School of Computing
UNIVERSITY OF GEORGIA

CSCI 4380/6380 DATA MINING

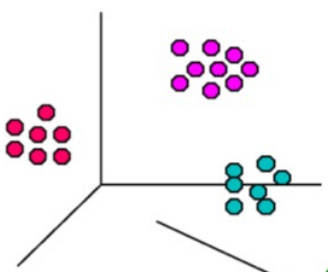
Fei Dou

Assistant Professor
School of Computing
University of Georgia

August 22, 2023

Data Representation

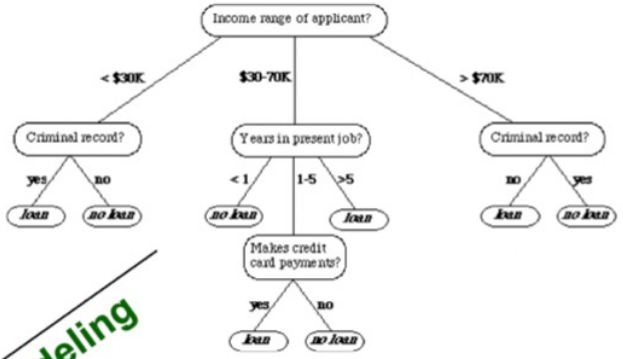
Data Mining



Clustering

Data

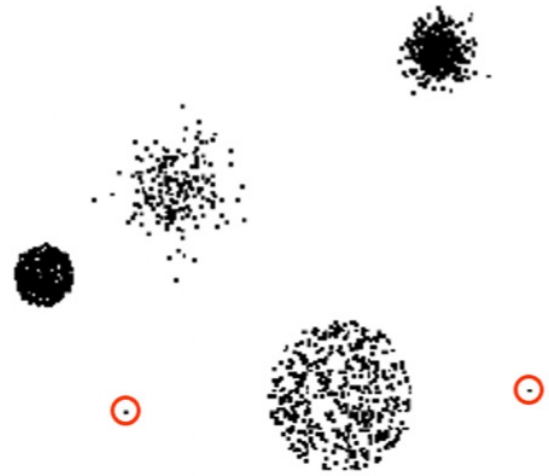
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes



Predictive Modeling

Association Rules

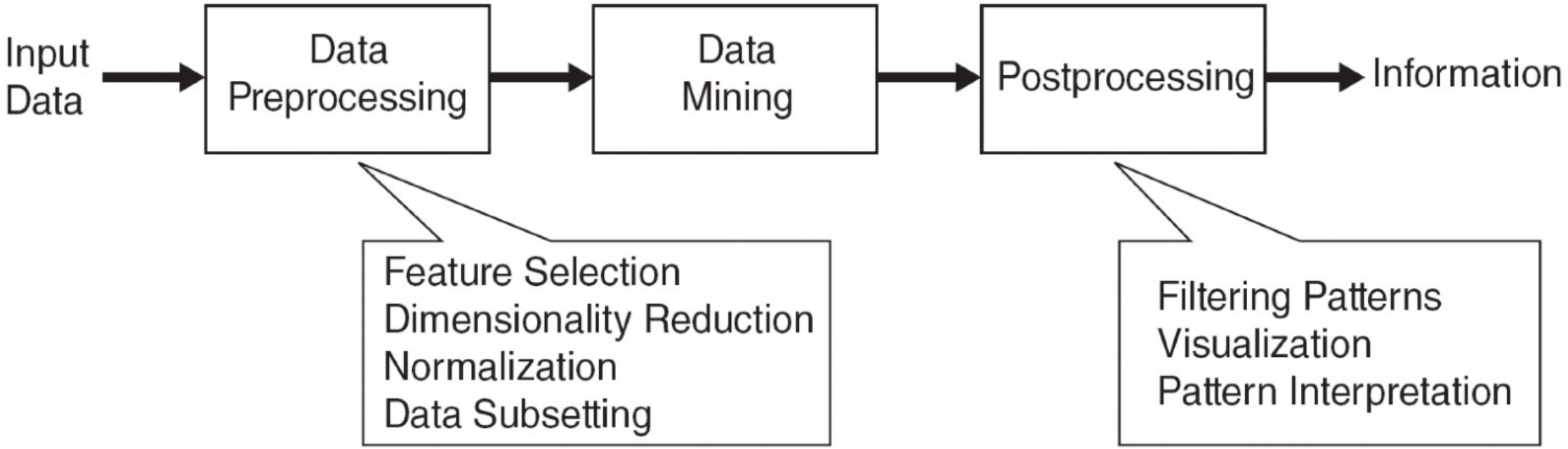
Anomaly Detection



Motivating Challenges

- Scalability
- High Dimensionality
- Heterogeneous and Complex Data
- Data Ownership and Distribution
- Non-traditional Analysis
  - hypothesis generation and evaluation

Data Mining Process



What is Data?

- Collection of ***data objects*** and their ***attributes***
- An ***attribute*** is a property or characteristic of an object
  - Examples: eye color of a person, etc.
  - Attribute is also known as variable, field, characteristic, dimension, or feature
- A collection of attributes describe an ***object***
  - Object is also known as record, point, case, sample, entity, or instance

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Attribute Values

- ***Attribute values*** are numbers or symbols assigned to an attribute for a particular object

- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
  - But properties of attribute can be different than the properties of the values used to represent the attribute

Types of Attributes

- There are different types of attributes
  - Nominal
    - Examples: ID numbers, eye color, zip codes
  - Ordinal
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height {tall, medium, short}
  - Interval
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - Ratio
    - Examples: temperature in Kelvin, length, counts, elapsed time (e.g., time to run a race)

Properties of Attribute Values

- The type of an attribute depends on which of the following properties/operations it possesses:
  - Distinctness: =, ≠
  - Order: <, >
  - Differences are meaningful : +, -
  - Ratios are meaningful: *, /

  - Nominal attribute: distinctness
  - Ordinal attribute: distinctness & order
  - Interval attribute: distinctness, order & meaningful differences
  - Ratio attribute: all 4 properties/operations

Discrete and Continuous Attributes

- Discrete Attribute, $x \in N$
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: binary attributes are a special case of discrete attributes

- Continuous Attribute, $x \in R$
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight.
  - Practically, real values can only be measured and represented using a finite number of digits.
  - Continuous attributes are typically represented as floating-point variables.

Important Characteristics of Data

- Dimensionality (number of attributes)
  - High dimensional data brings a number of challenges
- Sparsity
  - Only presence counts
- Resolution
  - Patterns depend on the scale
- Size
  - Type of analysis may depend on size of data

What Kind of Data Can be Mined?

- Database-oriented data sets and applications
  - Relational database, data warehouse, transactional database
- Advanced data sets and advanced applications
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks
  - Object-relational databases
  - Heterogeneous databases and legacy databases
  - Spatial data and spatiotemporal data
  - Text databases
  - The World-Wide Web

What Kind of Data Can be Mined?

- **Vector/Tabular Data**

	Sex	Race	Height	Income	Marital Status	Years of Educ.	Liberal-ness
R1001	M	1	70	50	1	12	1.73
R1002	M	2	72	100	2	20	4.53
R1003	F	1	55	250	1	16	2.99
R1004	M	2	65	20	2	16	1.13
R1005	F	1	60	10	3	12	3.81
R1006	M	1	68	30	1	9	4.76
R1007	F	5	66	25	2	21	2.01
R1008	F	4	61	43	1	18	1.27
R1009	M	1	69	67	1	12	3.25

What Kind of Data Can be Mined?

- **Vector/Tabular Data Data Matrix**

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute
- Such a data set can be represented by an m by n matrix

House #	Square Footage	Bedrooms	Price
1	1500	3	250000
2	1800	4	320000
3	1200	2	180000
...	...	...	...

What Kind of Data Can be Mined?

- **Set Data**
  - Each transaction involves a set of items.
  - Can represent transaction data as vector data

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

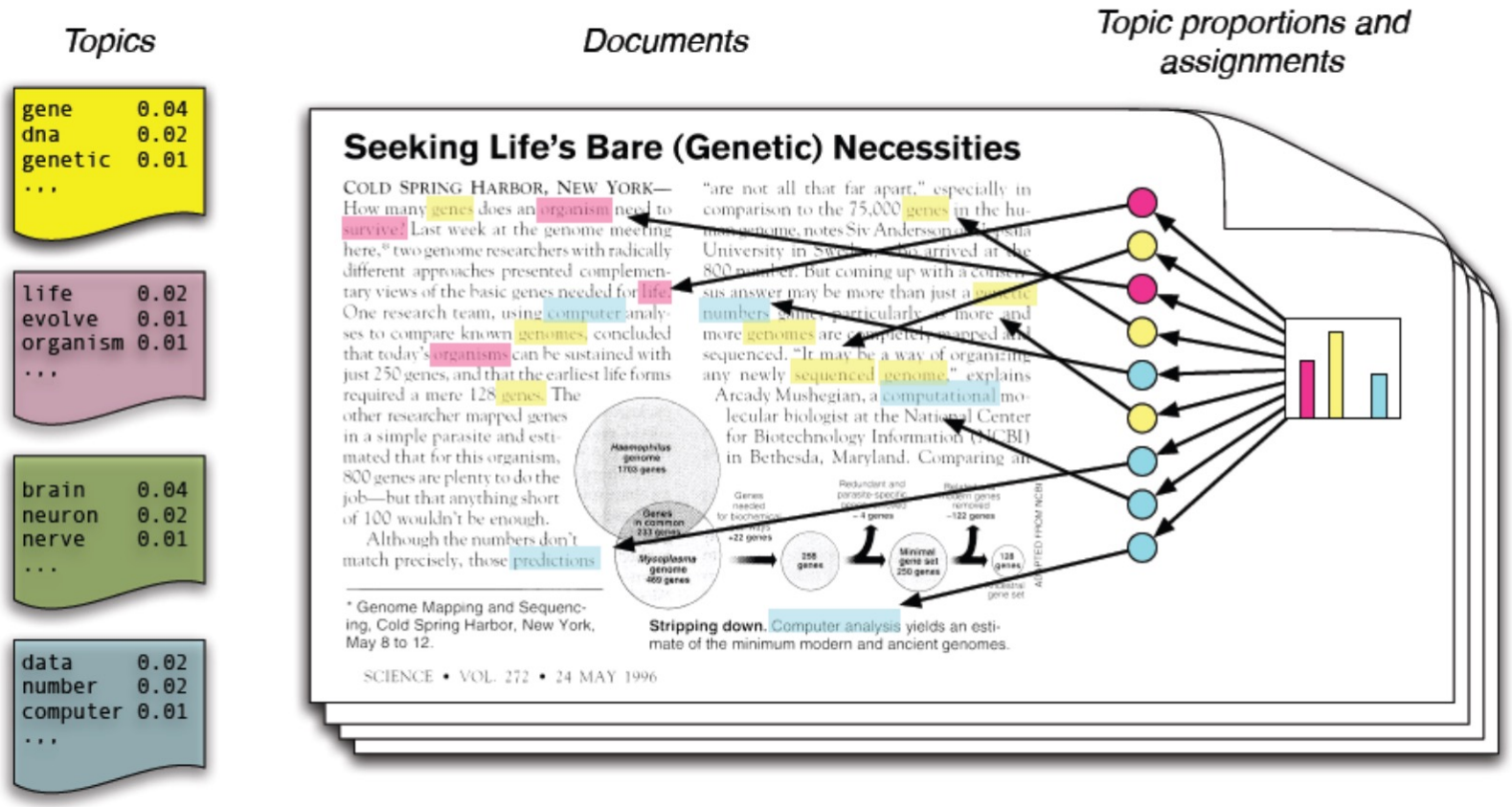
What Kind of Data Can be Mined?

- **Text Data**

"Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (i.e., learning relations between named entities)." –from wiki

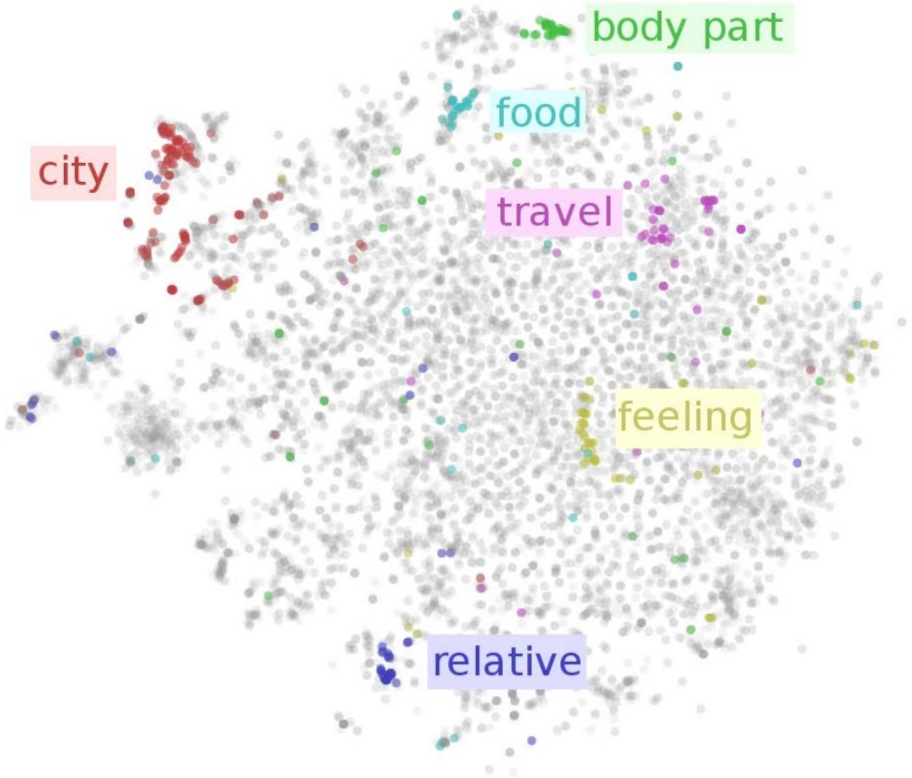
What Kind of Data Can be Mined?

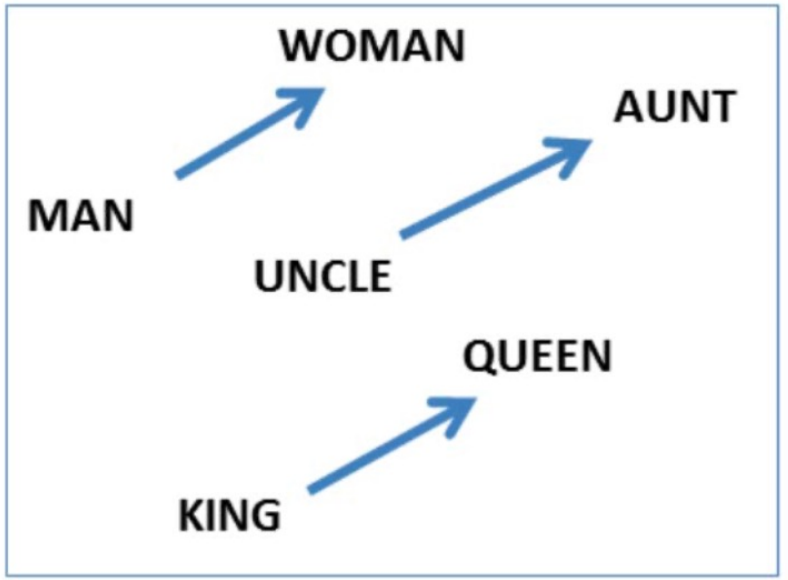
- **Text Data – Topic Modeling**



What Kind of Data Can be Mined?

- **Text Data – Word Embedding**

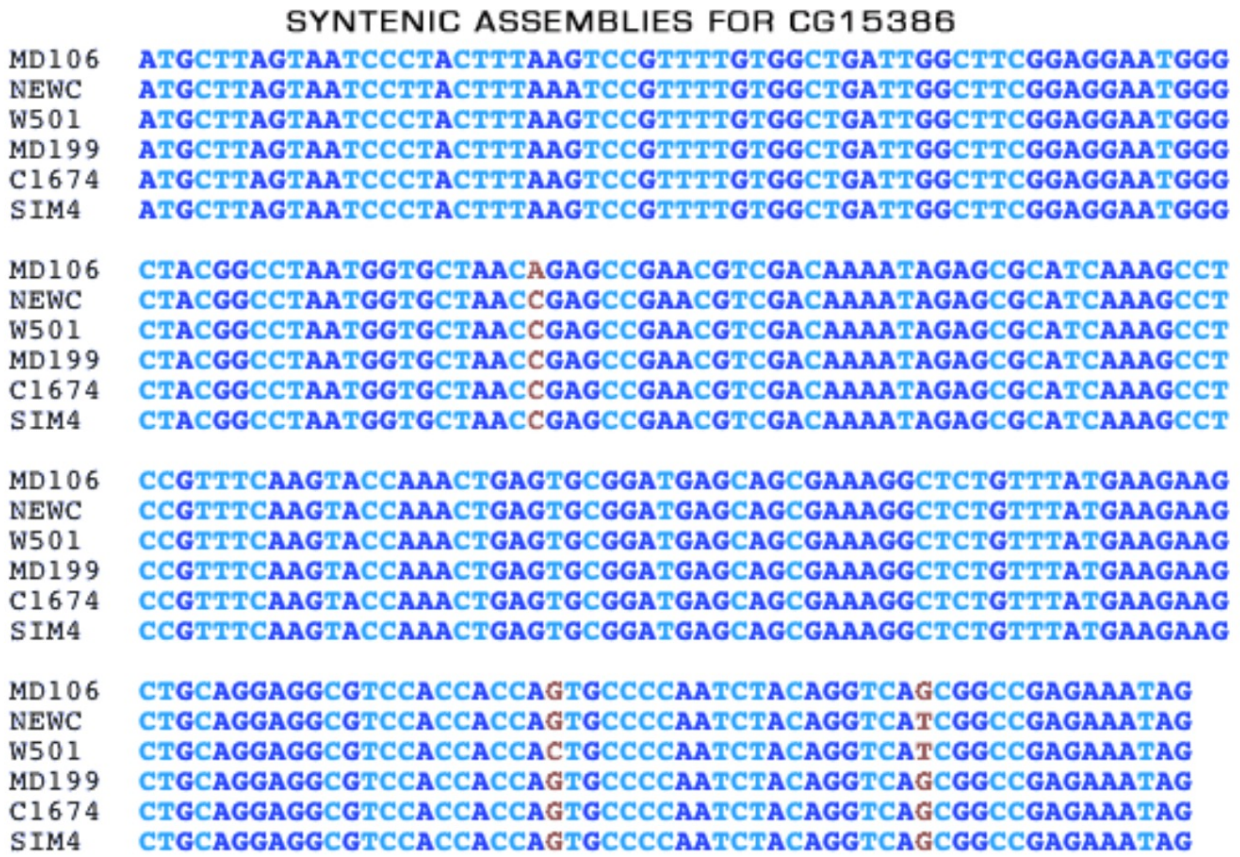




king - man + woman = queen

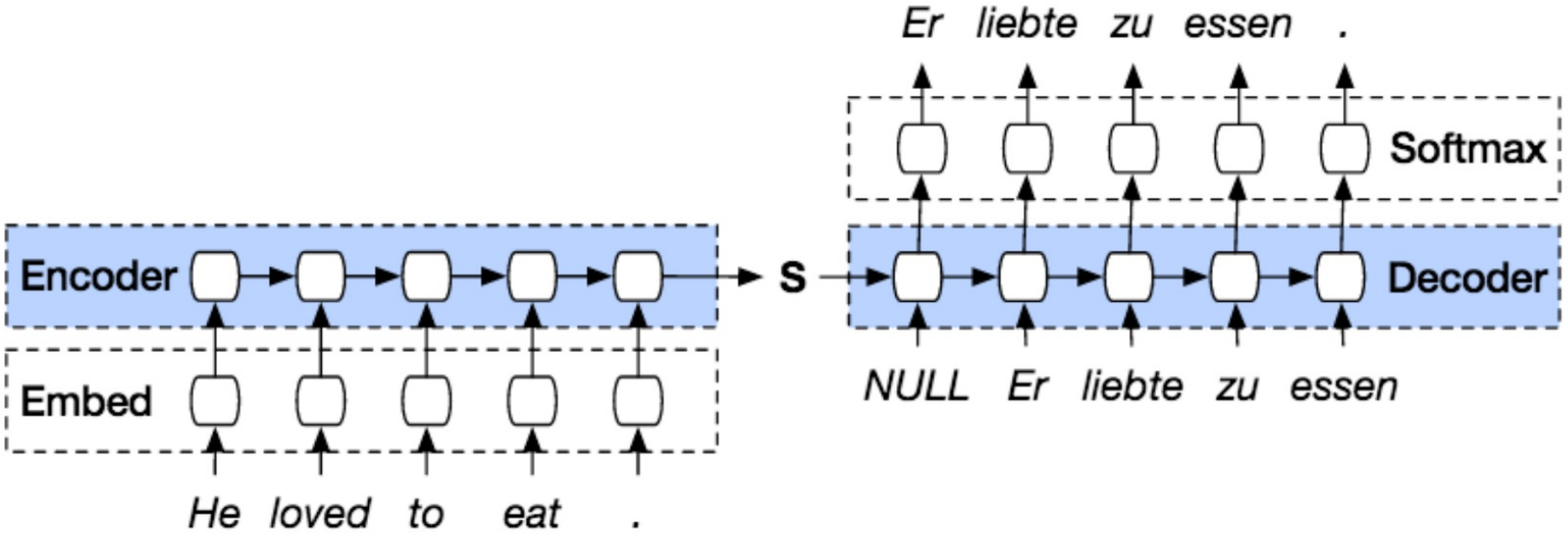
What Kind of Data Can be Mined?

- **Sequence Data**



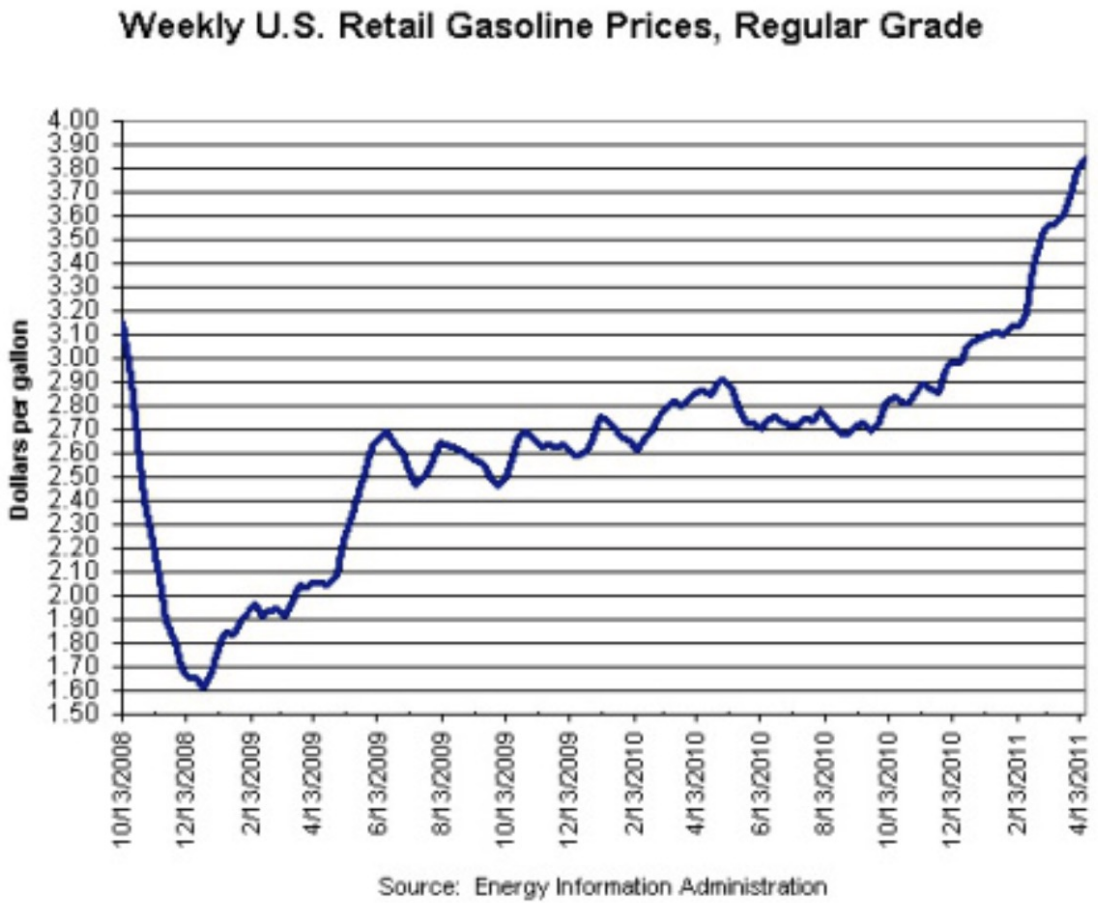
What Kind of Data Can be Mined?

- **Sequence Data – Seq2Seq**



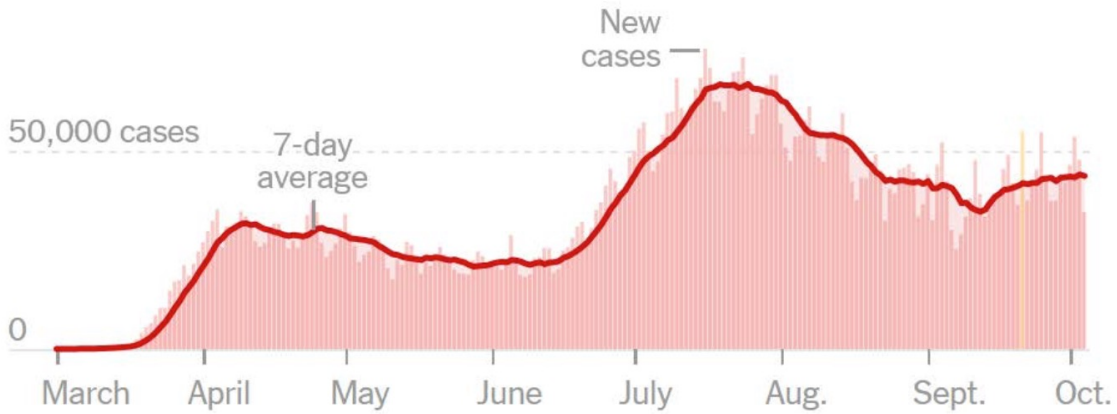
What Kind of Data Can be Mined?

- **Time Series**



What Kind of Data Can be Mined?

- **Time Series**



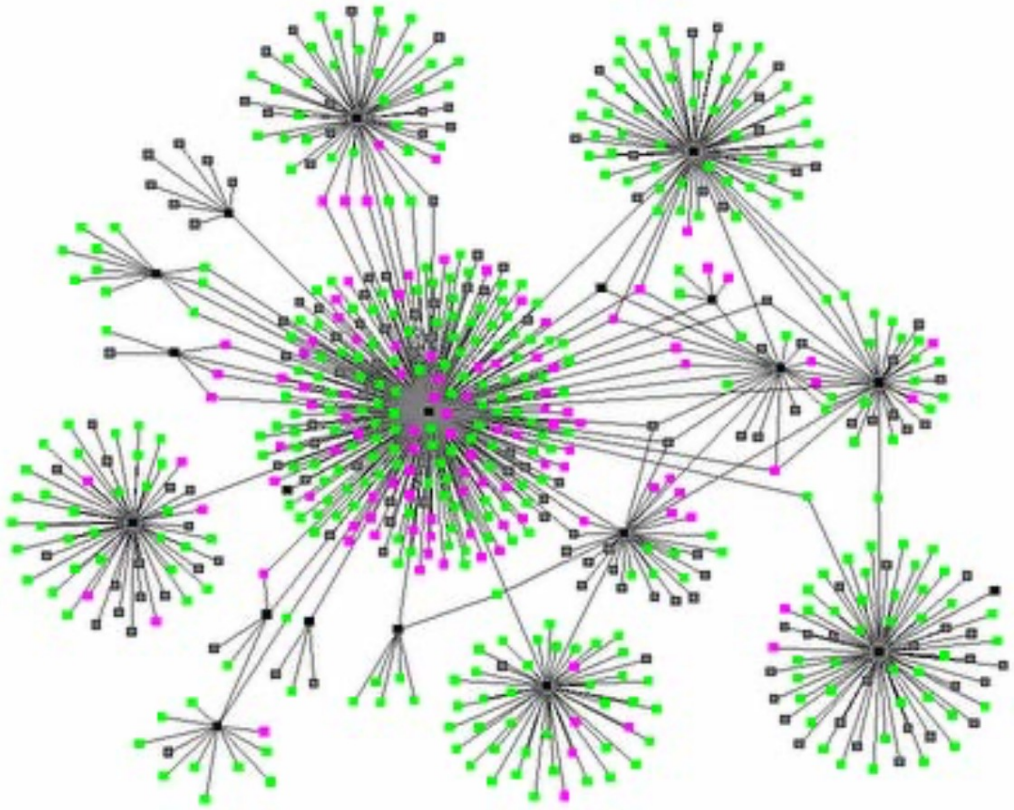
	TOTAL REPORTED	ON OCT. 4	14-DAY CHANGE
Cases	7.4 million+	34,491	+6% →
Deaths	209,603	332	−8% →

Day with data reporting anomaly.

Includes confirmed and probable cases where available. 14-day change trends use 7-day averages.

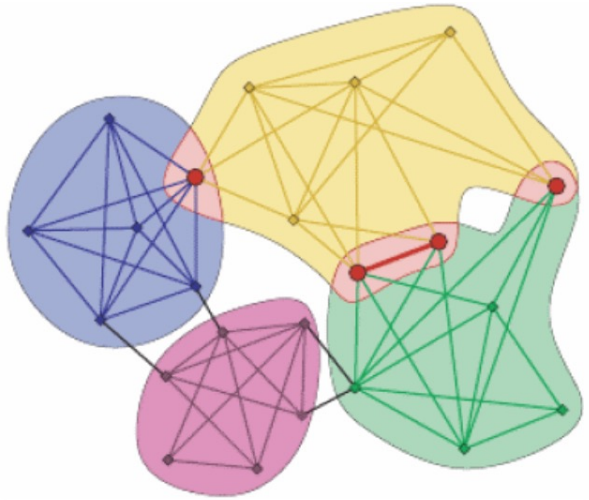
What Kind of Data Can be Mined?

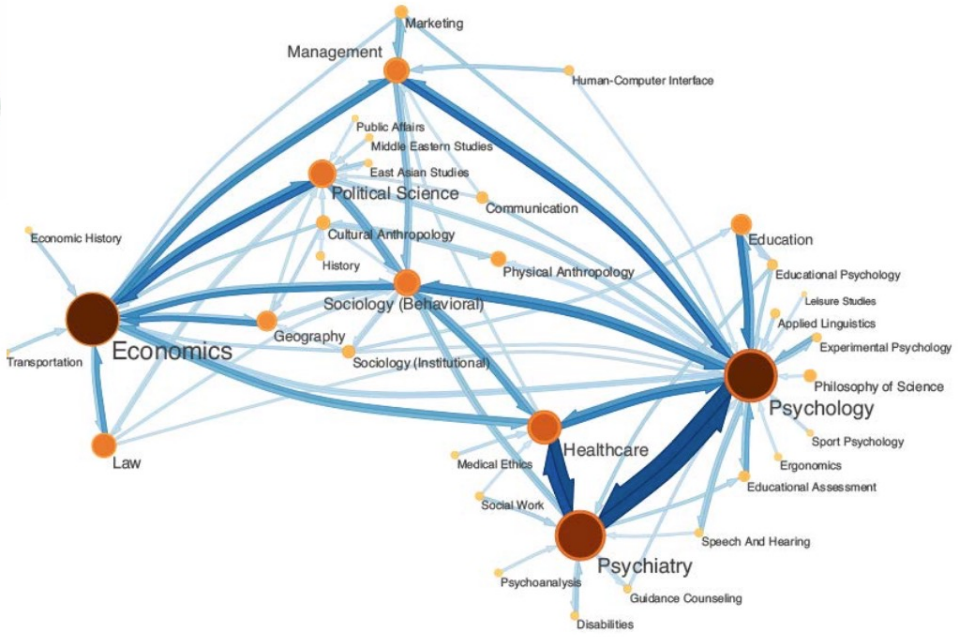
- **Graph / Network**



What Kind of Data Can be Mined?

- **Graph / Network – Community Detection**





What Kind of Data Can be Mined?

- **Image Data**

What Kind of Data Can be Mined?

- **Image Data – Neural Style Transfer**

What Kind of Data Can be Mined?

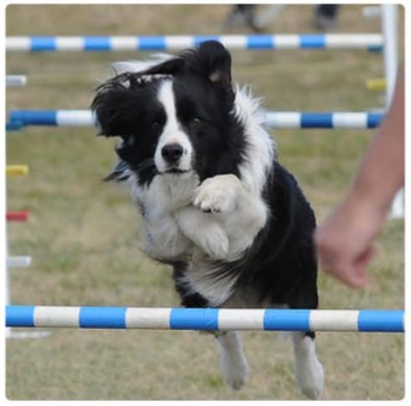
- **Image Data – Image Captioning**

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

"girl in pink dress is jumping in air."



"black and white dog jumps over bar."

"young girl in pink shirt is swinging on swing."