XGBoost

dataworkshop.eu

Alekseichenko Vladimir

Disclaimer

Data Workshop *[all time]* focuses on the **intuition** and **practical** tips.

For a formal treatment, see something else.*

* papers or classical machine learning books

Workshop Program

Wednsday, September 29

08:45 – 09:00 ○ **Opening**

09:00 – 10:30 ● **Workshop - part I**

Vladimir Alekseichenko

Introduction
Understand business and data
Overivew: decision tree, boosting, gradient boosting

10:30 – 11:00 ● **Coffee break**

11:00 – 12:30 ● **Workshop - part II**

Vladimir Alekseichenko

Simple model
Feature enginnering
Incremental learning

12:15 – 12:30 ○ **Summary**

Motivation

why boosting?

Motivation

XGboost is utilized

in many **winning** solutions.

XGBoost

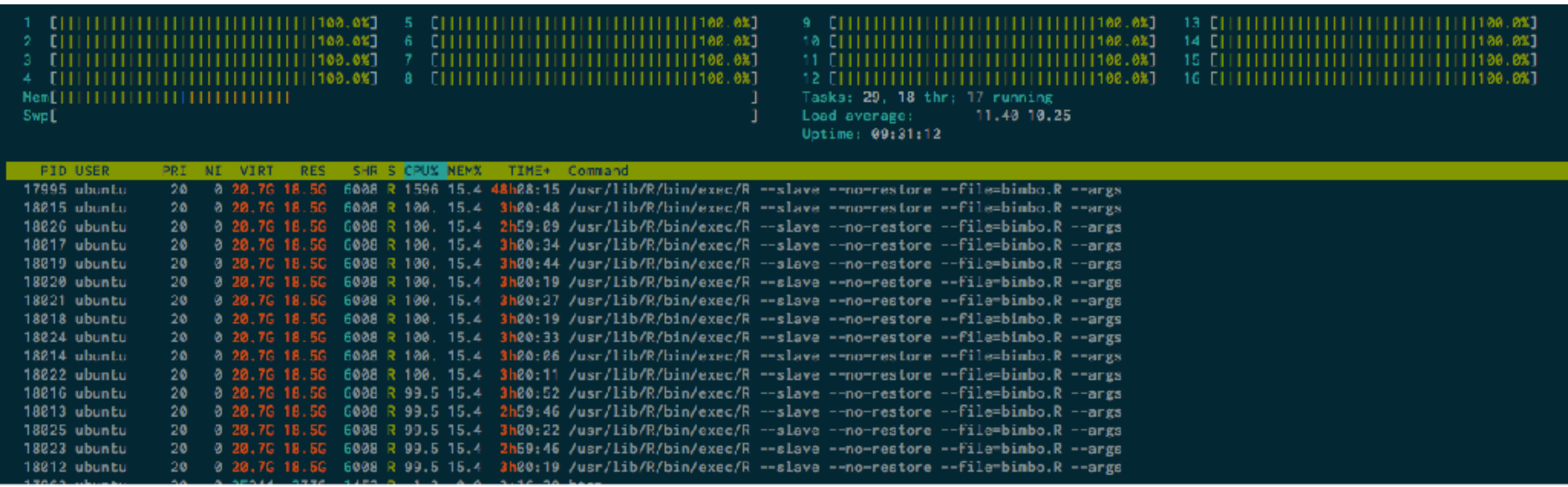
(eXtreme Gradient Boosting)

"When in doubt, use xgboost"

Owen Zhang

OWEN'S PROFILE ON KAGGLE

16 cores



1 [||||||||||||||||||||||||||||||100.0%]
2 [||||||||||||||||||||||||||||||100.0%]
3 [||||||||||||||||||||||||||||||100.0%]
4 [||||||||||||||||||||||||||||||100.0%]
5 [||||||||||||||||||||||||||||||100.0%]
6 [||||||||||||||||||||||||||||||100.0%]
7 [||||||||||||||||||||||||||||||100.0%]
8 [||||||||||||||||||||||||||||||100.0%]
9 [||||||||||||||||||||||||||||||100.0%]
10 [||||||||||||||||||||||||||||||100.0%]
11 [||||||||||||||||||||||||||||||100.0%]
12 [||||||||||||||||||||||||||||||100.0%]
13 [||||||||||||||||||||||||||||||100.0%]
14 [||||||||||||||||||||||||||||||100.0%]
15 [||||||||||||||||||||||||||||||100.0%]
16 [||||||||||||||||||||||||||||||100.0%]
Mem[||||||||||||||||||||||||||]
Swp[ ]
Tasks: 29, 18 thr; 17 running
Load average: 11.40 10.25
Uptime: 09:31:12
PID USER PRI NI VIRT RES SHR S CPU% MEM% TIME+ Command
17995 ubuntu 20 0 20.7G 18.5G 6008 R 1596 15.4 48h28:15 /usr/lib/R/bin/exec/R --slave --no-restore --file=bimbo.R --args
18015 ubuntu 20 0 20.7G 18.5G 6008 R 100. 15.4 3h20:48 /usr/lib/R/bin/exec/R --slave --no-restore --file=bimbo.R --args
18026 ubuntu 20 0 20.7G 18.5G 6008 R 100. 15.4 2h59:09 /usr/lib/R/bin/exec/R --slave --no-restore --file=bimbo.R --args
18017 ubuntu 20 0 20.7G 18.5G 6008 R 100. 15.4 3h20:34 /usr/lib/R/bin/exec/R --slave --no-restore --file=bimbo.R --args
18019 ubuntu 20 0 20.7G 18.5G 6008 R 100. 15.4 3h20:44 /usr/lib/R/bin/exec/R --slave --no-restore --file=bimbo.R --args
18020 ubuntu 20 0 20.7G 18.5G 6008 R 100. 15.4 3h20:19 /usr/lib/R/bin/exec/R --slave --no-restore --file=bimbo.R --args
18021 ubuntu 20 0 20.7G 18.5G 6008 R 100. 15.4 3h20:27 /usr/lib/R/bin/exec/R --slave --no-restore --file=bimbo.R --args
18018 ubuntu 20 0 20.7G 18.5G 6008 R 100. 15.4 3h20:19 /usr/lib/R/bin/exec/R --slave --no-restore --file=bimbo.R --args
18024 ubuntu 20 0 20.7G 18.5G 6008 R 100. 15.4 3h20:33 /usr/lib/R/bin/exec/R --slave --no-restore --file=bimbo.R --args
18014 ubuntu 20 0 20.7G 18.5G 6008 R 100. 15.4 3h20:26 /usr/lib/R/bin/exec/R --slave --no-restore --file=bimbo.R --args
18022 ubuntu 20 0 20.7G 18.5G 6008 R 100. 15.4 3h20:11 /usr/lib/R/bin/exec/R --slave --no-restore --file=bimbo.R --args
18016 ubuntu 20 0 20.7G 18.5G 6008 R 99.5 15.4 3h20:52 /usr/lib/R/bin/exec/R --slave --no-restore --file=bimbo.R --args
18013 ubuntu 20 0 20.7G 18.5G 6008 R 99.5 15.4 2h59:46 /usr/lib/R/bin/exec/R --slave --no-restore --file=bimbo.R --args
18025 ubuntu 20 0 20.7G 18.5G 6008 R 99.5 15.4 3h20:22 /usr/lib/R/bin/exec/R --slave --no-restore --file=bimbo.R --args
18023 ubuntu 20 0 20.7G 18.5G 6008 R 99.5 15.4 2h59:46 /usr/lib/R/bin/exec/R --slave --no-restore --file=bimbo.R --args
18012 ubuntu 20 0 20.7G 18.5G 6008 R 99.5 15.4 3h20:19 /usr/lib/R/bin/exec/R --slave --no-restore --file=bimbo.R --args

Are you ready?

github.com/dataworkshop/whyr_xgboost

Allstate®
You're in good hands.

kaggle™

Claims Severity

kaggle.com/c/allstate-claims-severity

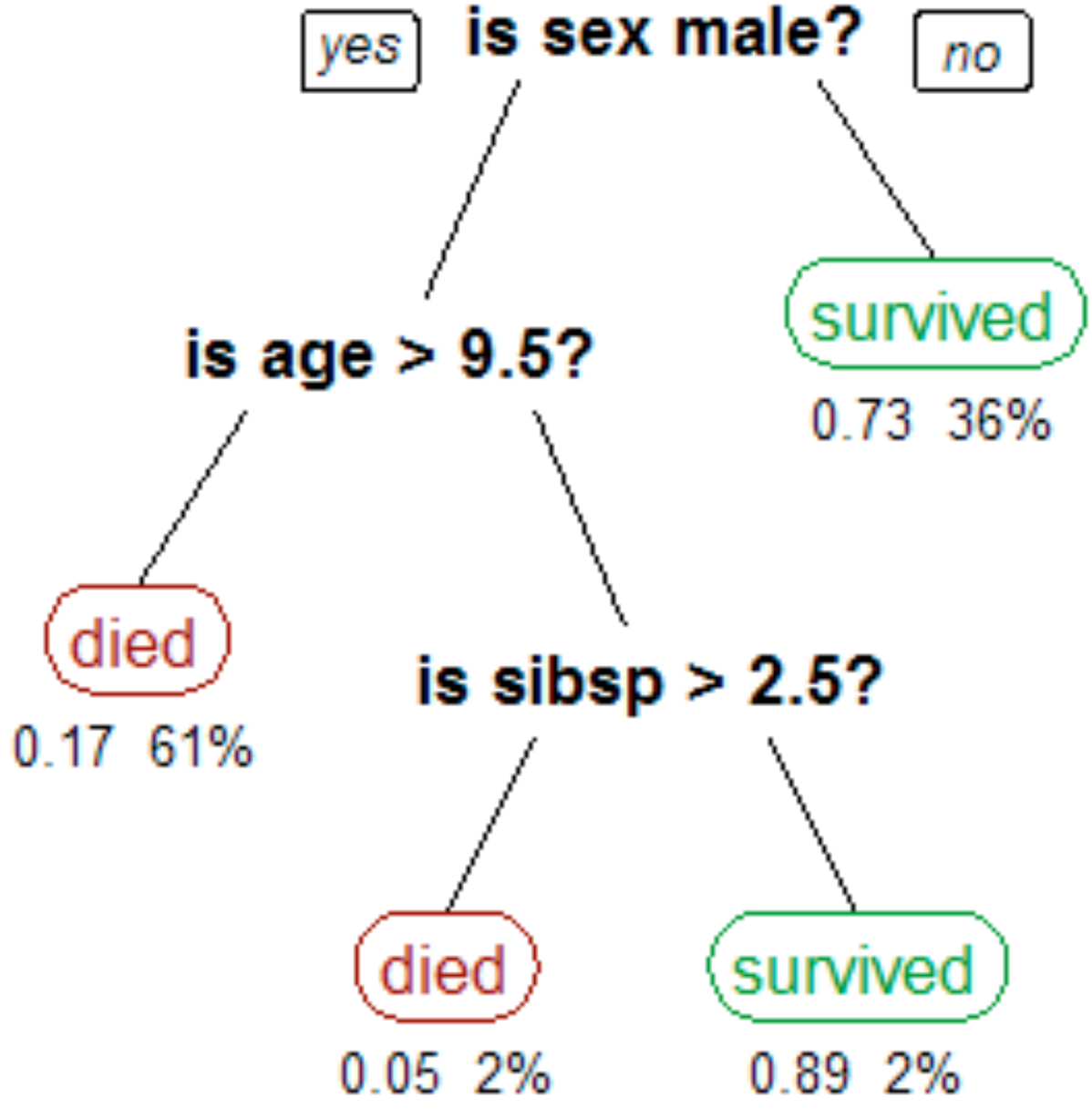
```
vova at va in ~/src/github/whyr_xgboost (master
$ tree .
.
├── LICENSE
├── README.md
├── input
│   ├── sample_submission.csv
│   ├── test.csv
│   └── train.csv
├── models
├── output
└── src
    ├── advanced_model.ipynb
    └── simple_xgboost.ipynb

4 directories, 7 files
```

Understand business and data

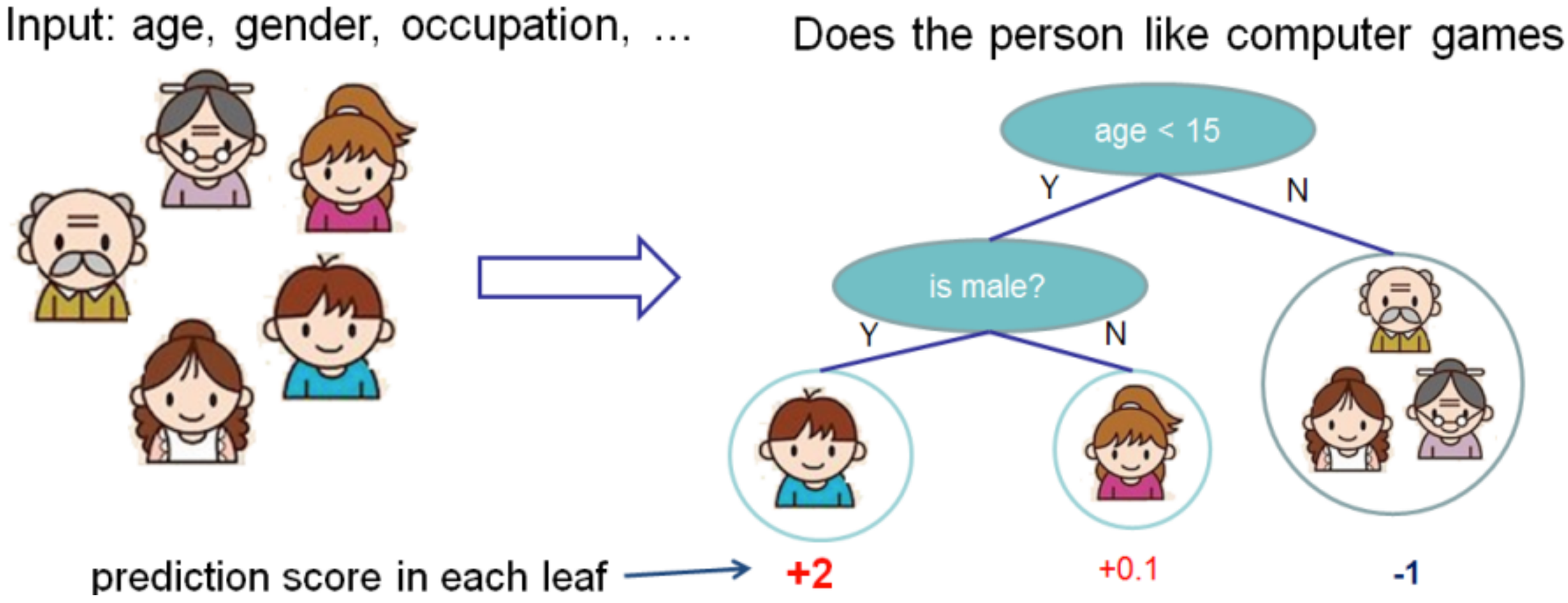
Decision trees

A bit history, why DT is not enough?

Decision Trees (CART)



Decision Tree



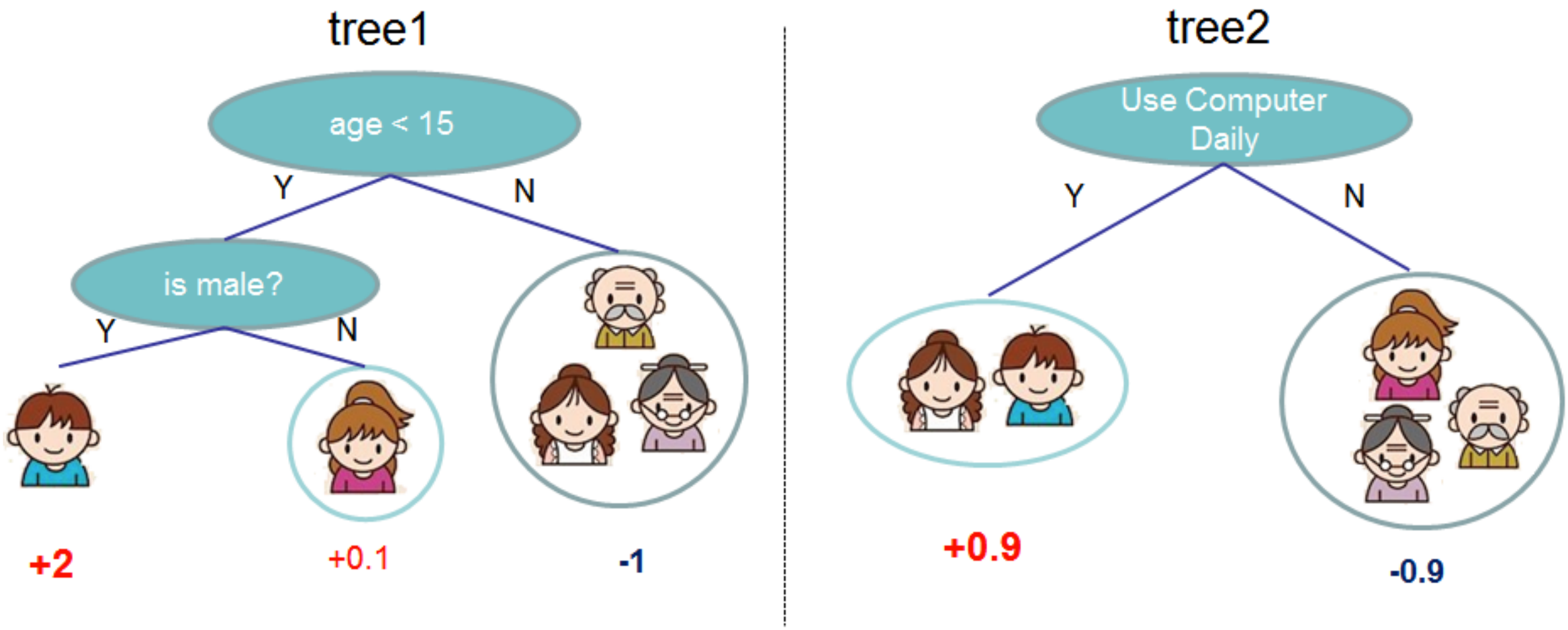
http://xgboost.readthedocs.io/en/latest/model.html

Decision Trees (CART)

pros - interpretable

cons - relatively a poor quality

Ensemble trees



f( ) = 2 + 0.9= 2.9 f( )= -1 - 0.9 = -1.9

http://xgboost.readthedocs.io/en/latest/model.html

Ensembles

Bagging, Random Forest, Bootstrap…

Ensembles

Two heads are better than one

Co dwie głowy, to nie jedna

Одна голова хорошо, а две лучше

Boosting

grant powers to machine learning

Boosting (what)

refers to a family of algorithms which converts **weak learner** *(aka base learner)* to **strong learners**

Boosting (how)

- Using average/ weighted average
- Considering prediction has higher vote

XGBoost

eXtreme Gradient Boosting
or **regularized** gradient boosting

XGBoost is used for **supervised** learning problems

- classification
- regression
- ranking

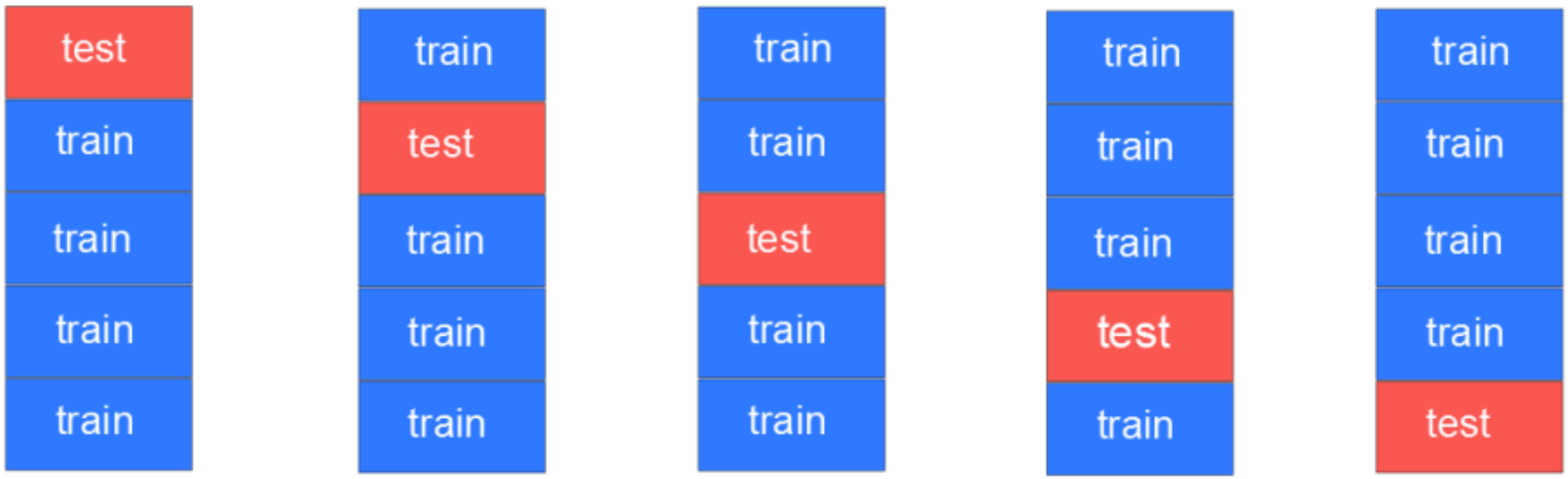
Tianqi Chen

the author of xgboost

Cross validation

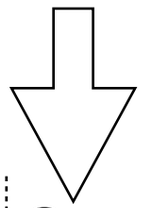
k-fold, leave-one-out, stratified

5-fold cross-validation



Cross-Validation: 3-fold

Age	Gender	Height
20	male	170
30	female	163
24	male	198
28	female	169



Age	Gender	Height
20	male	170
30	female	163
24	male	198

Age	Gender	Height
28	female	169
20	male	170
30	female	163

Age	Gender	Height
24	male	198
28	female	169
20	male	170

Age	Gender	Height
28	female	169

Age	Gender	Height
24	male	198

Age	Gender	Height
30	female	163

Feature Engineering

help a model to do better a job

Feature Engineering

- **Continuous =>** from 1 to 100…
- **dates =>** day, month, year, hour, is weekend…
- **categorical** (red, green, white)
  - assign an unique ID (1, 2, 3)
  - create n-binary columns (is red? etc)
  - probability with target variable (if red 20% then…)

Feature Selection

keep only relevant ones

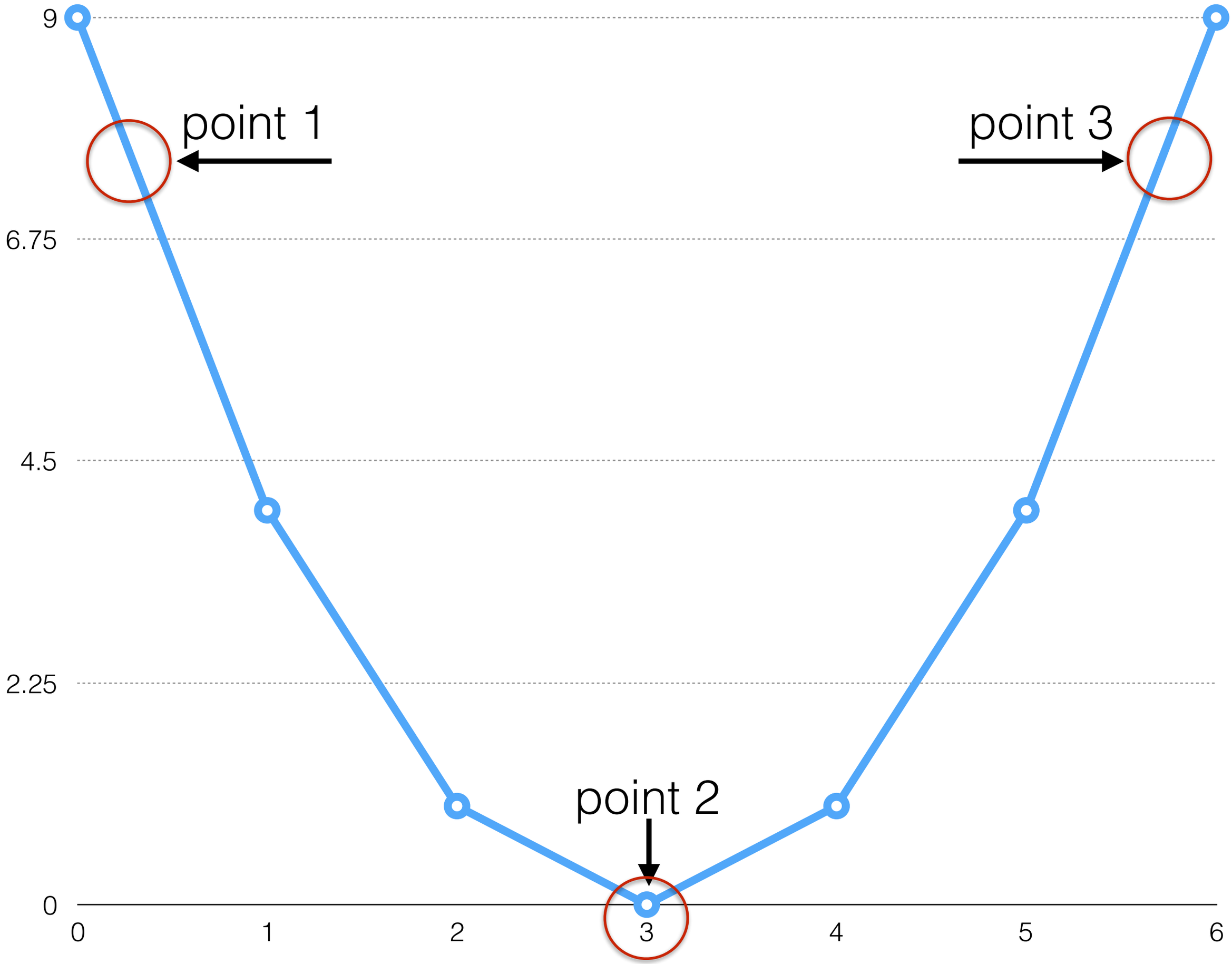
Feature selection

- Less means better (a simple model)
- Keep only the most valuable features (the ideal case - one :)
- Features (usually) depends on each other, be careful… (verify in an empiric way)
- Fewer features - faster solution

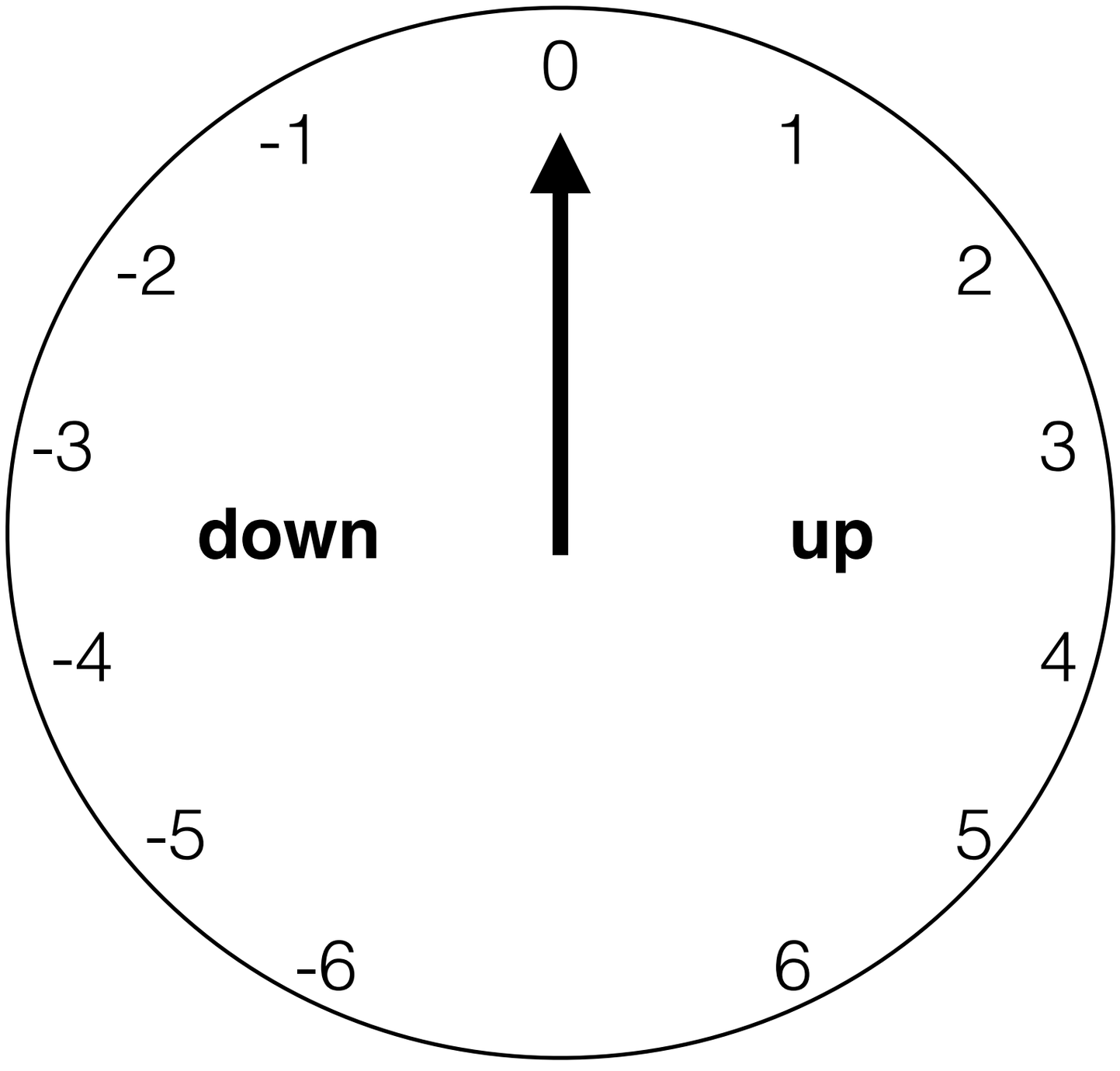
Derivative

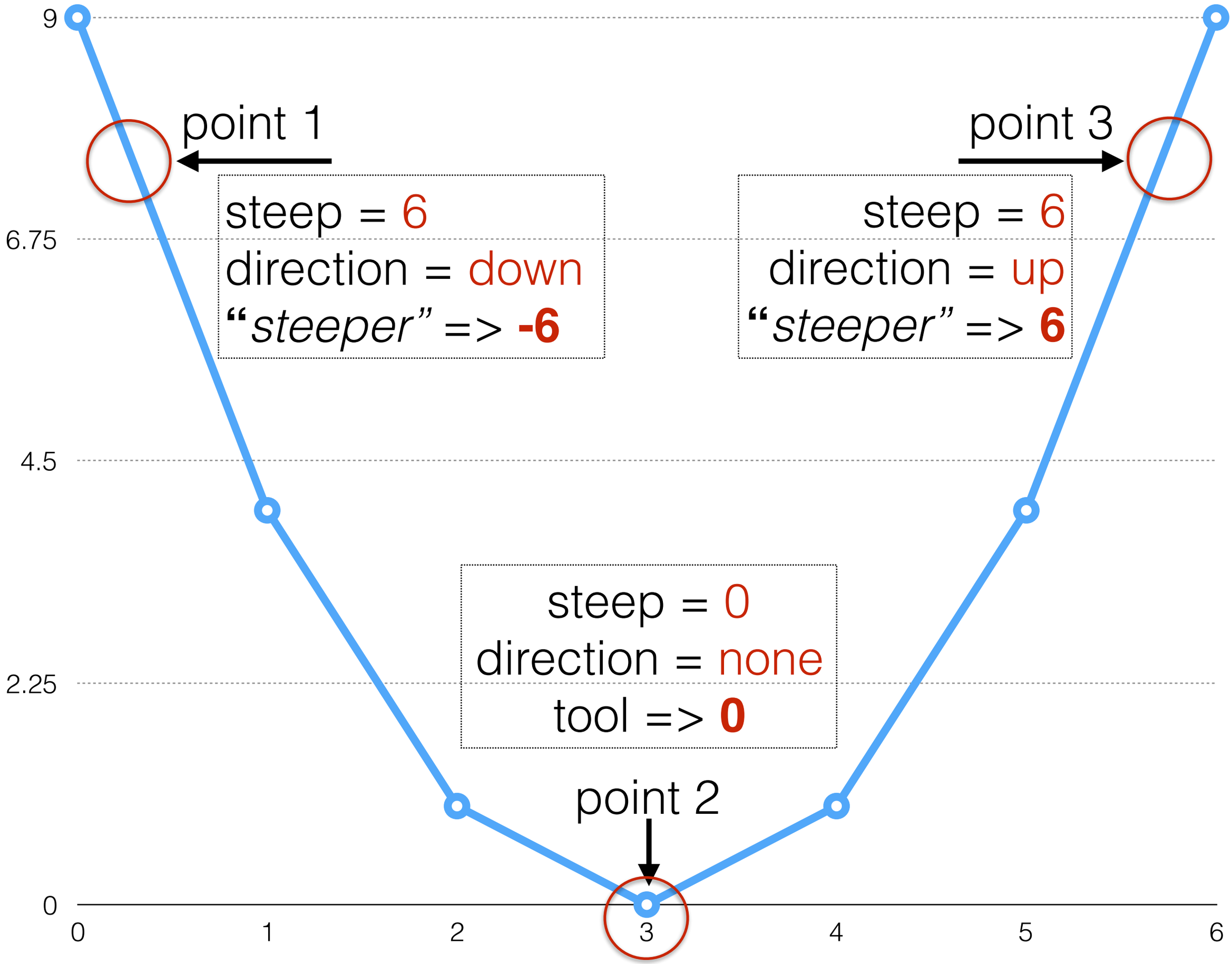
Intuitive example(s)

How **steep** and in which **direction** is **changing height** while car is going to forward?


9
6.75
4.5
2.25
0
0
1
2
3
4
5
6
point 1
point 2
point 3

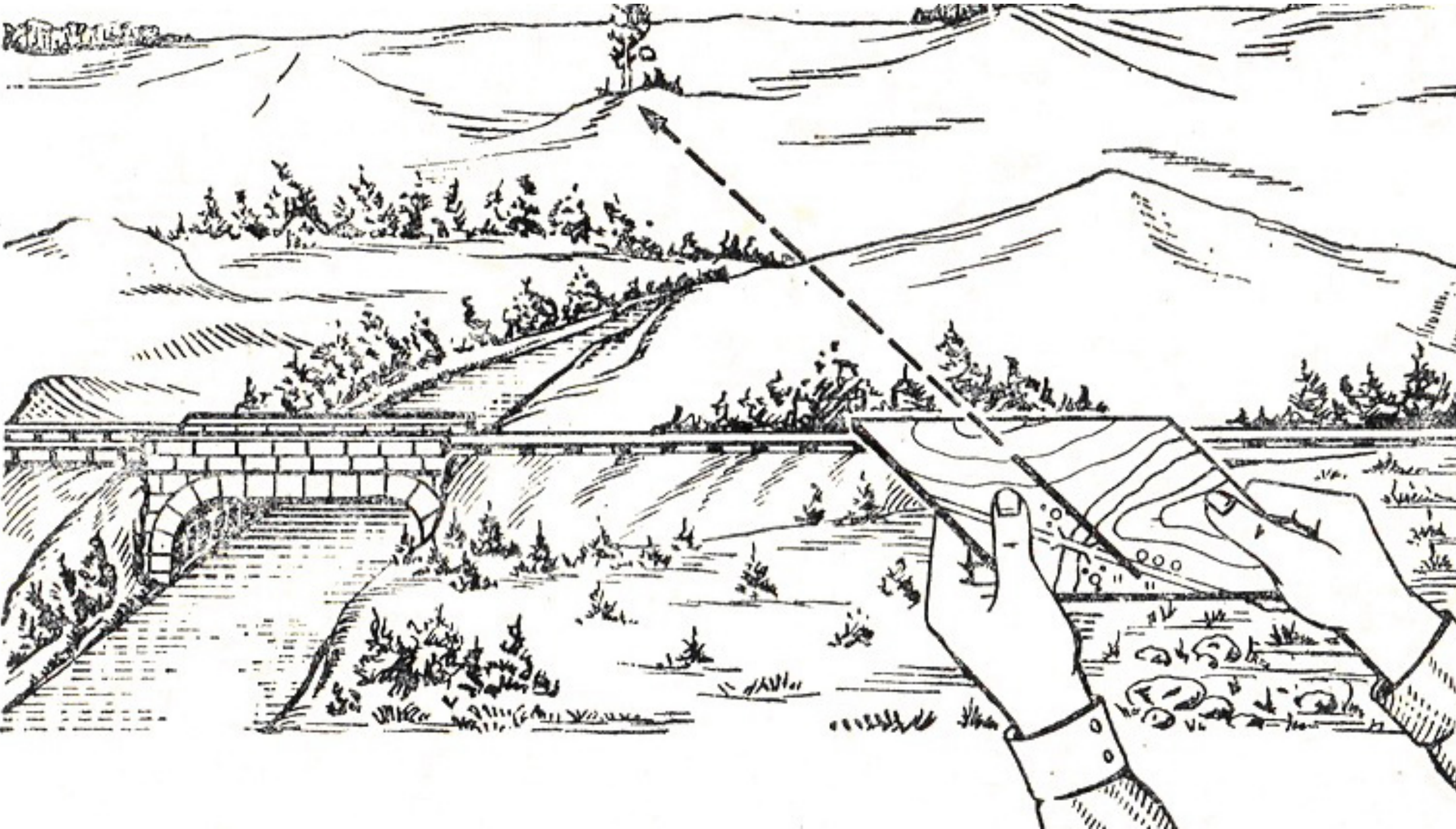
A tool (*steeper*) for measure **steep** and **direction**



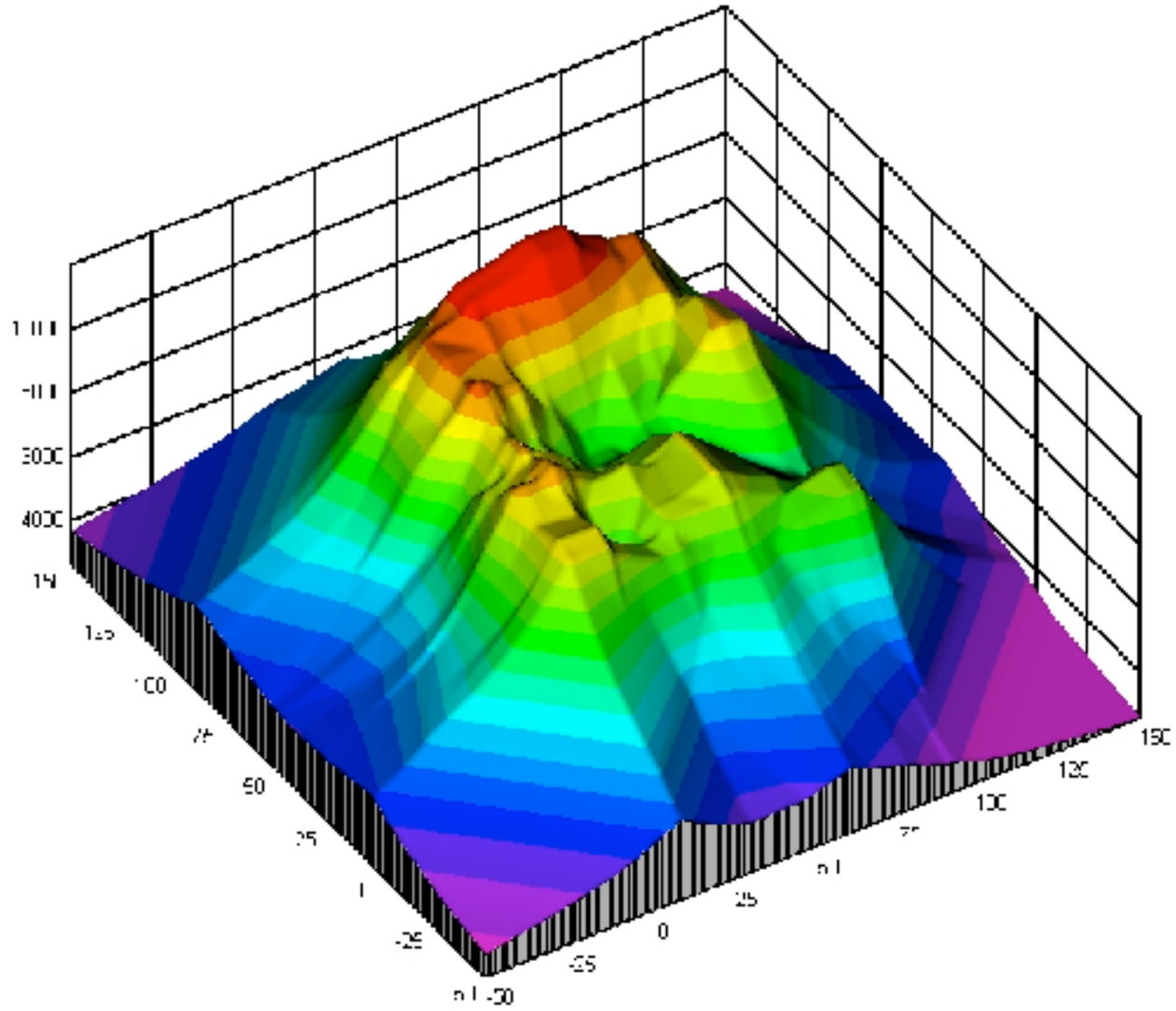

9
6.75
4.5
2.25
0
0
1
2
3
4
5
6
point 1
steep = 6
direction = down
“steeper” => -6
point 3
steep = 6
direction = up
“steeper” => 6
steep = 0
direction = none
tool => 0
point 2

Gradient Descent

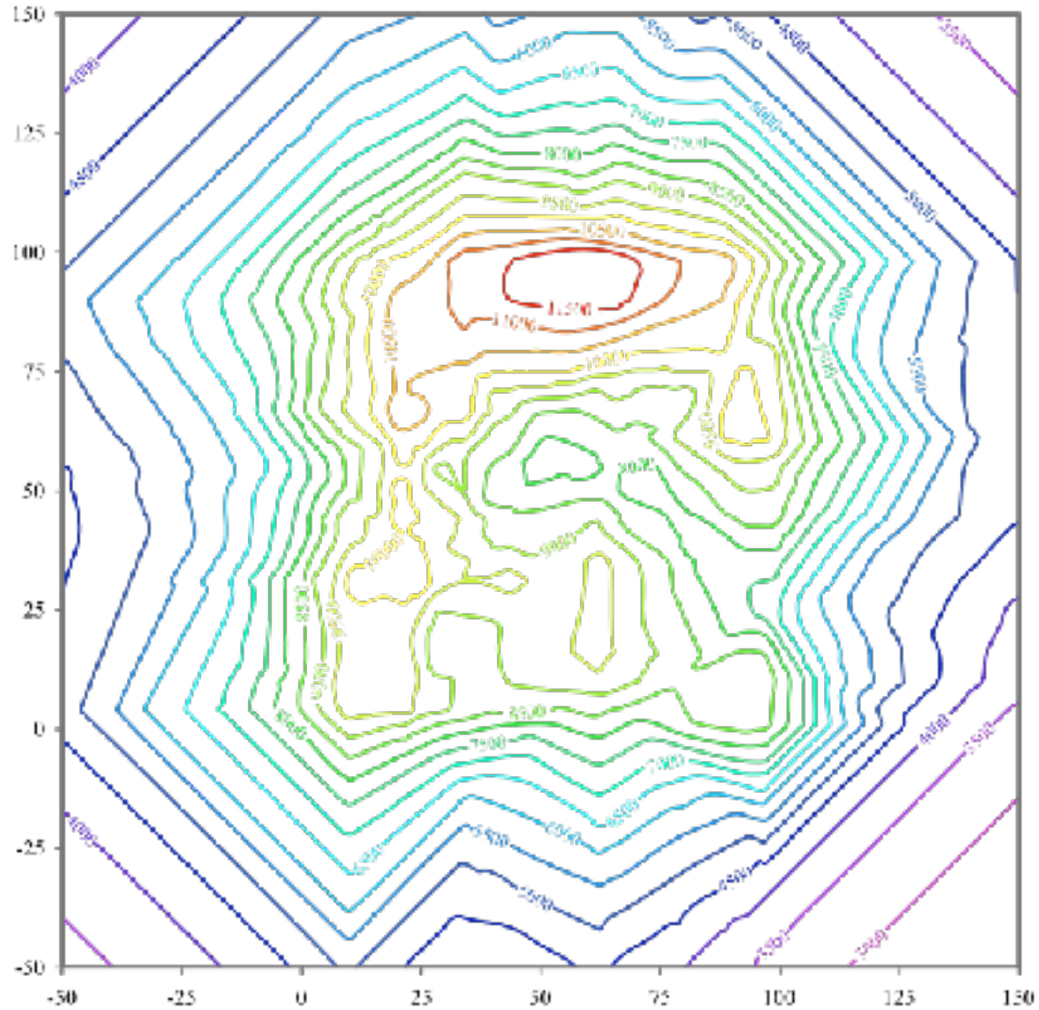
Contour Line



Contour Line

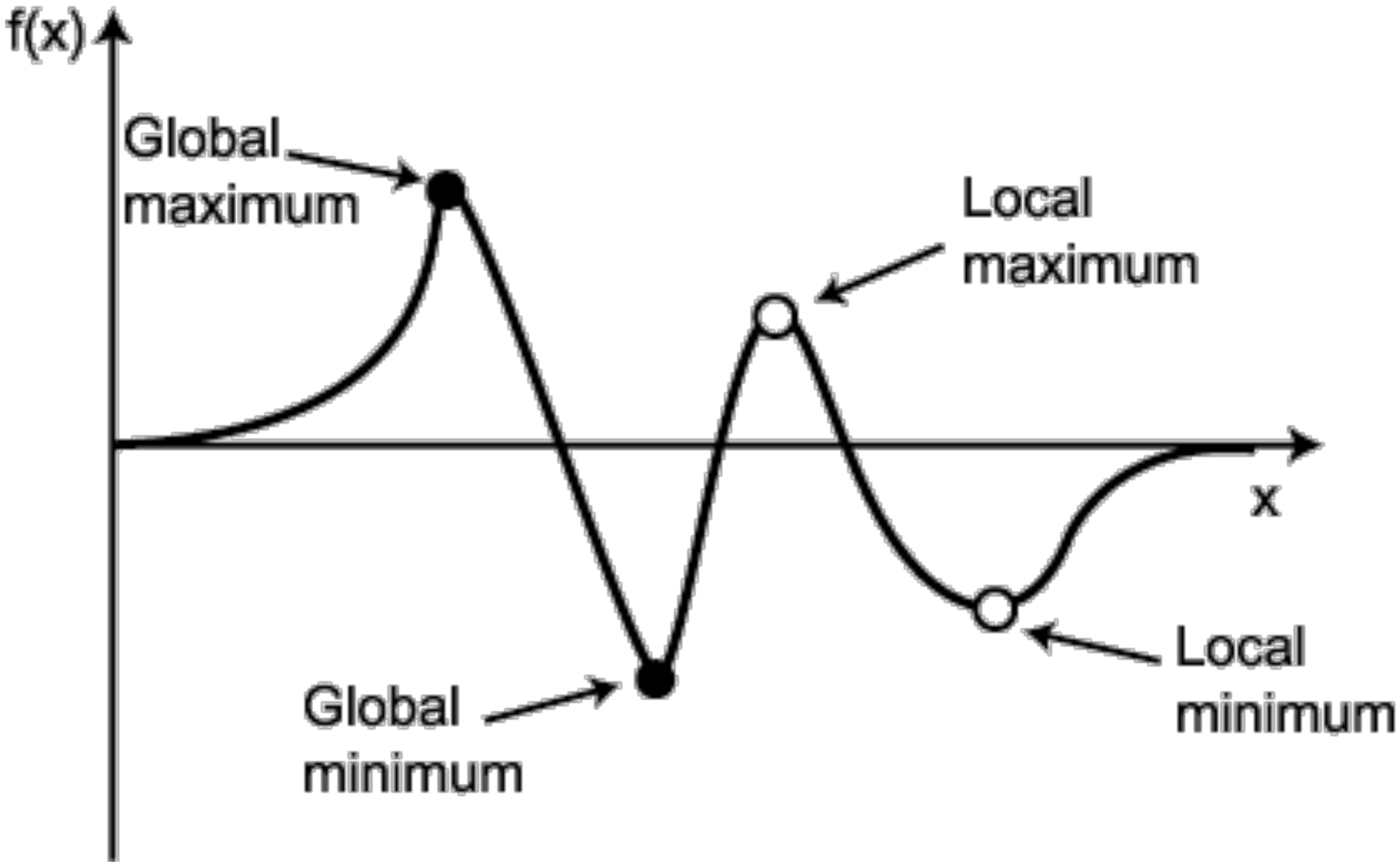


3D



2D

Global/local min/max



Summary

Three things

if you can remember only three…

- **XGboost** is **regularized** gradient boosting
- If you want to **win** try XGBoost
- Be careful with **overfitting**

Thank you

Data Workshop .eu

Praktyczne uczenie maszynowe dla programistów*

* Dla osób, które znają język Python (*jeden z najłatwiejszych języków dla początkujących)*

Uczenie poprzez **przykłady**, bez skomplikowanych i zawiłych (matematycznych) detali

Naucz się, jak zaimplementować **rozwiązanie end-to-end**

Start kursu: 30 października 2017 roku

33	14	38	12
Dni	Godzin	Minut	Sekund

30 WRZEŚNIA O 19:00 BEZPŁATNY WEBINAR

Praktyczne uczenie maszynowe dla programistów

Wszystko co chcesz wiedzieć o zawodzie przyszłości

Dołącz na dataworkshop.eu/free-webinar

biznesmysli.pl

Vladimir Alekseichenko

Architect
General Electric


vova.me


slon1024


hello@vova.me

dataworkshop.eu

DATA
WORKSHOP