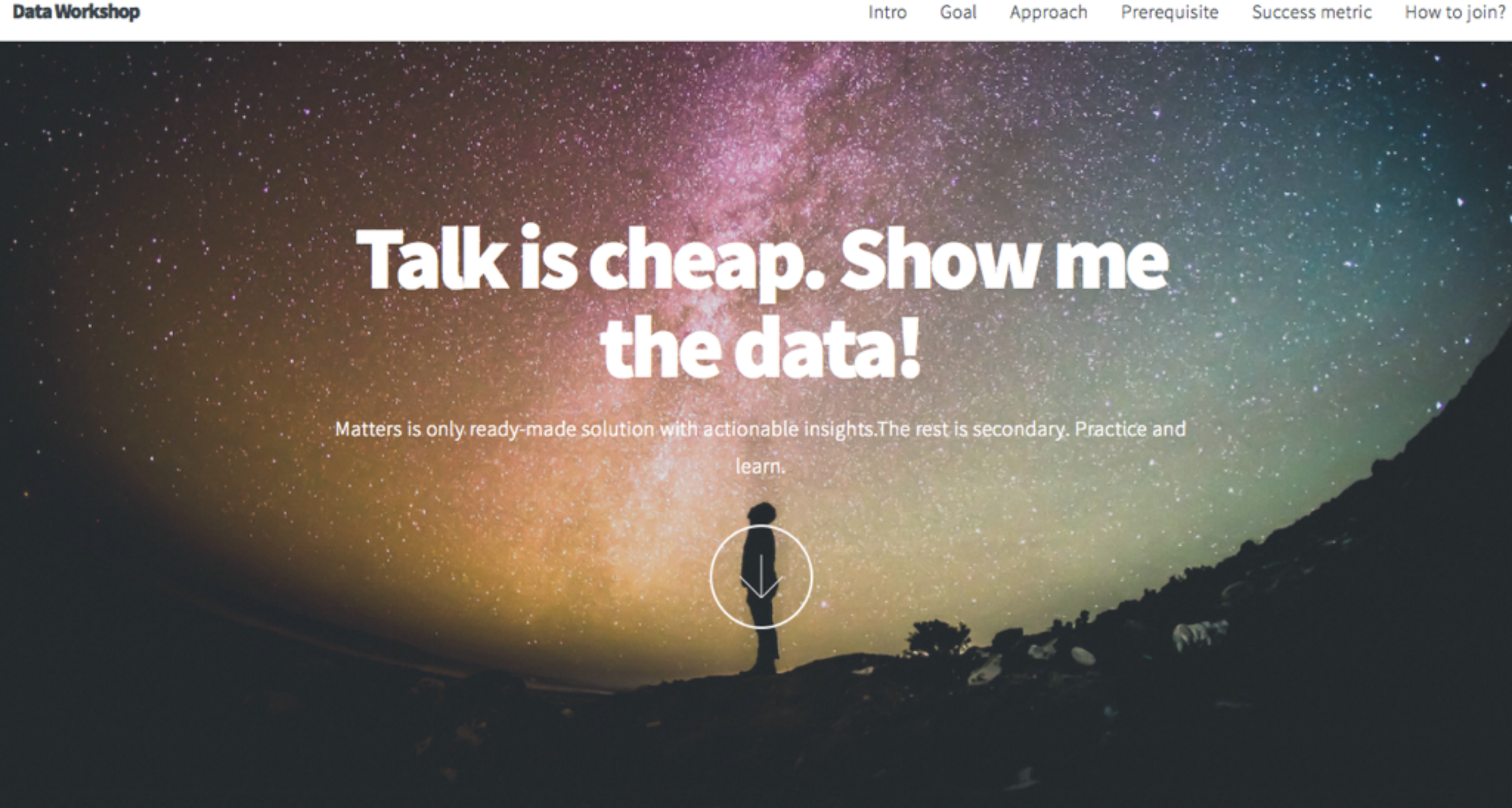
Data Workshop #2

When in doubt, use **xgboost**

dataworkshop.eu

Alekseichenko Vladimir

DataWorkshop.eu



About me

Vladimir Alekseichenko

Love analyze data

Architect

slon1024

slon1024

vova@vova.me

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

Agenda

- Boosting
- Gradient Boosting
- Extreme Gradient Boostring (xgboost)

Environment

github.com/dataworkshop/prerequisite

github.com/dataworkshop/xgboost

Packages

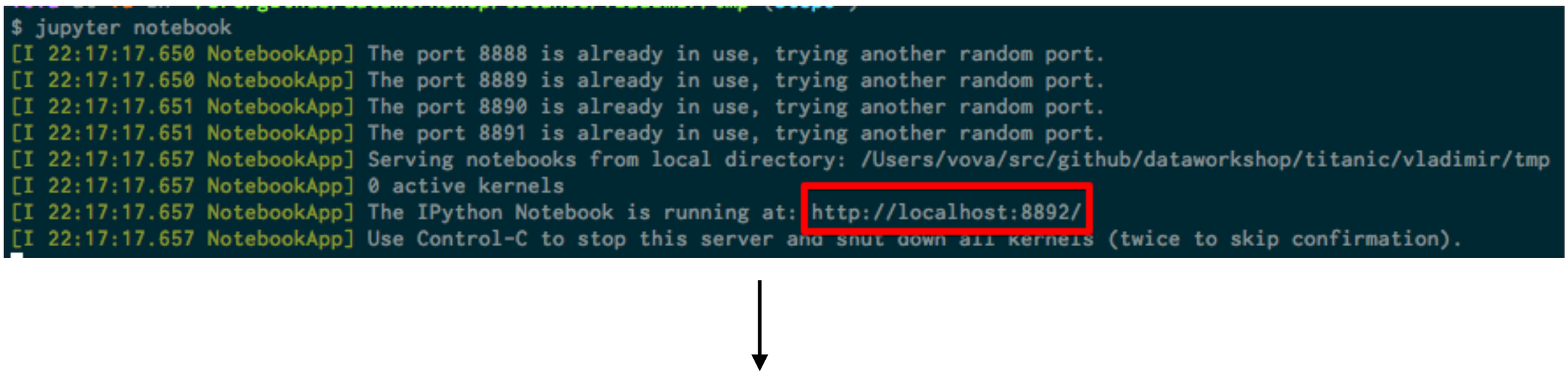
github.com/**dataworkshop/prerequisite**

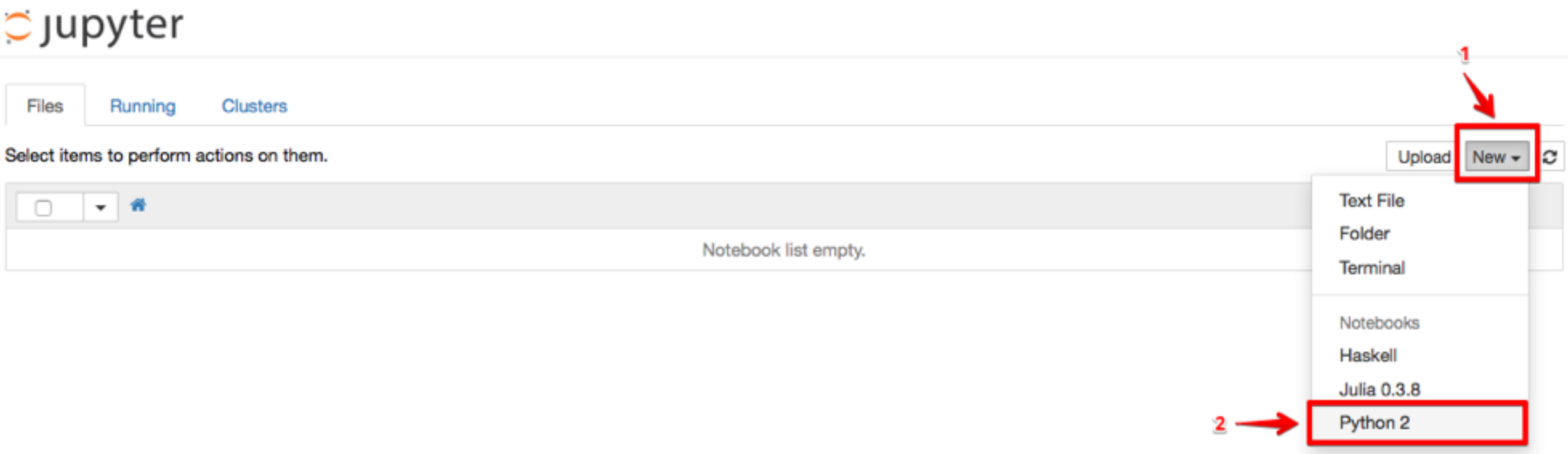
```
$ python run.py
seaborn-0.7.0 - OK
xgboost-0.4 - OK
matplotlib-1.5.1 - OK
IPython-4.1.2 - OK
numpy-1.11.0 - OK
pandas-0.18.0 - OK
sklearn-0.17.1 - OK

==================================================
All right, you are ready to go on Data Workshop!
```

```
$ python run.py
seaborn-0.6 should be upgraded to seaborn-0.7
xgboost-0.4 - OK
matplotlib-1.5.1 - OK
IPython-4.1.2 - OK
numpy-1.11.0 - OK
pandas-0.18.0 - OK
sklearn-0.17.1 - OK

==================================================
RECOMENDATION (without upgrade some needed features could be missing)
pip install --upgrade seaborn
```

```
$ python run.py
seaborn-0.7.0 - OK
xgboost - missing
matplotlib-1.5.1 - OK
IPython-4.1.2 - OK
numpy-1.11.0 - OK
pandas-0.18.0 - OK
sklearn-0.17.1 - OK

==================================================
REQUIRED
Please install those packages before Data Workshop: xgboost
pip install xgboost
More info how to install xgboost: http://xgboost.readthedocs.org/en/latest/build.html
```

jupyter notebook





Motivation

why boosting?

Motivation

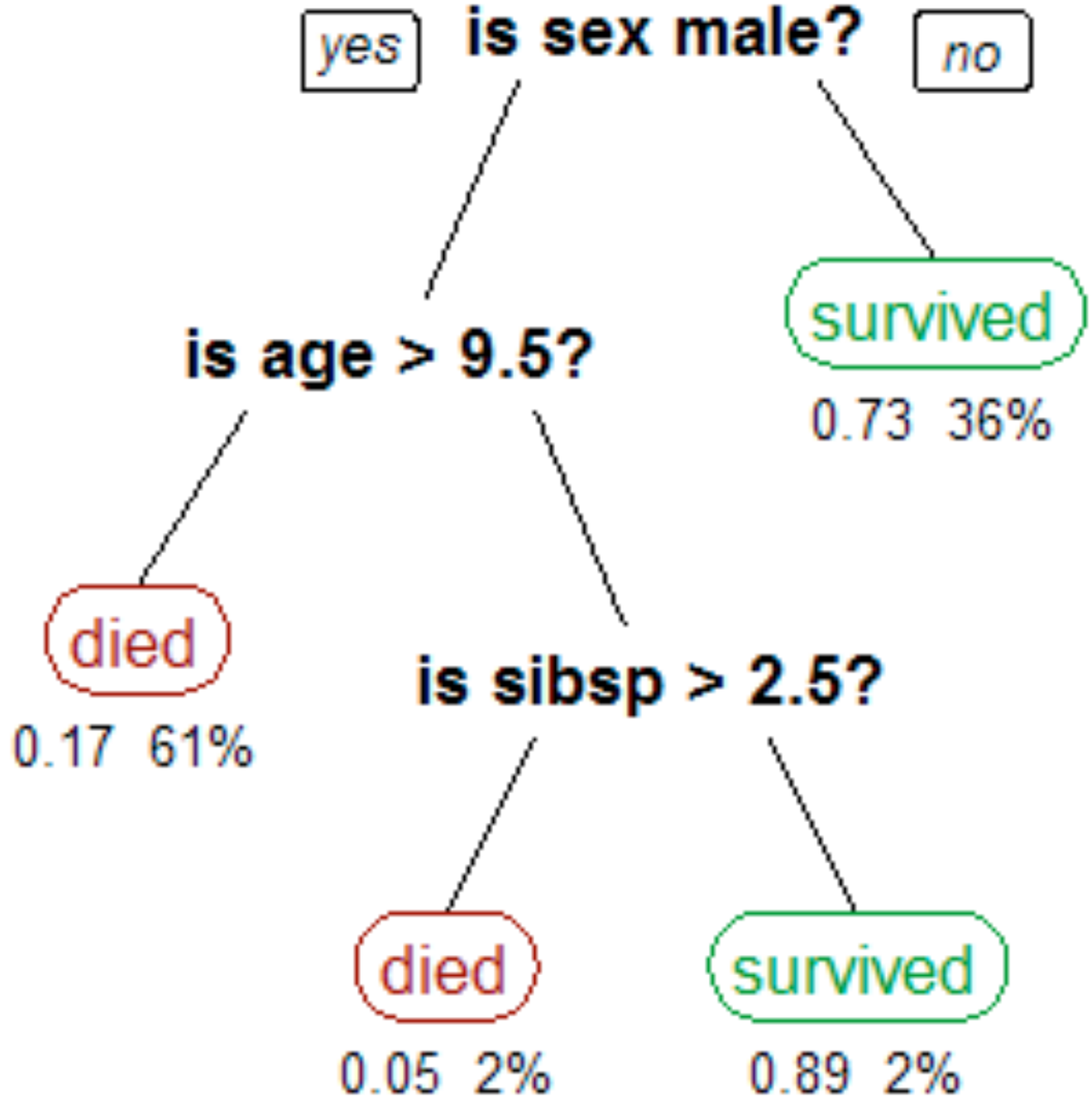
XGboost is utilized

in many **winning** solutions.

Decision trees

A bit history, DT is not enough?

Decision Trees (CART)



Decision Trees (CART)

pros - interpretable

cons - relatively a poor quality

Ensembles

Bagging, Random Forest, Bootstrap…

```
sklearn.ensembles.*
xgboost.*
```

Ensembles

Two heads are better than one

Одна голова хорошо, а две лучше

Co dwie głowy, to nie jedna

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

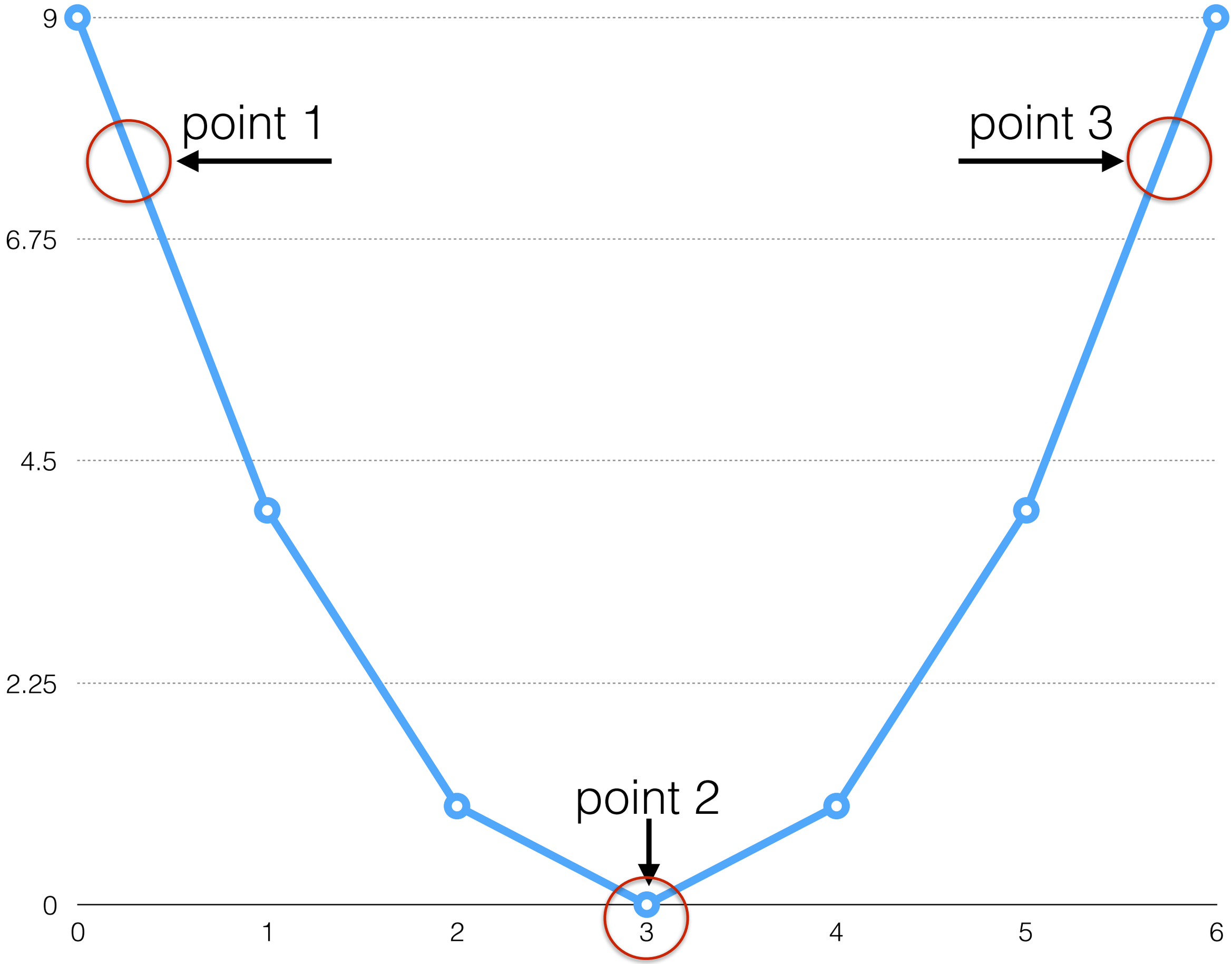
Tianqi Chen

the author of xgboost

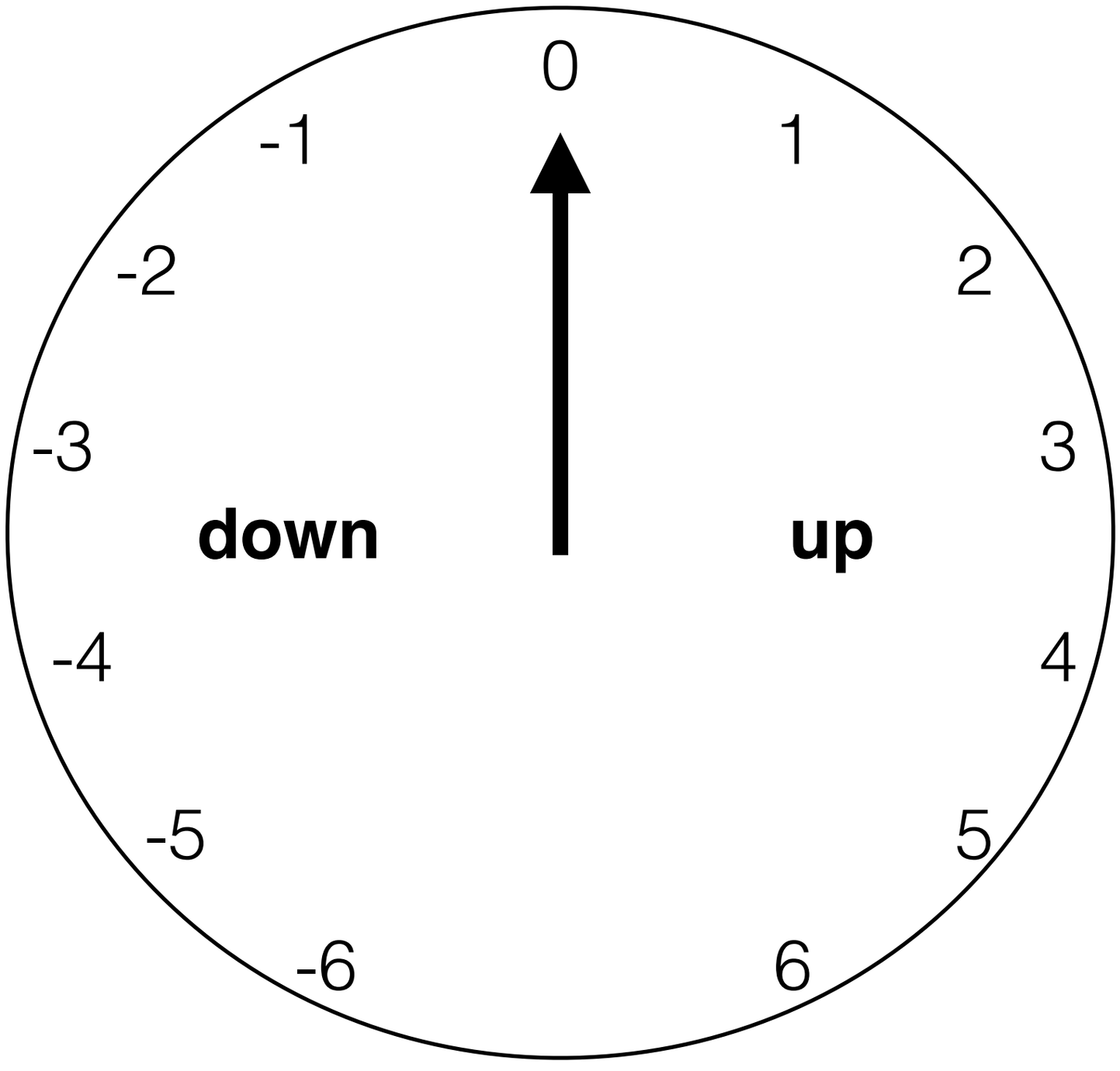
Derivative

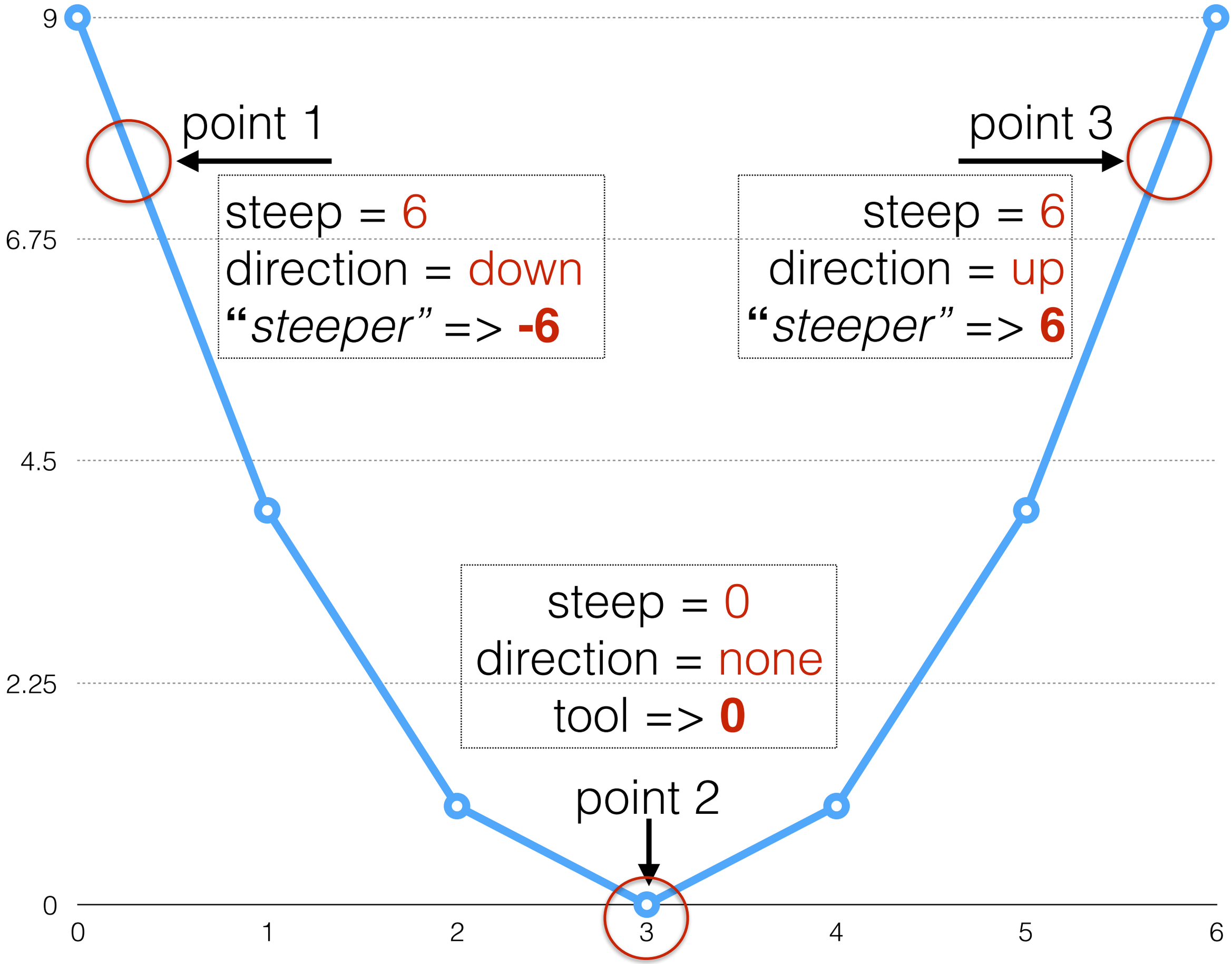
Intuitive example(s)

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

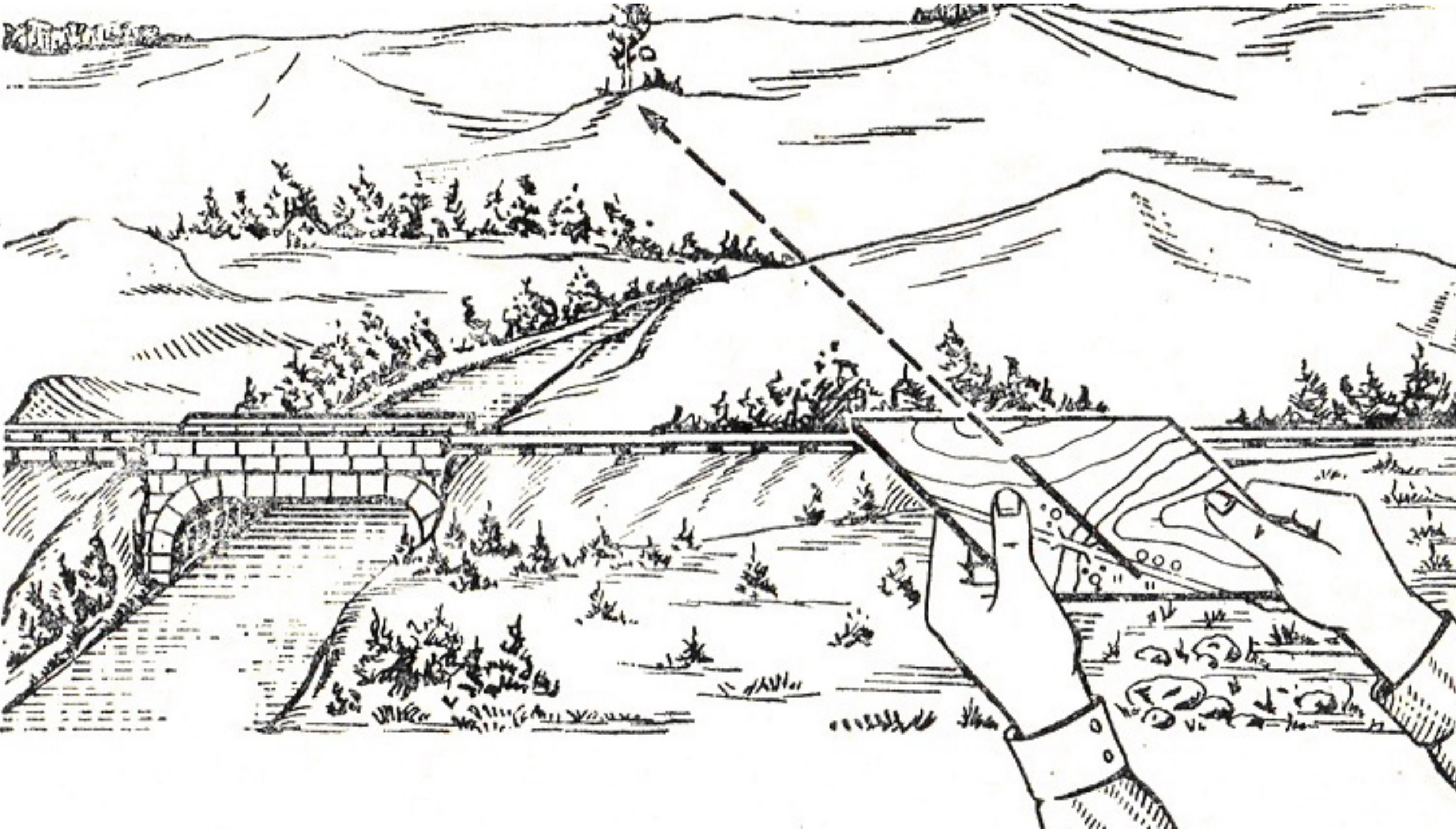
A tool (*steeper*) for measure **steep** and **direction**



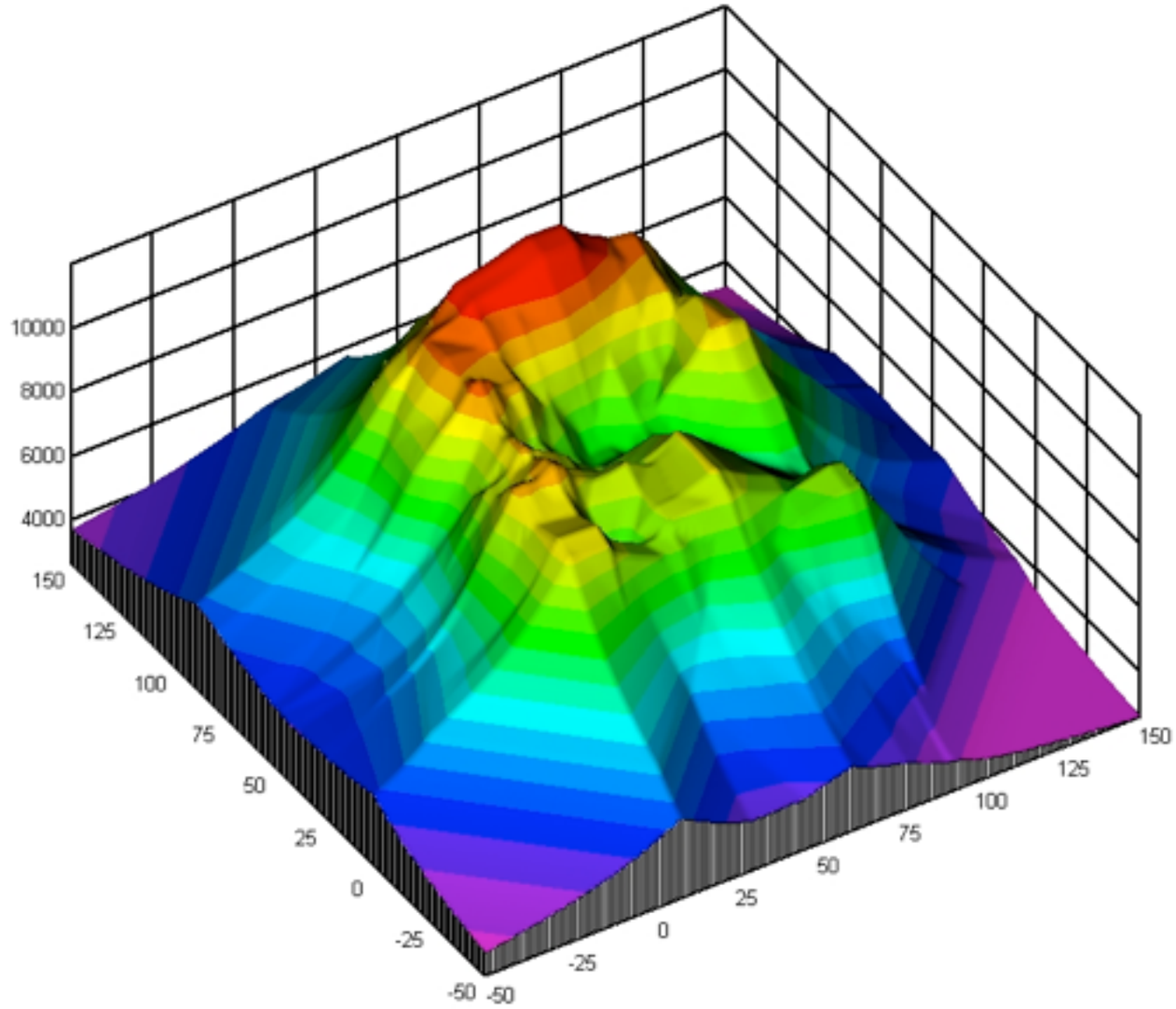

9
6.75
4.5
2.25
0
0
1
2
3
4
5
6
point 1
steep = 6
direction = down
“steeper” => -6
point 3
steep = 6
direction = up
“steeper” => 6
steep = 0
direction = none
tool => 0
point 2

Gradient Descent

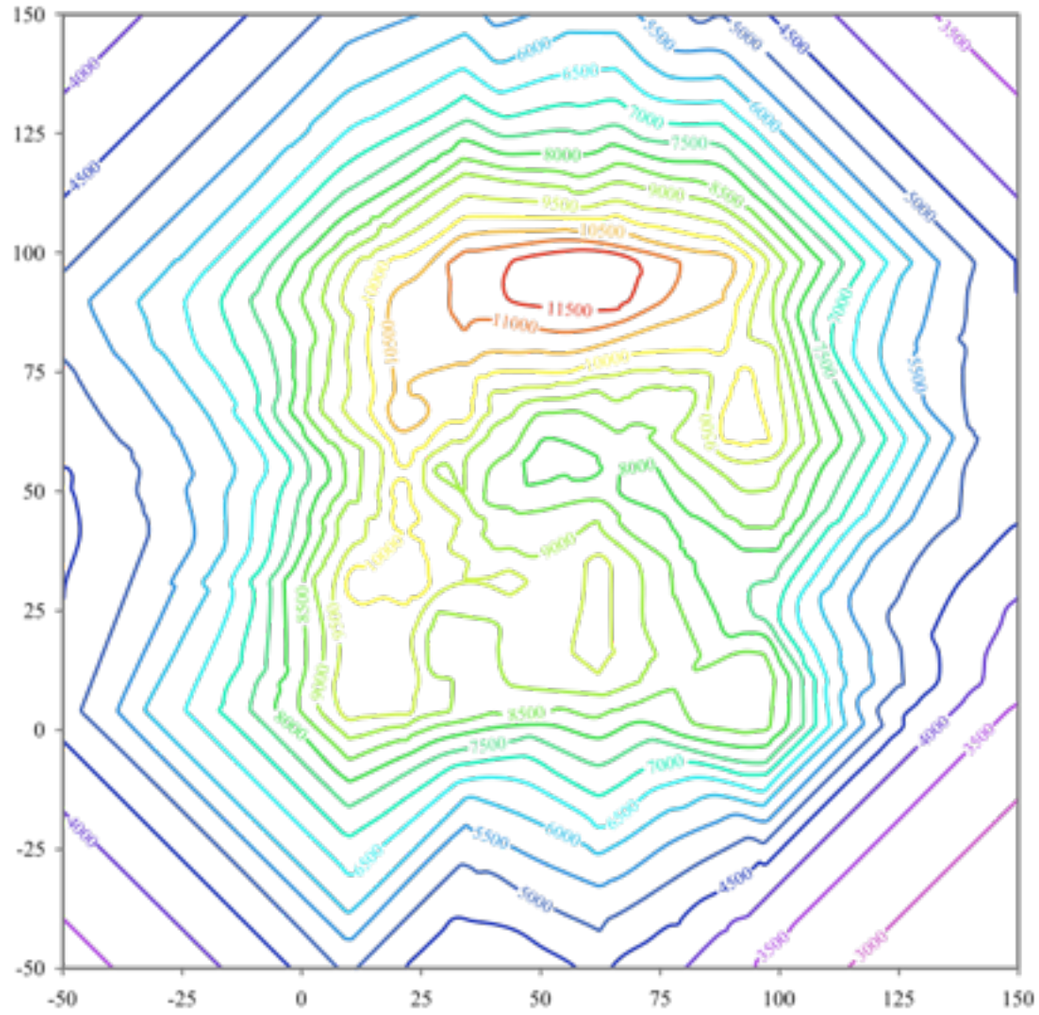
Contour Line



Contour Line

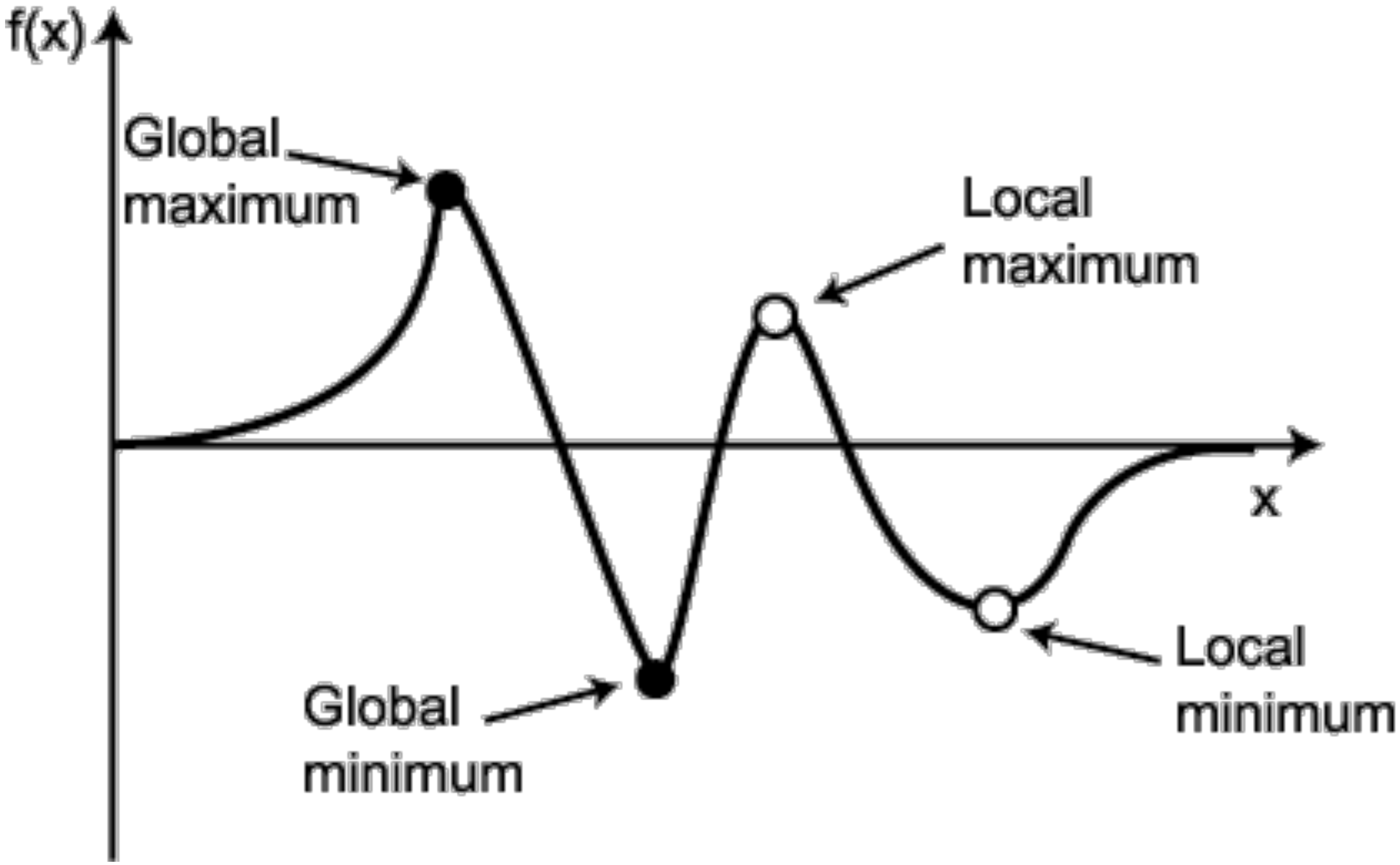


3D



2D

Global/local min/max



Summary

Three things

if you can remember only three…

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

Thank you