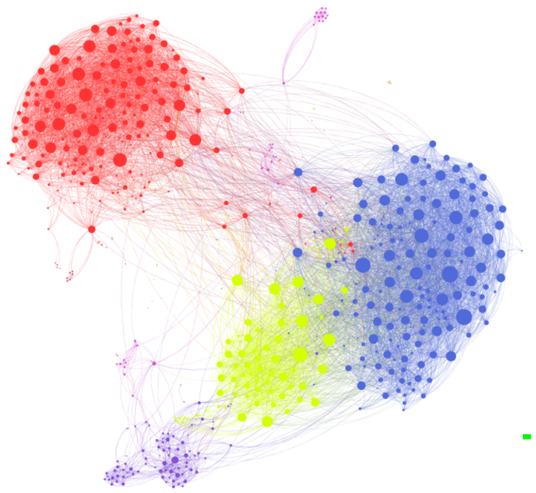
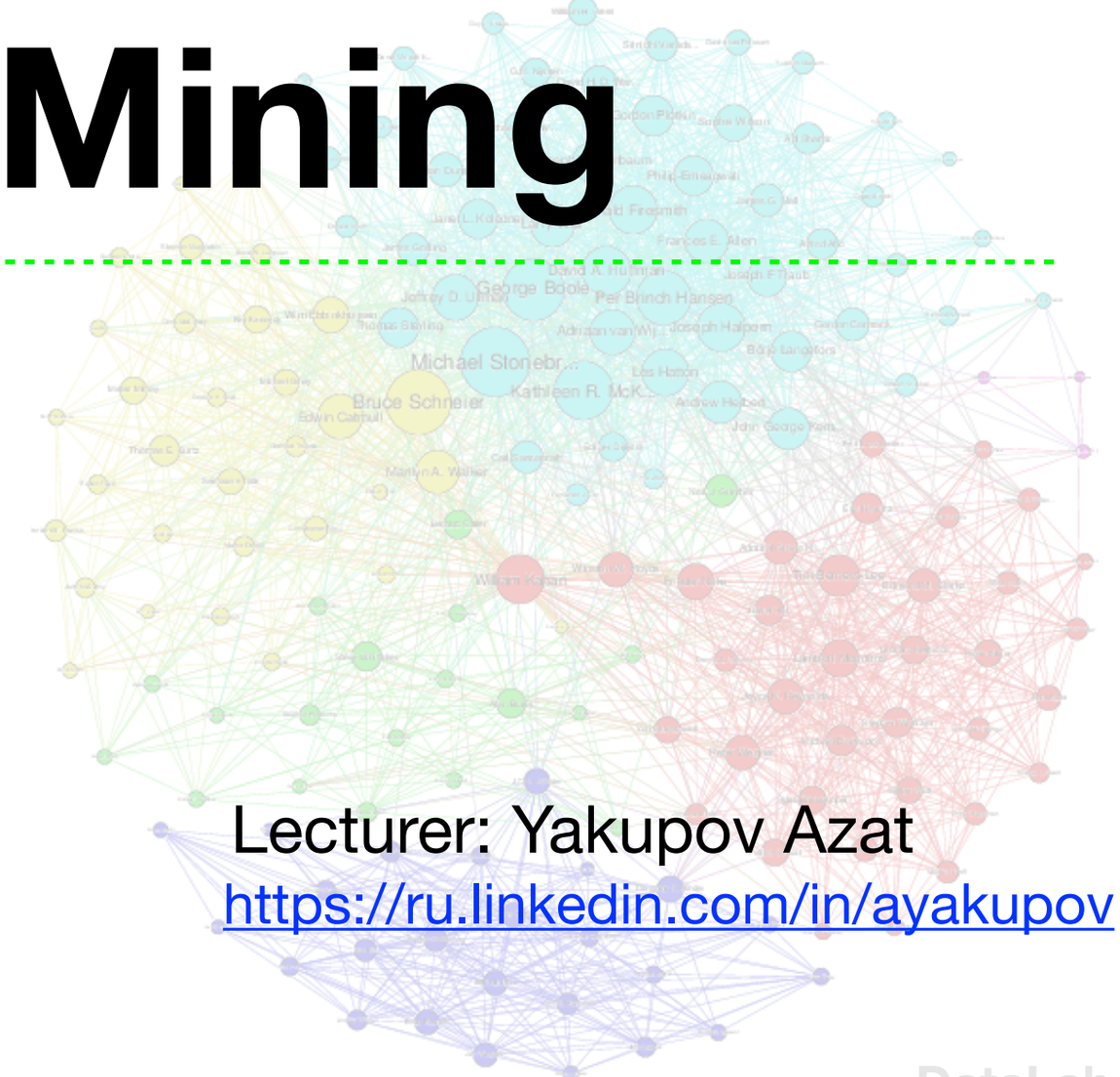


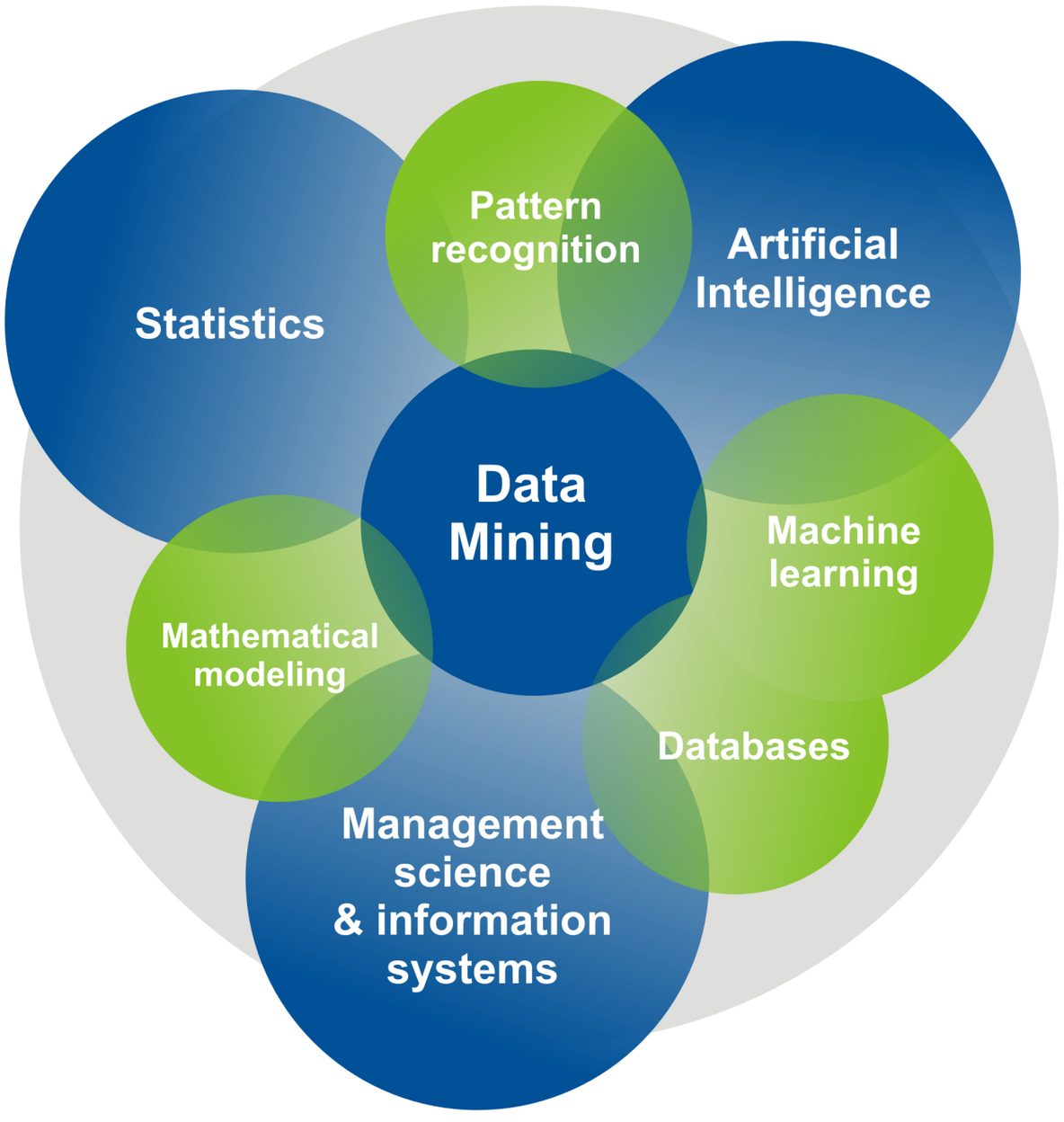
Data Mining



Lecturer: Yakupov Azat

https://ru.linkedin.com/in/ayakupov

Introduction



“Education is not pilling on of learning, information, data, facts, skills, or abilities - that’s training or instruction - but is rather making visible what is hidden as a seed.”

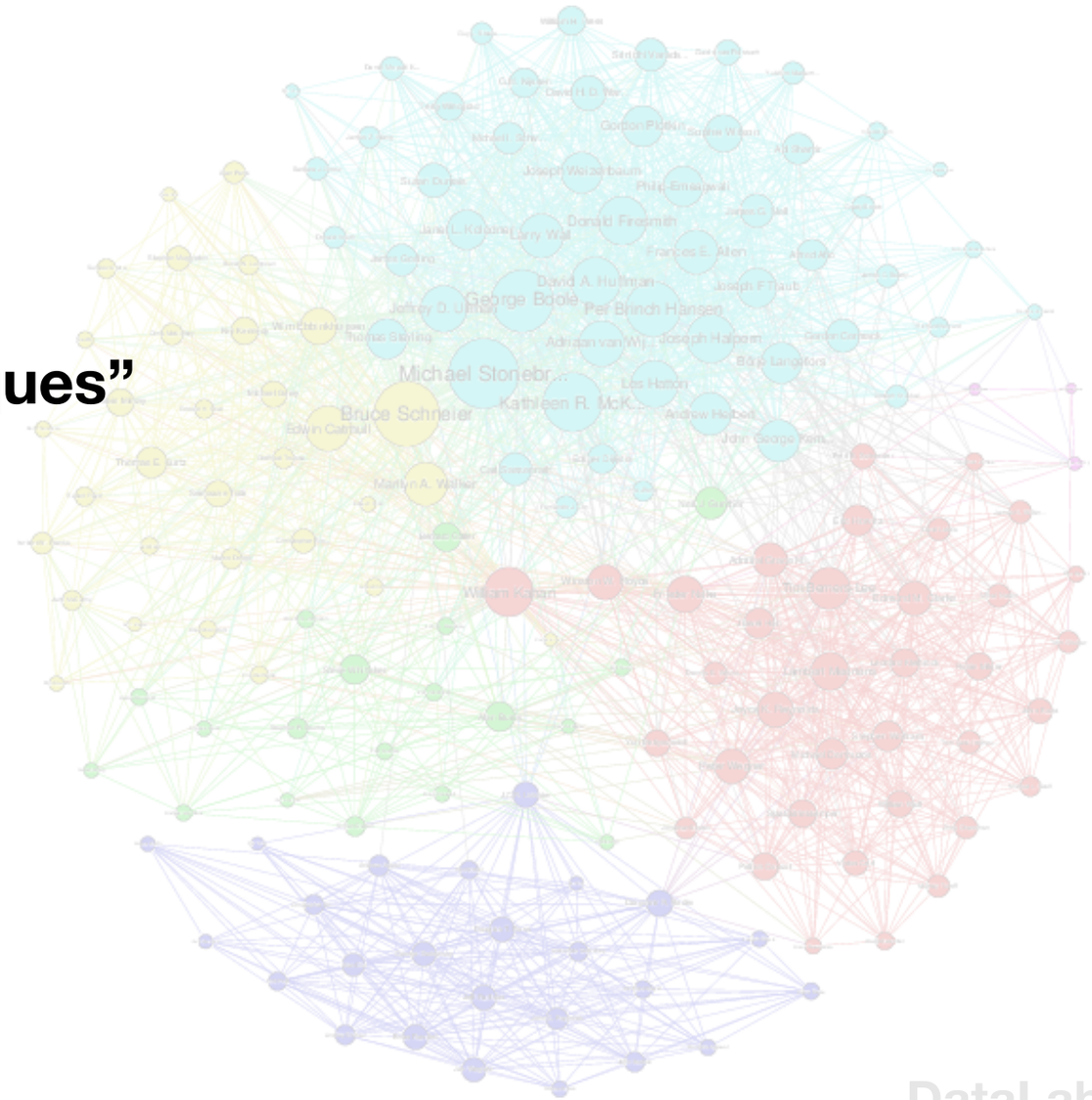
Thomas More

Literature

“Data Mining”
Charu C. Aggarwal

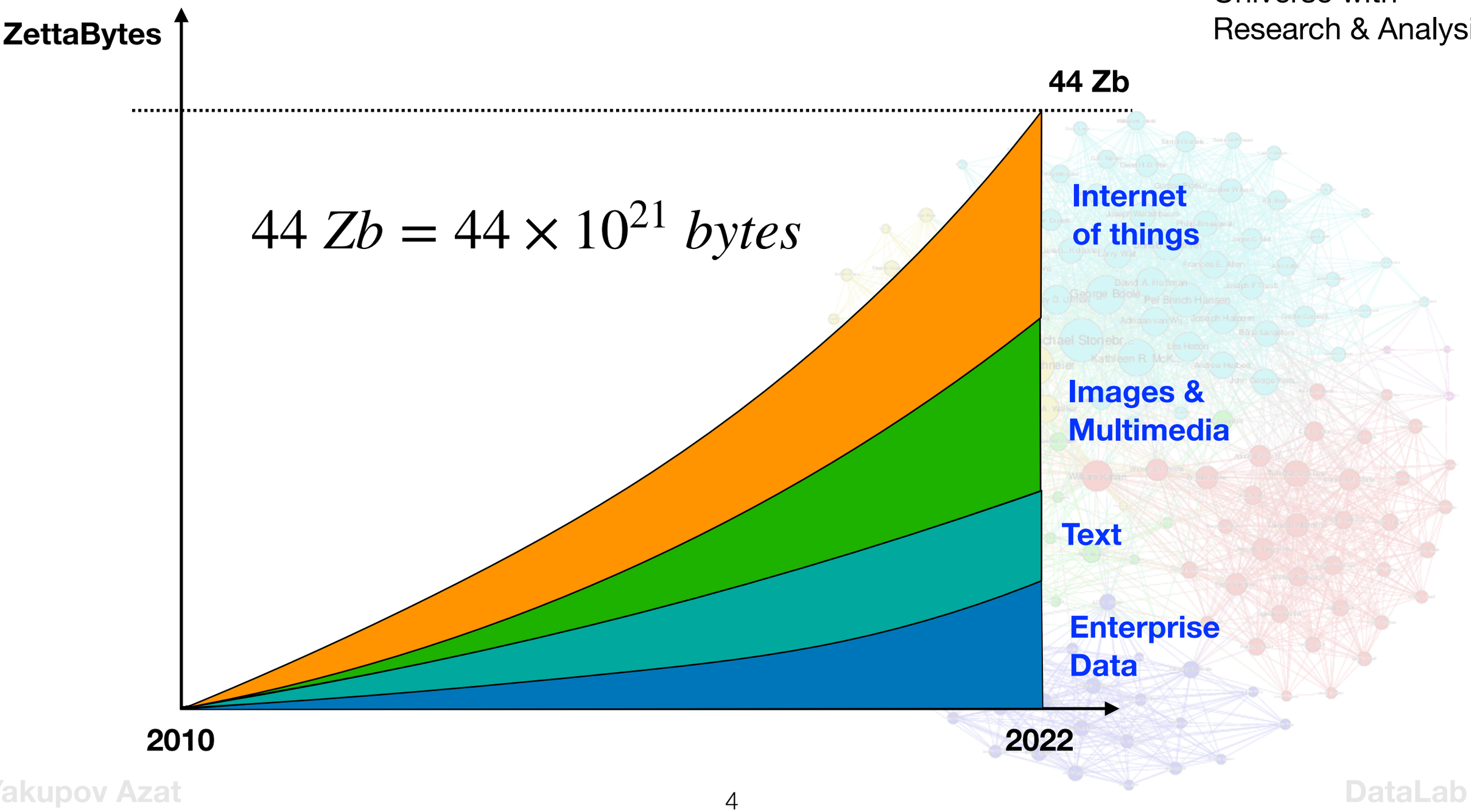
“Outlier Detection”
Charu C. Aggarwal

“Data Mining. Concepts and Techniques”
Jiawei Han, Micheline Kamber, Jian Pei



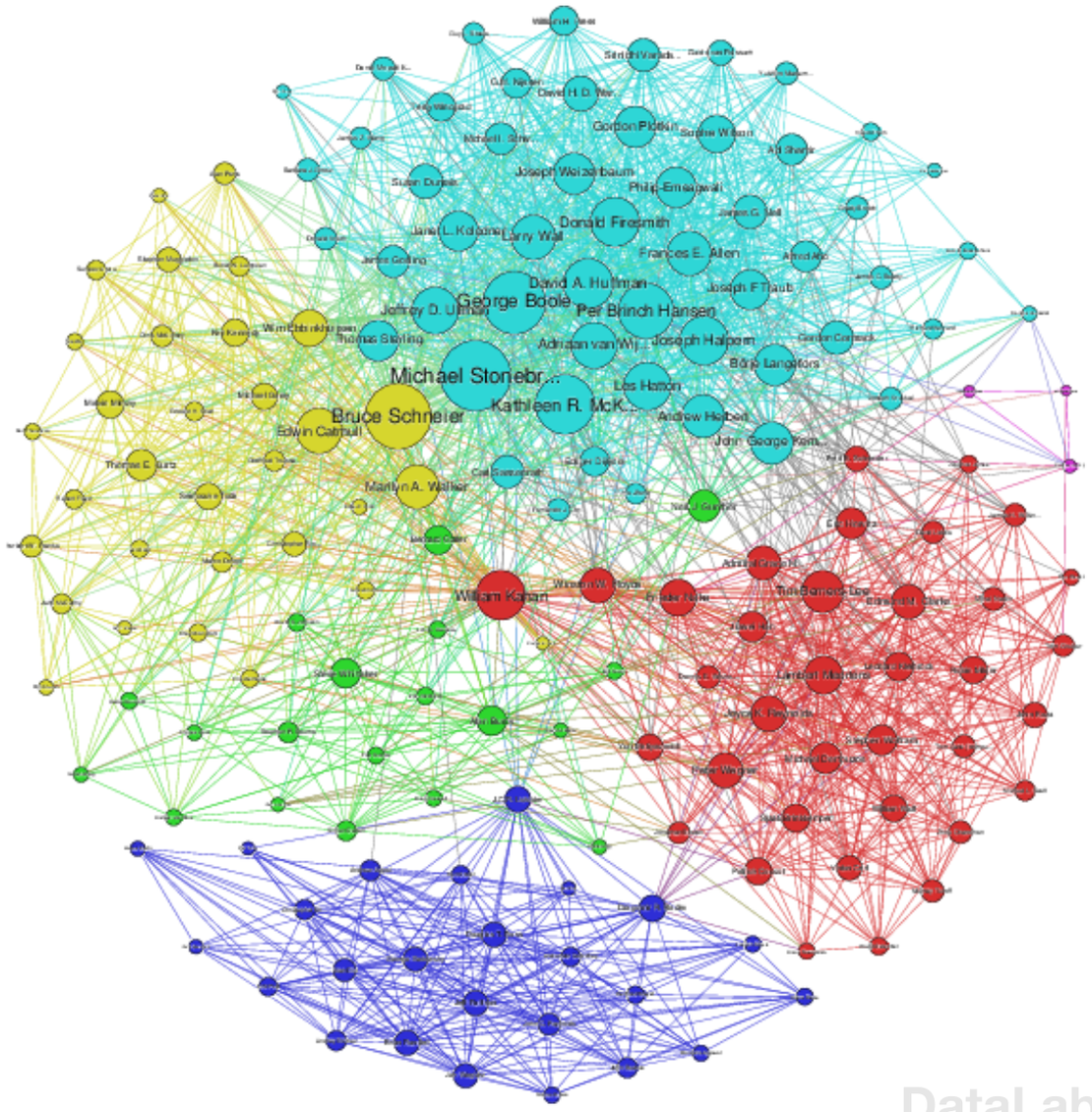
Data Mining

by EMC Digital Universe with Research & Analysis



Data Mining

Data Mining = **K**nowledge **D**iscovery from **D**ata (**KDD**)



Evolution of Databases

Data Collection and Database Creation
(1960s and earlier)

Primitive file processing

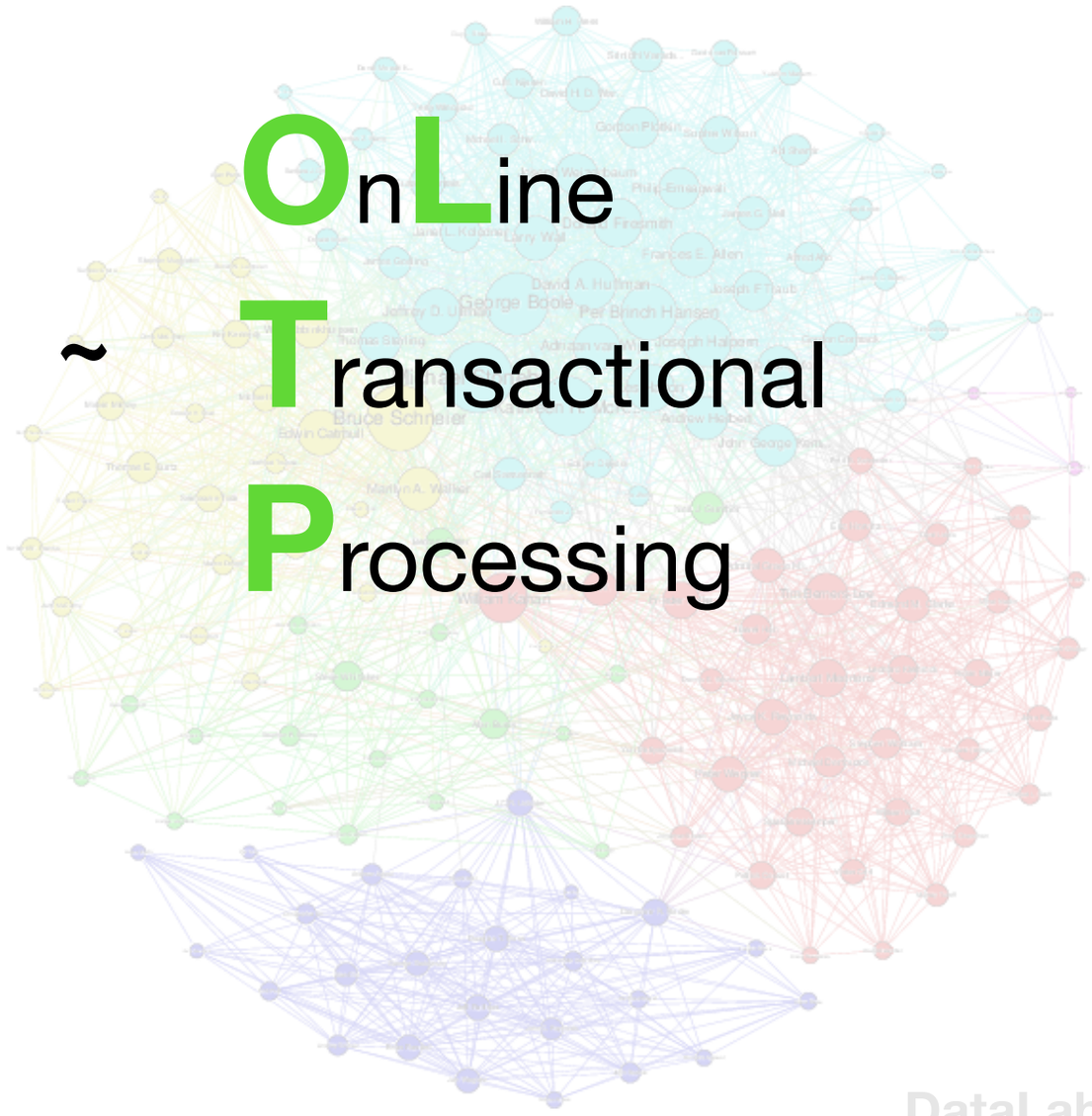
Database Management Systems
(1970s to early 1980s)

- Hierarchical and network database systems
- Relational database systems
- Data modeling: entity-relationship models, etc.
- Indexing and accessing methods
- Query languages: SQL, etc.
- User interfaces, forms, and reports
- Query processing and optimization
- Transactions, concurrency control, and recovery
- Online transaction processing (OLTP)

Evolution of Databases

Advanced Database Systems
(mid-1980s to present)

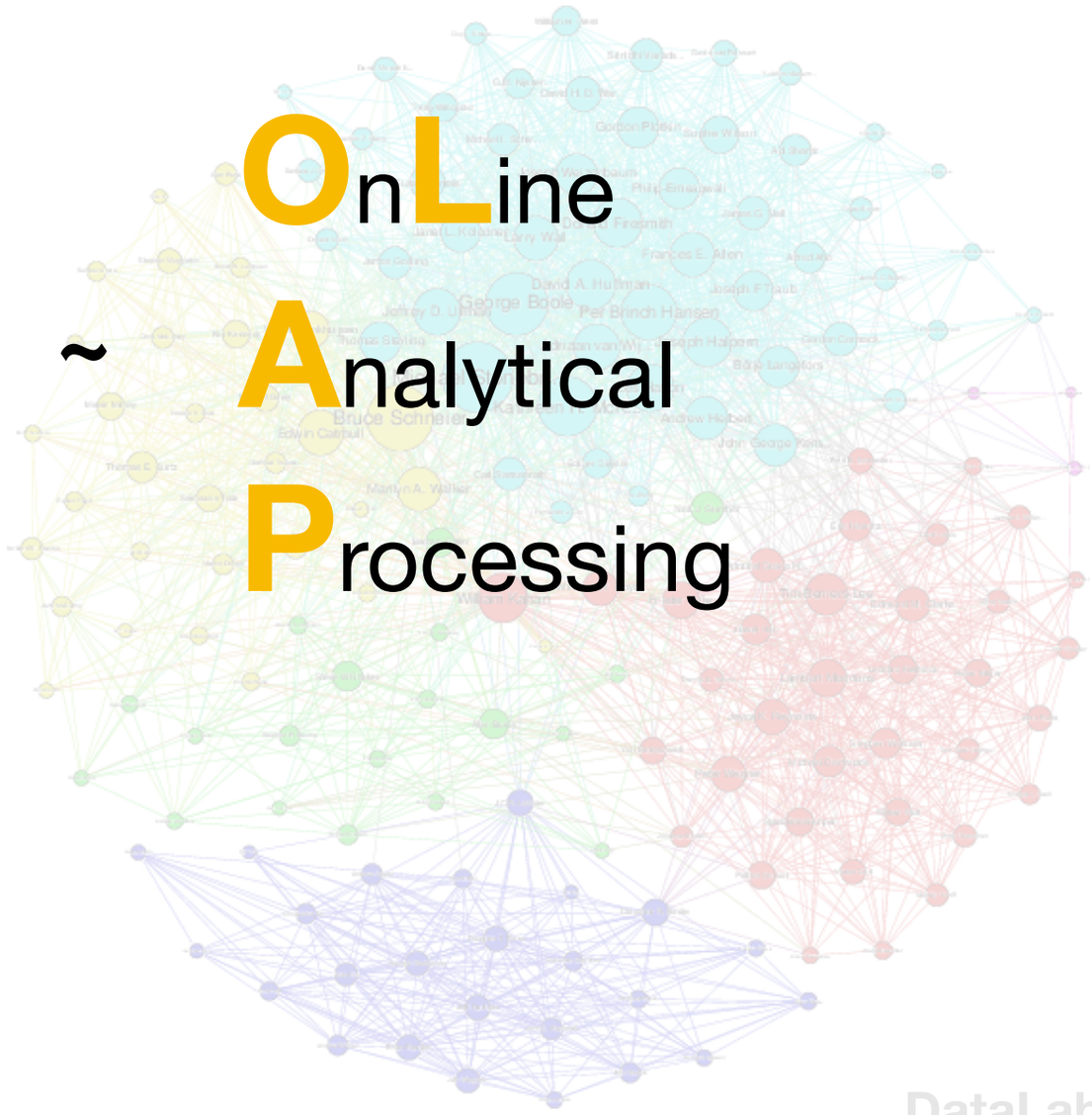
- Advanced data models: extended-relational, object relational, deductive, etc.
- Managing complex data: spatial, temporal, multimedia, sequence and structured, scientific, engineering, moving objects, etc.
- Data streams and cyber-physical data systems
- Web-based databases (XML, semantic web)
- Managing uncertain data and data cleaning
- Integration of heterogeneous sources
- Text database systems and integration with information retrieval
- Extremely large data management
- Database system tuning and adaptive systems
- Advanced queries: ranking, skyline, etc.
- Cloud computing and parallel data processing
- Issues of data privacy and security



Evolution of Databases

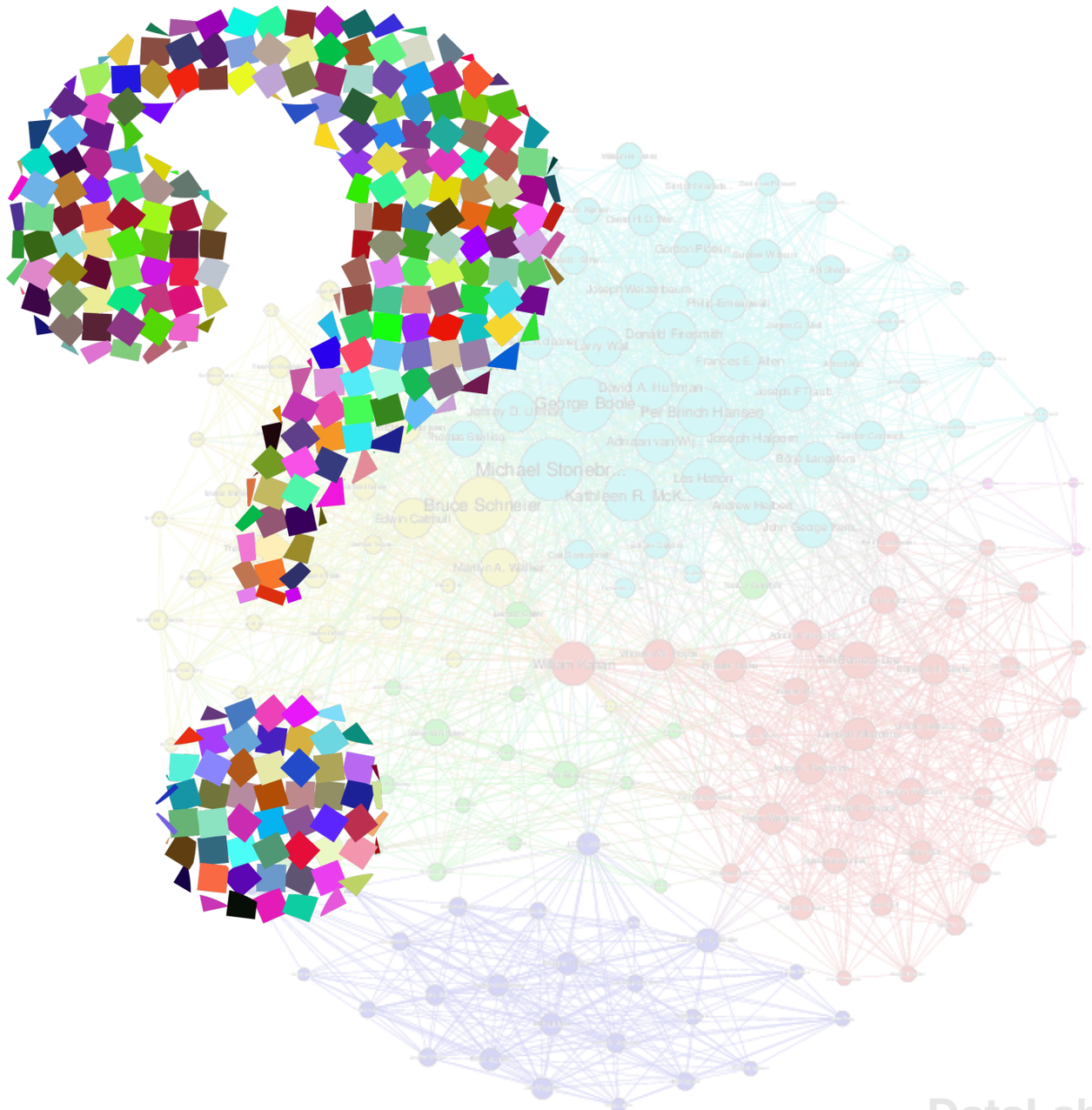
Advanced Data Analysis
(late-1980s to present)

- Data warehouse and OLAP
- Data mining and knowledge discovery: classification, clustering, outlier analysis, association and correlation, comparative summary, discrimination analysis, pattern discovery, trend and deviation analysis, etc.
- Mining complex types of data: streams, sequence, text, spatial, temporal, multimedia, Web, networks, etc.
- Data mining applications: business, society, retail, banking, telecommunications, science and engineering, blogs, daily life, etc.
- Data mining and society: invisible data mining, privacy-preserving data mining, mining social and information networks, recommender systems, etc.



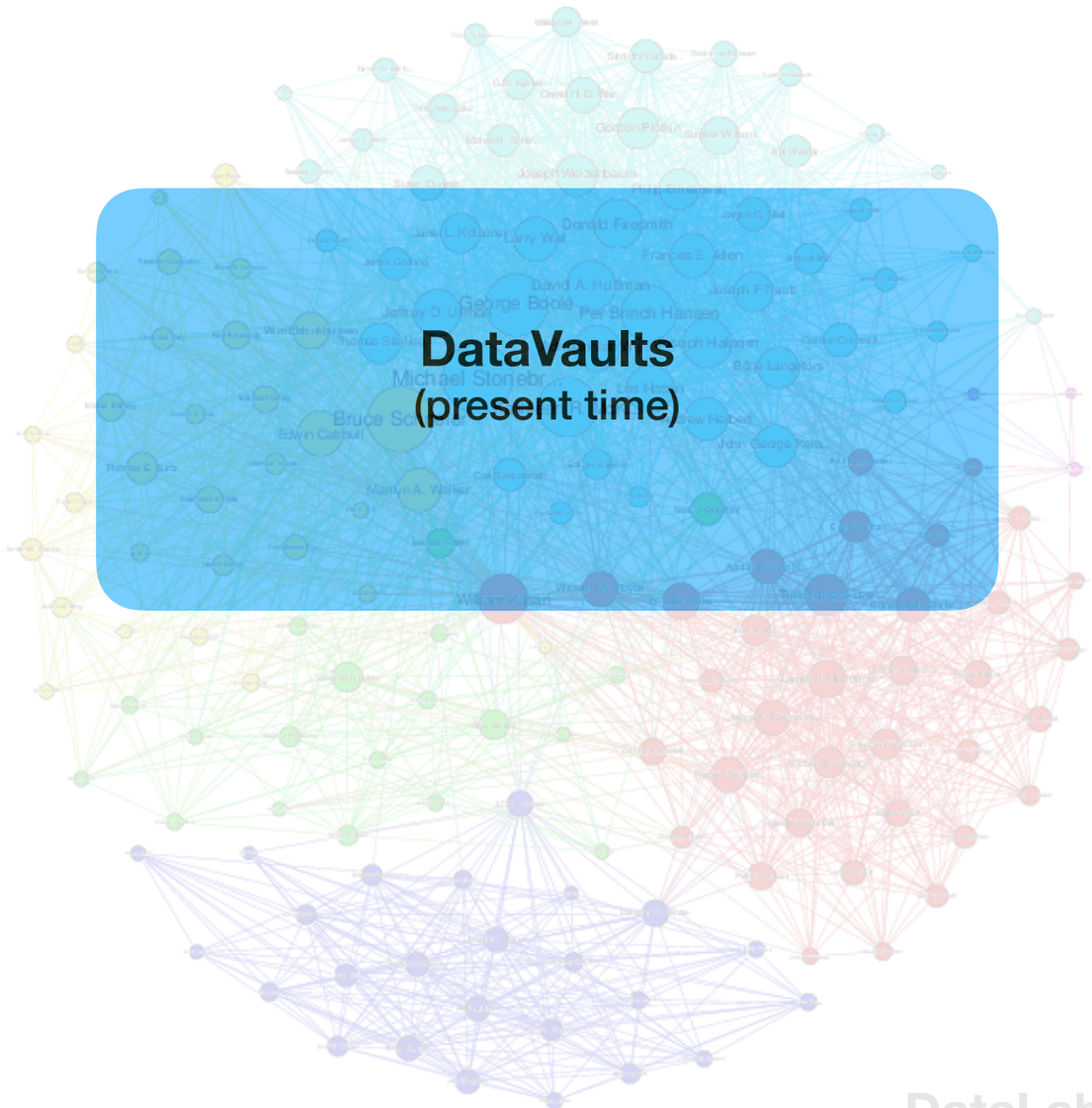
Evolution of Databases

What is the next step



Evolution of Databases

DataLakes
(present time)

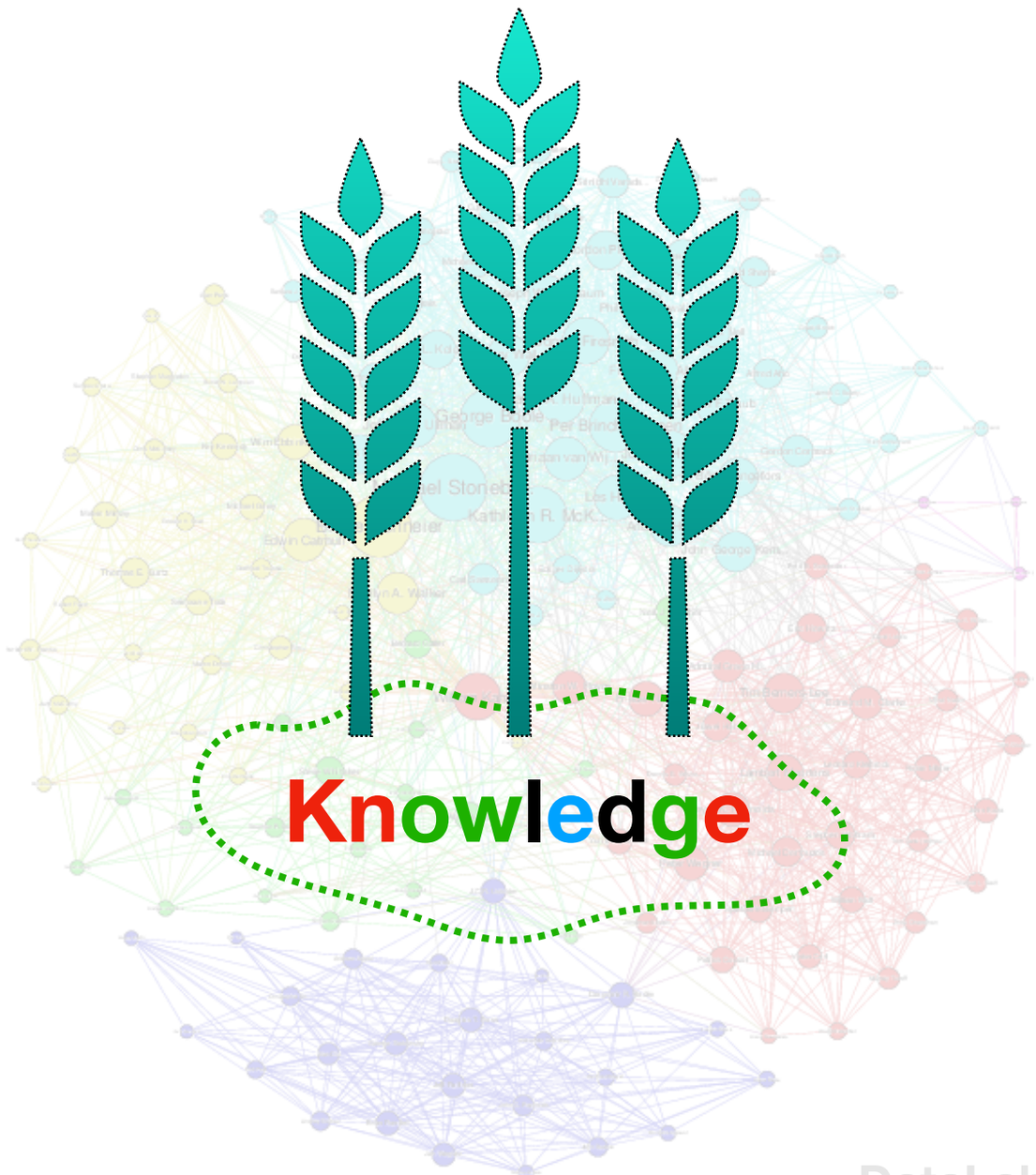


* my another course

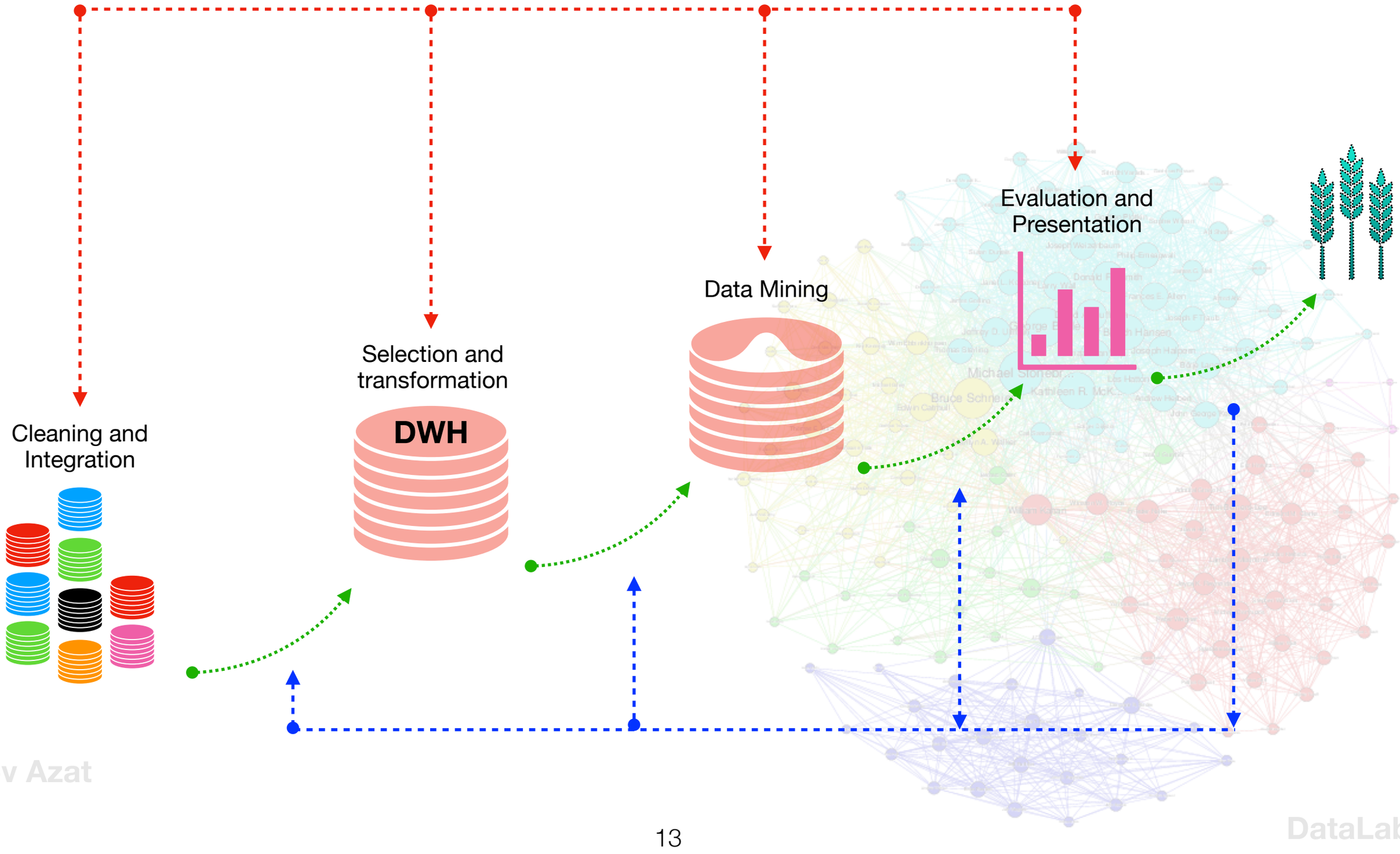
Why we need “Data Mining”

The world is **data rich**
but information **poor**

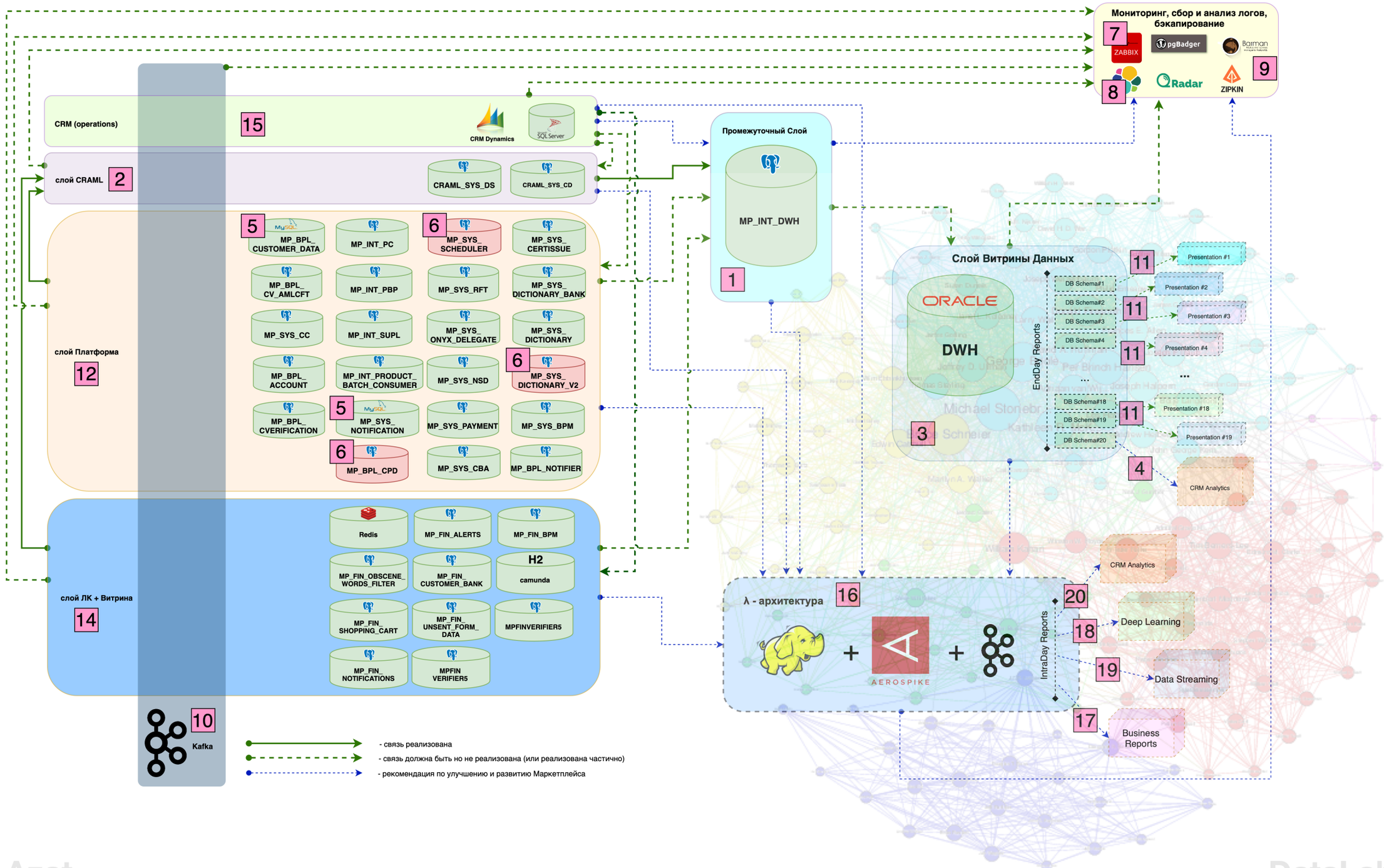
Why we need “Data Mining”



Data Mining Processes



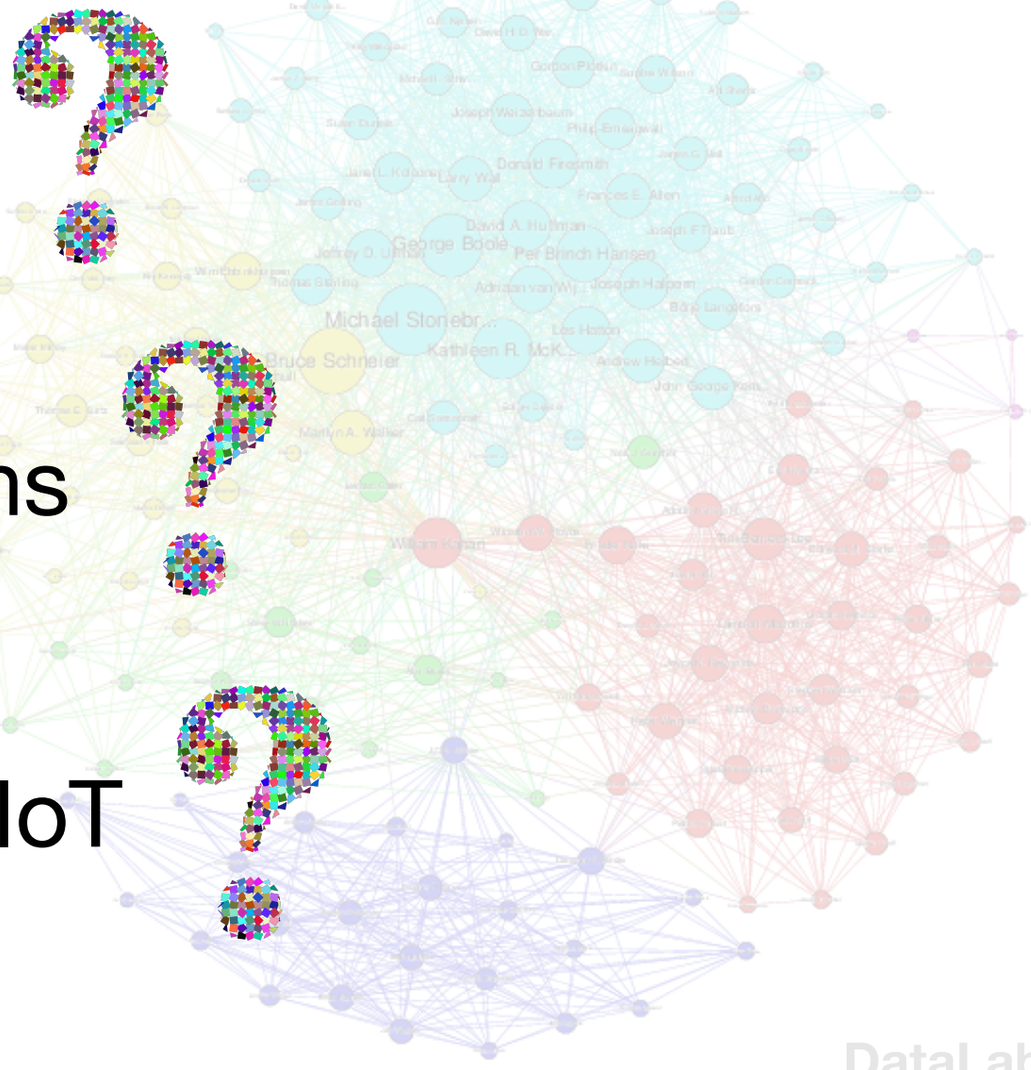
Data Mining Processes



Data Mining Tasks

- World Wide Web

- Financial interactions



- Phone user interactions
- Sensor technologies, IoT

Sample of Data Mining Task

Retailer has Web logs corresponding to customer accesses to Web pages at his or her site. Each of these Web pages corresponds to a product, and a customer access to a page may often be indicative of interest in that particular product. The retailer also stores demographic profiles for the different customers. The retailer wants to make targeted product recommendations to customers using the customer demographic and buying behaviour.

Sample of Data

98.206.207.157 - - [31/Jul/2020:18:09:38 -0700] “GET /productA.htm HTTP/1.1” 200 328177 “-” “Mozilla/5.0 (Mac OS X) AppleWebKit/536.26. (KHTML, like Gecko) Version/6.0 Mobile/10B329 Safari/8536.25” “retailer.net”

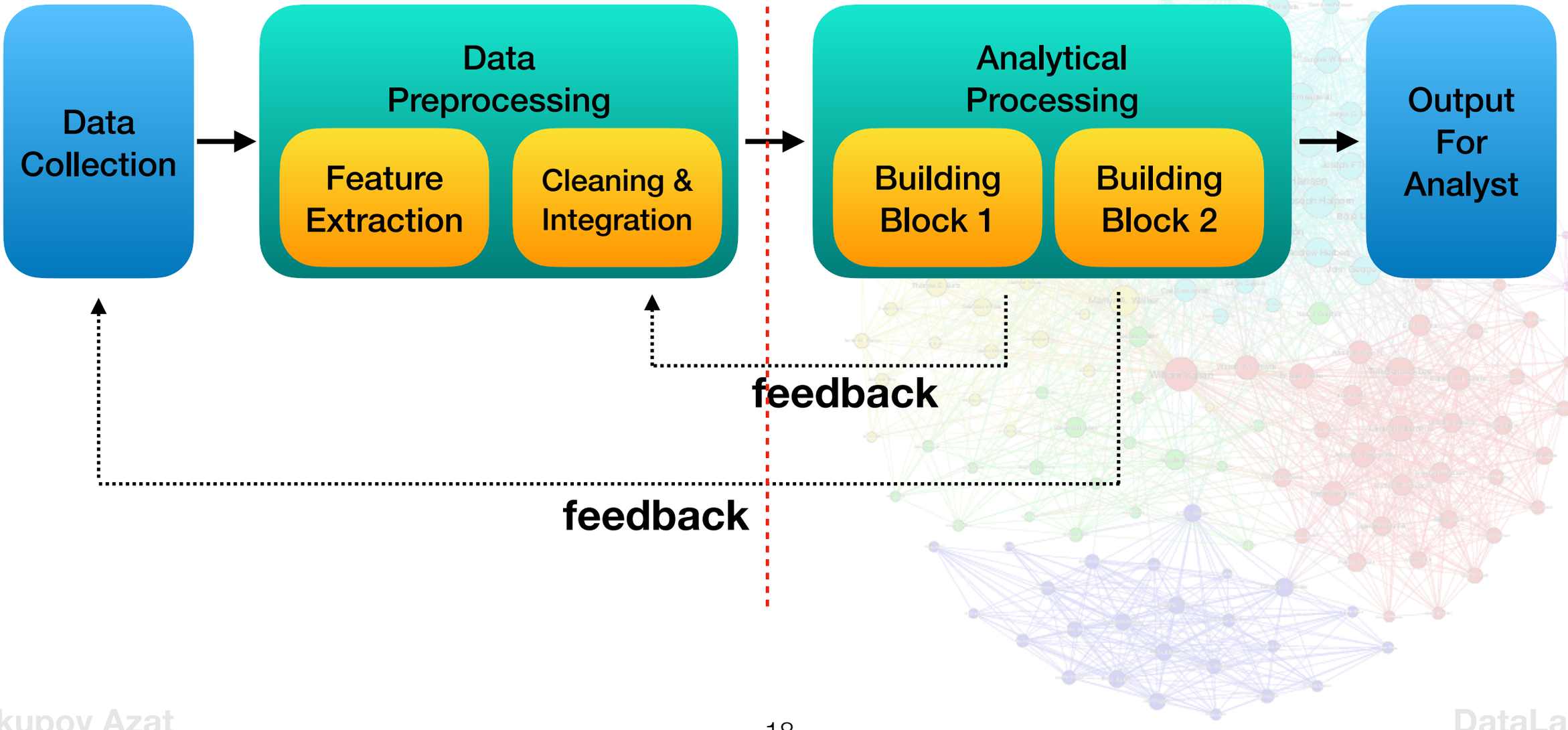
Data Mining Example

```
SELECT product, count(*)
  FROM Bag
GROUP BY product
ORDER BY product
```

Data Mining Process

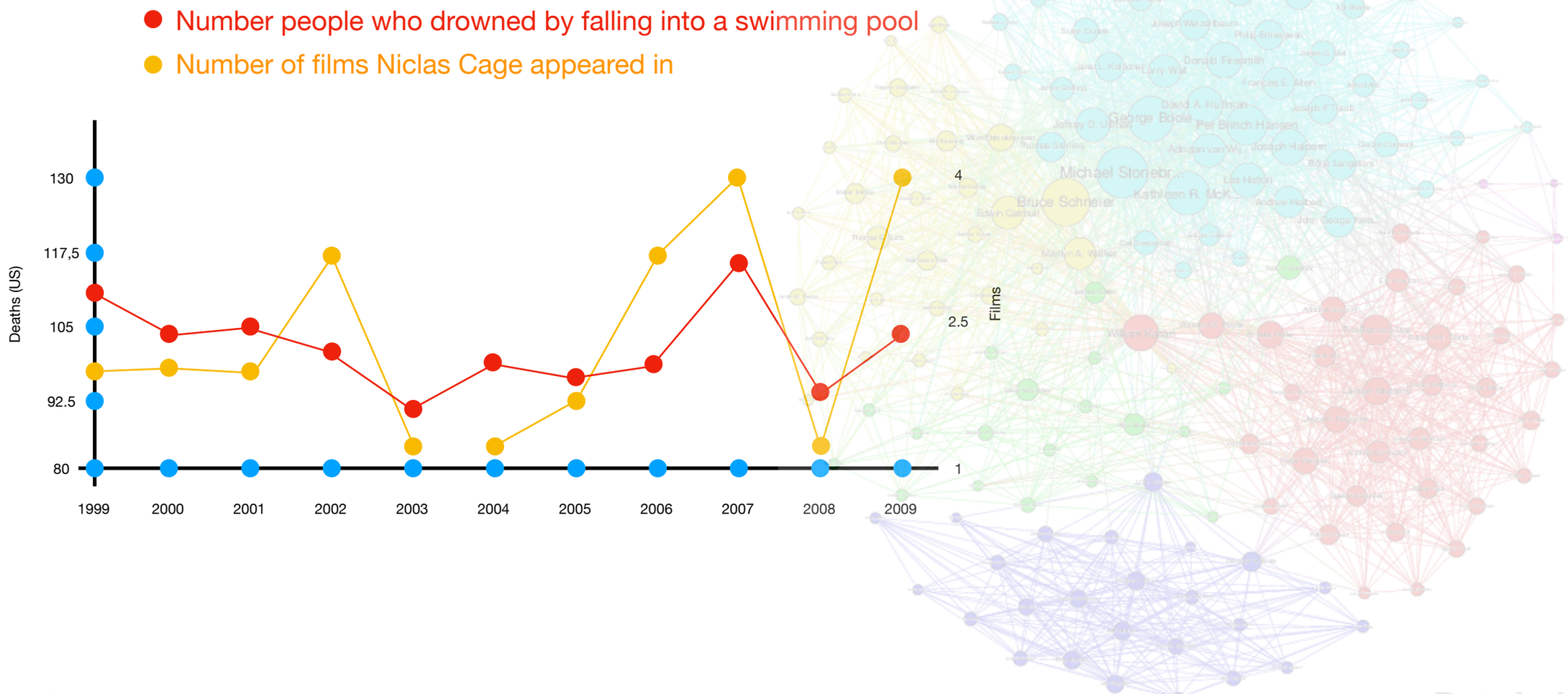
Data Preparation Phase

Analytical Phase



Bonferroni principle

Bonferroni's principle **helps to avoid** finding **bogus artefacts** in the data versus something what is truly there. In other words, it avoids finding simply random occurrences in data.



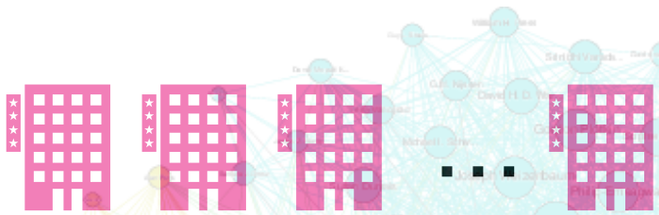
Bonferroni principle

(example)

Suppose, there are some ***terrorists*** and we want to detect them.
Also suppose that periodically they get together at a hotel to plot something bad

-  10^9 people with ***terrorists***

and

 $100\ 000$ hotels

- Everyone [] goes to a hotel  one day in 100 days
- Every  holds 100 people. (100 000 hotels - enough to hold 1% of a 10^9 people per day)
- Let’s examine hotel records for 1000 days

Bonferroni principle

(example)

The task is to find 2 ***terrorists*** who on 2 different days, were both at the same hotel

Suppose, everyone is a good man (**no any *terrorists***), meaning that everyone behaves at **random**, deciding with $p = 0.01$ to visit a hotel on a given day, choosing one of the $100\ 000$ hotels at **random**.

$$p(\text{person } \textit{and} \text{ person} \mid \text{sun}) = 0.01 \times 0.01 = 0.0001$$

at the **same** given Day

$$p(\text{person } \textit{and} \text{ person} \mid \text{sun} \mid \text{hotel}) = 0.0001 \times \frac{1}{100000} = 10^{-9}$$

at the **same** given Day at the **same** Hotel

Bonferroni principle

(example)

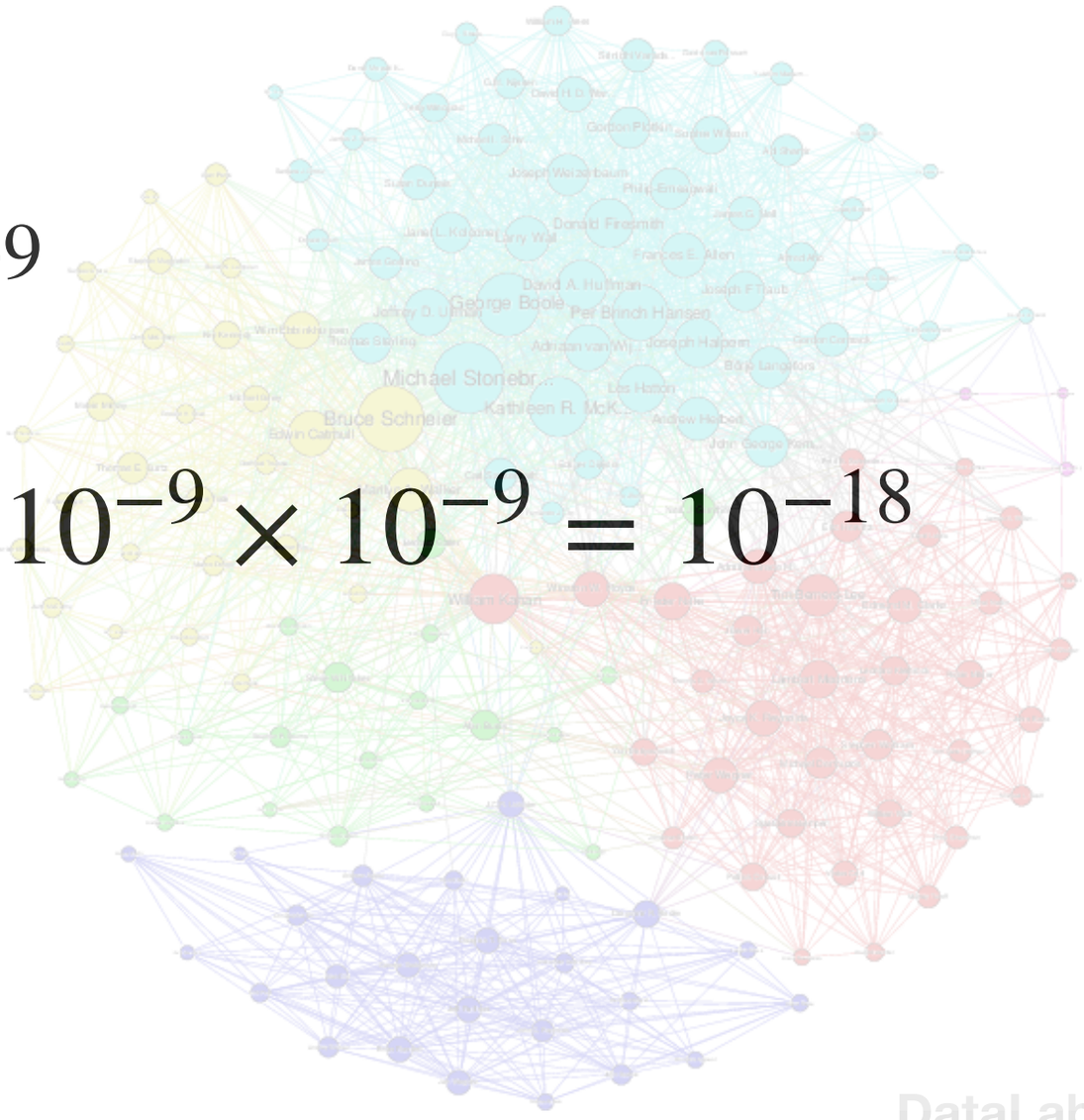
The task is to find 2 ***terrorists*** who on 2 different days, were both at the same hotel

$$p(\text{♀ and ♂} \mid \text{☀} \mid \text{🏨}) = 10^{-9}$$

at the **same** given Day at the **same** Hotel

$$p(\text{♀ and ♂} \mid \text{1st} \mid \text{2nd} \mid \text{🏨}) = 10^{-9} \times 10^{-9} = 10^{-18}$$

at the **same** Hotel



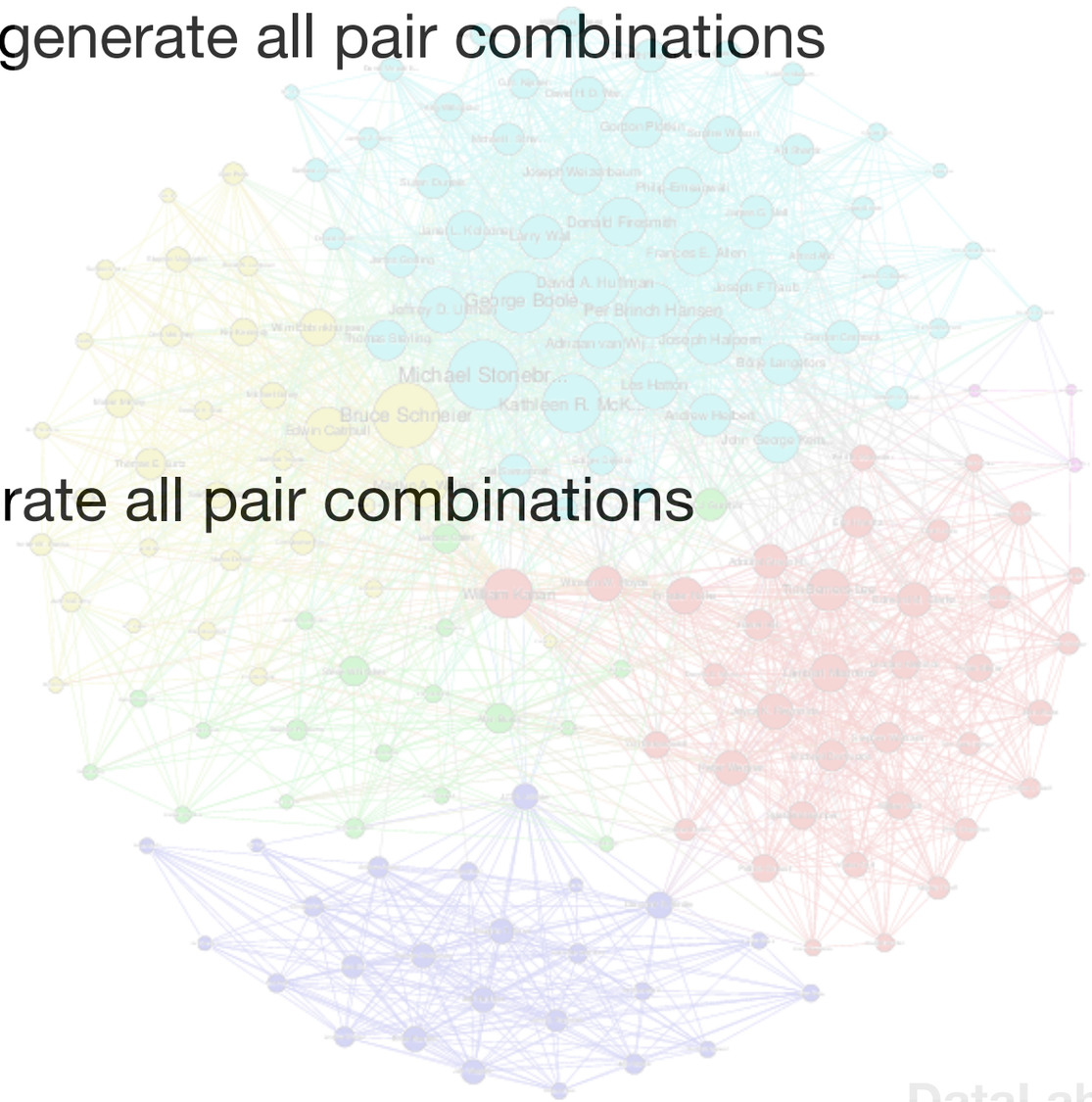
Bonferroni principle

(example)

To find a couple of ***terrorists*** we should generate all pair combinations

$$\binom{10^9}{2} \approx \frac{(10^9)^2}{2} = 5 \cdot 10^{17}$$

if n is greater number



To find a couple of ***days*** we should generate all pair combinations

$$\binom{10^3}{2} \approx \frac{(10^3)^2}{2} = 5 \cdot 10^5$$

Bonferroni principle

(example)

The final solution is

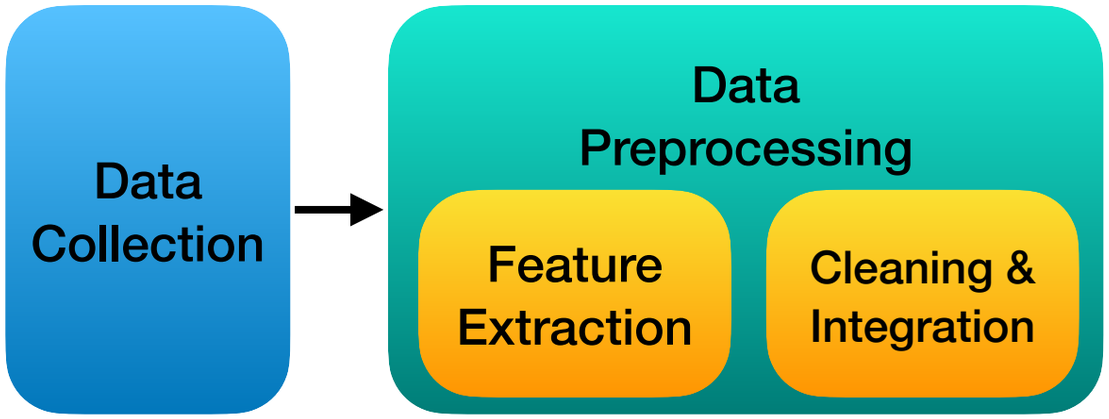
pairs of x ***pairs of*** x $p($ *and* | 1st | 2nd | at the **same** Hotel $) =$

$$5 \cdot 10^{17} \times 5 \cdot 10^{5} \times 10^{-18} = 250\,000$$

people who look like ***terrorists*** but are not

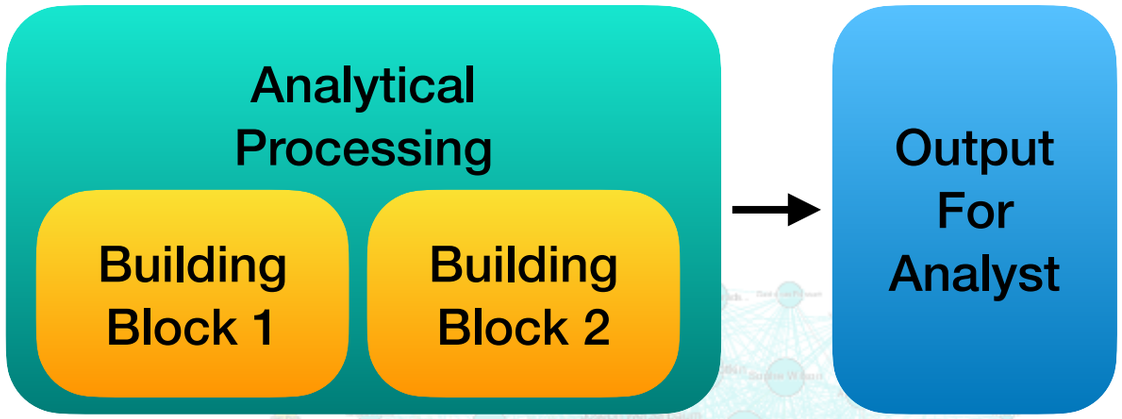
≈ 0.025 %

Data Preparation Phase



- Data Sources
- Relational Model
- PostRelational Model
- Extract Transform Load processes

Analytical Phase



- Mining Algorithms
- Mining Approaches
- Repeatable Methods

Data Mining Pattern Tasks

- Association Pattern Mining
- Clustering
- Classification
- Outlier detection

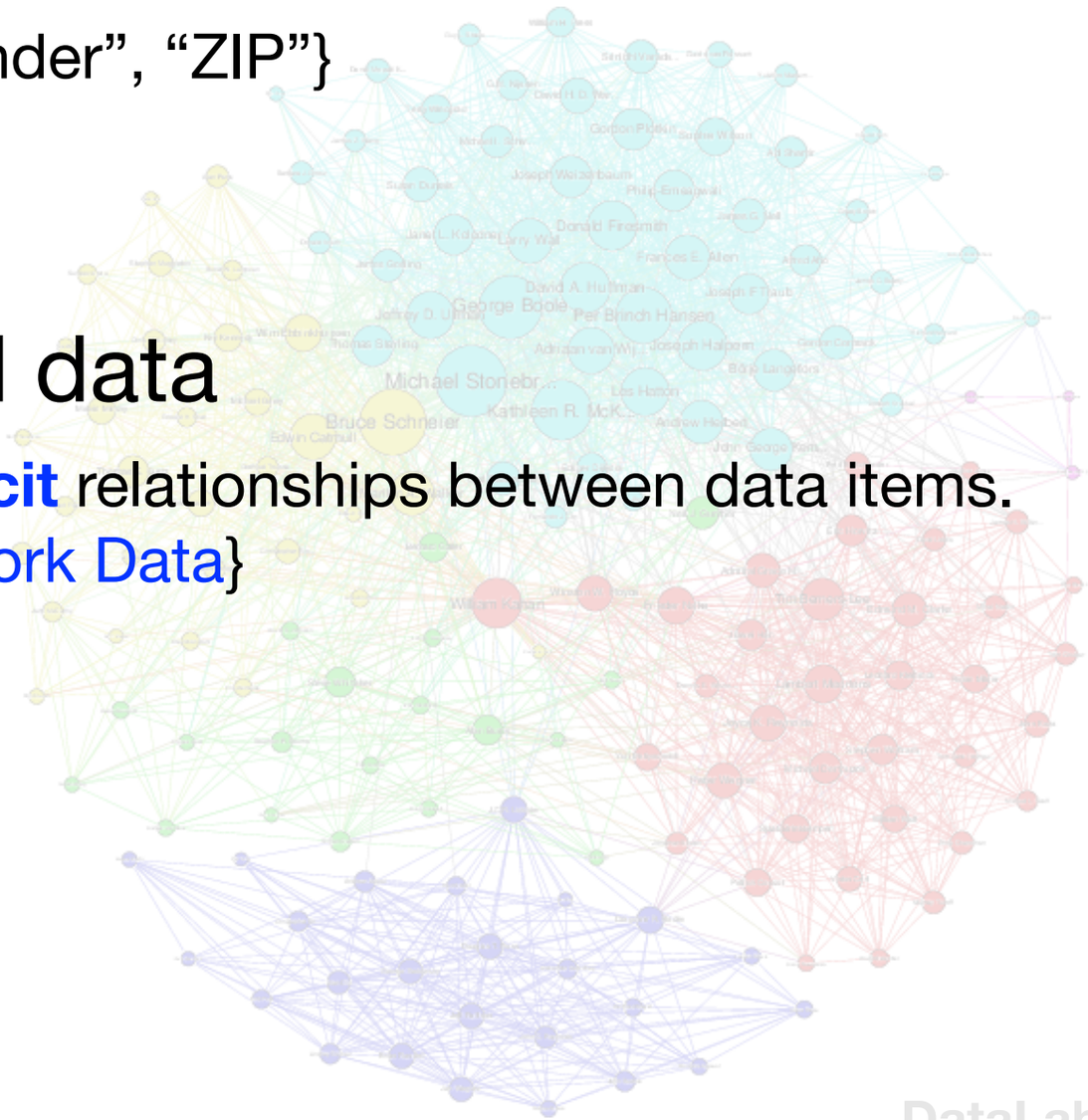
The Basic Data Types

- Nondependency-oriented data

  Simple data types. {“Age”, “Gender”, “ZIP”}

- Dependency-oriented data

  There are some **implicit** / **explicit** relationships between data items.
  {Time Series Data, Social Network Data}



Nondependency-oriented data

Name	Age	Gender	Race	ZIP
Ivan	45	M	Russian	10648
Peter	29	M	Native American	19467
Hen	13	F	Asian	98731
Kate	38	F	EU	28388

Set of Tuples ~
Multidimensional Data

A **multidimensional data set** $\mathscr{D}$ is a set of *n* records, $\overline{X_1} \ldots \overline{X_n}$

such that each record $\overline{X_i}$ contains a set of d features denoted by $(x_i^1 \ldots x_i^d)$

Nondependency-oriented data

Name	Age	Gender	Race	ZIP
Ivan	45	M	Russian	10648
Peter	29	M	Native American	19467
Hen	13	F	Asian	98731
Kate	38	F	EU	28388

- Quantitative Multidimensional Data: {Age}
- Categorical Data: {Gender, Race, ZIP}
- Mixed Attribute Data: {Age + Gender + Race + ZIP}
- Binary and Set Data: {Gender}
- Text Data: {Name}

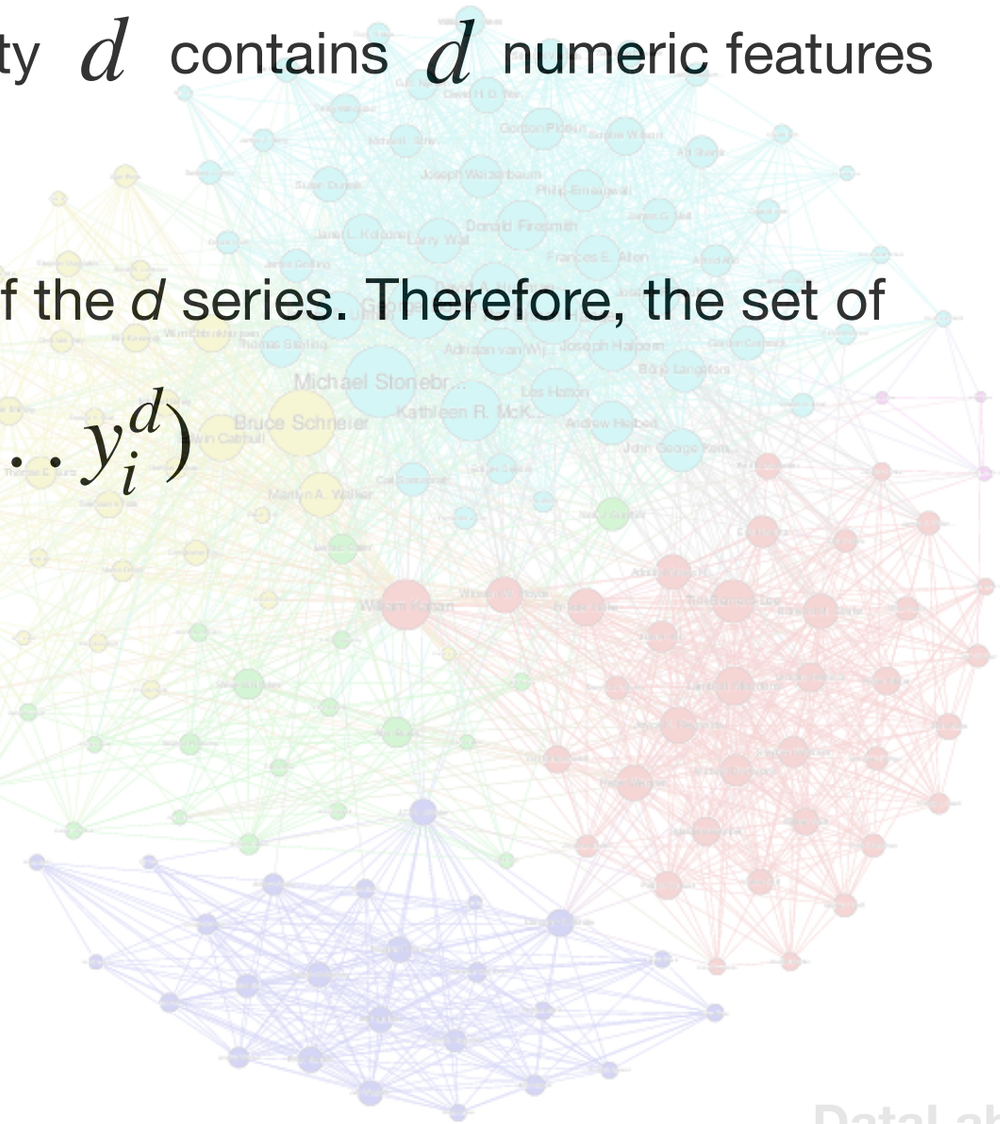
Dependency-oriented data

- Time-Series Data

A **time series** of length *n* and dimensionality d contains d numeric features at each of *n* time stamps $t_1 \ldots t_n$

Each time stamp contains a component for each of the *d* series. Therefore, the set of values received at time stamp t_i is $\overline{Y_i} = (y_i^1 \ldots y_i^d)$

The value of the *j*th series at time stamp t_i is y_i^j

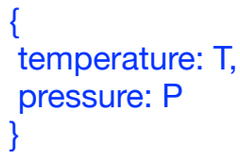


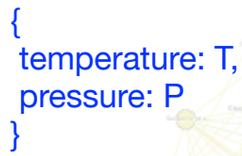
Dependency-oriented data

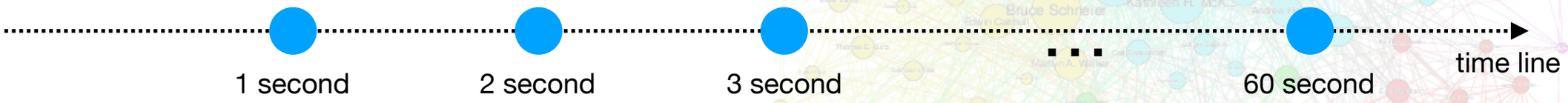
- Time-Series Data Sample

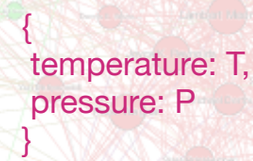
$$d = 2$$

$$n = 60$$









Dependency-oriented data

- Discrete Sequences and Strings

A **discrete sequence** of length *n* and dimensionality d contains d discrete feature values at each of *n* different time stamps $t_1 \ldots t_n$

Each of the *n* components $\overline{Y_i}$ contains d discrete behavioural attributes $(y_i^1 \ldots y_i^d)$ collected at the *i*-th time-stamp.

Behavioural attributes are values that are measured in a particular context. (temperature, pressure)

Dependency-oriented data

- Discrete Sequences Sample

$$d = 2$$

$$n = 100$$

{
Web page: P1,
client IP: C1
}

{
Web page: P2,
client IP: C2
}

{
Web page: P3,
client IP: C3
}

{
Web page: P4,
client IP: C4
}

1 request

2 request

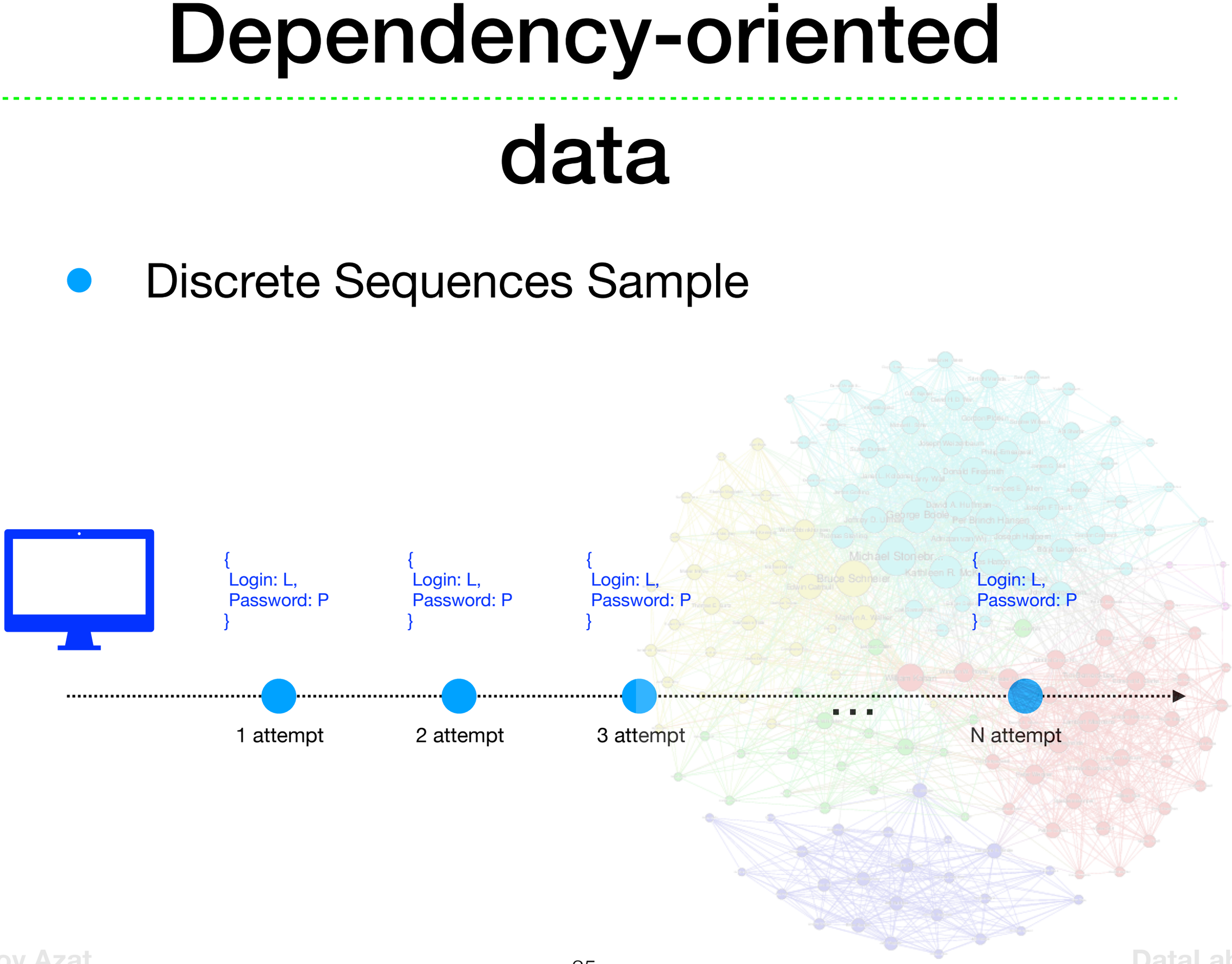
3 request

...

100 request

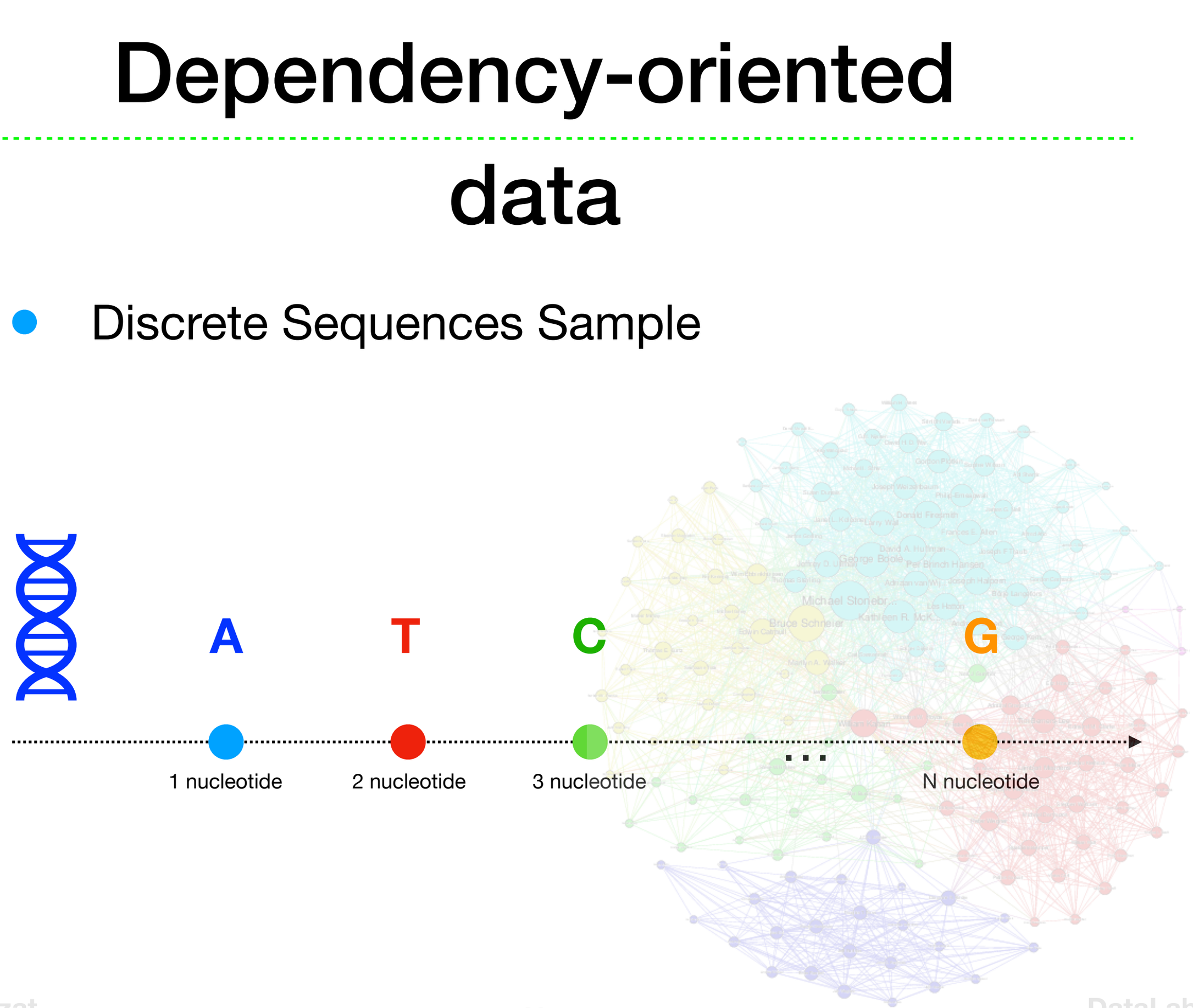
Dependency-oriented data

- Discrete Sequences Sample



Dependency-oriented data

- Discrete Sequences Sample



Dependency-oriented data

- Spatial Data

A d - dimensional **spatial data** record contains d behavioural attributes and one or more contextual attributes containing the spatial location.

Therefore, a d-dimensional spatial data set is a set of d dimensional records $\overline{X_1} \ldots \overline{X_n}$ together with a set of *n* locations $L_1 \ldots L_n$, such that the record $\overline{X_i}$ is associated with the location L_i

Behavioural attributes are values that are measured in a particular context. (temperature, pressure)

Contextual attributes are values define the context of the basis. (time stamp for sensor data)

Dependency-oriented data

- Spatial Data Sample

$$d = 2$$

$$n = 60$$

{
temperature: T,
pressure: P,
geolocation: 1
}

{
temperature: T,
pressure: P,
geolocation: 1
}

{
temperature: T,
pressure: P,
geolocation: 1
}

{
temperature: T,
pressure: P,
geolocation: 1
}

1 second

2 second

3 second

. . .

60 second

time line

{
temperature: T,
pressure: P,
geolocation: 2
}

{
temperature: T,
pressure: P,
geolocation: 2
}

{
temperature: T,
pressure: P,
geolocation: 2
}

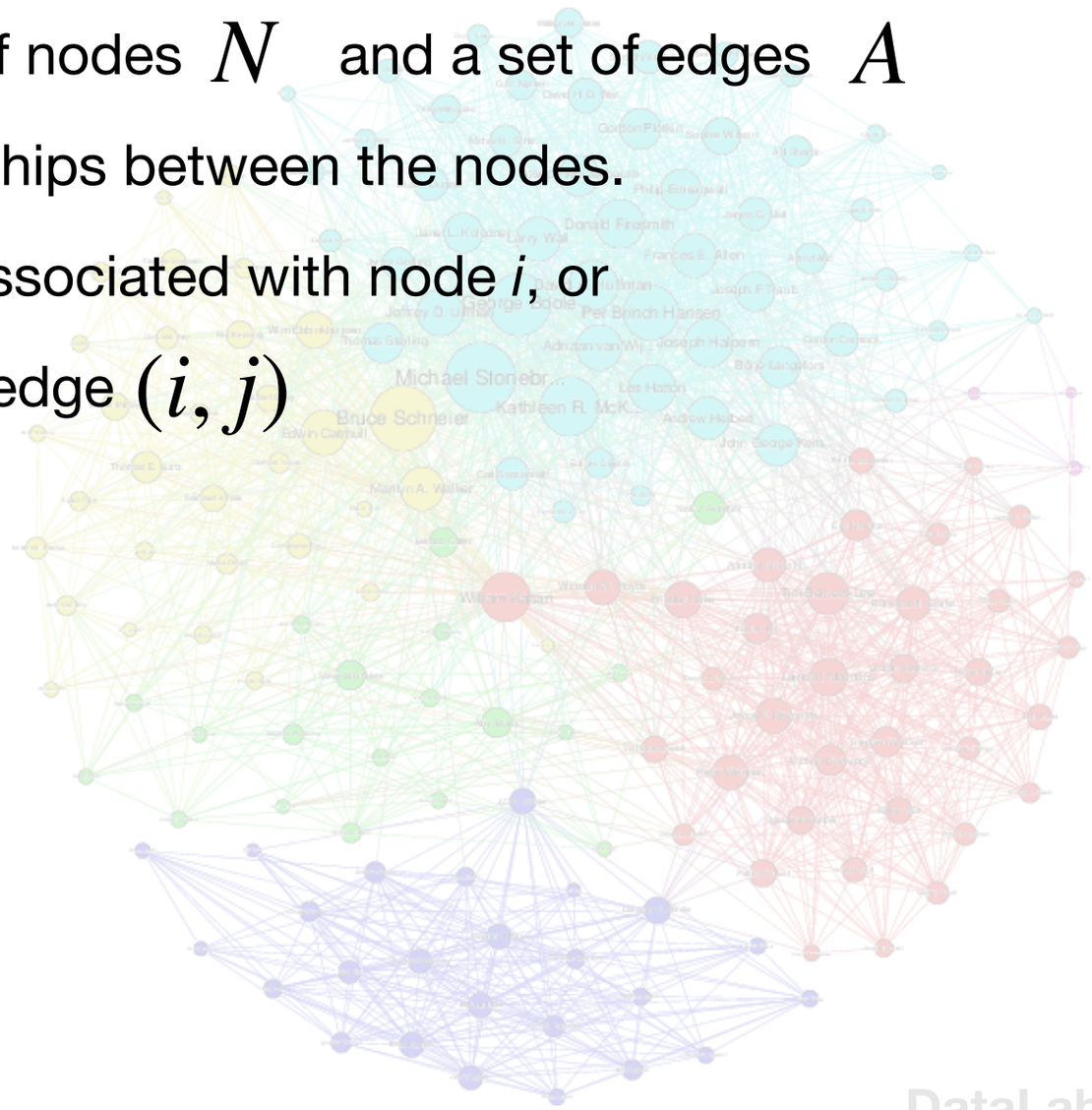
{
temperature: T,
pressure: P,
geolocation: 2
}

Dependency-oriented data

- Network and Graph Data

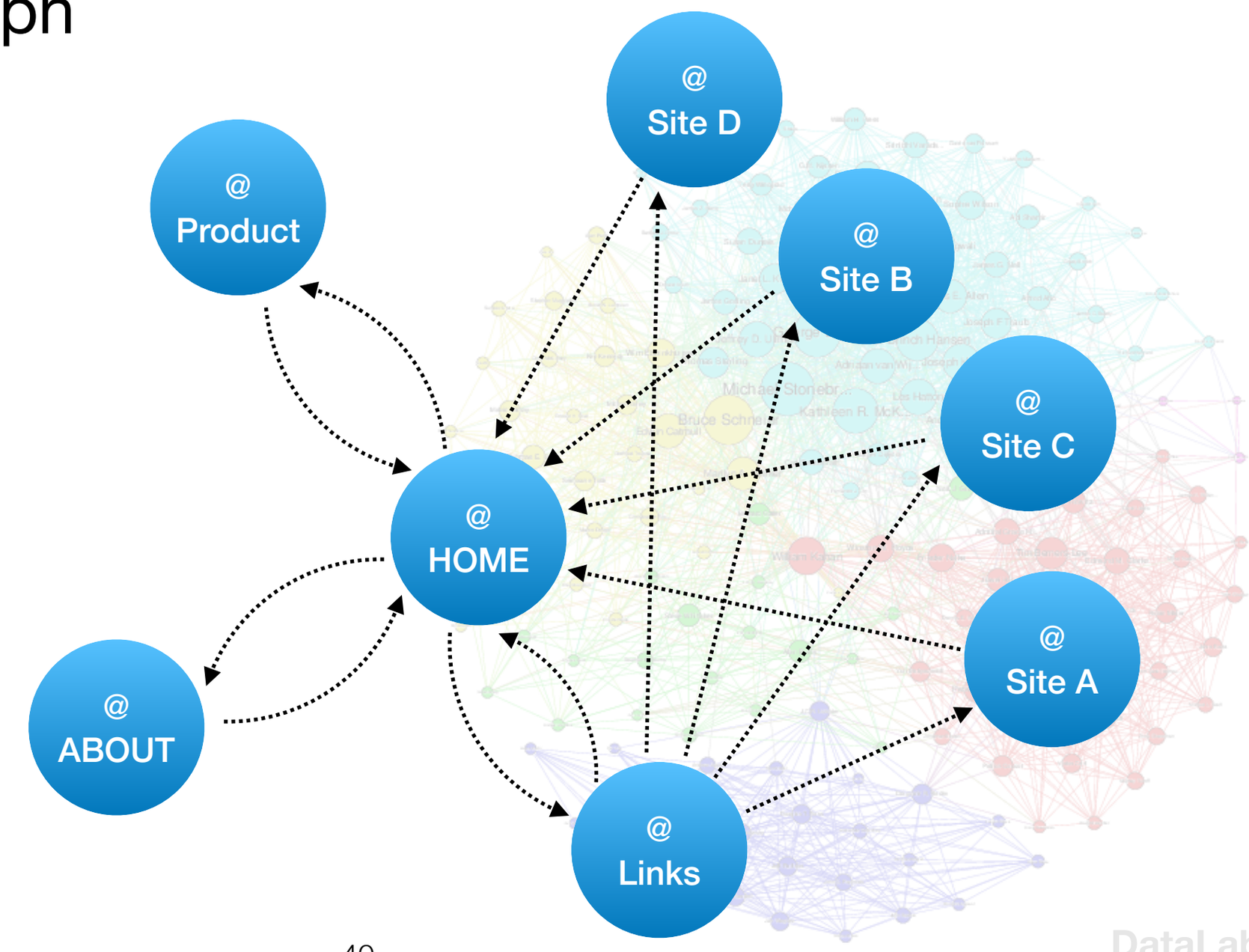
A network $G = (N, A)$ contains a set of nodes N and a set of edges A where the edges in A represent the relationships between the nodes.

In some cases, an attribute set $\overline{X_i}$ may be associated with node *i*, or an attribute set $\overline{Y_{ij}}$ may be associated with edge (i, j)



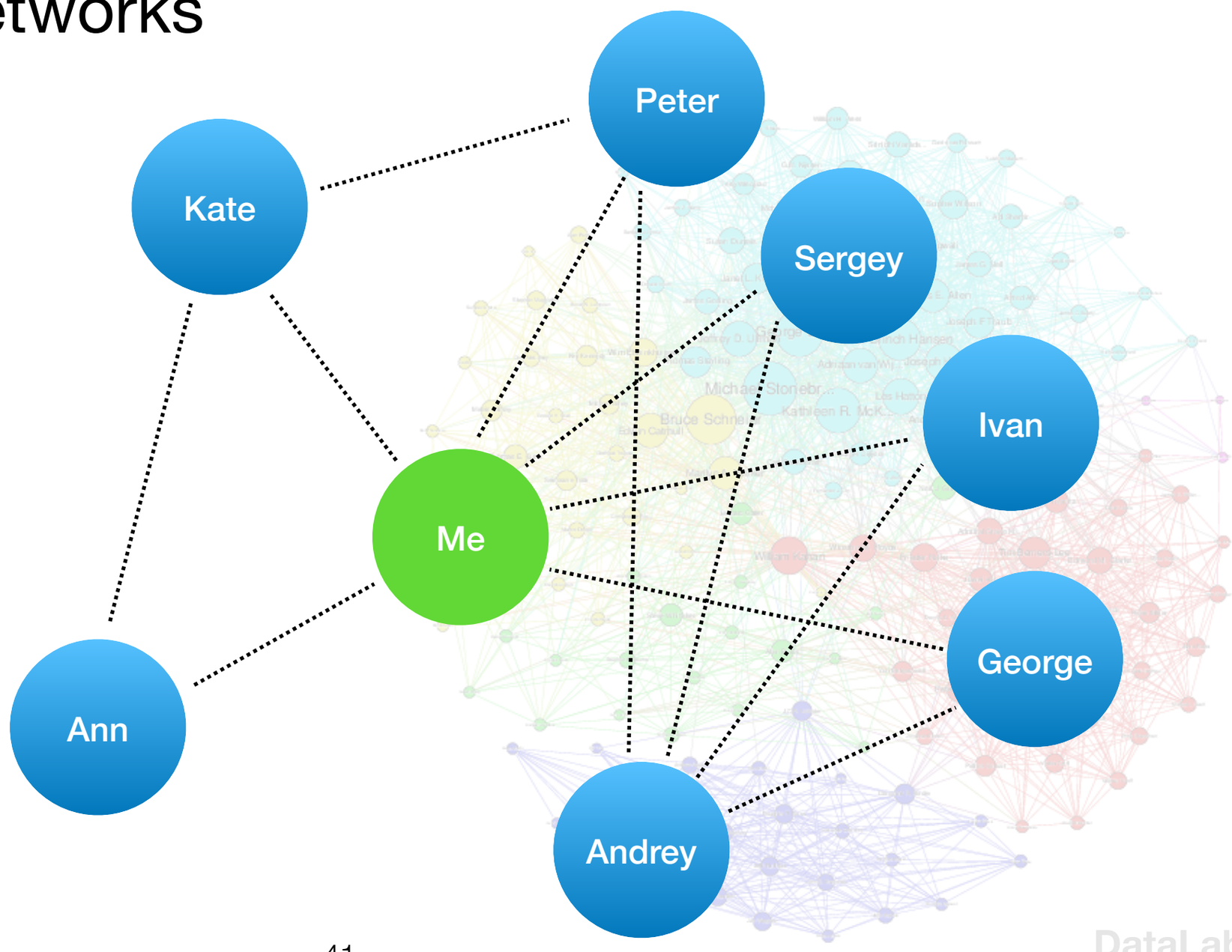
Dependency-oriented data

- Web graph



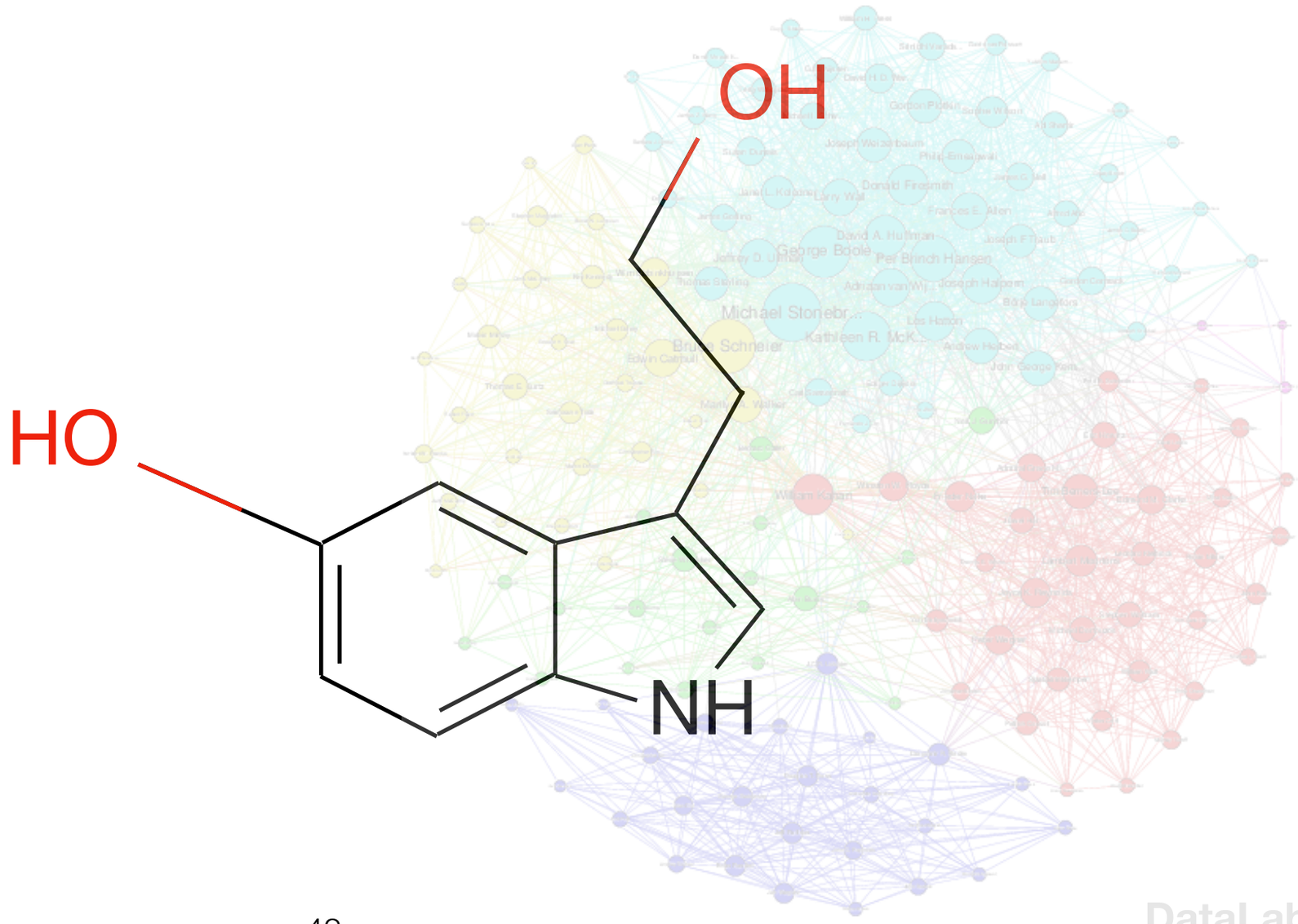
Dependency-oriented data

- Social networks



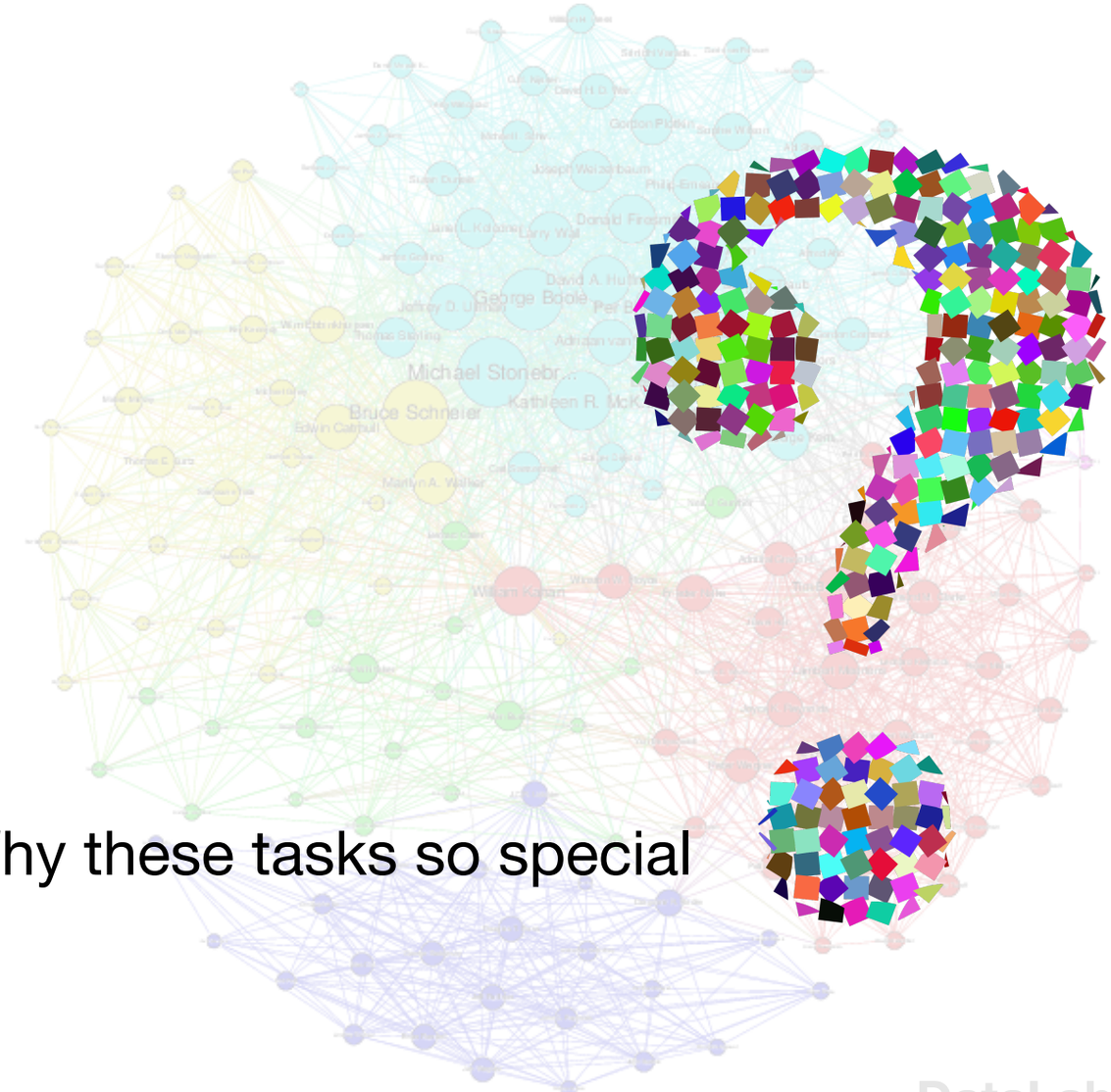
Dependency-oriented data

- Chemical compound databases



Data Mining Pattern Tasks

- Association Pattern Mining
- Clustering
- Classification
- Outlier detection



Why these tasks so special

Data Mining Pattern Tasks

A **multidimensional database** $\mathscr{D}$ with n records, and d attributes.

A **data matrix** D with n rows, and d columns.

$$D = \begin{array}{|cccccccc|} \hline & & & & & & & \\ & & & & & & & \\ & & & & & & & \\ & & & & & & & \\ & & & & & & & \\ & & & & & & & \\ & & & & & & & \\ & & & & & & & \\ \hline \end{array} \begin{array}{l} n_1 \\ n_2 \\ n_3 \\ n_4 \\ \\ \vdots \\ \\ n_n \end{array}$$

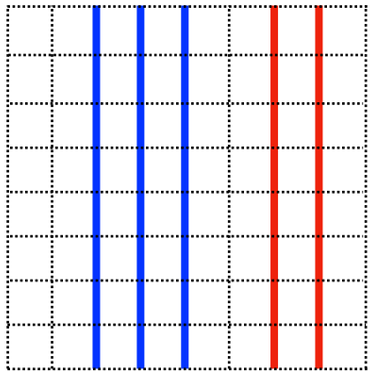
$$d_1 \quad d_2 \quad d_3 \quad d_4 \quad \cdots \quad d_d$$

$$D_{n \times d} \sim \mathscr{D}$$

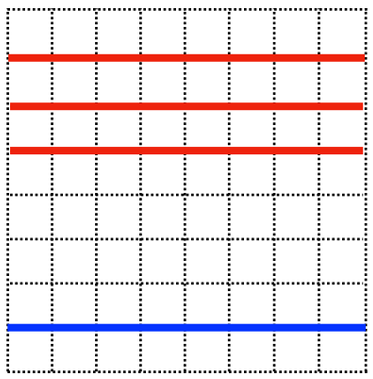
Data Mining Pattern Tasks

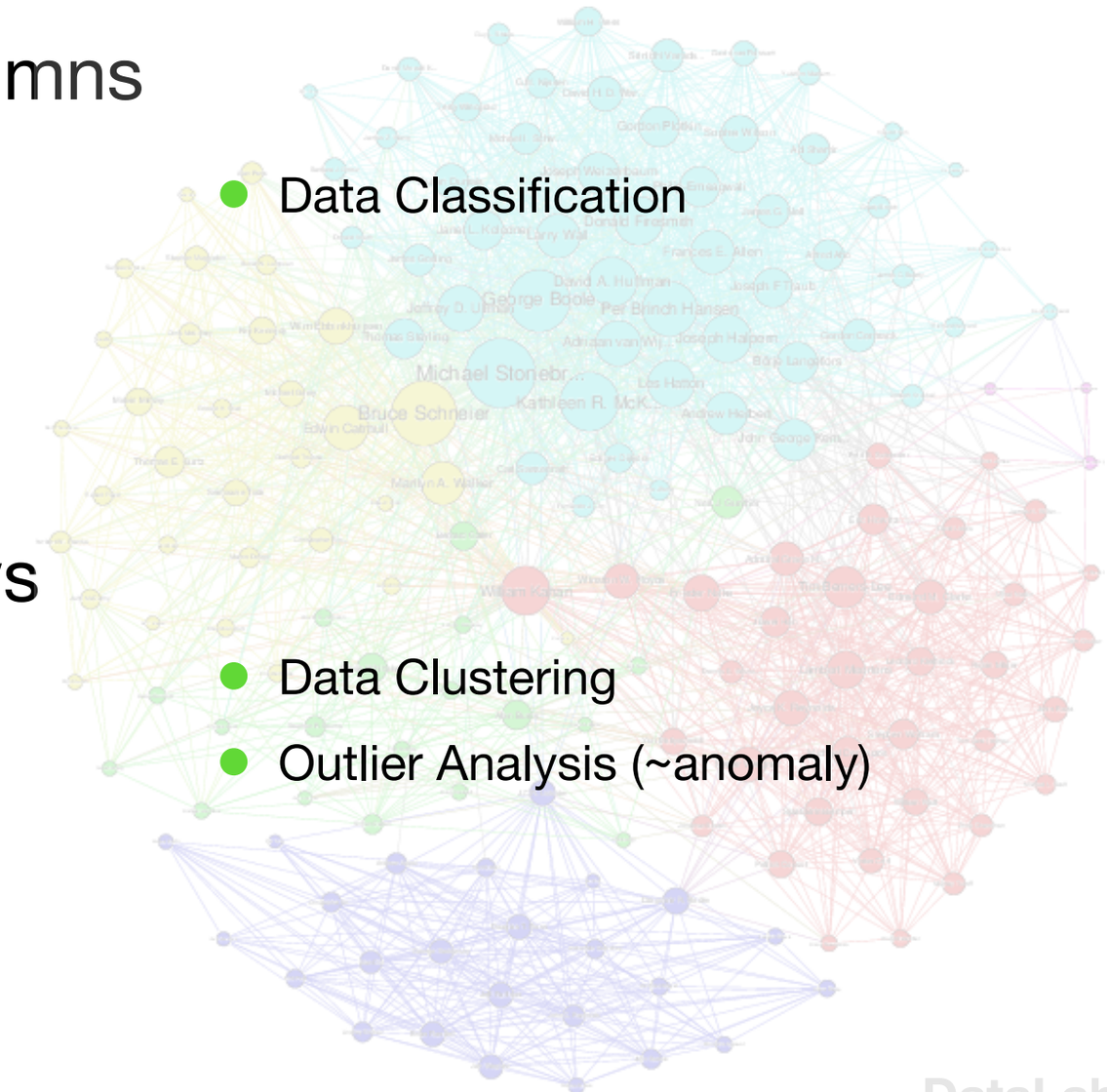
Data mining is finding summary relationships between the entries in the Data Matrix D that are either unusually frequent or unusually infrequent.

- Relationships between columns



- Relationships between rows





- Data Classification
- Data Clustering
- Outlier Analysis (~anomaly)

Association Pattern Mining

- in “Frequent Pattern Mining” term

Given a binary $n \times d$ data matrix D, determine all subsets of columns such that all the values in these columns take on the value of 1 for at least a fraction s of the rows in the matrix. The relative frequency of a pattern is referred to as its support. The fraction s is referred to as the minimum support.

$D =$

	Bread	Butter	Milk	Beer	Salad	
fraction s	1	1	1	0	0	Client #1
	1	1	1	1	0	Client #2
	1	1	1	1	0	Client #3
	1	1	1	0	1	Client #4

looks like these items are often bought together

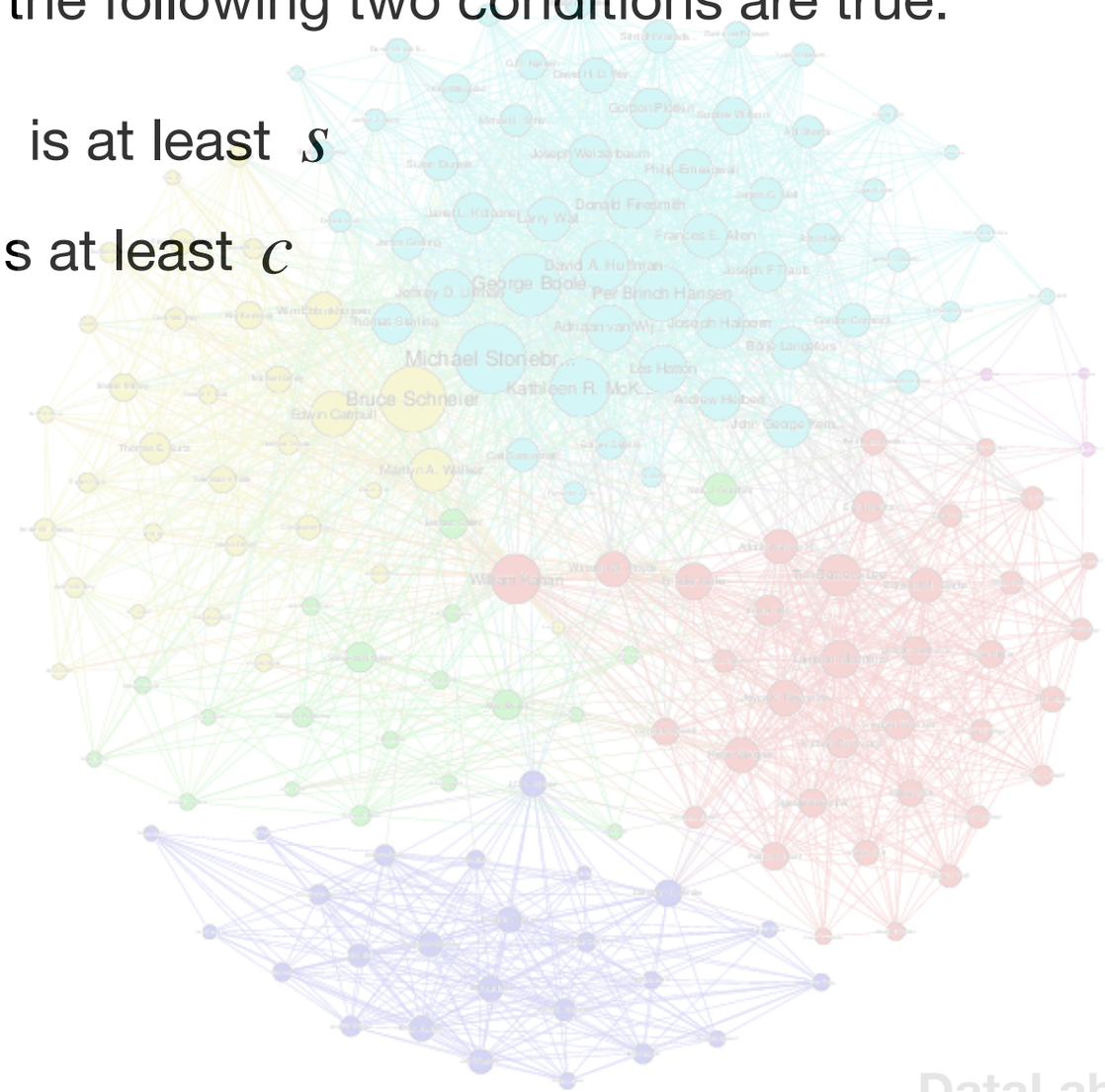
Association Pattern Mining

- in “Association Rules” term

Let A and B be two sets of items. The rule $A \Rightarrow B$ is said to be valid

at support level s and confidence level c, if the following two conditions are true:

- The support of the item set A is at least s
- The confidence of $A \Rightarrow B$ is at least c



Association Pattern Mining

$Wine \Rightarrow Cheese$ [**Support**: 9 %, **Confidence**: 65 %]

Support is the percentage of transactions (rows in data matrix D)
that contain both *Wine* and *Cheese* together.
(9% of all baskets had these 2 items together.)

$$Support(Wine \Rightarrow Cheese) = P(A \cup B)$$

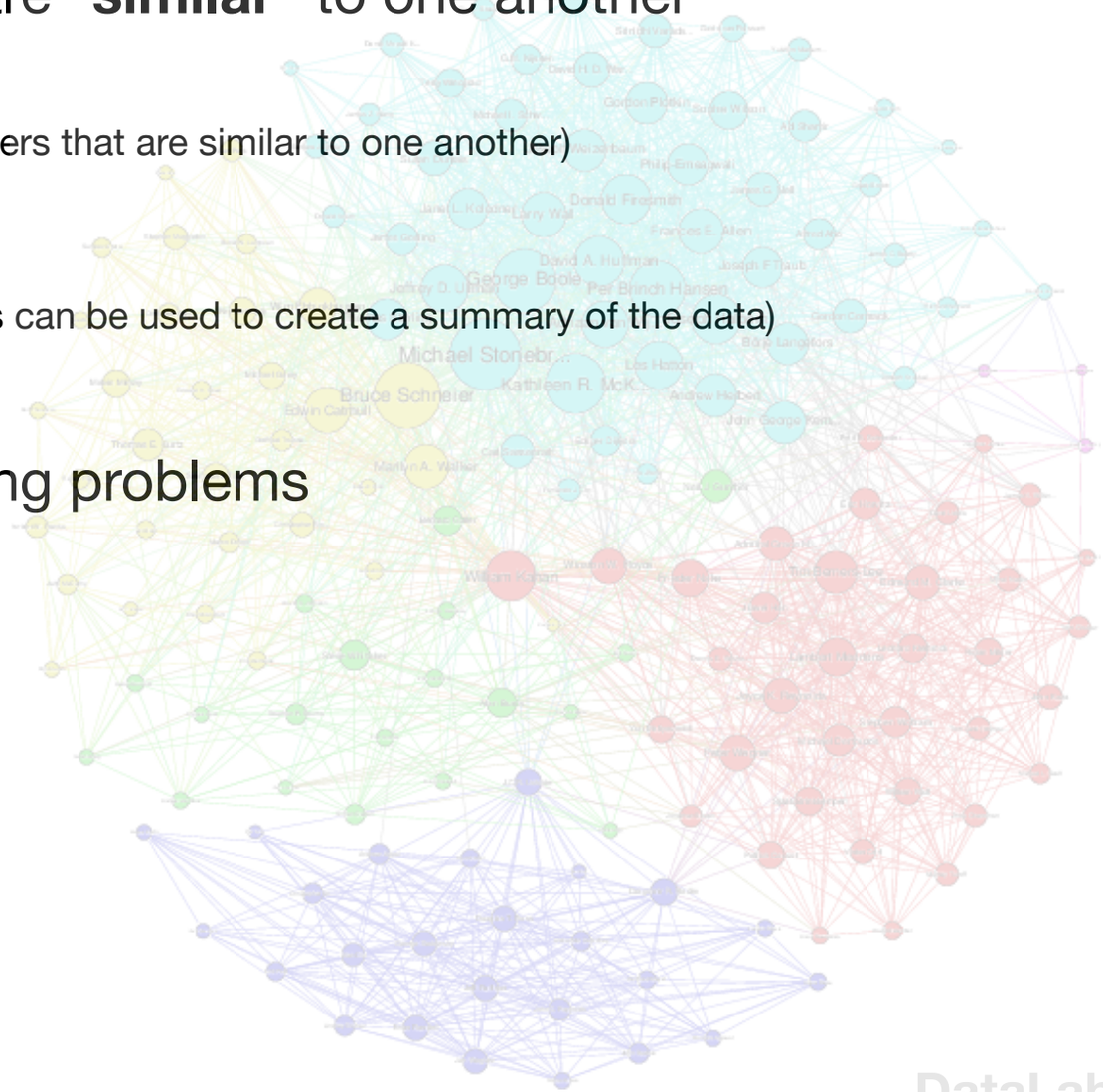
Confidence is the percentage of transactions (rows in data matrix D),
containing *Wine*, that also contain *Cheese*. In other words,
the probability of having *Cheese*, given that *Wine* is already in the basket.
(65% of all those who bought *Wine*, also bought *Cheese*.)

$$Confidence(Wine \Rightarrow Cheese) = P(A \mid B)$$

Data Clustering

Given a data matrix D , partition its rows (records) into sets $C_1, \ldots, C_k$ such that the rows (records) in each cluster are "**similar**" to one another

- Customer segmentation (customers that are similar to one another)
- Data summarization (similar groups can be used to create a summary of the data)
- Application to other data mining problems



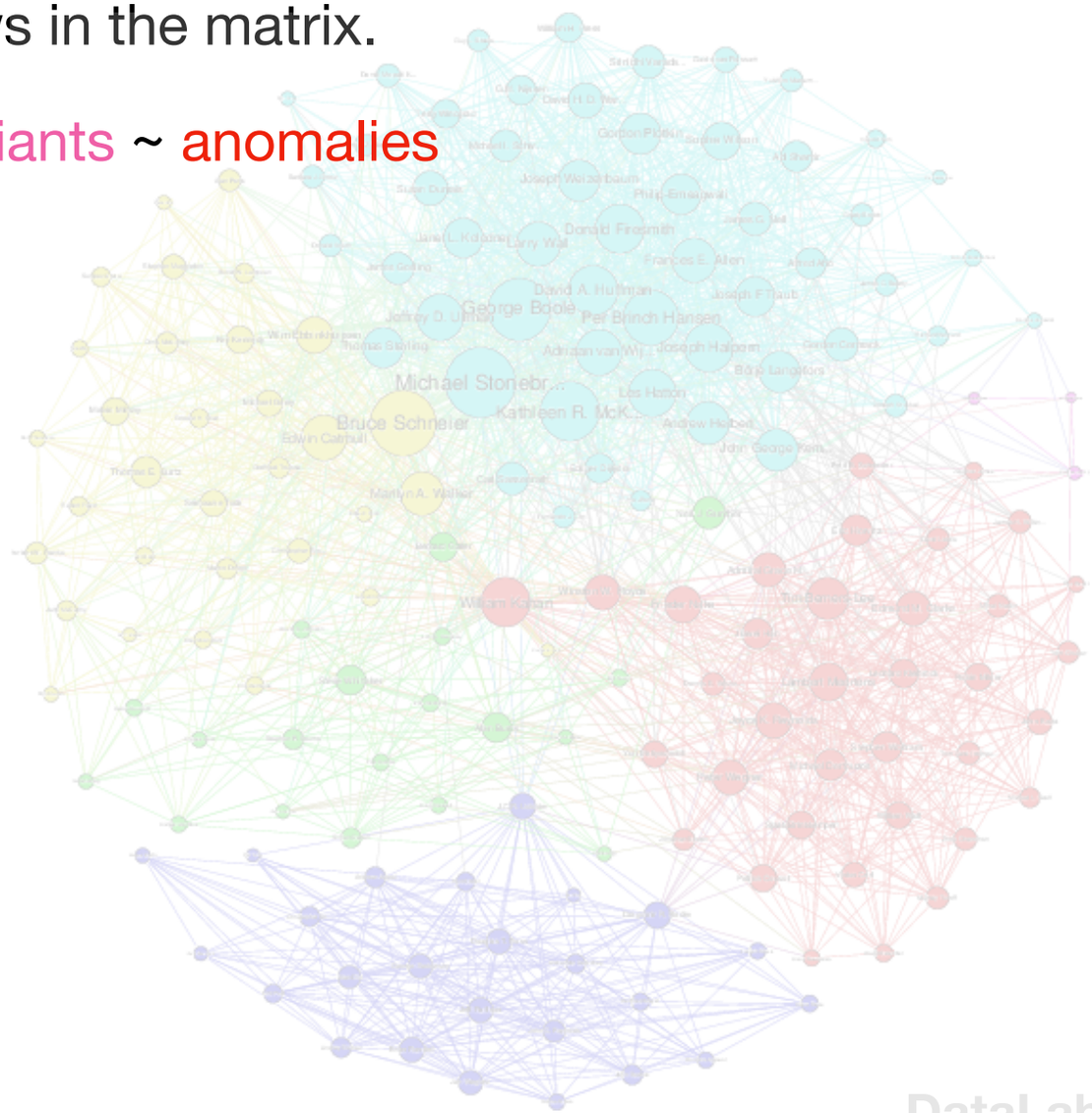
Outlier Detection

Given a data matrix D , determine the rows of the data matrix

that are very different from the remaining rows in the matrix.

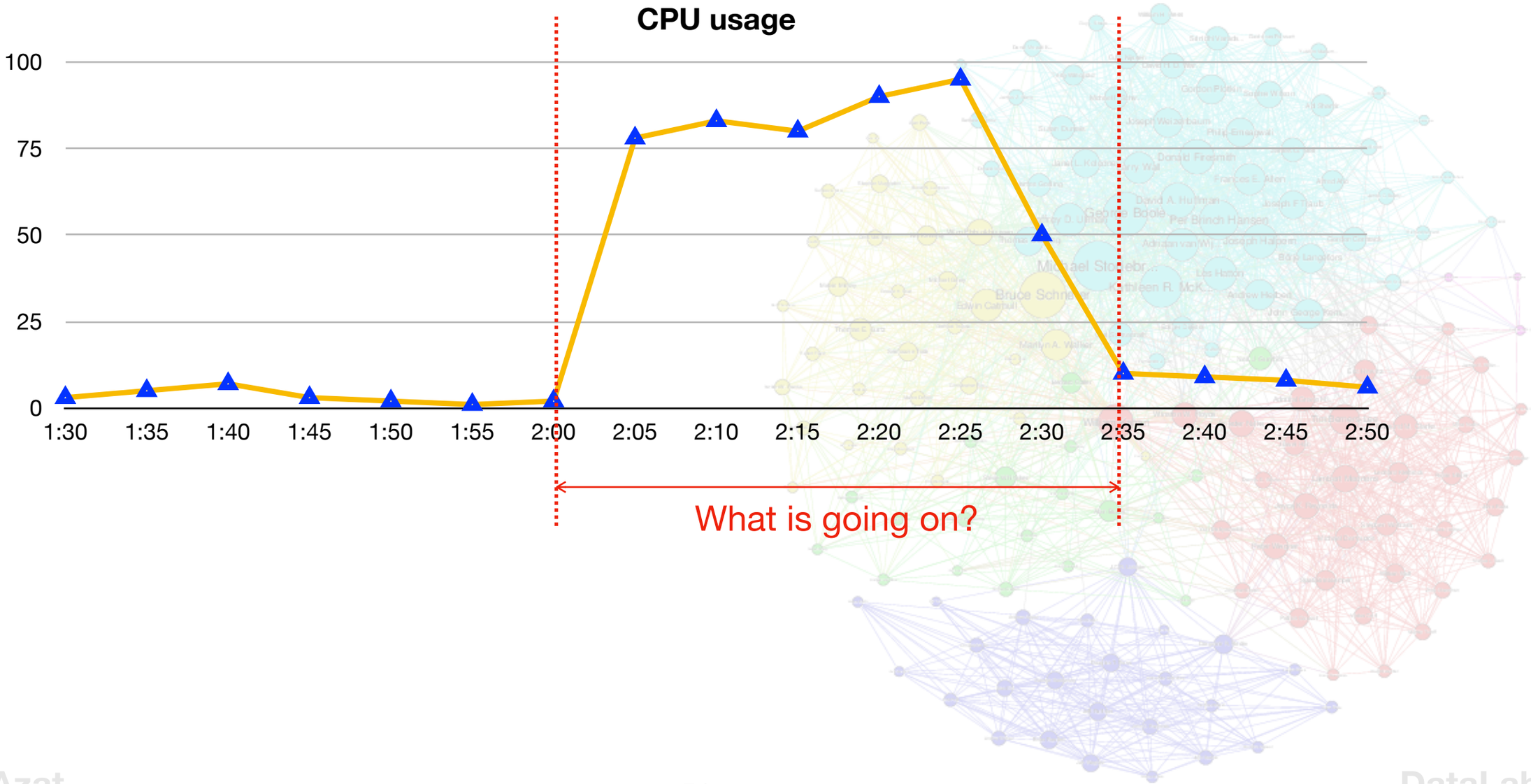
Outliers ~ abnormalities ~ discordants ~ deviants ~ anomalies

- Intrusion-detection systems
- Credit card fraud
- Sensor events
- Medical diagnosis
- Law enforcement
- Earth science



Outlier Detection Sample

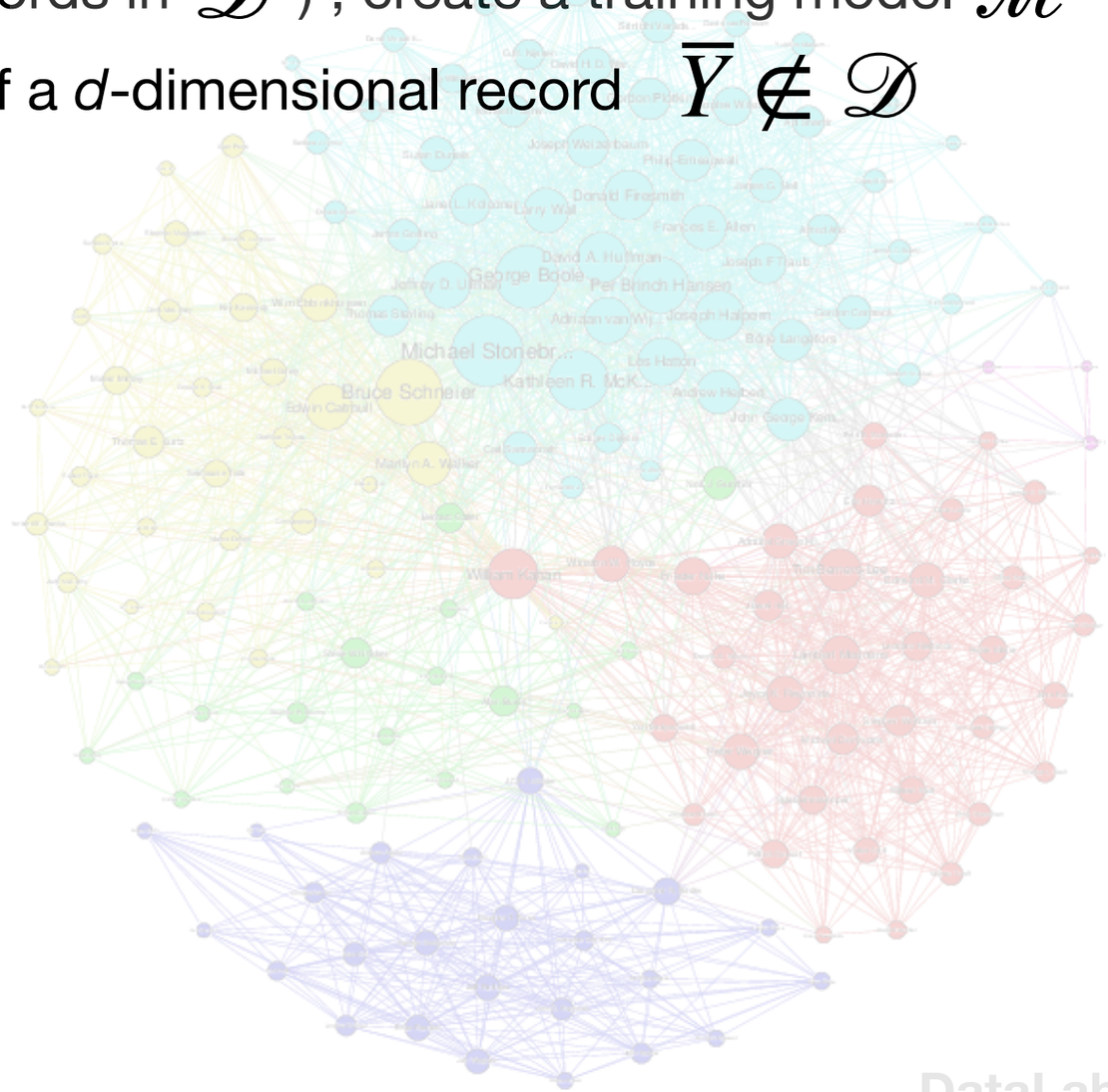
- System monitoring service



Data Classification

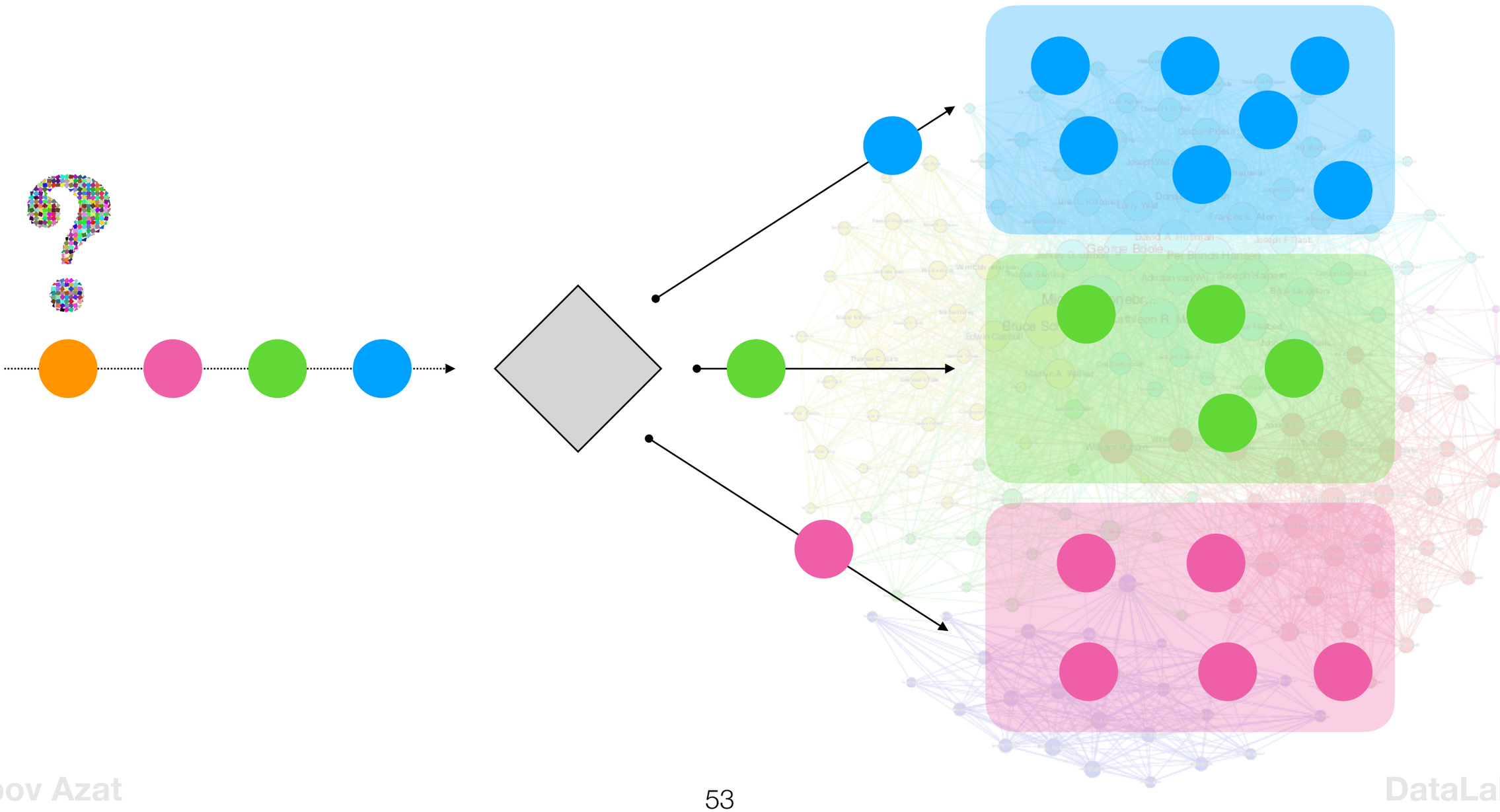
Given an $n \times d$ training matrix D , and a class label value in $\{1...k\}$ associated with each of the *n* rows in D (records in $\mathscr{D}$) , create a training model $\mathscr{M}$ which can be use to predict the class label of a *d*-dimensional record $\overline{Y} \notin \mathscr{D}$

- Target marketing
- Intrusion detection
- Supervised anomaly detection



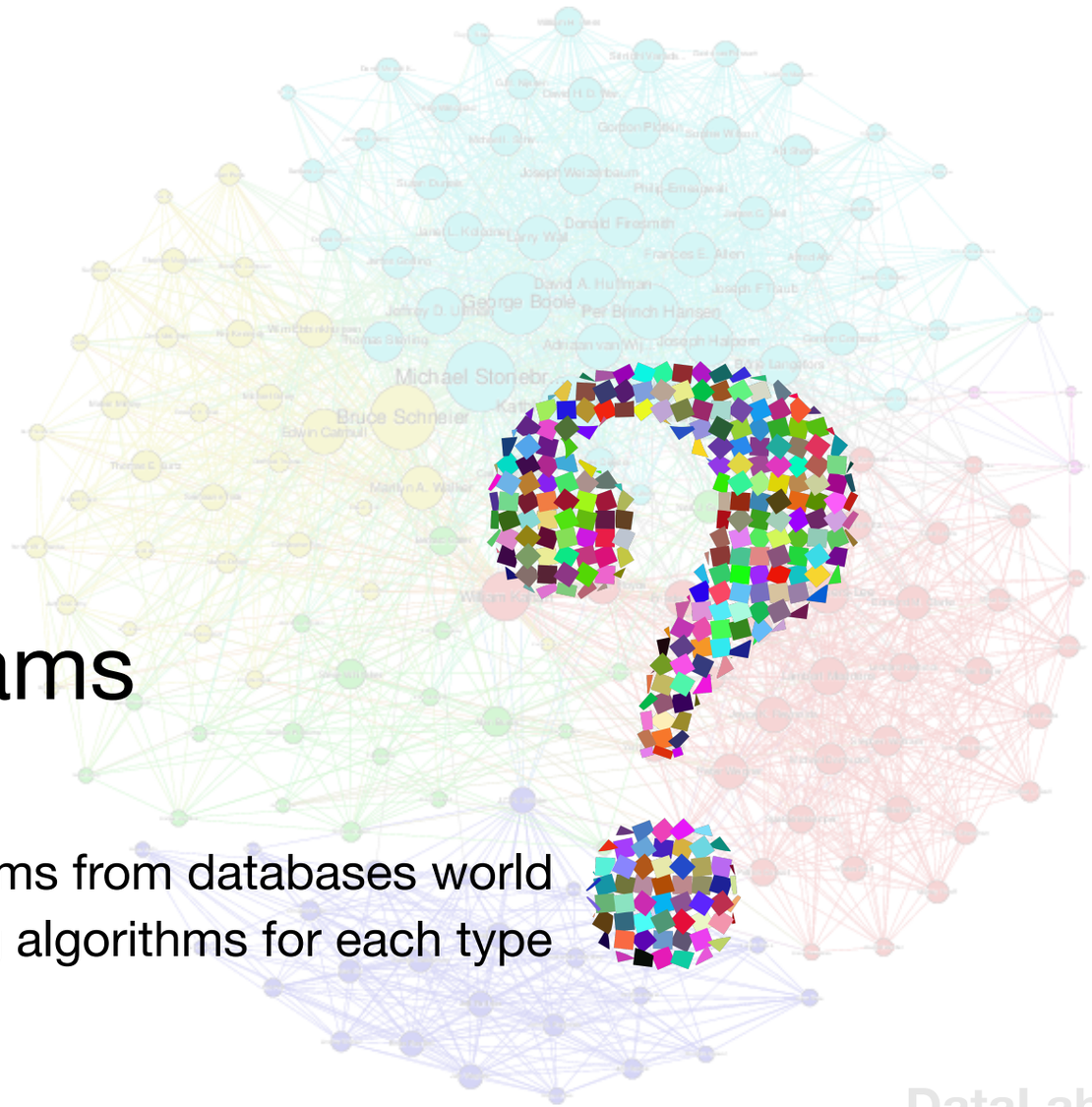
Data Classification Sample

- Classify goods by category



Data Mining Scalability

- Mining on Static Data

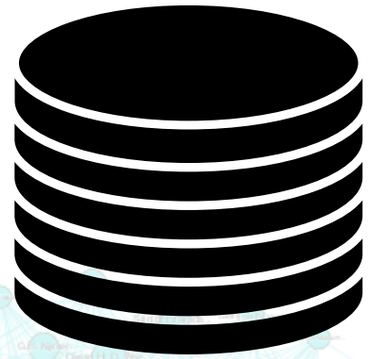


- Mining on Data Streams

Which are existing corresponding terms from databases world
Which are you knowing algorithms for each type

Data Mining Scalability

- Mining on Static Data

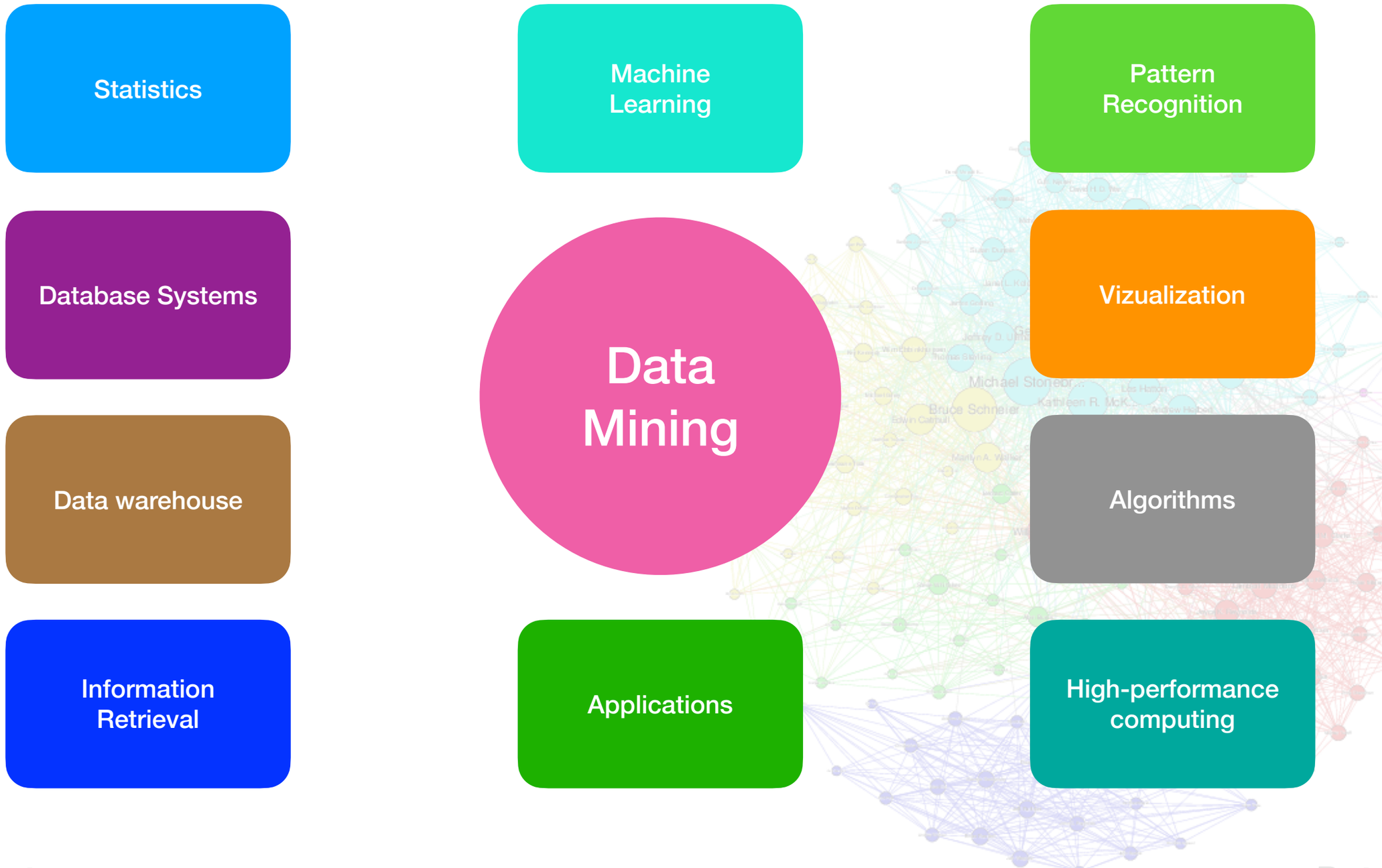


Data WareHouse

- Mining on Data Streams

- architecture

Data Mining


Statistics
Machine Learning
Pattern Recognition
Database Systems
Data Mining
Vizualization
Data warehouse
Algorithms
Information Retrieval
Applications
High-performance computing

Data Mining Sample Tasks

Store Product Placement

A merchant has a set of d products together with previous transactions from the customers containing baskets of items bought together. The merchant would like to know how to place the product on the shelves to increase the likelihood that items that are frequently bought together are placed on adjacent shelves.

Product Recommendations

A merchant has an (n x d) binary matrix representing the buying behaviour of n customers across d items. It is assumed that the matrix is sparse, and therefore each customer may have bought only a few items. It is desirable to use the product associations to make recommendations to customers.

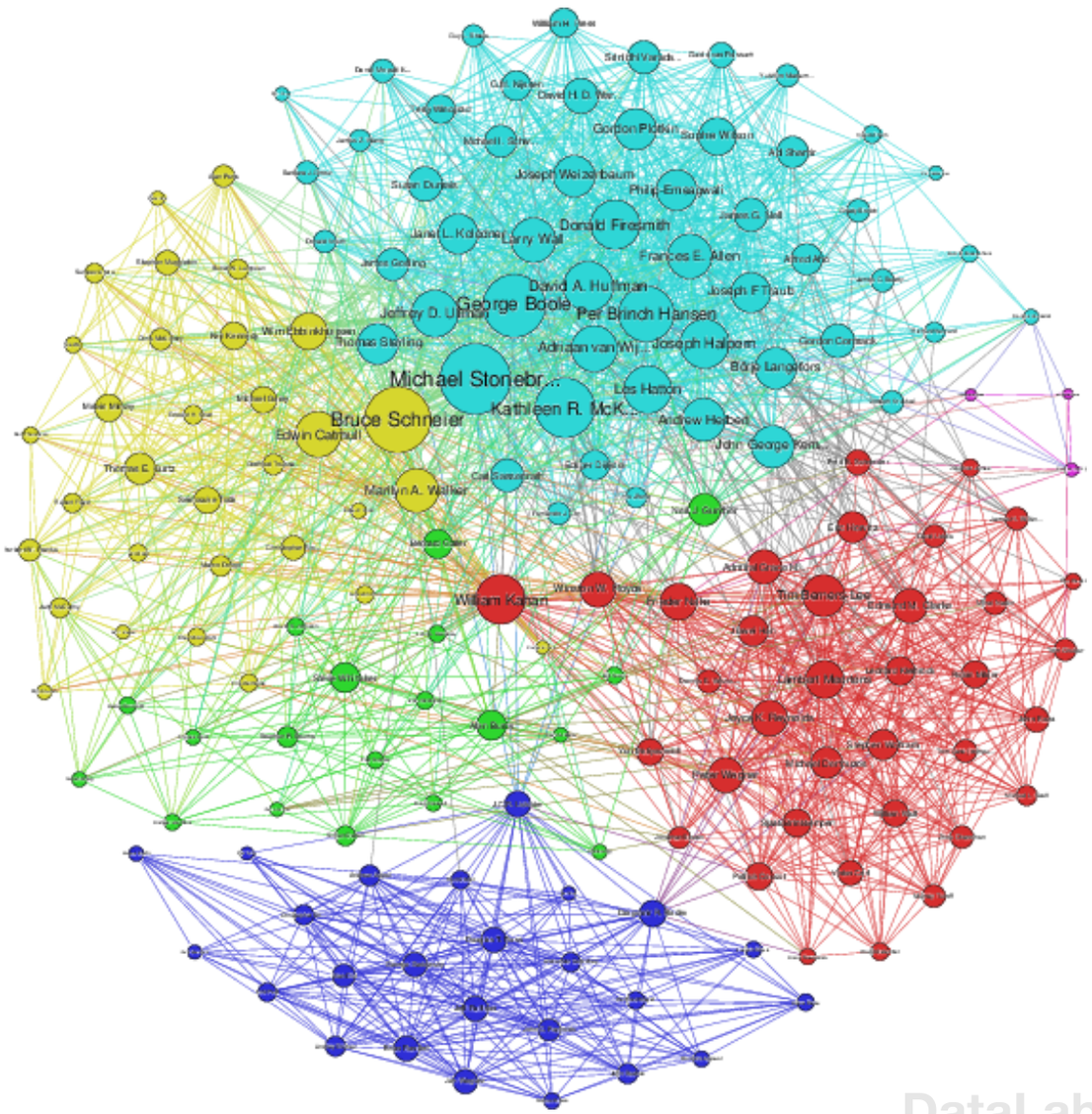
Medical Diagnosis

Consider a set of Medical metrics time series that are collected from different patients. It is desirable to determine the anomalous series from this set.

Web Log Anomalies

A set of Web logs is available. It is desired to determine the anomalous sequences from the logs.

Thanks!


Gordon Pask
Joseph Weizenbaum
Donald Firesmith
George Boole
Per Brinch Hansen
Michael Stonebr...
Bruce Schneier
Kathleen R. McK...
William Kahan