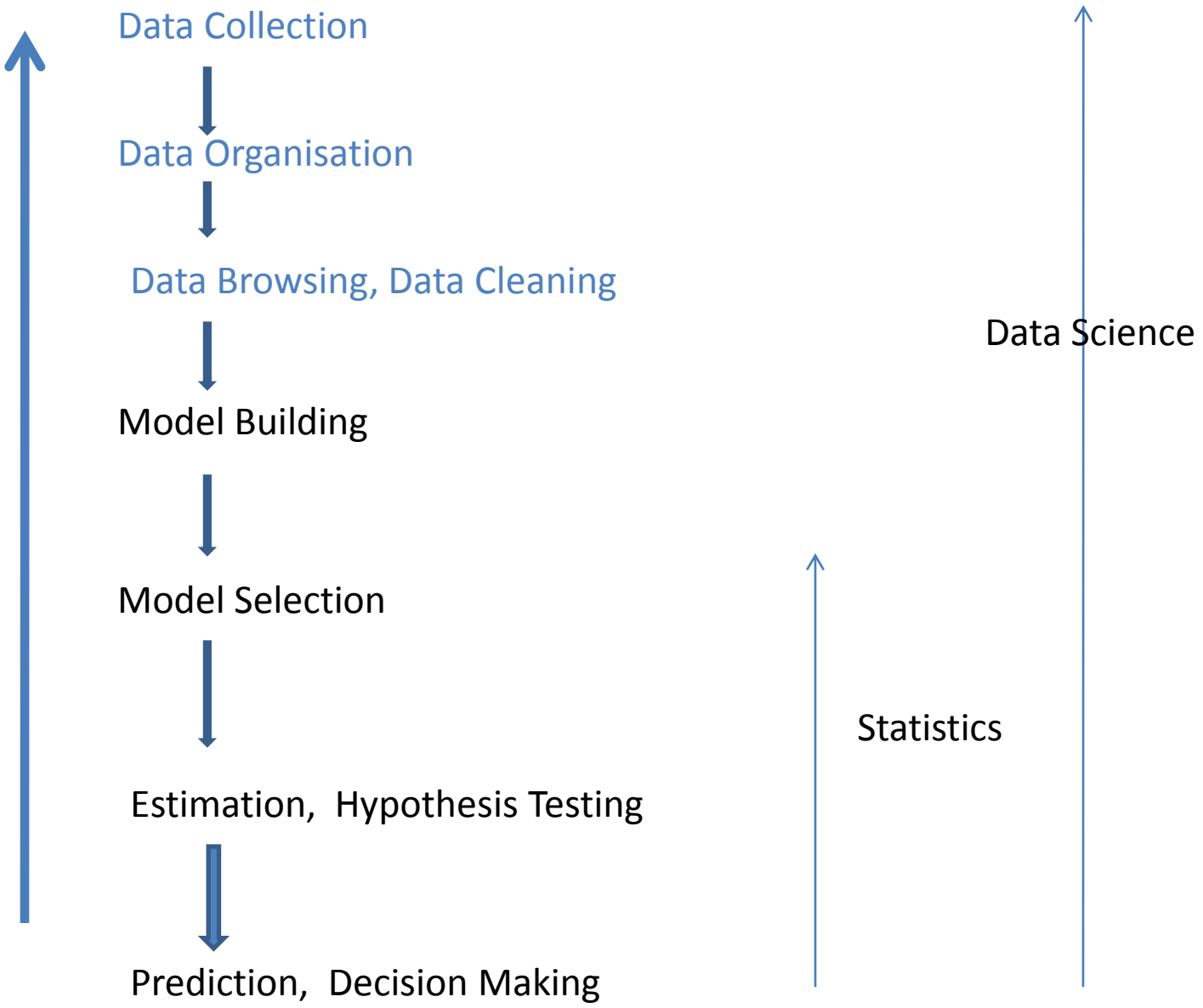
Ideas of DandD

Ritei Shibata

Dept. of Math., Keio University


Data Collection
Data Organisation
Data Browsing, Data Cleaning
Model Building
Model Selection
Estimation, Hypothesis Testing
Prediction, Decision Making
Data Science
Statistics

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

D and D

- Data and Description
  - Support Exchange of Data For Discovery
    - Self Explanatory, handy
    - Enough Background Information
    - Multi-Language
  - Support Metamorphism of Data
    - Traceback
    - Accumulate Experience in a formal way
  - Fudamentals of Data Science
    - Abstraction of Data General
    - Necessary and Sufficient Attributes

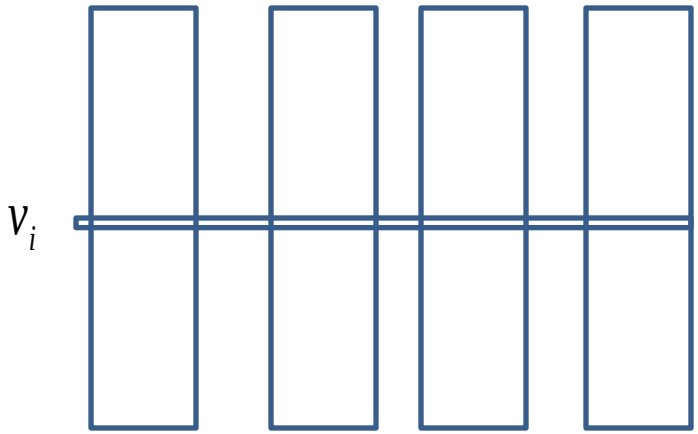
From Rule to Environment

- DandD Rule
  - DataBody+Attributes
- DandD Instance
  - An XML document
- Client-Server System
  - Server open for public on the Internet
- Environment
  - An Integrative Environment for Discovery through Data

Data Body

- Relational Scheme
  - Collection of Data Vectors: Relation
  - Collection of Records

$$R = \{v_1, v_2, ..., v_n\} \subset D_1 \times D_2 \times \cdots \times D_p$$



HypoRelation

- Overlooked in RDBMS
  - Native, Given, Profile attributes of the target object of the records
  - Radix
    - YYMMDD
    - Degree-Minutes-Seconds
  - Sum Constraint in a record
  - Exclusive in a record

HyperRelation

- Relation among Relations
  - Foreign Key in RDBMS

Company	Product	Price	Sold
A	Pen	10	10
B	Eraser	100	100
B	Ink	3	1000

Company	Address	Phone
A	Hiyoshi	045-561-0000
B	Yagami	045-555-1111
C	Fujisawa	046-322-1211
D	Mita	03-3473-1234

Example: Time Series and Point Process

- Amount of breathing (Sampling rate: 0.5 second)
- Heart Beat (peak time point)

Foreign Key is not enough to make a link between two relations

Original Data

	Beat Time	BFlow
1	0.516	0.04608150
2	1.200	-0.06378170
3	1.892	-0.00762939
4	2.596	0.10101300
5	3.268	-0.03997800
6	3.966	-0.09307860
7	4.656	-0.09979250
8	5.314	-0.09552000
9	5.976	0.03875730
10	6.650	0.13580300
11	7.328	0.12176500
12	7.982	-0.03204350
13	8.646	-0.08392330
14	9.342	-0.08178710
15	10.068	0.06744380
16	10.744	0.13641400
17	11.434	0.08575440
18	12.136	-0.02410890
19	12.836	-0.06469730
20	13.526	-0.07659910

sampling rate : 0.5 sec.

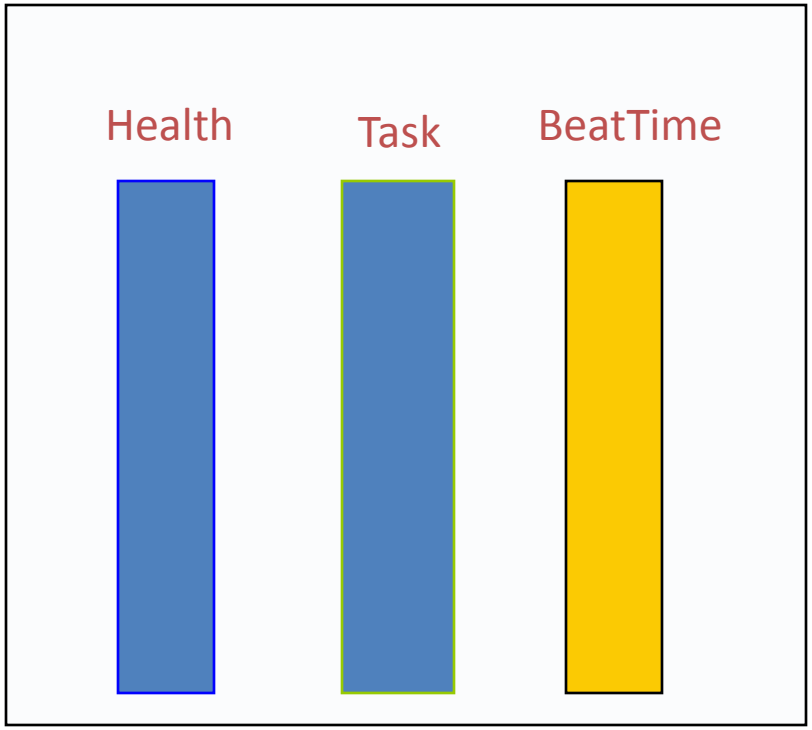
Point Process

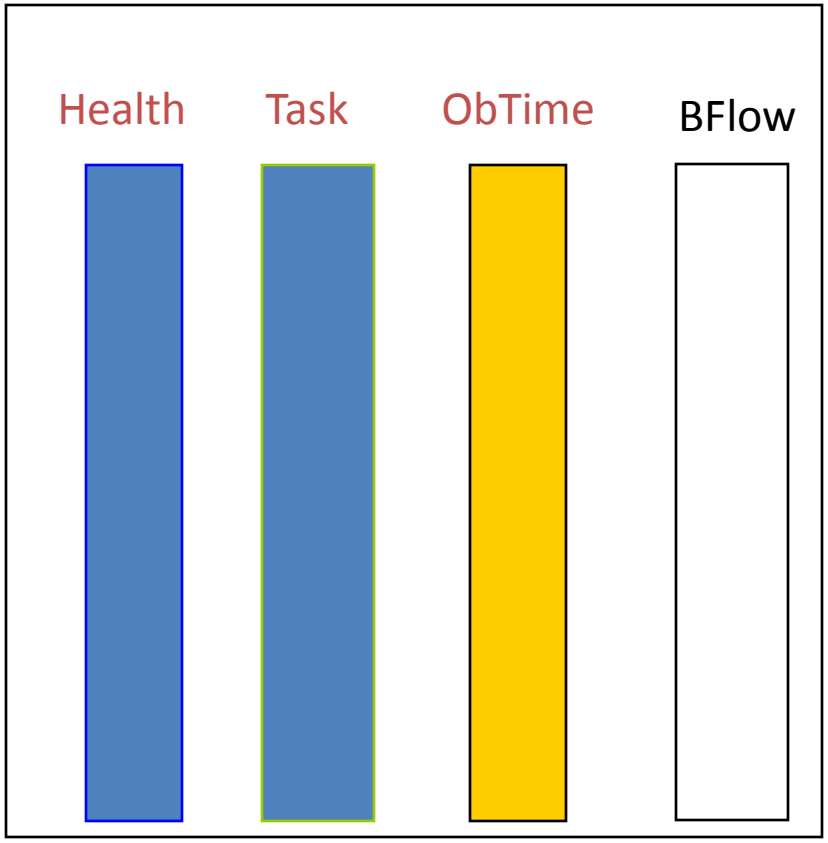
Beat Time
0.516
1.200
1.892
2.596
3.268
3.966
4.656
5.314
5.976
6.650
7.328
7.982
8.646
9.342
10.068
10.744
11.434
12.136
12.836
13.526

Time Series

ObTime	BFlow
0.5	0.04608150
1.0	-0.06378170
1.5	-0.00762939
2.0	0.10101300
2.5	-0.03997800
3.0	-0.09307860
3.5	-0.09979250
4.0	-0.09552000
4.5	0.03875730
5.0	0.13580300
5.5	0.12176500
6.0	-0.03204350
6.5	-0.08392330
7.0	-0.08178710
7.5	0.06744380
8.0	0.13641400
8.5	0.08575440
9.0	-0.02410890
9.5	-0.06469730
10.0	-0.07659910

HyperRelation

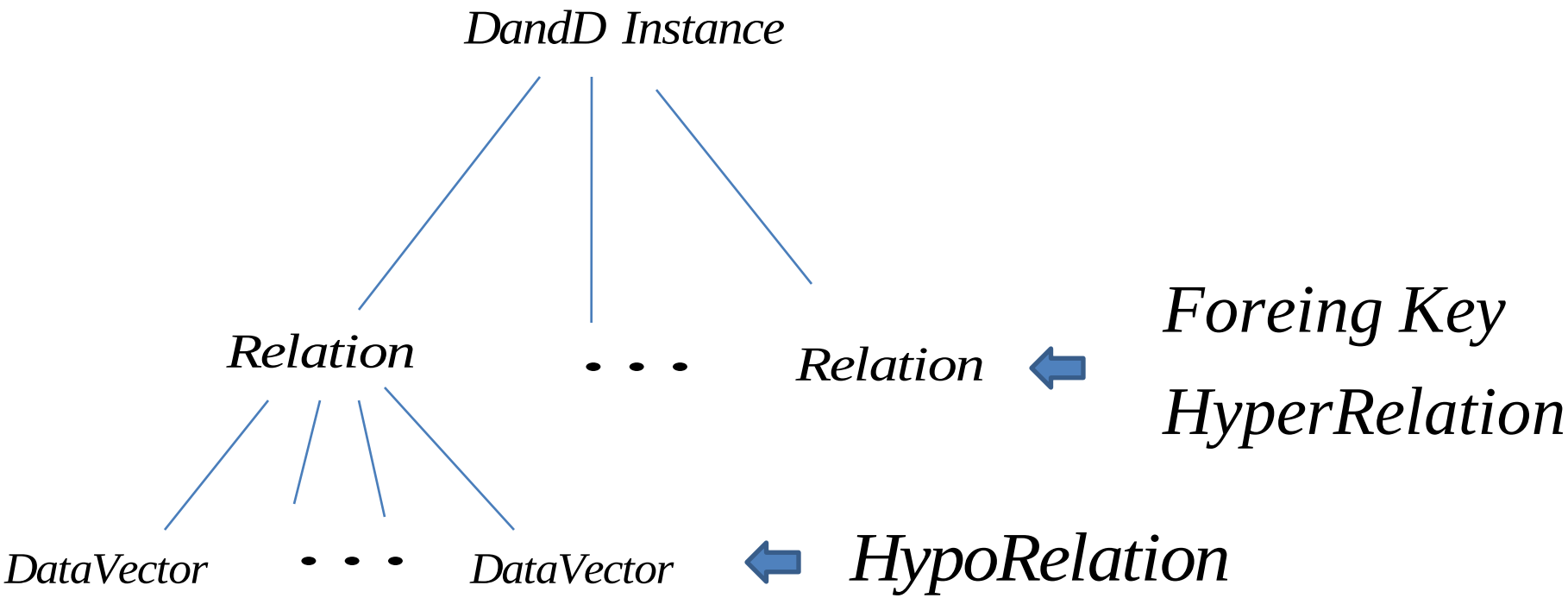




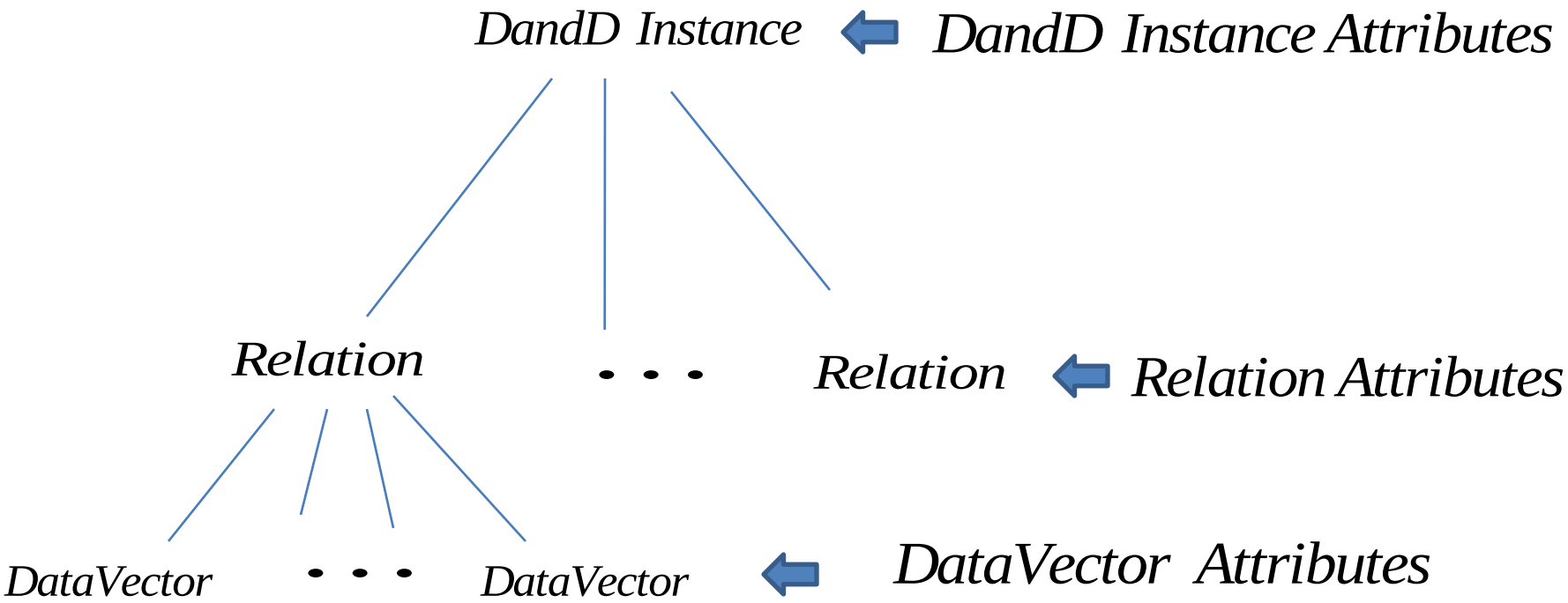
Health, Task: **Shared Value**

BeatTime, ObTime: **Common Measurement**

Structure of Data Body



Attributes


DandD Instance
DandD Instance Attributes
Relation
Relation
Relation Attributes
DataVector
DataVector
DataVector Attributes

Data Type: A Data Vector Attribute

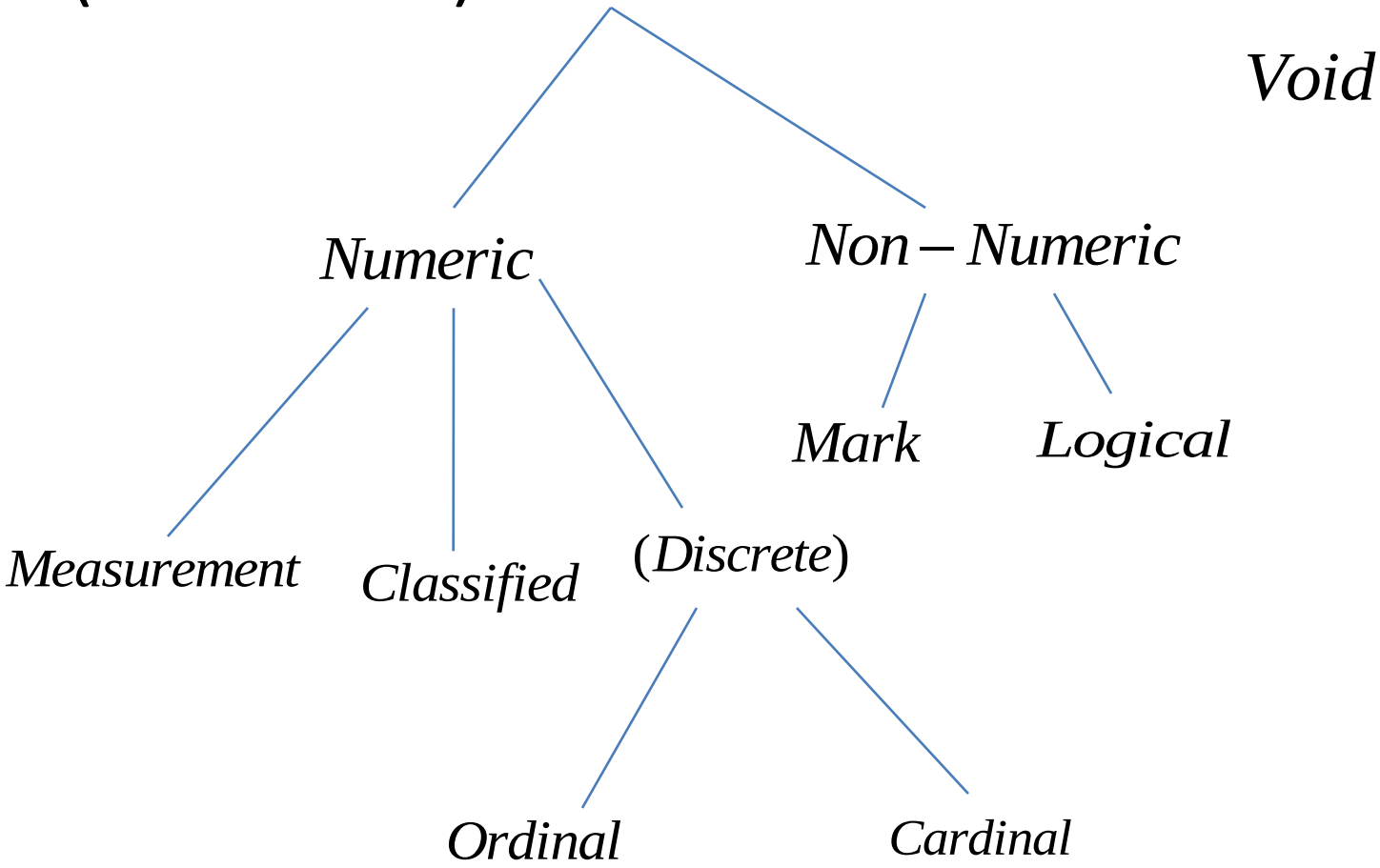
- Type common through a data vector
  - Known Classifications
    - Categorical, Numerical (For modelling)
    - Measurement, Count (Way of observation)
    - Continuous, Discrete (Computation)
    - Quatitative, Qualitative (Semantic)

$$0,1,\ldots : Count\,?\;\; Discrete\,?$$

$$0,1,1,0,1 :\;\; Logical\,?$$

$$0,1,3,3.2,1.2\ldots : Measurement\,?$$

- Data Type is not completely free from syntax
- Systematic Determination of Data Type (Thesaurus)?



Numeric

- Measurement
  - Always accompany with a unit
- Classified
  - Always accompany with the total
  - The result of aggregation
  - e.g. number of responses to each questionary
- Cardinal
  - Cardinal numeral(Count) , Non-Negative Integer
  - e.g. Number of people, Number of words, ...
- Ordinal
  - Ordinal Numeral, Positive(Non-Negative) Integer
  - e.g. Car No. in a Train, Batting Order, Age
  - Special case of Ordered Category

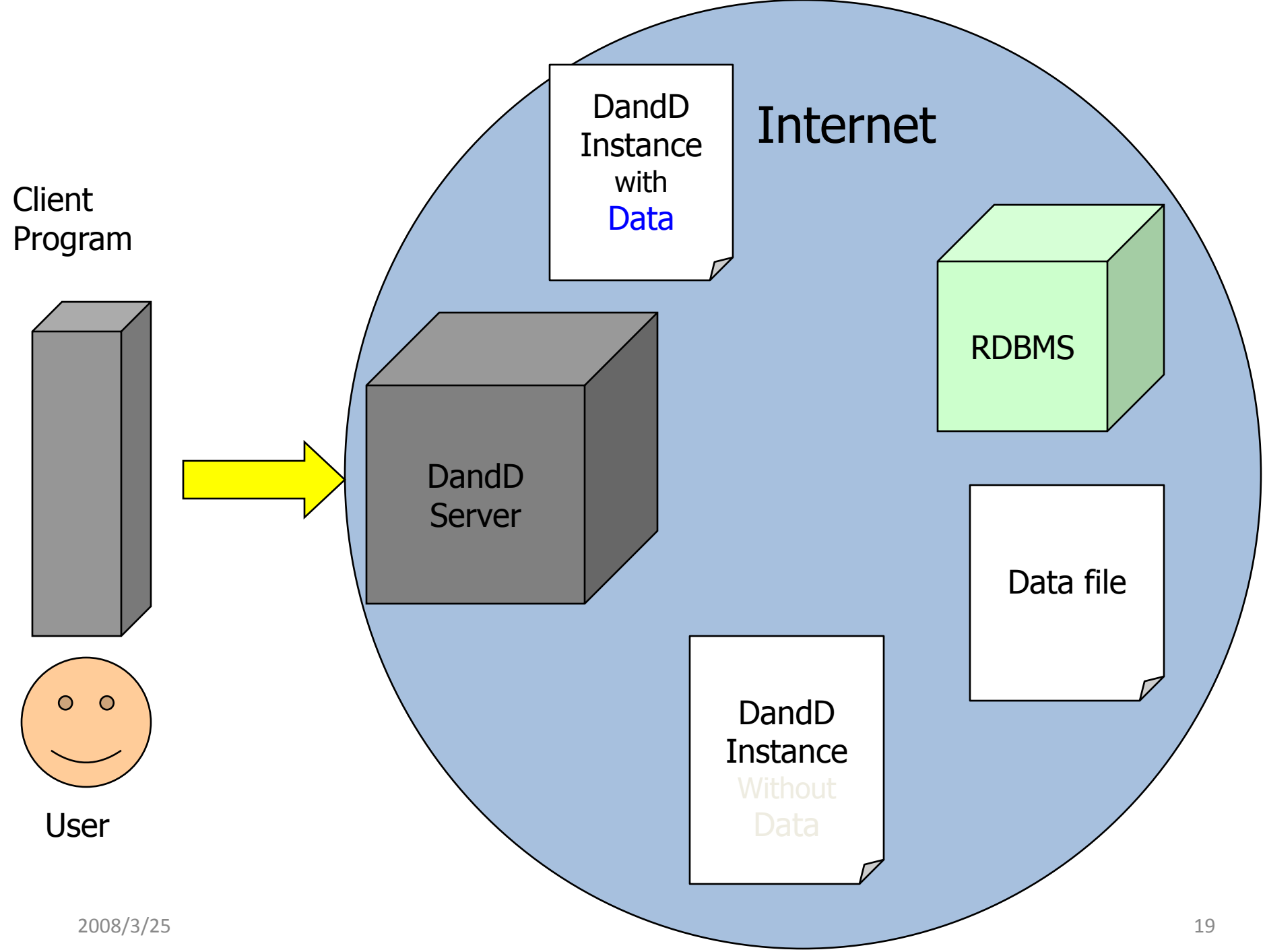
Non-Numeric

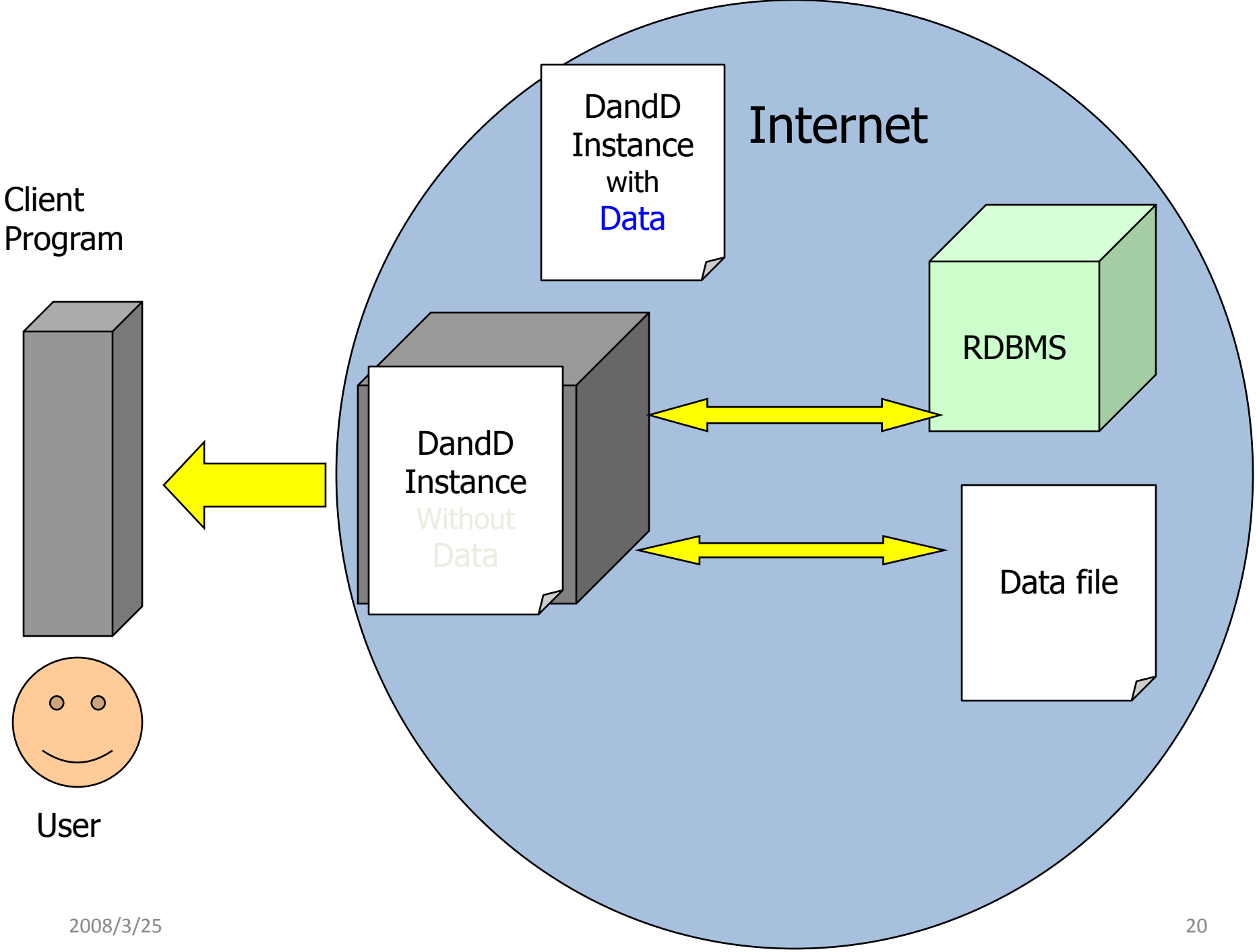
- Logical
  - 0 or 1, True or False
- Mark
  - Category
  - 0,1,2,3,.... (Non Numeric Number!)
  - Does not distinguish Ordered or Not
    - Too Semantic
    - Changeable

$Attribute : Ordered$

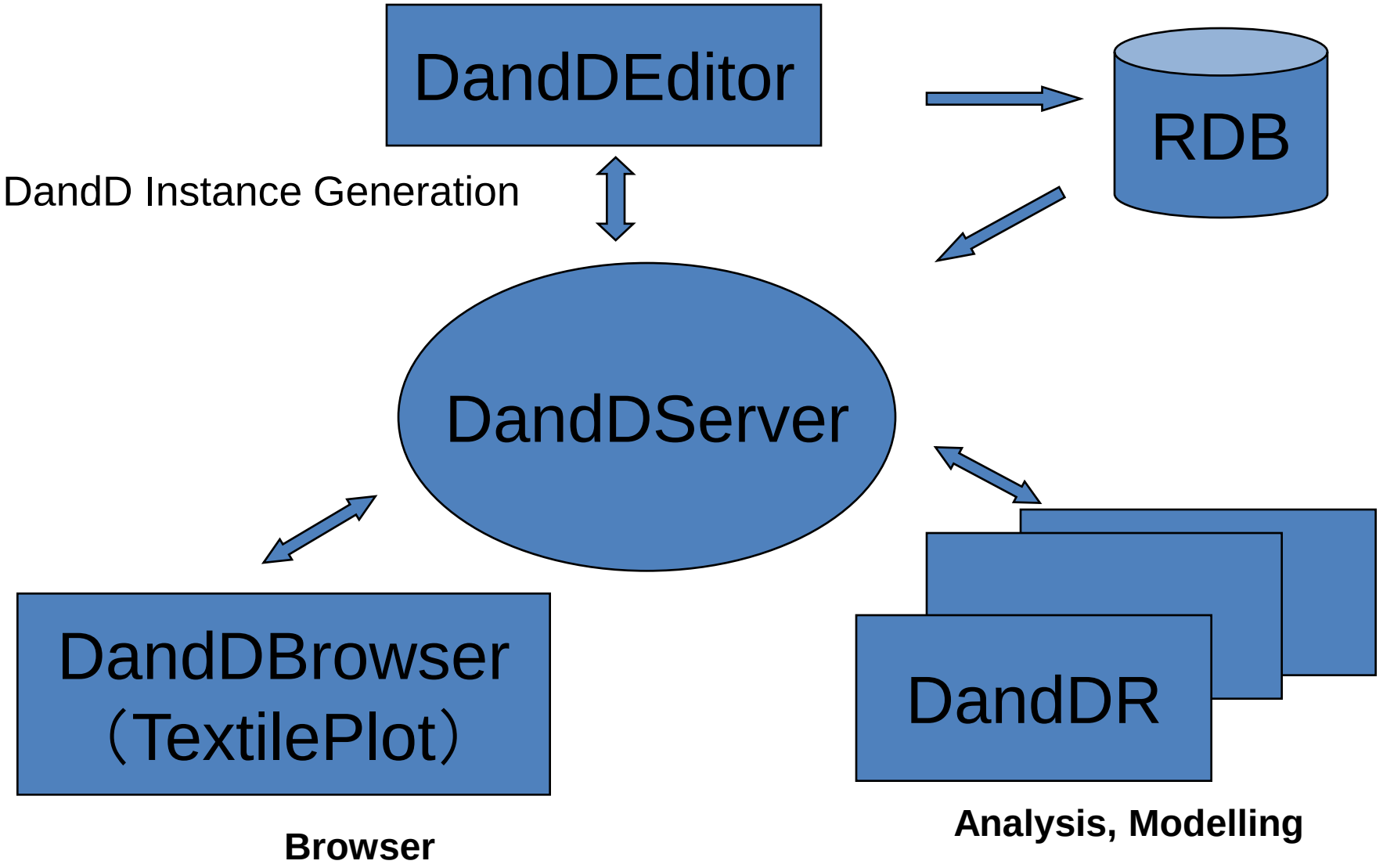
DandD Instance and Client-Server System

- Values in DataVector can be anywhere
  - RDB
  - Web
  - File
- DandD Instance Keeps Structure of Data Body and its Attributes as an XML instance


Client
Program
User
DandD
Server
DandD
Instance
with
Data
Internet
RDBMS
Data file
DandD
Instance
Without
Data


Internet
DandD Instance with Data
Client Program
DandD Instance Without Data
RDBMS
Data file
User

DandD Environment



DandD Home Page:

http://www.stat.math.keio.ac.jp/DandDIV/index.html

TextilePlot Home Page:

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

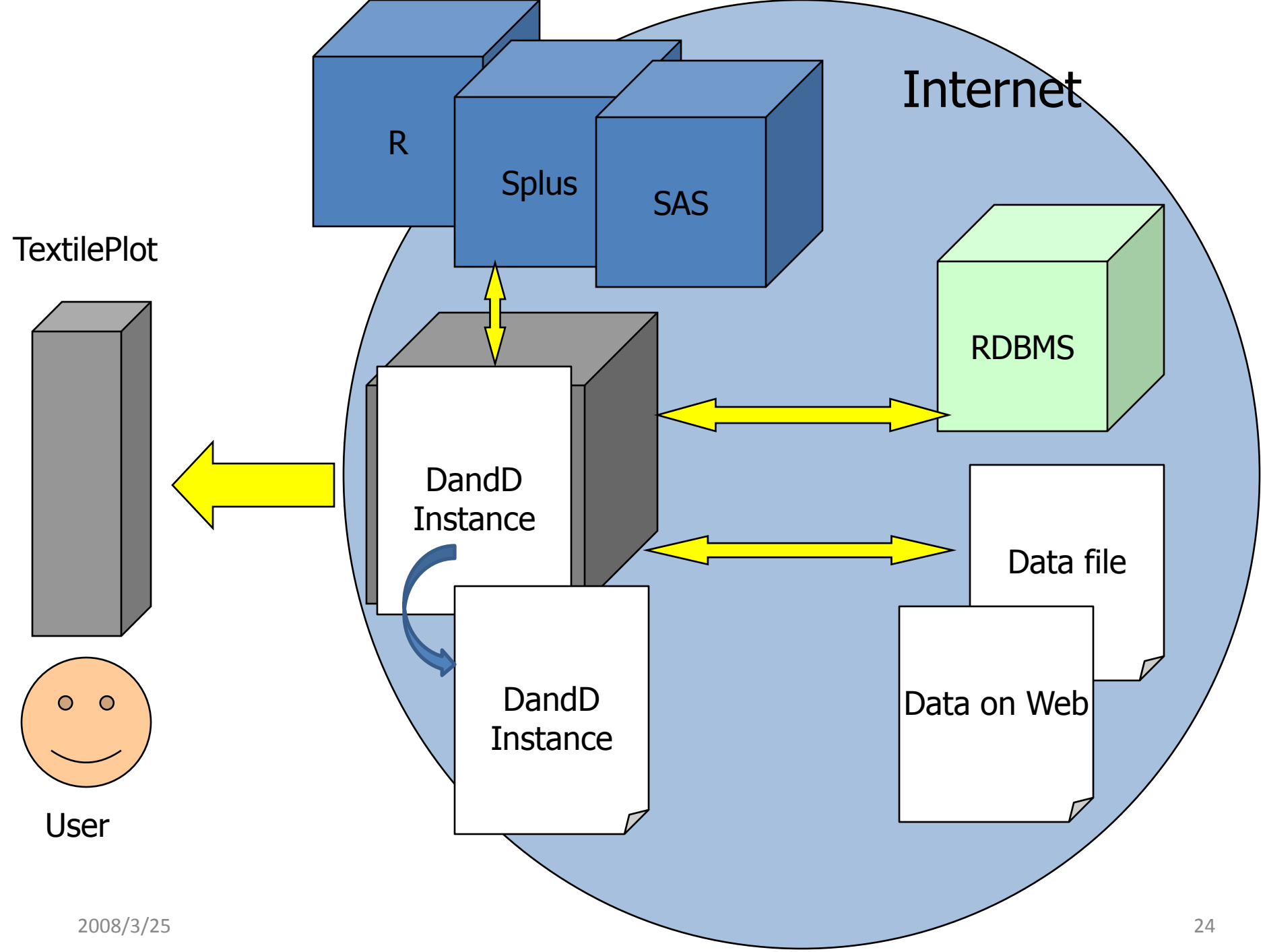
17:00~ 18:00 : Demonstrations: Exploration of DandD World

DandD environment for financial data by Daisuke Yokouchi

DandD instance generation in the textile plot environment by Natsuhiko Kumasaka

Future Works: 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


Internet
R
Splus
SAS
TextilePlot
RDBMS
DandD
Instance
Data file
DandD
Instance
Data on Web
User