Introduction to Data Mining

Project ENABLE

May 20, 2019

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

Course Structure

- The course has three parts:
  - **Lectures** will introduce the main topics
  - **Hands-on activities** will provide with the opportunity to practice those topics (usually come after the lectures)
  - **Project** will promote learning-by-doing
- Lecture slides and tools will be available at the course web page

Project

- One group project
- You will be given sample datasets
- Every Friday morning, you will have project update talk where you can share your progress and get help from instructors

Teaching Materials

- Text

- Data
  - Subset of MEPS data
- Tools
  - JupyterHub

Topics

- Introduction
- Data
- Data treatment
- Descriptive analysis
  - Mean, median, standard deviation, and so on
  - Bivariate analysis
- Prescriptive analysis
  - Linear regression
  - Association rule
  - Neural network
  - Feature selection
- Designing experiments
- Evaluation
- Visualization

Any Question and Suggestions?

- Your feedback is essential
  - We are building up the Summer Boot Camp together
- Share you questions and concerns with class
- Do ask
- No pain no gain – no magic

Rules and Policies

- Certificate

Introduction to Data Mining

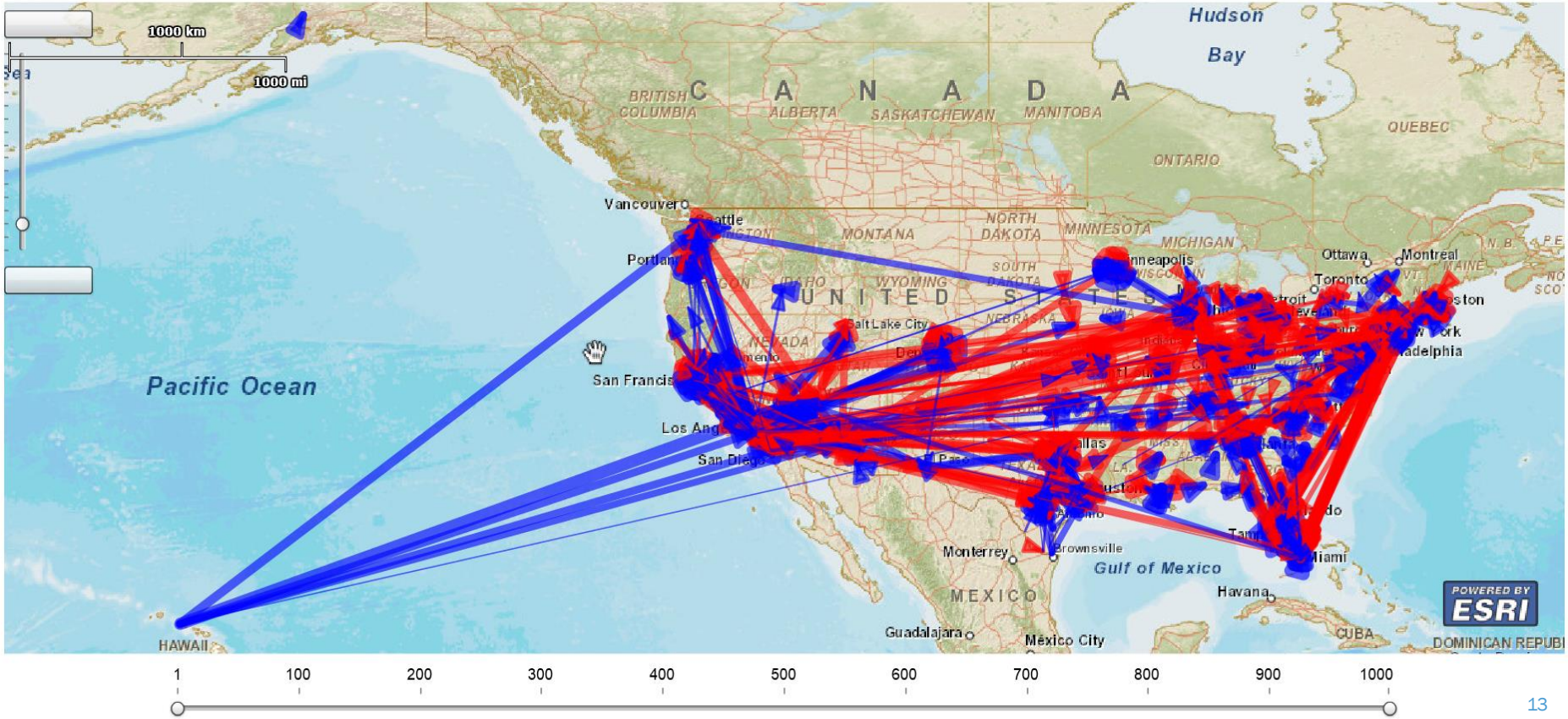
What is Data Mining?

Examples of Data Mining

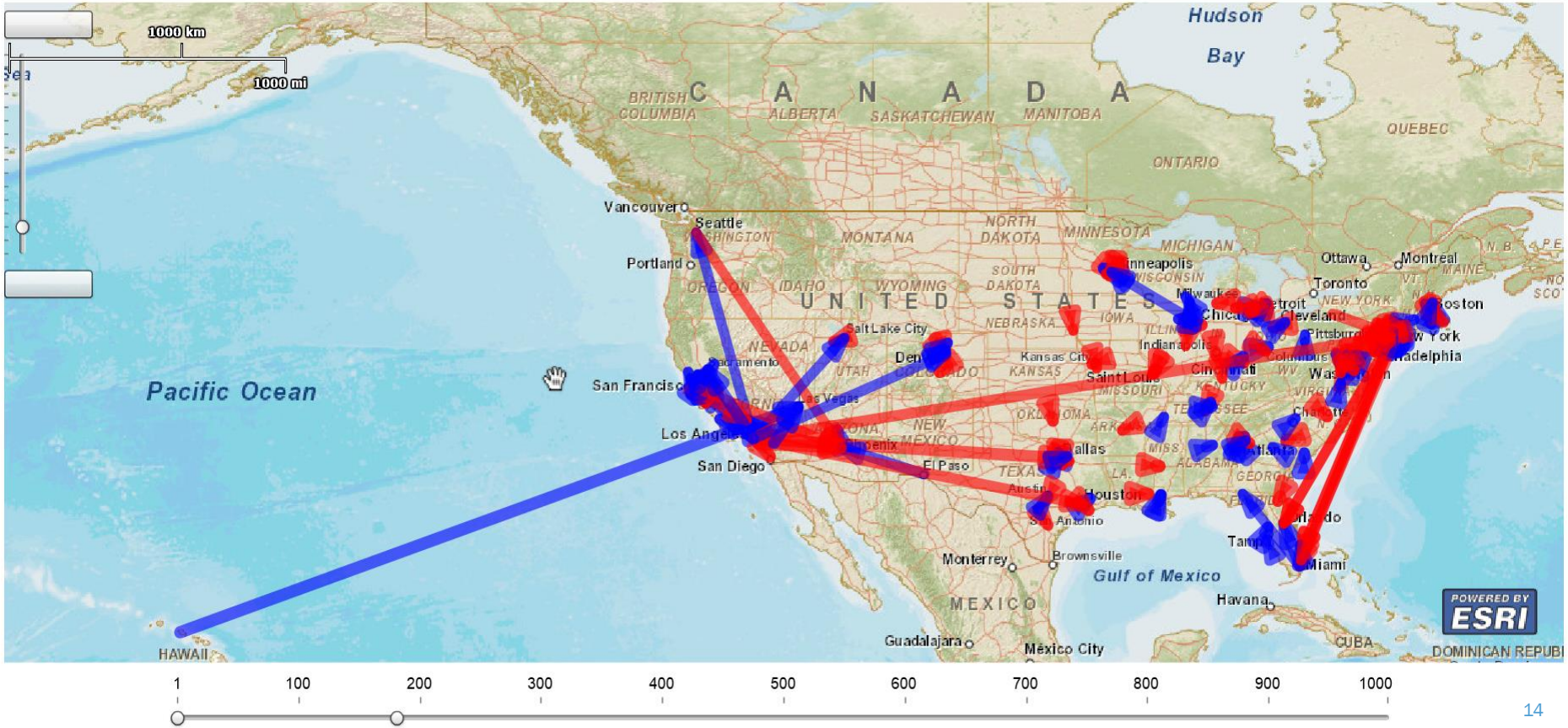
Data Mining Tasks

- Classification
- Association rule mining
- Clustering
- Visualization

Example of Visualization (Flow Map)



Example of Visualization (Flow Map, cont ...)



Why is Data Mining Important?

- A huge amount of data (from gigabytes to terabytes, and to petabytes)
- Human's capability?
- More computing resources and access to them get cheaper
  - Cloud computing (e.g., Amazon Web Service, Microsoft Azure, and Google Cloud)

Why is Data Mining Important? (cont ...)

MARKETS BUSINESS INVESTING TECH POLITICS CNBC TV

Amazon's joint health-care venture finally has a name: Haven

PUBLISHED WED, MAR 6 2019 • 4:05 PM EST | UPDATED WED, MAR 6 2019 • 5:09 PM EST

Angelica LaVito @ANGELICALAVITO | Christina Farr @CHRISSYFARR | Hugh Son @HUGH_SON

SHARE

KEY POINTS

- The joint health-care venture between Amazon, J.P. Morgan and Berkshire Hathaway finally has a name. And that's Haven.
- Amazon CEO Jeff Bezos, J.P. Morgan CEO Jamie Dimon and Berkshire Hathaway CEO Warren Buffett last January announced they were teaming up to tackle rising health-care costs.
- They named Dr. Atul Gawande as CEO in June.

RELATED


Amazon's joint health-care venture finally has a name: Haven


Mental health app makers join forces to go after multibillion-dollar market

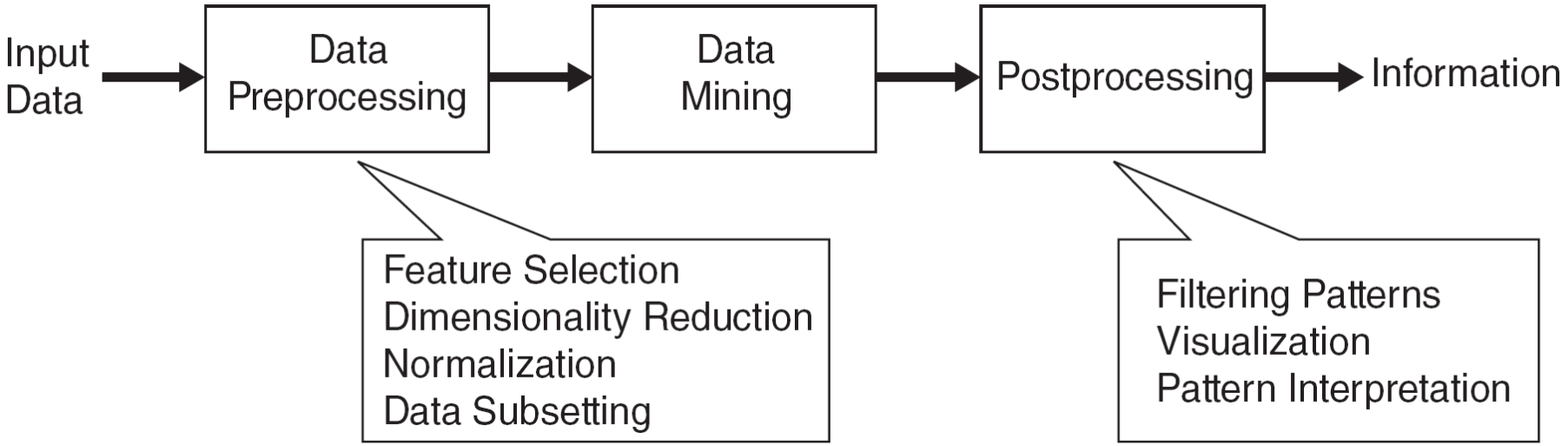

This hospital modeled itself after the Apple Store, lets mothers use gadgets to monitor pregnancies


Amazon continues its push into the pharmacy business, and has appointed a 14-year vet to run it

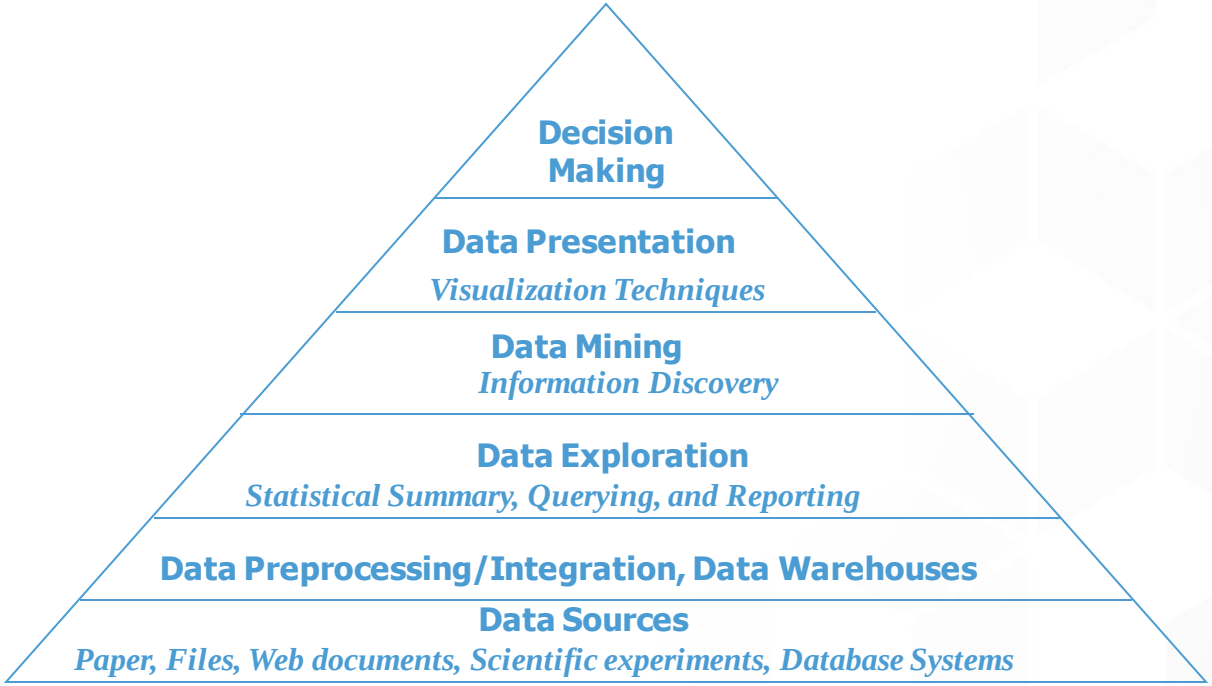

Epic CEO Judy Faulkner: We would never sell to

Why is Data Mining Necessary

Data Mining Process



Data Mining Process



Data Mining Applications

- Marketing: Target
- Fraud detection: credit card fraud
- Precision medicine: health care & medical data mining

Data Mining Tasks

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.

- Description Methods
  - Find human-interpretable patterns that describe the data.

Predictive Modeling: Classification

Regression

Clustering

Common Problems in Data Mining

- Tremendous amount of data
  - Efficiency and scalability
- Data quality (e.g., handling noise and incomplete data)
- High-dimensionality of data
- High-complexity of data
  - Mining networked, dynamic, and global repositories (e.g., fine dust)

Resources for Studying Data Mining

- Conferences
  - ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (KDD)
  - SIAM Data Mining Conf. (SDM)
  - (IEEE) Int. Conf. on Data Mining (ICDM)
  - European Conf. on Machine Learning and Principles and practices of Knowledge Discovery and Data Mining (ECML-PKDD)
  - Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)
  - Int. Conf. on Web Search and Data Mining (WSDM)
- Journals
  - Data Mining and Knowledge Discovery (DAMI or DMKD)
  - IEEE Trans. On Knowledge and Data Eng. (TKDE)
  - KDD Explorations
  - ACM Trans. on KDD

Course Resources

- Class website

- Piazza

Seeking help

- E-mail
- Office hours

Tips for Success

- Work hard
- Be patient and have reasonable expectations
- you're not supposed to understand everything we cover in class during class
- Seek help sooner rather than later
- Remember the golden rule: no pain, no gain

Any Questions?

What’s Data?

Next Class

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL