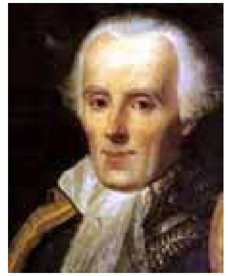
Andreas Holzinger

185.A83 Machine Learning for Health Informatics

2017S, VU, 2.0 h, 3.0 ECTS

Module 02 – Week 13

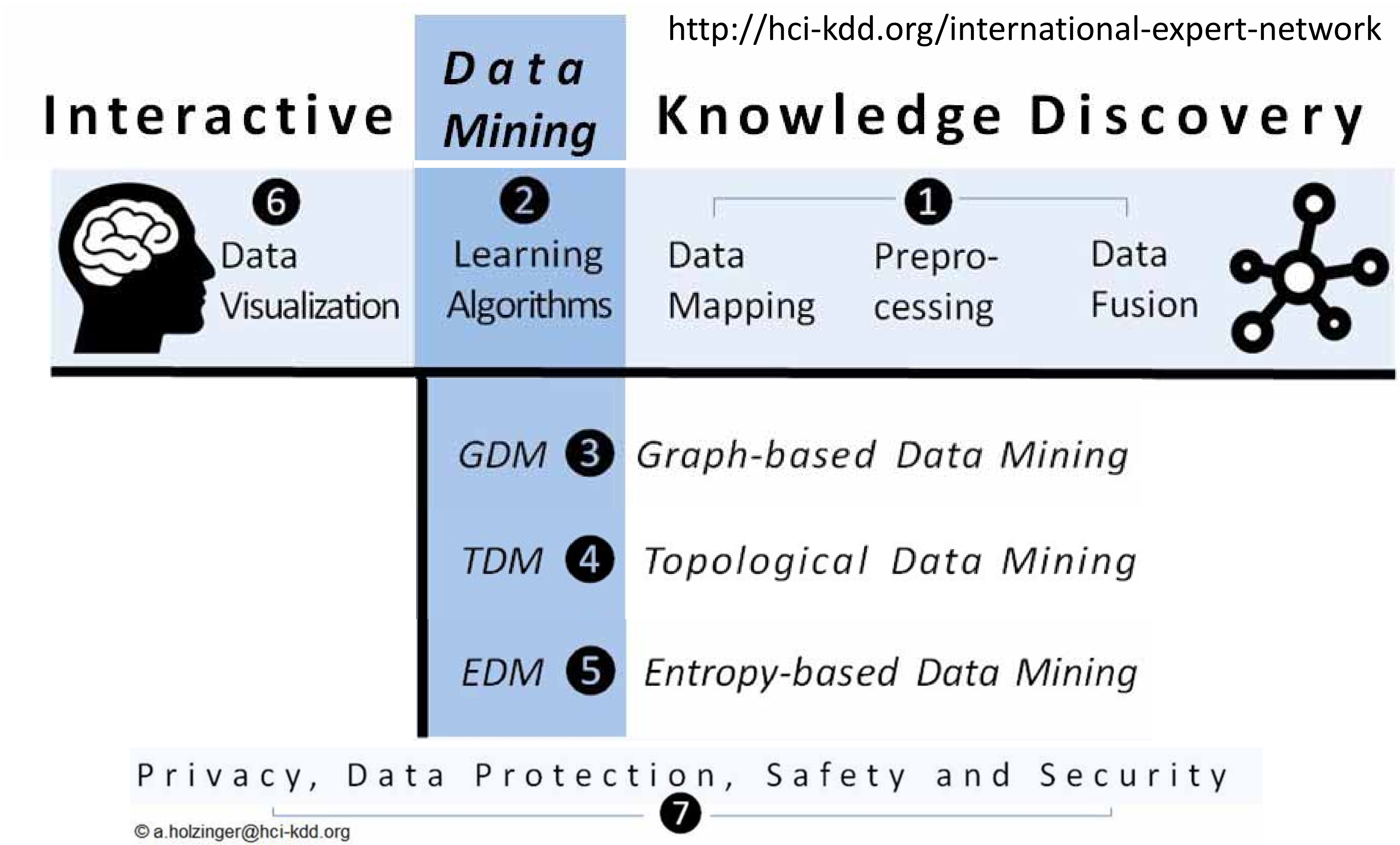


Probabilistic Graphical Models Part 1: From Decision Making under uncertainty to MCMC

a.holzinger@hci-kdd.org

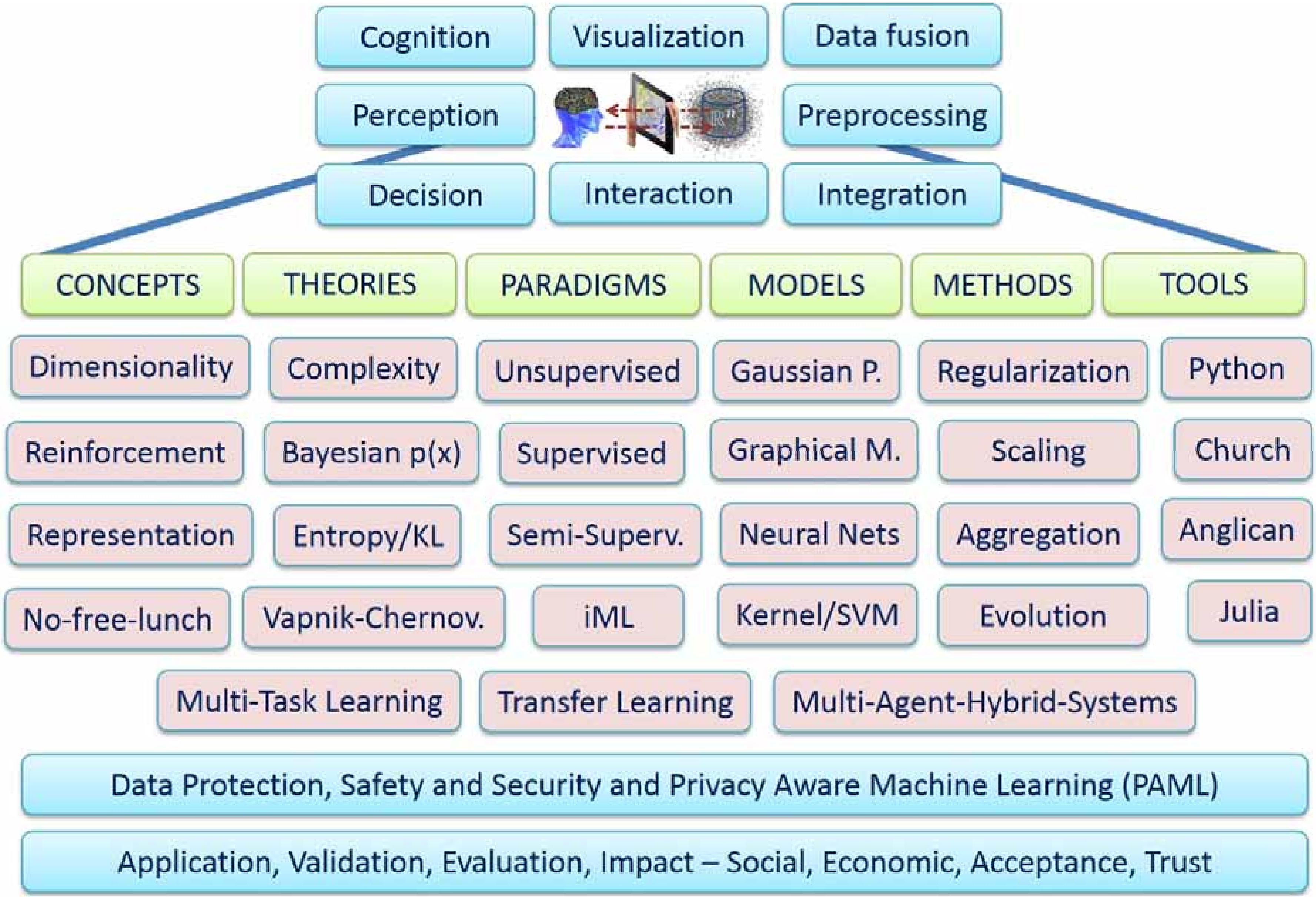
http://hci-kdd.org/machine-learning-for-health-informatics-course

ML needs a concerted effort fostering integrated research



Holzinger, A. 2014. Trends in Interactive Knowledge Discovery for Personalized Medicine: **Cognitive Science meets Machine Learning.** IEEE Intelligent Informatics Bulletin, 15, (1), 6-14.

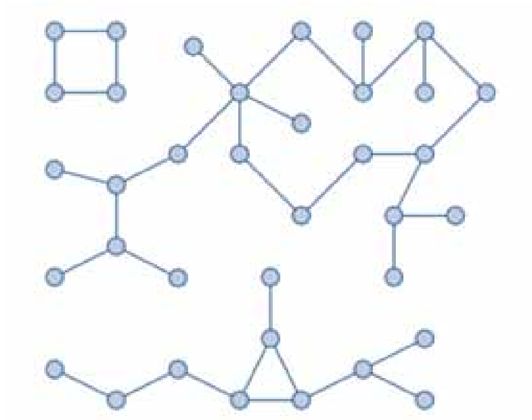
Machine Learning Jungle Top-Level View



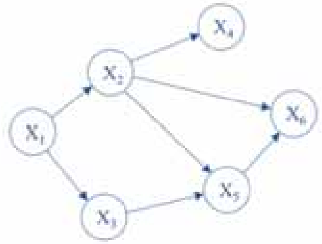
Holzinger, A. 2016. Machine Learning for Health Informatics. In: LNCS 9605, pp. 1-24, doi:10.1007/978-3-319-50478-0_1.

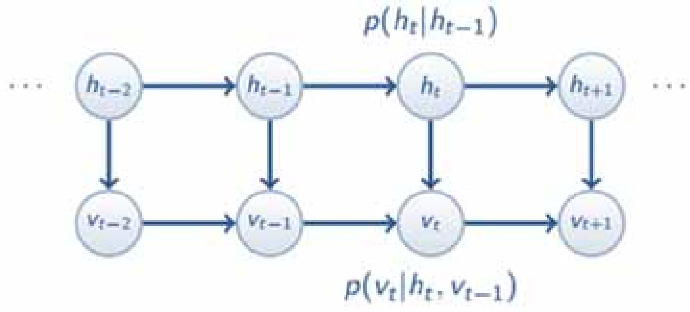
- 01 Decision Making under uncertainty
- 02 Graphs – Networks
- 03 Example Medical Knowledge Representation
- 04 Graphical Models and Decision Making
- 05 Bayes Networks
- 06 Graphical Model Learning
- 07 Probabilistic Programming
- 08 Markov Chain Monte Carlo (MCMC)
- 09 Metropolis Hastings Algorithm

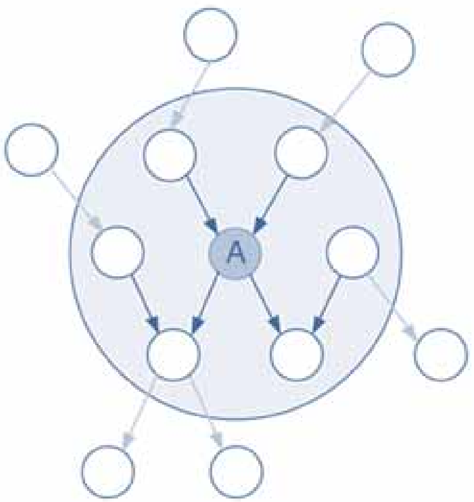

01 Reflection

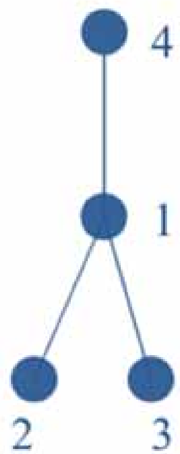


$$p(x) = \prod_i p(x_i | x_{pa_i})$$







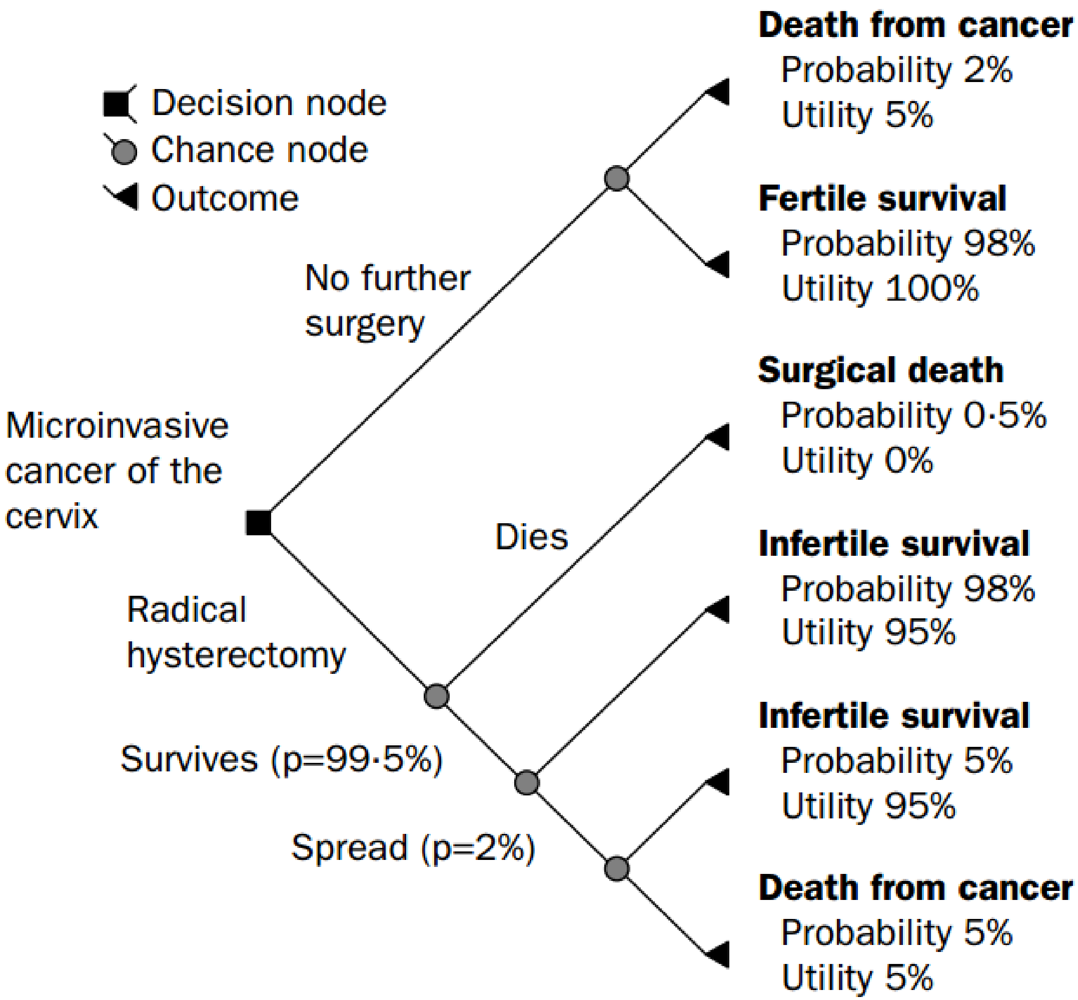


$$P(\mathbf{x}) = \prod_{i \in V} P_i(x_i) \prod_{(i,j) \in E} \frac{P_{i,j}(x_i, x_j)}{P_i(x_i) P_j(x_j)}$$

$$= P_1(x_1) P_{2|1}(x_2|x_1) P_{3|1}(x_3|x_1) P_{4|1}(x_4|x_1)$$

Graphical models are graphs where the nodes represent random variables and the links represent statistical dependencies between variables; This provides us with a tool for **reasoning under uncertainty**

Dgecision 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

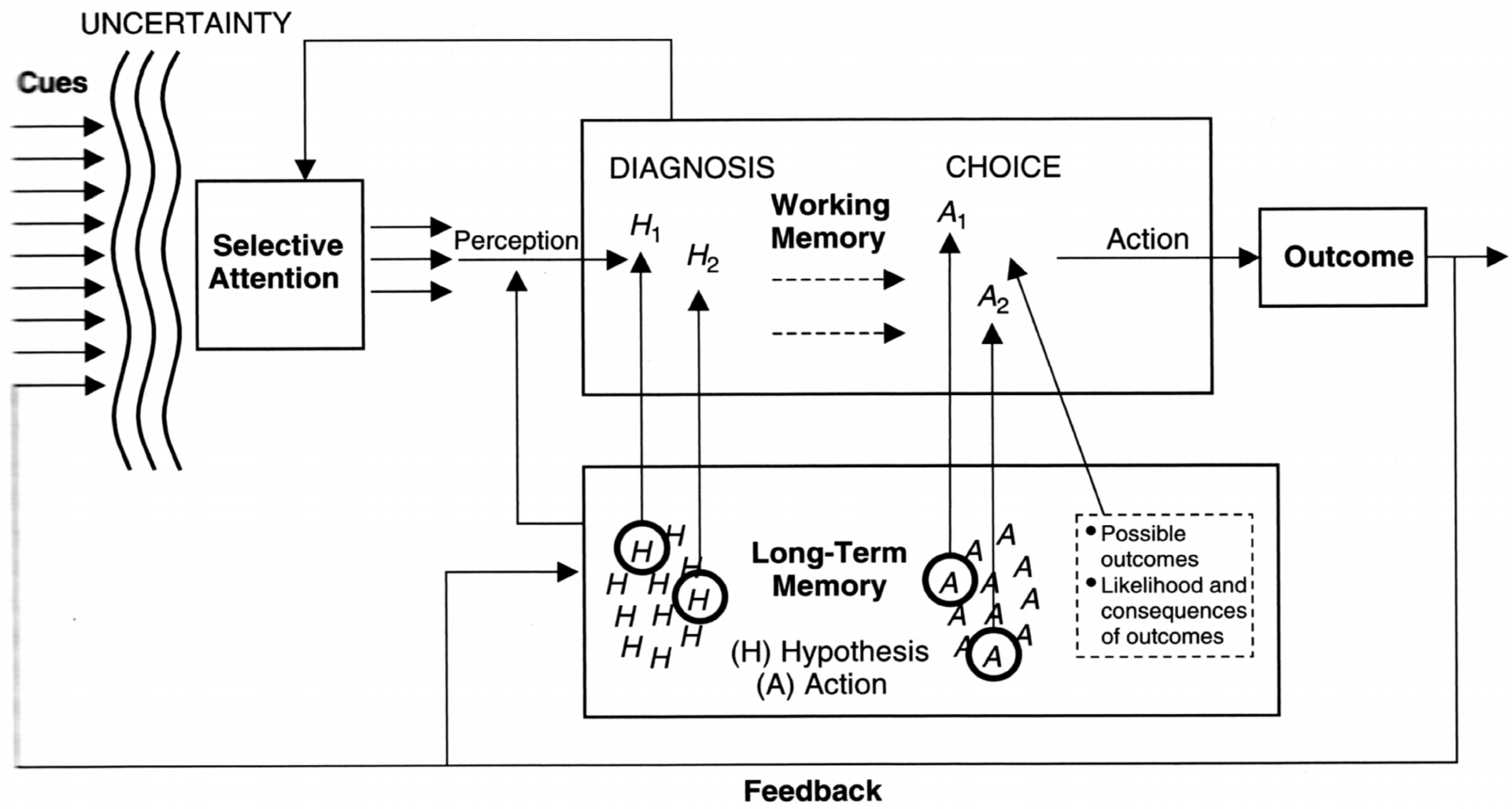
Elwyn, G., Edwards, A., Eccles, M. & Rovner, D. 2001. Decision analysis in patient care. The Lancet, 358, (9281), 571-574.

Who is Who?

01 Decision Making under uncertainty

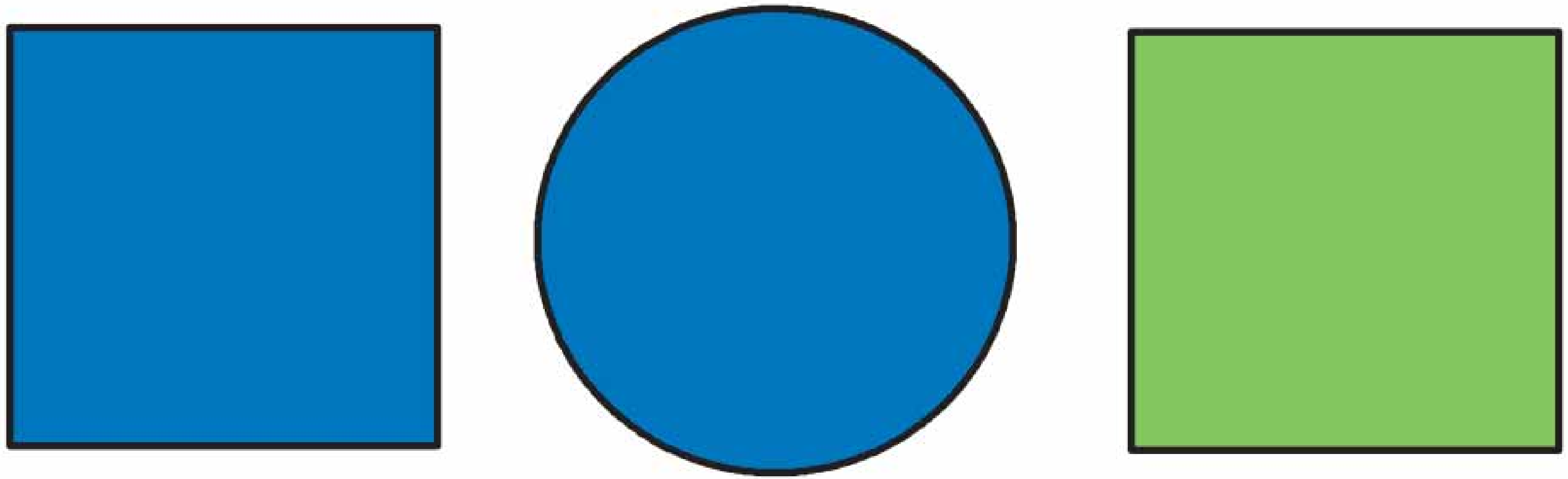
Laplace, P.-S. 1781. Mémoire sur les probabilités. *Mémoires de l'Académie Royale des sciences de Paris,* 1778, 227-332.

Medical action is …

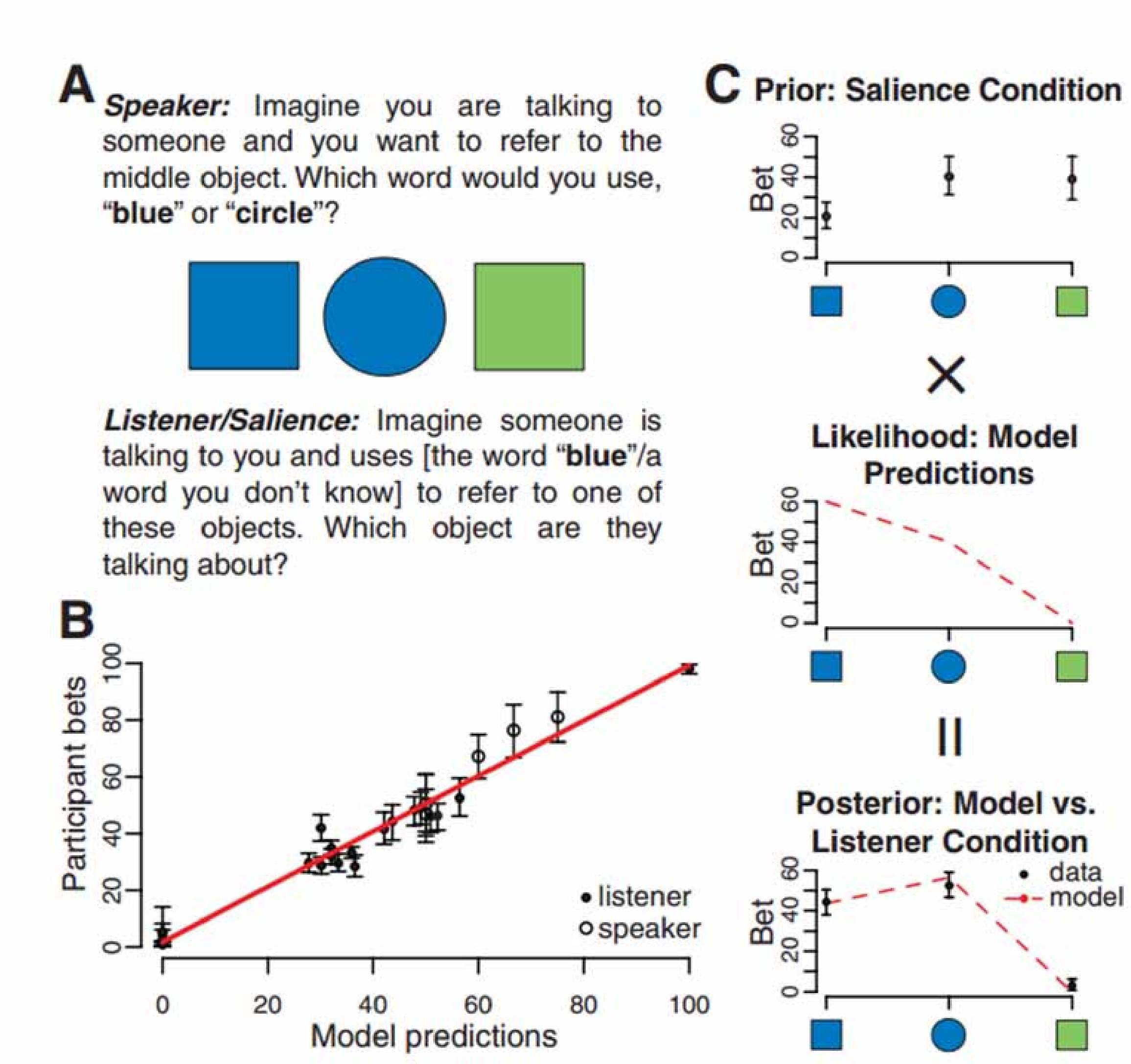


Wickens, C. D. (1984) *Engineering psychology and human performance. Columbus (OH), Charles Merrill.*

Predicting Pragmatic Reasoning in Language Games



Frank, M. C. & Goodman, N. D. 2012. Predicting pragmatic reasoning in language games. Science, 336, (6084), 998-998, doi:10.1126/science.1218633.



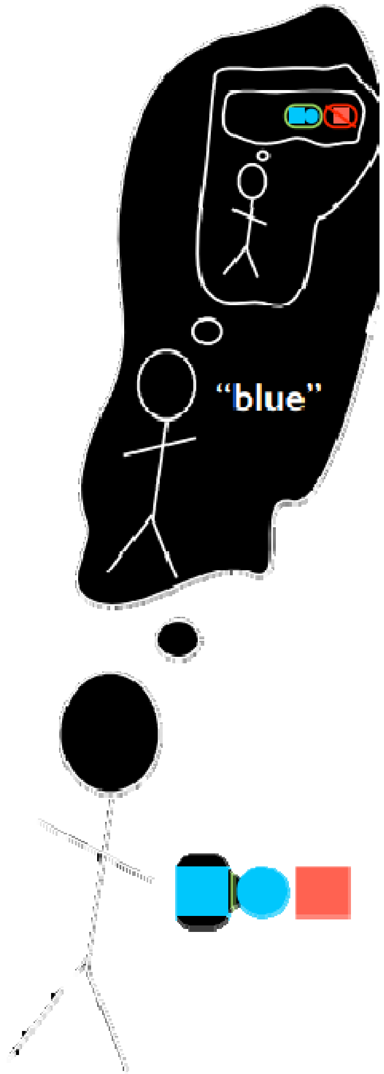
Frank, M. C. & Goodman, N. D. 2012. Predicting pragmatic reasoning in language games. Science, 336, (6084), 998-998, doi:10.1126/science.1218633.

Recursive reasoning: a case for probabilistic programming

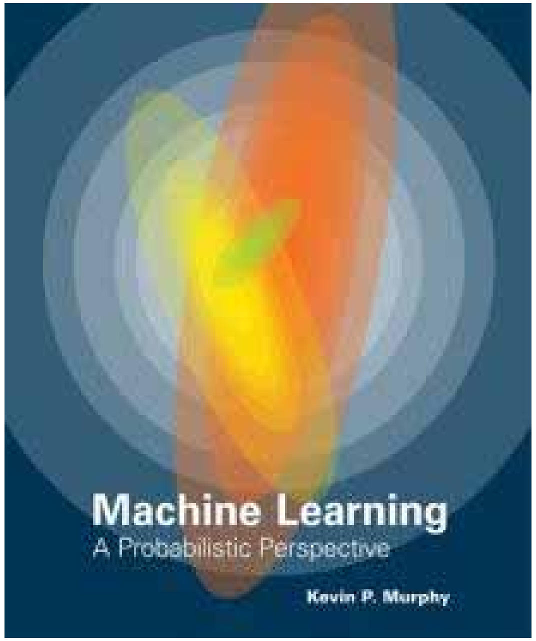

```
var literalListener = function(property){
 Infer(function(){
    var object = refPrior(context)
    condition(object[property])
    return object
```

```
var speaker = function(object) {
  Infer(function(){
    var property = propPrior()
    condition(
       object ==
```

```
var listener = function(property) {
  Infer(function(){
    var object = refPrior(context)
    condition(utterance ==
                 sample(speaker(object)))
    return object
  })}
```



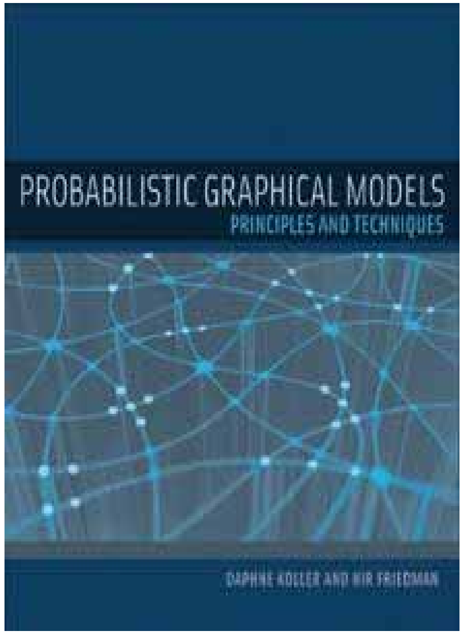
Goodman, N. D. & Frank, M. C. 2016. Pragmatic language interpretation as probabilistic inference. Trends in Cognitive Sciences, 20, (11), 818-829.



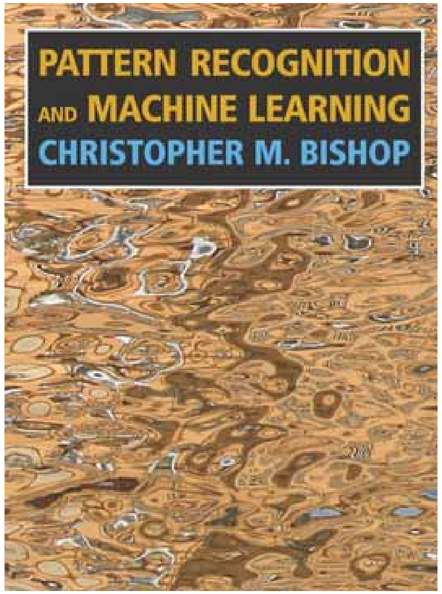
Murphy, K. P. 2012. Machine learning: a probabilistic perspective, MIT press.

Barber, D. 2012. *Bayesian reasoning and machine learning*, Cambridge University Press.

http://web4.cs.ucl.ac.uk/staff/D.Barber/textbook/181115.pdf



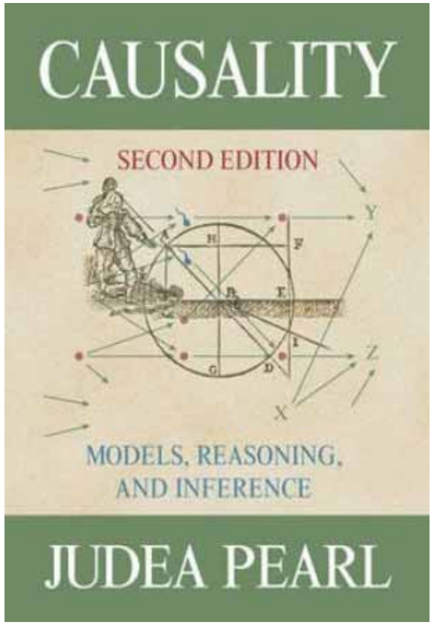
Koller, D. & Friedman, N. 2009. Probabilistic graphical models: principles and techniques, MIT press.



https://goo.gl/6a7rOC

Chapter 8 Graphical Models is as sample chapter fully downloadable for free

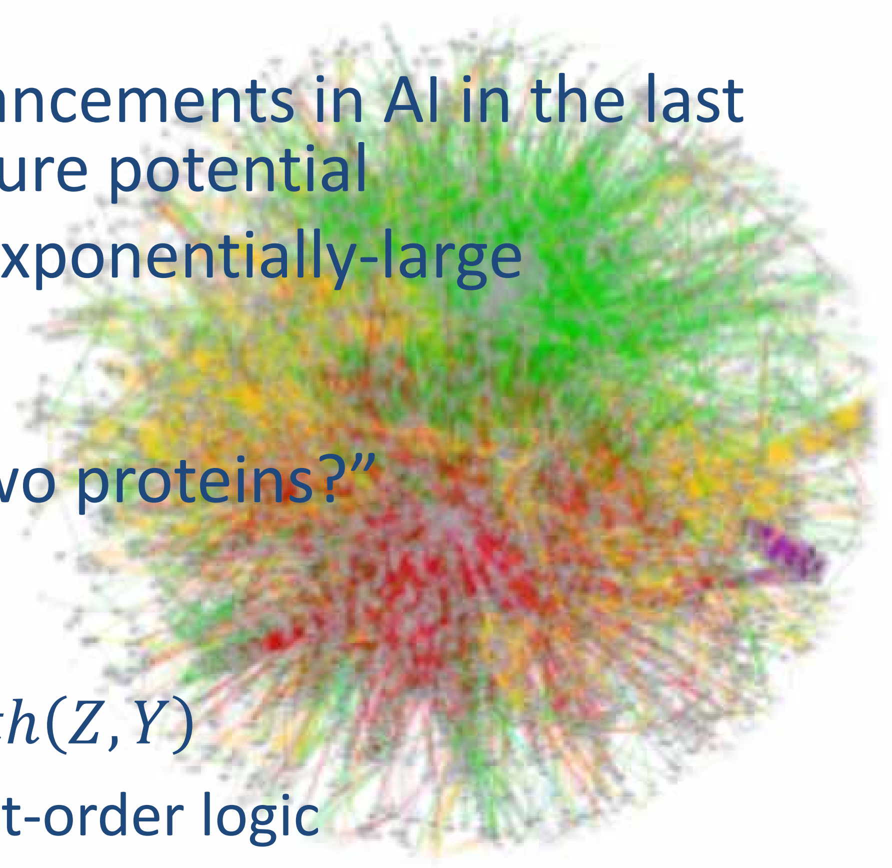
Bishop, C. M. 2006. Pattern Recognition and Machine Learning, Heidelberg, Springer.



http://bayes.cs.ucla.edu/BOOK-2K/

Pearl, J. 2009. Causality: Models, Reasoning, and Inference (2nd Edition), Cambridge, Cambridge University Press.

- PGM can be seen as a combination between
- **Graph Theory + Probability Theory + Machine Learning**
- One of the most exciting advancements in AI in the last decades – with enormous future potential
- 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

Key Challenges

- Medicine is an extremely complex application domain – dealing most of the time with uncertainties -> **probable information!**
- Key: Structure learning and prediction in large-scale biomedical networks with probabilistic graphical models
- Causality and Probabilistic Inference
- Uncertainties are present at all levels in health related systems
- Data sets from which ML learns are noisy, mislabeled, atypical, etc. etc.
- Even with data of high quality, gauging and combining a multitude of data sources and constraints in usually imperfect models of the world requires us to represent and process **uncertain knowledge** in order to make **viable decisions in context and within reasonable time!**
- In the increasingly complicated settings of modern science, model structure or causal relationships may not be known a-priori [1].
- Approximating probabilistic inference in Bayesian belief networks is NP-hard [2] -> here we need the "human-in-the-loop" [3]

[1] Sun, X., Janzing, D. & Schölkopf, B. Causal Inference by Choosing Graphs with Most Plausible Markov Kernels. ISAIM, 2006.
[2] Dagum, P. & Luby, M. 1993. Approximating probabilistic inference in Bayesian belief networks is NP-hard. Artificial intelligence, 60, (1), 141-153.
[3] Holzinger, A. 2016. Interactive Machine Learning for Health Informatics: When do we need the human-in-the-loop? Springer Brain Informatics (BRIN), 3, 1-13, doi:10.1007/s40708-016-0042-6.

02 Graphs=Networks

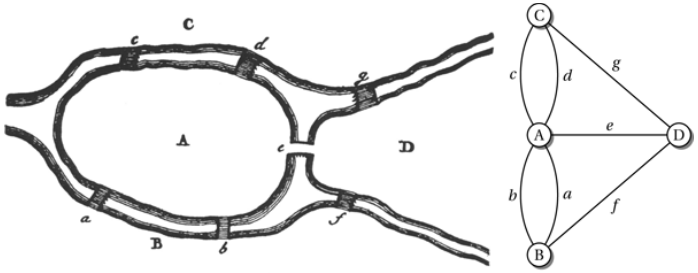
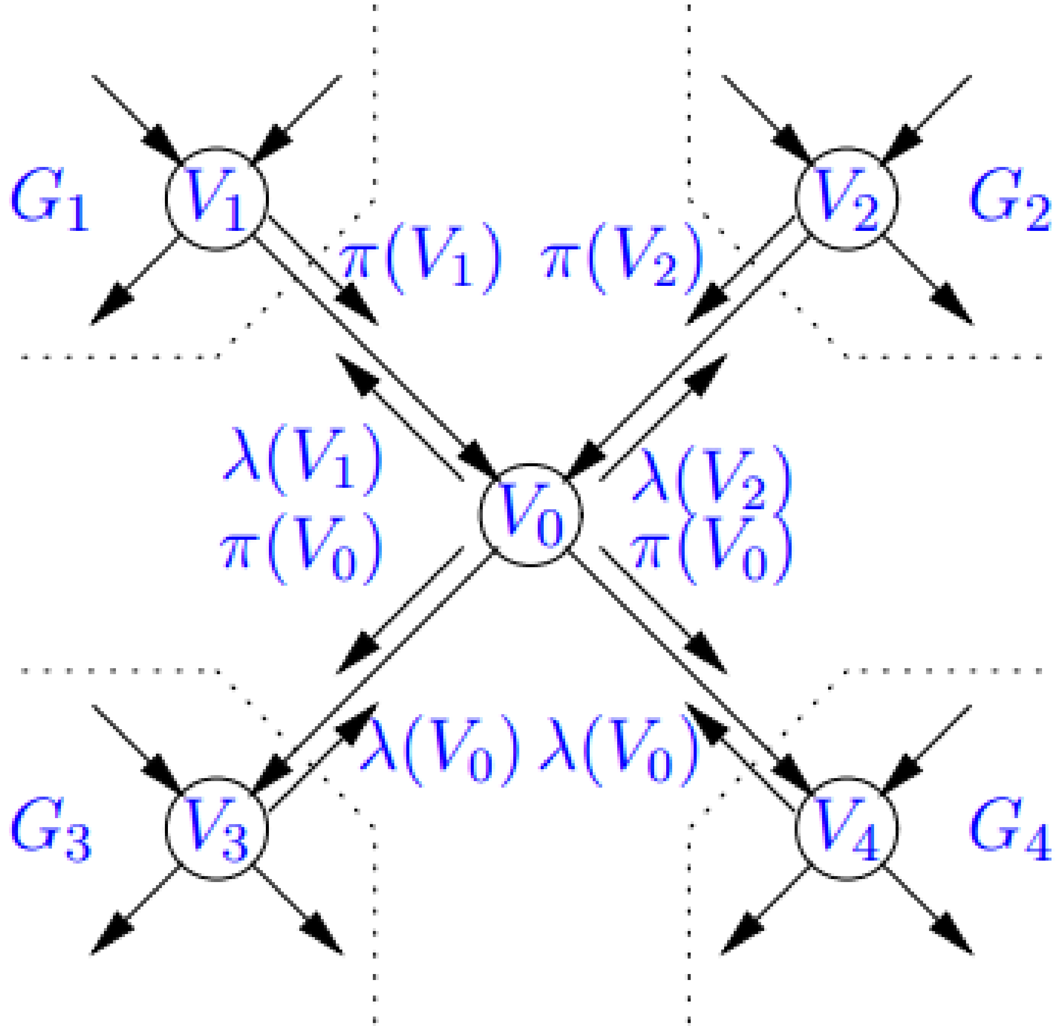


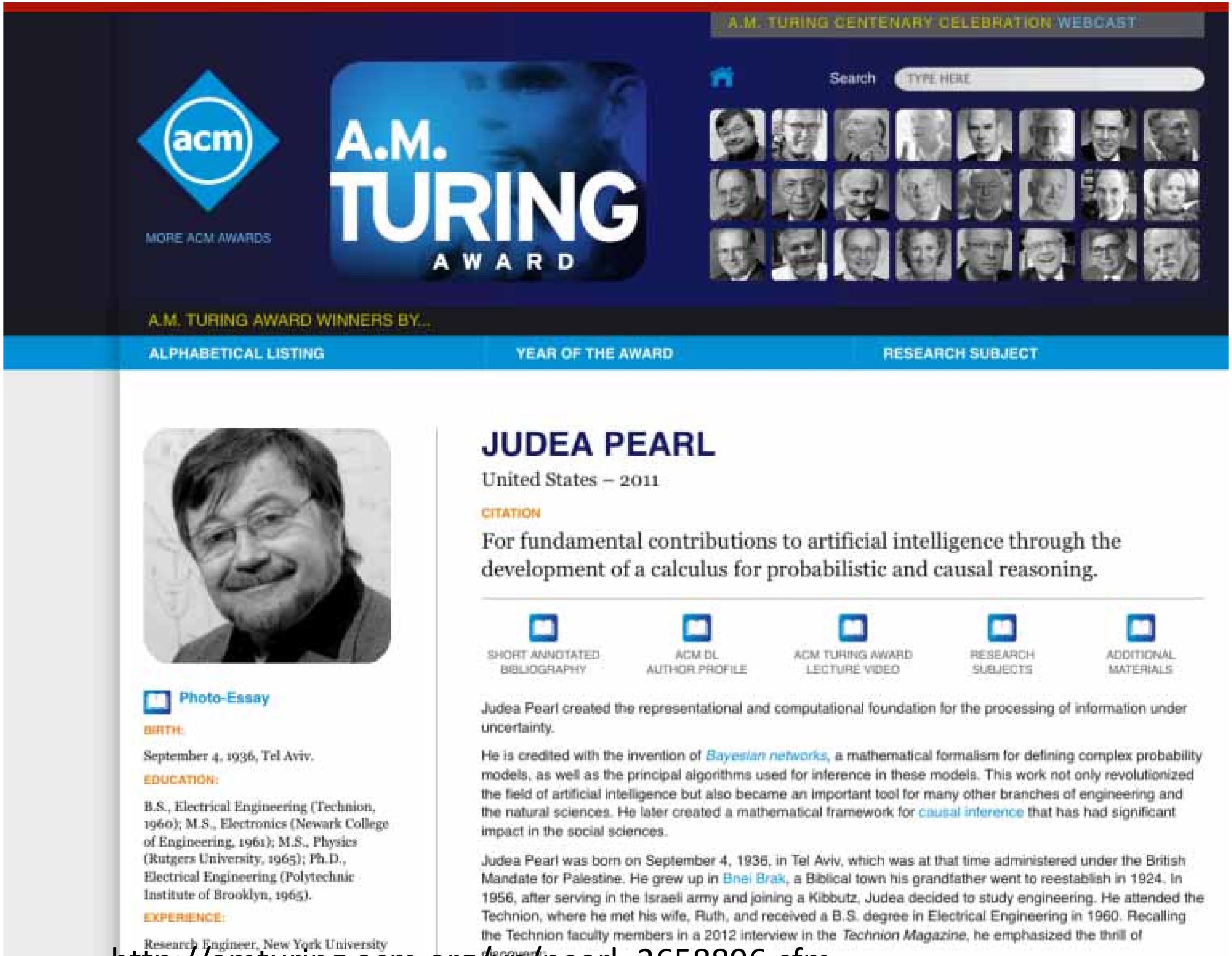
Image from https://people.kth.se/~carlofi/teaching/FEL3250-2013/courseinfo.html

252 years later: Belief propagation algorithm

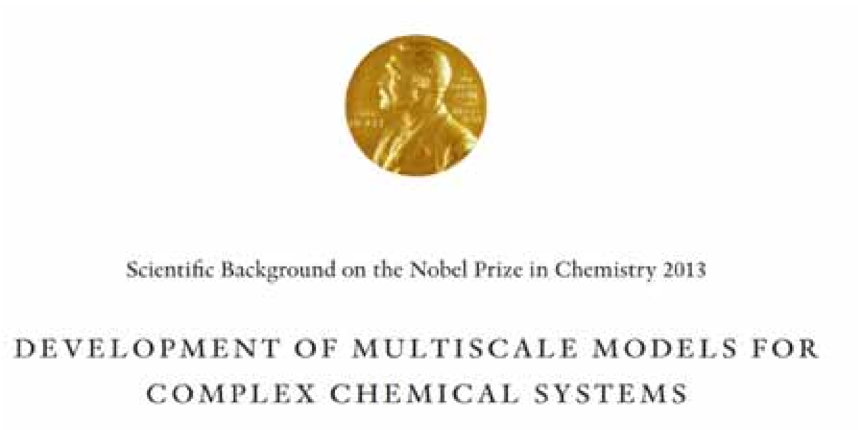


Pearl, J. 1988. Embracing causality in default reasoning. Artificial Intelligence, 35, (2), 259-271.

275 years later … the “Nobel-prize in Computer Science”



http://amturing.acm.org/vp/pearl_2658896.cfm


Scientific Background on the Nobel Prize in Chemistry 2013

DEVELOPMENT OF MULTISCALE MODELS FOR COMPLEX CHEMICAL SYSTEMS


Photo: A. Mahmoud
Martin Karplus
Prize share: 1/3


Photo: A. Mahmoud
Michael Levitt
Prize share: 1/3

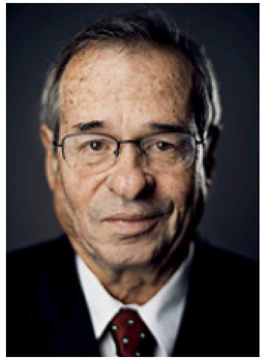
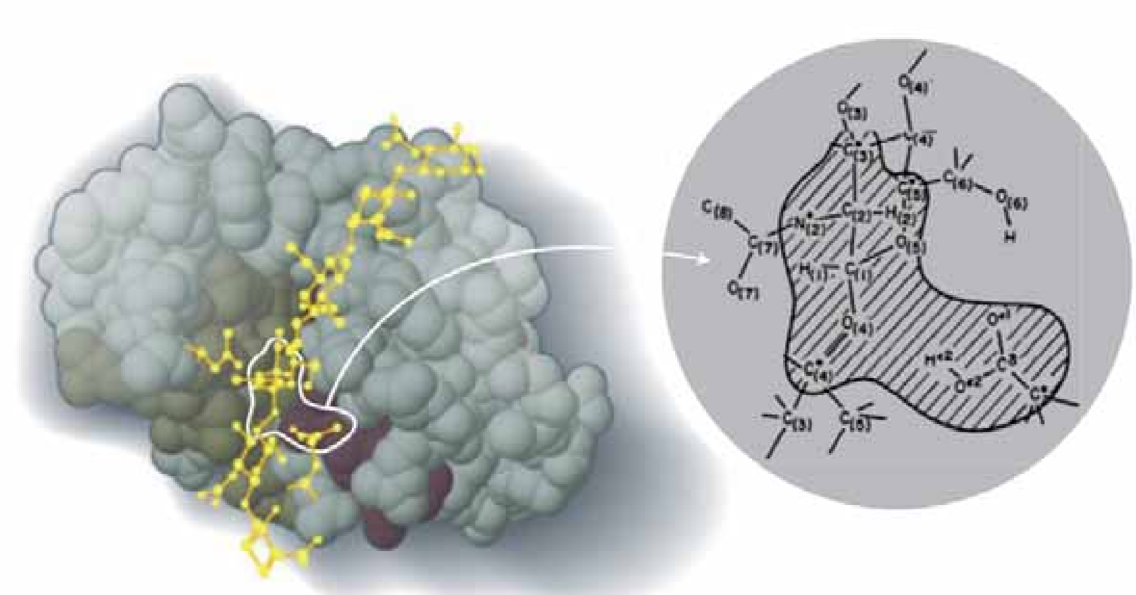


Photo: A. Mahmoud
Arieh Warshel
Prize share: 1/3

http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2013

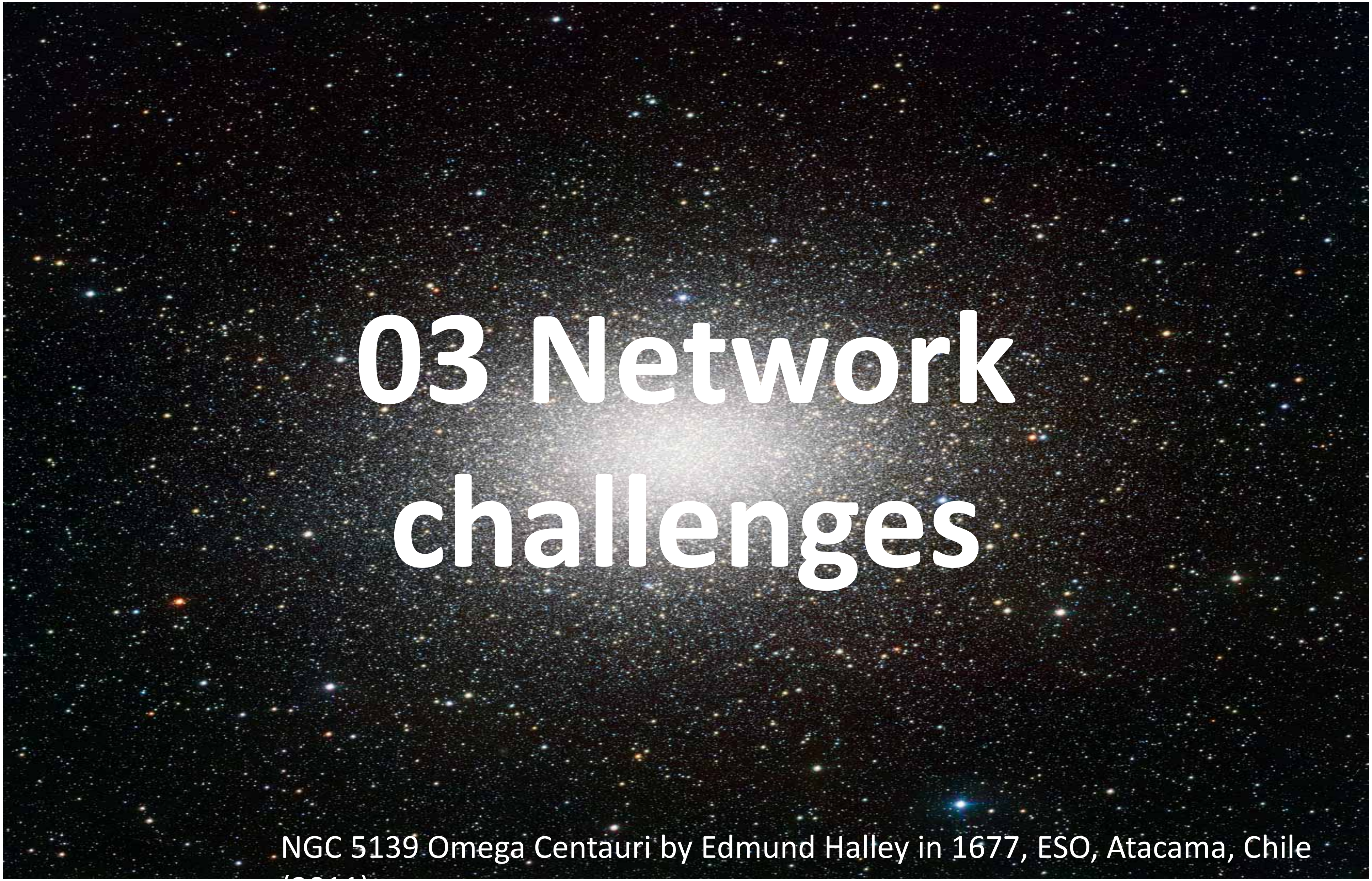


http://news.harvard.edu/gazette/story/2013/10/nobel_prize_awarded_2013/

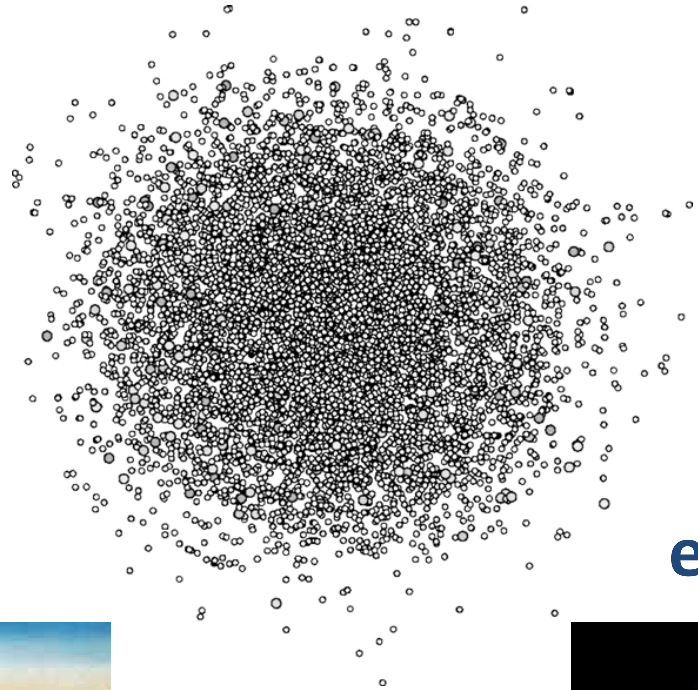
- **Graphs as models for networks**
- given as direct input (point cloud data sets)
- Given as properties of a structure
- Given as a representation of information (e.g. Facebook data, viral marketing, etc., …)

- **Graphs as nonparametric basis**
- we learn the structure from samples and infer
- flat vector data, e.g. similarity graphs
- encoding structural properties (e.g. smoothness, independence, …)

We skip this interesting chapter for now …



NGC 5139 Omega Centauri by Edmund Halley in 1677, ESO, Atacama, Chile (2011)



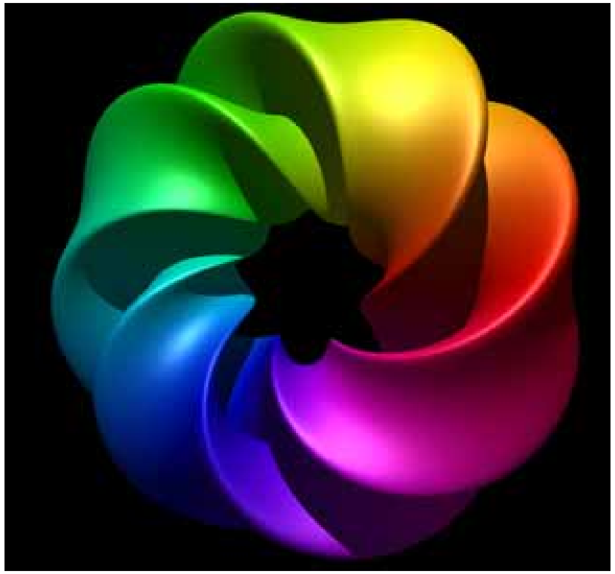
Time

Space

e.g. Entropy

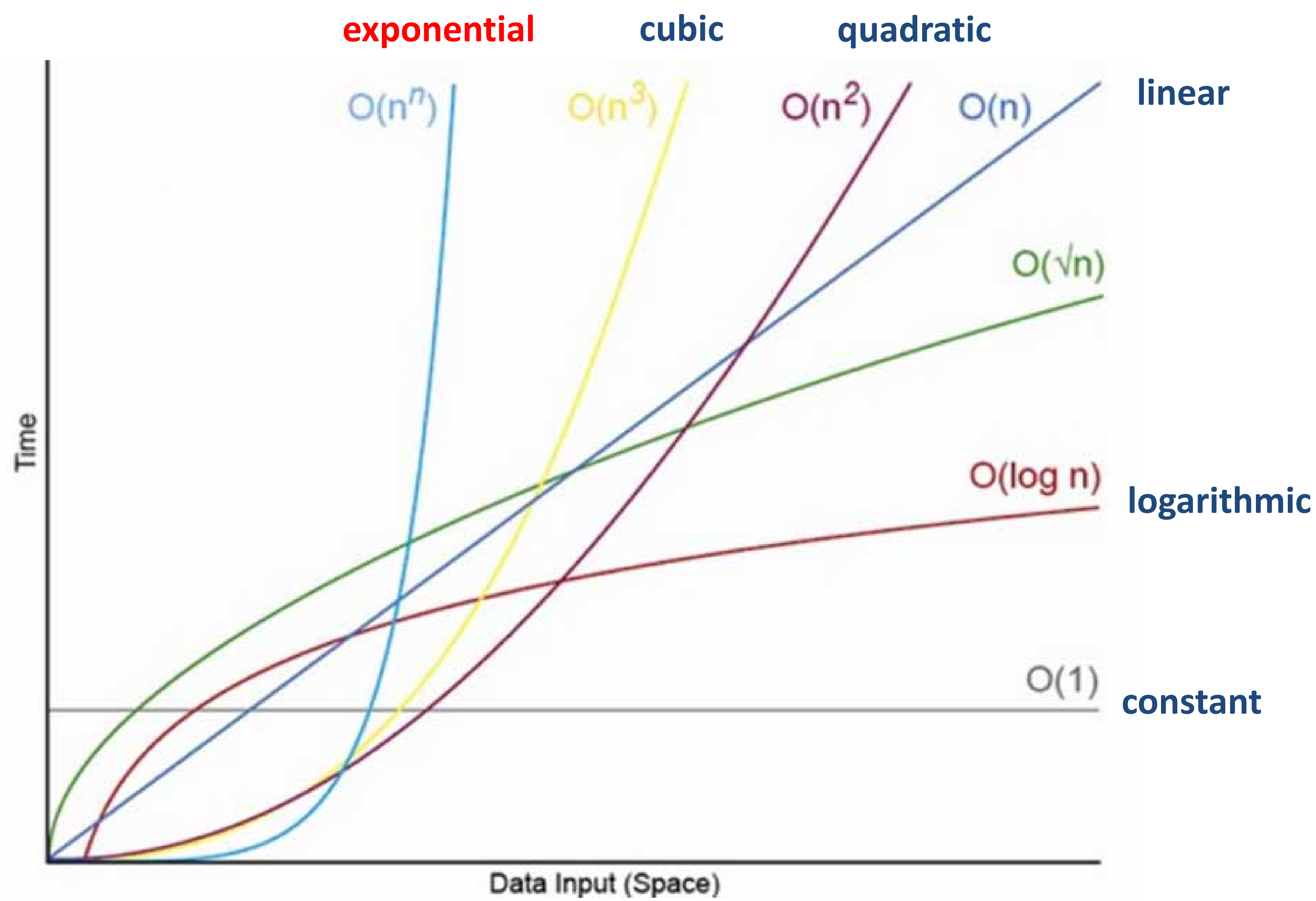
e.g. Topology

Dali, S. (1931) The persistence of memory



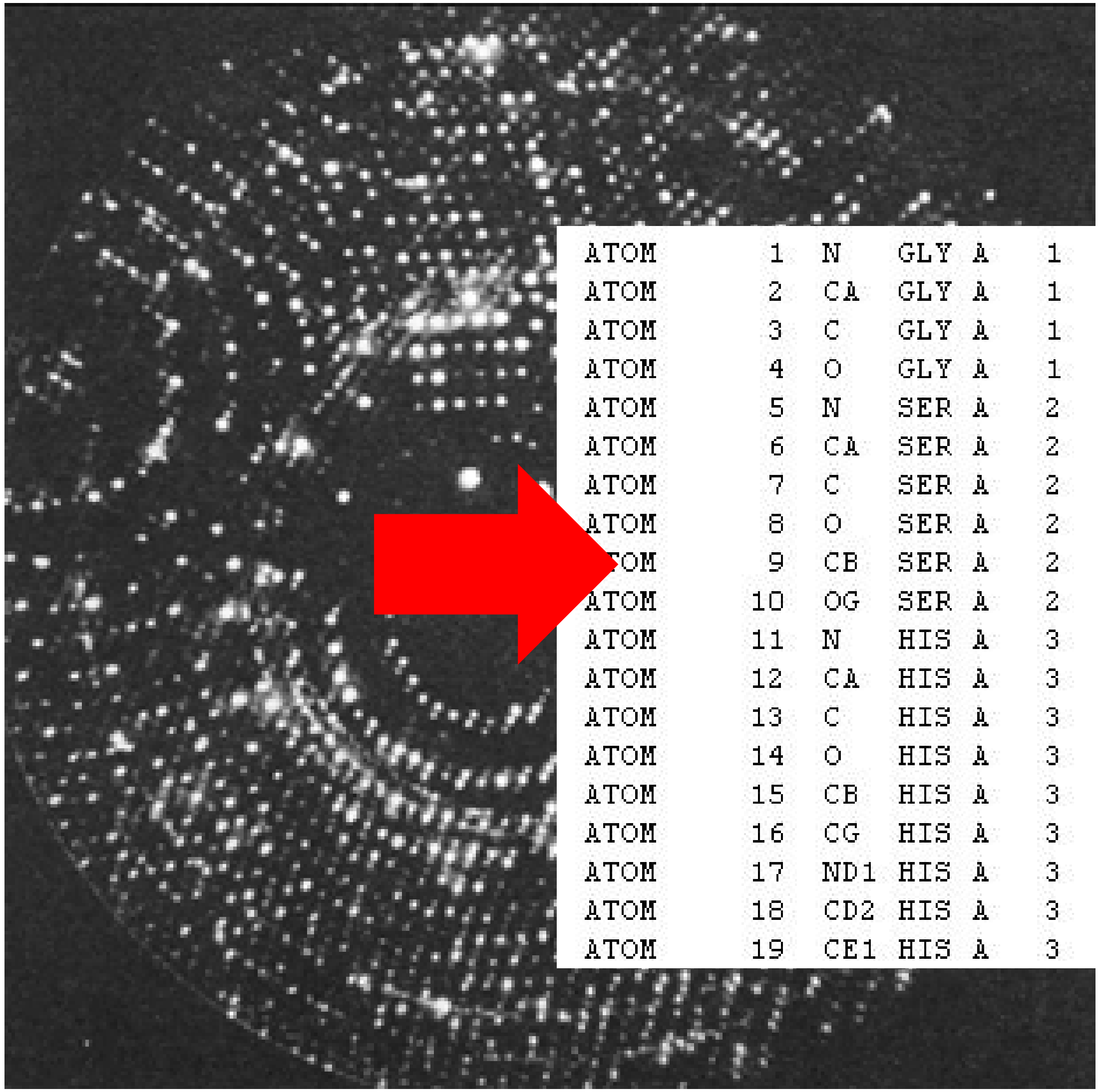
Bagula & Bourke (2012) Klein-Bottle

Complexity Problem: Time versus Space



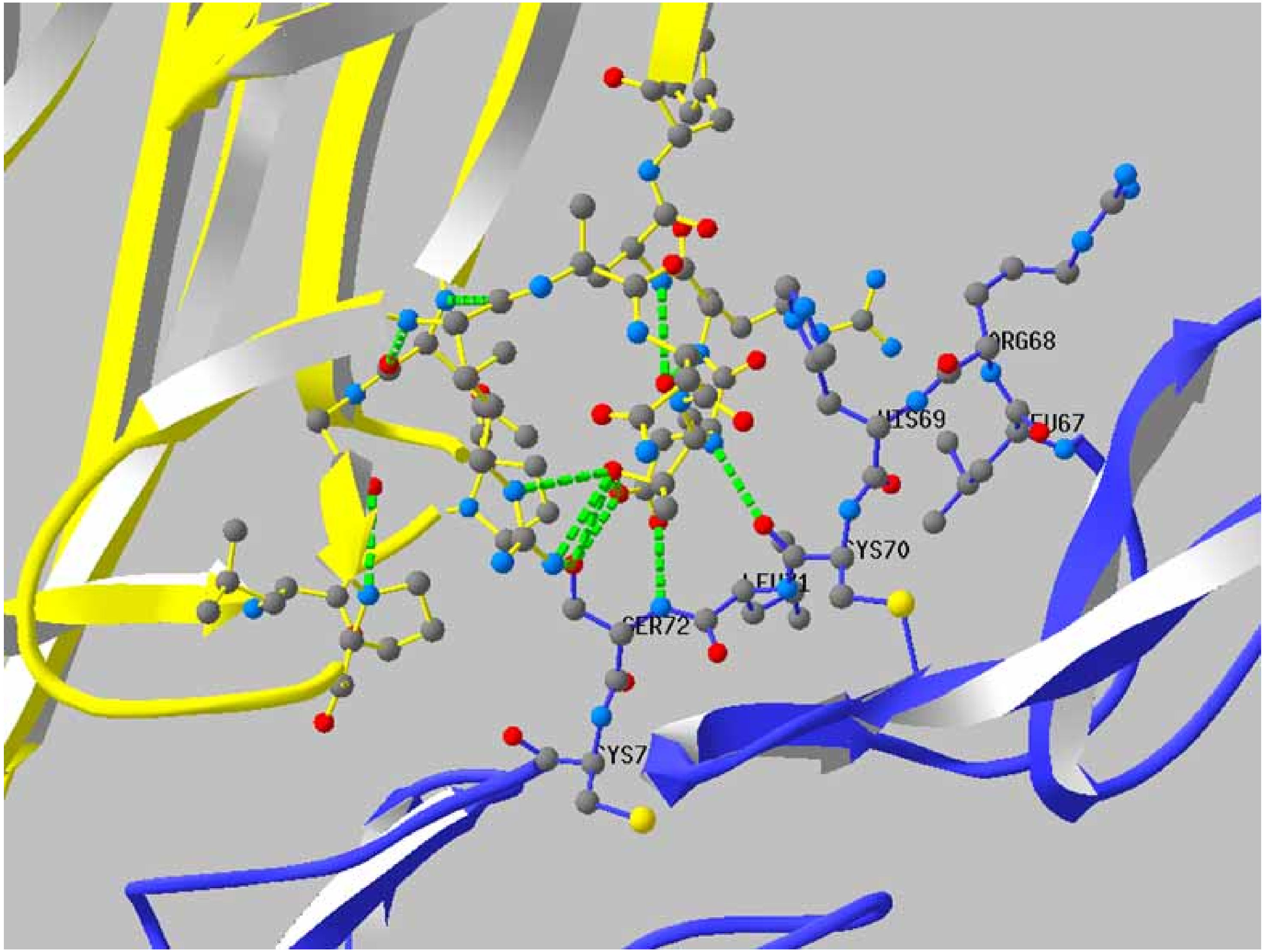
P versus NP and the Computational Complexity Zoo, please have a look at https://www.youtube.com/watch?v=YX40hbAHx3s

Our World in Data – Microscopic Structures



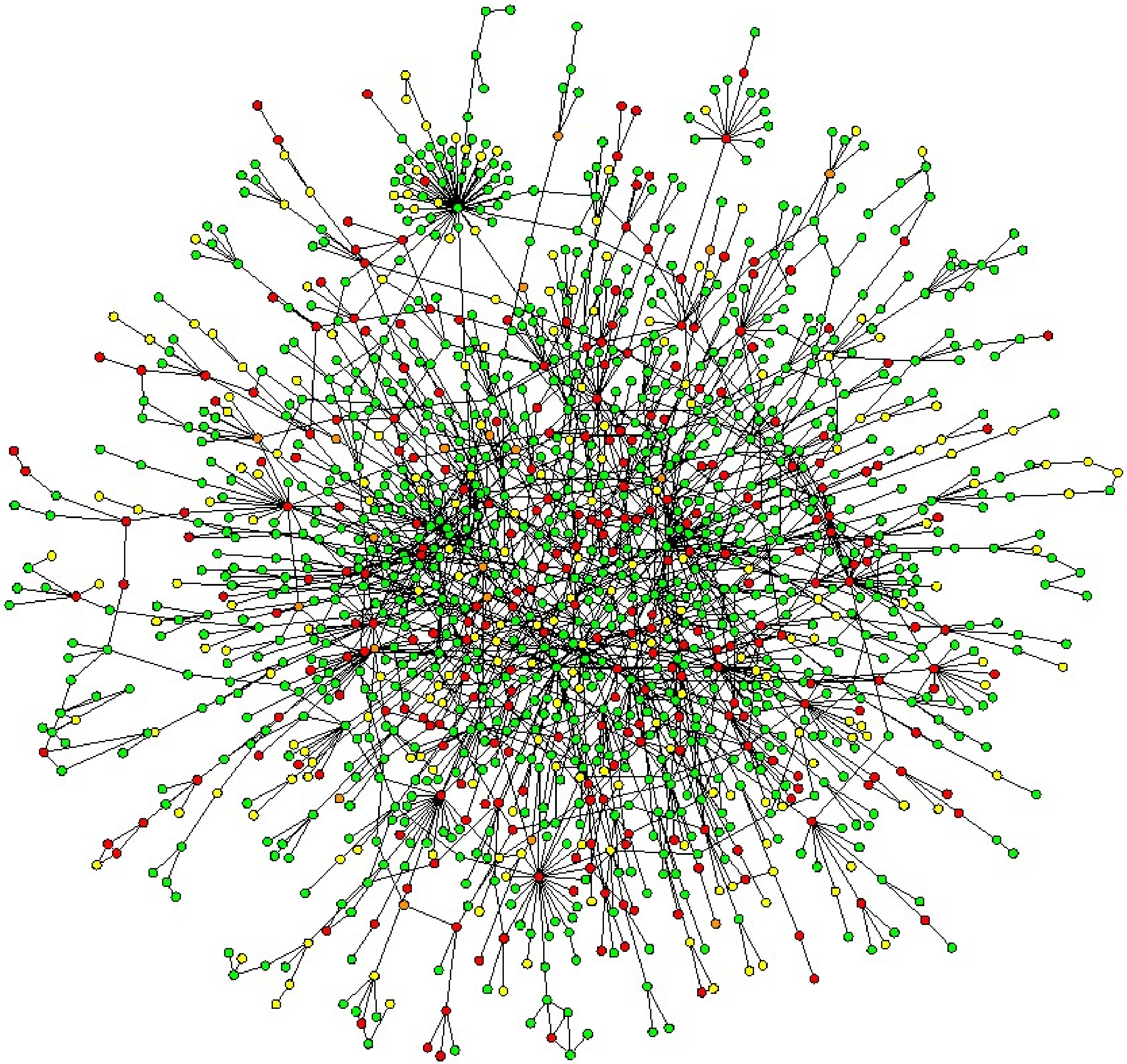
ATOM	1	N	GLY	A	1	44.842	51.034	101.284	0.01	27.20
ATOM	2	CA	GLY	A	1	45.640	50.230	100.389	0.01	26.99
ATOM	3	C	GLY	A	1	46.692	49.648	101.308	0.01	26.80
ATOM	4	O	GLY	A	1	46.895	50.222	102.381	0.01	26.91
ATOM	5	N	SER	A	2	47.283	48.516	100.951	1.00	26.26
ATOM	6	CA	SER	A	2	48.277	47.866	101.761	1.00	26.17
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*



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.*

First yeast protein-protein interaction network

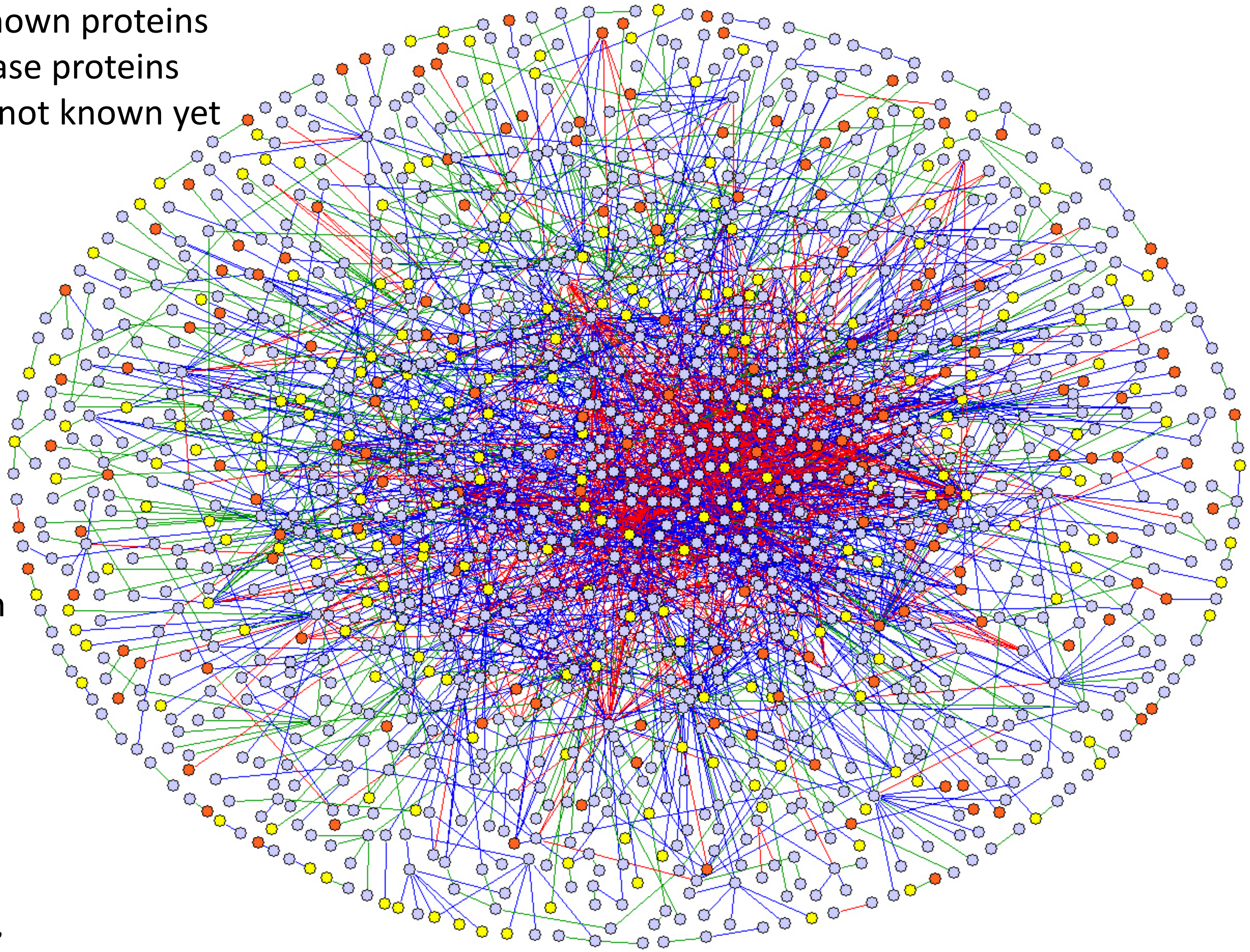


Nodes = proteins
Links = physical interactions (bindings)
Red Nodes = lethal
Green Nodes = non-lethal
Orange = slow growth
Yellow = not known

Jeong, H., Mason, S. P., Barabasi, A. L. & Oltvai, Z. N. (2001) Lethality and centrality in protein networks. *Nature, 411, 6833, 41-42.*

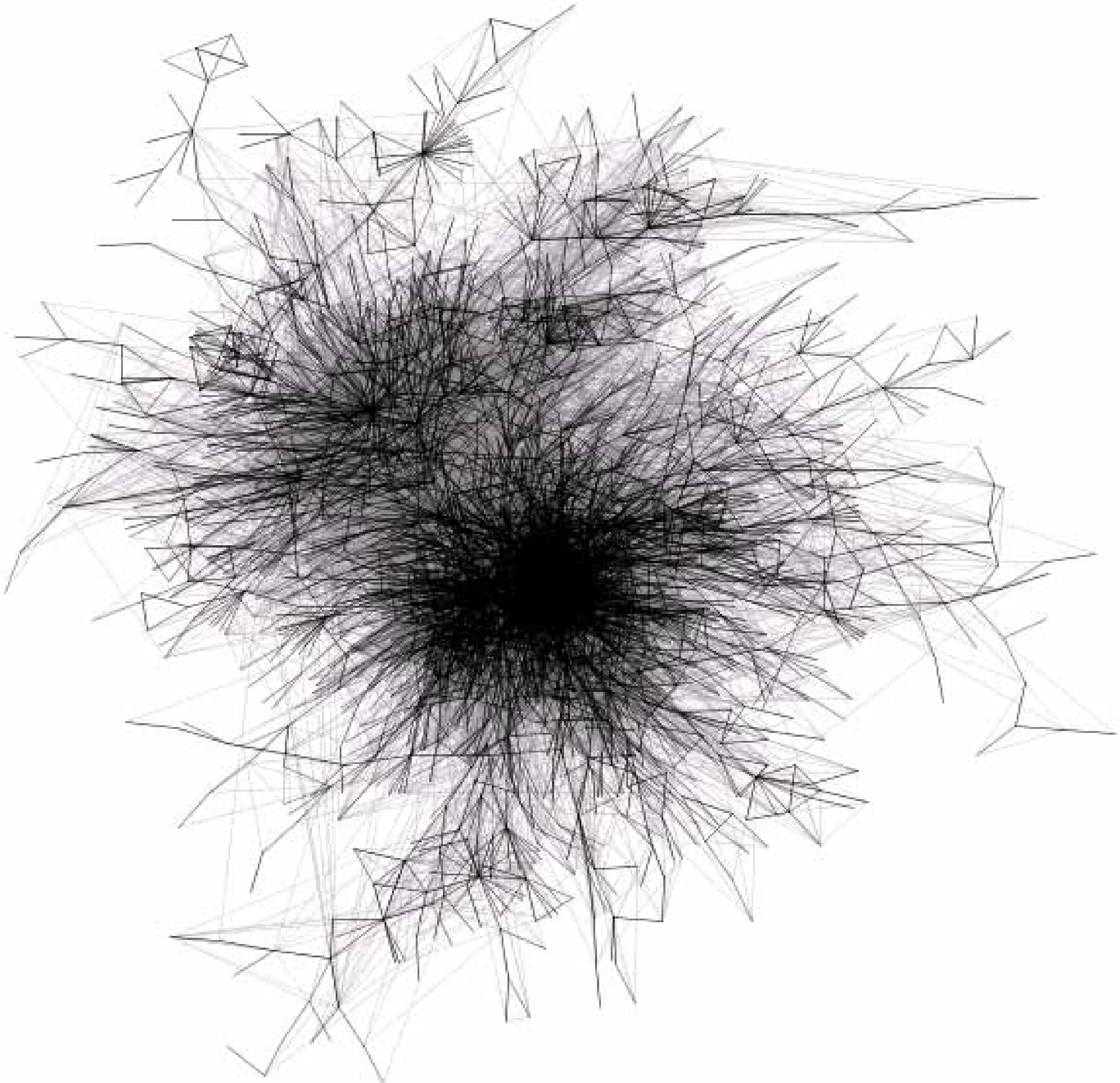
First human protein-protein interaction network

Light blue = known proteins
Orange = disease proteins
Yellow ones = not known yet



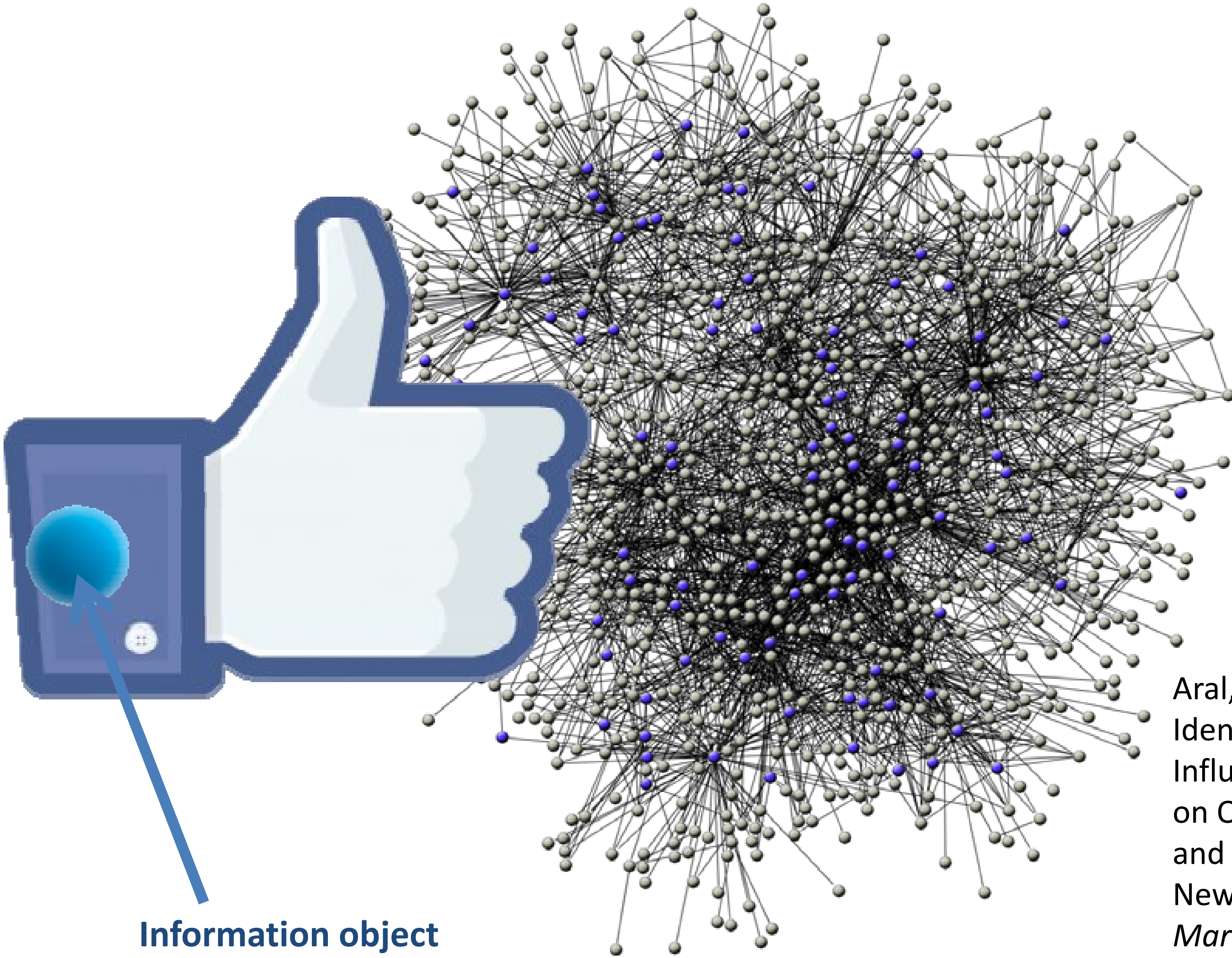
Stelzl, U. et al. (2005) A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome. *Cell, 122, 6, 957-968.*

Non-Natural Network Example: Blogosphere

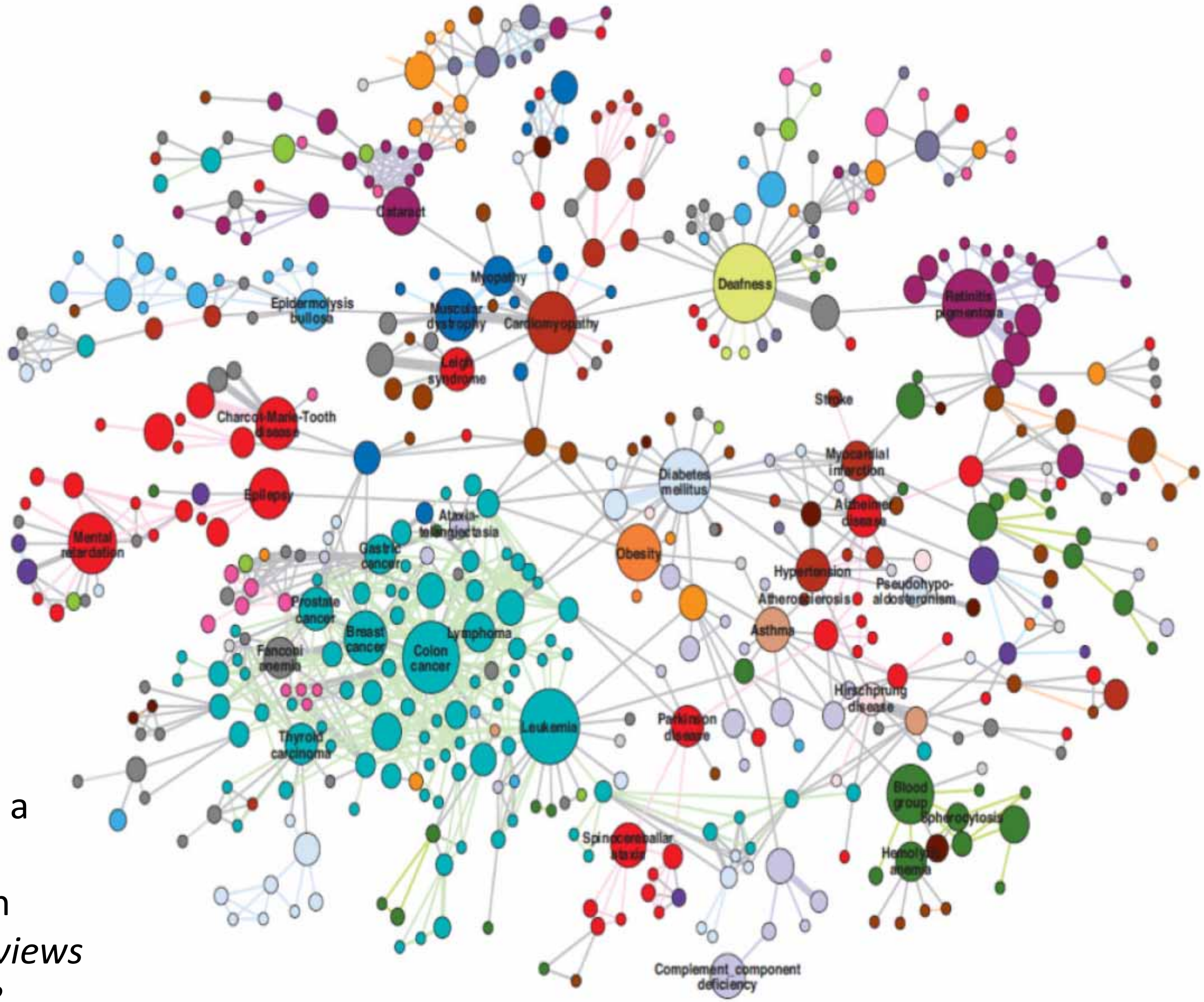


Hurst, M. (2007), Data Mining: Text Mining, Visualization and Social Media. Online available: http://datamining.typepad.com/data_mining/2007/01/the_blogosphere.html, last access: 2011-09-24

Social Behavior Contagion Network

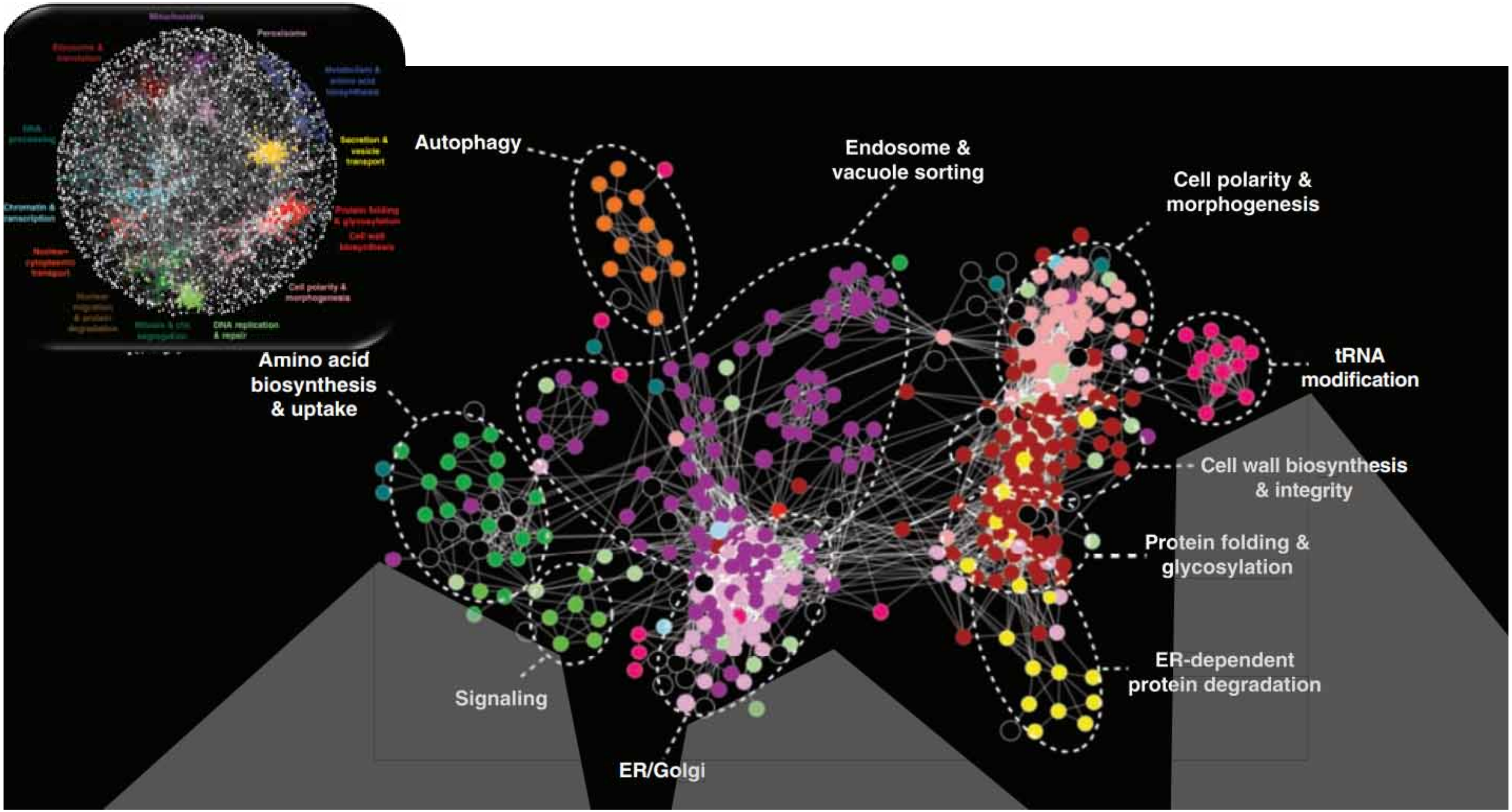


Aral, S. (2011) Identifying Social Influence: A Comment on Opinion Leadership and Social Contagion in New Product Diffusion. *Marketing Science, 30, 2, 217-223.*



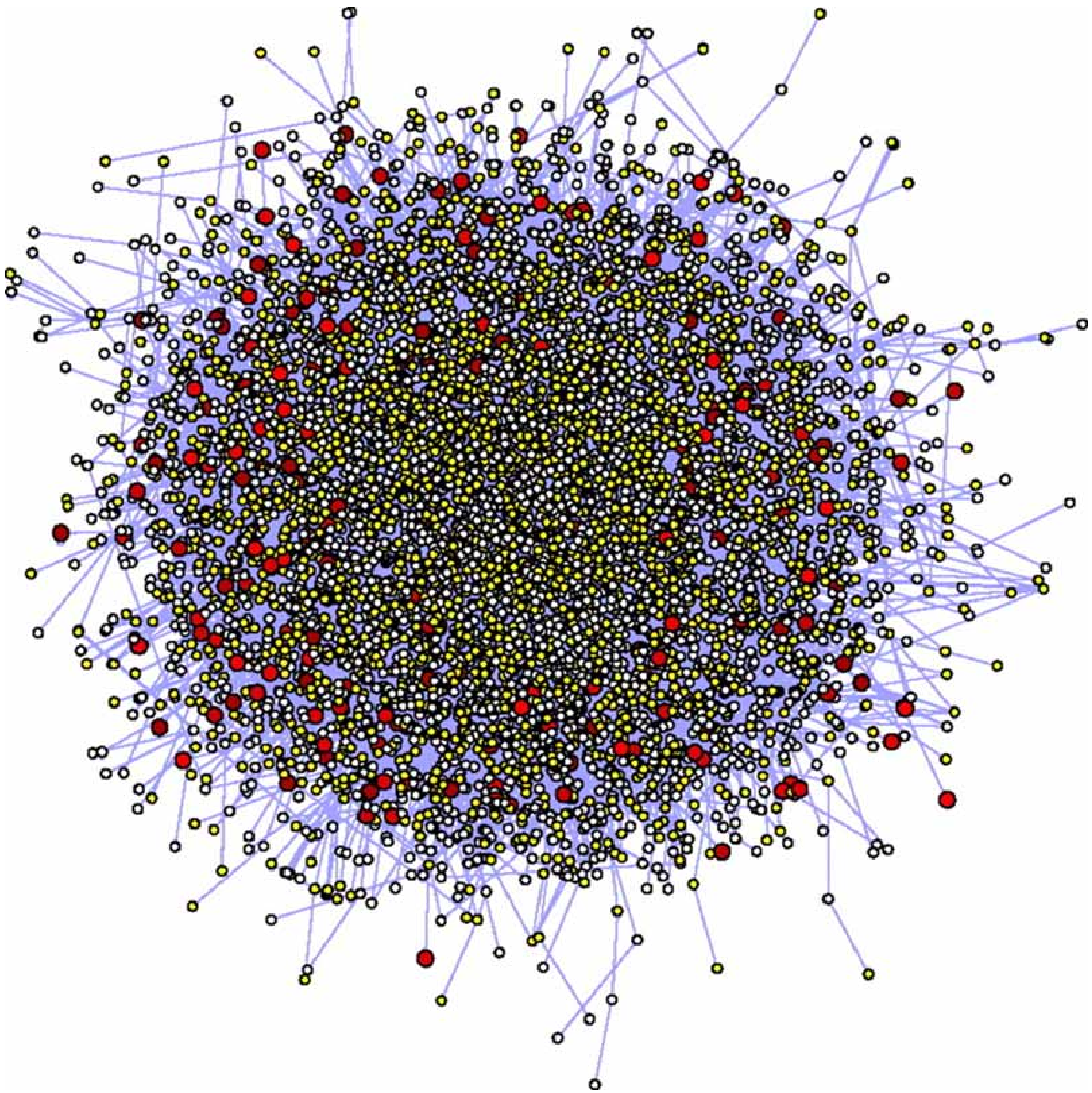
Barabási, A. L., Gulbahce, N. & Loscalzo, J. 2011. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics, 12****, 56-68.***

The Genetic Landscape of a cell



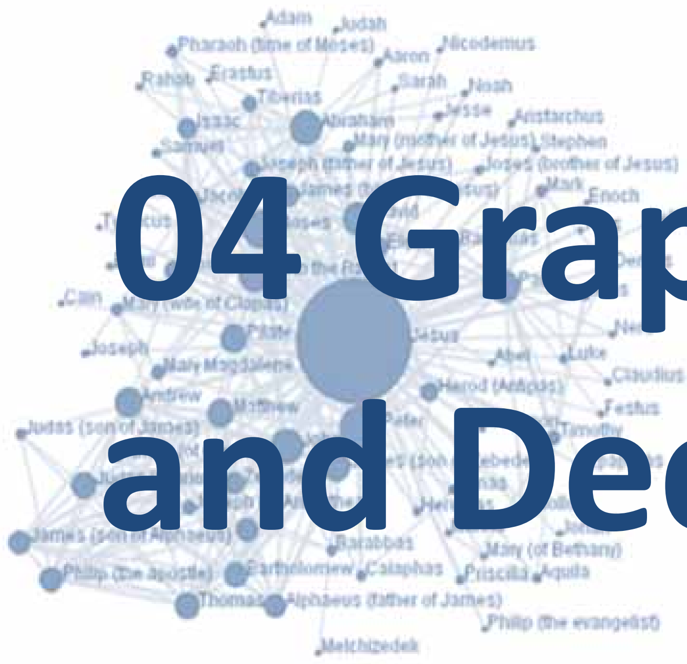
Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E. D., Sevier, C. S., Ding, H., Koh, J. L., Toufighi, K. & Mostafavi, S. 2010. The genetic landscape of a cell. science, 327, (5964), 425-431.

Example for a weakly structured data set - PPI



Kim, P. M., Korbel, J. O. & Gerstein, M. B. 2007. Positive selection at the protein network periphery: Evaluation in terms of structural constraints and cellular context. Proceedings of the National Academy of Sciences, 104, (51), 20274-20279.

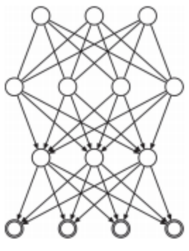
04 Graphical Models and Decision Making

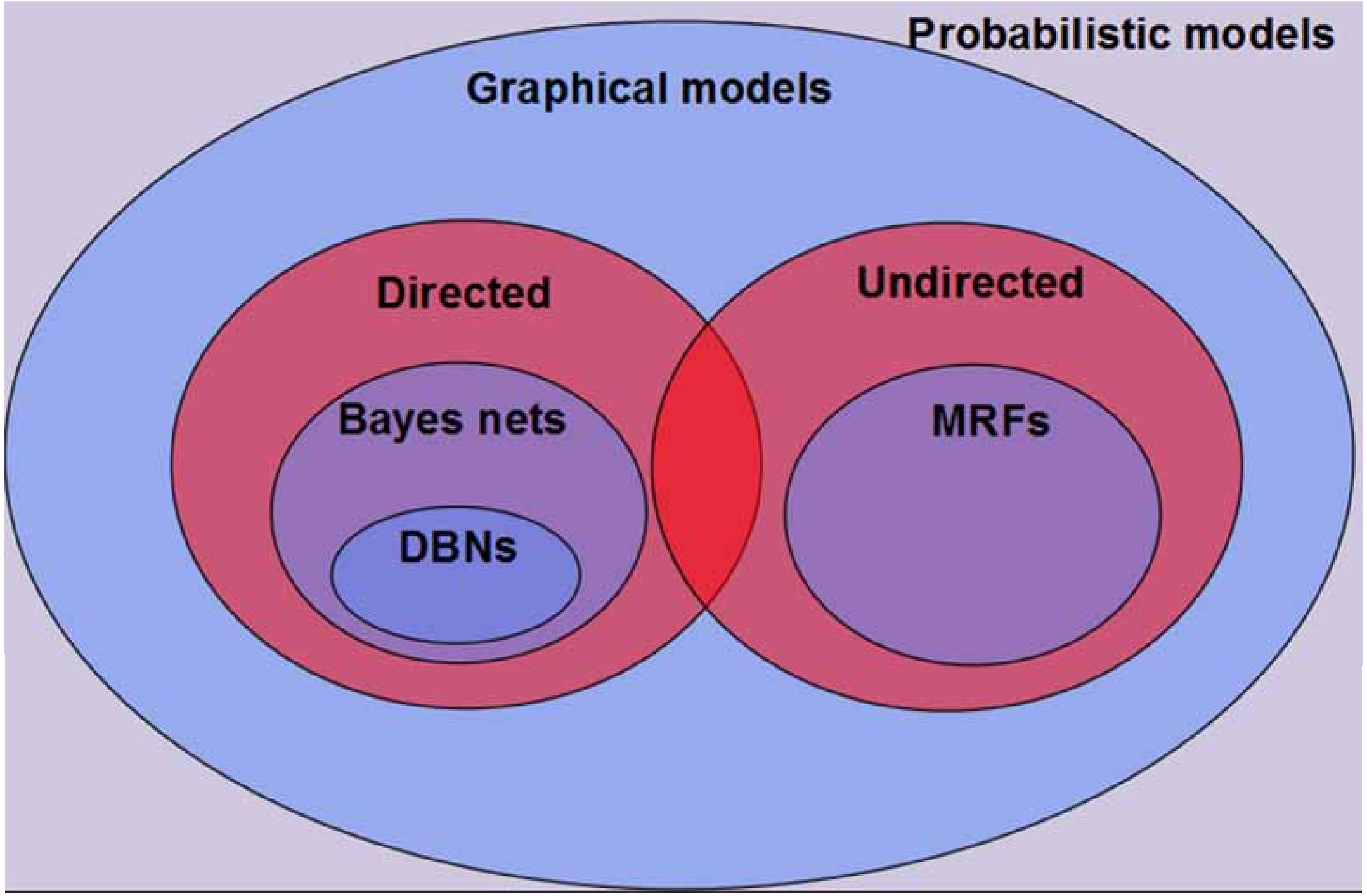


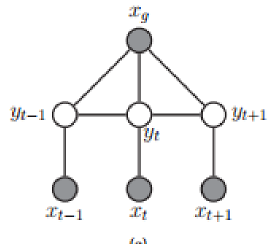
$$\mathcal{M}$$

$$\mathcal{D} \equiv \{X_1^{(i)}, X_2^{(i)}, \ldots, X_m^{(i)}\}_{i=1}^{N}$$

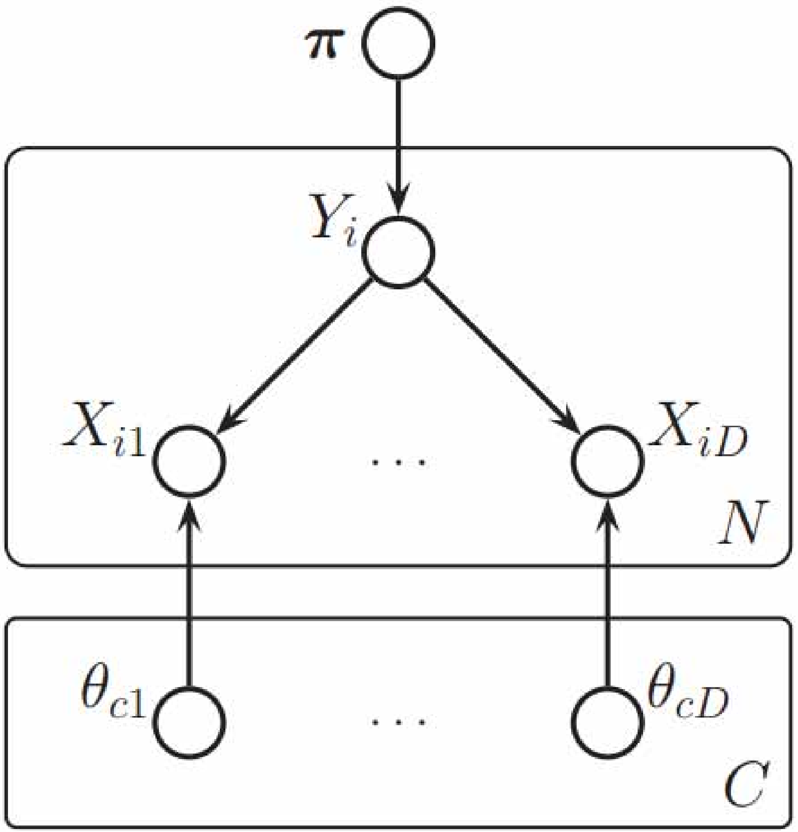
Classes of Graphical Models

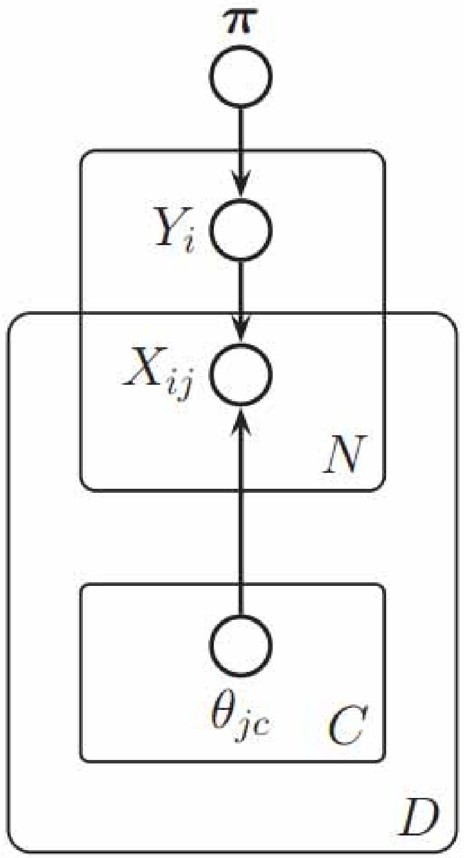






Murphy, K. P. 2012. Machine learning: a probabilistic perspective, Cambridge (MA), MIT press.





π … multinomial parameter vector, Stationary distribution of Markov chain

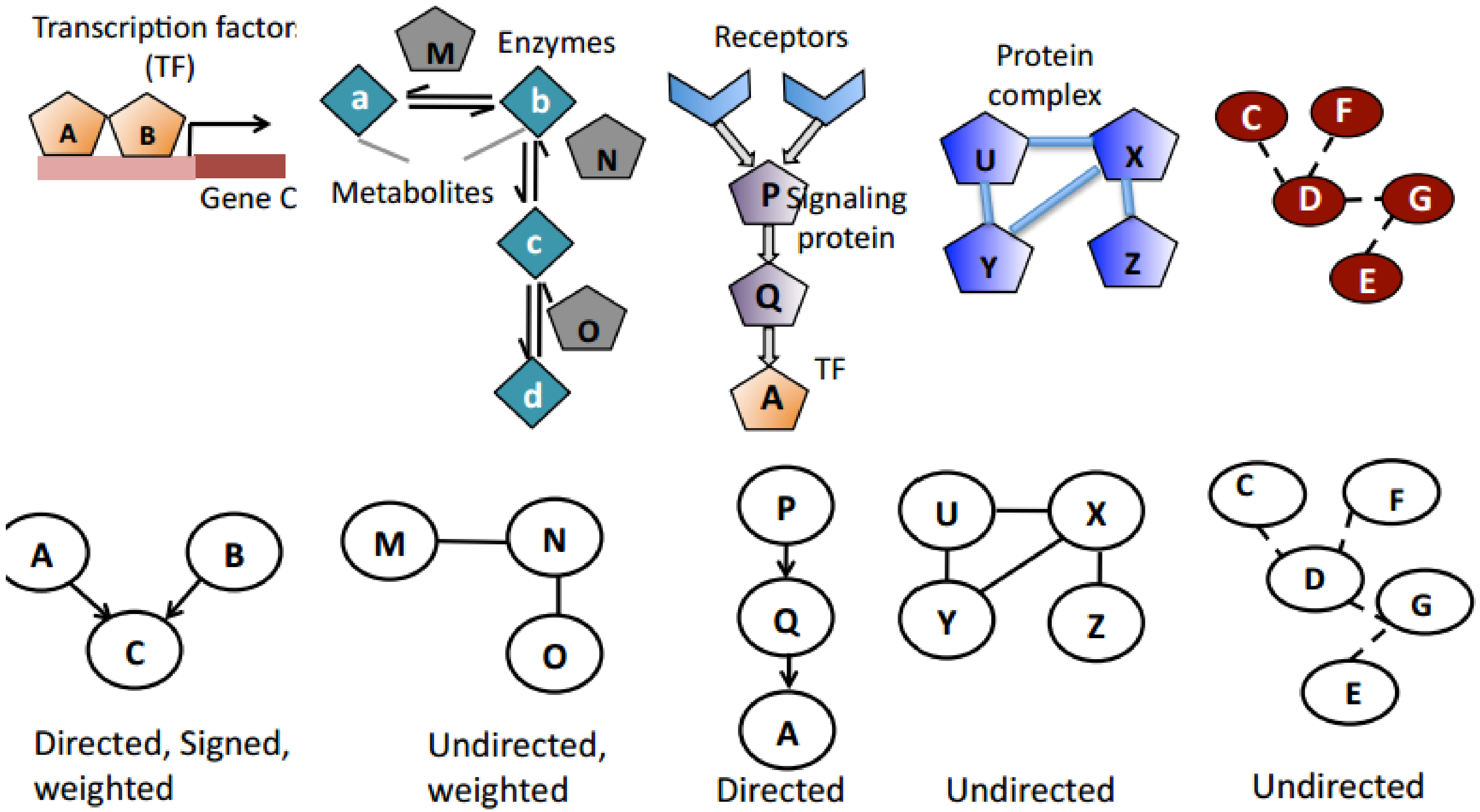
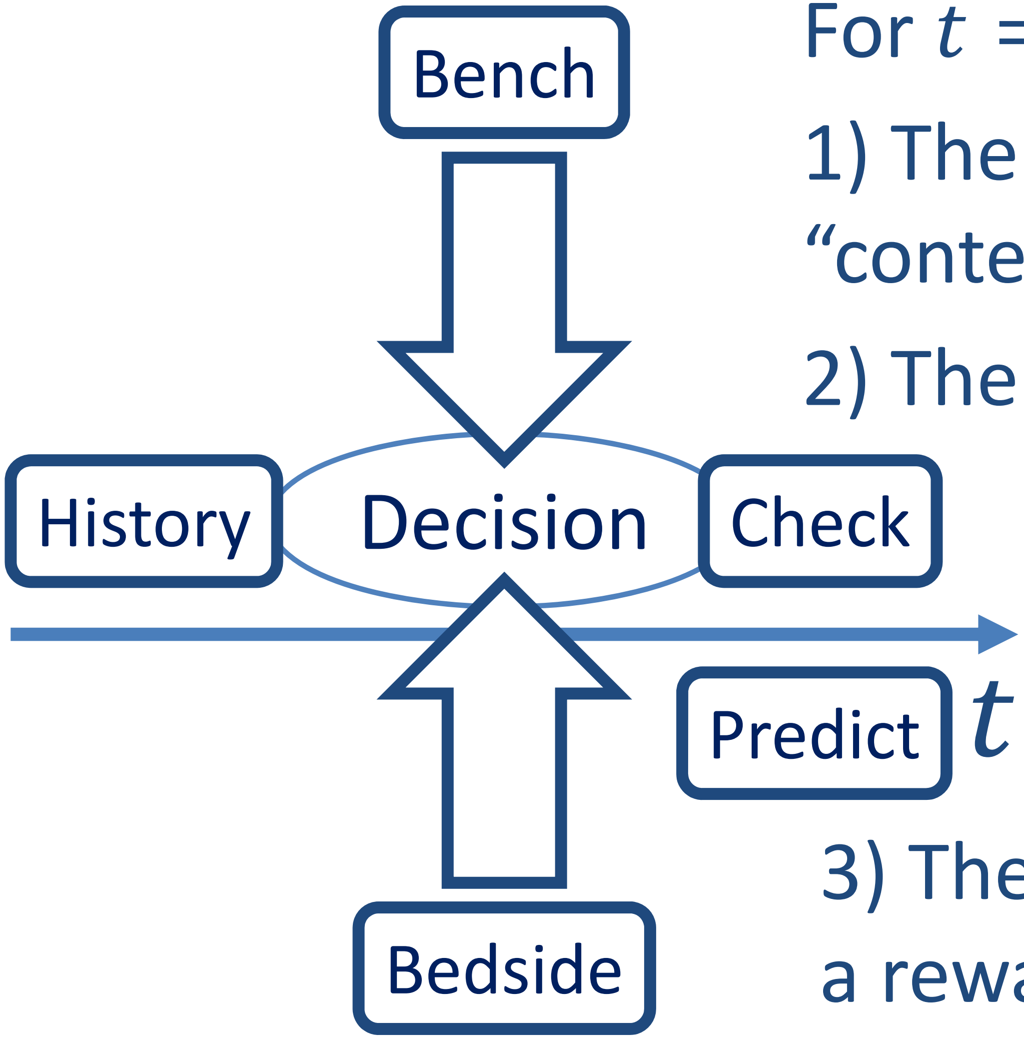


Image credit to Anna Goldenberg, Toronto

Goal: Learn an **optimal policy** for selecting best actions within a given **context**



For $t = 1, \dots, T$

1) The world produces a "context" $x_t \in X$

2) The learner selects an action $a_t \in \{1, \dots, K\}$

3) The world reacts with a reward $r_t(a_t) \in [0,1]$

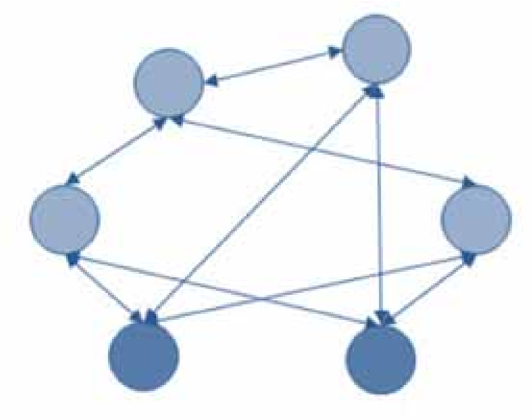
- Key Idea: Conditional independence assumptions are very useful – however: Naïve Bayes is extreme!
- X is *conditionally independent* of Y, given Z, if the P(X) governing X is independent of value Y, given value of Z:

$$(\forall i, j, k) P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k)$$

can be abbr. with $P(X|Y, Z) = P(X|Z)$

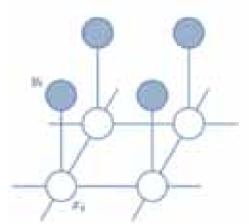
- Graphical models express sets of conditional independence assumptions via graph structure
- The graph structure plus associated parameters define joint probability distribution over the set of variables

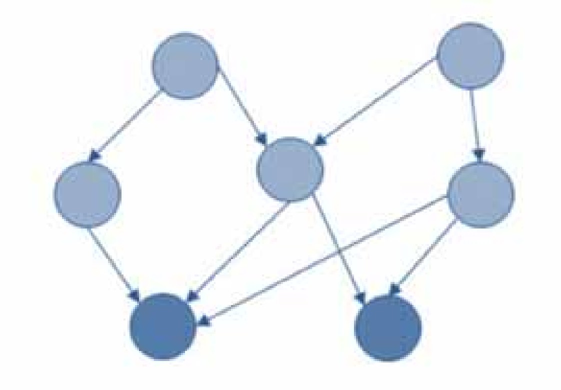
- Medicine is an extremely complex application domain – dealing most of the time with uncertainties -> **probable information!**
- When we have big data but little knowledge automatic ML can help to gain insight:
- **Structure learning and prediction in large-scale biomedical networks with probabilistic graphical models**
- If we have little data and deal with NP-hard problems we still need the human-in-the-loop



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

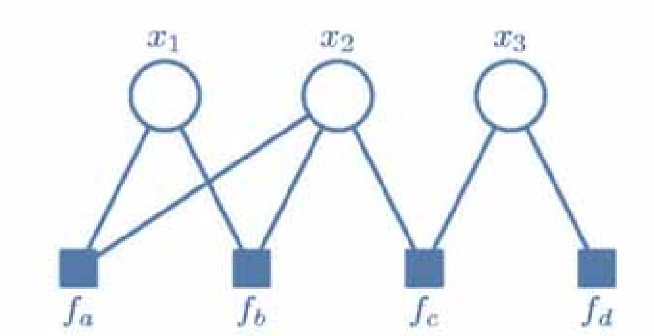
$$P(\boldsymbol{X}) = \frac{1}{Z} \exp\left(\sum_{ij} W_{ij}\, x_i x_j + \sum_i x_i b_i \right)$$





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

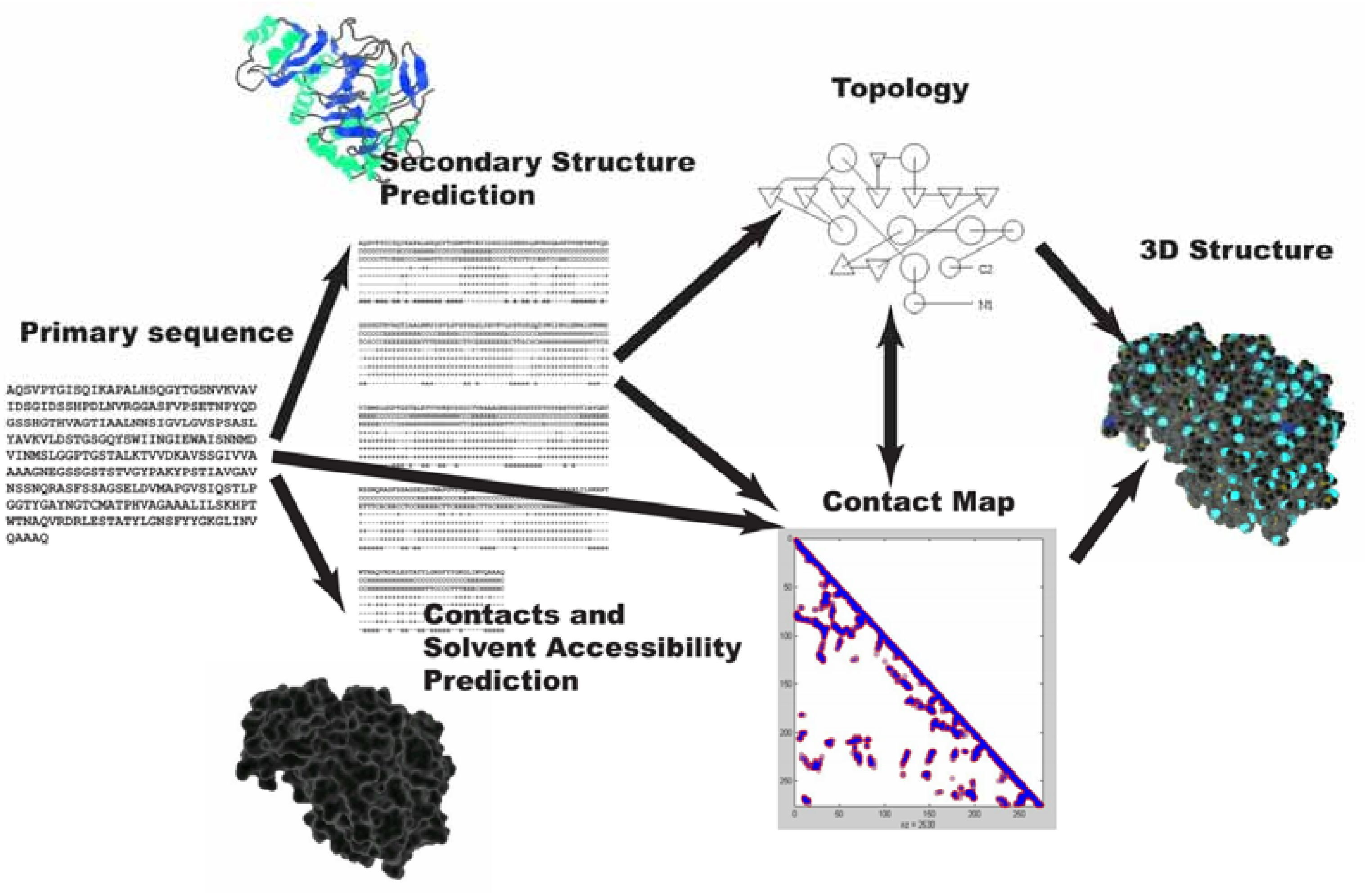
$$p(\mathbf{x}) = \prod_s f_s(\mathbf{x}_s)$$

- What is the advantage of factor graphs?

	Dependency	Efficient Inference	Usage
Bayesian Networks	Yes	Somewhat	Ancestral Generative Process
Markov Networks	Yes	No	Local Couplings and Potentials
Factor Graphs	No	Yes	Efficient, distributed inference

Table credit to Ralf Herbrich, Amazon

From structure to function prediction

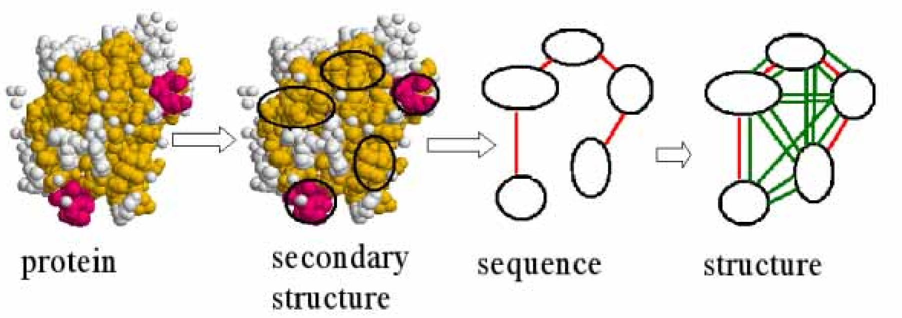


Baldi, P. & Pollastri, G. 2003. The principled design of large-scale recursive neural network architectures--dag-rnns and the protein structure prediction problem. The Journal of Machine Learning Research, 4, 575-602.

- Hypothesis: most biological functions involve the interactions between many proteins, and the complexity of living systems arises as a result of such interactions.
- In this context, the problem of inferring a global protein network for a given organism,
- - using all (genomic) data of the organism,
- is one of the main challenges in computational biology

Yamanishi, Y., Vert, J.-P. & Kanehisa, M. 2004. Protein network inference from multiple genomic data: a supervised approach. Bioinformatics, 20, (suppl 1), i363-i370.

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

Example: Protein Network Inference

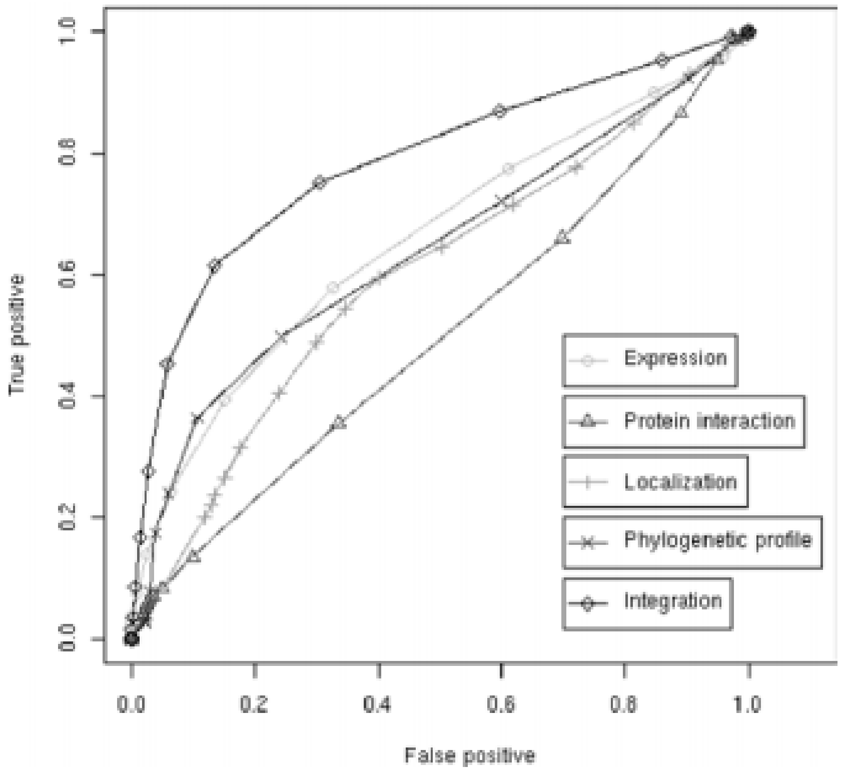
BIOINFORMATICS

Vol. 20 Suppl. 1 2004, pages i363–i370
DOI: 10.1093/bioinformatics/bth910

Protein network inference from multiple genomic data: a supervised approach

Y. Yamanishi[1,], J.-P. Vert[2] and M. Kanehisa[1]*

[1]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan and [2]Computational Biology group, Ecole des Mines de Paris, 35 rue Saint-Honoré, 77305 Fontainebleau cedex, France



K_{exp} (Expression)
K_{ppi} (Protein interaction)
K_{loc} (Localization)
K_{phy} (Phylogenetic profile)
$K_{\text{exp}} + K_{\text{ppi}} + K_{\text{loc}} + K_{\text{phy}}$ (Integration)

Example: Data fusion and Protein Annotation

BIOINFORMATICS

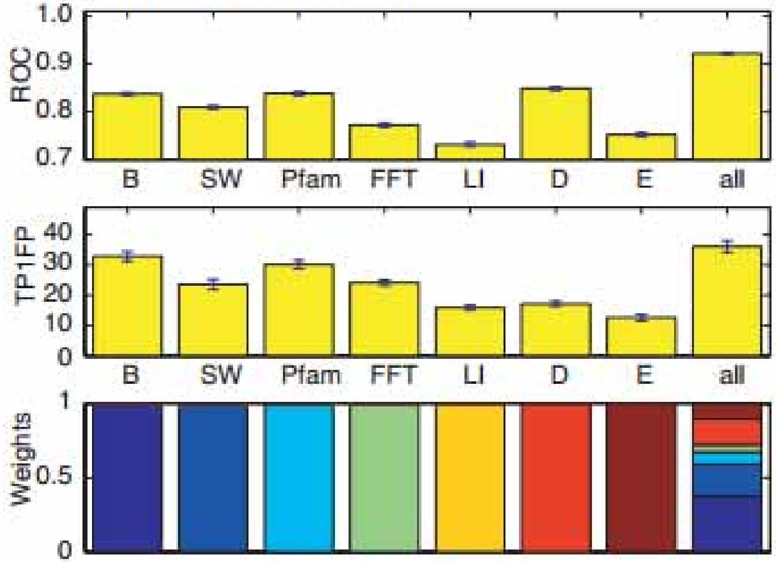
Vol. 20 no. 16 2004, pages 2626–2635
doi:10.1093/bioinformatics/bth294

A statistical framework for genomic data fusion

Gert R. G. Lanckriet[1], Tijl De Bie[3], Nello Cristianini[4], Michael I. Jordan[2] and William Stafford Noble[5,*]

[1]Department of Electrical Engineering and Computer Science, [2]Division of Computer Science, Department of Statistics, University of California, Berkeley 94720, USA, [3]Department of Electrical Engineering, ESAT-SCD, Katholieke Universiteit Leuven 3001, Belgium, [4]Department of Statistics, University of California, Davis 95618, USA and [5]Department of Genome Sciences, University of Washington, Seattle 98195, USA

Kernel	Data	Similarity measure
K_{SW}	protein sequences	Smith-Waterman
K_{B}	protein sequences	BLAST
K_{Pfam}	protein sequences	Pfam HMM
K_{FFT}	hydropathy profile	FFT
K_{LI}	protein interactions	linear kernel
K_{D}	protein interactions	diffusion kernel
K_{E}	gene expression	radial basis kernel
K_{RND}	random numbers	linear kernel



(B) Membrane proteins

Lanckriet, G. R., De Bie, T., Cristianini, N., Jordan, M. I. & Noble, W. S. 2004. A statistical framework for genomic data fusion. Bioinformatics, 20, (16), 2626-2635.

05 Bayesian Networks “Bayes’ Nets”

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

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

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.

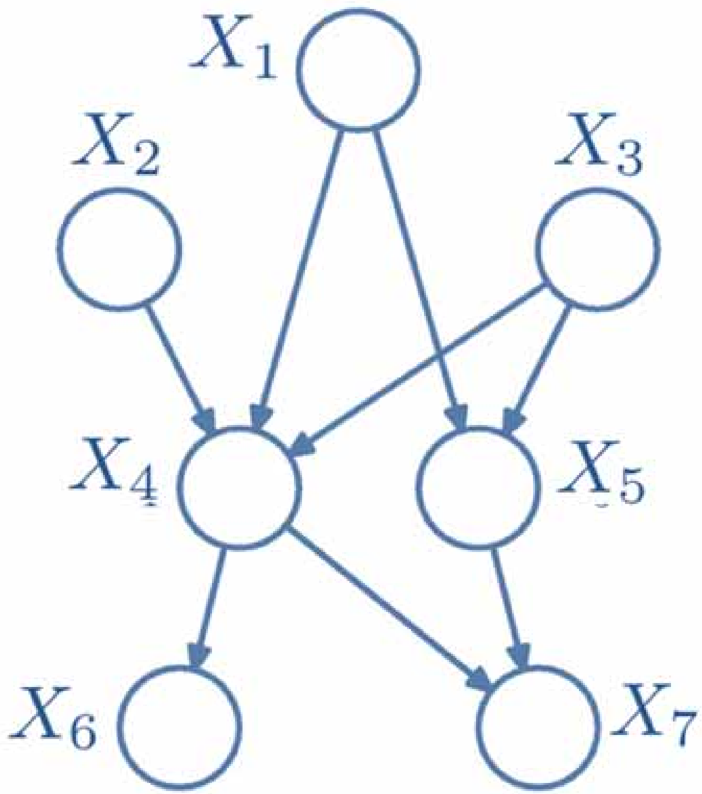
Pearl, J. (1988) *Probabilistic reasoning in intelligent systems: networks of plausible inference. San Francisco, Morgan Kaufmann.*

Example: Directed Bayesian Network with 7 nodes

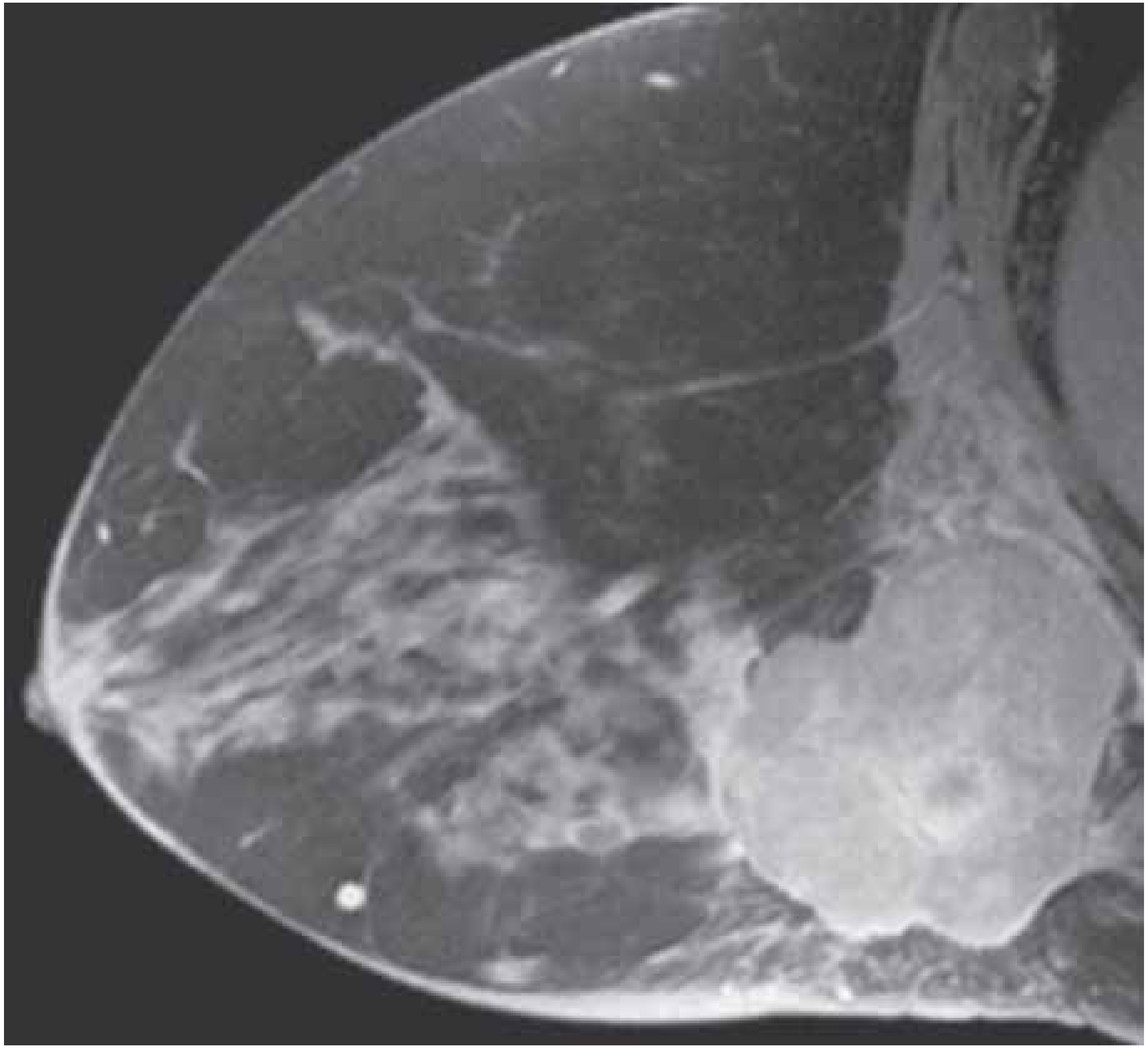
$$p(X_1, \ldots, X_7) =$$

$$p(X_1)p(X_2)p(X_3)p(X_4|X_1, X_2, X_3) \cdot$$

$$p(X_5|X_1, X_3)p(X_6|X_4)p(X_7|X_4, X_5)$$



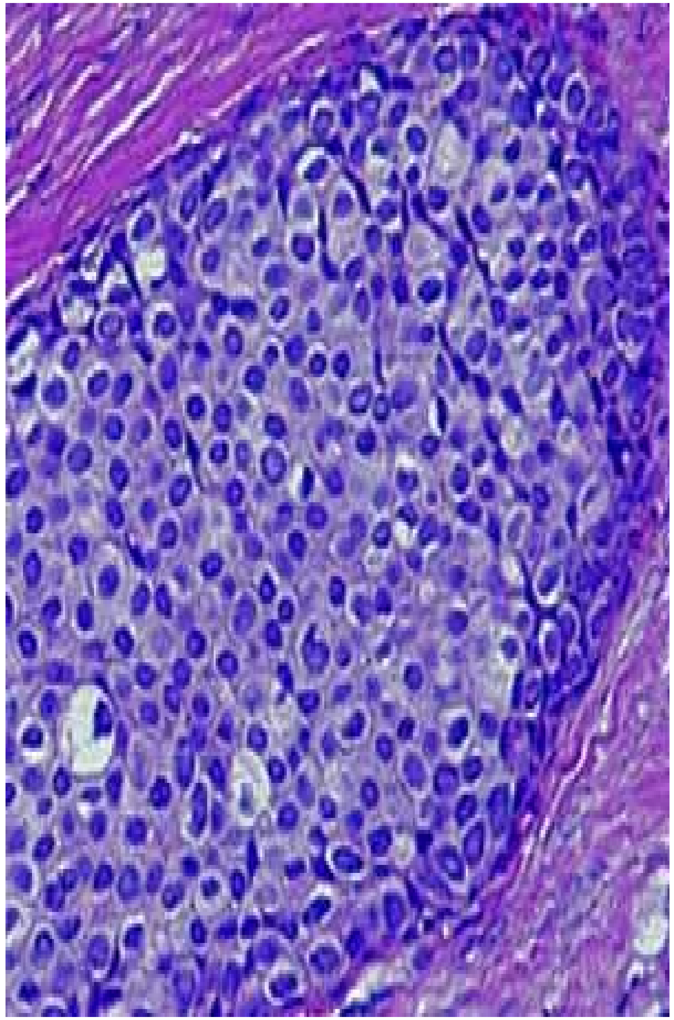
Clinical Case Example



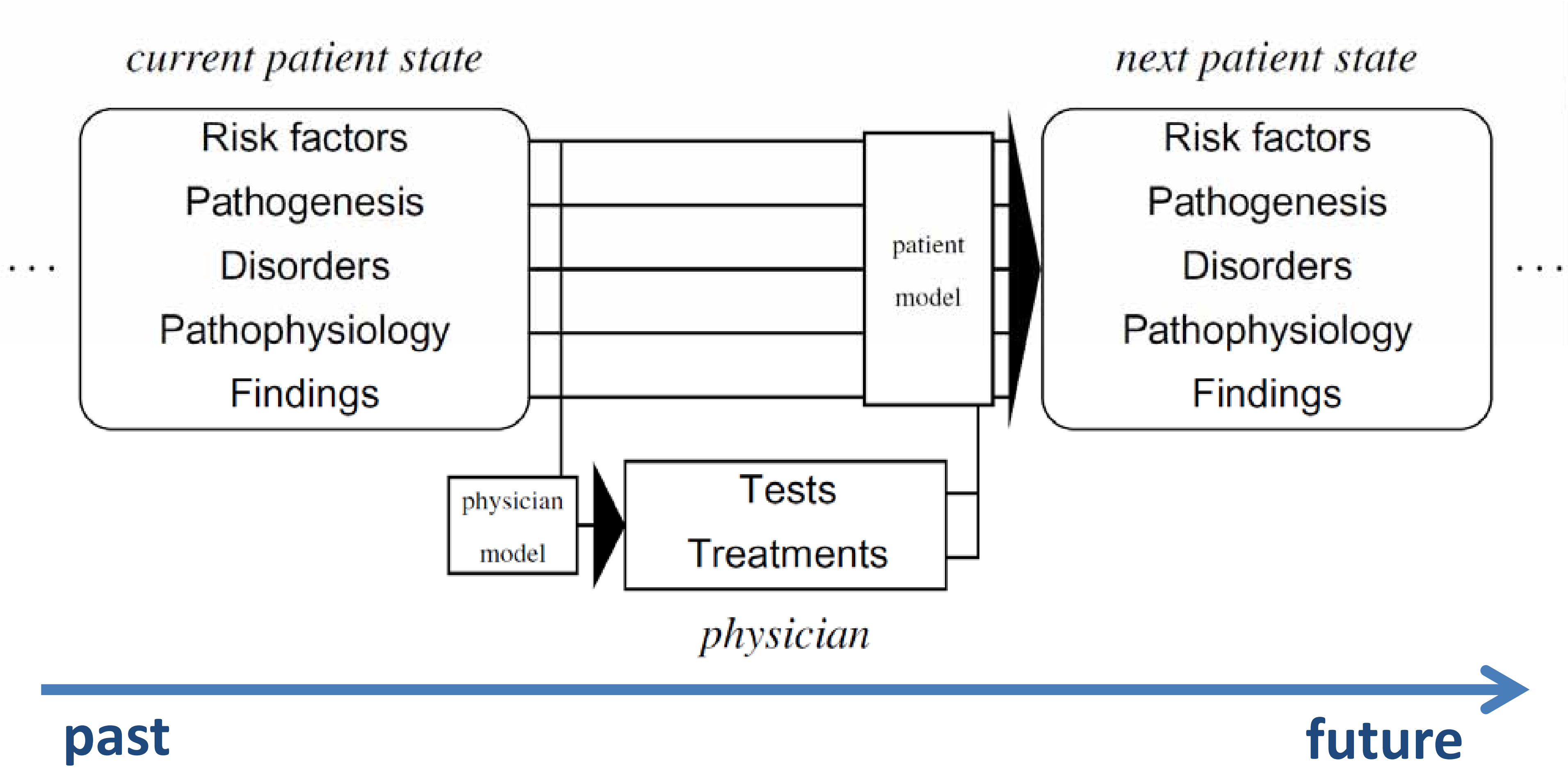
Overmoyer, B. A., Lee, J. M. & Lerwill, M. F. (2011) Case 17-2011 A 49-Year-Old Woman with a Mass in the Breast and Overlying Skin Changes. *New England Journal of Medicine, 364, 23, 2246-2254.*

Important in Clinical practice -> prognosis !

- = the prediction of the future course of a disease conditional on the patient's history and a projected treatment strategy
- Danger: probable Information !
- Therefore valid prognostic models can be of great benefit for clinical decision making and of great value to the patient, e.g., for notification and quality of-life decisions



Knaus, W. A., Wagner, D. P. & Lynn, J. (1991) Short-term mortality predictions for critically ill hospitalized adults: science and ethics. *Science, 254, 5030, 389.*

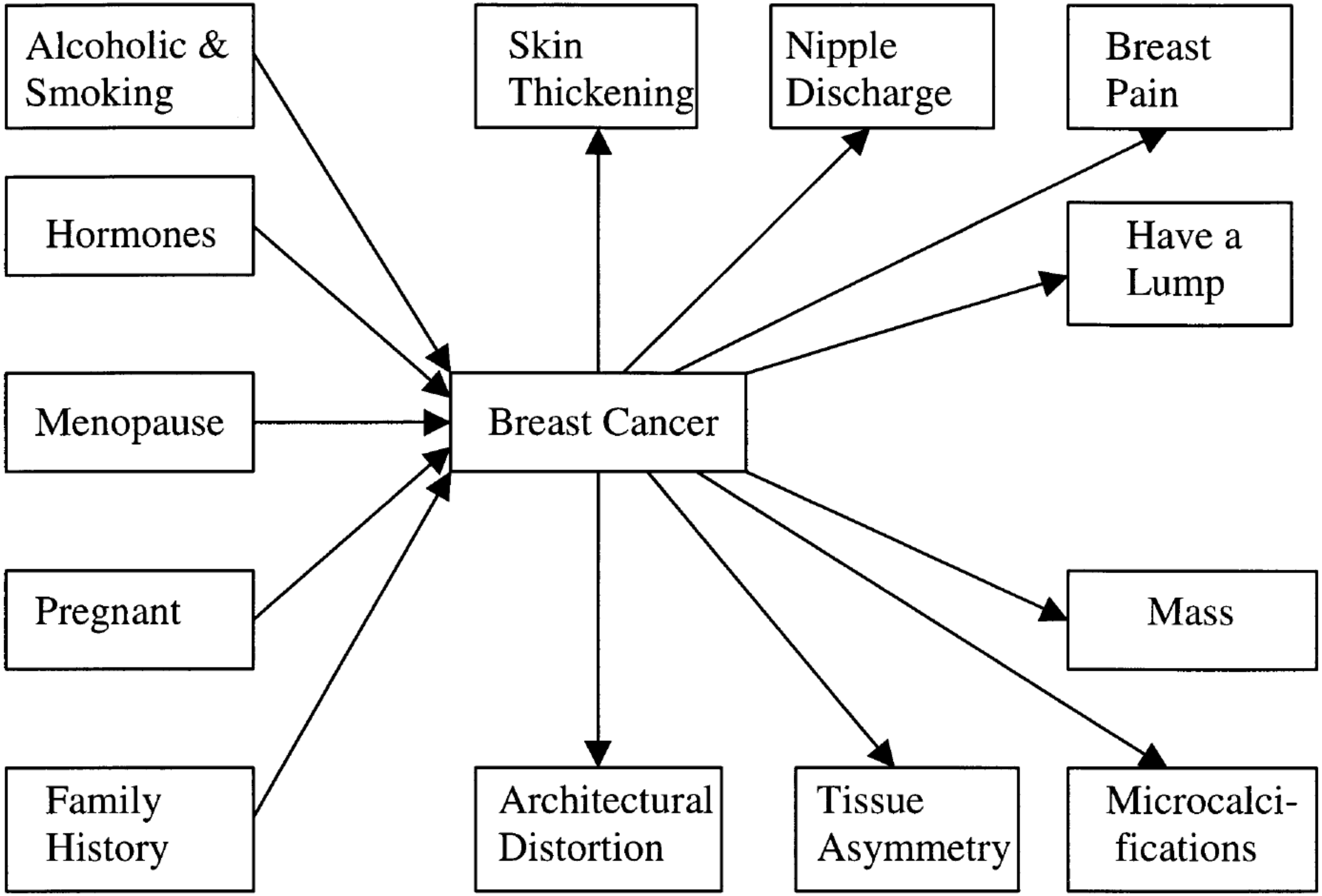


van Gerven, M. A. J., Taal, B. G. & Lucas, P. J. F. (2008) Dynamic Bayesian networks as prognostic models for clinical patient management. *Journal of Biomedical Informatics, 41, 4, 515-529.*

Example: Breast cancer - Probability Table

Category	Node description	State description
Diagnosis	Breast cancer	Present, absent.
Clinical history	Habit of drinking alcoholic beverages and smoking	Yes, no.
	Taking female hormones	Yes, no.
	Have gone through menopause	Yes, no.
	Have ever been pregnant	Yes, no.
	Family member has breast cancer	Yes, no.
Physical findings	Nipple discharge	Yes, no.
	Skin thickening	Yes, no.
	Breast pain	Yes, no.
	Have a lump(s)	Yes, no.
Mammographic findings	Architectural distortion	Present, absent.
	Mass	Score from one to three, score from four to five, absent
	Microcalcification cluster	Score from one to three, score from four to five, absent
	Asymmetry	Present, absent.

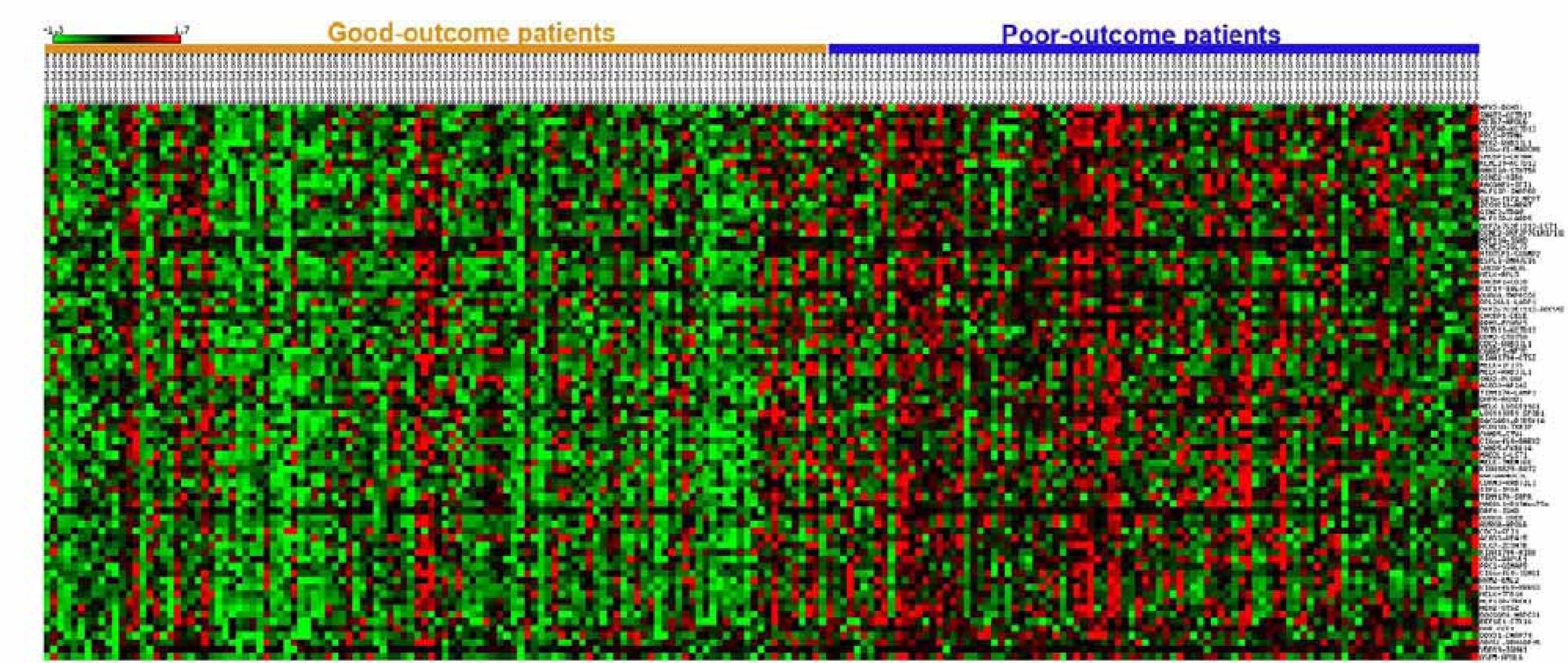
Wang, X. H., et al. (1999) Computer-assisted diagnosis of breast cancer using a data-driven Bayesian belief network. *International Journal of Medical Informatics, 54, 2, 115-126.*



Wang, X. H., et al. (1999) Computer-assisted diagnosis of breast cancer using a data-driven Bayesian belief network. *International Journal of Medical Informatics, 54, 2, 115-126.*

10 years later: Integration of microarray data

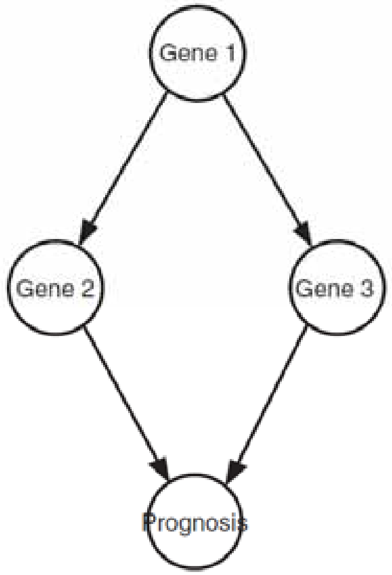
- Integrating microarray data from multiple studies to increase sample size;
- = approach to the development of more robust prognostic tests



Xu, L., Tan, A., Winslow, R. & Geman, D. (2008) Merging microarray data from separate breast cancer studies provides a robust prognostic test. *BMC Bioinformatics, 9, 1, 125-139.*

Example: BN 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

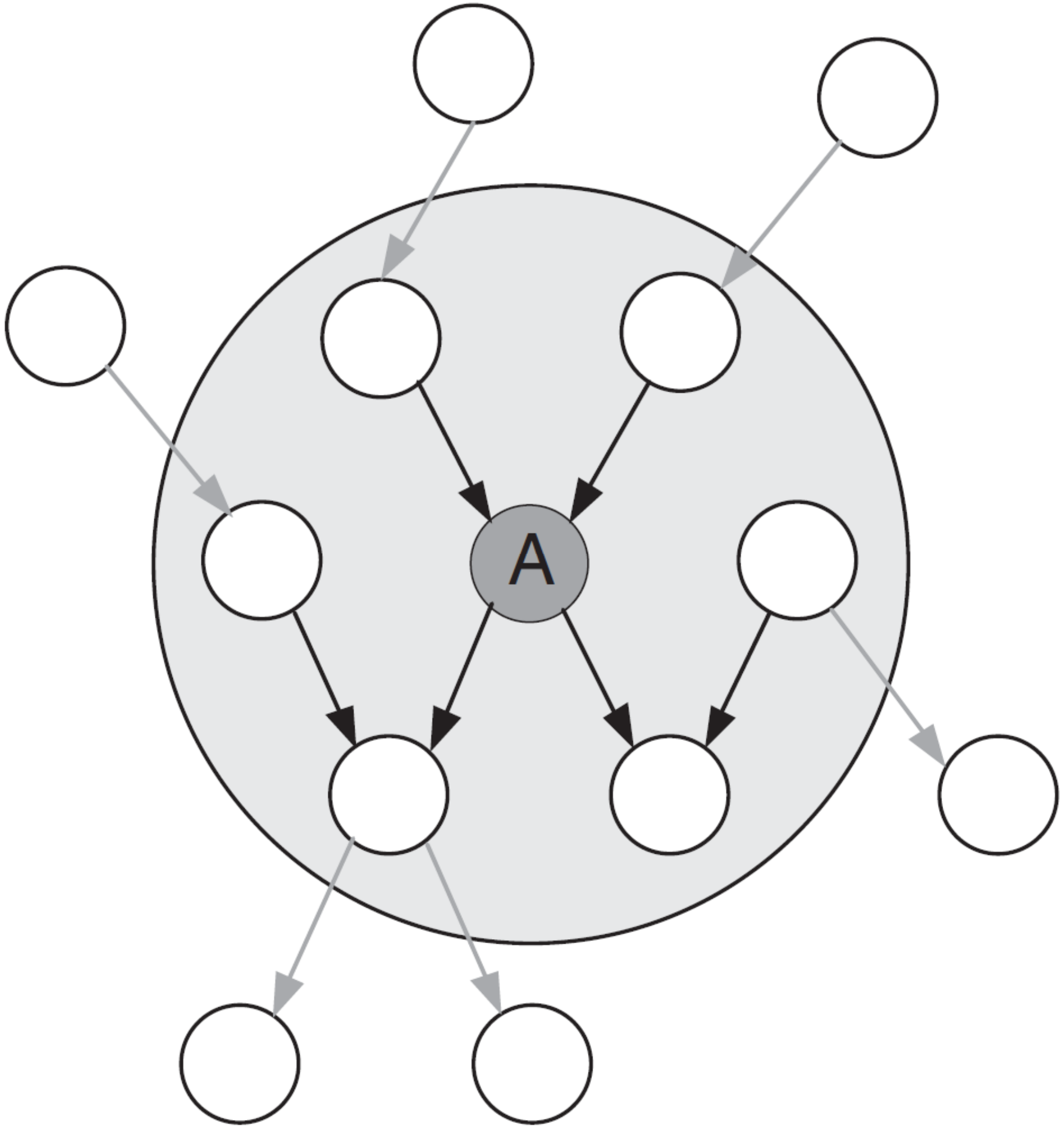
Gene 2	Gene 1 on	Gene 1 off
P(on)	0.3	0.6
P(off)	0.7	0.4

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.*

Concept Markov-Blanket

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.*



Dependency Structure -> first step (1/2)

- First the structure is learned using a search strategy.
- Since the number of possible structures increases super exponentially with the number of variables,
- the well-known greedy search algorithm K2 can be used in combination with the Bayesian Dirichlet (BD) scoring metric:

$$p(S|D) \propto p(S) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \left[\frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ijk})} \right]$$

N_{ijk} … number of cases in the data set D
having variable i in state k associated with the j-th instantiation of its parents in current structure S.
n is the total number of variables.

Dependency Structure – first step (2/2)

- Next, N_{ij} is calculated by summing over all states of a variable:
- $N_{ij} = \sum_{k=1}^{r_i} N_{ijk} \cdot N'_{ijk}$ and N'_{ij} have similar meanings but refer to prior knowledge for the parameters.
- When no knowledge is available they are estimated using $N_{ijk} = N/(r_i q_i)$
- with N the equivalent sample size,
- r_i the number of states of variable i and
- q_i the number of instantiations of the parents of variable i.
- $\Gamma(.)$ corresponds to the gamma distribution.
- Finally $p(S)$ is the prior probability of the structure.
- $p(S)$ is calculated by:
- $p(S) = \prod_{i=1}^{n} \prod_{l_i=1}^{p_i} p(l_i \longrightarrow x_i) \prod_{m_i=1}^{o_i} p(m_i x_i)$
- with p_i the number of parents of variable x_i and o_i all the variables that are not a parent of x_i.
- Next, $p(a \longrightarrow b)$ is the probability that there is an edge from a to b while $p(ab)$ is the inverse, i.e. the probability that there is no edge from a to b

Parameter learning -> second step

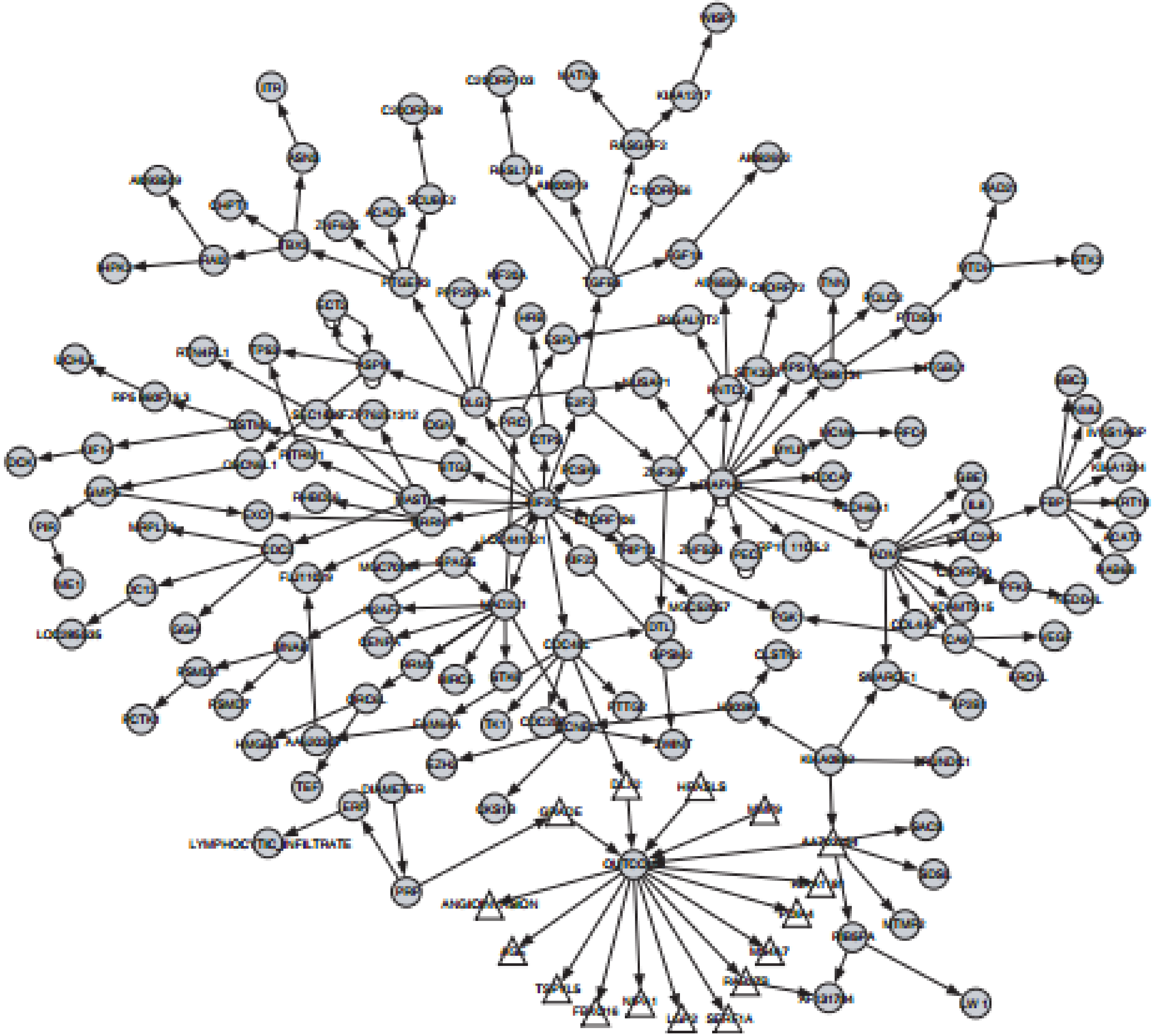
- Estimating the parameters of the local probability models corresponding with the dependency structure.
- CPTs are used to model these local probability models.
- For each variable and instantiation of its parents there exists a CPT that consists of a set of parameters.
- Each set of parameters was given a uniform Dirichlet prior:

$$\mathrm{p}\left(\theta_{ij} \middle| S\right) = Dir(\theta_{ij} | N'_{ij1}, \dots, N'_{ijk}, \dots, N'_{ijr_i})$$

Note: With θ_{ij} a parameter set where i refers to the variable and j to the j-th instantiation of the parents in the current structure. θ_{ij} contains a probability for every value of the variable x_i given the current instantiation of the parents. Dir corresponds to the Dirichlet distribution with $(N'_{ij1}, \dots, N'_{ijr_i})$ as parameters of this Dirichlet distribution. Parameter learning then consists of updating these Dirichlet priors with data. This is straightforward because the multinomial distribution that is used to model the data, and the Dirichlet distribution that models the prior, are conjugate distributions. This results in a Dirichlet posterior over the parameter set:

$$\mathrm{p}\left(\theta_{ij} \middle| D, S\right) = Dir(\theta_{ij} | N'_{ij1} + N_{ij1}, \dots, N'_{ijk} + N_{ijk}, \dots, N'_{ijr_i} + N_{ijr_i})$$

with N_{ijk} defined as before.



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.*

- For certain cases it is tractable if:
  - Just one variable is unobserved
  - We have singly connected graphs (no undirected loops -> belief propagation)
  - Assigning probability to fully observed set of variables
- Possibility: Monte Carlo Methods (generate many samples according to the Bayes Net distribution and then count the results)
- Otherwise: approximate solutions, NOTE:

Sometimes it is better to have an approximate solution to a complex problem – than a perfect solution to a simplified problem

Often it is better to have a good solution within time – than an perfect solution (much) later …

06 Graphical Model Learning

- Remember: GM are a marriage between probability theory and graph theory and provide a tool for dealing with our two grand challenges in the biomedical domain:

Uncertainty and complexity

- The learning task is two-fold:
  1) Learning unknown probabilities
  2) Learning unknown structures

Jordan, M. I. 1998. Learning in graphical models, Springer

Learning the Structure of GM from data

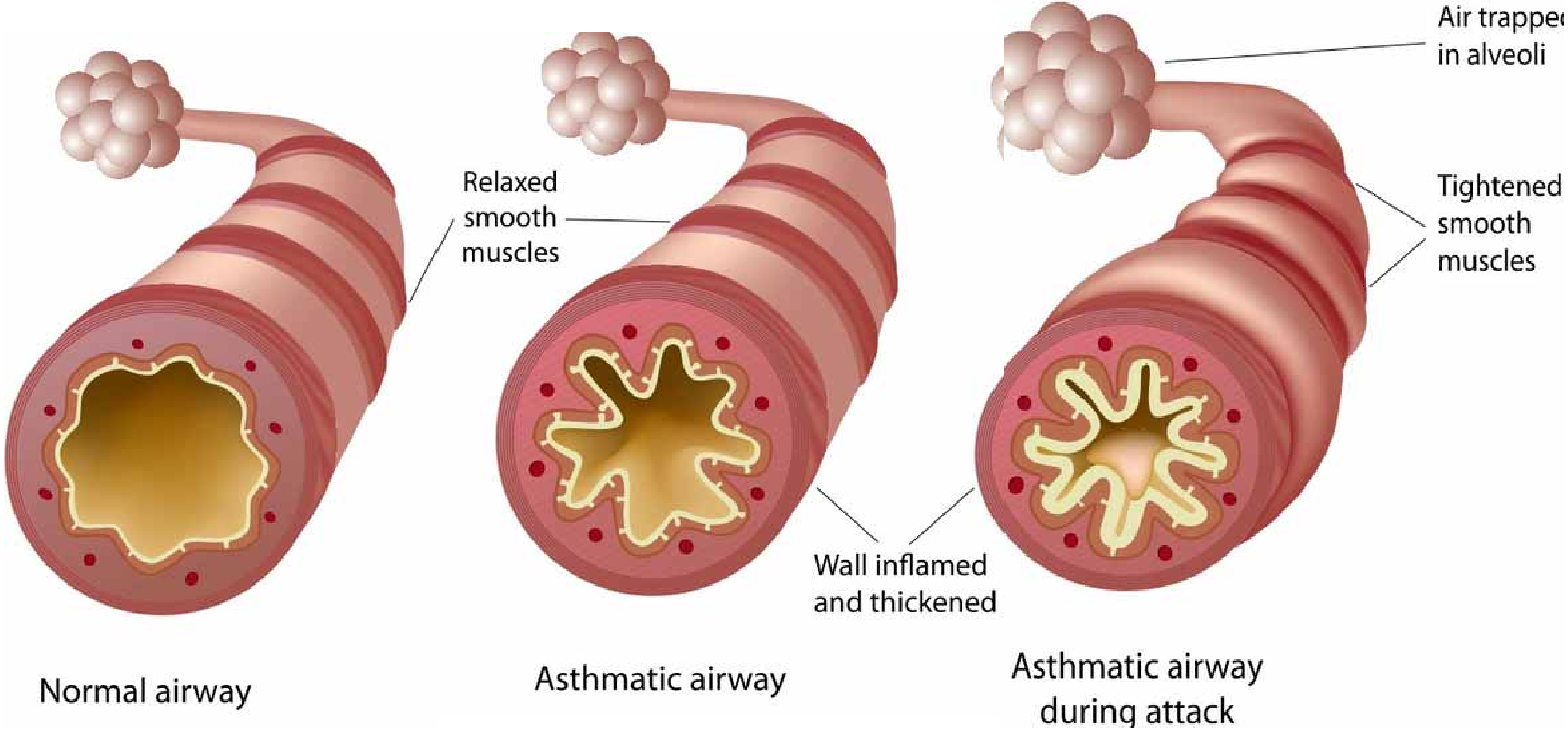
1) Test if a distribution is decomposable with regard to a given graph.
   - This is the most direct approach. It is not bound to a graphical representation,
   - It can be carried out w.r.t. other representations of the set of subspaces to be used to compute the (candidate) decomposition of a given distribution.
2) Find a suitable graph by measuring the strength of dependences.
   - This is a heuristic, but often highly successful approach, which is based on the frequently valid assumption that in a conditional independence graph an attribute is more strongly dependent on adjacent attributes than on attributes that are not directly connected to them.
3) Find an independence map by conditional independence tests.
   - This approach exploits the theorems that connect conditional independence graphs and graphs that represent decompositions.
   - It has the advantage that a single conditional independence test, if it fails, can exclude several candidate graphs. Beware, because wrong test results can thus have severe consequences.

Borgelt, C., Steinbrecher, M. & Kruse, R. R. 2009. Graphical models: representations for learning, reasoning and data mining, John Wiley & Sons.

A Question

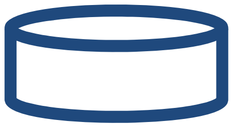
Who of you is NON-Smoker ?

Who of you is Smoker ?



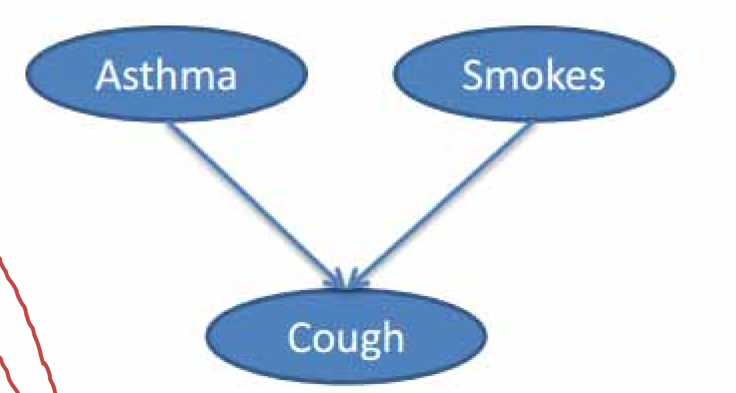
Beasley, R. 1998. Worldwide variation in prevalence of symptoms of asthma, allergic rhinoconjunctivitis, and atopic eczema: ISAAC. The Lancet, 351, (9111), 1225-1232, doi:http://dx.doi.org/10.1016/S0140-6736(97)07302-9.

Example for Graphical Model Learning



Bayesian Network

Patient	J46	Tussis	Smoker
Florian	1	1	0
Tamas	0	0	0
Matthias	1	0	0
Benjamin	0	1	1
Dimitrios	0	1	0
…			
…			

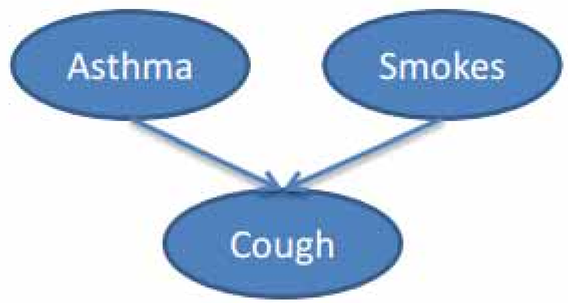


Florian	0	?	?

Rows are independent during learning and inference!

Florian	0	0.3	0.2

- Asthma can be hereditary
- Friends may have similar smoking habits
- Augmenting graphical model with relations between the entities – Markov Logic



2.1 Asthma ⇒ Cough

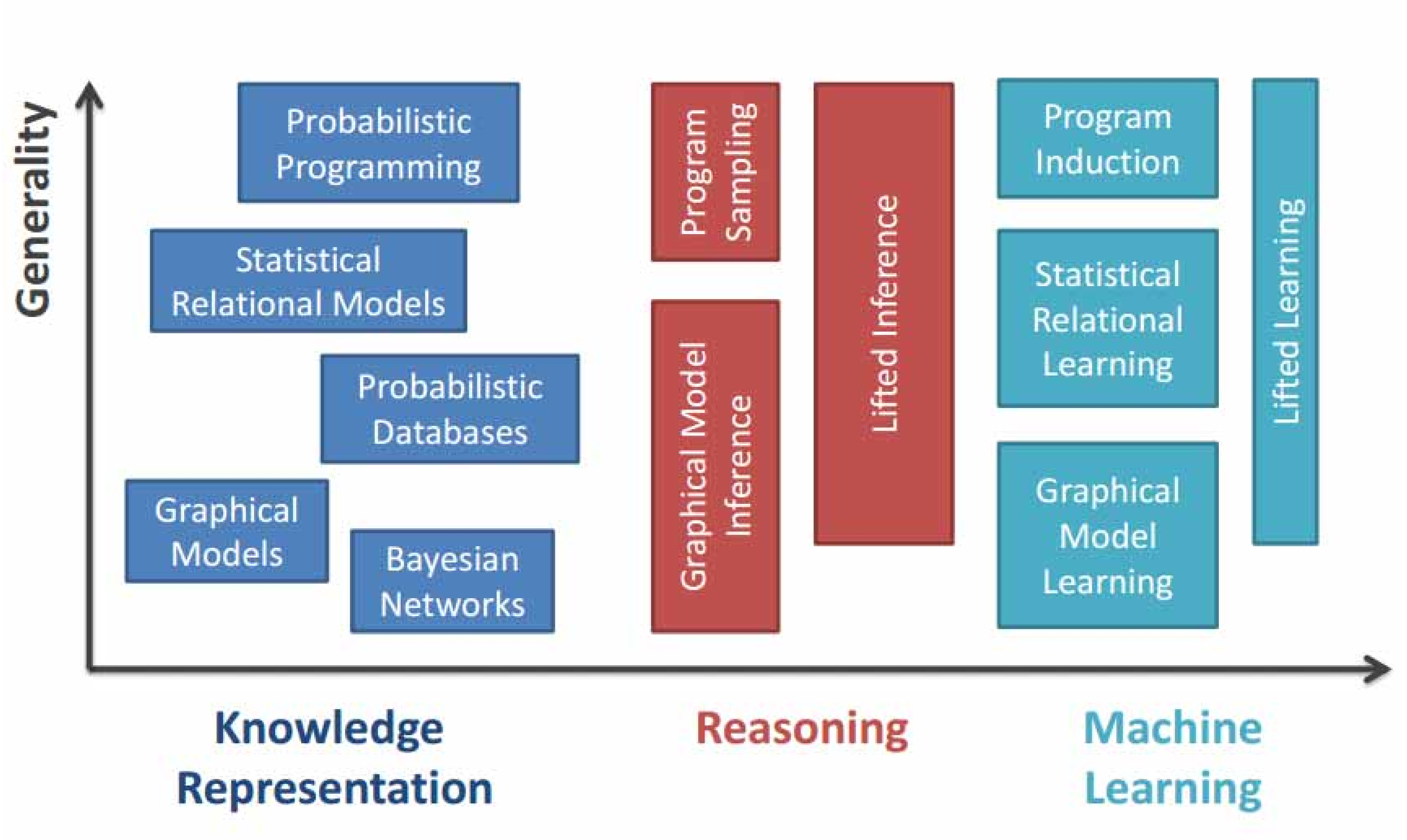
3.5 Smokes ⇒ Cough

2.1 Asthma(x) ⇒ Cough(x)

3.5 Smokes(x) ⇒ Cough(x)

1.9 Smokes(x) ∧ Friends(x,y) ⇒ Smokes(y)

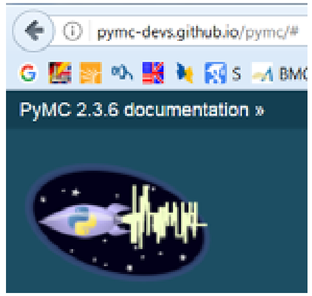
1.5 Asthma (x) ∧ Family(x,y) ⇒ Asthma (y)



Example for probabilistic rule learning, in which probabilistic rules are learned from probabilistic examples: The ProbFOIL+ Algorithm solves this problem by combining the principles of the rule learner FOIL with the probabilistic Prolog called ProbLog, see: De Raedt, L., Dries, A., Thon, I., Van Den Broeck, G. & Verbeke, M. 2015. Inducing probabilistic relational rules from probabilistic examples. International Joint Conference on Artificial Intelligence (IJCAI).

07 Probabilistic Programming

- C → Probabilistic-C
- Scala → Figaro
- Scheme → Church
- Excel → Tabular
- Prolog → Problog
- Javascript → webPP
- → Venture
- **Python → PyMC**



Probabilistic programs vs. graphical models

Probabilistic Program	Graphical Model
Variables	Variable nodes
Functions/operators	Factor nodes/edges
Fixed size loops/arrays	Plates
If statements	Gates (Minka & Winn)
Variable sized loops, Complex indexing, jagged arrays, mutation, recursion, objects/ properties…	No common equivalent

Medical Example

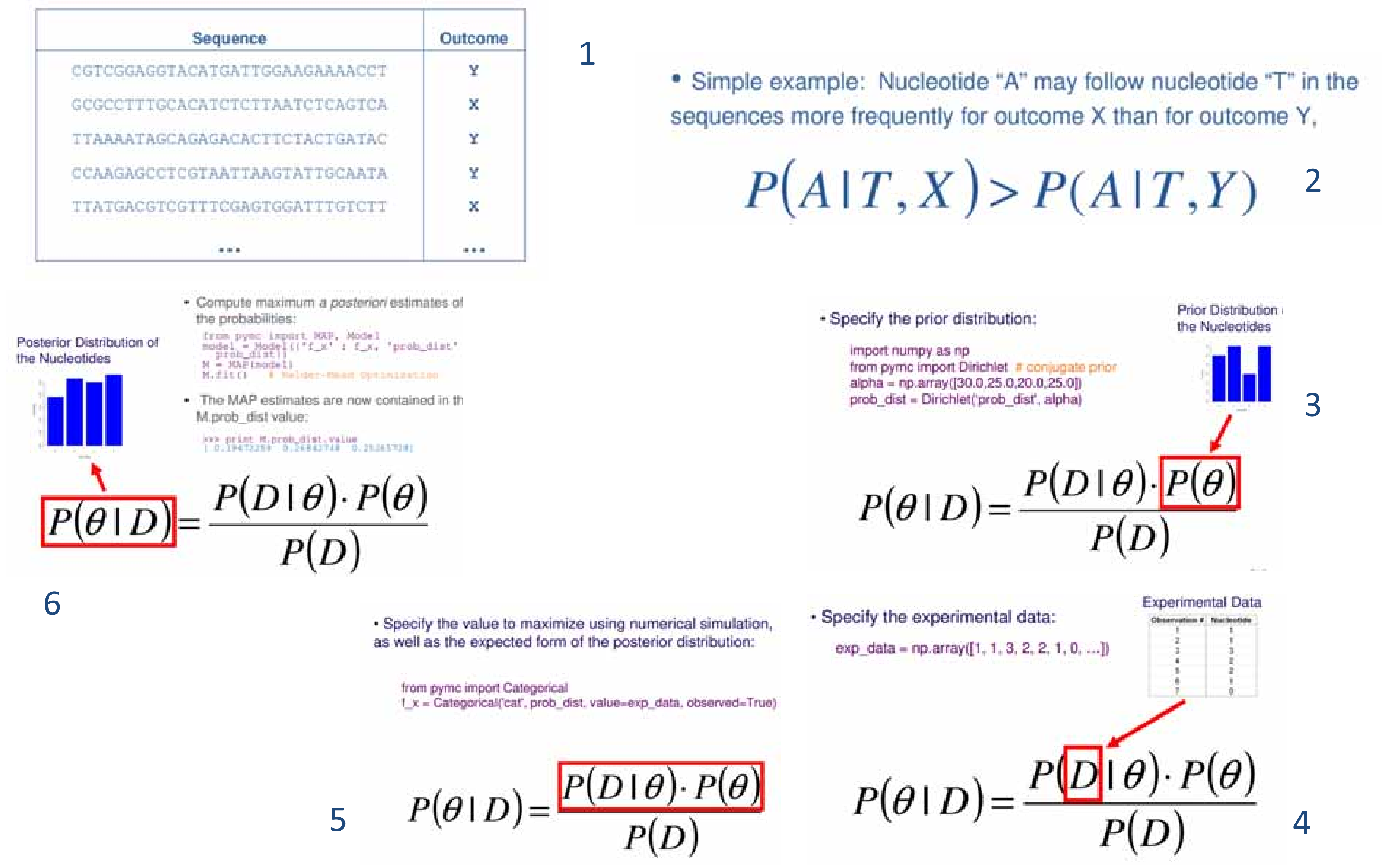
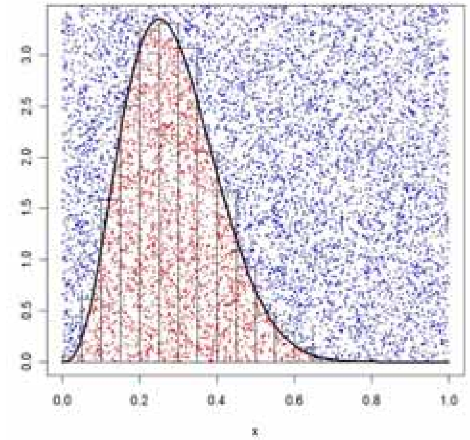
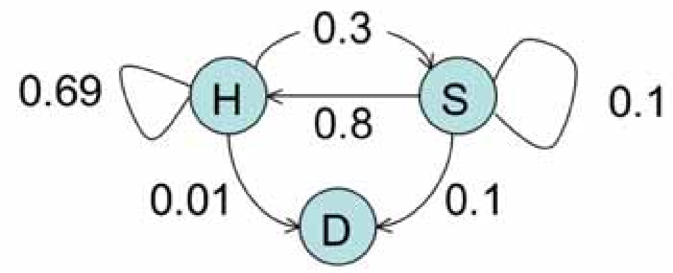


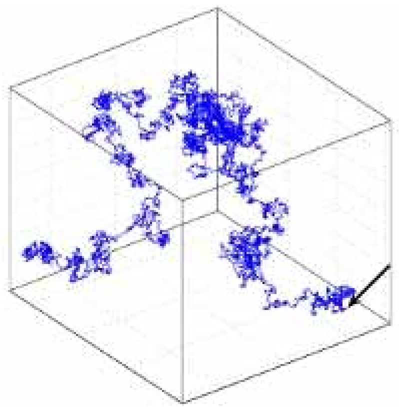
Image Source: Dan Williams, Life Technologies, Austin TX

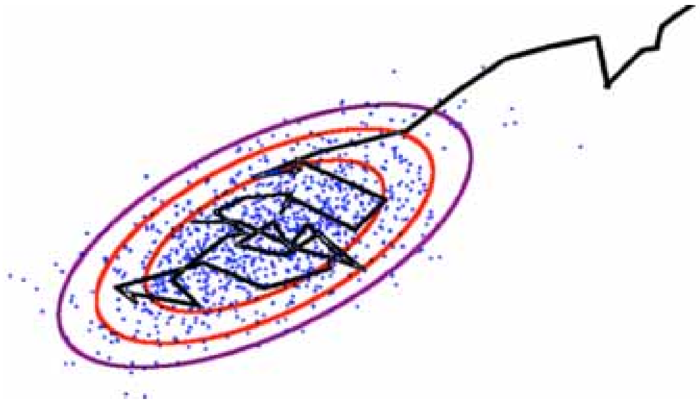
08 Markov Chain Monte Carlo (MCMC)

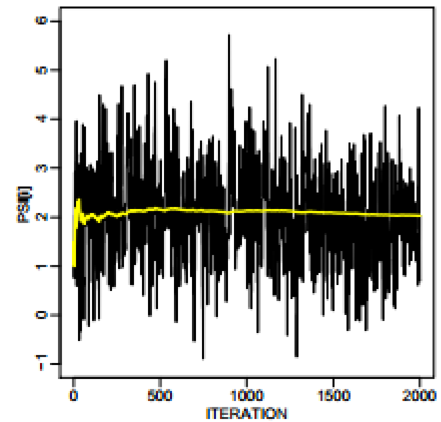
Monte Carlo Method (MC)
Monte Carlo Sampling
Markov Chains (MC)
MCMC
Metropolis-Hastings











- often we want to calculate characteristics of a **high-dimensional** probability distribution … $p(\mathcal{D}|\theta)$

$$p(h|d) \propto p(\mathcal{D}|\theta) * p(h)$$

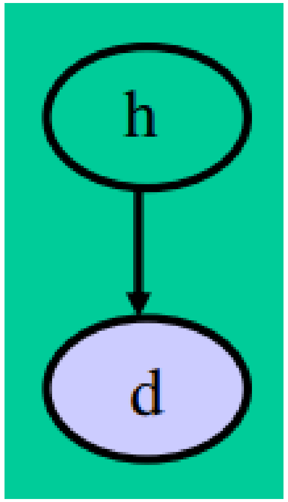
Posterior integration problem: (almost) all statistical inference can be deduced from the posterior distribution by calculating the appropriate sums, which involves an integration:

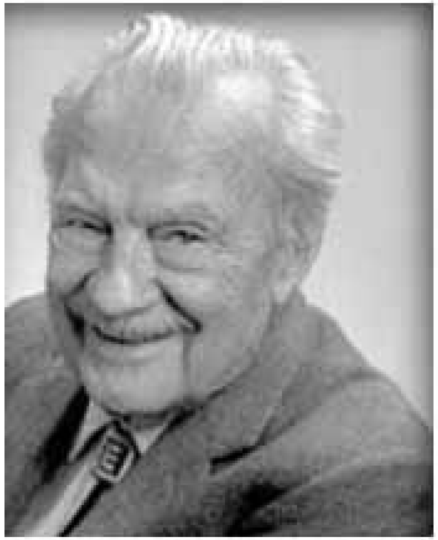
$$J = \int f(\theta) * p(\theta|\mathcal{D}) d\theta$$

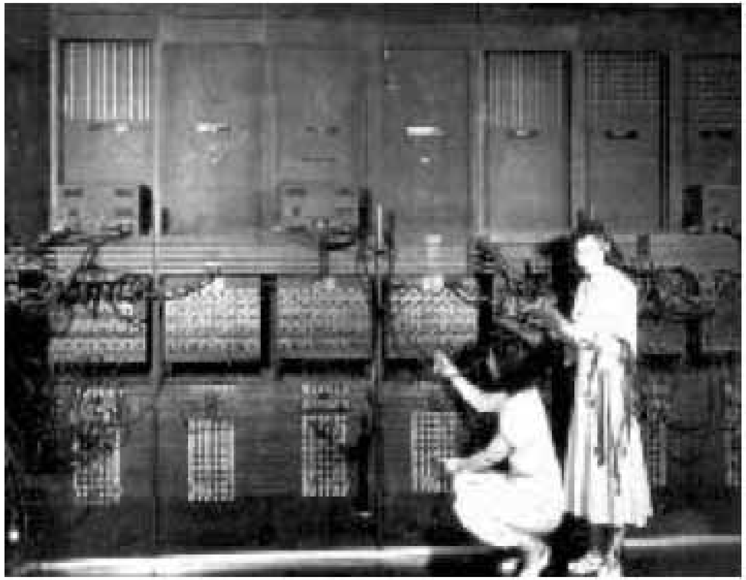
- **Statistical physics:** computing the partition function – this is evaluating the posterior probability of a hypothesis and this requires summing over all hypotheses … remember:

$$\mathcal{H} = \{H_1, H_2, ..., H_n\} \quad \forall (h,d)$$

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





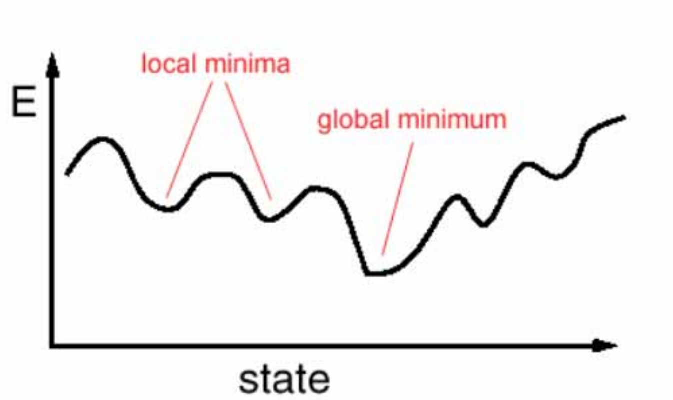
named after

- Class of algorithms that rely on **repeated random sampling**
- Basic idea: using **randomness** to solve problems with high uncertainty (Laplace, 1781)
- For solving **multidimensional integrals** which would otherwise intractable
- For simulation of systems with **many dof**
- e.g. fluids, gases, particle collectives, **cellular structures** - see our last tutorial on Tumor growth simulation!

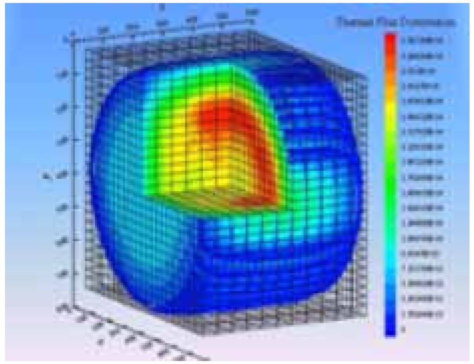
- for solving problems of probabilistic inference involved in developing computational models
- as a source of hypotheses about how the human mind might solve problems of inference
- For a function $f(x)$ and distribution $P(x)$, the expectation of f with respect to P is generally the average of f, when x is drawn from the probability distribution $P(x)$

$$\mathbb{E}_{p(x)}(f(x)) = \sum_{X} f(x)P(x)dx$$

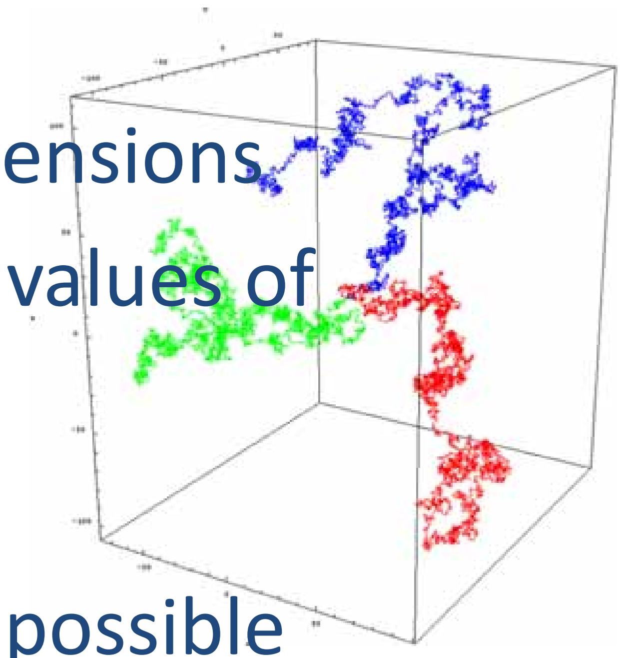
- Solving intractable integrals
- Bayesian statistics: **normalizing** constants, expectations, marginalization



- Stochastic Optimization
- Generalization of simulated annealing
- Monte Carlo expectation maximization (EM)



- Physical simulation
- estimating neutron diffusion time
- Computing expected utilities and best responses toward Nash equilibria
- Computing volumes in high-dimensions
- Computing eigen-functions and values of operators (e.g. Schrödinger)
- Statistical physics
- Counting many things as fast as possible



5,223 citations as of 26.03.2017

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

Number 247 SEPTEMBER 1949 *Volume 44*

THE MONTE CARLO METHOD

NICHOLAS METROPOLIS AND S. ULAM
Los Alamos Laboratory

We shall present here the motivation and a general description of a method dealing with a class of problems in mathematical physics. The method is, essentially, a statistical approach to the study of differential equations, or more generally, of integro-differential equations that occur in various branches of the natural sciences.

ALREADY in the nineteenth century a sharp distinction began to appear between two different mathematical methods of treating physical phenomena. Problems involving only a few particles were studied in classical mechanics, through the study of systems of ordinary differential equations. For the description of systems with very many particles, an entirely different technique was used, namely, the method of statistical mechanics. In this latter approach, one does not concentrate on the individual particles but studies the properties of *sets of particles*. In pure mathematics an intensive study of the properties of sets of points was the subject of a new field. This is the so-called theory of sets, the basic theory of integration, and the twentieth century development of the theory of probabilities prepared the formal apparatus for the use of such models in theoretical physics, i.e., description of properties of aggregates of points rather than of individual points and

Image Source:
http://www.manhattanprojectvoices.org/oral-histories/nicholas-metropolis-interview

34,140 citations (as of 26.03.2017)

THE JOURNAL OF CHEMICAL PHYSICS VOLUME 21, NUMBER 6 JUNE, 1953

Equation of State Calculations by Fast Computing Machines

NICHOLAS METROPOLIS, ARIANNA W. ROSENBLUTH, MARSHALL N. ROSENBLUTH, AND AUGUSTA H. TELLER,
Los Alamos Scientific Laboratory, Los Alamos, New Mexico

AND

EDWARD TELLER,* *Department of Physics, University of Chicago, Chicago, Illinois*

(Received March 6, 1953)

A general method, suitable for fast computing machines, for investigating such properties as equations of state for substances consisting of interacting individual molecules is described. The method consists of a modified Monte Carlo integration over configuration space. Results for the two-dimensional rigid-sphere system have been obtained on the Los Alamos MANIAC and are presented here. These results are compared to the free volume equation of state and to a four-term virial coefficient expansion.

I. INTRODUCTION

THE purpose of this paper is to describe a general method, suitable for fast electronic computing machines, of calculating the properties of any substance which may be considered as composed of interacting individual molecules. Classical statistics is assumed, only two-body forces are considered, and the potential field of a molecule is assumed spherically symmetric. These are the usual assumptions made in theories of liquids. Subject to the above assumptions, the method is not restricted to any range of temperature or density. This paper will also present results of a preliminary two-dimensional calculation for the rigid-sphere system. Work on the two-dimensional case with a Lennard-Jones potential is in progress and will be reported in a later paper. Also, the problem in three dimensions is being investigated.

* Now at the Radiation Laboratory of the University of California, Livermore, California.

II. THE GENERAL METHOD FOR AN ARBITRARY POTENTIAL BETWEEN THE PARTICLES

In order to reduce the problem to a feasible size for numerical work, we can, of course, consider only a finite number of particles. This number N may be as high as several hundred. Our system consists of a square† containing N particles. In order to minimize the surface effects we suppose the complete substance to be periodic, consisting of many such squares, each square containing N particles in the same configuration. Thus we define d_{AB}, the minimum distance between particles A and B, as the shortest distance between A and any of the particles B, of which there is one in each of the squares which comprise the complete substance. If we have a potential which falls off rapidly with distance, there will be at most one of the distances AB which can make a substantial contribution; hence we need consider only the minimum distance d_{AB}.

† We will use the two-dimensional nomenclature here since it is easier to visualize. The extension to three dimensions is obvious.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. 1953. Equation of State Calculations by Fast Computing Machines. The Journal of Chemical Physics, 21, (6), 1087-1092, doi:10.1063/1.1699114.

10,624 citations as of 26.03.2017

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika, 57, (1), 97-109.

Biometrika (1970), **57**, 1, *p.* 97
Printed in Great Britain

Monte Carlo sampling methods using Markov chains and their applications

BY W. K. HASTINGS
University of Toronto

SUMMARY

A generalization of the sampling method introduced by Metropolis *et al.* (1953) is presented along with an exposition of the relevant theory, techniques of application and methods and difficulties of assessing the error in Monte Carlo estimates. Examples of the methods, including the generation of random orthogonal matrices and potential applications of the methods to numerical problems arising in statistics, are discussed.

1. INTRODUCTION

For numerical problems in a large number of dimensions, Monte Carlo methods are often more efficient than conventional numerical methods. However, implementation of the Monte Carlo methods requires sampling from high dimensional probability distributions and this may be very difficult and expensive in analysis and computer time. General methods for sampling from, or estimating expectations with respect to, such distributions are as follows.

(i) If possible, factorize the distribution into the product of one-dimensional conditional distributions from which samples may be obtained.

(ii) Use importance sampling, which may also be used for variance reduction. That is, in order to evaluate the integral

$$J = \int f(x)p(x)\,dx = E_p(f),$$

where $p(x)$ is a probability density function, instead of obtaining independent samples $x_1, \ldots, x_N$ from $p(x)$ and using the estimate $\hat{J}_1 = \Sigma f(x_i)/N$, we instead obtain the sample from

- Expectation of a function $f(x, y)$ with respect to a random variable x is denoted by $\mathbb{E}_x\ [f(x, y)]$
- In situations where there is no ambiguity as to which variable is being averaged over, this will be simplified by omitting the suffix, for instance $\mathbb{E}x$.
- If the distribution of x is conditioned on another variable z, then the corresponding conditional expectation will be written $Ex[f(x)|z]$
- Similarly, the variance is denoted $var[f(x)]$, and for vector variables the covariance is written $cov[x, y]$

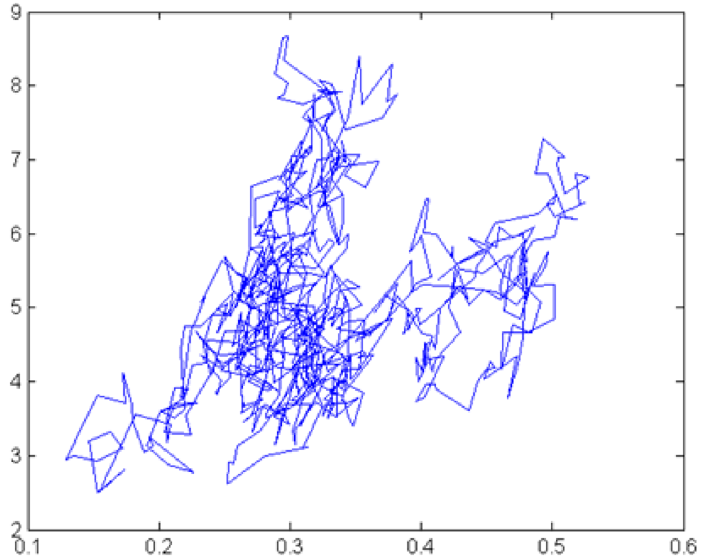
$$\underset{x}{\operatorname{argmax}} f(x)$$

Normalization: $$p(x|y) = \frac{p(y|x) * p(x)}{\int_X p(y|x) * p(x)dx}$$

Marginalization: $$p(x) = \int_Z p(x, z)dz$$

Expectation: $$\mathbb{E}_{p(x)}(f(x)) = \int_X f(x)p(x)dx$$

09 Metropolis-Hastings Algorithm



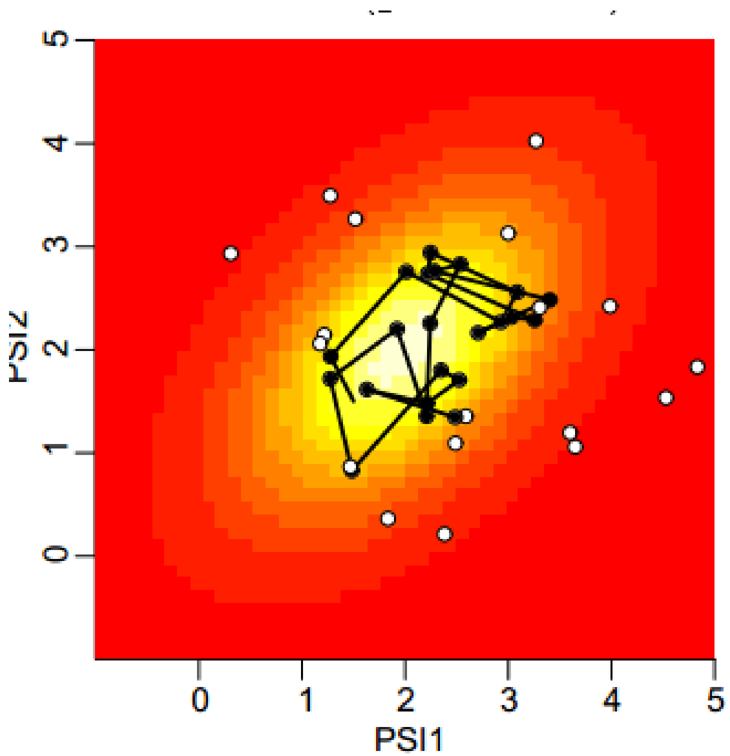
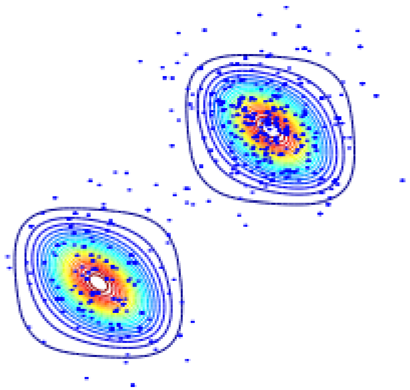


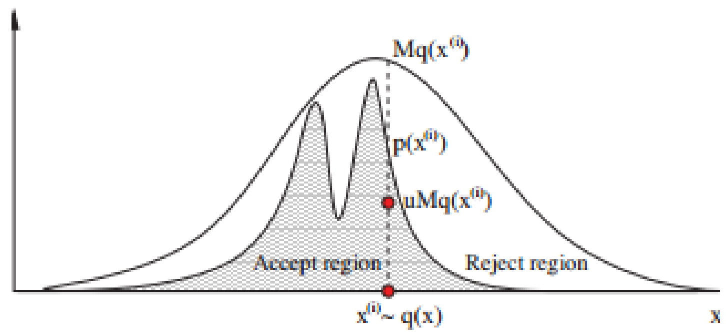
Image Source: Peter Mueller, Anderson Cancer Center

Metropolis Hastings MCMC sampling

Barber, D. 2012. Bayesian reasoning and machine learning, Cambridge, Cambridge University Press, p. 500

Choose a starting point x^1.
for $i = 2$ to L **do**
  Draw a candidate sample x^{cand} from the proposal $\tilde{q}(x'|x^{l-1})$.
  Let $a = \frac{\tilde{q}(x^{l-1}|x^{cand})p(x^{cand})}{\tilde{q}(x^{cand}|x^{l-1})p(x^{l-1})}$
  if $a \geq 1$ **then** $x^l = x^{cand}$
  else
    draw a random value u uniformly from the unit interval $[0, 1]$.
    if $u < a$ **then** $x^l = x^{cand}$
    else
      $x^l = x^{l-1}$
    end if
  end if
end for





Importance sampling

- Importance sampling is a technique to approximate averages with respect to an intractable distribution $p(x)$.
- The term 'sampling' is arguably a misnomer since the method does not attempt to draw samples from $p(x)$.
- Rather the method draws samples from a simpler importance distribution $q(x)$ and then reweights them
- such that averages with respect to $p(x)$ can be approximated using the samples from $q(x)$.

- The Gibbs Sampler is an interesting special case of MH:

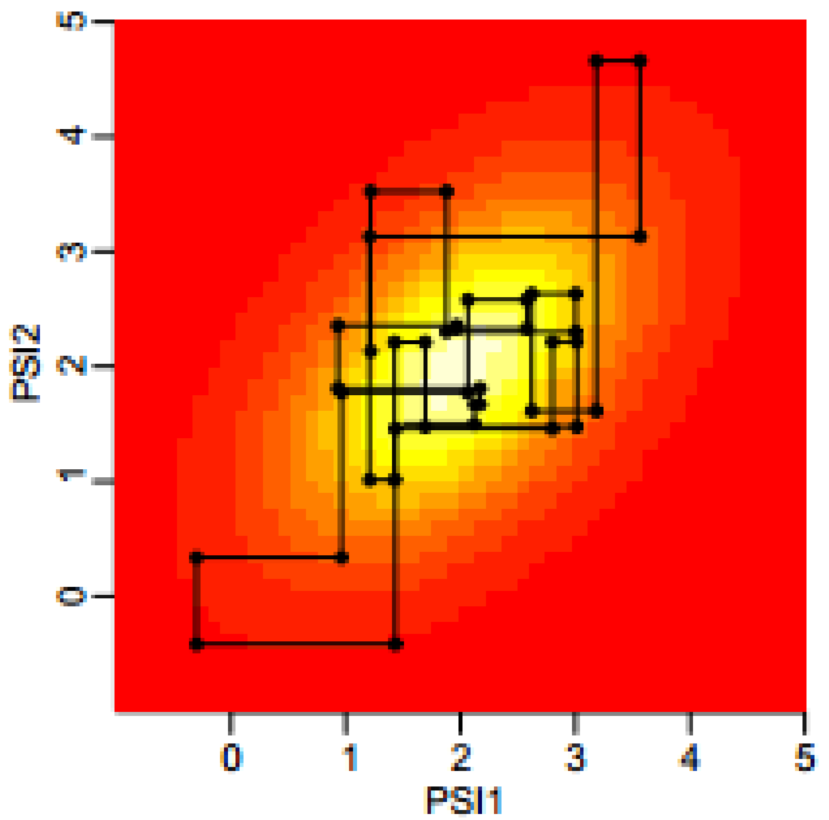
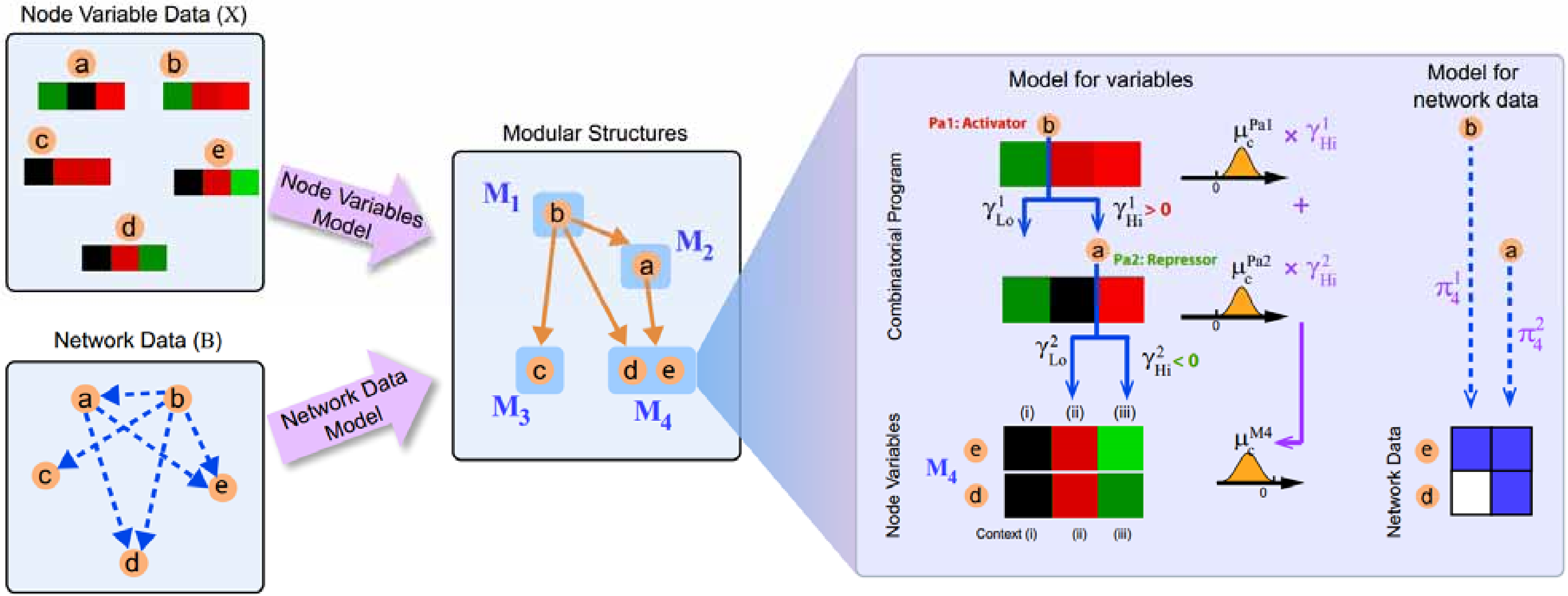
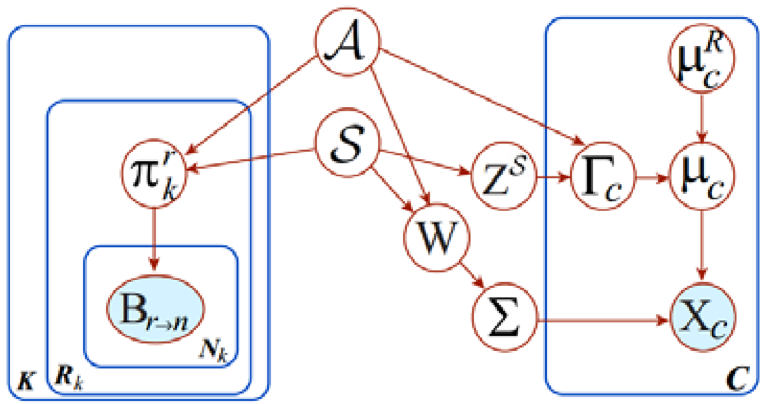


Image Source: Peter Mueller, Anderson Cancer Center

Sample

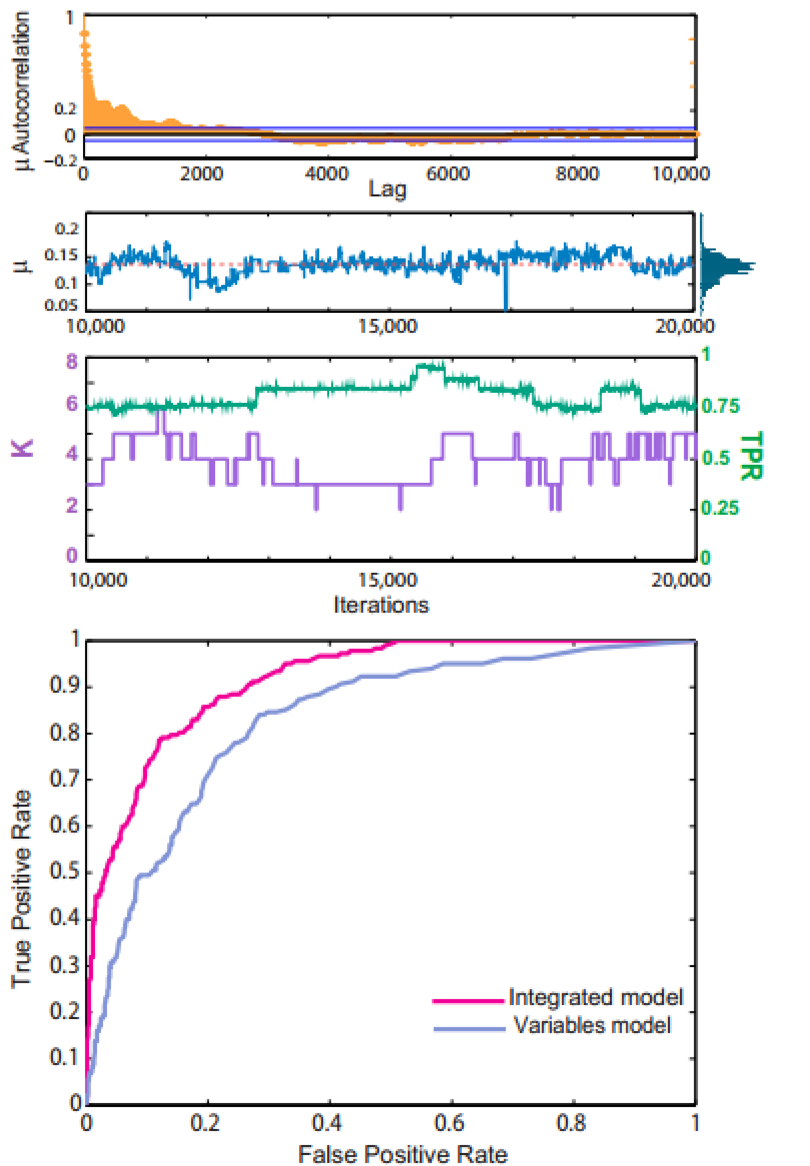


Azizi, E., Airoldi, E. M. & Galagan, J. E. 2014. Learning Modular Structures from Network Data and Node Variables. Proceedings of the 31st International Conference on Machine Learning (ICML). Beijing: JMLR. 1440-1448.



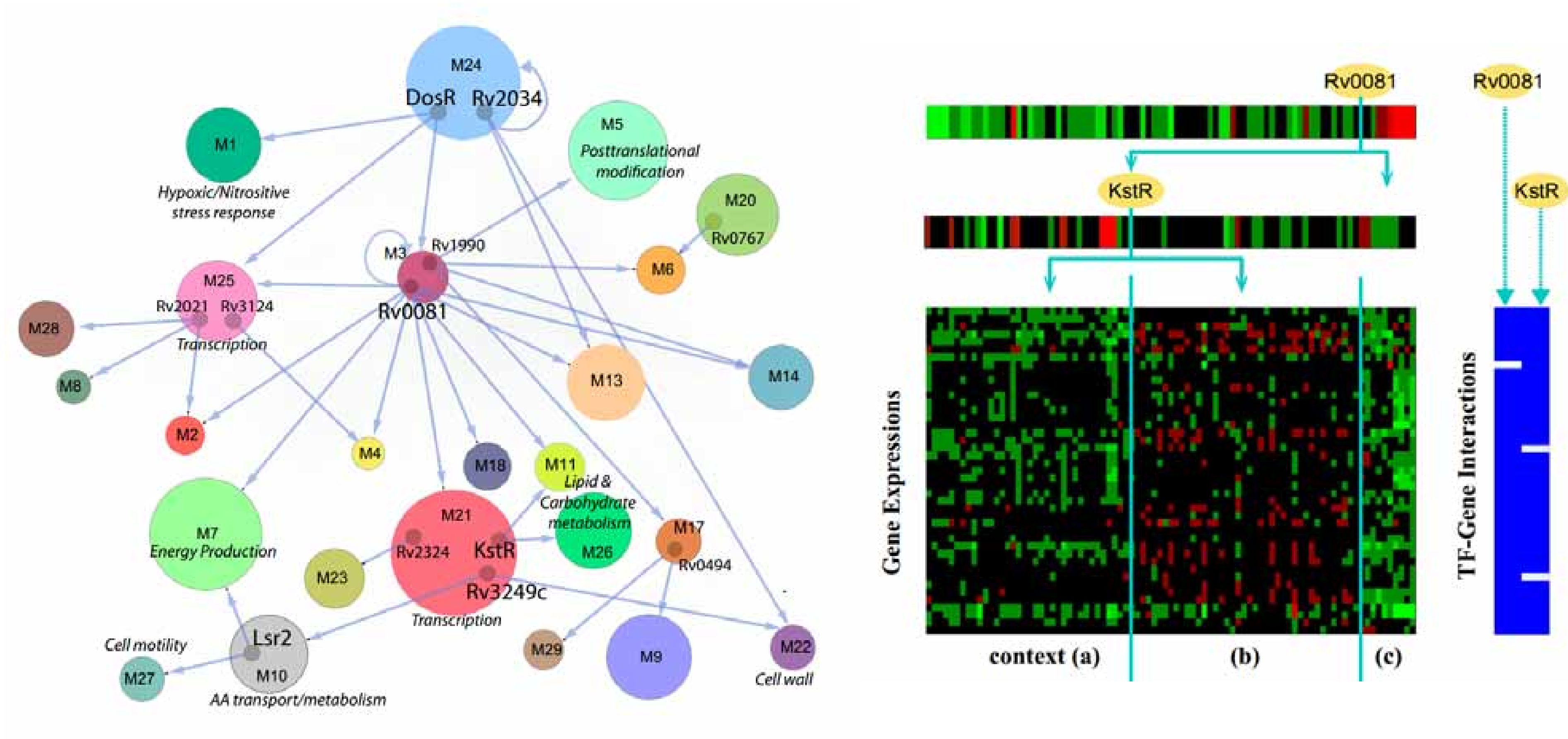
Algorithm 1 RJMCMC for sampling parameters

Inputs:
Node Variables Data $\mathbf{X}$
Network Data $\mathbf{B}$
for iterations $j = 1$ **to** J **do**
  Sample $\mathcal{A}^{(j+1)}$ given $\mathcal{A}^{(j)}$ using Alg 2 in (Azizi et al., 2014)
  Sample $\mathcal{S}^{(j+1)}$ given $\mathcal{S}^{(j)}$ using Alg 3 in (Azizi et al., 2014)
  for modules $k = 1$ **to** $K^{(j)}$ **do**
    Propose $w_k^{(j+1)} \sim \mathcal{N}(w_k^{(j)}, I)$
    Accept with probability P_{mh}; update $\Sigma^{(j+1)}$
    for parents $r = 1$ **to** R_k **do**
      Propose $z_k^{r(j+1)} \sim \mathcal{N}(z_k^{r(j)}, I)$; accept with P_{mh}
      Propose $\pi_k^{r(j+1)} \sim \mathcal{N}(\pi_k^{r(j)}, I)$; accept with P_{mh}
    end for
  end for
  for condition $c = 1$ **to** C **do**
    Propose $\boldsymbol{\mu}_c^{\mathbf{R}(j+1)} \sim \mathcal{N}(\boldsymbol{\mu}_c^{\mathbf{R}(j)}, I)$; accept with P_{mh}
    Propose $\gamma_c^{\mathbf{R}(j+1)} \sim \mathcal{N}(\gamma_c^{\mathbf{R}(j)}, I)$; accept with P_{mh}
  end for
end for

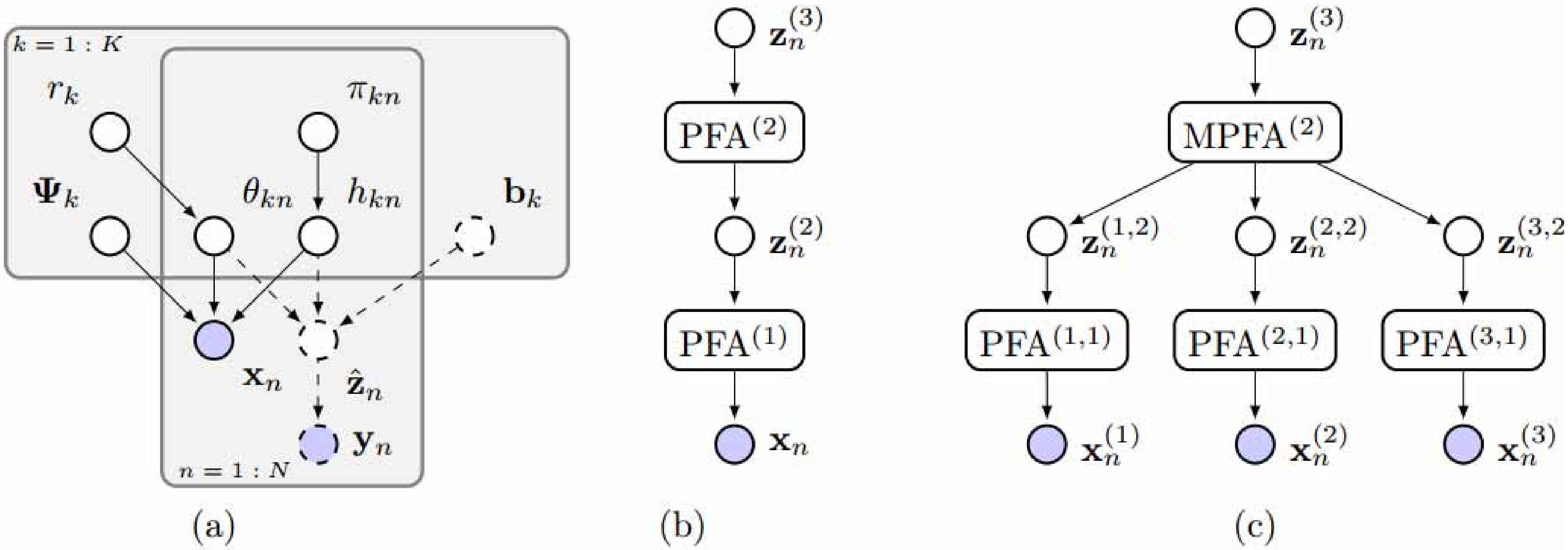


Azizi, E., Airoldi, E. M. & Galagan, J. E. 2014. Learning Modular Structures from Network Data and Node Variables. Proceedings of the 31st International Conference on Machine Learning (ICML). Beijing: JMLR. 1440-1448.

Myobacterium tuberculosis Gene Regulatory Network

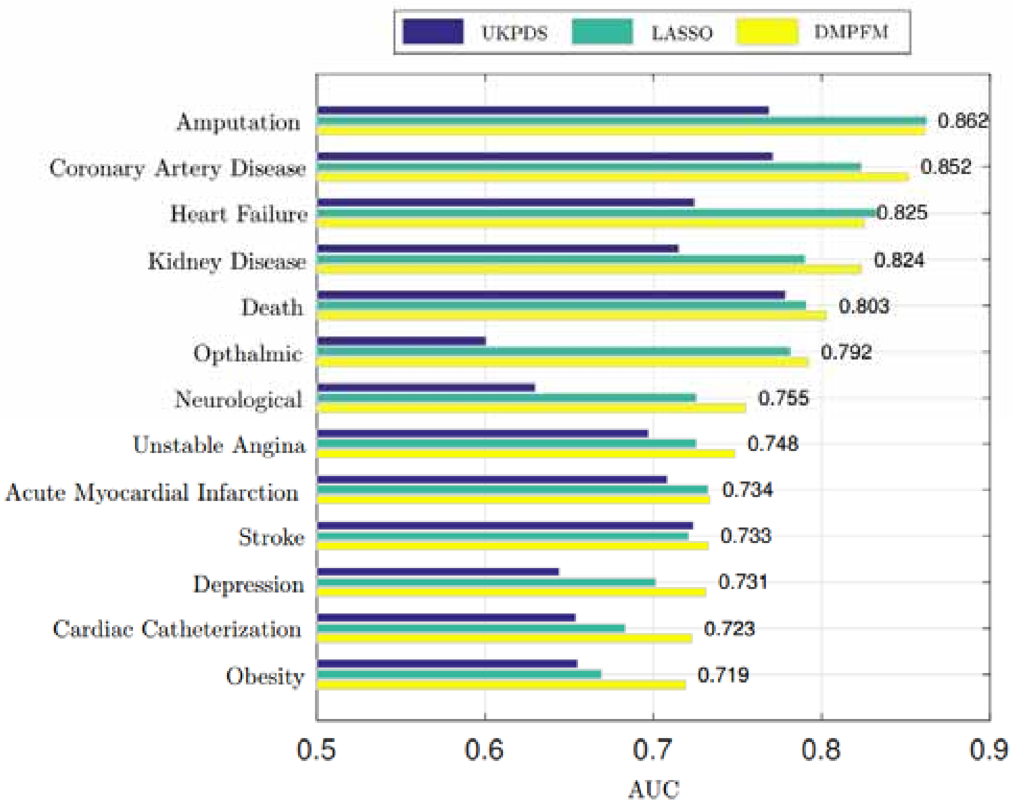


Azizi, E., Airoldi, E. M. & Galagan, J. E. 2014. Learning Modular Structures from Network Data and Node Variables. Proceedings of the 31st International Conference on Machine Learning (ICML). Beijing: JMLR. 1440-1448.

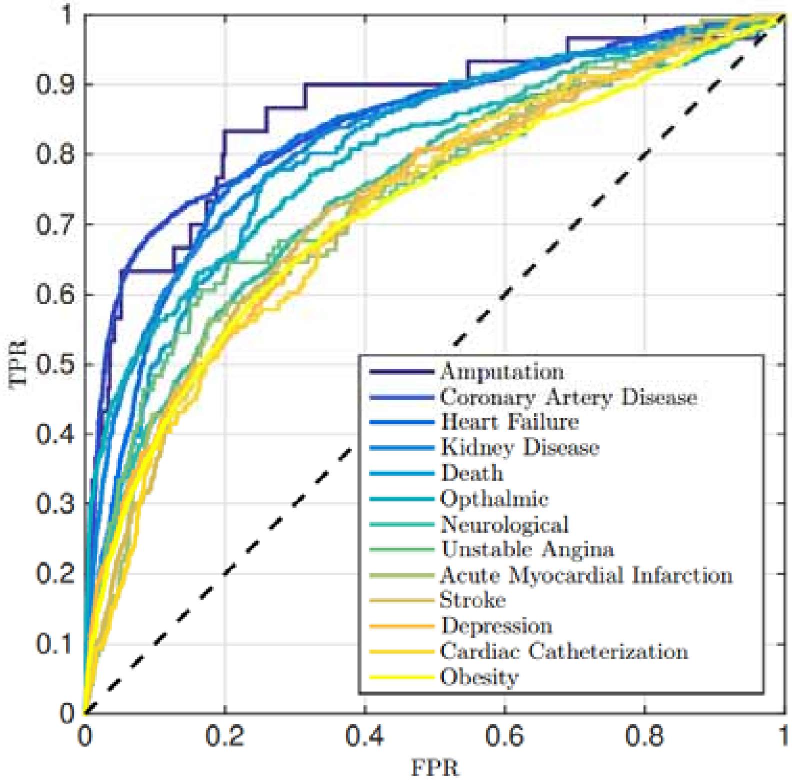


Henao, R., Lu, J. T., Lucas, J. E., Ferranti, J. & Carin, L. 2016. Electronic health record analysis via deep poisson factor models. Journal of Machine Learning Research JMLR, 17, 1-32.

MCMC based DPFM outperforms other approaches



(a)



(b)

Henao, R., Lu, J. T., Lucas, J. E., Ferranti, J. & Carin, L. 2016. Electronic health record analysis via deep poisson factor models. Journal of Machine Learning Research JMLR, 17, 1-32.

Still … there are a lot of open problems and challenges to solve …

no chance to retire!

Thank you!

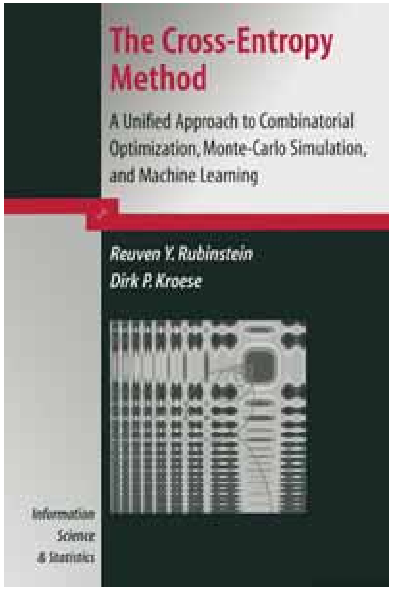
Questions

- What is the main difference between the ideas of Pierre Simon de Laplace and Lady Lovelace?
- What is medical action consiting most of the time?
- How does a human make a decision - as far as we know?
- What is the main idea of a probabilistic programming language?
- Why did Judea Pearl receive the Turing Award (Noble Prize in Computer Science)?
- What fields are coming together in PGM?
- What are the challenges in network structures?
- Give a classification of Graphical Models!
- What are plates and nested plates?
- Provide corresponding examples of metabolic networks!

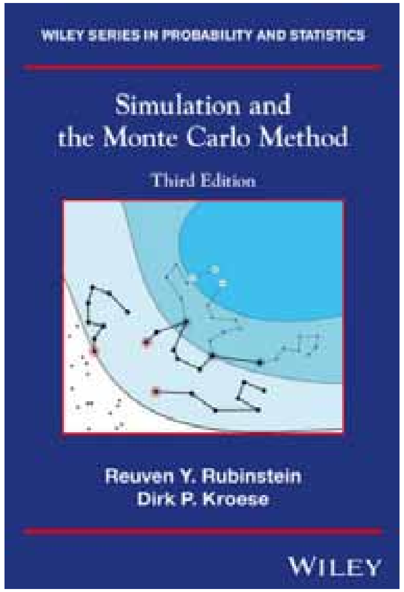
- What is a factored graph?
- Describe the protein structure prediction problem! Why is it hard?
- Why are protein-protein interactions so important?
- Describe the problem of graph-isomorphism!
- How does a Bayes Net work?
- Why is predicting important in clincial medicine?
- What is a Markov-Blankett?
- Which two tasks do we have in Graphical Model Learning?
- Why would we need probabilistic programming lanugages?
- Describe the main idea of MCMC!
- What is the main problem in marginalization?
- What is the benefit of the MH Algorithm?

Appendix

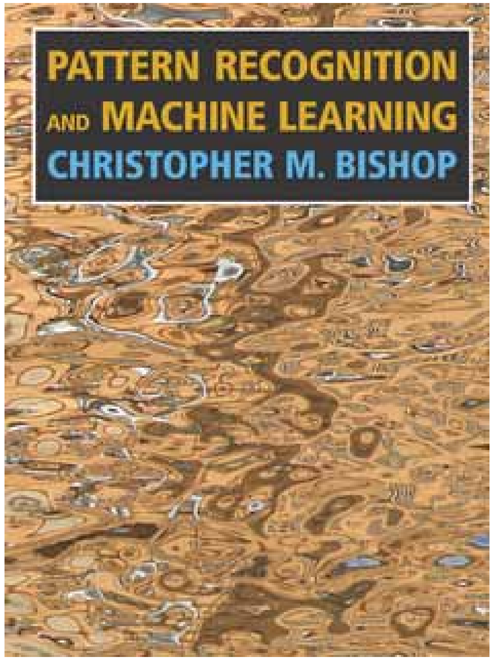
Some more specialist literature

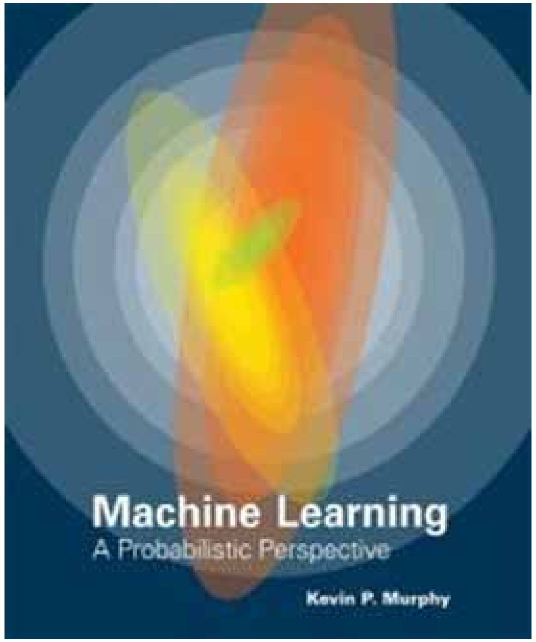


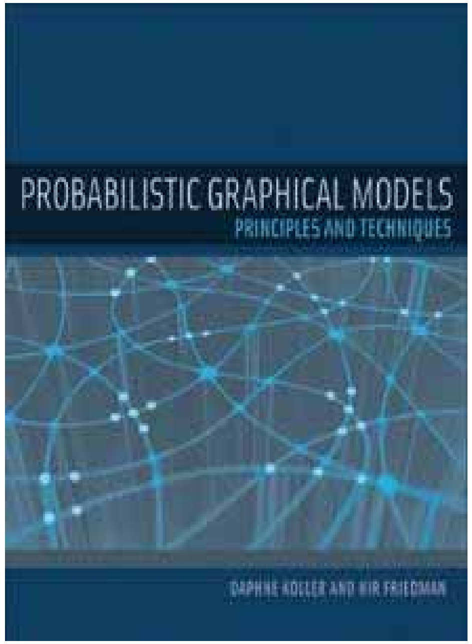
Rubinstein, R. Y. & Kroese, D. P. 2013. The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning, Springer



Rubinstein, R. Y. & Kroese, D. P. 2013. Simulation and the Monte-Carlo Method, Wiley



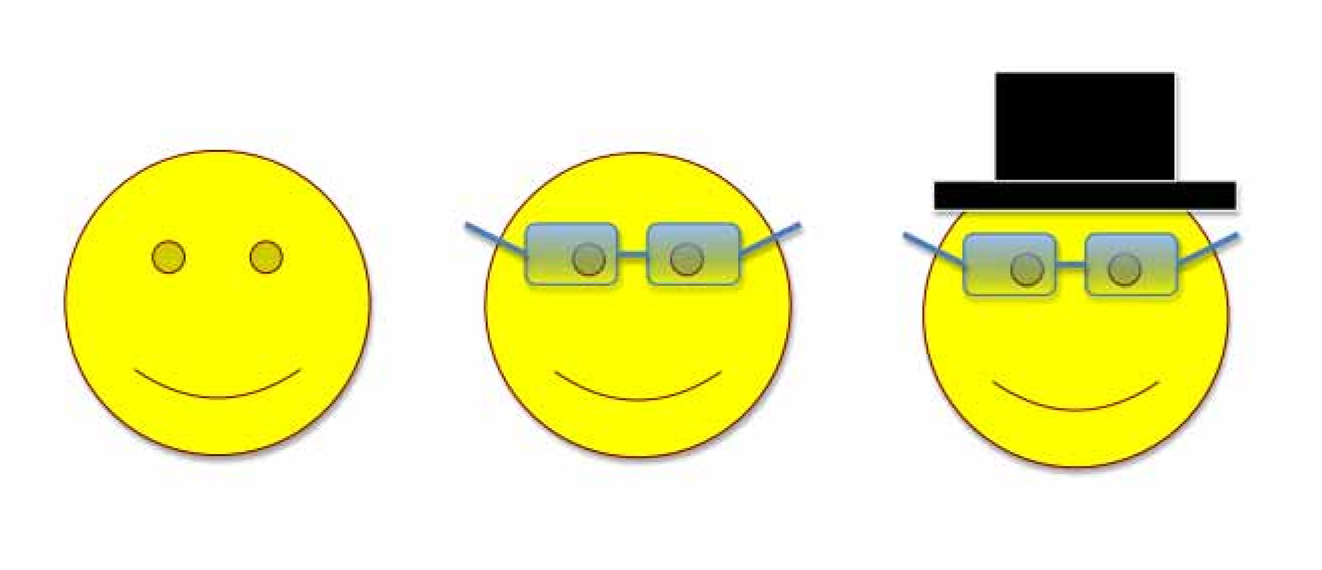




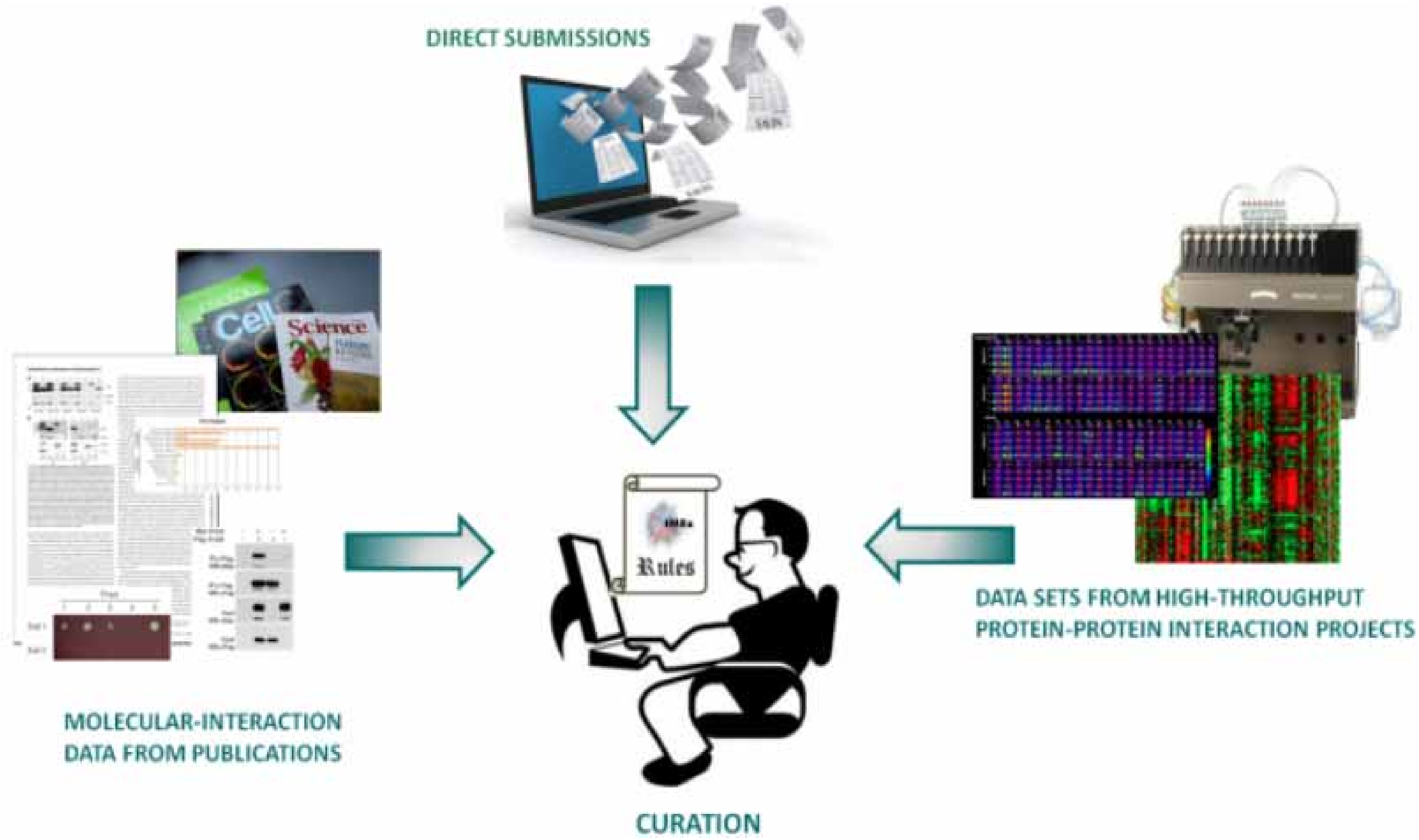
Bishop, C. M. 2007. Pattern Recognition and Machine Learning, Heidelberg, Springer. Chapter 8 on graphical models openly available: http://research.microsoft.com/en-us/um/people/cmbishop/prml/

Murphy, K. P. 2012. Machine learning: a probabilistic perspective, MIT press. Chapter 26 (pp. 907) – Graphical model structure learning

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



Stiller, A., Goodman, N. & Frank, M. C. Ad-hoc scalar implicature in adults and children. CogSci, 2011.



http://www.ebi.ac.uk/intact/