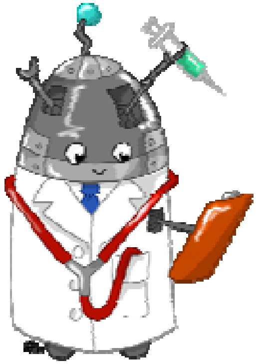


Andreas Holzinger

VO 709.049 Medical Informatics

01.02.2017 11:15-12:45

Lecture 12 Course Summary and Future Outlook (Reflection Lecture)

a.holzinger@tugraz.at

Tutor: markus.plass@student.tugraz.at

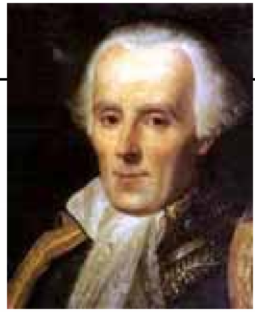
http://hci-kdd.org/biomedical-informatics-big-data

This whole lecture is a reflection course … so please


KEEP
CALM
AND
FASTEN YOUR
SEAT BELTS

Lecture 1: Computer Science meets Life Sciences

Repetition of Bayes - on the work of Laplace



What is the simplest mathematical operation for us?

$$p(x) = \sum_x (p(x, y)) \tag{1}$$

How do we call repeated adding?

$$p(x, y) = p(y|x) * p(y) \tag{2}$$

Laplace (1773) showed that we can write:

$$p(x, y) * p(y) = p(y|x) * p(x) \tag{3}$$

Now we introduce a third, more complicated operation:

$$\frac{p(x, y) * p(y)}{p(y)} = \frac{p(y|x) * p(x)}{p(y)} \tag{4}$$

We can reduce this fraction by $p(y)$ and we receive what is called Bayes rule:

$$p(x, y) = \frac{p(y|x) * p(x)}{p(y)} \qquad p(h|d) = \frac{p(d|h)p(h)}{p(d)} \tag{5}$$

d … $data$

h … $hypotheses$

$\mathcal{H}$ …$\{H_1, H_2, \ldots, H_n\}$ $\forall\, h, d$ …

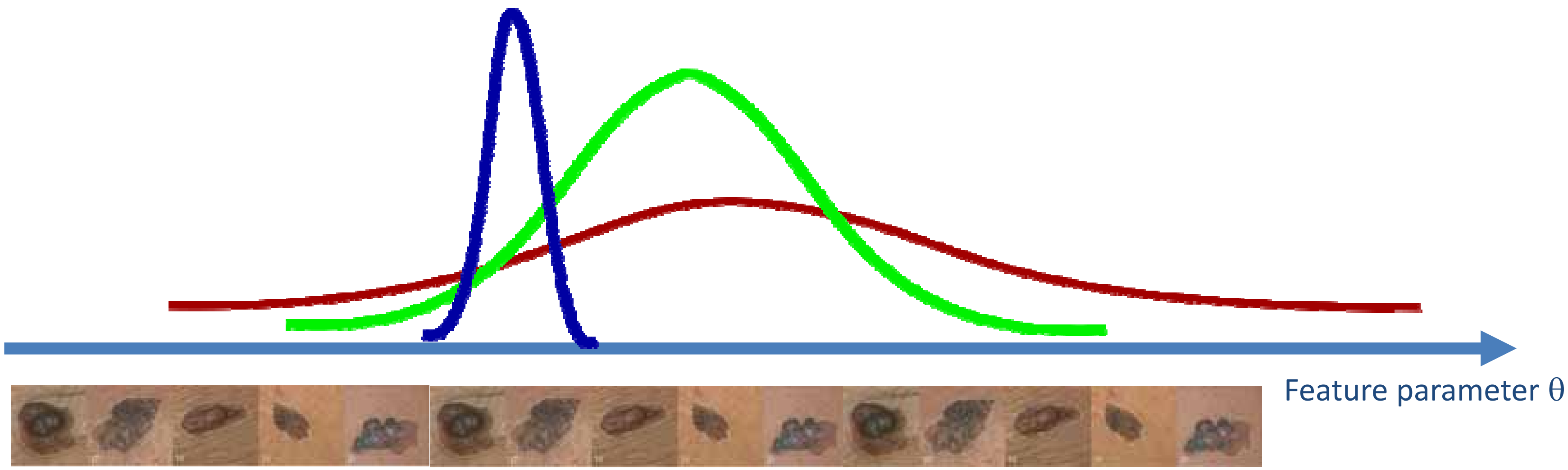
Likelihood

Prior Probability

$$p(h|d) = \frac{p(d|h) * p(h)}{\sum_{h \in H} p(d|h') \, p(h')}$$

Posterior Probability

Evidence = marginal likelihood = Normalization



Example for Learning and Bayesian Inference

- Your MD has bad news and good news for you.
- Bad news first: You are tested positive for a serious disease D, and this test T is 99% accurate
- Good news: It is a rare disease, striking only 1 in 10,000 (D)
- **How worried would you now be – or: what is the posterior?**

$$posterior = \frac{likelihood * prior}{evidence}$$

$$p(h|d) = \frac{p(d|h) * p(h)}{\sum(p(d|\bar{h}) * p(\bar{h}))}$$

$$p(T = 1|D = 1) = p(d|h) = 0{,}99 \ \ and$$

$$p(D = 1) = p(h) = 0{,}0001$$

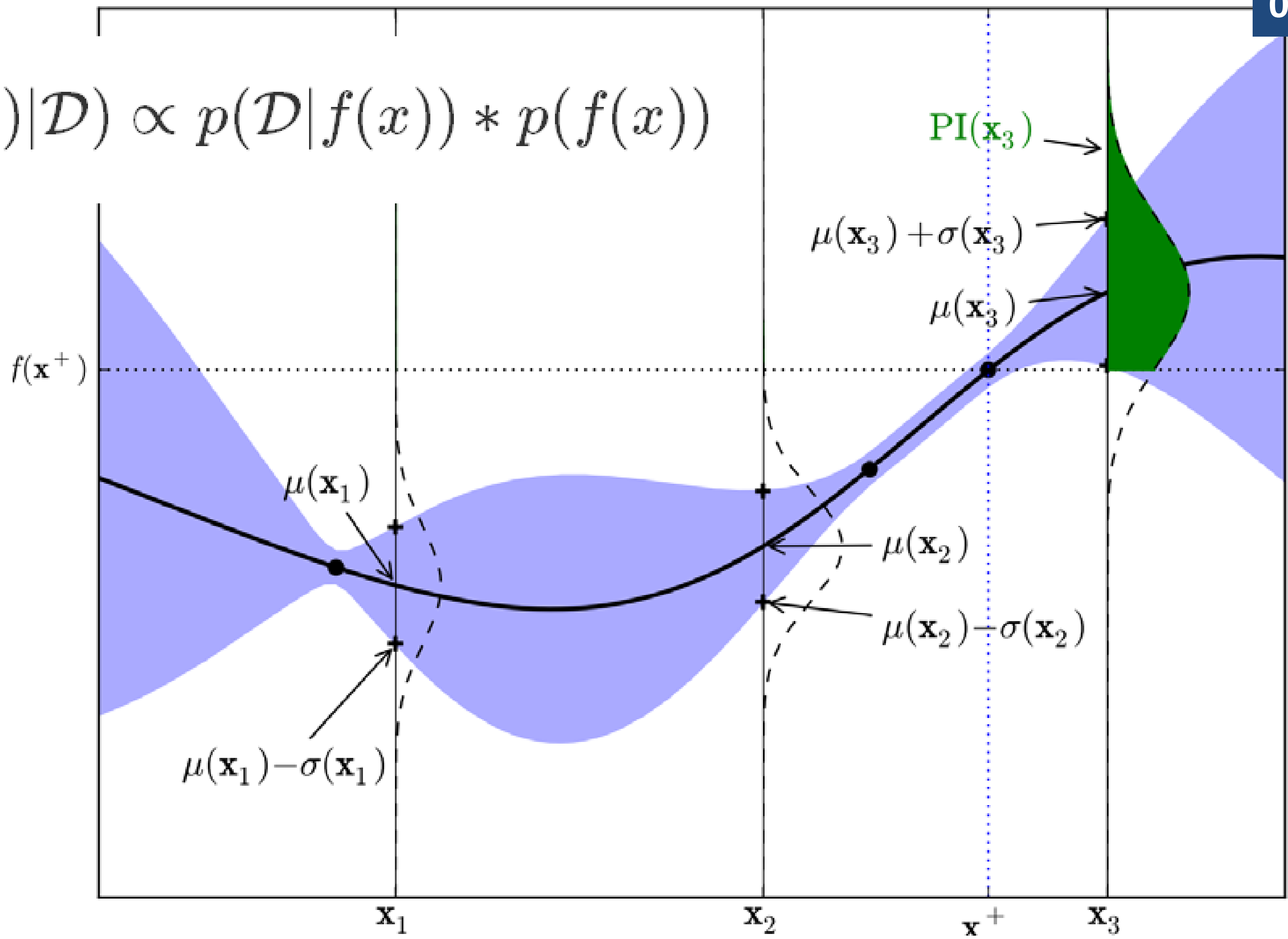
and $p(T = 0|D = 0) = 0{,}99$

$$p(D = 1 \,\big|\, T = 1) = \frac{(0{,}99)*(0{,}0001)}{(1-0{,}99)*(1-0{,}0001)+0{,}99*0{,}0001} =$$

= 0,0098 = 0,9%

Where is the highest certainty in the signal below?

$$p(f(x)|\mathcal{D}) \propto p(\mathcal{D}|f(x)) * p(f(x))$$

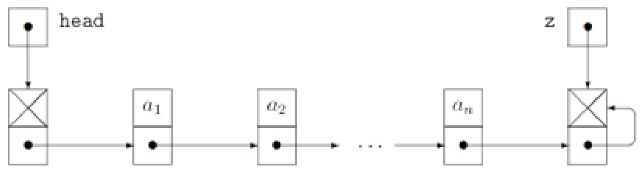


Brochu, E., Cora, V. M. & De Freitas, N. 2010. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. arXiv:1012.2599.

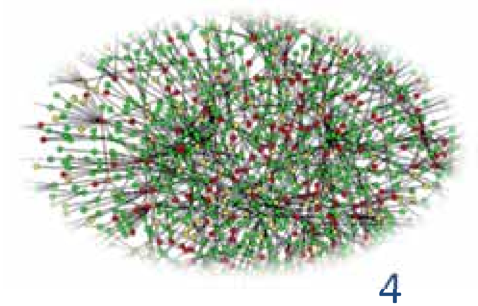
Lecture 2: Data, Information, Knowledge; Entropy and Kullback-Leibler Divergence

1

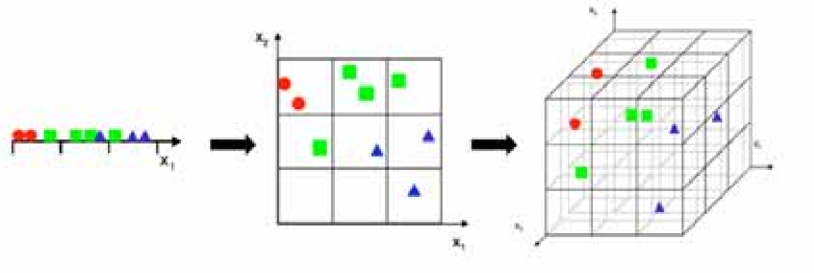
2



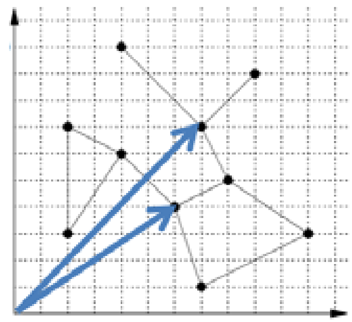
3



4



5



6

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_b P(x_i)$$

7

$$D_{\mathrm{KL}}(P\|Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x$$

8

9

What data do we have in biomedical informatics?

Electronic health record data
Physiological data
Laboratory results

Imaging data
X-Ray, ultrasound, MR, CT, PET, cams, observation (e.g. sleep laboratory), gait (walking)

Genomics

Proteomics
Protein-Protein Interactions

Metabolomics
Chemical processes
Cellular reactions
Enzymatic reactions

Transcriptomics
RNA, mRNA, rRNA, tRNA

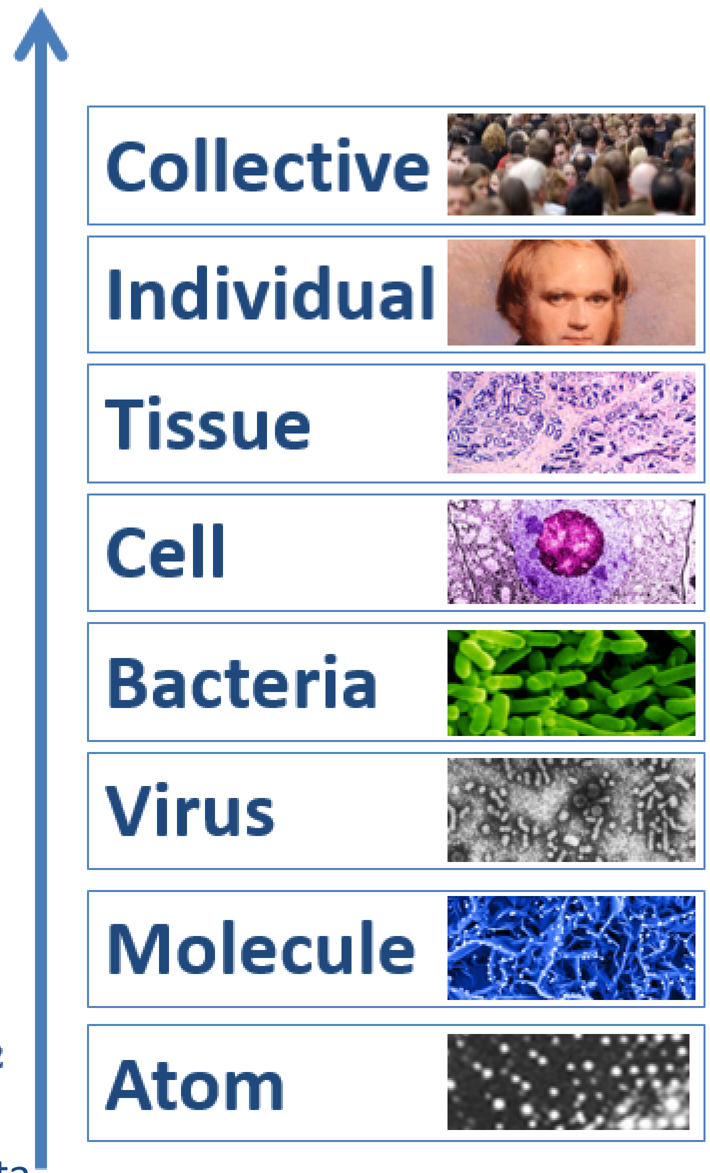
Epigenetics
Epigenetic modifications

Foodomics, Lipidomics
Nutrition data (Nutrigenomics)
Diet data (allergenics)

Exposome
Environmental data
Air pollution
Exposure (toxicants)

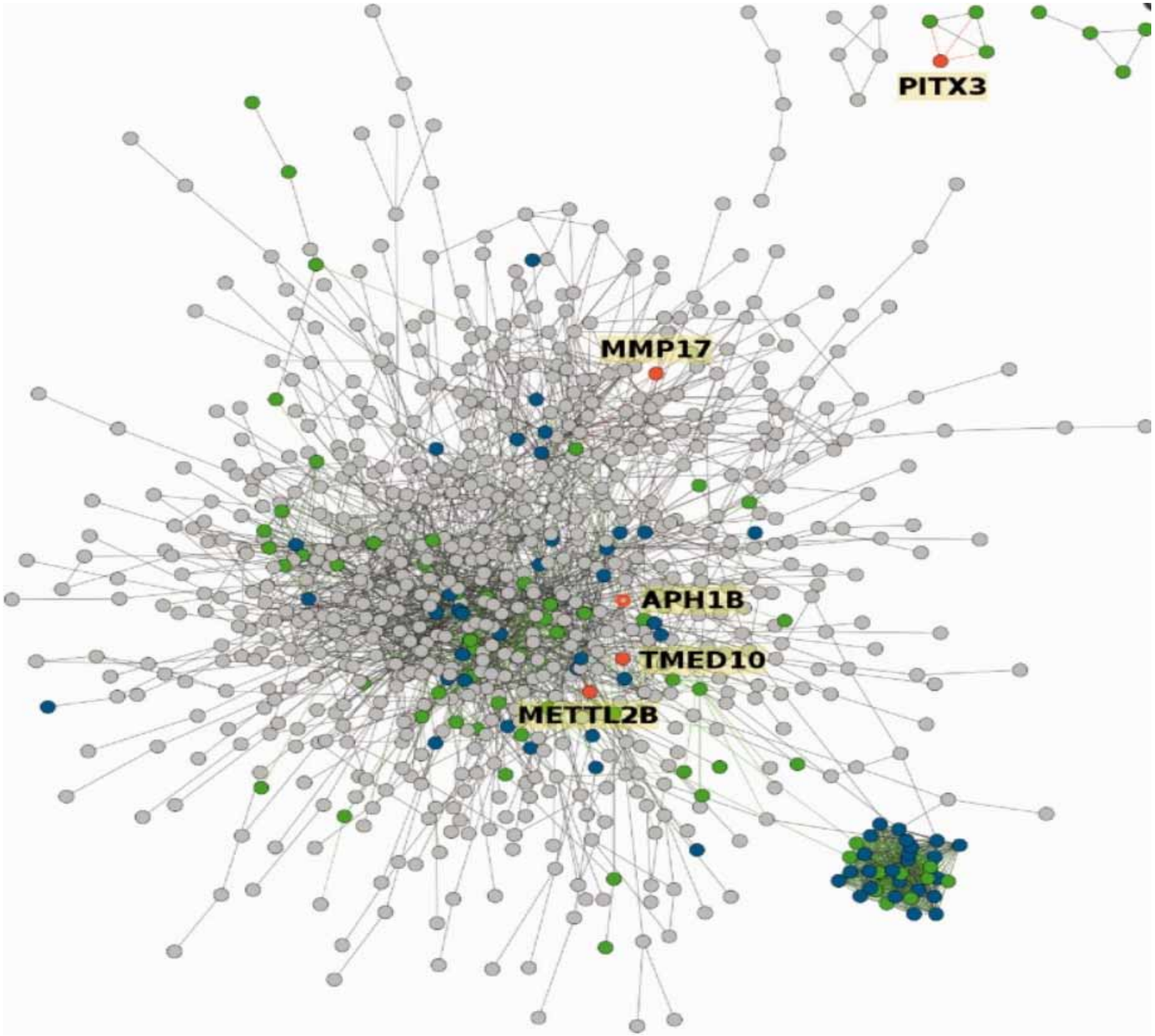
Collective data
Social data
Fitness, Wellness data
Ambient Assisted Living data
(Non-medical) personal data

10^{-12}



Question: Why is it so important to know such structures?

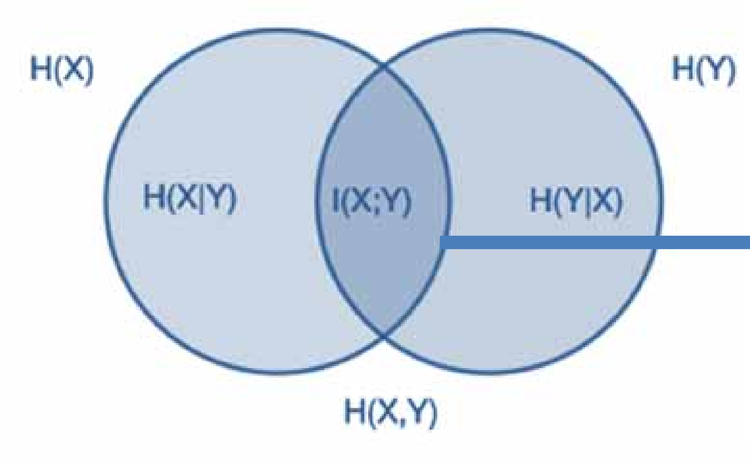
Winterhalter, C., Widera, P. & Krasnogor, N. 2014. JEPETTO: a Cytoscape plugin for gene set enrichment and topological analysis based on interaction networks. Bioinformatics, 30, (7), 1029-1030, doi:10.1093/bioinformatics/btt732.



What is/When do we need Shannon-Entropy?

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_b P(x_i)$$

- Measuring uncertainty, complexity, randomness, surprise, …, = **information!**



$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right)$$

$$I(X;Y) = H(X) - H(X|Y)$$

- In ML we need often to measure the **difference between two probability distributions**

Kullback-Leibler Divergence - discriminative information

For discrete distributions

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

For continuous distributions

$$D_{\mathrm{KL}}(P\|Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x$$

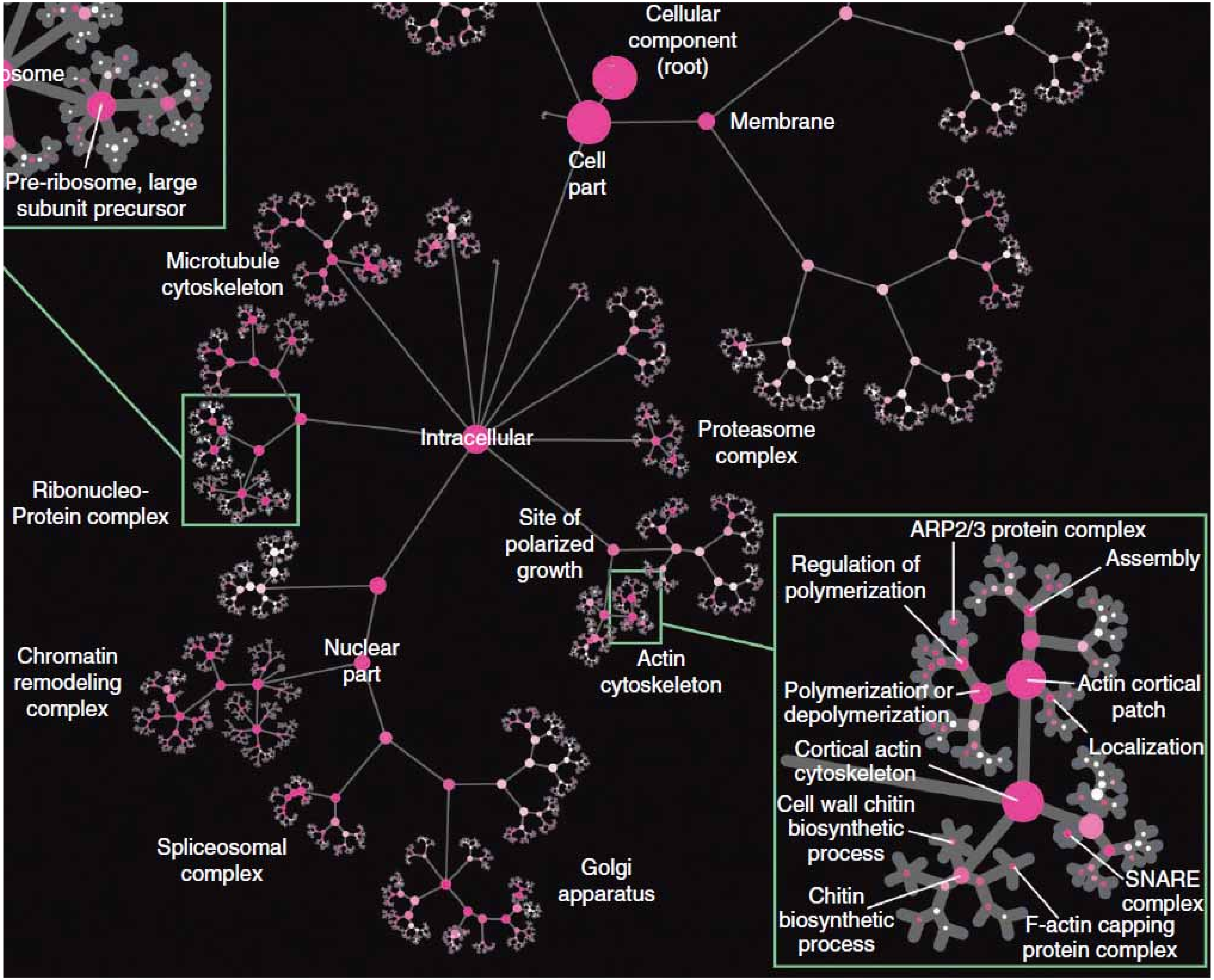
$$\mathrm{KL}(p\|q) \geqslant 0 \qquad \mathrm{KL}(p\|q) \not\equiv \mathrm{KL}(q\|p)$$

KL-divergence can also be used to measure the **distance between two distributions**

Kullback, S. & Leibler, R. A. 1951. On information and sufficiency. The annals of mathematical statistics, 22, (1), 79-86, doi:http://www.jstor.org/stable/2236703

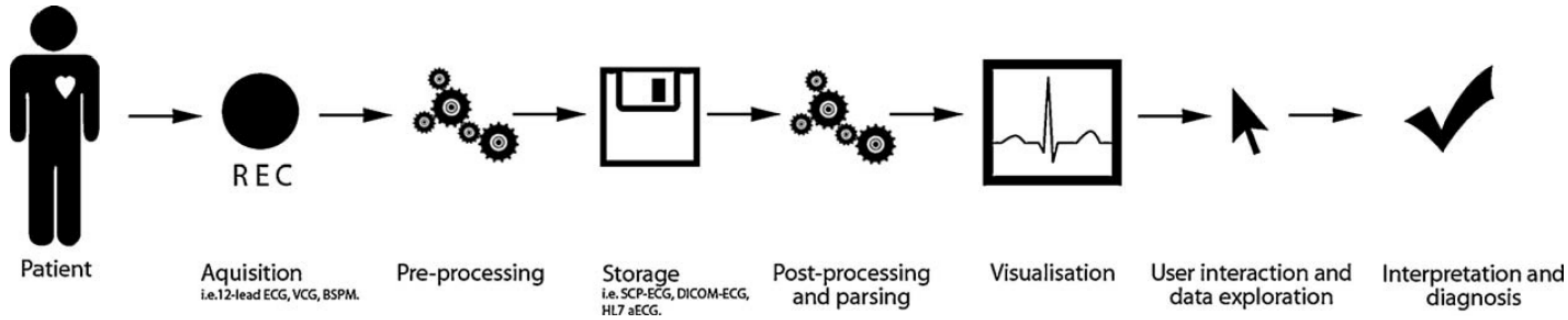
Lecture 3: Knowledge Representation, Ontologies & Classifications

What is this, and why is it important?

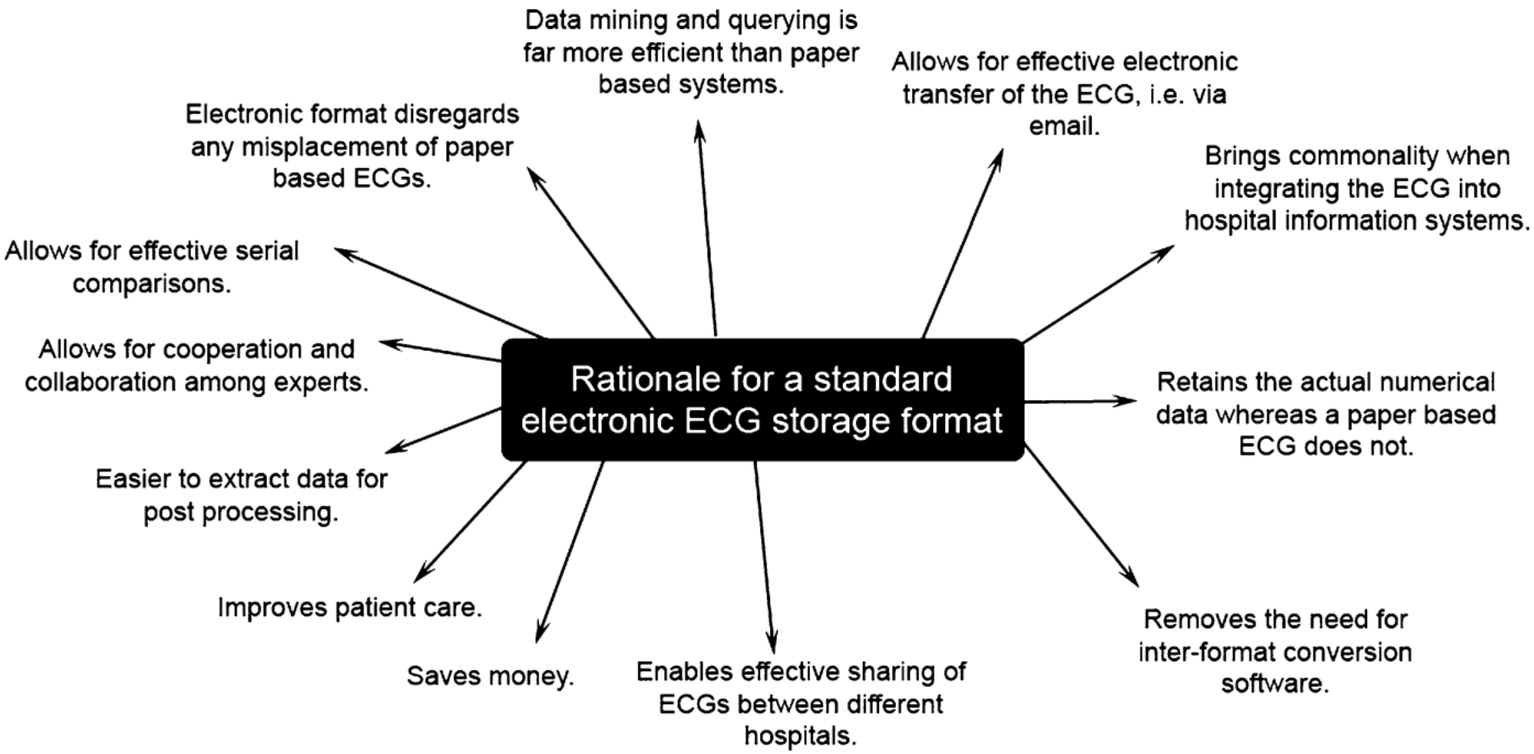


http://www.kurzweilai.net/images/cell-model.png (Credit: UC San Diego School of Medicine)

Slide 3-7: Standardized workflow of ECG data processing



Bond, R. R., Finlay, D. D., Nugent, C. D. & Moore, G. (2011) A review of ECG storage formats. *International Journal of Medical Informatics 80, 10, 681-697.*

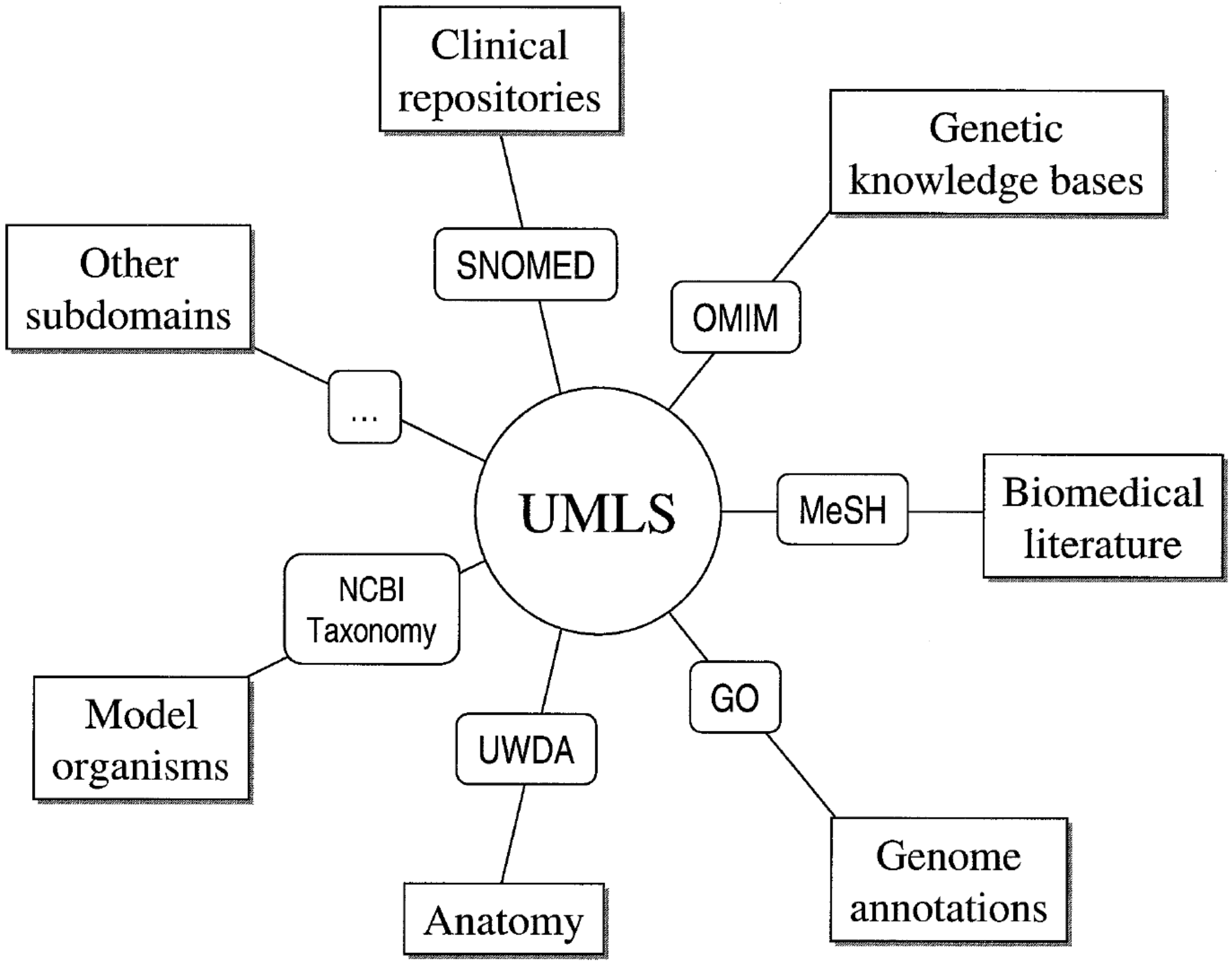


What is a knowledge representation?

Mathematical Logic	Psychology	Biology	Statistics	Economics
Aristotle				
Descartes				
Boole	James		Laplace	Bentham
				Pareto
Frege			Bernoullii	Friedman
Peano				
	Hebb	Lashley	Bayes	
Goedel	Bruner	Rosenblatt		
Post	Miller	Ashby	Tversky,	Von Neumann
Church	Newell,	Lettvin	Kahneman	Simon
Turing	Simon	McCulloch, Pitts		Raiffa
Davis		Heubel, Weisel		
Putnam				
Robinson				
Logic PROLOG	SOAR **KBS, Frames**	**Connectionism**	**Causal Networks**	**Rational Agents**

Davis, R., Shrobe, H. , Szolovits, P. 1993 What is a knowledge representation? AI Magazine, 14, 1, 17-33.

Why is this image important?



Lecture 4: Decision, Cognition, Uncertainty, Bayesian Statistics, Probabilistic Modelling

When is the human *) better?

*) human intelligence/natural intelligence/human mind/human brain/ learning

- **Natural Language Translation/Curation**

  Computers cannot understand the context of sentences [3]
- **Unstructured problem solving**

  Without a pre-set of rules, a machine has trouble solving the problem, because it lacks the creativity required for it [1]
- **NP-hard Problems**

  Processing times are often exponential and makes it almost impossible to use machines for it, but human make heuristic decisions which are often not perfect but sufficiently good [4]

When is the computer **) better?

**) Computational intelligence, Artificial Intelligence/soft computing/ML

- **High-dimensional data processing**

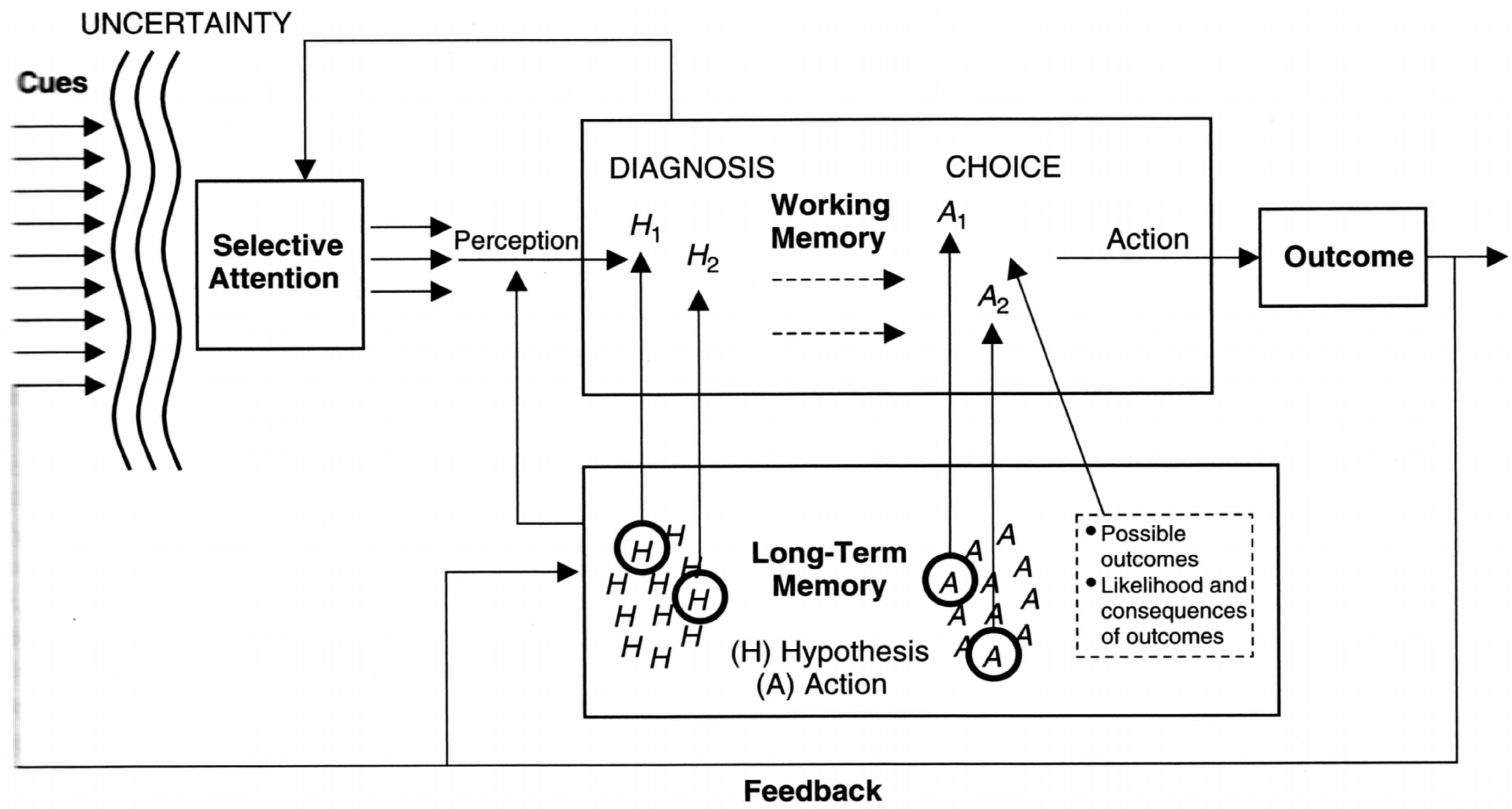
  Humans are very good at dimensions less or equal than 3, but computers can process data in arbitrarily high dimensions
- **Rule-Based environments**

  Difficulties for humans in rule-based environments often come from not recognizing the correct goal in order to select the correct procedure or set of rules [2]
- **Image optimization**

  Machine can look at each pixel and apply changes without human personal biases, and with more speed [1]

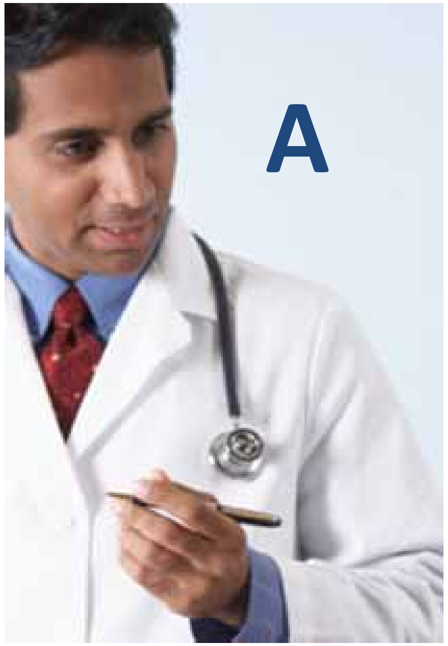
[1] Kipp, M. 2006. Creativity Meets Automation: Combining Nonverbal Action Authoring with Rules and Machine Learning. In: LNCS 4133, pp. 230-242, doi:10.1007/11821830_19.
[2] Cummings, M. M. 2014. Man versus Machine or Man + Machine? IEEE Intelligent Systems, 29, (5), 62-69, doi:10.1109/MIS.2014.87.
[3 Pizlo, Z., Joshi, A. & Graham, S. M. 1994. Problem Solving in Human Beings and Computers. Purdue TR 94-075.
[4] Griffiths, T. L. Connecting human and machine learning via probabilistic models of cognition. Interspeech, 2009, ISCA, 9-12
See also: Holzinger, A., Plass, M., Holzinger, K., Crisan, G., Pintea, C. & Palade, V. 2016. Towards interactive Machine Learning (iML): Applying Ant Colony Algorithms to solve the Traveling Salesman Problem with the Human-in-the-Loop approach. Springer Lecture Notes in Computer Science LNCS 9817. Heidelberg, Berlin, New York: Springer, pp. 81-95, doi:10.1007/978-3-319-45507-56.

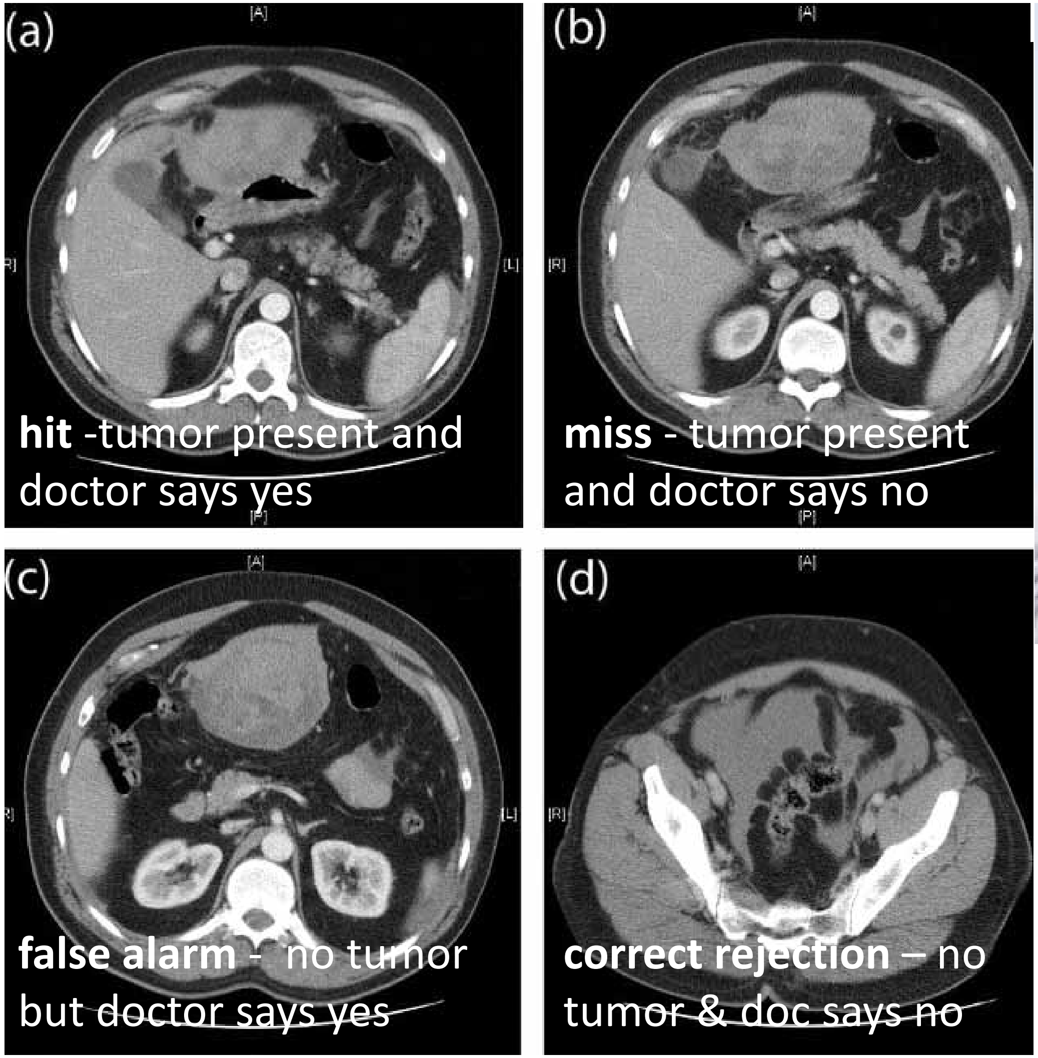
Slide 7-16 Human Decision Making

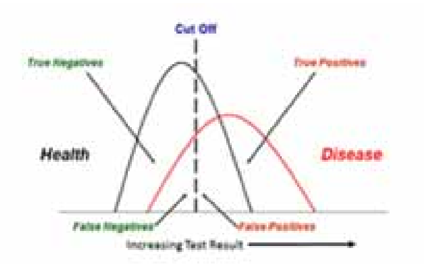


Wickens, C. D. (1984) *Engineering psychology and human performance. Columbus (OH), Charles Merrill.*

What does the Signal Detection Theory describe?







Two doctors, with equally good training, looking at the same CT scan, will have the same information … but they may have a different bias/criteria!

For a single decision variable an agent can select $D = d$ for any $d \in dom(D)$.

The expected utility of decision $D = d$ is

http://www.eoht.info/page/Oskar+Morgenstern

$$E(U \mid d) = \sum_{x_1,\ldots,x_n} P(x_1,\ldots,x_n \mid d)U(x_1,\ldots,x_n,d)$$

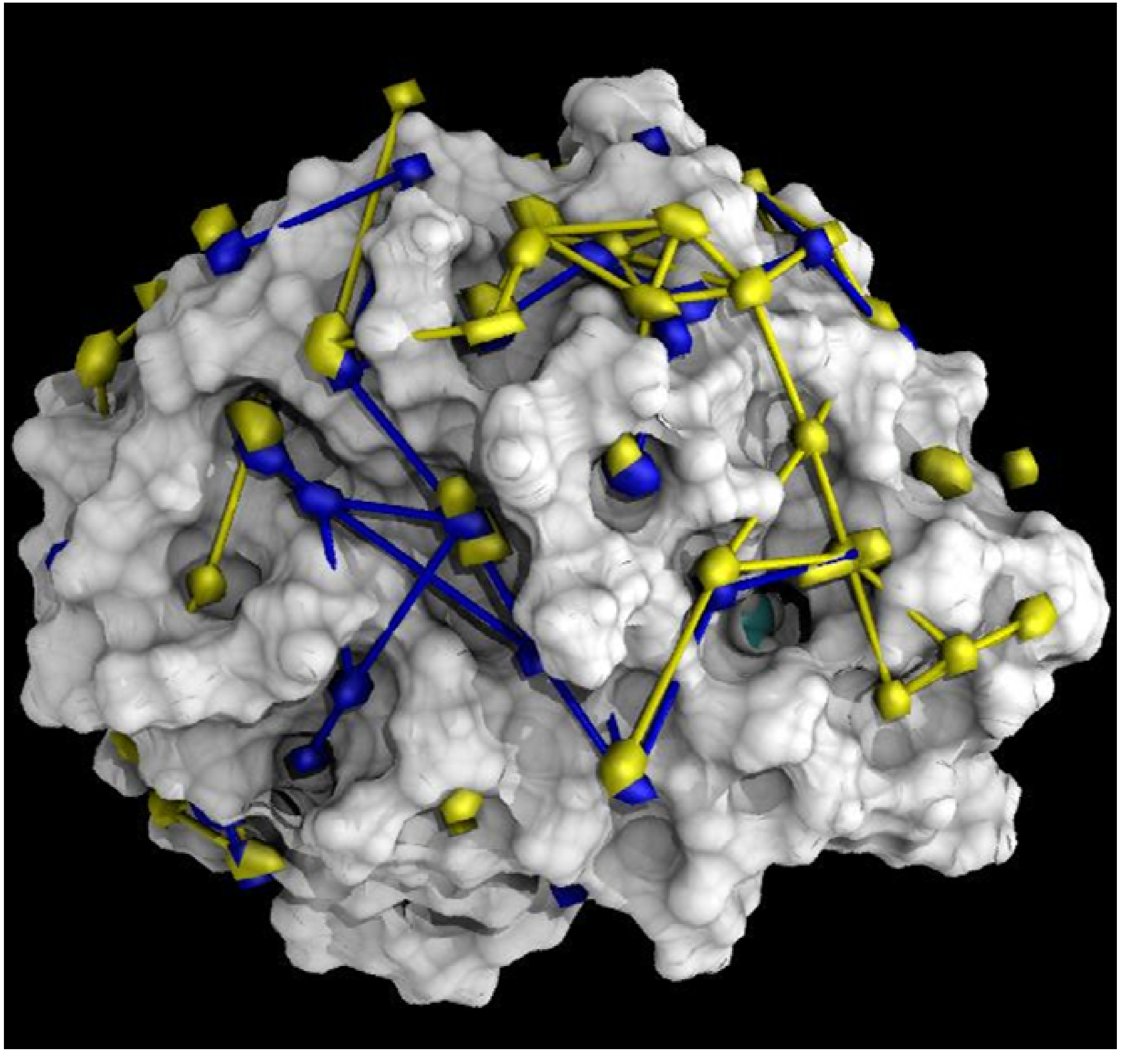
An optimal single decision is the decision $D = dmax$ whose expected utility is maximal:

$$d_{\max} = \arg \max_{d \in \mathrm{dom}(D)} E(U \mid d)$$

Von Neumann, J. & Morgenstern, O. 1947. Theory of games and economic behavior, Princeton university press.

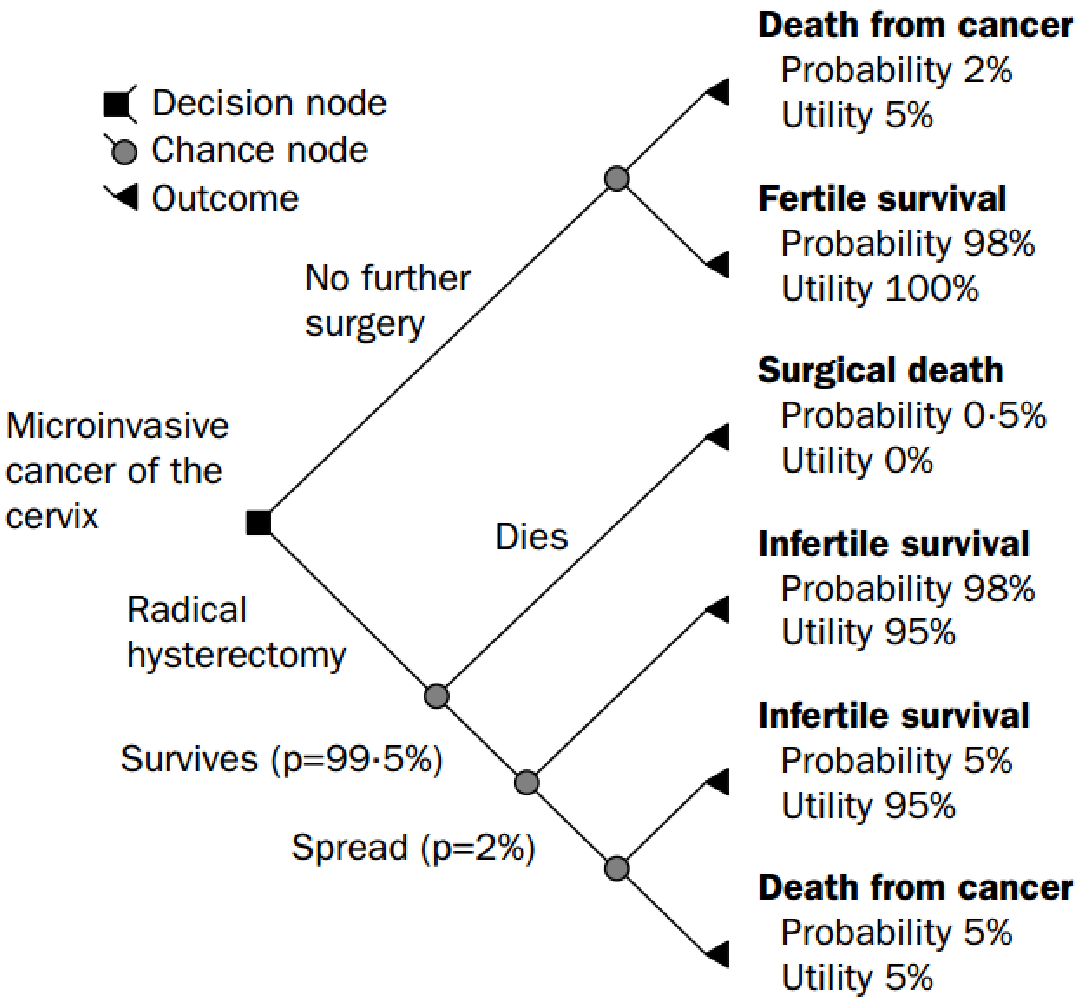
Lecture 5: Probabilistic Graphical Models I: From Knowledge Representation to Graph Learning

This protein graph model was worth the Nobel Prize 2013 …



http://sbcb.bioch.ox.ac.uk/users/oliver/software/

Remember: Decision trees are coming from Clinical Practice

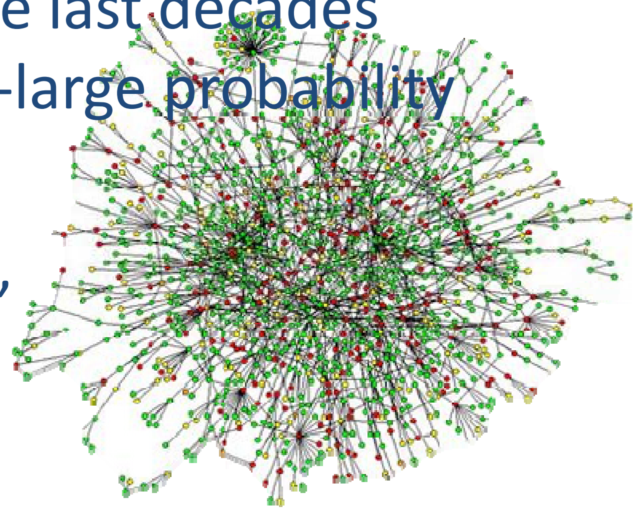


Physician treating a patient approx. 480 B.C.
Beazley (1963), Attic Red-figured Vase-Painters, 813, 96.
Department of Greek, Etruscan and Roman Antiquities, Sully, 1st floor, Campana Gallery, room 43 Louvre, Paris

Elwyn, G., Edwards, A., Eccles, M. & Rovner, D. 2001. Decision analysis in patient care. The Lancet, 358, (9281), 571-574.

What are Probabilistic Graphical Models?

- PGM can be seen as a combination between
- **Graph Theory + Probability Theory + Machine Learning**
- One of the most exciting AI advances in the last decades
- Compact representation for exponentially-large probability distributions



- Example Question:
  “Is there a path connecting two proteins?”
- $Path\ (X,Y) := \ edge\ (X,Y)$
- $Path\ (X,Y) := \ edge\ (X,Y), path\ (Z,Y)$
- This can NOT be expressed in first-order logic
- Need a Turing-complete fully-fledged language

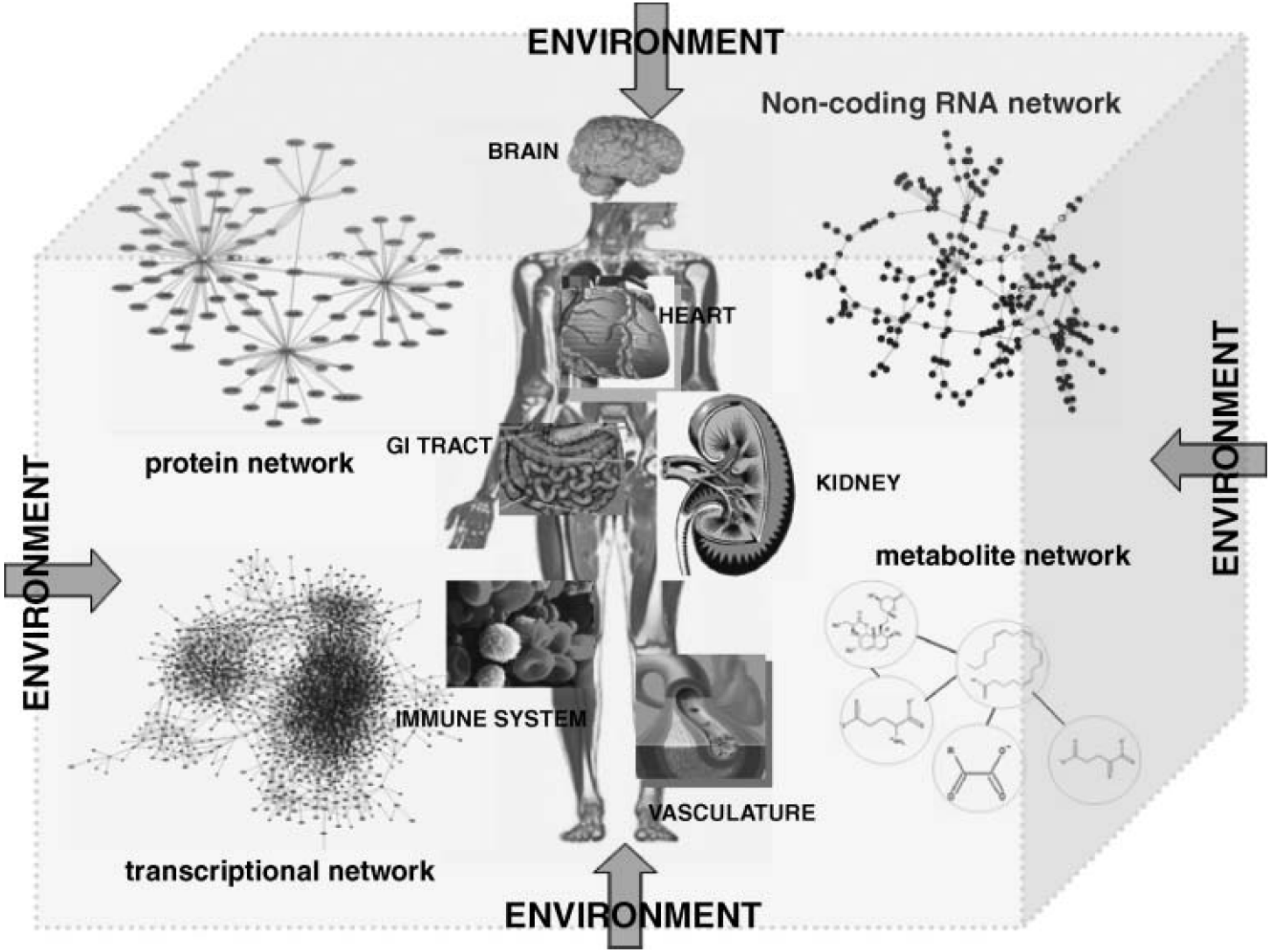
Friedman, N. 2004. Inferring cellular networks using probabilistic graphical models. Science, 303, (5659), 799-805.

Koller, D. & Friedman, N. 2009. Probabilistic graphical models: principles and techniques, MIT press.

- Take patient information, e.g., observations, symptoms, test results, -omics data, etc. etc.
- Reach conclusions, and predict into the future, e.g. how likely will the patient be re-admissioned
- **Prior** = belief before making a particular observation
- **Posterior** = belief after making the observation and is the prior for the next observation – intrinsically incremental

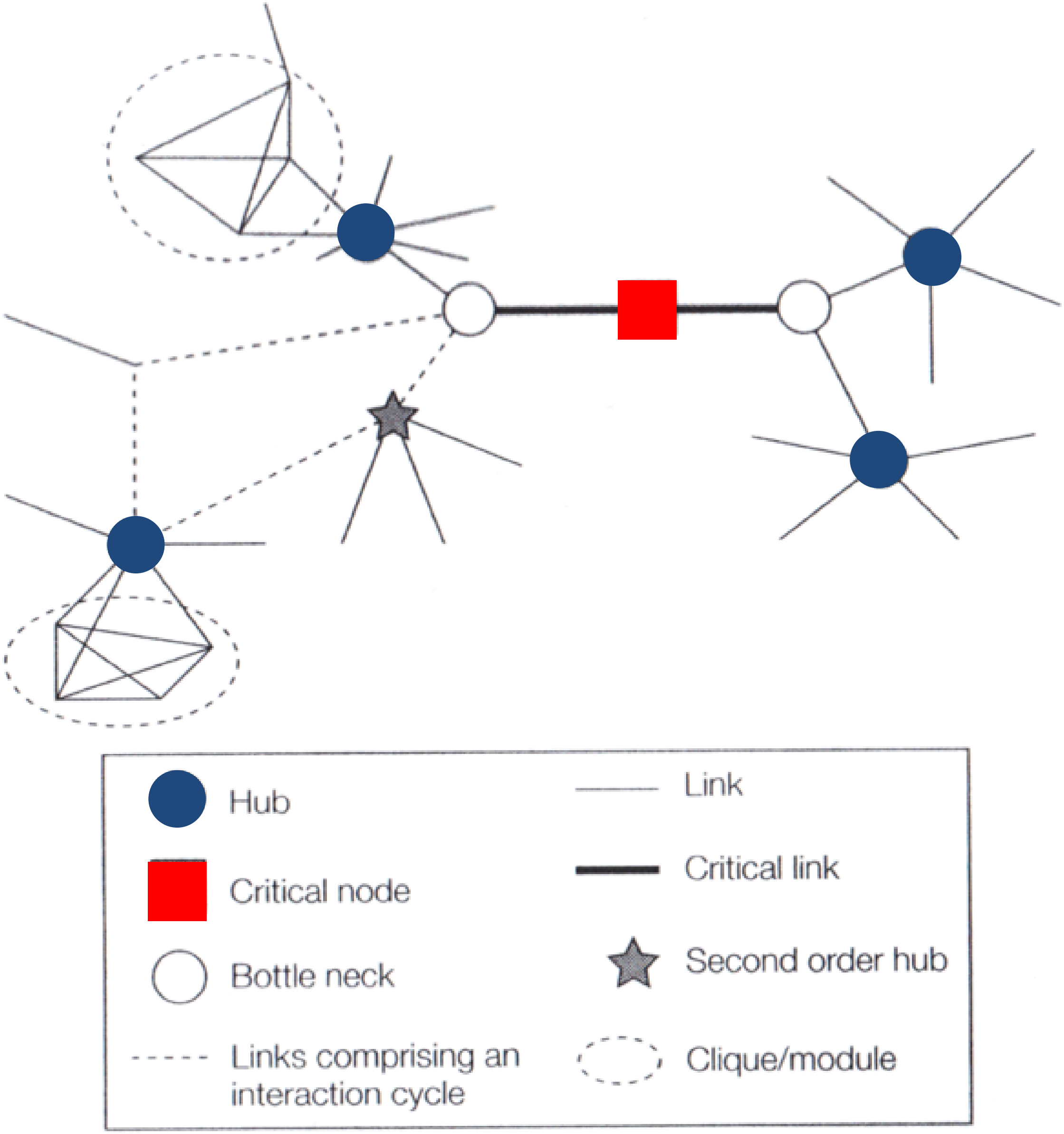
$$p(x_i|y_j) = \frac{p(y_j|x_i)p(x_i)}{\sum p(x_i, y_j)p(x_i)}$$

Identifying Networks in Disease Research



Schadt, E. E. & Lum, P. Y. (2006) Reverse engineering gene networks to identify key drivers of complex disease phenotypes. *Journal of lipid research, 47, 12, 2601-2613*.

$$G(V,E)\ Graph$$
$$V\ \ldots vertex$$
$$E\ \ldots edge\ \{a,b\}$$
$$\mathbf{a},\mathbf{b}\ \in V; a \neq b$$

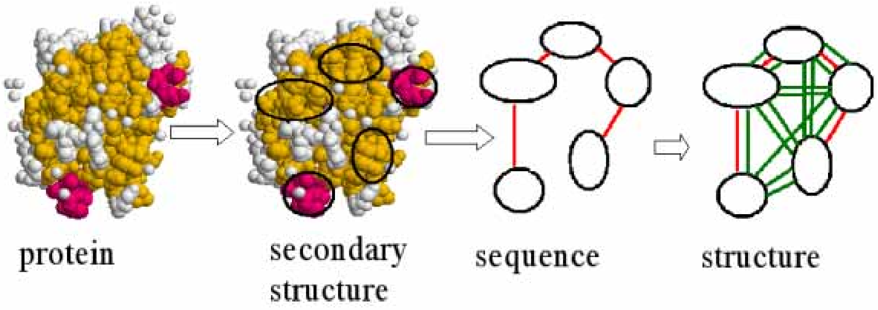


Hodgman, C. T., French, A. & Westhead, D. R. (2010) *Bioinformatics. Second Edition. New York, Taylor & Francis.*

Lecture 6: Probabilistic Graphical Models II: From Bayesian Networks to Graph Bandits

Problem: Is Graph Isomorphism NP-complete ?

Borgwardt, K. M., Ong, C. S., Schönauer, S., Vishwanathan, S., Smola, A. J. & Kriegel, H.-P. 2005. Protein function prediction via graph kernels. Bioinformatics, 21, (suppl 1), i47-i56.



- Important for health informatics: Discovering relationships between biological components
- Unsolved problem in computer science:
- Can the graph isomorphism problem be solved in polynomial time?
  - So far, no polynomial time algorithm is known.
  - It is also not known if it is NP-complete
  - We know that subgraph-isomorphism is NP-complete

Bayesian Network (BN) - Definition

- is a **probabilistic model**, consisting of two parts:
- 1) a dependency structure and
- 2) local probability models.

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i \,|\, Pa(x_i))$$

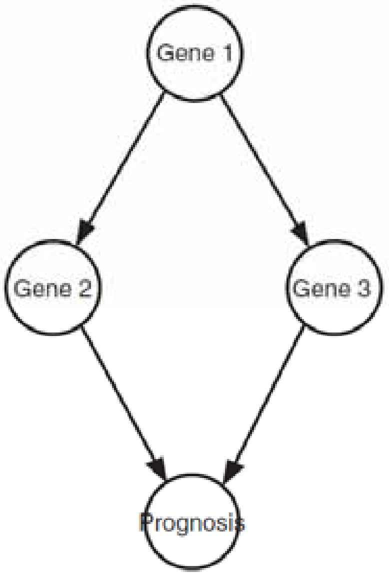
Where $Pa(x_i)$ are the parents of x_i

BN inherently model the **uncertainty in the data.** They are a successful marriage between probability theory and graph theory; allow to model a multidimensional probability distribution in a sparse way by searching independency relations in the data. Furthermore this model allows different strategies to integrate two data sources.

Pearl, J. (1988) *Probabilistic reasoning in intelligent systems: networks of plausible inference. San Francisco, Morgan Kaufmann.*

Example: Bayes Net with four binary variables

Gene 1	
P(on)	0.8
P(off)	0.2



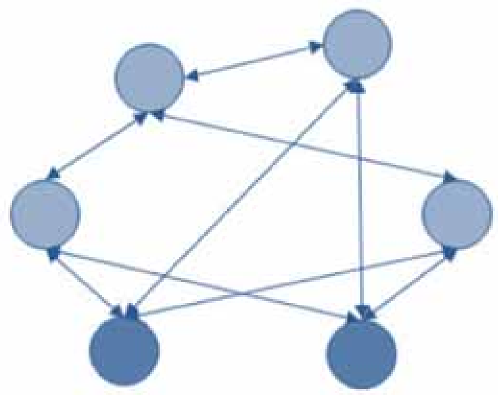
Gene 2	Gene 1 on	Gene 1 off
P(on)	0.3	0.6
P(off)	0.7	0.4

Gene 2	Gene 1 on	Gene 1 off
P(on)	0.3	0.6
P(off)	0.7	0.4

Prognosis	Gene 2 on Gene 3 on	Gene 2 on Gene 3 off	Gene 2 off Gene 2 on	Gene 2 off Gene 3 off
P(good)	0.6	0.1	0.9	0.5
P(poor)	0.4	0.9	0.1	0.5

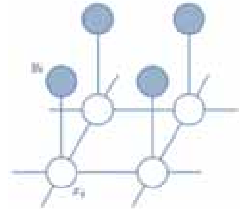
Gevaert, O., Smet, F. D., Timmerman, D., Moreau, Y. & Moor, B. D. (2006) Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics, 22, 14, 184-190.*

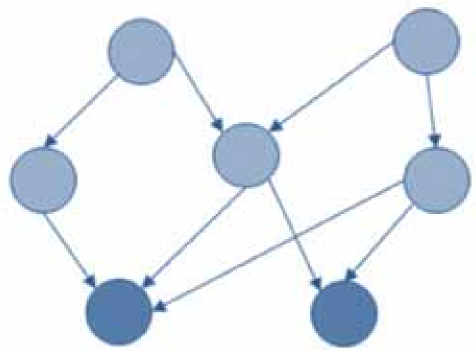
Three types of Probabilistic Graphical Models



Undirected: Markov random fields, useful e.g. for computer vision (Details: Murphy 19)

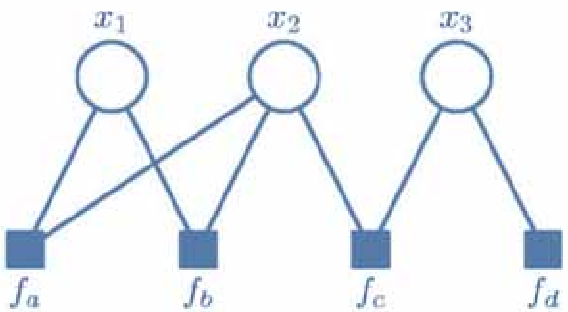
$$P(\boldsymbol{X}) = \frac{1}{Z}\exp\left(\sum_{ij} W_{ij}\, x_i x_j + \sum_i x_i b_i\right)$$





Directed: Bayes Nets, useful for designing models (Details: Murphy 10)

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k)$$



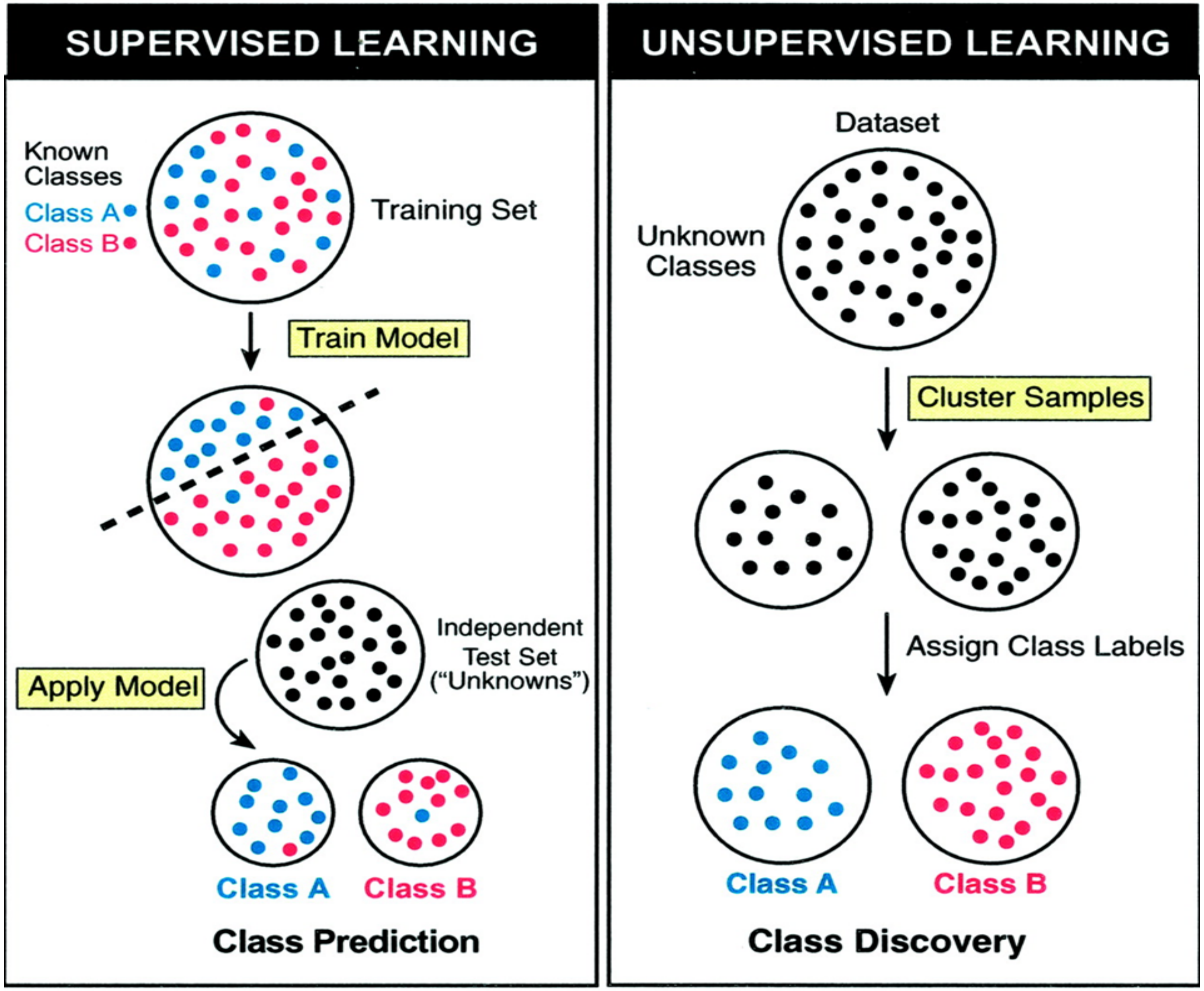
Factored: useful for inference/learning

$$p(\mathbf{x}) = \prod_s f_s(\mathbf{x}_s)$$

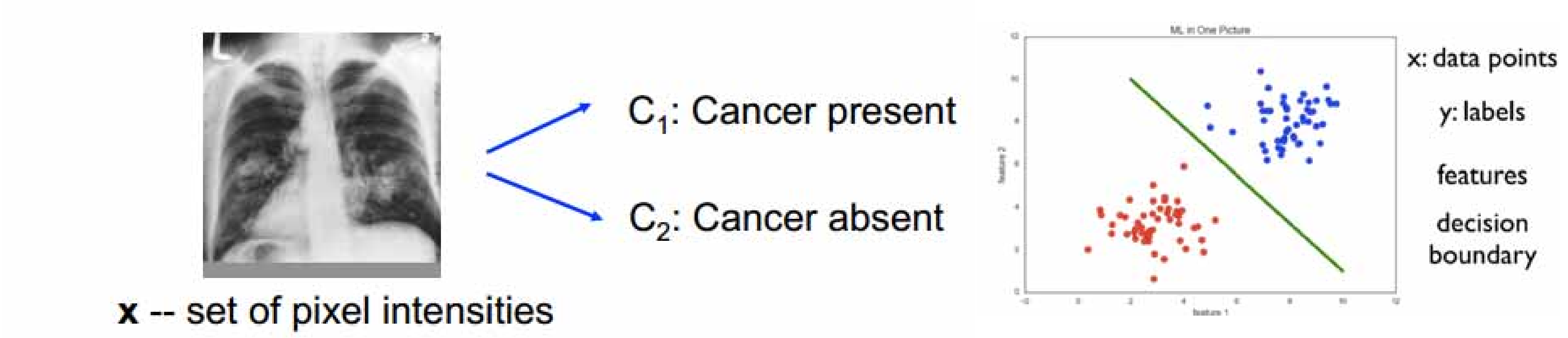
Tutorial on Factor Graphs http://deepdive.stanford.edu/inference

Lecture 7: Dimensionality Reduction and Subspace Clustering with the Doctor-in-the-Loop

Discovery vs. Prediction

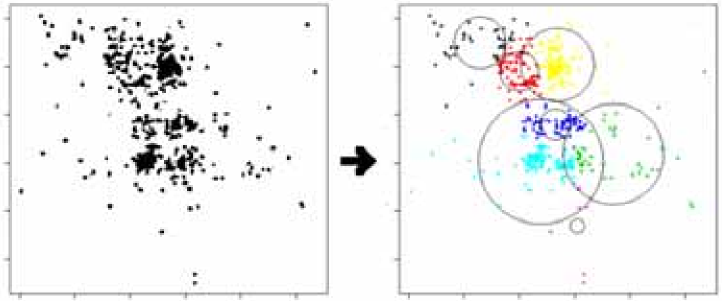


Ramaswamy, S. & Golub, T. R. (2002) DNA Microarrays in Clinical Oncology. *Journal of Clinical Oncology, 20, 7, 1932-1941.*



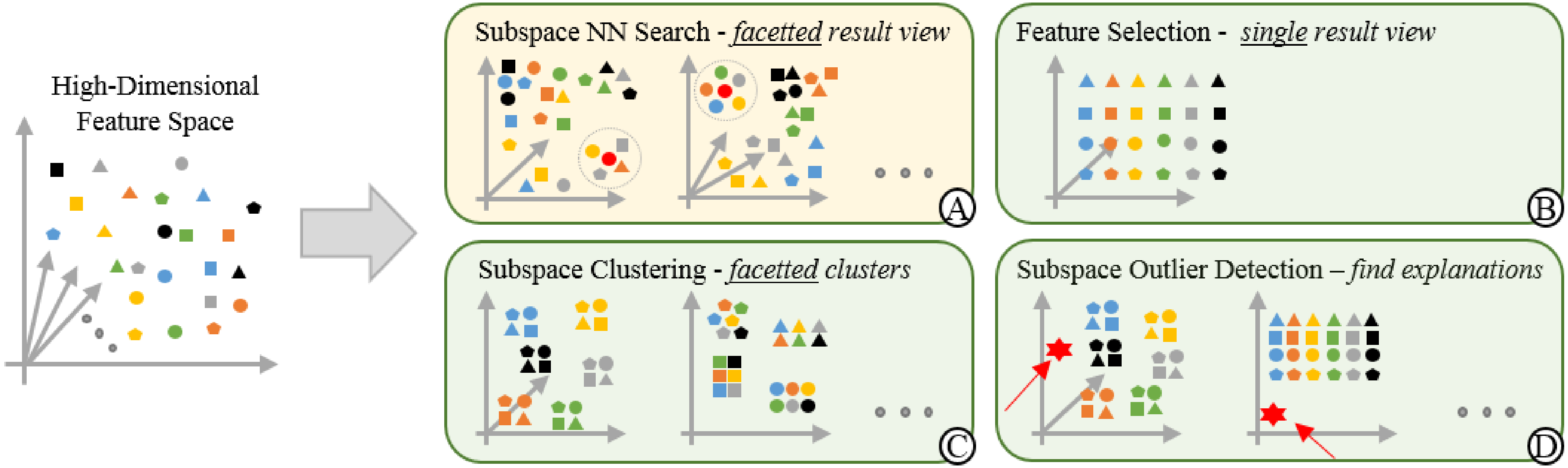
- Typical questions include:
  - Is this protein functioning as an enzyme?
  - Does this gene sequence contain a splice site?
  - Is this melanoma malign?
- Given object x – predict the class label y
  - If $y \in \{0,1\} \rightarrow$ binary classification problem
  - If $y \in \{1, \ldots, n\}$ and is $n \in \mathbb{N} \rightarrow$ multiclass problem
  - If $y \in \mathbb{R} \rightarrow$ regression problem

Why do we need Clustering?



- Group similar objects into clusters together, e.g.
  - For image segmentation
  - Grouping genes similarly affected by a disease
  - Clustering patients with similar diseases
  - Cluster biological samples for category discovery
  - Finding subtypes of diseases
  - Visualizing protein families
- Inference: given x_i, predict y_i by learning f
- No training data set – learn model and apply it

Why do we need Subspace Clustering in medical data?



Hund, M., Sturm, W., Schreck, T., Ullrich, T., Keim, D., Majnaric, L. & Holzinger, A. 2015. Analysis of Patient Groups and Immunization Results Based on Subspace Clustering. In: Guo, Y., Friston, K., Aldo, F., Hill, S. & Peng, H. (eds.) Brain Informatics and Health, Lecture Notes in Artificial Intelligence LNAI 9250. Cham: Springer International Publishing, pp. 358-368, doi:10.1007/978-3-319-23344-4_35.

Lecture 8: Decision Making under Uncertainty: Decision Support Systems

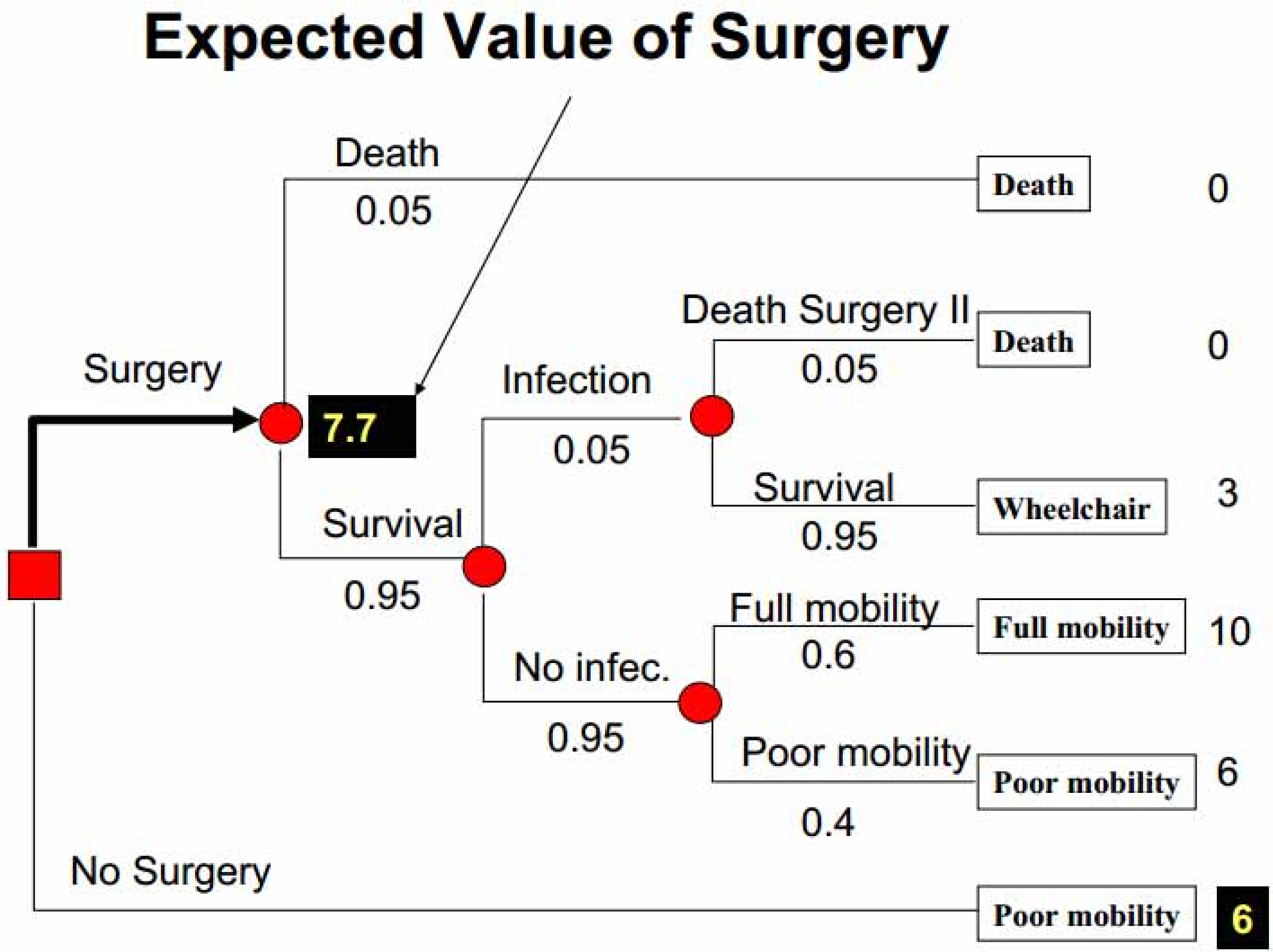
Two types of decisions (Diagnosis vs. Therapy)

- **Type 1 Decisions:** related to the **diagnosis,** i.e. computers are used to assist in diagnosing a disease on the basis of the individual patient data. Questions include:
  - What is the probability that this patient has a myocardial infarction on the basis of given data (patient history, ECG, …)?
  - What is the probability that this patient has acute appendices, given the signs and symptoms concerning abdominal pain?

- **Type 2 Decisions:** related to **therapy,** i.e. computers are used to select the best therapy on the basis of clinical evidence, e.g.:
  - What is the best therapy for patients of age x and risks y, if an obstruction of more than z % is seen in the left coronary artery?
  - What amount of insulin should be prescribed for a patient during the next 5 days, given the blood sugar levels and the amount of insulin taken during the recent weeks?

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

Helps to make rational decisions (risks vs. success)



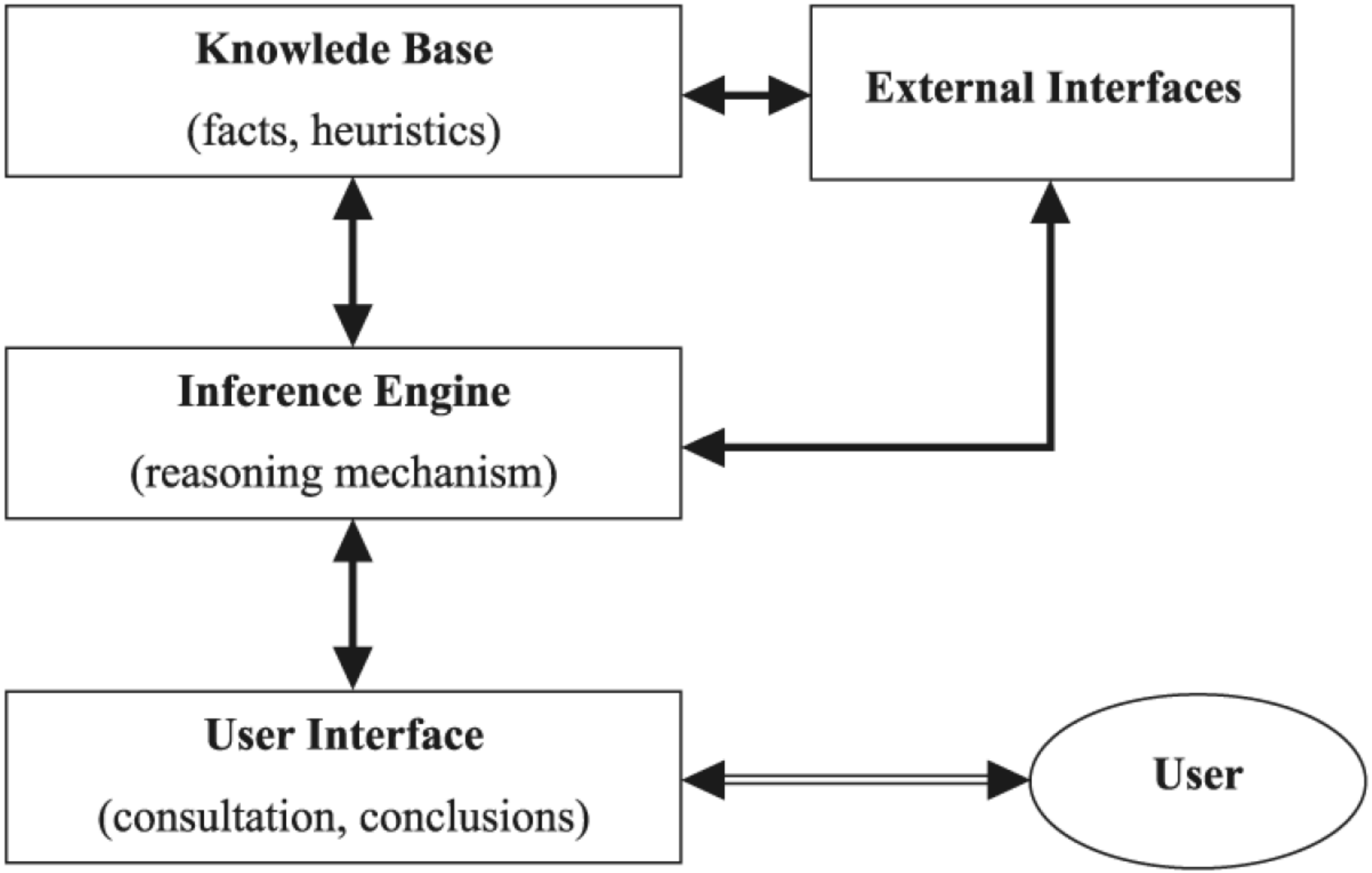
Original Example from MYCIN

h_1 = The identity of ORGANISM-1 is streptococcus
h_2 = PATIENT-1 is febrile
h_3 = The name of PATIENT-1 is John Jones

$CF[h_1,E] = .8$	:	There is strongly suggestive evidence (.8) that the identity of ORGANISM-1 is streptococcus
$CF[h_2,E] = -.3$	:	There is weakly suggestive evidence (.3) that PATIENT-1 is not febrile
$CF[h_3,E] = +1$	:	It is definite (1) that the name of PATIENT-1 is John Jones

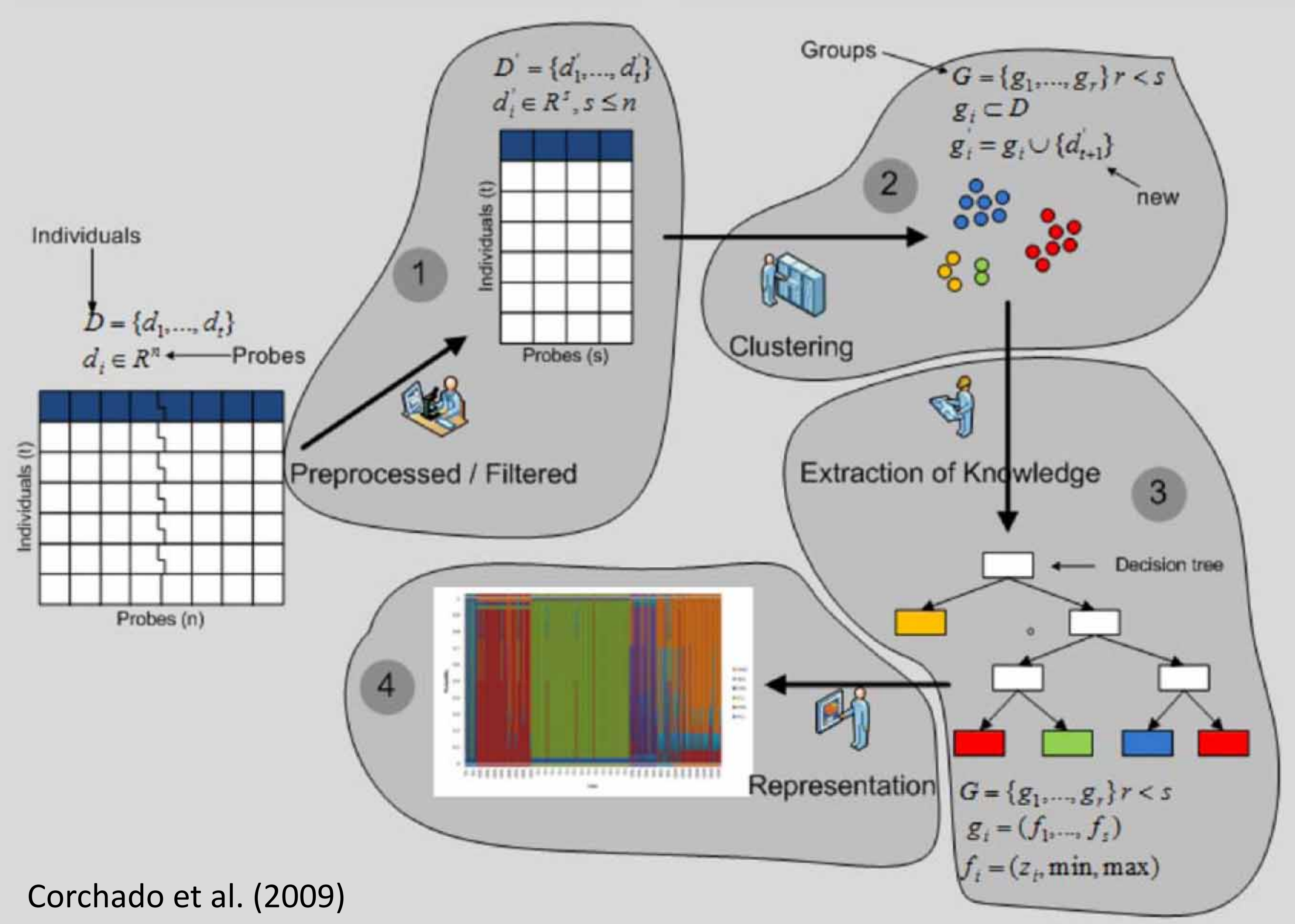
Shortliffe, E. H. & Buchanan, B. G. (1984) *Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project. Addison-Wesley.*

State-of-the-art architecture of DSS



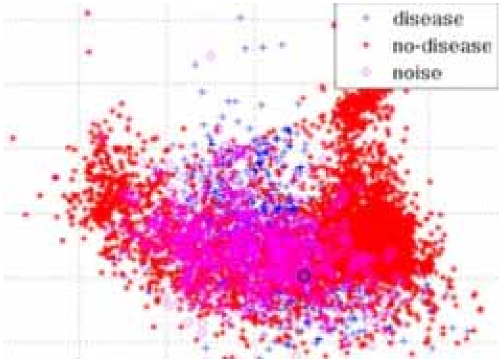
Metaxiotis, K. & Psarras, J. (2003) Expert systems in business: applications and future directions for the operations researcher. *Industrial Management & Data Systems, 103, 5, 361-368.*

Why is this model so interesting for us?



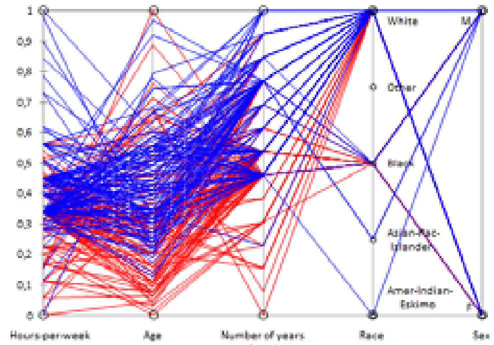
Corchado et al. (2009)

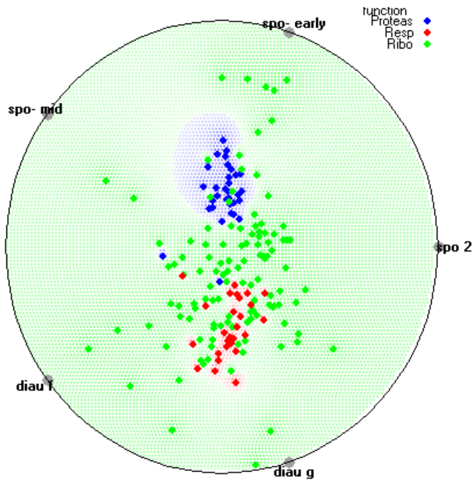
Lecture 9: Interactive Visualization and Visual Analytics



Scatterplot = oldest, point-based technique, projects data from n-dim space to an arbitrary k-dim display space;

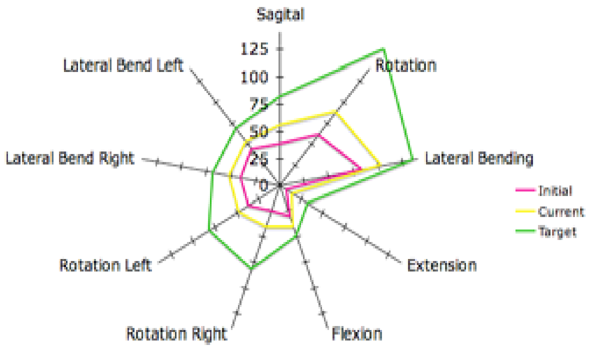
Parallel coordinates = (PCP), originally for the study of high-dimensional geometry, data point plotted as polyline;



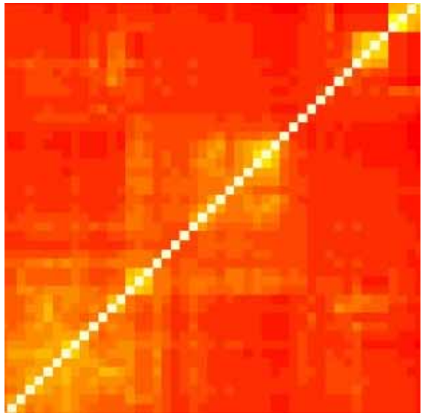


RadViz = Radial Coordinate visualization, is a “force-driven” point layout technique, based on Hooke’s law for equilibrium;

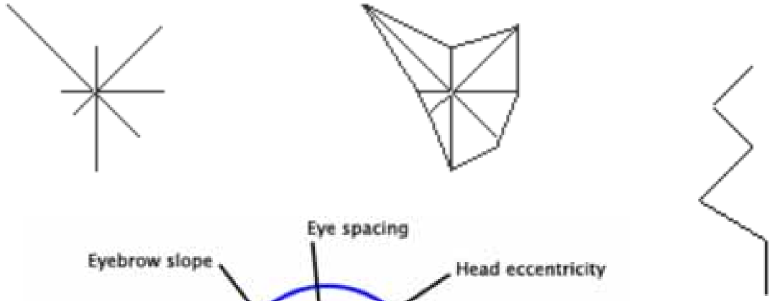
Slide 9-23 Visualizations for multivariate data Overview 2/2



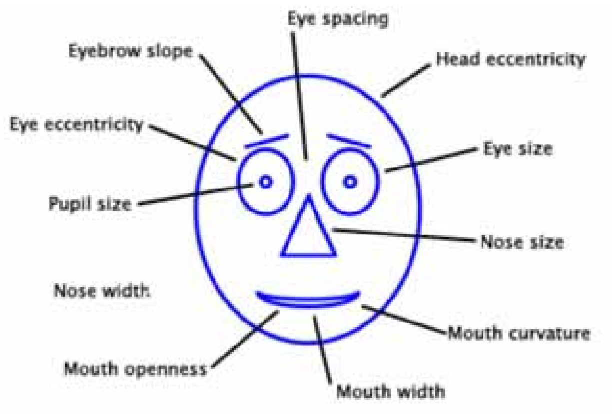
Radar chart (star plot, spider web, polar graph, polygon plot) = radial axis technique;



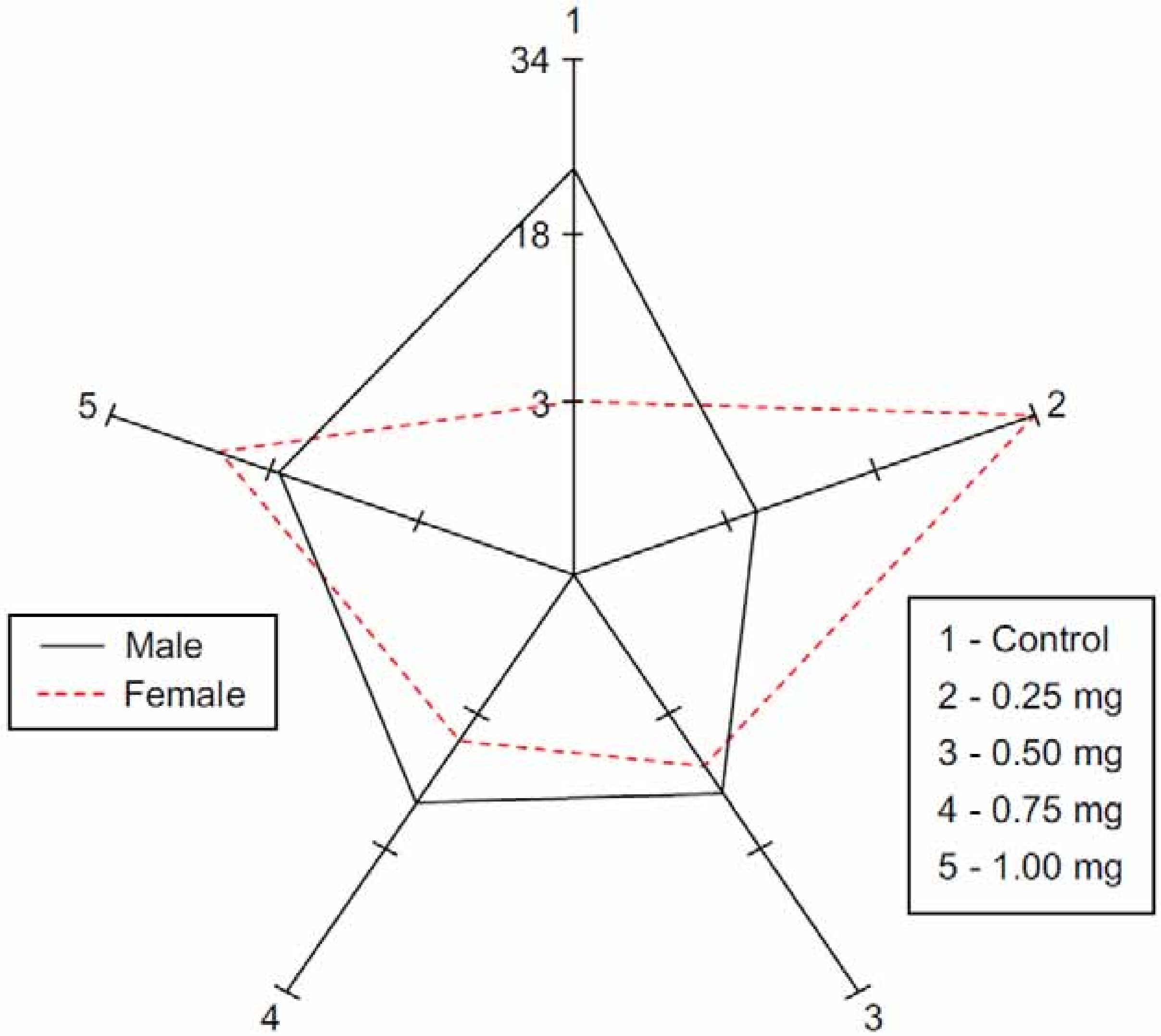
Heatmap = a tabular display technique using color instead of figures for the entries;



Glyph = a visual representation of the entity, where its attributes are controlled by data attributes;



Chernoff face = a face glyph which displays multivariate data in the shape of a human face



Saary, M. J. (2008) Radar plots: a useful way for presenting multivariate health care data. *Journal Of Clinical Epidemiology, 61, 4, 311-317.*

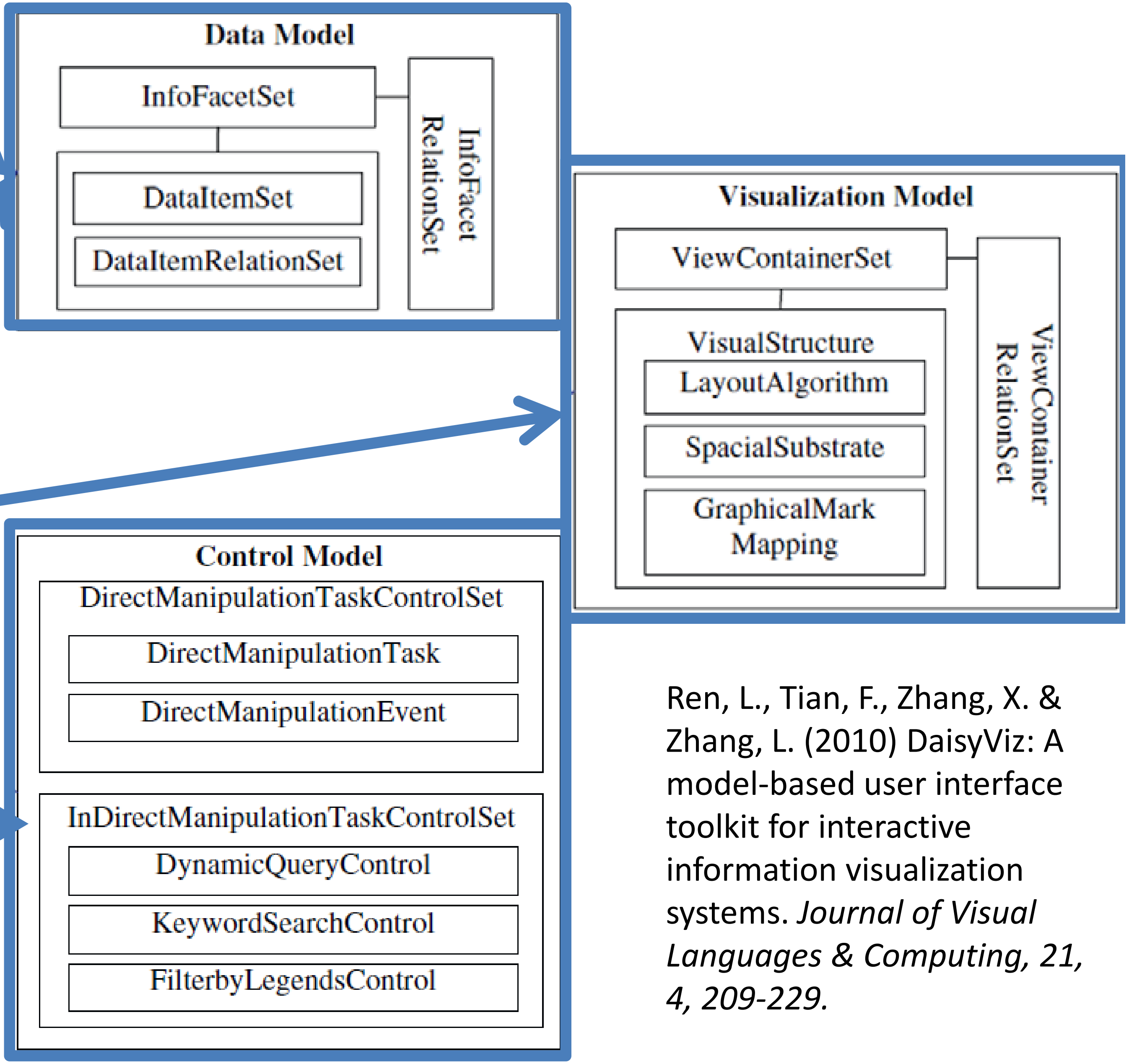
Slide 9-40 Design of Interactive Information Visualization

1) What facets of the target information should be visualized?

2) What data source should each facet be linked to and what relationships these facets have?

3) What layout algorithm should be used to visualize each facet?

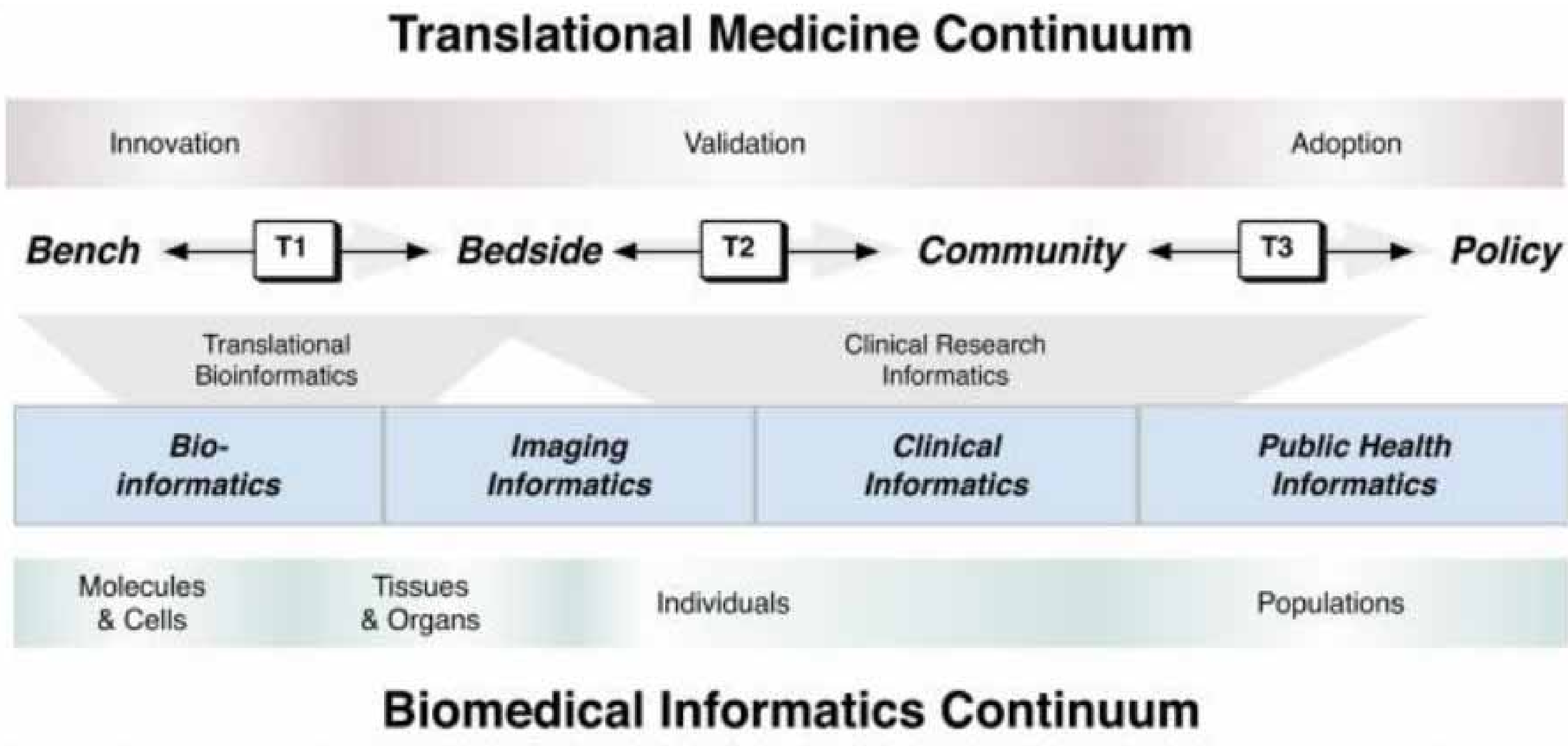
4) What interactive techniques should be used for each facet and for which infovis tasks?



Ren, L., Tian, F., Zhang, X. & Zhang, L. (2010) DaisyViz: A model-based user interface toolkit for interactive information visualization systems. *Journal of Visual Languages & Computing, 21, 4, 209-229.*

Lecture 10: Biomedical Information Systems and Knowledge Management

Translational Health Informatics Continuum



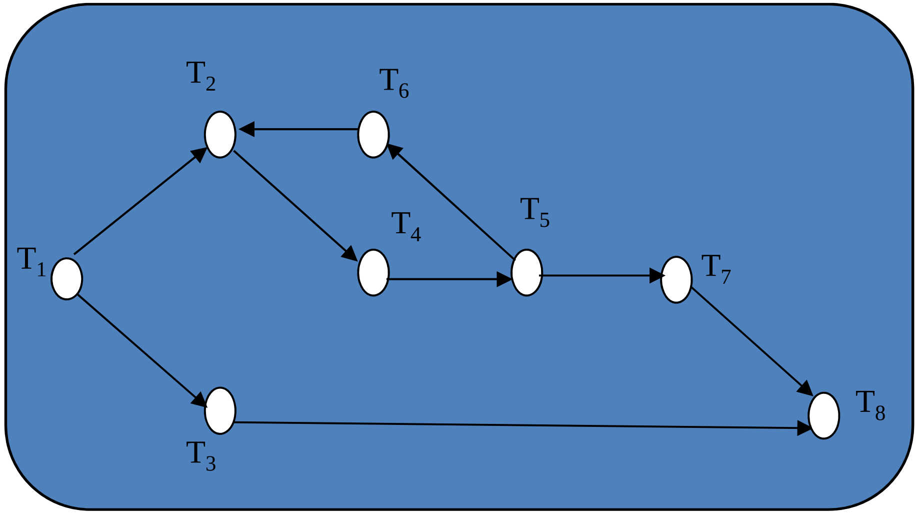
Sarkar, I. 2010. Biomedical informatics and translational medicine. Journal of Translational Medicine, 8, (1), 2-12.

Slide 10-6 Example: Formal workflow modeling 1/2

- A workflow is defined as a process that contains tasks T, and the respective rules on how those tasks are executed:
- Workflow W:= (T, P, C, A, S_0) where
  - T = {T_1, T_2, … T_m} A set of tasks, m≥ 1
  - P = $(p_{ij})_{m\,x\,m}$ **Precedence matrix of the task set**
  - C = $(c_{ij})_{m\,x\,m}$ **Conflict matrix of the task set**
  - A = (A(T_1), A(T_2), …,A(T_m)) Pre-Condition set for each task
  - $S_0 \in \{0, 1, 2, 3\}_m$ is the initial state

J. Wang, D. Rosca, W. Tepfenhart & A. Milewski (2006) Dynamic Workflow Modeling and Analysis, Monmouth University

Slide 10-7 Example: Formal workflow modeling 2/2



$T = \{T_1, T_2, \ldots, T_8\}$,

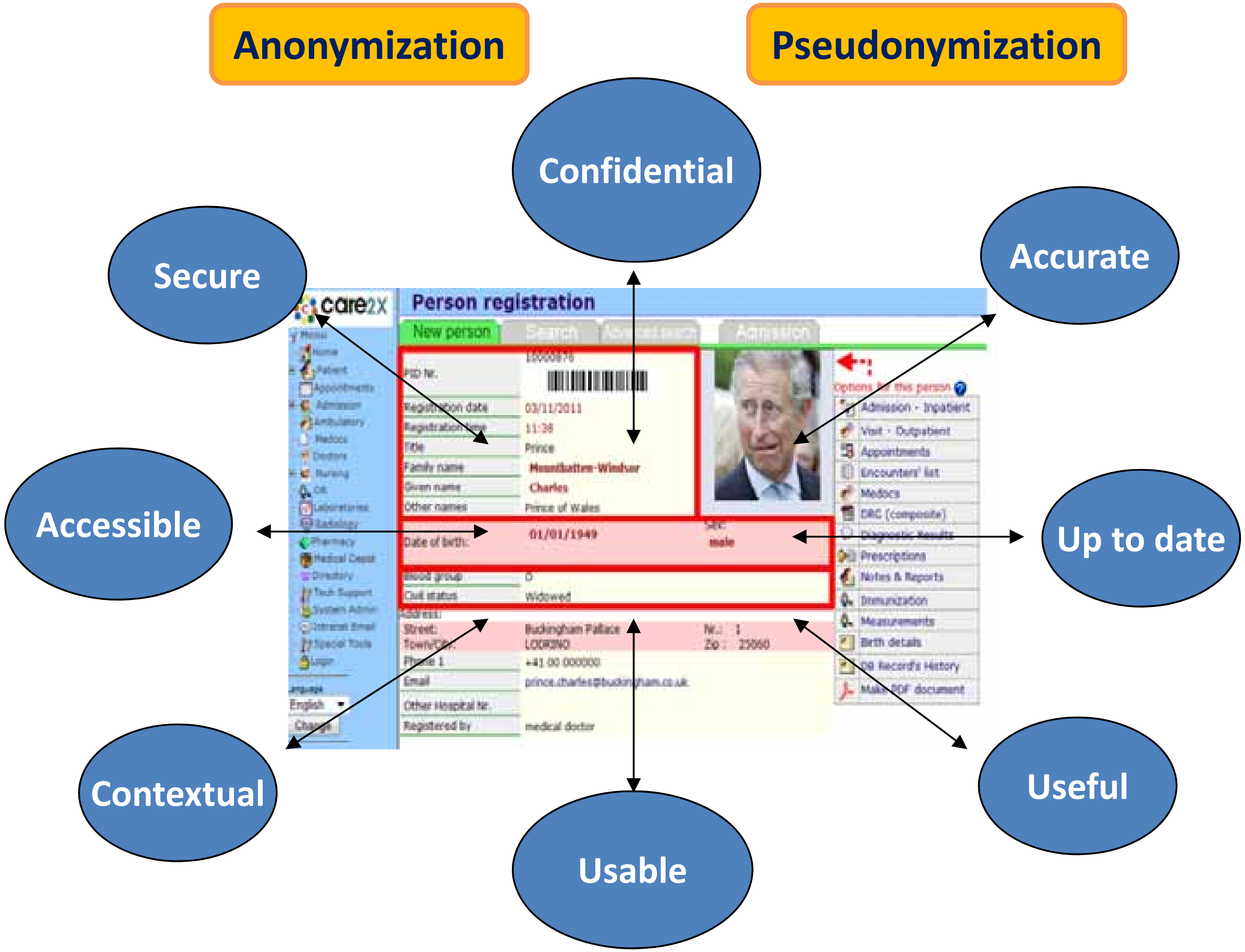
$A(T_1) = \emptyset$, $A(T_2) = \{\{T_1\}, \{T_6\}\}$, $A(T_3) = \{\{T_1\}\}$,
$A(T_4) = \{\{T_2\}\}$, $A(T_5) = \{\{T_4\}\}$,
$A(T_6) = A(T_7) = \{\{T_5\}\}$, $A(T_8) = \{\{T_3, T_7\}\}$.
$S_0 = (1, 0, 0, 0, 0, 0, 0, 0)$.

$$P = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \qquad C = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

J. Wang, D. Rosca, W. Tepfenhart & A. Milewski (2006) Dynamic Workflow Modeling and Analysis, Monmouth University

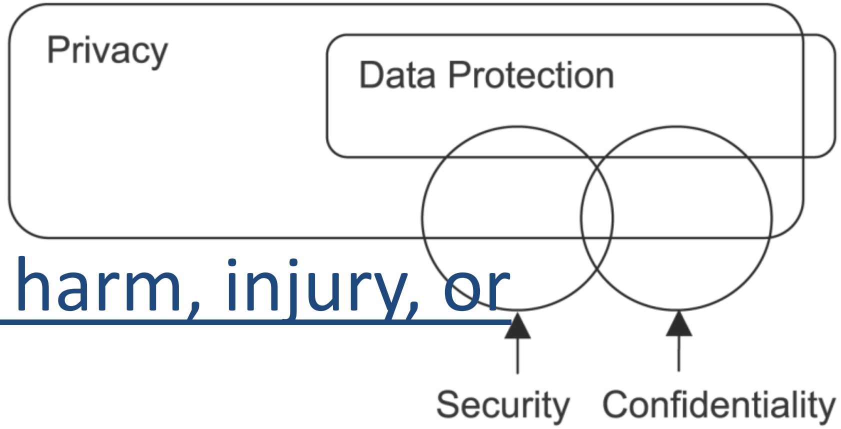
Lecture 11:
Privacy, Data Protection, Safety, Security & Privacy Aware Machine Learning

What are the Requirements of an electronic patient record



Anonymization: Personal data cannot be re-identified (e.g. k-Anonymization)
Pseudonymization: The personal data is replaced by a "pseudonym“, which allows later tracking back to the source data (re-identification)

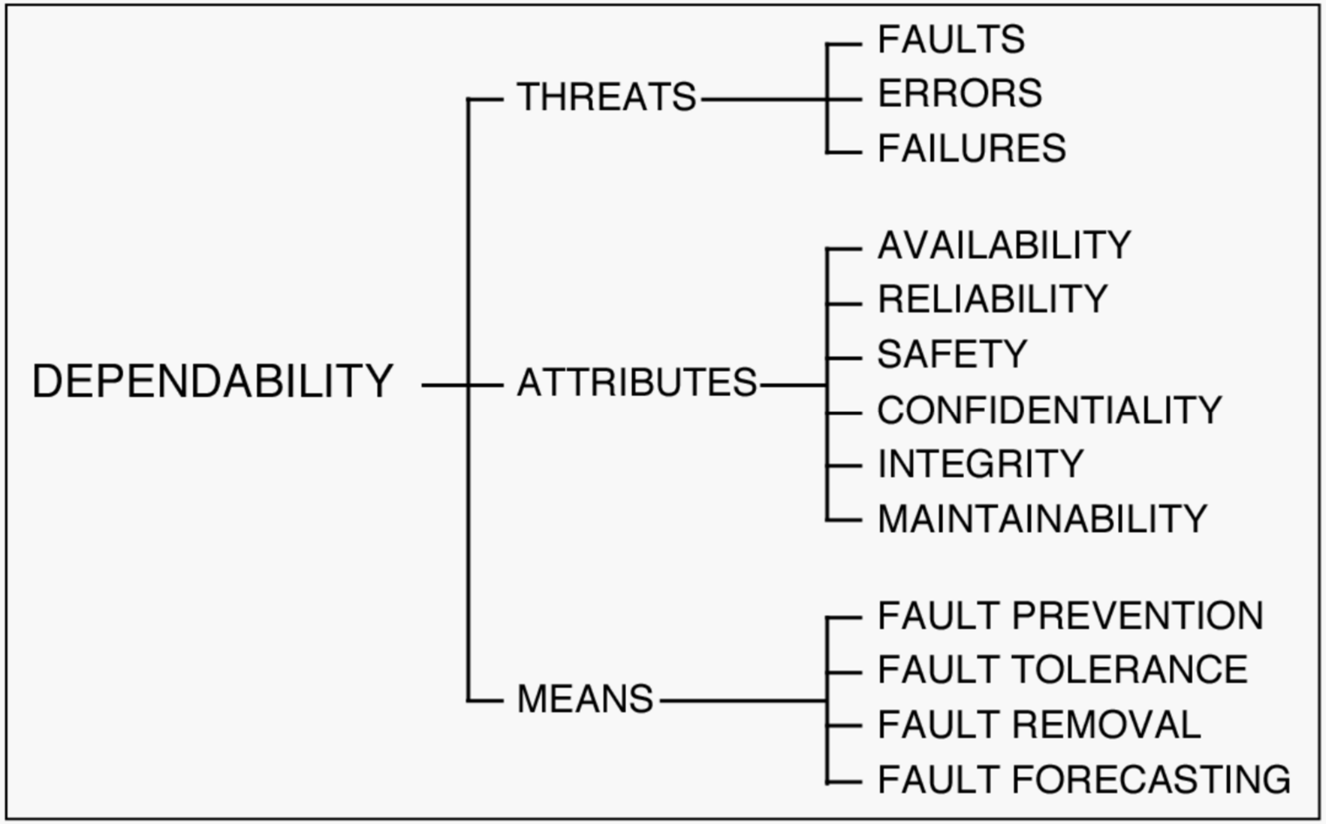
Which Definitions do we need to know?



- **Safety** = any <u>protection from harm, injury, or damage</u>;
- Data Protection = all measures to ensure availability and integrity of data
- **Privacy** = (US pron. "prai …"; UK pron. "pri …"; from Latin: privatus "separated from the rest", are the individual <u>rights of people</u> to protect their personal life and matters Confidentiality = secrecy ("ärztliche Schweigepflicht")

Mills, K. S., Yao, R. S. & Chan, Y. E. (2003) Privacy in Canadian Health Networks: challenges and opportunities. *Leadership in Health Services, 16, 1, 1-10.*

Safety, Security -> Technical Dependability



Avizienis, A., Laprie, J. C. & Randell, B. (2001) Fundamental concepts of dependability. *Technical Report Computing Science University of Newcastle, 1145, CS-TR-739, 7-12.*

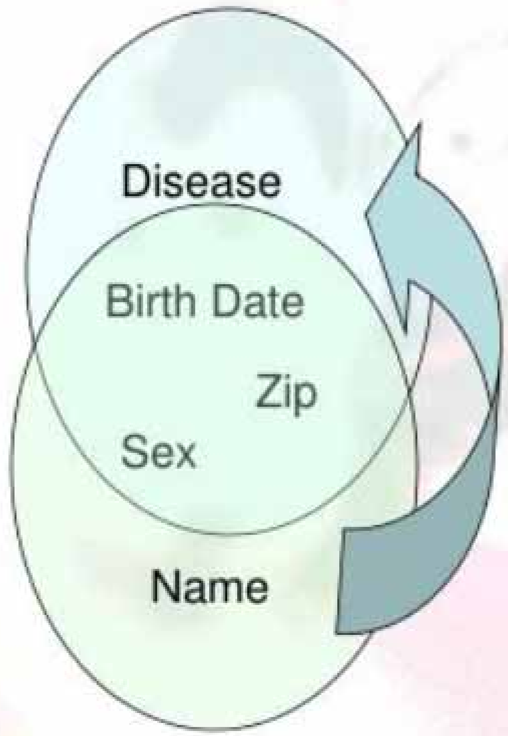
87 % of the population in the USA can be uniquely re-identified by Zip-Code, Gender and date of birth

Hospital Patient Data

Birthdate	Sex	Zipcode	Disease
1/21/76	Male	53715	Flu
4/13/86	Female	53715	Hepatitis
2/28/76	Male	53703	Brochitis
1/21/76	Male	53703	Broken Arm
4/13/86	Female	53706	Sprained Ankle
2/28/76	Female	53706	Hang Nail

Voter Registration Data

Name	Birthdate	Sex	Zipcode
Andre	1/21/76	Male	53715
Beth	1/10/81	Female	55410
Carol	10/1/44	Female	90210
Dan	2/21/84	Male	02174
Ellen	4/19/72	Female	02237



Samarati, P. 2001. Protecting respondents identities in microdata release. IEEE Transactions on Knowledge and Data Engineering, 13, (6), 1010-1027, doi:10.1109/69.971193.

Sweeney, L. 2002. Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10, (05), 571-588.

- **K-Anonymity** … a release of data is said to have the *k-anonymity property* if the information for each person contained in the release cannot be distinguished from at least $k-1$ individuals whose information also appear in the release.
- **L-Diversity** … extension requiring that the values of all confidential attributes within a group of k sets contain at least L clearly distinct values
- **t-Closeness** … extension requiring that the distribution of the confidential attribute within a group of k records is similar to the confidential attribute's distribution in the whole data set (local distribution must resemble the global distribution)

Example Privacy Aware Machine Learning (PAML)

The Right to Be Forgotten: Towards Machine Learning on Perturbed Knowledge Bases

Bernd Malle[1,2], Peter Kieseberg[1,2], Edgar Weippl[2], and Andreas Holzinger[1](✉)

[1] Holzinger Group HCI-KDD, Institute for Medical Informatics, Statistics and Documentation, Medical University Graz, Graz, Austria
{b.malle,a.holzinger}@hci-kdd.org
[2] SBA Research gGmbH, Favoritenstrae 16, 1040 Vienna, Austria
PKieseberg@sba-research.org

Abstract. Today's increasingly complex information infrastructures represent the basis of any data-driven industries which are rapidly becoming the 21st century's economic backbone. The sensitivity of those infrastructures to disturbances in their knowledge bases is therefore of crucial interest for companies, organizations, customers and regulating bodies. This holds true with respect to the direct provisioning of such information in crucial applications like clinical settings or the energy industry, but also when considering additional insights, predictions and personalized services that are enabled by the automatic processing of those data. In the light of new EU Data Protection regulations applying from 2018 onwards which give customers the right to have their data deleted on request, information processing bodies will have to react to these changing jurisdictional (and therefore economic) conditions. Their choices include a re-design of their data infrastructure as well as preventive actions like anonymization of databases per default. Therefore, insights into the effects of perturbed/anonymized knowledge bases on

Malle, B., Kieseberg, P., Weippl, E. & Holzinger, A. 2016. The right to be forgotten: Towards Machine Learning on perturbed knowledge bases. Springer Lecture Notes in Computer Science LNCS 9817. Heidelberg, Berlin, New York: Springer, pp. 251-256, doi:10.1007/978-3-319-45507-5_17.

Keywords: Machine learning · Knowledge bases · Right to be forgotten · Perturbation · Anonymization · k-anonymity · SaNGreeA ·

Very recent: News from 28.01.2017

Europäischer Datenschutz in der Big-Data-Welt

28.01.2017 12:30 Uhr – Monika Ermert, dpa

vorlesen

(Bild: Håkan Dahlström CC BY 2.0
)

Datensparsamkeit ist schwierig in Big Data-Zeiten. Der Beirat der Datenschutzkonvention des Europarats hat eine Reihe von Richtlinien für Dataminer vorgelegt. Derweil warnt der SAP-Finanzchef vor Risiken von EU-Datenschutzregeln.

Der Beirat der Datenschutzkonvention des Europarat legte zum Internationalen Datenschutztag „Richtlinien zum Schutz persönlicher Daten in einer Big Data-Welt“ vor.

https://www.heise.de/newsticker/meldung/Europaeischer-Datenschutz-in-der-Big-Data-Welt-3609737.html

Good luck for the Exam