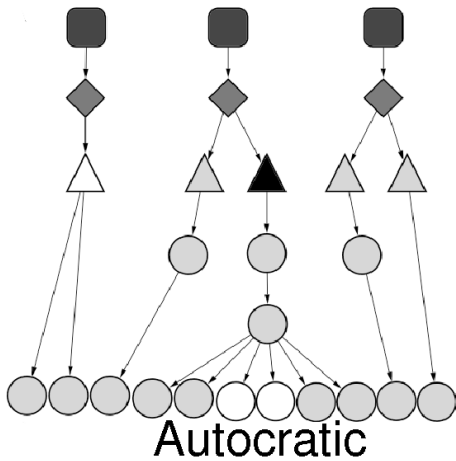


In a hierarchy, middle level (M) involves most collaboration (vs Top & Bot)
 (e.g, human transcription regulatory network)

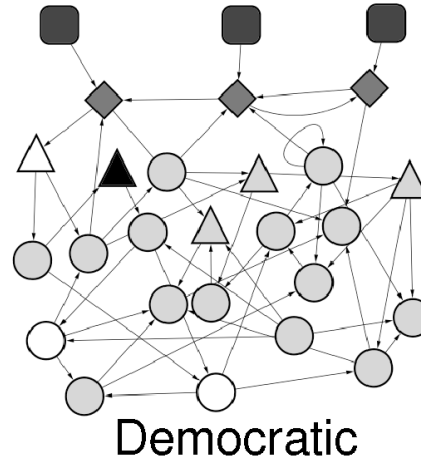


[Bhardwaj et al., PNAS (2010), in press]

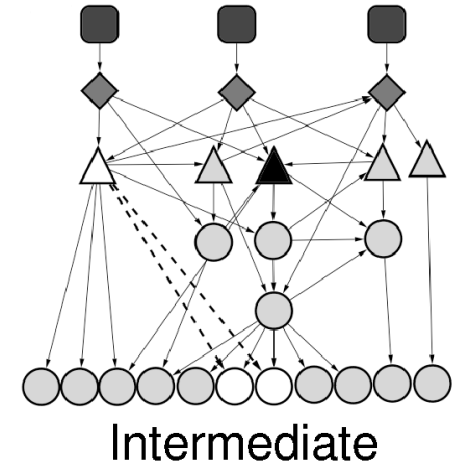
Different kinds of Hierarchies



- **Well-defined levels and a clear chain of command**
- *E.g., a military hierarchy*



- **Without well-defined levels & with more co-regulatory partnerships**
- *E.g., a club or a scientific collaboration network*



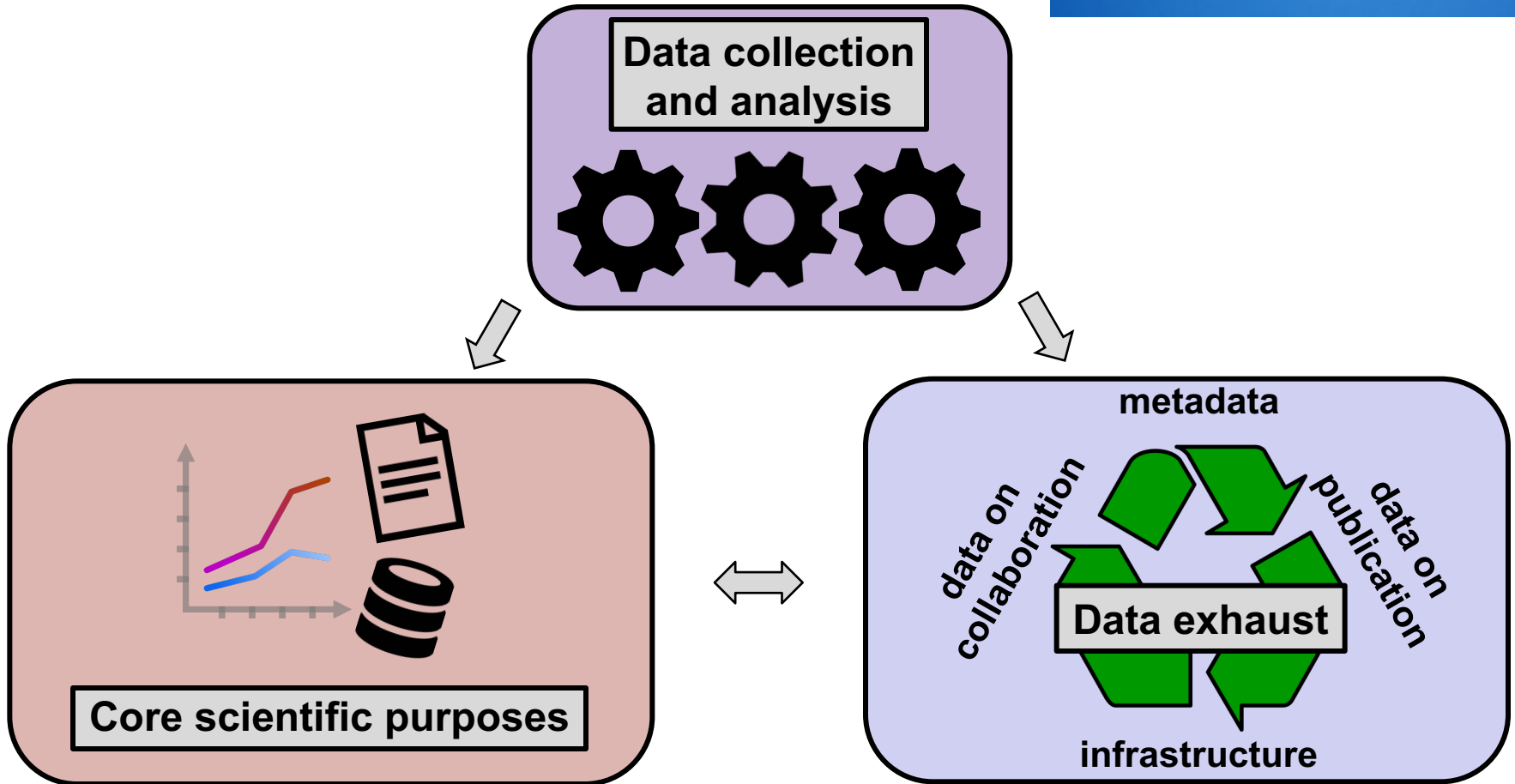
- **High degree of co-regulation and can be organized into hierarchies**
- *E.g., a law firm*

**The Other Side of the
Data Science Coin:**

**The Data Exhaust from
Personal Genomics
(SOS)**

Data Exhaust

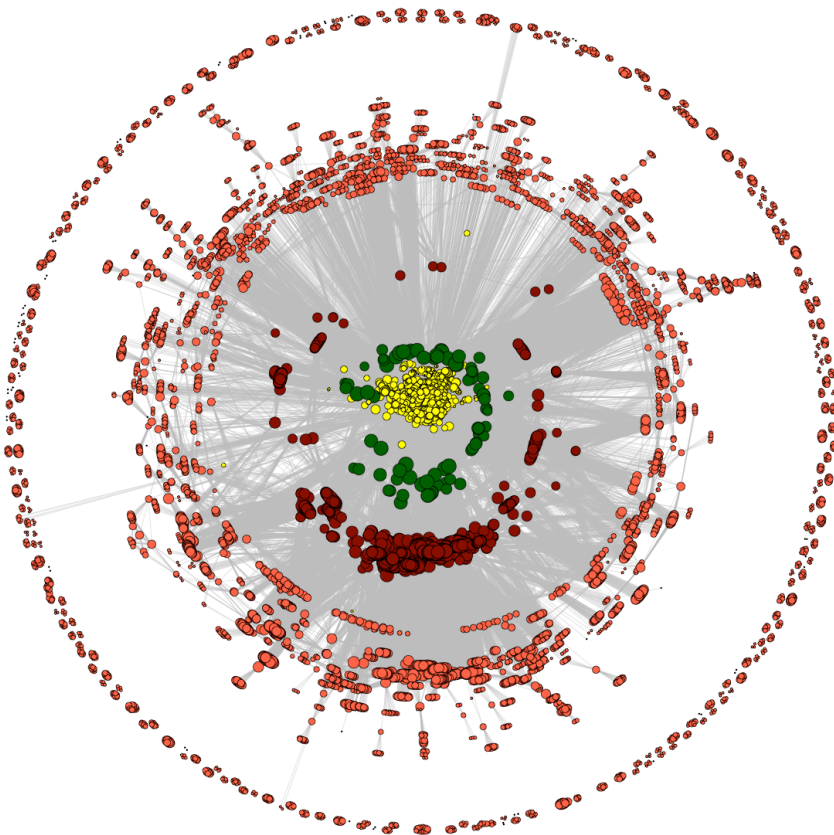
- Creative use of data is key to data science!
- Data exhaust = exploitable byproducts of big data collection and analysis



[photos: wikipedia/wikimedia]

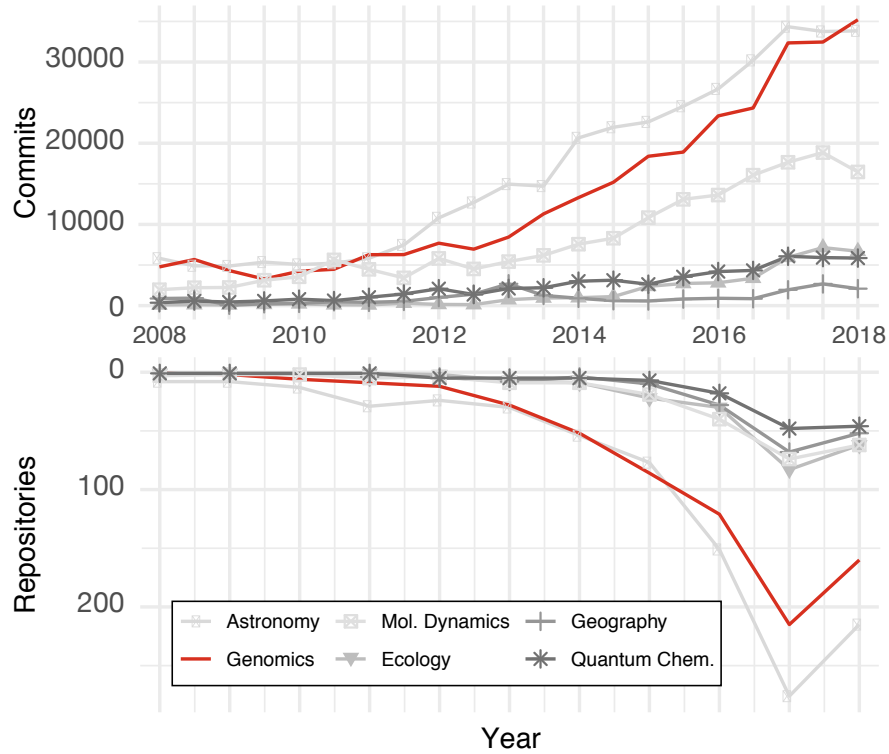
Exhaust Mining Application: Using Science to Study Science (SOS)

- ENCODE member
- non-member
- ENCODE member broker
- non-member broker
- co-authorship



- 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

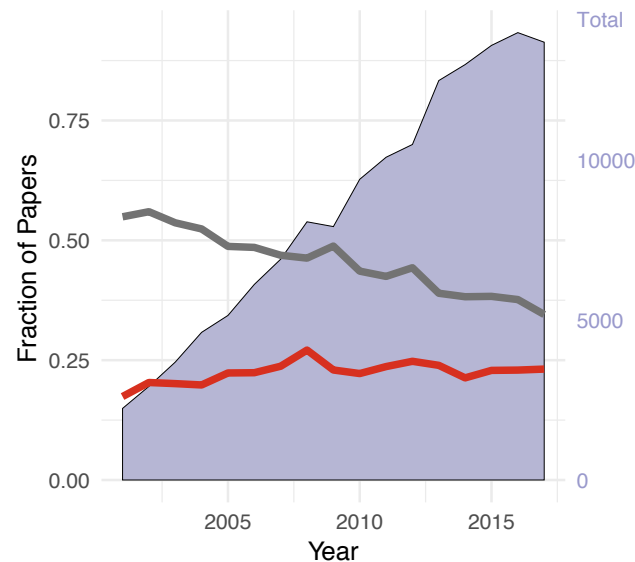
Quantifying Imports & exports from/to Genomics & other Data Science subfields



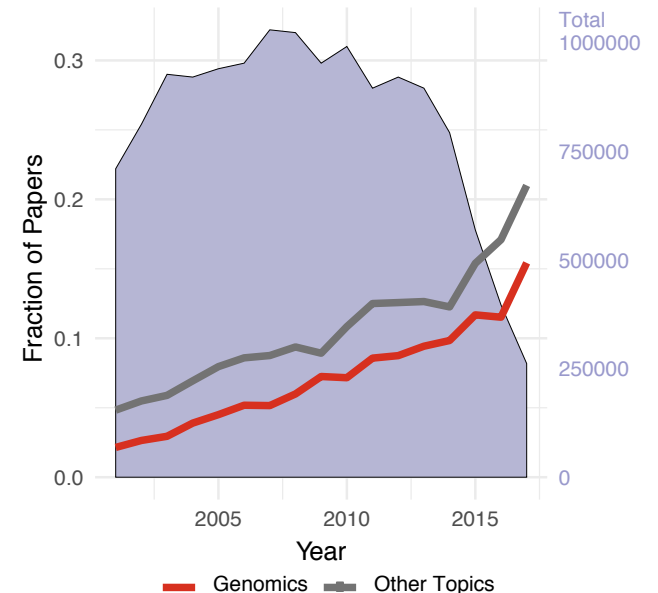
Genomics adoption of open source

Genomics contribution of HMM and Network papers

A) Hidden Markov Models



B) Network



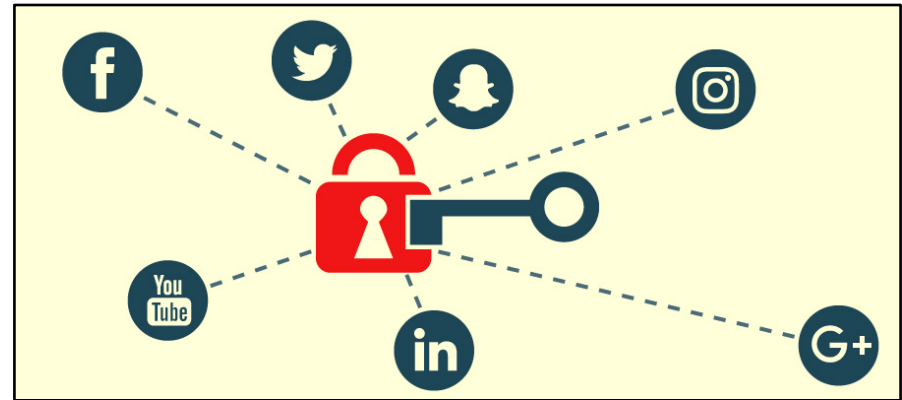
[Navarro et al. GenomeBiol. ('19, in press)]

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?

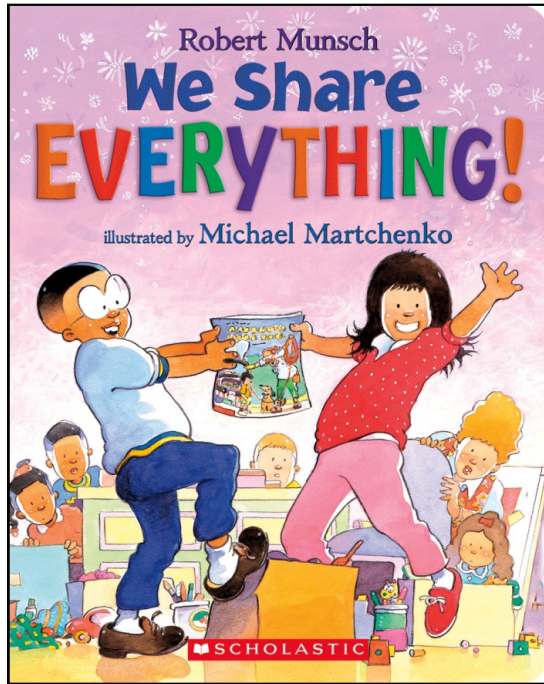
Genomic sequence very revealing about one's children. Is true consent possible?

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?

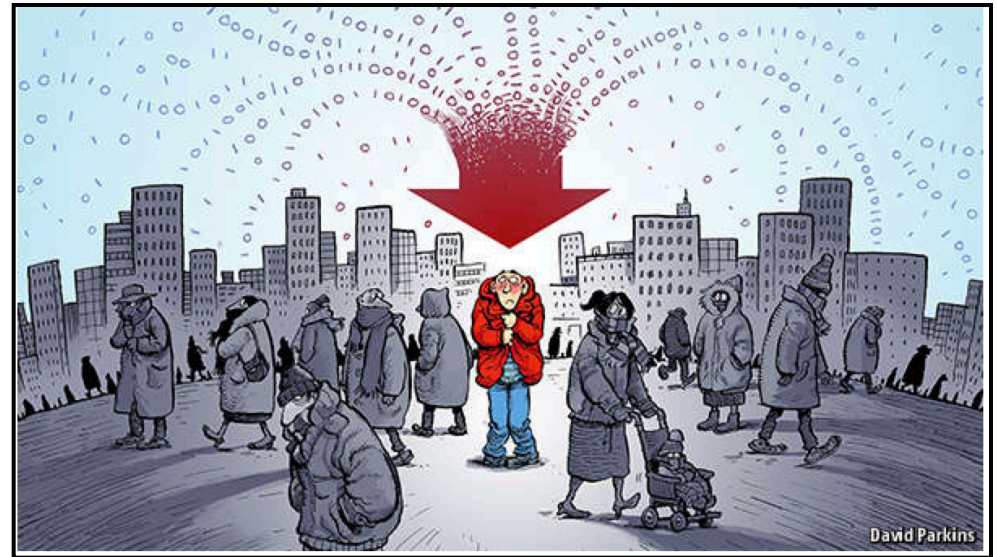


The Dilemma

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

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]

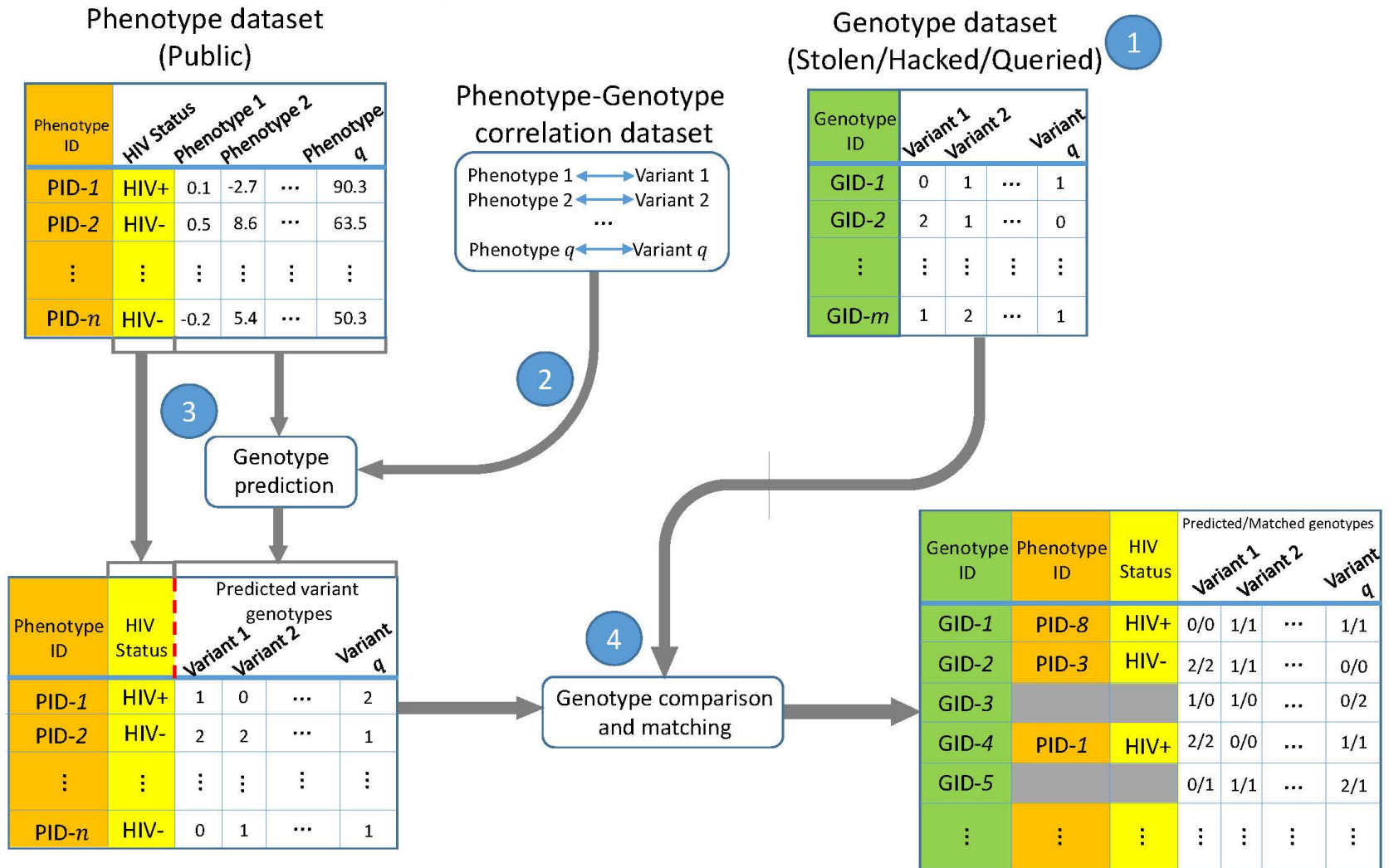
[Yale Law Roundtable ('10). Comp. in Sci. & Eng. 12:8; D Greenbaum & M Gerstein ('09). Am. J. Bioethics; D Greenbaum & M Gerstein ('10). SF Chronicle, May 2, Page E-4; Greenbaum et al. PLOS CB ('11)]

Current Social & Technical Solutions: The quandary where are now

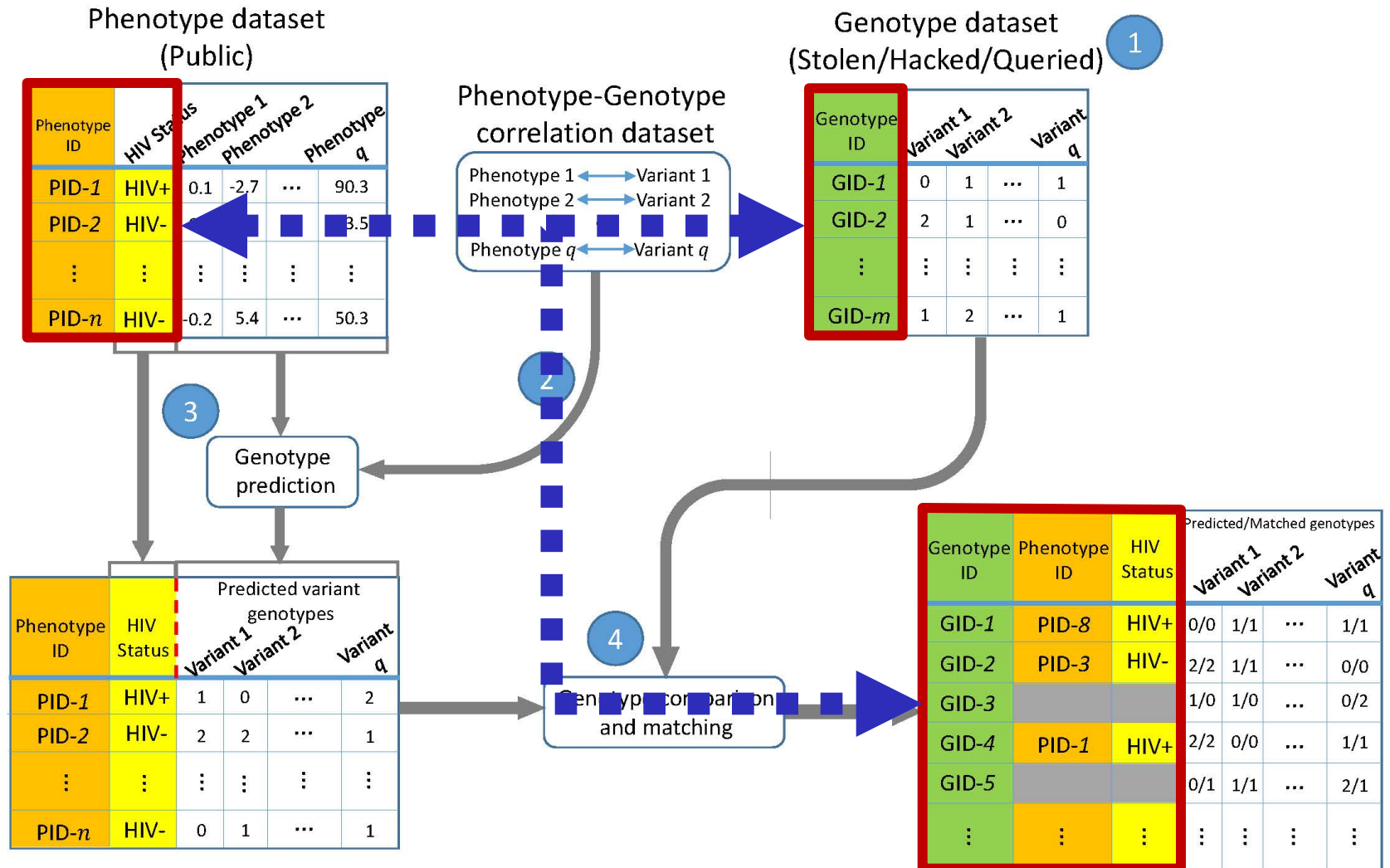
- **Closed Data** Approach
 - Consents
 - “Protected” distribution via dbGAP
 - Local computes on secure computer
- Issues with Closed Data
 - Non-uniformity of consents & paperwork
 - Different, confusing int'l norms
 - Computer security is burdensome
 - Many schemes get “hacked” .
 - **Tricky aspects 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



Linking Attack Scenario



Linking Attack Scenario



Linking Attacks: Case of Netflix Prize



Names available for many users!

User (ID)	Movie (ID)	Date of Grade	Grade [1,2,3,4,5]
NTFLX-0	NTFLX-19	10/12/2008	1
NTFLX-1	NTFLX-116	4/23/2009	3
NTFLX-2	NTFLX-92	5/27/2010	2
NTFLX-1	NTFLX-666	6/6/2016	5
...
...

User (ID)	Movie (ID)	Date of Grade	Grade [0-10]
IMDB-0	IMDB-173	4/20/2009	5
IMDB-1	IMDB-18	10/18/2008	0
IMDB-2	IMDB-341	5/27/2010	-
...
...
...

- Many users are shared
- The grades of same users are correlated
- A user grades one movie around the same date in two databases

Anonymized Netflix Prize Training Dataset
made available to contestants

Linking Attacks: Case of Netflix Prize



User (ID)	Movie (ID)	Date of Grade	Grade [1,2,3,4,5]
NTFLX-0	NTFLX-19	10/12/2008	1
NTFLX-1	NTFLX-116	4/23/2009	3
NTFLX-2	NTFLX-92	5/27/2010	2
NTFLX-1	NTFLX-666	6/6/2016	5
...
...

User (ID)	Movie (ID)	Date of Grade	Grade [0-10]
IMDB-0	IMDB-173	4/20/2009	5
IMDB-1	IMDB-18	10/18/2008	0
IMDB-2	IMDB-341	5/27/2010	-
...
...
...

- Many users are shared
- The grades of same users are correlated
- A user grades one movie around the same date in two databases
- IMDB users are public
- NetFLIX and IMdB moves are public

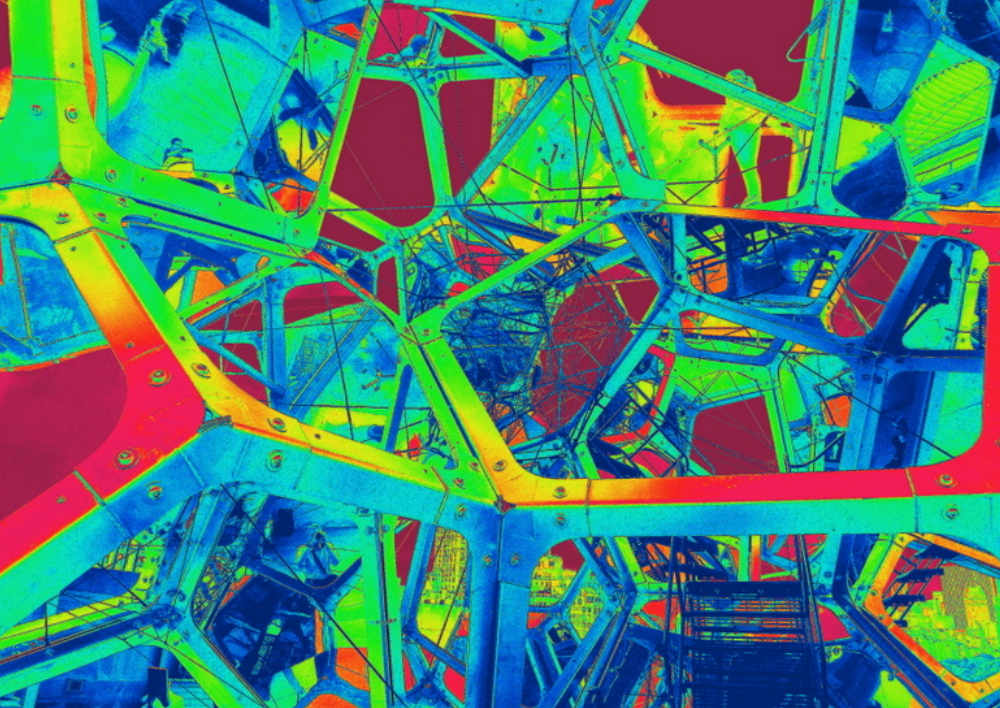
Linking Attacks: Case of Netflix Prize



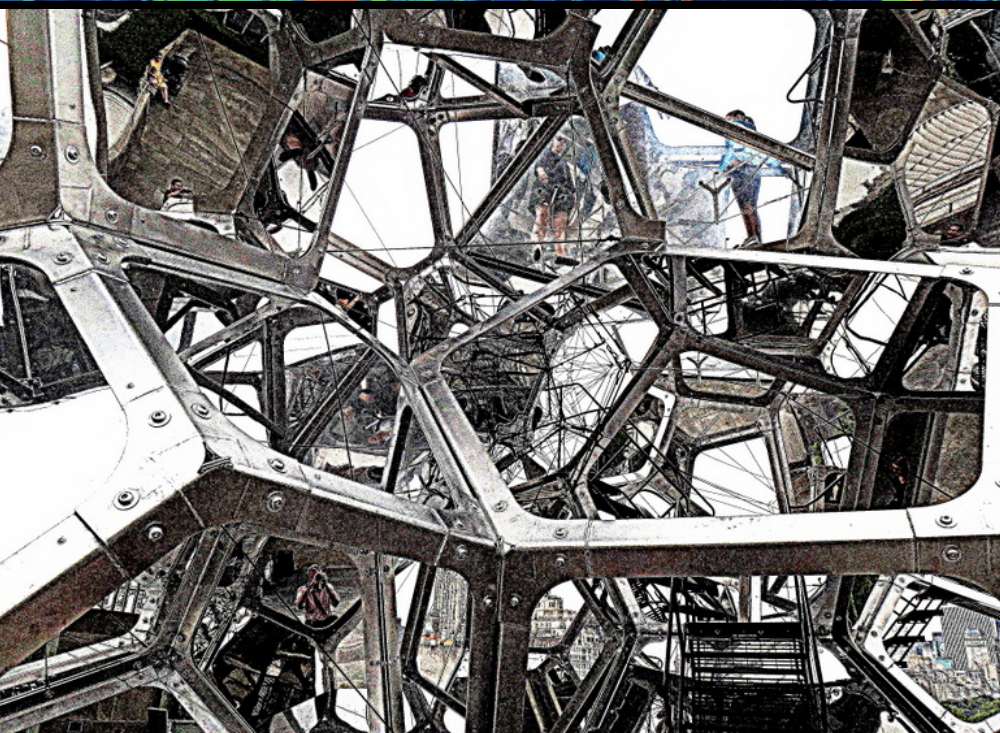
User (ID)	Movie (ID)	Date of Grade	Grade [1,2,3,4,5]
NTFLX-0	NTFLX-19	10/12/2008	1
NTFLX-1	NTFLX-116	4/23/2009	3
NTFLX-2	NTFLX-92	5/27/2010	2
NTFLX-1	NTFLX-666	6/6/2016	5
...
...

User (ID)	Movie (ID)	Date of Grade	Grade [0-10]
IMDB-0	IMDB-173	4/20/2009	5
IMDB-1	IMDB-18	10/18/2008	0
IMDB-2	IMDB-341	5/27/2010	-
...
...
...

- Many users are shared
- The grades of same users are correlated
- A user grades one movie around the same date in two databases



Biomedical Data
Science:
An Introduction



**Thanks for your
attention**

Teaching pack with no explicit credits.
Credits & references on each slide

Extra



Info about content in this slide pack

- General PERMISSIONS
 - This Presentation is copyright Mark Gerstein, Yale University, 2019.
 - Please read permissions statement at www.gersteinlab.org/misc/permissions.html .
 - Feel free to use slides & images in the talk with PROPER acknowledgement (via citation to relevant papers or link to gersteinlab.org).
 - Paper references in the talk were mostly from Papers.GersteinLab.org.
- PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see <http://streams.gerstein.info> .
 - In particular, many of the images have particular EXIF tags, such as kwpotppt , that can be easily queried from flickr, viz: <http://www.flickr.com/photos/mbgmbg/tags/kwpotppt>