CS (STAT) 5525: Data Analytics I

Introduction to Data Mining Problems, Concepts, and Algorithms

(3 credits, CRNs: 13294, 13295, 19845, 19846)

Anuj Karpatne

Assistant Professor, Computer Science

Virginia Tech

Torgersen Hall 2160F,

karpatne@vt.edu

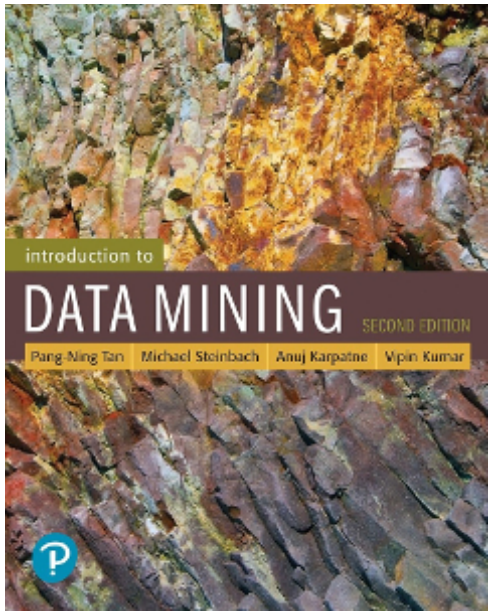
https://people.cs.vt.edu/karpatne/

Data Mining: Introduction

Lecture Notes for Chapter 1

Introduction to Data Mining, 2nd Edition

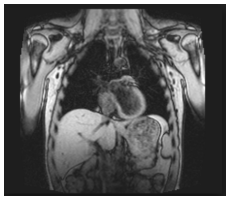
By Tan, Steinbach, Karpatne, Kumar

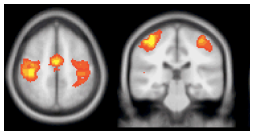


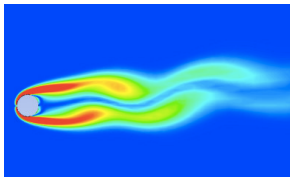
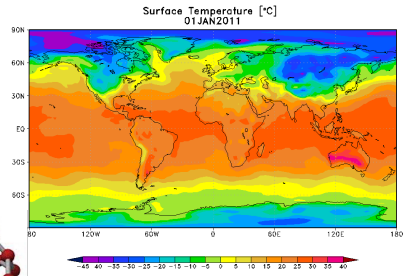
Visit the book webpage at

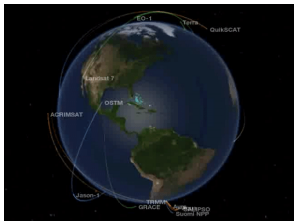
www.cs.umn.edu/~kumar/dmbook

Large-scale Data is Everywhere!





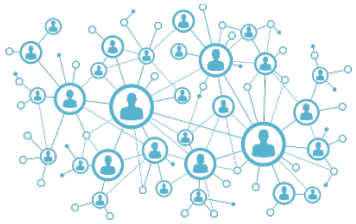




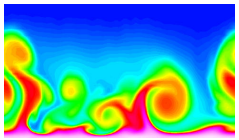


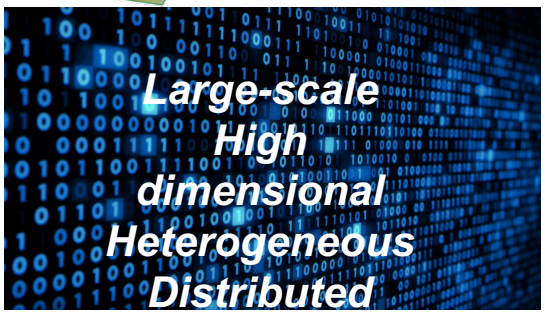
There has been enormous growth of data in both commercial and scientific arena due to advances in data generation, storage, and retrieval technologies

Golden Age of Data Science

Data → ***Data Science*** → Knowledge





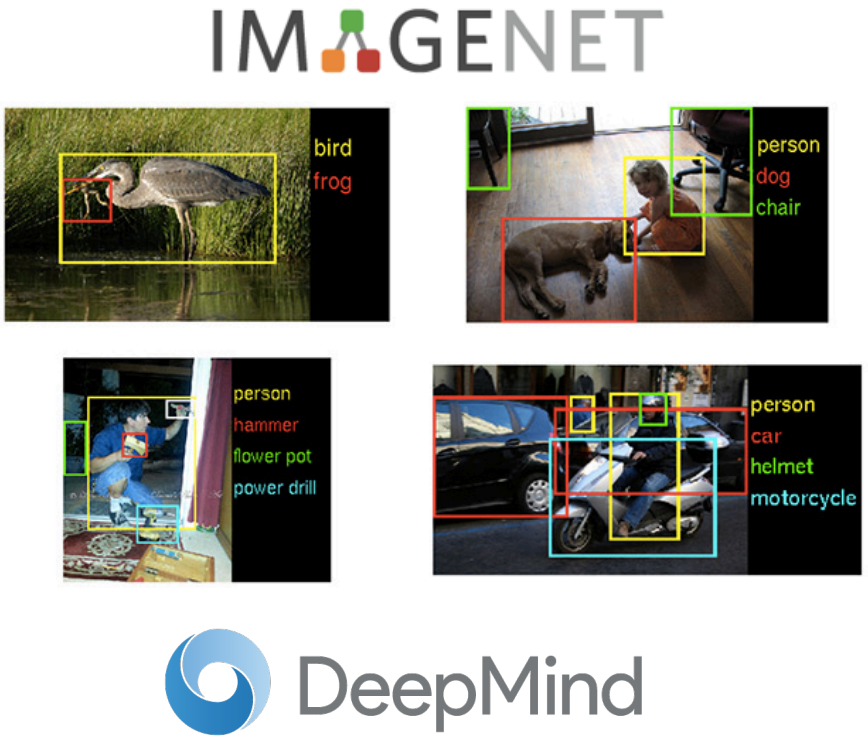


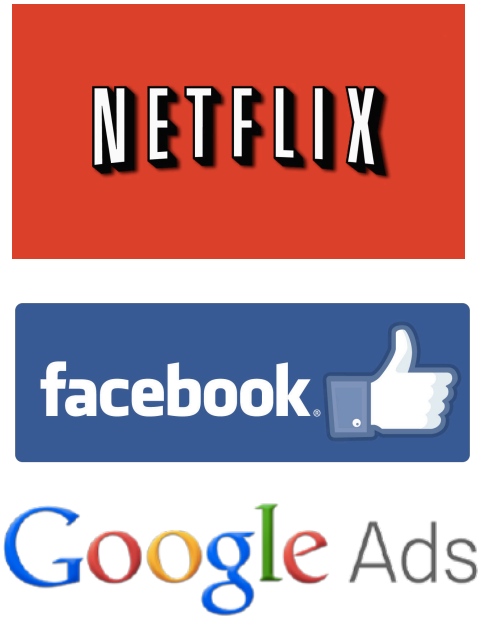
Machine Learning
Artificial Intelligence
Pattern Recognition
Data Mining / Data Analytics

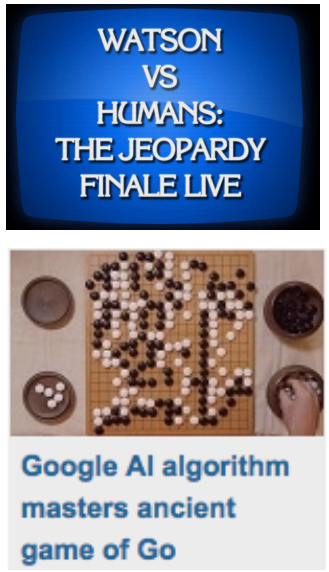
- **Patterns**
- **Models**
- **Relationships**

Automated tools for knowledge extraction from large volumes of data

Why Data Mining? Commercial Viewpoint



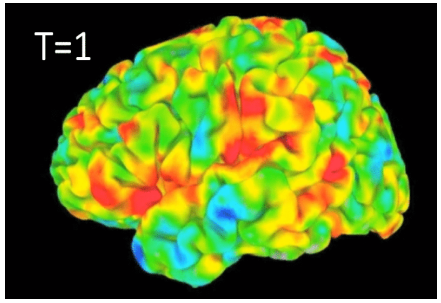




- Lots of data is being collected and warehoused
- Competitive pressure is strong

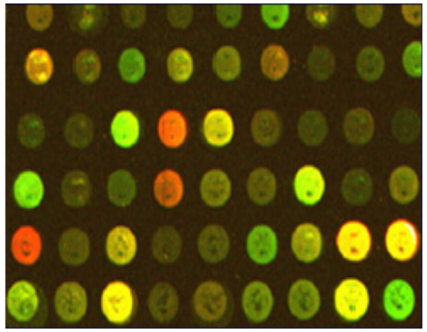
Why Data Mining? Scientific Viewpoint

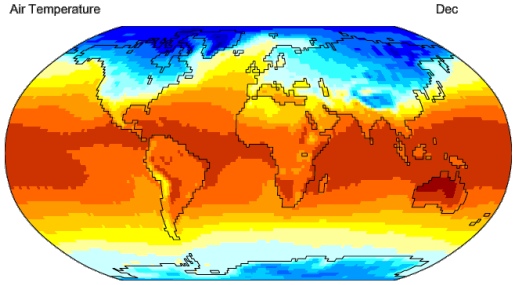
- Data collected and stored at enormous speeds
  - remote sensors on a satellite
    - NASA EOSDIS archives over petabytes of earth science data / year
  - telescopes scanning the skies
    - Sky survey data
  - High-throughput biological data
  - scientific simulations
    - terabytes of data generated in a few hours
- Data mining helps scientists
  - in automated analysis of massive datasets
  - In hypothesis formation



fMRI Data from Brain


Sky Survey Data

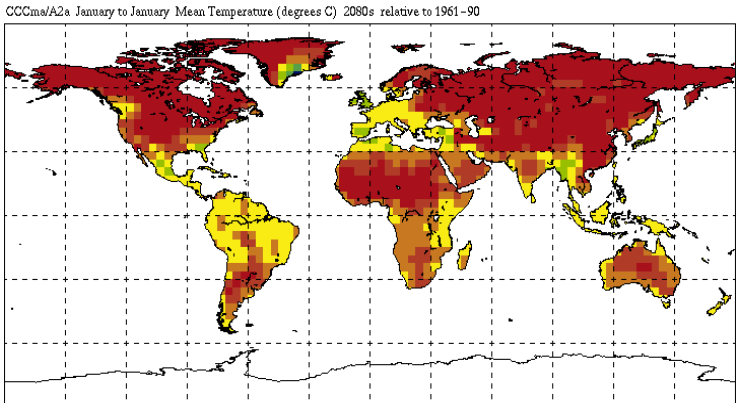

Gene Expression Data



Surface Temperature of Earth

Great Opportunities to Solve Society's Major Problems

Improving health care and reducing costs



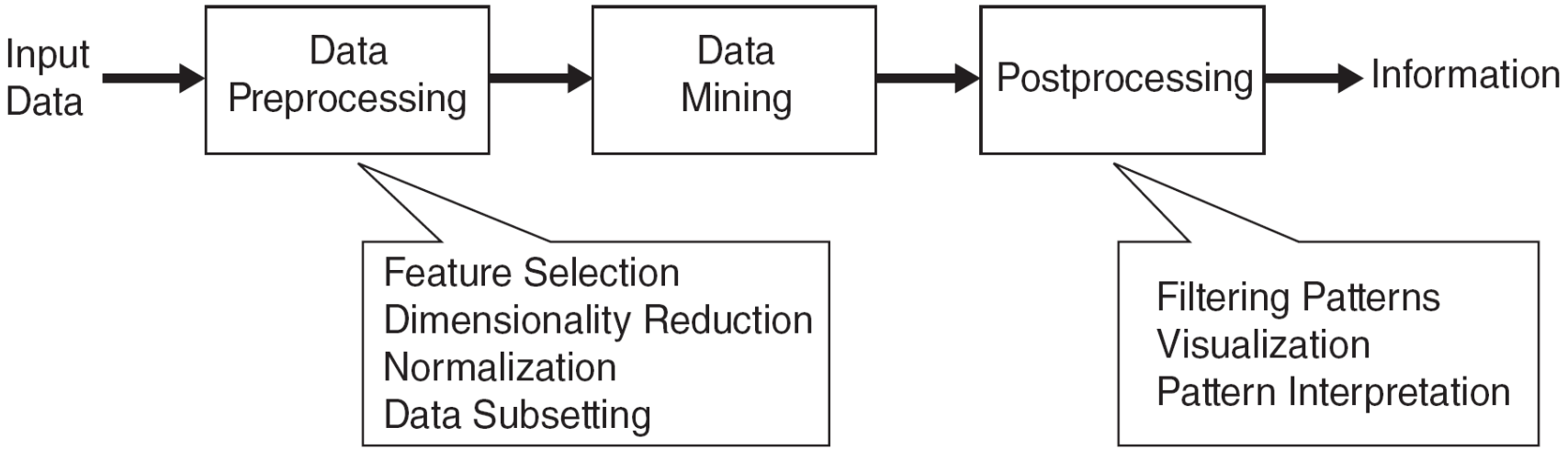
Predicting the impact of climate change

Finding alternative/ green energy sources

Reducing hunger and poverty by increasing agriculture production

What is Data Mining?

- Many Definitions
  - **Non-trivial** extraction of **previously unknown**, **useful,** and **interpretable** patterns from data



What is not Data Mining?

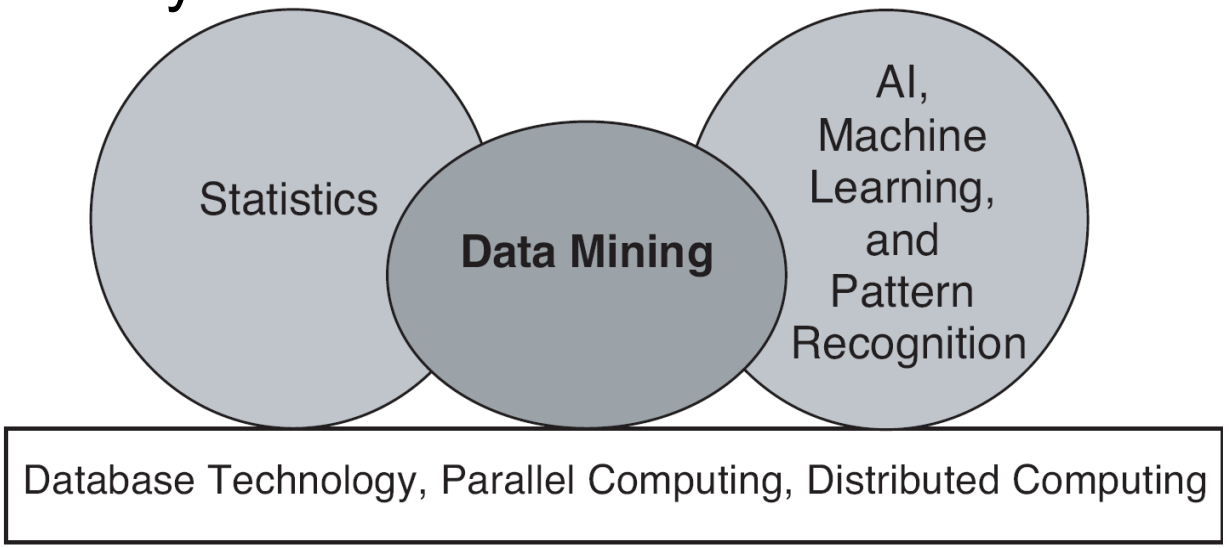
- **What is not Data Mining?**
  - Look up phone number in phone directory
  - Query a Web search engine for information about "Amazon"

- **What is Data Mining?**
  - Certain names are more prevalent in certain US locations (O'Brien, O'Rourke, O'Reilly… in Boston area)
  - Group together similar documents returned by search engine according to their context (e.g., Amazon rainforest, Amazon.com)

Origins of Data Mining

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems

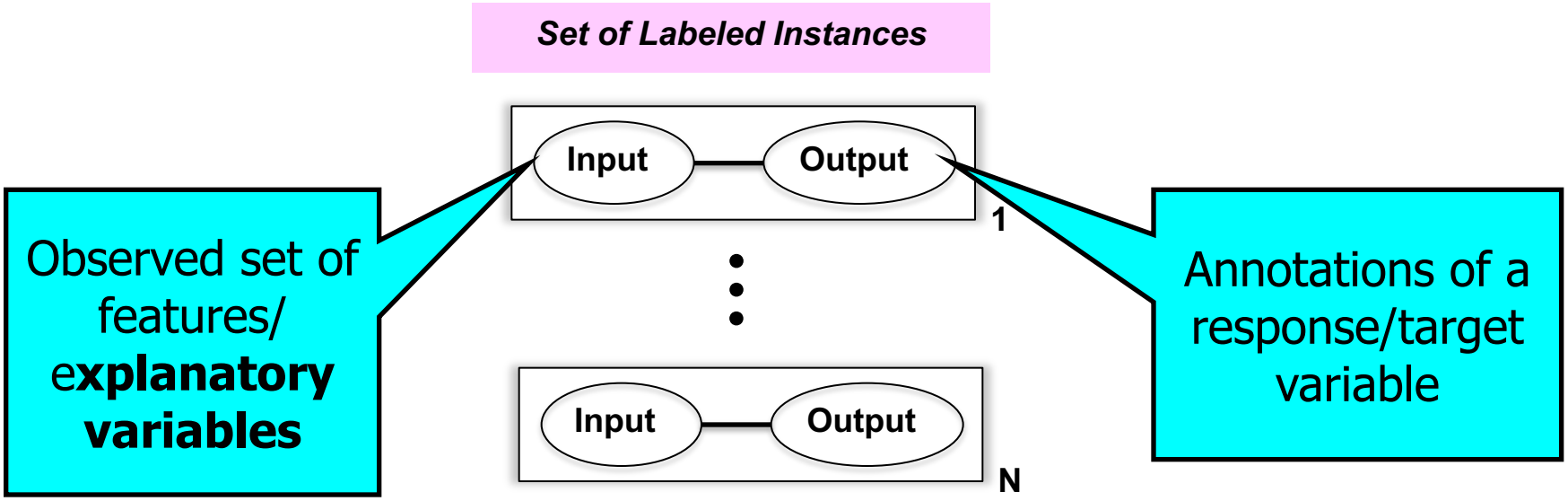
- Traditional techniques may be unsuitable due to data that is
  - Large-scale
  - High dimensional
  - Heterogeneous
  - Complex
  - Distributed



- A key component of the emerging field of data science and data-driven discovery

Key Areas of Data Mining

1. Predictive Modeling / Supervised Learning



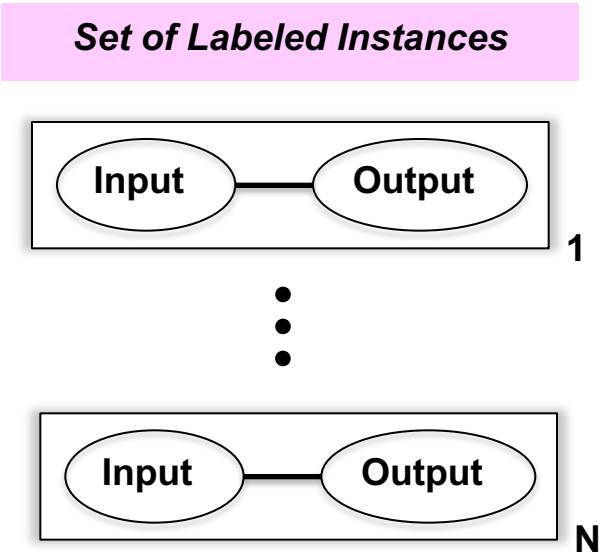
Basic Goal:

- **Model relationship between input and output variables to predict the output on unseen (new) instances**

Key Areas of Data Mining

1. Predictive Modeling

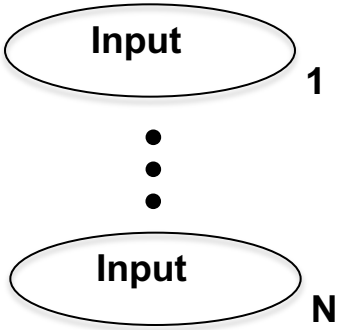
- **Classification**
  - **Target takes discrete values: {0,1,2,…}**
- **Regression**
  - **Target takes continuous values**



Key Areas of Data Mining

1. Predictive Modeling

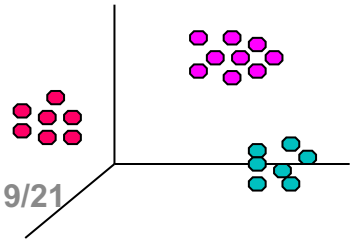
- Classification
- Regression



2. Descriptive Modeling / Unsupervised Learning

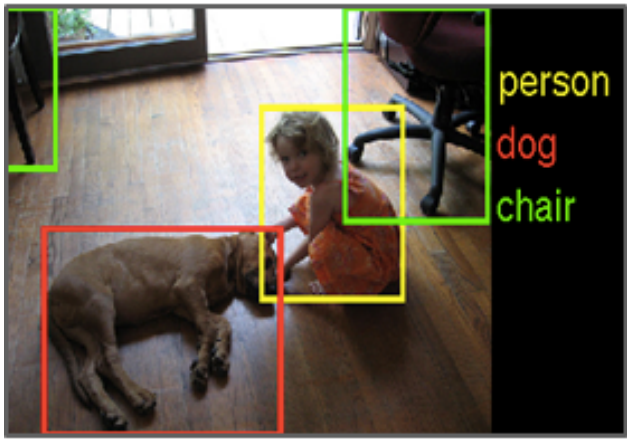
– Find human-interpretable patterns from "unlabeled" data

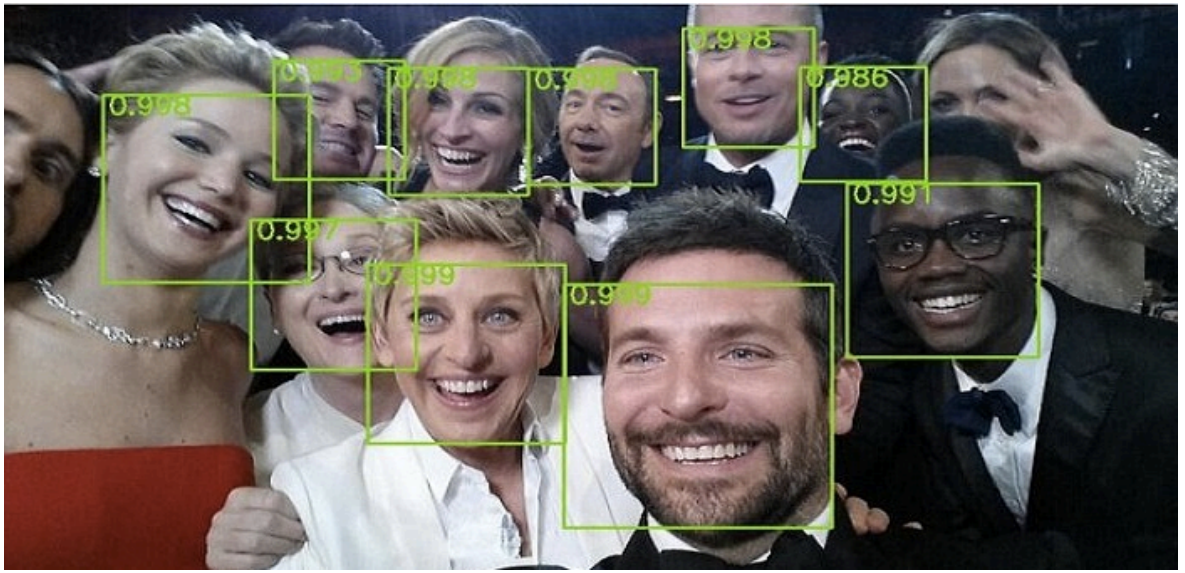
- Clustering
  - Find groups with similar properties
- Association Analysis
  - Find frequent associations
- Anomaly Detection
  - Find unusual instances



Classification: Illustrative Examples

- Image Recognition
  - Given the pixel values of an image region *(features)*, identify the type of object *(class)*



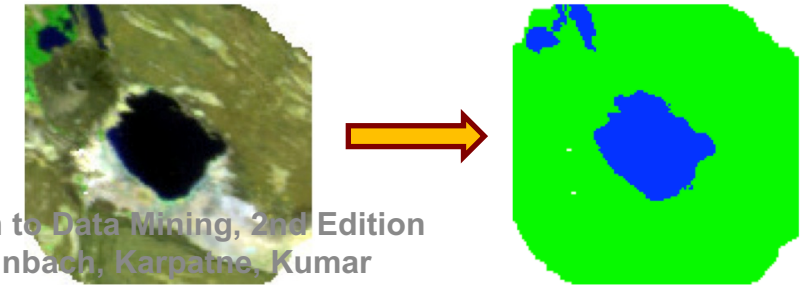


Classification: Illustrative Examples

- Image Recognition
  - Given the pixel values of an image region *(features)*, identify the type of object *(class)*
- Spam Filtering
  - Given the message header and content of an email *(features)*, classify spam or no spam *(class)*

Classification: Illustrative Examples

- Image Recognition
  - Given the pixel values of an image region *(features)*, identify the type of object *(class)*
- Spam Filtering
  - Given the message header and content of an email *(features)*, classify spam or no spam *(class)*
- Land Cover Mapping
  - Given the multi-spectral values *(features)*, classify land cover: water, vegetation, urban, etc. *(class)*



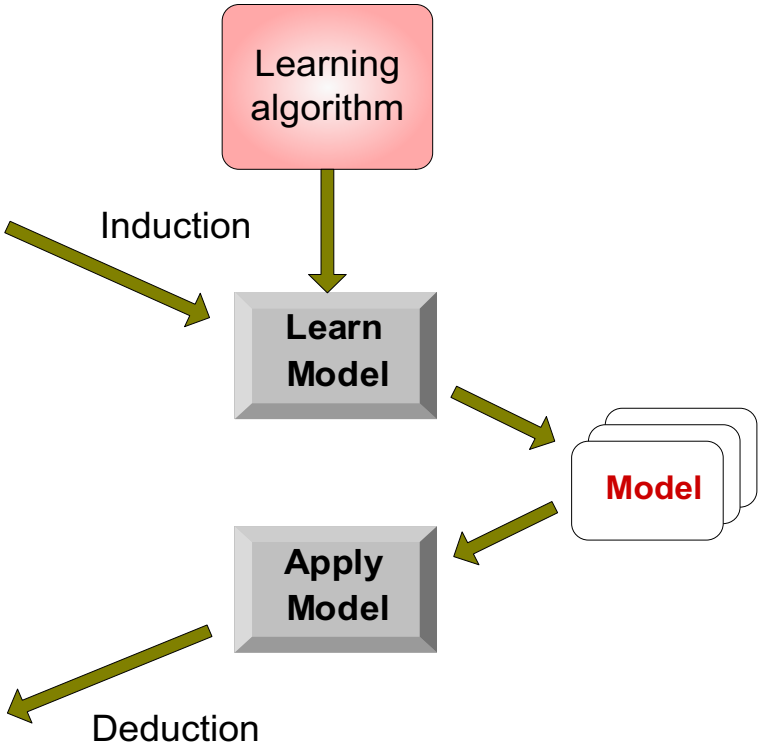
Predictive Modeling: General Approach

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

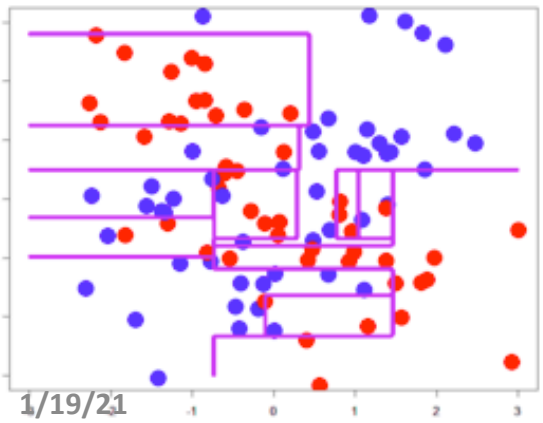


Classification Models

- Decision Trees
- Support Vector Machines (SVM)
- Nearest-neighbor Classifier
- Naïve Bayes and Probabilistic Graphical Models
- Artificial Neural Networks

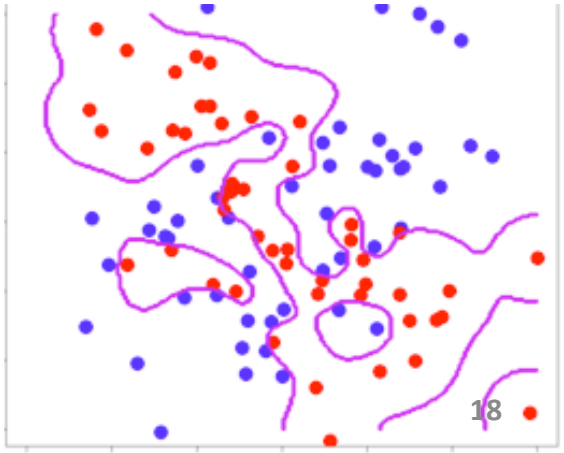
Models with varying *complexity*:
Capacity to represent complex boundaries

Decision Tree



SVM (less complex)

SVM (more complex)

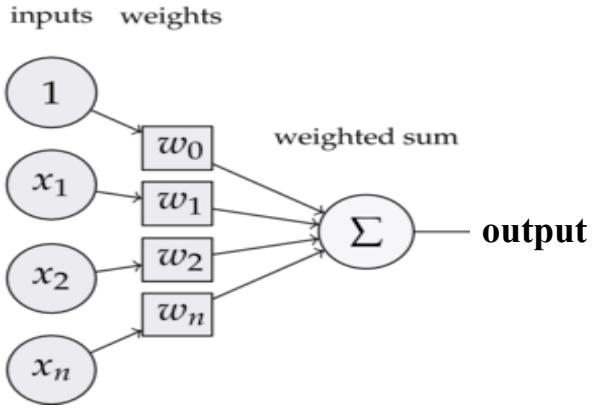


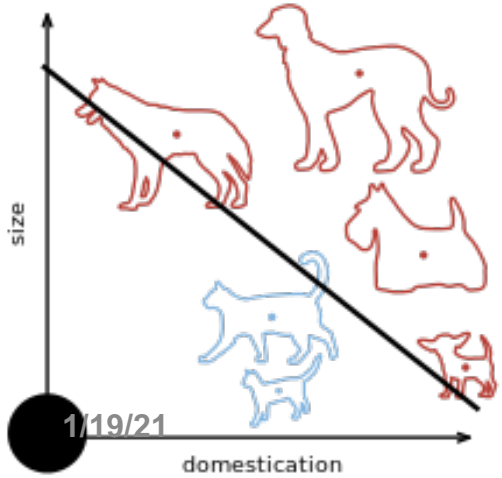
Example of Classification Model: Deep Learning

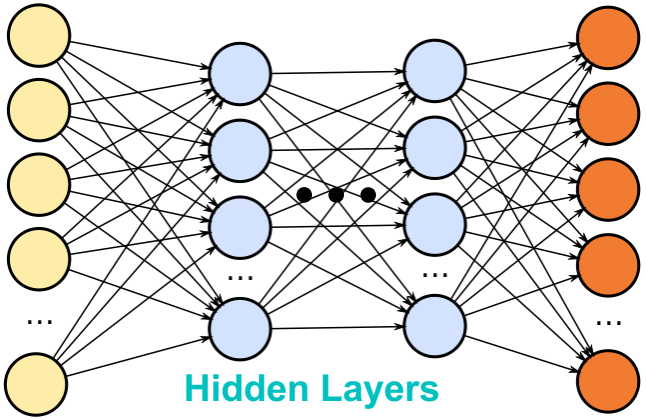
Perceptron (1970s) → **Deep Learning (~2010+)**

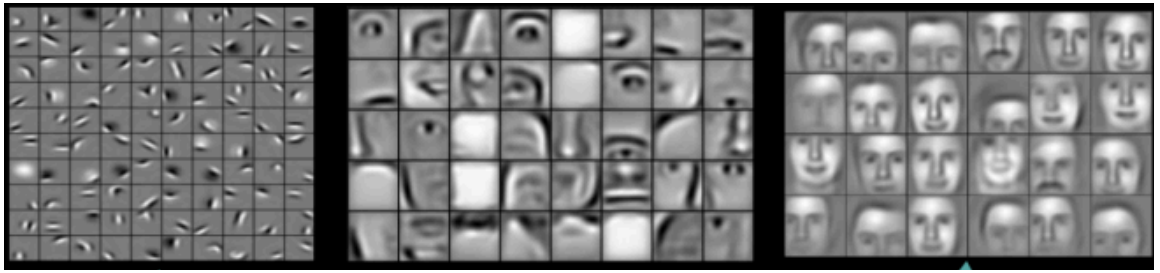
- Single processing unit
- Can only learn linear decision boundaries

- Composition of a large number of processing units
- Can learn highly complex decision boundaries



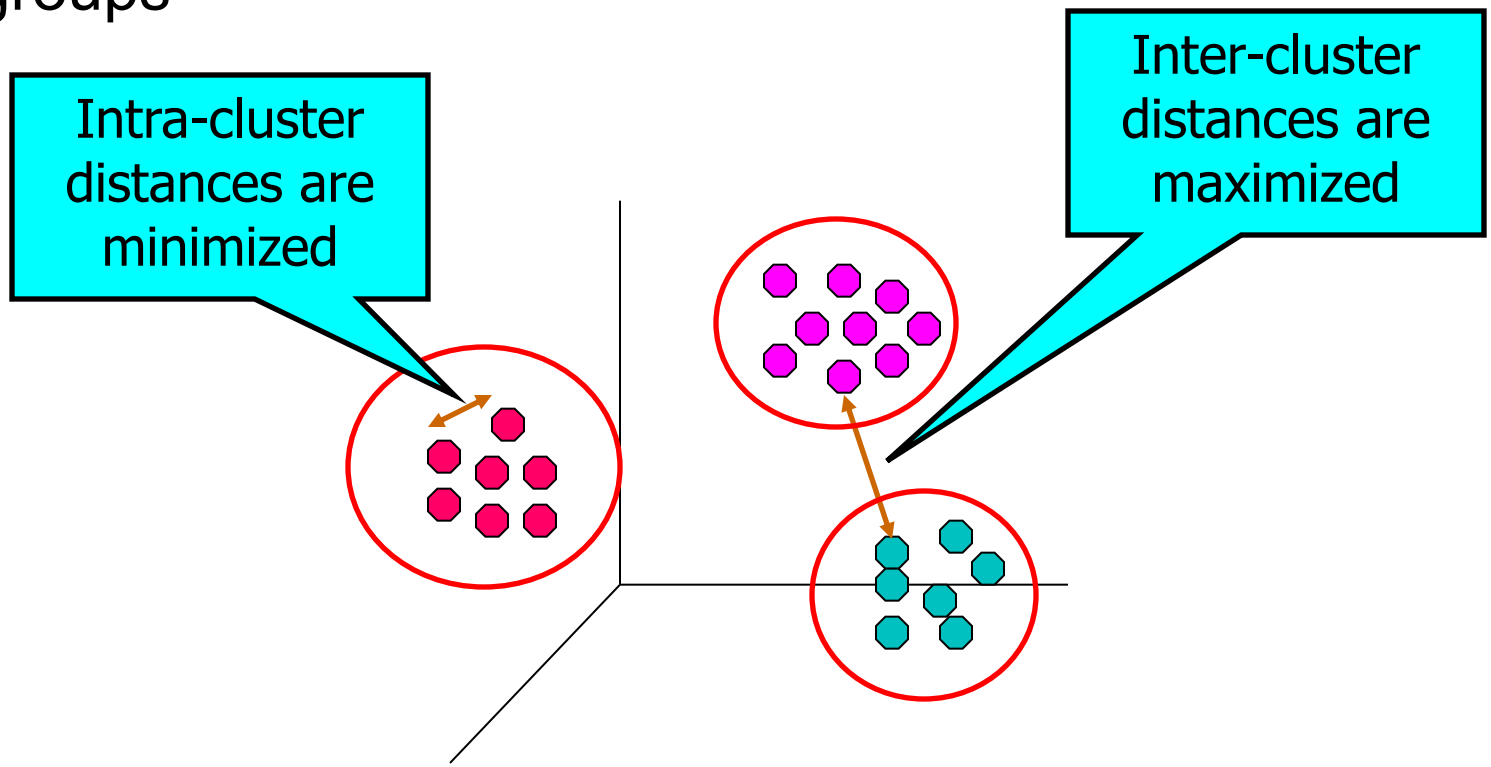






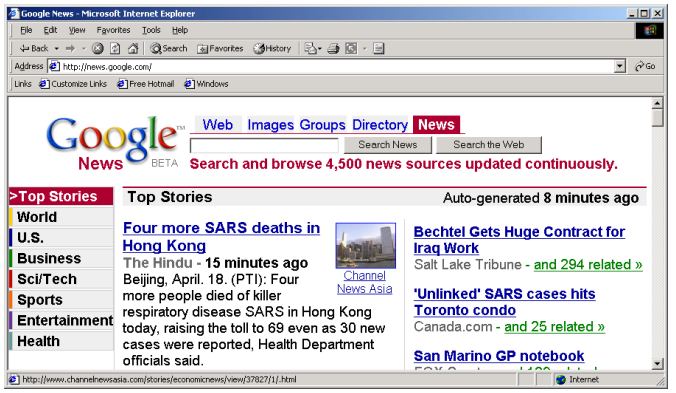
Clustering

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups

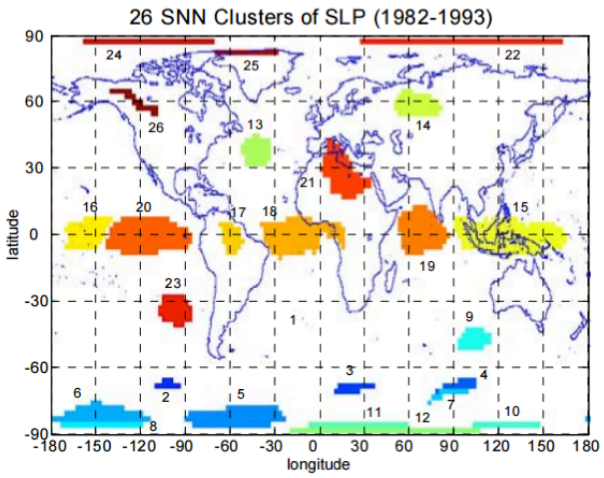


Clustering: Illustrative Examples

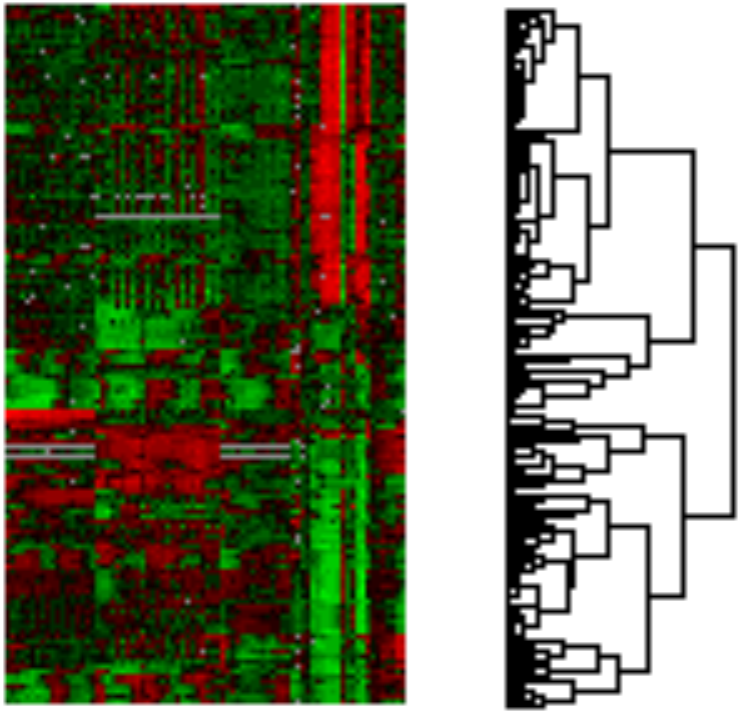
- **Understanding**
  - Group related documents for browsing
  - Group genes that have similar functionality
  - Group regions with similar climate activity



- **Summarization**
  - Reduce the size of large data sets



Clusters found using Sea Level Pressure Data



Courtesy: Michael Eisen

Association Analysis

- Given a set of records each of which contain some number of items from a given collection
  - Find patterns of co-occurrence of items

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

Rules Discovered:

{Milk} --> {Coke}

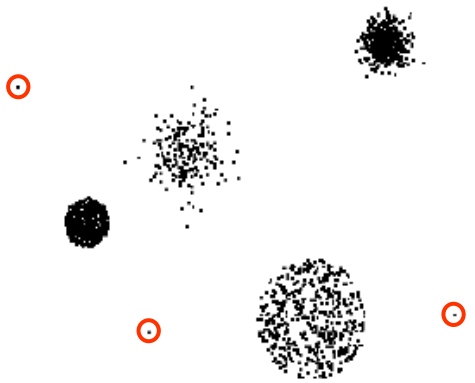
{Diaper, Milk} --> {Beer}

Association Analysis: Applications

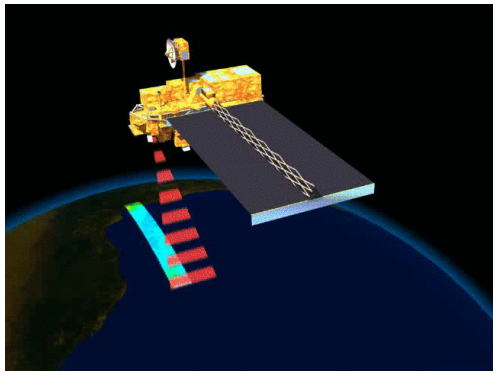
- Market-basket analysis
  - Rules are used for sales promotion, shelf management, and inventory management
- Medical Informatics
  - Rules are used to find combination of patient symptoms and test results associated with certain diseases

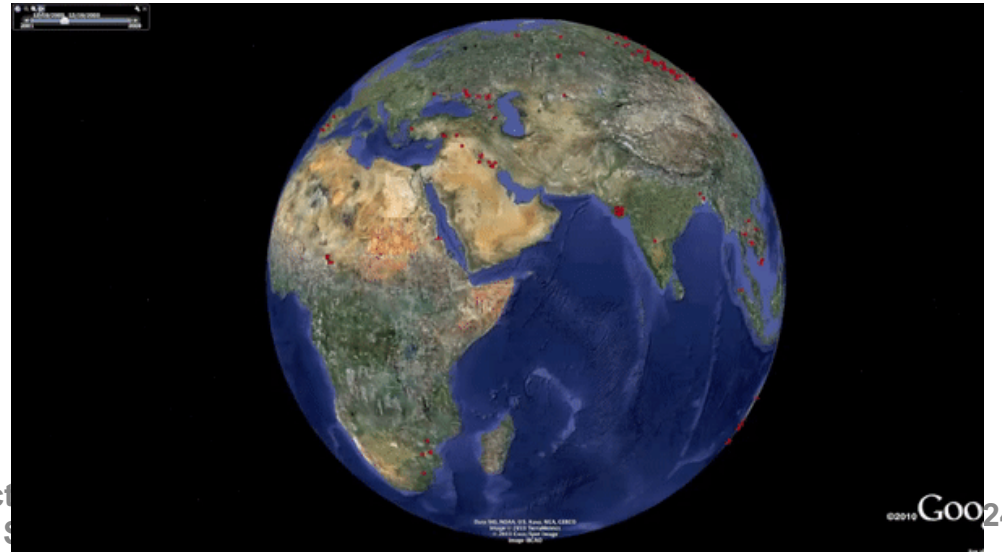
Anomaly Detection

- Detect significant deviations from normal behavior



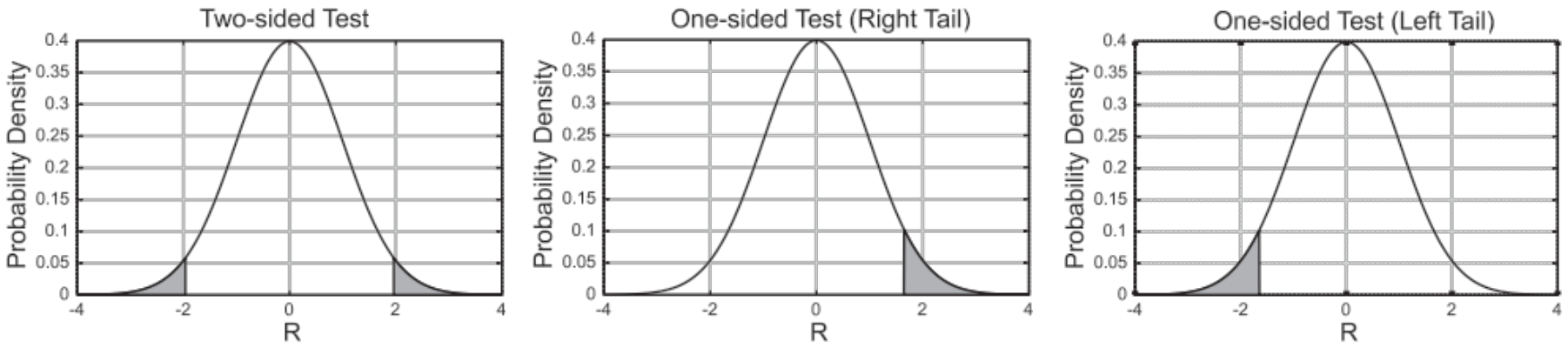
- Applications:
  - Credit Card Fraud Detection
  - Network Intrusion Detection
  - Detecting changes in the Global Forest Cover



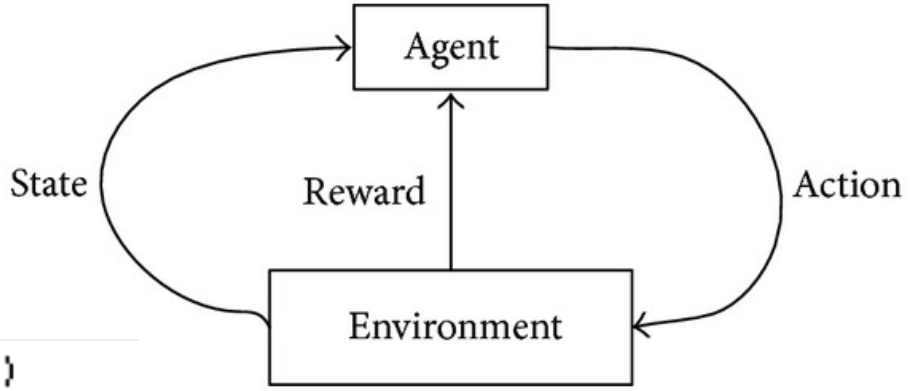


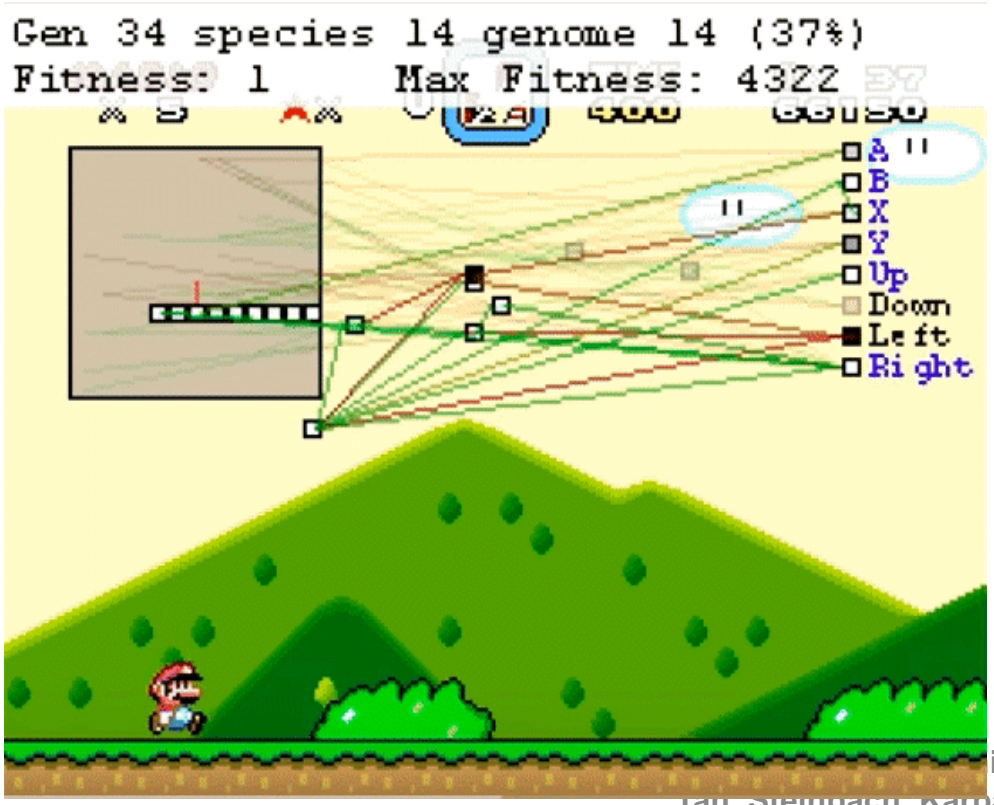
Avoiding False Discoveries

- Goal: To assess the statistical significance of a data mining result beyond random chance
  - Avoid discovery of *spurious* patterns and models
  - Especially important when testing multiple hypotheses
- Cross-cutting theme across all areas of data mining:
  - prediction, clustering, association analysis, anomaly detection



Additional Topics: Reinforcement Learning



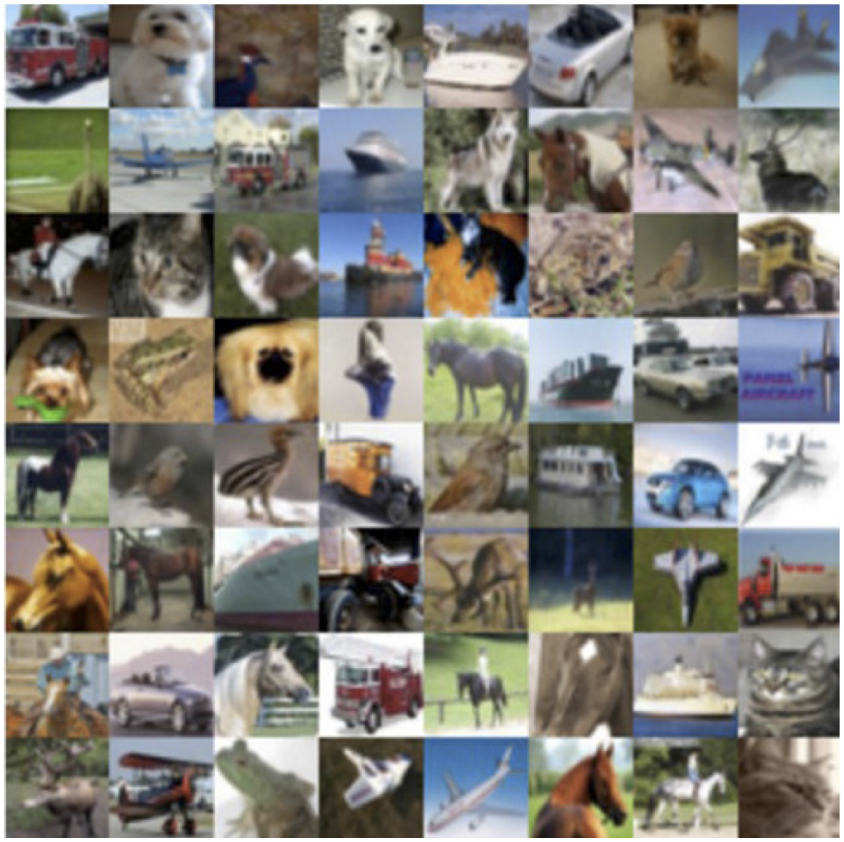


Google AI algorithm masters ancient game of Go

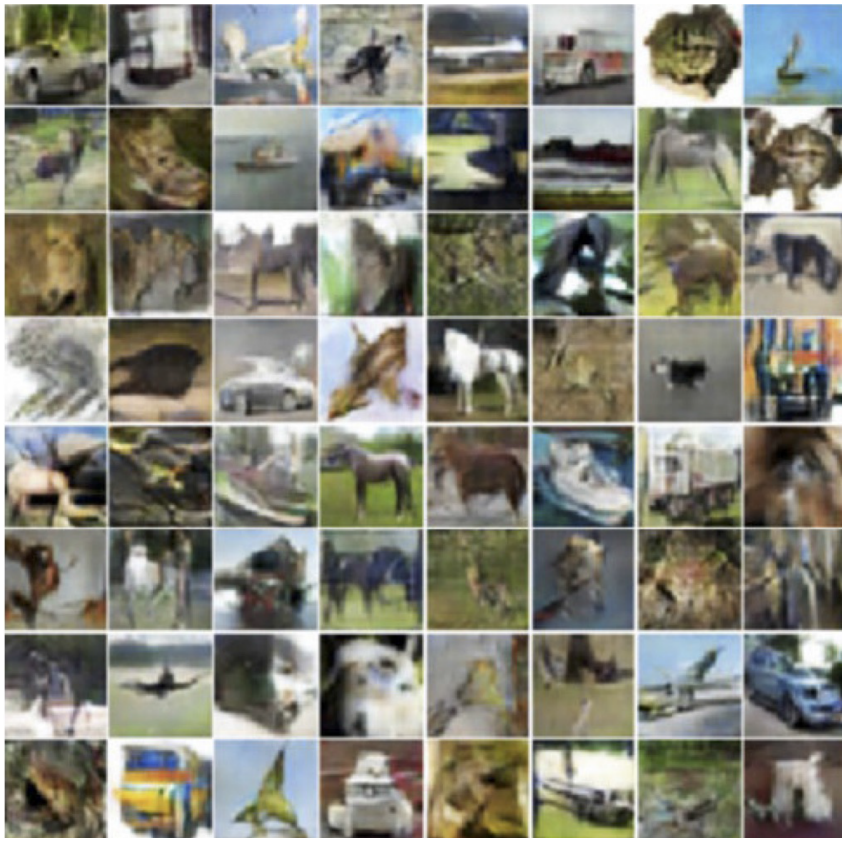
MarI/O: http://pastebin.com/ZZmSNaHX

Additional Topics: Generative Modeling

Generative Adversarial Networks (GANs)[1]:



Real-world images used for training



Synthetic images generated by GAN

[1]Goodfellow et al., 2014

Motivating Challenges

- Scalability
- High Dimensional, Heterogeneous, and Complex Data
- Paucity of Labeled Data
- Privacy and Security
- Interpretability / Scientific Consistency

What is Coming Up Next?

- HW1 (Posted: Jan 21, Due: Feb 4)
- Next Class: Understanding Data (Ch2)

Background Survey (Assignment 0)

- https://virginiatech.qualtrics.com/jfe/form/SV_6stCsvEwU52oHvo

(for students requesting force-add to the course, please use the passcode mentioned in the class)