

CS57300
PURDUE UNIVERSITY
AUGUST 23, 2021

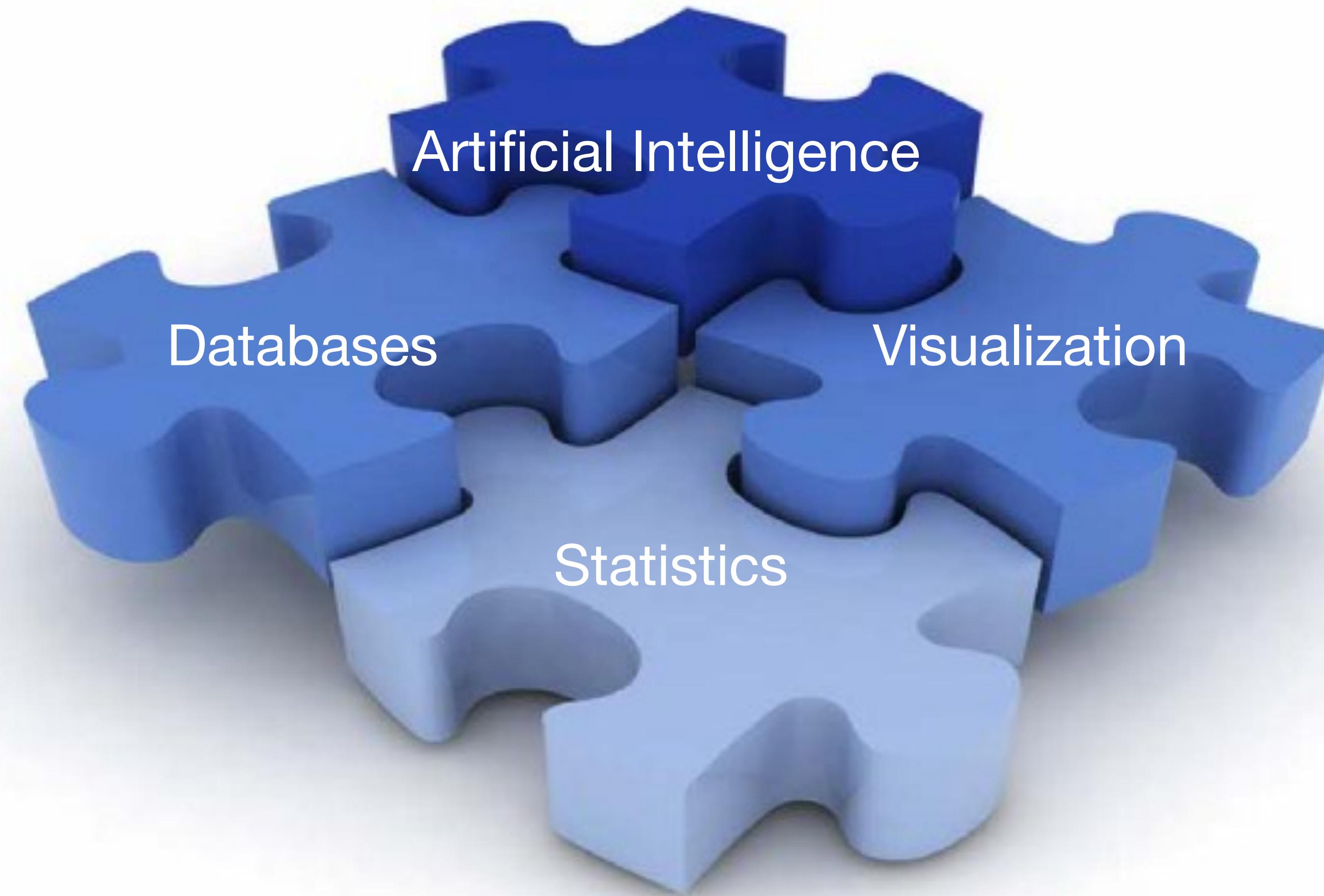
DATA MINING

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*)



COURSE OVERVIEW

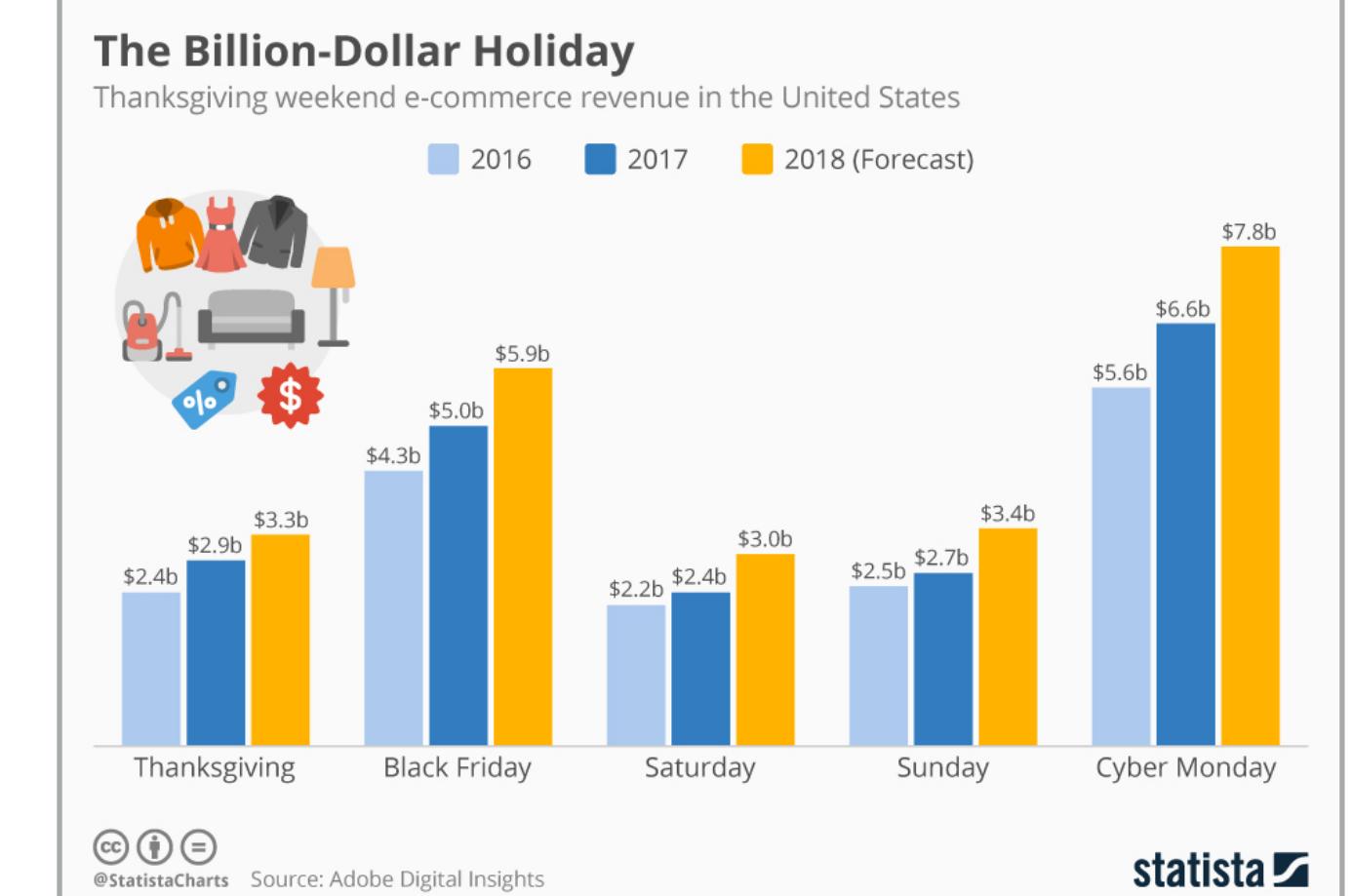
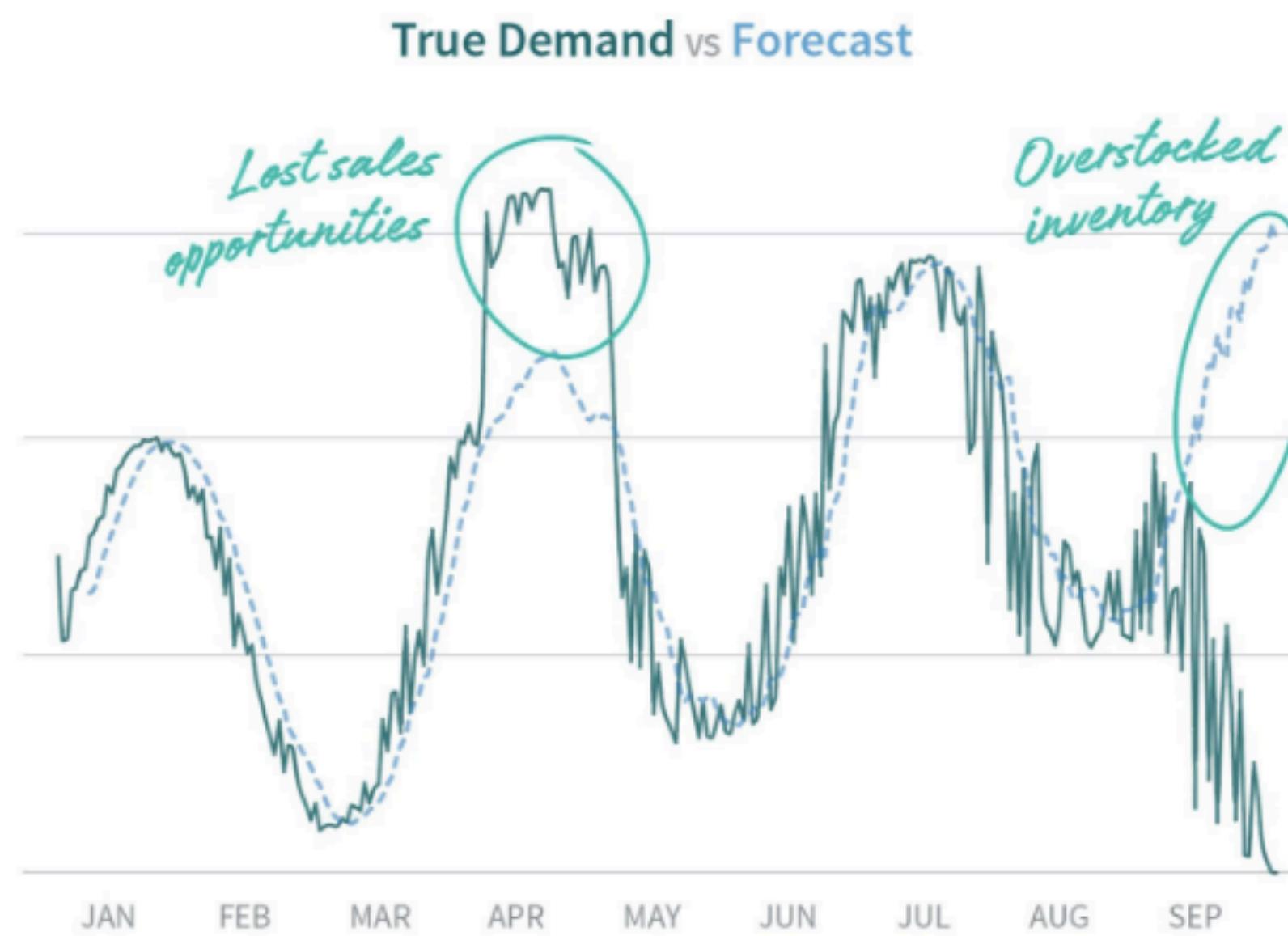
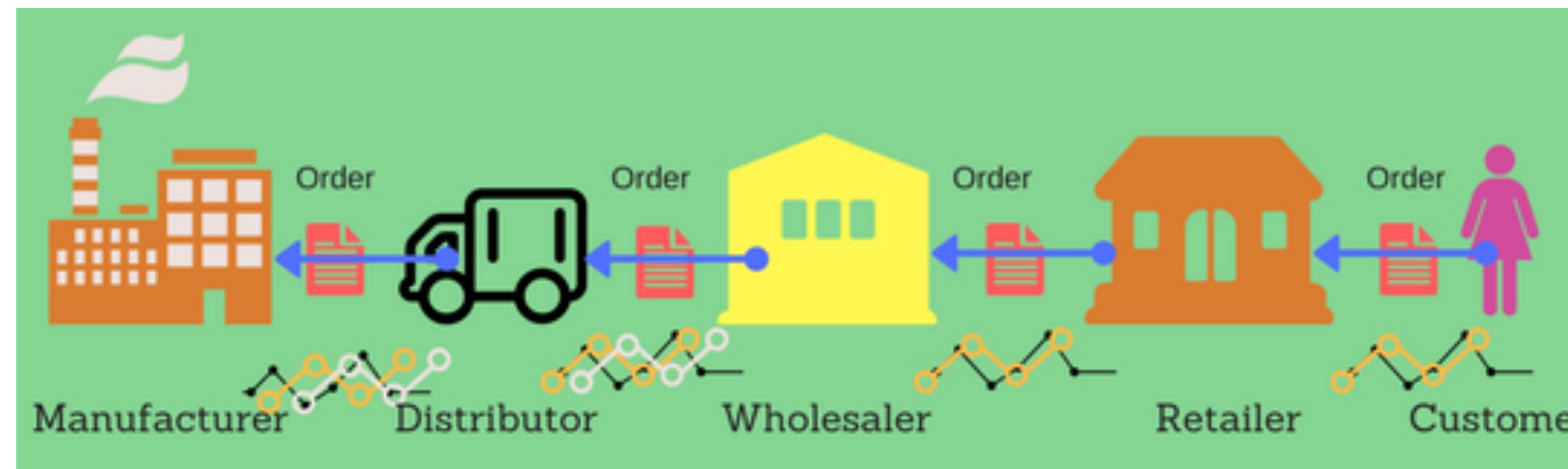
LET'S TALK ABOUT SHOPPING!



COURSE OVERVIEW

EXAMPLES

Sales and inventory forecast



EXAMPLES

Pregnant customer prediction



How Companies Learn Your Secrets



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.

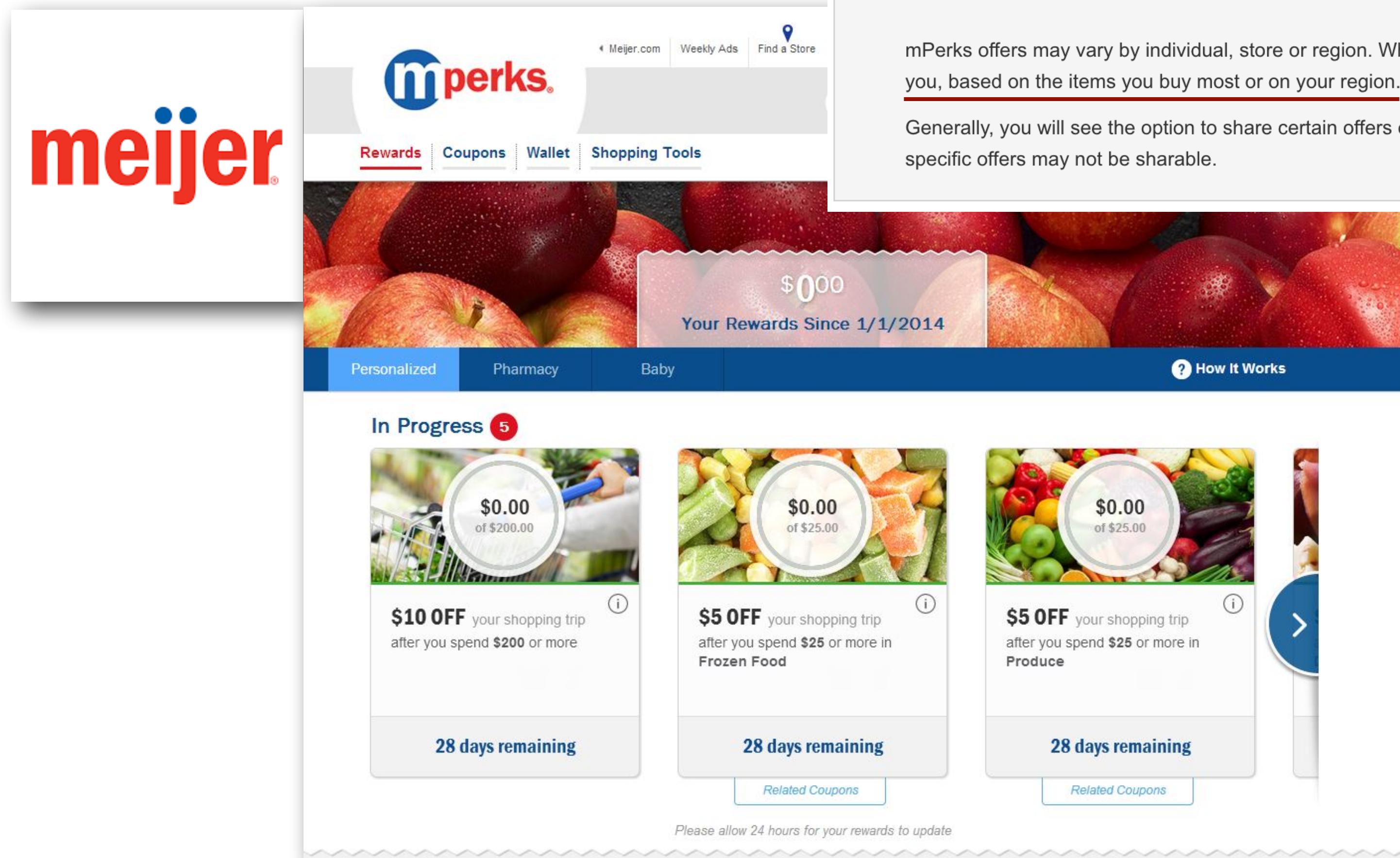
As Pole's computers crawled through the data, he was able to identify about 25 products that, when analyzed together, allowed him to assign each shopper a "pregnancy prediction" score. More important, he could also estimate her due date to within a small window, so Target could send coupons timed to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

COURSE OVERVIEW

EXAMPLES

Customer segmentation

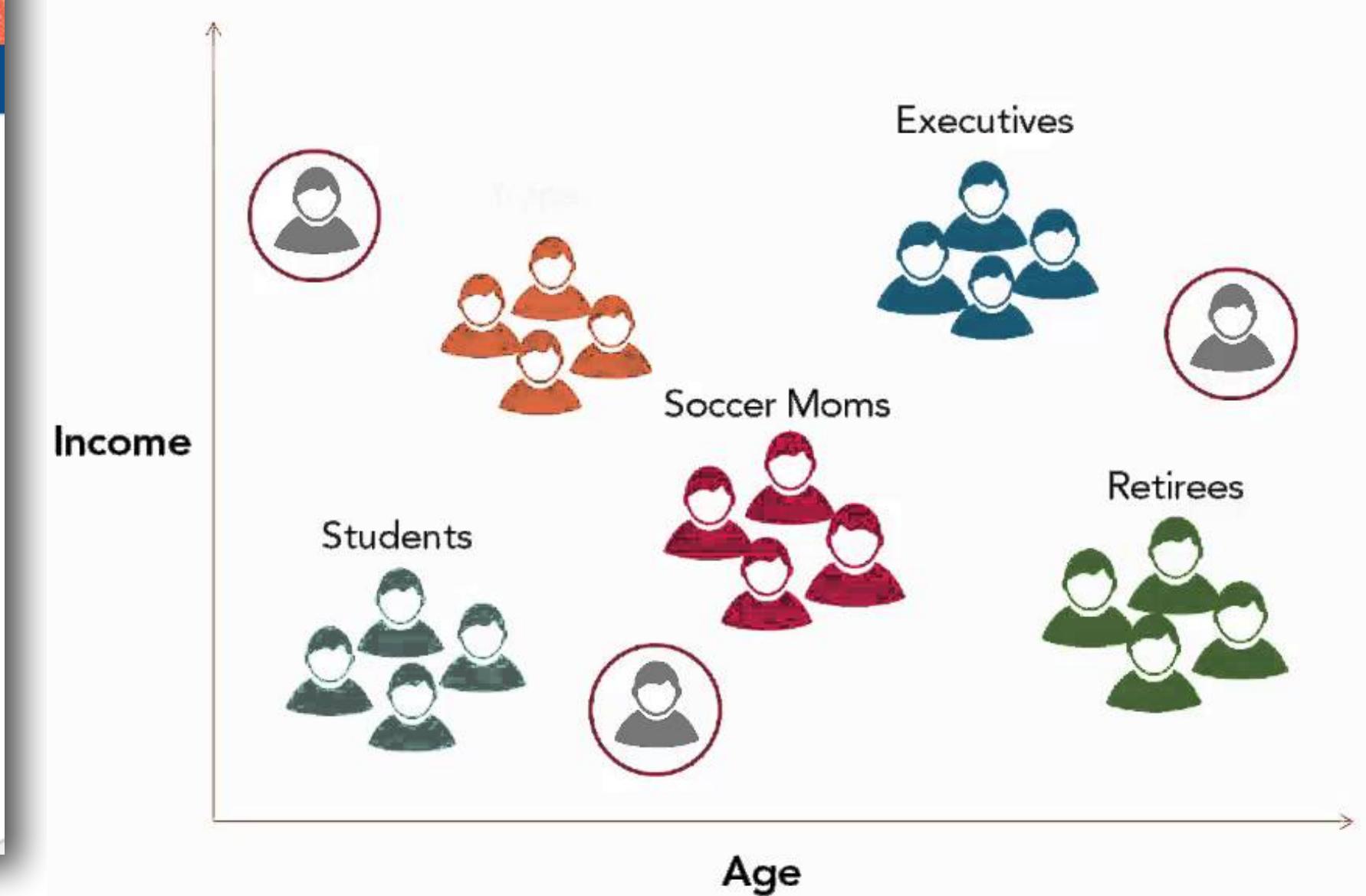


The image shows the Meijer mPerks rewards dashboard. At the top, there's a navigation bar with links to Meijer.com, Weekly Ads, Find a Store, mPerks logo, Rewards, Coupons, Wallet, and Shopping Tools. Below the navigation is a banner featuring a close-up of red apples with the text '\$0.00 Your Rewards Since 1/1/2014'. A blue header bar contains tabs for Personalized, Pharmacy, and Baby, along with a 'How It Works' link. The main content area is titled 'In Progress' and shows three offers: '\$10 OFF your shopping trip after you spend \$200 or more' (28 days remaining), '\$5 OFF your shopping trip after you spend \$25 or more in Frozen Food' (28 days remaining), and '\$5 OFF your shopping trip after you spend \$25 or more in Produce' (28 days remaining). Each offer includes a 'Related Coupons' button. At the bottom, a note says 'Please allow 24 hours for your rewards to update'.

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.

COURSE OVERVIEW

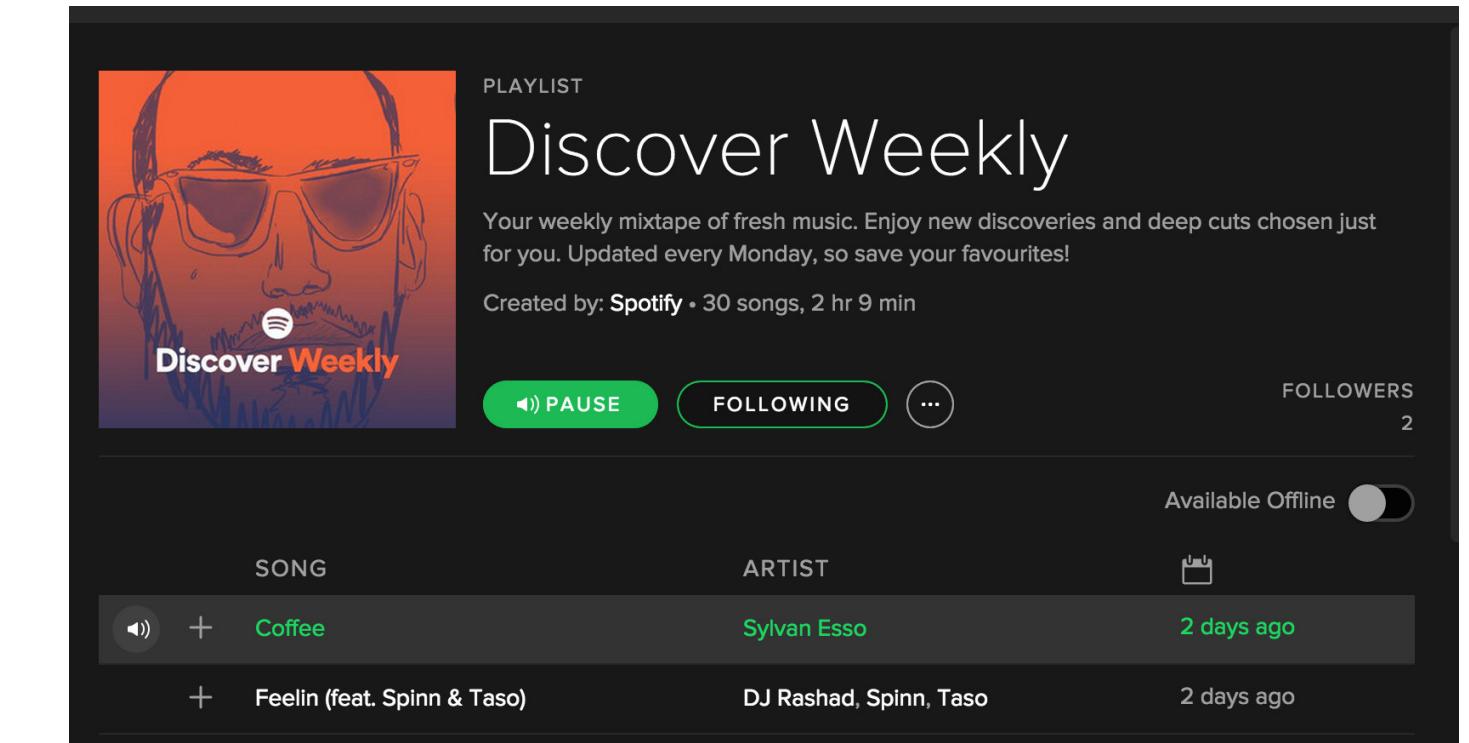
NOT JUST FOR MARKETING...



Spam email filter



Churn prediction



Music recommendation



Loan default risk assessment



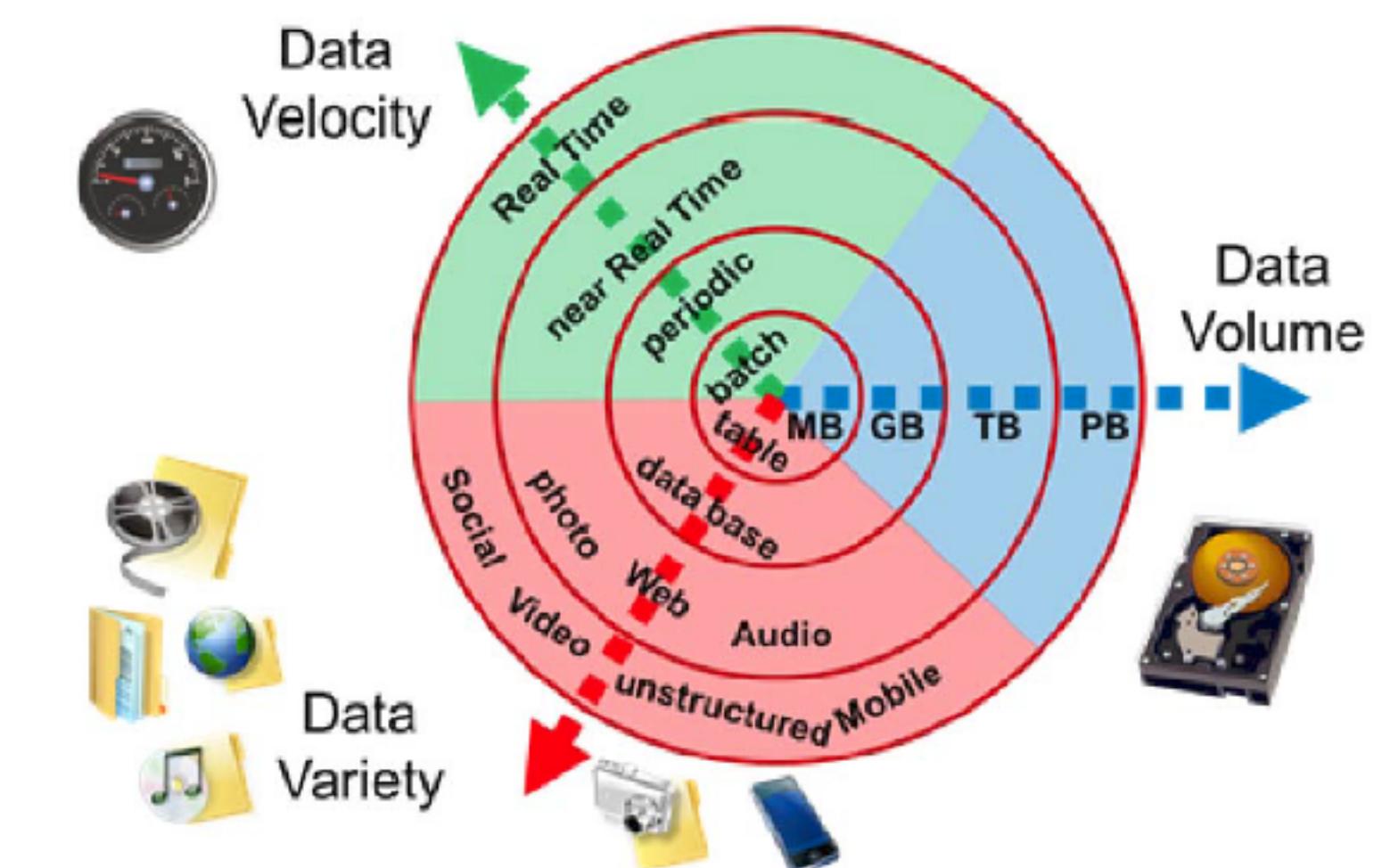
Personalized medicine

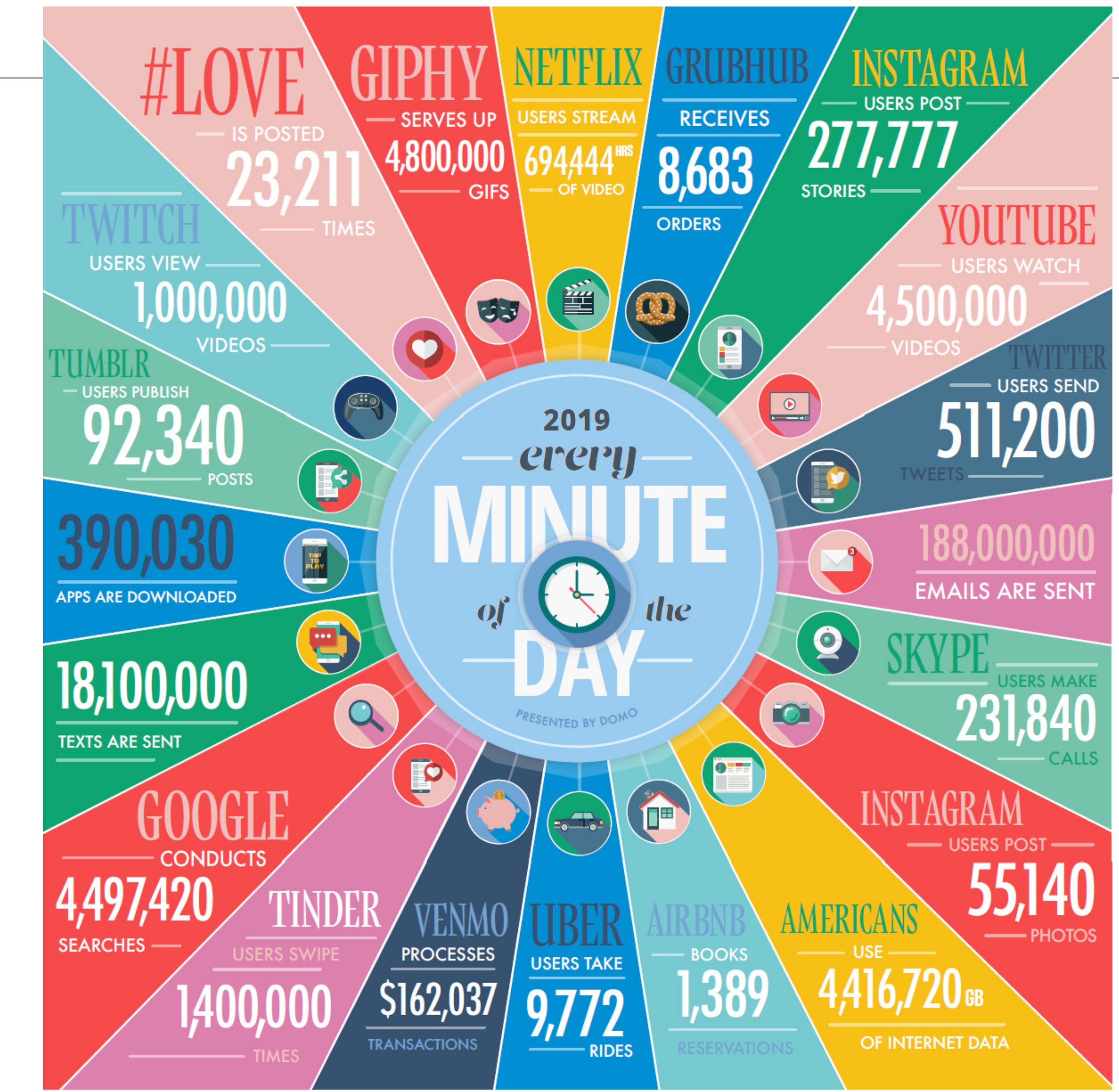
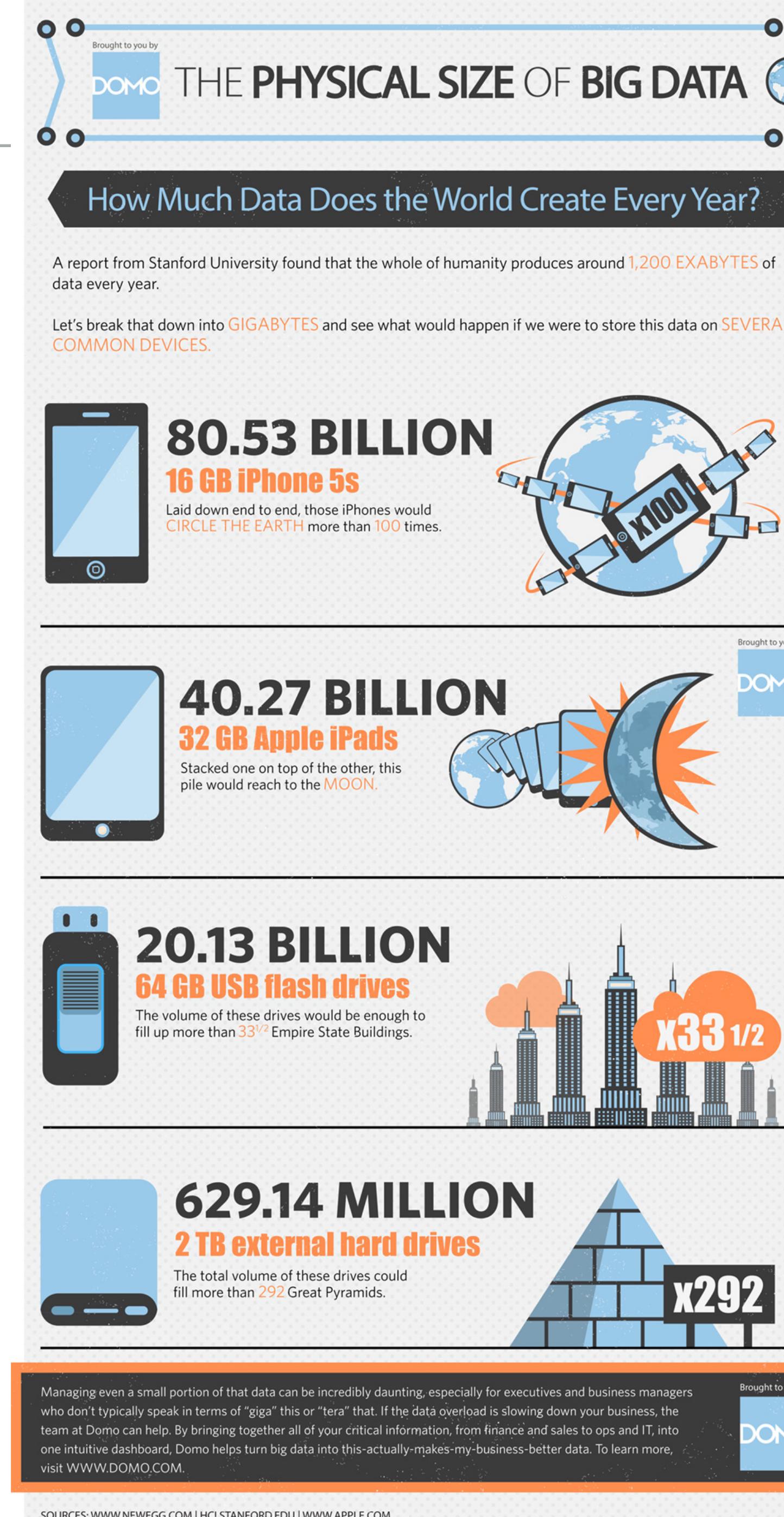
WHY NOW?



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

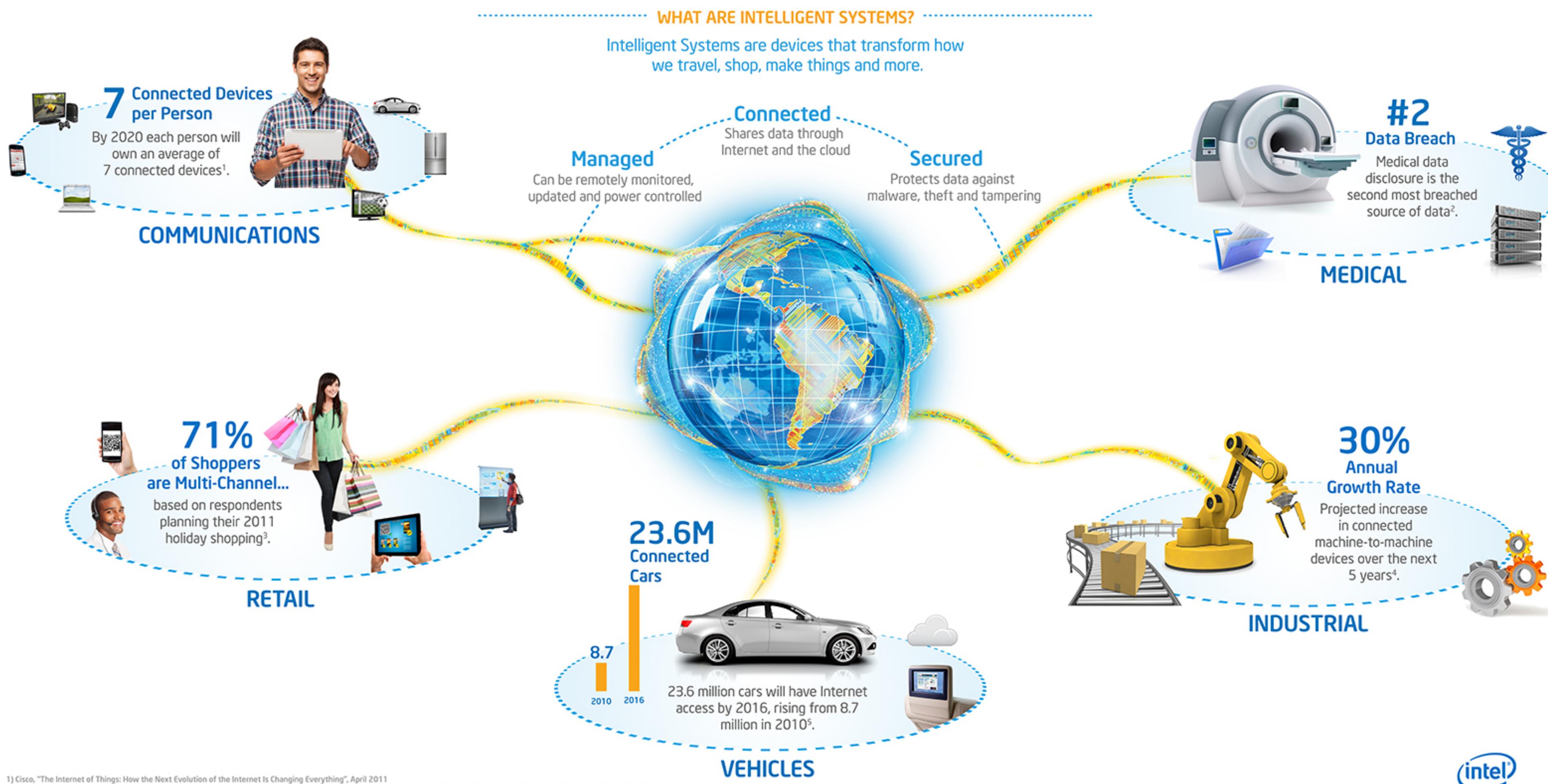
- ▶ Data volume: huge amount of data
- ▶ Data velocity: high speed at which the data is generated and processed
- ▶ Data variety: diverse type of the data





Data is everywhere

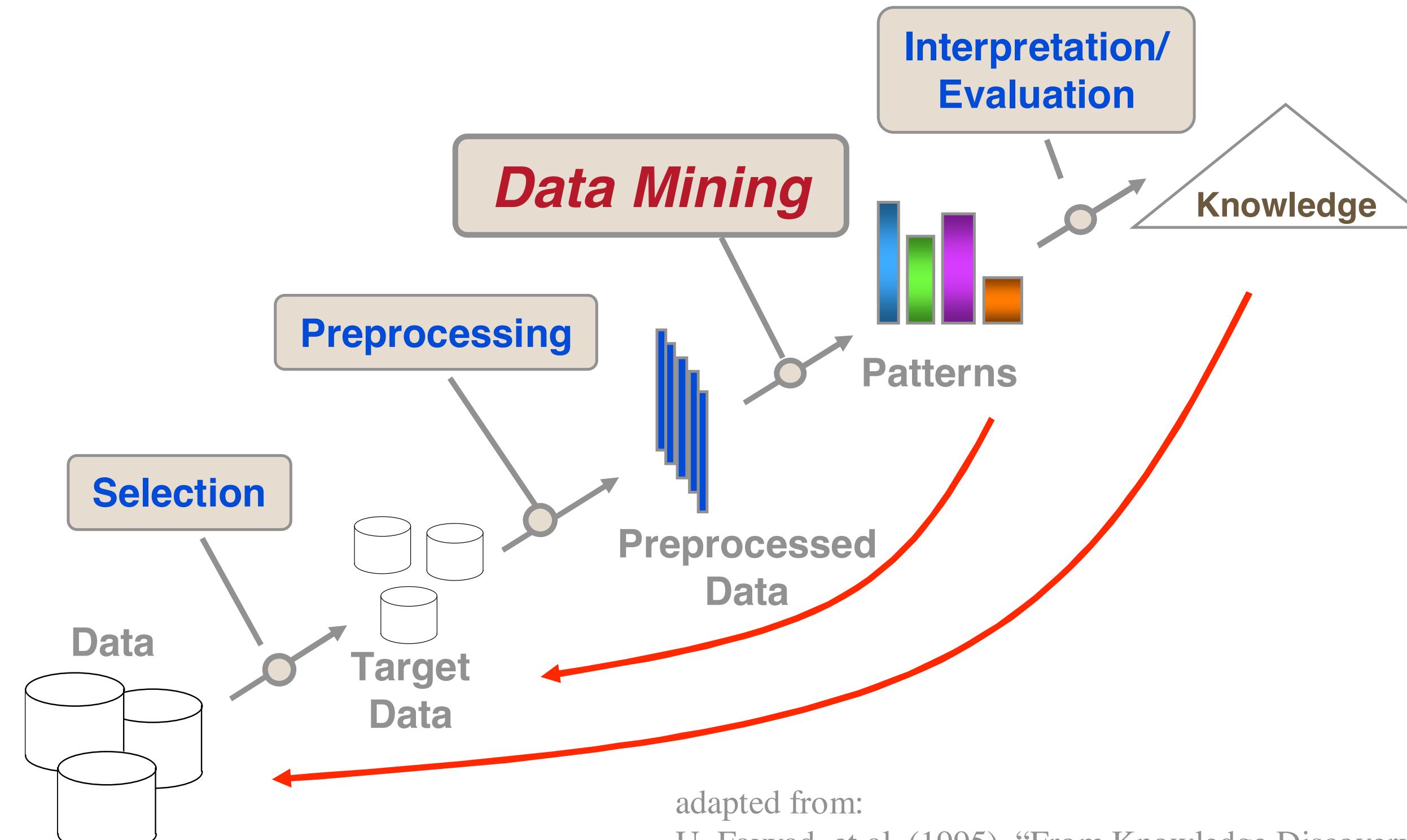
Intelligent Systems for a More Connected World



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

*2013 Intel Corporation. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries. Other names and brands may be claimed as the property of others.

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



The screenshot shows a real estate listing for a modern two-story house located at 111 Archer Ave, New York, NY 10031. The house features large glass windows, a wooden facade, and a balcony. A play button icon is overlaid on the image. To the right of the main image is a map showing the location's surroundings, including Lakeview Blvd E, Belmont Pl E, and Prospect St. Below the main image is a detailed description of the property: "111 Archer Ave, New York, NY 10031" with "4 beds • 3 baths • 3,410 sqft". A large white callout bubble contains the price information: "FOR SALE \$1,175,000" and "Zestimate®: \$1,275,448". Below the price, there is a section for "EST. MORTGAGE" showing "\$4,461/mo" with a calculator icon, and a link "Get pre-qualified". On the far right, there is a "CONTACT" form with fields for name, phone, email, and interest in the area, along with a "I want f" checkbox.

111 Archer Ave,
New York, NY 10031
4 beds • 3 baths • 3,410 sqft

FOR SALE
\$1,175,000
Zestimate®: \$1,275,448

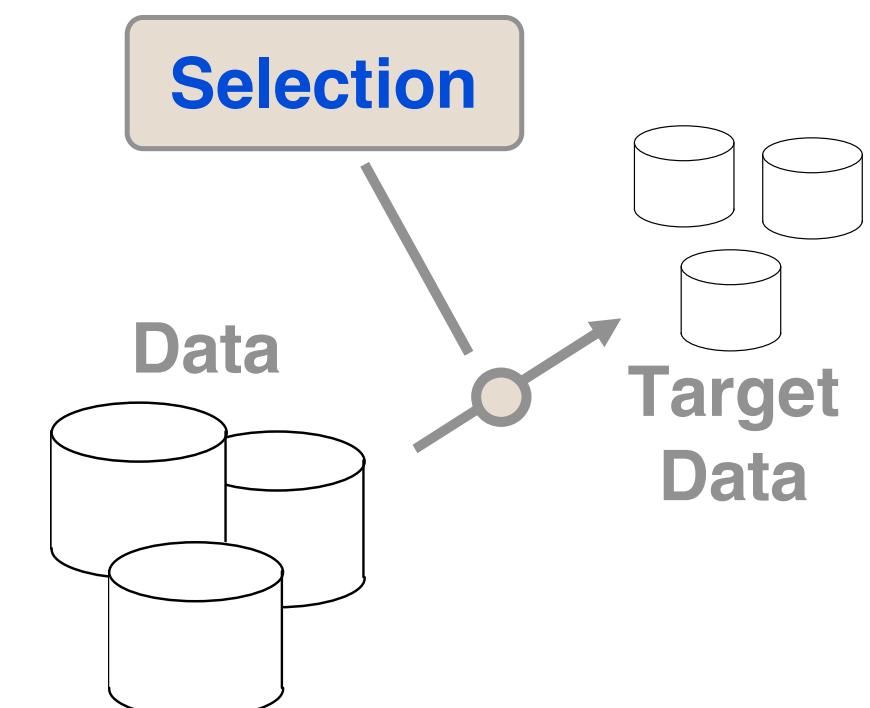
EST. MORTGAGE
\$4,461/mo
[Get pre-qualified](#)

CONTACT

Your n
Phone
Email
I am interested in this property in NY 10031.
 I want f

DATA SELECTION

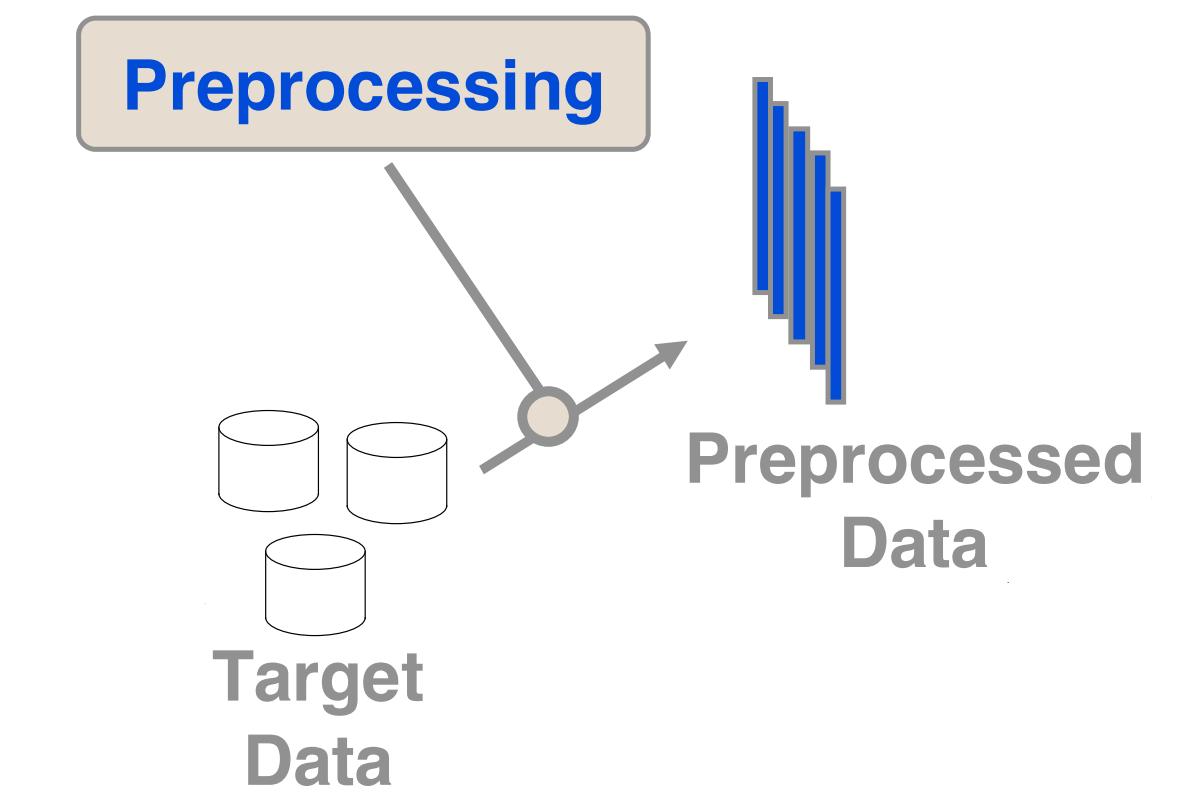
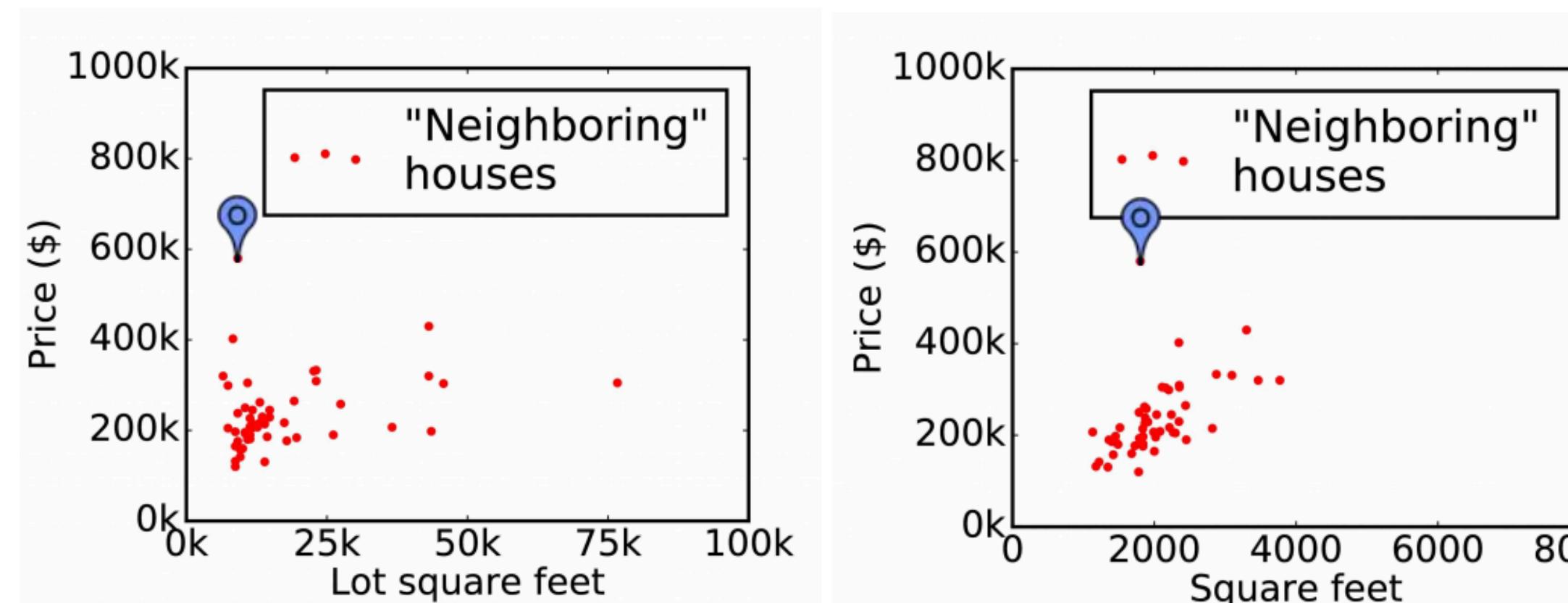
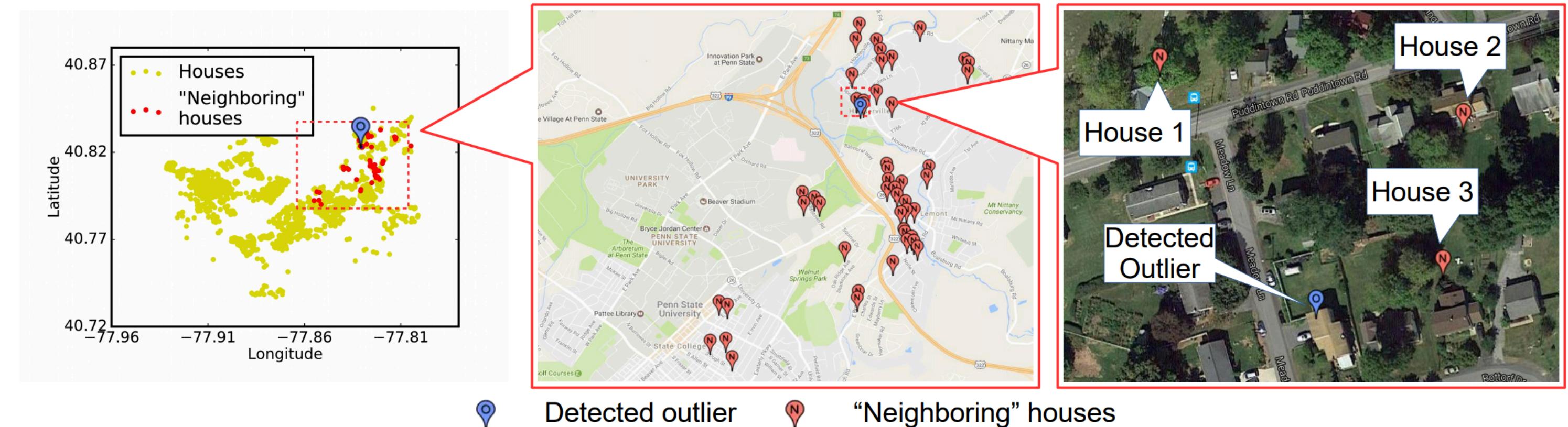
- ▶ Choose data sources
 - ▶ Seller-generated data on house property: lot size, square footage, number of bedrooms, etc.
 - ▶ Public records: property tax information, historic sale prices, recent sales of nearby homes, etc.
- ▶ Identify relevant attributes
- ▶ Sample data
 - ▶ Up until March 2013, Zillow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



COURSE OVERVIEW

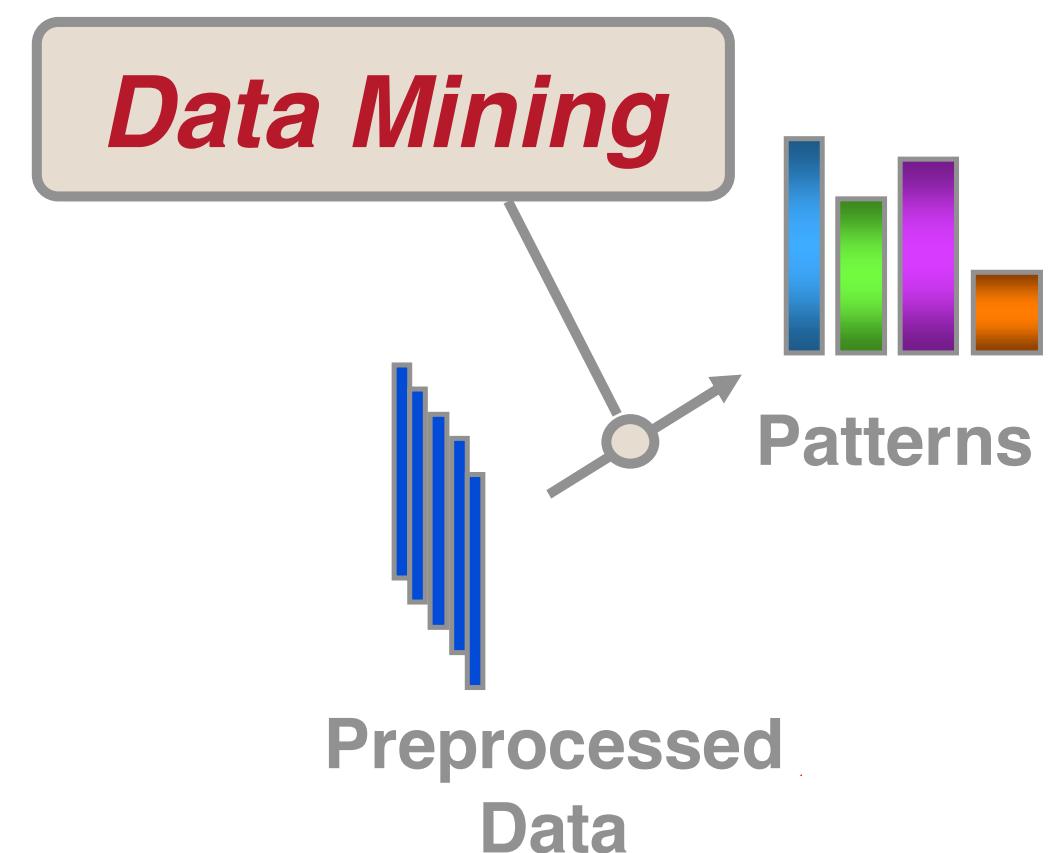
DATA PREPROCESSING

- ▶ 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



DATA MINING

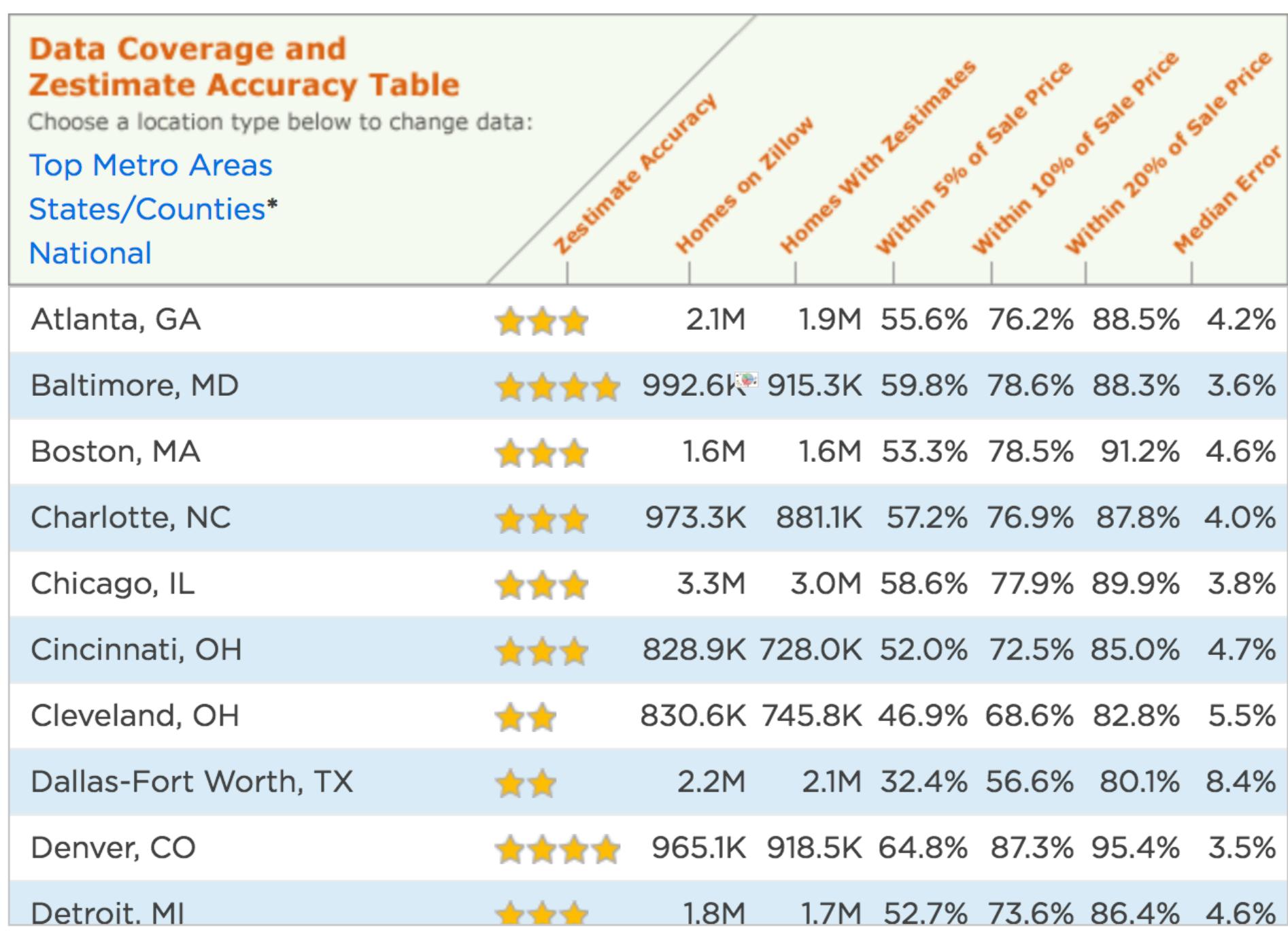
- ▶ Identify task
 - ▶ Predictive modeling: regression
- ▶ Choose models for learning and inference
 - ▶ For example, linear regression $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \epsilon$
- ▶ Learn model parameters
- ▶ Apply the learned model



COURSE OVERVIEW

INTERPRETATION / EVALUATION

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



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

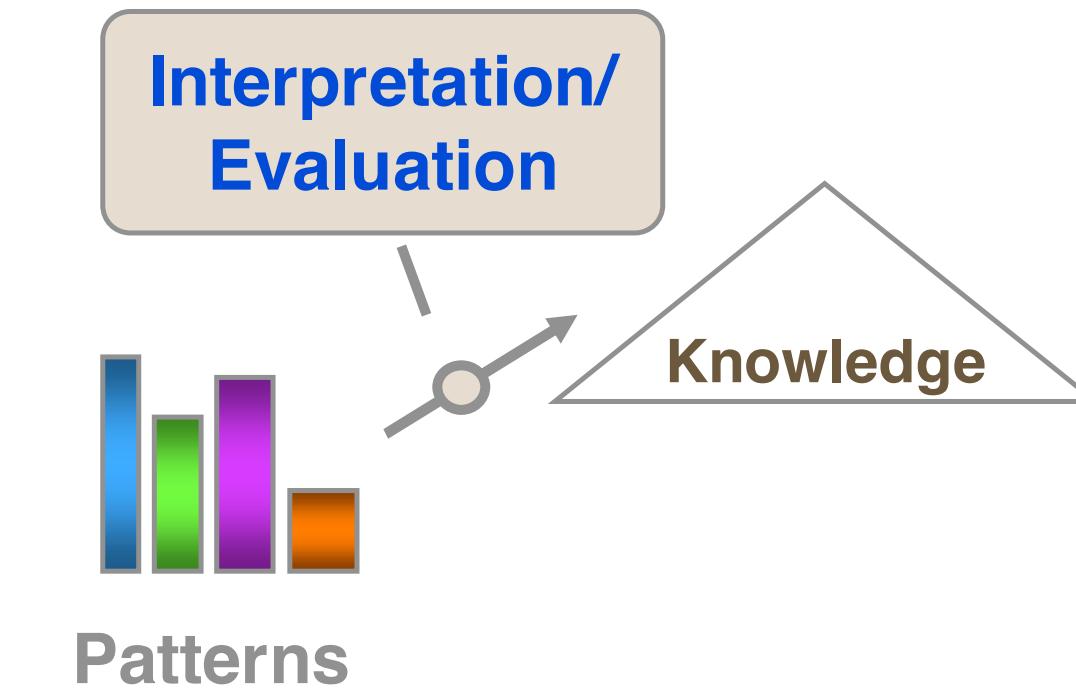
Homes within a mile of either high-end grocery store begin appreciating faster than other homes after the stores open, Zillow has discovered

Jan 25, 2016

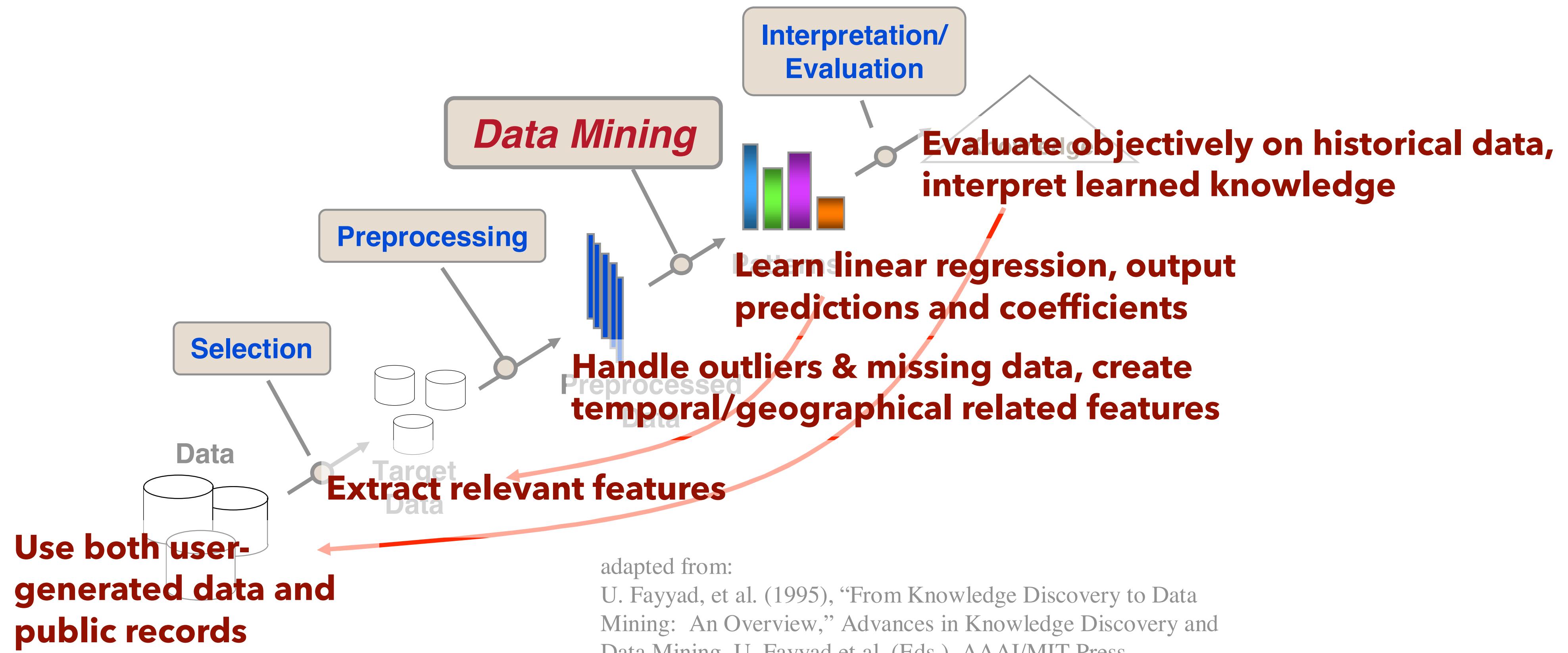
SEATTLE, Jan. 25, 2016 /PRNewswire/ -- Your local grocery market has a lot to do with what happens in your local housing market, according to a new analysis by Zillow featured in the paperback edition of *Zillow Talk: Rewriting the Rules of Real Estate* (Grand Central Publishing, Jan. 26).

Specifically, Zillow found that homes grow more rapidly in value if they are closer to a Trader Joe's or Whole Foods¹. 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



COURSE OVERVIEW

COURSE OVERVIEW

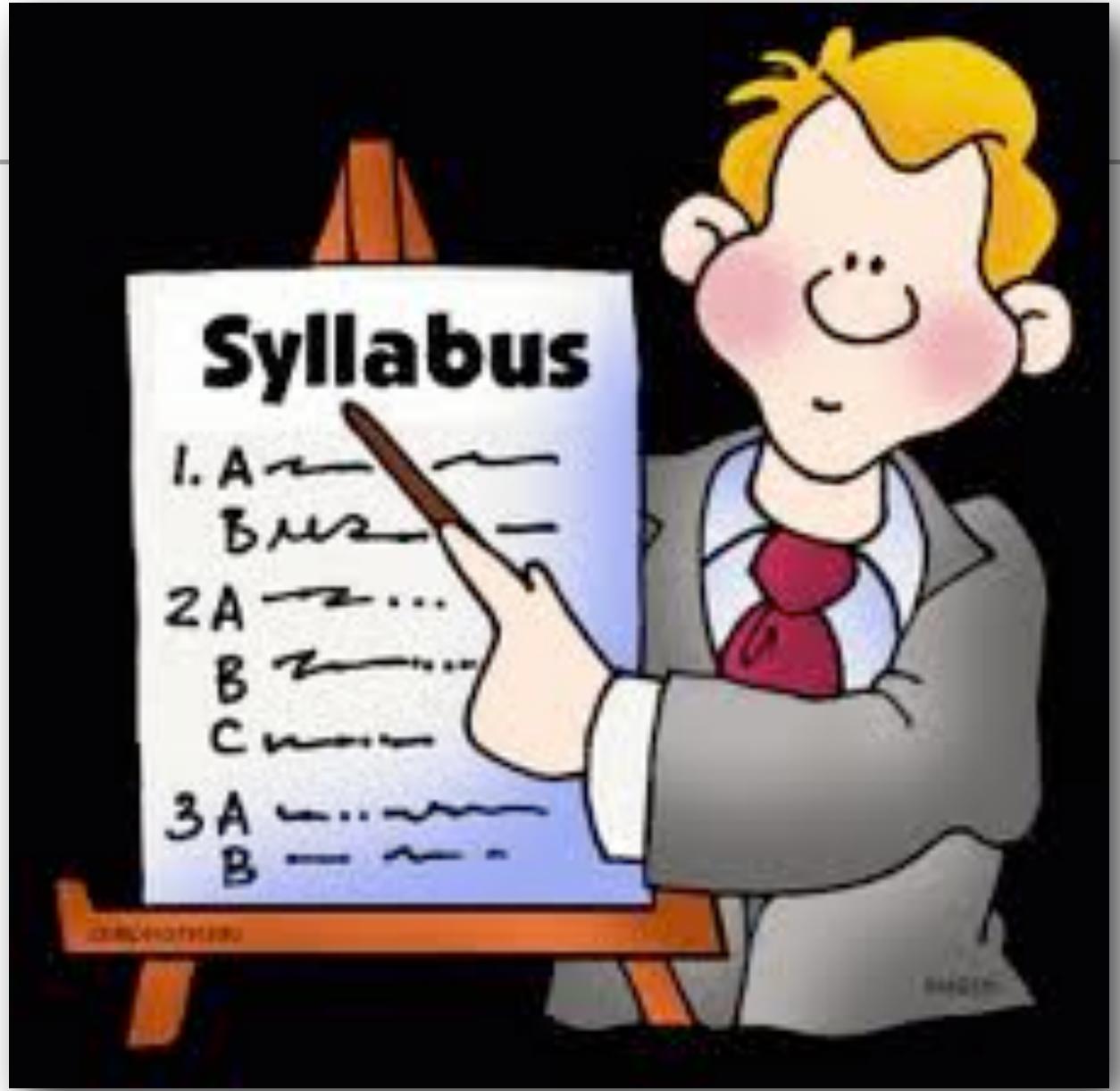
COURSE GOALS

- ▶ 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

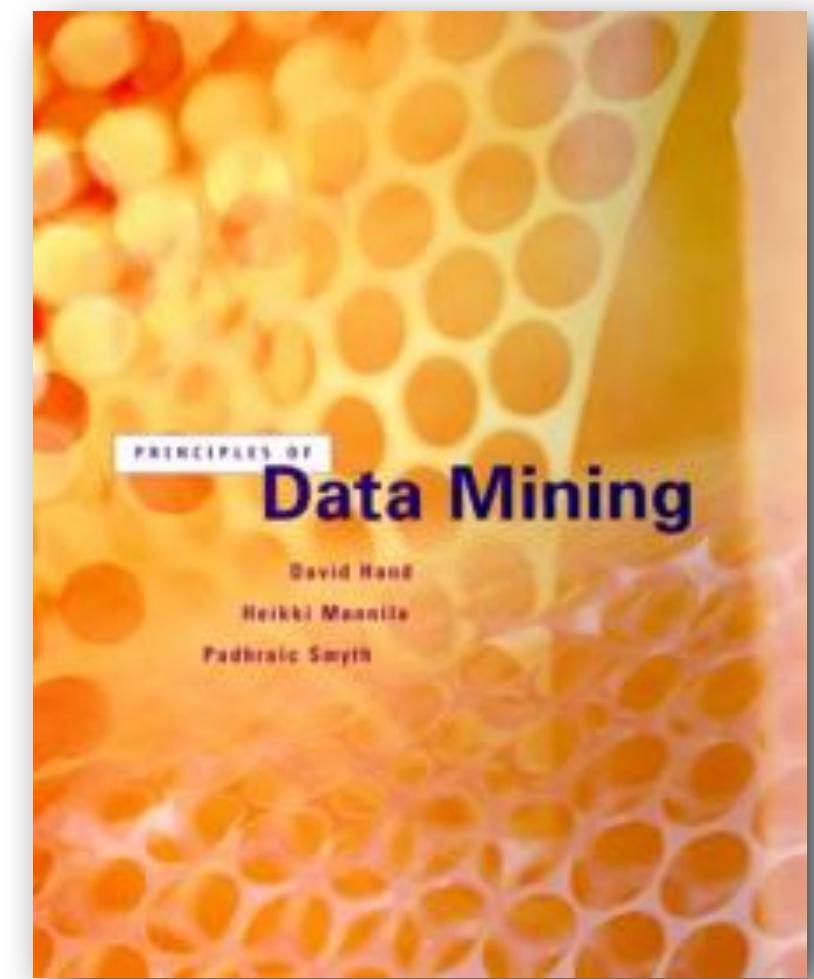
- ▶ Time and location: MW 4:30-5:45pm
- ▶ Instructor: Ming Yin, mingyin@purdue.edu
- ▶ Course videos and slides will be uploaded to BrightSpace after each class.
- ▶ Pandemic related policy:
 - ▶ Please wear your mask in class! This is a requirement of the university: <https://protect.purdue.edu/required-covid-19-protocols/>
 - ▶ Keep social distancing as much as possible when choosing your seats in the classroom!
 - ▶ If you have any questions after the class time is over, please post it on Piazza and/or sign up a virtual office hour with me.
- ▶ The class has reached the maximum capacity, so we will not enroll new students for this class any more.

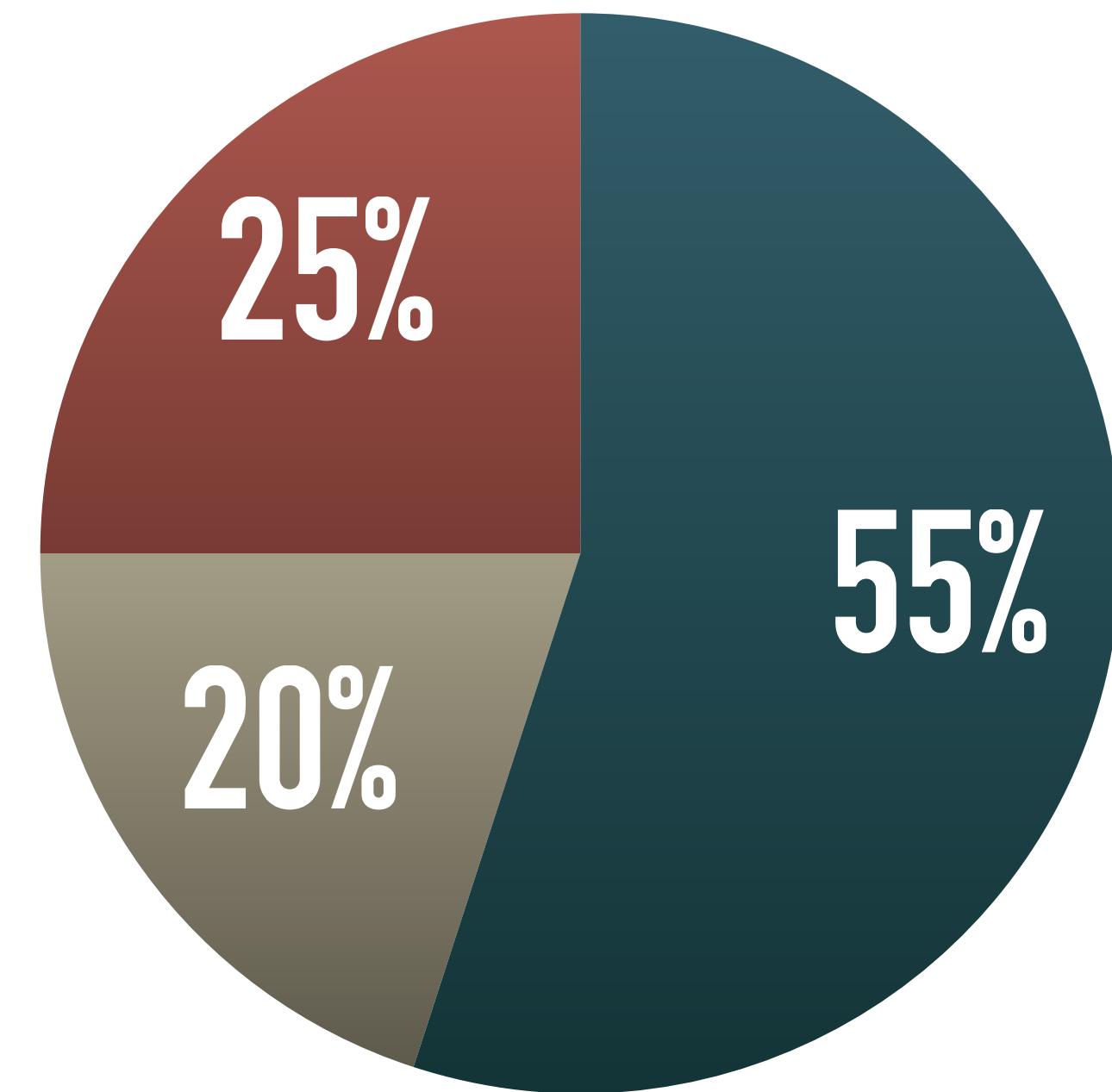
LOGISTICS

- ▶ Teaching assistants: **Meric Altug Gemalmaz, Yonghan Jung, Leonardo Vilela Teixeira**
 - ▶ Office hours: TBD
- ▶ Piazza: piazza.com/purdue/fall2021/cs57300
 - ▶ Please sign up to Piazza as soon as possible!
 - ▶ All Q&A will be on Piazza. Emails will be SLOW at best and unresponsive at worst
- ▶ Webpage: <http://mingyin.org/CS573/Fall2021/index.html>
- ▶ 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





● Homework

● Midterm

● Final Project

WORKLOAD

- ▶ Assignments
 - ▶ Five assignments including written/math exercises and programming assignments
- ▶ Exams
 - ▶ Online midterm: October 25 (24-hour)
- ▶ Final Project
 - ▶ Open topics
 - ▶ 1 proposal, 1 midterm report, 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 2-5 students
- ▶ Start to think about your project **NOW!**
 - ▶ Can consider data mining challenges on competition platforms like Kaggle

LATE POLICY

- ▶ Three **extension days** (scores not affected) can be applied on any chosen assignment **except for Assignment 1**
 - ▶ 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

ACADEMIC HONESTY

- ▶ Check Purdue's policy on academic integrity: [https://www.purdue.edu/odos/
osrr/academic-integrity/index.html](https://www.purdue.edu/odos/osrr/academic-integrity/index.html)
- ▶ DO NOT copy classmate's assignments / let your classmates copy your assignment / share answers during the online midterm exam
- ▶ The teaching team will closely monitor potential academic dishonesty!

NEXT CLASS

- ▶ Review basic knowledge on probability