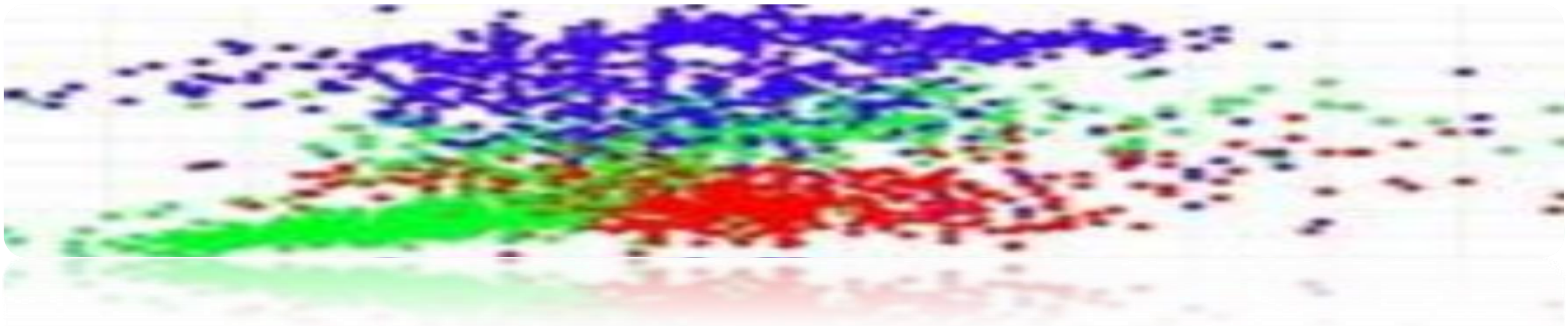


Dr.-Ing. Gjergji Kasneci

gjergji.kasneci@hpi.uni-potsdam.de

HPI Potsdam, winter term 2013/14

DATA MINING & PROBABILISTIC REASONING

Organization

- Timetable
  - Lectures
    - Tuesdays 13:30-15:00 in Room H-E.51
    - Every second Thursday 11:00-12:30 in Room H-2.57
  - Exercises
    - Every second Thursday 11:00-12:30 in Room H-2.57
- Teaching assistant
  - Maximilian Jenders (M.Sc.)
  - Expertise: Recommendations, Web Mining, Opinion Mining
- Exam
  - Condition for admission: Oral presentation of at least two solutions during the tutorials
  - Form of exam: oral exam at the end of the term

What is this lecture about?

- Data Mining
  - Analyzing data
  - Finding patterns/structure
  - Detecting outliers
  - Learning predictive models
  - Discovering knowledge

- Probabilistic Reasoning
  - Representing and quantifying uncertainty in data
  - Predicting likely outcomes of random variables, i.e., occurrence of events
  - Choosing the right model

Application areas

- Web mining (e.g., find documents for a given query or topic, group users by interest, recommendations, spam detection, …)
- Medicine/Bioinformatics (e.g., analyze the effect of drugs, derive diagnose based on symptoms, analyze protein-protein interactions, discover sequence similarities, detect mutations, …)
- Market analysis (e.g., market baskets, opinion mining, stock value prediction, influence propagation, …)
- Physics (e.g., multivariate data analysis, modeling motion of particles, i.e., Brownian motion, event classification, noise detection, …)
- Video games (e.g., AI game characters, matching players in online gaming, speech/shape recognition, …)
- …

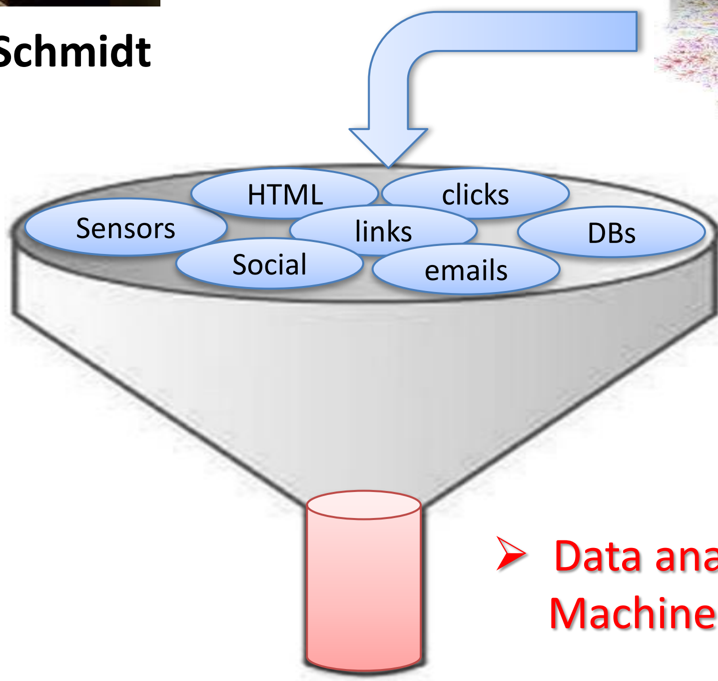
A Big Data perspective

"[...] every two days we create as much information as we did from the dawn of civilization up until 2003!"

Eric Schmidt

Large amounts of structured and unstructured data (often incomplete and ambiguous)

- Texts
- Lists, tables, graphs
- Images, audio, videos



- Distributed databases
- Key-value stores
- Column stores
- Document databases

- Data analytics , Data Mining,
  Machine Learning, and Knowledge Discovery

Example: Part-of-speech tagging (1)

- Task: Find the correct grammatical tag for terms in natural language text

- Difficulties arise from ambiguous grammatical meanings

- Examples

word		**tag**
flies	→	verb / noun
heat	→	verb / noun
like	→	verb / prep
water	→	noun / verb
in	→	prep / adv

Example: Part-of-speech tagging (2)

1. This/DT is/VBZ only/RB a/DT simple/JJ example/NN sentence/NN for/IN the/DT sake/NN of/IN presentation/NN
2. They/PRP are/VBP hunting/VBG dogs/NNS
3. Fruit/NNP flies/VBZ like/IN a/DT banana/NN

CC - Coordinating conjunction
CD - Cardinal number
DT - Determiner
EX - Existential there
FW - Foreign word
IN - Preposition or subordinating conjunction
JJ - Adjective
JJR - Comparative adjective
JJS - Superlative adjective
LS - List Item Marker
MD - Modal verb
NN - Singular noun
NNS - Plural noun
NNP - Proper singular noun
NNPS - Proper plural noun

PDT - Predeterminer
POS - Possesive ending
PRP - Personal pronoun
PRPS - Possesive pronoun
RB - Adverb
RBR - Comparative adverb
RBS - Superlative Adverb
RP - Particle
SYM - Symbol
TO - to
UH - Interjection
VB - Verb, base form
VBD - Verb, past tense
VBG - Verb, gerund/present participle
VBN - Verb, past participle

VBP - Verb, non 3rd ps. sing. present
VBZ - Verb, 3rd ps. sing. present
WDT - wh-determiner
WP - wh-pronoun
WPS - Possesive wh-pronoun
WRB - wh-adverb
S - Dollar sign
. - Sentence-break punctuation . ? !
- Pound sign
- - Dash sign
, - Comma
: - Colon, semi-colon
(- Open parenthesis)] }
) - Close parenthesis)] }
`` - Open quote
'' - Close quote

From: http://smile-pos.appspot.com/

Other important text analysis tasks

- Role labeling
- Entity recognition
- Entity disambiguation
- Relationship extraction
- Topic assignment (classification)
- Clustering

Example: Email classification

- Example classes
  - Spam vs. non-spam
  - Important vs. less important
  - Work-related / social / family / ads /…

- Simple model

  Assign email $\mathbf{m}$ to class

$$C^* = \mathrm{argmax}_C P(C|\mathbf{m}) = \mathrm{argmax}_C P\big(C|\overbrace{x_1(\mathbf{m}), x_2(\mathbf{m}), \dots, x_k(\mathbf{m})}^{\text{Features of } \mathbf{m}}\big)$$

e.g., email domain

e.g., indicates whether certain word appears

Example: Click prediction

flowers

Ungefähr 153.000.000 Ergebnisse (0,19 Sekunden)

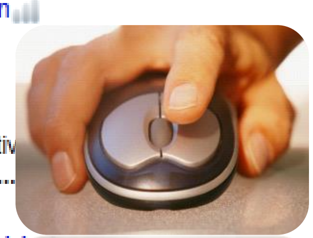
Rank ads by: $P(C = 1 | Q = q, A = a)$

Anzeige - Warum diese Anzeige?

Flowers to USA for $19.99 | ProFlowers.com
www.pro**flowers**.com
Send **Flowers** to Your Loved Ones. Free Vase & Satisfaction Guarantee.
Birthday Flowers - Valentine's Flowers - Free Delivery - 20% Off

Flowers.de
www.**flowers**.de/
Blumenversand-Vieles lässt sich mit Blumen leiter sagen! Deshalb bietet unser Blumenversand von **Flowers**.de das passende Blütenarrangement für jeden ...
Blumensträuße - Kundenlogin - Impressum - Geschenkideen

Flower - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Flower** - Diese Seite übersetzen
A **flower**, sometimes known as a bloom or blossom, is the reproductiv
in flowering plants (plants of the division Magnoliophyta, also called ...



Valentine's Day **Flowers** & Gifts | 1-800-**FLOWERS**.COM
www.1800**flowers**.com/ - Vereinigte Staaten - Diese Seite übersetzen
Find the perfect Valentine's Day **flowers** and gifts for your sweetheart at 1-800-**FLOWERS**.COM. Order roses, **flowers**, and other gifts for delivery on Valentine's ...
Birthday Flowers and Gifts - Sympathy - Roses - Sale

FTD.COM - **Flowers** Online | Roses, Fresh **Flowers**, Plants and Gift ...
www.ftd.com/ - Diese Seite übersetzen
22 Dec 2011 – Order **flowers** online for same day floral delivery. Shop for **flowers**, chocolates, roses, gifts and gift baskets by occasion, season or get beautiful ...

Anzeigen - Warum diese Anzeigen?

Blumen - Heute auf morgen
www.blumengruss.de
Inserent ist mit ★★★★★ bewertet
Bis 14 Uhr bestellt und am nächsten Tag bundesweit geliefert. Frisch!

Fleurop - echte Blumen
www.fleurop.de/blumenversand
fleurop.de ist mit ★★★★★ bewertet
von ECHTEN Floristen! Auf die Qualität kommt es an.

UK **Flower** Delivery
www.arena**flowers**.com/UK
Inserent ist mit ★★★★★ bewertet
Free Delivery & Fantastic Prices!
Send Beautiful **flowers** to the UK.

Send **Flowers** Online
www.euroflorist.de/_Send_**flowers**
Hand delivered fresh **flowers**.
Order by 3pm for same day service!

Fleurop Switzerland
www.fleurop.ch
Flowers within hours all over the world - Satisfaction guaranteed.

Example: Image categorization

IMAGENET Large Scale Visual Recognition Challenge 2013 (ILSVRC2013)

Introduction

This challenge evaluates algorithms for object detection and image classification at large scale. This year there will be three competitions:

1. A PASCAL-style detection challenge on fully labeled data for 200 categories of objects, NEW
2. An image classification challenge with 1000 categories, and
3. An image classification plus object localization challenge with 1000 categories.

Animal, animate being, beast, brute, creature, fauna

A living organism characterized by voluntary movement

1571 pictures | 87.44% Popularity Percentile | Wordnet IDs

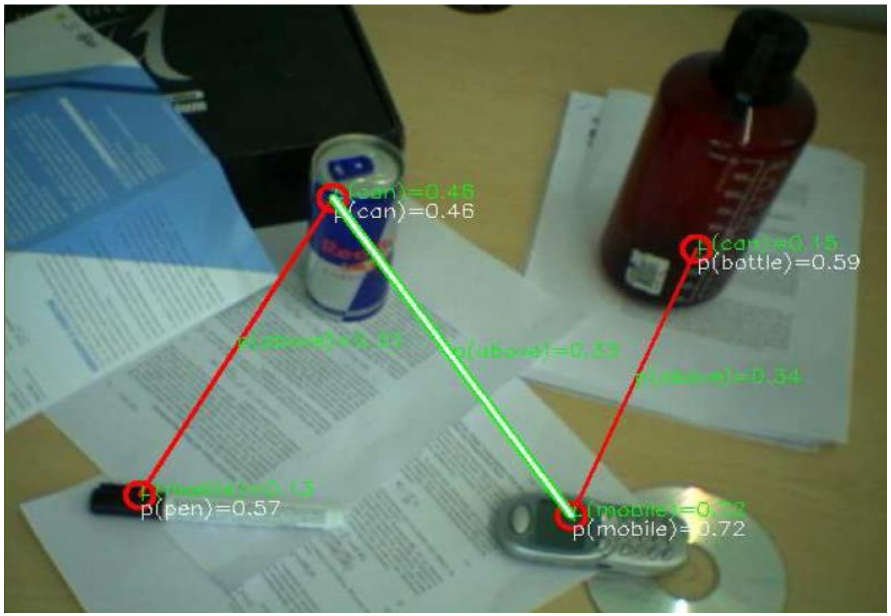
Numbers in brackets: (the number of synsets in the subtree).

- ImageNet 2011 Fall Release (21841)
  - animal, animate being, beast, brute,
  - plant, flora, plant life (3775)
  - person, individual, someone, someb
  - fungus (298)
  - natural object (551)
  - artifact, artefact (7894)
  - sport, athletics (165)
  - geological formation, formation (150)
  - Misc (13098)

Treemap Visualization | Images of the Synset | Downloads

Source: http://image-net.org/

Example: Object recognition and vision support



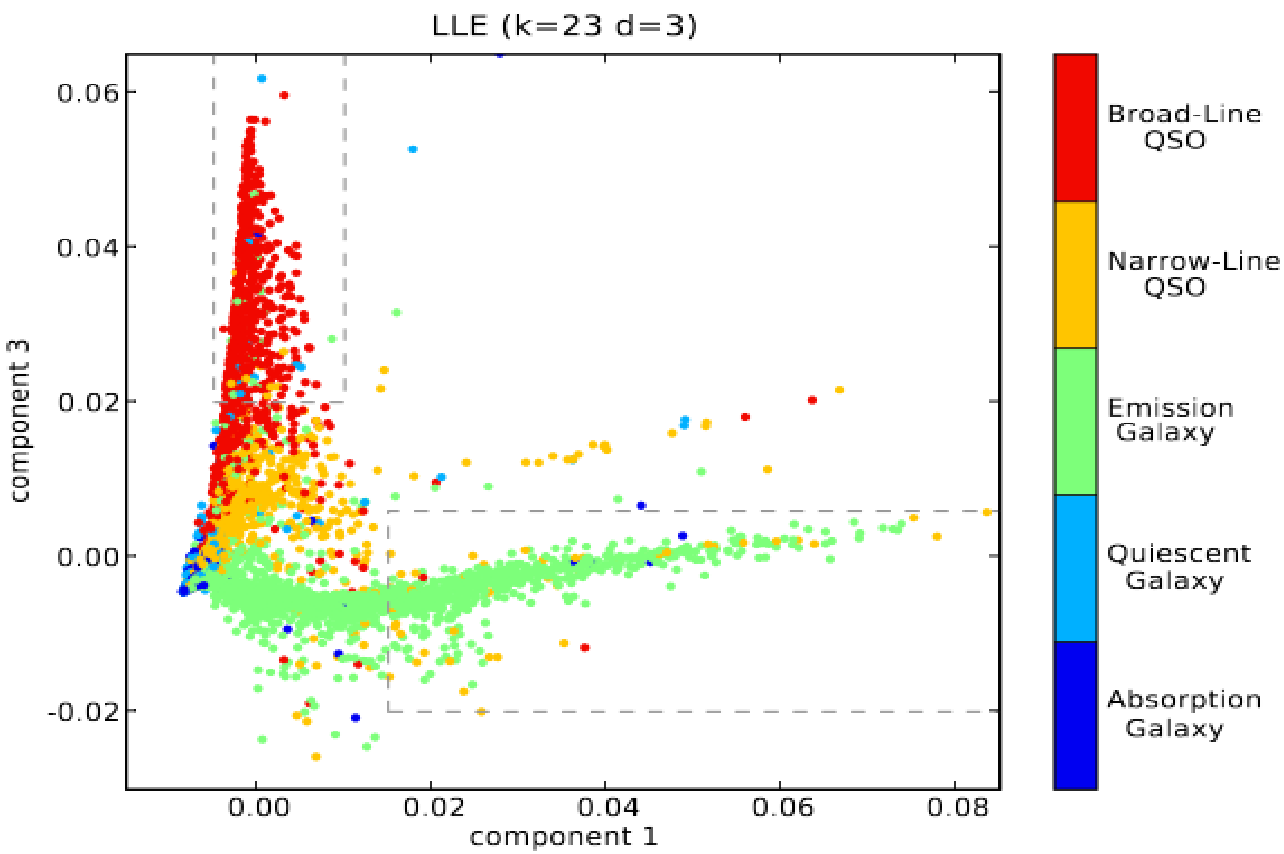
From: http://www.cognitivesystems.org

From: Tafaj et al.: ICANN'12

Example: Shape and speech recognition

Source: http://www.computerweekly.com

Example: Clustering astrophysical objects



From: http://ssg.astro.washington.edu/research.shtml?research/galaxies

Example: Recommendation

Amazon recommendations

Collaborative filtering

… see also the Netflix Challenge

Example: Movie recommendation

M1: The Shawshank Redemption
M2: The Usual Suspects
M3: The Godfather
M4: The Big Lebowski

$$
\begin{array}{c} \\ \text{User 1} \\ \text{User 2} \\ \text{User 3} \\ \text{User 4} \\ \text{User 5} \\ \text{User 6} \end{array}
\begin{array}{c} \begin{array}{cccc} \text{M1} & \text{M2} & \text{M3} & \text{M4} \end{array} \\ \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 2 & 2 & 2 \\ 0 & 0 & 0 & 1 \\ 1 & 2 & 3 & 2 \\ 1 & 0 & 1 & 1 \\ 0 & 2 & 2 & 3 \end{pmatrix} \end{array}
=
\begin{array}{c} \begin{array}{ccc} \text{T1} & \text{T2} & \text{T3} \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \end{array}
*
\begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}
*
\begin{array}{c} \begin{array}{cccc} \text{M1} & \text{M2} & \text{M3} & \text{M4} \end{array} \\ \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}
$$

e.g., drama e.g., crime e.g., comedy

Matrix factorization

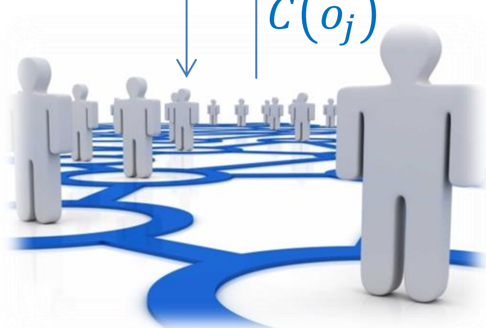
Example from: [Machine Learning by P. Flach]

Example: Learning from crowds

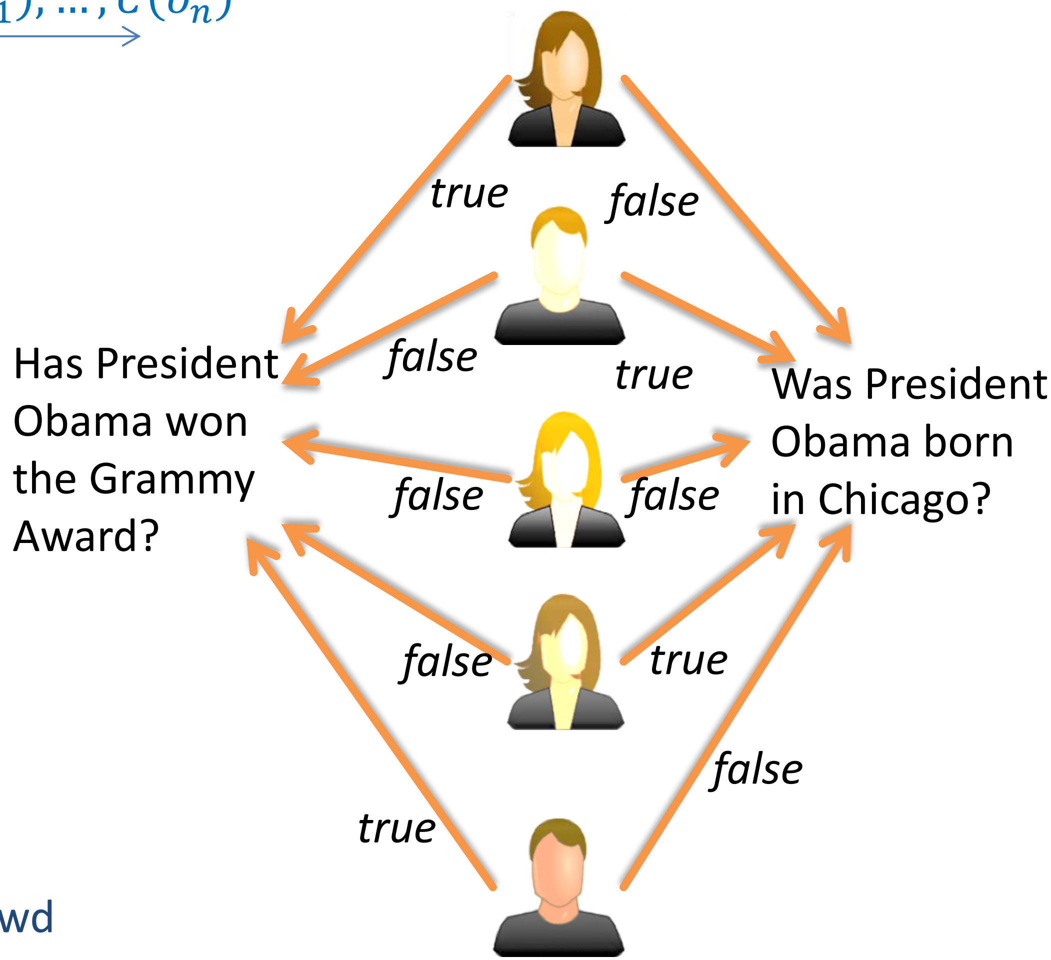
$o_1, \dots, o_n$ → Classification system → $C(o_1), \dots, C(o_n)$

Active learning scenario

$o_j \vdots o_i$ ↓ ↑ $C(o_i) \vdots C(o_j)$



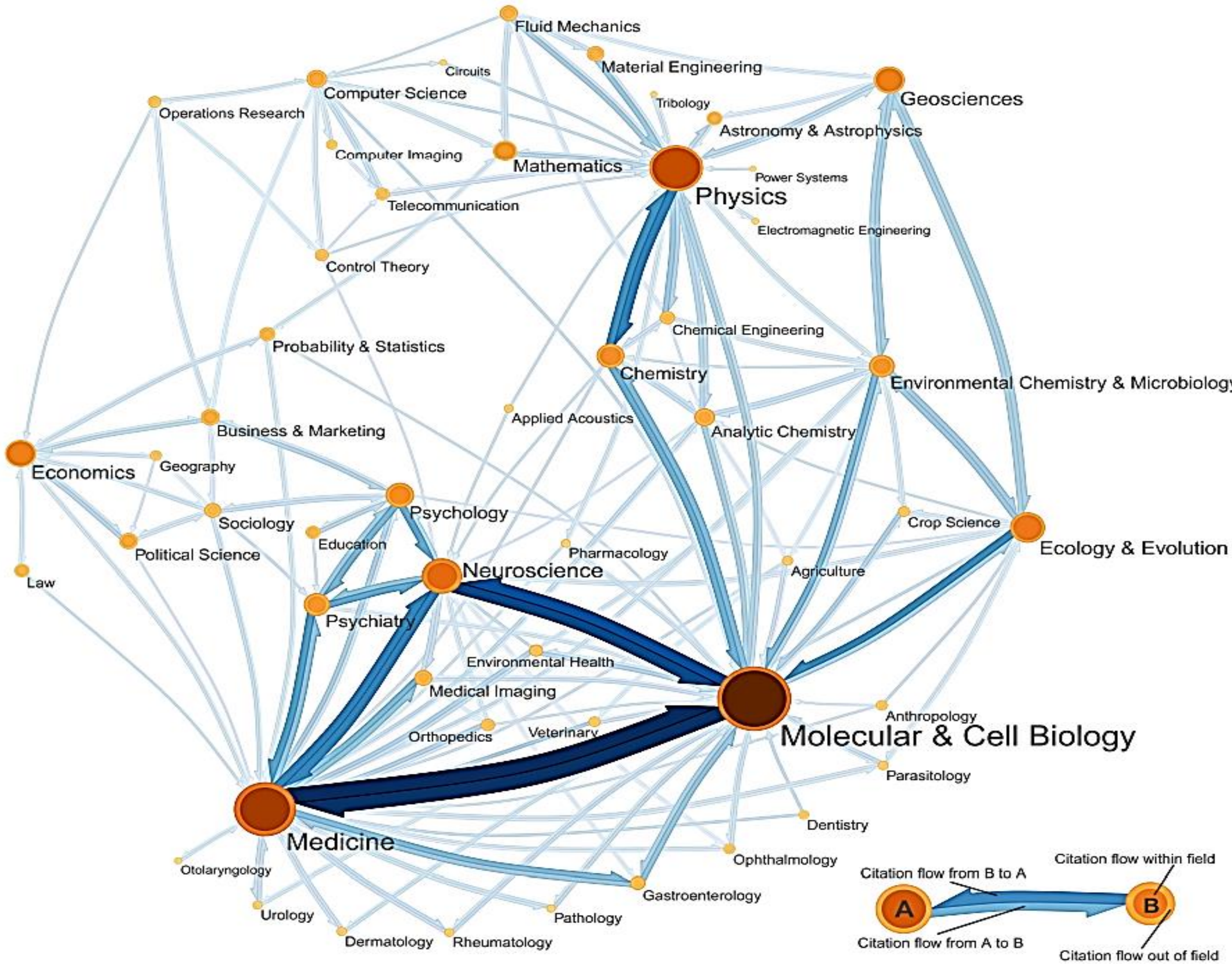
Social Web



Challenges:

1. As few labels as possible from crowd
2. Identify and give higher weight to experts
3. Derive a (globally) optimal labelling

Example: Community detection in social networks

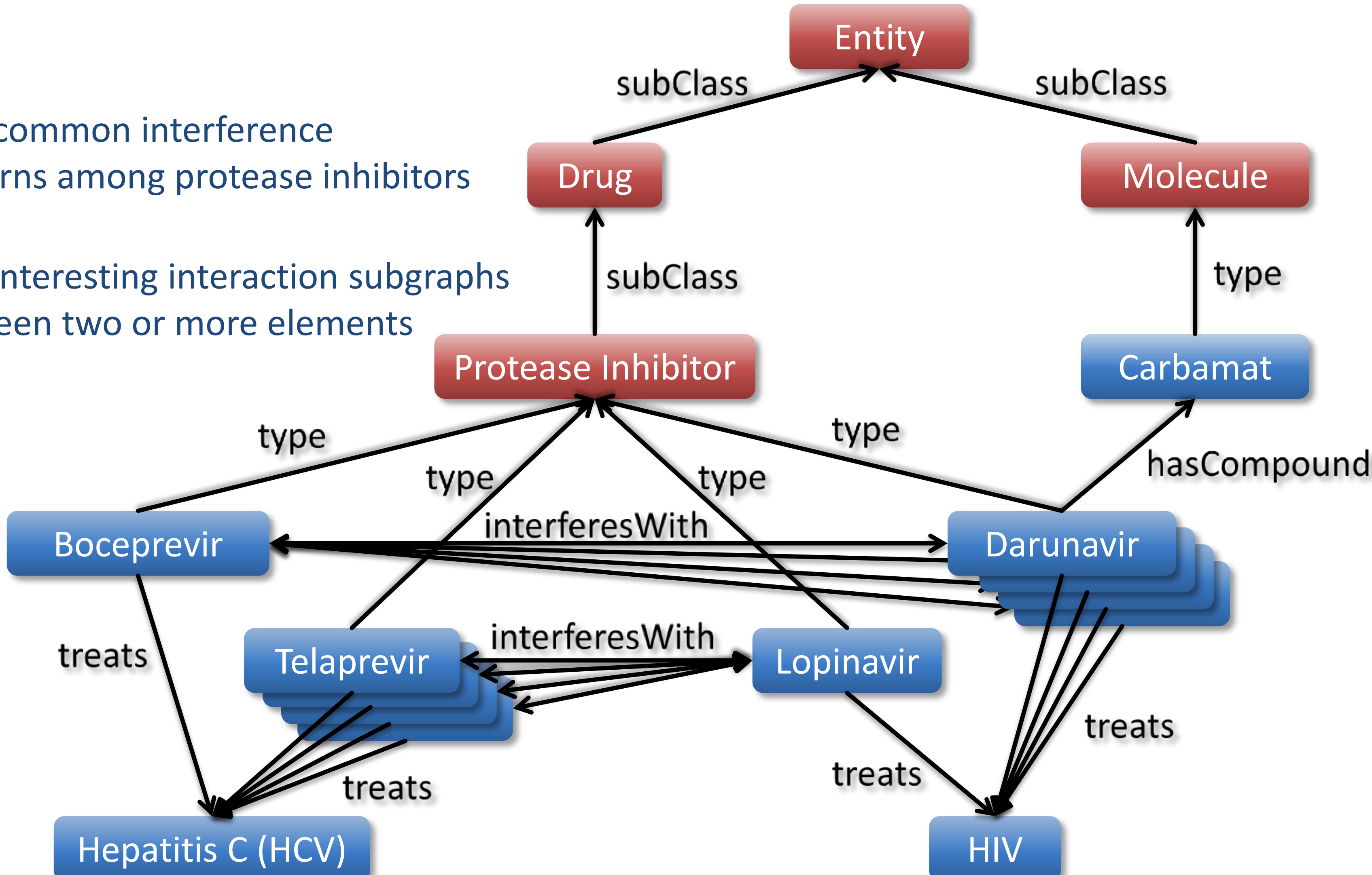


Source: S. Fortunato, Physics Reports 2010

Example: Knowledge discovery

Find common interference patterns among protease inhibitors

Find interesting interaction subgraphs between two or more elements



Important terms (1)

- **Predictive model / hypothesis:** Formalization of relationships between input and output variables with the goal of prediction

  Examples
  - $w_i = a + b * h_i + \epsilon$, e.g., weight is linearly dependent on height
  - $y \sim N(x, \sigma^2)$, i.e., y is normally distributed with mean x and variance σ^2
  - $P(\underbrace{l_1, \ldots, l_n}_{\text{grammatical labels}}, \underbrace{x_1, \ldots, x_n}_{n \text{ consecutive words}}) = P(x_1)P(x_1|l_1) \prod_{i \geq 2} P(x_i|x_{i-1})P(x_i|l_i)$

- **Parameterized statistical model**: Set of parameters and corresponding distributions that govern the data of interest

- **Learning**: Improvement on a task (measured by a **target function**) with growing experience

Important terms (2)

- **Training**: Sequence of observations from which experience can be gained
- **Target function**: Formal definition for the goal that has to be achieved

Possible goals

- Identify the “best next” item to label in active learning
- Maximize the joint probability of two or more observations (given some parameters)
- Predict the “best next” move in a chess game

Often, only an approximation of the “ideal” target function is considered

Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Example of a target function

- Task: Predict the number of retweets $V(\mathbf{t}_i)$ for a tweet $\mathbf{t}_i$

$$V(\mathbf{t}_i) \approx \hat{V}(\mathbf{t}_i = \underbrace{(t_1, t_2, \dots, t_k)^T}_{\text{features}}) = w_0 + w_1 t_{i1} + w_2 t_{i2} + \cdots + w_k t_{ik} = \mathbf{w}^T \mathbf{t}_i$$

Number of URLs

Number of hashtags

Number of possible readers

- Choosing an approximation algorithm
  - Learn a function $\hat{V}$ that predicts R_i based on $\mathbf{t}_i$ from training examples of the form $(\mathbf{t}_1 = (37,0,\dots,1)^T, R_1 = 0), \dots, (\mathbf{t}_n = (23879,3,\dots,0)^T, R_n = 214)$
  - $\hat{V}$ should minimize the training error $\frac{1}{2}\sum_{i=1}^{n}(R_i - V(\mathbf{t}_i))^2$

Inductive learning hypothesis and Occam's razor

- Suppose a learning algorithm performs well on the training examples
- How do we know that it will perform well on other unobserved examples?
- Lacking any further information, we assume the following hypothesis holds

  Any algorithm approximating the target function well over a sufficiently large set of training examples will also approximate it well over unseen examples ***(Inductive Learning Hypothesis)****.*

- But there may be many different algorithms that approximate the target function similarly well … Which one should be chosen?

  *Other things being equal, prefer the simplest hypothesis (****Occam's Razor****)*

Interesting questions related to learning algorithms

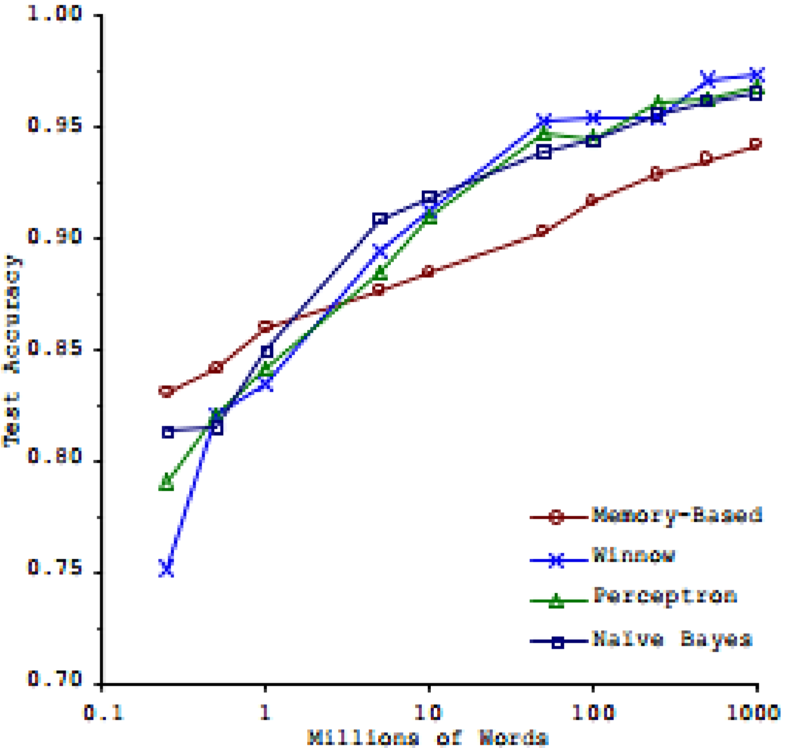
- How to (formally) represent training examples?
- How many examples are sufficient?
- What algorithms can be used for a given target function?
- How complex is a given learning algorithm?
- How can a learning algorithm quickly adept to new observations?

Learning with labeled data

- Which algorithm works best for Confusion Set Disambiguation (Banko & Brill ACL’01)?

  - Problem: Choose the correct use of a word, given a set of words with which it is commonly confused

  - Examples: {principle, principal}, {then, than}, {to, two, too}, and {weather, whether}



- Often, what matters is data!

Inductive bias is fine, there's no free lunch!

- **Inductive bias of a learning algorithm:** Set of assumptions that allow the algorithm to predict well on unseen examples

  Examples of inductive bias
  - (Conditional) independence assumption
  - Item belongs to same class as its neighbors
  - Select features that are highly correlated with the class (but uncorrelated with each other)
  - Choose the model that worked best on test data according to some measure

- **No Free Lunch Theorem (D. H. Wolpert & W. G. Macready 1997)**

  For any leaning algorithm, any elevated performance over one class of problems is offset by the performance over another class

Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Areas of learning theory

- Supervised Learning
  - Classification problems
  - Input: feature vector
  - Output: one of a finite number of discrete categories
- Unsupervised Learning
  - Clustering, dimensionality reduction, density estimation
  - Input: feature vectors
  - Output: similar groups of vectors, reduced vectors, or distribution of data from the input space
- Regression
  - Like classification but output is continuous
- Reinforcement Learning
  - Find suitable actions to maximize reward
  - Trade-off between exploration (trying out new actions) and exploitation (choose action with maximal reward)

Topics of this lecture

- Basics from probability theory, statistics, information theory
- Evaluation measures
- Hierarchical classifiers
- Linear classifiers
- Artificial neural networks
- Regression
- Clustering and topic models
- Graphical models (directed vs. undirected models)
- Factor graphs and inference
- Reinforcement learning

Related literature

- Literature
  - I. H. Witten, E. Frank, M. A. Hall: Data Mining - Practical Machine Learning Tools and Techniques (Chapters 1 – 6)
  - C. Bishop: Pattern Recognition and Machine Learning (Chapters 1 – 4, 8, 9)
  - T. M. Mitchell: Machine Learning (Chapters 3 – 6, 8, 10)
  - P. Flach: Machine Learning – The Art and Science of Algorithms that make Sense of Data (Chapters 1 – 3, 5 – 11)
  - D. J. C. MacKay: Information Theory, Inference and Learning Algorithms (Chapters 1 – 6)

- Important conferences
  - KDD, WSDM, ICDM, WWW, CIKM, ICML, ECML, ACL, EMNLP, NIPS, …

- Tools
  - The Weka Toolkit (http://www.cs.waikato.ac.nz/ml/weka/)
  - The **R** Project for Statistical Computing (http://www.r-project.org/)