Web-Mining Agents

Prof. Dr. Ralf Möller
Dr. Özgür Özçep

Universität zu Lübeck
Institut für Informationssysteme

Tanya Braun (Lab Class)

Organizational Issues: Lab Exercises

- **Start**: Wed, 18.10., 2-4pm, IFIS 2032, Class also Thu 2-4pm, IFIS 2032 (2rd floor)
- **Lab**: Fr. 2-4pm, Building 64, IFIS 2032 (2rd floor) (registration via Moodle right after this class)
- **Lab sheet** provided via Moodle after class on Thu.

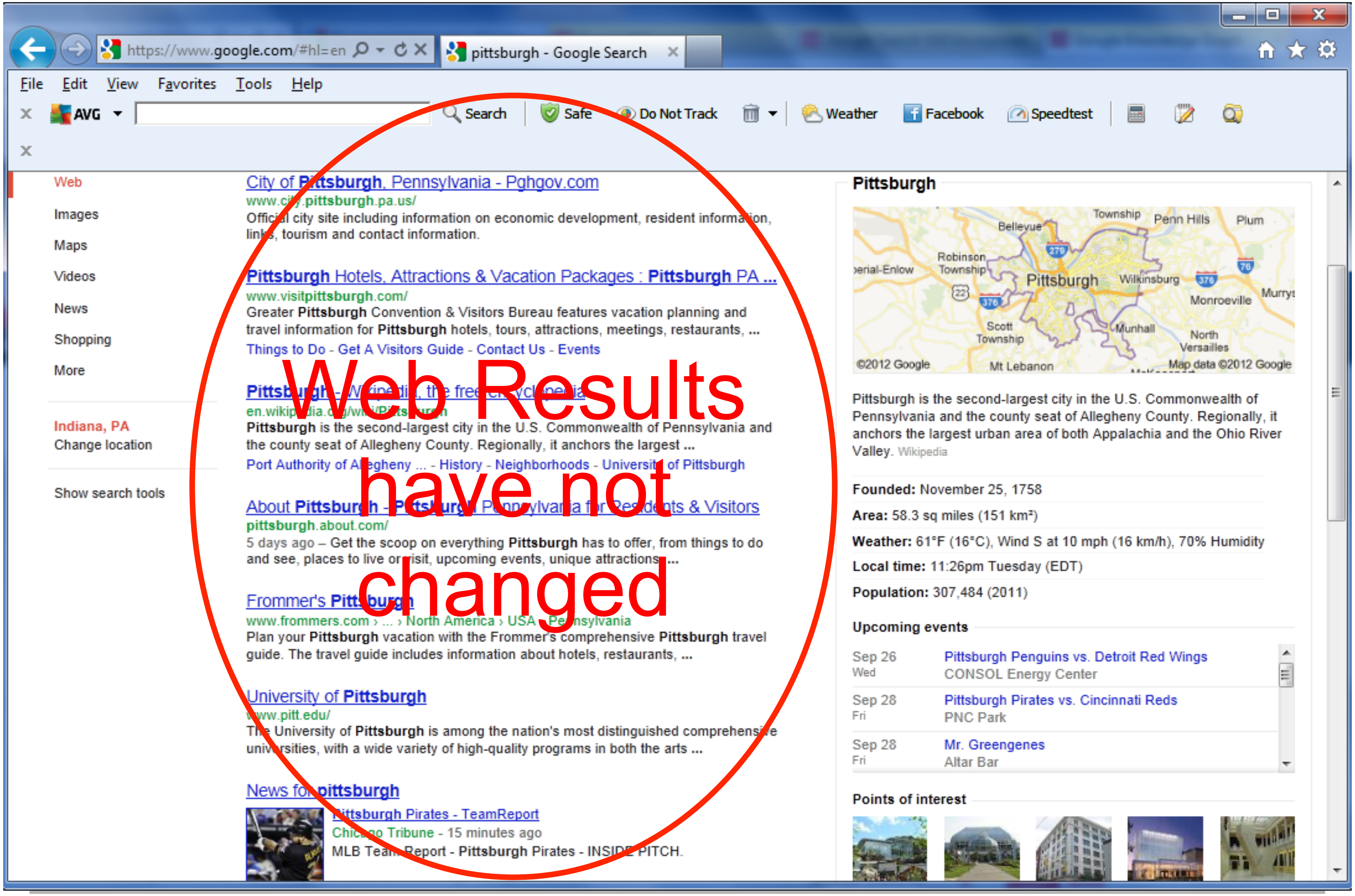
Organizational Issues: Exam

- Registration in class required to be able to participate in oral exam at the end of the semester (2 slots)

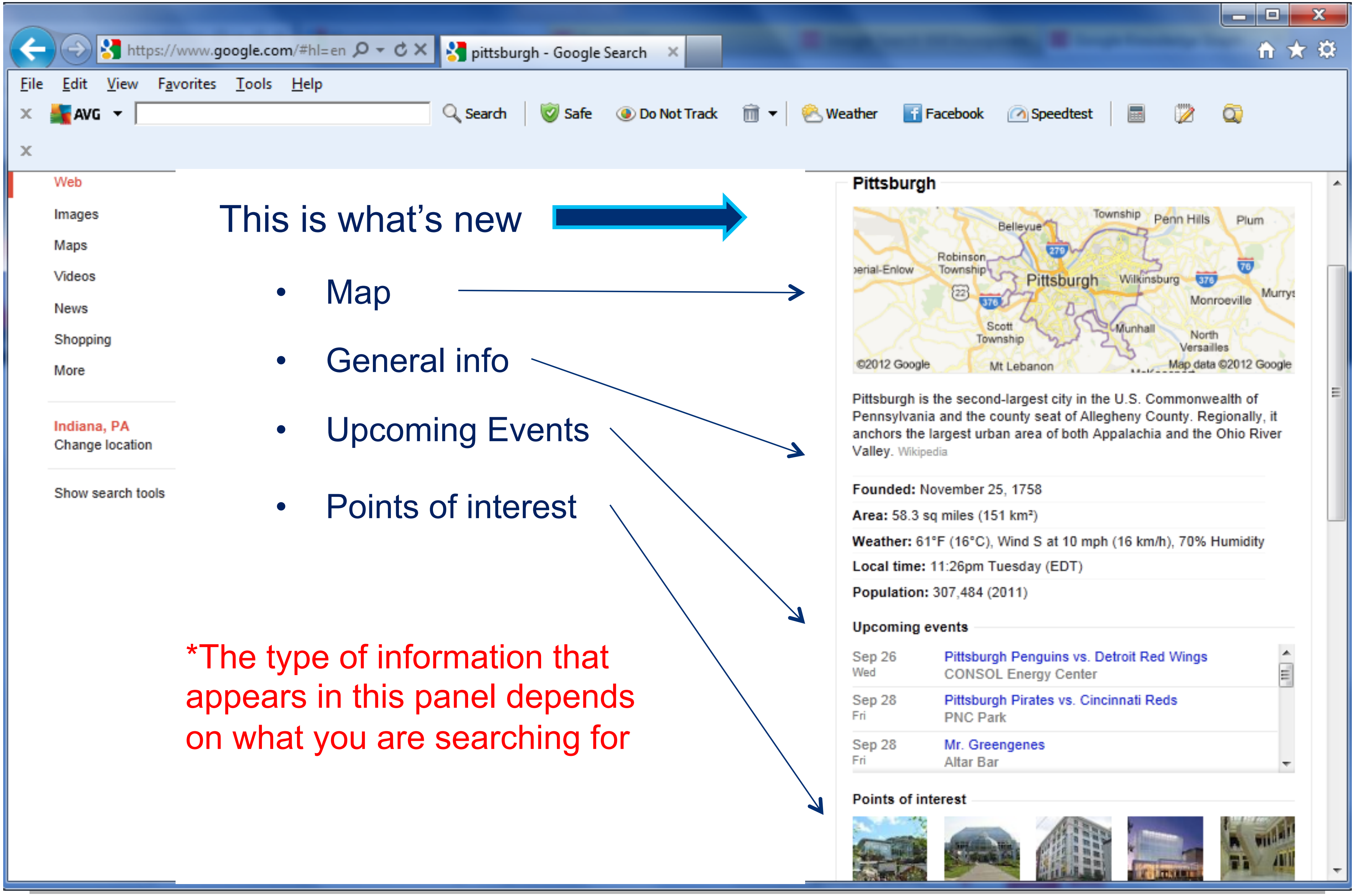
Search Engines: State of the Art

- Input: Strings (typed or via audio), images, ...
- Public services:
  - Links to web pages plus mini synopses via GUI
  - Presentations of structured information via GUI excerpts from the Knowledge Vault
    http://videolectures.net/kdd2014_murphy_knowledge_vault/
    (previously known as Knowledge Graph)
- NSA services: ?
- Methods: Information retrieval, machine learning
- Data: Grabbed from free resources (win-win suggested)

Search Results



Search Results



This is what's new

- Map
- General info
- Upcoming Events
- Points of interest

*The type of information that appears in this panel depends on what you are searching for

Search Engines: State of the Art

- Input: Strings (typed or via audio), images, ...
- Public services:
  - Links to web pages plus mini synopses via GUI
  - Presentations of structured information via GUI excerpts from the Knowledge Vault (previously known as Knowledge Graph)
- NSA services: ?
- Methods: Information retrieval, machine learning
- Data: Grabbed from many resources (win-win suggested):
  - Web, Wikipedia (DBpedia, Wikidata, …), DBLP, Freebase, ...

Search Engines

- Find documents: Papers, articles, presentations, ...
  - Extremely cool
  - But…
- Hardly any support for interpreting documents w.r.t. certain goals (Knowledge Vault is just a start)
- No support for interpreting data

- Claim: Standard search engines provide services but copy documents (and possibly data)
- Why can't individuals provide similar services on their document collections and data?

Personalized Information Engines

- Keep data, provide information
- Invite „agents“ to „view“ (i.e., interpret) local documents and data, without giving away all data
- Let agents take away „their“ interpretation of local documents and data (just like in a reference library).
- Doc/data provider benefits from other agents by (automatically) interacting with them
  - Agents should be provided with incentives to have them „share“ their interpretations
- No GUI-based interaction, but …
  … semantic interaction via agents

Courses@IFIS

- **Web and Data Science**
  - Module: Web-Mining Agents
    - Machine Learning / Data Mining (Wednesdays)
    - Agents / Information Retrieval (Thursdays)
    - Requirements:
      - Algorithms and Data Structures, Logics, Databases, Linear Algebra and Discrete Structures, Stochastics
  - Module: Foundations of Ontologies and Databases
    - (Wednesdays 16.00-18.30)
- Web-based Information Systems
- Data Management
  - Mobile and Distributed Databases
  - Semantic Web

Complementary Courses@UzL

- Algorithmics, Logics, and Complexity
- Signal Processing / Computer Vision
- Machine Learning
- Pattern Recognition
- Artificial Neural Networks (Deep Learning)

Introduction

Overview ML, Data Mining, Uncertainty, Probability

Literature

- Stuart Russell, Peter Norvig, Artificial Intelligence – A Modern Approach, Pearson, 2009 (or 2003 ed.)
- Ian H. Witten, Eibe Frank, Mark A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2011
- Ethem Alpaydin, Introduction to Machine Learning, MIT Press, 2009

- Numerous additional books, presentations, and videos

What We Mean by "Learning"

- Machine learning (ML) is programming computers / developing algorithms for
  - optimizing a performance criterion
  - using example data or "past experience"
  - by constructing general models that are good and useful approximations of the data

- Role of Statistics: Building mathematical models, core task is inference from a sample
- Role of CS: Efficient algorithms to
  - solve the optimization problem
  - and represent and evaluate the model for inference

Why and When “Learn” ?

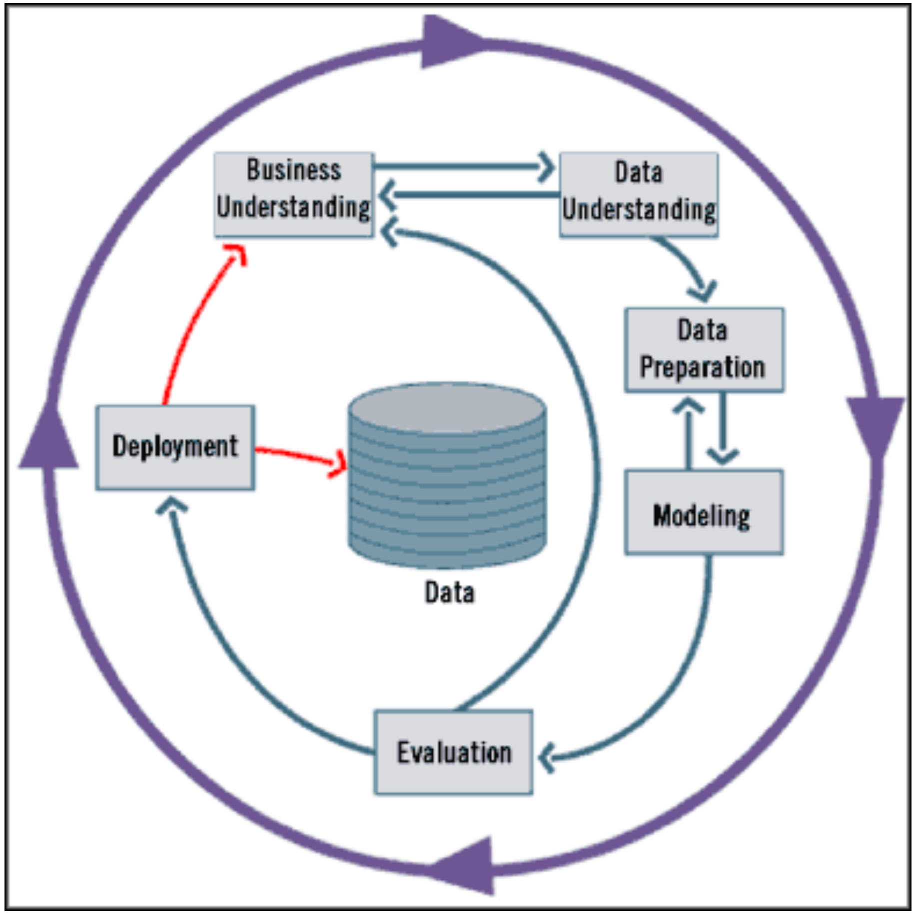
- There is no need to “learn” to calculate payrolls
- Learning is used in the following cases:
  - No human expertise
    (navigating on planet X)
  - Humans are unable to explain their expertise
    (speech recognition)
  - Solution changes in time
    (routing on a computer network)
  - Solution needs to be adapted to particular cases
    (user biometrics)

Data Mining

Application of machine learning methods to large databases is called ‘’Data mining”.

- Retail: Market basket analysis, customer relationship management (CRM, also relevant for wholesale)
- Finance: Credit scoring, fraud detection
- Manufacturing: Optimization, troubleshooting
- Medicine: Medical diagnosis
- Telecommunications: Quality of service optimization
- Bioinformatics: Sequence or structural motifs, alignment
- Web mining: Search engines
- ...

Standard Data Mining Life Cycle



But don‘t let the schema fool you

- No full automatism
- Handling DM tools require expert knowledge & intervention

Sample of ML Applications

- Learning Associations
- Supervised Learning
  - Classification
  - Regression
- Unsupervised Learning
- Reinforcement Learning

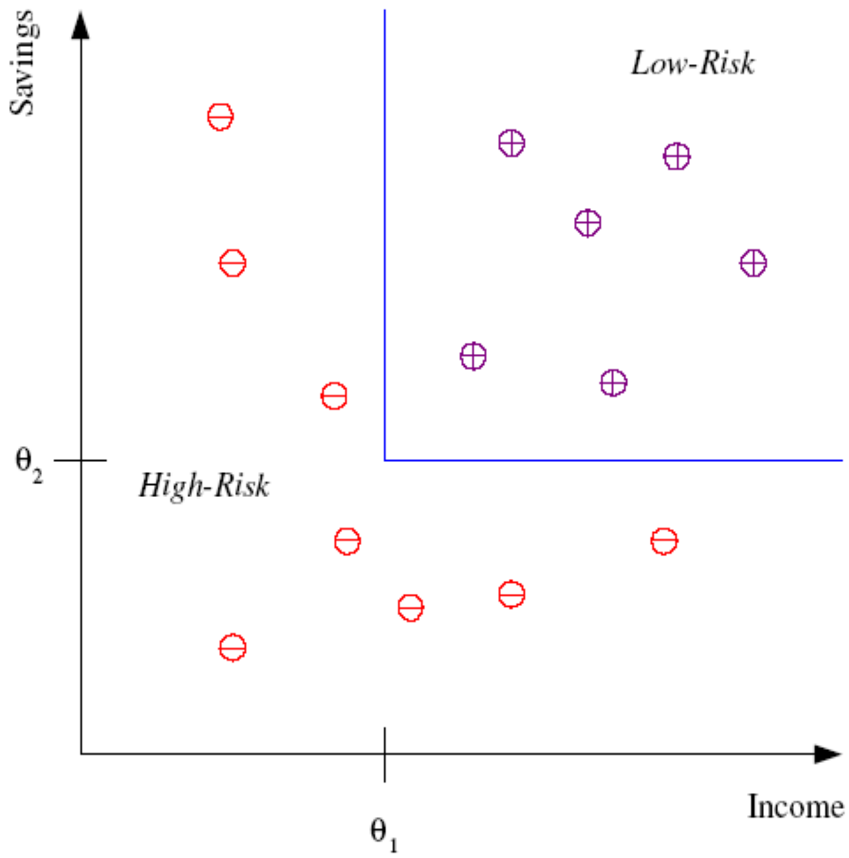
Learning Associations

- Basket analysis
  $P(Y \mid X)$ probability that somebody who buys X also buys Y where X and Y are products/services.

  Example: $P(\text{chips} \mid \text{beer}) = 0.7$
- If we know more about customers or make a distinction among them:
  - $P(Y \mid X, D)$
    where D is the customer profile (age, gender, marital status, …)
  - In case of a web portal, items correspond to links to be shown/prepared/downloaded in advance

Classification

- Example: Credit scoring
- Differentiating between low-risk and high-risk customers from their *income* and *savings*



Discriminant: IF *income* > θ_1 AND *savings* > θ_2

THEN low-risk ELSE high-risk

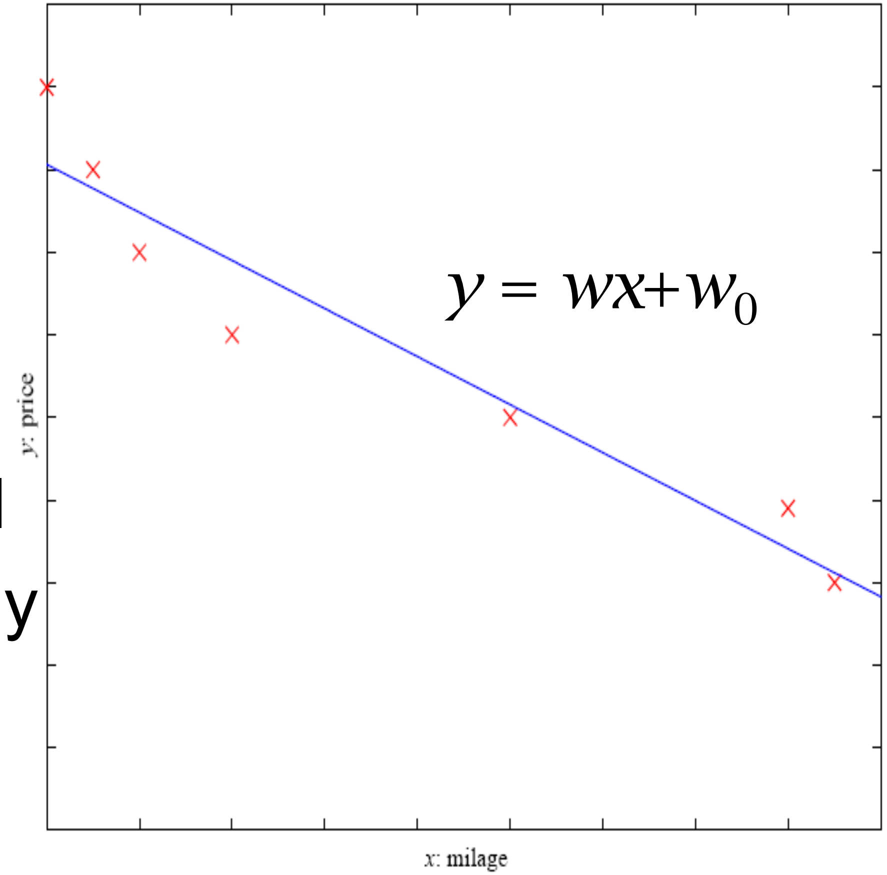
Classification: Applications

- Aka Pattern recognition
- Character recognition: Different handwriting styles.
- Face recognition: Pose, lighting, occlusion (glasses, beard), make-up, hair style
- Speech recognition: Temporal dependency
  - Use of a dictionary for the syntax of the language
  - Sensor fusion: Combine multiple modalities; e.g., visual (lip image) and acoustic for speech
- Medical diagnosis: From symptoms to illnesses
- Reading text:
- ...

Regression

In classification and regression the learned data structures
Are straight forward (concepts, function resp.
In next lecture we will consider other more complex structures
(bayesian networks) for learning

- Example: Price of a used car
- x : car attribute
  y : price
  $y = g(x \mid \theta)$: functional dependeny
  $g(\,)$: model,
  θ: parameters



Supervised Learning: Uses

- Prediction of future cases: Use the rule to predict the output for future inputs
- Knowledge extraction: The rule is easy to understand
- Compression: The rule is simpler than the data it explains
- Outlier detection: Exceptions that are not covered by the rule, e.g., fraud

Unsupervised Learning

- Learning “what normally happens”
- No output (we do not know the right answer)
- Clustering: Grouping similar instances
- Example applications
  - Customer segmentation in CRM (customer relationship manag.)
    - Company may have different marketing approaches for different groupings of customers
  - Document classification in unknown domains
  - Image compression: Color quantization
    - Instead of using 24 bits to represent 16 million colors, reduce to 6 bits and 64 colors, if the image only uses those 64 colors
  - Bioinformatics: Learning motifs (sequences of amino acids in proteins)

Reinforcement Learning

- Learning a policy: A sequence of actions/outputs
- No supervised output but delayed reward
- Credit assignment problem
- Game playing
- Robot in a maze
- Multiple agents, partial observability, ...

New Trends in ML

- In older ML approaches one assumes that the relevant features (of data and target) are given

(i.e., human-hand-made)

- Finding the right features is not trivial

- Learn features automatically

  (-> Deep Learning)

- Find (computationally) appropriate feature space
  - Transform (reduce) feature space

(-> SVMs, Kernels)

Overview

Supervised Learning

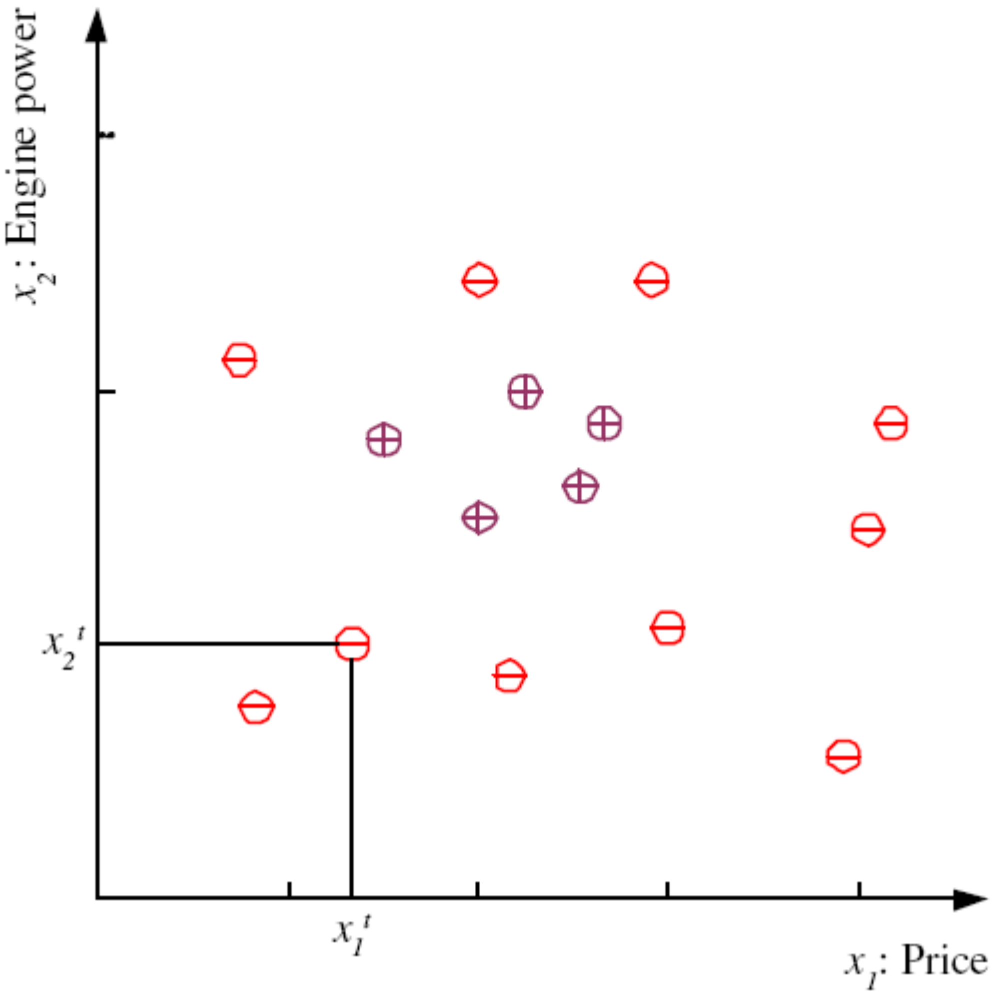
Learning a Class from Examples

- Class C of a "family car"
  - Prediction: Is car *x* a family car?
  - Knowledge extraction: What do people expect from a family car?
- Output:

  Positive (+) and negative (–) examples
- Input representation:

  x_1: price, x_2: engine power

Training set $\mathcal{X}$



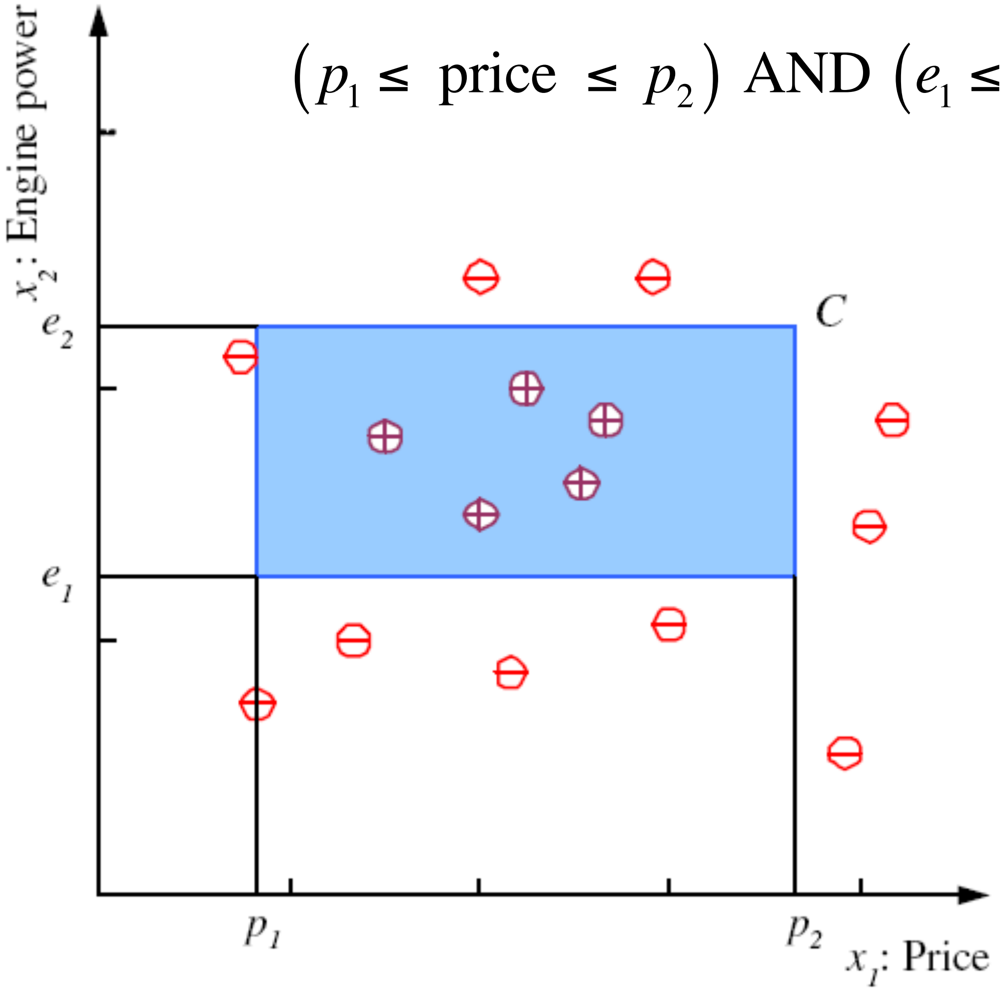
$$\mathcal{X} = \{\mathbf{x}^t, r^t\}_{t=1}^N$$

$$r = \begin{cases} 1 \text{ if } \boldsymbol{x} \text{ is positive} \\ 0 \text{ if } \boldsymbol{x} \text{ is negative} \end{cases}$$

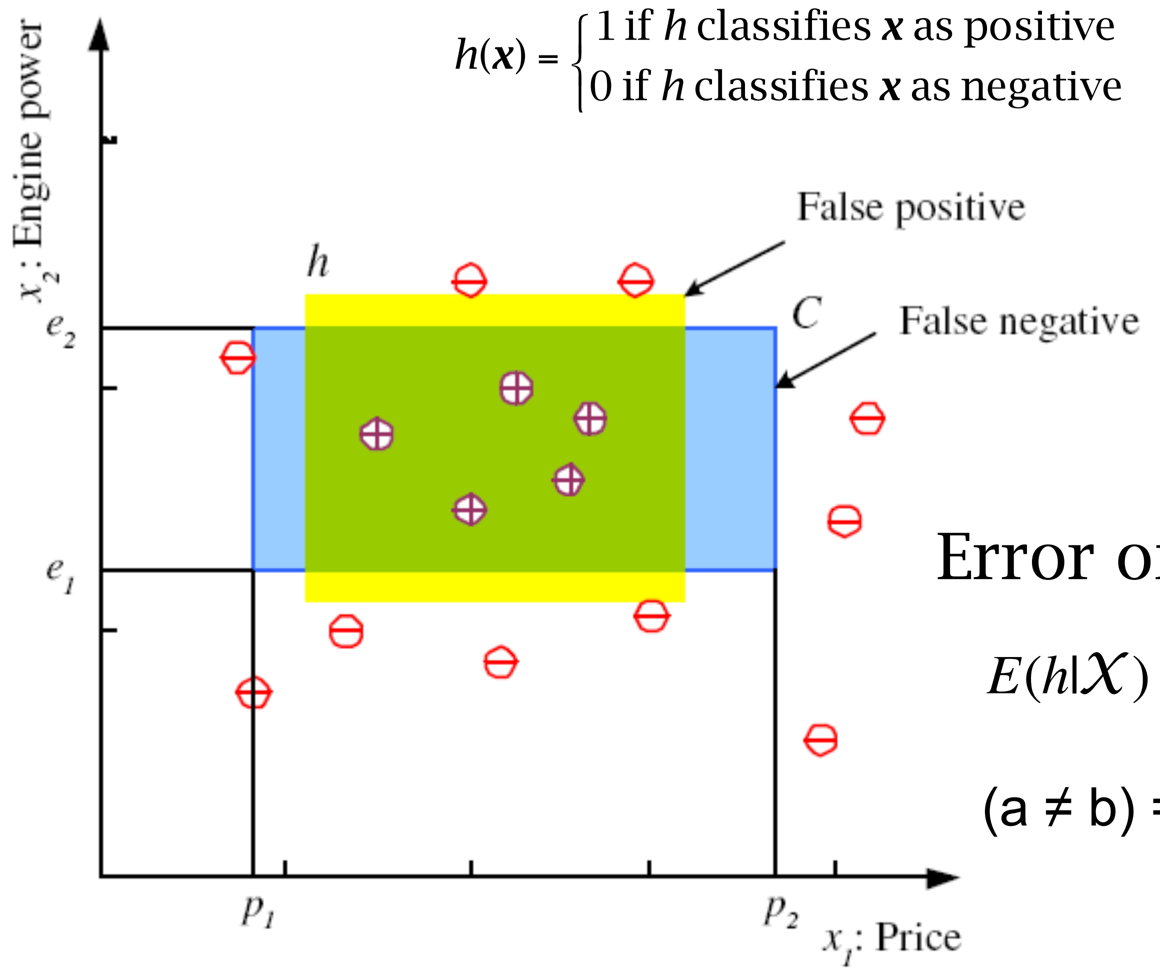
$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Class C

$$(p_1 \leq \text{ price } \leq p_2) \text{ AND } (e_1 \leq \text{ engine power } \leq e_2)$$



Hypothesis class $\mathcal{H}$



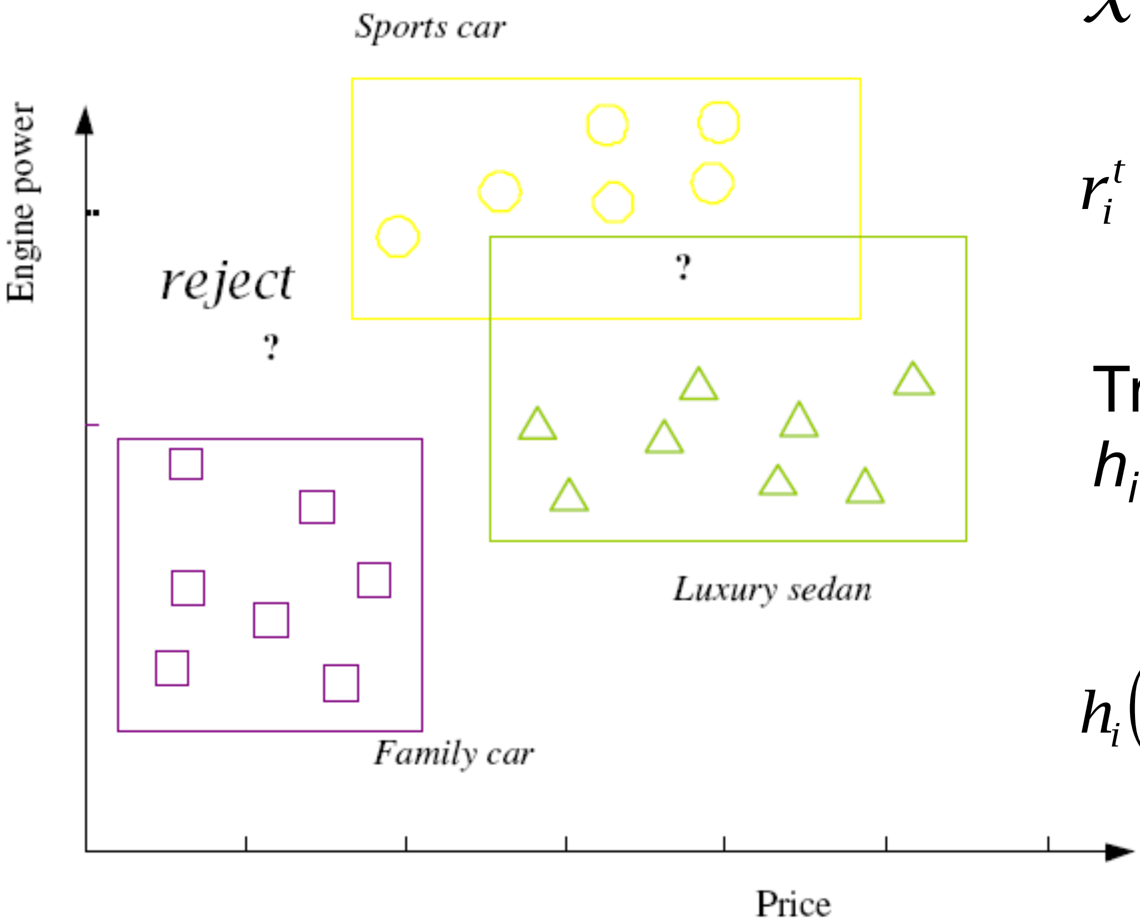
Sometimes one one consideres generalized approach via bounds in so called version spaces (all hypotheses fitting to the data)

Error of h on X

$$E(h|\mathcal{X}) = (1/N)\sum_{t=1}^{N}\left(h\left(\mathbf{x}^t\right) \neq r^t\right)$$

(a ≠ b) = 1 if ≠, 0 otherwise

Multiple Classes, C_i i=1,...,K



$$\mathcal{X} = \{\boldsymbol{x}^t, r^t\}_{t=1}^N$$

$$r_i^t = \begin{cases} 1 \text{ if } \boldsymbol{x}^t \in C_i \\ 0 \text{ if } \boldsymbol{x}^t \in C_j, j \neq i \end{cases}$$

Train hypotheses $h_i(\boldsymbol{x})$, i =1,...,K:

$$h_i(\boldsymbol{x}^t) = \begin{cases} 1 \text{ if } \boldsymbol{x}^t \in C_i \\ 0 \text{ if } \boldsymbol{x}^t \in C_j, j \neq i \end{cases}$$

Regression

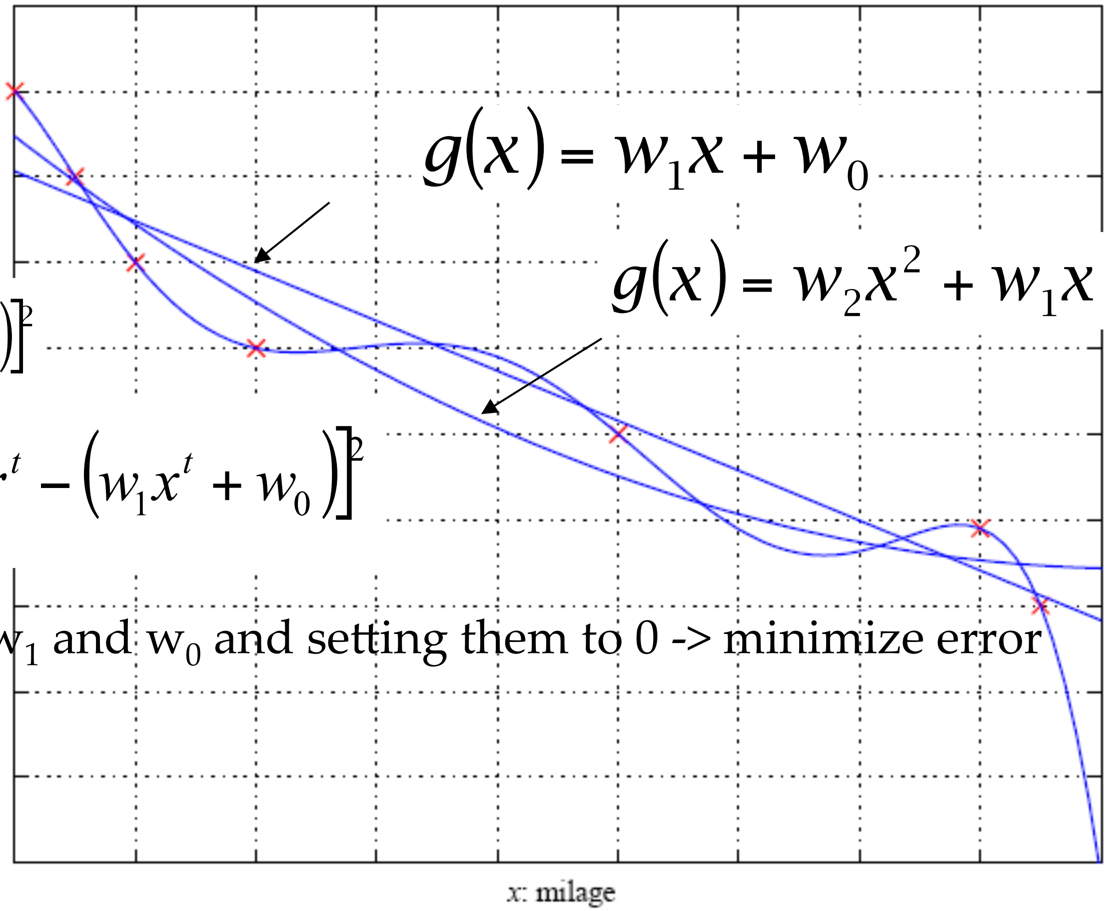
$$\mathcal{X} = \{x^t, r^t\}_{t=1}^N$$

$$r^t \in \Re$$

$$r^t = f(x^t)$$

$$E(g \mid \mathcal{X}) = \frac{1}{N}\sum_{t=1}^{N}\left[r^t - g(x^t)\right]^2$$

$$E(w_1, w_0 \mid \mathcal{X}) = \frac{1}{N}\sum_{t=1}^{N}\left[r^t - (w_1 x^t + w_0)\right]^2$$



Partial derivatives of E w.r.t w_1 and w_0 and setting them to 0 -> minimize error

$$w_1 = \frac{\sum_t x^t r^t - \overline{x}\overline{r}N}{\sum_t (x^t)^2 - N\overline{x}^2}$$

$$w_0 = \overline{r} - w_1 \overline{x}$$

Dimensions of a Supervised Learner

1. Model: $g(\boldsymbol{x} \mid \theta)$

2. Loss function: $E(\theta \mid \mathcal{X}) = \sum_t L\left(r^t, g(\boldsymbol{x}^t \mid \theta)\right)$

3. Optimization procedure: $\theta^* = \arg \min_\theta E(\theta \mid \mathcal{X})$

In most of ML: It‘s all about optimization

Model Selection & Generalization

- Learning is an ill-posed problem;
  data is not sufficient to find a unique solution
- The need for inductive bias, assumptions about hypothesis space H
- Generalization: How well a model performs on new data
- Overfitting: H more complex than concept C
  (function f, resp.)
- Underfitting: H less complex than C (resp. f)

Triple Trade-Off

There is a trade-off between three factors (Dietterich, 2003):

1. Complexity of H, for short *c*(H),
2. Training set size, *N*,
3. Generalization error, E, on new data

- *As N*↑, E↓
- As *c* (H) ↑, first *E*↑ and then *E*↓

How then do we chose the right hypothesis space?

Dietterich, T. G. 2003. "Machine Learning." In *Nature Encyclopedia of Cognitive Science*. London: Macmillan

Cross-Validation

- To estimate generalization error, we need data unseen during training. We split the data as
  - Training set (50%)

    [training, say, n models $g_1(\theta^*_1), \ldots g_n(\theta^*_n)$]
  - Validation set (25%)

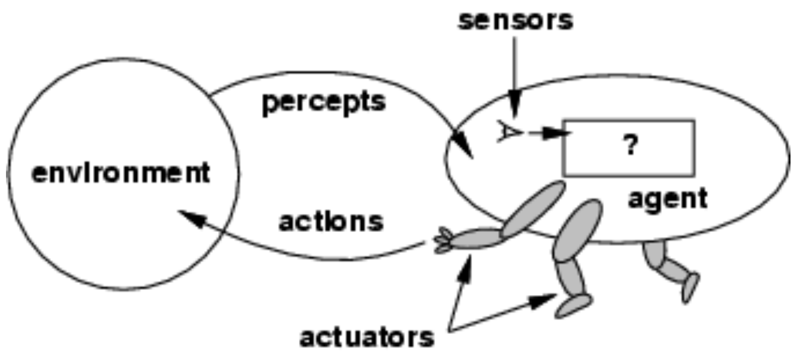
    [choosing best model:

    $$g_j(\theta^*_j) = \min \arg_{g_i(\theta^* i)} E(g_i(\theta^*_i) | VS)$$]
  - Test (publication) set (25%)

    [estimating generalization error of best model:

    $$E(g(\theta^*_j) | TS)$$]
- Resampling when there is few data

Agents and Environments



- The agent function maps from percept histories to actions: [*f*: P* → A]
- The agent program runs on the physical architecture to produce *f*
- agent = architecture + program
  architecture: PC, robotic car, …

Special case: Information Retrieval Agents

- Agent should be allowed to "extract information" from their host environment …
- ... and "make the information available" to their creator (owner)
- Very simple example: Linear regression
  – Percepts = Tuples from a database table
  – Extract information = compute a model (here: a line with parameters w_0, w_1) of the data
  – "Make information available" = send the model (w_0, w_1) to the creator, not the data
  – Data would be too large anyway in general settings

Handling Uncertainty with Probability

Uncertainty

Let action A_t = leave for airport t minutes before flight
Will A_t get me there on time?

Problems:

1. partial observability (road state, other drivers' plans, etc.)
2. noisy sensors (traffic reports)
3. uncertainty in action outcomes (flat tire, etc.)
4. immense complexity of modeling and predicting traffic

Hence, it seems that a purely logical approach either

1. risks falsehood: “A_{25} will get me there on time”, or
2. leads to conclusions that are too weak for decision making:

“A_{25} will get me there on time if there's no accident on the bridge and it doesn't rain and my tires remain intact etc etc.”

(A_{1440} might reasonably be said to get me there on time but I'd have to stay overnight in the airport …)

Methods for handling uncertainty

- Logic:
  - Assume my car does not have a flat tire
  - Assume A_{25} works unless contradicted by evidence
- Issues: What assumptions are reasonable? How to handle contradiction?
- Rules with fudge factors (belief in the rule):
  - $A_{25} \vdash_{0.3}$ get there on time
  - Sprinkler $\vdash_{0.99}$ WetGrass
  - WetGrass $\vdash_{0.7}$ Rain
- Issues: Problems with combination, e.g., Sprinkler causes Rain??

- Probability
  - Model agent's degree of belief
  - Given the available evidence,
    A_{25} will get me there on time with probability 0.04

Probability

Probabilistic assertions summarize effects of

- laziness: failure to enumerate exceptions, qualifications, etc.
- theoretical ignorance: no complete theory
- practical ignorance: lack of relevant facts, initial conditions, tests, etc.

Subjective probability:

- Probabilities relate propositions to agent's own state of knowledge

  e.g., $P(A_{25} \mid \text{no reported accidents}) = 0.06$

These are not assertions about the world

Probabilities of propositions change with new evidence:

e.g., $P(A_{25} \mid \text{no reported accidents, 5 a.m.}) = 0.15$

Probability theory: Representation formalism

- Basic element: **Random variable (RV)**
- Similar to propositional logic: possible worlds defined by assignment of values to random variables.
- Boolean random variables
  e.g., *Cavity* (do I have a cavity?). Domain is < true , false >
- Discrete random variables
  e.g., *Weather* is one of < *sunny, rainy, cloudy, snow* >
- Domain values must be exhaustive and mutually exclusive
- Elementary propositions are constructed by assignment of a value to a random variable: e.g.,
  - *Weather* = *sunny*,
  - *Cavity* = *false* (abbreviated as ¬*cavity*)
  - *Cavity* = *true* (abbreviated as *cavity*)
- Complex propositions formed from elementary propositions and standard logical connectives, e.g., *Weather* = *sunny* ∨ *Cavity* = *false*

Syntax

- Atomic event: A complete specification of the state of the world about which the agent is uncertain

E.g., if the world is described by only two Boolean variables: Cavity and Toothache, then there are 4 distinct atomic events:

Cavity = false ∧ Toothache = false
Cavity = false ∧ Toothache = true
Cavity = true ∧ Toothache = false
Cavity = true ∧ Toothache = true

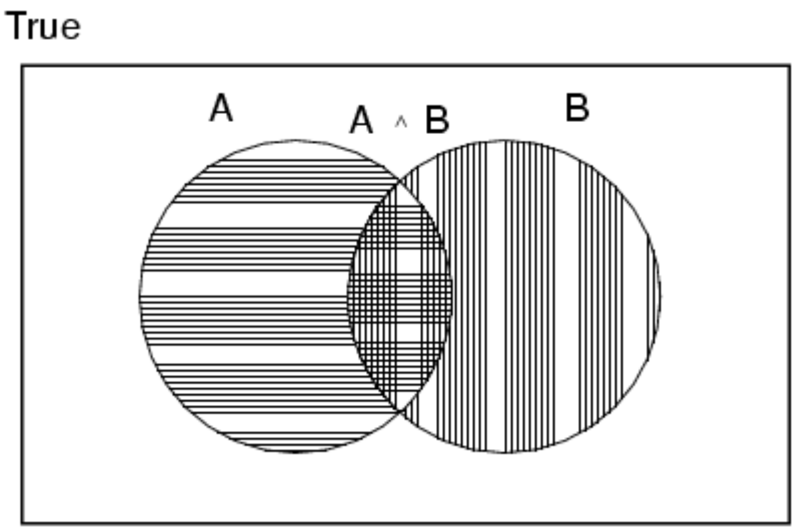
- Atomic events are mutually exclusive and exhaustive

Axioms of probability

For any propositions *A, B*

- $0 \leq P(A) \leq 1$
- $P(true) = 1$ and $P(false) = 0$

  (*true* stands for a tautology such as $(A \vee \neg A)$;
  false stands for a contradiction such as $A \,\&\, \neg A$))
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$



Example World

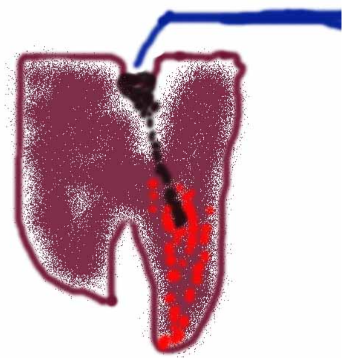
Example (Dentist problem with four var

Toothache (I have a toothache)

Cavity (I have a cavity)

Catch (steel probe catches in my tooth

Weather (sunny,rainy,cloudy,snow)



Prior probability

- Prior or unconditional probabilities of propositions

  e.g., $P(Cavity = \text{true}) = 0.1$ and $P(Weather = \text{sunny}) = 0.72$

  correspond to belief prior to arrival of any (new) evidence

- Probability distribution
  gives values for all possible assignments:

  $\mathbf{P}(Weather) = \langle 0.72, 0.1, 0.08, 0.1 \rangle$

  (normalized, i.e., sums to 1 because one must be the case)

Full joint probability distribution

- Joint probability distribution for a set of random variables gives the probability of every atomic event on those random variables
  P(Cavity,Whether) describes a 2 × 4 matrix of values:

Weather =	sunny	rainy	cloudy	snow
Cavity = true	0.144	0.02	0.016	0.02
Cavity = false	0.576	0.08	0.064	0.08

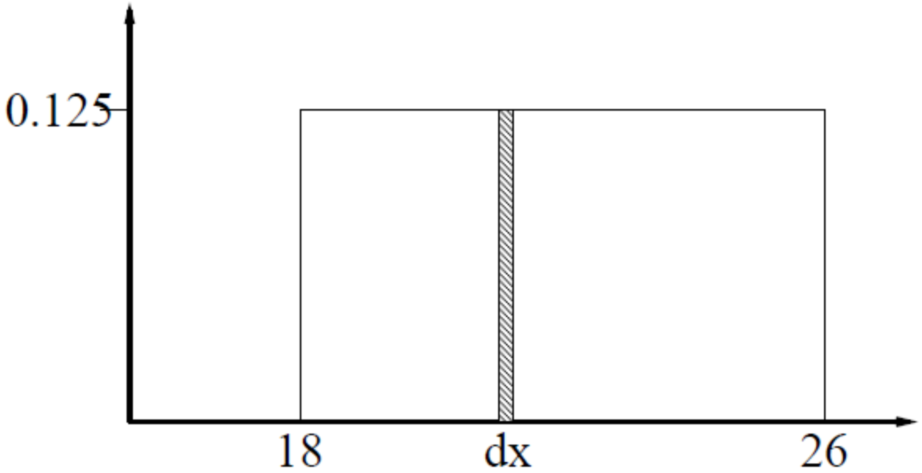
- Full joint probability distribution: all random variables involved
  - **P**(Toothache, Catch, Cavity, Weather)

- Every query about a domain can be answered by the full joint distribution

Probability for continuous variables

Express distribution as a parameterized function of value:

$P(X = x) = U[18, 26](x)$ = uniform density between 18 and 26



Here P is a density; integrates to 1.

$P(X = 20.5) = 0.125$ really means

$$\lim_{dx \to 0} P(20.5 \leq X \leq 20.5 + dx)/dx = 0.125$$

$P(a <= X <= b) = \int_a^b U(x)\, dx$

Discrete random variables: Notation

- Dom(W) = {sunny, rainy, cloudy, snow} and Dom(W) disjoint from domain of other random variables:
  – Atomic event "W=rainy" often written as "rainy"
  – Example: P(rainy), the random variable W is instantiated by the value rainy
- Boolean variable C
  – Atomic event "C=true" written as "c"
  – Atomic event "C=false" written as "¬c"
  – Examples: P(c) or P(¬c)

Conditional probability

- Conditional or posterior probabilities
  e.g., P(cavity | toothache) = 0.8
  or: <0.8>
  i.e., given that *toothache* is all I know

- (Notation for conditional distributions:
  P(Cavity | Toothache) = 2-element vector of 2-element vectors)

- If we know more, e.g., cavity is also given, then we have
  P(cavity | toothache,cavity) = 1

- New evidence may be irrelevant, allowing simplification, e.g.,
  P(cavity | toothache, sunny) = P(cavity | toothache) = 0.8

- This kind of inference, sanctioned by domain knowledge, is crucial

Conditional probability

- Definition of conditional probability (in terms of unconditional probability):
  $P(a \mid b) = P(a \wedge b) / P(b)$ if $P(b) > 0$
- Product rule gives an alternative formulation ($\wedge$ is commutative):
  $P(a \wedge b) = P(a \mid b)\, P(b) = P(b \mid a)\, P(a)$
- A general version holds for whole distributions, e.g.,
  $\mathbf{P}(Weather, Cavity) = \mathbf{P}(Weather \mid Cavity)\, \mathbf{P}(Cavity)$

  View as a set of 4 × 2 equations, not matrix mult.
  (1,1) $P(Weather{=}sunny \mid Cavity{=}true)\; P(Cavity{=}true)$
  (1,2) $P(Weather{=}sunny \mid Cavity{=}false)\; P(Cavity{=}false)$, ….
- Chain rule is derived by successive application of product rule:

$$
\begin{aligned}
\mathbf{P}(X_1, \ldots, X_n) &= \mathbf{P}(X_1,\ldots,X_{n-1})\, \mathbf{P}(X_n \mid X_1,\ldots,X_{n-1}) \\
&= \mathbf{P}(X_1,\ldots,X_{n-2})\, \mathbf{P}(X_{n-1} \mid X_1,\ldots,X_{n-2})\, \mathbf{P}(X_n \mid X_1,\ldots,X_{n-1}) \\
&= \ldots \\
&= \prod_{i=1}^{n} \mathbf{P}(X_i \mid X_1, \ldots, X_{i-1})
\end{aligned}
$$

Bayes‘ Rule (…is at the heart of everything)

Product rule $P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$

$\Rightarrow$ Bayes' rule $P(a|b) = \dfrac{P(b|a)P(a)}{P(b)}$

Useful for assessing diagnostic probability from causal probability:

$$P(Cause|Effect) = \frac{P(Effect|Cause)P(Cause)}{P(Effect)}$$

E.g., let M be meningitis, S be stiff neck:

$$P(m|s) = \frac{P(s|m)P(m)}{P(s)} = \frac{0.8 \times 0.0001}{0.1} = 0.0008$$

Note: posterior probability of meningitis still very small!

Inference by enumeration

- Start with the joint probability distribution:

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

- For any proposition φ, sum the atomic events where φ is true: $P(\varphi) = \Sigma_{\omega:\omega\models\varphi} P(\omega)$

Inference by enumeration

- Start with the joint probability distribution:

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

- For any proposition φ, sum the atomic events where it is true:
  $P(\varphi) = \Sigma_{\omega:\omega \models \varphi} P(\omega)$

- P(toothache) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2
- Unconditional or **marginal probability** of toothache

- Marginalization: $\mathbf{P}(\mathbf{Y}) = \Sigma_{\mathbf{z} \in \mathbf{Z}} \mathbf{P}(\mathbf{Y},\mathbf{z})$
- Conditioning on Z: $\mathbf{P}(\mathbf{Y}) = \Sigma_{\mathbf{z} \in \mathbf{Z}} \mathbf{P}(\mathbf{Y}|\mathbf{z})P(\mathbf{z})$

Inference by enumeration

- Start with the joint probability distribution:

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

For any proposition φ, sum over the atomic events ω where it is true:
$P(\varphi) = \Sigma_{\omega:\omega \models \varphi} P(\omega)$

- P(cavity ∨ *toothache*) = 0.108 + 0.012 + 0.072 + 0.008+ 0.016 + 0.064 = 0.28

(P(*cavity* ∨ *toothache*) = P(*cavity*) + P(*toothache*) – P(*cavity* ∧ *toothache*))

Inference by enumeration

- Start with the joint probability distribution:

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ cavity	**.016**	**.064**	**.144**	**.576**

- Can also compute conditional probabilities:

$$P(\neg cavity \mid toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)}$$

$$= \frac{0.016+0.064}{0.108 + 0.012 + 0.016 + 0.064}$$

$$= 0.4$$

P(cavity | toothache) = 0.108+0.012/0.2 = 0.6

Normalization

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

- Denominator **P**(**z**) (or P(toothache) in the example before) can be viewed as a normalization constant α

P(Cavity | toothache) = α **P**(Cavity,toothache)

= α [**P**(Cavity,toothache,catch) + **P**(Cavity,toothache,¬ catch)]

= α [<0.108,0.016> + <0.012,0.064>]

= α <0.12,0.08> = <0.6,0.4>

Inference by enumeration, contd.

- Typically interested in **P**(**Y** | **E** = e) = ?
  - Posterior joint distribution of **Y** under evidence E = e
  - Y = query variables
  - E = evidence variables
  - X := Y ∪ Y ∪ H = all RVs
  - H = hidden variables

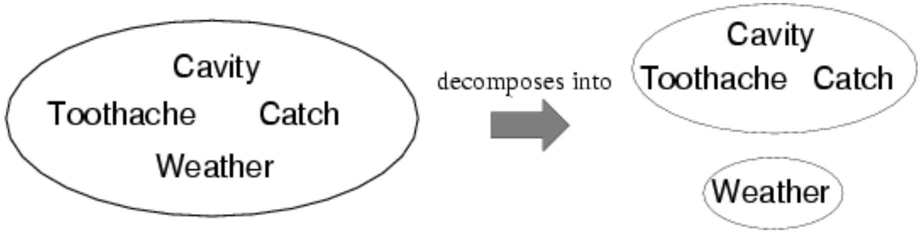
- Calculated by summing out the hidden variables:

$$\mathbf{P}(\mathbf{Y} \mid \mathbf{E} = \mathbf{e}) = \mathbf{P}(\mathbf{Y},\mathbf{E} = \mathbf{e})/P(E=e) = \alpha\mathbf{P}(\mathbf{Y},\mathbf{E} = \mathbf{e}) =$$
$$= \alpha\Sigma_h \mathbf{P}(\mathbf{Y},\mathbf{E}= \mathbf{e}, \mathbf{H} = \mathbf{h})$$

- Obvious problems
  1. Worst-case time complexity $O(d^n)$ (d = largest domain cardinality and n = #RVs)
  2. Space complexity $O(d^n)$ to store the joint distribution
  3. How to calculate joint distribution?

Independence

- A and B are independent iff $\mathbf{P}(A, B) = \mathbf{P}(A)\,\mathbf{P}(B)$
  (or alternatively: iff one of the following holds:
  1. $\mathbf{P}(A|B) = \mathbf{P}(A)$ and $P(A) \neq 0$ and $P(B) \neq 0$
  2. $P(A) = 0$ or $P(B) = 0$)



$\mathbf{P}$(Toothache, Catch, Cavity, Weather)
= $\mathbf{P}$(Toothache, Catch, Cavity) $\mathbf{P}$(Weather)

- 32 (= 2 x 2 x2 x 4) entries reduced to 12 (= 2x2x2 + 4);
- for n independent biased coins, $O(2^n) \rightarrow O(n)$
- Absolute independence powerful but rare
- Dentistry is a large field with hundreds of variables, none of which are independent. What to do?

Conditional independence

- **P**(Toothache, Cavity, Catch) has $2^3 - 1 = 7$ independent entries
- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
  (1) **P**(catch | toothache, cavity) = **P**(catch | cavity)
- The same independence holds if I haven't got a cavity:
  (2) **P**(catch | toothache,¬cavity) = **P**(catch | ¬cavity)
- Catch is conditionally independent of Toothache given Cavity:
  P(Catch | Toothache,Cavity) = **P**(Catch | Cavity)
- Equivalent statements:
  P(Toothache | Catch, Cavity) = **P**(Toothache | Cavity)
  P(Toothache, Catch | Cavity) = **P**(Toothache | Cavity) **P**(Catch | Cavity)

Conditional independence contd.

- Write out full joint distribution using chain rule:
  P(Toothache, Catch, Cavity)
  = **P**(Toothache | Catch, Cavity) **P**(Catch, Cavity)

  = **P**(Toothache | Catch, Cavity) **P**(Catch | Cavity) **P**(Cavity)

  = **P**(Toothache | Cavity) **P**(Catch | Cavity) **P**(Cavity)
  (with conditional independence)

  i.e., 2 + 2 + 1 = 5 independent numbers
- In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in n to linear in n.
- Conditional independence is our most basic and robust form of knowledge about uncertain environments.