Andreas Holzinger
185.A83 Machine Learning for Health Informatics
2019S, VU, 2.0 h, 3.0 ECTS
Lecture 04 – Dienstag, 02.04.2019

From Decision Making under uncertainty to graphical models and MCMC

andreas.holzinger AT tuwien.ac.at
https://hci-kdd.org/machine-learning-for-health-informatics-class-2019

Red thread through the lecture today

- 00 Reflection from last lecture
- 01 Decision Making under uncertainty
- 02 Some Basics of Markov Processes
- 03 Some Basics of Concept Learning
- 04 Some Basics of Graphs/Networks and Challenges
- 05 Bayes Nets
- 06 Probabilistic Programming
- 07 Markov Chain Monte Carlo (MCMC)
- 08 Metropolis Hastings Algorithm

Why Explainability? Why Causability?

Remember always

Causability :=
a property of a person (Human Intelligence)

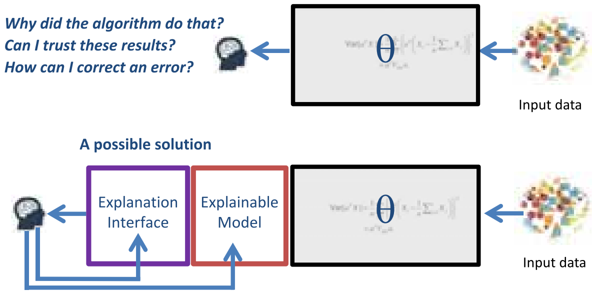
Explainability :=
a property of a system (Artificial Intelligence)

Andreas Holzinger et al. 2019. Causability and Explainability of AI in Medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, doi:10.1002/widm.1312.

https://onlinelibrary.wiley.com/doi/full/10.1002/widm.1312

We need effective tools for Human-AI Interaction

Why did the algorithm do that?
Can I trust these results?
How can I correct an error?



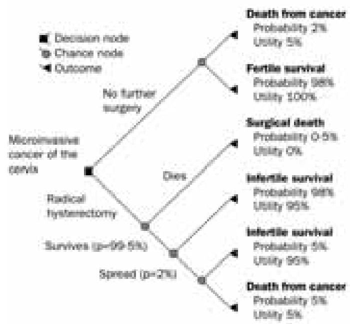
The domain expert can understand why …
The domain expert can learn and correct errors …
The domain expert can re-enact on demand …

Which Methods of ex-AI do you know?

- 1) Gradients
- 2) Sensitivity Analysis
- 3) Decomposition Relevance Propagation
  - Pixel-RP, Layer-RP, Deep Taylor Decomposition, …
- 4) Optimization
  - Local Interpretable Model-Agnostic Explanations (LIME)
  - Black Box Explanations tr. Transparent Approximations (BETA)
- 5) Deconvolution and Guided Backpropagation
- 6) Concept Activation Vectors CAV

Andreas Holzinger LV 706.315 From explainable AI to Causability, 3 ECTS course at Graz University of Technology
https://hci-kdd.org/explainable-ai-causability-2019 (course given since 2016)

Decision trees are coming from Clinical Practice



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

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

Our next goal is to bring those two together …

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.

Remember always: Medical action is …

Learning and Inference

04

d … data

h … hypotheses

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

Likelihood

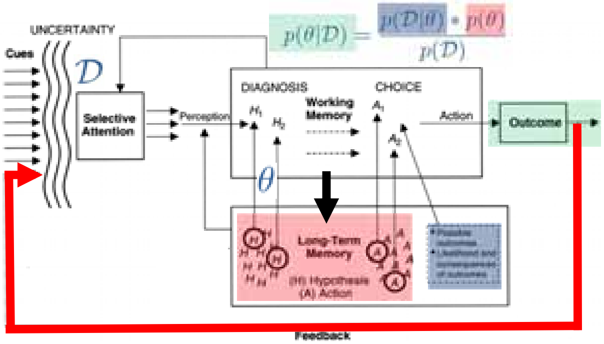
Prior Probability

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

Posterior Probability

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

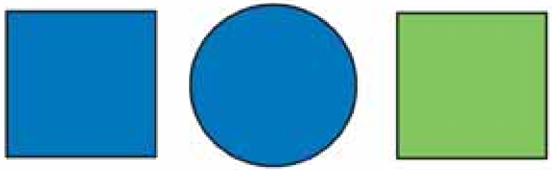
Connection to Cognitive Science: Decision Making



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

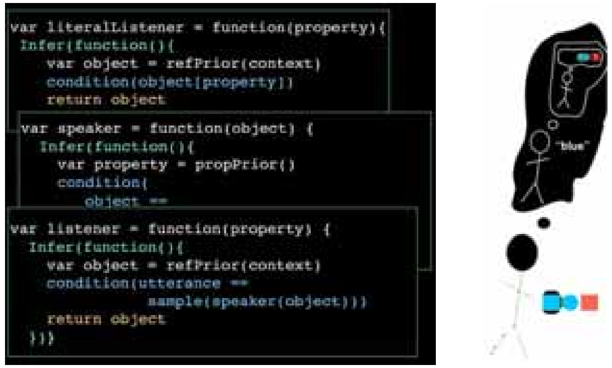
Predicting Pragmatic Reasoning in Language Games

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



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

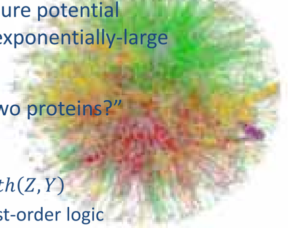
Recursive reasoning: a case for probabilistic programming



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

What are Probabilistic Graphical Models?

- 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?"



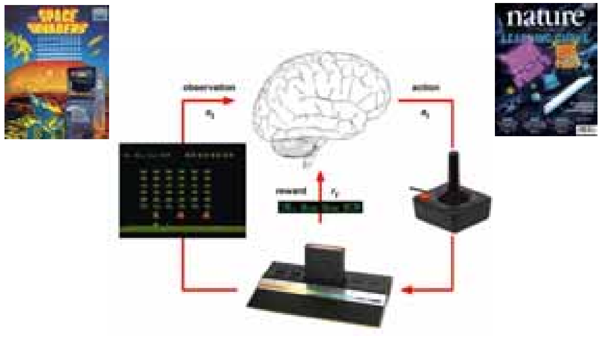
- $Path\ (X,Y) := edge\ (X,Y)$
- $Path\ (X,Y) := edge\ (X,Y), path(Z,Y)$
- This can NOT be expressed in first-order logic
- Would need a Turing-complete fully-fledged language

2) Some basics of Markov Processes in Machine Learning

Please remember:

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

Learning to play an Atari Game



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

Example Video Atari Game

Goal: Select actions to maximize total future reward

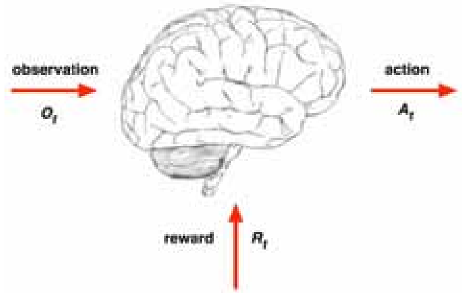
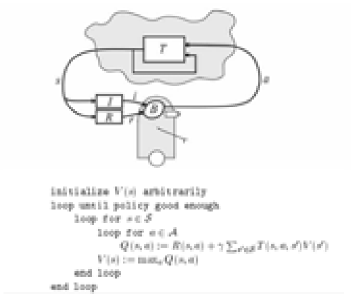


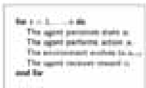
Image credit to David Silver, UCL

Standard RL-Agent Model goes back to Cybernetics 1950

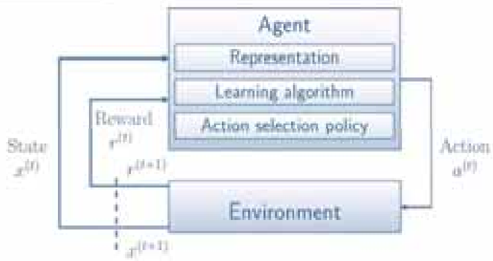


Kaelbling, L. P., Littman, M. L. & Moore, A. W. 1996. Reinforcement learning: A survey. Journal of Artificial Intelligence Research, 4, 237-285.

RL-Agent seeks to maximize rewards



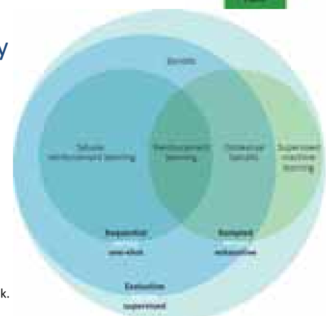
Intelligent behavior arises from the actions of an individual seeking to **maximize its received reward** signals in a **complex and changing world**



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

RL – Types of Feedback (crucial!)

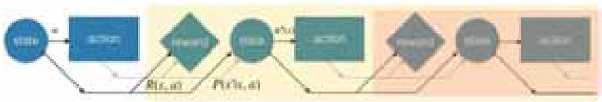
- Supervised: Learner told best a
- Exhaustive: Learner shown every possible x
- One-shot: Current x independent of past a



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)



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

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)

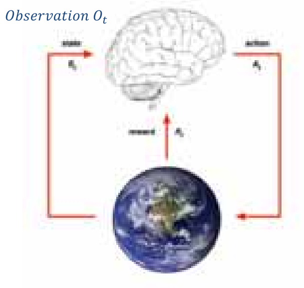
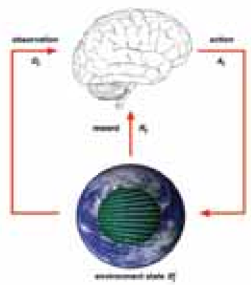


Image credit to David Silver, UCL

Environmental State is the current representation

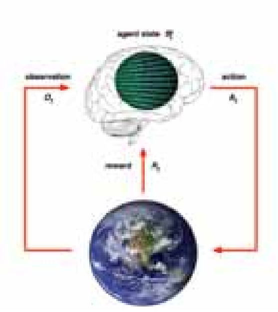
- i.e. whatever data the environment uses to pick the next observation/reward
- The environment state is not usually visible to the agent
- Even if S is visible, it may contain irrelevant information
- A State S_t is Markov iff:

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



Agent State is the agents internal representation

- 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:
- $S = \ f\ (H)$



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

Components of RL Agents and Policy of Agents

- 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:

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

What if the environment is only partially observable?

- Partial observability: when agent only indirectly observes environment (robot which is not aware of its current location; good example: Poker play: only public cards are observable for the agent):
- 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: $S_t^a = \sigma(S_{t-1}^a W_s + O_t W_o)$

3) Some basics of Concept Learning

Remember: Reasoning = “Sensemaking”

- **Deductive Reasoning** = Hypothesis > Observations > Logical Conclusions (general → 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 → 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.

When is a cup a cup? (When is a cat a cat?)

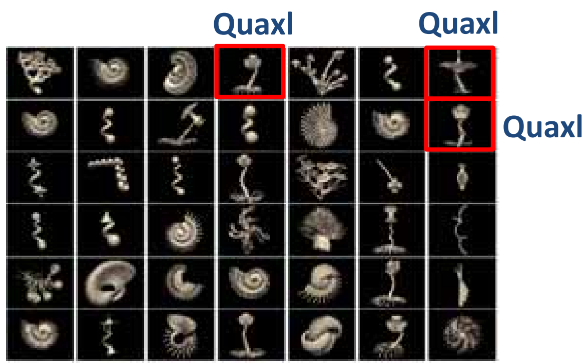
- Bruner, Goodnow, and Austin (1956) published “A Study of Thinking”, which became a landmark in cognitive science and has much influence on machine learning.
  - Rule-Based Categories
  - A concept specifies conditions for membership

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

How does our mind get so much out of it …

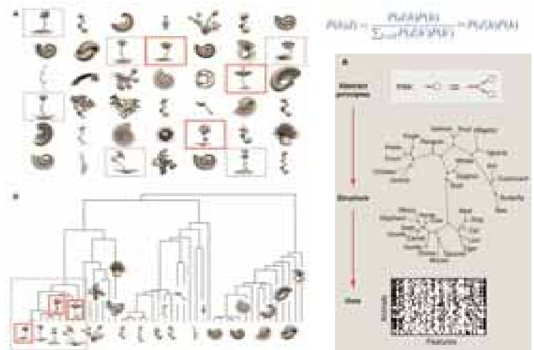
Salakhutdinov, R., Tenenbaum, J. & Torralba, A. 2012. One-shot learning with a hierarchical nonparametric Bayesian model. Journal of Machine Learning Research, 27, 195-207.

Learning 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 …



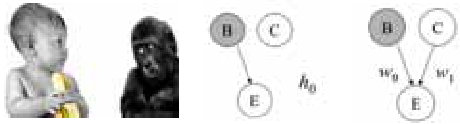
Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. 2011. How to grow a mind: Statistics, structure, and abstraction. Science, 331, (6022), 1279-1285.

Concept learning

- 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.

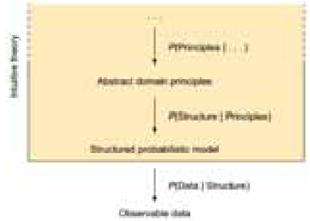
Does a relationship exist? If yes … how strong?



- 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)

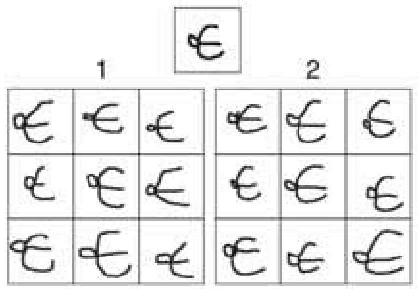
Modeling basic cognitive capacities as intuitive Bayes

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



Tenenbaum, J. B., Griffiths, T. L. & Kemp, C. 2006. Theory-based Bayesian models of inductive learning and reasoning. Trends in cognitive sciences, 10, (7), 309-318.

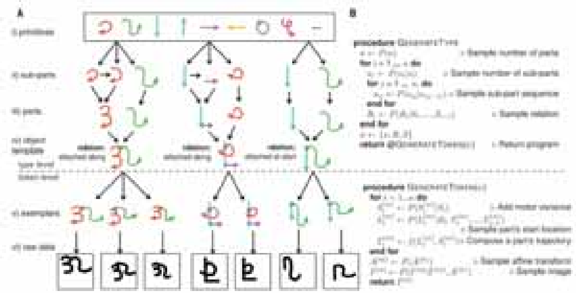
Drawn by Human or Machine Learning Algorithm?



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

Human-Level concept learning – probabilistic induction

A Bayesian program learning (BPL) framework, capable of learning a large class of visual concepts from just a single example and generalizing in ways that are mostly indistinguishable from people



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

4) Graphs=Networks

Leonhard Euler 1736 …

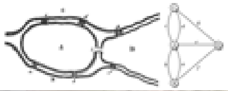
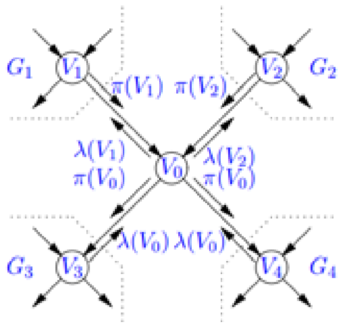


Image from https://people.kth.se/~carlofi/teaching/FEL3250-2013/courseinfo.html

252 years later: Belief propagation algorithm

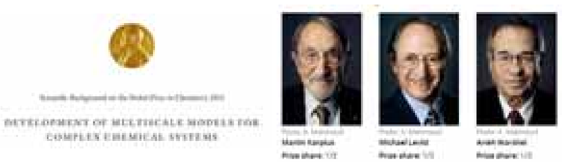


Pearl, J. 1988. Embracing causality in default reasoning. Artificial Intelligence, 35, (2), 259-271.

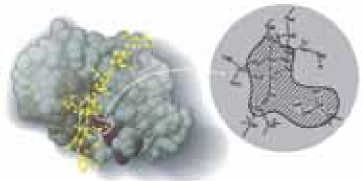
275 years later … the "Nobel-prize in Computer Science"

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

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 does 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., …)

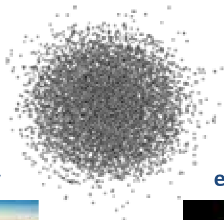
- **Graphs as nonparametric basis**
- we learn the structure from samples and infer
- flat vector data, e.g. similarity graphs
- encoding structural properties (e.g. smoothness, independence, …)

We skip this interesting chapter for now …

Our World in Data (1/2) – Macroscopic Structures

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

Two thematic mainstreams in dealing with data …



Time

e.g. Entropy

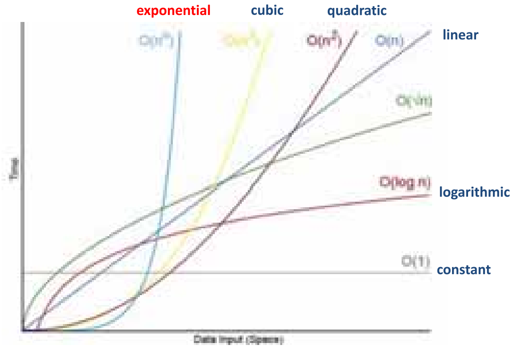
Space

e.g. Topology

Dali, S. (1931) The persistence of memory

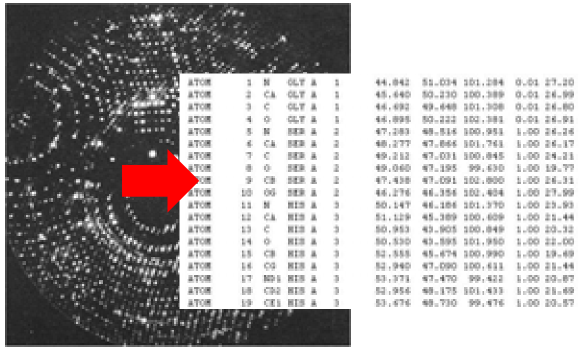
Bagula & Bourke (2012) Klein-Bottle

Complexity Problem: Time versus Space



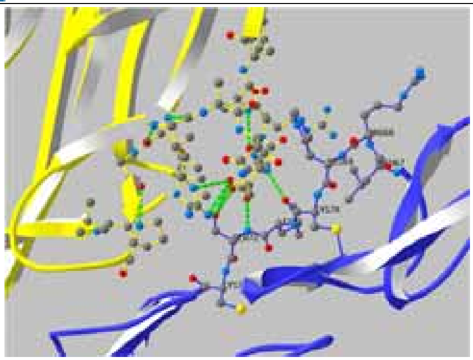
P versus NP and the Computational Complexity Zoo, please have a look at https://www.youtube.com/watch?v=YX40hbAHx3s

Our World in Data – Microscopic Structures



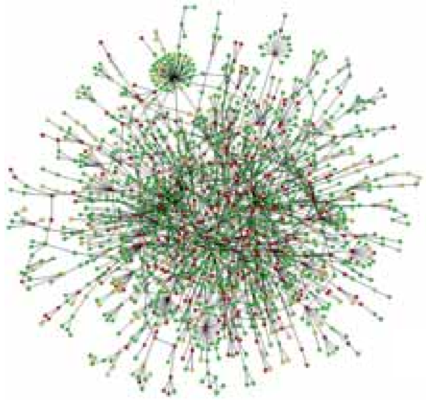
Wiltgen, M. & Holzinger, A. (2005) Visualization in Bioinformatics: Protein Structures with Physicochemical and Biological Annotations. In: *Central European Multimedia and Virtual Reality Conference. Prague, Czech Technical University (CTU), 69-74*

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

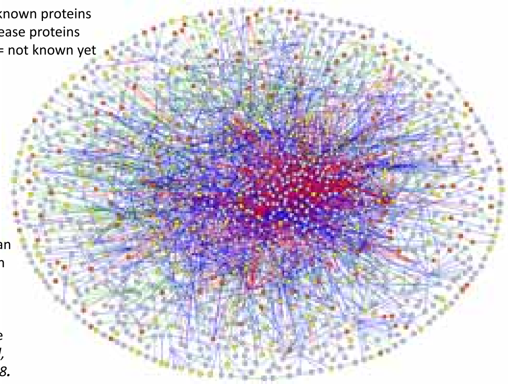


Nodes = proteins
Links = physical interactions (bindings)
Red Nodes = lethal
Green Nodes = non-lethal
Orange = slow growth
Yellow = not known

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

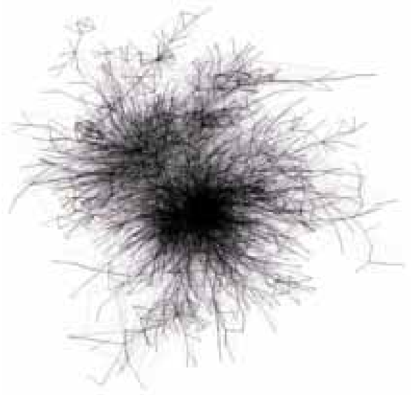
First human protein-protein interaction network

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



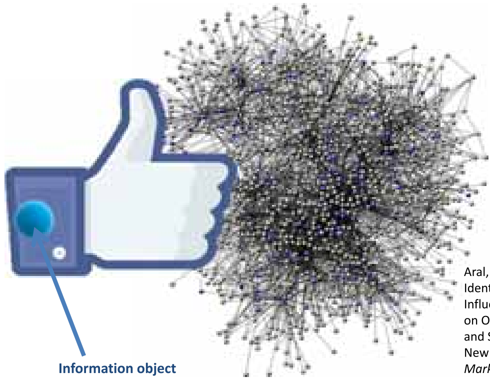
Stelzl, U. et al. (2005) A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome. *Cell, 122, 6, 957-968.*

Non-Natural Network Example: Blogosphere



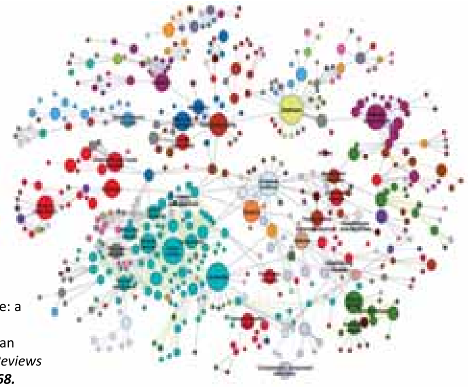
Hurst, M. (2007), Data Mining: Text Mining, Visualization and Social Media. Online available: http://datamining.typepad.com/data_mining/2007/01/the_blogosphere.html, last access: 2011-09-24

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.***

05 Bayesian Networks "Bayes' Nets"

Bayesian Network (BN) - Definition

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

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

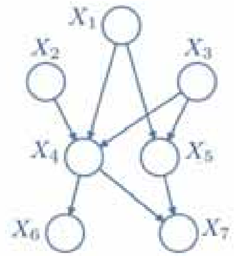
Where $Pa(x_i)$ are the parents of x_i

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.

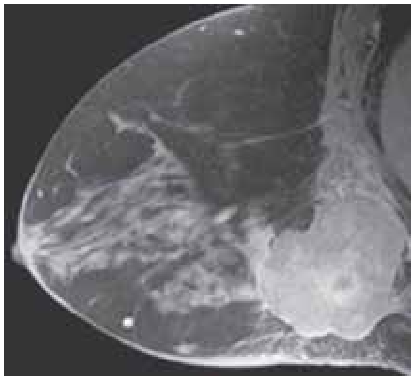
Pearl, J. (1988) *Probabilistic reasoning in intelligent systems: networks of plausible inference. San Francisco, Morgan Kaufmann.*

Example: Directed Bayesian Network with 7 nodes

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



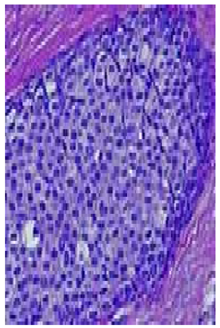
Clinical Case Example



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.*

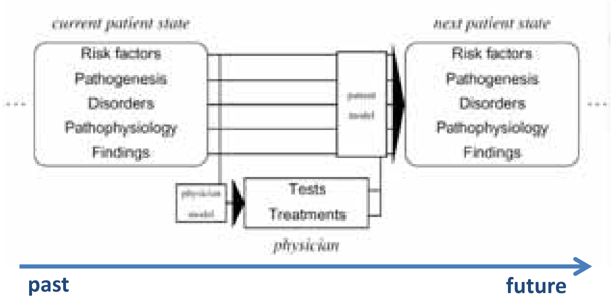
Important in Clinical practice -> prognosis !

- = 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.*

Predicting the future on past data and present status



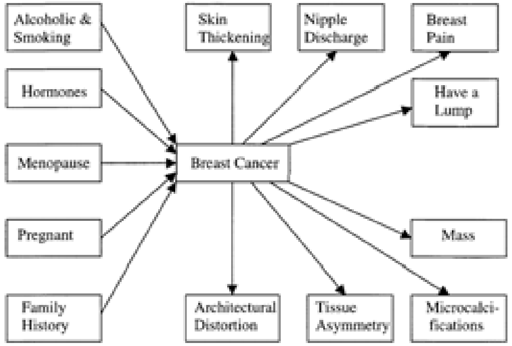
van Gerven, M. A. J., Taal, B. G. & Lucas, P. J. F. (2008) Dynamic Bayesian networks as prognostic models for clinical patient management. *Journal of Biomedical Informatics, 41, 4, 515-529.*

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.*

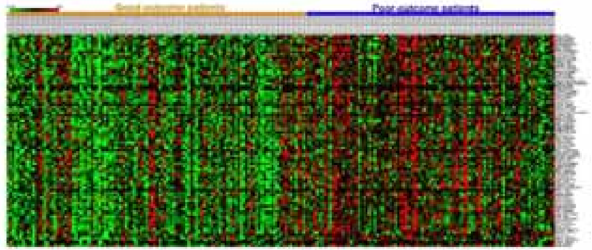
Breast cancer – big picture – state of 1999



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

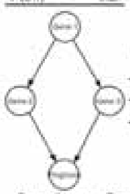


Xu, L., Tan, A., Winslow, R. & Geman, D. (2008) Merging microarray data from separate breast cancer studies provides a robust prognostic test. *BMC Bioinformatics, 9, 1, 125-139.*

Example: 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



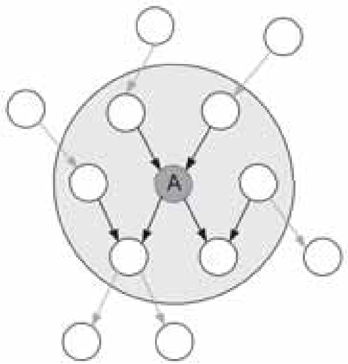
Gene 2	Gene 1 on	Gene 1 off
P(on)	0.3	0.6
P(off)	0.7	0.4

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:

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

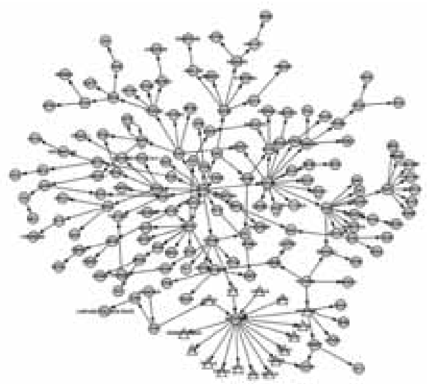
Note: With θ_{ij} a parameter set where i refers to the variable and j to the j-th instantiation of the parents in the current structure. θ_{ij} contains a probability for every value of the variable x_i given the current instantiation of the parents. Dir corresponds to the Dirichlet distribution with $(N'_{ij1}, \dots, N'_{ijr_i})$ 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:

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

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, NOTE:

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

My name is Andreas Holzinger …

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

Now we connect the two …

06 Probabilistic Programming

Probabilistic-programming.org

- 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

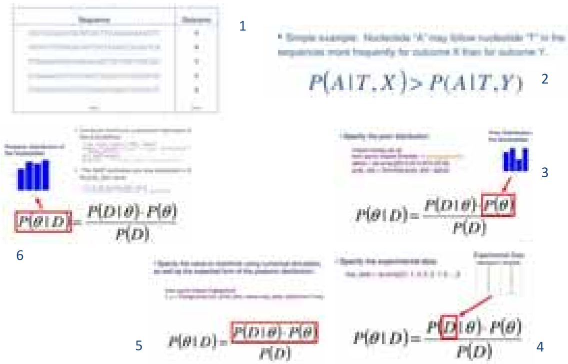
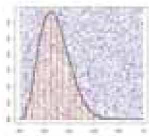


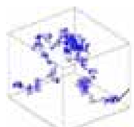
Image Source: Dan Williams, Life Technologies, Austin TX

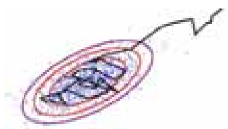
Finally a practical example

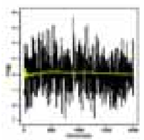
07 Markov Chain Monte Carlo (MCMC)

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









In real-world systems you have observable data $\mathcal{D}$

- 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:

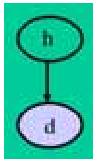
$$J = \int f(\theta) * p(\theta|\mathcal{D}) d\theta$$

Origin

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

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

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



Simulation of samples …

named after

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!

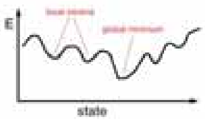
MC connects Computer Science with Cognitive Science

- 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 $P(x)$

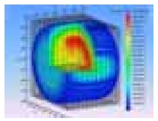
$$\mathbb{E}_{p(x)}(f(x)) = \sum_X f(x)P(x)dx$$

Mathematical simulation via MC

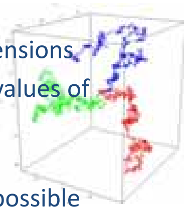
- Solving intractable integrals
- Bayesian statistics: **normalizing** constants, expectations, marginalization
- Stochastic Optimization
- Generalization of simulated annealing
- Monte Carlo expectation maximization (EM)



Physical simulation via MC



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



Facts on one slide

- 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 $cov[x, y]$

Global optimization: What is the main problem?

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

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

Finally a practical example

08 Metropolis-Hastings Algorithm

Metropolis, Rosenbluth et al. (1953), Hastings (1970)

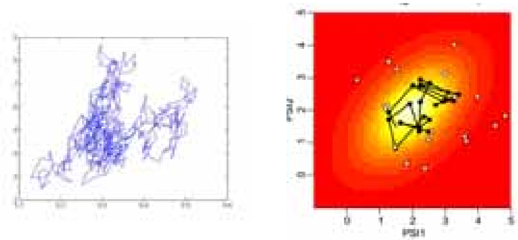


Image Source: Peter Mueller, Anderson Cancer Center

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 q̃(x'|x^{l-1}).
    Let a = q̃(x^{l-1}|x^cand) p(x^cand) / q̃(x^cand|x^{l-1}) p(x^{l-1})
    if a ≥ 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 $q(x)$.

Gibbs Sampling

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

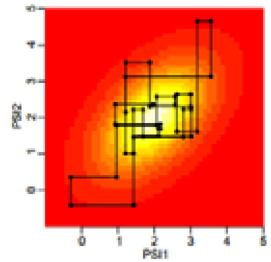
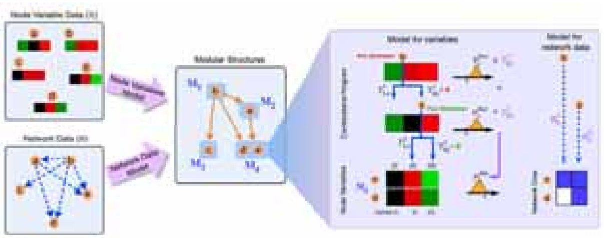
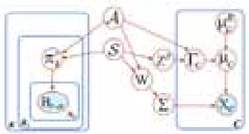


Image Source: Peter Mueller, Anderson Cancer Center

Example

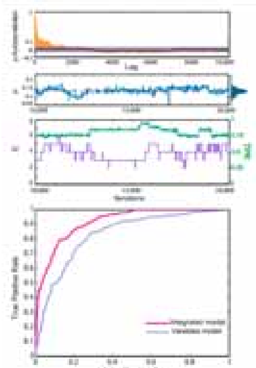


Azizi, E., Airoldi, E. M. & Galagan, J. E. 2014. Learning Modular Structures from Network Data and Node Variables. Proceedings of the 31st International Conference on Machine Learning (ICML). Beijing: JMLR. 1440-1448.



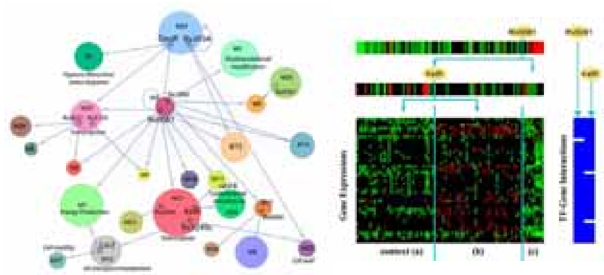
Algorithm 1 RJMCMC for sampling parameters

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_{trick}; 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_{trick}
Propose $\pi_k^{r(j+1)} \sim \mathcal{N}(\pi_k^{r(j)}, I)$; accept with P_{trick}
end for
end for
for condition $c = 1$ **to** C **do**
Propose $\mu_c^{\mathbf{B}(j+1)} \sim \mathcal{N}(\mu_c^{\mathbf{B}(j)}, I)$; accept with P_{trick}
Propose $\gamma_c^{\mathbf{B}(j+1)} \sim \mathcal{N}(\gamma_c^{\mathbf{B}(j)}, I)$; accept with P_{trick}
end for
end for



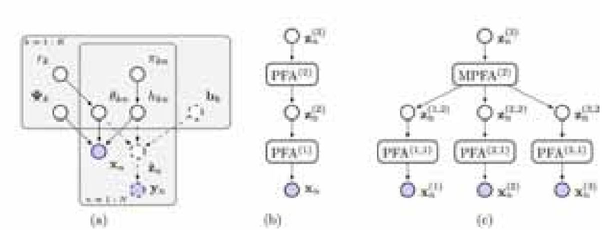
Azizi, E., Airoldi, E. M. & Galagan, J. E. 2014. Learning Modular Structures from Network Data and Node Variables. Proceedings of the 31st International Conference on Machine Learning (ICML). Beijing: JMLR. 1440-1448.

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.

Electronic Health Record Analysis via Deep Poisson



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.

Thank you!

Questions

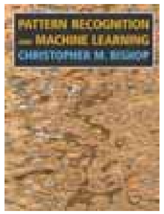
Sample Questions

- 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!

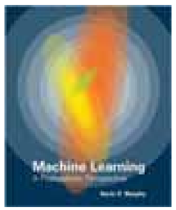
- What is a factored graph?
- Describe the protein structure prediction problem! Why is it hard?
- Why are protein-protein interactions so important?
- Describe the problem of graph-isomorphism!
- How does a Bayes Net work?
- Why is predicting important in clincial medicine?
- What is a Markov-Blankett?
- 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

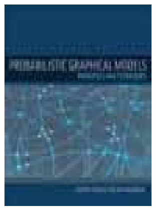
Basics and Background reading



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/



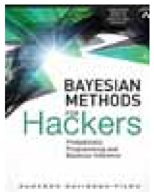
Murphy, K. P. 2012. Machine learning: a probabilistic perspective, MIT press. Chapter 26 (pp. 907) – Graphical model structure learning



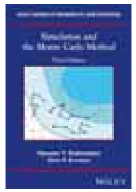
Koller, D. & Friedman, N. 2009. Probabilistic graphical models: principles and techniques, MIT press.

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



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

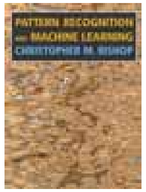


Rubinstein, R. Y. & Kroese, D. P. 2013. Simulation and the Monte-Carlo Method, Wiley

My friend has glasses: can you show me my friend?

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

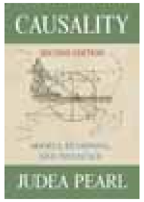
Pattern Recognition and Machine Learning Chapter 8



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.

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.

Why is RL interesting?

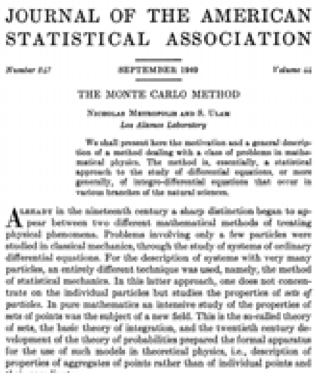
- Reinforcement Learning is the **oldest approach,** with the longest history and can provide insight into understanding human learning [1]
- RL is the **"AI problem in the microcosm"** [2]
- Future opportunities are in Multi-Agent RL (MARL), Multi-Task Learning (MTL), Generalization and **Transfer-Learning** [3], [4].

[1] Turing, A. M. 1950. Computing machinery and intelligence. Mind, 59, (236), 433-460.
[2] Littman, M. L. 2015. Reinforcement learning improves behaviour from evaluative feedback. Nature, 521, (7553), 445-451, doi:10.1038/nature14540.
[3] Taylor, M. E. & Stone, P. 2009. Transfer learning for reinforcement learning domains: A survey. The Journal of Machine Learning Research, 10, 1633-1685.
[4] Pan, S. J. & Yang, Q. A. 2010. A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering, 22, (10), 1345-1359, doi:10.1109/tkde.2009.191.

Remember: Three main types of Machine Learning

- **I) Supervised learning (classification)**
  - $y = f(x)$
  - Given x, y pairs; find a f that map a new x to a proper y
  - Regression, logistic regression, classification
  - Expert provides examples e.g. classification of clinical images
  - Disadvantage: Supervision can be expensive
- **II) Unsupervised learning (clustering)**
  - $f(x)$
  - Given x (features only), find f that gives you a description of x
  - Find similar points in high-dim X
  - E.g. clustering of medical images based on their content
  - Disadvantage: Not necessarily task relevant
- **III) Reinforcement learning**
  - $y = f(x)$
  - more general than supervised/unsupervised learning
  - learn from interaction to achieve a goal
  - Learning by direct interaction with environment (automatic ML)
  - Disadvantage: broad difficult approach, problem with high-dim data

5,223 citations as of 26.03.2017



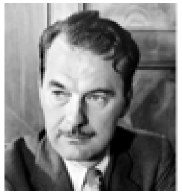


Image Source: http://www.manhattanprojectvoices.org/oral-histories/nicholas-metropolis-interview

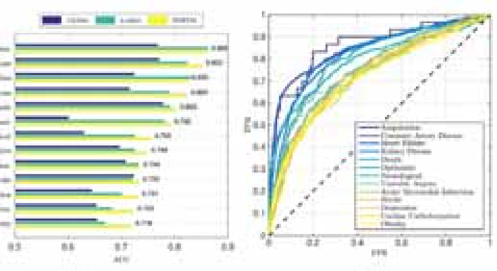
37,202 (as of 10.4.2018) - 34,140 citations (26.03.2017)

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.

12,081 as of 10.4.2018 - 10,624 citations 26.03.2017

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

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.

Finally a practical example

Graphical Model Learning

Learning Graphical Models from data

- 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

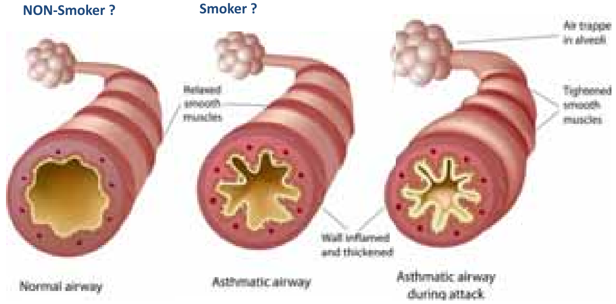
1) Test if a distribution is decomposable with regard to a given graph.
   - This is the most direct approach. It is not bound to a graphical representation,
   - It can be carried out w.r.t. other representations of the set of subspaces to be used to compute the (candidate) decomposition of a given 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	?	?

Florian	0	0.3	0.2

Rows are independent during learning and inference!

Relational Representation Learning and Prediction

- 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

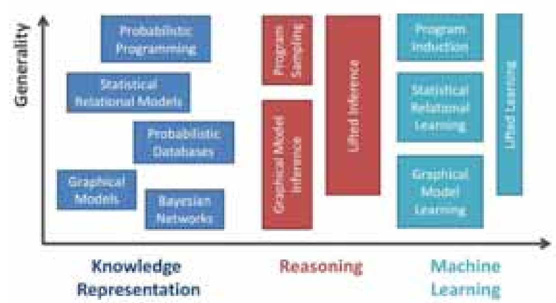
2.1 Asthma(x) ⇒ Cough(x)

3.5 Smokes(x) ⇒ Cough(x)

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

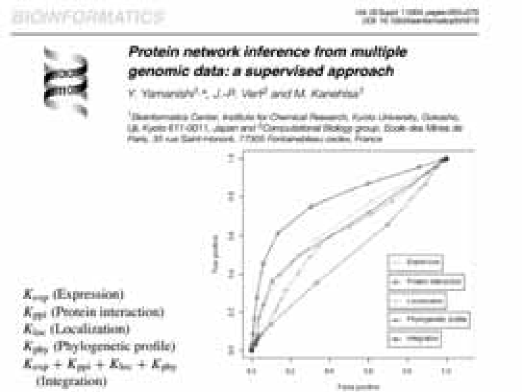
1.5 Asthma (x) ∧ Family(x,y) ⇒ Asthma (y)

Knowledge Representation > Reasoning > Learning

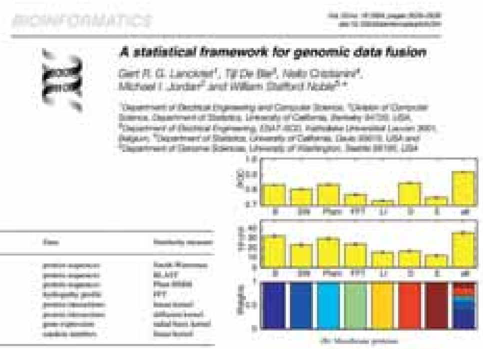


Example for probabilistic rule learning, in which probabilistic rules are learned from probabilistic examples: The ProbFOIL+ Algorithm solves this problem by combining the principles of the rule learner FOIL with the probabilistic Prolog called ProbLog, see: De Raedt, L., Dries, A., Thon, I., Van Den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

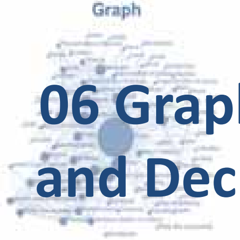
Example: Protein Network Inference



Example: Data fusion and Protein Annotation



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.

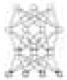
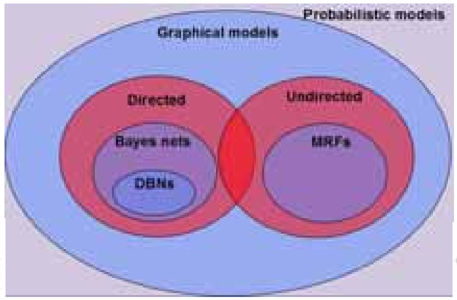


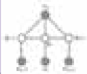
06 Graphical Models and Decision Making

$$\mathcal{M}$$

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

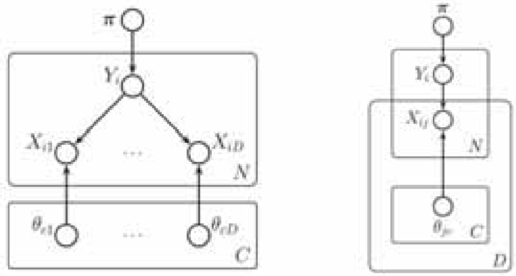
Classes of Graphical Models



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

Naïve Bayes classifier as DGM (single/nested plates)



π ... multinomial parameter vector, Stationary distribution of Markov chain

Regulatory>Metabolic>Signaling>Protein>Co-expression

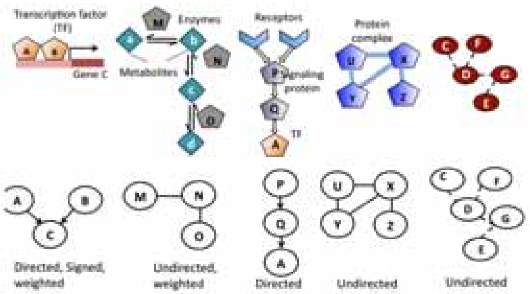
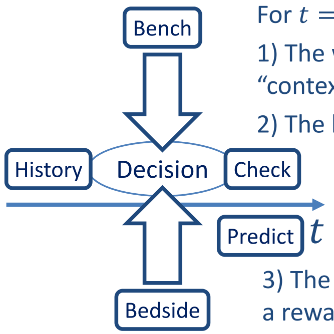


Image credit to Anna Goldenberg, Toronto

Decision Making: Learn good policy for selecting actions

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

GM are amongst the most important ML developments

- 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:

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

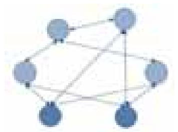
can be abbr. with $P(X|Y,Z) = P(X|Z)$

- 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

Remember

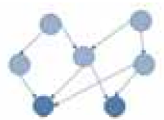
- Medicine is an extremely complex application domain – dealing most of the time with uncertainties -> **probable information!**
- When we have big data but little knowledge automatic ML can help to gain insight:
- **Structure learning and prediction in large-scale biomedical networks with probabilistic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

Three types of Probabilistic Graphical Models



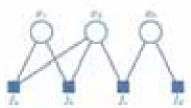
Undirected: Markov random fields, useful e.g. for computer vision (Details: Murphy 19)

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



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

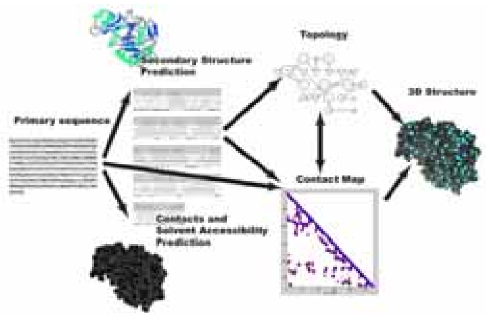
$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k | \mathrm{pa}_k)$$



Factored: useful for inference/learning

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

From structure to function prediction



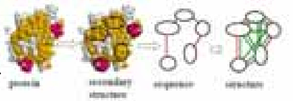
Baldi, P. & Pollastri, G. 2003. The principled design of large-scale recursive neural network architectures--dag-rnns and the protein structure prediction problem. The Journal of Machine Learning Research, 4, 575-602.

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.

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