

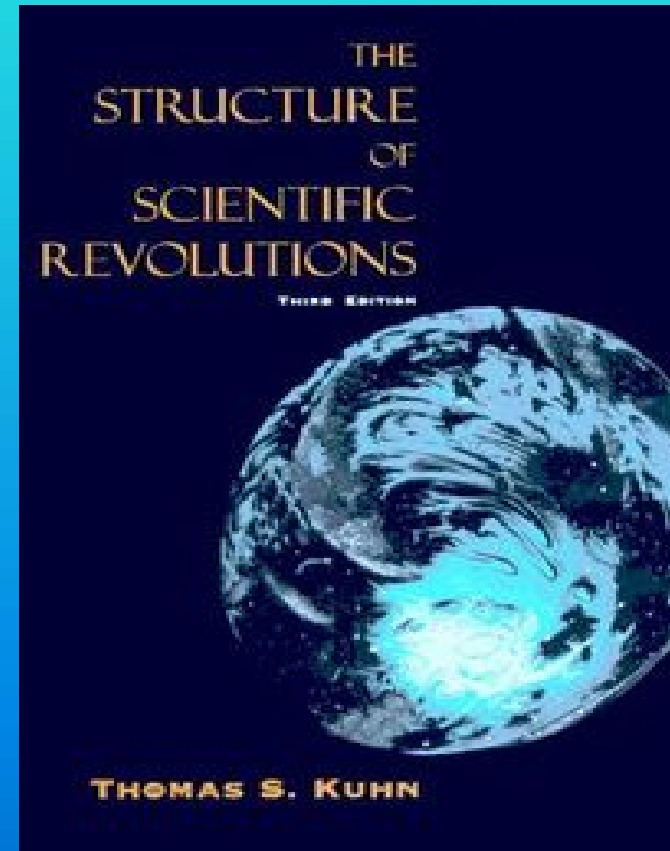
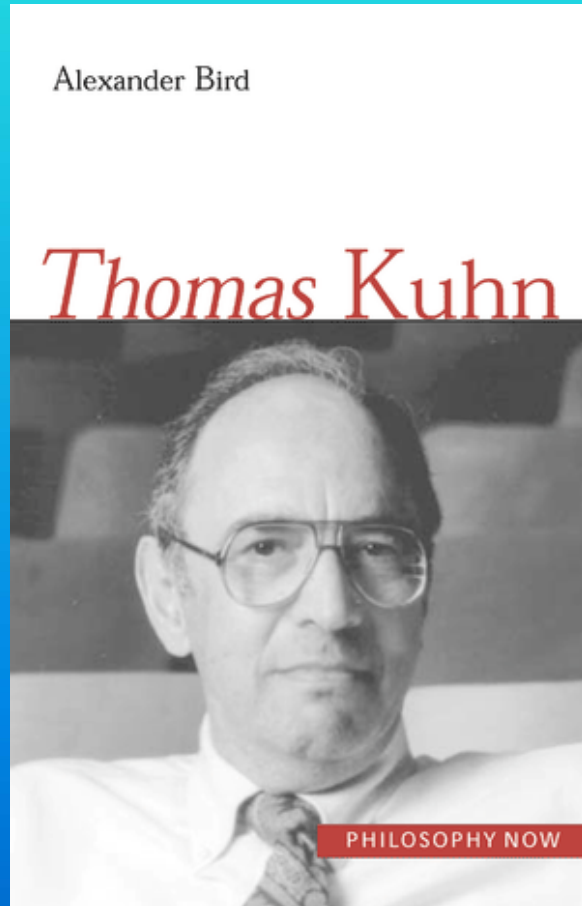
Statistics in Data Science

Ritei Shibata

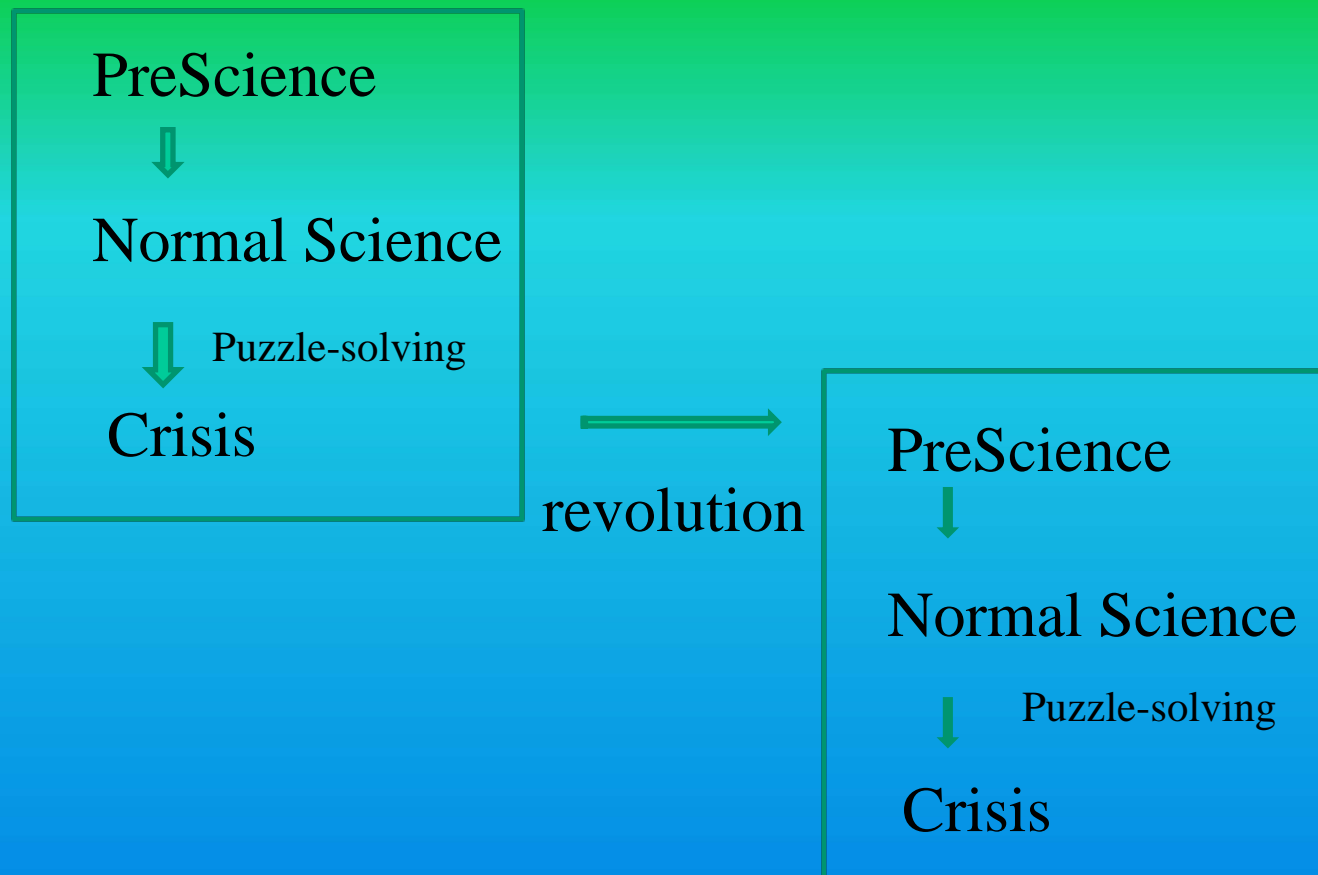
(Keio University, Yokohama, Japan)

Paradigm Shift

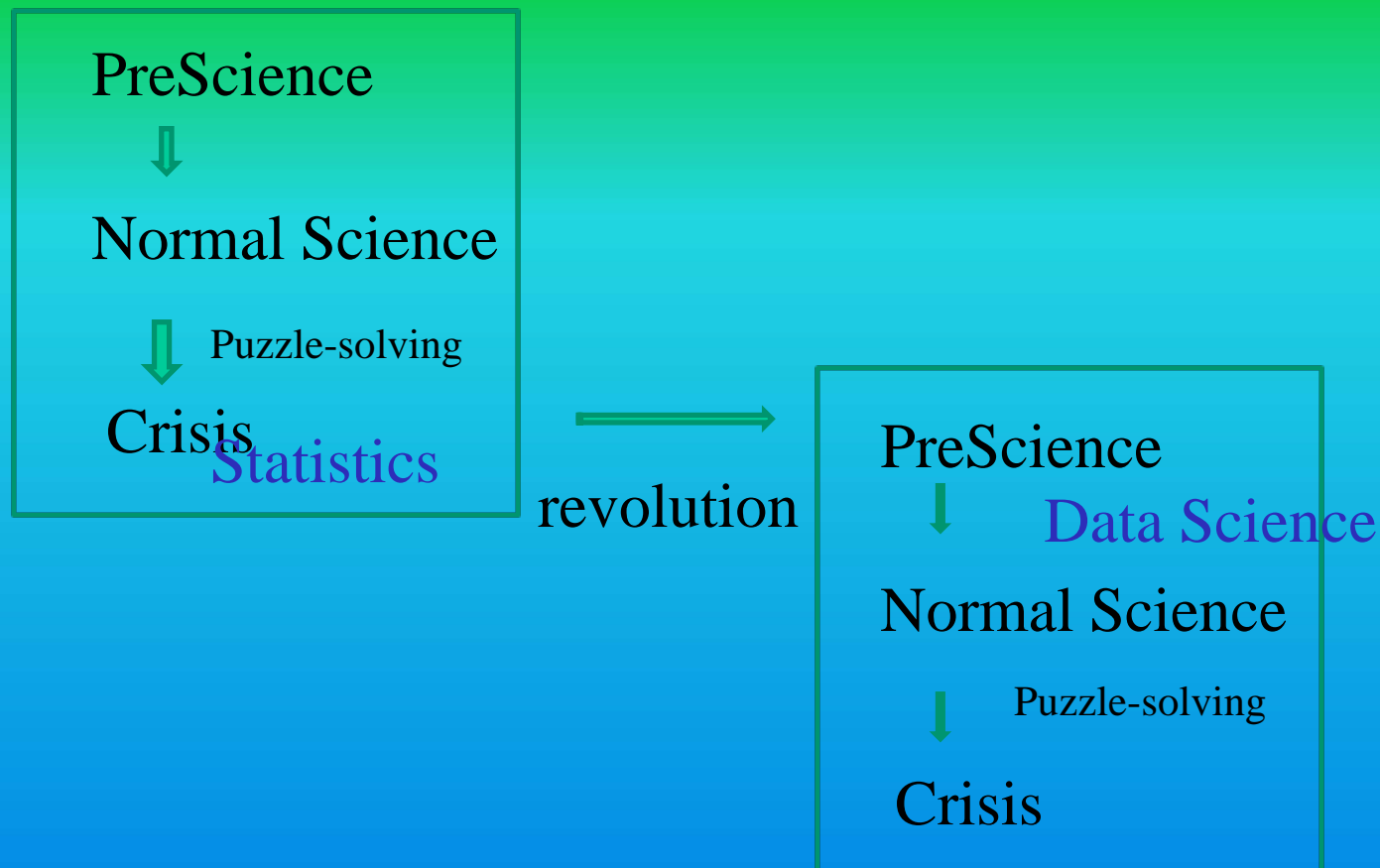
Thomas S. Kuhn (1922-1996)



University of Chicago Press, 1962



The failure of a result to conform to the paradigm
The failure of a result to conform to the paradigm
The failure of a result to conform to the paradigm
The failure of a result to conform to the paradigm
⋮



The failure of a result to conform to the paradigm of modern statistics

- Insists on Randomness

$$X_1, X_2, \dots, X_n : i.i.d.$$

- Sticks on Methodologies or Formal Procedures

Analysis of Variance, Discriminant Analysis,

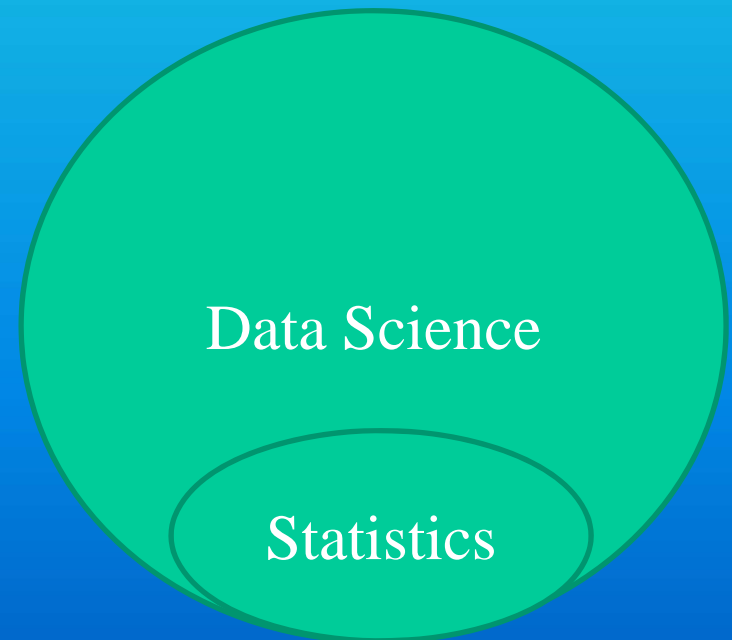


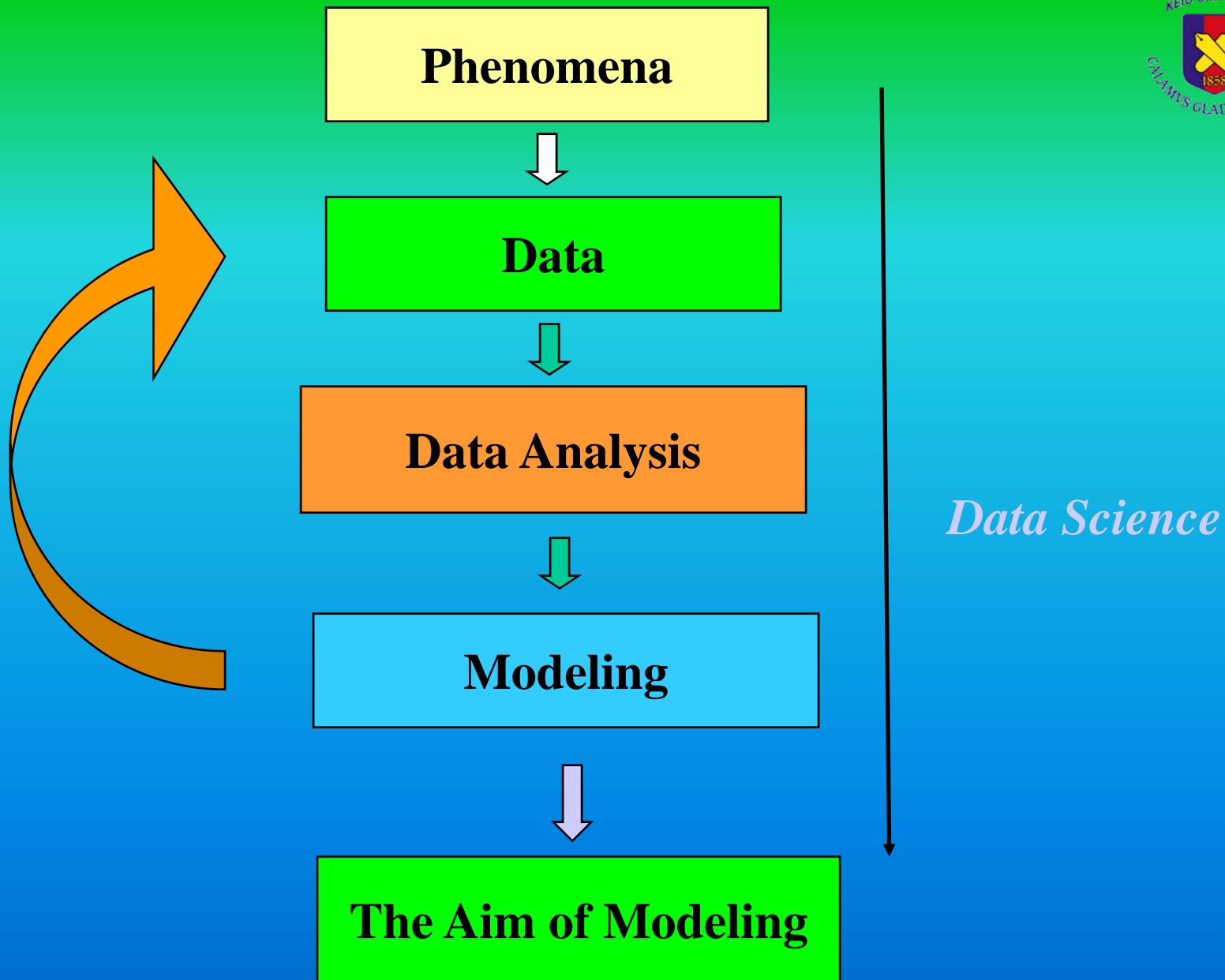
- Loosing Power of Finding Something New
- Subject to Other Sciences
- Loosing Charm and Respect
 - ✓ Decrease of Students
 - ✓ Decrease of Professional Statisticians

Paradigm of Data Science



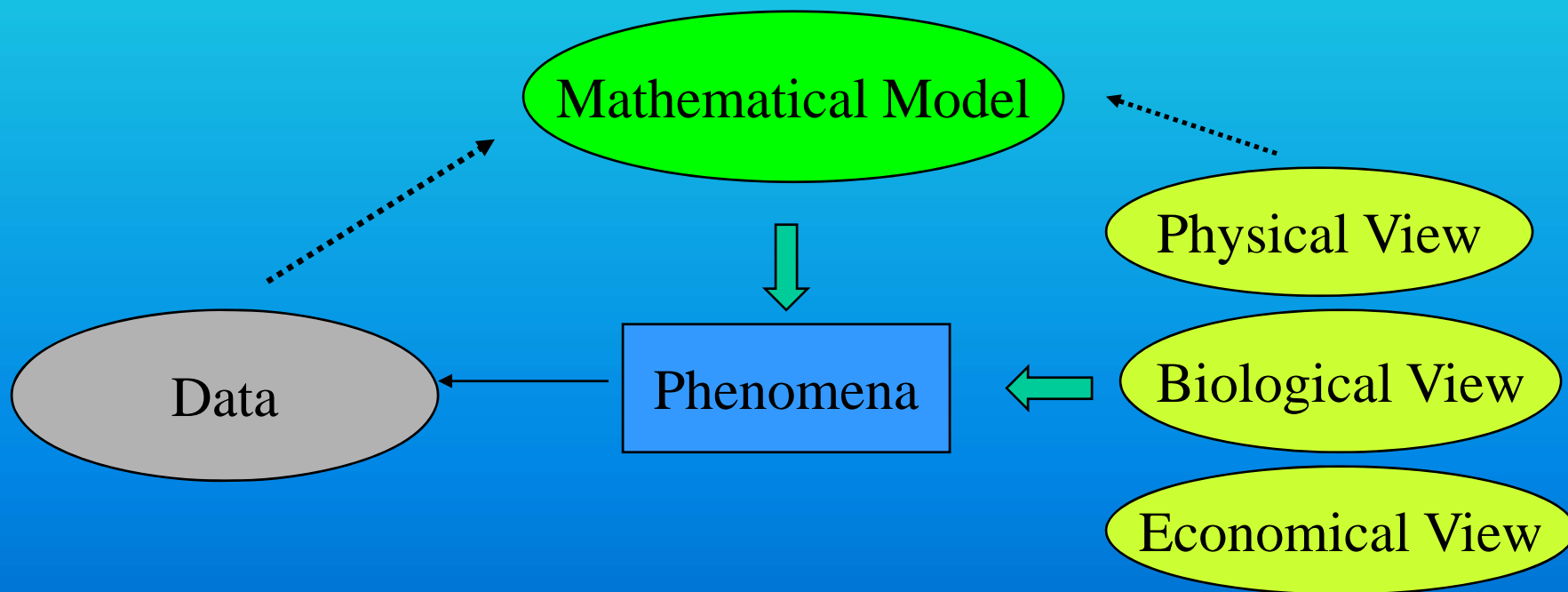
- Science of Data
 - Not a science of methodology
 - Not an application of Probability Theory
- Interest in
 - Diversity of data
 - Quality of data
 - Attributes of data
 - Flow of data
 - Metamorphosis of data
 - Structure of data
- From Data to Model
 - Stochastic and deterministic
- Human Interface to Data
 - Data Visualisation
 - Visualisation of the result





Modeling in Data Science

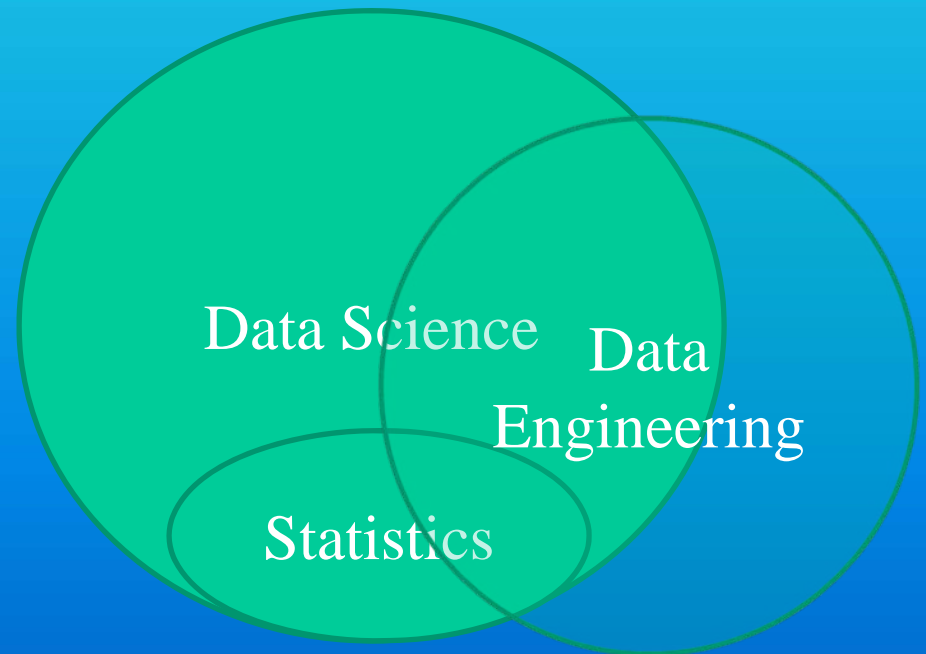
- Integration



Data Engineering



Signal Processing
Data Assimilation
Data Base Management
Data Mining
Text Mining



Data Science Series of Books, 2001~, Kyoritsu Pub.



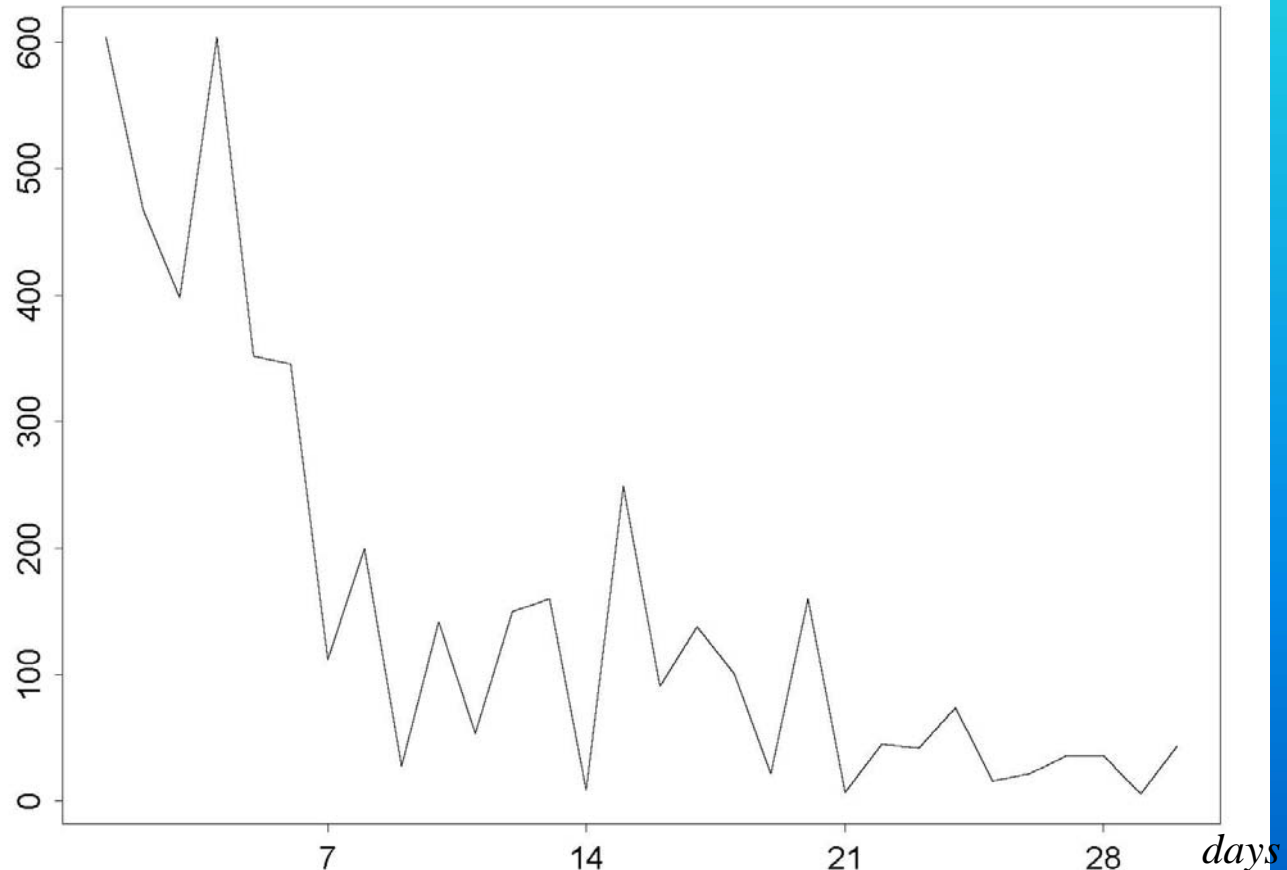
Contents of the Series

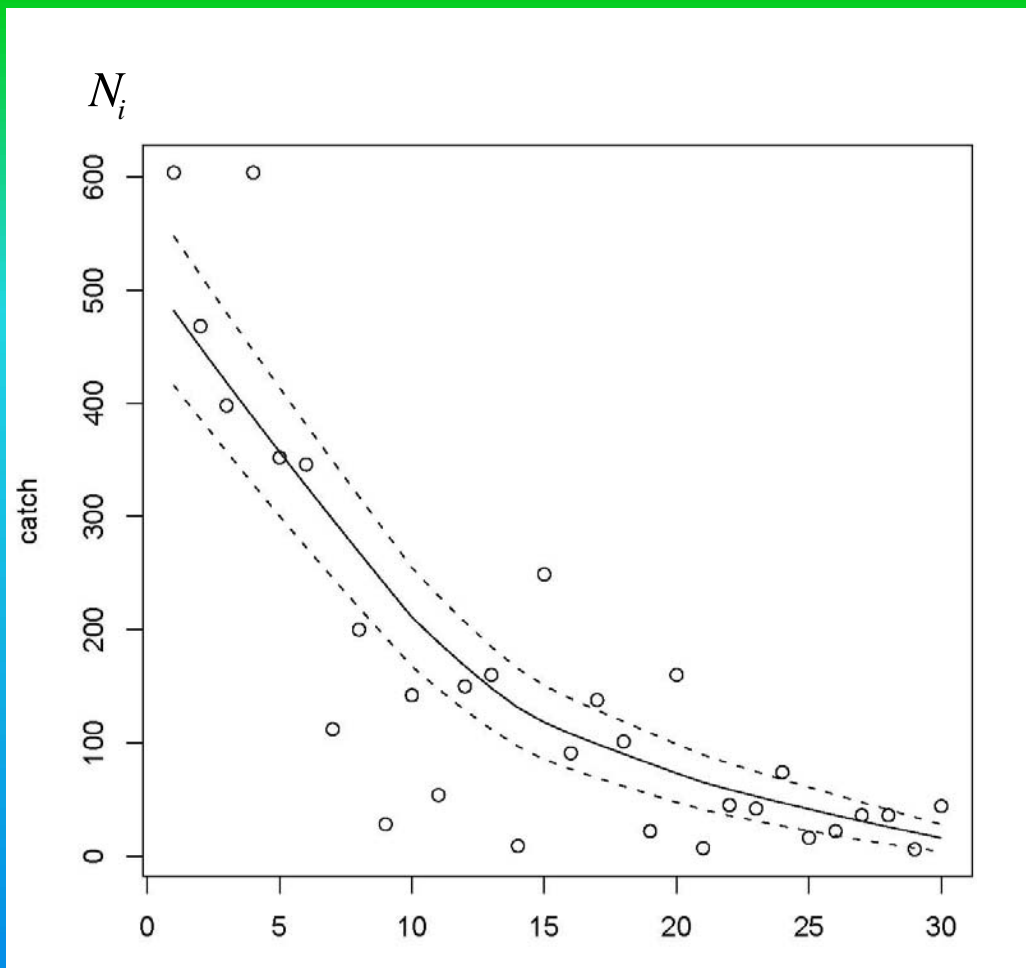
1. Data Literacy *by Ritei Shibata*
2. Data Sampling *by Masakazu Jimbo et al.*
3. Data Mining *by Takashi Fukuda et al.*
4. Data Modeling *to appear*
5. Model Validation *by G. Kitagawa et al.*
6. Data Learning Algorithm *by Sumio Watanabe*
7. Spatial Data Modeling *by Shigeru Mase et al.*
8. Earth Environmental Data *by Kunio Shimizu et al.*
9. Environment and Health Data *by Takashi Yanagawa*
10. Clinical Data *by Toshio Tango*
11. Sports Data *by Yuji Ohgi*
12. Financial Data *to appear*

The Number of Red Sea Breams Caught

n=40000 red sea breams released

The number






Smoothed line by lowess
 $\pm 2\sigma_i$

$$E(N_i) = np_i$$

$$\sigma_i = Sd(N_i) = \sqrt{np_i(1-p_i)}$$

Statistician  “Over Dispersion!”

What Statistician Does

- Change Distribution!
 - Binomial  Normal $N(np_i, \sigma_i^2)$, $i = 1, 2, \dots, 30$
- Introduce dependency!
 - Catches
 - Group of Brems
- Brems may escape from the region. Model it!
- Ignore the over dispersion!
- Something wrong! No way to do

What Data Scientist Does

- Check Background of Data

Release Date: 1989-09-30

Survey Period: 1989-10-01~1989-10-30

The numbers: Reported by Fishermen:

How many marked sea breams caught

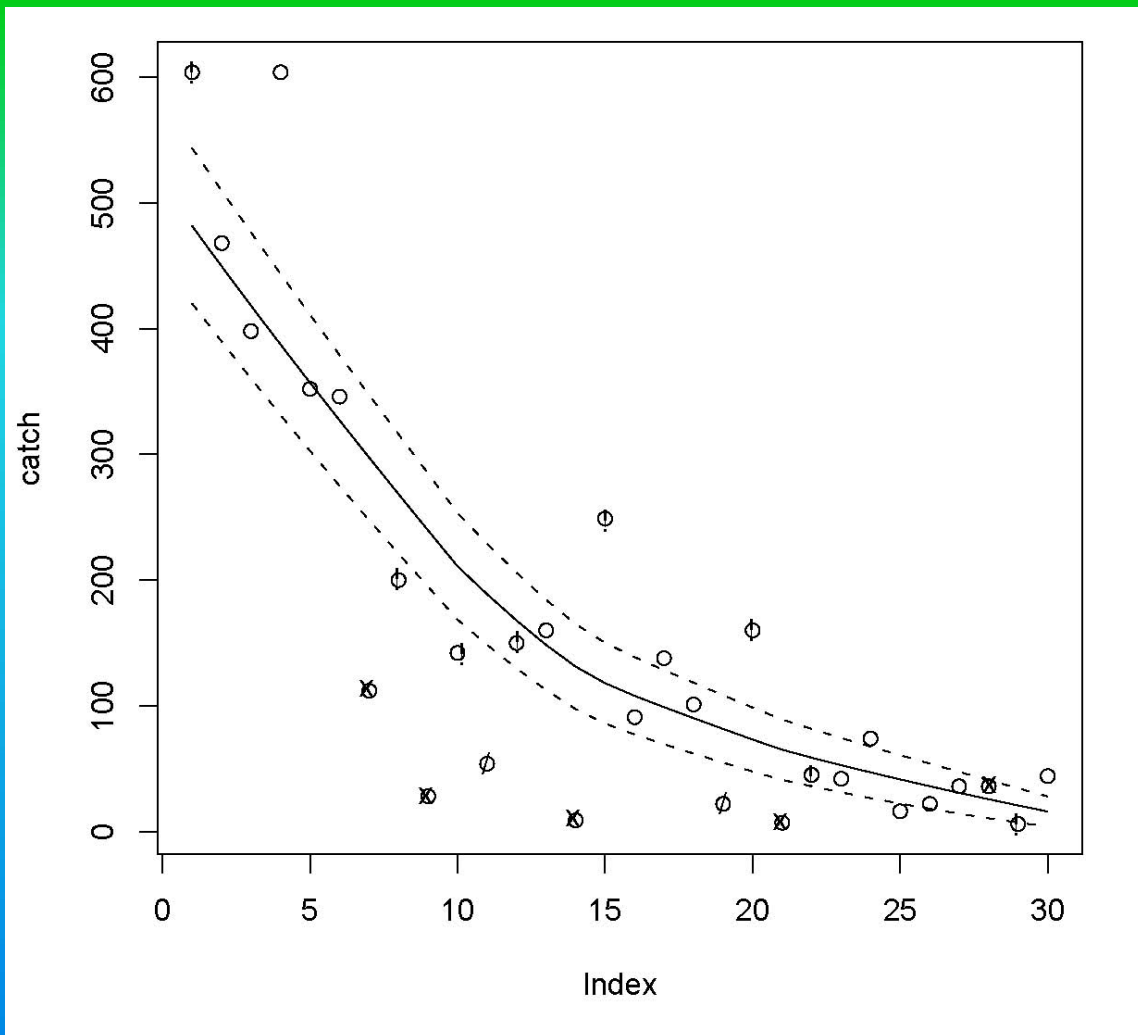
Sun	Mon	Tue	Wed	Thu	Fri	Sat
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

Sun	Mon	Tue	Wed	Thu	Fri	Sat
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

7,9,14,21: Saturday or Before Holiday

Fish Market Closes on Sunday and Holidays

11,19: Cold Days



\times : *Saturday or Before Holiday*

$|$: *Sunday, Holiday, Cold Days and Next Days*

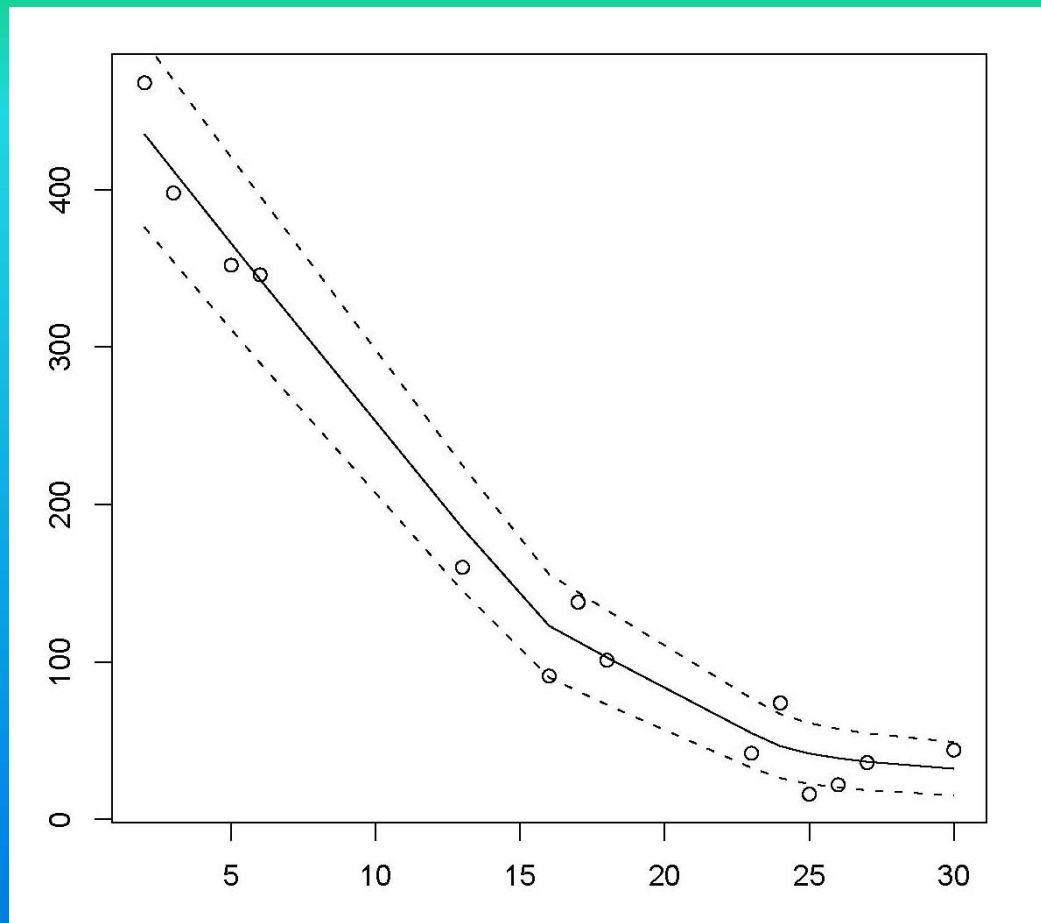
1989-09-04

Check Original Figure!



604	468	398	604	352	346	112	200	28
142	54	150	160	9	249	91	138	101
22	160	7	45	42	74	16	22	36
36	6	44						

Data Scientist's Job



Bernoulli Model Still Works!

Fisherman is also happy!



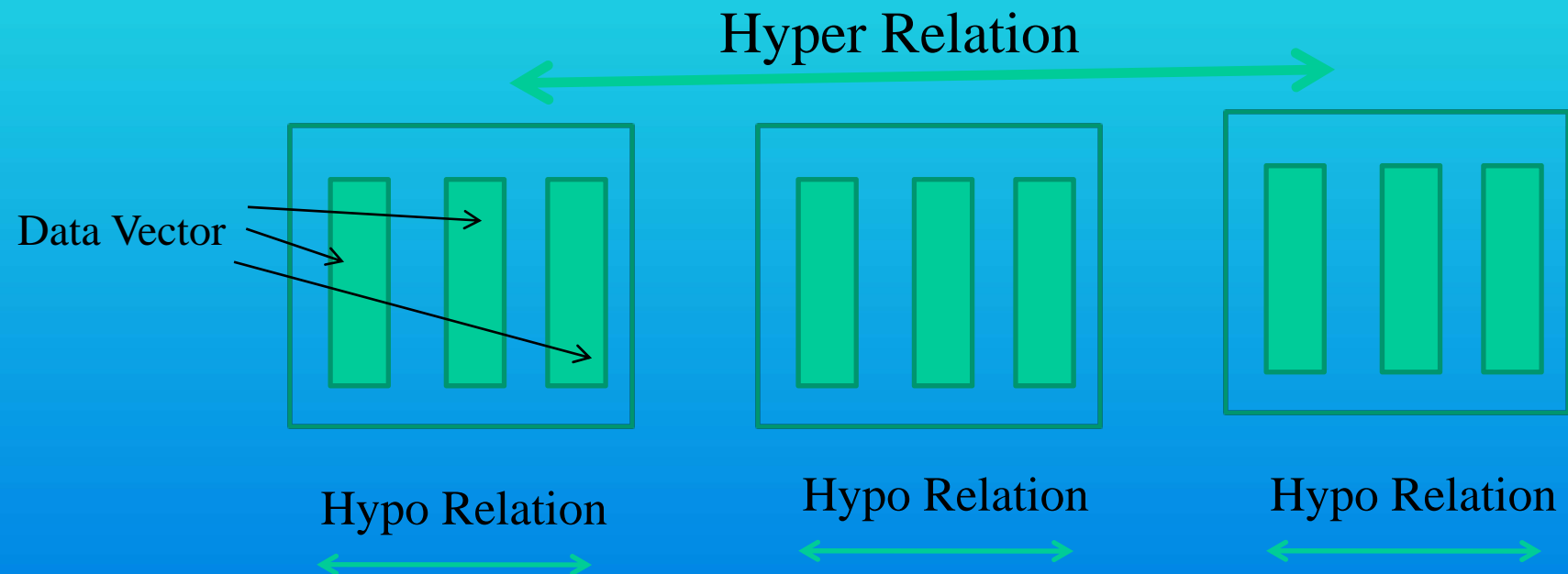
This case study suggests

- Better to know the background of data
 - Fish Market, Cold Days
- Possible miss-operation of data
- Total number of catches missing; p_i : *not real survival rate*
 - Need of redesign of the sampling
- Data is first, stochastic is next
- Good reasoning is necessary in any stage as a Science

Need of Data Literacy Beyond Statistics

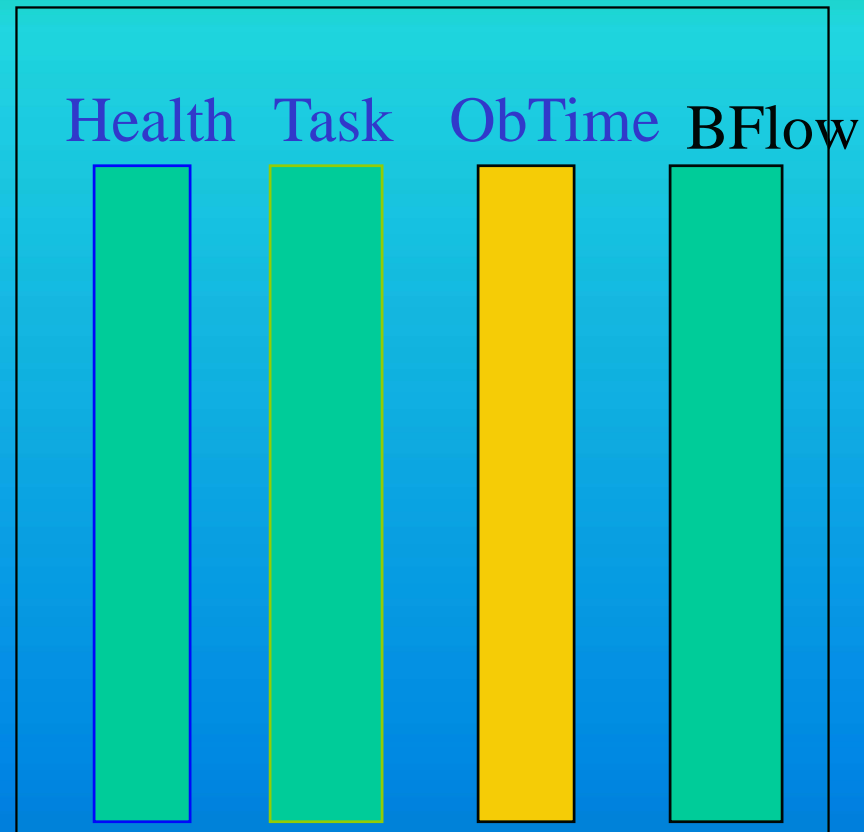
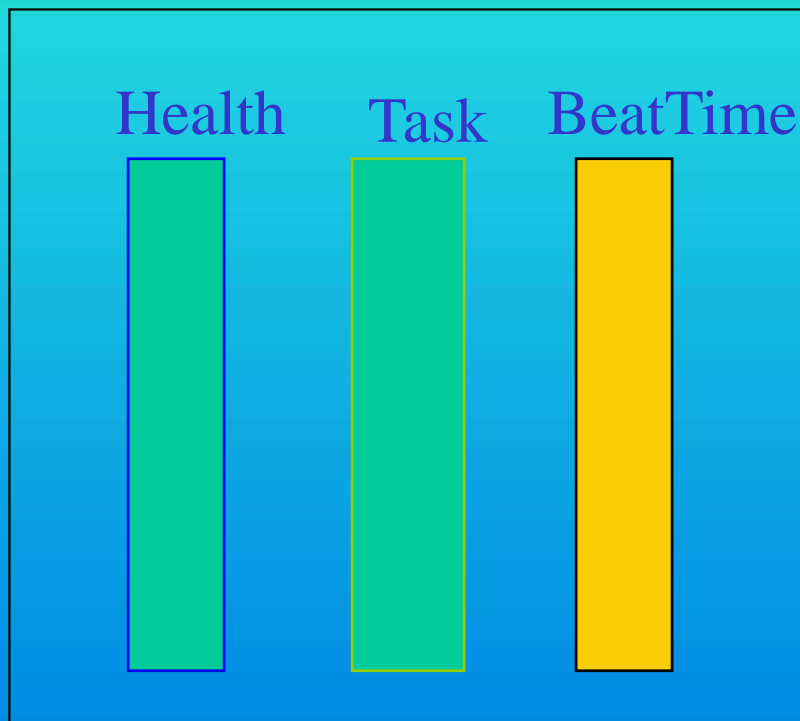
Fundamentals of Data Science

- Abstraction of Data



- Hierarchy of Attributes

An example of Hyper Relation



Health, Task: **Shared Value**
BeatTime, BFlow: **Common Measurement**

Infrastructure of Data Science

- DandD (Data and Description) rule

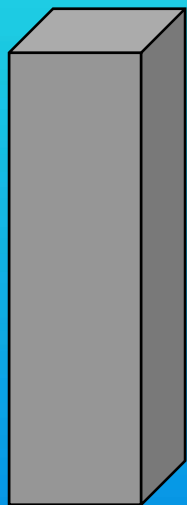
DandD
Instance
with
Data

DandD
Instance
Without
Data

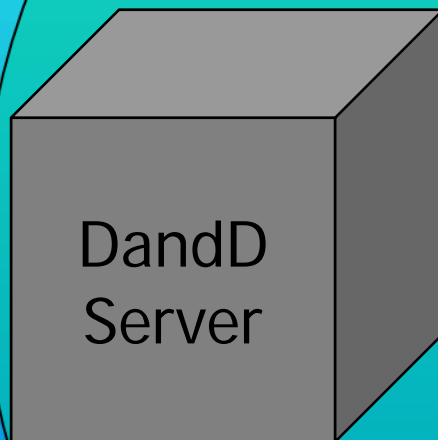
An XML document for organising Data with Description

Everything is described in it!
Works as an agent.

Client
Program

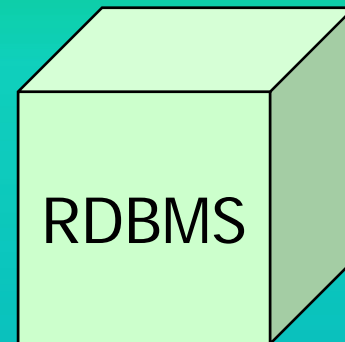


User



DandD
Instance
with
Data

Internet



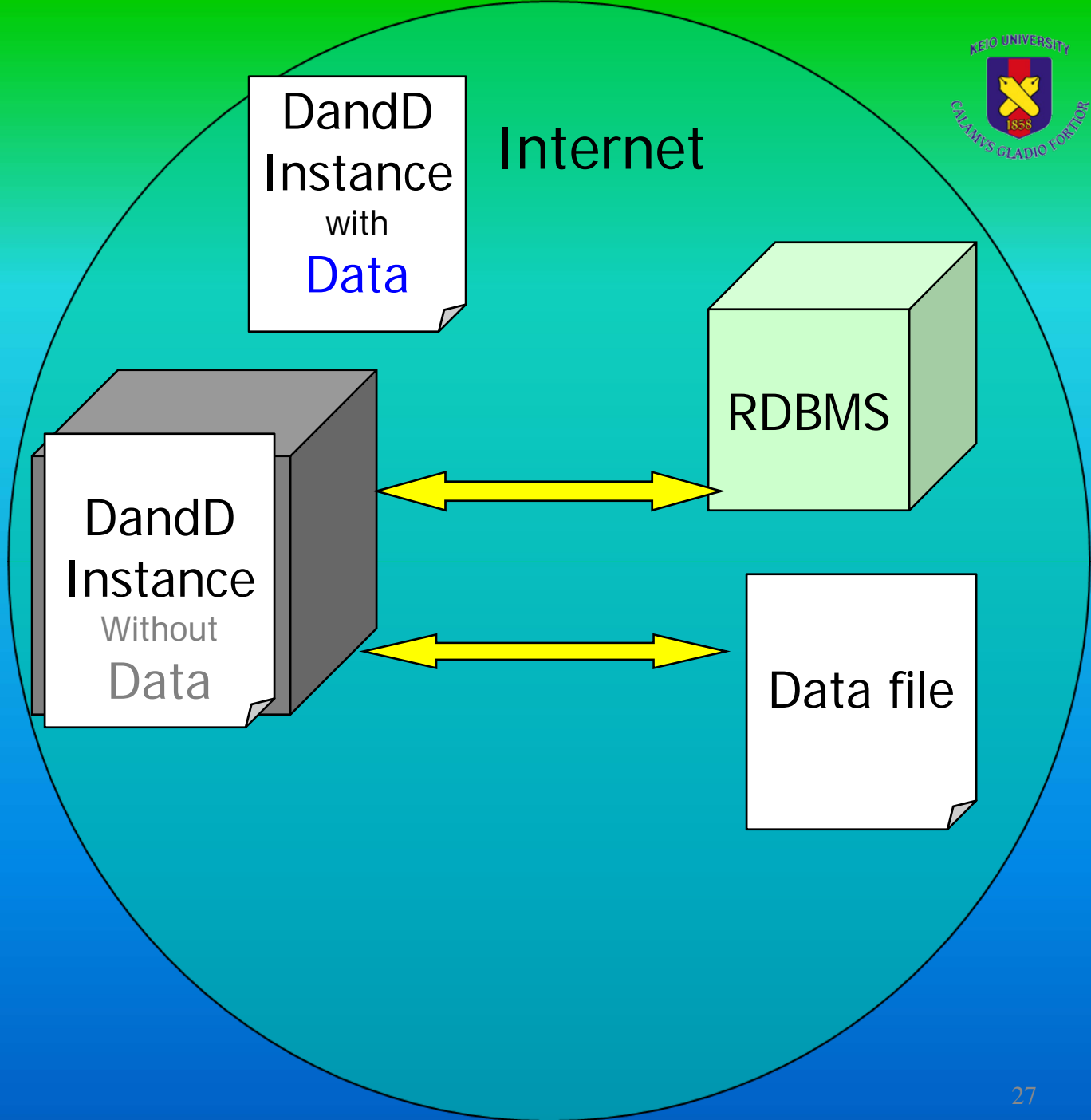
Data file

DandD
Instance
Without
Data

Client
Program



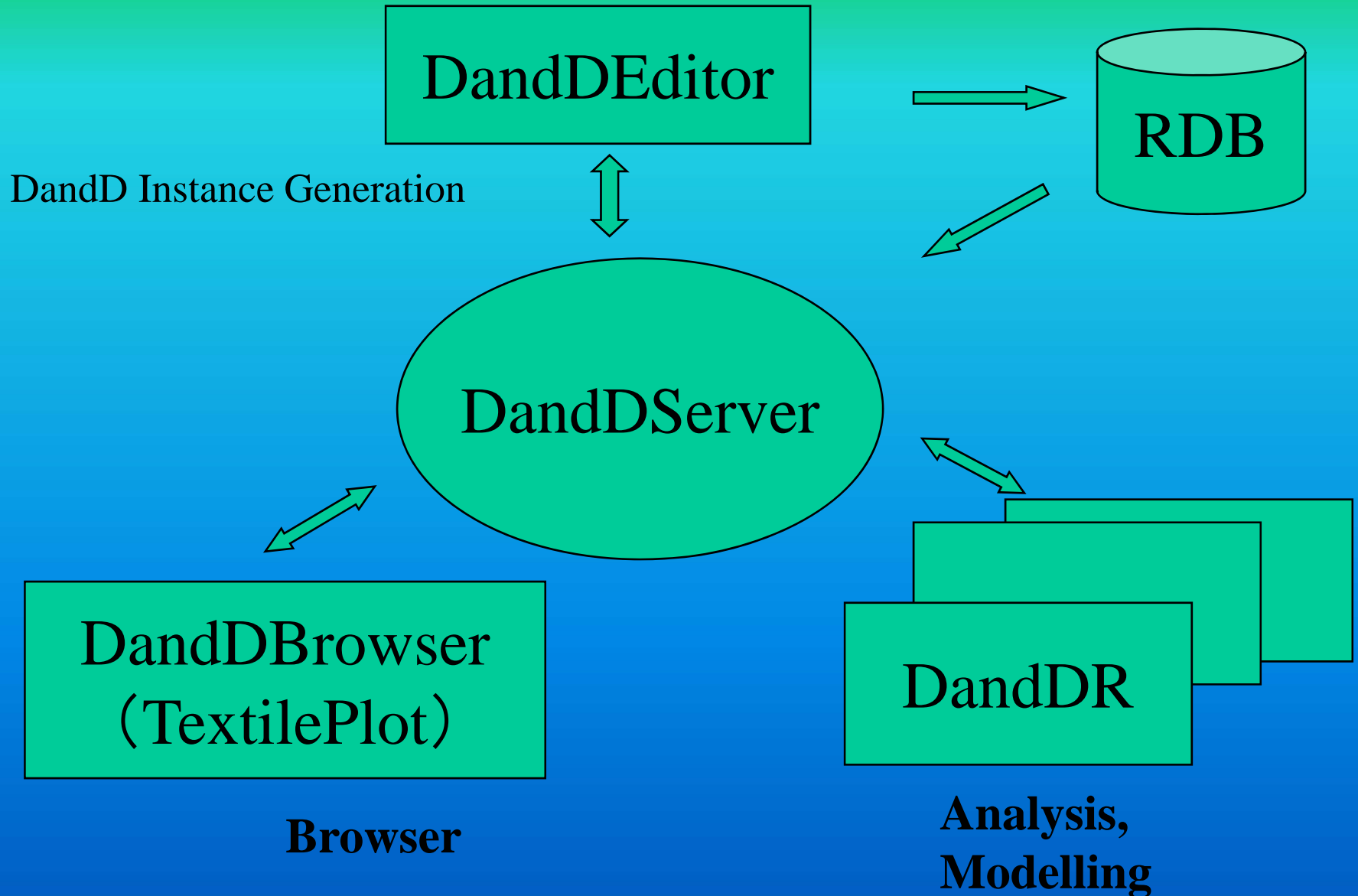
User



DandD Environment

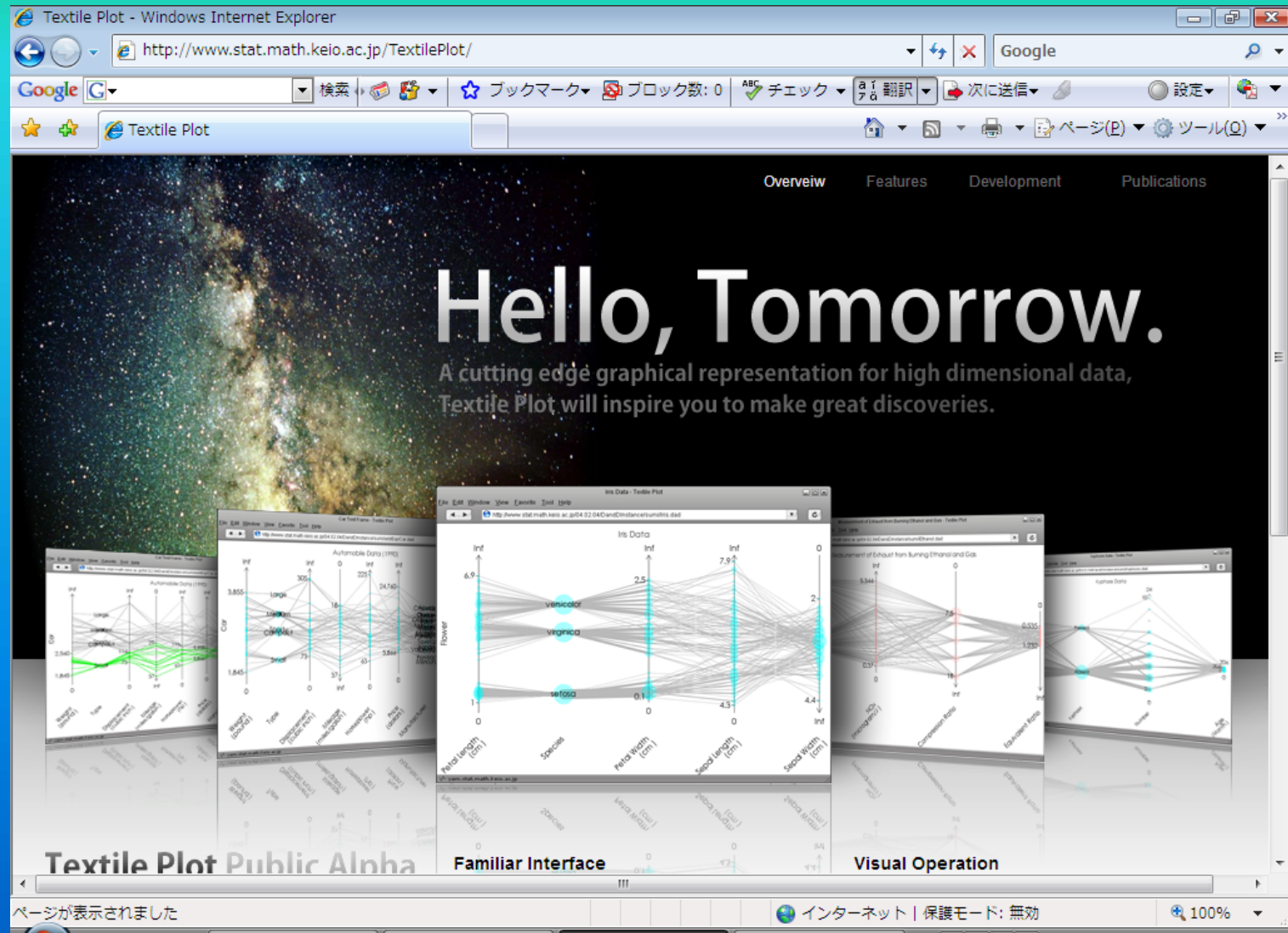


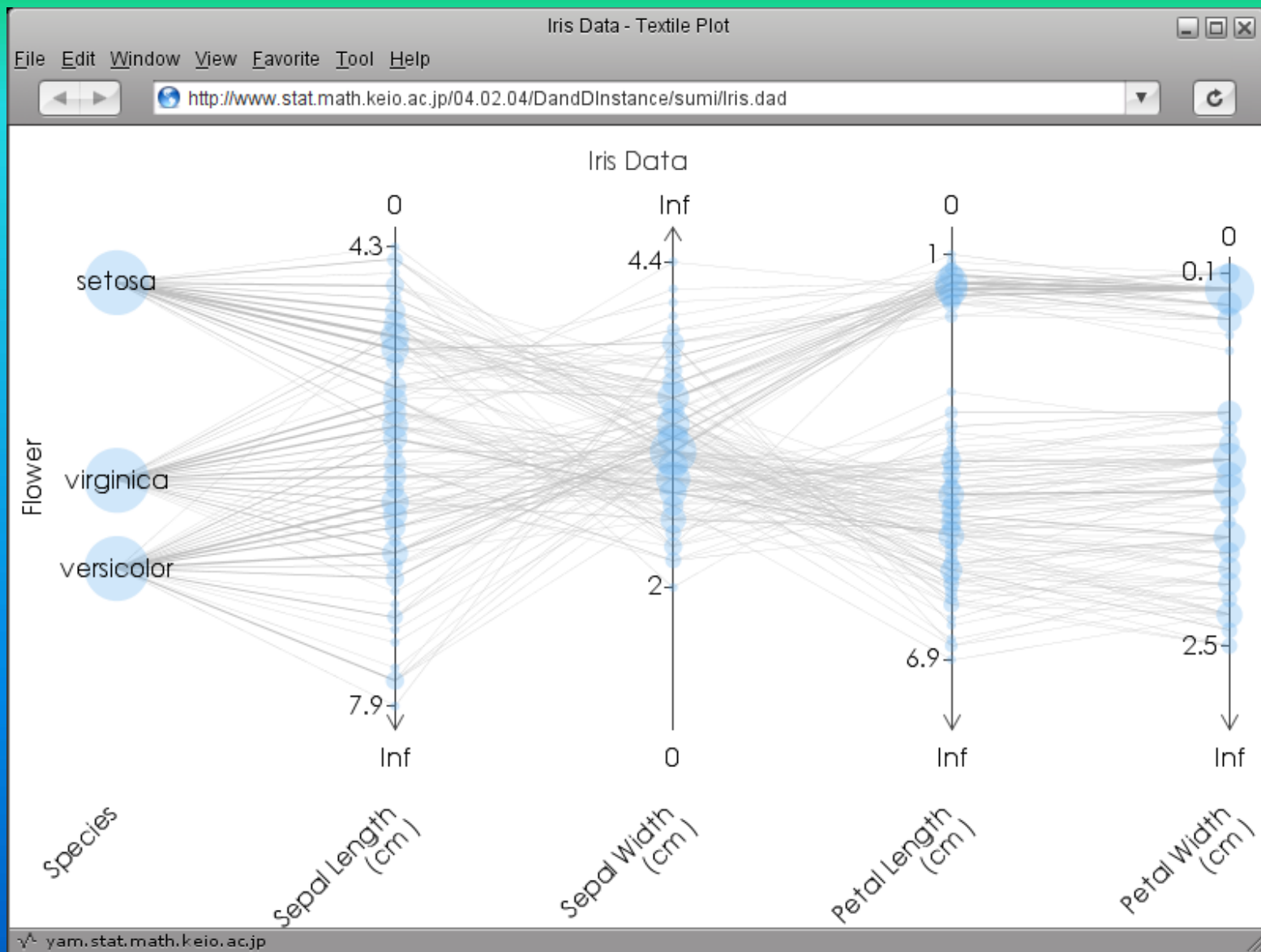
<http://www.stat.math.keio.ac.jp/DandDIV/>

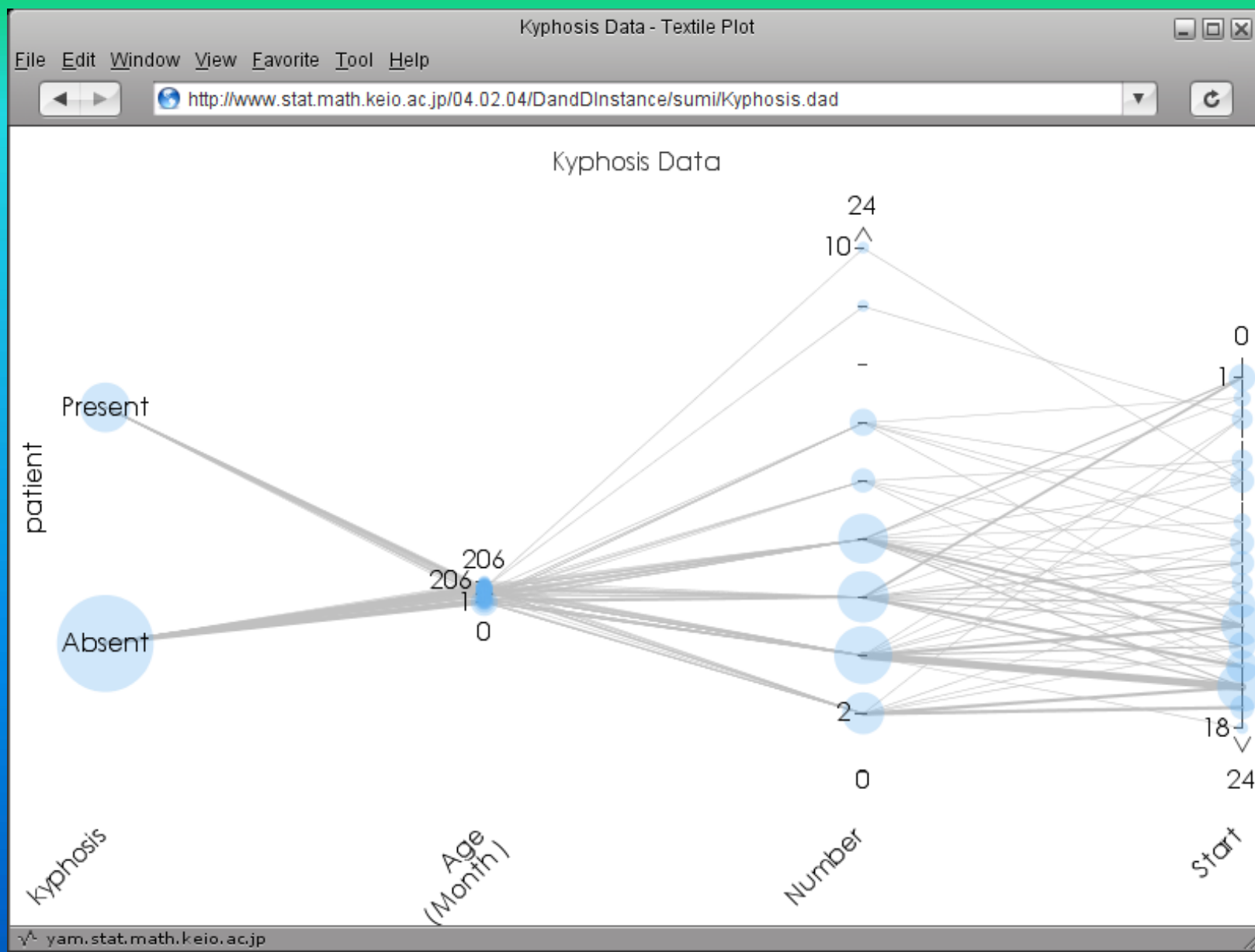


High Dimensional Data Visualisation

<http://www.stat.math.keio.ac.jp/TextilePlot/>







Integrative Environment for Discovery Through Data Science

- User Interface
 - Textile Plot
 - Data Manipulation
 - Model Fitting
 - Model Evaluation
- Hide Analysis Software
 - From Science of Methodology to Science of Data
 - Complex Huge Data

