

<https://bit.ly/2KBGl56>

KYOTO UNIVERSITY

Statistical Machine Learning Theory

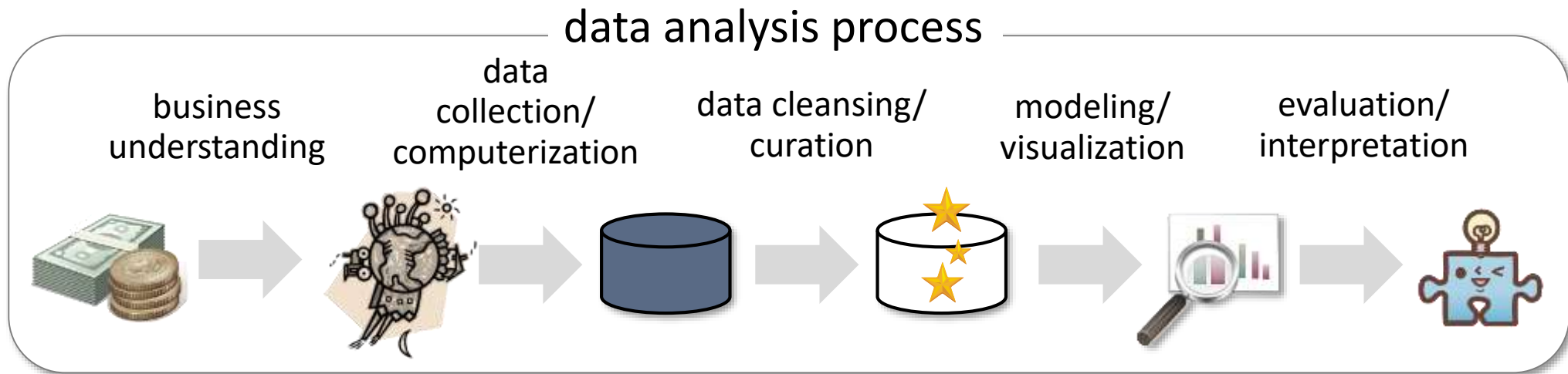
Predictive Modeling Challenge

Hisashi Kashima / Makoto Yamada

DEPARTMENT OF INTELLIGENCE SCIENCE
AND TECHNOLOGY

A serious issue in data analytics: Manpower bottleneck

- Automatic data analysis techniques (e.g. machine learning) are often considered as main components of data analytics
- Data analysis is heavily labor intensive
 - Manual processing dominates a large part of data analysis process
 - Data analysis process standards (e.g., CRISP-DM)



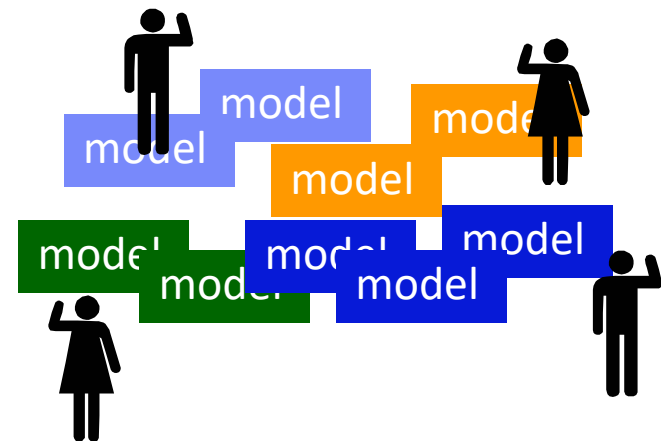
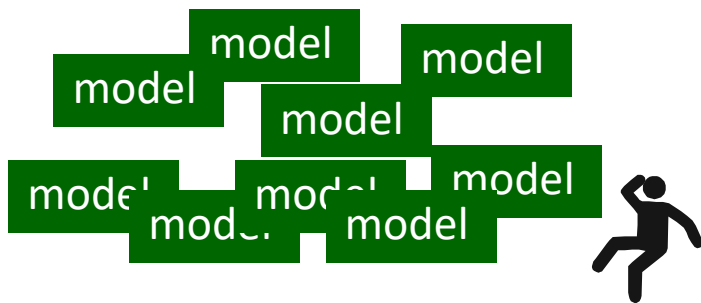
Big shortage of data scientists: Implies labor intensity in data analysis

- *“By 2015, 4.4 million IT jobs globally will be created to support big data”, but “only one-third of the IT jobs will be filled”*
- Peter Sondergaard (Senior VP at Gartner)
- *“Data Scientist: The Sexiest Job of the 21st Century”*
- Thomas H. Davenport and D.J. Patil, Harvard Business Review
- These statements imply the labor intensity of data analysis



Labor intensity of data modeling: Exploring huge model space is labor-intensive

- Predictive modeling is labor-intensive
 - Requires extensive model selection + feature engineering
 - “No free lunch”: there is no universally good model
- *Crowds of data scientists* can explore the huge model space
 - Hard for a *single data scientist*



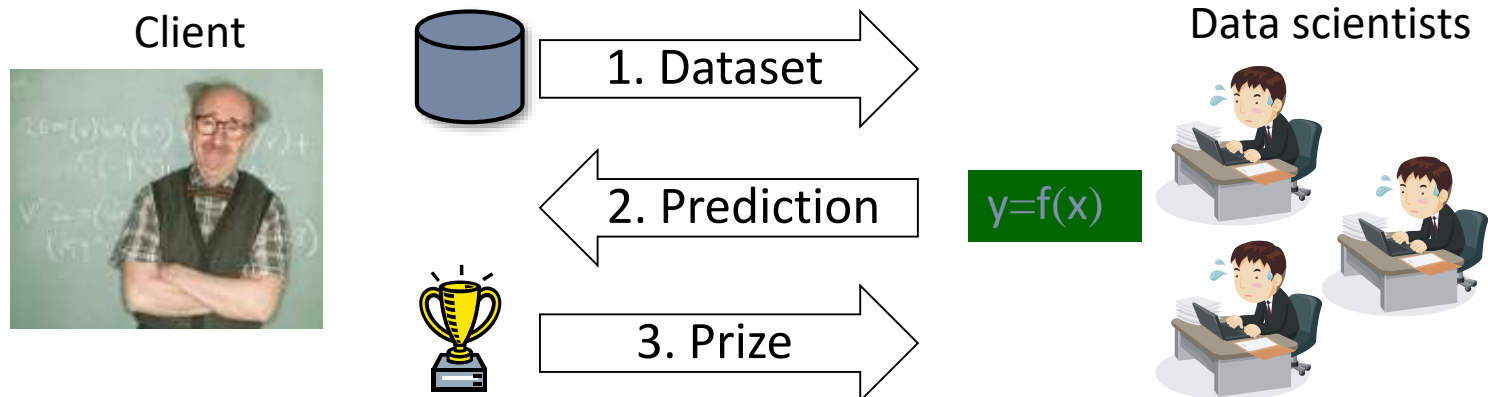
Predictive modeling competition: Crowdsourcing of data scientists

■ Predictive modeling competition:

1. Training dataset is published
2. Participants submit predictions for test dataset
3. Winner is determined by results on test data (and gets awarded)


} Several weeks
to months

■ Supporting platforms (e.g. Kaggle)



Predictive modeling challenge:

Supervised classification competition

- A supervised classification problem:
 - Implementing some algorithms by yourself is recommended, but you can use publicly available implementations (e.g. scikit.learn)
- Participate into a competition at <http://universityofbigdata.net>
 - Online ad click prediction (by courtesy of  CyberAgent®)
 - Will start at **May 20th** and ends at **June 30th**
- Submit a report summarizing your work
 - Due: **July 9th noon**

How to participate:

Register to University Of Big Data

- The competition is held at the educational competition platform *University of Big Data*:
<http://universityofbigdata.net/?lang=en>
- Register with your Google account (if you have not)
 - With registration code ‘SML2018challenge’



- Challenge to the competition requires a permission (which may take a few hours to days)
 - If you still cannot access to the competition page, contact the instructor

Submit your prediction: <https://bit.ly/2L9Z74C>

- See the instructions at <http://universityofbigdata.net/competition/5723788444434432?lang=en>

The screenshot shows the competition page for 'Ad data Click Prediction Challenge' on the University of Big Data website. The page includes a header with the university logo and navigation links. The main content area provides details about the competition, including the problem type (Classification), evaluation metric (Area under the ROC curve (AUC)), and competition status (Coming). It also lists the start and end dates, public/private status, and invitation setting. A 'Submission' modal is open, showing a file upload field with 'Select file' and 'Submit' buttons, a note field, and an intermediate ranking table.

Ad data Click Prediction Challenge

In this competition, you are asked to predict whether an ad is clicked. More specifically, using the probability of click of an ad.

The competition dataset has been provided by **Cyber Agent AI Lab**.

Tutorial for this competition is available.

Problem type	Classification
Evaluation metric	Area under the ROC curve (AUC)
Competition status	Coming
Started	2018/05/20 00:00 (Japan Standard Time)
Ends	2018/06/30 23:59 (Japan Standard Time)
Public/Private	Private
Invitation setting	Invitation only

[Show console](#) [Edit](#)

Dataset

[Download](#)
un2018-ad-click.zip

Submission

[Select file](#) [Submit](#)

管理権アカウントには追加投稿権限はありません

You can upload a file of up to 20MB. You can compress your submission using the .zip compression format.

Note (optional)

You can add a note to your submission. Notes are shown in the bottom of this page and only you can see your note.

Intermediate ranking

Intermediate rank	Nickname	Intermediate score
1	University of Big Data	0.0240

This leaderboard is calculated on the latest submissions. The intermediate scores are calculated using 50% of the test dataset, and the final scores are calculated using the other 50%. Final ranks are determined according to the final scores.

Report submission:

Submit a report summarizing your work (in English)

- Submission:
 - Due: July 9th noon
 - Send your report to `statisticallearningtheory2018@gmail.com` and confirm you receive an ack on 9th
- The report must include:
 - Idea behind your approach, analysis pipeline, results, and discussions
(Do not include your source codes)
 - At least 3 pages, but do not exceed 6 pages in LNCS format

The competition task: Advertisement click prediction

- Predict whether the advertisement will be clicked

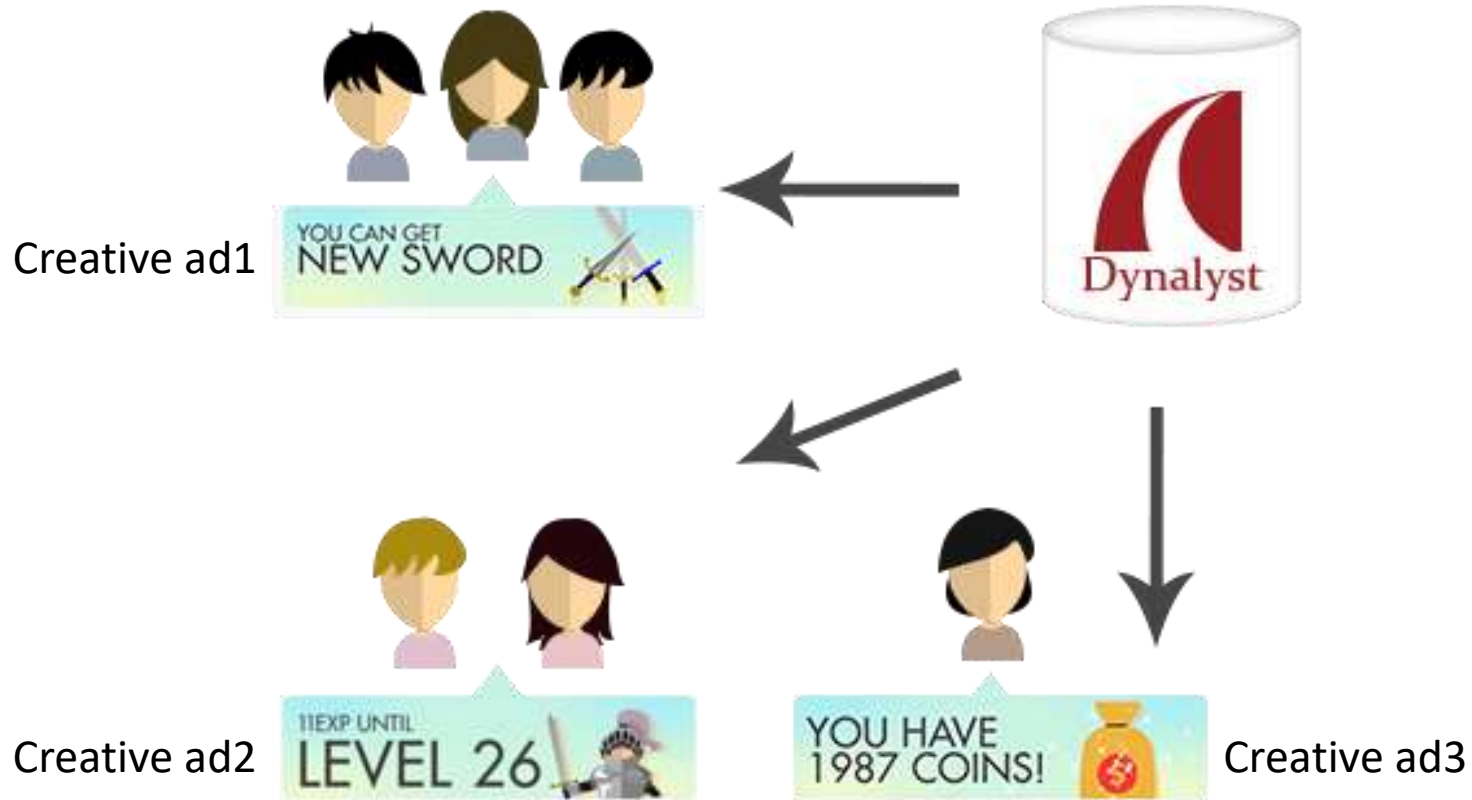


Image at <https://www.dynalyst.jp/>

Dataset:

Training data, test data

- Training data: data_train.csv

Advertisement feature input $\mathbf{x}^{(i)}$

Correct labels $\mathbf{y}^{(i)}$

	Logged_at	Advertiser_id	Campaign_id		click
1	2018-03-15 00:00:00.125	1909	7942	...	0
2	2018-03-15 00:00:29.917	2088	10668	...	1

- Test data: data_test.csv

	Logged_at	Advertiser_id	Campaign_id		click
1	2018-03-15 15:28:09.221	1953	8687	...	Predict this
2	2018-03-15 15:28:13.202	1909	7948	...	

Submission:

Submit your predictions for the test data

- Predict the probability of each advertisement information.

	Logged_at	Advertiser_id	Campaign_id		click
1	2018-03-15 15:28:09.221	1953	8687	...	0.3384
2	2018-03-15 15:28:13.202	1909	7948	...	0.4951

Submission



```
0.3384
0.4951
...
```

Example submission file:
sample-submission.dat

- You can make submissions at most three times a day

Evaluation measure:

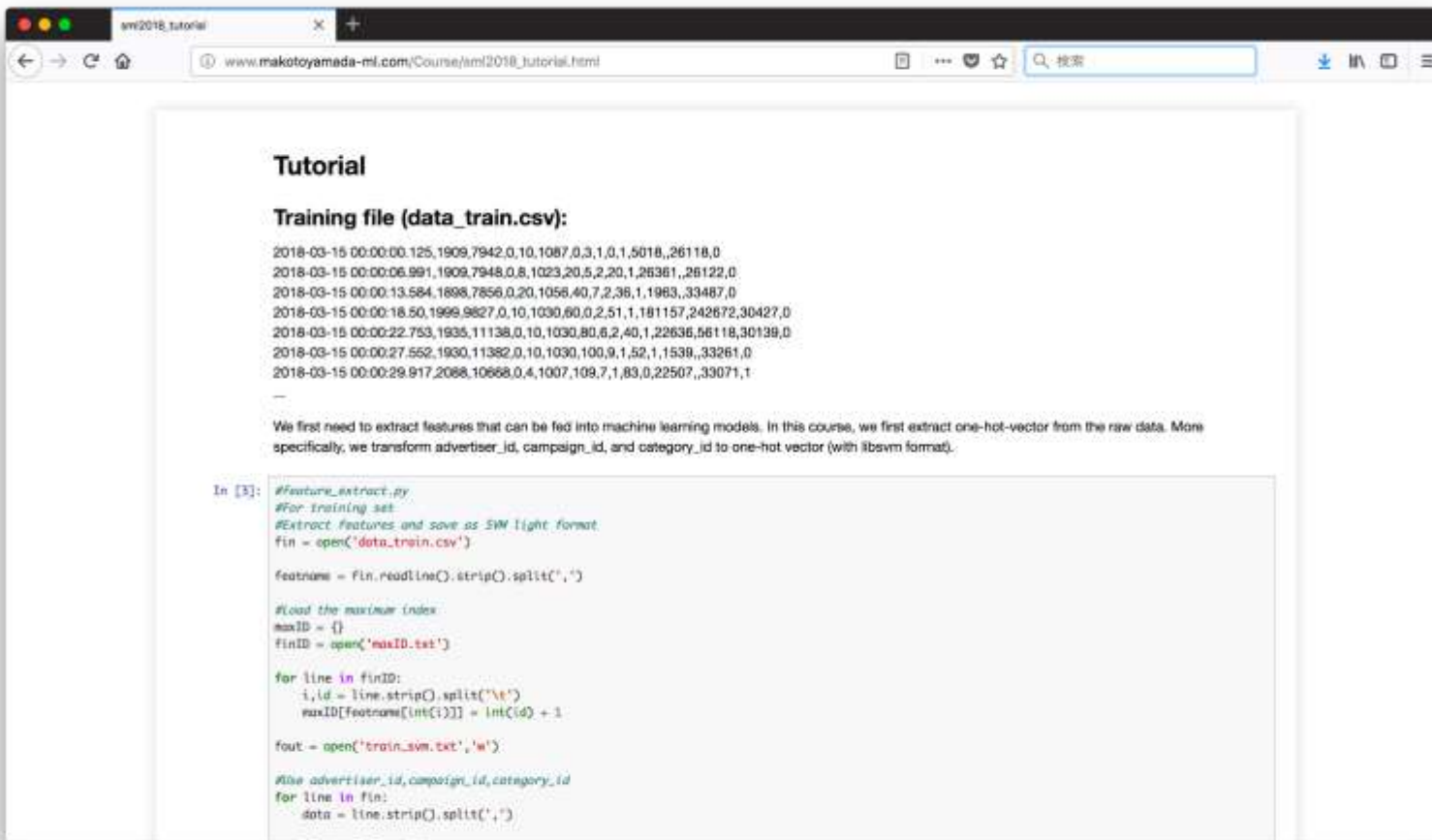
ROC-AUC

- ROC-AUC is a evaluation measure of two-class classification
- See http://scikit-learn.org/stable/modules/model_evaluation.html#roc-metrics

Tutorial:

Quick start guide for making the first predictions

- Find the tutorial at: http://www.makotoyamada-ml.com/Course/sml2018_tutorial.html



Tutorial

Training file (data_train.csv):

```
2018-03-15 00:00:00.125,1909,7942,0,10,1087,0,3,1,0,1,5018,,26118,0
2018-03-15 00:00:06.991,1909,7948,0,8,1023,20,5,2,20,1,26361,,26122,0
2018-03-15 00:00:13.584,1898,7856,0,20,1056,40,7,2,36,1,1963,,33487,0
2018-03-15 00:00:18.50,1999,9827,0,10,1030,60,0,2,51,1,181157,242672,30427,0
2018-03-15 00:00:22.753,1935,11138,0,10,1030,80,6,2,40,1,22636,56118,30139,0
2018-03-15 00:00:27.552,1930,11382,0,10,1030,100,9,1,52,1,1538,,33261,0
2018-03-15 00:00:29.917,2068,10668,0,4,1007,109,7,1,83,0,22507,,33071,1
```

—

We first need to extract features that can be fed into machine learning models. In this course, we first extract one-hot-vector from the raw data. More specifically, we transform advertiser_id, campaign_id, and category_id to one-hot vector (with libsvm format).

```
In [3]: #feature_extract.py
#For training set
#Extract features and save as SVM light format
fin = open('data_train.csv')

featname = Fin.readline().strip().split(',')

#Load the maximum index
maxID = {}
finID = open('maxID.txt')

for line in finID:
    i, id = line.strip().split('\t')
    maxID[featname[int(i)]] = int(id) + 1

fout = open('train_svm.txt', 'w')

#Use advertiser_id, campaign_id, category_id
for line in fin:
    data = line.strip().split(',')
```