

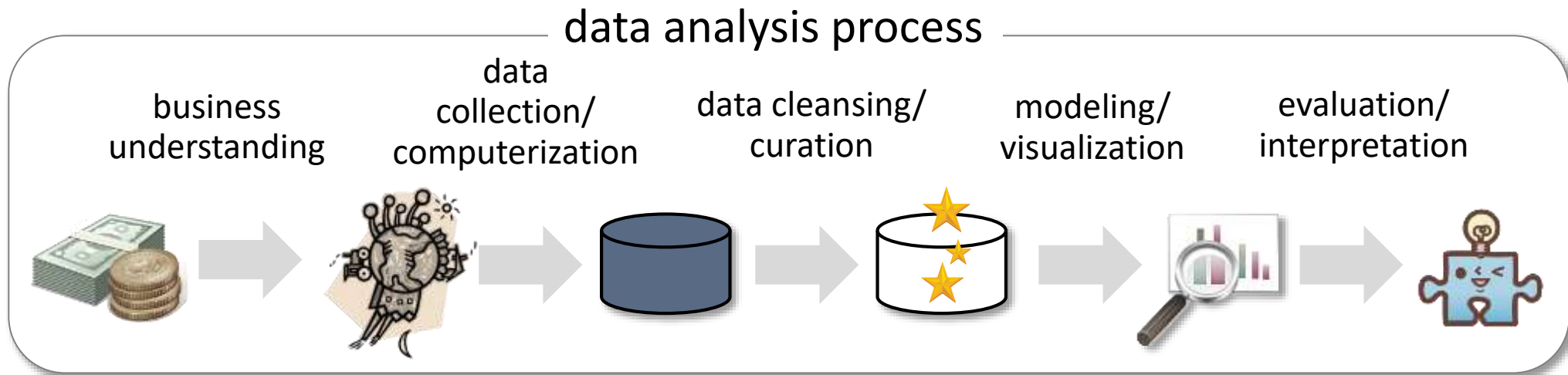
*Statistical Machine Learning Theory*

**Predictive Modeling Challenge**

Hisashi Kashima  
kashima@i.Kyoto-u.ac.jp

# 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

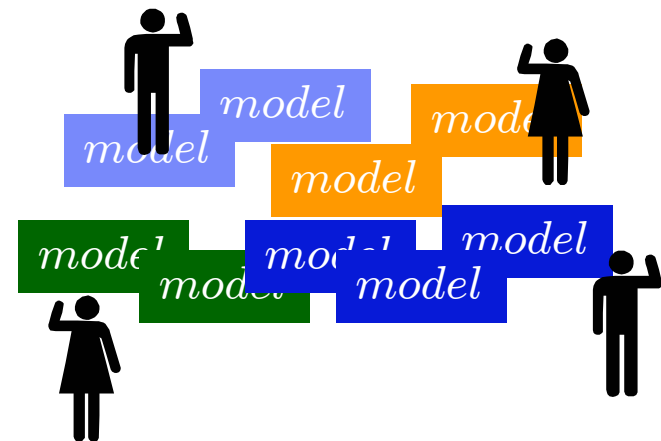
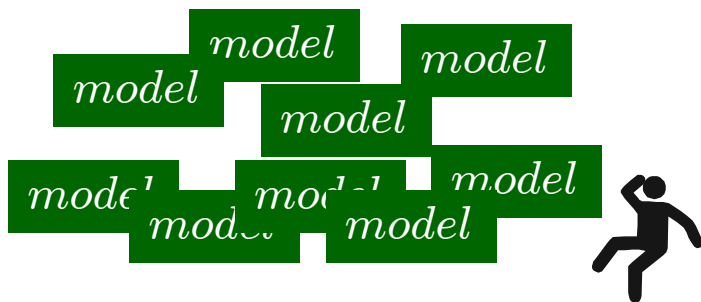
---

- *“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*



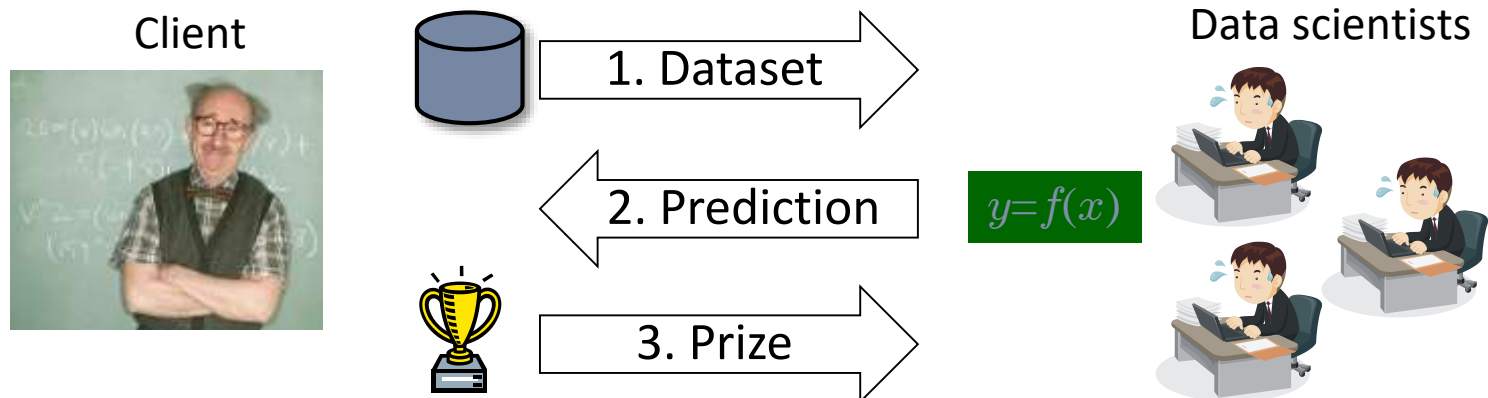
# Predictive modeling competition: Crowdsourcing of data scientists

## ■ 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 (multi-label) 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>
  - Name card recognition task
  - Will start at **May 12th** 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 ‘SML2017challenge’



# Submit your prediction: <https://goo.gl/iAGRnA>

- See the instructions at <http://universityofbigdata.net/competition/572378844443>

The screenshot shows the competition page with a submission form and an intermediate ranking table.

**Submission**

管理者アカウントには提出回数制限はありません。  
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

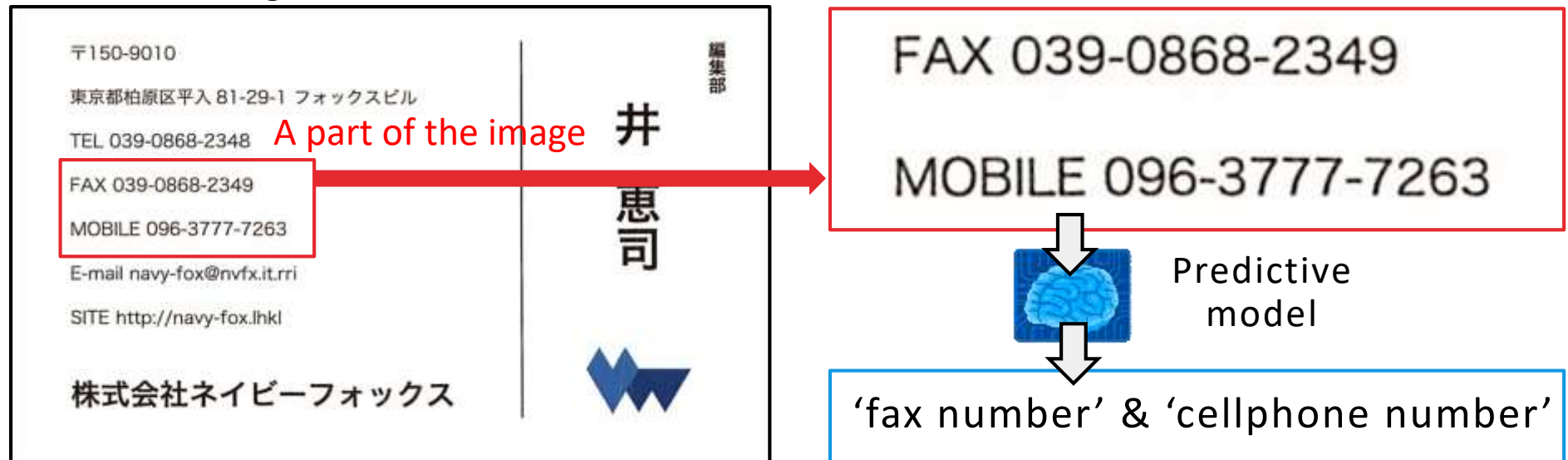
---

- Submission:
  - Due: July 9th noon
  - Send your report to `statisticallearningtheory2017@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 source codes)
  - At least 3 pages, but do not exceed 6 pages in LNCS format

# The competition task: Name card recognition

- A part of an image of a name card
- Tell whether or not the part includes some of the following items:  
company name, name, position, address, phone number, fax number, cellphone number, e-mail address, and URL

Name card image



# Dataset:

## Training data, test data, and image data

- Training data: train.csv

A part of a name card image  $\mathbf{x}^{(i)}$

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

	filename	left	top	right	bottom	Company_name		url
1	2842.png	491	455	796	485	1	...	0
2	182.png	24	858	311	886	0	...	0

- Test data: test.csv

	filename	left	top	right	bottom
1	1942.png	66	359	361	386
2	101.png	58	373	519	422

Predict this part

- Image data: images.zip

# Submission:

## Submit your predictions for the test data

- Predict the probability of each image part including each item (e.g. Company\_name)

	filename	left	top	right	bottom	Company_name		url
1	1942.png	491	455	796	485	0.068	...	0.102
2	101.png	24	858	311	886	0.555	...	0.003



Submission

```
0.068 0.208 ... 0.102
0.555 0.109 ... 0.003
...
```

Example submission file:  
sample-submission.dat

- You can make submissions at most three times a day

## Evaluation measure:

### Macro averaged ROC-AUC

---

- ROC-AUC is a evaluation measure of two-class classification
- The task is a multi-label classification problem, a collection of two-class classification problems
- Macro-averaged AUC is the average of the AUC scores for the two-class classification problems
- See [http://scikit-learn.org/stable/modules/model\\_evaluation.html#roc-metrics](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://universityofbigdata.net/competition/tutorial/5723788444434432?lang=en>



The screenshot shows the top navigation bar of the University of Big Data website. It includes the university logo, the text "UNIVERSITY OF BIG DATA", and navigation links for "Competitions" and "Enroll". There are also buttons for "Sign up with Google Account" and "Sign in", along with a settings icon. Below the navigation bar, a dark blue header contains the title "[Sansan Data Analysis Challenge] Business card field labeling tutorial". The main content area starts with an "Introduction" section, followed by a paragraph explaining the tutorial's purpose. Below this, a list of required libraries is provided, and a "Reference (In Japanese):" section contains two links.

UNIVERSITY OF  
BIG DATA

Competitions Enroll

Sign up with Google Account Sign in

[Sansan Data Analysis Challenge] Business card field labeling tutorial

Introduction

This tutorial introduces an exemplar implementation for solving the competition of [Sansan Data Analysis Challenge] Business card field labeling, by using Python and the existing implementations of machine learning classifiers provided by `scikit-learn`.

The source code provided has been tested in Python 2.7.8, and requires the following libraries:

- `scikit-learn 0.18.1`
- `NumPy 1.11.3`
- `pandas 0.19.2`
- `Python Imaging Library (PIL) 1.1.7`

Reference (In Japanese):

- [https://deepanalytics.jp/contents/sansan\\_tutorial\\_1](https://deepanalytics.jp/contents/sansan_tutorial_1)
- [https://github.com/takagiwa-ss/deepanalytics\\_compe26\\_benchmark](https://github.com/takagiwa-ss/deepanalytics_compe26_benchmark)