Baby Steps to Being a Data Scientist

CodePot
29th August 2015

About me

Vladimir Alekseichenko

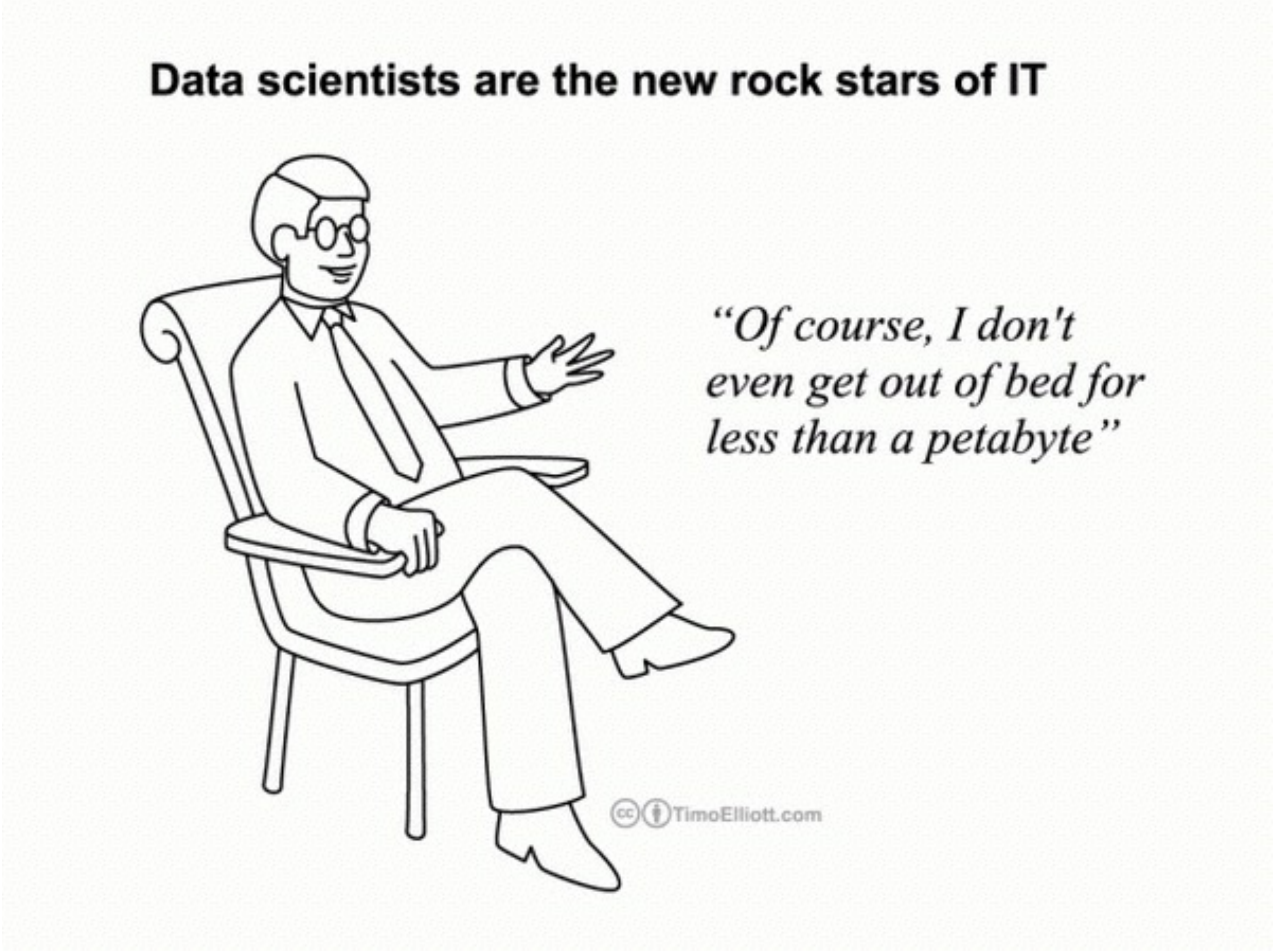
Love analyze data

Architect Search Platform

slon1024

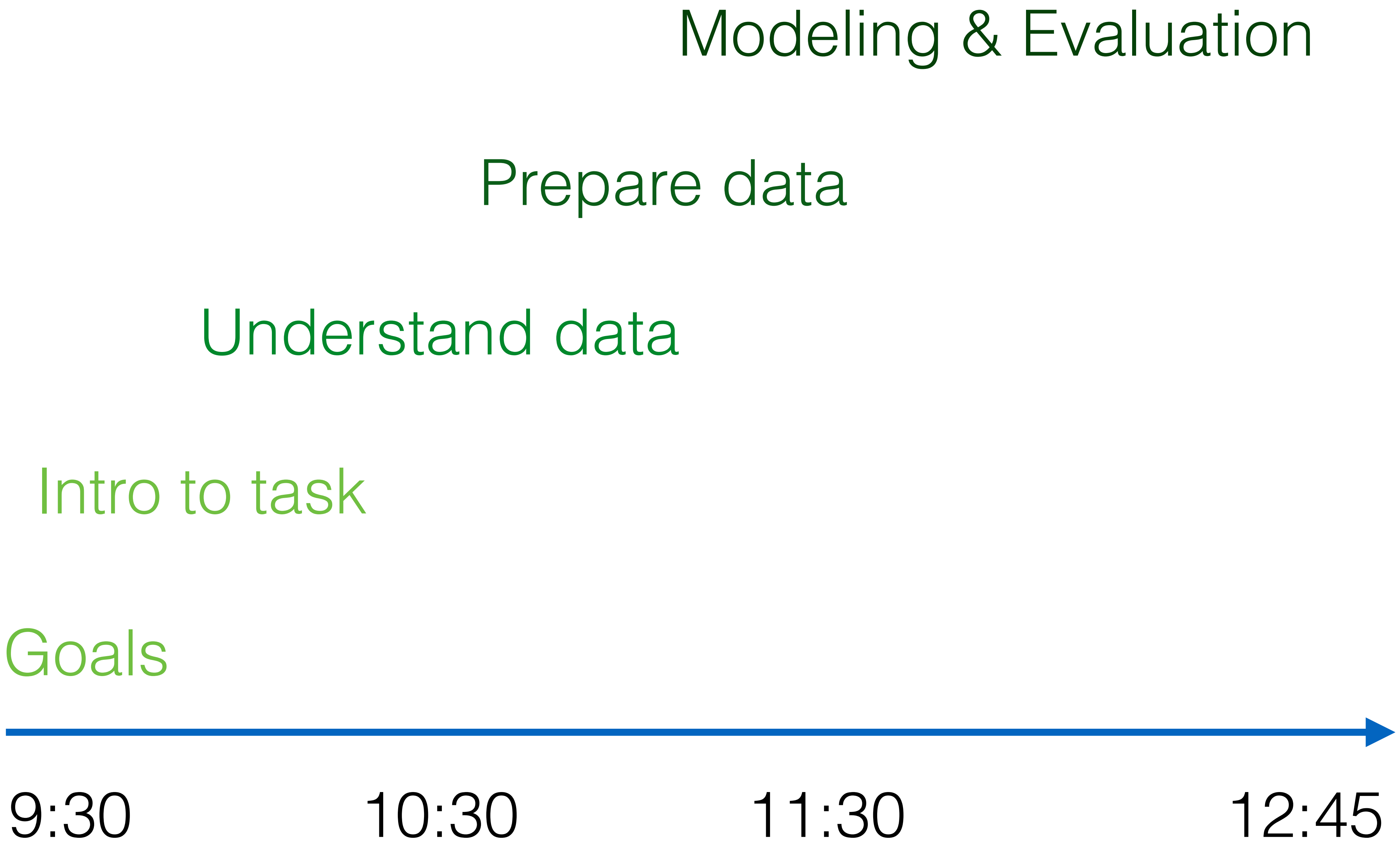
slon1024

GET IN LOSERS
WE’RE GOING TO
DO SCIENCE.


Data scientists are the new rock stars of IT
"Of course, I don't even get out of bed for less than a petabyte"
TimoElliott.com

Your Goals?

Plan



Modeling & Evaluation

Prepare data

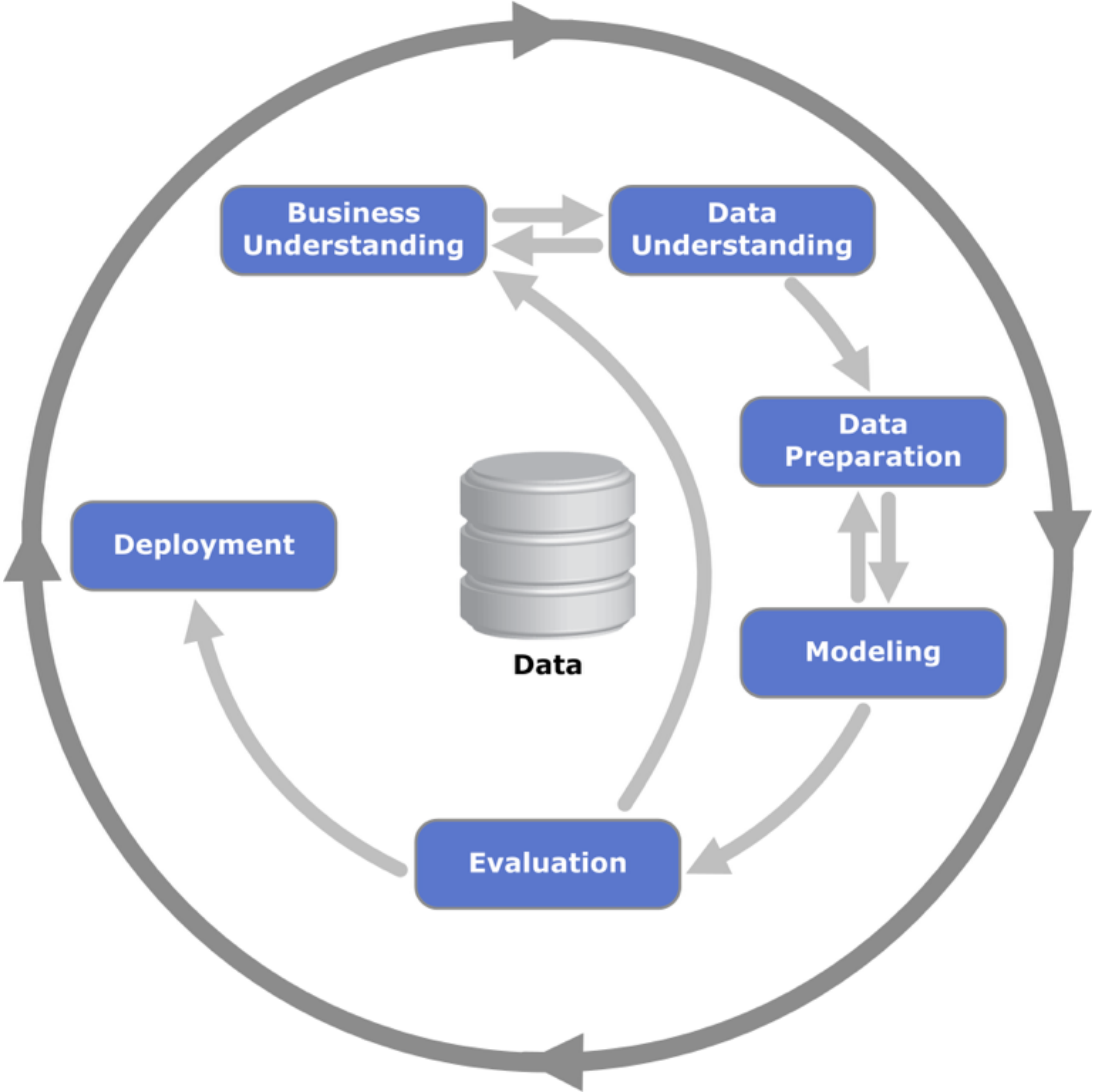
Understand data

Intro to task

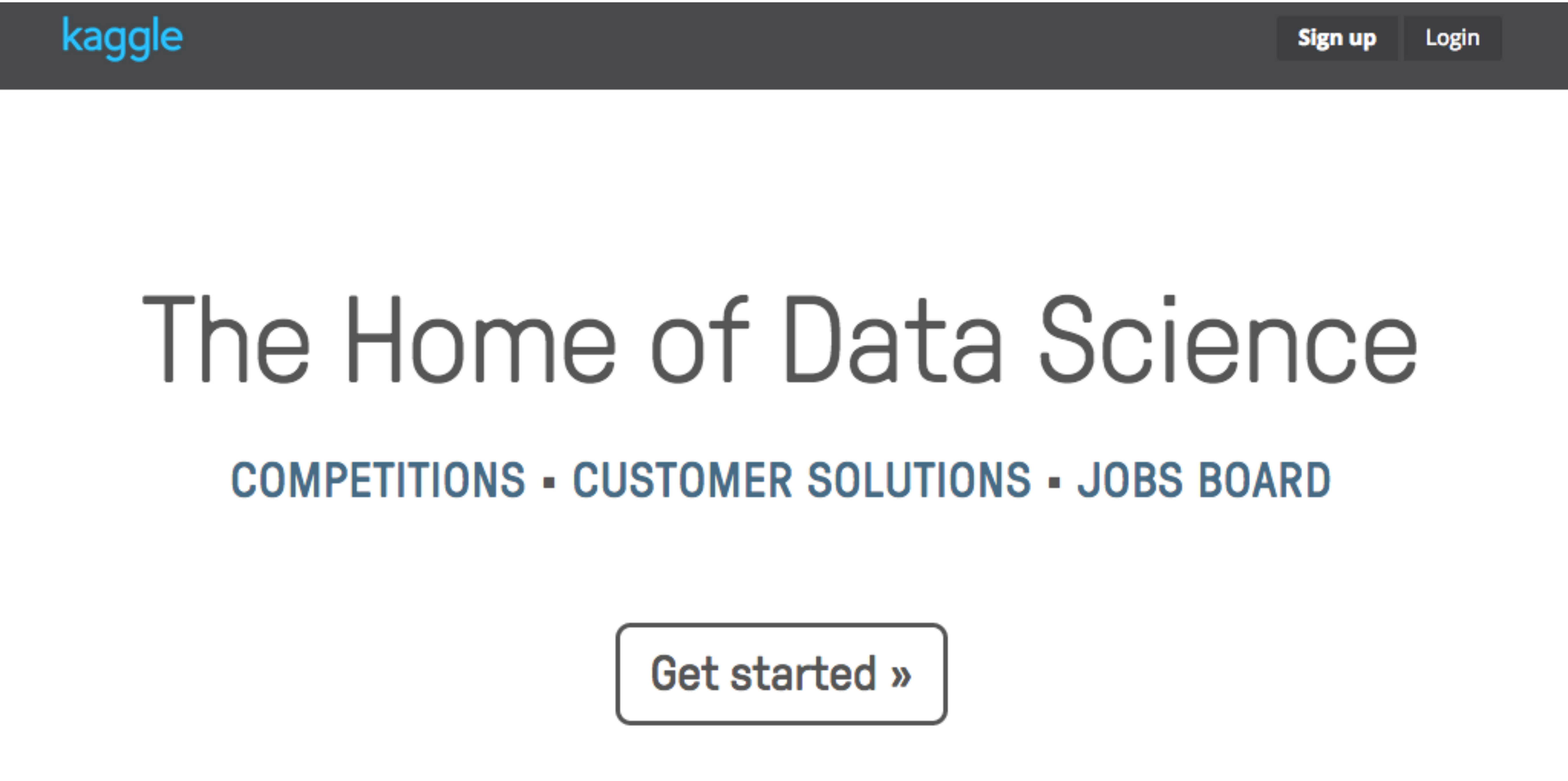
Goals

9:30 10:30 11:30 12:45

Process


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

How to practice?



The Home of Data Science

COMPETITIONS ▪ CUSTOMER SOLUTIONS ▪ JOBS BOARD

Bike Sharing Demand

kaggle.com/c/bike-sharing-demand

Tools

IPython notebook

pandas

numpy

scikit-learn

matplotlib

ggplot

seaborn

Example

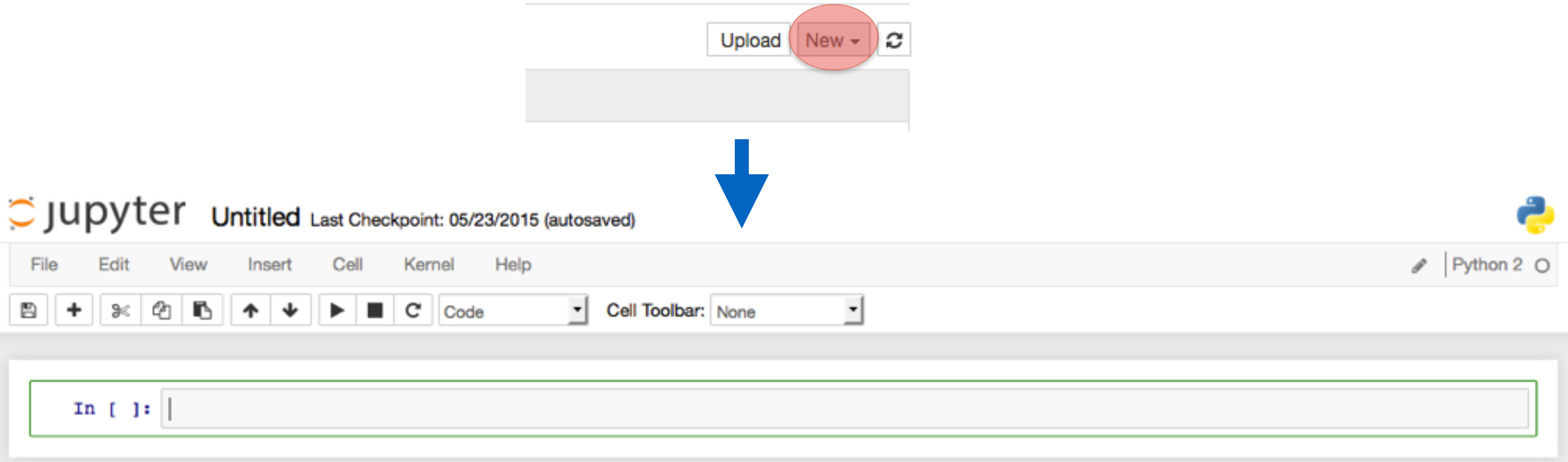
http://bit.ly/1MTw6pN

My Solution

http://bit.ly/1NENz6Z

ipython notebook

IPython notebook



Shortcuts:

1. **Ctrl + M + H** - help
2. **Ctrl + M + A** - a new cell above
3. **Ctrl + M + B** - a new cell bellow
4. **Shift + Enter** - run cell, select bellow

Understand Data

Understand Data
Input

- Read data from train.csv
- http://bit.ly/1MTq4FM

```
import pandas as pd

train = pd.read_csv(‘train.csv’)
```

Understand Data
Output

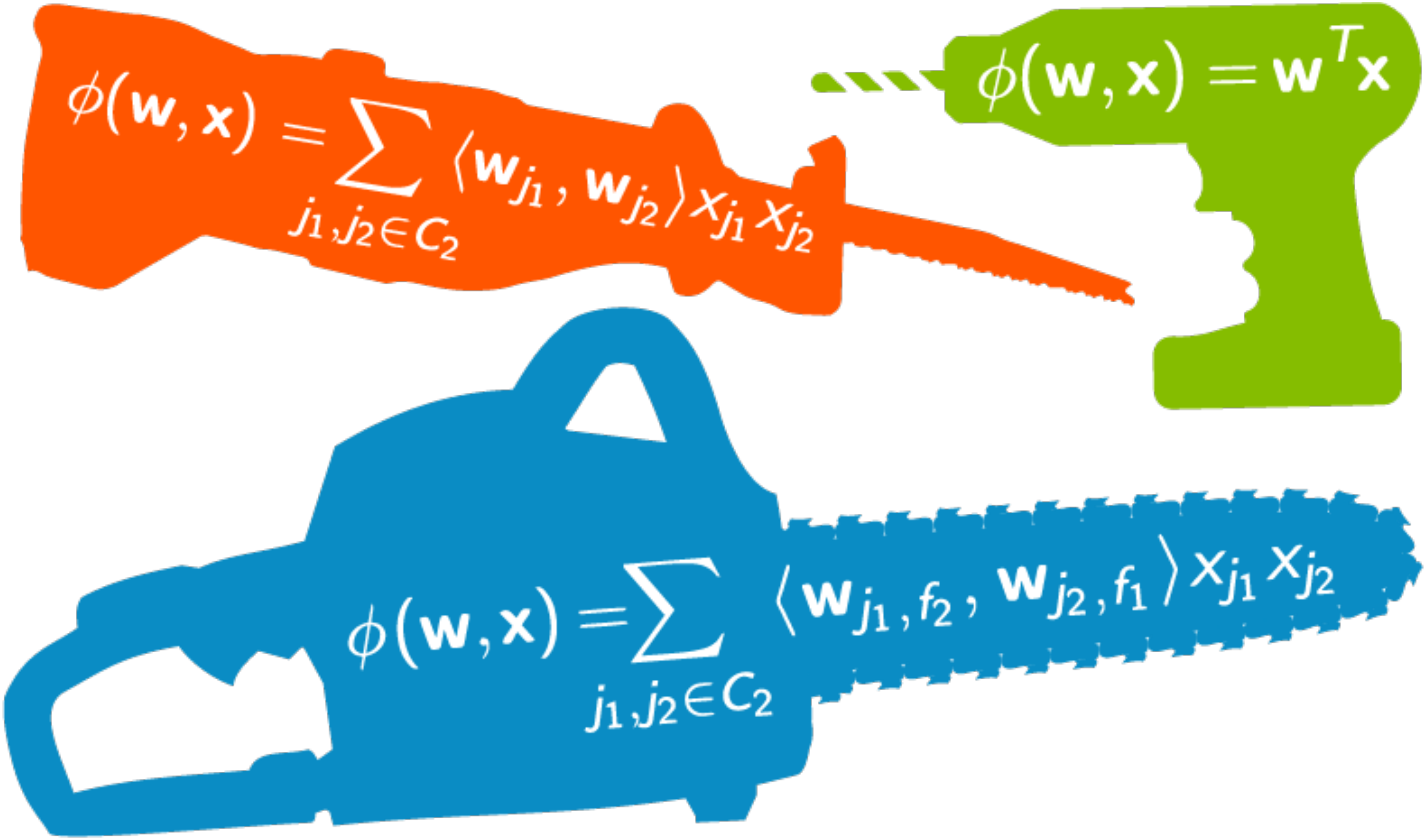
- Understand target (predict) value
- Statistic info about cols/rows
- Strategy about missing values

```
train.describe()

train.info()

#train.fillna(…)
```

Prepare Data

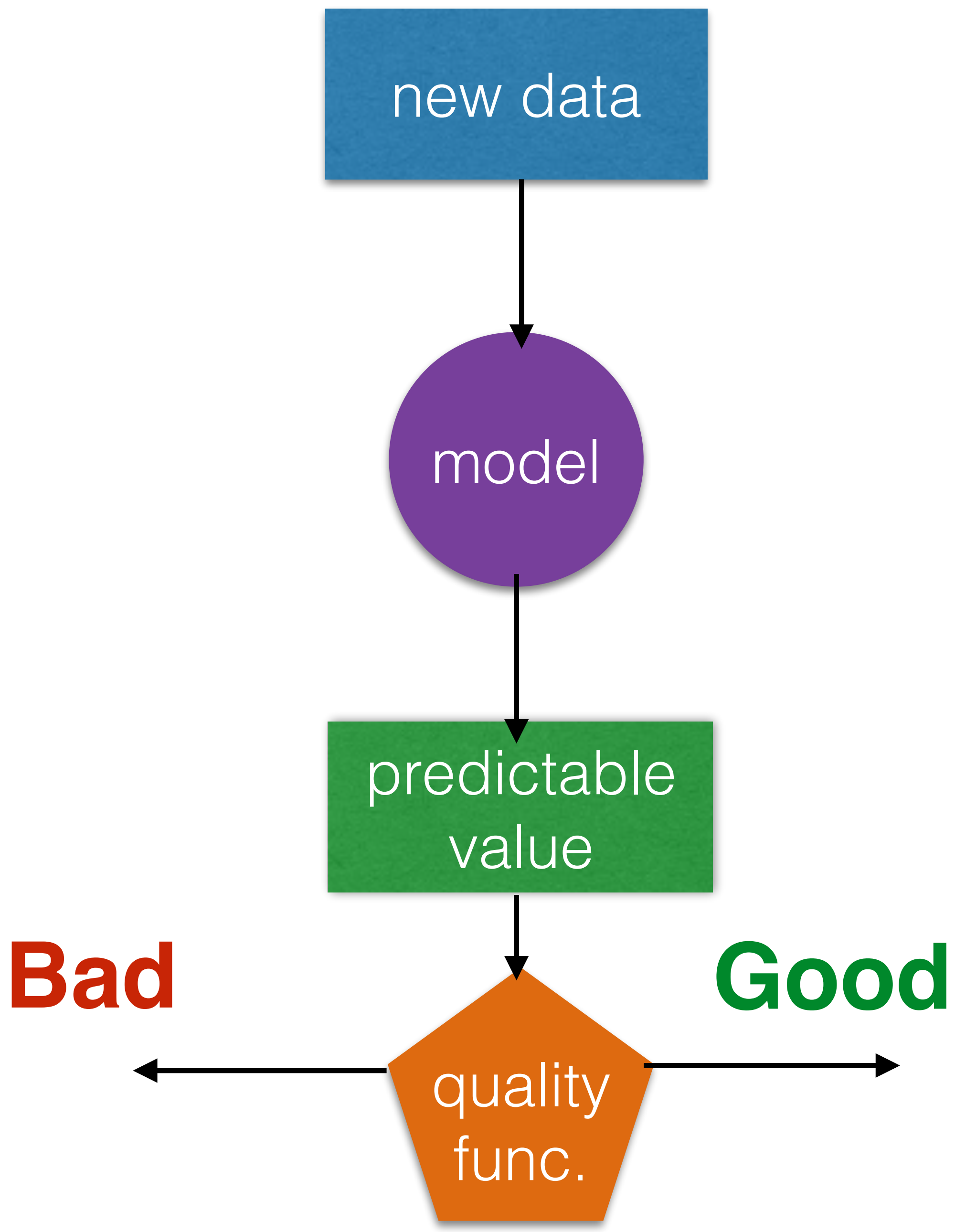

$\phi(\mathbf{w},\mathbf{x}) = \sum_{j_1,j_2\in C_2} \langle \mathbf{w}_{j_1}, \mathbf{w}_{j_2}\rangle x_{j_1} x_{j_2}$
$\phi(\mathbf{w},\mathbf{x}) = \mathbf{w}^T\mathbf{x}$
$\phi(\mathbf{w},\mathbf{x}) = \sum_{j_1,j_2\in C_2} \langle \mathbf{w}_{j_1,f_2}, \mathbf{w}_{j_2,f_1}\rangle x_{j_1} x_{j_2}$

Modeling & Evaluation

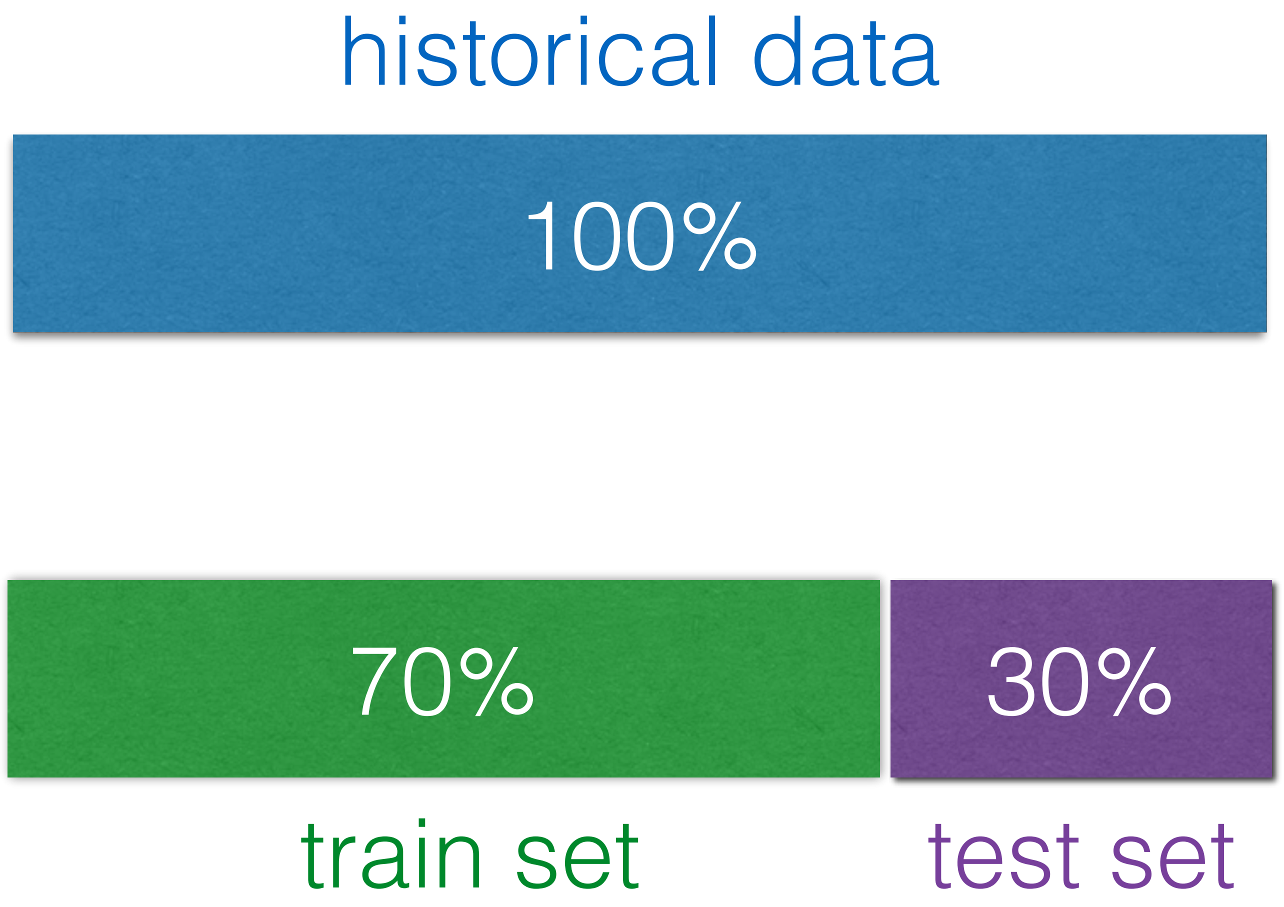
Madame Zaza
Fortune Teller
Madame Zaza
PREDICTIVE ANALYTICS
TimoElliott.com

"Why the change? Well, I could see where the future was going..."

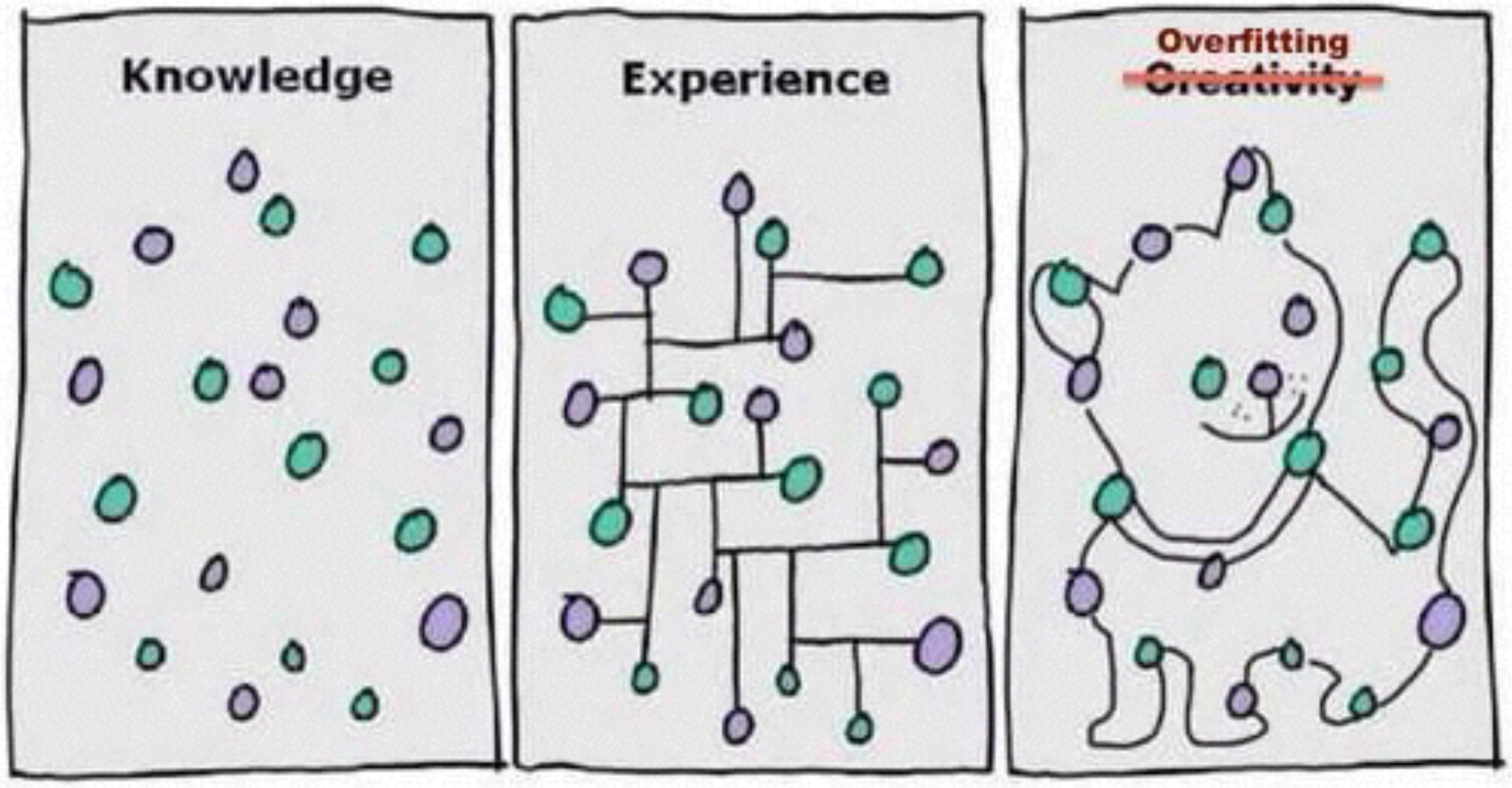
Build a model



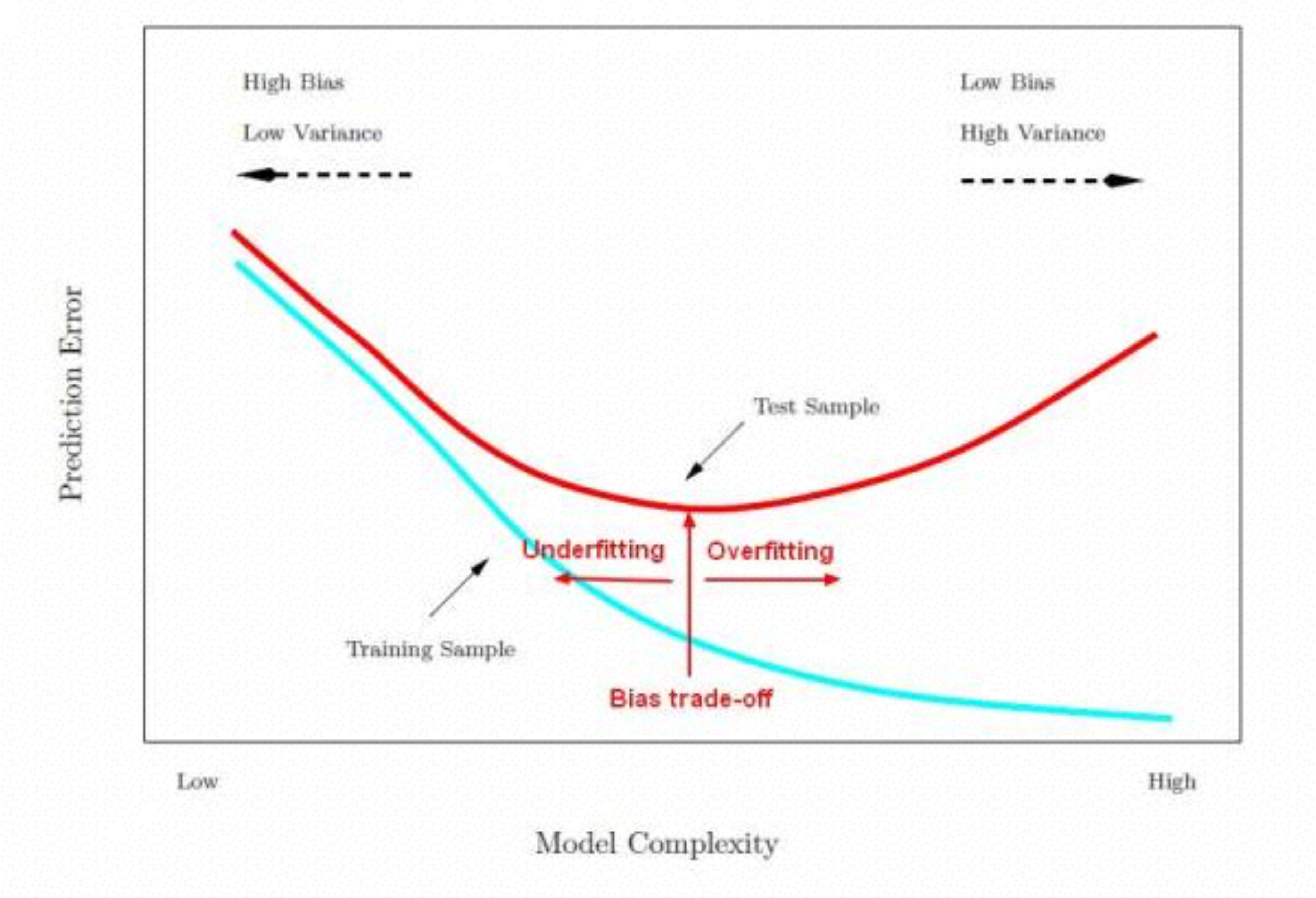
train set & **test** set



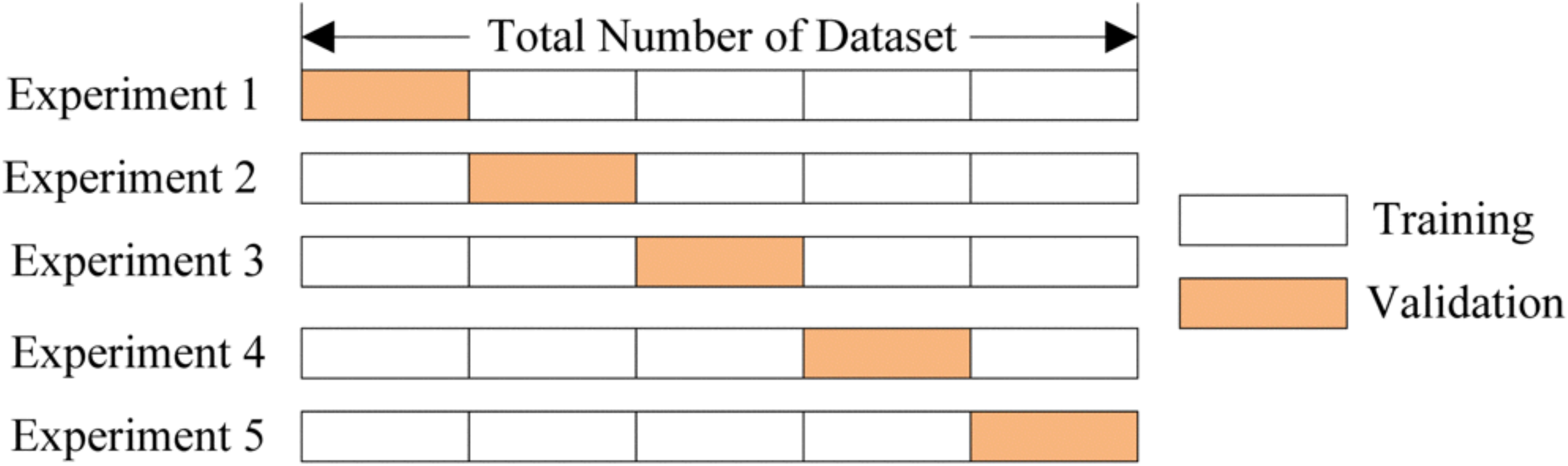
Overfitting



Overfitting



Cross validation



General example

```
#model = …

model.fit(X_train, y_train)

y_pred = model.predict(X_test)

#quality_model(y_pred, y_test)
```

http://scikit-learn.org/stable/modules/classes.html

Summary

- Understand your data (including a target value)
- Understand function of quality
- Experiment a lot :)