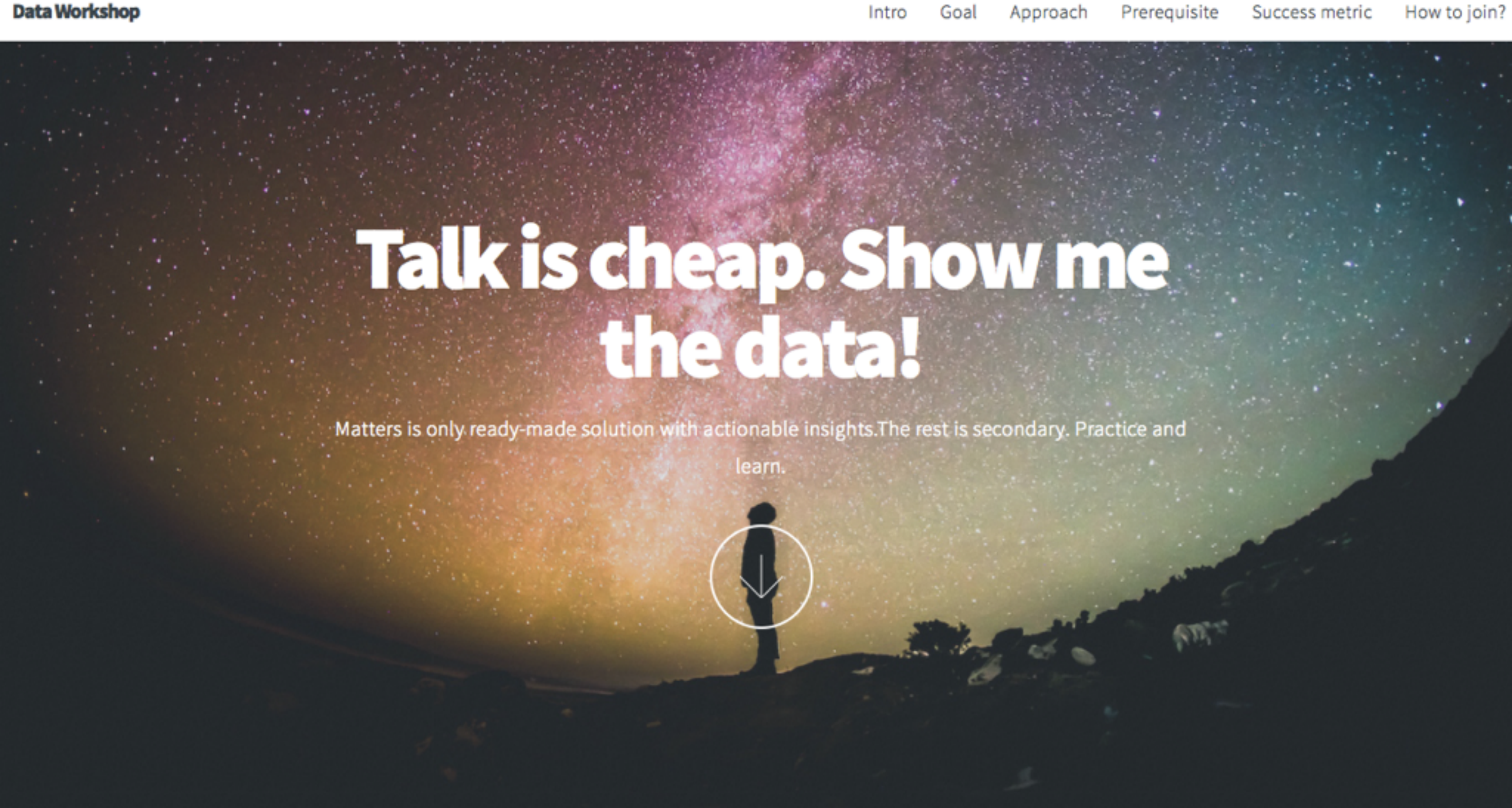
Data Workshop #1

dataworkshop.eu

Alekseichenko Vladimir

DataWorkshop.eu



About me

Vladimir Alekseichenko

Love analyze data

Architect

slon1024

slon1024

vova@vova.me

Agenda

- Prepare environment
- Build & evaluate a simple model
- while True:
  - Submit result to Kaggle
  - Improve model

Environment

github.com/**dataworkshop/prerequisite**

Packages

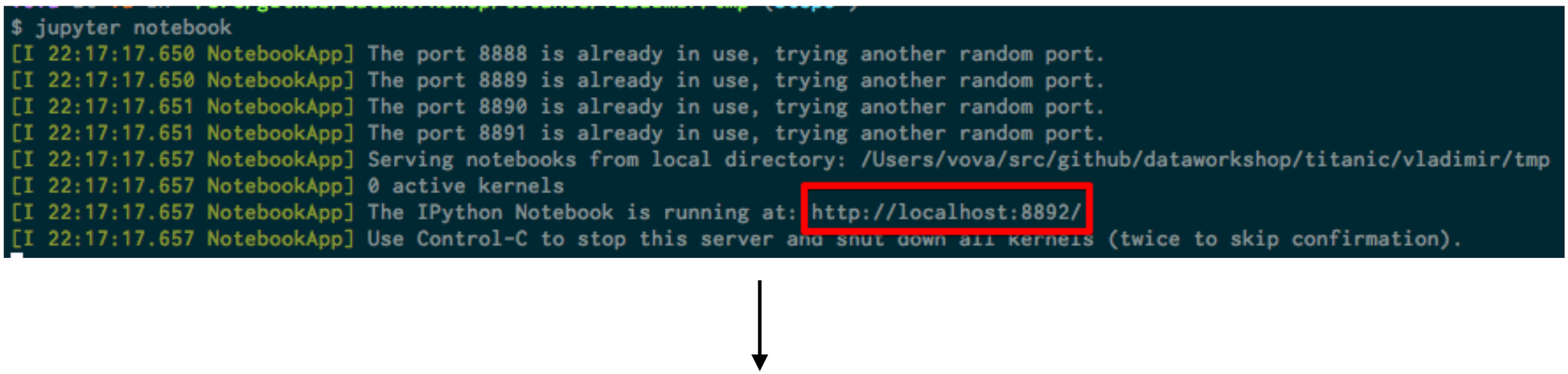
github.com/**dataworkshop/prerequisite**

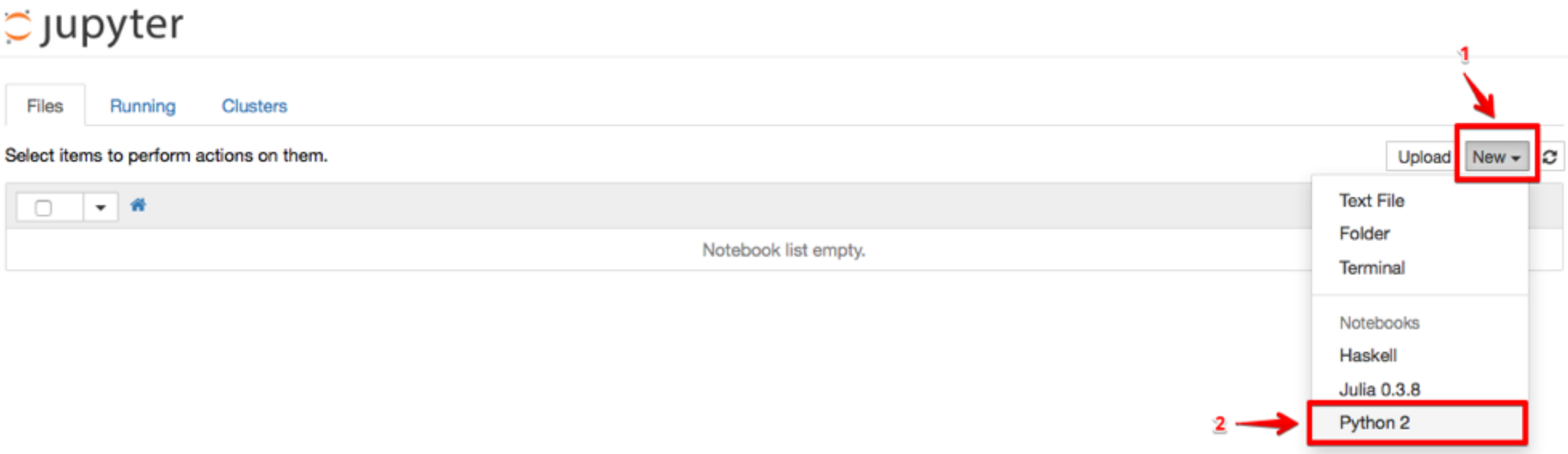
```
$ python run.py
seaborn-0.7.0 - OK
xgboost-0.4 - OK
matplotlib-1.5.1 - OK
IPython-4.1.2 - OK
numpy-1.11.0 - OK
pandas-0.18.0 - OK
sklearn-0.17.1 - OK

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





API review

sklearn

- .fit(X_train, y_train)
    - .transform(X)
- .predict(X_test)

Build a simple model

github.com/**dataworkshop/titanic**

Titanic

Titanic - what happened?

https://www.youtube.com/watch?v=FSGeskFzE0s

- **PassengerId** -- A numerical id assigned to each passenger.
- **Survived** -- Whether the passenger survived (1), or didn't (0). Target variable.
- **Pclass** -- The class the passenger was in -- first, second and third.
- **Name** -- the name of the passenger.
- **Sex** -- The gender of the passenger -- male or female.
- **Age** -- The age of the passenger. Fractional.
- **SibSp** -- The number of siblings and spouses the passenger had on board.
- **Parch** -- The number of parents and children the passenger had on board.
- **Ticket** -- The ticket number of the passenger.
- **Fare** -- How much the passenger paid for the ticker.
- **Cabin** -- Which cabin the passenger was in.
- **Embarked** -- Where the passenger boarded the Titanic.

Build more advanced model

github.com/**dataworkshop/titanic**

How to handle with…

- Missing values
- Categorical variables
- Pipeline
- Hyperparameter optimisation

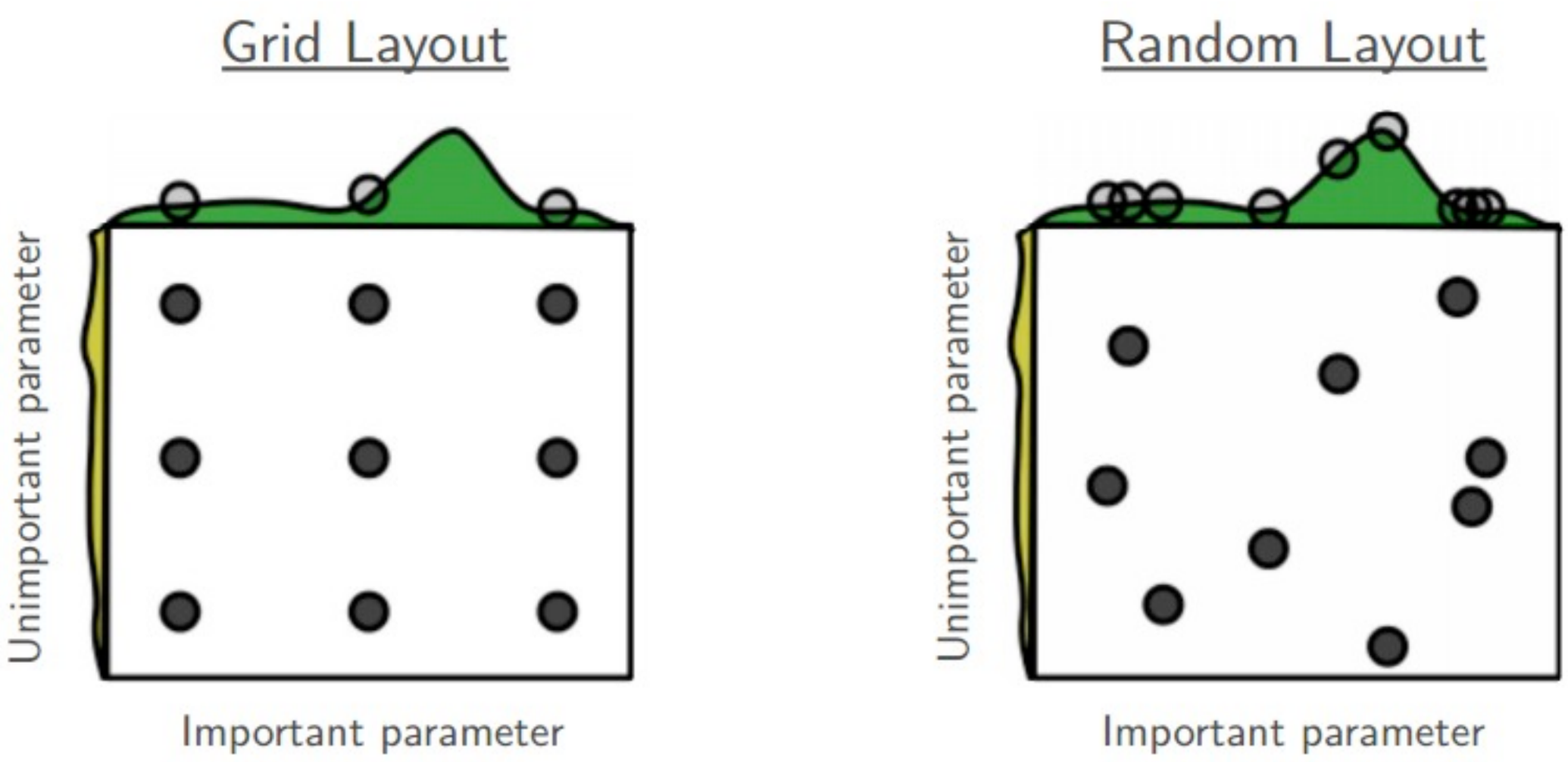
Missing values

- Remove rows () or columns (features)
- Replace by value (-1, 0 and so on)
  - `pandas.fillna(value)`
- Use mean, median, most frequent
  - `sklearn.preprocessing.Imputer`

Categorical Variables

- Convert into unique ID (integer)
  - “a”, “b”, “c” => 1, 2, 3
    sklearn.preprocessing.**LabelEncoder**
- One hot encoding
  - “a”, ”b”, “c” => [1, 0, 0], [0, 1, 0], [0, 0, 1]
    pandas**.get_dummies**
- Probability
  - (#samples == {“a”, “b”,”c”}) / #samples ALL

Hyperparameter optimization



http://blog.kaggle.com/2015/07/16/scikit-learn-video-8-efficiently-searching-for-optimal-tuning-parameters/

Summary

Best results…
but this **not** about ML :)

#	Δ1w	Team Name * in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	—	Prabhudatta Das *	1.00000	1	Wed, 24 Feb 2016 15:14:43
2	—	tryn2win	1.00000	16	Wed, 02 Mar 2016 16:15:52 (-47.1h)
3	—	norwayT	1.00000	1	Thu, 03 Mar 2016 02:24:21
4	—	determinedEurasians	1.00000	1	Sat, 05 Mar 2016 15:25:43
5	—	Formandos Puc Barreiro	1.00000	2	Sun, 06 Mar 2016 17:13:36 (-22.7h)
6	—	Nijida	1.00000	4	Tue, 22 Mar 2016 08:43:34 (-12.9d)
7	—	Eric_Chen	1.00000	1	Thu, 17 Mar 2016 06:46:43
8	—	Sajid Umair 2	1.00000	1	Sun, 27 Mar 2016 13:10:32
9	—	WooJungKim	1.00000	3	Sat, 02 Apr 2016 16:58:46
10	—	DaiHeping	1.00000	13	Wed, 06 Apr 2016 09:52:51
11	↑4	Prerit Ahuja	1.00000	2	Mon, 18 Apr 2016 12:48:26

Three things

if you can remember only three…

- **Feature engineering** is important

- Manage **categorical variables** and **missing values** is important as well :)

- **Ensemble** is important

What Next?

Active Competitions

State Farm Distracted Driver Detection Can computer vision spot distracted drivers?	**3 months** **410** teams **235** scripts **$65,000**
Santander Customer Satisfaction Which customers are happy customers?	**9.8 days** **4662** teams **3129** scripts **$60,000**
Home Depot Product Search Relevance Predict the relevance of search results on homedepot.com	**2.8 days** **2147** teams **1648** scripts **$40,000**
Expedia Hotel Recommendations Which hotel type will an Expedia customer book?	**48 days** **313** teams **371** scripts **$25,000**

Thank you