Data Polishing

A New Approach to Data Abstraction by Clarifying Meaningful Hidden Structures in Data

Takeaki Uno

(National Institute of Informatics, Japan)

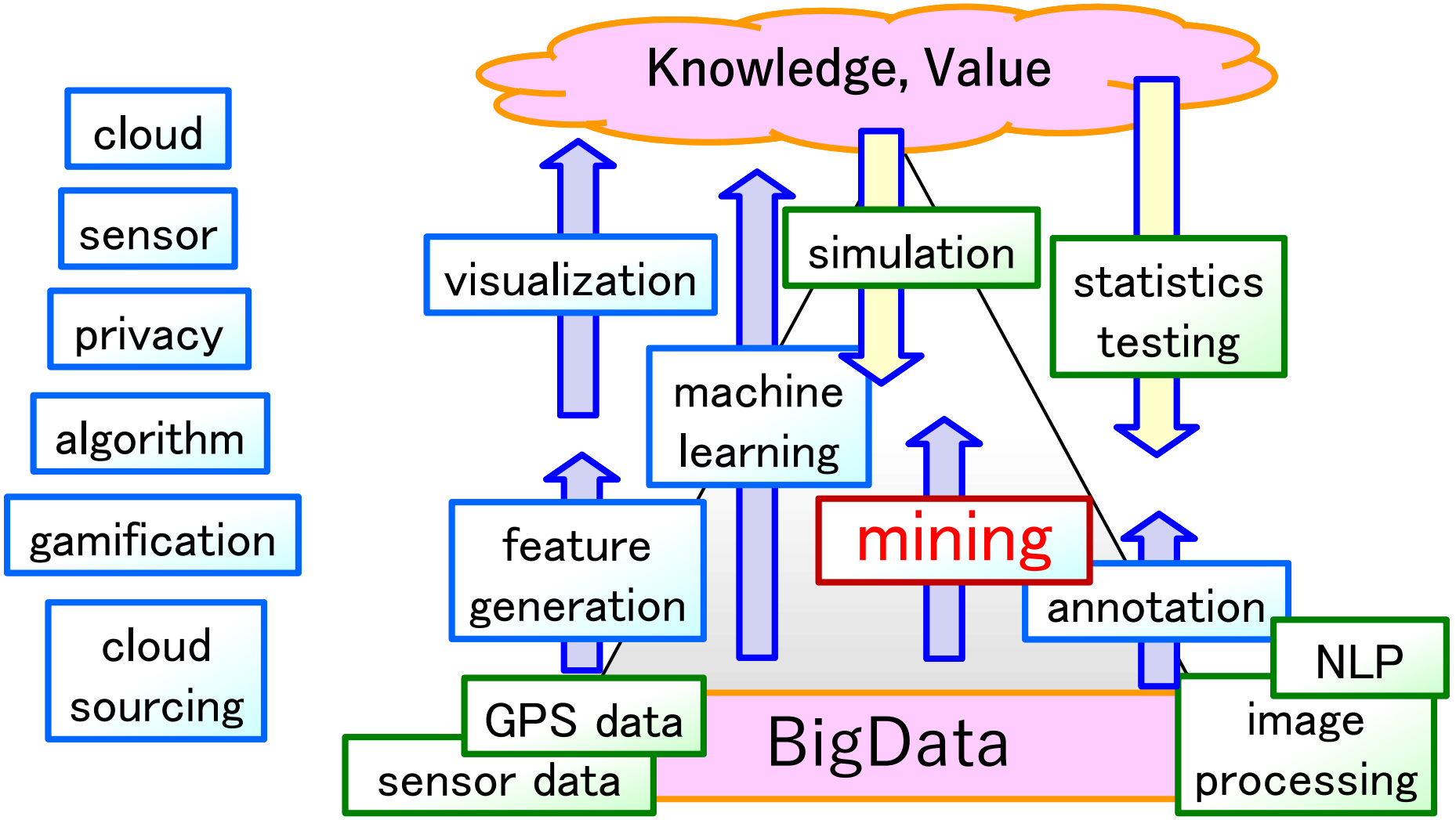
http://research.nii.ac.jp/~uno/index-j.html
e-mail: uno@nii.ac.jp

12/July/2018 DATAIA-JST

International Symposium on Data Science and AI

Mining & other Methodologies on BigData

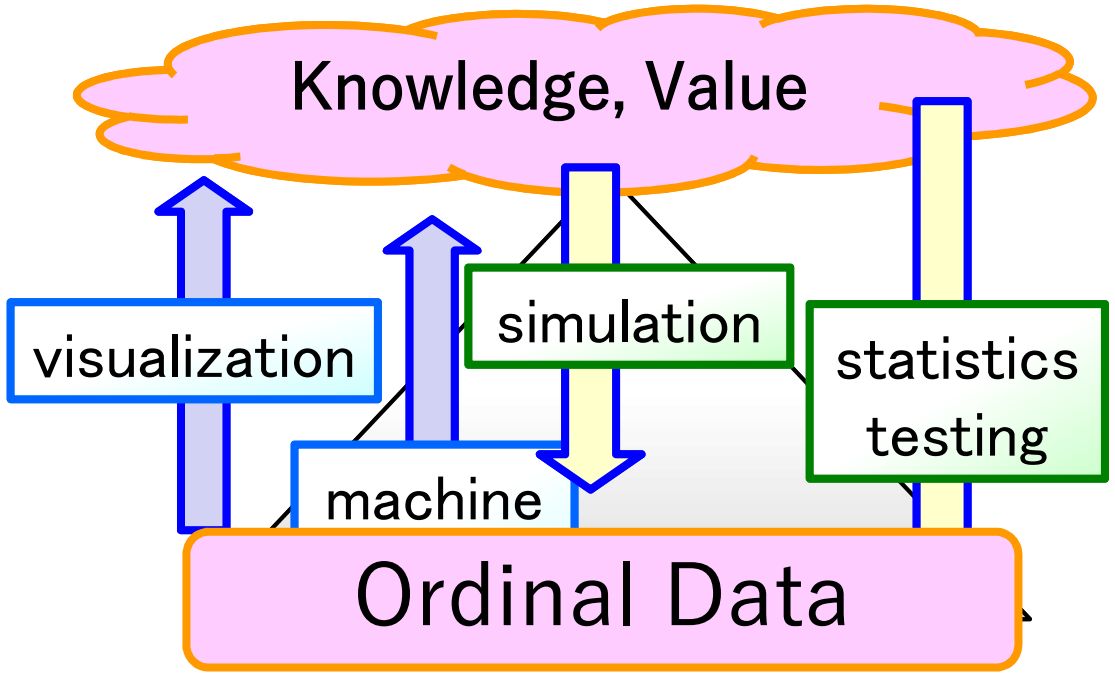
Mining gives materials to upper layers, from big, shallow meaning, sparse, noisy, and ill-granularity data; an approach is **abstraction**



Traditional Situations

Data is close to the knowledge; the entities of the data have

- cloud
- sensor
- privacy
- algorithm
- gamification
- cloud sourcing

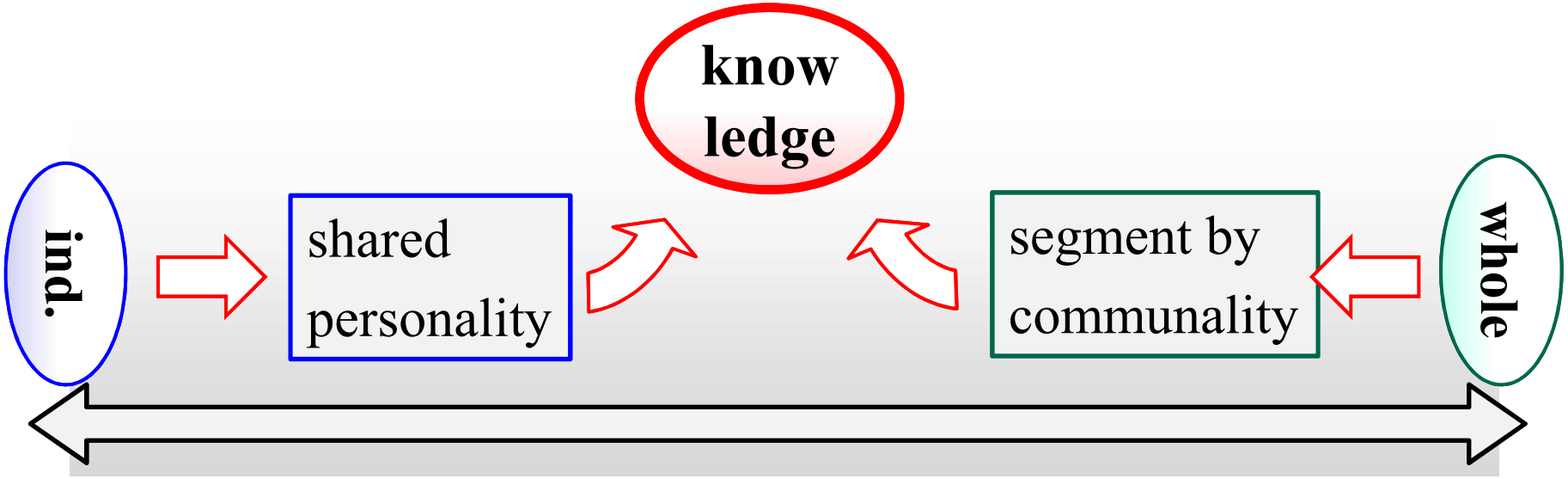


Analysis along Human Cognition

- Two human approaches for understanding the data

  + start from the overview, segments the data in some ways, and understand the comparison or combinations of them

  + deeply observe several individuals, and extract abstracted strong personality which leads some feasible hypothesis



Abstraction and Specialization

- Segmentation is specification, and grouping is abstraction

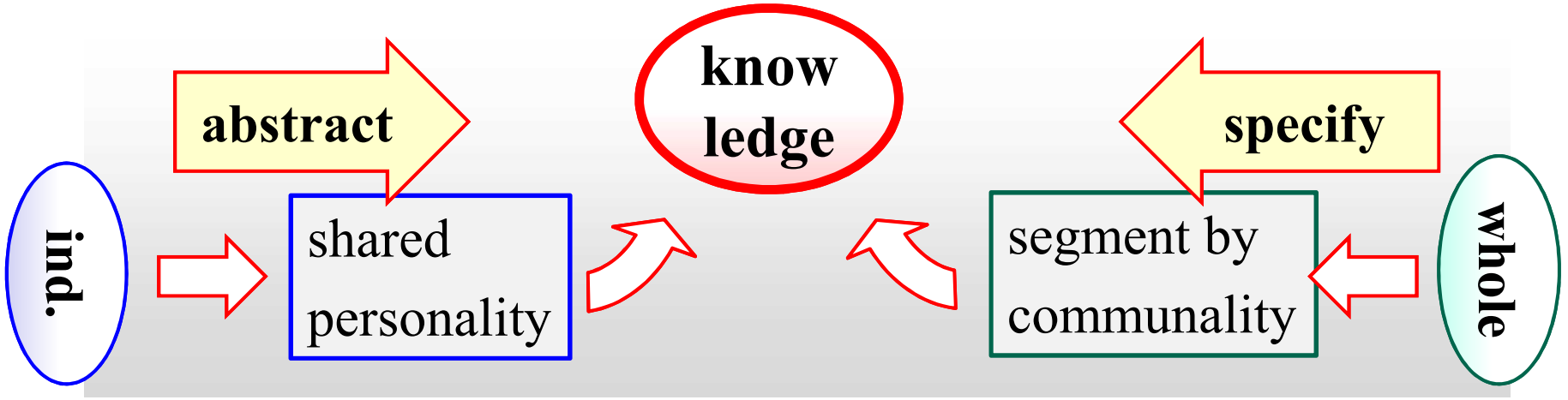
  Many analysis including classification directs specification

  directed along some hypothesis which users already have

- Abstraction is basically mining type, such as pattern mining, community mining, etc.

  But such mining usually aims to completely find all possibility

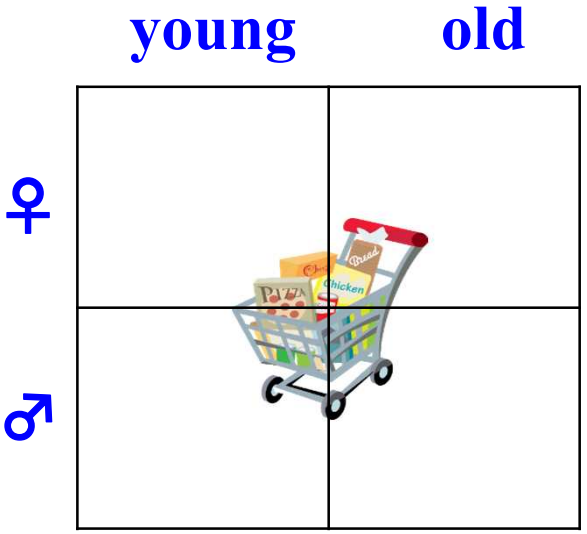
  Not so many methodologies are proposed nor used



Basket Data in Supermarket

- Segment the customers in a supermarket

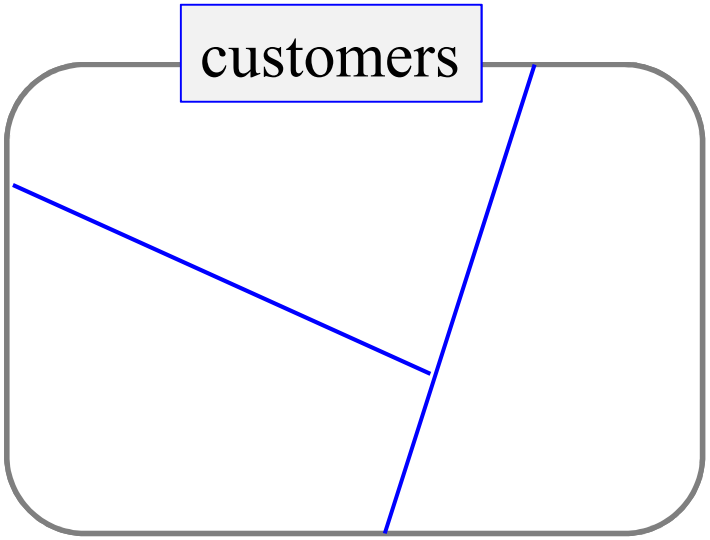
Features of the groups, segmented by
gender, age, etc… **We know already!**



We want to know their lifestyles
Baskets may represent lifestyles, so look at clusters given by the similarity of one's baskets

Clustering algorithms gives clusters

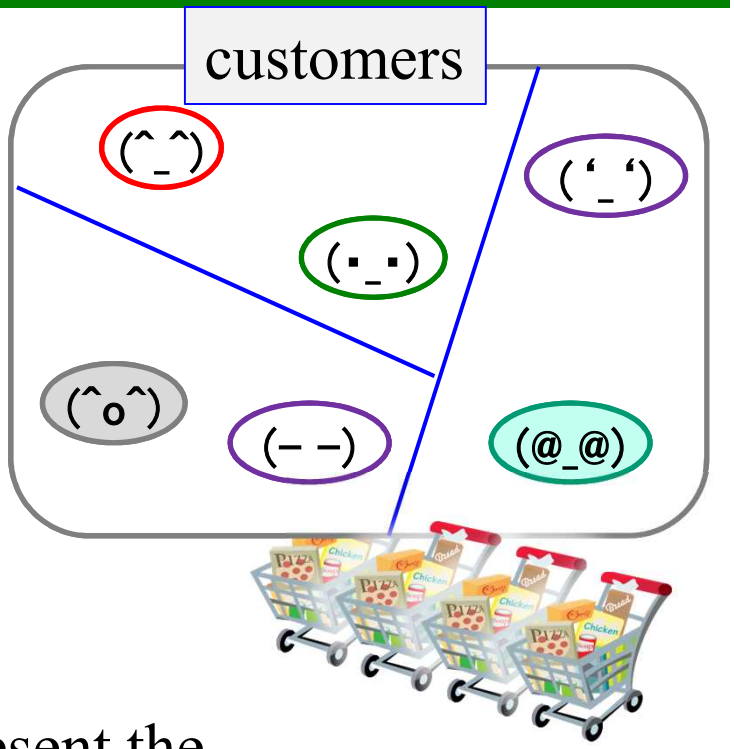
… but the features and meanings of the groups are hardly seen (what's this?)
they includes so many customers



Look at Particles

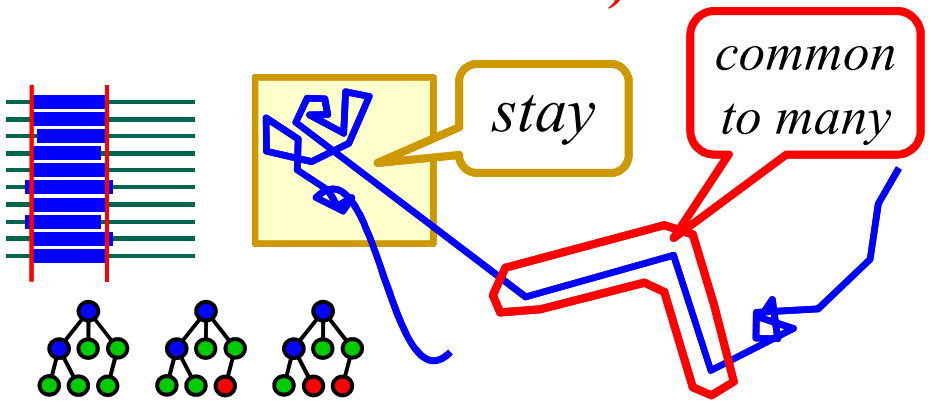
Clusters includes **small groups of similar customers** sharing common properties

Finding groups composed of similar customers will help understanding
We call such groups **“particle”** of data
（not finding boundary, but **center of dense**）



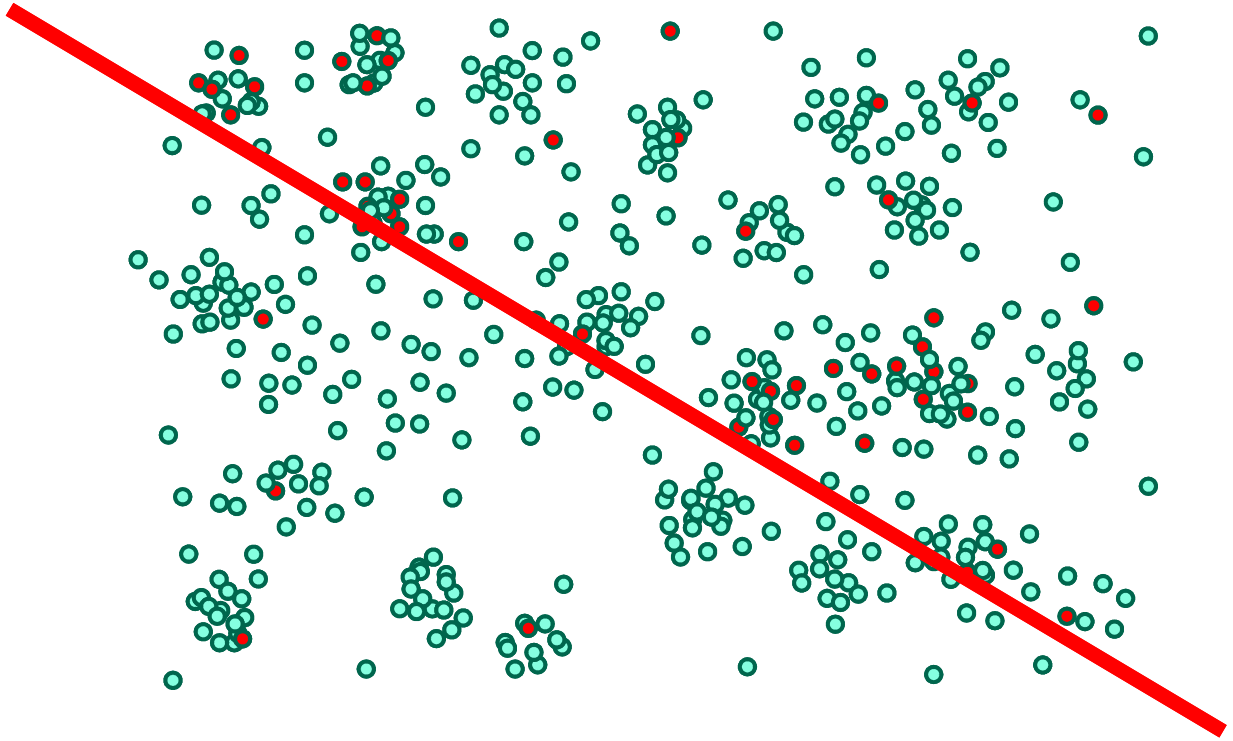
In general, particles means objects that represent the **minimal units of local meanings, (abstracted individuals)**

burst, pattern, boundary, flow,
common sequence, group,
local center, repetition, …



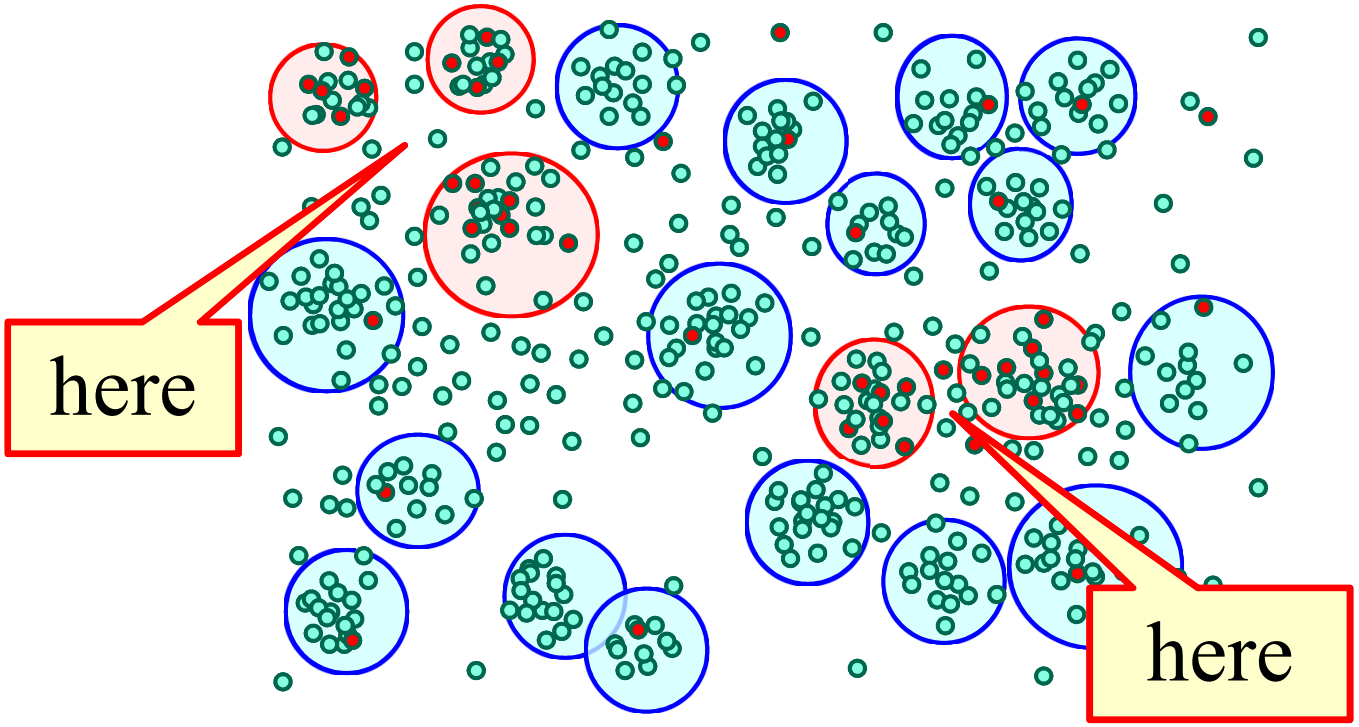
Machine Learning without Particles

- Partition the data into two areas, including more reds, and not
- Even though attains high accuracy, the solution is "**hard to understand**" the mechanism



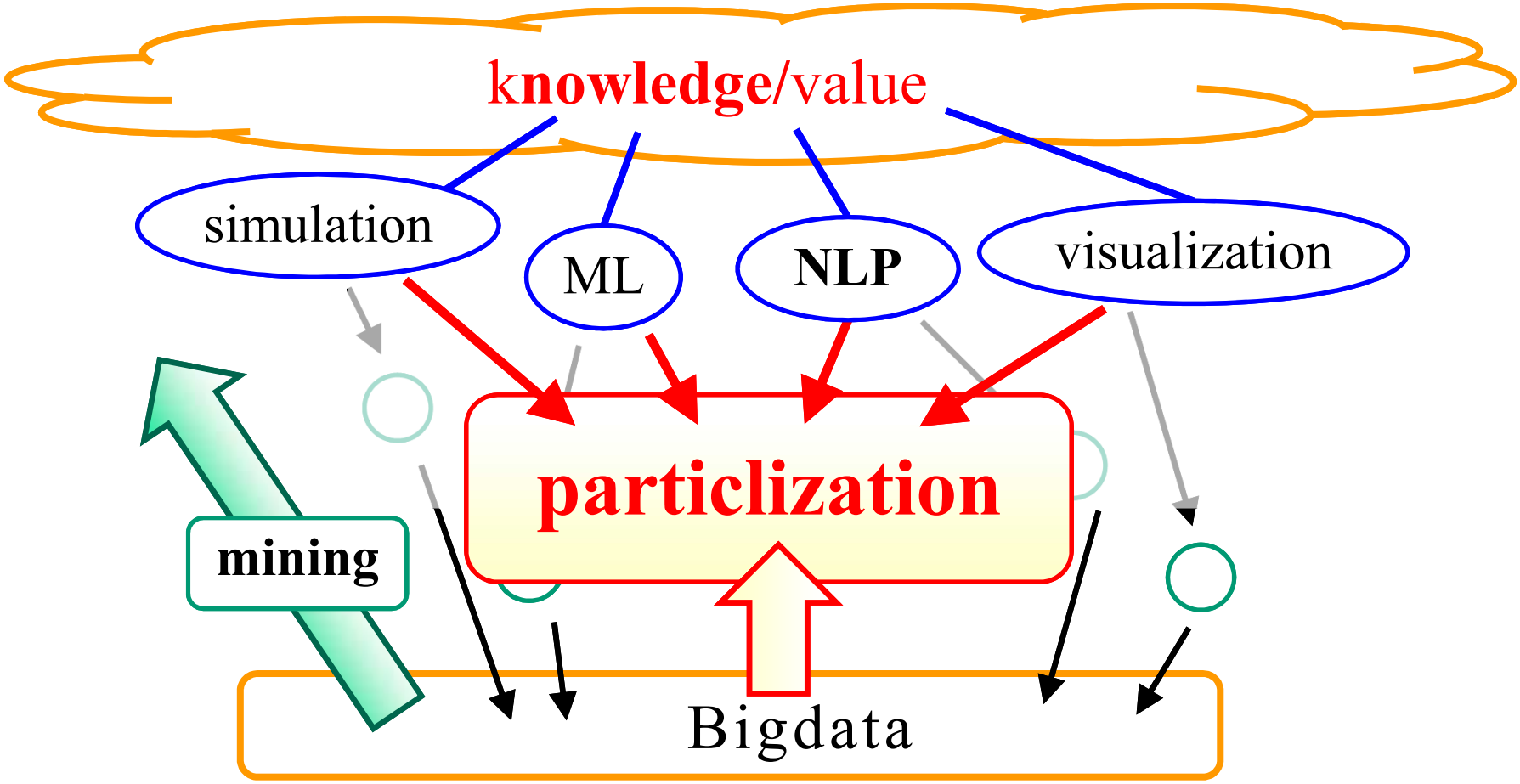
Particles make it Understandable

Easier to get some meanings, or inspires



The Data Particlization Approach

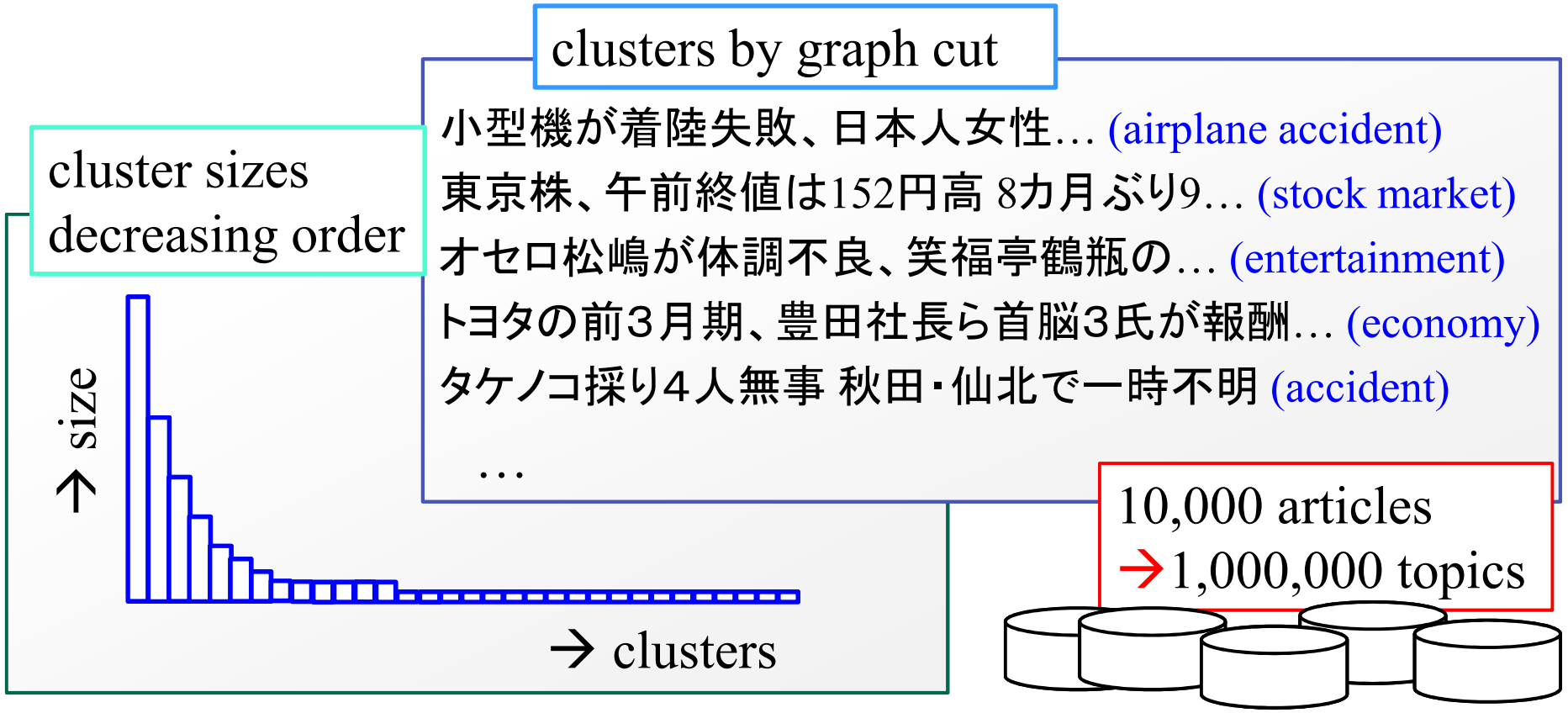
- Existing methods design “particle-like structures” independently
- Mining is not directed to good utilities of the methods
- Data particlization serves as the basis for the data analysis tasks



Micro-Clustering is actually Hard

Clustering finds (global) **"classes"**, but has many problems

huge small solutions, unbalanced sizes, skewed granularity



Experiment on Reuter Articles

- The sizes follow Zipf's distribution.

AMSTERDAM 1997-01-16 NETHERLANDS: AOT PROVISIONAL 1996 NET ALMOST DOUBLES.

MANILA 1997-01-16 PHILIPPINES: Planters sets 275 to 300 pct stock dividend.

SHANGHAI 1997-01-16 CHINA: Shanghai's tax income up 26.9 pct in 1996.

TOKYO 1997-01-16 JAPAN: Japan bank lending fell 0.4 pct in December -- BOJ.

CONCORD, Mass. 1996-10-16 USA: GenRad Inc Q3 diluted shr rises to $0.21.

BANGKOK 1997-01-16 THAILAND: Thai 1996 sugar exports 4.34 mln tonnes - surveyors.

HOFFMAN ESTATES, Ill. 1996-10-16 USA: Sears, Roebuck and Co Q3 net higher.

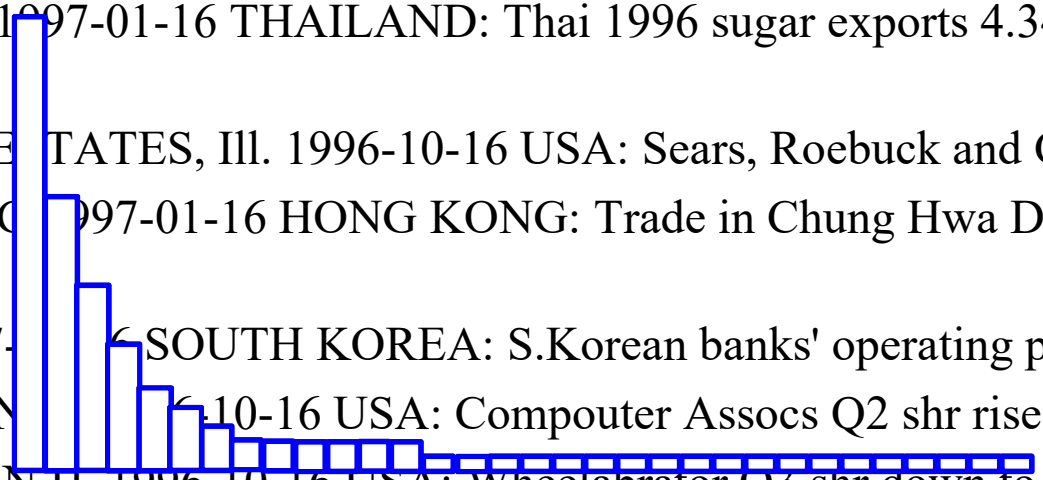
HONG KONG 1997-01-16 HONG KONG: Trade in Chung Hwa Dev shares suspended.

SEOUL 1997-01-16 SOUTH KOREA: S.Korean banks' operating profits rise 16.2 pct.

ISLANDIA, N.Y. 1996-10-16 USA: Compouter Assocs Q2 shr rises to $0.59.

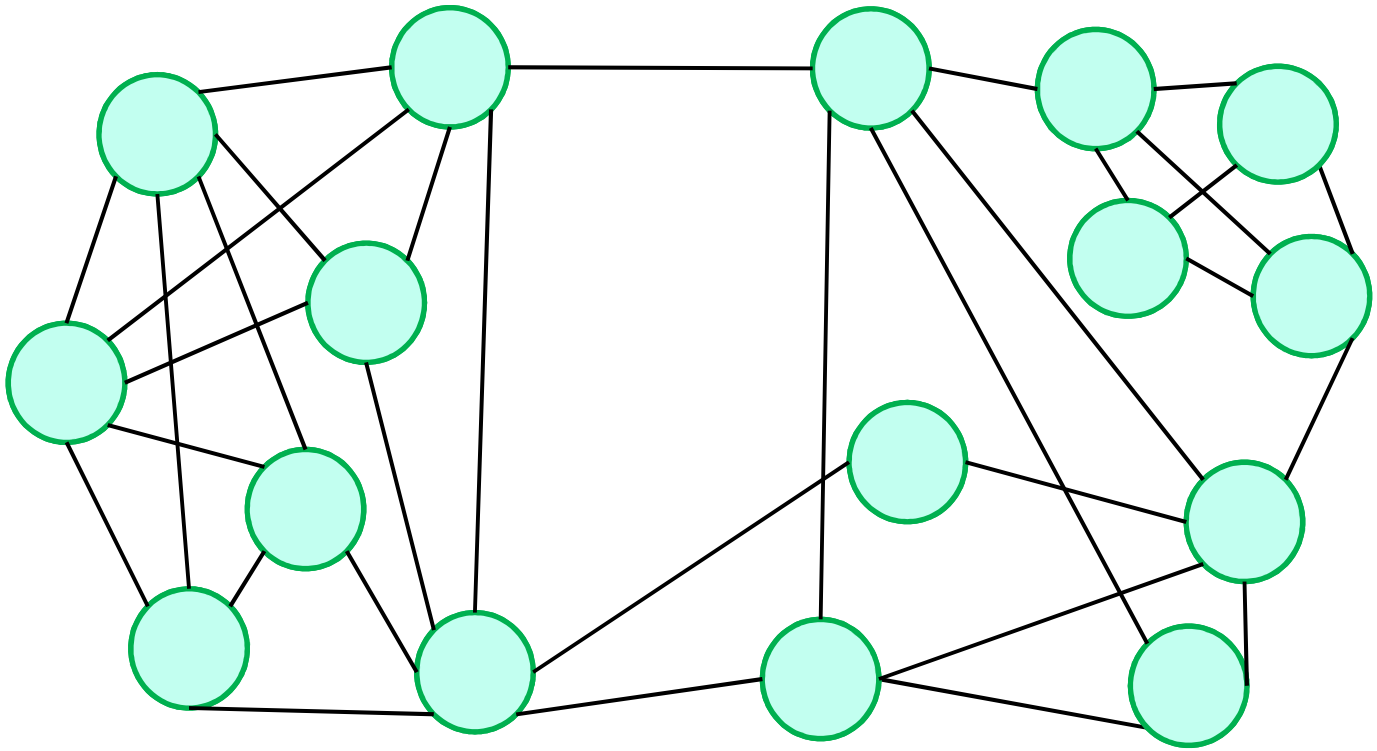
HAMPTON, N.H. 1996-10-16 USA: Wheelabrator Q3 shr down to $0.28.

BASLE, Switzerland 1997-01-16 SWITZERLAND: Roche sees higher 1996 profit.



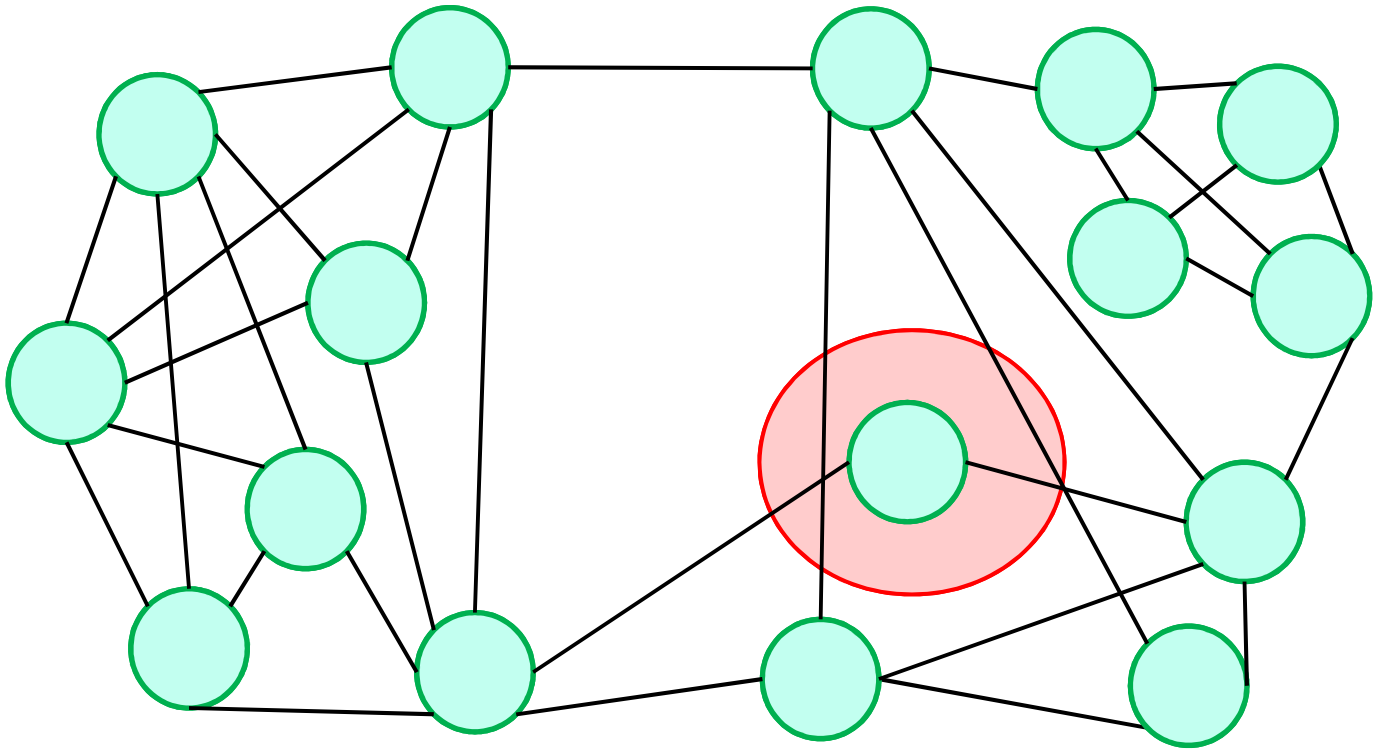
Graph Cut

Partition the graph into two parts



Graph Cut

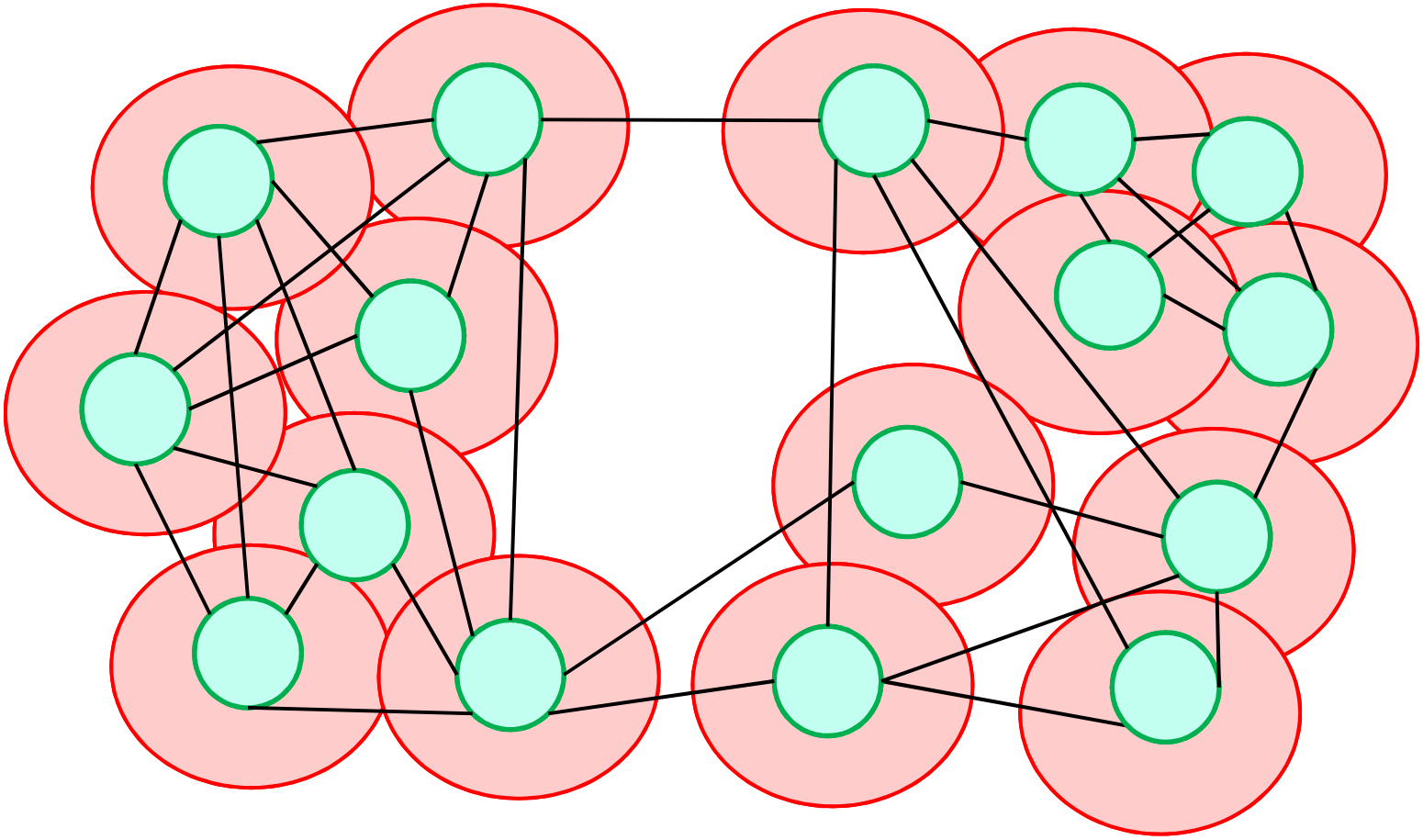
Partition the graph into two parts



Ex. Modularity Maximization

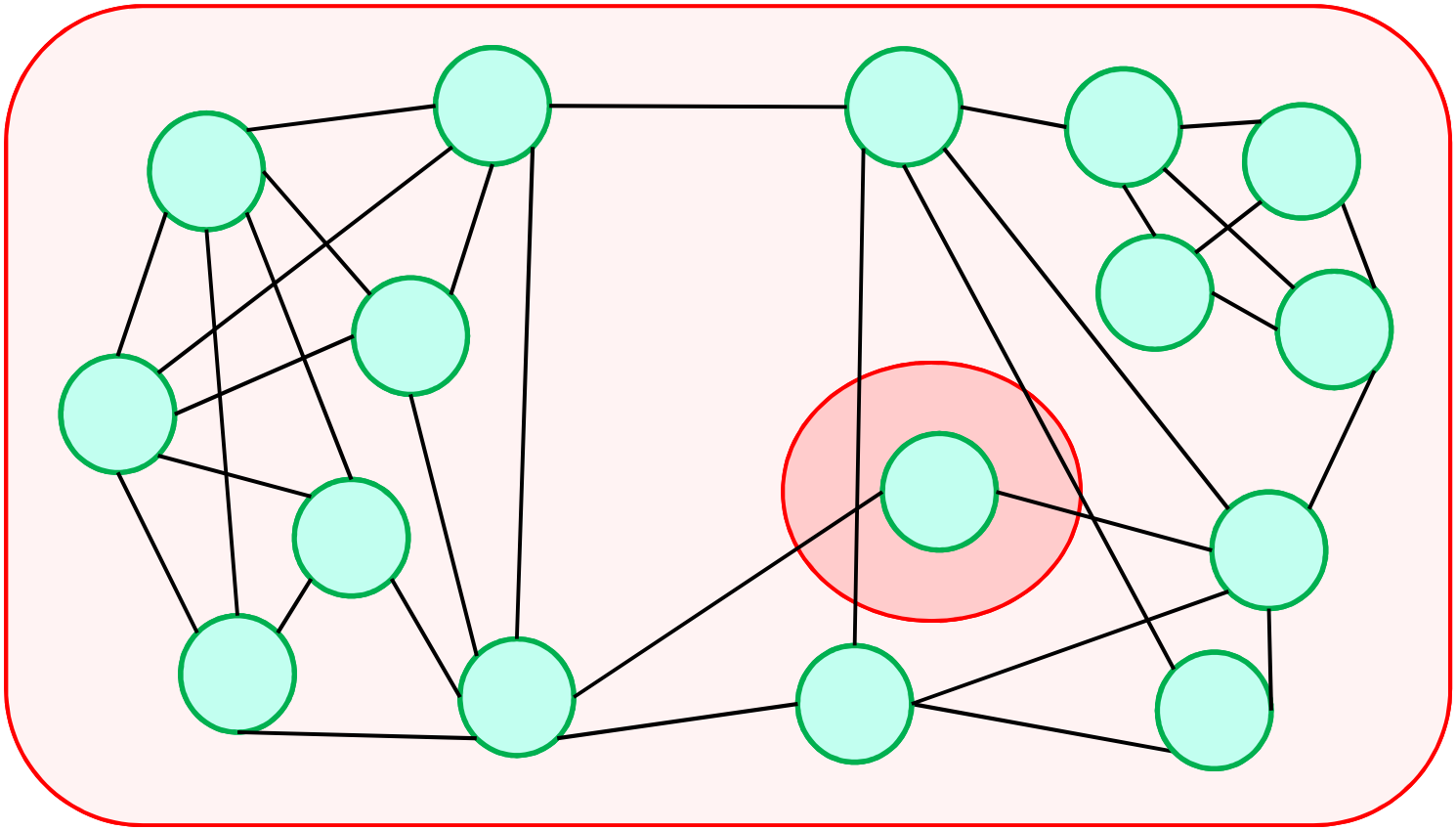
Partition the graph while maximizing modularity

Modularity = inside density / boundary density



Ex. Modularity Maximization

Give incentives to increase the size



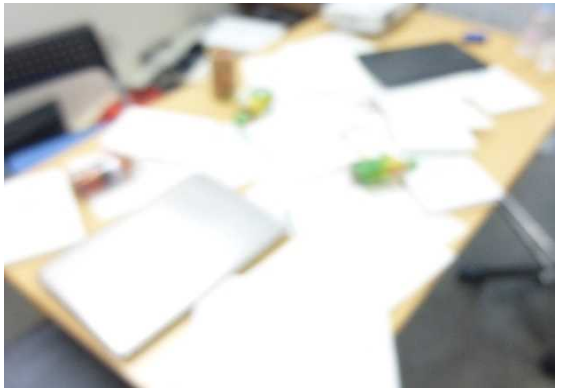
Basic Idea : Clarify Structures

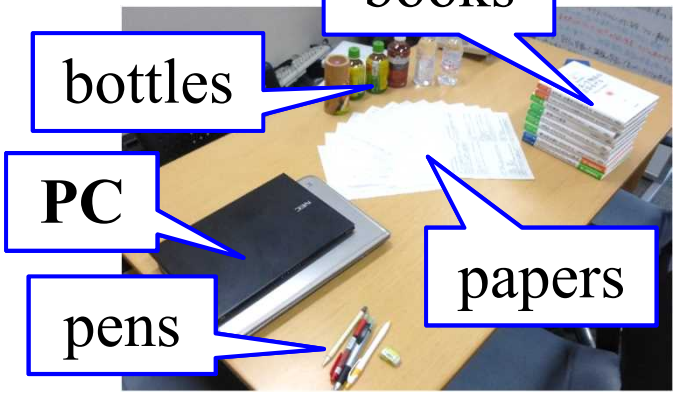
Why bad? ... because, the boundaries of the structures are not clear

The analogy: making the picture visually clearer
sharpening edges, erasing noise, removing shadows, … and **rearranging** objects

At the same time, the accuracy in recognizing, classifying, and segmenting of the objects in the picture can be increased

Do the same in Bigdata!



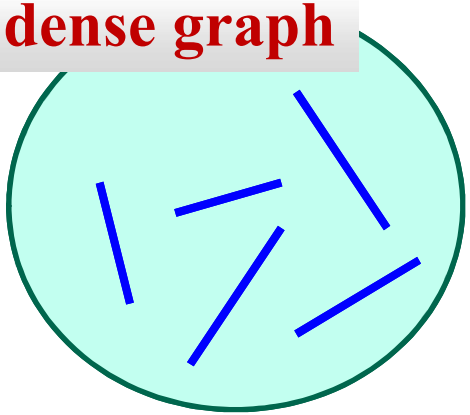


New Approach: Data Polishing

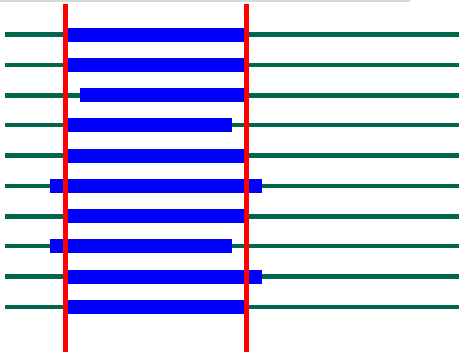
DataPolishing

Remove ambiguity by local change from **Feasible Hypothesis**

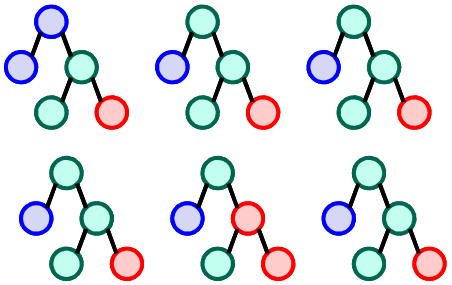
dense graph



sequence/string



pattern



- Modify parts that should be so, thus no loss of solution
- Clear the ambiguities, unify similar solutions, decrease #solutions
- No loss of comprehension

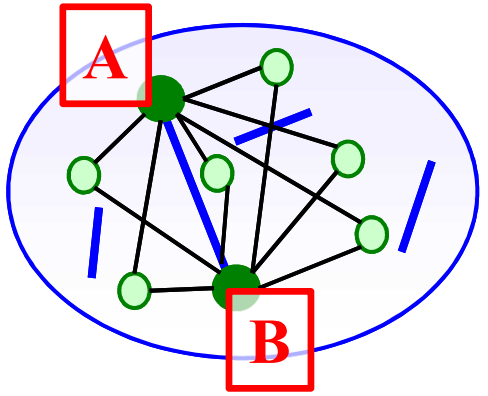
Data Polishing for Graph Clustering

Use a necessary condition instead of hard dense graph enumeration

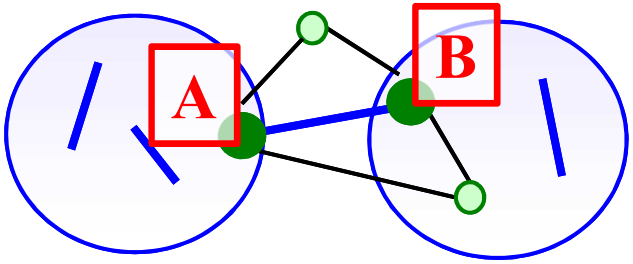
A and **B** belong to a group ⇔ **A** and **B** share many neighbors
(neighbors are similar)

+ if **#common neighbors** of **A** and **B**,
  - ≥ **k** → make edge **(A,B)**
  - < **k** → remove edge **(A,B)**

(neighbor similarity ≥ **θ**, to uniform granularity)



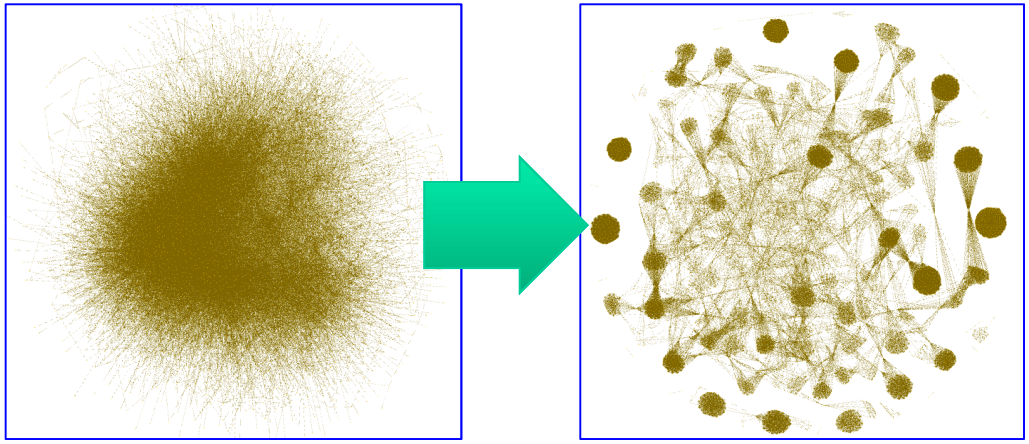
- Apply this to all vertex pairs at once
- Repeat this until convergence



(the density of a dense subgraph is always increased)

Preliminary Study for Graph Clustering

the scale	original	polished
#nodes	3,282	3,282
#edges	35,168	73,132
density	3.3‰	6.8‰
#cliques	32,953	343



Companies and their business relations

Prediction accuracy:

accuracy on customer attribute prediction by clustering methods

	clique	Newman	graph cut
original	60.60%	59.70%	60.03%
particle	**71.36%**	**62.76%**	**67.78%**

Noise robustness:

discovery rates of clusters (particles) by clustering methods

	polishing	Newman	graph cut
noise 10%	**100.00%**	68.74%	76.10%
noise 40%	**99.69%**	7.91%	77.03%

Result: Data Polishing

- Number, size and distribution are appropriate

  the contents are also so

PARIS 1996-12-02 FRANCE: PRESS DIGEST - Algeria - Dec 2.
PARIS 1996-12-01 FRANCE: PRESS DIGEST - Algeria - Dec 1.
PARIS 1996-11-27 FRANCE: PRESS DIGEST - Algeria - Nov 27.
PARIS 1996-11-24 FRANCE: PRESS DIGEST - Algeria - Nov 24.
…

PARIS 1996-10-02 FRANCE: Single currency to be engine for Europe - Juppe.
BUENOS AIRES 1997-08-06 ARGENTINA: TABLE - ARGENTINA POSTS $129 MLN TRADE GAP IN JUNE.
BRASILIA 1997-04-01 BRAZIL: Brazil, Argentina resume talks over import rule.
BUENOS AIRES 1996-12-20 ARGENTINA: Fiat 'pens $600 mln plant in Argentina.
BRASILIA 1996-10-14 BRAZIL: Mercosur union to hold contest for logo - diplomat.
…

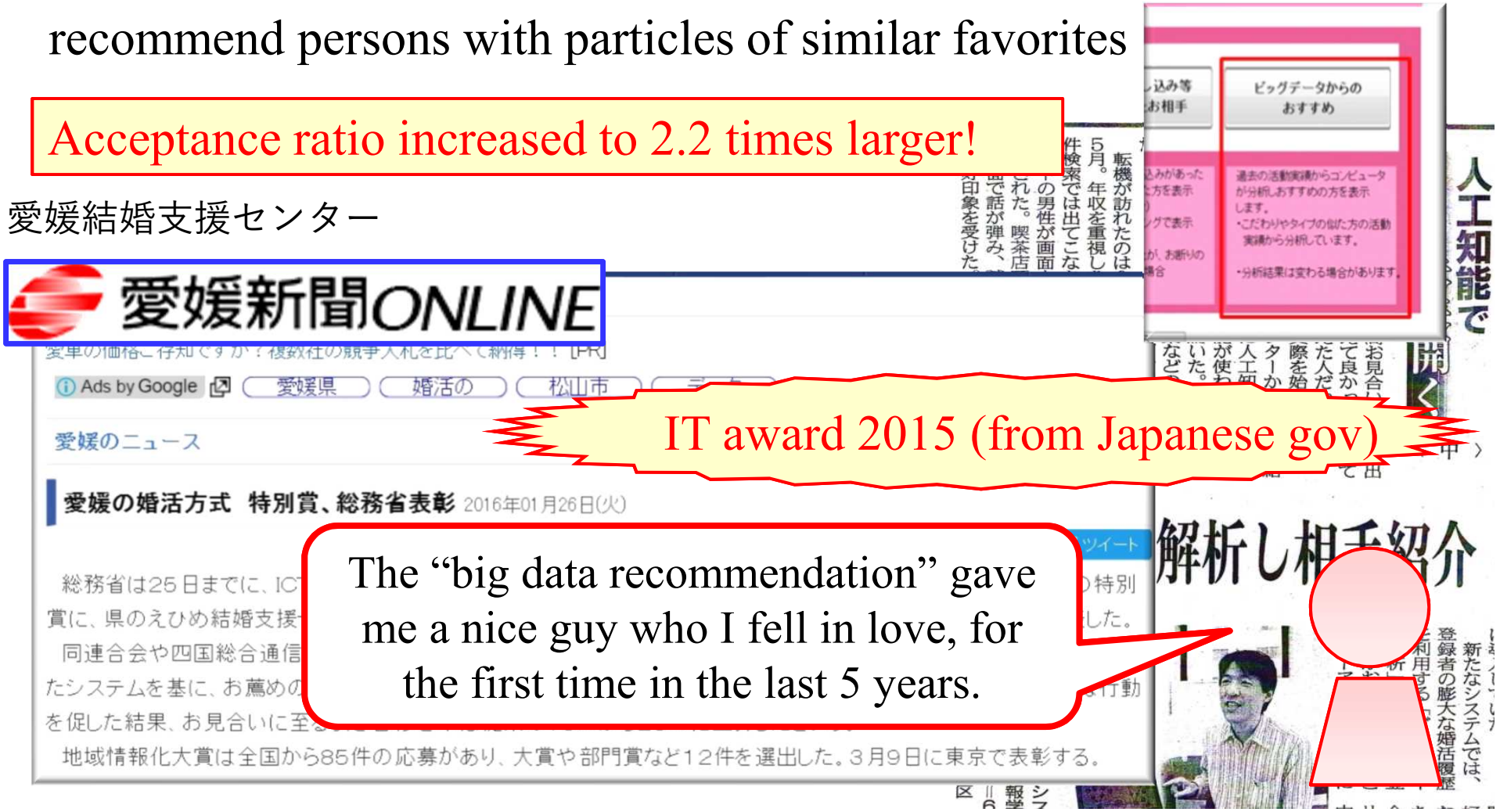
Marriage Matching Recommendation

- Profile search limits the candidates for marriage

→ based on ” similar favorites, similar personality” hypothesis, recommend persons with particles of similar favorites

Acceptance ratio increased to 2.2 times larger!

愛媛結婚支援センター



IT award 2015 (from Japanese gov)

The “big data recommendation” gave me a nice guy who I fell in love, for the first time in the last 5 years.

Targeting on Internet Ads.

- Learn users who may have interest to a particular sites, by using the user's web browsing log data

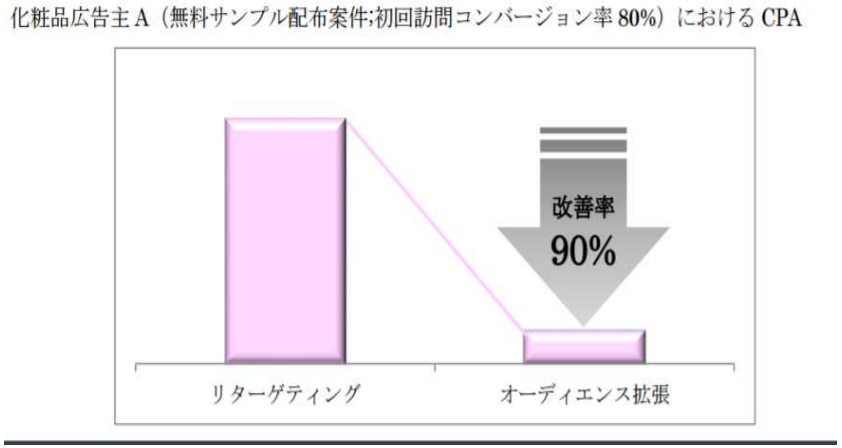
  Apply data polishing to sites clustering, the features of user behaviors had become helpful in learning process

Up to 90% improvements, on the decrease of necessary ads to get required number of clicks

- Similarity of the sites is defined by the similarity of the sets of users visiting the sites

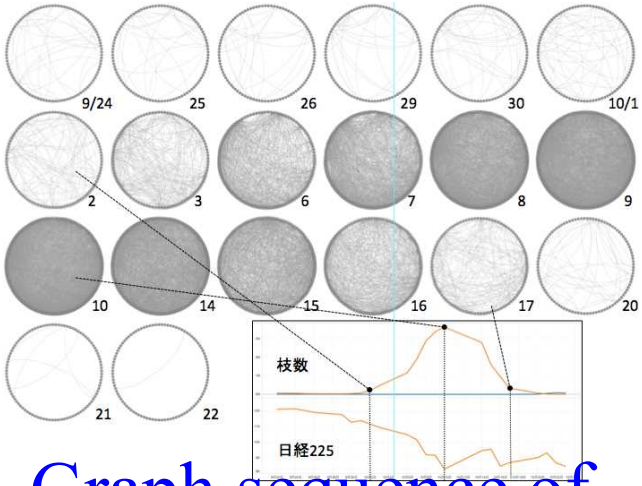
PLATFORM ONE

News Release

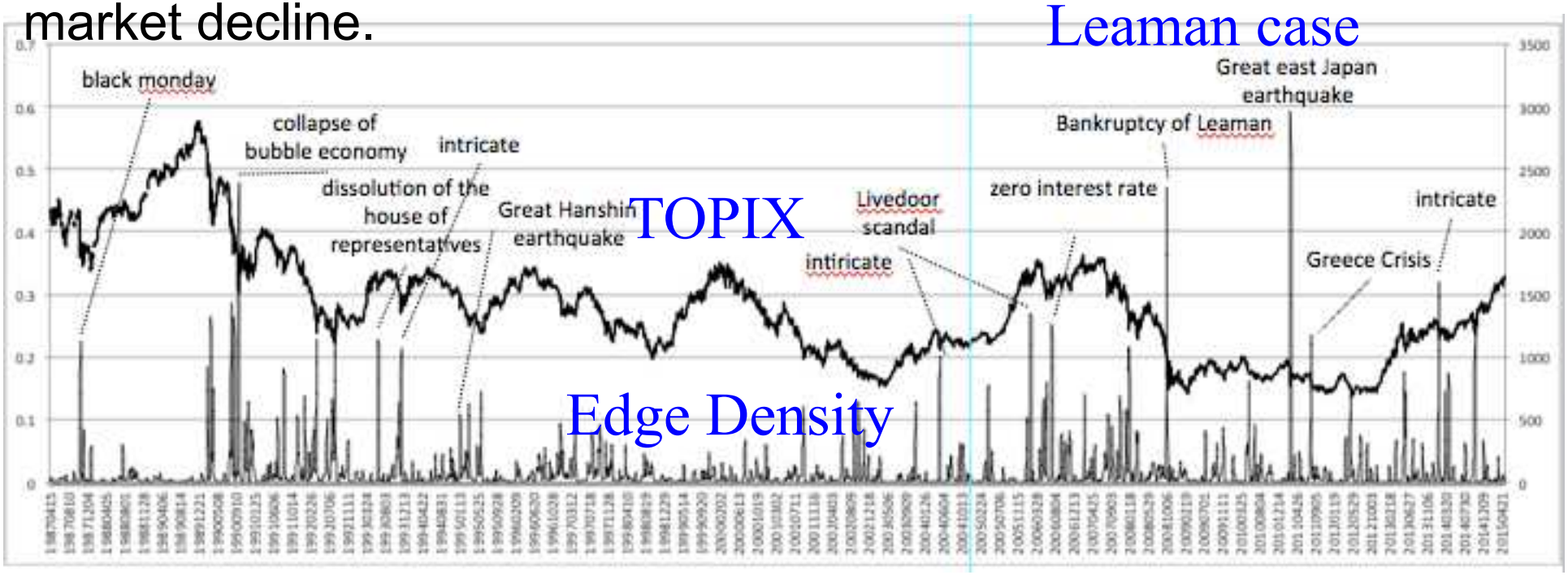


Herding in Financial Markets

- Model building of "herding" phenomenon using individual stock price data.
- Representation by graph sequence
- Similarity graph with pairs of equities having price co-movement
- Edge density predict major bottoms of market decline.



Graph sequence of Leaman case



Coming Future

AI development makes people easy to analyze data

--- then, they would begin to post the result of analysis

Without clear understanding, it brings chaos

When data analysis becomes easy to understand, the analysis with deep understanding would give a new and wide view of the world, system and society

Making the analysis understandable is so important