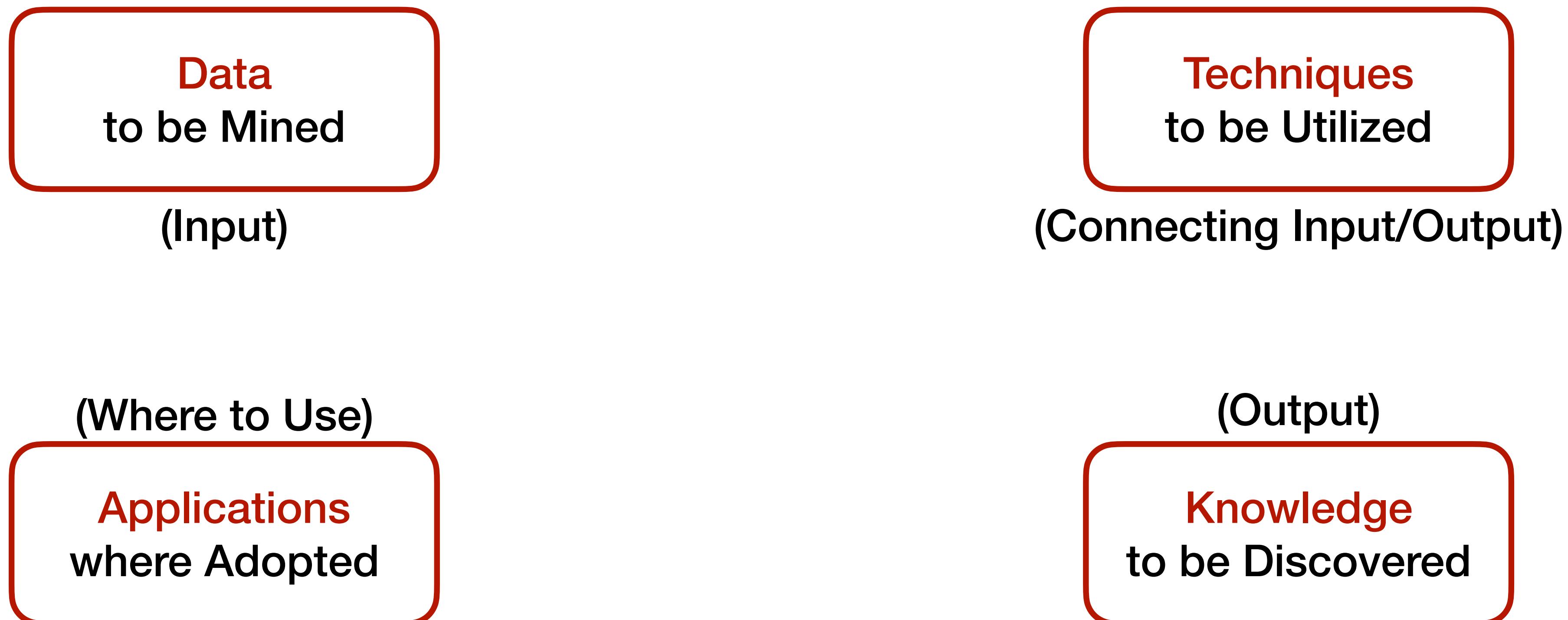


Data Representation Abstraction

INST447 Data Source and manipulation

Wei Ai
aiwei@umd.edu

Four Dimensions of Data Mining Problems



Towards Real-World Data

- We have learned how to collect, store, and process data with Python data structures and tools (e.g., arrays, lists, hash tables, Numpy, Pandas, databases)
- How can they be used in downstream tasks, such as data mining.
- There is a big gap between real data and analytics.
- Data representation bridges this gap.
 - A mathematical way to describe what data look like.

Real World Scenario I

The screenshot shows a product page for a book on Amazon. The top section displays a 'Frequently bought together' recommendation, featuring three books: 'Structure and Interpretation of Computer Programs', 'The Elements of Computing Systems: Building a Modern Computer from First Principles', and 'The Algorithm Design Manual'. Below this, a customer review by 'BH' is shown, rating it 3 stars and stating 'High hopes but a little disappointed'. The review discusses the book's content and code examples. The bottom section shows a 'Customers who bought this item also bought' recommendation, listing 'The Elements of Computing Systems: Building a Modern Computer from First Principles' and 'The Pragmatic Programmer'. A second customer review by 'Sean Lawrence' is displayed, rating it 5 stars and mentioning receiving counterfeit copies.

Frequently bought together

BH **High hopes but a little disappointed**
October 7, 2015
Format: Paperback | Verified Purchase

I had high hopes for this book given the reviews here. I bought this book brush up on my existing knowledge and learn new things and was very excited to receive it. There are some interesting points and I liked that it has a section on heuristics. It introduced me to the idea of simulated annealing, which I am using for a problem at work now actually, but like most things in the book I had to turn to the internet for a better explanation. The book is a bit wordy and hard to read at times.

The other major gripe I had is with the code examples. Most variables and method parameters are named a, k, c .. etc. For instance here is a method header in the backtracking section: `construct_candidates(int a[], int k, int n, int c[], int *ncandidates)`

This makes the code very hard to read and follow as well when you are a lines deep in a method trying to figure out what `k[i] = c + a` is. It has been my experience that most math/engineer types and professors tend to write code this way vs giving things descriptive names so the code is easily understandable. The book would be 10X more valuable had the code samples been clearer. I ended up sending it back.

100 people found this helpful

Helpful | 1 comment | Report abuse

Customers who bought this item also bought

Sean Lawrence **Great book but I received counterfeit copies**
July 15, 2017
Format: Hardcover | Verified Purchase

Buyer beware! Amazon appears to be selling counterfeit textbooks. I have notified Amazon of this problem and re purchased and still received what I believe to be counterfeits due to the very poor quality of the color, binding, printing, and horrible misspellings on the back of the book in particular.

The contents of the book is great BTW. However I'd recommend buying direct from publisher.

50 people found this helpful

Helpful | 2 comments | Report abuse

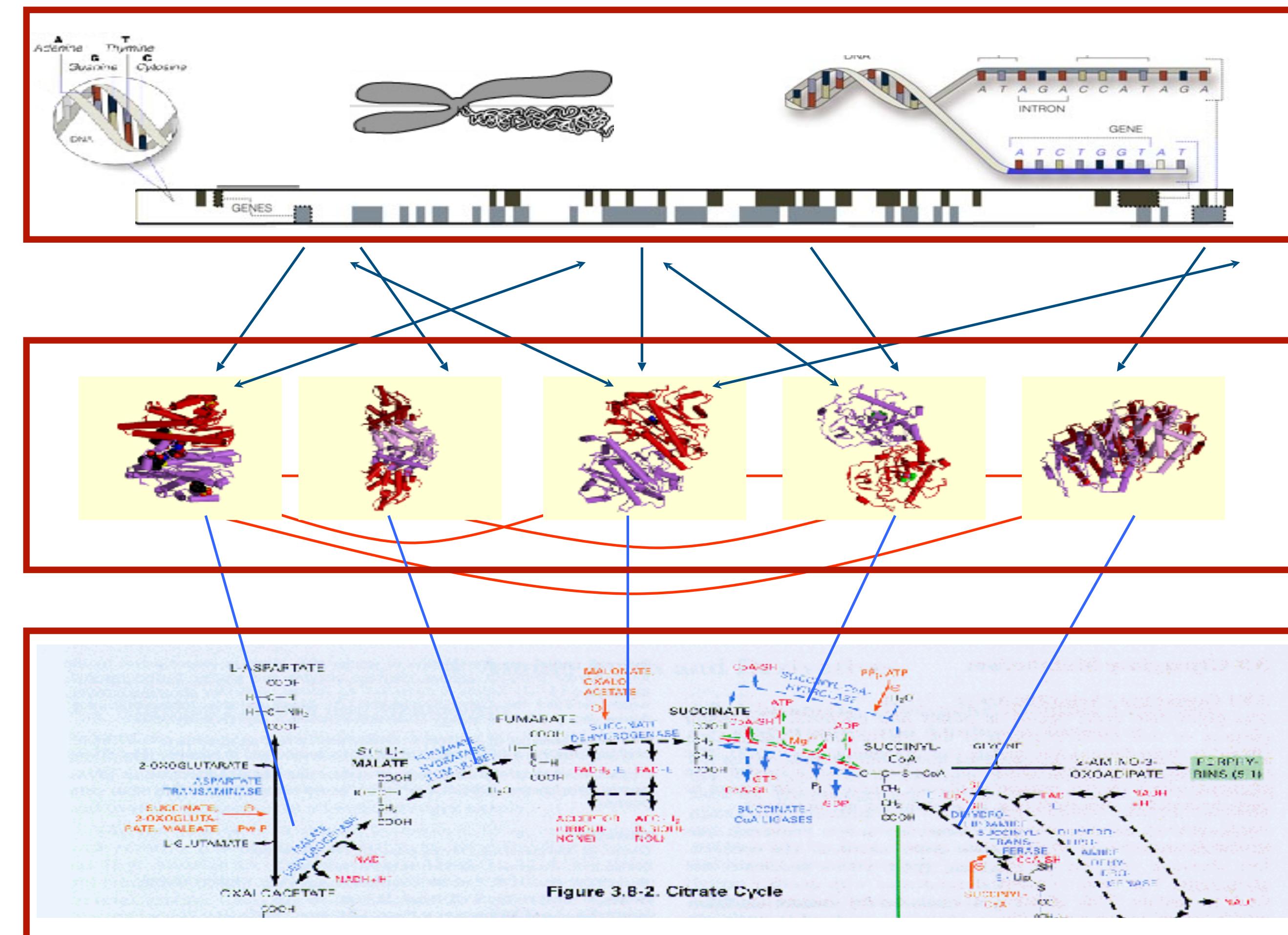
\$35.00 **prime**

Hardcover \$40.18 **prime**

Page 1 of 13

How do we generate recommendations from such complex and messy information?

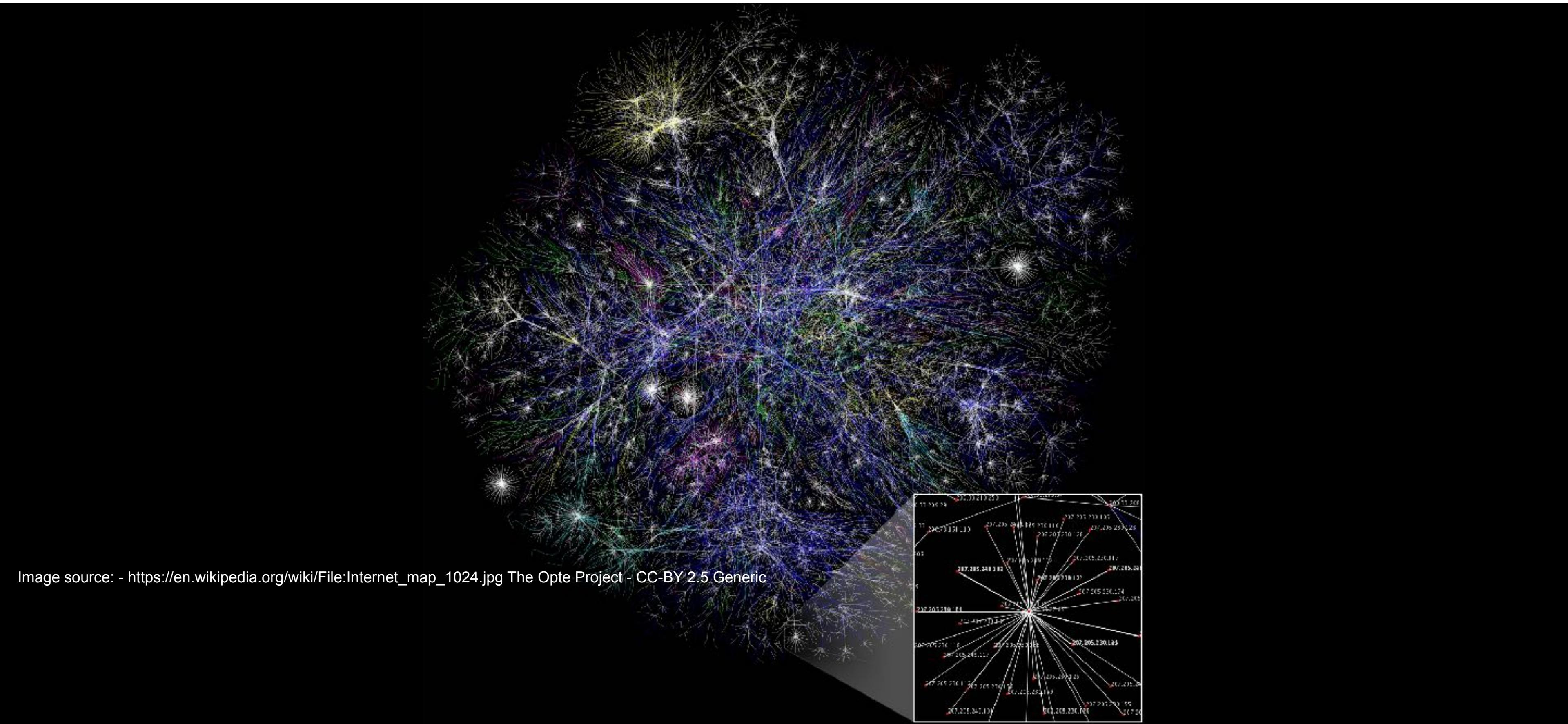
Real World Scenario II



What python data structure would you use to load this complex data?

Image source: Albert-László Barabási's SIGIR09 keynote

Real World Scenario III

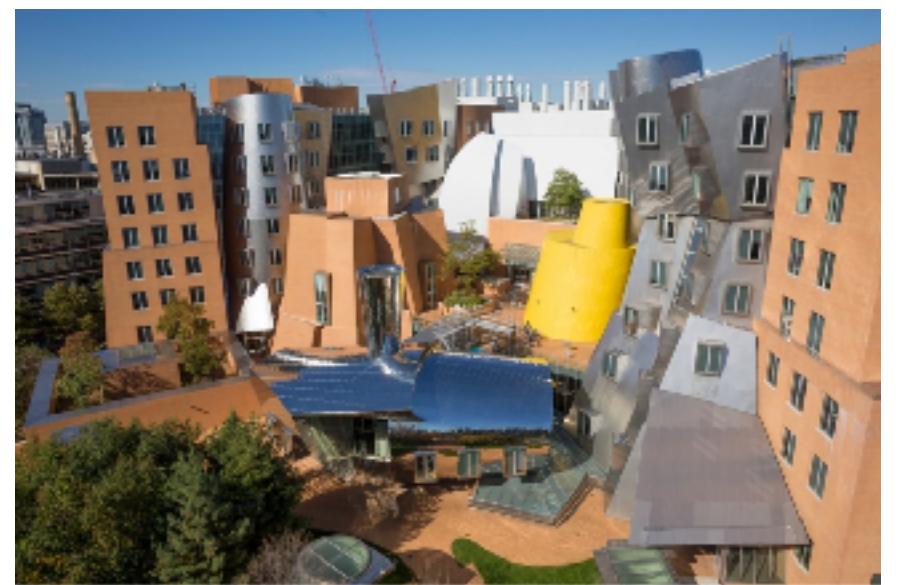


Challenge of Real Data

What we are used to:



What the reality is:



So, how to deal with the wild thing?

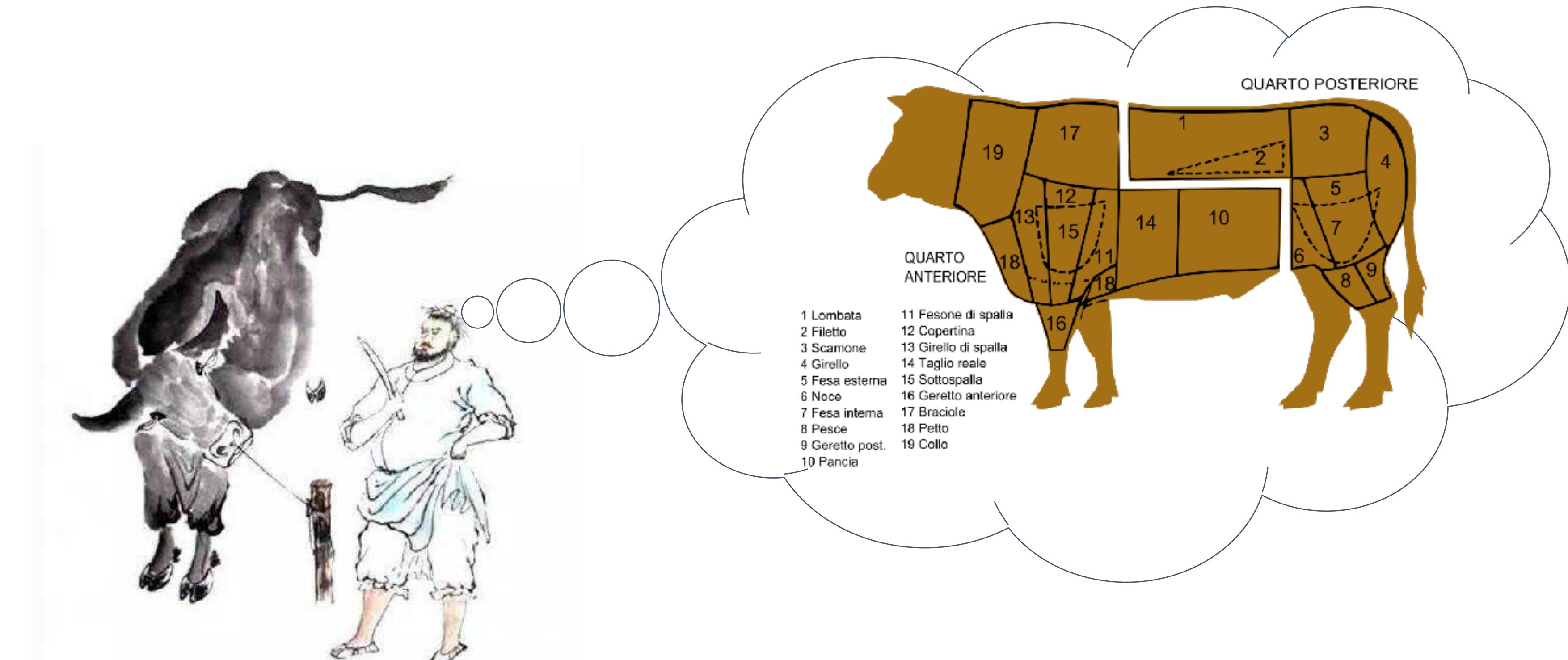


Volume, Variety, and Velocity

Data Formulation

- There are more data science applications than you may expect.
- But there aren't so many basic data types.
- How shall we abstract, formulate, represent the data in real applications?
- Data formulation is usually the first task of data mining

The Skillful Butcher (Story by Chuang-Tzu, ~300 BC)



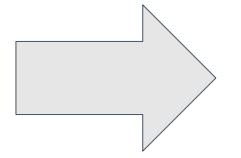
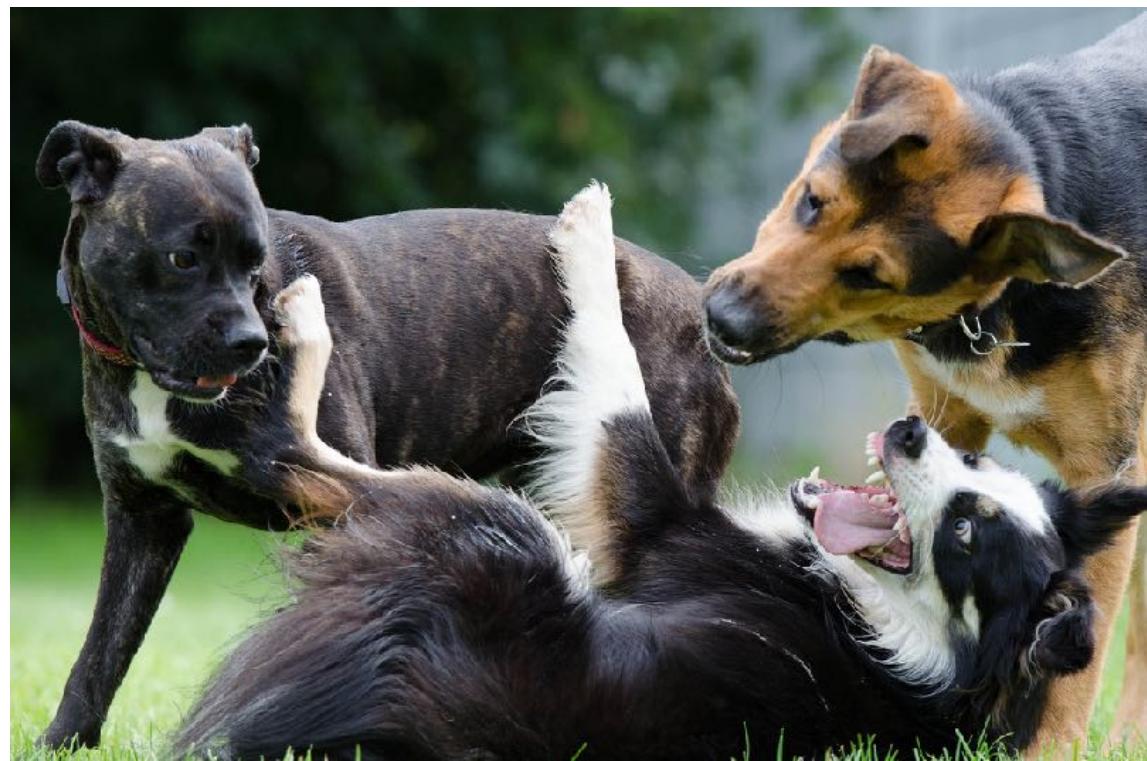
What does a Data Scientist See?

- What is a basic **object** of information?
- What are the **properties/attributes** of the data object?
- How are the **attributes** structured?
- How to assign **values** to the attributes?
- How are different data objects **related**?

Data scientists must be able to answer these questions in a mathematical way

Representation of data allow data scientists to answer in a mathematical way

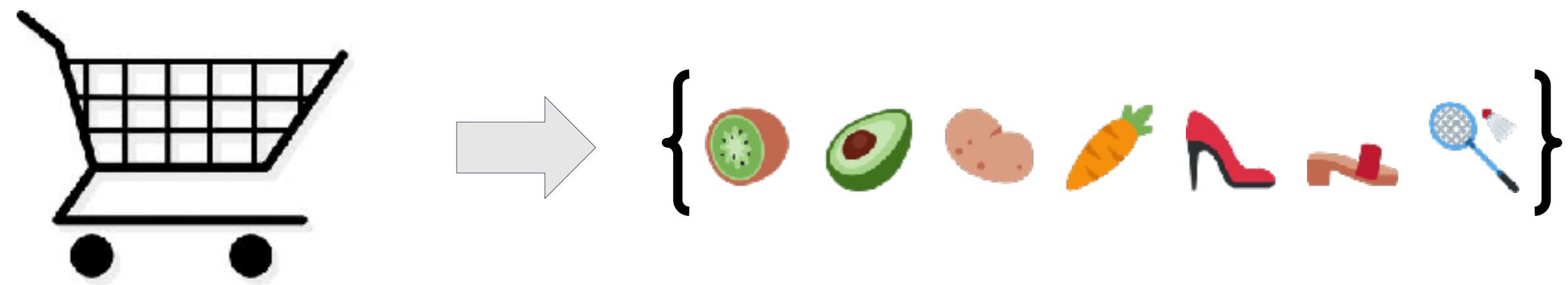
Data Representations



- Item Set
- Vector / Matrix
- Sequence
- Time Series
- Spatial
- Spatiotemporal
- Graph / Network
- Stream

Itemset Data

- Data Object: a shopping basket, a piece of text, a board of directors, ...



- Attribute: **appearance** of a categorical item
 - a product, a word, a person, etc.

The Itemset Representation

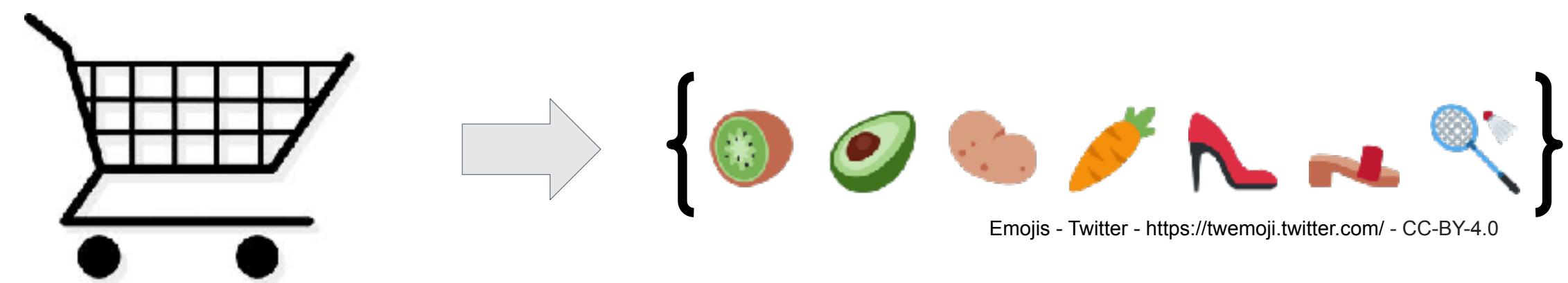
- Each data object is represented as a set of items:

$$X = \{x_1, x_2, \dots, x_k\}$$

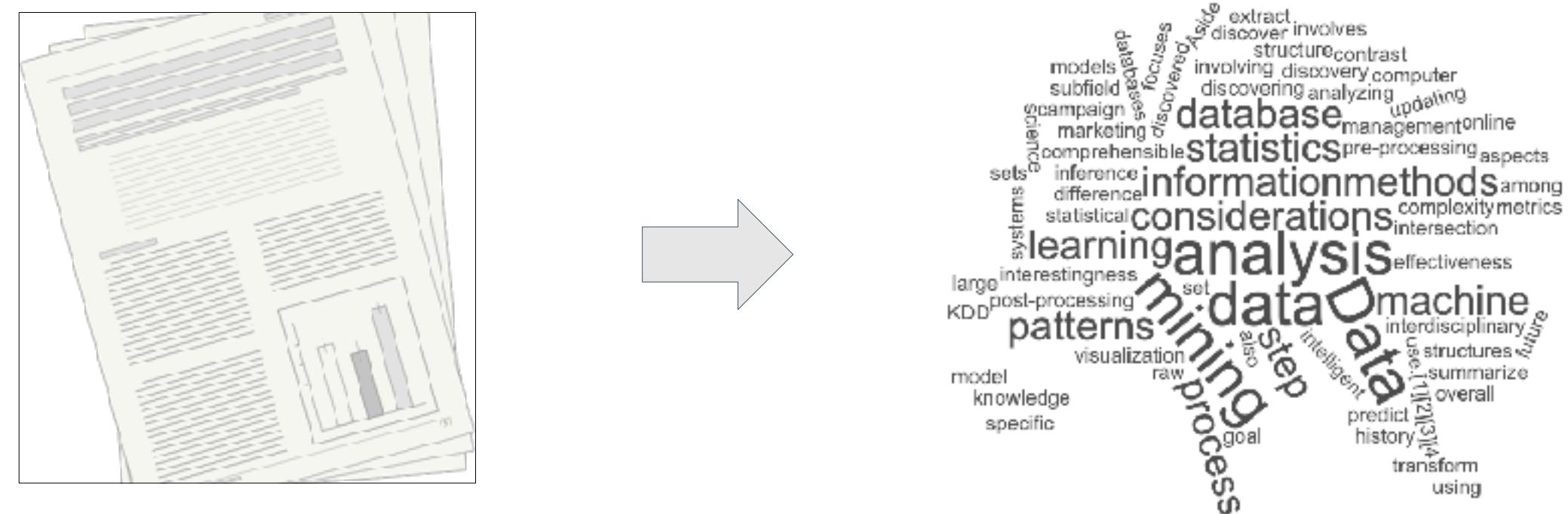
- x_i belongs to X if and only if that **categorical item** appears in the set.
- Order or counts of the items don't matter

Example of Itemsets

- # • Shopping Baskets:

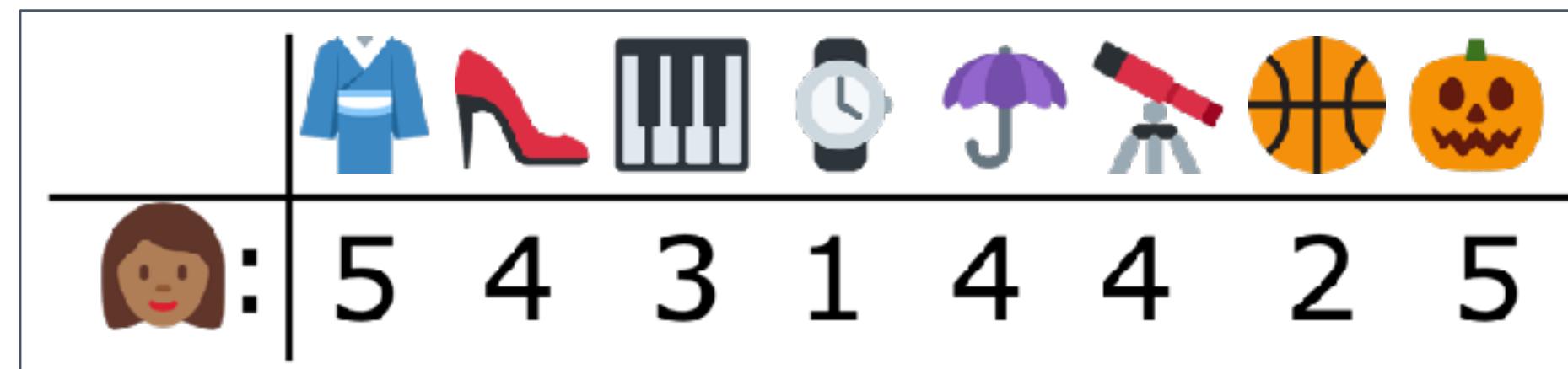


- Text (as bag-of-words):



Vector Data

- **Data Object:** E.g., a user's ratings of products, or course grades of a student



- **Attribute:** a **numerical** property of the object.
- E.g., Kimono=5; Shoe=4; Piano=3, etc.

The Vector Representation

- Data represented as n-dimensional vectors:
 - Each dimension corresponds to one attribute

$$\vec{X} = \langle x_1, x_2, \dots, x_n \rangle$$

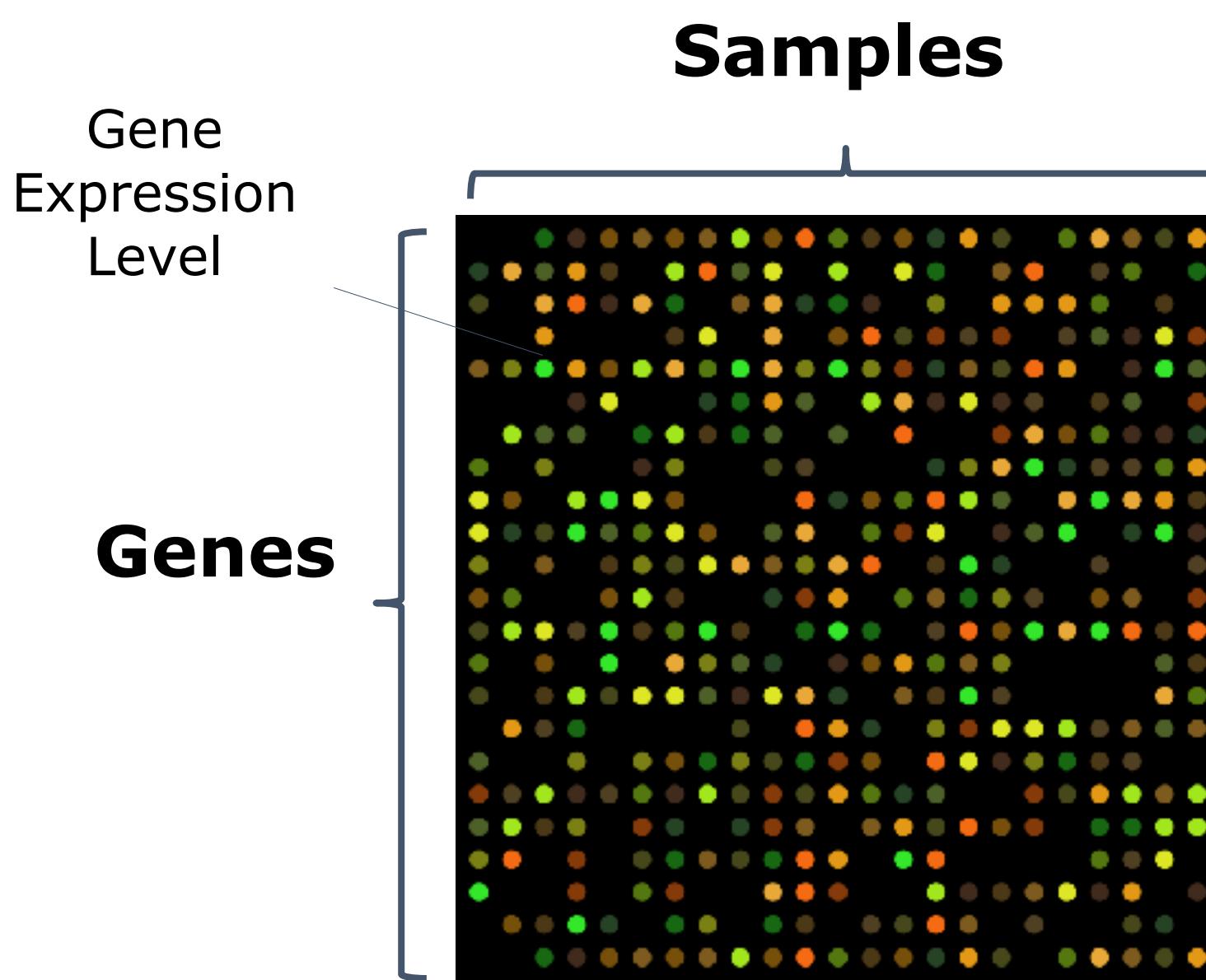
- x_i is the **numerical value** of X at the i -th dimension (attribute).
- Each attribute is unique; cannot change order.
- Multiple objects → a **matrix** (a collection of vectors).

Example of Matrices

Product Ratings

	💻	🛍️	🎧	🎮
A	✓	✗	✓	✓
B	✓	✓	✗	✗
C	✓	✓	✗	✗
D	✗		✓	
E	✓	✓	?	✗

Microarrays



Sequence Data

- Data object: a DNA sequence, a session of search queries, a sentence (of words), a trace of user actions
- Attributes: pairs of **positions** and **categorical** item, in a sequential order

The Sequence Representation

- Data represented as a **sequence of items**:

$$X = \{(x_1, 1), (x_2, 2), \dots, (x_k, k)\}$$

- x_i is the categorical item appeared at the i -th position of X .

Example of Sequences

DNA sequences

Unaligned sequences

Human	ACAT	TATGGACAGGTAAAGTAAAAAACATATT
Chimpanzee	ACAT	TATGGACAGGTAAAGTAAAAAACATATT
Macaque	ATATA	CATTACGGACAGGTAAAGTAAAAAACAT

Aligned sequences

Human	ACA	TTATGGACAGGTAAAGTAAAAAACATATT
Chimpanzee	ACA	TTATGGACAGGTAAAGTAAAAAACATATT
Macaque	ATATA	CATTACGGACAGGTAAAGTAAAAAACAT

Search sequence

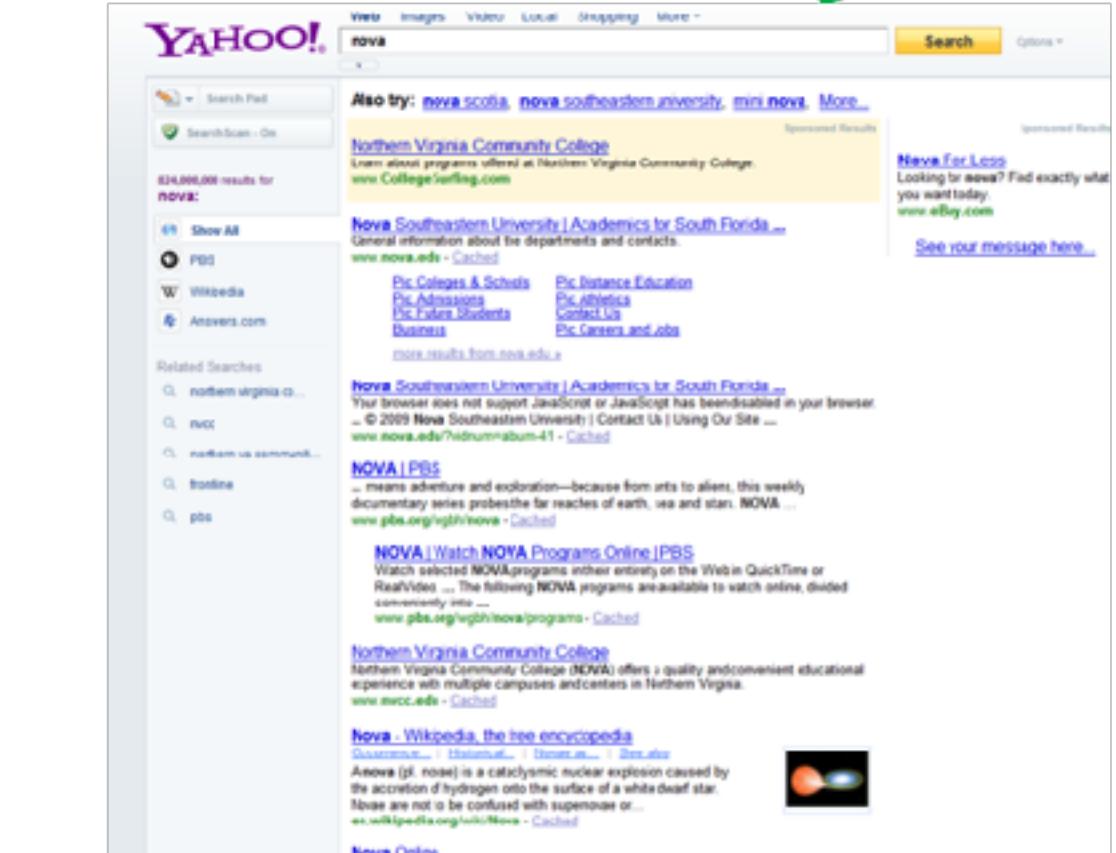
mustang

[www.fordvehicles.com/
cars/mustang](http://www.fordvehicles.com/cars/mustang)

ford mustang

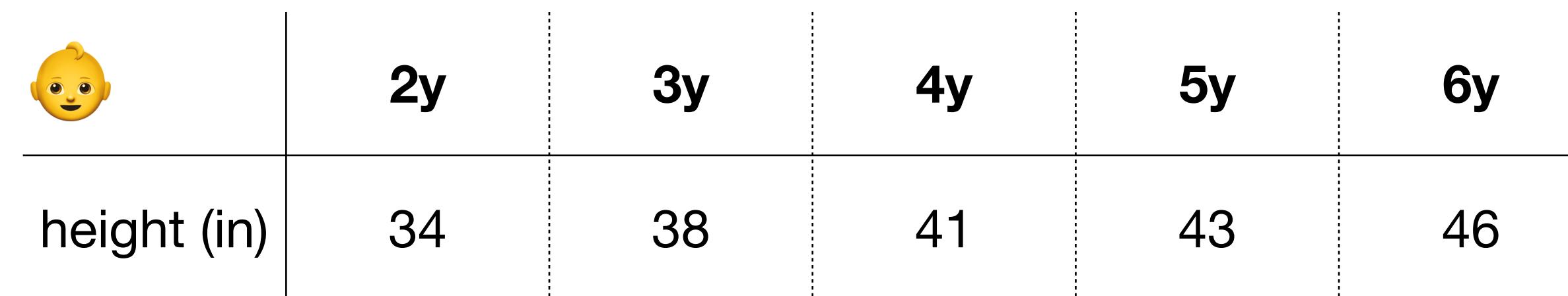
[en.wikipedia.org/wiki/
Ford_Mustang](http://en.wikipedia.org/wiki/Ford_Mustang)

Nova



Time Series Data

- Data Object: growth chart, stock price over time, battery life over time



- Attribute: the measurement of a (**numerical**) property observed at a given time point

The Time Series Representation

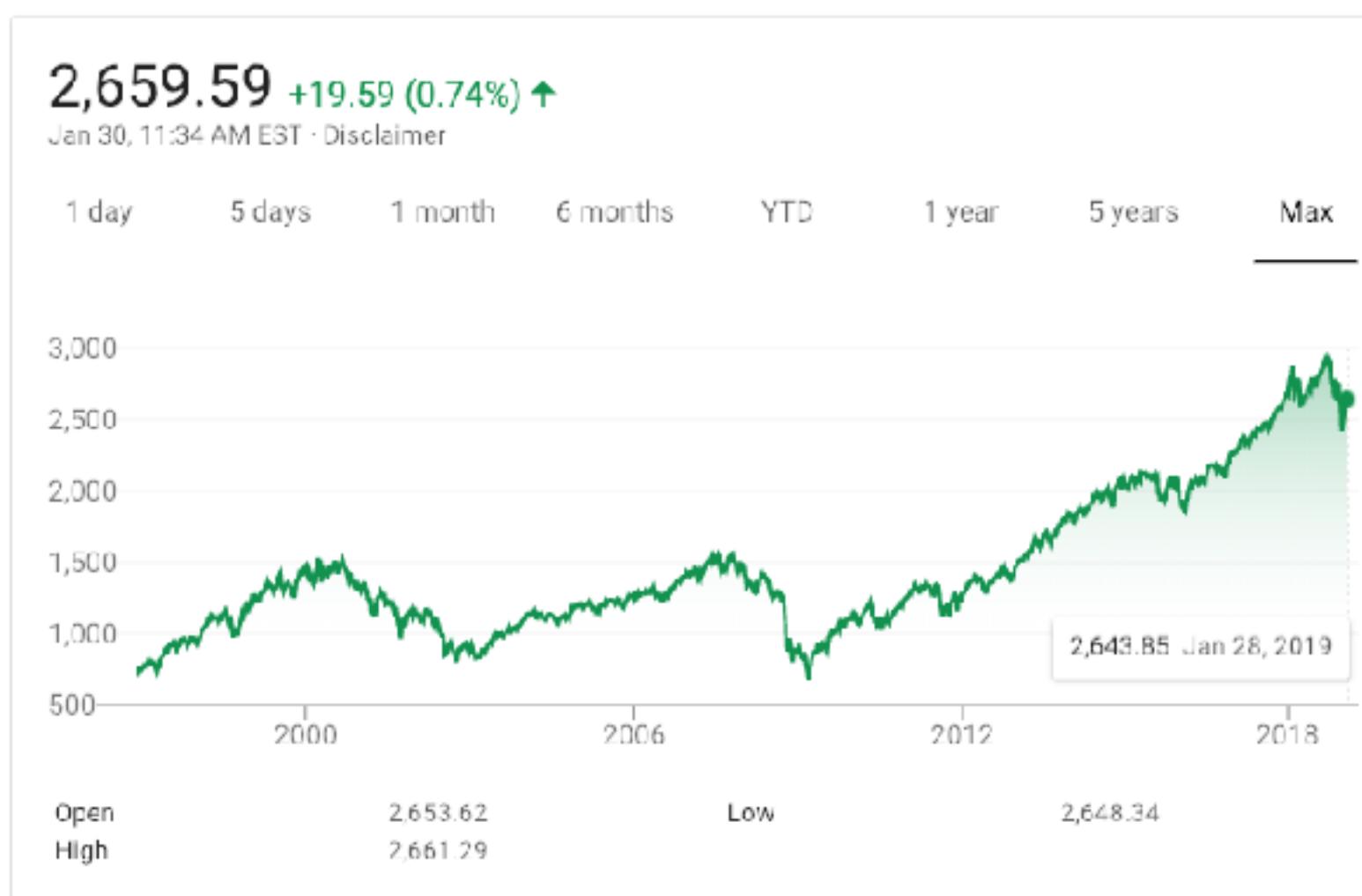
- A list of timestamped measurements:

$$X = \{(x_1, t_1), (x_2, t_2), \dots, (x_n, t_n)\}$$

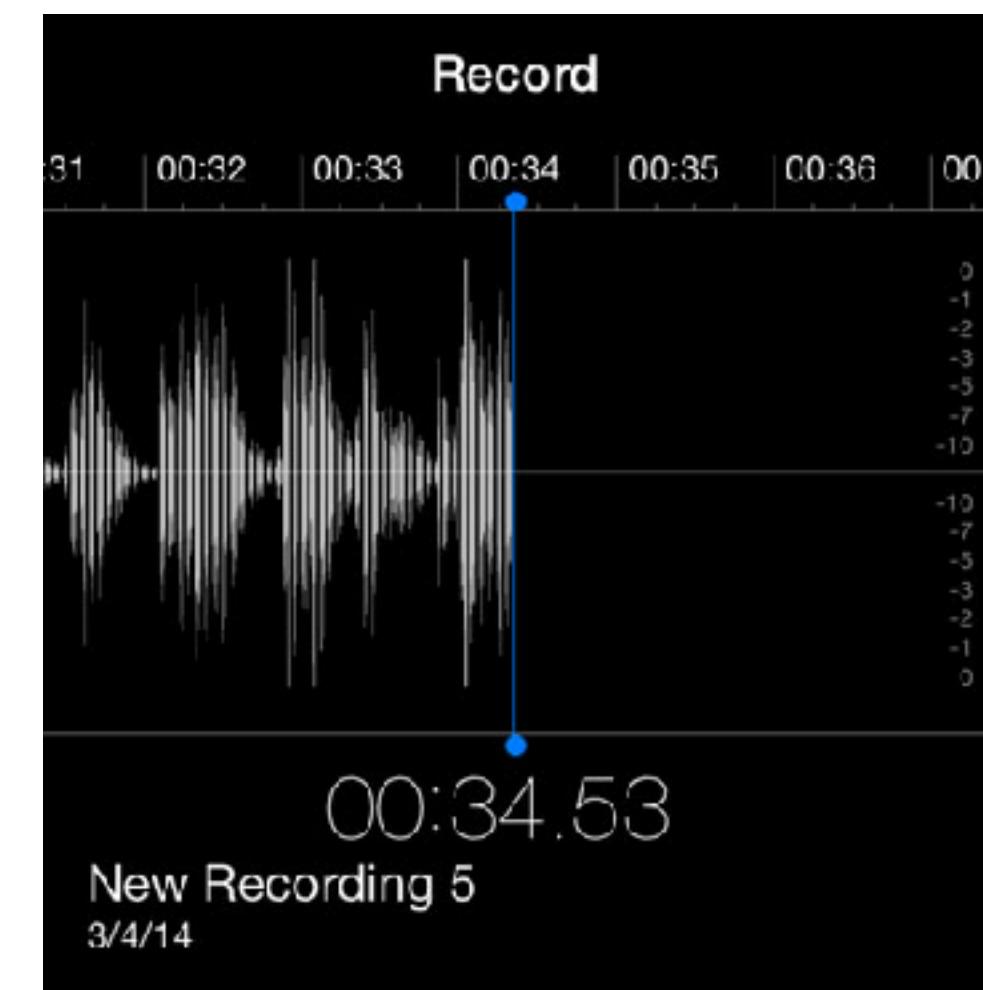
- x_i is the (numerical) measurement of a property of X observed at time stamp t_i .
- Alternative representation: $x = f(t)$

Examples of Time Series

Stock Market (SP 500)



Voice / Speech data



Spatial/Spatiotemporal Data

- Data Object: GPS trajectory of a vehicle, spread of a disease, a heat map
- Attribute: measurement of a (**numerical**) property at a given location is **spatial data**.
- If measurement also includes a given time point, it is **spatiotemporal data**.

Spatial/Spatiotemporal Data

- 1D = just name of location
- 2D = longitude and latitude
- 3D = longitude, latitude, altitude

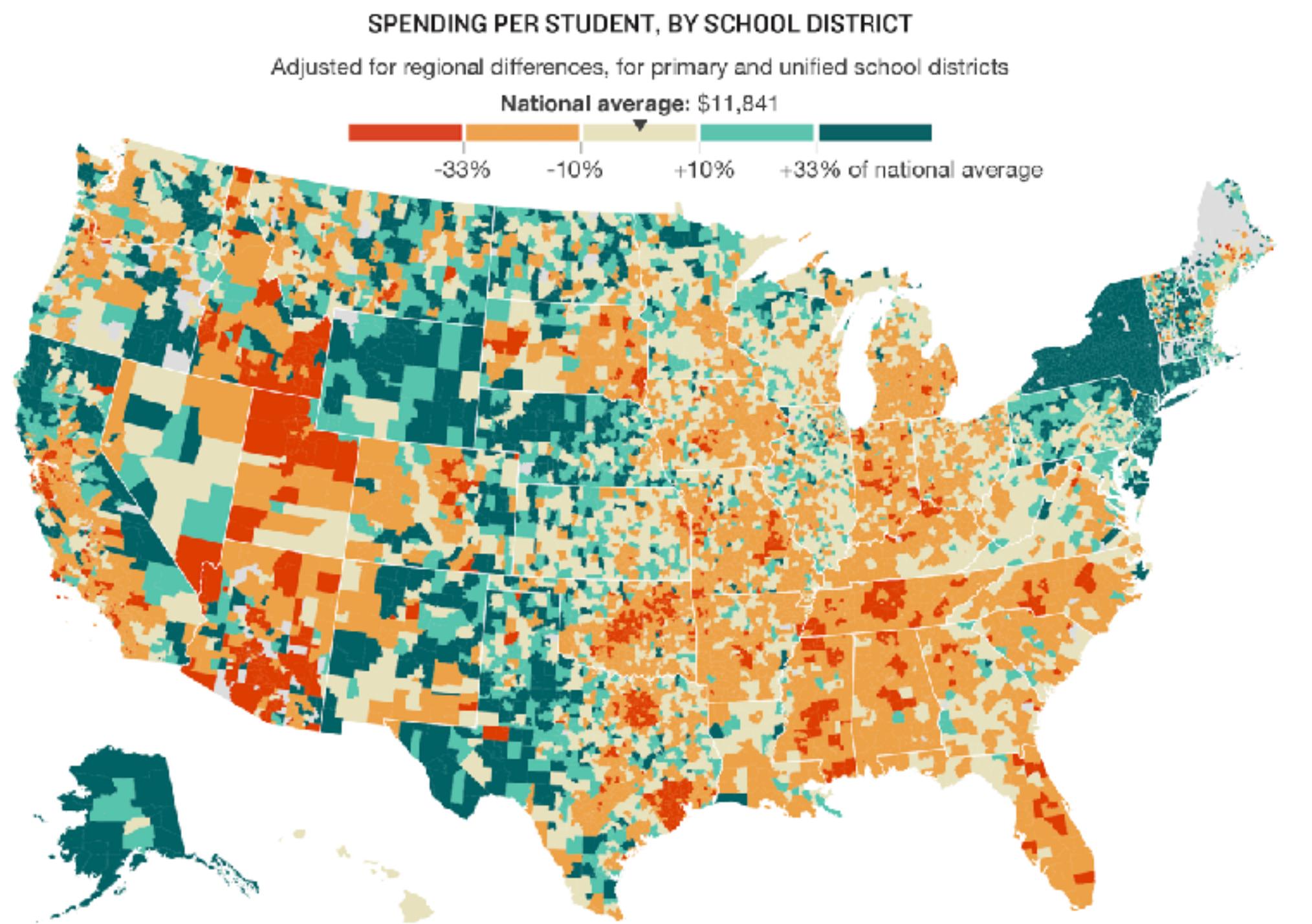
The Spatial/Spatiotemporal Representation

- List of location-labeled measurements (2D):

$$X = \{(x_1, \underset{\text{longitude}}{\lambda}_1, \underset{\text{latitude}}{\phi}_1), (x_2, \lambda_2, \phi_2), \dots, (x_n, \lambda_n, \phi_n)\}$$

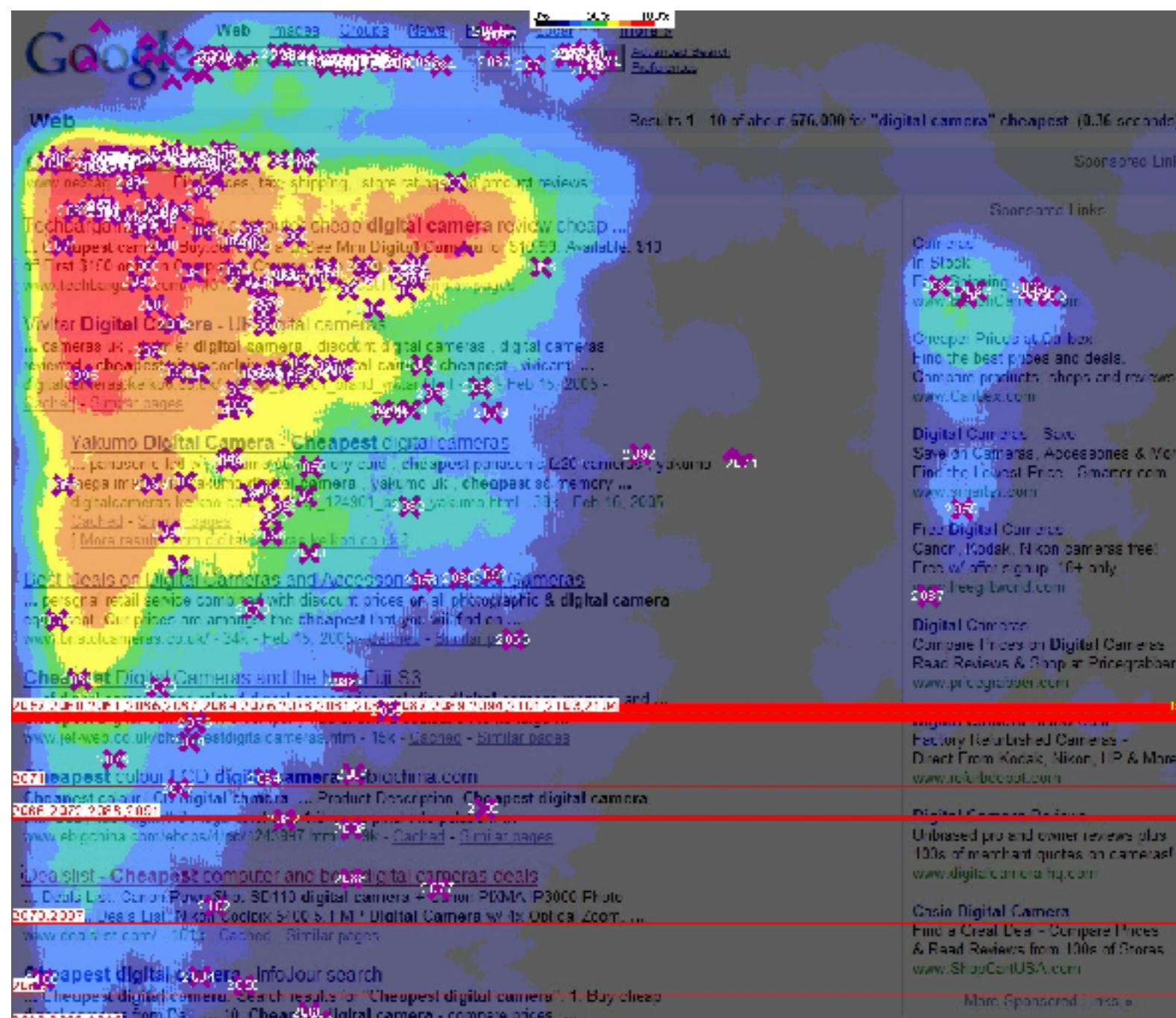
- Alternative Representation (2D): $x = f(\lambda, \phi)$

Examples of Spatial Data



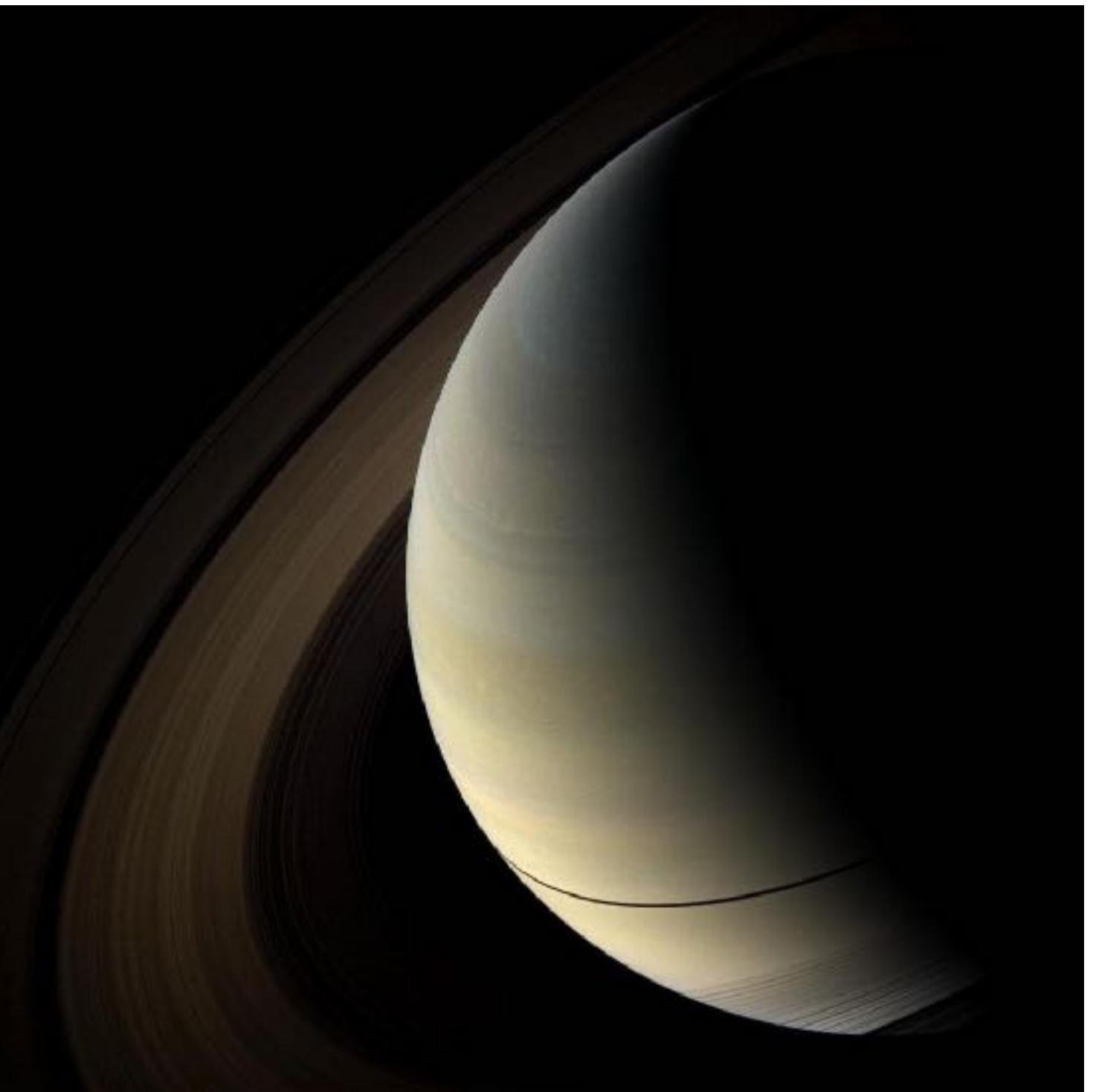
Examples of Spatial Data

- Average time users' eyes are fixed at locations on a Web page.
- Locations are indicated by horizontal (X) and vertical (Y) positions.



Examples of Spatial Data

- 1000 x 1000 px image:
 - X: 460
 - Y: 137
- RGB values:
 - R: 151
 - G: 149
 - B: 138

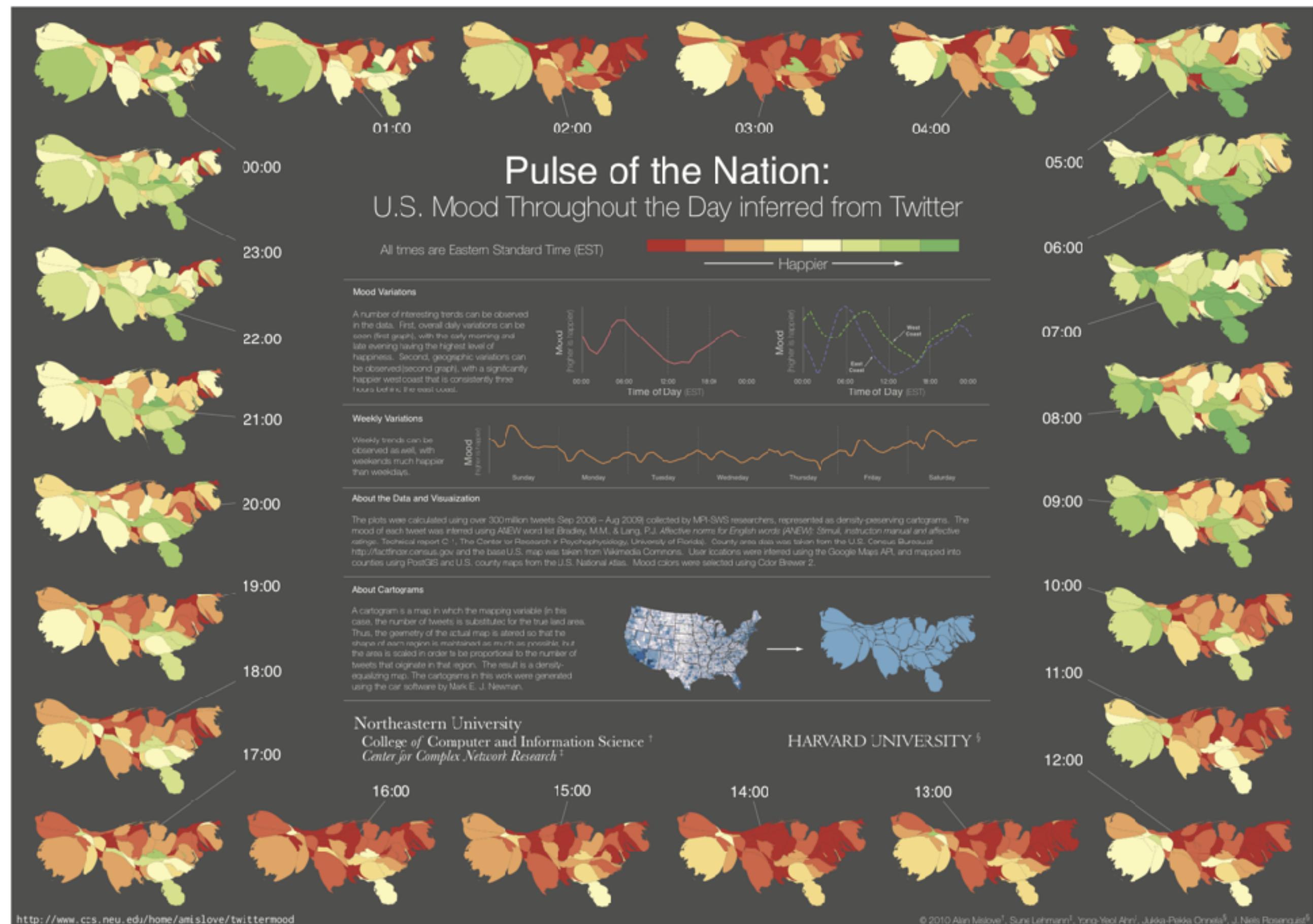


Spatiotemporal Data

- Simply add the **time dimension** to a spatial representation to describe spatiotemporal data.

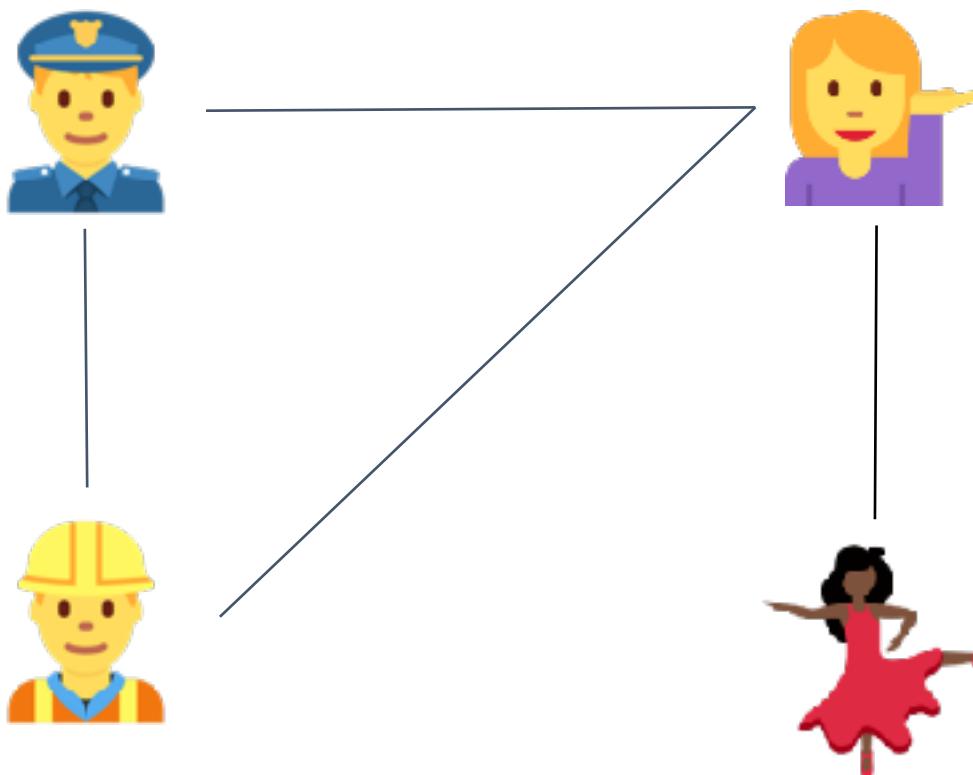
$$X = \{(x_1, \lambda_1, \phi_1, t_1), \dots\}$$
$$x = f(\lambda, \phi, t)$$

Example of Spatiotemporal Data



Graph (Network) Data

- Data objects: an online social network, the Internet, the Web
- Attribute: **nodes** and **links**



The Graph (Network) Representation

- Data formulation: $G(V, E)$
- V is a set of nodes (vertices, entities)
 - A node can be a categorical item or a complex data object.

The Graph (Network) Representation

- Data formulation: $G(V, E)$
- V is a set of nodes (vertices, entities):
 - A node can be a categorical item or a complex data object.

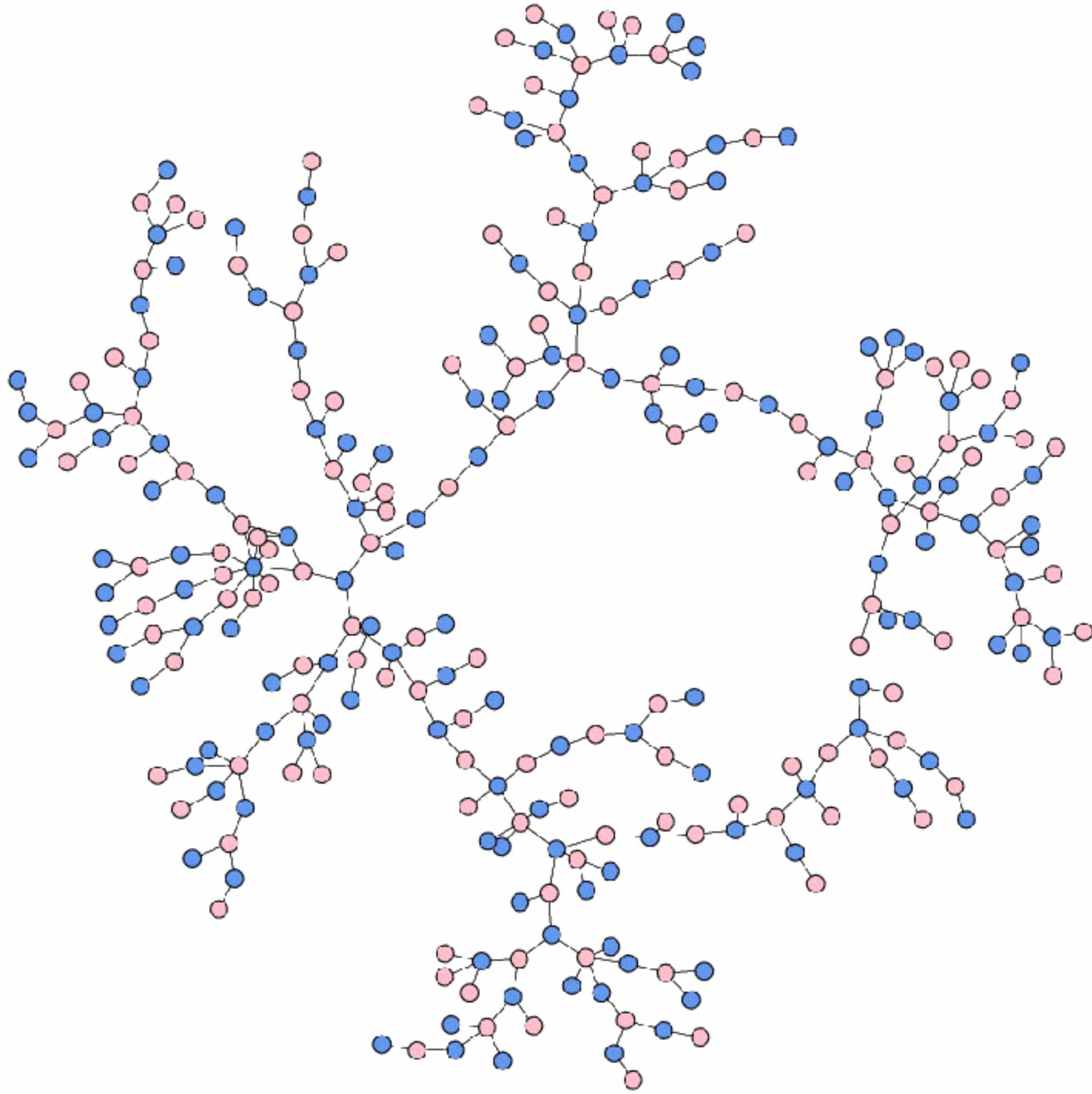
$$V = \{v_1, v_2, \dots, v_n\}$$

- E is a set of links (edges, relations) between two nodes:

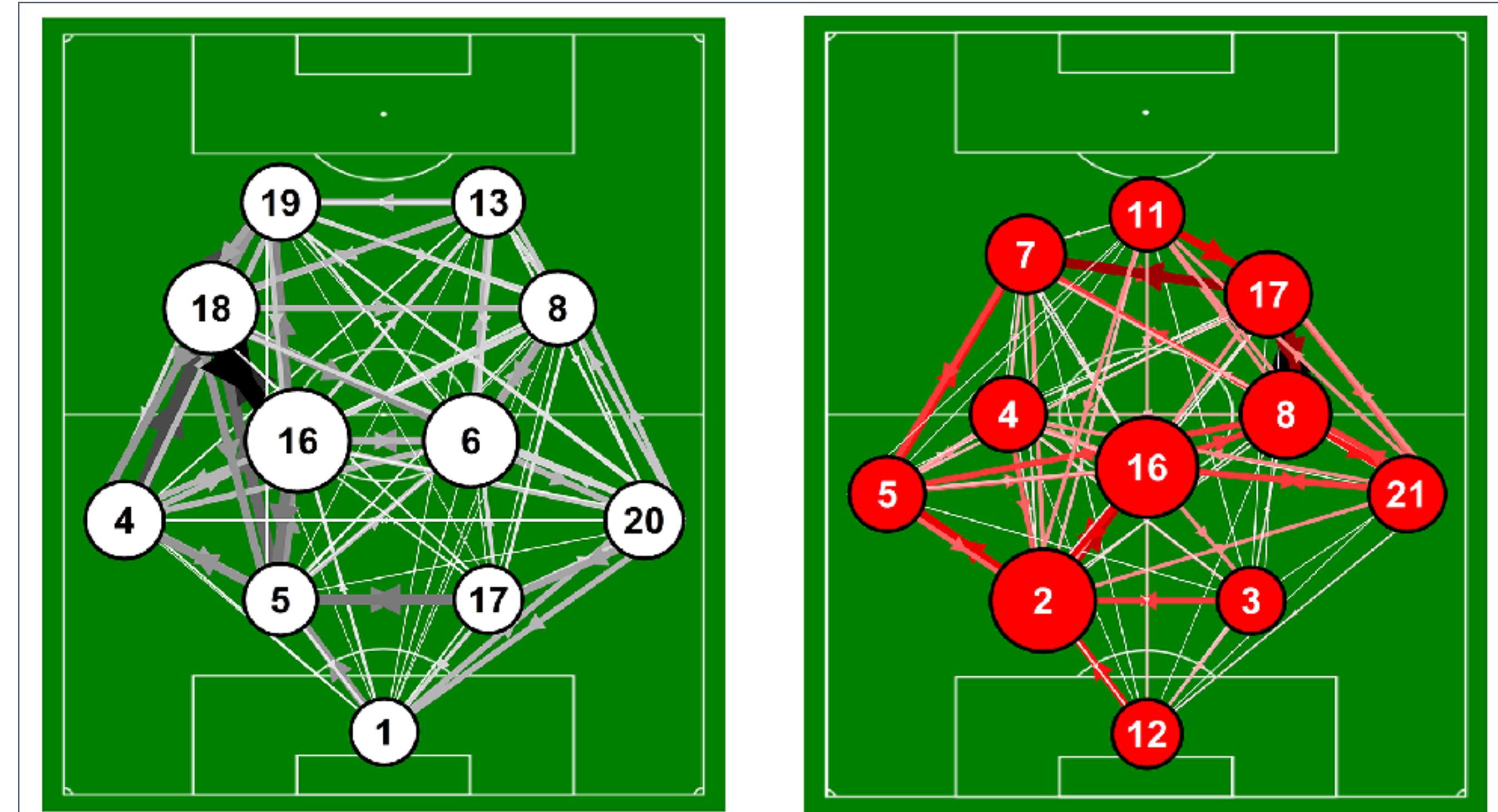
$$E = \{(v_i, v_j), \dots\}$$

Examples of Networks

High school
dating network



Examples of Networks



Soccer passing network

Stream Data

- Objects arrive with continuous time stamps
 - Example: Email inbox, news feeds
- Data objects: emails, network packages
- Attributes: arrival time (or order) as one specific attribute

The Stream Representation

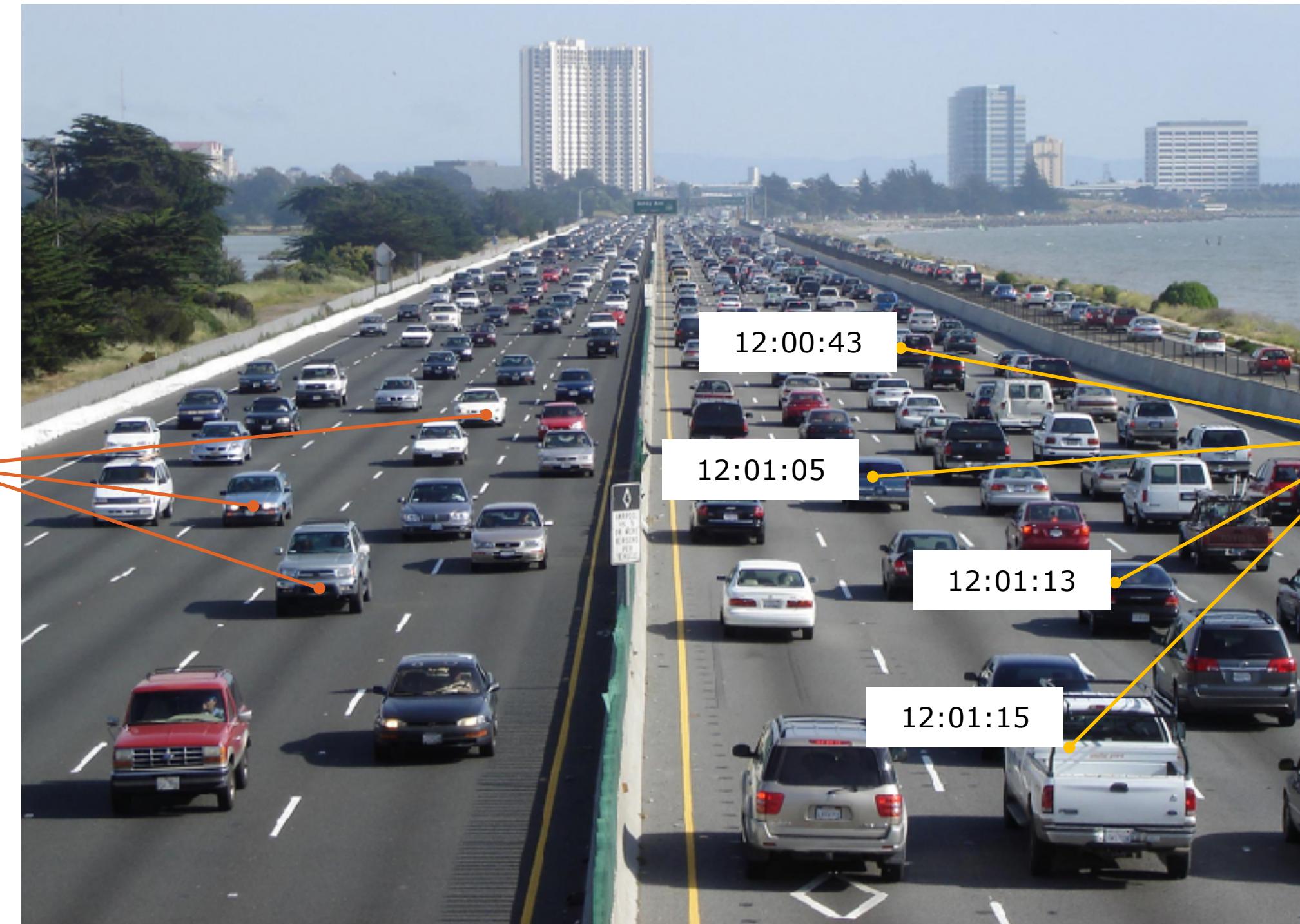
- Formulation of Data ($t_k \leq t_{k+1} \leq t_{k+2} \dots$):

$$D = \{..., (X_k, t_k), (X_{k+1}, t_{k+1}), ..., (X_n, t_n), \dots\}$$

- X_k can be any **simple or complex** data object.

Examples of Data Streams

Each vehicle
is an object
of this view
stream

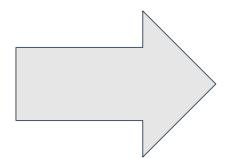
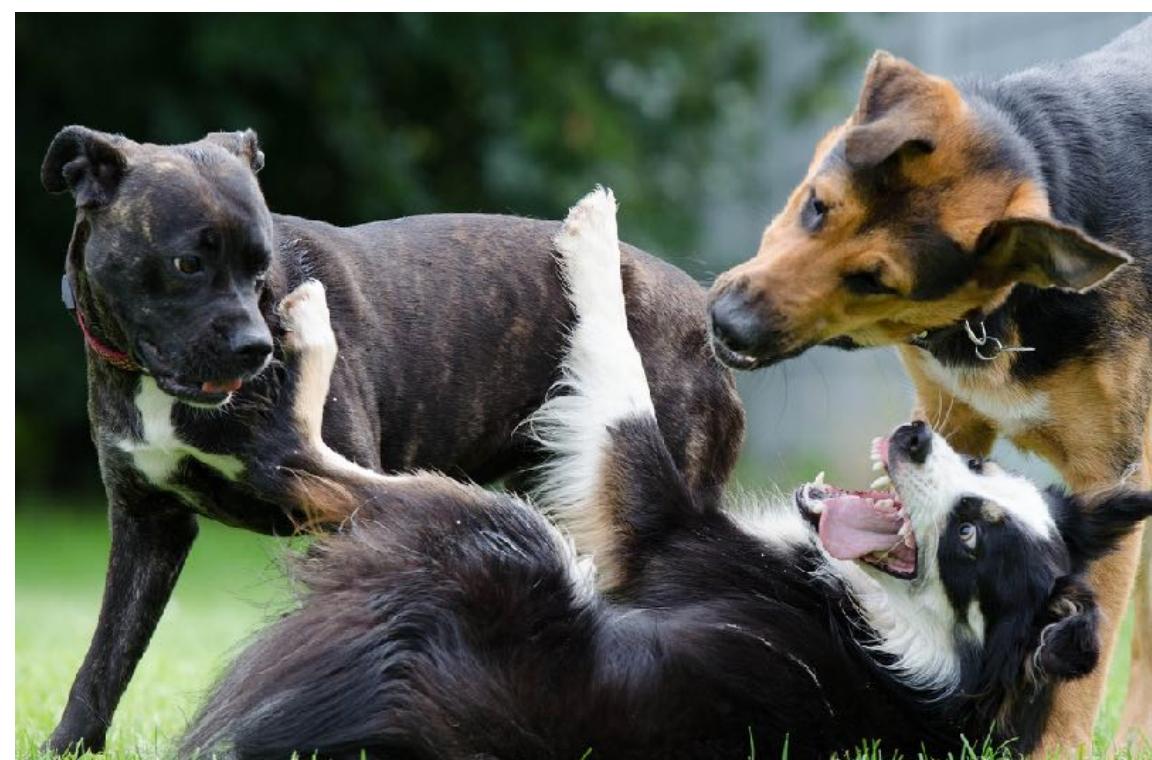


Arrival attribute
identified by the
time a car
appeared in the
view stream

Different Representations for “To be or not to be”

- As bag of words: {to, be, or, not}
- As vector of words: $\langle 2, 2, 1, 1 \rangle$
- As sequence of characters:
 - t → o → _ → b → e → _ → o → ...
- As sequence of words:
 - to → be → or → not → to → be
- As network of words:
 -
- Choice of representation depends on task

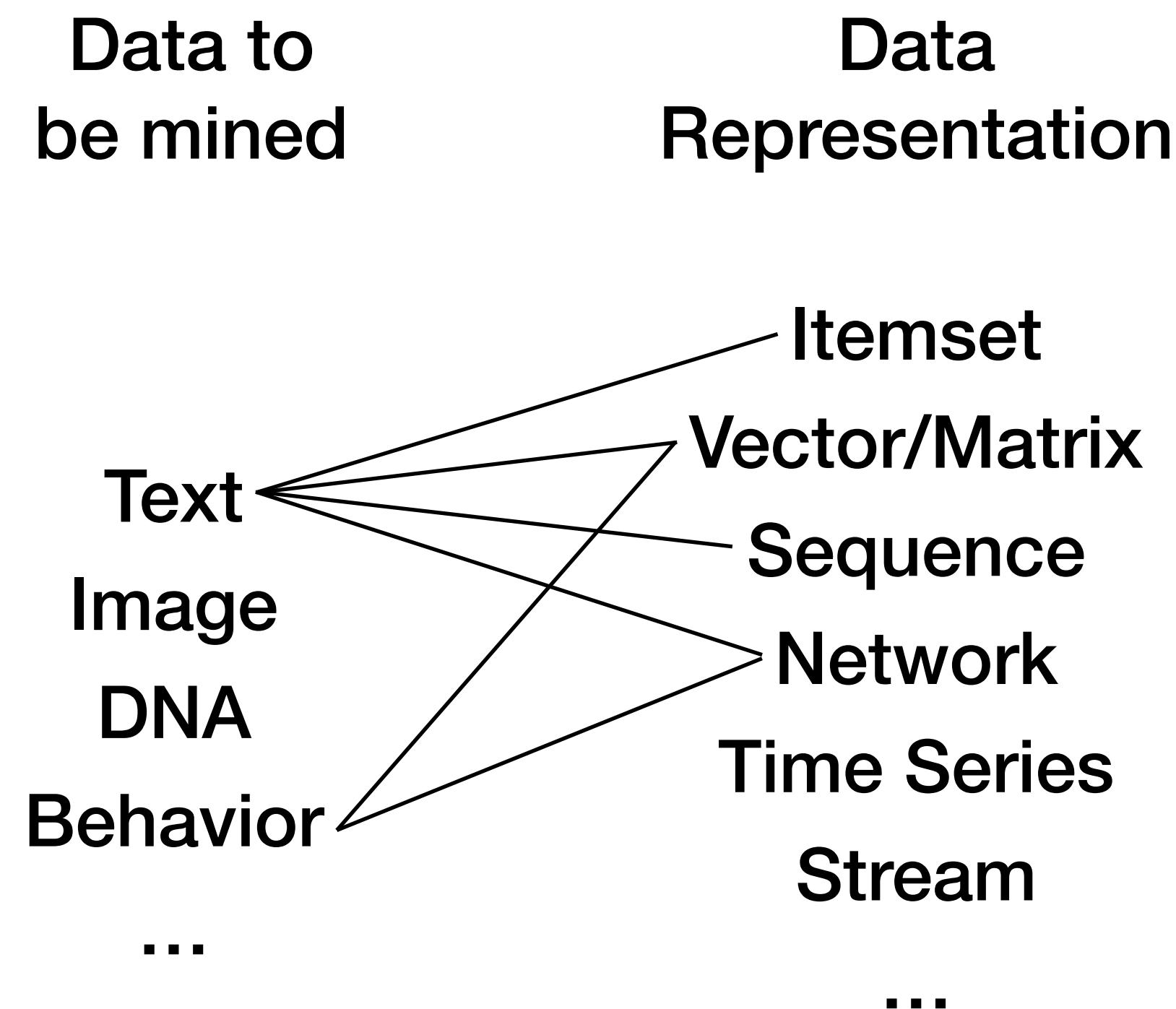
Data Source and Manipulation



- Item Set
- Vector / Matrix
- Sequence
- Time Series
- Spatial
- Spatiotemporal
- Graph / Network
- Stream

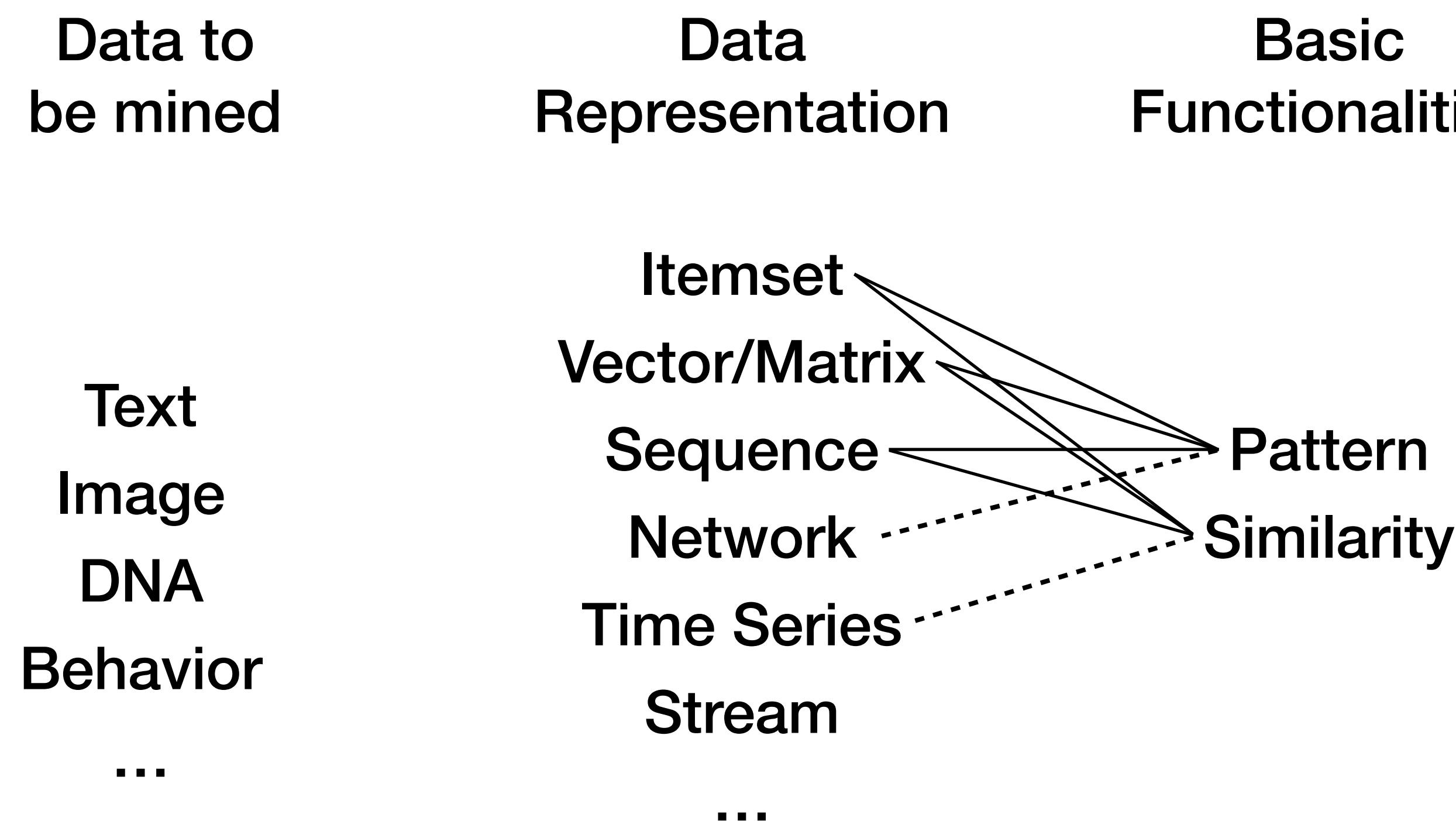
Filtering
Transforming
Pivot
Merging
Group
Aggregation
...

What's next?



- Text -> Itemset:
 - {"to", "be", "or", "not", ...}
- Text -> Vector:
 - $<1, 0, 0, 1>$
- Behavior -> Vector:
 - User-Product Rating Matrix
- Behavior -> Network:
 - Twitter “Following” network

Basic Data Mining Functionalities: Pattern & Similarity



- Itemset + Patterns:
 - Frequent Pattern Mining
- Itemset + Similarity:
 - Jaccard Similarity
- Vector + Similarity:
 - Dot Product
 - Manhattan/Euclidean distance
 - Cosine Similarity
- Sequence + Similarity:
 - Edit distance
 - Shingling
- Network + Pattern:
 - ...

Data Mining Functionalities

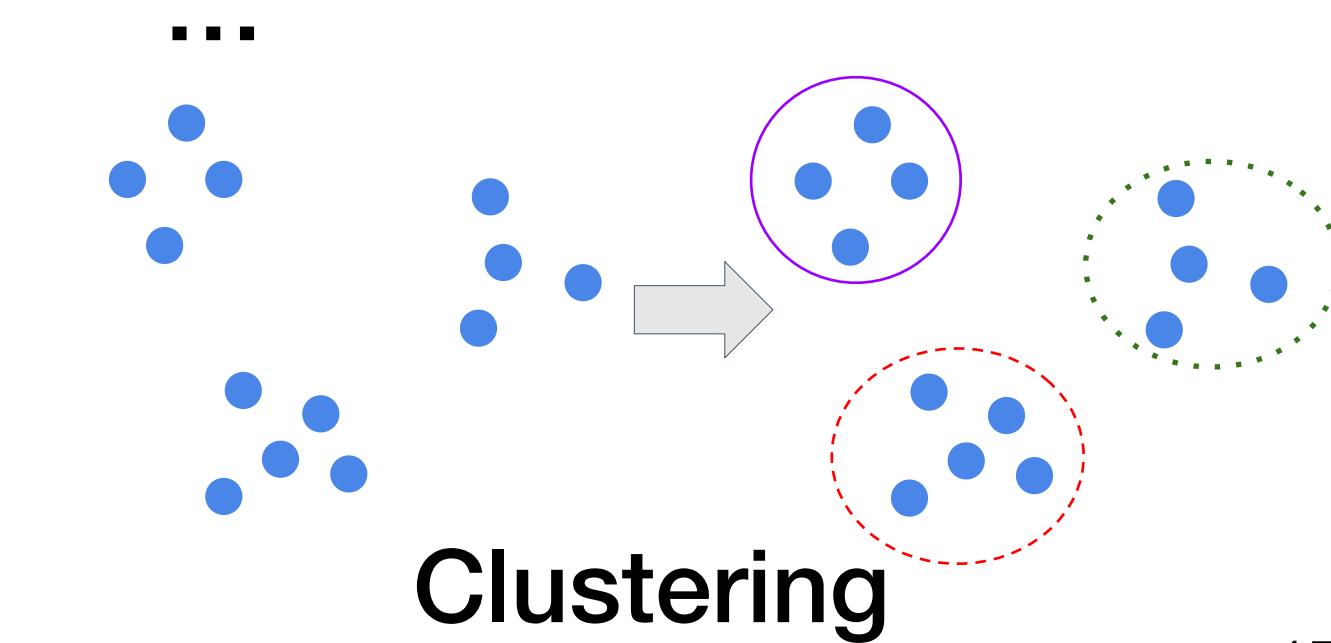
Data to be mined

Text
Image
DNA
Behavior

...

Data Representation

Itemset
Vector/Matrix
Sequence
Network
Time Series
Stream



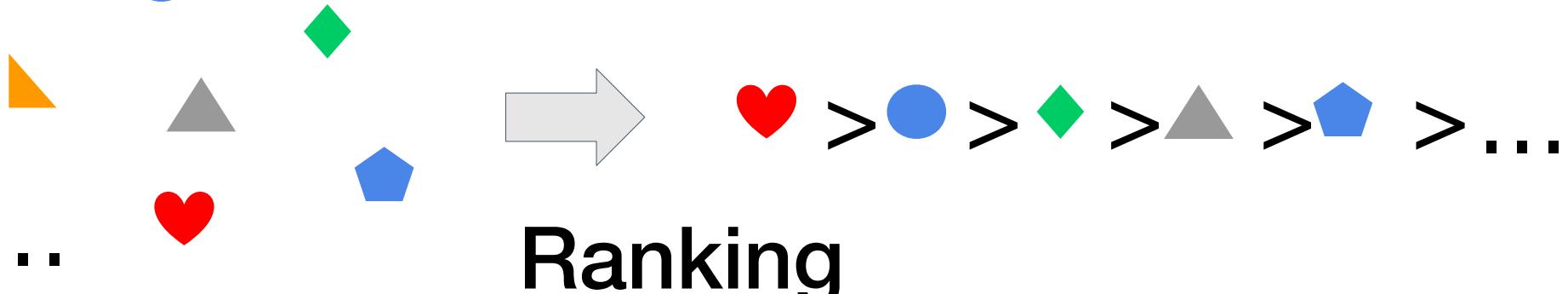
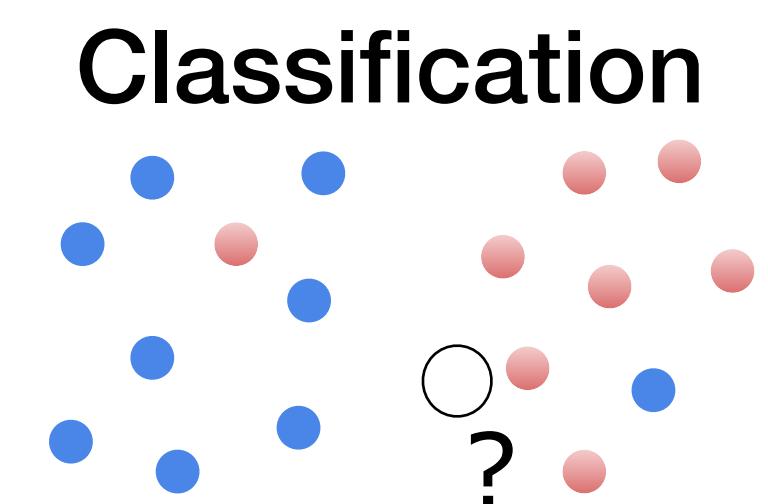
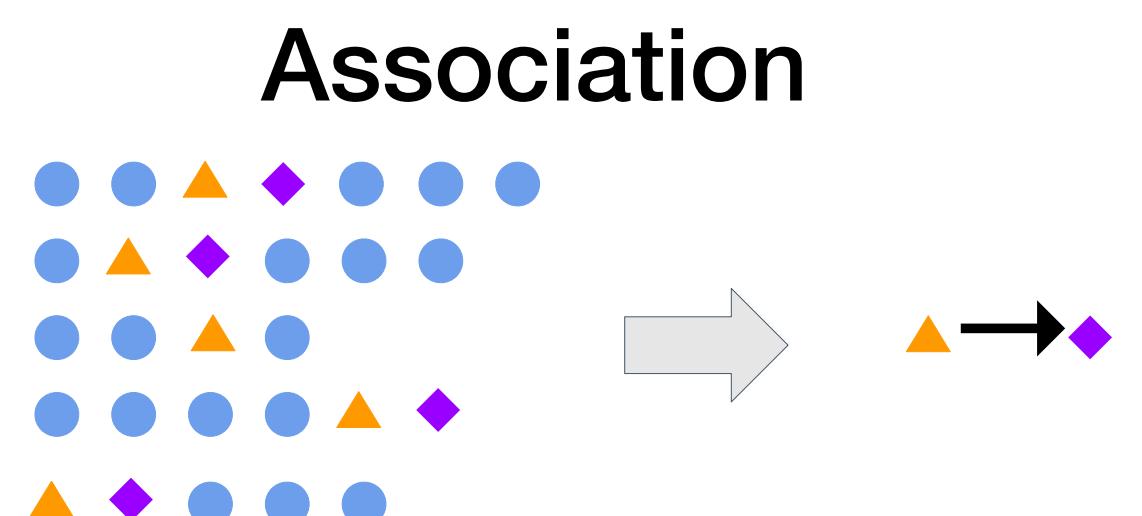
Basic Functionalities

Pattern Similarity

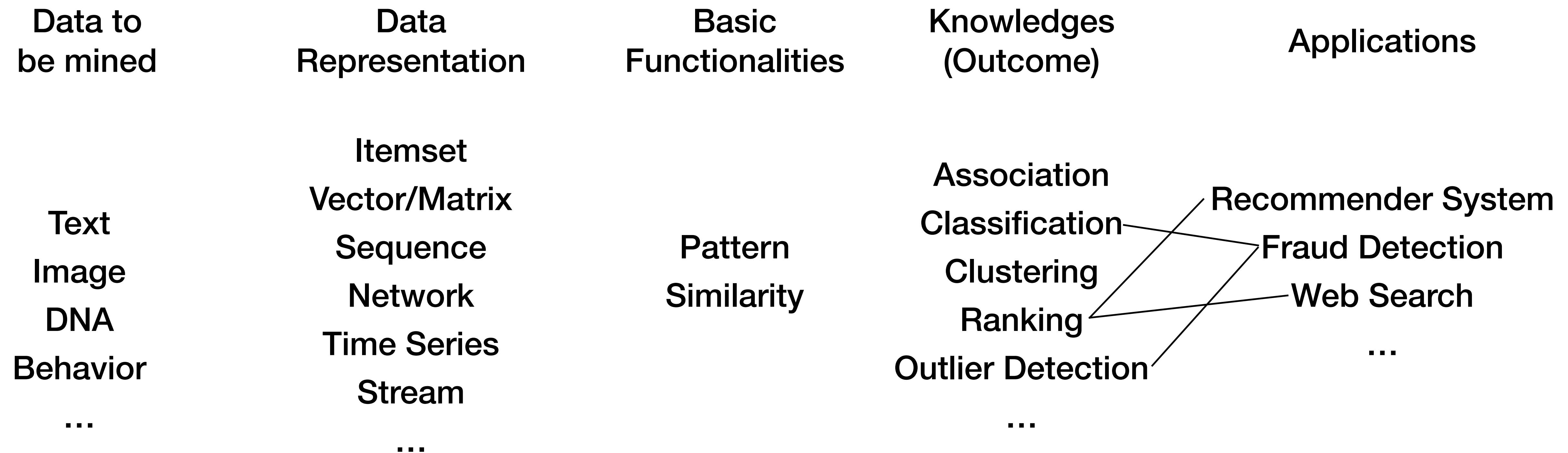
Knowledges (Outcome)

Association
Classification
Clustering
Ranking
Outlier Detection

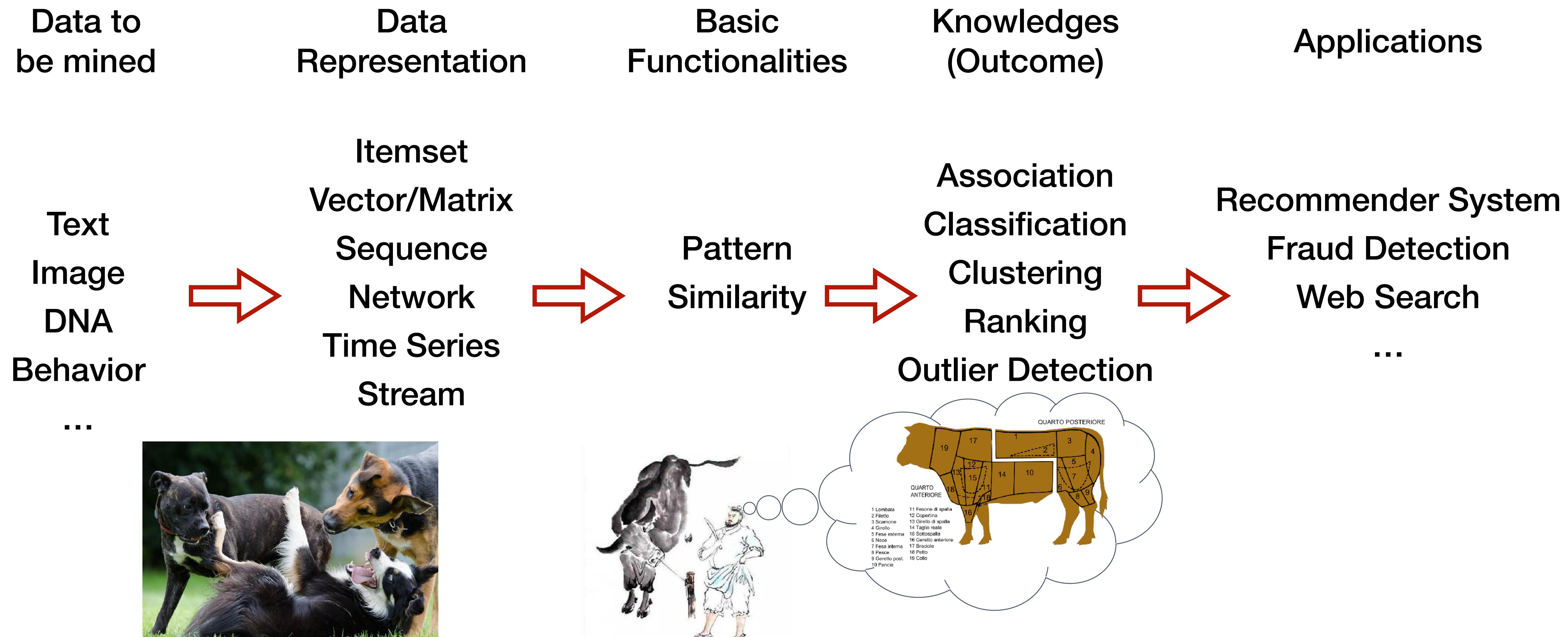
...



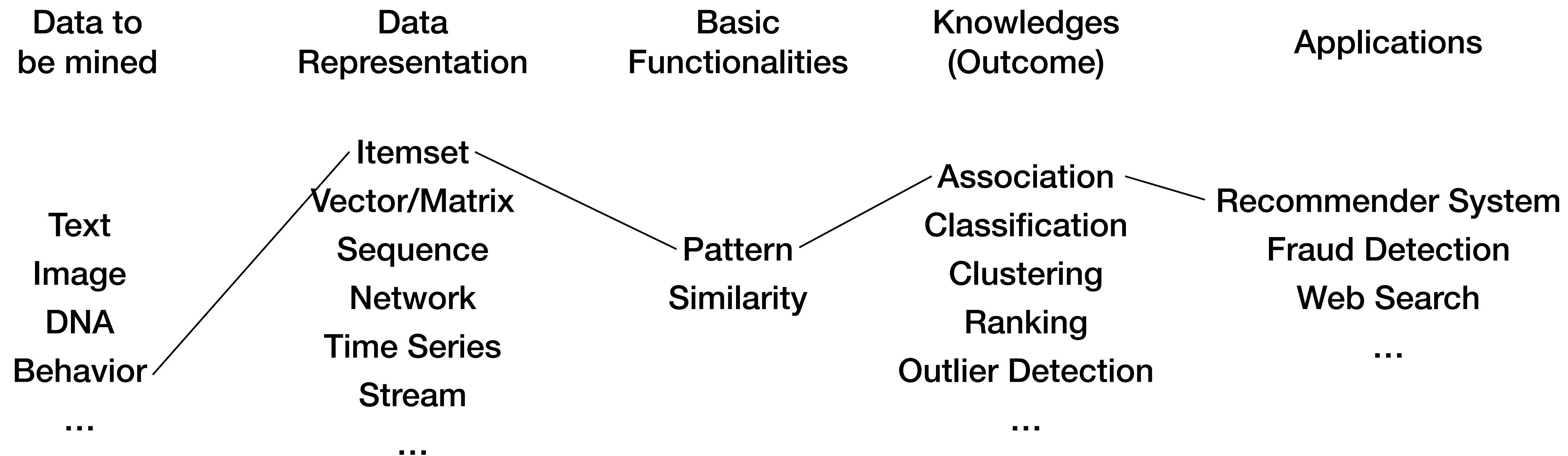
Functionalities -> Applications



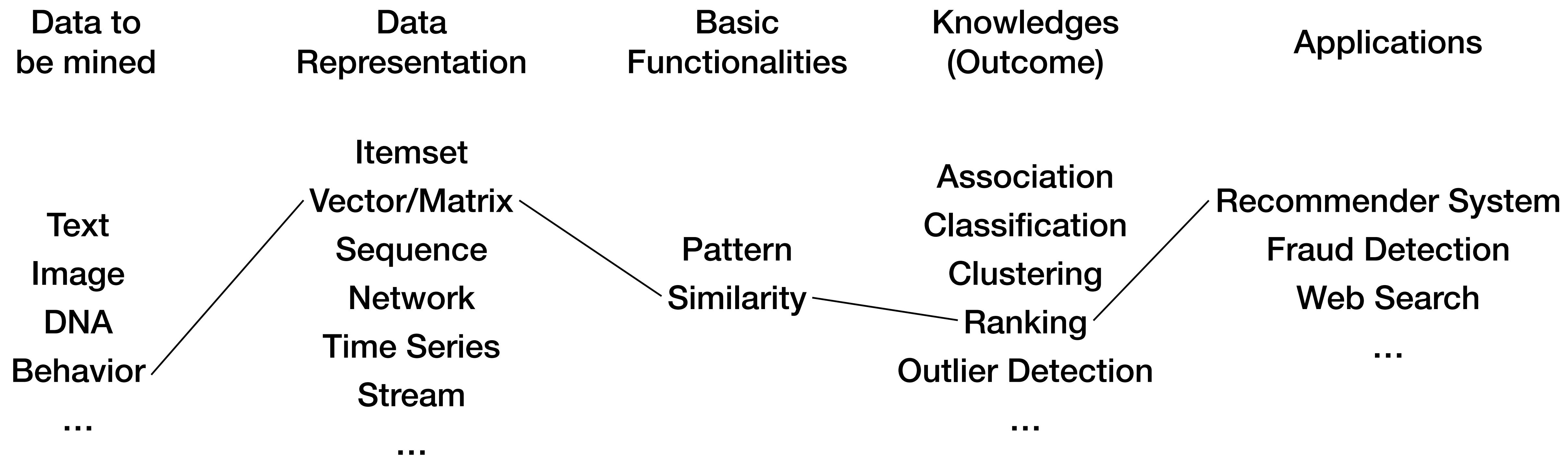
Data Mining: Knowledge Discovery from Data



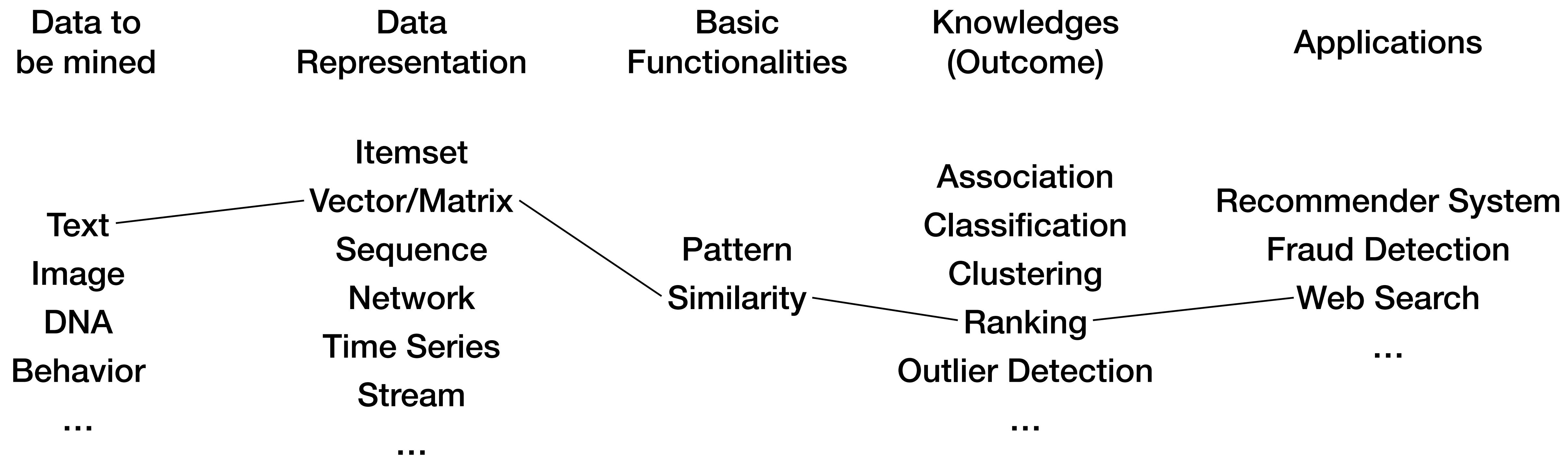
Example: Amazon “Frequently Bought Together”



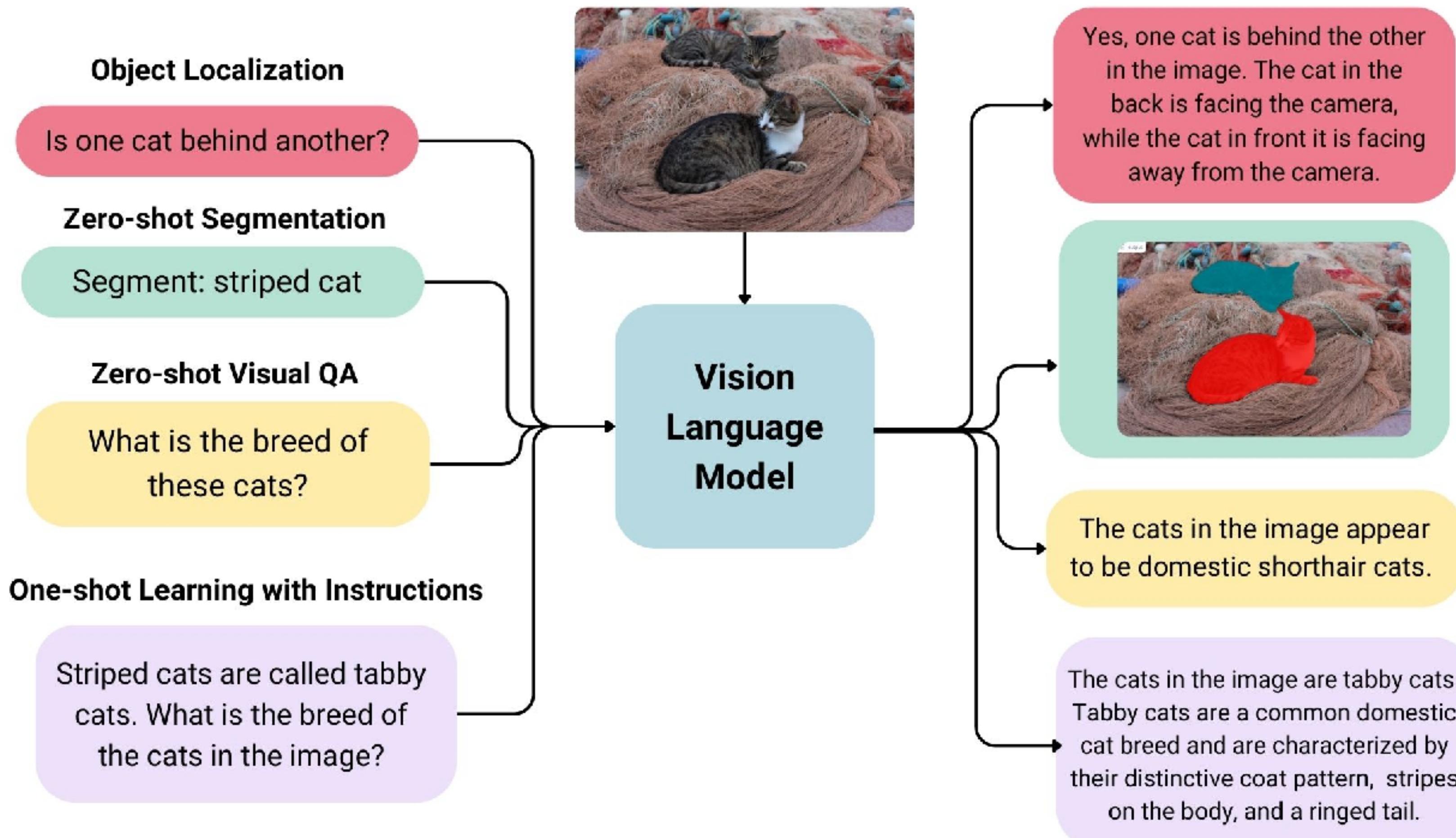
Example: Netflix Movie Recommender System



Example: Text Retrieval



How does LLM change the game?



Audio-Language Model (Qwen-Audio)

(a) Basic Interaction

User: I'm trying to understand how this model handles different audio inputs.

Model: OK. There are two main ways: **recognize** and **discuss**.

User: Can you show me an example of each?

Model: Sure! (a) **recognize**: "OK. T 有什么 discuss build" (b) **discuss**: "OK. T 和 f Engli find"

(b) Multiple Inputs

User: I have multiple audio files. How can I handle them?

Model: You can use the **discuss** function to handle multiple inputs simultaneously. For example, you can upload multiple audio files and ask for a summary or analysis.

(c) Image Recognition

User: I uploaded an image of a bird. What's the model's response?

Model: The tone of the audio is cheerful, possibly something pleasant. It might be a woman in audio, possibly discussing something.

(d) Audio-Motivated Creative Writing

User: I want to write a poem about the audio I just heard.

Model: Write a poem about the above audio

Birds are singing in the trees,
A gentle breeze is blowing free,
The rustling leaves are music,
As nature's symphony goes on.
The sun is shining bright and warm,
A peaceful atmosphere is sown,
In the midst of nature's beauty,
We can find serenity and duty.
The birds take flight with ease,
Their chirping voices fill the trees,
A symphony of nature's grace,
A moment to pause and embrace.
The wind whispers through the branches,
As the leaves dance in the branches,
A peaceful moment in time,
A chance to let go and unwind.
So let us take a moment here,
To appreciate the beauty so dear,
In the midst of nature's wonder,
We can find a sense of wonder.

(e) Music Appreciation

User: I uploaded a piano piece. What's the model's response?

Model: what is it

A solo piano piece.

Describe this music in detail

This is a solo piano piece. It is fast tempo with a playful and humorous mood. The piece is in the key of C major and is in 4/4 time signature. The melody is played on the right hand with the left hand playing chords. The piece is a classic ragtime style.

Give me some recommendations for similar music.

some similar music to "Solo Piano Ragtime" could be:

1. "The Entertainer" by Scott Joplin
2. "Maple Leaf Rag" by Scott Joplin
3. "Ave Maria" by Franz Schubert
4. "Für Elise" by Ludwig van Beethoven
5. "The Blue Danube" by Johann Strauss II

How Recent ML development is Changing the game?

- Vector representation becomes prevalent
 - Word2Vec, Doc2Vec, Wav2Vec,
 - Everything can be “embedded” as a vector
- With LLM, a new paradigm is emerging:
 - Text data becomes more prevalent intermediate data representation,
 - Stronger ability to extract information from multimodal data source.
 - But “newer” development are contending the argument (e.g. DeepSeek OCR)
- “Agentic” AI is blurring the boundary.

Recap

- Data formulation is the first task of data mining
- Different representations of data may be applied to the same data.
- How to represent data as itemsets, matrix, time series, sequences, networks, and streams
- Recent Advancement in GenAI bring stronger capacity in extracting information from wild data.