COSC-254 Data Mining

Times & Location: MW 2—3.20pm, Science Center E110

Website: http://rionda.to/courses/cosc-254-s19/, **Moodle** for assignments and forum

Instructor: Matteo Riondato (he/his)

TA: Alexander Einarsson
Office Hours: Th 3—5.00pm, Science Center E210

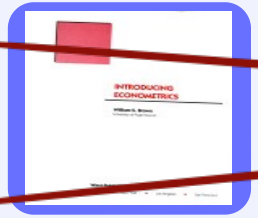
Prerequisites: COSC-211 Data Structures

What comes to mind when you think of "Data Mining"?

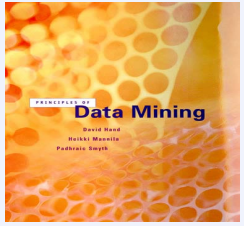
63

What is Data Mining?

"The process of extracting hidden patterns from data."

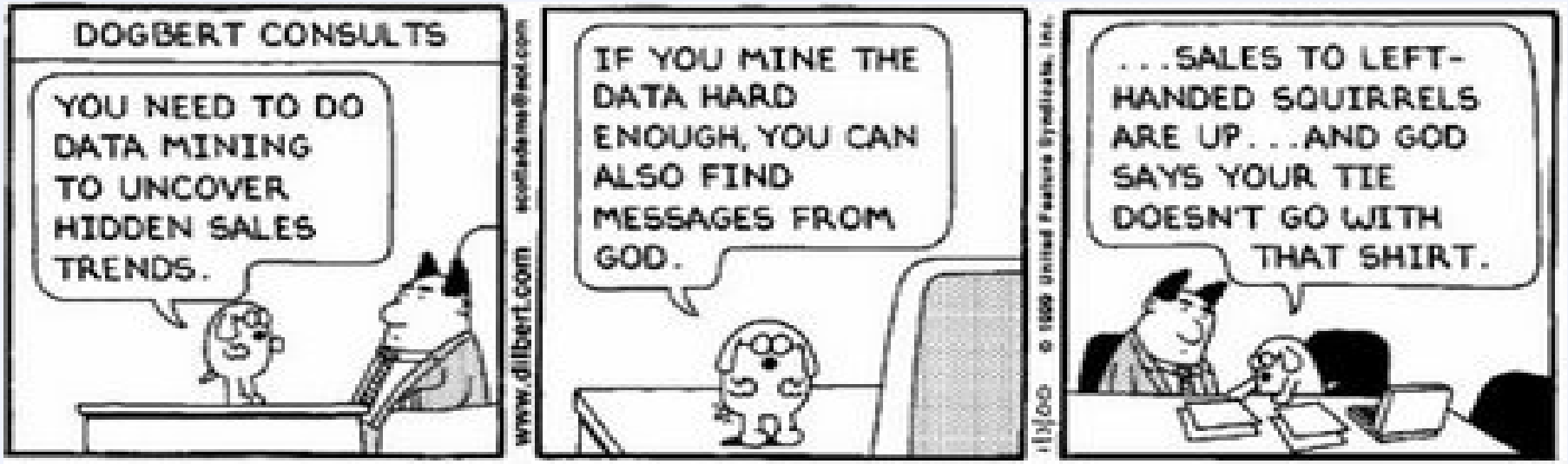


~~"An Unethical Econometric practice of massaging and manipulating the data to obtain the desired results."~~



"The analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are understandable and useful to the data owner."

Meaningfulness of Analytic Answers



G. Smith - “The Exaggerated Promise of So-Called Unbiased Data Mining”, Wired, 1/11/2019
https://www.wired.com/story/the-exaggerated-promise-of-data-mining/

DM Research

Important Conferences on DM

KDD, WSDM, WWW, CIKM, ICDM, SDM, ECML PKDD, VLDB

Important Journals on DM

DAMI, TKDD, TKDE, KAIS, TODS, VLDBJ

Performance Evaluation Initiatives / Benchmarks

KDD Cup, http://www.sigkdd.org/kddcup/index.php

(see DBLP bibliography for full detail, http://www.dblp.org/)

What will we learn?

- **We will learn to mine different types of data:**
  - Data is high dimensional
  - Data is a graph
  - Data is infinite/never-ending
  - Data is labeled
- **We will learn to use different models of computation:**
  - MapReduce
  - Streams and online algorithms
  - Single machine in-memory

What will we learn?

- **We will learn to solve real-world problems:**
  - Recommender systems
  - Market Basket Analysis
  - Spam detection
  - Duplicate document detection
  - Finding important information in a graph/network
- **We will learn various “tools”:**
  - Linear algebra (SVD, Rec. Sys., Community detection)
  - Dynamic programming (frequent itemsets)
  - Hashing (LSH, Bloom filters)