Machine Learning Overview

Key questions

- What can you do with machine learning?
  (and what can’t you do?)
- Where does ML fit in a data-to-decision workflow?

A short history of the world…

- The earth cooled
- (… time passes…)
- Humans arrived* and made decisions
- Language and Math/Logic is created and people write *rules* for making decisions
- Computers invented: able to make *faster* decisions
- Humans write fixed programs to make decisions using (probabilistic) distilled judgements about relationships in the data [Expert systems]
- Humans decided that *the computer would be better at learning statistical relationships* between attributes of the data [Machine Learning]

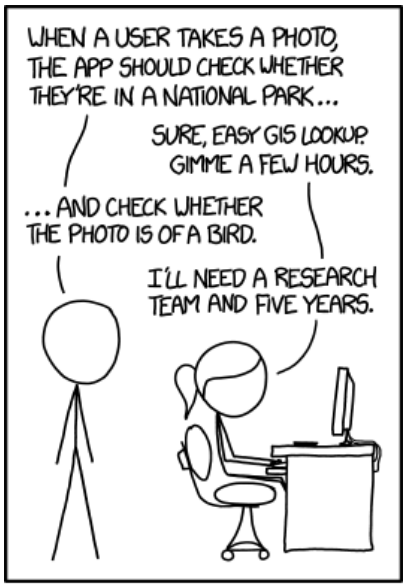
*Insert scientific and/or religious word of your choosing here

What can you do with Machine Learning?

- Infer how data elements are related
  - Does ice cream consumption depend on outdoor temperature?
- Make predictions about a target variable within a dataset
  - How much ice cream will be consumed this summer in Ohio?
- Determine the category something belongs to
  - Bird vs non-bird

https://www.pexels.com/search/birds/

https://en.wikipedia.org/wiki/Strelitzia



xkcd #1425 (View original here)

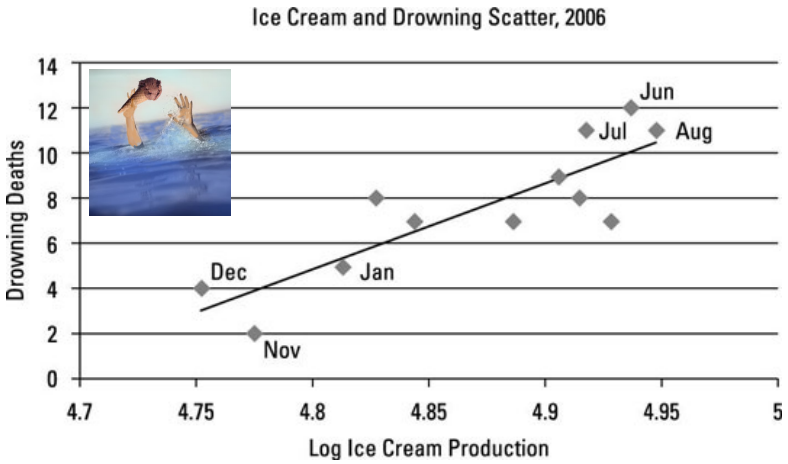
What can you do with Machine Learning(2) ?

- Optical Character/Number recognition
- Translate Text between languages
- Audio<->Text
- Find Components of an image
- Describe the contents of an image
- Pick the best <restaurant; movie; product; …> for me

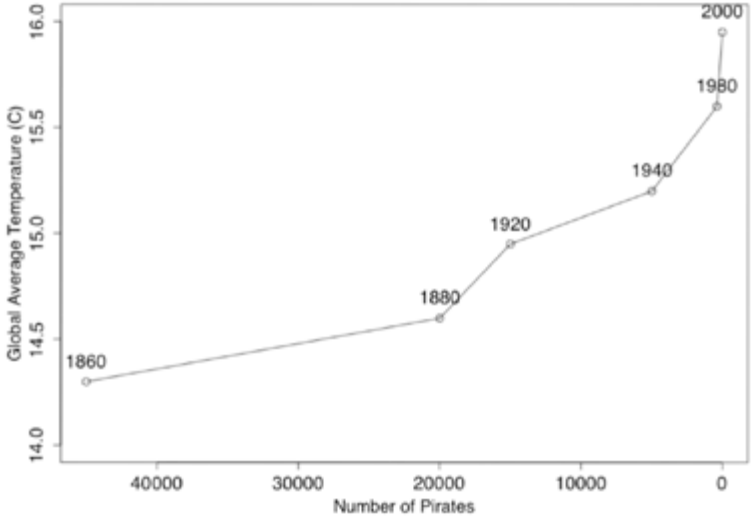
https://www.technologyreview.com/s/523326/how-google-cracked-house-number-identification-in-street-view/

What can't you do with ML?

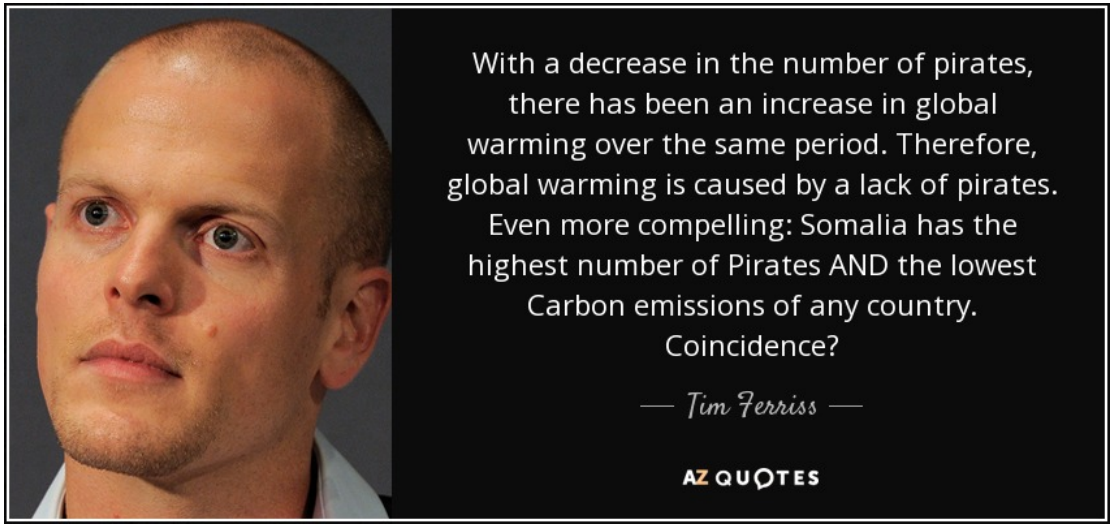
- Determine causality
  - Production of ice cream is correlated with drowning. Which way is the causality?
- Determine whether data is Evidence or Coincidence



http://www.dummies.com/education/economics/econometrics/the-role-of-casuality-in-econometrics/



http://sourcesandmethods.blogspot.com/2011/03/passport-ownership-cures-diabetes.html



What can't you do with *just* ML (2)

- Find the fastest route to an address on a map
  (use a cost-based pathfinding algorithm)
- Learn to Play Chess, Checkers, Go (efficiently)
  (use a heuristic deep search algorithm)
- Determine geocoordinates from only a photo
  (this only works for some common locations…)
- Drive a vehicle autonomously
  (this requires much more than ML)

Motivation for a Data-to-Decision Workflow

- Goal is to make better decisions
- Many ways to make decisions
  - Heuristic-based Human Judgement
  - Human-built computational models (e.g. expert systems)
  - Data Analysis (correlation, trends)
- Statistical Machine Learning suggests learning from Data
  - But where does the data come from?
  - And what resulting decision activities will the data support?

Where does ML fit in a workflow?

- A workflow is a defined sequence of steps used to process something / do something / answer a question
- ML is often in the middle of the workflow.
  For example, ML can help:
  - Determine what objects an image contains
  - Decide which category an observation in the data belongs to
  - Predict a value of a variable based on other observed values
- Before ML can occur, data* is gathered, wrangled, cleaned…
- After ML gives an output, decisions are made

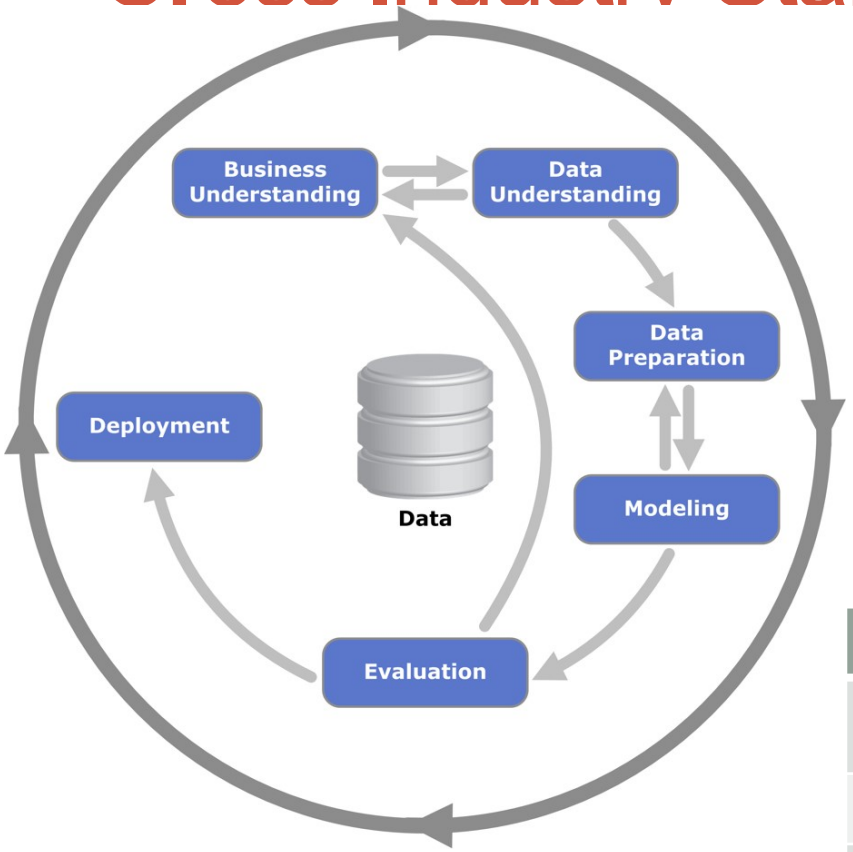
*Data can be numbers, text, audio, images, video….

Workflow Motivation

Since decisions derived from Data are only as good as the data, we must ensure the data collection, management, and analysis process is sound

- CRISP-DM
- KDD
- Audience Participation (x 2)

Cross Industry Standard Process for Data Mining (CRISP-DM)



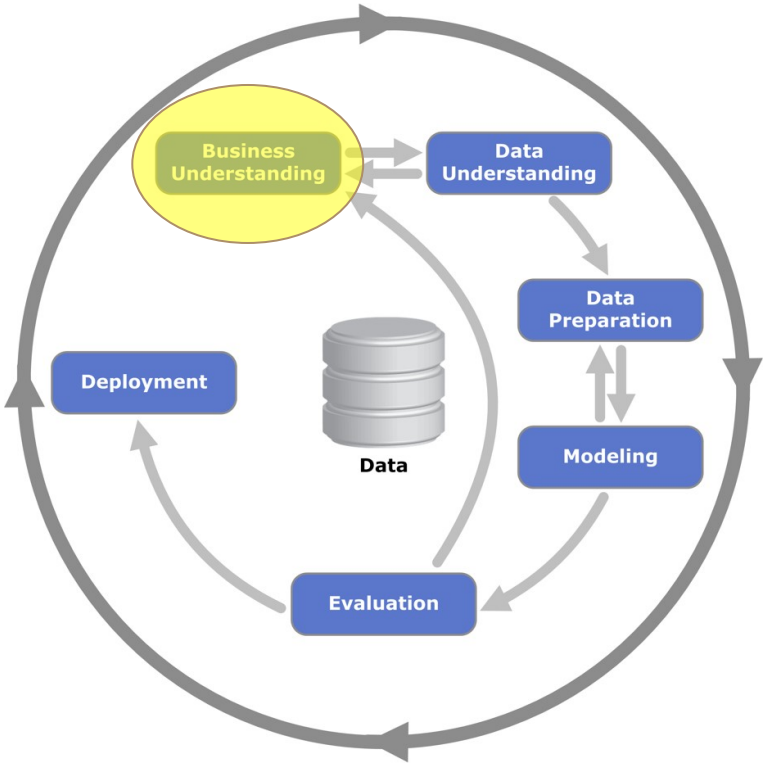
Process diagram showing the relationship between the different phases of CRISP-DM
https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining#/media/File:CRISP-DM_Process_Diagram.png
CC-SA Kenneth Jansen

Business understanding	Data understanding	Data Prep	Modeling	Evaluation	Deployment
Determine business objectives	Collect initial data	Select data	Select modeling techniques	Evaluate results	Plan deployment
Assess situation	Describe data	Clean data	Generate test design	Review Process	Plan monitoring & maintenance
Determine DM objectives	Explore data	Construct data	Build model	Determine next steps	Produce final report
Produce project plan	Verify data quality	Integrate data	Assess model		Review project
		Format data			

Ó. Marbán et al., "A Data Mining & Knowledge Discovery Process Model," in Data Mining and Knowledge Discovery in Real Life Applications, no. February, J. Ponce and A. Karahoca, Eds. Vienna, Austria: I-Tech, 2009, pp. 483–453.

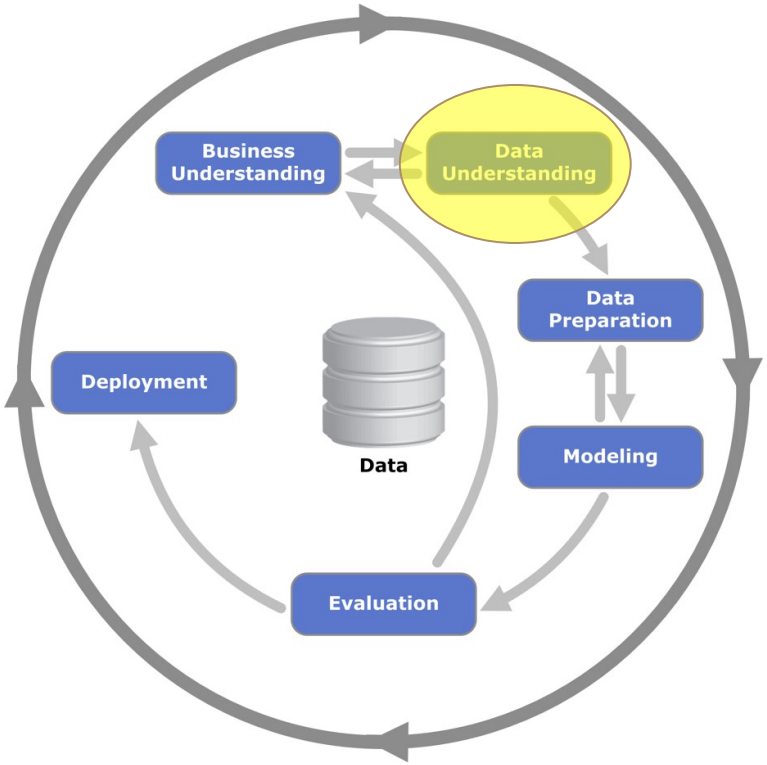
CRISP-DM: Business Understanding

- Understand project objectives and requirements
- Develop a data mining definition and plan



Business understanding	Data understanding	Data Prep	Modeling	Evaluation	Deployment
Determine business objectives	Collect initial data	Select data	Select modeling techniques	Evaluate results	Plan deployment
Assess situation	Describe data	Clean data	Generate test design	Review Process	Plan monitoring & maintenance
Determine DM objectives	Explore data	Construct data	Build model	Determine next steps	Produce final report
Produce project plan	Verify data quality	Integrate data	Assess model		Review project
		Format data			

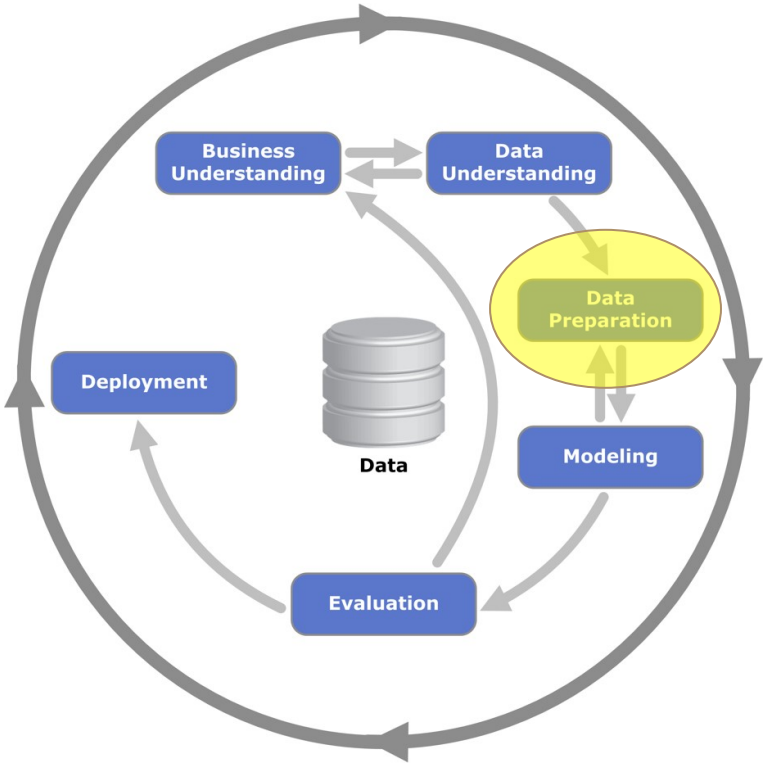
CRISP-DM: Data Understanding



- Collect Data
- Document / Describe data
- Become familiar / Explore data
- Identify data quality problems
- Determine data subsets
- Form investigative hypotheses

Business understanding	Data understanding	Data Prep	Modeling	Evaluation	Deployment
Determine business objectives	Collect initial data	Select data	Select modeling techniques	Evaluate results	Plan deployment
Assess situation	Describe data	Clean data	Generate test design	Review Process	Plan monitoring & maintenance
Determine DM objectives	Explore data	Construct data	Build model	Determine next steps	Produce final report
Produce project plan	Verify data quality	Integrate data	Assess model		Review project
		Format data			

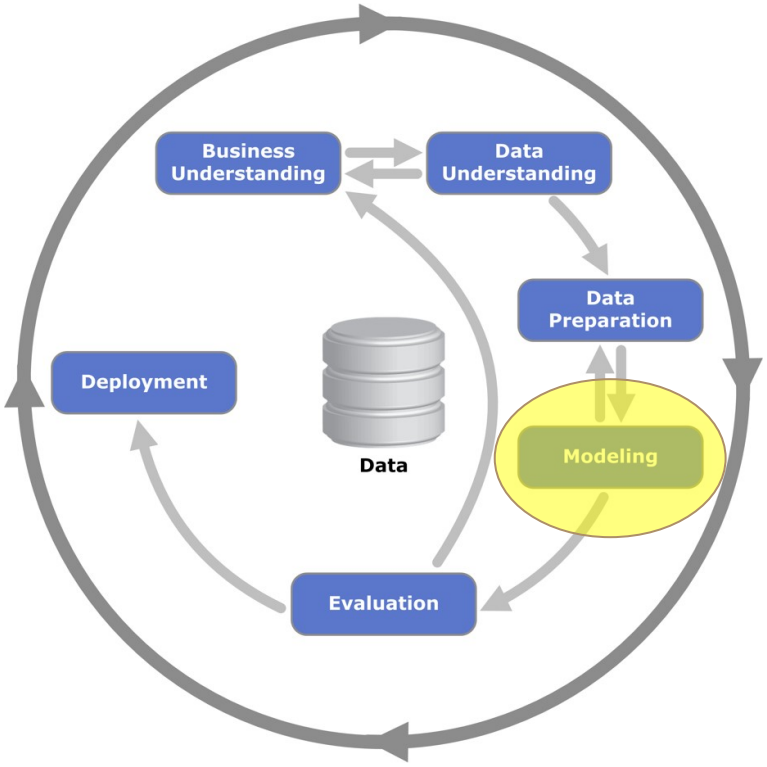
CRISP-DM: Data Preparation



- Construct final dataset from initial raw data
  - Select / Subset / Combine / Join data
  - Clean data (remove bad data / outliers?)
  - Construct missing data (impute)
  - Format data

Business understanding	Data understanding	Data Prep	Modeling	Evaluation	Deployment
Determine business objectives	Collect initial data	Select data	Select modeling techniques	Evaluate results	Plan deployment
Assess situation	Describe data	Clean data	Generate test design	Review Process	Plan monitoring & maintenance
Determine DM objectives	Explore data	Construct data	Build model	Determine next steps	Produce final report
Produce project plan	Verify data quality	Integrate data	Assess model		Review project
		Format data			

CRISP-DM: Modeling

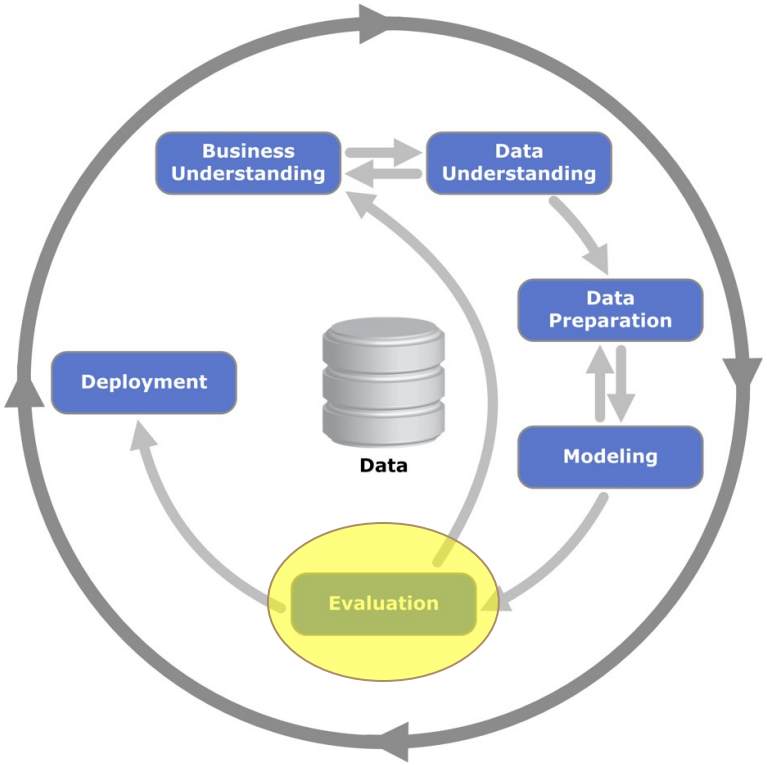


- Select modeling techniques
- Create the test/eval/assess design
- Build model
  - Fit & tune/calibrate model parameters
- Assess models

Business understanding	Data understanding	Data Prep	Modeling	Evaluation	Deployment
Determine business objectives	Collect initial data	Select data	Select modeling techniques	Evaluate results	Plan deployment
Assess situation	Describe data	Clean data	Generate test design	Review Process	Plan monitoring & maintenance
Determine DM objectives	Explore data	Construct data	Build model	Determine next steps	Produce final report
Produce project plan	Verify data quality	Integrate data	Assess model		Review project
		Format data			

CRISP-DM: Evaluation

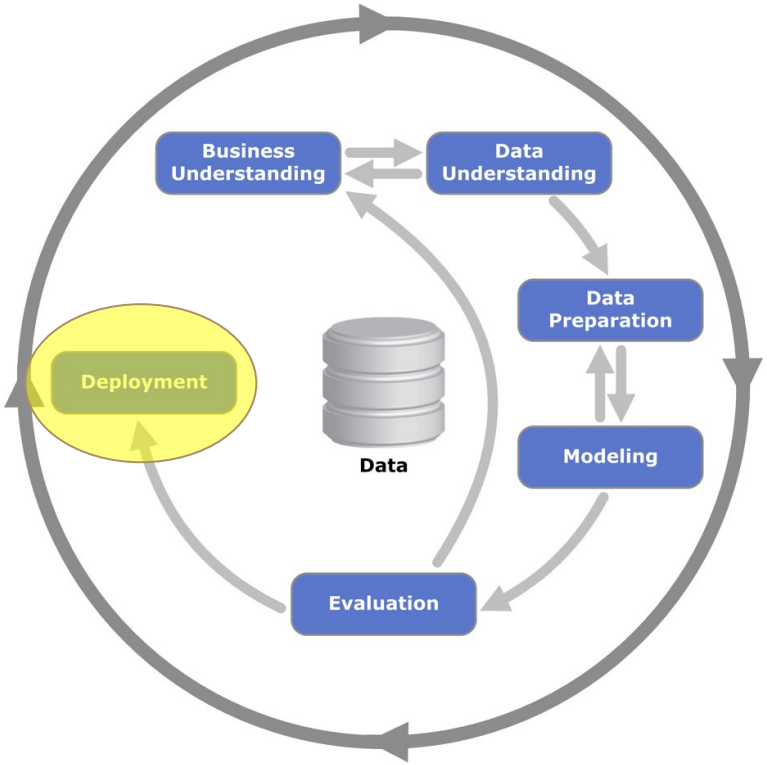
- Evaluate quality using performance criteria on unseen data
- Evaluate model's ability to answer business objectives
  - Example – model might require more data than we can collect
- Determine how to use the output of the model to make decisions



Business understanding	Data understanding	Data Prep	Modeling	Evaluation	Deployment
Determine business objectives	Collect initial data	Select data	Select modeling techniques	Evaluate results	Plan deployment
Assess situation	Describe data	Clean data	Generate test design	Review Process	Plan monitoring & maintenance
Determine DM objectives	Explore data	Construct data	Build model	Determine next steps	Produce final report
Produce project plan	Verify data quality	Integrate data	Assess model		Review project
		Format data			

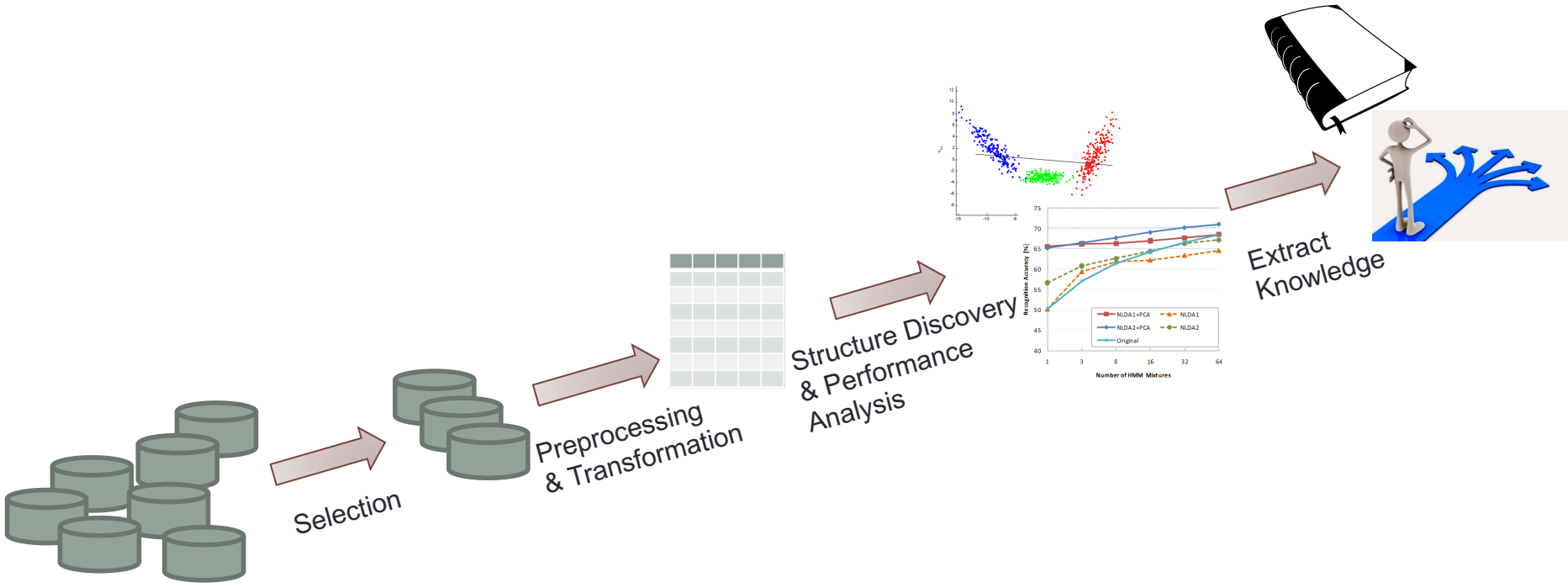
CRISP-DM: Deployment

- Integrate the model into the decision-making process
- Continuous Assessment: ensure model works well over time
- Maintenance: when to update/fix/retune the model



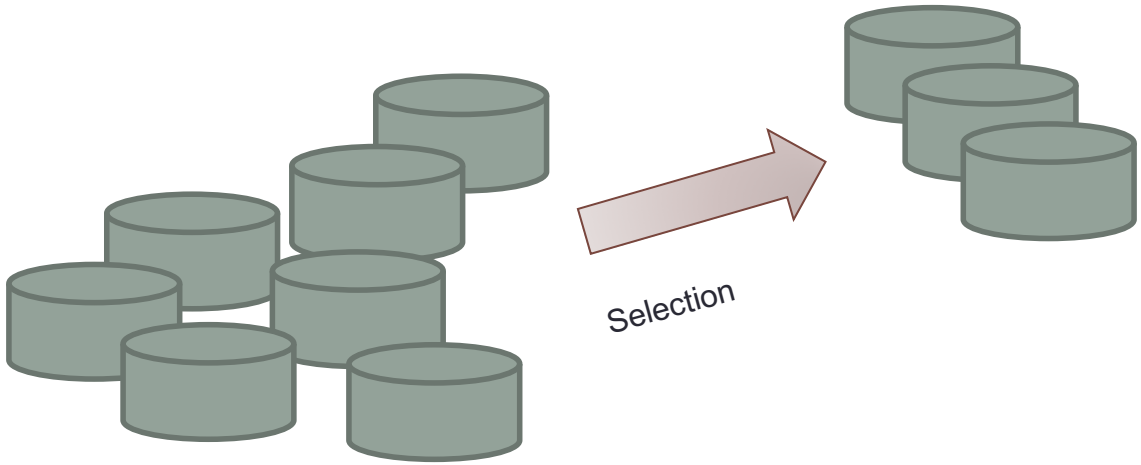
Business understanding	Data understanding	Data Prep	Modeling	Evaluation	Deployment
Determine business objectives	Collect initial data	Select data	Select modeling techniques	Evaluate results	Plan deployment
Assess situation	Describe data	Clean data	Generate test design	Review Process	Plan monitoring & maintenance
Determine DM objectives	Explore data	Construct data	Build model	Determine next steps	Produce final report
Produce project plan	Verify data quality	Integrate data	Assess model		Review project
		Format data			

Knowledge Discovery in Databases



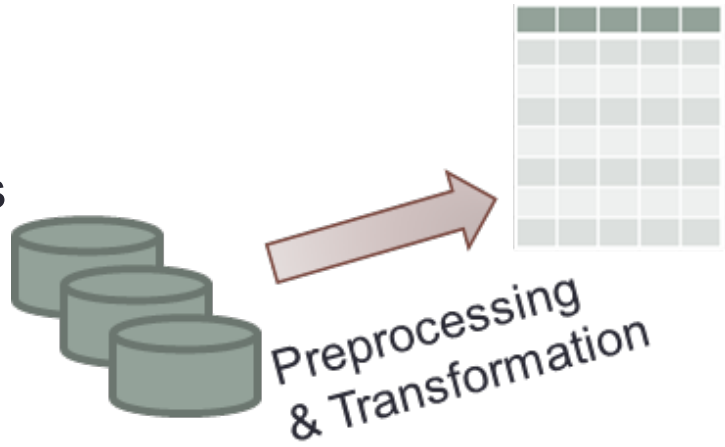
KDD: Selection

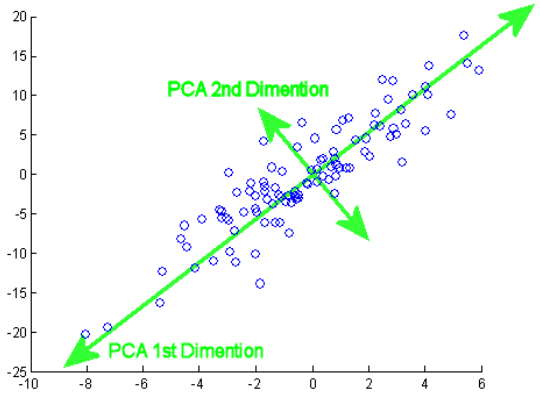
- Determine the question which needs to be answered
  - Scientific / Hypothesis Driven
- Choose the data which supports the question



KDD: Preprocessing and Transformation (Data Wrangling)

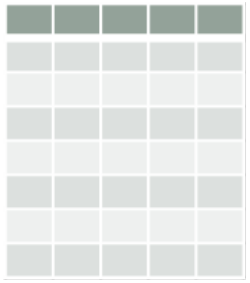
- Preprocessing:
  - Selected Data may be in many forms
    - Raw text / Image / Video / Markup Language / Time-Series Signal
  - Extract desired *Features* from data
    - Dimensionality reduction / filtering
    - Event counts, measurements, pixel values
    - Generate *observations* (rows) with *feature values* (columns)
    - Impute missing or incorrect values?
- Transformation:
  - Scale the feature values
  - Project observations into a different space/subspace



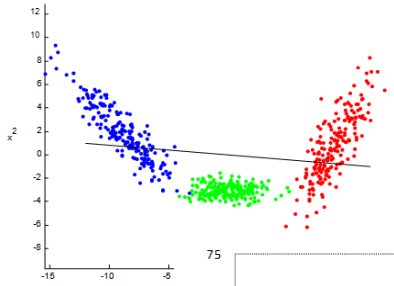


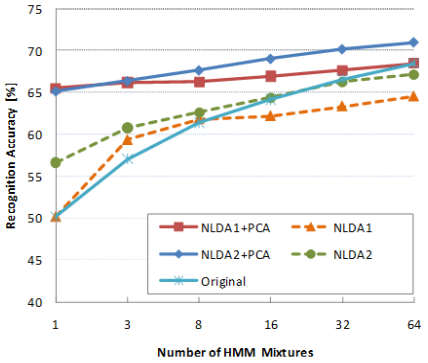
KDD: Structure Discovery and Performance Analysis

- Structure Discovery (Model Fitting)
  - Data Exploration – find trends
  - Regression – estimate a value
  - Classification – determine membership
  - Clustering – determine groupings
  - Inference – determine important features
  - Cross-Validation – tune the model
- Performance Analysis
  - Assess quality of model on unseen data



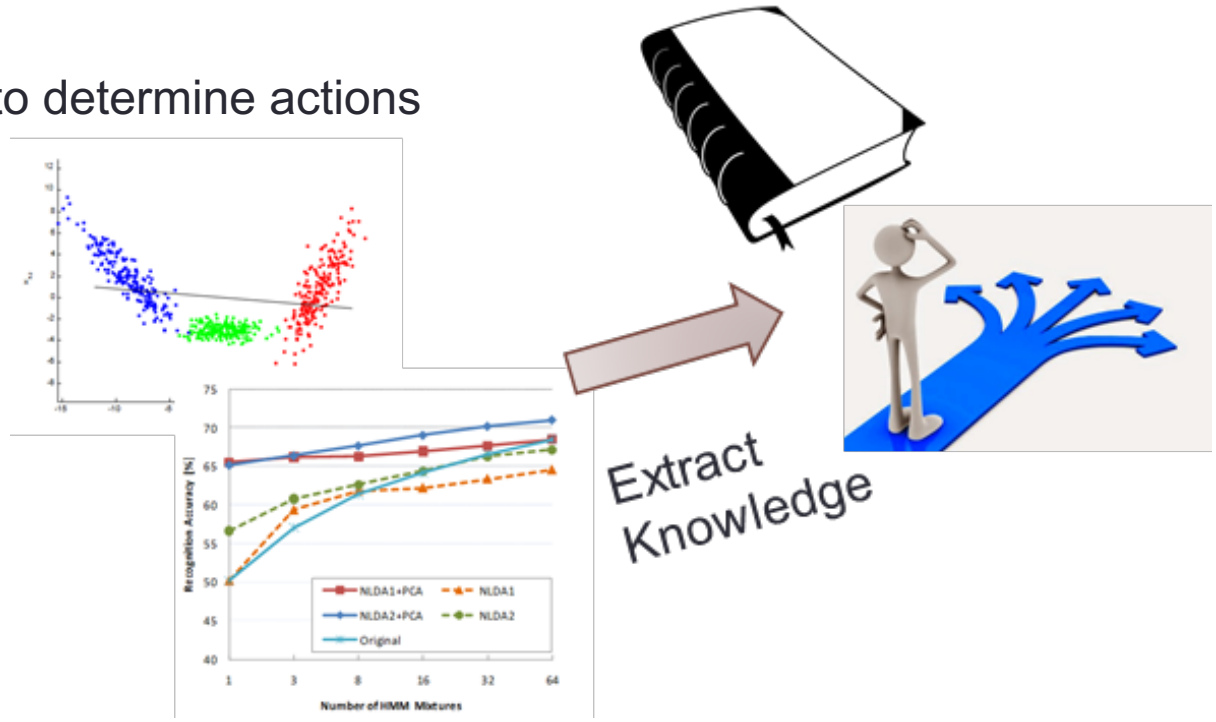
Structure Discovery & Performance Analysis





KDD: Extract Knowledge

- The ultimate goal of machine learning
  - Knowledge:
    - Generate understanding of the relationships within the data
    - Understand how features affect the output
  - Decision-making
    - Using knowledge to determine actions

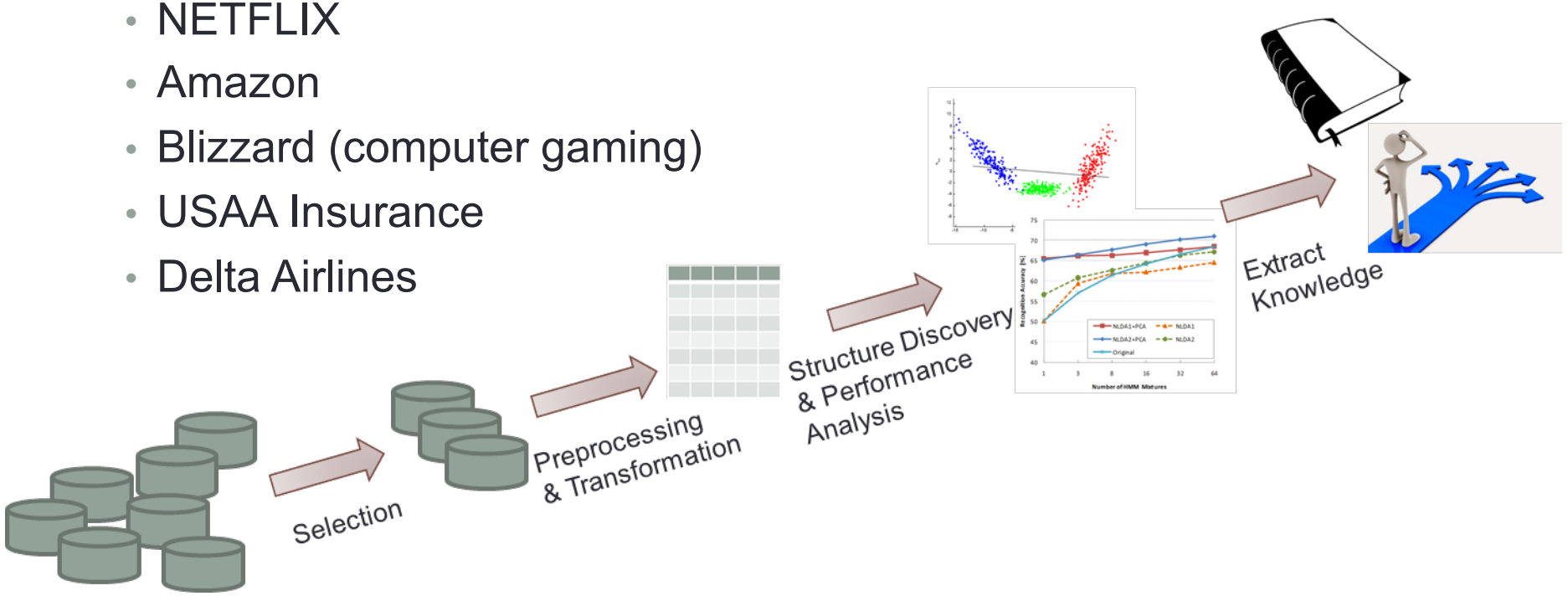


KDD: Challenges and Cautions

- Data mining vs. Data dredging
- Missing data
- Massive datasets / High dimensionality
- Overfitting & Statistical significance
- Concept drift / Nonstationary data

Audience Participation – Minute Paper #1

- Choose one of the following companies and describe how the company might use the KDD process to make a business decision. Describe details of the steps
  - NETFLIX
  - Amazon
  - Blizzard (computer gaming)
  - USAA Insurance
  - Delta Airlines



Audience Participation – Minute Paper #2

- How do you envision using the KDD process in ***your*** AFIT research?
  - Outline what needs to occur in each step
- What portion of the research could you work on for your class project?