Machine Learning

Pak Alexander Alexandrovich

Candidate of tech. sciences

a.pak@kbtu.kz

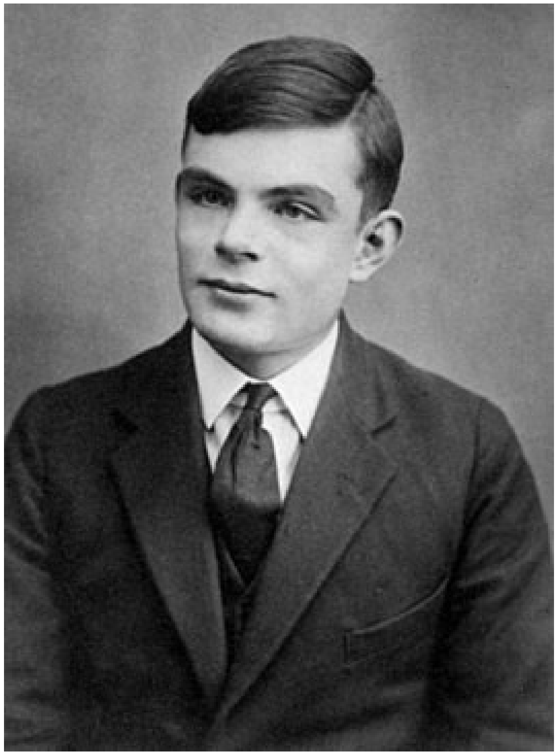
“I keep saying that the sexy job in the next 10 years will be statisticians”
Hal Varian, Chief Economist Google

Lecture 1: Introduction

History of Artificial Intellegence

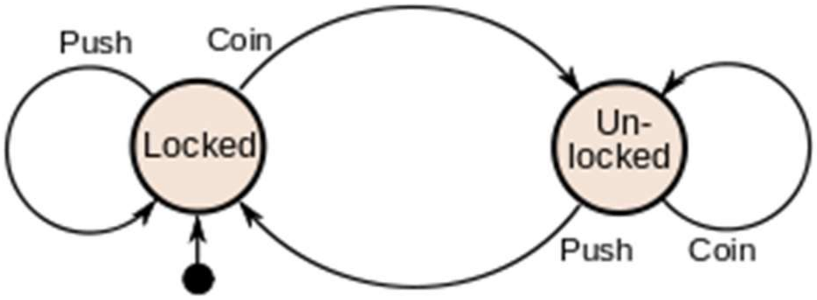
- 1936-1956 The birth of theory
- 1956-1976 Golden Age
- 1969-1980 Crisis of ANN

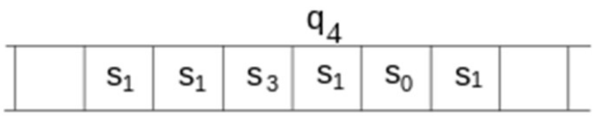
The birth of AI



On computable numbers, with an app to the Entscheidungsproblem (1936)
Computing machinery and intelligence (1950)

Finite state machine or A-machine of A.Turing





The head is always over a particular square of the tape;

Finite state machine or A-machine of A.Turing

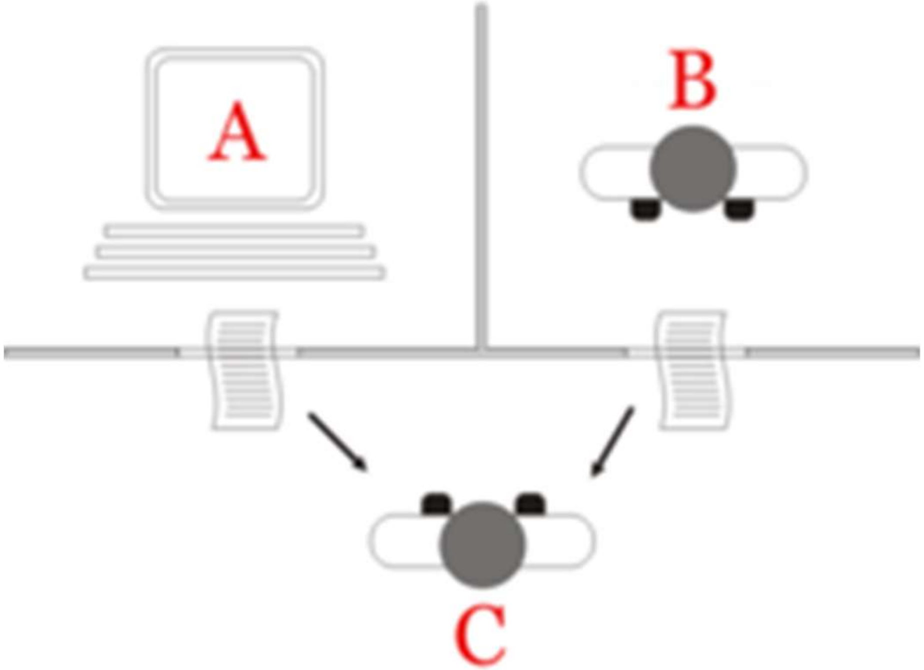
The behaviour of the computer at any moment is determined by the symbols which **he** is observing and **his** ‘state of mind’ at that momen,

Alan Turing

Example of calculation with the help of Turing Machine

23 56 8	23 56 138	23 56 138 115 1288

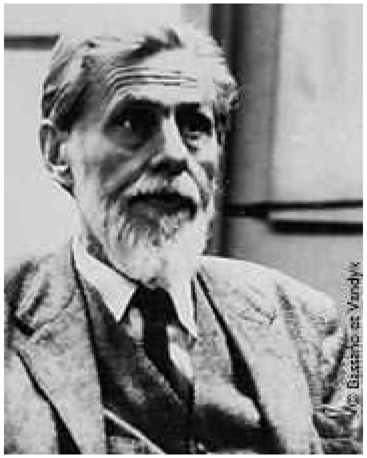
Turing Test



The Death of the Father of AI sci

Turing was prosecuted in 1952 for homosexual acts, when by the Labouchere Amendment, "gross indecency" was criminal in the UK. He accepted chemical castration treatment, with DES, as an alternative to prison. Turing died in 1954, 16 days before his 42nd birthday, from cyanide poisoning. An inquest determined his death as suicide, but it has been noted that the known evidence is also consistent with accidental poisoning.

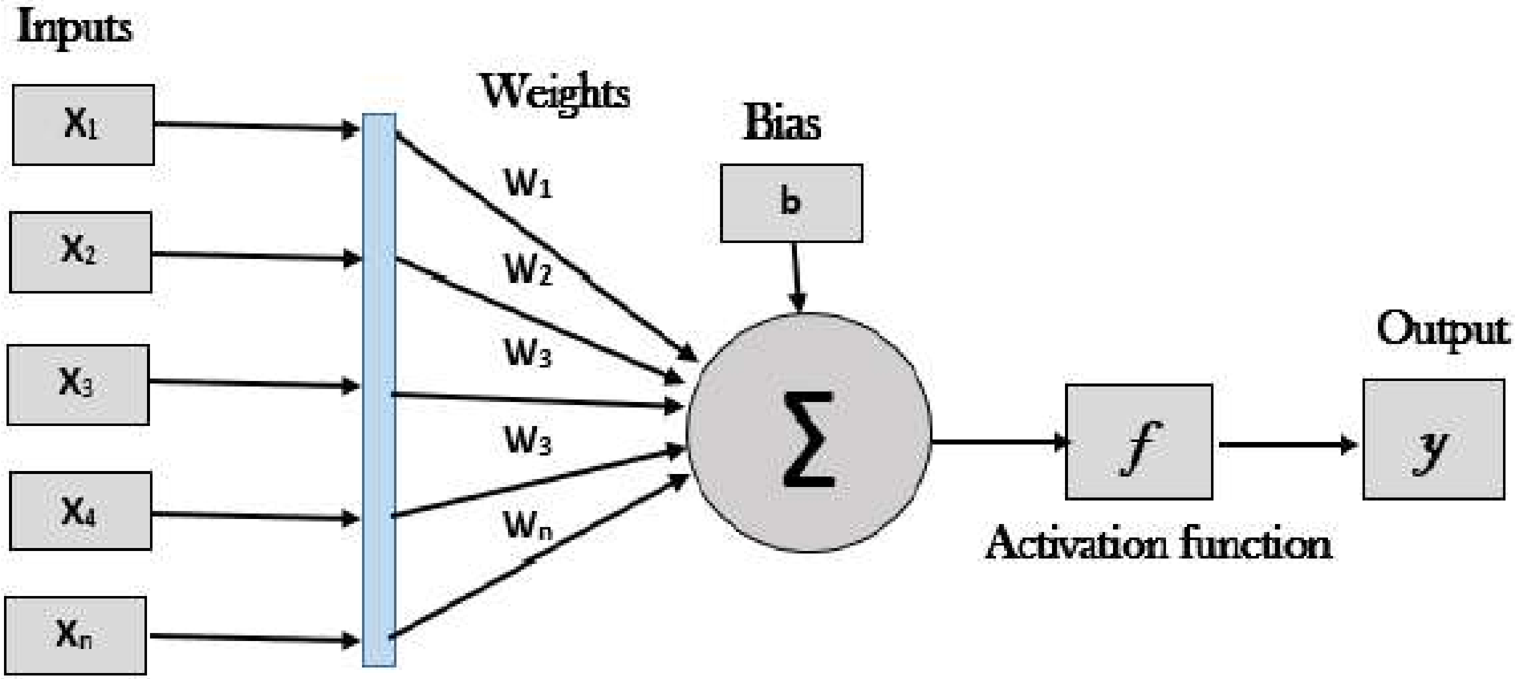
A LOGICAL CALCULUS OF THE IDEAS IMMANENT IN NERVOUS ACTIVITY* (1943)



Warren McCulloch

Walter Pitts

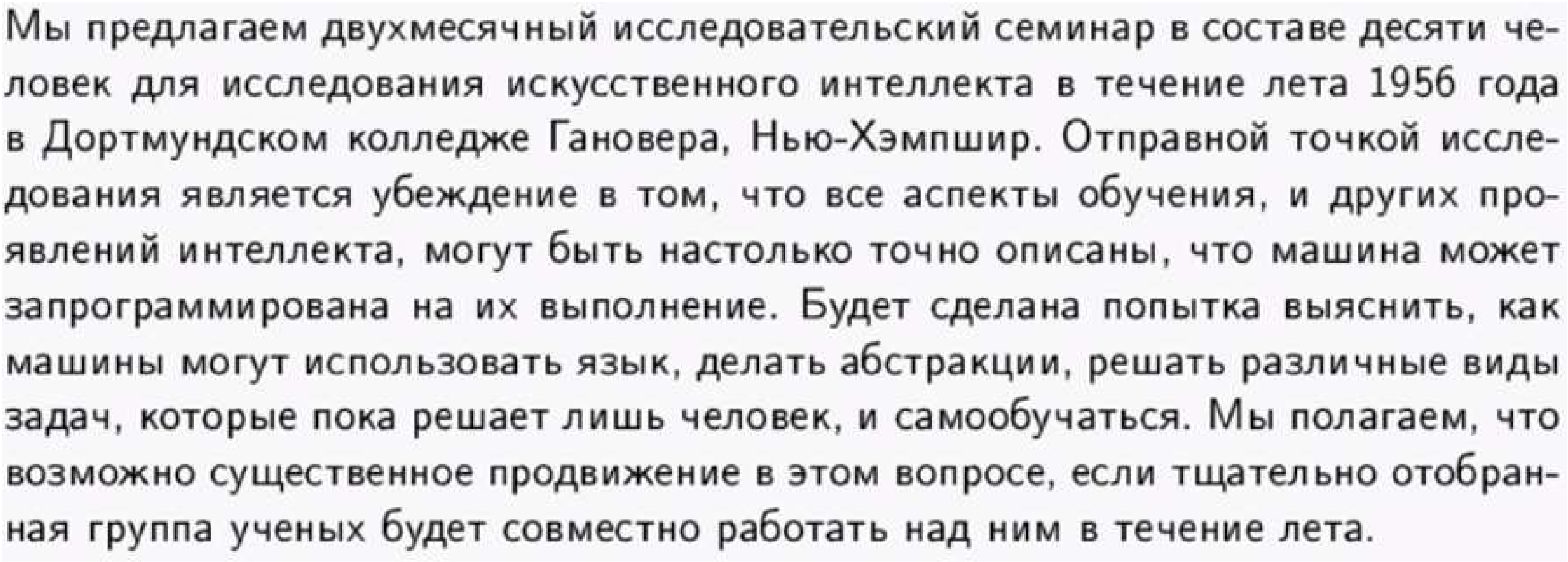
Formal Neuron of McCulloch and Pitts



Evolutional calculus

Nils Aall Barricelli
Genetic Algo (1954)

McCarthy et al. Proposal for the project



Мы предлагаем двухмесячный исследовательский семинар в составе десяти человек для исследования искусственного интеллекта в течение лета 1956 года в Дортмундском колледже Гановера, Нью-Хэмпшир. Отправной точкой исследования является убеждение в том, что все аспекты обучения, и других проявлений интеллекта, могут быть настолько точно описаны, что машина может запрограммирована на их выполнение. Будет сделана попытка выяснить, как машины могут использовать язык, делать абстракции, решать различные виды задач, которые пока решает лишь человек, и самообучаться. Мы полагаем, что возможно существенное продвижение в этом вопросе, если тщательно отобранная группа ученых будет совместно работать над ним в течение лета.

Golden Age 1956-1976

John Alan Robinson

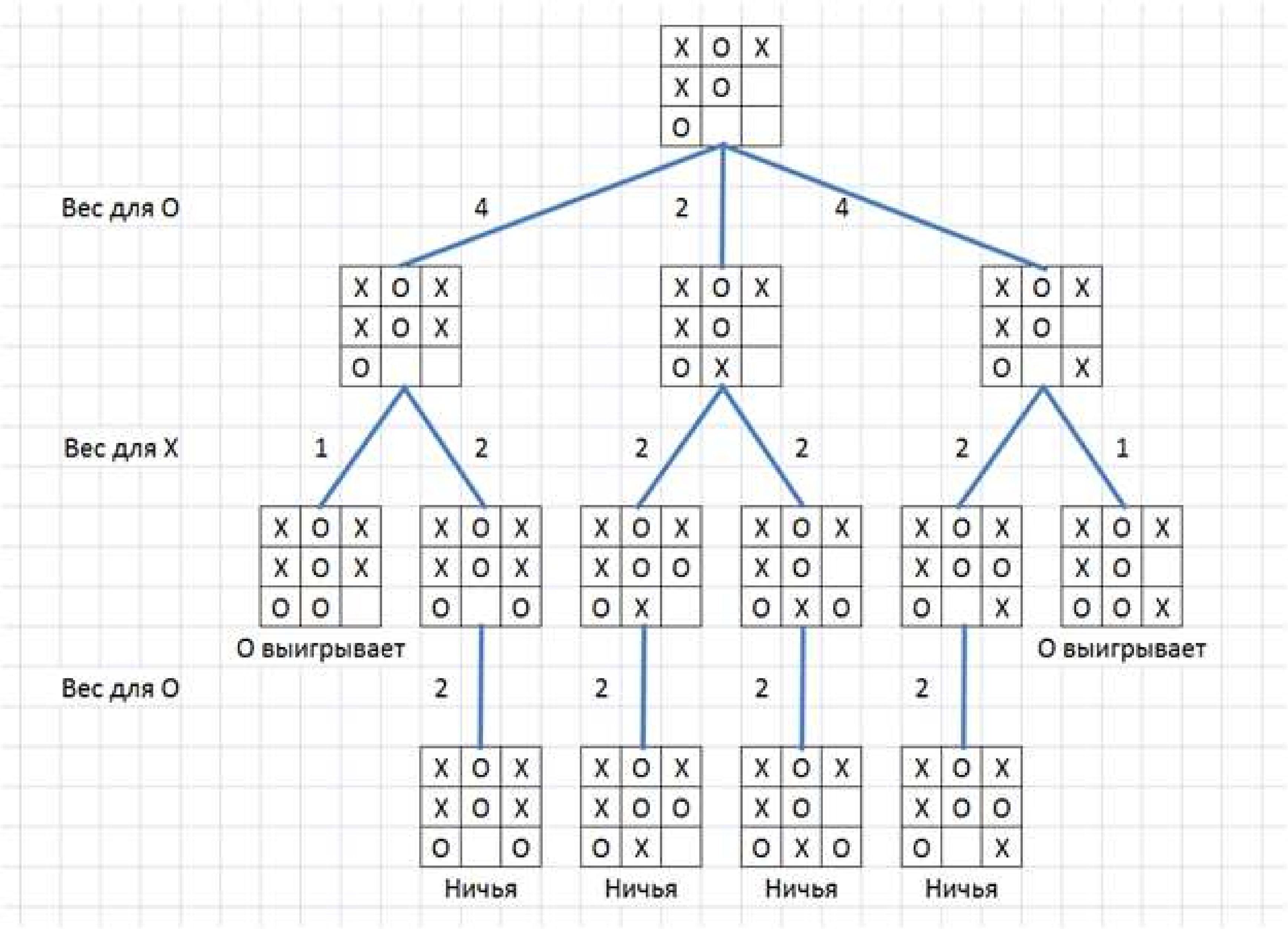
Prolog language (1972)

Alan Colmerauer

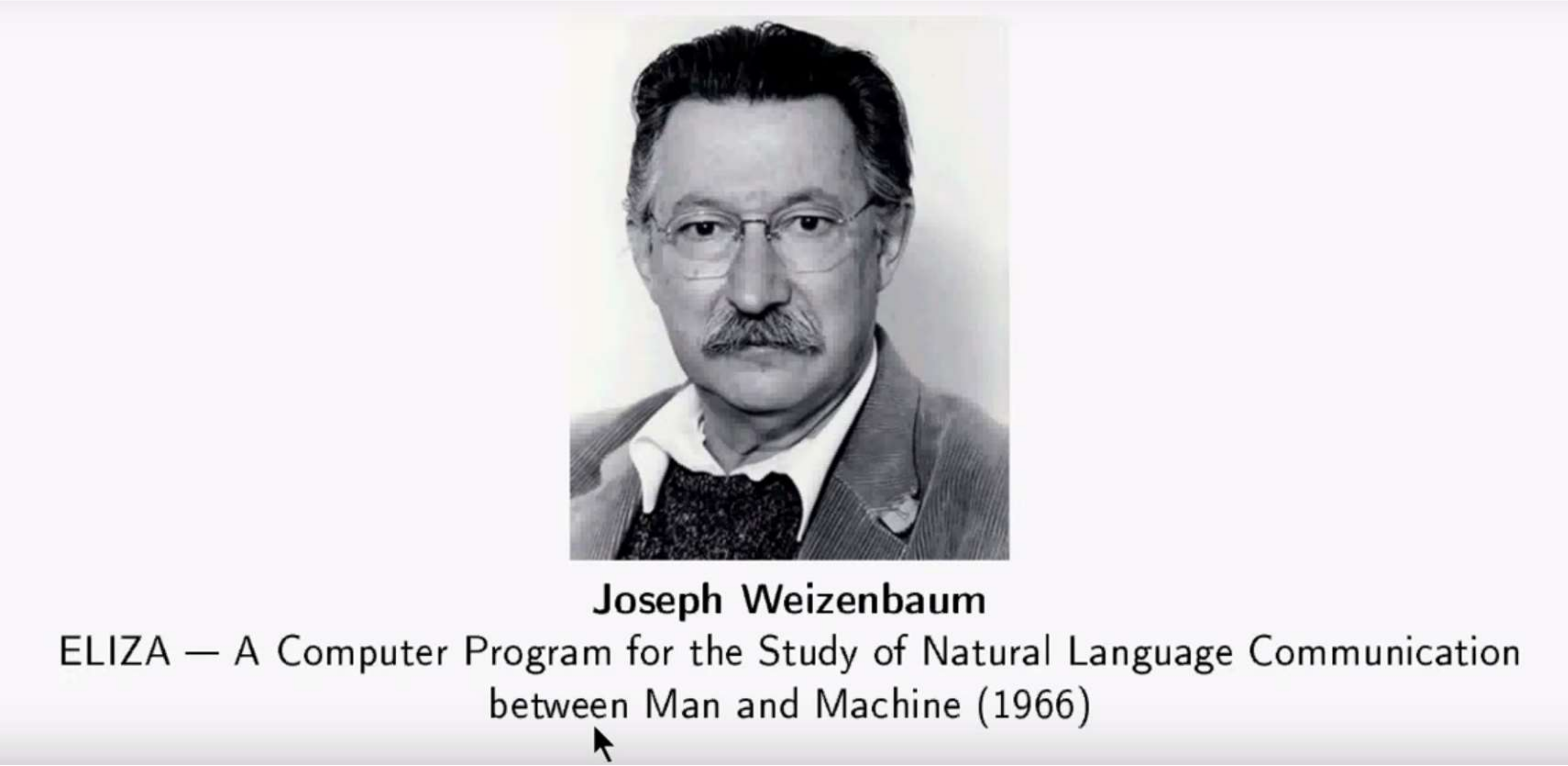
Philippe Roussel

Lotfi Zadeh

Fuzzy Logics (1965)


Вес для О
4
2
4
Вес для Х
1
2
2
2
2
1
О выигрывает
О выигрывает
Вес для О
2
2
2
2
Ничья
Ничья
Ничья
Ничья

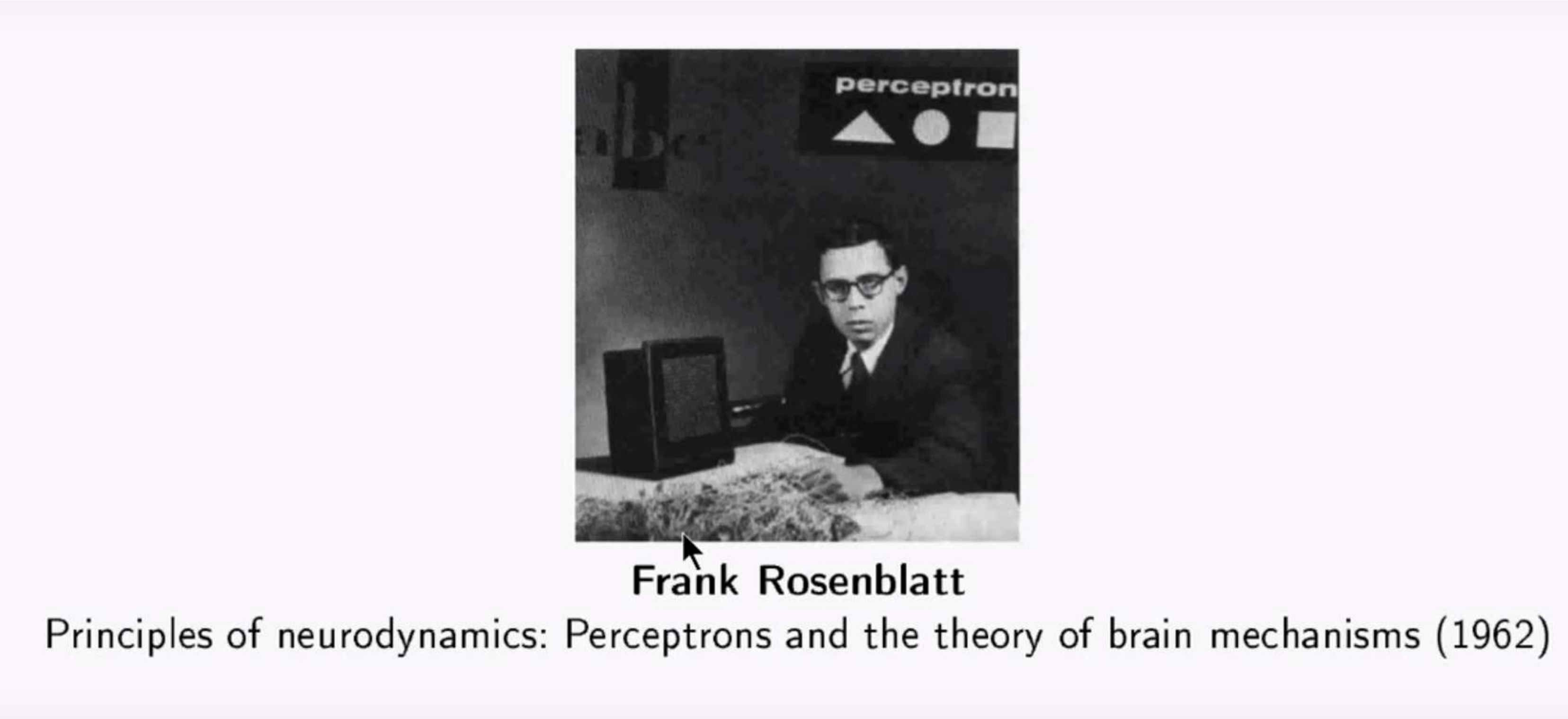
Natural Language Processing



Joseph Weizenbaum

ELIZA — A Computer Program for the Study of Natural Language Communication between Man and Machine (1966)

Neural Nets



Frank Rosenblatt

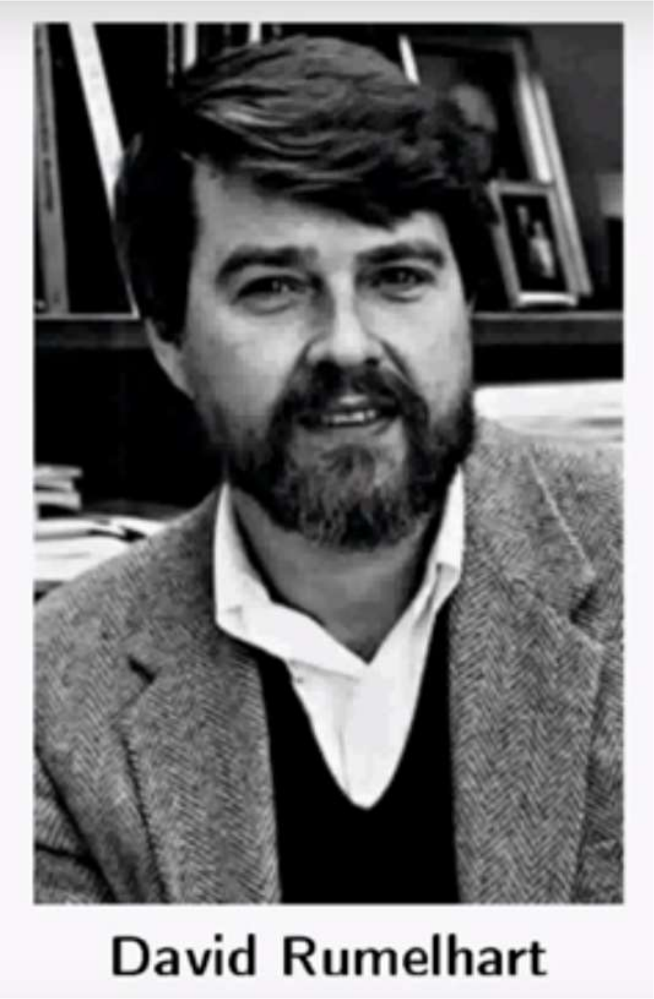
Principles of neurodynamics: Perceptrons and the theory of brain mechanisms (1962)

Evolutional calculus

John Henry Holland

Minsky Papert

Back Prop (1974)



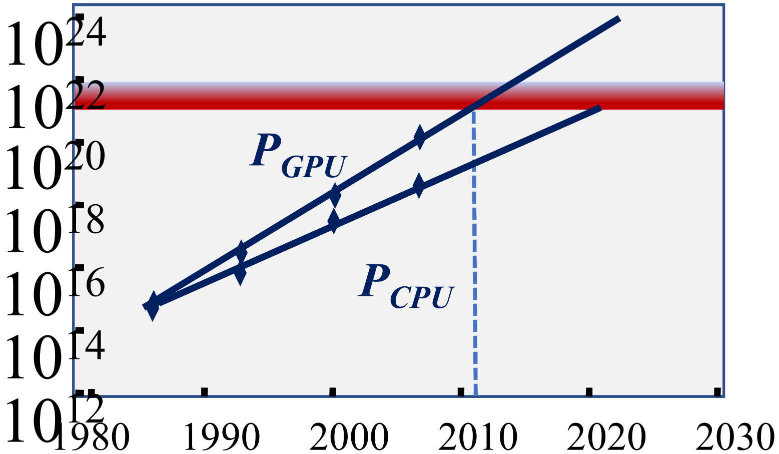
1986

Сенсорный ИИ человеческого уровня
на базе глубокого обучения

- Новая парадигма **ИИ на базе машинного обучения**
- Распознавание образов **на уровне человека**
- Закономерное **следствие закона Мура**

Science, 20

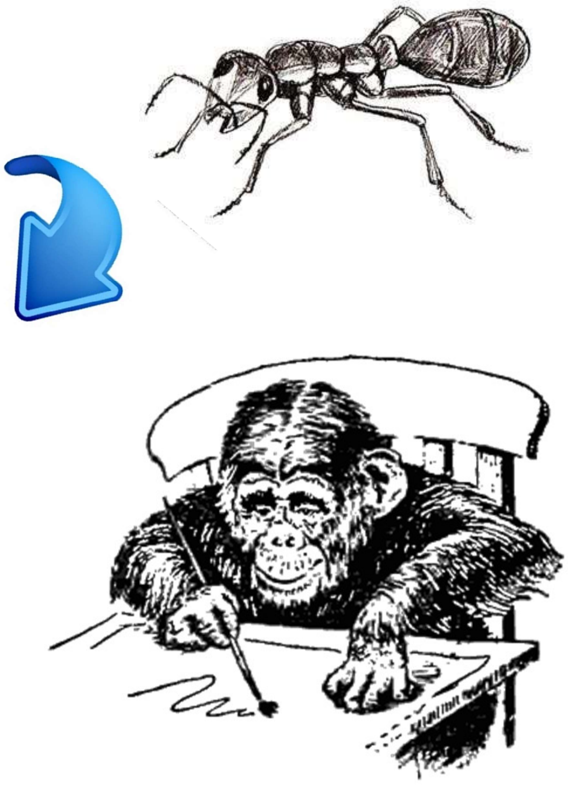
$P_{чел}$ ~ 10^{10} чел × 10^{10} байт × 10 Гц ~ 10^{21} Flops ➡



Революция в Искусственном Интеллекте

- “Good Old-Fashioned” Artificial Intelligence
  - Интеллект можно запрограммировать

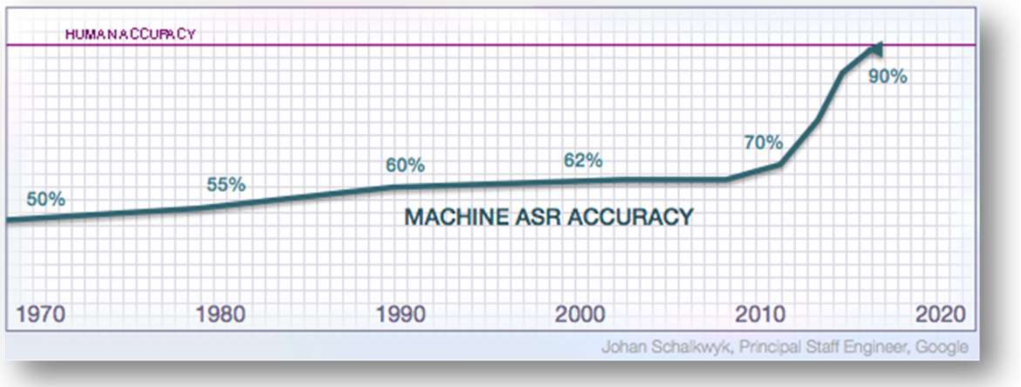
- Machine Learning Artificial Intelligence
  - Интеллект возникает в процессе обучения



Распознавание образов на уровне человека

- Машинное зрение
- Распознавание лиц
- Распознавание речи
- Понимание речи
- Машинный перевод …

Глубокие нейронные сети



Johan Schalkwyk, Principal Staff Engineer, Google

Сложность распознавания образов на уровне человека

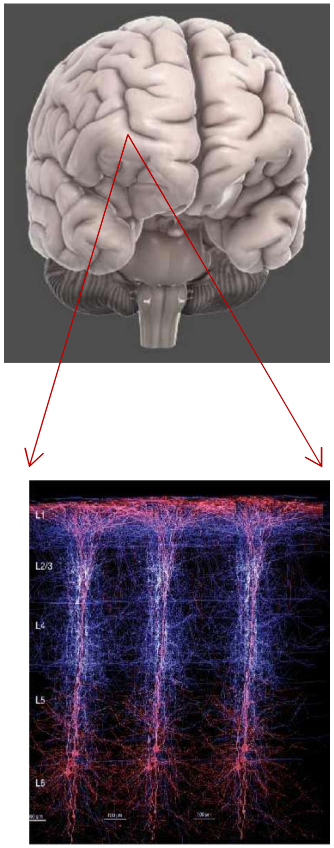
- Сознательное мышление (весь мозг) $W \lesssim 10^{10}$

  $C \sim mW^2 < 10^{21}$

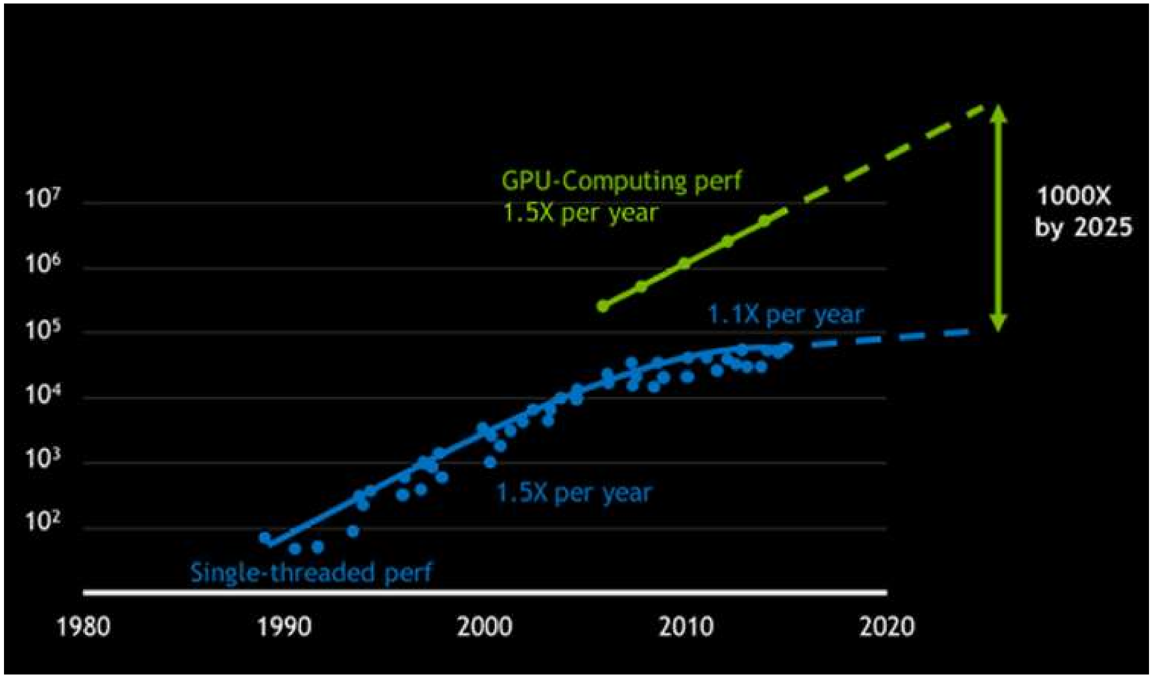
- Сенсорный интеллект (зрение, слух, …) $W \lesssim 10^{9}$

  $C \sim mW^2 < 10^{19}$

10^6 колонок
по 10^4 нейронов



Доступные мощности ×10 каждые 5 лет



$$t \sim \frac{10^{19}}{10^{13\div14}} < 10^{5\div6}\text{сек}$$

Сегодня: 3 дня

⇦ 10^{13}-10^{14} FLOPS

Конец 90-х: 100 лет

⇦ 10^{9}-10^{10} FLOPS

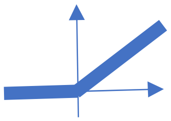
Новации глубокого обучения

Наследство 20-го века

- Базовые архитектуры
  - **CNN** (LeCun, 1989)
  - **LSTM** (Hochreiter, 1997)
- Алгоритмы обучения
  - **S**tochastic **G**radient **D**escent

Новации 21-го века

- Регуляризация обучения
  - **ReLU** (Nair, 2010)
  - **Dropout** (Hinton, 2012)
  - **Batch normalization** (Ioffe, 2015)



- Нейросетевая схемотехника
  - Прикладные задачи **узкого ИИ**

Краткая история Deep Learning

Le, 2013
Large scale unsupervised learning (16K CPU, W = 10^9)

Hochreiter, 1997
Long Short-Term Memory (**LSTM**)

LeCun, 1989
Handwritten zip-code recognition (**CNN**)

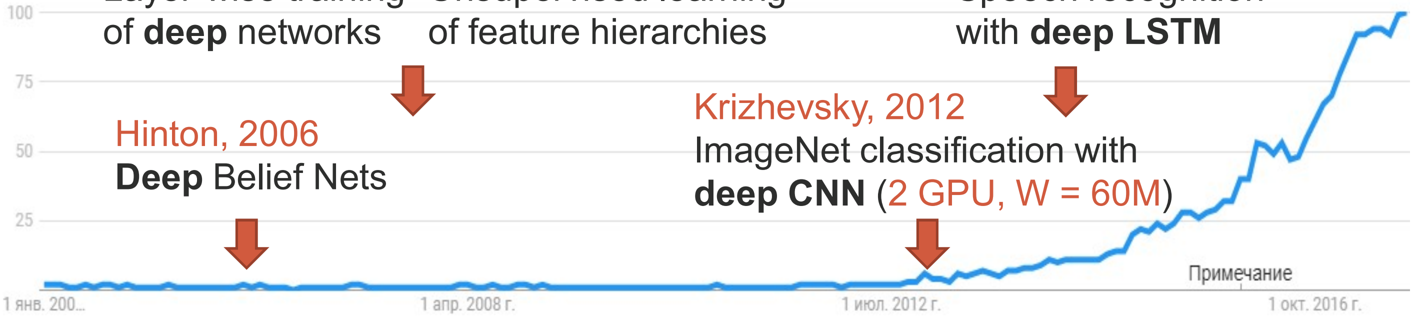
Bengio, 2007
Layer-wise training of **deep** networks

Hinton, 2006
Deep Belief Nets

LeCun, 2007
Unsupervised learning of feature hierarchies

Krizhevsky, 2012
ImageNet classification with **deep CNN** (2 GPU, W = 60M)

Graves, 2013
Speech recognition with **deep LSTM**



t ~ столетия

t ~ годы

t ~ месяцы

t ~ недели

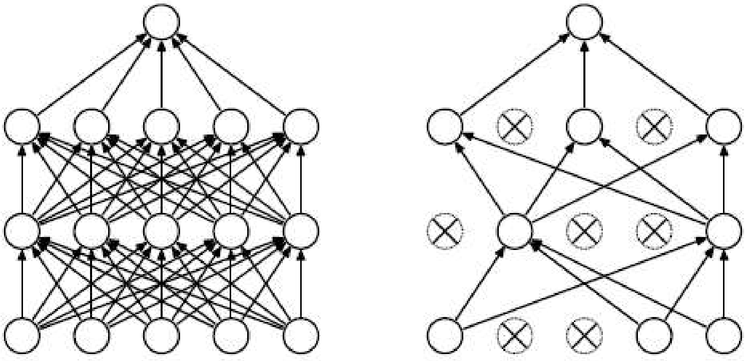
t ~ дни

Shallow nets

Layer-wise Deep nets

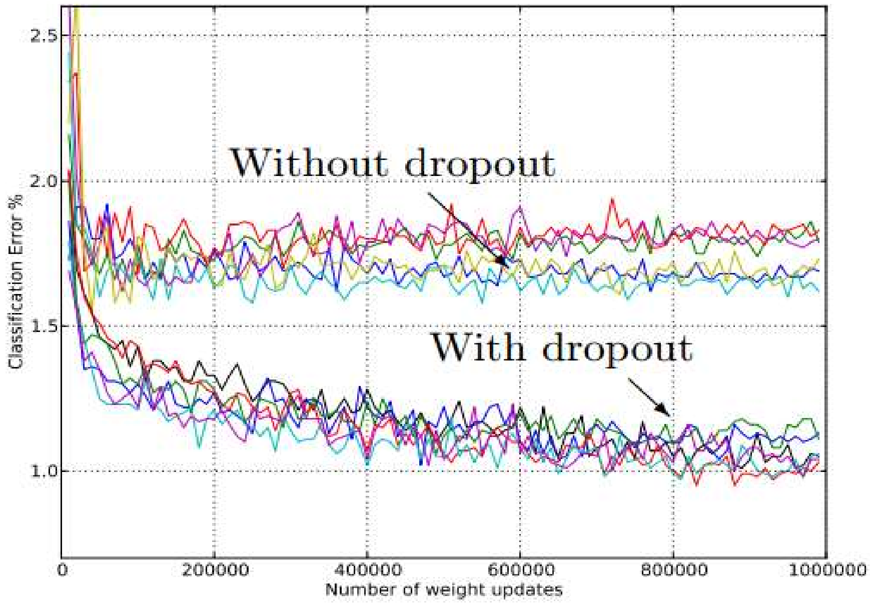
End-to-end Deep nets

Dropout



(a) Standard Neural Net

(b) After applying dropout.



Hinton (2012) *Improving neural networks by preventing co-adaptation of feature detectors*

Batch normalization

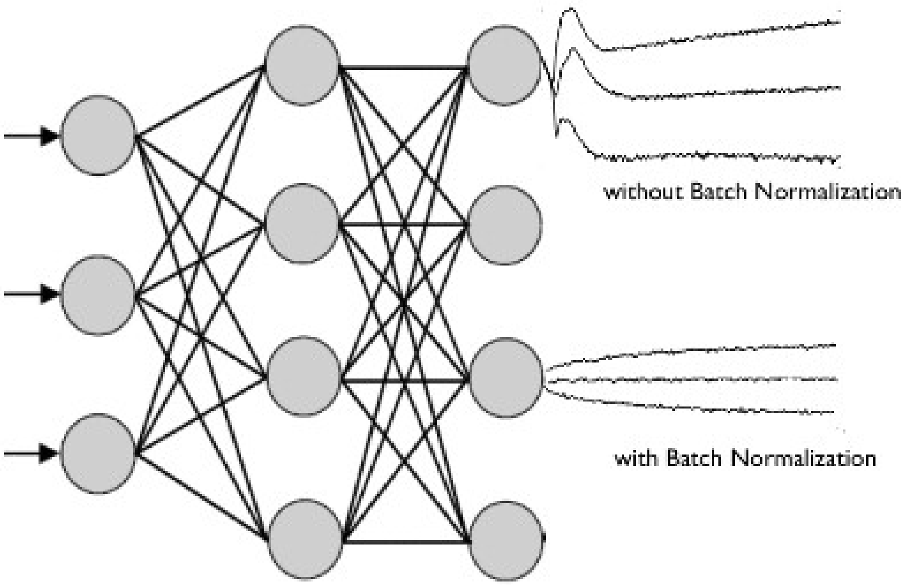
Ioffe (2015) *Batch normalization: Accelerating deep network training by reducing internal covariate shift*

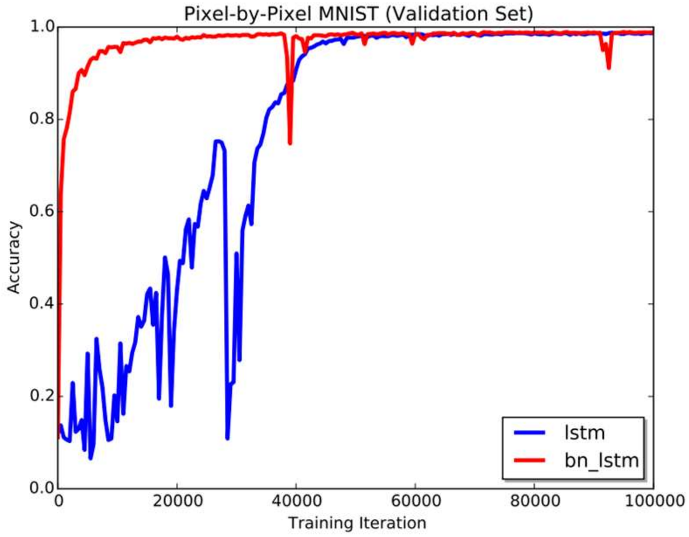
$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}}$$

$$y_i \leftarrow \gamma\widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i)$$





Big Data

- Widespread use of personal computers and wireless communication leads to “big data”
- We are both producers and consumers of data
- Data is not random, it has structure, e.g., customer behavior
- We need “big theory” to extract that structure from data for

  (a) Understanding the process

  (b) Making predictions for the future

Why “Learn” ?

- Machine learning is programming computers to optimize a performance criterion using example data or past experience.
- There is no need to “learn” to calculate payroll
- Learning is used when:
  - Human expertise does not exist (navigating on Mars),
  - 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)

What We Talk About When We Talk About “Learning”

- Learning general models from a data of particular examples
- Data is cheap and abundant (data warehouses, data marts); knowledge is expensive and scarce.
- Example in retail: Customer transactions to consumer behavior:

  People who bought “Blink” also bought “Outliers” (www.amazon.com)
- Build a model that is *a good and useful approximation* to the data.

Data Mining

- Retail: Market basket analysis, Customer relationship management (CRM)
- Finance: Credit scoring, fraud detection
- Manufacturing: Control, robotics, troubleshooting
- Medicine: Medical diagnosis
- Telecommunications: Spam filters, intrusion detection
- Bioinformatics: Motifs, alignment
- Web mining: Search engines
- …

What is Machine Learning?

- Optimize a performance criterion using example data or past experience.
- Role of Statistics: Inference from a sample
- Role of Computer science: Efficient algorithms to
  - Solve the optimization problem
  - Representing and evaluating the model for inference

Applications

- Association
- Supervised Learning
  - Classification
  - Regression
- Unsupervised Learning
- Reinforcement Learning

Learning Associations

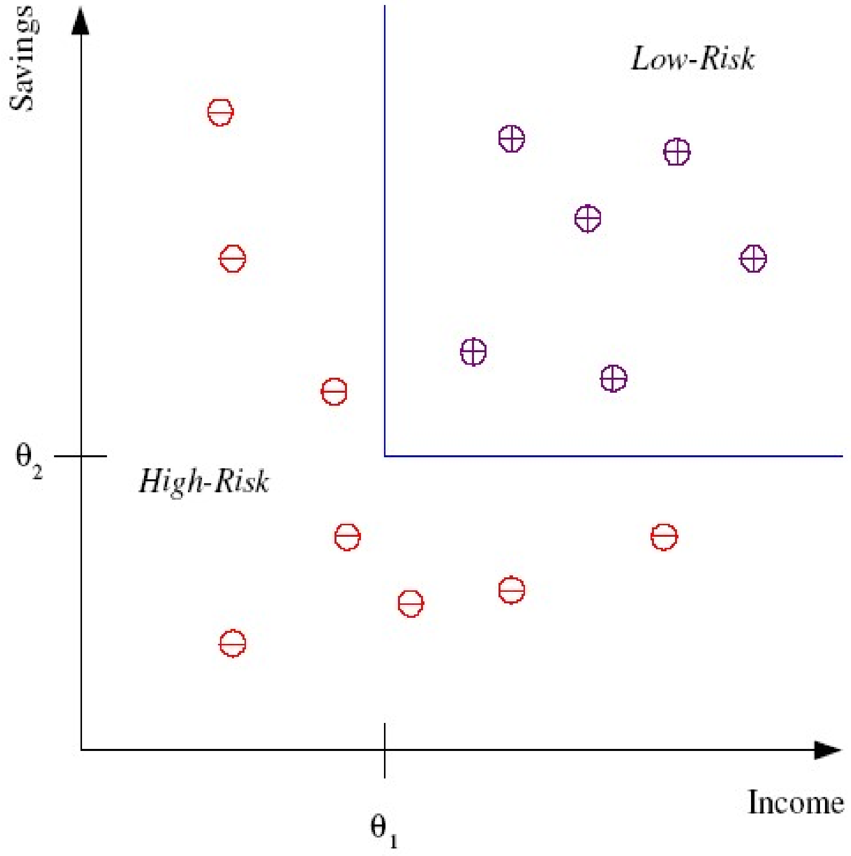
- Basket analysis:

  P (*Y* | *X*) probability that somebody who buys *X* also buys *Y* where *X* and *Y* are products/services.

  Example: *P* (chips | beer) = 0.7

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

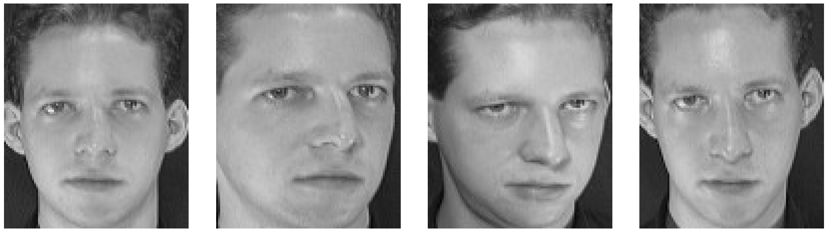
THEN low-risk ELSE high-risk

Classification: Applications

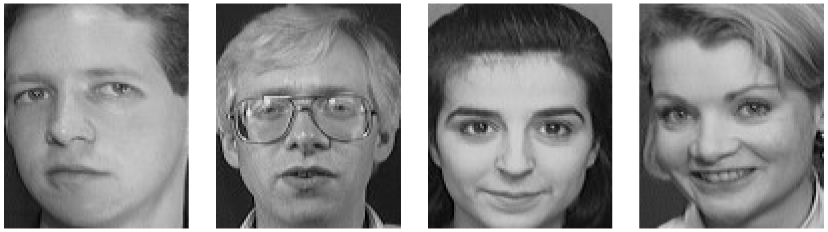
- Aka Pattern recognition
- Face recognition: Pose, lighting, occlusion (glasses, beard), make-up, hair style
- Character recognition: Different handwriting styles.
- Speech recognition: Temporal dependency.
- Medical diagnosis: From symptoms to illnesses
- Biometrics: Recognition/authentication using physical and/or behavioral characteristics: Face, iris, signature, etc
- Outlier/novelty detection:

Face Recognition

Training examples of a person



Test images



ORL dataset,
AT&T Laboratories, Cambridge UK

Regression

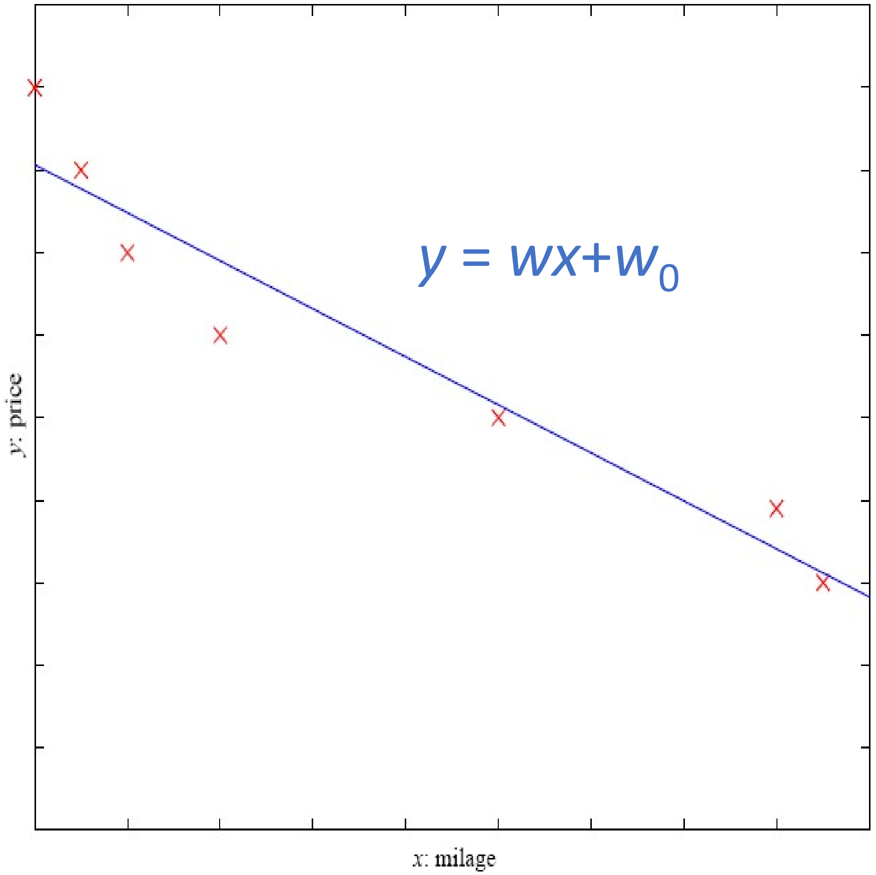
- Example: Price of a usec car
- x : car attributes

  y : price

  $$y = g\,(x \mid \theta)$$

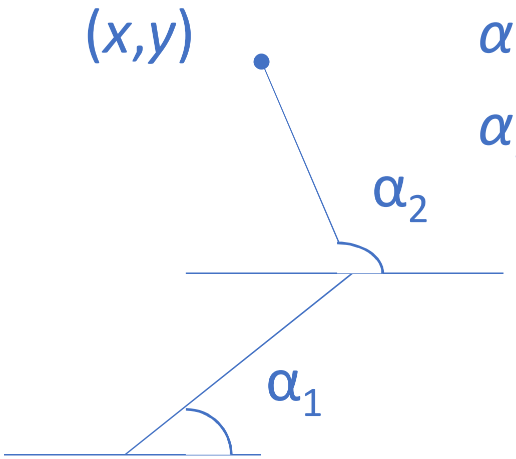
  g ( ) model,

  θ parameters



Regression Applications

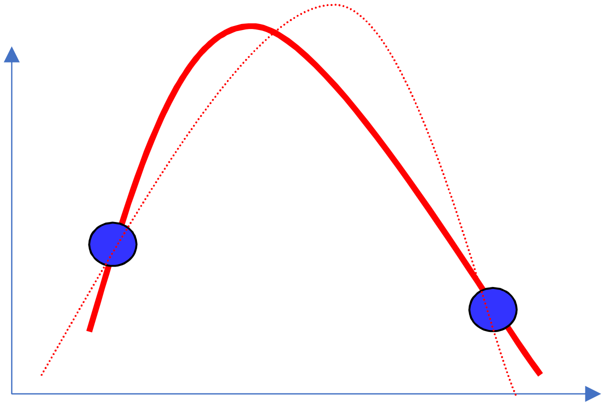
- Navigating a car: Angle of the steering
- Kinematics of a robot arm



$\alpha_1 = g_1(x,y)$

$\alpha_2 = g_2(x,y)$

- Response surface design



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
- Clustering: Grouping similar instances
- Example applications
    - Customer segmentation in CRM
    - Image compression: Color quantization
    - Bioinformatics: Learning motifs

Reinforcement Learning

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

Resources: Datasets

- UCI Repository: http://www.ics.uci.edu/~mlearn/MLRepository.html
- Statlib: http://lib.stat.cmu.edu/
- Kaggle: http://www.Kaggle.com/

Datasets

country	incomeperperson	alcconsumption	armedforcesrate	breastcancerper100th
Afghanistan		0.03	0.5696534	26.8
Albania	1914.996551	7.29	1.0247361	57.4
Algeria	2231.993335	0.69	2.306817	23.5
Andorra	21943.3399	10.17		
Angola	1381.004268	5.57	1.4613288	23.1

Codebook

Variable name: The name or number assigned to each variable in the data collection. Some researchers prefer to use mnemonic abbreviations (e.g., EMPLOY1), while others use alphanumeric patterns (e.g., VAR001). For survey data, try to name variables after the question numbers - e.g., Q1, Q2b, etc. [In above example, H40-SF12-2]

Variable label: A brief description to identify the variable for the user. Where possible, use the exact question or research wording. ["SF12 - ASSESSMENT OF R'S GENERAL HEALTH"]

Question text: Where applicable, the exact wording from survey questions. ["In general, would you say your health is . . ."]

Values: The actual coded values in the data for this variable. [1, 2, 3, 4, 5]

Value labels: The textual descriptions of the codes. [Excellent, Very Good, Good, Fair, Poor]

Codebook

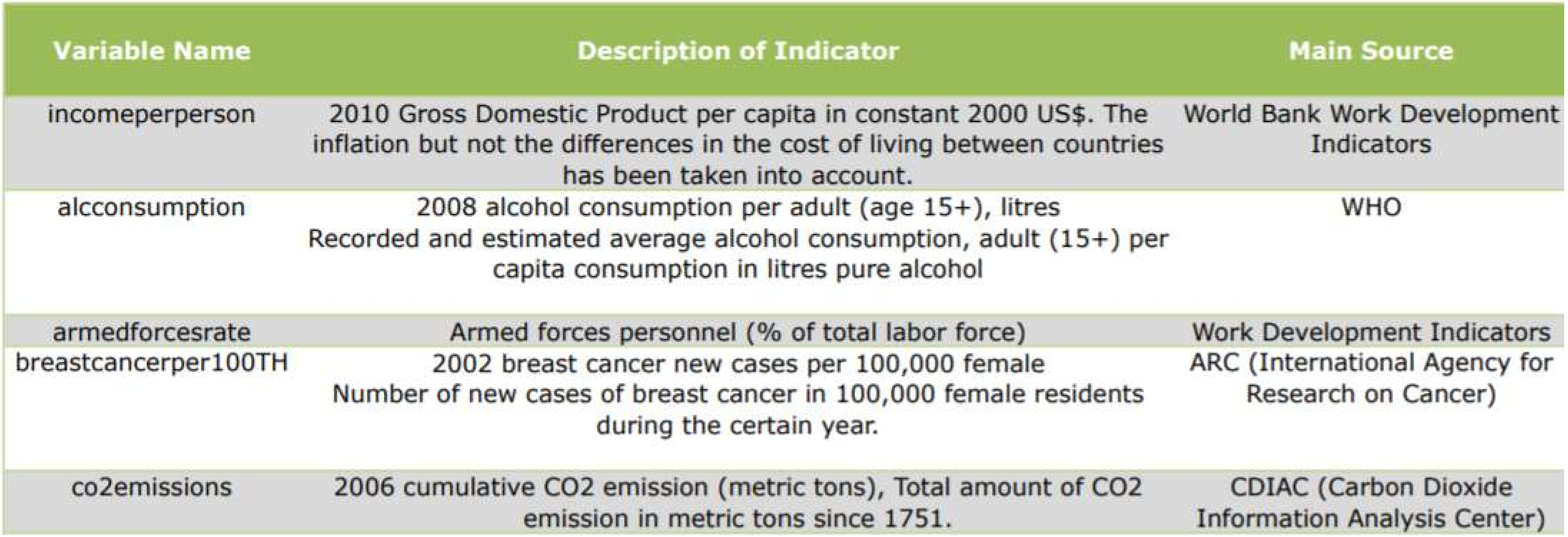
Summary statistics: Where appropriate and depending on the type of variable, provide unweighted summary statistics for quick reference. For categorical variables, for instance, frequency counts showing the number of times a value occurs and the percentage of cases that value represents for the variable are appropriate. For continuous variables, minimum, maximum, and median values are relevant.

Missing data: Where applicable, the values and labels of missing data. Missing data can bias an analysis and is important to convey in study documentation. Remember to describe all missing codes, including "system missing" and blank. [e.g., Refusal (-1)]

Universe skip patterns: Where applicable, information about the population to which the variable refers, as well as the preceding and following variables. [e.g., Default Next Question: H00035.00]

Notes: Additional notes, remarks, or comments that contextualize the information conveyed in the variable or relay special instructions. For measures or questions from copyrighted instruments, the notes field is the appropriate location to cite the source.

Codebook



Variable Name	Description of Indicator	Main Source
incomeperperson	2010 Gross Domestic Product per capita in constant 2000 US$. The inflation but not the differences in the cost of living between countries has been taken into account.	World Bank Work Development Indicators
alcconsumption	2008 alcohol consumption per adult (age 15+), litres Recorded and estimated average alcohol consumption, adult (15+) per capita consumption in litres pure alcohol	WHO
armedforcesrate	Armed forces personnel (% of total labor force)	Work Development Indicators
breastcancerper100TH	2002 breast cancer new cases per 100,000 female Number of new cases of breast cancer in 100,000 female residents during the certain year.	ARC (International Agency for Research on Cancer)
co2emissions	2006 cumulative CO2 emission (metric tons), Total amount of CO2 emission in metric tons since 1751.	CDIAC (Carbon Dioxide Information Analysis Center)

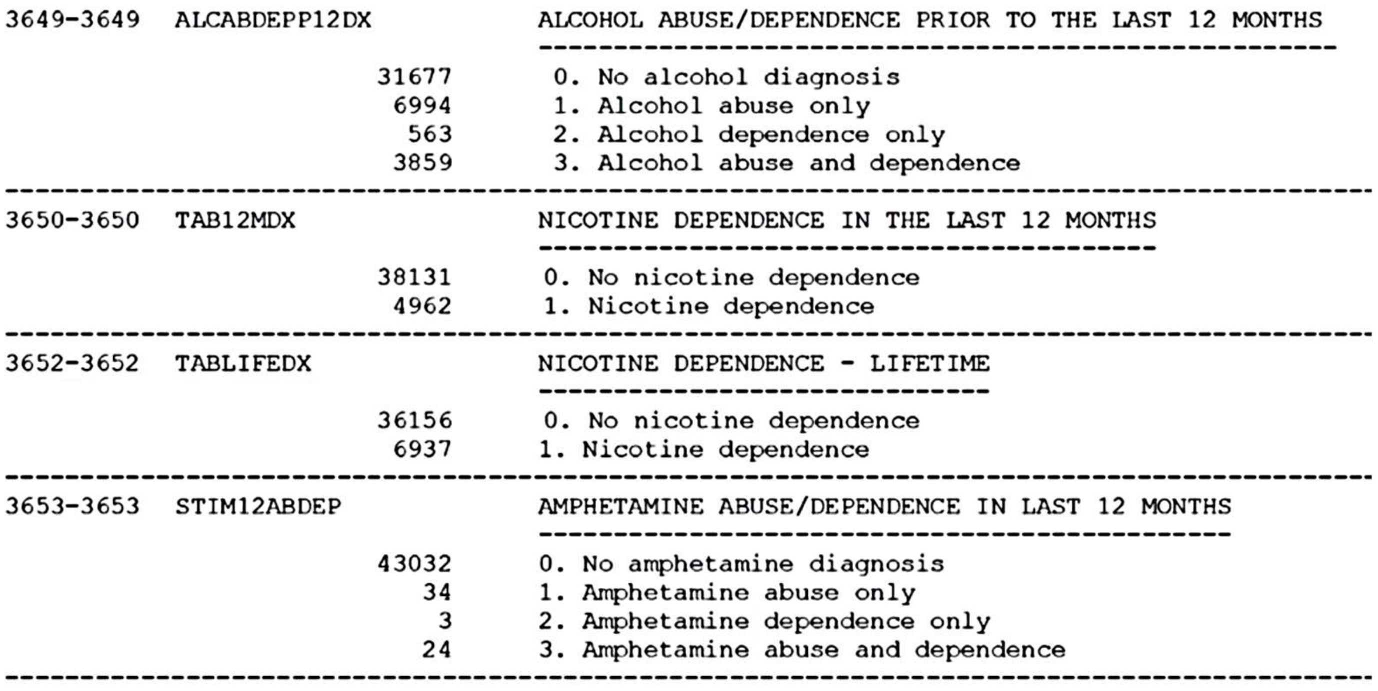
Variable: Categorical

- The blood type of a person: A, B, AB or O.
- The state that a person lives in.
- The political party that a voter in a European country might vote for: Christian Democrat, Social Democrat, Green Party, etc.
- The type of a rock: igneous, sedimentary or metamorphic.
- The identity of a particular word (e.g., in a language model): One of *V* possible choices, for a vocabulary of size *V*.

Variable: Quantitative

Variable	Measurement
Height	Inches, feet, centimeters,
Tempeture	Celsius, Fahrenheit, Kelvin, Réaumur...
Age	Years, months, decades, minutes
Weight	Pounds, tons, ounces, grams
Area	Acres, square miles, square feet
Speed	Miles per hour, light years, feet per second

Example with NESARC



```
3649-3649   ALCABDEPP12DX            ALCOHOL ABUSE/DEPENDENCE PRIOR TO THE LAST 12 MONTHS
                                     ----------------------------------------------------
                          31677       0. No alcohol diagnosis
                           6994       1. Alcohol abuse only
                            563       2. Alcohol dependence only
                           3859       3. Alcohol abuse and dependence
-----------------------------------------------------------------------------------------
3650-3650   TAB12MDX                 NICOTINE DEPENDENCE IN THE LAST 12 MONTHS
                                     -----------------------------------------
                          38131      0. No nicotine dependence
                           4962      1. Nicotine dependence
-----------------------------------------------------------------------------------------
3652-3652   TABLIFEDX                NICOTINE DEPENDENCE - LIFETIME
                                     ------------------------------
                          36156      0. No nicotine dependence
                           6937      1. Nicotine dependence
-----------------------------------------------------------------------------------------
3653-3653   STIM12ABDEP              AMPHETAMINE ABUSE/DEPENDENCE IN LAST 12 MONTHS
                                     ----------------------------------------------
                          43032      0. No amphetamine diagnosis
                             34      1. Amphetamine abuse only
                              3      2. Amphetamine dependence only
                             24      3. Amphetamine abuse and dependence
-----------------------------------------------------------------------------------------
```

Resources: Journals

- Journal of Machine Learning Research www.jmlr.org
- Machine Learning
- Neural Computation
- Neural Networks
- IEEE Trans on Neural Networks and Learning Systems
- IEEE Trans on Pattern Analysis and Machine Intelligence
- Journals on Statistics/Data Mining/Signal Processing/Natural Language Processing/Bioinformatics/...

Resources: Conferences

- International Conference on Machine Learning (ICML)
- European Conference on Machine Learning (ECML)
- Neural Information Processing Systems (NIPS)
- Uncertainty in Artificial Intelligence (UAI)
- Computational Learning Theory (COLT)
- International Conference on Artificial Neural Networks (ICANN)
- International Conference on AI & Statistics (AISTATS)
- International Conference on Pattern Recognition (ICPR)
- …