Welcome

185.A83 Machine Learning for Health Informatics
2021S, VU, 2.0 h, 3.0 ECTS

Andreas Holzinger, Rudolf Freund
Marcus Bloice, Florian Endel, Anna Saranti

From Decision Making under Uncertainty to Probabilistic Graphical Models

Contact: andreas.holzinger AT tuwien.ac.at

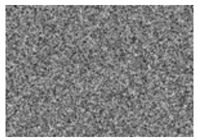
https://human-centered.ai/lv-185-a83-machine-learning-for-health-informatics-class-of-2021

What will we learn today ?

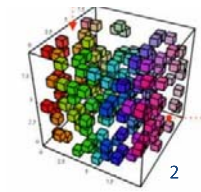
- 00 Reflection from last lecture
- 01 Decision Making under uncertainty
- 02 Some Basics of Graphs/Networks
- 03 Bayesian Networks (BN)
- 04 Markov Chain Monte Carlo (MCMC)
- 05 Metropolis Hastings Algorithm (MH)
- 06 Probabilistic Programming (PP)

00 Reflection

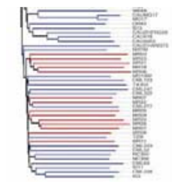
Warm-up Quiz



1

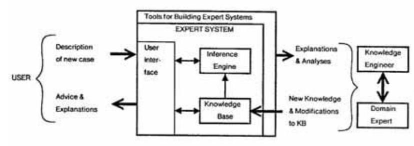


2



3

4



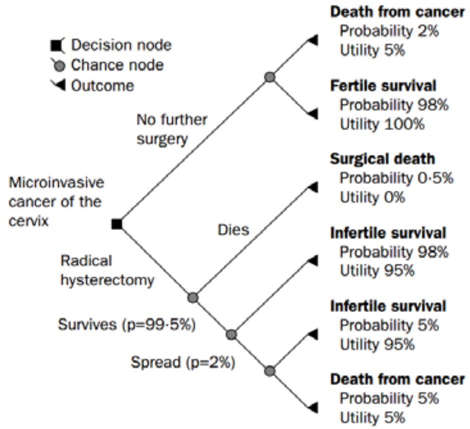
5

On top-level - which machine learning approaches do we know ?

- Symbolic ML
  - First order logic, inverse deduction, knowledge composition
  - Tom Mitchell, Steve Muggleton, Ross Quinlan, …
- Bayesian ML
  - Statistical learning, probabilistic inference
  - Judea Pearl, Michael Jordan, David Heckermann, …
- Cognitive ML
  - Analogisms from Psychology, Kernel machines
  - Vladimir Vapnik, Peter Hart, Douglas Hofstaedter, …
- Connectionist ML
  - Neuroscience, Backpropagation
  - Geoffrey Hinton, Yoshua Bengio, Yann LeCun, …
- Evolutionary ML
  - Nature-inspired concepts, genetic programming
  - John Holland (1929-2015), John Koza, Hod Lipson, …

Pedro Domingos 2015. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, Penguin UK.
https://learning.acm.org/techtalks/machinelearning

Why is this important for us?



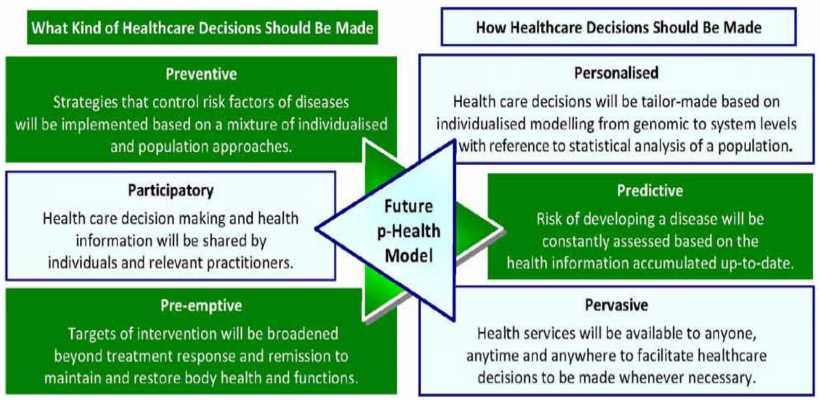
Physician treating a patient approx. 480 B.C.
Beazley (1963), Attic Red-figured Vase-Painters, 813, 96.
Department of Greek, Etruscan and Roman Antiquities, Sully, 1st floor, Campana Gallery, room 43 Louvre, Paris

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

A short historical digression: whom do you see here?

Both Images are in the public domain

Remember: What is "personalized medicine" ?

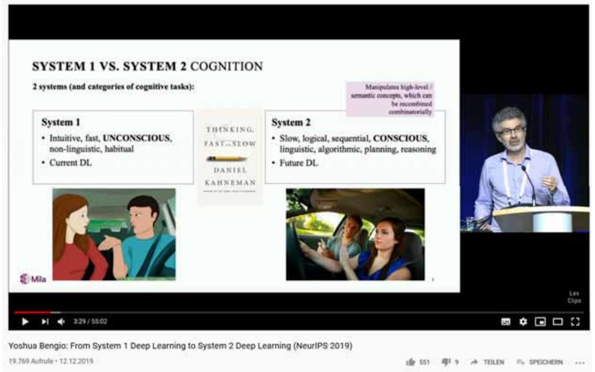


Zhang, Y. T. & Poon, C. C. Y. (2010) Editorial Note on Bio, Medical, and Health Informatics. *Information Technology in Biomedicine, IEEE Transactions on, 14, 3, 543-545.*

01 Decision Making under uncertainty

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

How to go from System 1 deep learning to System 2 deep learning?



https://www.youtube.com/watch?v=T3sxeTgT4qc

Daniel Kahneman 2011. Thinking, fast and slow, New York, Macmillan.

Amos Tversky & Daniel Kahneman 1974. Judgment under uncertainty: Heuristics and biases. Science, 185, (4157), 1124-1131, doi:10.1126/science.185.4157.1124.

What is medical action?

Decision Making: Learn good policy for selecting actions

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

Bench

History Decision Check

Predict t

Bedside

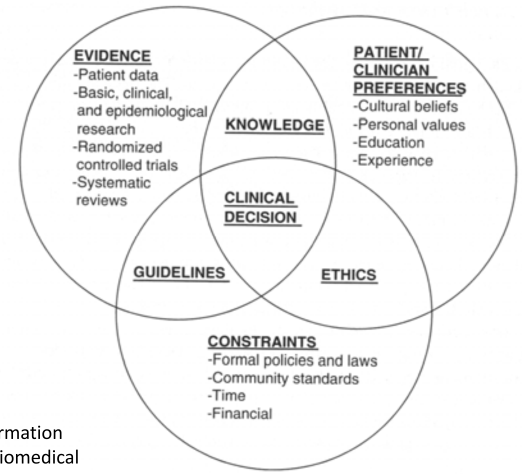
For $t = 1, \dots, T$

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 $r_t(a_t) \in [0,1]$

Why is decision making so difficult ?



William Hersh 2010. Information Retrieval: A Health and Biomedical Perspective, New York, Springer.

How does a medical doctor make a decision ?

3 July 1959, Volume 130, Number 3366 SCIENCE

Reasoning Foundations of Medical Diagnosis

Symbolic logic, probability, and value theory aid our understanding of how physicians reason.

Robert S. Ledley and Lee B. Lusted

The purpose of this article is to analyze the complicated reasoning processes inherent in medical diagnosis. The importance of this problem has received recent emphasis by the increasing interest in the use of electronic computers as an aid to medical diagnostic processes

fitted into a definite disease category, or that it may be one of several possible diseases, or else that its exact nature cannot be determined." This, obviously, is a greatly simplified explanation of the process of diagnosis, for the physician might also comment that after seeing a

ance are the ones who do remember and consider the most possibilities."

Computers are especially suited to help the physician collect and process clinical information and remind him of diagnoses which he may have overlooked. In many cases computers may be as simple as a set of hand-sorted cards, whereas in other cases the use of a large-scale digital electronic computer may be indicated. There are other ways in which computers may serve the physician, and some of these are suggested in this paper. For example, medical students might find the computer an important aid in learning the methods of differential diagnosis. But to use the computer thus we must understand how the physician makes a medical diagnosis. This, then, brings us to the subject of our investigation: the reasoning foundations of medical diagnosis and treatment.

Medical diagnosis involves processes that can be systematically analyzed, as well as those characterized as "intangible." For instance, the reasoning foundations of medical diagnostic procedures

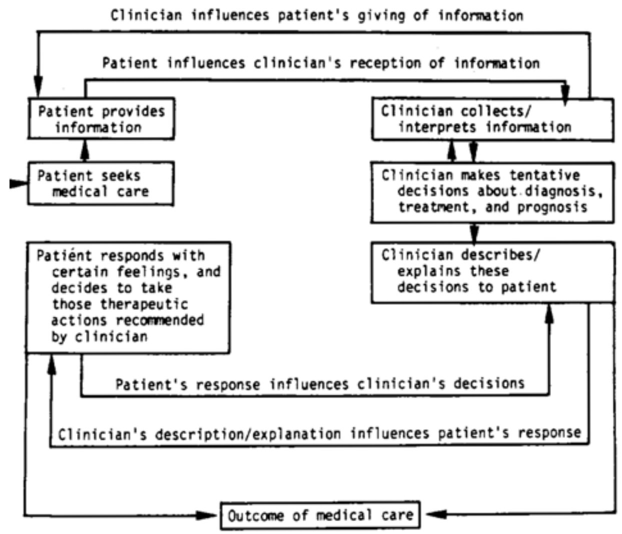
What about the accuracy and uncertainty in medicine ?

- Medical (clinical) data are defined and detected disturbingly "soft" …
- … having an obvious degree of **variability** and **inaccuracy**.
- Taking a medical history, the performance of a physical examination, the interpretation of laboratory tests, even the definition of diseases … are surprisingly **inexact**.
- Data is defined, collected, and interpreted with a degree of variability and inaccuracy which falls far short of the standards **which engineers do expect from most data**.
- Moreover, standards might be **interpreted variably** by different medical doctors, different hospitals, different medical schools, different medical cultures, …

Anthony L. Komaroff 1979. The variability and inaccuracy of medical data. Proceedings of the IEEE, 67, (9), 1196-1207.

Why is the patient-doctor dialogue so important ?

Anthony L. Komaroff 1979. The variability and inaccuracy of medical data. Proceedings of the IEEE, 67, (9), 1196-1207.



d … *data*

h … *hypotheses*

$\mathcal{H}$ … $\{h_1, h_2, \ldots, h_n\}$

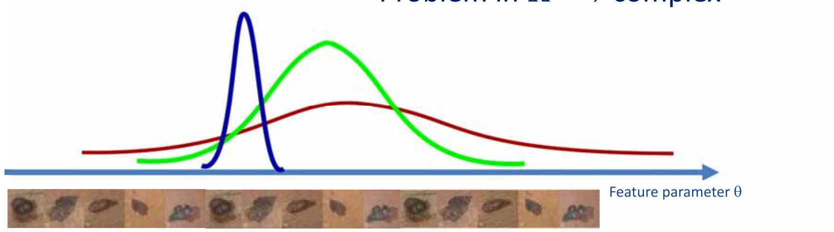
Likelihood

Prior Probability

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

Posterior Probability

Problem in $\mathbb{R}^n \to$ complex



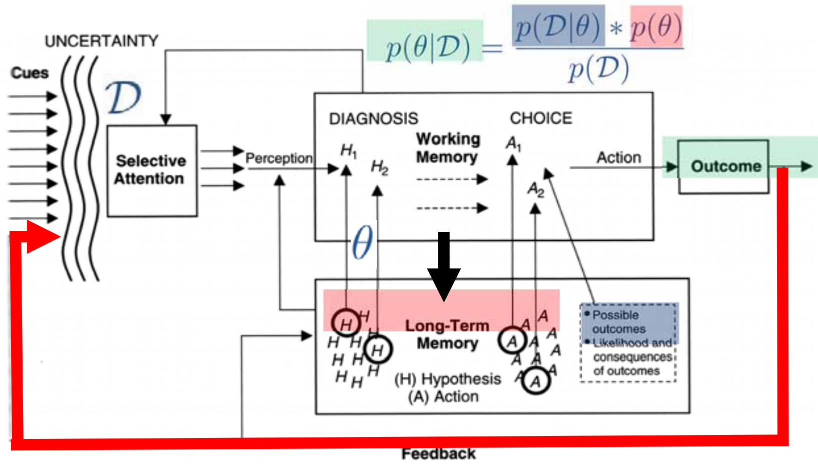
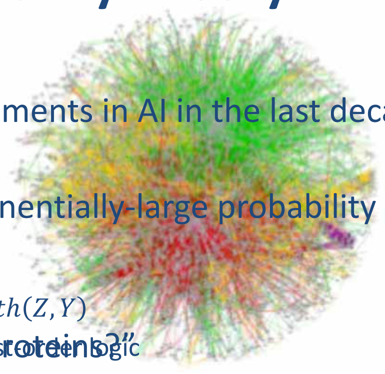


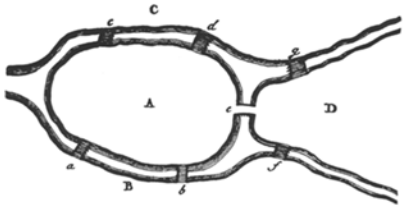
Image by Christopher D. Wickens 1984, modified by Andreas Holzinger 2004

02 Graphs = Networks

- 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
  - $Path\ (X,Y) := edge\ (X,Y)$
  - $Path\ (X,Y) := edge\ (X,Y), path(Z,Y)$
  - This can NOT be expressed in first-order logic
- Example Question: "Is there a path connecting two proteins?"



We start in 1736



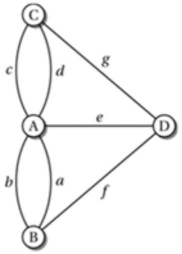
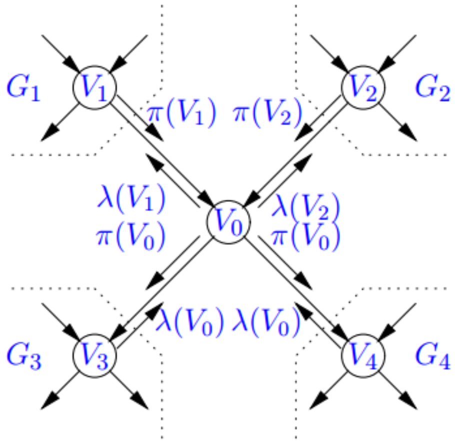


Image from https://people.kth.se/~carlofi/teaching/FEL3250-2013/courseinfo.html

Leonhard Euler 1741. Solutio problematis ad geometriam situs pertinentis. Commentarii academiae scientiarum Petropolitanae, 8, 128-140.

252 years later: Belief propagation algorithm



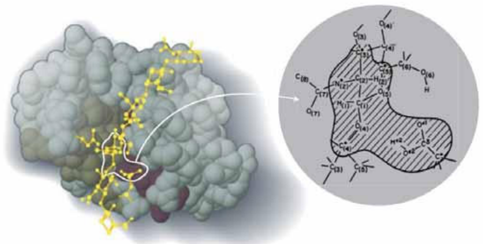
Pearl, J. 1988. Embracing causality in default reasoning. Artificial Intelligence, 35, (2), 259-271.

275 years later … the "Nobel-prize in Computer Science"

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

Nobel Prize in Chemistry 2013

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



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

First Question: Where do graphs come from?

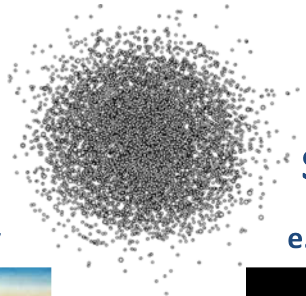
- **Graphs as models for networks**
- given as direct input (point cloud data sets)
- Given as properties of a structure
- Given as a representation of information (e.g. Facebook data, viral marketing, etc., …)

- **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, …)

What do you see here ?

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

Why is time and space important in health informatics ?



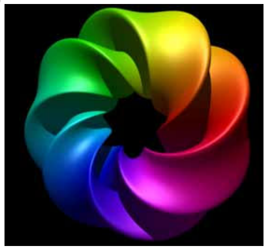
Time

e.g. Entropy

Space

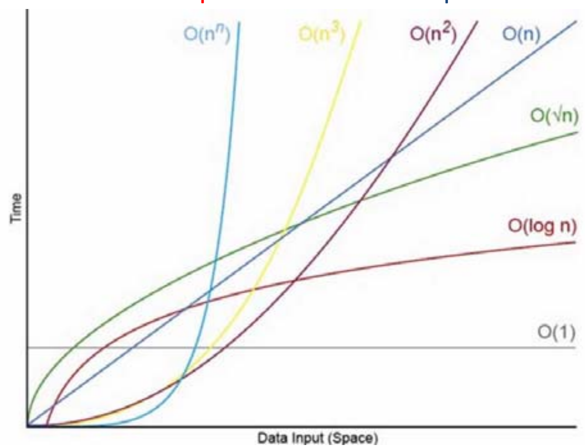
e.g. Topology

Dali, S. (1931) The persistence of memory



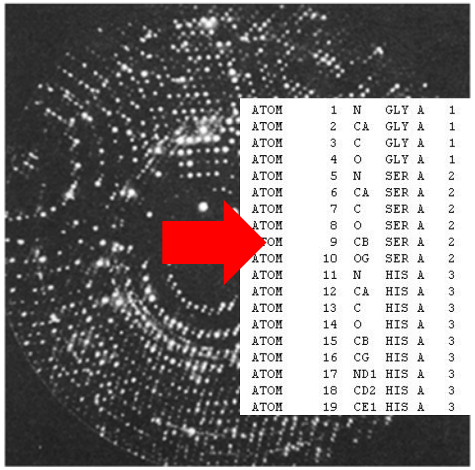
Bagula & Bourke (2012) Klein-Bottle

Why is the Complexity Problem: Time versus Space affecting us ?



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

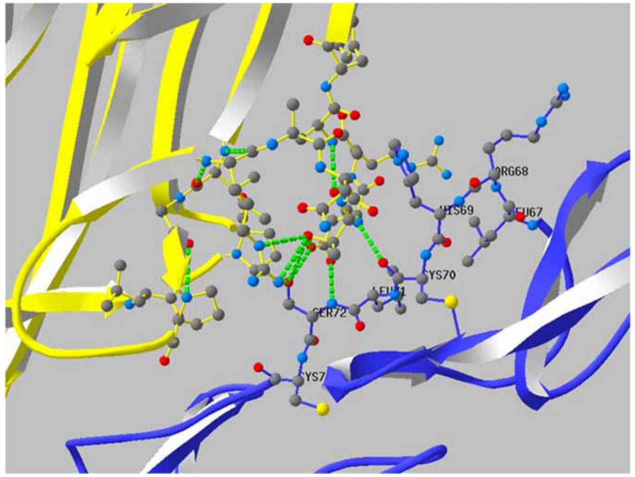
Why are protein structures so important ?



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

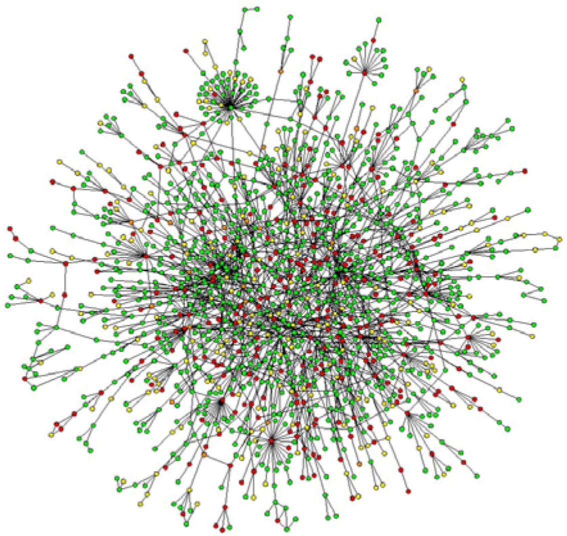
Wiltgen, M. & Holzinger, A. (2005) Visualization in Bioinformatics: Protein Structures with Physicochemical and Biological Annotations. In: *Central European Multimedia and Virtual Reality Conference. Prague, Czech Technical University (CTU), 69-74*

Getting Insight: Knowledge Discovery from Data



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

First yeast protein-protein interaction network

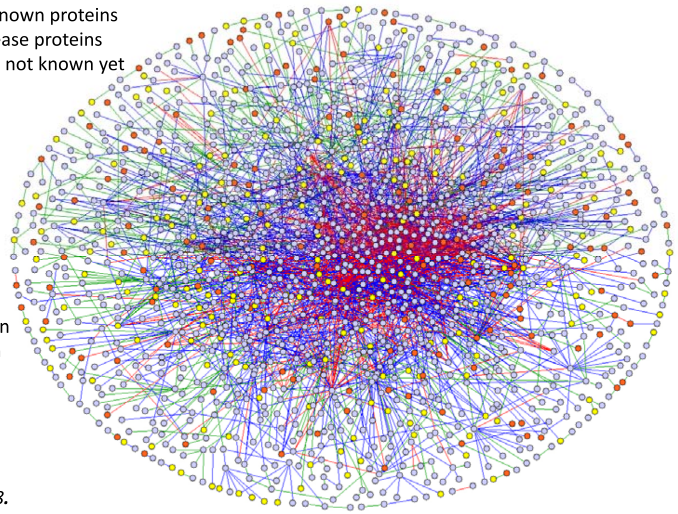


Nodes = proteins
Links = physical interactions (bindings)
Red Nodes = lethal
Green Nodes = non-lethal
Orange = slow growth
Yellow = not known

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

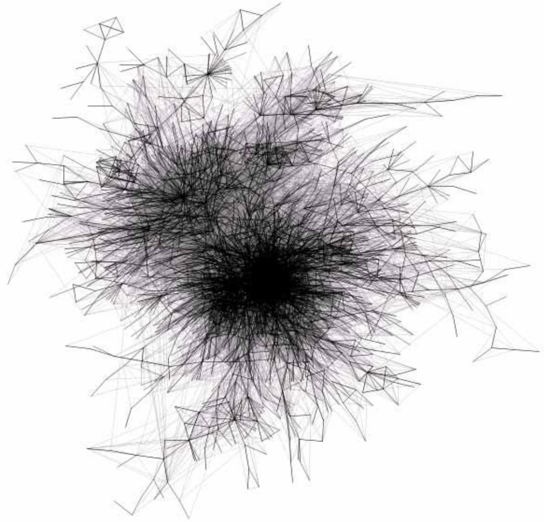
First human protein-protein interaction network

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



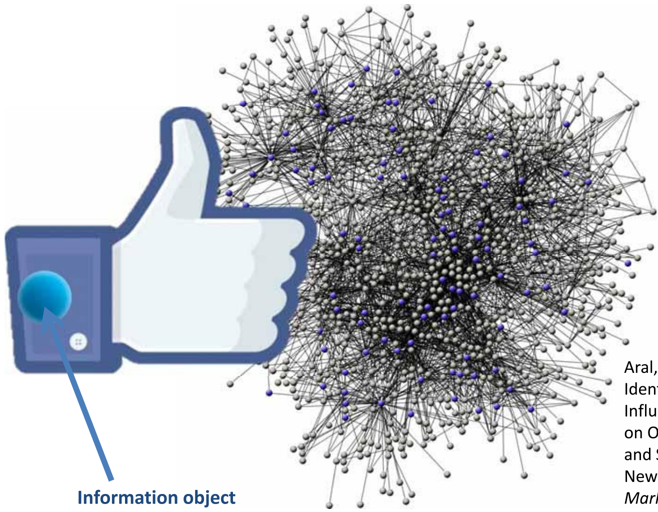
Stelzl, U. et al. (2005) A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome. *Cell, 122, 6, 957-968.*

Non-Natural Network Example: Blogosphere



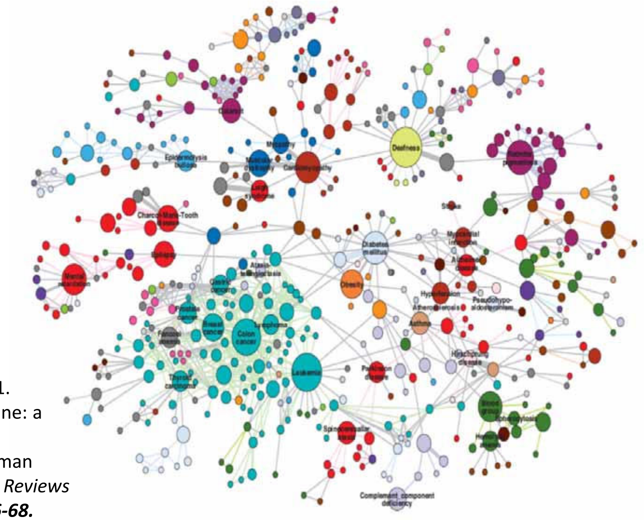
Hurst, M. (2007), Data Mining: Text Mining, Visualization and Social Media. Online available: http://datamining.typepad.com/data_mining/2007/01/the_blogosphere.html, last access: 2011-09-24

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.*

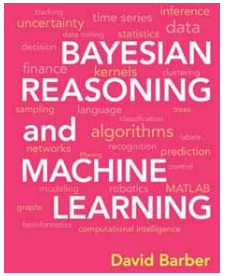
Human Disease Network -> Network Medicine



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

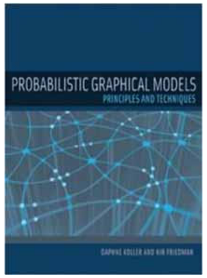
03 Bayesian Networks “Bayes’ Nets”

Book recommendations 1



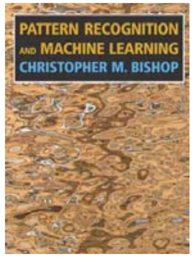
David Barber 2012. Bayesian reasoning and machine learning, Cambridge, Cambridge University Press.

http://www.cs.ucl.ac.uk/staff/d.barber/brml/



Daphne Koller & Nir Friedman 2009. Probabilistic graphical models: principles and techniques, MIT press.

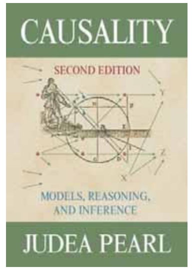
Book recommendations 2



https://goo.gl/6a7rOC

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

Chris Bishop 2006. Pattern Recognition and Machine Learning, Heidelberg, Springer.



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

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

What are the rules of probability ?

$$P(x) = \sum_y P(x,y)$$

$$P(x,y) = P(y|x)P(x)$$

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

$$P(x) = \sum_y P(x|y)P(y)$$

Digression: Markov Processes in Machine Learning

Why are Markov decision processes so important ?

- Markov decision processes (MDP) are …
- random processes in which the future, given the present, is independent of the past!
- one of the most important classes of random processes!

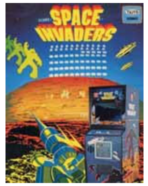
Med Decis Making
Vol. 3, No. 4, 1983

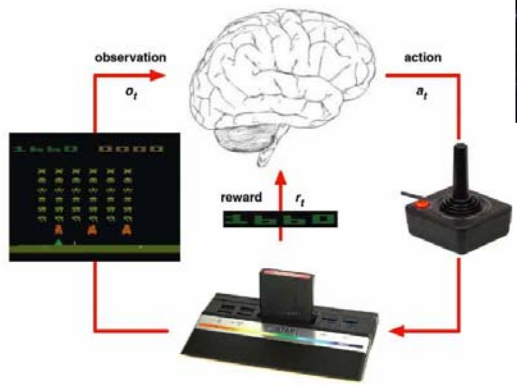
The Markov Process in Medical Prognosis

J. Robert Beck, M.D., and Stephen G. Pauker, M.D.

The physician's estimate of prognosis under alternative treatment plans is a principal factor in therapeutic decision making. Current methods of reporting prognosis, which include five-year survivals, survival curves, and quality-adjusted life expectancy, are crude estimates of natural history. In this paper we describe a general-purpose model of medical prognosis based on the Markov process and show how this simple mathematical tool may be used to generate detailed and accurate assessments of life expectancy and health status. (Med Decis Making 3:419–458, 1983)

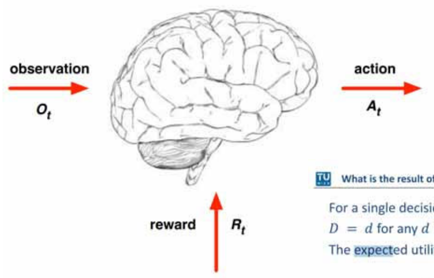
How can MDP be useful for machine learning ?





Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S. & Hassabis, D. 2015. Human-level control through deep reinforcement learning. Nature, 518, (7540), 529-533, doi:10.1038/nature14236

From where do we know such behaviour ?



What is the result of the Expected Utility Theory $E(U|d)$?

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

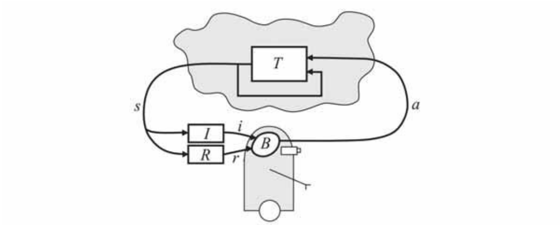
$$E(U \mid d) = \sum_{x_1,\ldots,x_n} P(x_1,\ldots,x_n \mid d) U(x_1,\ldots,x_n,d)$$

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

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

John Von Neumann & Oskar Morgenstern 1944. Theory of games and economic behavior, Princeton university press.

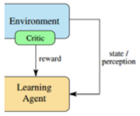
Standard RL-Agent Model goes back to Cybernetics 1950




```
initialize V(s) arbitrarily
loop until policy good enough
    loop for s ∈ S
        loop for a ∈ A
            Q(s,a) := R(s,a) + γ Σ_{s'∈S} T(s,a,s')V(s')
        V(s) := max_a Q(s,a)
    end loop
end loop
```

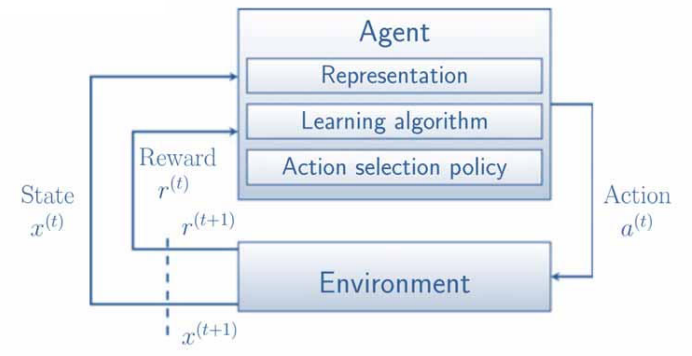
Kaelbling, L. P., Littman, M. L. & Moore, A. W. 1996. Reinforcement learning: A survey. Journal of Artificial Intelligence Research, 4, 237-285.

What has an RL-agent to do with MDP ?




```
for t = 1,...,n do
  The agent perceives state s_t
  The agent performs action a_t
  The environment evolves to s_{t+1}
  The agent receives reward r_t
end for
```

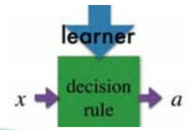
Intelligent behavior arises from the actions of an individual seeking to **maximize its received reward** signals in a **complex and changing world**

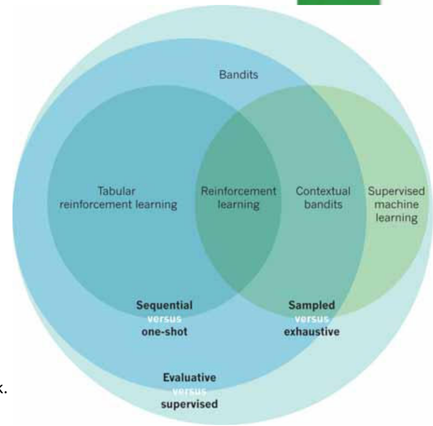


Sutton, R. S. & Barto, A. G. 1998. Reinforcement learning: An introduction, Cambridge MIT press

RL – Types of Feedback (crucial!)

- Supervised: Learner told best a
- Exhaustive: Learner shown every possible x
- One-shot: Current x independent of past a

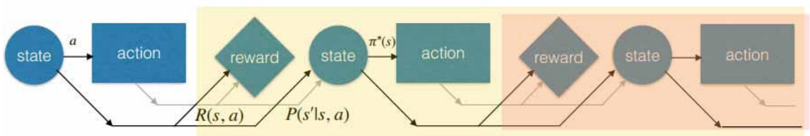




Littman, M. L. 2015. Reinforcement learning improves behaviour from evaluative feedback. Nature, 521, (7553), 445-451.

Problem Formulation in a MDP

- Markov decision processes specify setting and tasks
- Planning methods use knowledge of P and R to compute a good policy π
- Markov decision process model captures both sequential feedback and the more specific one-shot feedback (when $P(s'|s,a)$ is independent of both s and a



$$Q^*(s,a) = R(s,a) + \gamma \Sigma P(s'|s,a) \max_{a'} Q^*(s',a')$$

Littman, M. L. 2015. Reinforcement learning improves behaviour from evaluative feedback. Nature, 521, (7553), 445-451.

Agent observes environmental state at each step t

- 1) Overserves
- 2) Executes
- 3) Receives Reward
- Executes action A_t:
- $O_t = sa_t = se_t$
- Agent state = environment state = information state
- Markov decision process (MDP)

Observation O_t

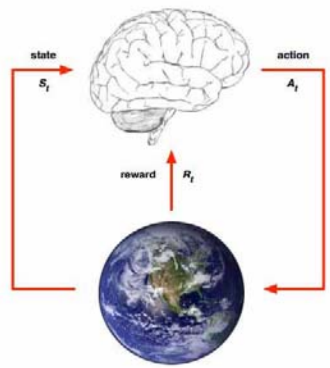
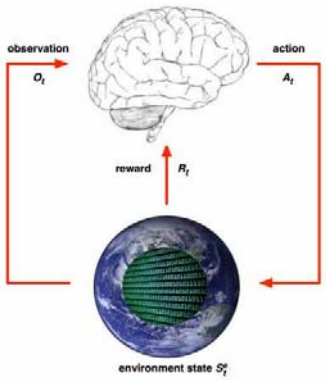


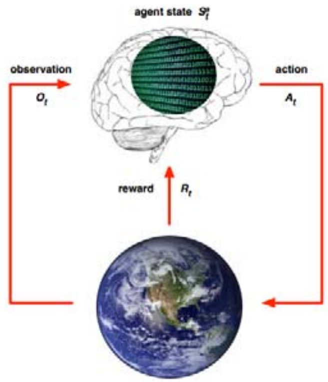
Image credit to David Silver, UCL

- i.e. whatever data the environment uses to pick the next observation/reward
- The environment state is not usually visible to the agent
- Even if S is visible, it may contain irrelevant information
- A State S_t is Markov iff:

$$\mathbb{P}[S_{t+1}|S_t] = \mathbb{P}[S_{t+1}|S_1, \dots, S_t]$$



- i.e. whatever information the agent uses to pick the next action
- it is the information used by reinforcement learning algorithms
- It can be any function of history:
- $S = f(H)$



$$H_t = O_1, R_1, A_1, \dots, A_{t-1}, O_t, R_t$$

- RL agent components:
  - Policy: agent's behaviour function
  - Value function: how good is each state and/or action
  - Model: agent's representation of the environment
    - Policy as the agent's behaviour
      - is a map from state to action, e.g.
      - Deterministic policy: a = (s)
      - Stochastic policy: (ajs) = P[At = ajS t = s
    - Value function is prediction of future reward:

$$v_\pi(s) = \mathbb{E}_\pi \left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \mid S_t = s \right]$$

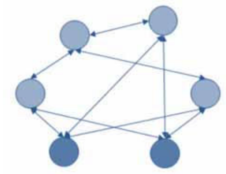
- Partial observability: when agent only indirectly observes environment
- Formally this is a Partially Observable Markov Decision Process (POMDP):
  - Agent must construct its own state representation S,

for example:

- Complete history: $S_t^a = H_t$
- Beliefs of environment state: $S_t^a = (\mathbb{P}[S_t^e = s^1], \dots, \mathbb{P}[S_t^e = s^n])$
- Recurrent neural network: $S_t^a = \sigma(S_{t-1}^a W_s + O_t W_o)$

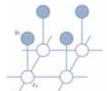
Back to Bayesian Networks

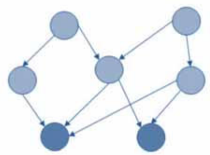
Three types of Probabilistic Graphical Models



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

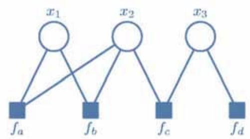
$$P(\boldsymbol{X}) = \frac{1}{Z}\exp\left(\sum_{ij} W_{ij}\, x_i x_j + \sum_i x_i b_i\right)$$





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

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

So, what is a directed Bayesian Network (BN) ?

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

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

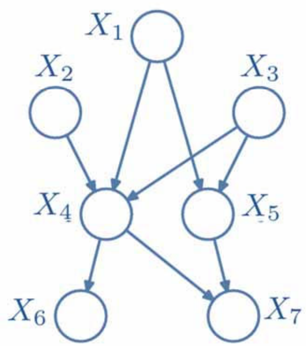
Where $Pa(x_i)$ are the parents of x_i

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.

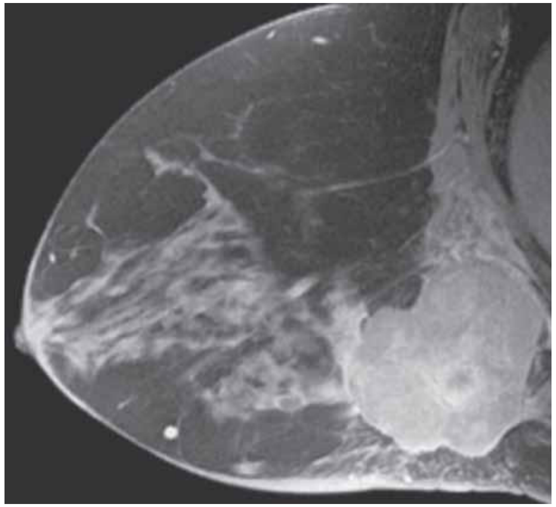
Pearl, J. (1988) *Probabilistic reasoning in intelligent systems: networks of plausible inference. San Francisco, Morgan Kaufmann.*

Example: Directed Bayesian Network with 7 nodes

$$p(X_1, \ldots, X_7) =$$
$$p(X_1)p(X_2)p(X_3)p(X_4|X_1, X_2, X_3)\cdot$$
$$p(X_5|X_1, X_3)p(X_6|X_4)p(X_7|X_4, X_5)$$



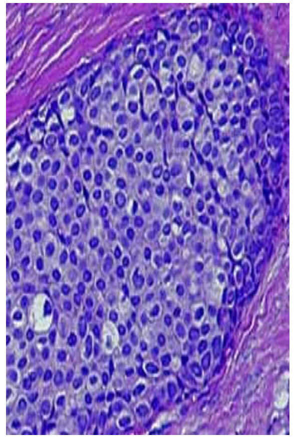
Clinical Case Example



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.*

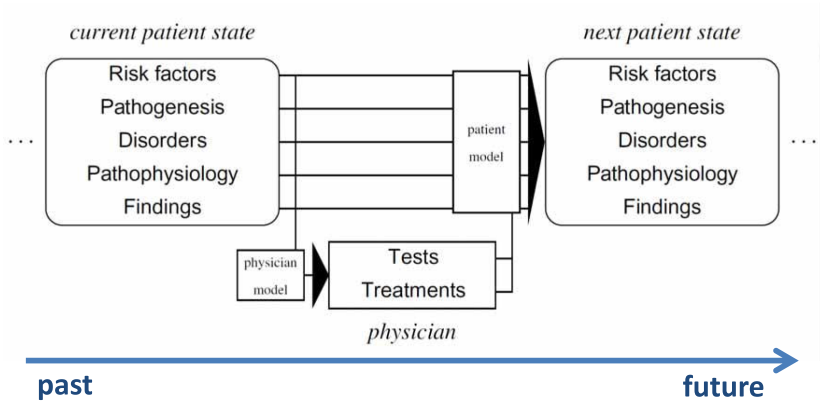
What is important in clinical decision making ?

- = 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.*

How can one predict the future on past data and present status ?



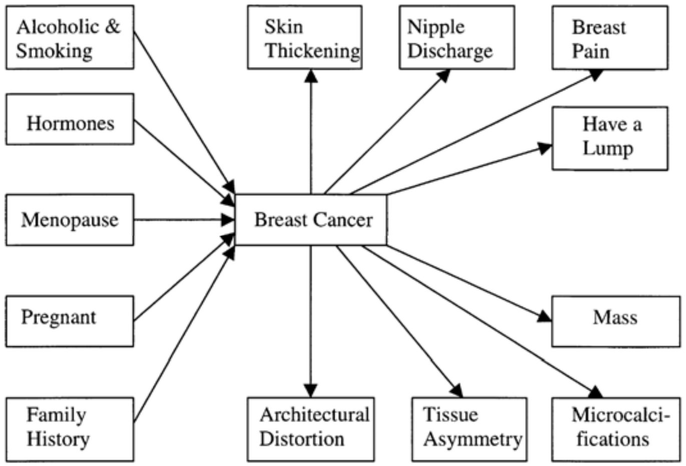
van Gerven, M. A. J., Taal, B. G. & Lucas, P. J. F. (2008) Dynamic Bayesian networks as prognostic models for clinical patient management. *Journal of Biomedical Informatics, 41, 4, 515-529.*

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.*

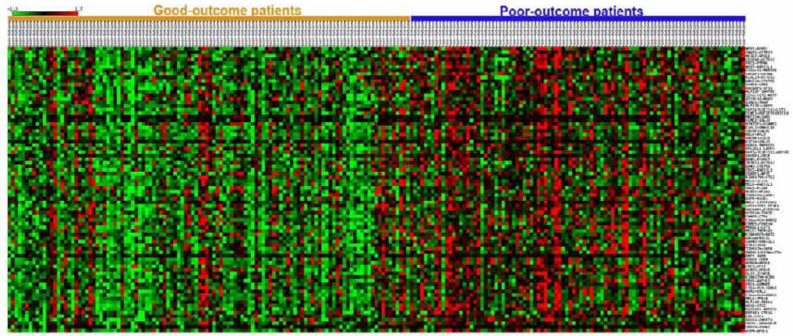
Breast cancer – big picture – state of 1999



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

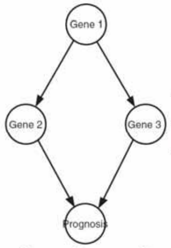


Xu, L., Tan, A., Winslow, R. & Geman, D. (2008) Merging microarray data from separate breast cancer studies provides a robust prognostic test. *BMC Bioinformatics, 9, 1, 125-139.*

Example: 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



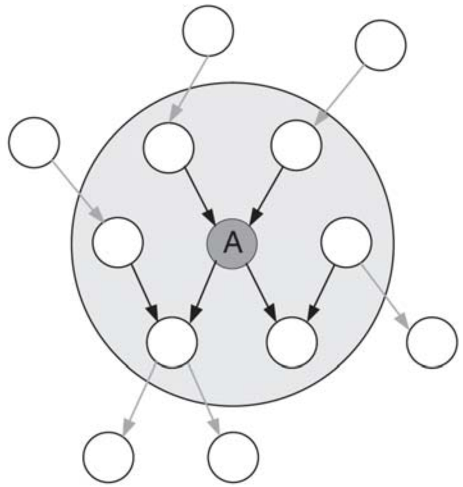
Gene 2	Gene 1 on	Gene 1 off
P(on)	0.3	0.6
P(off)	0.7	0.4

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.*

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:

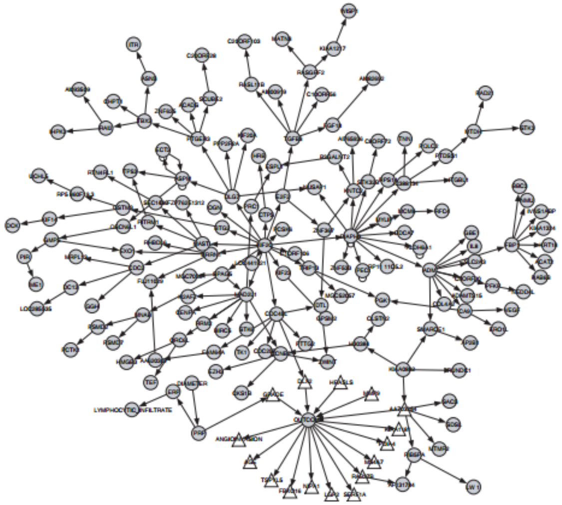
$$\mathrm{p}(\theta_{ij}|S) = Dir(\theta_{ij}|N'_{ij1}, \dots, N'_{ijk}, \dots, N'_{ijr_i})$$

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:

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

with N_{ijk} defined as before.

Predicting the prognosis of breast cancer (integrated a.)



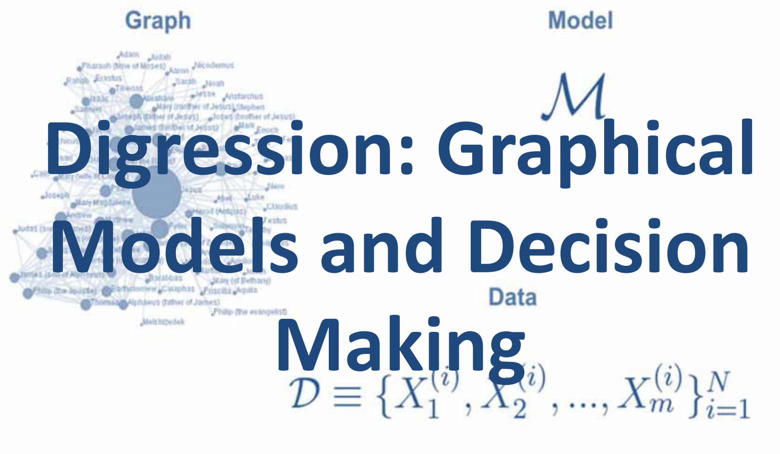
Gevaert, O., Smet, F. D., Timmerman, D., Moreau, Y. & Moor, B. D. (2006) Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics, 22, 14, 184-190.*

Inference in Bayes Nets is intractable (NP-complete!)

- 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 …

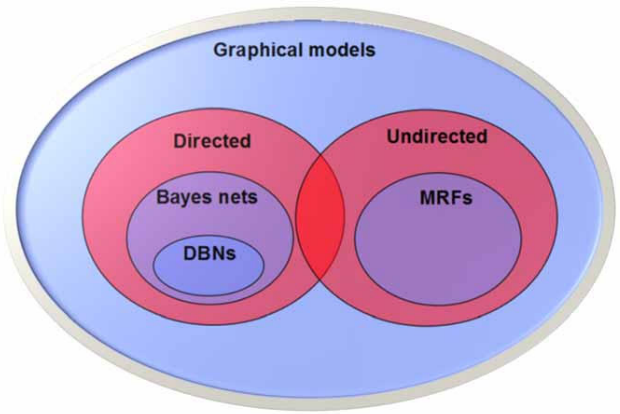
A practical advise

Often it is better to have a good solution within time – than an perfect solution too late …



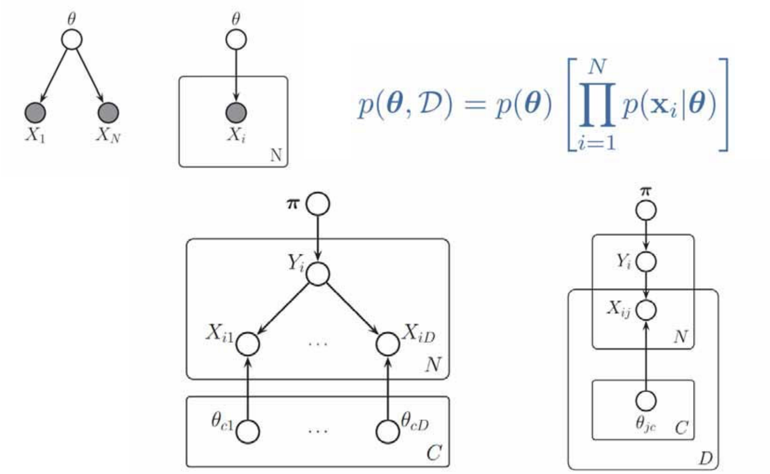
Digression: Graphical Models and Decision Making

What Classes of Graphical Models do we know ?



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

Naïve Bayes classifier as DGM (single/nested plates)



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

Regulatory>Metabolic>Signaling>Protein>Co-expression

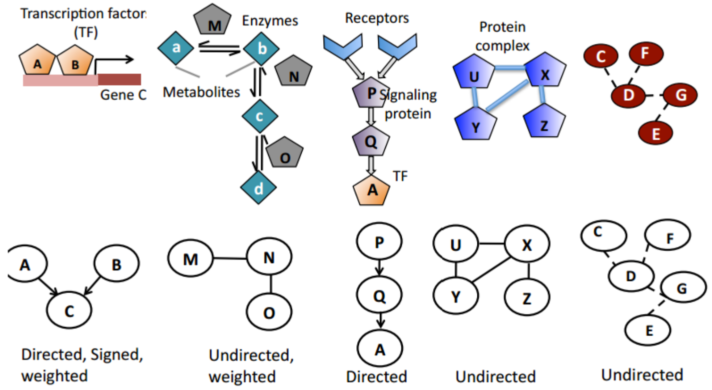
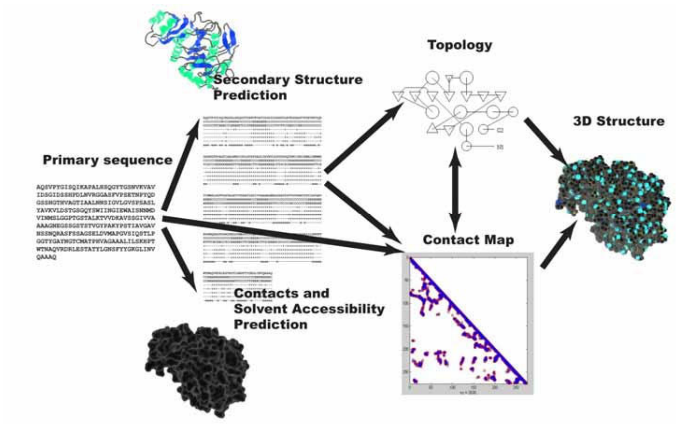


Image credit to Anna Goldenberg, Toronto

Remember

- 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

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.

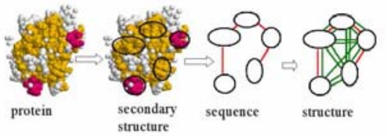
Protein Network Inference

- 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.

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.

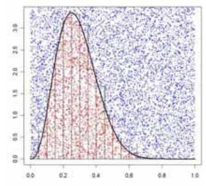


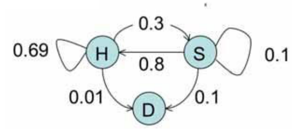
- Important for health informatics: Discovering relationships between biological components
- Unsolved problem in computer science:
- Can the graph isomorphism problem be solved in polynomial time?
  - So far, no polynomial time algorithm is known.
  - It is also not known if it is NP-complete
  - We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

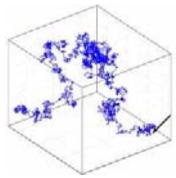
Finally a practical example

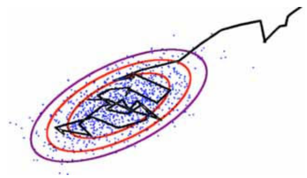
04 Markov Chain Monte Carlo (MCMC)

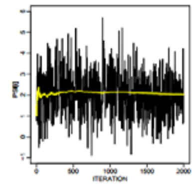
Monte Carlo Method (MC)
Monte Carlo Sampling
Markov Chains (MC)
MCMC
Metropolis-Hastings











What is the problem with observable data $\mathcal{D}$ in the real-world ?

- 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:

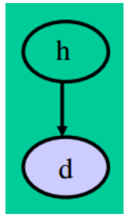
$$J = \int f(\theta) * p(\theta|\mathcal{D}) d\theta$$

What is the problem of learning and inference ?

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

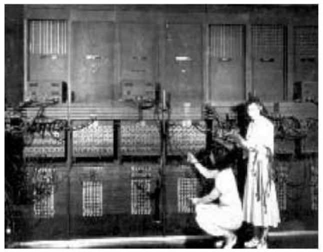
$$\mathcal{H} = \{H_1, H_2, ..., H_n\} \quad \forall(h,d)$$

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



What was the origin of MCMC ?



Summary: What are Monte Carlo methods?

- 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!

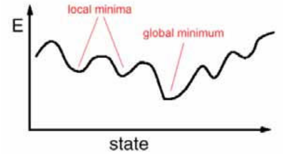
MC connects Computer Science with Cognitive Science

- 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 $P(x)$

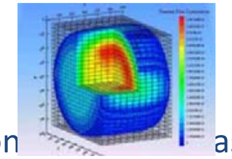
$$\mathbb{E}_{p(x)}(f(x)) = \sum_{X} f(x)P(x)dx$$

Mathematical simulation via MC

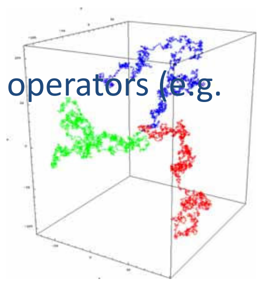
- Solving intractable integrals
- Bayesian statistics: **normalizing** constants, expectations, marginalization
- Stochastic Optimization
- Generalization of simulated annealing
- Monte Carlo expectation maximization (EM)



Physical simulation via MC



- Physical simulation
- estimating neutron diffusion time
- Computing expected utilities and best respon ash 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



Notations

- 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 $\mathbb{E}_x[f(x)|z]$
- 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: $$\mathbb{E}_{p(x)}(f(x)) = \int_X f(x)p(x)dx$$

05 Metropolis-Hastings Algorithm

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

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.

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}.

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

† 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.

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 97

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.

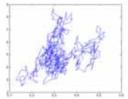
(ii) Use importance sampling, which may also be used for variance reduction. That is, in order to evaluate the integral

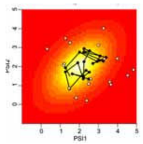
$$J = \int f(x)p(x)dx = E_p(f),$$

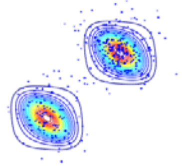
where $p(x)$ is a probability density function, instead of obtaining independent samples $x_1, \ldots, x_N$ from $p(x)$ and using the estimate $\hat{J}_1 = \Sigma f(x_i)/N$, we instead obtain the sample from

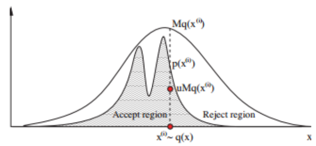
Barber, D. 2012. Bayesian reasoning and machine learning, Cambridge, Cambridge University

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 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:

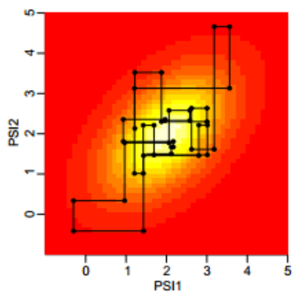
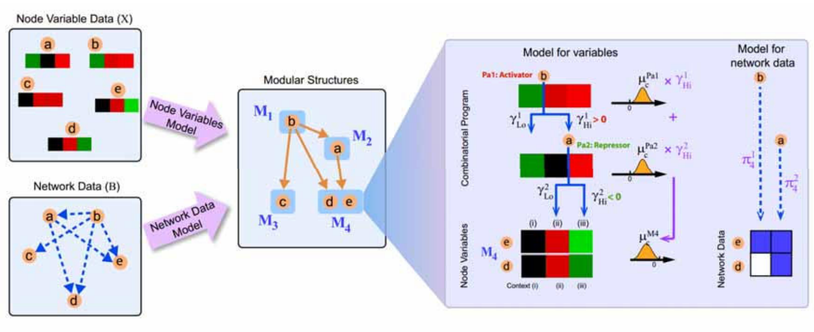


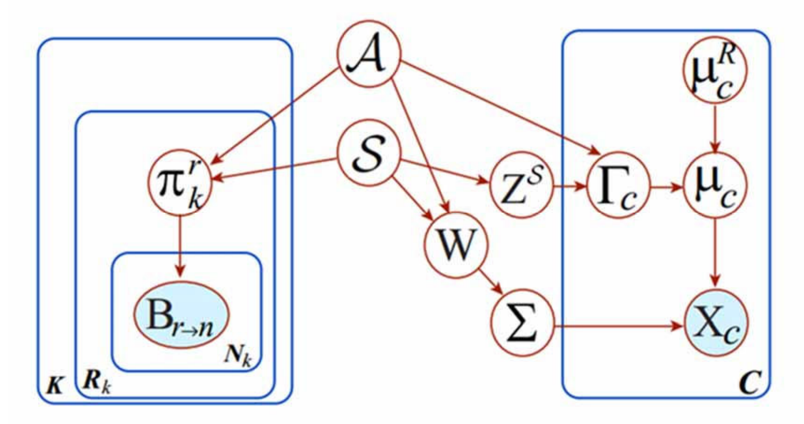
Image Source: Peter Mueller, Anderson Cancer Center

How to learn modular structures from Network Data ?



Elham Azizi, Edoardo M. Airoldi & James E. Galagan. Learning Modular Structures from Network Data and Node Variables. In: Xing, Eric P. & Jebara, Tony, eds. Proceedings of the 31st International Conference on Machine Learning (ICML), 2014 Beijing. JMLR, 1440-1448.

Graphical Model

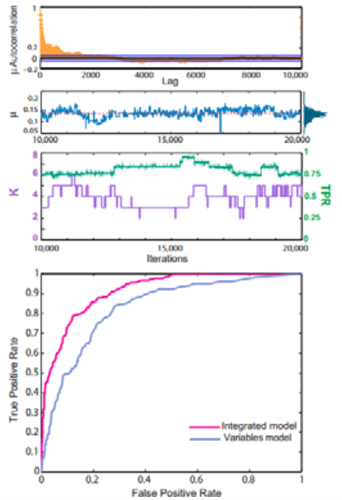


Elham Azizi, Edoardo M. Airoldi & James E. Galagan. Learning Modular Structures from Network Data and Node Variables. In: Xing, Eric P. & Jebara, Tony, eds. Proceedings of the 31st International Conference on Machine Learning (ICML), 2014 Beijing. JMLR, 1440-1448.

How can you estimate the parameters using RJMCMC ?

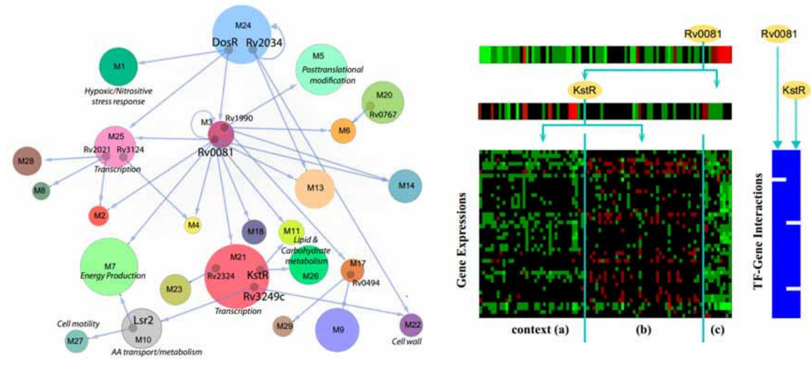
Algorithm 1 RJMCMC for sampling parameters

Inputs:
Node Variables Data **X**
Network Data **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 $\boldsymbol{\gamma}_c^{\mathbf{R}(j+1)} \sim \mathcal{N}(\boldsymbol{\gamma}_c^{\mathbf{R}(j)}, I)$; accept with P_{mh}
end for
end for



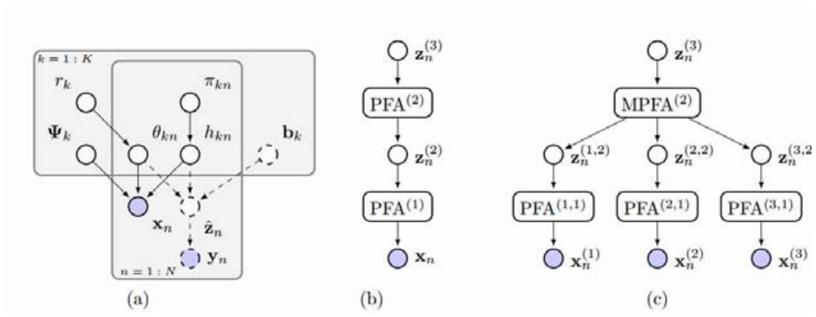
Azizi, E., Airoldi, E. M. & Galagan, J. E. 2014. Learning Modular Structures from Network Data and Node Variables. Proceedings of the 31st International Conference on Machine Learning (ICML). Beijing: JMLR. 1440-1448.

Myobacterium tuberculosis Gene Regulatory Network



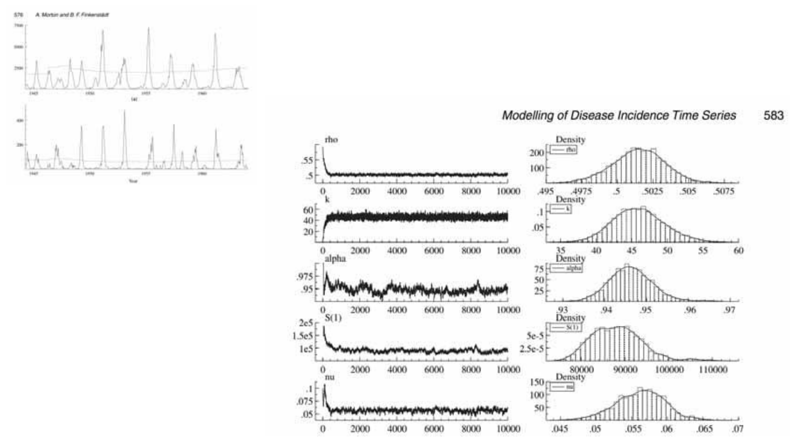
Azizi, E., Airoldi, E. M. & Galagan, J. E. 2014. Learning Modular Structures from Network Data and Node Variables. Proceedings of the 31st International Conference on Machine Learning (ICML). Beijing: JMLR. 1440-1448.

An alternative approach



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.

Discrete time modelling of disease incidence time series by MCMC

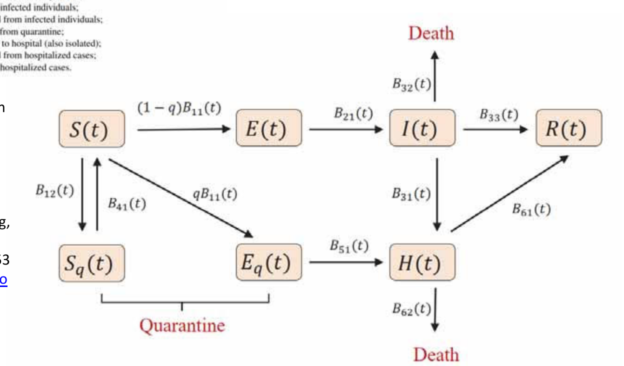


Alexander Morton & Bärbel F. Finkenstädt 2005. Discrete time modelling of disease incidence time series by using Markov chain Monte Carlo methods. Journal of the Royal Statistical Society: Series C (Applied Statistics), 54, (3), 575-594, doi:10.1111/j.1467-9876.2005.05366.x

Discrete-time stochastic epidemic model of COVID-19

- $B_{11}(t)$ is the number of susceptible individuals who become newly infected;
- $B_{12}(t)$ is the number of quarantined susceptible individuals who have contact with infected individuals but are not infected;
- $B_{21}(t)$ is the number of new cases with symptom onset;
- $B_{31}(t)$ is the number of new confirmed and admitted patients;
- $B_{32}(t)$ is the number of new death from infected individuals;
- $B_{33}(t)$ is the number of newly recovered from infected individuals;
- $B_{41}(t)$ is the number of people released from quarantine;
- $B_{51}(t)$ is the number of people admitted to hospital (also isolated);
- $B_{61}(t)$ is the number of newly recovered from hospitalized cases;
- $B_{62}(t)$ is the number of new death from hospitalized cases.

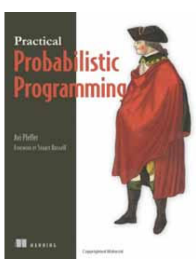
Sha He, Sanyi Tang & Libin Rong 2020. A discrete stochastic model of the COVID-19 outbreak: Forecast and control. Journal of Mathematical Biosciences & Engineering, 17, (4), 2792-2804, doi:10.3934/mbe.2020153 https://www.aimspress.com/MBE/2020/4/2792 (Online open available)



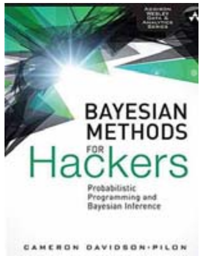
$$L(B_{11}(t), B_{12}(t), B_{21}(t), B_{31}(t), B_{32}(t), B_{33}(t), B_{41}(t), B_{51}(t), B_{61}(t), B_{62}(t)|\Theta) = \prod_{t=0}^{T^n} g_{i,j}(B_{ij}(t)|\cdot)$$

06 Probabilistic Programming

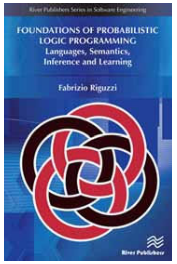
Book recommendations



Avi Pfeffer 2016. Practical probabilistic programming, Shelter Island (NY), Manning.



Cameron Davidson-Pilon 2015. Bayesian methods for hackers: probabilistic programming and Bayesian inference, Addison-Wesley Professional.



Fabrizio Riguzzi 2018. Foundations of Probabilistic Logic Programming, River Publishers.

Arnaud N. Fadja & Fabrizio Riguzzi 2017. Probabilistic Logic Programming in Action. In: Holzinger, Andreas, Goebel, Randy, Ferri, Massimo & Palade, Vasile (eds.) Towards Integrative Machine Learning and Knowledge Extraction: BIRS Workshop, Banff, AB, Canada, July 24-26, 2015, Revised Selected Papers. Cham: Springer, pp. 89-116, doi:10.1007/978-3-319-69775-8_5.

TU WIEN

So, what is probabilistic programming ?

HCAI HUMAN-CENTERED.AI

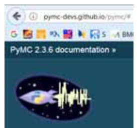
- Probabilistic thinking is a valuable tool for decision making
- Overcoming uncertainties is the huge success currently in machine learning (and for AI ;-)
- Probabilistic reasoning is a versatile tool
- PPLs are domain specific languages that use probabilistic models and the methods to make inferences in those models
- The "magic" is in combining "probability methods" with "representational power"

TU WIEN

Probabilistic-programming.org

HCAI HUMAN-CENTERED.AI

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



PyMC3

PyMC Pythonic Markov chain Monte Carlo

TU WIEN

Medical Example

HCAI HUMAN-CENTERED.AI

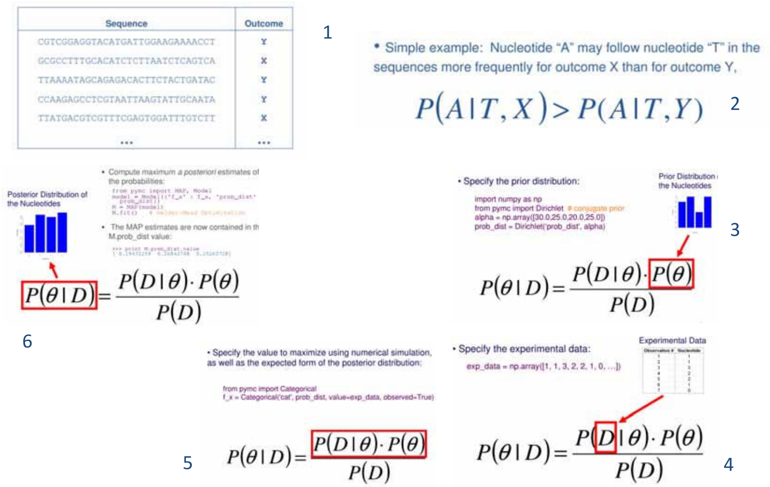
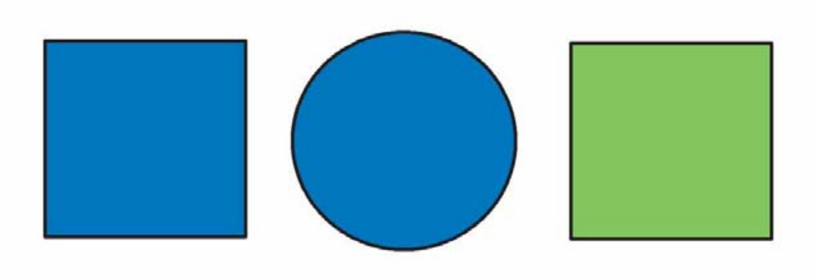


Image Source: Dan Williams, Life Technologies, Austin TX

Digression on Concept Learning

You are talking to you colleague and want to refer to the middle object – which wording would you prefer: circle or blue?

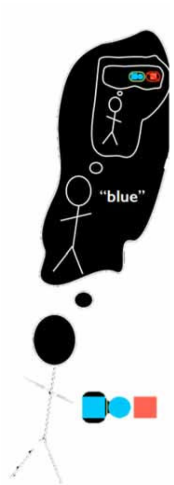


Michael C. Frank & Noah D. Goodman 2012. Predicting pragmatic reasoning in language games. *Science,* 336, (6084), 998-998, doi:10.1126/science.1218633.


```
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
  })}
```

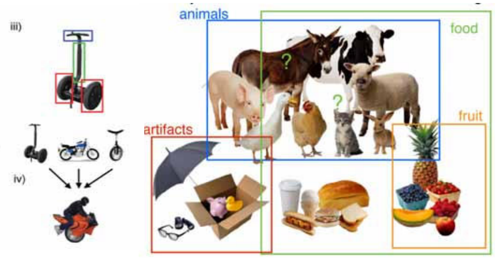


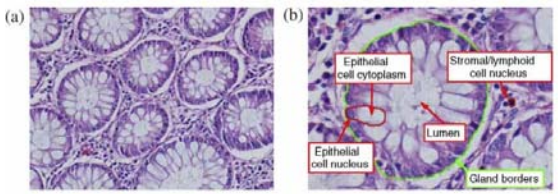
Noah D. Goodman & Michael C. Frank 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences,* 20, (11), 818-829, doi:10.1016/j.tics.2016.08.005.

Why do we need concepts ?

Concepts can be defined as a category membership

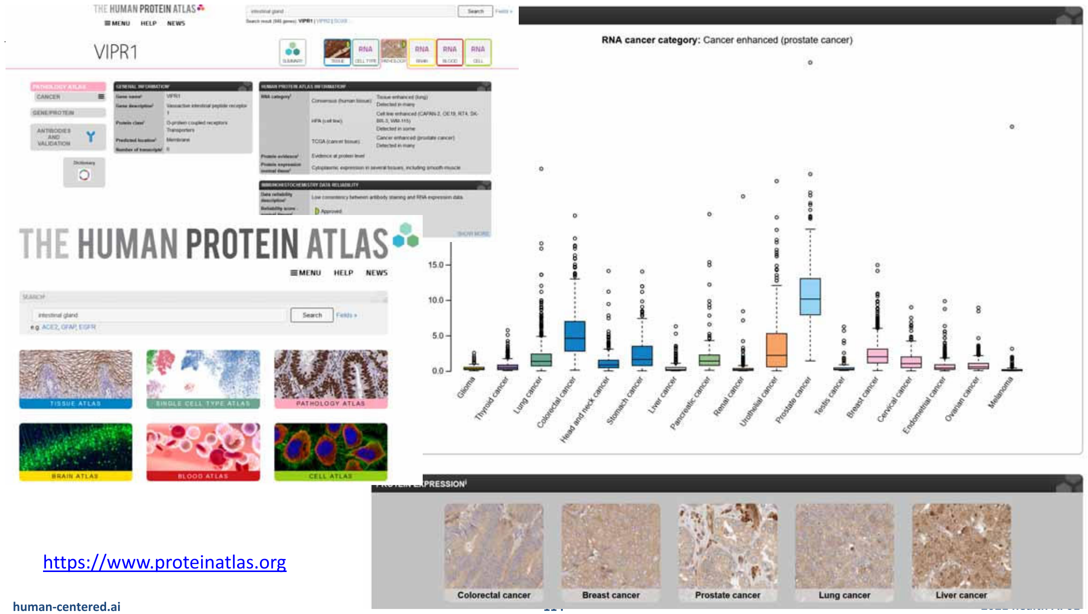
- can be relational and abstract
- category = set of objects that have commonalities
- concept = mental representation of categories
- concepts can be defined, e.g. triangle = a polygon with three sides, a gland = group of cells





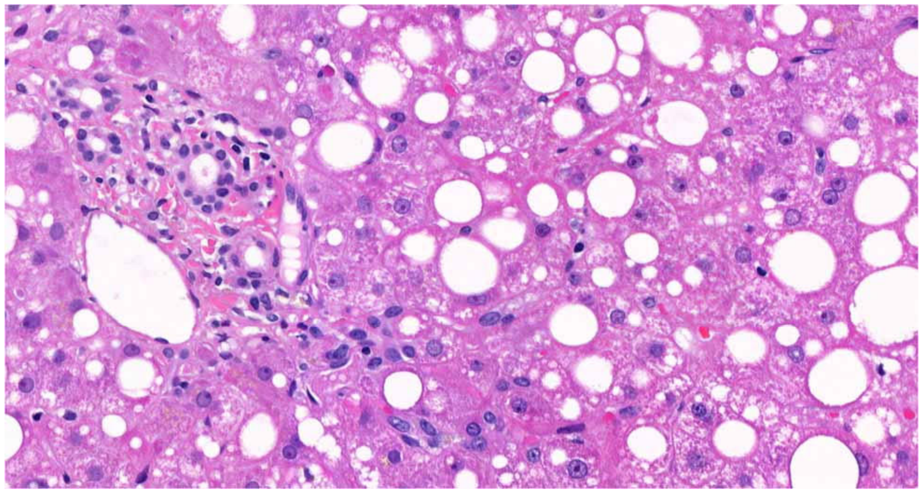
Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum & Samuel J. Gershman (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, (e253), doi:10.1017/S0140525X16001837.

Cigdem Gunduz-Demir, Melih Kandemir, Akif Burak Tosun & Cenk Sokmensuer (2010). Automatic segmentation of colon glands using object-graphs. *Medical image analysis,* 14, (1), 1-12, doi:10.1016/j.media.2009.09.001.



https://www.proteinatlas.org

Example: How do human pathologists make diagnoses ?

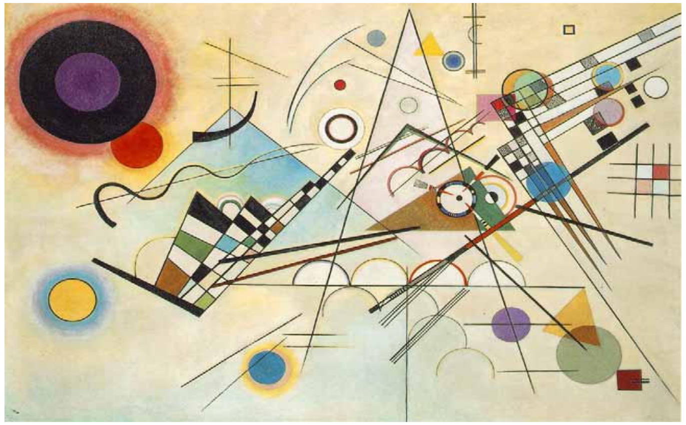


Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal & Heimo Müller (2019). Causability and Explainability of Artificial Intelligence in Medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9, (4), 1-13, doi:10.1002/widm.1312.

What is ground truth ? Where is the ground truth ?

- := information provided by direct observation (empirical evidence) in contrast to information provided by inference
  - *Empirical evidence* = information acquired by observation or by experimentation in order to verify the truth (fit to reality) or falsify (non-fit to reality).
  - *Empirical inference* = drawing conclusions from empirical data (observations, measurements)
  - *Causal inference* = drawing conclusions about a causal connection based on the conditions of the occurrence of an effect
  - *Causal machine learning* is key to ethical AI in health to model explainability for bias avoidance and algorithmic fairness for decision making

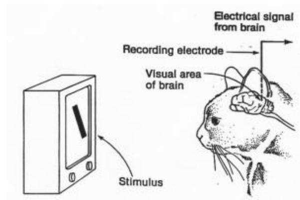
Mattia Prosperi, Yi Guo, Matt Sperrin, James S. Koopman, Jae S. Min, Xing He, Shannan Rich, Mo Wang, Iain E. Buchan, Jiang Bian (2020). Causal inference and counterfactual prediction in machine learning for actionable healthcare. Nature Mach.Intelligence, 2, (7), 369-375, doi:10.1038/s42256-020-0197-y

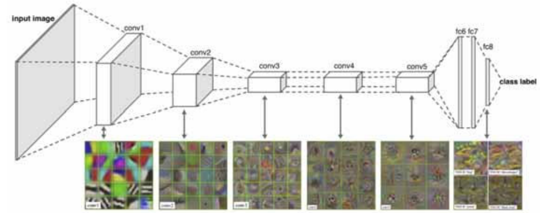


Note: Image is in the public domain and is used according UrhG §42 lit. f Abs 1 as "Belegfunktion" for discussion with students

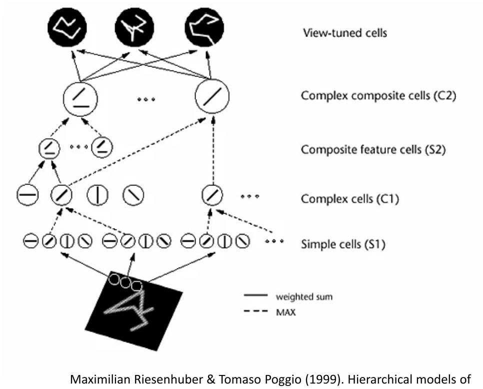
Komposition VIII, 1923, Solomon R. Guggenheim Museum, New York. Source: https://de.wikipedia.org/wiki/Wassily_Kandinsky

David H. Hubel & Torsten N. Wiesel 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of Physiology, 160, (1), 106-154, doi:10.1113/jphysiol.1962.sp006837





David G.T. Barrett, Ari S. Morcos & Jakob H. Macke (2019). Analyzing biological and artificial neural networks: challenges with opportunities for synergy? *Current opinion in neurobiology,* 55, 55-64.



Maximilian Riesenhuber & Tomaso Poggio (1999). Hierarchical models of object recognition in cortex. Nature Neuroscience, 2, (11), 1019-1025, doi:10.1038/14819.

- Bruner, Goodnow, and Austin (1956) published "A Study of Thinking", which became a landmark in cognitive science and has much influence on machine learning.
  - Rule-Based Categories
  - A concept specifies conditions for membership

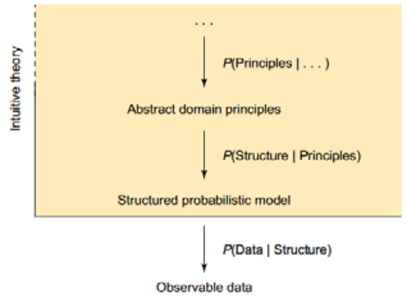
Jerome S. Bruner, Jacqueline J. Goodnow & George A. Austin 1986. A Study of Thinking, Transaction Books.

- which is highly relevant for ML research, concerns the factors that determine the subjective difficulty of concepts:
- Why are some concepts psychologically extremely simple and easy to learn,
- while others seem to be extremely difficult, complex, or even incoherent?
- These questions have been studied since the 1960s but are still unanswered …

Feldman, J. 2000. Minimization of Boolean complexity in human concept learning. Nature, 407, (6804), 630-633, doi:10.1038/35036586.

How can we model basic cognitive capacities as intuitive Bayes ?

- Similarity
- Representativeness and evidential support
- Causal judgement
- Coincidences and causal discovery
- Diagnostic inference
- Predicting the future



$$P(h|x,T) = \frac{P(x|h,T)P(h|T)}{\sum_{h' \in H_T} P(x|h',T)P(h'|T)}$$

Joshua B. Tenenbaum, Thomas L. Griffiths & Charles Kemp 2006. Theory-based Bayesian models of inductive learning and reasoning. Trends in cognitive sciences, 10, (7), 309-318, doi:10.1016/j.tics.2006.05.009.

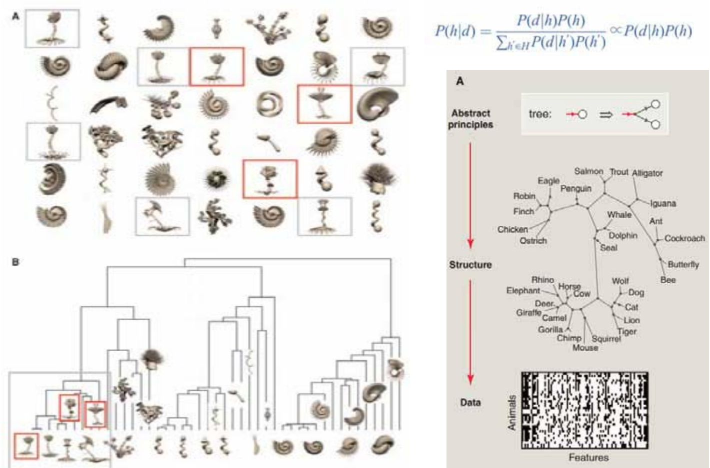
How does our mind get so much out of it ?

Salakhutdinov, R., Tenenbaum, J. & Torralba, A. 2012. One-shot learning with a hierarchical nonparametric Bayesian model. Journal of Machine Learning Research, 27, 195-207.

How can we learn words for objects – concepts from examples

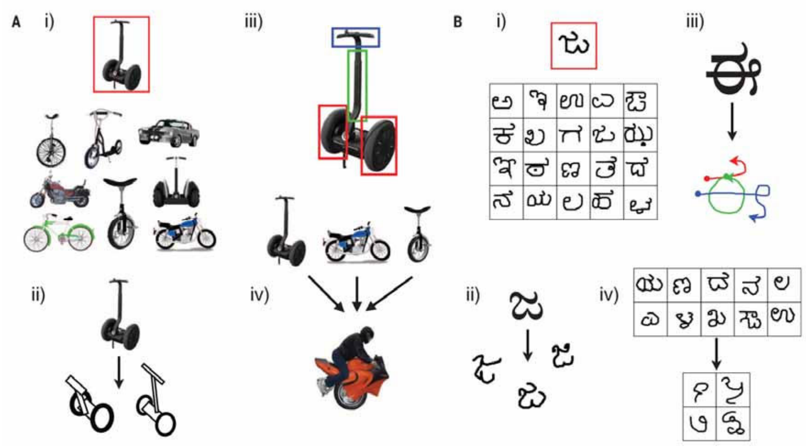
Salakhutdinov, R., Tenenbaum, J. & Torralba, A. 2012. One-shot learning with a hierarchical nonparametric Bayesian model. Journal of Machine Learning Research, 27, 195-207.

How do we understand our world ?



Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. 2011. How to grow a mind: Statistics, structure, and abstraction. Science, 331, (6022), 1279-1285.

What is probabilistic program induction ?

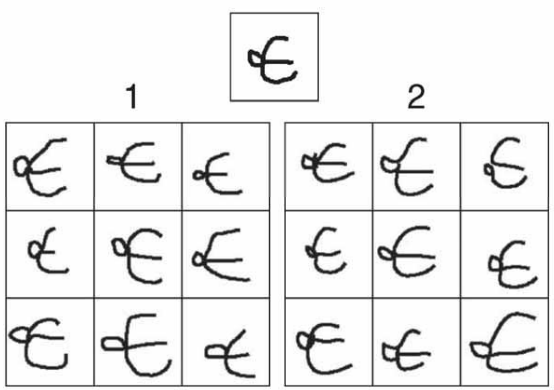


Brenden M. Lake, Ruslan Salakhutdinov & Joshua B. Tenenbaum 2015. Human-level concept learning through probabilistic program induction. Science, 350, (6266), 1332-1338, doi:10.1126/science.aab3050.

What is the difference between deduction, induction, abduction ?

- **Deductive Reasoning** = Hypothesis > Observations > Logical Conclusions (general $\rightarrow$ specific – proven correctness)
  - DANGER: Hypothesis must be correct! DR defines whether the truth of a conclusion can be determined for that rule, based on the truth of premises: A=B, B=C, therefore A=C
- **Inductive reasoning** = makes broad generalizations from specific observations (specific $\rightarrow$ general – not proven correctness)
  - DANGER: allows a conclusion to be false if the premises are true
  - generate hypotheses and use DR for answering specific questions
- **Abductive reasoning** = inference = to get the best explanation from an incomplete set of preconditions.
  - Given a true conclusion and a rule, it attempts to select some possible premises that, if true also, may support the conclusion, though not uniquely.
  - Example: "When it rains, the grass gets wet. The grass is wet. Therefore, it might have rained." This kind of reasoning can be used to develop a hypothesis, which in turn can be tested by additional reasoning or data.

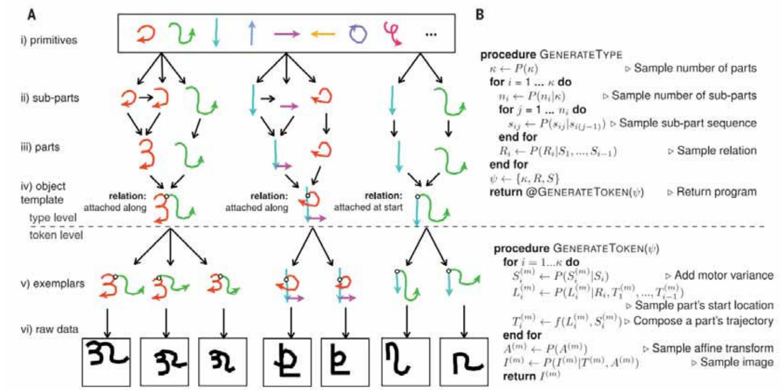
Drawn by Human or Machine Learning Algorithm ?



Brenden M. Lake, Ruslan Salakhutdinov & Joshua B. Tenenbaum 2015. Human-level concept learning through probabilistic program induction. Science, 350, (6266), 1332-1338, doi:10.1126/science.aab3050.

What can a Bayesian program learning (BPL) framework do ?

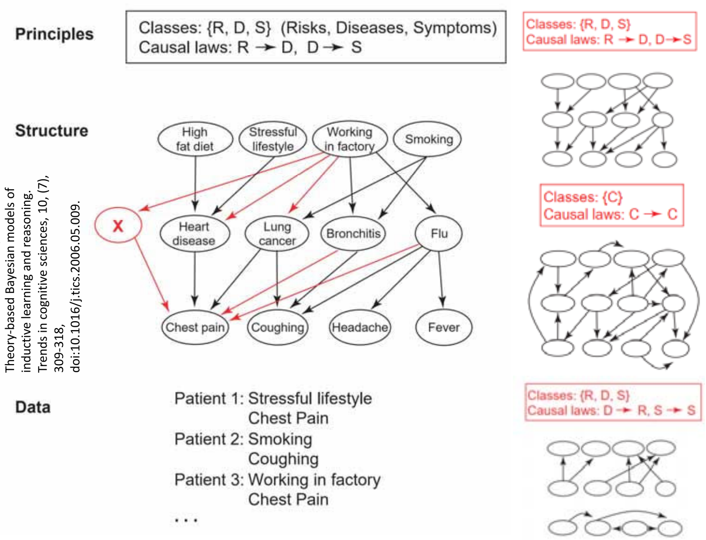
A Bayesian program learning (BPL) framework, capable of learning a
and
om people



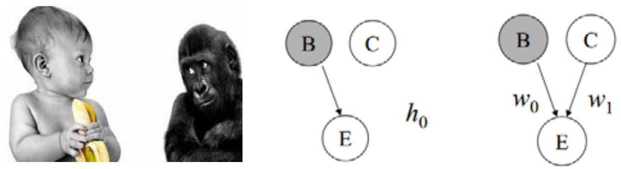
Brenden M. Lake, Ruslan Salakhutdinov & Joshua B. Tenenbaum 2015. Human-level concept learning through probabilistic program induction. Science, 350, (6266), 1332-1338, doi:10.1126/science.aab3050.

What can we do with graphical models ?

Joshua B. Tenenbaum, Thomas L. Griffiths & Charles Kemp 2006. Theory-based Bayesian models of inductive learning and reasoning. Trends in cognitive sciences, 10, (7), 309-318, doi:10.1016/j.tics.2006.05.009.



Does a relationship exist? If yes … how strong?



- Cognition as probabilistic inference
  - Visual perception, language acquisition, motor learning, associative learning, memory, attention, categorization, reasoning, causal inference, decision making, theory of mind
- Learning concepts from examples
- Learning causation from correlation
- Learning and applying intuitive theories (balancing complexity vs. fit)

Thank you!

Appendix