PyData Ecosystem Webinar:
Introduction and Best Practices

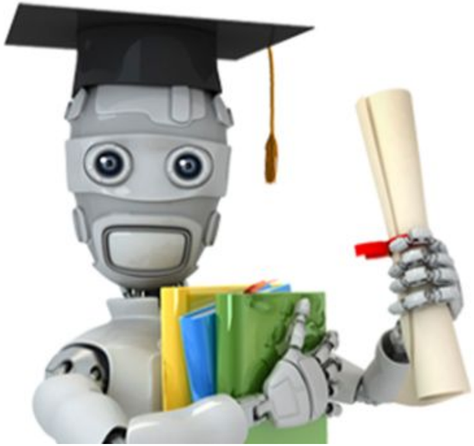
Tim Babych

Sphere Software Senior Engineer

TIM BABYCH

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

--Arthur Samuel, 1959



SUPERVISED LEARNING

“Right answers” do exist

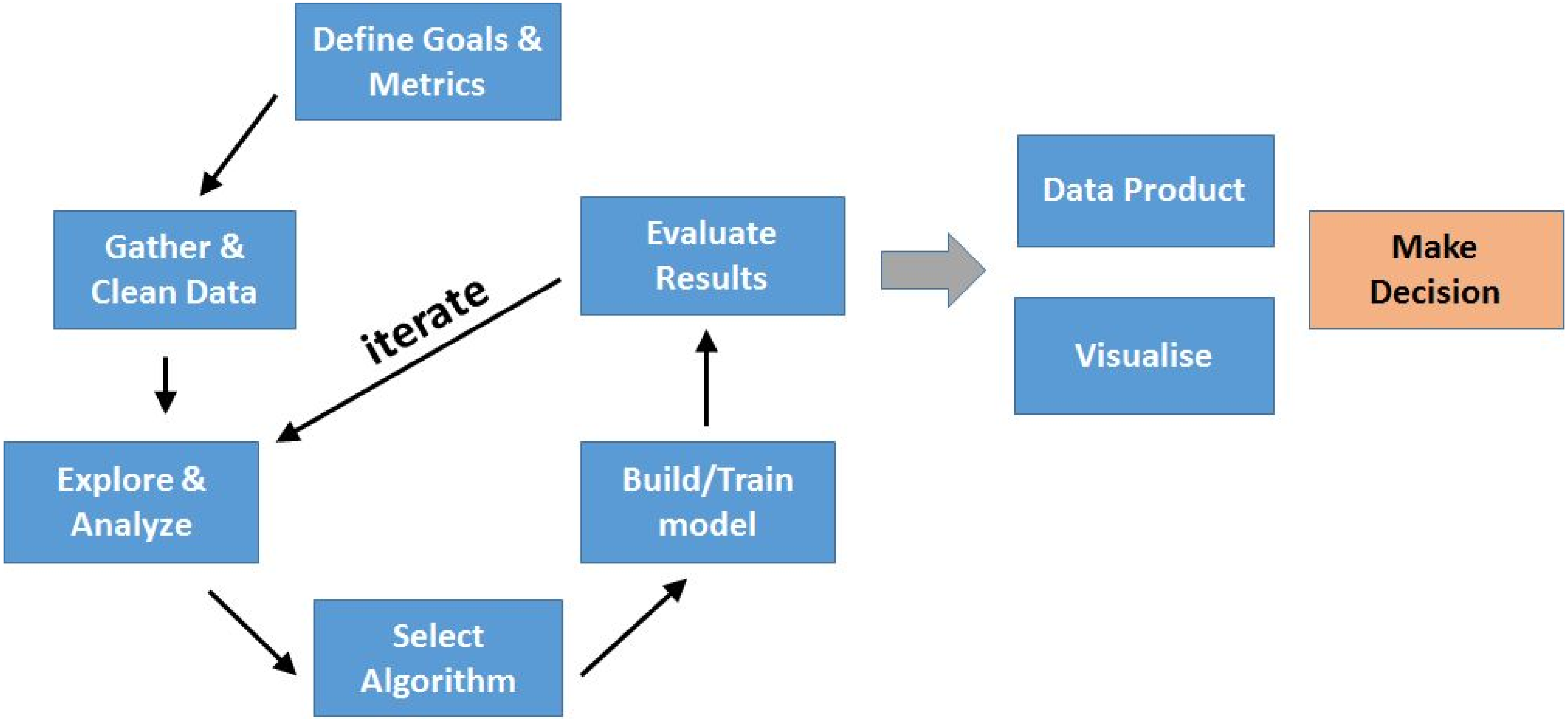
- Spam detectors
- Weather prediction
- Game outcomes
- Medical diagnosis
- Insurance
- Object detection
- Speech recognition

UNSUPERVISED LEARNING

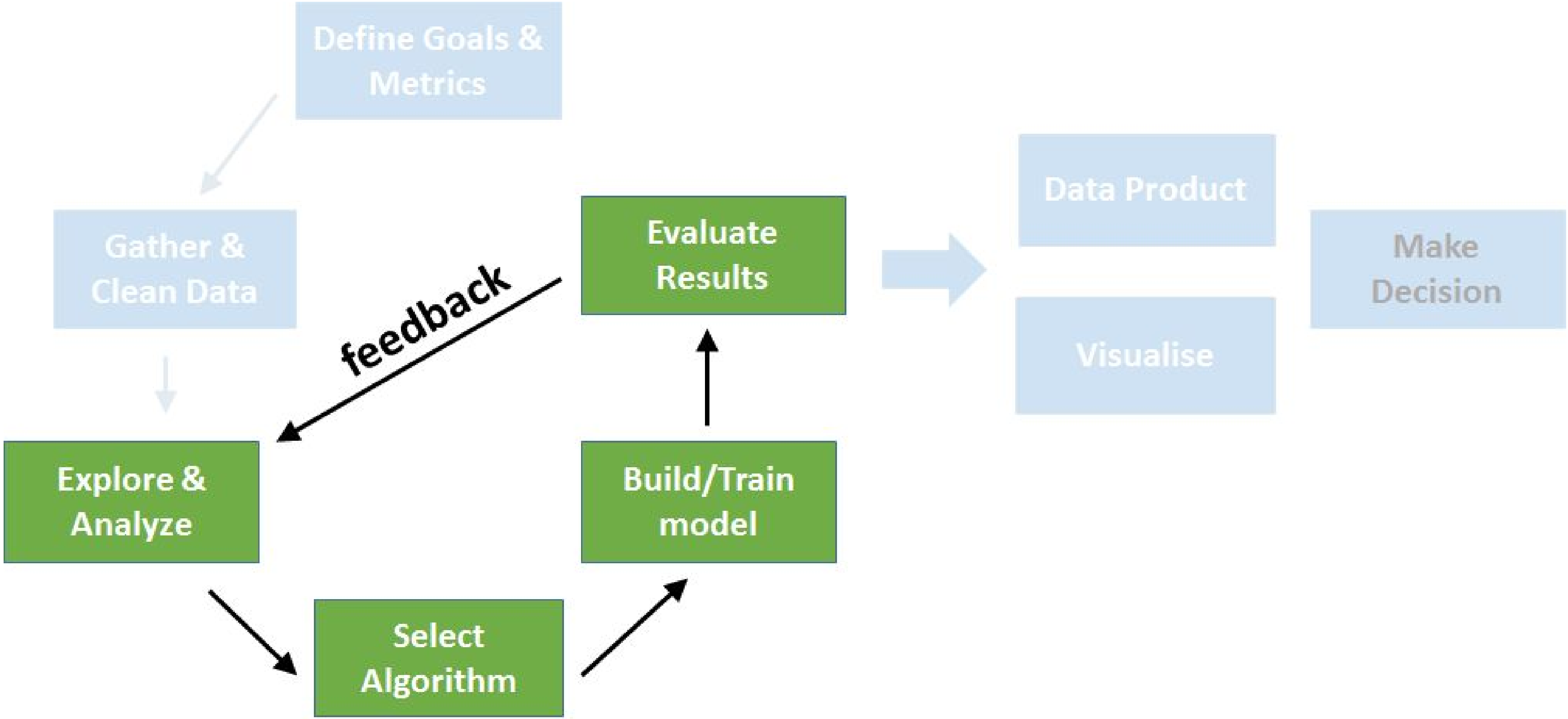
There are no right answers! Much harder.

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

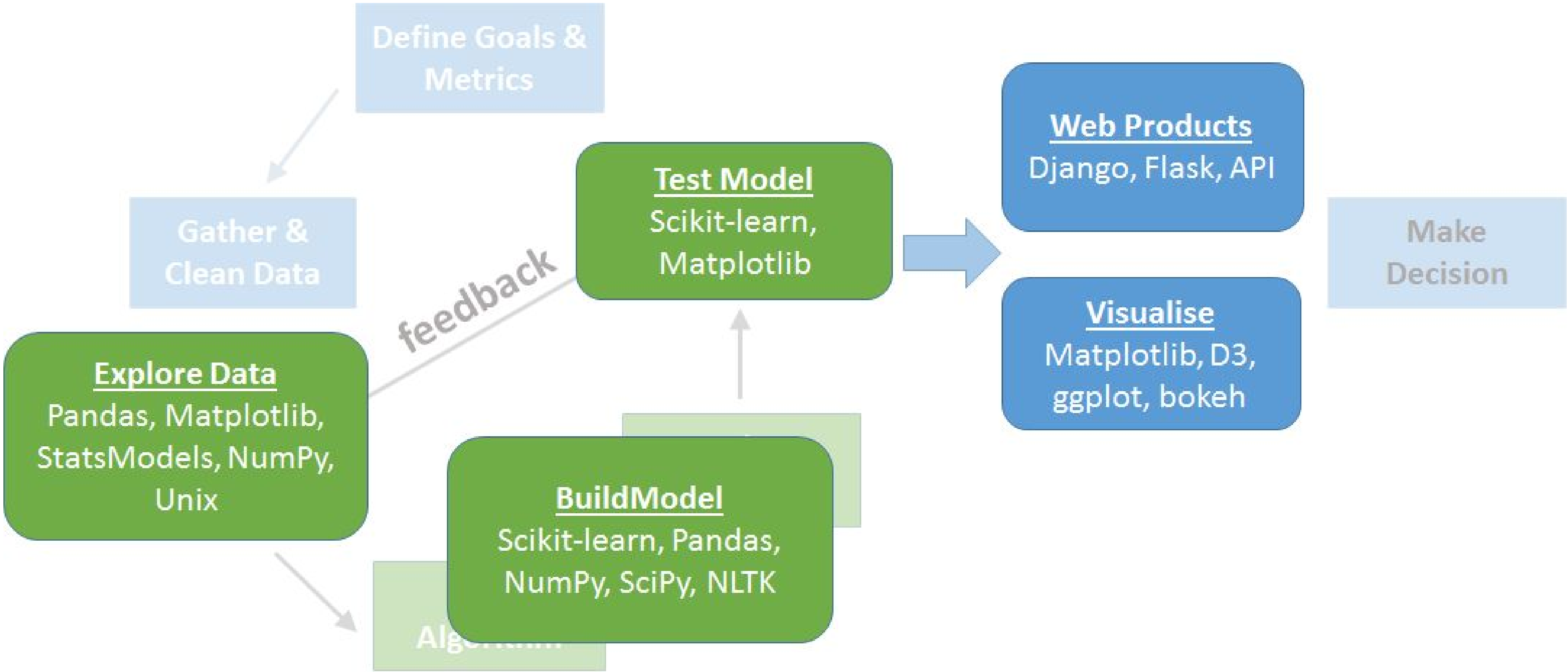
DATA SCIENCE FLOW



MACHINE LEARNING PART



POPULAR TOOLS FOR THE TASKS



Define Goals & Metrics

Gather & Clean Data

feedback

Explore Data
Pandas, Matplotlib, StatsModels, NumPy, Unix

Alg

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

Test Model
Scikit-learn, Matplotlib

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

TITANIC SURVIVORS

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)

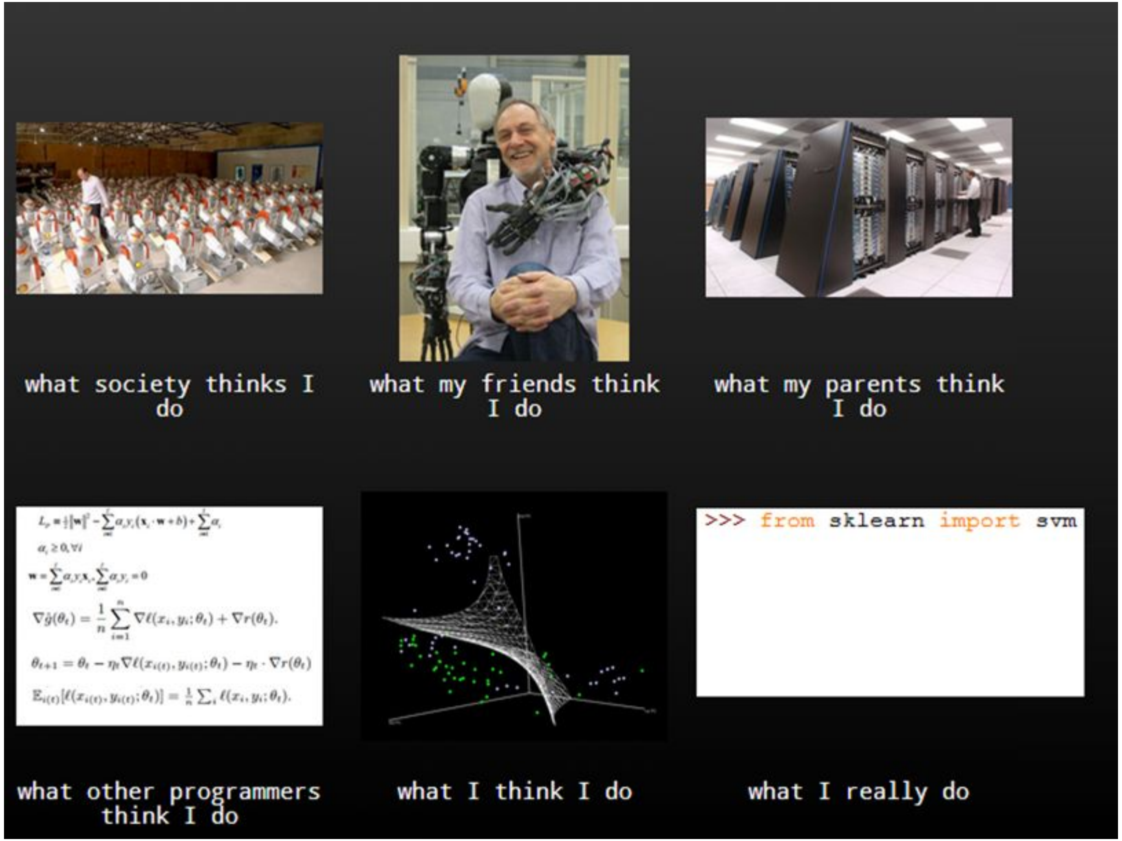
TOOLKIT

```
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
what other programmers think I do
what I think I do
what I really do

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

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

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

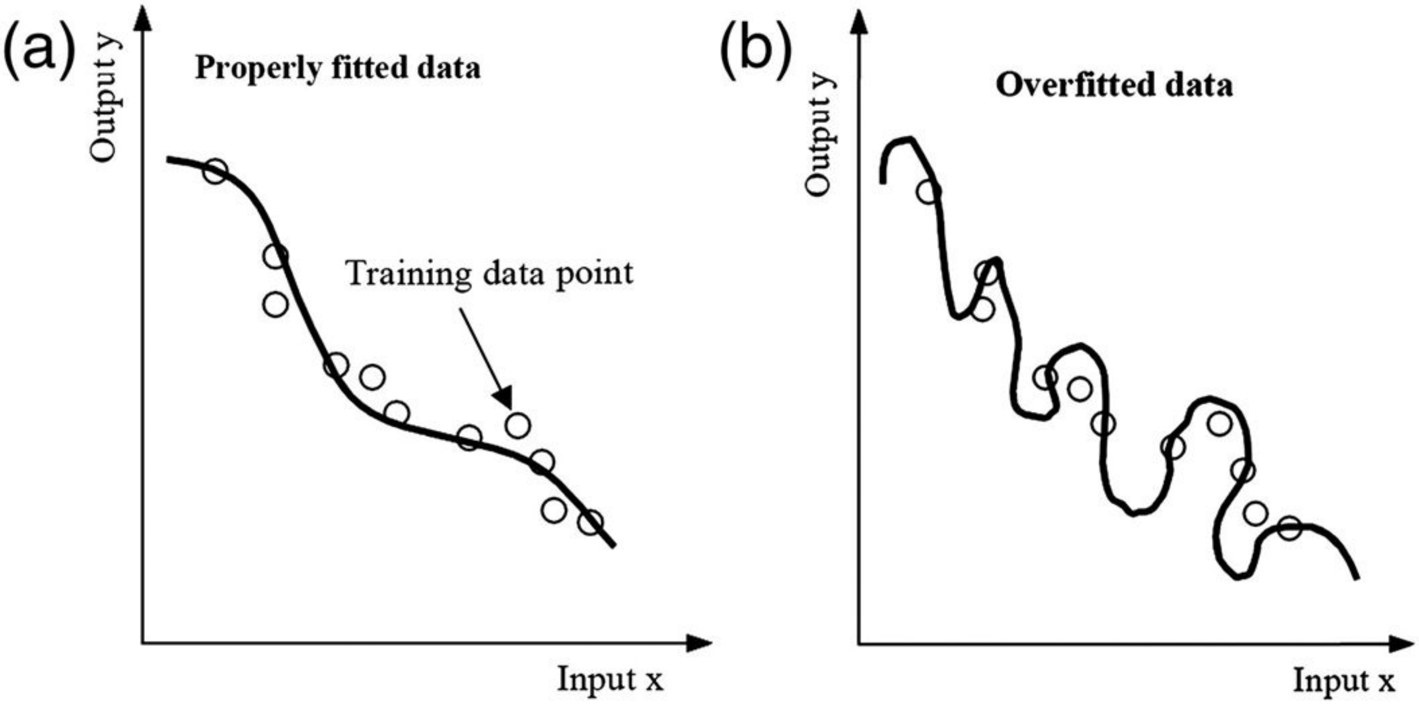
$$\nabla \hat{g}(\theta_t) = \frac{1}{n} \sum_{i=1}^{n} \nabla \ell(x_i, y_i; \theta_t) + \nabla r(\theta_t).$$

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

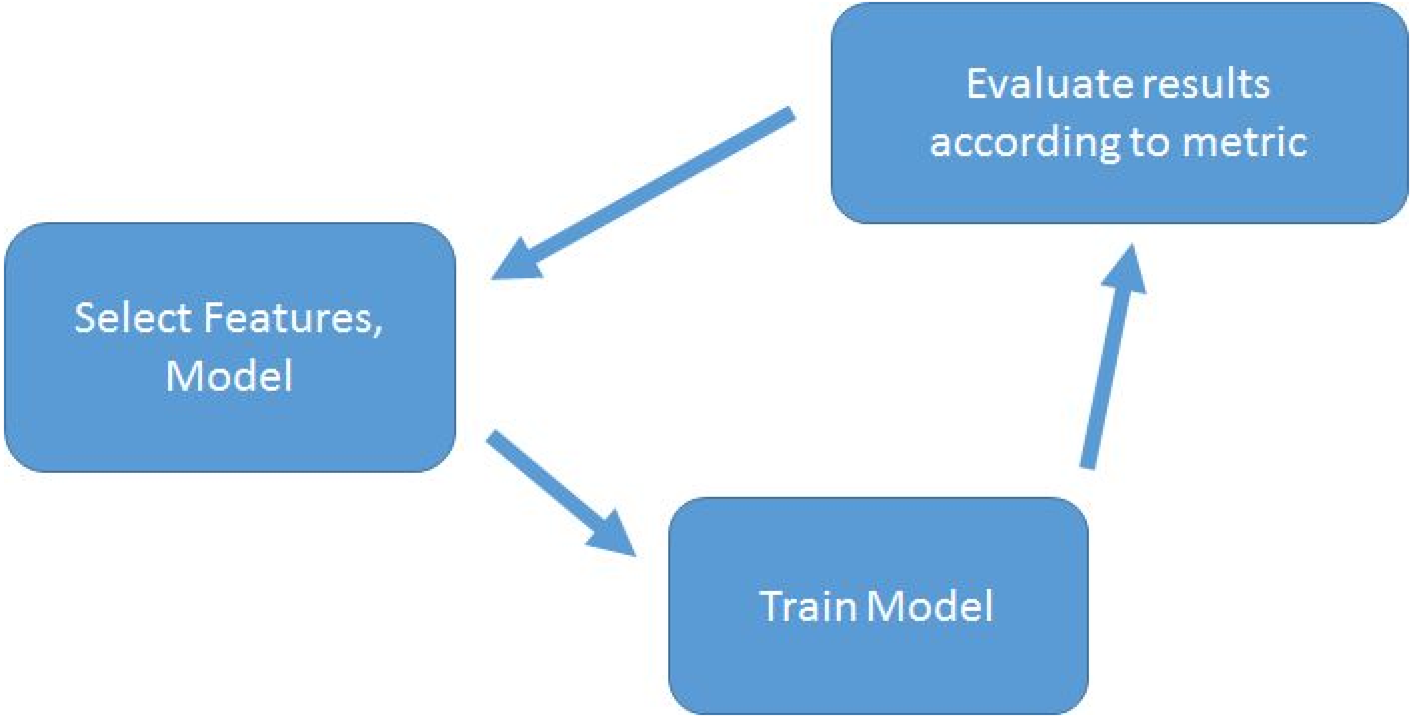
$$\mathbb{E}_{i(t)}[\ell(x_{i(t)}, y_{i(t)}; \theta_t)] = \frac{1}{n} \sum_i \ell(x_i, y_i; \theta_t).$$

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

Don’t Overdo it:



A DAY IN A DATA SCIENTIST’S LIFE



Evaluate results according to metric

Select Features, Model

Train Model

SELECTED BOOKS

Title	Author	About
Learning From Data	Yaser S. Abu-Mostafa	Small, good for beginners and has an online course
Machine Learning: A Probabilistic Perspective	Kevin P. Murphy	Larger, current, and very popular
The Elements of Statistical Learning: Data Mining, Inference, and Prediction	Trevor Hastie, Robert Tibshirani, Jerome Friedman	A lot of theory, and has a free PDF edition

SELECTED MACHINE LEARNING RESOURCES

Title	Author	Website
Machine Learning	Andy Ng	Coursera.org
Intro to Machine Learning	Sebastian Thrun	Udacity.com
DataQuest courses		DataQuest.io
Kaggle Competitions and Tutorials		Kaggle.com