Genomics & Data Science

M Gerstein, Yale

(See last slide for more info.)

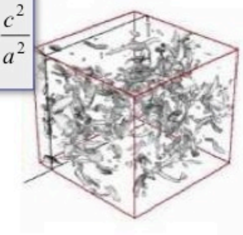
Credits & references on each slide

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

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

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



#3 - Simulation

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

Emphasis on: Supercomputers

#4 - Data Science

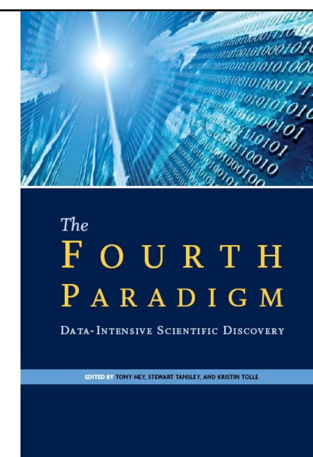
Data gathering and storing

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…..



[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 "Big Tech": a buzz-word for successful Ads

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

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

CIO Network
INSIGHTS AND IDEAS FOR TECHNOLOGY LEADERS.
+ Follow (469)

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

Why Big Data Is All Retailers Want for Christmas

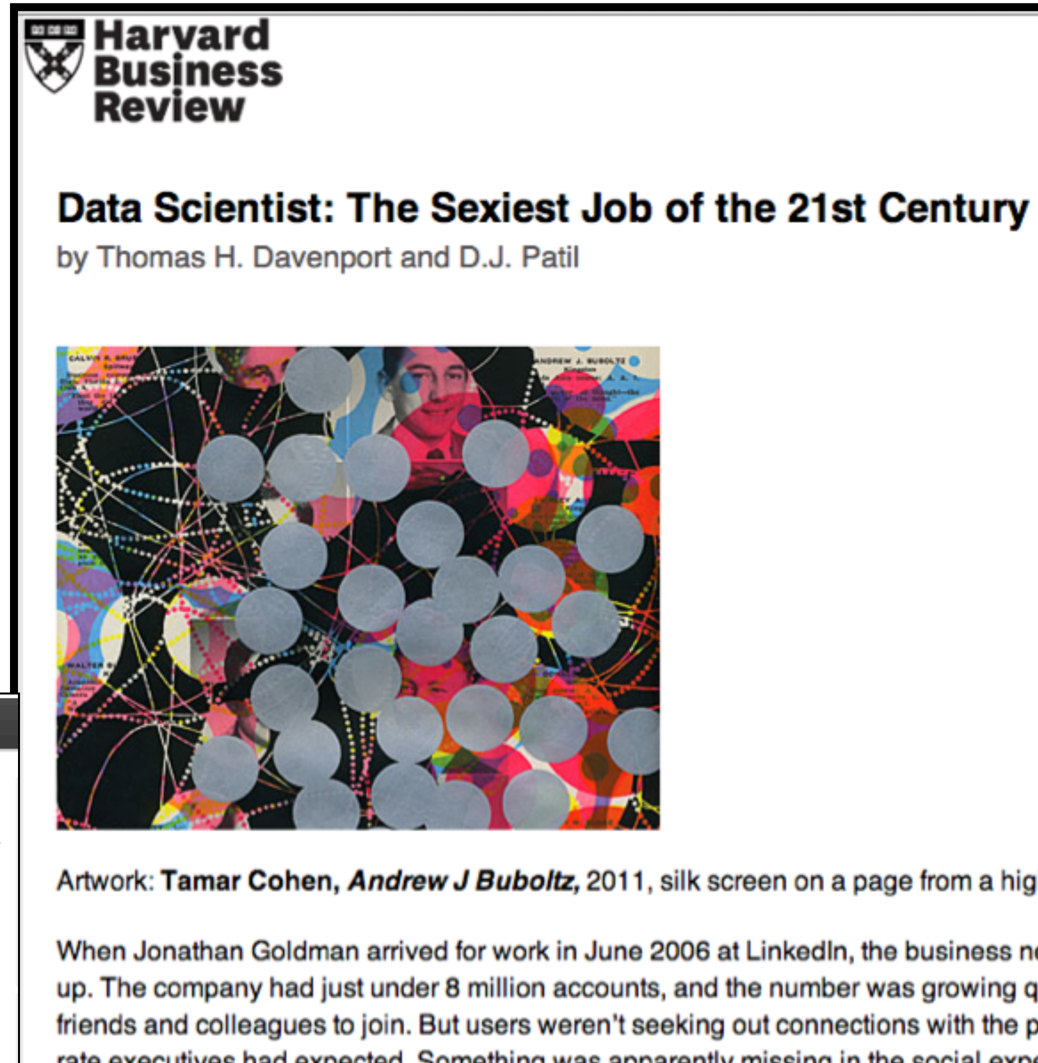
Eric Savitz, Forbes Staff

+ 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.

108 Share
349 Tweet
193 Share
353 Submit
12 +1



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

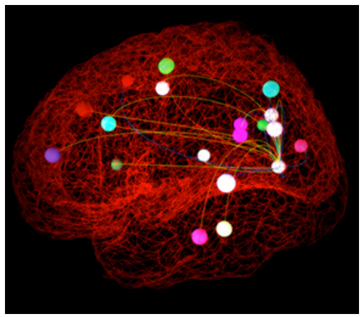
When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business ne
up. The company had just under 8 million accounts, and the number was growing qu
friends and colleagues to join. But users weren't seeking out connections with the pe
rate executives had expected. Something was apparently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

[Oct. '12 issue]

Data Science in Traditional Science

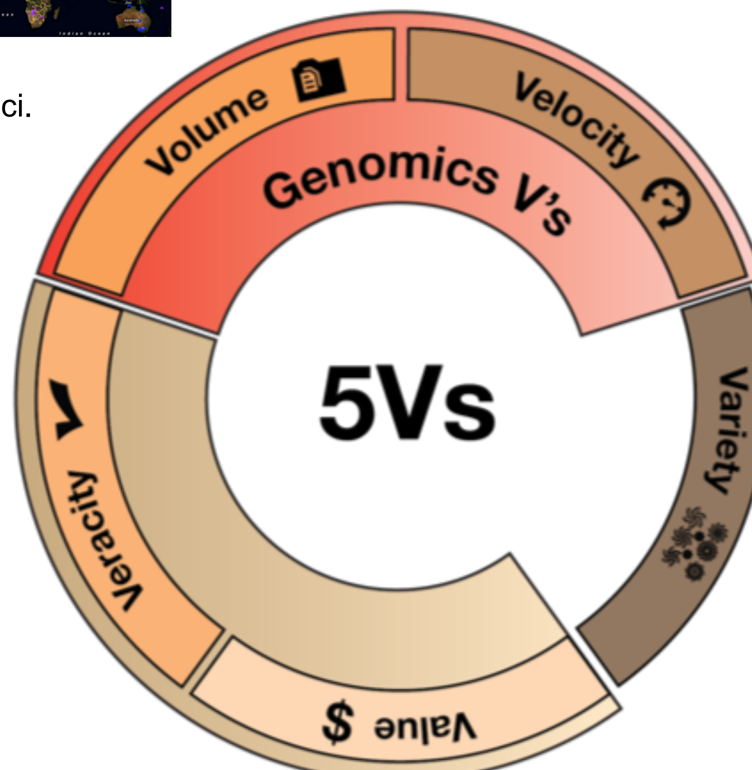
- Pre-dated commercial mining
- Instrument generated
- Large data sets often created by large teams not to answer one Q but to be mined broadly
- Often coupled to a physical/biological model
- Interplay w/ experiments

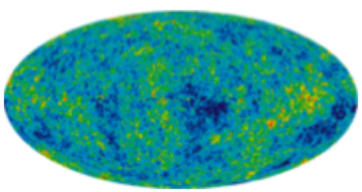
High energy physics - Large Hadron Collider



Neuroscience - The Human Connectome Project

Ecology & Earth Sci. - Fluxnet





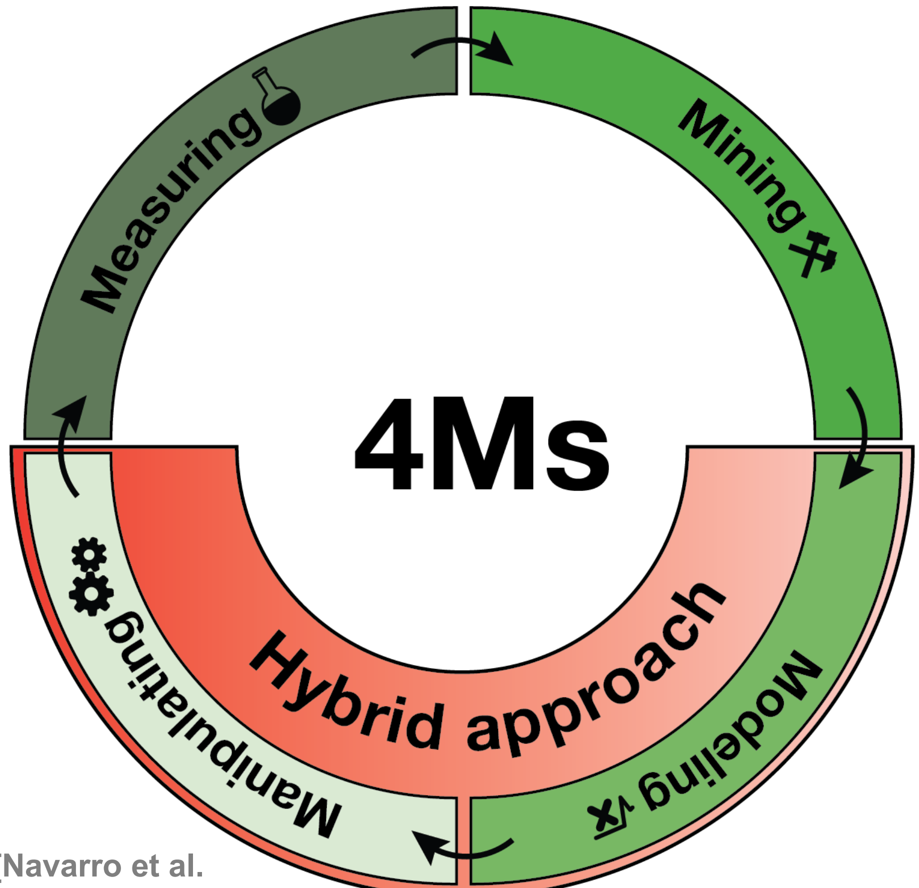
Astronomy - Sloan Digital Sky survey

Genomics DNA sequencer

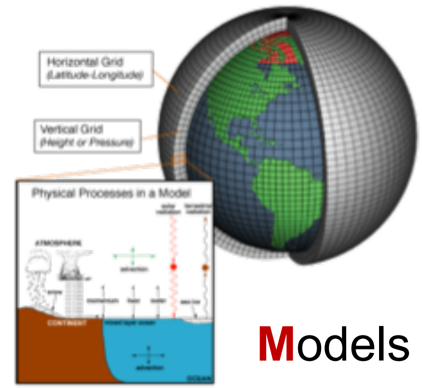
[Navarro et al. GenomeBiol. ('19, in press)]

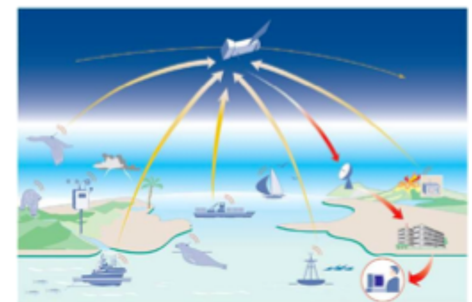
Coupling of Scientific Data to Models & Experiments

- 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)]





Models + Data **M**ining

↓

Forecasts

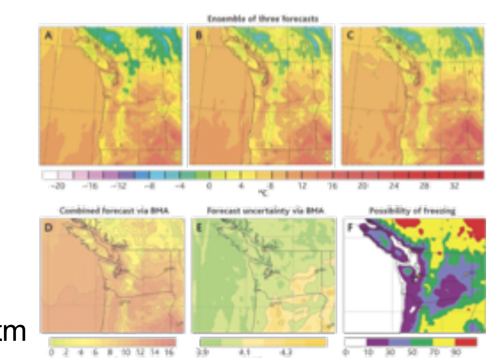


Image from http://web.aibn.uq.edu.au/cssb/ResearchProjects.htm

NOAA

Genomics & Data Science

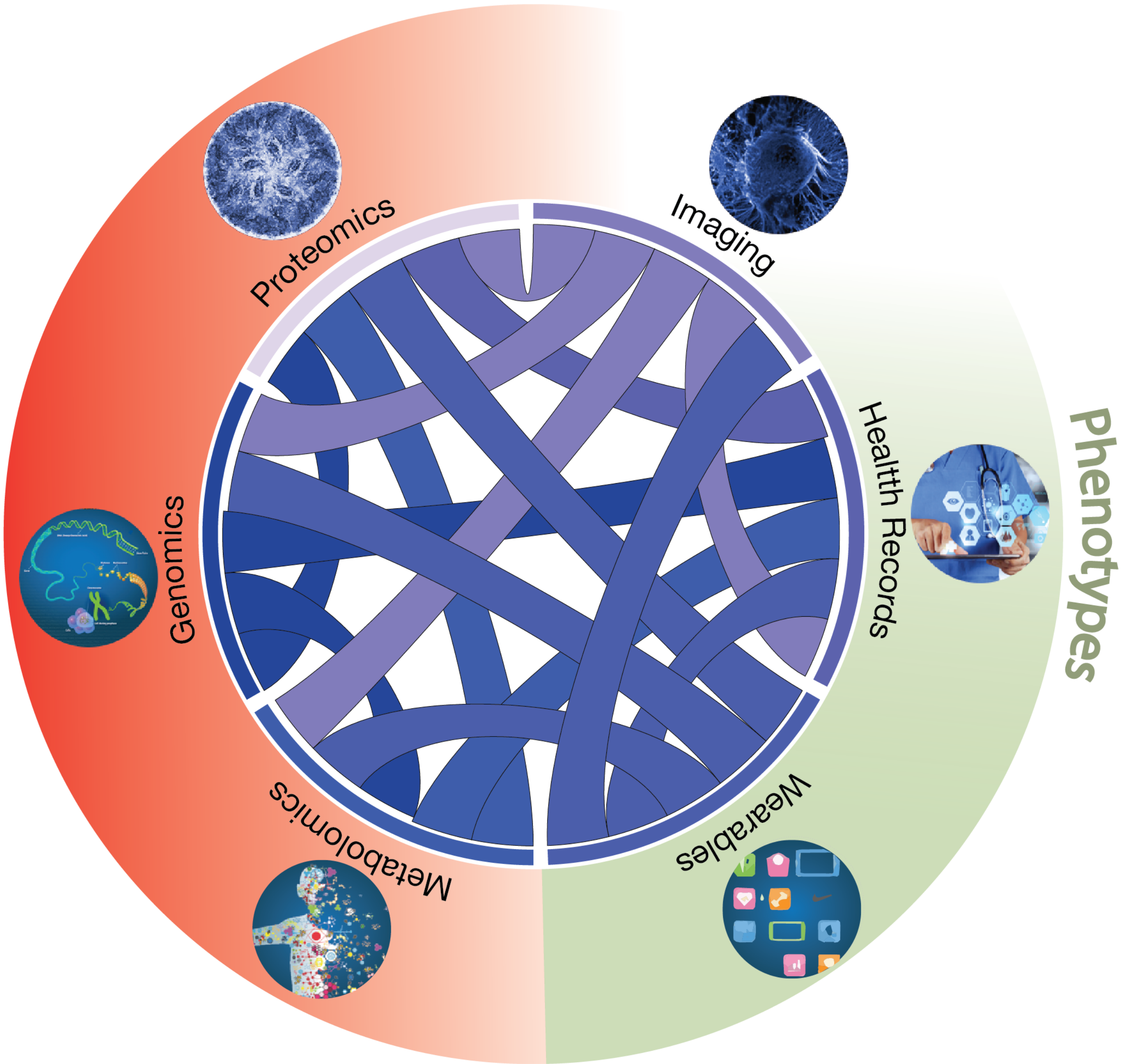
- **Background** on data science & the biomed. sub-domain
  - Placing genomics in this context -- **3V & 4M** frameworks
  - Types of data & integration betw. types + **Moore's law scaling**
- Main **practical heath applications**
  - Designing drugs, customizing treatments….
  - Focused example on **identifying drug targets** for neuropsych. diseases
- Another application: **"Bioinspired" insights on systems** design
  - Reducing bottlenecks in the middle of a network hierarchy
  - Different patterns of connectivity & constraint in natural v engineered systems
- Mining the **data exhaust** (from personal genomics)
  - SOS
  - **Privacy**

Genomics & Data Science

- **Background** on data science & the biomed. sub-domain
  - Placing genomics in this context -- **3V & 4M** frameworks
  - Types of data & integration betw. types + **Moore's law scaling**
- Main **practical heath applications**
  - Designing drugs, customizing treatments….
  - Focused example on **identifying drug targets** for neuropsych. diseases
- Another application: **"Bioinspired" insights on systems** design
  - Reducing bottlenecks in the middle of a network hierarchy
  - Different patterns of connectivity & constraint in natural v engineered systems
- Mining the **data exhaust** (from personal genomics)
  - SOS
  - **Privacy**

Drivers of Biomedical Data Science

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

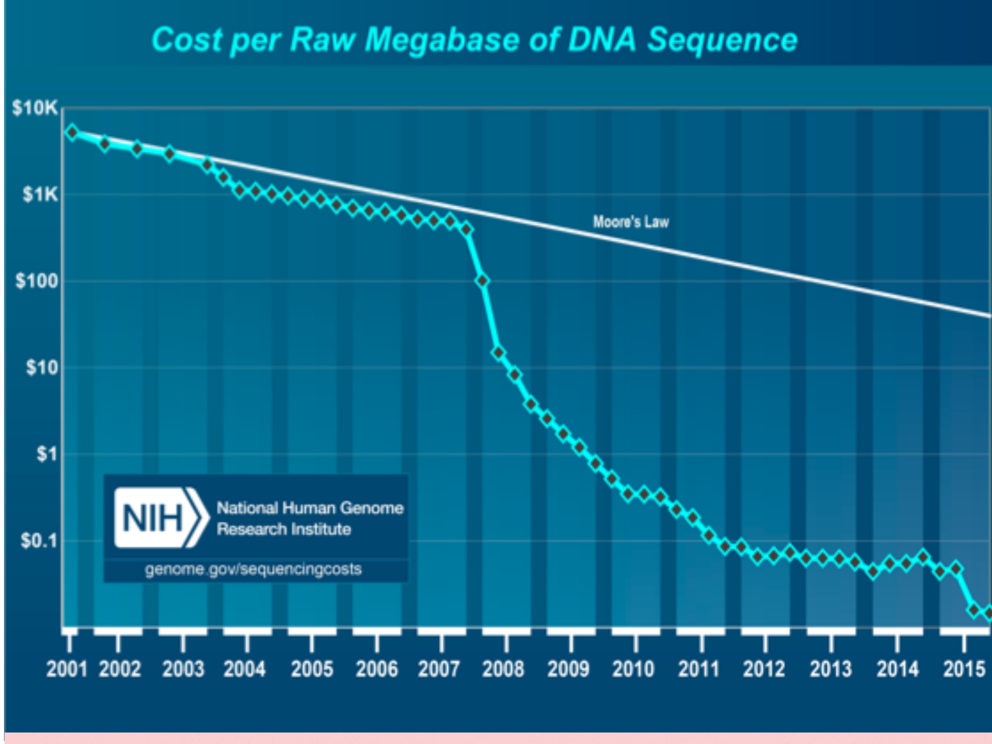


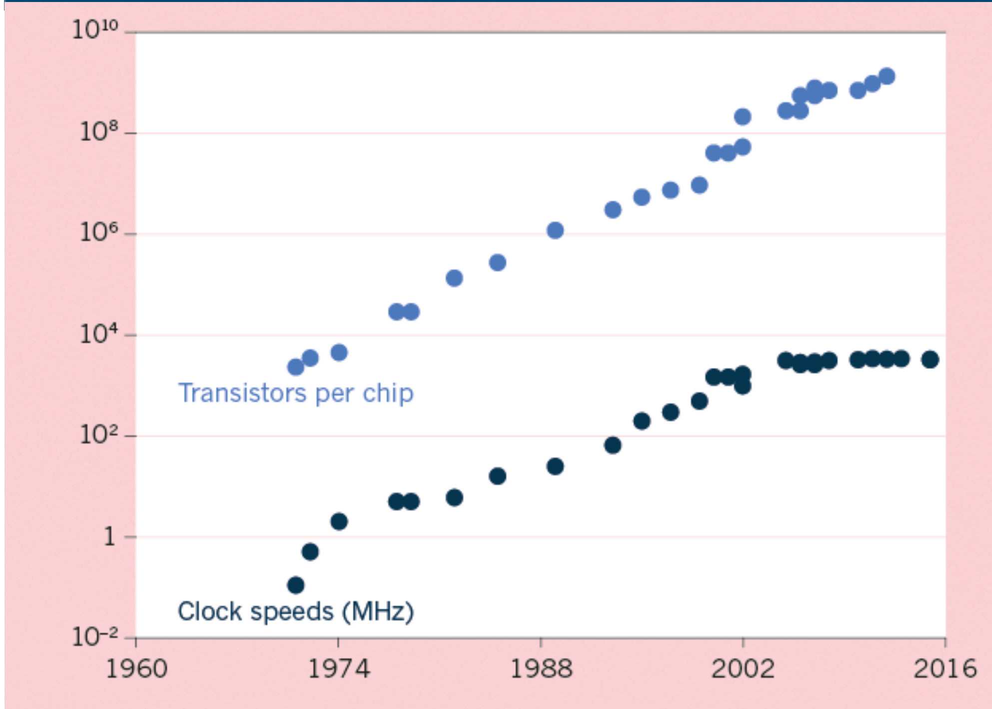
[Navarro et al. GenomeBiol. ('19, in press)]

The Scaling of Genomic Data Science:

Powered by exponential increases in data & computing

(Moore’s Law)

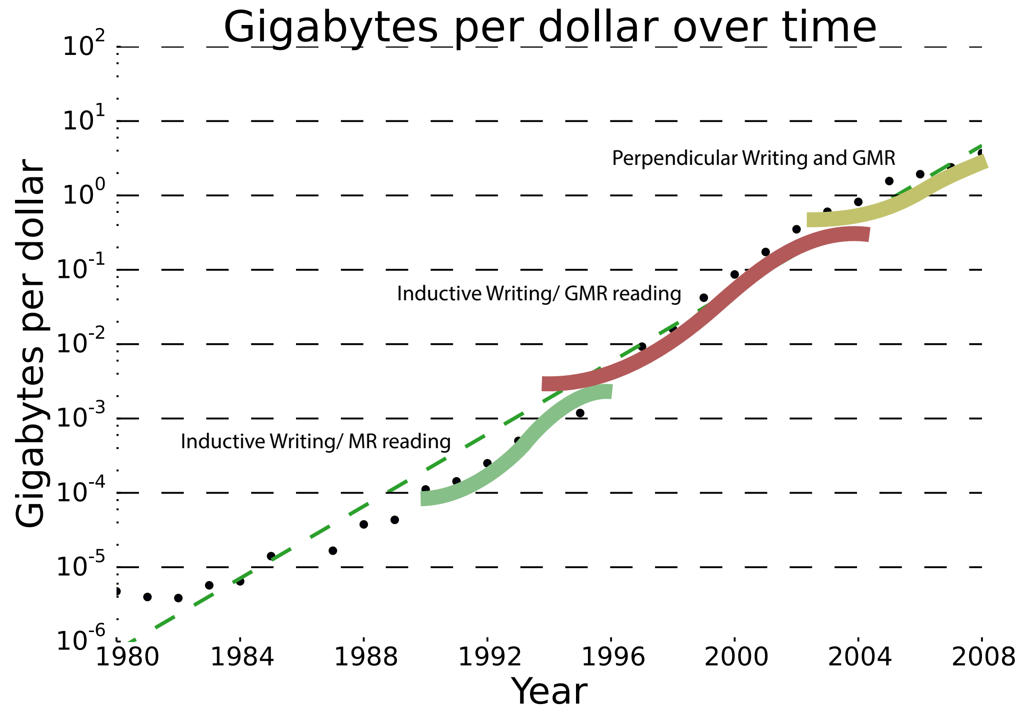


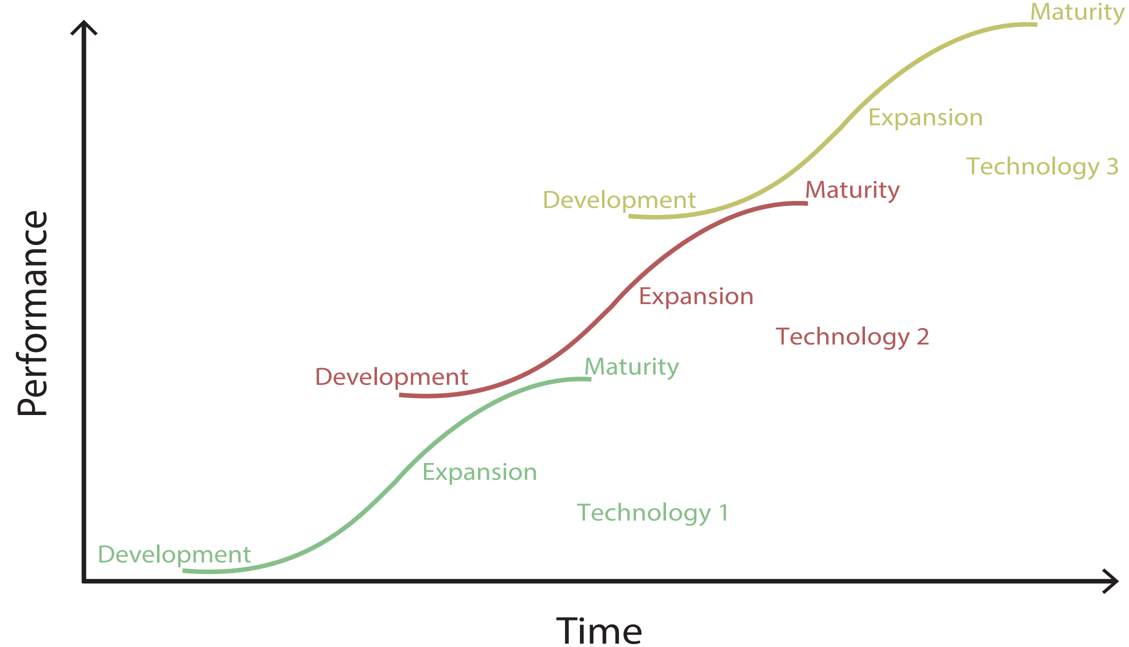


[NHGRI website + Waldrop (‘15) Nature]

Kryder's Law and S-curves underlying exponential growth

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

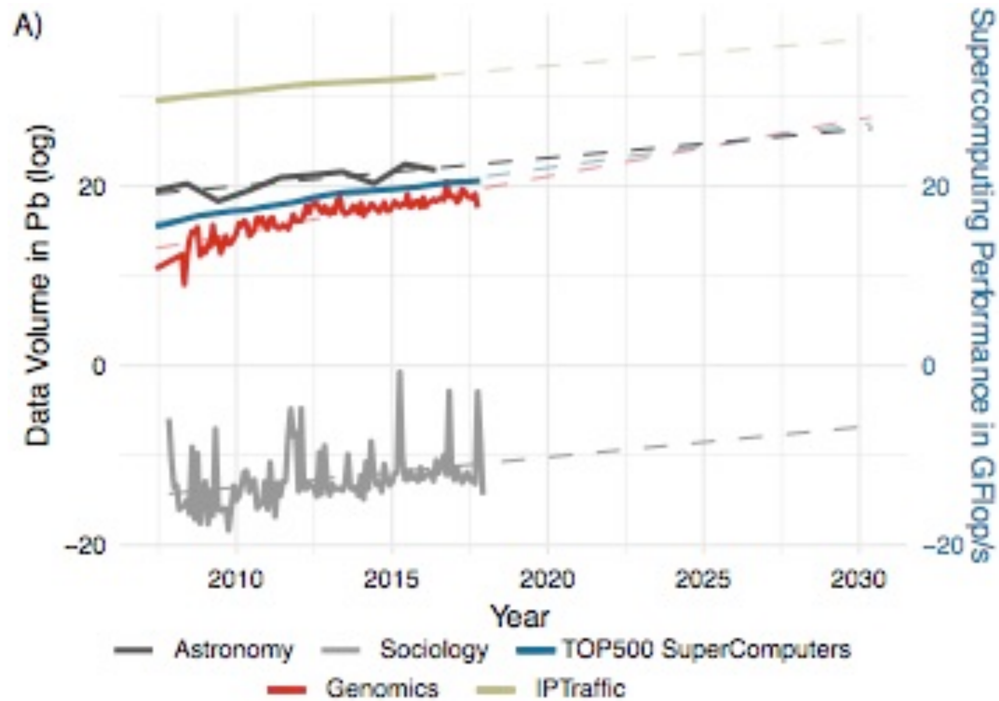


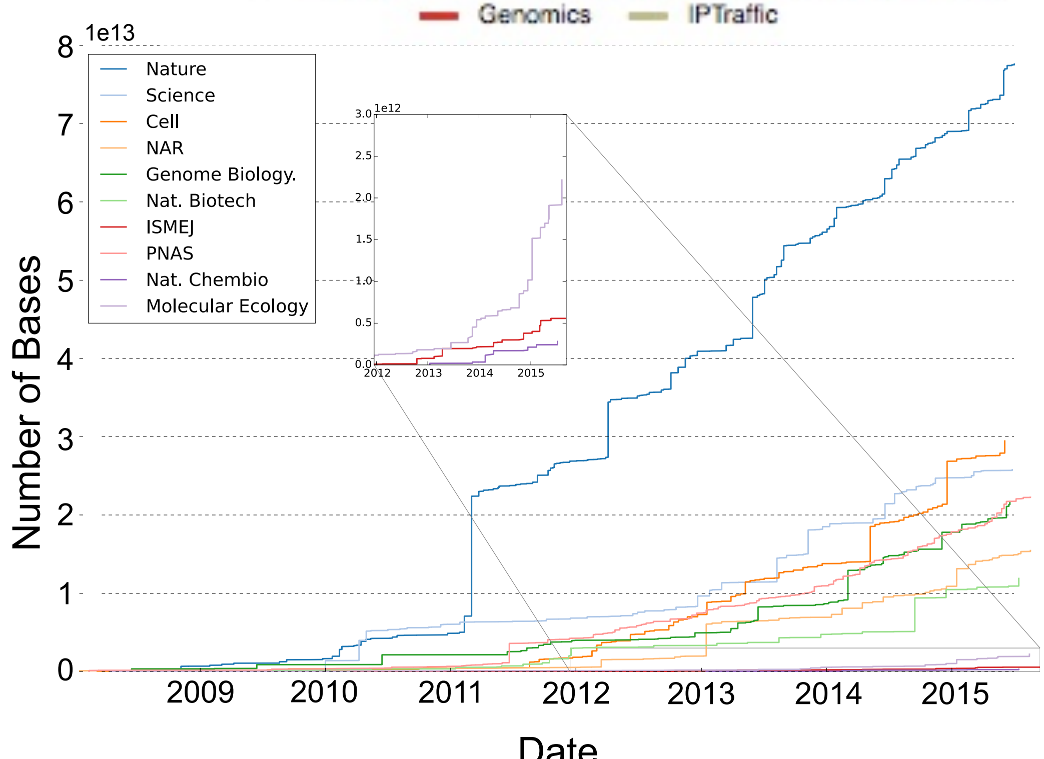


[Muir et al. ('15) GenomeBiol.]

Sequencing cost reductions have resulted in an explosion of data

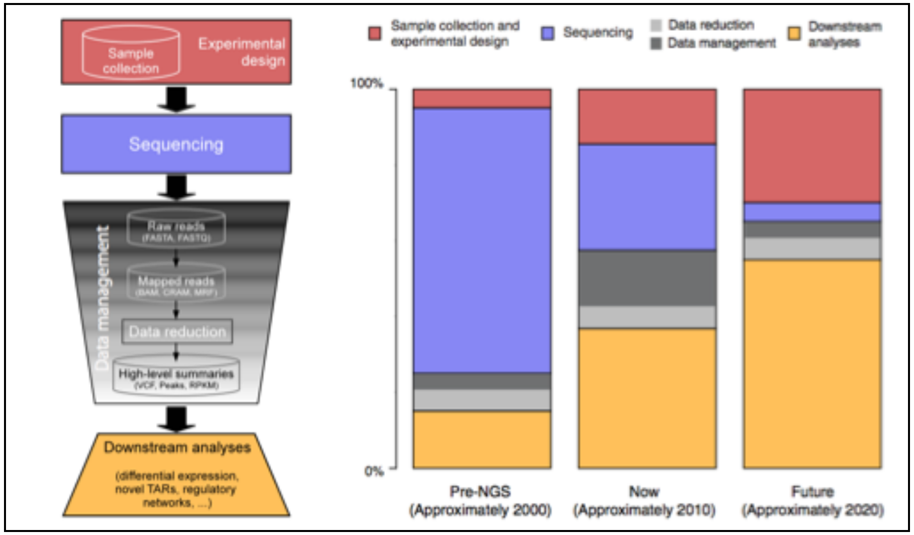
[Muir et al. ('15) GenomeBiol. Navarro et al. GenomeBiol. ('19, in press)]





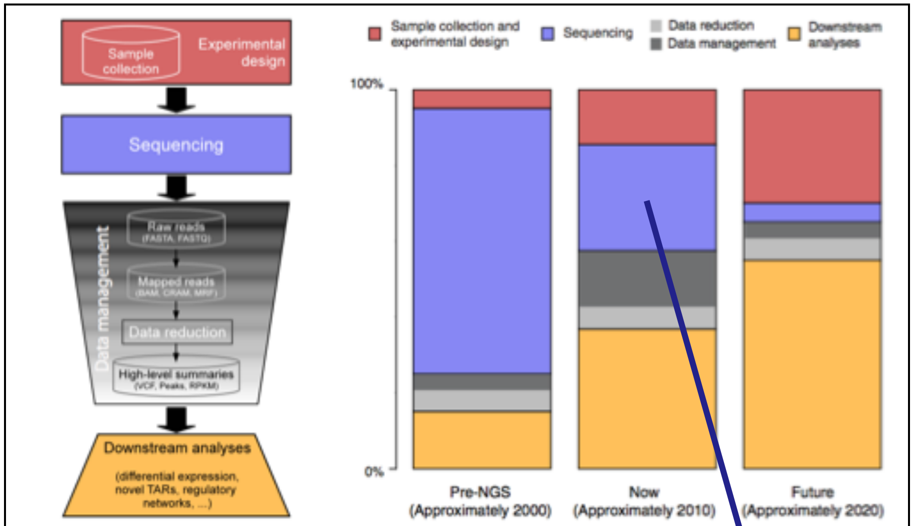
- The type of sequence data deposited has changed as well.

The changing costs of a sequencing pipeline

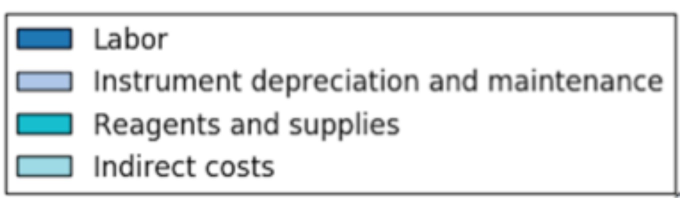


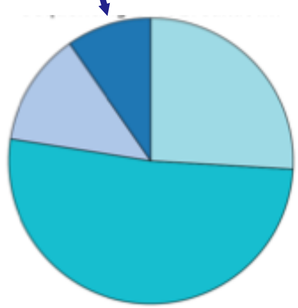
From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from
the actual seq. to sample
collection & analysis

The changing costs of a sequencing pipeline



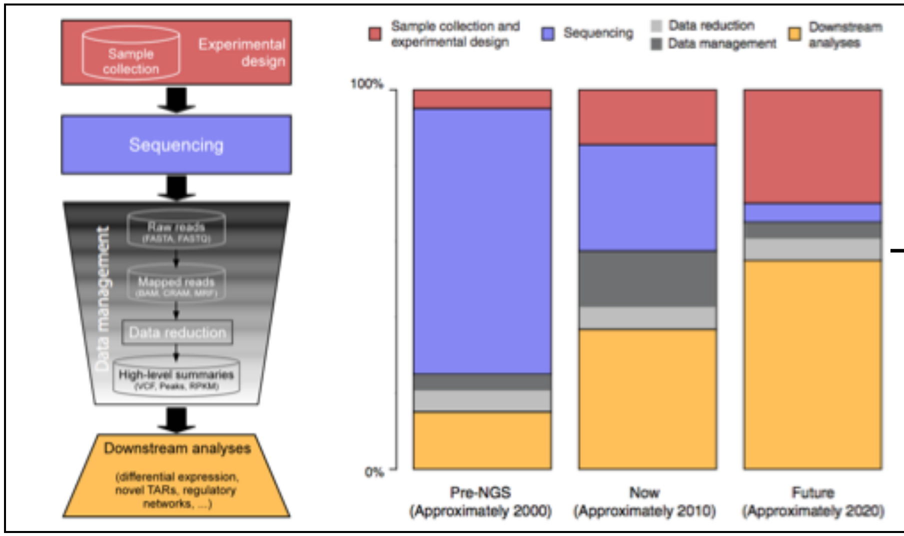
From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

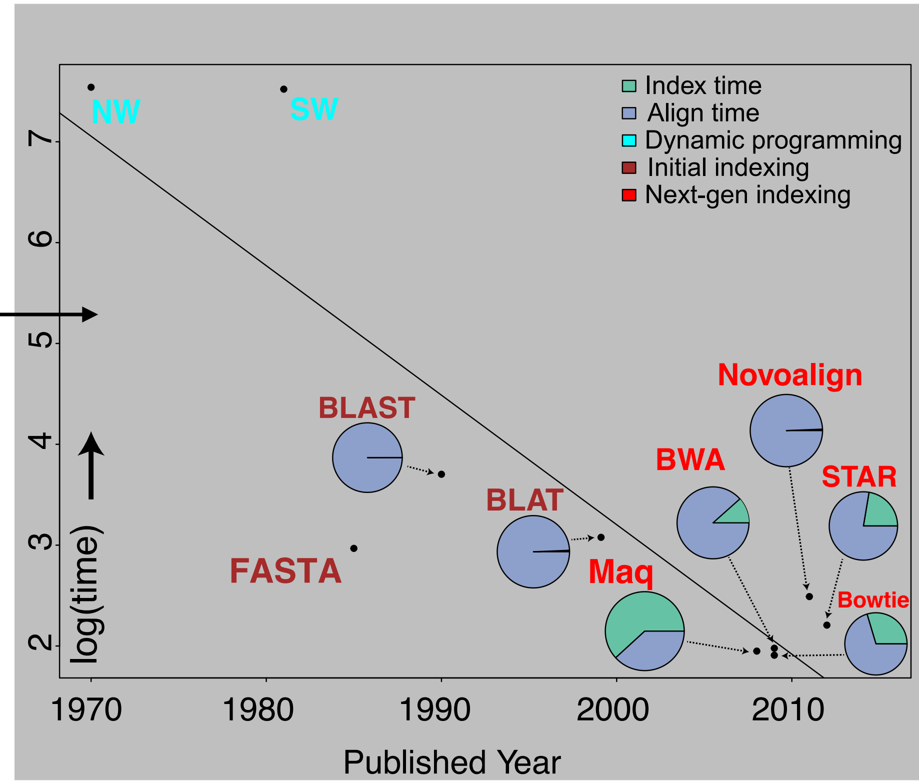




[Sboner et al. (‘11), Muir et al. (‘15) Genome Biology]

The changing costs of a sequencing pipeline



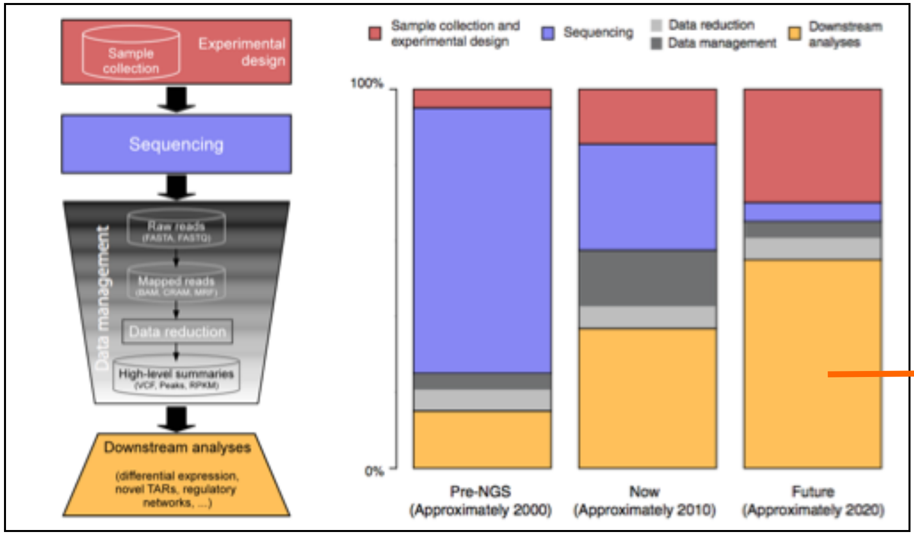


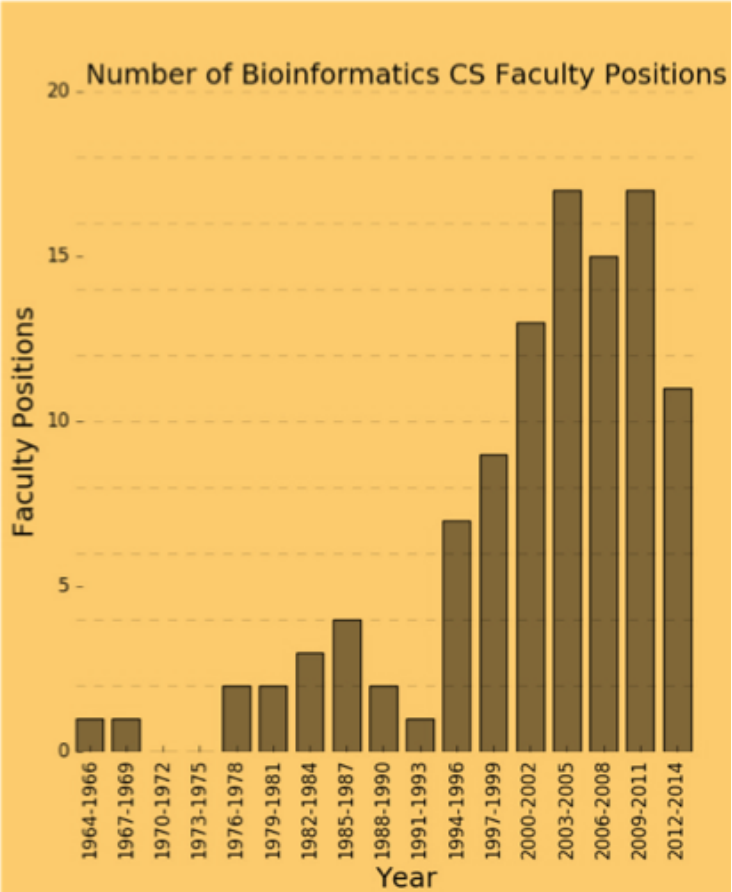
From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

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

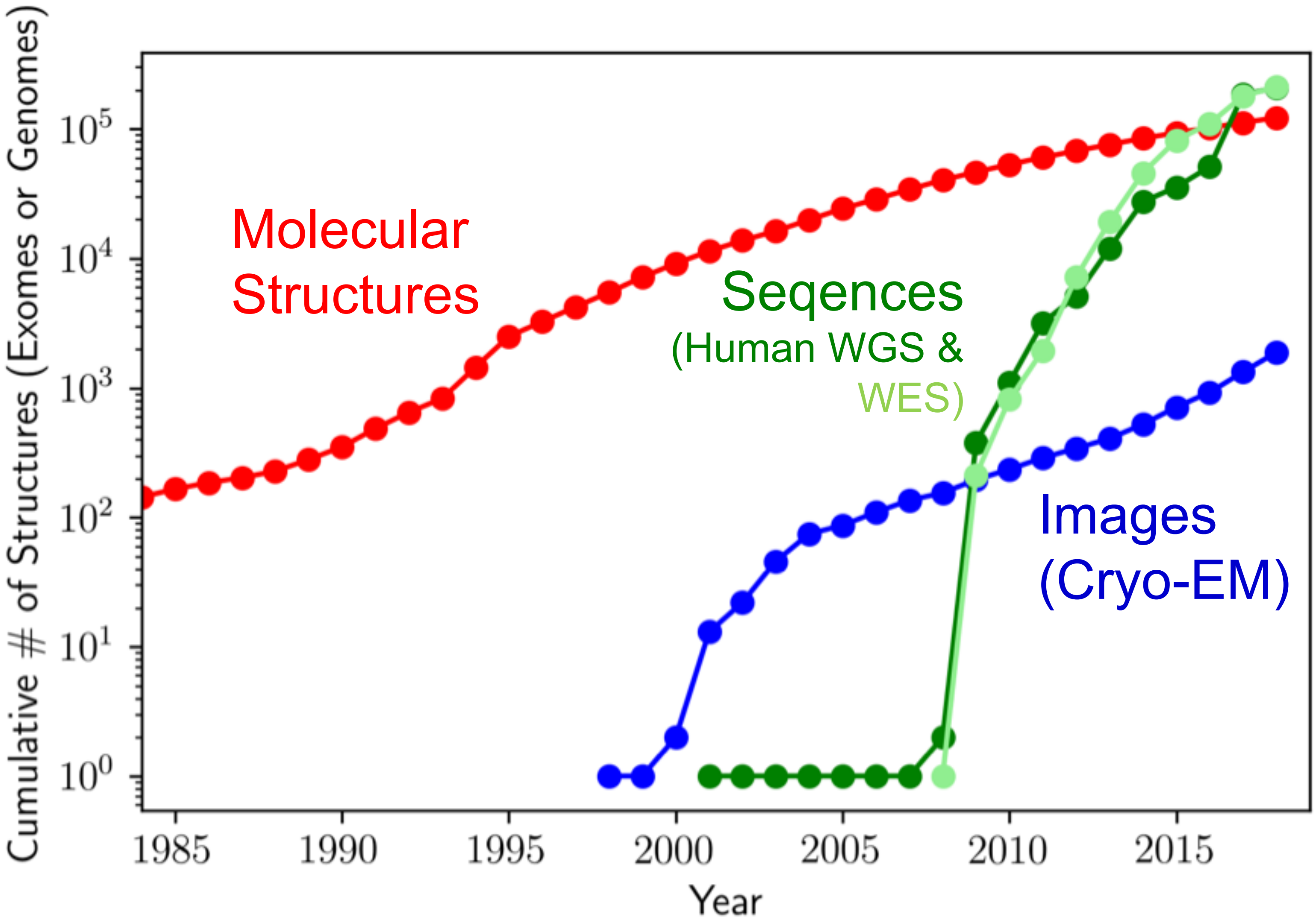




From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

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

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

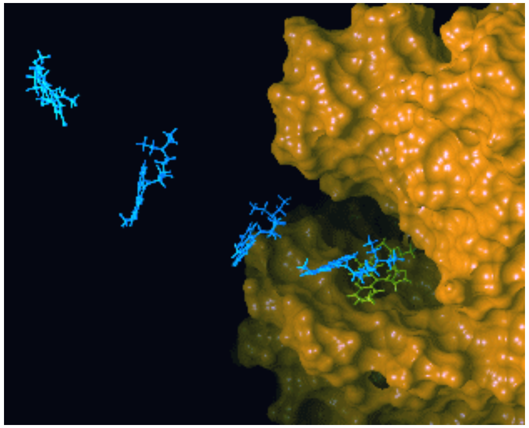


[Navarro et al. GenomeBiol. ('19, in press)]

Genomics & Data Science

- **Background** on data science & the biomed. sub-domain
  - Placing genomics in this context -- **3V & 4M** frameworks
  - Types of data & integration betw. types + **Moore's law scaling**
- Main **practical heath applications**
  - Designing drugs, customizing treatments….
  - Focused example on **identifying drug targets** for neuropsych. diseases
- Another application: **"Bioinspired" insights on systems** design
  - Reducing bottlenecks in the middle of a network hierarchy
  - Different patterns of connectivity & constraint in natural v engineered systems
- Mining the **data exhaust** (from personal genomics)
  - SOS
  - **Privacy**

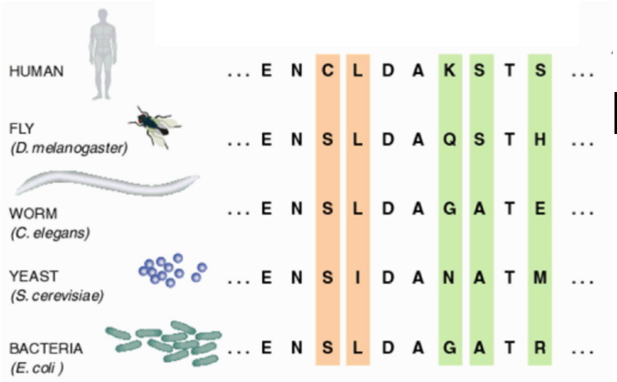
Major Bioinformatics Applications



Adapted from Olsen Group Docking Page at Scripps

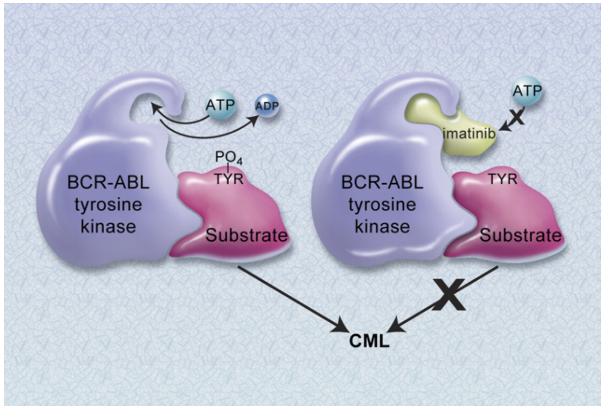
I. Designing Drugs from Structural Targets

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



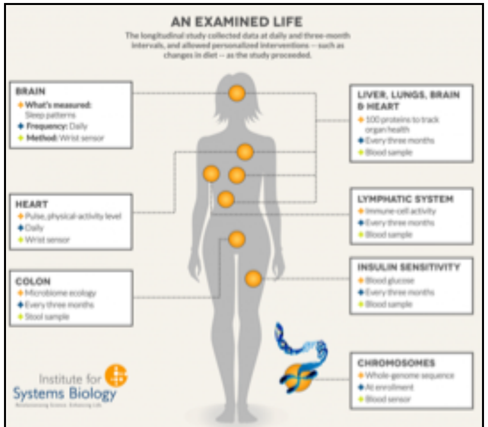
II. Finding Homologs

- 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



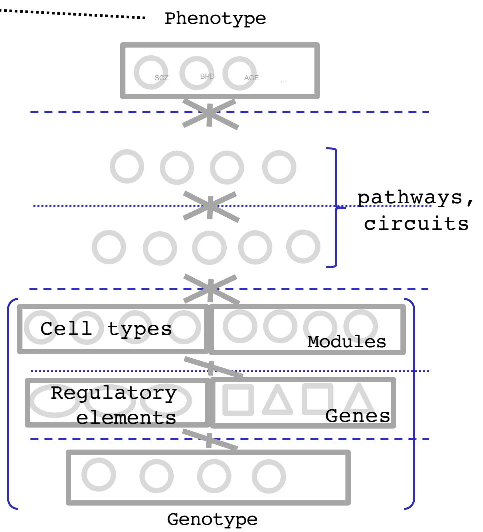
IV. Personal Genome Characterization

- 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 BJ. 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



*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)

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%**	-
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

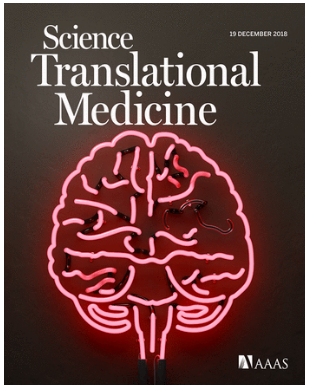
(in contrast to many ... e)

Thus, interested in devel ... ch:

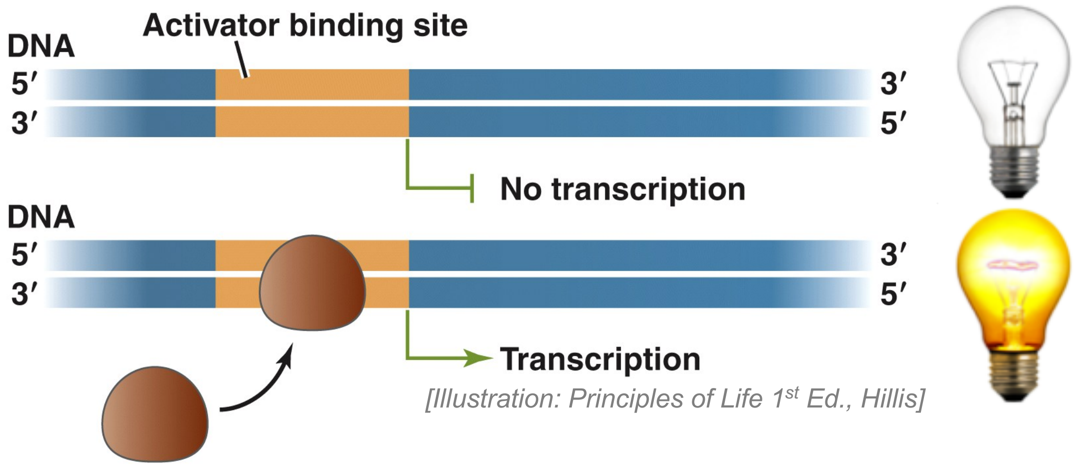
Use observations ... n latent structure.

Use the predictive ... to begin to suggest actors inv...

Recent Rollout in Science addressing this, involving many Yale Researchers

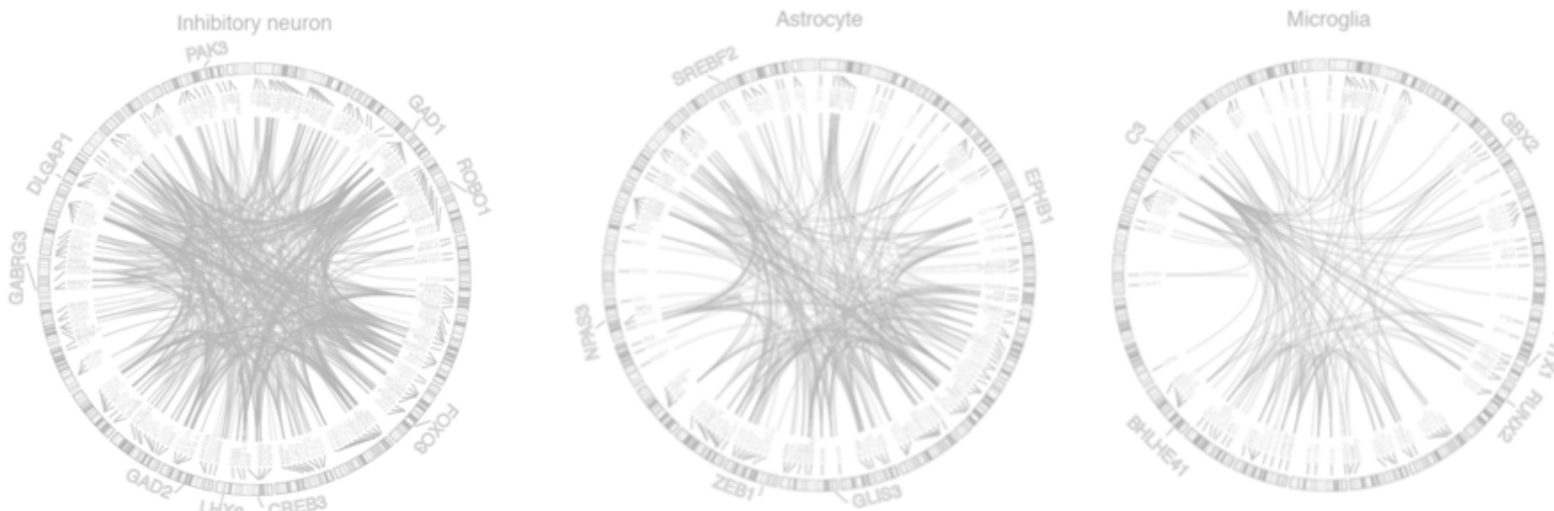


Developing a gene regulatory network for the human brain



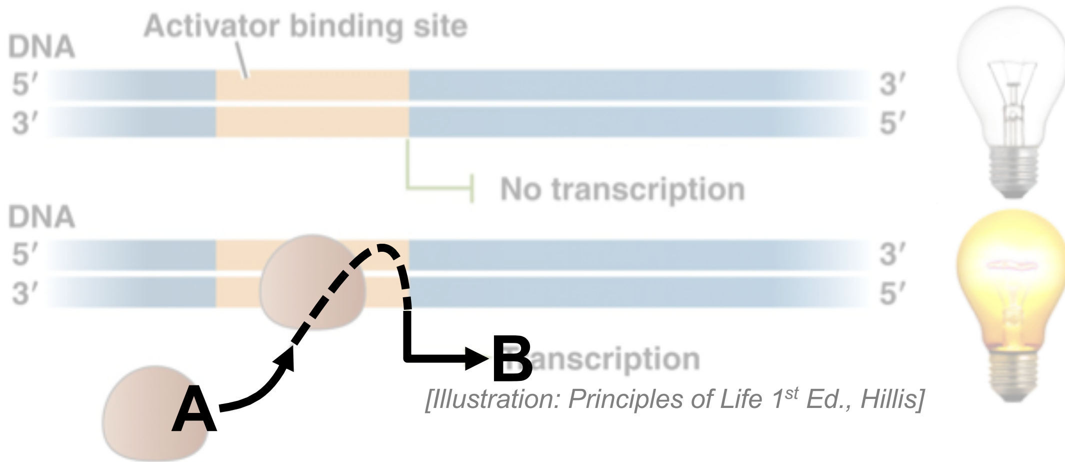
[Illustration: Principles of Life 1st Ed., Hillis]

Many such gene regulatory relationships form a network



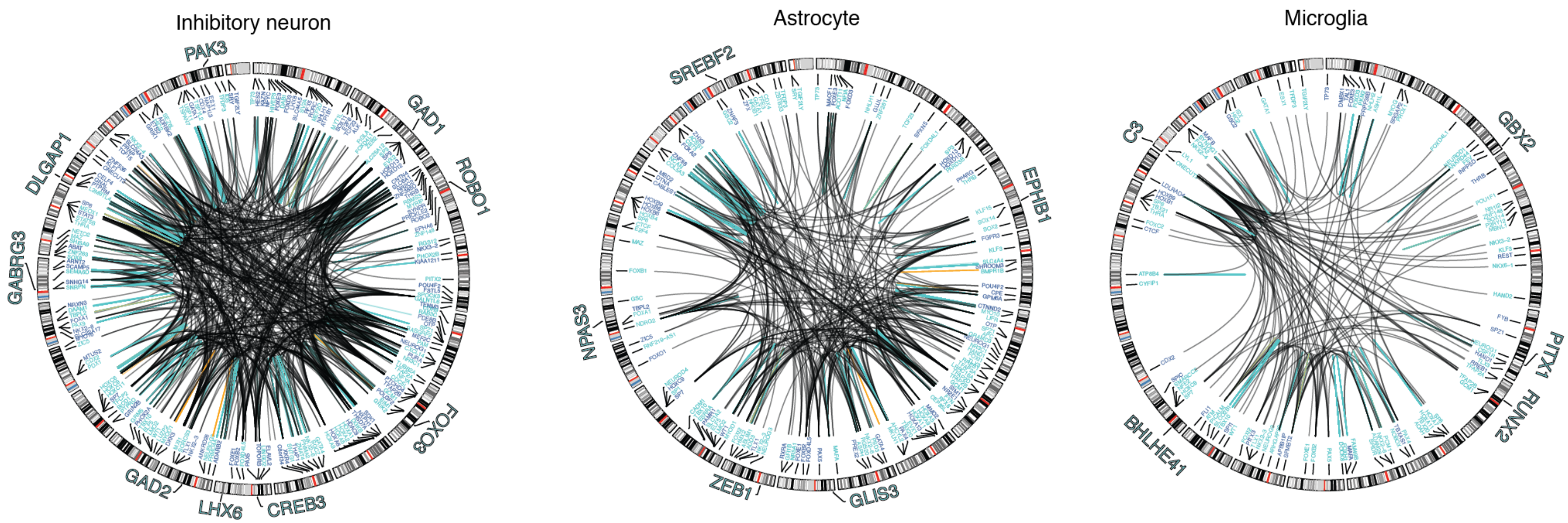
[Wang et al. ('18) Science]

Developing a gene regulatory network for the human brain



[Illustration: Principles of Life 1st Ed., Hillis]

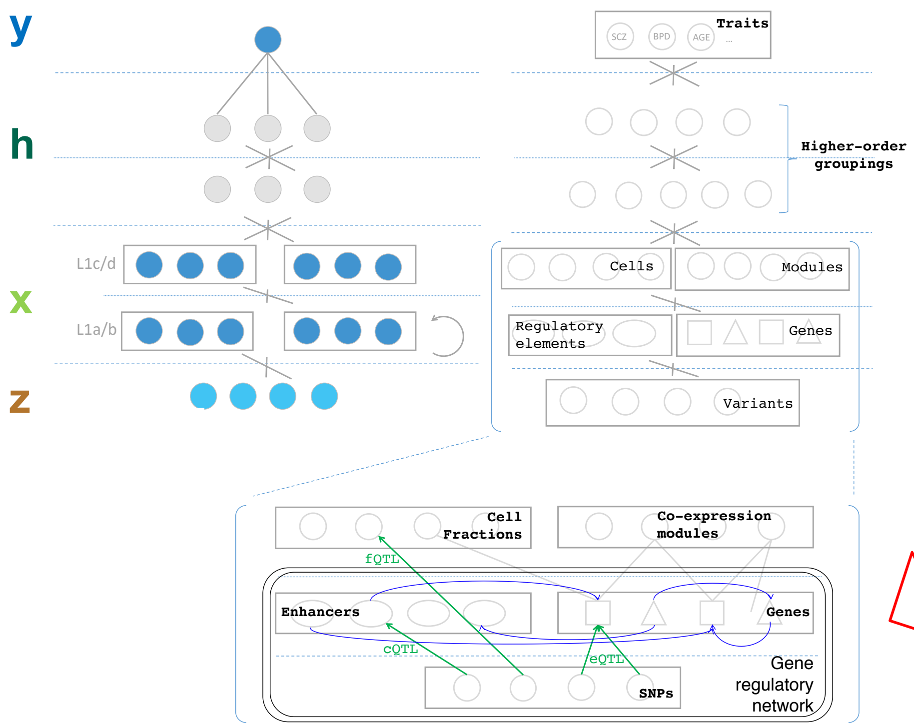
Many such gene regulatory relationships form a network

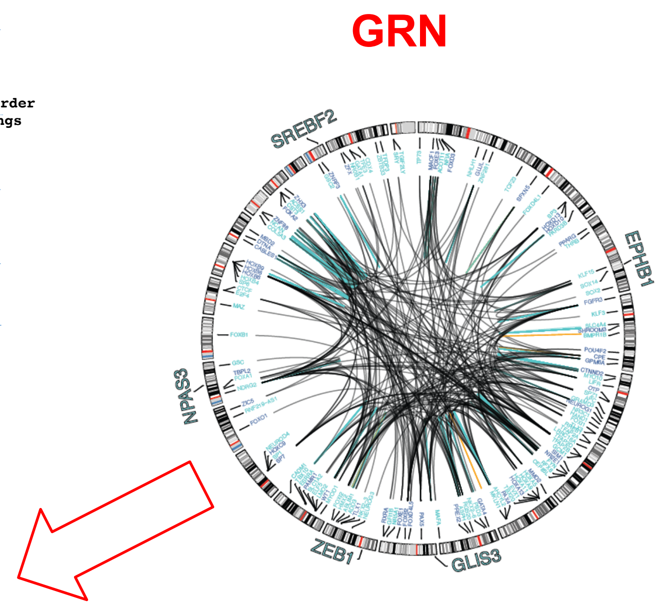


[Wang et al. ('18) Science]

Deep Structured Phenotype Network (DSPN)

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





y: phenotypes

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

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

z: genotypes (e.g., SNPs)

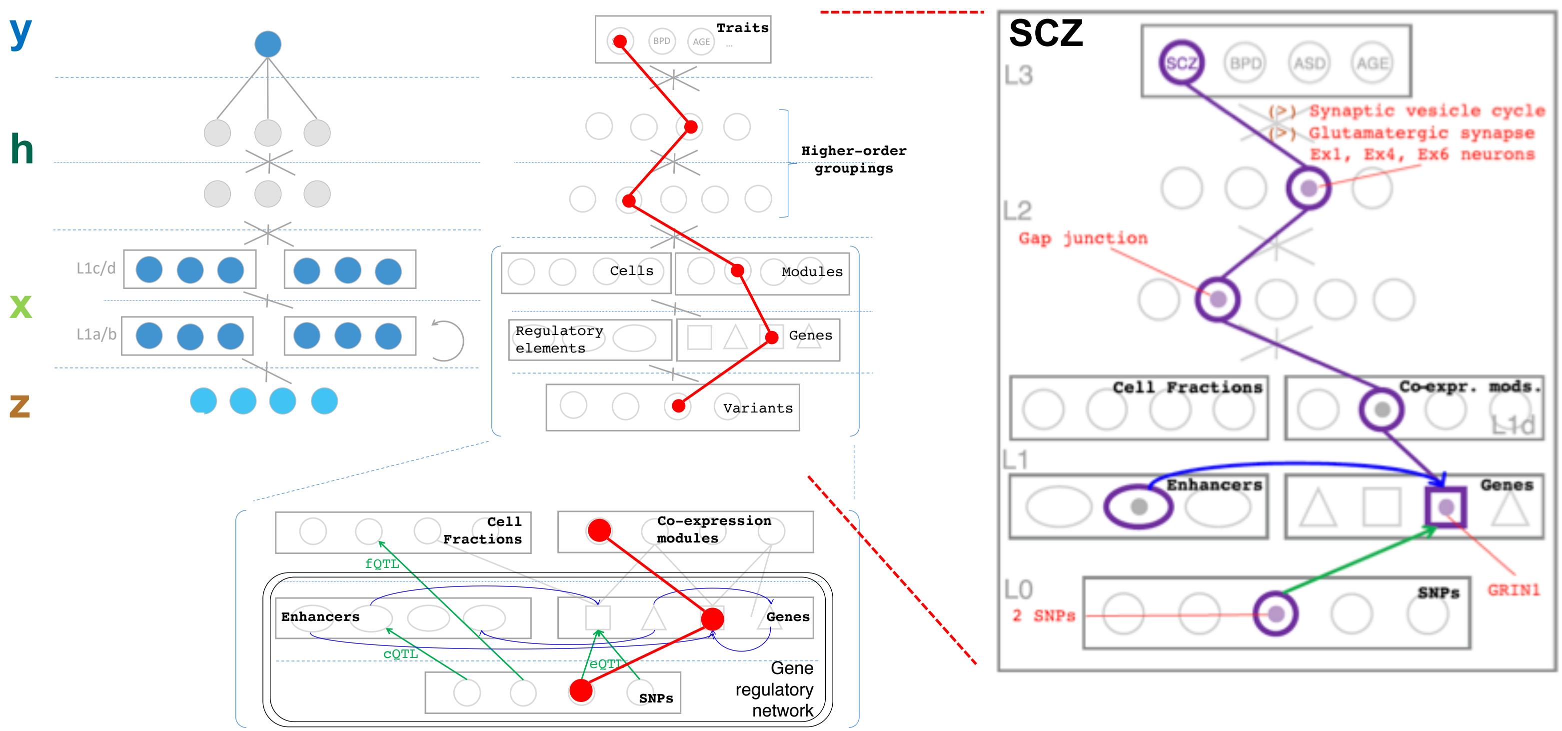
Deep Boltzmann Machine Energy model:

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

[Wang et al. ('18) Science]

Deep Structured Phenotype Network (DSPN)

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



y: phenotypes

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

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

z: genotypes (e.g., SNPs)

Deep Boltzmann Machine Energy model:

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

[Wang et al. ('18) Science]

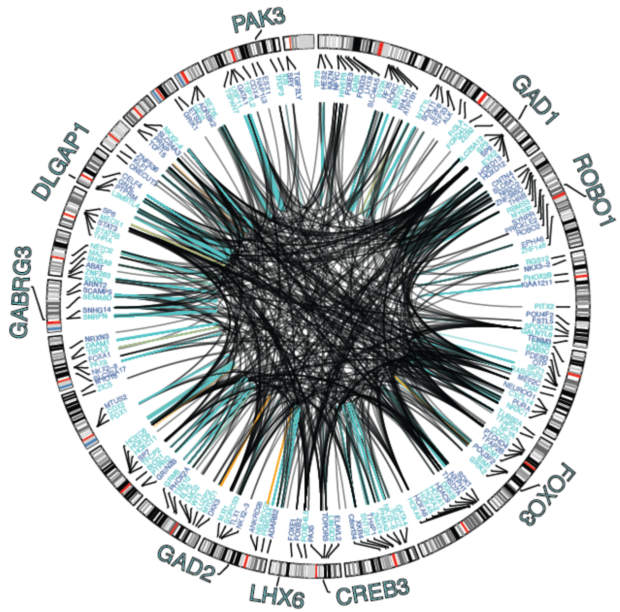
Genomics & Data Science

- **Background** on data science & the biomed. sub-domain
  - Placing genomics in this context -- **3V & 4M** frameworks
  - Types of data & integration betw. types + **Moore's law scaling**
- Main **practical heath applications**
  - Designing drugs, customizing treatments….
  - Focused example on **identifying drug targets** for neuropsych. diseases
- Another application: **"Bioinspired" insights on systems** design
  - Reducing bottlenecks in the middle of a network hierarchy
  - Different patterns of connectivity & constraint in natural v engineered systems
- Mining the **data exhaust** (from personal genomics)
  - SOS
  - **Privacy**

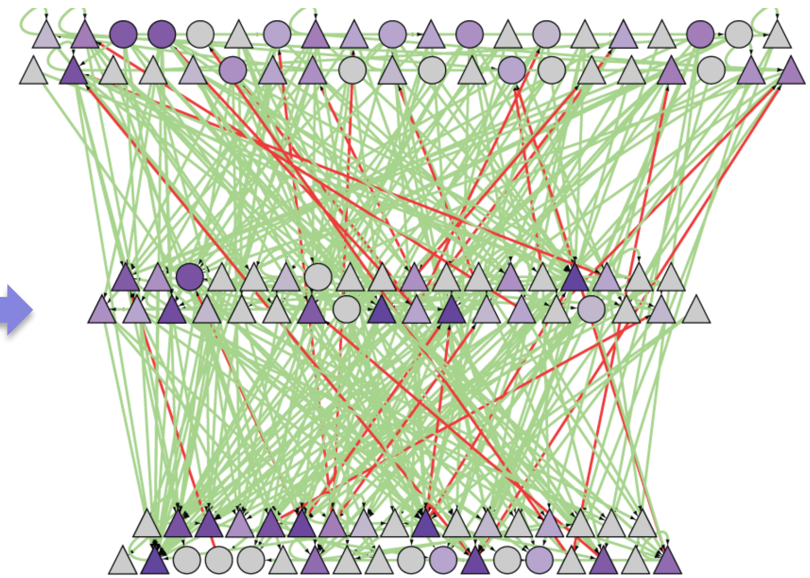
Transforming a Regulatory Network into a Hierachy

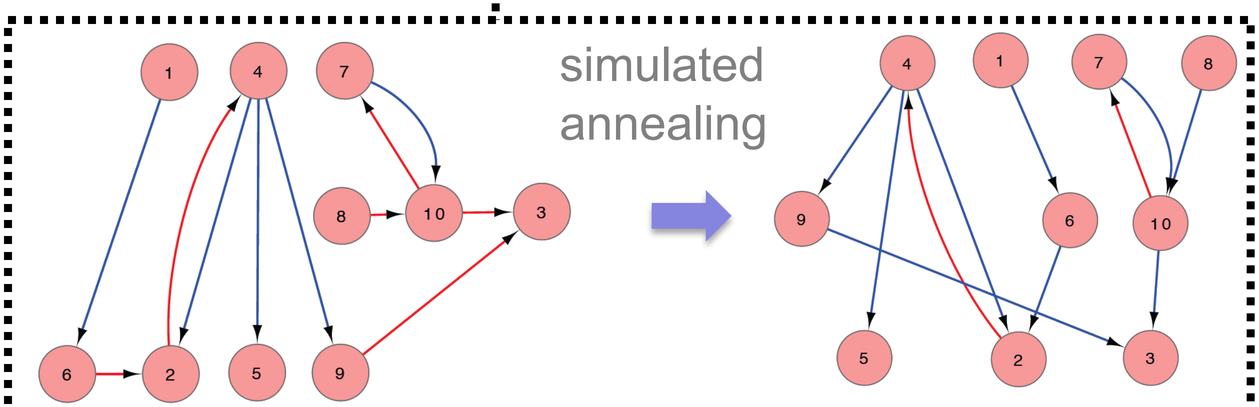
Regulatory Network

Inhibitory neuron



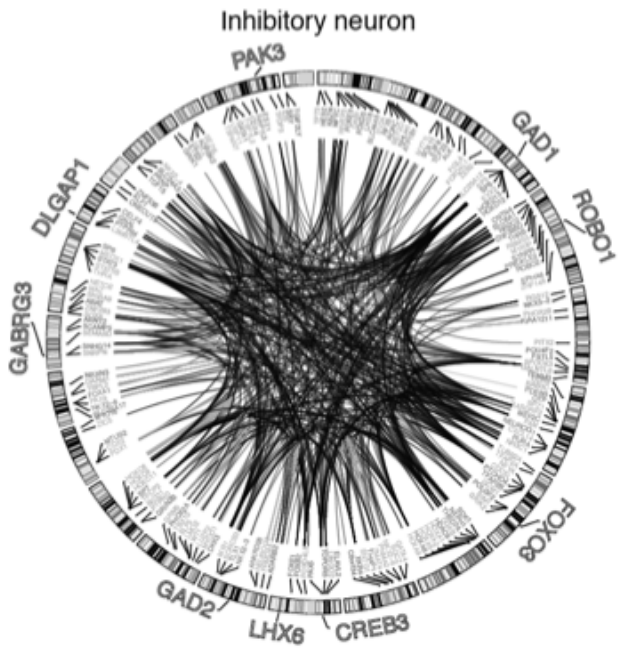
Network Hierarchy

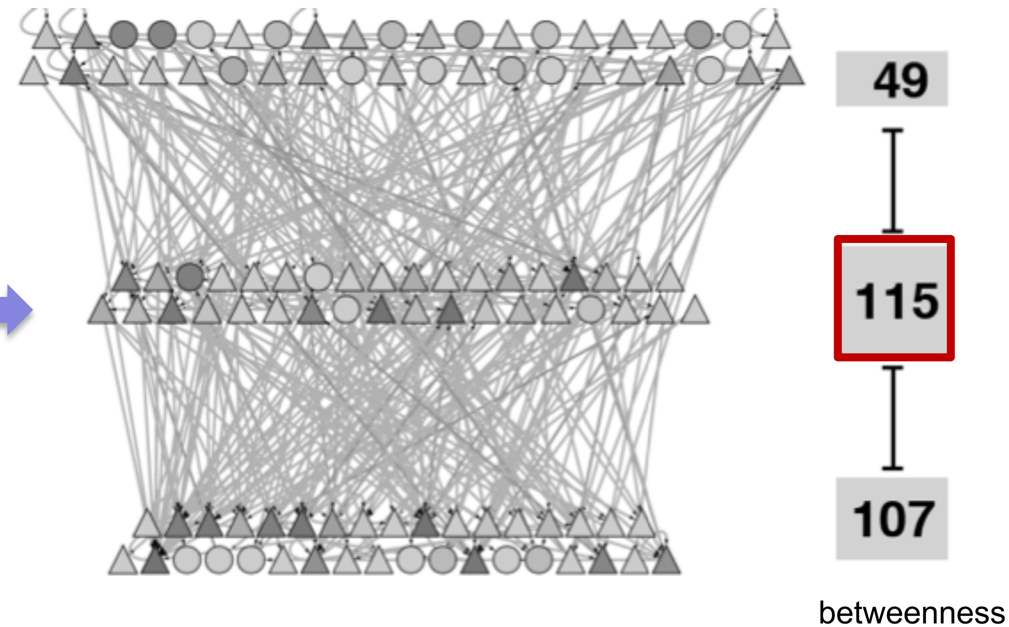


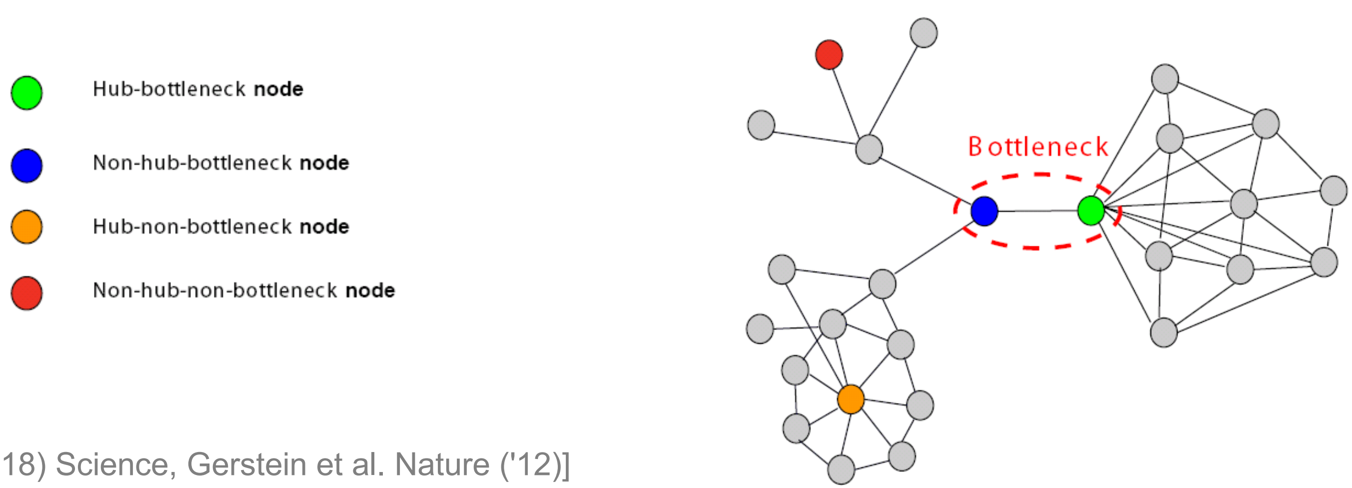


[Wang et al. ('18) Science, Gerstein et al. Nature ('12)]

Mid-level of the hierarchy has many high-connectivity bottlenecks

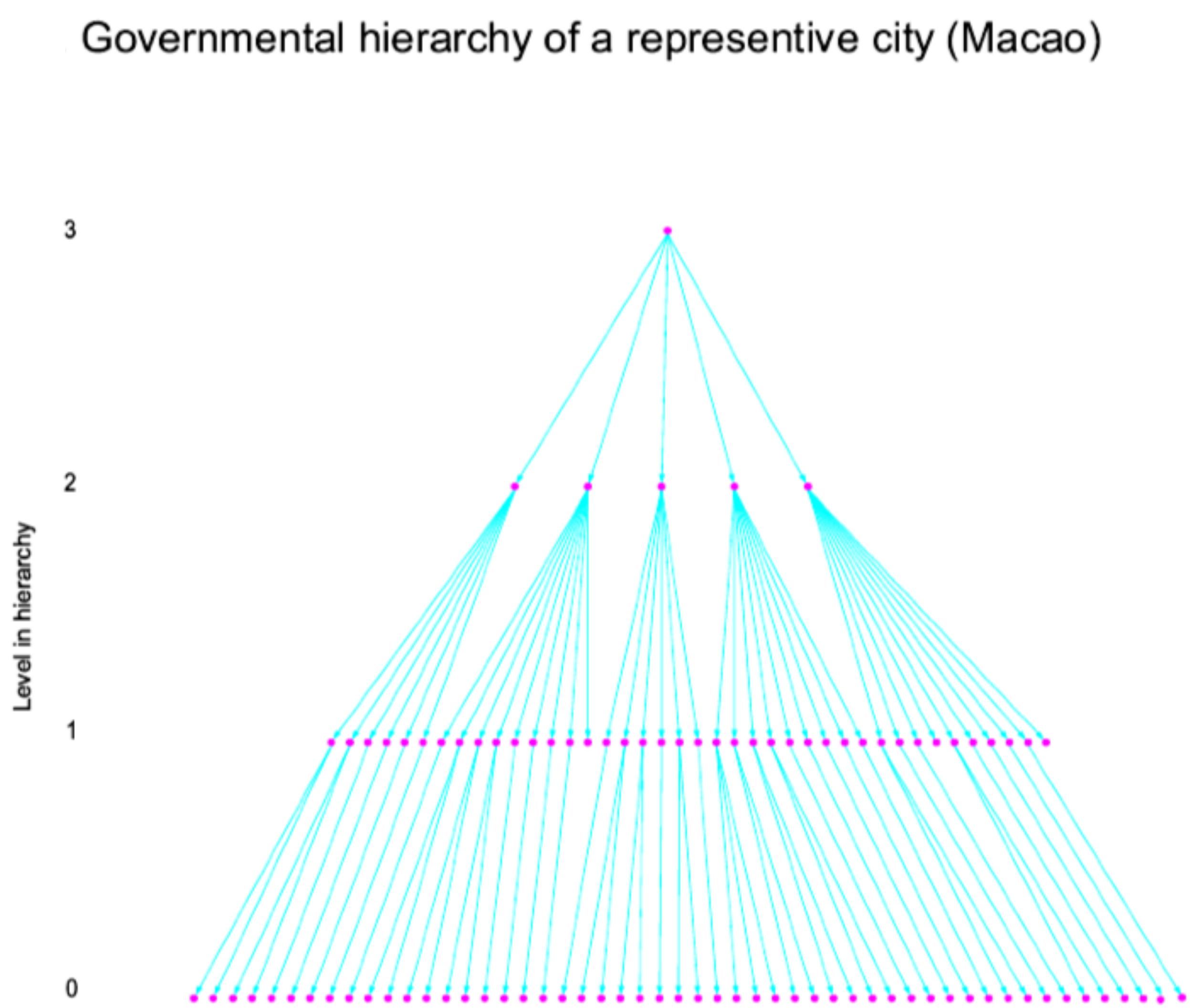


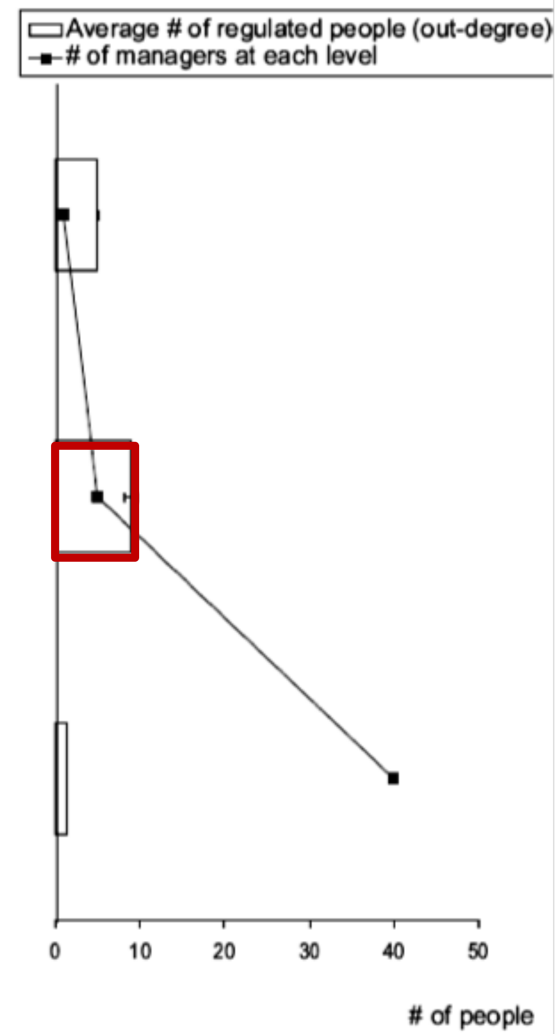




[Wang et al. ('18) Science, Gerstein et al. Nature ('12)]

Governmental hierachies exhibit even more pronounced mid-level bottlenecks

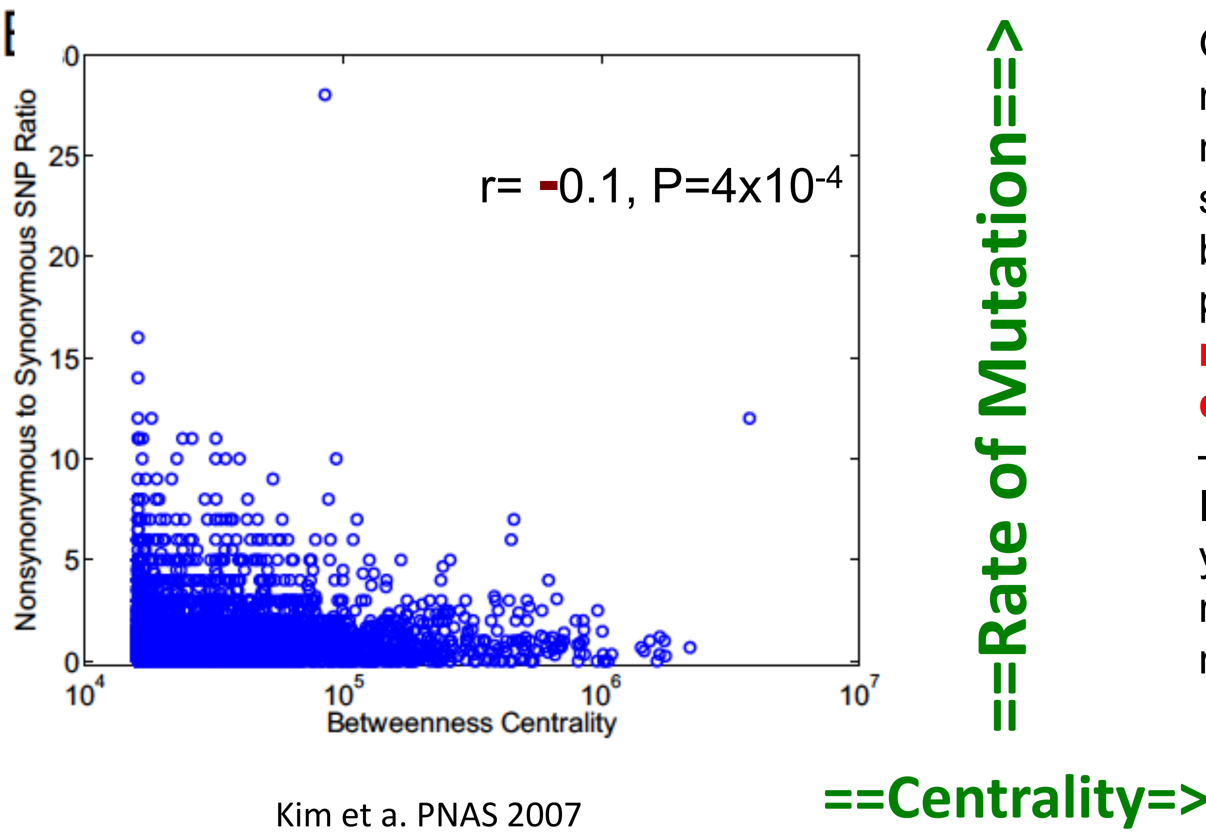




[Yu et al., PNAS ('06)]

[Yu et al., *PNAS* '06]

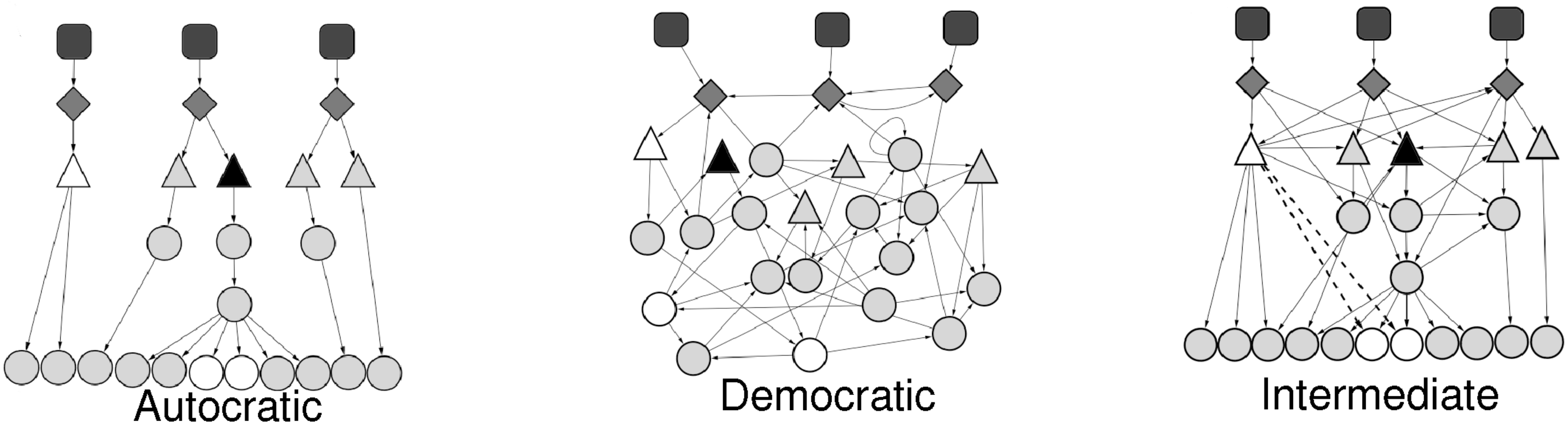
Connectivity v. Constraint: in biological networks, less mutations tolerated at hubs & bottlenecks

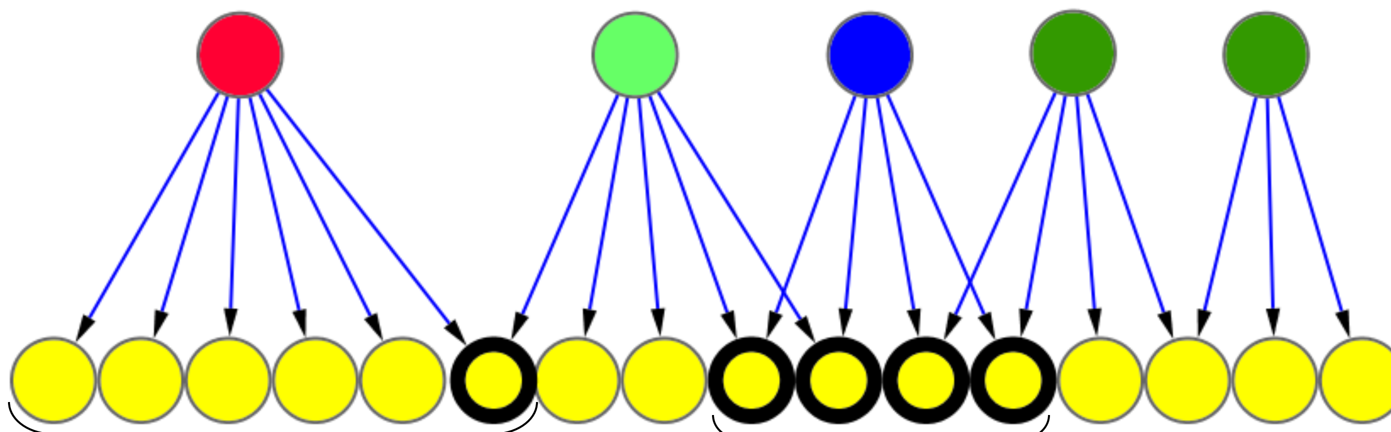


Kim et a. PNAS 2007

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** – e.g. Protein-Protein Interaction Networks (human, yeast, E coli, worm, fly), miRNA networks, regulatory networks, &c

Biological Networks tend to have more "democratic" hierarchies, easing bottlenecks, than many social ones





Degree of Collaboration

1/6=0.16
"Autocratic"

4/4=1
"Democratic"

[Bhardwaj et al., PNAS (2010)]

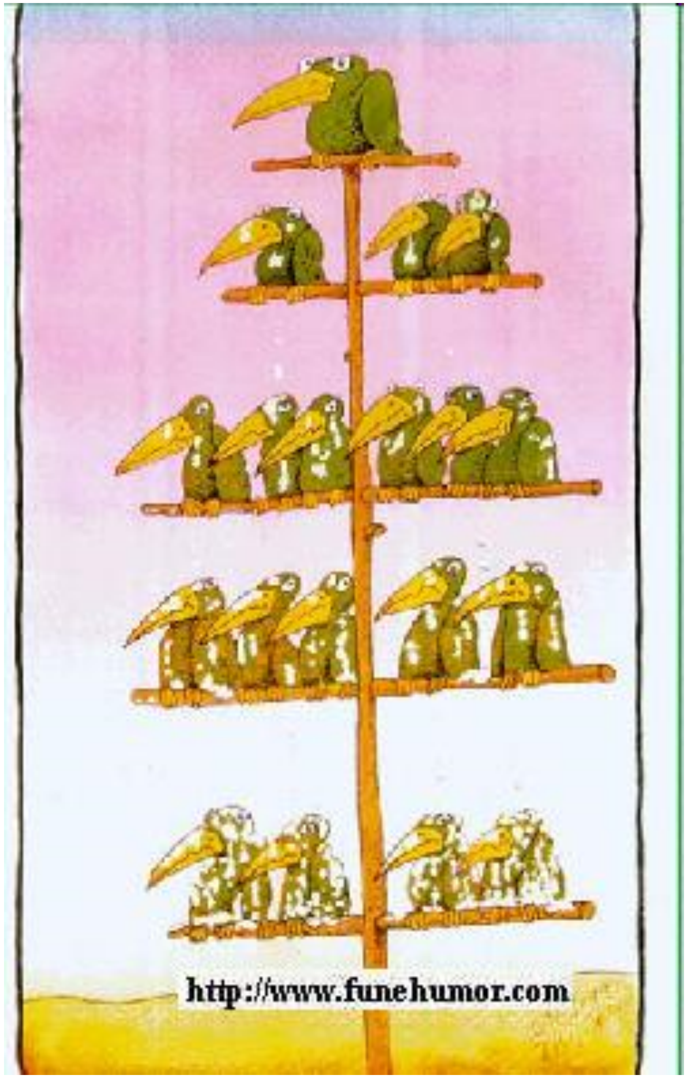
Middle Managers Interact the Most in Efficient Corporate Settings

- Floyd, S. W. et al (1992)
  Middle management involvement in strategy and its association with strategic type *Strategic Management Journal* 13, 153-167.

- Woodward, J. (1982) Industrial Organization: Theory and Practice (Oxford University Press, Oxford).

- Floyd, S. W. et al (1993)
  Dinosaurs or Dynamos? Recognizing Middle Management's Strategic Role
  The Academy of Management Executive 8, 47-57.

- Floyd, S. W. et al (1997)
  Middle management's strategic influence and organizational performance
  Journal of Management Studies 34, 465-485.



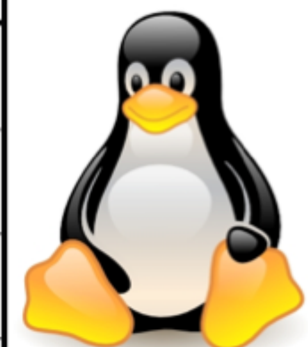
[Bhardwaj et al., PNAS (2010), in press]

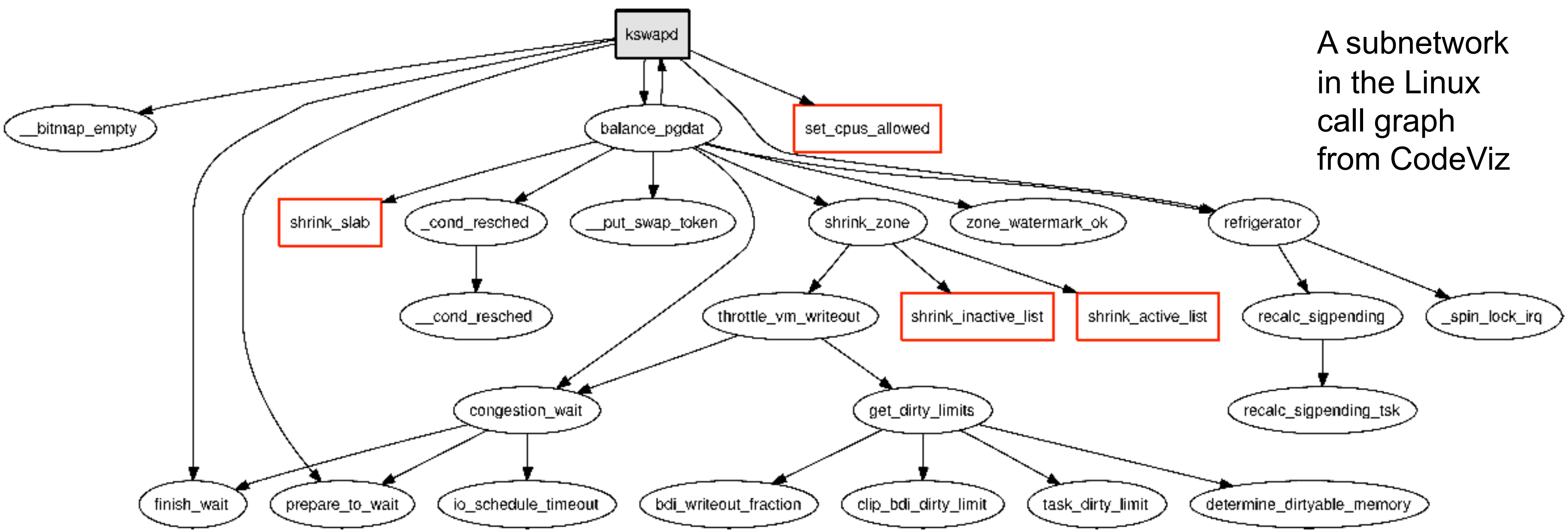
Genomics & Data Science

- **Background** on data science & the biomed. sub-domain
  - Placing genomics in this context -- **3V & 4M** frameworks
  - Types of data & integration betw. types + **Moore's law scaling**
- Main **practical heath applications**
  - Designing drugs, customizing treatments….
  - Focused example on **identifying drug targets** for neuropsych. diseases
- Another application: **"Bioinspired" insights on systems** design
  - Reducing bottlenecks in the middle of a network hierarchy
  - Different patterns of connectivity & constraint in natural v engineered systems
- Mining the **data exhaust** (from personal genomics)
  - SOS
  - **Privacy**

Transcriptional regulatory network vs Linux call graph

		E. coli transcriptional regulatory network	Linux call graph
Basic properties of systems	Nodes	Genes (TFs & targets)	Functions (subroutines)
	Edges	Transcriptional regulation	Function calls
	External constraints	Natural environment	Hardware architecture, customer requirements
	Origin of evolutionary changes	Random mutation & natural selection	Designers' fine tuning

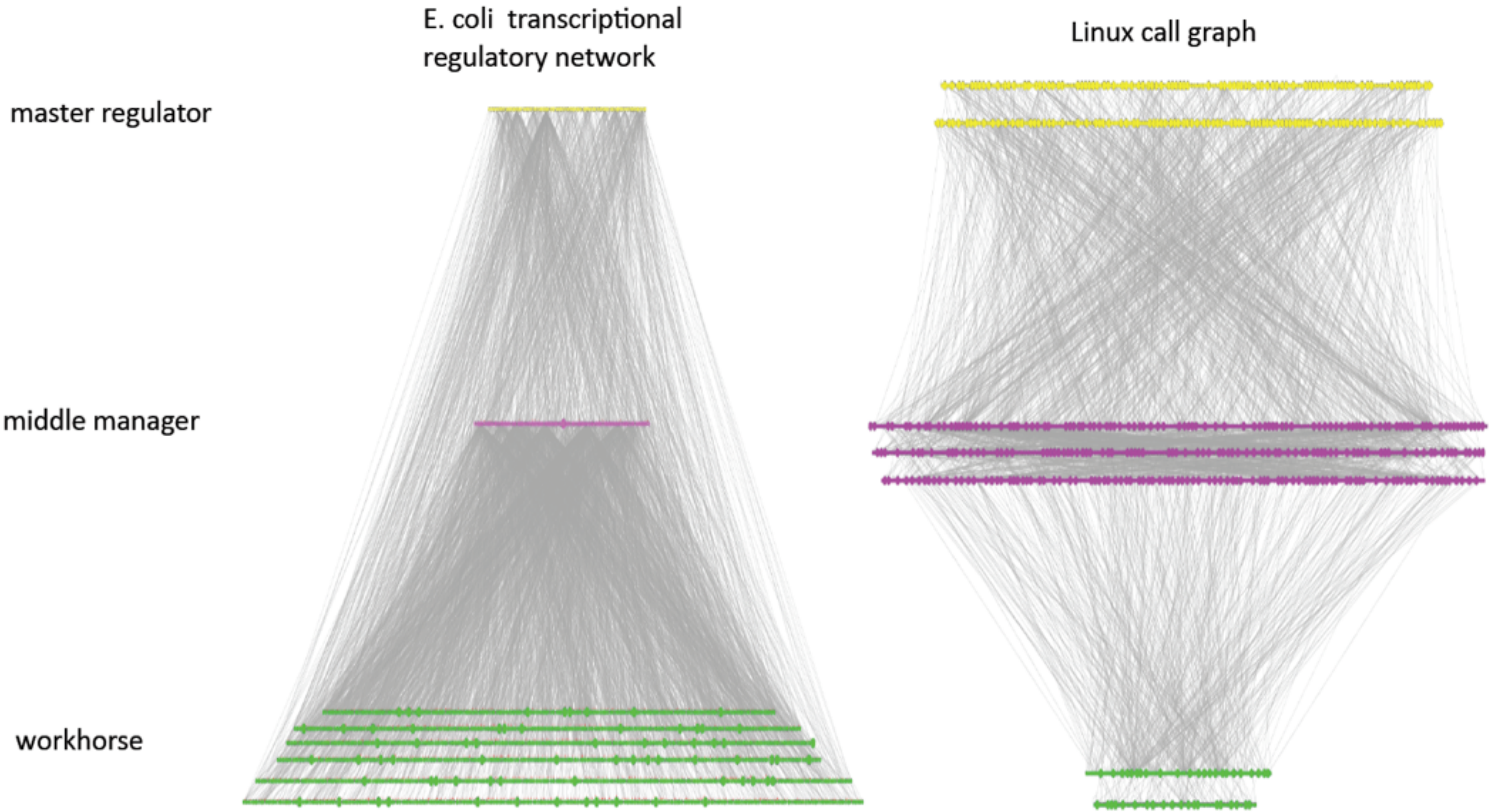




A subnetwork in the Linux call graph from CodeViz

[Yan et al., *PNAS* '10]

Dominated by “out hubs” - crp

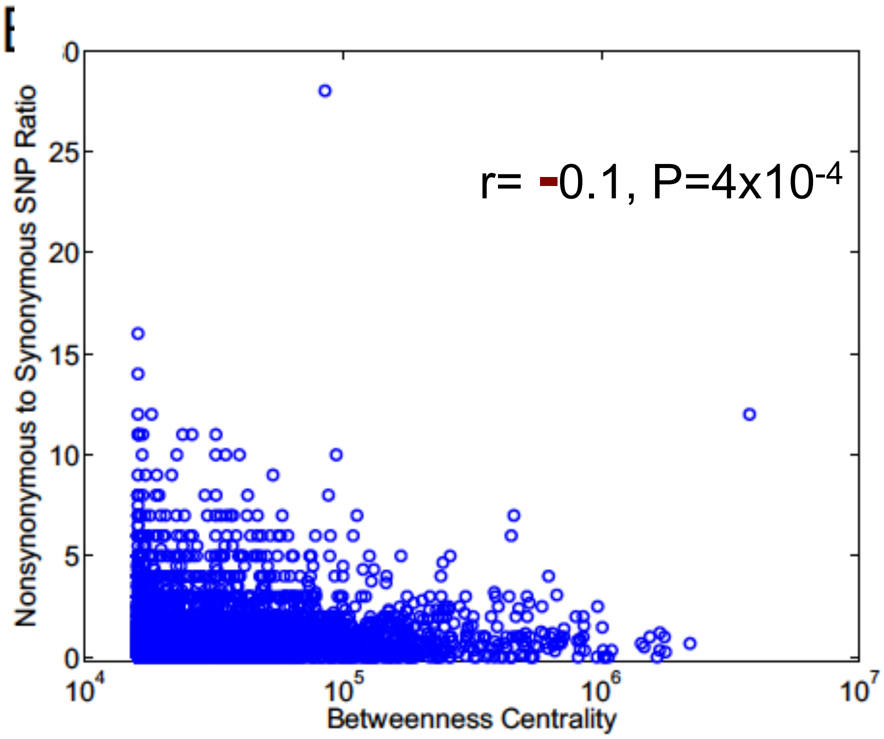


Dominated by “in hubs” - printf

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

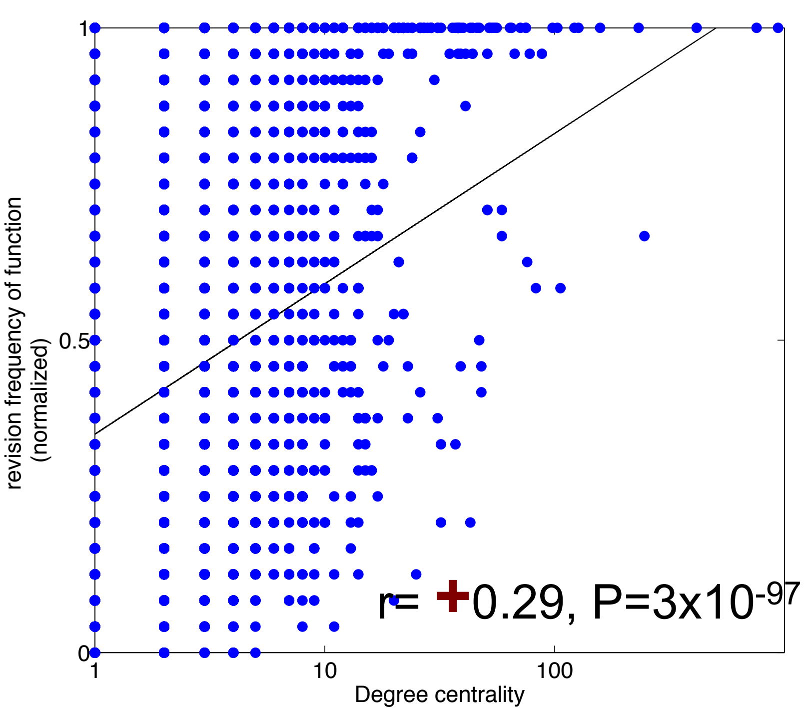
Connectivity v. Constraint: differences betw. biological & technological networks

More Connectivity, More Constraint : A theme borne out in many evolutionary studies of biological network



Kim et a. PNAS 2007

Variation in call graph measured from number of edits to Linux source



Yan KK et al. PNAS 2010

==Rate of Mutation==>

==Centrality=>

Perspectives on Random Change v Intelligent Design

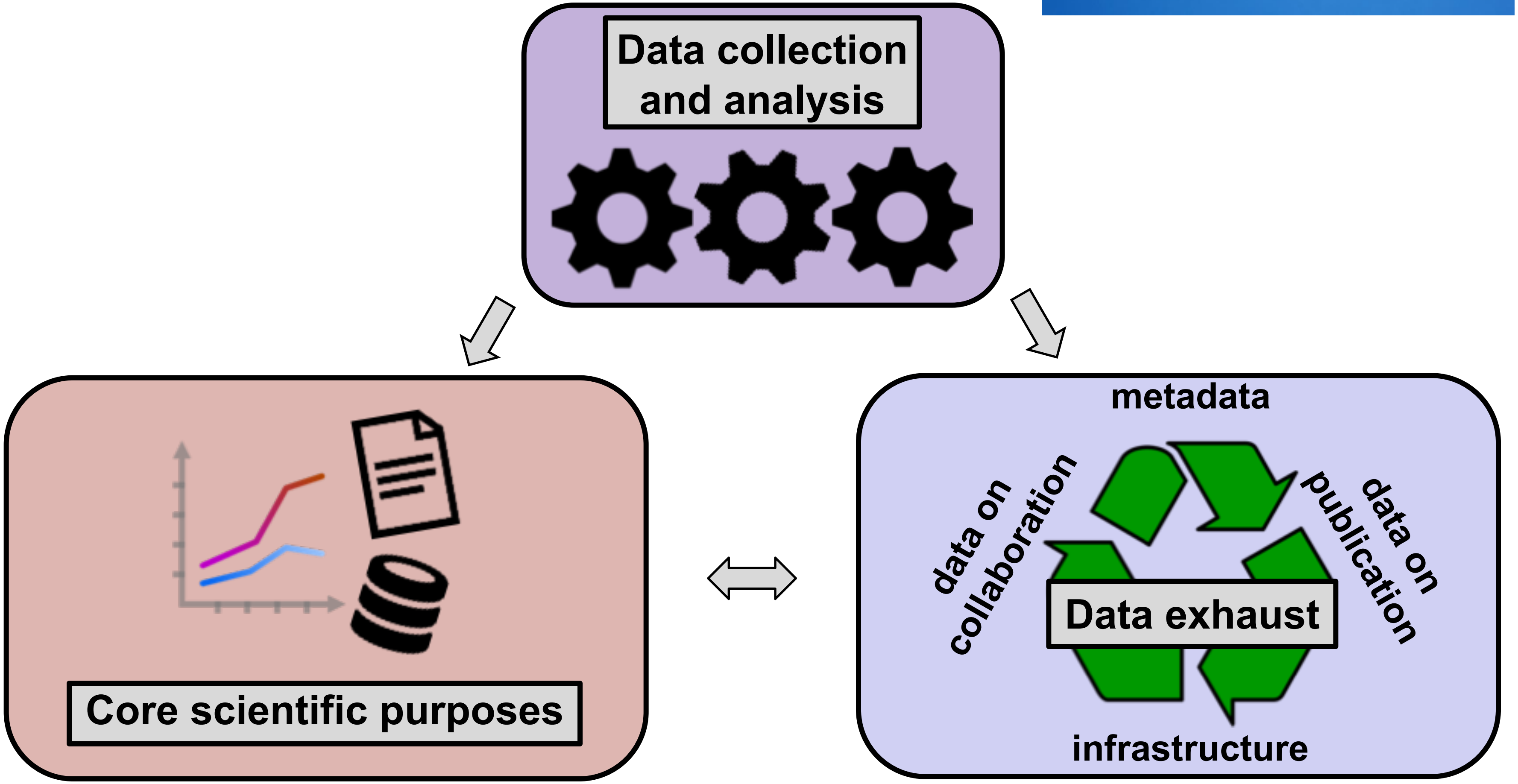
- Central points = hubs & bottlenecks
- If changes random, best not to put them in central pts.
- If changes made rationally, can put them into central pts.
  - Moreover, good to do this, as these more often used – i.e more efficient
  - Why there's so much GWB construction

Genomics & Data Science

- **Background** on data science & the biomed. sub-domain
  - Placing genomics in this context -- **3V & 4M** frameworks
  - Types of data & integration betw. types + **Moore's law scaling**
- Main **practical heath applications**
  - Designing drugs, customizing treatments….
  - Focused example on **identifying drug targets** for neuropsych. diseases
- Another application: **"Bioinspired" insights on systems** design
  - Reducing bottlenecks in the middle of a network hierarchy
  - Different patterns of connectivity & constraint in natural v engineered systems
- Mining the **data exhaust** (from personal genomics)
  - SOS
  - **Privacy**

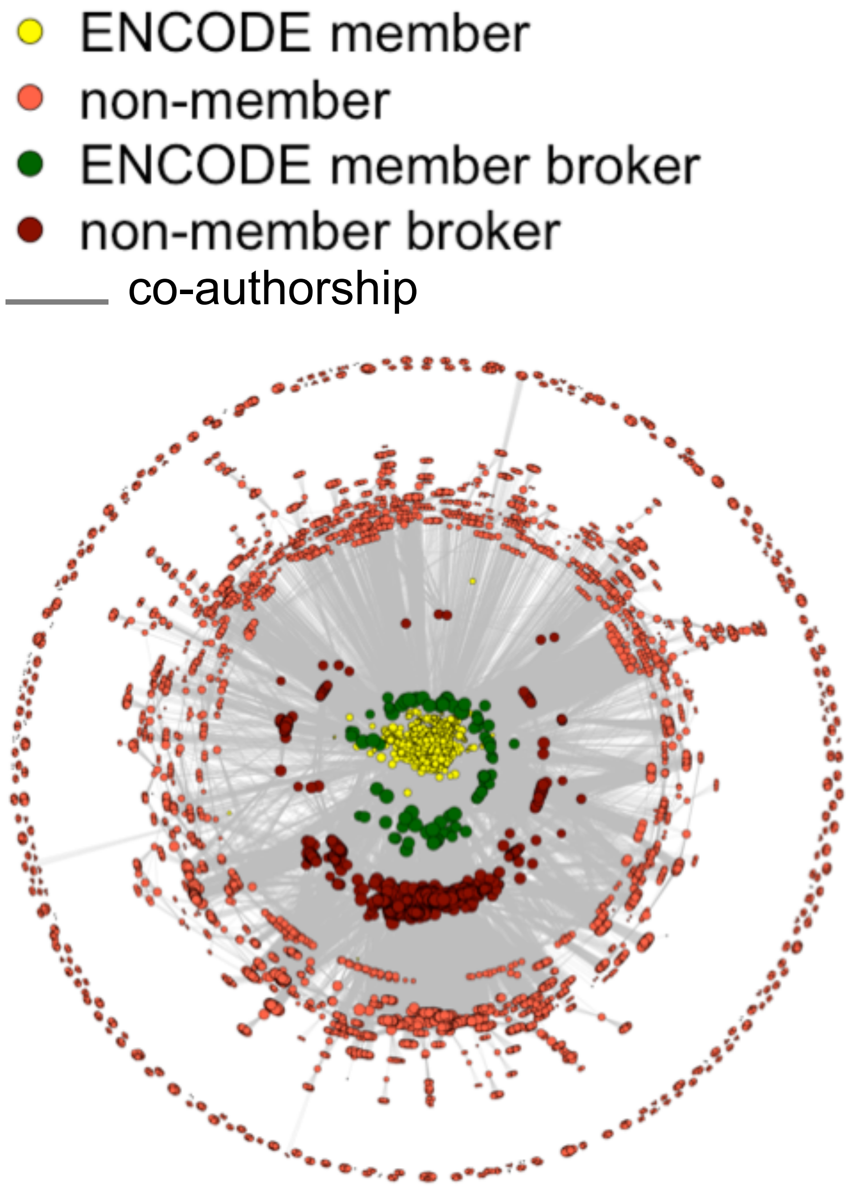
Data Exhaust

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



[photos: wikipedia/wikimedia]

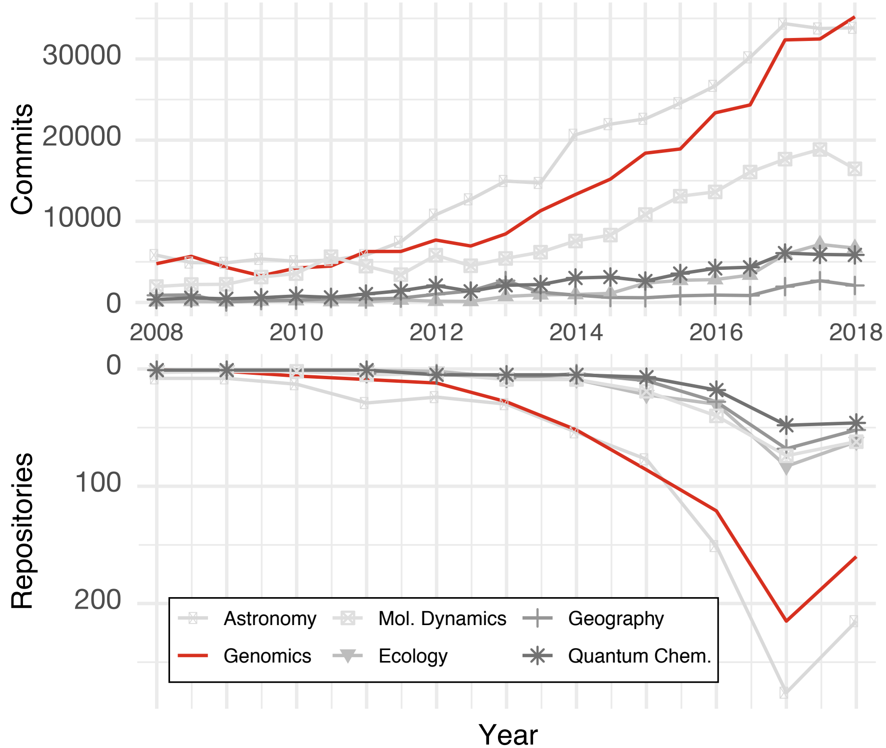
Exhaust Mining Application:
Using Science to Study Science (SOS)



- 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

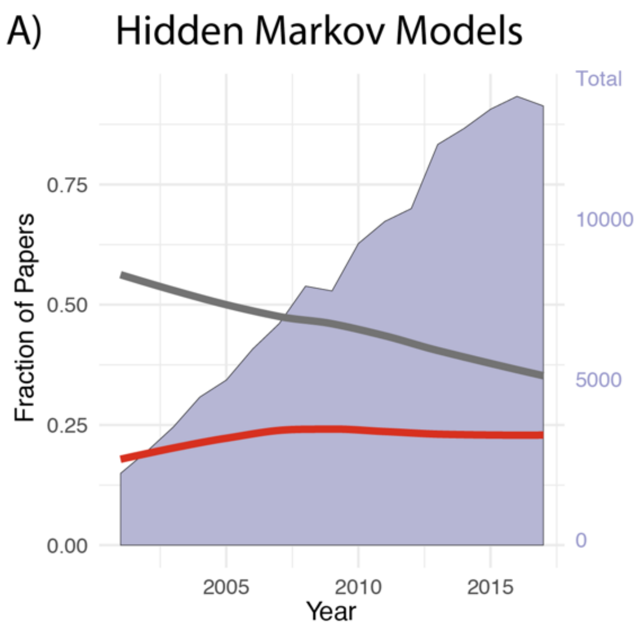
[Wang et al., TIG ('16)]

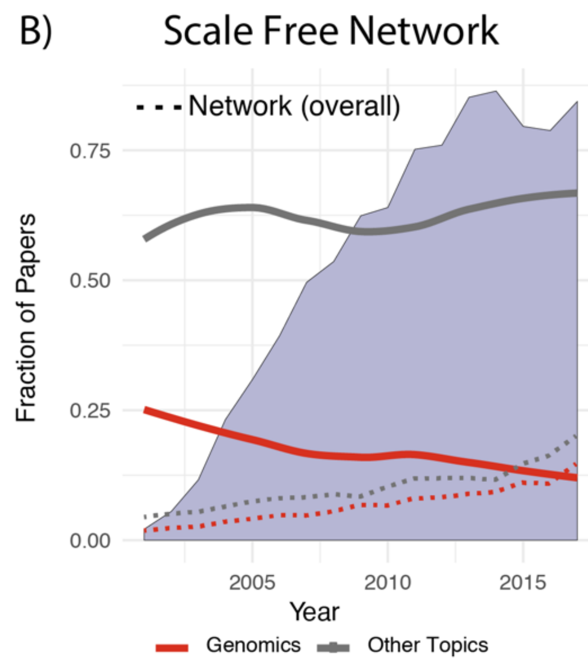
Quantifying Imports & exports from/to Genomics & other Data Science subfields



Genomics adoption of open source

Genomics contribution of HMM and Network papers





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

Genomics & Data Science

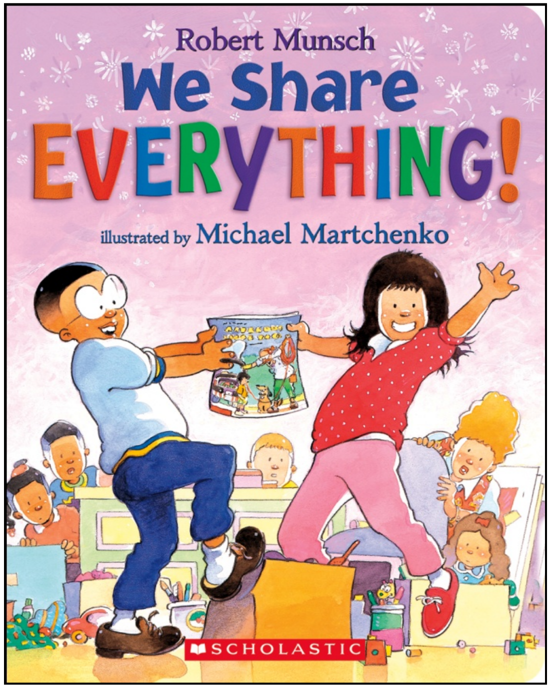
- **Background** on data science & the biomed. sub-domain
  - Placing genomics in this context -- **3V & 4M** frameworks
  - Types of data & integration betw. types + **Moore's law scaling**
- Main **practical heath applications**
  - Designing drugs, customizing treatments….
  - Focused example on **identifying drug targets** for neuropsych. diseases
- Another application: **"Bioinspired" insights on systems** design
  - Reducing bottlenecks in the middle of a network hierarchy
  - Different patterns of connectivity & constraint in natural v engineered systems
- Mining the **data exhaust** (from personal genomics)
  - SOS
  - **Privacy**

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

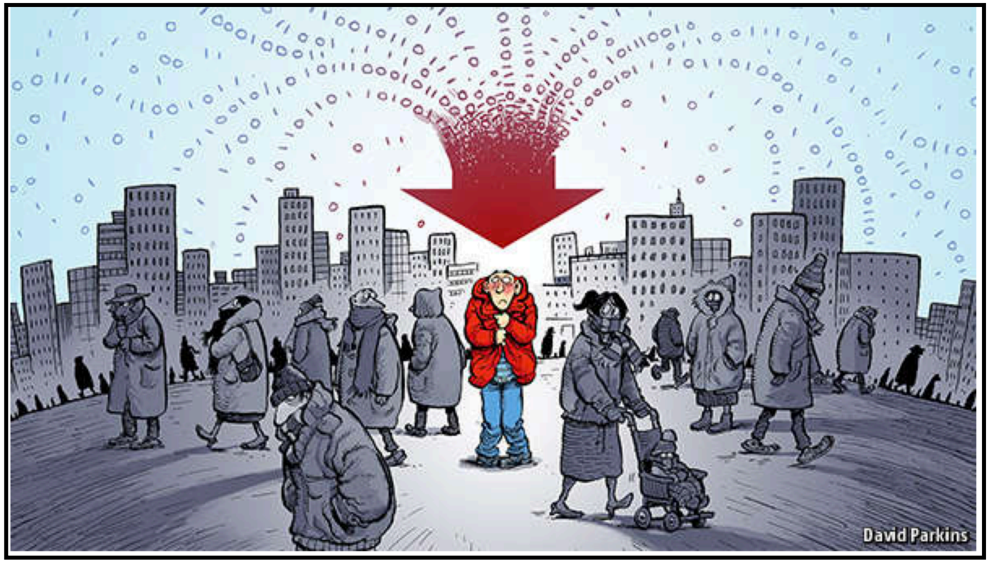
Genetic Exceptionalism :
The Genome is very fundamental data, potentially very revealing about one's identity & characteristics
Personal Genomic info. essentially meaningless currently but will it be in 20 yrs? 50 yrs?
Genomic sequence very revealing about one's children. Is true consent possible?
Once 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)]

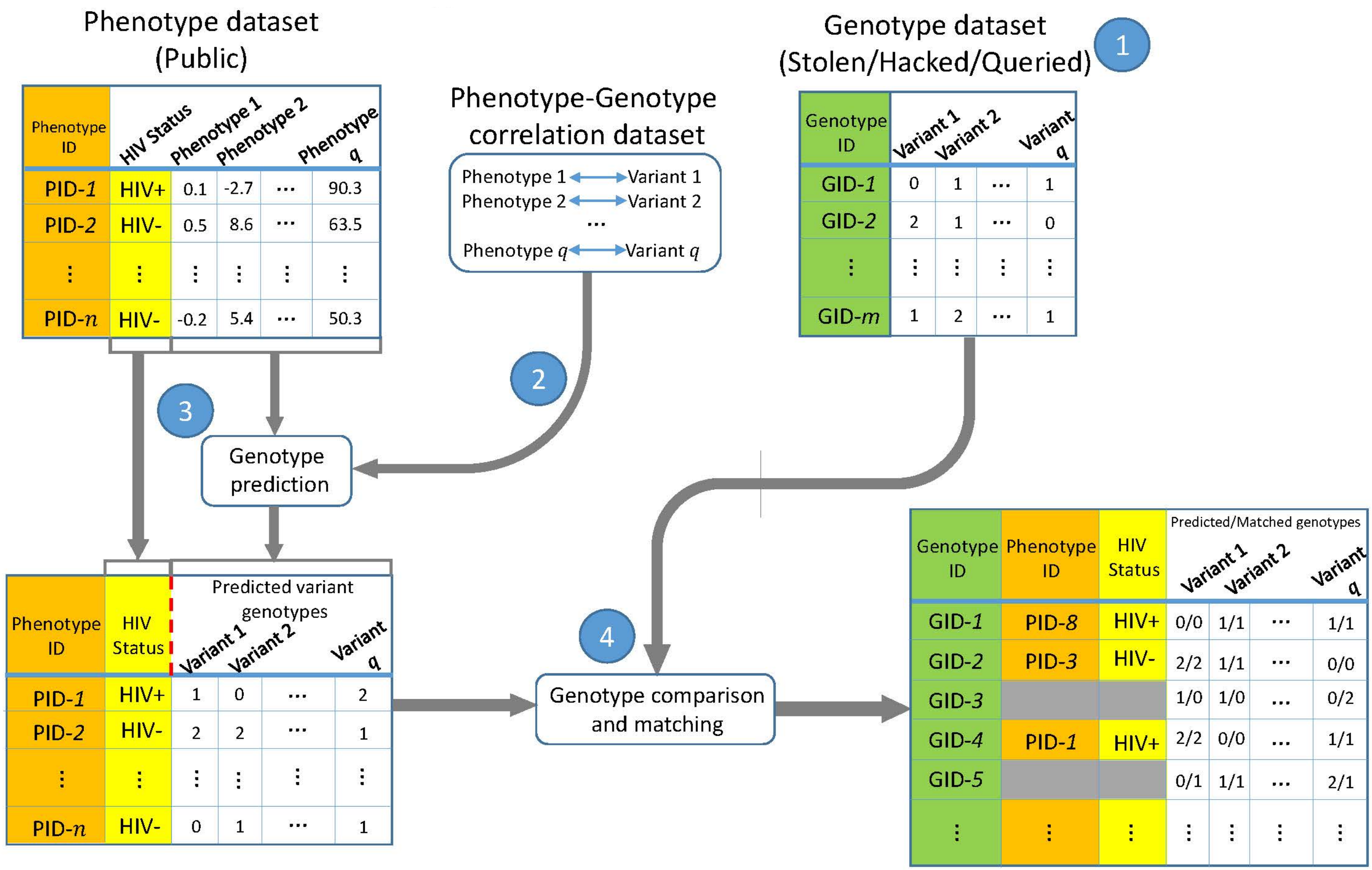
Current Social & Technical Solutions: The quandary where are now

- 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

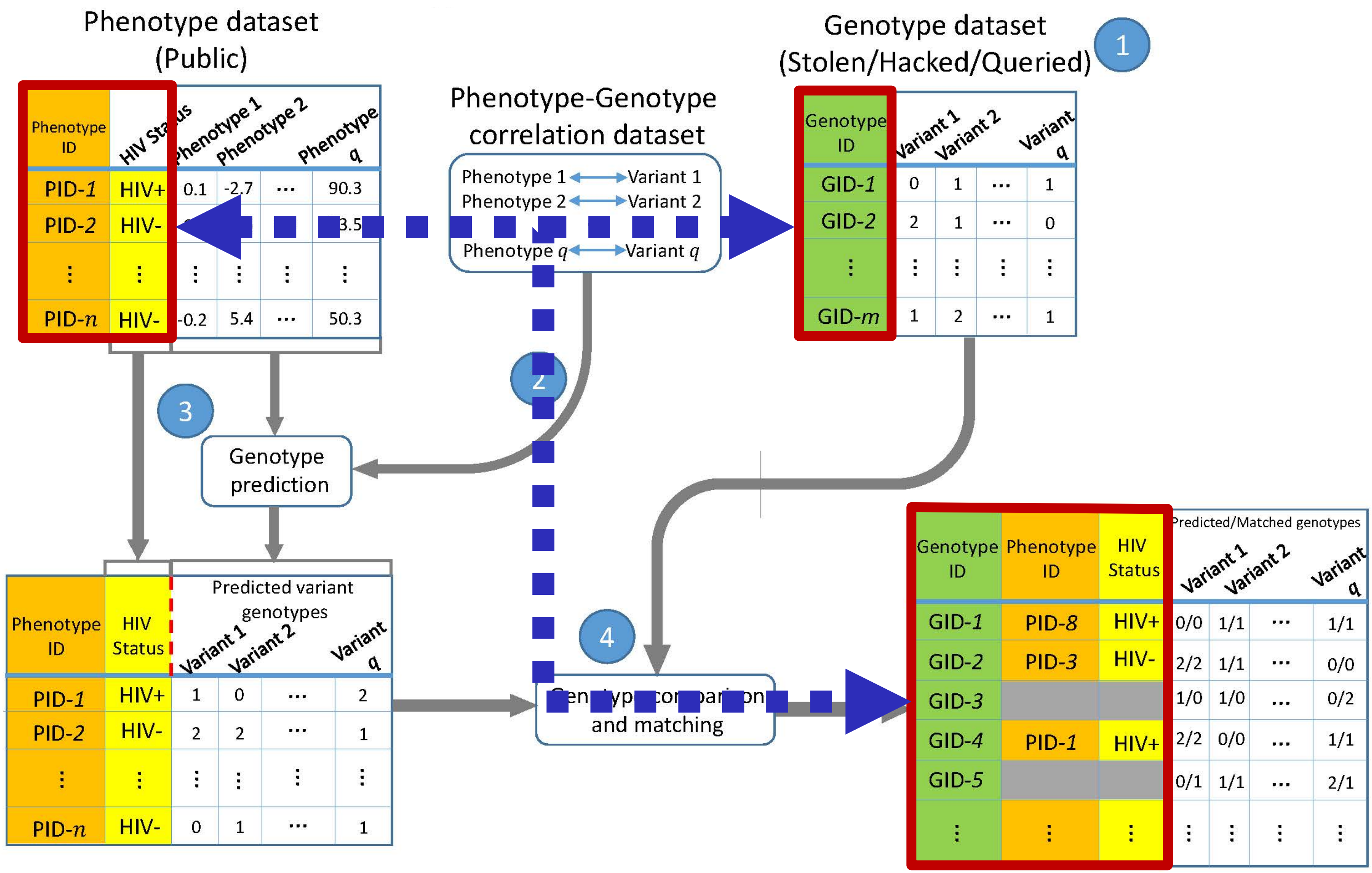
[Greenbuam et al ('04), Nat. Biotech; Greenbaum & Gerstein ('13), The Scientist; Photo: futurism.com]

Linking Attack Scenario



[Harmanciet al. Nat. Meth. (in revision)]

Linking Attack Scenario



[Harmanciet al. Nat. Meth. (in revision)]

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

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

- IMDB users are public

- NetFLIX and IMdB moves are public

Linking Attacks: Case of Netflix Prize

NETFLIX ⟷ IMDb

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

Genomics & Data Science

- **Background** on data science & the biomed. sub-domain
  - Placing genomics in this context -- **3V & 4M** frameworks
  - Types of data & integration betw. types + **Moore's law scaling**
- Main **practical heath applications**
  - Designing drugs, customizing treatments….
  - Focused example on **identifying drug targets** for neuropsych. diseases
- Another application: **"Bioinspired" insights on systems** design
  - Reducing bottlenecks in the middle of a network hierarchy
  - Different patterns of connectivity & constraint in natural v engineered systems
- Mining the **data exhaust** (from personal genomics)
  - SOS
  - **Privacy**

Genomics & Data Science

- **Background** on data science & the biomed. sub-domain
  - Placing genomics in this context -- **3V & 4M** frameworks
  - Types of data & integration betw. types + **Moore's law scaling**
- Main **practical heath applications**
  - Designing drugs, customizing treatments….
  - Focused example on **identifying drug targets** for neuropsych. diseases
- Another application: **"Bioinspired" insights on systems** design
  - Reducing bottlenecks in the middle of a network hierarchy
  - Different patterns of connectivity & constraint in natural v engineered systems
- Mining the **data exhaust** (from personal genomics)
  - SOS
  - **Privacy**

Genomics
&
Data Science

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