SI 671/721:
Introduction to Data Mining (I)

Lecture 1
Fall 2021

Instructor: Prof. Paramveer Dhillon
dhillonp@umich.edu
University of Michigan

UMSI

What is Data Mining?

- With the rapid growth of data over the last couple of decades various terms have gained popularity—**Data Mining** being one of them.

- What really is data mining? Is it about techniques? Is it about analyzing large scale data? What do you think?
- Is it the same as machine learning, data science, and big data analytics? If not, how is it different?

Alternative Names of Data Mining

- Knowledge Discovery in Databases
- Knowledge Extraction
- Data/Pattern Analysis
- Data Archeology
- Data Dredging
- Information Harvesting

Explosive Growth of data due to the advent of Internet

By 2025 ~100 zettabytes of data will be generated worldwide.

- 1 zettabyte = 1000 exabytes
- 1 exabyte = 1000 petabyte
- 1 petabyte = 1000 terabytes
- 1 zettabyte = 1 trillion terabytes

E.g., Tweets: ~6,000 generated per sec

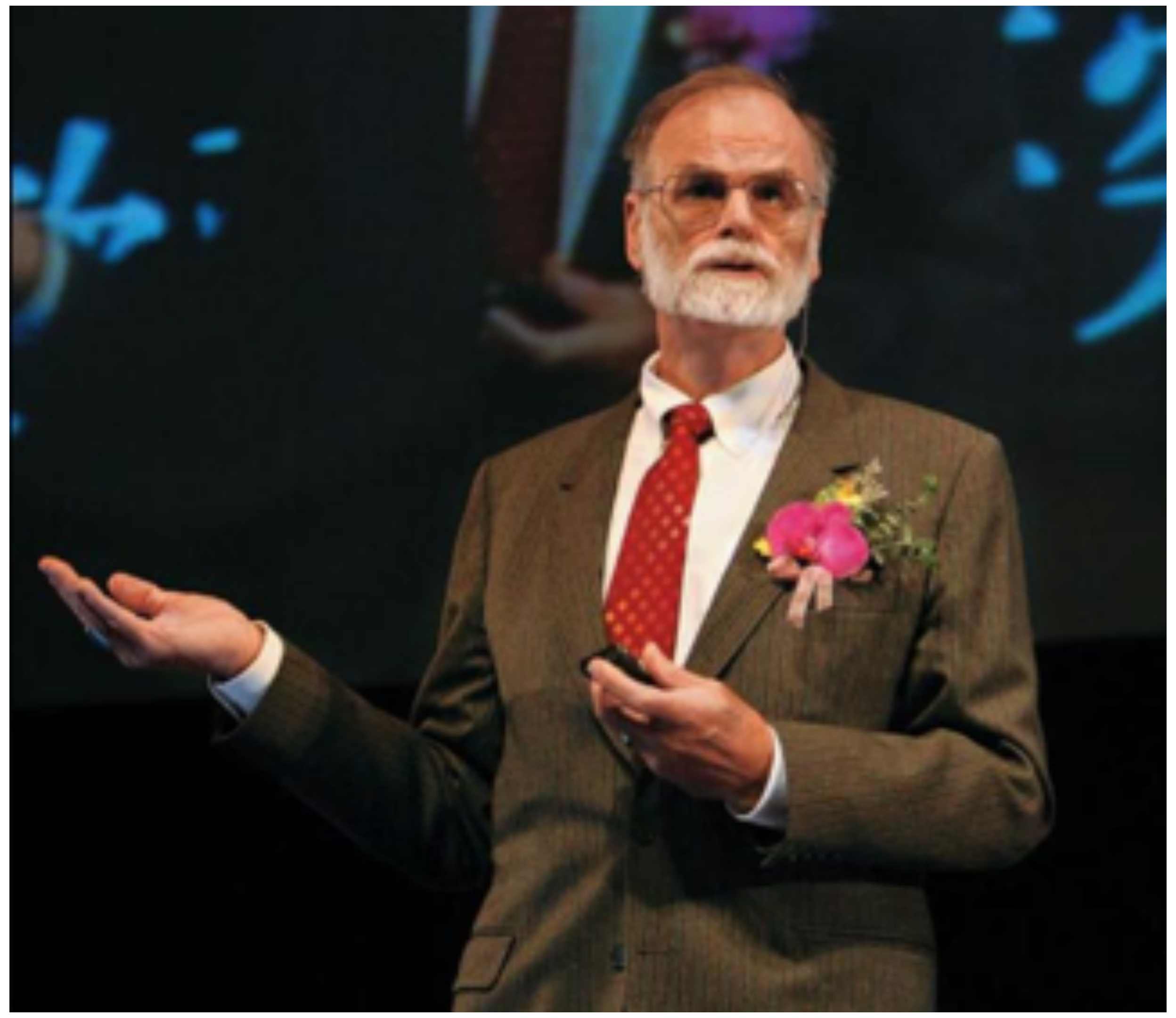
- 200 billion Tweets generated per year!

Explosive Growth of data

"We are drowning in data, but starving for knowledge."

*Inspired from quote by John Naisbitt in 1982: "We are drowning in information but starved for knowledge."

The Fourth Paradigm of Science



Jim Gray (1944 - 2007)
Computer Scientist
Turing Award Winner (1998)

First Paradigm: Empirical/Experimental Science (~1600)
Second Paradigm: Theoretical Science (1600~1950s)
Third Paradigm: Computational Science (1950s-1990s)

Fourth Paradigm: Data-intensive Science called "eScience" (2000s ~)

- Use of data-driven discovery
- Closely related to "data science"
- *Data mining is the major challenge*

What is Data Mining?

"Knowledge Discovery from Data"

"Knowledge Discovery from Data": What do the experts say?

<u>Jiawei Han:</u>

Extraction of ***interesting*** (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from ***huge amounts of data***.

"Knowledge Discovery from Data": What do the experts say?

Sunita Sarawagi:

Process of semi-automatically analyzing ***large databases*** to find ***patterns*** that are:

- valid: hold on new data with some certainty
- novel: non-obvious to the system
- useful: should be possible to act on the item
- understandable: humans should be able to interpret the pattern

“Knowledge Discovery from Data”: What do the experts say?

Vipin Kumar:

Exploration and analysis, by automatic or semi-automatic means, of ***large quantities of data*** in order to discover ***meaningful*** patterns.

Not Everything is Data Mining!

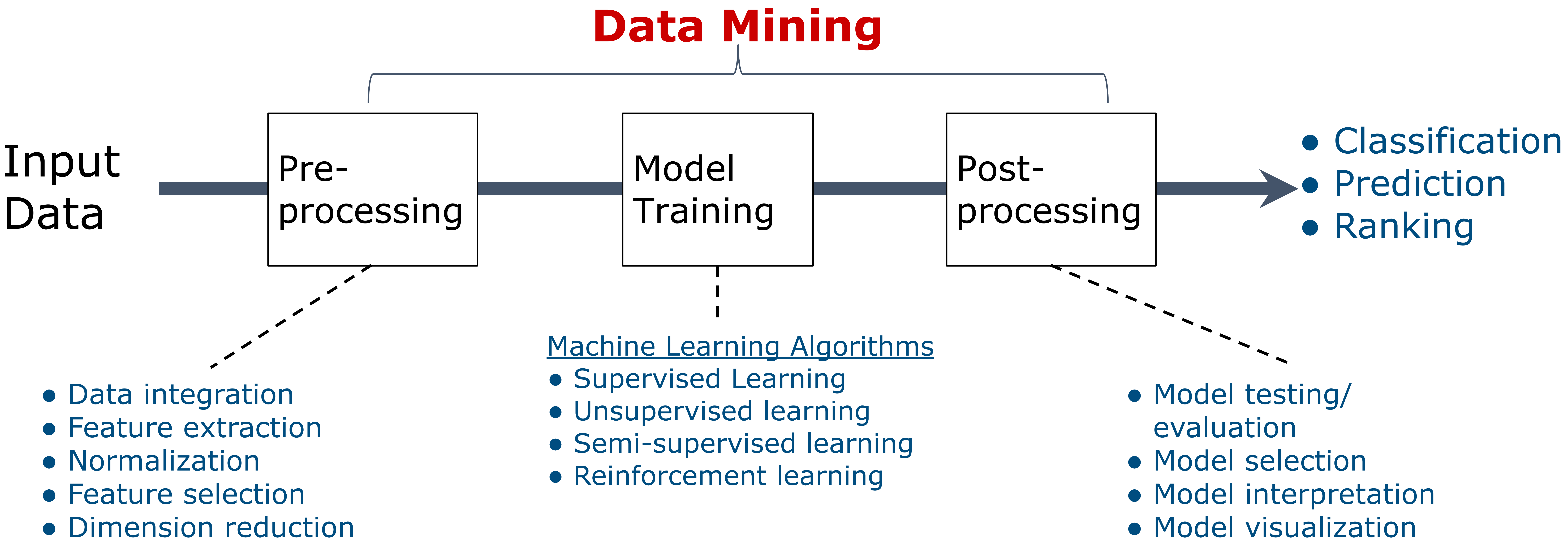
- Looking up a phone number in phone directory. Data mining?
- Query a search engine for pages that contain "Amazon." Data mining?
- Collecting and storing data in a database. Data Mining?

Concepts Related to Data Mining

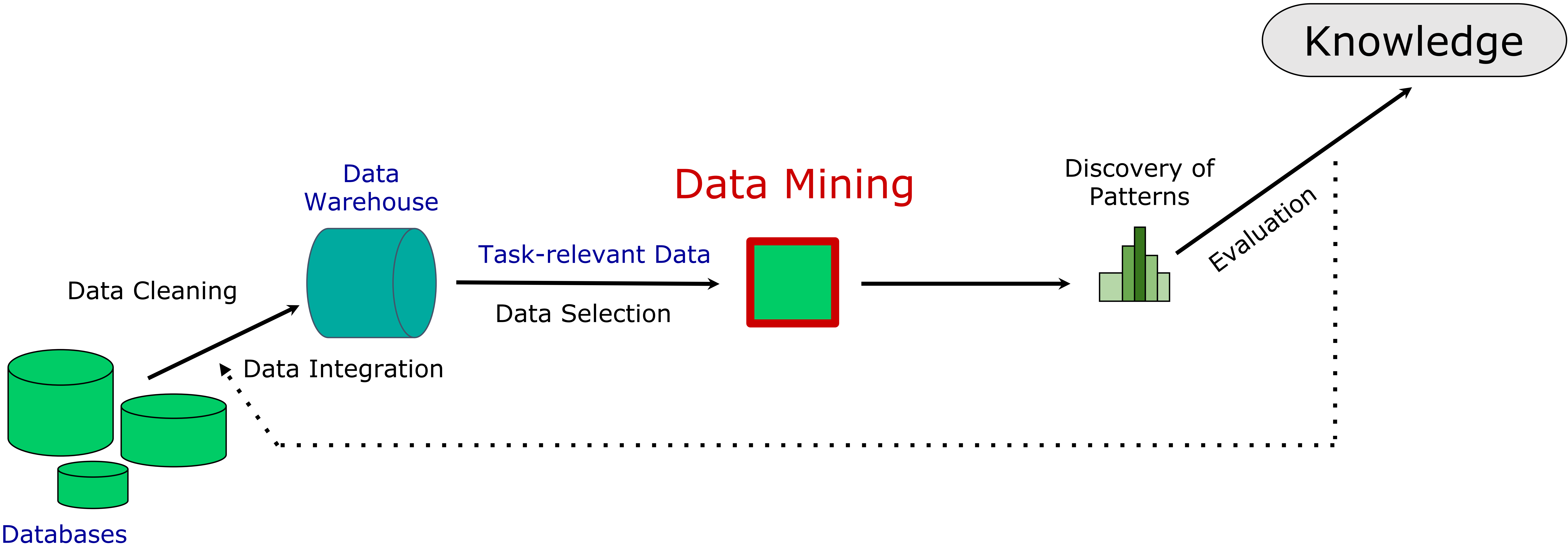
• Machine Learning • Pattern Recognition	Techniques utilized in data mining process.
• Database Management Systems • Data Warehouses	The systems that support data mining.
• Big Data Analytics • Data Science	Data Mining is a key component to these broad fields.
• Business Intelligence	A particular application of data mining.

Different Views of Data Mining

A Machine Learning View

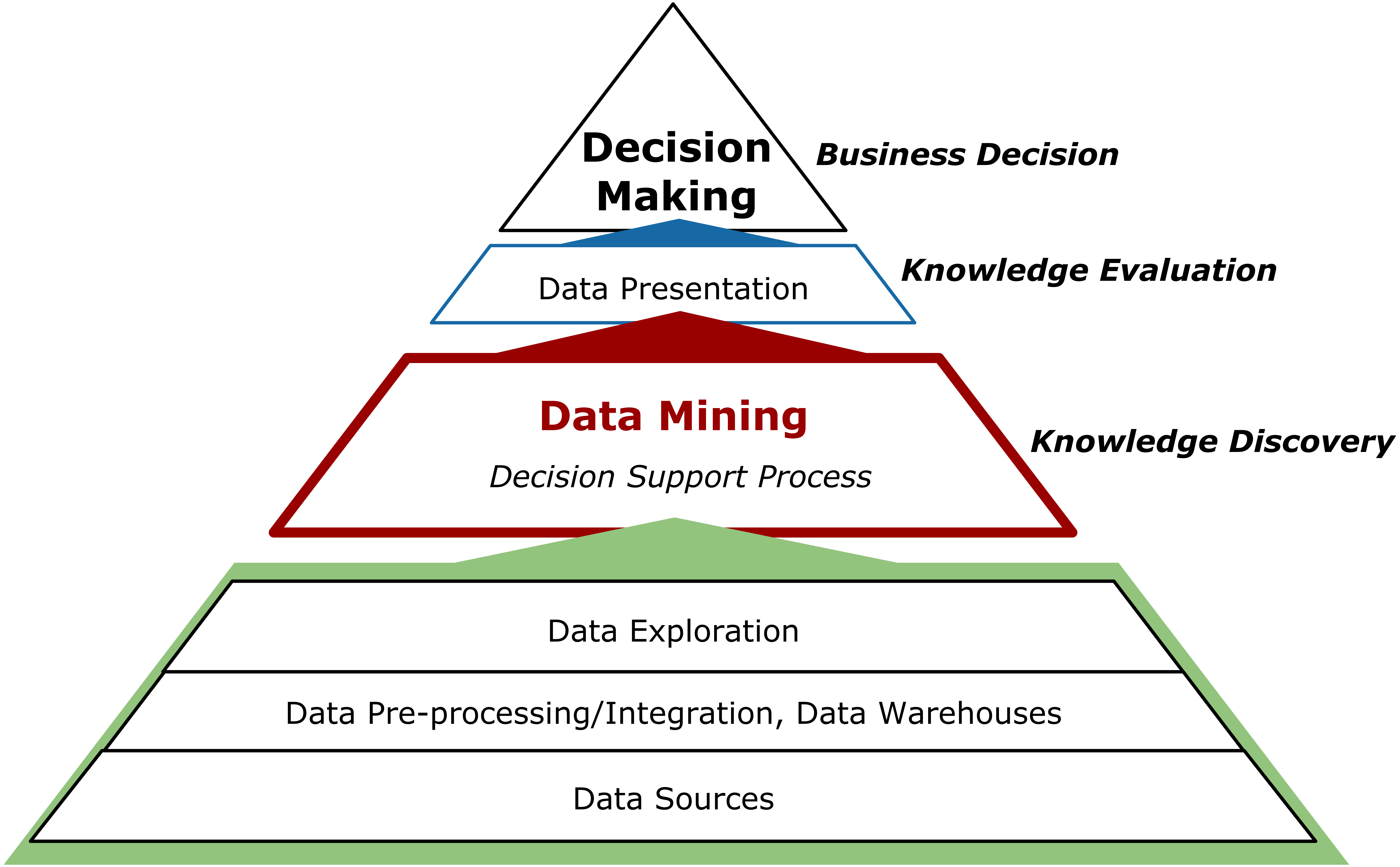


A Database View



Modified figure from Jiawei Han - https://hanj.cs.illinois.edu/bk3/

An Application View: Business Intelligence



Modified figure from Jiawei Han

Four dimensions of Data Mining

Four-Dimensions of Data Mining

- Data to be mined (Input)
- Knowledge to be discovered (Output)
- Techniques utilized (Connects Input-Output)
- Applications adopted (Where to use?)

1. Data to be Mined

Real world data can be characterized by:

Type/representation, such as itemsets, vectors/ matrices, sequence, time-series, spatiotemporal, data streams, or graphs.

Genre/application, such as transactional data, text and web, multimedia, social and information networks, biological data, or user behaviors.

2. Knowledge to be Discovered

(also known as ***data mining functionalities***)

Functionalities include:

- Lower-level output, such as patterns of data, similarity of data, or associations of data.

- Decision-driven output, such as classification, clustering, trend/deviation, prediction, and outlier analysis.

- Descriptive or predictive data mining.

3. Techniques Utilized

Data cubing, machine learning, statistics, pattern recognition, user modeling, visualization, and data-intensive computing.

4. Data Mining Applications

- Retail (advertising, market segmentation)
- Telecommunication (spam call detection)
- Banking (loan approvals, estimate credit scores)
- Social networks (Facebook, Twitter)
- Scientific discoveries (Biology data mining)
- Web search (smart question answering)
- Stock market analysis (make stock picks)
- Text mining (natural language processing)
- Clinics (health informatics)

Four-Dimensions of Data Mining

- Data to be mined
- Knowledge to be discovered
- Techniques utilized
- Applications adopted

Towards Real-World Data

- Python data structures & tools for collecting, storing, and processing data are not sufficient for data mining!
- Data, in reality, are not simple.
- There is a big gap between real data and analytics.
- Data representation bridges this gap.

Data Representation: A mathematical way to describe what data looks like.

Challenge posed by Real Data

What we are used to:

Binary Code - Christiaan Colen - https://www.flickr.com/photos/christiaancolen/20607150556 - CC BY SA 2.0

What the reality is:

Stata Center MIT - King of Hearts - https://commons.wikimedia.org/wiki/File:Stata_Center_MIT_October_2014.jp g - CC-BY-SA-3.0

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.

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.

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?

Suitable data representations allow data scientists to answer these questions in a mathematical way.

Data Representations

Messy data needs to be represented in a clear mathematical fashion before performing data mining.

Some common 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.

1. The Itemset Representation

Each data object is represented as a set of items:

$$X = \{x_1, x_2, x_3, \ldots, x_k\}$$

- 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.

	👘	👠	🎹	⌚	☂️	🔭	🏀	🎃
👩:	5	4	3	1	4	4	2	5

Attribute: a numerical property of the object.

- E.g., Kimono=5; Shoe=4; Piano=3, etc.

2. The Vector Representation

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

$$\overrightarrow{X} = \langle x_1, x_2, x_3, \ldots, x_n \rangle$$

- x_i is the numerical value of X at the i^{th} dimension (attribute).

- Each attribute is unique; cannot change order.

2. The Vector Representation

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

$$\overrightarrow{X} = \langle x_1, x_2, x_3, \ldots, x_n \rangle$$

- 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

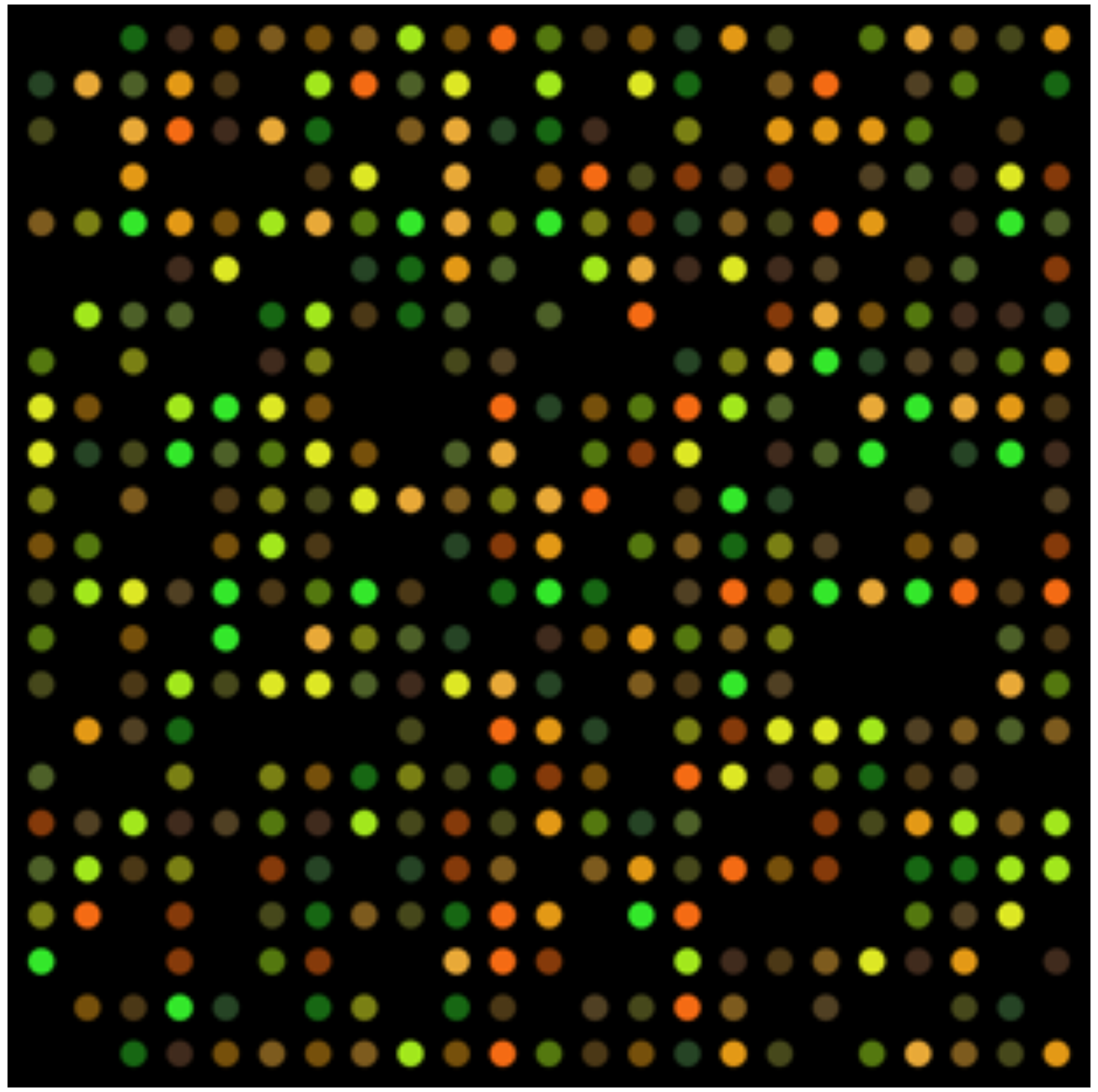
https://www.incubegroup.com/blog/recommender-system-for-private-banking/

Microarrays

Samples

Gene
Expression
Level

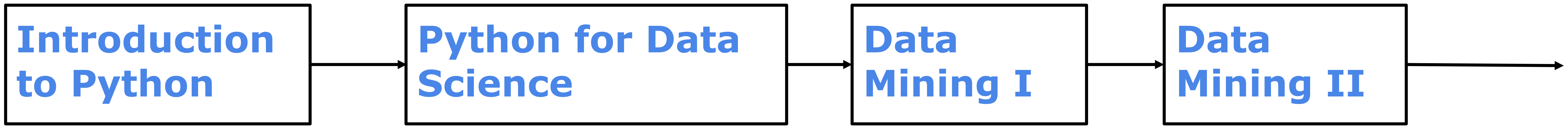
Genes



Sequence Data

Data object: curriculum paths, 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



(For a degree program, each course and its position are set in a sequential order)

3. The Sequence Representation

Data represented as a sequence of items:

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

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

Example of Sequences

DNA sequences

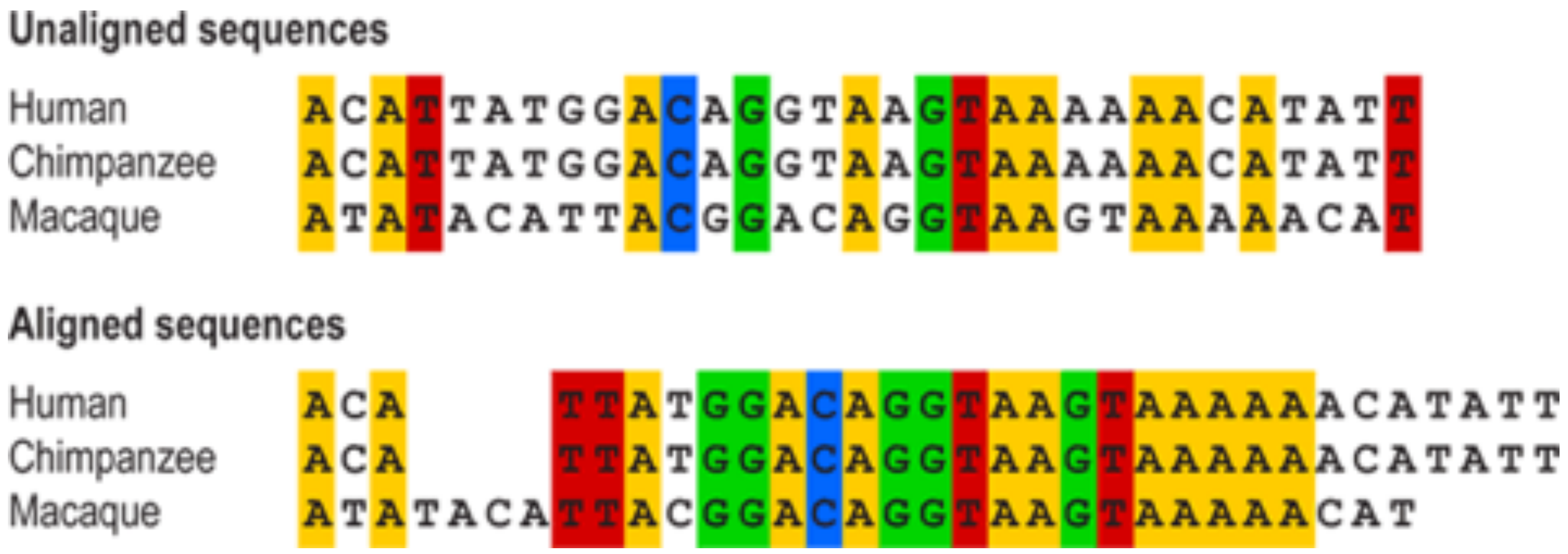
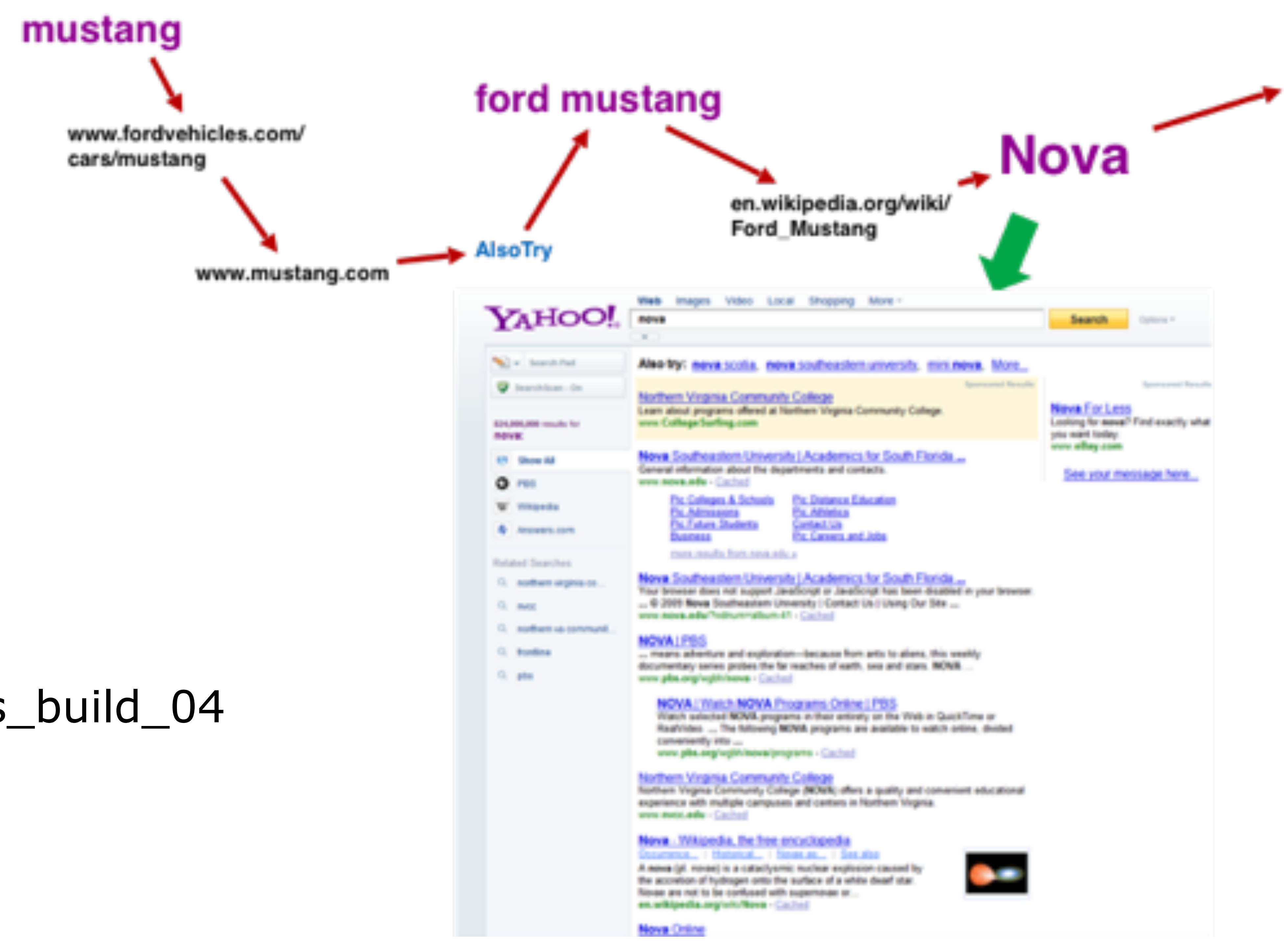


Image source: https://evolution.berkeley.edu/evolibrary/article/0_0_0/evotrees_build_04

Search sequence



Time Series Data

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

👶:	2y	3y	4y	5y	6y
height (in):	34	38	41	43	46

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

4. The Time Series Representation

- A list of timestamped measurements:

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

- 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 (S&P 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.**

5. The Spatial Representation

List of location-labeled measurements (2D):

$$X = \{(x_1, \lambda_1, \phi_1), (x_2, \lambda_2, \phi_2), \ldots (x_n, \lambda_n, \phi_n)\}$$

Longitude Latitude

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

Examples of Spatial Data

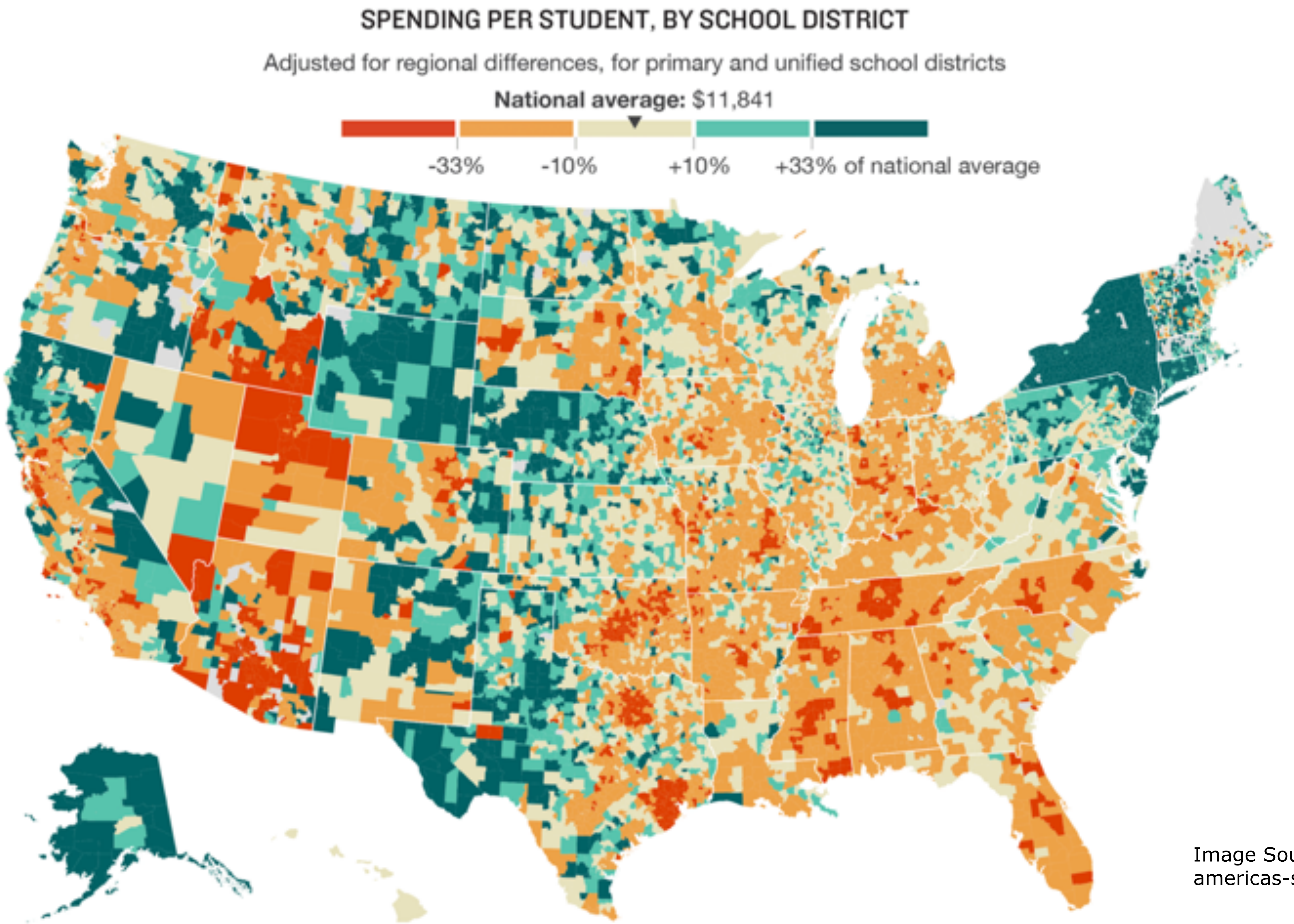


Image Source: https://www.npr.org/2016/04/18/474256366/why-americas-schools-have-a-money-problem

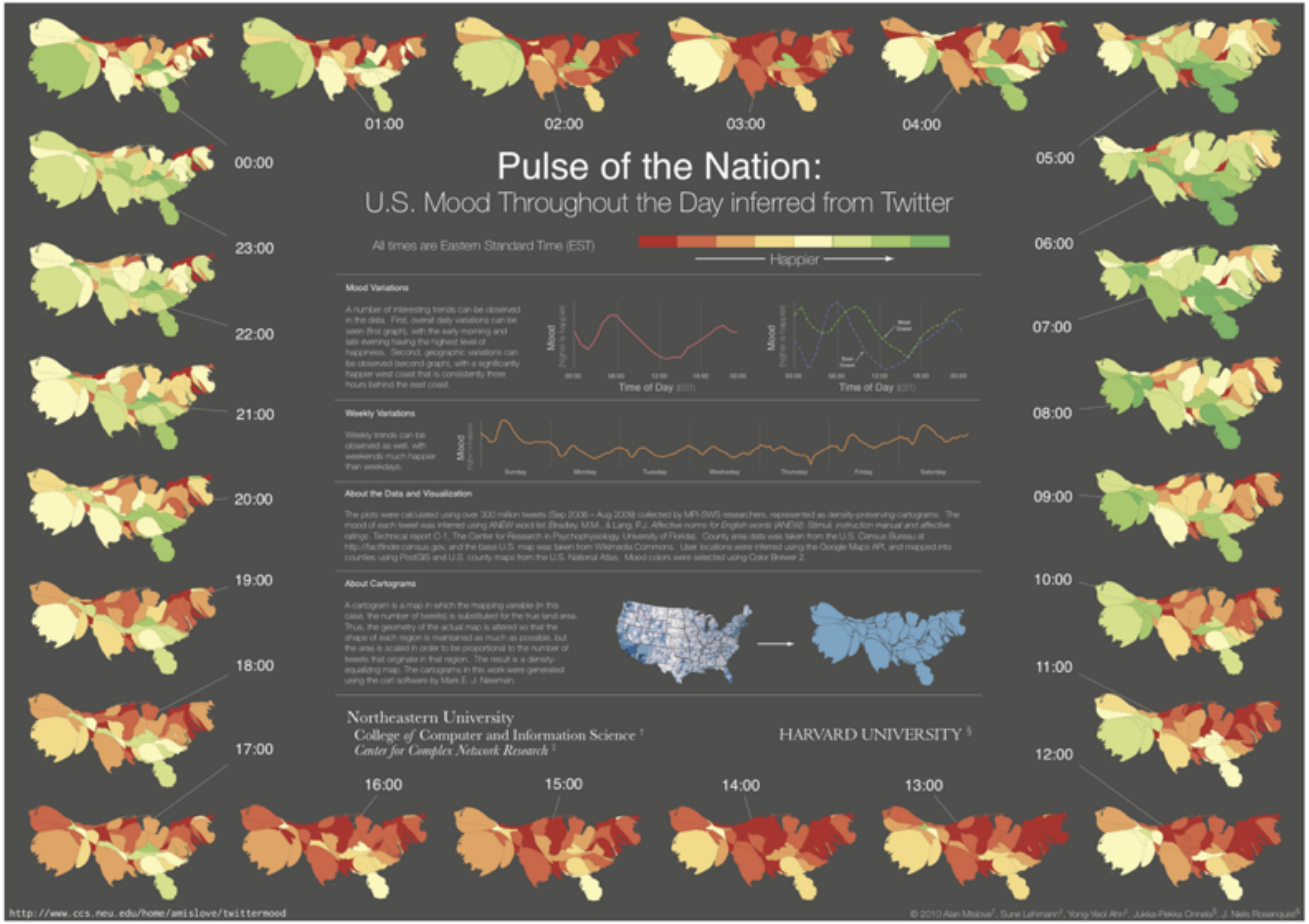
6. Spatiotemporal Data Representation

$$X = \{(x_1, \lambda_1, \phi_1, t_1), (x_2, \lambda_2, \phi_2, t_2), \ldots, (x_n, \lambda_n, \phi_n, t_n)\}$$

$$x = f(\lambda, \phi, t)$$

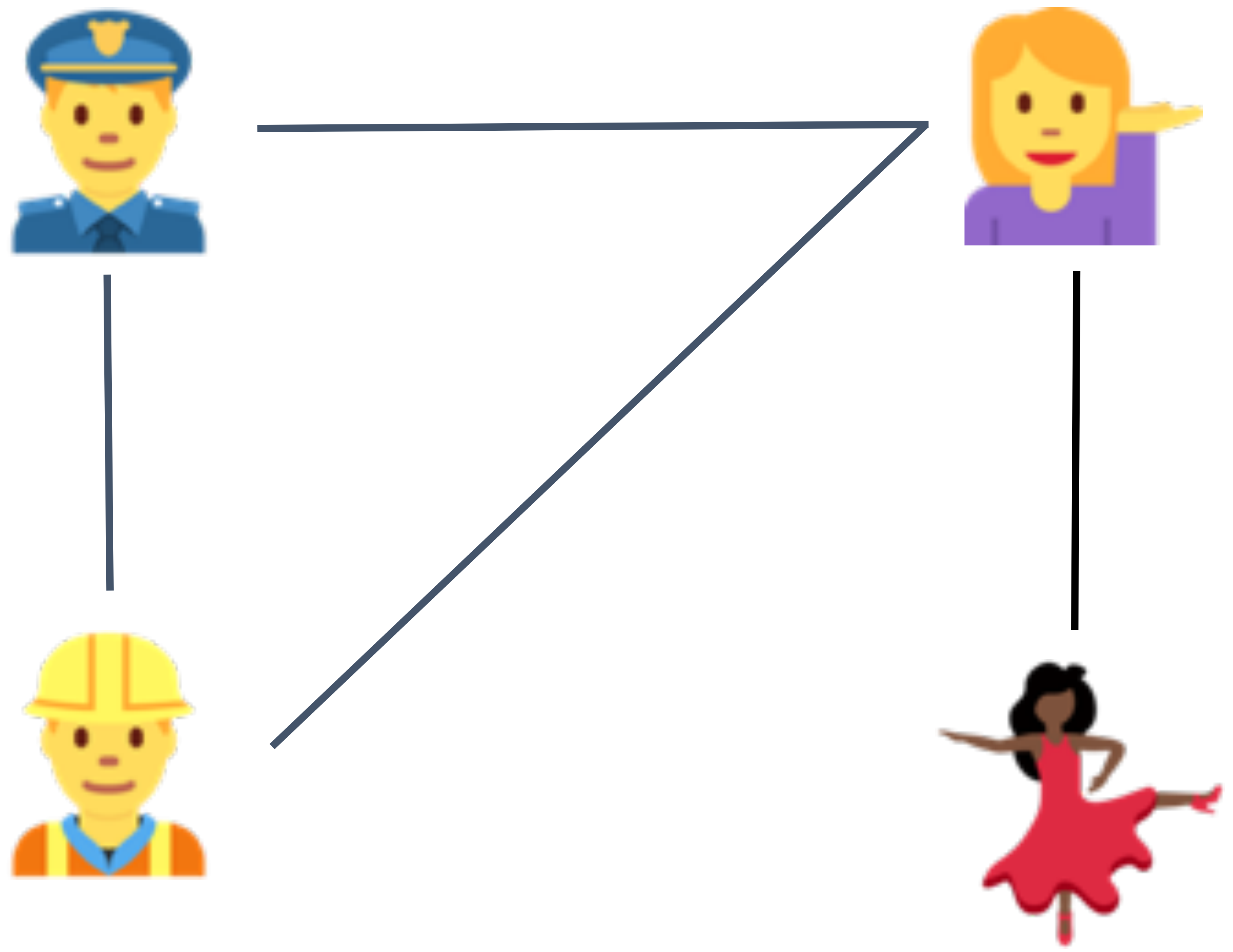
Simply add the time dimension to a spatial representation to describe spatiotemporal data.

Example of Spatiotemporal Data



Graph (Network) Data

Data objects: an online social network, the Internet, the Web.



Attribute: nodes and links

7. The Graph (Network) Representation

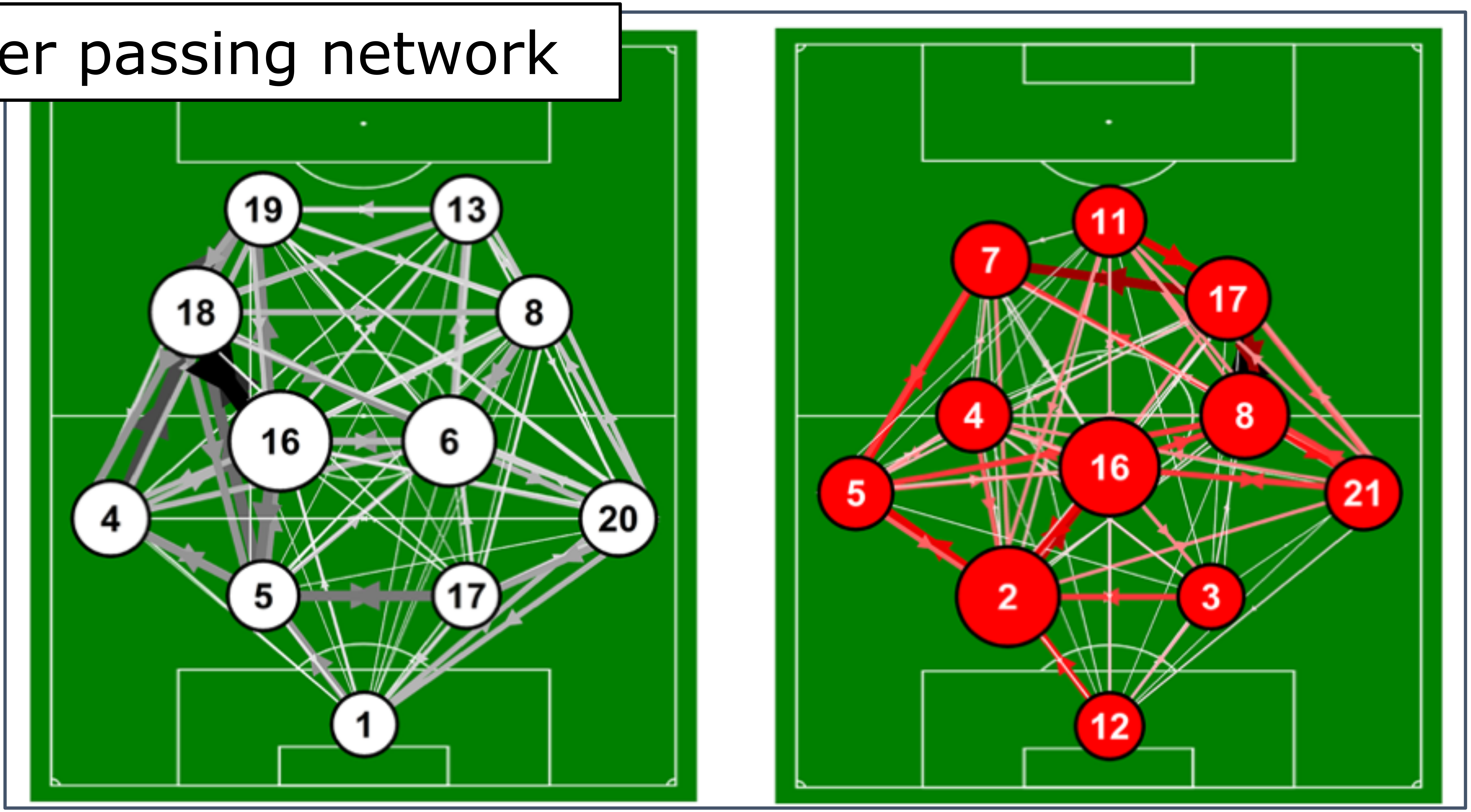
- Data formulation: $G = (V, E)$

- V is a set of nodes (vertices, entities): $V = \{v_1, v_2, \ldots, v_n\}$

- E is a set of links (edges, relations) between two nodes: $E = \{(v_i, v_j), \ldots\}$

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.

8. The Stream Representation

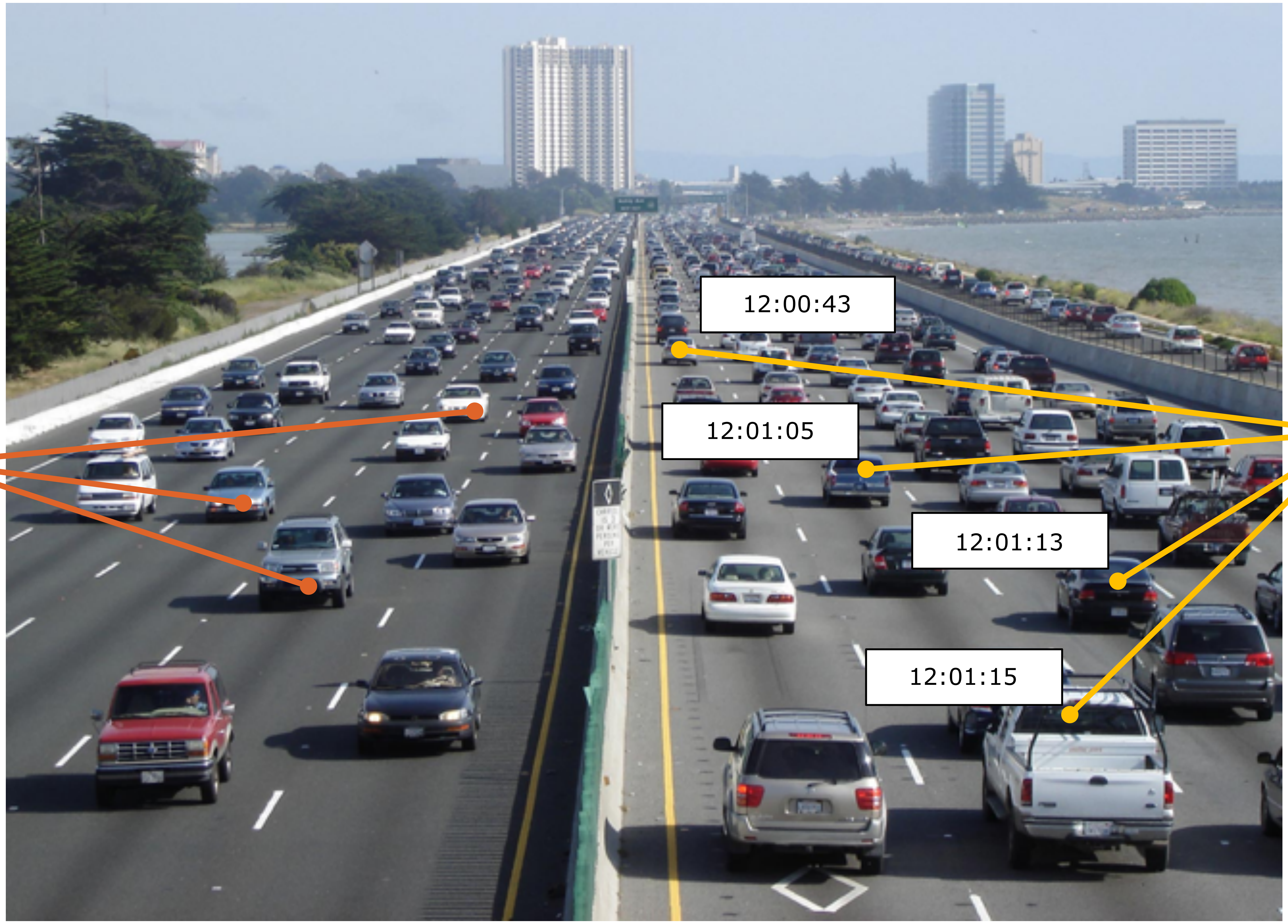
Formulation of Data $(t_k \leq t_{k+1} \leq t_{k+2}, \ldots)$:

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

X_k can be any simple or complex data object.

Examples of Data Streams

Each vehicle is an object of this view stream



https://en.wikipedia.org/wiki/Transport_economics#/media/File:I-80_Eastshore_Fwy.jpg

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

Four-Dimensions of Data Mining

- Data to be mined
- Knowledge to be discovered
- Techniques utilized
- Applications adopted

Welcome to SI 671/721!

SI 671/721- Data Mining: Methods & Applications

- Advanced graduate level course.

- Introduce the state-of-the-art of data mining.

- Different from most data mining courses:
  - Organized by different genres of data.
  - Focus on data mining applications instead of machine learning and statistical models.

- Prepare students for doing data mining research or applying data mining to other fields of research.

- Related to machine learning, statistics, database, information retrieval, natural language processing, network theory, etc.

Who should take this course?

- Graduate students who are interested in doing research in the field of data mining (providers of data mining).

- Graduate students who encounter data mining tasks in their own field (consumers of data mining)
  - E.g., business intelligence, bioinformatics, health informatics, Web analysis, social networks, …

- Graduate students who are interested in data mining applications and solving data mining challenges.

- Students who want to get a job as data scientists.

Who am I? Who are you?

<u>Me:</u>

- Assistant Professor at UMSI.
- Previously a post-doctoral researcher at MIT.
- Got Ph.D in Computer Science from U. of Pennsylvania.
- *<u>Research Interests:</u>* Some combination of Statistics, Machine Learning, NLP, and Computational Social Science.

<u>You:</u>

- *<u>Program of Study</u>:* BSI, MSI, PhD? SI or outside?
- Background?
- Why do you want to learn Data Mining?

Prerequisites

- Linear algebra: vectors and matrices.
- Probability/statistics: random variables, discrete and continuous distributions, Bayes theorem, …
  - SI 544 or equivalent (e.g., STAT 250)
- Programming: proficiency in at least one programming language (Java, C++, Perl, **Python**, etc.)
- Data manipulation skills:
  - Take SI601 and SI618 first if you don’t have such skills.

Beware ...

- This is NOT a programming course.
  - We will not teach/learn how to program, but assume that you are fluent in programming.

- This is NOT a math/statistics course.

- We will focus on (practical) algorithms and their applications.

- Check with me if you think you do not have the right prerequisites or have concerns.

Grading

2 multiple-choice quizzes (24-hour time limit): 5 x 2 = **10%**

3 Homework assignments (all programming/data analysis): 20 x 3 = **60%**

Course Project: **30%**

- –Proposal: 10% (due 11/1 Week 10)
- –Final presentation: 10% (UMSI Fall exposition ~12/10)
- –Final report: 10% (Due in finals week ~12/13)

Extra grade: **2%** for students who help answer others' questions and further the discussion on a topic on Slack.

What are homeworks like?

- Each homework has a large programming component designed around a particular task/dataset
  - Datasets will be medium size and workable on a slow laptop
  - They can take a while so start early!

- Each homework has a written component describing your results and analysis
- We recommend you use Jupyter Notebooks for Homeworks and python libraries such as numpy, scikit-learn, pandas etc.

Course Project

- Research project or Software tool development
- Example:
  - ✓ A public opinion/health/topic monitor in social media
  - ✓ A de-identification tool for health records using conditional random fields
  - ✓ An efficient network clustering method for very large scale networks
  - ✓ A comparative study of community detection algorithms
  - ✓ Mining frequent sequential patterns in Twitter diffusion paths
  - ✓ A primitive study of correlating social media with the stock market
  - ✓ Author identification of historical literature (essays and fiction)
- Replication of recent research papers
- Option to work in small groups (2-3 people)

Administrivia (I)

Regular meetings: Mondays, 8:30 – 10:00 am ET, via Zoom.

Office hours: Mondays 1-3pm ET, in-person @ 3389 NQ or via Zoom.

GSIs: Two amazing GSIs
(1) Yulin Yu (yulinyu@umich.edu)
(2) Anmol Panda (anmolp@umich.edu)

Zoom links for instructor and GSI office hours as well as discussion sections are in syllabus.

Administrivia (II)

The course has required discussion sections accompanying most lectures.

They will implement concepts covered in the class via Python Jupiter notebooks.

They will go a long way in helping you in homeworks/ final project and provide background for understanding material covered in lectures. So please attend them!

Administrivia (II)

There are four discussion sections (two led by Yulin and two by Anmol).

Three of the four are in-person. You need to attend ONLY 1 of them.

1. (Remote) Mondays 10-1130 AM
2. (2185 NQ) Mondays 10-1130 AM
3. (1245 NQ) Mondays 530-7 PM
4. (B124 MLB) Mondays 530-7 PM

Administrivia (III)

- Use right channels for communications
  - Most questions -> Slack (you should all have been added to the course slack channel. Please monitor Slack regularly.)
  - Complex technical questions -> Office hours

- Deadline for submitting assignments: Monday 11:59pm Eastern Time.
- 3 Day grace buffer period (overall)

Going forward

- Everything will be posted to Canvas including syllabus.
- We'll use Slack for discussions.
- Come to our office hours if you have any questions.
- Most Importantly: **PLEASE DON'T PANIC.**
  - Let's fight these tough times together.
  - We understand that not everyone is local. So, we will try to make accommodations for everyone to make things work & ensure that you learn something useful in this course.

Thank You

Questions?