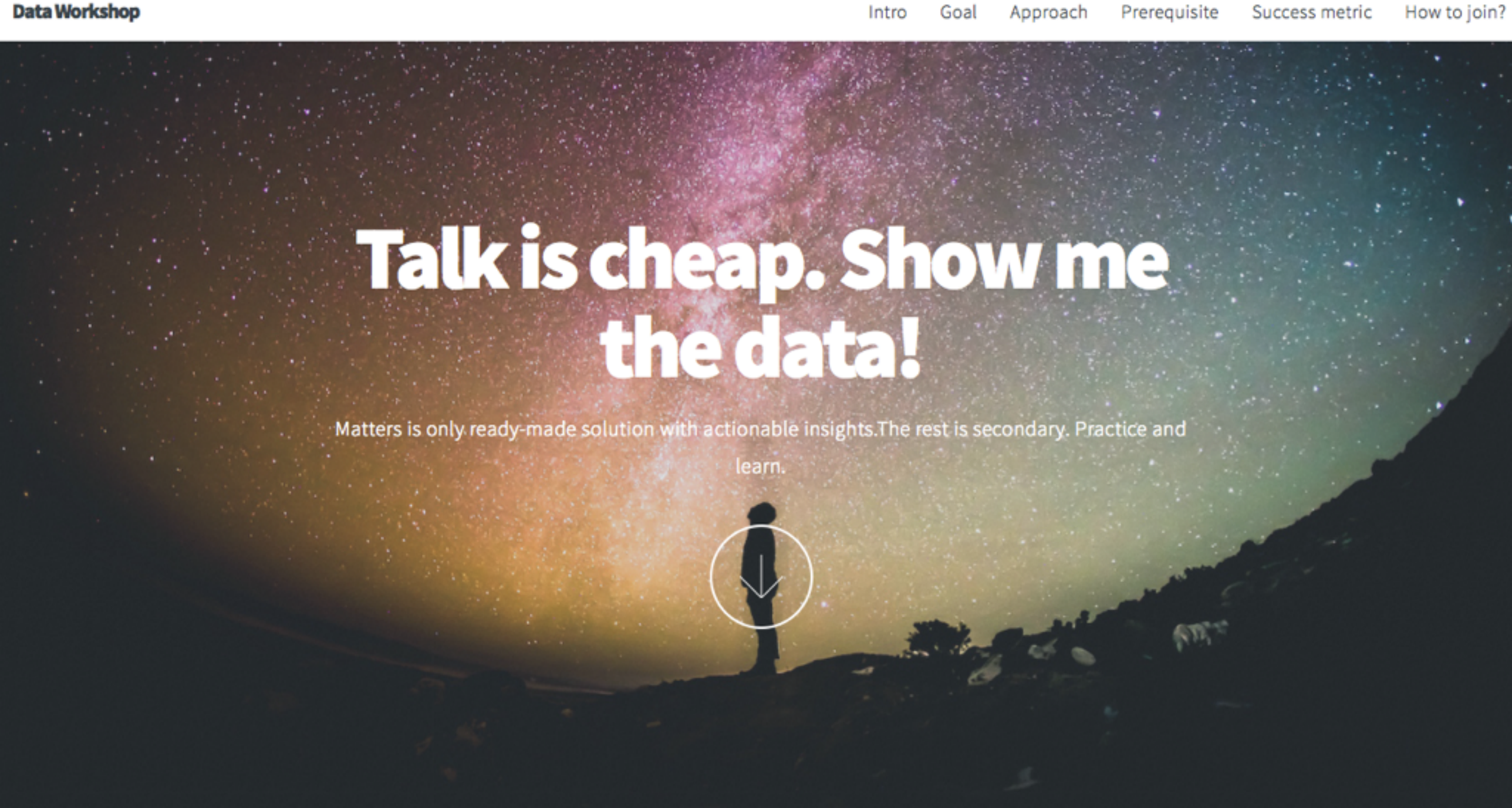
Data Workshop #4

Bike Sharing Demand

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

Environment

github.com/dataworkshop/prerequisite

github.com/dataworkshop/bike_sharing_demand

Packages

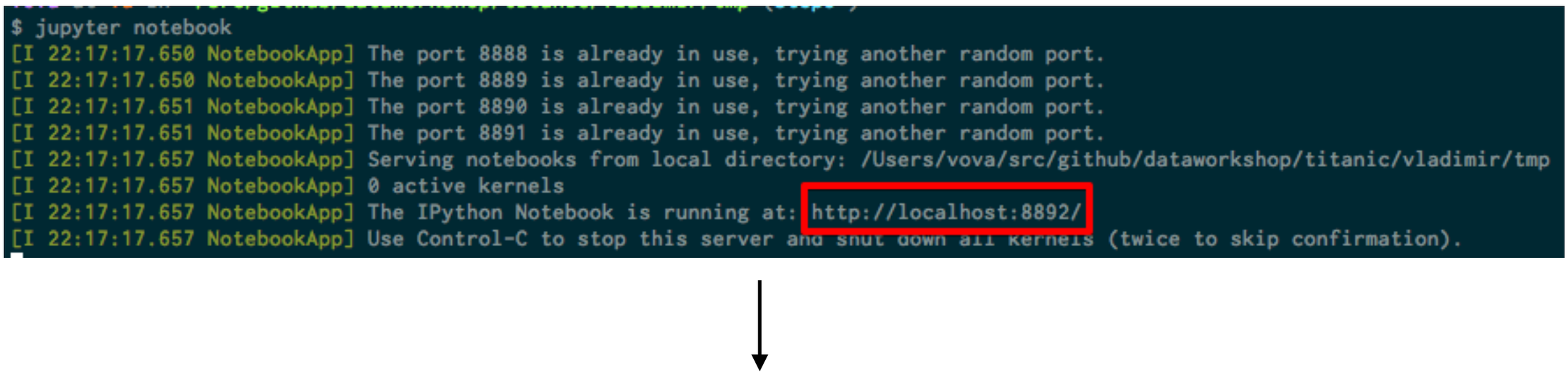
github.com/**dataworkshop/prerequisite**

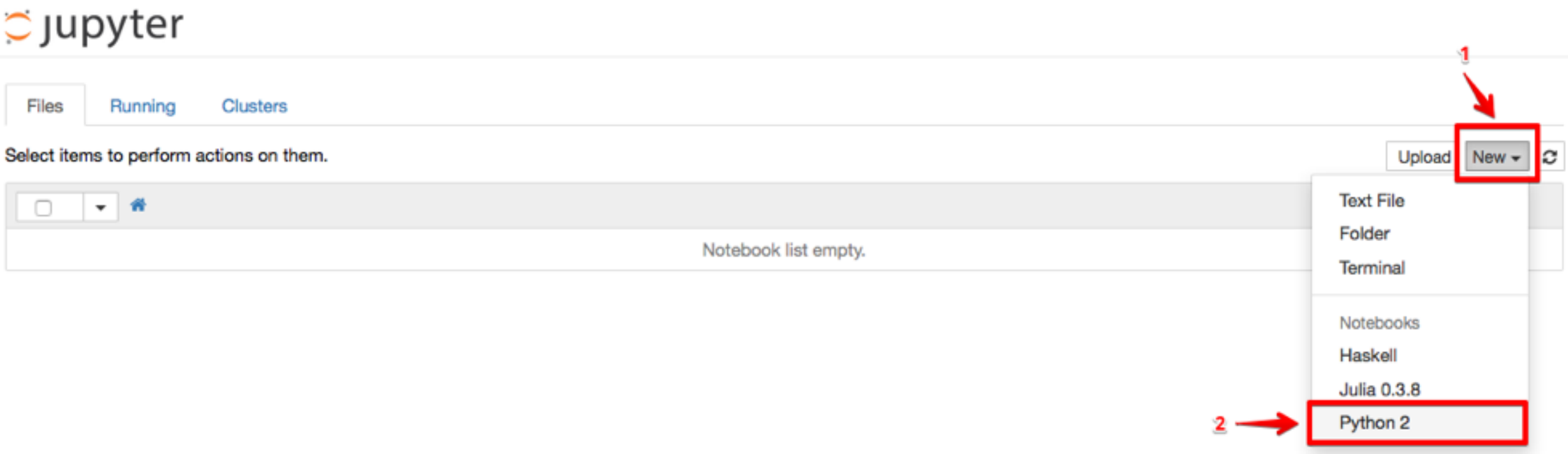
```
$ python run.py
seaborn-0.7.0 - OK
xgboost-0.4 - OK
matplotlib-1.5.1 - OK
IPython-4.1.2 - OK
numpy-1.11.0 - OK
pandas-0.18.0 - OK
sklearn-0.17.1 - OK

==================================================
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

Big Picture

Understand Business & Data

Read and explore data

Chanel_ID	Client_ID	Product_ID	Demand
1	3	2	10
1	3	5	15
3	3	6	12

Feature Engineering

Create a new ones based on already exists

Chanel_ID	Client_ID	Product_ID	…	Demand	DemandLog
1	3	2	…	10	2.303
1	3	5	…	15	2.708
3	3	6	…	12	2.485

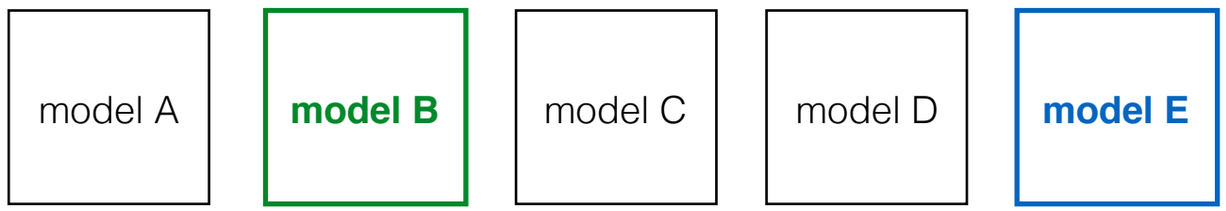
Feature Selection

Select only useful features

~~Chanel_ID~~	Client_ID	Product_ID	~~…~~	~~Demand~~	DemandLog
	3	2	…		2.303
	3	5	…		2.708
	3	6	…		2.485

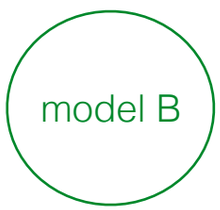
Model Selection

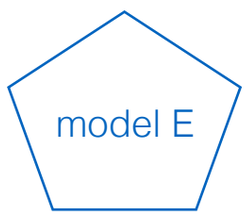
Find the best model(s)



Tuning Hyperparameters

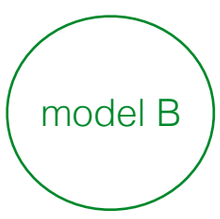
Find the best hyperparameters for given model





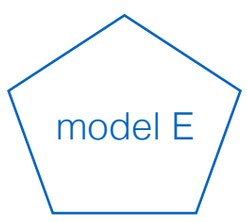
Ensemble Modeling

Combine few models into one more better



x0.6

+



x0.4

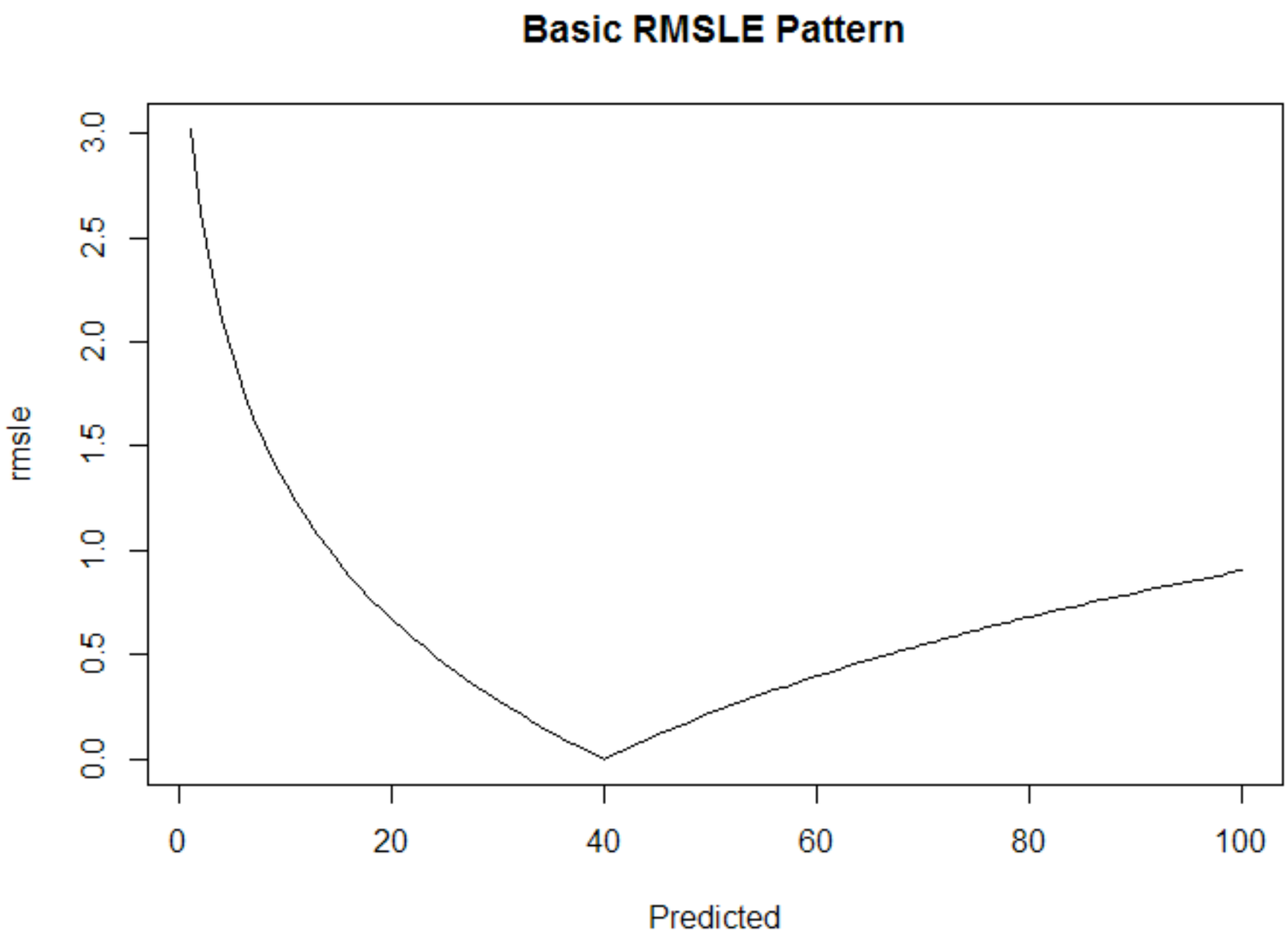
Understand Business & Data

Understand Metric

Simple points about metrics

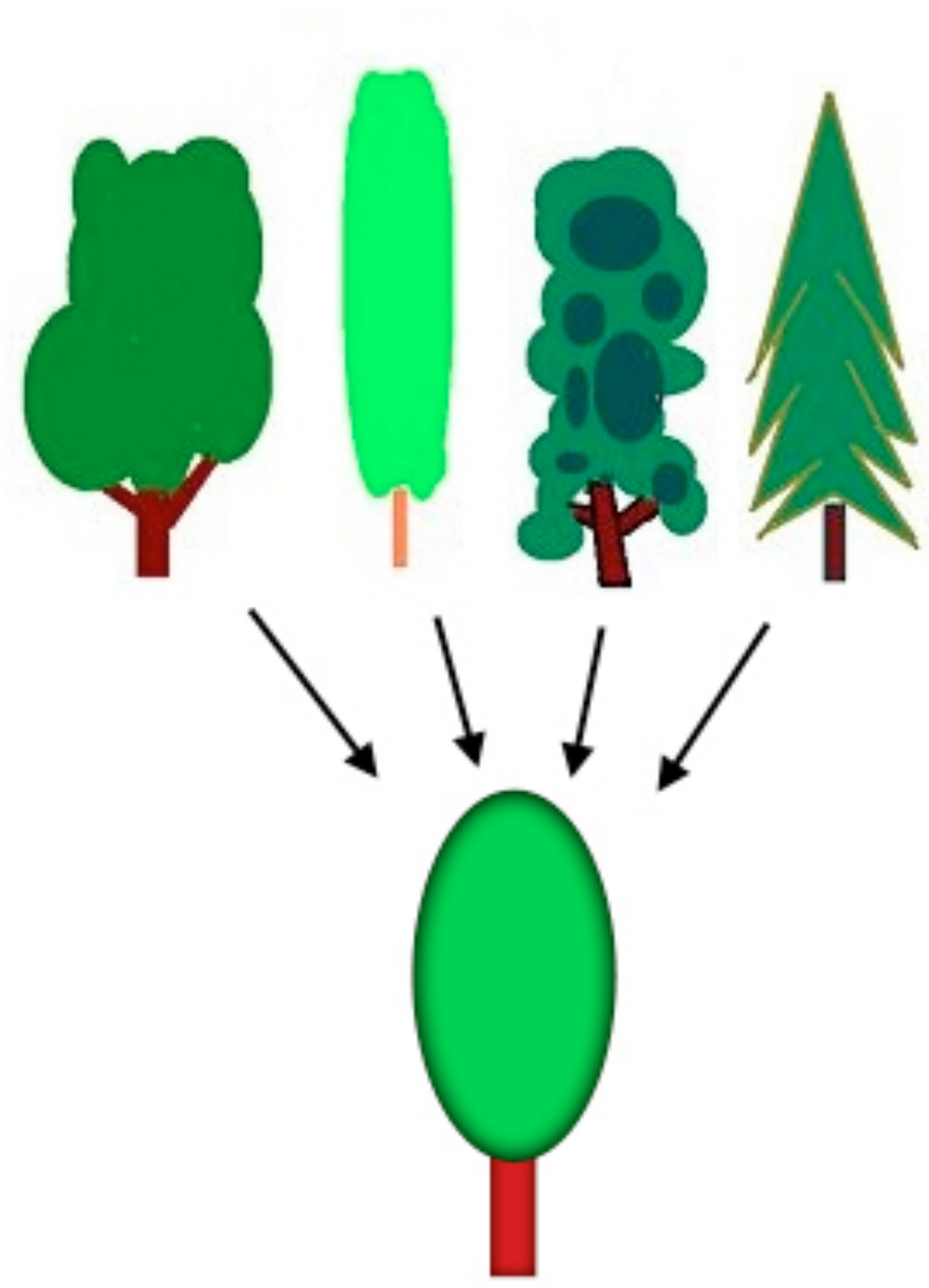
- Range
- Outliers
- Negative & Positive

RMSLE

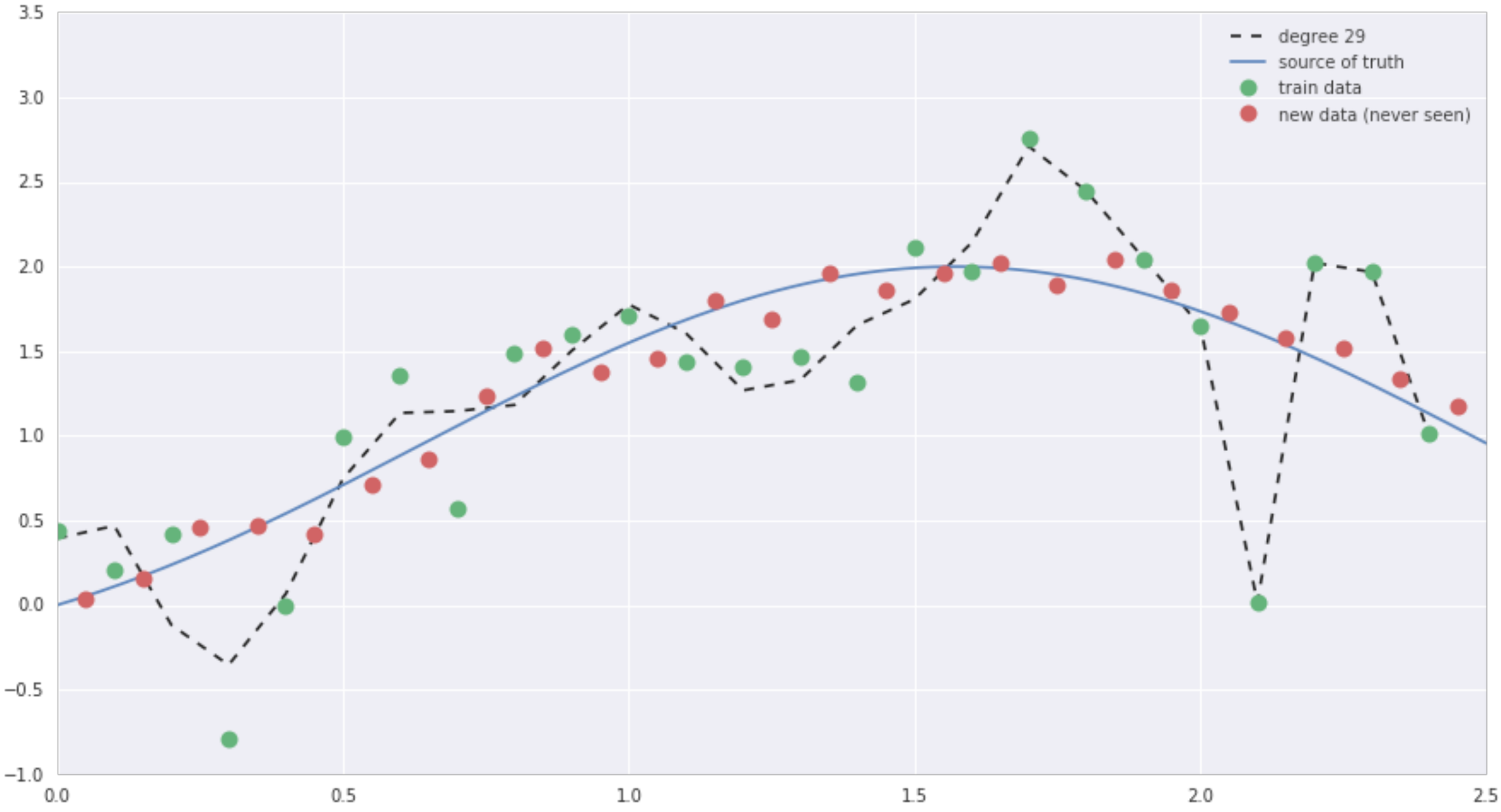

Basic RMSLE Pattern
rmsle
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
20
40
60
80
100
Predicted

Validation

Generalization



Overfitting



Summary

Three things

if you can remember only three…

- Understand your business and data
- Understand expected success by metric[s]
- Experiment a lot

Thank you