X-Informatics: Health Informatics

July 5 2013

Geoffrey Fox
gcf@indiana.edu
http://www.infomall.org/

Associate Dean for Research,
School of Informatics and Computing
Indiana University Bloomington
2013

Big Data Ecosystem in One Sentence

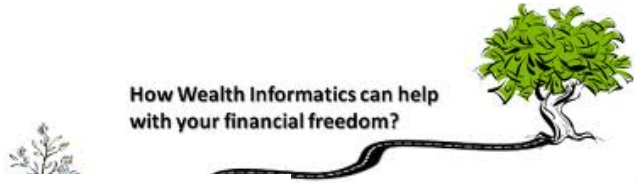
Use Clouds running Data Analytics Collaboratively processing Big Data to solve problems in X-Informatics (or e-X)

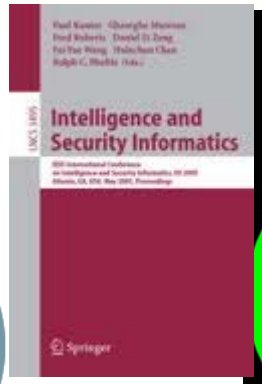
X = Astronomy, Biology, Biomedicine, Business, Chemistry, Climate, Crisis, Earth Science, Energy, Environment, Finance, Health, Intelligence, Lifestyle, Marketing, Medicine, Pathology, Policy, Radar, Security, Sensor, Social, Sustainability, Wealth and Wellness with more fields (physics) defined implicitly

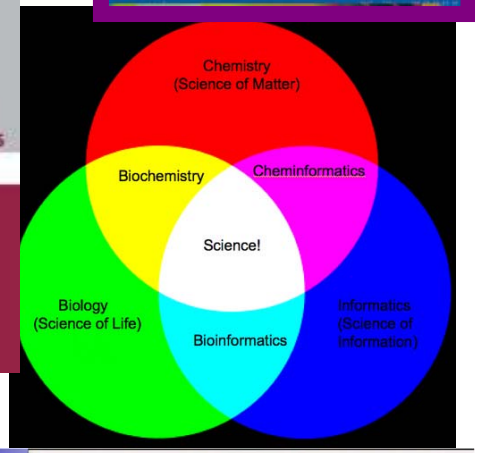
Spans Industry and Science (research)

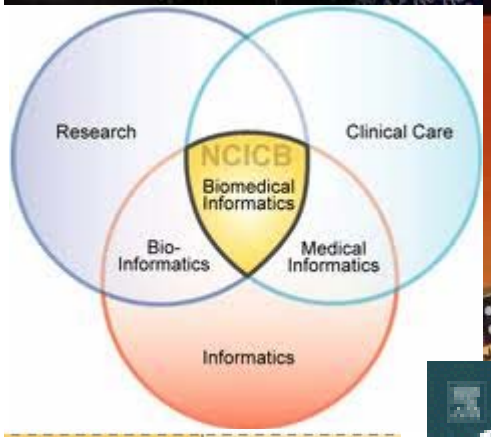
Education: Data Science see recent New York Times articles

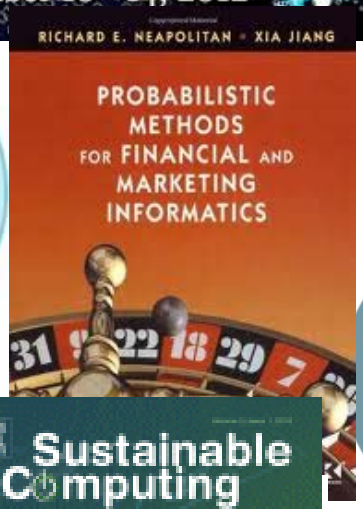
http://datascience101.wordpress.com/2013/04/13/new-york-times-data-science-articles/

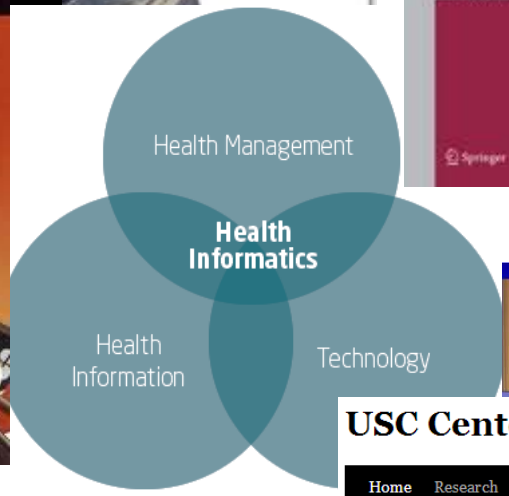


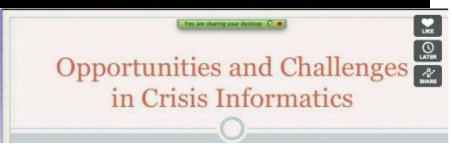


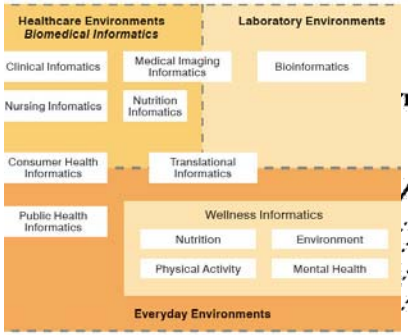


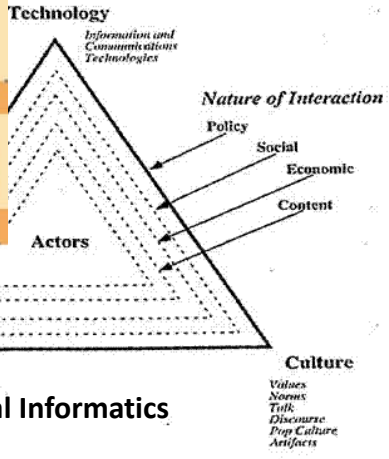




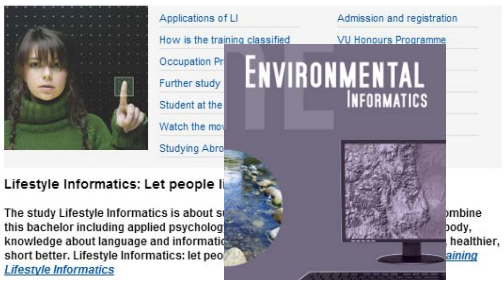








Social Informatics



Big Data and Health

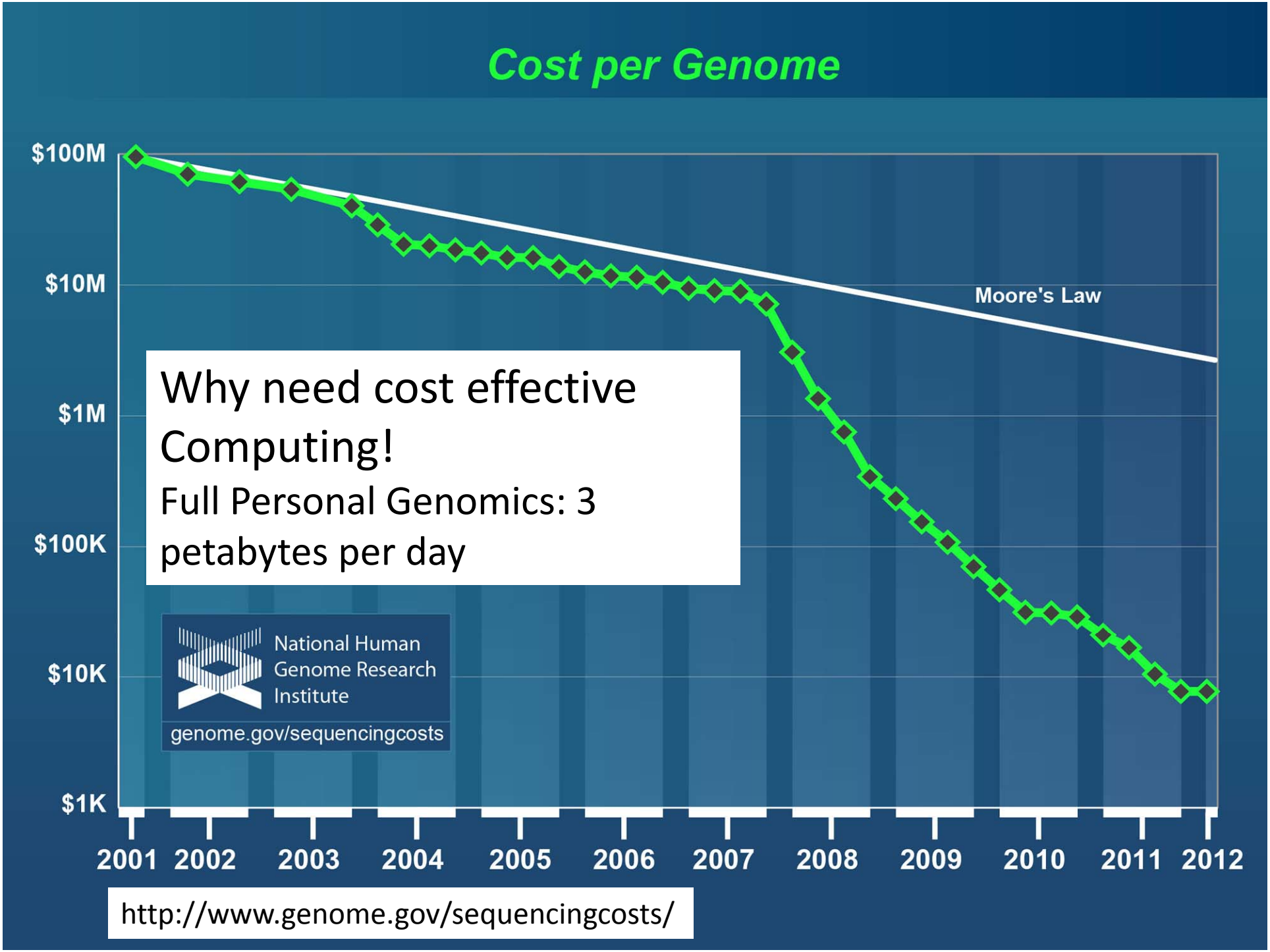
Types of Biomedical Big Data Problems

- Pervasive Health Sensors including data entered into or from smart phones (events)
- Radiology (images)
- Genomics/Proteomics
- Electronic medical records sizewise dominated by omics and images?
  - Updated by events
- Classic data access and sophisticated datamining

Ninety-six percent of radiology practices in the USA are filmless and Table below illustrates the annual volume of data across the types of diagnostic imaging; this does not include cardiology which would take the total to over 10^9 GB (an Exabyte).

http://grids.ucs.indiana.edu/ptliupages/publications/Where%20does%20all%20the%20data%20come%20from%20v7.pdf

Modality	Part B non HMO	All Medicare	All Population	Per 1000 persons	Ave study size (GB)	Total annual data generated in GB
CT	22 million	29 million	87 million	287	0.25	21,750,000
MR	7 million	9 million	26 million	86	0.2	5,200,000
Ultrasound	40 million	53 million	159 million	522	0.1	15,900,000
Interventional	10 million	13 million	40 million	131	0.2	8,000,000
Nuclear Medicine	10 million	14 million	41 million	135	0.1	4,100,000
PET	1 million	1 million	2 million	8	0.1	200,000
Xray, total incl. mammography	84 million	111 million	332 million	1,091	0.04	13,280,000
All Diagnostic Radiology	174 million	229 million	687 million	2,259	0.1	68,700,000 **68.7 PETAbytes**


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
Why need cost effective Computing!
Full Personal Genomics: 3 petabytes per day
National Human Genome Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
http://www.genome.gov/sequencingcosts/

Genomics and Personalized Medicine

Adapting treatments to a person's specific genetic make-up:

- Targeting patients who can benefit (*e.g.* 10% of people cannot respond to codeine), and not develop toxicities (*e.g.* Abacavir for HIV).
- Appropriate dosage of a drug by using genetic variants to understand drug metabolism (*e.g.* anti-depressants, beta blockers, opioid analgesics)
- More drug approvals (re-approvals) because can now target the right sub-group based on genetics.

http://www.ieee-icsc.org/ICSC2010/Tony%20Hey%20-%2020100923.pdf

Larry Smarr: Where There Is Data There Is Hope

Posted on February 12, 2013 by Ernesto Ramirez

"I never thought I would be getting into this business."

This is the first sentence in a mind-expanding talk by Larry Smarr about his self-tracking journey.

In 1999 Larry moved from Illinois to La Jolla, CA to take a position at the University of California, San Diego (UCSD). Like most Southern California transplants he quickly adapted to the local norms and began looking for ways to improve his fitness and health. Continue reading →

Posted in Conference, Discussions, Videos | Tagged conference, Crohn's Disease, Larry Smarr, QS 2012, qstop | 7 Comments

Quantified Self and the Future of Health

Posted on February 16, 2012 by Ernesto Ramirez

On February 7th, 2012 there was an amazing "meeting of the minds" at CALIT2 down in San Diego, CA. The local San Diego Quantified Self meetup group working in collaboration with CALIT2, the Center for Wireless and Population Health Systems, and the West Wireless Health Institute brought together Gary Wolf (Quantified Self founder), Larry Smarr (CALIT2 founding director), Dr. Eric Topol (Scrips Translational Research Institute director and world-renowned digital health evangelist), and Dr. Joseph Smith (Chief Medical Officer of the West Wireless Health Institute) for a great panel discussion. As you'll see and hear below, it was a lively discussion surrounding the topics of Quantified Self, personal health, the future of the medical profession, and patient-provider communication. There was also a great round of questions from the audience (and twitter) and I highly suggest you stick around to hear the very last question!

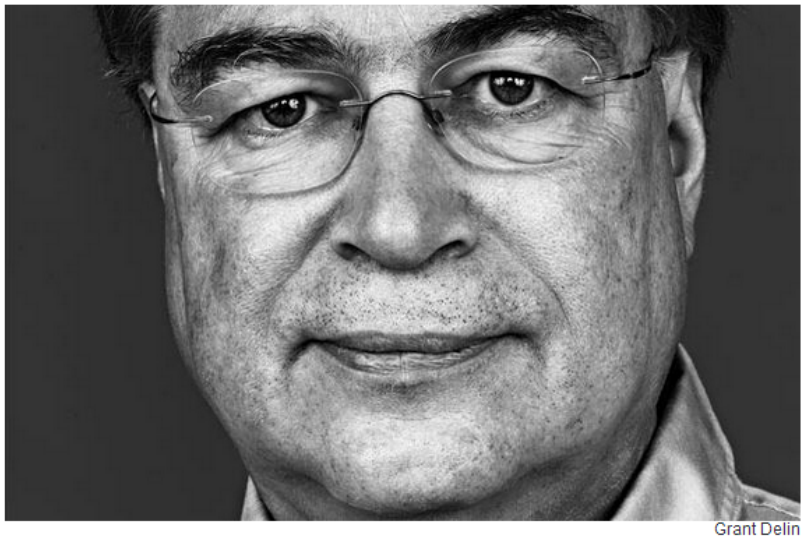
http://quantifiedself.com/larry-smarr/

JULY/AUGUST 2012

The Measured Man

Larry Smarr, an astrophysicist turned computer scientist, has a new project: charting his every bodily function in minute detail. What he's discovering may be the future of health care.

MARK BOWDEN | JUN 13 2012, 10:15 AM ET



Grant Delin

LIKE MANY PEOPLE who are careful about their weight, Larry Smarr once spent two weeks measuring everything he put in his mouth. He charted each serving of food in grams or teaspoons, and broke it down into these categories: protein, carbohydrates, fat, sodium, sugar, and fiber.

Larry used the data to fine-tune his diet. With input nailed down, he turned to output. He started charting the calories he burns, in workouts on an elliptical trainer and in the steps he takes each day. If the number on his pedometer falls short of his prescribed daily 7,000, he will find an excuse to go for a walk. Picture a tall, slender man with the supple, slightly deflated look of someone who has lost a lot of weight, plodding purposefully in soft shoes along the sunny sidewalks of La Jolla, California.

VIDEO

A Brief Visual History of Jordan
90,000 years in 134 seconds

WRITERS

Garance Franke-Ruta
How to Evolve on Gays in Public 8:45 AM ET

Alexis C. Madrigal
The Sounds of the Fastest Plane in the World, an ICBM Missile, and 28 Other Jets, Rockets, and Weapons 8:26 AM ET

Matt Schiavenza
Why You Haven't Heard of Any Chinese Brands ... 7:30 AM ET

Conor Friedersdorf
It's Come to This: Debating Death by Autopilot 7:03 AM ET

Eleanor Barkhorn
Lilly Pulitzer Knew a Secret to Women's Clothing: Dresses Are Practical APR 7, 2013

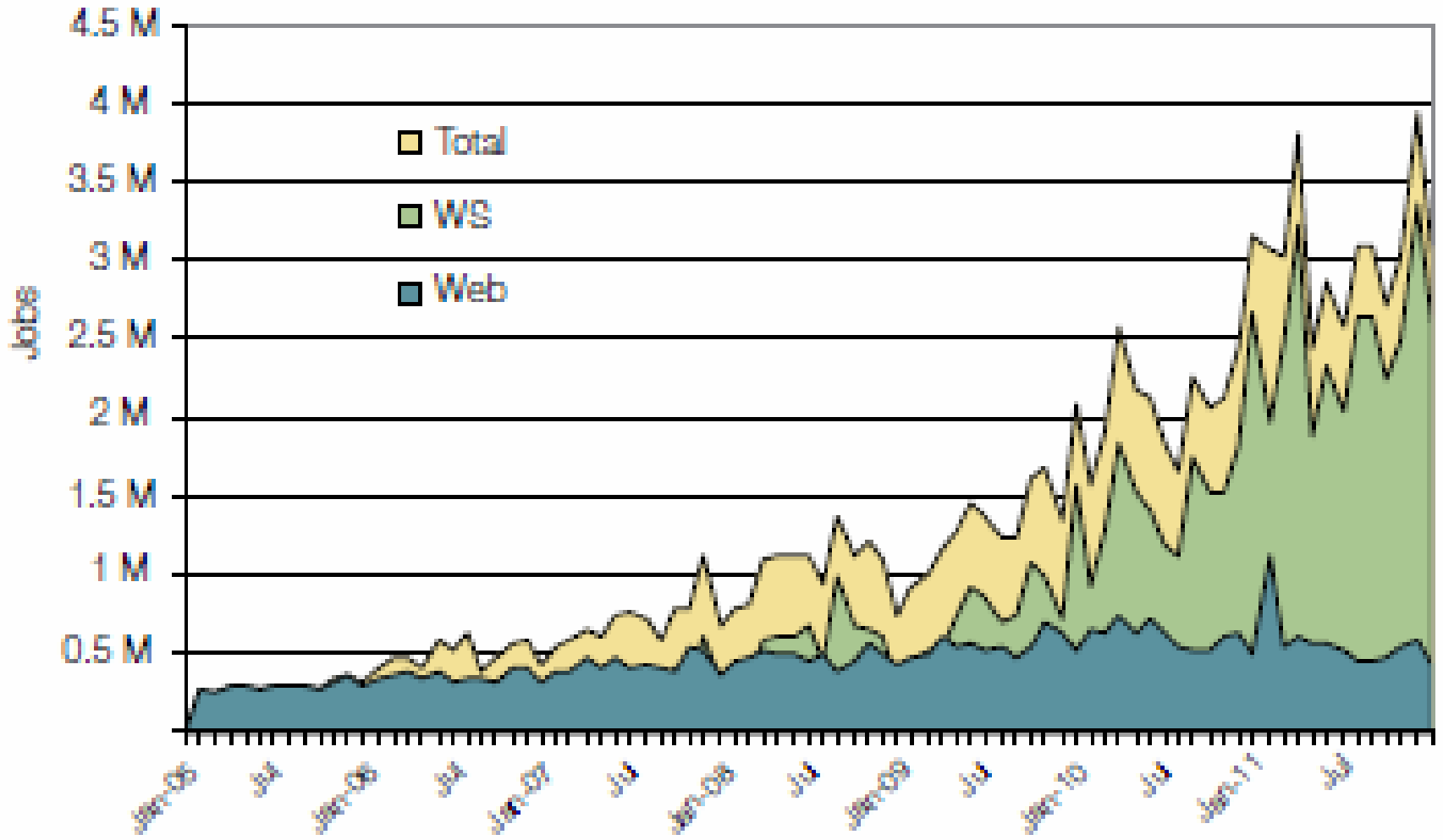
James Fallows
Ten Years Ago: The al-Dura Case APR 7, 2013

Andrew Cohen

Quantified Self

1. New approach to democratic science
2. "Open access" for medical data
3. Flip side of privacy

http://quantifiedself.com/larry-smarr/



2005-20011 Job request at European Bioinformatics Institute EBI for Web hits and automated services WS
http://www.ebi.ac.uk/Information/Brochures/

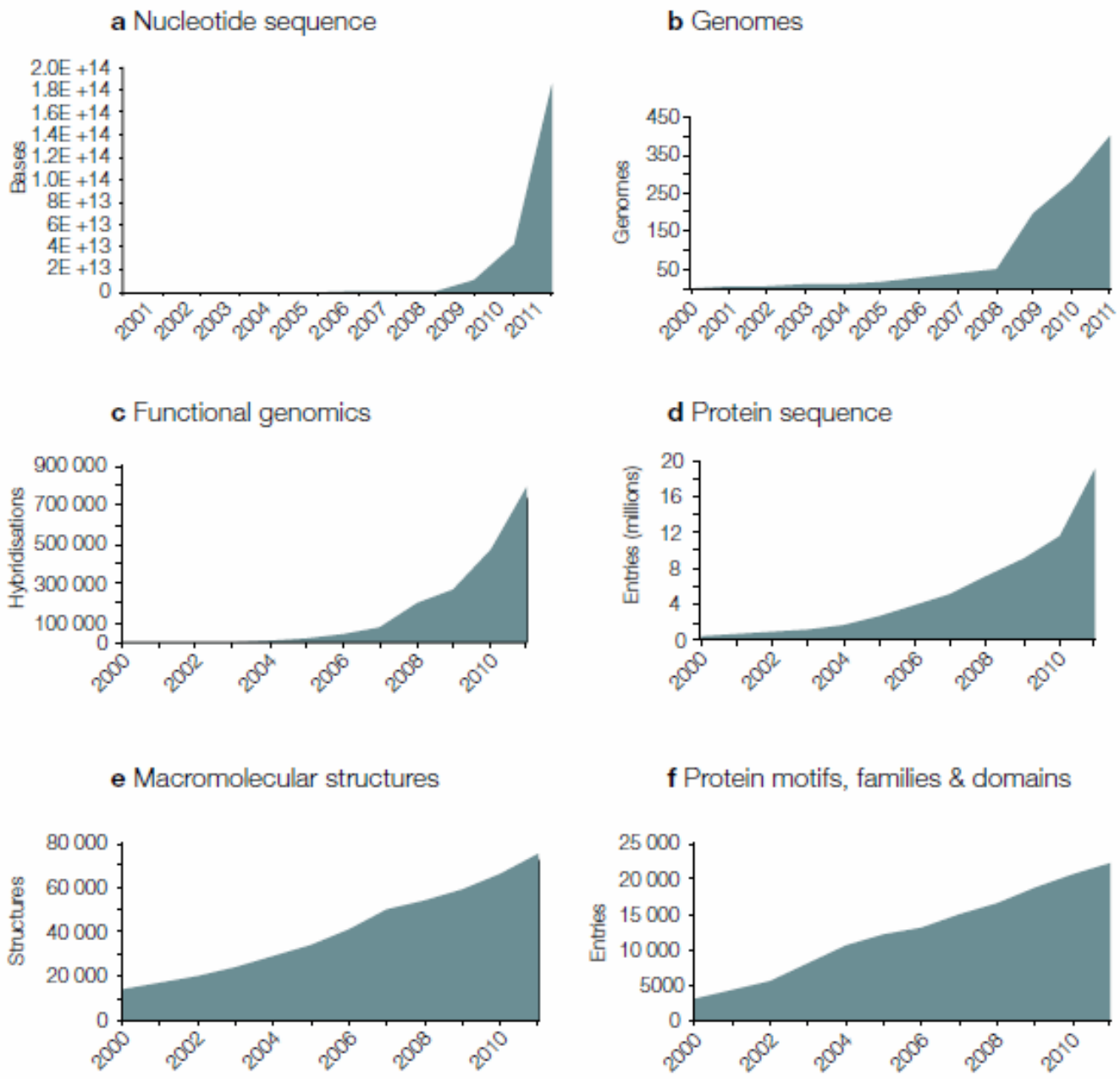


Figure 2. Growth of EMBL-EBI's core data resources from 2000 to 2011. (a) Nucleotide sequence (bases in the European Nucleotide Archive); (b) genomes (entire genomes in Ensembl plus Ensembl Genomes combined); (c) functional genomics (assays in the ArrayExpress Archive,); (d) protein sequence (protein sequences in UniParc); (e) macromolecular structures (structures in PDBe); (f) protein families, motifs and domains (entries in InterPro).

2005-20011 Data stored at European Bioinformatics Institute EBI

http://www.ebi.ac.uk/Information/Brochures/

McKinsey Report on the big-data revolution in US health care

McKinsey Report on The big-data revolution in US health care: Accelerating value and innovation

April 2013

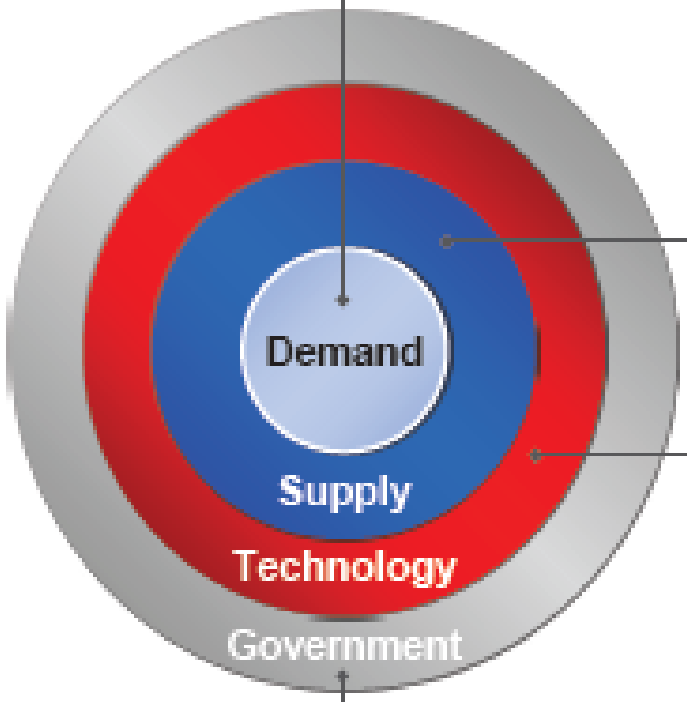
Full

http://www.mckinsey.com/insights/health_systems_and_services/~/media/mckinsey/dotcom/insights/health%20care/the%20big-data%20revolution%20in%20us%20health%20care/the_big_data_revolution_in_healthcare.ashx

Summary

http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights/Health%20care/The%20big-data%20revolution%20in%20US%20health%20care/The%20big-data%20revolution%20in%20US%20health%20care%20Accelerating%20value%20and%20innovation.ashx

Exhibit 1: The convergence of multiple positive changes has created a tipping point for innovation.



Demand for better data, for example:

- Huge cost pressure in the context of reform, economic climate, payment innovation
- First movers showing impact; risk of being "beaten to the punch"

Supply of relevant data at scale, for example:

- Clinical data will become "liquid" thanks to EMRs and information exchanges
- Non-healthcare consumer data are increasingly aggregated and accessible

Technical capability, for example:

- Significant advances in the ability to combine claims and clinical data and protect patient privacy
- Analytical tools now prevalent in front line across all functions

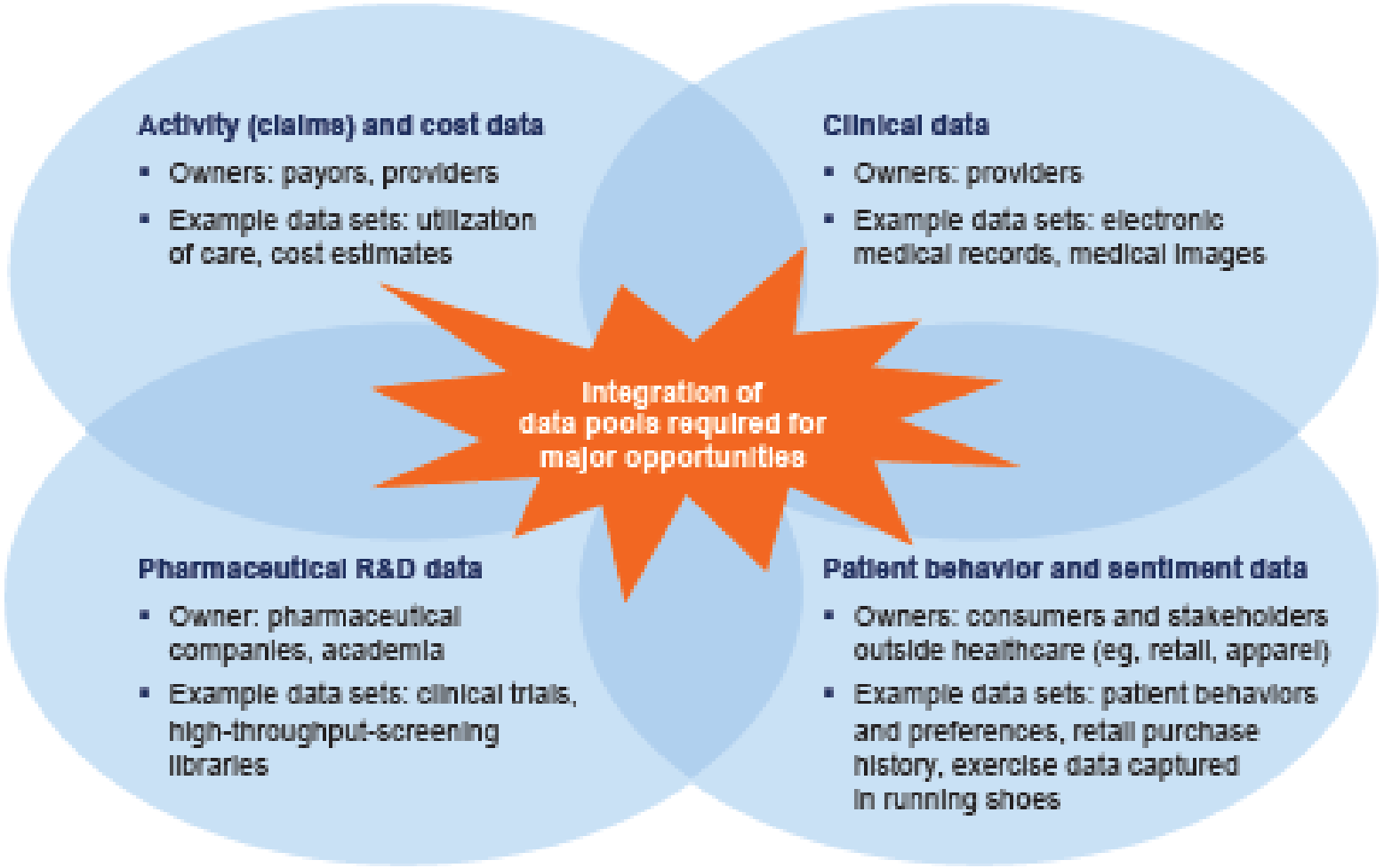
Government catalyzing market change, for example:

- Continued commitment to making data publicly available
- Government is enabling private sector participants to create interoperable standards

Source: McKinsey analysis

McKinsey

Exhibit 2: Primary data pools are at the heart of the big-data revolution in healthcare.

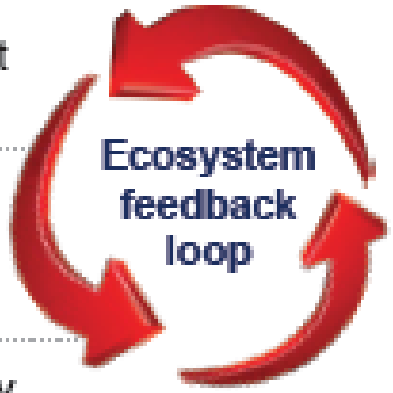


Source: McKinsey Global Institute analysis

McKinsey

Exhibit 3: Big data is changing the paradigm: these are the new value pathways.

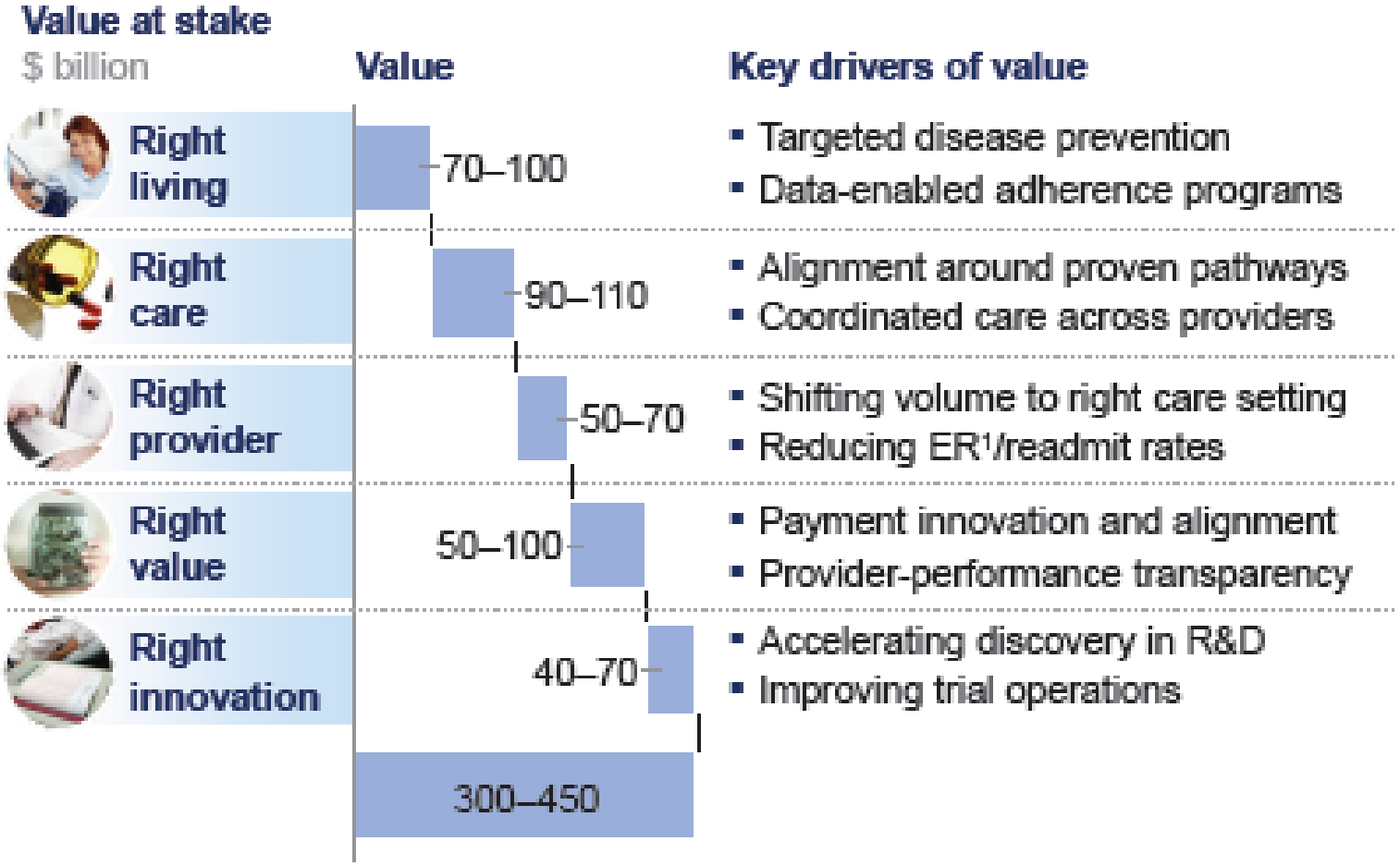
	Description
 **Right living**	Informed lifestyle choices that promote well-being and the active engagement of consumers in their own care
 **Right care**	Evidence-based care that is proven to deliver needed outcomes for each patient while ensuring safety
 **Right provider**	Care provider (eg, nurse, physician) and setting that is most appropriate to deliver prescribed clinical impact
 **Right value**	Sustainable approaches that continuously enhance healthcare value by reducing cost at the same or better quality
 **Right innovation**	Innovation to advance the frontiers of medicine and boost R&D productivity in discovery, development, and safety



Source: McKinsey analysis

McKinsey

Exhibit 4: Applying early successes at scale could reduce US healthcare costs by $300 billion to $450 billion.



1 Emergency room.

Source: American Diabetes Association; American Hospital Association; HealthPartners Research Foundation; McKinsey Global Institute; National Bureau of Economic Research; US Census Bureau

McKinsey

Exhibit 5: Most new big-data applications target consumers and providers across pathways.

Number of innovations observed, by value pathway and target user[1]

	Consumers	Providers	Payors	Manufacturers	Unique applications across the value pathway[1]
Right living	51	32	20	6	68
Right care	41	64	17	10	78
Right provider	10	38	14	5	41
Right value	41	56	19	9	78
Right innovation	6	11	4	11	20
Total for customer type[1]	149	201	74	41	*Totals do not align because of scoring method*[1]

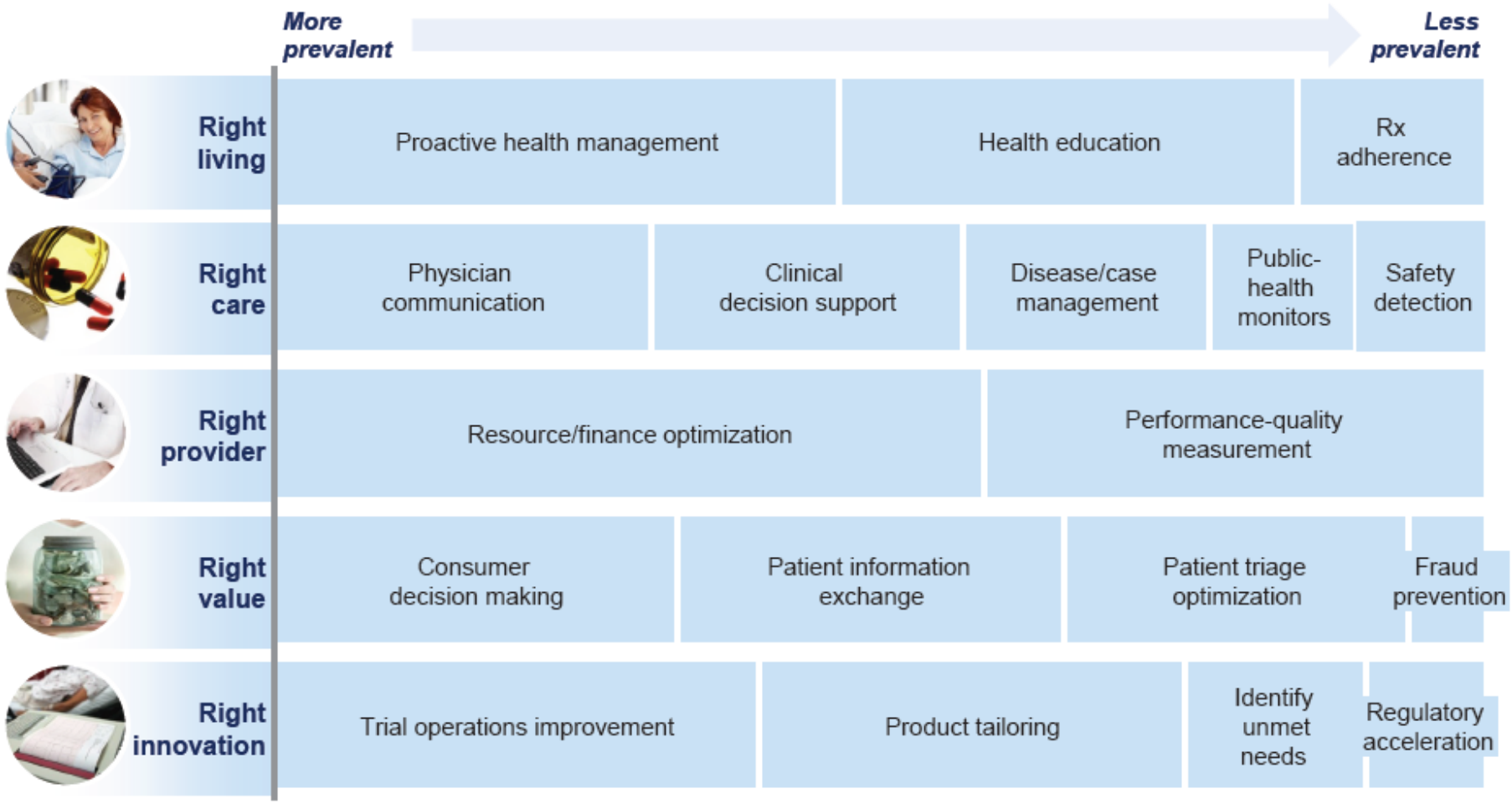
1 Applications fitting in multiple customer categories were counted multiple times; applications were scored for a single, primary value pathway.

Source: N=132, from (1) top 100 submissions to HDI Forum and (2) health-technology companies receiving $2 million+ in venture-capital funding in 2011–12, according to Rock Health/Capital IQ databases; excludes ideas not relevant to big data/analytical application

McKinsey

Exhibit 6: Innovations are weighted toward influencing individual decision-making levers.

Total size of the bar = 100% of ideas in that value pathway; sections are proportional to the % of ideas with specific applications[1]

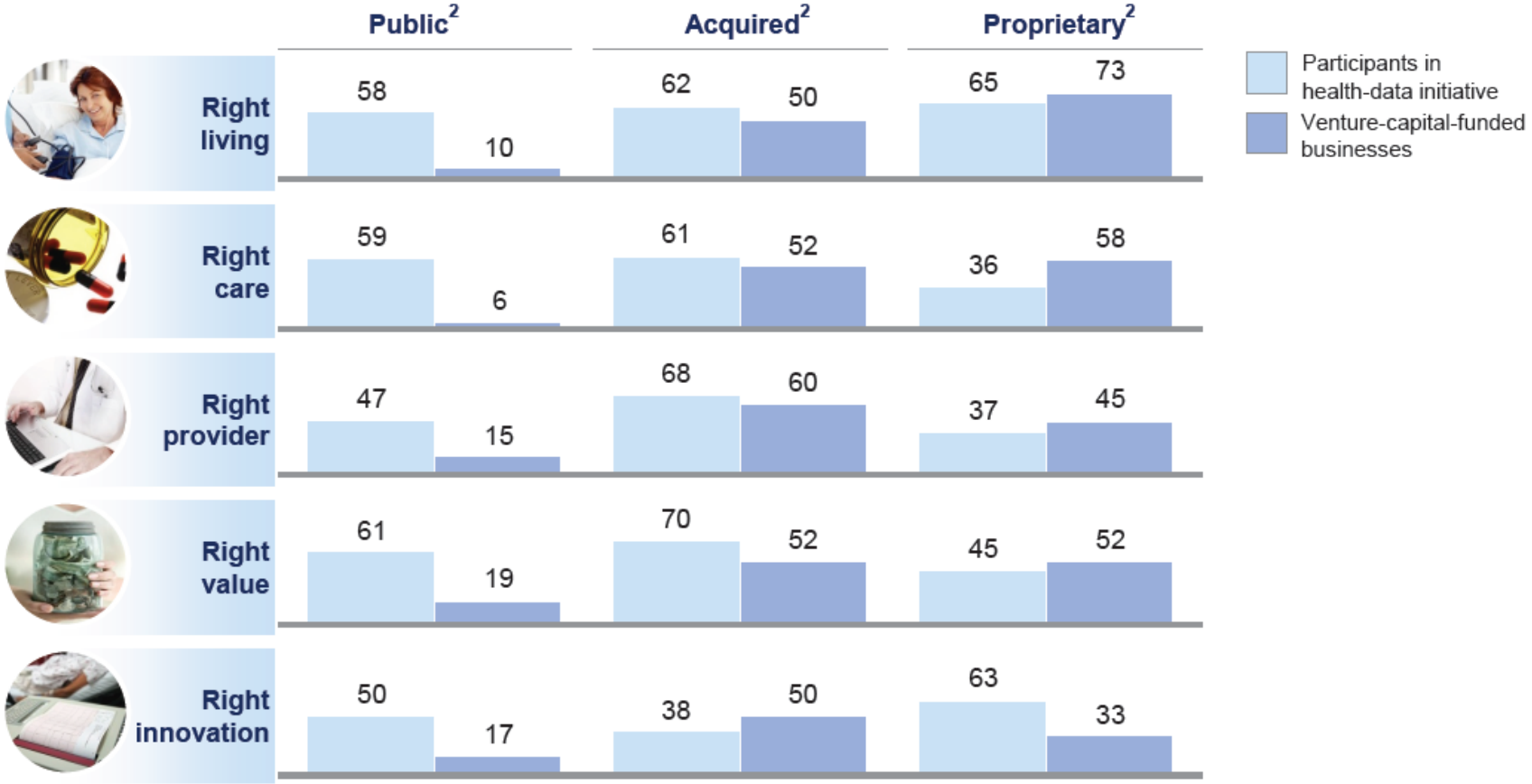


1 Applications fitting in multiple customer categories were counted multiple times.

Source: N=132, from (1) top 100 submissions to HDI Forum and (2) health-technology companies receiving $2 million+ in venture-capital funding in 2011–12, according to Rock Health/Capital IQ databases; excludes ideas not relevant to big-data/analytical application

Exhibit 7: Big-data innovations use a range of public, acquired, and proprietary data types.

Primary data types used:
% of total innovations in each pathway[1]



1 Each idea could use multiple data types.
2 We define data sources as: public: accessible without purchase or partnership required; may be restricted by user or use; acquired: existing data sets purchased or obtained from nonpublic third parties (eg, private payors, electronic health records); proprietary: generated or captured by the company; data documented for the first time by the company or application.

Source: N=132, from (1) top 100 submissions to HDI Forum and (2) health-technology companies receiving $2 million+ in venture-capital funding in 2011–12, according to Rock Health/Capital IQ databases; excludes ideas not relevant to big data/analytical application

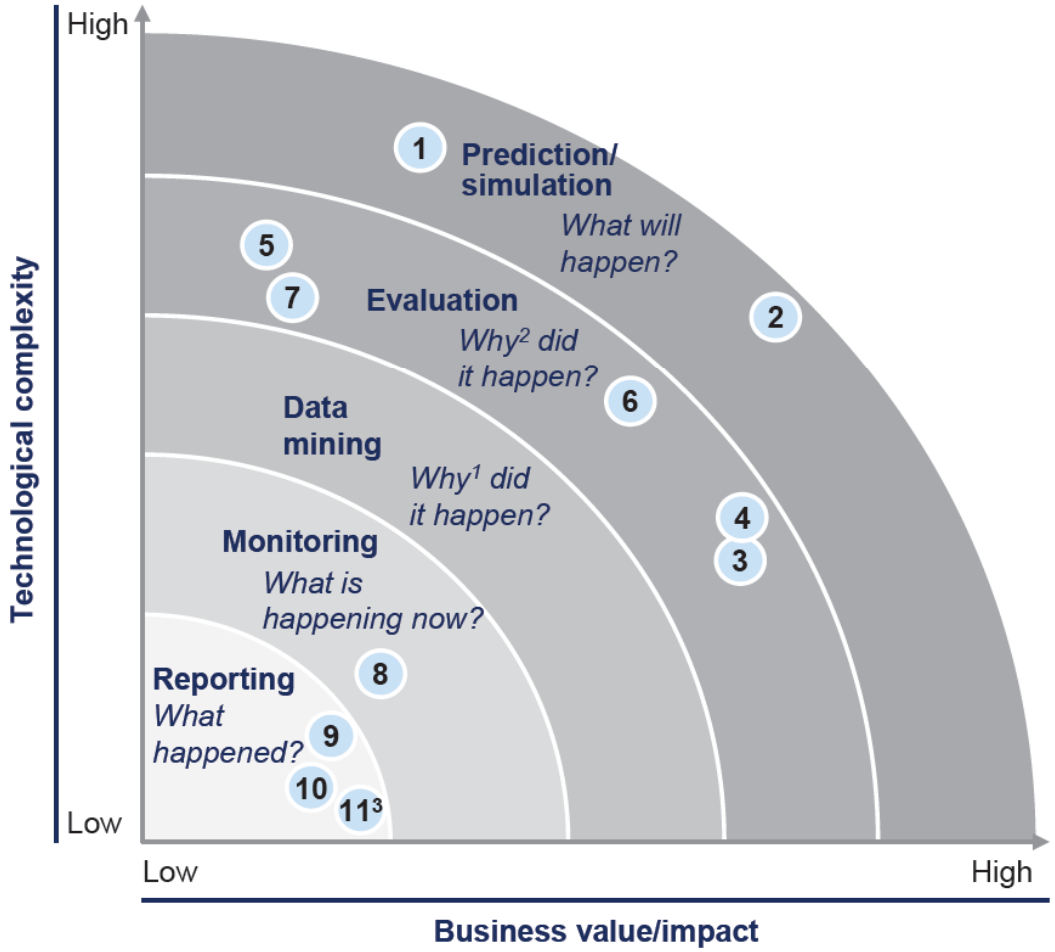
Exhibit 8: Organizations implementing a big-data transformation should provide the leadership required for the associated cultural transformation.

Role for senior leaders

	Performance	Health
Aspire Where do we want to go?	Setting the performance goals	Defining explicit organizational aspirations with the same rigor
Assess How ready are we to go there?	Determining gaps across technical, managerial, and behavioral systems	Understanding the mind-set shifts needed in the organization
Architect What do we need to do to get there?	Developing a portfolio of initiatives to improve performance	Designing the implementation along the levers that drive people to change
Act How do we manage the journey?	Designing the approach to rolling out initiatives in the organization	Building broad ownership, taking a structured approach, and measuring impact
Advance How do we keep moving forward?	Setting up mechanisms to drive continuous improvement	Developing leaders to enable them to drive change

Source: Scott Keller and Colin Price, Beyond Performance: How Great Organizations Build Ultimate Competitive Advantage, Hoboken, NJ: John Wiley & Sons, 2011

Exhibit 9: Companies must develop a range of big-data capabilities.



Examples

1. Risk stratification/patient identification for integrated-care programs
2. Risk-adjusted benchmark/simulation of hospital productivity
3. Identification of patients with negative drug–drug interactions
4. Identification of patients with potential diseases ("patient finder")
5. Evaluation of clinical pathways
6. Evaluation of drug efficacy based on real-world data
7. Performance evaluation of integrated-care programs and contracts
8. Identification of inappropriate medication
9. Systematic reporting of misuse of drugs
10. Systematic identification of obsolete-drug usage
11. Personal health records

1 Machine based: evaluation of data correlations only.
2 Hypothesis based: integration of advanced analytics to determine causation, interdependencies.
3 Higher business value expected if further enhanced and rolled out as personal health record.

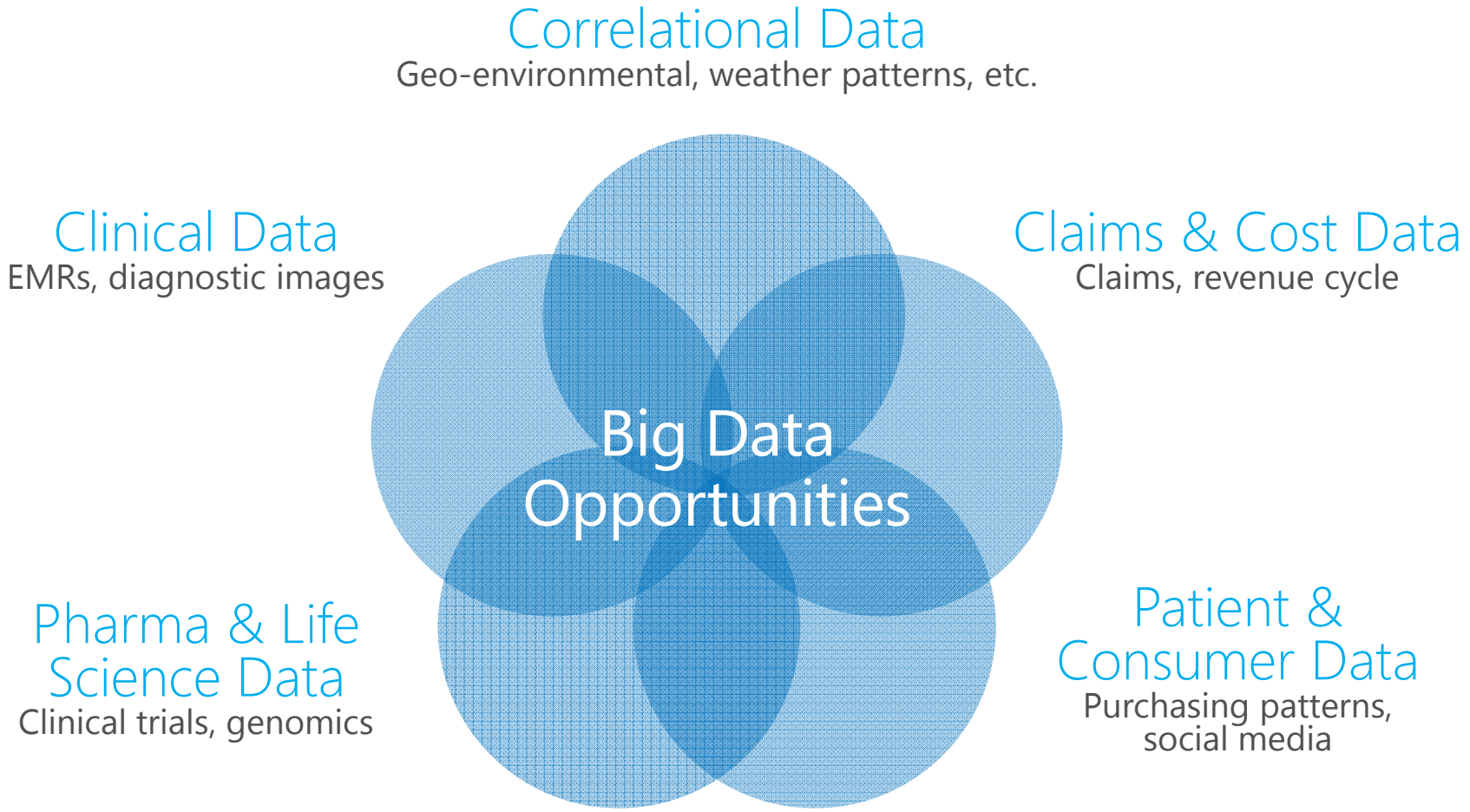
McKinsey

Microsoft Report on Big Data in Health

https://partner.microsoft.com/download/global/40193764

Big Data opportunities in health

The promise of Big Data to transform health and social services comes from new capabilities to increases "Data Convergence" opportunities.

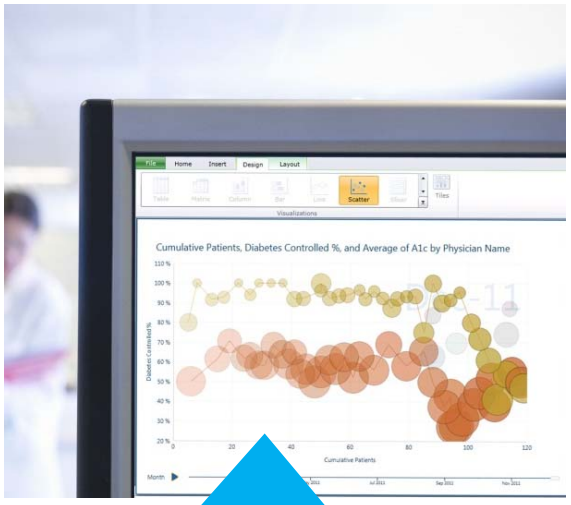


New possibilities with Big Data

Live data feed

Based on a "snapshot" of large and different datasets right now, what can I predict will happen going forward?

For example, how do I optimize my resources to see and plan for the emerging outbreak of infectious diseases?



Social analytics

What are the top health topics and issues on the minds of consumers now?

What can I predict based on social media topics?

Advanced analytics

How might I determine the best care pathways for treating a population of patients with the same medical condition?

What actions might I take to decrease the likelihood of a medical condition?

https://partner.microsoft.com/download/global/40193764

Table of Contents

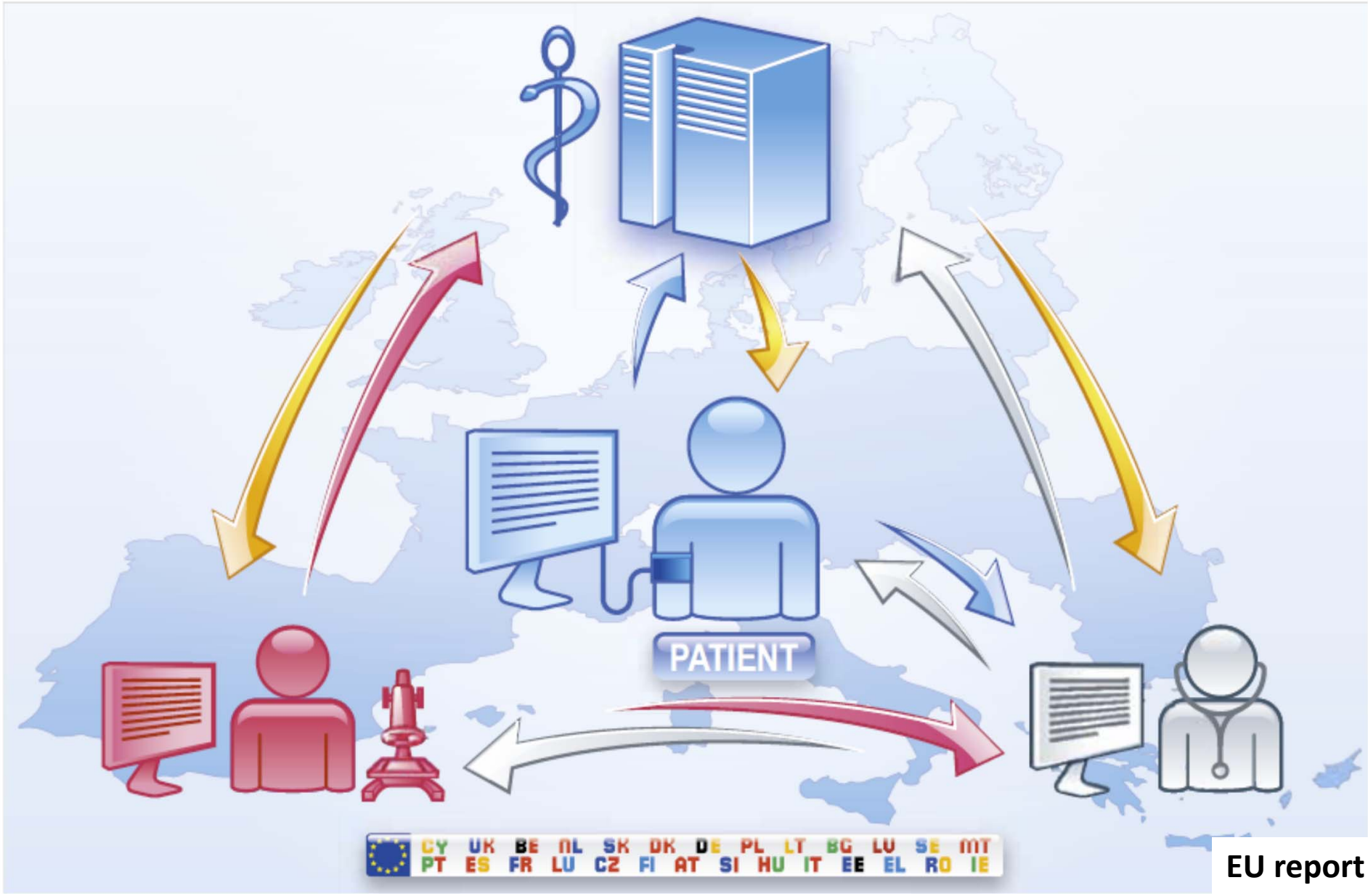
EU Report on Redesigning health in Europe for 2020

http://ec.europa.eu/information_society/activities/health/docs/policy/taskforce/redesigning_health-eu-for2020-ehtf-report2012.pdf

Use the power of data

- Data often sits in silos in primary, secondary and tertiary health institutions. This silo mentality mirrors the way that health professionals guard their own competence and areas of expertise. In the new era of eHealth, this has to end.
- Multidisciplinary teams of different actors, not all of whom are healthcare professionals, are part of future picture of health. Currently there is a sharp divide between 'official' medical data and the wealth of other health information generated by users that is not used for care. We need to find a way of making this data more trustworthy.
- The key question is what people do with this information and how they can use it. New rules are needed to define how to integrate official data and user data to create a more holistic picture of patient situation for health care as well provide early feedback for preventive care. Certification of applications is one way forward but it should be based on a set of principles for how health related data should be treated rather than regulation.
- Health institutions must publish the data on their performance and health outcomes. This information should be regularly collected, comparable and publicly available. This will support a drive to the top as high performing organisations and individuals can be identified and used as an example to inspire change. In health, performance is not just how efficiently the system operates but also the patient experience of the care. Publication of such data in other sectors has led to strong public demand for better performance and a greater focus on accountability and results.
- http://ec.europa.eu/information_society/activities/health/docs/policy/taskforce/redesigning_health-eu-for2020-ehtf-report2012.pdf

Lever for change #1:
My data, my decisions



EU report

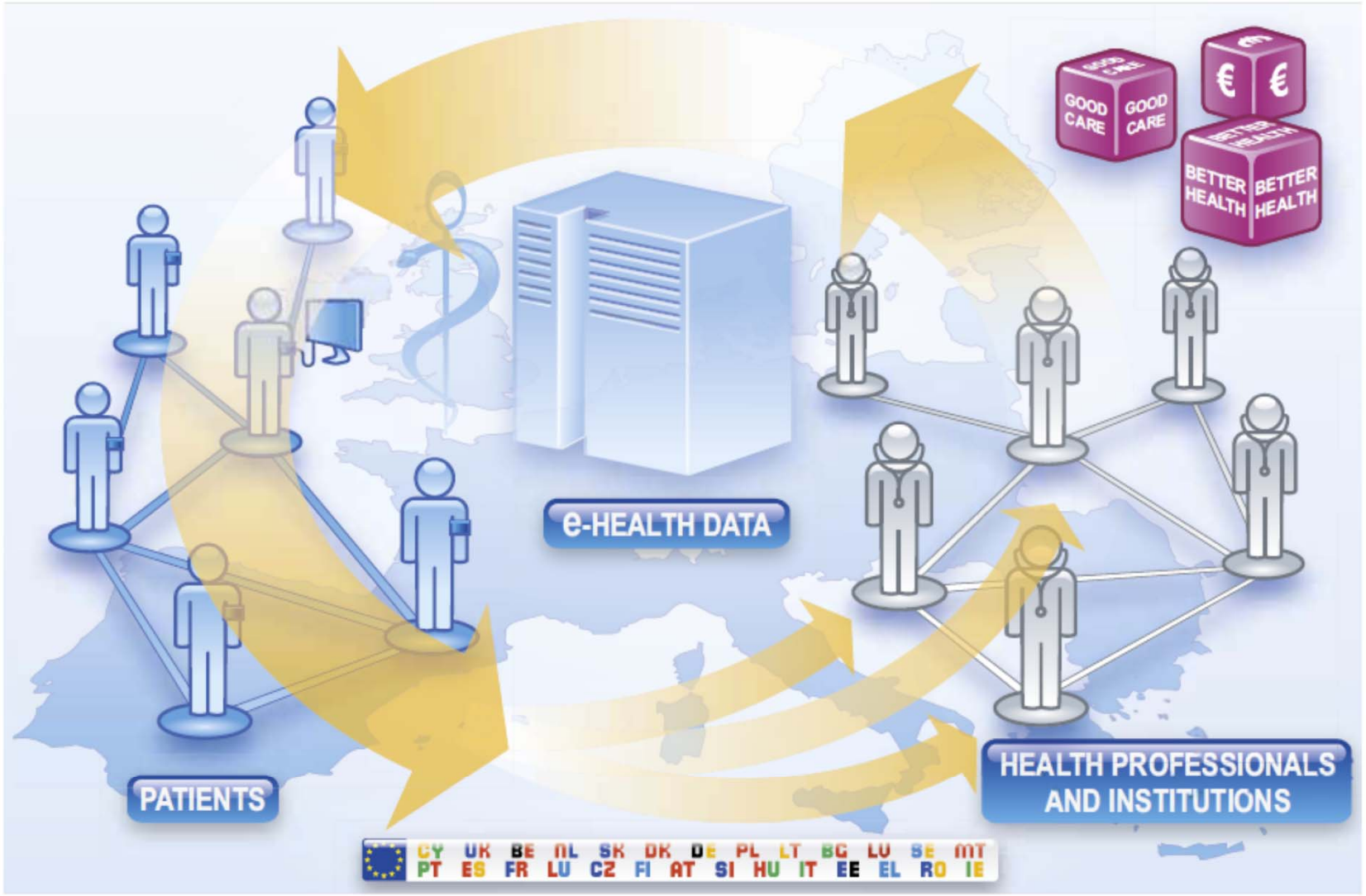
Lever for change #2: **Liberate the data**

Data can be compared to oil: In the ground it is unusable and worthless. Extracted and refined, it has huge value. Large amounts of data currently sit in different silos within health and social care systems. If this data is released in an appropriate manner and used effectively it could transform the way that care is provided.

Lever for change #3:
Connect up everything

EU report

Lever for change #4:
Revolutionise health

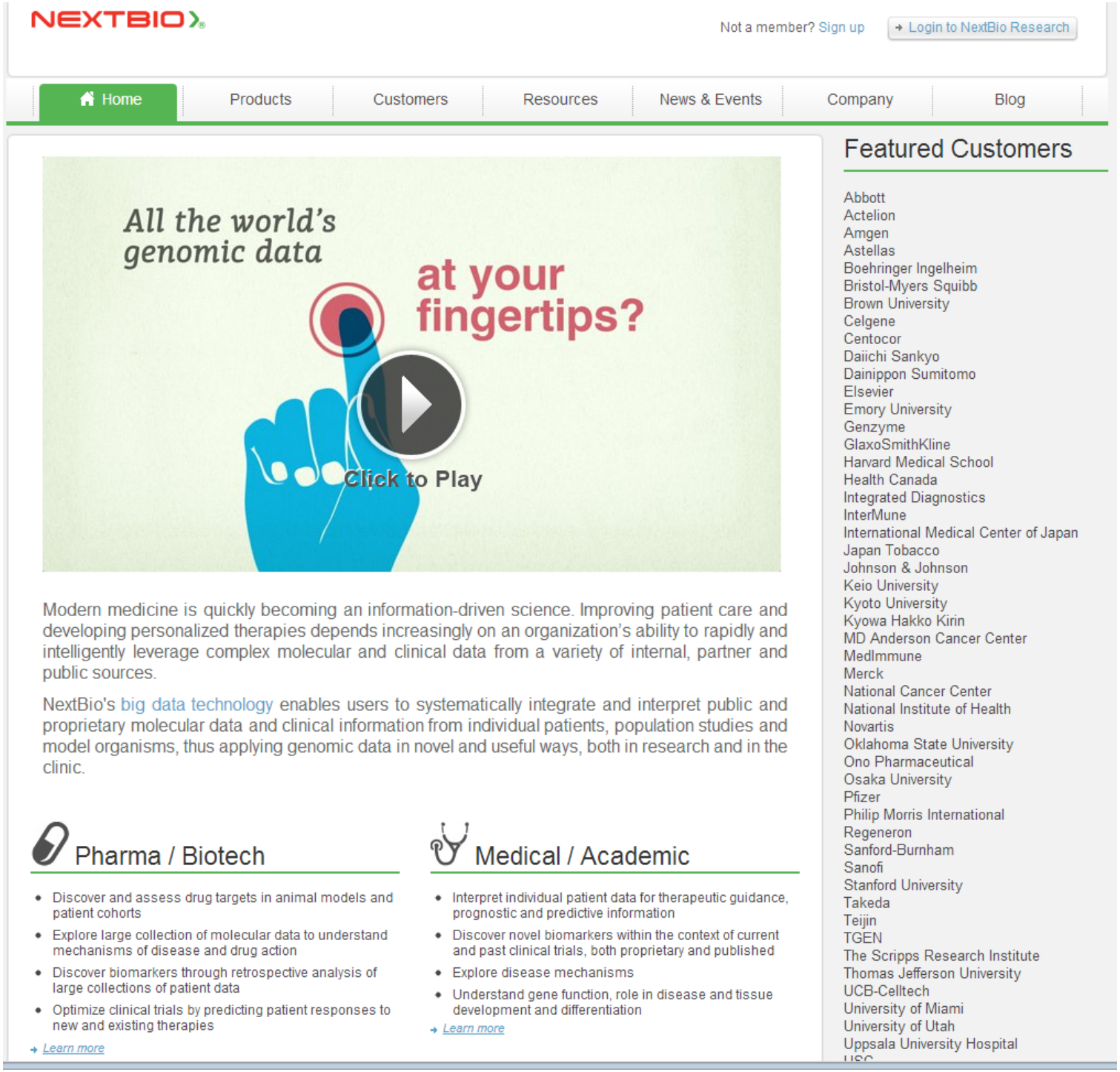


Full transparency will unleash disruptive innovation across the health sector. Armed with data about the performance of health professionals and institutions – and how these differ from one another – patients will be able to make more informed choices about where and how they want to be treated. This will have real impact on resource allocation in health, as funding follows the patients. This bottom-up process contributes towards an enabling environment for eHealth and will be the momentum driving the pace of change.

EU report

People in unequal societies have poorer health, as the WHO Commission on the Social Determinants of Health revealed. Within and between EU countries there are entrenched health inequalities that result in differences in life expectancies of up to 15 years between the wealthy and the poor. Those without the skills, capacity and opportunity to use eHealth risk being further excluded. New ICT tools have the potential to reduce these inequalities but they need to be designed to actively promote and enhance equity. This means ensuring that rural communities have access to services and that products are usable for patients with a diverse range of literacy and technical abilities.

Clouds and Health



Medical Cloud Start up

Healthcare & Cloud Computing

- Patient's information would be stored in a cloud
- Accessed and managed over the Internet
- Since we are on a paperless route, this is a great idea to store information
- Authorized users
- Information on one cloud is connected to bigger clouds
  - Ex. Big Bend RHIO connected to the NHIN

http://healthinformatics.wikispaces.com/file/view/cloud_computing.ppt

Considerations With Cloud Computing in Healthcare

- Since information is stored over the Internet, precautions must be taken

- Cloud system must conform to the HIPAA act
  - Personal Health Information
  - Secure transmission of PHI over the Internet
  - Need to maintain a secure, safe, and authorized environment for the prevention of information leakage

http://healthinformatics.wikispaces.com/file/view/cloud_computing.ppt

Advantages of Cloud Computing

- Low costs
  - Outsourcing information reduces amount spent on new technology
  - Easier to maintain
- More secure
  - Companies are hired to watch over the information
- Interoperability
  - Access information from anywhere
  - Can be accessed using different devices

http://healthinformatics.wikispaces.com/file/view/cloud_computing.ppt

Advantages of Cloud Computing

- Increases the adoption of EMRs
- Beneficial for small companies
- Easy to share information among different organizations and doctors

http://healthinformatics.wikispaces.com/file/view/cloud_computing.ppt

Cloud Computing Disadvantages

- Security is the main disadvantage of cloud computing
- Consumers are worried about Insurance companies getting a hold of there information and discriminating based upon current medical conditions they may have or medical conditions that they could develop later in life.
- They are also worried about government agencies getting a hold of there information and exploiting it to third party vendors, or there place of work.

http://healthinformatics.wikispaces.com/file/view/cloud_computing.ppt

Disadvantages Cont.

- The cloud companies do not always handle all of the security themselves and sometimes pass it off to third party vendors

- Consumers need to make sure to thoroughly check out these companies to see who else they are involved with and check out there reputation to see if you trust them to not share your information with any other outside sources. If the third party vendor looks trustworthy then you are probably safe to send your medical records over the cloud safely.

http://healthinformatics.wikispaces.com/file/view/cloud_computing.ppt

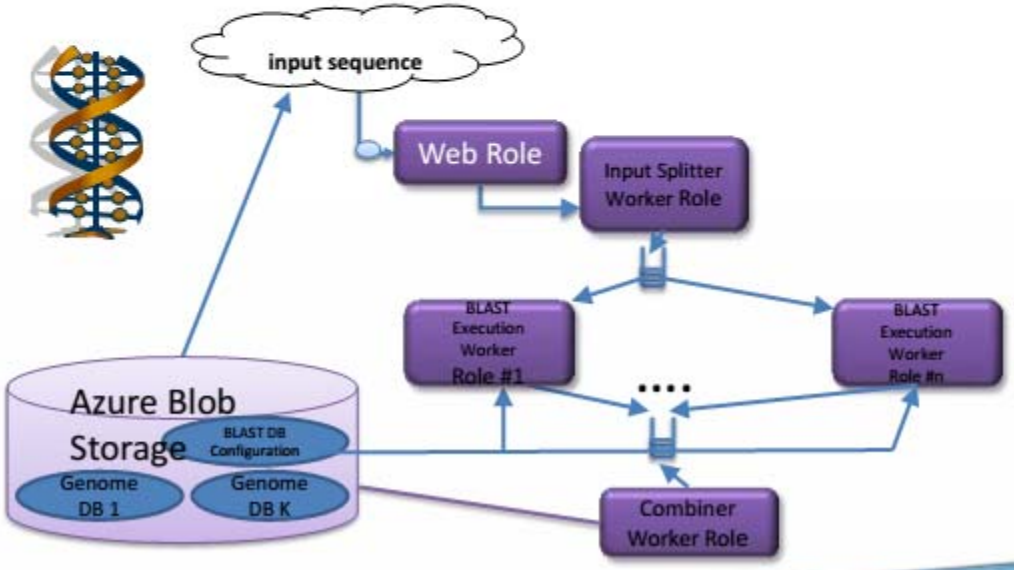
Genomics, Proteomics and Information Visualization

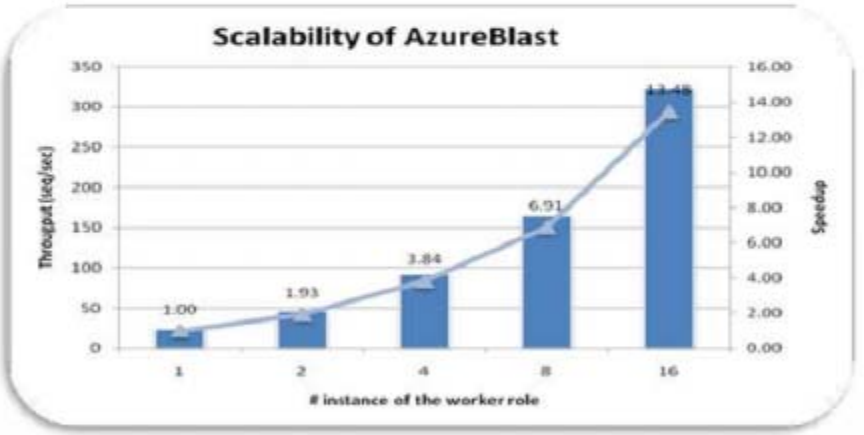
So far genome not a major part of Big Open Data in medicine

NCBI Blast for Azure

Seamless Experience

- Evaluate data and invoke computational models from Excel.
- Computationally heavy analysis done close to large database of curated data.
- Scalable for large, surge computationally heavy analysis.
- Test local, run on the cloud.



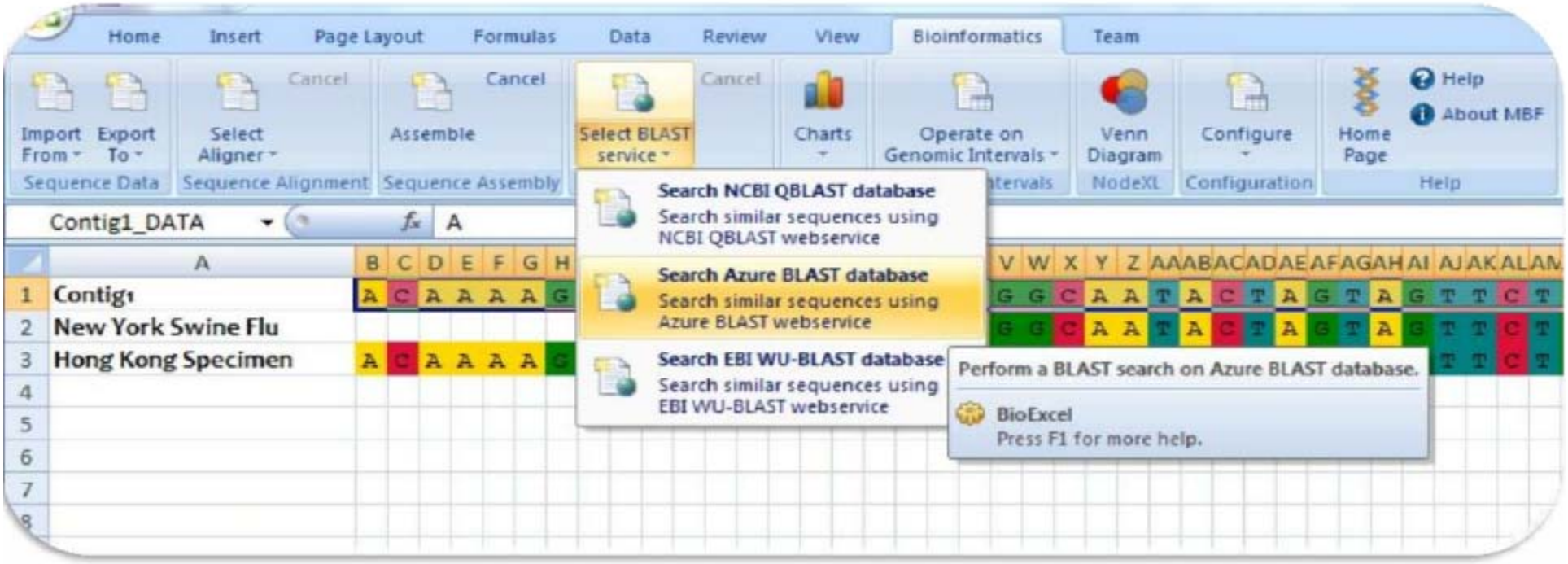


Now available for public release
http://research.microsoft.com/azure

Microsoft

http://www.oerc.ox.ac.uk/downloads/presentations-from-events/microsoftworkshop/gannon

Making Excel the user interface to the cloud



Microsoft

http://www.oerc.ox.ac.uk/downloads/presentations-from-events/microsoftworkshop/gannon

Genomics in Personal Health

- Suppose you measured everybody's genome every 2 years
  - 30 petabits of new gene data per day
  - factor of 100 more for raw reads with coverage
- Data surely distributed
- $1.5*10^8$ to $1.5*10^{10}$ continuously running present day cores to perform a simple Blast analysis on this data
  - Amount depends on clever hashing and maybe Blast not good enough as field gets more sophisticated
- Analysis requirements not well articulated in many fields – See http://www.delsall.org for life sciences
  - LHC data analysis well understood – is it typical?
  - LHC Pleasing parallel (PP) – some in Life Sciences like Blast also PP

DNA Sequencing Pipeline

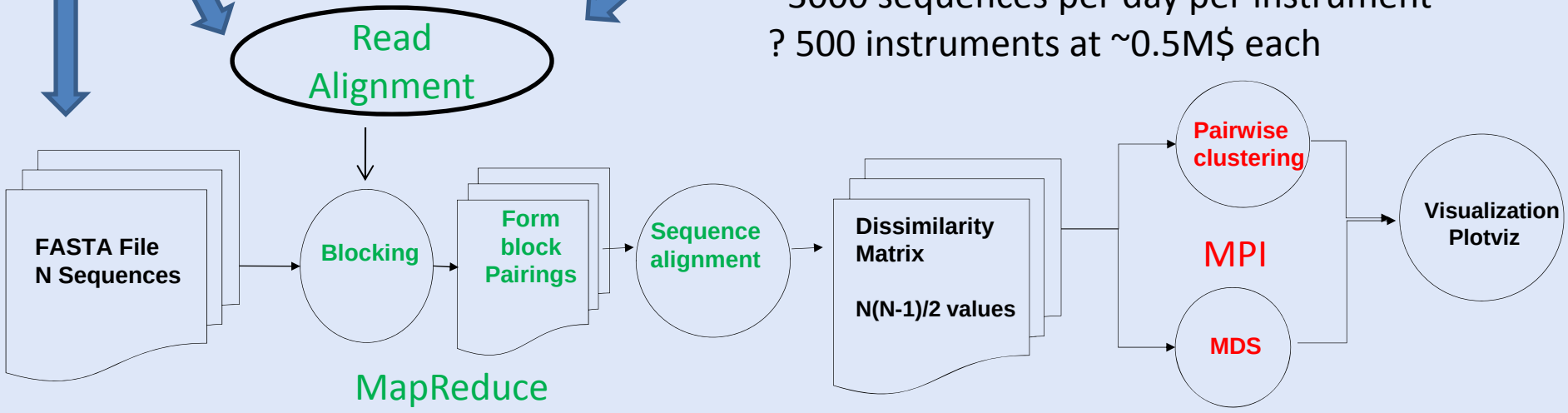
Illumina/Solexa

Roche/454 Life Sciences

Applied Biosystems/SOLiD

Internet

~300 million base pairs per day leading to
~3000 sequences per day per instrument
? 500 instruments at ~0.5M$ each



Protein Sequence Universe

- PSU Goal: Enhance annotation resources with analytic and visualization (browser) tools.
- One component of PSU is to project sequence data into 3D using multidimensional scaling (MDS).
- MDS **interpolation** allows expanding the universe without time consuming all vs all $O(N^2)$
  - 3D map allows much faster interpolation
- Use set of pairwise dissimilarities – don't do MSA – so don't have vectors in some space

High Performance Dimension Reduction and Visualization

- Need is pervasive
  - Large and high dimensional data are everywhere: biology, physics, Internet, …
  - **Visualization can help data analysis**
- Visualization of large datasets with high performance
  - Map high-dimensional data into low dimensions (2D or 3D).
  - Need Parallel programming for processing large data sets
  - Developing high performance dimension reduction algorithms:
    - MDS(Multi-dimensional Scaling)
    - GTM(Generative Topographic Mapping)
    - DA-MDS(Deterministic Annealing MDS)
    - DA-GTM(Deterministic Annealing GTM)
  - Interactive visualization tool PlotViz

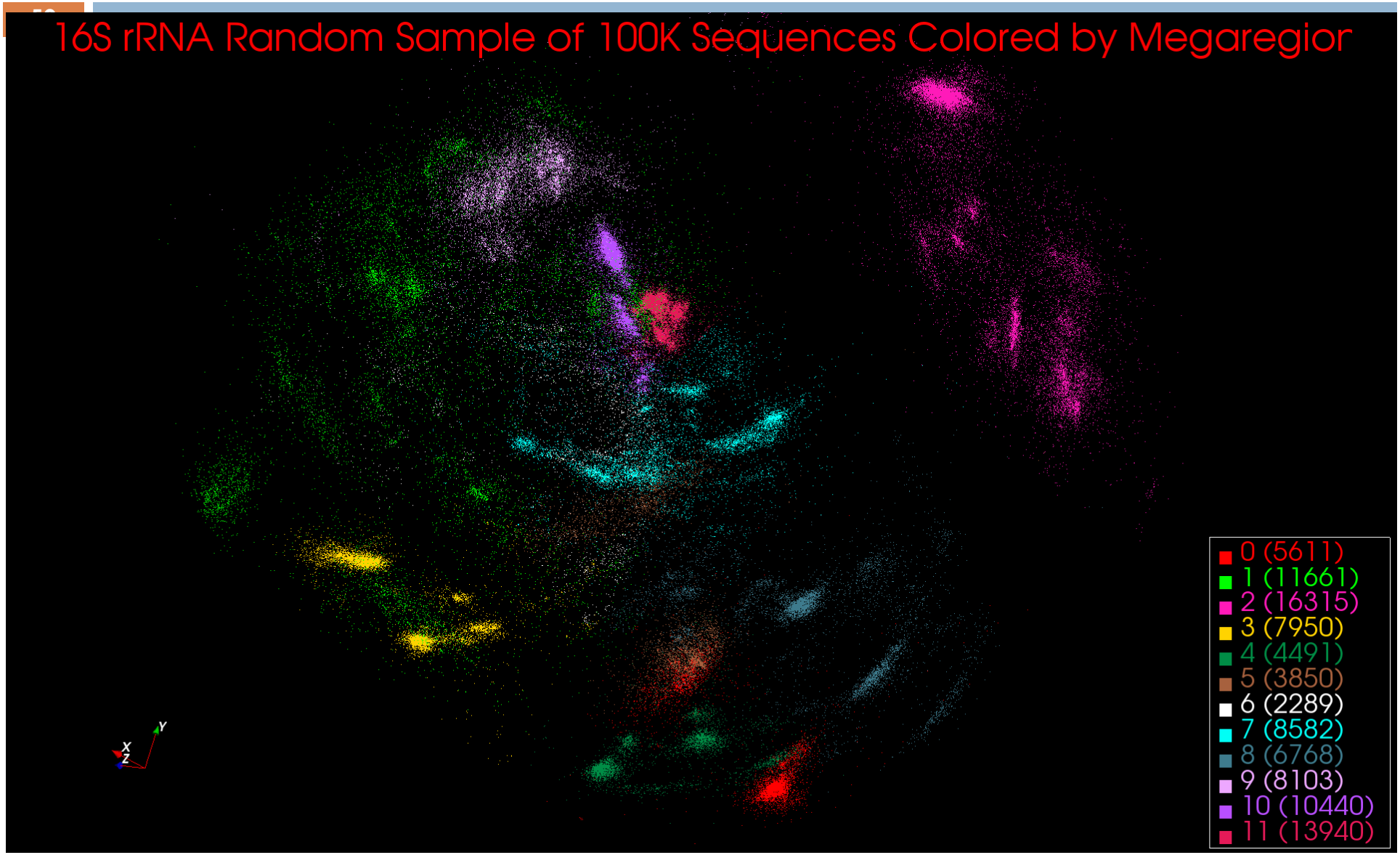
Multi-Dimensional Scaling (MDS)

- Sammon's objective function

$$H = \sum_{i<j}^{n} \frac{\left(f(\delta_{ij}) - d(x_i, x_j)\right)^2}{f(\delta_{ij})}$$

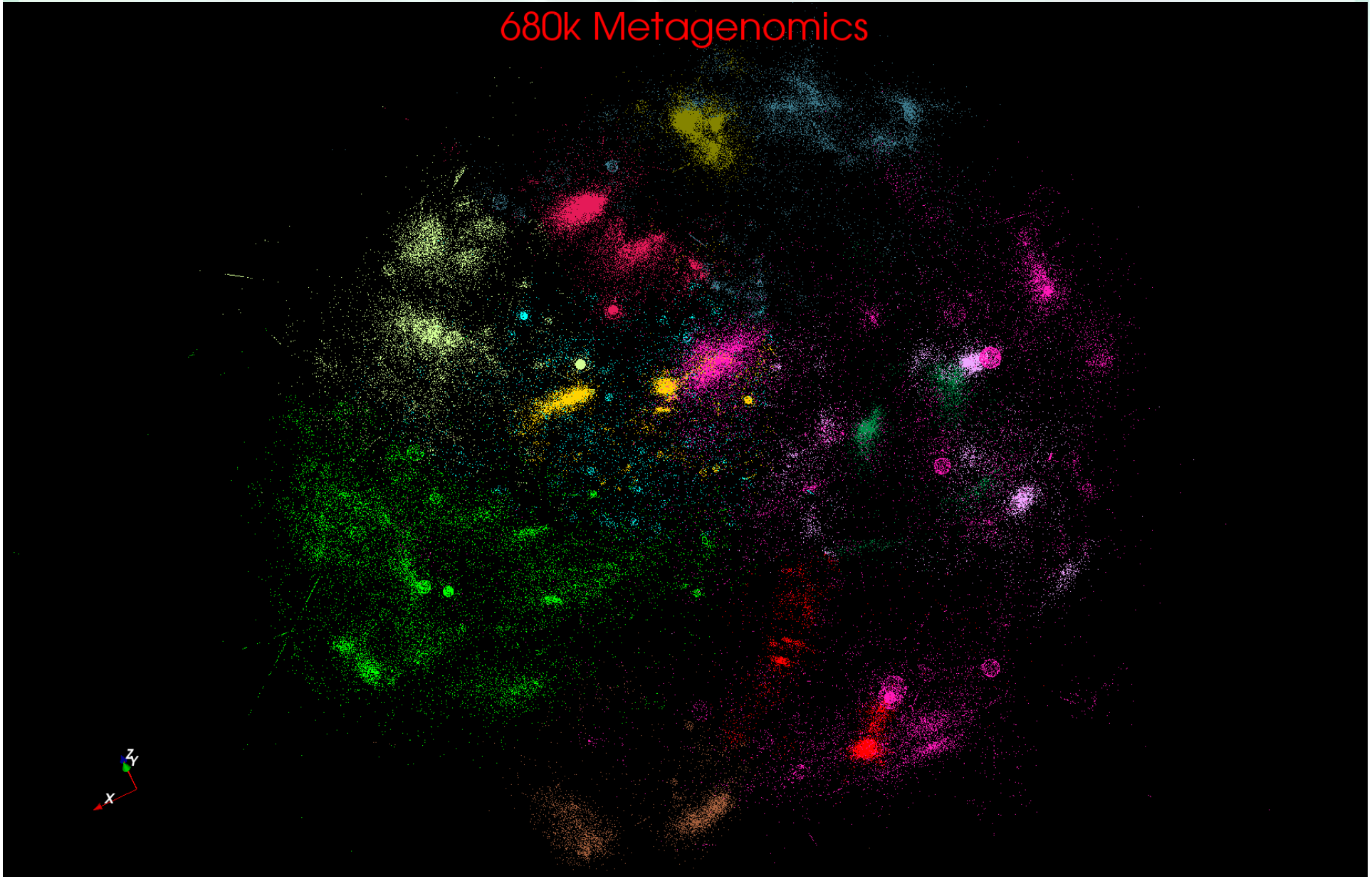
- δ_{ij} is dissimilarity **measure** between sequences ***i*** and ***j***
- ***d*** is Euclidean distance (here in 3D for visualization) between projections $\mathbf{x}_i$ and $\mathbf{x}_i$
- Denominator chosen to get larger contribution in objective function from smaller dissimilarities
- ***f*** is monotone transformation of dissimilarity measure chosen "artistically"

Typical Metagenomics MDS



Metagenomics

http://salsahpc.indiana.edu/millionseq/mina/16SrRNA_index.html



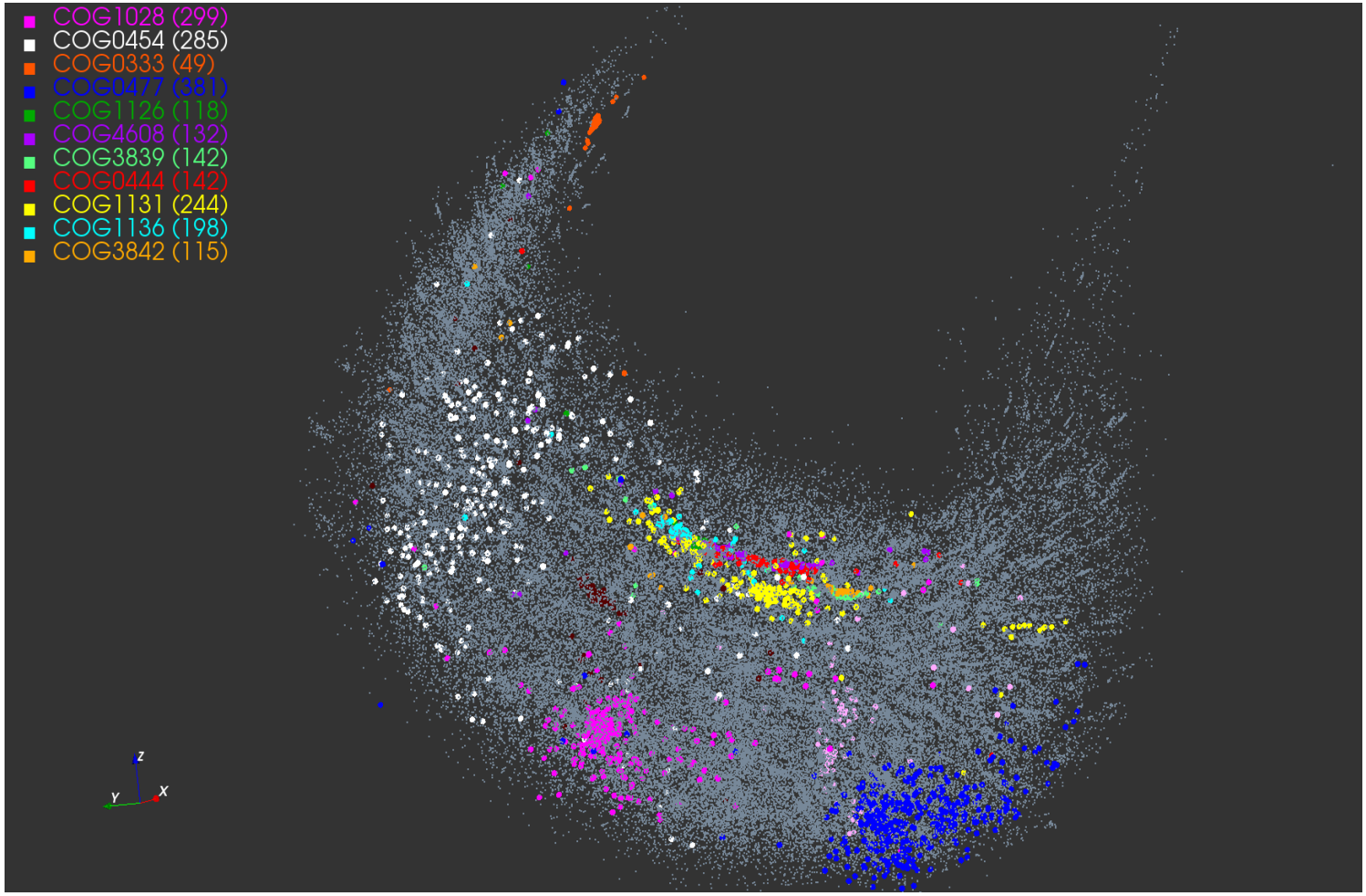
MDS Details

- ***f*** chosen heuristically to increase the ratio of standard deviation to mean for $f(\delta_{ij})$ and to increase the range of dissimilarity measures.
- ***O(n²)*** complexity to map ***n*** sequences into 3D.
- MDS can be solved using EM (SMACOF – fastest but limited) or directly by Newton's method (it's just χ^2)
- Used robust implementation of nonlinear χ^2 minimization with Levenberg-Marquardt
- 3D projections visualized in PlotViz

MDS Details

- Input Data: 100K sequences from well-characterized prokaryotic COGs.
- Proximity measure: sequence alignment % scores
- Scores calculated using Needleman-Wunsch
- Scores "sqrt 4D" transformed and fed into MDS
  - Analytic form for transformation to 4D
  - δ_{ij}^{n} *decreases dimension* $n > 1$*; increases* $n < 1$
- "sqrt 4D" reduced dimension of distance data from 244 for δ_{ij} to14 for $f(\delta_{ij})$
- Hence more uniform coverage of Euclidean space

3D View of 100K COG Sequences



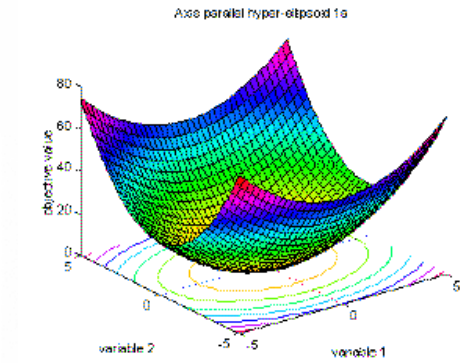
Technology Comment

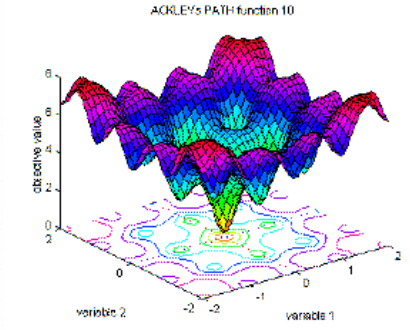
- MDS is minimizing the sum of squares, This is classic optimization problem called χ^2 (chisq) or Least Squares
  $\chi^2 = \Sigma_{i=1}^{M}(\text{Data-value}(i) - \text{Prediction}(i))^2/\text{Error(i)}^2$
- For MDS, M = N(N-1)/2 and
  Data-value is original value of distance and Prediction is distance after 3D projection
- We need to minimize χ^2 as function of 3D positions.
- The good news is that χ^2 has a well defined minimum as it is positive – sum of squares
- Bad news – very hard to minimize
- Good News – Problem is naturally parallel over points i and initial starting points

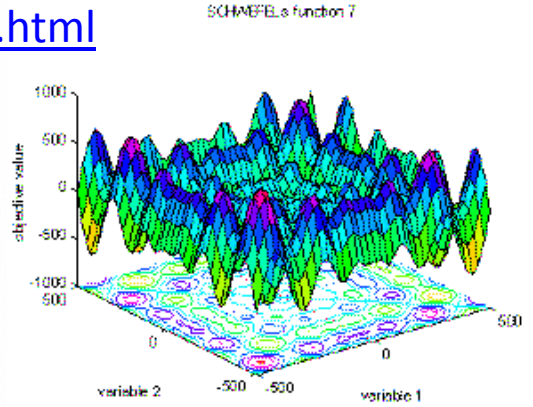
Solving Least Squares

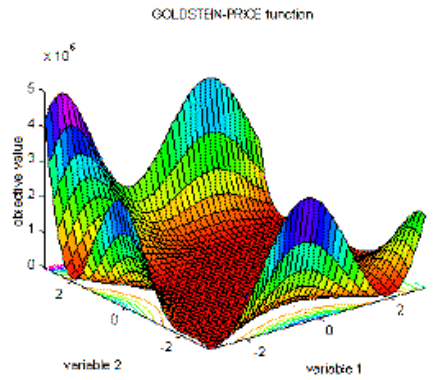
- Error(*i*) set to 1
- We have many more terms N(N-1)/2 in χ^2 than unknowns (3N) so that says there are enough constraints to find unknowns
- However hard to find components on which χ^2 depend little on
  - For N ~ 10^5, there are a lot of ill defined components

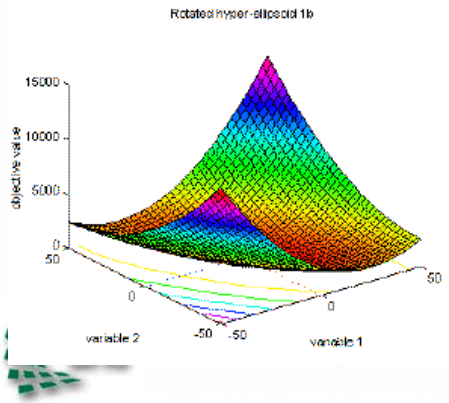
http://www.geatbx.com/docu/fcnindex-01.html

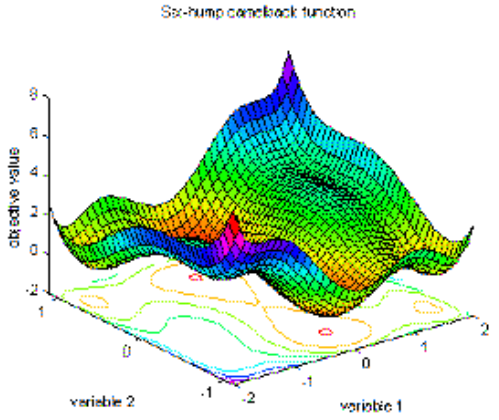












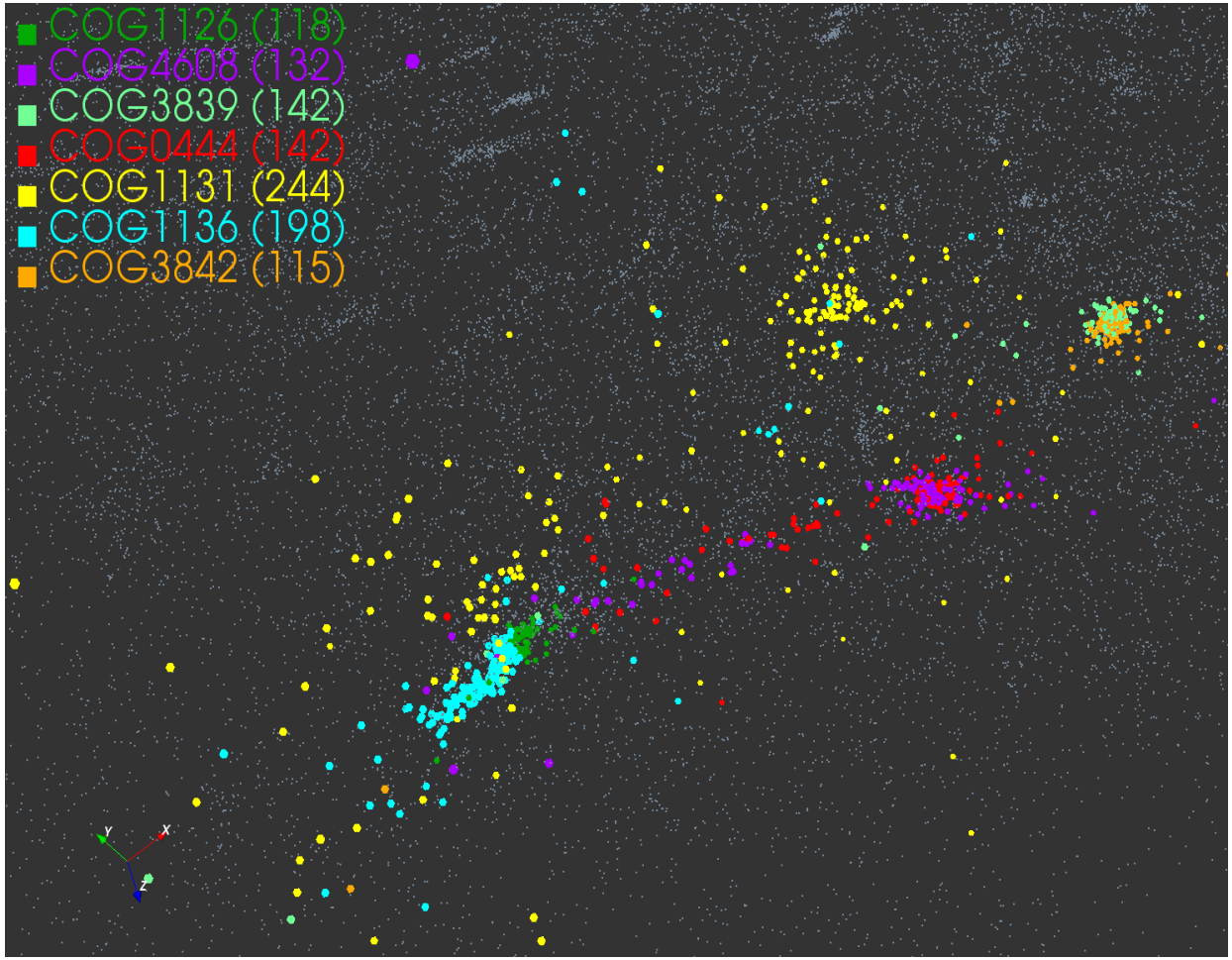
COG: Clusters of Orthologous Groups

- COG database was developed by NCBI.
- Proteins classified into groups with common function encoded in complete genomes.
- Prokaryotes (COG): 66 genomes, 200K proteins, 5K clusters.
- Eukaryotes (KOG): 7 genomes, 113K proteins, 5K clusters.
- Valuable scientific resource: 5K citations.
- Last updated: 2006.

Cluster Annotation

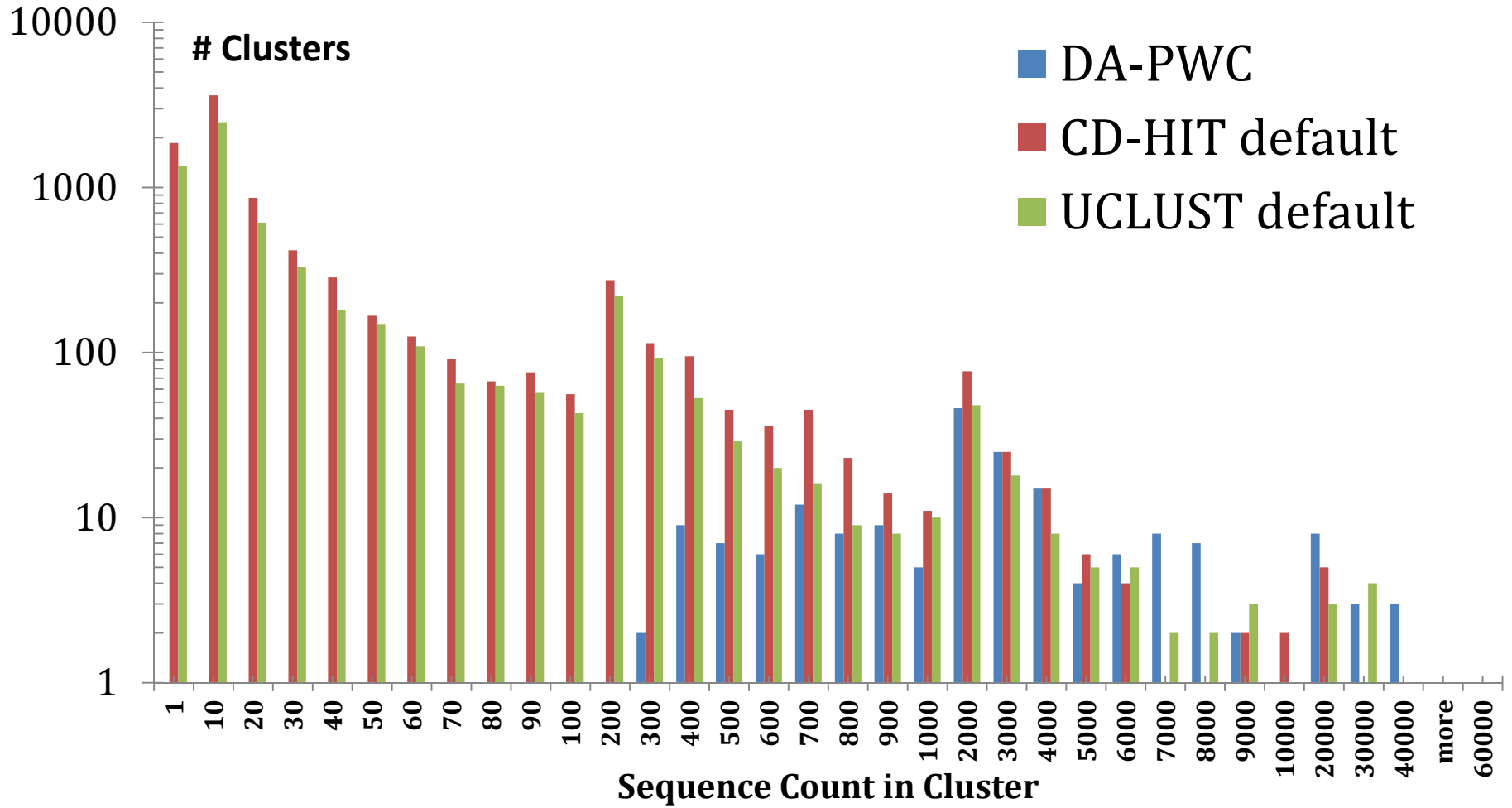
COG	Annotation	Uniref100
COG1131	ABC-type multidrug transport system, ATPase component	14406
COG1136	ABC-type antimicrobial peptide transport system, ATPase component	7306
COG1126	ABC-type polar amino acid transport system, ATPase component	4061
COG3839	ABC-type sugar transport systems, ATPase component	4121
COG0444	ABC-type dipeptide/oligopeptide/nickel transport system ATPase comp	3520
COG4608	ABC-type oligopeptide transport system, ATPase component	3074
COG3842	ABC-type spermidine/putrescine transport systems, ATPase comp	3665
COG0333	Ribosomal protein L32	1148
COG0454	Histone acetyltransferase HPA2 and Related acetyltransferases	14085
COG0477	Permeases of the major facilitator superfamily	48590
COG1028	Dehydrogenases with different specificities	37461

Selected Clusters



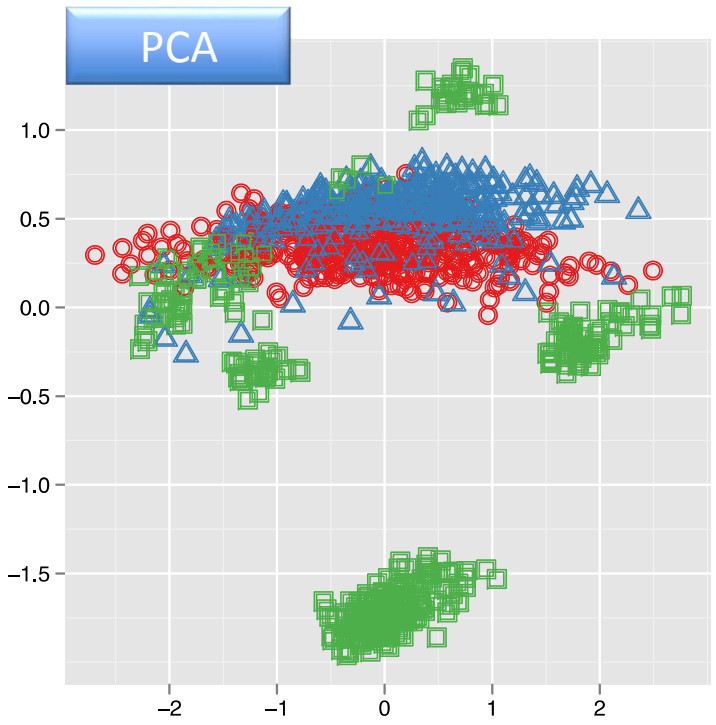
Metagenomics with 3 Clustering Methods

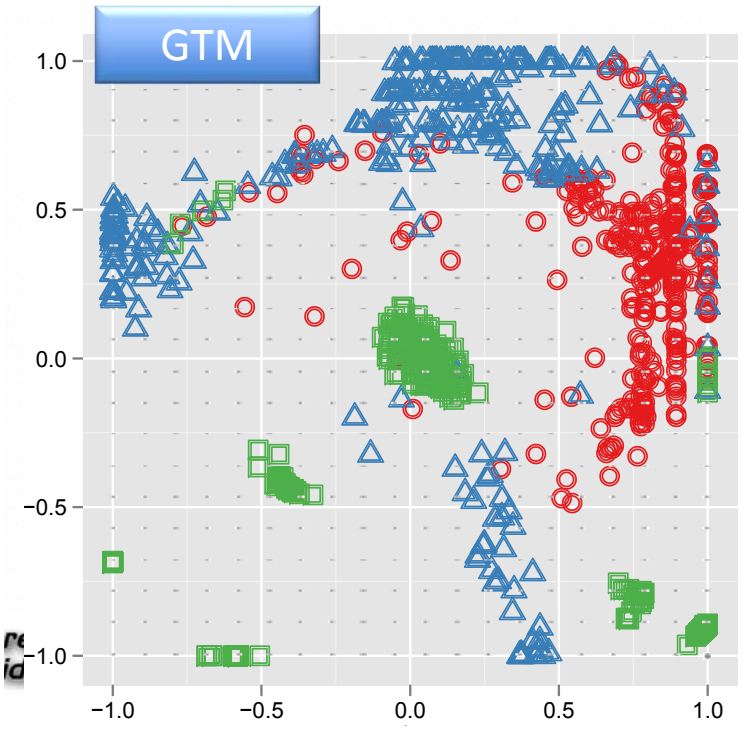
- DA-PWC 188 Clusters; CD-Hit 6000; UCLUST 8418
- DA-PWC doesn't need seeding like other methods – All clusters found by splitting



Advantages of GTM

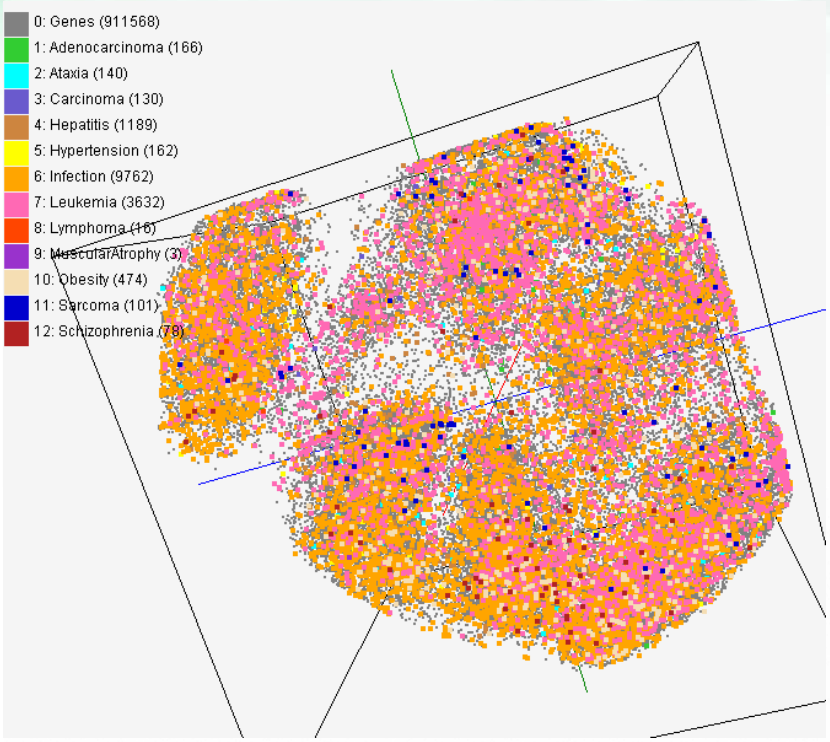
- Computational complexity is *O*(KN), where
  - N is the number of data points
  - K is the number of latent variables or *clusters*. K << N
- Efficient, compared with MDS which is *O*(N^2)
- Produce more separable map (right) than PCA (left)





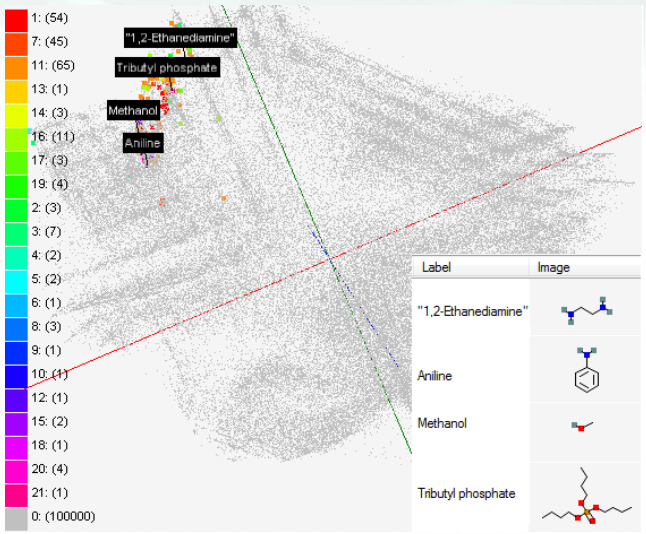
Oil flow data
1000 points
12 Dimensions
3 Clusters

Data Mining Projects using GTM



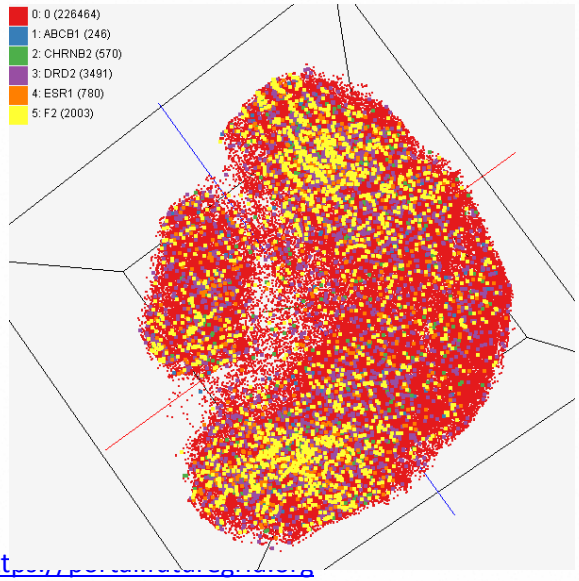
PubChem data with CTD visualization

About 930,000 chemical compounds are visualized in a 3D space, annotated by the related genes in Comparative Toxicogenomics Database (CTD)



Visualizing 215 solvents by GTM-Interpolation

215 solvents (colored and labeled) are embedded with 100,000 chemical compounds (colored in grey) in PubChem database



Chemical compounds reported in literatures

Visualized 234,000 chemical compounds which may be related with a set of 5 genes of interest (ABCB1, CHRNB2, DRD2, ESR1, and F2) based on the dataset collected from major journal literatures

DA-PLSA with DA-GTM

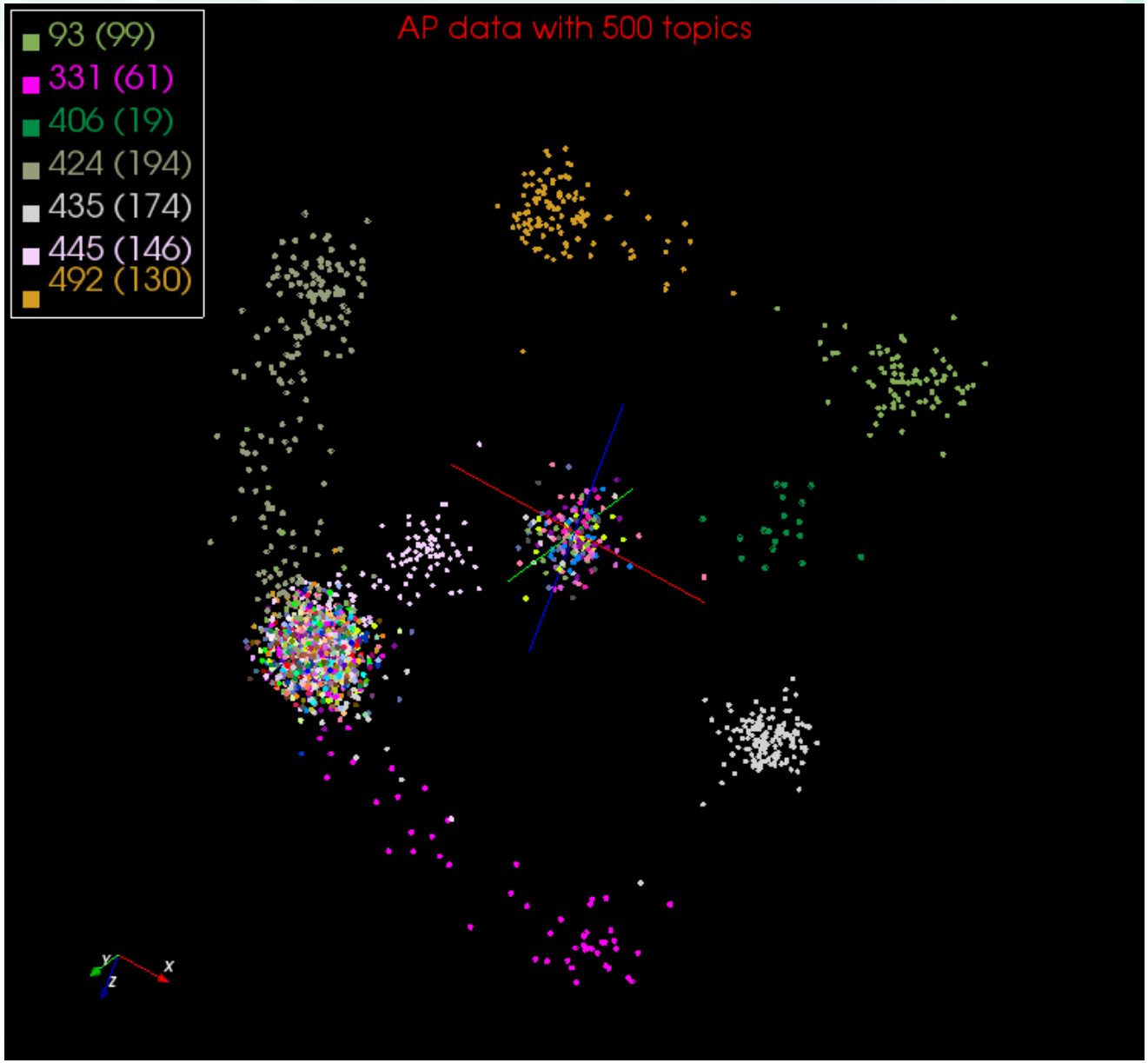
Corpus
(Set of documents)

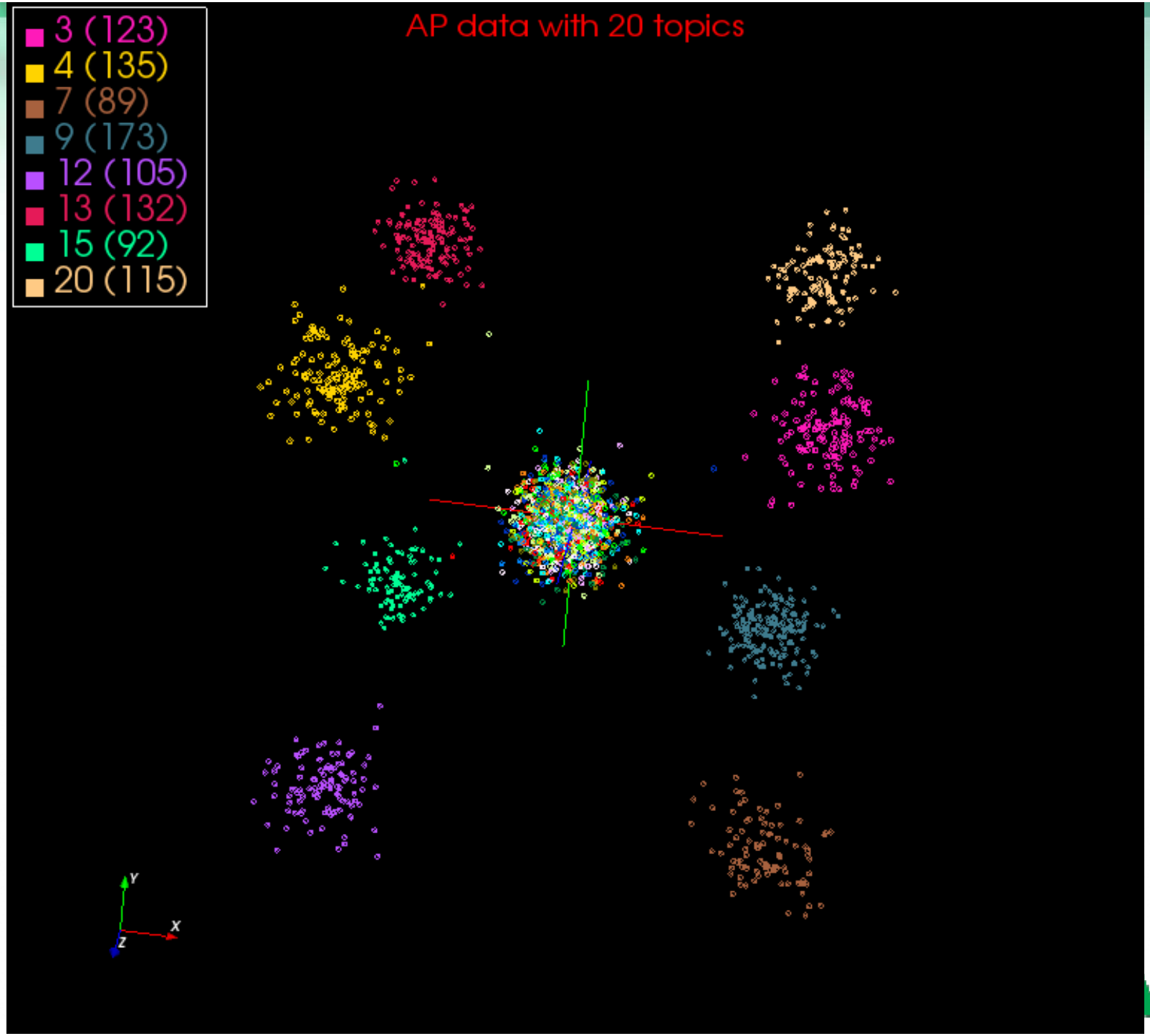
DA-PLSA

Corpus
in K-dimension

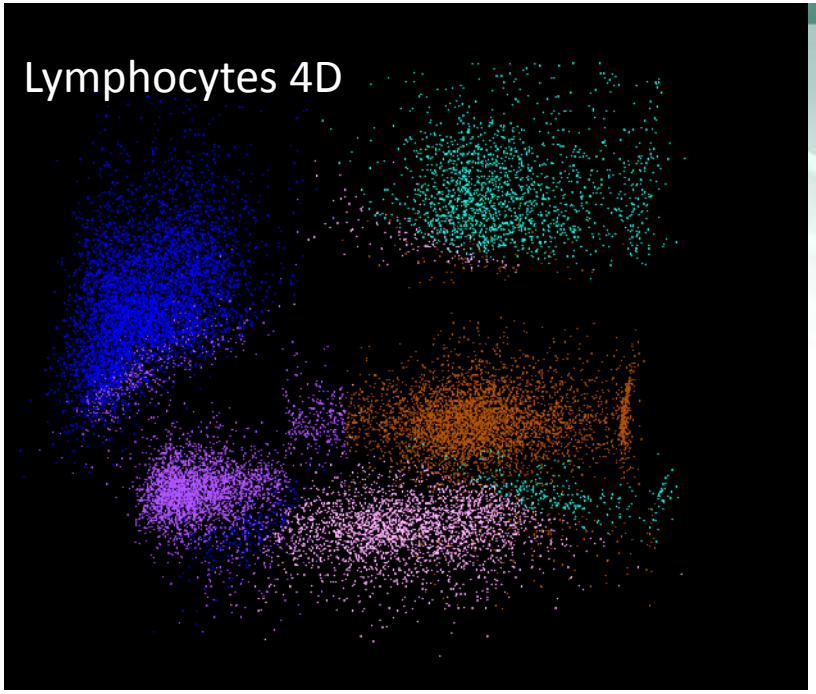
DA-GTM

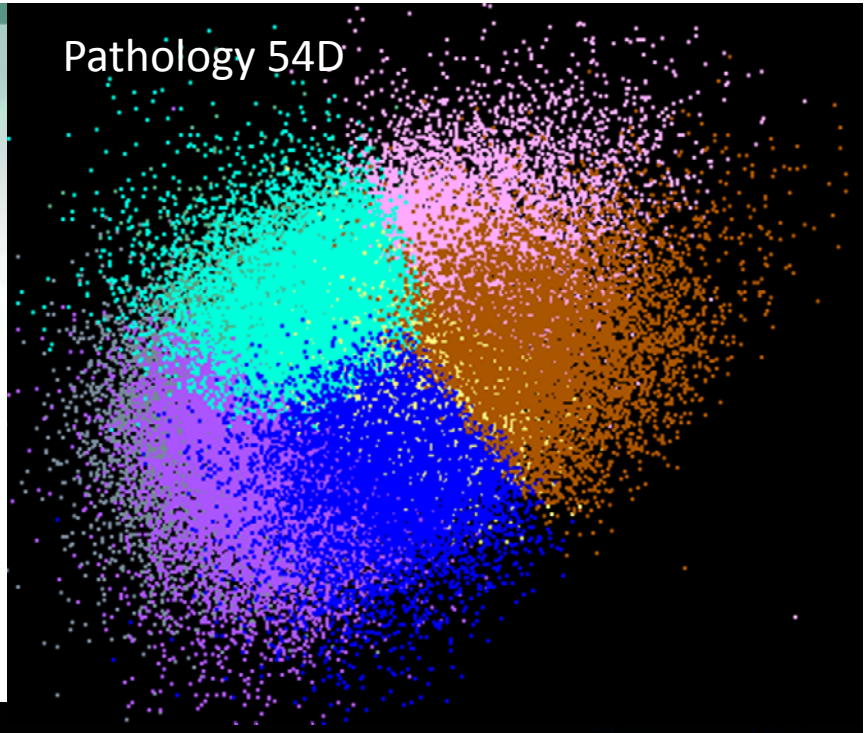
Embedded Corpus
in 3D



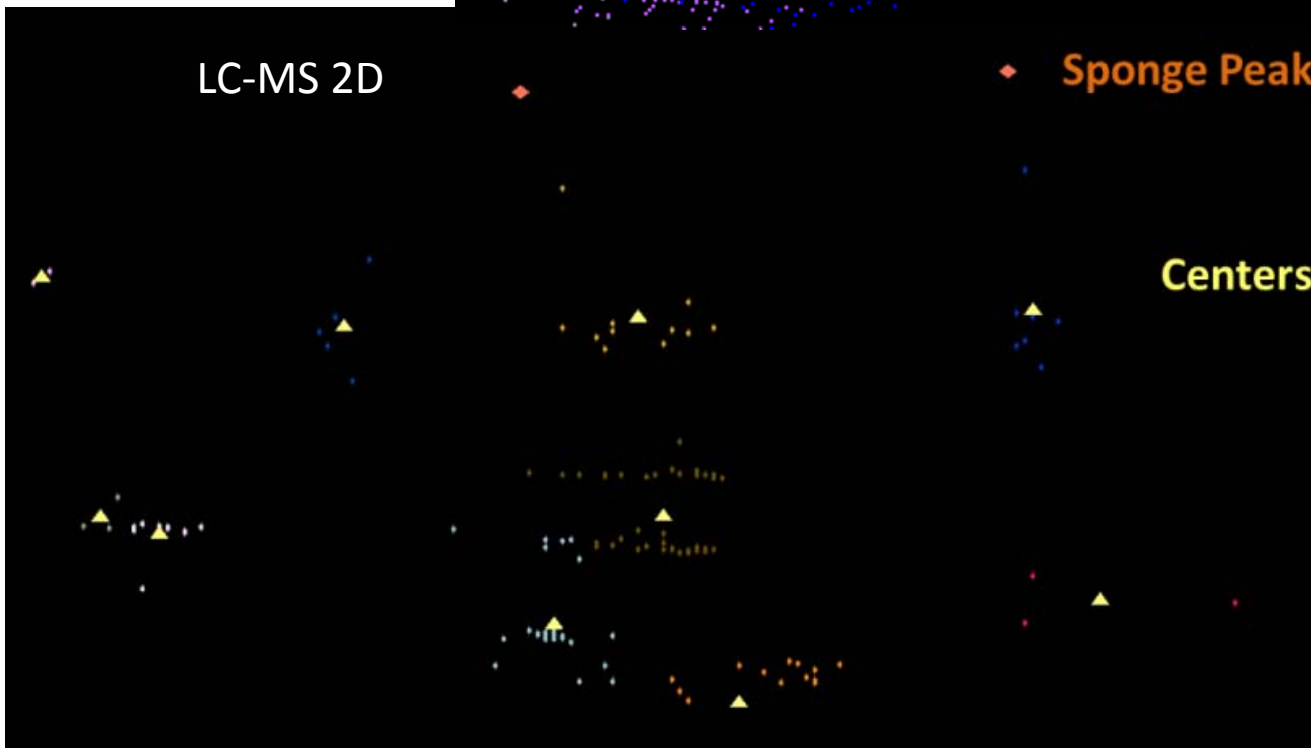

AP data with 20 topics
3 (123)
4 (135)
7 (89)
9 (173)
12 (105)
13 (132)
15 (92)
20 (115)
Y
X
Z

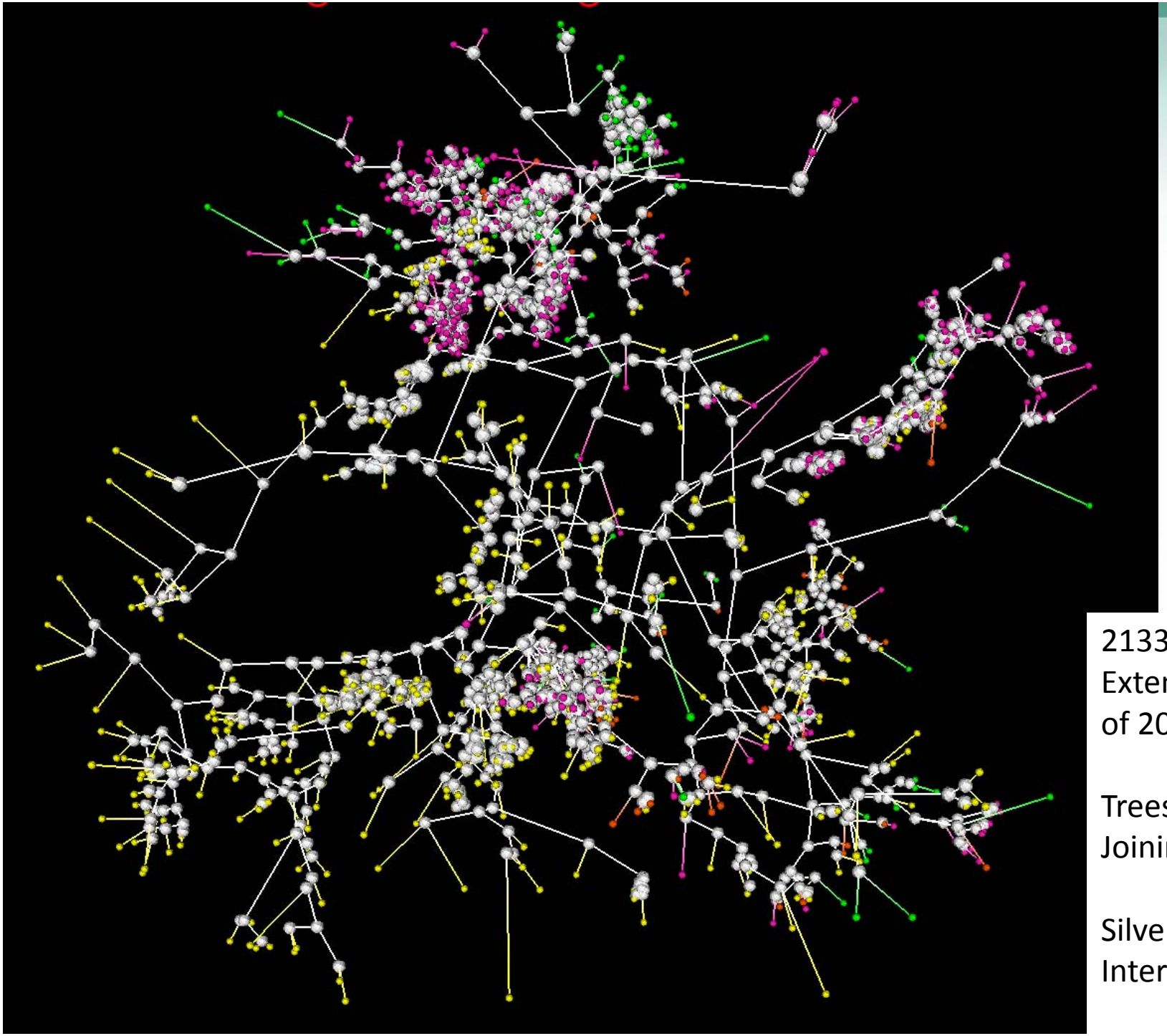
HPC





- Dimension Reduction/MDS helps address
- You can get answers (from clustering) but do and how do you believe them!





2133 Sequences
Extended from set of 200

Trees by Neighbor Joining in 3D map

Silver Spheres
Internal Nodes