UMD

UNIVERSITY OF MINNESOTA DULUTH

Driven to Discover℠

CS 4232/5232 – Data Mining and Machine Learning

Fall 2023

Instructor

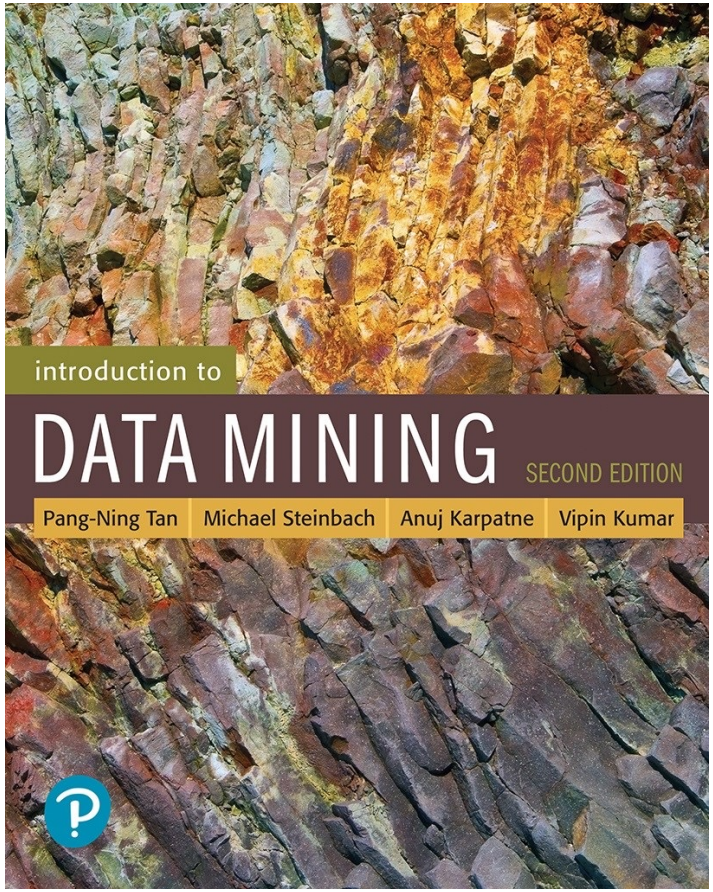
- Eleazar Leal
  - Heller Hall 331
  - Email: eleal@d.umn.edu
  - Office hours:
    - Tuesday, Thursday:
      - 9:15 am to 9:45 am,
      - 3:15 pm to 4:15 pm
    - and by email appointment
  - Research Interests: Spatial databases, Parallel Algorithms, Data Mining

Teaching Assistant

- Nam Phung
  - MWAH 187
  - Email: phung025@d.umn.edu
  - Office Hours:
    - Monday 11 am–11:50 am, 3 pm–3:50 pm
    - Tuesday 2 pm–2:50 pm
    - Wednesday 3 pm–4:50 pm
    - Friday 11 am–11:50 am

Textbooks

- Introduction to Data Mining
  - 2nd Edition
  - Authors:
    - Pang-Ning Tan
    - Michael Steinbach
    - Anuj Karpatne
    - Vipin Kumar



- Slides will be posted online, but it is expected that students will read the book

Course Outline

- Introduction to Machine Learning and Data Mining
- Exploratory Data Analysis
- Supervised Learning
  - Model Assessment and Selection
  - Regression
    - Least-squares regression
    - Regularized linear models
  - Classification
    - K-Nearest neighbors
    - Decision trees
    - Logistic regression
    - Bayesian networks
    - Neural networks and deep learning
    - Support vector machines
    - Ensemble learning: bagging, random forests, boosting
    - Among others…
- Unsupervised Learning
- Association Rule Mining

Course Outline

- This course **is not**:
  - A course on R, Caret, Python, scikit-learn, TensorFlow, Pytorch, Weka, etc.
  - A course on **how to use** machine learning / data mining packages

- In this course, **you will learn the inner working of many machine learning algorithms from a Computer Science perspective**. This means that there will be math.

About the Math

- There will be math in this course:
  - Derivatives
  - Linear Algebra
    - Matrices
    - Vectors
  - Probabilities and statistics

About the Math (Cont'd)

- So we need to compute the gradient of *E(w)* with respect to *w*: $\nabla E(w) = \left(\frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}, \ldots, \frac{\partial E}{\partial w_n} \right)$

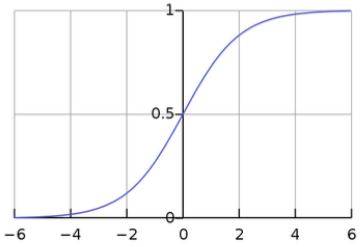
$$\begin{aligned}
\frac{\partial E}{\partial w_i} &= \frac{\partial}{\partial w_i} \frac{1}{2} \sum_{d \in D} (t_d - o_d)^2 \\
&= \frac{1}{2} \sum_{d \in D} \frac{\partial}{\partial w_i} (t_d - o_d)^2 \\
&= \frac{1}{2} \sum_{d \in D} 2(t_d - o_d) \frac{\partial}{\partial w_i} (t_d - o_d) \\
&= \sum_{d \in D} (t_d - o_d) \frac{\partial}{\partial w_i} (t_d - w^T x_d) \\
&= \sum_{d \in D} (t_d - o_d)(-x_{id})
\end{aligned}$$

$$o(x_1, \ldots, x_n) = \begin{cases} 1 \text{ if } w_0 + w_1 x_1 + \cdots + w_n x_n > 0 \\ -1 \text{ otherwise.} \end{cases}$$

Where x_d is the d-th training example:

$$x_d = (1, x_{1d}, x_{2d}, \ldots, x_{nd})^T$$

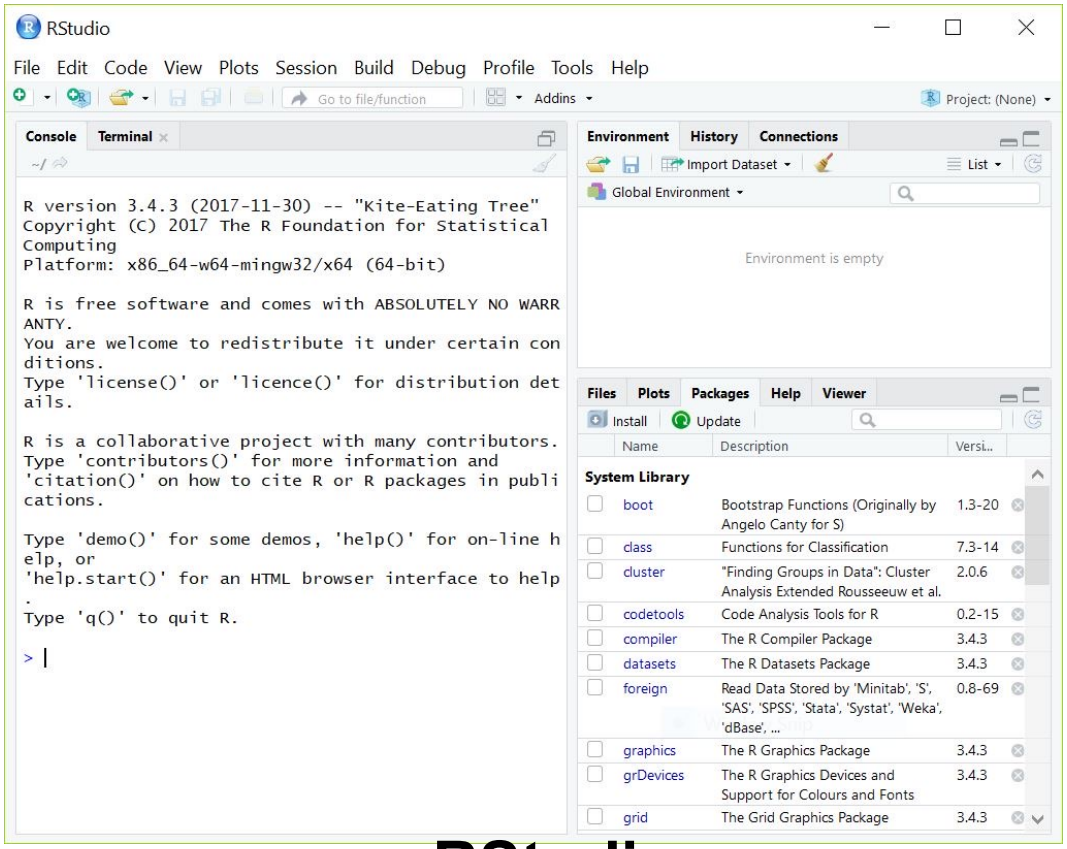
$$f(x) = \frac{e^x}{1 + e^x} = \frac{1}{1 + e^{-x}}$$



$$\begin{aligned}
\frac{d\sigma(x)}{dx} &= \frac{d}{dx} \left(\frac{e^x}{1 + e^x} \right) \\
&= \frac{e^x (1 + e^x) - (e^x)^2}{(1 + e^x)^2} \\
&= \frac{e^x}{(1 + e^x)^2} \\
&= \sigma(x) * (1 - \sigma(x))
\end{aligned}$$

Programming Language

- The official programming language in the course is R



RStudio

Course Rubric

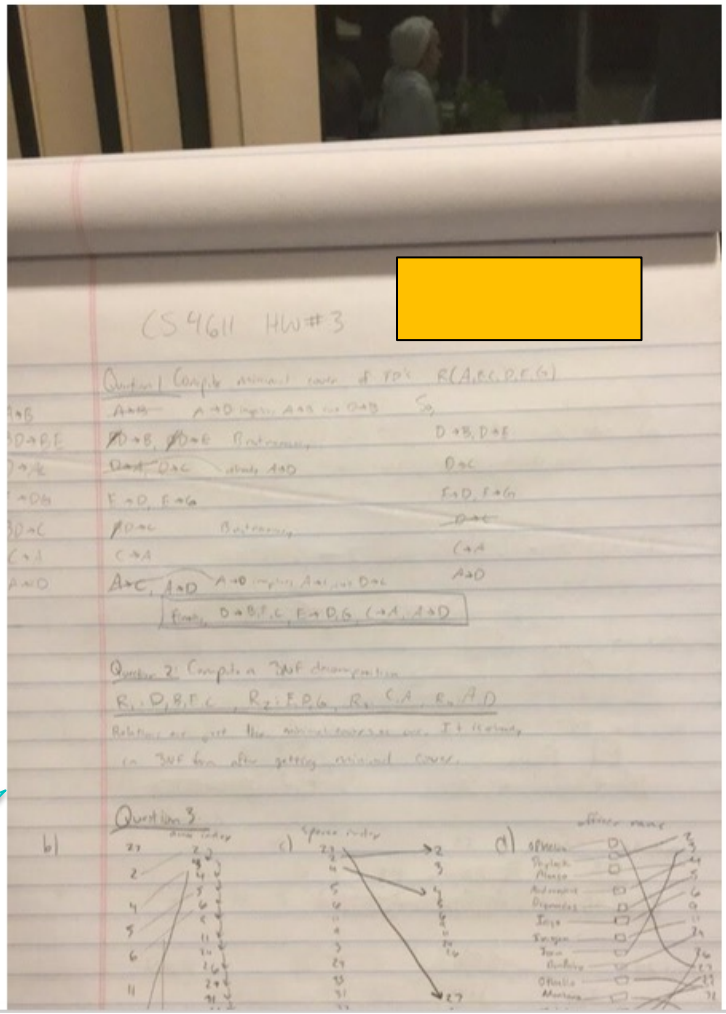
- Final 30%
- Homework Assignments 15%
- Project 40%
- Midterm 15%

Homework

- Homework assignments are to be done individually. Homework will be pencil-and-paper and/or programming assignments.
- Students will work on the homework assignments in the lab and at home.
- Submit homework to Canvas

Homework (Cont'd)

- Make your canvas submission readable!!
- High contrast for the text, make it bright
- Make it look nice

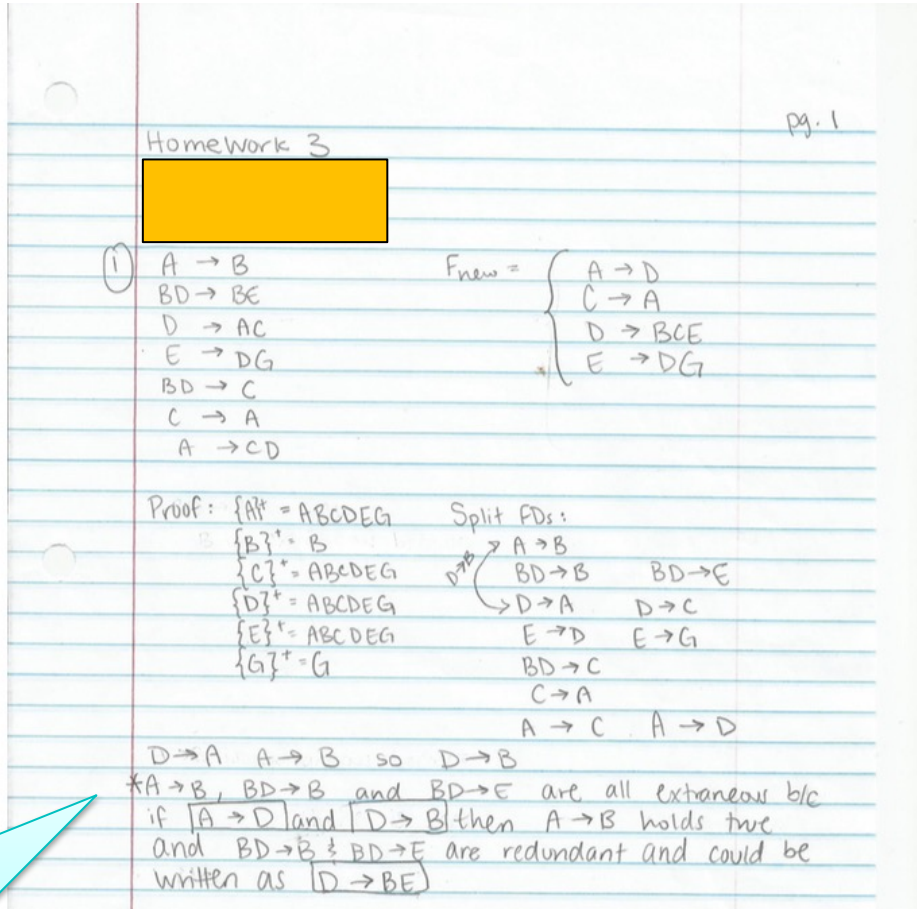


Example of a bad submission

How many things are wrong in this submission?

Homework (Cont'd)

- Make your canvas submission readable!!
- High contrast for the text.
- Make it look nice



Example of a good submission

Project

- More information about the project(s) coming in the future
- The project is a group assignment
- This is a semester-long assignment
- The project will involve solving a machine learning/data mining problem
- There will be a project proposal, progress reports, and a final presentation
- You are free to choose any programming language

Late Policy

- Two 24-hour free extensions with no penalty during the whole semester

- Otherwise, submitting an assignment late, but not later than 24 hours late, will carry a penalty of 15% of the grade.

Midterm and Final

- Exams are closed everything: books, phones, etc.
- The final exam is *comprehensive*: it covers everything seen in the course
- It is expected that students will memorize all formulas that we will see in class
- The book and the slides provided are the study guide
- Exams will focus on the contents seen in class and in the homework/labs and quizzes

For CS 5232 Only

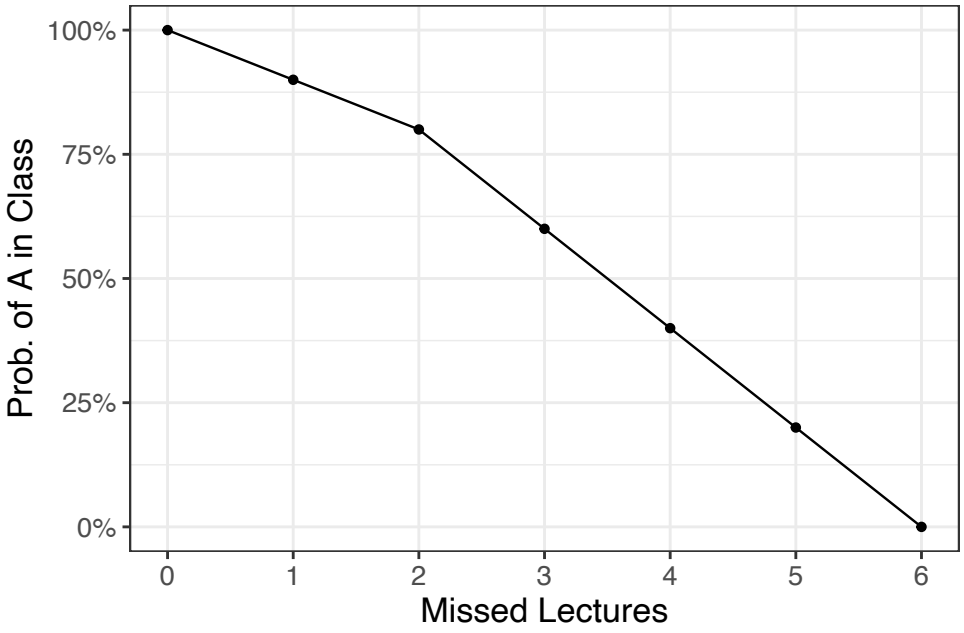
- Homework assignments, the project, the midterm and the final exam for CS 5232 students might involve additional and/or more complex activities than those required for CS 4232 students.

Class Participation

- If there is something you don't understand in class, ask right away!
- If the class is moving too fast, ask questions!
- I expect you to participate and ask questions in class (or in office hours)

Importance of Attending Class

- This course is designed with the expectation that students **attend all lectures.**
- There are exercises covered in th lectures but not in the slides.
- We review material during class.
- Coming to class makes studying later easier
- The greater the number of classes you miss, the more likely it is that you won't do well in the class. This is not because each lecture builds on previous ones.



Ok, it does not work exactly like in this figure, but you get the point! The course is harder if you don't come to class

Make the effort to come to every lecture!

Academic Misconduct

- **Plagiarism will not be tolerated in this course**
- Submitting an assignment containing code that you did not create is considered academic dishonesty.
- All assignments submitted for individual grading (i.e., homework and exams) **must be the result of the work of that individual only**.
- **All** assignments submitted for group evaluation **must be the result of the work of that group exclusively**.
- If you have doubts about whether something constitutes plagiarism or not, consult me.

Academic Misconduct

- Students **may not distribute, via the Internet or other means, lecture notes or instructor-provided materials**, except to other members of the same class or with the express written consent of the instructor**.**

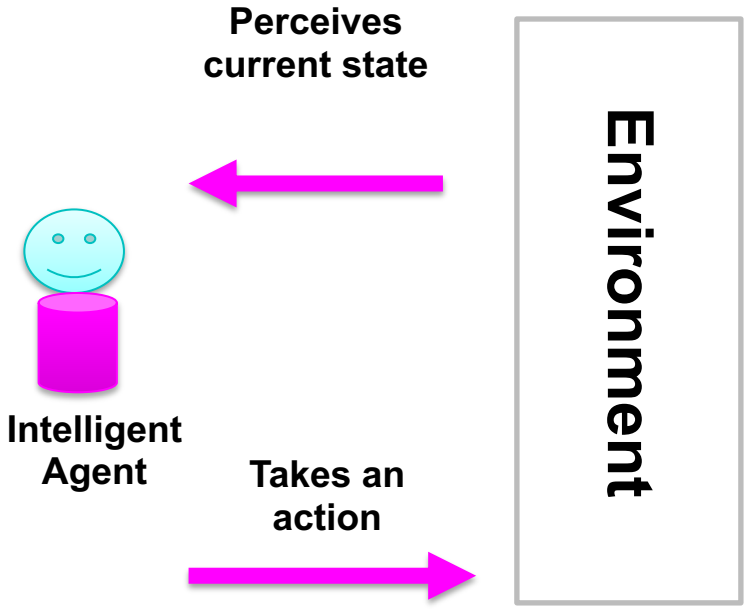
Additional Expectations

- As is the case for **any** college class:

It is expected that for each hour of lecture time, students will spend a minimum of 2 to 3 hours outside of class studying the course material on their own.

An Intelligent Agent in its Environment

- Environment
- What is an agent?
  - Perceives its environment and can act on it through actions.
- Performance measure



An Intelligent Agent in its Environment

- What is the environment?
- What are the agents?
- What are the actions of the agent?
- How do you measure the performance of an agent?

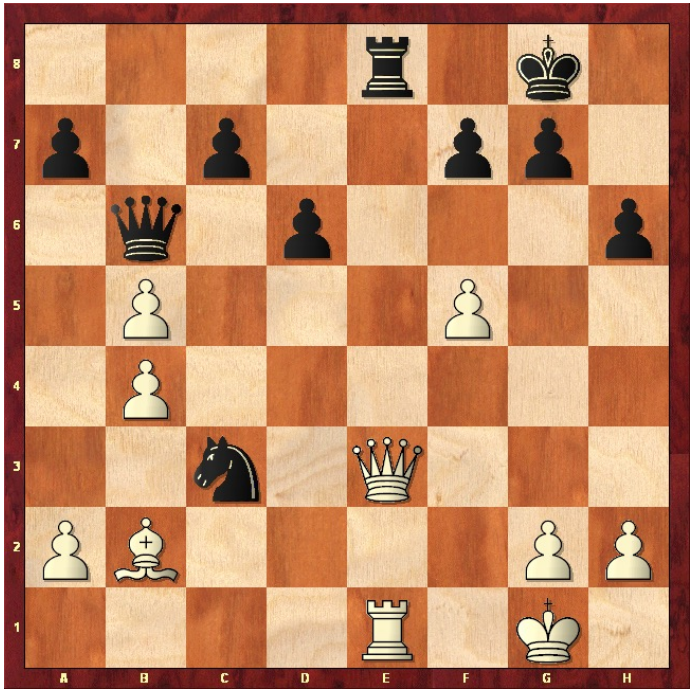


Figure from https://es.wikipedia.org/wiki/Ajedrez

What is a Machine Learning?

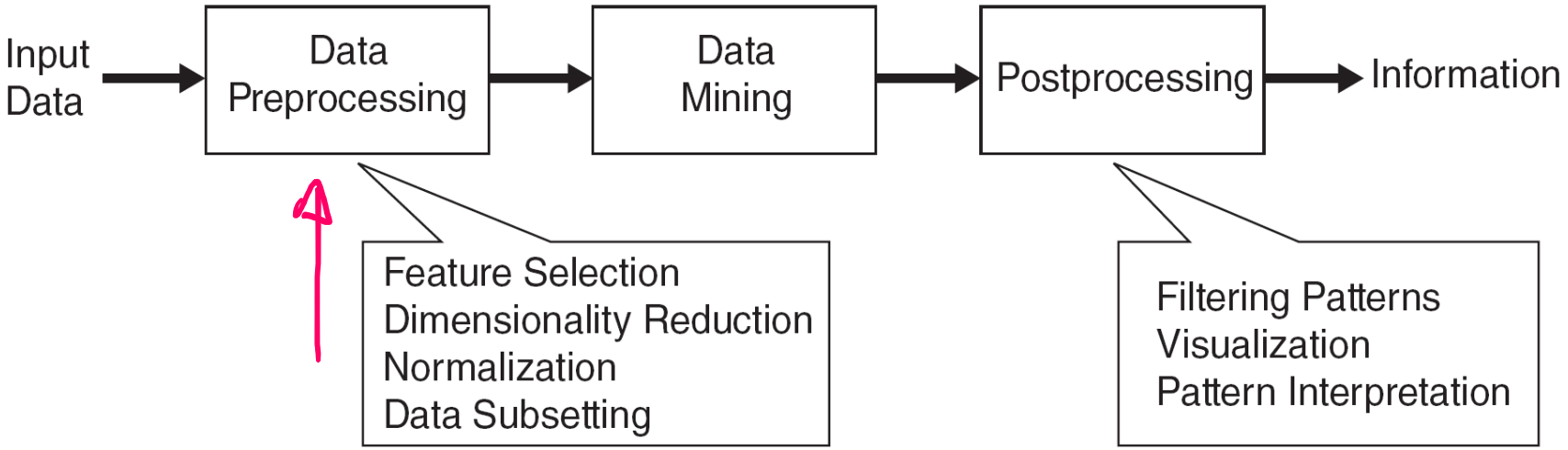
"A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P* if its performance at tasks in *T*, as measured by *P*, improves with experience *E*."

Tom Mitchell, *Machine Learning*, 1998

What is Data Mining?

- Many Definitions

  - Non-trivial extraction of implicit, previously unknown and potentially useful information from data
  - Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns



What is a Machine Learning?

"A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P* if its performance at tasks in *T*, as measured by *P*, improves with experience *E*."

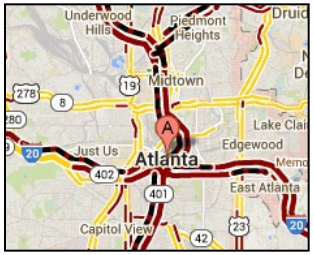
Tom Mitchell, *Machine Learning*, 1998

Large-scale Data is Everywhere!

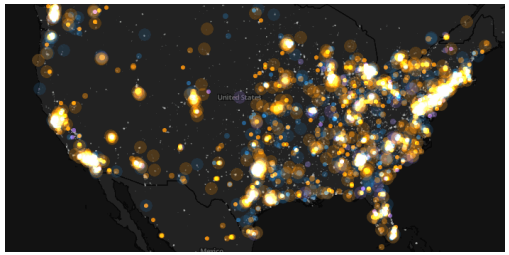
- There has been enormous data growth in both commercial and scientific databases due to advances in data generation and collection technologies
- New mantra
  - Gather whatever data you can whenever and wherever possible.
- Expectations
  - Gathered data will have value either for the purpose collected or for a purpose not envisioned.

Cyber Security

E-Commerce

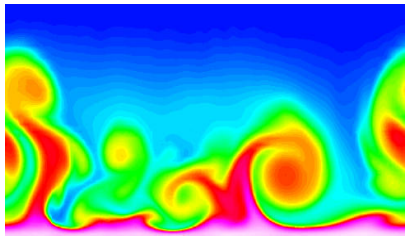


Traffic Patterns



Social Networking: Twitter

Sensor Networks

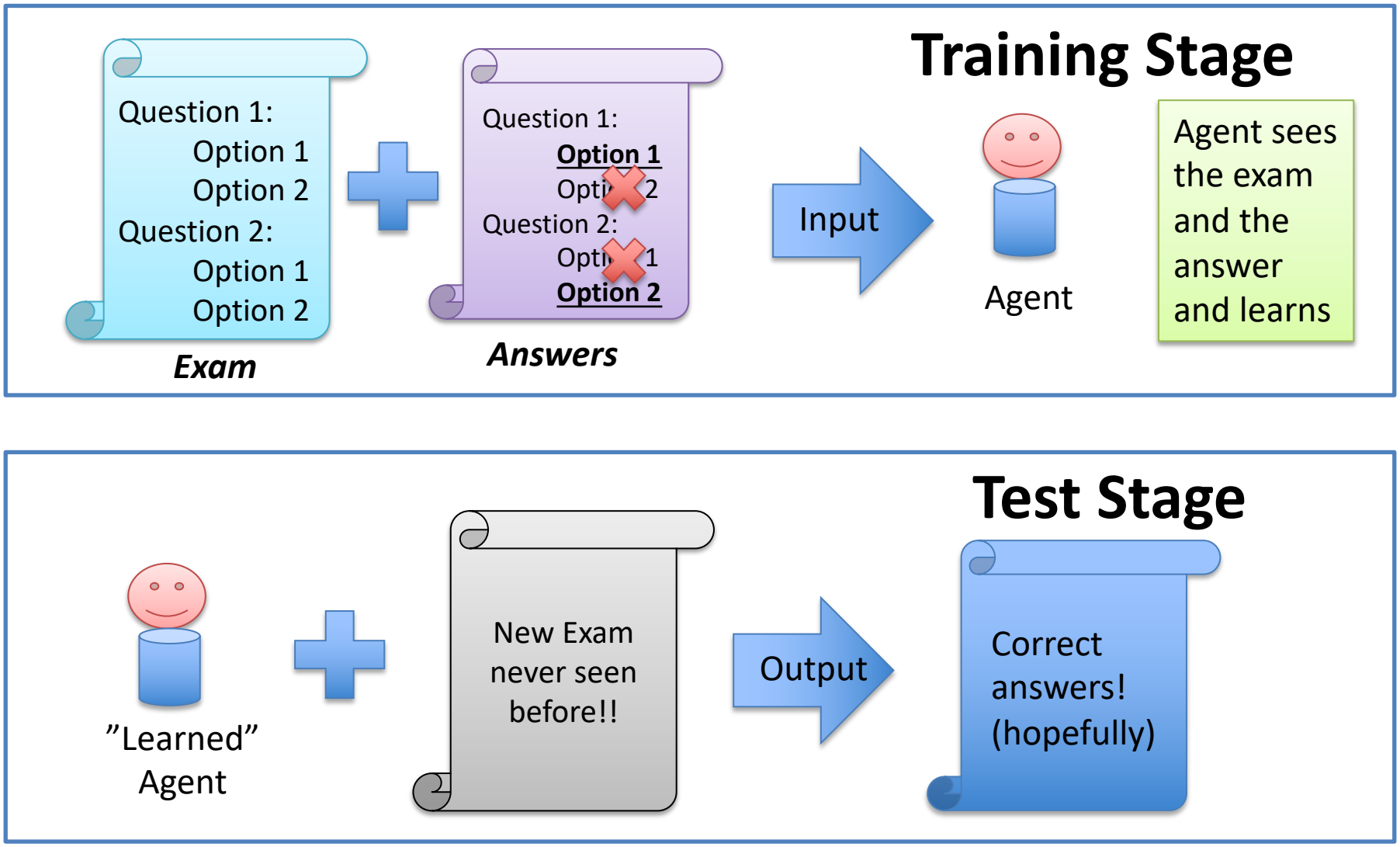


Computational Simulations

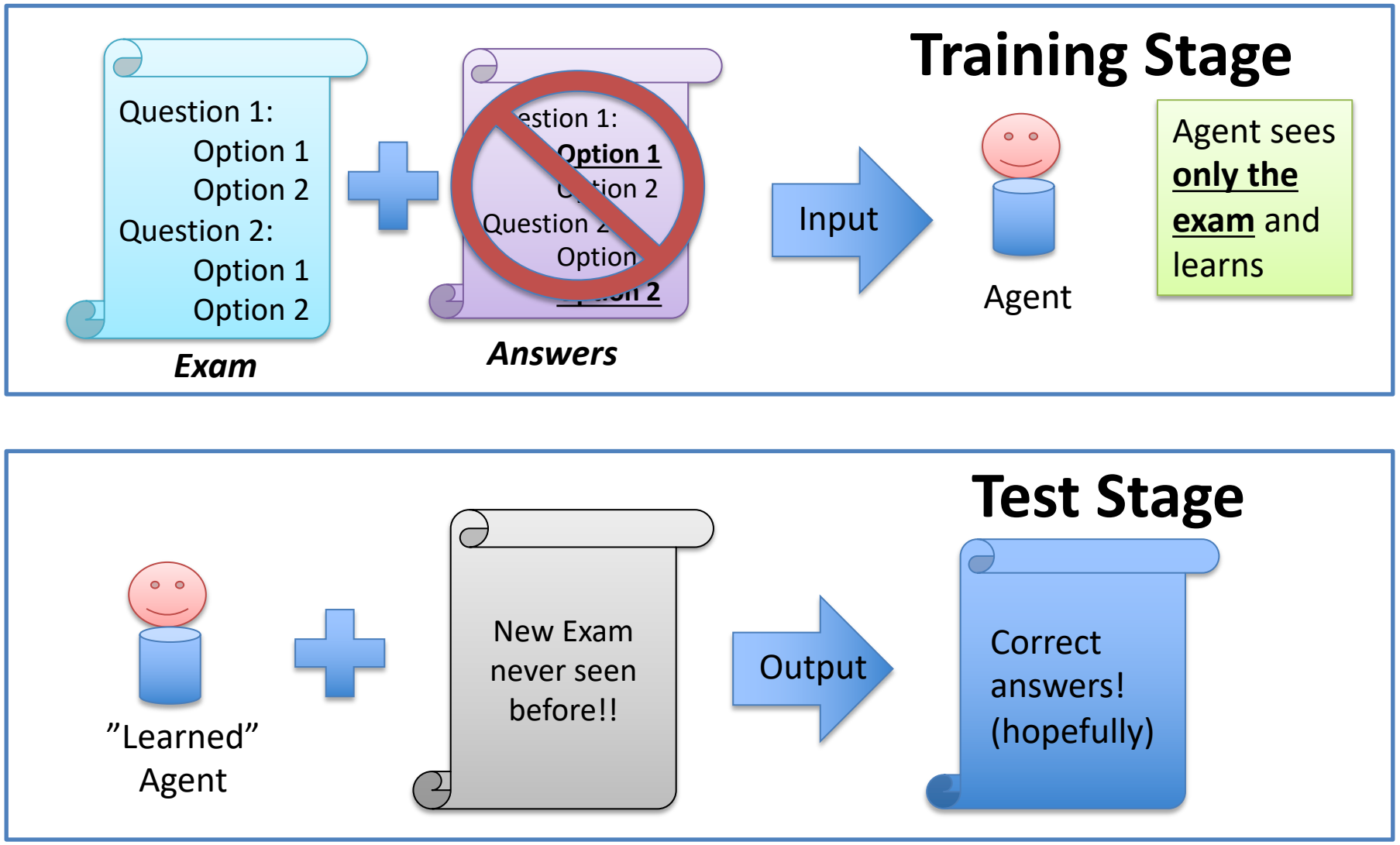
Classification of Machine Learning Tasks

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning
- Semi-supervised learning

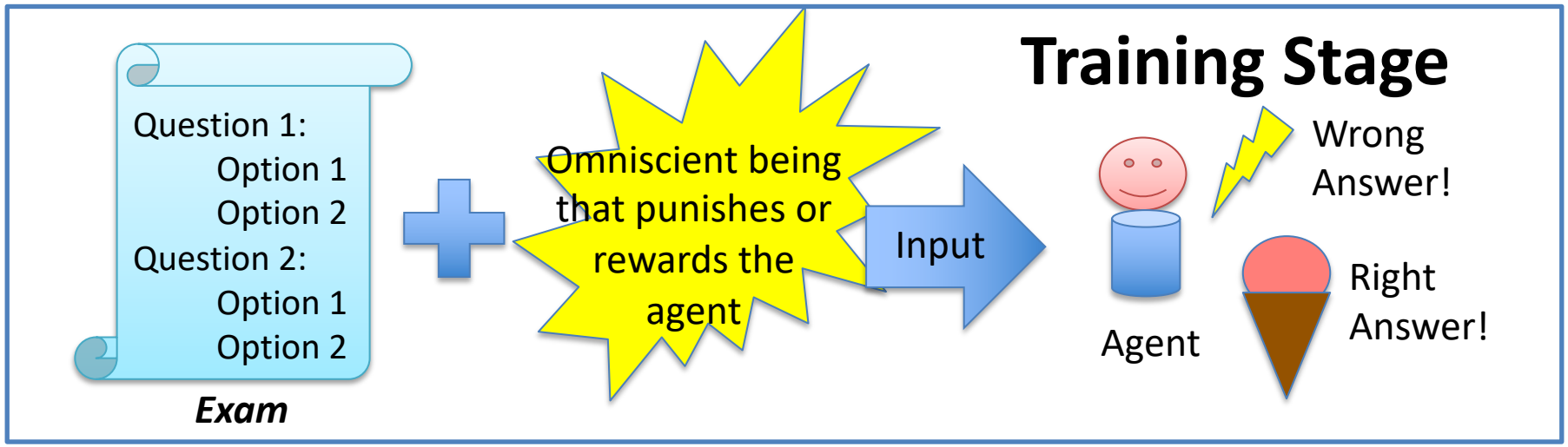
Supervised Learning

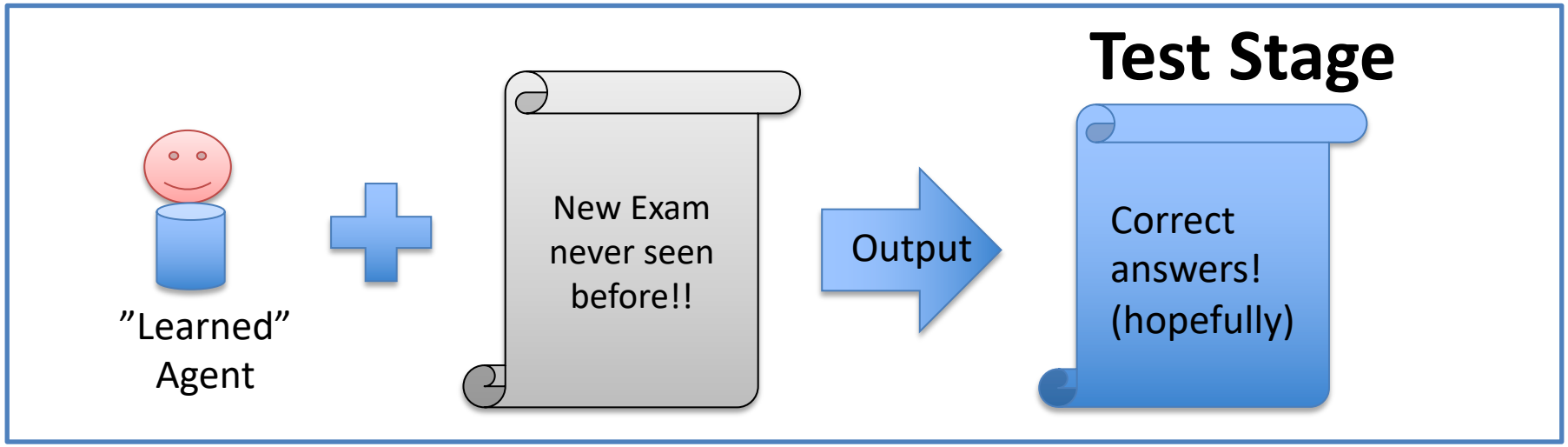


Unsupervised Learning

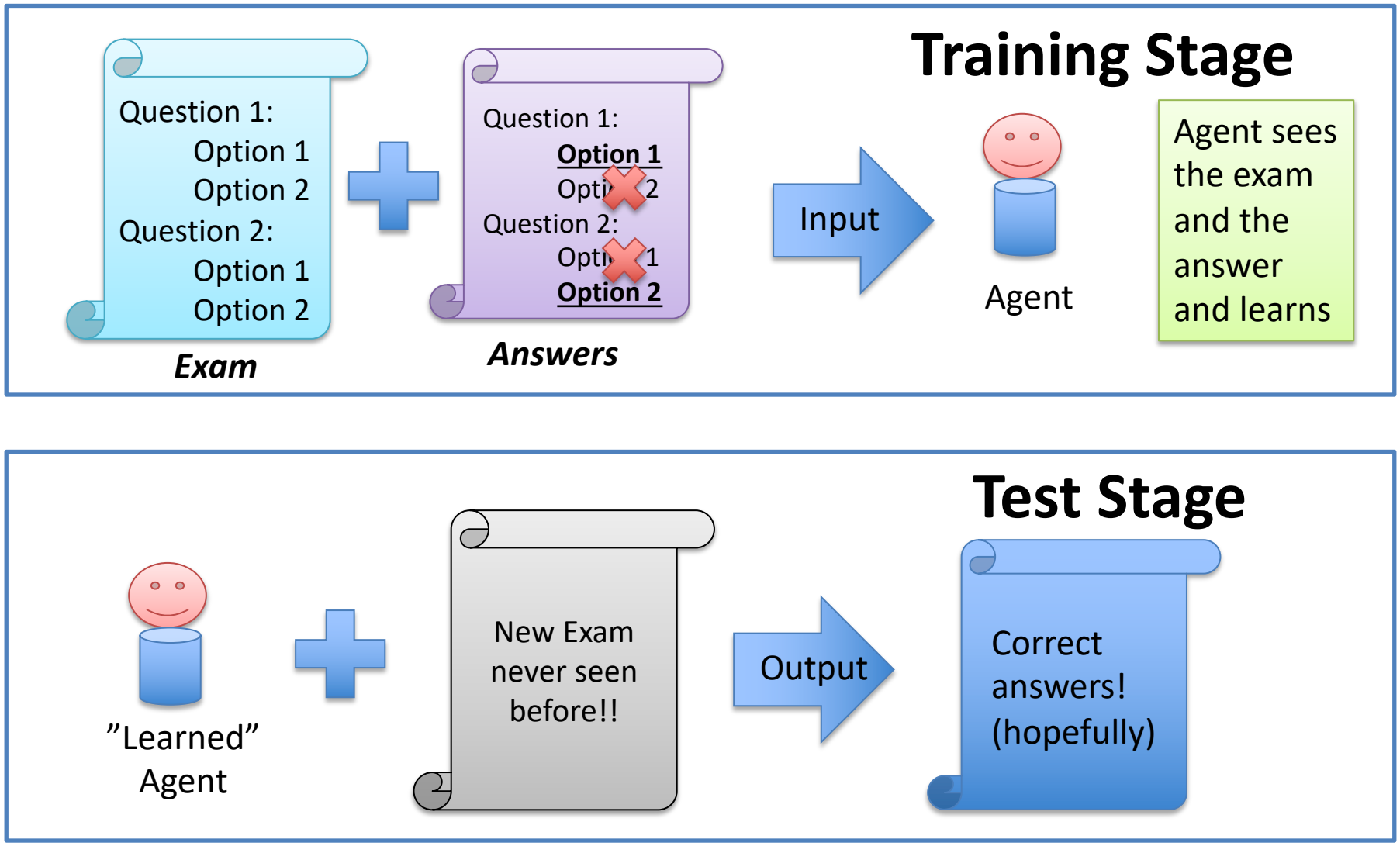


Reinforcement Learning



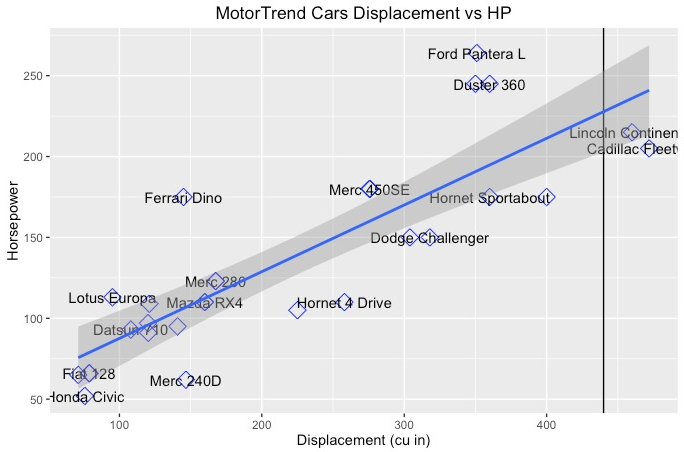


Supervised Learning

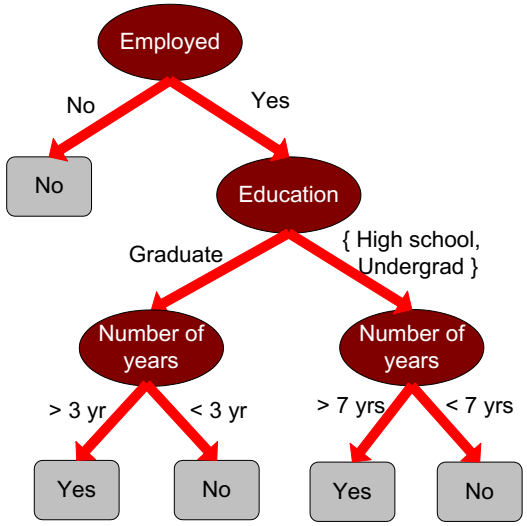


Types of Supervised Learning

- Regression
  - Linear Regression
  - Regression Trees
- Classification
  - Decision Trees
  - Logistic Regression
  - Neural Networks
  - SVMs



Linear Regression

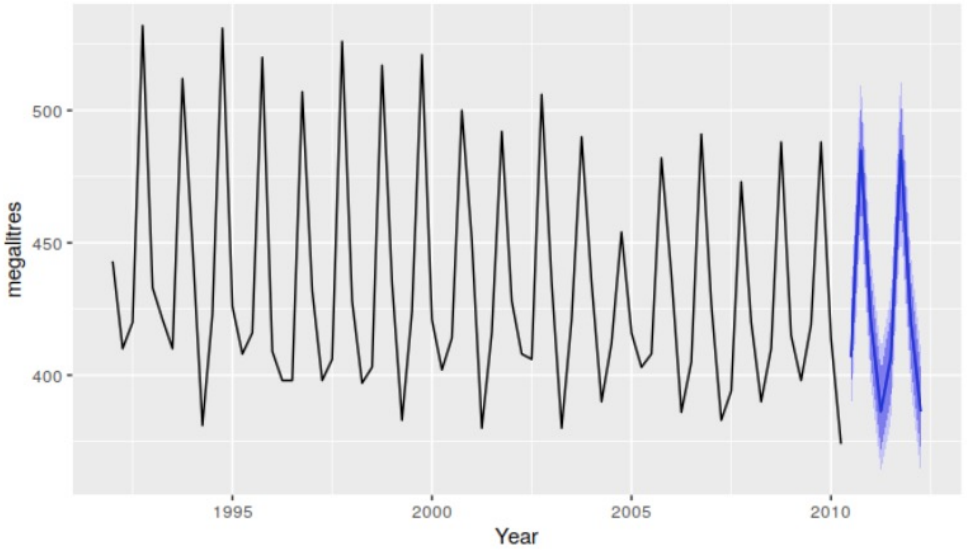


Decision Trees

Poll 2!

Predictive Modeling: Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency
- Extensively studied in statistics
- Examples:
  - Predicting sales amounts of new product based on advertising expenditure
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices



Quarterly production of beer in Australia 1992–2010 with a 2 year forecast

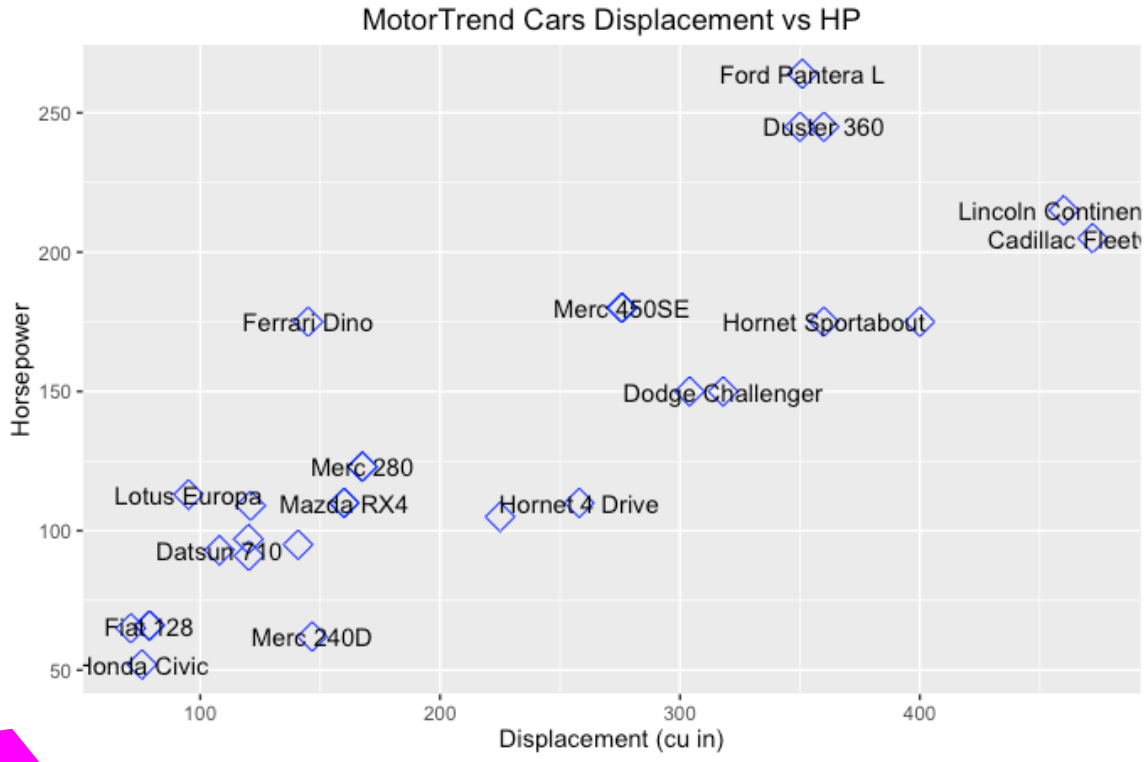
A ”hello, World!” in Machine Learning

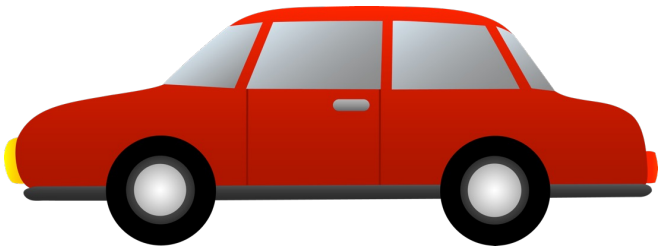
	mpg	# of Cylinders	Displacement (cu. In.)	Horsepower
Mazda Rx4	21	6	160	**110**
Cadillac Fleetwood	10.4	8	460	**205**
Dodge Challenger	15.5	8	318	**150**

Sample of the MTCars dataset

A ”hello, World!” in Machine Learning

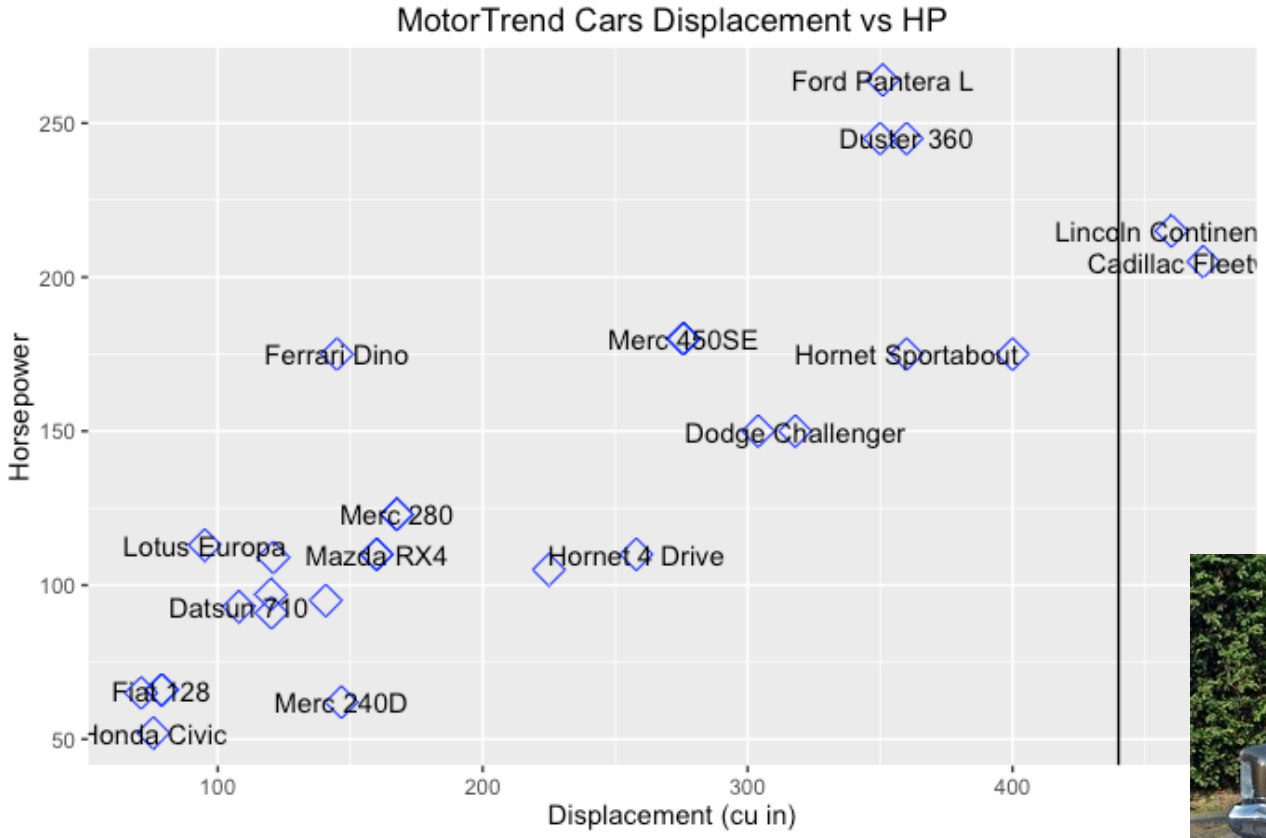
Horsepower is usually proportional to the size of the engine





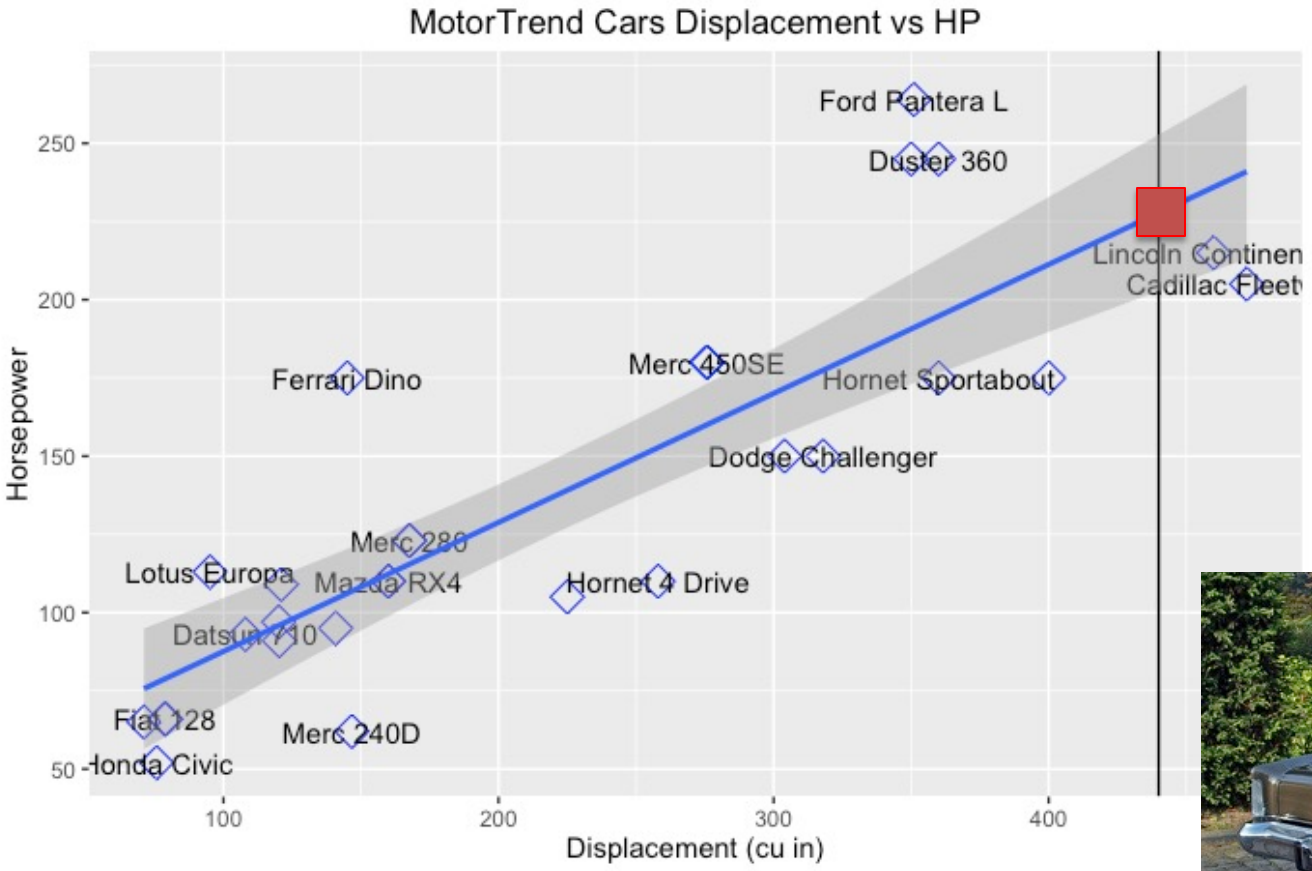
Little car credits:
https://clipartfox.com/categories/view/1043146d7fe5f5fda642973f60a4af14524ccf97/cartoon-car-free-clipart.html

A ”hello, World!” in Machine Learning



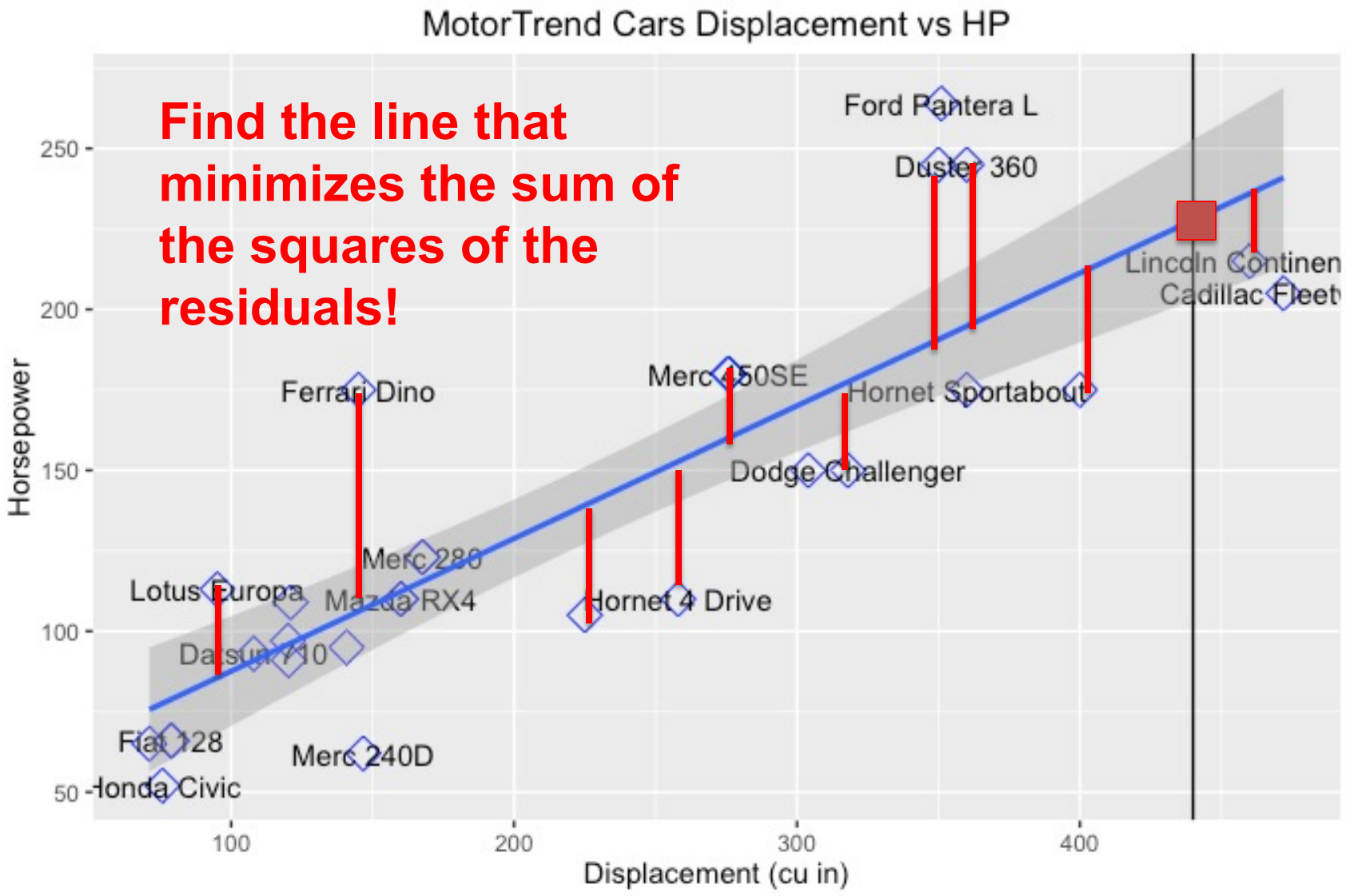
Can we predict the power of a Chrysler Imperial with an engine of 440 cu in?

A ”hello, World!” in Machine Learning

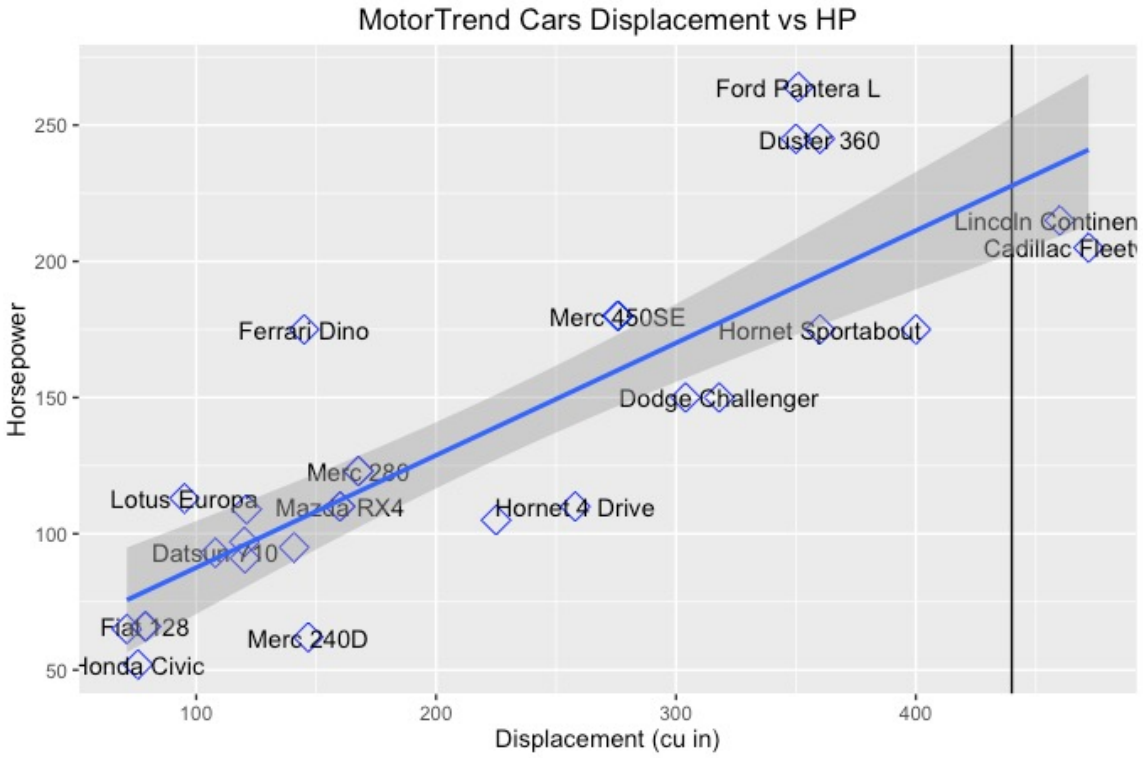


Can we predict the power of a Chrysler Imperial with an engine of 440 cu in?

A ”hello, World!” in Machine Learning



A ”hello, World!” in Machine Learning



Given a set of pairs

$$(disp_1, hp_1), \ldots, (disp_n, hp_n)$$

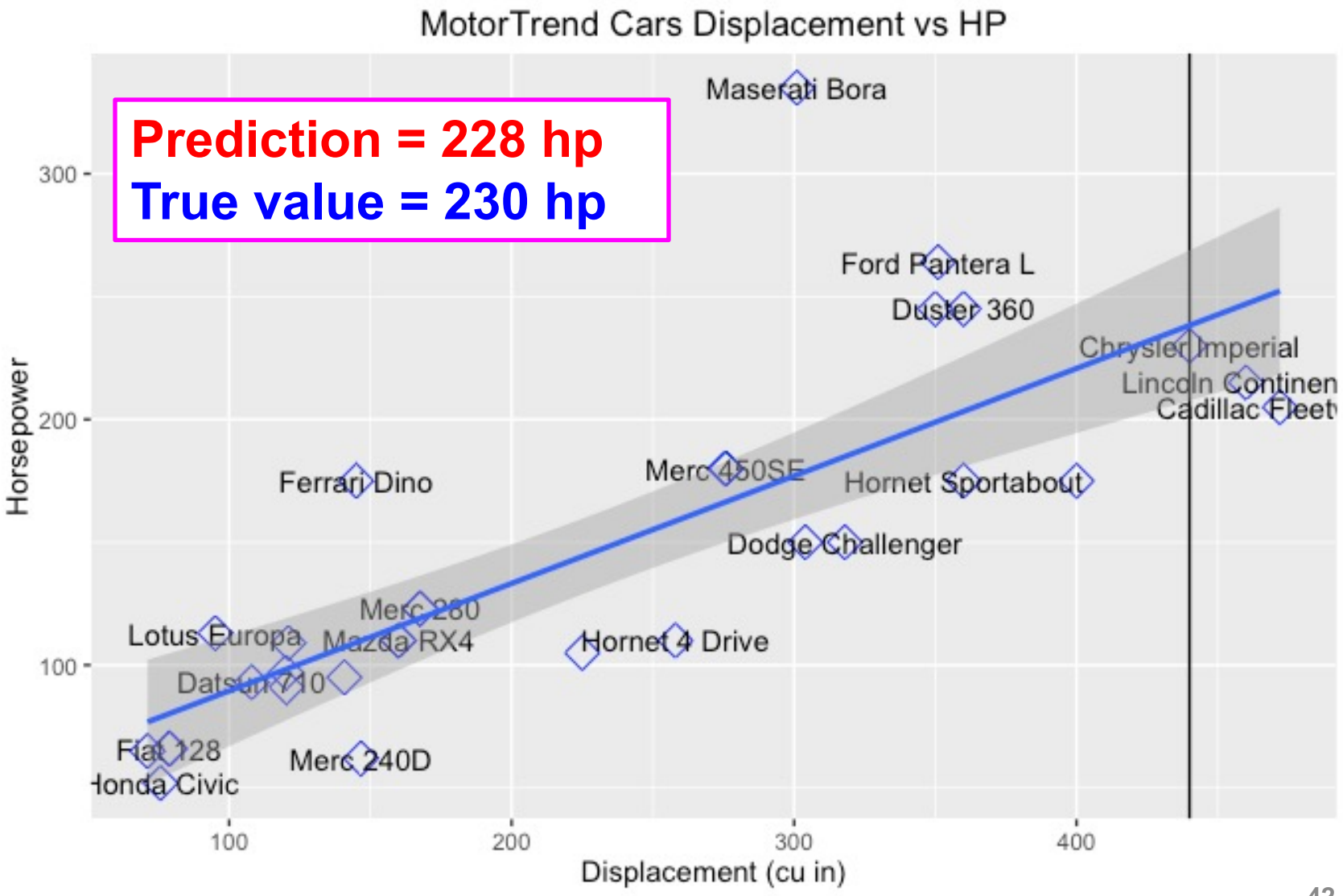
Find a line

$$\widehat{hp}_i = a \cdot disp_i + b$$

Minimize

$$\sum_{i=1}^{n} \left(hp_i - \widehat{hp}_i \right)^2$$

A ”hello, World!” in Machine Learning



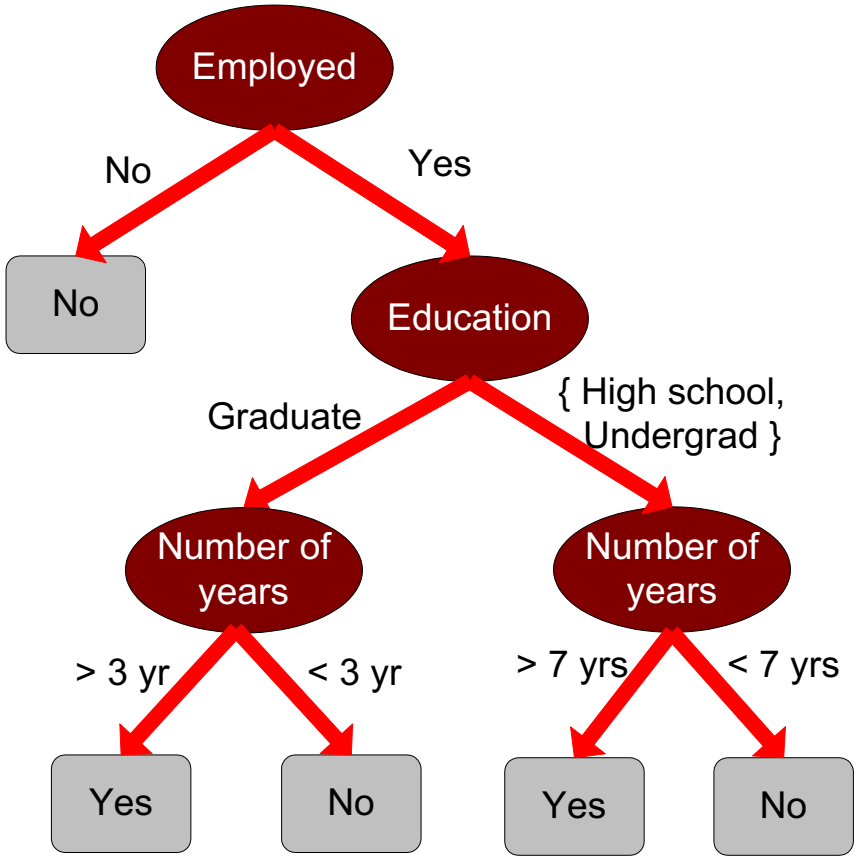
Predictive Modeling: Classification

- Find a model for class attribute as a function of the values of other attributes

Class

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Model for predicting credit worthiness

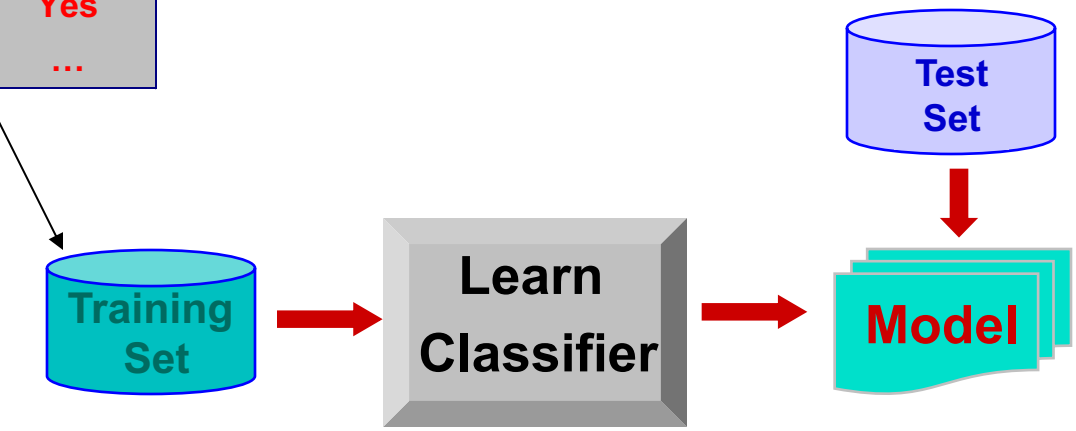


Classification Example

categorical categorical quantitative class

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
…	…	…	…	…

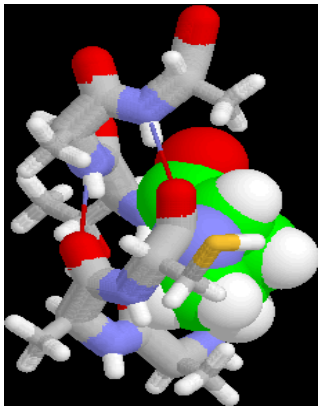


Examples of Classification Task

- Classifying credit card transactions as legitimate or fraudulent

- Classifying land covers (water bodies, urban areas, forests, etc.) using satellite data

- Categorizing news stories as finance, weather, entertainment, sports, etc
- Identifying intruders in the cyberspace
- Predicting tumor cells as benign or malignant



- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil

Classification: Application 1

- Fraud Detection
  - **Goal:** Predict fraudulent cases in credit card transactions.
  - **Approach:**
    - Use credit card transactions and the information on its account-holder as attributes
      - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions as fraud or fair transactions. This forms the class attribute
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account

Classification: Application 2

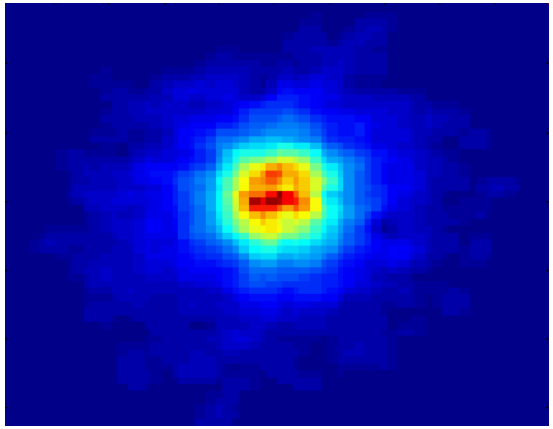
- Churn prediction for telephone customers
  - **Goal:** To predict whether a customer is likely to be lost to a competitor
  - **Approach:**
    - Use detailed record of transactions with each of the past and present customers, to find attributes
      - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
    - Label the customers as loyal or disloyal
    - Find a model for loyalty

From [Berry & Linoff] Data Mining Techniques, 1997

Classifying Galaxies

Courtesy: http://aps.umn.edu

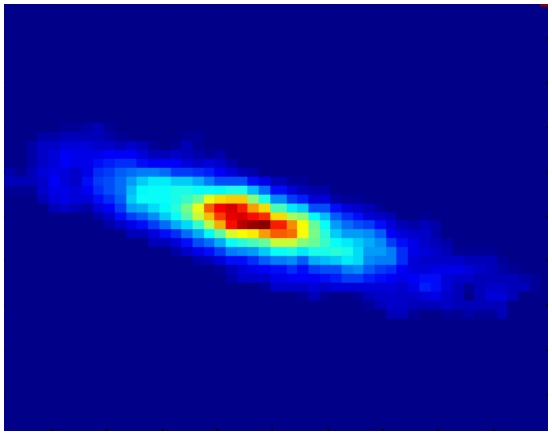
Early



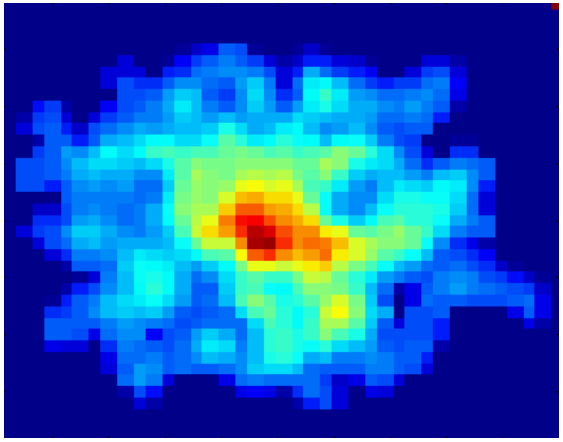
Class:
- **Stages of Formation**

Attributes:
- **Image features,**
- **Characteristics of light waves received, etc.**

Intermediate

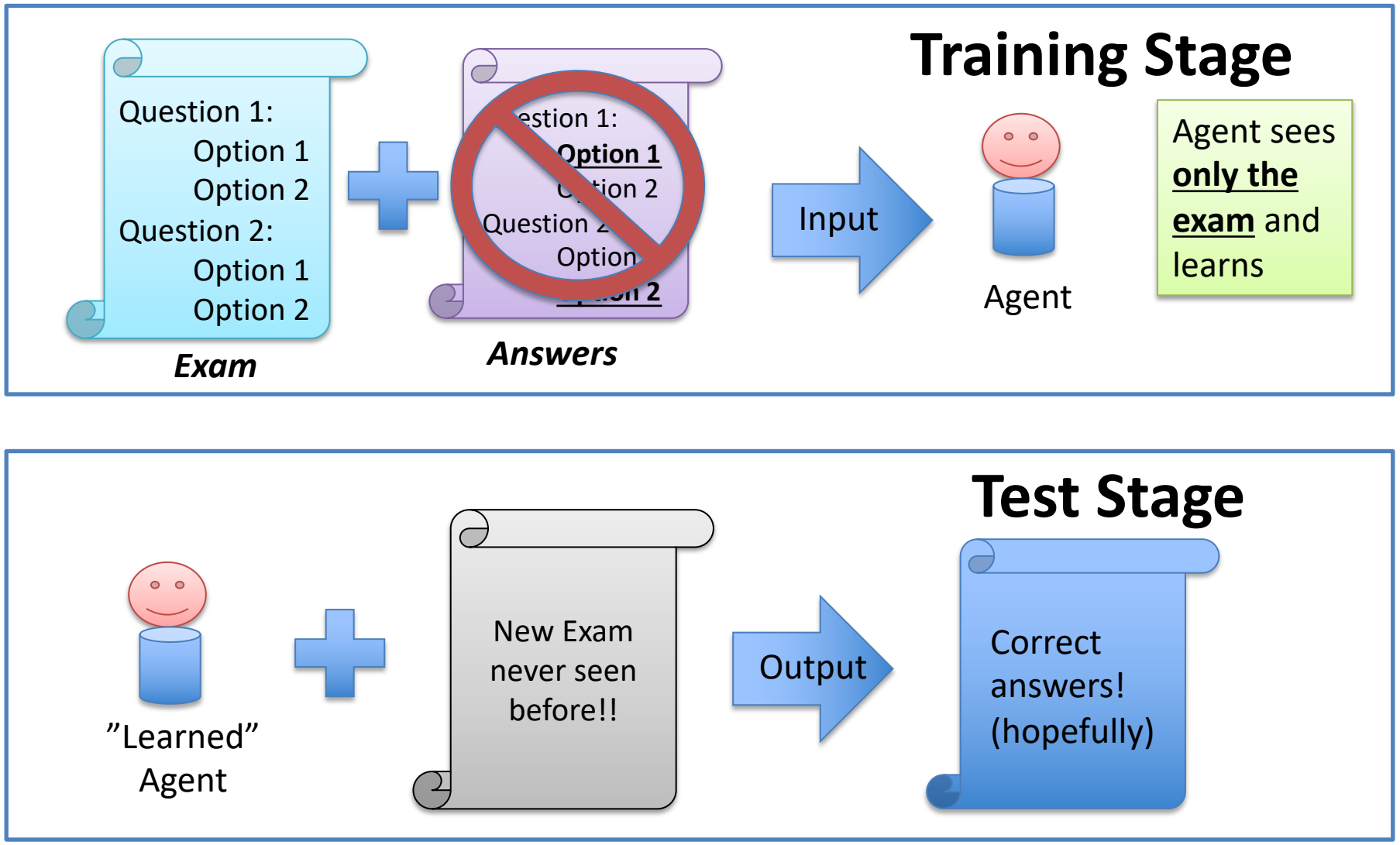


Late



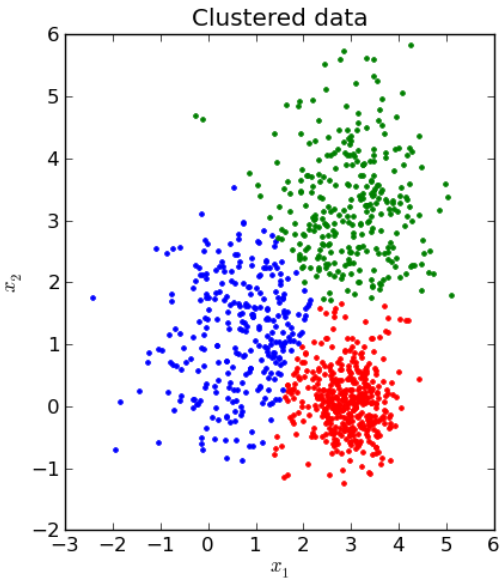
Data Size:
- **72 million stars, 20 million galaxies**
- **Object Catalog: 9 GB**
- **Image Database: 150 GB**

Unsupervised Learning

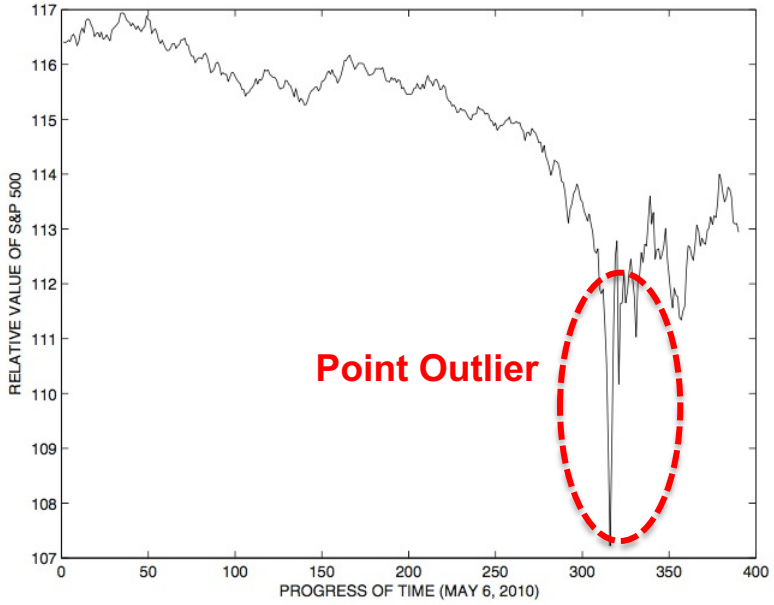


Types of Unsupervised Learning

- Clustering
- Principal Component Analysis (PCA)
- Association Analysis
- Anomaly Detection



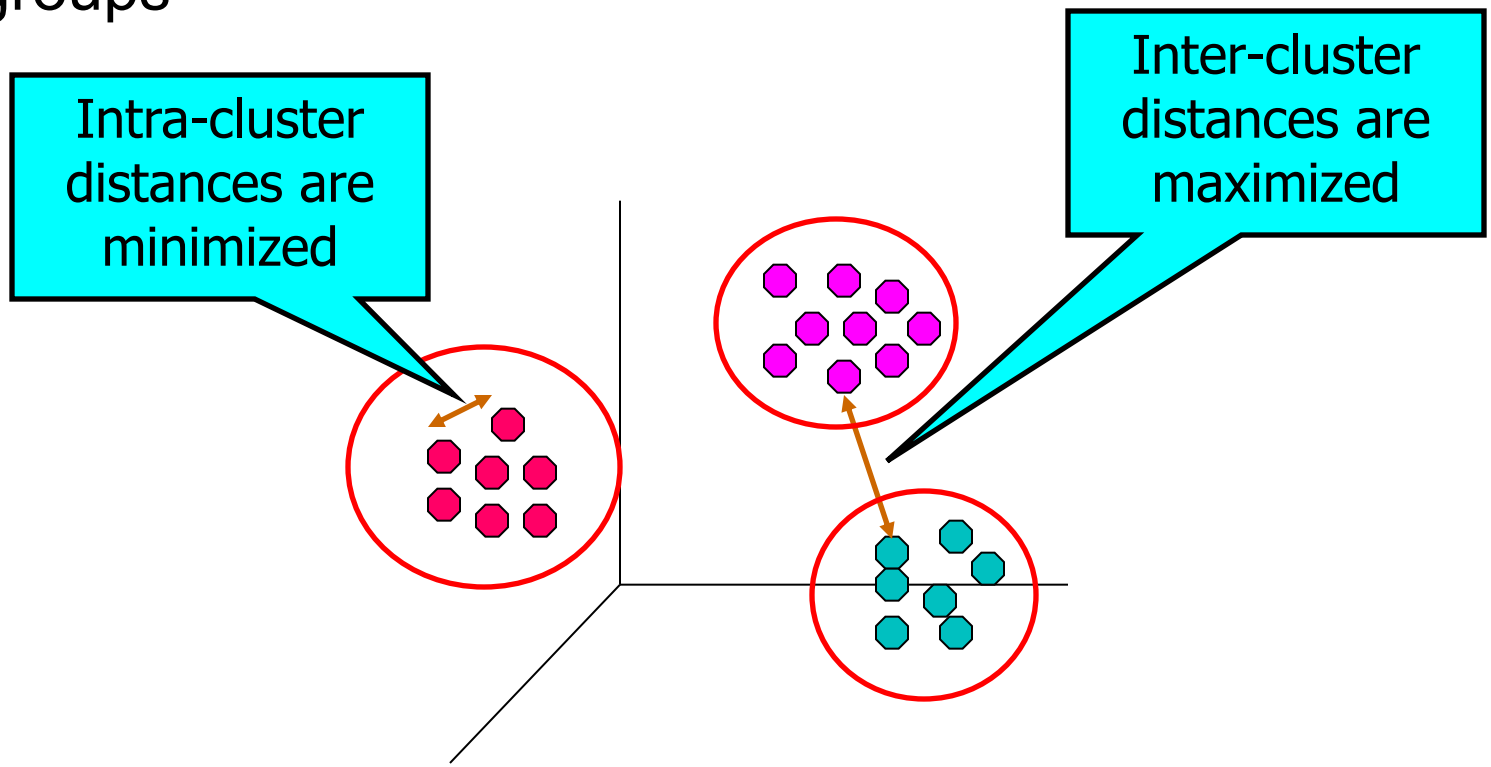
Clustering



Anomaly Detection

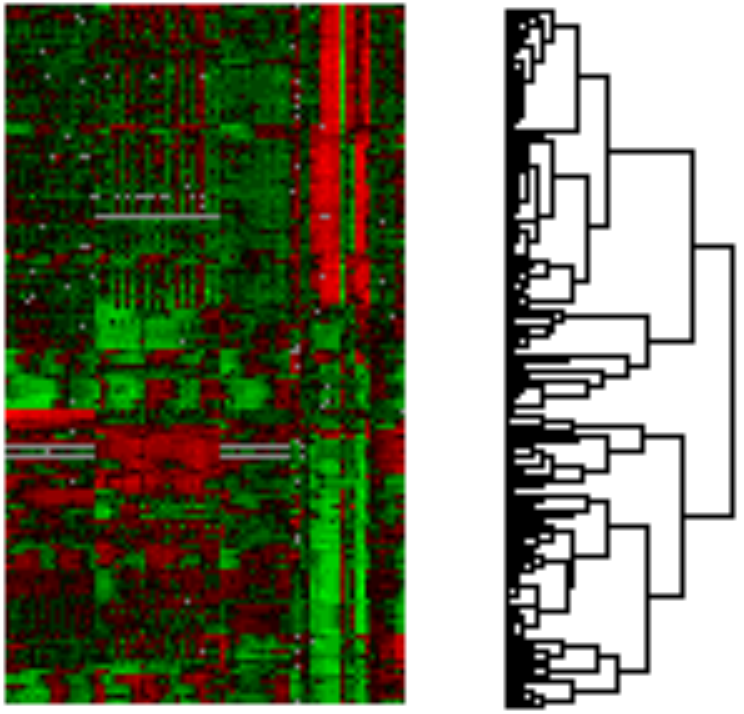
Clustering

- 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

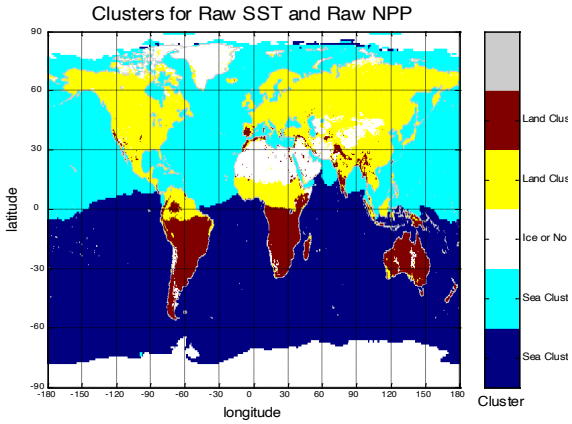


Applications of Cluster Analysis

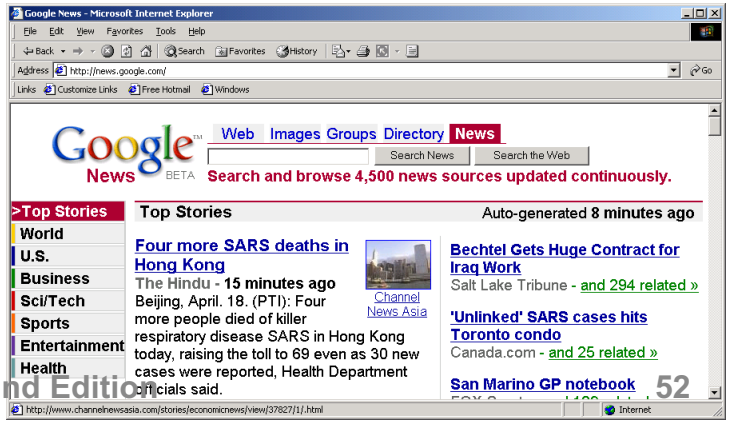
- **Understanding**
  - Custom profiling for targeted marketing
  - Group related documents for browsing
  - Group genes and proteins that have similar functionality
  - Group stocks with similar price fluctuations
- **Summarization**
  - Reduce the size of large data sets



Courtesy: Michael Eisen



Use of K-means to partition Sea Surface Temperature (SST) and Net Primary Production (NPP) into clusters that reflect the Northern and Southern Hemispheres.



Unsupervised Learning Examples

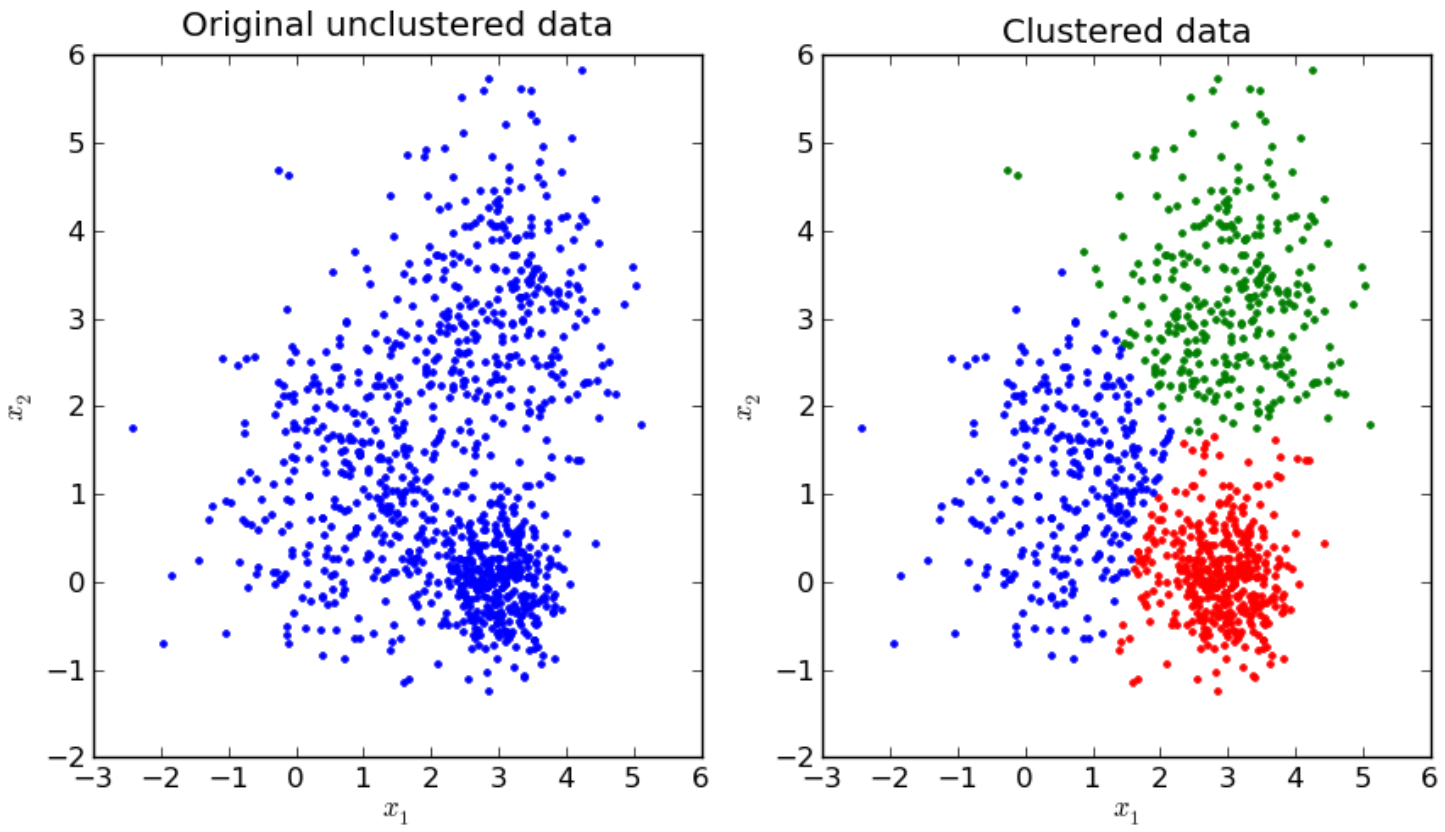


Figure from http://atreecodesinbrooklyn.tumblr.com/post/95407362512/k-means-clustering

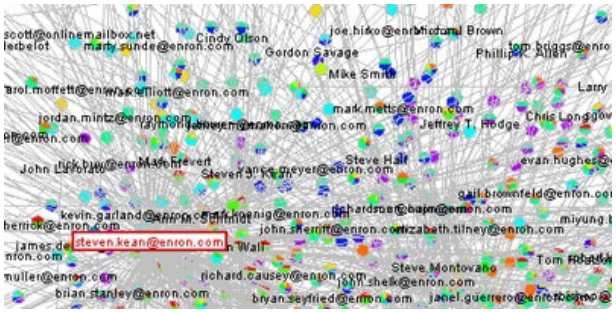
Clustering: Application 1

- Market Segmentation:
  - **Goal:** subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - **Approach:**
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering: Application 2

- Document Clustering:
  - **Goal:** To find groups of documents that are similar to each other based on the important terms appearing in them.
  - **Approach:** To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.

Enron email dataset



Data Mining Tasks

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.
- Description Methods
  - Find human-interpretable patterns that describe the data.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other 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}

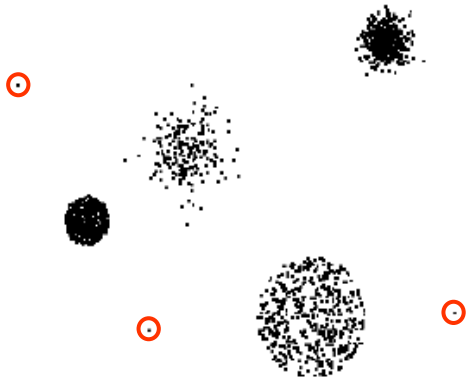
{Diaper, Milk} --> {Beer}

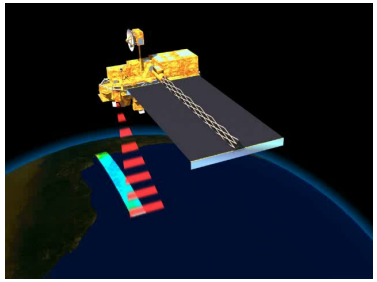
Association Analysis: Applications

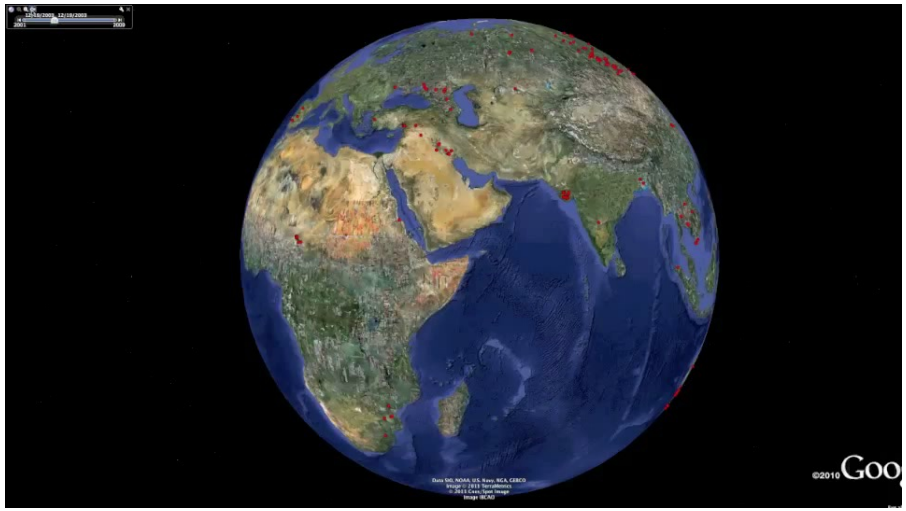
- Market-basket analysis
  - Rules are used for sales promotion, shelf management, and inventory management
- Telecommunication alarm diagnosis
  - Rules are used to find combination of alarms that occur together frequently in the same time period
- Medical Informatics
  - Rules are used to find combination of patient symptoms and test results associated with certain diseases

Deviation/Anomaly/Change Detection

- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection
  - Network Intrusion Detection
  - Identify anomalous behavior from sensor networks for monitoring and surveillance.
  - Detecting changes in the global forest cover.



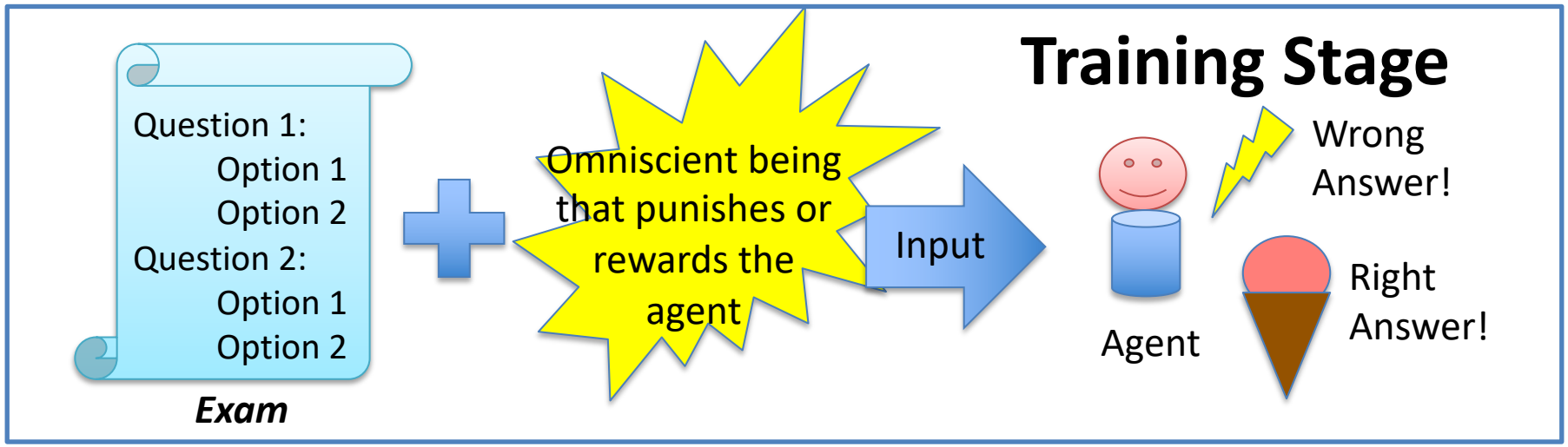


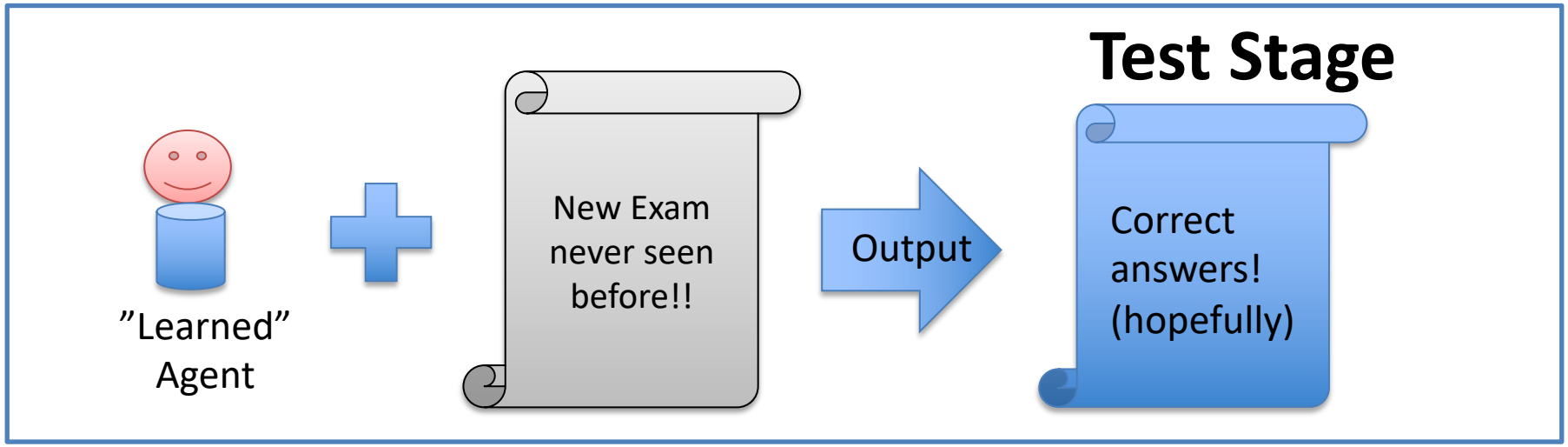


Motivating Challenges

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

Reinforcement Learning





Reinforcement Learning Examples

- Learning to park a car
  - https://www.youtube.com/watch?v=VMp6pq6_QjI
- Learning to flip pancakes
  - https://www.youtube.com/watch?v=W_gxLKSsSIE&list=PL5nBAYUyJTrM48dViibyi68urttMlUv7e
- Learning to play Atari Games
  - https://www.youtube.com/watch?v=V1eYniJ0Rnk