Andreas Holzinger
185.A83 Machine Learning for Health Informatics
2018S, VU, 2.0 h, 3.0 ECTS
Lecture 06 - Module 04 – Week 20 - 15.05.2018

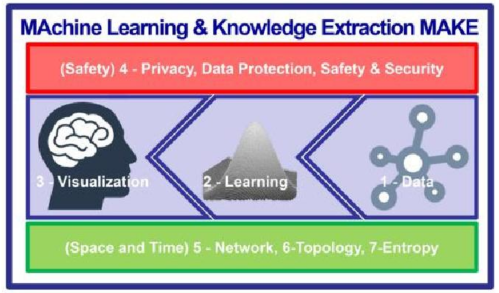
Probabilistic Graphical Models Part 2: From Bayesian Networks to Probabilistic Topic Models

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

Crazy Ideas > Science > Engineering > Business

Science **is to test crazy ideas –**
Engineering is to put these ideas into Business
Lucky Students ☺

ML needs a concerted effort fostering integrated research



Machine Learning Jungle Top-Level View

Cognition | Visualization | Data fusion
Perception | Preprocessing

Decision | Interaction | Integration

CONCEPTS	THEORIES	PARADIGMS	MODELS	METHODS	TOOLS
Dimensionality	Complexity	Unsupervised	Gaussian P.	Regularization	Python
Reinforcement	Bayesian p(x)	Supervised	Graphical M.	Scaling	Church
Representation	Entropy/KL	Semi-Superv.	Neural Nets	Aggregation	Anglican
No-free-lunch	Vapnik-Chernov.	iML	Kernel/SVM	Evolution	Julia

Multi-Task Learning | Transfer Learning | Multi-Agent-Hybrid-Systems

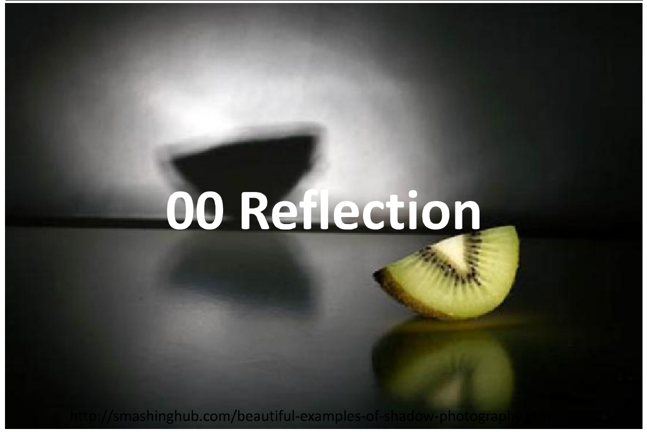
Data Protection, Safety and Security and Privacy Aware Machine Learning (PAML)

Application, Validation, Evaluation, Impact – Social, Economic, Acceptance, Trust

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

Red thread through the lecture today

- **00 Reflection**
- **01 Probabilistic Decision Making**
- **02 Probabilistic Programming Part II**
- **03 Probabilistic Topic Models**
- **04 Knowledge Representation in Net Medicine**
- 05 ML on Graphs Examples
- 06 Digression: Similarity
- 07 Graph Measures
- 08 Point Clouds from Natural Images



To reach a level of <u>usable</u> intelligence we need to …

04

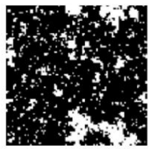
- 1) **learn** from prior data
- 2) **extract** knowledge
- 2) **generalize,**
  - i.e. guessing where a probability mass function concentrates
- 4) fight the curse of **dimensionality**
- 5) **disentangle** underlying explanatory factors of data, i.e.
- 6) **understand** the data in the **context** of an application domain

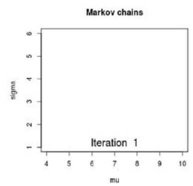
Sampling from big data is an important topic

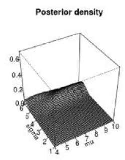
$$\mathrm{E}[f] = \int f(z)p(z)dz$$

$$\hat{f} = \frac{1}{L}\sum_{l=1}^{L} f(z^{(l)})$$

Compute $a_i := \sum_j J_{ij} x_j$
Draw u from Uniform$(0,1)$
If $u < 1/(1+e^{-2a_i})$
$x_i := +1$
Else
$x_i := -1$

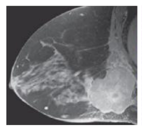


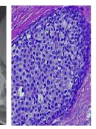


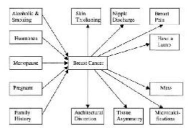


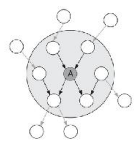
Propp, J. G. & Wilson, D. B. 1996. Exact sampling with coupled Markov chains and applications to statistical mechanics. Random structures and Algorithms, 9, (1-2), 223-252.

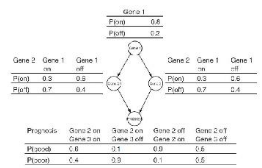
Medical Example: Breast cancer prognosis incl. Genetics











$$P(z_1,\dots,z_n) = \prod_{i=1}^{n} P(z_i|pa(z_i))$$

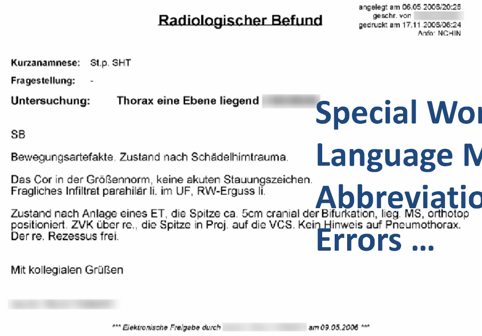
Gevaert, O., Smet, F. D., Timmerman, D., Moreau, Y. & Moor, B. D. (2006) Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics, 22, 14, 184-190.*

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

The medical report is the most important medium


Radiologischer Befund

angelegt am 06.05.2006/20:26
geschr. von
gedruckt am 17.11.2006/06:24
Anfo: NCHIN

Kurzanamnese: St.p. SHT

Fragestellung: -

Untersuchung: **Thorax eine Ebene liegend**

SB

Bewegungsartefakte. Zustand nach Schädelhirntrauma.

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

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

Mit kollegialen Grüßen

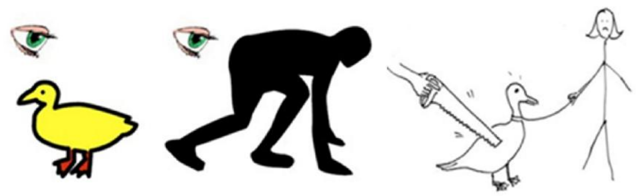
**** Elektronische Freigabe durch am 09.05.2006 ****

Special Words
Language Mix
Abbreviations
Errors …

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

Why is this sentence complex? What do I need?

"I saw her duck"



German Local Hospital Abbreviations … (example)

- HWI =
  - Harnwegsinfekt
  - Hinterwandinfarkt
  - Hinterwandischämie
  - Hakenwurminfektion
  - Halswirbelimmobilisation
  - Hip Waist Index
  - Height-Width Index
  - Heart-Work Index
  - Hemodynamically weighted imaging
  - High Water Intake
  - Hot water irrigation
  - Hepatitic weight index
  - Häufig wechselnder Intimpartner
- Leitung = Nervenleitung, Abteilungsleitung, Stromleitung, Wasserleitung, Harnleitung, Ableitung, Vereinsleitung ☺…

Final Quiz

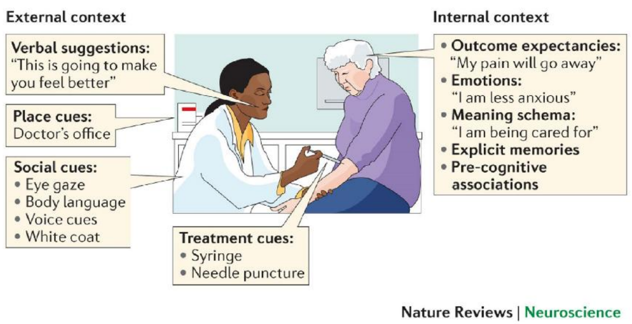
- Intelligence?
  - Hundreds of controversial definitions – very hard to define;
  - For us: ability to solve problems, to make decisions and to acquire and apply knowledge and skills
- Learning?
  - Different definitions – relatively hard to define
  - basically acquisition of knowledge through prior experience
- Problem Solving?
  - Process of finding solutions to complex issues
- Reasoning?
  - ability of our mind to think and understand things
- Sense Making?
  - Process of giving meaning to experience
- Causality?
  - Relationship between cause and effect
- Decision Making?
  - Process of "de-ciding" ("ent-scheiden") between alternative options

01 Probabilistic Decision Making

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

Why is the topic "decision making" so important …

Where can you apply artificial intelligence here?



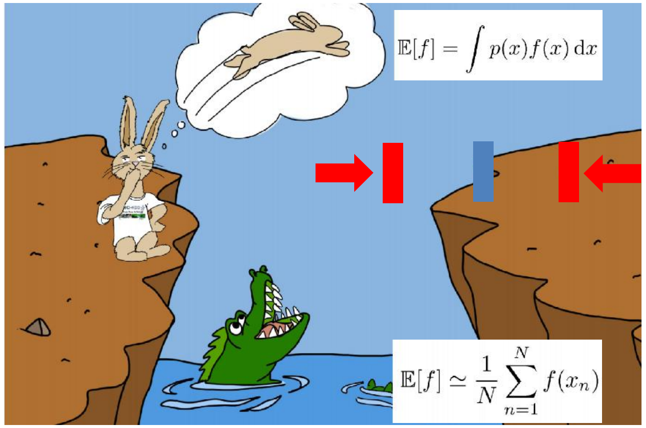
Wager, T. D. & Atlas, L. Y. 2015. The neuroscience of placebo effects: connecting context, learning and health. Nat Rev Neurosci, 16, (7), 403-418, doi:10.1038/nrn3976

Remember: 2 types of decisions (Diagnosis vs. Therapy)

- **Type 1 Decisions:** related to the **diagnosis**, i.e. computers are used to assist in diagnosing a disease on the basis of the individual patient data. Questions include:
  - What is the probability that this patient has a myocardial infarction on the basis of given data (patient history, ECG, …)?
  - What is the probability that this patient has acute appendices, given the signs and symptoms concerning abdominal pain?
- **Type 2 Decisions:** related to **therapy**, i.e. computers are used to select the best therapy on the basis of clinical evidence, e.g.:
  - What is the best therapy for patients of age x and risks y, if an obstruction of z % is seen in the left coronary artery?
  - What amount of insulin should be prescribed for a patient during the next 5 days, given the blood sugar levels and the amount of insulin taken during the recent weeks?

Harold C. Sox, Michael C. Higgins & Douglas K. Owens 1988. Medical decision making, Second Edition, Chichester, Wiley.

Decision Making under Uncertainty



Remember: 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

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

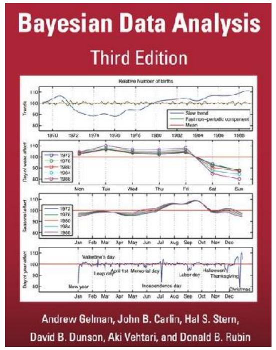
$$E(U \mid d) = \sum_{x_1,\ldots,x_n} P(x_1,\ldots,x_n \mid d) U(x_1,\ldots,x_n,d)$$

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

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

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

http://www.stat.columbia.edu/~gelman/book/



https://github.com/avehtari/BDA_py_demos

http://www.stat.columbia.edu/~gelman/book/data/

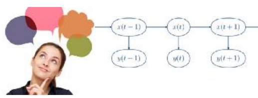
Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari & Donald B. Rubin 2014. Bayesian data analysis, Boca Raton (FL), CRC press.

Human learning vs. Machine Learning

- Example 1: Inverse Probability
- Example 2: Diagnosis
- Example 3: Language understanding

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

$$P(words|sounds) \propto P(sounds|words) * P(words)$$

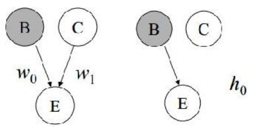


Recognize speech

Wreck a nice beach

- Learning ensures that new observations (d) match our previous hypotheses (h)

Cognition as probabilistic inference

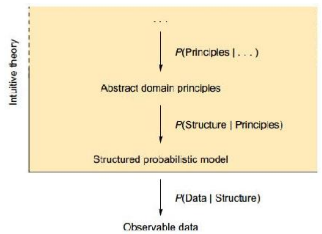


- Visual perception, language understanding, motor learning, associative learning, categorization, concept learning, reasoning, causal inference, …
- Learning concepts from (few!) examples
- Learning and applying intuitive theories (balancing complexity vs. fit optimality)

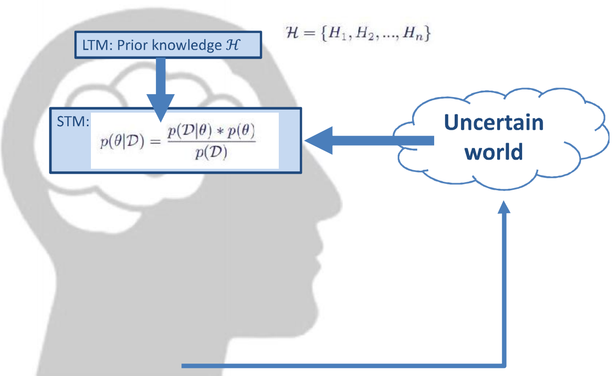
Modeling basic cognitive capacities as intuitive Bayes

- 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.



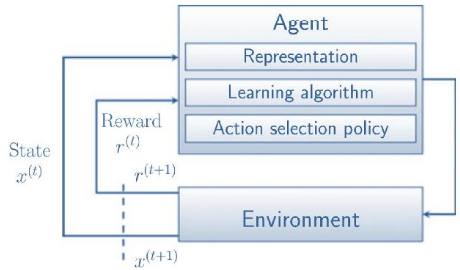
Human brains as probabilistic reasoning machines



Similar as our RL-Agent seeks to maximize rewards

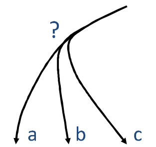
```
for t = 1, ..., n do
    The agent perceives state s_t
    The agent performs action a_t
    The environment evolves to s_{t+1}
    The agent receives reward r_t
end for
```

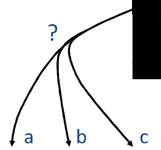
Intelligent behavior arises from the actions of an individual seeking to **maximize its received reward** signals in a **complex and changing world**



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

De-cision (Ent-scheidung) between alternatives





{a,b,c}

→ decision that is best for worst case

Non-deterministic model

~ Adversarial search

{a(p_a),b(p_b),c(p_c)}

→ **decision that maximizes expected utility value**

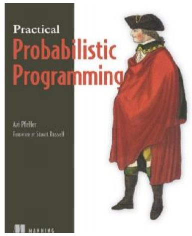
Probabilistic model

02 Probabilistic Programming

Recommended Resources for Probabilistic Programming

- Dan ROY: Probabilistic Programming Wiki
  http://www.probabilistic-programming.org/wiki/Home
- Frank WOOD, many tutorials, slides, code and papers
  http://www.robots.ox.ac.uk/~fwood/teaching/index.html
- Avi PFEFFER 2016. Practical probabilistic programming, Shelter Island (NY), Manning
  https://www.manning.com/books/practical-probabilistic-programming

Look also for work of:
Andrew GORDON
Noah GOODMAN
Josh TENENBAUM
John WINN
Rob ZINKOV
Vikash MANSINGHA
David WINGATE



Reasoning under uncertainty: Decision Making

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

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})} \qquad p(x_i|y_j) = \frac{p(y_j|x_i)p(x_i)}{\sum p(x_i, y_j)p(x_i)}$$

Learning representations (θ, h) from observed data

Observed data: 

≈ Training data: $\mathcal{D} = x_{1:n} = \{x_1, x_2, ..., x_n\}$ $\quad x, y \;\; A, B, ...$

Feature Parameter: θ **or hypothesis** ***h*** $\quad h \in \mathcal{H}$

Prior belief ≈ prior probability of hypothesis ***h*****:** $p(\theta) \quad p(h)$

Likelihood ≈ ***p(x)*** **of the data that** ***h*** **is true** $p(\mathcal{D}|\theta) \quad p(d|h)$

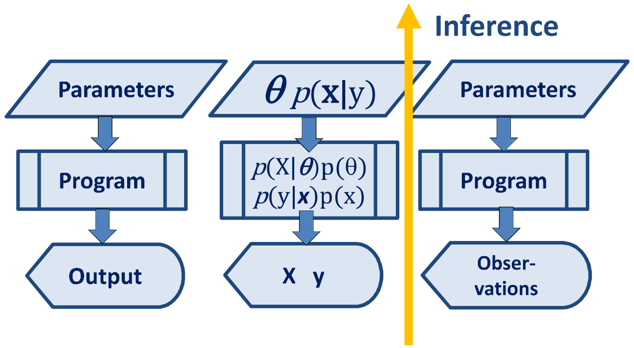
Data evidence ≈ marginal ***p(x)*** **that** ***h*** **= true** $p(\mathcal{D}) \quad \sum_{h \in \mathcal{H}} p(d|h) * p(h)$

Posterior ≈ ***p(x)*** **of** ***h*** **after seen ("learn") data** ***d*** $p(\theta|\mathcal{D}) \quad p(h|d)$

$$posterior = \frac{likelihood * prior}{evidence} \qquad p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta) * p(\theta)}{p(\mathcal{D})}$$

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

Probabilistic Programming Concept



Frank Wood, Jan-Willem Van De Meent & Vikash Mansinghka. A New Approach to Probabilistic Programming Inference. AISTATS 2014, Reykjavik, JMLR, 1024-1032

Comparison

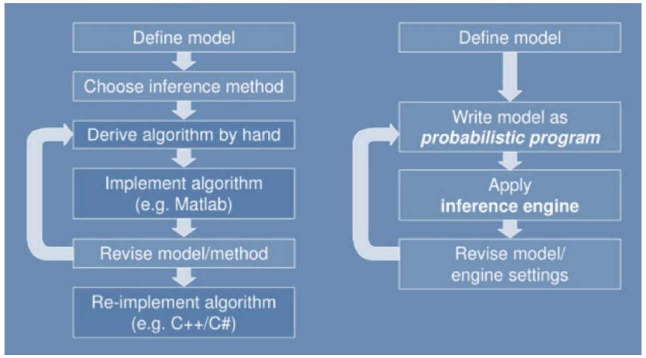
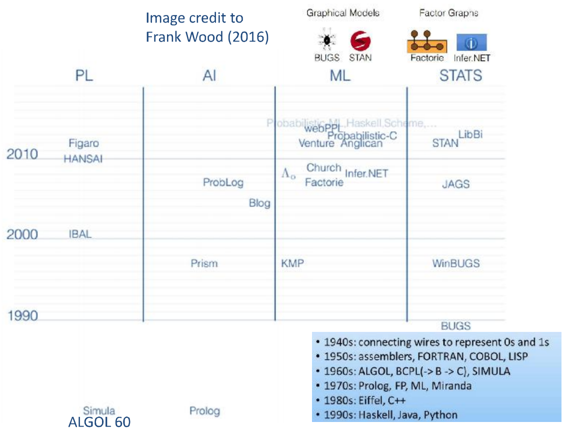


Image credit to John WINN (2010)

Probabilistic Programming Languages

Image credit to Frank Wood (2016)

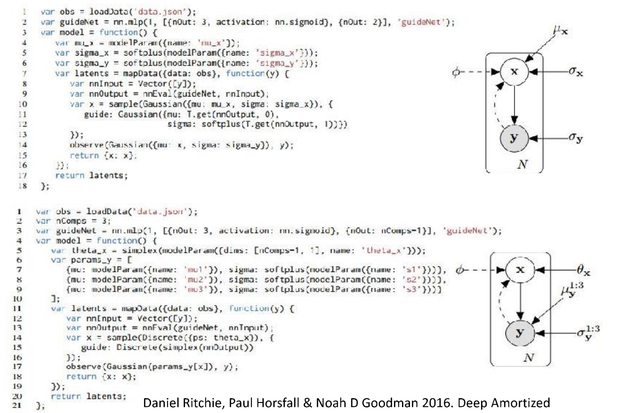


- 1940s: connecting wires to represent 0s and 1s
- 1950s: assemblers, FORTRAN, COBOL, LISP
- 1960s: ALGOL, BCPL(-> B -> C), SIMULA
- 1970s: Prolog, FP, ML, Miranda
- 1980s: Eiffel, C++
- 1990s: Haskell, Java, Python

Some selected PPLs

- https://github.com/pymc-devs/pymc
- http://infernet.azurewebsites.net/
- http://mc-stan.org/
- https://github.com/p2t2/figaro
- https://sites.google.com/site/bloginference/
- http://projects.csail.mit.edu/church/wiki/Church
- http://factorie.cs.umass.edu/
- http://www.openbugs.net/w/FrontPage
- http://mcmc-jags.sourceforge.net/

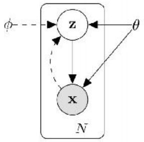
Try out WebPPL ("web-people") http://dippl.org



Daniel Ritchie, Paul Horsfall & Noah D Goodman 2016. Deep Amortized Inference for Probabilistic Programs. arXiv:1610.05735.

Remember: directed graphical model

Diederik P Kingma & Max Welling 2013. Auto-encoding variational Bayes. arXiv:1312.6114 (1983 citations as of 13.05.2018 07:00)



Algorithm 1 Minibatch version of the Auto-Encoding VB (AEVB) algorithm. Either of the two SGVB estimators in section 2.3 can be used. We use settings $M = 100$ and $L = 1$ in experiments.

$\boldsymbol{\theta}, \boldsymbol{\phi} \leftarrow$ Initialize parameters
repeat
  $\mathbf{X}^M \leftarrow$ Random minibatch of M datapoints (drawn from full dataset)
  $\boldsymbol{\epsilon} \leftarrow$ Random samples from noise distribution $p(\boldsymbol{\epsilon})$
  $\mathbf{g} \leftarrow \nabla_{\boldsymbol{\theta},\boldsymbol{\phi}} \widetilde{\mathcal{L}}^M(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{X}^M, \boldsymbol{\epsilon})$ (Gradients of minibatch estimator (8))
  $\boldsymbol{\theta}, \boldsymbol{\phi} \leftarrow$ Update parameters using gradients $\mathbf{g}$ (e.g. SGD or Adagrad [DHS10])
until convergence of parameters $(\boldsymbol{\theta}, \boldsymbol{\phi})$
return $\boldsymbol{\theta}, \boldsymbol{\phi}$

Explainability gets increasingly important …

Deep Probabilistic Programming Languages: A Qualitative Study

Guillaume Baudart, IBM Research, guillaume.baudart@ibm.com
Martin Hirzel, IBM Research, hirzel@us.ibm.com
Louis Mandel, IBM Research, lmandel@us.ibm.com

v1 [cs.AI] 17 Apr 2018

ABSTRACT

Deep probabilistic programming languages try to combine the advantages of deep learning with those of probabilistic programming languages. If successful, this would be a big step forward in machine learning and programming languages. Unfortunately, as of now, this new crop of languages is hard to use and understand. This paper addresses this problem directly by explaining deep probabilistic programming languages and indirectly by characterizing their current strengths and weaknesses.

CCS CONCEPTS

- **Theory of computation → Probabilistic computation;**
- **Computing methodologies → Neural networks;**
- **Software and its engineering → Domain specific languages;**

KEYWORDS

DL, PPL, DSL

1 INTRODUCTION

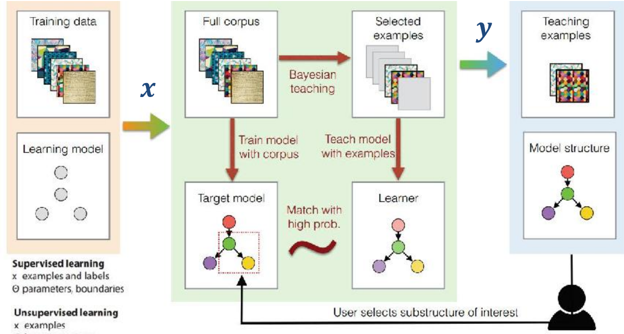
A deep probabilistic programming language (PPL) is a language for specifying both deep neural networks and probabilistic models. In other words, a deep PPL draws upon programming languages,

These frameworks provide automatic differentiation (users need not manually calculate gradients for gradient descent), GPU support (to efficiently execute vectorized computations), and Python-based embedded domain-specific languages [18].

Deep PPLs, which have emerged just recently [29–32], aim to combine the benefits of PPLs and DL. Ideally, programs in deep PPLs would overtly represent uncertainty, yield explainable models, and require only a small amount of training data; be easy to write in a well-designed programming language; and match the breakthrough accuracy and fast training times of DL. Realizing all of these promises would yield tremendous advantages. Unfortunately, this is hard to achieve. Some of the strengths of PPLs and DL are seemingly at odds, such as explainability vs. automated feature engineering, or learning from small data vs. optimizing for large data. Furthermore, the barrier to entry for work in deep PPLs is high, since it requires non-trivial background in fields as diverse as statistics, programming languages, and deep learning. To tackle this problem, this paper characterizes deep PPLs, thus lowering the barrier to entry, providing a programming-languages perspective early when it can make a difference, and shining a light on gaps that the community should try to address.

This paper uses the Stan PPL as a representative of the state of the art in regular (not deep) PPLs [9]. Stan is a main-stream, mature,

Explainable AI via Bayes : $P_T(x|\Theta) \propto P_L(\Theta|x)$



Supervised learning
x examples and labels
Θ parameters, boundaries

Unsupervised learning
x examples
Θ latent structures

Reinforcement learning
x actions, observations, rewards
Θ learned policy & world model

Deep learning
x training examples
Θ network weights

Scott Cheng-Hsin Yang & Patrick Shafto 2017. Explainable Artificial Intelligence via Bayesian Teaching. NIPS 2017 Workshop Machine Teaching. Long Beach (CA).

What are people doing with PPL

x	y
program source code	program output
scene description	image
policy and world	rewards
cognitive process	behavior
simulation	constraint

Image credit to Frank Wood (2016)

Image Interpretation

Vikash K. Mansinghka, Tejas D. Kulkarni, Yura N. Perov & Josh Tenenbaum. Approximate Bayesian image interpretation using generative probabilistic graphics programs. In: Burges, Christopher J. C., Bottou, Leon, Welling, Max, Ghahramani, Zhoubin & Weinberger, Kilian Q., eds. Advances in Neural Information Processing Systems, 2013 Lake Tahoe. NIPS, 1520-1528.

Approximate Bayesian Image Interpretation using Generative Probabilistic Graphics Programs

Vikash K. Mansinghka* [1,2], Tejas D. Kulkarni* [1,2], Yura N. Perov [1,2,3], and Joshua B. Tenenbaum [1,2]

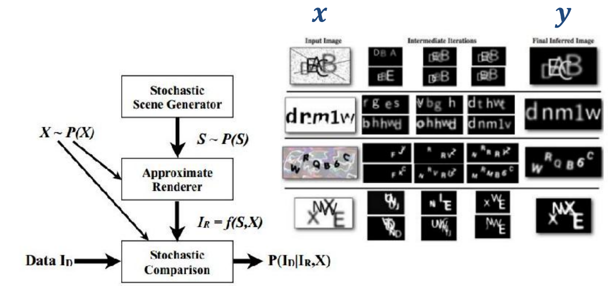
[1] Computer Science and Artificial Intelligence Laboratory, MIT
[2] Department of Brain and Cognitive Sciences, MIT
[3] Institute of Mathematics and Computer Science, Siberian Federal University

Abstract

The idea of computer vision as the Bayesian inverse problem to computer graphics has a long history and an appealing elegance, but it has proved difficult to directly implement. Instead, most vision tasks are approached via complex bottom-up processing pipelines. Here we show that it is possible to write short, simple probabilistic graphics programs that define flexible generative models and to automatically invert them to interpret real-world images. Generative probabilistic graphics programs (GPGP) consist of a stochastic scene generator, a renderer based on graphics software, a stochastic likelihood model linking the renderer's output and the data, and latent variables that adjust the fidelity of the renderer and the tolerance of the likelihood. Representations and algorithms from computer graphics are used as the deterministic backbone for highly approximate and stochastic generative models. This formulation combines probabilistic programming, computer graphics, and approximate Bayesian computation, and depends only on general-purpose, automatic inference techniques. We describe two applications: reading sequences of degraded and adversarially obscured characters, and inferring 3D road models from vehicle-mounted camera images. Each of the probabilistic graphics programs we present relies on under 20 lines of probabilistic code, and

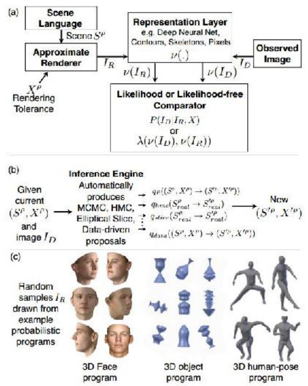
Scene Description -> Image

$$P(S|I_D) \propto \int P(S)P(X)\delta_{f(S,X)}(I_R)P(I_D|I_R, X)dX$$



Vikash K. Mansinghka, Tejas D. Kulkarni, Yura N. Perov & Josh Tenenbaum. Approximate Bayesian image interpretation using generative probabilistic graphics programs. In: Burges, Christopher J. C., Bottou, Leon, Welling, Max, Ghahramani, Zhoubin & Weinberger, Kilian Q., eds. Advances in Neural Information Processing Systems, 2013 Lake Tahoe. NIPS, 1520-1528.

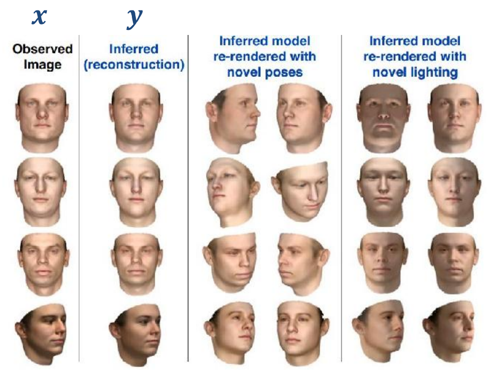
Another Example




```
function PROGRAM(MU, PC, EV, VERTEX_ORDER)
  # Scene Language: Stochastic Scene Gen
  face=Dict();shape = []; texture = [];
  for S in ["shape", "texture"]
   for p in ["nose", "eyes", "outline", "lips"]
    coeff = MvNormal(0,1,1,99)
    face[S][p] = MU[S][p]+PC[S][p].*(coeff.*EV[S][p])
   end
  end
  shape=face["shape"][:]; tex=face["texture"][:];
  camera = Uniform(-1,1,1,2); light = Uniform(-1,1,1,2)

  # Approximate Renderer
  rendered_img= MeshRenderer(shape,tex,light,camera)

  # Representation Layer
  ren_ftrs = getFeatures("CNN_Conv6", rendered_img)

  # Comparator
  #Using Pixel as Summary Statistics
  observe(MvNormal(0,0.01), rendered_img-obs_img)
  #Using CNN last conv layer as Summary Statistics
  observe(MvNormal(0,10), ren_ftrs-obs_cnn)
end

global obs_img = imread("test.png")
global obs_cnn = getFeatures("CNN_Conv6", img)
#Load args from file
TR = trace(PROGRAM,args=[MU,PC,EV,VERTEX_ORDER])
# Data-Driven Learning
learn_datadriven_proposals(TR,100000,"CNN_Conv6")
load_proposals(TR)
# Inference
infer(TR,CB,20,["DATA-DRIVEN"])
infer(TR,CB,200,["ELLIPTICAL"])
```

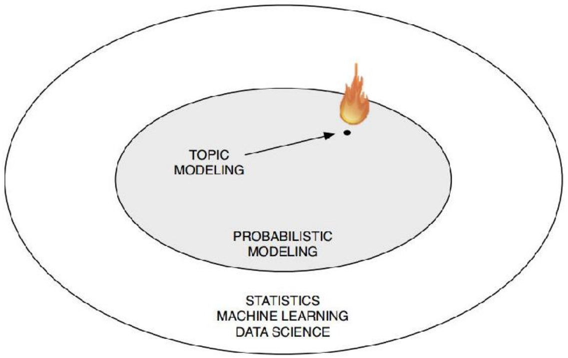
Kulkarni, Kohli, Tenenbaum & Mansinghka. Picture: A probabilistic programming language for scene perception. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015. 4390-4399.

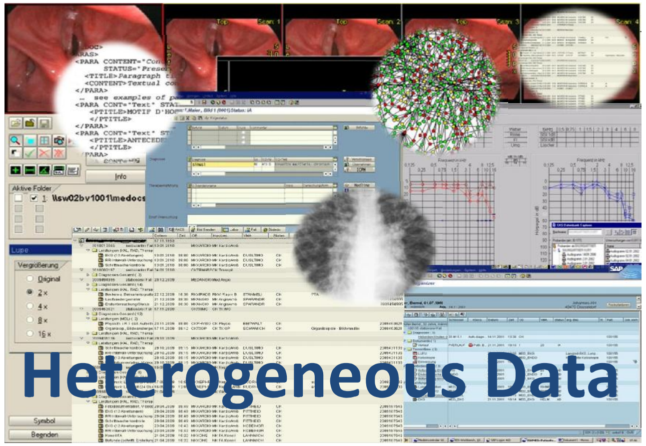
Probabilistic Program for Inference

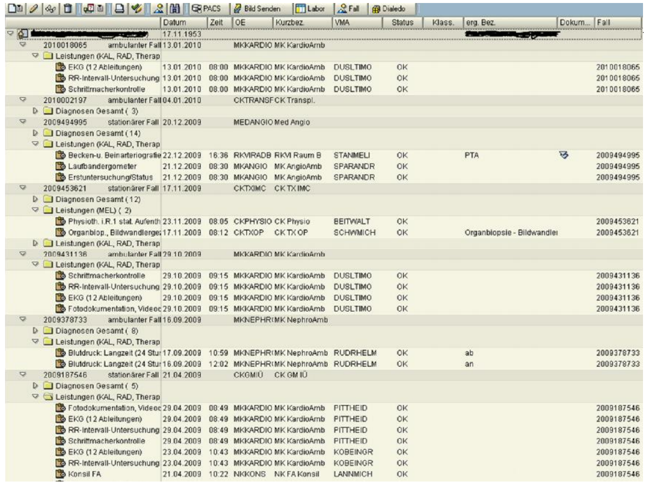


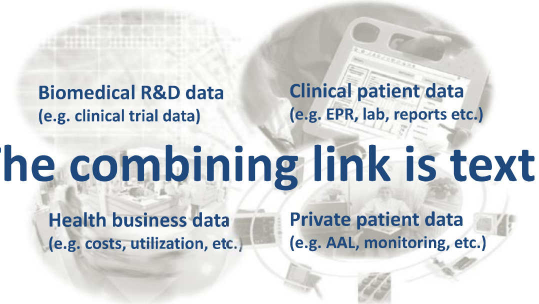
Kulkarni, Kohli, Tenenbaum & Mansinghka. Picture: A probabilistic programming language for scene perception. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015. 4390-4399.

Topic modelling – small topic but hot topic in ML









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

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

The combining link is text

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

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

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

Salton, Wong, Yang, Cornell University 1975

Information Retrieval and Language Processing — C.A. Montgomery, Editor

A Vector Space Model for Automatic Indexing

G. Salton, A. Wong and C. S. Yang
Cornell University

In a document retrieval, or other pattern matching environment where stored entities (documents) are compared with each other or with incoming patterns (search requests), it appears that the best indexing (property) space is one where each entity lies as far away from the others as possible; in these circumstances the value of an indexing system may be expressible as a function of the density of the object space; in particular, retrieval performance may correlate inversely with space density. An approach based on space density computations is used to choose an optimum indexing vocabulary for a collection of documents. Typical evaluation results are shown, demonstating the usefulness of the model.

Key Words and Phrases: automatic information retrieval, automatic indexing, content analysis, document space

CR Categories: 3.71, 3.73, 3.74, 3.75

1. Document Space Configurations

Consider a document space consisting of documents D_i, each identified by one or more index terms T_j; the terms may be weighted according to their importance, or unweighted with weights restricted to 0 and 1.[1] A typical three-dimensional index space is shown in Figure 1, where each item is identified by up to three distinct terms. The three-dimensional example may be extended to t dimensions when t different index terms are present. In that case, each document D_i is represented by a t-dimensional vector

$D_i = (d_{i1}, d_{i2}, \ldots, d_{it})$,

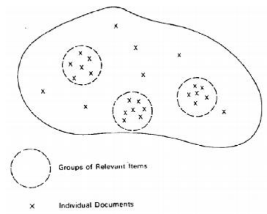
d_{ij} representing the weight of the jth term.

Given the index vectors for two documents, it is possible to compute a similarity coefficient between them, $s(D_i, D_j)$, which reflects the degree of similarity in the corresponding terms and term weights. Such a similarity measure might be the inner product of the two vectors, or alternatively an inverse function of the angle between the corresponding vector pairs; when the term assignment for two vectors is identical, the angle will be zero, producing a maximum similarity measure.

Instead of identifying each document by a complete vector originating at the 0-point in the coordinate system, the relative distance between the vectors is preserved by normalizing all vector lengths to one, and considering the projection of the vectors onto the envelope of the space represented by the unit sphere. In that case, each document may be depicted by a single

Vector representation of document space

$$F = \sum_{i=1}^{n} \sum_{\substack{j=1 \\ i \neq j}}^{n} s(D_i, D_j)$$



Gerard M. Salton, Andrew Wong & Chungshu S. Yang 1975. Vector-Space Model for automatic indexing. Communications of the ACM, 18, (11), 613-620, doi:10.1145/361219.361220.

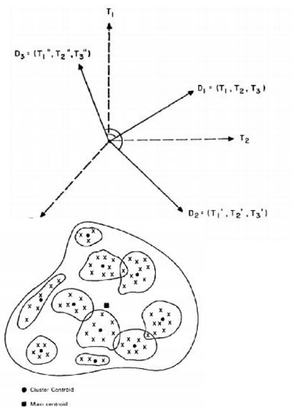


Table I. Effect of Performance Change on Space Density

	(a) Effect of performance improvement on space density				(b) Effect of performance deterioration on space density			
	Cluster organization A (155 clusters; 2.1 overlap)		Cluster organization B (83 clusters; 1.3 overlap)		Cluster organization A (155 clusters; 2.1 overlap)		Cluster organization B (83 clusters; 1.3 overlap)	
Type of indexing	Standard term frequency weights (f_i^k)	Term frequency with inverse doc. freq. ($f_i^k \cdot IDF_k$)	Standard term frequency weights (f_i^k)	Term frequency with inverse doc. freq. ($f_i^k \cdot IDF_k$)	Standard term frequency weights (f_i^k)	Term frequency with document frequency ($f_i^k \cdot DF_k$)	Standard term frequency weights (f_i^k)	Term frequency with document frequency ($f_i^k \cdot DF_k$)
Recall-precision output*	—	+14%	—	+14%	—	−10.1%	—	−10.1%
Average similarity between documents and corresponding cluster centeroids (x)	.712	.668 (−.044)	.650	.589 (−.061)	.712	.741 (+.029)	.650	.696 (+.046)
Average similarity between cluster centroids and main centroid	.500	.454 (−.046)	.537	.492 (−.045)	.500	.555 (+.055)	.537	.574 (+.037)
Average similarity between pairs of cluster centroids (y)	.273	.209 (−.046)	.315	.252 (−.063)	.273	.329 (+.056)	.315	.362 (+.047)
Ratio y/x	.273/.712 = .383	.209/.668 = .318 (−19%)	.315/.650 = .485	.252/.589 = .428 (−12%)	.273/.712 = .383	.329/.741 = .444 (+16%)	.315/.650 = .485	.362/.696 = .520 (+7%)

* From [2].

Gerard M. Salton, Andrew Wong & Chungshu S. Yang 1975. Vector-Space Model for automatic indexing. Communications of the ACM, 18, (11), 613-620, doi:10.1145/361219.361220.

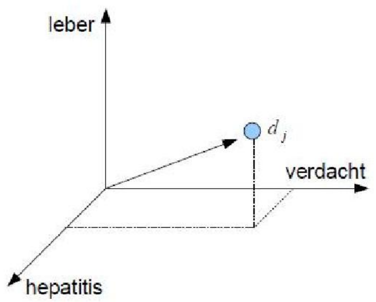
Example (1)

- $D = \langle d_1, d_2, \ldots d_n \rangle$
- $d_i = t_1, t_2, \ldots t_k$
- $w_{i,j} = \begin{cases} 1, & t_i \in d_j \\ 0, & t_i \notin d_j \end{cases} \rightarrow d_j = (0, 1, 1, 0, 1, \ldots, 1)^T$
- $w_{i,j} = \begin{cases} (1 + \log f_{i,j}) * \log \frac{N}{n_i}, & \text{if } f_{i,j} > 0 \\ 0 & \text{otherwise} \end{cases}$

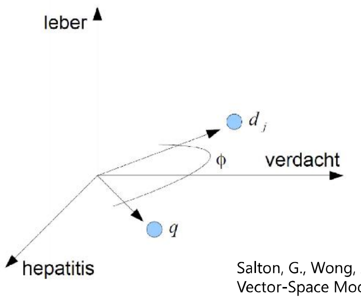
Example (2)

$$D_{m \times n} = \begin{Bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,n-1} & w_{1,n} \\ w_{2,1} & w_{2,2} & & w_{2,n-1} & w_{2,n} \\ \vdots & & \ddots & & \vdots \\ w_{m-1,1} & w_{m-1,2} & & w_{m-1,n} & w_{m-1,n} \\ w_{m,1} & w_{m,2} & \cdots & w_{m,n-1} & w_{m,n} \end{Bmatrix}$$

Example (3)



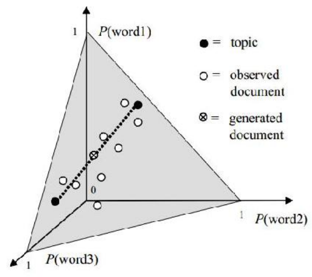
Vector Space Model



$$\cos(\phi) = \frac{q \cdot d_j}{\| q \| \| d_j \|}$$

Salton, G., Wong, A. & Yang, C. S. 1975. Vector-Space Model for automatic indexing. *Communications of the ACM*, 18, (11), 613-620.

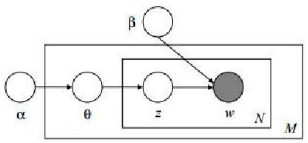
Geometry of Topic Models

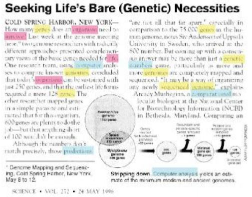


- Documents = categorical distributions over a large space of predefined vocabulary
- Topics = categorical distributions
- Generative model = each document can be seen as a convex combination of the topic distributions

Teh, Y. W., Jordan, M. I., Beal, M. J. & Blei, D. M. 2006. Hierarchical dirichlet processes. Journal of the american statistical association, 101, (476), 1566-1581.

Generative statistical model for natural language



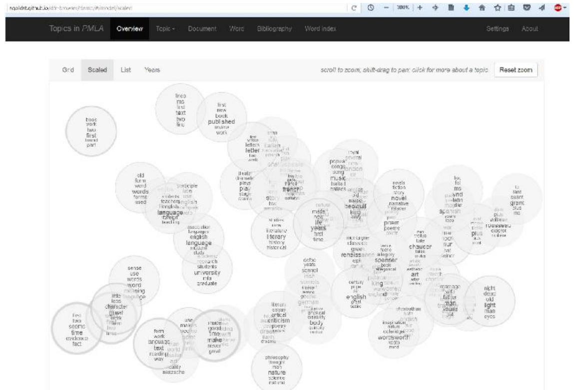


Given the parameters α and β, the joint distribution of a topic mixture θ, a set of N topics z, and a set of N words w is given by:

$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta)$$

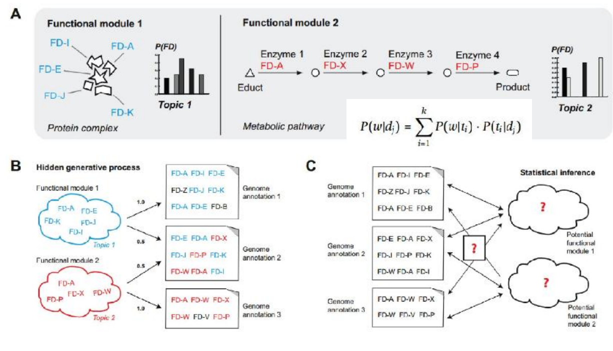
Blei, D. M., Ng, A. Y. & Jordan, M. I. 2003. Latent dirichlet allocation. The Journal of machine Learning research, 3, 993-1022.

Motivation: to get insight into unknown document sets



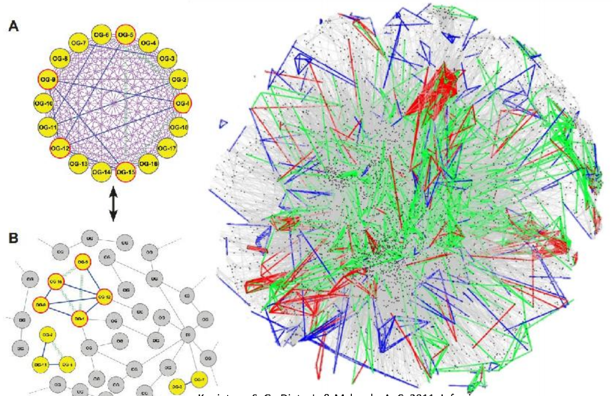
http://agoldst.github.io/dfr-browser/demo/#/model/scaled

Example from Bioinformatics



Konietzny, S. G., Dietz, L. & Mchardy, A. C. 2011. Inferring functional modules of protein families with probabilistic topic models. BMC bioinformatics, 12, (1), 1.

Eval. scheme for inferred potential functional modules

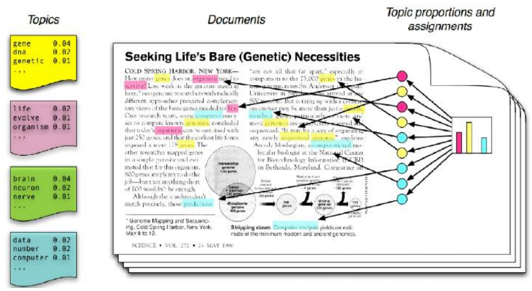


Konietzny, S. G., Dietz, L. & Mchardy, A. C. 2011. Inferring functional modules of protein families with probabilistic topic models. BMC bioinformatics, 12, (1), 1.

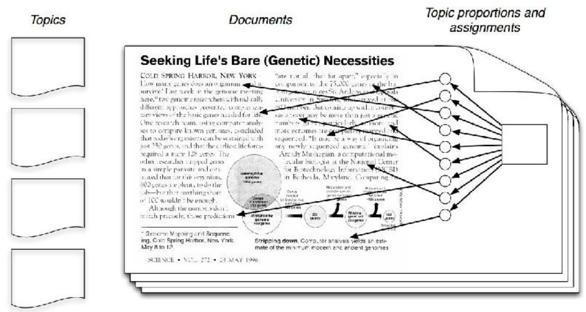
Generative Probabilistic Model

Goal: to get insight in unknown document collections

See a nice demo http://agoldst.github.io/dfr-browser/demo/#/model/grid



Each doc is a random mix of corpus-wide topics and each word is drawn from one of these topics



We only observe the docs – the other structure is hidden; then we compute the posterior p(t,p,a|docs)

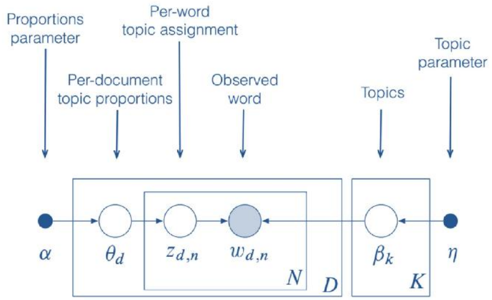
Output Example: 4 learned topics (credit to Blei, 2008)

human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Columns sorte probability word given tc

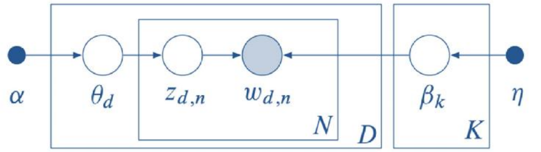
D. Blei

LDA is an example for a probabilistic graphical model



- Encodes assumptions on data with a factorization of the joint
- Connects assumptions to algorithms for computing with data
- Defines the posterior (through the joint)

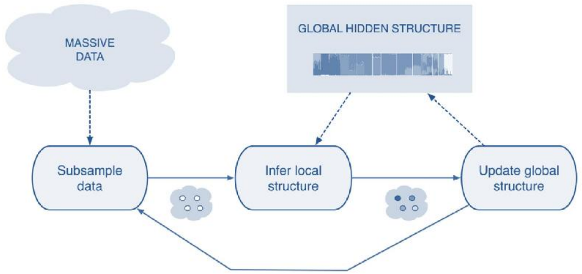
Posterior inference



$$p(\beta, \boldsymbol{\theta}, \mathbf{z} \mid \mathbf{w}) = \frac{p(\beta, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w})}{\int_\beta \int_{\boldsymbol{\theta}} \sum_{\mathbf{z}} p(\beta, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w})}$$

We can't compute the denominator, the marginal $p\ (w)$, therefore we use **approximate inference**; However, this do not scale well …

For "big data" stochastic variational inference



Hoffman, M. D., Blei, D. M., Wang, C. & Paisley, J. 2013. Stochastic variational inference. The Journal of Machine Learning Research, 14, (1), 1303-1347.

Stochastic variational inference

Initialize $\lambda^{(0)}$ randomly.
Set the step-size schedule ρ_t appropriately.
repeat
  Sample a document w_d uniformly from the data set.
  Initialize $\gamma_{dk} = 1$, for $k \in \{1, \ldots, K\}$.
  repeat
    For $n \in \{1, \ldots, N\}$ set

$$\phi_{dn}^k \propto \exp\{\mathbb{E}[\log \theta_{dk}] + \mathbb{E}[\log \beta_{k,w_{dn}}]\},\ k \in \{1, \ldots, K\}.$$

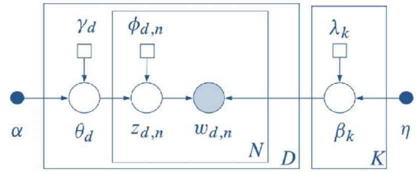
    Set $\gamma_d = \alpha + \sum_n \phi_{dn}$.
  until local parameters ϕ_{dn} and γ_d converge.
  For $k \in \{1, \ldots, K\}$ set intermediate topics

$$\hat{\lambda}_k = \eta + D \sum_{n=1}^{N} \phi_{dn}^k w_{dn}.$$

  Set $\lambda^{(t)} = (1 - \rho_t)\lambda^{(t-1)} + \rho_t \hat{\lambda}$.
until forever

Hoffman, M. D., Blei, D. M., Wang, C. & Paisley, J. 2013. Stochastic variational inference. The Journal of Machine Learning Research, 14, (1), 1303-1347.

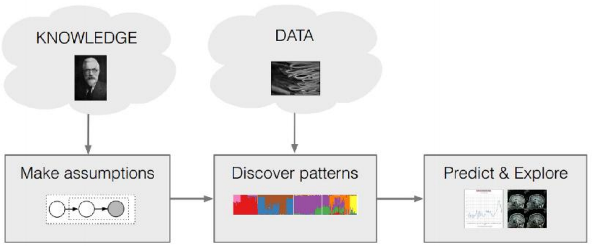
Stochastic variational inference in LDA



1. Sample a document
2. Estimate the local variational parameters using the current topics
3. Form intermediate topics from those local parameters
4. Update topics as a weighted average of intermediate and current topics

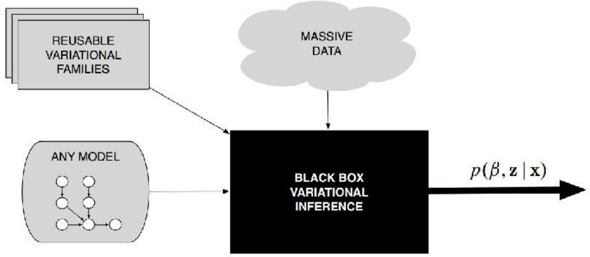
Hoffman, M. D., Blei, D. M., Wang, C. & Paisley, J. 2013. Stochastic variational inference. The Journal of Machine Learning Research, 14, (1), 1303-1347.

Approximate inference can be difficult to achieve



Hoffman, M. D., Blei, D. M., Wang, C. & Paisley, J. 2013. Stochastic variational inference. The Journal of Machine Learning Research, 14, (1), 1303-1347.

Black Box Approach



Hoffman, M. D., Blei, D. M., Wang, C. & Paisley, J. 2013. Stochastic variational inference. The Journal of Machine Learning Research, 14, (1), 1303-1347.

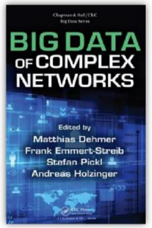
Conclusion: What is needed …

- Flexible and expressive components for building models
- Scalable and generic inference algorithms
- Easy to use software to stretch probabilistic modeling into the health domain
- Topic models are only one approach towards detection of topics in text collections
- More general: Identify re-occurring patterns in data collections generally …
- Much open work for you in the future ☺

Please look at some toolkits

- Particular topic models
  - Stanford topic model toolbox http://nlp.stanford.edu/software/tmt
  - Topic modeling at Princeton http://www.cs.princeton.edu/~blei/topicmodeling.html
  - MALLET (Java) http://mallet.cs.umass.edu
  - Network topic models: Bayes-stack https://github.com/bgamari/bayes-stack
  - Gensim (Python) http://radimrehurek.com/gensim/
  - R package for Topic models. http://epub.wu.ac.at/3987/
- Frameworks for generative models
  - Variational inference: Infer.net http://research.microsoft.com/infernet/
  - Gibbs sampling: OpenBUGS http://openbugs.net/

Dehmer, M., Emmert-Streib, F., Pickl, S. & Holzinger, A. (eds.) 2016. Big Data of Complex Networks, Boca Raton, London, New York: CRC Press Taylor & Francis Group.



04 Knowledge Representation in Network Medicine

Network of Networks in Biology

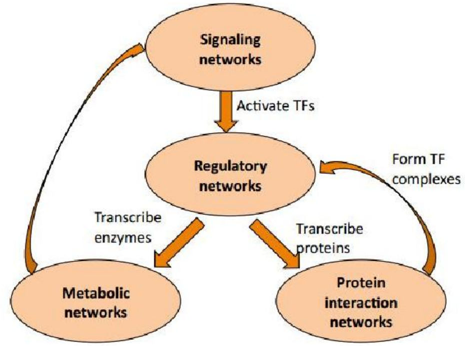


Image credit to Anna Goldenberg, Toronto

Genome-Phenome association in complex diseases

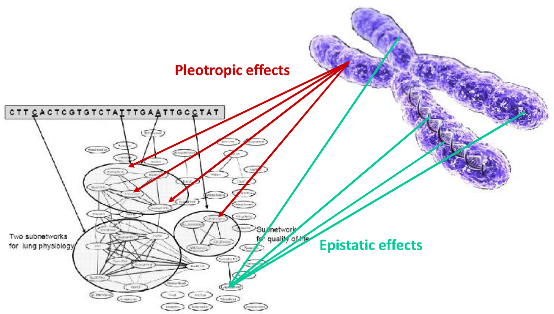


Image credit to Eric Xing, Carnegie Mellon University, Pittsburgh

From data sets to networks

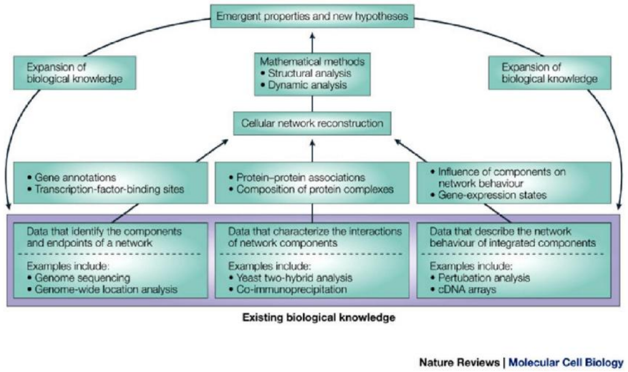


Image description find here: http://www.nature.com/nrm/journal/v6/n2/fig_tab/nrm1570_F1.html

Regulatory>Metabolic>Signaling>Protein>Co-expression

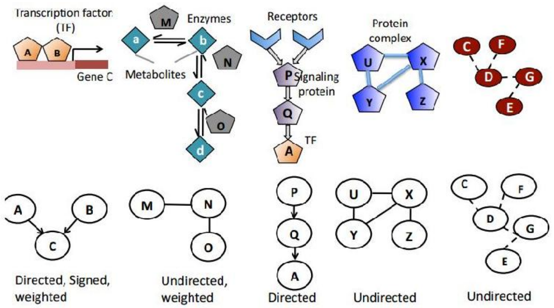
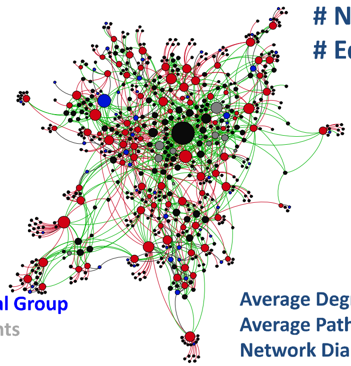


Image credit to Anna Goldenberg, Toronto

Example for a Medical Knowledge Space



 # Nodes: 641
 # Edges: 1250

Agent
Condition
Pharmacological Group
Other Documents

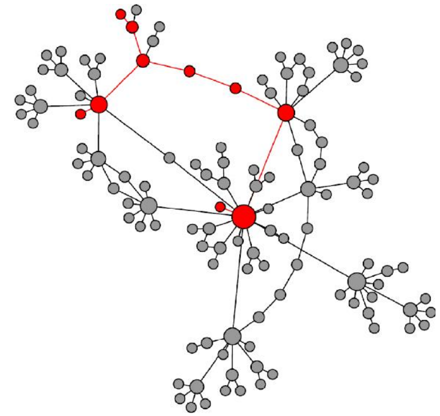
Average Degree: 3.888
Average Path Length: 4.683
Network Diameter: 9

Holzinger, A., Ofner, B., Dehmer, M.: Multi-touch Graph-Based Interaction for Knowledge Discovery on Mobile Devices: State-of-the-Art and Future Challenges. In: LNCS 8401, pp. 241–254, (2014)

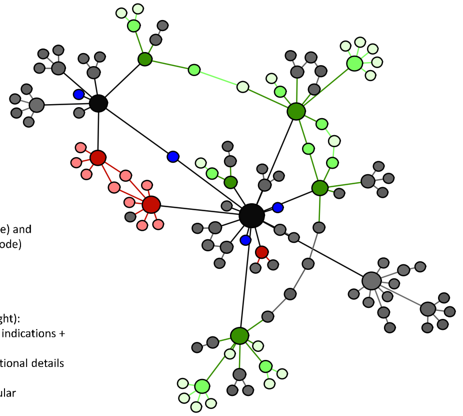
Medical Details of the Graph

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

Example for the shortest path



Example for finding related structures

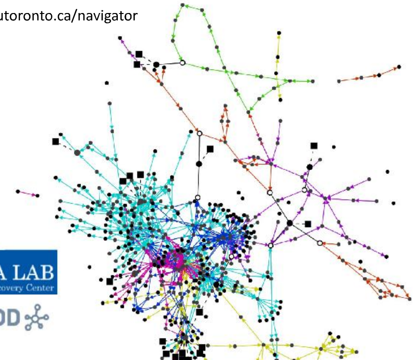


Relationship between
Adrenaline (center black node) and
Dobutamine (top left black node)
Blue: Pharmacological Group
Dark red: Contraindication;
Light red: Condition

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

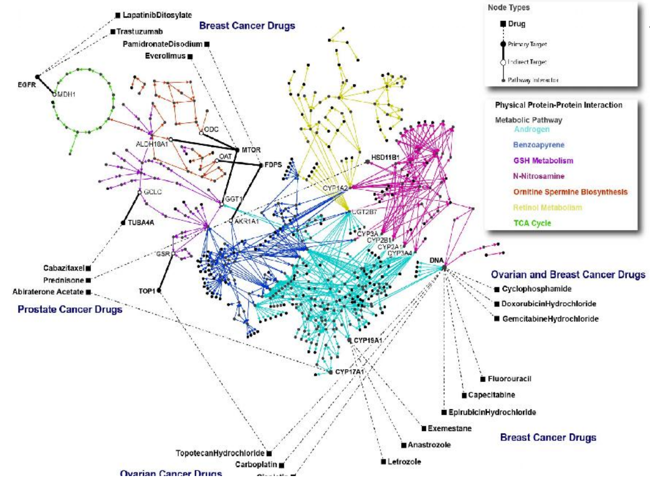
Interactive Visual Data Mining

http://ophid.utoronto.ca/navigator

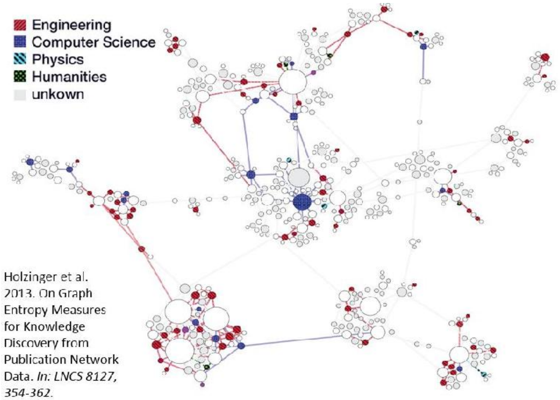


JURISICA LAB
IBM Life Sciences Discovery Center

Otasek, D., Pastrello, C., Holzinger, A. & Jurisica, I. 2014. Visual Data Mining: Effective Exploration of the Biological Universe. In: Holzinger, A. & Jurisica, I. (eds.) Interactive Knowledge Discovery and Data Mining in Biomedical Informatics: State-of-the-Art and Future Challenges. Lecture Notes in Computer Science LNCS 8401. Heidelberg, Berlin: Springer, pp. 19–34, doi:10.1007/978-3-662-43968-5_2.



Example: Graph Entropy Measures



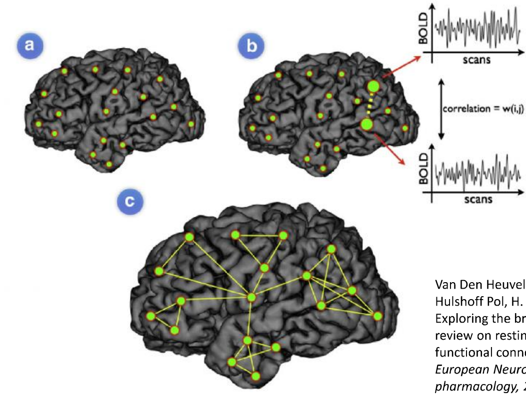
Holzinger et al. 2013. On Graph Entropy Measures for Knowledge Discovery from Publication Network Data. *In: LNCS 8127, 354-362.*

Some selected open problems

- **Problem:** What is the max. number of edges of an Relative Neighborhood Graph in R3 ? No supra-linear lower bound is known.
- **Problem:** What is the structural interpretation of graph measures ? They are mappings which maps graphs to the reals. Thus, they can be understood as graph complexity measures and investigating their structural interpretation relates to understand what kind of structural complexity they detect.
- **Problem:** It is important to visualize large networks meaningfully. So far, there has been a lack of interest to develop efficient software beyond the available commercial software.
- **Problem:** Are multi-touch interaction graphs structurally similar to other graphs (from known graph classes)? This calls for a comparison of graph classes and their structural characteristics.
- **Problem:** Which graph measures are suitable to determine the complexity of multi-touch interaction graphs? Does this lead to any meaningful classification based on their topology?
- **Problem:** What is interesting? Where to start the interaction?

Holzinger, A., Ofner, B., & Dehmer, M. (2014). Multi-touch Graph-Based Interaction for Knowledge Discovery on Mobile Devices: State-of-the-Art and Future Challenges. LNCS 8401 (pp. 241–254). Berlin, Heidelberg: Springer.

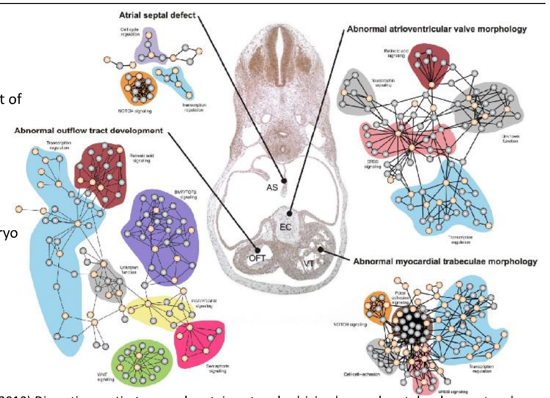
Example: The brain is a complex network



Van Den Heuvel, M. P. & Hulshoff Pol, H. E. (2010) Exploring the brain network: a review on resting-state fMRI functional connectivity. *European Neuropsycho-pharmacology, 20, 8, 519-534.*

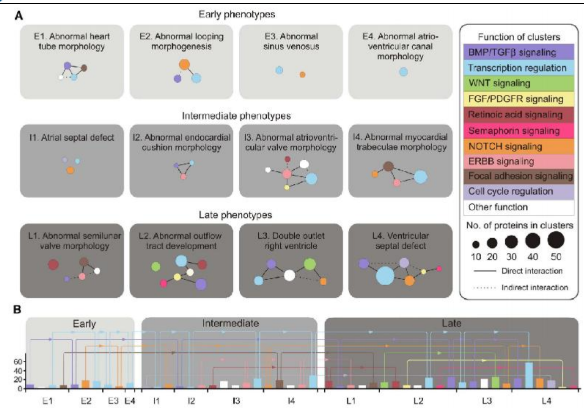
Representative Examples of disease complexes

Examples of 4 functional networks driving the development of different anatomical structures in the human heart of a 37-day old human embryo



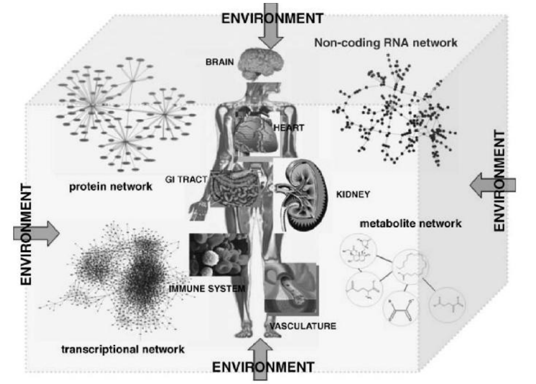
Lage, K. et. al (2010) Dissecting spatio-temporal protein networks driving human heart development and related disorders. *Molecular systems biology, 6, 1, 1-9.*

Example: Cell-based therapy



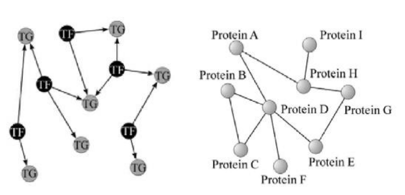
Lage et. al (2010)

Identifying Networks in Disease Research



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

Three main types of biomedical networks



Transcriptional regulatory network with two components:
TF = transcription factor
TG = target genes
(TF regulates the transcription of TG)

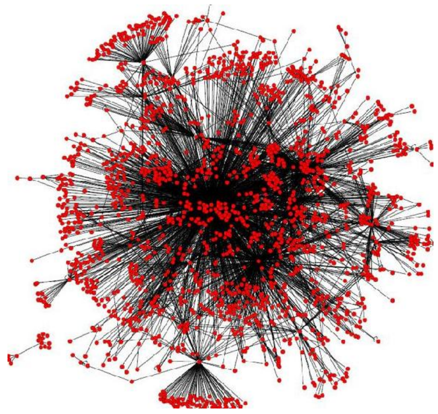
Protein-Protein interaction network

Metabolic network (constructed considering the reactants, chemical reactions and enzymes)

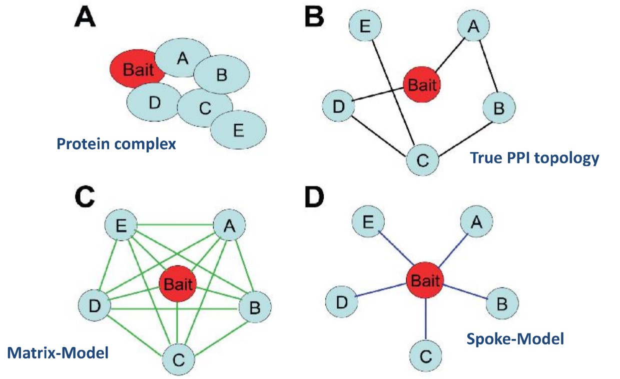
Costa, L. F., Rodrigues, F. A. & Cristino, A. S. (2008) Complex networks: the key to systems biology. *Genetics and Molecular Biology, 31, 3, 591–601.*

Example Transcriptional Regulatory Network

Salgado, H., Santos-Zavaleta, A., Gama-Castro, S., Peralta-Gil, M., Peñaloza-Spínola, M. I., Martínez-Antonio, A., Karp, P. D. & Collado-Vides, J. 2006. The comprehensive updated regulatory network of Escherichia coli K-12. *BMC bioinformatics, 7, (1), 5.*

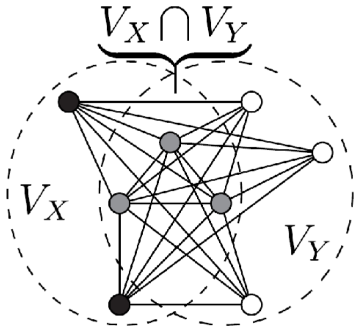


Network Representations of Protein Complexes



Wang, Z. & Zhang, J. Z. (2007) In search of the biological significance of modular structures in protein networks. *PLoS Computational Biology, 3, 6, 1011-1021.*

Correlated Motif Mining (CMM)



Boyen, P., Van Dyck, D., Neven, F., van Ham, R. C. H. J. & van Dijk, A. (2011) SLIDER: A Generic Metaheuristic for the Discovery of Correlated Motifs in Protein-Protein Interaction Networks. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on, 8, 5, 1344-1357.*

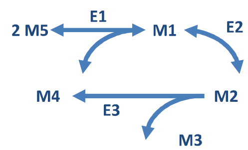
Steepest Ascent Algorithm applied to CMM

Input: PPI-network $G = (V, E, \lambda)$, $\ell, d \in \mathbb{N}$, $d < \ell$
Output: $\{X^*, Y^*\}$ best correlated motif pair found in G

$\{X^*, Y^*\} \leftarrow \mathrm{randomMotifPair}()$
$maxsup \leftarrow f(\{X^*, Y^*\}, G)$
$sup \leftarrow -\infty$
while $maxsup > sup$ **do**
  $\{X, Y\} \leftarrow \{X^*, Y^*\}$
  $sup \leftarrow maxsup$
  for all $\{X', Y'\} \in N(\{X, Y\})$ **do**
    if $f(\{X', Y'\}, G) > maxsup$ **then**
      $\{X^*, Y^*\} \leftarrow \{X', Y'\}$
      $maxsup \leftarrow f(\{X', Y'\}, G)$

Boyen et al. (2011)

Metabolic Network



	M1	M2	M3	M4	M5
M1	0	1	0	1	1
M2	1	0	1	1	0
M3	0	0	0	0	0
M4	1	0	0	0	0
M5	1	0	0	0	0

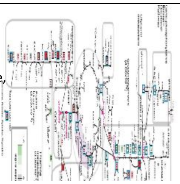
M1	M2
M1	M4
M1	M5
M2	M1
M2	M3
M2	M4
M4	M1
M5	M1

Matrix contains many sparse elements - In this case it is computationally more efficient to represent the graph as an adjacency list

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

Metabolic networks are usually big … big data ☺

Schmid, A. K., Reiss, D. J., Pan, M., Koide, T. & Baliga, N. S. (2009) A single transcription factor regulates evolutionarily diverse but functionally linked metabolic pathways in response to nutrient availability. *Molecular Systems Biology, 5, 1-9.*

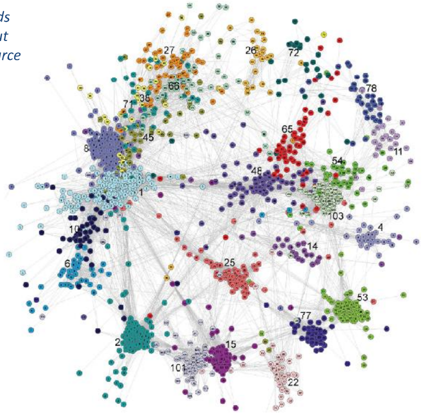


http://www.nature.com/msb/journal/v5/n1/fig_tab/msb200940_F6.html

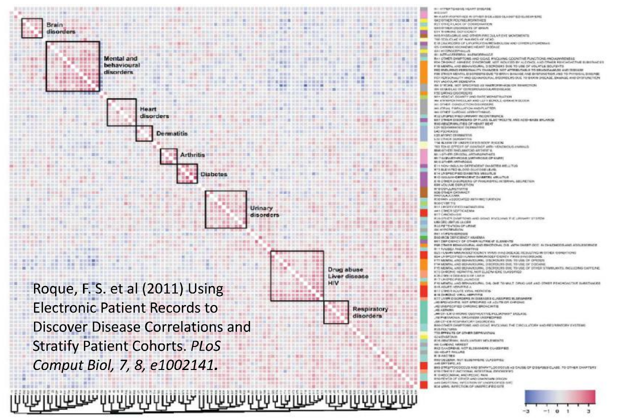
Using EPRs to Discover Disease Correlations

Electronic patient records remain a unexplored, but potentially rich data source for example to discover correlations between diseases.

Roque, F. S., Jensen, P. B., Schmock, H., Dalgaard, M., Andreatta, M., Hansen, T., Søeby, K., Bredkjær, S., Juul, A., Werge, T., Jensen, L. J. & Brunak, S. (2011) Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts. *PLoS Computational Biology, 7, 8, e1002141.*



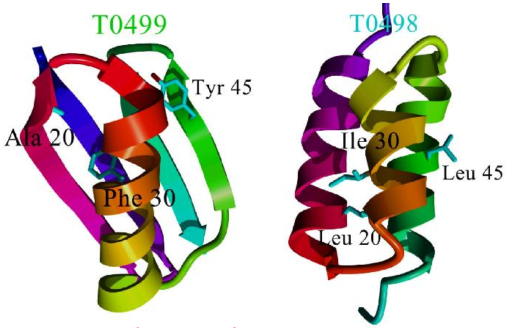
Heatmap of disease-disease correlations (ICD)



Roque, F. S. et al (2011) Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts. *PLoS Comput Biol, 7, 8, e1002141.*

Example: ὁμολογέω (homologeo)

He, Y., Chen, Y., Alexander, P., Bryan, P. N. & Orban, J. (2008) NMR structures of two designed proteins with high sequence identity but different fold and function. Proceedings of the National Academy of Sciences, 105, 38, 14412.



T0499 TTYKLILNLKQAKEEAIKEAVDAGTAEKYFKLIANAKTVEGVWTYKDEIKTFTVTE

T0498 TTYKLILNLKQAKEEAIKELVDAGTAEKYIKLIANAKTVEGVWTLKDEIKTFTVTE

Conclusion

- Homology modeling is a knowledge-based prediction of protein structures.
- In homology modeling a protein sequence with an unknown structure (the target) is aligned with one or more protein sequences with known structures (the templates).
- The method is based on the principle that homologue proteins have similar structures.
- **Homology modeling will be extremely important to personalized and molecular medicine in the future.**

05 Machine Learning on Graphs Relevant for Health Informatics

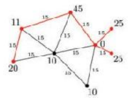
Example: Lymphoma is the most common blood cancer

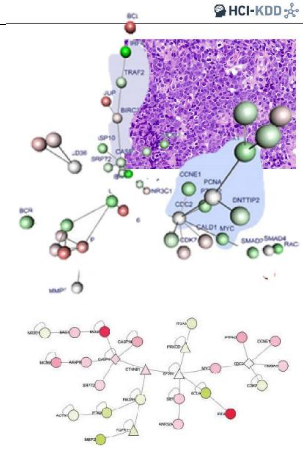
The two main forms of lymphoma are Hodgkin lymphoma and non-Hodgkin lymphoma (NHL). Lymphoma occurs when cells of the immune system called lymphocytes, a type of white blood cell, grow and multiply uncontrollably. Cancerous lymphocytes can travel to many parts of the body, including the lymph nodes, spleen, bone marrow, blood, or other organs, and form a mass called a tumor. The body has two main types of lymphocytes that can develop into lymphomas: B-lymphocytes (B-cells) and T-lymphocytes (T-cells).

www.lymphoma.org http://imagebank.hematology.org/

ML tasks on graphs

- Discover unexplored interactions in PPI-networks and gene regulatory networks
- Learn the structure
- Reconstruct the structure





Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T. & Müller, T. 2008. Identifying functional modules in protein–protein interaction networks: an integrated exact approach. Bioinformatics, 24, (13), i223-i231.

SCIENTIFIC REPORTS

OPEN

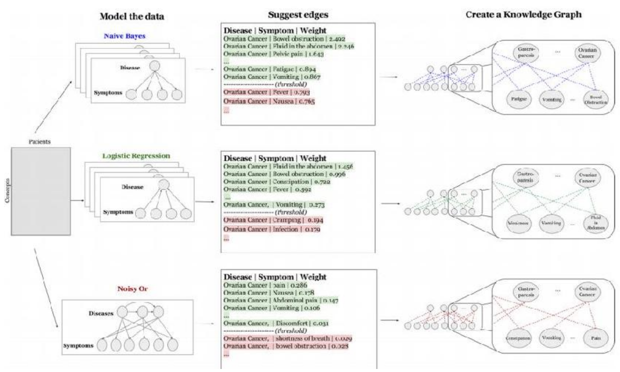
Learning a Health Knowledge Graph from Electronic Medical Records

Received: 3 March 2017
Accepted: 1 June 2017
Published online: 20 July 2017

Maya Rotmensch[1], Yoni Halpern[2], Abdulhakim Tlimat[3], Steven Horng[3,4] & David Sontag[5,6]

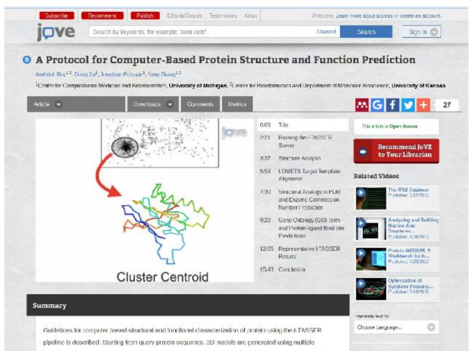
Demand for clinical decision support systems in medicine and self-diagnostic symptom checkers has substantially increased in recent years. Existing platforms rely on knowledge bases manually compiled through a labor-intensive process or automatically derived using simple pairwise statistics. This study explored an automated process to learn high quality knowledge bases linking diseases and symptoms directly from electronic medical records. Medical concepts were extracted from 273,174 de-identified patient records and maximum likelihood estimation of three probabilistic models was used to automatically construct knowledge graphs: logistic regression, naive Bayes classifier and a Bayesian network using noisy OR gates. A graph of disease-symptom relationships was elicited from the learned parameters and the constructed knowledge graphs were evaluated and validated, with permission, against Google's manually-constructed knowledge graph and against expert physician opinions. Our study shows that direct and automated construction of high quality health knowledge graphs from medical records using rudimentary concept extraction is feasible. The noisy OR model produces a high quality knowledge graph reaching precision of 0.85 for a recall of 0.6 in the clinical evaluation. Noisy OR significantly outperforms all tested models across evaluation frameworks ($p < 0.01$).

Workflow for modeling relationship disease-symptom



Maya Rotmensch, Yoni Halpern, Abdulhakim Tlimat, Steven Horng & David Sontag 2017. Learning a Health Knowledge Graph from Electronic Medical Records. Scientific Reports, 7, 5994, doi:10.1038/s41598-017-05778-z.

From structure to function



http://www.jove.com/video/3259/a-protocol-for-computer-based-protein-structure-function

Interesting: Hubs tend to link to small degree nodes

Nodes: proteins
Links: physical interactions (binding)

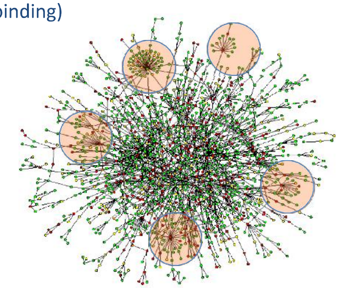
Puzzling pattern:
Hubs tend to link to small degree nodes.
Why is this puzzling?

In a random network, the probability that a node with degree k links to a node with degree k' is:

$$p_{kk'} = \frac{kk'}{2L}$$

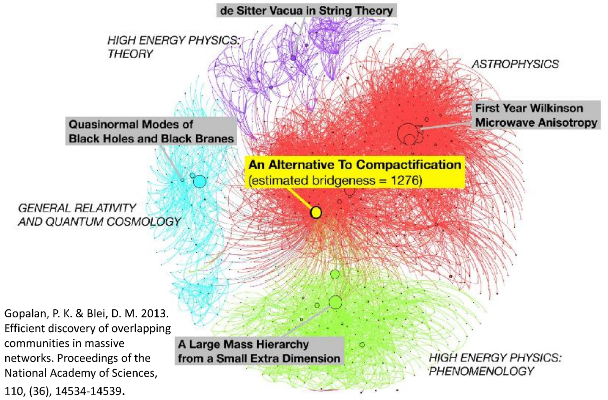
k=50, k'=13, N=1,458, L=1746

$p_{50,13} = 0.15$ $p_{2,1} = 0.0004$



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

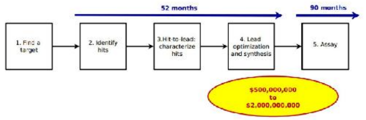
Example: Subgraph Discovery



Gopalan, P. K. & Blei, D. M. 2013. Efficient discovery of overlapping communities in massive networks. Proceedings of the National Academy of Sciences, 110, (36), 14534-14539.

Why do we want to apply ML to graphs

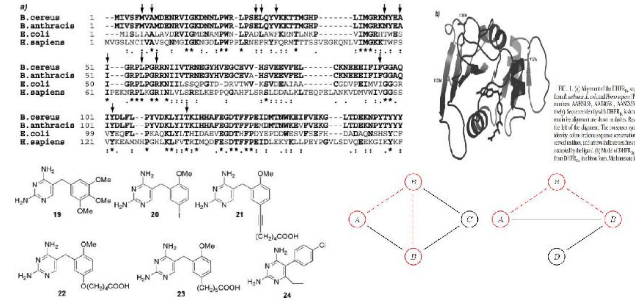
- A) Discovery of unexplored interactions
- B) Learning and Predicting the structure
- C) Reconstructing the structure
- Which joint probability distributions does a graphical model represent?
- How can we learn the parameters and structure of a graphical model?



The chemical space

- 10^{60} possible small organic molecules
- 10^{22} stars in the observable universe

Example Question: Predicting Function from Structure



How similar are two graphs? How similar is their structure? How similar are their node and edge labels?

Joska, T. M. & Anderson, A. C. 2006. Structure-activity relationships of Bacillus cereus and Bacillus anthracis dihydrofolate reductase: toward the identification of new potent drug leads. Antimicrobial agents and chemotherapy, 50, 3435-3443.

Graph Comparison

- Similar Property Principle: Molecules having similar structures should have similar activities.
- Structure-based representations: Compare molecules by comparing substructures, e.g.
  - Sets as vectors: Measure similarity by the cosine distance
  - Sets as sets: Measure similarity by the Jaccard distance
  - Sets as points: Measure similarity by Euclidean distance
- Problems: Dimensionality, Non-Euclidean cases

Thank you!

Questions

Sample Questions (1/3)

- Describe the clinical decision making process!
- Which type of graph is particularly useful for inference and learning?
- What is the key challenge in the application of graphical models for health informatics?
- What was Judea Pearl (1988) discussing in his paper, for which he received the Turing award?
- What main difficulties arise during breast cancer prognosis?
- What can be done to increase the robustness of prognostic cancer tests?
- Inference in Bayes Nets is NP-complete, but there are certain cases where it is tractable, which ones?

Sample Questions (2/3)

- Why do we want to apply ML to graphs?
- Describe typical ML tasks on the example of blood cancer cells!
- If you have a set of points – which similarity measures are useful?
- Why is graph comparison in the medical domain useful?
- Why is the Gromov-Hausdorff distance useful?
- What is the central goal of a generative probabilistic model?
- Describe the LDA-model and its application for topic modelling!

Sample Questions (2/3)

- Briefly describe the stochastic variational inference algorithms!
- What is the principle of a bandit?
- How does a multi-armed bandit (MAB) work?
- In which ways can a MAB represent knowledge?
- What is the main problem of a clinical trail – and maybe the main problem in clinical medicine?
- Why are rare diseases both important and relevant? Describe an example disease!
- What is the big problem in clinical trials for rare diseases?
- What did Richard Bellman (1956) describe with dynamic programming?
- Why are graph bandits a hot topic for ML research?

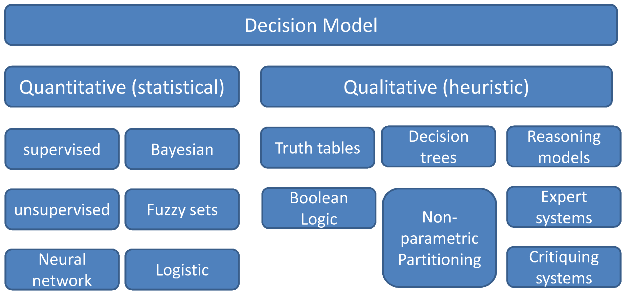
Solutions of the Quiz

- 1=this is a factor graph of an undirected graph – we have seen this in protein networks (refer to slide Nr. 70 in lecture 5). Factor graph is bipartite and has two types of nodes: Variables, which can be either evidence variables (when we know its value) or query variables (when the value is unknown and we want to predict the value); and factors, which define the relationship between variables in the graph. Each factor can be connected to many variables and comes with a factor function to define the relationship between these variables. For example, if a factor node is connected to two variables nodes A and B, a possible factor function could be imply(A,B), meaning that if the random variable A takes value 1, then so must the random variable B. Each factor function has a weight associated with it, which describes how much influence the factor has on its variables in relative terms. For more information please consult: http://deepdive.stanford.edu/inference
- 2= this is the decomposition of a tree, rooted at nodes into subtrees
- 3= an example for machine translation, Image credit to Kevin Gimpel, Carnegie Mellon University
- 4= the famous expectation-utility theory according to von Neumann and Morgenstern (1954): a decision-maker faced with risky (probabilistic) outcomes of different choices will behave as if he is maximizing the expected value of some function defined over the potential outcomes at some specified point in the future.
- 5= MYCIN –expert system that used early AI (rule-based) to identify bacteria causing severe infections, such as bacteremia and meningitis, and to recommend antibiotics, with the dosage adjusted for patient's body weight — the name derived from the antibiotics themselves, as many antibiotics have the suffix "-mycin".
- 6= metabolic and physical processes that determine the physiological and biochemical properties of a cell. These networks comprise the chemical reactions of metabolism, the metabolic pathways, as well as the regulatory interactions that guide these reactions.
- 7= With the sequencing of complete genomes, it is now possible to reconstruct the network of biochemical reactions in many organisms, from bacteria to human. Several of these networks are available online, e.g. Kyoto Encyclopedia of Genes and Genomes (KEGG), EcoCyc, BioCyc etc. Metabolic networks are powerful tools for studying and modelling metabolism.

Appendix

1) Reasoning under Uncertainty

Remember: Taxonomy of Decision Support Models



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

Dealing with uncertainty in the real world

- The information available to humans is often imperfect – imprecise - uncertain.
- This is especially in the medical domain the case.
- An **human agent** can cope with deficiencies.
- Classical logic permits only **exact reasoning**:
- IF A is true THEN A is non-false and
  IF B is false THEN B is non-true
- Most real-world problems do not provide this exact information, mostly it is inexact, incomplete, uncertain and/or **un-measurable!**

MYCIN – rule based system - certainty factors

- MYCIN is a rule-based Expert System, which is used for therapy planning for patients with bacterial infections
- Goal oriented strategy ("Rückwärtsverkettung")
- To every rule and every entry a certainty factor (CF) is assigned, which is between 0 und 1
- Two measures are derived:
- MB: measure of belief
- MD: measure of disbelief
- Certainty factor – CF of an element is calculated by:
  $CF[h] = MB[h] - MD[h]$
- CF is positive, if more evidence is given for a hypothesis, otherwise CF is negative
- CF[h] = +1 -> h is 100 % true
- CF[h] = –1 -> h is 100% false

Original Example from MYCIN

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

$CF[h_1,E] = .8$: There is strongly suggestive evidence (.8) that the identity of ORGANISM-1 is streptococcus

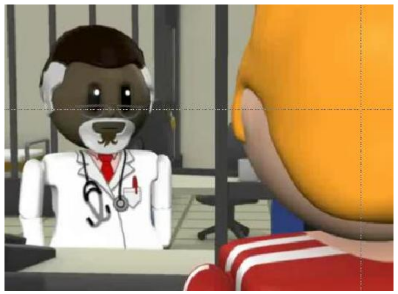
$CF[h_2,E] = -.3$: There is weakly suggestive evidence (.3) that PATIENT-1 is not febrile

$CF[h_3,E] = +1$: It is definite (1) that the name of PATIENT-1 is John Jones

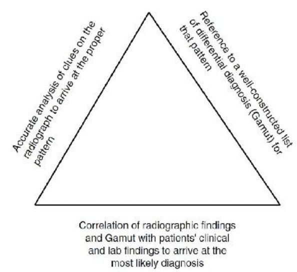
Shortliffe, E. H. & Buchanan, B. G. (1984) *Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project. Addison-Wesley.*

MYCIN was *no* success in the clinical practice

https://www.youtube.com/watch?v=IVGWM0CKNWA ("real nurse triage")



Gamuts: Triangulation to find diagnoses



Reeder, M. M. & Felson, B. 2003. *Reeder and Felson's gamuts in radiology: comprehensive lists of roentgen differential diagnosis, New York, Springer Verlag.*

Gamut F-137

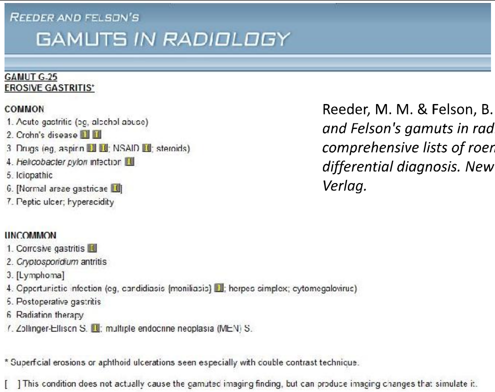
PHRENIC NERVE PARALYSIS OR DYSFUNCTION

COMMON
1. Iatrogenic (eg, surgical injury; chest tube; therapeutic avulsion or injection: subclavian vein puncture)
2. Infection (eg, tuberculosis; fungus disease; abscess)
3. Neoplastic invasion or compression (esp. carcinoma of lung)

UNCOMMON
1. Aneurysm$_g$, aortic or other
2. Birth trauma (Erb's palsy)
3. Herpes zoster
4. Neuritis, peripheral (eg, diabetic neuropathy)
5. Neurologic disease$_g$ (eg, hemiplegia; encephalitis; polio; Guillain-Barré S.)
6. Pneumonia
7. Trauma

Reference
1. Prasad S, Athreya BH: Transient paralysis of the phrenic nerve associated with head injury. JAMA 1976;236:2532–2533

Example - Gamuts in Radiology


REEDER AND FELSON'S
GAMUTS IN RADIOLOGY

GAMUT G-25
EROSIVE GASTRITIS*

COMMON
1. Acute gastritis (eg, alcohol abuse)
2. Crohn's disease
3. Drugs (eg, aspirin; NSAID; steroids)
4. *Helicobacter pylori* infection
5. Idiopathic
6. [Normal areae gastricae]
7. Peptic ulcer; hyperacidity

UNCOMMON
1. Corrosive gastritis
2. *Cryptosporidium* antritis
3. [Lymphoma]
4. Opportunistic infection (eg, candidiasis (moniliasis); herpes simplex; cytomegalovirus)
5. Postoperative gastritis
6. Radiation therapy
7. Zollinger-Ellison S.; multiple endocrine neoplasia (MEN) S.

* Superficial erosions or aphthoid ulcerations seen especially with double contrast technique.

[ ] This condition does not actually cause the gamuted imaging finding, but can produce imaging changes that simulate it.

Reeder, M. M. & Felson, B. (2003) *Reeder and Felson's gamuts in radiology: comprehensive lists of roentgen differential diagnosis.* New York, Springer Verlag.

http://rfs.acr.org/gamuts/data/G-25.htm

Reasoning under uncertainty

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

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

Remember: 2 types of decisions (Diagnosis vs. Therapy)

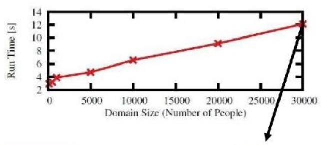
- **Type 1 Decisions:** related to the **diagnosis**, i.e. computers are used to assist in diagnosing a disease on the basis of the individual patient data. Questions include:
  - What is the probability that this patient has a myocardial infarction on the basis of given data (patient history, ECG, …)?
  - What is the probability that this patient has acute appendices, given the signs and symptoms concerning abdominal pain?
- **Type 2 Decisions:** related to **therapy**, i.e. computers are used to select the best therapy on the basis of clinical evidence, e.g.:
  - What is the best therapy for patients of age x and risks y, if an obstruction of more than z % is seen in the left coronary artery?
  - What amount of insulin should be prescribed for a patient during the next 5 days, given the blood sugar levels and the amount of insulin taken during the recent weeks?

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

Future Outlook

The future is in integrative ML, i.e. combining relational databases, ontologies and logic with probabilistic reasoning models and statistical learning – and algorithms that have good **scalability**

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



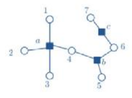
~~Big data~~

Big models

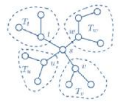
Learns a model over 900,030,000 random variables

Van Den Broeck, G., Taghipour, N., Meert, W., Davis, J. & De Raedt, L. Lifted probabilistic inference by first-order knowledge compilation. Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three, 2011. AAAI Press, 2178-2185.

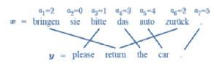
Quiz



1



2



3

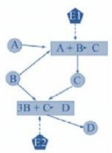
$$E(U \mid d) = \sum_{x_1,\ldots,x_n} P(x_1,\ldots,x_n \mid d)U(x_1,\ldots,x_n,d)$$

4

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

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

5



6

Probabilistic-programming.org

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

PyMC3

PyMC Pythonic Markov chain Monte Carlo

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

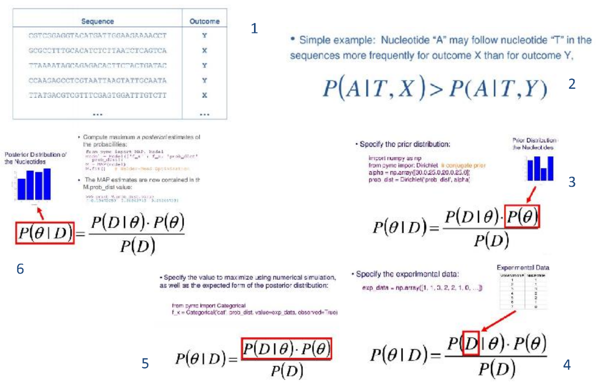
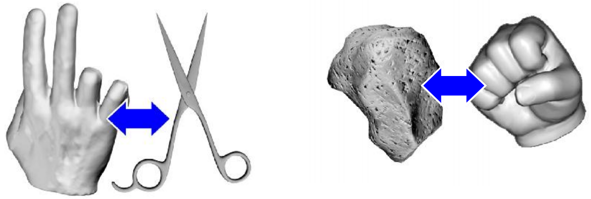


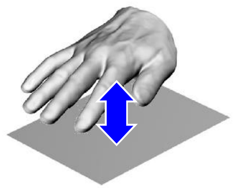
Image Source: Dan Williams, Life Technologies, Austin TX

05 Digression: What is similarity?

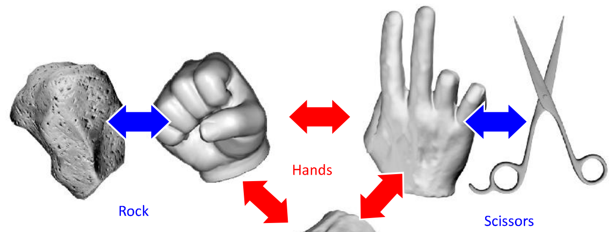
What is Similar?

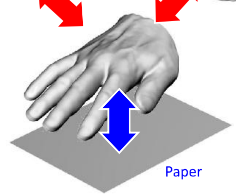
Image credit to Eamonn Keogh (2008)





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



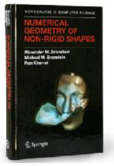


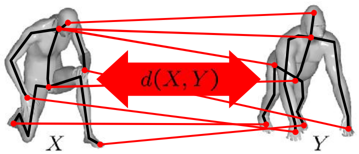
Bronstein, A. M., Bronstein, M. M. & Kimmel, R. 2008. *Numerical geometry of non-rigid shapes, New York, Springer.*

Similarity and Correspondence

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

http://www.inf.usi.ch/bronstein/



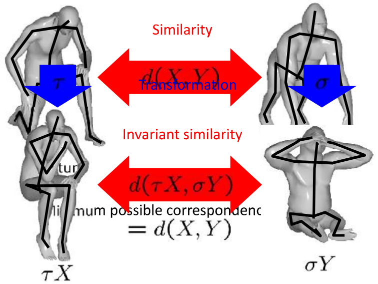


Structure Structure

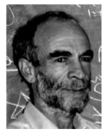
Correspondence quality = structure similarity (distortion)

Minimum possible correspondence distortion

Invariant Similarity



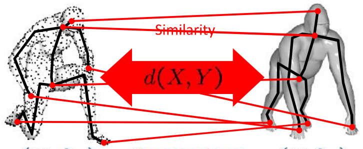
Gromov-Hausdorff dist: finding the opt. correspondence



Michail Gromov (1943-)

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

Felix Hausdorff (1868-1942)



(X, δ_X) Metric space — Correspondence — (Y, δ_Y) Metric space

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

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

Discrete optimization over correspondences is NP hard !

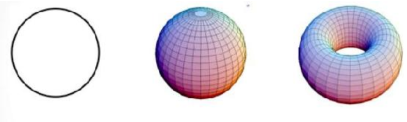
Distinguish topological spaces

Enrico Betti (1823-1892)

Counts the number of "i-dimensional holes"

bi is the "i-th Betti number"

Emmy Noether (1882-1935)



$b_1=1$	$b_1=0$	$b_1=2$
$b_2=0$	$b_2=1$	$b_2=1$

Betti numbers are computed as dimensions of Boolean vector spaces (E. Noether)

Zomorodian, A. & Carlsson, G. 2005. Computing Persistent Homology. *Discrete & Computational Geometry, 33, (2), 249-274.*

Structural Patterns are often hidden in weakly str. data

- Statement of Vin de Silva (2003), Pomona College:
- Let M be a topological or metric space, known as the *hidden parameter space*;
- let $\mathbb{R}^d$ be a Euclidean space, the *observation space*,
- and let $f: M \longrightarrow \mathbb{R}^d$ be a continuous embedding.
- Furthermore, let $X \subset M$ be a finite set of data points, perhaps the realization of a stochastic process, i.e., a family of random variables $\{X_i, i \in I\}$ defined on a probability space $(\Omega, \mathcal{F}, P)$, and denote $Y = f(X) \subset \mathbb{R}^d$ the images of these points under the mapping f.
- We refer to X as *hidden data*, and Y as the *observed data*.
- M, f and X are unknown, but Y is - so can we identify M?

Topological Data Mining

- Mega Problem: To date none of our known methods, algorithms and tools scale to the massive amount and dimensionalities of data we are confronted in practice;
- we need much more research efforts towards making computational topology successful as a general method for data mining and knowledge discovery

Holzinger, A. 2014. On Topological Data Mining. In: Lecture Notes in Computer Science, LNCS 8401. Berlin Heidelberg: Springer, pp. 331-356, doi:10.1007/978-3-662-43968-5_19.

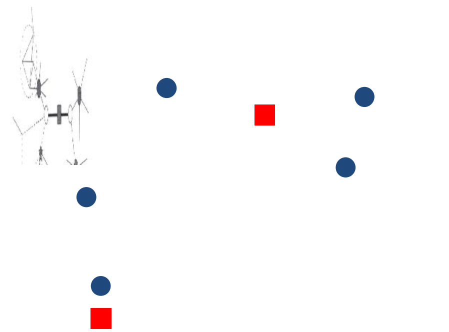
06 Review of basic concepts, metrics and measures

Complex Biological Systems key concepts

- In order to understand complex biological systems, the three following key concepts need to be considered:
- (i) **emergence,** the discovery of links between elements of a system because the study of individual elements such as genes, proteins and metabolites is insufficient to explain the behavior of whole systems;
- (ii) **robustness,** biological systems maintain their main functions even under perturbations imposed by the environment; and
- (iii) **modularity,** vertices sharing similar functions are highly connected.
- Network theory can largely be applied for biomedical informatics, because many tools are already available

Network Basics on the Example of Bioinformatics

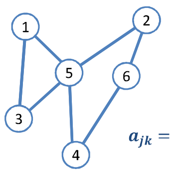
$G(V, E)$ *Graph*
V … *vertex*
E … *edge* $\{a, b\}$
$a, b \in V; a \neq b$



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

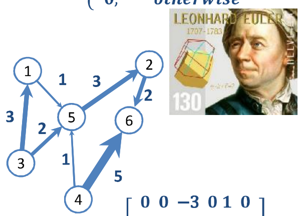
Baby Stuff: Computational Graph Representation

Adjacency (ə-ˈjā-sᵊn(t)-sē) **Matrix** $A = (a_{jk})$ $\quad a_{jk} = \begin{cases} 1, & \text{if } \{j,k\} \in E \\ 0, & \text{otherwise} \end{cases}$



$$a_{jk} = \begin{bmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \end{bmatrix}$$

Simple graph, symmetric, binary

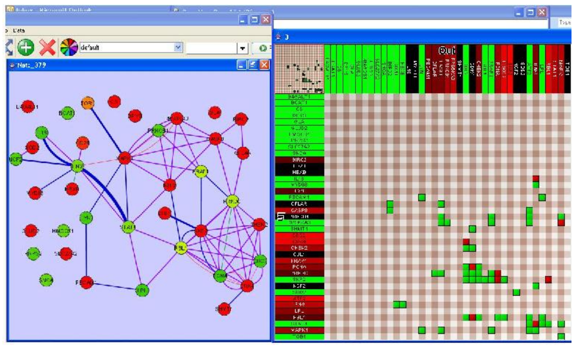


$$w_{jk} = \begin{bmatrix} 0 & 0 & -3 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 \\ 3 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 1 & -5 \\ 1 & -2 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 5 & 0 \end{bmatrix}$$

Directed and weighted

For more information: Diestel, R. (2010) *Graph Theory, 4th Edition. Berlin, Heidelberg, Springer.*

Example: Tool for Node-Link Visualization

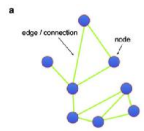


Jean-Daniel Fekete http://wiki.cytoscape.org/InfoVis_Toolkit

Fekete, J.-D. The infovis toolkit. Information Visualization, INFOVIS 2004, 2004. IEEE, 167-174.

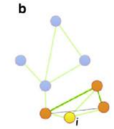
Some Network Metrics (1/2)

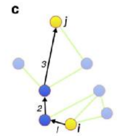
Order = total number of nodes n; **Size** = total number of links (a): $\sum_i \sum_j a_{ij}$



Clustering Coefficient (b) = the degree of concentration of the connections of the node's neighbors in a graph and gives a measure of local inhomogeneity of the link density:

$$C_i = \frac{2t_i}{k(k_i - 1)} \qquad C = \frac{1}{n} \sum_i C_i$$





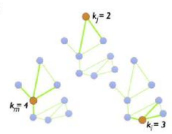
Path length (c) = is the arithmetical mean of all the distances:

$$l = \frac{1}{n(n-1)} \sum_{i \neq j} d_{ij}$$

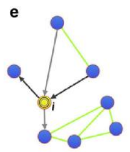
Costa, L. F., Rodrigues, F. A., Travieso, G. & Boas, P. R. V. (2007) Characterization of complex networks: A survey of measurements. *Advances in Physics, 56, 1, 167-242.*

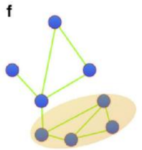
Some Network Metrics (2/2)

- **Centrality** (d) = the level of "betweenness- centrality" of a node I ("hub-node in Slide 28);



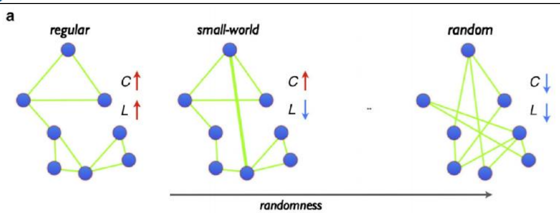
- **Nodal degree** (e) = number of links connecting i to its neighbors: $k_i = \sum_i a_{ij}$

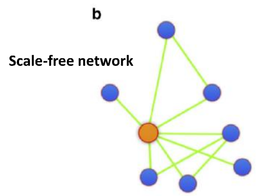




Modularity (f) = describes the possible formation of communities in the network, indicating how strong groups of nodes form relative isolated sub-networks within the full network (refer also to Slide 5-8).

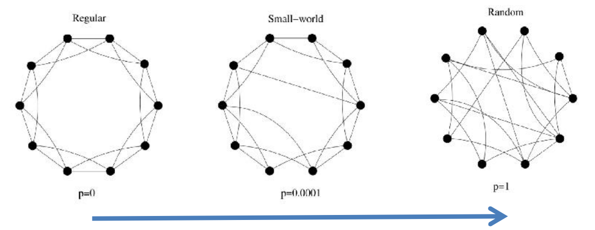
Network Topologies





Van Heuvel & Hulshoff (2010)

Small-World Networks



Increasing randomness

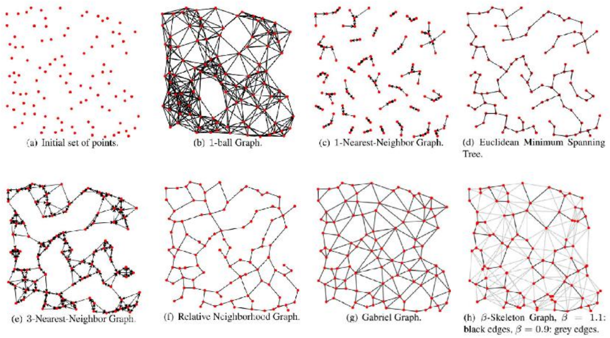
29.000 citations …

Watts, D. J. & Strogatz, S. (1998) Collective dynamics of small-world networks. *Nature, 393, 6684, 440-442.*

Milgram, S. 1967. The small world problem. *Psychology today, 2, (1), 60-67.*

Slide 5-15 Graphs from Point Cloud Data Sets

HCI-KDD



Lézoray, O. & Grady, L. 2012. Graph theory concepts and definitions used in image processing and analysis. *In: Lézoray, O. & Grady, L. (eds.) Image Processing and Analysing With Graphs: Theory and Practice. Boca Raton (FL): CRC Press, pp. 1-24.*

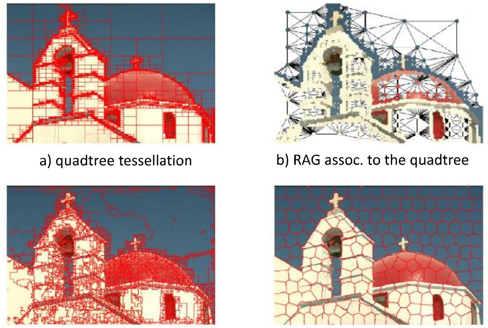
Finally a practical example

HCI-KDD

07 How do you get point cloud data from natural images?

Graphs from Images

HCI-KDD



a) quadtree tessellation

b) RAG assoc. to the quadtree

c) Watershed Algorithm

d) SLIC superpixels

Lézoray, O. & Grady, L. 2012. Graph theory concepts and definitions used in image processing and analysis. *In: Lézoray, O. & Grady, L. (eds.) Image Processing and Analysing With Graphs: Theory and Practice. Boca Raton (FL): CRC Press, pp. 1-24.*

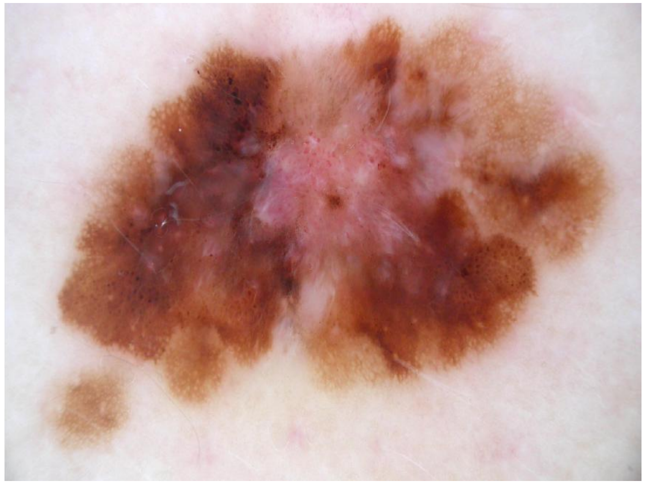
Example Watershed Algorithm

HCI-KDD

Algorithm 4.2 Watershed transform w.r.t. topographical distance based on image integration via the Dijkstra-Moore shortest paths algorithm.

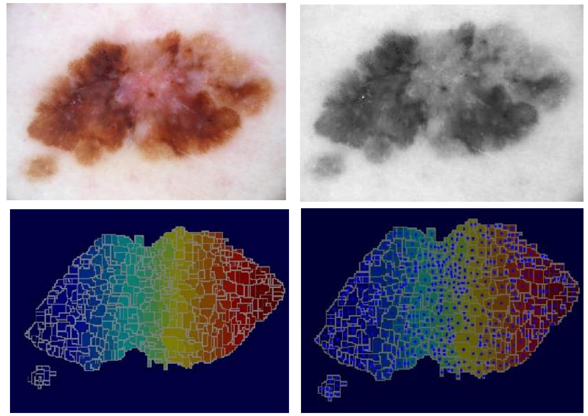
```
procedure ShortestPathWatershed;
INPUT: lower complete digital grey scale image G = (V, E, im) with cost function cost.
OUTPUT: labelled image lab on V.
#define WSHED 0                          (* label of the watershed pixels *)
(* Uses distance image dist. On output, dist[v] = im[v], for all v ∈ V. *)

for all v ∈ V do        (* Initialize *)
  lab[v] ← 0 ; dist[v] ← ∞
end for
for all local minima m_i do
  for all v ∈ m_i do
    lab[v] ← i ; dist[v] ← im[v]        (* initialize distance with values of minima *)
  end for
end for
while V ≠ ∅ do
  u ← GetMinDist(V)        (* find u ∈ V with smallest distance value dist[u] *)
  V ← V \ {u}
  for all v ∈ V  with (u, v) ∈ E do
    if dist[u] + cost[u, v] < dist[v] then
      dist[v] ← dist[u] + cost(u, v)
      lab[v] ← lab[u]
    else if lab[v] ≠ WSHED and dist[u] + cost[u, v] = dist[v] and lab[v] ≠ lab[u] then
      lab[v] = WSHED
    end if
  end for
end while
```

Meijster, A. & Roerdink, J. B. A proposal for the implementation of a parallel watershed algorithm. Computer Analysis of Images and Patterns, 1995. Springer, 790-795.



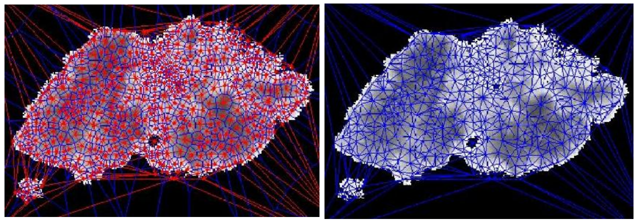
Graphs from Images: Watershed + Centroid

HCI-KDD



Slide 5-20 Graphs from Images: Voronoi <> Delaunay

HCI-KDD



Holzinger, A., Malle, B. & Giuliani, N. 2014. On Graph Extraction from Image Data. In: Slezak, D., Peters, J. F., Tan, A.-H. & Schwabe, L. (eds.) Brain Informatics and Health, BIH 2014, Lecture Notes in Artificial Intelligence, LNAI 8609. Heidelberg, Berlin: Springer, pp. 552-563.

For Voronoi please refer to: Aurenhammer, F. 1991. Voronoi Diagrams - A Survey of a fundamental geometric data structure. *Computing Surveys,* 23, (3), 345-405.

For Delaunay please refer to: Lee, D.-T. & Schachter, B. J. 1980. Two algorithms for constructing a Delaunay triangulation. Intl. Journal of Computer & Information Sciences, 9, (3), 219-242.

Are graphs better than feature vectors ?

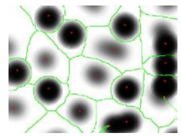
HCI-KDD

- More expressive data structures
- Find novel connections between data objects
- Fit for applying graph based machine learning techniques
- New approaches (Belief Propagation, global understanding from local properties)

Bunke, H.: Graph-based tools for data mining and machine learning. In Perner, P., Rosenfeld, A., eds.: Machine Learning and Data Mining in Pattern Recognition, Proceedings. Volume 2734 of Lecture Notes in Artificial Intelligence. Springer-Verlag Berlin, (Berlin) 7–19
Holzinger, A., Blanchard, D., Bloice, M., Holzinger, K., Palade, V., Rabadan, R.: Darwin, lamarck, or baldwin: Applying evolutionary algorithms to machine learning techniques. In: The 2014 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2014), IEEE (2014) in print

Watershed methods

HCI-KDD



- Topographic maps => landscapes with height structures
- Segmentation into regions of pixels
- Assuming drops of water raining on the map
- Following paths of descent
- Lakes called catchment basins
- Also possible: Flooding based
- Needs Topographical distance measures (MST)

Vincent, L. & Soille, P. 1991. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. IEEE transactions on pattern analysis and machine intelligence, 13, (6), 583-598.

Watershed 4 Steps

- 1) Transformation into a topographic map
  - Convert gray values into height information
- 2) Finding local minima
  - Inspecting small regions in sequence
- 3) Finding catchment basins
  - Algorithm simulating flooding
  - Graph algorithms such as Minimum Spanning Trees
- 4) Erecting watersheds
  - Artificial divide between catchment basins
  - Final segmentation lines

Watershed Algo based on connected components

7	4	8	12	11	3
7	7	8	12	11	7
13	13	15	16	16	13
19	19	18	17	15	7
20	18	17	16	15	5

(a) The original image

→	m	←	←	→	m
↗	↑	↖	←	↗	↑
↑	↑	↖	↖	↗	↑
↑	↑	↑	→	↘	↓
→	→	→	→	→	m

(b) Each pixel connect to lowest minimum

0	0	0	0	1	1
0	0	0	0	1	1
0	0	0	0	1	1
0	0	0	2	2	2
2	2	2	2	2	2

(c) The Image with labels

Connects each pixel to the lowest neighbor pixel, all pixel connected to same lowest neighbor pixel form a segment

Region Merging (from here see Tutorial Bernd Malle)

- Region Merging
  - Based on Kruskals MST algorithm
  - Takes input image as natural graph with vertices := pixels and edges := pixel neighborhoods
  - Visits edges in ascending order of weight and merges regions if they satisfy a certain criterion
  - Flexible as merging criterion can be adapted as desired (for amount, size, or shape of resulting regions)

Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. International Journal of Computer Vision 59 (2004) 167–181

Challenges

- We want to find "interesting" novel patterns (rules, anomalies, outliers, similarities, …)
- Problem #1: How to get a graph?
- Problem #2: How do graphs evolve?
- Problem #3: What tools to apply?
- Problem #4: Scalability to TB, PB, EB …
- **Success is in repeatability and scalability**

State-of-the-Art Facts

- Study of complex networks started in the 1990s with the insight that real networks contain properties not present in random (Erdös-Renyi) networks.
- Meanwhile networks and network-based approaches form an integral part of many studies throughout the sciences.
- Graph-Theory provides powerful tools to organize data structurally and in combination with statistical and machine learning methods allows a meaningful analysis of underlying processes.
- For instance, a mapping of causal disease genes and disorders as made available by the OMIM database provided novel insights into disease patterns, as recently demonstrated by investigating the diseasome (http://diseasome.eu).

Future Outlook