Kaggle workshop: Shelter Animal Outcomes

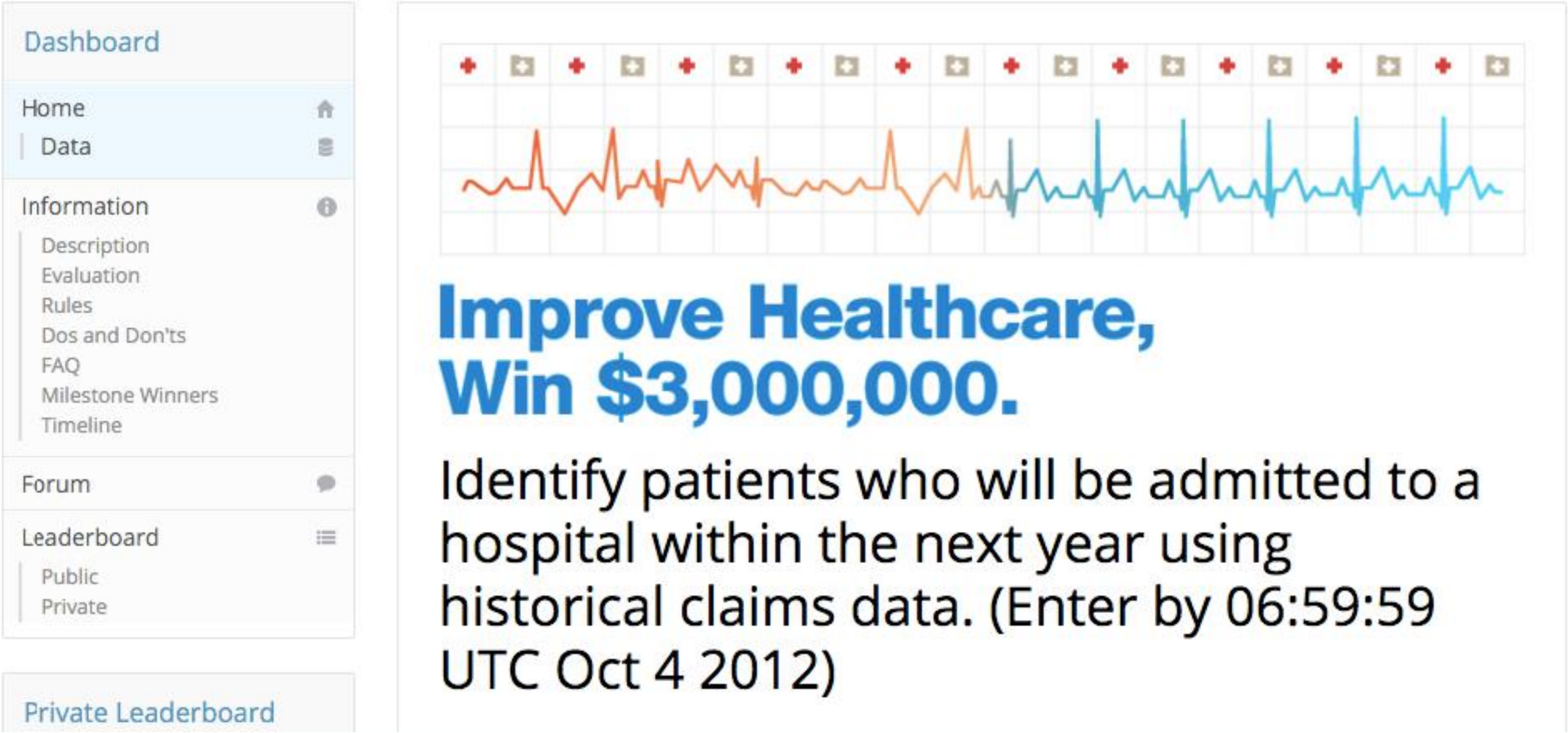
Ilya Ezepov

Agenda

- Intro to Kaggle

- Data Analysis routine:
  - Exploratory data analysis
  - Feature engineering
  - Machine learning
  - Ensemble construction

Kaggle

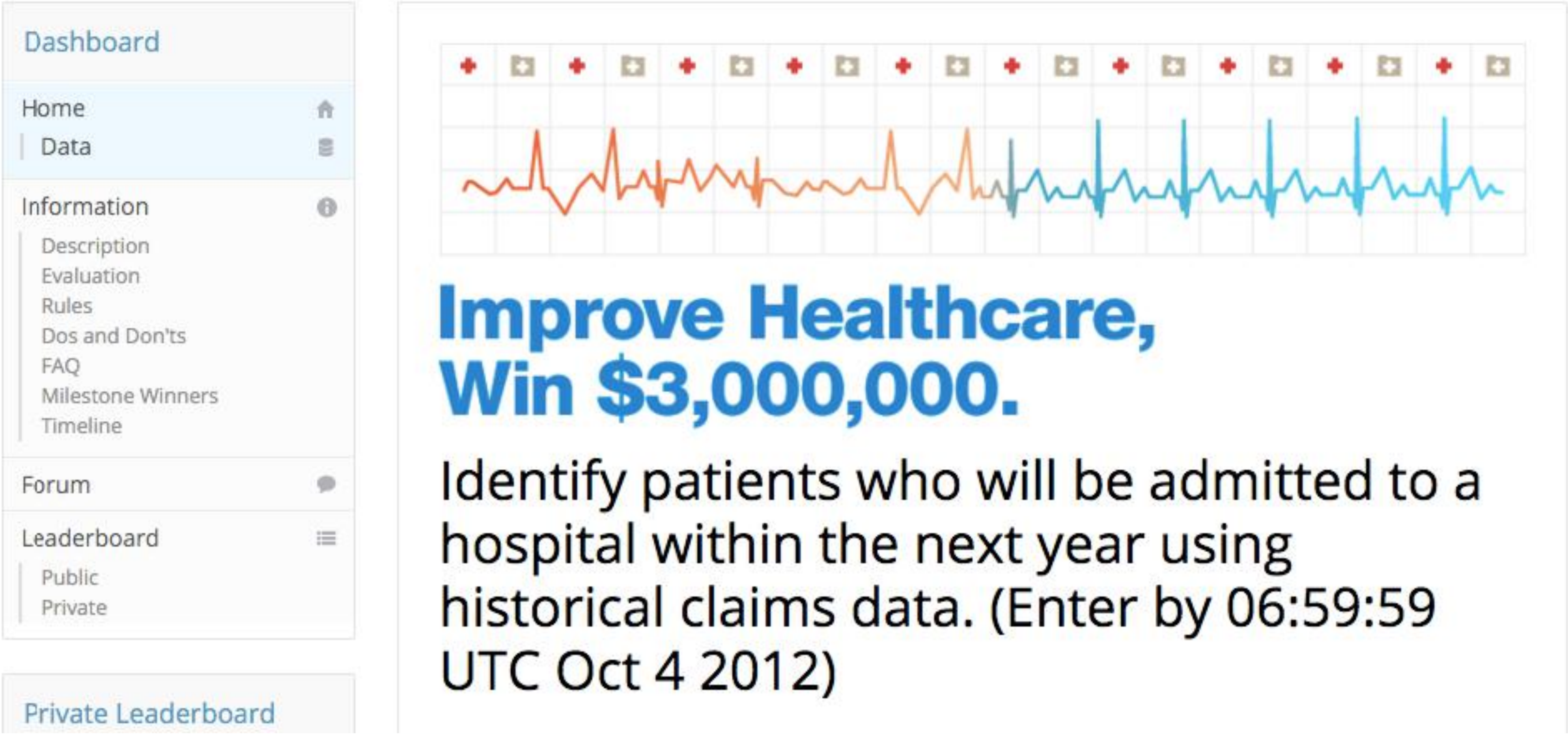


Improve Healthcare, Win $3,000,000.

Identify patients who will be admitted to a hospital within the next year using historical claims data. (Enter by 06:59:59 UTC Oct 4 2012)

As of July 2015, Kaggle claims approximately 332,000 data scientists on its job boards.

Kaggle



Improve Healthcare,
Win $3,000,000.

Identify patients who will be admitted to a hospital within the next year using historical claims data. (Enter by 06:59:59 UTC Oct 4 2012)

As of July 2015, Kaggle claims approximately 332,000 data scientists on its job boards.

Idea: In 1998 Rob McEwen asked data scientist for $500,000 to find best places to mine gold. In a year he got $3 billion .

Shelter Animal Outcomes

- Due to the public nature of the data, this competition does not count towards Kaggle ranking points.

- We ask that you respect the spirit of the competition and do not cheat. You should not submit entries based on test-set answers or train your model on the test set. Hand labeling is also forbidden.

- Your model should only use information which was available prior to the time for which it is forecasting.

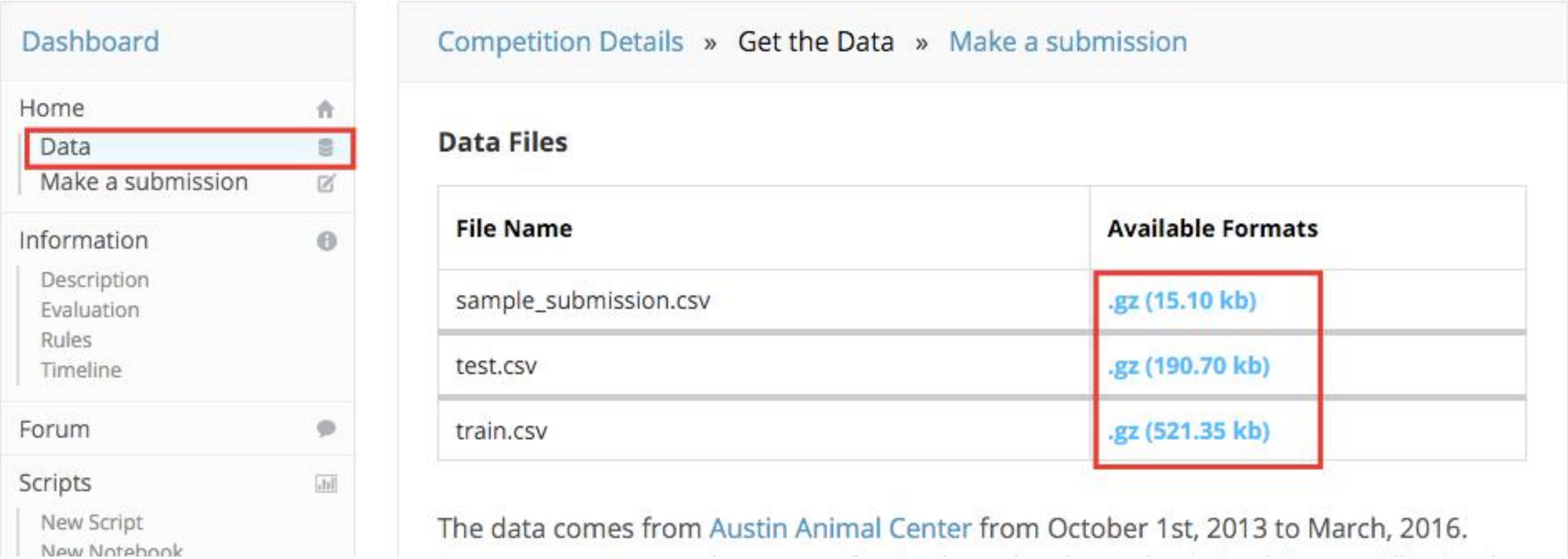
Shelter Animal Outcomes

do not cheat. You should not submit entries based on test-set answers or train your model on the test set. Hand labeling is also forbidden.

- Your model should only use information which was available prior to the time for which it is forecasting.

Shelter Animal Outcomes

1. Learn the Data



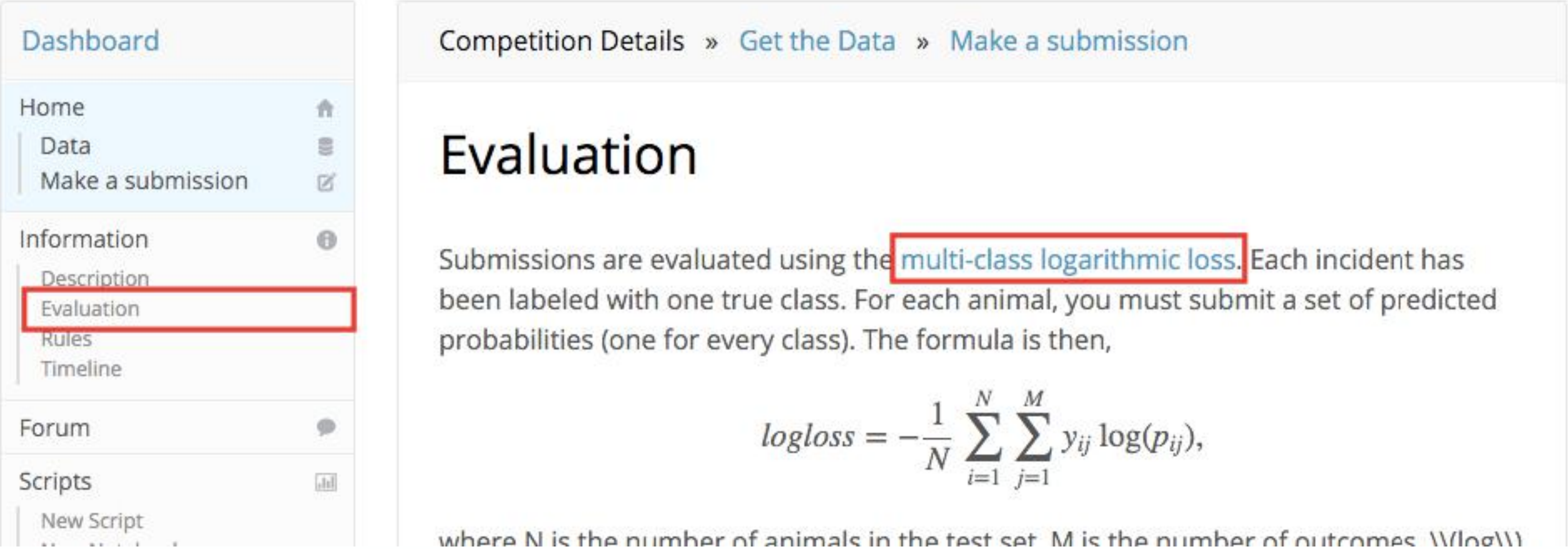
Data Files

File Name	Available Formats
sample_submission.csv	.gz (15.10 kb)
test.csv	.gz (190.70 kb)
train.csv	.gz (521.35 kb)

The data comes from Austin Animal Center from October 1st, 2013 to March, 2016.

In this competition, you are going to predict the outcome of the animal as they leave the Animal Center. These outcomes include: Adoption, Died, Euthanasia, Return to owner, and Transfer.

2. Learn evaluation



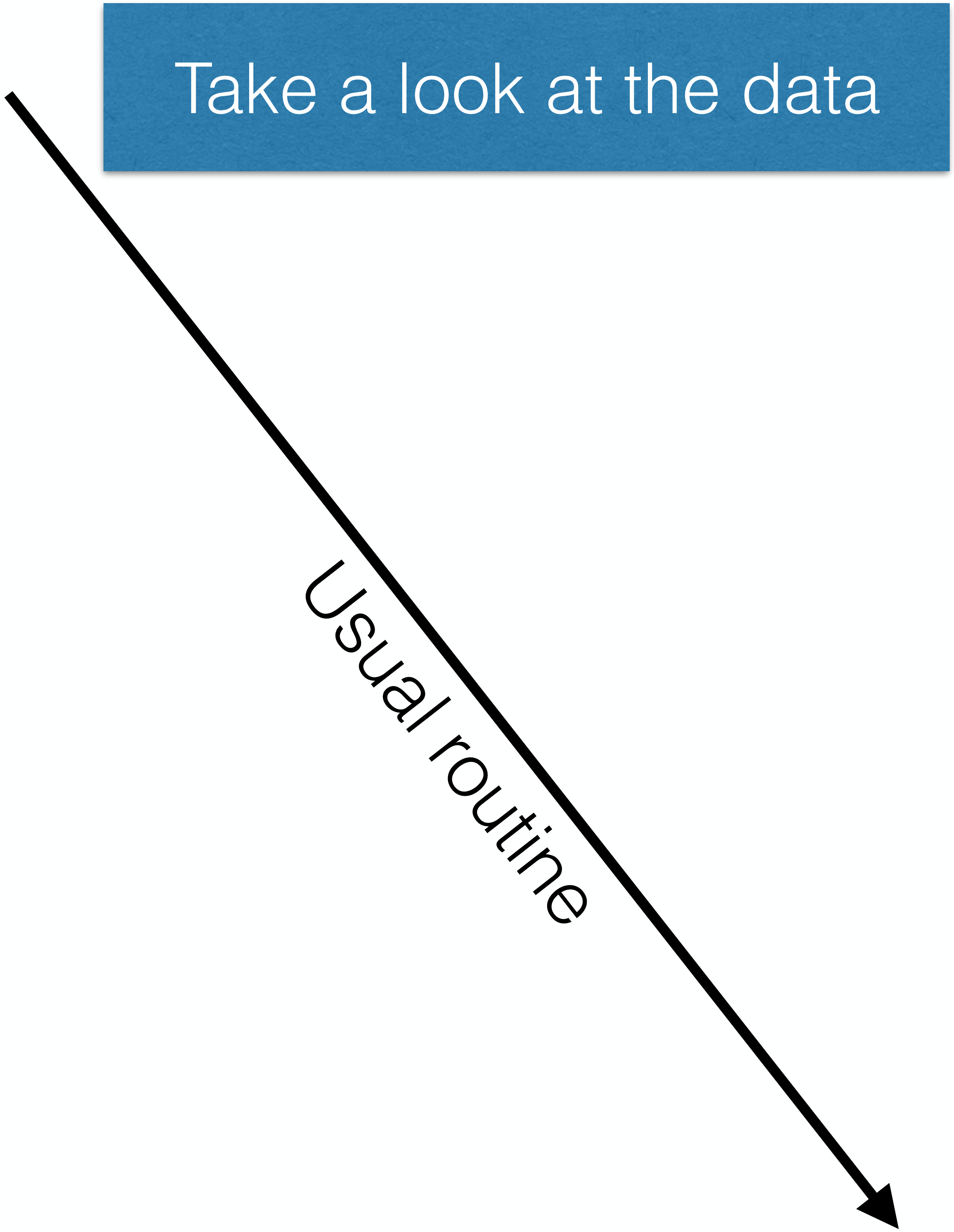
Evaluation

Submissions are evaluated using the multi-class logarithmic loss. Each incident has been labeled with one true class. For each animal, you must submit a set of predicted probabilities (one for every class). The formula is then,

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij}),$$

where N is the number of animals in the test set, M is the number of outcomes, \\(log\\)

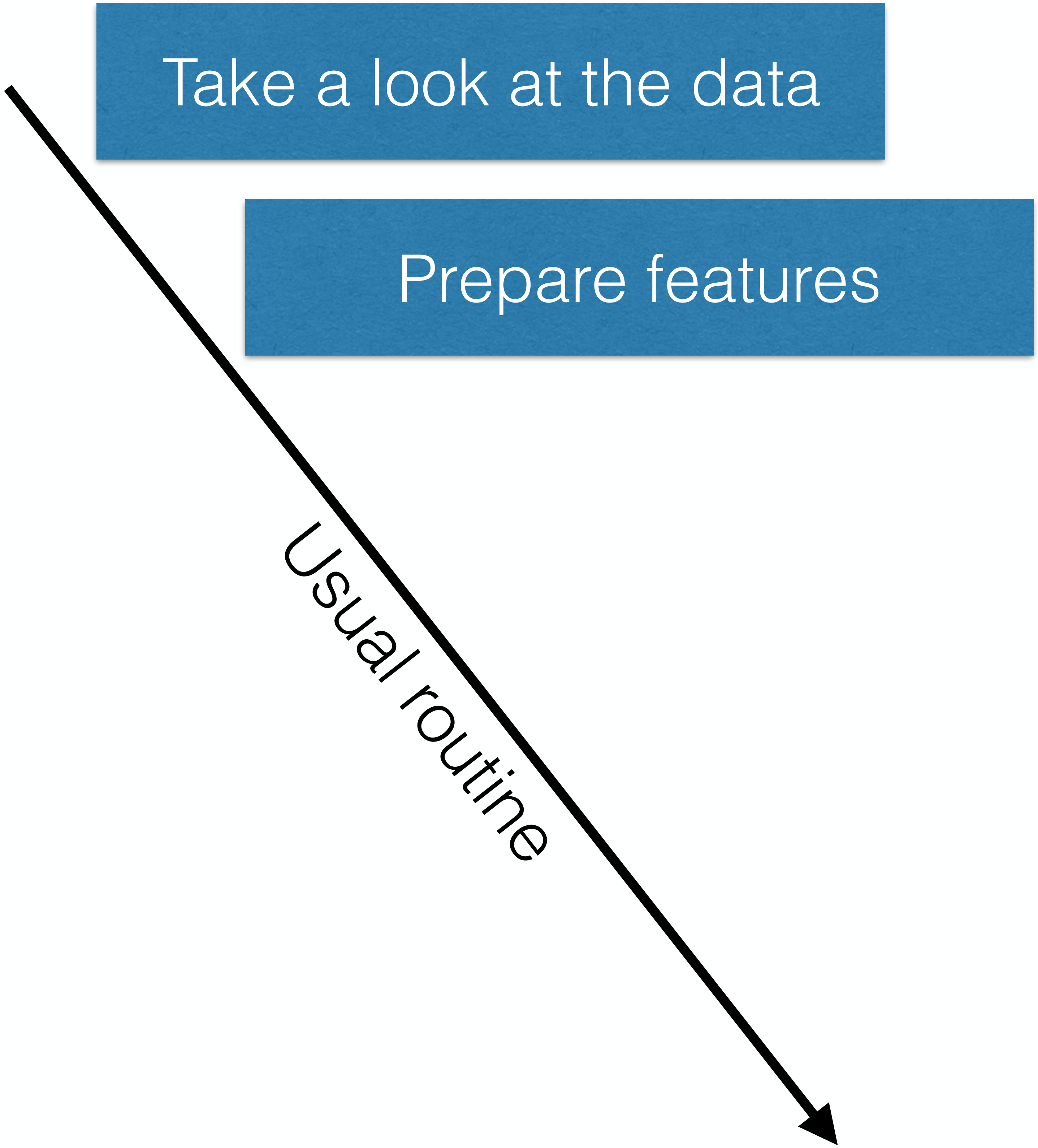
3. Do some Data Science



Take a look at the data

Usual routine

3. Do some Data Science

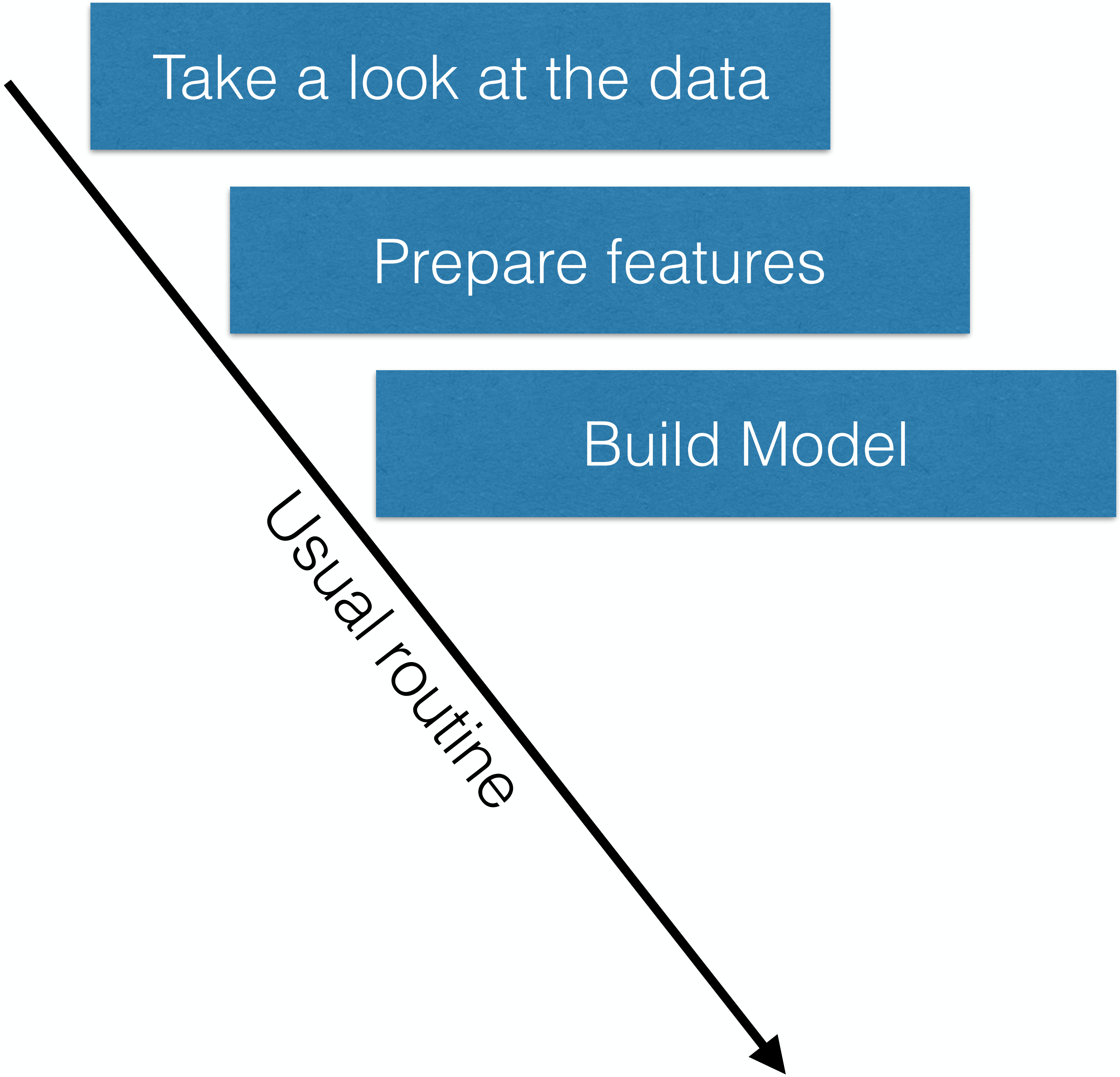


Take a look at the data

Prepare features

Usual routine

3. Do some Data Science



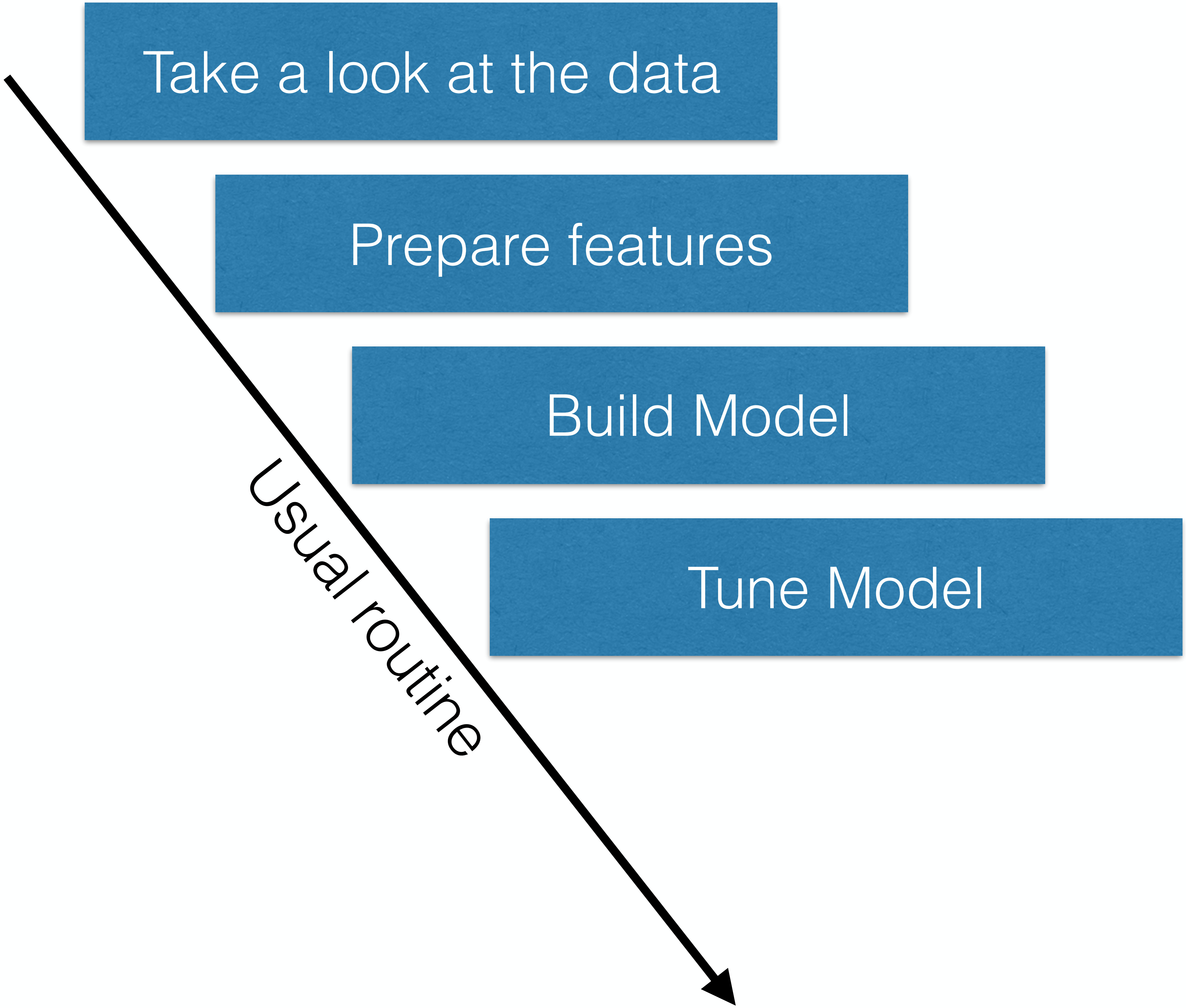
Take a look at the data

Prepare features

Build Model

Usual routine

3. Do some Data Science



Take a look at the data

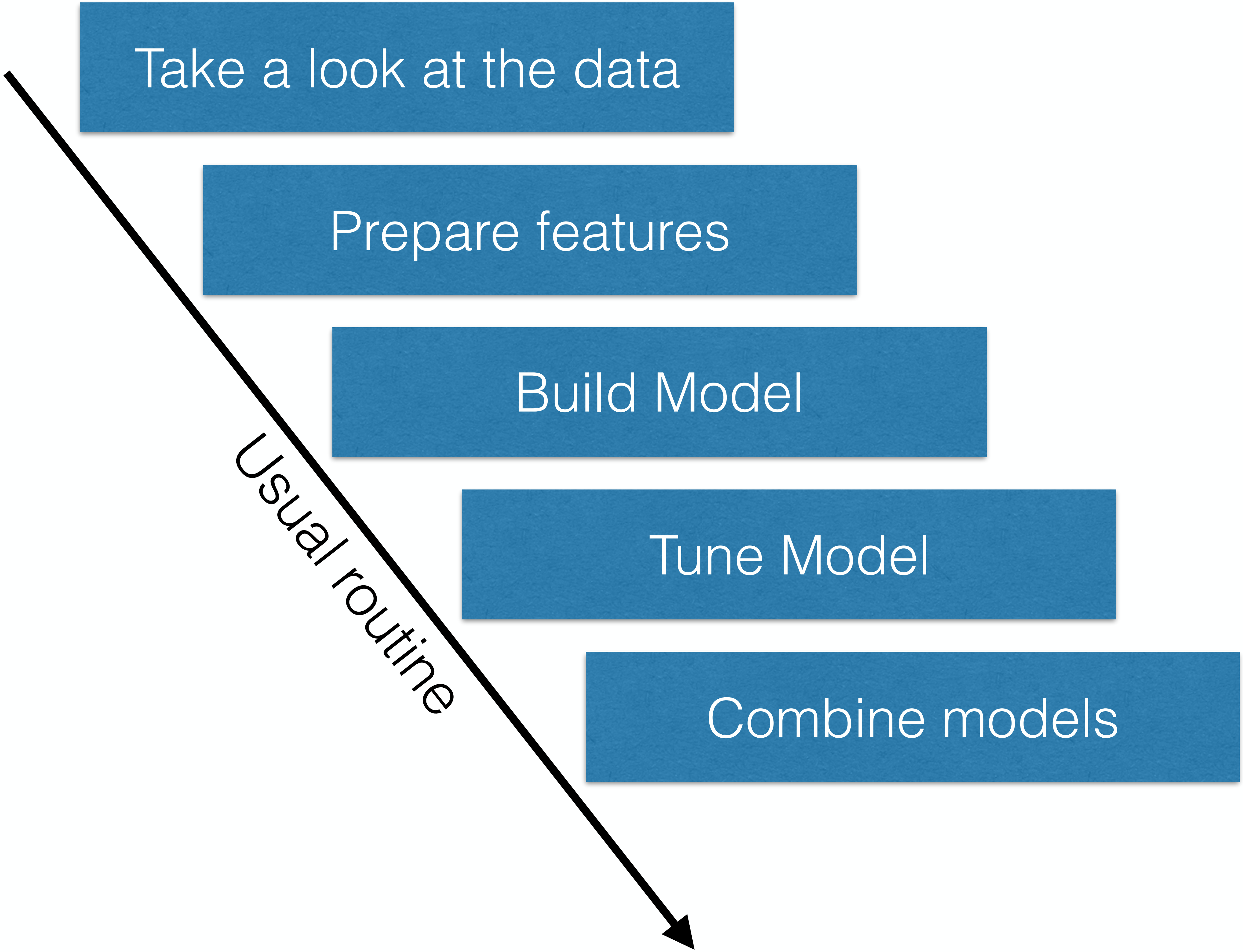
Prepare features

Build Model

Tune Model

Usual routine

3. Do some Data Science



Take a look at the data

Prepare features

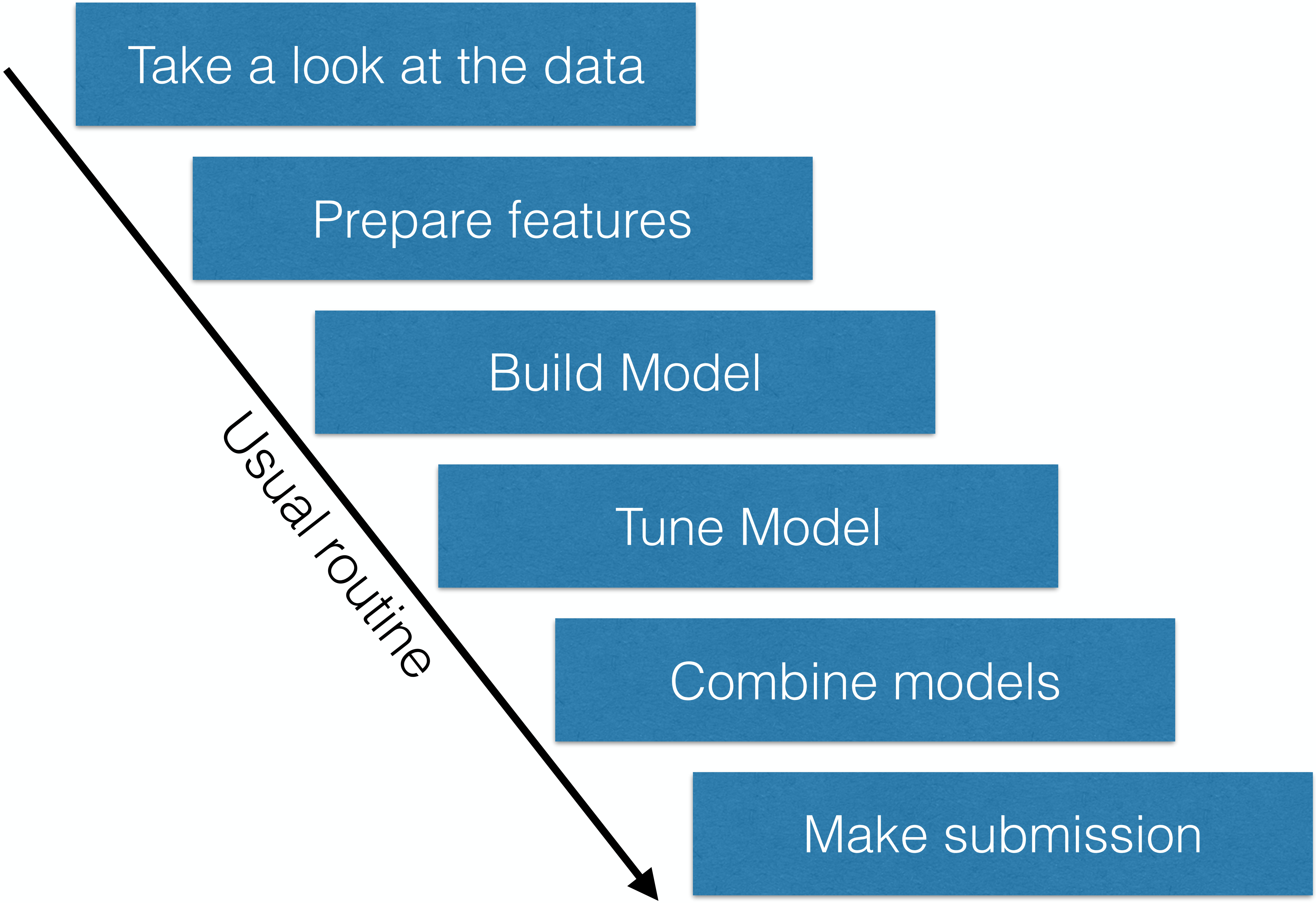
Build Model

Tune Model

Combine models

Usual routine

3. Do some Data Science



Take a look at the data

Prepare features

Build Model

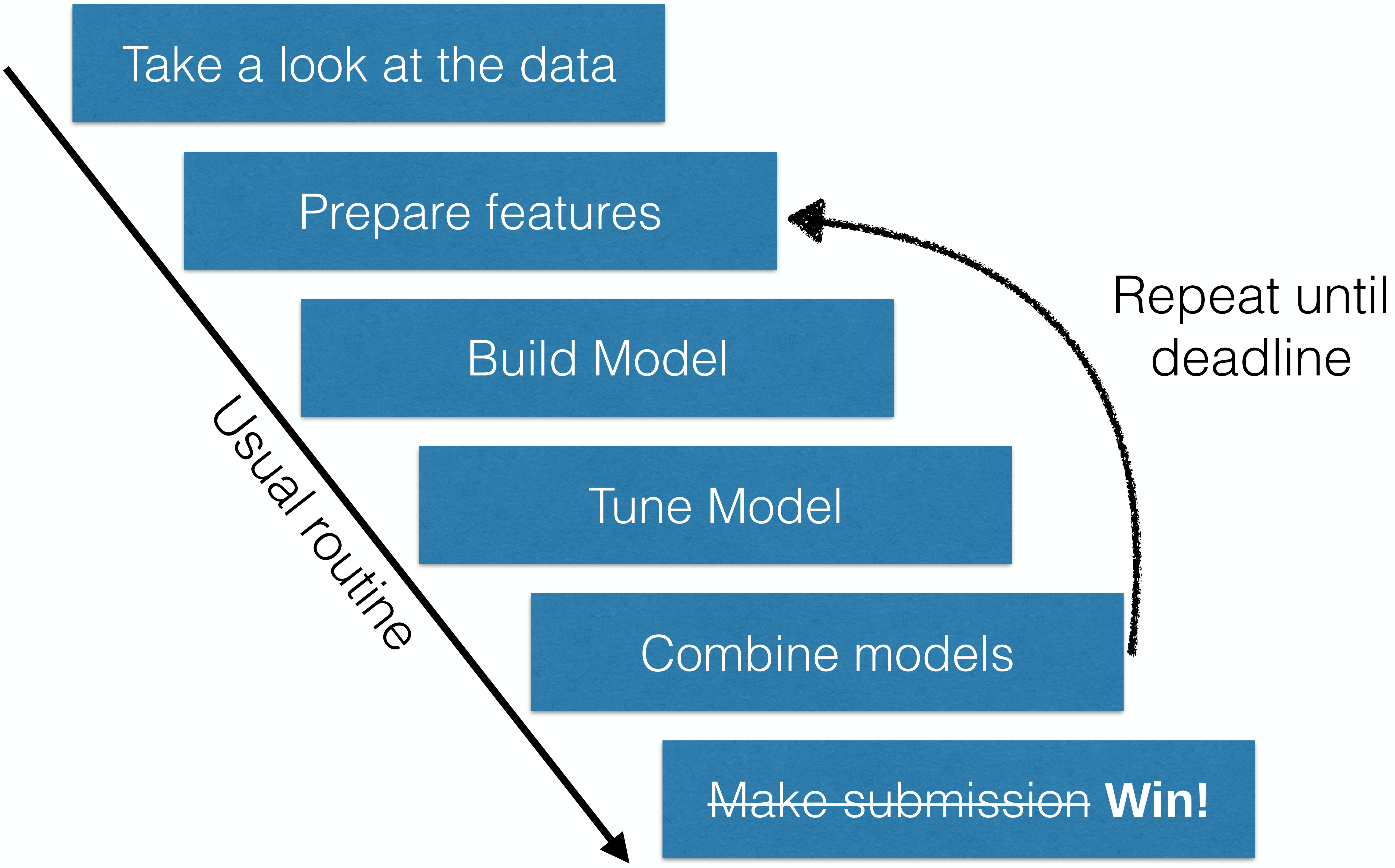
Tune Model

Combine models

Make submission

Usual routine

3. Do some Data Science



Take a look at the data

Prepare features

Build Model

Tune Model

Combine models

~~Make submission~~ **Win!**

Usual routine

Repeat until deadline

You will know after workshop

- Data wrangling with pandas
- Basics of matplotlib
- NA data imputation
- Importance of cross-validation
- Basics of ML libraries: sciki-learn, keras, xgboost
- Hyperparameter tuning
- Making ensembles
- Going to Kaggle top-10