CS57300
PURDUE UNIVERSITY

JANUARY 8, 2019

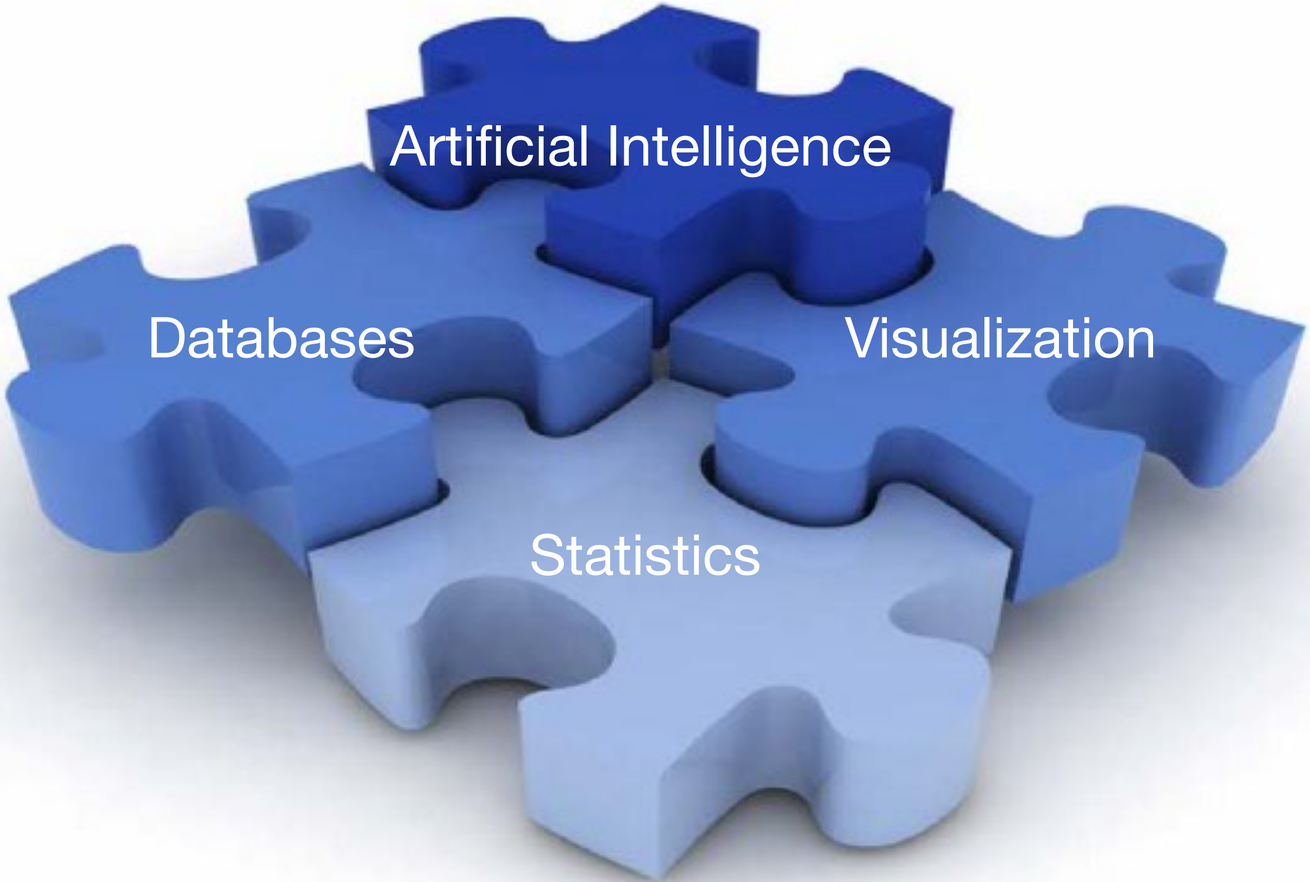
DATA MINING

INTRODUCTION

- What is data mining?
- Data mining examples
- Why now?
- Data mining process
- Course overview

DATA MINING

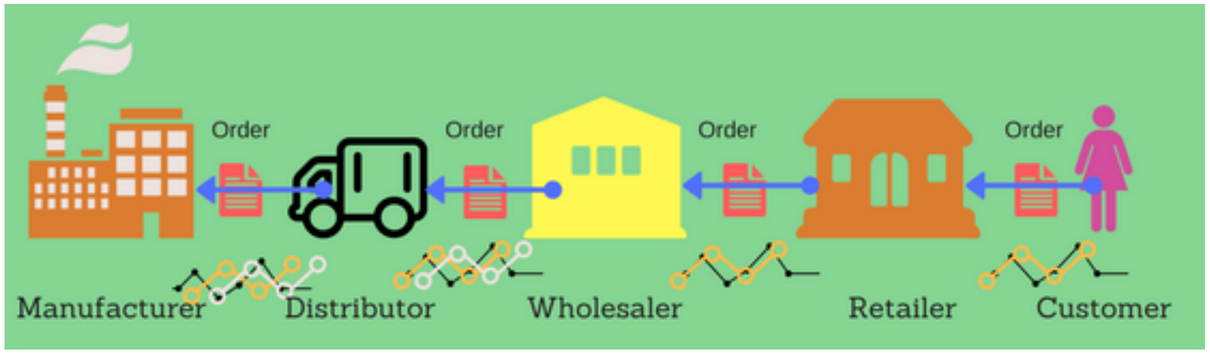
The process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data *(Fayyad, Piatetsky-Shapiro & Smith 1996)*

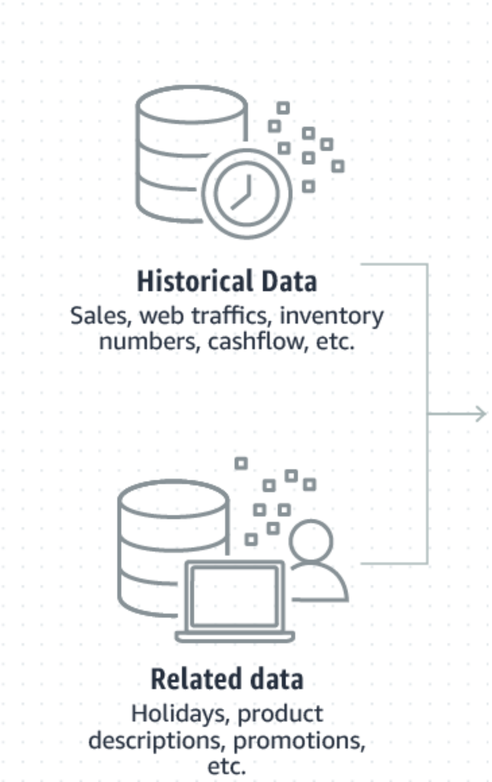


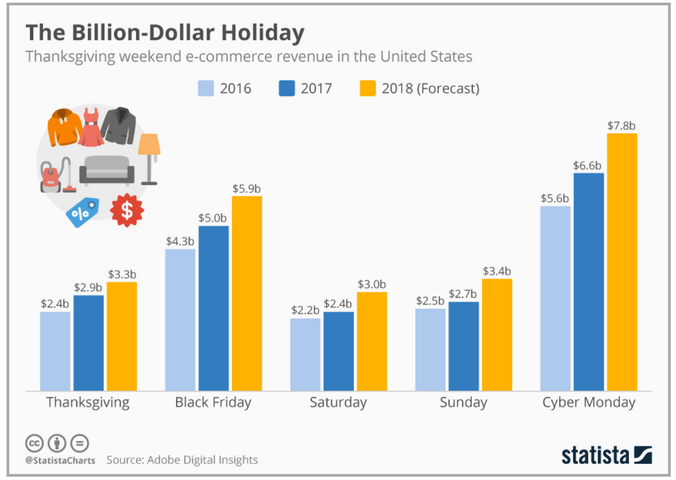
LET'S TALK ABOUT SHOPPING!

EXAMPLES

Sales and inventory forecast







EXAMPLES

Pregnant customer prediction

How Companies Learn Your Sec

By CHARLES DUHIGG
Published: February 16, 2012 | 570 Comments

And among life events, none are more important than the arrival of a baby. At that moment, new parents' habits are more flexible than at almost any other time in their adult lives. If companies can identify pregnant shoppers, they can earn millions.

sounds. Target has a baby-shower registry, and Pole started there, observing how shopping habits changed as a woman approached her due date, which women on the registry had willingly disclosed. He ran test after test, analyzing the data, and before long some useful patterns emerged. Lotions, for example. Lots of people buy lotion, but one of Pole's colleagues noticed that women on the baby registry were buying larger quantities of unscented lotion around the beginning of their second trimester.

As Pole's computers crawled through the data, he was able to identify about 25 products that, when analyzed together, allowed him to assign each shopper a "pregnancy prediction" score. More important, he could also estimate her due date to within a small window, so Target could send coupons timed to very specific stages of her pregnancy.

Soon after the new ad campaign began, Target's Mom and Baby sales exploded. The company doesn't break out figures for specific divisions, but between 2002 — when Pole was hired — and 2010, Target's revenues grew from $44 billion to $67 billion. In 2005, the company's president, Gregg Steinhafel, boasted to a room of investors about the company's "heightened focus on items and categories that appeal to specific guest segments such as mom and baby."

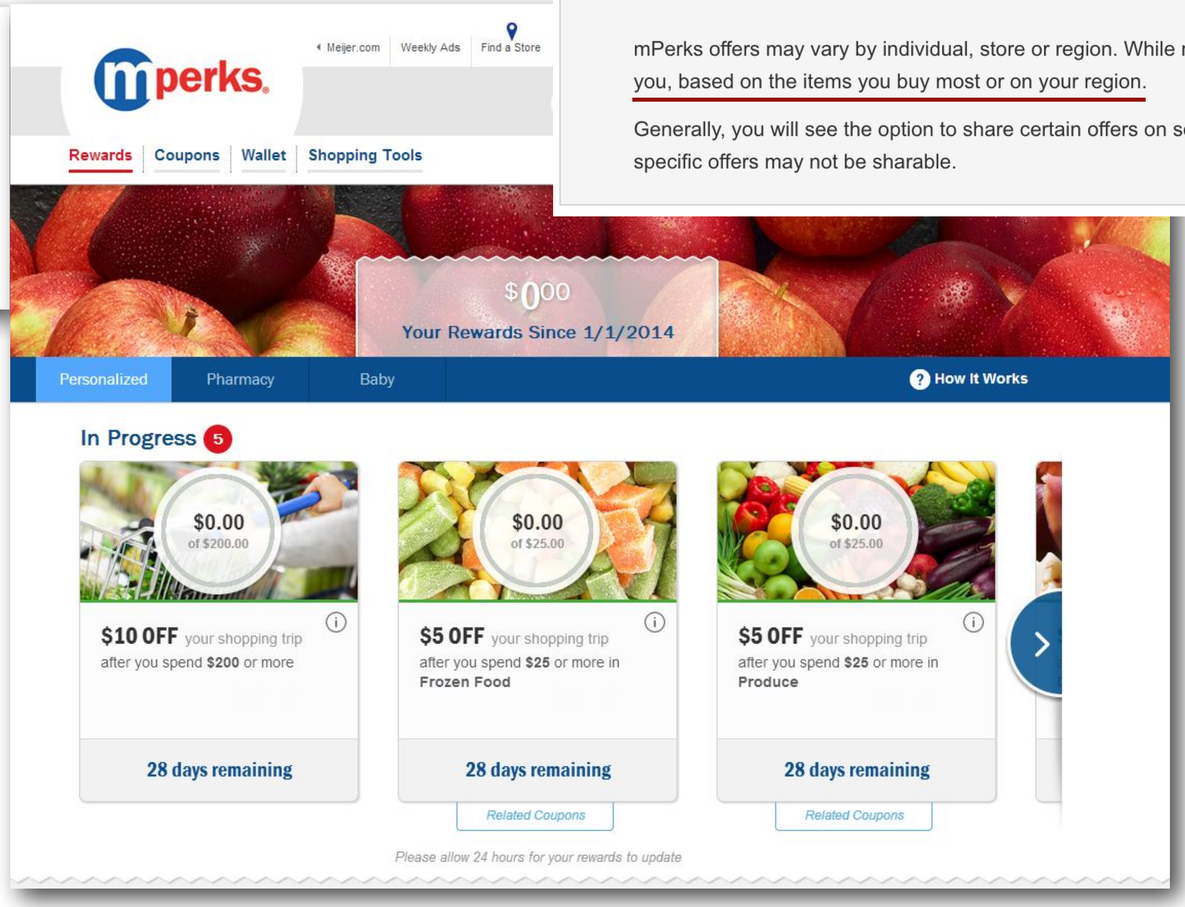
EXAMPLES

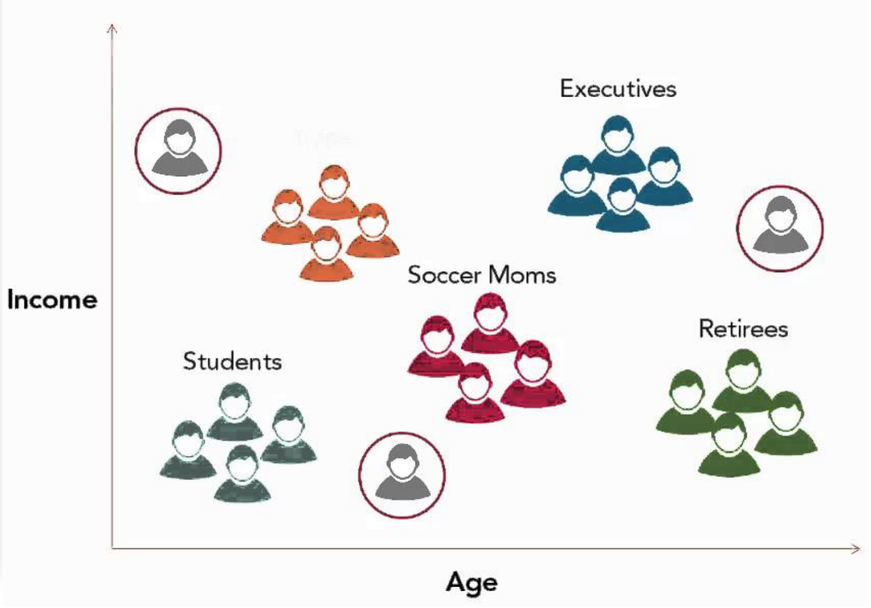
Customer segmentation

Why do I see different mPerks offers and coupons than my friends?

mPerks offers may vary by individual, store or region. While many members may see similar coupons, some coupons are personalized just for you, based on the items you buy most or on your region.

Generally, you will see the option to share certain offers on social media networks such as Twitter and Facebook, whereas individual or regionally specific offers may not be sharable.



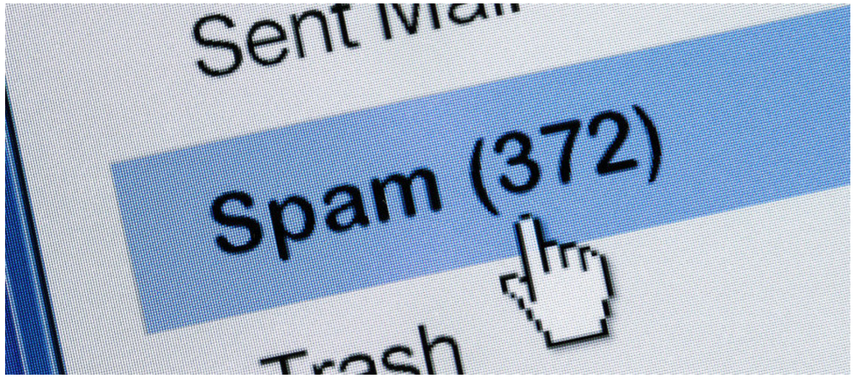


EXAMPLES

Purchase pattern mining

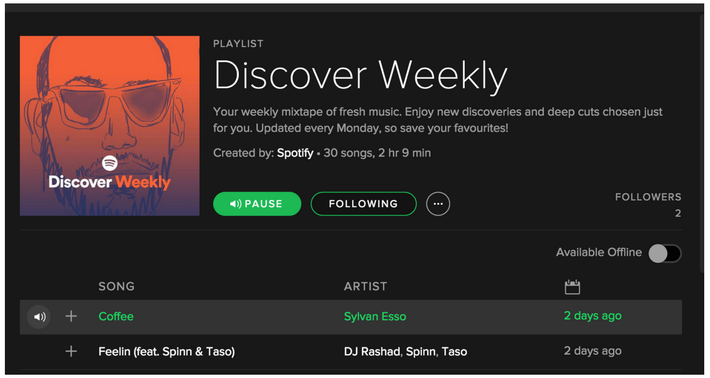
- A retail chain put all its checkout-counter data into a giant database and found a unexpected correlation: sales of diapers and beer
- It appears that young fathers would make a late-night run to the store to pick up Pampers and get some Bud Light while they were there.
- The store placed the disparate items together. Sales zoomed.

NOT JUST FOR MARKETING…



Spam email filter

Churn prediction

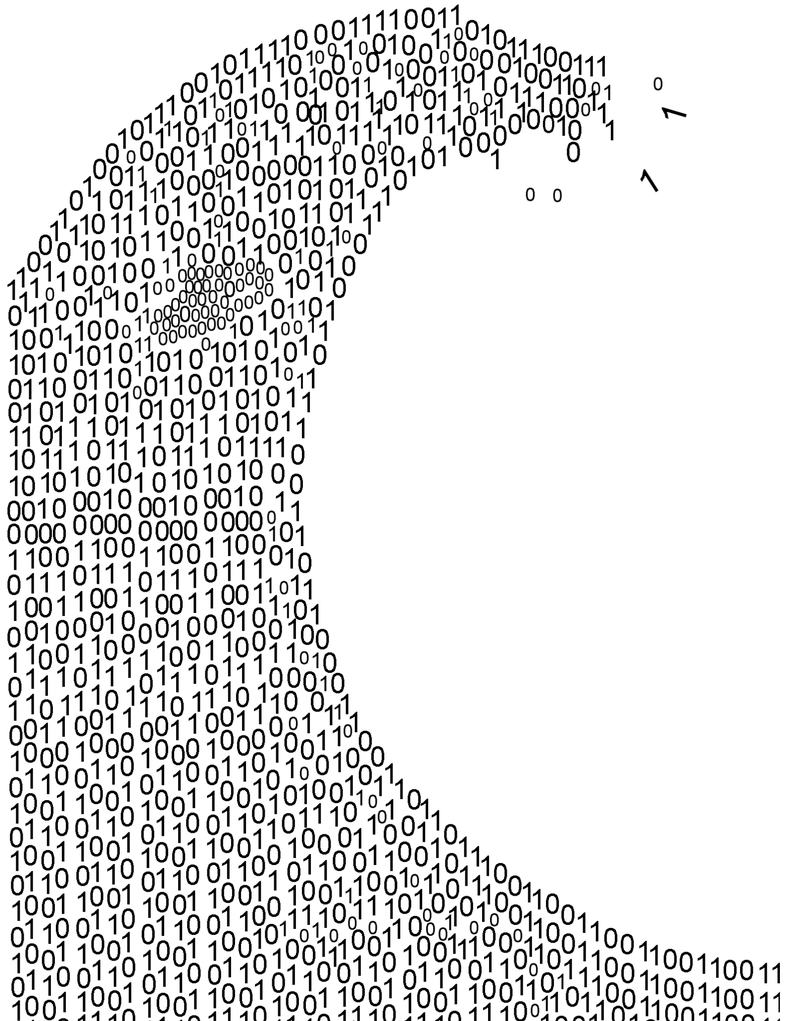


Music recommendation

Loan default risk assessment

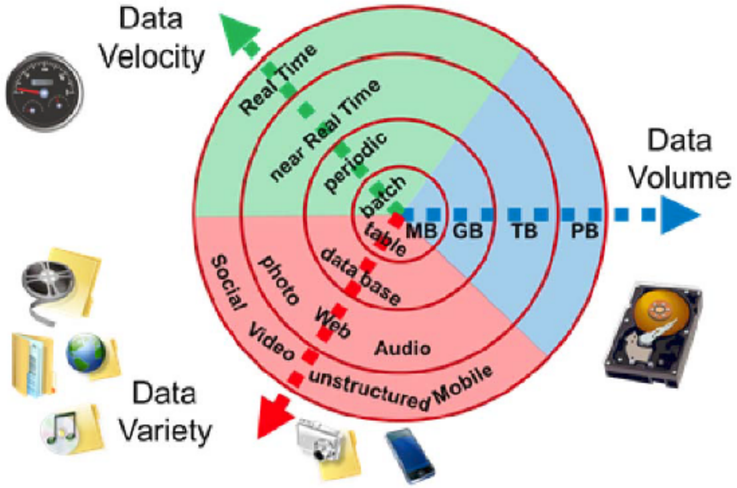
Personalized medicine

WHY NOW?

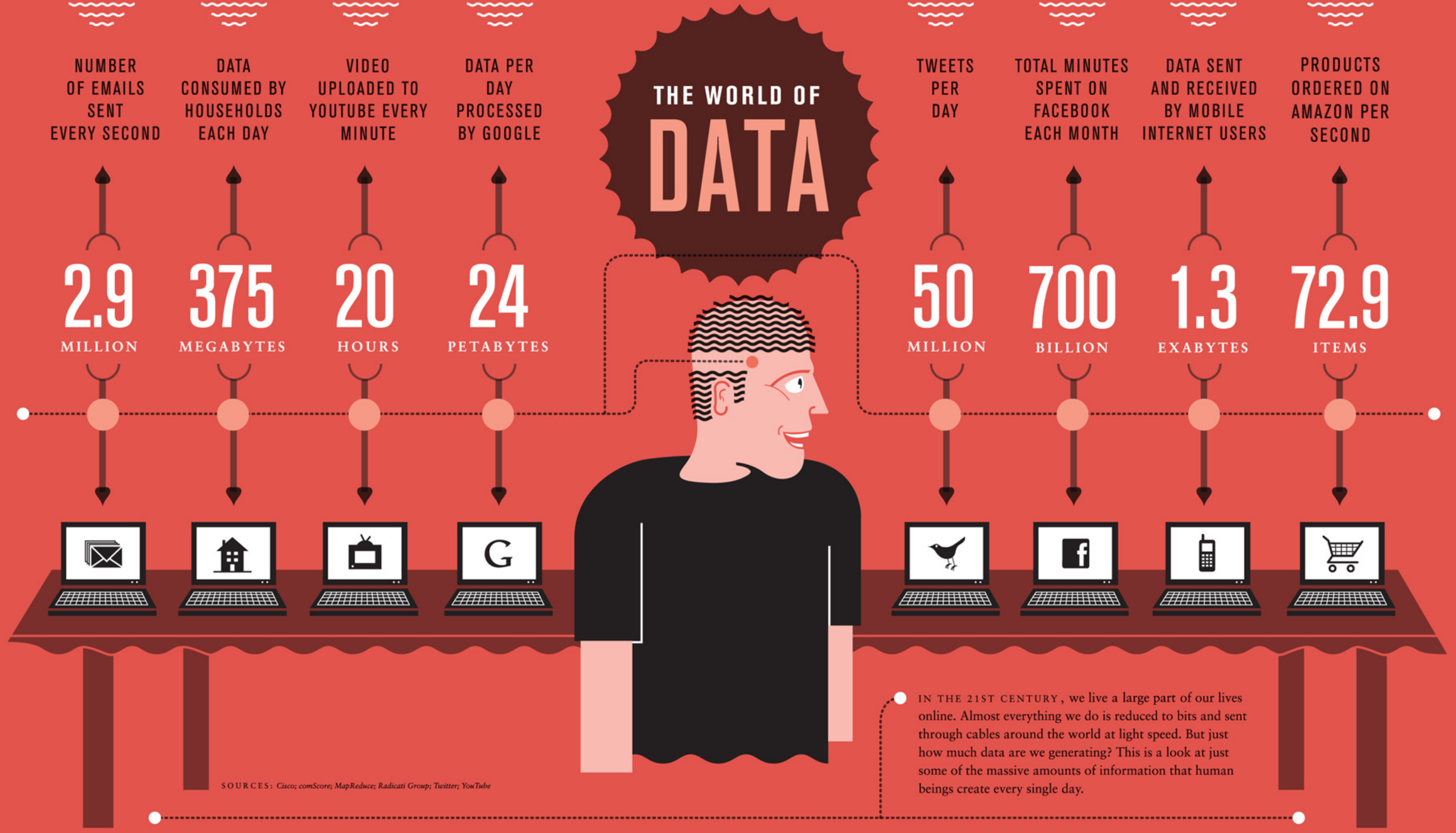


We are entering the **BIG DATA** age!

- Data volume: huge amount of data
- Data velocity: high speed at which the data is generated and processed
- Data variety: diverse type of the data



THE WORLD OF DATA



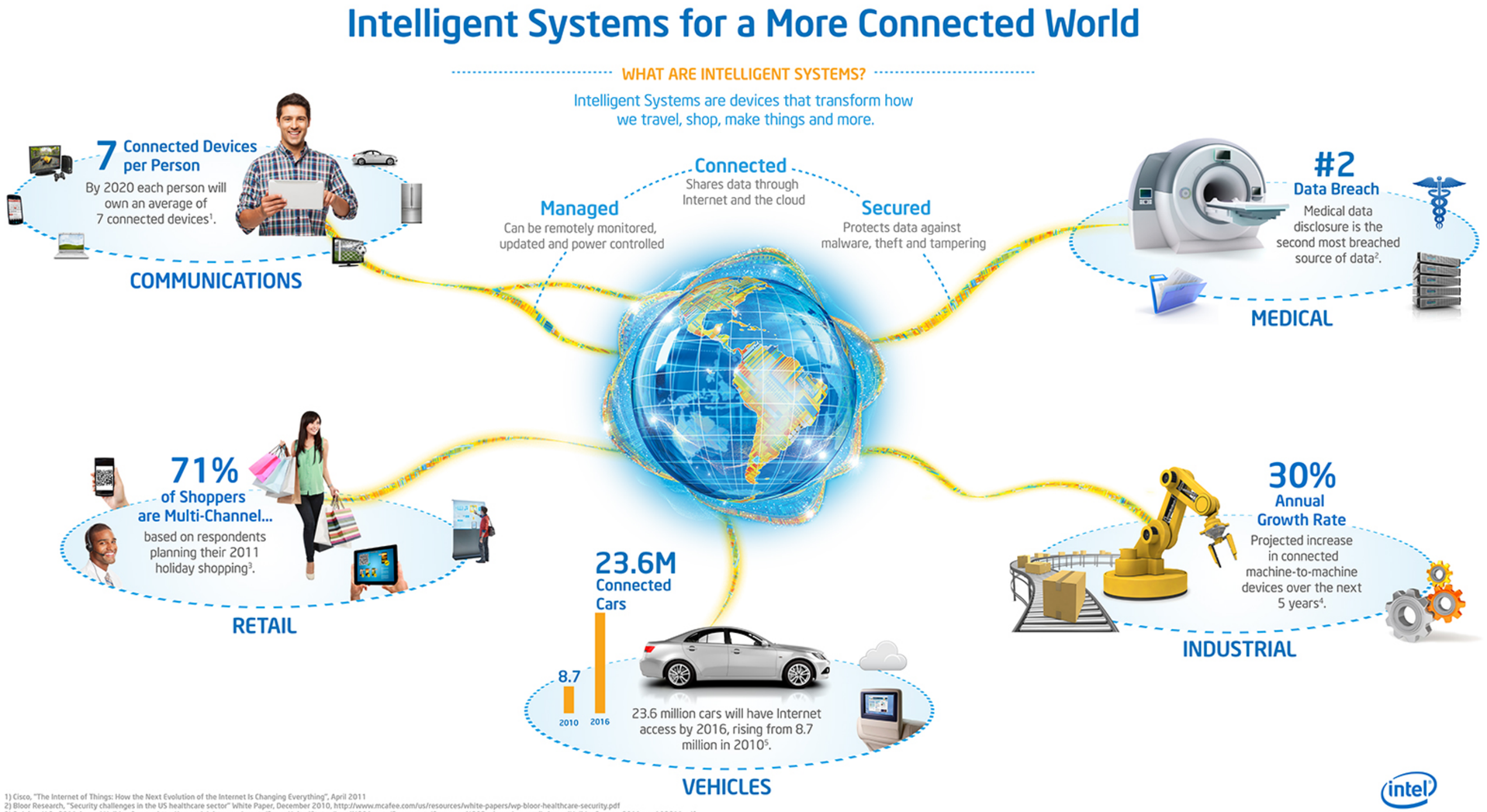
SOURCES: *Cisco; comScore; MapReduce; Radicati Group; Twitter; YouTube*

IN THE 21ST CENTURY, we live a large part of our lives online. Almost everything we do is reduced to bits and sent through cables around the world at light speed. But just how much data are we generating? This is a look at just some of the massive amounts of information that human beings create every single day.

A COLLABORATION BETWEEN GOOD AND OLIVER MUNDAY

IN PARTNERSHIP WITH IBM

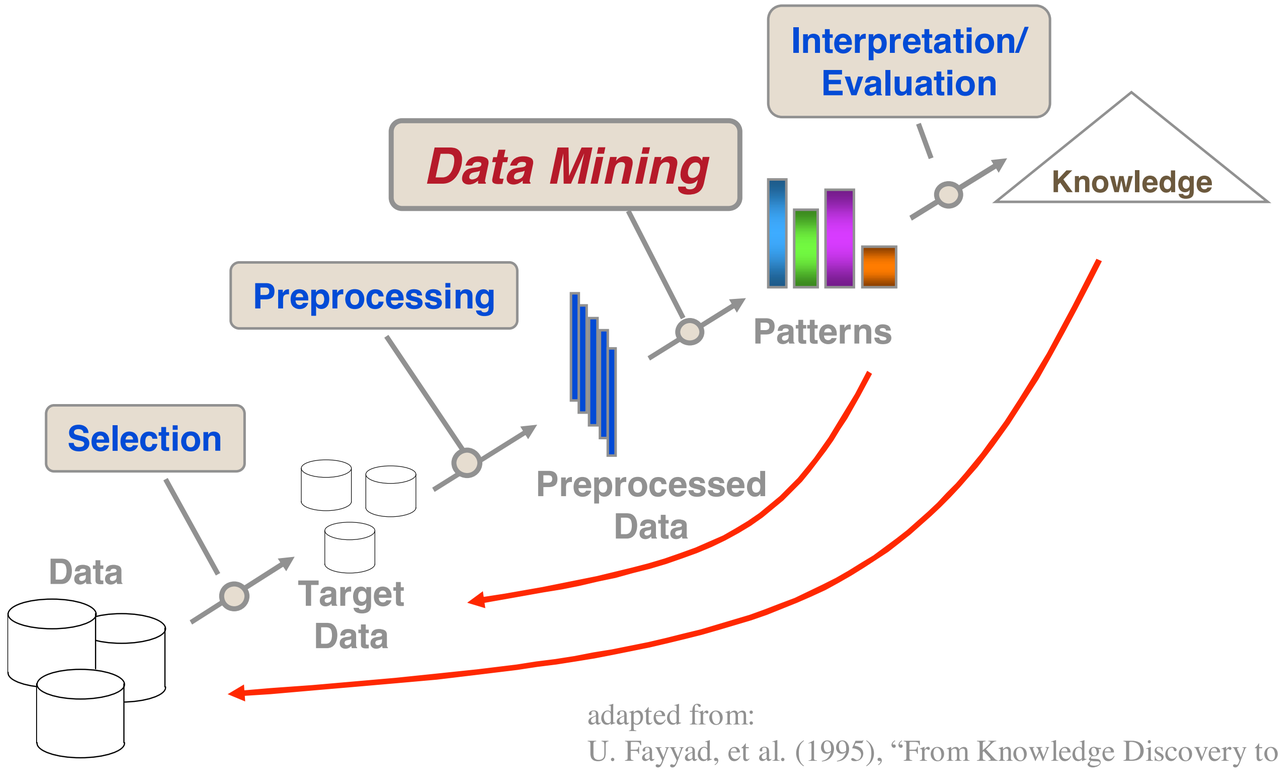
Data is everywhere



1) Cisco, "The Internet of Things: How the Next Evolution of the Internet Is Changing Everything", April 2011
2) Bloor Research, "Security challenges in the US healthcare sector" White Paper, December 2010, http://www.mcafee.com/us/resources/white-papers/wp-bloor-healthcare-security.pdf
3) Deloitte U.S., 2011 Annual Holiday Survey, http://www.deloitte.com/assets/Dcom-UnitedStates/Local%20Assets/Documents/Consumer%20Business/us_retail_AnnualHolidaySurvey_2011_pr_102611.pdf
4) McKinsey Global Institute analysis, "Big data: The next frontier for innovation, competition, and productivity", June 2011)
5) Wall Street Journal, http://online.wsj.com/article/SB10001424052702304066504576349763614933844.html, estimate from research firm, Frost & Sullivan

intel Intelligent Systems

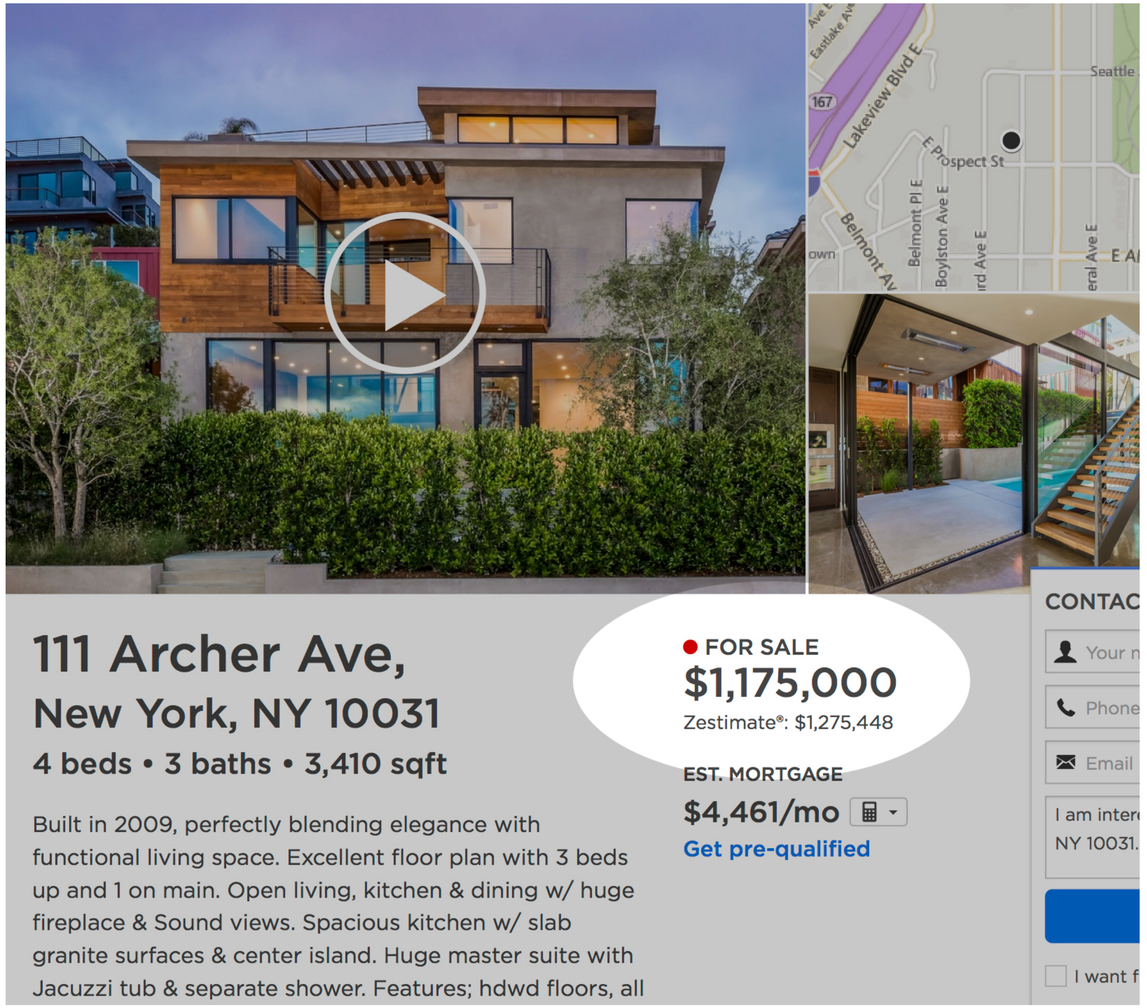
DATA MINING PROCESS



adapted from:
U. Fayyad, et al. (1995), "From Knowledge Discovery to Data Mining: An Overview," Advances in Knowledge Discovery and Data Mining, U. Fayyad et al. (Eds.), AAAI/MIT Press

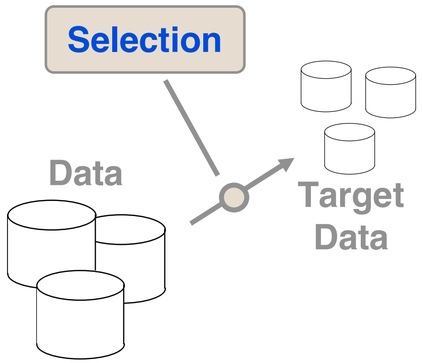
DATA MINING PROCESS IN THE CONTEXT OF HOUSE PRICE PREDICTION

Zillow® Zestimates



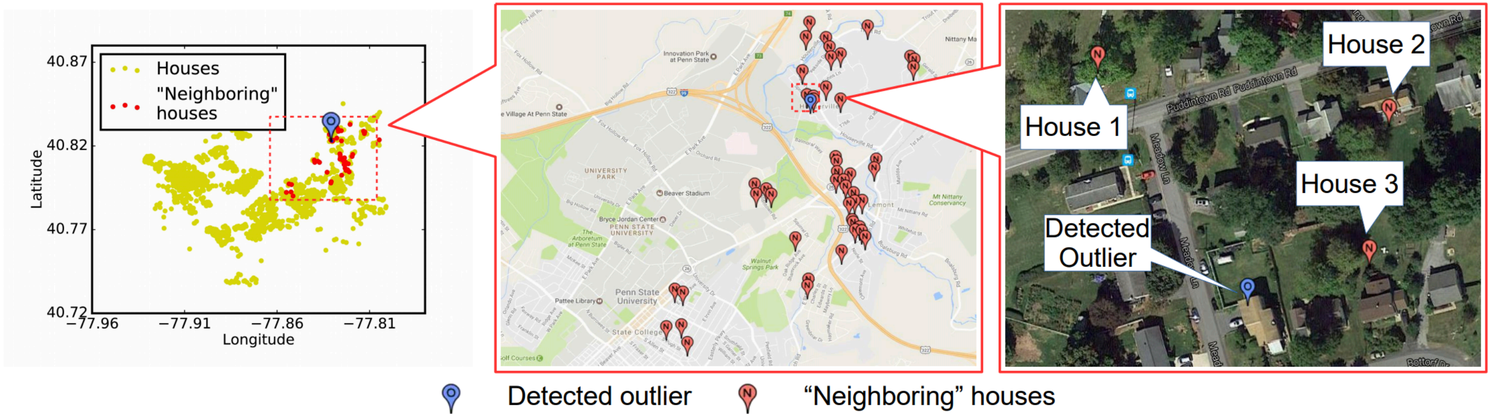
DATA SELECTION

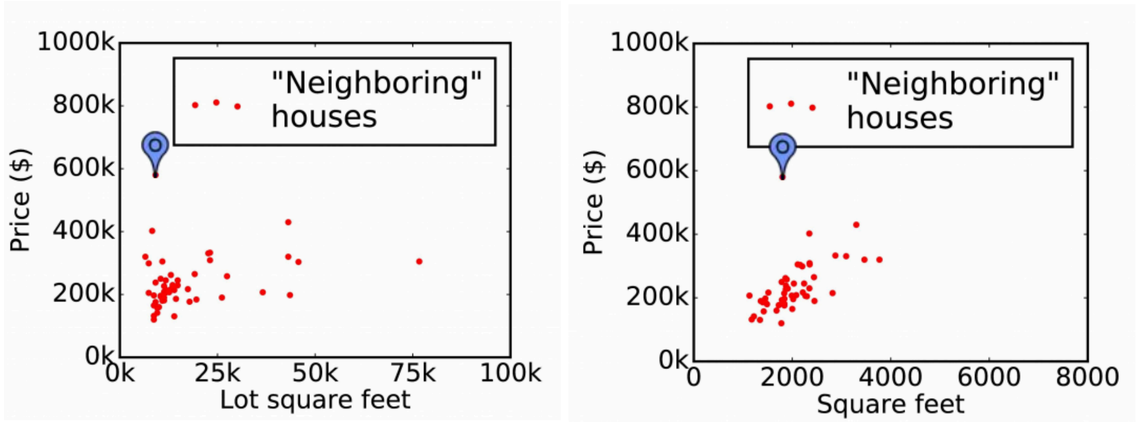
- Choose data sources
  - Seller-generated data on house property: lot size, square footage, number of bedrooms, etc.
  - Public records: property tax information, historic sale prices, recent sales of nearby homes, etc.
- Identify relevant attributes
- Sample data
  - Up until March 2013, Zillow has data on about 100 million U.S. homes…

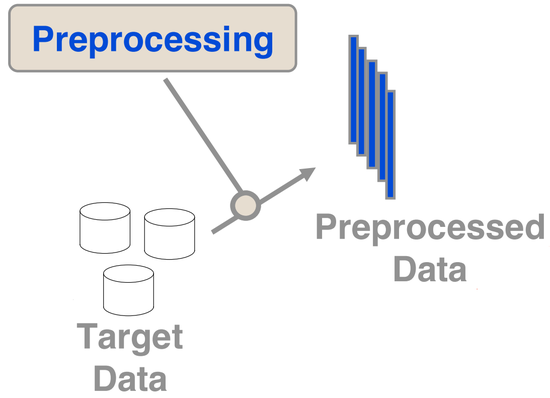


DATA PREPROCESSING

- Remove noise or outliers
- Handle missing values
- Account for time or other changes
  - Historic house sale prices are likely on different scales compared to their prices today
  - Attach nearby home information (measured by geographical distance) to each house



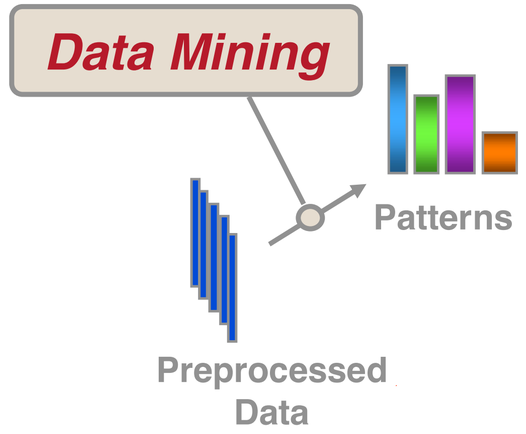




Picture credit: Zheng et al. Contextual Spatial Outlier Detection with Metric Learning, KDD 2017

DATA MINING

- Identify task
  - Predictive modeling: regression
- Choose models for learning and inference
  - For example, linear regression $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$
- Learn model parameters
- Apply the learned model



INTERPRETATION / EVALUATION

- Assess accuracy of model/results
- Interpret model for end-users
- Consolidate knowledge

Data Coverage and Zestimate Accuracy Table

Choose a location type below to change data:

Top Metro Areas
States/Counties*
National

	Zestimate Accuracy	Homes on Zillow	Homes With Zestimates	Within 5% of Sale Price	Within 10% of Sale Price	Within 20% of Sale Price	Median Error
Atlanta, GA	★★★	2.1M	1.9M	55.6%	76.2%	88.5%	4.2%
Baltimore, MD	★★★★	992.6K	915.3K	59.8%	78.6%	88.3%	3.6%
Boston, MA	★★★	1.6M	1.6M	53.3%	78.5%	91.2%	4.6%
Charlotte, NC	★★★	973.3K	881.1K	57.2%	76.9%	87.8%	4.0%
Chicago, IL	★★★	3.3M	3.0M	58.6%	77.9%	89.9%	3.8%
Cincinnati, OH	★★★	828.9K	728.0K	52.0%	72.5%	85.0%	4.7%
Cleveland, OH	★★	830.6K	745.8K	46.9%	68.6%	82.8%	5.5%
Dallas-Fort Worth, TX	★★	2.2M	2.1M	32.4%	56.6%	80.1%	8.4%
Denver, CO	★★★★	965.1K	918.5K	64.8%	87.3%	95.4%	3.5%
Detroit, MI	★★★	1.8M	1.7M	52.7%	73.6%	86.4%	4.6%

Last updated: December 12, 2018

Zillow

Press Room Home
Press Releases
Company Info
In the News
Awards & Recognition
Images and B-Roll
Media Contacts
Zillow Research
Zillow Group

CONTACT US

Press Releases

Homes Near Trader Joe's, Whole Foods Stores Appreciate Faster

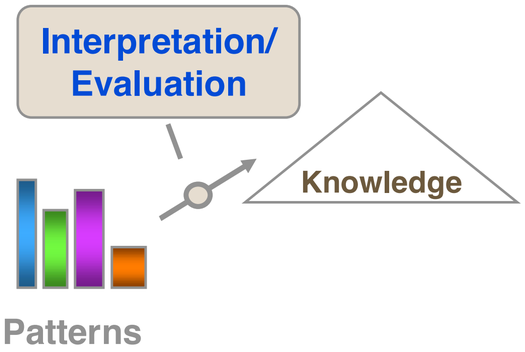
Homes within a mile of either high-end grocery store begin appreciating faster than other homes after the stores open, Zillow has discovered

Jan 25, 2016

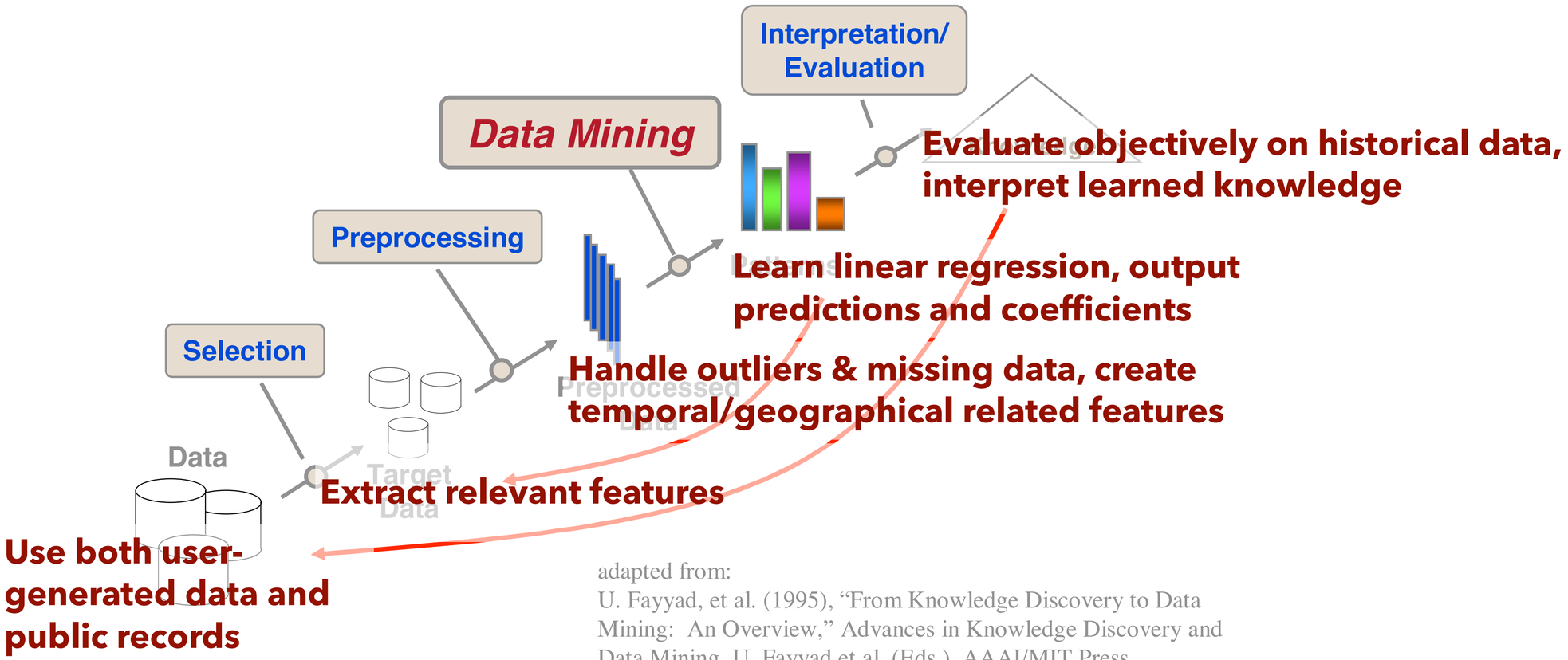
SEATTLE, Jan. 25, 2016 /PRNewswire/ -- Your local grocery market has a lot to do with what happens in your local housing market, according to a new analysis by Zillow featured in the paperback edition of *Zillow Talk: Rewriting the Rules of Real Estate* (Grand Central Publishing, Jan. 26).

Specifically, Zillow found that homes grow more rapidly in value if they are closer to a Trader Joe's or Whole Foods[i]. Between 1997 and 2014, homes near the two grocery chains were consistently worth more than the median U.S. home. By the end of 2014, homes within a mile of either store were worth more than twice as much as the median home in the rest of the country.

"Like Starbucks, the stores have become an amenity in their own right – a signal to the home-buying public that the neighborhood they're located in is desirable, perhaps up-and-coming, and definitely improving," said Zillow Group Chief Economist Stan Humphries. "Like a self-fulfilling prophecy, the stores may actually drive home prices. Even if they open in neighborhoods where home prices have lagged those in the wider city, they start to outperform the city overall once the stores arrive."



DATA MINING PROCESS: RECAP



adapted from:
U. Fayyad, et al. (1995), "From Knowledge Discovery to Data Mining: An Overview," Advances in Knowledge Discovery and Data Mining, U. Fayyad et al. (Eds.), AAAI/MIT Press

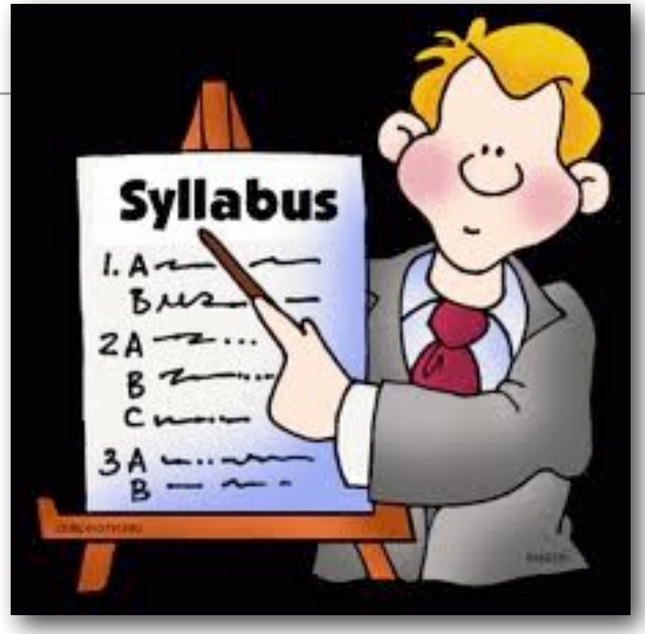
COURSE OVERVIEW

COURSE GOALS

- Identify key elements of data mining systems and the knowledge discovery process
- Understand how algorithmic elements interact
- Recognize various types of data mining tasks
- Familiarity with standard models/algorithms
- Implement and apply basic algorithms
- Understand how to evaluate performance

TOPICS

- Statistical basics and background
- Elements of data mining algorithms
- Data preparation, exploration, and visualization
- Predictive modeling
- Methodology, evaluation, theory
- Descriptive modeling
- Pattern mining

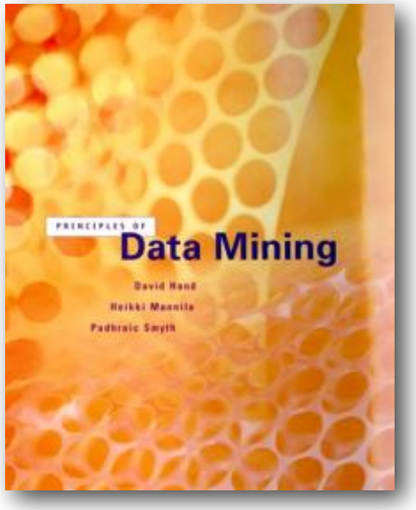


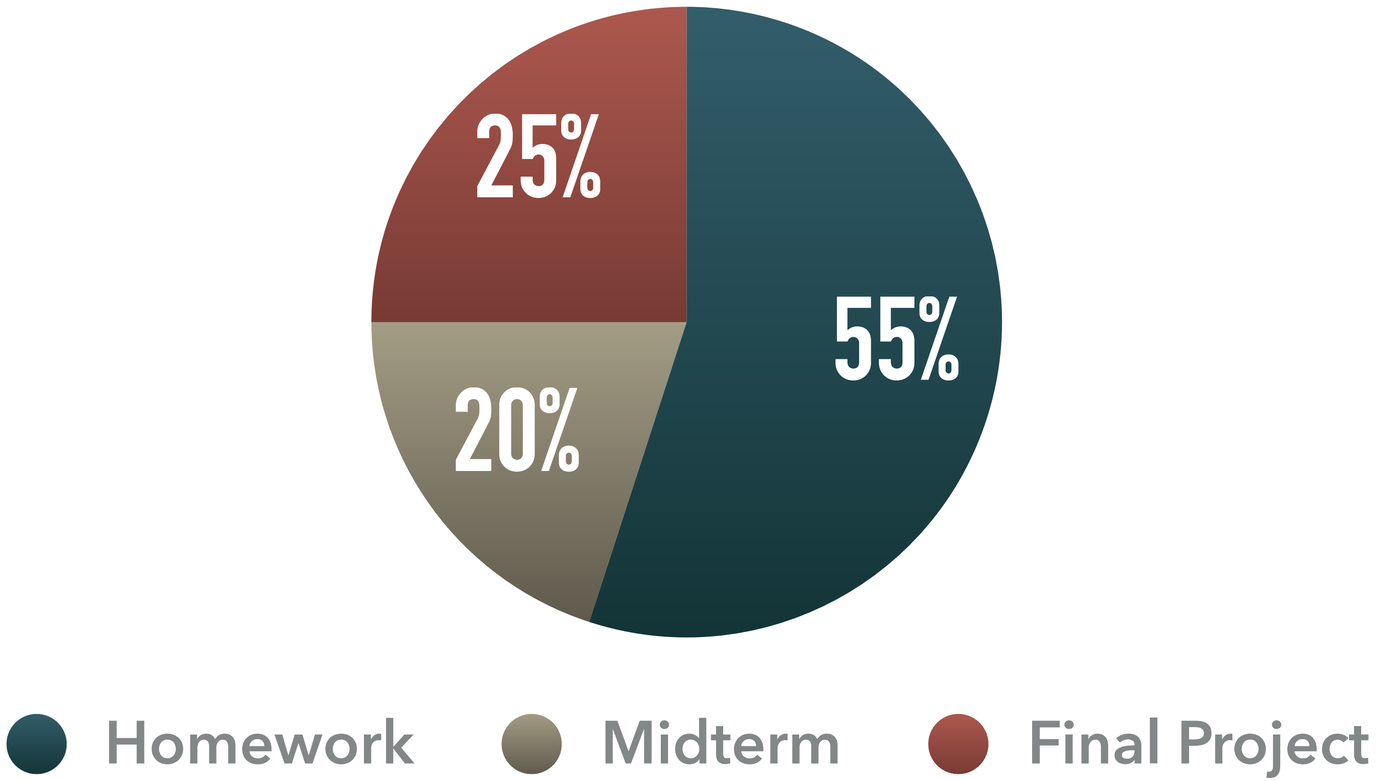
LOGISTICS

- Time and location: TTh 4:30-5:45, WANG 2599
- Instructor: Ming Yin
  mingyin@purdue.edu, LWSN 2142B, office hours: Wednesdays 4-5pm
- Teaching assistants: Hao Ding, Mahak Goindani, Omkar S. Patil
  office hours: TBD
- Webpage: http://mingyin.org/CS573/Spring2019/index.html
- Piazza: https://piazza.com/purdue/spring2019/cs57300/home
  - Please sign up to Piazza as soon as possible!
  - All communication will be on Piazza. Emails will be SLOW at best and unresponsive at worst
- Prerequisites: introductory statistics course (e.g., STAT 516), adequate programming skills (e.g., CS381, STAT598G)

READINGS

- *Principles of Data Mining*, Hand, Mannila, and Smyth, MIT Press, 2001.
- Other useful references:
  - C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006




25%
55%
20%
Homework
Midterm
Final Project

WORKLOAD

- Assignments
  - Five assignments including written/math exercises and programming assignments
- Exams
  - One in-class midterm
- Final Project
  - Open topics
  - 1 final project pitch, 1 final project presentation, 1 final report

MORE ON FINAL PROJECT

- Apply learned techniques and algorithms to real-life data mining tasks
- Identify your own topics, tasks and datasets
- Complete in teams of 2-4 students
- Start to think about your project **NOW**!
  - Can consider data mining challenges on competition platforms like Kaggle

LATE POLICY

- **Three *extension days*** can be applied on any chosen assignment ***except for Assignment 1*** (scores not affected)
  - Each chosen assignment can get more than one extension day
  - You must explicitly state in your assignment submission the number of extension days used
  - Cannot be rearranged after they are applied
  - Cannot be used on any project-related due dates
- Beyond extension days, you get 10% off per day late, up to a max of 5 days late
  - Assignments submitted 5 days late or more will not be accepted

NEXT CLASS

- Review basic knowledge on probability