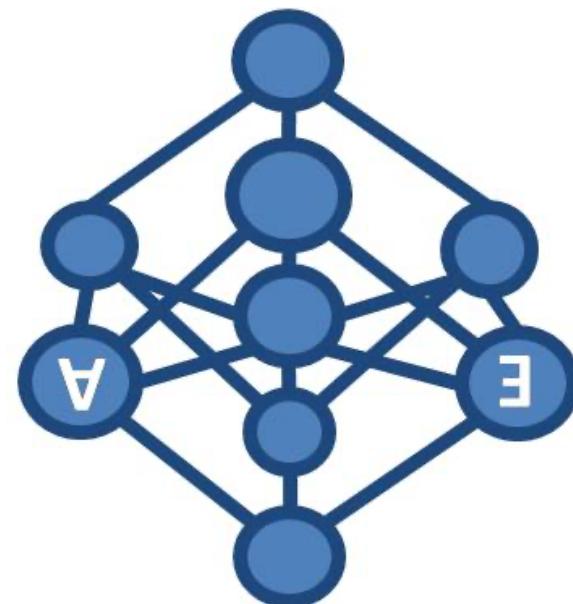


Probabilistic Graphical Models*

Introduction



TECHNISCHE
UNIVERSITÄT
DARMSTADT



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

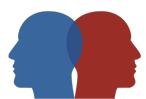


This Course

One of the **most exciting developments** in

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

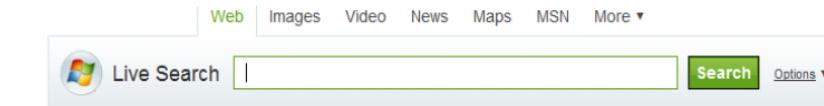
in the last two (or three, or more) decades...



SOME EXAMPLE APPLICATIONS



Web Search



YAHOO!

Web | Images | Video | Local | Shopping | more ▾

Search

Google™

Google Search I'm Feeling Lucky Advanced Search Preferences Language Tools

Relative Feedback:
Clicks reflect preference between observed links.

(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>

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

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-efficient inference in relational 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). Efficient 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-efficient inference in relational 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). Efficient inference. In Proceedings of the Twenty-First National Conference on Artificial Intelligence.

[Domingos et al.]

Entity Resolution

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

Singla, P., & Domingos, P. (2006). Memory-efficient inference in relational 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. Hofnung (2006). Efficient inference. In Proceedings of the Twenty-First National Conference on Artificial Intelligence.

[Domingos et al.]

Entity Resolution

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

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

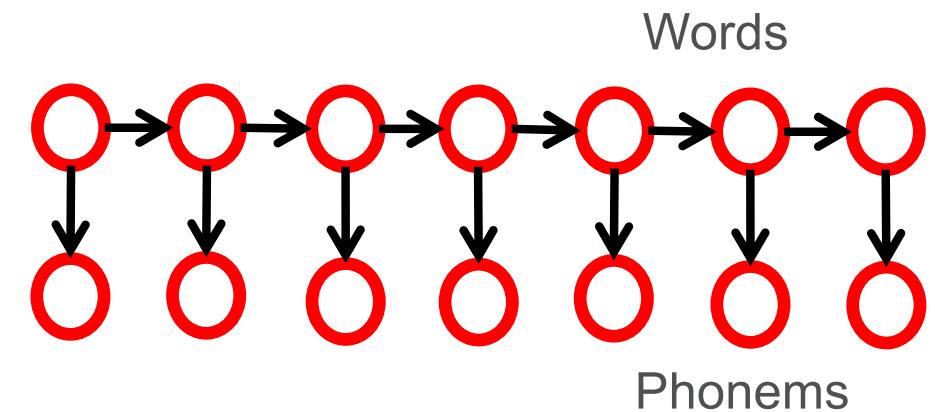
H. Poon & P. Domingos, Sound and Efficient Inference with Probabilistic and Deterministic Dependencies", in Proc. AAAI-06, Boston, MA, 2006

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

[Domingos et al.]

Speech Recognition

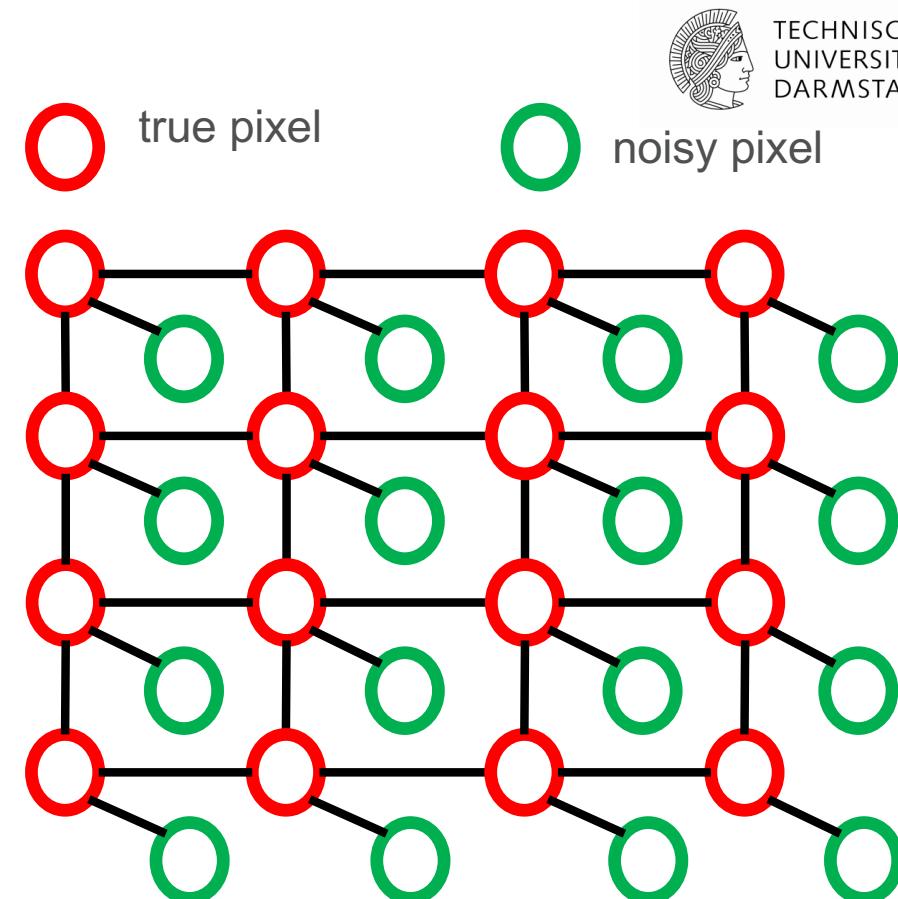
- Estimate the words from audio signal
- Hidden Markov Model



„He ate the cookies on the couch“

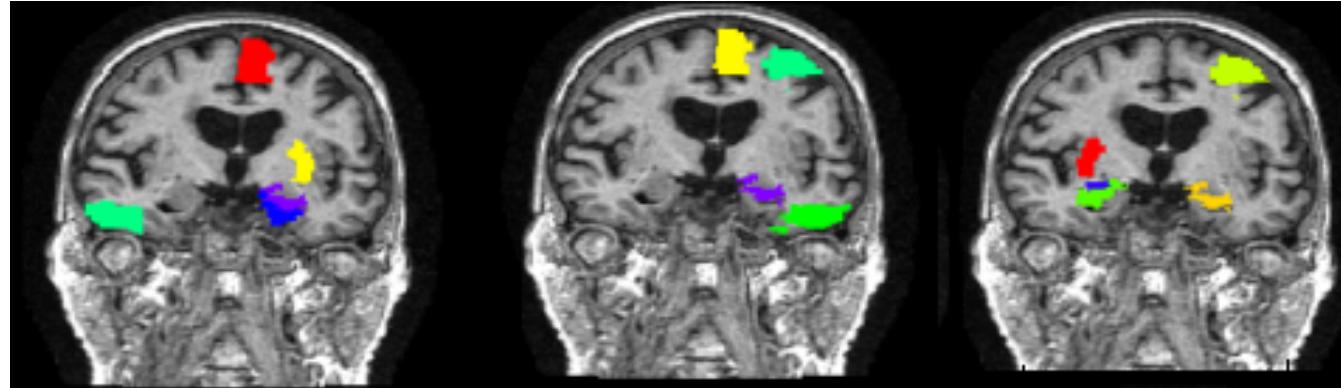


Image Denoising



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

Helping to Cure Alzheimer

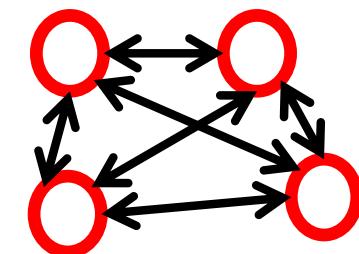
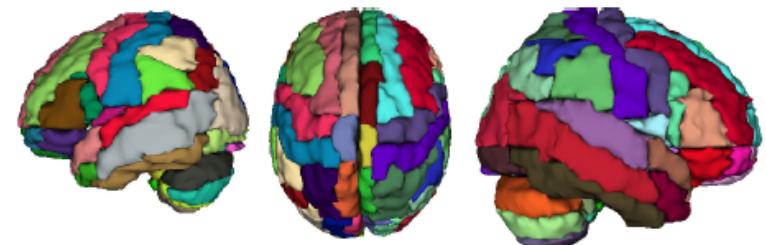


Alzheimer

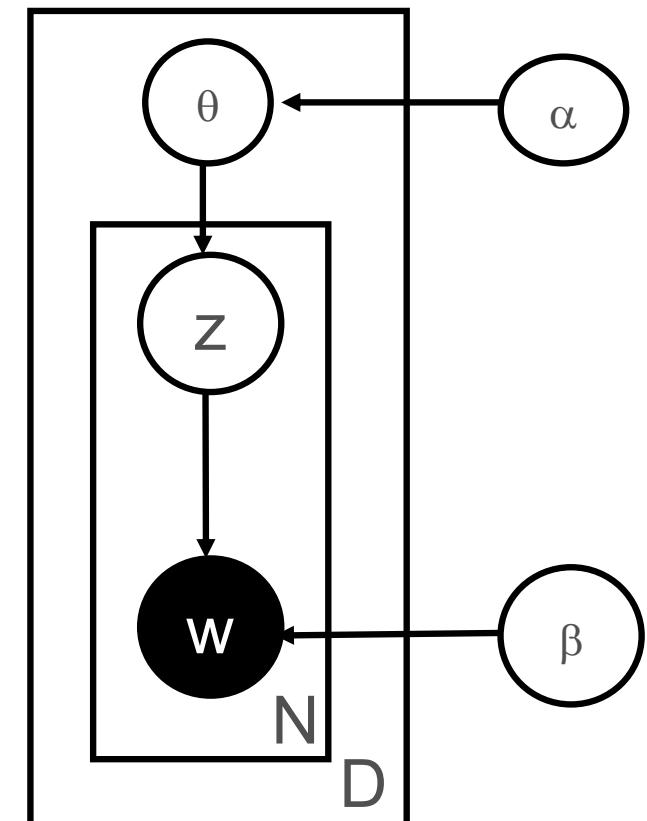
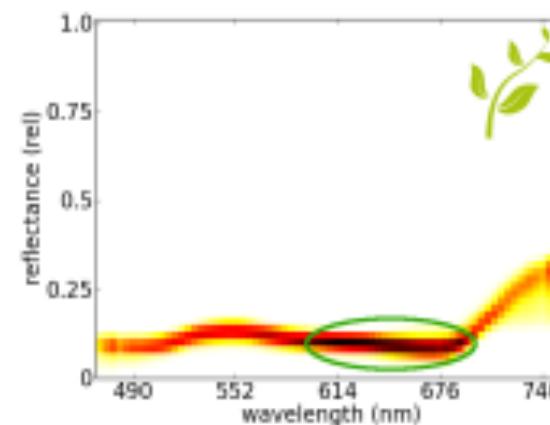
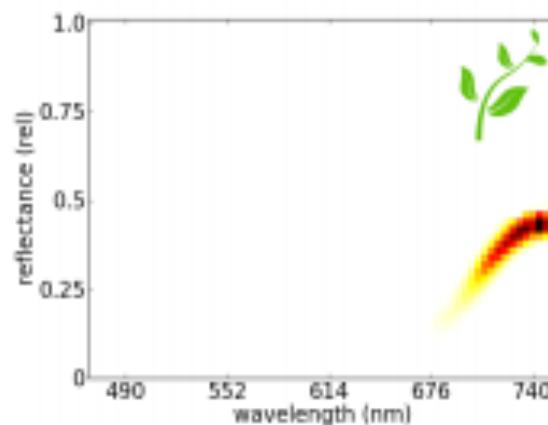
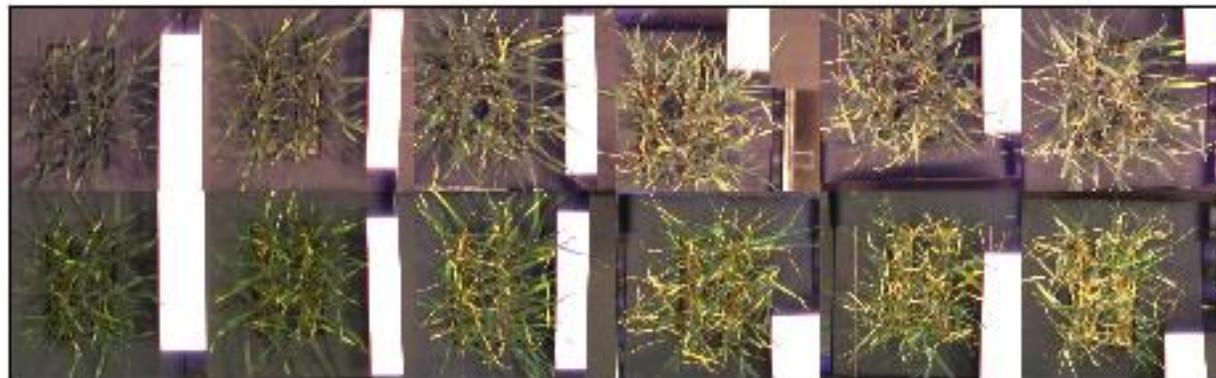
Mild-cognitive
impairment

Cognitive normal

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



Precision Agriculture

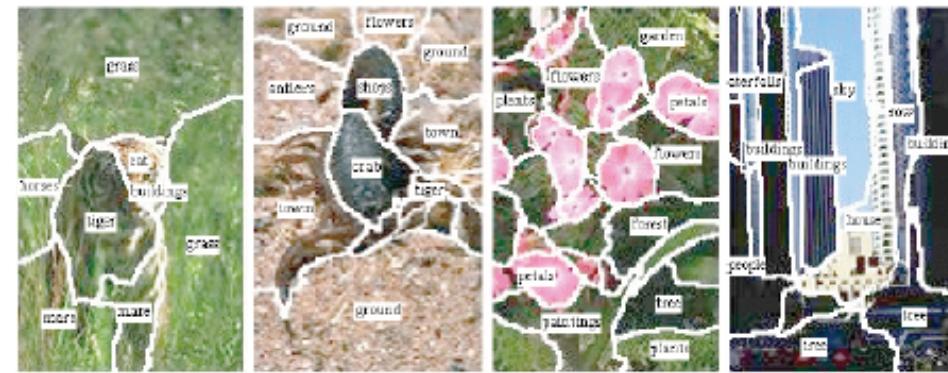
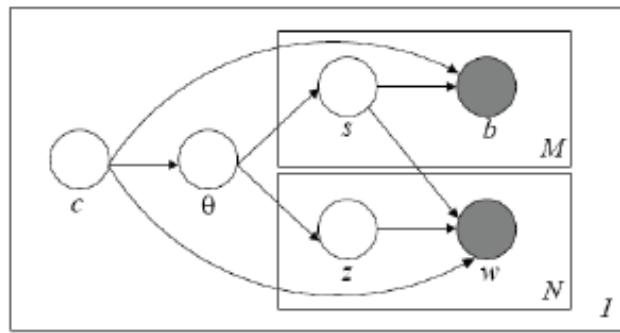


- How do plants react on stress?
- Non-parametric Bayesian models such as Latent Dirichlet Allocation

Images and Text

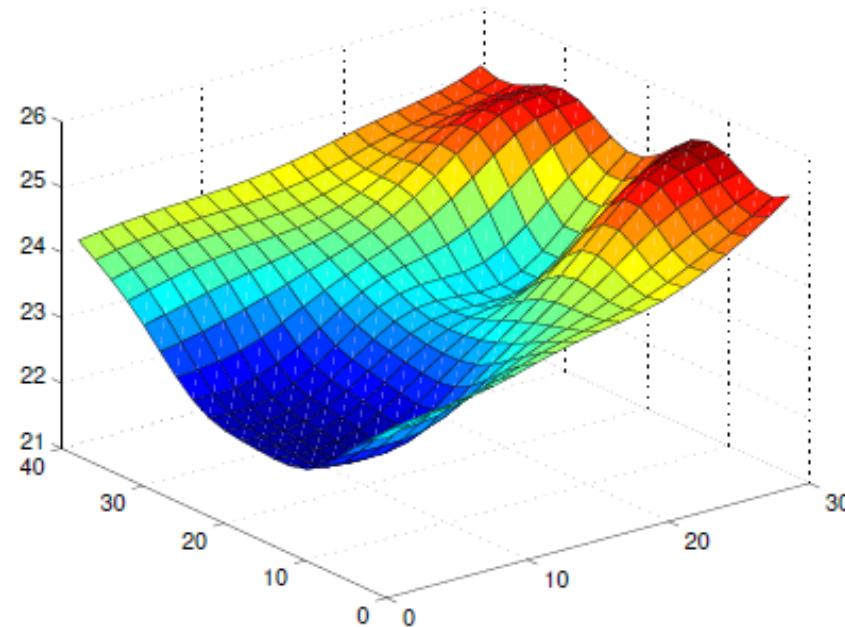
Hierarchical Bayesian models

[Barnard et al.]

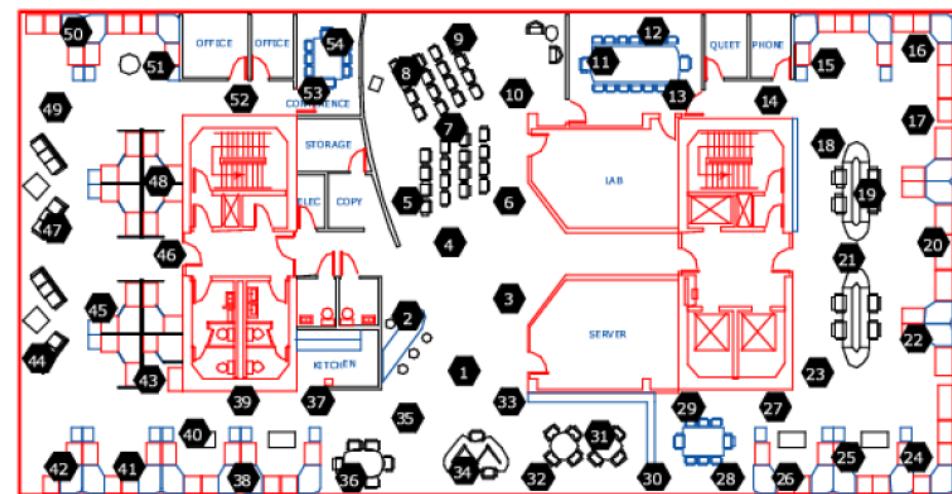


- Describe the objects captured by an image

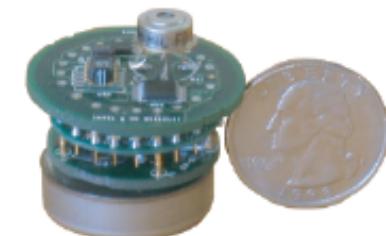
Modeling Sensor Data



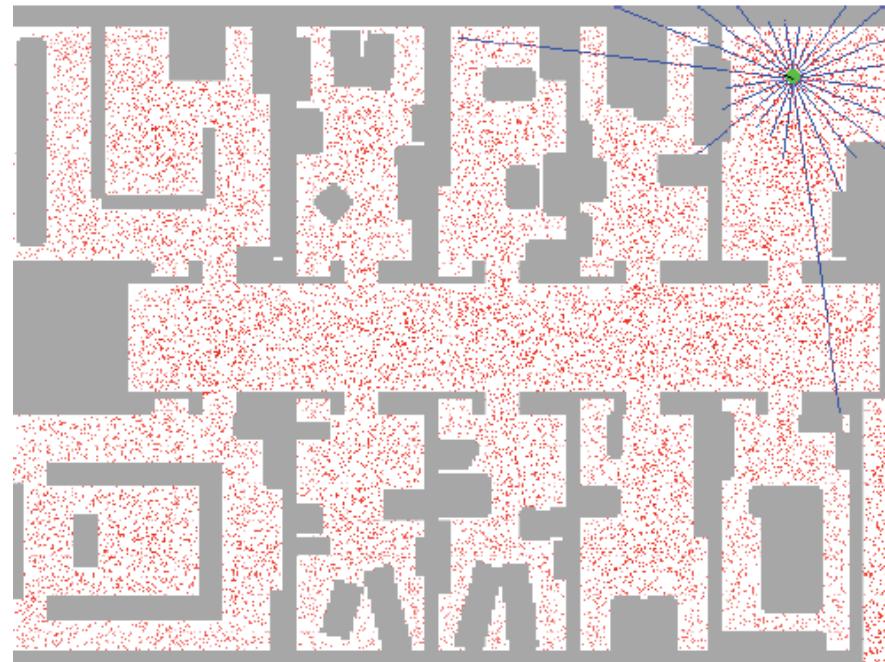
[Guestrin et al.]



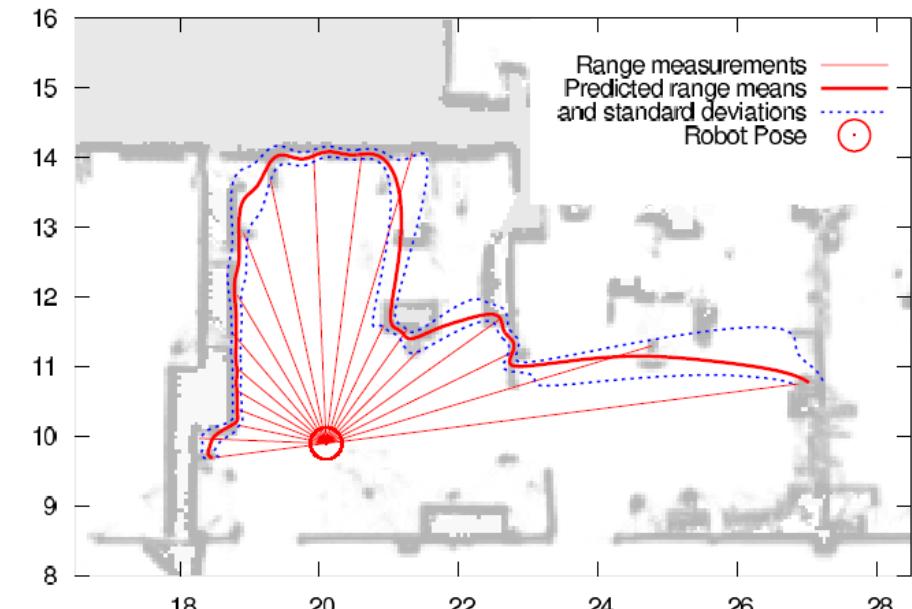
- Where to place sensors to get the best model?
- Non-parametric Bayesian models such as Gaussian Processes



Tracking and Robot Localization



[Fox et al.]

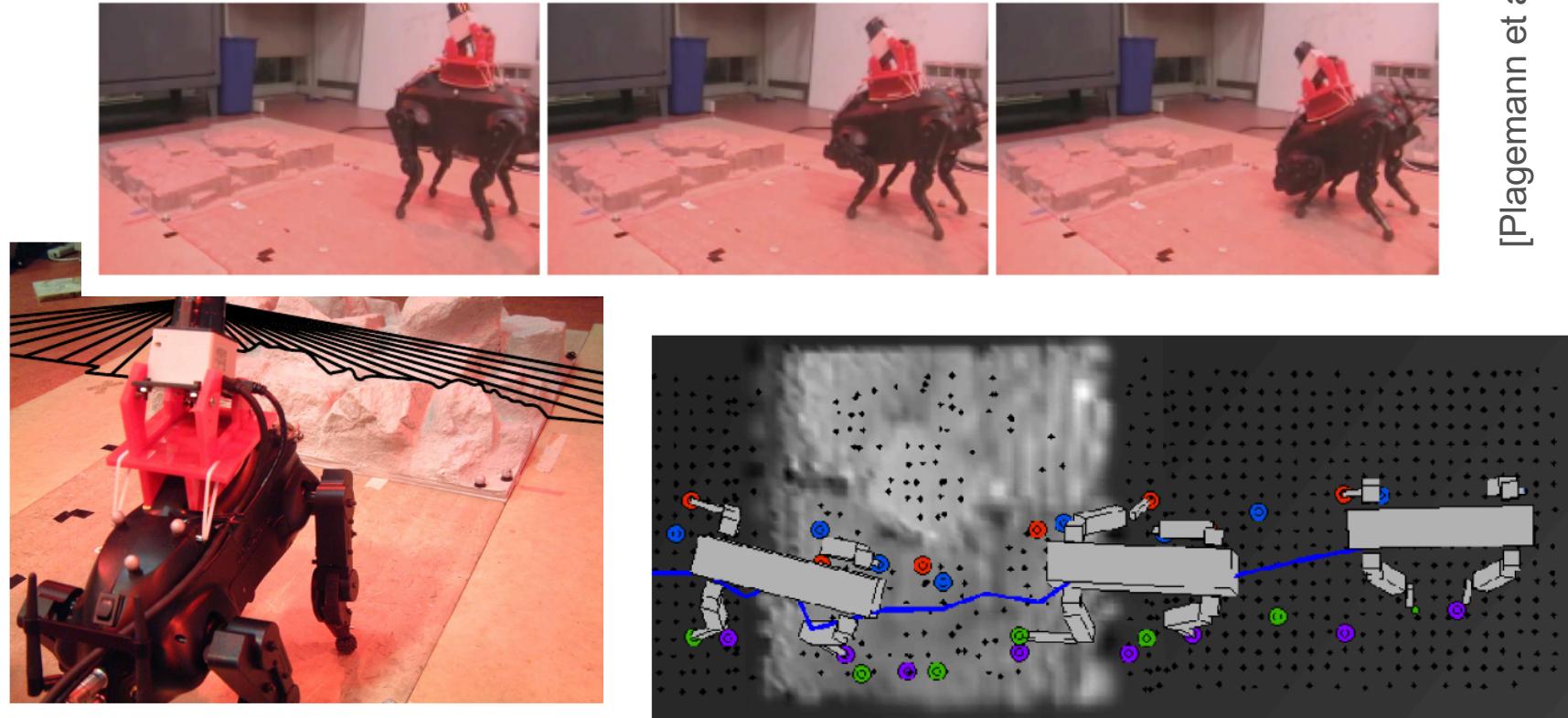


[Plagemann et al.]

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



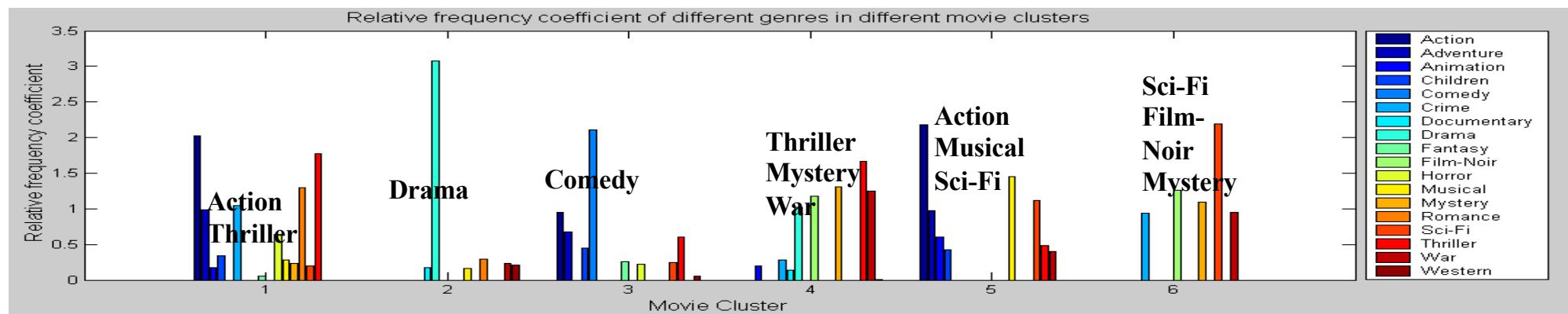
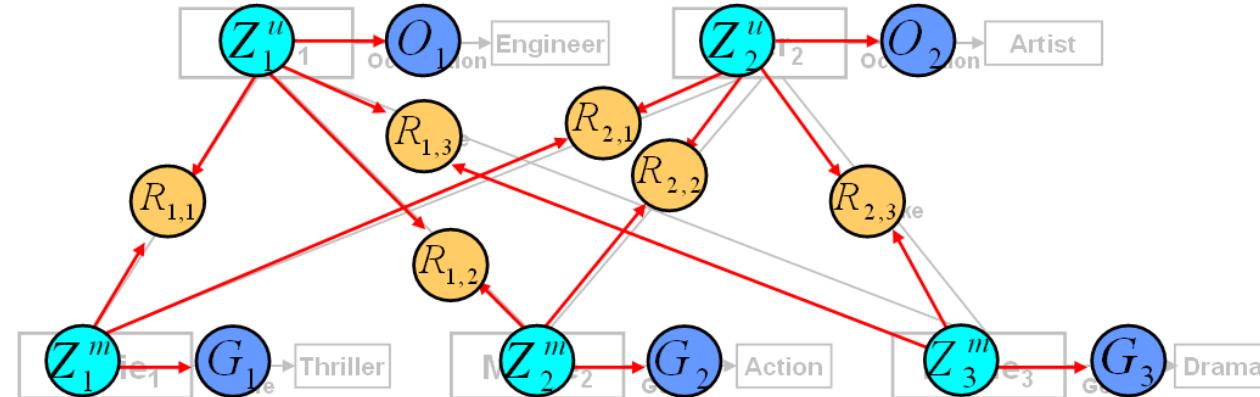
3D Terrain Modeling



[Plagemann et al.]

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

Social Network Analysis



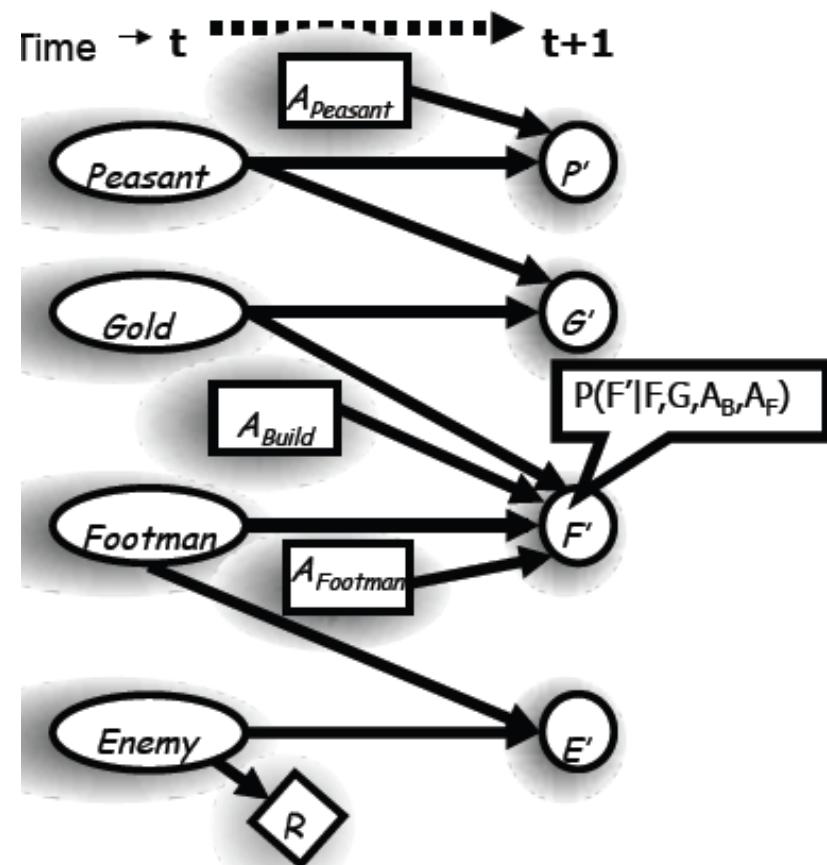
- Do social groups share interests?
- Collaborative filtering



Planning Under Uncertainty

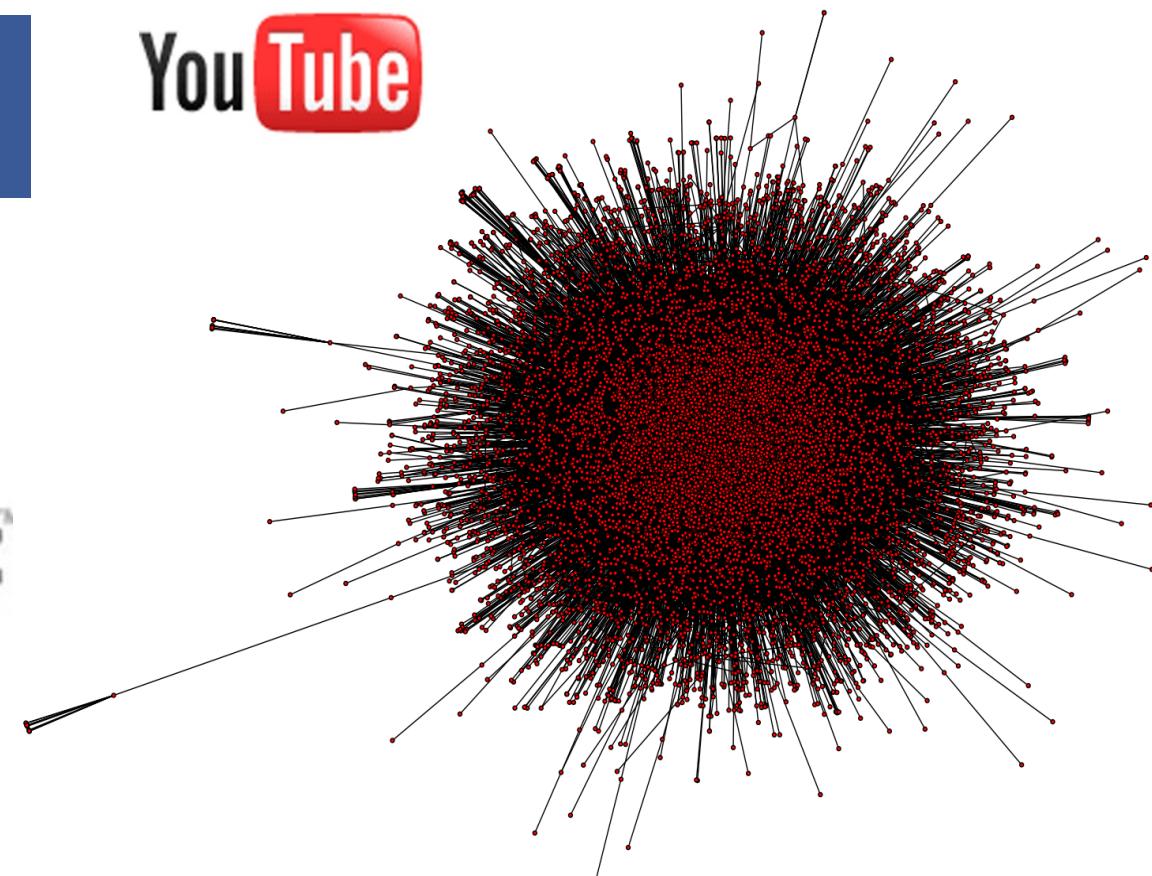
- How do act optimally in noisy environments?
- (Partially observed) Markov Decision Processes

Dynamic Bayesian networks
Factored Markov decision problems



[Guestrin et al.]

How do you efficiently broadcast information?



[Kersting et al. UAI 2010]



Combinatorial Problems



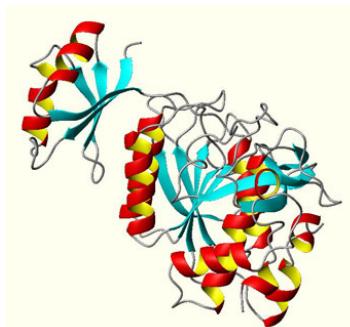
TECHNISCHE
UNIVERSITÄT
DARMSTADT



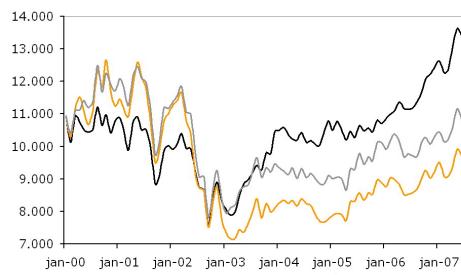
logistics



chip design



protein folding



portfolio optimization

TimeTable		03:37	23:32:23
たま	プラ	サ	駅 東急 渋谷方面(平日)
19	02	07	09 15 20 26 28 33 40 45
			47 53 59
20	06	08	14 21 28 30 36 42 50 56
21	02	08	14 20 27 34 41 48 55
22	02	11	20 30 39 49
23	00	12	25 36 47
0	01	15	44
23:36	水田		
23:47	青山		
00:01	渋谷		
00:15	二子		

03:37

timetabling



production planning



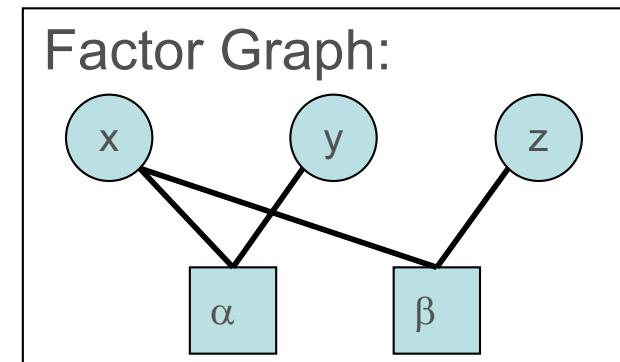
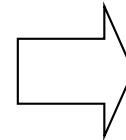
air traffic routing

Encoding CSPs

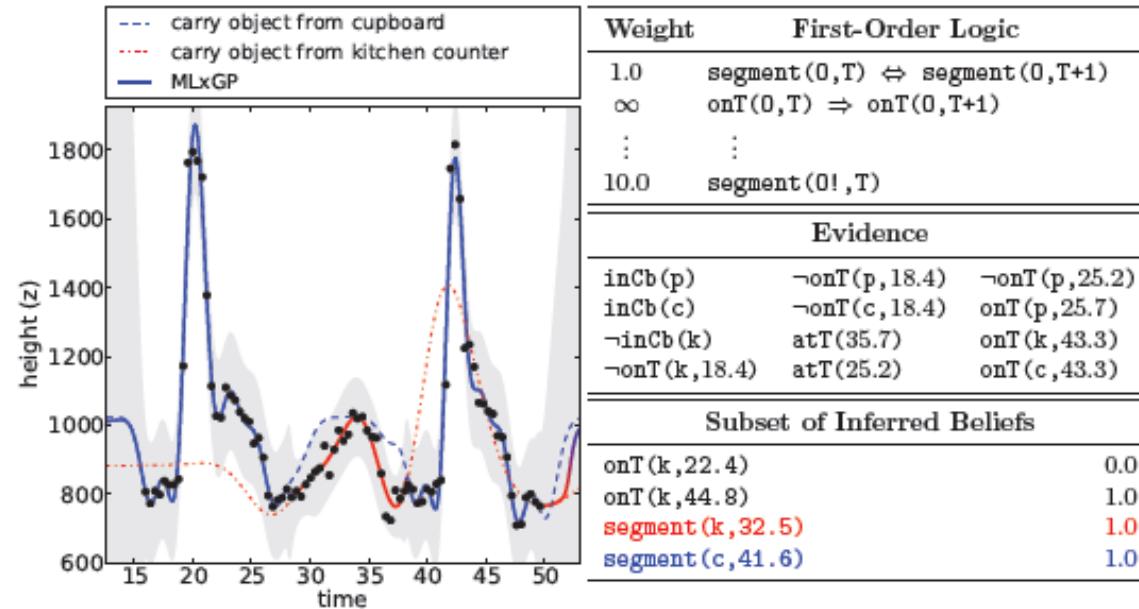
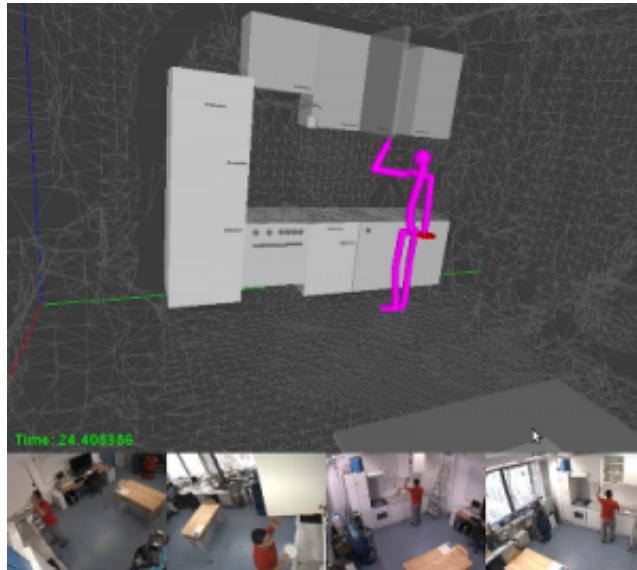
- A CSP is a problem of finding a **configuration** (values of discrete variables) that is globally **consistent** (all constraints are satisfied)
- One can visualize the connections between variables and constraints in so called **factor graph**:

e.g. SAT Problem:

$$\left(\underbrace{\mathbf{x} \vee \mathbf{y}}_{\alpha} \right) \wedge \left(\underbrace{\neg \mathbf{x} \vee \mathbf{z}}_{\beta} \right)$$

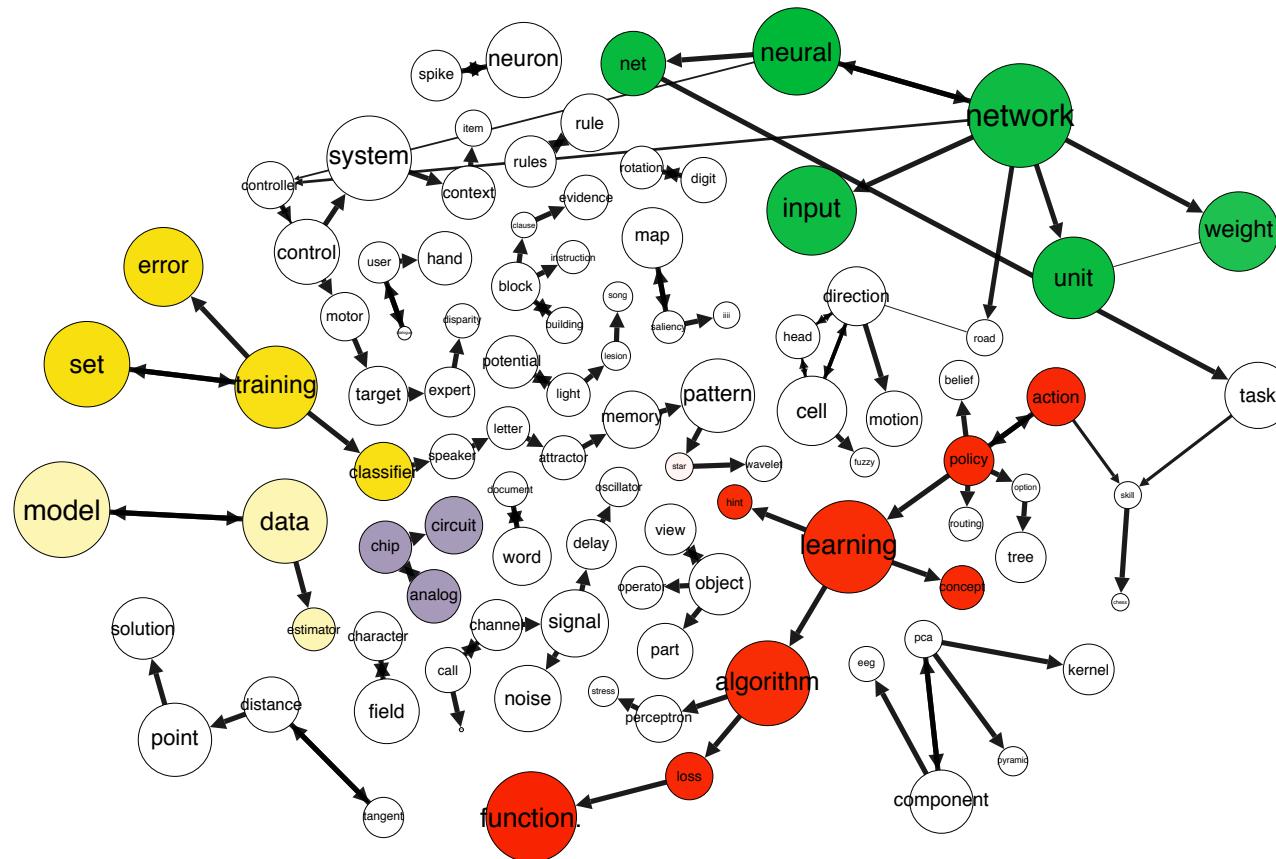


Machines Reading Regression Data



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

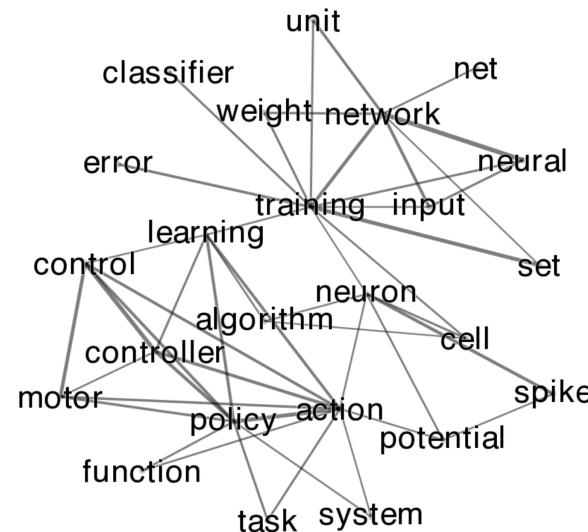
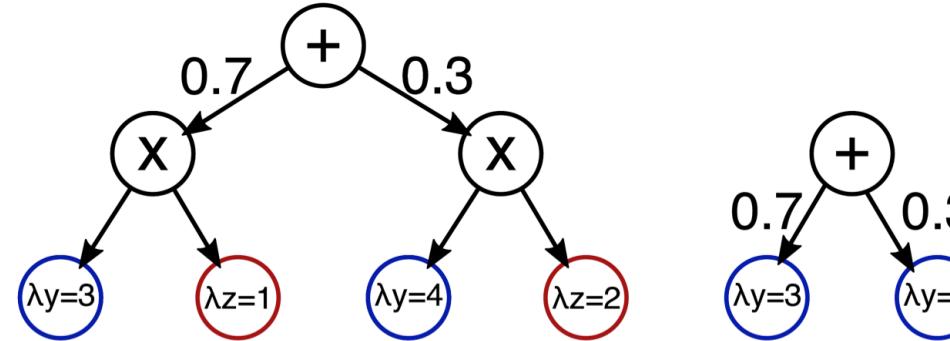
Machines Reading Texts



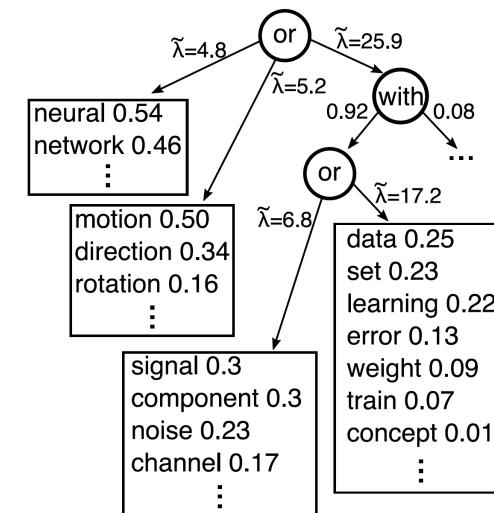
Let the machine read and understand text documents.
Here you see the main dependencies among words
estimated on a corpus of scientific papers



Deep and Tractable Probabilistic Models



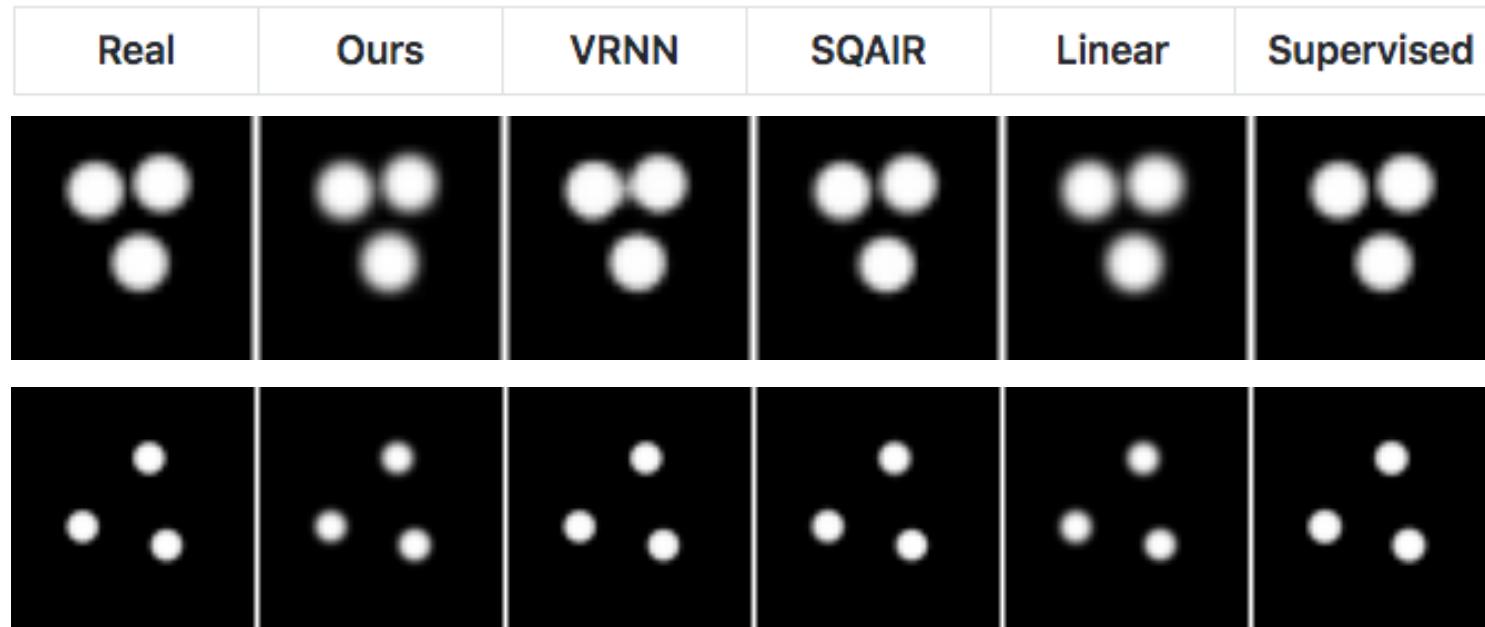
Mutual Information
(NIPS corpus)



Poisson Multinomial SPN
= hierachical topic model



Learning Physics using Deep Probabilistic Models



... 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

cigdem.turan@cs.tu-darmstadt.de
kersting@cs.tu-darmstadt.de



Exercise Sessions

- 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, ZOOM**



Hands-on project!

- 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, ...



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**



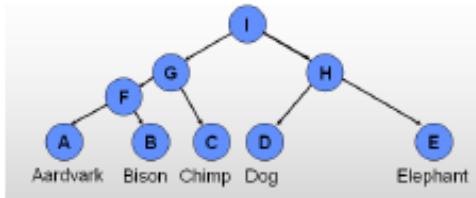
Enjoy !!

- 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?

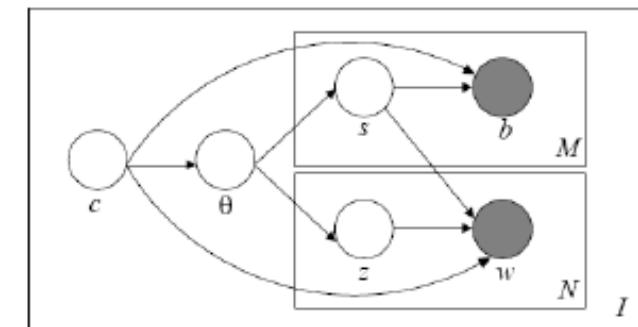
- What does the model mean/implies/assume?

- 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?

- **Representation**
 - Graphical models represent exponentially 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
 - Probabilities
 - Independence
 - Etc.
- Next time
 - Two nodes make a Bayesian network
 - Naive Bayes
 - Bayesian networks



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
- Continuous distributions, Moments, Entropy, MLE, ...



Sample space and Events

- Ω : **Sample Space**, result of an experiment
 - If you toss a coin twice $\Omega = \{\text{HH}, \text{HT}, \text{TH}, \text{TT}\}$
 - Event: a subset of Ω
 - First toss is head = $\{\text{HH}, \text{HT}\}$
 - **S: event space, a set of events:**
 - Closed under finite union and complements
 - Entails other binary operation: union, diff, etc.
 - Contains the empty event and Ω

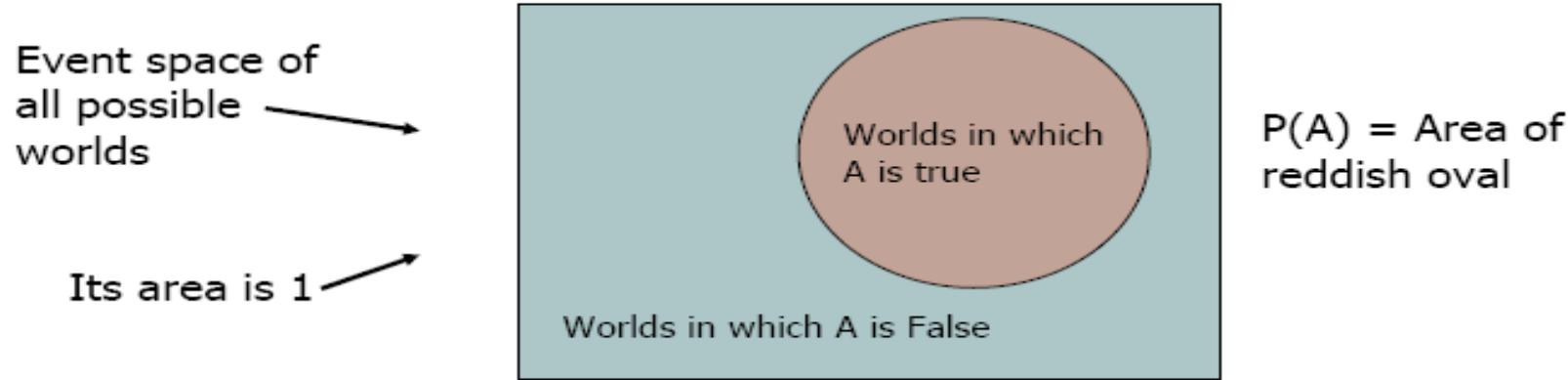


Probability Measure

- Defined over (Ω, \mathcal{S}) s.t.
 - $P(\alpha) \geq 0$ for all α in \mathcal{S}
 - $P(\Omega) = 1$
 - If α, β are disjoint, then
 - $P(\alpha \cup \beta) = P(\alpha) + P(\beta)$
- We can deduce other axioms from the above ones
 - Ex.: $P(\alpha \cup \beta)$ for non-disjoint event



Illustration

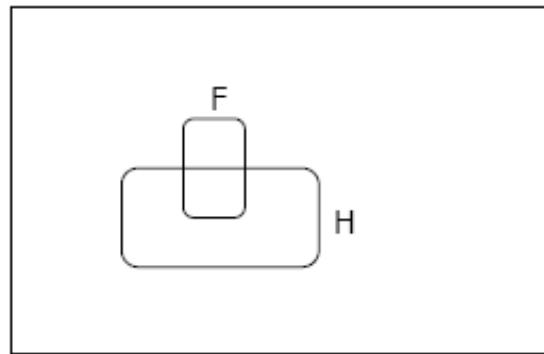


- We can go on and define conditional probability, using the above visualization

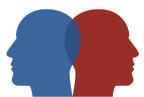


Conditional Probability

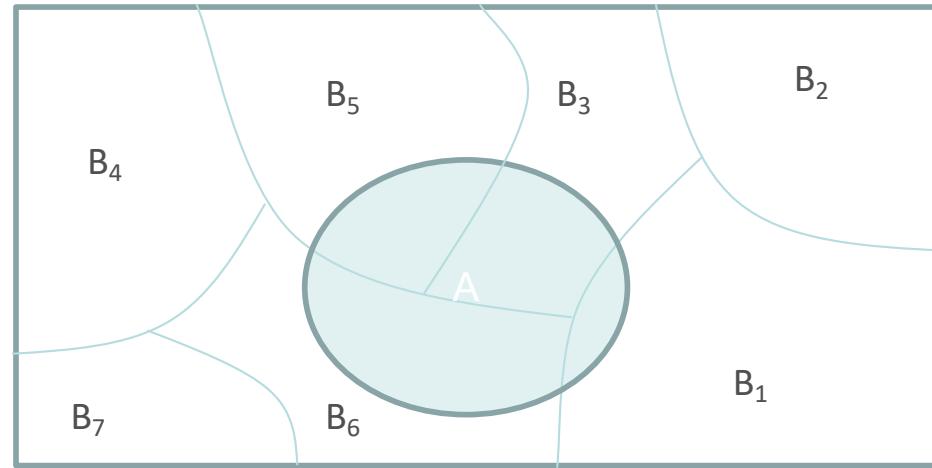
$P(F|H)$ = Fraction of worlds in which H is true that also have F true



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



Rule of total probability



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

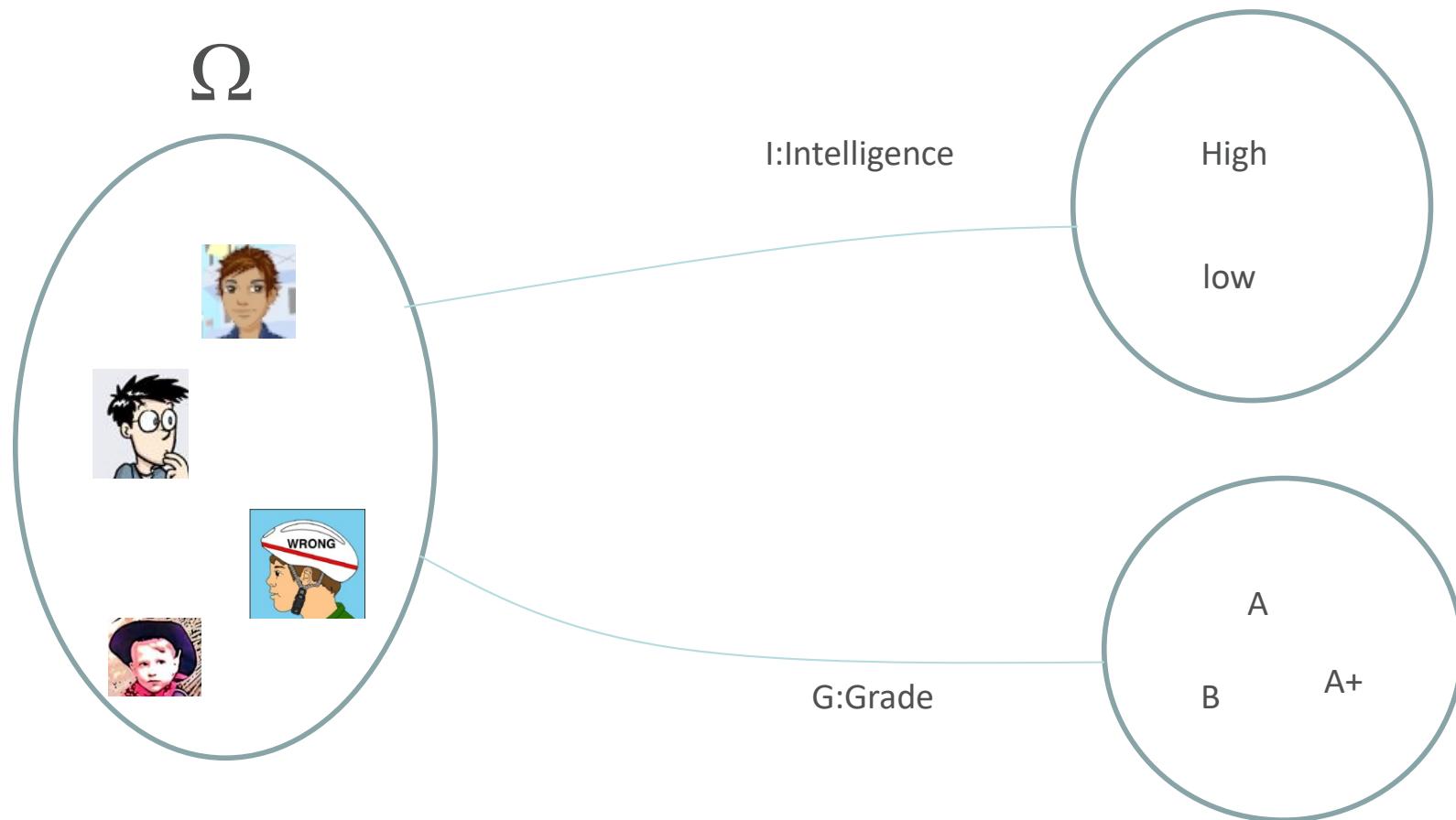


From Events to Random Variable (RV)

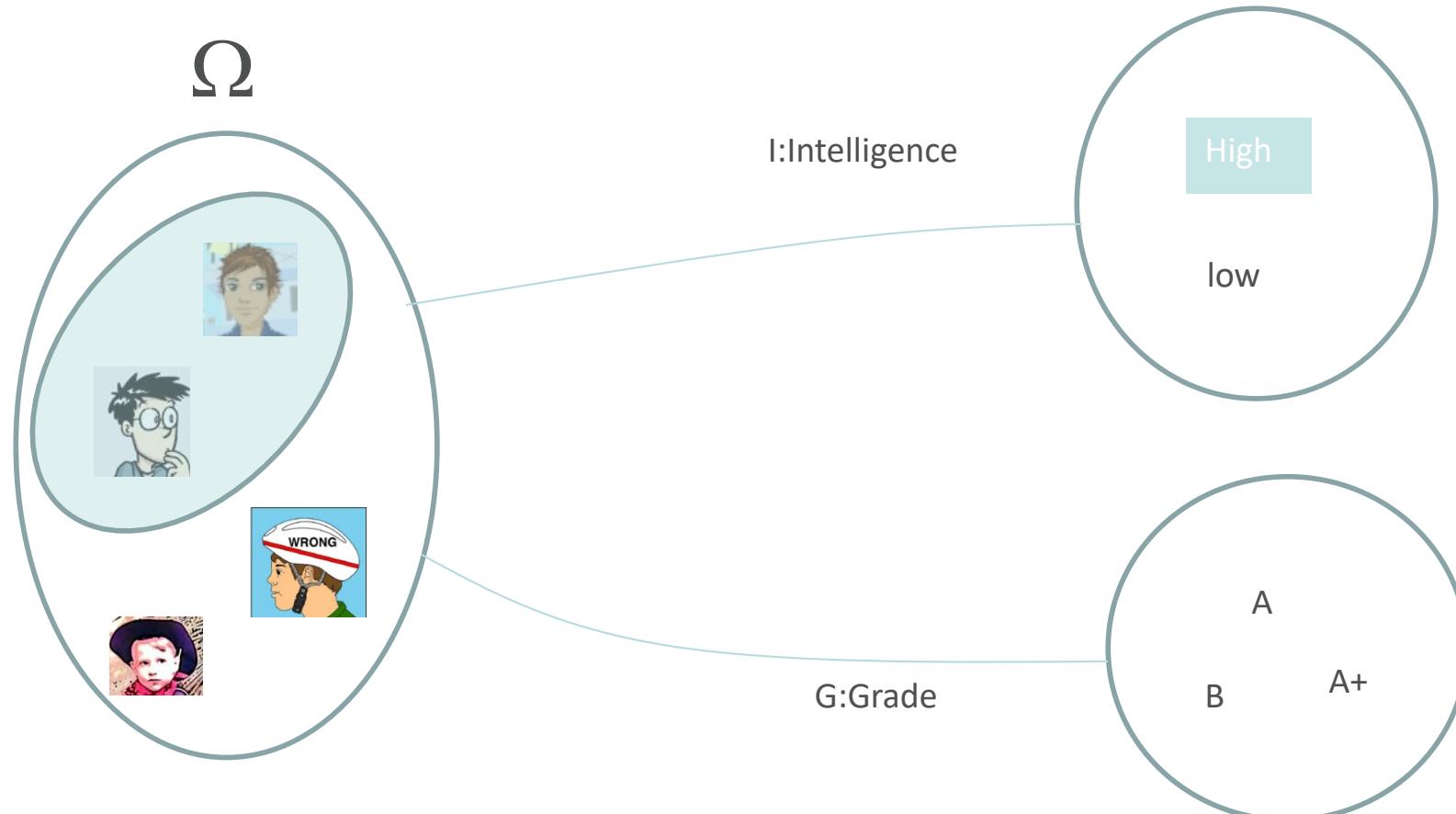
- Almost all the term we will be dealing with **RV**
- **Concise way of specifying attributes of outcomes**
- Modeling students (Grade and Intelligence):
 - Ω = all possible students
 - What are events
 - Grade_A = all students with grade A
 - Grade_B = all students with grade B
 - Intelligence_High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



Random Variables



Random Variables



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

- Always true

- $$\begin{aligned} P(x,y,z) &= p(x) \text{ } p(y|x) \text{ } p(z|x, y) \\ &= p(z) \text{ } p(y|z) \text{ } p(x|y, z) \\ &= \dots \end{aligned}$$



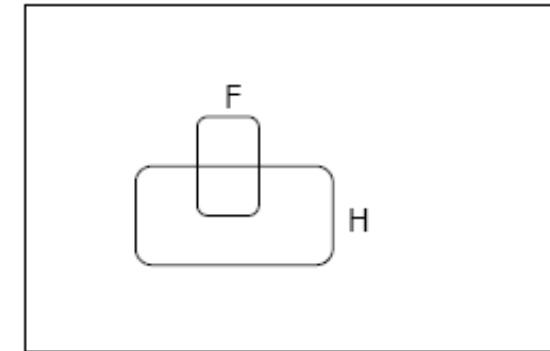
Conditional Probability

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

events

But we will always write it this way:

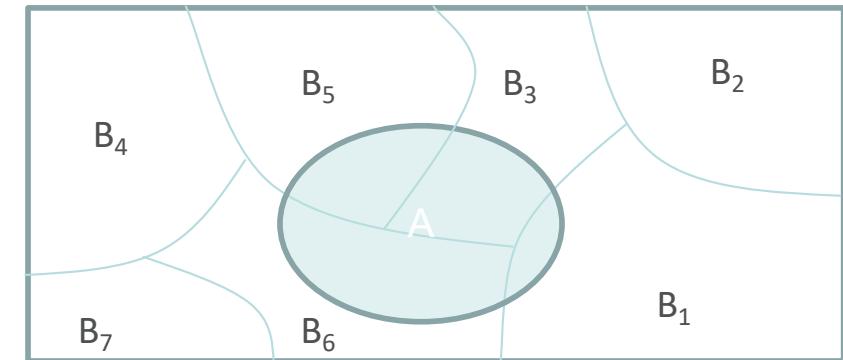
$$P(x | y) = \frac{P(x, y)}{P(y)}$$



Marginalization

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

$$\begin{aligned} P(x) &= \sum_y P(x,y) \\ &= \sum_y P(y) P(x|y) \end{aligned}$$



Marginalization Cont.

- Another example

$$\begin{aligned} P(x) &= \sum_{y,z} P(x, y, z) \\ &= \sum_{z,y} P(y, z) P(x | y, z) \end{aligned}$$



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



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

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

- Where does this come from?



Bayes Rule cont.

- You can condition on more variables

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



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.
- $X \perp Y | Z$ 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
 - $P(X=x | Z=z, Y=y) = P(X=x | Z=z)$
 - $P(Y=y | Z=z, X=x) = P(Y=y | Z=z)$
 - $P(X=x, Y=y | Z=z) = P(X=x | Z=z) P(Y=y | Z=z)$

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



Properties of CI

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



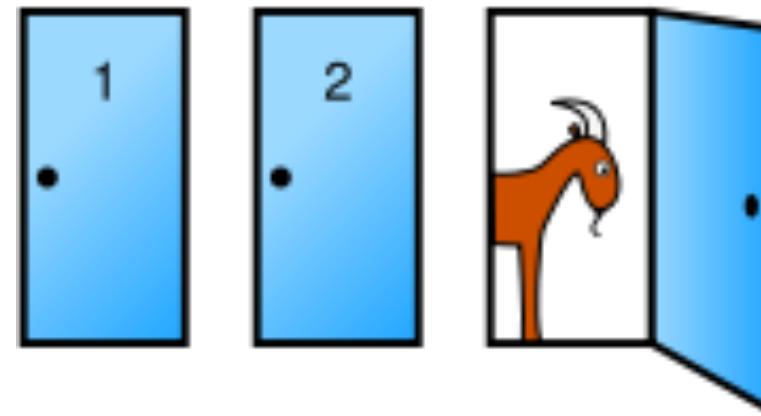
Probability Review

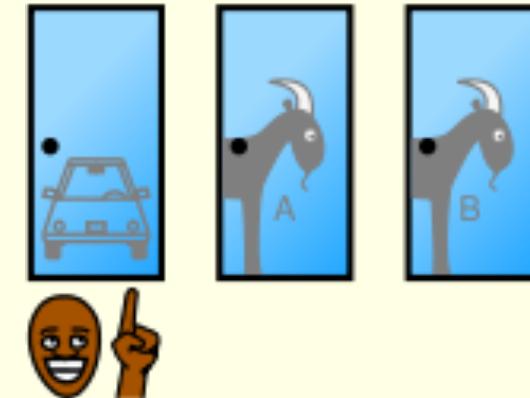
- Events and Event spaces
- Random variables
- Joint probability distributions
 - Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties
 - Independence, conditional independence
- Examples
- Moments



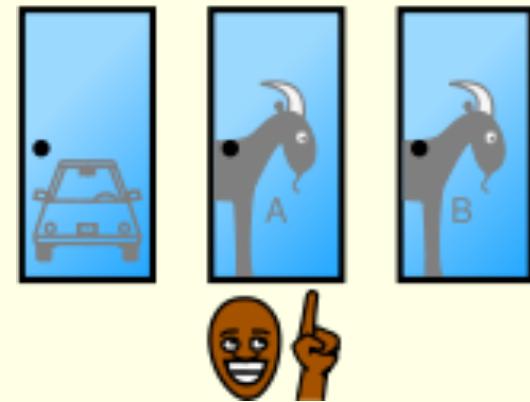
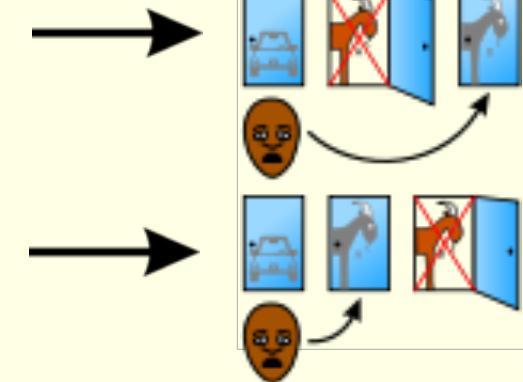
Monty Hall Problem

- 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?

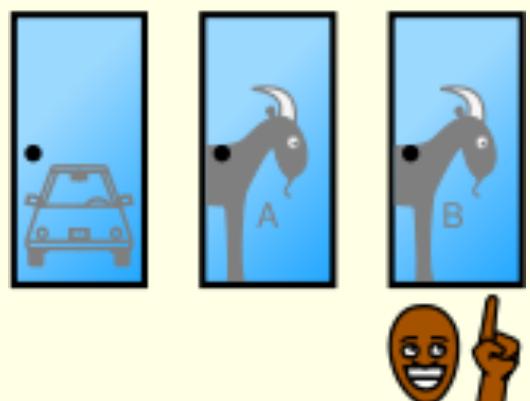
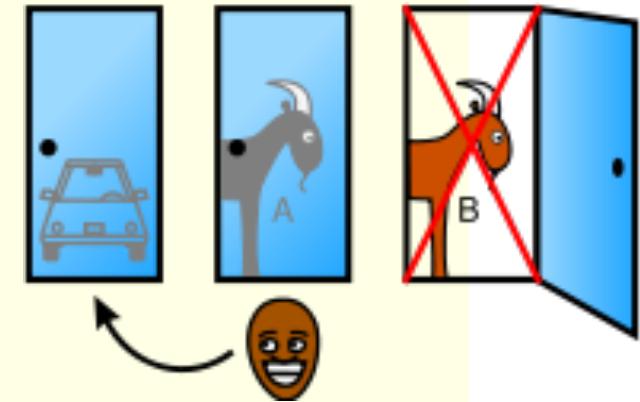




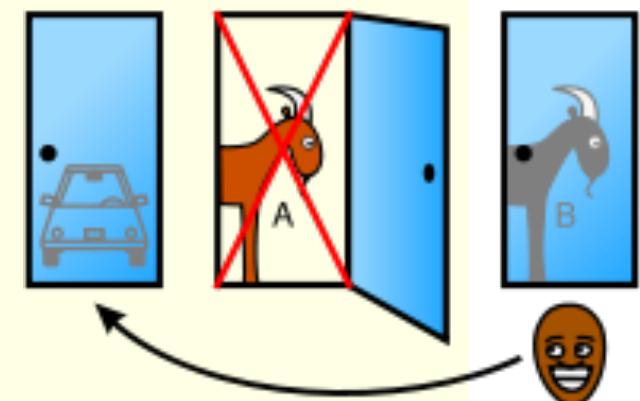
*Host reveals
Goat A
or
Host reveals
Goat B*



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

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





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

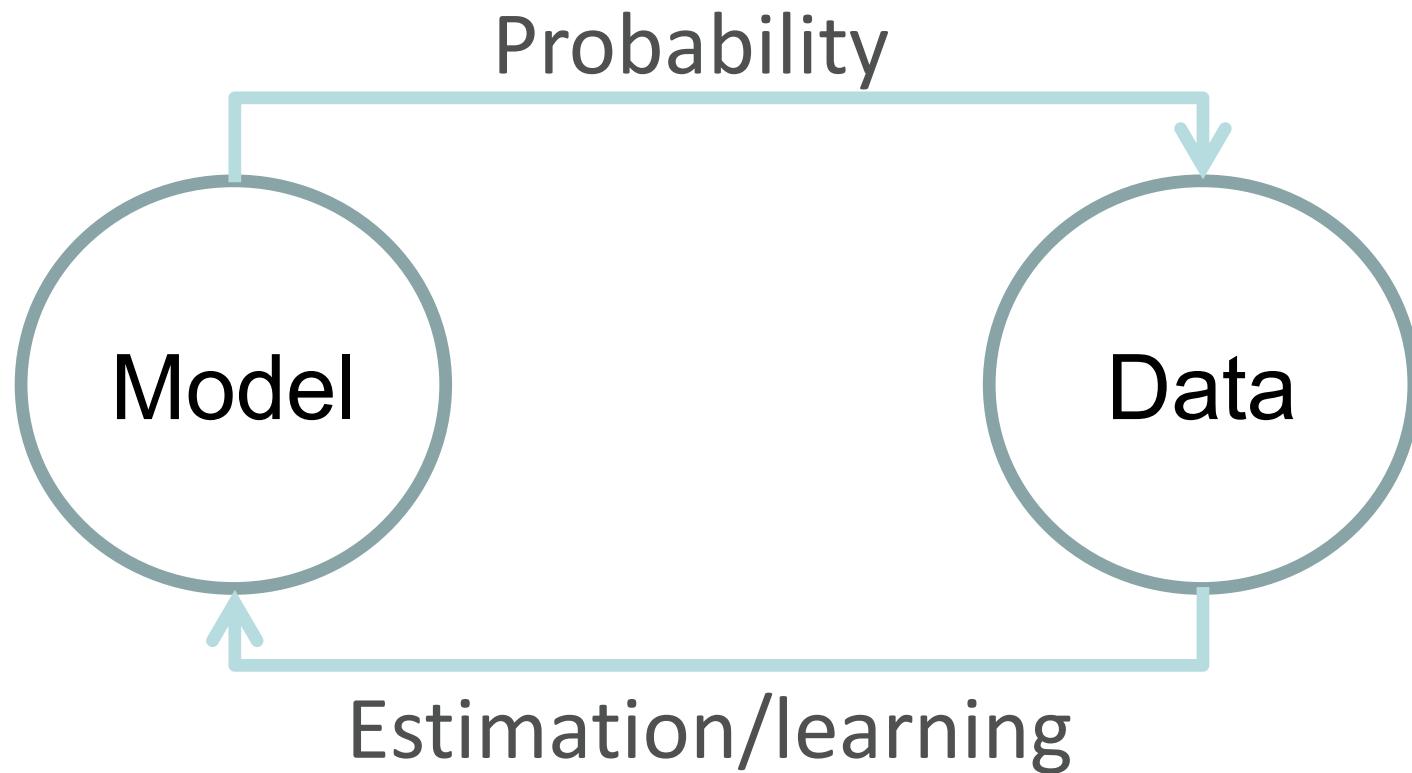
- $V(aX+b) = a^2V(X)$

- 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

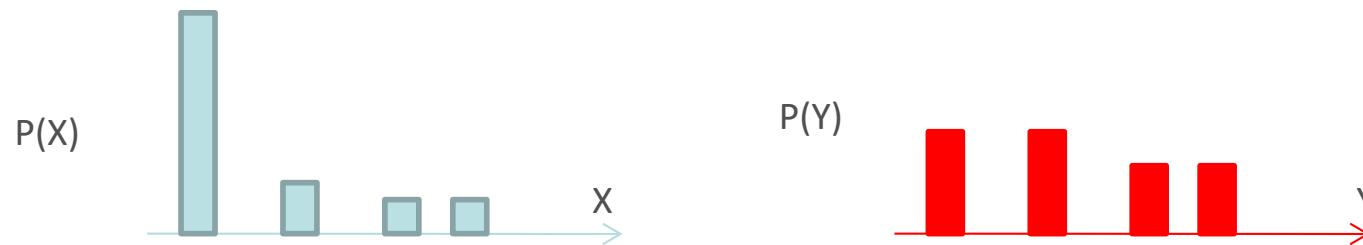


- Maximum Likelihood estimation
 - Example on board
 - Given N coin tosses, what is the coin bias (θ)?
- Sufficient Statistics:
 - Useful concept that we will make use later
 - In solving the above estimation problem, we only care about N_h , N_t , these are called the sufficient statistics of this model.
 - 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 than 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

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

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

$$H_p(X, Y, Z) = H_p(X) + H_p(Y|X) + H_p(Z|X, Y)$$



Mutual Information: MI

- Remember independence?
 - If $X \perp Y$ then knowing Y won't change our belief about X
 - Mutual information can help quantify this! (not the only way though)
- MI: $I_P(X;Y) = H_P(X) - H_P(X|Y)$
 - 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 $f(x)$ 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

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



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**

