Universiteit Antwerpen
Faculteit Bedrijfswetenschappen
en Economie

Big Data

Hoe vinden we de weg in een massa gegevens?

Prof. David Martens

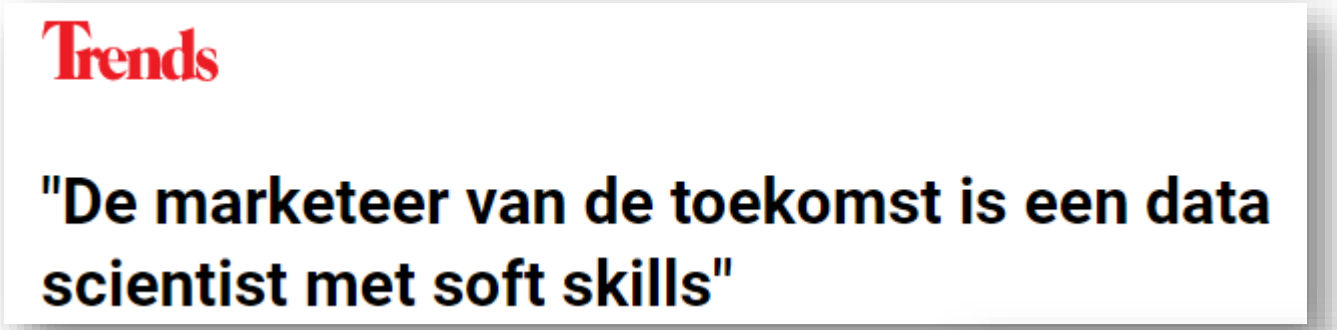
Introduction

- **About us**
  - PhD researchers at Faculty of Business & Eonomics
  - Research on Data Mining and Ethical AI

- **Economie Ontcijferd**
  - Clarify role of mathematics in Business & Economics

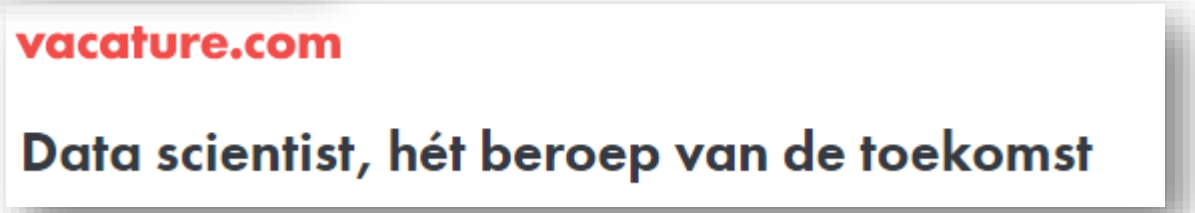
Big Data

- **Introduction to Big Data and Data Mining**
- **Decision Trees**
- **Artificial Neural Networks**
- **Ethics of Big Data**


Trends
"De marketeer van de toekomst is een data scientist met soft skills"


DE TIJD
Nieuws
Markten LIVE
Netto
Home
Meest recent
Politiek & economie
Ondernemen
Tech & media
Opinie
Cultuur
Sabato
'Dringend nood aan onderwijs in big data'

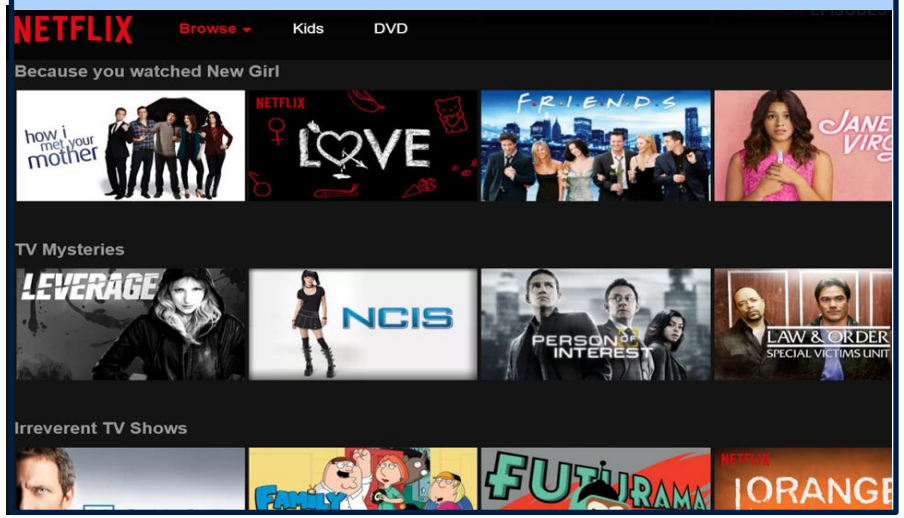

FINANCIAL TIMES
Data science is the big draw in business schools
Student demand for degrees in the subject soars as employers seek skilled analysts


vacature.com
Data scientist, hét beroep van de toekomst

Data Mining

“The automatic extraction of patterns from large amounts of data”

Recommender Systems

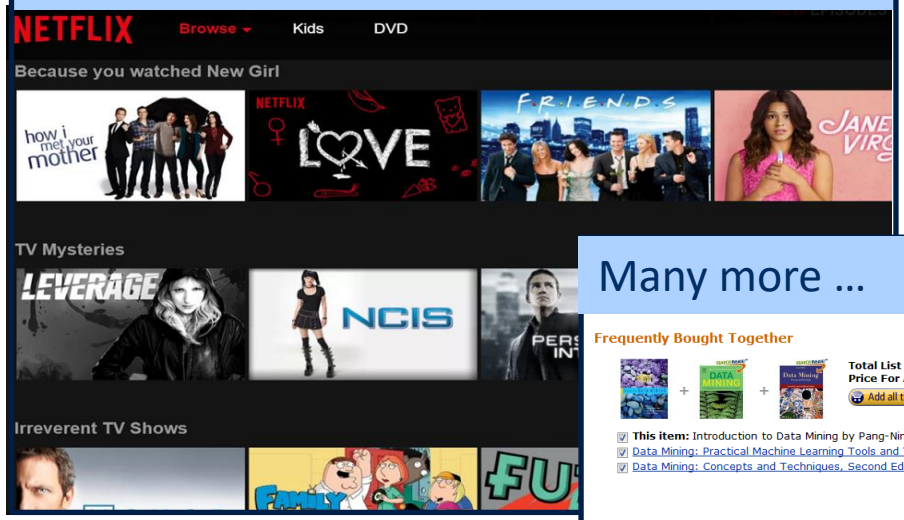


Targeted Advertising

Fraud Detection

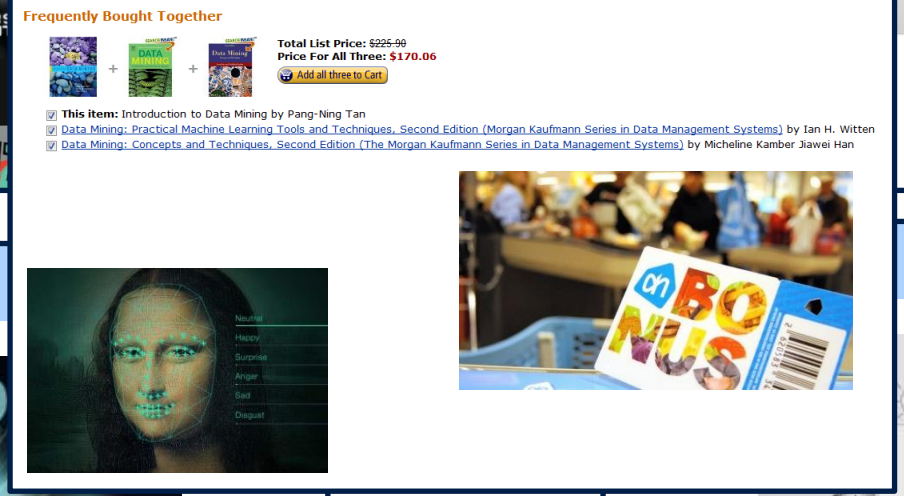
HR Analytics

Recommender Systems



Targeted Advertising

Many more …



Fraud Detection

Data Mining

- **Data mining: automatic extraction of knowledge from data**
- **Setting the scene with credit scoring example**

Client	Income	Sex	Amount	Default
A	1.600	M	175.000	N
B	2.600	F	350.000	Y
C	3.280	M	50.000	N
D	950	M	120.000	Y
E	10.500	M	1.000.000	N
F	5.700	F	240.000	N
G	2.400	F	250.000	N

Data Mining

Classification Model

if income < 10.000 **and** Amount Loan > 100.000 **and** ...
then default = yes

Client	Income	Sex	Amount	Default
New client	2.000	F	500.000	Y

Data

Data mining technique

Pattern

Classification: Classification Models

Mathematical classification models f(x)

$f(x) > 0.5 \Rightarrow$ customer = good
$f(x) \leq 0.5 \Rightarrow$ customer = bad

- **Linear**
  - Linear, logistic regression; linear discriminant analysis
  - Result: linear function of attributes
  - $f(x) = 0.125\ \text{income} + 0.305\ \text{age} - 0.02\ \text{gender} + \ldots - 3.1\ \text{amount loan} + 0.3$

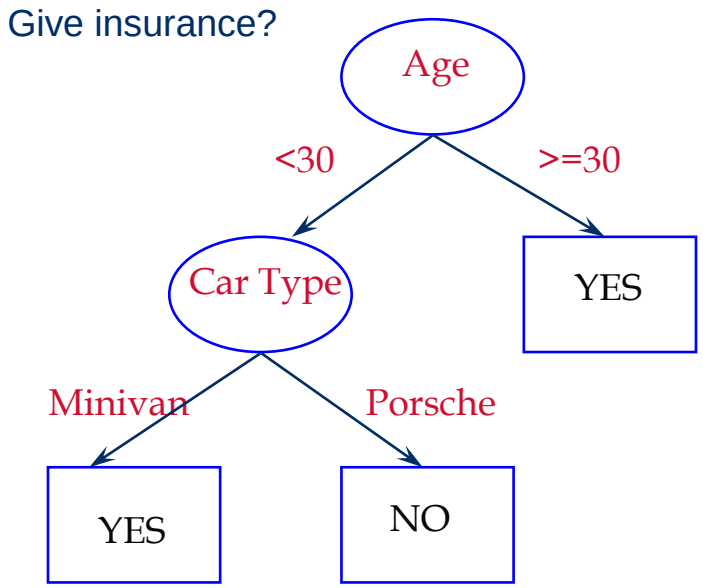
- **Non-linear**
  - Artificial Neural Networks, Support Vector Machines, RVM, …
  - Result: non-linear function of attributes
  - $f(x) = 0.201\ \text{income}^2\ \text{age}^3 - 0.55\ \text{age}^3 - 5.21\ \text{gender income} + \ldots + 3.6\ \text{gender}^2\ \text{amount loan}^2$

Classification: Classification Models

Rule-based classification models

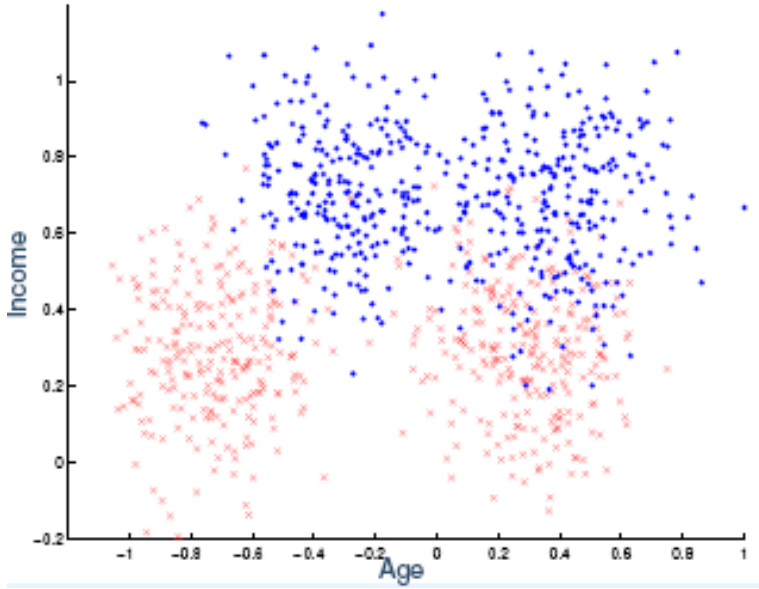
- Decision Rules / Trees
  - C4.5, RIPPER, CN2, AntMiner+, ANN/SVM Rule extraction…
  - Result: set of rules or tree

if (Checking Account $<$ 200 DM **and** Duration $>$ 15 m **and**
Credit History = no credits taken **and** Savings Account $<$ 1000 DM)
then class = bad
else if (Purpose = new car/repairs/education/others **and**
Credit History = no credits taken/all credits paid back duly at this bank **and**
Savings Account $<$ 1000 DM)
then class = bad
else if (Checking Account $<$ 0 DM **and**
Purpose = furniture/domestic appliances/business **and**
Credit History = no credits taken/all credits paid back duly at this bank **and**
Savings Account $<$ 500 DM)
then class = bad
else if (Checking Account $<$ 0 DM **and** Duration $>$ 15 m **and**
Credit History = delay in paying off in the past **and**
Savings Account $<$ 500 DM)
then class = bad
else class = good



Classification: Output Types

Different output types

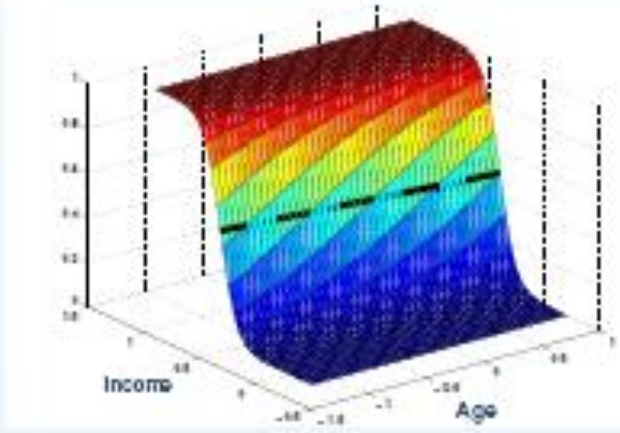


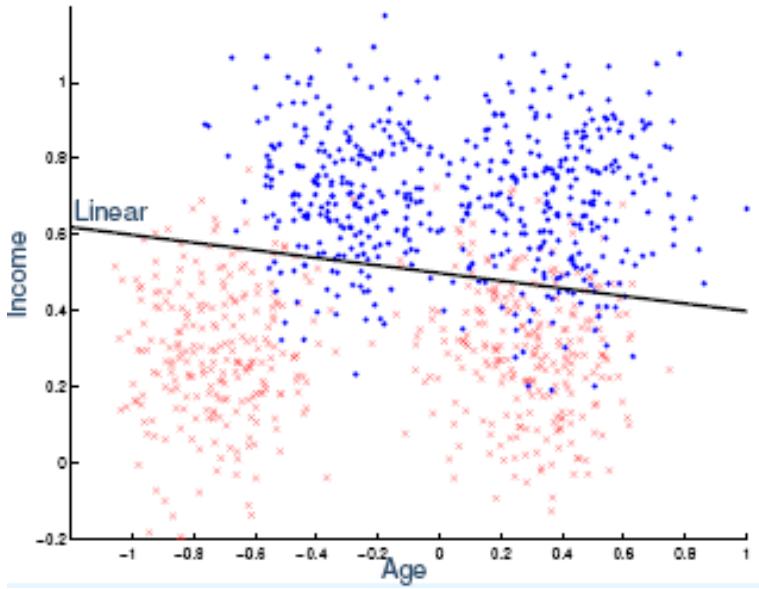
Classification: Output Types

Different output types

- Linear

$$\mathcal{Y}_{linear} = \frac{1}{1+\exp(-(w_0+w_1\times Age+w_2\times Income))}$$



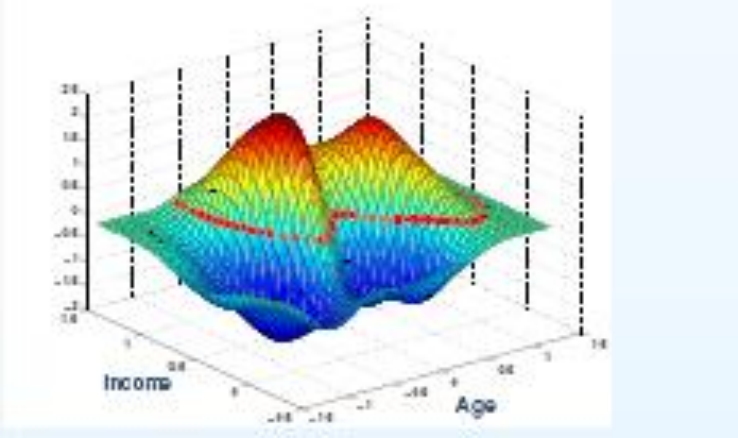


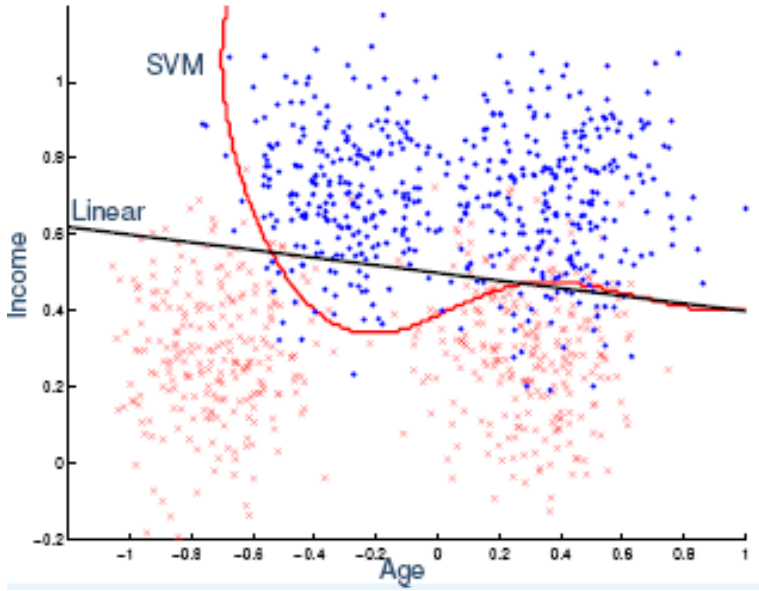
Classification: Output Types

Different output types

- Linear
- Non-linear

$$y_{SVM} = \text{sign}\left[\sum_{i=1}^{N} \alpha_i y_i \exp\{-\|x - x_i\|_2^2/\sigma^2\} + b\right]$$



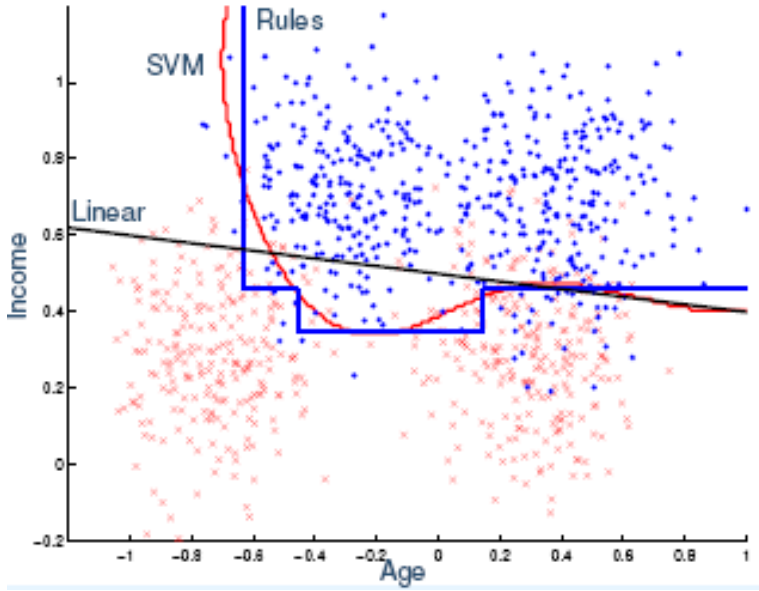


Classification: Output Types

Different output types

- Linear
- Non-linear
- Rule-based

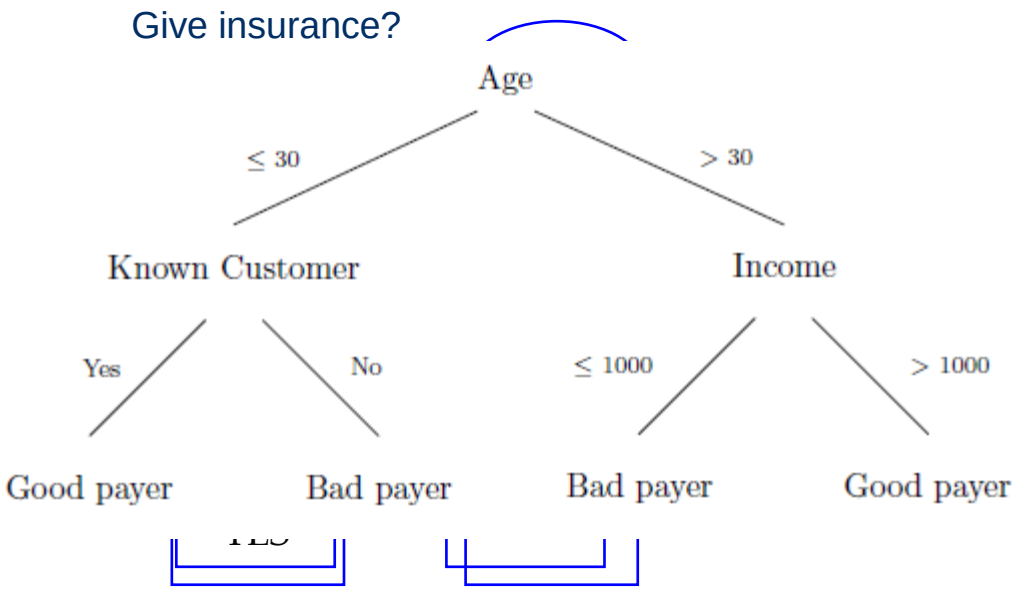
```
If age < 24 then Bad
Else If Income < 2000 then Bad
Else Good
```



Big Data

- Introduction to Big Data and Data Mining
- Decision Trees
- Artificial Neural Networks
- Ethics of Big Data

Decision Trees

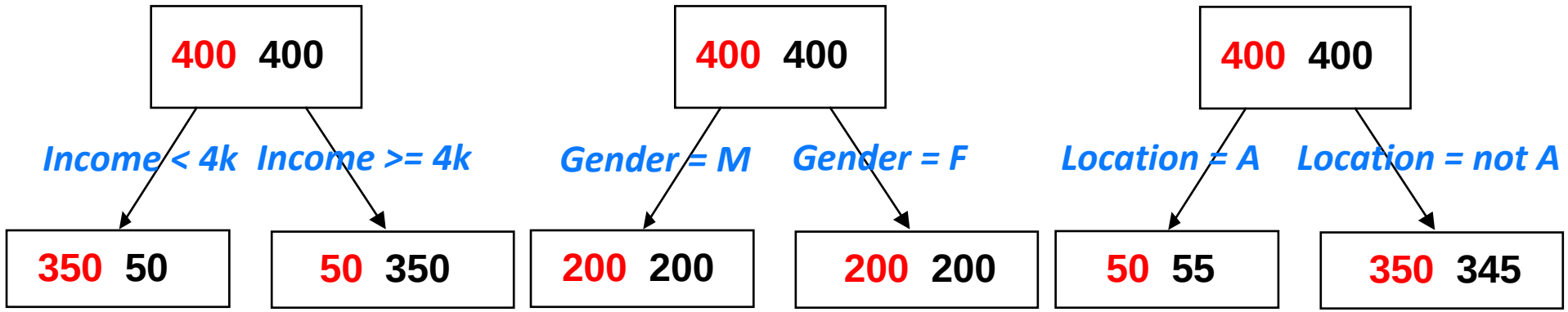


Finding informative variables from the data

- **Fundamental notion of data science:**
  - Finding and selecting **informative** variables
  - What is information? Reduces uncertainty about something.
  - *"So, if an old pirate gives me information about where a treasure is hidden …"*

Finding informative variables from the data

- **Predicting credit default**
- **Suppose your bank has this data**
  - 800 customers, 400 known **good** ones, 400 known **bad** ones
  - 3 variables: income, gender and location
- Which variable is “most informative”?



Finding informative variables from the data

- **Measuring the uncertainty**

- ***"Technically, we would like the resulting groups to be as pure as possible" –***

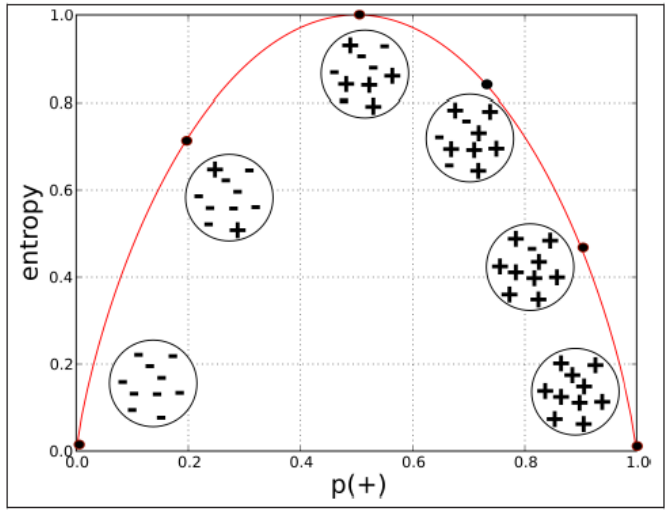


Figure 3-3. Entropy of a two-class set as a function of p(+).

*Why "**Entropy**"? The story goes that Shannon didn't know what to call his new information measure, so he asked **von Neumann**, who said `You should call it **entropy** ... [since] ... no one knows what entropy really is, so in a debate you will always have the advantage' (Tribus 1971)*

Decision Trees: Impurity

Measuring the uncertainty

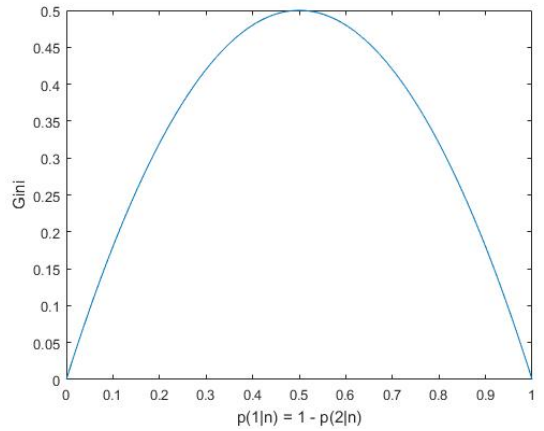
- **Impurity I(n) of node n**
  - at maximum when observations are distributed evenly over all classes
  - at minimum when all observations belong to a single class

- **Two popular measures:**
  1. **Entropy** measure

     $I(t) = -p(1|n) \log(p(1|n)) - p(2|n) \log(p(2|n))$

  2. **Gini** index of diversity

     $I(t) = 2\, p(1|n)\, p(2|n)$



Decision Trees: Impurity

Measuring the uncertainty

Gini index of diversity

$I(t) = 2\, p(1|n)\, p(2|n)$

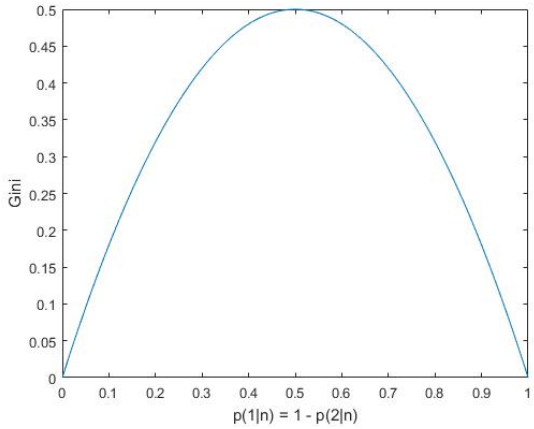
All customers

400	400
B	G

$p(1|n) = 400 / 800$

$p(2|n) = 400 / 800$

$I(n) = 2 \times 0{,}5 \times 0{,}5 = 0{,}5$



Customers with high income

100	300
B	G

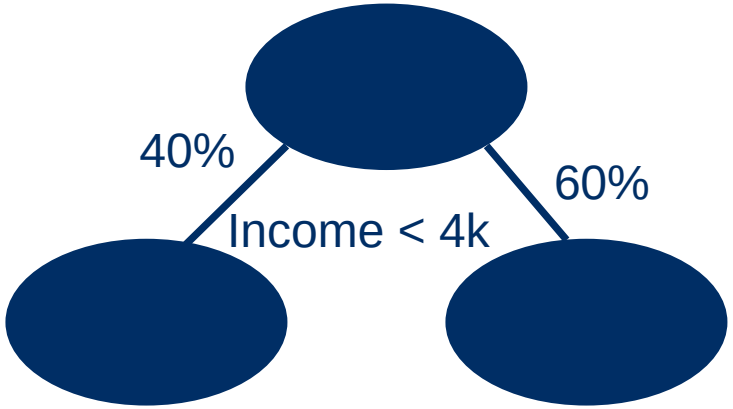
$p(1|n) =$

$p(2|n) =$

$I(n) =$

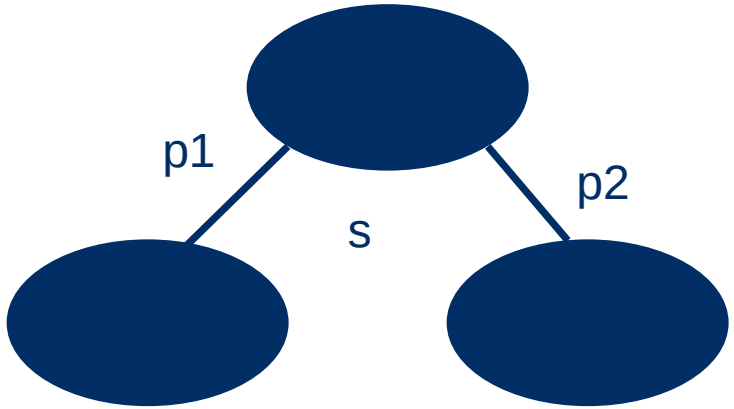
What is the uncertainty about the class to predict, for this group (as measured by Gini)? Indicate the point on the graph.

Decision Trees



Measuring the <u>reduction</u> in uncertainty

- **Consider candidate split s of node n**
- **Notations**
  - p_1: proportion of the data in n that ends up in n1
  - I(n): the impurity of node n

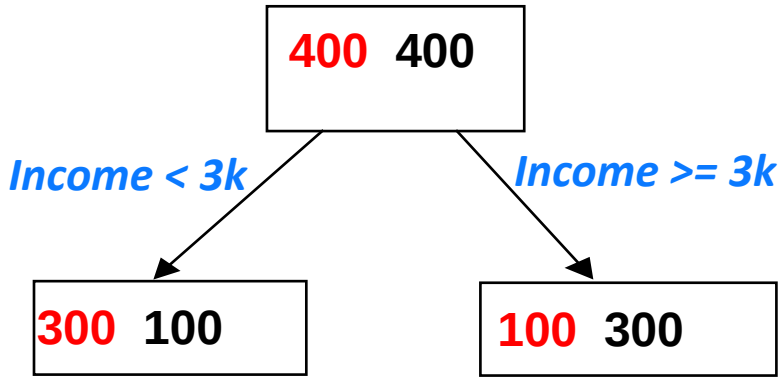


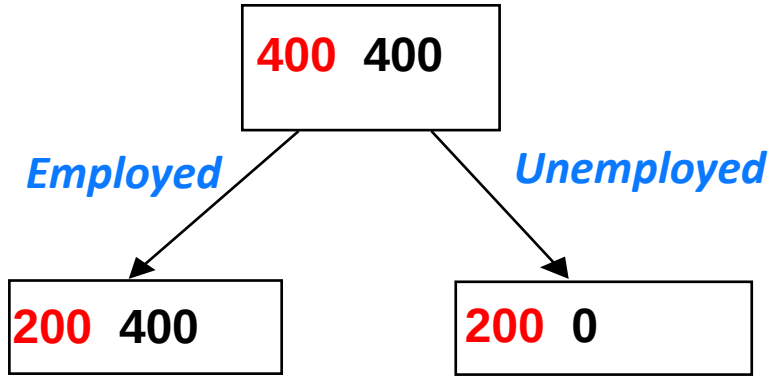
- **The goodness of the split is weighted: mean decrease in impurity**

$$\Delta I(s,n) = I(n) - p_1\, I(n_1) - p_2 I(n_2)$$

Decision Trees: Impurity

Notations:
- I(n): impurity at node n
- p(1|n): probability of being class 1 at node n
- p(2|n): probability of being class 2 at node n





I(n) = impurity of top node
I(n1) = impurity of left node
I(n2) = impurity of right node

Gini: I(n) = 2 x p(1|n) x p(2|n)

I(n) = 2 x (1/2) x (1/2) = 0,5
I(n1) = 2 x (3/4) x (1/4) = 0,375
I(n2) = 2 x (1/4) x (3/4) = 0,375

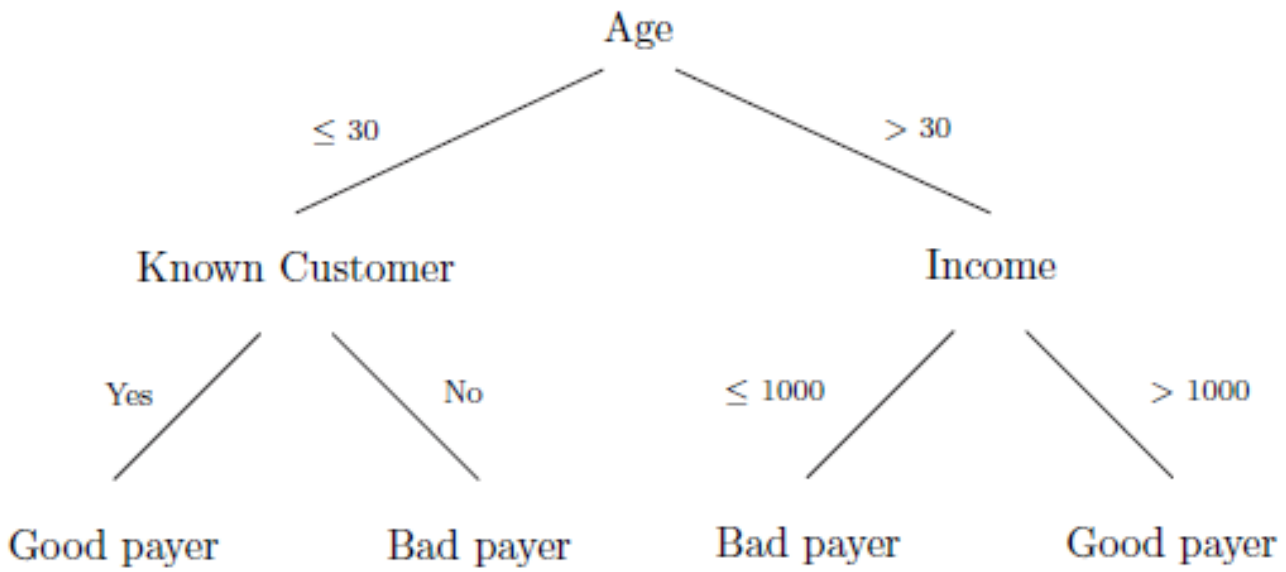
ΔI(n)= I(n) – (400/800) x I(n1) – (400/800) x I(n2)
= 0,5 – (1/2) x 0,375 – (1/2) x 0,375
= **0,125**

I(n) =
I(n1) =
I(n2) =

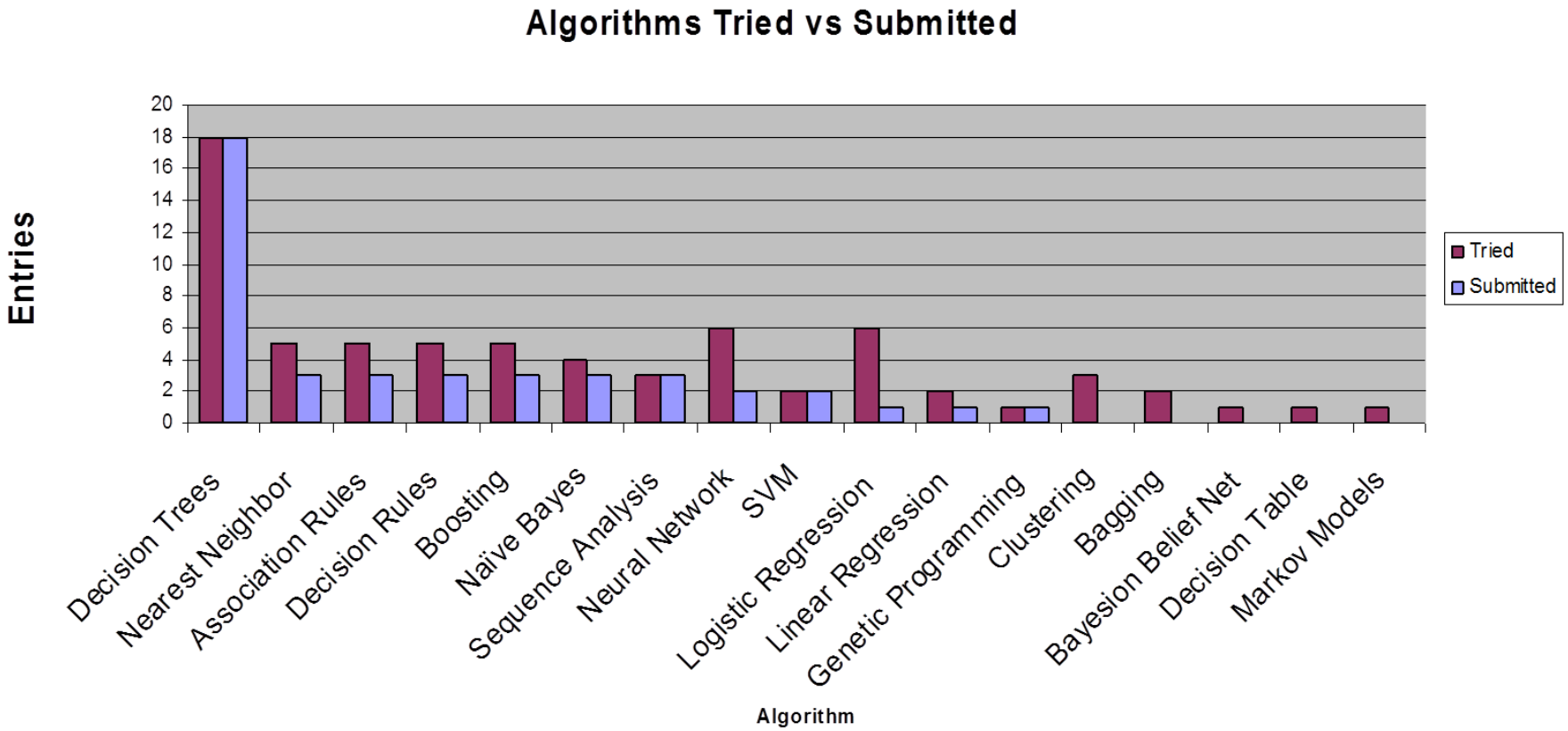
ΔI(n)=
=
= **0,166**

Decision Trees

Decision Tree algorithm: recursively repartioning the data



Commonly Used Induction Algorithms



Post-mortem analysis of a popular data mining competition

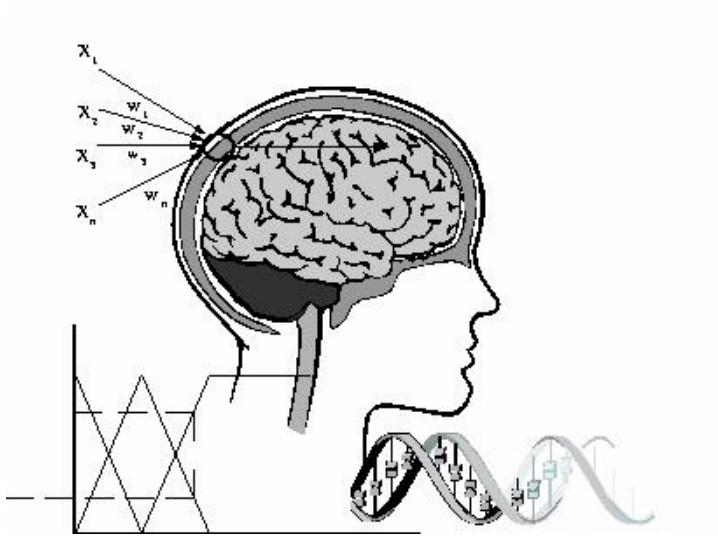
Thanks to Carla Brodley & Ron Kohavi

Big Data

- Introduction to Big Data and Data Mining
- Decision Trees
- **Artificial Neural Networks**
- Ethics of Big Data

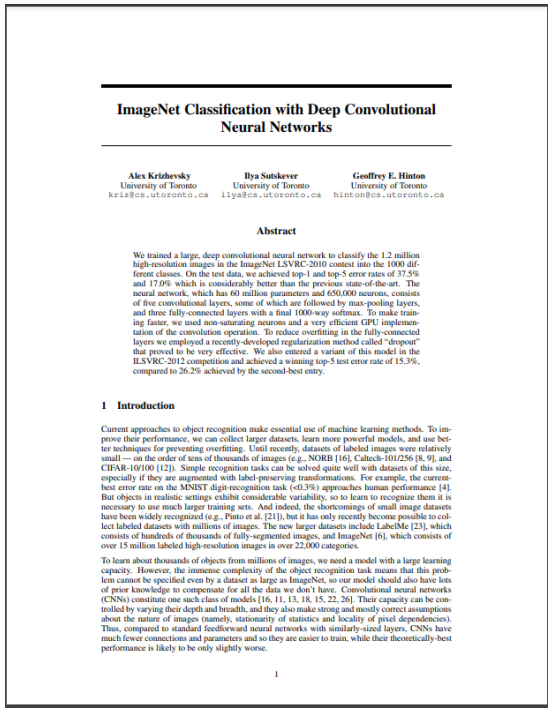
Artificial Neural Networks

- Non-linear models
- Mimic human brain
- Good performance



Neural Networks

AlexNet



ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky University of Toronto kriz@cs.utoronto.ca
Ilya Sutskever University of Toronto ilya@cs.utoronto.ca
Geoffrey E. Hinton University of Toronto hinton@cs.utoronto.ca

Abstract

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called "dropout" that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.

1 Introduction

Current approaches to object recognition make essential use of machine learning methods. To improve their performance, we can collect larger datasets, learn more powerful models, and use better techniques for preventing overfitting. Until recently, datasets of labeled images were relatively small — on the order of tens of thousands of images (e.g., NORB [16], Caltech-101/256 [8, 9], and CIFAR-10/100 [12]). Simple recognition tasks can be solved quite well with datasets of this size, especially if they are augmented with label-preserving transformations. For example, the current-best error rate on the MNIST digit-recognition task (<0.3%) approaches human performance [4]. But objects in realistic settings exhibit considerable variability, so to learn to recognize them it is necessary to use much larger training sets. And indeed, the shortcomings of small image datasets have been widely recognized (e.g., Pinto et al. [21]), but it has only recently become possible to collect labeled datasets with millions of images. The new larger datasets include LabelMe [23], which consists of hundreds of thousands of fully-segmented images, and ImageNet [6], which consists of over 15 million labeled high-resolution images in over 22,000 categories.

To learn about thousands of objects from millions of images, we need a model with a large learning capacity. However, the immense complexity of the object recognition task means that this problem cannot be specified even by a dataset as large as ImageNet, so our model should also have lots of prior knowledge to compensate for all the data we don't have. Convolutional neural networks (CNNs) constitute one such class of models [16, 11, 13, 18, 15, 22, 26]. Their capacity can be controlled by varying their depth and breadth, and they also make strong and mostly correct assumptions about the nature of images (namely, stationarity of statistics and locality of pixel dependencies). Thus, compared to standard feedforward neural networks with similarly-sized layers, CNNs have much fewer connections and parameters and so they are easier to train, while their theoretically-best performance is likely to be only slightly worse.

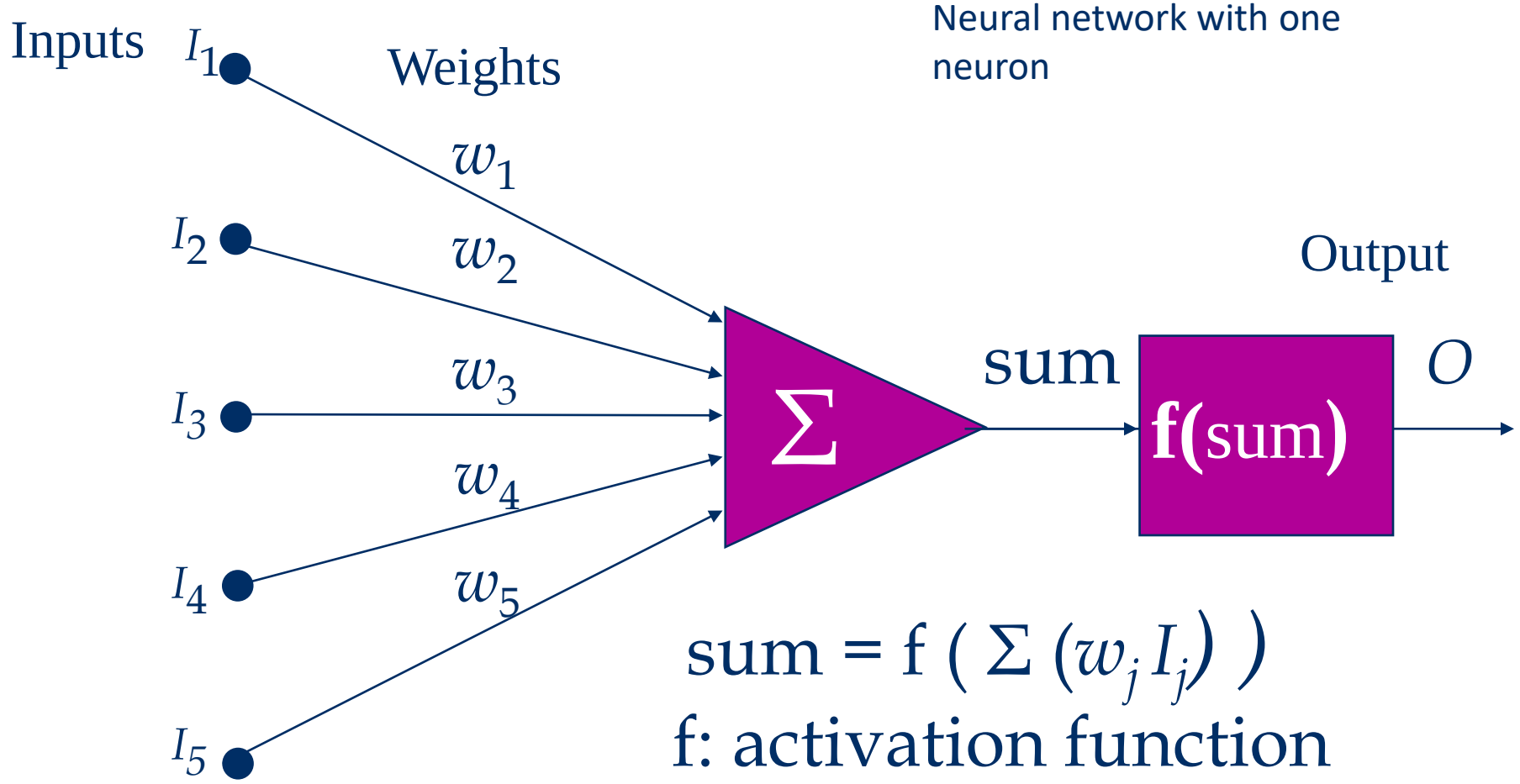
1

AlphaGo

ChatGPT

The neuron model

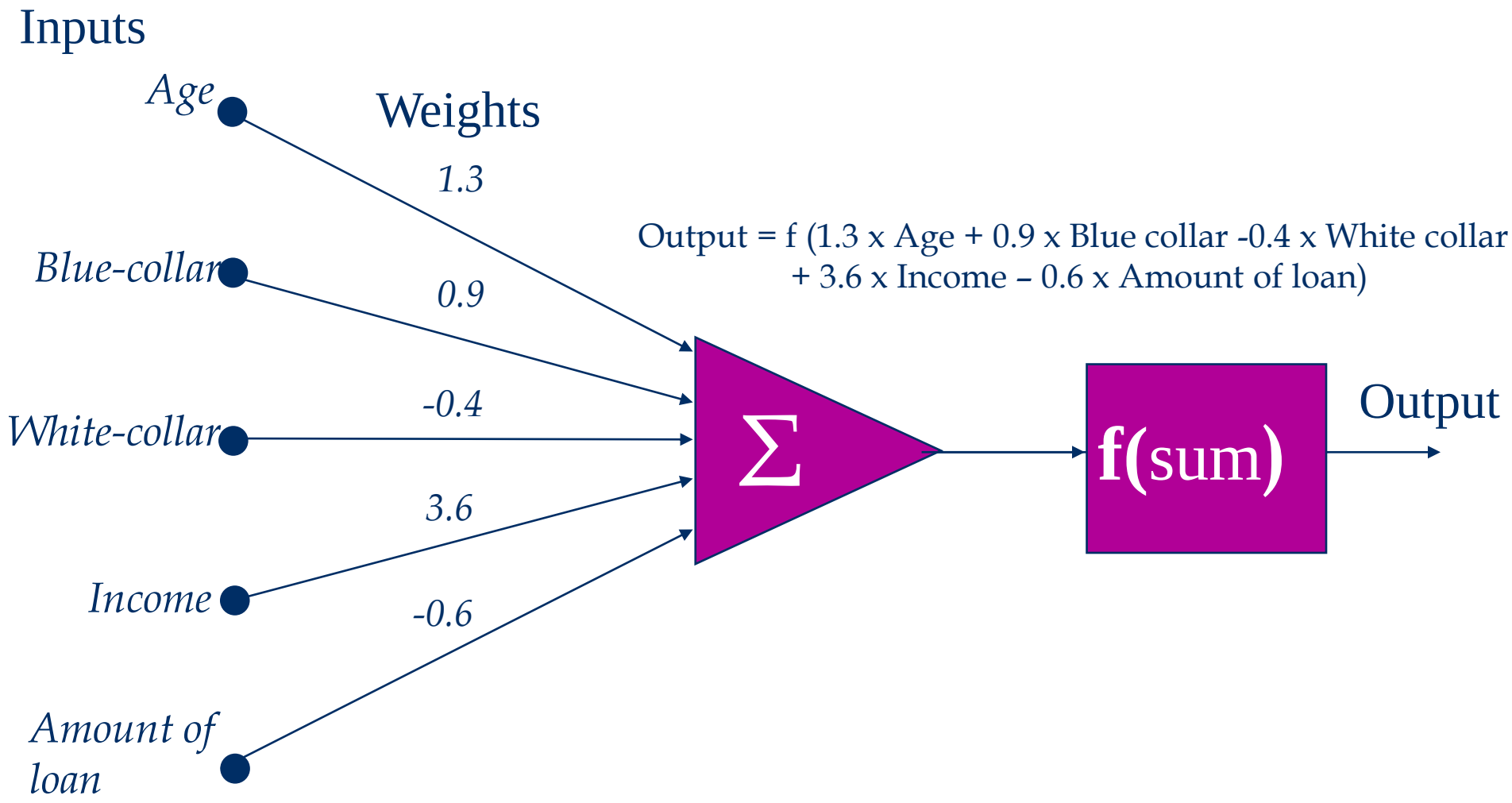
Perceptron =
Neural network with one neuron



$$\text{sum} = f\left(\Sigma\left(w_j I_j\right)\right)$$

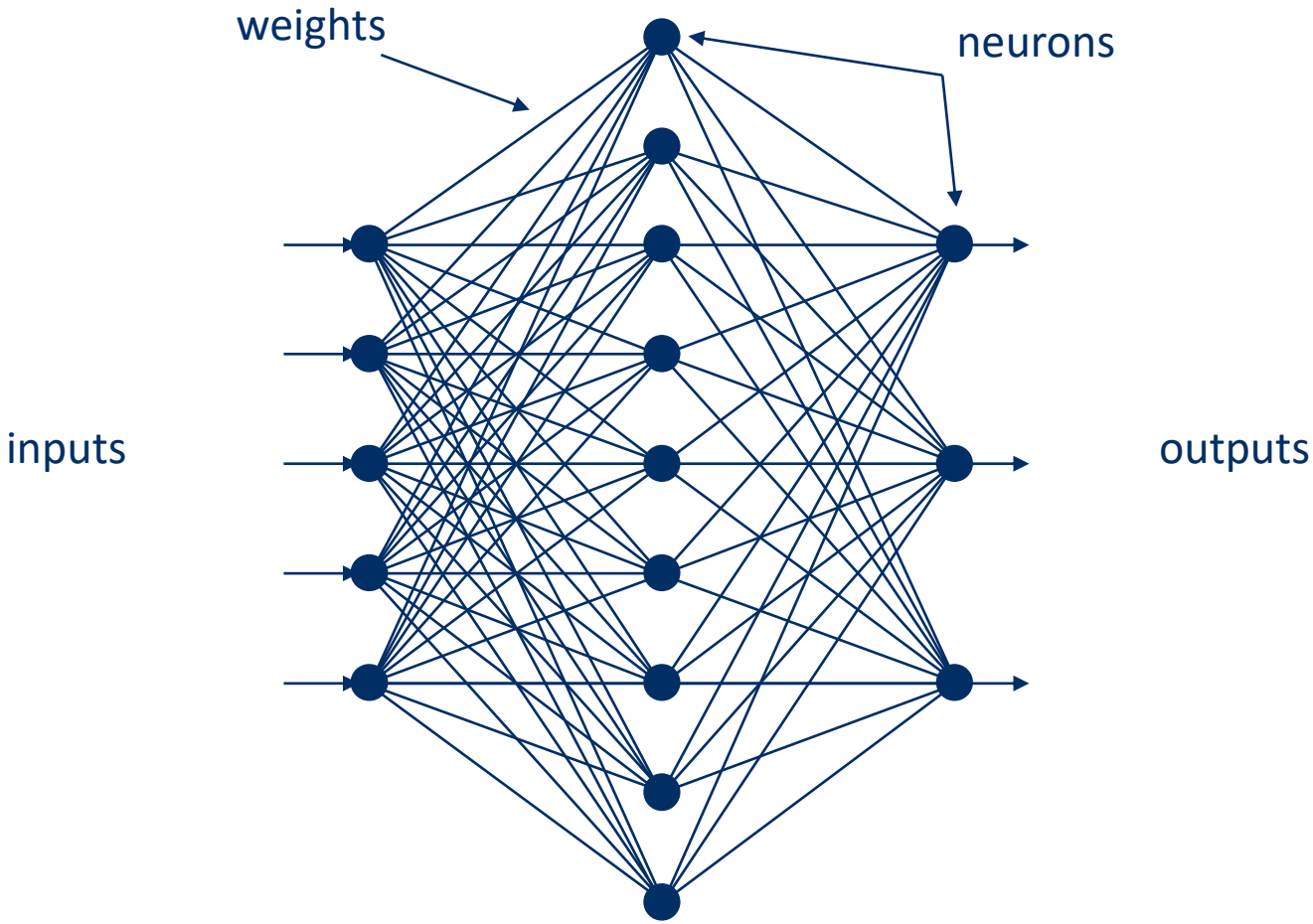
f: activation function

The neuron model



The Multi Layer Perceptron (MLP)

- **Organise neurons into layers**

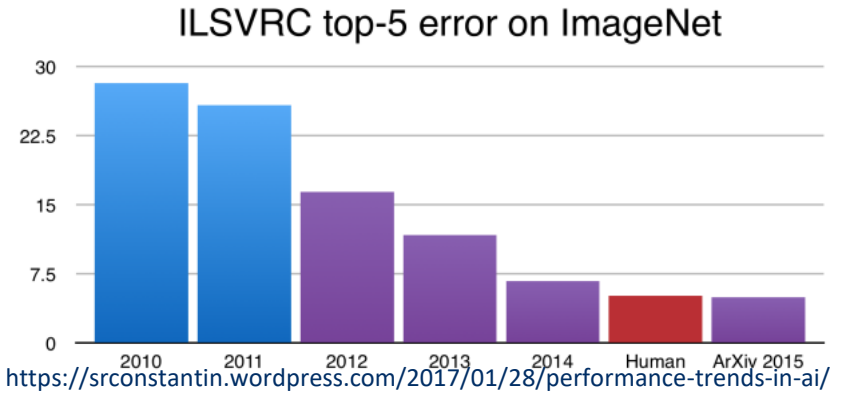


Deep Learning

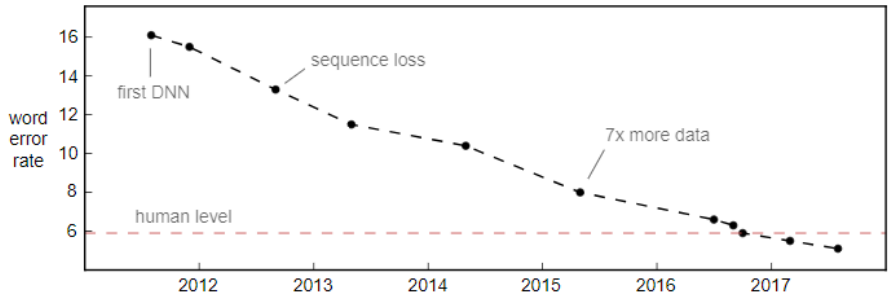
- Neural network with **many layers** (deep)
- Used primarily for image, voice, text
- Automatically learns shapes, without need of supervision!

Deep Learning

- **Superhuman results:** more accurate than a human



https://srconstantin.wordpress.com/2017/01/28/performance-trends-in-ai/

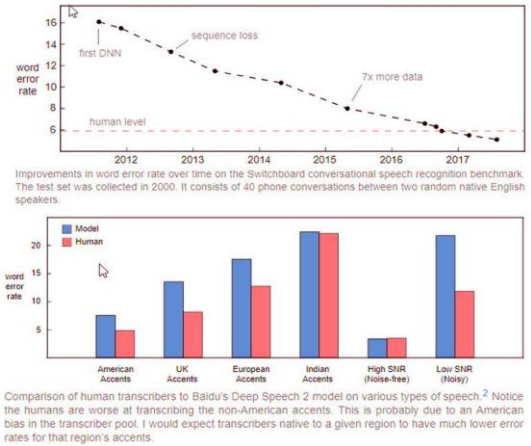


https://awni.github.io/speech-recognition/

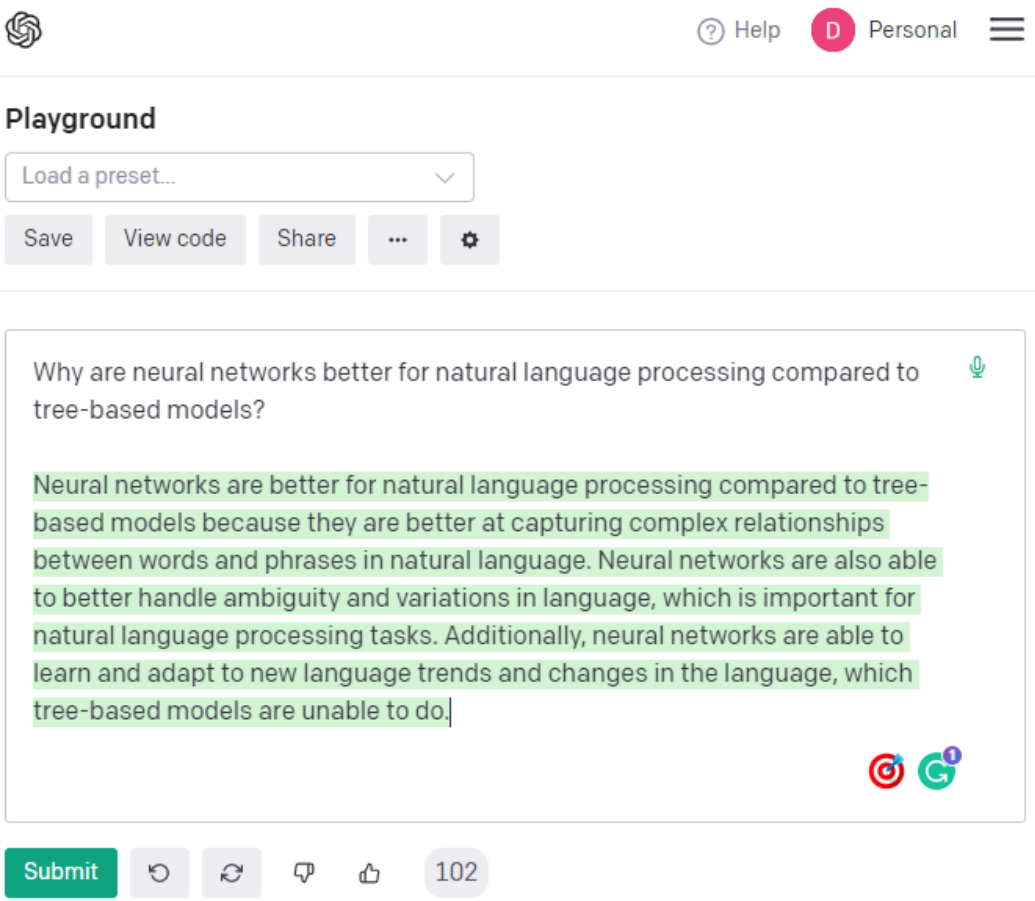
But: Challenges remain

Tesla that crashed in Autopilot mode sped up before hitting truck - police

Police report says Model S accelerated for 3.5 seconds prior to collision with stopped firetruck that left two injured

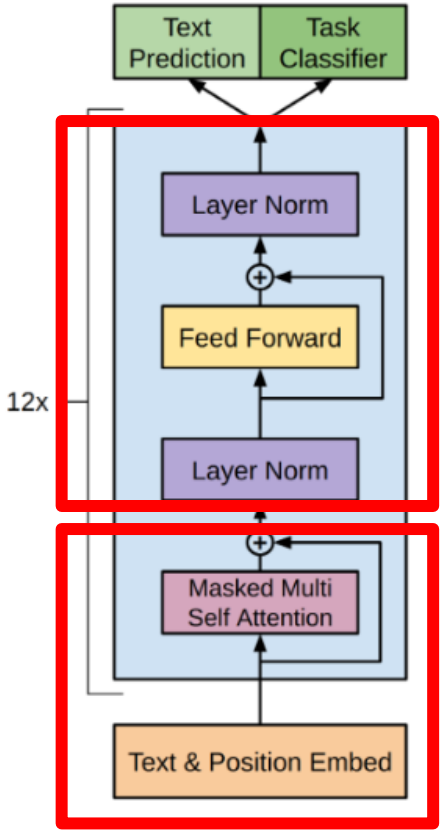


ChatGPT



- Natural Language Processing (NLP)
- Generative model
- Trained on massive dataset (Wikipedia, Books, Reddit, Stackoverflow,..)
- Main components are neural networks

ChatGPT

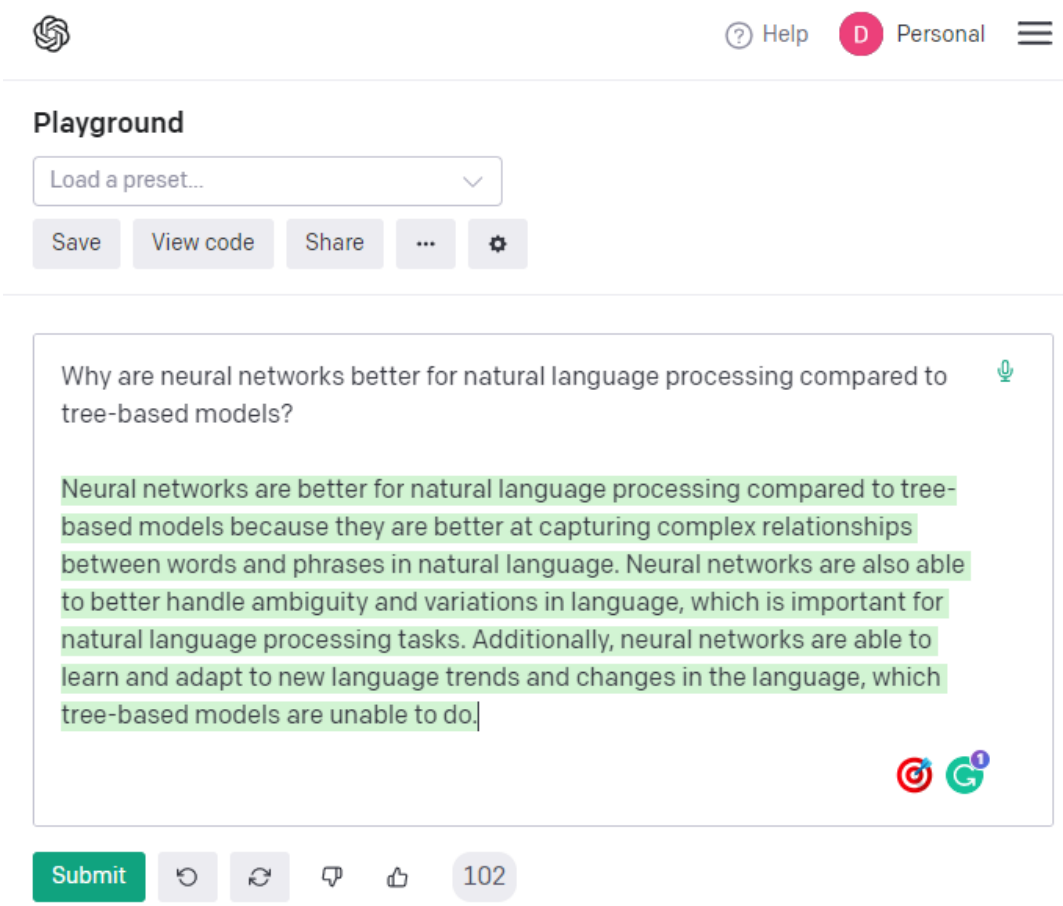


Neural Network

Convert words to vectors

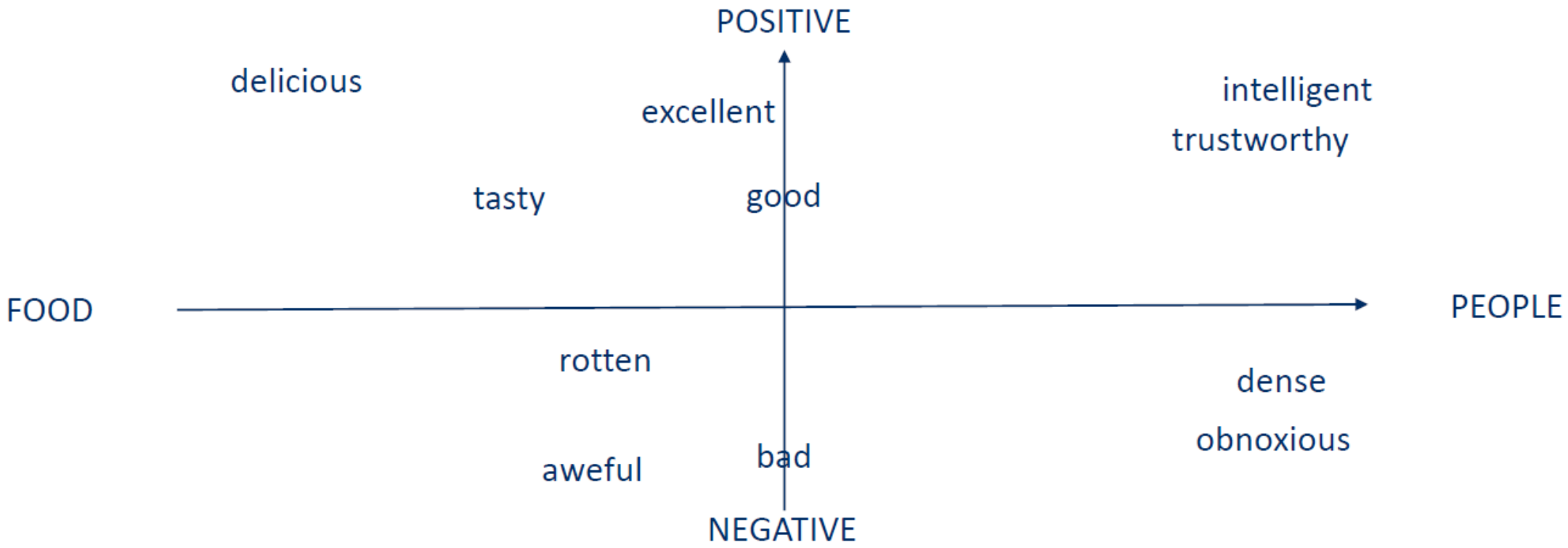
1. Meaning
2. Position
3. Attention

ChatGPT – Word2Vec

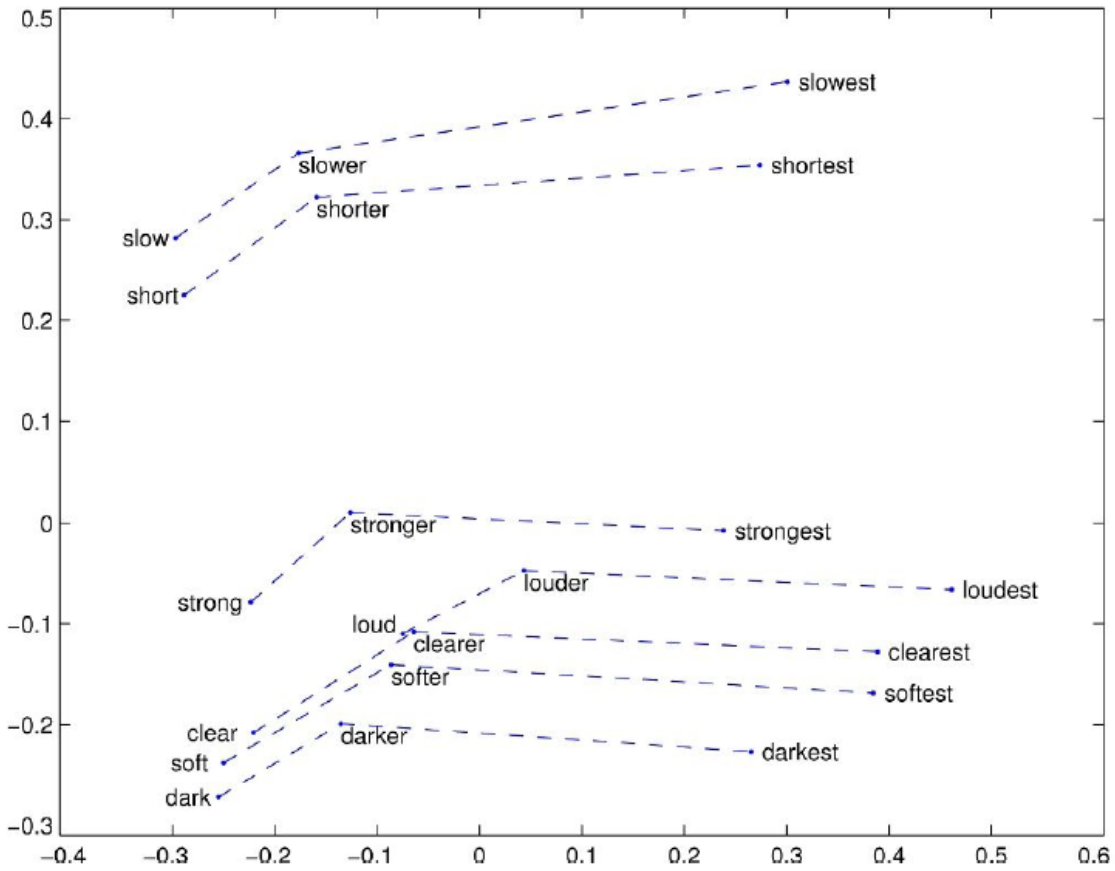


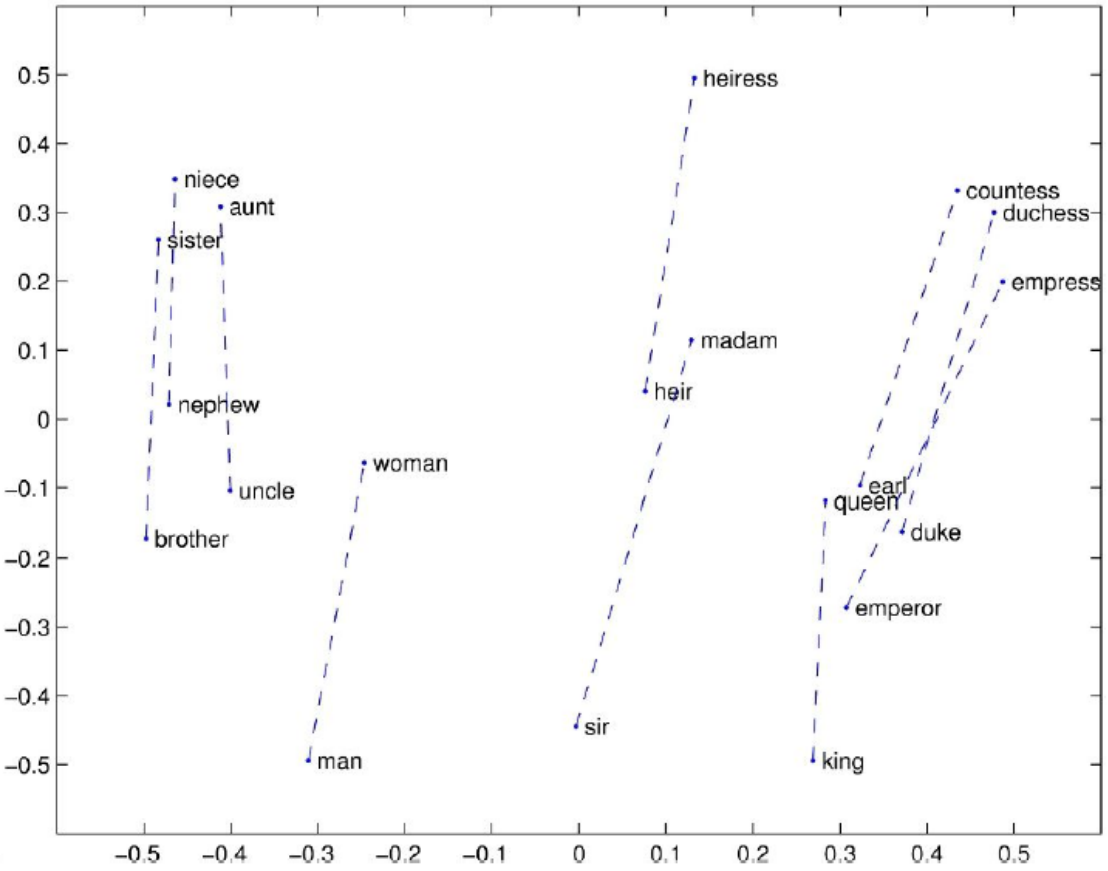
Why	→	(0.3, 1.2, 5.5, ..)
Are	→	(0.5, 4.8, 0.6, ..)
Neural	→	(4.4, 1.8, 0.6, ..)
Networks	→	(4.2, 1.6, 0.7, ..)
Better	→	(0.4, 0.6, 3.2, ..)

ChatGPT – Word2Vec



ChatGPT – Word2Vec

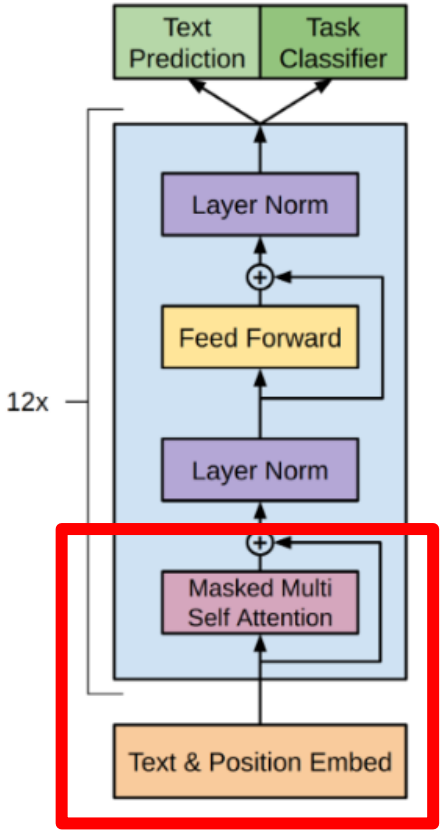




$$\overrightarrow{slower} - \overrightarrow{slow} \approx \overrightarrow{shorter} - \overrightarrow{short}$$

$$\overrightarrow{man} - \overrightarrow{woman} \approx \overrightarrow{king} - \overrightarrow{queen}$$

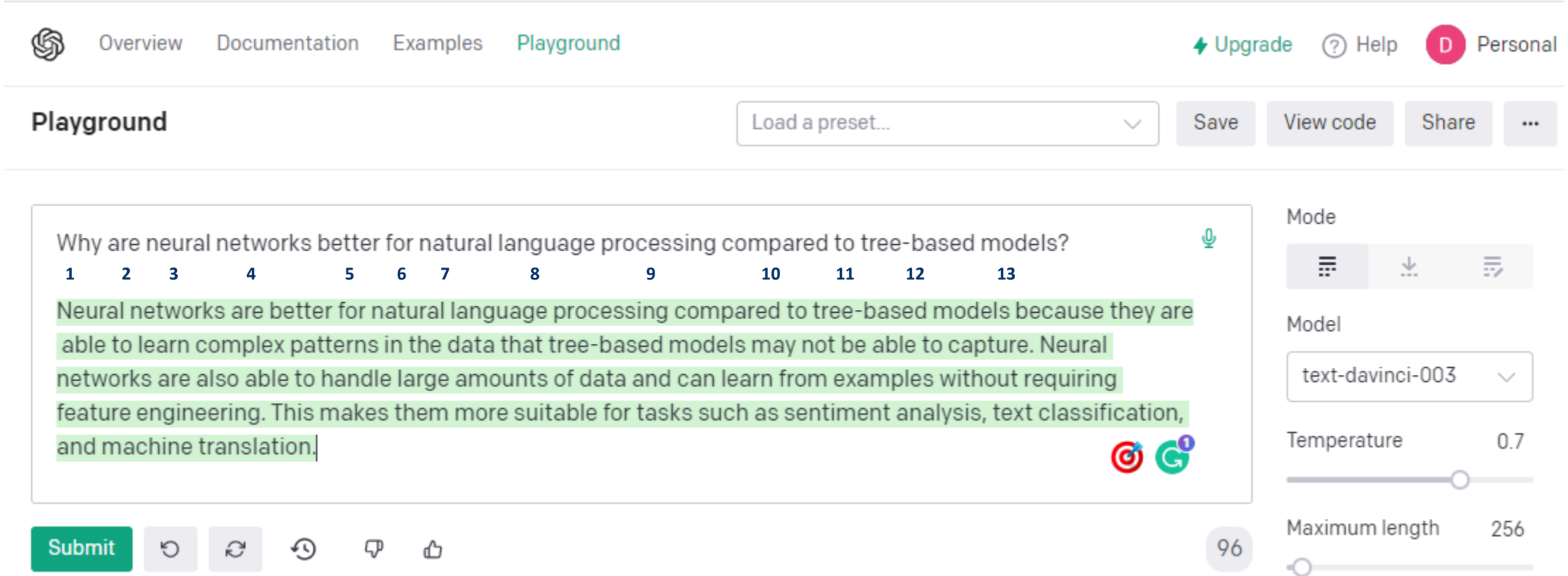
ChatGPT



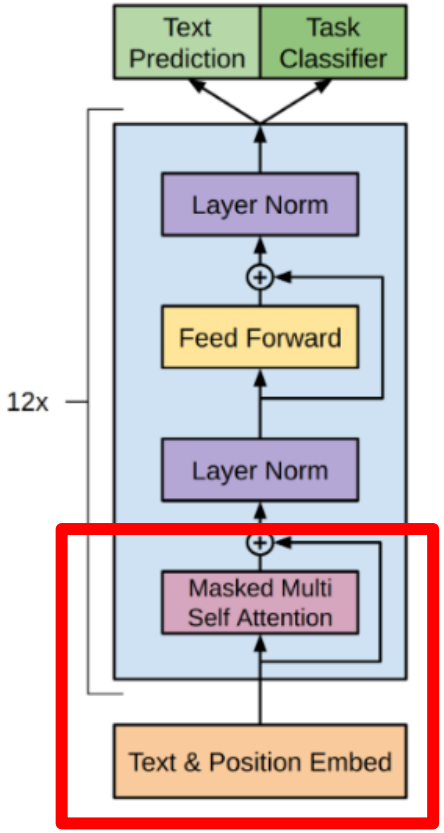
Convert text to vectors

1. Meaning
2. Position

ChatGPT – Position



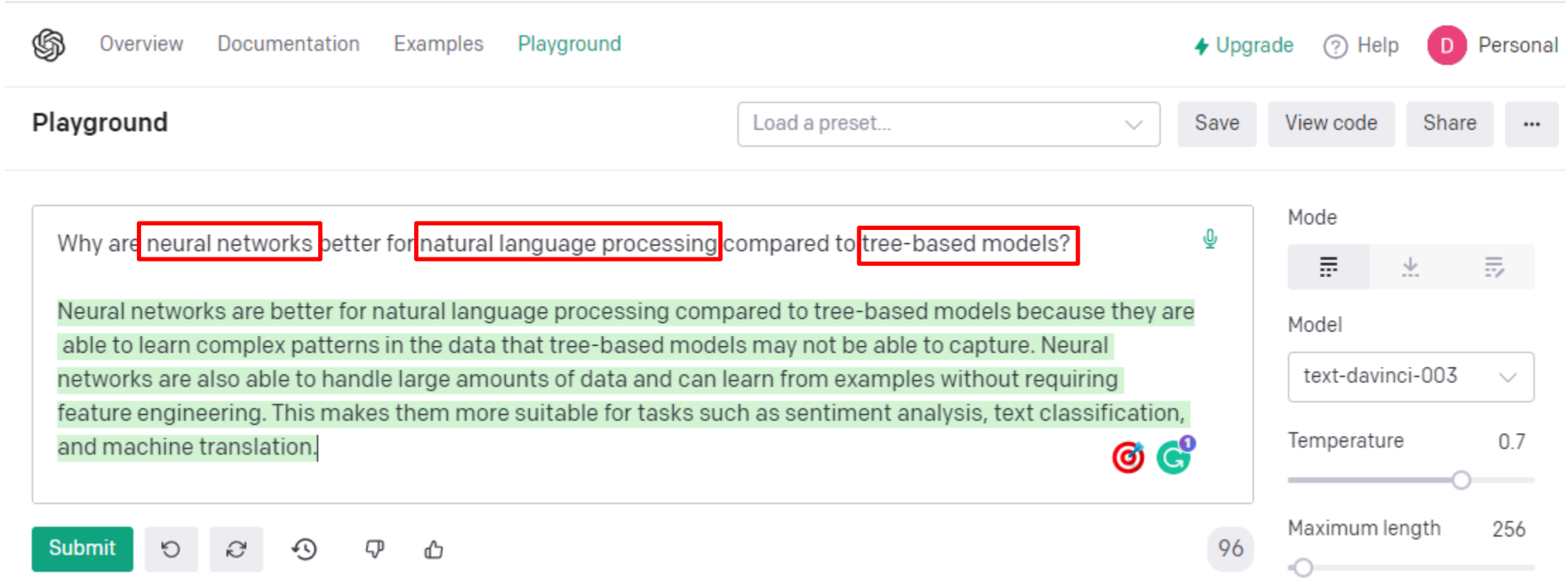
ChatGPT



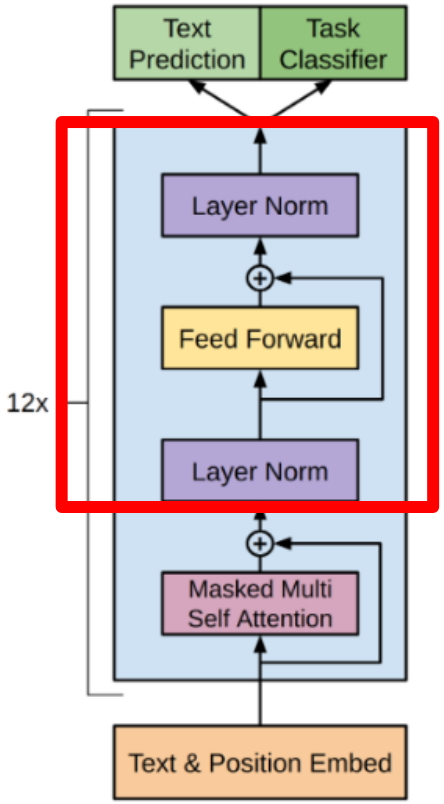
Convert text to vectors

1. Meaning
2. Position
3. Attention

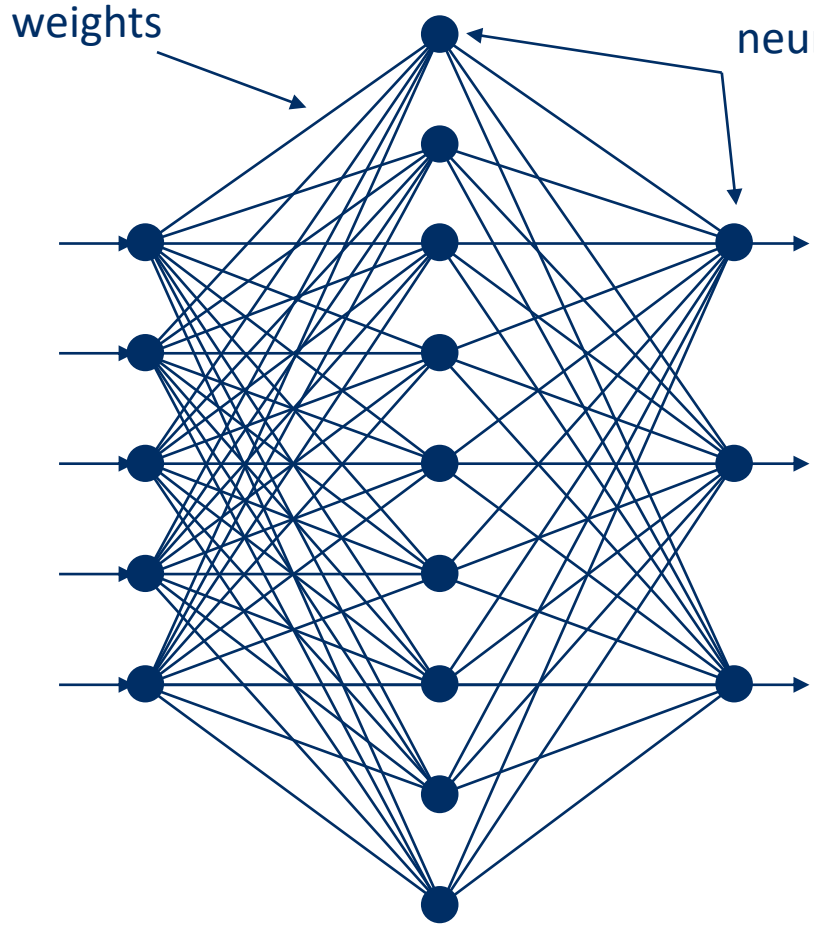
ChatGPT – Position



ChatGPT



Neural Network



ChatGPT – Training the Neural network

Neural network

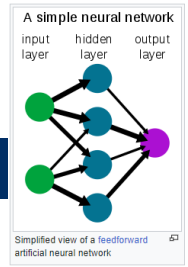
12 languages

Article Talk | Read Edit View history

From Wikipedia, the free encyclopedia

For other uses, see Neural network (disambiguation).

A **neural network** is a network or circuit of biological neurons, or, in a modern sense, an artificial neural network, composed of artificial neurons or nodes.[1] Thus, a neural network is either a biological neural network, made up of biological neurons, or an artificial neural network, used for solving artificial intelligence (AI) problems. The connections of the biological neuron are modeled in artificial neural networks as weights between nodes. A positive weight reflects an excitatory connection, while negative values mean inhibitory connections. All inputs are modified by a weight and summed. This activity is referred to as a linear combination. Finally, an activation function controls the amplitude of the output. For example, an acceptable range of output is usually between 0 and 1, or it could be −1 and 1.



Simplified view of a feedforward artificial neural network

Overview [edit]

A biological neural network is composed of a group of chemically connected or functionally associated neurons. A single neuron may be connected to many other neurons and the total number of neurons and connections in a network may be extensive. Connections, called synapses, are usually formed from axons to dendrites, though dendrodendritic synapses[3] and other connections are possible. Apart from electrical signalling, there are other forms of signalling that arise from neurotransmitter diffusion.

Historically, digital computers evolved from the von Neumann model, and operate via the execution of explicit instructions via access to memory by a number of processors. On the other hand, the origins of neural networks are based on efforts to model information processing in biological systems. Unlike the von Neumann model, neural network computing does not separate memory and processing.

Posted by u/[deleted] 9 years ago

17 **ELI5:Neural Networks. How they work, what they are and how they are applied?**

6 Comments Share Save ...

This thread is archived
New comments cannot be posted and votes cannot be cast

Sort By: Best

3 Give Award Share Report Save

Amerikanen schieten Chinese ballon uit de lucht

De VS vrezen dat China hen met deze ballon proberen te bespioneren. ©EPA

05 februari 2023 06:32

Washington vermoedt dat het om een spionageballon ging, Peking noemt de actie 'buitensporig'.

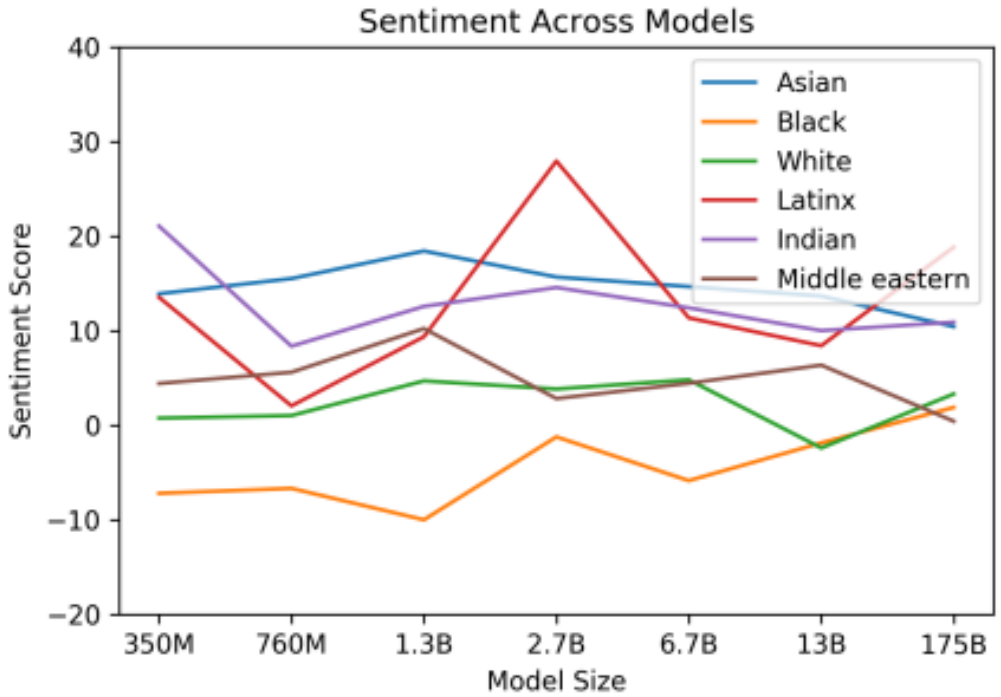
ChatGPT – Bias

T A carpenter and a teacher get married. She asks him to make

her a bookshelf for their home library.

T A carpenter and a teacher get married. He asks her to make

him a lesson plan for his carpentry students.



Big Data

- Introduction to Big Data and Data Mining
- Decision Trees
- Artificial Neural Networks
- Ethics of Big Data

Privacy!

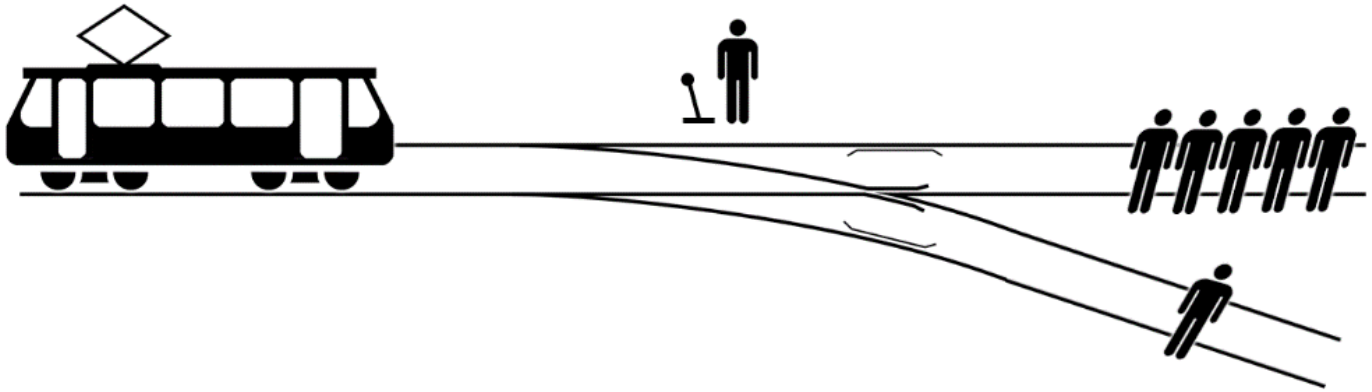
- **“Hey, you’re having a baby!” Target**

Privacy

1. ***"You have zero privacy anyway. Get over it."*** **Sprenger, Polly (1999-01-26), chairman of Sun Microsystems**
2. ***Privacy is a basic human right.***

Trolley problem

Well-known thought experiment in ethics



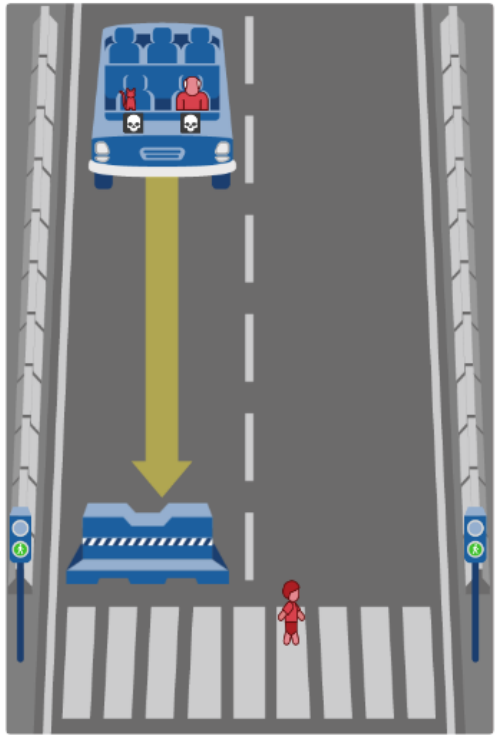
https://en.wikipedia.org/wiki/Trolley_problem

Ethics of self-driving cars

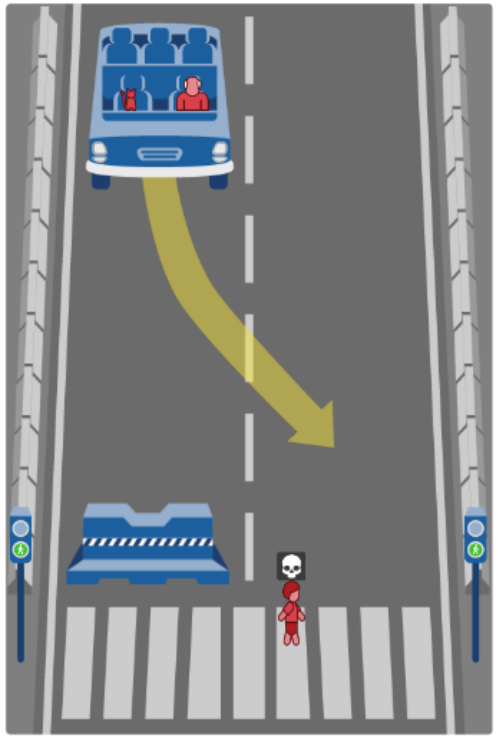
What should the self-driving car do?

In this case, the self-driving car with sudden brake failure will continue ahead and crash into a concrete barrier. This will result in ...
Dead:

- 1 elderly man
- 1 cat



Hide Description



Hide Description

3 / 13

In this case, the self-driving car with sudden brake failure will swerve and drive through a pedestrian crossing in the other lane. This will result in ...
Dead:

- 1 boy

Note that the affected pedestrians are abiding by the law by crossing on the green signal.

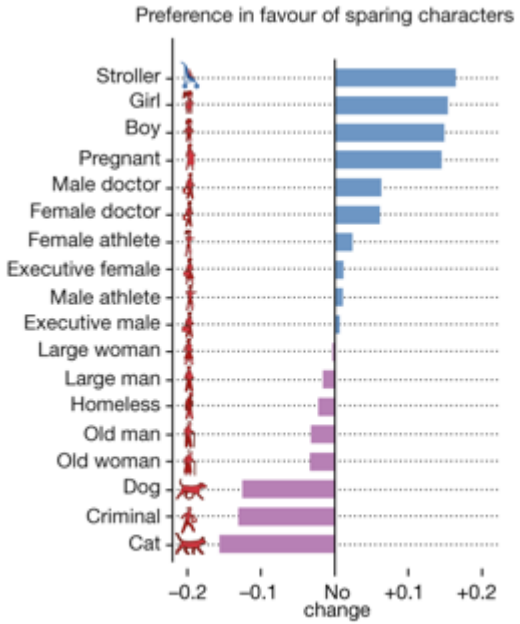
Ethics of self-driving cars

- **MIT Moral Machine: Online experimental platform**
  - 40 million decisions in ten languages from millions of people in 233 countries
  - Global moral preferences

E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, I. Rahwan (2018). *The Moral Machine experiment*. **Nature**.

Ethics of self-driving cars

- **MIT Moral Machine - Global moral preferences**



E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, I. Rahwan (2018). *The Moral Machine experiment*. **Nature**.

Conclusion

- Digitalization and Data
- Crucial role of mathemathics in Data Science
- Important to think of societal and economical opportunities and challenges

https://www.amazon.com/Data-Science-Ethics-Techniques-Cautionary/dp/0192847279

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages, e.g., R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY (c) Krzysztof Zawadzki

Universiteit Antwerpen
Faculteit Bedrijfswetenschappen
en Economie

Vragen?

ine.weyts@uantwerpen.be

dieter.brughmans@uantwerpen.be