CSCI 4502/5502

Data Mining - Fall 2023 - Lecture 2 - September 5

Ravi Starzl, PhD

Introduction to Data Mining

The Cornerstone of Data Science

Ravi Starzl, PhD - Data Mining Fall 2023

Into the Digital Era

1. People's daily lives
   - 5 billion Internet users
   - 500 million+ tweets a day
2. Scientific Discovery
   - LHC: 15 PB/year; LSST: 20 TB/night
3. Industrial and Systems:
   - 1 Boeing 787 10 TB/flight
4. IDC Digital Universe Report
   - 2009: 0.8ZB => 35ZB (2020)
   - 2013: 4.4ZB => 44ZB (2020)

Ravi Starzl, PhD - Data Mining Fall 2023

Why Data Mining?

1. Data Explosion: KB,MB,GB,TB,PB,EB,ZB...
   - Data creation, transmission, storage, sharing, processing
   - We are drowning in data but starving for knowledge

2. Need automated analysis of massive data

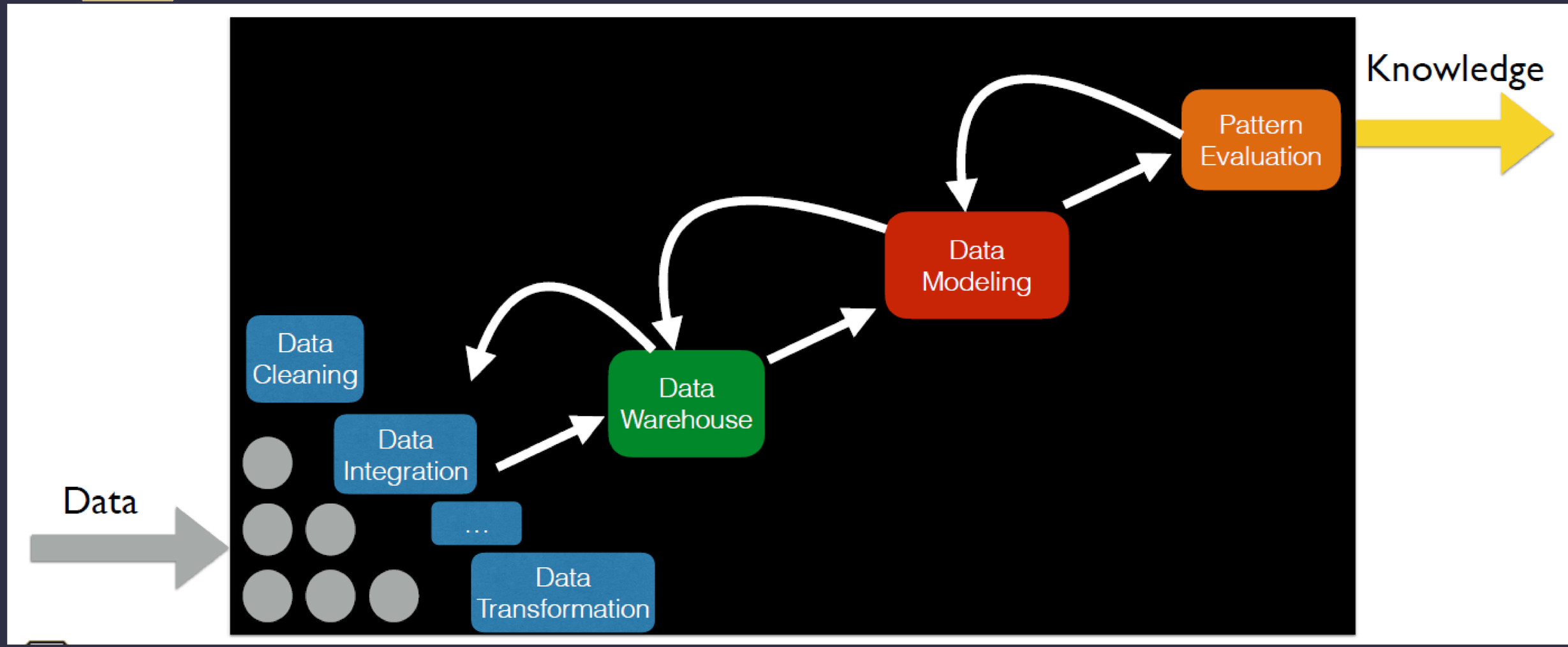
What is Data Mining?

- Data Mining (knowledge discovery from data)
  - Extraction of interesting patterns or knowledge from huge amounts of data
  - Interesting: Valid, previously unknown, potentially useful, ultimately understandable by human
  - Huge amounts of data: Scalability, efficiency

Data Mining Application Areas

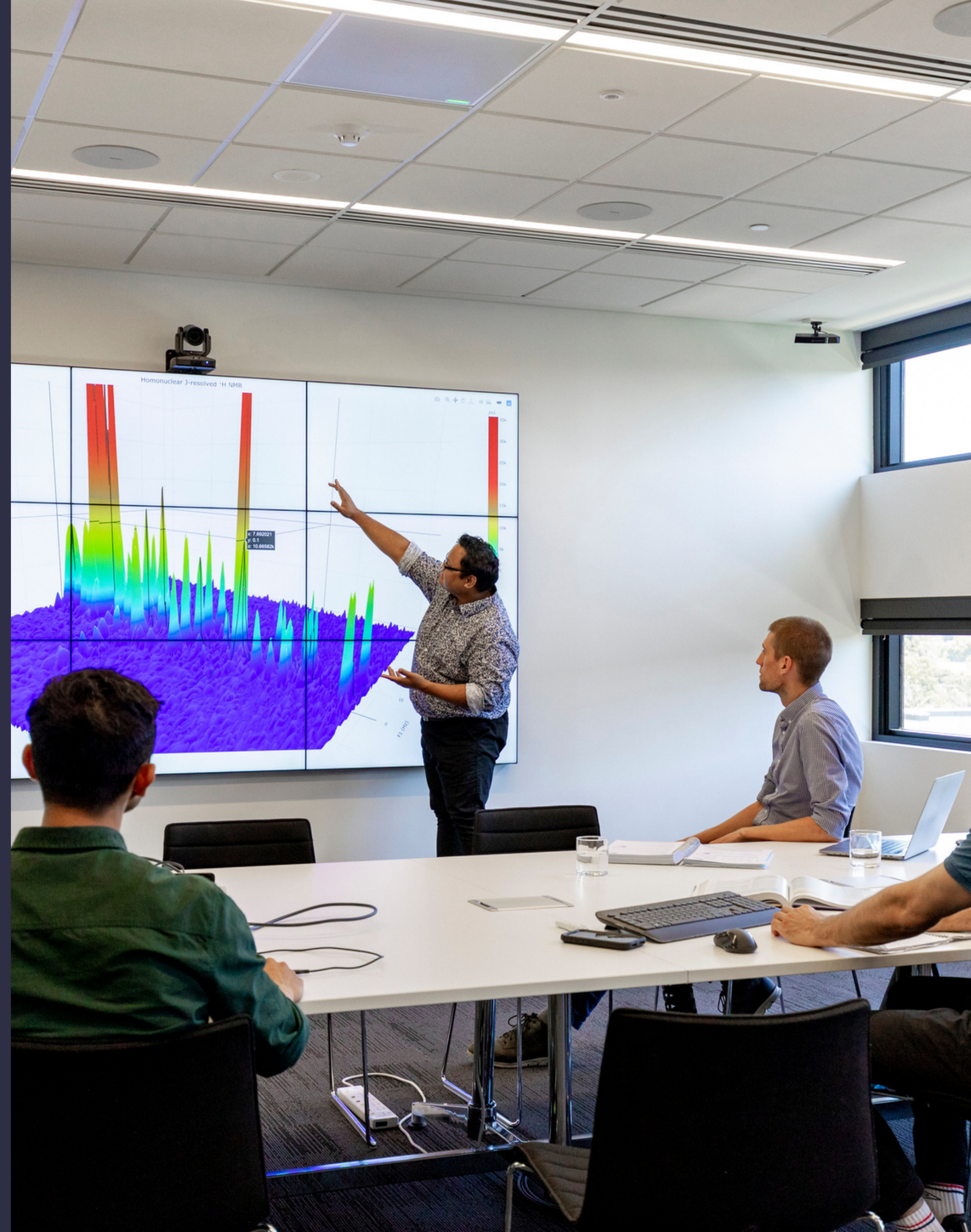
- Science
  - astrophysics, bioinformatics, drug discovery, sustainable energy, oceanography, seismology, etc.
- Business
  - market analysis, fraud detection, target marketing, churn prediction, product recommendation, etc.
- Web
  - search engines, advertising, online social networks, trending, etc.
- Government
  - surveillance, crime detection, transportation, development, etc.
- A lot more!

Data Mining Pipeline



Data Mining: Various Views

- Data View
  - types of data to be mined
- Knowledge View
  - types of knowledge to be discovered
- Method View
  - types of techniques utilized
- Application View
  - types of applications adapted



Data View

- The 3Vs, 4Vs, and 5Vs

Volume

Variety

Veracity

Velocity

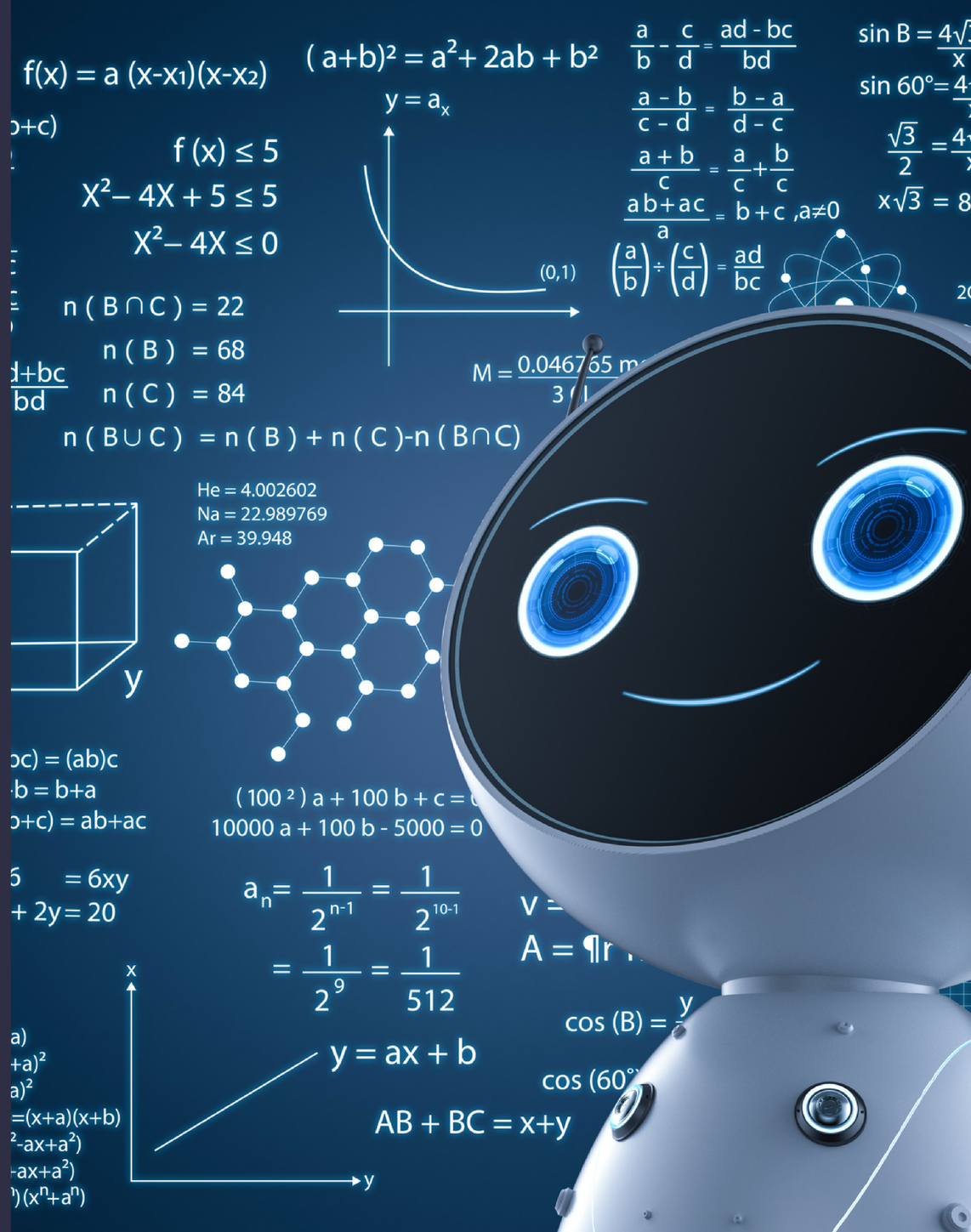
Value

Data View

- Database Oriented
  - relational, transactional
  - data warehouse, NoSQL
- Sequence, stream, temporal, time-series data
  - Trend analysis, anomaly
  - Spatial, spatial-temporal data
- Text, multimedia, Web data
  - topic detection, similarity, popularity, sentiment
- Graph, social networks data
  - substructures, shared interests, influencers, information diffusion

Knowledge View

- Concept/class description
- Frequent patterns, associations, correlations
- Classification and prediction
- Cluster analysis
- Outlier analysis
- Evolution analysis



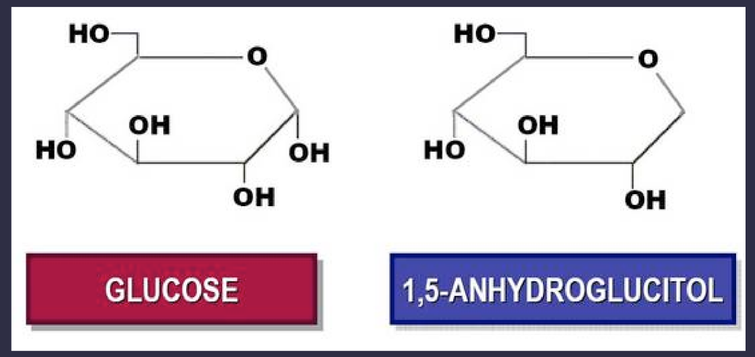
Ravi Starzl, PhD - Data Mining Fall 2023

Concept / Class Description

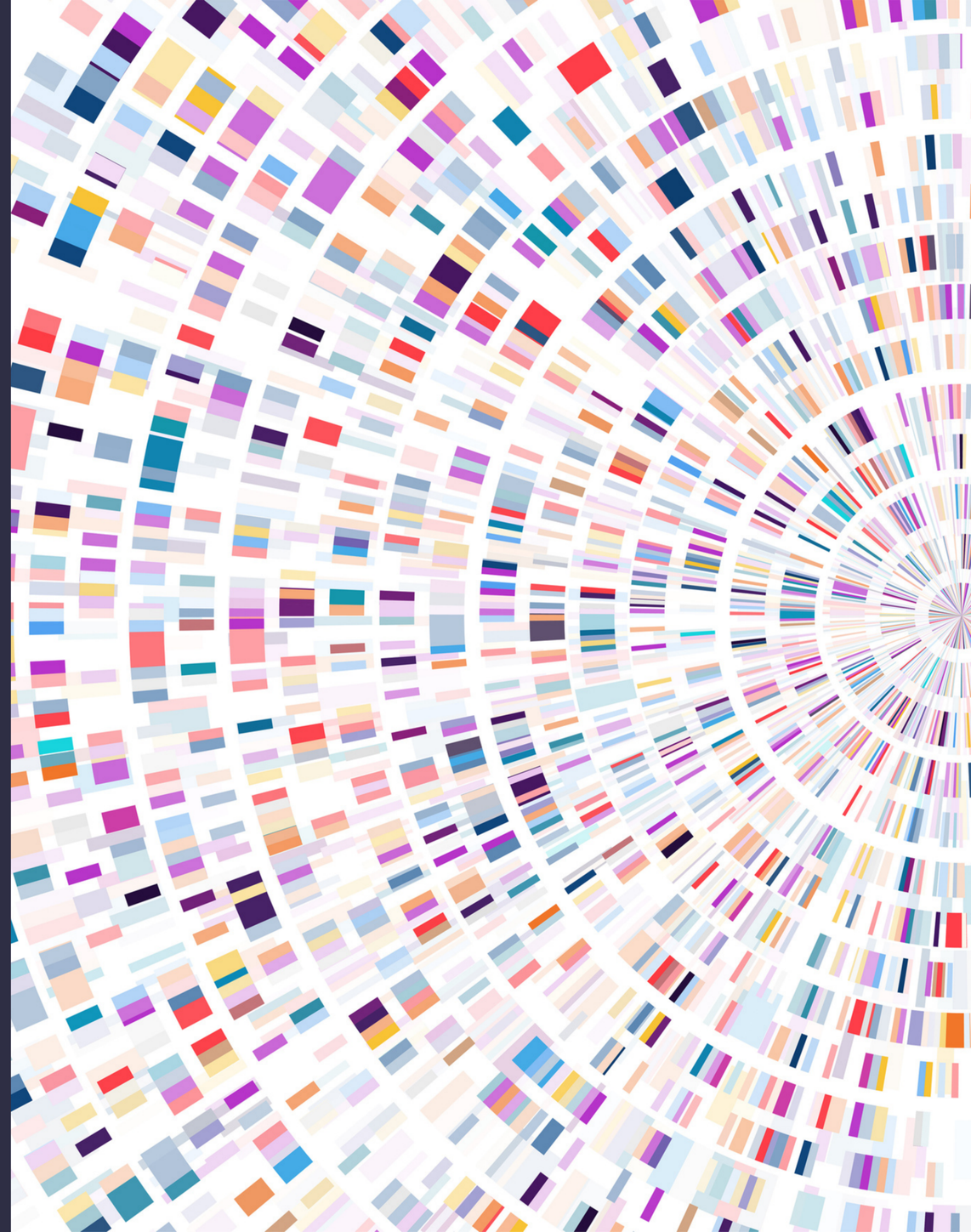
- Data characterization (summarization)
- Customers who spend $1000 a year
- Age 40-50, employed, good credit ratings
- Data discrimination (contrast)
- Frequent vs. infrequent customers: e.g., age, education, employed
- Dry vs. wet regions: e.g., precipitation, humidity, temperature

Frequent Patterns

- Frequent itemsets
  - e.g., (milk, bread, egg), (beer, diaper)
- Frequent sequences
  - e.g., <printer, paper>, <dinner, movie>
- Frequent structures



http://www.endotext.org/diabetes/diabetes12/figures/figure12.jpg



Associations

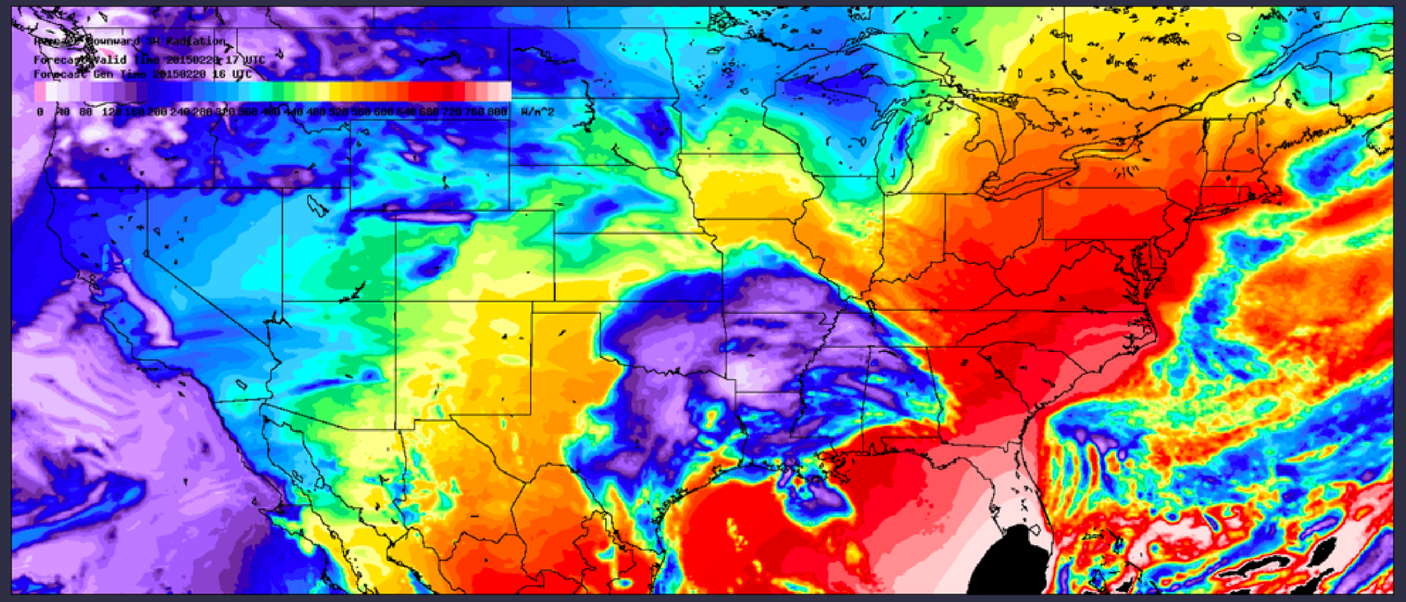
- Association analysis
  - buys (X, milk) => buys (X, bread)
  - [support = 0.5%, confidence = 75%]
- Minimum support (or confidence) threshold
- Support
  - chance of A and B appearing together
- Confidence
  - if A appears, chance of B appears

Classification

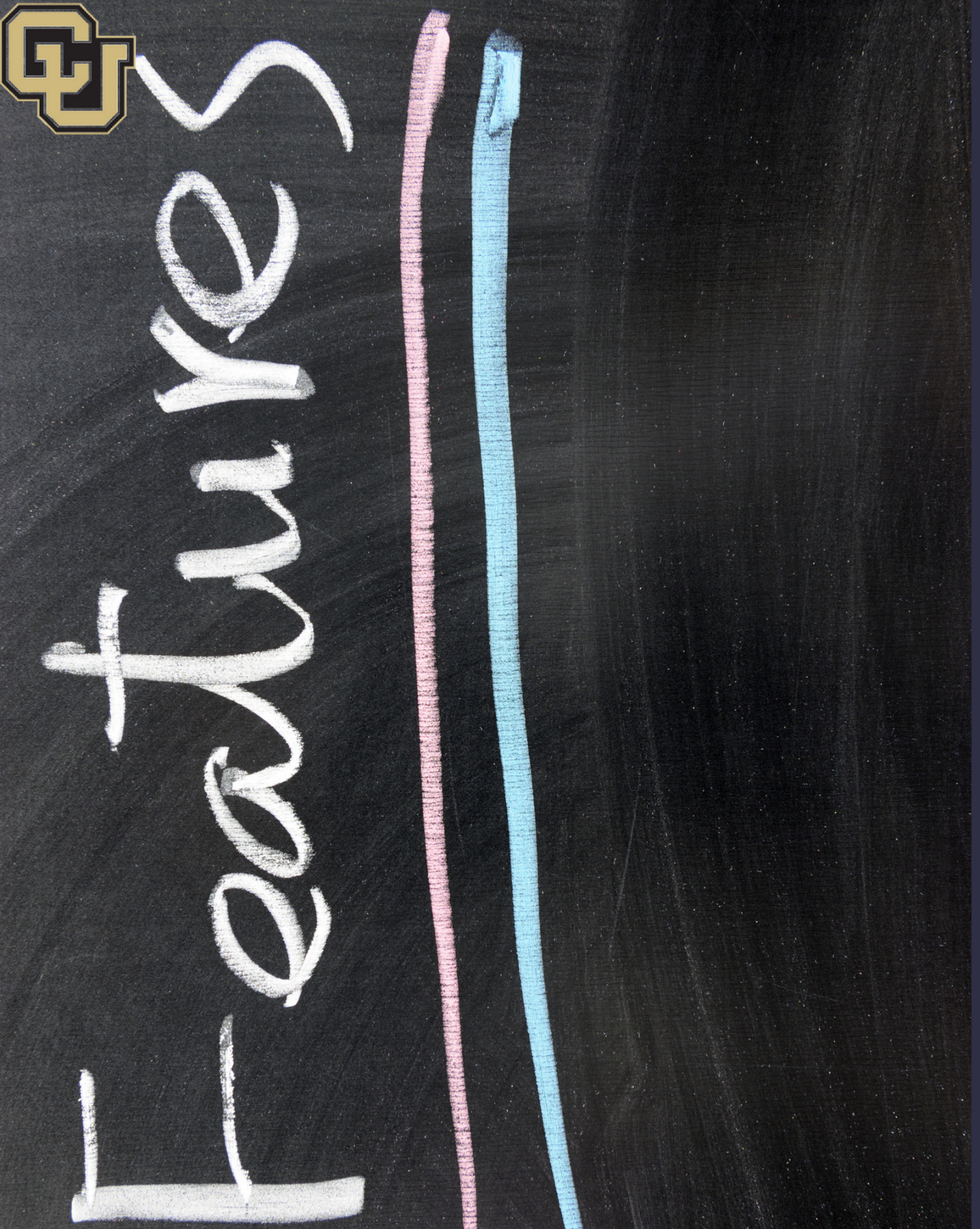
- Finding a model that describes and distinguishes data classes or concepts
- Training data
- IF-THEN rules, decision tree, neural network

Prediction

- Numerical prediction: continuous-valued instead of class labels
- E.g., weather, stock price, traffic

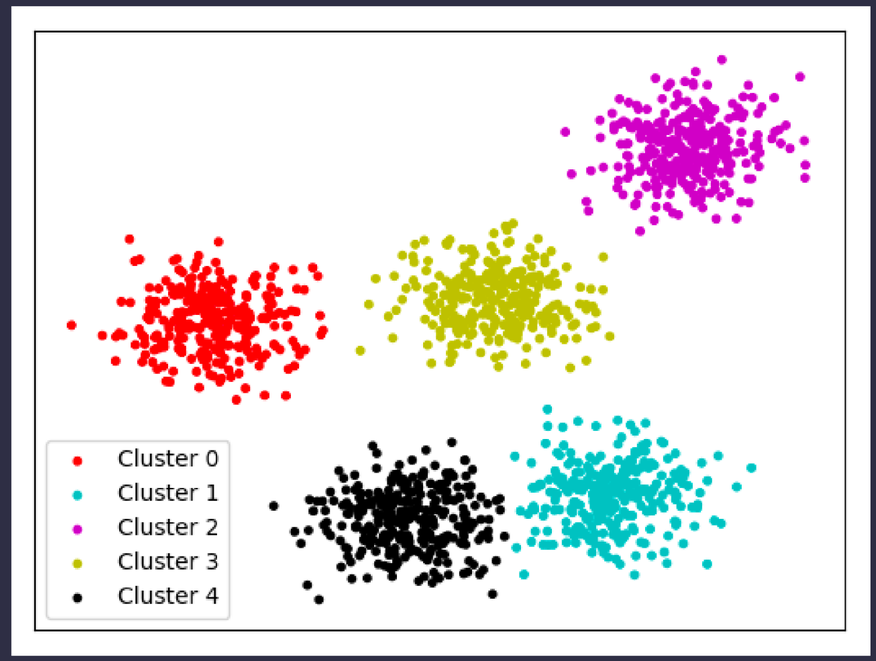


https://nar.ucar.edu/sites/default/files/labs/ral/2017/2.3.weather_prediction_3.png



Cluster Analysis

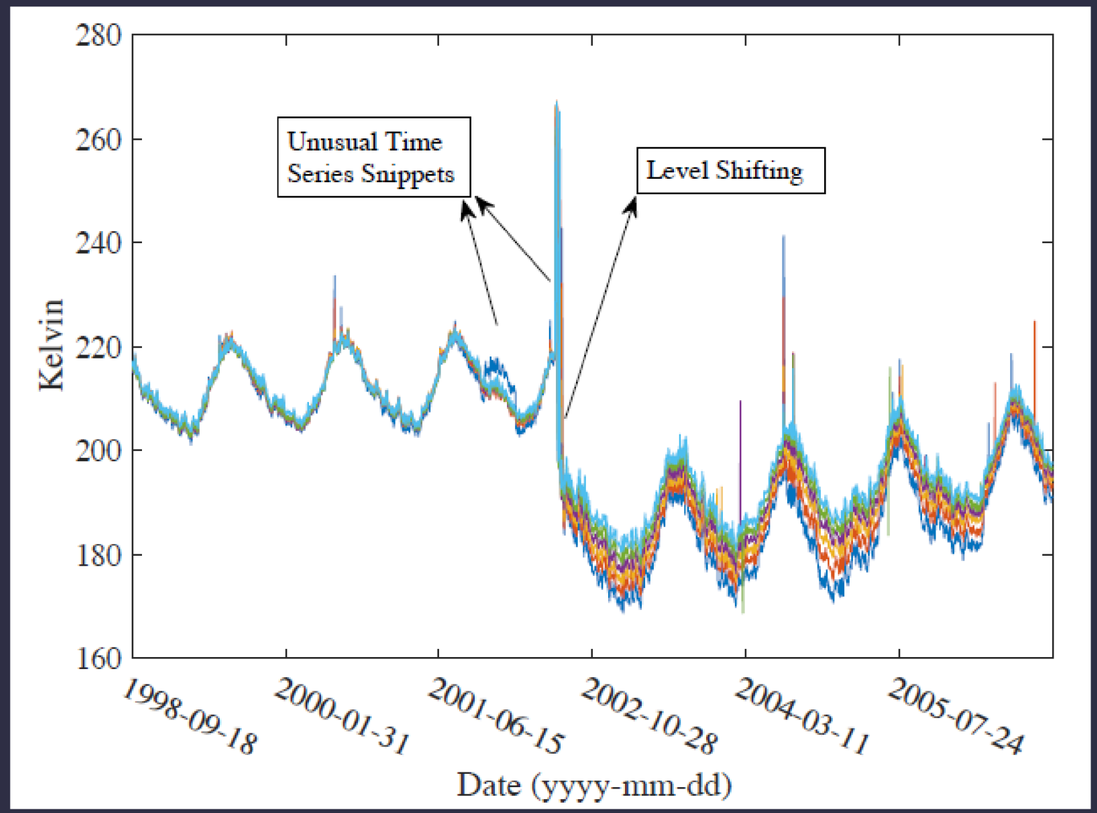
- Class labels unknown
- Intraclass similarity
  - maximize, closeness
- Interclass similarity
  - minimize, separation
- Hierarchical



https://miro.medium.com/max/1052/1*RsF6MMkuv0eECd_6m0-otw.png

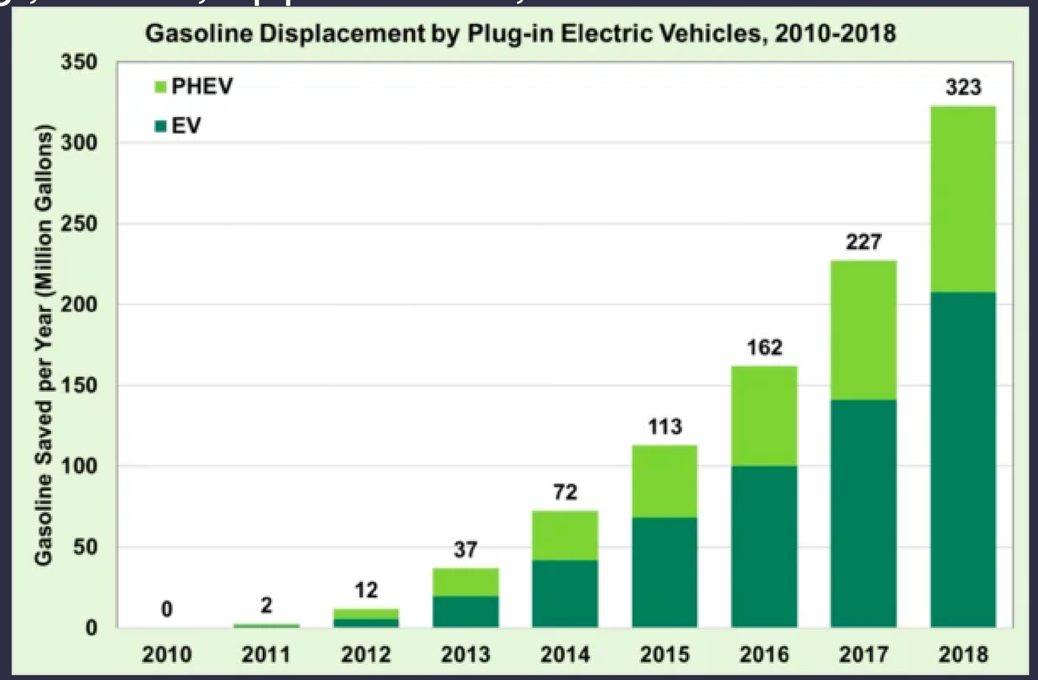
Outlier Analysis

- Outliers
  - do not comply with the general model
- Noise or exception
- Fraud detection, rare event analysis
- E.g. credit fraud analysis



Trend and Evolution Analysis

- Trends, deviations
- Sequential pattern mining
  - e.g., traffic congestion
- Periodicity analysis
- E.g., music, applications, etc.



https://www.osti.gov/biblio/1506474-assessmentlight-duty-plug-electric-vehicles-united-states

Market Analysis / Management

- Data sources: credit card transactions, club cards, customer calls, etc.
- What types of customers buy what products
- What factors attract new customers
- Target marketing, product recommendation, discount
- Fraud detection

Are The Patterns Interesting?

- Interesting pattern
  - valid on new/test data with some certainty
  - novel
  - potentially useful
  - ultimately understandable by humans
- Objective measures
  - e.g., support, confidence, false positive/negative, accuracy
- Subjective measures
- Completeness, exclusiveness

Major Issues In Data Mining

- Mining technology
  - mining different knowledge from diverse data (maybe noisy or incomplete)
  - pattern evaluation: interestingness
  - efficiency, effectiveness, scalability
  - parallel, distributed, incremental mining
  - incorporation of background knowledge
  - integration of discovered knowledge with existing knowledge

Major Issues In Data Mining

- User interaction
  - data mining query languages, ad-hoc mining
  - expression and visualization of results
  - interactive mining at multiple granularities
- Applications and social impacts
  - domain-specific data mining
  - applications of data mining results
  - protect data security, integrity, privacy

Data Science Ethics

- Data ownership
- Privacy, anonymity
- Data and model validity
- Data and model bias (algorithmic fairness)
- Interpretation, application, societal consequence

Ravi Starzl, PhD - Data Mining Fall 2023

Data Mining Resources

- ACM SIGKDD: https://www.kdd.org/
- Conferences
  - KDD: tutorials, research, applied data science, KDD Cup, sponsors
  - SDM, ICDM, WSDM, CIKM, ICDE, TheWebConference (formerly WWW),
  - SIGIR, ICML, CVPR, NeurIPS (formerly NIPS), SIGMOD, VLDB, and more
- Journals
  - TKDE, TKDD, DMKD, TPAMI, and more

Summary

- Chapter 1: Introduction to data mining
  - Data mining: discover interesting patterns in huge amounts of data
  - Data mining pipeline
  - Different views: data, knowledge, method, application
  - Measure of pattern interestingness
  - Major issues in data mining

Thank you

A special thank you to Qin Lv for her slides,
on which this lecture is based