Data Science Toolbox:

NumPy, Pandas, Scikit-learn

Tim Babych, PL.PyCON 2015

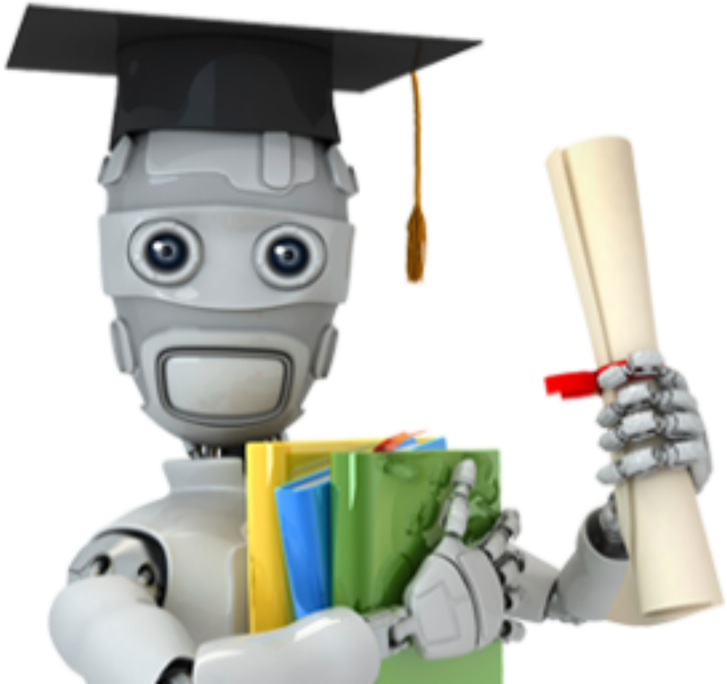
About

- Machine Leaning overview
- Data Science & Big Data
- Algorithms and tools
- Skills and Courses
- Questions

Machine Learning

Field of study that gives computers the ability to learn without being explicitly programmed

Arthur Samuel, 1959



Supervised learning

“Right answers” do exist

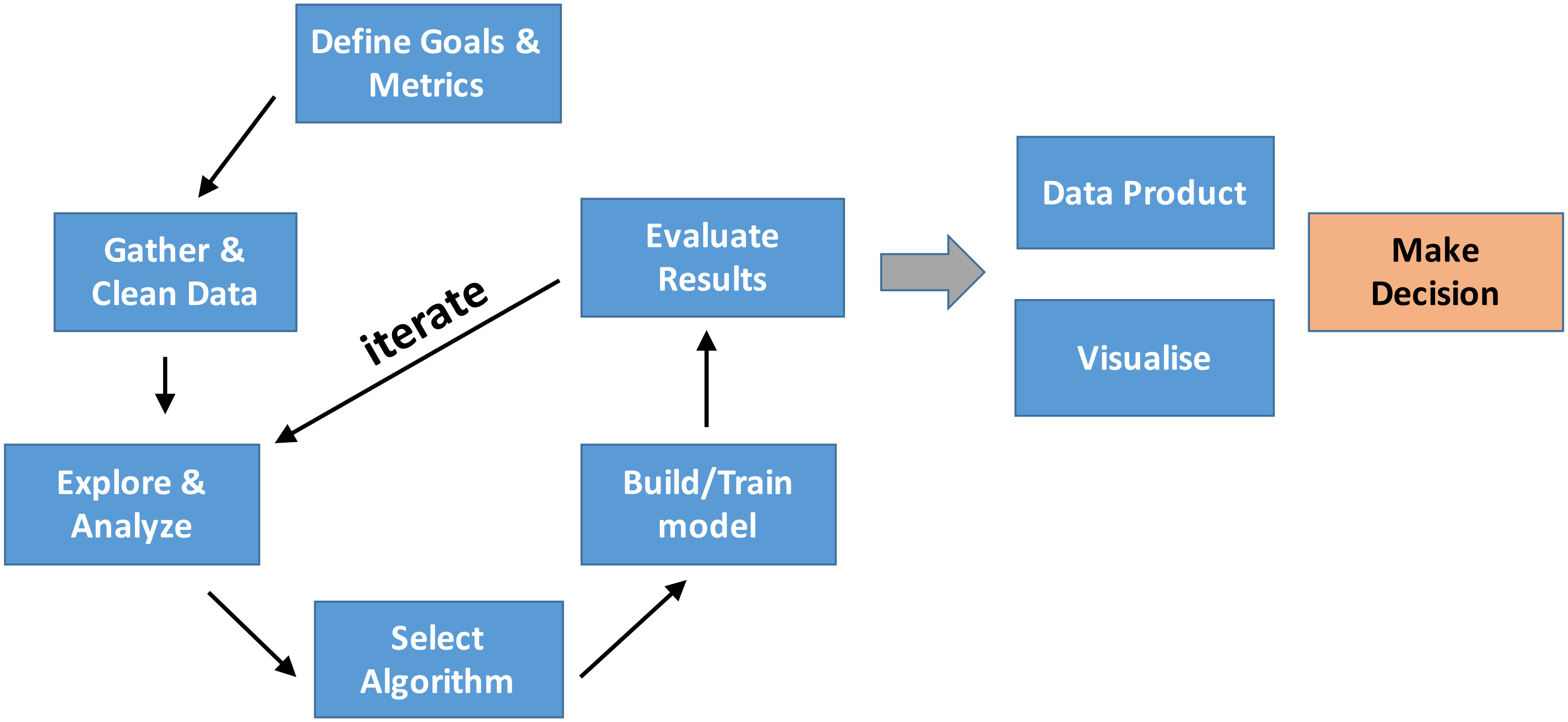
- Spam detectors
- Weather prediction
- Game outcomes
- Medical diagnosis
- Insurance
- Object detection
- Speech recognition

Unsupervised learning

There are no right answers! Much harder

- Find some structure in given data
- Cluster data into groups
- Playing games

Data Science Flow



Define Goals & Metrics

Gather & Clean Data

Explore & Analyze

Select Algorithm

Build/Train model

Evaluate Results

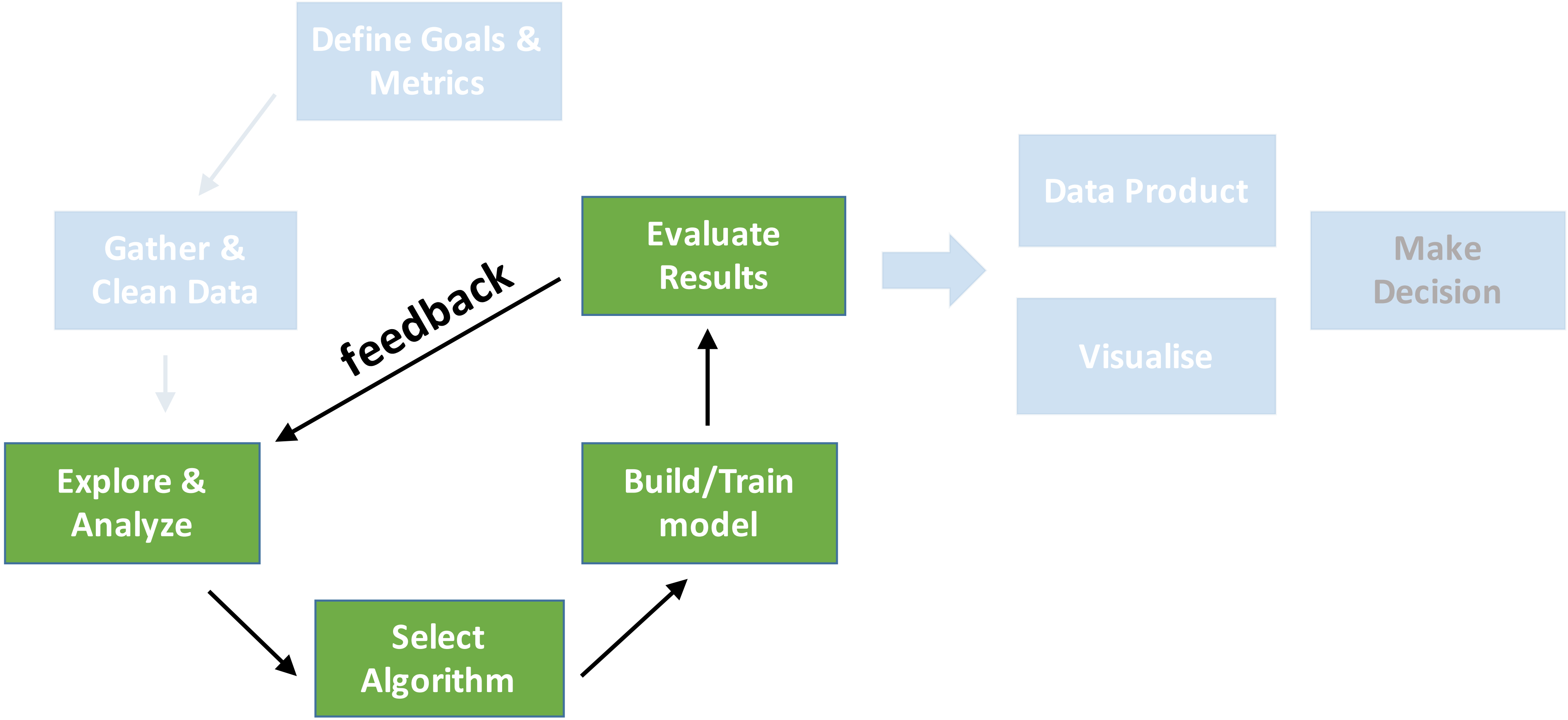
iterate

Data Product

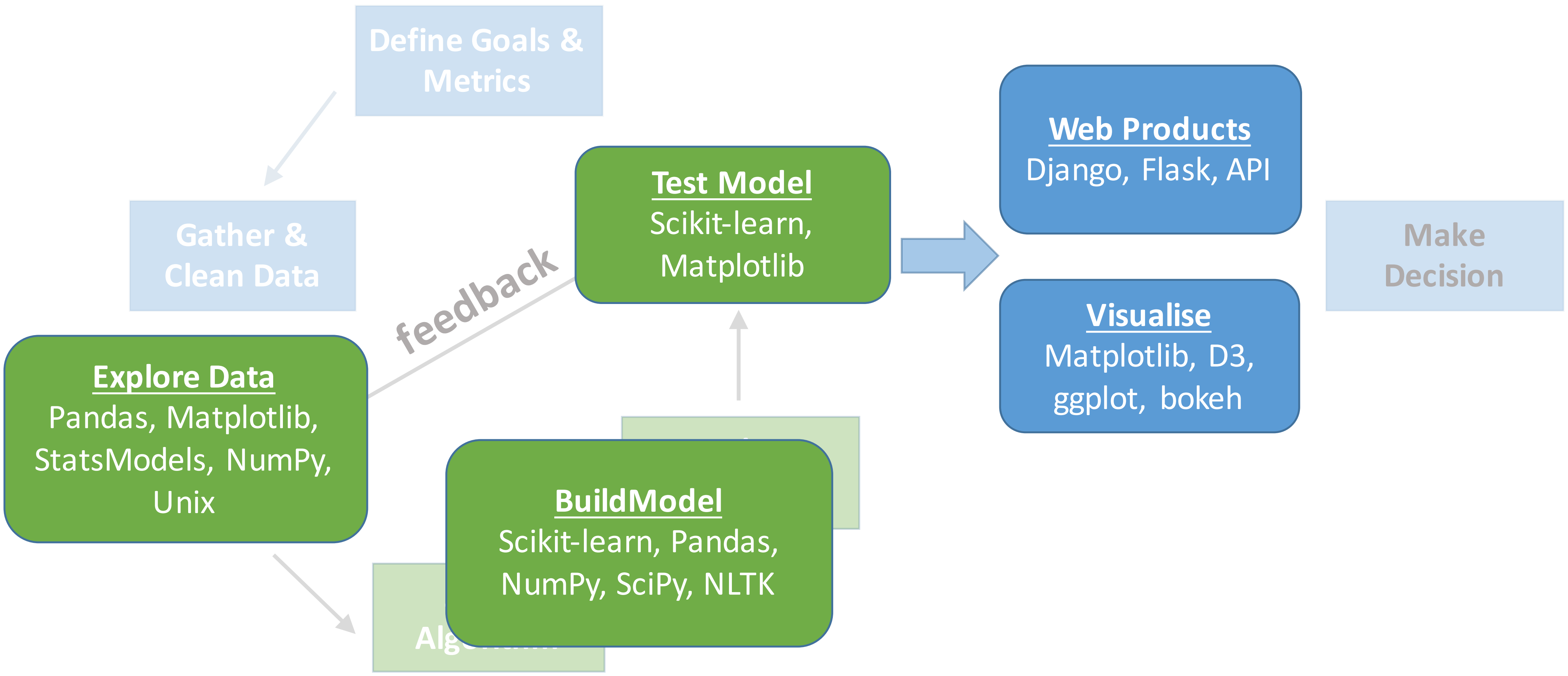
Visualise

Make Decision

Machine Learning part



Tools for the tasks



Define Goals & Metrics

Gather & Clean Data

Explore Data
Pandas, Matplotlib, StatsModels, NumPy, Unix

feedback

Test Model
Scikit-learn, Matplotlib

Alg

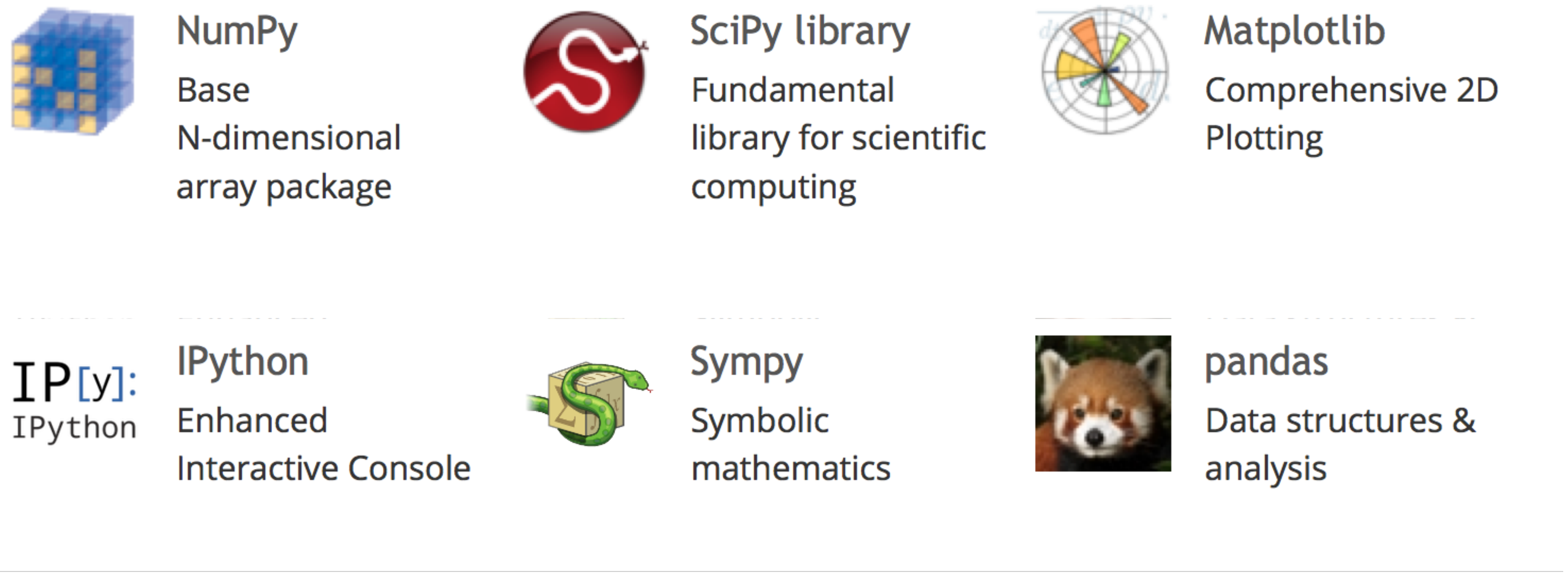
BuildModel
Scikit-learn, Pandas, NumPy, SciPy, NLTK

Web Products
Django, Flask, API

Visualise
Matplotlib, D3, ggplot, bokeh

Make Decision

SciPy stack



NumPy

Base N-dimensional array package

SciPy library

Fundamental library for scientific computing

Matplotlib

Comprehensive 2D Plotting

IPython

Enhanced Interactive Console

Sympy

Symbolic mathematics

pandas

Data structures & analysis

Example: Titanic passengers

survival	(0 = No; 1 = Yes)
pclass	Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
name	
sex	
age	
sibsp	Number of Siblings/Spouses Aboard
parch	Number of Parents/Children Aboard
ticket	Ticket Number
fare	
cabin	
embarked	Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

Toolkit

```
pip install numpy scikit-learn pandas matplotlib
pip install "ipython[notebook]"
```

OR

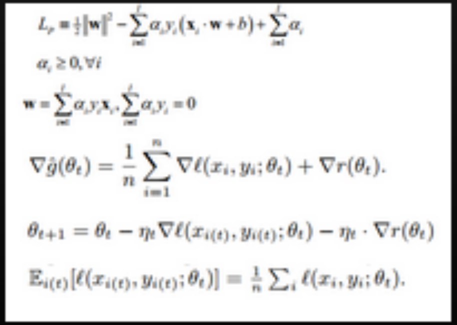
Use Anaconda distribution

Machine Learning

what society thinks I do

what my friends think I do

what my parents think I do



$$L_p = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{l} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{w} + b) + \sum_{i=1}^{l} \alpha_i$$

$$\alpha_i \geq 0, \forall i$$

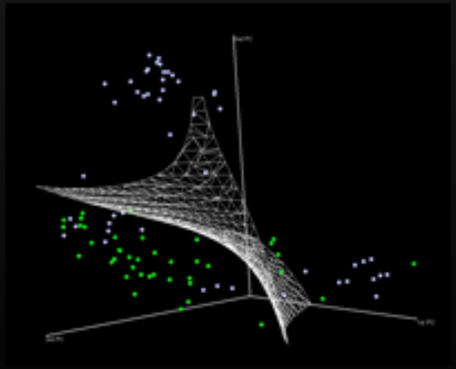
$$\mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i, \sum_{i=1}^{l} \alpha_i y_i = 0$$

$$\nabla \hat{g}(\theta_t) = \frac{1}{n} \sum_{i=1}^{n} \nabla \ell(x_i, y_i; \theta_t) + \nabla r(\theta_t).$$

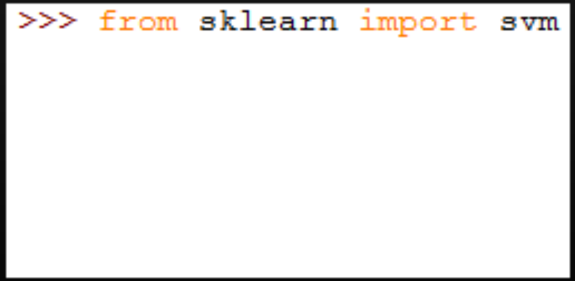
$$\theta_{t+1} = \theta_t - \eta_t \nabla \ell(x_{i(t)}, y_{i(t)}; \theta_t) - \eta_t \cdot \nabla r(\theta_t)$$

$$\mathbb{E}_{i(t)}[\ell(x_{i(t)}, y_{i(t)}; \theta_t)] = \frac{1}{n} \sum_i \ell(x_i, y_i; \theta_t).$$

what other programmers think I do



what I think I do



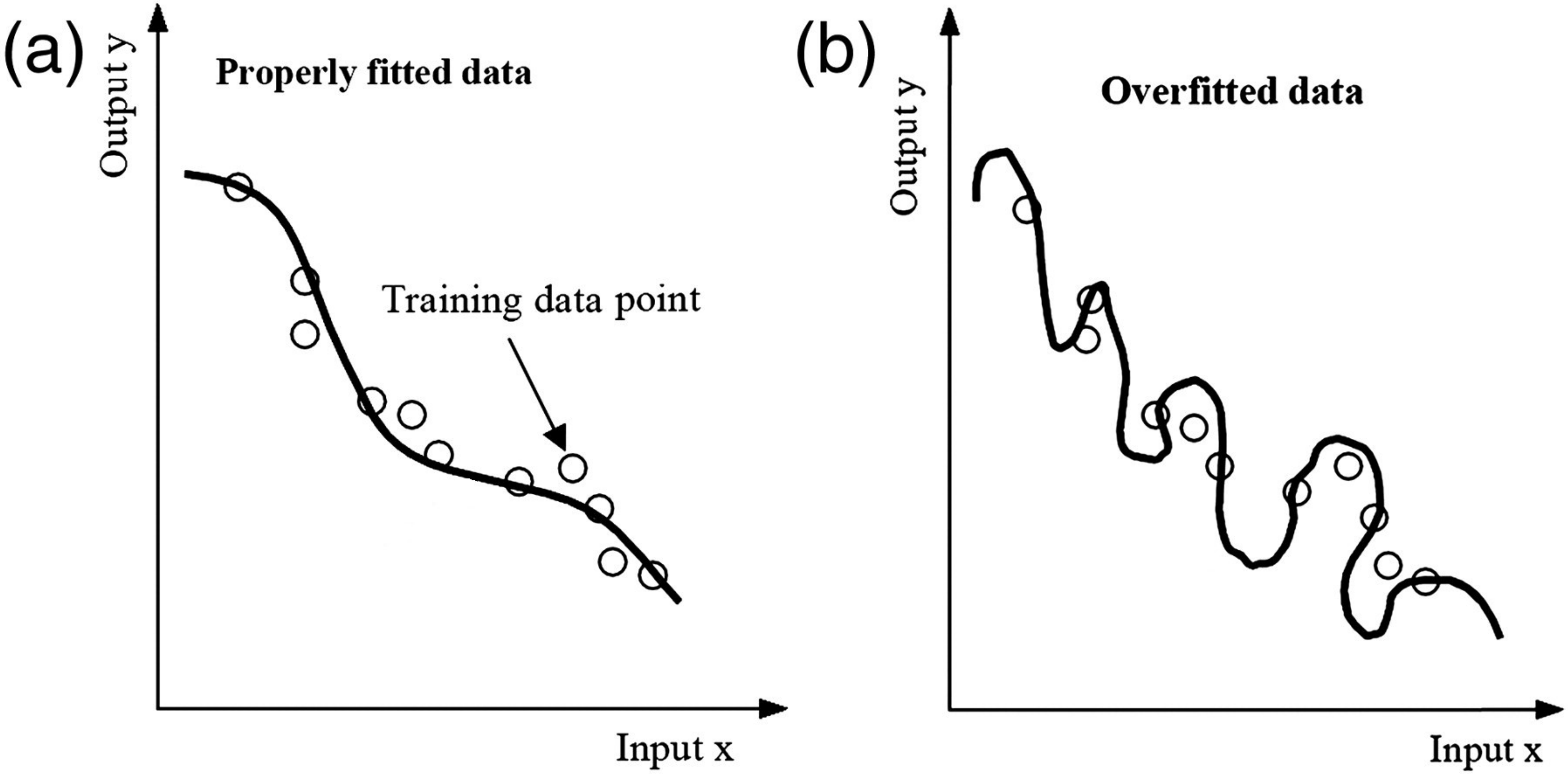
```
>>> from sklearn import svm
```

what I really do

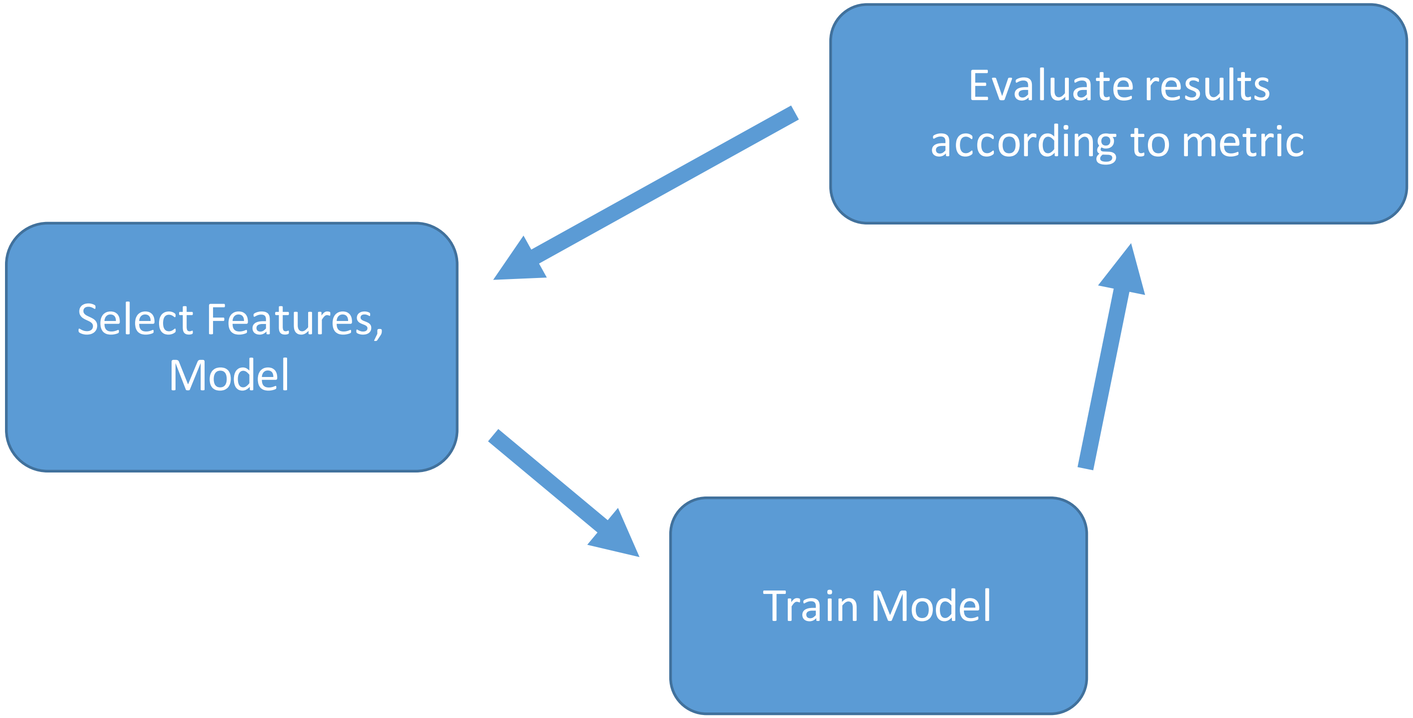
Let’s dive into Jupyter notebooks: they are **awesome**!

https://github.com/tymofij/datascience-pandas-talk-pycon-pl

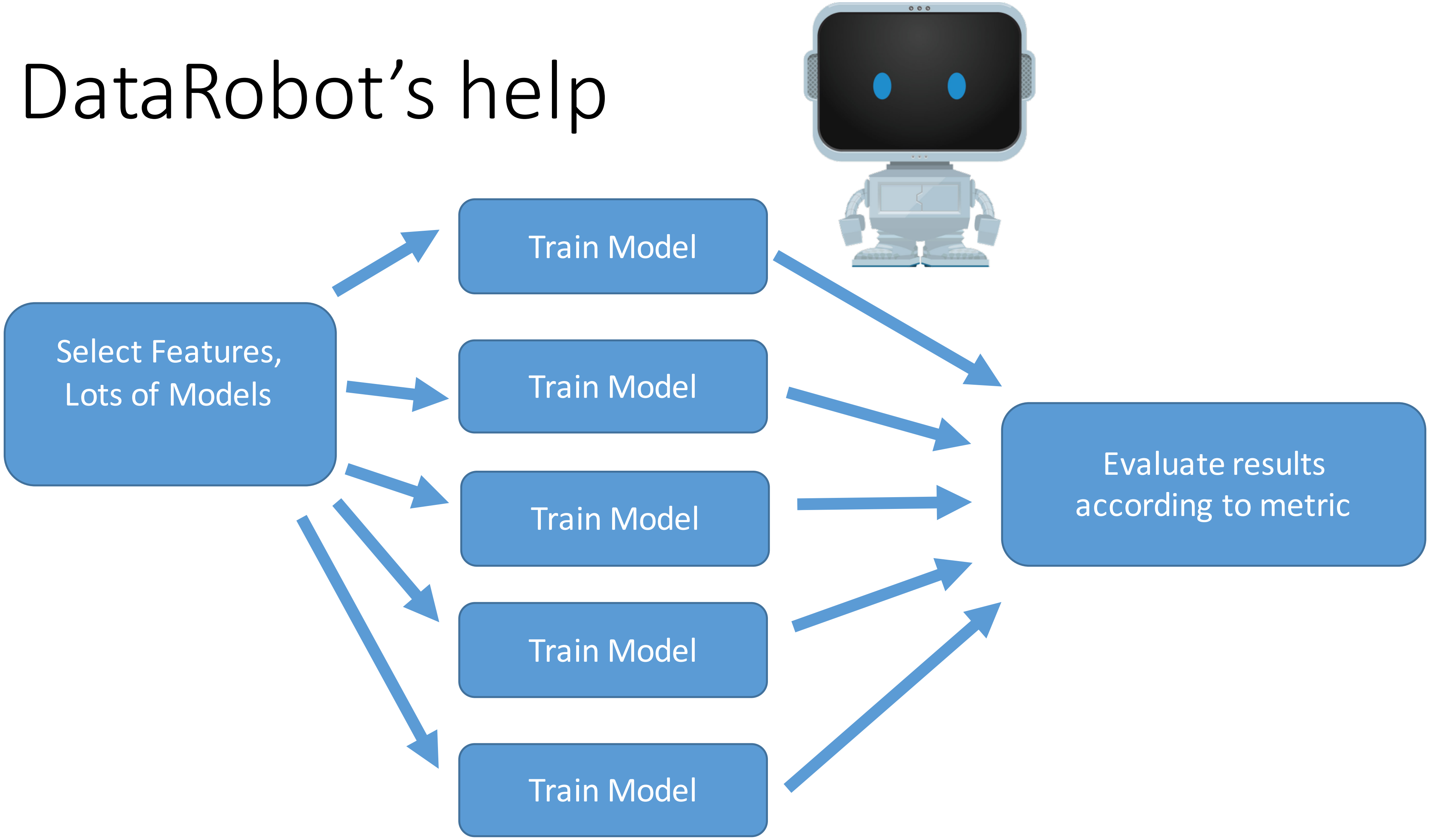
Do not overdo it.



A day in datascientist’s life



DataRobot’s help



Select Features, Lots of Models

Train Model

Train Model

Train Model

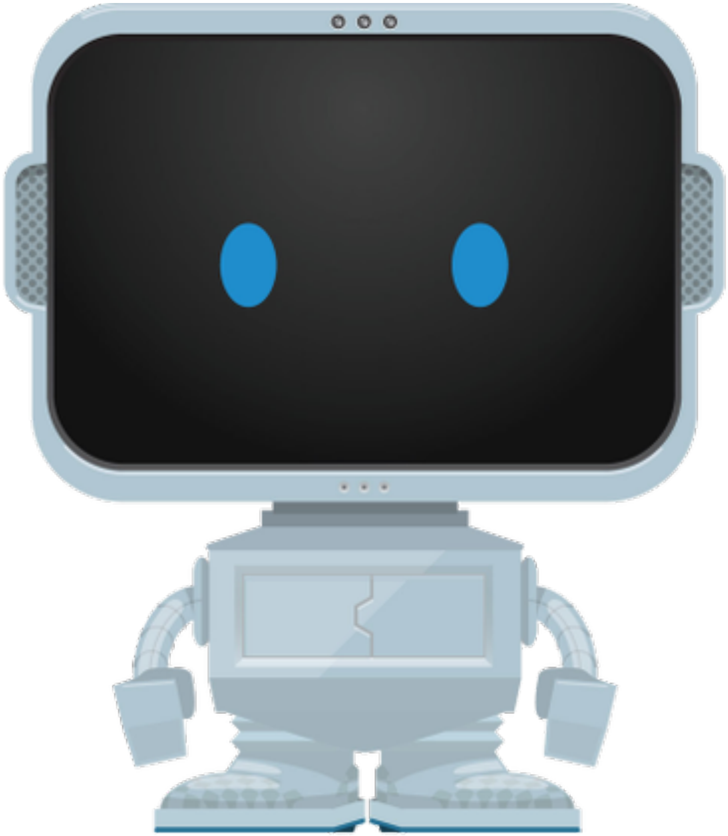
Train Model

Train Model

Evaluate results according to metric

DataRobot is hiring!

- Software Engineers
- Data Scientists
- DevOps
- Boston and Kyiv offices



http://www.datarobot.com/careers/

Selected Books

- Learning From Data -
  small, good for beginners and has an online course

- Machine Learning: A Probabilistic Perspective
  larger and still current and very popular

- The Elements of Statistical Learning: Data Mining, Inference, and Prediction
  a lot of theory, has free PDF edition

Selected Machine Learning resources

- Machine Learning by Andy Ng (Coursera)
- Intro to Machine Learning by Sebastian Thrun (Udacity)
- dataquest.io
- Kaggle competitions and tutorials

Thanks!
Questions?

Tim Babych

tim.babych@gmail.com
http://clear.com.ua
http://github.com/tymofij
http://twitter.com/tymofiy