EBERHARD KARLS UNIVERSITÄT TÜBINGEN
FACULTY OF SCIENCE
Data Science & Analytics

Data Mining and Probabilistic Reasoning

Organization

- **Lectures**: Mondays 14:00 - 16:00 in Room F119
- **Exercises**: Every 3rd week, (Oct. 29th; Nov. 19th; Dec. 10th; Jan. 14th 19; Feb. 4th 19)
- **Resources** on ILIAS:
  Veranstaltungen (Magazin)> Wintersemester 2018-2019>
  7 Mathematisch-Naturwissenschaftliche Fakultät> Informatik>
  Data Science & Analytics
- **Teaching assistants**:
  Vadim Borisov (vadim.borisov@uni-tuebingen.de),
  Johannes Haug (johannes-christian.haug@uni-tuebingen.de)
- **Exam**: End of the term; date and form tbd.

What is this lecture about?

Data Mining

- Analyzing data
- Processing and indexing data
- Finding patterns/structure
- Detecting outliers
- Learning predictive models
- Discovering knowledge

Probabilistic Reasoning

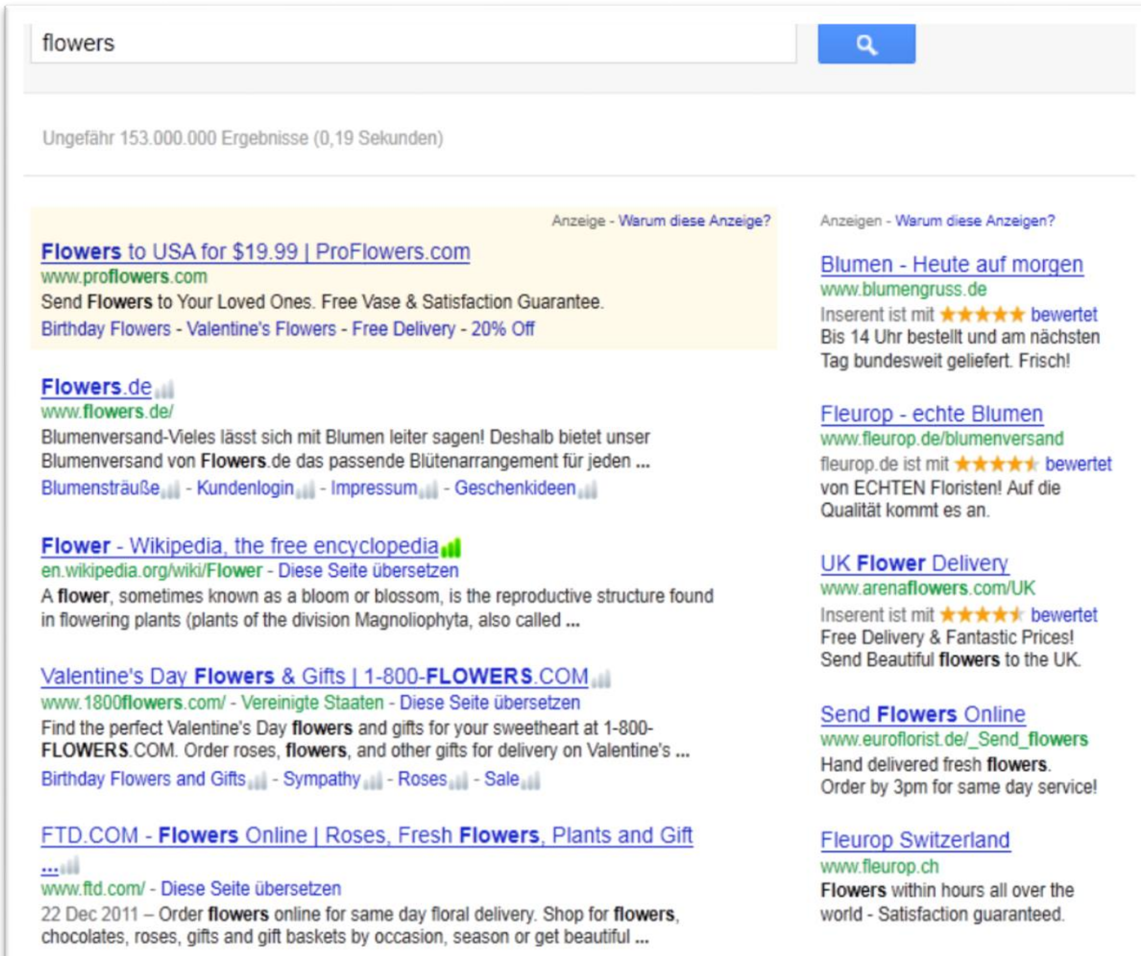
- Representing and quantifying uncertainty in data
- Computing probabilities and predicting outcomes of random variables, i.e., occurrence of events
- Choosing the model that best explains the data

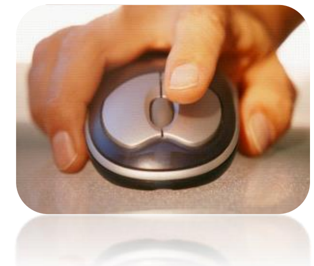
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Application areas

- **Web mining** (e.g., find documents for a given query or topic, group users by interest, ad ranking and recommendations, spam detection, …)
- **Medicine, Bioinformatics, Pharmaceutics** (e.g., diagnostics, analyze the effect of drugs, derive diagnose based on symptoms, analyze protein-protein interactions, discover sequence similarities, detect mutations, …)
- **Financial services & market analysis** (e.g., credit scoring and prediction of default, fraud detection, recommendation, market baskets, opinion mining, stock value prediction, influence propagation, …)
- **Automotive** (e.g. driving assistance, car diagnostics, self-driving cars, …)
- **Video games** (e.g., AI game characters, matching players in online gaming, speech/shape recognition, …)
- **Science, esp. Physics** (e.g., multivariate data analysis, modeling motion of particles, i.e., Brownian motion, event classification, noise detection, …)
- **Behavior analysis** (e.g., typical behavioral patterns, situation-based, socially driven, technology driven, …)

Example: Click prediction (ad ranking)





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

Example: Recommendation

Amazon recommendations

Collaborative filtering

… see also the Netflix Challenge

Example: Movie recommendation through matrix factorization

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

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

T1: Drama?
T2: Crime?
T3: Comedy?

Latent dimensions

Source: *Machine Learning* by P. Flach

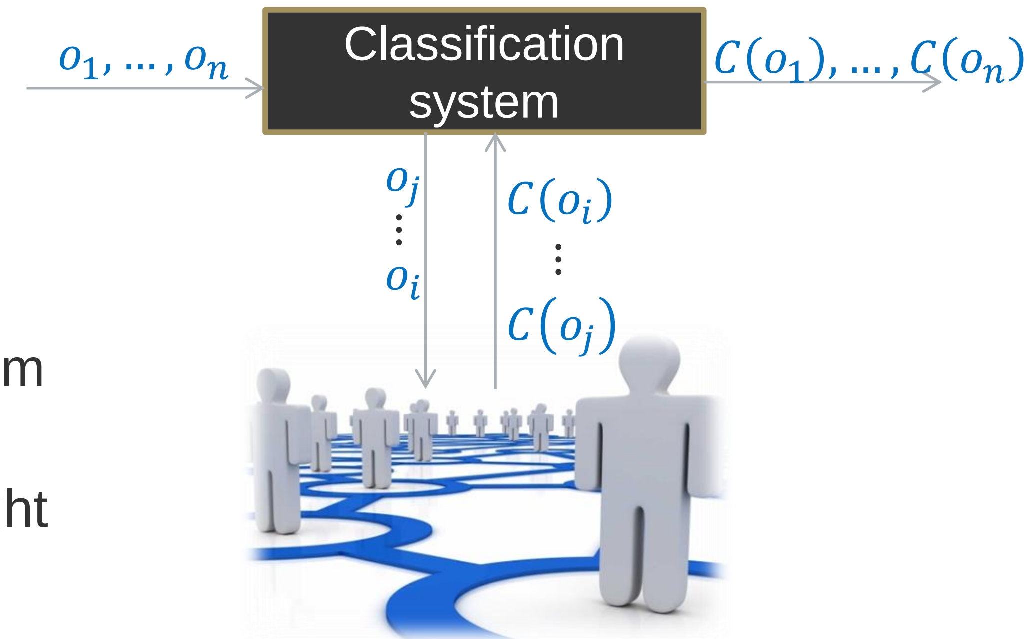
Example: Learning from crowds

Applications

- Label enrichment
- Truth discovery
- Opinion mining
- Data curation

Challenges

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



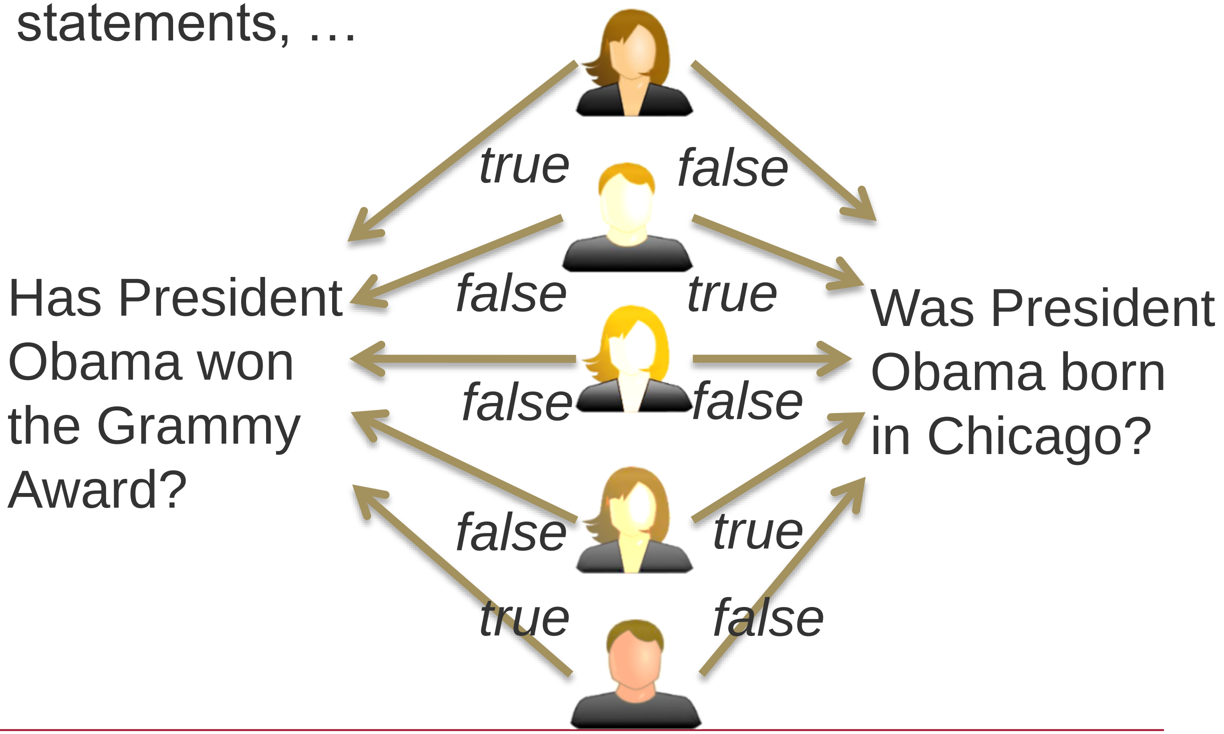
Social Web

Active learning scenario

Example: Latent truth discovery

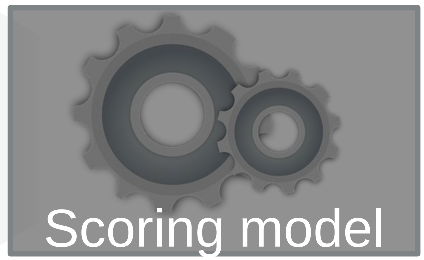
Task: Establish reliability of information sources and the truthfulness of statements made by those sources

Challenges: Inconsistent statements, missing statements, temporal changes, corrupted statements, …



Example: Credit scoring

Input
Credit history
Types of credit
Payment history
Credit cards
Length of history
Age
…

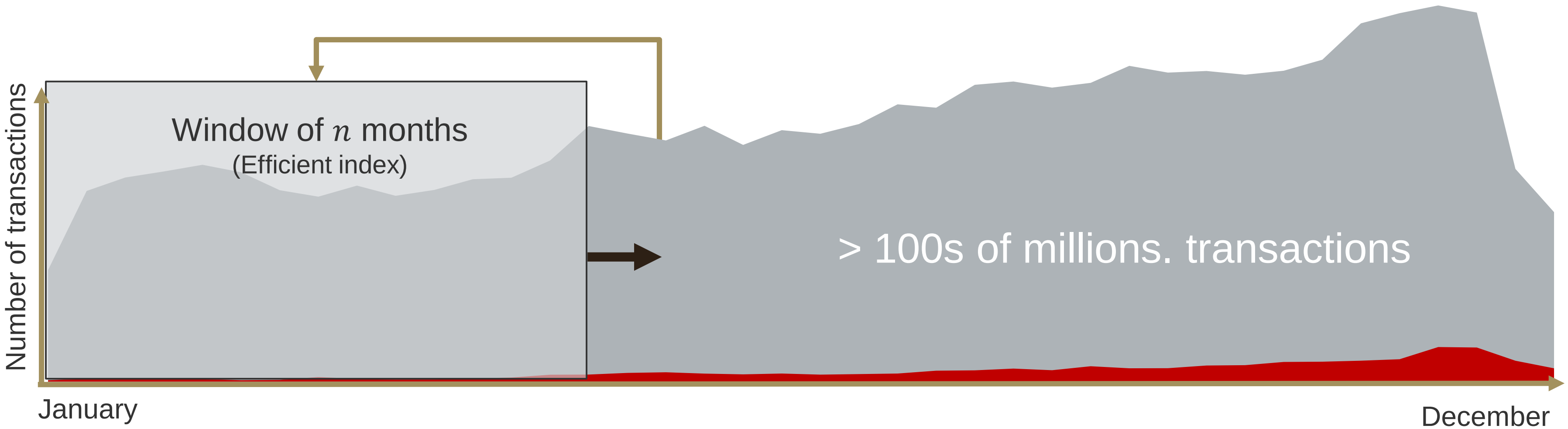


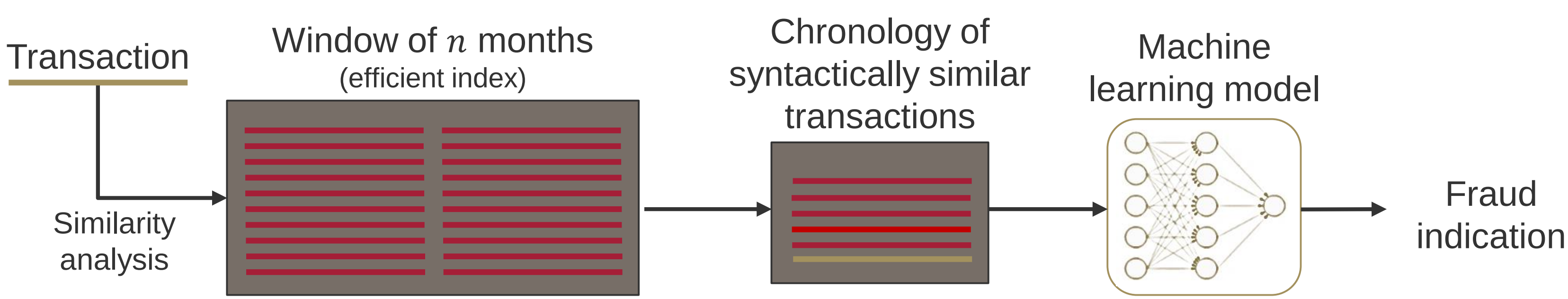
Category	Score	Population
A	9.863 – 9.999	0,80 %
B	9.772 – 9.862	1,64 %
C	9.709 – 9.771	2,47 %
D	9.623 – 9.708	3,10 %
E	9.495 – 9.622	4,38 %
F	9.282 – 9.494	6,21 %
G	8.774 – 9.281	9,50 %
H	8.006 – 8.773	16,74 %
I	7.187 – 8.005	25,97 %
K	6.391 – 7.186	32,56 %
L	4.928 – 6.390	41,77 %
M	1 – 4.927	60,45 %

Challenges
- Calibration (realistic predictions)
- Robustness (model performs well over time)
- Data minimization constraint (use only data that is relevant)

Source: https://www.schufa.de/de/unternehmenskunden/leistungen/bonitaet/geschaeft-privatkunden/schufa-branchenscores/

Example: Real-time fraud detection

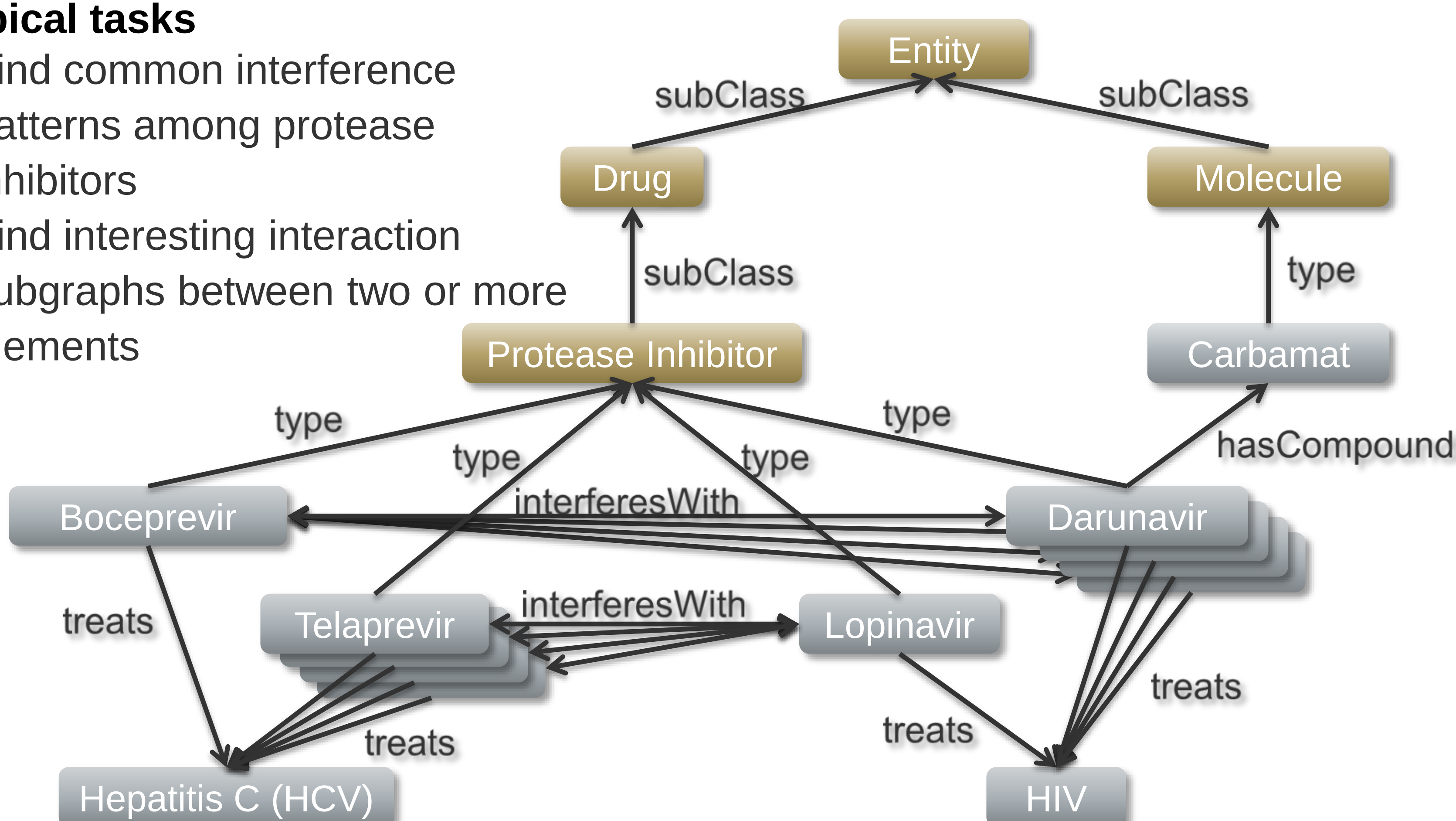




Example: Knowledge discovery

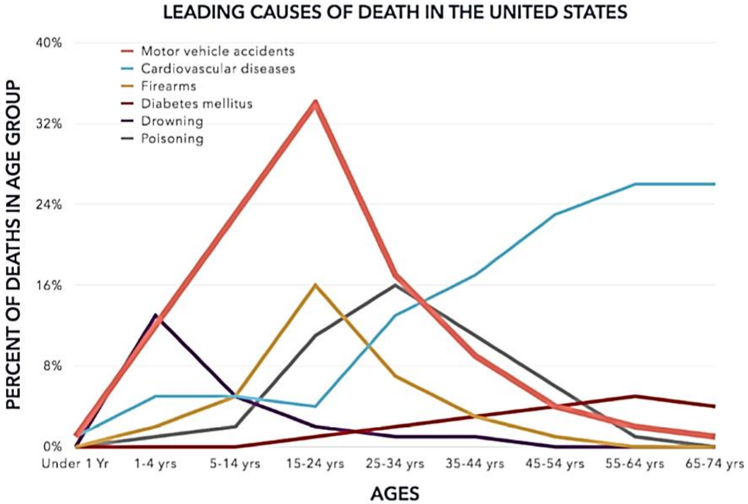
Typical tasks

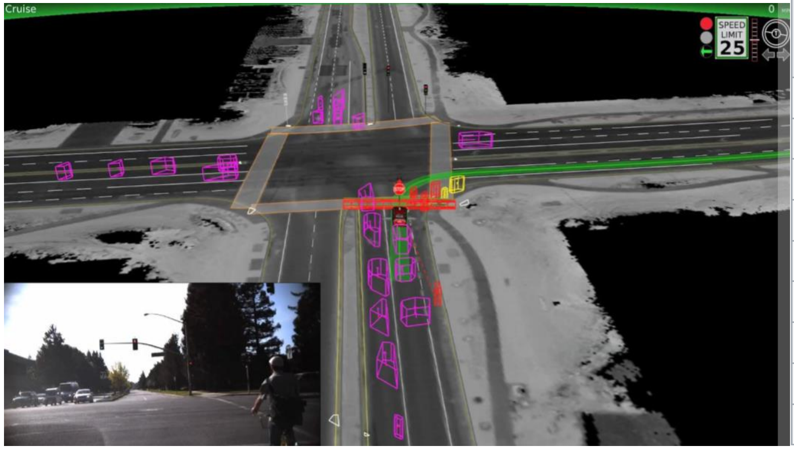
- Find common interference patterns among protease inhibitors
- Find interesting interaction subgraphs between two or more elements



Example: Self-driving cars

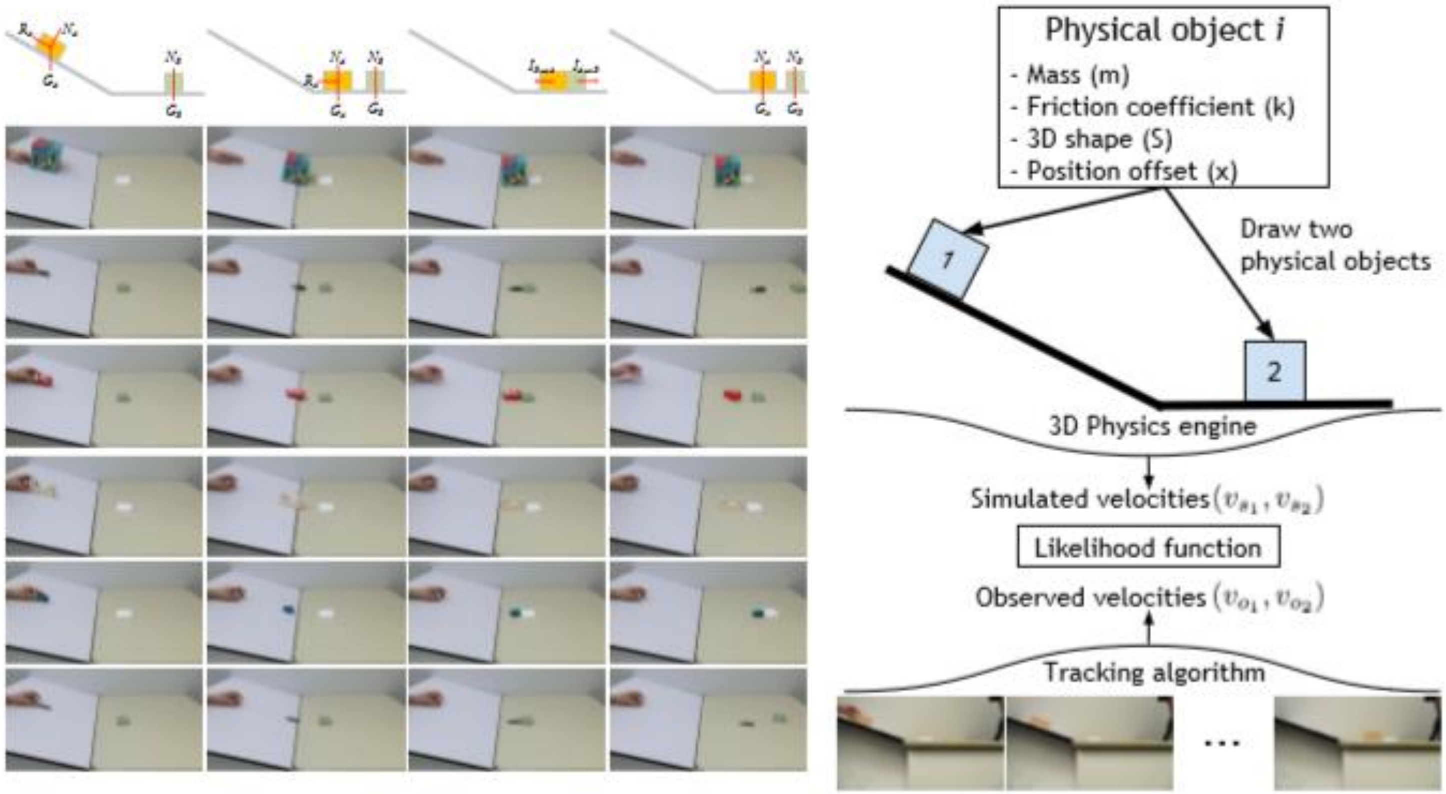
Source: *Google Self-driving Car Project*
Table source*:* Wikipedia





Maker	2016	
	Distance between disengagements	Distance
Google	5,127.9 miles (8,252.6 km)	635,868 miles (1,023,330 km)
BMW	638 miles (1,027 km)	638 miles (1,027 km)
Nissan	263.3 miles (423.7 km)	6,056 miles (9,746 km)
Ford	196.6 miles (316.4 km)	590 miles (950 km)
General Motors	54.7 miles (88.0 km)	8,156 miles (13,126 km)
Delphi Automotive Systems	14.9 miles (24.0 km)	2,658 miles (4,278 km)
Tesla	2.9 miles (4.7 km)	550 miles (890 km)
Mercedes Benz	2 miles (3.2 km)	673 miles (1,083 km)
Bosch	0.68 miles (1.09 km)	983 miles (1,582 km)

Example: Parameter estimation in physical systems



(a) (b)

Source: *Galileo: Perceiving Physical Object Properties by Integrating a Physics Engine with Deep Learning.* J. Wu, I. Yildirim, J.J. Lim, W.T. Freeman, J.B. Tenenbaum

A Big Data perspective

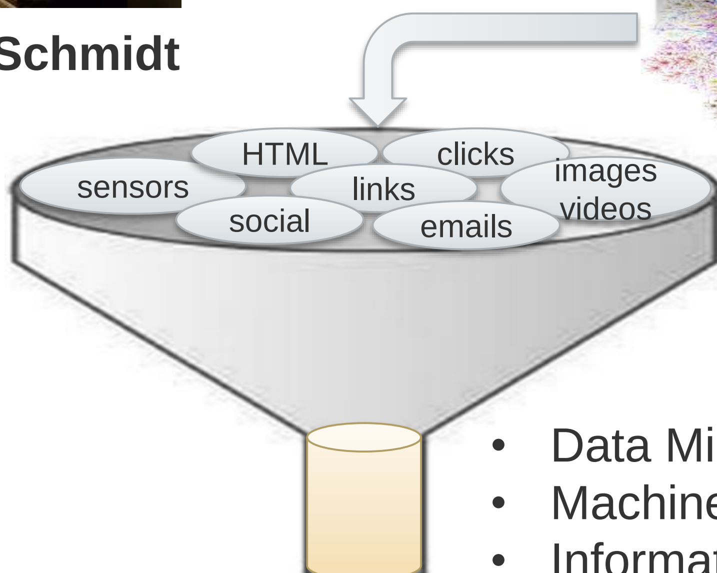
Sources:
Quote: https://techcrunch.com/2010/08/04/schmidt-data/
Photo: Wikipedia

"[…] 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



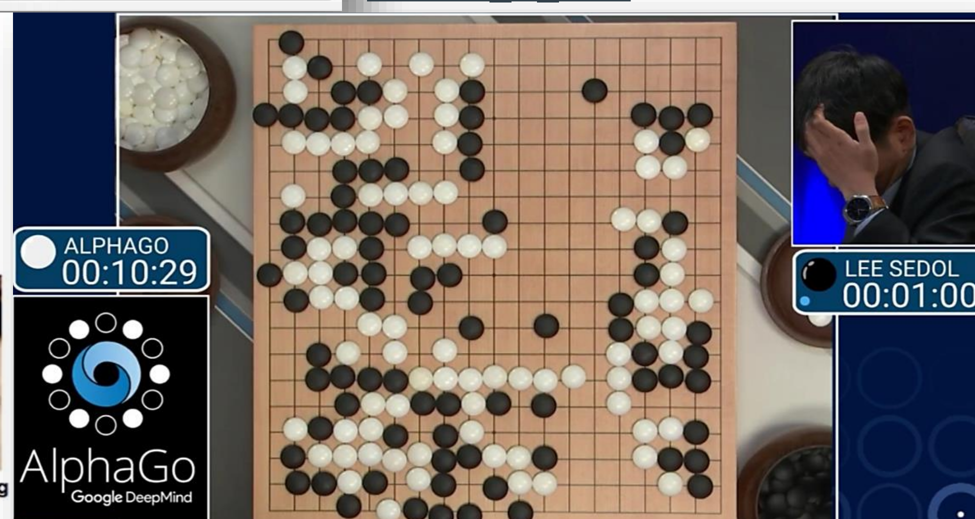
- Distributed databases
- Key-value stores
- Column stores
- Document databases

- Data Mining,
- Machine Learning,
- Information Retrieval

} Subfields of Artificial Intelligence (AI)

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Recent breakthroughs in AI



Sources:
(1)https://www.skype.com/en/features/skype-translator/
(2)https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Vinyals_Show_and_Tell_2015_CVPR_paper.html
(3)https://www.analyticsvidhya.com/blog/2017/01/introduction-to-reinforcement-learning-implementation/alphago-vs-lee-sedol-2_w_600/

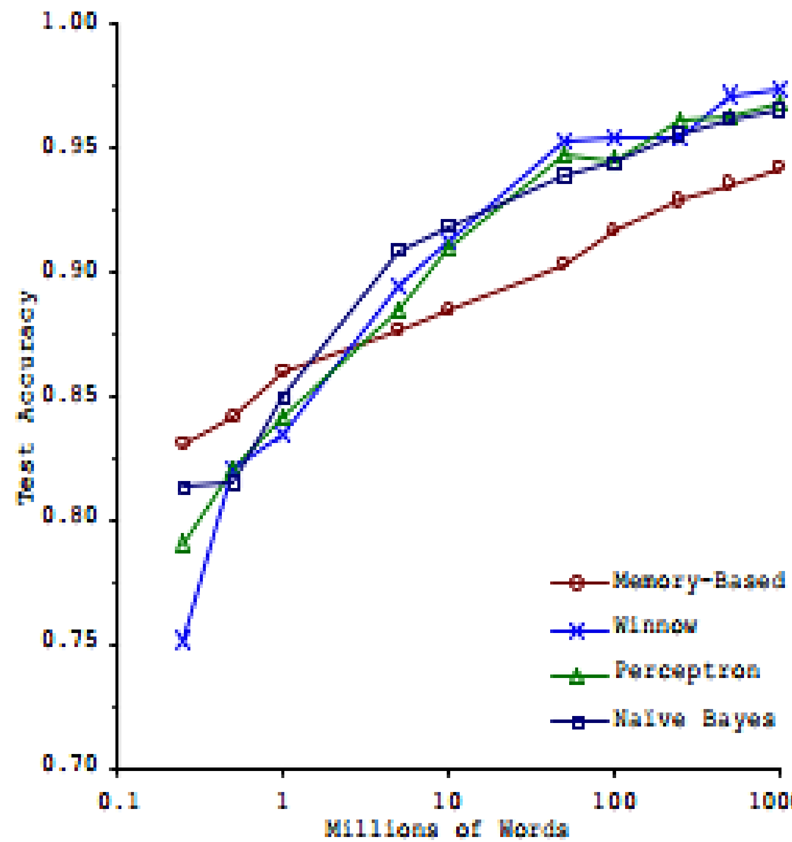
Main reasons for such breakthroughs

- **Large labelled datasets** (through social labelling and social engineering, crowd services, open datasets from public sector and industry, …)
- **Higher computing power** (better multi-core CPUs, GPUs, larger RAM, …)
- **Complex and refined algorithms** (deep learning, ensemble techniques, complex combinations of learning algorithms, e.g., deep reinforcement learning, …)

Learning with labelled data works better when more data is available

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}, {weather, whether}, …



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

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

Other important text analysis tasks

- Role labeling
- Entity recognition
- Entity disambiguation and extraction
- Relationship extraction
- Topic assignment (classification, clustering)
- Semantic understanding („AI-complete“ problem)

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Deep Learning architectures achieve lowest error rates

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, c
  - 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/

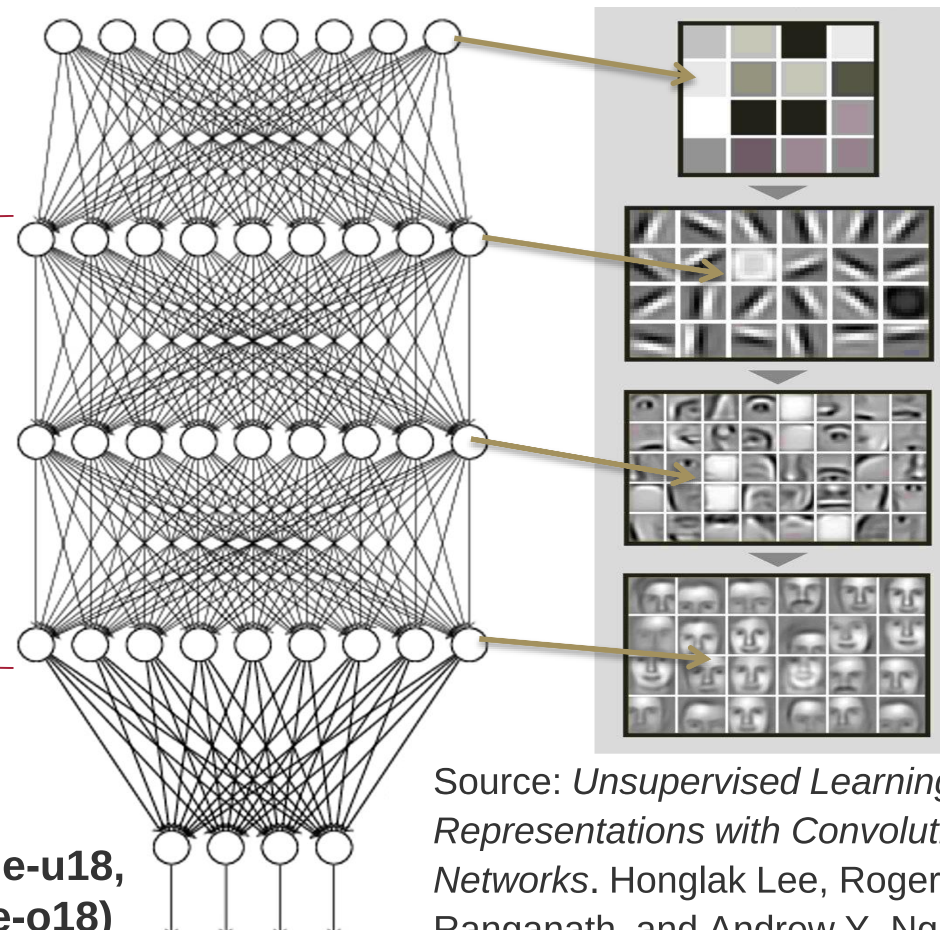
Deep neural networks (schematic overview)

Input layer

Hidden layers

Output layer

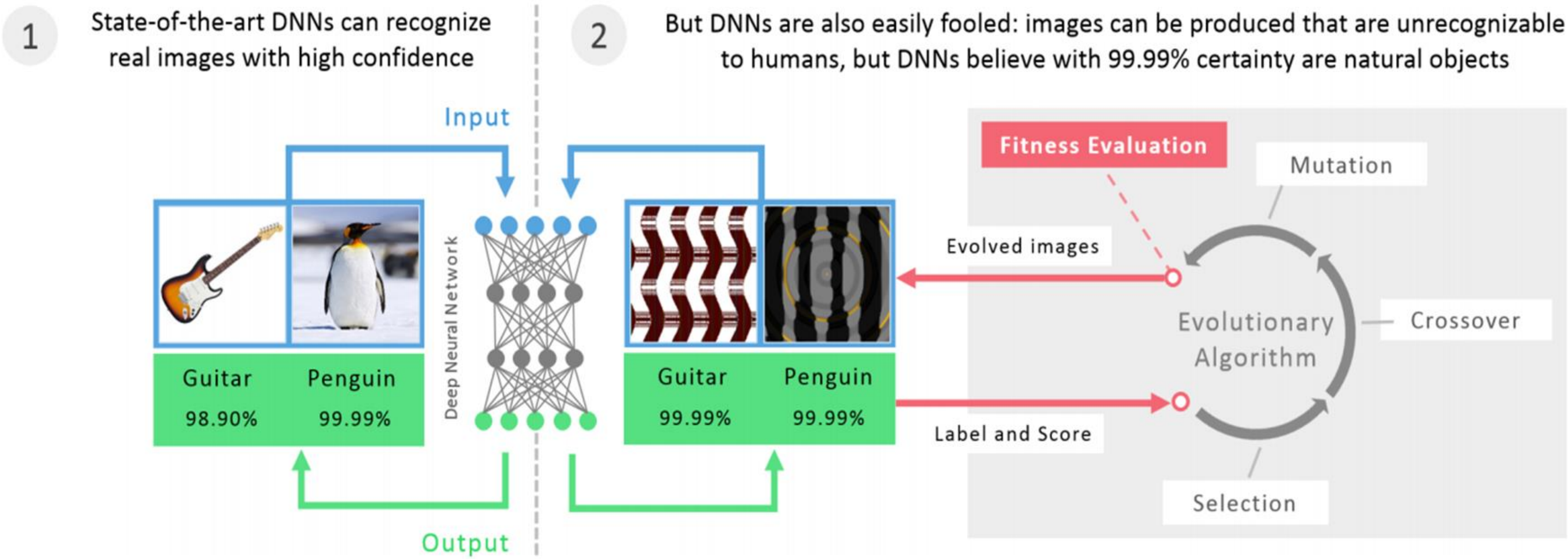
(male-u18, female-u18, male-o18, female-o18)



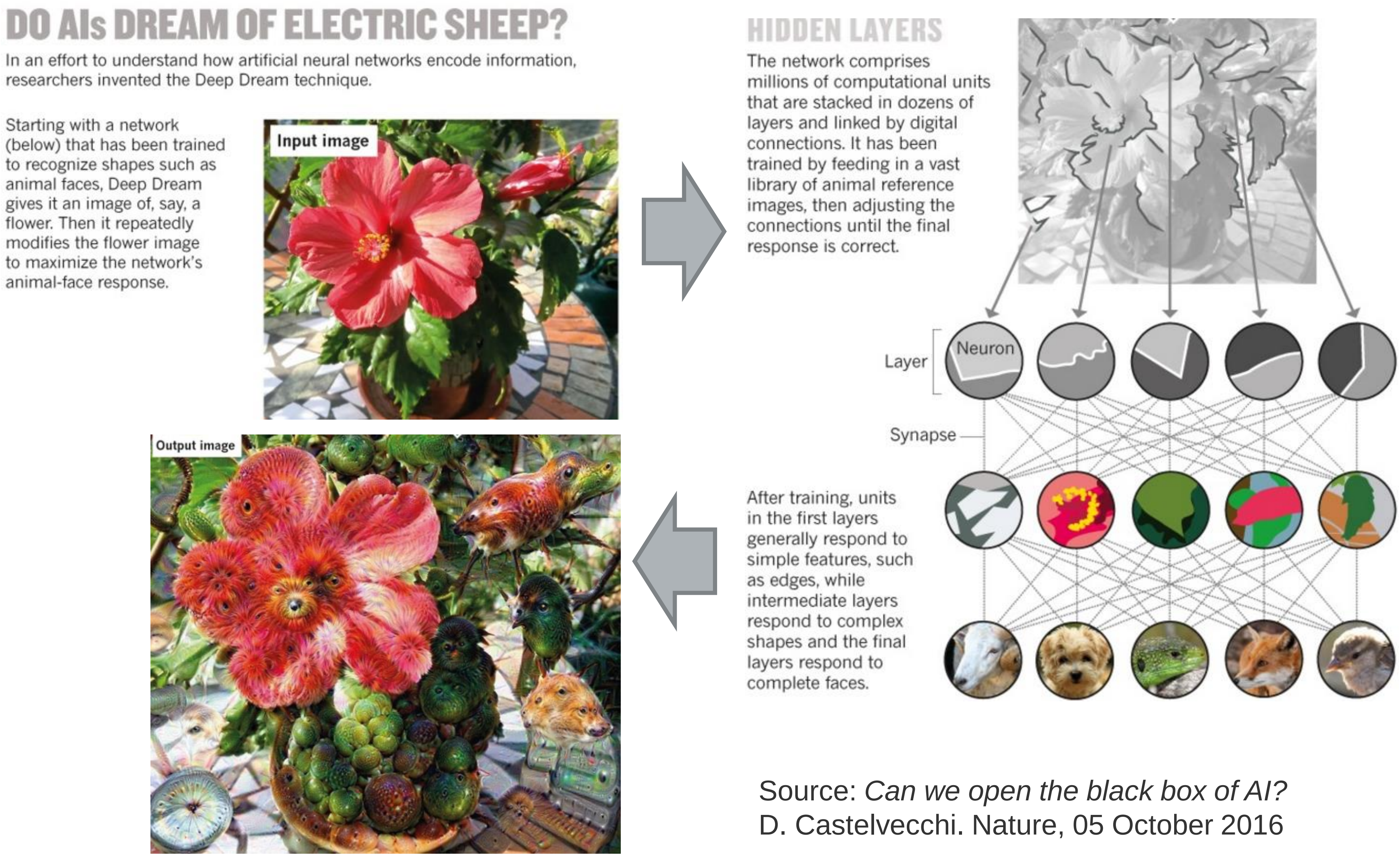
Source: *Unsupervised Learning of Hierarchical Representations with Convolutional Deep Belief Networks*. Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y. Ng; Com. ACM 2011

Foundation for a new research field: Adversarial Learning

Which patterns are recognized?



Source: *Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images.* A. Nguyen, J. Yosinski, J. Clune. Computer Vision and Pattern Recognition (CVPR ’15), IEEE, 2015



Source: *Can we open the black box of AI?*
D. Castelvecchi. Nature, 05 October 2016

Other issues concerning complex models

- **Explainability**
  - What is the influence of the features on the produced output?
  - How will the model react to certain changes to the underlying distribution of the input variables?
  - Is it possible to construct a simpler model with the same predictive power as the complex one?
- **Fairness, bias and variance**
  - What is fairness? How can conflicting definitions be handled?
  - How is the bias-variance tradeoff handled?
- **Model updates**
  - How is the model trained and updated?
  - How can quality assurance be handled?

Important terms (1)

- **Predictive model / hypothesis:** Formalization of relationships between input and output variables with the goal of prediction
- **Examples**
  - $w_i = a + b * h_i + \epsilon_i$, 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
  - $P(\underbrace{l_1, \dots, l_n}_{\text{grammatical labels}}, \underbrace{x_1, \dots, x_n}_{n \text{ consecutive words}}) = P(x_1)P(x_1|l_1)\prod_{i \geq 2} P(x_i|x_{i-1})P(x_i|l_i)$
- **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

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 most probable class given the observation
- $C^* = \text{argmax}_C P(C|\mathbf{m}) = \text{argmax}_C P\big(C|\underbrace{x_1(\mathbf{m}), x_2(\mathbf{m}), \dots, x_k(\mathbf{m})}_{\text{Feature values of } \mathbf{m}}\big)$

e.g., sender domain

e.g., indicates whether certain word appears

Important terms (2)

Training set: 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

Example of a target function

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

Number of URLs

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

Number of possible readers   Number of hashtags

Developing 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 $\frac{1}{2}\sum_{i=1}^{n}\left(\log R_i - \hat{V}(\mathbf{t}_i)\right)^2$

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 so-called **Inductive Learning 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.*
- But there may be many different algorithms that approximate the target function similarly well … Which one should we choose? **Occam’s Razor:** *Other things being equal, prefer the simplest hypothesis that explains your observations*

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?
- What is the computational complexity of a given learning algorithm?
- How can a learning algorithm quickly adapt to new observations?

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.

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
- Data preprocessing
- Indexing for efficient similarity search
- Evaluation measures for supervised learning
- Linear classifiers
- Non-linear classifiers
- Regression
- Clustering and topic models
- Graphical models (directed vs. undirected models)
- Factor graphs and inference

Related literature

- I. H. Witten, E. Frank, M. A. Hall: *Data Mining - Practical Machine Learning Tools and Techniques*
- J. Han, M. Kamber, J. Pei: *Data Mining: Concepts and Techniques*
- C. Bishop: *Pattern Recognition and Machine Learning*
- T. M. Mitchell: *Machine Learning*
- P. Flach: *Machine Learning – The Art and Science of Algorithms that make Sense of Data*
- D. J. C. MacKay: *Information Theory, Inference and Learning Algorithms*
- I. Goodfellow, Y. Bengio, A. Courville: *Deep Learning*
- L. Wasserman: *All of statistics*
- J. Leskovec, A. Rajaraman, JD Ullman: *Mining Massive Datasets*

Conferences, tools and datasets

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

Tools

- Scikit-Learn (http://scikit-learn.org/stable/)
- SciPy (https://www.scipy.org/)
- The Weka Toolkit (http://www.cs.waikato.ac.nz/ml/weka/)
- The **R** Project for Statistical Computing (http://www.r-project.org/)

Open datasets

- Kaggle datasets (https://www.kaggle.com/datasets)
- UCI datasets (https://archive.ics.uci.edu/ml/datasets.html)
- Weka datasets

Contact Information

Gjergji Kasneci
gjergji.kasneci@uni-tuebingen.de

Consultation hours: By appointment