

CSE 590
DATA SCIENCE FUNDAMENTALS

DATA SCIENCE
COMPONENTS AND TASKS

KLAUS MUELLER

COMPUTER SCIENCE DEPARTMENT
STONY BROOK UNIVERSITY AND SUNY KOREA

Lecture	Topic	Projects
1	Intro, schedule, and logistics	
2	Data Science components and tasks	
3	Data types	Project #1 out
4	Introduction to R, statistics foundations	
5	Introduction to D3, visual analytics	
6	Data preparation and reduction	
7	Data preparation and reduction	Project #1 due
8	Similarity and distances	Project #2 out
9	Similarity and distances	
10	Cluster analysis	
11	Cluster analysis	
12	Pattern miming	Project #2 due
13	Pattern mining	
14	Outlier analysis	
15	Outlier analysis	Final Project proposal due
16	Classifiers	
17	Midterm	
18	Classifiers	
19	Optimization and model fitting	
20	Optimization and model fitting	
21	Causal modeling	
22	Streaming data	Final Project preliminary report due
23	Text data	
24	Time series data	
25	Graph data	
26	Scalability and data engineering	
27	Data journalism	
	Final project presentation	Final Project slides and final report due

TASK #1: CLASSIFICATION

Predict which class a member of a certain population belongs to

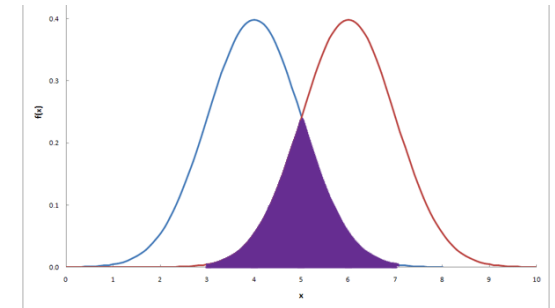
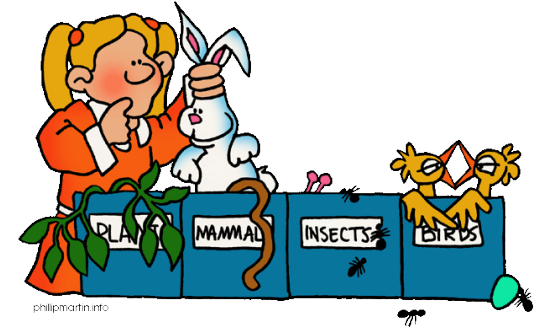
- absolute
- probabilistic

Require a classification model

- absolute
- probabilistic (likelihood)

Scoring with a model

- each population member gets a score for a particular class/category
- sort each class or member scores to assign
- scoring and classification are related

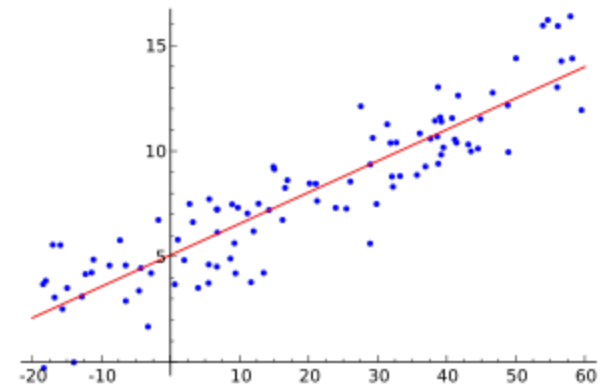


TASK #2: REGRESSION

Regression = value estimation

Fit the data to a function

- often linear, but does not have to be
- quality of fit is decisive



Regression vs. classification

- classification predicts that something will happen
- regression predicts how much of it will happen

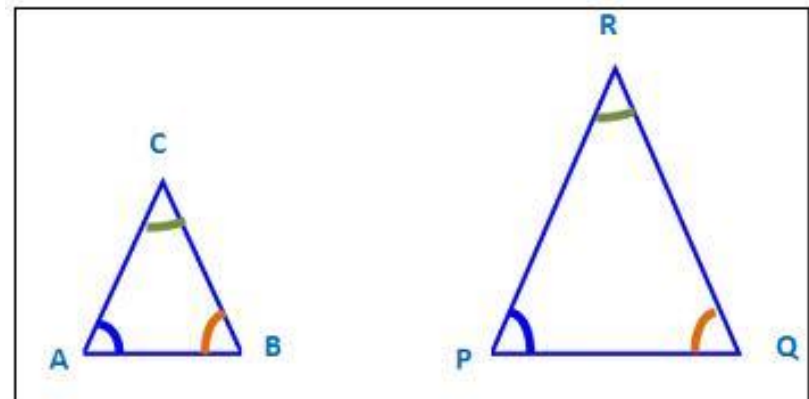
TASK #3: SIMILARITY MATCHING

Identify similar individuals based on data known about them

- need a measure of similarity
- features that define similarity
- characteristics

Similarity often part of

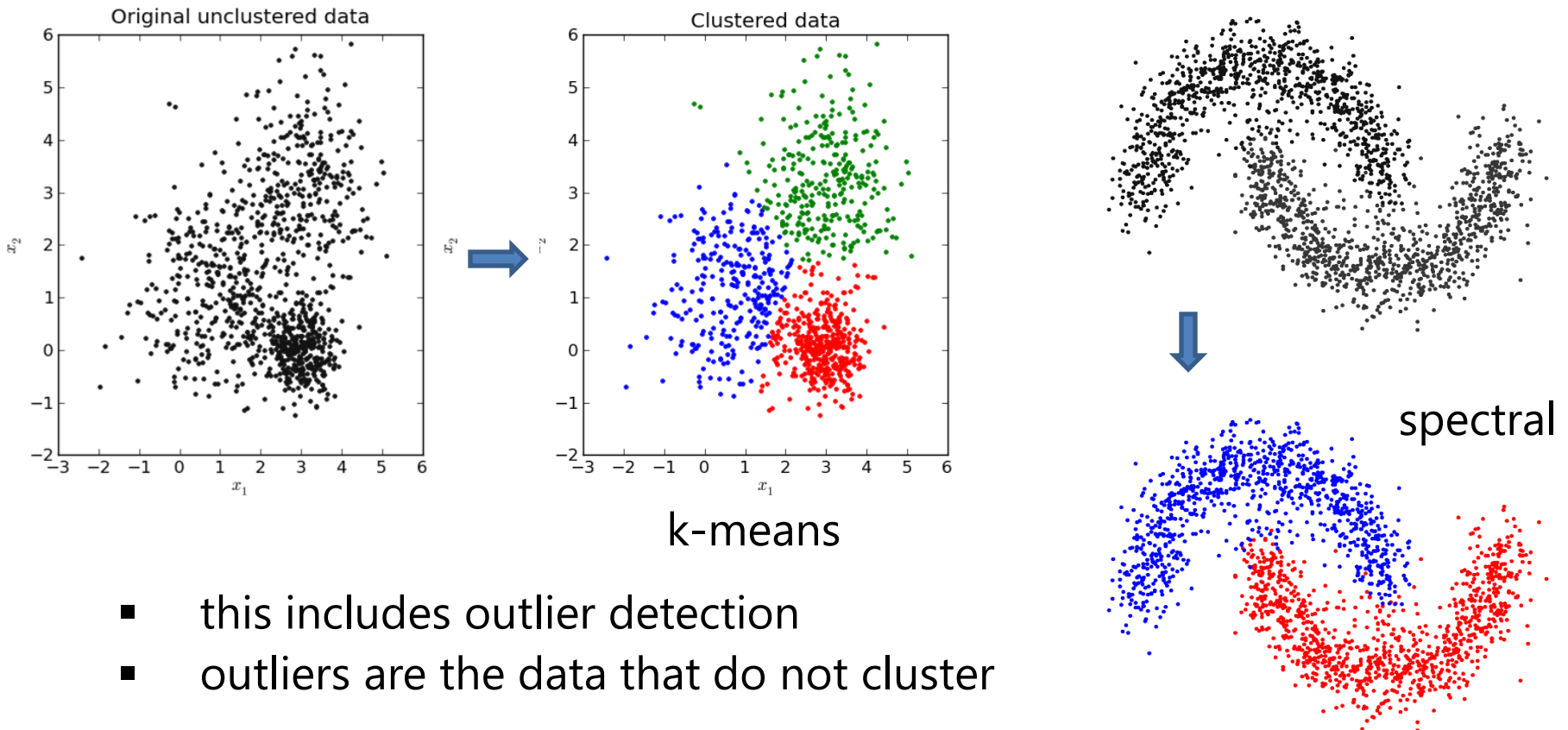
- classification
- regression
- clustering



TASK #4: CLUSTERING

Group individuals in a population together by their similarity

- preliminary domain exploration to see which natural groups exist



- this includes outlier detection
- outliers are the data that do not cluster

TASK #5: CO-OCCURRENCE GROUPING

Find associations between entities based on transactions involving them

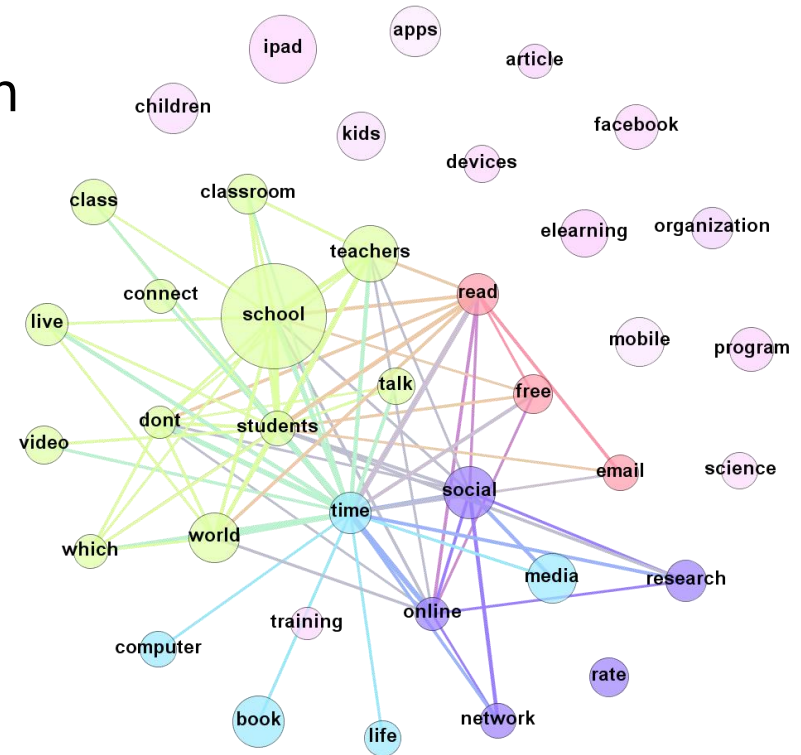
- what products are commonly purchased together?

Applications

- basket analysis
- recommender systems

Difference to clustering

- in clustering similarity is based on the object's attributes
- in co-occurrence similarity is based on objects appearing together



TASK #6: PROFILING

Also known as behavior description

- attempts to **characterize** the typical behavior of an individual, group, or population


Often used to establish behavioral norms for **anomaly detection**

- fraud detection
- intrusion detection

Examples:

- credit card fraud
- airport security

Example User Profile #2



Demographics

- Mid-20s
- Single; no children
- College degree
- Assistant Manager
- Earns \$45K

Psychographics

- Humanistic persona
- Gets recommendations from friends & others on social networks
- Values time with friends
- Likes to save money / get a good deal

Buying Habits

- Doesn't rush to make a purchasing decision
- Spends time on Facebook, Etsy, and fashion sites
- Uses mobile phone for texting, talking, and apps

Content & Messaging

- Customer testimonials
- Photos of people
- Conversion goals = Sign up for email newsletter, like Facebook page, and download mobile app

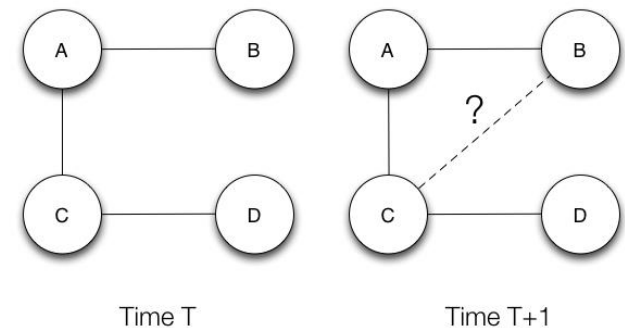
Kim Harris

Carol Morgan Cox | InterMedia4Web.com | @CivicLink

TASK #7: LINK PREDICTION

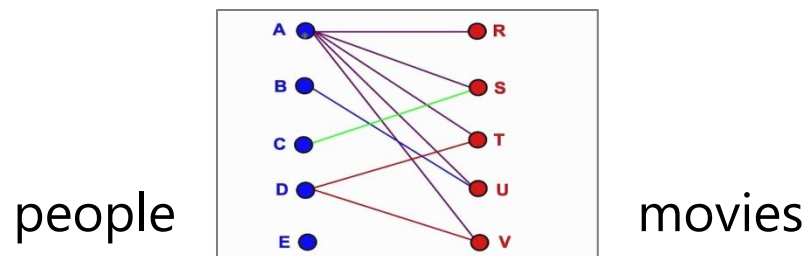
Predict connections between data items

- usually works within a graph
- predict missing links
- estimate link strength



Applications

- in recommendation systems
- friend suggestion in Facebook (social graph)
- link suggestion in LinkedIn (professional graph)
- movie suggestion in Netflix (bipartite graph people – movies)



TASK #8: DATA REDUCTION

Take a large dataset and substitute it with a smaller one

- keep loss of information minimal
- clustering and cleaning
- importance sampling
- dimension reduction
- data abstraction
- big data → small data
- find latent variables



Example – Movie *Taste*

- not directly measurable – latent variable
- derive from movie viewing preferences
- can reveal genre, etc.

TASK #9: CAUSAL MODELING

Understand what events or actions influence others



Different from predictive modeling

- tries **to explain why** the predictive model worked (or not)

Potentially unreliable when done from observational data

- conducting a targeted experiment is better
- even with big data...

Builds on counterfactual analysis

- an event is causal if mutating it will lead to undoing the outcome
- "If only I hadn't been speeding, my car wouldn't have been wrecked"
- downward vs. upward counterfactual thinking
- can explain happiness of bronze medalists vs. silver medalists
- just making the grade vs. just missing the grade

CASE STUDY: WHAT CAUSES LOW MPG

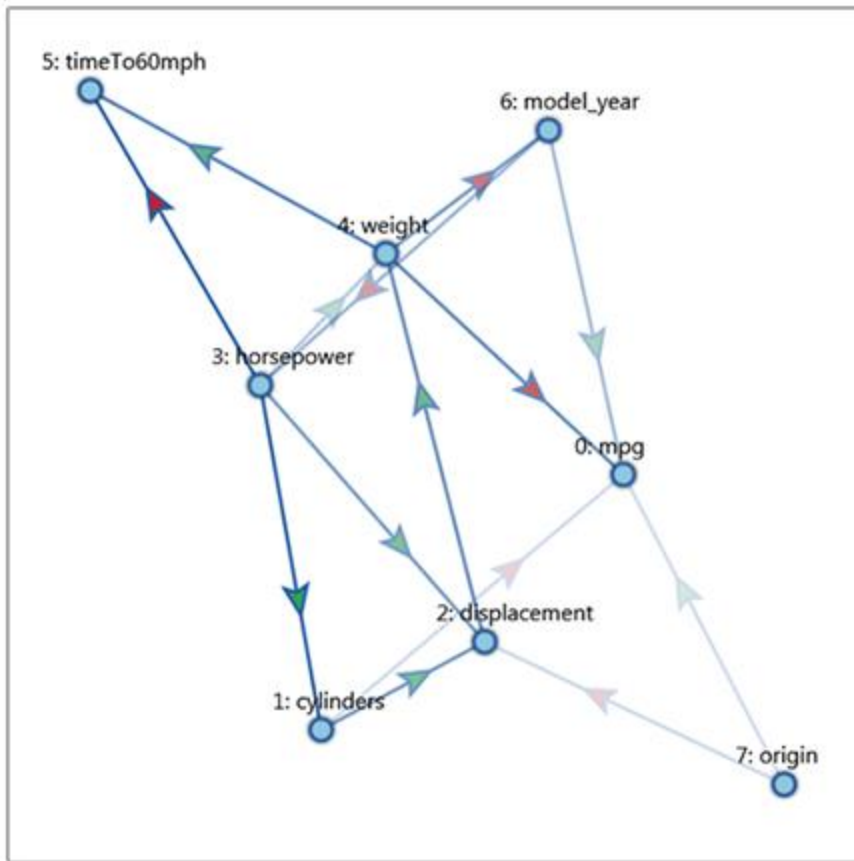
THE CAR DATA SET

Consider the salient features of a car (not really big data):

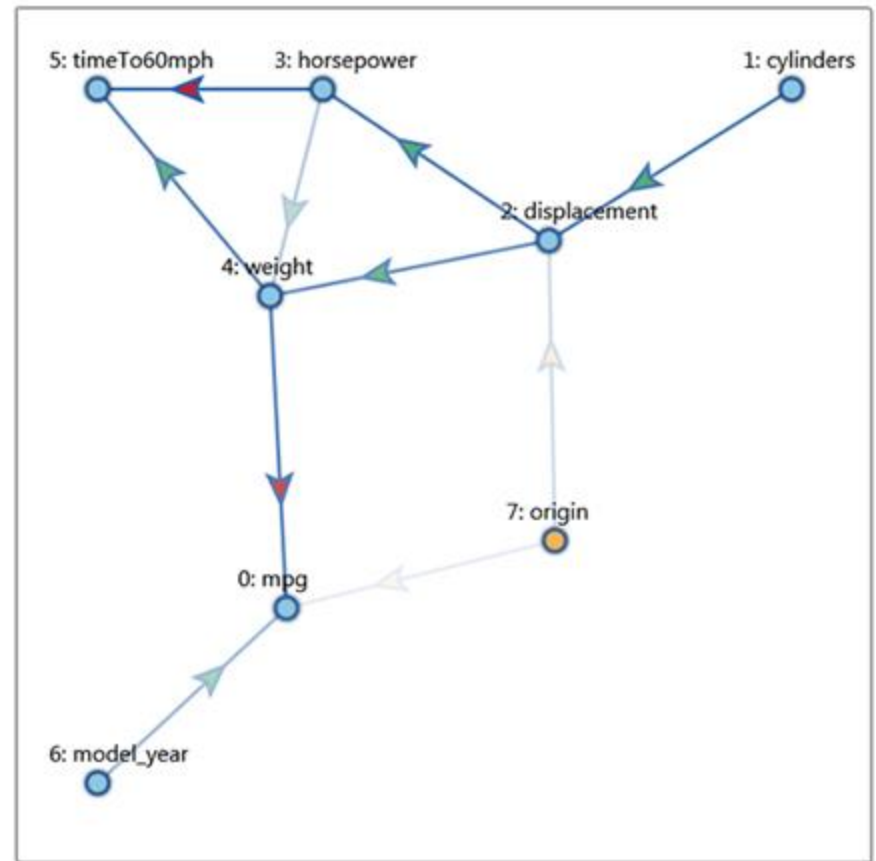
- miles per gallon (MPG)
- top speed
- acceleration (time to 60 mph)
- number of cylinders
- horsepower
- weight
- country origin

400 cars from the 1980s

GLOBAL LAYOUT OF THE CAR DATA

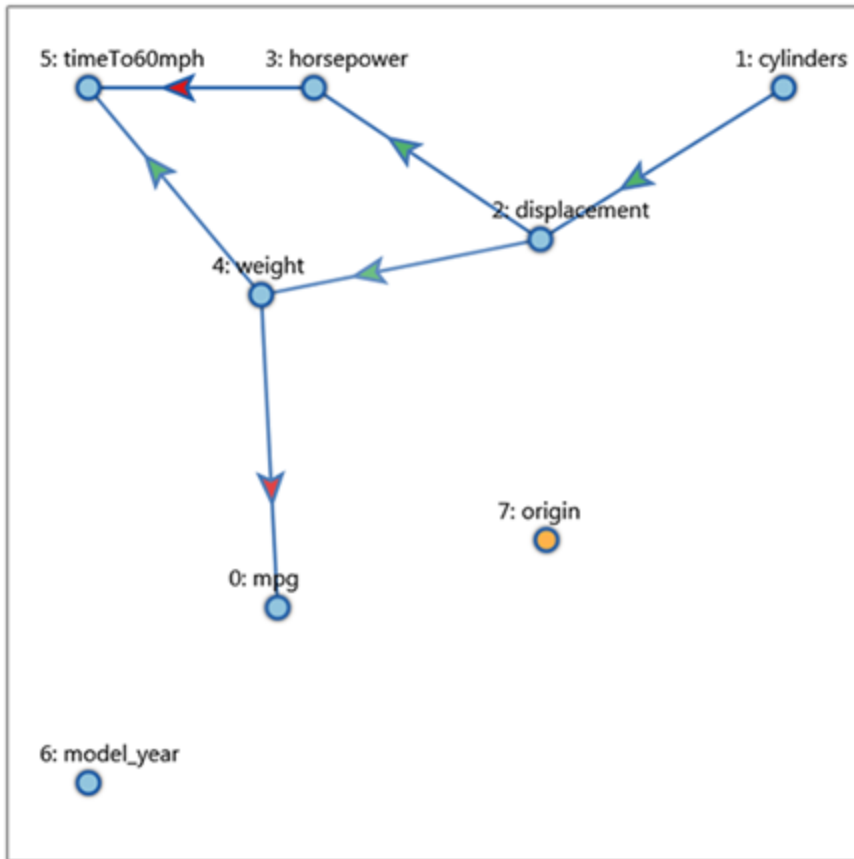


Random

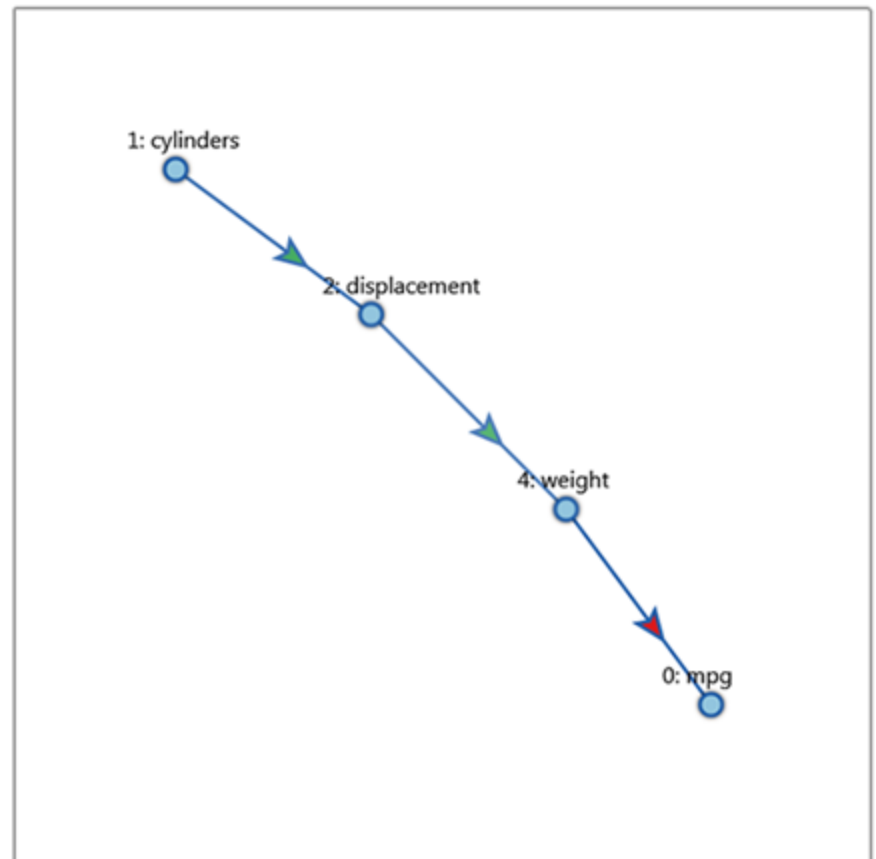


Causal

SEEKING THE CAUSE OF LOW MPG



Isolating MPG

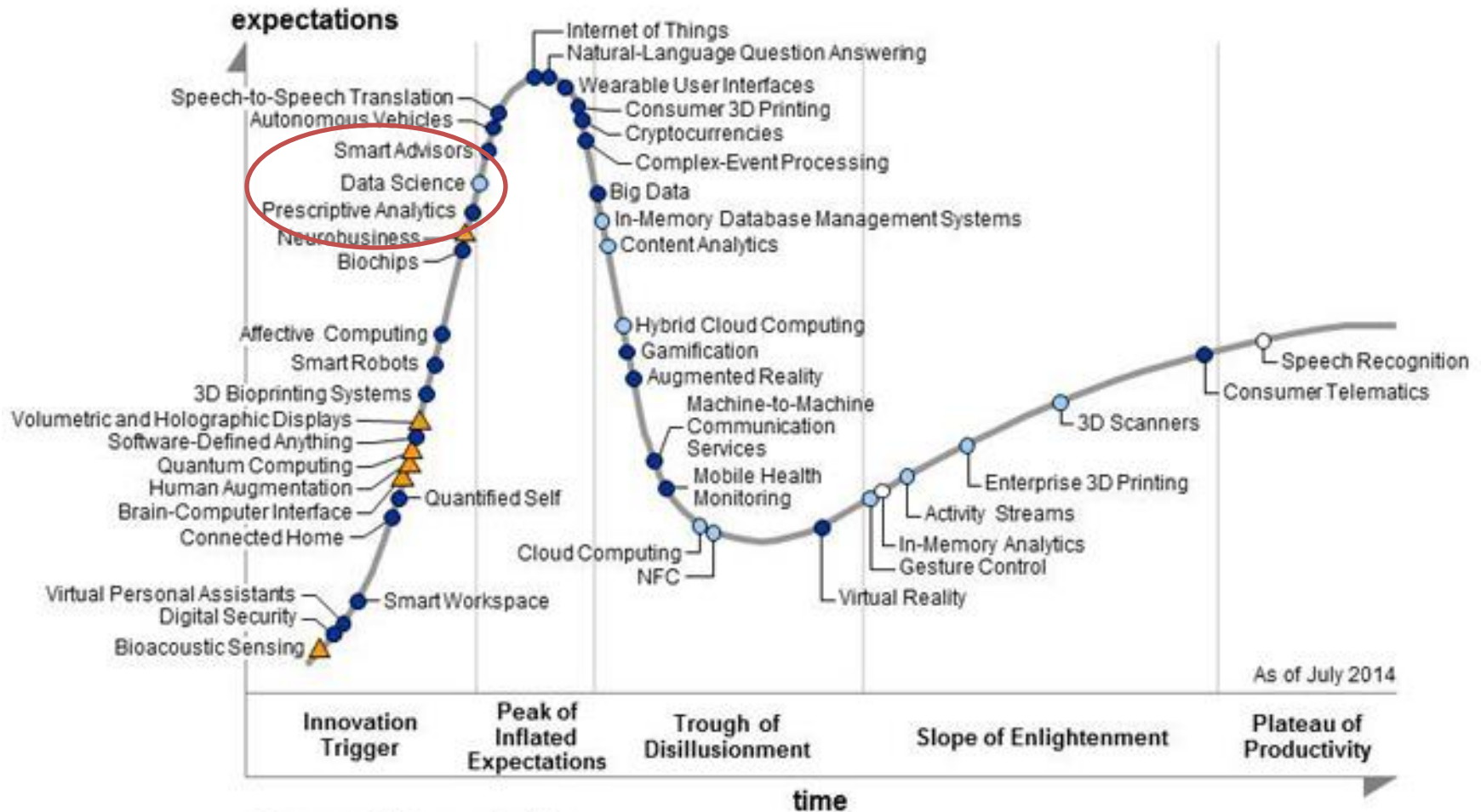


Causal Chain

VIDEO

[video](#)

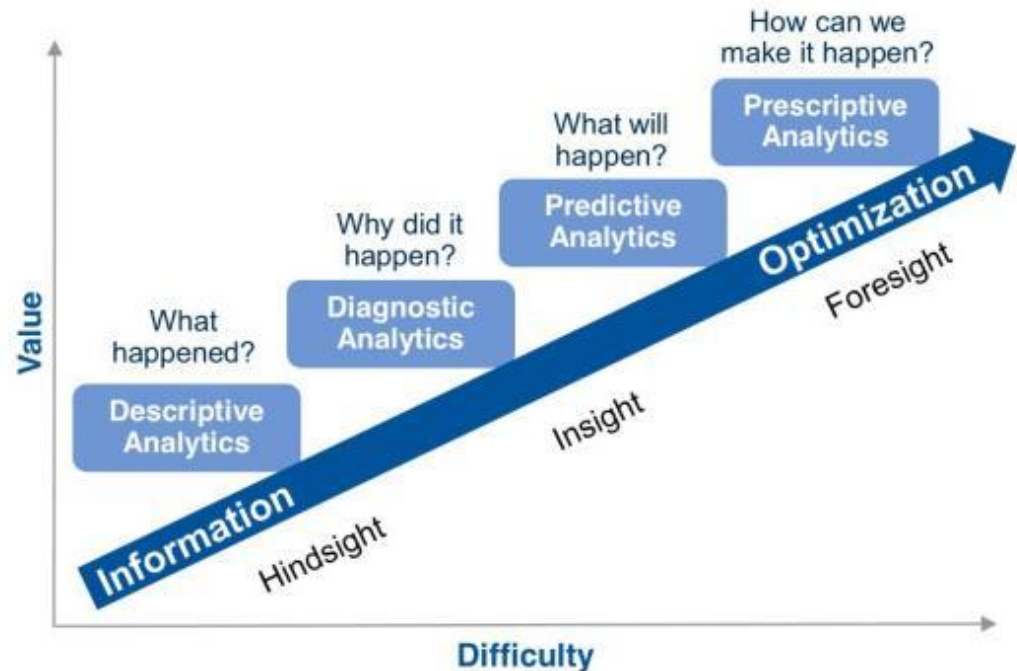
GARTNER HYPE CURVE



PRESCRIPTIVE ANALYTICS

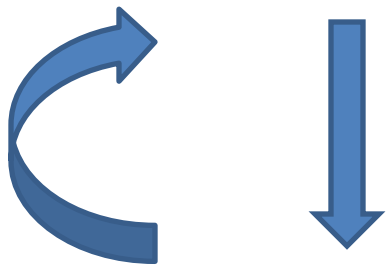
Prescriptive analytics – much related to data science

- suggests actions to benefit from the predictions
- shows decision makers the implications of each decision option
- synthesizes big data, math & business rules, machine learning to make predictions

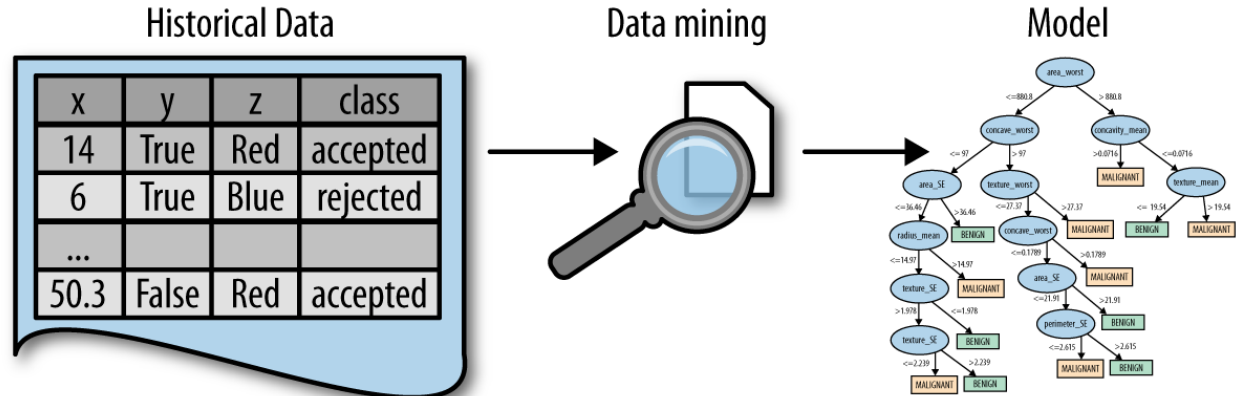


TWO DISTINCT PROCESSES

Mining existing data to produce a model



Using the model to make predictions on new data



Training data have all values specified

Model is deployed

Mining

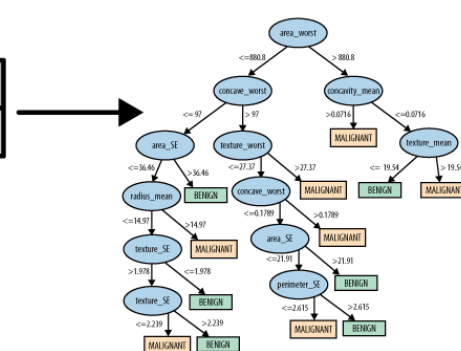
Use

New data item

x	y	z	class
30	false	Red	?

New data item has class value unknown (e.g. will customer accept?)

Model



Class: accepted, Probability: 0.88

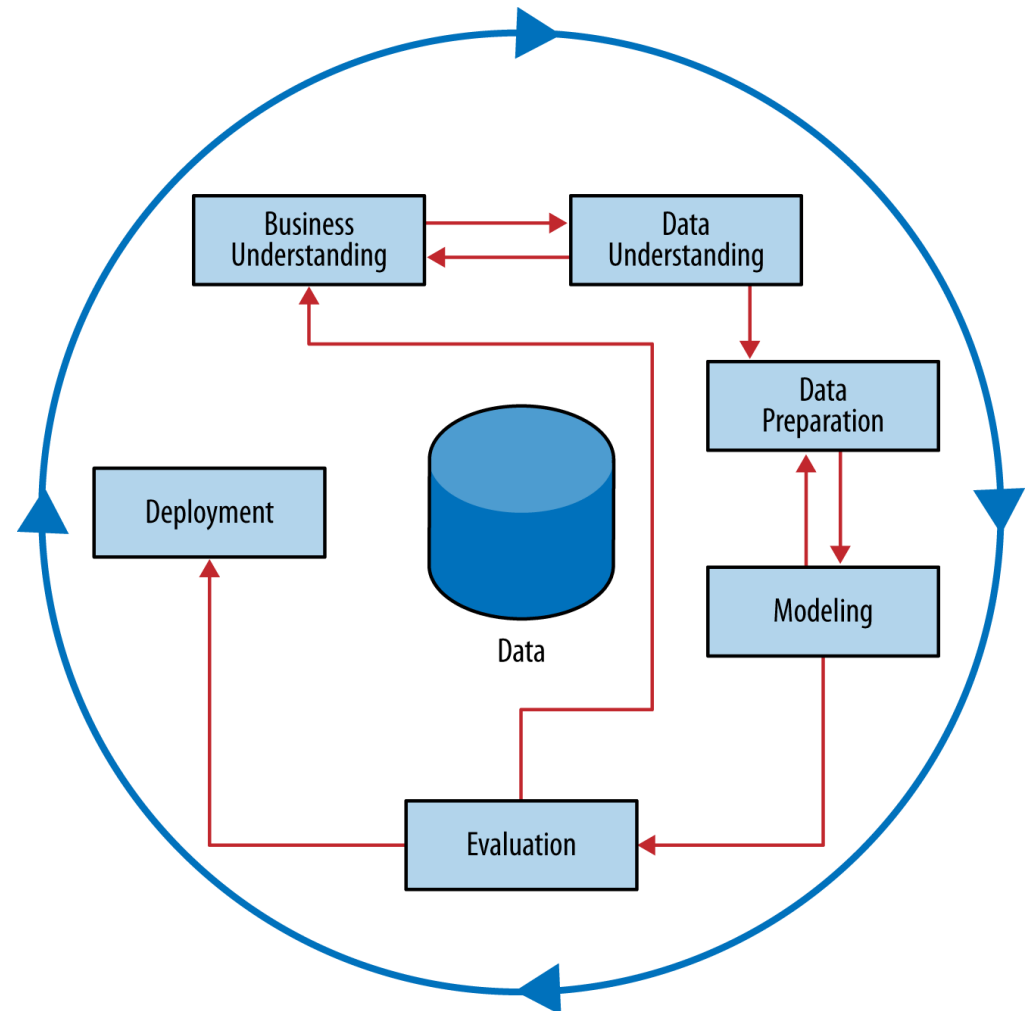
THE CRISP DATA MINING PROCESS

CRISP = Cross Industry Standard
Process for Data Mining

It's an iterative process

BU and DU cast the application
domain problems into one more
Data Science (DS) problems/tasks

Knowledge of DS fundamentals is
very important to come up with
novel solutions



EVALUATION AND DEPLOYMENT

Repeated evaluation with new data gains confidence

- it also readies the system for deployment

Deployment

- usually the evaluated system/model is just a prototype
- deploying a model into a production system typically requires that the model be recoded for the production environment
- usually for greater speed or compatibility with an existing system.
- this may incur substantial expense and investment
- usually the data science team is responsible for producing a working prototype, along with its evaluation
- then the development team takes over

SOME SAMPLE BUSINESS QUESTIONS

WHO ARE THE MOST PROFITABLE CUSTOMERS?

If “profitable” can be defined clearly based on existing data

- then this is a **straightforward database query**
- use a standard query tool to retrieve a set of customer records from a database
- **sort the results** by indicator of profitability
- **select** the highest ranked customers

So not really data science



IS THERE REALLY A DIFFERENCE BETWEEN THE PROFITABLE CUSTOMERS AND THE AVERAGE CUSTOMER?

This is a question about a hypothesis

“There is a difference in value to the company between the profitable customers and the average customer”

This can be statistically tested with confidence intervals

- can use the list derived before
- need to define *value* and *profitable* and score each customer
- they might be different
- then run the hypothesis test

BUT WHO REALLY ARE THESE CUSTOMERS? CAN I CHARACTERIZE THEM?

Data Mining:

- **extract the characteristics** of individual customers from a database via queries
- the WHAT, WHERE, and WHEN

Data Science:

- **determine what characteristics differentiate** profitable customers from unprofitable ones
- the WHY and HOW

WILL SOME PARTICULAR NEW CUSTOMER
BE PROFITABLE?

HOW MUCH REVENUE SHOULD I EXPECT
THIS CUSTOMER TO GENERATE?

Use data mining to retrieve historical data records

Produce predictive models of profitability

Apply to the new customer to generate the prediction

Subject of this course