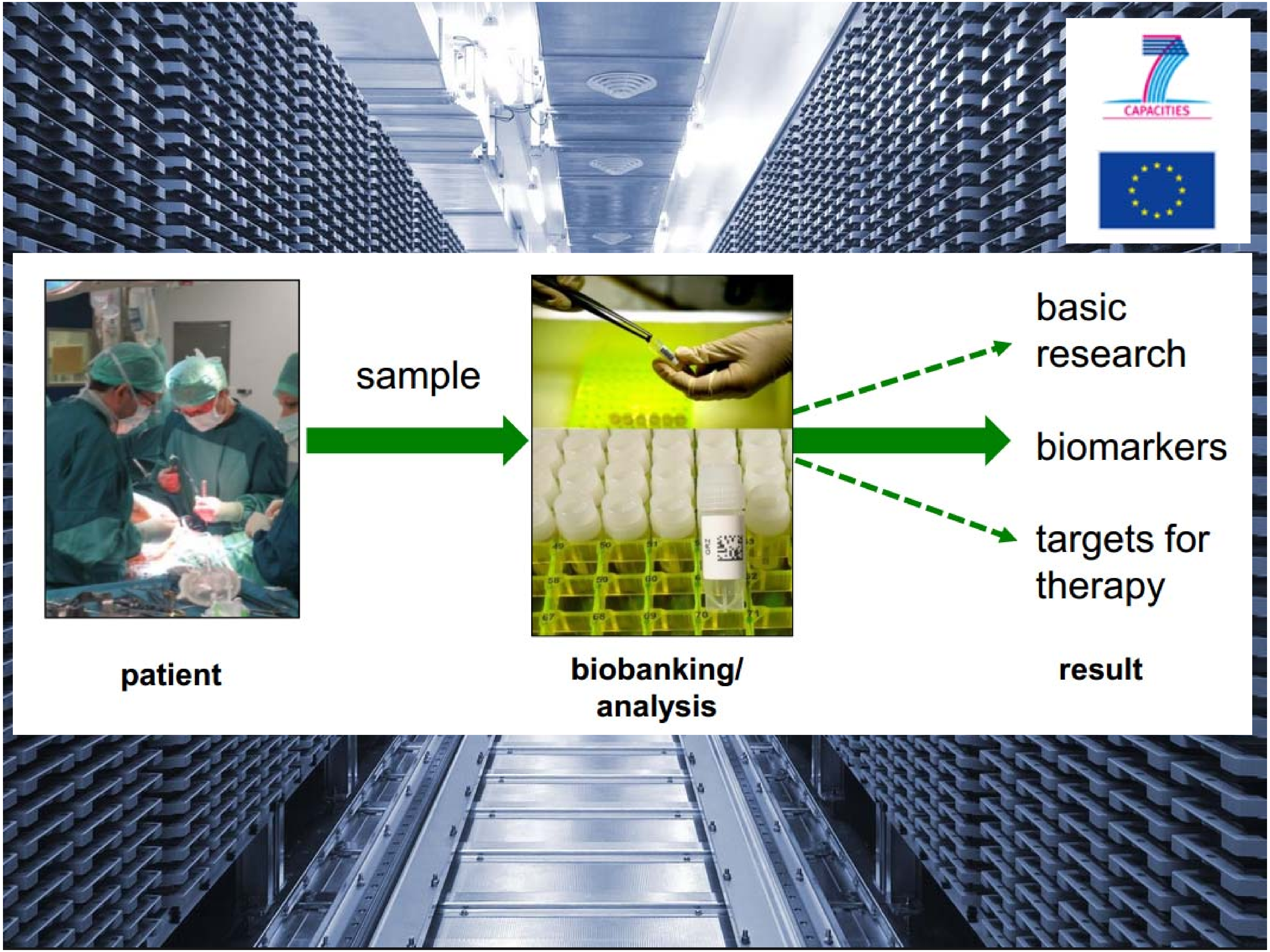
Institute for Medical Informatics, Statistics and Documentation
Research Unit Human-Computer Interaction
Medical University of Graz

On Interactive Data Mining

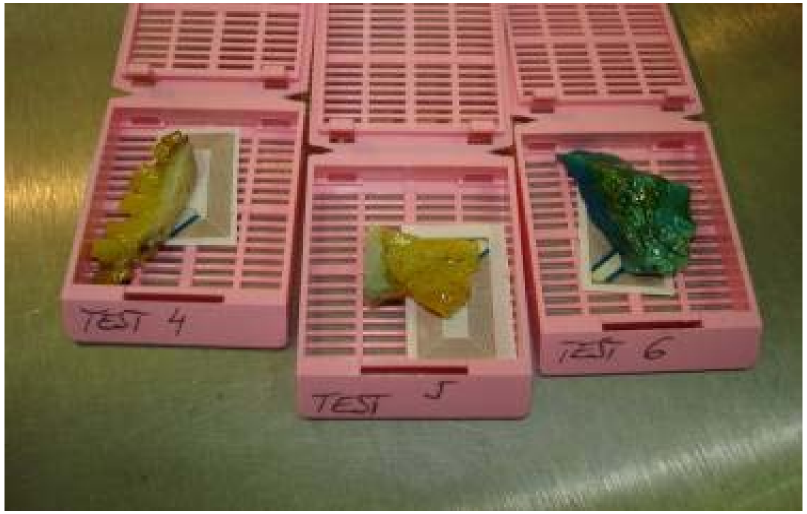
Andreas Holzinger

Lisboa, 07.01.2014
www.hci4all.at


HELP


CAPACITIES
sample
basic research
biomarkers
targets for therapy
patient
biobanking/ analysis
result

Biggest Pathology in Europe produces massive data sets …



Holzinger, A., Schwaberger, K. & Weitlaner, M. 2005. Ubiquitous Computing for Hospital Applications: RFID-Applications to enable research in Real-Life environments, *IEEE COMPSAC*

1970 …
Collecting data …

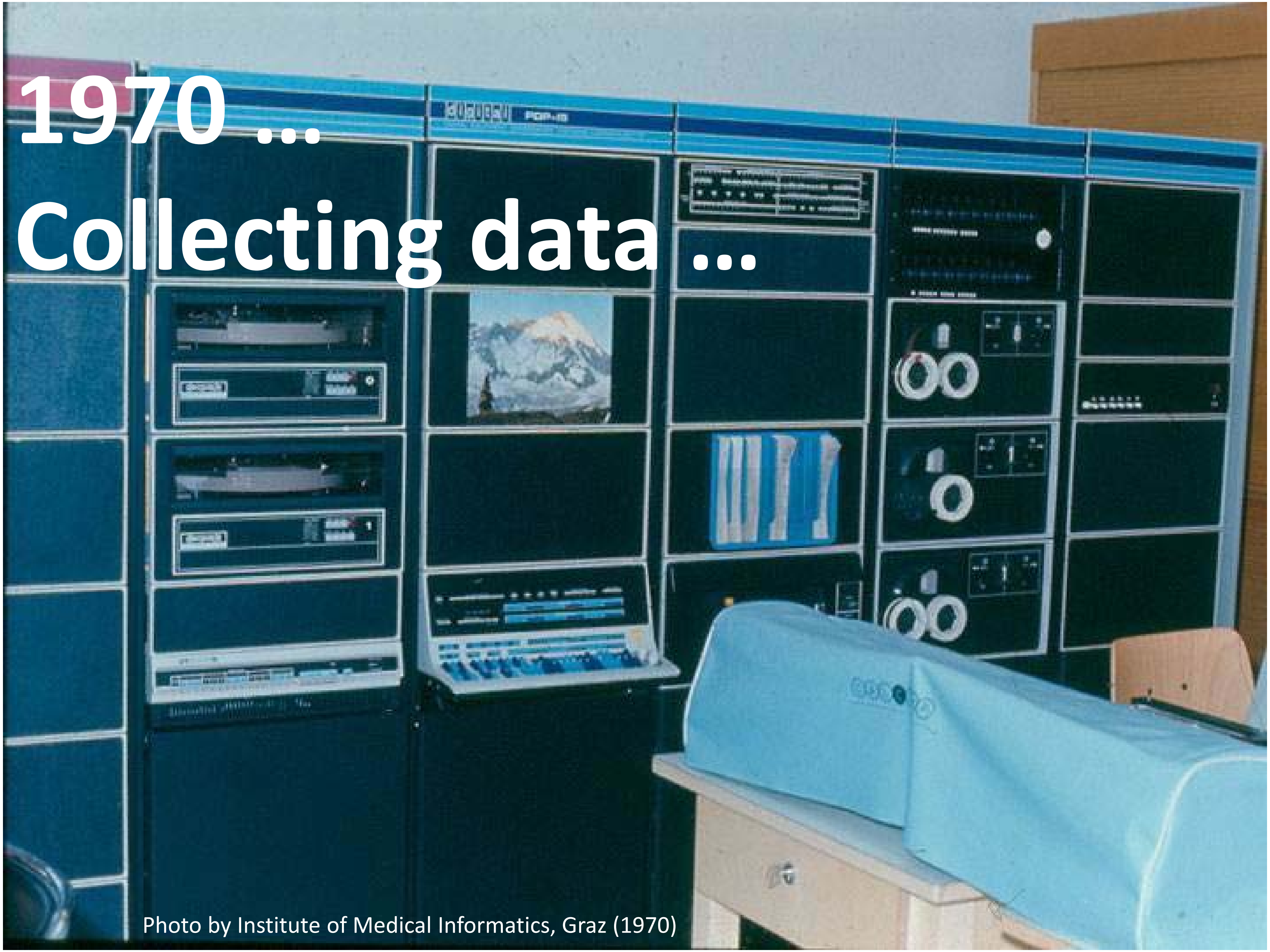
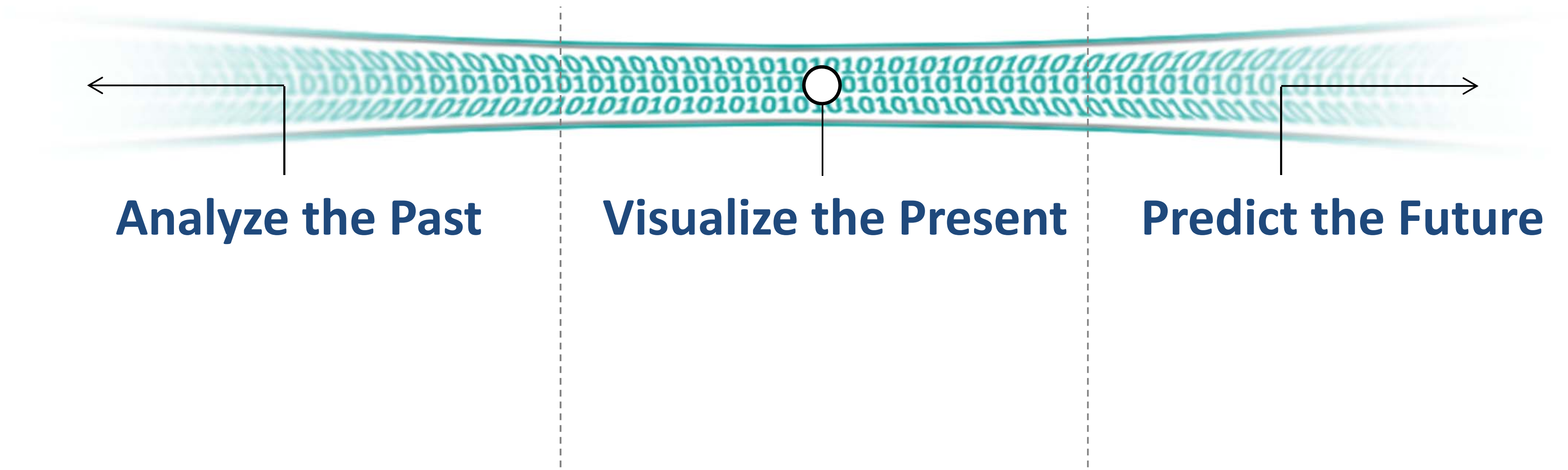
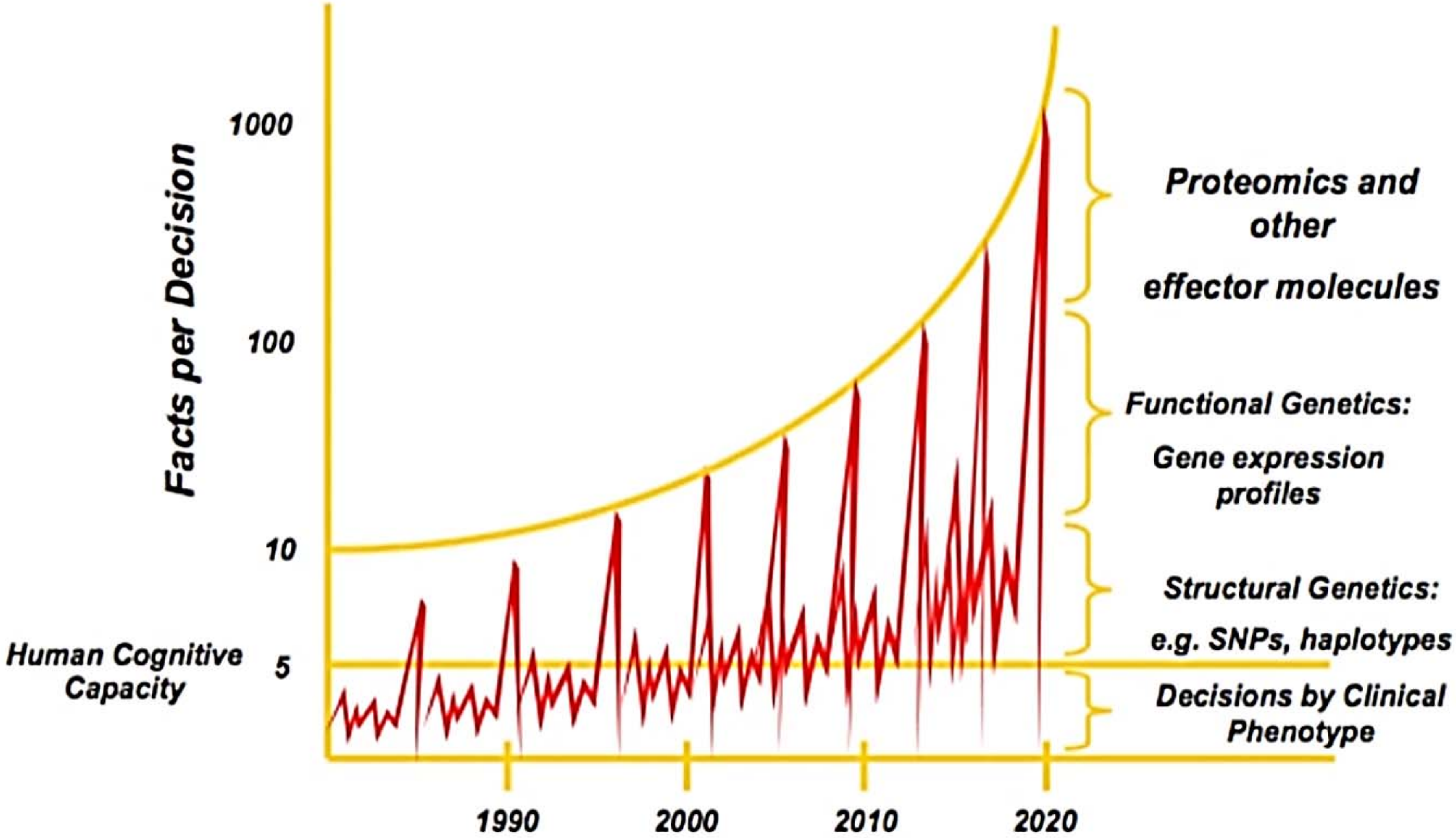


Photo by Institute of Medical Informatics, Graz (1970)


four decades later …
… data, data, data …


Analyze the Past
Visualize the Present
Predict the Future

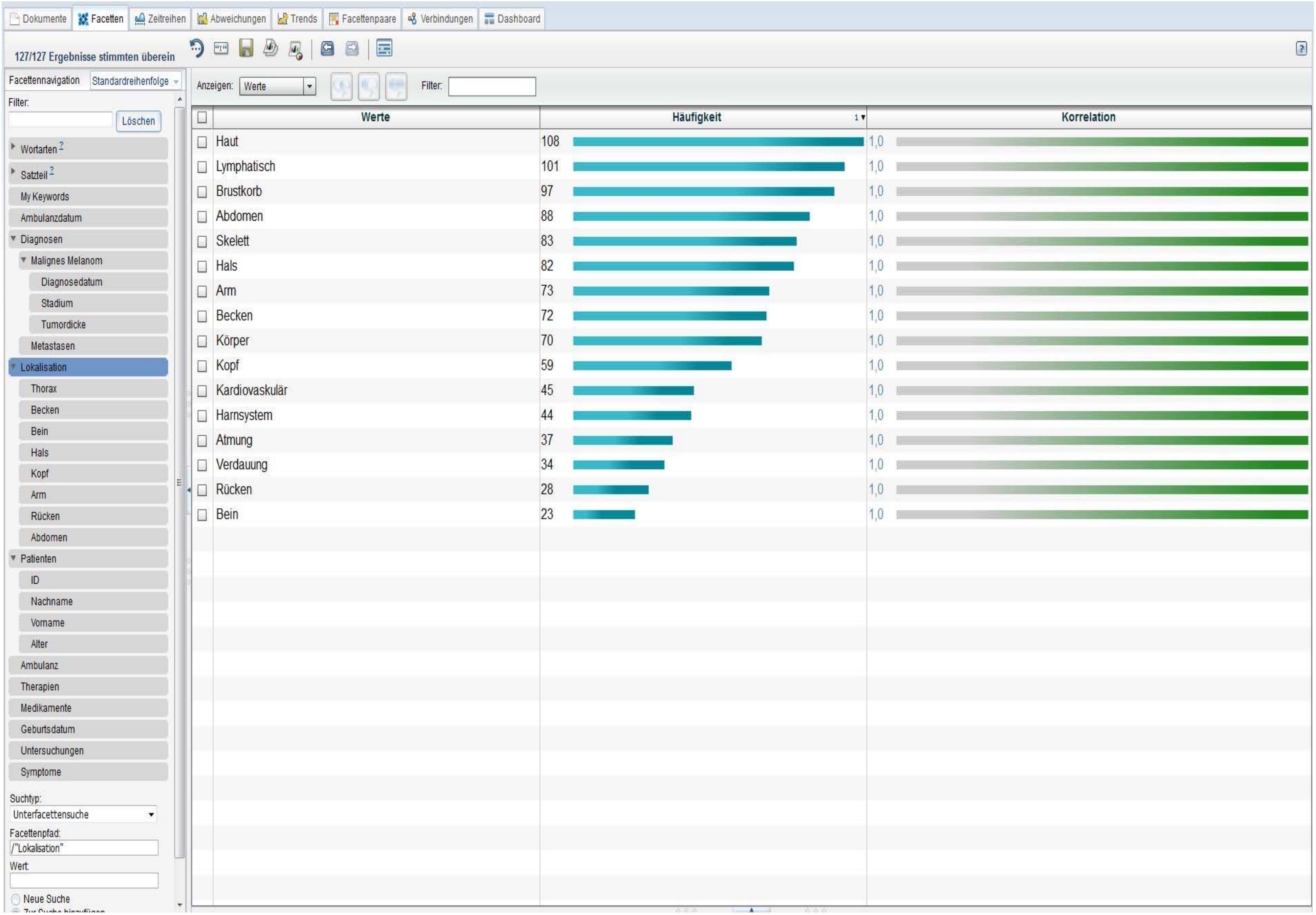
No medical doctor can remember ALL the factual data …



Stead, W. (2007) Growth in facts affecting provider decisions versus human cognitive capacity, IOM Meeting

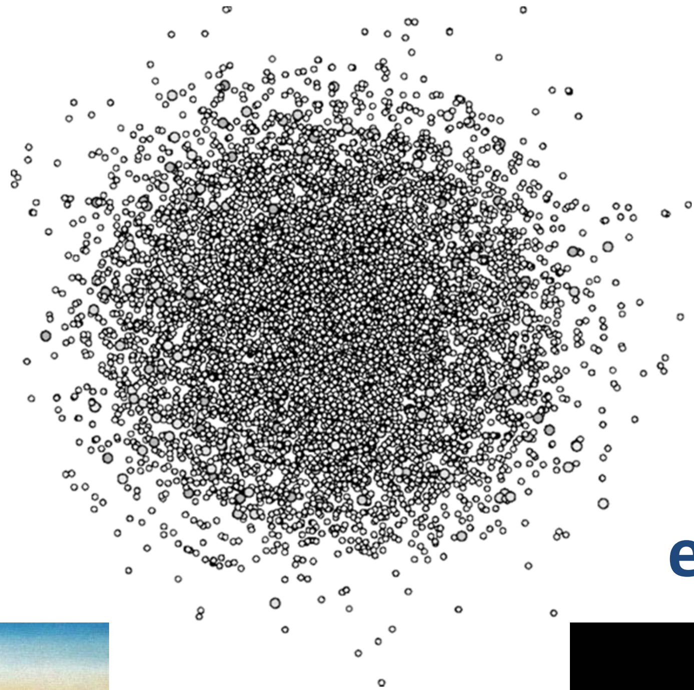

IBM
POWER7
2020

Very first steps …



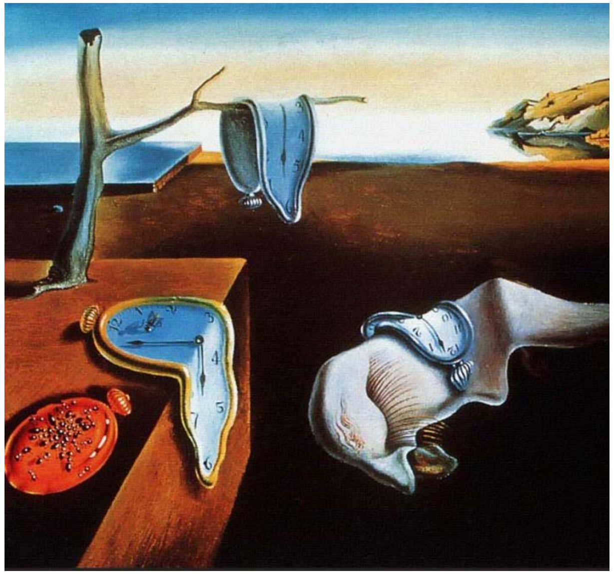
Holzinger, A., Stocker, C., Ofner, B., Prohaska, G., Brabenetz, A. & Hofmann-Wellenhof, R. 2013. Combining HCI, Natural Language Processing, and Knowledge Discovery - Potential of IBM Content Analytics as an assistive technology in the biomedical domain. Springer Lecture Notes in Computer Science LNCS 7947. Heidelberg, Berlin, New York: Springer, 13-24.

Two issues …



Time

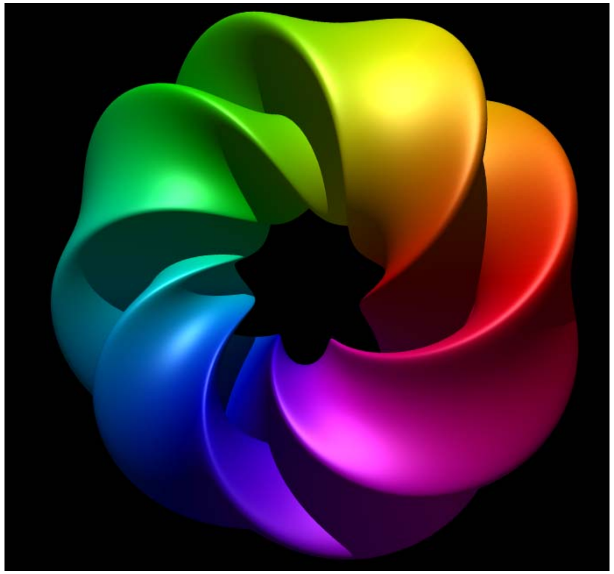
e.g. Entropy



Dali, S. (1931) The persistence of memory

Space

e.g. Topology



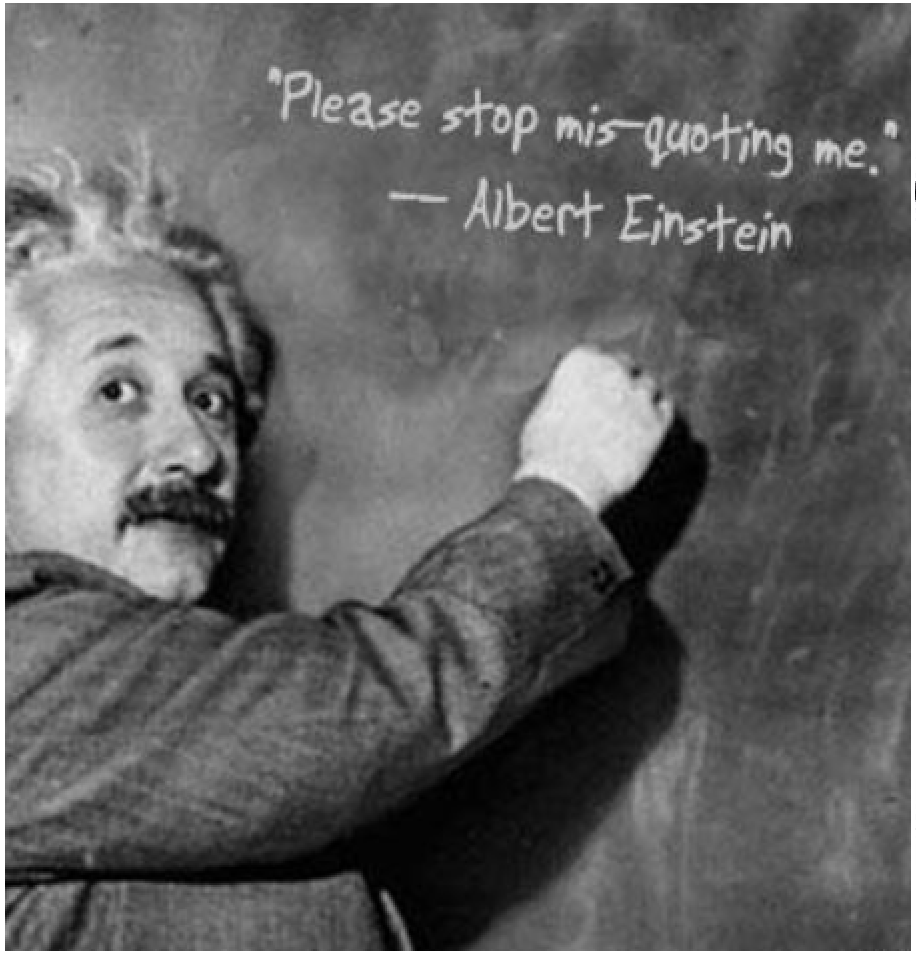
Bagula & Bourke (2012) Klein-Bottle

Human Interaction Computer

Human intelligence harnesses machine intelligence

Holzinger, A. 2012. On Knowledge Discovery and interactive intelligent visualization of biomedical data.
In: DATA - International Conference on Data Technologies and Applications.

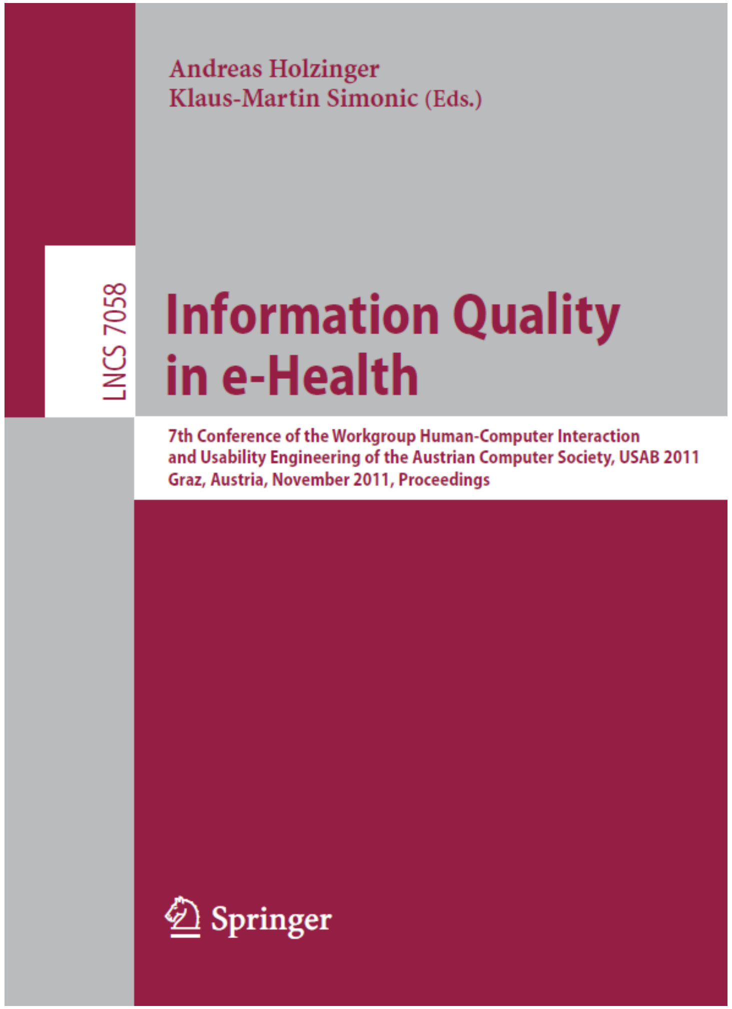
… BOTH together are powerful beyond imagination …



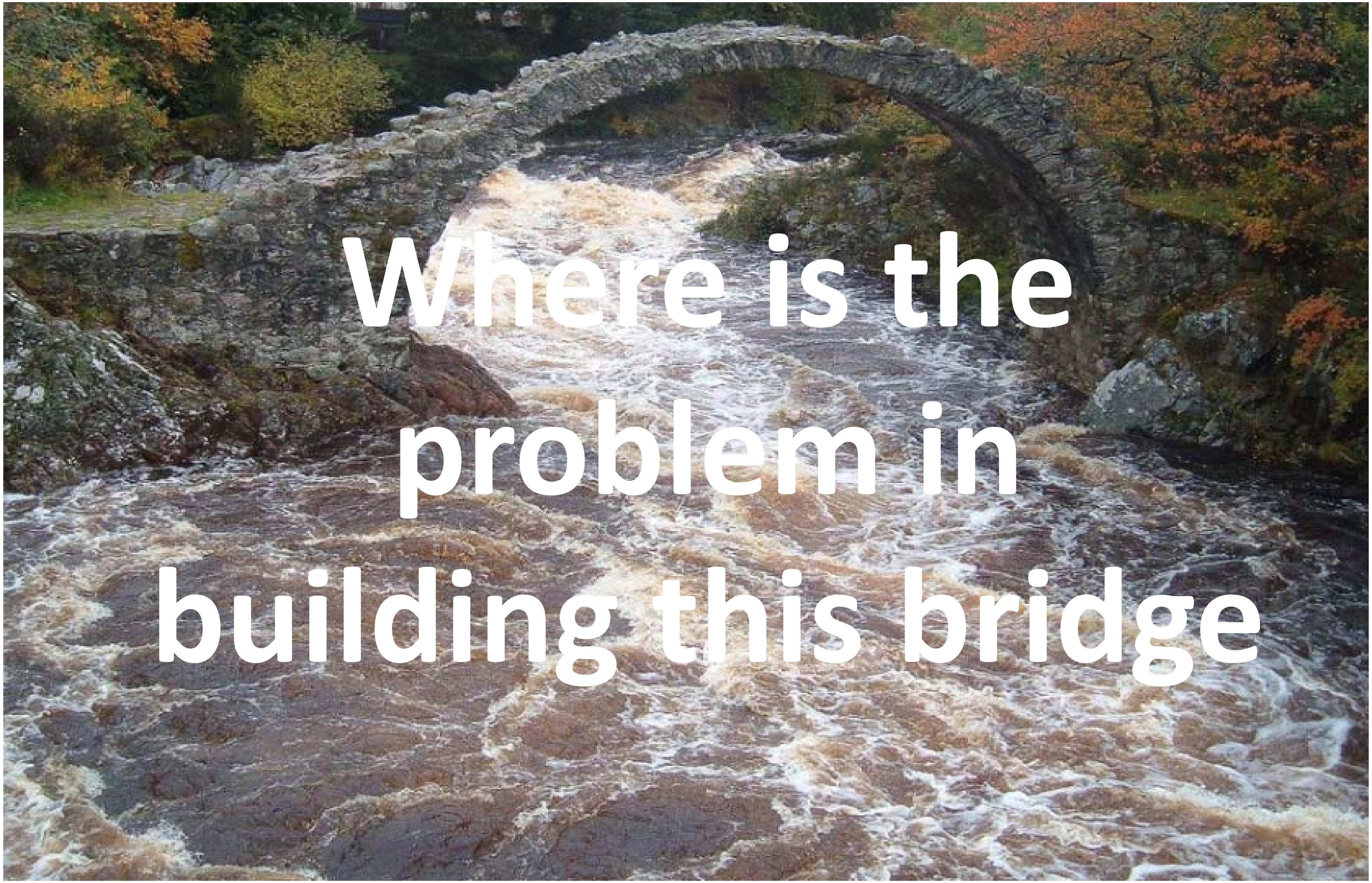
Our central hypothesis: Information bridges this gap

Holzinger, A. & Simonic, K.-M. (eds.) 2011. *Information Quality in e-Health, Springer LNCS 7058*

Modern information management can <u>bridge the hiatus theoreticus</u>, the gap between (scientific) knowledge and its application



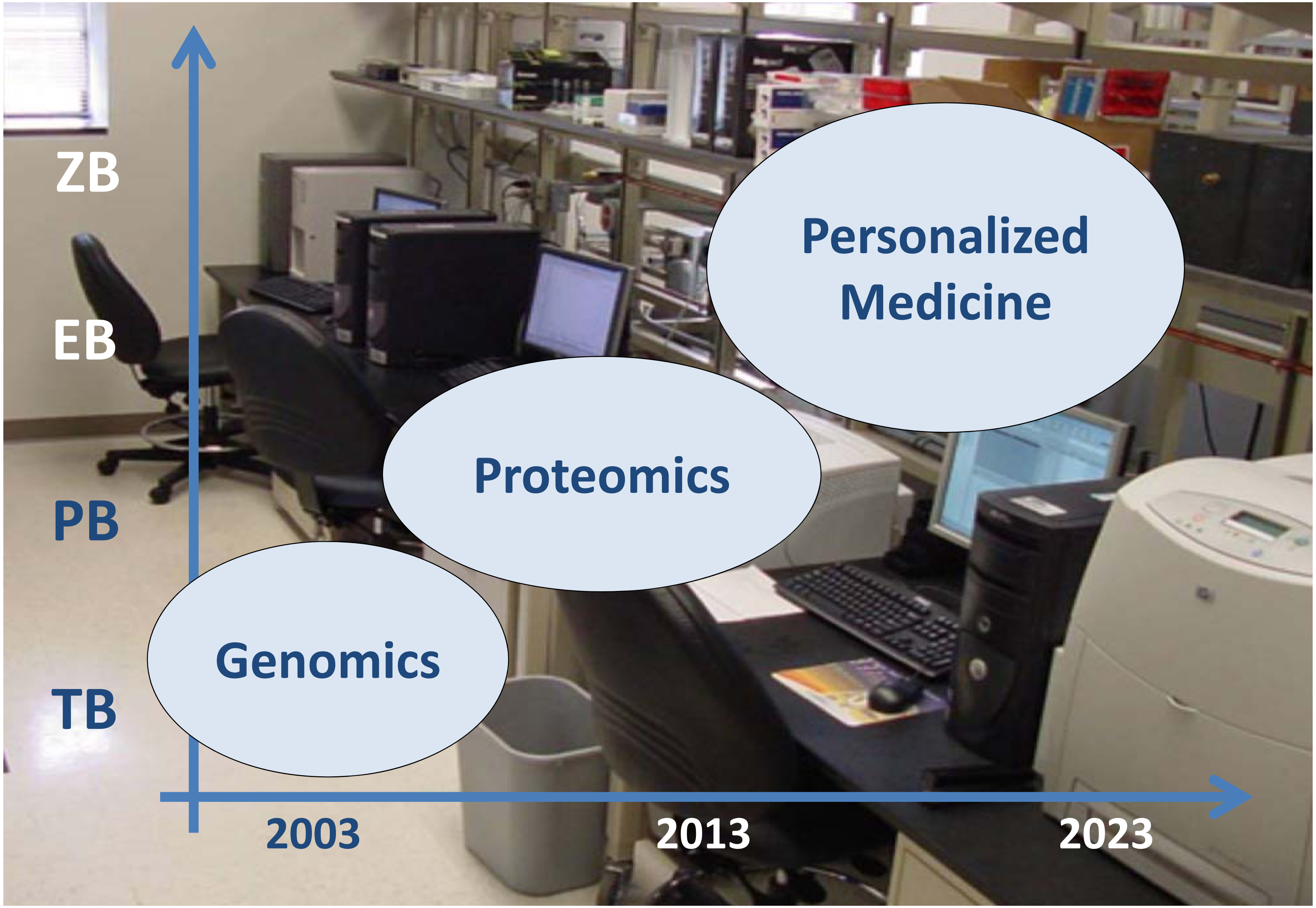
Holzinger, A. & Simonic, K.-M. (Eds.) (2011) *Information Quality in e-Health. Lecture Notes in Computer Science LNCS 7058, Heidelberg, New York, Springer.*

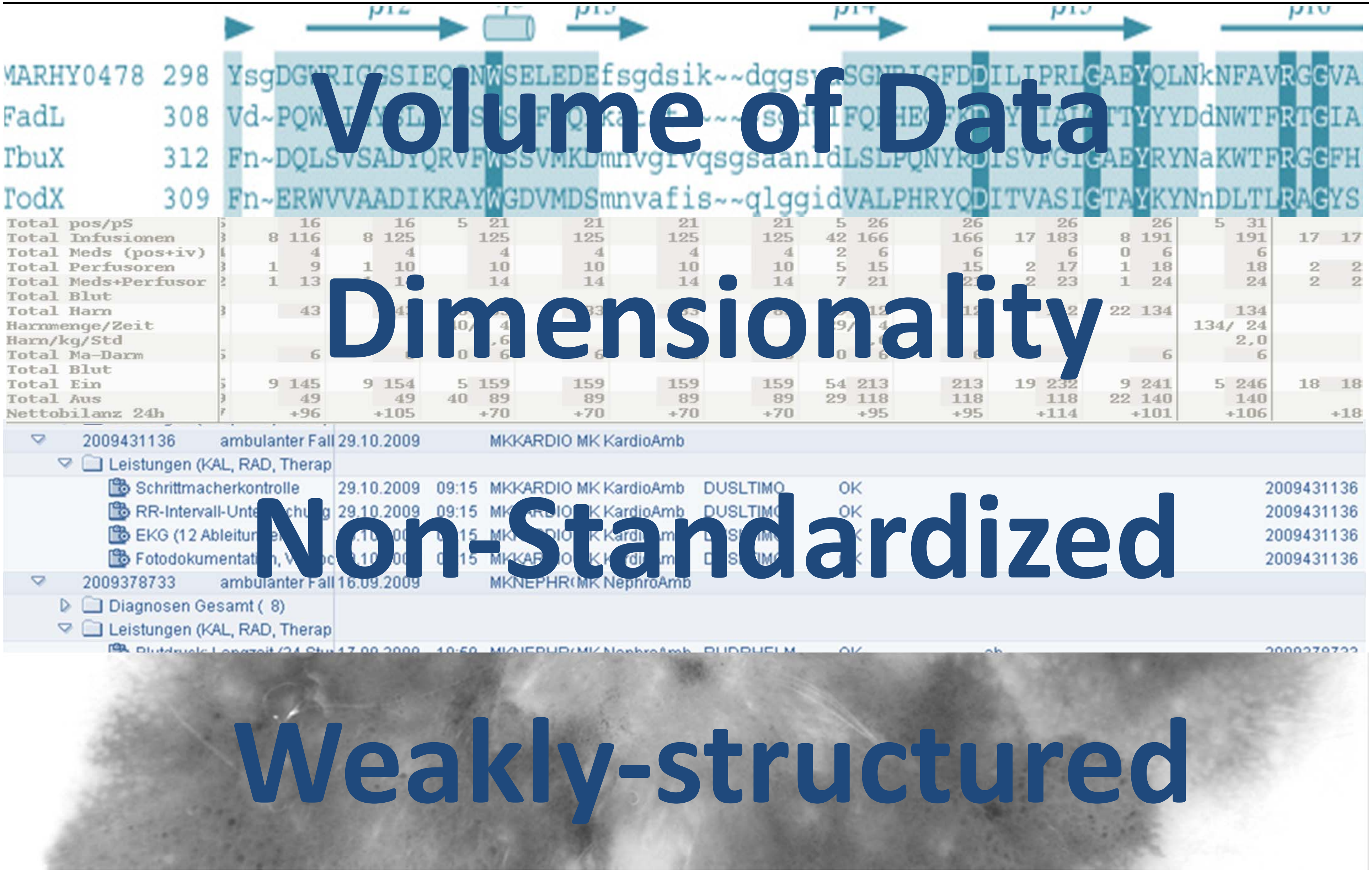

Where is the
problem in
building this bridge

Ok, let us ask an representative CEO …


4 September 2008 | www.nature.com/nature | £10
HBR.ORG
Harvard Business Review
OCTOBER 2012
The Big Idea
The True Measures of Success
Michael J. Mauboussin
International Business
10 Rules for Managing Global Innovation
Keeley Wilson and Yves L. Doz
Leadership
What Ever Happened To Accountability?
Thomas E. Ricks
GETTING CONTROL OF
BIG DATA
How vast new streams of information are changing the art of management
PAGE XX

Let us look into the Future …





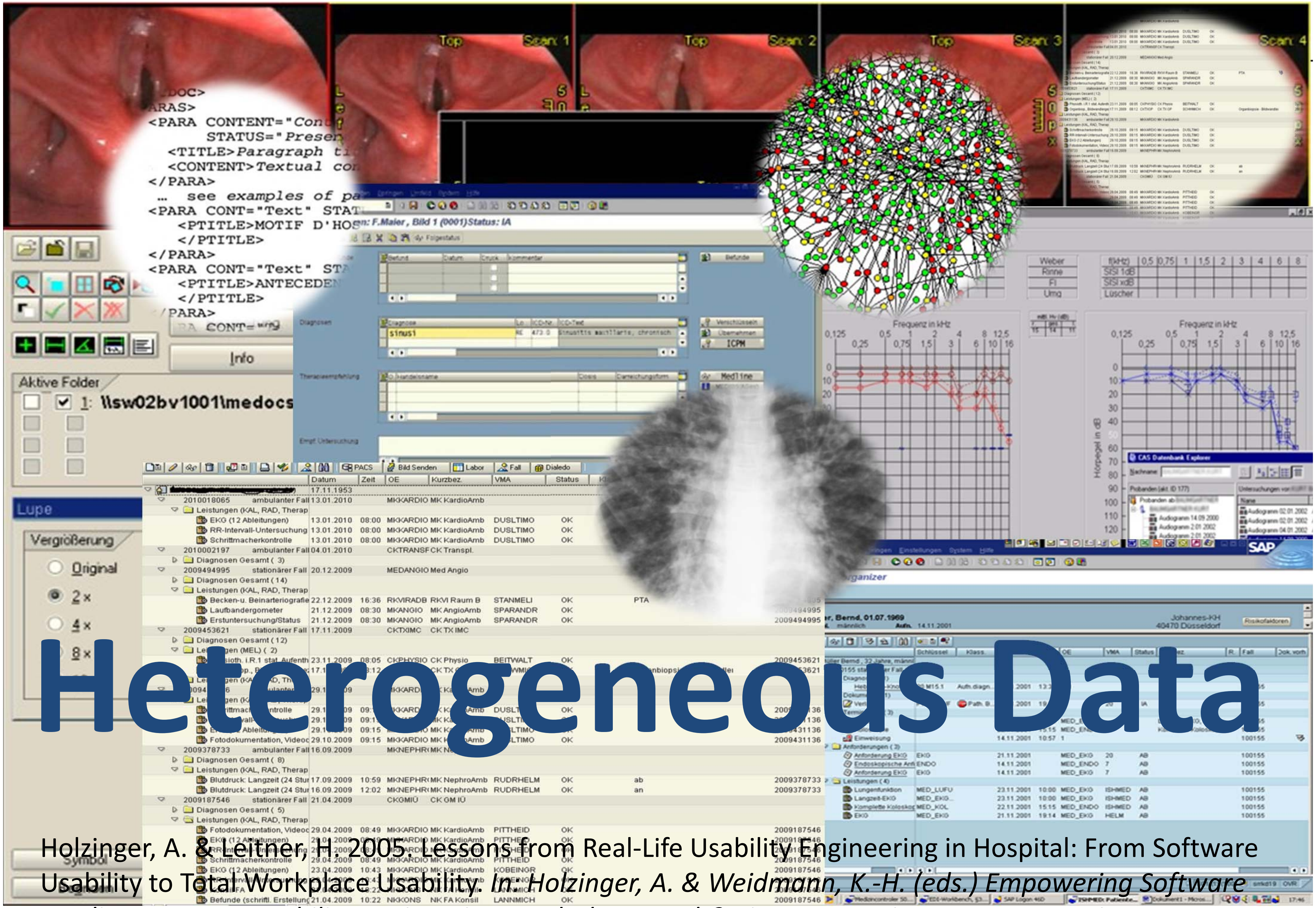
Volume of Data

Dimensionality

Non-Standardized

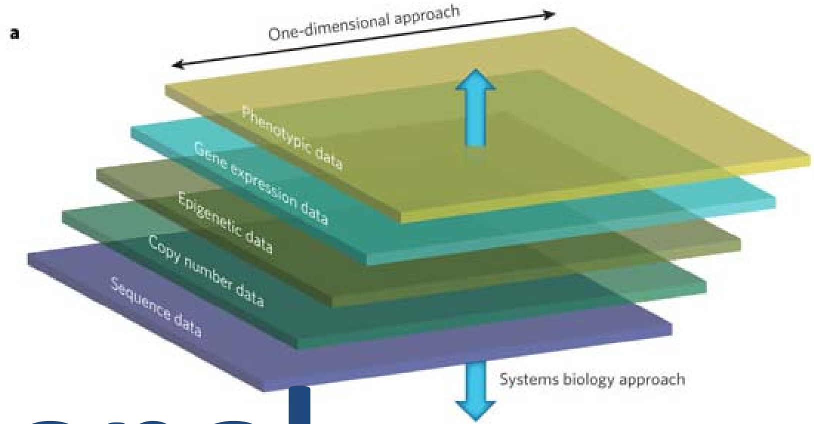
Weakly-structured

Holzinger, A. (2011) Weakly Structured Data in Health-Informatics. In: *INTERACT 2011, Lisbon, IFIP, 5-7.*

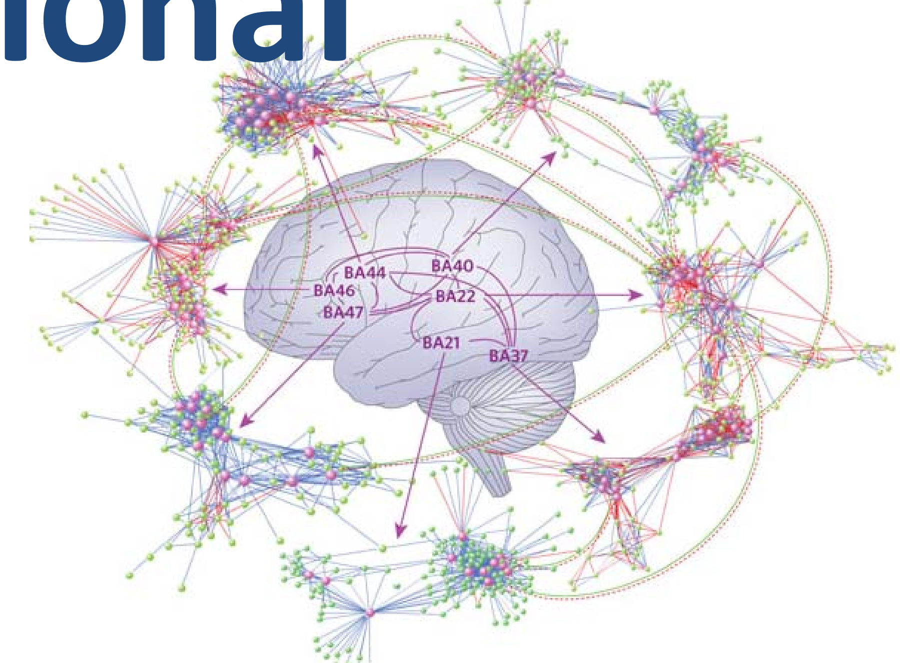


Heterogeneous Data

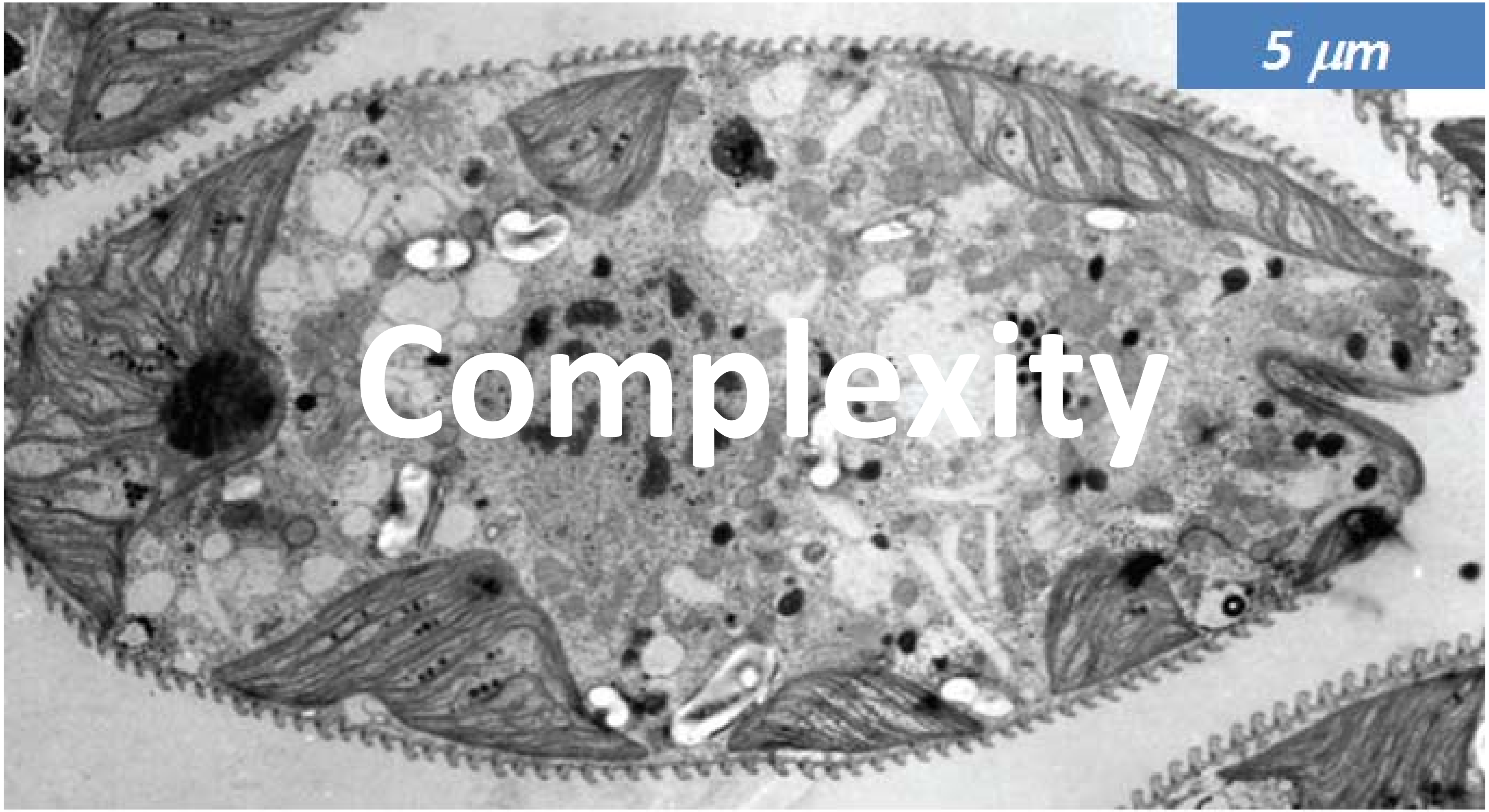
Holzinger, A. & Leitner, H. 2005. Lessons from Real-Life Usability Engineering in Hospital: From Software Usability to Total Workplace Usability. *In: Holzinger, A. & Weidmann, K.-H. (eds.) Empowering Software Quality: How can Usability Engineering reach these goals? Vienna: Austrian Computer Society, pp. 153-160.*



Complex and High dimensional



Geschwind, D. H. & Konopka, G. 2009. Neuroscience in the era of functional genomics and systems biology. *Nature, 461, (7266), 908-915.*



Lane, N. & Martin, W. (2010) The energetics of genome complexity. *Nature, 467, 7318, 929-934.*


FOR A FAIR SELECTION EVERYBODY HAS TO TAKE THE SAME EXAM: PLEASE CLIMB THAT TREE

Standards ?

EBM CPG

Standardized Medicine

Person-alized Medicine

GBM GPM

Pervasive Healthcare

Preventive Health Integration

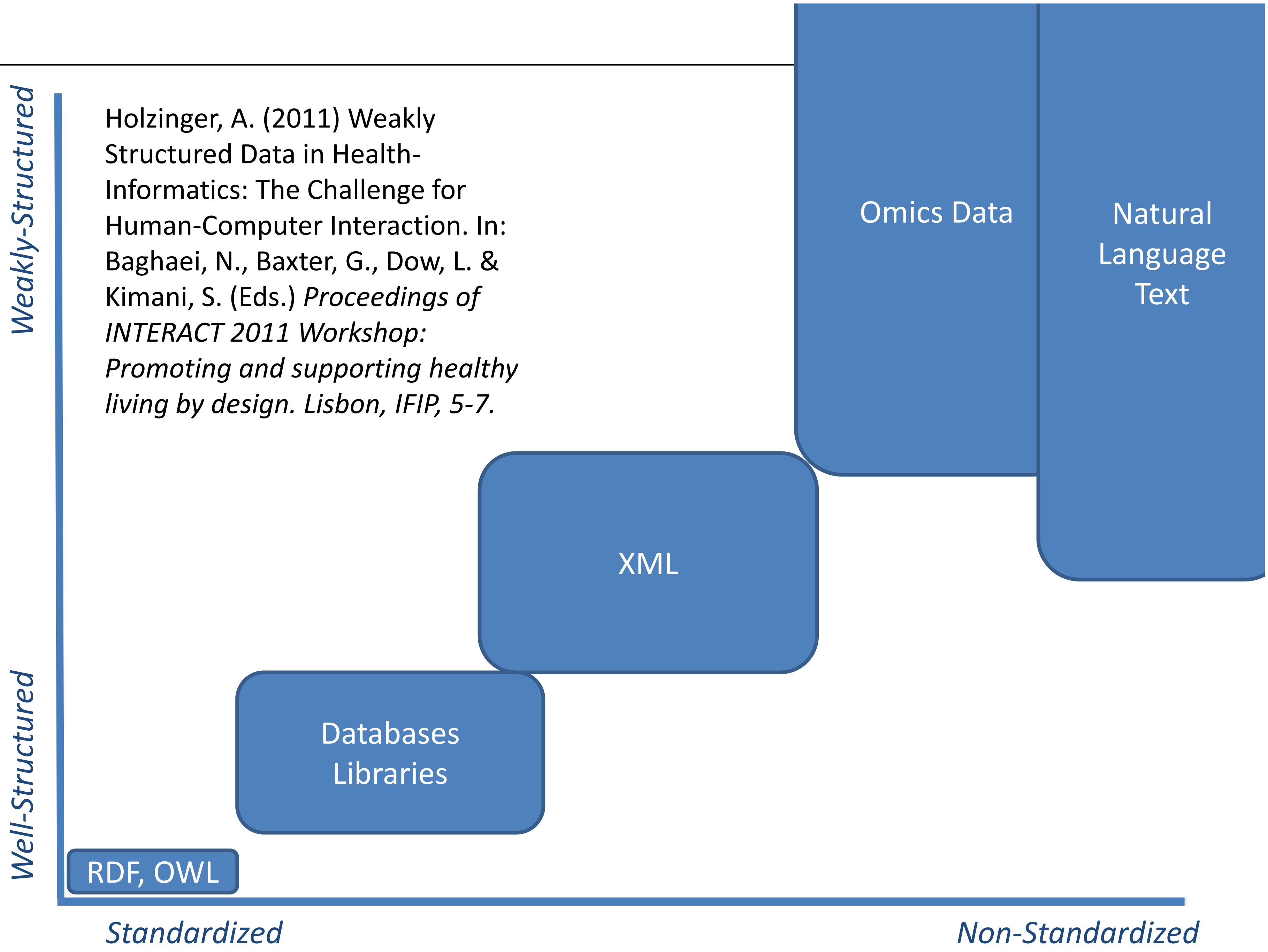
EBM = Evidence Based Medicine
CPG = Clinical Practice Guideline
GBM = Genome Based Medicine
GPM = Genetic Polymorphism

Tanaka, H. (2010)

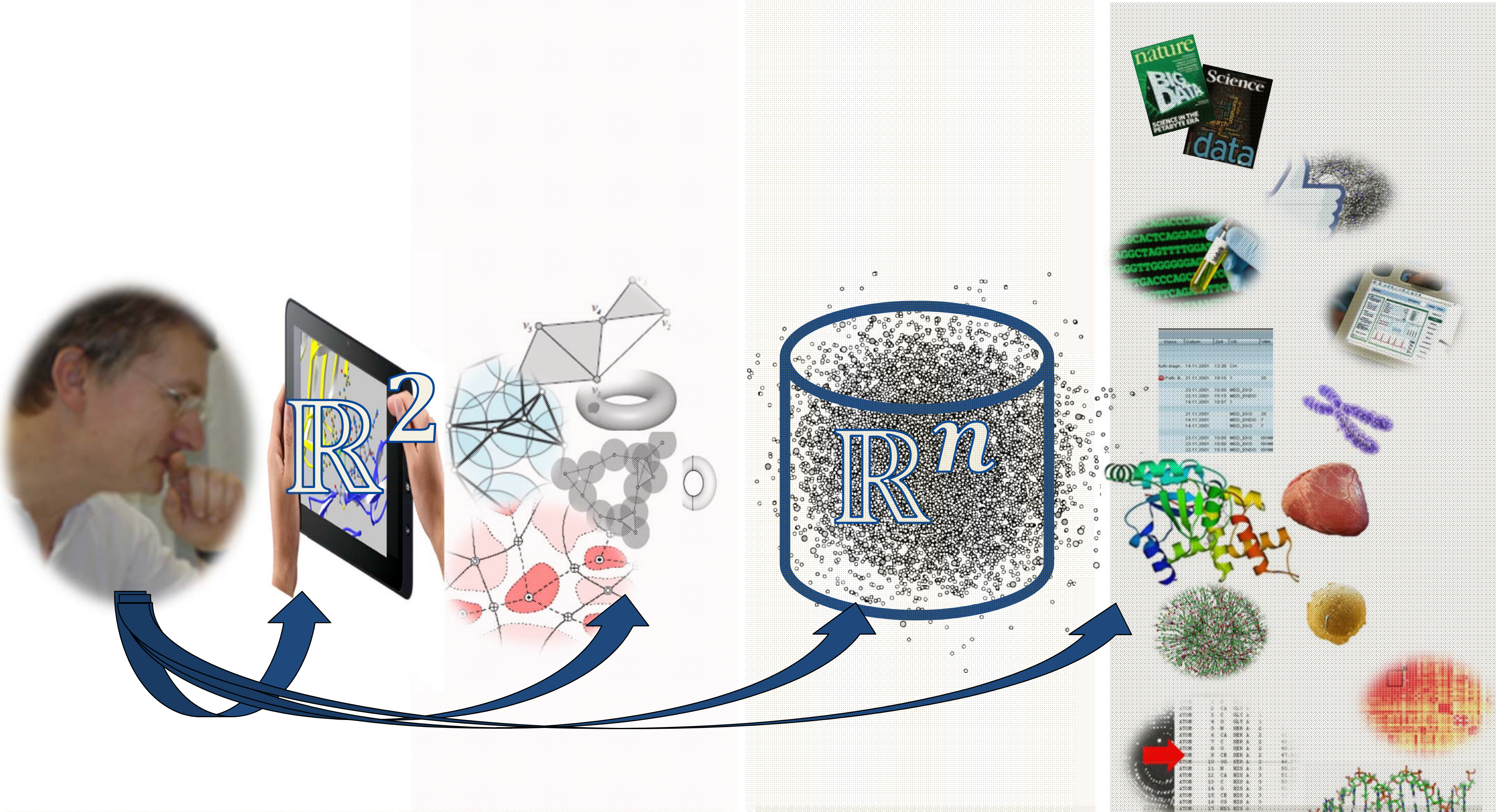
Privacy,
Security, Safety,
Data Protection,
Anonymity,
Fair Use, …

… Data Integration in the Life Sciences!

Holzinger, A. (2011) Weakly Structured Data in Health-Informatics: The Challenge for Human-Computer Interaction. In: Baghaei, N., Baxter, G., Dow, L. & Kimani, S. (Eds.) *Proceedings of INTERACT 2011 Workshop: Promoting and supporting healthy living by design. Lisbon, IFIP, 5-7.*



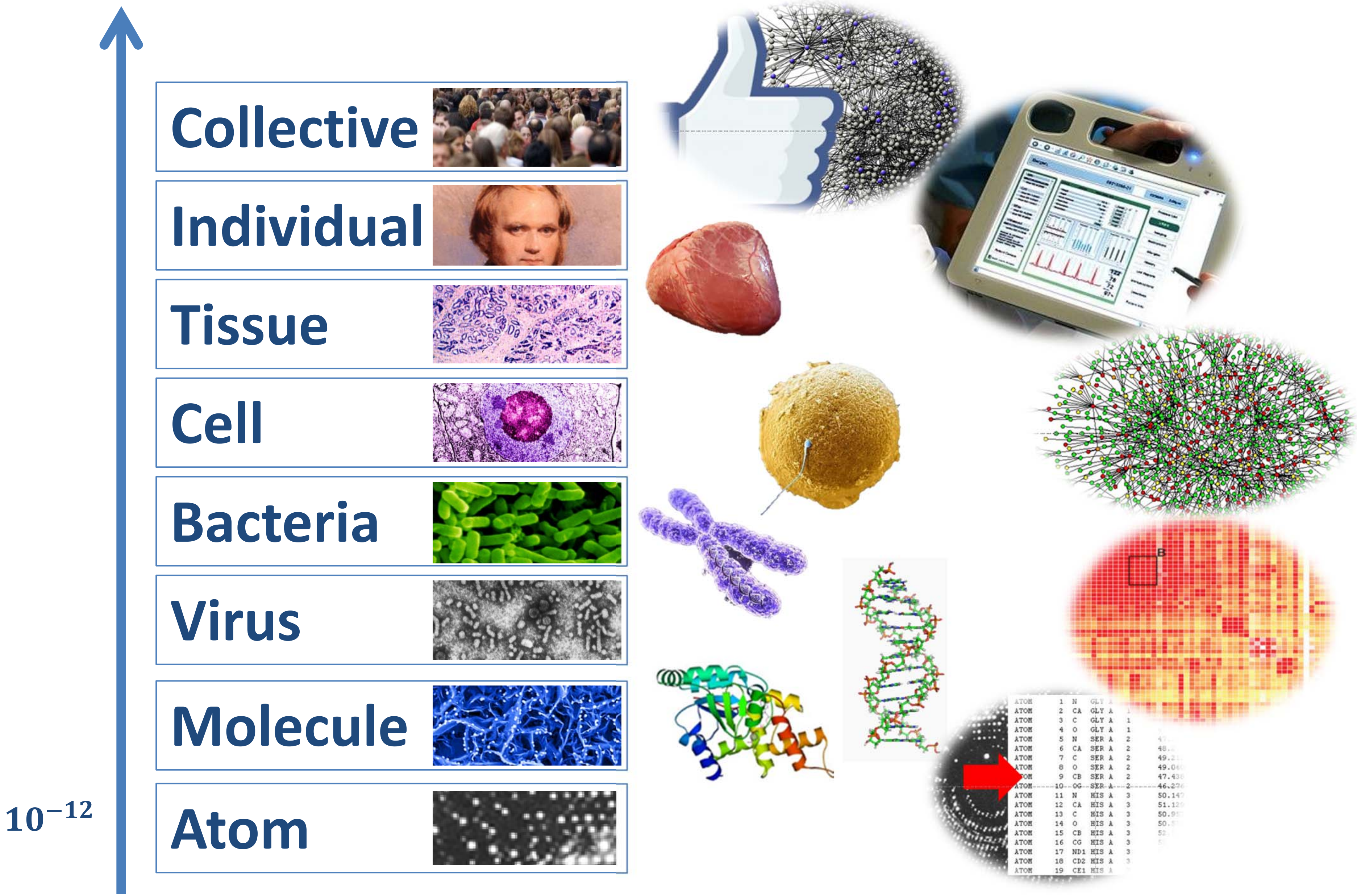
The Knowledge Discovery Process Chain



HCI, Interactive Visualization, Analytics, Decision Support	Topological Data Mining - Pattern Discovery	Sampling, Cleansing, Preprocessing, Mapping	Data Integration Data Fusion Pre-selection

Privacy, Data Protection, Data Security, Data Safety

Data sources in the Life Sciences

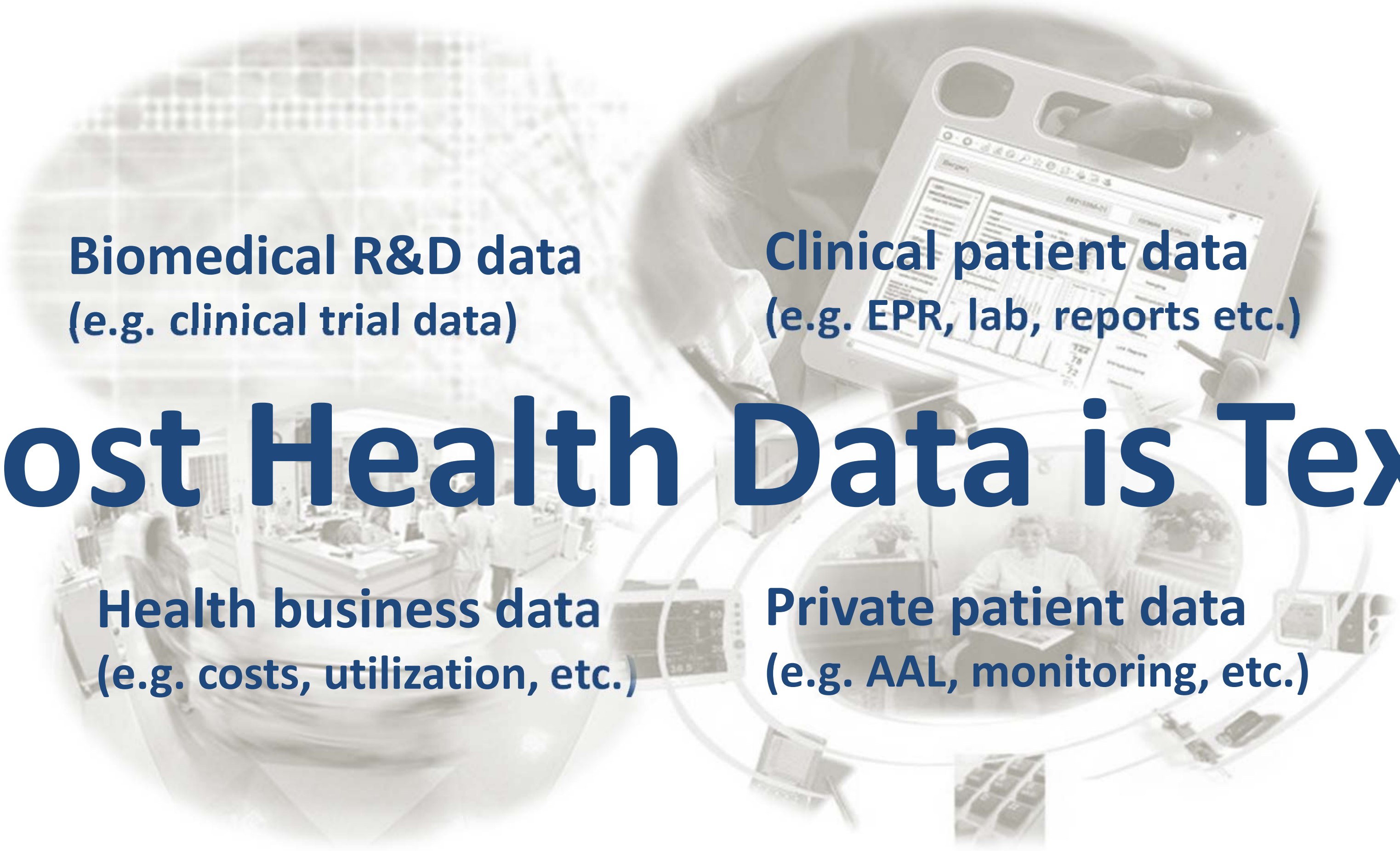


Holzinger, A. (2012). *Biomedical Informatics: Computational Sciences meets Life Sciences.*

Example: Omics-data …

Genomics	Transcriptomics	Proteomics	Metabolomics	Protein–DNA interactions	Protein–protein interactions	Fluxomics	Phenomics
Genomics (sequence annotation)	• ORF validation • Regulatory element identification[74]	• SNP effect on protein activity or abundance	• Enzyme annotation	• Binding-site identification[75]	• Functional annotation[79]	• Functional annotation	• Functional annotation[71,103] • Biomarkers[125]
	Transcriptomics (microarray, SAGE)	• Protein: transcript correlation[20]	• Enzyme annotation[109]	• Gene-regulatory networks[76]	• Functional annotation[89] • Protein complex identification[82]		• Functional annotation[102]
		Proteomics (abundance, post-translational modification)	• Enzyme annotation[99]	• Regulatory complex identification	• Differential complex formation	• Enzyme capacity	• Functional annotation
			Metabolomics (metabolite abundance)	• Metabolic-transcriptional response		• Metabolic pathway bottlenecks	• Metabolic flexibility • Metabolic engineering[109]
				Protein–DNA interactions (ChIP–chip)	• Signalling cascades[89,102]		• Dynamic network responses[84]
					Protein–protein interactions (yeast 2H, coAP–MS)		• Pathway identification activity[89]
						Fluxomics (isotopic tracing)	• Metabolic engineering
							Phenomics (phenotype arrays, RNAi screens, synthetic lethals)

Joyce, A. R. & Palsson, B. Ø. 2006. The model organism as a system: integrating 'omics' data sets. *Nature Reviews Molecular Cell Biology, 7**, 198-210.***



Biomedical R&D data
(e.g. clinical trial data)

Clinical patient data
(e.g. EPR, lab, reports etc.)

Most Health Data is Text

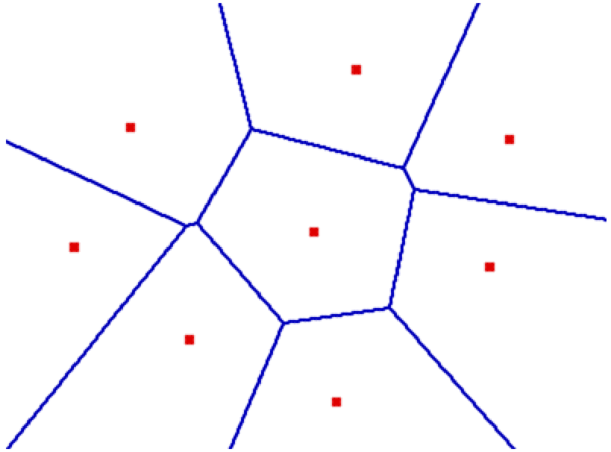
Health business data
(e.g. costs, utilization, etc.)

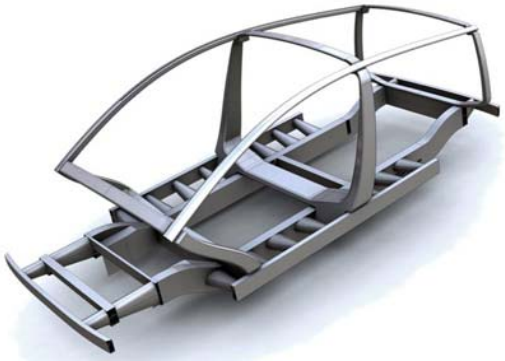
Private patient data
(e.g. AAL, monitoring, etc.)

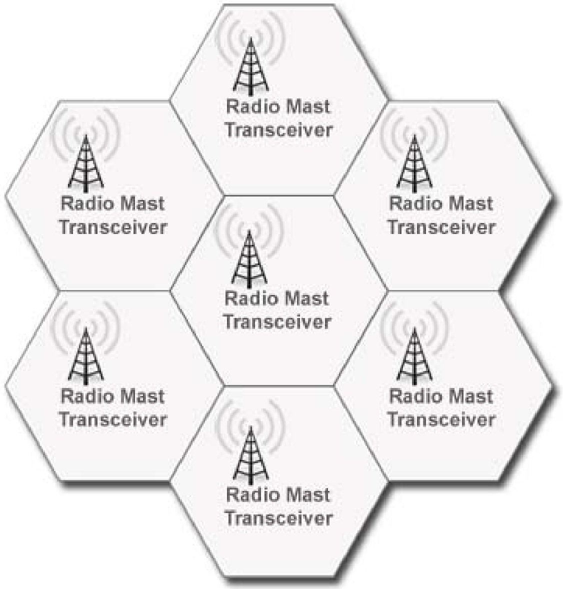
Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. & Byers, A. H. (2011) *Big data: The next frontier for innovation, competition, and productivity. Washington (DC), McKinsey Global Institute.*

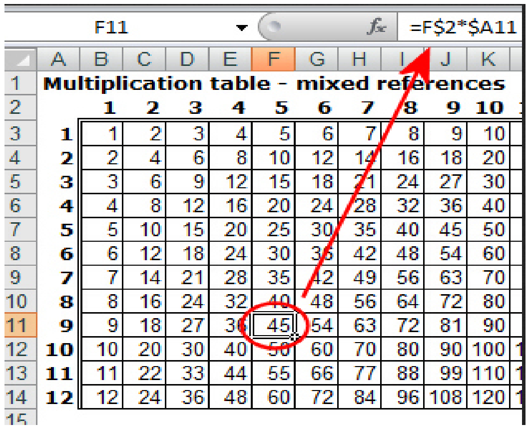
Context!

Semantic Ambiguity – Missing Context



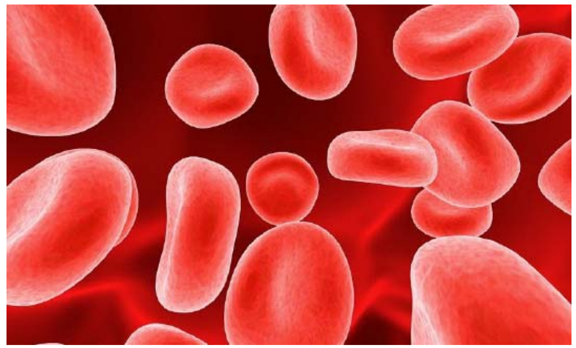




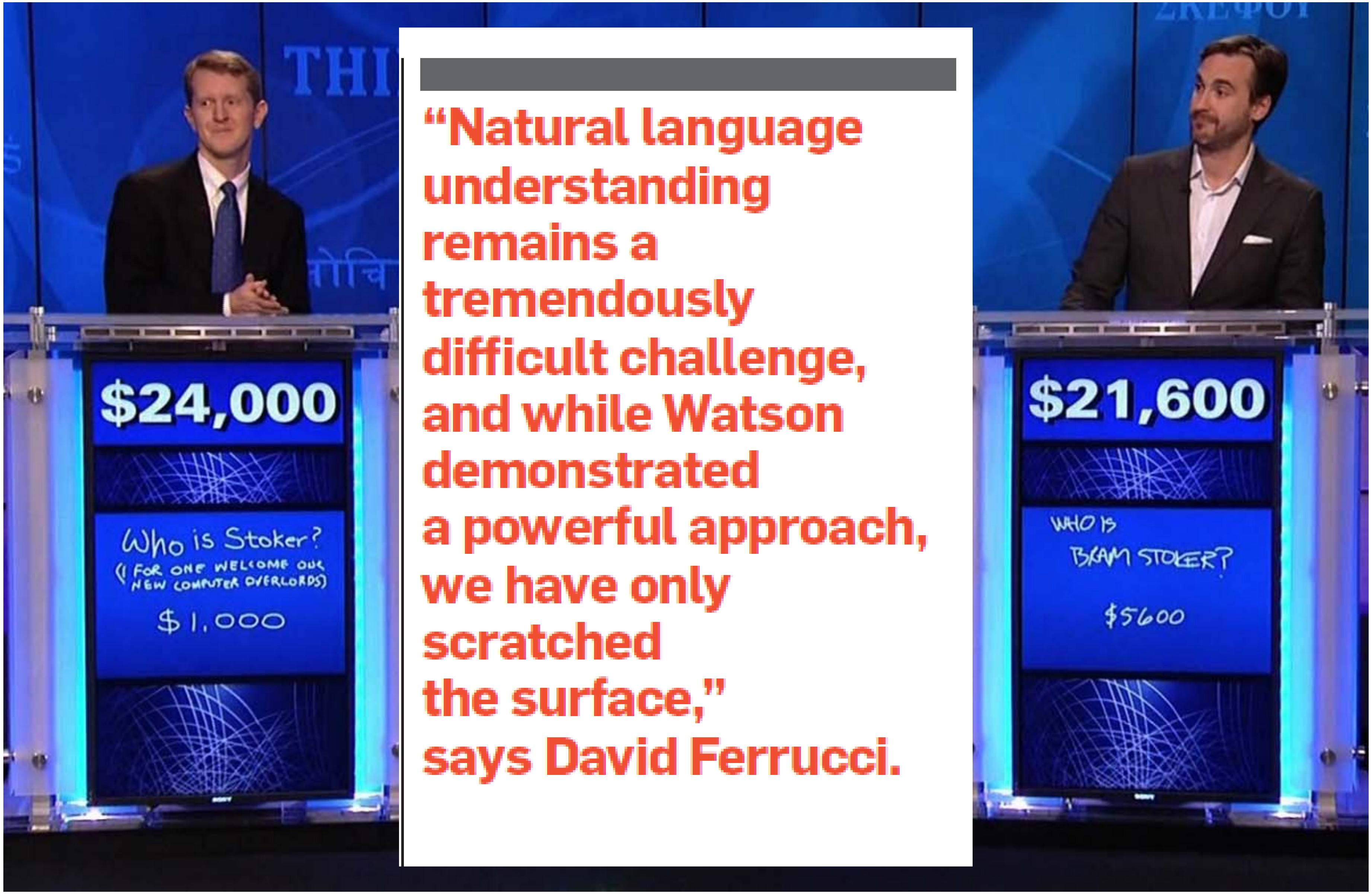


F11 =F$2*$A11

	A	B	C	D	E	F	G	H	I	J	K
1	Multiplication table - mixed references										
2		1	2	3	4	5	6	7	8	9	10
3	1	1	2	3	4	5	6	7	8	9	10
4	2	2	4	6	8	10	12	14	16	18	20
5	3	3	6	9	12	15	18	21	24	27	30
6	4	4	8	12	16	20	24	28	32	36	40
7	5	5	10	15	20	25	30	35	40	45	50
8	6	6	12	18	24	30	36	42	48	54	60
9	7	7	14	21	28	35	42	49	56	63	70
10	8	8	16	24	32	40	48	56	64	72	80
11	9	9	18	27	36	45	54	63	72	81	90
12	10	10	20	30	40	50	60	70	80	90	100
13	11	11	22	33	44	55	66	77	88	99	110
14	12	12	24	36	48	60	72	84	96	108	120
15											



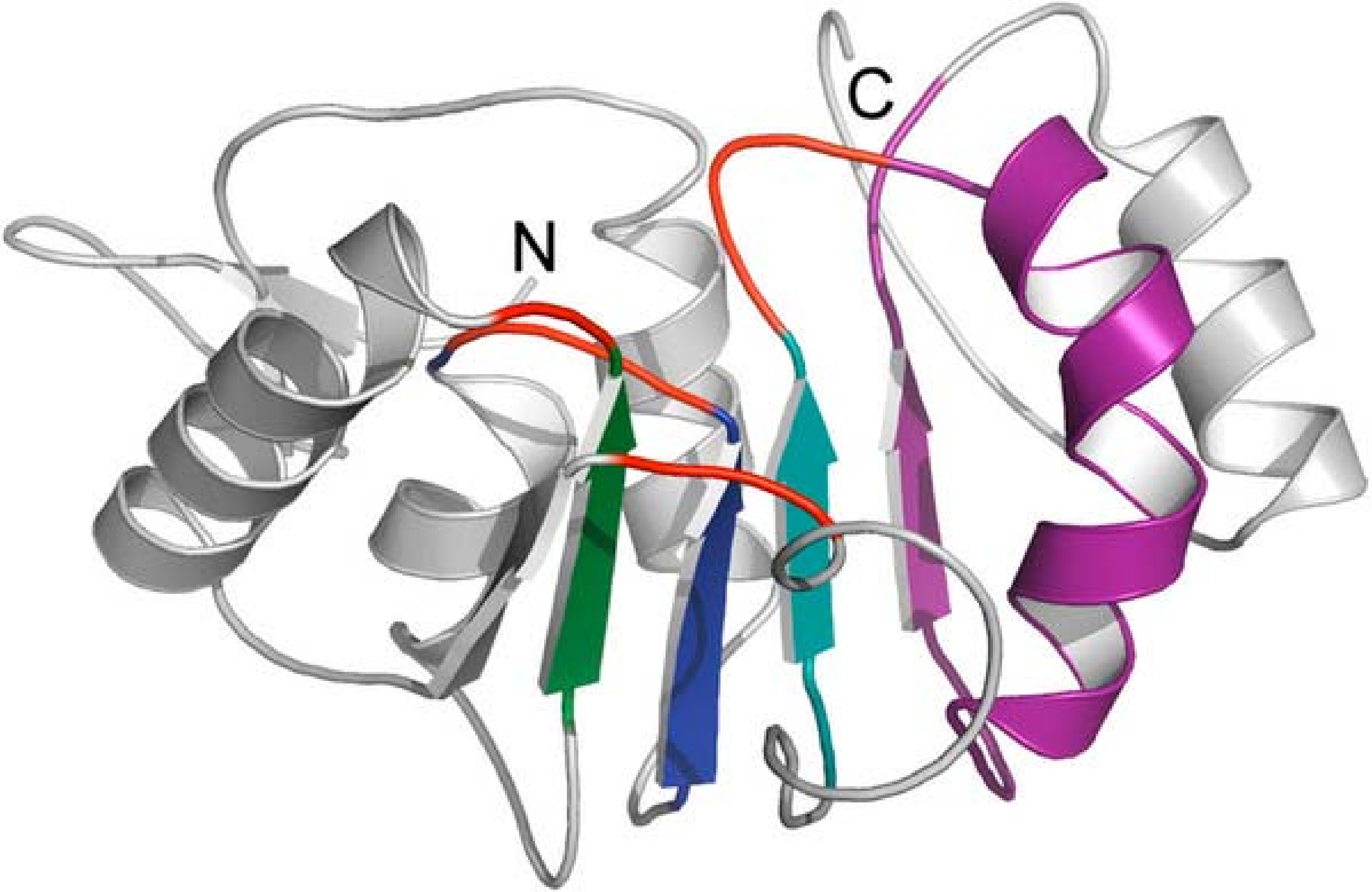
Much open work to do …



"Natural language understanding remains a tremendously difficult challenge, and while Watson demonstrated a powerful approach, we have only scratched the surface," says David Ferrucci.

Is a picture really worth a thousand words?

Example: Ribbon Diagram of a Protein Structure



Magnani, R., et al. 2010. Calmodulin methyltransferase is an evolutionarily conserved enzyme that trimethylates Lys-115 in calmodulin. *Nature Communications, 1, 43.*

angelegt am 06.05.2006/20:2
geschr. von
gedruckt am 17.11.2006/08:2
Anfo: NCHI

Radiologischer Befund

Kurzanamnese: St.p. SHT

Fragestellung: -

Untersuchung: **Thorax eine Ebene liegend**

SB

Bewegungsartefakte. Zustand nach Schädelhirntrauma.

Das Cor in der Größennorm, keine akuten Stauungszeichen.
Fragliches Infiltrat parahilär li. im UF, RW-Erguss li.

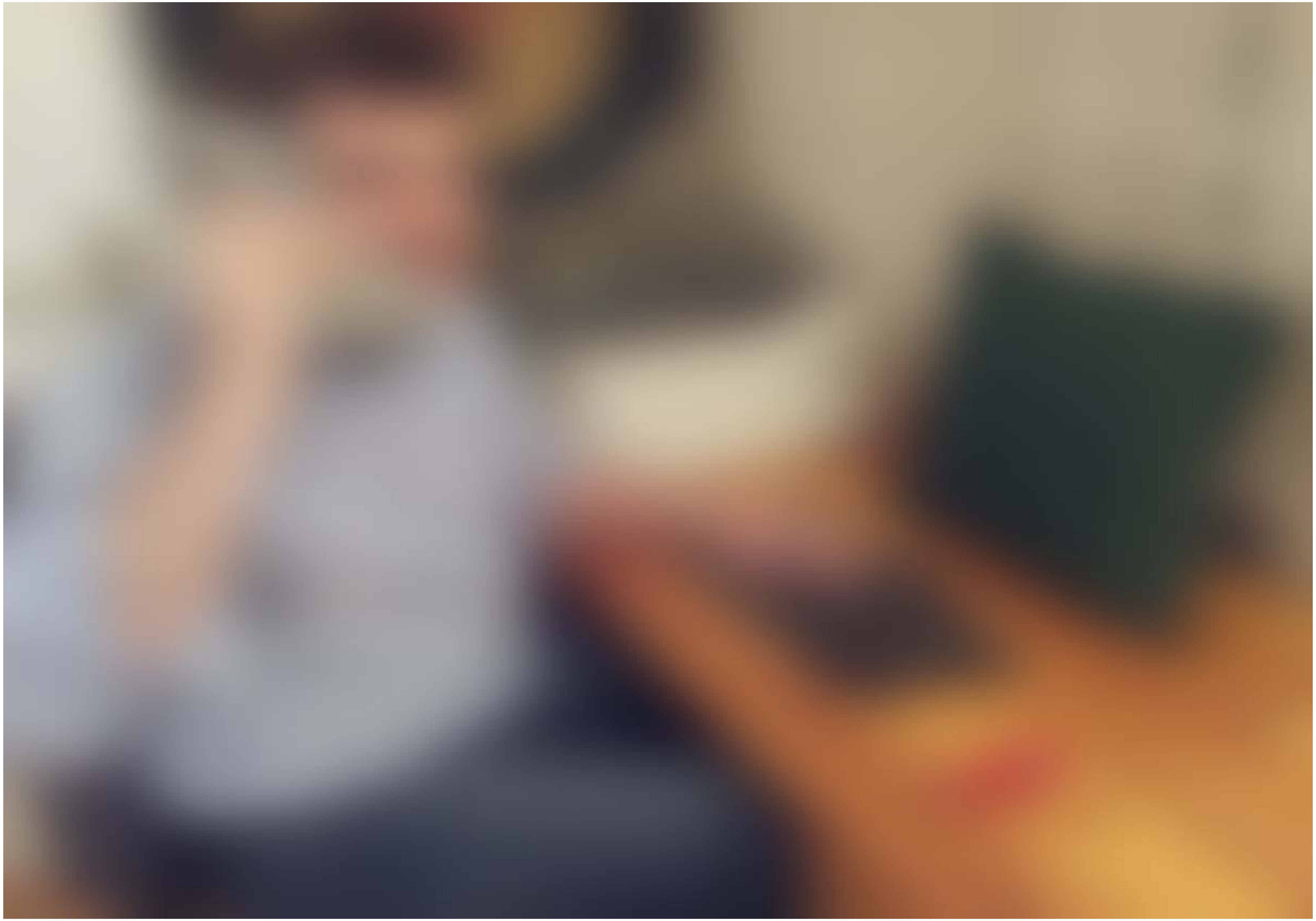
Zustand nach Anlage eines ET, die Spitze ca. 5cm cranial der Bifurkation, lieg. MS, orthot
positioniert. ZVK über re., die Spitze in Proj. auf die VCS. Kein Hinweis auf Pneumothorax
Der re. Rezessus frei.

Mit kollegialen Grüßen

**** Elektronische Freigabe durch am 09.05.2006 ****

Holzinger, A., Geierhofer, R. & Errath, M. 2007. Semantische Informationsextraktion in medizinischen Informationssystemen. *Informatik Spektrum, 30, (2), 69-78.*

A picture is worth a thousand words ...

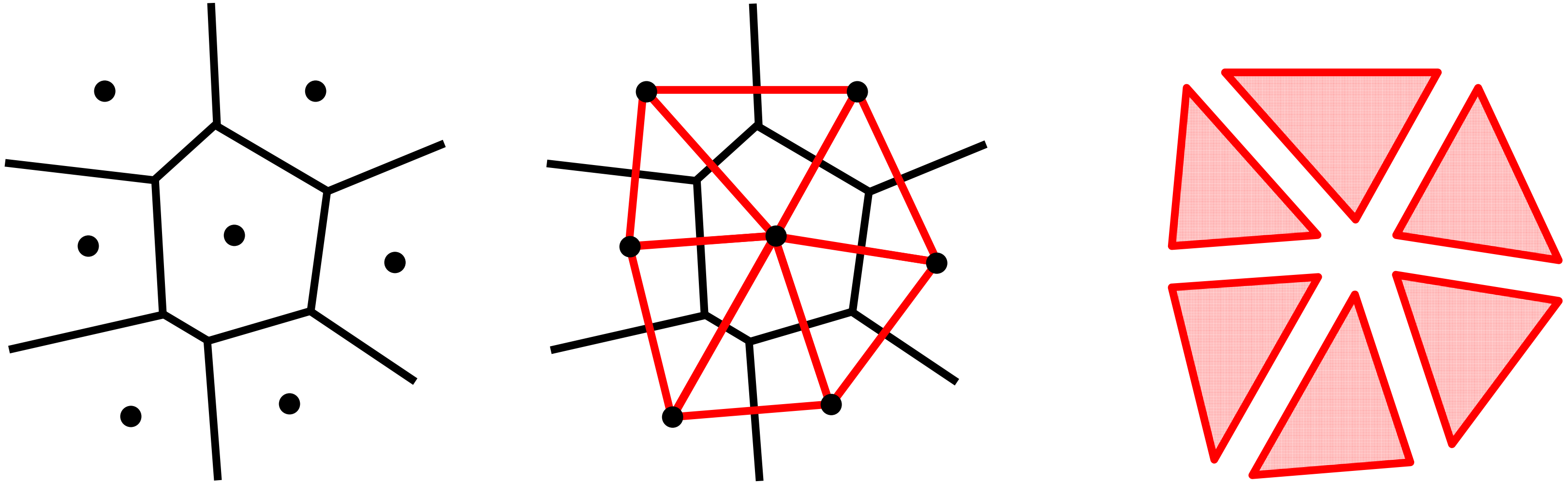


… ooops … Genius Malignus (Rene Descartes 1596 – 1650)

Boris Delaunay (1890 – 1980)

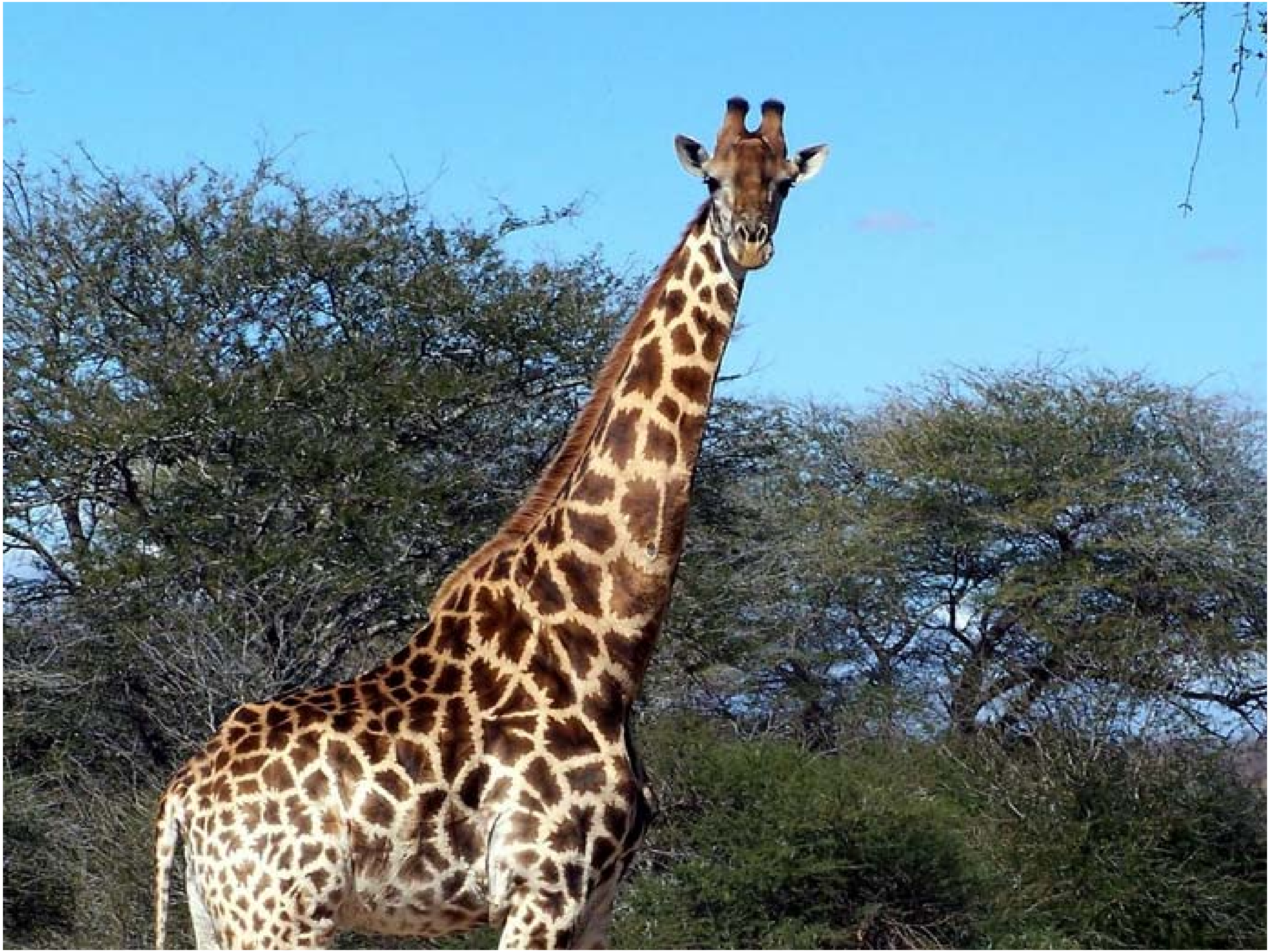
Define connectivity: a pair of points whose Voronoi cells are adjacent are connected

The connectivity graph is <u>dual</u> to the Voronoi diagram and is called Delaunay tessellation

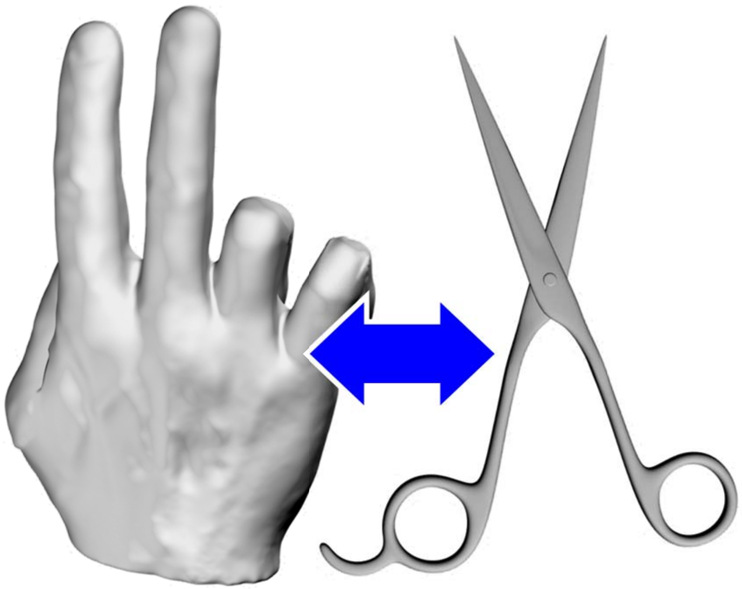


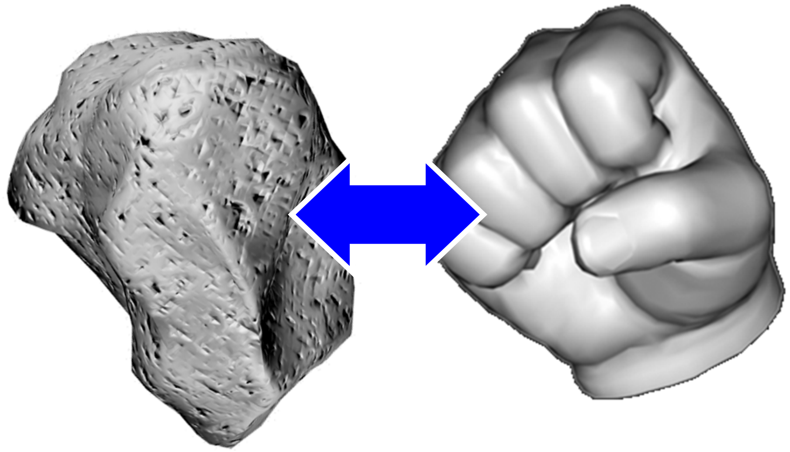
Bronstein, A. M., Bronstein, M. M. & Kimmel, R. 2008. *Numerical geometry of non-rigid shapes, New York, Springer.*


Context

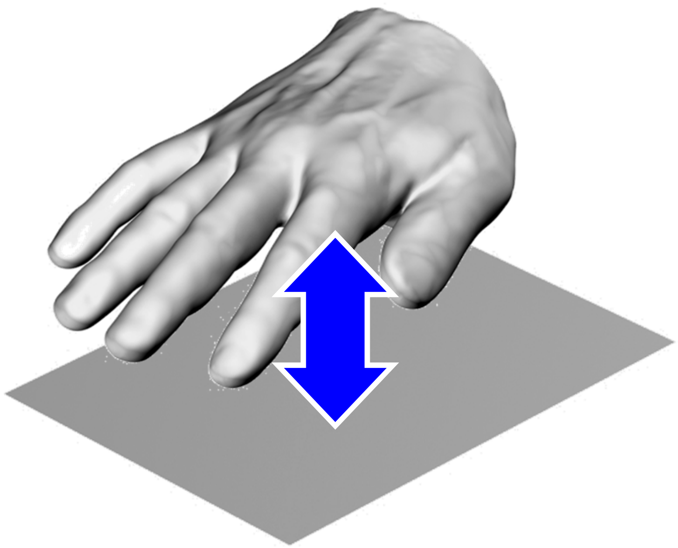


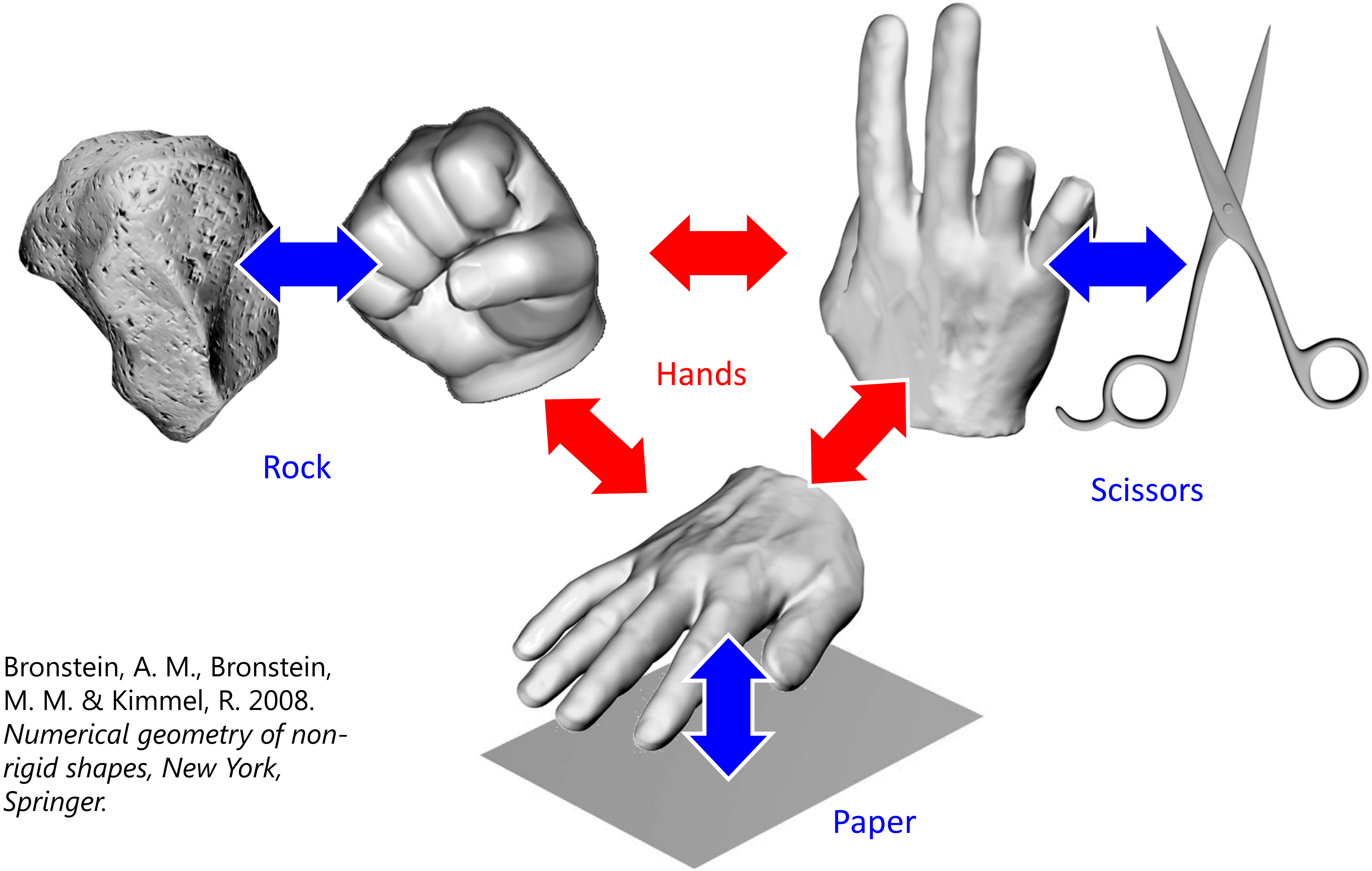
What is similarity?





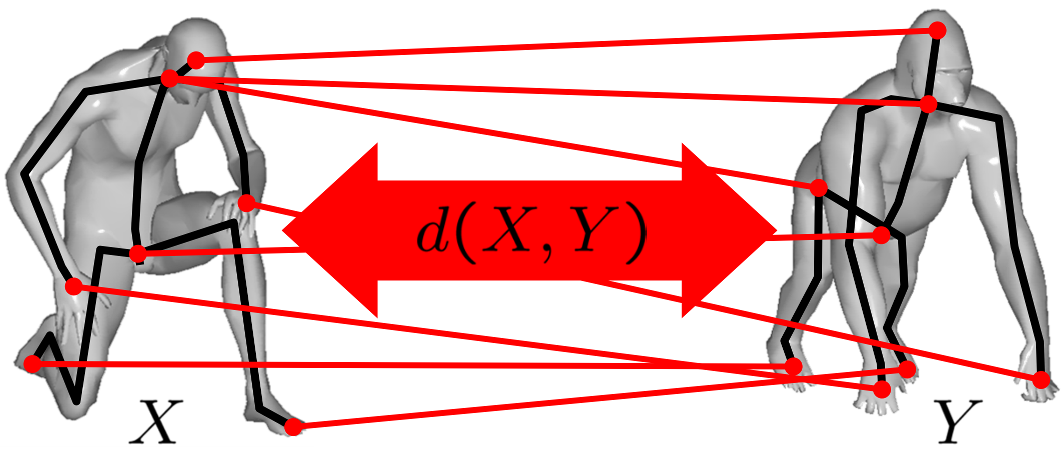
Bronstein, A. M., Bronstein, M. M. & Kimmel, R. 2008. *Numerical geometry of non-rigid shapes, New York, Springer.*





Bronstein, A. M., Bronstein, M. M. & Kimmel, R. 2008. *Numerical geometry of non-rigid shapes, New York, Springer.*

Similarity and Correspondence



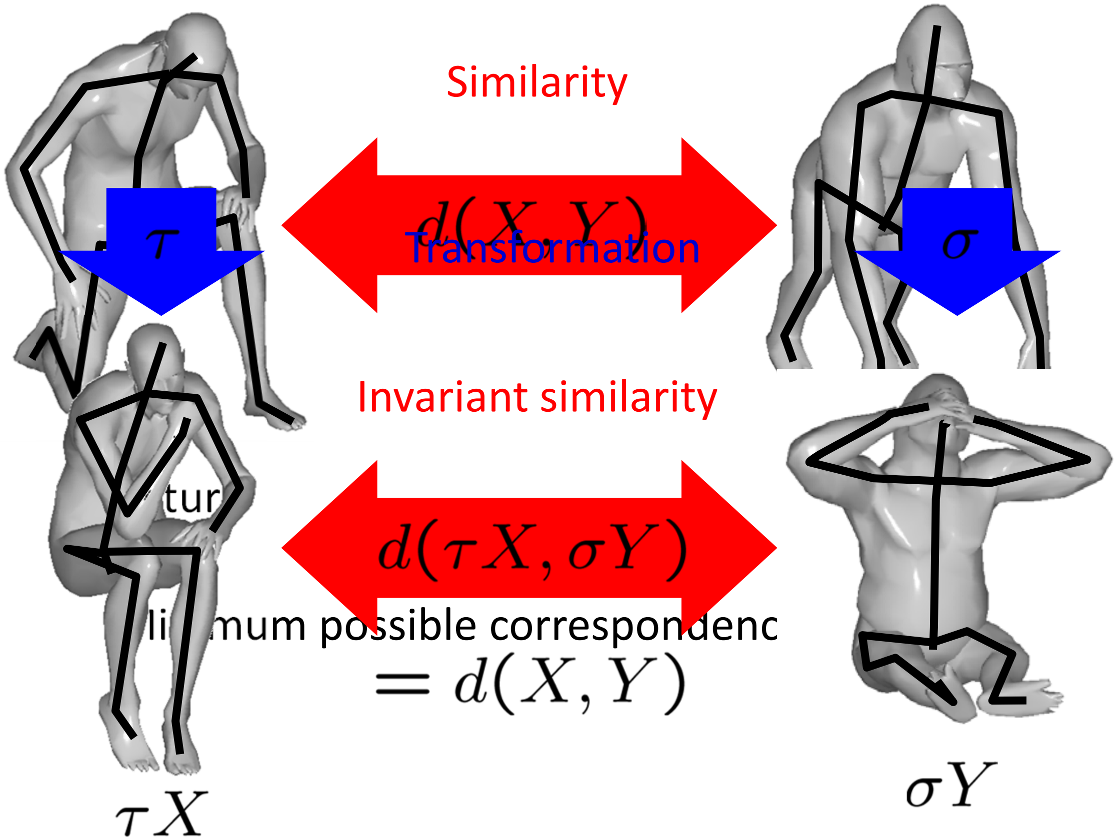
Structure

Structure

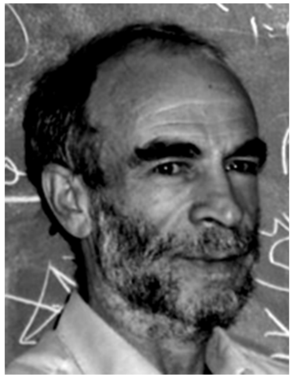
Bronstein, A. M., Bronstein, M. M. & Kimmel, R. 2008. *Numerical geometry of non-rigid shapes, New York, Springer.*

Correspondence quality = structure similarity (distortion)

Minimum possible correspondence distortion

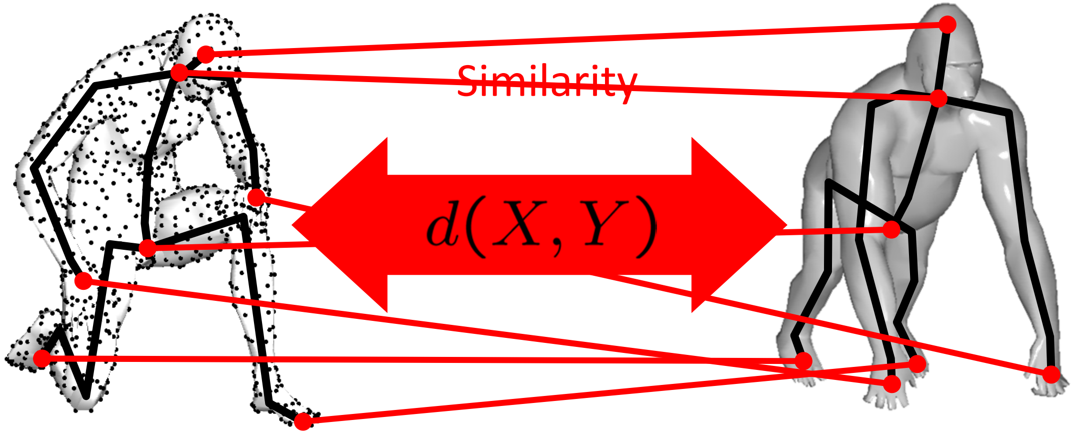

Similarity
$d(X,Y)$
Transformation
τ
σ
Invariant similarity
$d(\tau X,\sigma Y)$
tur
li mum possible correspondenc
$= d(X,Y)$
τX
σY

Gromov-Hausdorff dist: finding the opt. correspondence



Michail Gromov
(1943-)

Gromov, M. (1984) Infinite groups as geometric objects.



(X, δ_X) Metric space

Correspondence

(Y, δ_Y) Metric space

Felix Hausdorff
(1868-1942)

$$d_{\text{GH}}(X,Y) = \frac{1}{2} \min_{\mathcal{C}} \max_{\substack{(x_i,y_i)\in\mathcal{C} \\ (x_j,y_j)\in\mathcal{C}}} \left| \delta_X(x_i,x_j) - \delta_Y(y_i,y_j) \right|$$

$$\forall x_i \, \exists y_i \ \text{ s.t.} (x_i, y_i) \in \mathcal{C} \qquad \forall y_i \, \exists x_i \ \text{ s.t.} (x_i, y_i) \in \mathcal{C}$$

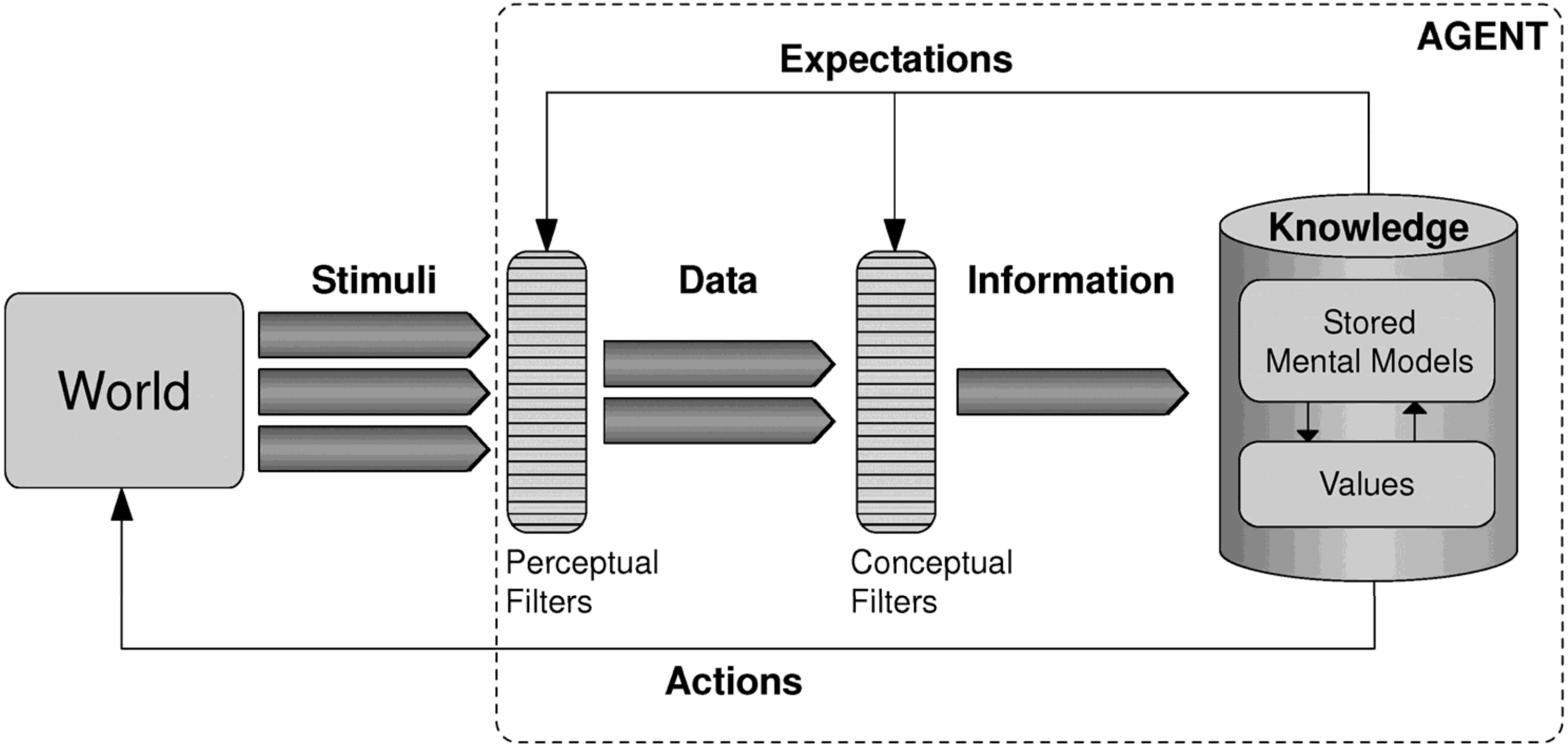
Discrete optimization over correspondences (NP hard)

Huber (1985): “What is interesting?”

- **Projection pursuit** : Find a subset of coordinates of the data which display “interesting” features. Often the selection of the subset of coordinates is manual, but there are automated algorithms which can find these subsets automatically also. Finally one has to inspect each projection and decide if its “interesting”.

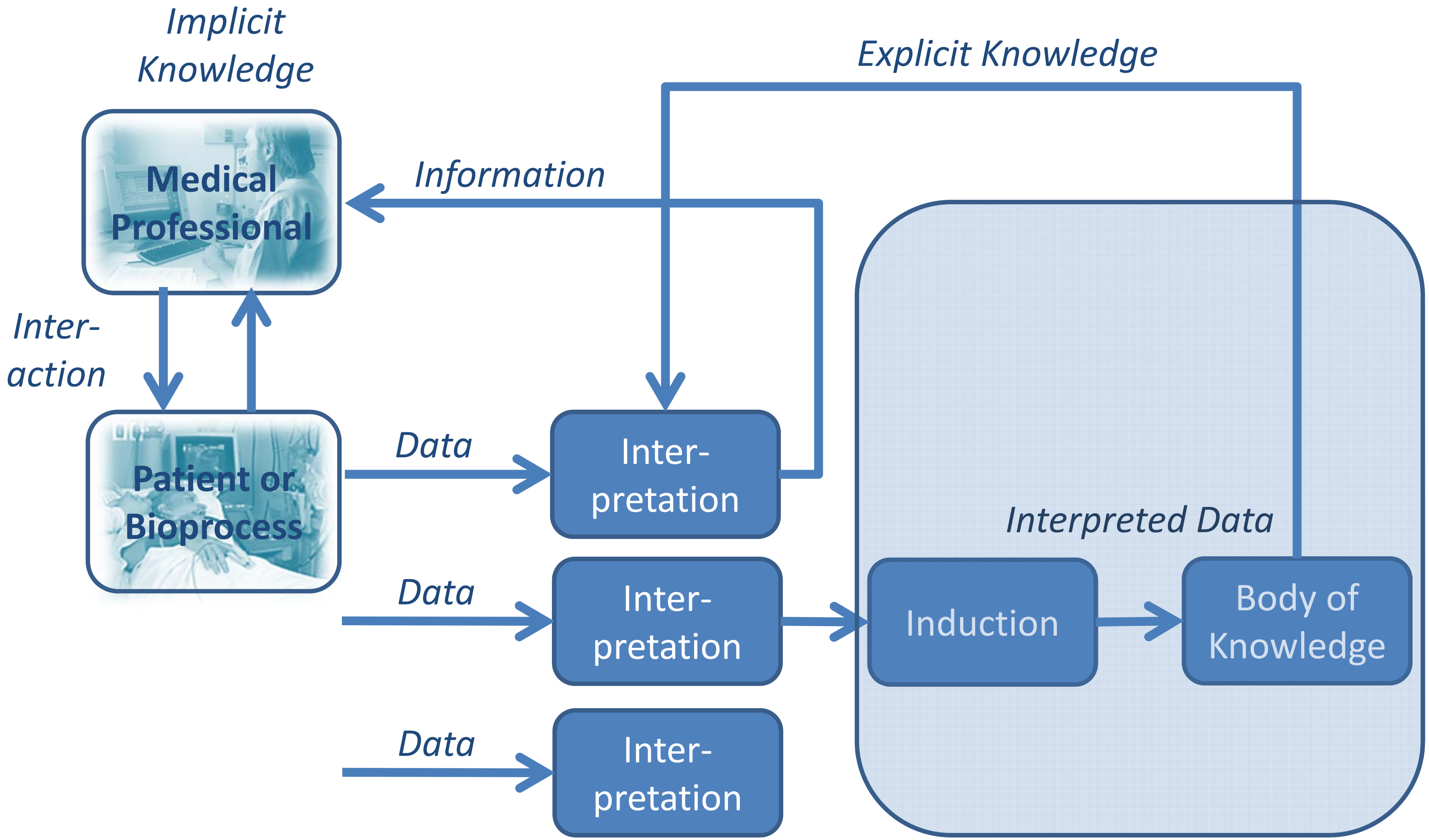
Huber P.J.: Projection pursuit. *Ann. Statist. 13*, 2 (1985), 435-525.

Our definition of Knowledge

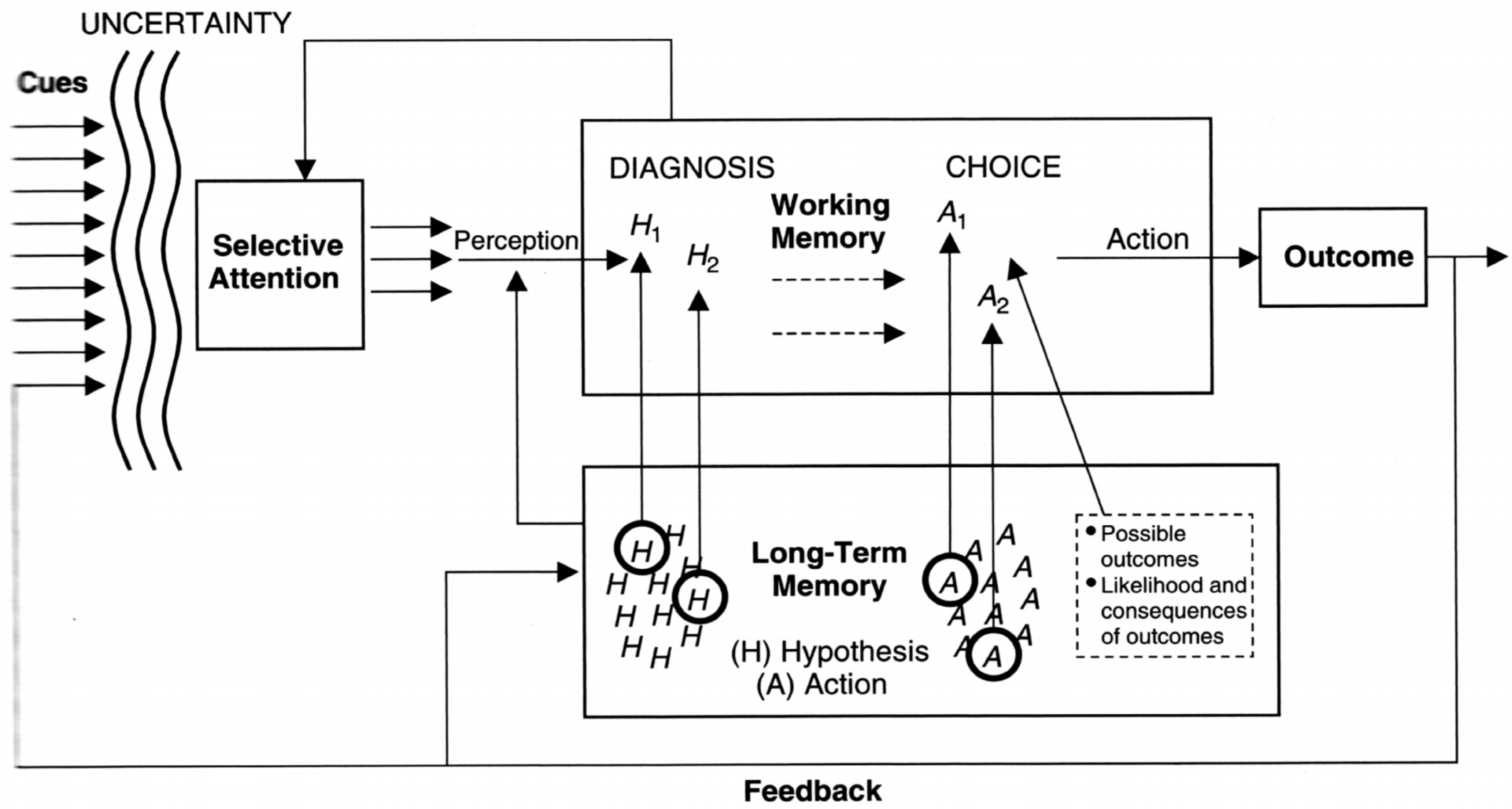


Boisot, M. & Canals, A. 2004. Data, information and knowledge: have we got it right? *Journal of Evolutionary Economics, 14, (1), 43-67.*

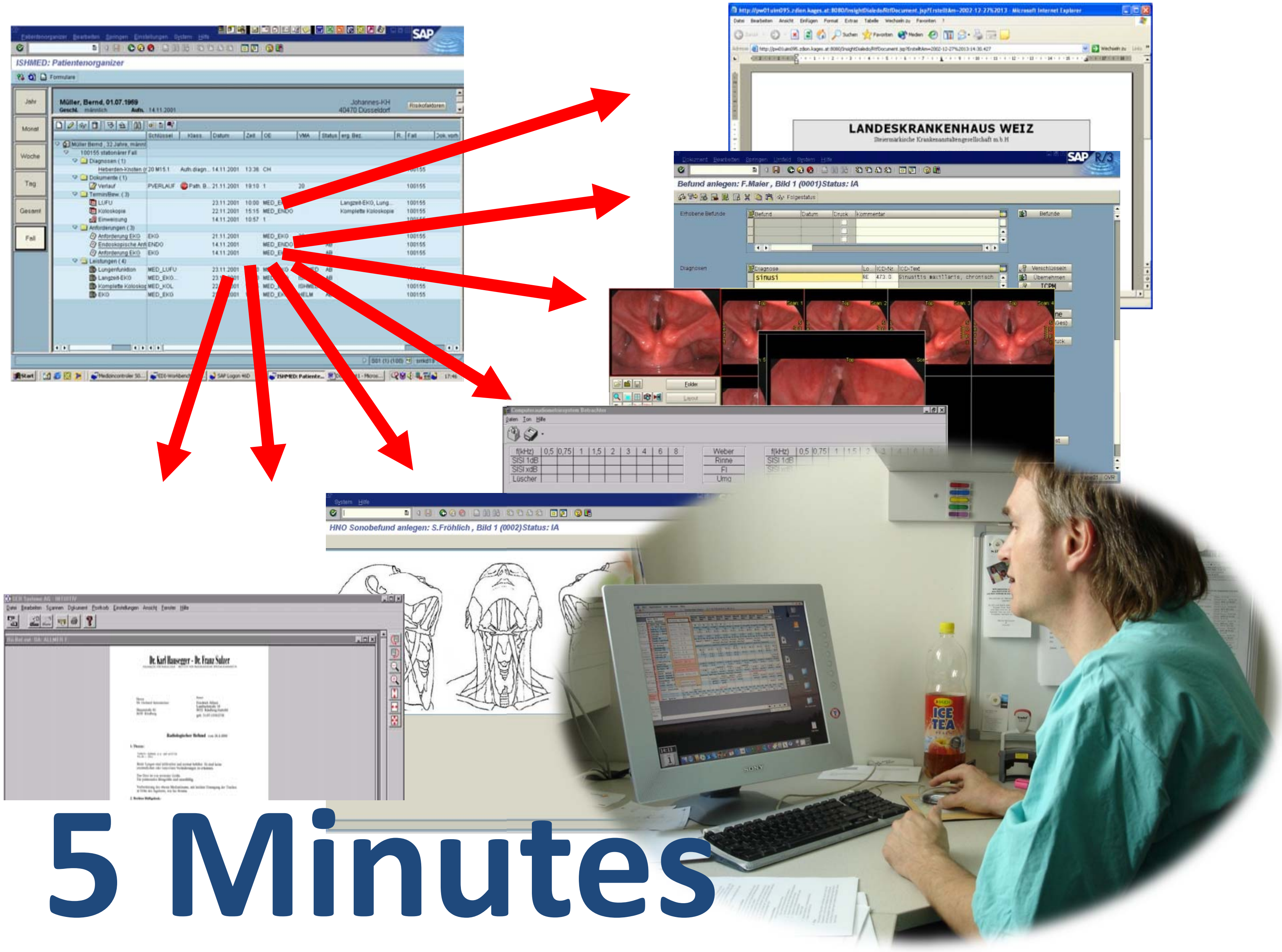
Implicit vs. Explicit Knowledge



Bemmel, J. H. v. & Musen, M. A. (1997) *Handbook of Medical Informatics. Heidelberg, Springer.*



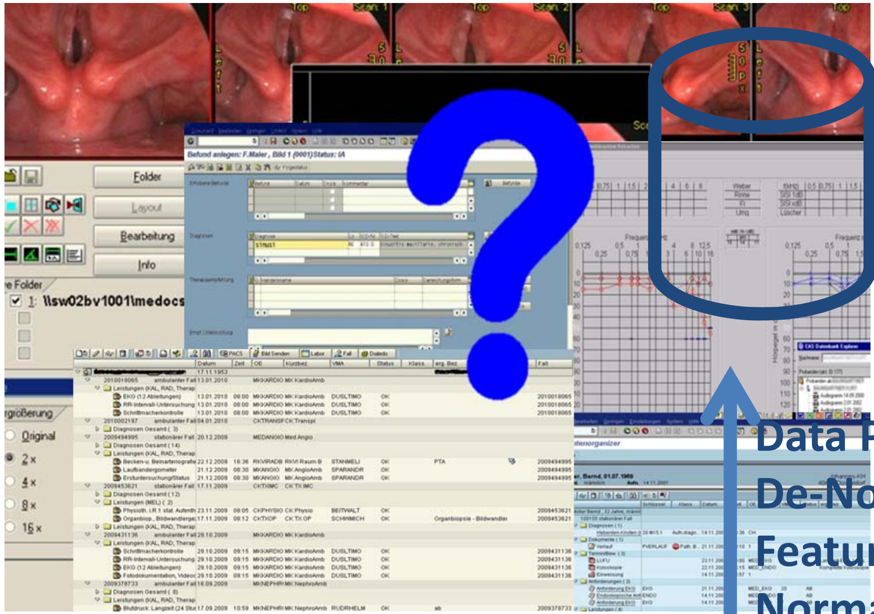
Wickens, C. D. (1984) *Engineering psychology and human performance. Columbus (OH), Charles Merrill.*


ISHMED: Patientenorganizer
LANDESKRANKENHAUS WEIZ
Befund anlegen: F.Maier , Bild 1 (0001)Status: IA
HNO Sonobefund anlegen: S.Fröhlich , Bild 1 (0002)Status: IA
Dr. Karl Hausegger - Dr. Franz Sulzer

5 Minutes

Selected
Target Data

Visualization,
Validation
Etc.



Data Integration, Data Fusion
Privacy, Security, Safety
Data Protection
Anonymization-> Open Data

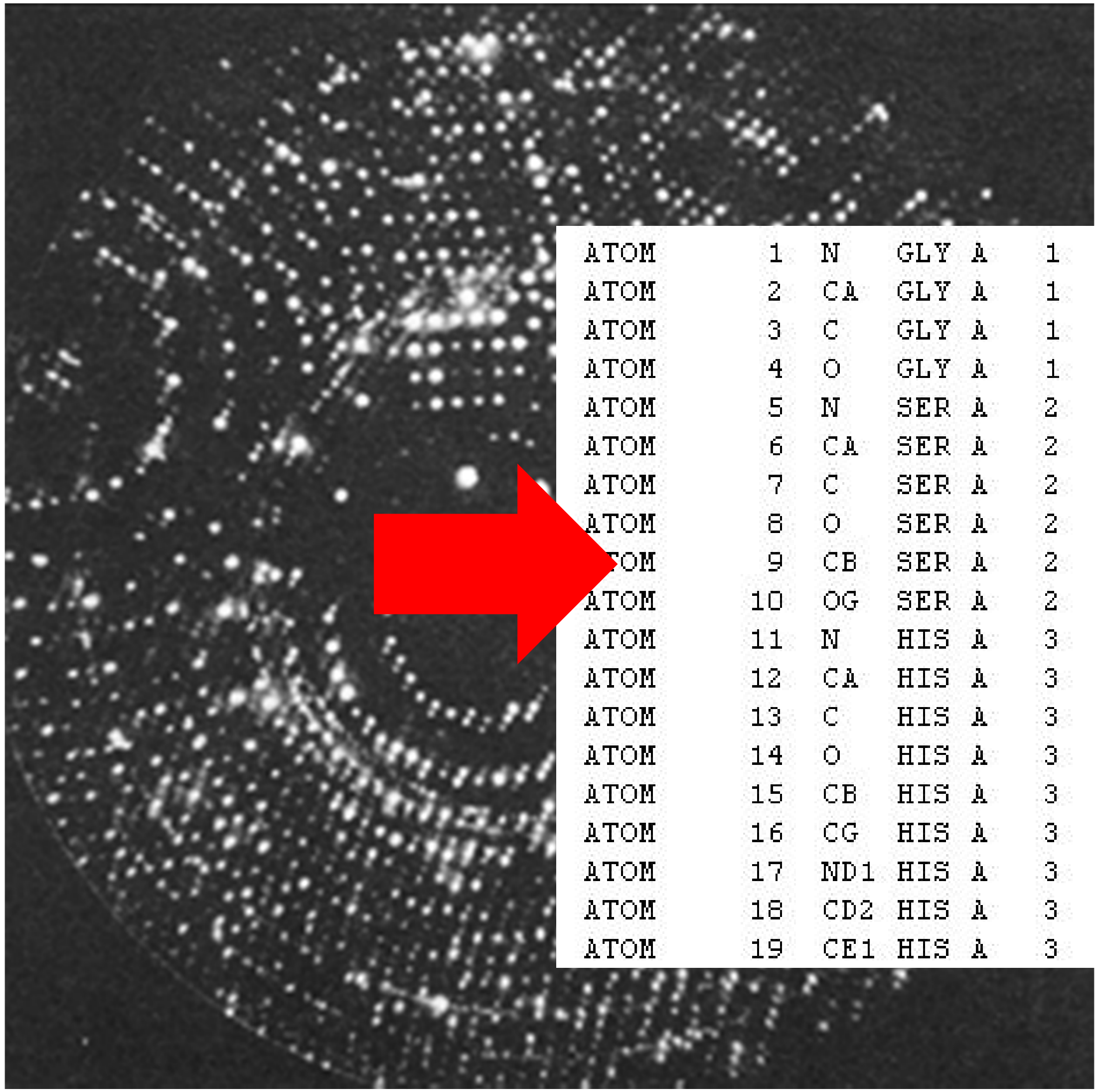
Data Preprocessing
De-Noising
Feature extraction
Normalization etc.

Classification
Clustering etc.

INTERACTION ?

Holzinger, A. & Zupan, M. 2013. KNODWAT: A scientific framework application for testing knowledge discovery methods for the biomedical domain. *BMC Bioinformatics, 14, (1), 191.*

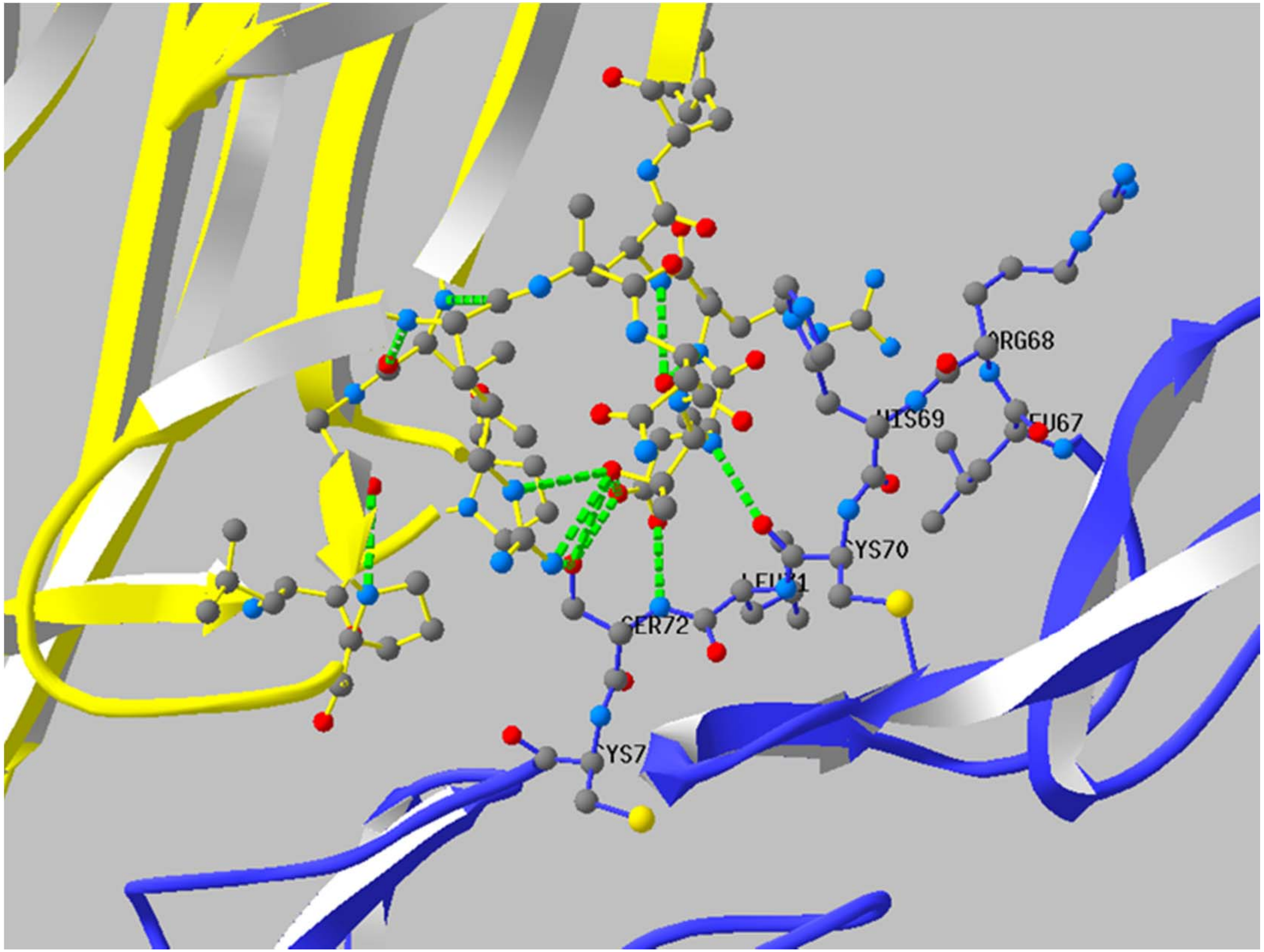
Medical People are NOT interested in data …



ATOM	1	N	GLY	A	1	44.842	51.034	101.284	0.01	27.20
ATOM	2	CA	GLY	A	1	45.640	50.230	100.389	0.01	26.99
ATOM	3	C	GLY	A	1	46.692	49.648	101.308	0.01	26.80
ATOM	4	O	GLY	A	1	46.895	50.222	102.381	0.01	26.91
ATOM	5	N	SER	A	2	47.283	48.516	100.951	1.00	26.26
ATOM	6	CA	SER	A	2	48.277	47.866	101.761	1.00	26.17
ATOM	7	C	SER	A	2	49.212	47.031	100.845	1.00	24.21
ATOM	8	O	SER	A	2	49.060	47.195	99.630	1.00	19.77
ATOM	9	CB	SER	A	2	47.438	47.091	102.800	1.00	26.31
ATOM	10	OG	SER	A	2	46.276	46.356	102.404	1.00	27.99
ATOM	11	N	HIS	A	3	50.147	46.186	101.370	1.00	23.93
ATOM	12	CA	HIS	A	3	51.129	45.389	100.609	1.00	21.44
ATOM	13	C	HIS	A	3	50.953	43.905	100.849	1.00	20.32
ATOM	14	O	HIS	A	3	50.530	43.595	101.950	1.00	22.00
ATOM	15	CB	HIS	A	3	52.555	45.674	100.990	1.00	19.69
ATOM	16	CG	HIS	A	3	52.940	47.090	100.611	1.00	21.44
ATOM	17	ND1	HIS	A	3	53.371	47.470	99.422	1.00	20.87
ATOM	18	CD2	HIS	A	3	52.956	48.175	101.433	1.00	21.69
ATOM	19	CE1	HIS	A	3	53.676	48.730	99.476	1.00	20.57

Wiltgen, M. & Holzinger, A. (2005) Visualization in Bioinformatics: Protein Structures with Physicochemical and Biological Annotations. In: *Central European Multimedia and Virtual Reality Conference. Prague, Czech Technical University (CTU), 69-74*

… they are interested in knowledge gained from this data



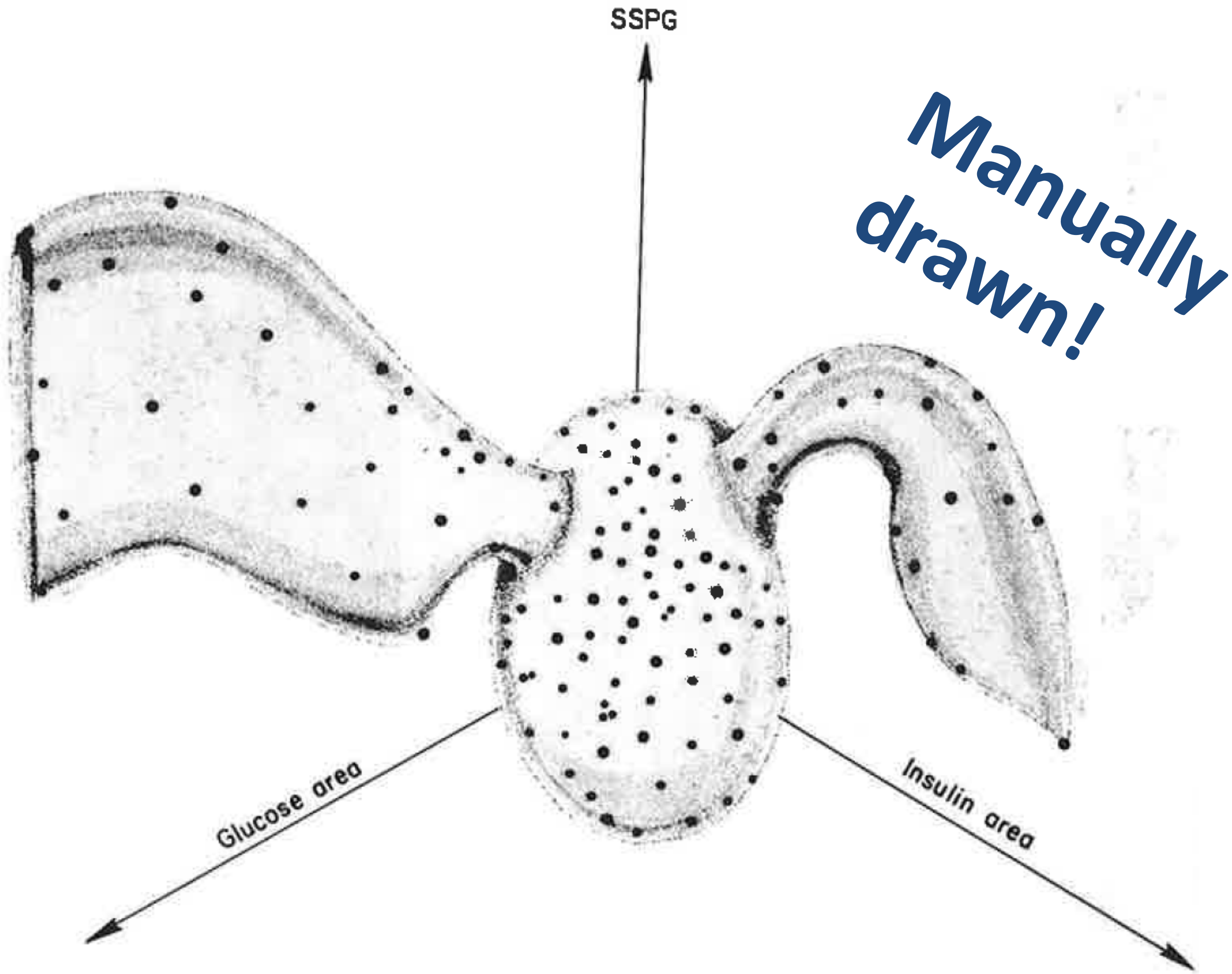
Wiltgen, M., Holzinger, A. & Tilz, G. P. (2007) Interactive Analysis and Visualization of Macromolecular Interfaces Between Proteins. In: *Lecture Notes in Computer Science (LNCS 4799). Berlin, Heidelberg, New York, Springer, 199-212.*

Miller-Reavens Diabetes Study from 1979 (1/2)

- 145 diabetes patients
- 6 dimensional data set:
  - 1) age,
  - 2) relative weight,
  - 3) fasting plasma glucose,
  - 4) area under the plasma glucose curve for the three hour glucose tolerance test (OGTT),
  - 5) area under the plasma insulin curve for the OGTT,
  - 6) steady state plasma glucose response.
- Method: **Projection Pursuit (PP)**
- Result: $\mathbb{R}^6 \longrightarrow \mathbb{R}^3$

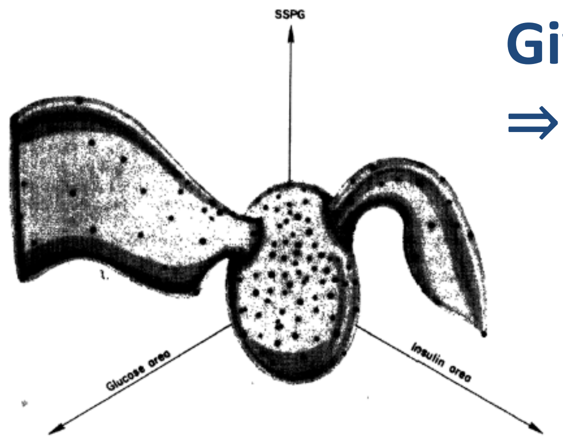
Reaven, G. & Miller, R. (1979) An attempt to define the nature of chemical diabetes using a multidimensional analysis. *Diabetologia, 16, **1, 17-24.***

Miller-Reaves Diabetes Study (2/2)



Reaven, G. & Miller, R. (1979) An attempt to define the nature of chemical diabetes using a multidimensional analysis. *Diabetologia, 16,* ***1, 17-24.***

Mapping the data from $\mathbb{R}^6$ to $\mathbb{R}^2$

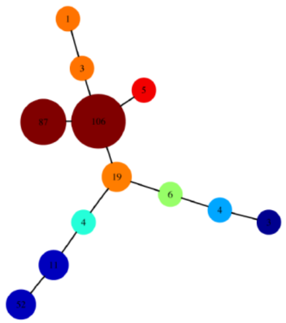


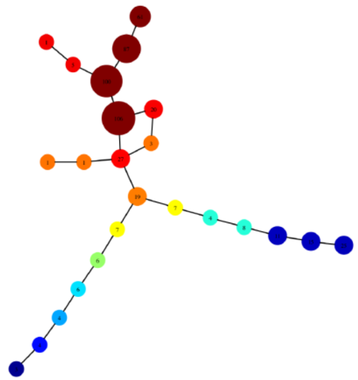
Given a point cloud data set X and a covering U
$\Rightarrow$ find the simplicial complex!

$$f : \boldsymbol{X} \to \mathbb{R}$$

$$f : \boldsymbol{X} \to \boldsymbol{Z}$$

$$\boldsymbol{\mathcal{U}} = \{\boldsymbol{U}_\alpha\}_{\alpha \in A}$$

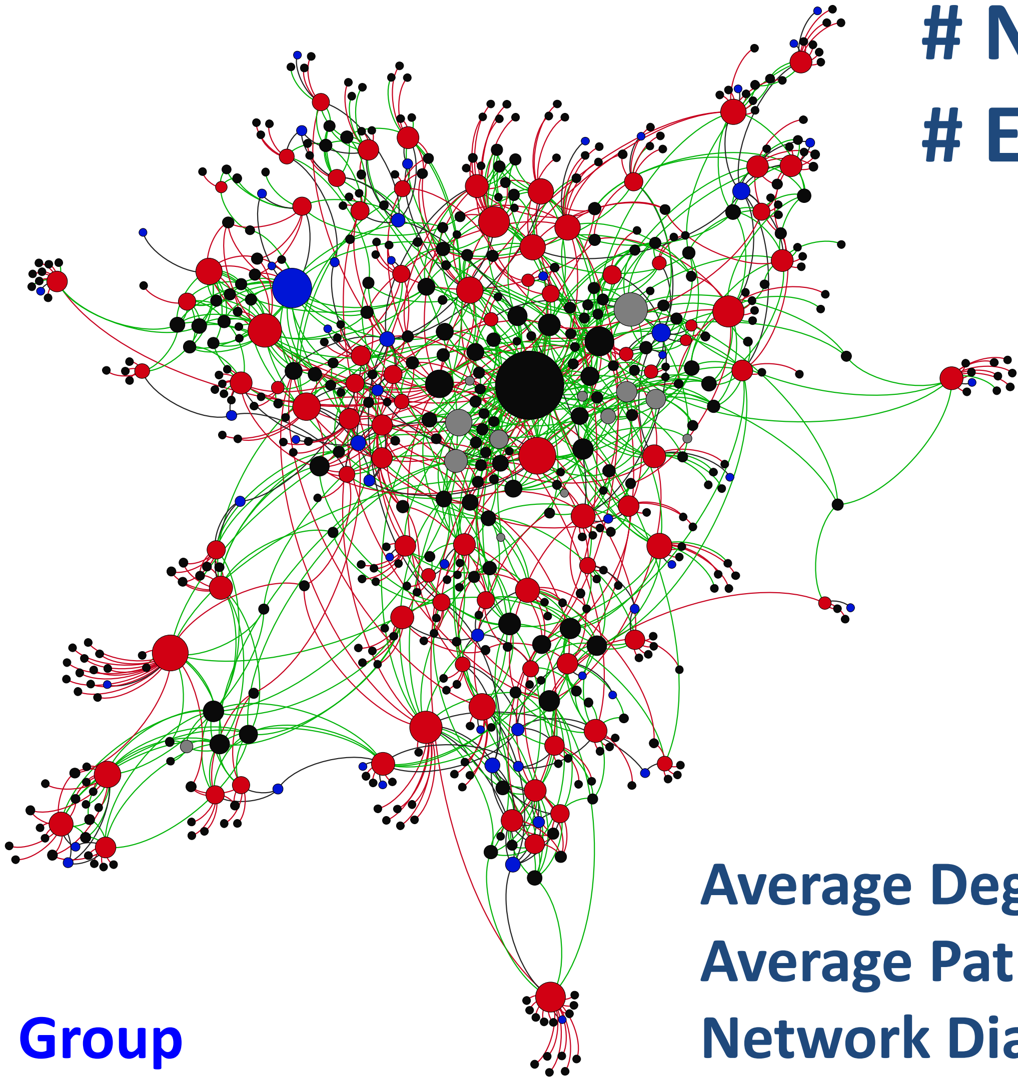




$$f_\varepsilon(x) = C_\varepsilon \sum_y \exp\left(\frac{-d(x,y)^2}{\varepsilon}\right)$$

Singh, G., Mémoli, F. & Carlsson, G. (2007). *Topological methods for the analysis of high dimensional data sets and 3D object recognition. Eurographics Symposium on Point-Based Graphics, Euro Graphics Society, 91-100. Density Estimation*

Example for a Medical Knowledge Space



Nodes: 641

Edges: 1250

Agent

Condition

Pharmacological Group

Other Documents

Average Degree: 3,888

Average Path Length: 4.683

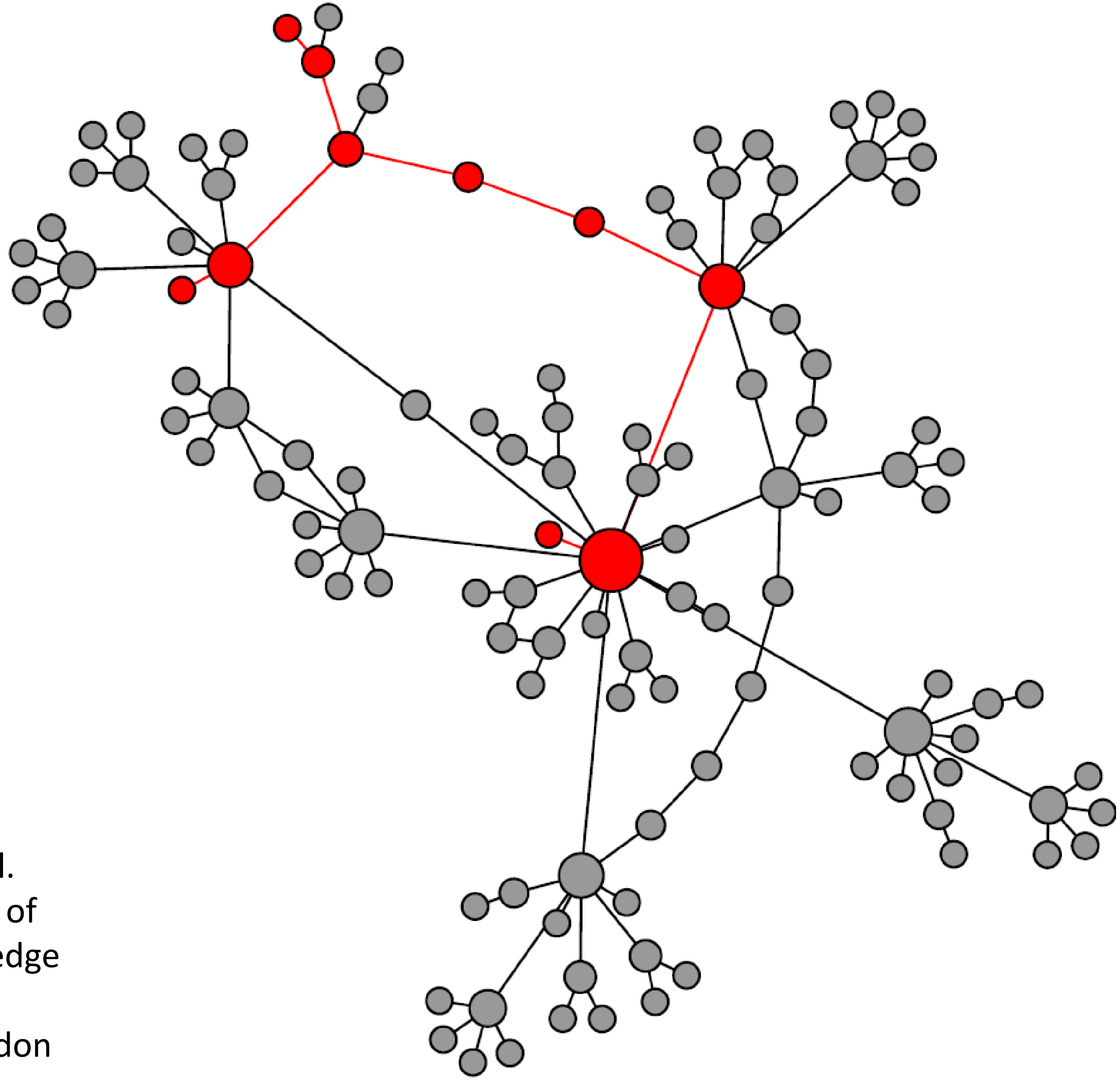
Network Diameter: 9

Holzinger, A., et al. 2013. Constraints of List-based Knowledge Interaction. In: Medicine 2.0 London

Medical Details of the Graph

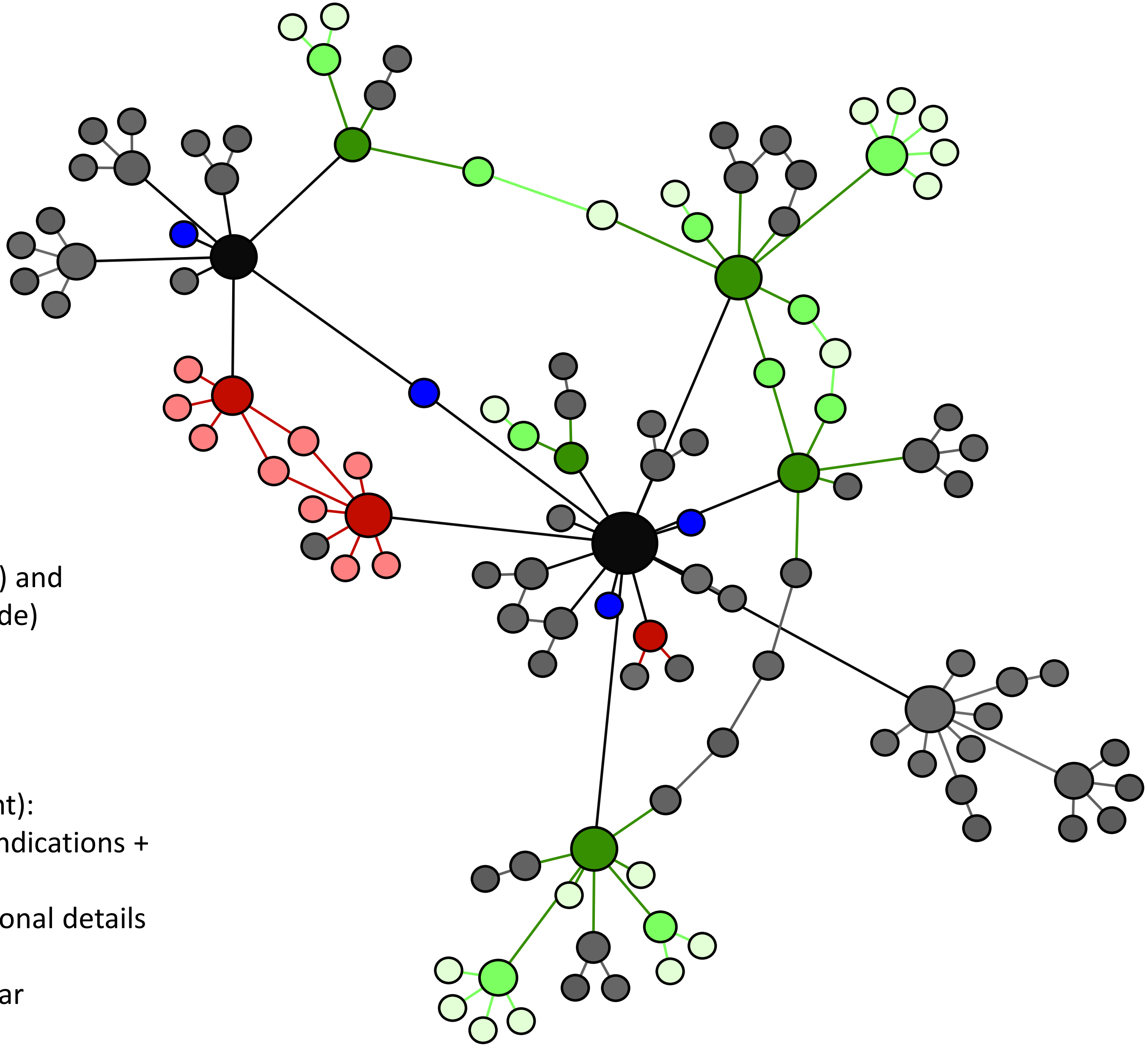
- **Nodes**
  - **drugs**
  - clinical guidelines
  - **patient conditions** (indication, contraindication)
  - **pharmacological groups**
  - tables and calculations of medical scores
  - algorithms and other medical documents
- **Edges**: 3 crucial types of relations inducing medical relevance between two active substances
  - pharmacological groups
  - indications
  - contra-indications

Example for the shortest path



Holzinger, A., et al.
2013. Constraints of
List-based Knowledge
Interaction. In:
Medicine 2.0 London

Example for finding related structures



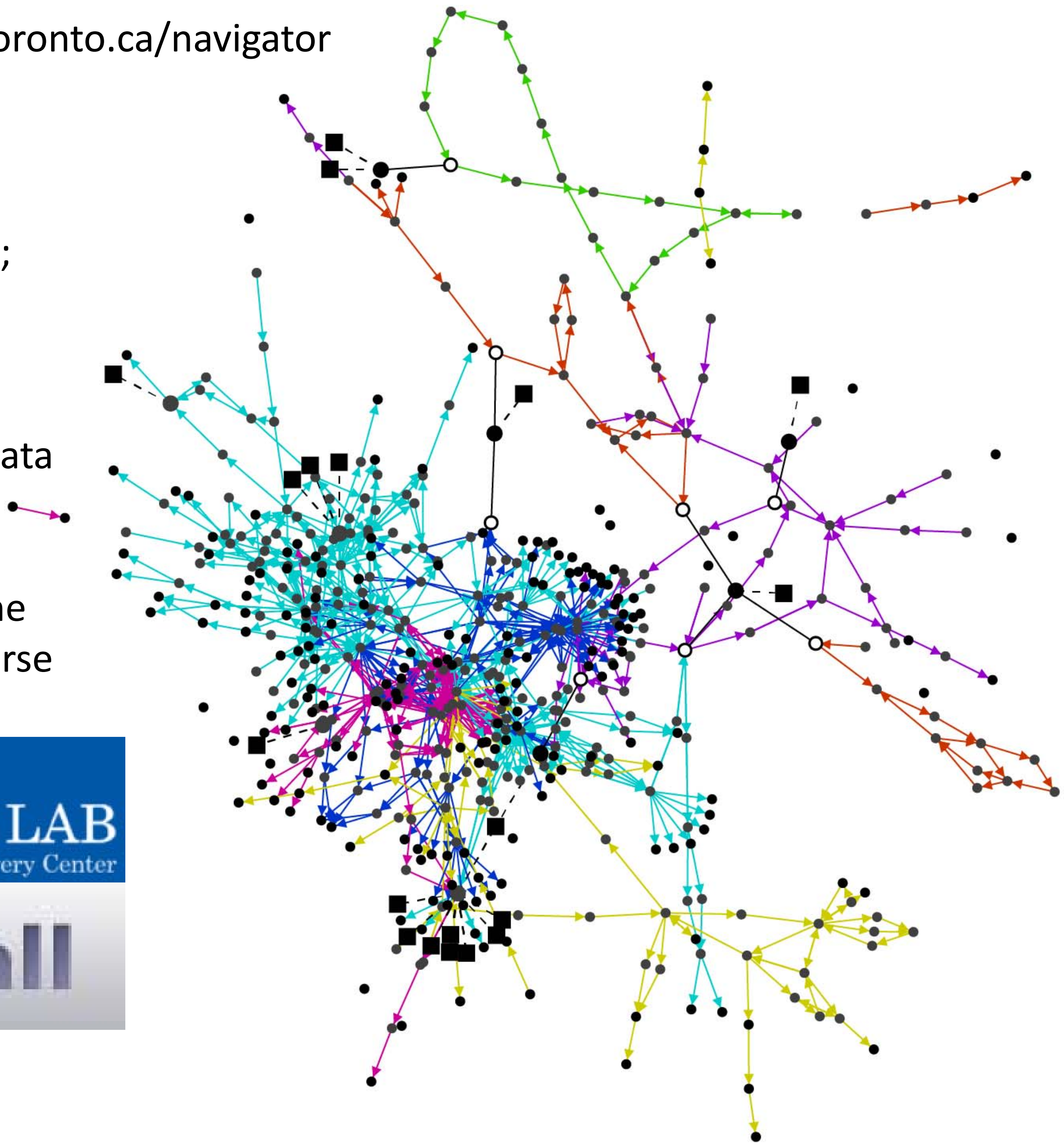
Relationship between
Adrenaline (center black node) and
Dobutamine (top left black node)
Blue: Pharmacological Group
Dark red: Contraindication;
Light red: Condition

Green nodes (from dark to light):
1. Application (one ore more indications + corresponding dosages)
2. Single indication with additional details (e. g. “VF after 3rd Shock”)
3. Condition (e.g. VF, Ventricular Fibrillation)

Interactive Visual Data Mining to connect Drugs and MB

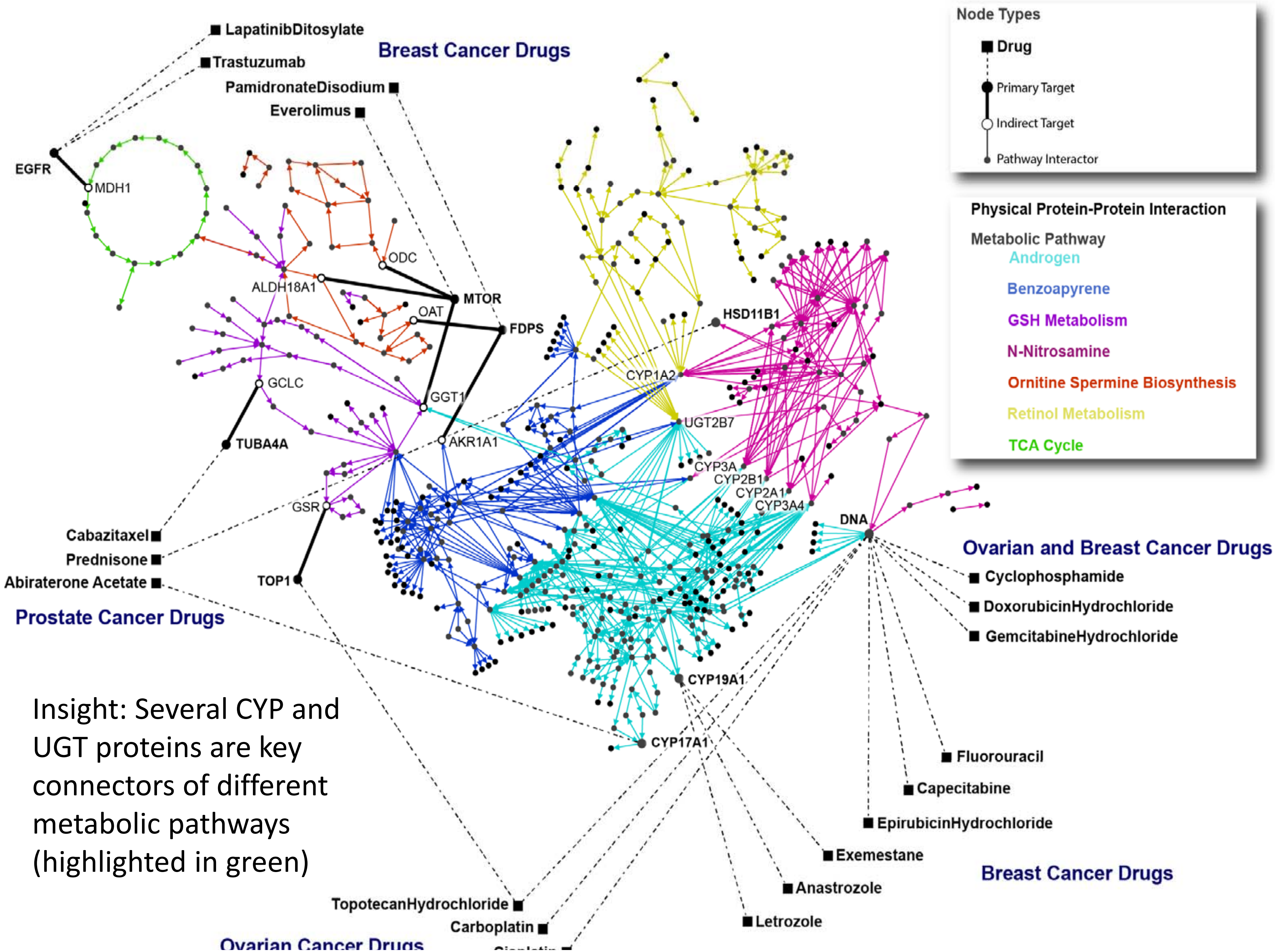
http://ophid.utoronto.ca/navigator

David Otasek, D;
Pastrello, C.;
Jurisica, I,
Holzinger, A
(2014): Visual Data
Mining:
Effective
Exploration ofthe
Biological Universe



JURISICA LAB
IBM Life Sciences Discovery Center

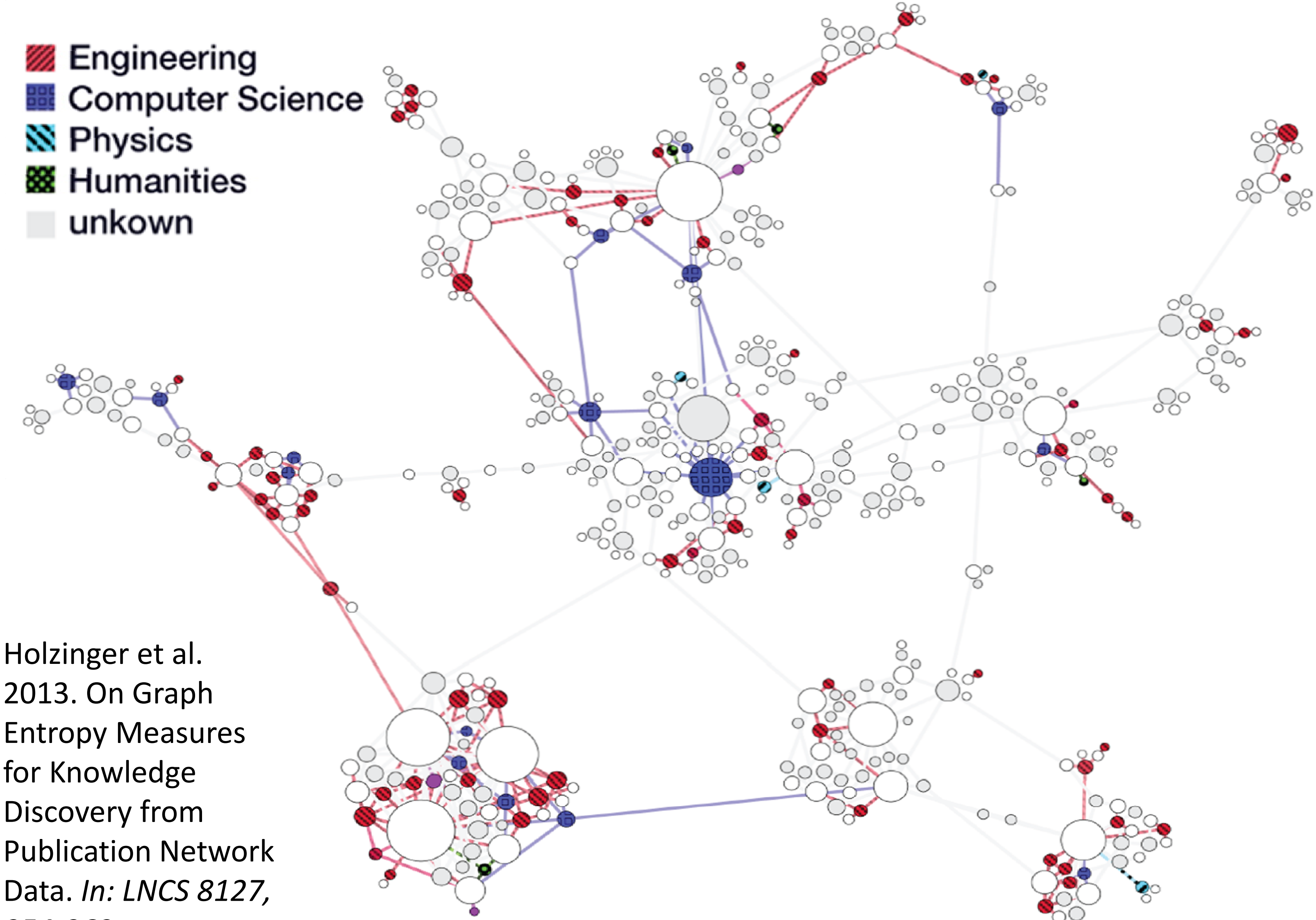
hci4all



Insight: Several CYP and UGT proteins are key connectors of different metabolic pathways (highlighted in green)

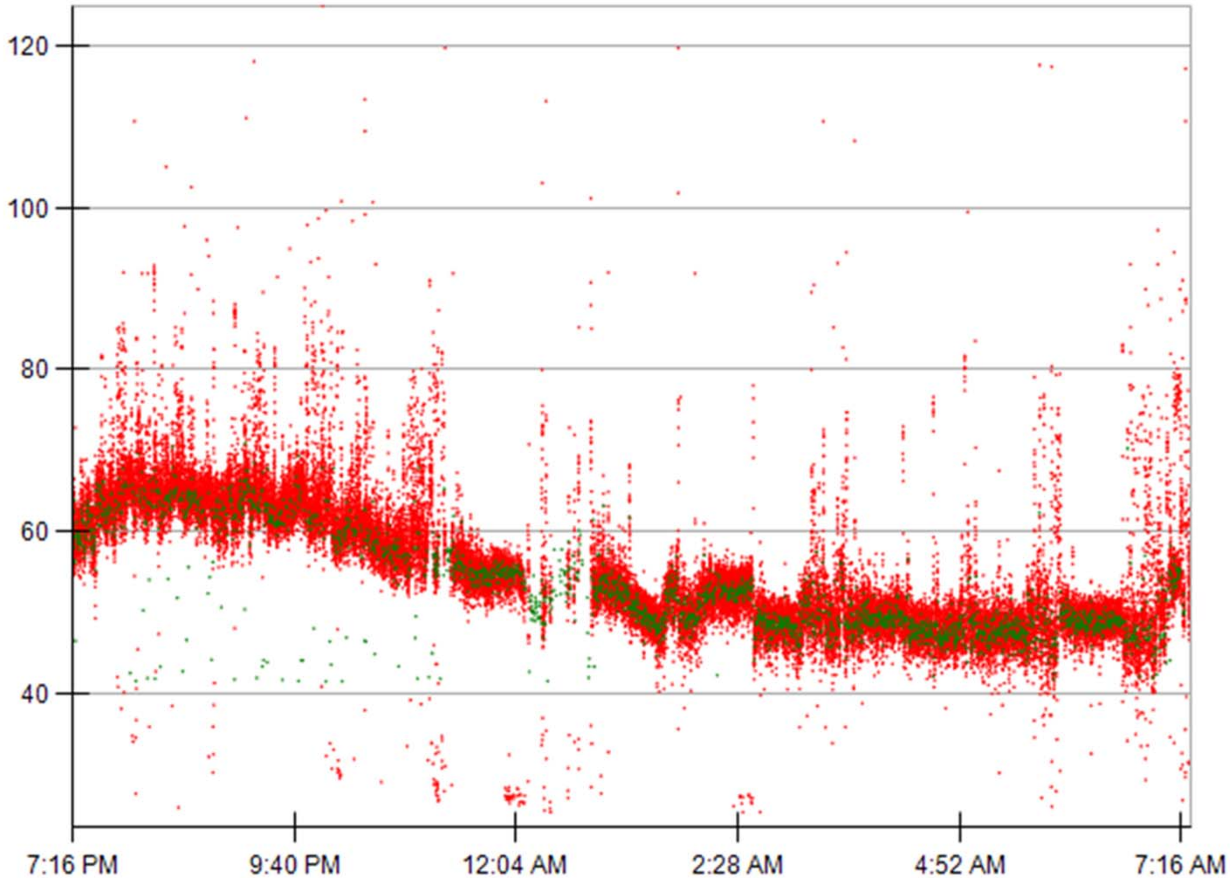
Three promising areas in interactive Data Mining

Example: Graph Entropy Measures



Holzinger et al. 2013. On Graph Entropy Measures for Knowledge Discovery from Publication Network Data. *In: LNCS 8127, 354-362.*

Example: Time Series Data



EU Project EMERGE (2007-2010)

Holzinger, A., Stocker, C., Bruschi, M., Auinger, A., Silva, H. & Fred, A. 2012. On Applying Approximate Entropy to ECG Signals for Knowledge Discovery on the Example of Big Sensor Data. In: R. Huang, E. a. E. (ed.) Active Media Technologies AMT 2012, LNCS 7669. Macau: Springer, pp. 646–657.

$$Let: \langle x_n \rangle = \{x_1, x_2, \dots, x_N\}$$

$$\vec{X}_i = (x_i, x_{(i+1)}, \dots, x_{(i+m-1)})$$

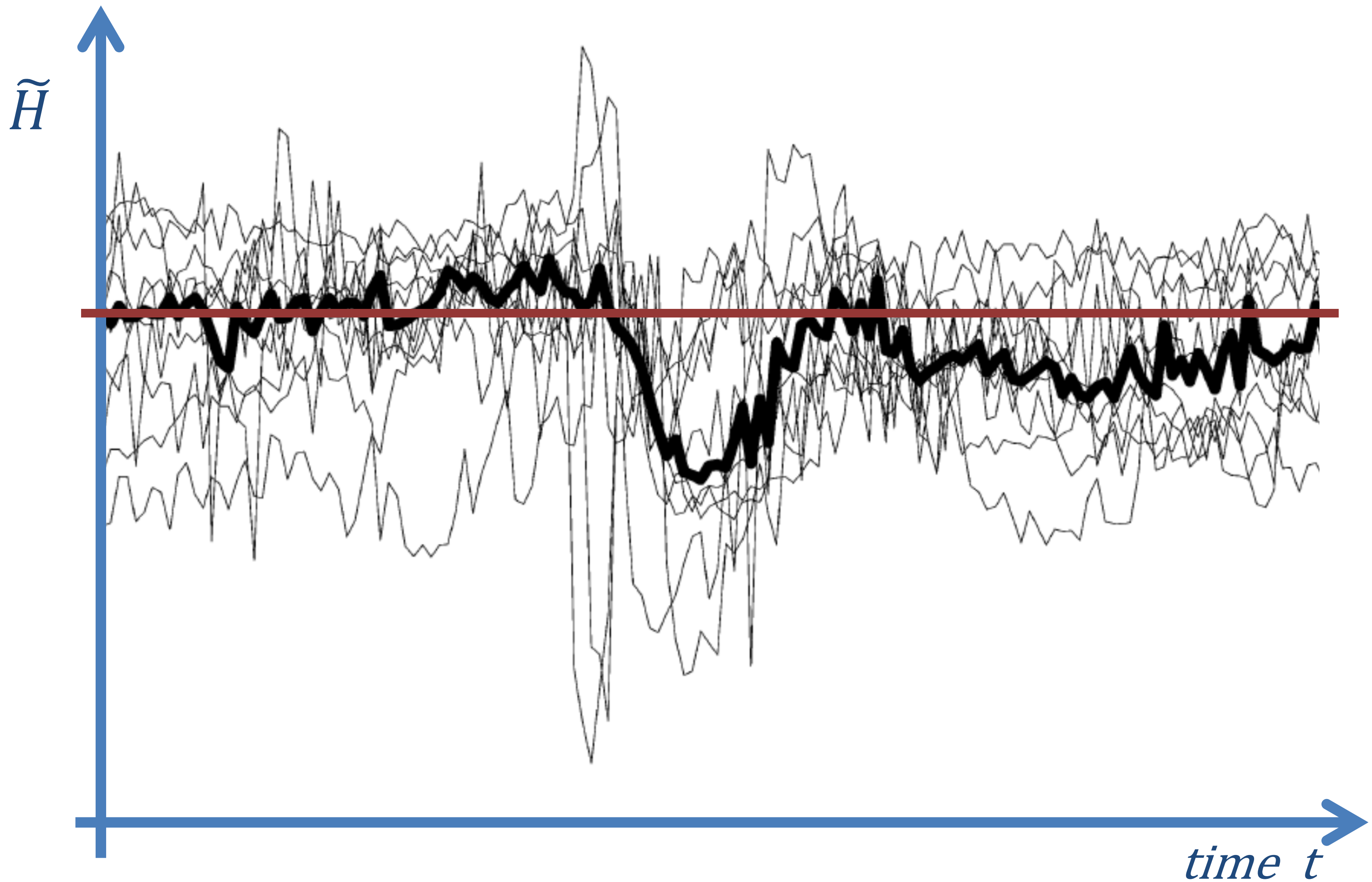
$$\|\vec{X}_i, \vec{X}_j\| = \max_{k=1,2,\dots,m} (|x_{(i+k-1)} - x_{(j+k-1)}|)$$

$$\boxed{\widetilde{\mathrm{H}}(m,r) = \lim_{N\to\infty} [\phi^m(r) - \phi^{m+1}(r)]}$$

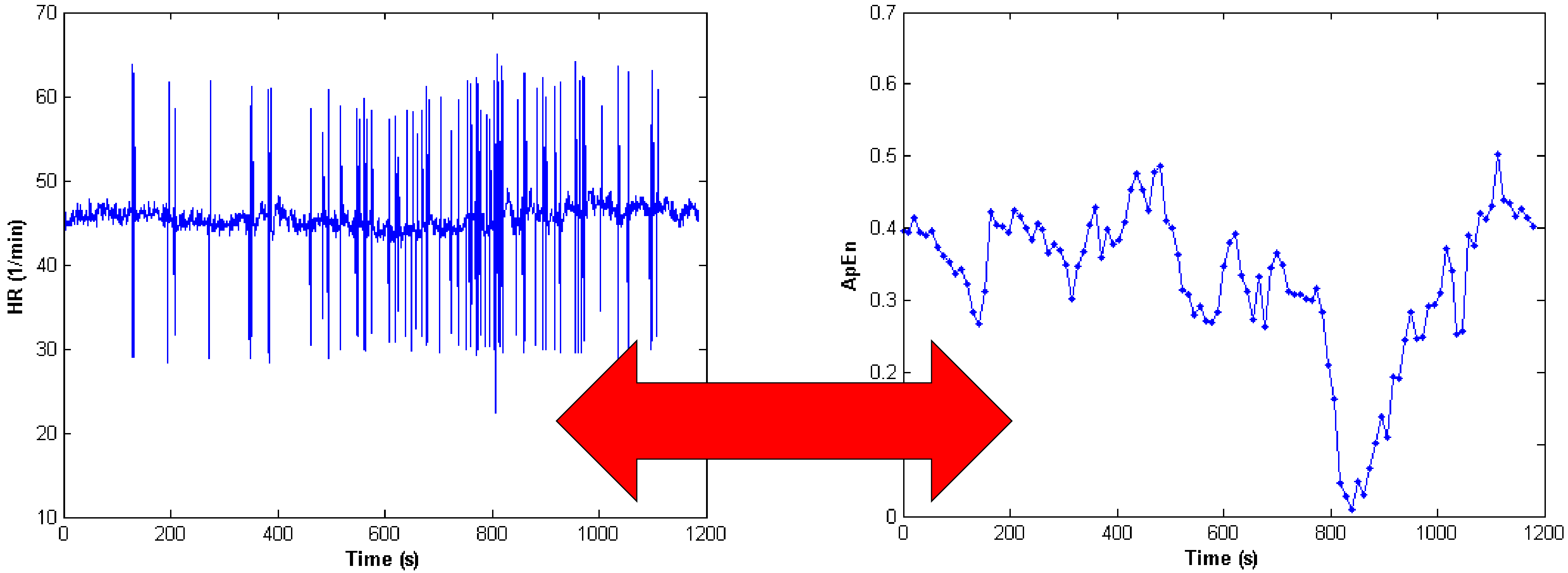
$$C_r^m(i) = \frac{N^m(i)}{N-m+1} \qquad \phi^m(r) = \frac{1}{N-m+1} \sum_{t=1}^{N-m+1} \ln C_r^m(i)$$

m is the run length and r is the tolerance window r (we must assume that m is equal to r!)

Example: ApEn (2/3)



ApEn (3/3)



- **1) Computational Geometric Topology**
  - Interaction with Graphs
  - Surfaces (e.g. manifolds)
  - Complexes (e.g. simplical complexes)
- 2) **Computational Algebraic Topology**
  - Homology (e.g. Betti numbers)
  - Duality (e.g. for projective spaces)
  - Morse Functions (e.g. analyze the topology of a manifold)
- 3) **Computational Persistent Topology**
  - Persistence (e.g. finding persistent homologies)
  - Stability
  - Examples: Gene Expression Data, Protein docking, Image Segmentation etc.

Conclusion: Let the user interactively manipulate the data

- **Focus Selection** = via direct manipulation and selection tools, e.g. multi-touch (in data space a n-dim location might be indicated); see a recent work by Randy Goebel
- **Attention Routing** = anomaly detection, draws people's attention to interesting areas to start their analyses;
- **Extent Selection** = specifying extents for an interaction, e.g. via a vector of values (a range for each data dimension or a set of constraints;
- **Interaction type selection** = e.g. a pair of menus: one to select the space, and the other to specify the general class of the interaction;
- **Interaction level selection** = e.g. the magnitude of scaling that will occur at the focal point (via a slider, along with a reset button);


Thank you!