

Data Workshop #4

Bike Sharing Demand

dataworkshop.eu

DataWorkshop.eu

Data Workshop

[Intro](#)

[Goal](#)

[Approach](#)

[Prerequisite](#)

[Success metric](#)

[How to join?](#)

**Talk is cheap. Show me
the data!**

Matters is only ready-made solution with actionable insights. The rest is secondary. Practice and learn.



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



```
$ jupyter notebook
[I 22:17:17.650 NotebookApp] The port 8888 is already in use, trying another random port.
[I 22:17:17.650 NotebookApp] The port 8889 is already in use, trying another random port.
[I 22:17:17.651 NotebookApp] The port 8890 is already in use, trying another random port.
[I 22:17:17.651 NotebookApp] The port 8891 is already in use, trying another random port.
[I 22:17:17.657 NotebookApp] Serving notebooks from local directory: /Users/vova/src/github/dataworkshop/titanic/vladimir/tmp
[I 22:17:17.657 NotebookApp] 0 active kernels
[I 22:17:17.657 NotebookApp] The IPython Notebook is running at: http://localhost:8892/
[I 22:17:17.657 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
```



 jupyter

Files Running Clusters

Select items to perform actions on them.

Notebook list empty.

Upload New

Text File
Folder
Terminal

Notebooks
Haskell
Julia 0.3.8
Python 2

2

Motivation

Big Picture

Understand Business & Data

Read and explore data



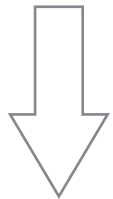
Feature Engineering

Create a new ones based on already exists



Feature Selection

Select only useful features



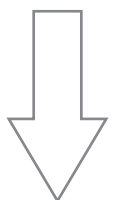
Model Selection

Find the best model(s)



Tuning Hyperparameters

Find the best hyperparameters for given model



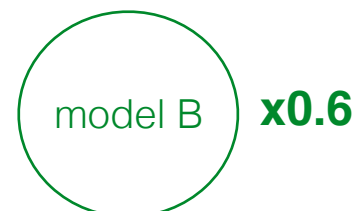
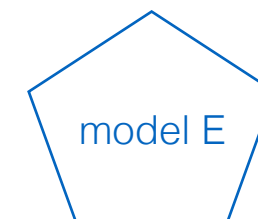
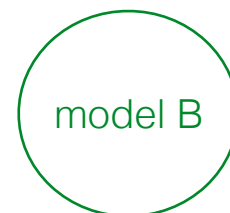
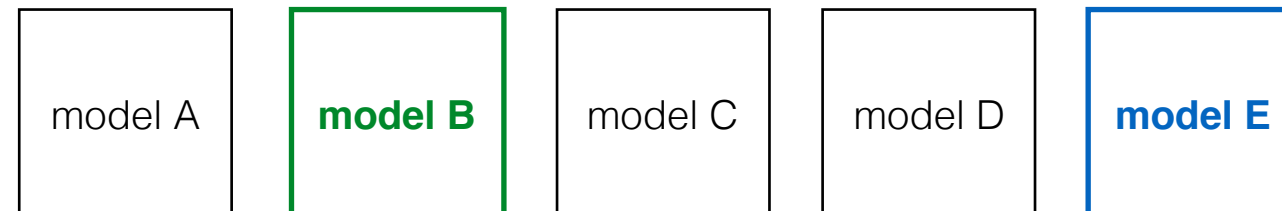
Ensemble Modeling

Combine few models into one more better

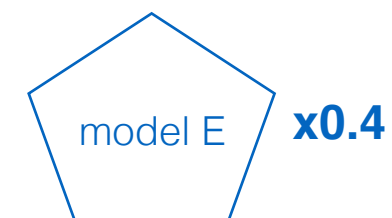
Chanel_ID	Client_ID	Product_ID	Demand
1	3	2	10
1	3	5	15
3	3	6	12

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

	Client_ID	Product_ID	...	DemandLog
	3	2	...	2.303
	3	5	...	2.708
	3	6	...	2.485



+



Understand Business & Data

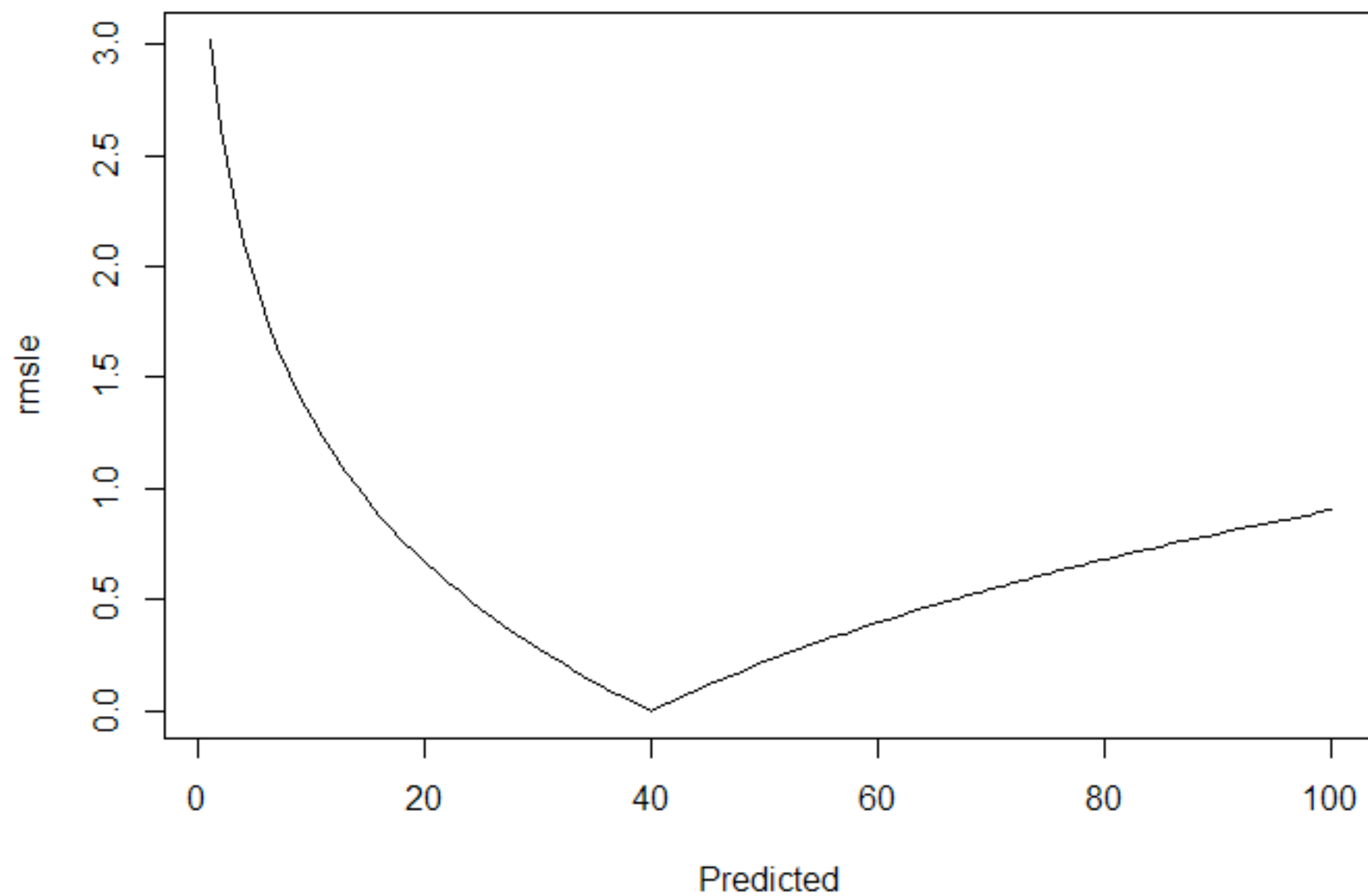
Understand Metric

Simple points about metrics

- Range
- Outliers
- Negative & Positive

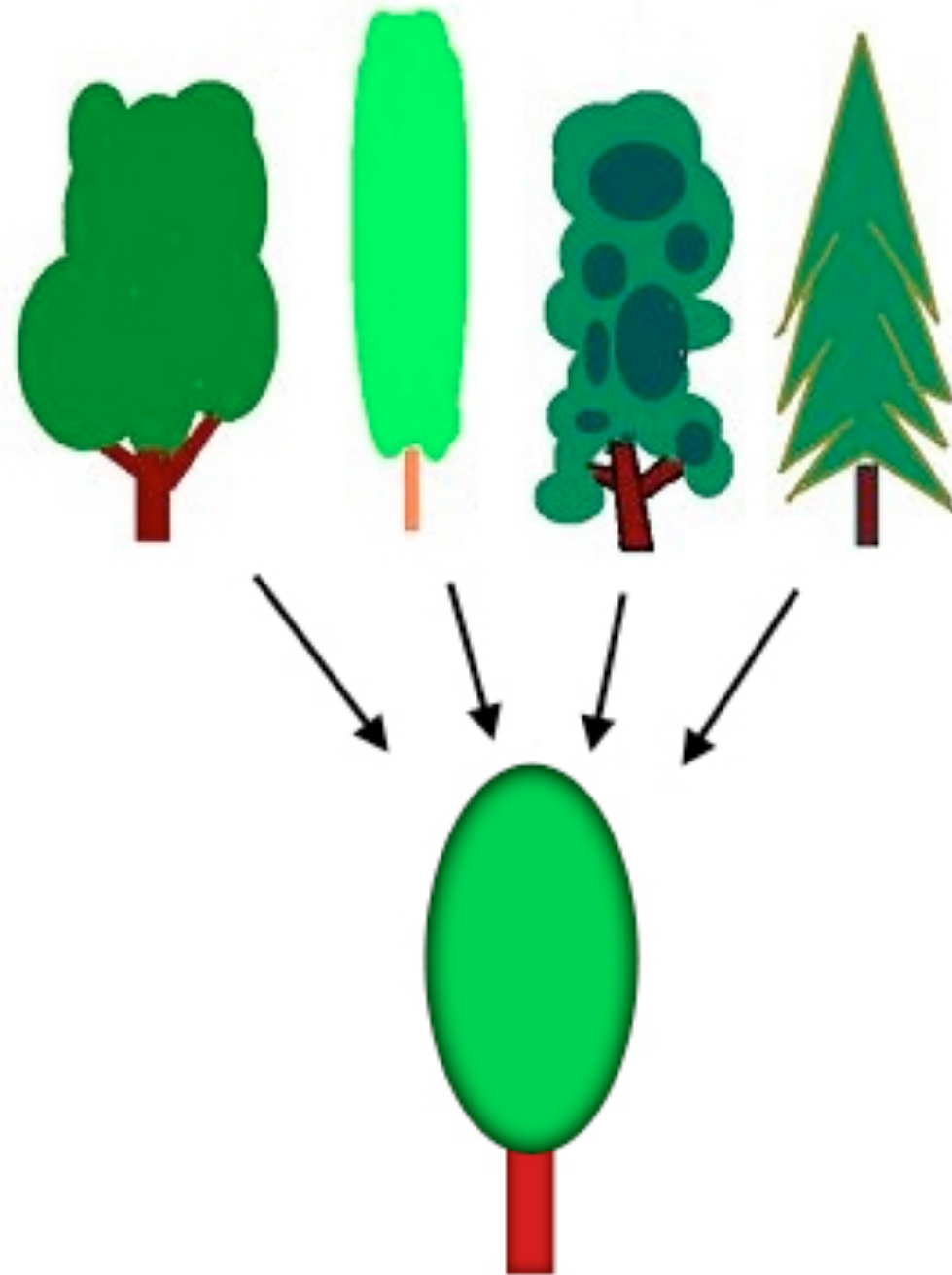
RMSLE

Basic RMSLE Pattern

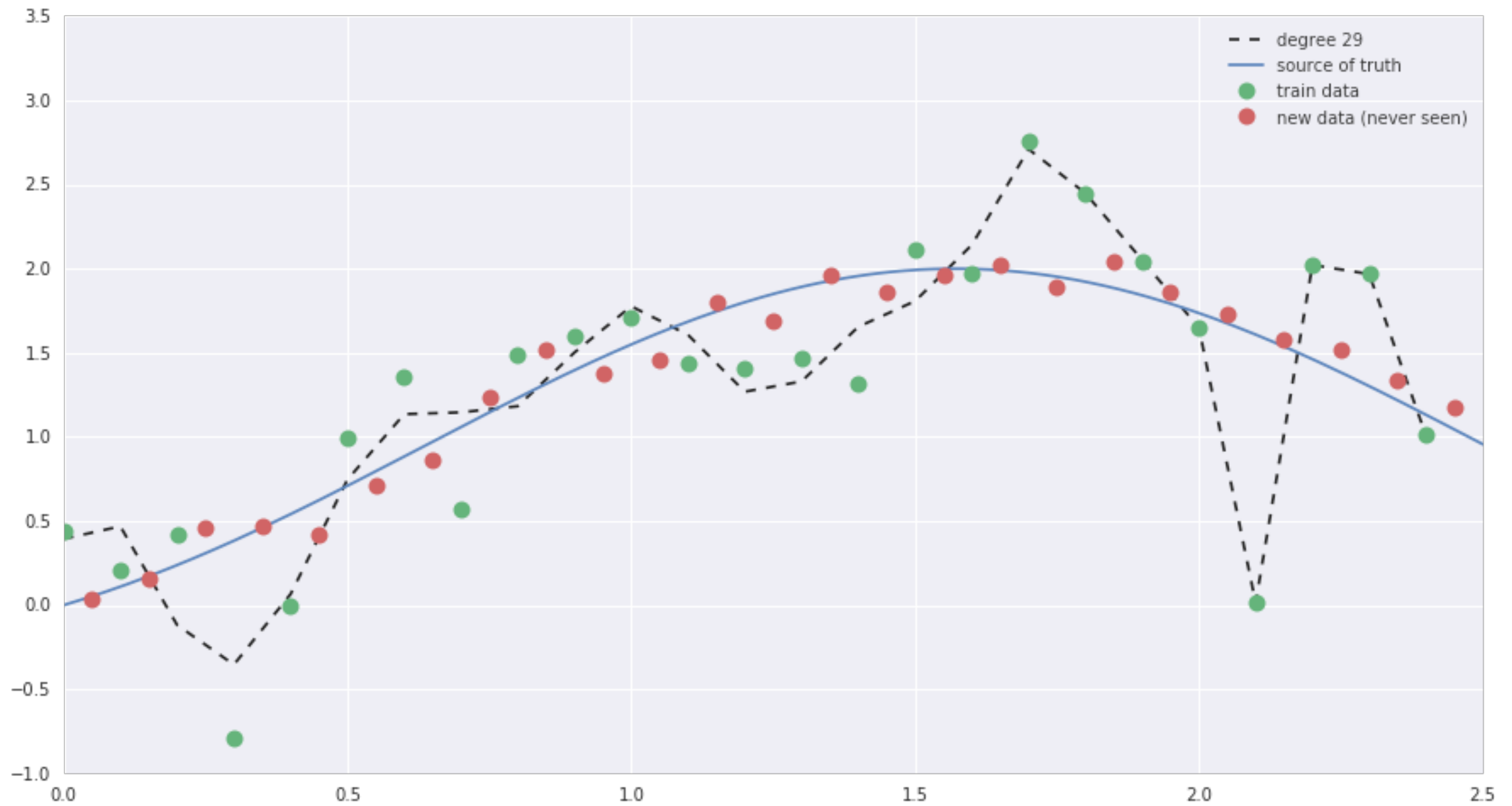


Validation

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