Data Particlization for Next Generation Data Mining

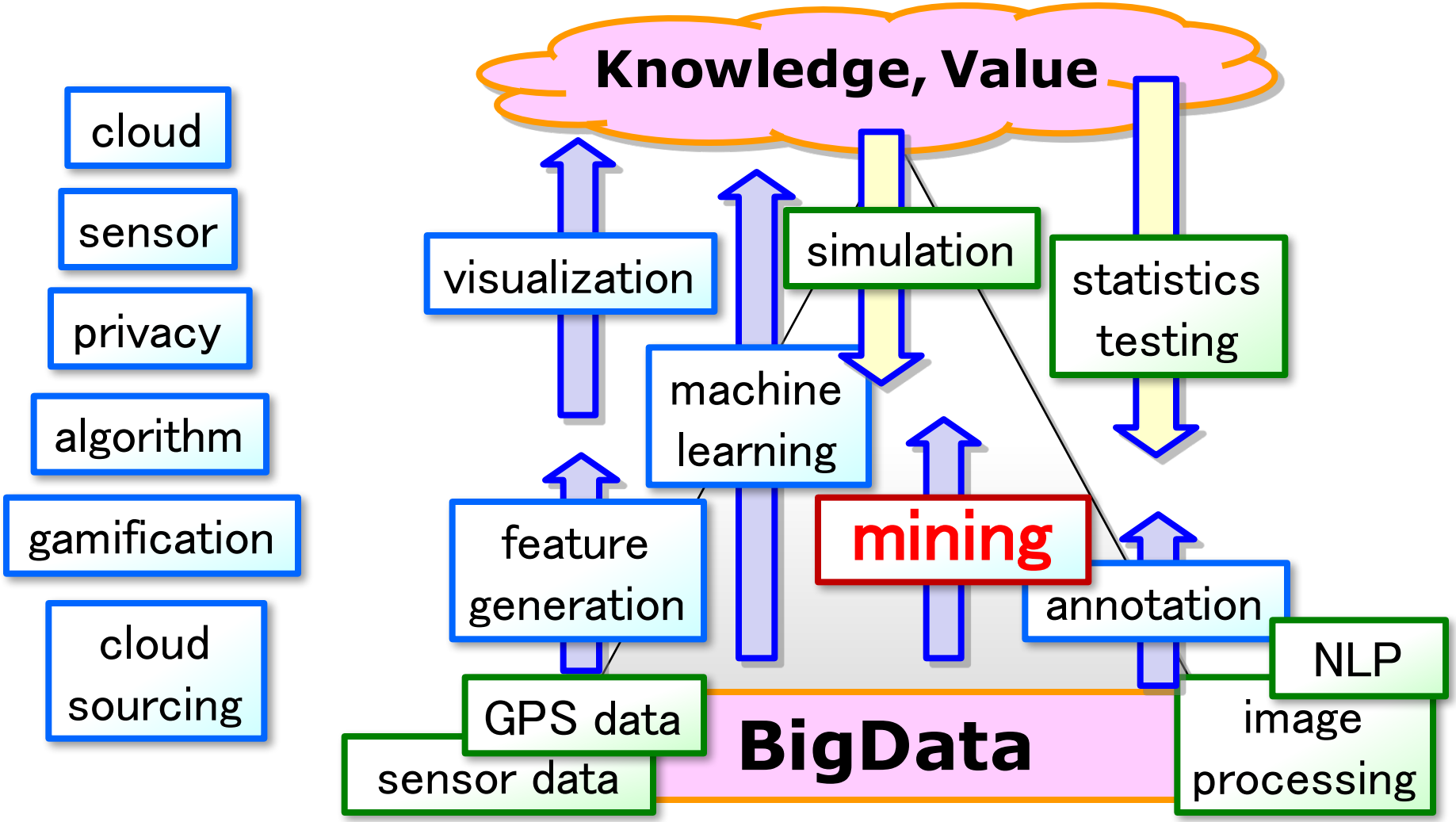
Takeaki Uno National Institute of Informatics

Akihiro Yamamoto Kyoto University
Yukinobu Hamuro Kwansei Gakuin University
Kumiyo Nakakoji Kyoto University

21/Apr/2016

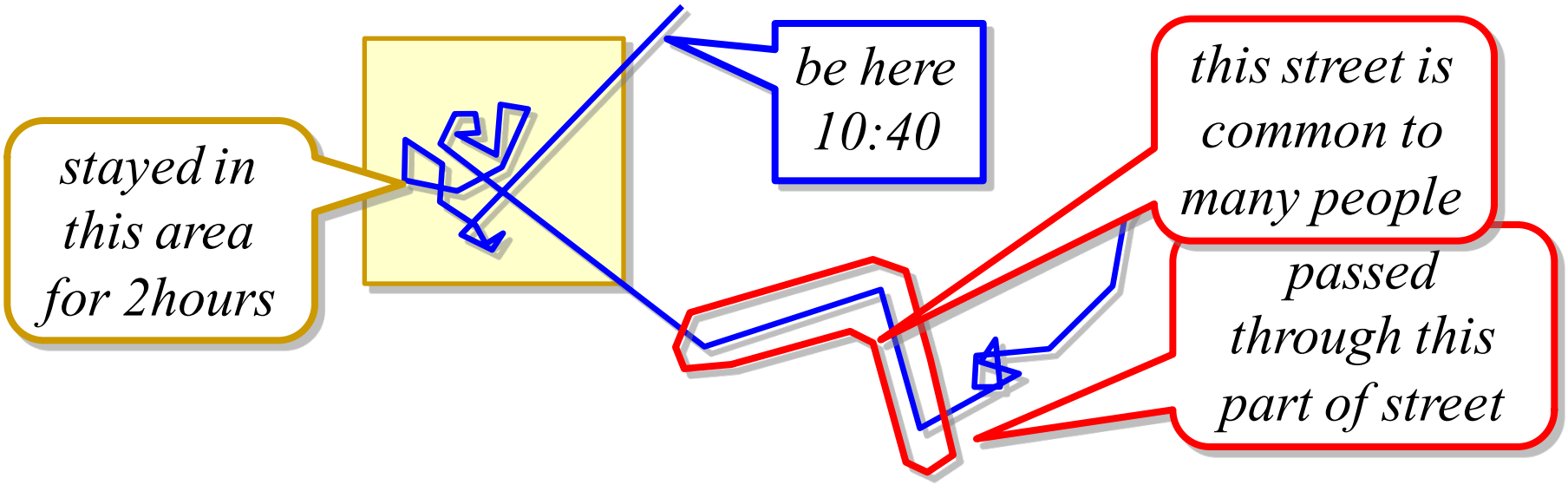
Mining & other Methodologies on BigData

Mining finds **structures** that are used by methods in upper layers from big, shallow meaning, sparse, noisy, and ill-granularity data



Data Abstraction

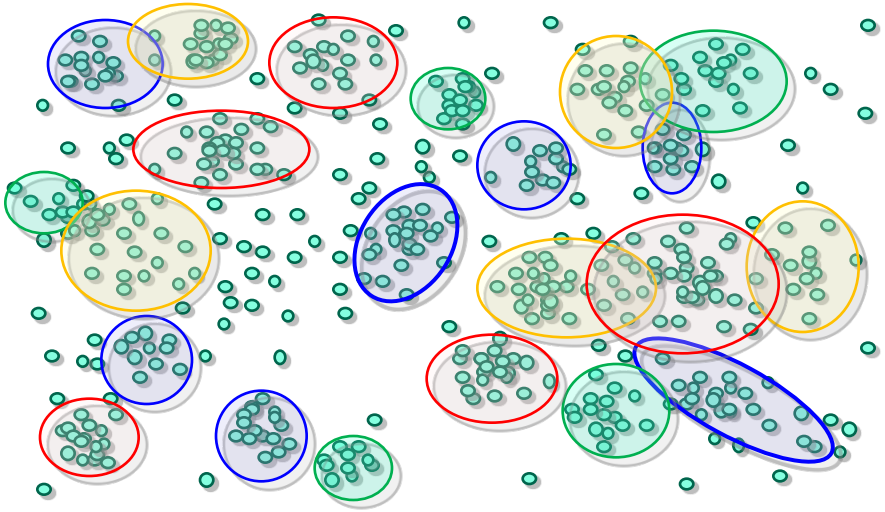
ex.) trajectory: sequence of points → sequence of few places



ex.) data points:

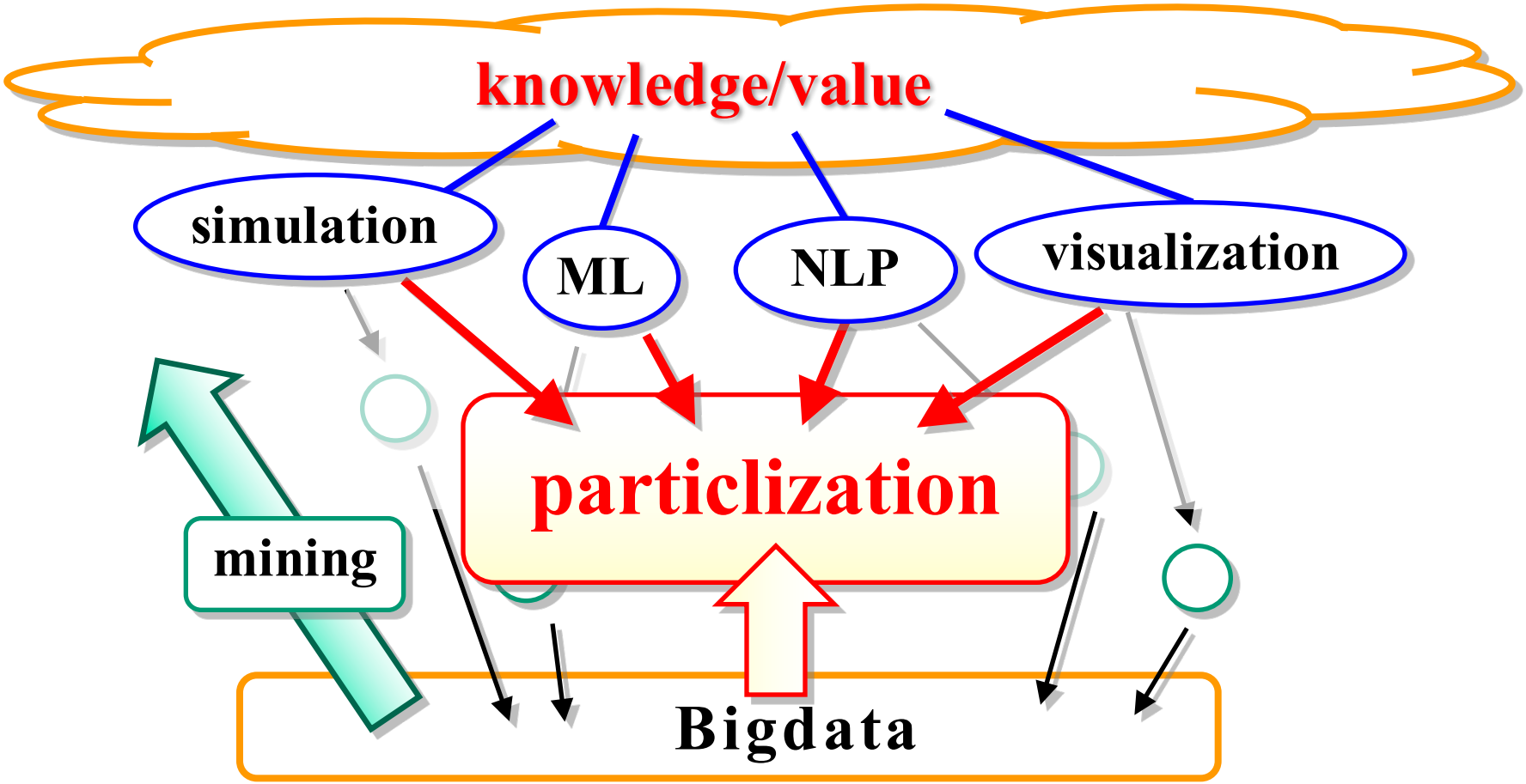
data objects →

local groups of similar objects



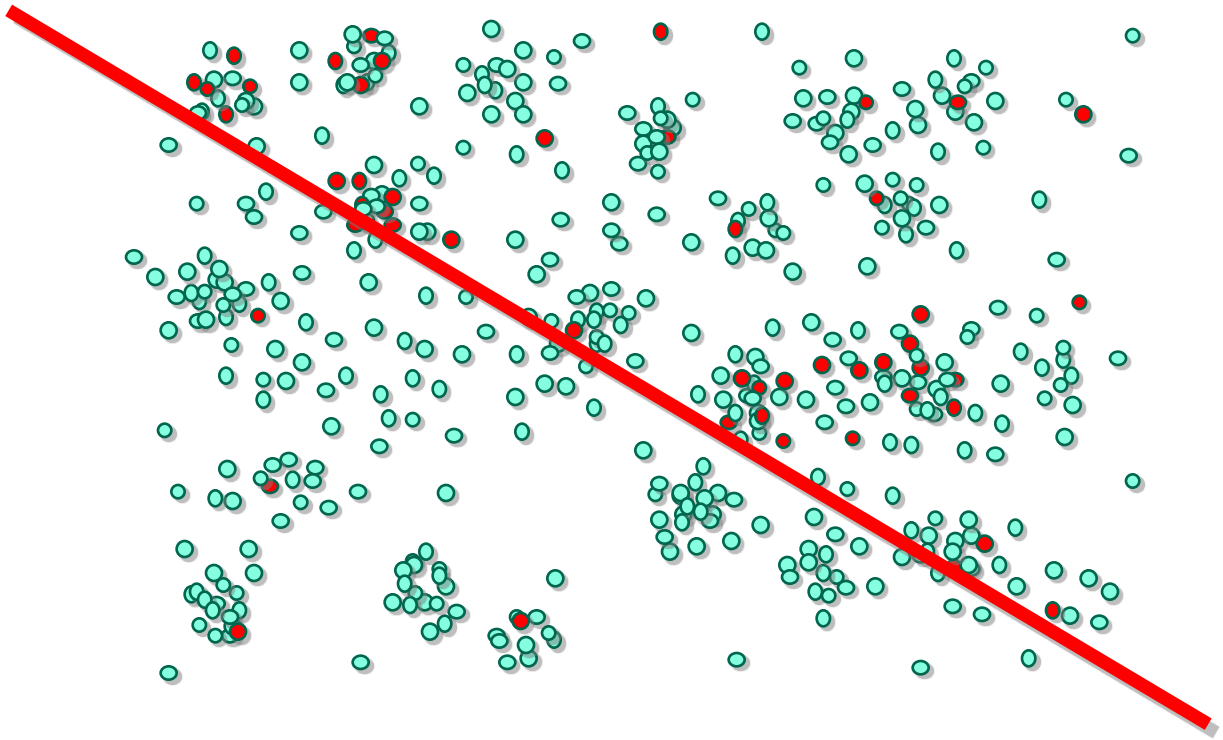
The Data Particlization Approach

- 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



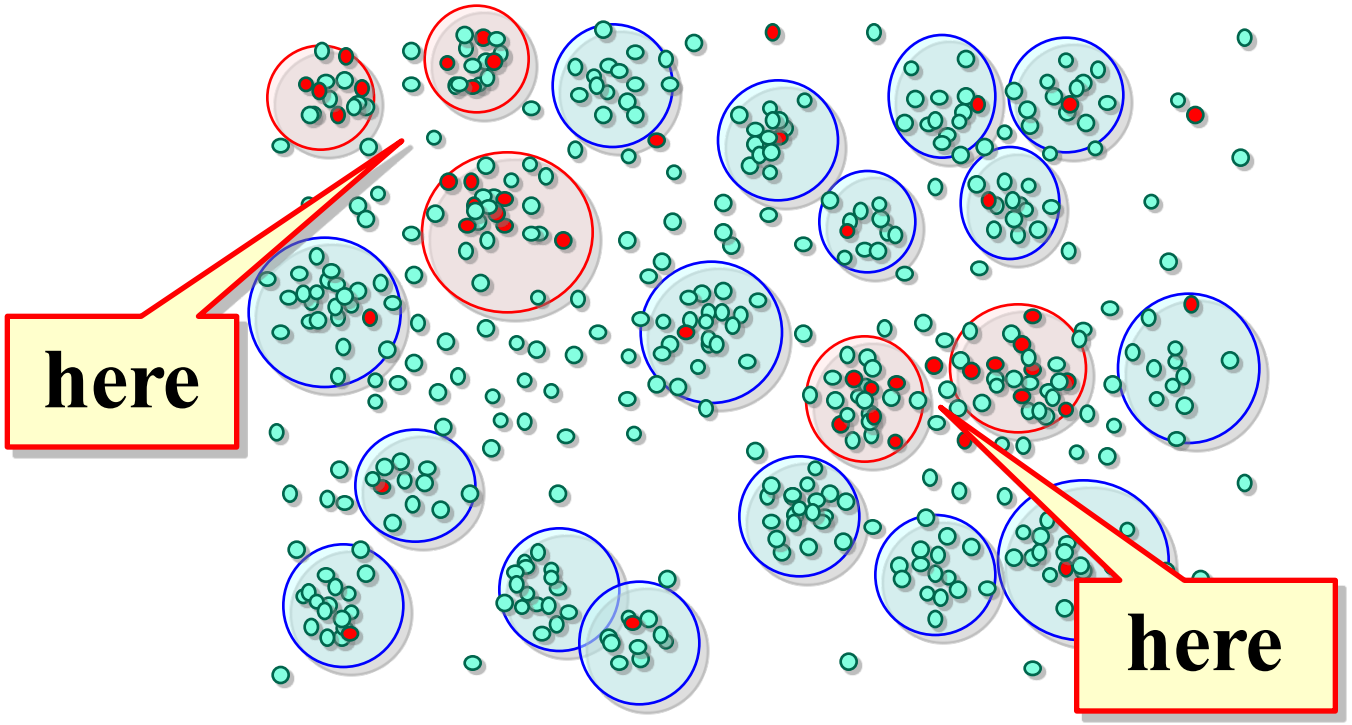
Machine Learning without Abstract

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



With Particles

Easier to get some meanings, or inspires

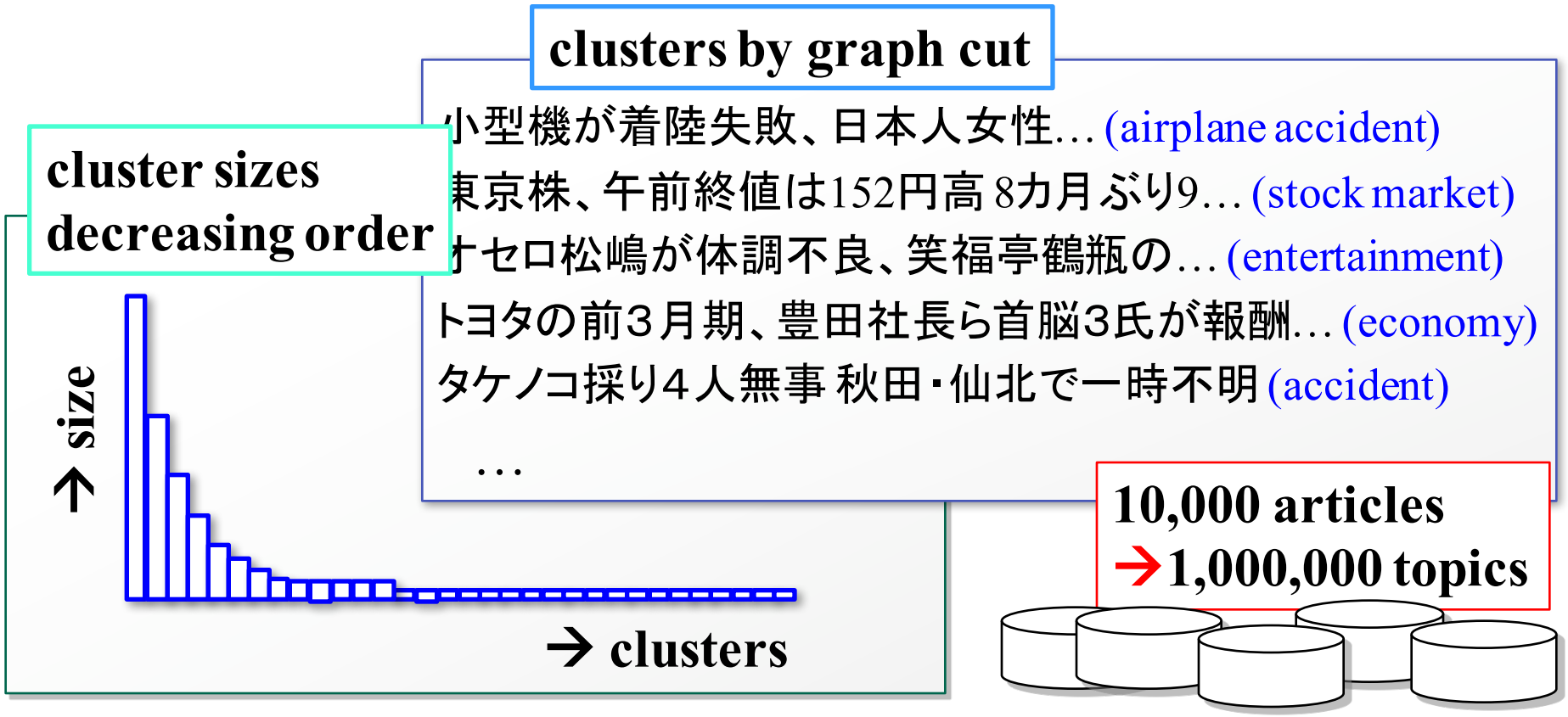


Why not Clustering?

Clustering finds (global) **"classes"**, but particles are **"structures"**

… so, has many problems

huge small solutions, unbalanced sizes, skewed granularity



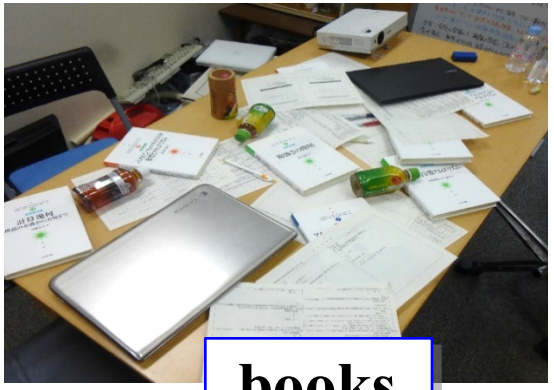
Basic Idea : Clarify Structures

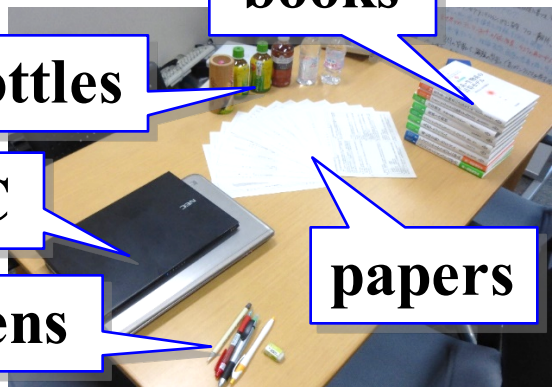
Why bad? ... because, the boundaries of the structures are not clear

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!

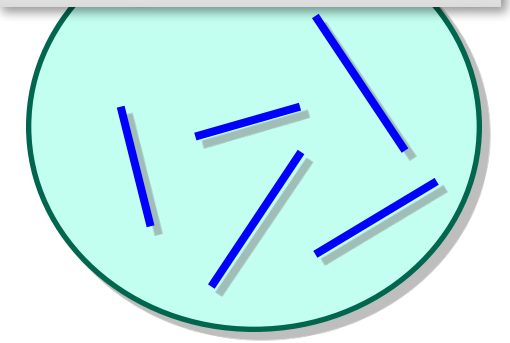




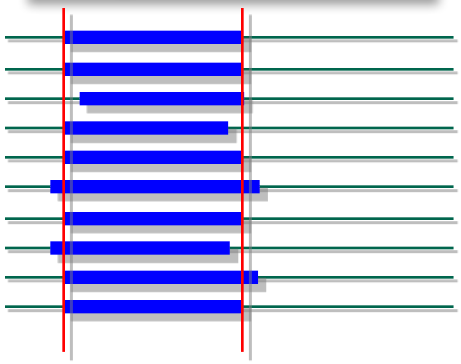
A Proposed Method: Data Polishing

Reveal hidden structures by modifying the data based on feasible hypotheses

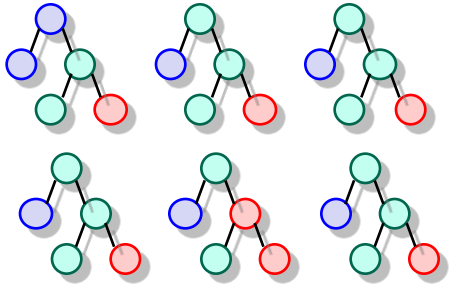
Graph clustering



Segmentation



Pattern mining

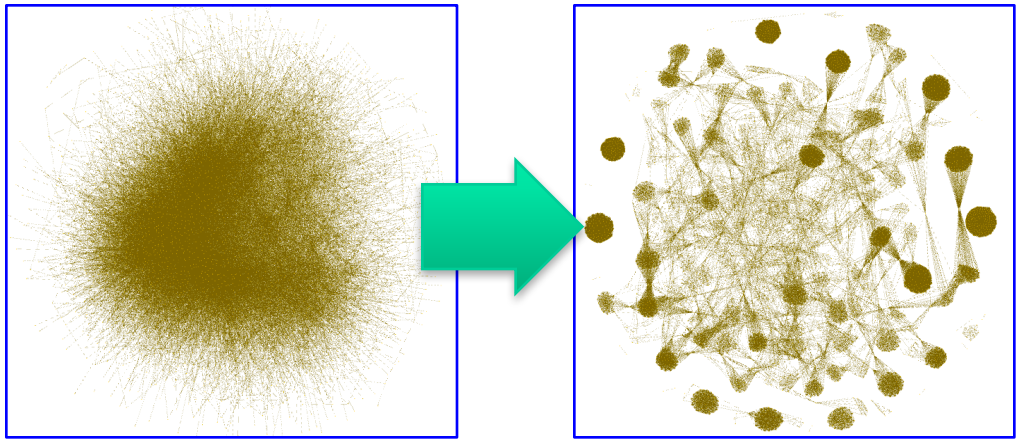


so that

- parts of the data are modified in such a way that any solution and structure would not be lost
- ambiguities are resolved, similar solutions are unified, and the number of solutions is reduced
- the quality of the data analysis will not be deteriorated

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%

+ acceptance ratio for dating proposal in marriage support : **13% → 29%**

+ target size (users to show ads) without loss on internet advertisement : **→ 1/10**

Organization

Modeling & Algorithm

T. Uno NII

Real World Applications

Y. Hamuro Kwansei-Gakuin Univ.

Semantics Analysis

A. Yamamoto Kyoto Univ.

Data Interaction

K. Nakakoji Kyoto Univ.

Extracting **needs** and **importance** from **data/user analysis**, and algorithms for **data polishing** and **semantic structures** of particles