


A General Framework for Big Data Analysis and Some of Our Work

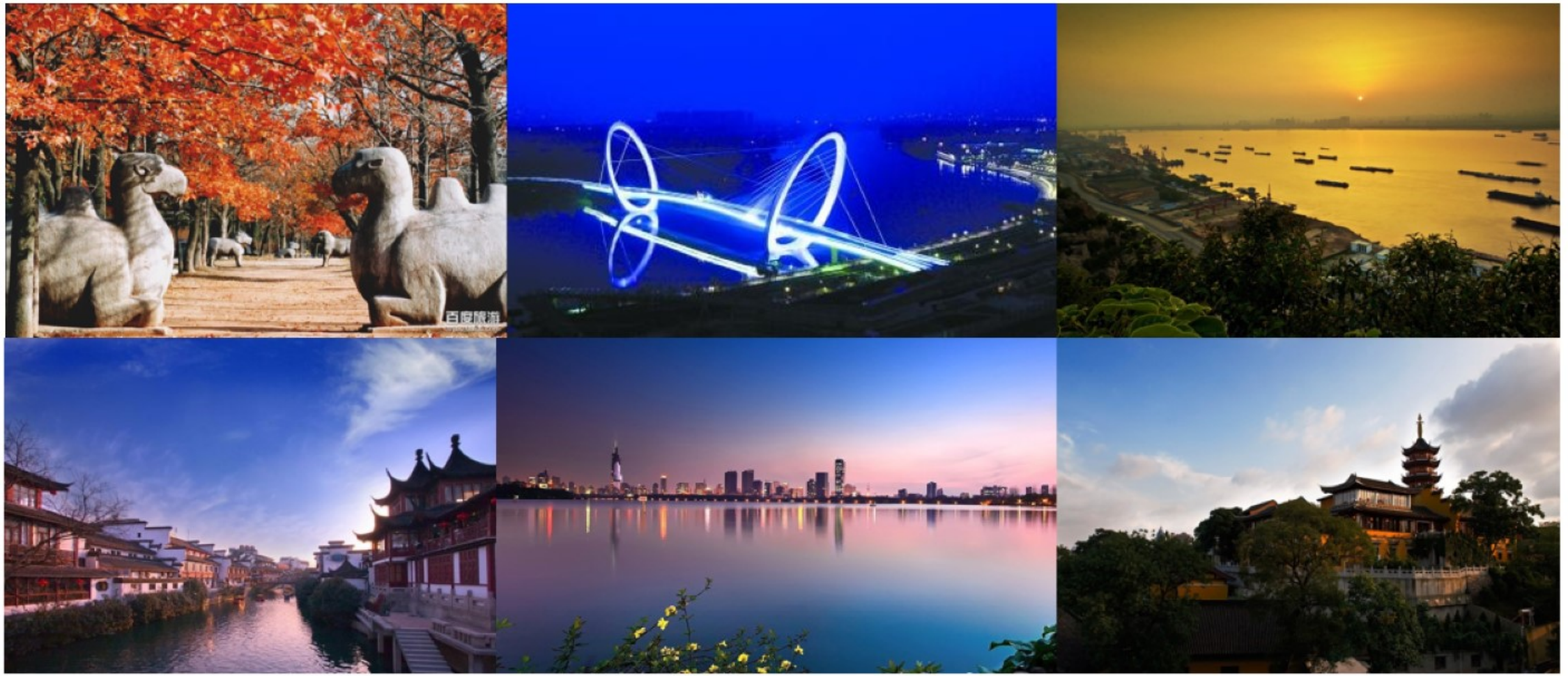
Department of Computer Science and Technology, Nanjing University
National Key Laboratory for Novel Software Technology
Collaborative Innovation Center of Novel Software Technology and Industrialization

Chongjun WANG

Arras

2016-6-6

Beautiful Scenery of Nanjing



Beautiful Campus of NJU

About Dept. of CS.

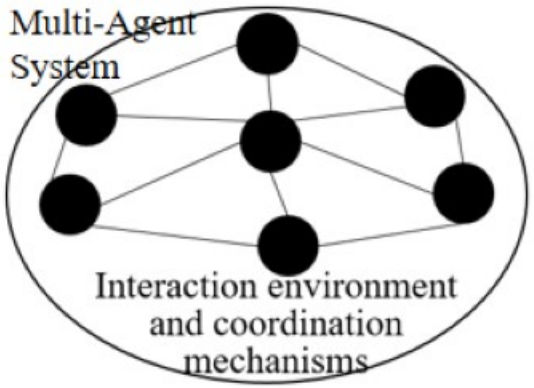
- Department of Computer Science and Technology, Nanjing University
  - There are over 120 full-time staff, including
    - 43 professors;
    - 42 associate professors;
    - Over 20 lecturers;
    - Over 20 office staff;
  - Research areas of our dept. are listed as follows
    - Distributed Computing
    - Network space security
    - Database
    - Software Methodology
    - Software Engineering
    - Nature Language Processing
    - Artificial Intelligence
    - Multimedia processing

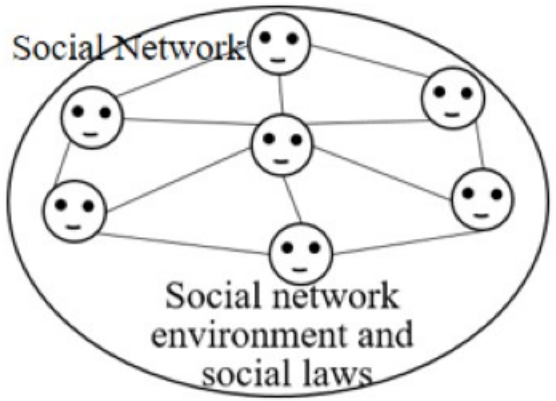
About IIP Group, NJU

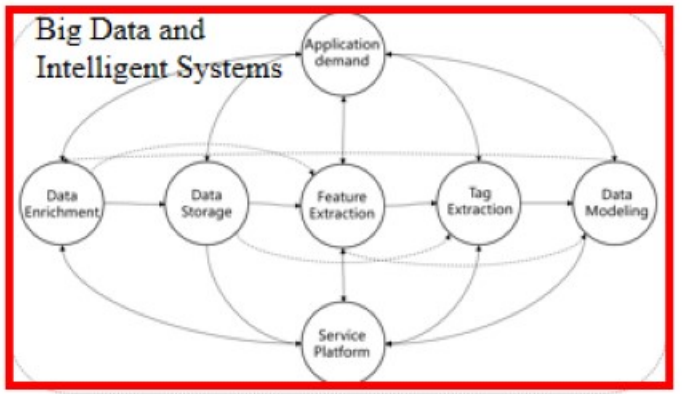
- Intelligent Information Processing Group, Nanjing University
  - There are 5 full-time staff, including
    - Dr. Prof. Chongjun, WANG
    - Prof. Junyuan, XIE
    - Dr. Associate Prof. Ning, LI
    - Dr. Assistant Prof. Jun, Wu
    - Dr. Assistant Prof. Lei, ZHANG
  - There are 50 students, including
    - 8 Ph.D. Students
    - 42 Master Students

About IIP Group, NJU

- Research Interest
  - Agent and Multi-Agent Systems
  - Complex Network Analysis
  - Big Data and Intelligent Systems







Outline

- Introduction
- Our thoughts on Big Data
- A General Framework for Big Data Processing
- Application Cases based on Big Data
- Summary

Introduction

Data → (extracted) → Information → (discovered) → Knowledge

The record used to describe the things, such as graph, voice, image, text, number, character and symbol.

Generally refers to the significance of the data message used to decrease the uncertainty of the thing.

$$H = \sum_{k=1}^{n} p_k \log_2 p_k$$

Those summarized from the social practice, which is believed to be correct.

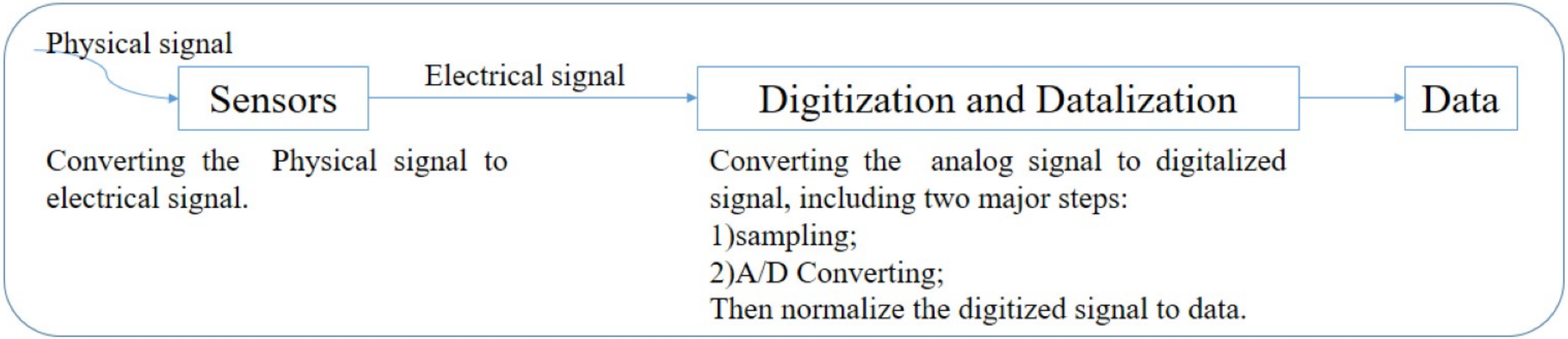
The knowledge or insight hidden in the data can bring more value.

More Data Needed

- Discovering knowledge or insight hidden in the data can bring more value.
  - –We can better know the world and the earth
  - –We can better know the society we are living in
  - –We can better know ourselves
  - –Etc
- In fact, every day, we are living in the data
  - –we produce all kinds of data
  - –we use the data to observe the world

People need more data to discover knowledge or insight…

From Physical Signal to Data



- With the advancement of technology
  - Digital means become more and more convenient.
  - Digital cost is getting lower and lower.
  - ……

This means that more and more types data can be digitalized easily.

Data to Big Data

- Above two reasons lead to a new times tagged as “Big Data” (not limited)
  - People need more data for their different value expectation.
  - More and more types of data can be digitalized easily.
  - ……

Big Data

- "Big Data" has become an increasingly topical subject, then what's Big data?
  - Big data usually includes datasets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.

X-V	Description
Volume	The quantity of generated and stored data. The size of the data determines the value and potential insight- and whether it can actually be considered big data or not.
Variety	The type and nature of the data. This helps people who analyze it to effectively use the resulting insight.
Velocity	In this context, the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development.
Variability	Inconsistency of the data set can hamper processes to handle and manage it.
Veracity	The quality of captured data can vary greatly, affecting accurate analysis.
Value	Data is useful and data is assets, knowledge discovered from data is useful and can create new value.

Examples

WAL★MART®

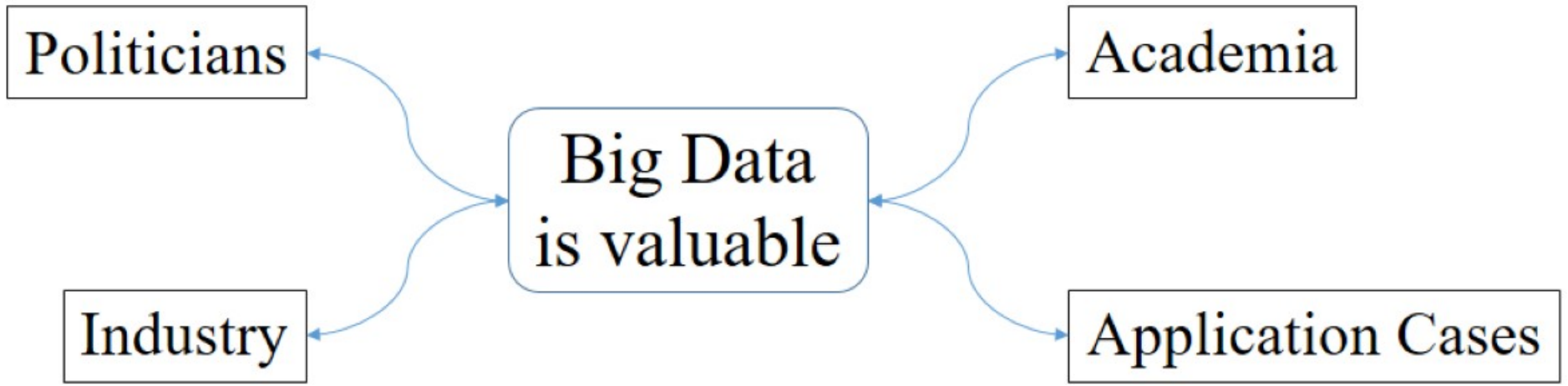
.....

Value of Big Data

- Value oriented from Data
  - Data is valuable, the original data is digitalized and stored only because of the original value expectation;
  - Then how to reuse, expand and promote the value of the original data.
  - Connecting and integrating all data from each data source maybe the most important.
- Value oriented from the big data platform
  - The big data platform is for data gathering, data modeling, data severing, computation severing and so on.
  - Application cases need data and the value of the platform comes from different application cases.
  - needing driven from data.

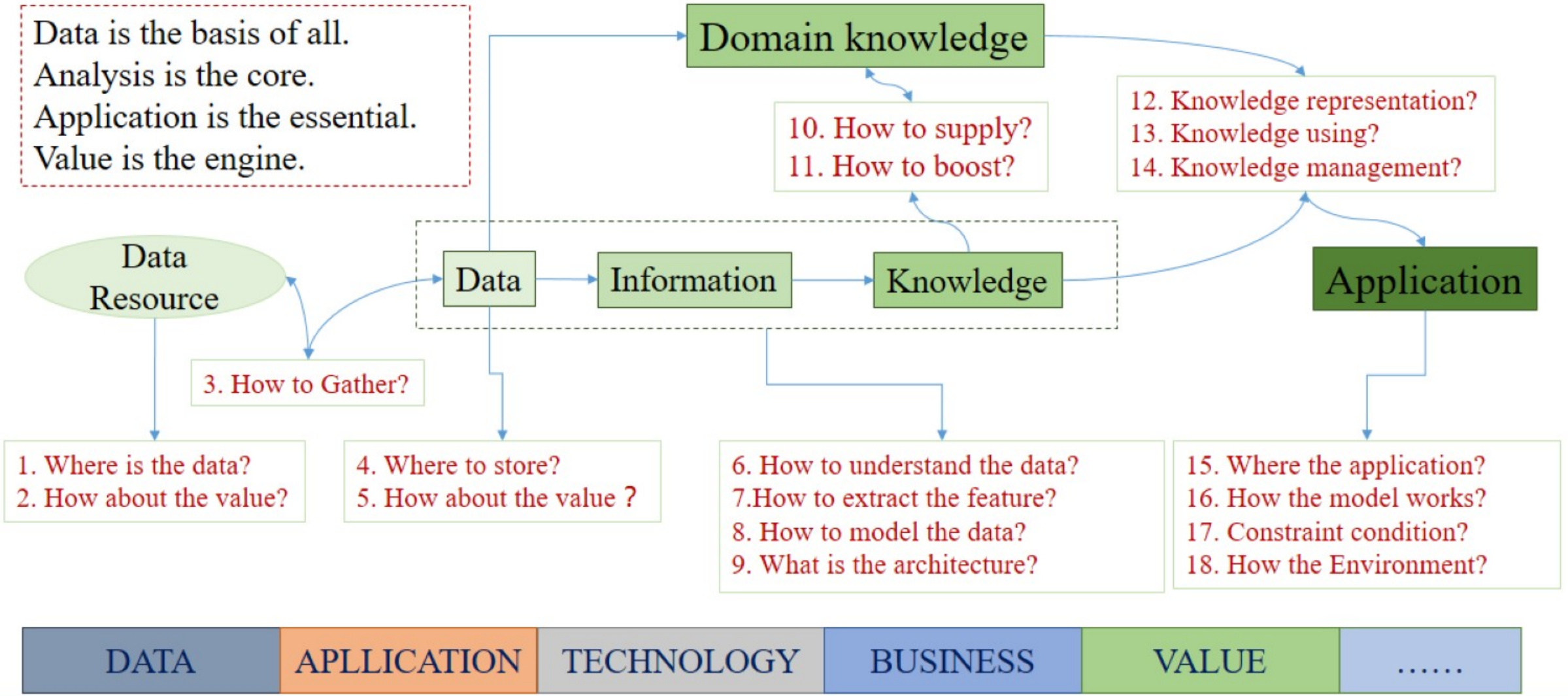
Value of Big Data

- Certainly, value is subjective, which means,
  - Different people/roles have different value expectations, thus,
  - Different people/roles use different strategies and actions
  - etc.



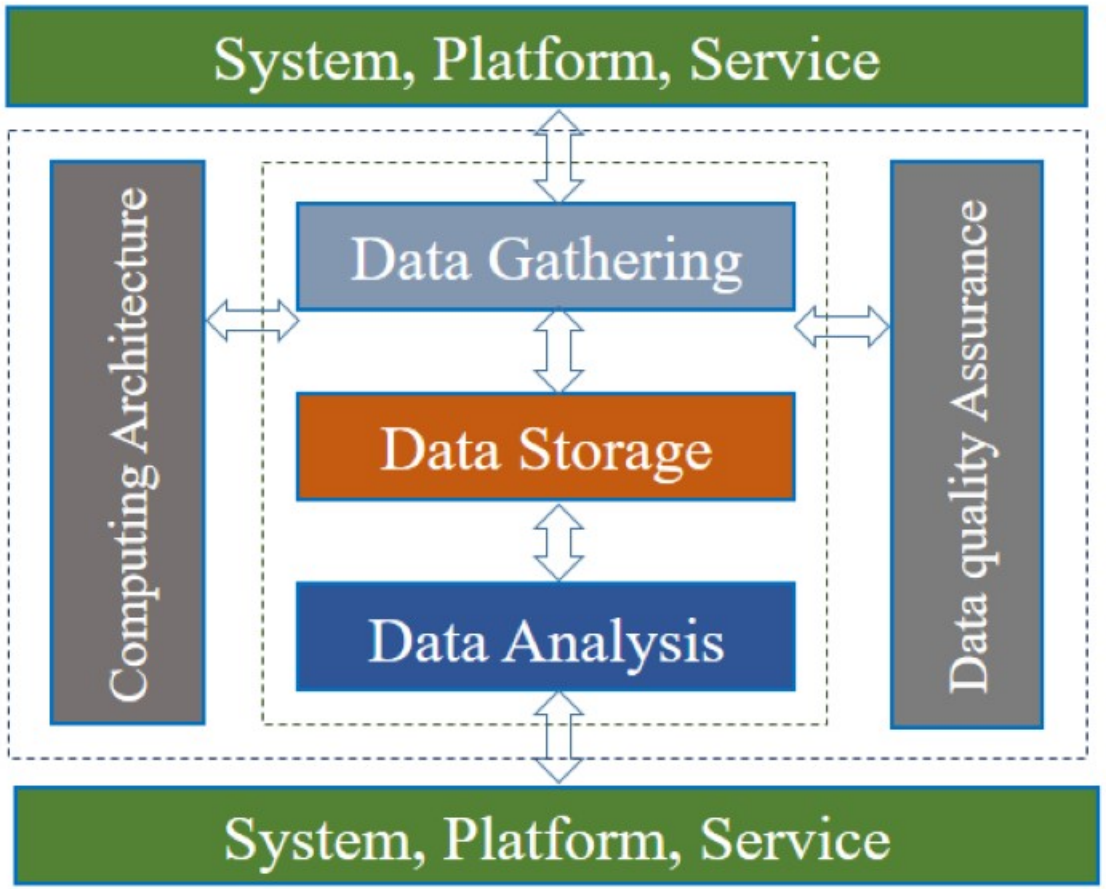
How to realize the value expectation?

Big Data for Application Cases



DATA | APLLICATION | TECHNOLOGY | BUSINESS | VALUE | ……

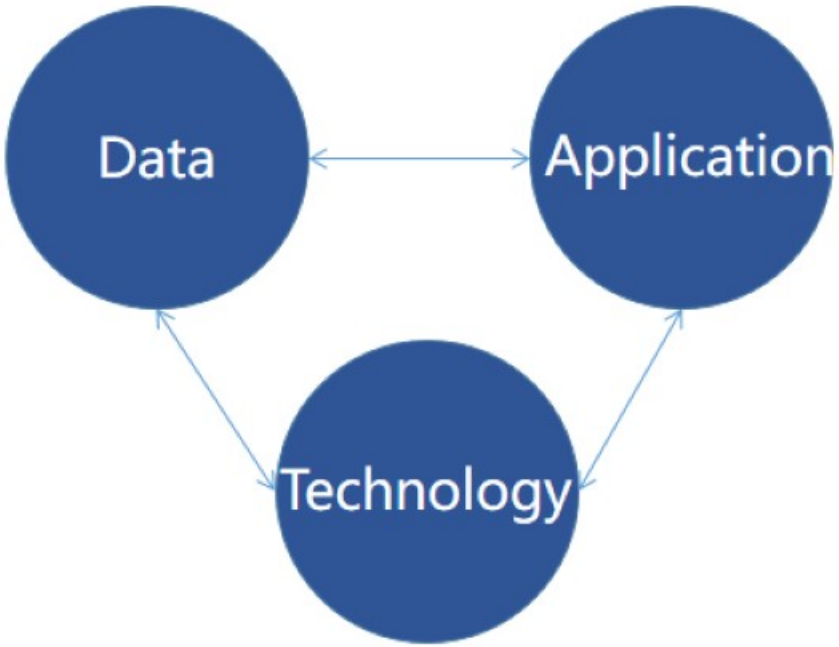
Relevant technology



Computation model	Data quality
Network security	Work flow
Software architecture	Web Service
Information security	Law service
Ubiquitous network	Data security
Software engineering	……

What we are doing…

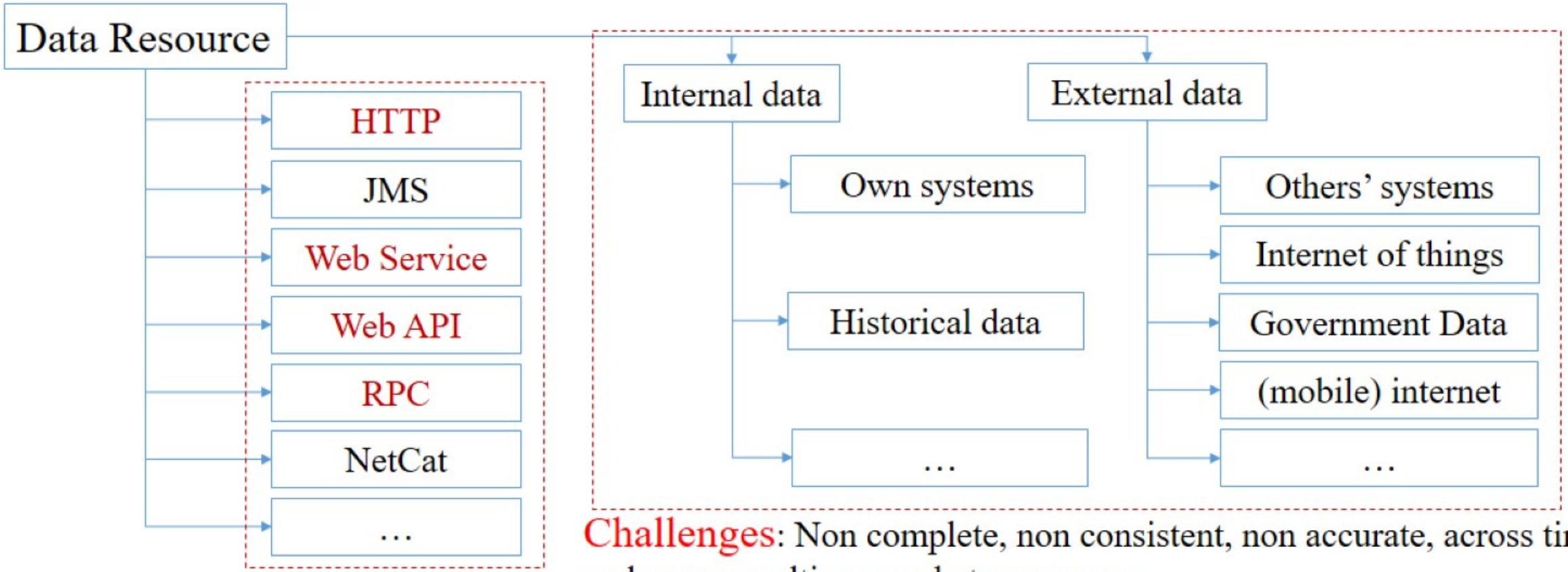
- Common issues
  - For a specific application
    - –How to do and what data needed and where is the data ?
  - Using the data
    - –What to do and how to do?
  - With some technology
    - –What data needed and what to do ?
  - Others
    - –Process ?
    - –Framework ?



Gathering, Enriching and Connecting the relevant data for potential demand;
Researching and developing analysis tool kit for the potential demand;
Providing solution for potential demand, usually, data service or computaiton service is given;

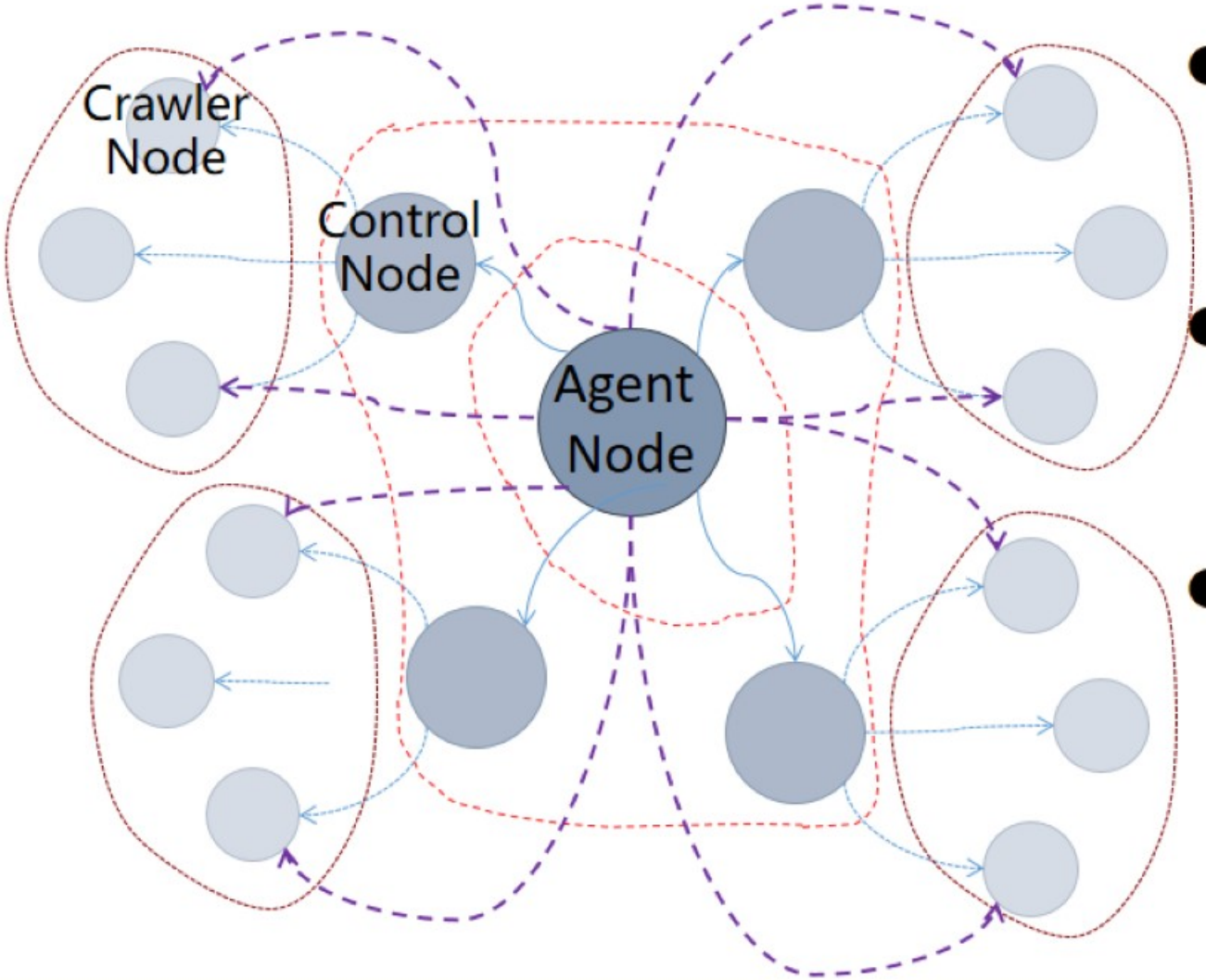
Data Gathering

Objective: Research on data gathering or enriching method and constructing data center or platform



Challenges: Non complete, non consistent, non accurate, across time and space, multi-source heterogeneous

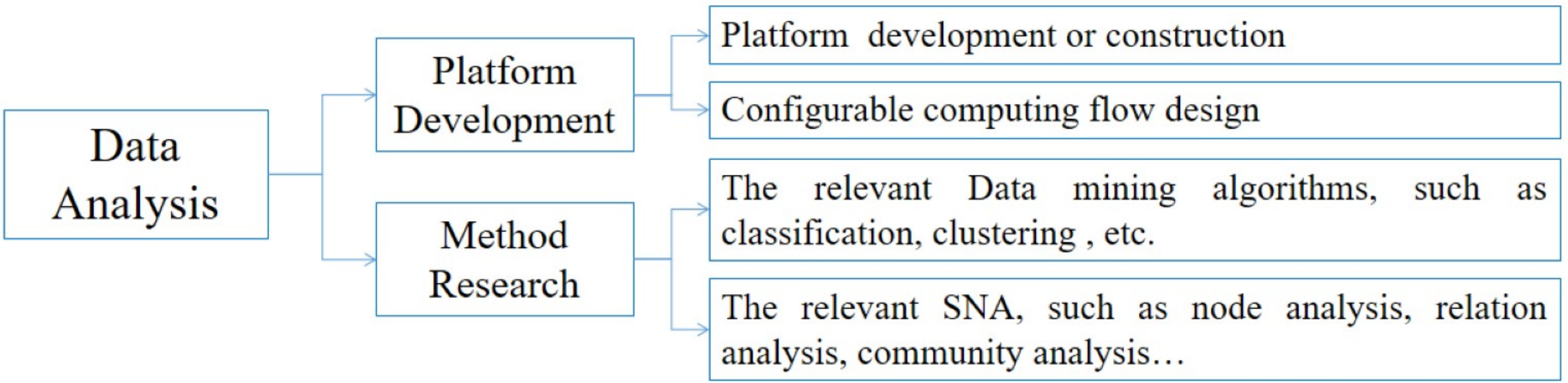
Web Crawler



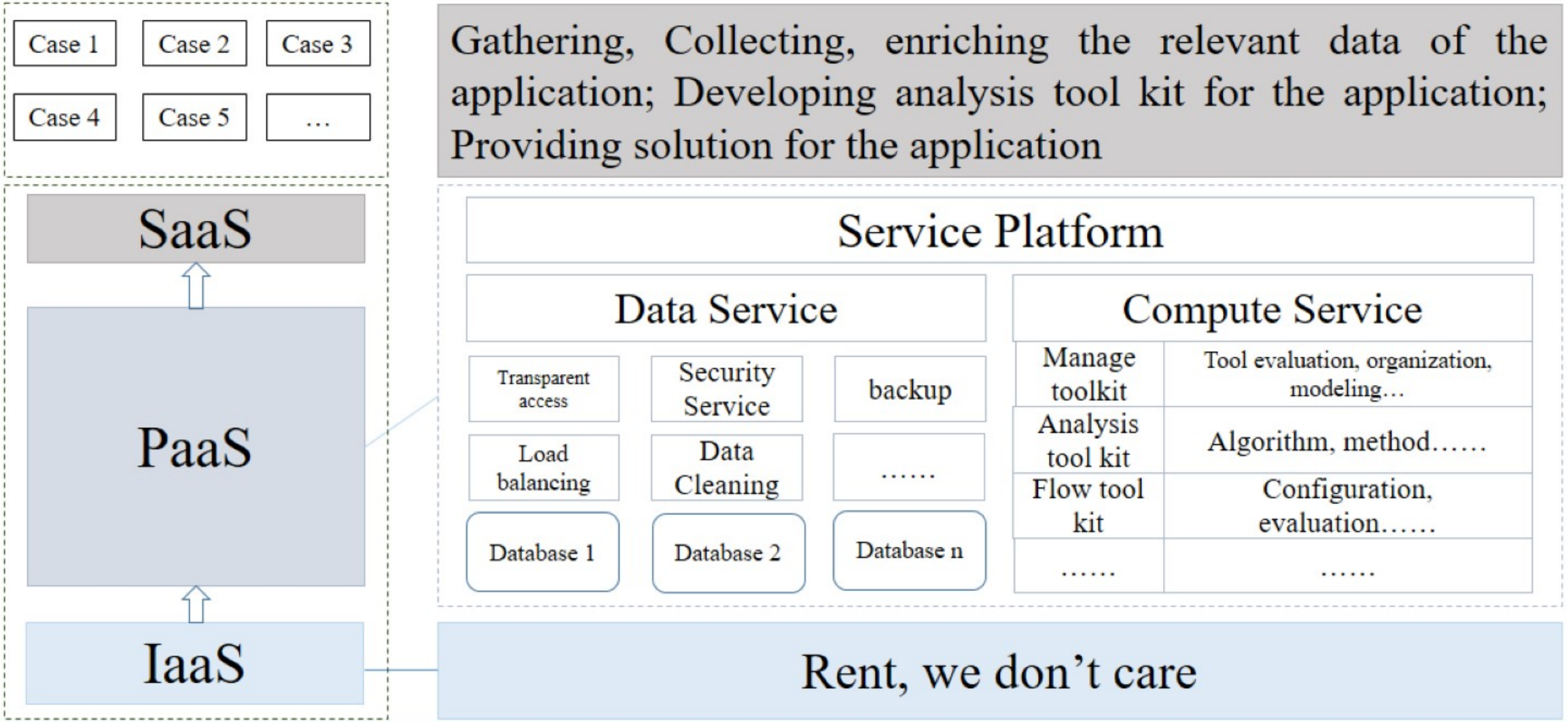
- Agent Node
  - Communication with control node
  - Communication with Crawler node
  - Assign the crawler node to control node
- Control Node
  - Configure the task
  - Communication with others control node
  - Manage or evaluate the linked crawler node
- Crawler Node
  - Communication with agent node to be assigned to control node
  - Communication with control node to get task and report the status
  - Report to the linked control node

Data Analysis

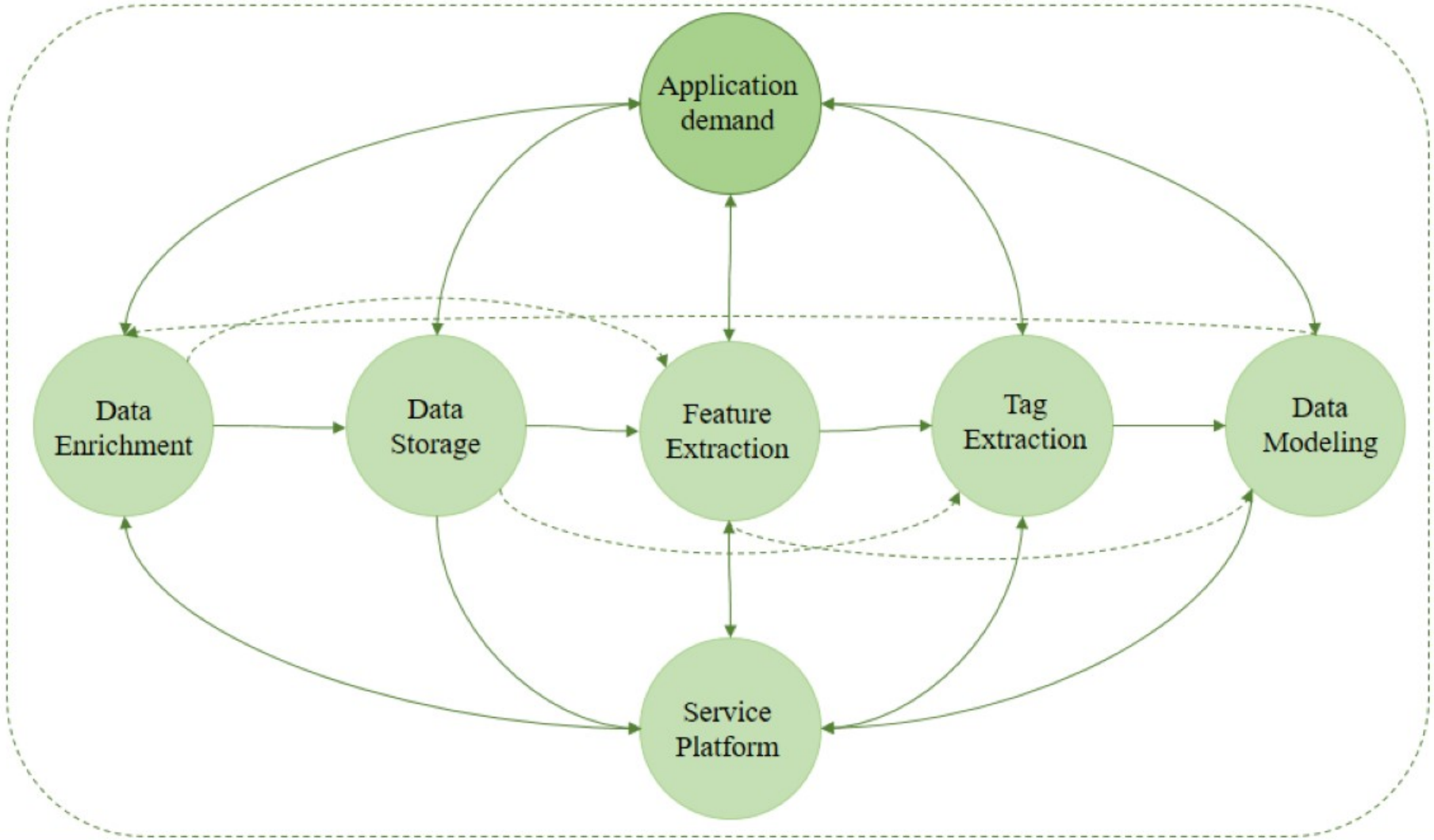
Objective: Research on data analysis method and constructing computation center or platform



System Development



A general Framework of Big Data Analysis



Application Cases

- Enterprise Credit Investigation Service
  - Loan Risk Evaluation
  - Public Opinion Data Service Platform
  - ……
- Individual Analysis Service
  - Crime Analysis
  - Telecom Data Analysis for Precision marketing
  - …..
- Intelligent Information Service
  - Research On Traditional Chinese Medicine
  - Intelligent service for Education
  - ……

Loan Risk Evaluation and Our More Work

Credit Investigation Service

Nowdays, what does the financial institutions expect to buy?

Infrastructure

- Credit Investigation
- Payment
- Accounts Settling
- Others

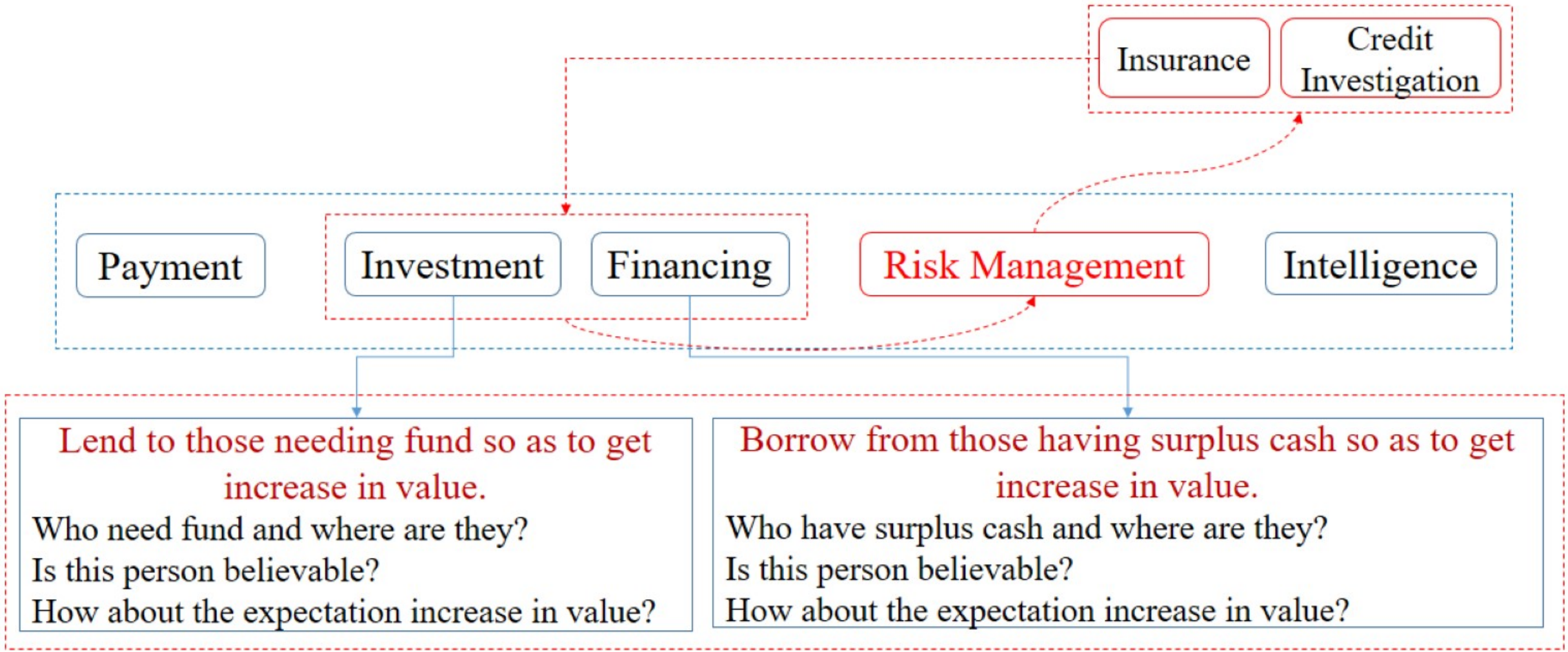
Produce

- Hardware
- Software
- Information
- Office supplies
- Others

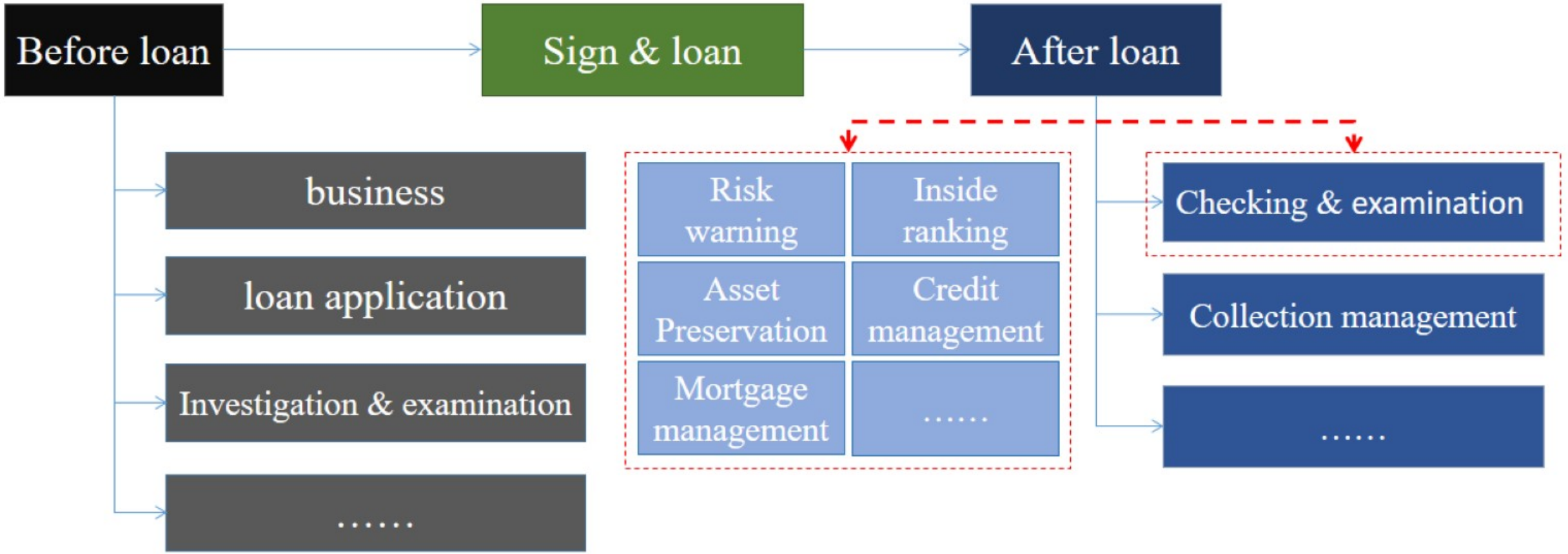
Consulting & training

- Management and business consulting
- IT consulting
- Management and business training
- Others

Risk Management



Pain spot of financial institutions

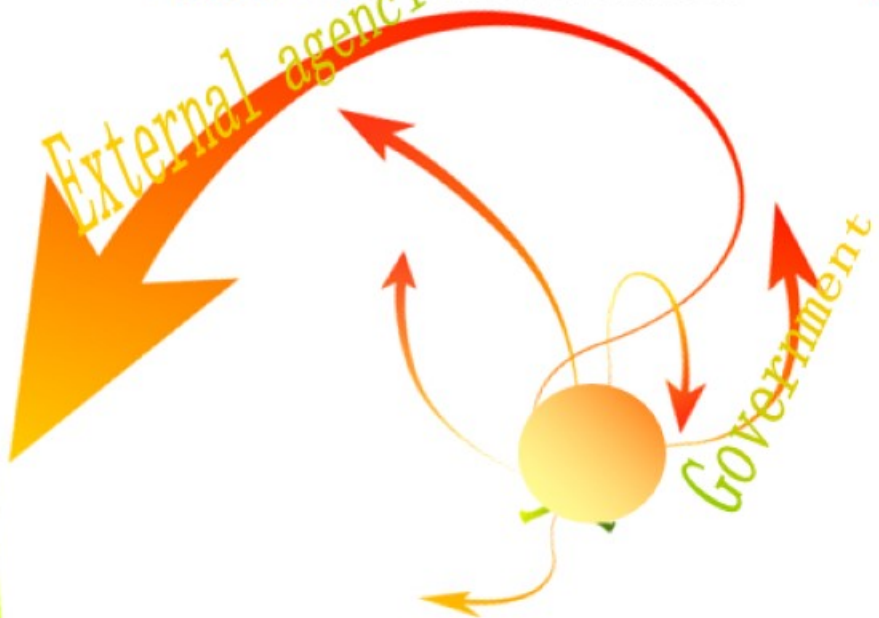


Loan Risk Evaluation

Internal Information

1) Business side
2) Antifraud
3) Credit Approval
4) Credit Service or Asset evaluation
5) Asset portfolio
6) etc

External Information



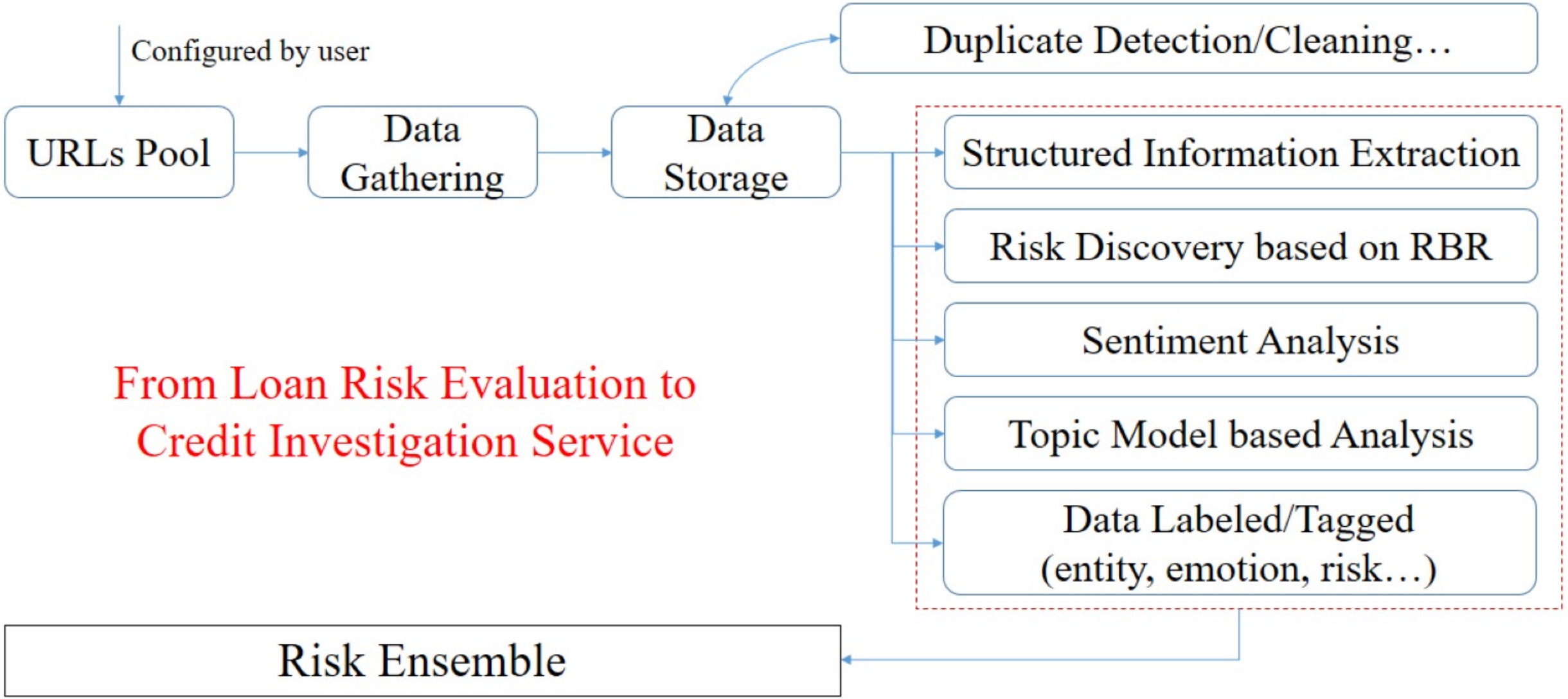
1) From External agencies
   (1)Credit Information
   (2)Debt Information
   (3)Legal Action Information
   (4)etc
2) From Others Bank
3) From Internet (Negative information)
4) From Government
5) From Others.

Gathering External data from internet, which is relevant to the loan risk evaluation;

Developing the relevant tool kit for risk evaluation;

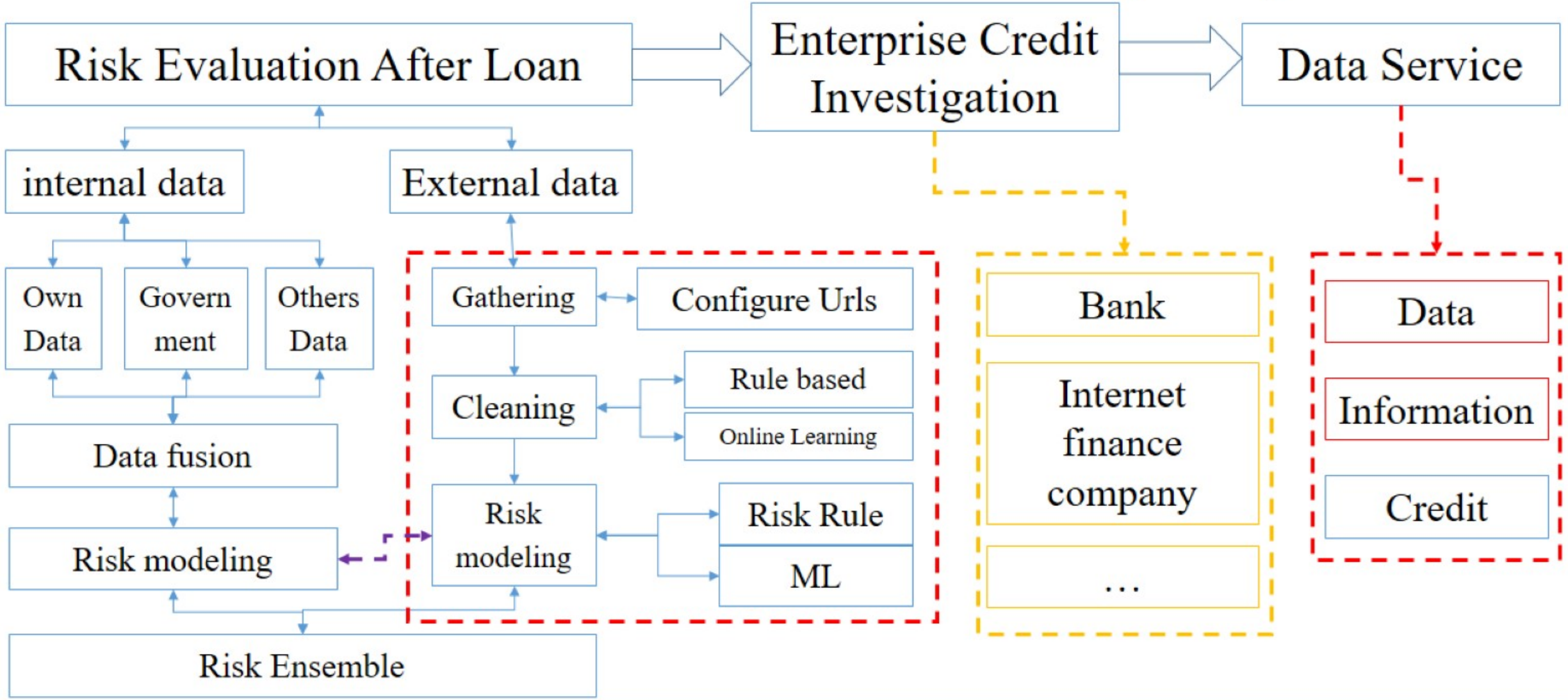
Providing data service for customer;

Analysis Example

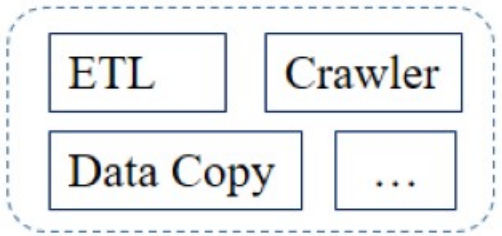


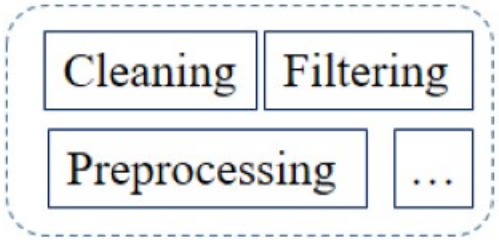
From Loan Risk Evaluation to Credit Investigation Service

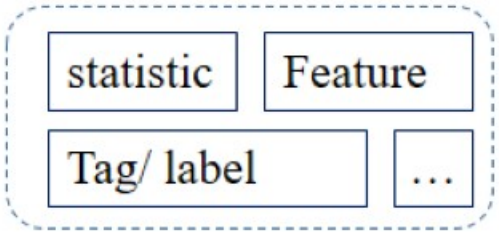
Some of Our Work(1/2)

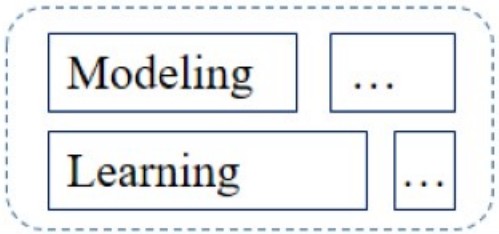


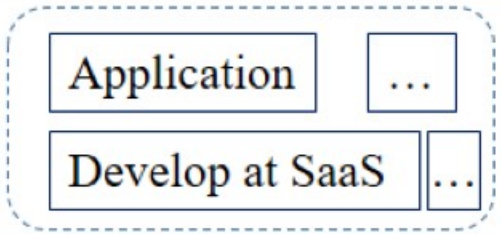
Some of Our Work(2/2)

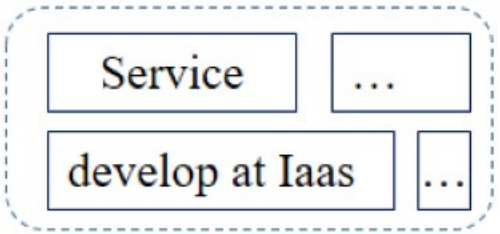












The Gathered data can be sold as well without any preprocessing..

Preprocessing the Gathered data is cleaned, filtered and preprocessed, and these preprocessed data can be sold also.

Extracting Information from the data, and the information can be sold so as to help buyer understand the data/object clearly.

Data Trading Platform
Credit Investigation
Trading Platform

Summary(1/4)

1. "data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges.", *Oxford English Dictionary* (OED).
2. "an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using on-hand data management tools or traditional data processing applications.", *Wikipedia, 2014*
3. "datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze," . *McKinsey*
4. "The broad range of new and massive data types that have appeared over the last decade or so.", *Davenport*
5. The belief that the more data you have the more insights and answers will rise automatically from the pool of ones and zeros.
6. A new attitude by businesses, non-profits, government agencies, and individuals that combining data from multiple sources could lead to better decisions.
7. "The ability of society to harness information in novel ways to produce useful insights or goods and services of significant value" and "…things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value.", *Viktor Mayer-Schönberger and Kenneth Cukier*
8. The new tools helping us find relevant data and analyze its implications.
9. The convergence of enterprise and consumer IT.
10. The shift (for enterprises) from processing internal data to mining external data.
11. The shift (for individuals) from consuming data to creating data.
12. The merger of Madame Olympe Maxime and Lieutenant Commander Data.

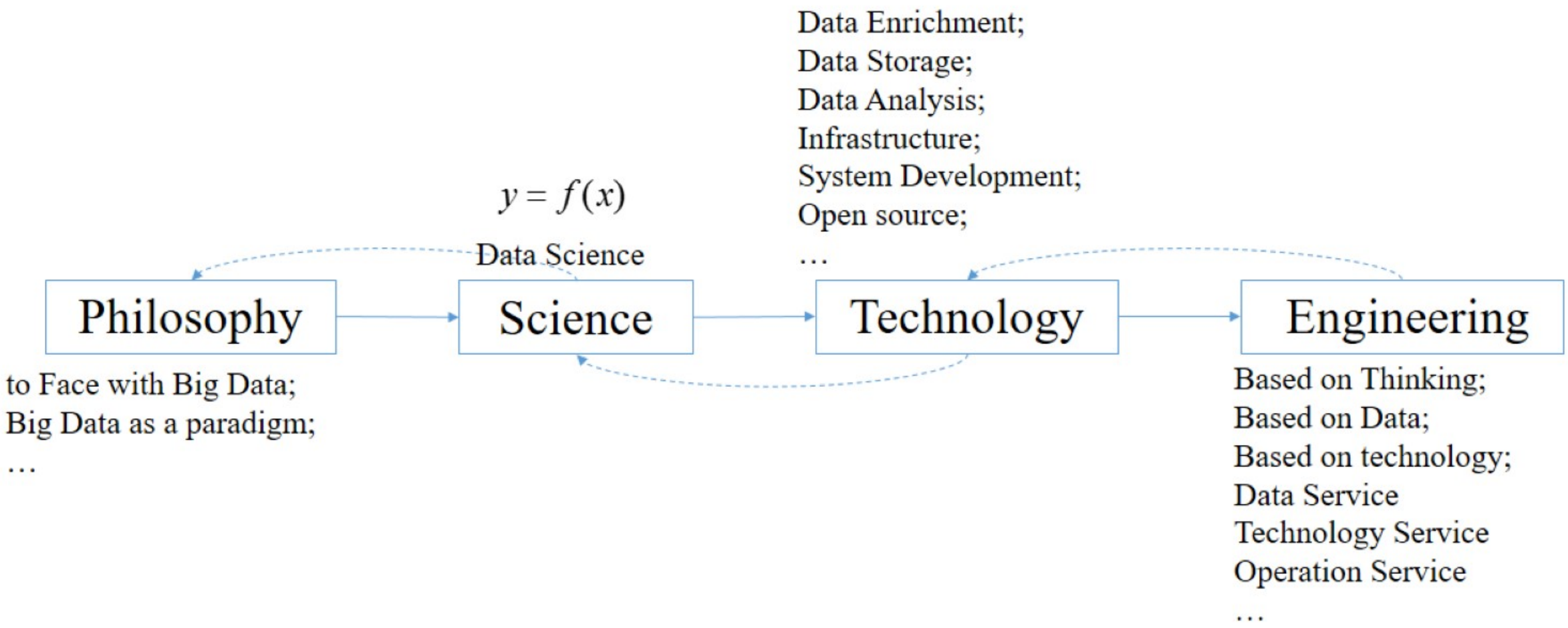
Def. Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.

Summary(2/4)

- **WHAT** the Big Data is
  - No matter what to say, the size is usually large and massive.
  - Shown as 1-4.
- **Whether** to bring value
  - An expectation that data can bring/create new value
  - Shown as 5-6.
- **HOW** to use the Big Data
  - A tool kit for gathering, connecting and analyzing data
  - Shown as 7-12.

Everyone may/should give his own definition for Big Data.

Summary(3/4)


Data Enrichment;
Data Storage;
Data Analysis;
Infrastructure;
System Development;
Open source;
…
$y = f(x)$
Data Science
Philosophy
Science
Technology
Engineering
to Face with Big Data;
Big Data as a paradigm;
…
Based on Thinking;
Based on Data;
Based on technology;
Data Service
Technology Service
Operation Service
…

Summary(4/4)

- This talk introduces some of our work on Big Data and Intelligent Systems.
- More information, please contact
  - Contact : Chongjun WANG
  - Email: chjwang@nju.edu.cn
  - Tel: 13913h922928


Thanks for Your Attention
and
Welcome to Nanjing/Nanjing University