UNION
COLLEGE

CSC-321: Data Mining & Machine Learning

Prof. Nick Webb

webbn@union.edu

UNION COLLEGE

Data Mining: What's it all about?

- It's a buzzword!
  - Good for me: Get you interested?
  - Bad for me: Maybe it's not what you think?

- A practical definition:
  - Finding *useful* patterns in large amounts of data
  - What certain agencies call *actionable intelligence*

UNION COLLEGE

Data Mining

- Two ‘modes’ of thought

- Looking for patterns in large amount of data
  - These might not be wildly complex patterns
- Looking for patterns or relationships that are too hard to describe
  - These might only exist in small amounts of data

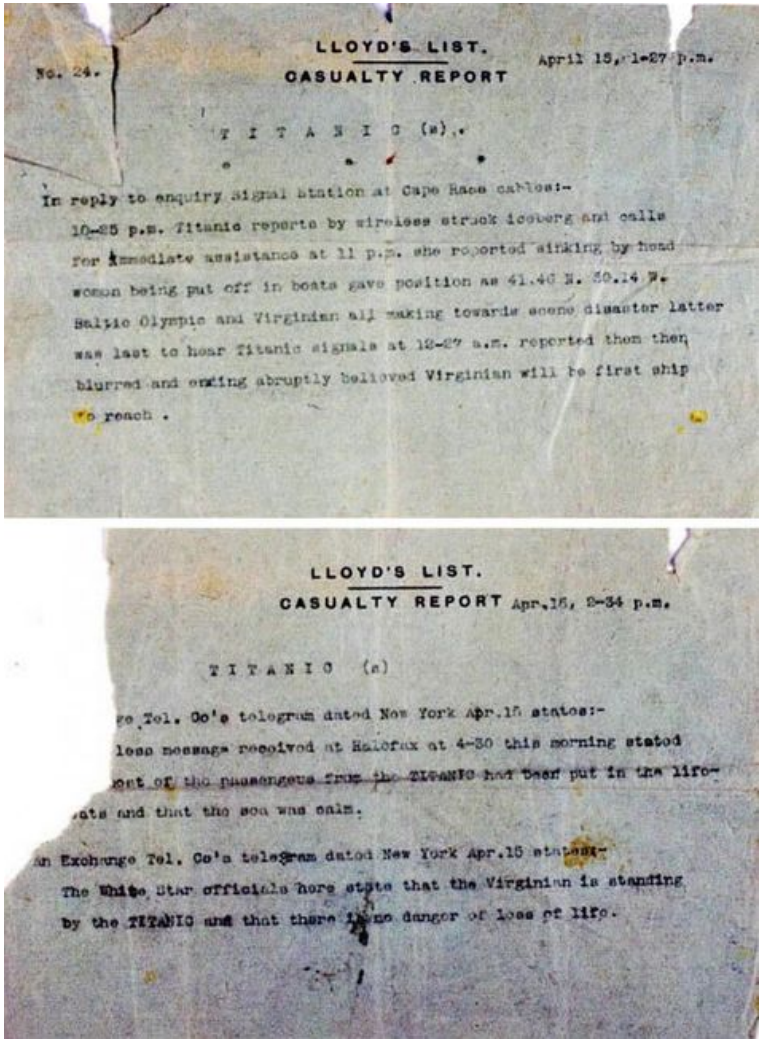
UNION
COLLEGE

The data fire hose…

UNION COLLEGE

…is not new

- Many people used to be employed at Lloyds of London to scour newspapers for reports of ship wrecks



No. 24. LLOYD'S LIST. April 15, 1-27 p.m.
CASUALTY REPORT

TITANIC (s).

In reply to enquiry Signal Station at Cape Race cables:-
10-25 p.m. Titanic reports by wireless struck iceberg and calls for immediate assistance at 11 p.m. she reported sinking by head women being put off in boats gave position as 41.46 N. 50.14 W. Baltic Olympic and Virginian all making towards scene disaster latter was last to hear Titanic signals at 12-27 a.m. reported them then blurred and ending abruptly believed Virginian will be first ship to reach .

LLOYD'S LIST.
CASUALTY REPORT Apr.15, 2-34 p.m.

TITANIC (s)

...e Tel. Co's telegram dated New York Apr.15 states:-
...less message received at Halifax at 4-30 this morning stated ...ort of the passengers from the TITANIC had been put in the life-...ats and that the sea was calm.

...n Exchange Tel. Co's telegram dated New York Apr.15 states:-
The White Star officials here state that the Virginian is standing by the TITANIC and that there is no danger of loss of life.

UNION COLLEGE

Machine Learning: What's it all about?

- We could find these patterns by hand
- Indeed, we did find these patterns by hand!

- As data increases, so too does the time, and the complexity
- Maybe our intuition is a good place to start, but maybe machines can do better.

UNION COLLEGE

Machine Learning

- Automated (or semi-automated) way for machine to 'discover' patterns in data
- CAUTION: There is NO magic
- You should know what you're looking for
- You should have an idea how to find it

UNION
COLLEGE

NO MAGIC

UNION COLLEGE

Machine Learning

- Automated (or semi-automated) way for machine to 'discover' patterns in data
- CAUTION: There is NO magic
- You should (mostly) know what you're looking for
- You should (mostly) have an idea how to find it

UNION COLLEGE

ML ≠ AI

- Machine Learning and Artificial Intelligence are NOT the same
- ML – find patterns in data
- AI – give machines wisdom and intelligence

- ML will find an answer, but will not know if it makes sense
- AI will know if an answer is sensible

UNION COLLEGE

How is it used?

- Academic Examples
  - Speech: Recognition, Production
  - Image: Recognition, Emotions
  - Language: Translation, Analysis
  - Bioinformatics
  - Robotics
  - Financial markets
  - Sports
  - Politics
  - And so on and so on…

UNION COLLEGE

How is it used?

- Real world
  - Amazon
  - Netflix
  - Google
  - Intelligence Community
  - Supermarkets
  - Self driving cars
  - Dating websites
  - And so on and so on…

UNION COLLEGE

ML of the day

- https://www.wired.com/story/machine-learning-march-madness/

UNION COLLEGE

So what will you learn?

- What ML is:
  - Key algorithms
  - How we implement them
  - How we use them
  - The theory behind them

UNION COLLEGE

Key topics

- Basic algorithms
  - Linear regression
  - Naïve Bayes
  - KNN
  - Decision Trees
- How we evaluate algorithms
- Practical application
- Ethics and data

UNION COLLEGE

Skills you need

- Python
  - Implementation of algorithms
  - Acquisition and manipulation of data
- Statistics
  - Data description
  - Model evaluation

UNION COLLEGE

Learning Objectives

- Theoretical and practical aspects of machine learning as a tool for mining data
- Knowledge of methods and techniques for learning, application and evaluation
- Familiarity with open source machine learning tools

UNION COLLEGE

Assignments

- This is a developing course
  - TRANSLATION: I can and WILL change my mind
- Homework Assignments
- Midterm
- Challenge Problem
- Final Project
  - Literature survey
  - Final research paper

UNION COLLEGE

Homework

- Getting familiar with specific Python functionality
- Implementing algorithms
- Exercises using including data sets
- Explore the features of open source tool kits and start to work with real data

UNION COLLEGE

Challenge Problem

- Working on a data set I give you
- WHAT can you do to it, how well can you classify the data?
- A chance to get familiar with the data pipeline

UNION COLLEGE

Final Project

- Find a data set
- Perform experiments
- Create 8 page research paper
  - Intro, method, results, conclusion
  - Tell me about important attributes
  - What learning experiments you used

UNION COLLEGE

Course Policies

- There are ZERO late assignments accepted in this course

- Academic Integrity
  - DO NOT use code from ANYWHERE

UNION COLLEGE

Tools

- Python 3
- scipy / numpy / matplotlib / scikit-learn
- Jupyter notebooks

- Easiest way is to use anaconda
  - https://www.anaconda.com/