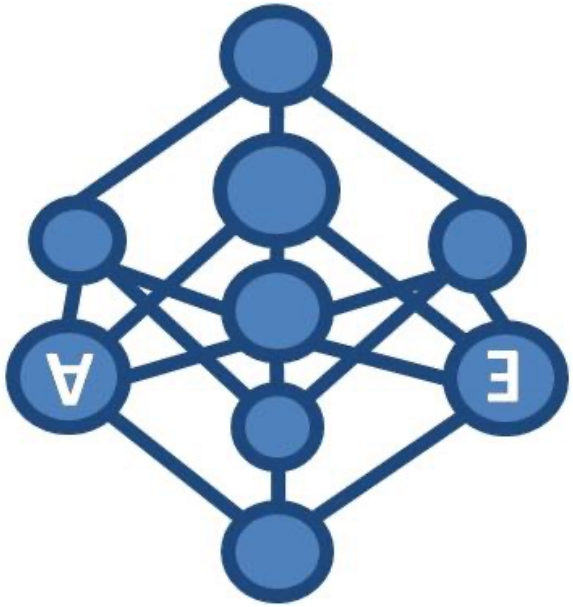
Probabilistic Graphical Models*

Introduction

TECHNISCHE UNIVERSITÄT DARMSTADT



*Thanks to Carlos Guestrin, Pedro Domingos and many others for making their slides publically available

This Course

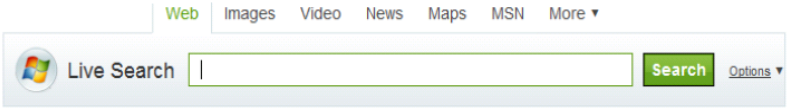
One of the **most exciting developments** in

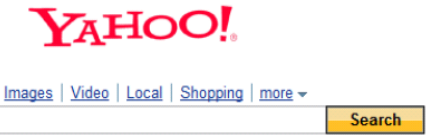
- Machine Learning
- Knowledge representation
- Artificial Intelligence,
- Statistics,
- Electrical Engineering,
- …

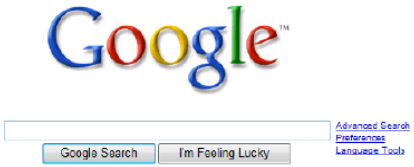
in the last two (or three, or more) decades…

SOME EXAMPLE APPLICATIONS

Web Search







Relative Feedback:
Clicks reflect preference between observed links.

Absolute Feedback:
The clicked links are relevant to the query.

(3 < 2),
(7 < 2),
(7 < 4),
(7 < 5),
(7 < 6)

1. Kernel Machines
   http://svm.first.gmd.de/
2. Support Vector Machine
   http://jbolivar.freeservers.com/
3. SVM-Light Support Vector Machine
   http://ais.gmd.de/~thorsten/svm light/
4. An Introduction to Support Vector Machines
   http://www.support-vector.net/
5. *Support Vector Machine and Kernel ... References*
   http://svm.research.bell-labs.com/SVMrefs.html
6. Archives of SUPPORT-VECTOR-MACHINES ...
   http://www.jiscmail.ac.uk/lists/SUPPORT...
7. Lucent Technologies: SVM demo applet
   http://svm.research.bell-labs.com/SVT/SVMsvt.html
8. Royal Holloway Support Vector Machine
   http://svm.dcs.rhbnc.ac.uk

Rel(1),
NotRel(2),
Rel(3),
NotRel(4),
NotRel(5),
NotRel(6),
Rel(7)

[Joachims]

Information Extraction

Parag Singla and Pedro Domingos, “Memory-Efficient Inference in Relational Domains” (AAAI-06).

Singla, P., & Domingos, P. (2006). Memory-efficent inference in relatonal domains. In Proceedings of the Twenty-First National Conference on Artificial Intelligence (pp. 500-505). Boston, MA: AAAI Press.

H. Poon & P. Domingos, Sound and Efficient Inference with Probabilistic and Deterministic Dependencies”, in Proc. AAAI-06, Boston, MA, 2006.

P. Hoifung (2006). Efficent inference. In Proceedings of the Twenty-First National Conference on Artificial Intelligence.

[Domingos et al.]

Segmentation

Parag Singla and Pedro Domingos, "Memory-Efficient Inference in Relational Domains" (AAAI-06).

Singla, P., & Domingos, P. (2006). Memory-efficent inference in relatonal domains. In Proceedings of the Twenty-First National Conference on Artificial Intelligence (pp. 500-505). Boston, MA: AAAI Press.

H. Poon & P. Domingos, Sound and Efficient Inference with Probabilistic and Deterministic Dependencies", in Proc. AAAI-06, Boston, MA, 2006.

P. Hoifung (2006). Efficent inference. In Proceedings of the Twenty-First National Conference on Artificial Intelligence.

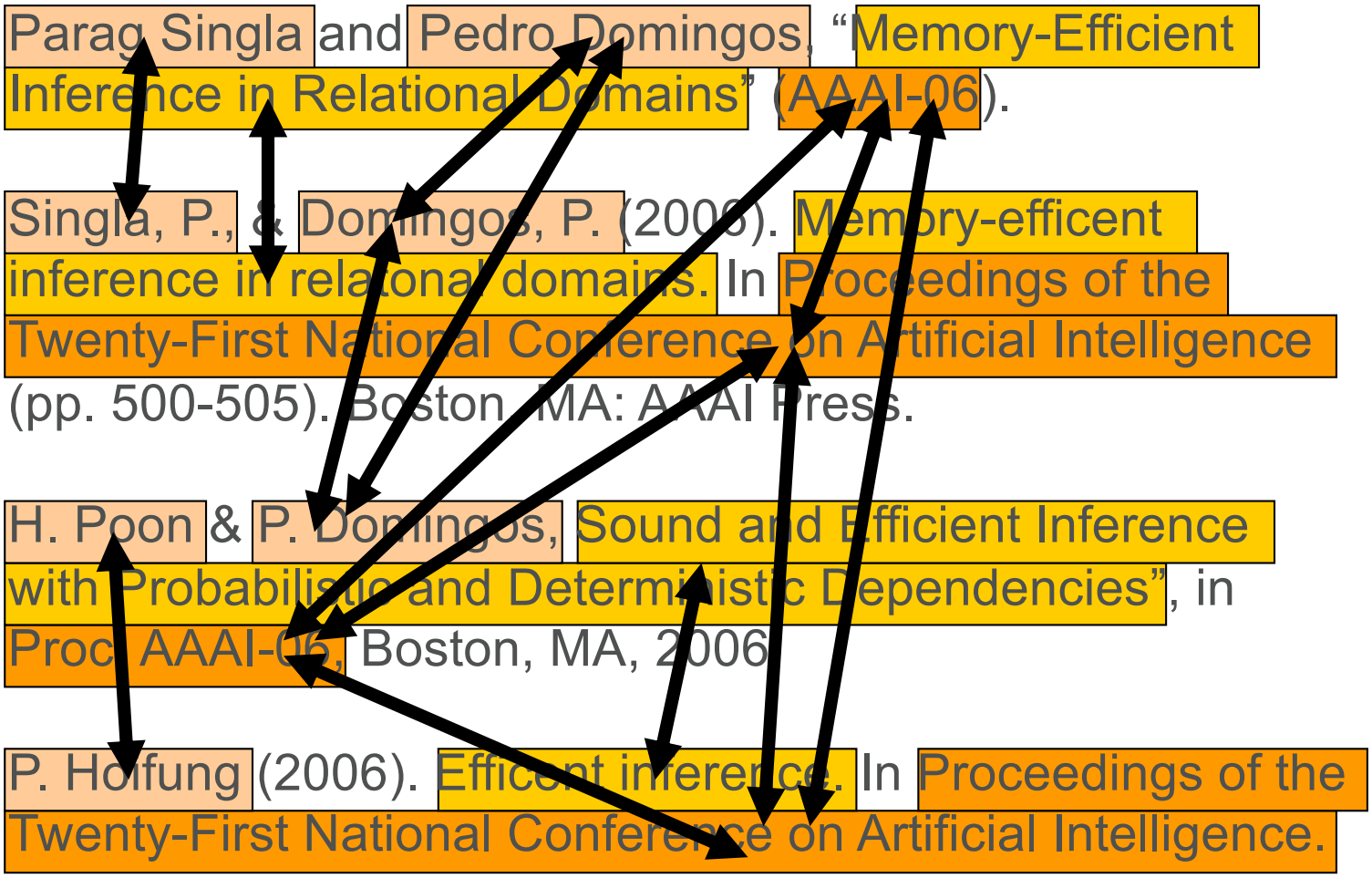
[Domingos et al.]

Author

Title

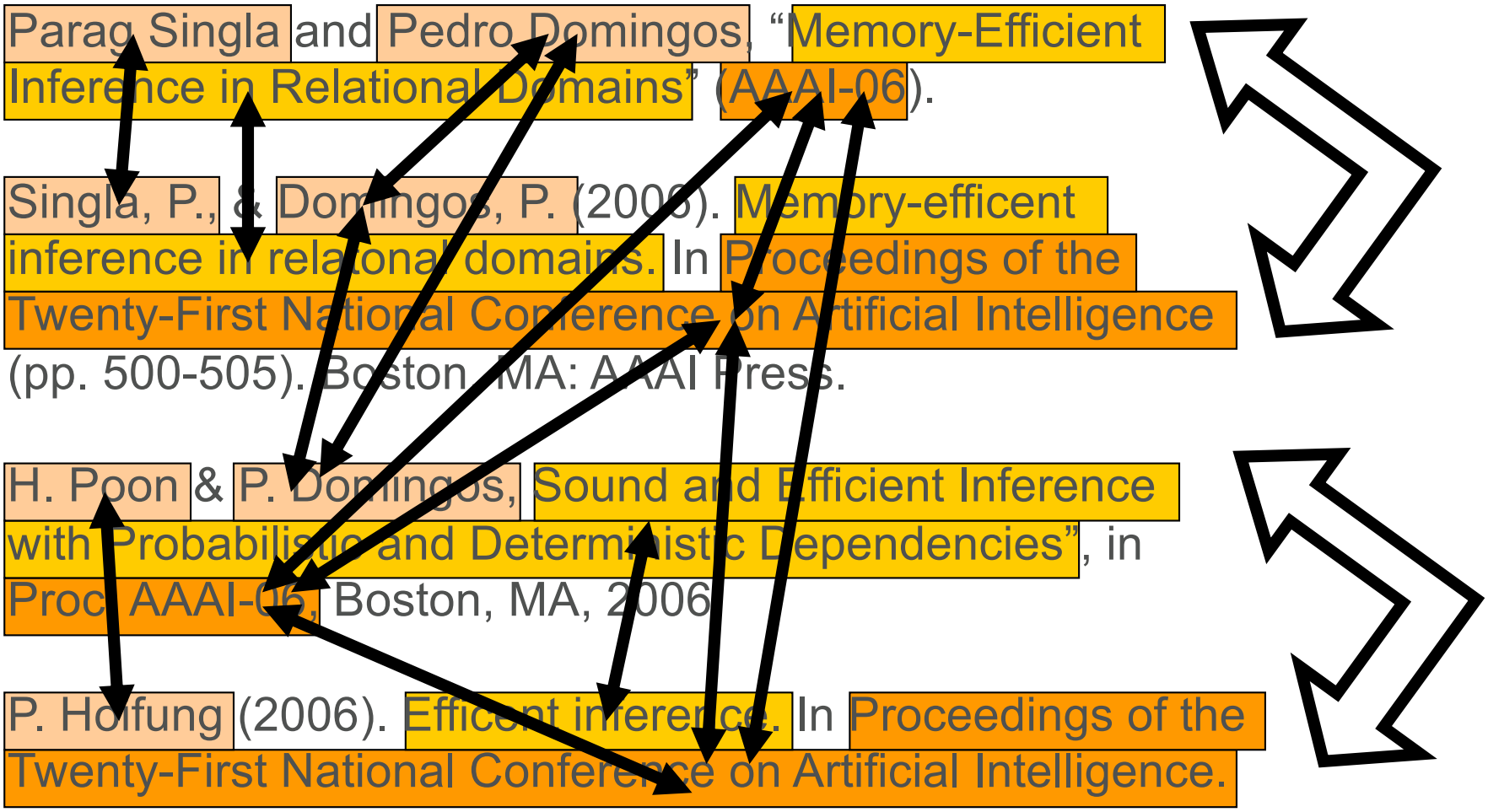
Venue

Entity Resolution



[Domingos et al.]

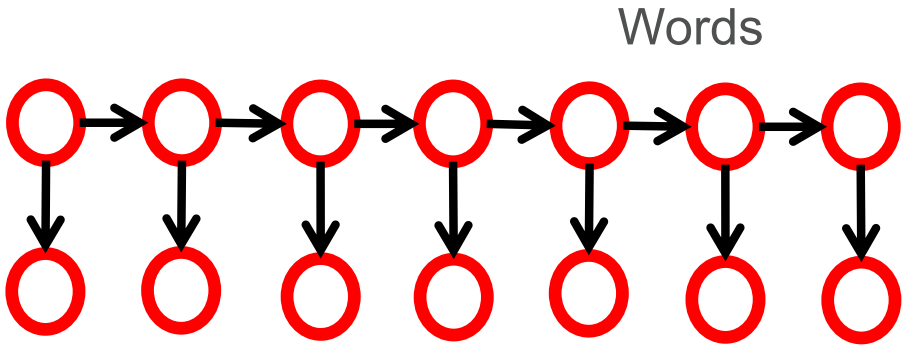
Entity Resolution



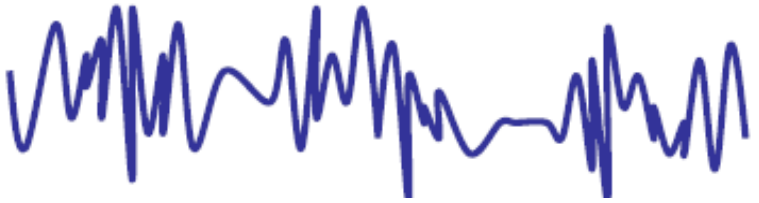
[Domingos et al.]

Speech Recognition

- Estimate the words from audio signal
- Hidden Markov Model

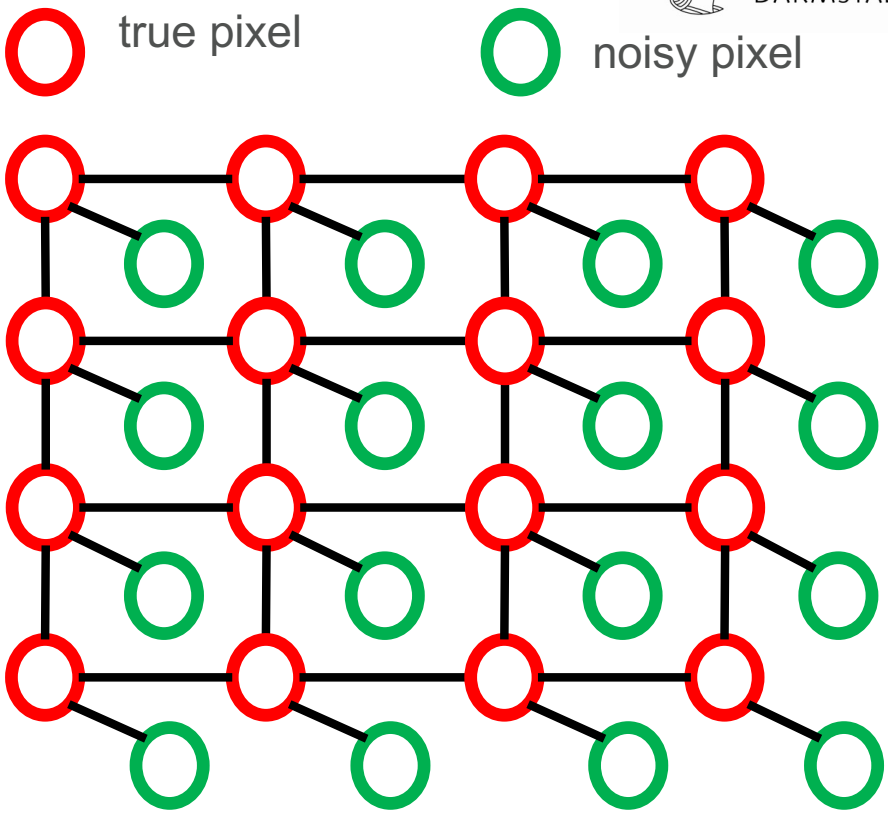


Phonems



„He ate the cookies on the couch“

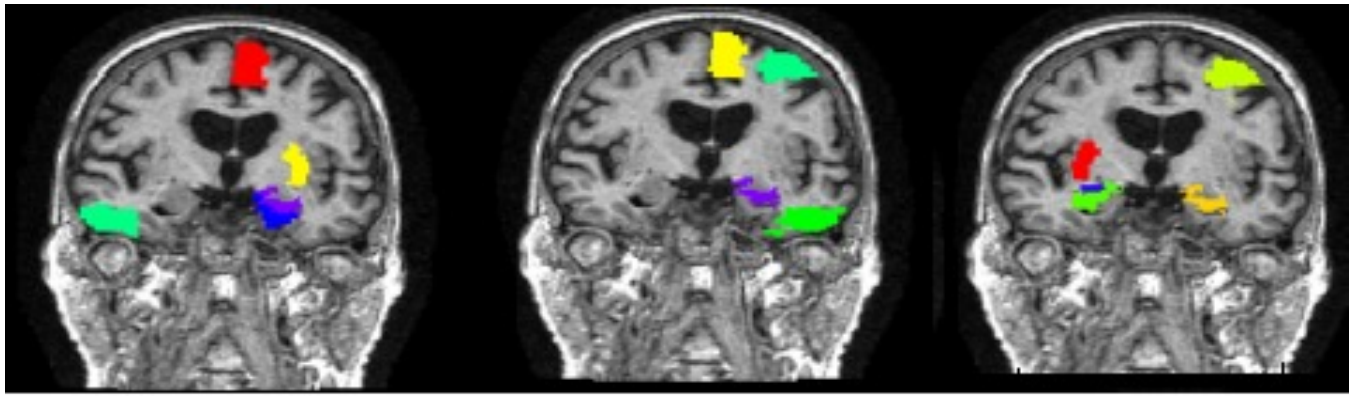
Image Denoising



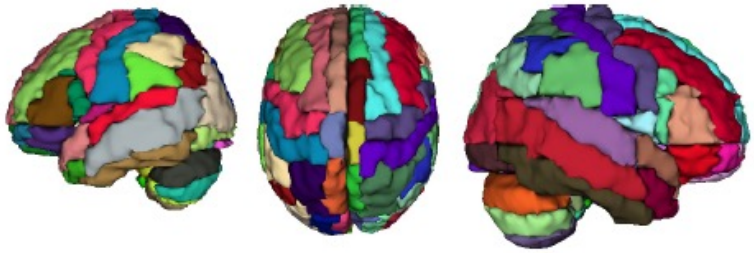
- Estimate the original image from the denoised one
- Markov Networks

[Natarajan, Joshi, Saha, Edwards, Khot, Moody, Kersting, Whitlow, Maldjian ICMLA 2012]

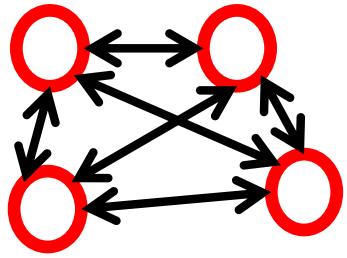
Helping to Cure Alzheimer



Alzheimer | Mild-cognitive impairment | Cognitive normal

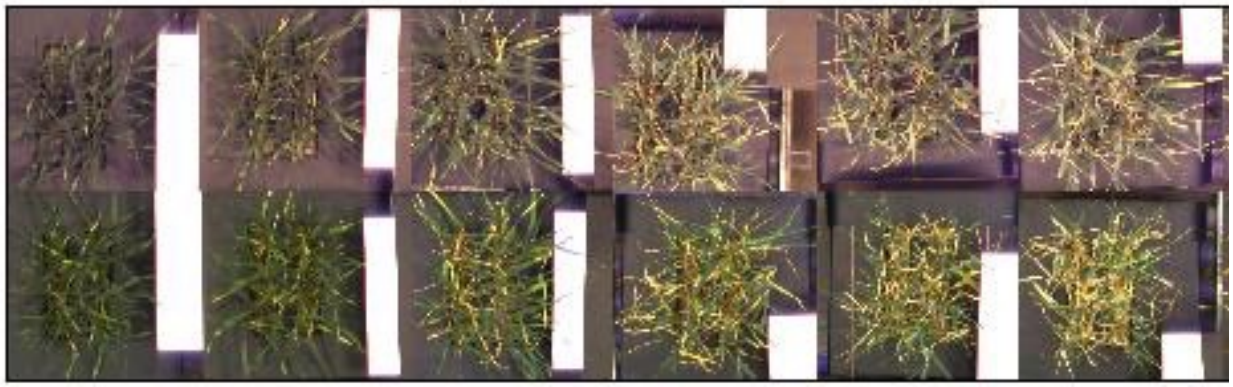


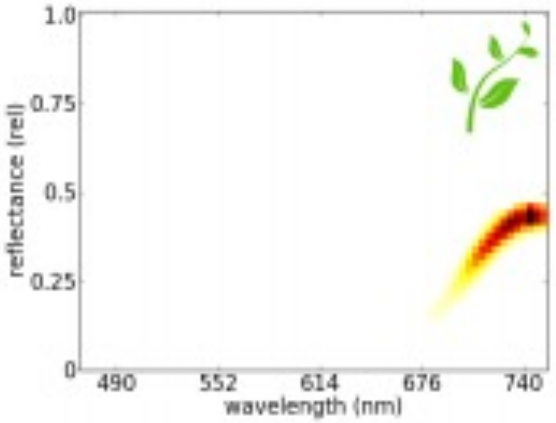
- Which areals are good for detecting different stages of Alzheimer?
- Relational dependency networks that capture both the brain and the medical knowledge

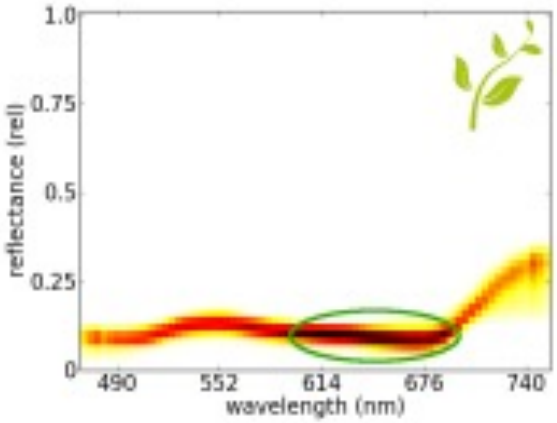


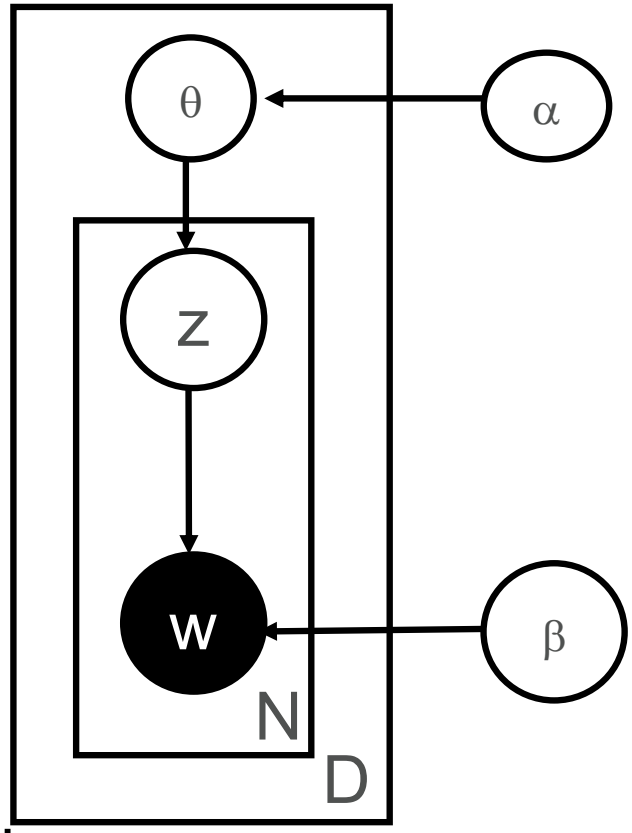
[Wahabzada, Kersting, UAI 2012, Whabzada et al. Plos One 2015]

Precision Agriculture







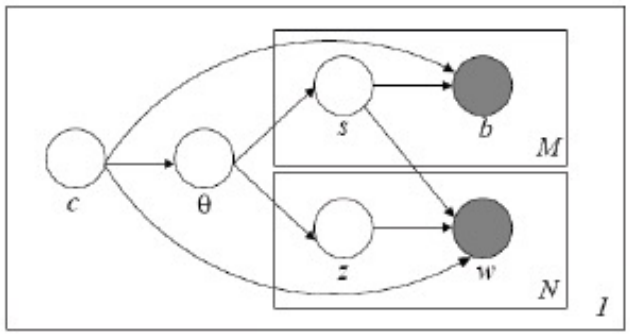


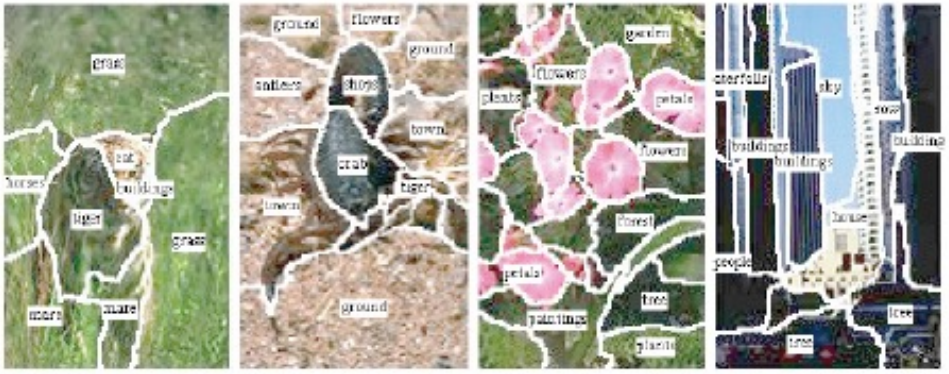
- How do plants react on stress?
- Non-paramteric Bayesian models such as Latent Dirichlet Allocation

Images and Text

Hierarchical Bayesian models

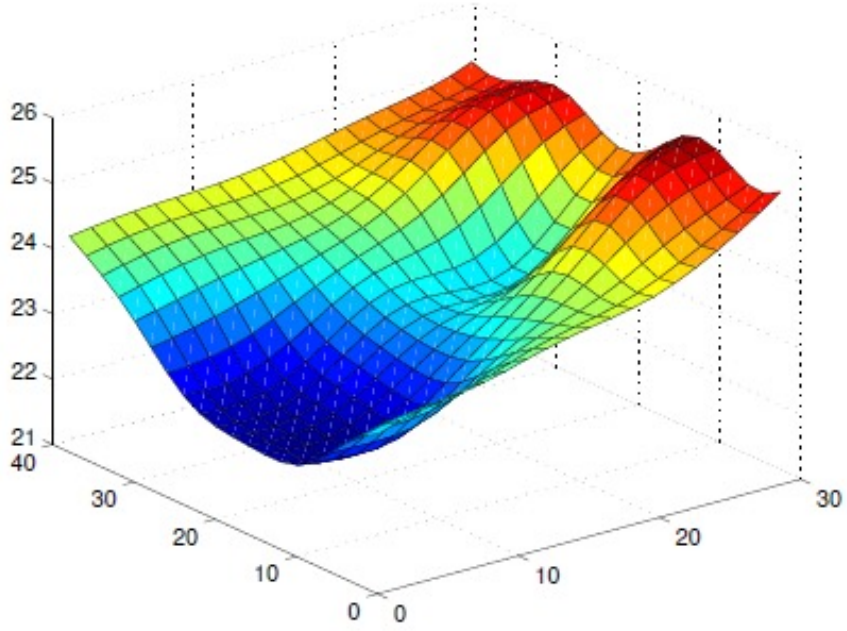
[Barnard et al.]



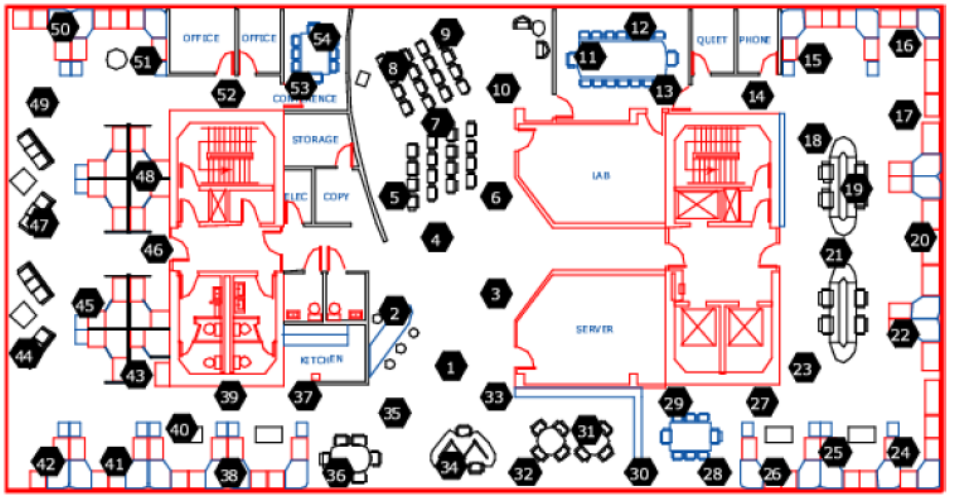


- Describe the objects captured by an image

Modeling Sensor Data

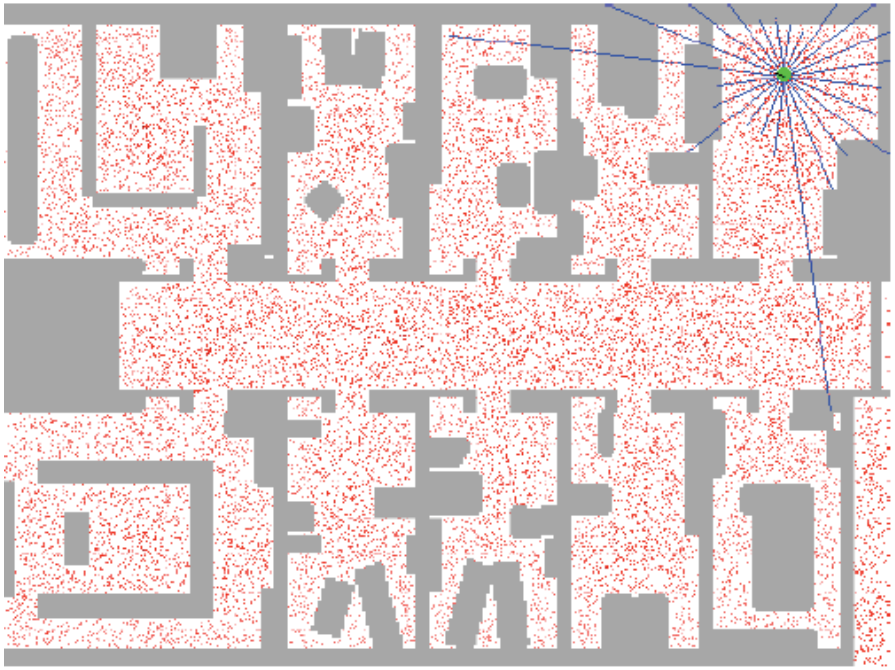


[Guestrin et al.]

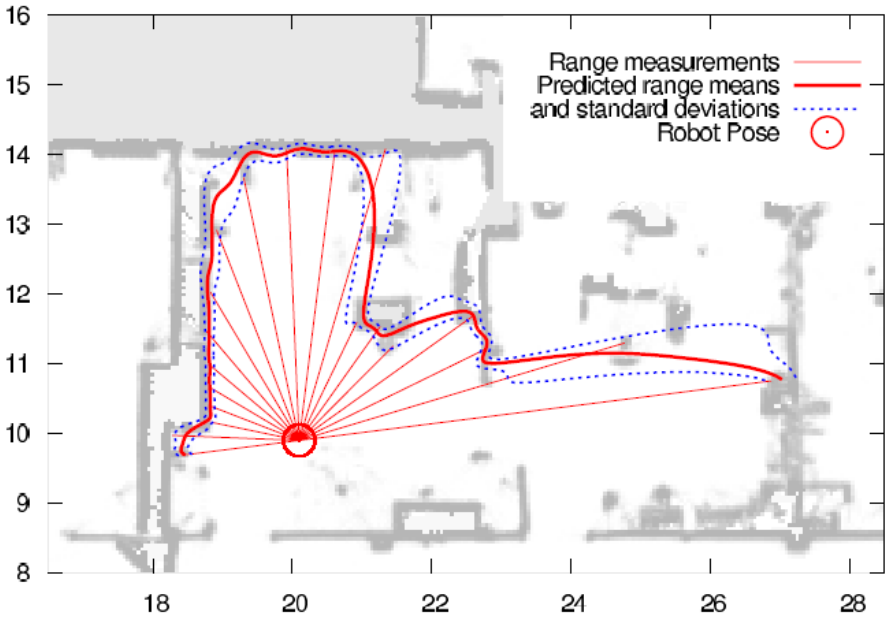


- Where to place sensors to get the best model?
- Non-paramteric Bayesian models such as Gaussian Processes

Tracking and Robot Localization



[Fox et al.]

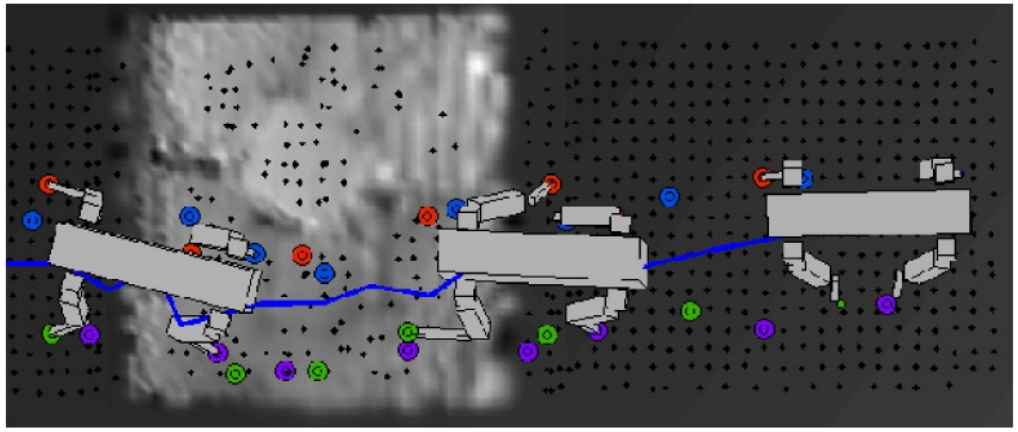


[Plagemann et al.]

- Estimate your location from observations over time
- Recursive Bayesian filters such as Kalman filter and particle filter

3D Terrain Modeling

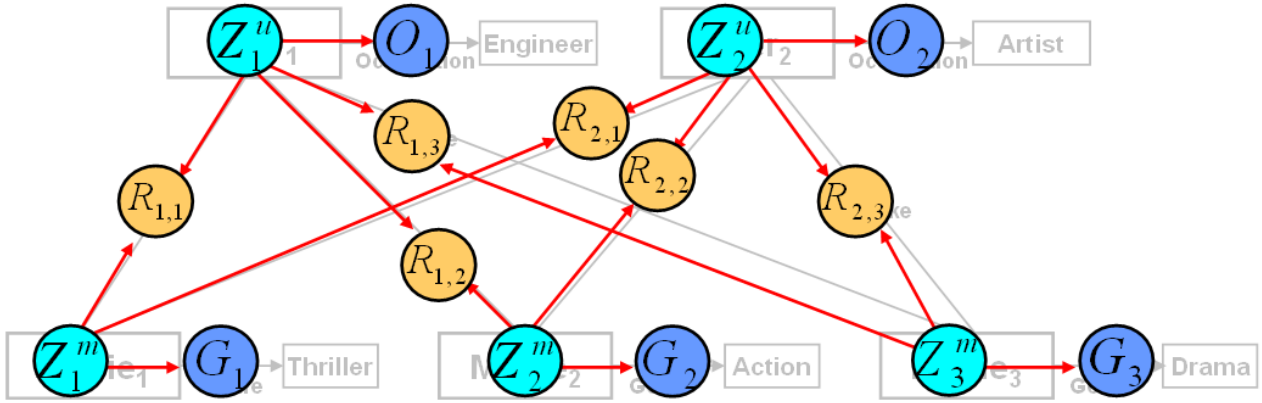
[Plagemann et al.]

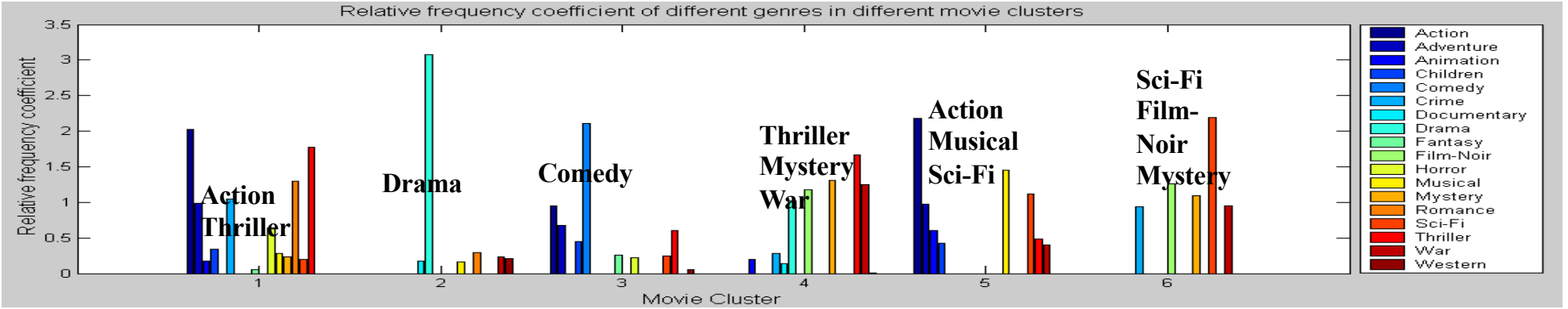


- How does the environment look like e.g. to plan foot placement?
- Large-scale Gaussian Processes

[Xu et al. Advanced in Social Networks Mining and Analysis 2010]

Social Network Analysis



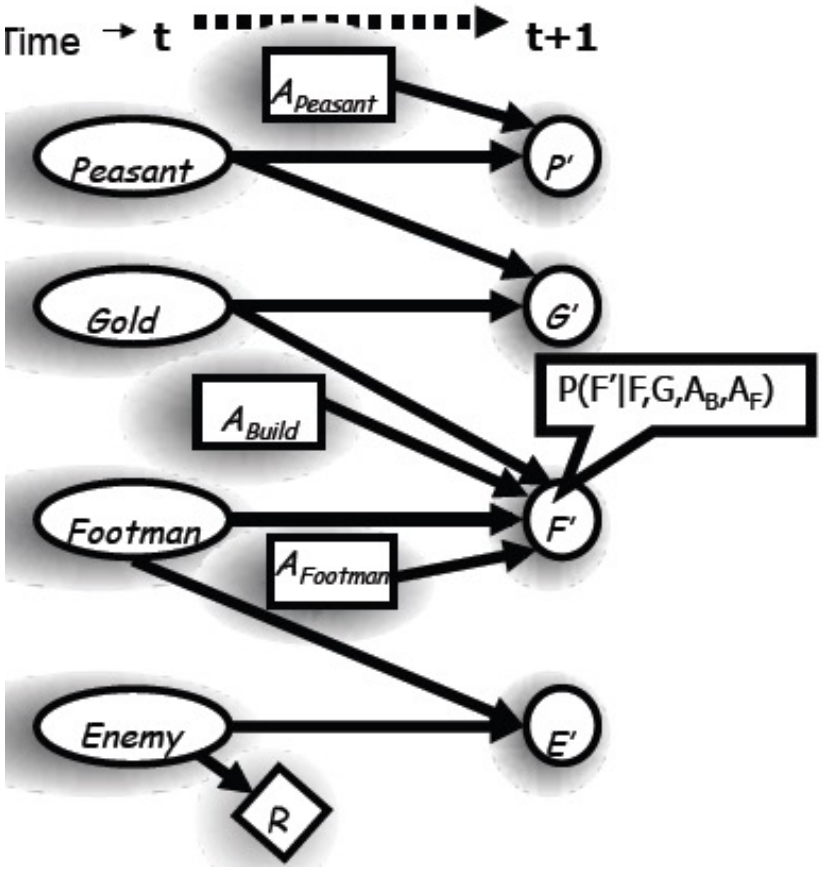


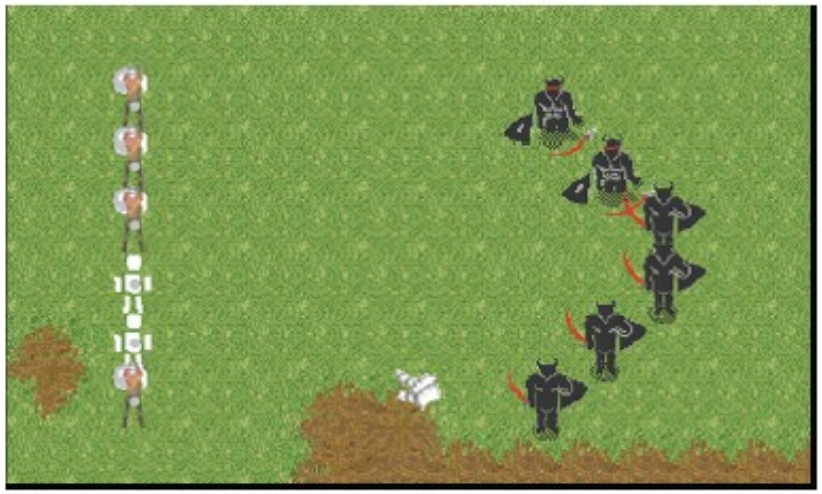
- Do social groups share interests?
- Collaborative filtering

Planning Under Uncertainty

- How do act optimally in noisy environments?
- (Partially observed) Markov Decision Processes

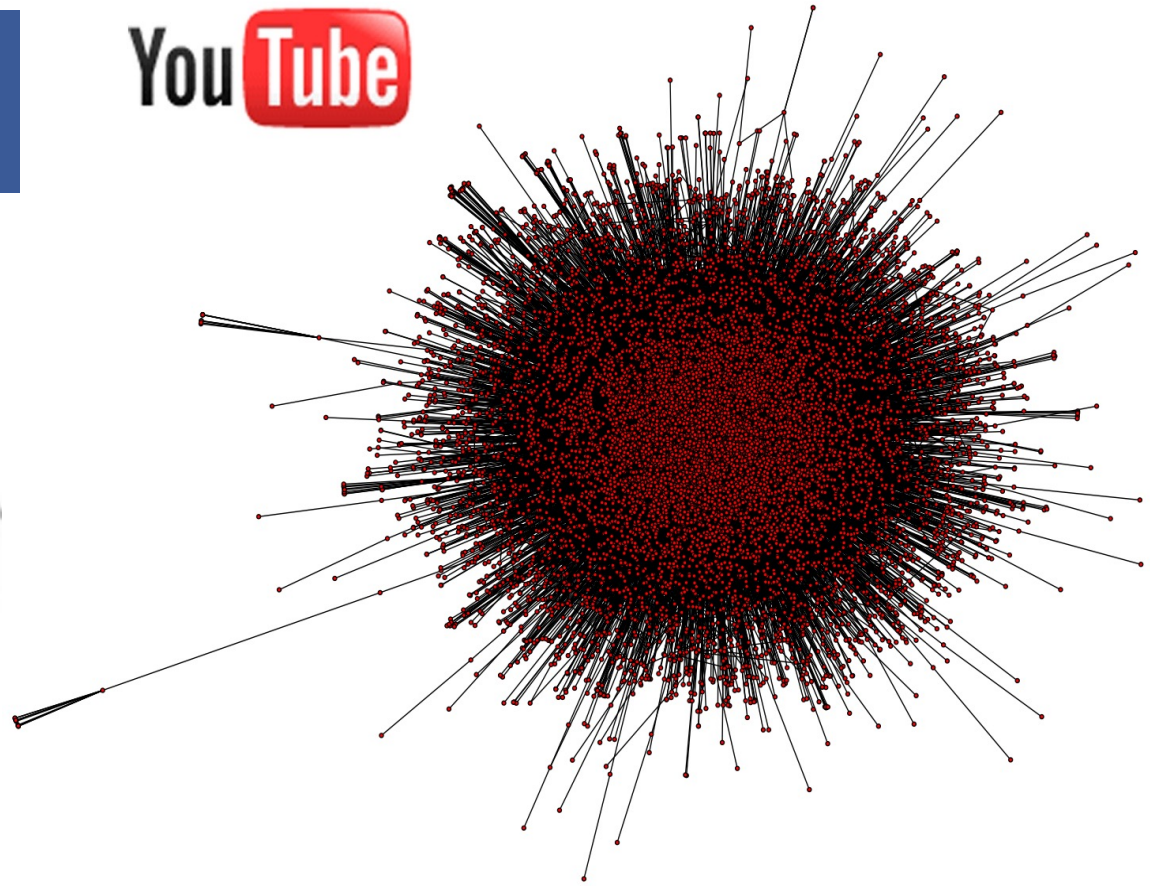
Dynamic Bayesian networks
Factored Markov decision problems





[Guestrin et al.]

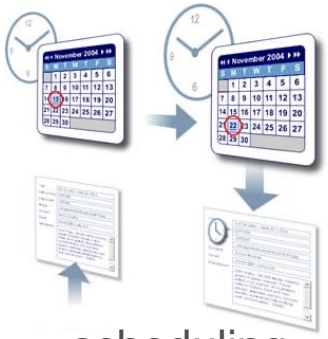
How do you efficiently broadcast information?



[Kersting et al. UAI 2010]

Combinatorial Problems

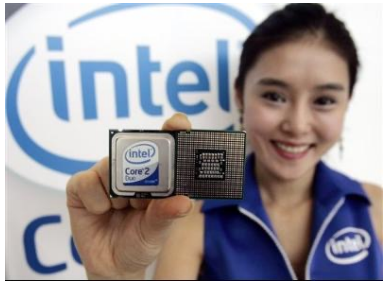
TECHNISCHE UNIVERSITÄT DARMSTADT



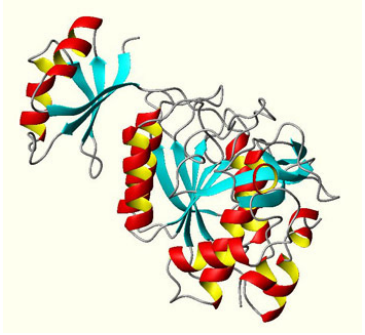
scheduling

logistics

supply chain management



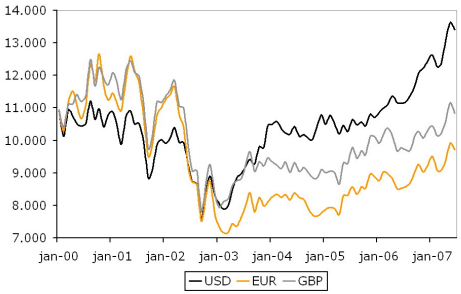
chip design



protein folding

network design

air traffic routing



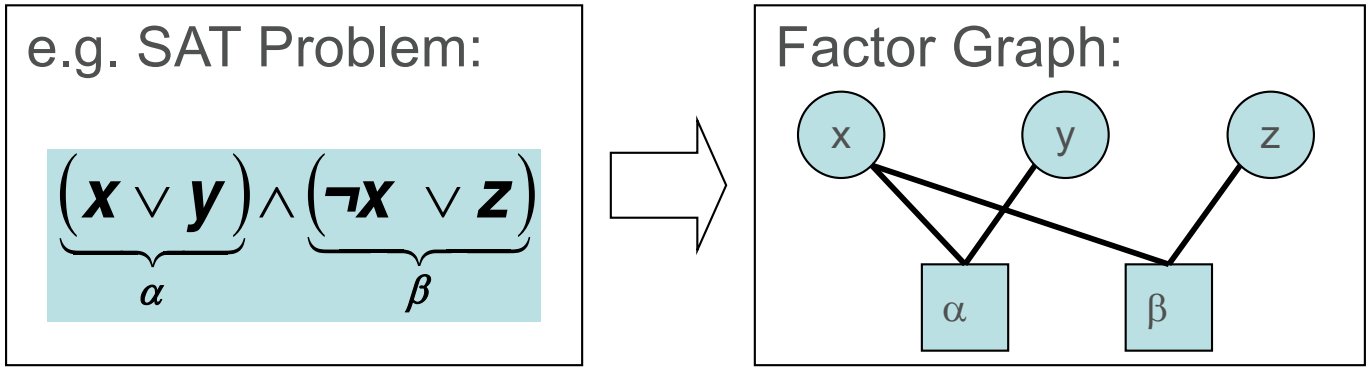
portfolio optimization

timetabling

production planning

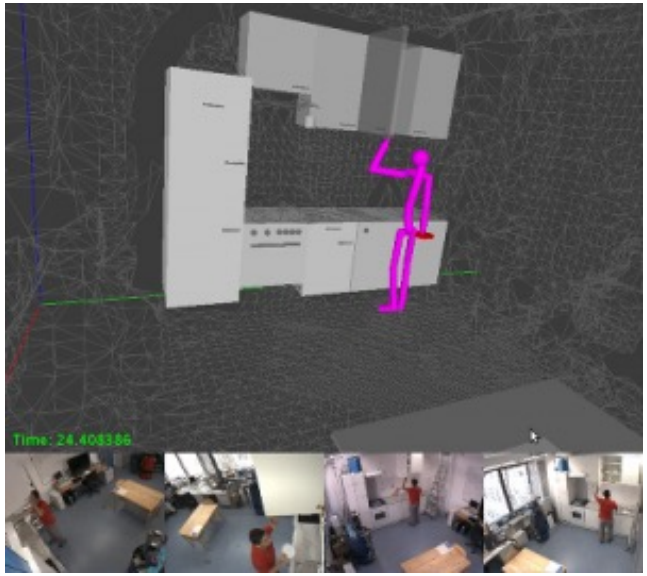
Encoding CSPs

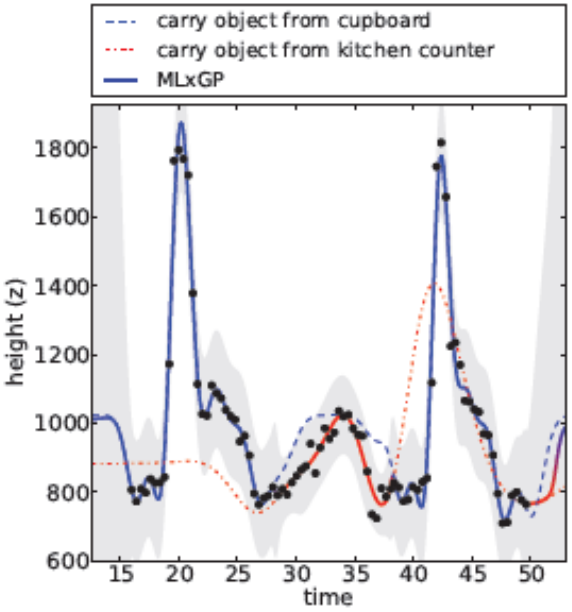
- A CSP is a problem of finding a configuration (values of discrete variables) that is globally consistent (all constraints are satisfied)
- One can visualize the connections between variables and constraints in so called **factor graph**:



[Schiegg, Neumann, Kersting 2012]

Machines Reading Regression Data





Weight	First-Order Logic
1.0	segment(0,T) ⇔ segment(0,T+1)
∞	onT(0,T) ⇒ onT(0,T+1)
⋮	⋮
10.0	segment(0!,T)

Evidence

inCb(p)	¬onT(p,18.4)	¬onT(p,25.2)
inCb(c)	¬onT(c,18.4)	onT(p,25.7)
¬inCb(k)	atT(35.7)	onT(k,43.3)
¬onT(k,18.4)	atT(25.2)	onT(c,43.3)

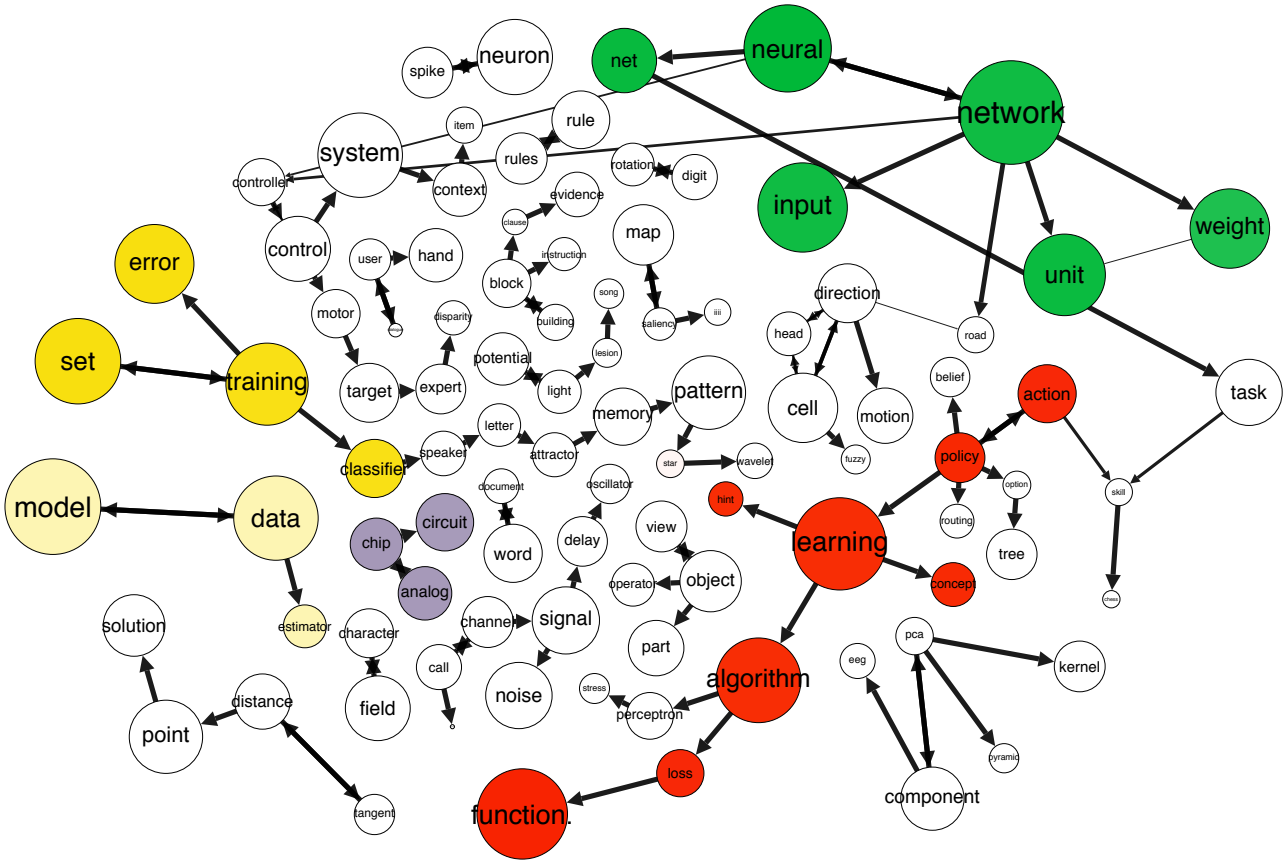
Subset of Inferred Beliefs

onT(k,22.4)	0.0
onT(k,44.8)	1.0
segment(k,32.5)	1.0
segment(c,41.6)	1.0

Let the machine read the regression data: “To set the table the subject first carried the plate, then the knife, followed by the cup."

[Hadiji, Molina, Natarajan, Kersting, 2015] and many others

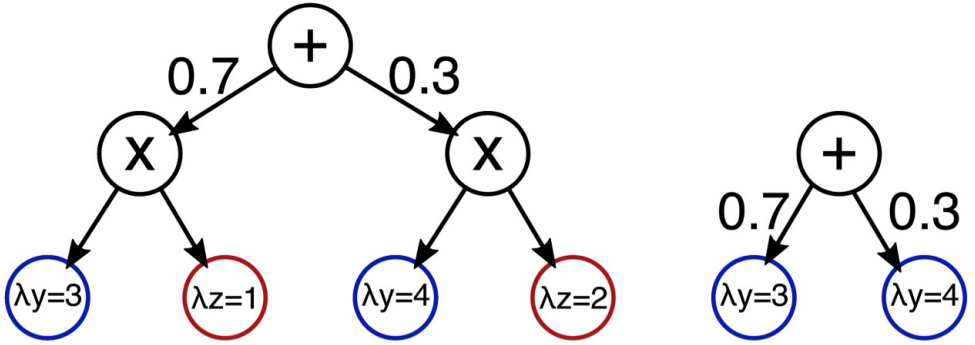
Machines Reading Texts

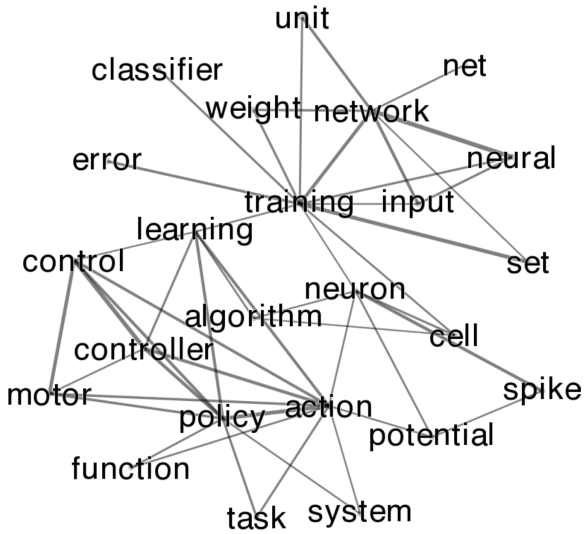


Let the machine read and understand text documents. Here you see the main dependencies among words estimated on a corpus of scientific papers

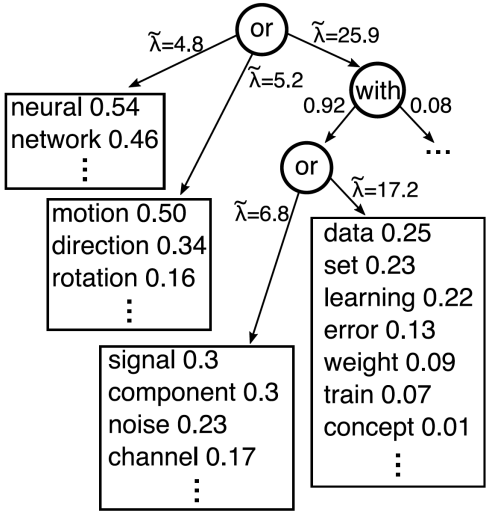
TECHNISCHE UNIVERSITÄT DARMSTADT

Deep and Tractable Probabilistic Models





Mutual Information
(NIPS corpus)

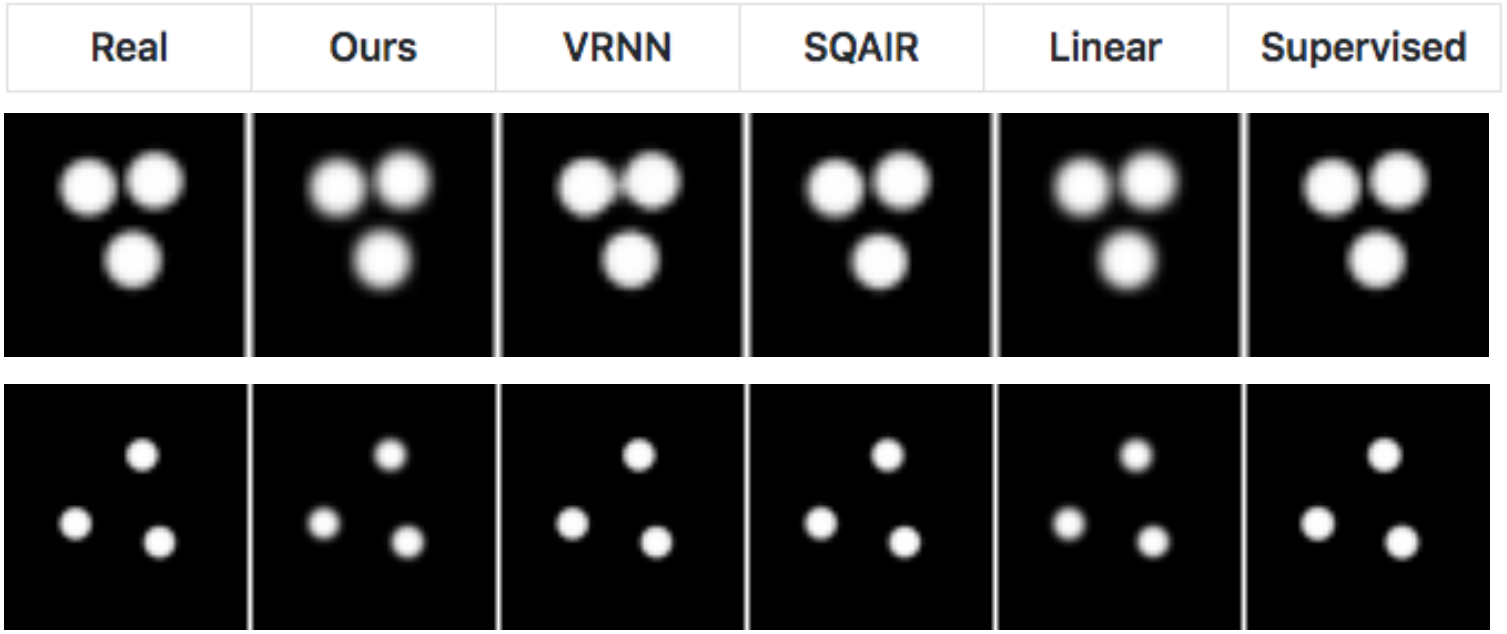


Poisson Multinomial SPN
= hierachical topic model

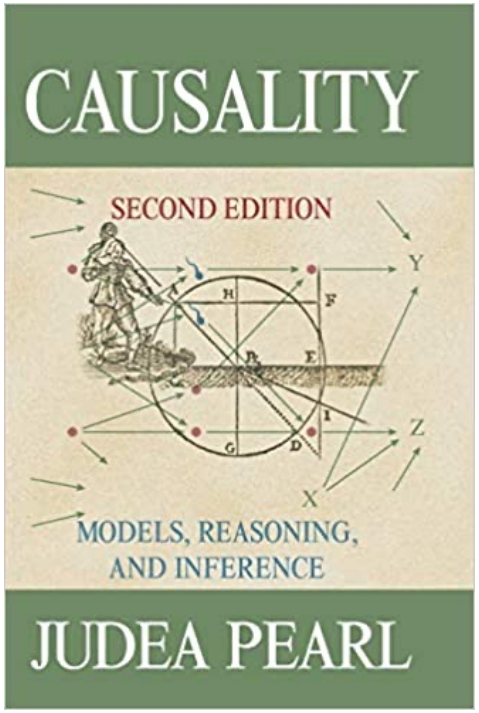
[Jannik Kossen, Karl Stelzner, Marcel Hussing, Claas Voelcker, Kristian Kersting ICLR2020]

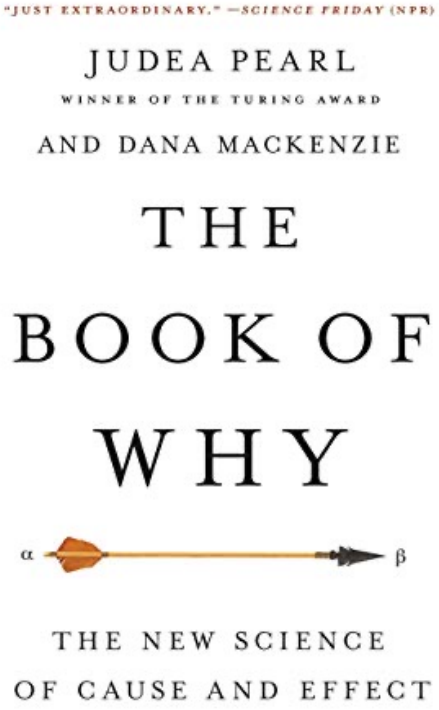
TECHNISCHE UNIVERSITÄT DARMSTADT

Learning Physics using Deep Probabilistic Models



(Tractable Causal) Models





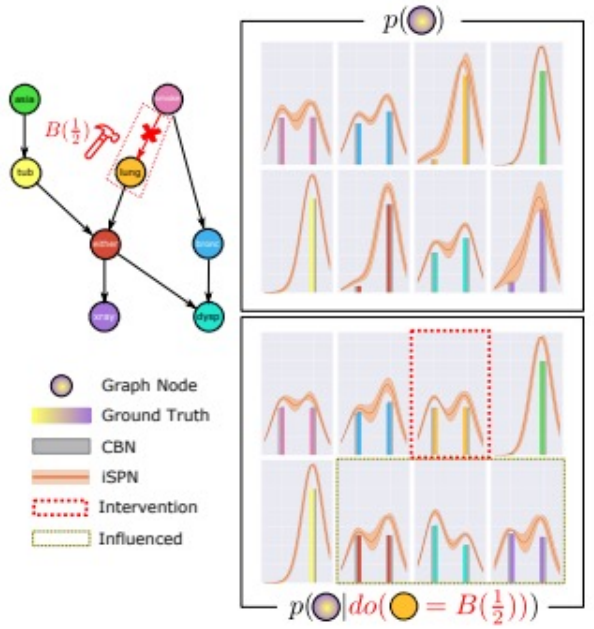


Figure 1: **Capturing interventional distributions using iSPN**. The interventional distributions for the ASIA data set using a causal Bayesian network (CBN, small-scale gold standard, gray bars) as well as an interventional SPN (iSPN) by intervening on *lung*. iSPN are sensible to the influences of the given intervention onto the system i.e., subsequent effects in the causal hierachy. (Best viewed in color.)

Based on graphical models, Judea Pearl presents and unifies the probabilistic, manipulative, counterfactual, and structural approaches to causation and devises simple mathematical tools for studying the relationships between causal connections and statistical associations

[Matej Zečević, Devendra Singh Dhami, Athresh Karanam, Sriraam Natarajan and Kristian Kersting NeurIPS]

… and many,

many,

many,

many

more !

… and they are serious business

- Probabilistic (relational) models are used by several million users

Syllabus

- Covers a wide range of **Probabilistic Graphical Models topics** – **from basic to state-of-the-art**
- You will learn things such as
  - Bayesian networks, Markov networks, factor graphs, conditional random fields, junction trees, parameter learning, structure learning, exact inference, variable elimination, approximate inference, sampling, importance sampling, MCMC, Gibbs, loopy belief propagation, Bayesian learning, missing data, EM, temporal Models, hidden Markov Models, Forwards-Backwards, Viterbi, Baum-Welch, assumed density filtering, DBNs, BK, Relational probabilistic models,…
- **Covers algorithms, some theory, and applications**
- **It's going to be fun but also some hard work**

Prerequisites

- We provide background, but class can be fast paced
- Ability to deal with “abstract mathematical concepts”
- Do not hesitate to ask questions. Let’s make this course as interactive as possible

Steven Lang, steven.lang@cs.tu-darmstadt.de
Wolfgang Stammer, wolfgang.stammer@cs.tu-darmstadt.de

Devendra Dhami, devendra.dhami@cs.tu-darmstadt.de
Kristian Kersting, kersting@cs.tu-darmstadt.de

Exercise Sessions

TECHNISCHE UNIVERSITÄT DARMSTADT

- Very useful!
  - Review material, present background, answer questions, Discussion, just chatting about everything.
- „Hands on“ Project
  - How does this work in practice?
  - Prepare yourself for probabilistic methods such as graphical models but also (Poisson) trees, generalized linear models, boosting ... In particular make use of fancy python packages. Check also Kaggle. Anyhow, convince me of a project of your choose!
- 1.5 hours per (every second?) week
- Date and place: **Fridays 11:40-13:20, S103/221**

Hands-on project!

TECHNISCHE UNIVERSITÄT DARMSTADT

- Use our own library SPflow for sum-product networks:

  https://github.com/SPFlow/SPFlow

This is a deep variant of graphical models. Apply it to something interesting! Or add functionalities such as other learning algorithms, inference algorithms, compilation to the new PyTorch via EinSums, combinations with DeepNetworks, extracting variable importance measures, causality, interactive learning, …

Grading

- Final exam will count
- Entry to the final exam
  - Regularly attending exercise sessions
  - Regularly working on the homework

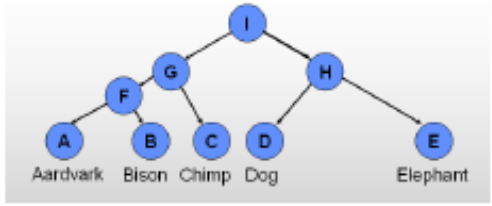
Homeworks

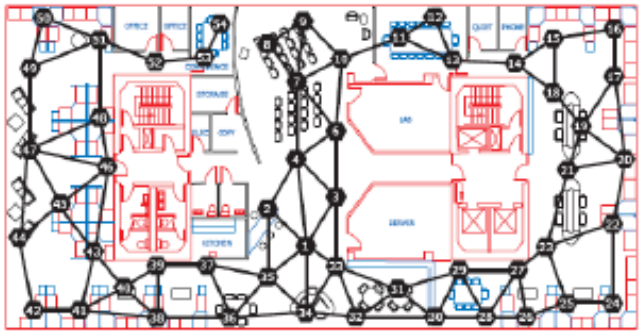
- Homeworks might be hard, start early ☺
- Due in the beginning of exercise sessions
- Collaboration
  - You may **discuss the questions**
  - Each student writes its own answers
  - Write on your homework anyone with whom you collaborated
- **IMPORTANT:** We may use some material from other classes or from papers for the homeworks. Unless otherwise specified, please only look at the readings when doing your homework ! **You are taking this class because you want to learn, so this rule is self-enforced**

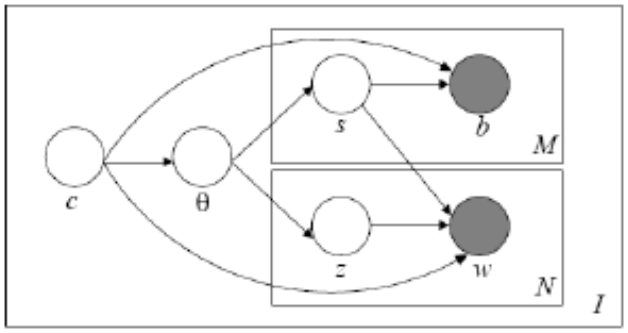
Enjoy !!

- **Probabilistic graphical models are having significant impact** in science, engineering and beyond
- This class should give you the **basic foundation for applying GMs and developing new methods**
- The fun begins…

What are the Fundamental Questions of Graphical Models?







- Representation
  - What are the types of models?
  - What does the model mean/imply/assume?
- Inference
  - How do I answer questions/queries with my model?
- Learning
  - I am lazy. Can data tell me the right model?
  - What model is the right for my data?

What are the Fundamental Questions of Graphical Models?

TECHNISCHE UNIVERSITÄT DARMSTADT

- Representation
  - Graphical models represent expontentially large probability distributions compactly
  - **Key concept:** Conditional Independence
- Inference
  - What is the probability of X given some observations?
  - What is the most likely explanation for what is happening?
  - What decision should I make?
- Learning
  - What are the right/good parameters for the model?
  - How do I obtain the structure of the model?

Where do we start?

- From Bayesian networks
- „Complete“ BN presentation first
  - Representation, Exact Inference, Learning
- We will focus on discrete domains
- Later in the semester
  - Undirected models, Approximate Inference, Temporal models, maybe Gaussian processes, i.e., continuous domains, Relational domains; let‘s see what we will manage
- Class focuses on fundamentals – Understand the basic concepts is important for me

But first

- Review of basic probability concepts
  - Probabilites
  - Independence
  - Etc.
- Next time
  - Two nodes make a Bayesian network
  - Naive Bayes
  - Bayesian networks

Probability Review

TECHNISCHE UNIVERSITÄT DARMSTADT

- Events and Event spaces
- Random variables
- Joint probability distributions
  - Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties
  - Independence, conditional independence
- Examples
- Continuous distributions, Moments, Entropy, MLE, …

Sample space and Events

TECHNISCHE UNIVERSITÄT DARMSTADT

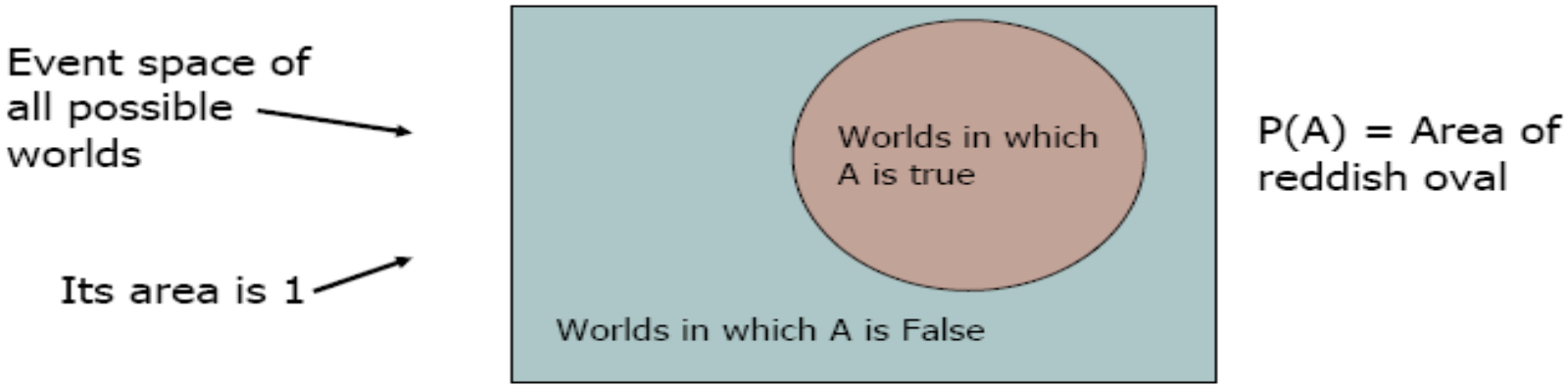
- **Ω : Sample Space**, result of an experiment
  - If the experiment is tossing a coin twice, $\Omega = \{HH, HT, TH, TT\}$
- Event: a subset of Ω
  - First toss is head = {HH,HT}
- **S: event space, set of events:**
  - Closed under finite union and complements
    - Entails other binary operation: union, diff, etc.
  - Contains the empty event (no event) and Ω (every event)

Probability Measure

TECHNISCHE UNIVERSITÄT DARMSTADT

- Defined over (Ω,S) s.t.
  - P(α) >= 0 for all α in S
  - P(Ω) = 1
  - If α, β are disjoint, then
    - P(α U β) = P(α) + P(β)
- We can deduce other axioms from the above ones
  - Ex.: P(α U β) for non-disjoint event

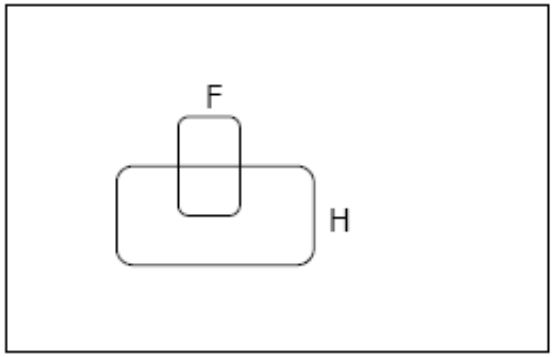
Illustration



- We can go on and define conditional probability, using the above visualization

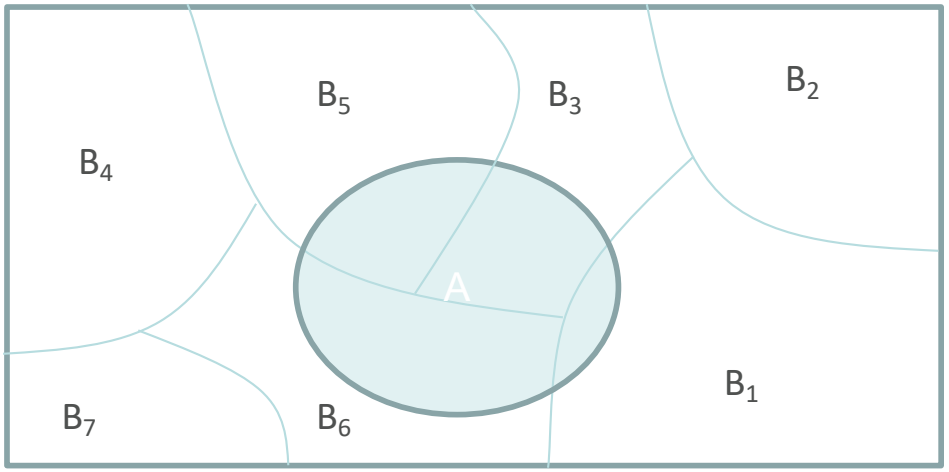
Conditional Probability

P(F|H) = Fraction of worlds in which H is true that also have F true



$$p(f \mid h) = \frac{p(F \cap H)}{p(H)}$$

Rule of total probability



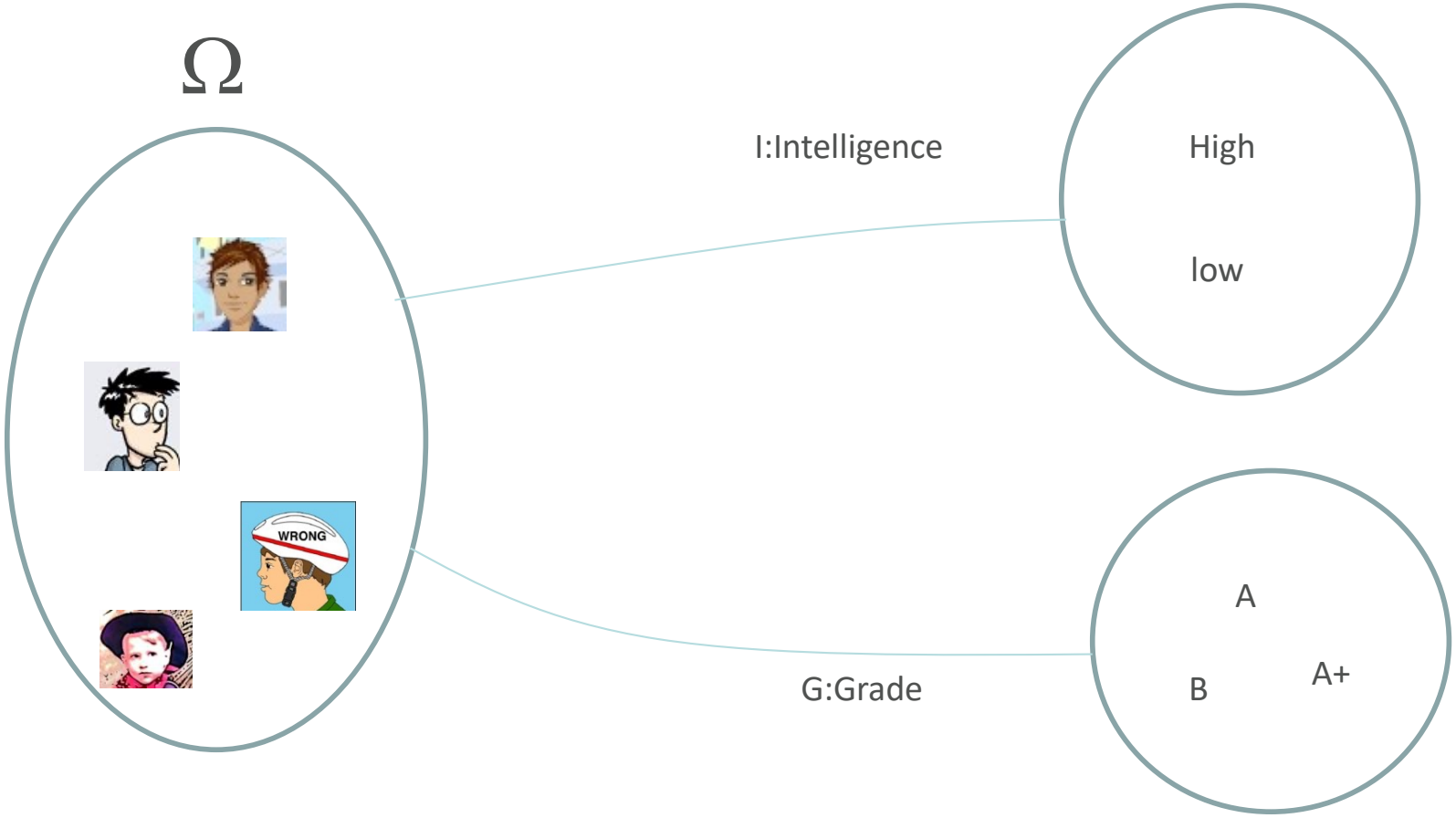
$$p(A) = \sum P(B_i) P(A \mid B_i)$$

From Events to Random Variable (RV)

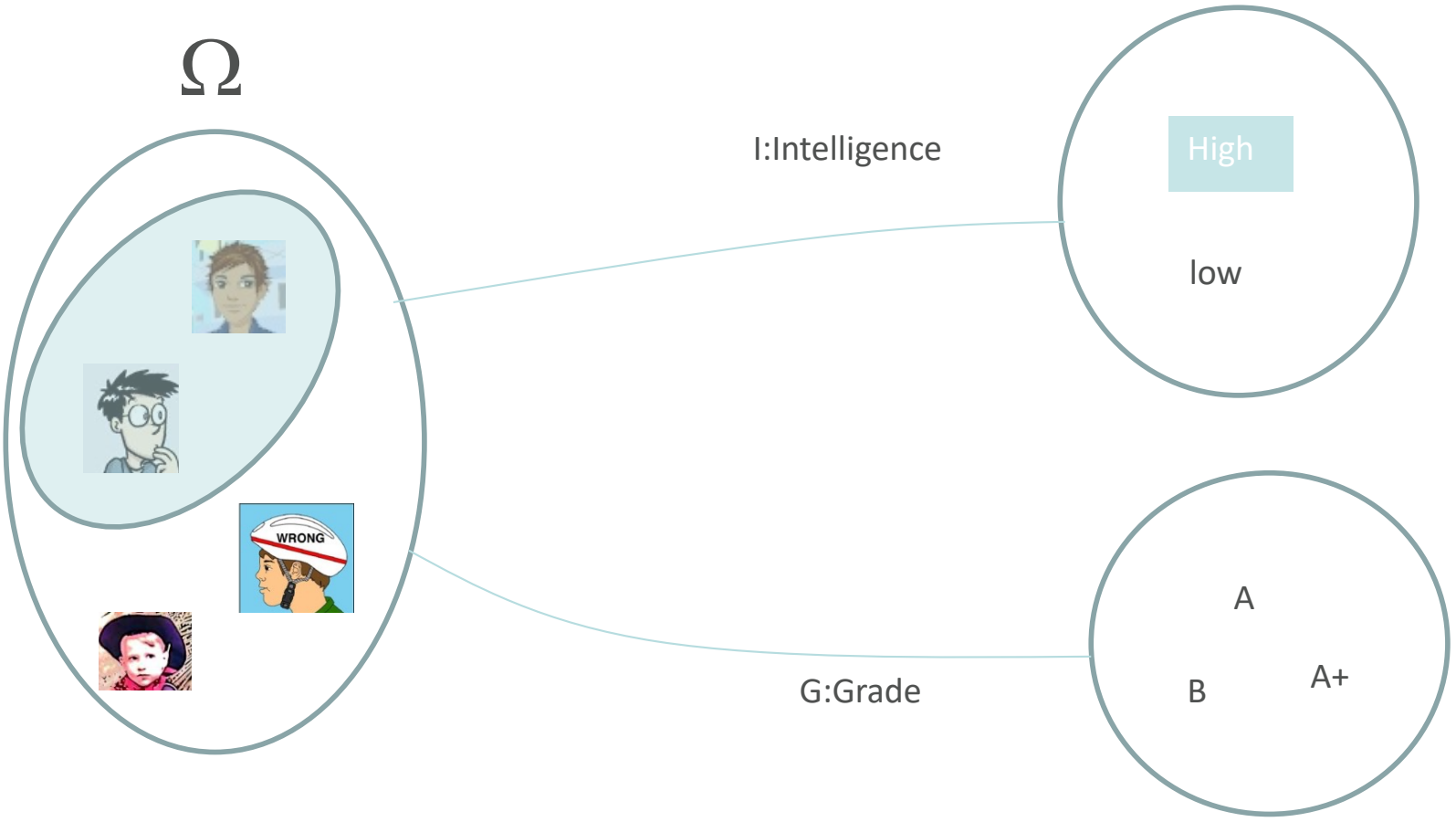
TECHNISCHE UNIVERSITÄT DARMSTADT

- Almost all the term we will be dealing with RV
- Concise way of specifying attributes of outcomes
- Modeling students (with Grade and Intelligence):
  - $\Omega =$ all possible students with grade and intelligence
  - What are events
    - Grade_A = all students with grade A
    - Grade_B = all students with grade B
    - Intelligence_High = … with high intelligence
  - Very cumbersome
  - We need "functions" that maps from Ω to an attribute space.

Random Variables



Random Variables



P(I = high) = P({all students whose intelligence is high})

Probability Review

- Events and Event spaces
- Random variables
- Joint probability distributions
  - Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties
  - Independence, conditional independence
- Examples
- Moments

Joint Probability Distribution

- **Random variables encode attributes**
- **Not all possible combination of attributes are equally likely**
  - **Joint probability distributions quantify this**

- P(X= x, Y= y) = P(x, y)
  - How probable is it to observe these two attributes together?
  - How can we manipulate joint probability distributions?

One of the most important rules of the class: Chain Rule

TECHNISCHE UNIVERSITÄT DARMSTADT

- Always true
  - $P(x,y,z) = p(x)\, p(y|x)\, p(z|x,y)$
    $= p(z)\, p(y|z)\, p(x|y,z)$
    $= \ldots$

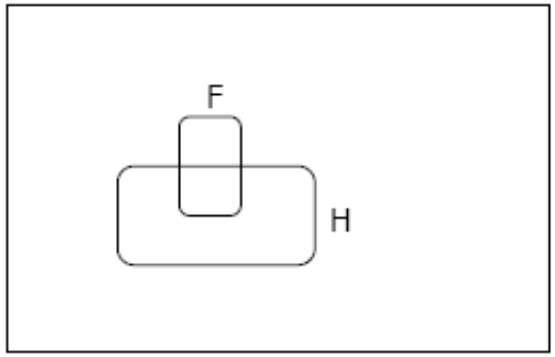
Conditional Probability

events

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

But we will always write it this way:

$$P(x \mid y) = \frac{P(x, y)}{P(y)}$$

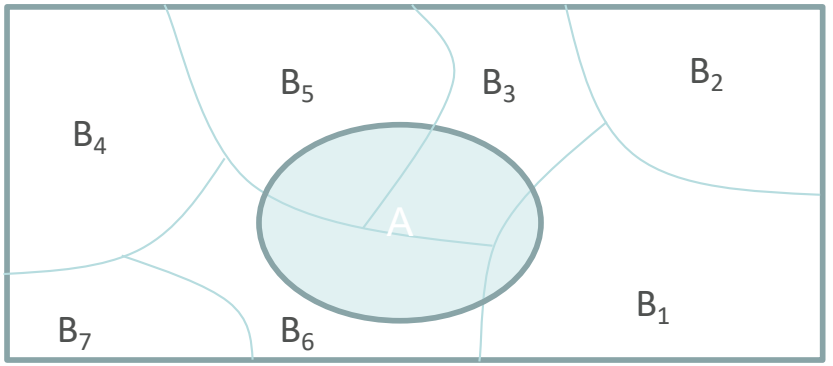


Marginalization

- We know P(X,Y), what is P(X=x)?
- We can use the low of total probability, why?

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

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



Marginalization Cont.

- Another example

$$P(x) = \sum_{y,z} P(x,y,z)$$

$$= \sum_{z,y} P(y,z) P(x \mid y,z)$$

One of the most important rules of the class: Bayes Rule

- Assume, we know that P(smart) = .7
  - **If we also know that the students grade is A+, then how does this affect our belief about his intelligence, P(smart|A+)?**

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

  - Where does this come from?

Bayes Rule cont.

- You can condition on more variables

$$P(x \mid y,z) = \frac{P(x \mid z)P(y \mid x,z)}{P(y \mid z)}$$

Probability Review

- Events and Event spaces
- Random variables
- Joint probability distributions
  - Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties
  - Independence, conditional independence
- Examples
- Moments

Independence

- **X is independent of Y** means that **knowing Y does not change our belief about X.**
- If they are independent, then
  - $P(X|Y=y) = P(X)$
  - [Definition] $P(X=x, Y=y) = P(X=x)\,P(Y=y)$
  - The above should hold for all x, y
  - **It is symmetric and written as $X \perp Y$**

CI: Conditional Independence

- **RV are rarely independent but we can still leverage local structural properties like CI.**
- **$X \perp Y \mid Z$** if **once Z is observed, knowing the value of Y does not change our belief about X**
  - The following should hold for all x,y,z
  - $P(X=x \mid Z=z, Y=y) = P(X=x \mid Z=z)$
  - $P(Y=y \mid Z=z, X=x) = P(Y=y \mid Z=z)$
  - $P(X=x, Y=y \mid Z=z) = P(X=x \mid Z=z)\, P(Y=y \mid Z=z)$

We call these **factors** : very useful concept !!

Properties of CI

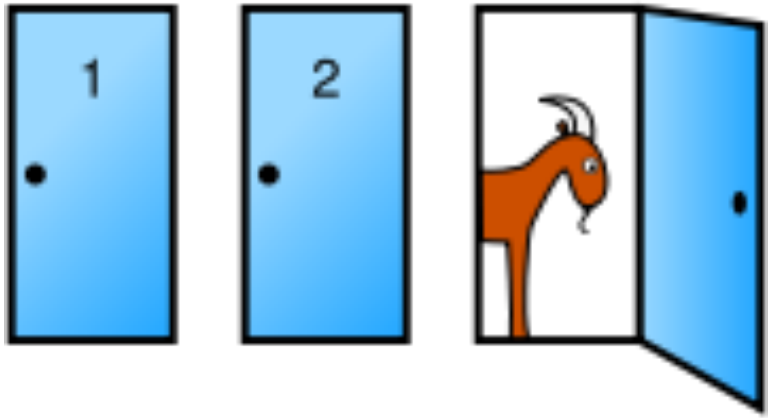
- **Symmetry:**
  - $(X \perp Y \mid Z) \Rightarrow (Y \perp X \mid Z)$
- **Decomposition:**
  - $(X \perp Y,W \mid Z) \Rightarrow (X \perp Y \mid Z)$
- **Weak union:**
  - $(X \perp Y,W \mid Z) \Rightarrow (X \perp Y \mid Z,W)$
- **Contraction:**
  - $(X \perp W \mid Y,Z)$ & $(X \perp Y \mid Z) \Rightarrow (X \perp Y,W \mid Z)$
- **Intersection:**
  - $(X \perp Y \mid W,Z)$ & $(X \perp W \mid Y,Z) \Rightarrow (X \perp Y,W \mid Z)$
  - Only for positive distributions!
  - $P(\alpha)>0$, 8α, $\alpha\neq$;

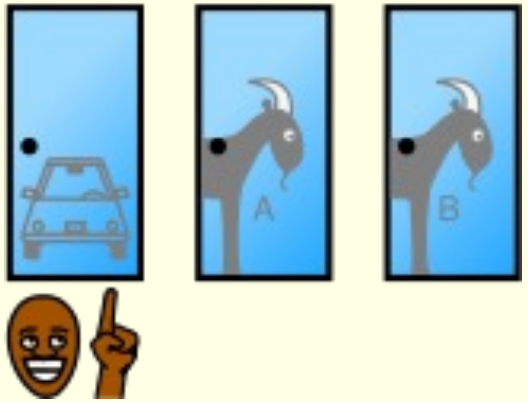
Probability Review

- Events and Event spaces
- Random variables
- Joint probability distributions
  - Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties
  - Independence, conditional independence
- Examples
- Moments

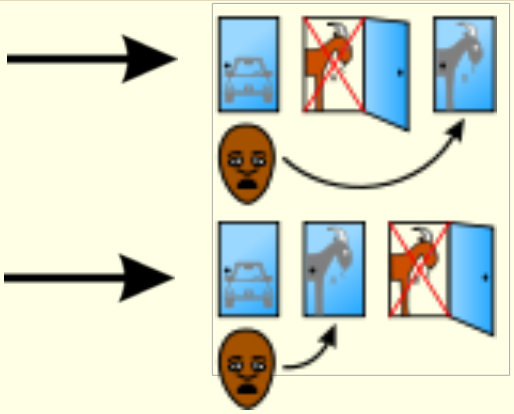
Monty Hall Problem

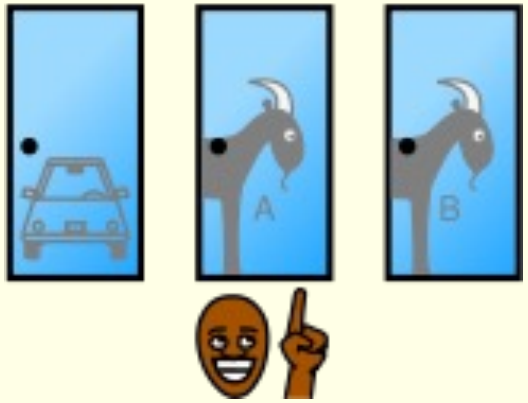
- You're given the choice of three doors: Behind one door is a car; behind the others, goats.
- You pick a door, say No. 1
- The host, who knows what's behind the doors, opens another door, say No. 3, which has a goat.
- Do you want to pick door No. 2 instead?



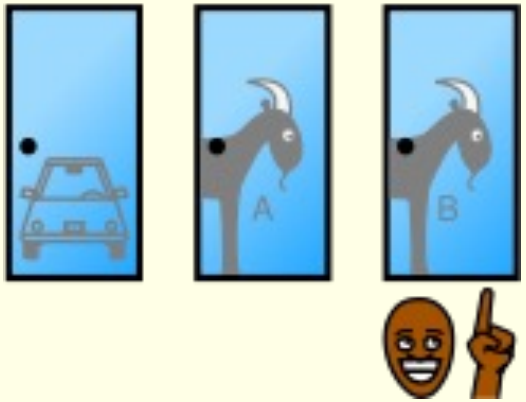


Host reveals Goat A
or
Host reveals Goat B

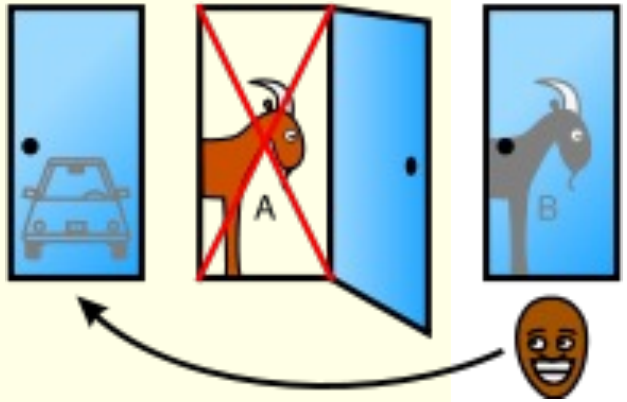




Host must reveal Goat B



Host must reveal Goat A



Monty Hall Problem in terms of probabilities having the Bayes rule in mind

- C_i : the car is behind door *i*, *i* = 1, 2, 3
- $P(C_i) = 1/3$
- H_{ij}: the host opens door *j* after you pick door *i*
- $$P(H_{ij} | C_k) = \begin{cases} 0 & i = j \\ 0 & j = k \\ 1/2 & i = k \\ 1 & i \neq k, j \neq k \end{cases}$$

  Your door will not be opened

  Host does not want to show the car

  You picked the car. The host will randomly select among the two goats

  The host will select the other goat

Monty Hall Problem: Bayes Rule

- WLOG, i=1, j=3

$$P(C_1 | H_{13}) = \frac{P(H_{13} | C_1) P(C_1)}{P(H_{13})}$$

$$P(H_{13} | C_1) P(C_1) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$$

Monty Hall Problem: Bayes Rule cont.

$$
\begin{aligned}
P(H_{13}) &= P(H_{13}, C_1) + P(H_{13}, C_2) + P(H_{13}, C_3) \\
&= P(H_{13} | C_1) P(C_1) + P(H_{13} | C_2) P(C_2) \\
&= \frac{1}{6} + 1 \cdot \frac{1}{3} \\
&= \frac{1}{2}
\end{aligned}
$$

$$
P(C_1 | H_{13}) = \frac{1/6}{1/2} = \frac{1}{3}
$$

Monty Hall Problem: Bayes Rule cont.

$$P(C_1|H_{13}) = \frac{1/6}{1/2} = \frac{1}{3}$$

$$P(C_2|H_{13}) = 1 - \frac{1}{3} = \frac{2}{3} > P(C_1|H_{13})$$

- *You should switch!*

Moments

- **Mean** (Expectation): $\mu = E(\mathrm{X})$
  - Discrete RVs: $E(\mathrm{X}) = \sum_{v_i} v_i \mathrm{P}(\mathrm{X} = v_i)$
  - Continuous RVs: $E(\mathrm{X}) = \int_{-\infty}^{+\infty} x f(x) dx$

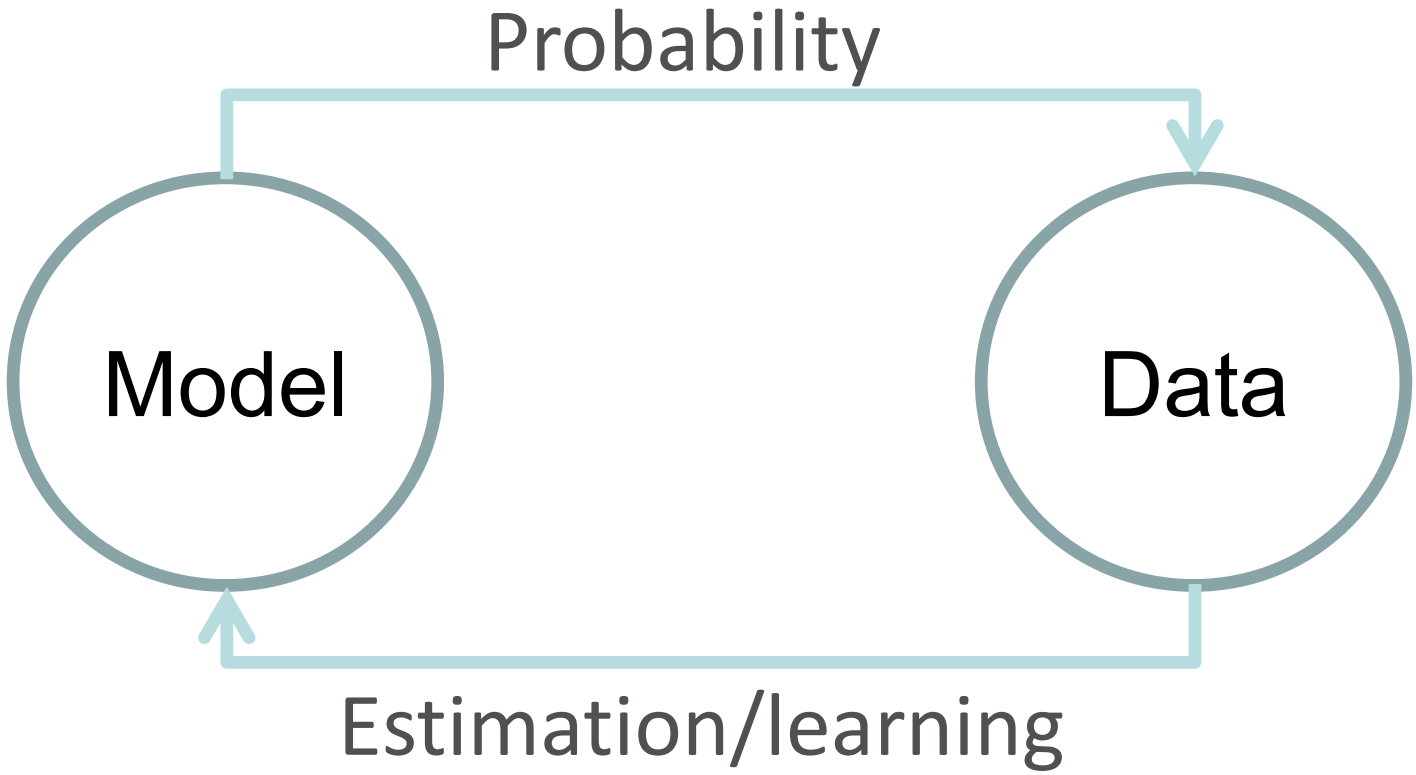
- **Variance**: $V(\mathrm{X}) = E(\mathrm{X} - \mu)^2$
  - Discrete RVs: $V(\mathrm{X}) = \sum_{v_i} (v_i - \mu)^2 \mathrm{P}(\mathrm{X} = v_i)$
  - Continuous RVs: $V(\mathrm{X}) = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx$

Properties of Moments

- **Mean**
  - $E(aX) = aE(X)$
  - $E(X+Y) = E(X) + E(Y)$
  - If X and Y are independent, $E(XY) = E(X) \cdot E(Y)$
- **Variance**
  - $V(aX+b) = a^2 V(X)$
  - If X and Y are independent,

$$V(X+Y) = V(X) + V(Y)$$

The Big Picture



Statistical Inference

- Given observations from a model
  - What (conditional) independence assumptions hold?
    - Structure learning
  - If you know the family of the model (ex, multinomial), what are the value of the parameters: MLE, Bayesian estimation.
    - Parameter learning

MLE

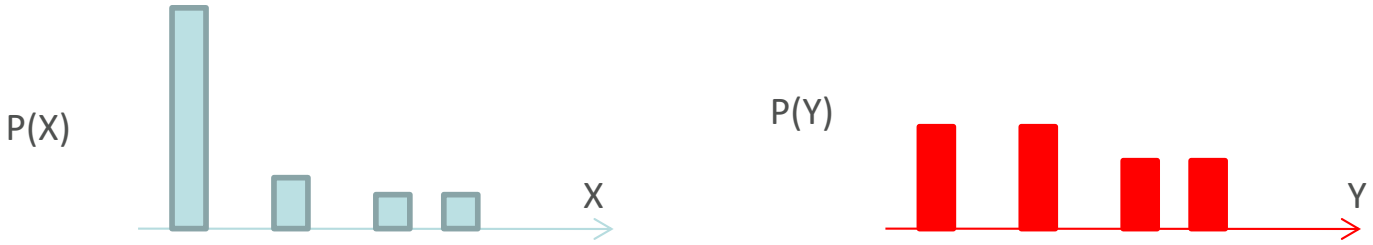
TECHNISCHE UNIVERSITÄT DARMSTADT

- Maximum Likelihood estimation
  - Example on board
    - Given N coin tosses, what is the coin bias (θ)?

- **Sufficient Statistics**:
  - Useful concept that we will make use later
  - In solving the above estimation problem, we only care about N_h, N_t , these are called the sufficient statistics of this model.
    - All coin tosses that have the same SS will result in the same value of θ

Information Theory

- P(X) encodes our uncertainty about X
  - Some variables are more uncertain that others



  - How can we quantify this intuition?
    - **Entropy: average number of bits required to encode X**

$$H_P(X) = E\left[\log \frac{1}{p(x)}\right] = \sum_x P(x) \log \frac{1}{P(x)}$$

Information Theory cont.

- **Entropy**: average number of bits required to encode X

$$H_P(X) = E\left[\log \frac{1}{p(x)}\right] = \sum_x P(x) \log \frac{1}{P(x)}$$

- We can define **conditional entropy** similarly

$$H_P(X \mid Y) = E\left[\log \frac{1}{p(x \mid y)}\right] = H_P(X,Y) - H_P(Y)$$

- We can also define the **chain rule for entropies** (not surprising)

$$H_P(X,Y,Z) = H_P(X) + H_P(Y \mid X) + H_P(Z \mid X,Y)$$

Mutual Information: MI

- Remember independence?
  - If X⊥Y then knowing Y won't change our belief about X
  - Mutual information can help quantify this! (not the only way though)
- MI: $I_P(X;Y) = H_P(X) - H_P(X \mid Y)$
  - Symmetric
  - I(X;Y) = 0 iff, X and Y are independent!

Continuous Random Variables

- What if X is continuous?
- Probability density function (pdf) instead of probability mass function (pmf)
- A pdf is any function $f\left(x\right)$ that describes the probability density in terms of the input variable *x*.

PDF

- Properties of pdf
  - $f(x) \geq 0, \forall x$
  - $\int_{-\infty}^{+\infty} f(x) = 1$
  - $f(x) \leq 1$???
- Actual probability can be obtained by taking the integral of pdf
  - E.g. the probability of X being between 0 and 1 is

$$\mathrm{P}(0 \leq X \leq 1) = \int_0^1 f(x)dx$$

Cumulative Distribution Function

- $F_X(v) = P(X \leq v)$
- Discrete RVs
  - $F_X(v) = \sum_{v_i} P(X = v_i)$
- Continuous RVs
  - $F_X(v) = \int_{-\infty}^{v} f(x)\,dx$
  - $\frac{d}{dx} F_X(x) = f(x)$

What is next: Bayesian Networks

- **One of the most exciting recent advancements in statistical AI**
- **Compact representation for exponentially-large probability distributions**
- Fast marginalization too
- **Exploit conditional independencies**