ビッグデータ応用の展開と課題
ガンの個別化医療と都市除排雪への応用を事例として

田中　譲

北海道大学大学院情報科学研究科
特任教授

ビッグデータとは

- [一般的定義]: 3V (Volume, Variety, Velocity), 4V (+Veracity), 5V (+Value)
- [科学技術応用における定義]: ミッション駆動型研究からデータ駆動型研究へのパラダイムシフトを象徴する旗印
  - ミッション駆動型：課題→仮説立案→実験（データ収集・解析）→仮説検証
  - データ駆動型：大規模データ収集⇒（（課題）→仮説立案→データ検索・解析→仮説検証）

パラダイムシフトとしてのビッグデータ

- ミッション駆動型研究からデータ駆動型研究へ
- このパラダイムシフトの契機
  - ウェブコンテンツの急増
  - サーチエンジン・サービスの台頭と利用者ログの獲得・蓄積　→　意図のデータベース
  - モバイル情報の獲得・蓄積（スマートフォン、プローブ・カー・システム）
  - 新世代DNAシーケンサの普及
  - IoT（物のネットワーク）の急速な進歩普及
    - 社会において
      - 多様なセンサーとアクチュエータのネットワーク
    - 先端科学技術研究環境において
      - 計測・観測・分析機器のディジタル化とネットワーク結合
  - SNSの普及　→　人の繋がりの分析
  - IBM Watson

領域趣旨

同じ戦略目標の下に2つの領域を設定

戦略目標

文部科学省 MINISTRY OF EDUCATION, CULTURE, SPORTS, SCIENCE AND TECHNOLOGY-JAPAN

分野を超えたビッグデータ利活用により新たな知識や洞察を得るための革新的な情報技術及びそれらを支える数理的手法の創出・高度化・体系化

達成目標①
各アプリケーション分野においてビッグデータの利活用を推進しつつ様々な分野に展開することを想定した次世代アプリケーション基盤技術の創出・高度化

達成目標②
様々な分野のビッグデータの統合解析を行うための次世代基盤技術の創出・高度化・体系化

研究領域1: ビッグデータ応用

CREST

科学的発見・社会的課題解決にむけた各分野のビッグデータ利活用推進のための次世代アプリケーション技術の創出・高度化

研究総括

田中 譲

北海道大学大学院情報科学研究科
特任教授

研究領域2: ビッグデータ基盤

CREST・さきがけ複合領域

ビッグデータ統合利活用のための次世代基盤技術の創出・体系化

研究総括

喜連川 優

国立情報学研究所　所長/
東京大学生産技術研究所 教授

副研究総括

柴山 悦哉

東京大学情報基盤センター 教授

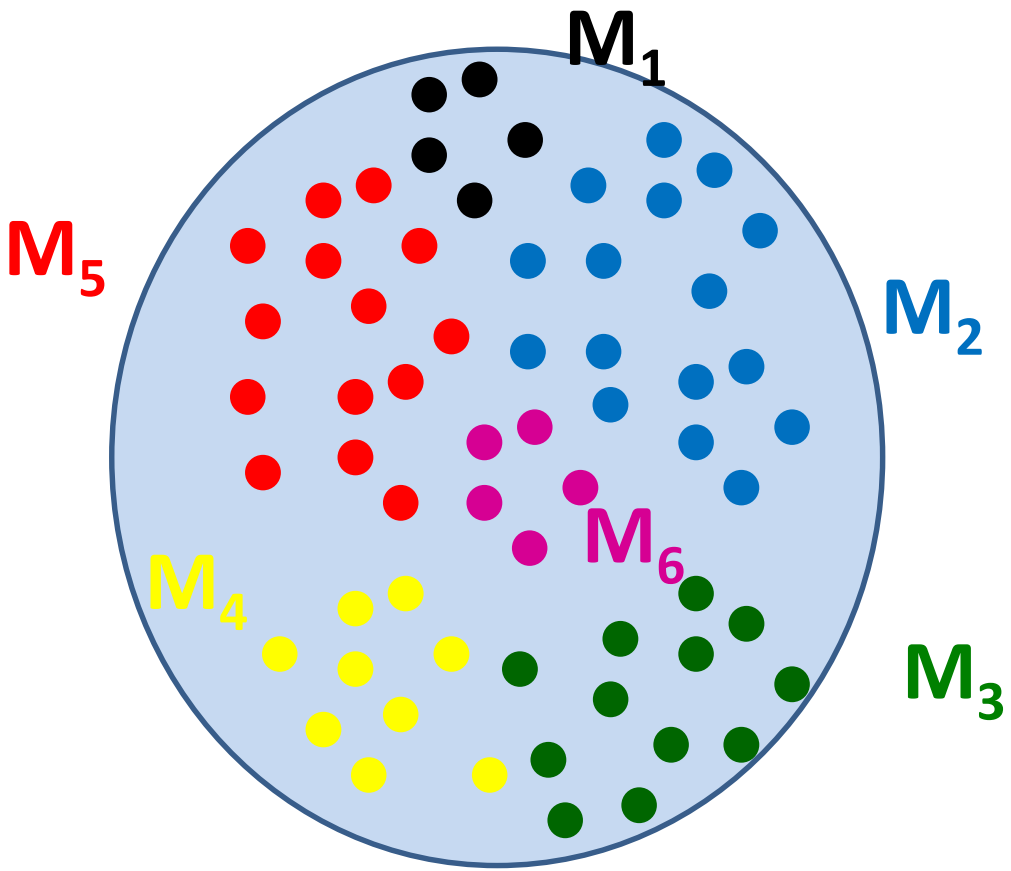
挑戦的ビッグデータ応用の特徴

- **Volume**に重点を置いた**単一種類大規模データ**の分析
- 対象系は比較的**単純**な**単一モデル**に従う（と仮定）

- **Variety**に重点を置いた、**多種類データ**を関連付けた分析
- 対象系は**複雑系（System of systems)**
  - 全体系は**相互作用する複数の要素系で構成**
    - 個々の要素系は**異なる機構**を持ち、**異なるモデル**に従う
    - 全系は機構は複雑すぎて**単一モデルによるモデリングは不可**
  - **複数異種データ**を関連付けた分析
    - 異なるパターンや規則性に従う**異種混合データ**の分析
  - **分析シナリオが定石として確立していない**
    - 特定分析毎に目的に応じた**適切なデータセグメンテーション**が**必要**
    - 試行錯誤的・即興的分析過程の支援が必要

複雑系としての対象系 (1)

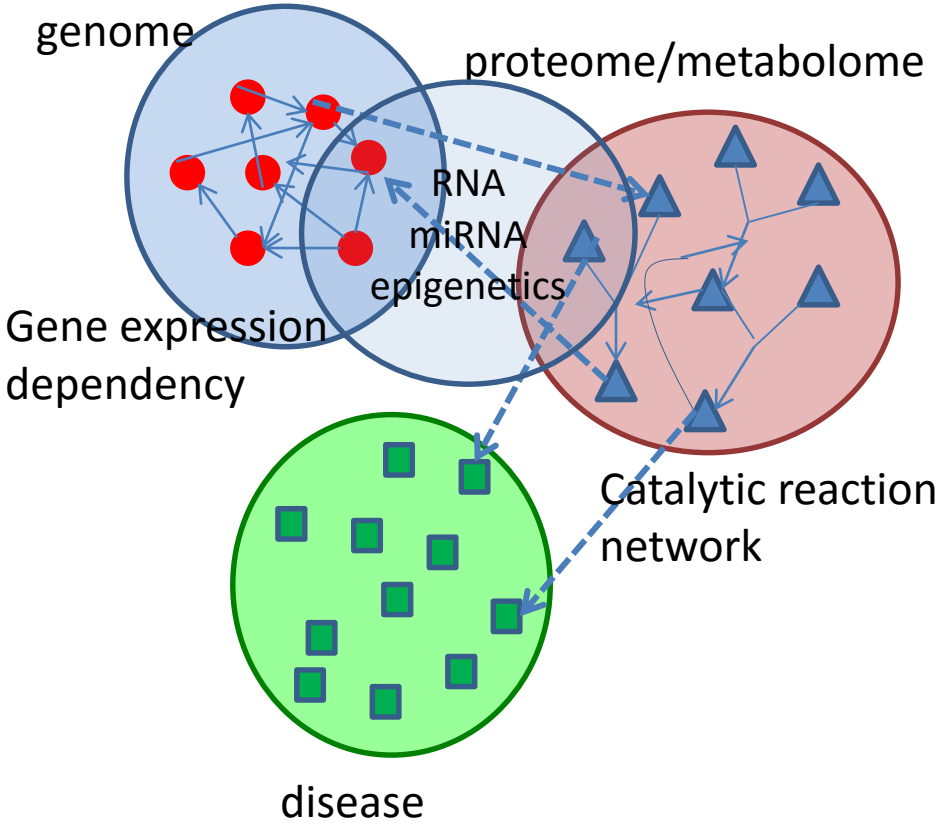
heterogeneous



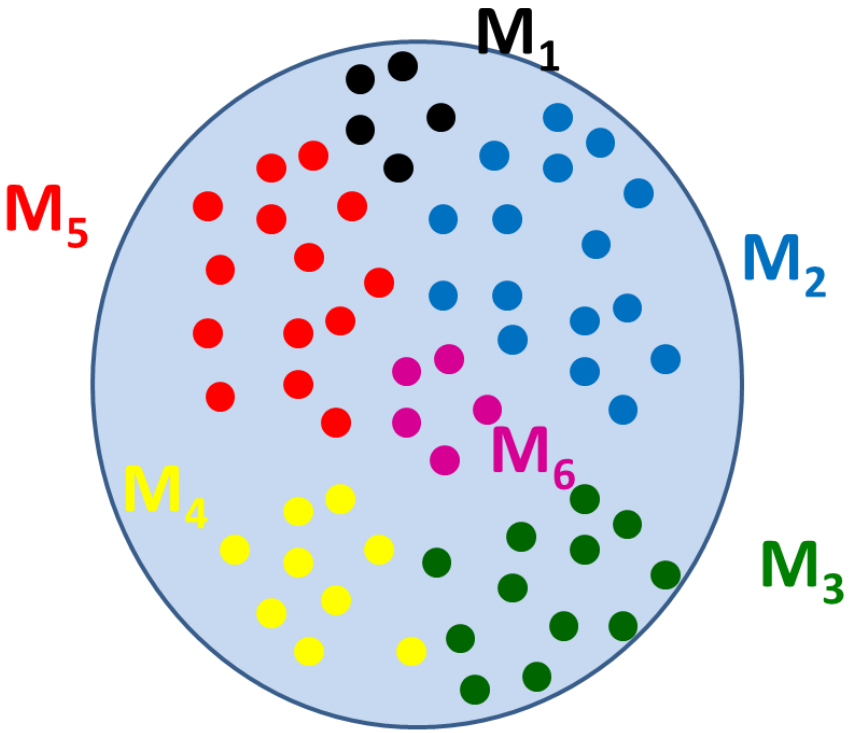
appropriate segmentation!

複雑系としての対象系 (2)

Mutually interacting multiple complex systems



Each subsystem is heterogeneous.



appropriate segmentation!

分析対象が複雑化

- 単一モデルでモデリング可能な対象（系の機構が単純）
  - POS データ, カード利用ログ
  - ウェブ情報, Google 検索ログ
  - etc.

- 異種混合系と考えられる複雑系が対象（系の機構が複雑)
  - 個人化医療
  - 都市規模の社会サービスの最適化
  - etc.

分析シナリオを見つけること自体が研究対象

典型的アプローチ

全体系の機構が複雑だが数学モデリング可の場合

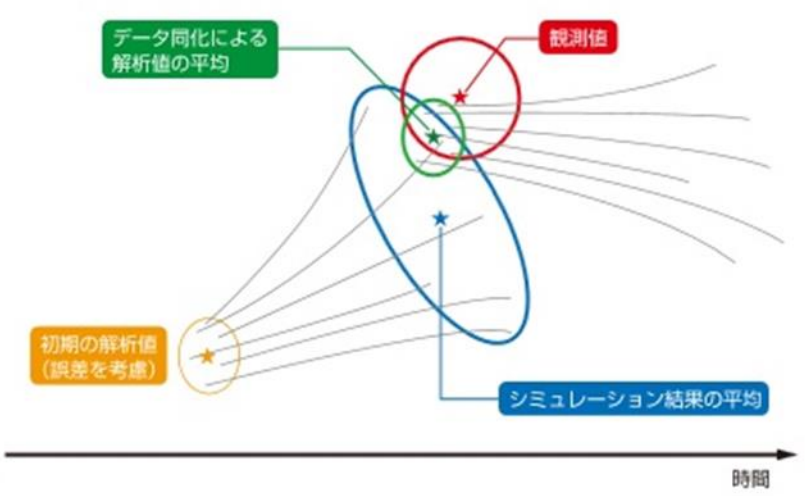
値が未知な多数のパラメータを含む**シミュレーション**
（値未知パラメータの多様な値の組み合わせで**アンサンブル・シミュレーション**）

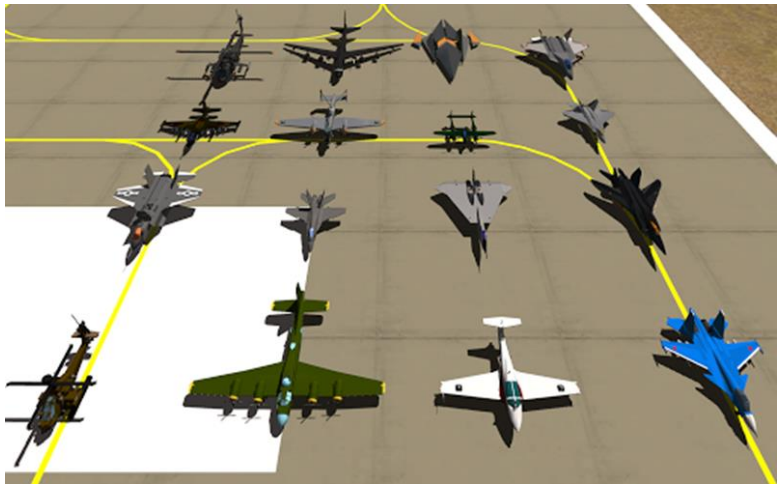
＋ 実データと**データ同化**によりシミュレーション結果を選択

＝ **予測（動的システム）**（例：気象予測）

or

パラメータ値同定（静的システム）(例：航空機設計）

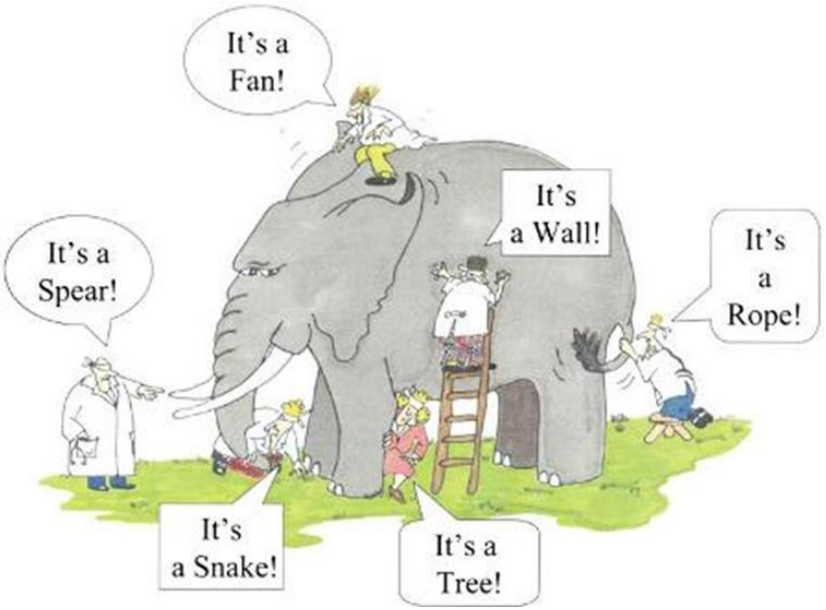




典型的アプローチ

全体系の機構が複雑で数学モデリング不可の場合

- 個々の対象を特徴づける**特徴量（説明変数）の集合を創生**
  ⇒ 個々のオブジェクトを特徴量の**多次元ベクトル表現**
  ⇒ **統計解析、クラスタリング、マイニング、機械学習**が適用可



$X_1=v_1, x_2=v_2, \ldots, x_n=v_n$ → 'elephant'
$X_1=v_1', x_2=v_2', \ldots, x_n=v_n'$ → 'giraffe'
.....
..... → 'elephant'
.....
..... → 'giraffe'
.....

機械学習

$X_1=u_1' x_2=u_2' \ldots, x_n=u_n$ → ?

- 適切な特徴量（説明変数、指標）の創生・定義が最も重要
- 特徴量の創出には、対象系の特定のアスペクトに注目した数理モデリングによる現象解釈が必要

適切な特徴量（説明変数、指標）の創生・定義

- 何故バイオ・インフォマティクスが最初のデータ駆動サイエンスの対象となったか？

- ゲノム＝指標集合
  - ACGTやアミノ酸の特定の並びや、それらの特定の並び、繰り返し回数、それらの間の距離
    - 連（run）、アイランド・パターンなど、すでに文字列分析における指標群と同様に扱える
  - 遺伝子発現：発現の有無は組織をトランザクション、遺伝子をアイテムとみなすと、バスケット分析が適用可能
  - 遺伝子発現の依存関係：　依存関係ネットワーク
- タンパク質／ペプチド
  - PROFEAT：アミノ酸配列からタンパク質とペプチドの構造的・生理化学的特徴量を求めるサービス

The Rise of Material & Catalyst Genome

Why "Catalyst Genome"?

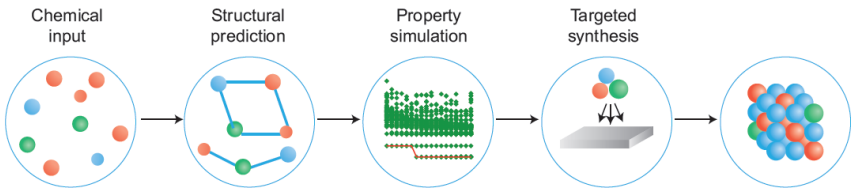
INORGANIC MATERIALS

The quest for new functionality

Building on our understanding of the chemical bond, advances in synthetic chemistry, and large-scale computation, materials design has now become a reality. From a pool of 400 unknown compositions, 15 new compounds have been realized that adopt the predicted structures and properties.

Aron Walsh

Materials chemists are spoiled for choice. The 98 naturally occurring elements of the periodic table give rise to 4,753 potential binary compounds (that is, C_2^{98}), 152,096 ternary compounds and 3,612,280 quaternary compounds, assuming equal amounts of each element in a single phase. The combinations exceed



Angewandte Editorial

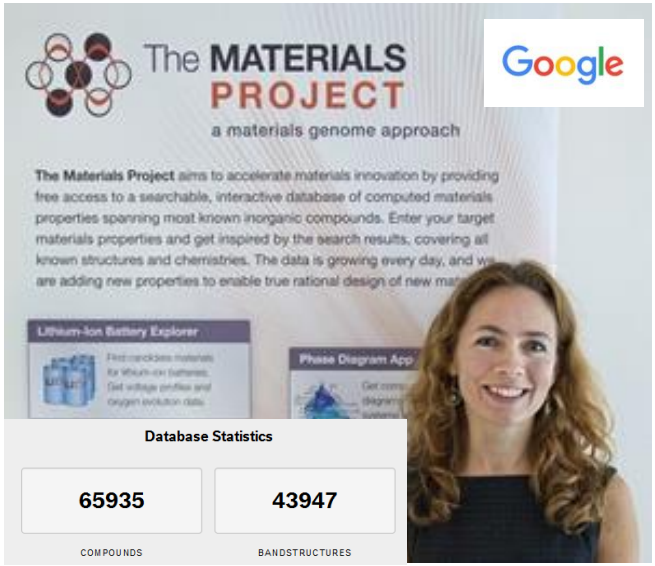
DOI: 10.1002/anie.201208487

The Catalyst Genome

Jens K. Nørskov and Thomas Bligaard**

The quest for the materials genome—the properties of a material that define its functional properties—has started. day's chemical industry, they are also essential for building a completely new,

Jens K. Nørskov
Professor, Stanford University and SLAC National Accelerator Laboratory



応用と基盤技術の間の溝

- 誤解
  - 応用分野の研究者：「データはある。どう分析すればよいか？」
  - CSの研究者：「良いデータ集合さえ用意してくれれば、分析は任せてもらってよい。」
- ともすれば忘れられてしまう観点
  - どのような説明変数に注目してデータを用意すべきか？
  - どのようなアスペクトに注目すべきか？
  - 各アスペクトはどのように数理モデリングすべきか？

応用と基盤技術の間の溝

- 対象系の機構が非常に複雑な応用分野が増えている。
  - 全形の数理モデリングが不可
- 実応用から対象系の本質的なアスペクトを明確に抽出し、アスペクトごとに数理モデリングを行い、適切な特徴量（説明変数、指標）を創生・定義
  - → 対象系の本質的なアスペクトに注目し、**数学モデリングが行える人材が不足**
  - → 対象系の数理モデリングに基づく理解が不足
- 実応用に即し、種々の分析手法の中から最適なものを選び最適分析シナリオを構築して適用
  - → **データサイエンティストの不足**
    - **アルゴリズムの研究者ではなく、そのプラグマティクスを理解し、与えられた課題に対して最適な分析シナリオを構築できる人材**
  - → 各種分析手法の意味理解力の不足
  - → 課題に即した**モデリング能力と分析シナリオ構築能力の不足**

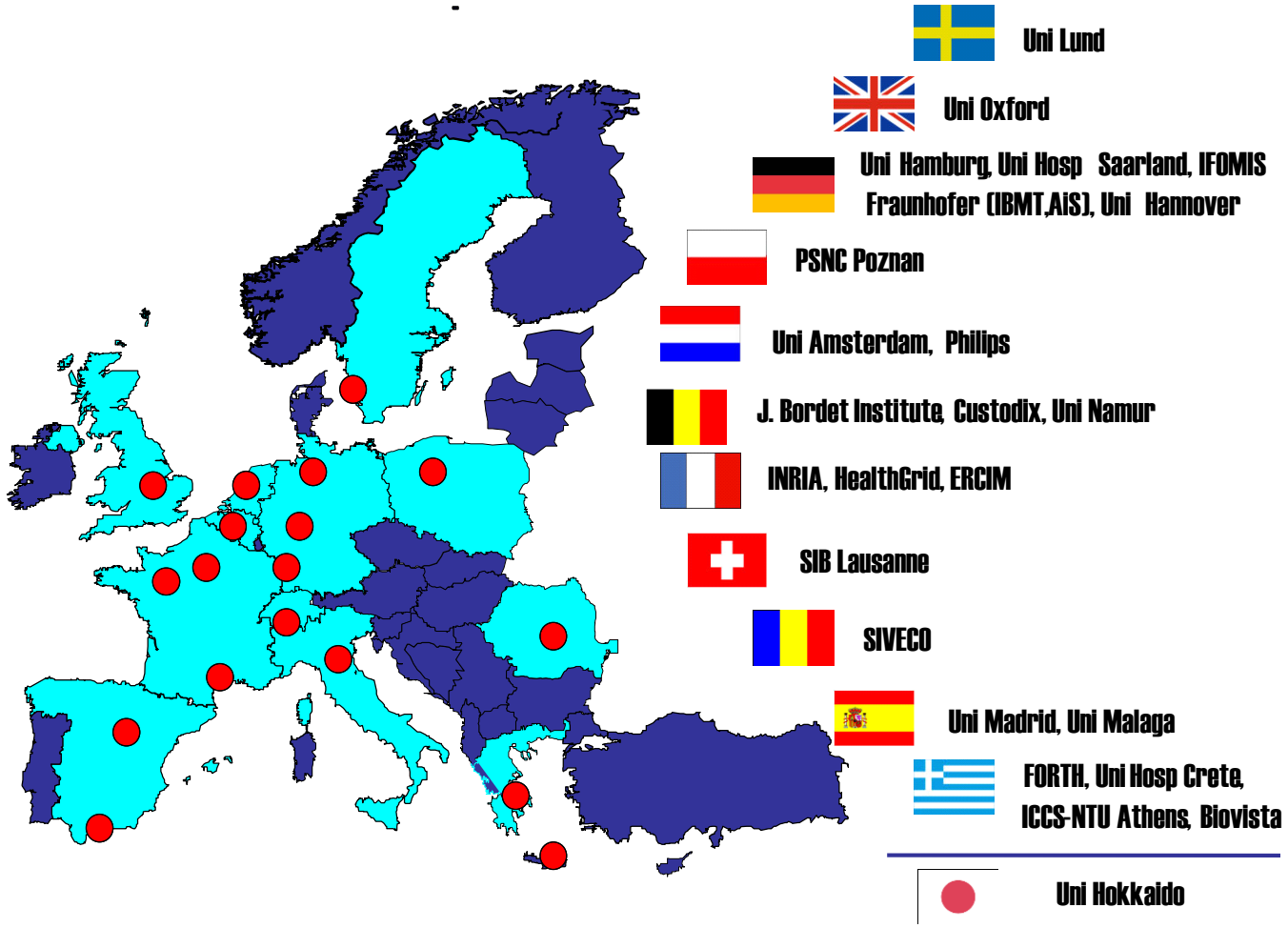
My involvements in Big Data projects

- Cutting-Edge Data-Based Science or e-Science
  - EU FP projects for integrated IT support of clinical trials on cancer
    - FP6 Integrated Project ACGT (Advancing Clinico-Genomic Trials on Cancer) (02/2006 – 07/2010)
      - 26 teams
    - FP7 Large-scale Integration Project p-medicine (personalized medicine) (02/2011 – 01/2015)
      - 29 teams
- Urban Monitoring and Social Service Management
  - MEXT initiative project on Social CPS (Cyber-Physical System) for Efficient Social Services (09/2012-03/2017)
    - Project Consortium (NII (National Institute of Informatics), Hokkaido Univ., Osaka Univ., Kyushu Univ.)
- Program Officer of the JST CREST Program on Big Data Applications (2013-2020)
- Collaboration with Dr. Keisuke Takahashi in Material Informatics (2014-)

My involvements in Big Data projects

- Cutting-Edge Data-Based Science or e-Science
  - EU FP projects for integrated IT support of clinical trials on cancer
    - FP6 Integrated Project ACGT (Advancing Clinico-Genomic Trials on Cancer) (02/2006 – 07/2010)
      - 26 teams
    - FP7 Large-scale Integration Project p-medicine (personalized medicine) (02/2011 – 01/2015)
      - 29 teams
- Urban Monitoring and Social Service Management
  - MEXT initiative project on Social CPS (Cyber-Physical System) for Efficient Social Services (09/2012-03/2017)
    - Project Consortium (NII (National Institute of Informatics), Hokkaido Univ., Osaka Univ., Kyushu Univ.)
- Program Officer of the JST CREST Program on Big Data Applications (2013-2020)
- Collaboration with Dr. Keisuke Takahashi in Material Informatics (2014-)

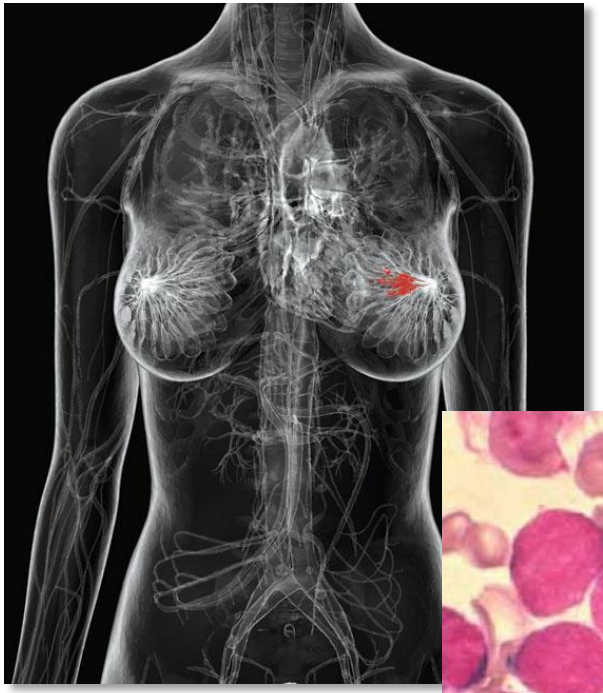
The ACGT Consortium



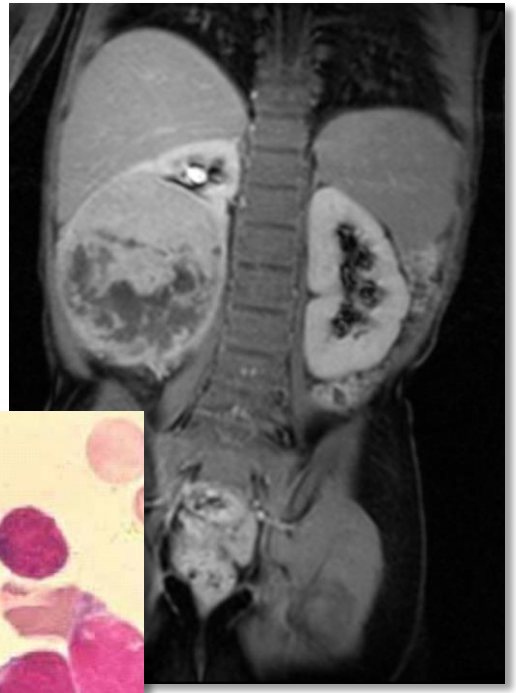
The p-medicine Consortium

Cancer Domains

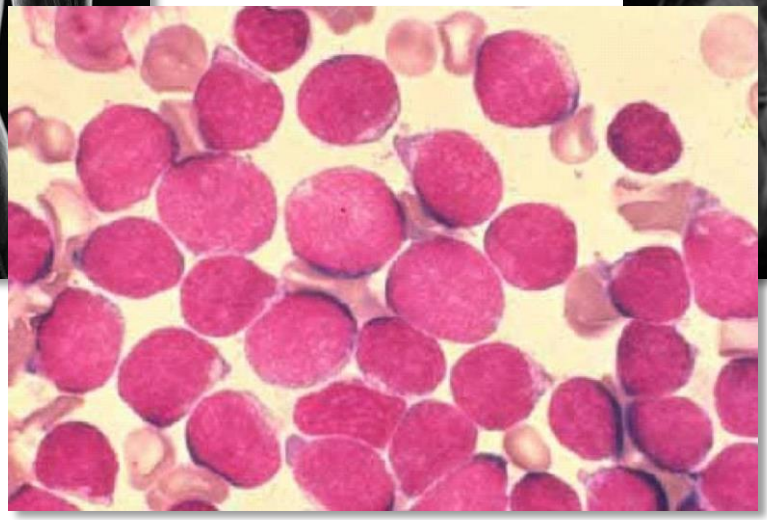
Breast Cancer



Nephroblastoma



ALL



Large scale data & computing

Seamless access and integration of distributed, heterogeneous data
in a data warehouse repeatedly over time (≈ 200 GB / patient and time point)

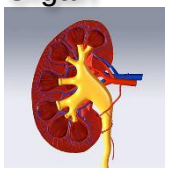
size

Organism

Clinical data

Patient characteristics, age, gender, treatment, outcome, acute toxicity, late effects, …

Organ



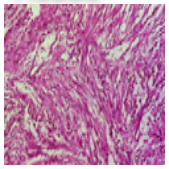
Imaging and segmentation data

CT, MRI, Positron Emission Tomography, …

Functionality

Kidney function, …

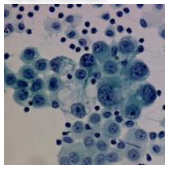
Tissue



Histopathology

Immunoassaying, in situ hybridization, cellularity, proliferation, …

cellular

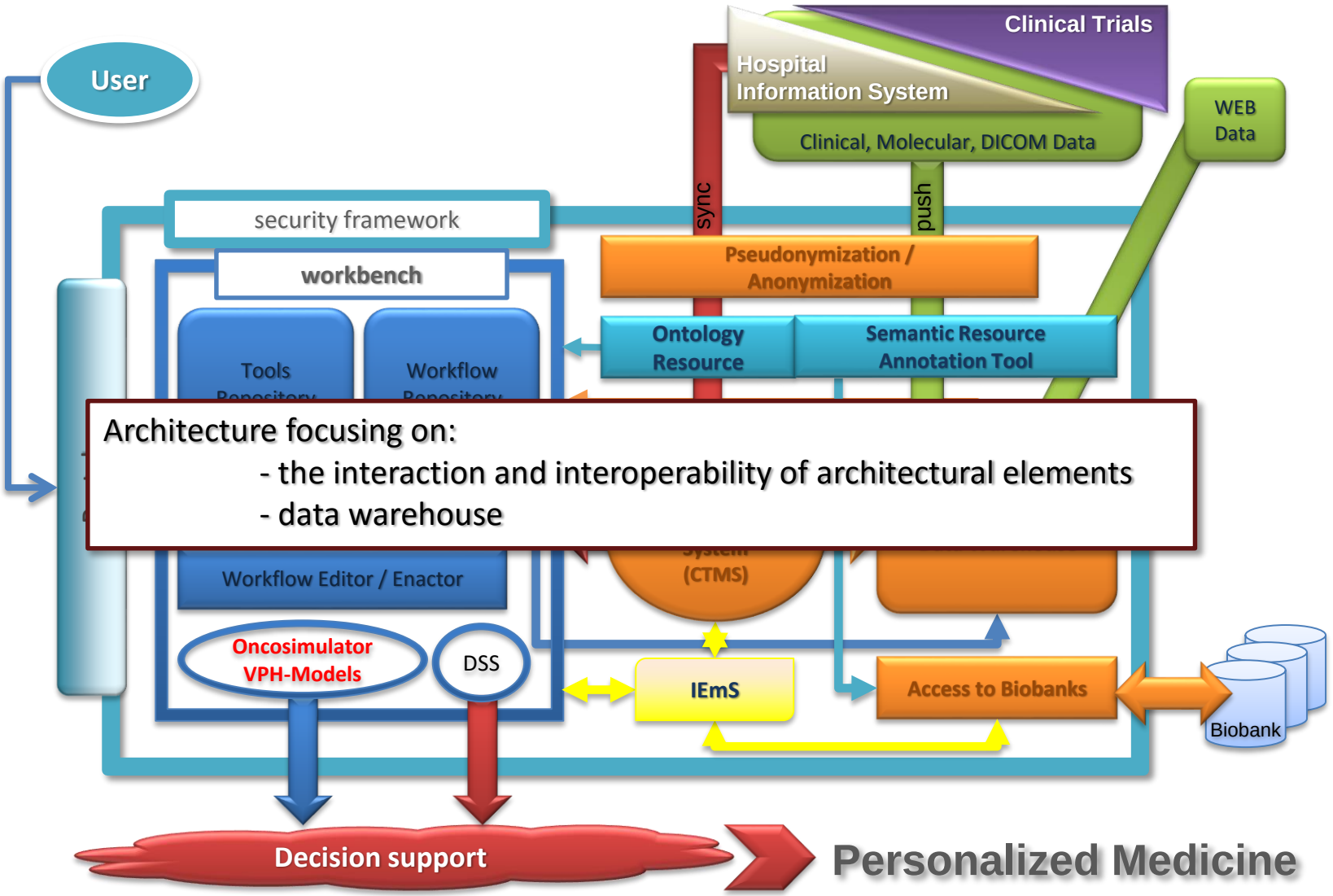


Immunology, Cytogenetics
circulating tumour cells
stem cells

sub-cellular

DNA Level → sequencing, epigenetic profiling, array CGH
RNA Level → miRNA profiling, gene expression profiling
Protein Level → Tumour specific auto-antibody profiling, proteomics

Infrastructure Project



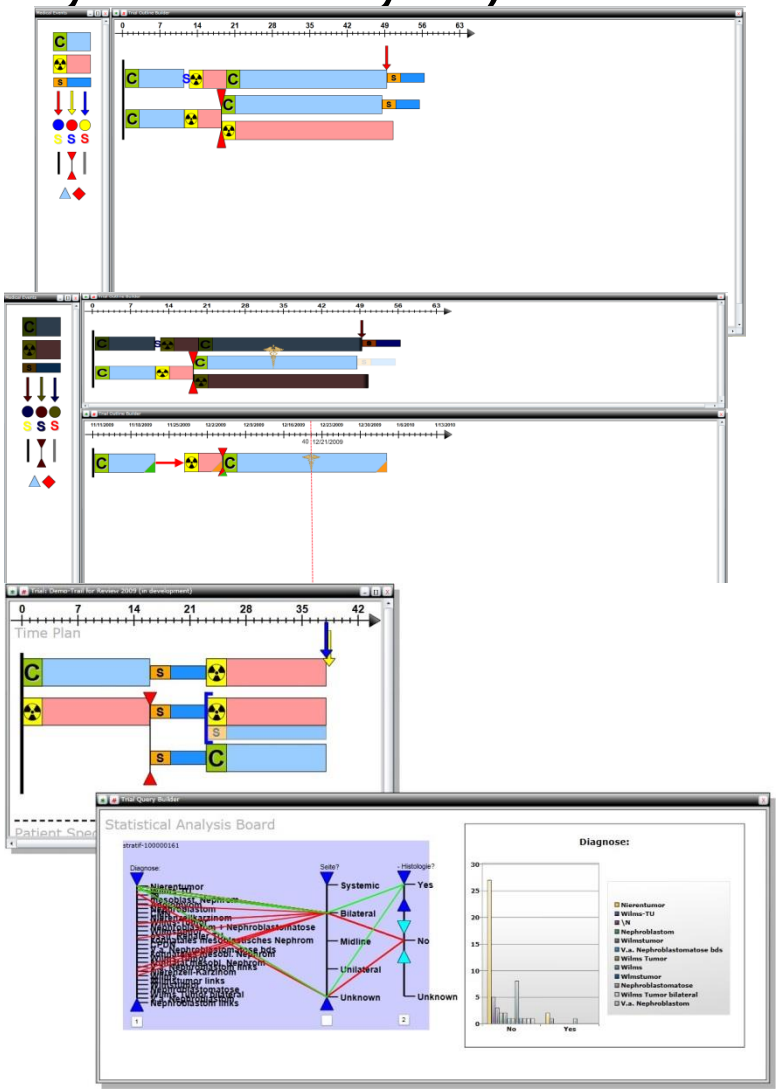
Led by IRI
(legal partner in p-medicine)

- established in 1983
- first Institute dedicated to this goal at a German University
- Directors of the IRI:
  - Prof. Dr. Nikolaus Forgó,
  - Prof. Dr. Axel Metzger, LL.M.
- Special focus on data protection in medical research and IP-issues with regard information technology

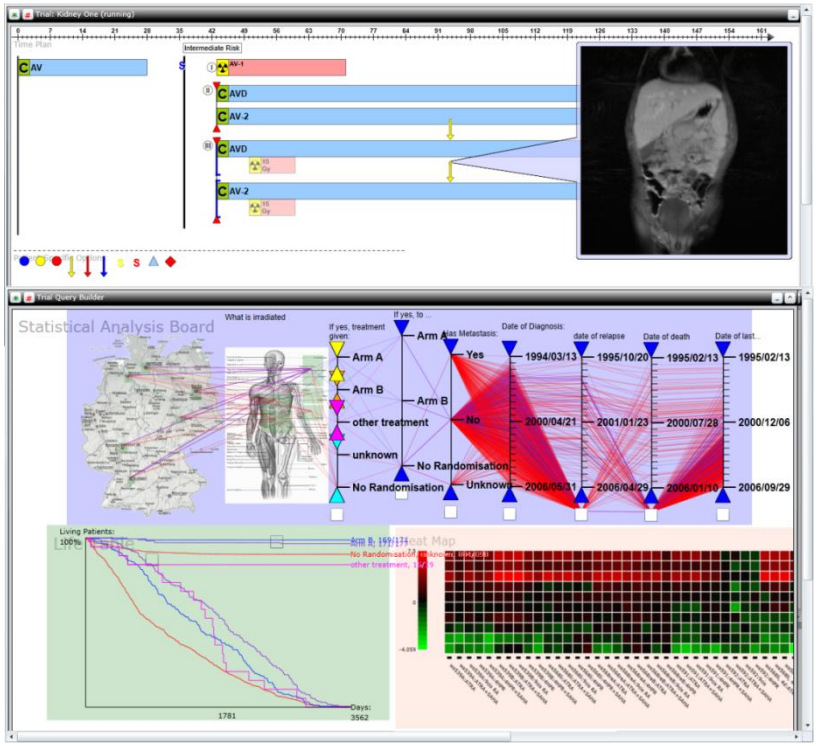
Trial Outline Builder (TOB) (2010)

(Web-top integrated environment for planning trials, patient data acquisition, and exploratory data analysis)

- Trial Plan Editing
  - Copy-and-paste of trial event types to design both a trial flow graph and a set of some additional events outside the flow.
  - A click of each event opens its CRF editor
- Patient Treatment View : CRF input for each patient through the TOB
  - Possibly with the specification of some additional outside-of-flow events
- Query & Analysis View: Querying the DB
  - for specific cases for their statistical analysis or the visualization of correlations among specified items



How to find out patient cases in which some treatment arm may show significantly better performance than others?



In traditional clinical trials, patient segmentations were *a priori* designed.

This is not the case in personalized medicines.

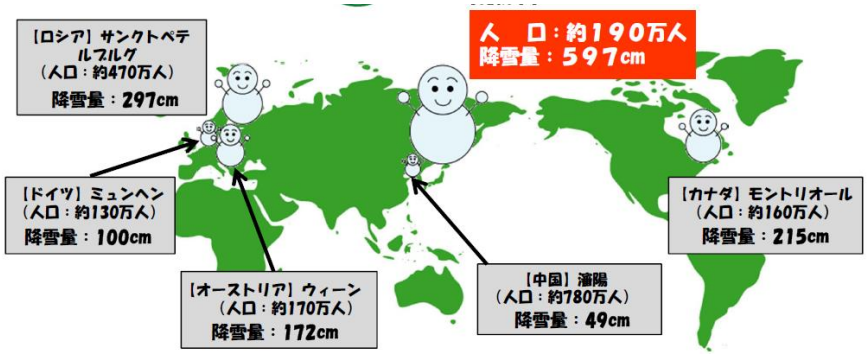
Each stratified patient group leading to a randomization of more than one candidate arm may show no significant difference of survival rates among them without further segmenting this patient group.

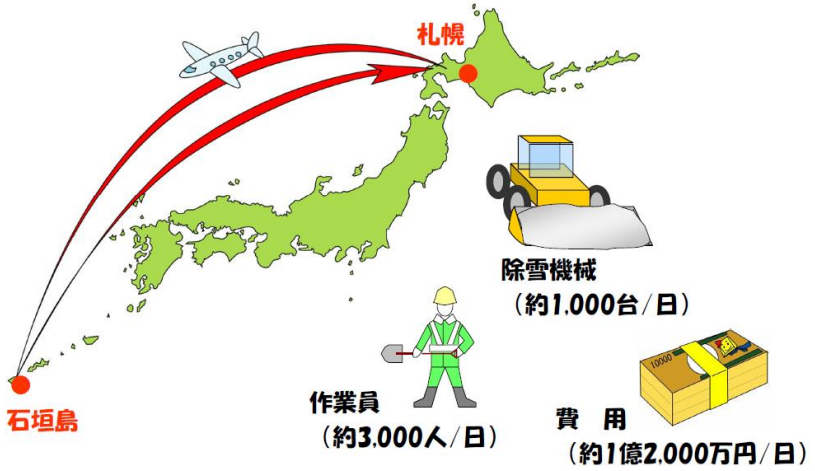
My involvements in Big Data projects

- Cutting-Edge Data-Based Science or e-Science
  - EU FP projects for integrated IT support of clinical trials on cancer
    - FP6 Integrated Project ACGT (Advancing Clinico-Genomic Trials on Cancer) (02/2006 – 07/2010)
      - 26 teams
    - FP7 Large-scale Integration Project p-medicine (personalized medicine) (02/2011 – 01/2015)
      - 29 teams
- Urban Monitoring and Social Service Management
  - MEXT initiative project on Social CPS (Cyber-Physical System) for Efficient Social Services (09/2012-03/2017)
    - Project Consortium (NII (National Institute of Informatics), Hokkaido Univ., Osaka Univ., Kyushu Univ.)
- Program Officer of the JST CREST Program on Big Data Applications (2013-2020)
- Collaboration with Dr. Keisuke Takahashi in Material Informatics (2014-)

Snow Removal in Sapporo as a Large-scale Complex Social Service

- **Population:** **1,920,739**
- **Annual snowfall:** **597cm**
  - The largest annual snowfall among the cities with more than 1M people in the world
- **Annual budget for snow plowing and removing (2010):**
  14,729,000,000 yen
  (147,000,000 $)
  2nd last season:
  22,000,000,000 yen
  (220,000,000 $)
- **Total distance of snow plowing and removing during a single night:** **5,328km**





Preparation of Data and Realtime Monitoring

- **Traffic Data**
  - **Probe-car data**: **private cars** (past 2 yrs), **taxi cars** (past 2 yrs + realtime: 12/2013-)
  - Realtime probe-car data from buses (20 buses)
  - Traffic jam sensor data (past 2 yrs)
  - Probe-person data (retrospective data +realtime data?)
  - Statistical subway passenger records (past 9 yrs)
- **Weather Data**
  - Meteorological multi-sensor data (52 locations) (past 10 yrs +realtime data?)
  - Weather mesh data (past 4 yrs)
  - X band MP radar data (realtime: 12/2013-)
- **Snow plowing and removing records**
  - **Retrospective records** (past 5 yrs +real-time data)
  - **Realtime probe-car data** (25 vehicles)
  - Complaints from residents (past 5 yrs +real-time data?)
- Traffic accident data
  - Injury or death traffic accident records from Hokkaido Police Office (10 past yrs)
- Road condition
  - 3D road measurement by a bus with a laser range scanner (realtime: 02/2014-)

Clustering of road links after snow removal (two days after the snowfall)

Each road link is represented as a vector of 288 dimension showing the change of the average taxi speed at every 5 min. during a day.

The segments from the major roads are generally clustered together.

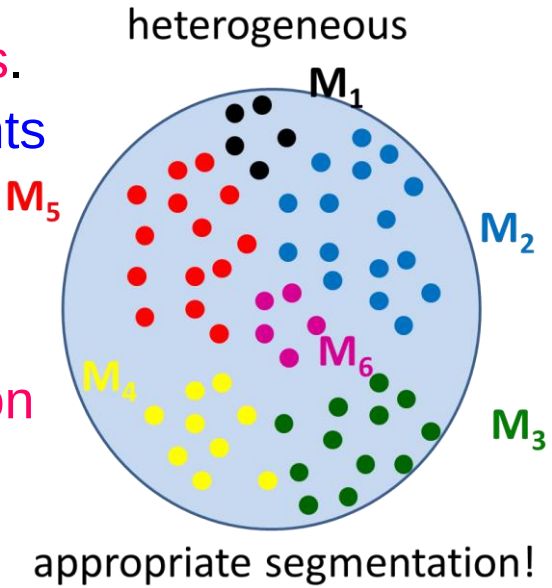
Clustering of road links on a day directly after a heavy snowfall

Each road link is represented as a vector of 288 dimension showing the change of the average taxi speed at every 5 min. during a day.

The segments from the major roads end up in several different clusters.

Macro Analysis vs Micro Analysis

- **Personalized medicine**
  - aims to find out a group of patients in which one candidate treatment may show better survival rate than the other candidate treatments.
  - The whole group of patients is heterogeneous.
  - The analysis of a heterogeneous set of patients cannot extract meaningful knowledge for personalized medicine.
- **Social CPS**
  - The influence of snowfall and snow removal on traffic is not uniform across different road segments.
  - We cannot apply macro analysis methods to obtain any meaningful knowledge about winter roads.



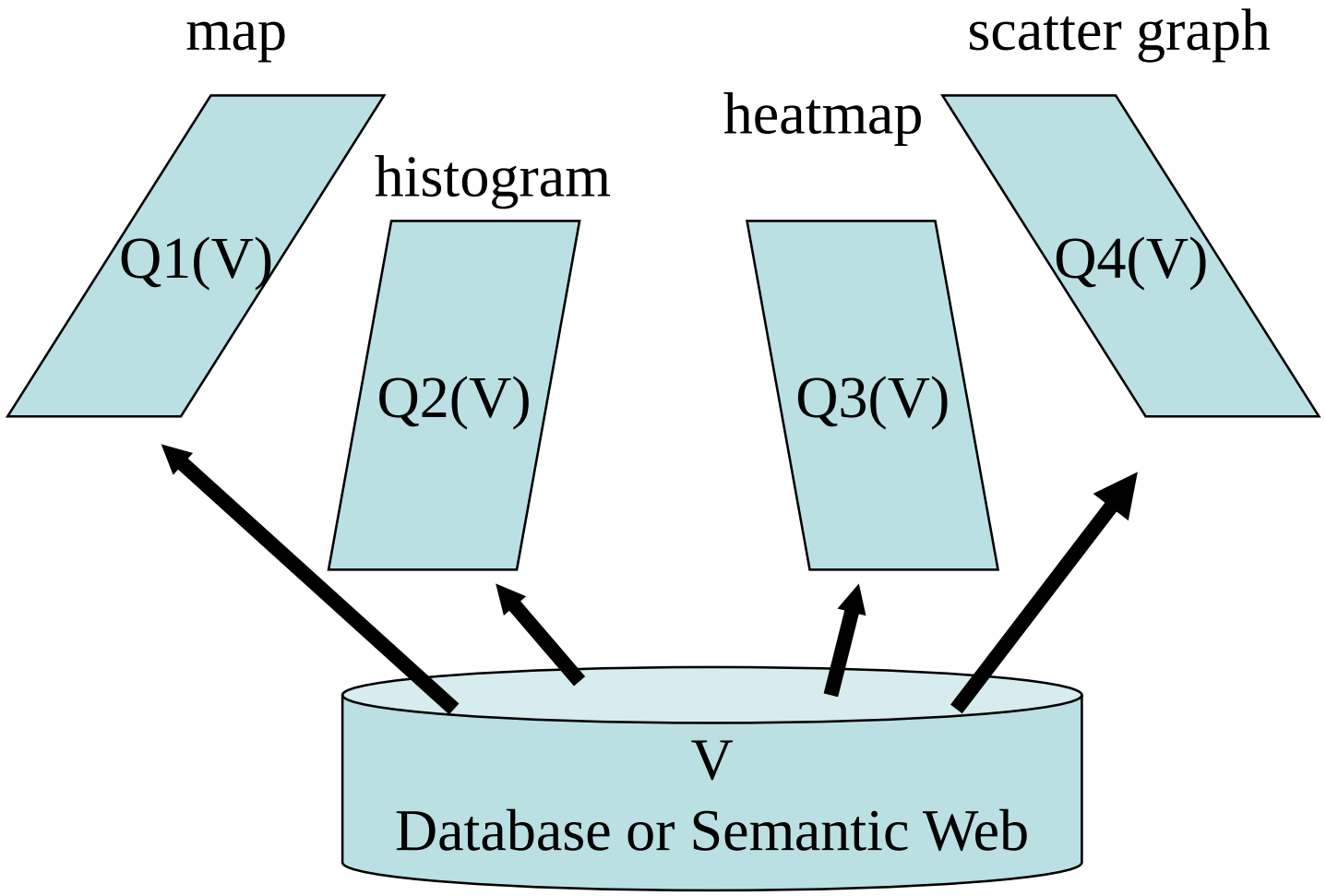
Interoperability among Data Segmentation and Micro Analysis

- Standard solution: Workflow
  - Data segmentation by a query → statistical analysis / data mining / visualization → decision making
  - One way from segmentation to analysis
  - Good for planned-for analysis scenarios
- In both cutting-edge scientific research processes and strategic planning of urban-scale social services, the finding of analysis scenarios itself is also a research goal, and this process is inherently exploratory and improvisational.

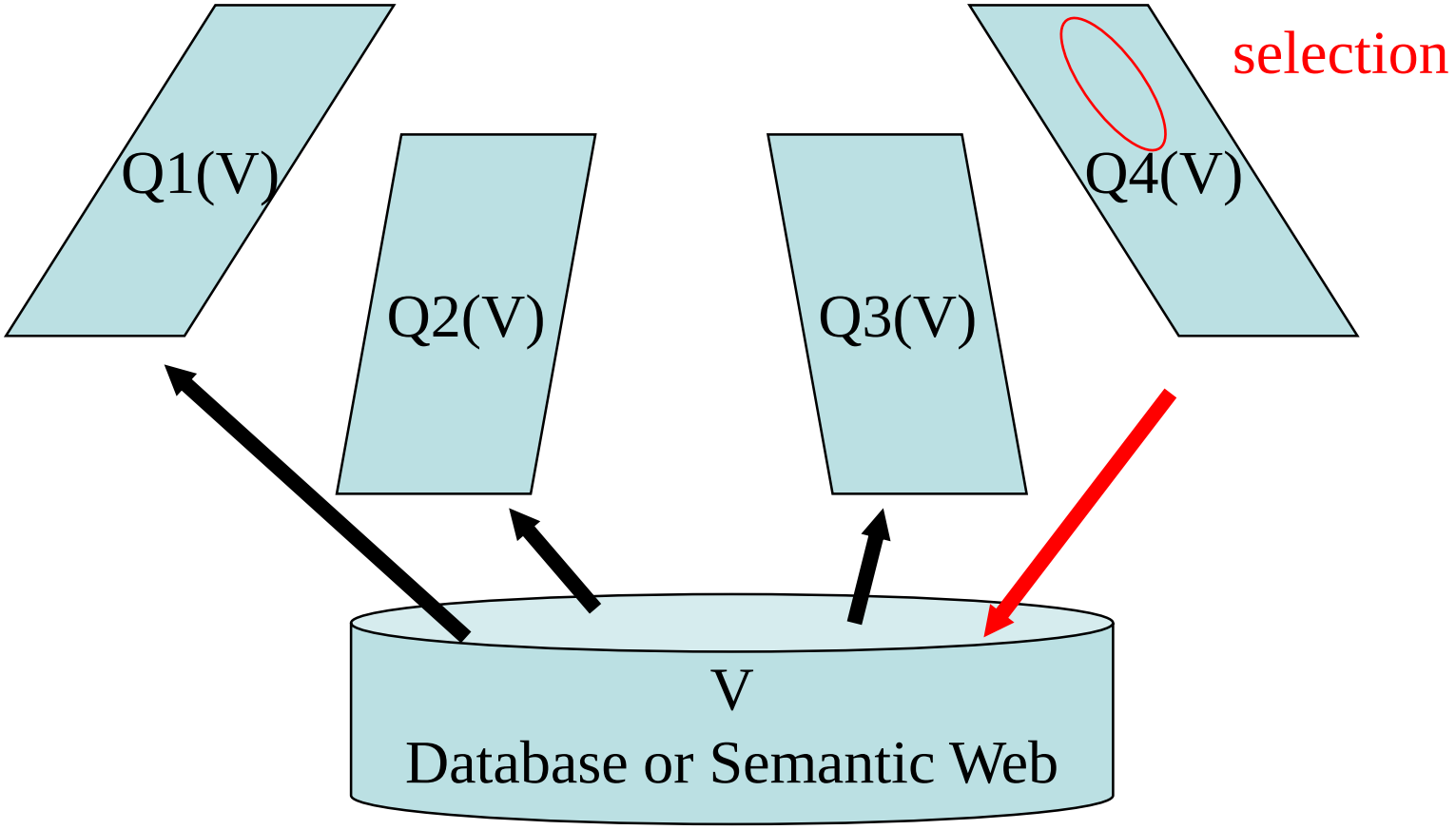
2 Types of Analysis Scenarios

- Planed-for analysis scenarios
  - use already-established analysis methods
  - mostly for routine analyses
- Exploratory and improvisational analysis scenarios
  - support the finding of analysis scenarios, which itself is also a research goal.

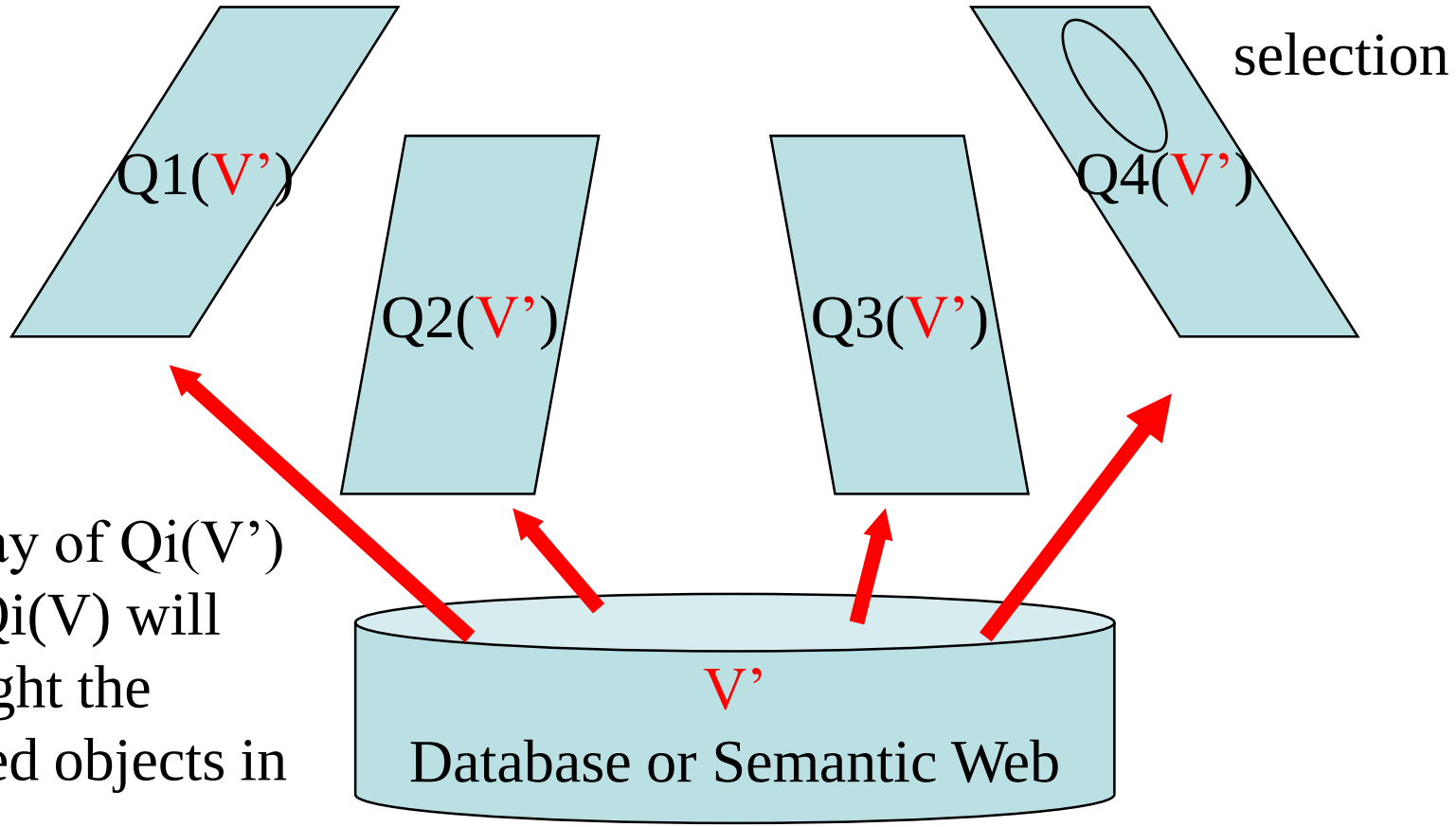
Coordinated Multiple Views as a Well-Known Framework for Exploratory Visual Analytics



Coordinated Multiple Views as a Well-Known Framework for Exploratory Visual Analytics

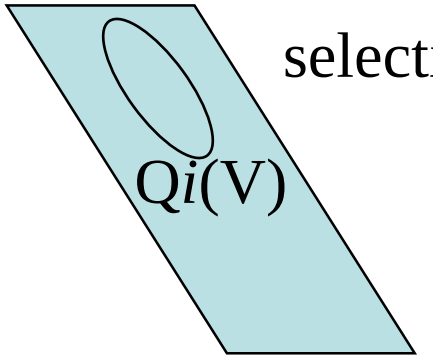


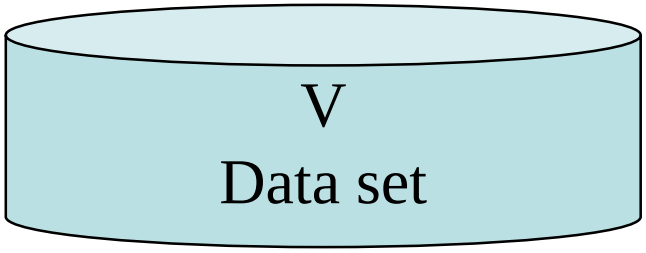
Coordinated Multiple Views as a Well-Known Framework for Exploratory Visual Analytics



Overlay of Qi(V’) with Qi(V) will highlight the selected objects in Qi(V).

Each visualization view





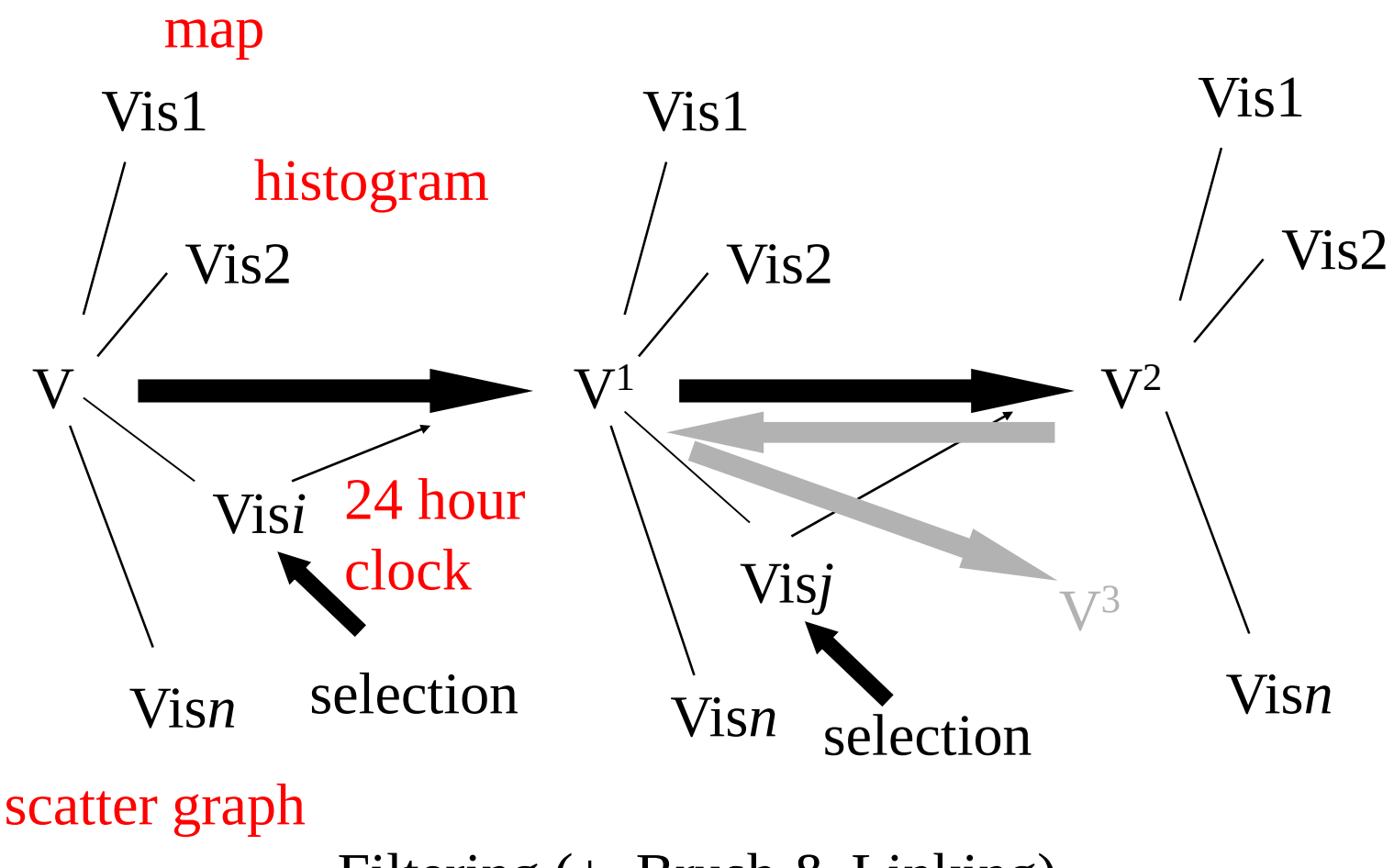
attr(Q*i*(V)):
list of (derived) attributes
cond(Q*i*(v)):
conditions in the where clause of Q*i*_(v), which should be true.
Direct selection is defined as a condition dsC(attr(Q*i*(V)) on the set of derived attributes attr(Q*i*(V))

This should further quantify V as

```
select  *
from V
where dsC(attr(Qi(V)))
```

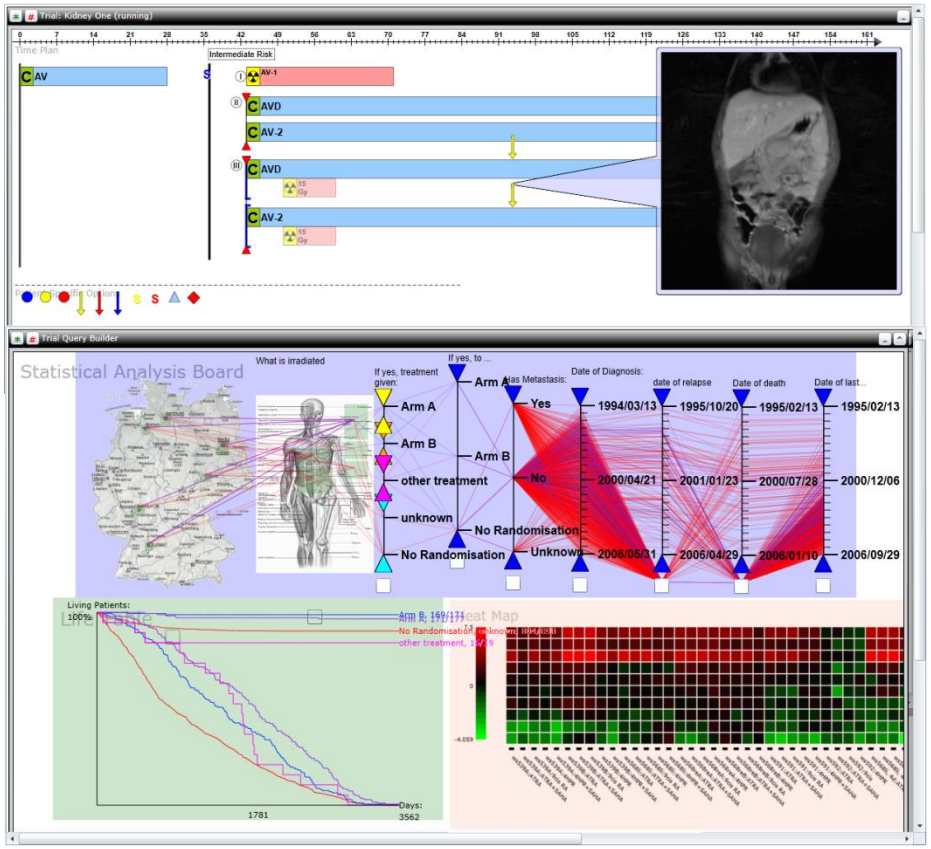
to make it V'

Exploratory Object Quantification

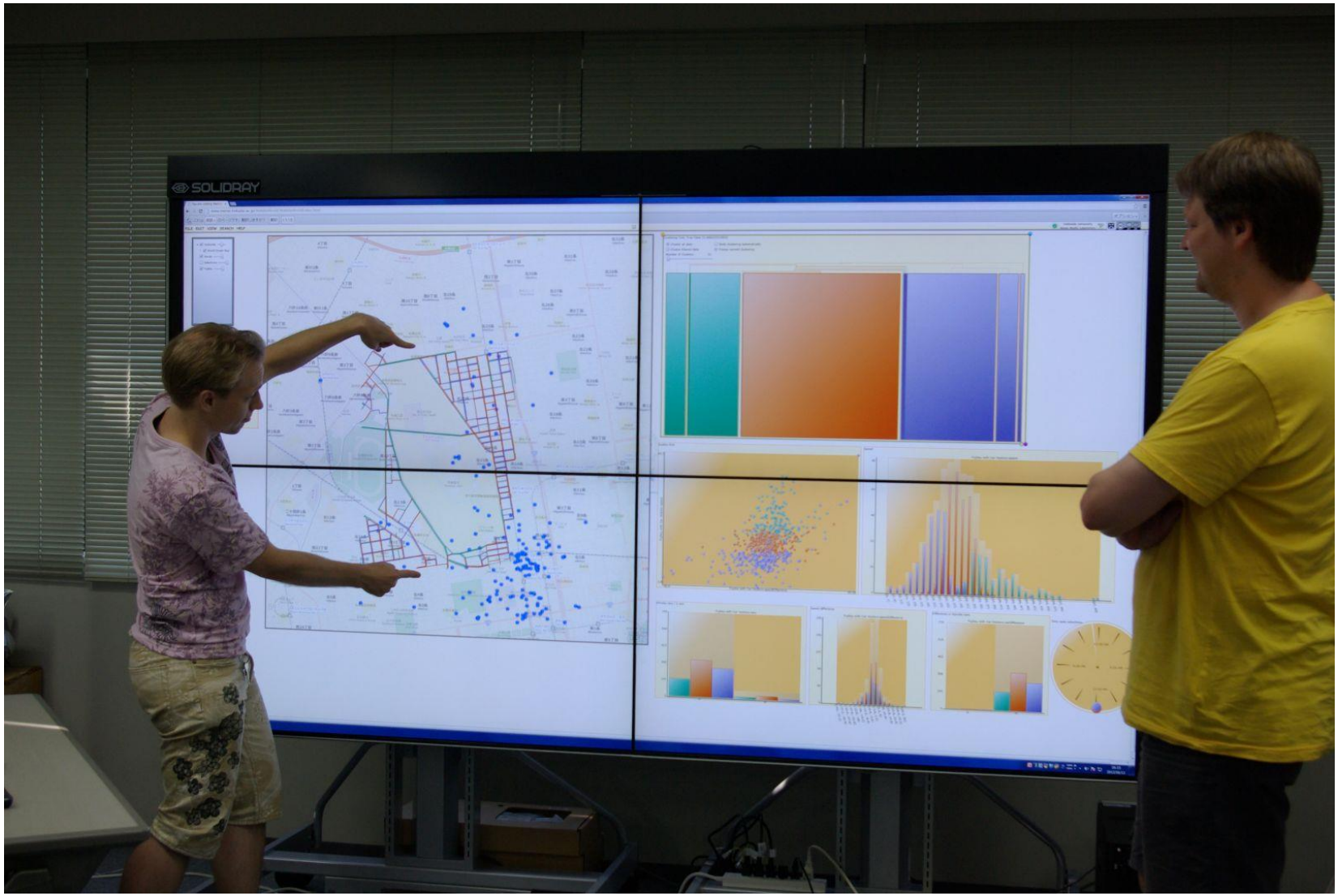


Filtering (+ Brush & Linking)

Query & Analysis View of TOB (Trial Outline Builder)

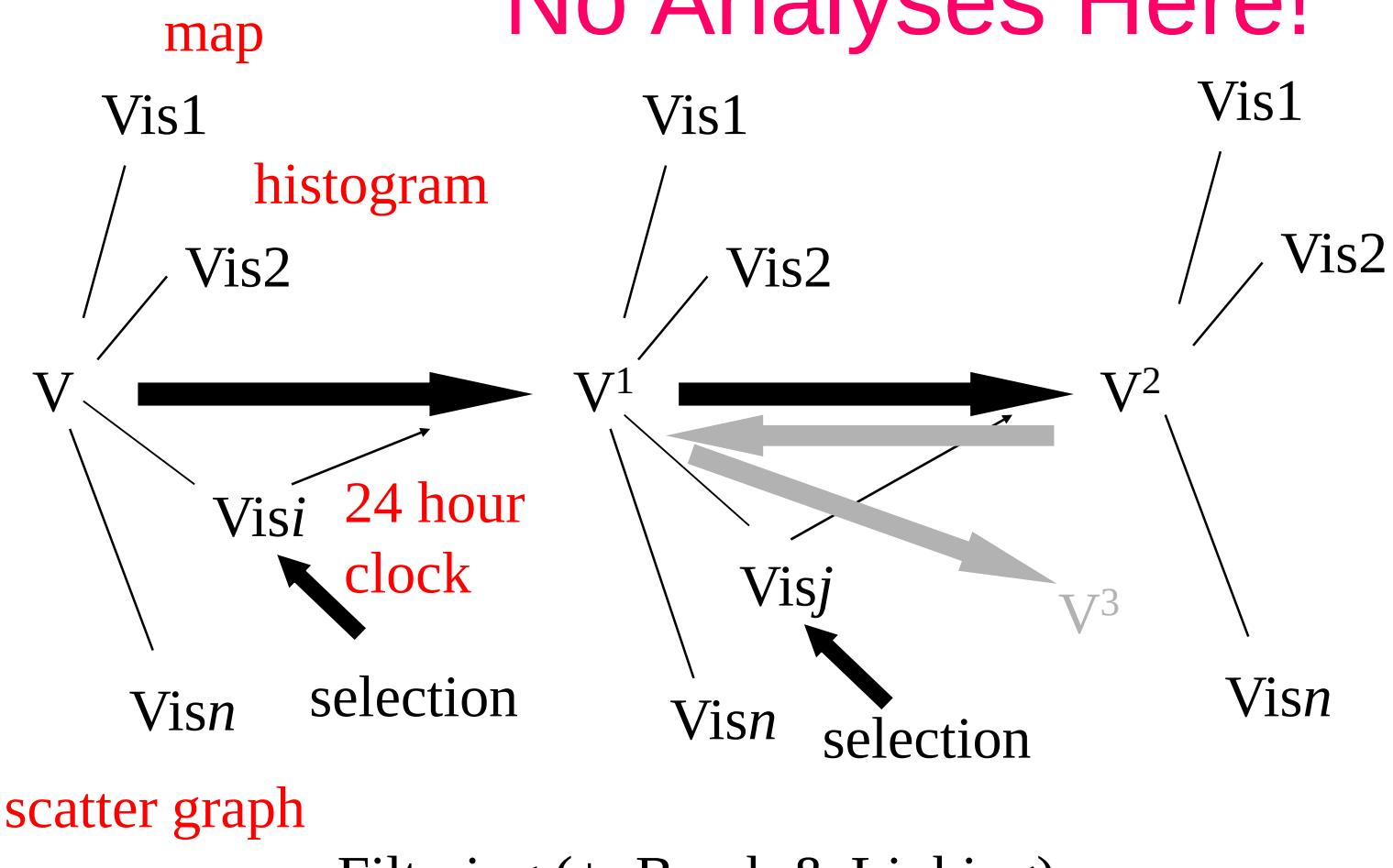


Geospatial Digital Dashboard for Exploratory Visual Analytics (2013)



Exploratory Visual Analytics with Coordinated Multiple Views

No Analyses Here!



Filtering (+ Brush & Linking)

How to introduce analyses and their visualizations?

- such as
  - –Clustering
  - –Pattern mining
  - –Statistical analysis
- into the coordinated multiple views framework

Analysis Results as Relations

- Clustering result:

  Cluster(Attr, ClusterID).

  – The values of Attr work as the object IDs of objects that are clustered.
- Mining result:

  Mining(Pattern, Supp (, Conf))

  Include(Attr, Pattern)

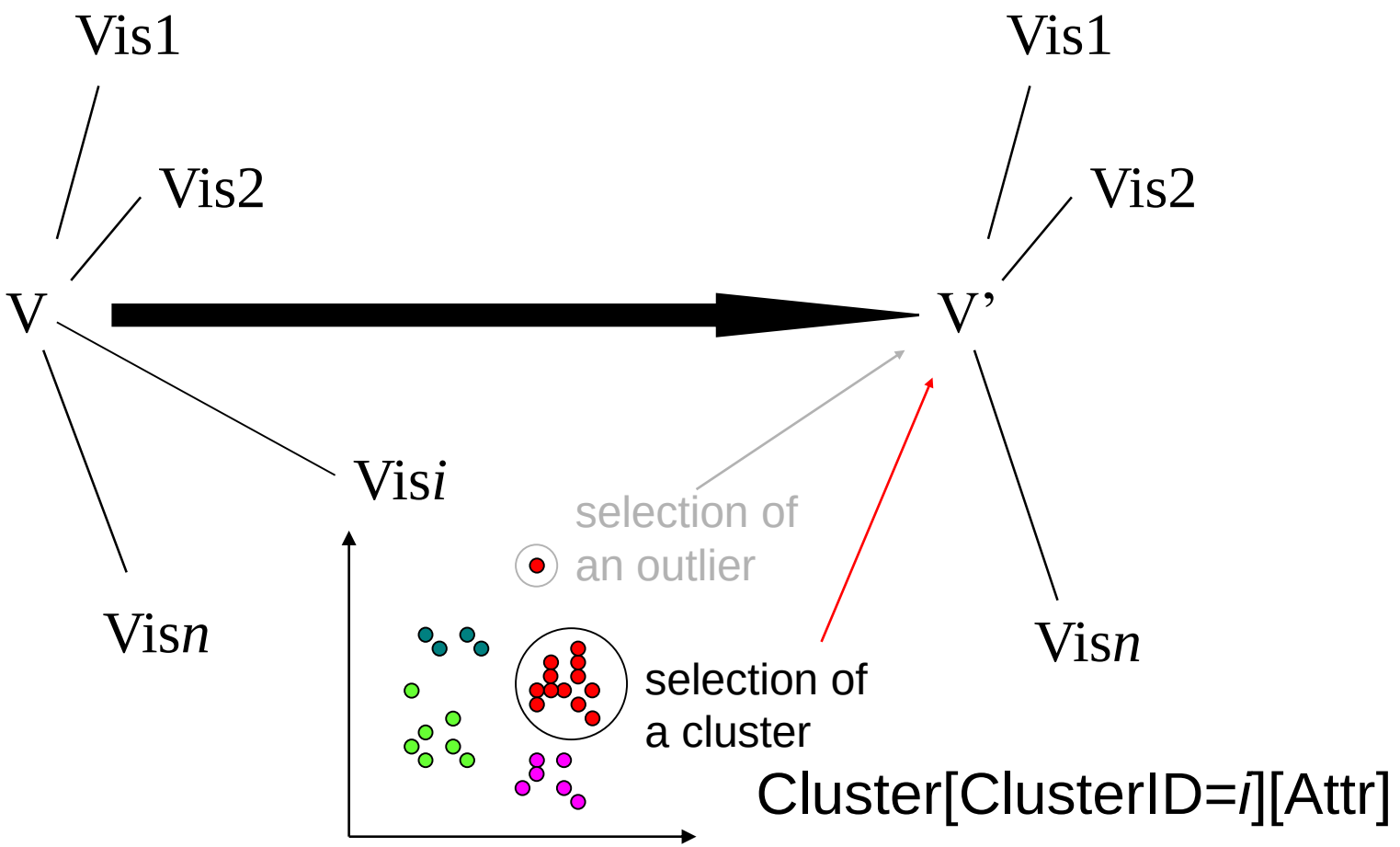
  – The second relation tells which objects among those identified by the attribute value of Attr include each mined pattern.
- Statistical analysis result:

  Stat(GBattributes, Afunction).

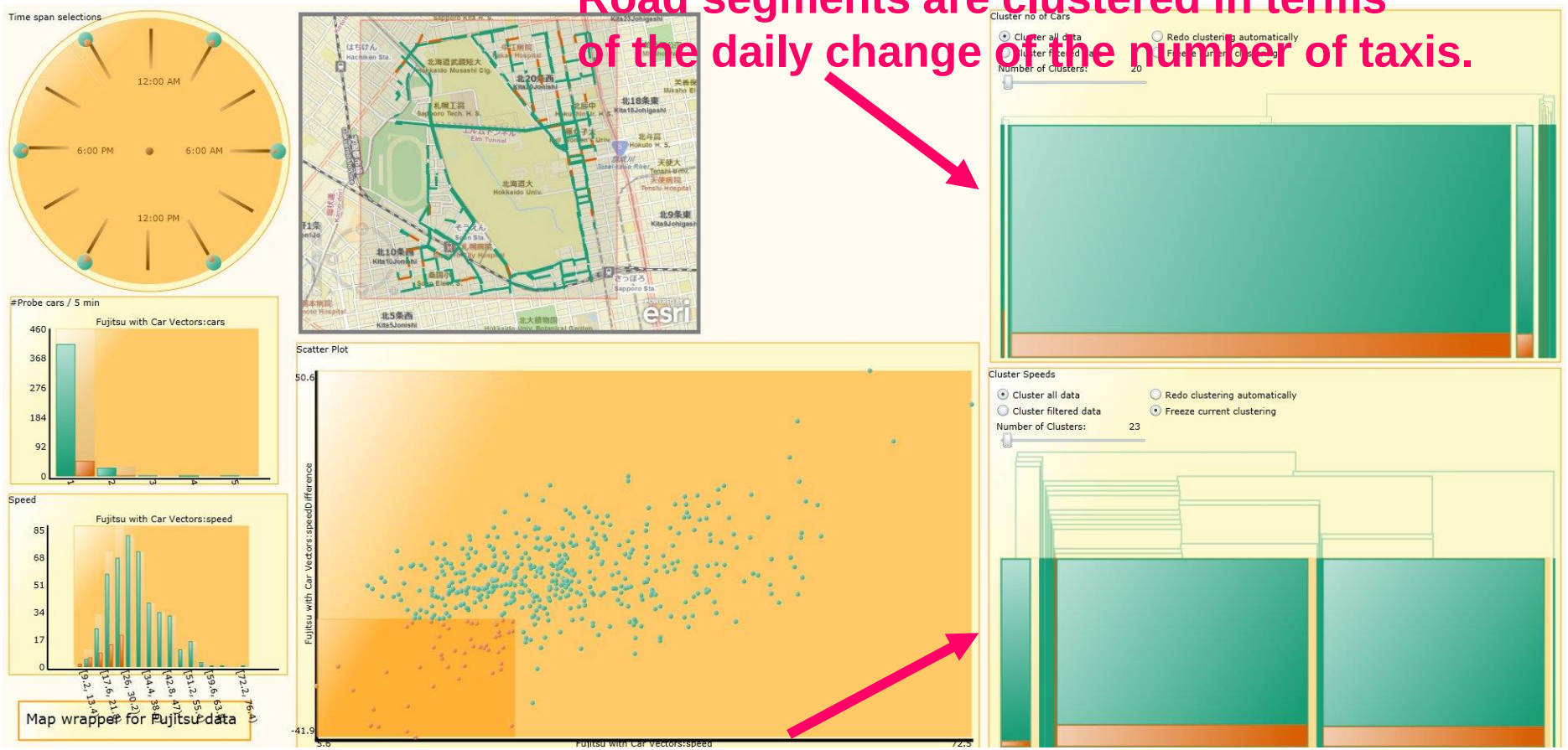
  – GBattributes : group-by attributes

  – Afunction : derived attribute calculated as the value of an aggregate function such as average, count, minimum, and maximum to the set of values of the specified attribute in each group

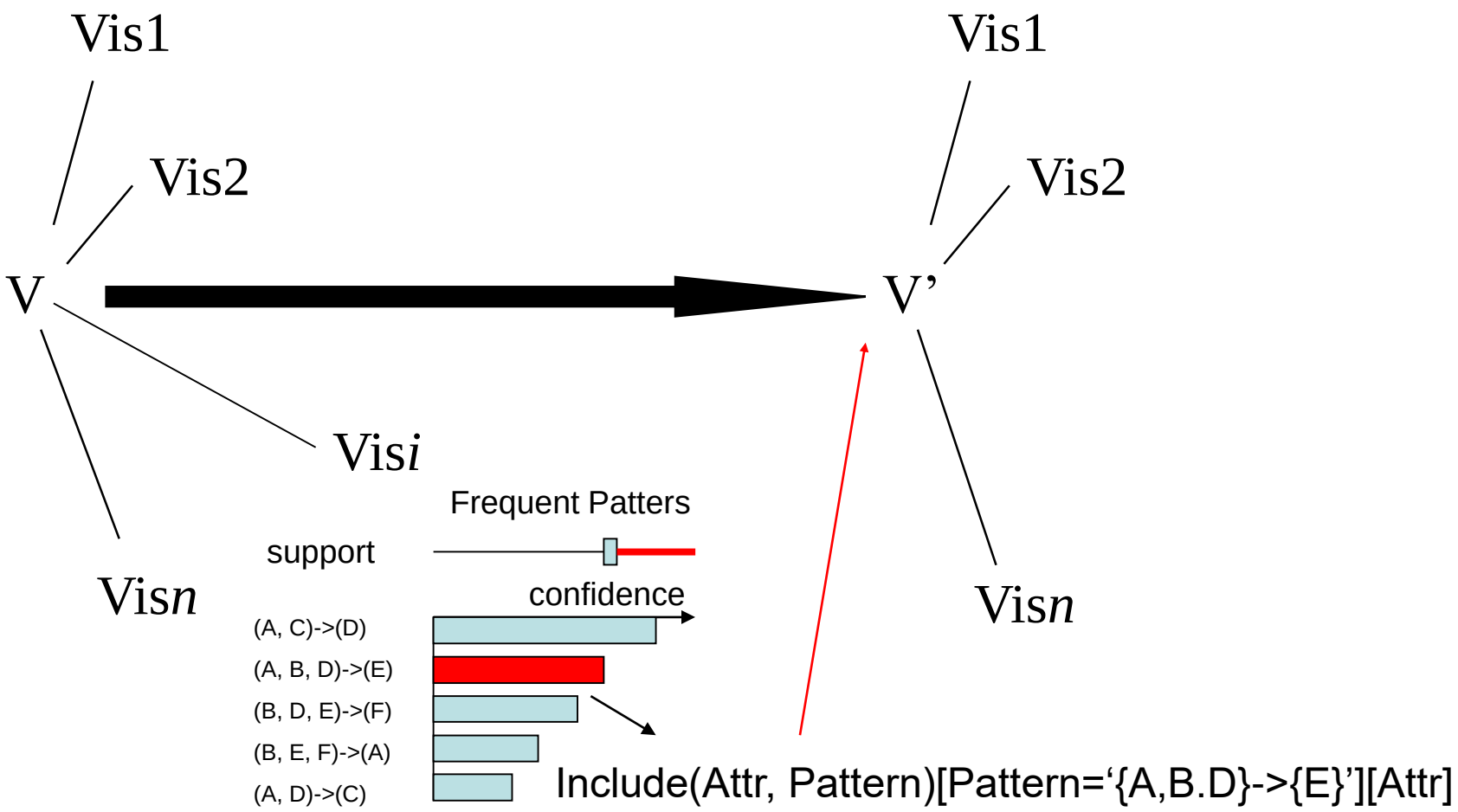
Exploratory Quantification and Analysis of Objects through a Clustering Result



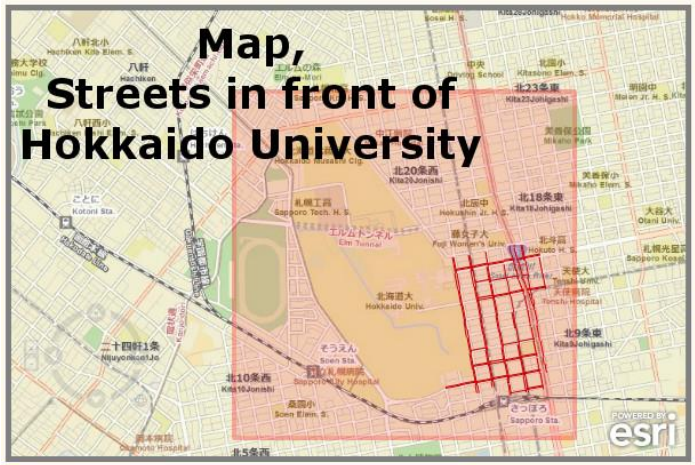
Integration of clustering tools into coordinated multiple visualizations in Geospatial Digital Dashboard (2013)

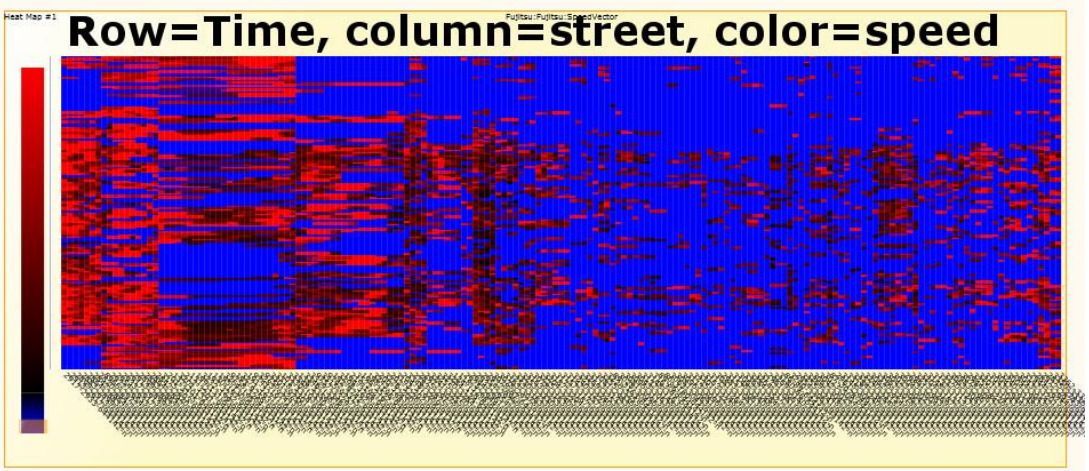


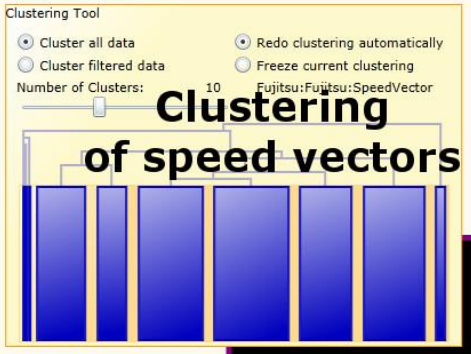
Exploratory Quantification and Analysis of Objects through a Mining Result



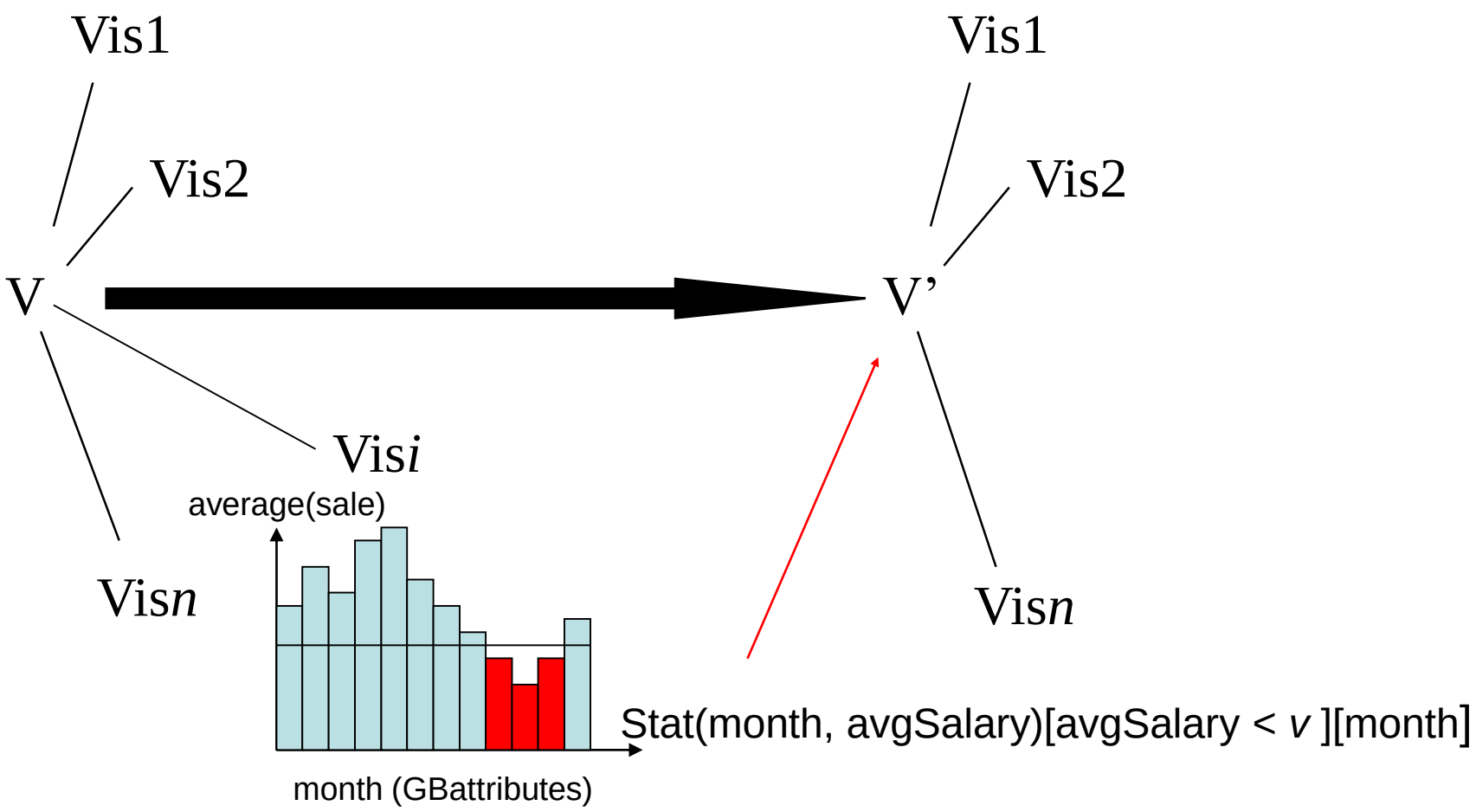
Integration of item set mining tools into coordinated multiple visualizations in Geospatial Digital Dashboard (2014)



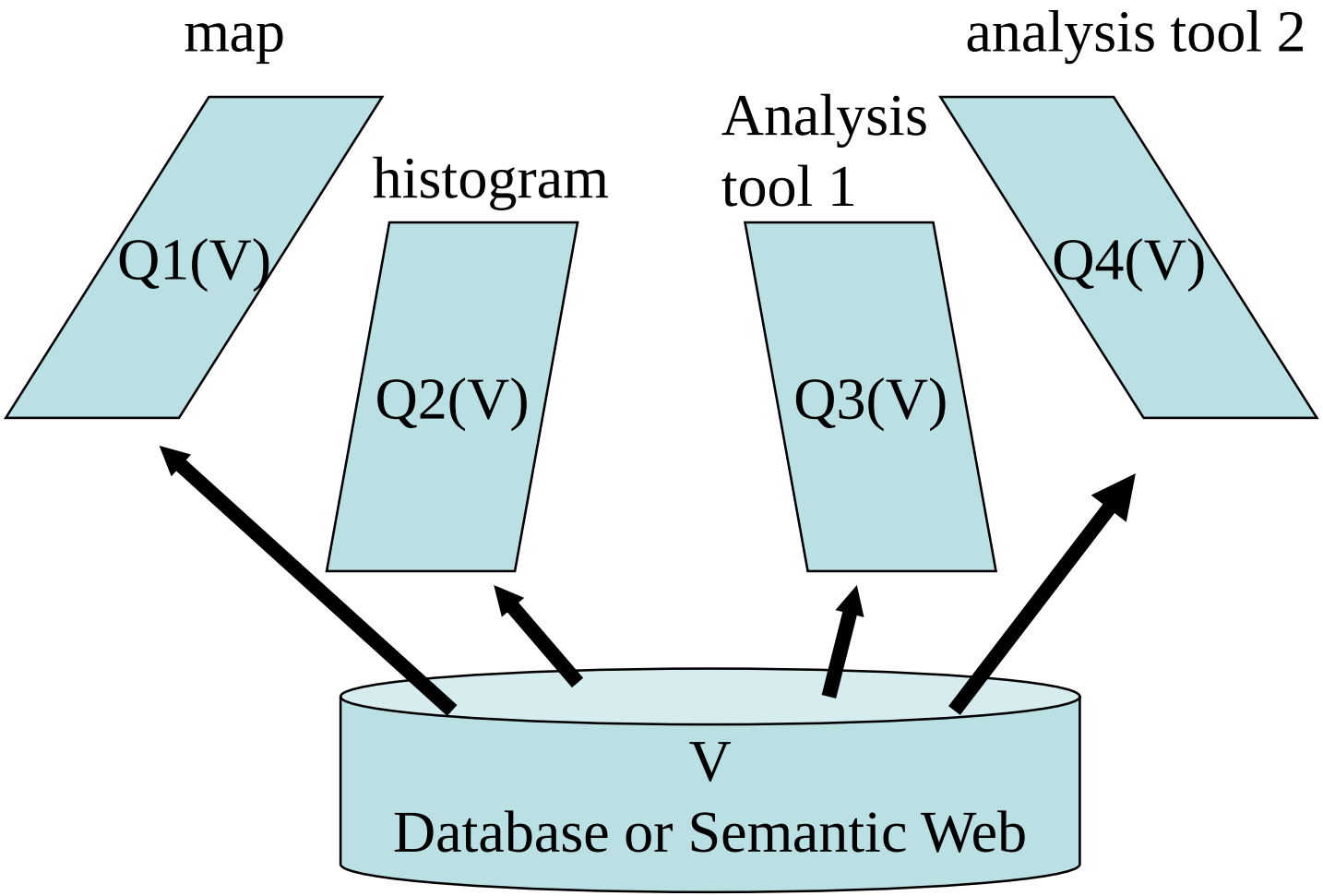




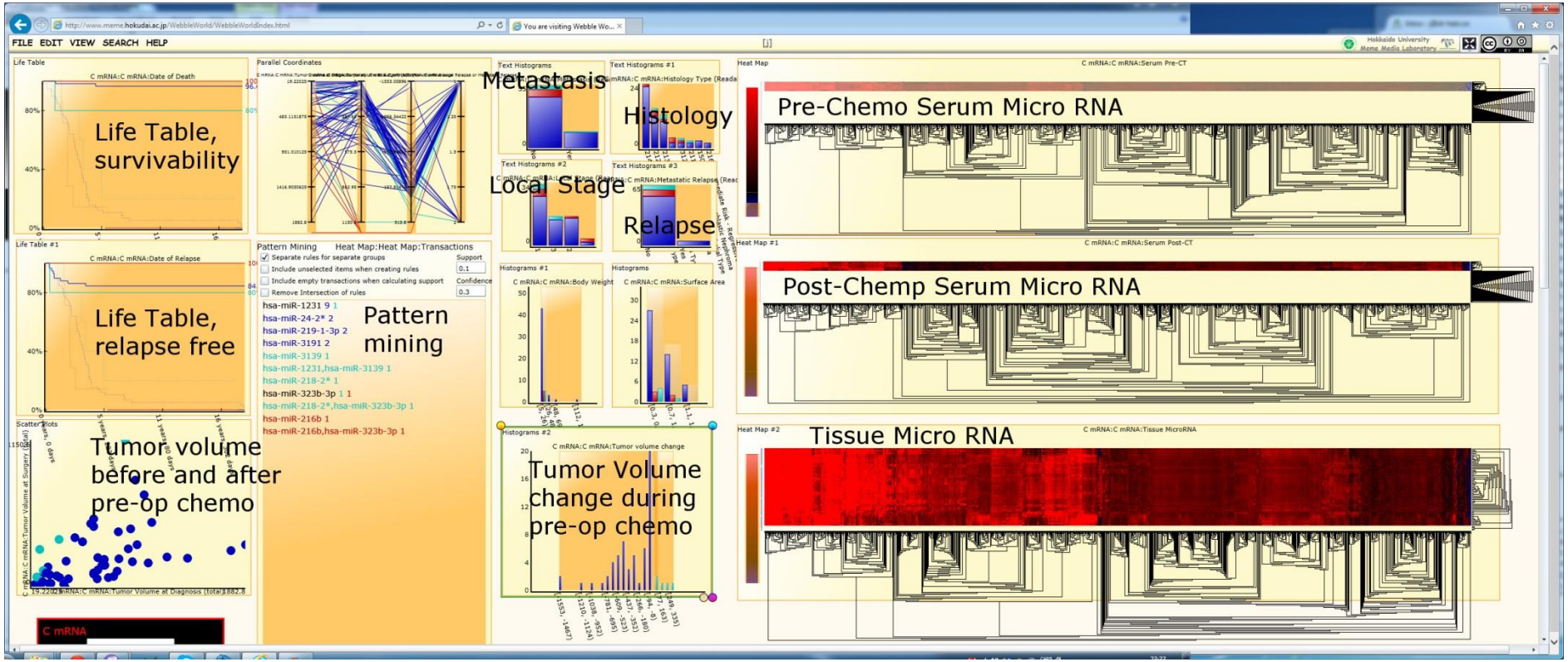
Exploratory Quantification and Analysis of Objects through a Statistical Chart



Coordinated Multiple Views and Analyses Framework



TOB for analyzing the Effect of Pre-op Chemotherapy



Each found pattern may work as a new biomarker to identify those patients who are helped or not helped by the preop chemotherapy

Our Goals of Trial Ouline Builder

- TOB for exploratory visual analytics of clinico-genomic trials on cancers
- To mine gene expression patterns for
  - Further segmentation of patients with respect to both phenotype and genotype characteristics
  - To find a patient group which shows meaningfully better recovery rate in one of the candidate treatment arms than the other arms.
- More patients are required to make the analysis on further segmented data still statistically significant.

新タイプの説明変数

- マニングやクラスタリングの結果見つかる頻出パターンやクラスタIDが、新しい説明変数となる場合がある！

My involvements in Big Data projects

- Cutting-Edge Data-Based Science or e-Science
  - EU FP projects for integrated IT support of clinical trials on cancer
    - FP6 Integrated Project ACGT (Advancing Clinico-Genomic Trials on Cancer) (02/2006 – 07/2010)
      - 26 teams
    - FP7 Large-scale Integration Project p-medicine (personalized medicine) (02/2011 – 01/2015)
      - 29 teams
- Urban Monitoring and Social Service Management
  - MEXT initiative project on Social CPS (Cyber-Physical System) for Efficient Social Services (09/2012-03/2017)
    - Project Consortium (NII (National Institute of Informatics), Hokkaido Univ., Osaka Univ., Kyushu Univ.)
- Program Officer of the JST CREST Program on Big Data Applications (2013-2020)
- Collaboration with Dr. Keisuke Takahashi in Material Informatics (2014-)

Toward Material Informatics

Collaboration with Dr. Keisuke Takahashi (2014 -)

Material synthesis and design from first principle calculation and machine learning

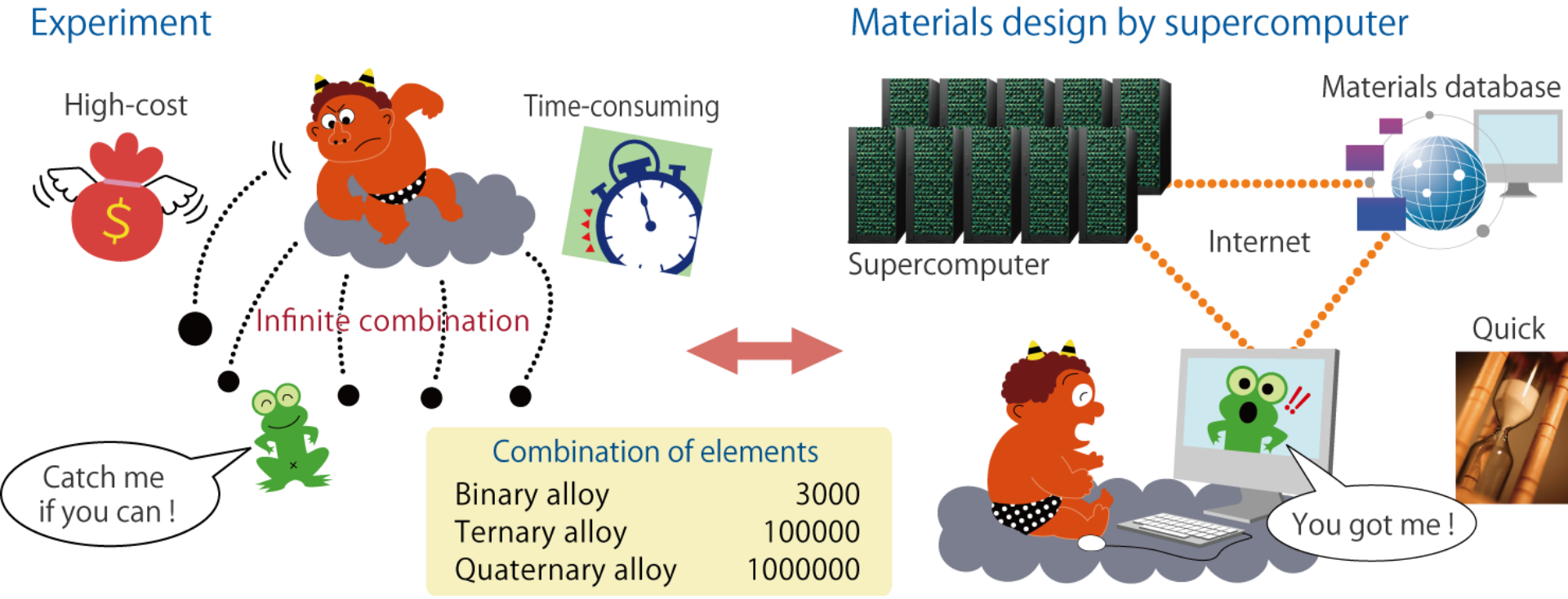
Keisuke Takahashi*
Graduate School of Engineering, Hokkaido University, N-13, W-8, Sapporo 060-8278, Japan

Yuzuru Tanaka
Meme Media Laboratory, Hokkaido University, N-13, W-8, Sapporo 060-8278, Japan
(Dated: June 10, 2015)

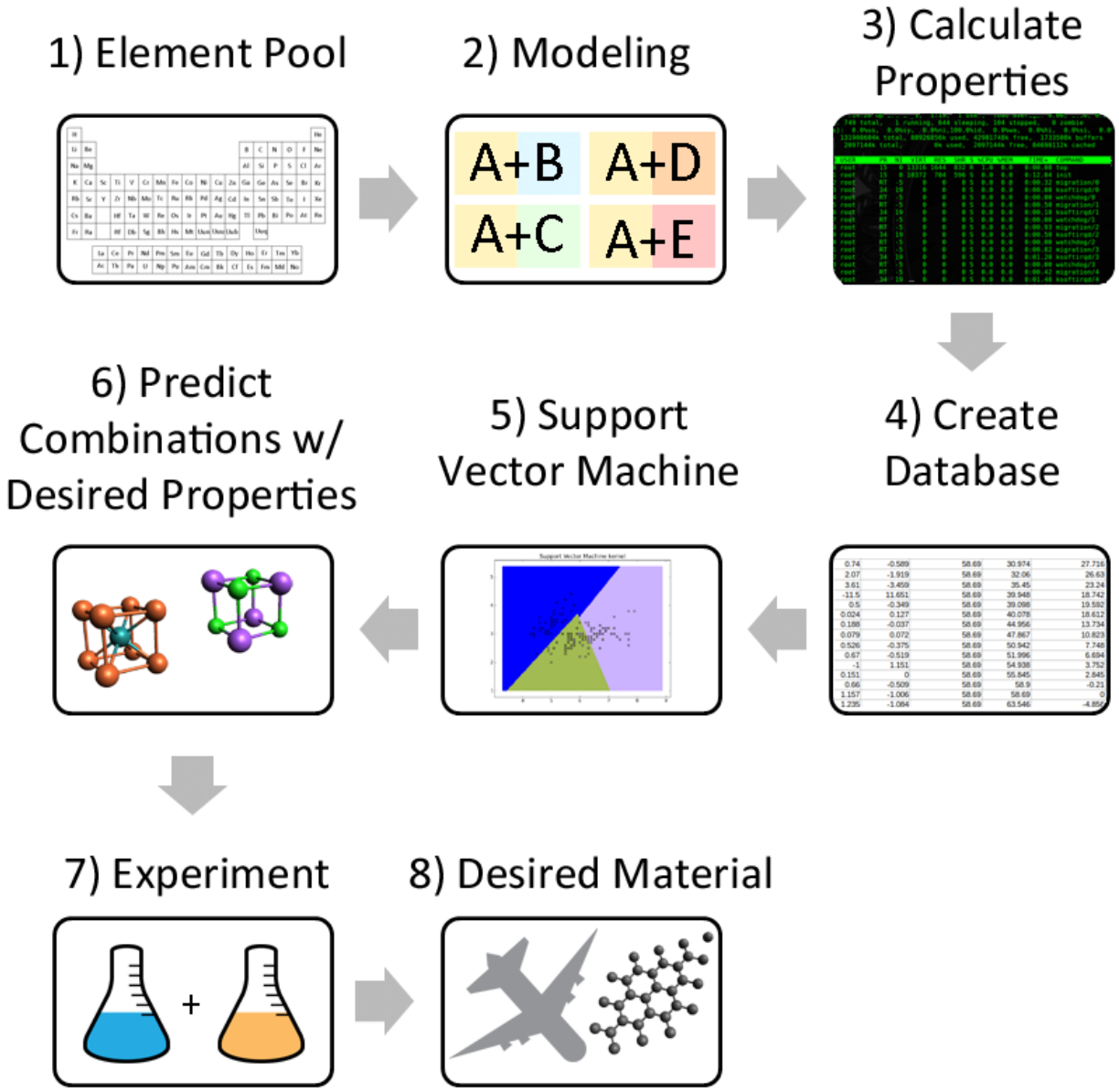
Desired material synthesis and design can be predicted on the basis of first principle calculation and machine learning. Material big data is constructed based on density functional theory and database is then trained using the support vector machine. The predicted material properties are comparative to experimental material properties. The proposed workflow become the bridge between the material database and designing materials. The approach enable the efficient material mining from big database and reveal the undiscovered desired material.

Keisuke Takahashi, Yuzuru Tanaka, ***Computational Materials Science*** (2016) 364-367 (In Press)

Why Material Informatics and Computational Material Science?



Proposed work-flow



Keisuke Takahashi, Yuzuru Tanaka, ***Computational Materials Science*** (2016) 364-367 (In Press)

Designing Explanatory Variables

- Primitive explanatory variables
  - defined as measurable features or simulation parameters.
- Derived explanatory variables
  - defines as functions of some primitive explanatory variables
- Marker variables
  - Defined as cluster ids or pattern ids

Primitive Explanatory Variables

- Measurable features
  - e.g., average velocity $v(t)$, number of cars $n(t)$, and length of each road link l
- Simulation parameters
  - of the whole system
    - e.g., temperature, atmospheric pressure, wind direction and speed, precipitation in meteorological simulation
  - of each aspect modeling
    - e.g., energy increase for reduced lattice length (aspect: solidness)

Derived Explanatory Variables

- e.g., traffic flow of each road link:

  $\rho(t) = v(t)\, n(t)/l$

- Depending on what analysis method you use, some type of derived variables are already implicitly considered as explanatory variables.
  - e.g., linear combination of higher order terms of original explanatory variable in case of using SVM

Marker Variables

- Clustering result
  - Each cluster id may work as an explanatory variable of further segmentation and analysis
- Mining result
  - Each frequent pattern id may work as an explanatory variable of further segmentation and analysis
    - e.g., mined miRNA expression pattern in preop chemotherapy

What are required in big data applications?

- Collaboration between open minded researchers, one from CS and the other from domain science.
- One from CS should guide the systematization of big data approach.
  - A total architecture with its compatible platform technologies for exploratory visual analytics to discover both analysis scenarios and new knowledge.
- One from the domain science should either mathematically model the target or define a set of appropriate indices to describe the mesoscopic model of the target.

重要なのは？

分析アルゴリズム ＞ データ　No!

データ ＞ 分析アルゴリズム　No!

データ ＞ モデリング ＞ 分析アルゴリズム　No!

モデリング ＞ データ ＞ 分析シナリオ
(アスペクト)

我が国独自の問題

- 殆どのアルゴリズムの研究者はシーズ・ドリブン
  - 欧米ではニーズ・ドリブン、ミッション・ドリブンで新しいアルゴリズムや統計学が創生されてきた。
  - 短期研究成果評価方式がこの傾向を一層促進
- 対象のメゾスコピック・レベルの現象の数学モデリングが行える人材が少ない。
  - アナリシスのための数学教育が中心で、シンセシスのための数学教育が不十分