

# Fizikusok az adattudományban

Pafka Szilárd

Epoch (USA)

AtomCsill, ELTE TTK

Budapest, 2018. október

**Adattudomány / adatbányászat**  
data science / data mining

# **Adattudomány / adatbányászat**

data science / data mining

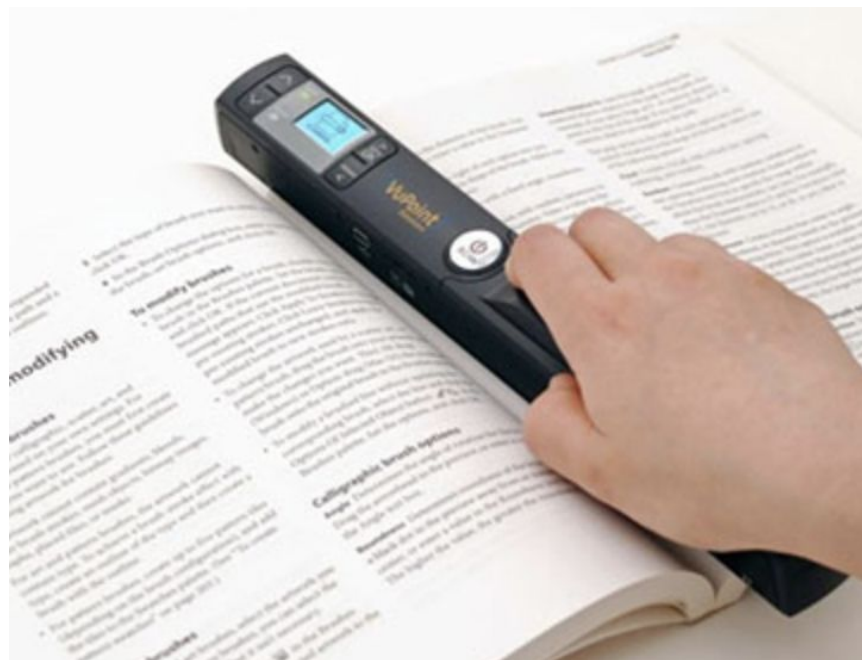
**adatok → érték**

# **Adattudomány / adatbányászat** data science / data mining

**adatok → érték**

adatok: számítógépes rendszerekben

érték: tudás, vmi hasznos, vállalatnál \$\$\$



Receipt image used for OCR testing



Tesseract output

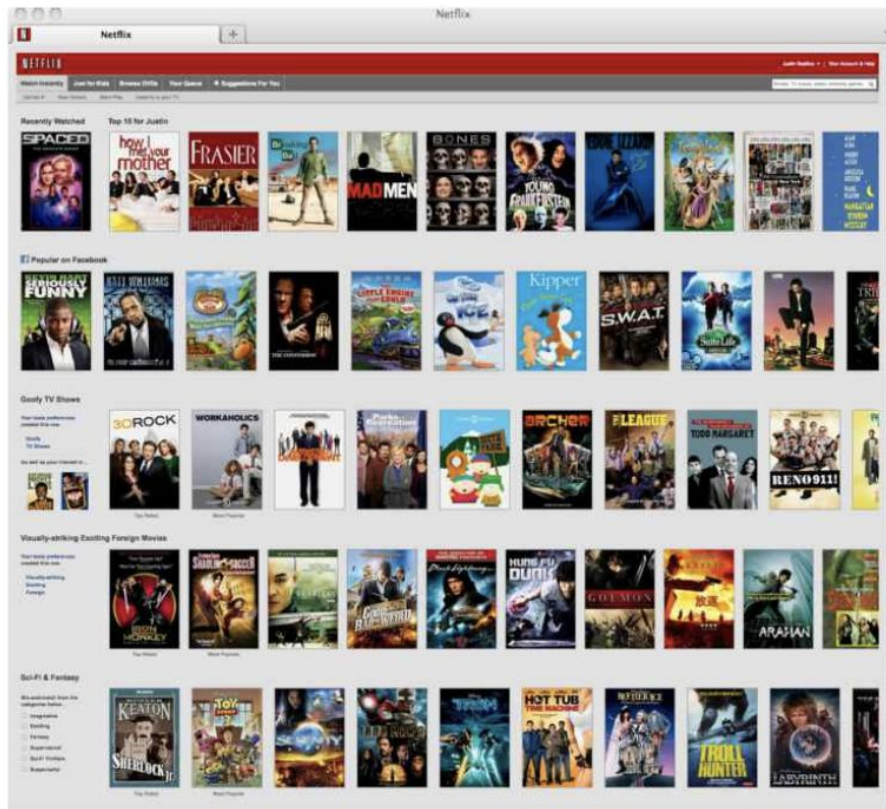
---

DARK CHOCOLATE  
MINTS, OZ Q,  
BLUEBERRIES, , OZ  
CANS, FLOZ FOR, CRV, ,

DARK CHOCOLATE MINTS	8.94
5OZ # 1.49/OZ	
BLUEBERRIES 1.5LB (24OZ)	5.99
JACK DANIELS WHISKY 750ML	17.99 T
HENNINGER BEER - 16 OZ CANS	5.99 T
8FLOZ FOR 5.99	
CRV	0.30 T
6 # 0.050	

Ranking

Rows





```
s/label_image.py /tf_files/diagnose/leg005.jpg
```

```
python /tf_files/label_image.py /tf_files/diagn
```



```
GraphDef version 9. Use tf.nn.  
brokenleg (score = 0.91144)  
healthyleg (score = 0.08856)
```



```
GraphDef version 9. Use tf.nn.  
healthyleg (score = 0.76665)  
brokenleg (score = 0.23335)
```





Google

in:spam

Gmail ▾



Delete forever

Not spam

COMPOSE

Delete all sp

Inbox (1)

Starred

Important

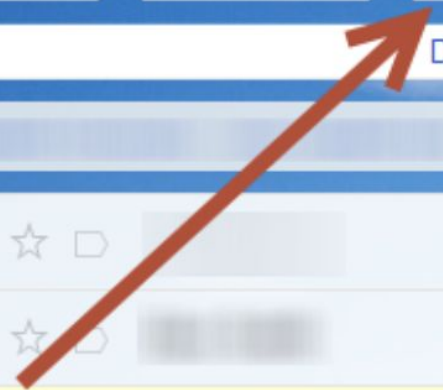
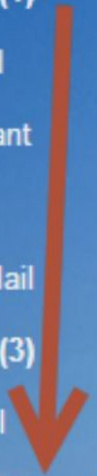
Chats

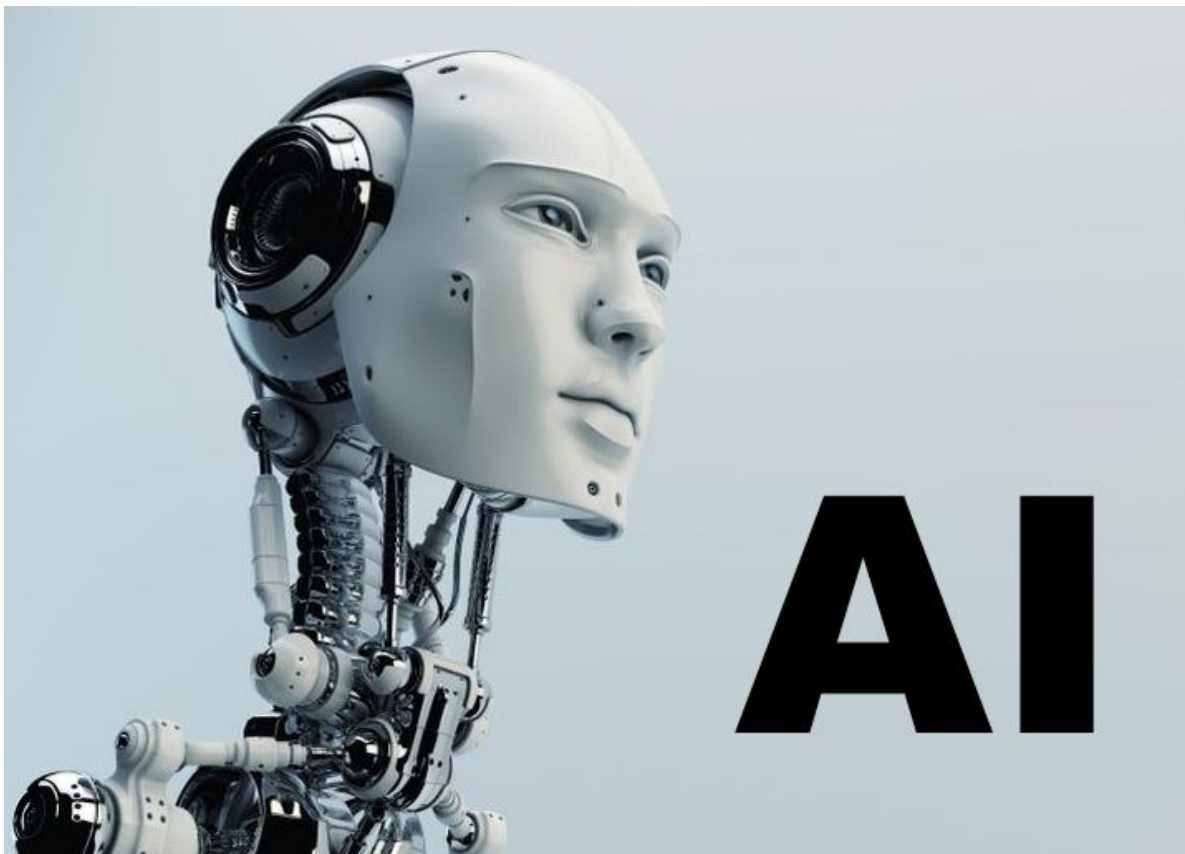
Sent Mail

Drafts (3)

All Mail

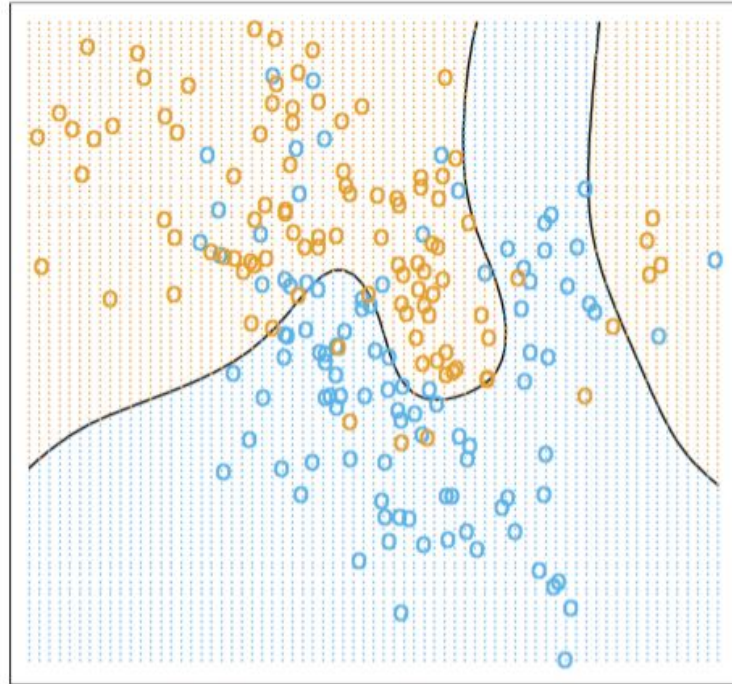
Spam (13)





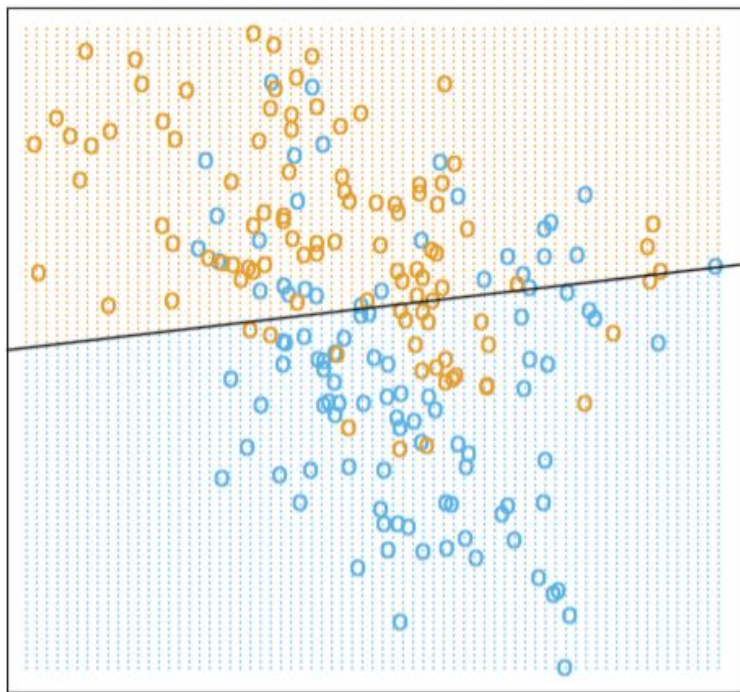
$$y = f(x_1, x_2, \dots, x_n)$$

Bayes Optimal Classifier

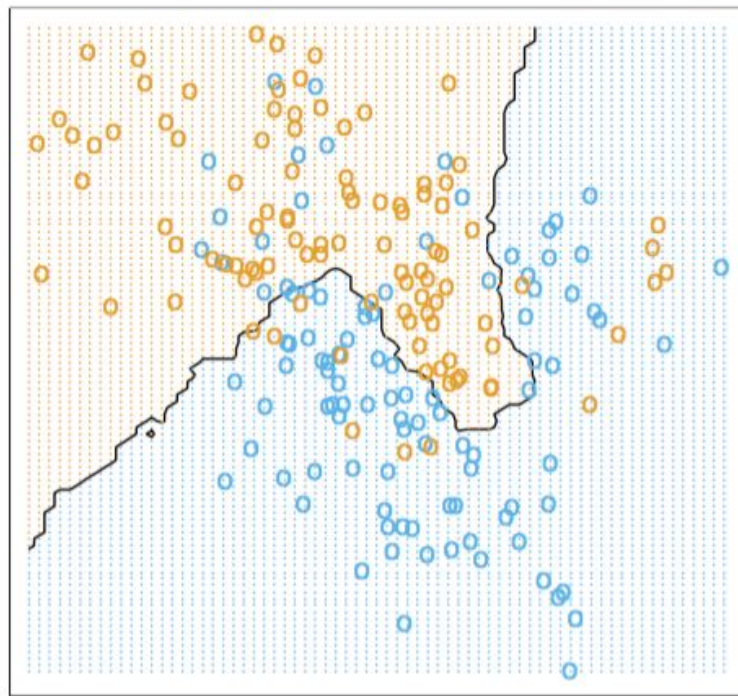


$$y = f(x_1, x_2, \dots, x_n)$$

Linear Regression of 0/1 Response



15-Nearest Neighbor Classifier



# Adattudomány?

The image shows a page of handwritten astronomical tables, possibly from a historical manuscript. The tables are organized into sections labeled with Roman numerals (X, XI, XII, XIII, XIV) and contain columns of numbers and text. The numbers appear to be celestial coordinates or planetary positions, with some entries including names of celestial bodies or locations. The handwriting is in a cursive script, and the paper shows signs of age and wear.

**X**

10.11.16	10.11.16	10.11.16	10.11.16	10.11.16
10.11.16	10.11.16	10.11.16	10.11.16	10.11.16
10.11.16	10.11.16	10.11.16	10.11.16	10.11.16
10.11.16	10.11.16	10.11.16	10.11.16	10.11.16
10.11.16	10.11.16	10.11.16	10.11.16	10.11.16

**XI**

10.11.16	10.11.16	10.11.16	10.11.16	10.11.16
10.11.16	10.11.16	10.11.16	10.11.16	10.11.16
10.11.16	10.11.16	10.11.16	10.11.16	10.11.16
10.11.16	10.11.16	10.11.16	10.11.16	10.11.16
10.11.16	10.11.16	10.11.16	10.11.16	10.11.16

**XII**

10.11.16	10.11.16	10.11.16	10.11.16	10.11.16
10.11.16	10.11.16	10.11.16	10.11.16	10.11.16
10.11.16	10.11.16	10.11.16	10.11.16	10.11.16
10.11.16	10.11.16	10.11.16	10.11.16	10.11.16
10.11.16	10.11.16	10.11.16	10.11.16	10.11.16

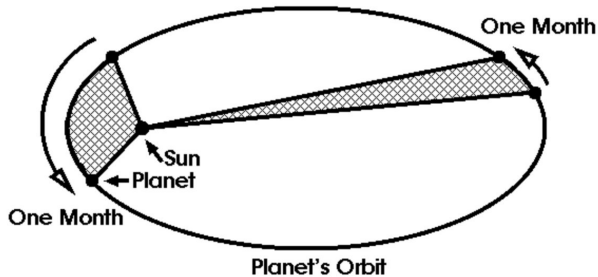
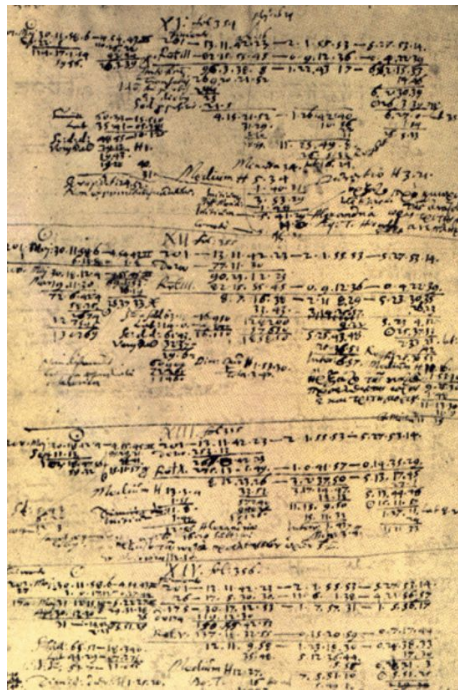
**XIII**

10.11.16	10.11.16	10.11.16	10.11.16	10.11.16
10.11.16	10.11.16	10.11.16	10.11.16	10.11.16
10.11.16	10.11.16	10.11.16	10.11.16	10.11.16
10.11.16	10.11.16	10.11.16	10.11.16	10.11.16
10.11.16	10.11.16	10.11.16	10.11.16	10.11.16

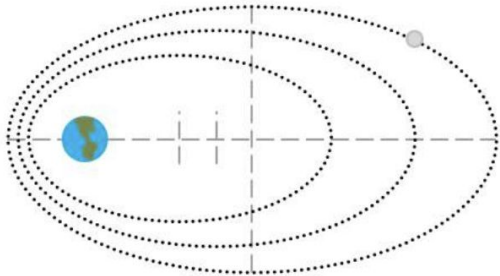
**XIV**

10.11.16	10.11.16	10.11.16	10.11.16	10.11.16
10.11.16	10.11.16	10.11.16	10.11.16	10.11.16
10.11.16	10.11.16	10.11.16	10.11.16	10.11.16
10.11.16	10.11.16	10.11.16	10.11.16	10.11.16
10.11.16	10.11.16	10.11.16	10.11.16	10.11.16

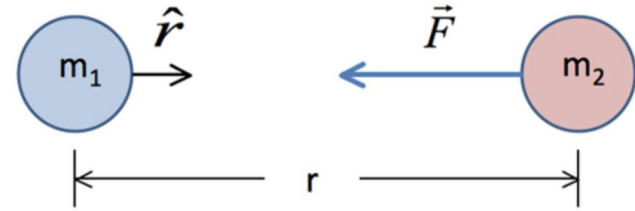
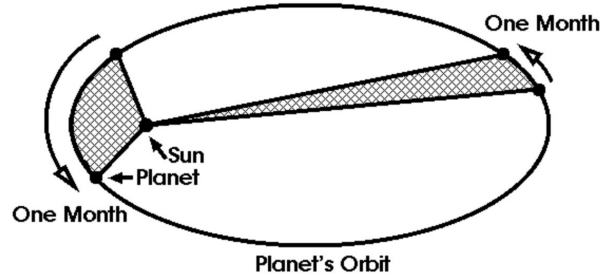
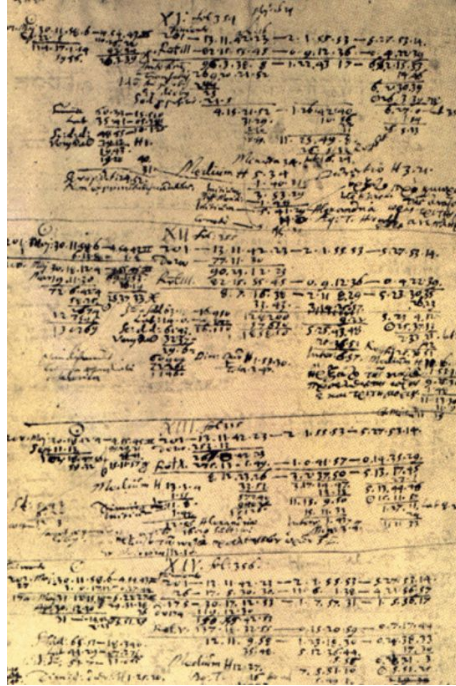
# Adattudomány?



$$\frac{T_1^2}{T_2^2} = \frac{R_1^3}{R_2^3}$$

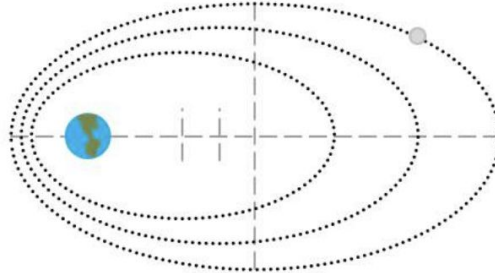


# Adattudomány?



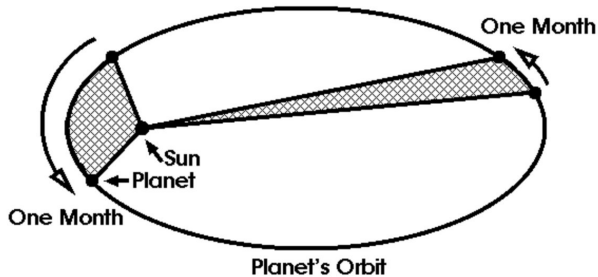
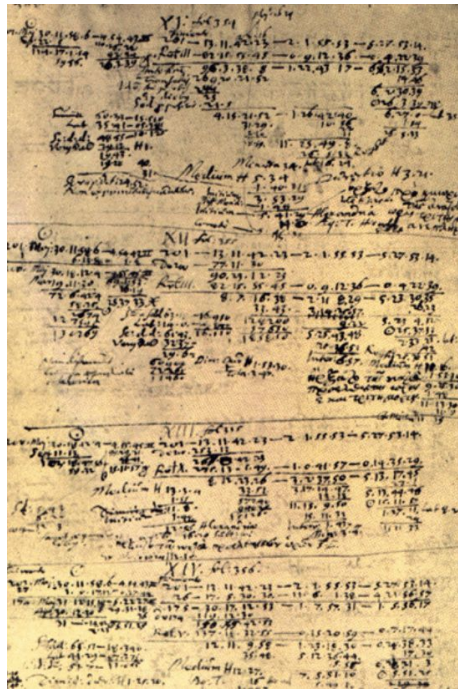
$$\vec{F} = -\frac{Gm_1m_2}{r^2}\hat{r}$$

$$\frac{T_1^2}{T_2^2} = \frac{R_1^3}{R_2^3}$$

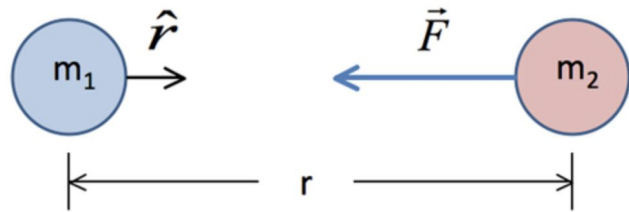
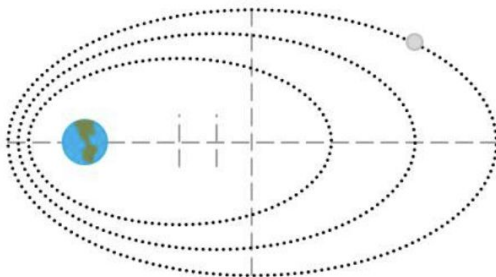




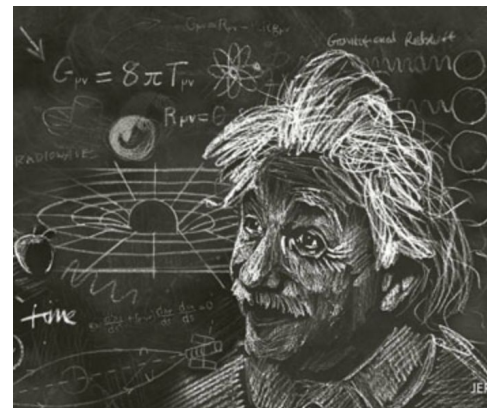
# Adattudomány?

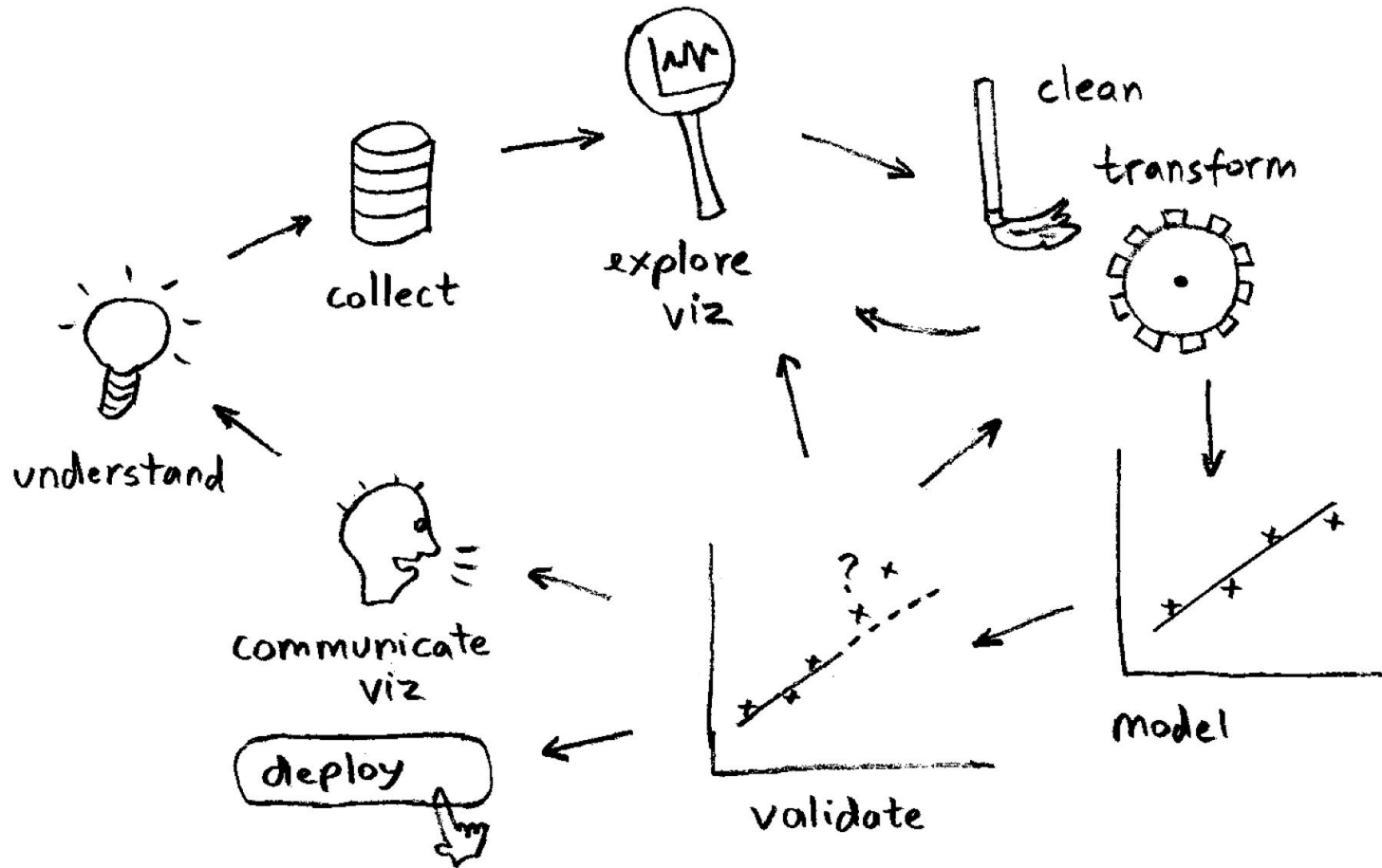


$$\frac{T_1^2}{T_2^2} = \frac{R_1^3}{R_2^3}$$



$$\vec{F} = -\frac{Gm_1m_2}{r^2}\hat{r}$$





 [szilard](#) / [teach-data-science-msc-analytics-ceu](#)

 [szilard](#) / [teach-data-science-msc-analytics-ceu](#)

### Sample projects from students:

- Laszlo Sallo: [insurance risk prediction](#) also in Kaggle competition
- Oliver Kocsis: [classification of body postures](#)

# Wearable Computing: Classification of Body Postures and Movements (PUC-Rio) Data Set

*Oliver Kocsis*

*February 23, 2016*

### Sample projects from students:

- Laszlo Sallo: [insurance risk prediction](#) also in Kaggle competition
- Oliver Kocsis: [classification of body postures](#)

# Wearable Computing: Classification of Body Postures and Movements (PUC-Rio) Data Set

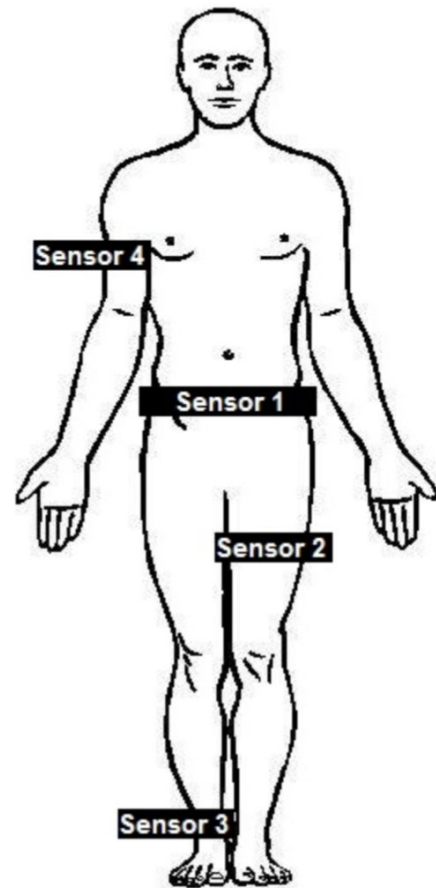
*Oliver Kocsis*

*February 23, 2016*



## Wearable Computing: Classification of Body Postures

Download: [Data Folder](#), [Data Set Description](#)





# The R Project for Statistical Computing



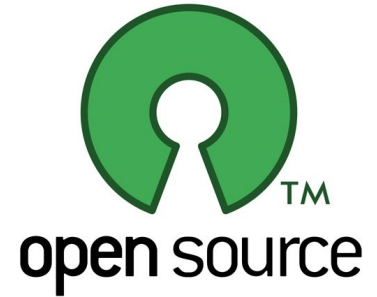
# The R Project for Statistical Computing

- adatok beolvasása
- adatok manipulációja
- adatvizualizáció (ábrák)
- statisztikai modellezés
- modellek felhasználása



# The R Project for Statistical Computing

- adatok beolvasása
- adatok manipulációja
- adatvizualizáció (ábrák)
- statisztikai modellezés
- modellek felhasználása

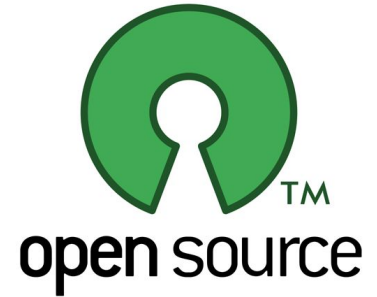






# The R Project for Statistical Computing

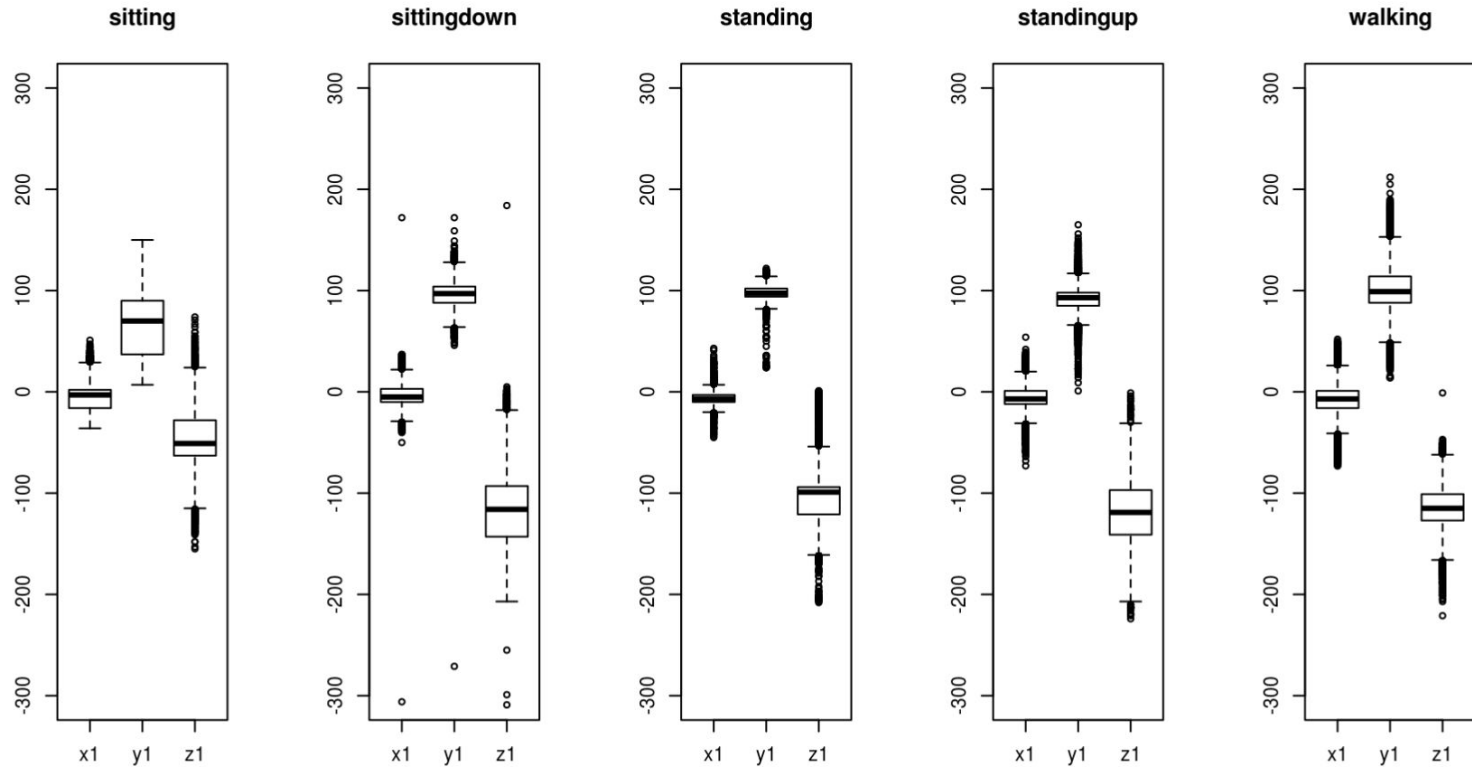
- adatok beolvasása
- adatok manipulációja
- adatvizualizáció (ábrák)
- statisztikai modellezés
- modellek felhasználása



```
data <- read.csv("dataset-har-PUC-Rio-ugolino.csv", sep = ";")
setDT(data)
colnames(data)[4] <- "height"
data[, gender := as.factor(gender)]
data[, age := as.integer(age)]
```

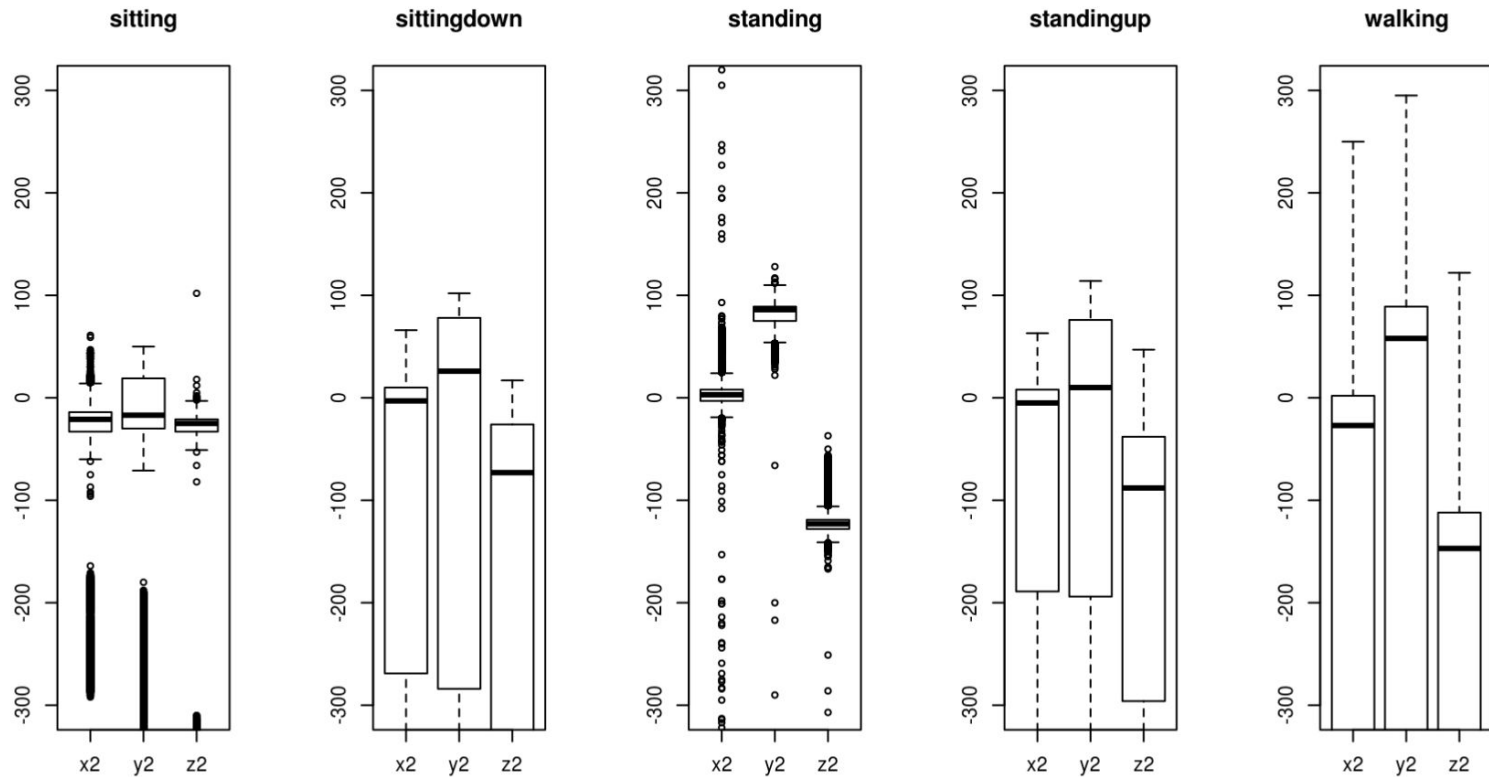
# Sensor 1 Waist

```
boxplot(sensor_1, NULL)
```



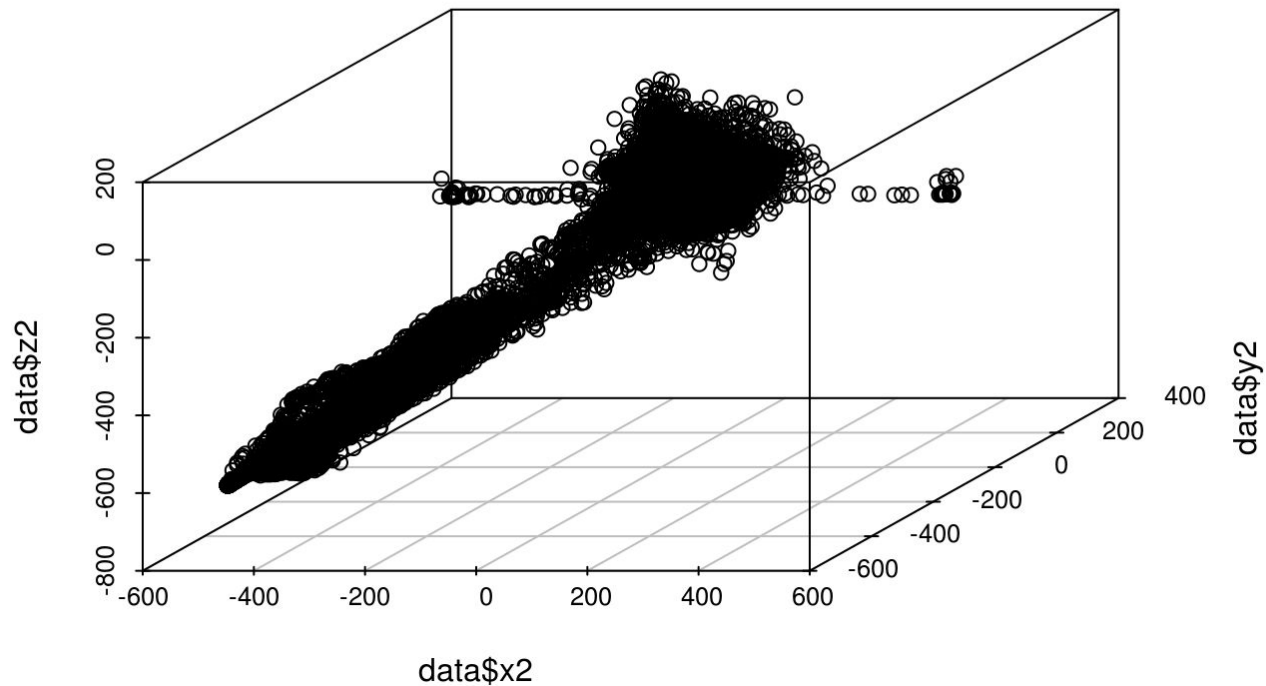
# Sensor 2 Left Thigh

```
boxplot.sensor(sensor_2, NULL)
```



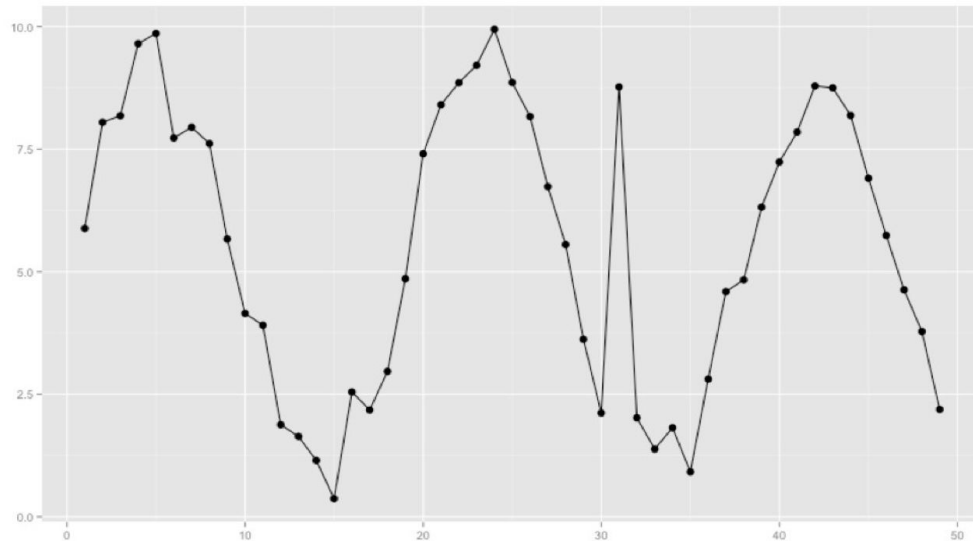
```
scatterplot3d(data$x2, data$y2, data$z2, main="Sensor 2")
```

## Sensor 2

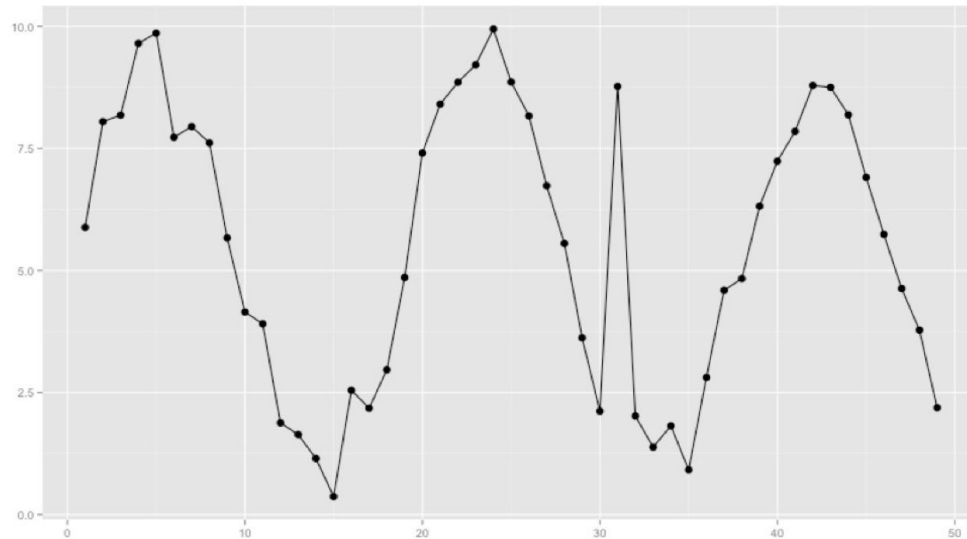


5.8839338 8.0500895 8.1838378 9.6537864 9.8625664 7.7283027 7.9485545  
7.6139286 5.6673501 4.1509576 3.9085585 1.8794583 1.6390833 1.1494471  
0.3701523 2.5463324 2.1793825 2.9664571 4.8563495 7.4056122 8.4070250  
8.8555943 9.2110448 9.9459725 8.8605880 8.1672658 6.7366060 5.5535530  
3.6201724 2.1181429 8.7715833 2.0190151 1.3814497 1.8169363 0.9166560  
2.8093003 4.5931840 4.8333278 6.3170302 7.2390577 7.8509665 8.7915221  
8.7507326 8.1899304 6.9060409 5.7413247 4.6312077 3.7803116 2.1903559

5.8839338 8.0500895 8.1838378 9.6537864 9.8625664 7.7283027 7.9485545  
7.6139286 5.6673501 4.1509576 3.9085585 1.8794583 1.6390833 1.1494471  
0.3701523 2.5463324 2.1793825 2.9664571 4.8563495 7.4056122 8.4070250  
8.8555943 9.2110448 9.9459725 8.8605880 8.1672658 6.7366060 5.5535530  
3.6201724 2.1181429 8.7715833 2.0190151 1.3814497 1.8169363 0.9166560  
2.8093003 4.5931840 4.8333278 6.3170302 7.2390577 7.8509665 8.7915221  
8.7507326 8.1899304 6.9060409 5.7413247 4.6312077 3.7803116 2.1903559



5.8839338 8.0500895 8.1838378 9.6537864 9.8625664 7.7283027 7.9485545  
7.6139286 5.6673501 4.1509576 3.9085585 1.8794583 1.6390833 1.1494471  
0.3701523 2.5463324 2.1793825 2.9664571 4.8563495 7.4056122 8.4070250  
8.8555943 9.2110448 9.9459725 8.8605880 8.1672658 6.7366060 5.5535530  
3.6201724 2.1181429 **8.7715833** 2.0190151 1.3814497 1.8169363 0.9166560  
2.8093003 4.5931840 4.8333278 6.3170302 7.2390577 7.8509665 8.7915221  
8.7507326 8.1899304 6.9060409 5.7413247 4.6312077 3.7803116 2.1903559





**Big Data Borat**

@BigDataBorat

 Follow

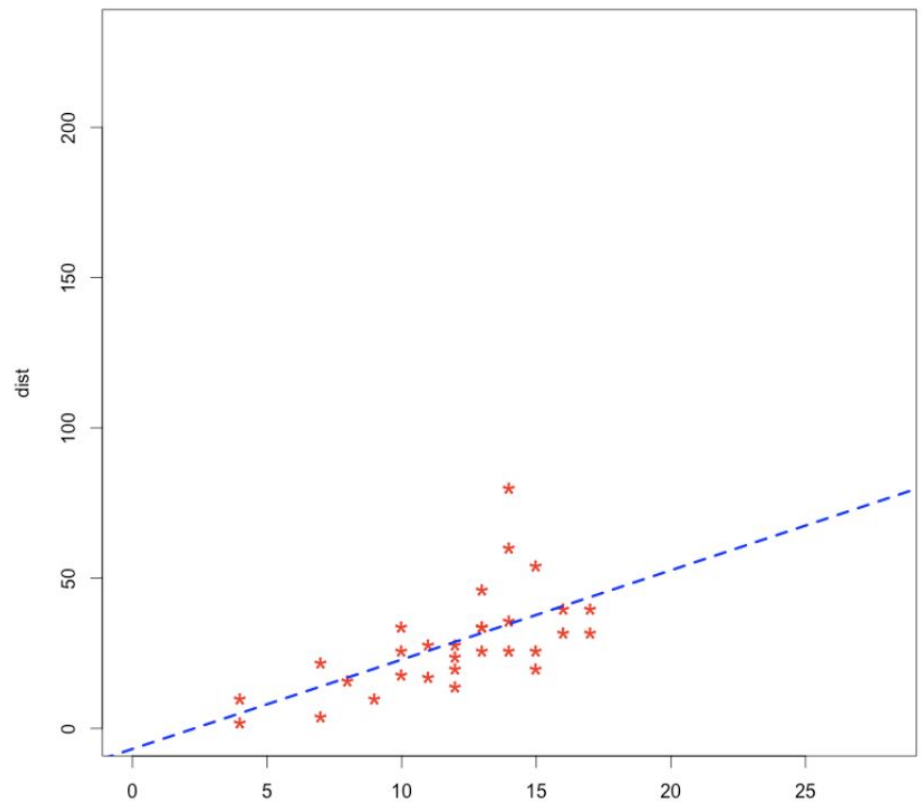
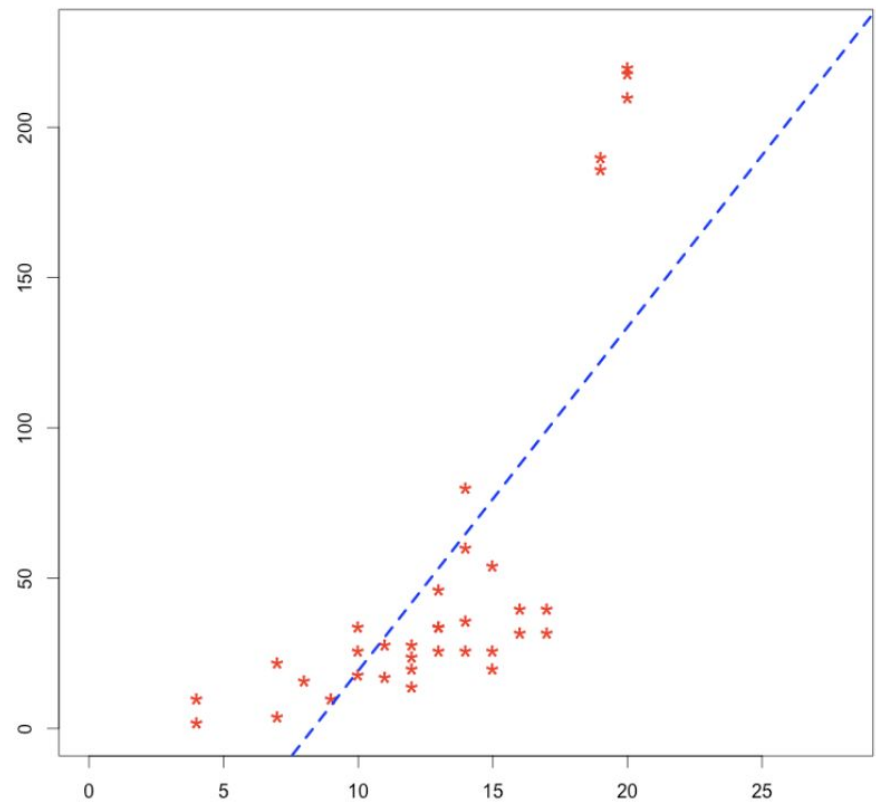
In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.

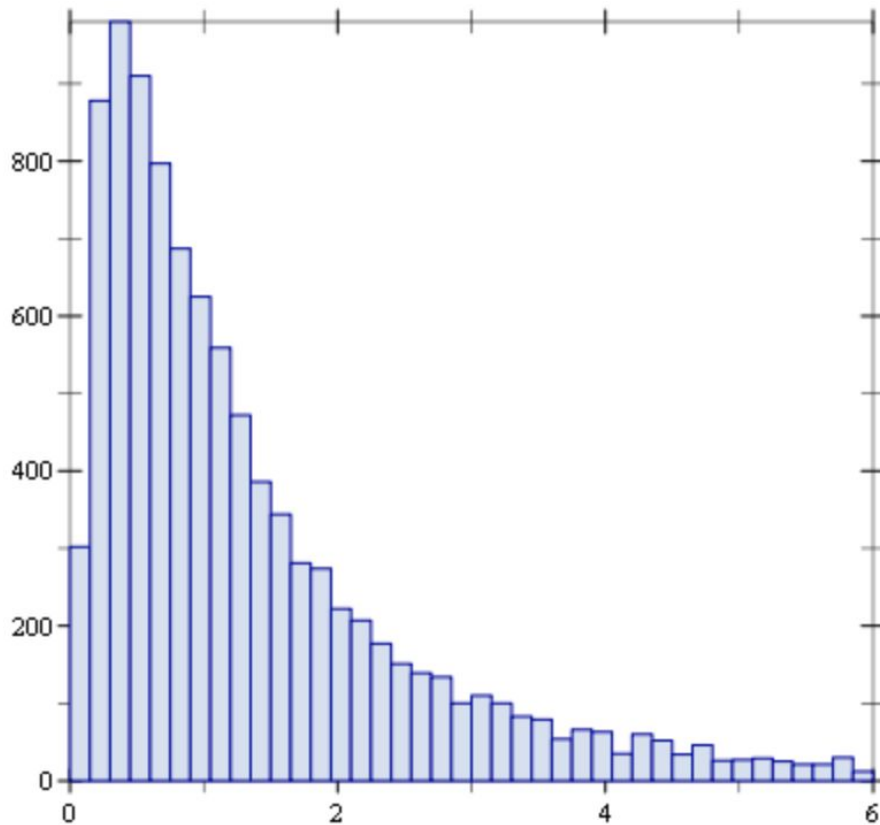
 Reply  Retweet  Favorite  More



# Adatbányászat?

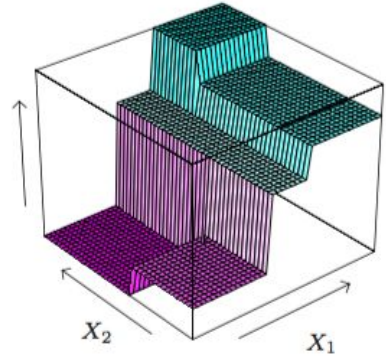
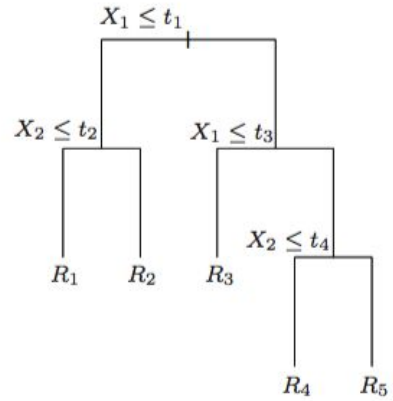
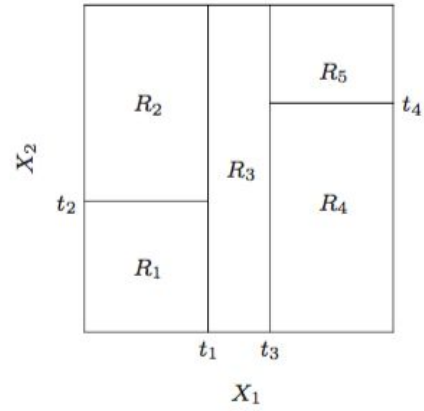
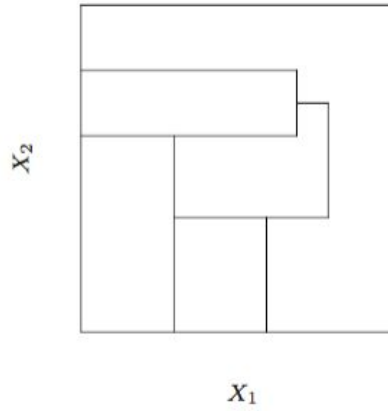


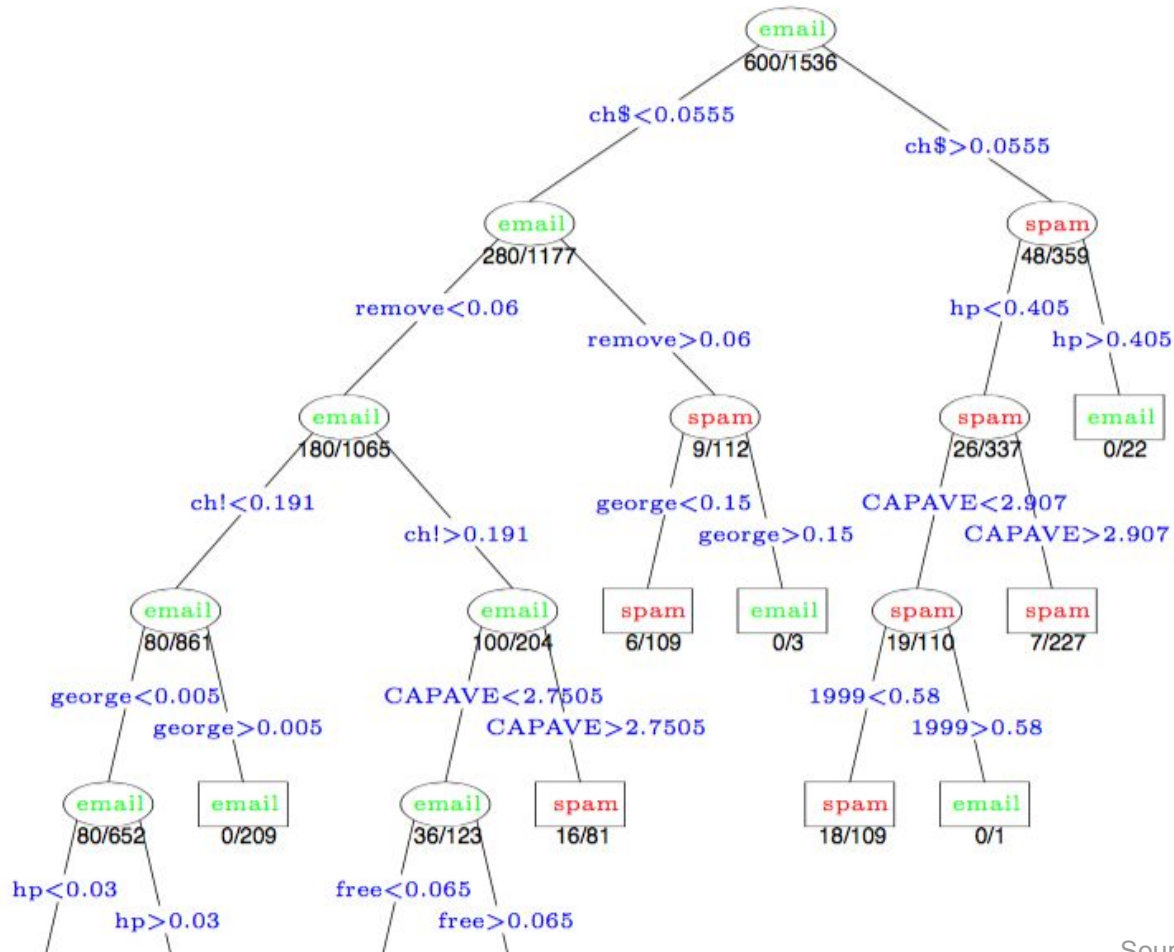




- valószínűségi eloszlások
- valószínűségi összefüggések
- valószínűségszámítás
- statisztika
- ...

```
gbm <- h2o.gbm(model_id = "gbm", x = d_x_idxxs, y = d_y_idxxs,  
  training_frame = dx_train, validation_frame = dx_valid,  
  ntrees = 300, max_depth = 110, nbins = 60, learn_rate = 0.5,  
  stopping_rounds = 5, stopping_tolerance = 1e-3, seed = seed)
```





---

**Algorithm 10.1** *AdaBoost.M1*.

---

1. Initialize the observation weights  $w_i = 1/N$ ,  $i = 1, 2, \dots, N$ .
2. For  $m = 1$  to  $M$ :
  - (a) Fit a classifier  $G_m(x)$  to the training data using weights  $w_i$ .
  - (b) Compute
$$\text{err}_m = \frac{\sum_{i=1}^N w_i I(y_i \neq G_m(x_i))}{\sum_{i=1}^N w_i}.$$
  - (c) Compute  $\alpha_m = \log((1 - \text{err}_m)/\text{err}_m)$ .
  - (d) Set  $w_i \leftarrow w_i \cdot \exp[\alpha_m \cdot I(y_i \neq G_m(x_i))]$ ,  $i = 1, 2, \dots, N$ .
3. Output  $G(x) = \text{sign} \left[ \sum_{m=1}^M \alpha_m G_m(x) \right]$ .

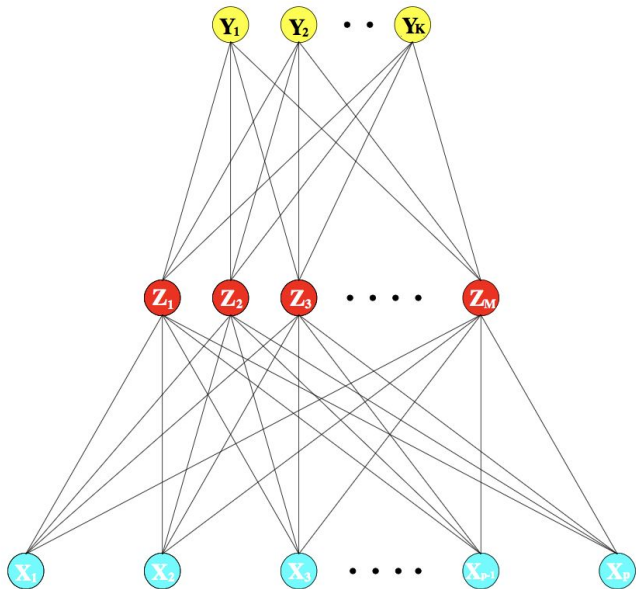
```
gbm <- h2o.gbm(model_id = "gbm", x = d_x_idx, y = d_y_idx,  
  training_frame = dx_train, validation_frame = dx_valid,  
  ntrees = 300, max_depth = 110, nbins = 60, learn_rate = 0.5,  
  stopping_rounds = 5, stopping_tolerance = 1e-3, seed = seed)
```

Confusion Matrix: vertical: actual; across: predicted

	sitting	sittingdown	standing	standingup	walking
sitting	12697	4	0	7	0
sittingdown	3	2751	17	35	22
standing	0	2	11779	8	39
standingup	6	62	27	2939	36
walking	1	14	27	7	10925



```
deephlearning <- h2o.deephlearning(model_id = "deephlearning", x = d_x_idxxs, y = d_y_idxxs,  
  training_frame = dx_train, validation_frame = dx_valid,  
  activation = "Tanh", hidden = c(200,200), epochs = 30,  
  stopping_rounds = 5, stopping_tolerance = 1e-3, seed = seed)
```

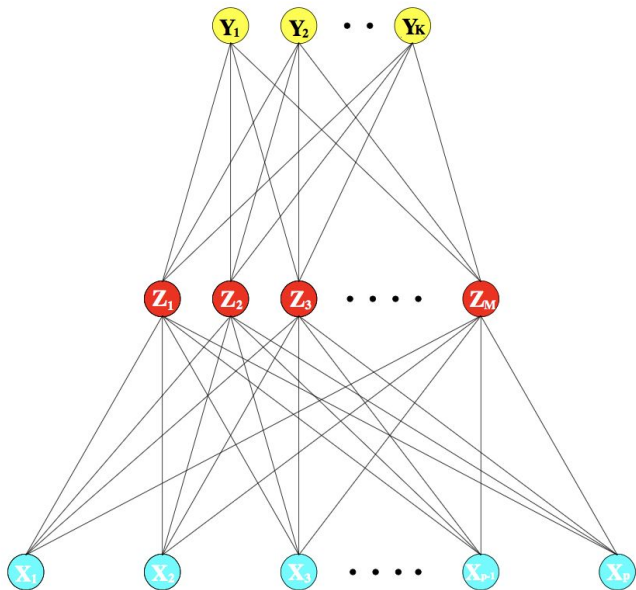


$$Z_m = \sigma(\alpha_{0m} + \alpha_m^T X), \quad m = 1, \dots, M,$$

$$T_k = \beta_{0k} + \beta_k^T Z, \quad k = 1, \dots, K,$$

$$f_k(X) = g_k(T), \quad k = 1, \dots, K,$$

**FIGURE 11.2.** Schematic of a single hidden layer, feed-forward neural network.



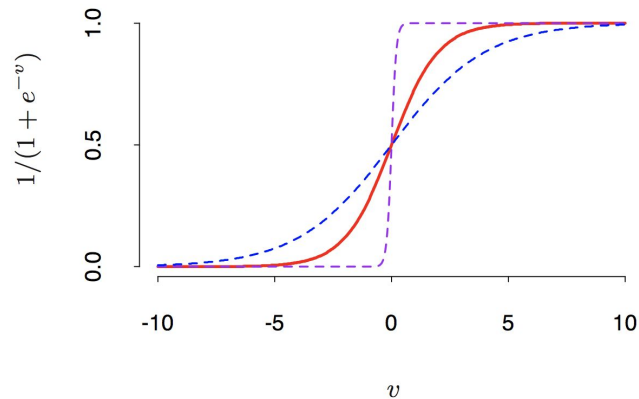
**FIGURE 11.2.** Schematic of a single hidden layer, feed-forward neural network.

$$Z_m = \sigma(\alpha_{0m} + \alpha_m^T X), \quad m = 1, \dots, M,$$

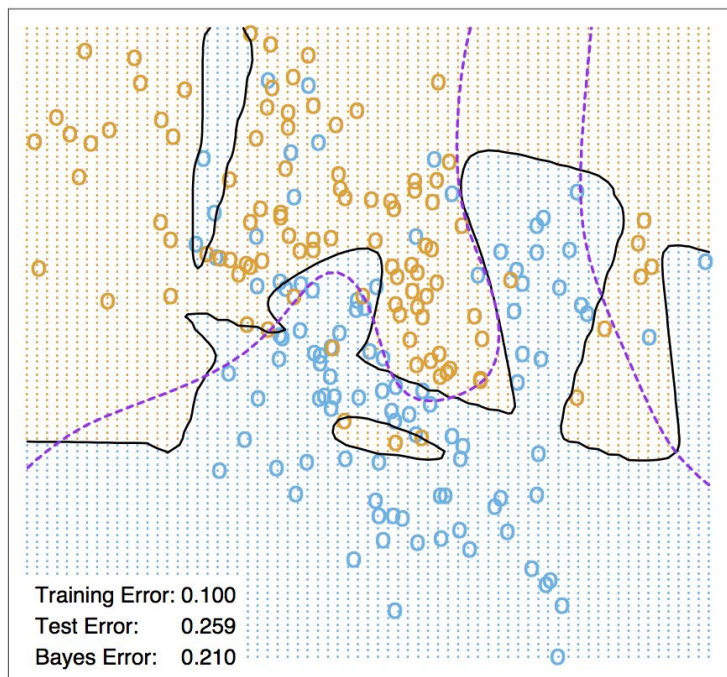
$$T_k = \beta_{0k} + \beta_k^T Z, \quad k = 1, \dots, K,$$

$$f_k(X) = g_k(T), \quad k = 1, \dots, K,$$

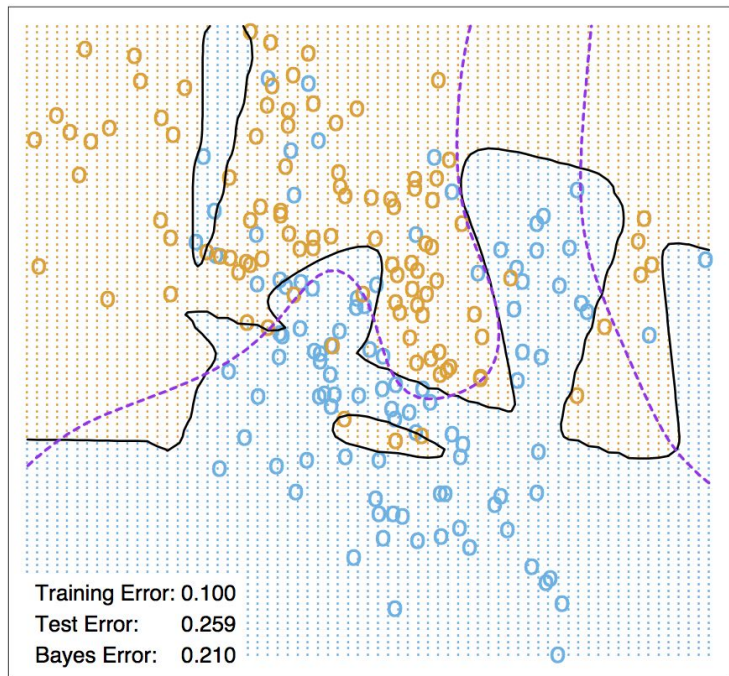
$$g_k(T) = \frac{e^{T_k}}{\sum_{\ell=1}^K e^{T_\ell}}.$$



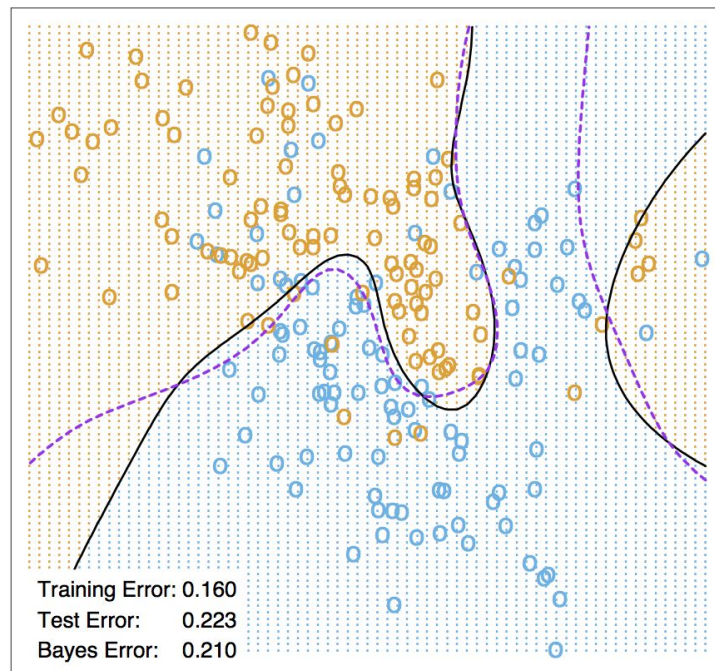
## Neural Network - 10 Units, No Weight Decay

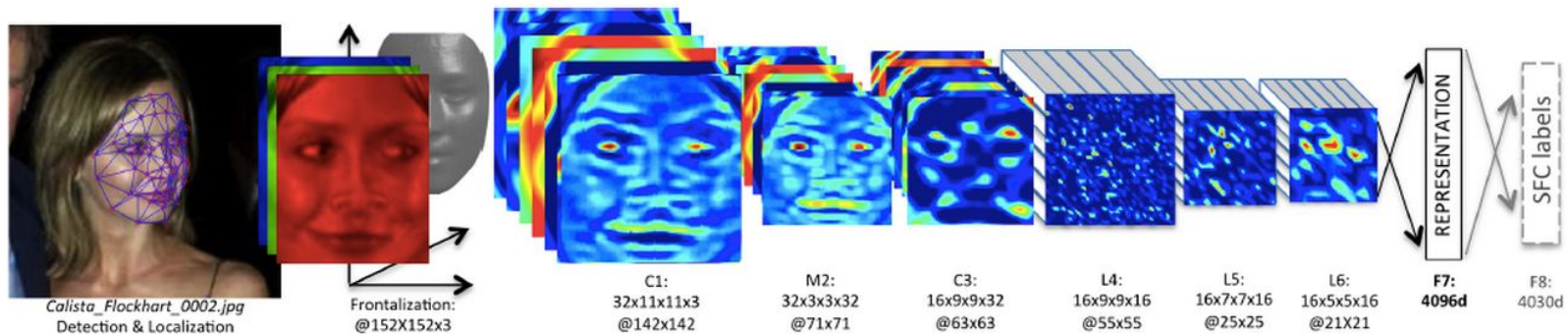


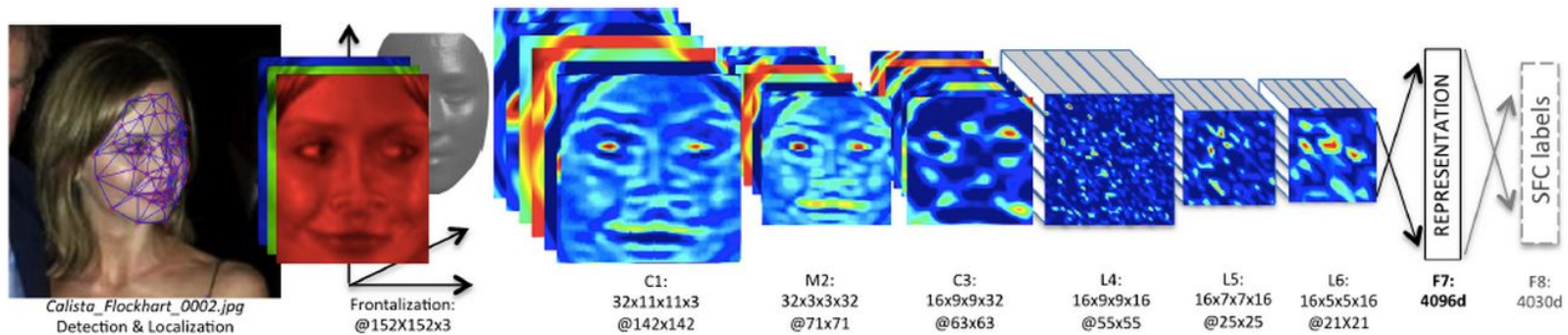
Neural Network - 10 Units, No Weight Decay

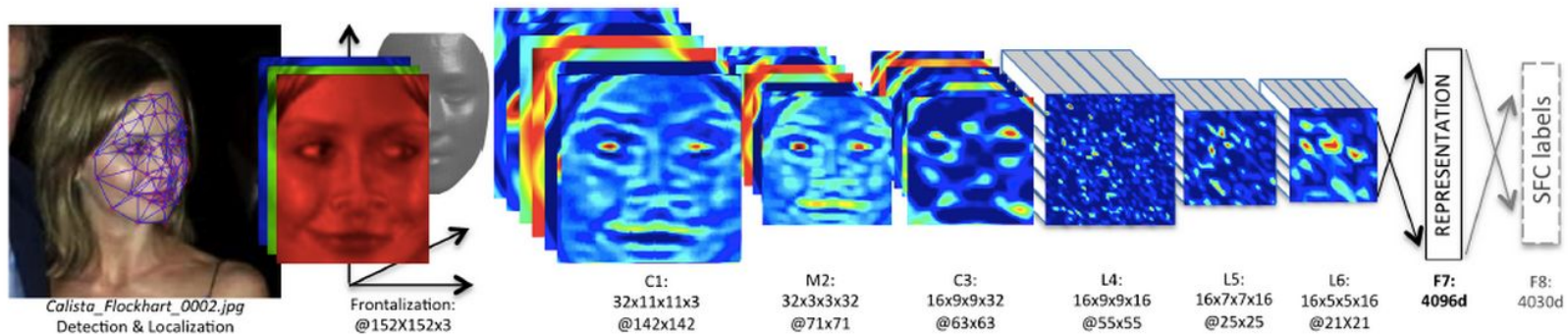


Neural Network - 10 Units, Weight Decay=0.02

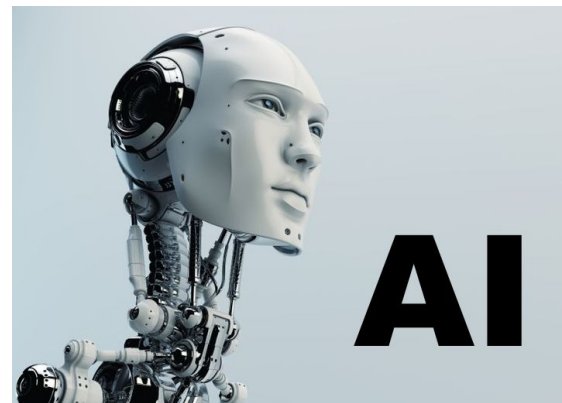
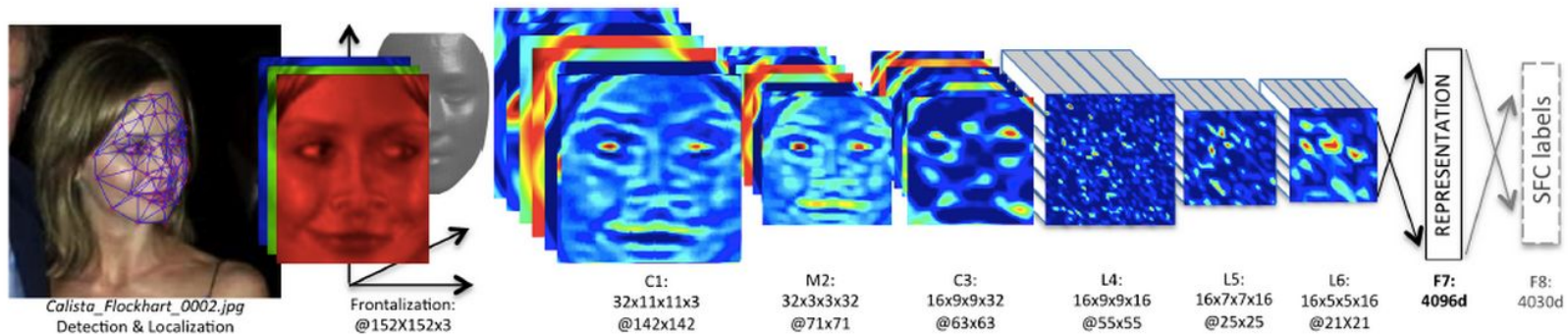












# kaggle

**Machine Learning Challenge Winning Solutions**

- The most frequently used tool by data science competition winners
  - 17 out of 29 winning solutions in kaggle last year used XGBoost
  - Solve wide range of problems: store sales prediction; high energy physics event classification; web text classification; customer behavior prediction; motion detection; ad click through rate prediction; malware classification; product categorization; hazard risk prediction; massive online course dropout rate prediction
- Present and Future of KDDCup. Ron Bekkerman (KDDCup 2015 chair): "Something dramatic happened in Machine Learning over the past couple of years. It is called XGBoost – a package implementing Gradient Boosted Decision Trees that works wonders in data classification. Apparently, every winning team used XGBoost, mostly in ensembles with other classifiers. Most surprisingly, the winning teams report very minor improvements that ensembles bring over a single well-configured XGBoost."
- A lot contributions from the kaggle community

56:01 / 1:16:29

**XGBoost A Scalable Tree Boosting System June 02, 2016**

DataScience.LA  
2.6K

5,632 views

+ Add to Share ... More

# What are the top academic backgrounds of data scientists?

## Master's

Rank	% of people	Field
1	12.86%	Computer Science
2	12.49%	Business Administration/ Management
3	10.98%	Statistics
4	10.20%	Mathematics
5	8.54%	<u>Physics</u>
6	5.25%	Machine Learning/ Data Science
7	4.50%	Electrical Engineering
8	4.21%	Economics & Finance
9	2.85%	Computer Engineering
10	2.48%	Biology

## PhD

Rank	% of people	Field
1	14.74%	<u>Physics</u>
2	14.46%	Computer Science
3	10.83	Mathematics
4	8.24%	Statistics
5	4.77%	Electrical Engineering
6	4.08%	Biology
7	4.06%	Machine Learning/Data Science
8	3.25%	Computer Engineering
9	3.09%	Neuroscience
10	2.74%	Economics & Finance

1992- ELTE fizikus

1996-98 Monte Carlo szim., Kosterlitz-Thouless

1999- pénzügyi alkalmazások

1992- ELTE fizikus

1996-98 Monte Carlo szim., Kosterlitz-Thouless

1999- pénzügyi alkalmazások

2001-05 CIB Bank kockázatkezelés

2004 PhD

1992- ELTE fizikus

1996-98 Monte Carlo szim., Kosterlitz-Thouless

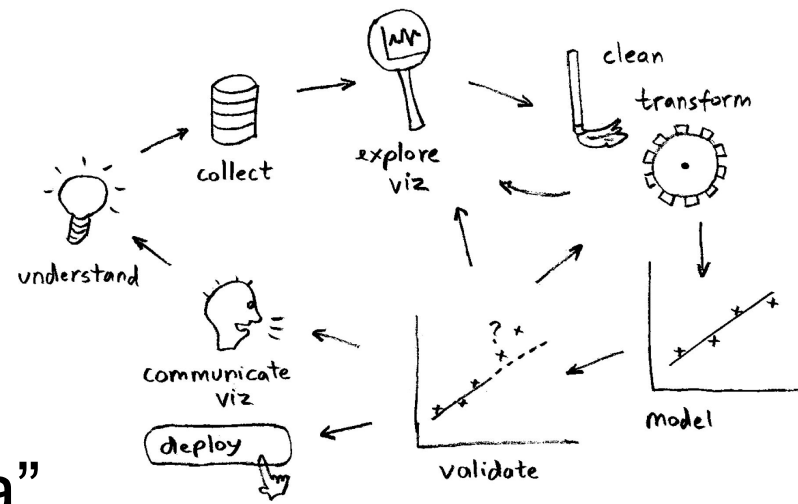
1999- pénzügyi alkalmazások

2001-05 CIB Bank kockázatkezelés

2004 PhD

2006- Kalifornia, data science

2016-17 CEU, UCLA (1-1 kurzus)



numerikus jártasság  
adatok manipulációja, “tisztítása”

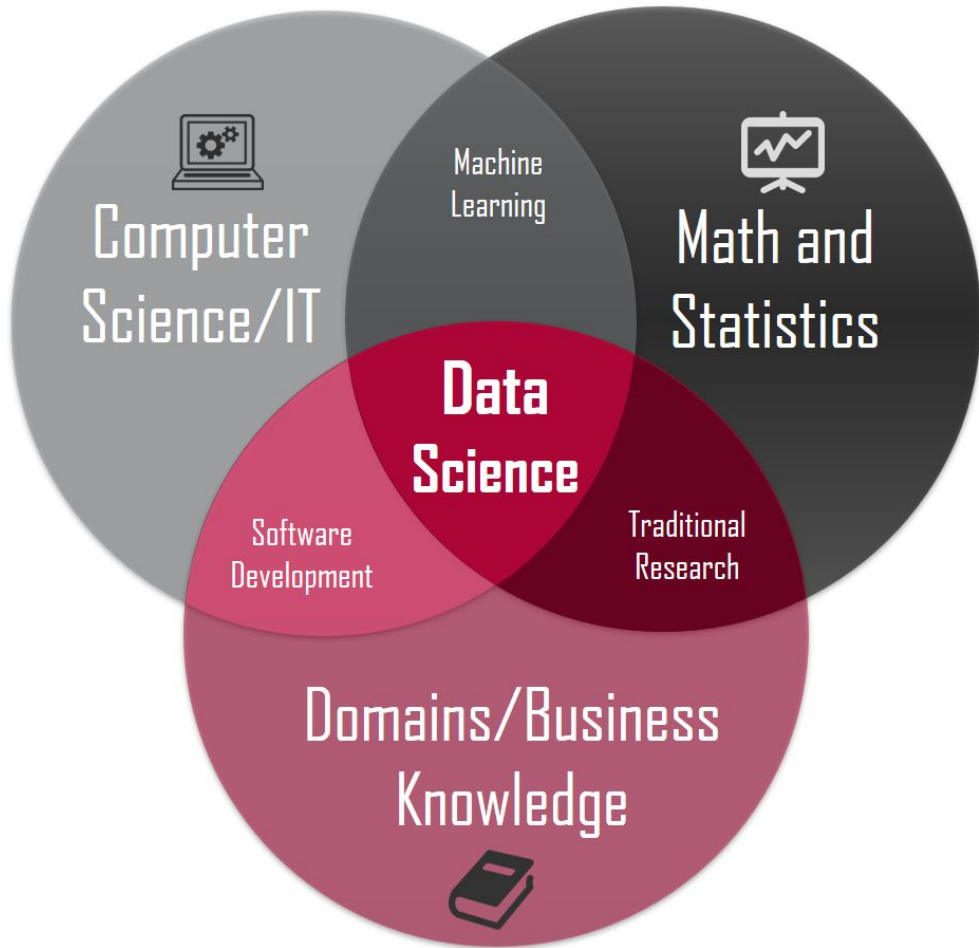
adatvizualizáció

eszközök erre (Unix/Linux, Matlab, R, Python)

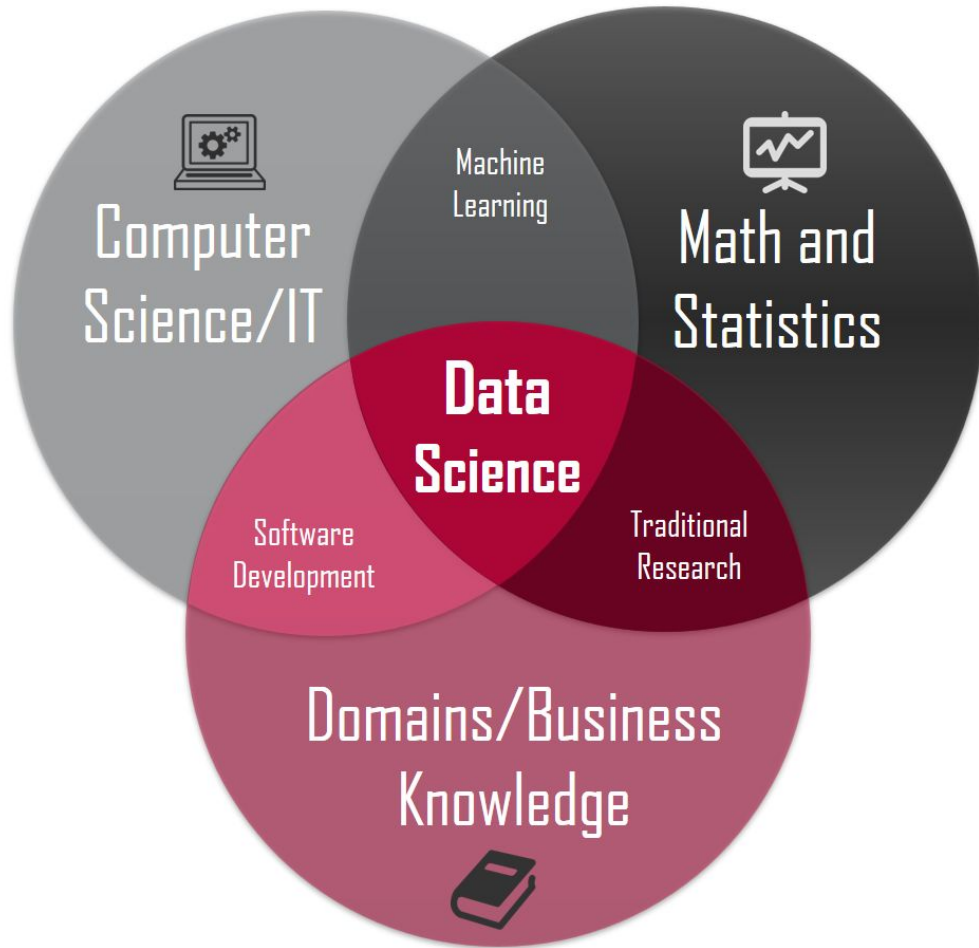
programozás

szimulációk

modellekben való gondolkodás







## Soft skills:

- curiosity
- creativity
- skepticism
- ask good questions
- communication

# Data Scientist: The Sexiest Job of the 21st Century

Harvard  
Business  
Review

By 2018, the US alone could face a shortage of 140,000 to 190,000 people with deep analytical skills - McKinsey, 2011



## Data Science Specialization

Launch Your Career in Data Science. A nine-course introduction to data science, developed and taught by leading professors.

 **GENERAL ASSEMBLY**

# DATA SCIENCE IMMERSIVE

**12-WEEK FULL-TIME CAREER ACCELERATOR IN LOS ANGELES**

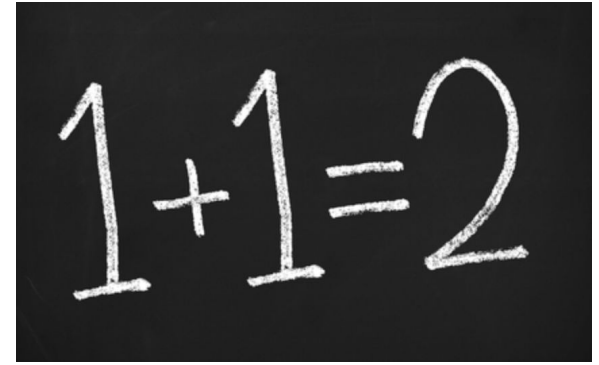
**UCLA** College  
Master of Applied Statistics

**MS in Business Analytics**










**I**



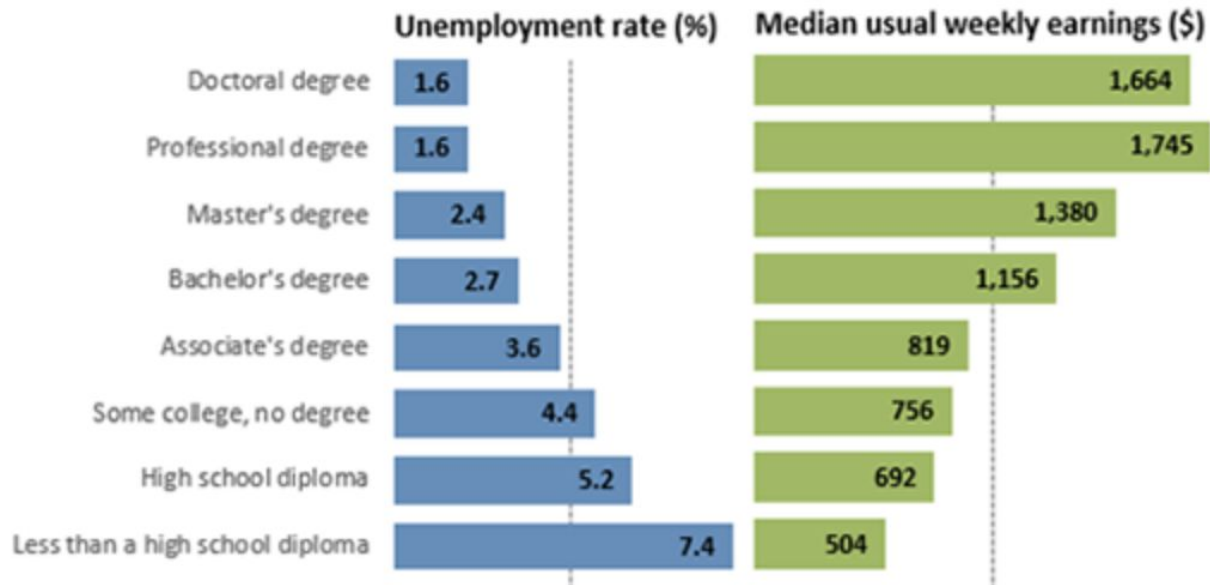
**PHYSICS**

**I**   
**PHYSICS**

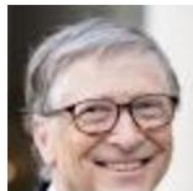




# I ❤️ PHYSICS



# I ❤️ PHYSICS



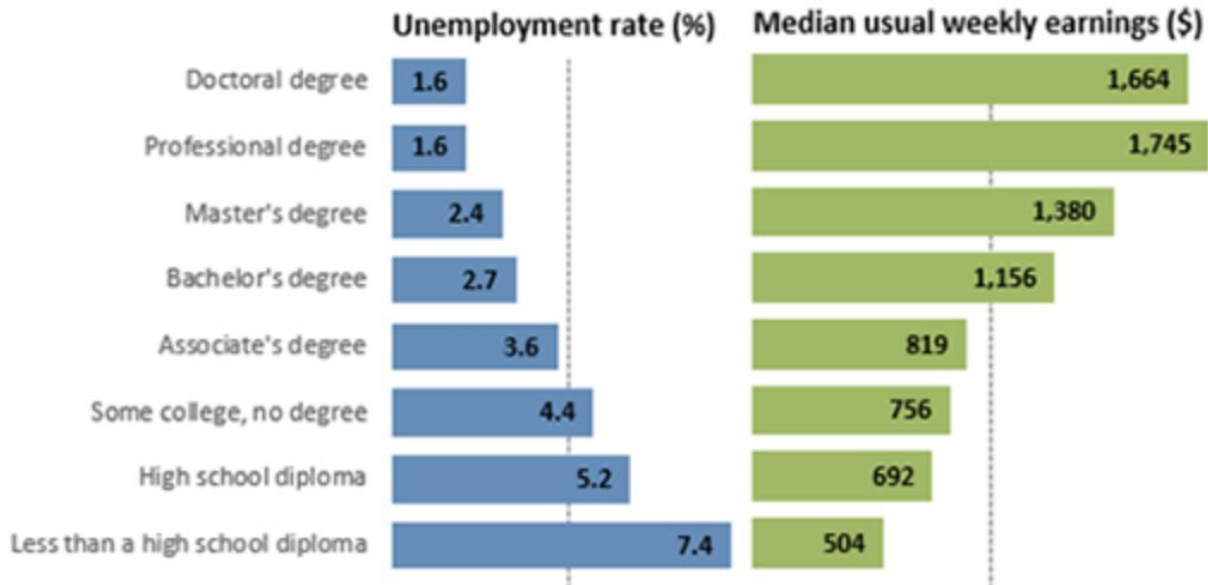
Bill Gates



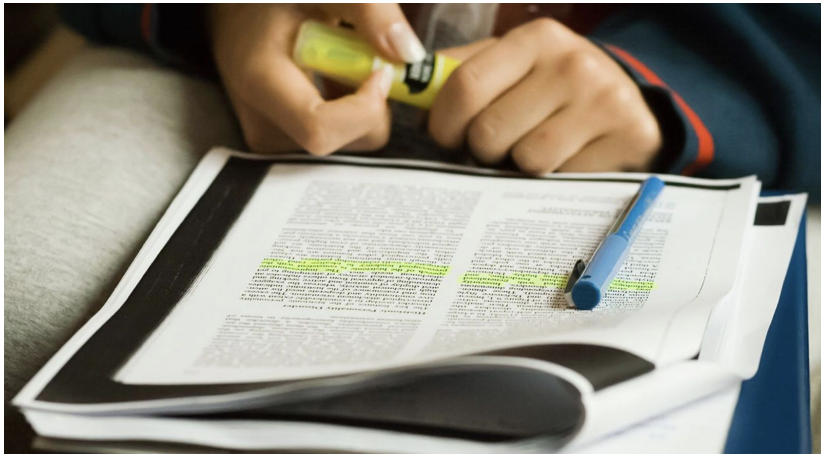
Steve Jobs



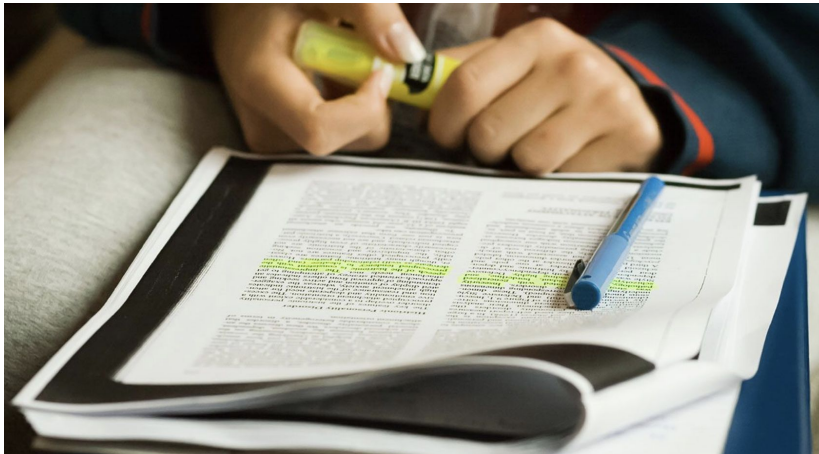
Mark  
Zuckerberg



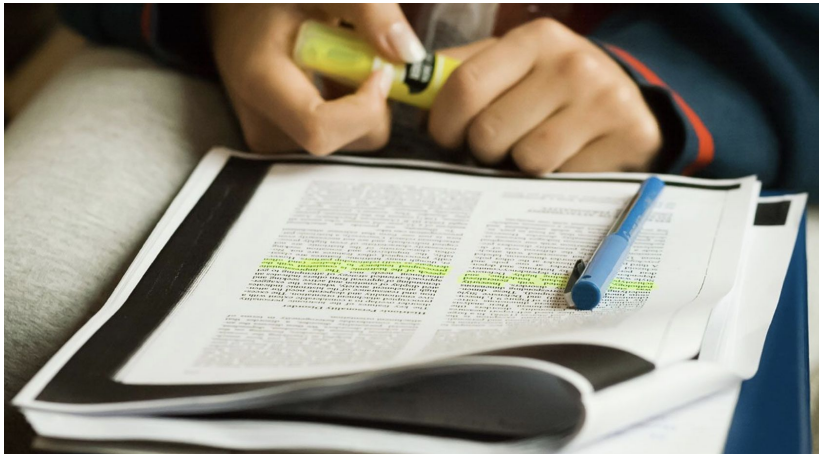




In  
English  
Please.



In  
English  
Please.



**MATH**

$\cos \frac{A}{2} = \sqrt{\frac{1 + \cos A}{2}}$     $\frac{a}{\sin A} = \frac{b}{\sin B} = \frac{c}{\sin C} = 2R$     $x^2 - a^2 = (x+a)(x-a)$     $\operatorname{arccoth}(z) = \frac{1}{2} \ln \frac{z+1}{z-1}$     $\sqrt{A} = \sqrt{r} \cdot \exp f(\log r) - f(\phi)$

$\sin(x) = \frac{e^{ix} - e^{-ix}}{2i}$     $\sinh(x) = \frac{e^x - e^{-x}}{2}$     $\cosh(x) = \frac{e^x + e^{-x}}{2}$     $\operatorname{sech}(x) = \frac{2}{e^x + e^{-x}}$     $\operatorname{csch}(x) = \frac{2}{e^x - e^{-x}}$

$\log_m n = \frac{\log_n m}{\log_n n}$     $\operatorname{arccsch}(z) = \ln \left( 1 + \sqrt{1+z^2} \right) / z$     $\operatorname{arcsch}(z) = \ln \left( \frac{1}{z} + \sqrt{1+z^2} \right)$

$\cos(-x) = \cos(x)$     $\sec(-x) = \sec(x)$     $\tan(-x) = -\tan(x)$

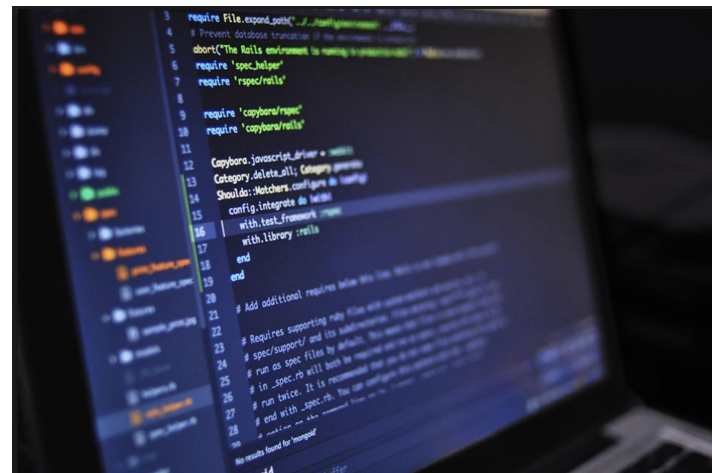
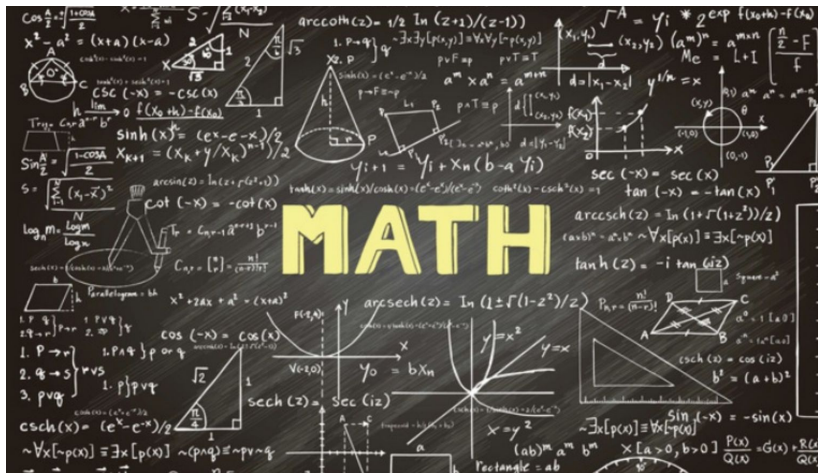
$\operatorname{csch}(x) = \frac{1}{\sinh(x)}$     $\operatorname{csch}(iz) = i \operatorname{csch}(x)$     $\operatorname{csch}(x) = \frac{1}{\sinh(x)}$     $\operatorname{csch}(iz) = \frac{1}{\sinh(iz)}$

$\operatorname{sech}(z) = \frac{2}{e^z + e^{-z}}$     $\operatorname{Sec}(iz) = \frac{1}{\cos(iz)}$     $\operatorname{csch}(x) = \frac{1}{\sinh(x)}$     $\operatorname{csch}(iz) = \frac{1}{\sinh(iz)}$

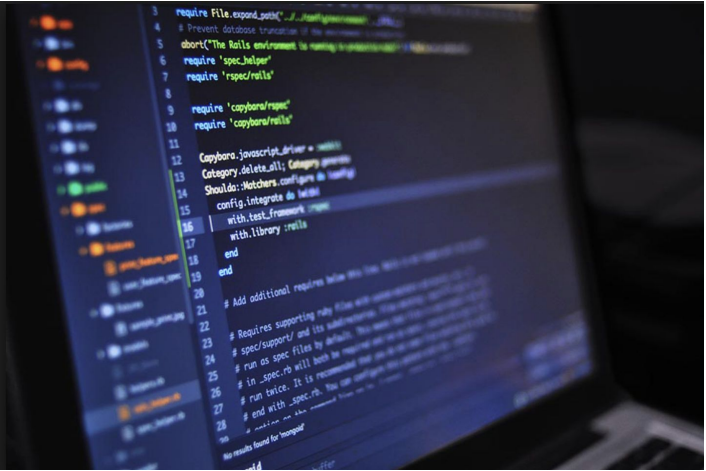
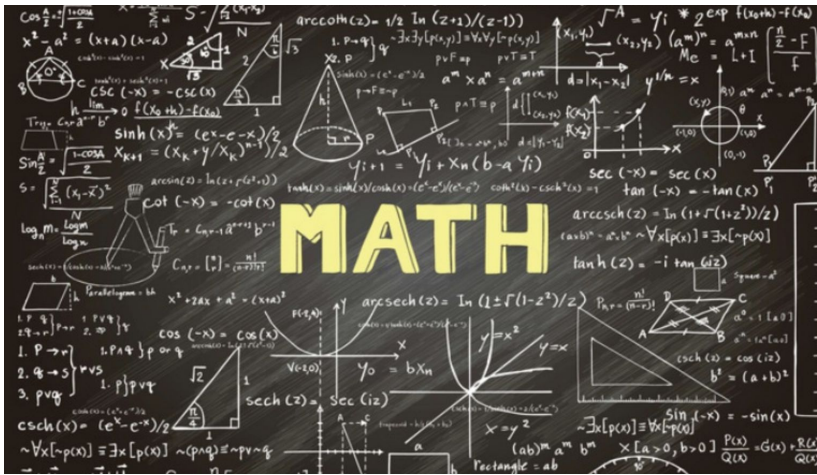
$\operatorname{csch}(x) = \frac{1}{\sinh(x)}$     $\operatorname{csch}(iz) = \frac{1}{\sinh(iz)}$     $\operatorname{csch}(x) = \frac{1}{\sinh(x)}$     $\operatorname{csch}(iz) = \frac{1}{\sinh(iz)}$

$\operatorname{csch}(x) = \frac{1}{\sinh(x)}$     $\operatorname{csch}(iz) = \frac{1}{\sinh(iz)}$     $\operatorname{csch}(x) = \frac{1}{\sinh(x)}$     $\operatorname{csch}(iz) = \frac{1}{\sinh(iz)}$

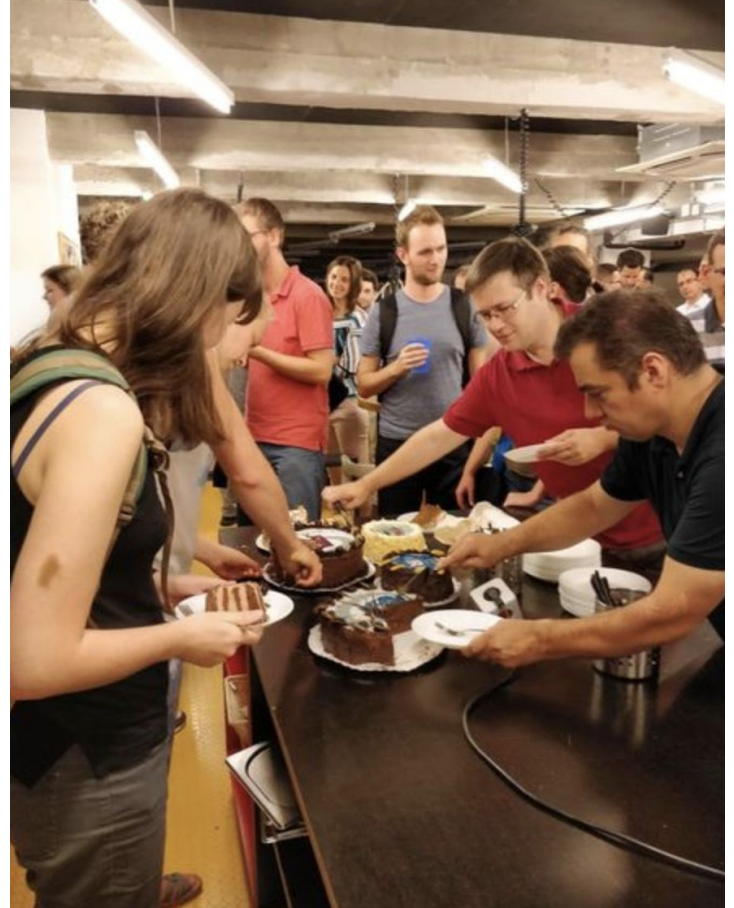
$\operatorname{csch}(x) = \frac{1}{\sinh(x)}$     $\operatorname{csch}(iz) = \frac{1}{\sinh(iz)}$     $\operatorname{csch}(x) = \frac{1}{\sinh(x)}$     $\operatorname{csch}(iz) = \frac{1}{\sinh(iz)}$











INTROVERT



EXTROVERT





---

## Santa Monica, CA

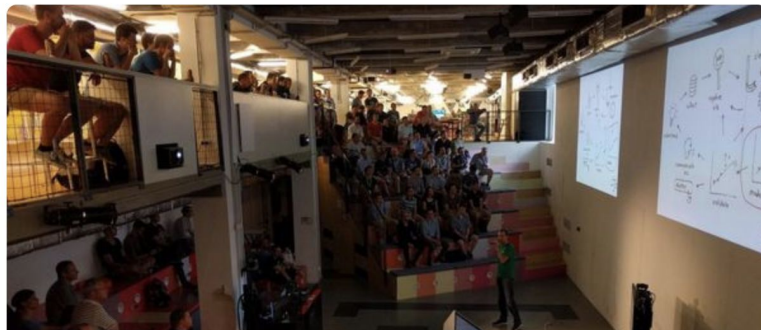
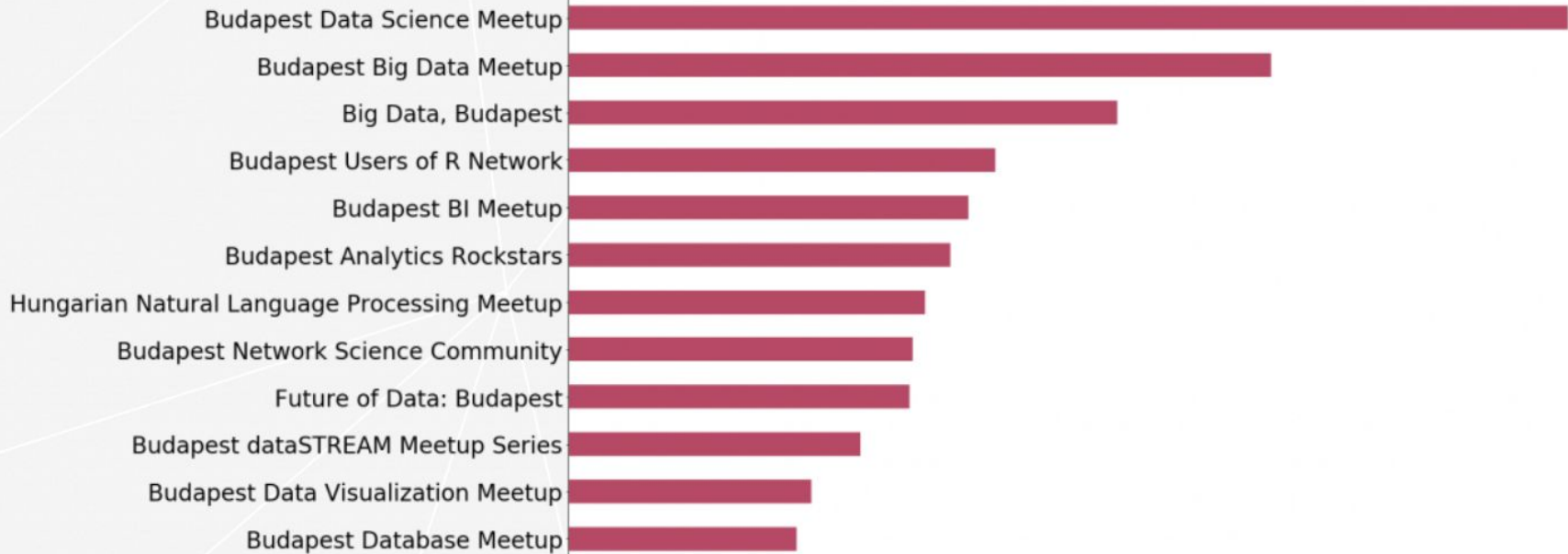
Founded Mar 20, 2009

---

R users

1,347





## Budapest Data Science Meetup

Location

**Budapest, Hungary**

Members

**2,915**



Organizers

**Zoltan C. T. and 5 others**



Search



Home



My Network



Jobs



**Szilard Pafka** • 1st

Chief Data Scientist

Santa Monica, California

Message

More...



EPOCH



Eötvös University,  
Budapest (ELTE)



See contact info



500+ connections



**Farhan Baluch**

Senior Data Scientist at Netflix

Connected 5 months ago

Message



**Thomas W. Dinsmore**

Driving competitive intelligence at DataRobot. Recruiters: I'm not in the market.

Connected 5 months ago

Message





Search



Home



My Network



Jobs



**Szilard Pafka** • 1st

Chief Data Scientist

Santa Monica, California

Message

More...



EPOCH



Eötvös University,  
Budapest (ELTE)



See contact info



500+ connections



**Farhan Baluch**

Senior Data Scientist at Netflix

Connected 5 months ago

Message



**Thomas W. Dinsmore**

Driving competitive intelligence at DataRobot. Recruiters: I'm not in the market.

Connected 5 months ago

Message





## Mikel Bober-Irizar

[Follow](#)

Hey, I'm a kid doing data science. <https://mxbi.net>

Editor of Imploding Gradients

55 Following 705 Followers · 



# Mikel Bober-Irizar

[Follow](#)

Hey, I'm a kid doing data science. <https://mxbi.net>

Editor of Imploding Gradients

55 Following 705 Followers · 



## How I got 3rd place in The Ultimate Student Hunt

*The Ultimate Student Hunt was a week-long machine learning competition hosted by AnalyticsVidhya, in which I came solo 3rd place. Here's a blog post going in-depth into my solution and thought process.*

### Step 1: Identify the problem

The moment the competition started, the first thing I did was click download on the data. As it was downloading, I had a quick look through the problem

# Összefoglaló:

1. Adattudomány: példák, 1 projekt röviden (eszközök, szükséges tudás)
2. Miért fizikusok?
3. Pár karriertanács