Bioinformatics

Péter Antal

Computational Biomedicine (Combine) workgroup
Department of Measurement and Information Systems,
Budapest University of Technology and Economics

Méréstechnika és
Információs Rendszerek
Tanszék

MŰEGYETEM 1782

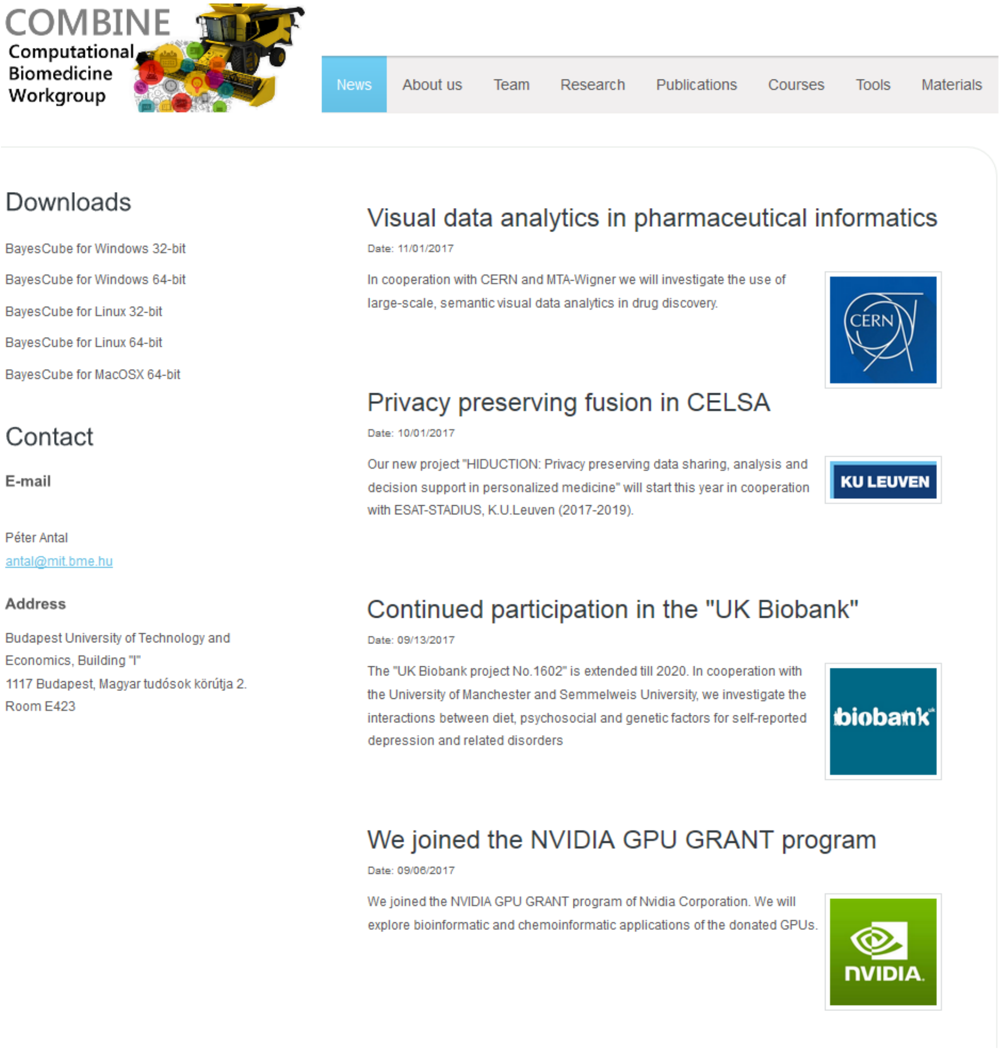
Course info

- Course site
  - https://www.mit.bme.hu/oktatas/targyak/vimiav10
- Lecturer
  - Péter Sárközy psarkozy@mit.bme.hu
  - Péter Antal, antal@mit.bme.hu
  - Bence Bruncsics bruncsics@mit.bme.hu
  - Bence Bolgár, bolgar@mit.bme.hu
- Schedule
  - Tuesday, Thursday 12.15-13.45, IE320, building I, wing E, 3rd floor
- Contact hour
  - By appointment, BME IE.412
- Book
  - P.Baldi: Bioinformatics
  - Antal et al.:Bioinformatics: http://www.tankonyvtar.hu/hu/tartalom/tamop412A/2011_0079_antal_bioinformatika/adatok.html
- Slides
  - At course site

Homework, midterm, …grading

- No midterm
- 1 homework
- Exam

Computational Biomedicine (ComBine) workgroup



Downloads

BayesCube for Windows 32-bit

BayesCube for Windows 64-bit

BayesCube for Linux 32-bit

BayesCube for Linux 64-bit

BayesCube for MacOSX 64-bit

Contact

E-mail

Péter Antal
antal@mit.bme.hu

Address

Budapest University of Technology and
Economics, Building "I"
1117 Budapest, Magyar tudósok körútja 2.
Room E423

Visual data analytics in pharmaceutical informatics

Date: 11/01/2017

In cooperation with CERN and MTA-Wigner we will investigate the use of large-scale, semantic visual data analytics in drug discovery.

Privacy preserving fusion in CELSA

Date: 10/01/2017

Our new project "HIDUCTION: Privacy preserving data sharing, analysis and decision support in personalized medicine" will start this year in cooperation with ESAT-STADIUS, K.U.Leuven (2017-2019).

Continued participation in the "UK Biobank"

Date: 09/13/2017

The "UK Biobank project No.1602" is extended till 2020. In cooperation with the University of Manchester and Semmelweis University, we investigate the interactions between diet, psychosocial and genetic factors for self-reported depression and related disorders

We joined the NVIDIA GPU GRANT program

Date: 09/06/2017

We joined the NVIDIA GPU GRANT program of Nvidia Corporation. We will explore bioinformatic and chemoinformatic applications of the donated GPUs.

Team
Bence Bolgár
Bence Bruncsics
András Gézsi
Gábor Hullám
András Millinghoffer
Péter Sárközy
Péter Antal

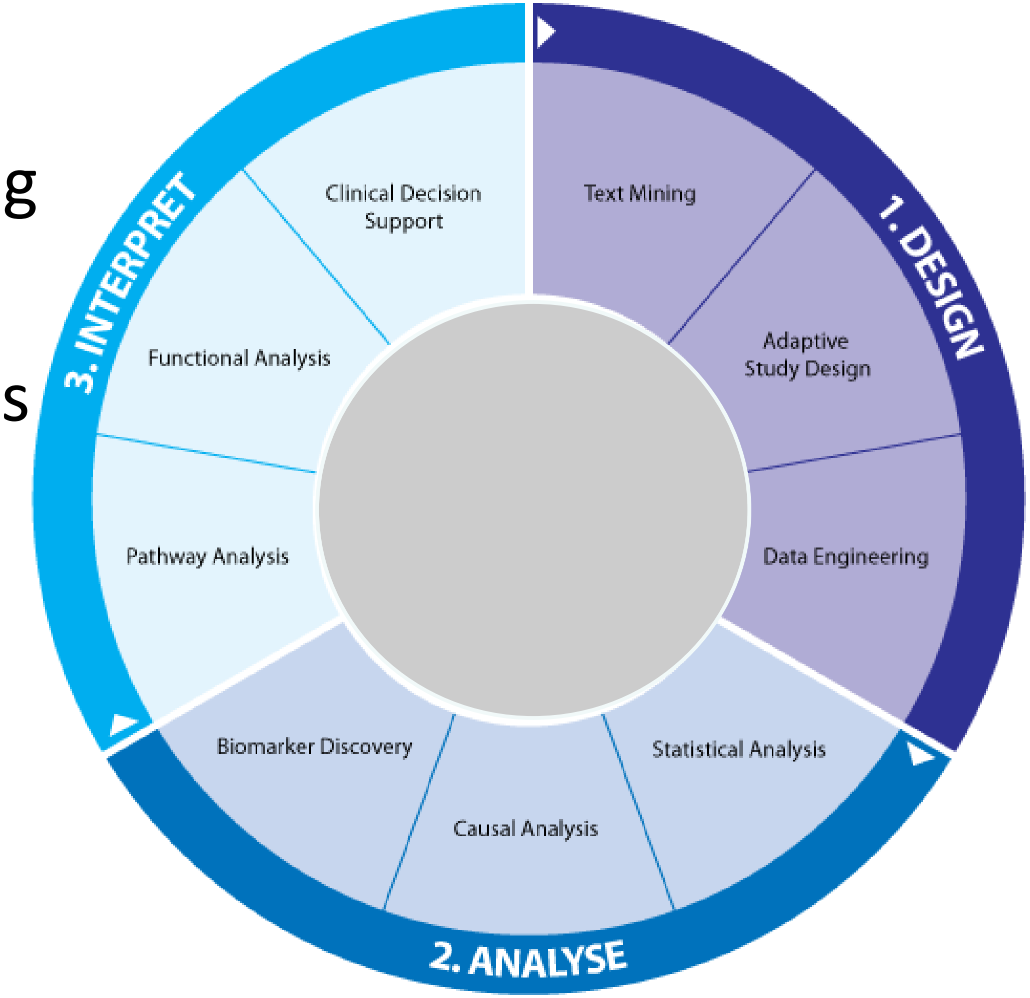
http://bioinfo.mit.bme.hu/

KOBAK: Bioinformatics

1. DNA genotyping and sequencing
2. Post-processing of genetic data
3. Protein modeling
4. Homology modeling
5. Functional effect of genetic variants
6. Gene regulatory networks
7. Genetic association studies
8. Gene expression studies
9. Biomarker analysis
10. Network science
11. Systems modeling
12. Causal inference in biomedicine
13. Text-mining in biomedicine
14. Study design
15. The biomedical big data
16. Data and knowledge fusion
17. Bayesian encyclopedia
18. Bioinformatic workflow systems
19. Drug discovery for personalized medicine
20. Metagenomics

ComBineLab.hu: Themes

- Knowledge engineering
- Study design
- Genetic measurements
- Data engineering
- Data analysis
- Interpretation
- Decision support



ComBineLab.hu: tools

- **BayesEye: Bayesian, systems-based data analysis**
  - Bayesian model averaging over Bayesian network structures.
- **BayesCube: Probabilistic decision support**
  - Semantically enriched Bayesian and decision network models.
- **BysCyc/QSF** (Bayesian Encyclopedia):
  - Large-scale probabilistic inference
- **QDF**: Kernel-based fusion methods for repositioning
  - Multi-aspect rankings and multi-aspect metrics in drug discovery
- **Variant Meta Caller**: precision NGS
  - Next-generation sequencing pipelines
- **VB-MK-LMF**: drug-target interaction prediction
  - Variational Bayesian Multiple Kernel Logistic Matrix Factorization

- ... see Tools @ http://bioinfo.mit.bme.hu/

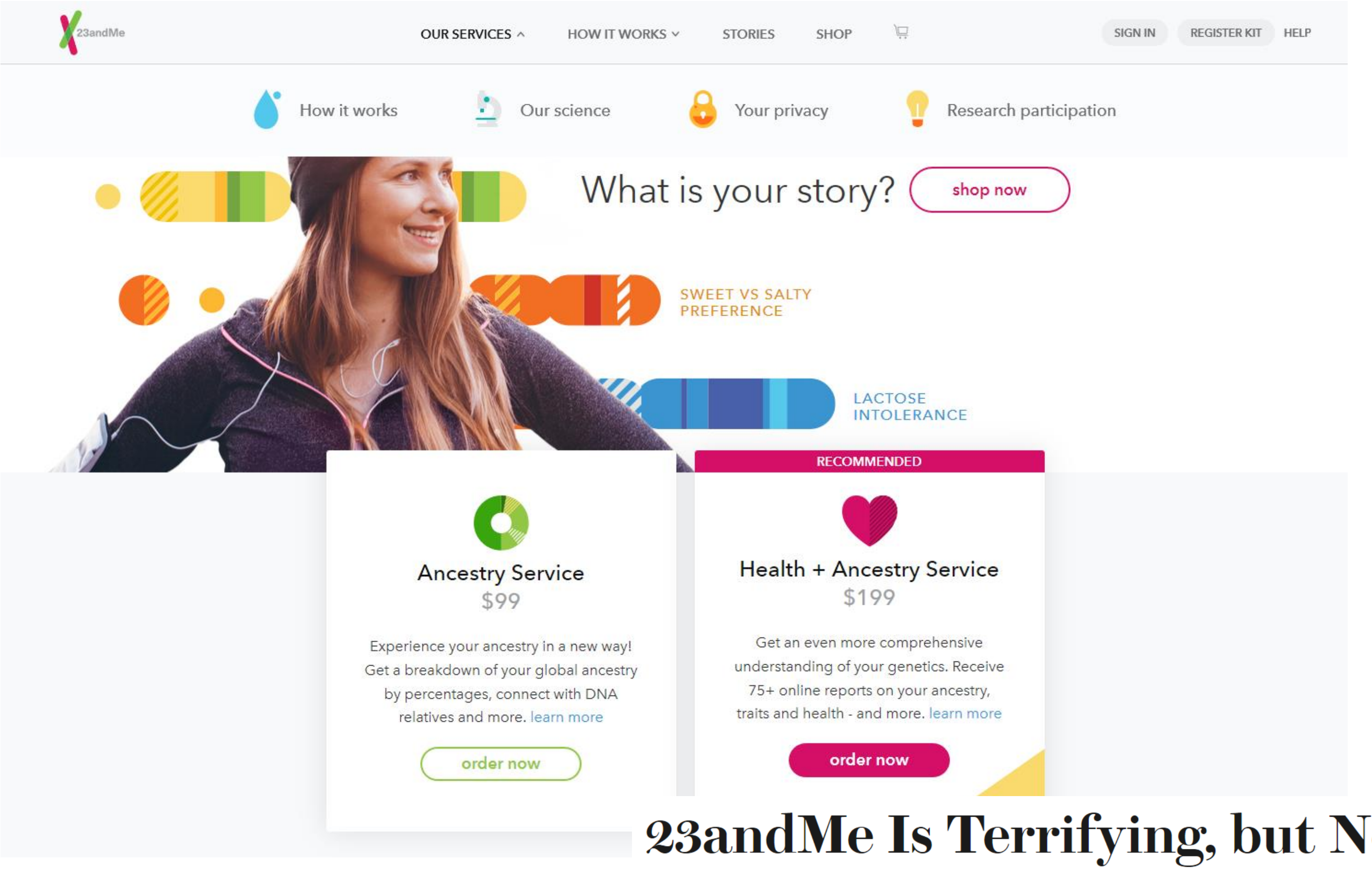
Course outline

- NGS data analysis: 5 weeks
- Semantic technologies: 1 week:
- Chemoinformatics, drug discovery: 1 week
- GWAS data analysis: 3 weeks
- Biomed decision support: 2 weeks
- Causal data analysis: 1 lecture
- Guest lectures: phylogeny,..
- Cases studies

Overview

- The „big data”/omic era of chemo- and bioinformatics
- Data and knowledge fusion in biomedicine
- The semantic unification of pharmacological spaces
- Research topics
  - Multi-aspect virtual screening
  - Drug repositioning
  - Aging research
  - Systems-based analysis of large health data sets
  - Medical decision support
  - Privacy preserving data analysis in chemo/bioinformatics

Direct2customer genetics



What is your story?

Ancestry Service
$99

Experience your ancestry in a new way! Get a breakdown of your global ancestry by percentages, connect with DNA relatives and more. learn more

Health + Ancestry Service
$199

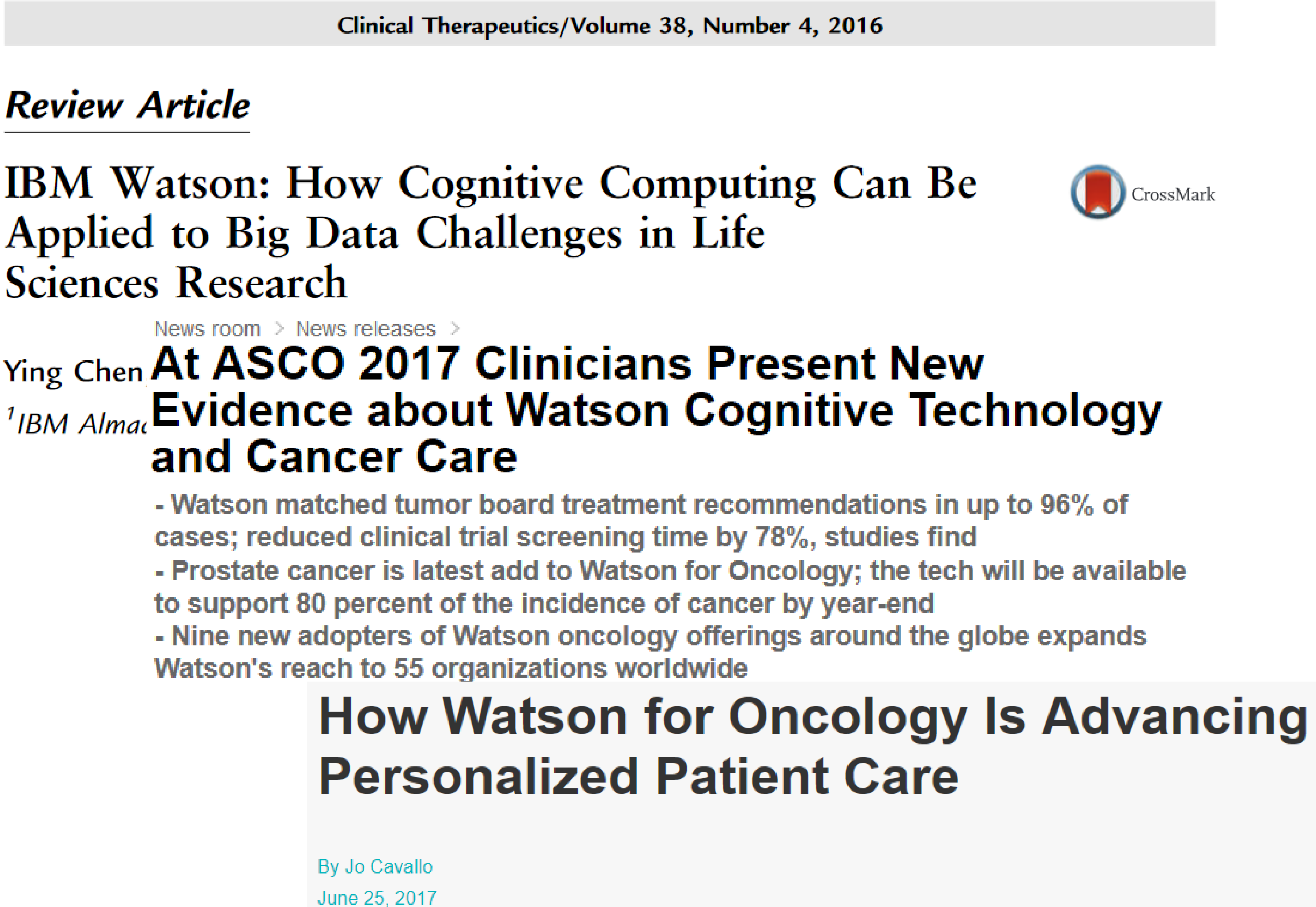
Get an even more comprehensive understanding of your genetics. Receive 75+ online reports on your ancestry, traits and health - and more. learn more

23andMe Is Terrifying, but Not for the Reasons the FDA Thinks

The genetic-testing company's real goal is to hoard your personal data

By Charles Seife on November 27, 2013

An era of a new *health care?*



Clinical Therapeutics/Volume 38, Number 4, 2016

Review Article

IBM Watson: How Cognitive Computing Can Be Applied to Big Data Challenges in Life Sciences Research

Ying Chen

[1]IBM Alma

At ASCO 2017 Clinicians Present New Evidence about Watson Cognitive Technology and Cancer Care

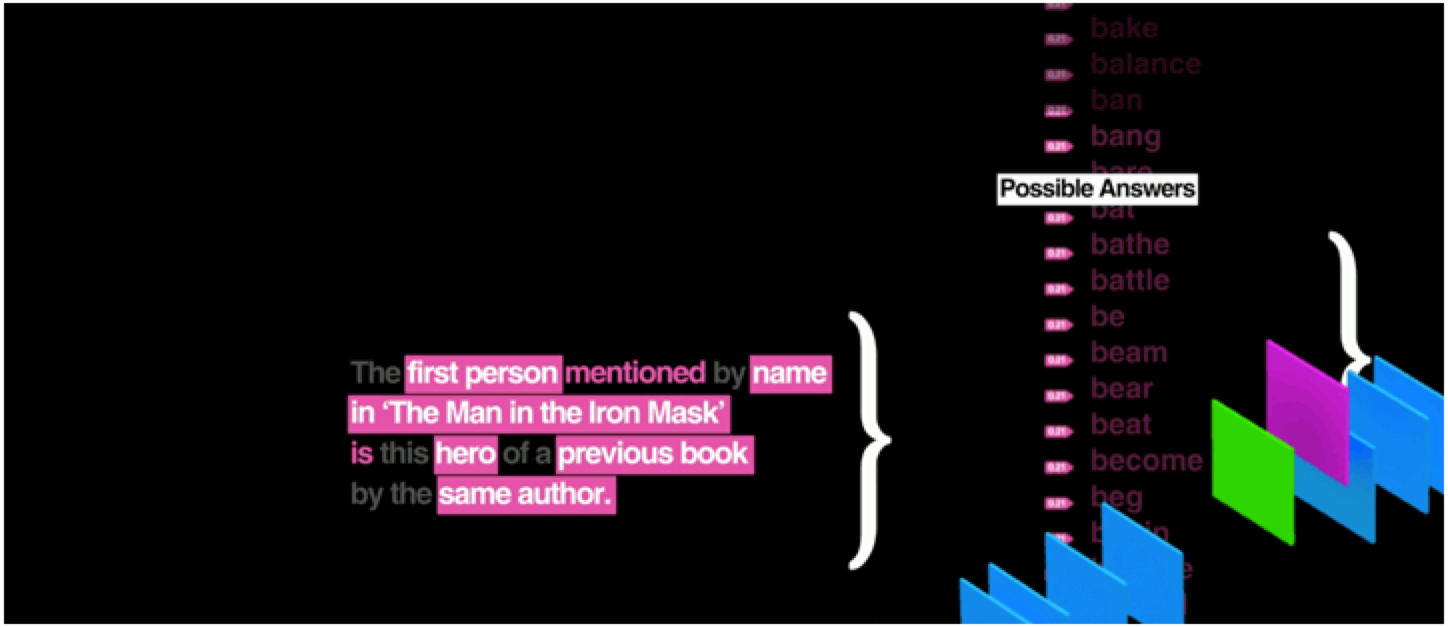
- Watson matched tumor board treatment recommendations in up to 96% of cases; reduced clinical trial screening time by 78%, studies find
- Prostate cancer is latest add to Watson for Oncology; the tech will be available to support 80 percent of the incidence of cancer by year-end
- Nine new adopters of Watson oncology offerings around the globe expands Watson's reach to 55 organizations worldwide

How Watson for Oncology Is Advancing Personalized Patient Care

By Jo Cavallo

June 25, 2017

Watson?
The Science Behind an Answer

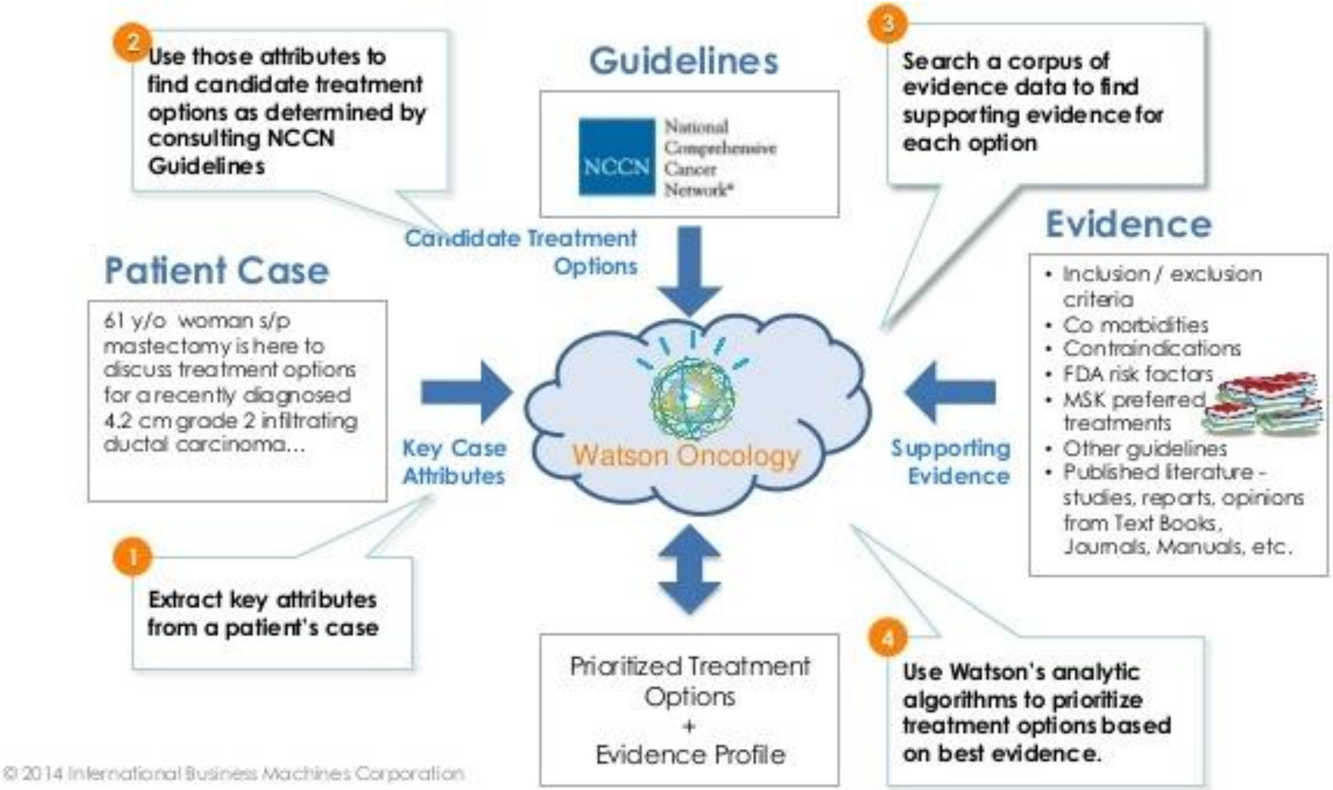


The first person mentioned by name in 'The Man in the Iron Mask' is this hero of a previous book by the same author.

- http://www-03.ibm.com/innovation/us/watson/what-is-watson/science-behind-an-answer.html

IBM WATSON

Biomedical decision support systems



2 Use those attributes to find candidate treatment options as determined by consulting NCCN Guidelines

Guidelines

NCCN National Comprehensive Cancer Network®

3 Search a corpus of evidence data to find supporting evidence for each option

Candidate Treatment Options

Evidence

Patient Case

61 y/o woman s/p mastectomy is here to discuss treatment options for a recently diagnosed 4.2 cm grade 2 infiltrating ductal carcinoma…

- Inclusion / exclusion criteria
- Co morbidities
- Contraindications
- FDA risk factors
- MSK preferred treatments
- Other guidelines
- Published literature - studies, reports, opinions from Text Books, Journals, Manuals, etc.

Key Case Attributes

Watson Oncology

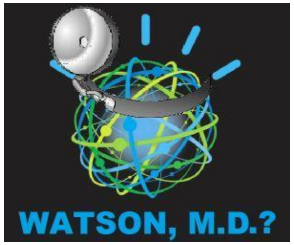
Supporting Evidence

1 Extract key attributes from a patient's case

Prioritized Treatment Options + Evidence Profile

4 Use Watson's analytic algorithms to prioritize treatment options based on best evidence.

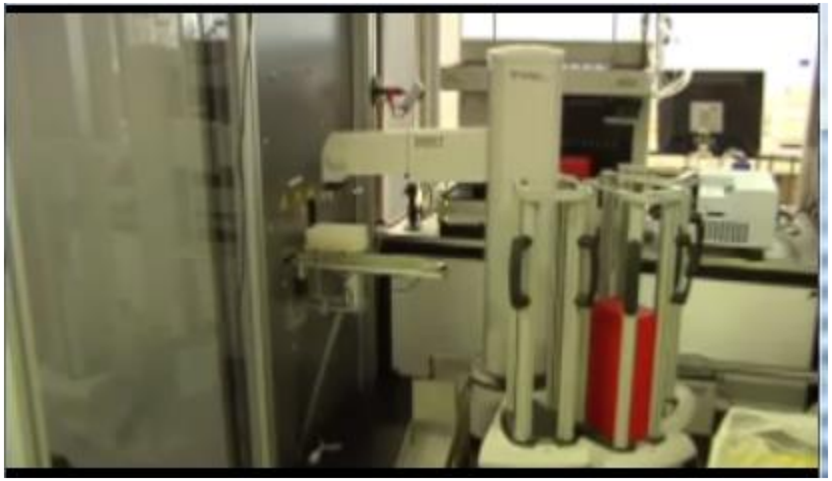
© 2014 International Business Machines Corporation



Watson for Oncology – assessment and advice cycle
www.avanteoconsulting.com/machine-learning-accelerates-cancer-research-discovery-innovation/

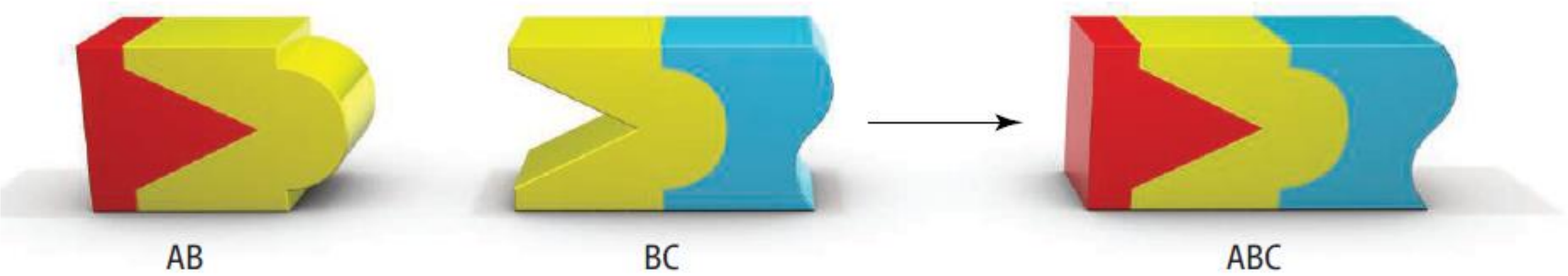
Automated discovery systems

- Langley, P. (**1978**). Bacon: A general discovery system. Proceedings of the Second Biennial Conference of the Canadian Society for Computational Studies of Intelligence (pp. 173-180). Toronto, Ontario.
- …
- Chrisman, L., Langley, P., & Bay, S. (**2003**). Incorporating biological knowledge into evaluation of causal regulatory hypotheses. Proceedings of the Pacific Symposium on Biocomputing (pp. 128-139). Lihue, Hawaii.
- (Gene prioritization…)
- R.D.King et al.: The Automation of Science, Science, **2009**



„Machine science”

- Swanson, Don R. "Fish oil, Raynaud's syndrome, and undiscovered public knowledge." *Perspectives in biology and medicine* 30.1 (1986): 7-18.
- Smalheiser, Neil R., and Don R. Swanson. "Using **ARROWSMITH**: a computer-assisted approach to formulating and assessing scientific hypotheses." *Computer methods and programs in biomedicine* 57.3 (1998): 149-153.
- D. R. Swanson et al.: **An interactive system for finding complementary literatures: a stimulus to scientific discovery**, Artificial Intelligence, 1997



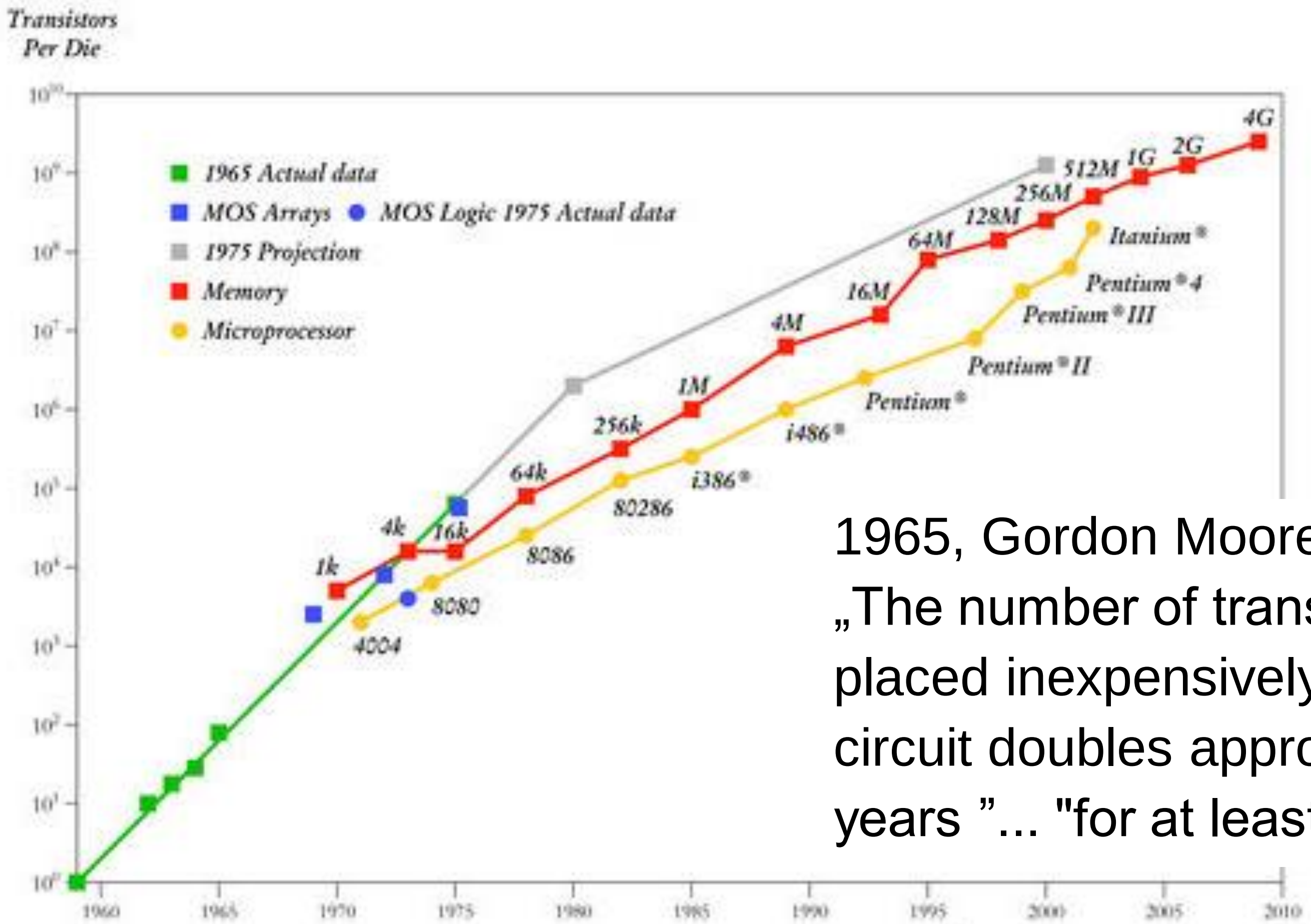
- James Evans and Andrey Rzhetsky: **Machine science**, Science, 2013

„Soon, computers could generate many useful hypotheses with little help from humans.”

Factors behind the new health care

- New measurements?
  - Ultimate diagnostics?
- New theories?
  - Unified theory of medicine?
  - Unified theory of artificial intelligence?
- New therapies?
  - Preventions?
  - Drugs?
  - Anti-aging, rejuvenation?

Computing power: Moore’s Law



Integration and parallelization wont bring us further. End of Moore’s law?

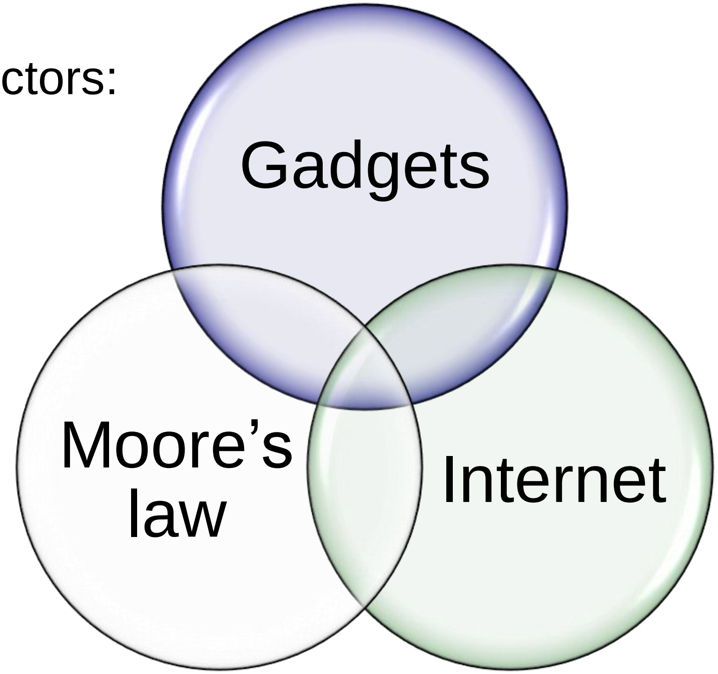
1965, Gordon Moore, founder of Intel: „The number of transistors that can be placed inexpensively on an integrated circuit doubles approximately every two years ”... "for at least ten years"

The „big” data

- Financial transaction data, mobile phone data, user (click) data, e-mail data, internet search data, social network data, sensor networks, ambient assisted living, intelligent home, wearable electronics,...

“The line between the virtual world of computing and our physical, organic world is blurring.” E.Dumbill: Making sense of big data, Big Data, vol.1, no.1, 2013

Factors:



Definitions of „big data”

M. Cox and D. Ellsworth, “Managing **Big Data** for Scientific Visualization,” Proc. ACM Siggraph, ACM, 1997

The 3xV: **volume, variety, and velocity** (2001).

The 8xV: Vast, Volumes of Vigorously, Verified, Vexingly Variable Verbose yet Valuable Visualized high Velocity Data (2013)

Not „conventional” data: „Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn’t fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it (E.Dumbill: Making sense of big data, Big Data, vol.1, no.1, 2013)

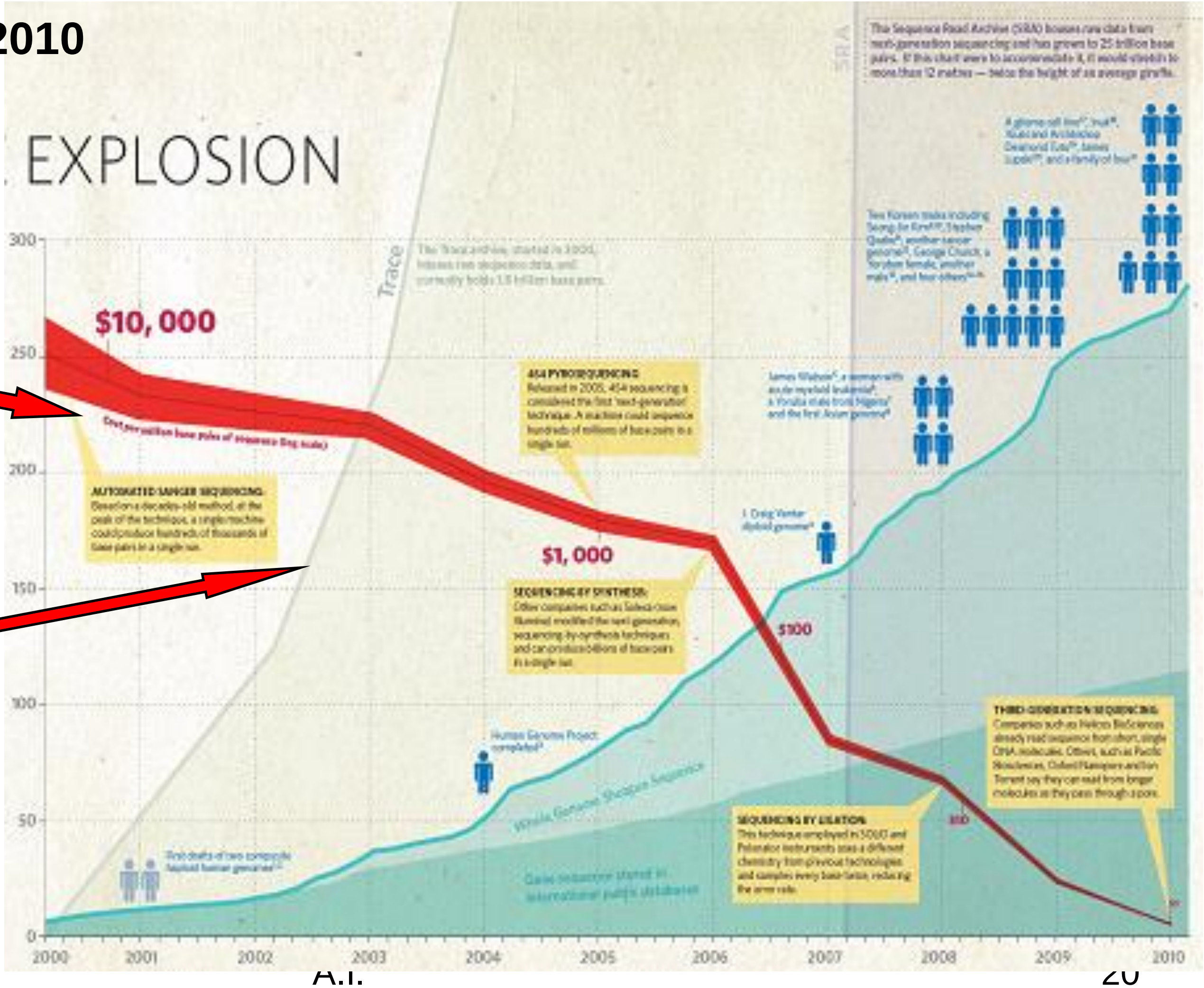
Carlson’s Law for Biological Data

NATURE, Vol 464, April 2010

Sequencing costs per mill. base

Publicly available genetic data

- x10 every 2-3 years
- Data volumes and complexity that IT has never faced before…



The Trace archive, started in 2000, houses raw sequence data, and currently holds 1.5 trillion base pairs.

The Sequence Read Archive (SRA) houses raw data from next-generation sequencing and has grown to 25 trillion base pairs. If this chart were to accommodate it, it would stretch to more than 12 metres — twice the height of an average giraffe.

AUTOMATED SANGER SEQUENCING: Based on a decades-old method, at the peak of the technique, a single machine could produce hundreds of thousands of base pairs in a single run.

454 PYROSEQUENCING: Released in 2005, 454 sequencing is considered the first 'next-generation' technique. A machine could sequence hundreds of millions of base pairs in a single run.

SEQUENCING BY SYNTHESIS: Other companies such as Solexa (now Illumina) modified the next generation, sequencing-by-synthesis techniques and can produce billions of base pairs in a single run.

SEQUENCING BY LIGATION: This technique employed in SOLiD and Polonator instruments uses a different chemistry from previous technologies and samples every base twice, reducing the error rate.

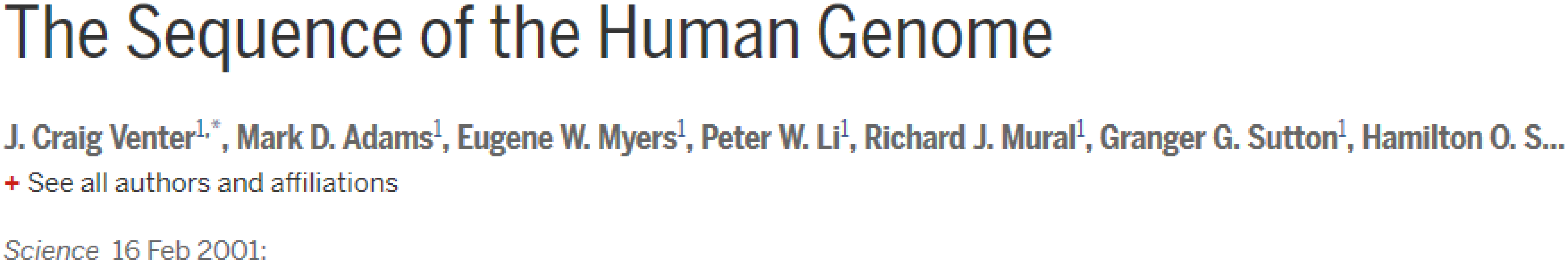
THIRD-GENERATION SEQUENCING: Companies such as Helicos BioSciences already read sequence from short, single DNA molecules. Others, such as Pacific Biosciences, Oxford Nanopore and Ion Torrent say they can read from longer molecules as they pass through a pore.

The new health care: personalized medicine

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W. and Funke, R., 2001. Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), pp.860-921.



The Sequence of the Human Genome

J. Craig Venter[1,*], Mark D. Adams[1], Eugene W. Myers[1], Peter W. Li[1], Richard J. Mural[1], Granger G. Sutton[1], Hamilton O. S...

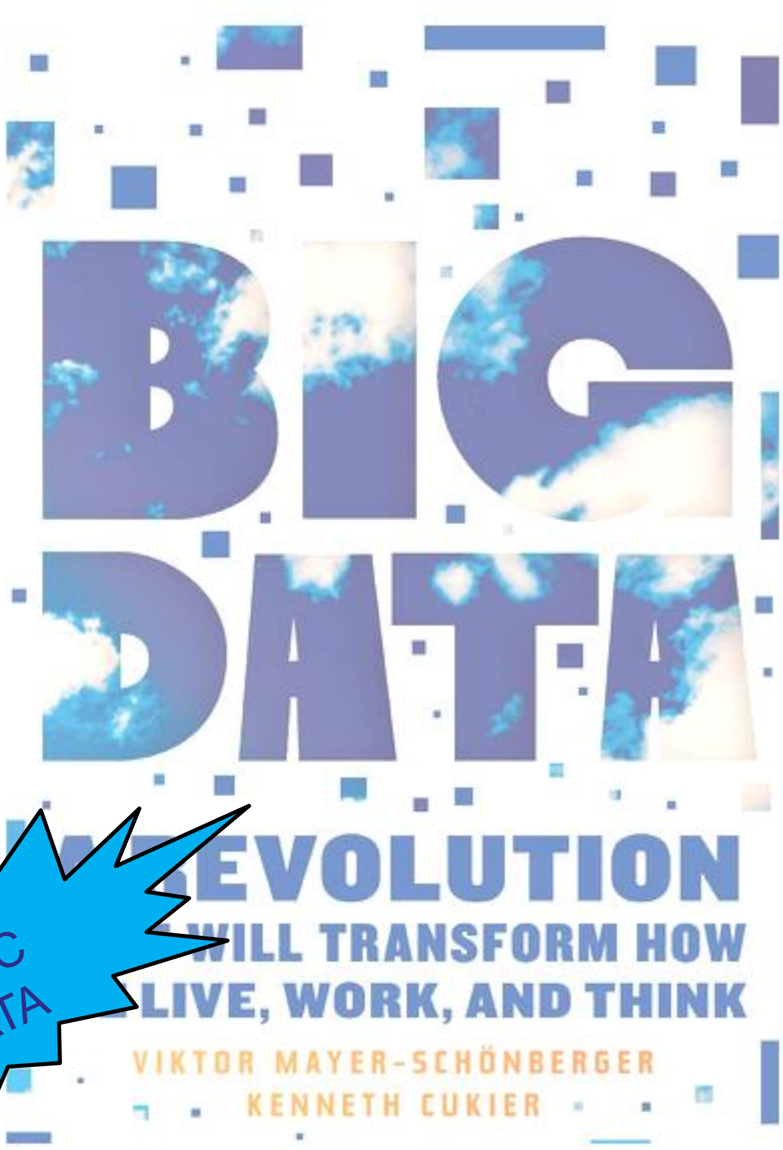
+ See all authors and affiliations

Science 16 Feb 2001:

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. and Gocayne, J.D., 2001. The sequence of the human genome. *Science*, *291*(5507), pp.1304-1351.

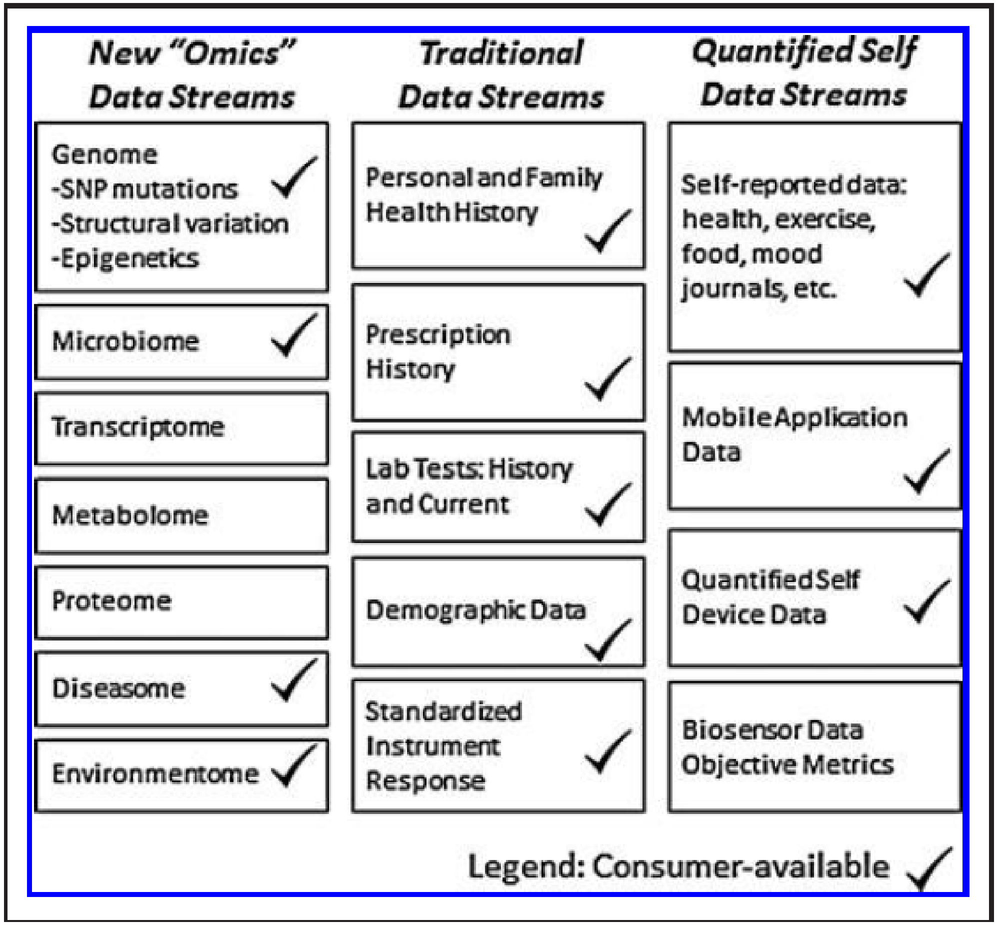
The „omic” definition of „big data”

.. [data] is often big in relation to the phenomenon that we are trying to record and understand. So, if we are only looking at 64,000 data points, but **that represents the totality or the universe of observations. That is what qualifies as big data. You do not have to have a hypothesis in advance before you collect your data**. You have collected all there is—**all the data there is about a phenomenon.**



OMIC DATA

Big health data streams



New "Omics" Data Streams

- Genome -SNP mutations -Structural variation -Epigenetics ✓
- Microbiome ✓
- Transcriptome
- Metabolome
- Proteome
- Diseasome ✓
- Environmentome ✓

Traditional Data Streams

- Personal and Family Health History ✓
- Prescription History ✓
- Lab Tests: History and Current ✓
- Demographic Data ✓
- Standardized Instrument Response ✓

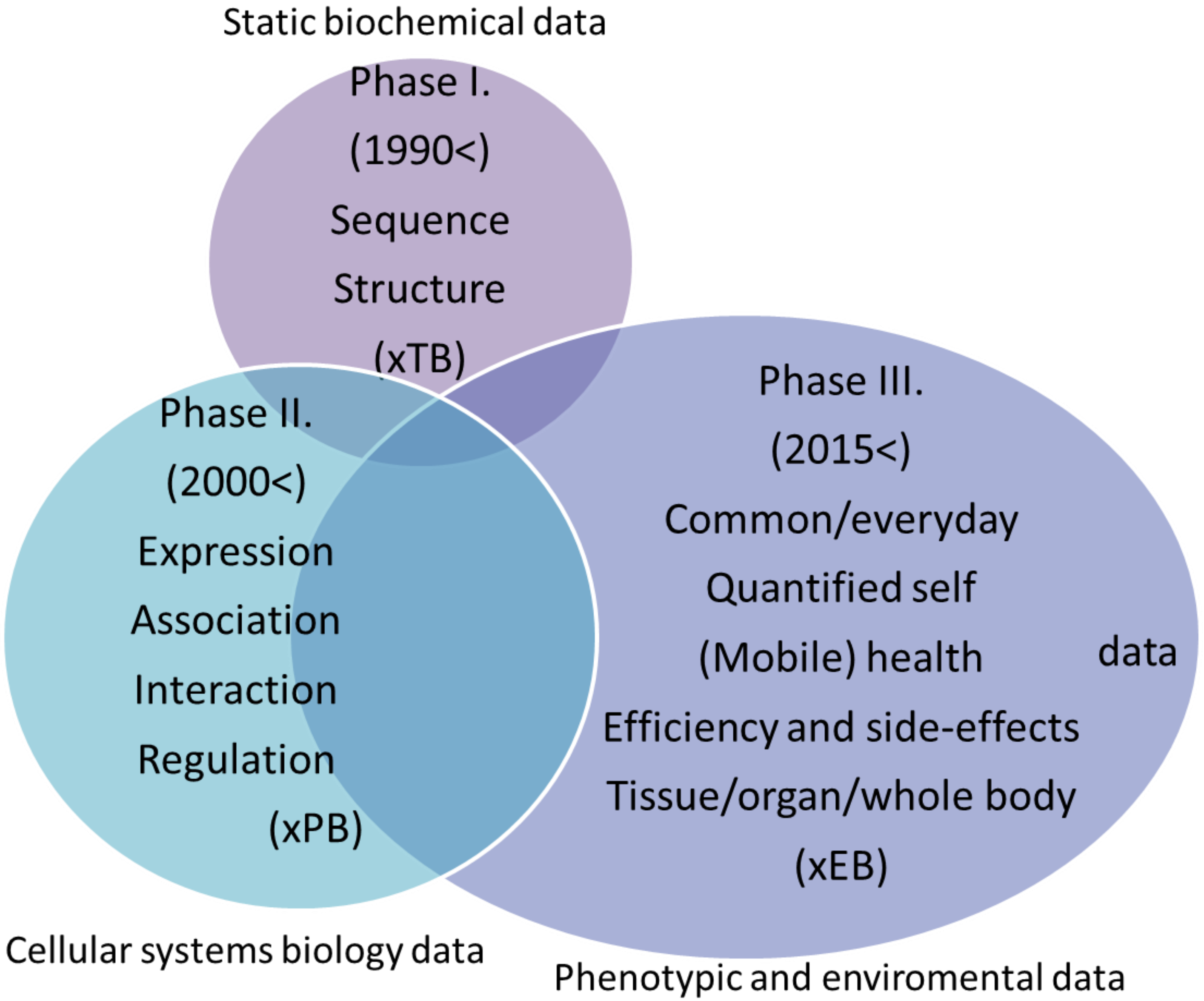
Quantified Self Data Streams

- Self-reported data: health, exercise, food, mood journals, etc. ✓
- Mobile Application Data ✓
- Quantified Self Device Data ✓
- Biosensor Data Objective Metrics

Legend: Consumer-available ✓

M.Swan: THE QUANTIFIED SELF: Fundamental Disruption in Big Data Science and Biological Discovery, Big data, Vol 1., No. 2., 2013

Big „omic” data sets in biomed.



Static biochemical data

Phase I.
(1990<)
Sequence
Structure
(xTB)

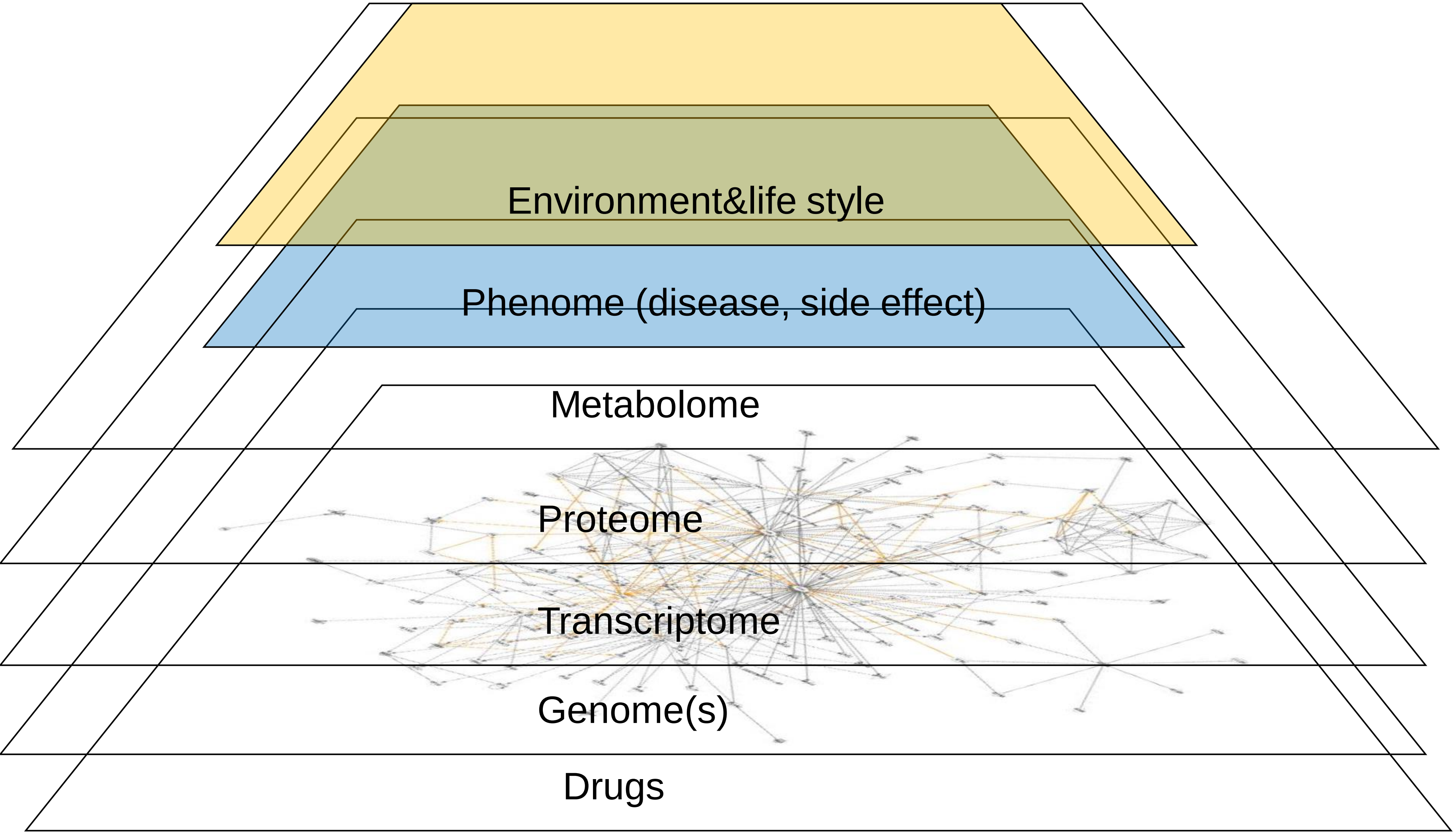
Phase II.
(2000<)
Expression
Association
Interaction
Regulation
(xPB)

Phase III.
(2015<)
Common/everyday
Quantified self
(Mobile) health data
Efficiency and side-effects
Tissue/organ/whole body
(xEB)

Cellular systems biology data

Phenotypic and enviromental data

Multiple levels in biomedicine



Environment&life style

Phenome (disease, side effect)

Metabolome

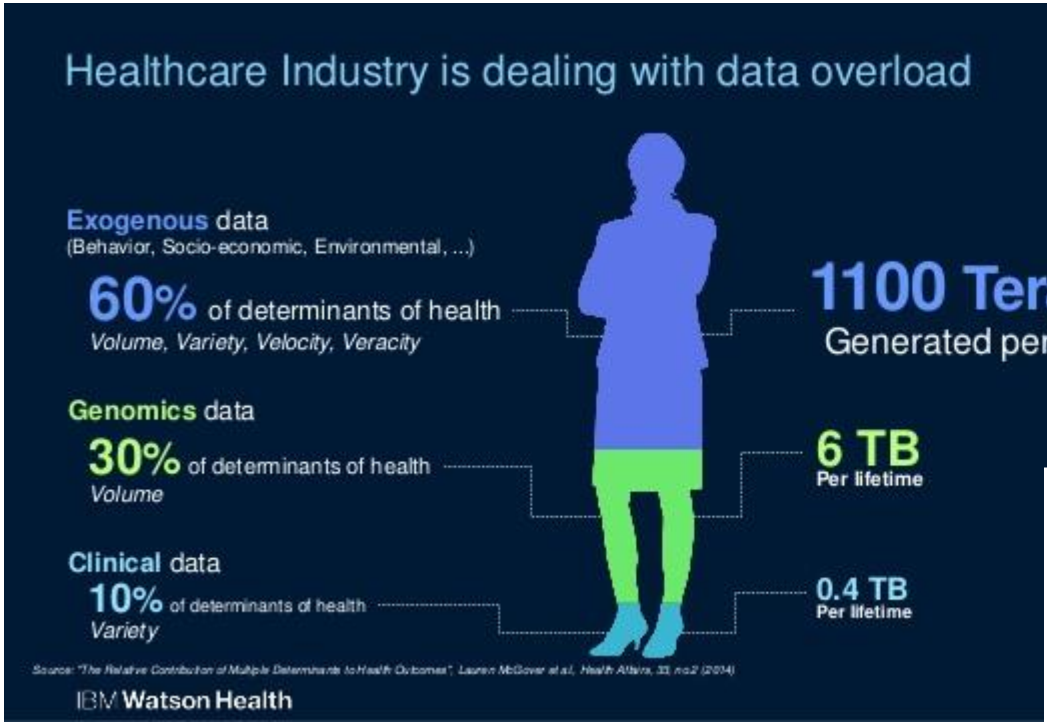
Proteome

Transcriptome

Genome(s)

Drugs

Data: „Big” data in life sciences



Healthcare Industry is dealing with data overload

Exogenous data
(Behavior, Socio-economic, Environmental, ...)
60% of determinants of health
Volume, Variety, Velocity, Veracity

Genomics data
30% of determinants of health
Volume

Clinical data
10% of determinants of health
Variety

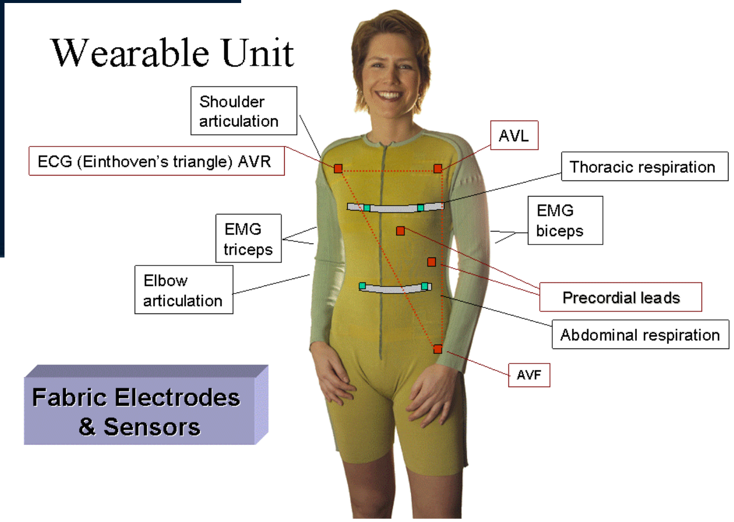
1100 Terabytes
Generated per lifetime

6 TB
Per lifetime

0.4 TB
Per lifetime

Source: "The Relative Contribution of Multiple Determinants to Health Outcomes", Lauren McGover et al., Health Affairs, 33, no.2 (2014)

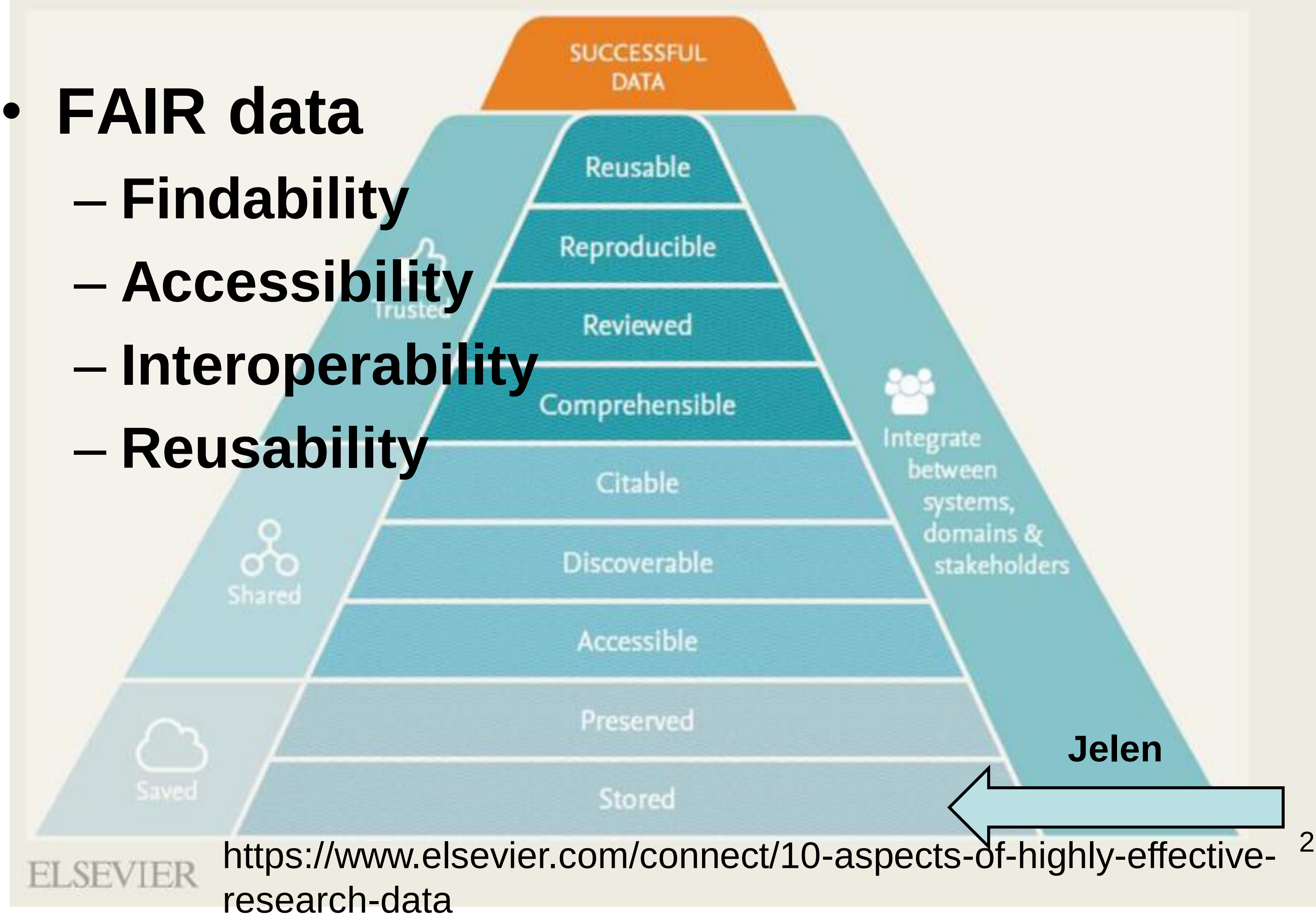
IBM Watson Health



Wearable Unit

Open data

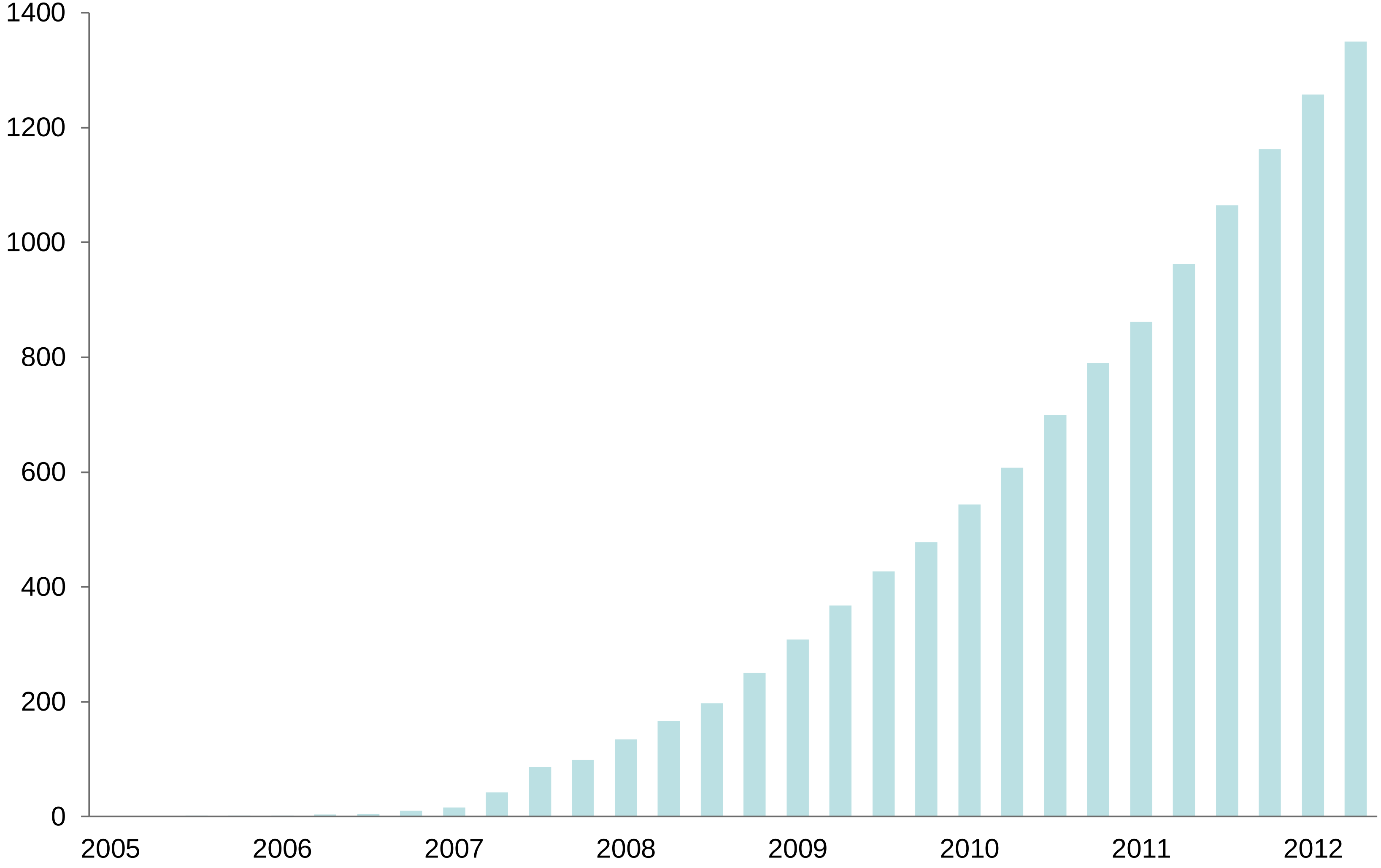
- **FAIR data**
  - **Findability**
  - **Accessibility**
  - **Interoperability**
  - **Reusability**



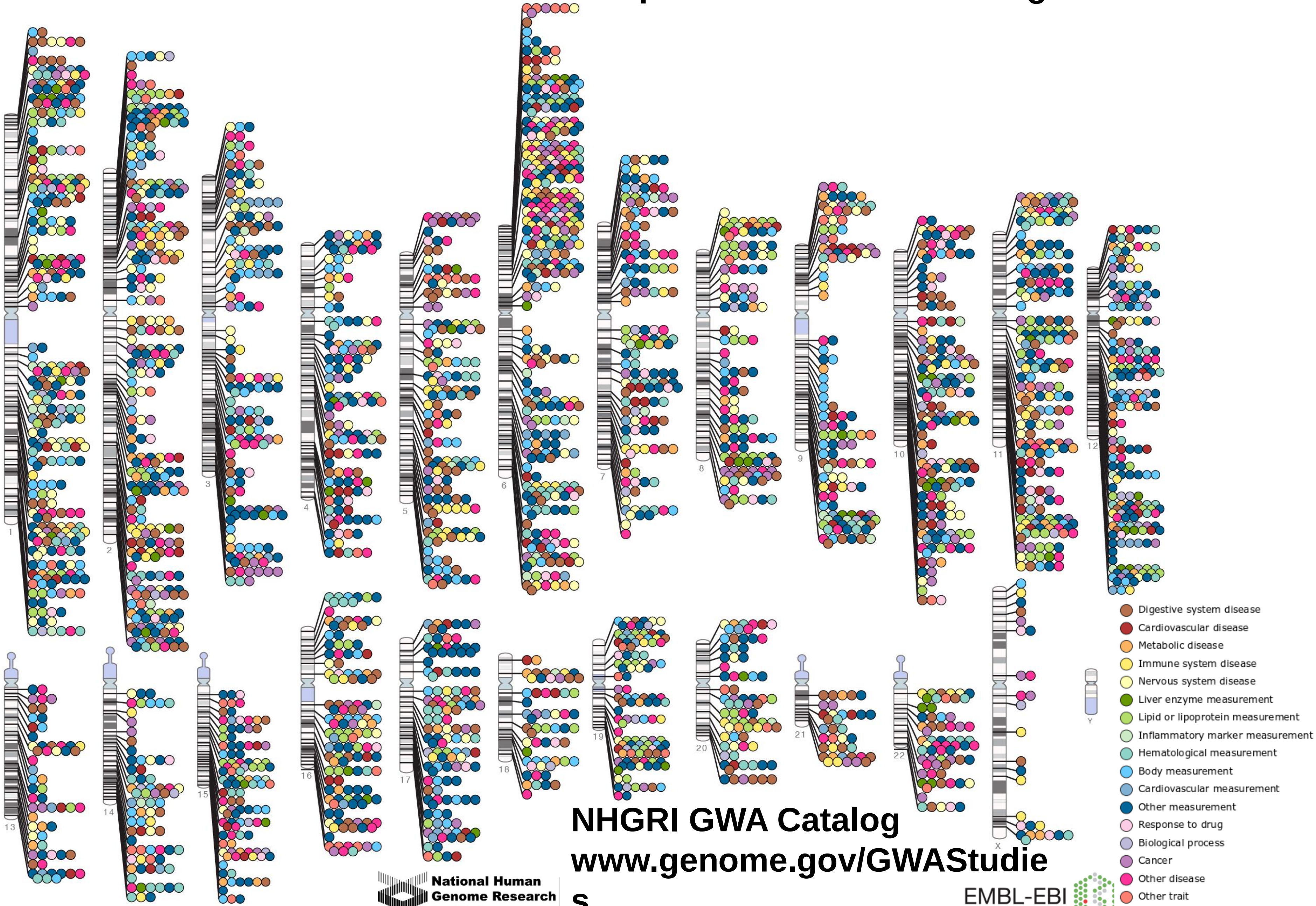
Jelen

https://www.elsevier.com/connect/10-aspects-of-highly-effective-research-data

Number of genome-wide association studies



Published Genome-Wide Associations through 12/2012
Published GWA at $p \leq 5X10^{-8}$ for 17 trait categories

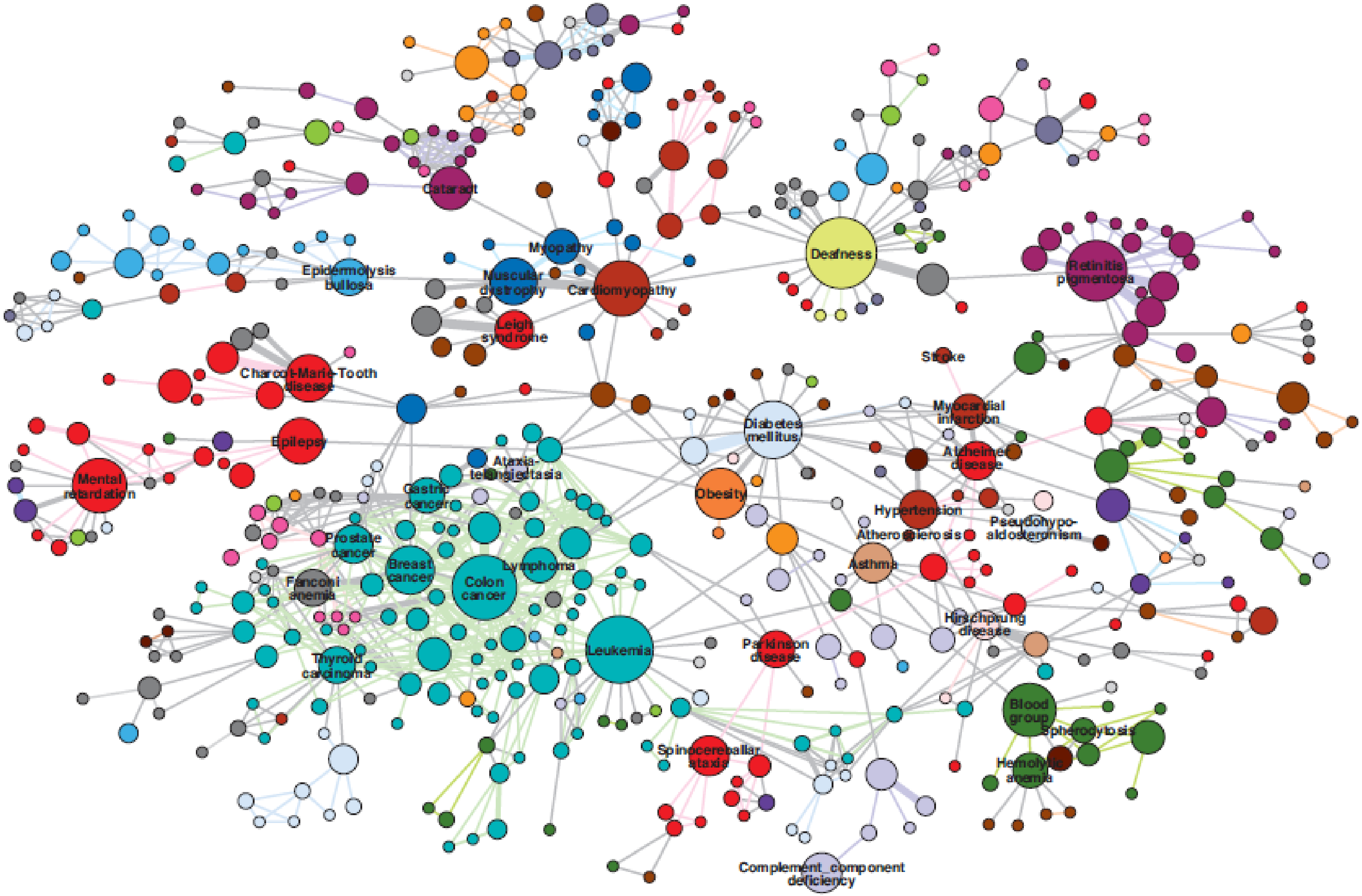


NHGRI GWA Catalog
www.genome.gov/GWAStudies

National Human Genome Research Institute

EMBL-EBI

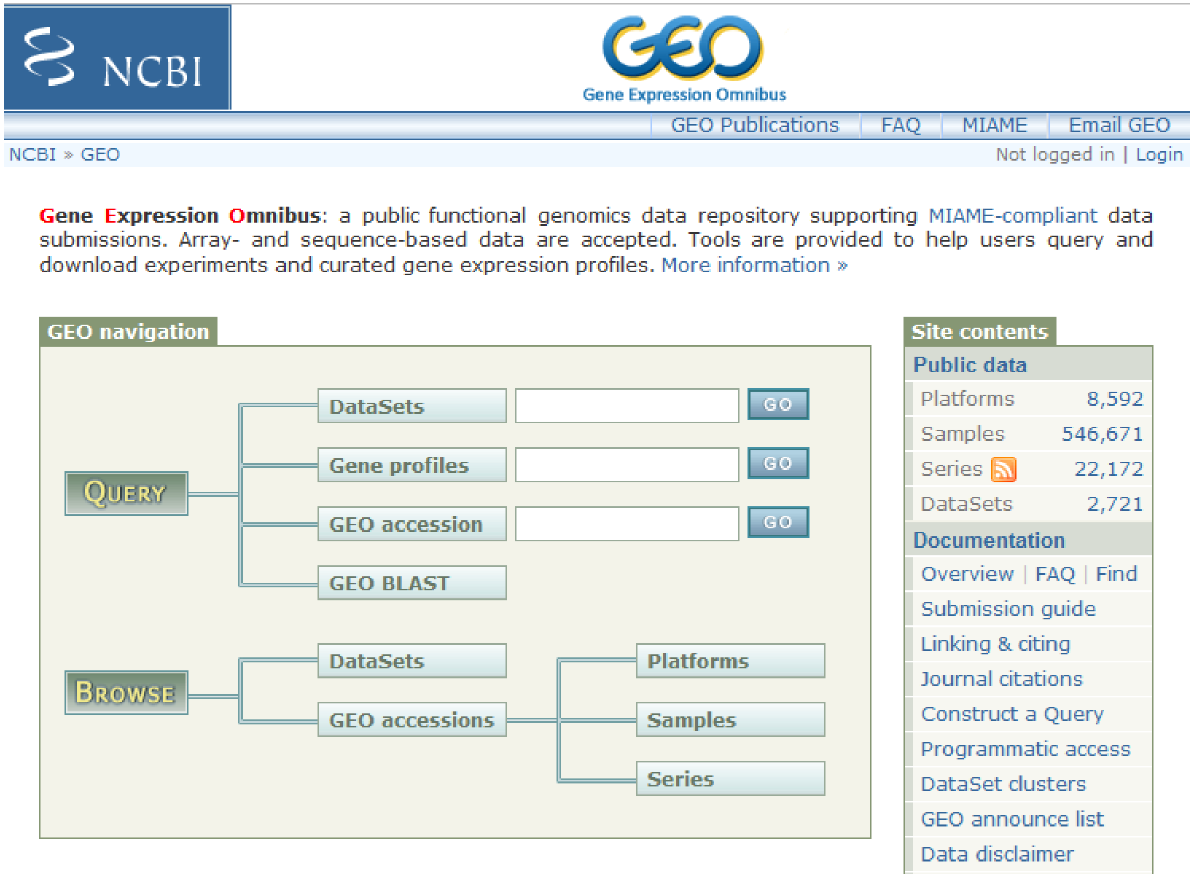
Genetic overlap based disease maps



L.A.Barabási:PNAS, 2007, The human disease network

Repositories for gene expression

- Gene Expression Omnibus (NCBI)
- http://www.ncbi.nlm.nih.gov/geo/



Gene Expression Omnibus: a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles. More information »

Public data	
Platforms	8,592
Samples	546,671
Series	22,172
DataSets	2,721

Documentation

Overview | FAQ | Find
Submission guide
Linking & citing
Journal citations
Construct a Query
Programmatic access
DataSet clusters
GEO announce list
Data disclaimer

Data: chemogenomics screening

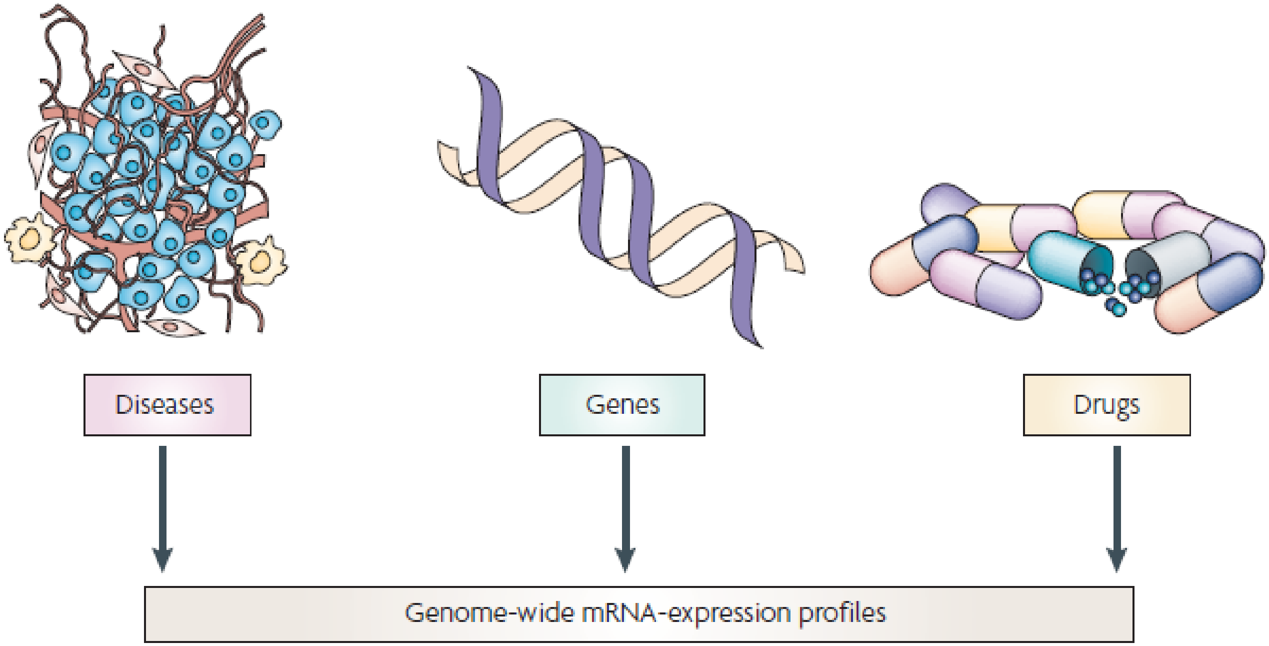
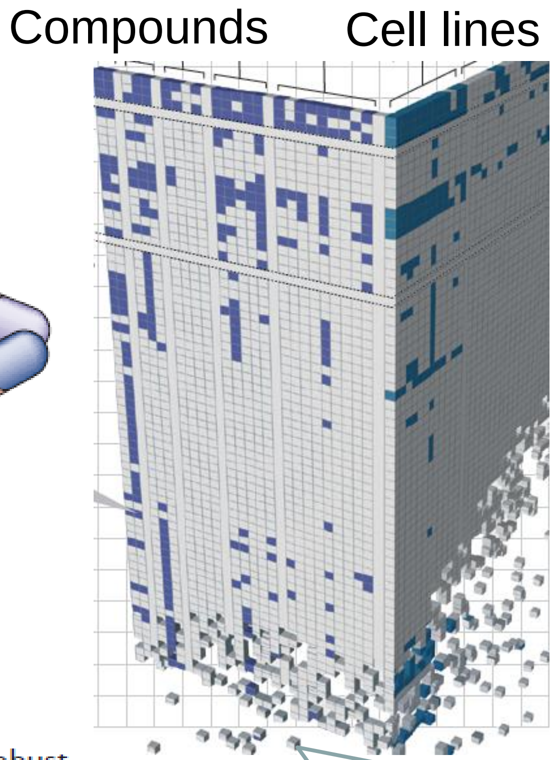


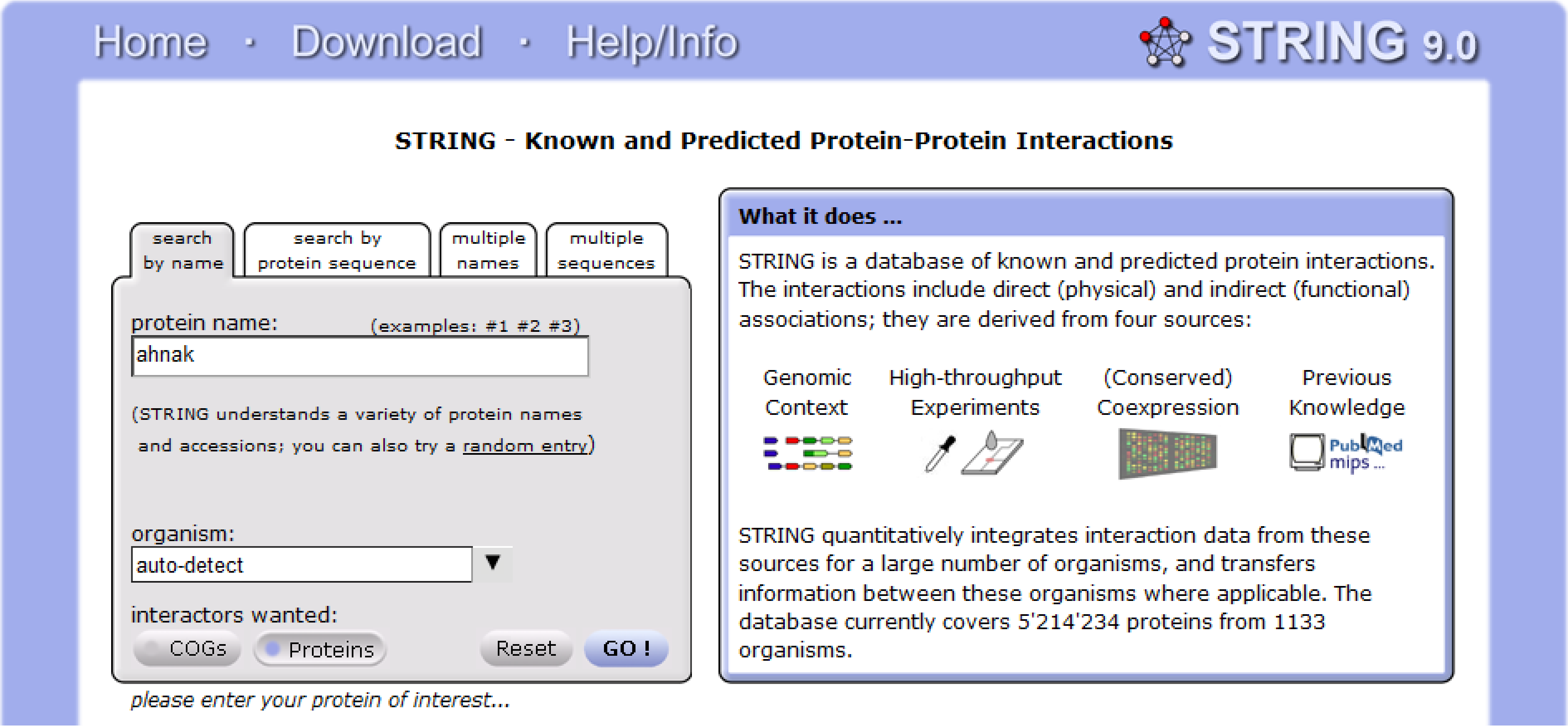
Figure 2 | **A universal functional bioassay.** As all of the transcripts are now known, and robust technologies for their simultaneous measurement are available, it is possible to capture objective high-dimensional depictions of all induced or organic biological conditions in a common global analytical space, and thereby readily appreciate similarities between them.



Each cell is transcriptional proifle

- Justin Lamb: The Connectivity Map: a new tool for biomedical research, Nature, 7,pp 54-60, 2007

STRING - Protein-Protein Interactions



STRING - Known and Predicted Protein-Protein Interactions

protein name: (examples: #1 #2 #3)

ahnak

(STRING understands a variety of protein names and accessions; you can also try a random entry)

organism:

auto-detect ▼

interactors wanted:

○ COGs ● Proteins

please enter your protein of interest...

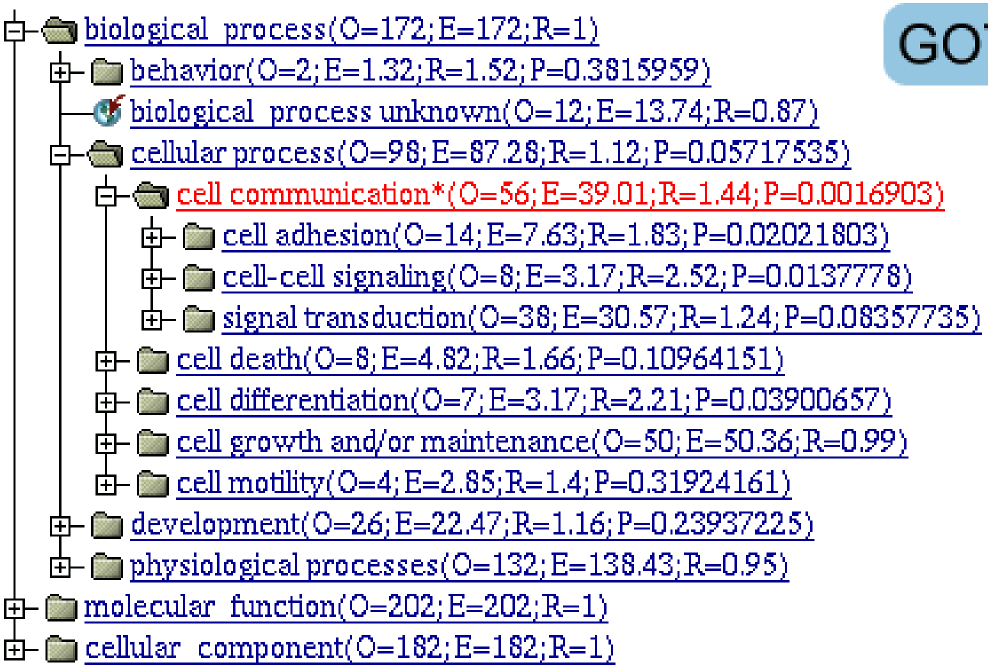
What it does ...

STRING is a database of known and predicted protein interactions. The interactions include direct (physical) and indirect (functional) associations; they are derived from four sources:

STRING quantitatively integrates interaction data from these sources for a large number of organisms, and transfers information between these organisms where applicable. The database currently covers 5'214'234 proteins from 1133 organisms.

- http://string-db.org/

Unification of biology: Gene Ontology

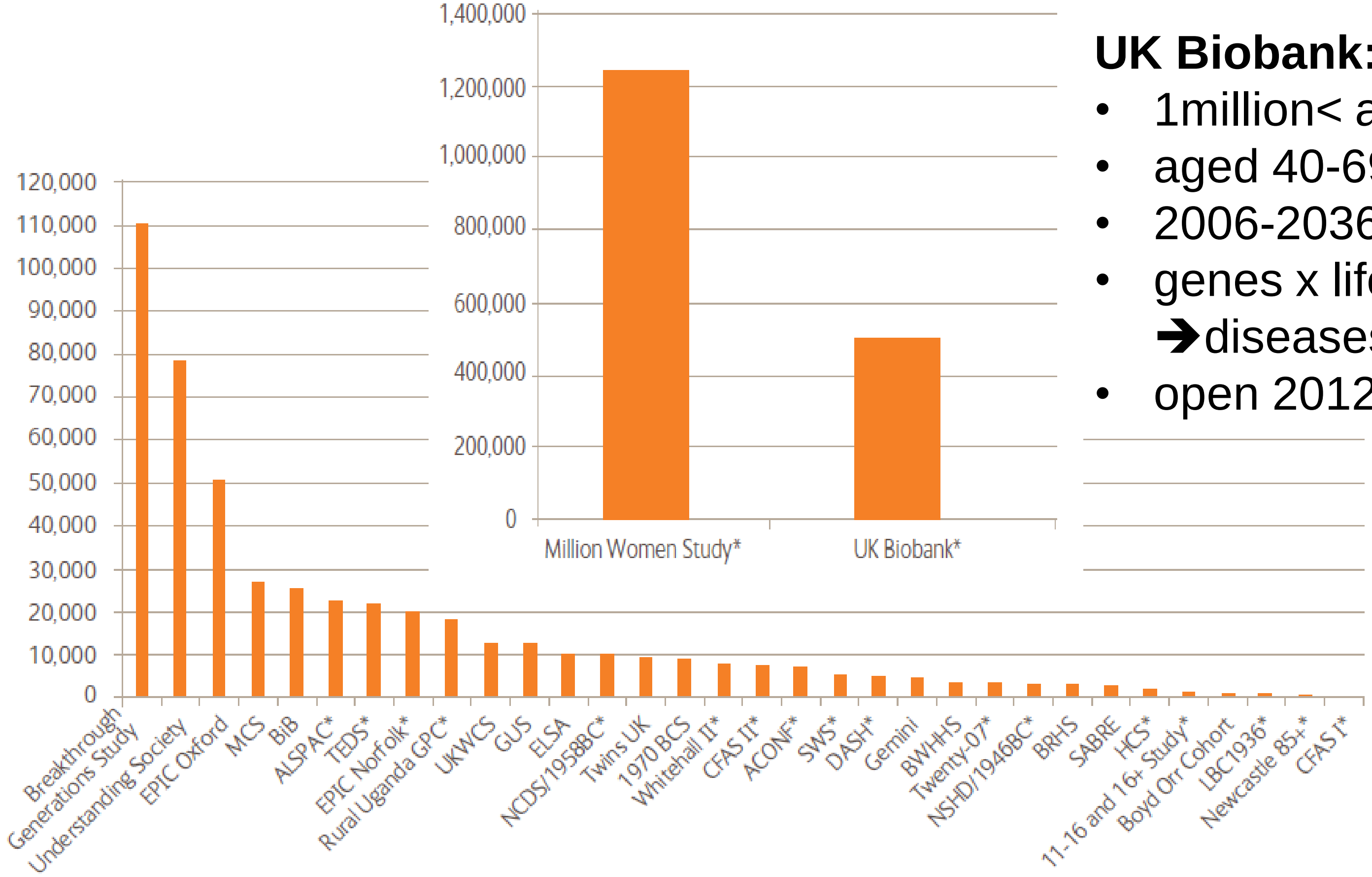


```
biological_process(O=172;E=172;R=1)
    behavior(O=2;E=1.32;R=1.52;P=0.3815959)
    biological_process unknown(O=12;E=13.74;R=0.87)
    cellular process(O=98;E=87.28;R=1.12;P=0.05717535)
        cell communication*(O=56;E=39.01;R=1.44;P=0.0016903)
            cell adhesion(O=14;E=7.63;R=1.83;P=0.02021803)
            cell-cell signaling(O=8;E=3.17;R=2.52;P=0.0137778)
            signal transduction(O=38;E=30.57;R=1.24;P=0.08357735)
        cell death(O=8;E=4.82;R=1.66;P=0.10964151)
        cell differentiation(O=7;E=3.17;R=2.21;P=0.03900657)
        cell growth and/or maintenance(O=50;E=50.36;R=0.99)
        cell motility(O=4;E=2.85;R=1.4;P=0.31924161)
    development(O=26;E=22.47;R=1.16;P=0.23937225)
    physiological processes(O=132;E=138.43;R=0.95)
molecular_function(O=202;E=202;R=1)
cellular_component(O=182;E=182;R=1)
```

GOTree Window

- Ontologies:
  - Gene Ontology (GO): http://www.geneontology.org/
  - Enzyme Classification (EC)
  - Unified Medical Language Systems (UMLS)
  - OBO

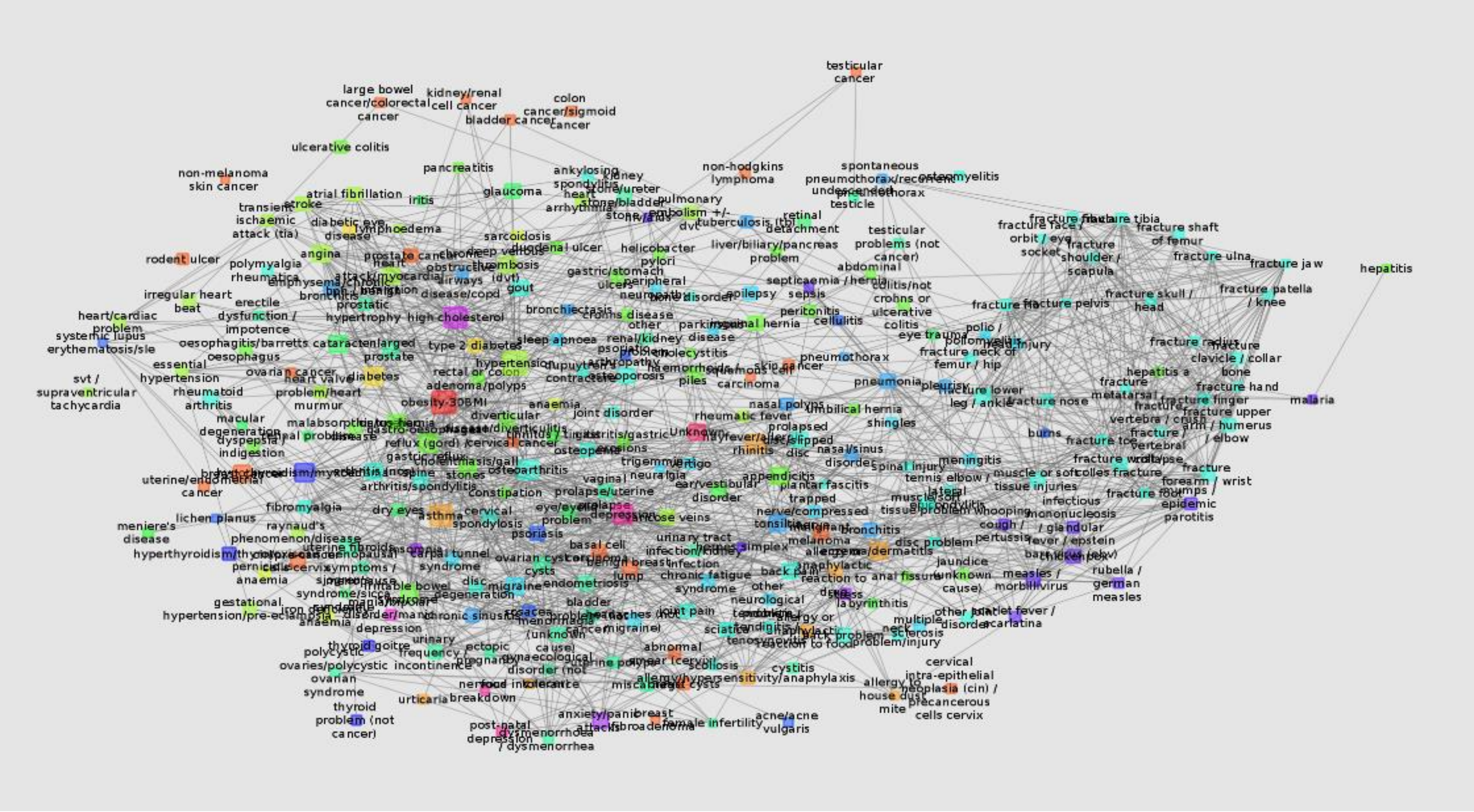
Large-scale cohorts in UK



UK Biobank:

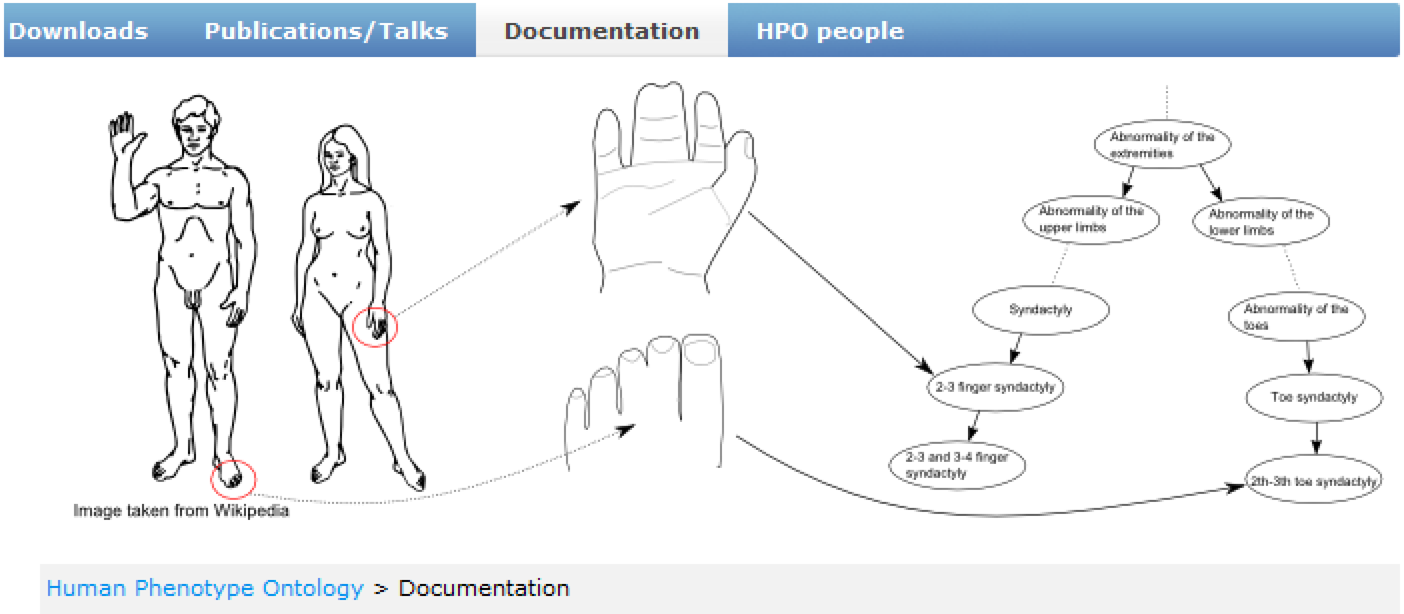
- 1million< adults
- aged 40-69,
- 2006-2036<
- genes x lifestyle x environment ➔diseases
- open 2012-

Epidemiologocal disease maps



Marx, P., Antal, P., Bolgar, B., Bagdy, G., Deakin, B. and Juhasz, G., 2017. Comorbidities in the diseasome are more apparent than real: What Bayesian filtering reveals about the comorbidities of depression. *PLoS computational biology*, *13*(6), p.e1005487.

The Human Phenotype Ontology


Downloads
Publications/Talks
Documentation
HPO people
Image taken from Wikipedia
Abnormality of the extremities
Abnormality of the upper limbs
Abnormality of the lower limbs
Syndactyly
Abnormality of the toes
2-3 finger syndactyly
Toe syndactyly
2-3 and 3-4 finger syndactyly
2th-3th toe syndactyly
Human Phenotype Ontology > Documentation

http://human-phenotype-ontology.github.io/

Number of biomedical publications

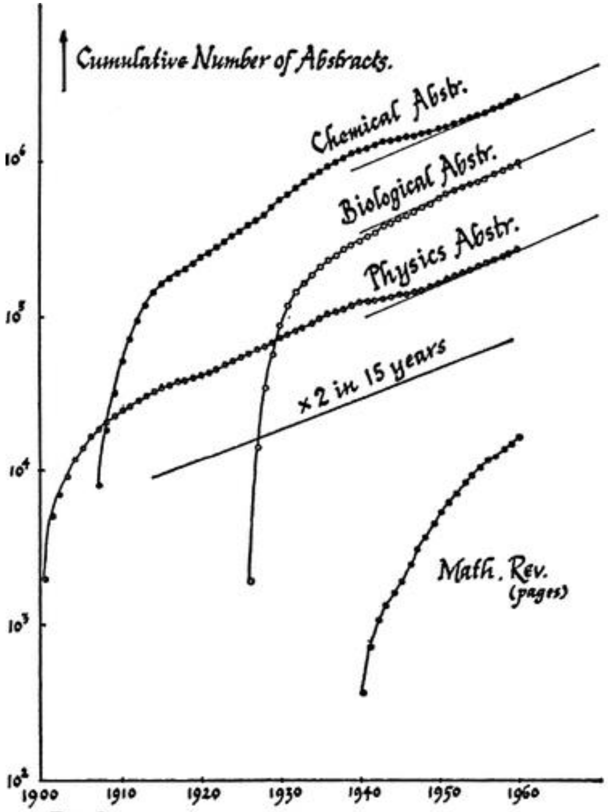
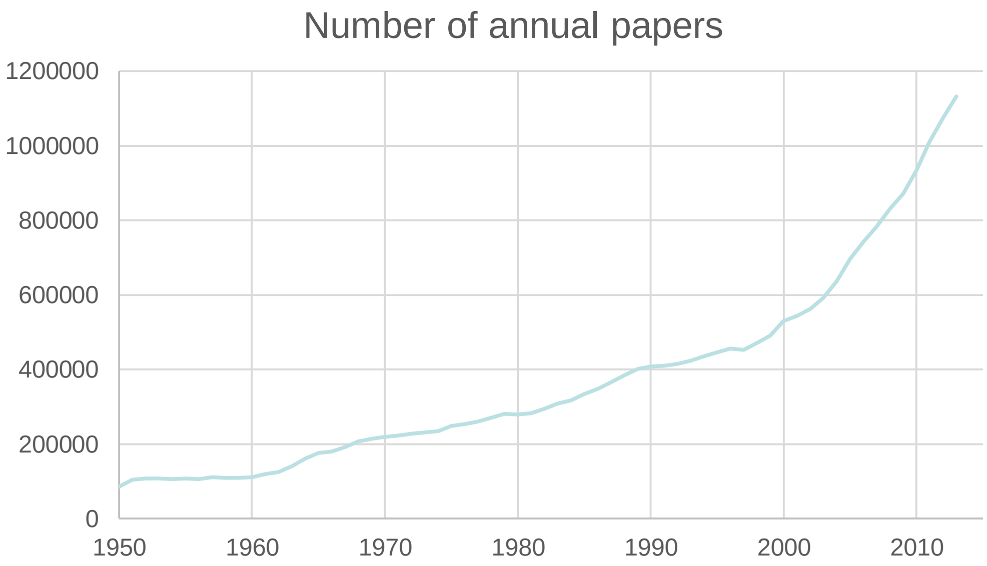


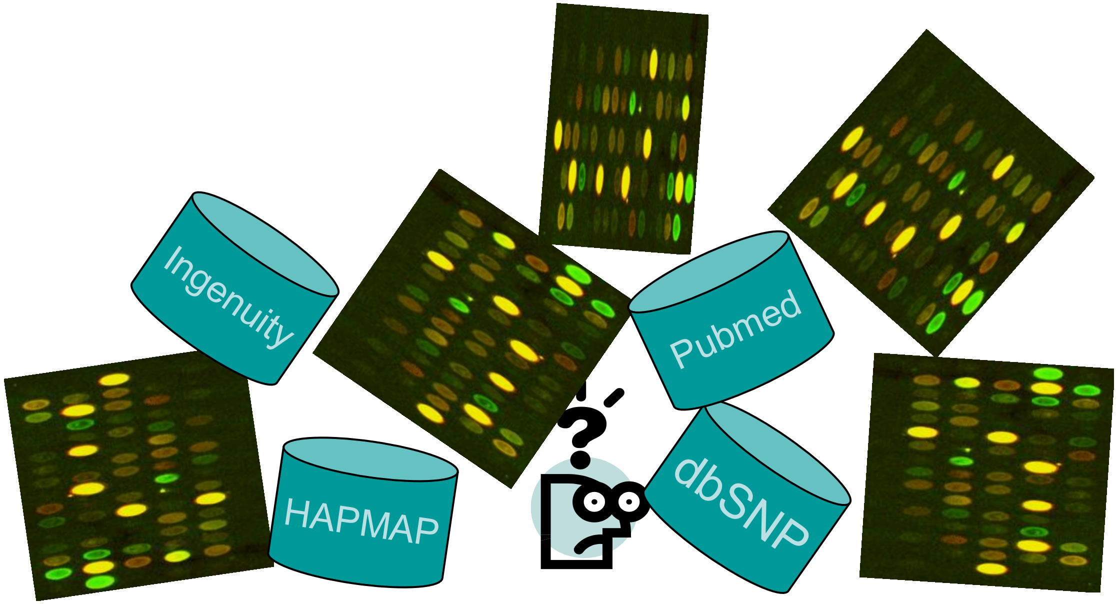
Fig. 2. CUMULATIVE NUMBER OF ABSTRACTS IN VARIOUS SCIENTIFIC FIELDS, FROM THE BEGINNING OF THE ABSTRACT SERVICE TO GIVEN DATE

It will be noted that after an initial period of rapid expansion to a stable growth rate, the number of abstracts increases exponentially, doubling in approximately 15 years.

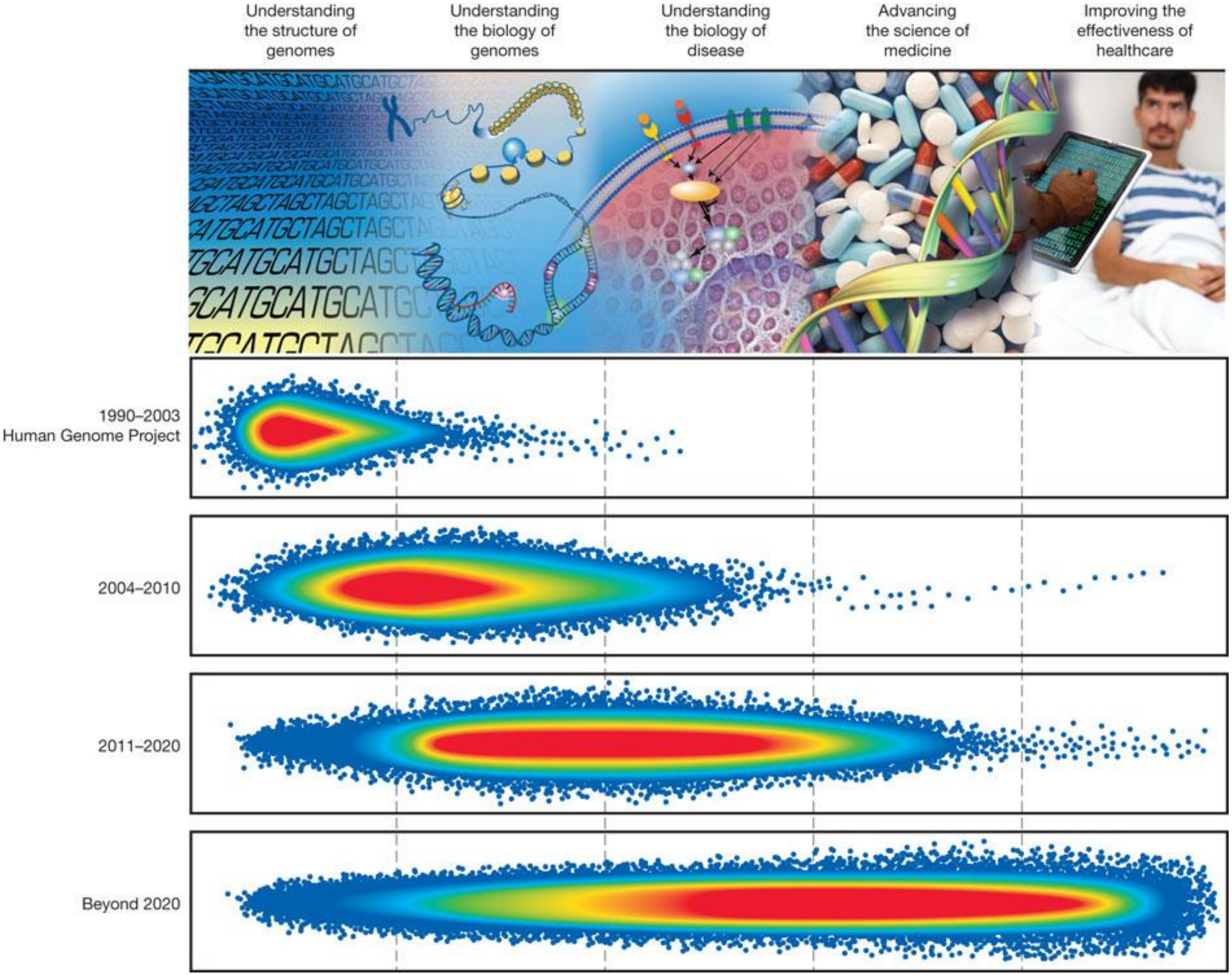
Little Science, Big Science, by Derek J. de Solla Price, 1963



The fusion bottleneck
(~limits of personal cognition)



Accomplishments of genomics research



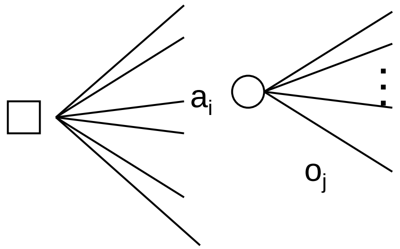
E D. Green *et al*. *Nature* **470**, 204-213 (2011) doi:10.1038/nature09764

Biomedical decision support

Optimal decision: decision theory probability theory+utility theory

- Decision situation:
  - Actions a_i
  - Outcomes o_j
  - Probabilities of outcomes $p(o_j \mid a_i)$
  - Utilities/losses of outcomes $U(o_j \mid a_i)$
  - Maximum Expected Utility Principle (MEU) $EU(a_i) = \sum_j U(o_j \mid a_i) p(o_j \mid a_i)$
  - Best action is the one with maximum expected utility $a^* = \arg\max_i EU(a_i)$

Actions a_i Outcomes Probabilities Utilities, costs Expected utilities



$P(o_j|a_i)$ $U(o_j)$, $C(a_i)$ $\Big\}$ $EU(a_i) = \sum P(o_j|a_i)U(o_j)$

$\vdots$ $\vdots$

Types of inference

- (Passive, observational) inference
  - P(Query|Observations, Observational data)
- Interventionist inference
  - P(Query|Observations, Interventions)
- Counterfactual inference
  - P(Query| Observations, Counterfactual conditionals)

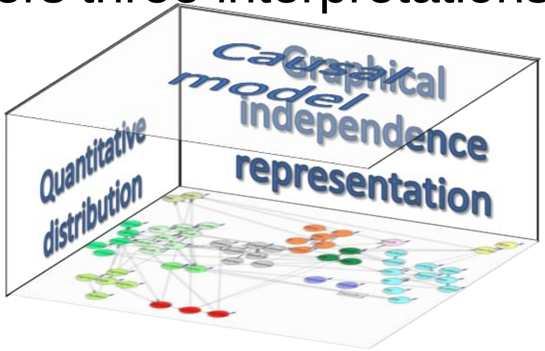
- Biomedical applications
  - Prevention
  - Screening
  - Diagnosis
  - Therapy selection
  - Therapy modification
  - Evaluation of therapic efficiancy

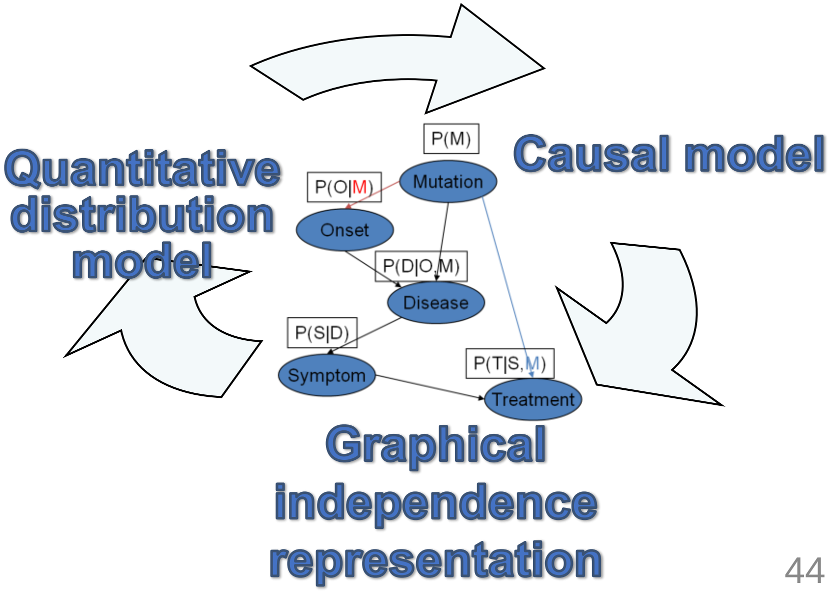
Probabilistic graphical models: Bayesian Networks

- A directed acyclic graph (DAG)
- Nodes are random variables
- Edges represent direct dependence (causal relationship)
- Local models: $P(X_i|Pa(X_i))$
- Offers three interpretations

Thomas Bayes (c. 1702 – 1761)

$$P(Model \mid Data) \propto P(Data \mid Model)P(Model)$$


Causal model
Graphical independence representation
Quantitative distribution

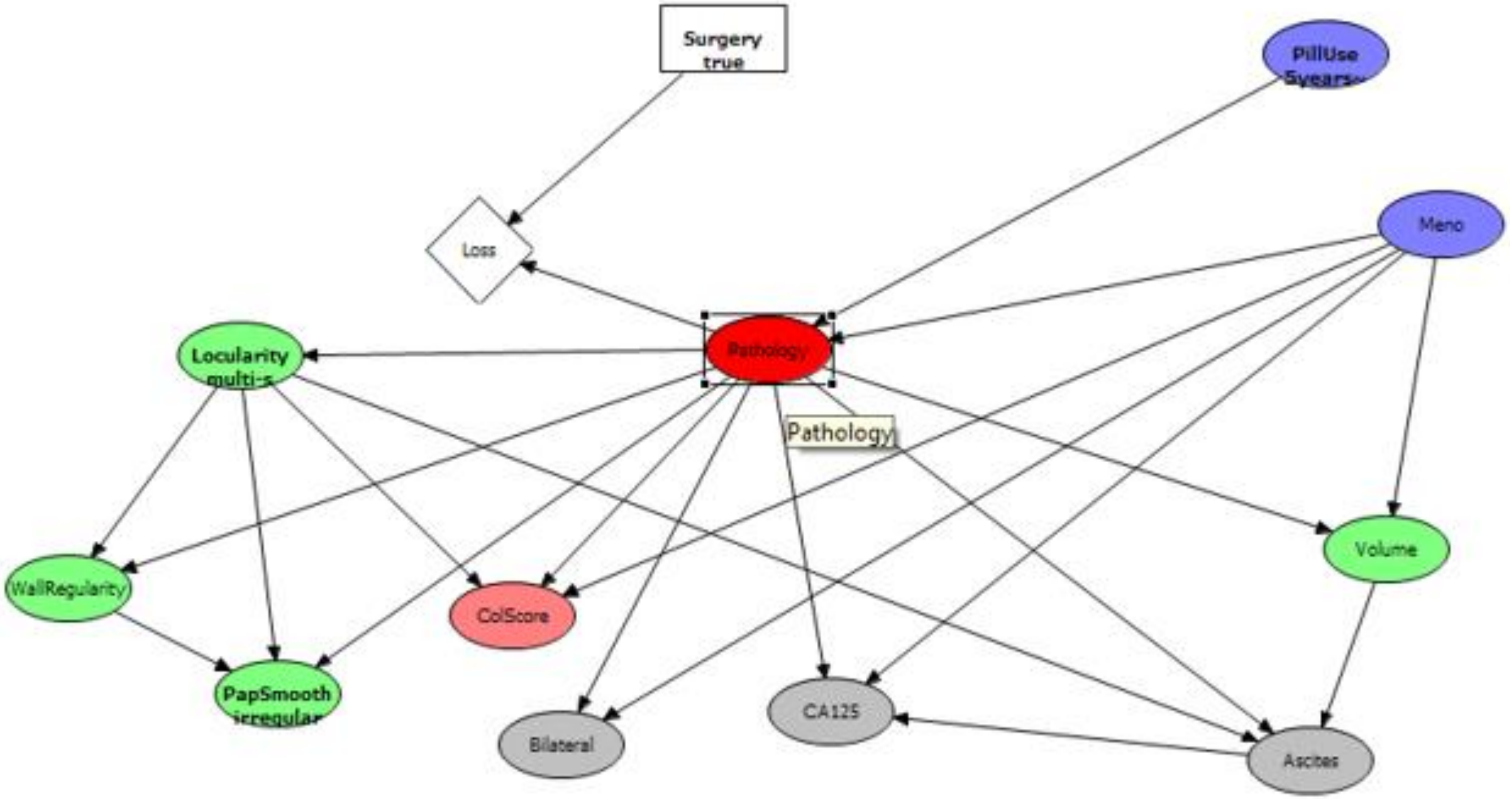


Quantitative distribution model

Causal model

Graphical independence representation

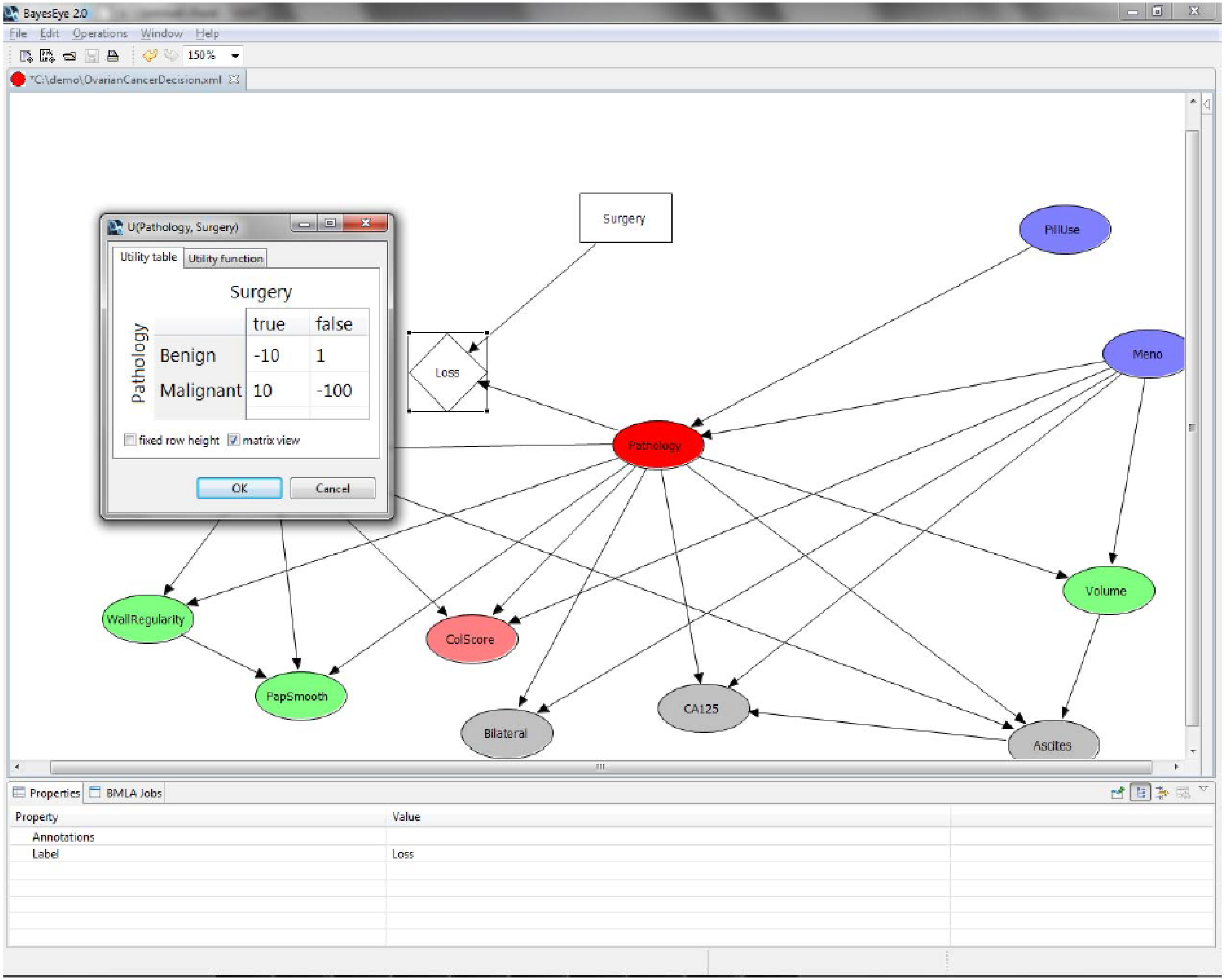
Ovarian tumor diagnostics



International Ovarian Tumor Analysis (IOTA, Dirk Timmerman)

Antal, P., Fannes, G., Timmerman, D., Moreau, Y. and De Moor, B., 2004. Using literature and data to learn Bayesian networks as clinical models of ovarian tumors. *Artificial Intelligence in medicine*, *30*(3), pp.257-281.

Decision networks

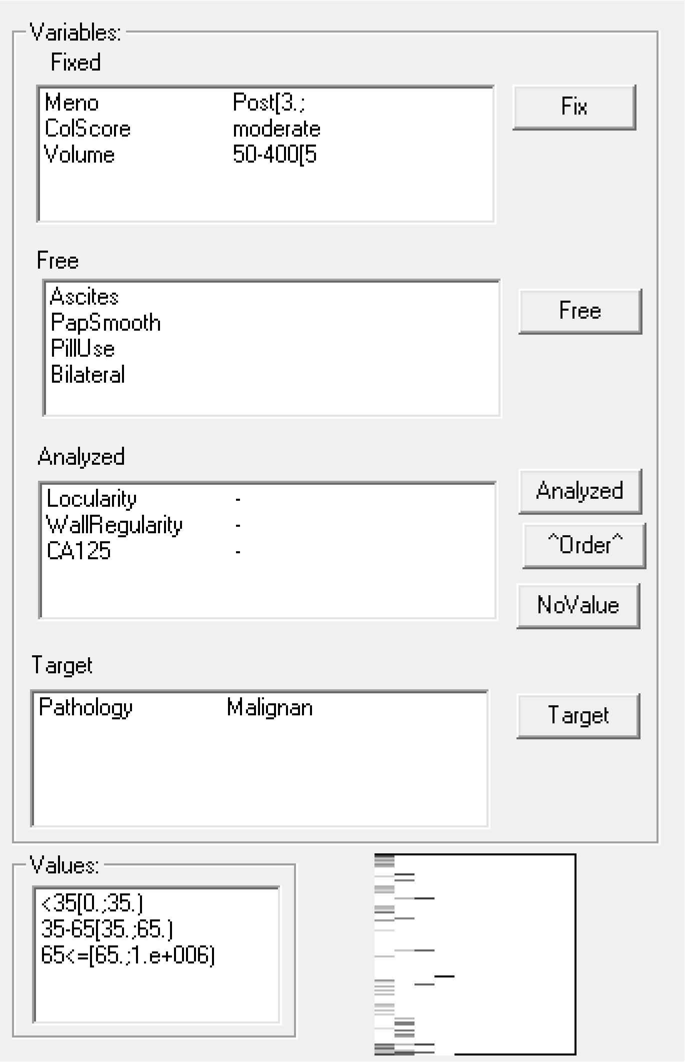


	Surgery	
Pathology	true	false
Benign	-10	1
Malignant	10	-100

□ fixed row height ☑ matrix view

Property	Value
Annotations	
Label	Loss

Sensitivity of the inference



Variables:

Fixed

Meno	Post[3.;
ColScore	moderate
Volume	50-400[5

Free

Ascites
PapSmooth
PillUse
Bilateral

Analyzed

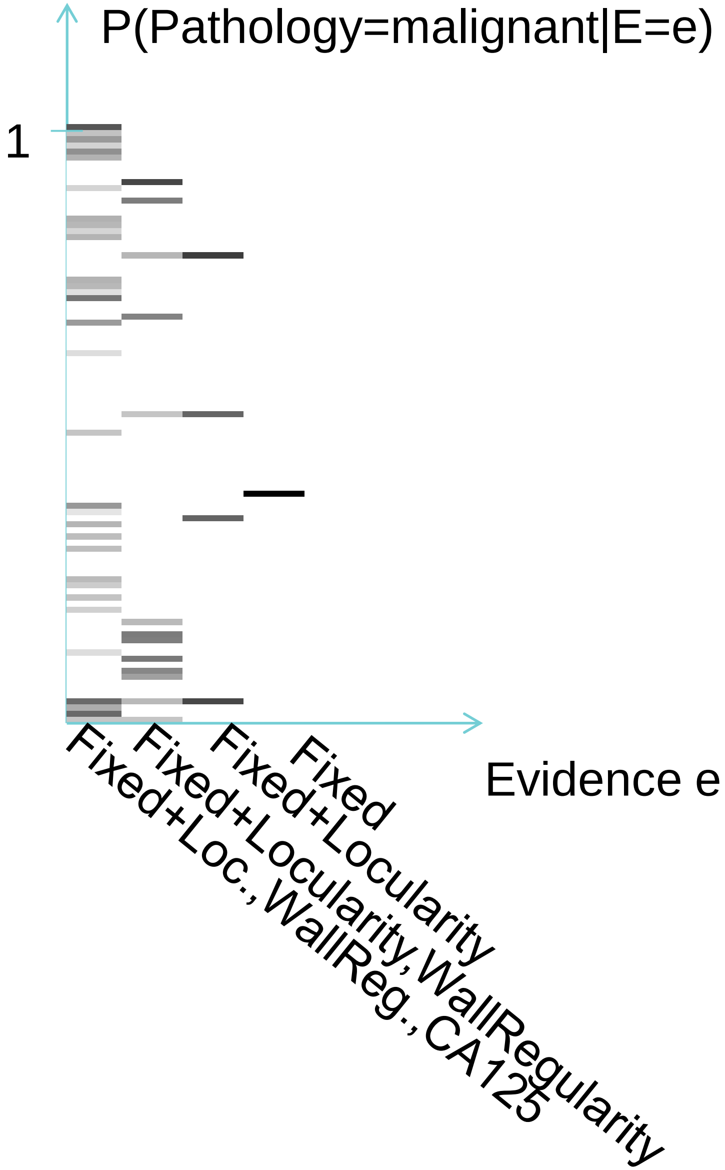
Locularity	-
WallRegularity	-
CA125	-

Target

Pathology	Malignan

Values:

<35[0.;35.)
35-65[35.;65.)
65<=[65.;1.e+006)



FUSION

Pharma productivity (~gap)

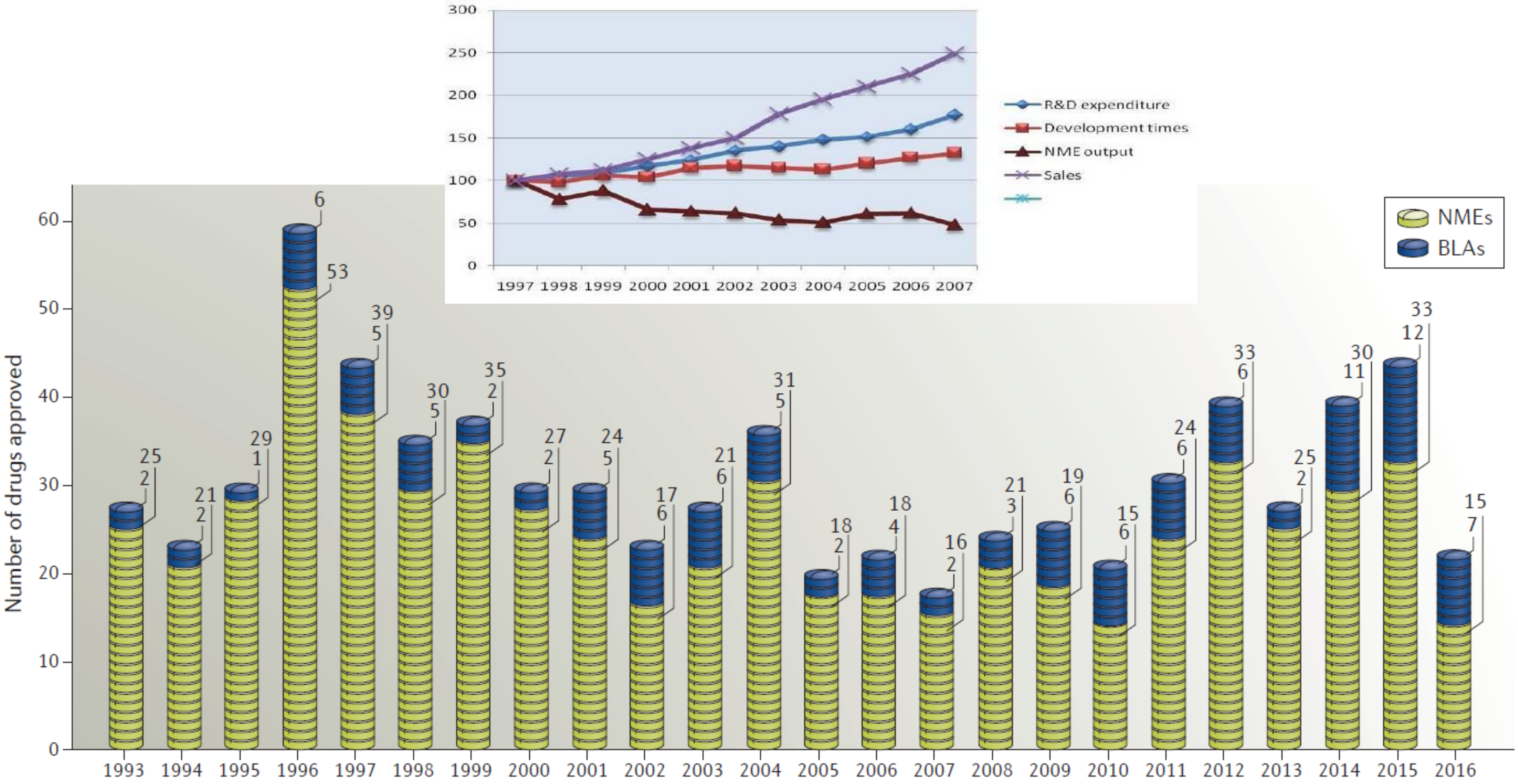
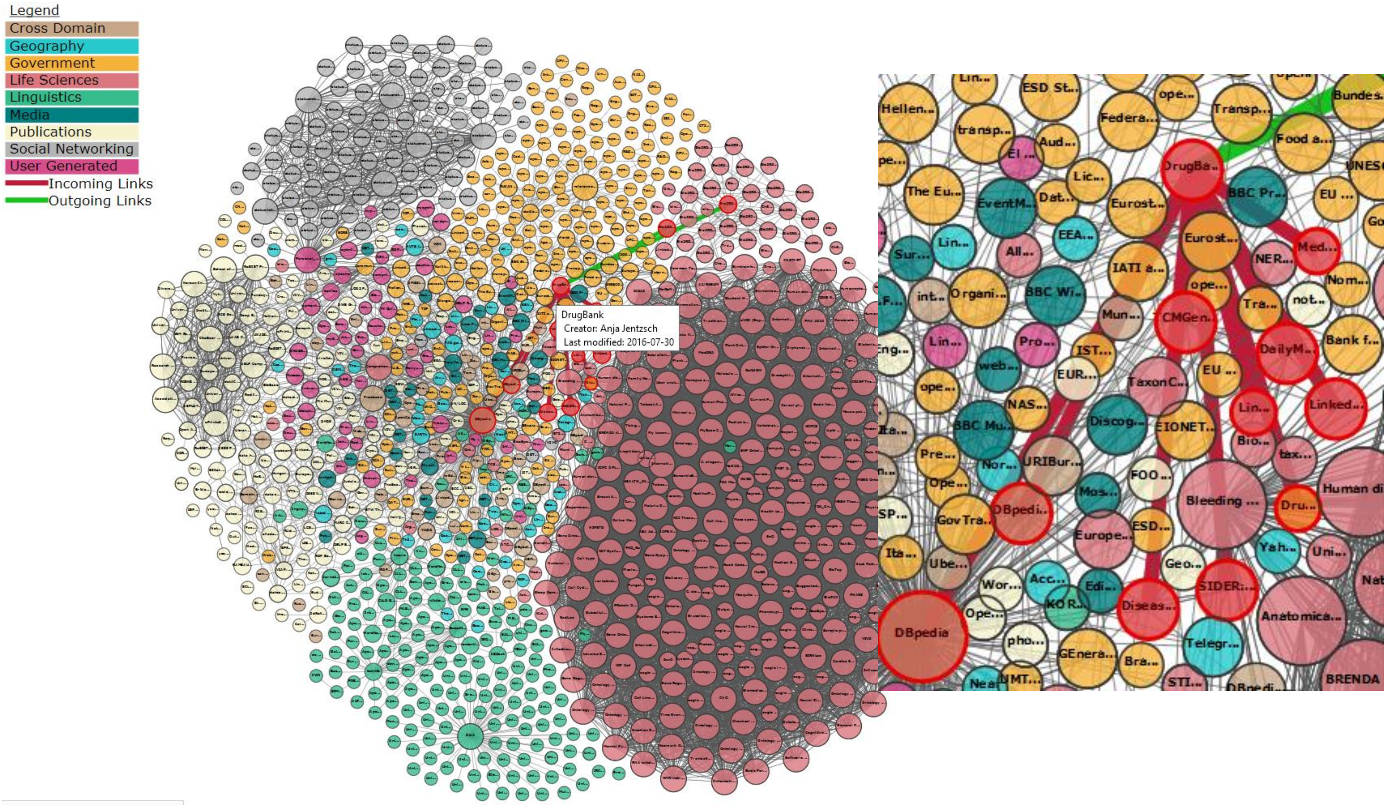


Figure 1 | **Novel FDA approvals since 1993.** New molecular entities (NMEs) and biologics licence applications (BLAs) approved by the Center for Drug Evaluation and Research (CDER) since 1993 (see also TABLE 1). Approvals by the Center for Biologics Evaluation and Research (CBER) are not included in this drug count (see TABLE 3). Data are from Drugs@FDA.

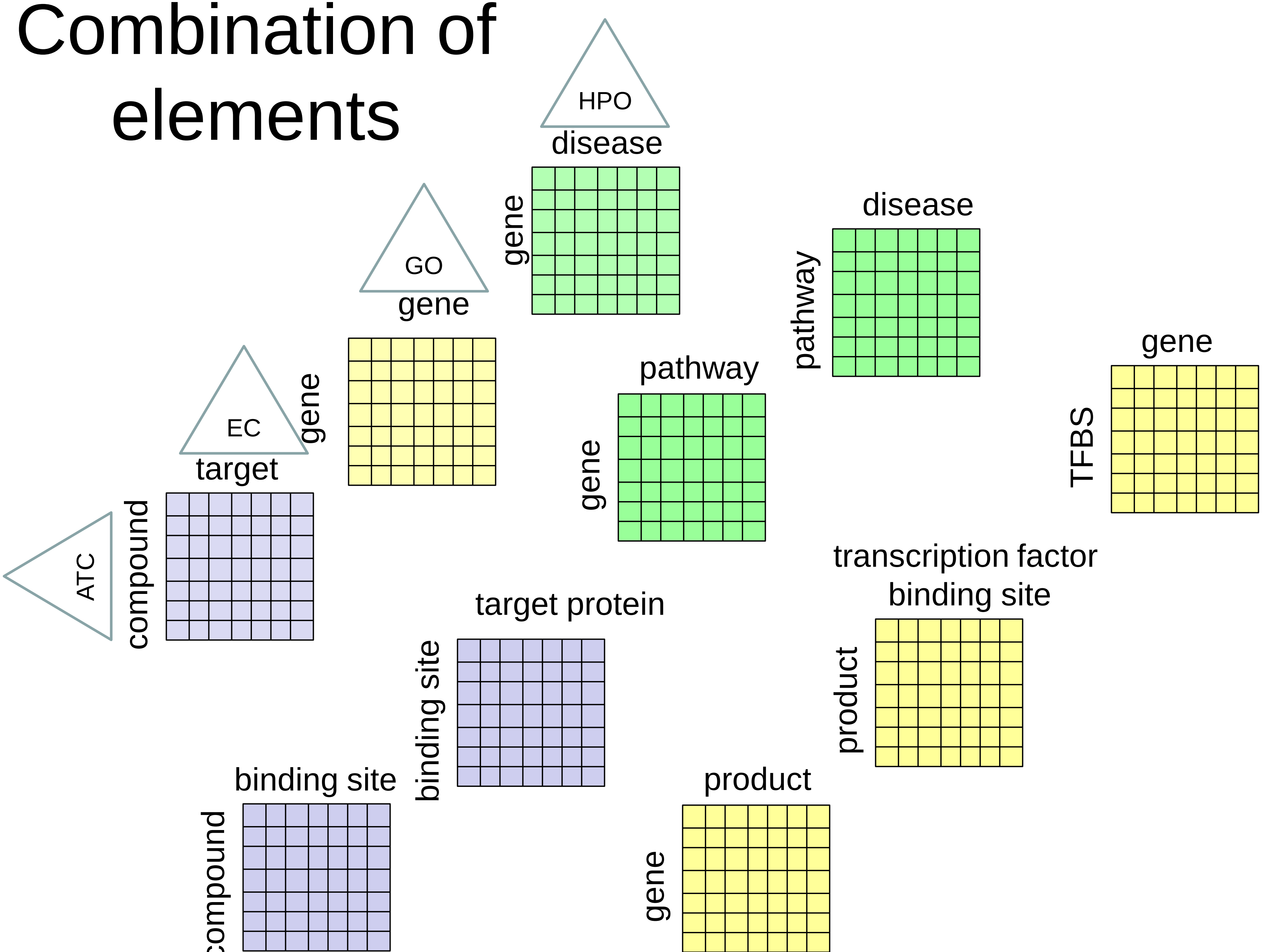
Mullard, A., 2017. 2016 FDA drug approvals. *Nature Reviews Drug Discovery*, *16*(2), pp.73-76.

Linked Open Data: ~2020



Linking Open Data cloud diagram 2017, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/

Combination of elements



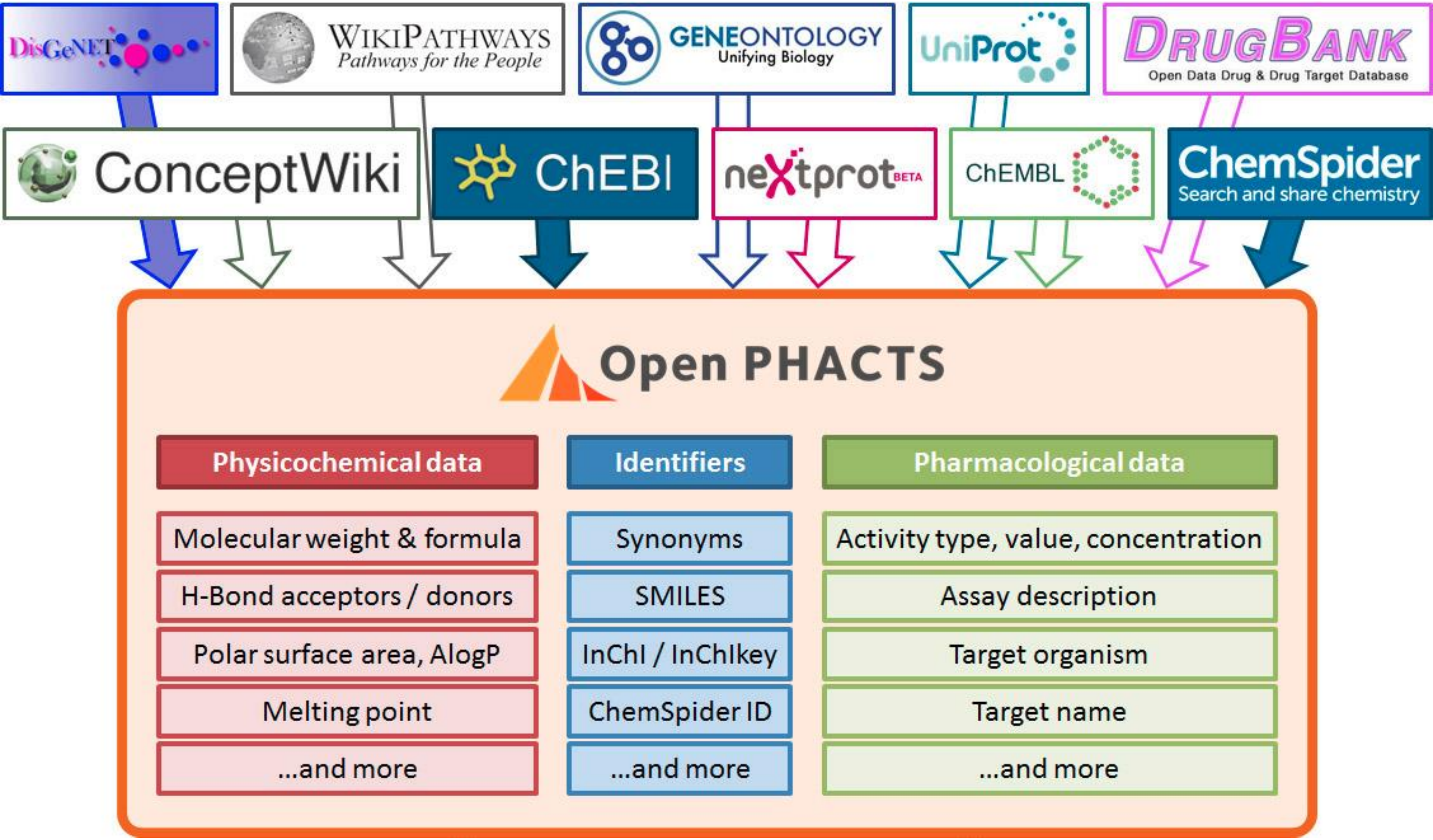
Open Pharmacological Space

Open PHACTS

Precursor: Gene Ontology: tool for the unification of biology, Nature, 2000

- Discovery Platform for **cross-domain** fusion.
- Public, curated, linked data.
  - The data sources you **already use**, **integrated** and **linked** together: *compounds*, *targets*, *pathways*, *diseases* and *tissues*.
- Everything in **triples**: **Subject-predicate-object**

Open PHACTS

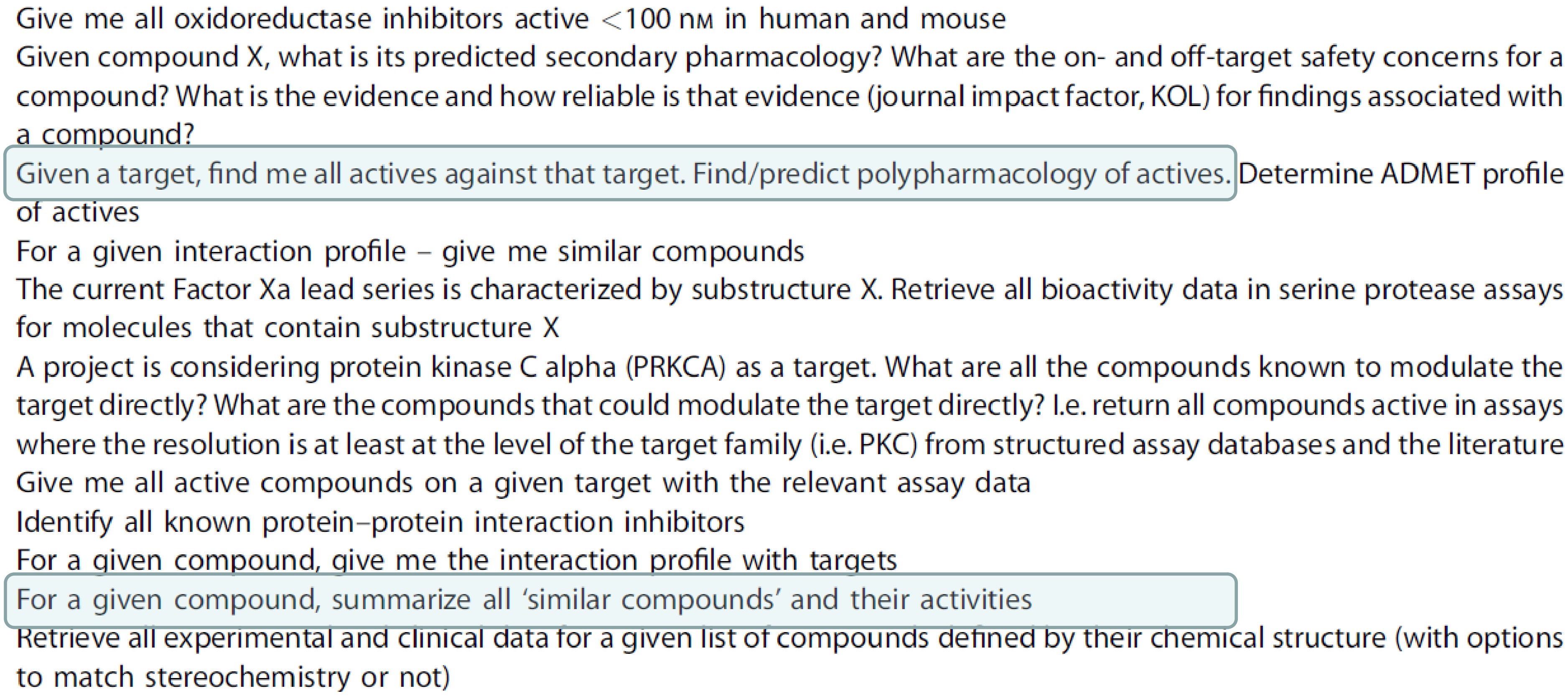


Physicochemical data	Identifiers	Pharmacological data
Molecular weight & formula	Synonyms	Activity type, value, concentration
H-Bond acceptors / donors	SMILES	Assay description
Polar surface area, AlogP	InChI / InChIkey	Target organism
Melting point	ChemSpider ID	Target name
...and more	...and more	...and more

Open PHACTS

- **Discovery Platform** to cross barriers.
- The data sources you **already use**, **integrated** and **linked** together: *compounds*, *targets*, *pathways*, *diseases* and *tissues*.
- ChEBI, ChEMBL, ChemSpider, ConceptWiki, DisGeNET, DrugBank, Gene Ontology, neXtProt, UniProt and WikiPathways.
- **For questions** in drug discovery, **answers** from publications in peer reviewed scientific journals.

Top questions in the pharma industry I. (Open PHACTS)



Give me all oxidoreductase inhibitors active <100 nM in human and mouse

Given compound X, what is its predicted secondary pharmacology? What are the on- and off-target safety concerns for a compound? What is the evidence and how reliable is that evidence (journal impact factor, KOL) for findings associated with a compound?

Given a target, find me all actives against that target. Find/predict polypharmacology of actives. Determine ADMET profile of actives

For a given interaction profile – give me similar compounds

The current Factor Xa lead series is characterized by substructure X. Retrieve all bioactivity data in serine protease assays for molecules that contain substructure X

A project is considering protein kinase C alpha (PRKCA) as a target. What are all the compounds known to modulate the target directly? What are the compounds that could modulate the target directly? I.e. return all compounds active in assays where the resolution is at least at the level of the target family (i.e. PKC) from structured assay databases and the literature

Give me all active compounds on a given target with the relevant assay data

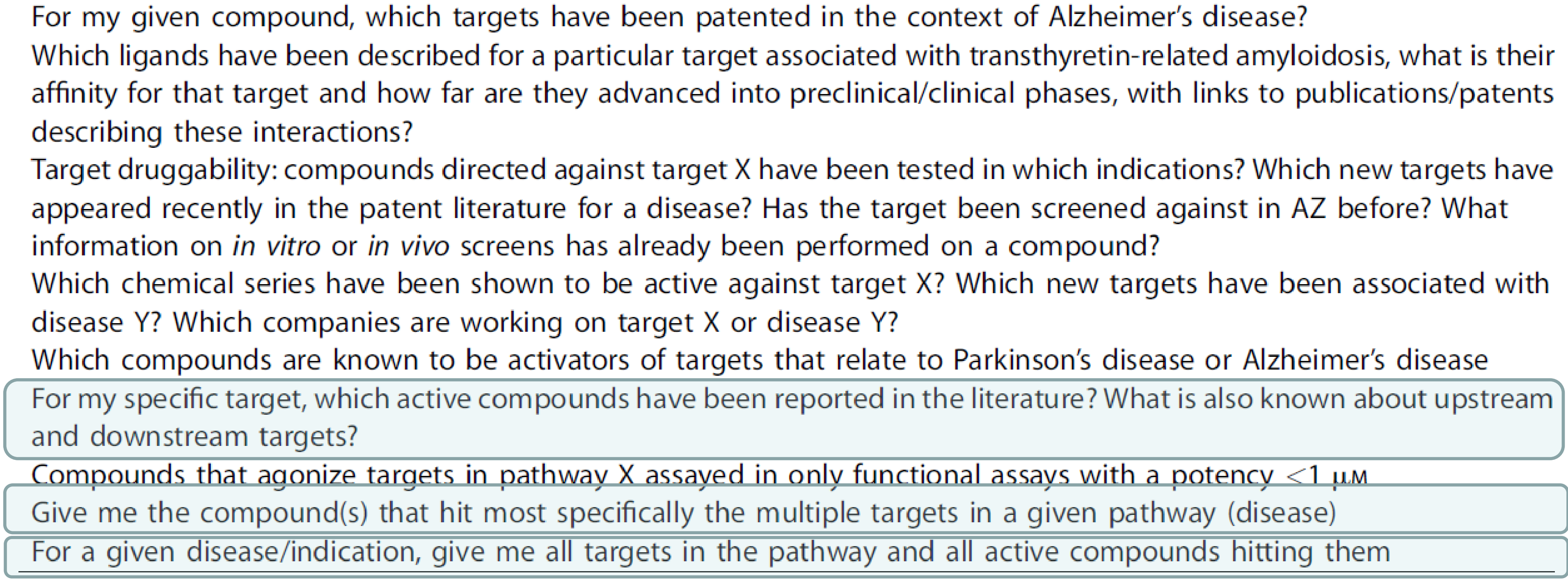
Identify all known protein–protein interaction inhibitors

For a given compound, give me the interaction profile with targets

For a given compound, summarize all 'similar compounds' and their activities

Retrieve all experimental and clinical data for a given list of compounds defined by their chemical structure (with options to match stereochemistry or not)

Top questions II.



For my given compound, which targets have been patented in the context of Alzheimer's disease?

Which ligands have been described for a particular target associated with transthyretin-related amyloidosis, what is their affinity for that target and how far are they advanced into preclinical/clinical phases, with links to publications/patents describing these interactions?

Target druggability: compounds directed against target X have been tested in which indications? Which new targets have appeared recently in the patent literature for a disease? Has the target been screened against in AZ before? What information on *in vitro* or *in vivo* screens has already been performed on a compound?

Which chemical series have been shown to be active against target X? Which new targets have been associated with disease Y? Which companies are working on target X or disease Y?

Which compounds are known to be activators of targets that relate to Parkinson's disease or Alzheimer's disease

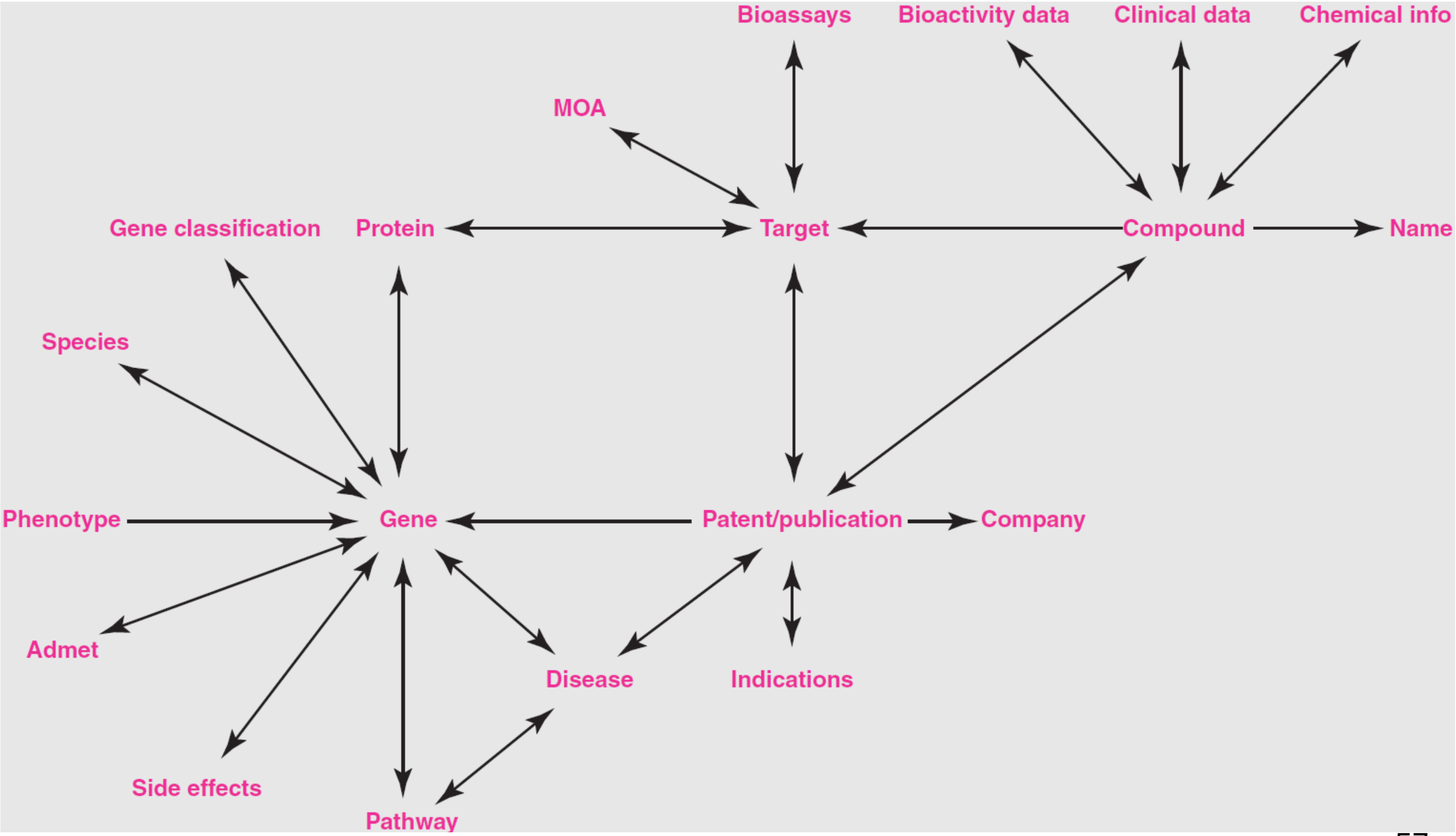
For my specific target, which active compounds have been reported in the literature? What is also known about upstream and downstream targets?

Compounds that agonize targets in pathway X assayed in only functional assays with a potency <1 μM

Give me the compound(s) that hit most specifically the multiple targets in a given pathway (disease)

For a given disease/indication, give me all targets in the pathway and all active compounds hitting them

Open PHACTS: databases



Open PHACTS

Dataset	Downloaded	Version	Licence	Triples
Bio Assay Ontology			CC-By	10,360
CALOHA	8 Apr 2015	2014-01-22	CC-By-ND	14,552
ChEBI	4 Mar 2015	125	CC-By-SA	1,012,056
ChEMBL	18 Feb 2015	20.0	CC-By-SA	445,732,880
ConceptWiki	12 Dec 2013		CC-By-SA	4,331,760
DisGeNET	31 Mar 2015	2.1.0	ODbL	15,011,136
Disease Ontology		2015-05-21	CC-By	188,062
DrugBank	19 Feb 2015	4.1	Non-commercial	4,028,767
ENZYME		2015_11	CC-By-ND	61,467
FDA Adverse Events	9 Jul 2012		CC0	13,557,070

Total: ~3 Billion triples

Open PHACTS

Dataset	Downloaded	Version	Licence	Triples
Gene Ontology	4 Mar 2015		CC-By	1,366,494
Gene Ontology Annotations	17 Feb 2015		CC-By	879,448,347
NCATS OPDDR	Nov 2015	Oct 2015		2,643
neXTProt (NP)	1 Feb 2014	1.0	CC-By-ND	215,006,108
OPS Chemical Registry		4 Nov 2014	CC-By-SA	241,986,722
HMDB		*3.6*	*HMDB*	
MeSH		*2015*	*MeSH*	
PDB Ligands		*2*	*PDB*	
OPS Metadata			CC-By-SA	2,053
UniProt		2015_11	CC-By-ND	1,131,186,434
WikiPathways		20151118	CC-By	11,781,627

Total: ~3 Billion triples

OPS: open tools for free academic use

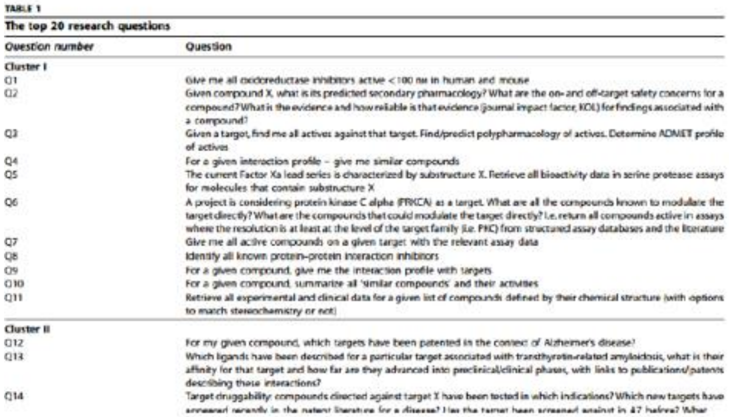
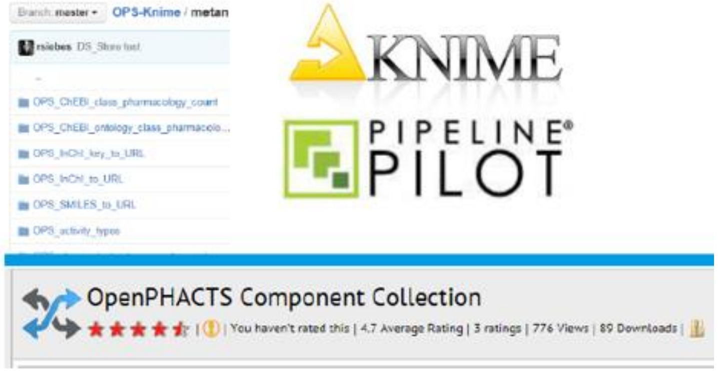


TABLE 1

The top 20 research questions

Question number	Question
Cluster I	
Q1	Give me all oxidoreductase inhibitors active <100 nM in human and mouse
Q2	Given compound X, what is its predicted secondary pharmacology? What are the on- and off-target safety concerns for a compound? What is the evidence and how reliable is that evidence (journal impact factor, KOL) for findings associated with a compound?
Q3	Given a target, find me all actives against that target. Find/predict polypharmacology of actives. Determine ADMET profile of actives
Q4	For a given interaction profile – give me similar compounds
Q5	The current Factor Xa lead series is characterized by substructure X. Retrieve all bioactivity data in serine protease assays for molecules that contain substructure X
Q6	A project is considering protein kinase C alpha (PRKCA) as a target. What are all the compounds known to modulate the target directly? What are the compounds that could modulate the target directly? I.e. return all compounds active in assays where the resolution is at least at the level of the target family (i.e. PKC) from structured assay databases and the literature
Q7	Give me all active compounds on a given target with the relevant assay data
Q8	Identify all known protein-protein interaction inhibitors
Q9	For a given compound, give me the interaction profile with targets
Q10	For a given compound, summarize all 'similar compounds' and their activities
Q11	Retrieve all experimental and clinical data for a given list of compounds defined by their chemical structure (with options to match stereochemistry or not)
Cluster II	
Q12	For my given compound, which targets have been patented in the context of Alzheimer's disease?
Q13	Which ligands have been described for a particular target associated with transthyretin-related amyloidosis, what is their affinity for that target and how far are they advanced into preclinical/clinical phases, with links to publications/patents describing these interactions?
Q14	Target druggability: compounds directed against target X have been tested in which indications? Which new targets have appeared recently in the patent literature for a disease? Has the target been screened against in AZ before? What

The list of relevant research questions



Branch: master ▾ OPS-Knime / metan

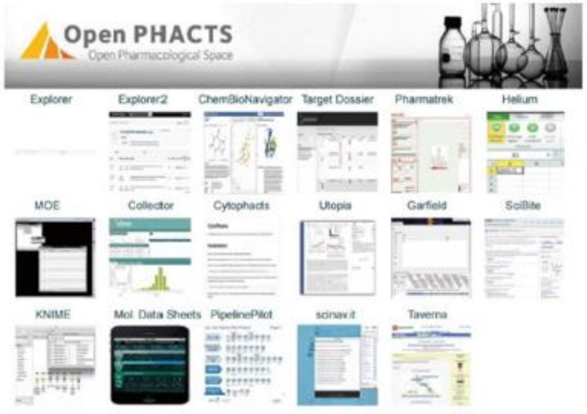
rsiebes DS_Store lost

..

- OPS_ChEBI_class_pharmacology_count
- OPS_ChEBI_ontology_class_pharmacolo...
- OPS_InChI_key_to_URL
- OPS_InChI_to_URL
- OPS_SMILES_to_URL
- OPS_activity_types

OpenPHACTS Component Collection

★★★★☆ | You haven't rated this | 4.7 Average Rating | 3 ratings | 776 Views | 89 Downloads |

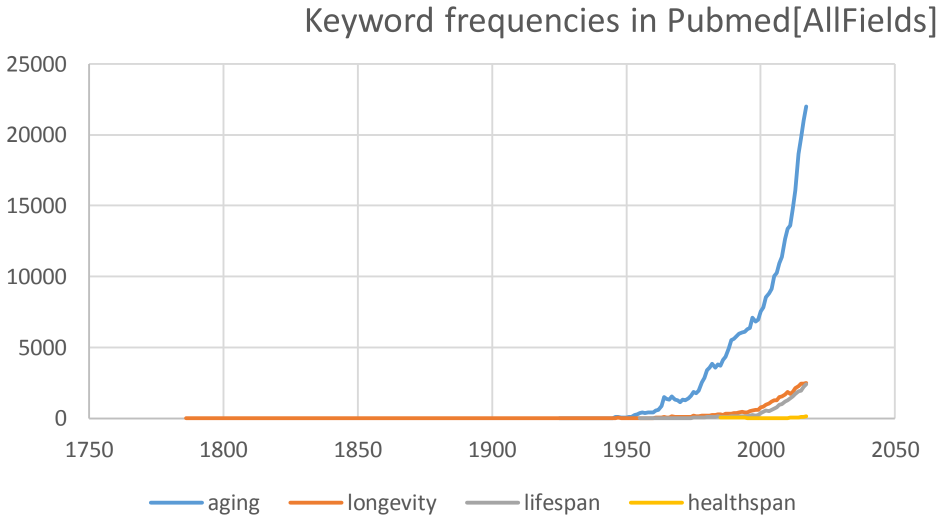


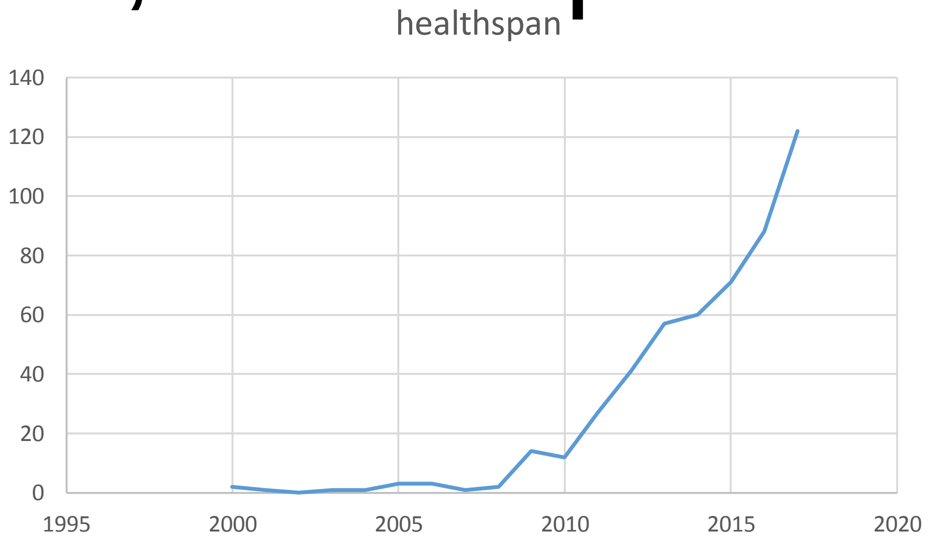
http://www.openphactsfoundation.org/apps.html

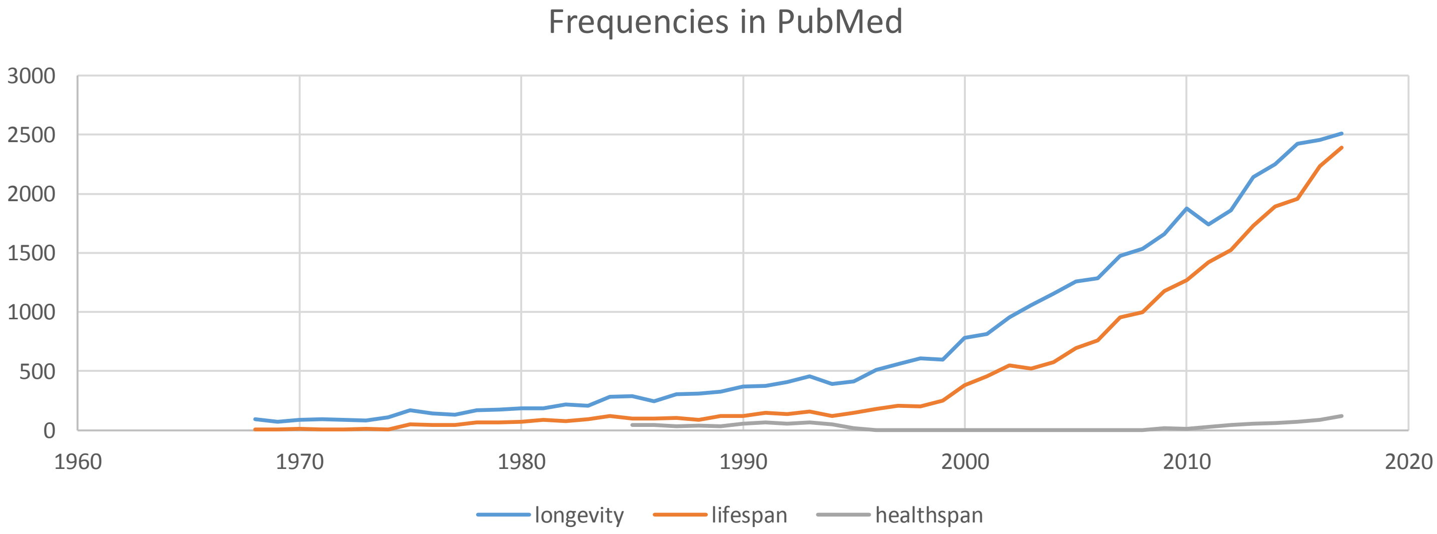
The solutions to answer them

FUSION in aging research

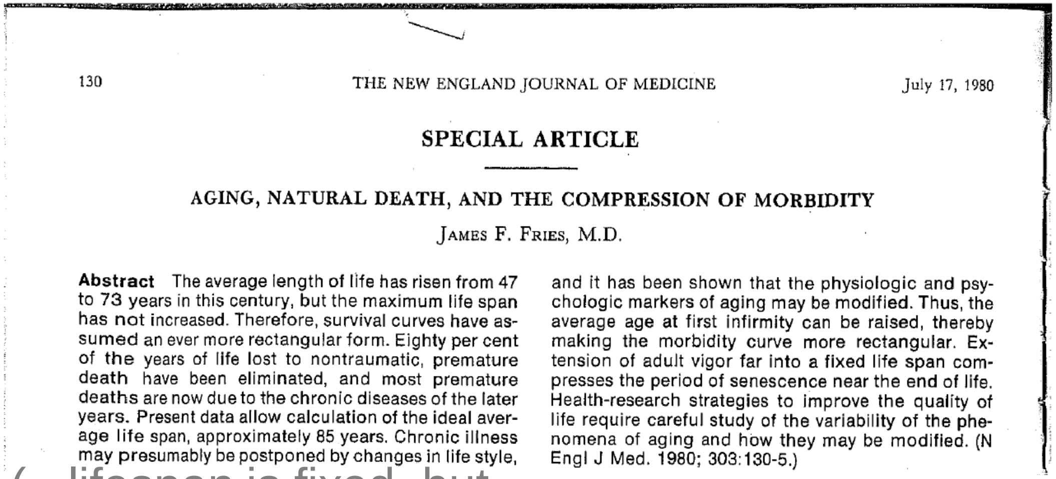
Number of publications related to aging, lifespan, healthspan







Healthspan



130 THE NEW ENGLAND JOURNAL OF MEDICINE July 17, 1980

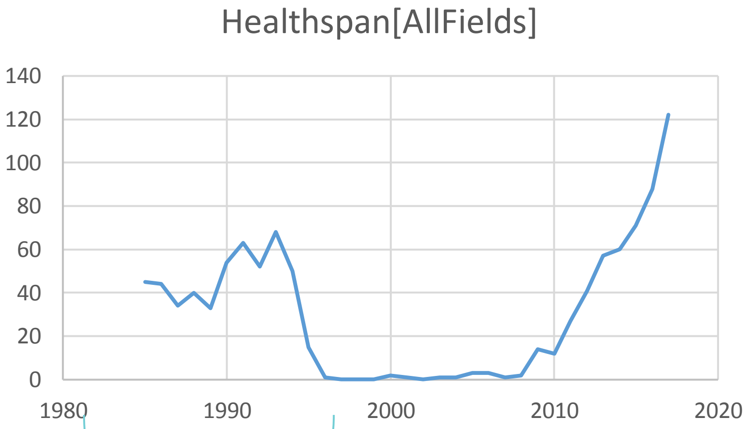
SPECIAL ARTICLE

AGING, NATURAL DEATH, AND THE COMPRESSION OF MORBIDITY

JAMES F. FRIES, M.D.

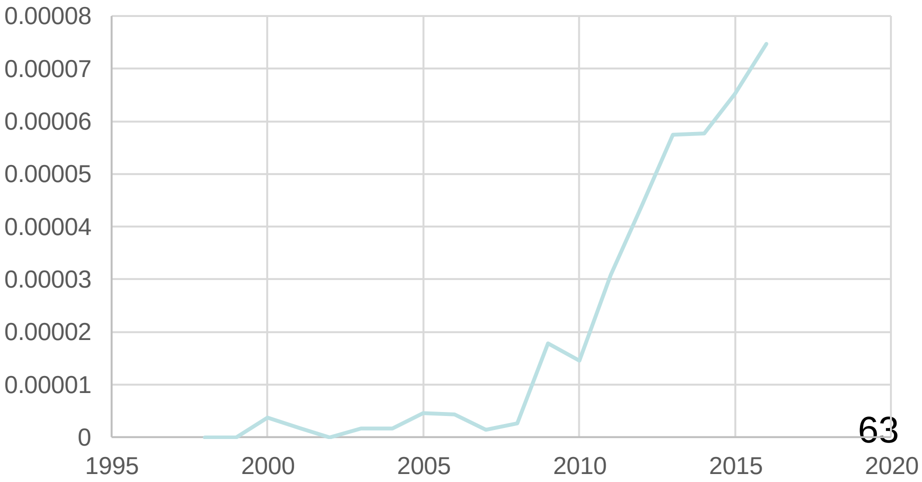
Abstract The average length of life has risen from 47 to 73 years in this century, but the maximum life span has not increased. Therefore, survival curves have assumed an ever more rectangular form. Eighty per cent of the years of life lost to nontraumatic, premature death have been eliminated, and most premature deaths are now due to the chronic diseases of the later years. Present data allow calculation of the ideal average life span, approximately 85 years. Chronic illness may presumably be postponed by changes in life style, and it has been shown that the physiologic and psychologic markers of aging may be modified. Thus, the average age at first infirmity can be raised, thereby making the morbidity curve more rectangular. Extension of adult vigor far into a fixed life span compresses the period of senescence near the end of life. Health-research strategies to improve the quality of life require careful study of the variability of the phenomena of aging and how they may be modified. (N Engl J Med. 1980; 303:130-5.)

(~ lifespan is fixed, but morbidity/senescence can be compressed by modifiable life style factors)



„Healthspan” journal

Healthspan[Title] ratio



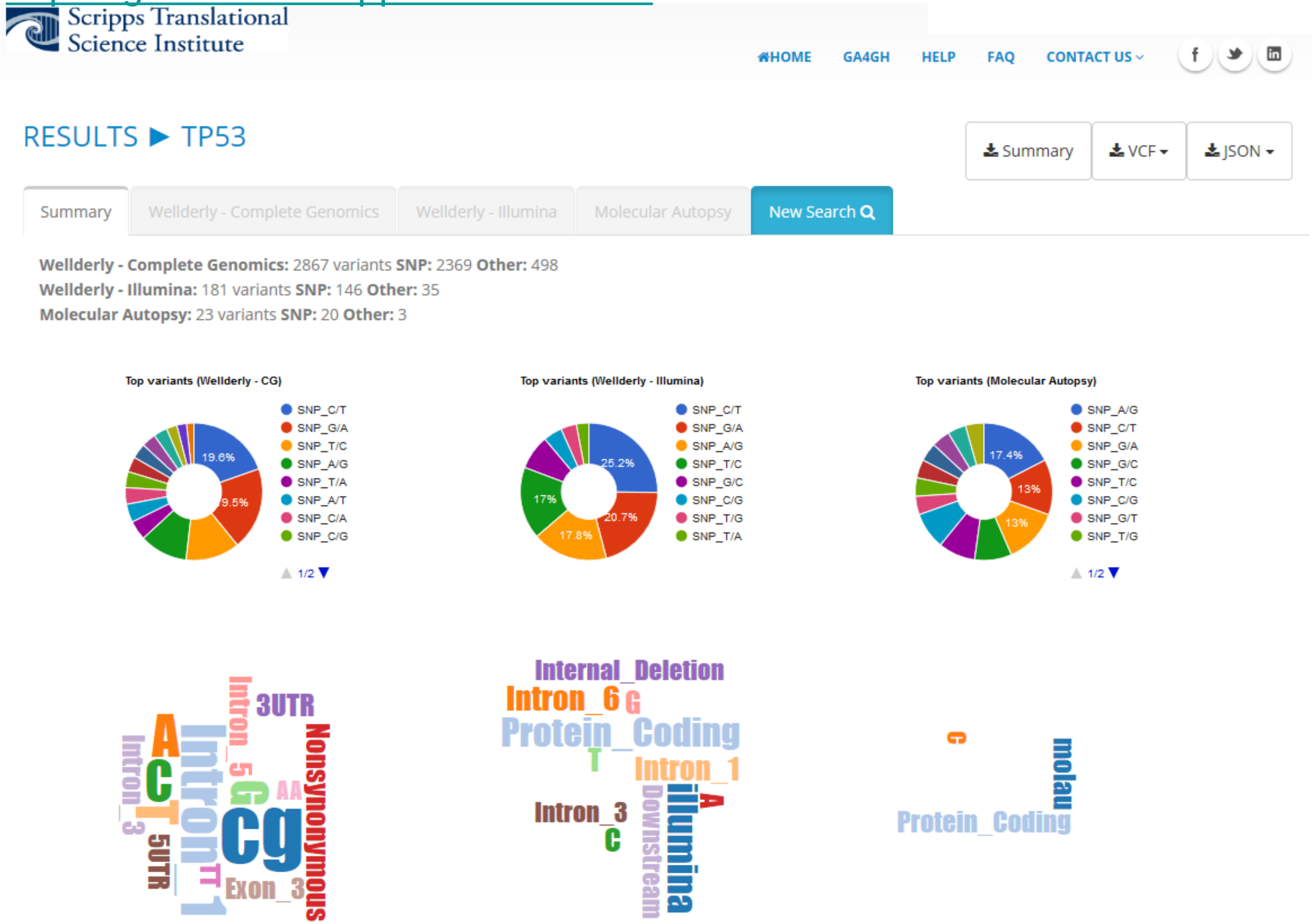
Scripps Translational
Science Institute

The Wellderly study

- Erikson G.A. et al.: Whole Genome Sequencing of a Healthy Aging Cohort, Cell, 2016
  - Definition of **the „Wellderly phenotype**: >80 years old with no chronic diseases and who are not taking chronic medications + exclusions with any autoimmune disease, blood clots, cancer, type I or II diabetes, dementia, myocardial infarction, renal failure, and stroke.
  - WGS 600 wellderly people (511 analyzed)

Wellderly variants on the web

https://genomics.scripps.edu/browser/



Scripps Translational Science Institute

HOME GA4GH HELP FAQ CONTACT US

RESULTS ▶ TP53

Summary VCF JSON

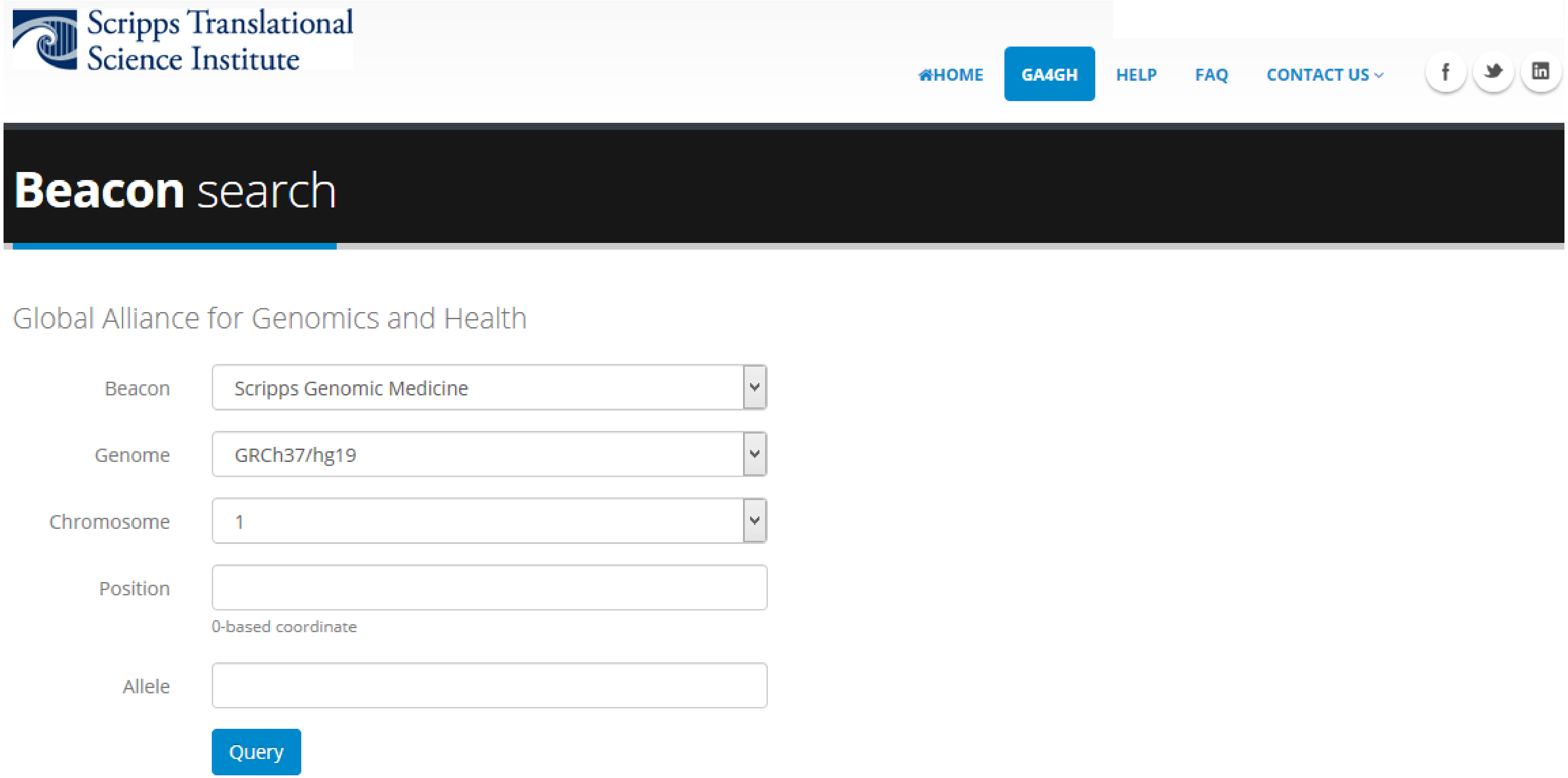
Summary | Wellderly - Complete Genomics | Wellderly - Illumina | Molecular Autopsy | New Search

Wellderly - Complete Genomics: 2867 variants **SNP:** 2369 **Other:** 498

Wellderly - Illumina: 181 variants **SNP:** 146 **Other:** 35

Molecular Autopsy: 23 variants **SNP:** 20 **Other:** 3

Sharing variants in research



Beacon search

Global Alliance for Genomics and Health

Beacon	Scripps Genomic Medicine
Genome	GRCh37/hg19
Chromosome	1
Position	
	0-based coordinate
Allele	

Global Alliance for Genomics and Health (GA4GH)

A Page, D Baker, M Bobrow, K Boycott, J Burn, S Chanock, et al. Genomics. a federated ecosystem for sharing genomic, clinical data. global alliance for genomics and health. Science, 352(6291):1278{1280, 2016.

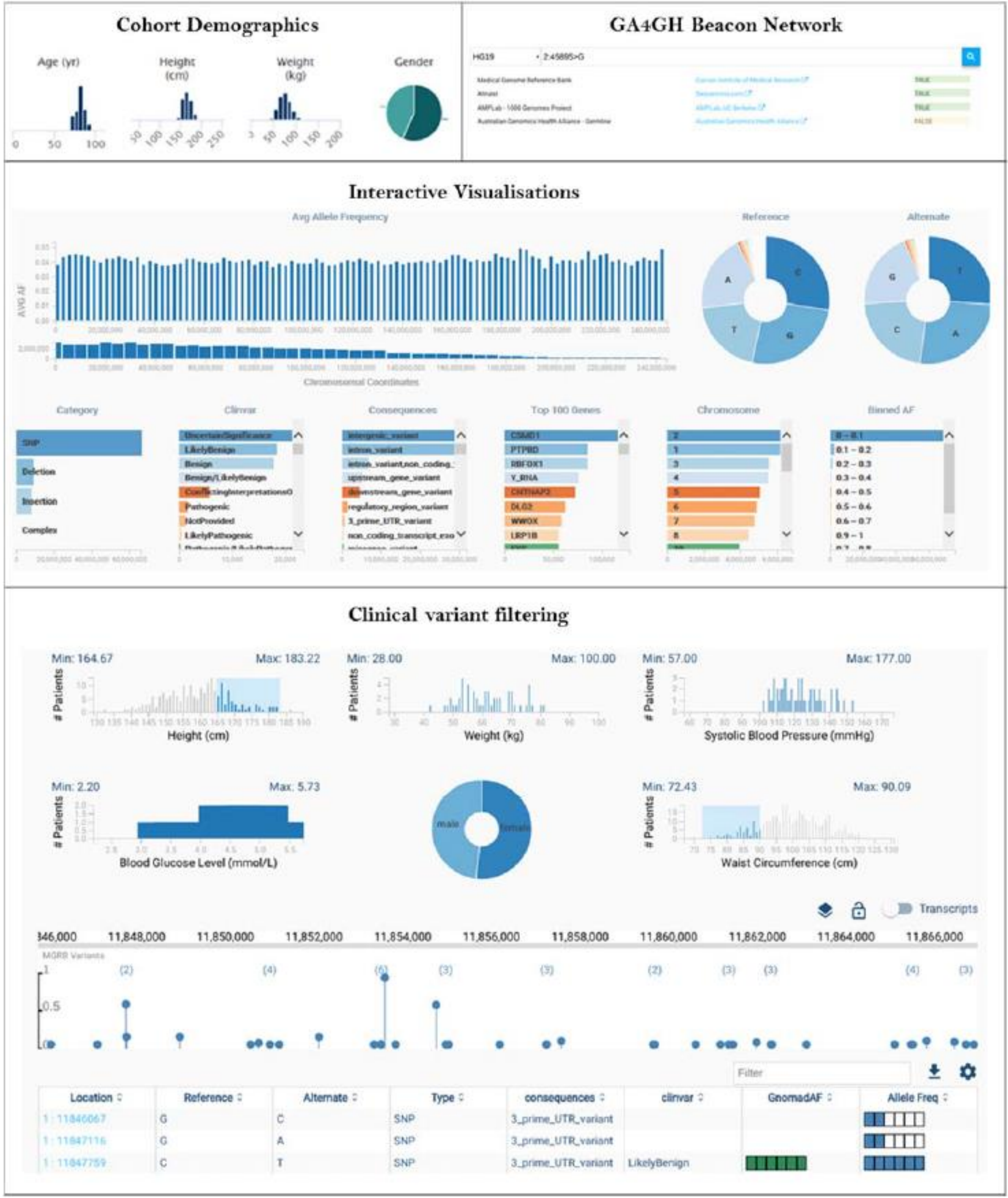
The largest healthspan cohort

Paul Lacaze, et al.: **The medical genome reference bank: a whole-genome data resource of 4,000 healthy elderly individuals. rationale and cohort design**. bioRxiv, page 274019, 2018.

Concept of non-penetrance

Lacaze, Paul, et al. "Pathogenic variants in the healthy elderly: unique ethical and practical challenges." *Journal of medical ethics* (2017)

Lacaze, Paul, et al.: "Penetrance and the Healthy Elderly." *Genetic testing and molecular biomarkers* 21.11 (2017)



HG19 2:45895>G

Medical Genome Reference Bank	Garvan Institute of Medical Research	TRUE
Altruist	Sequencing.com	TRUE
AMPLab - 1000 Genomes Project	AMPLab, UC Berkeley	TRUE
Australian Genomics Health Alliance - Germline	Australian Genomics Health Alliance	FALSE

Location	Reference	Alternate	Type	consequences	clinvar	GnomadAF	Allele Freq
1 : 11846067	G	C	SNP	3_prime_UTR_variant			
1 : 11847116	G	A	SNP	3_prime_UTR_variant			
1 : 11847759	C	T	SNP	3_prime_UTR_variant	LikelyBenign		

Reference or „control” genome?

- Genetic landscape of
  - age-associated/common diseases vs longevity
  - longevity/lifespan vs. healthspan

- M. Beekman et al. : **Genome-wide association study (GWAS)-identied disease risk alleles do not compromise human longevity**. PNAS, 107(42): 2010.
- Y. Freudenberg-Hua et al.: **Disease variants in genomes of 44 centenarians. Molecular genetics & genomic medicine**, 2(5):2014.
- M. Stevenson et al.: **Burden of disease variants in participants of the long life family study**. Aging (Albany NY), 7(2):123, 2015.
- L. C. Tindale et al.: **Burden of common complex disease variants in the exomes of two healthy centenarian brothers**. Gerontology, 62(1):58-62, 2016.
- S. CP Williams: **Genetic mutations you want**., PNAS, 113(10):2554-2557, 2016.
- P. Lacaze et al: **Penetrance and the healthy elderly**. Genetic testing and molecular biomarkers, 21(11):637-640, 2017.

Nemzeti Egészségtárház

- 2018-2020, Nemzeti Bionikai Program
- PI: Prof. Molnár Mária Judit (SE)
- SE-PPKE kooperáció
- Célok:
  - 100 „hosszú egészségű” **magyar** kohorsz
  - Teljes genom szekvenálás

    ➔ „referencia” genom/variáns készlet

    ➔ IT infrastruktúra eredmények megosztására

    ➔”egészséghossz”/healthspan kutatás

Summary

- NGS data analysis: 5 weeks
- Semantic technologies: 1 week:
- Chemoinformatics, drug discovery: 1 week
- GWAS data analysis: 3 weeks
- Biomed decision support: 2 weeks
- Causal data analysis: 1 lecture
- Guest lectures: phylogeny,..
- Cases studies

Thank you for your attention!