Executive Data Science Education

Claudia Perlich

Chief Scientist

@claudia_perlich

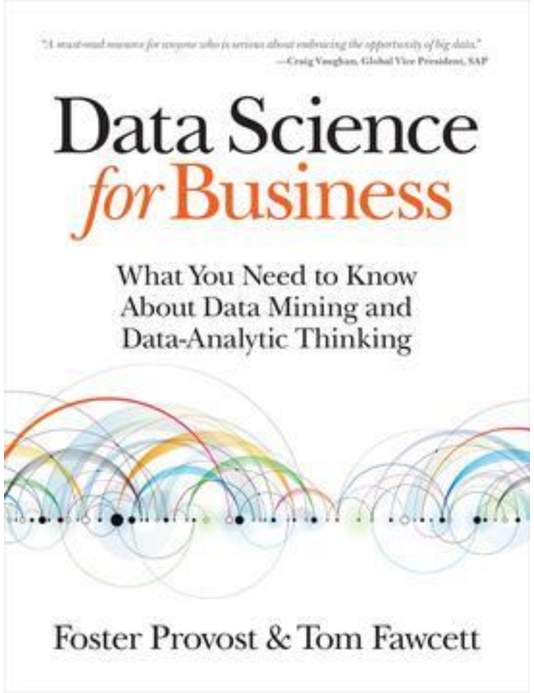
dstillery

Introductions

- Master from CU Boulder in Computer Science
- Master from TU Darmstadt in Computer Science
- PhD from Stern School of Business in Information Systems
- IBM T.J. Watson Research Lab: Research Staff member
- Chief Scientist at Dstillery
- Adjunct Professor at Stern NYU

Data Mining for Business Intelligence

- Course is offered since 2003 …
- Since 2015 2 Tracks: Managerial & Technical
- Managerial (MBA Elective)
  - No coding required
  - Use WEKA
  - Many ‘cases’ covered
  - Focus on managing DS
  - Project


"A must-read resource for anyone who is serious about embracing the opportunity of big data."
—Craig Vaughan, Global Vice President, SAP
Data Science for Business
What You Need to Know About Data Mining and Data-Analytic Thinking
Foster Provost & Tom Fawcett

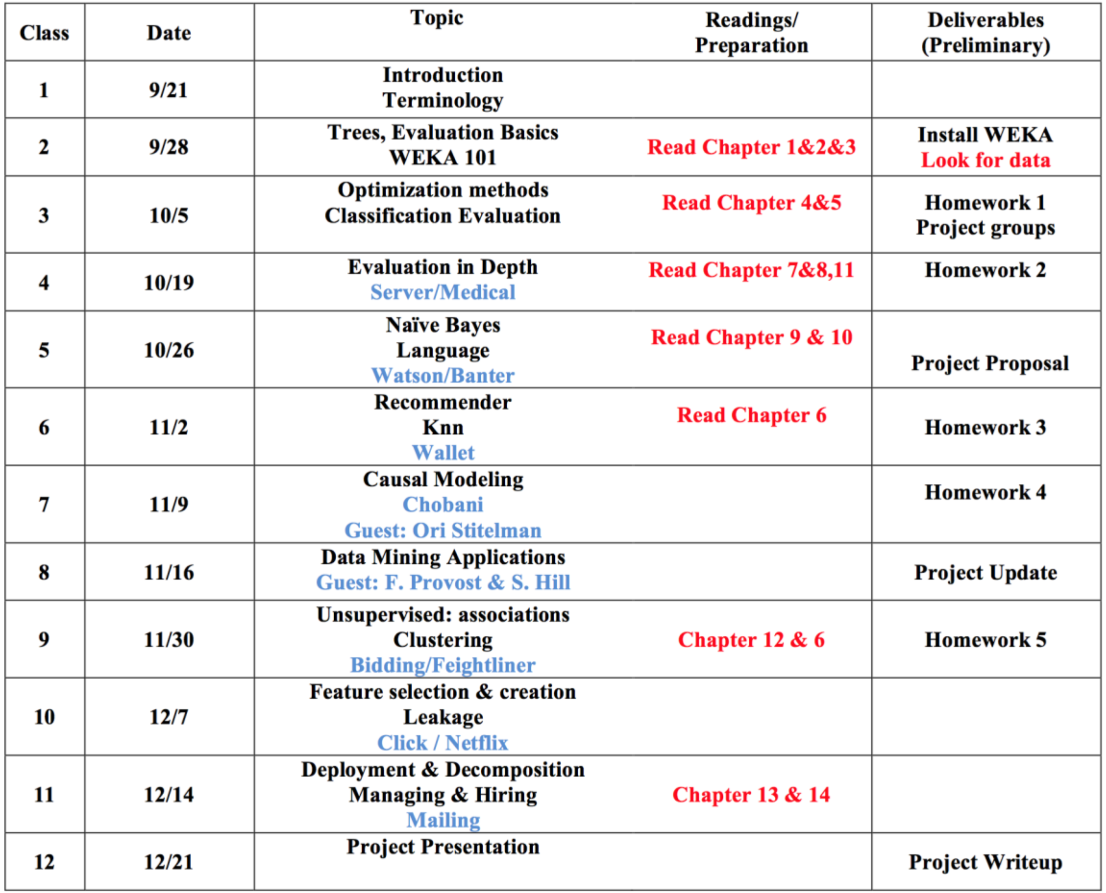
Course Format

- weekly 3 hour lectures
- 6 homeworks that are heavily scripted to enable project
- 2-3 guest speakers
- 2 hour take home final
- Project
  - Find your own predictive problem & data
  - Solve it (using any tool of your choice)
  - Demonstrate business value
  - 10-15 pages written report
  - (in class presentation)

Broad Outline

- Terminology
- Methods
  - Supervised & some unsupervised learning
  - Model evaluation
  - Importance
- Applications
- Managing DS
  - Deployment
  - Hiring & Interviews
  - Proposal evaluation

Syllabus Example from Fall 2016



Class	Date	Topic	Readings/ Preparation	Deliverables (Preliminary)
1	9/21	Introduction Terminology		
2	9/28	Trees, Evaluation Basics WEKA 101	Read Chapter 1&2&3	Install WEKA Look for data
3	10/5	Optimization methods Classification Evaluation	Read Chapter 4&5	Homework 1 Project groups
4	10/19	Evaluation in Depth Server/Medical	Read Chapter 7&8,11	Homework 2
5	10/26	Naïve Bayes Language Watson/Banter	Read Chapter 9 & 10	Project Proposal
6	11/2	Recommender Knn Wallet	Read Chapter 6	Homework 3
7	11/9	Causal Modeling Chobani Guest: Ori Stitelman		Homework 4
8	11/16	Data Mining Applications Guest: F. Provost & S. Hill		Project Update
9	11/30	Unsupervised: associations Clustering Bidding/Feightliner	Chapter 12 & 6	Homework 5
10	12/7	Feature selection & creation Leakage Click / Netflix		
11	12/14	Deployment & Decomposition Managing & Hiring Mailing	Chapter 13 & 14	
12	12/21	Project Presentation		Project Writeup

Typical Class Format

- 60% technical material
- Occasional demonstration using WEKA
- 40% case discussion
  - 10% business case brainstorming
  - 30% teacher presentation of solution
  - 10% critical evaluation of impact

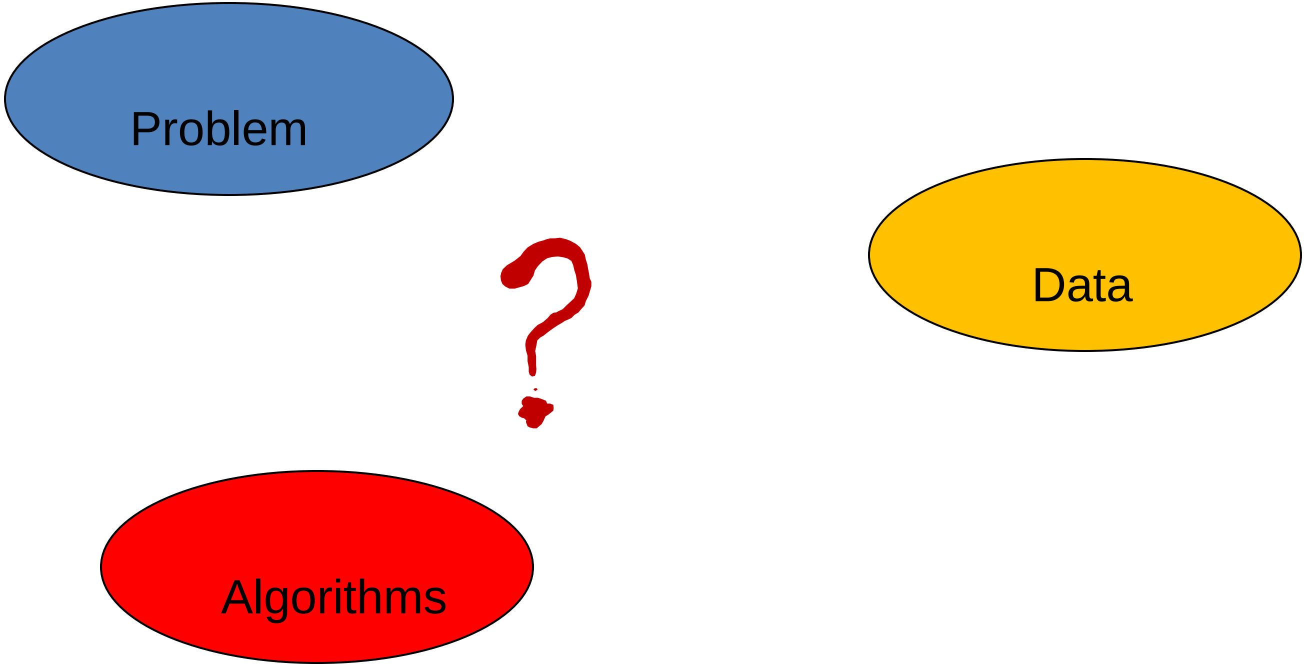
High-level Goals of the class

1. Approach business problems data-analytically
   - Think carefully & systematically about whether & how data can improve performance
2. Be able to interact competently on the topic of data mining for business intelligence
   - Know the basics of data mining processes, algorithms, & systems well enough
3. Receive hands-on experience mining data
   - You should be able to follow up on ideas or opportunities that present themselves

‘Managerial’ Objectives

- Appreciate the hardships of good DS
  - Data prep is key and takes as long as it takes
  - Sometimes the data does not support the solution of the problem
- Recognize DS opportunities in your business
  - Only some problems are DS problems
- Evaluate DS
  - BS detection …
- Train to think ‘backwards’ from the problem, not forwards from the data
- How to get started with DS in a (small) company ..
- How to hire a data scientist (what to look for)

Course Philosophy

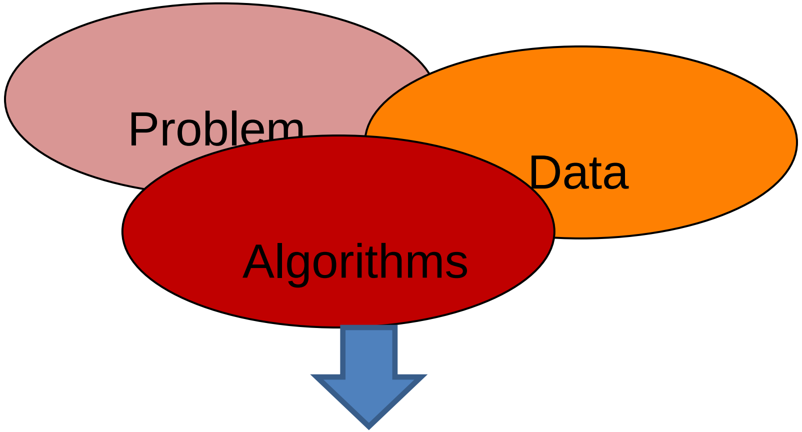


Problem

Data

Algorithms

What is Good Data Science?



Problem

Data

Algorithms

Case Discussion

- 3 types of cases/example projects:
  - Published work with technical details and business impacts. Please read them as best as you can (some details might be beyond your current knowledge)
  - Brainstorming problems and possible solutions. Think about the problem and possible solutions. Be prepared to present your ideas in class and defend them
  - Problems I have worked on that demonstrate some concept from class

Cases Instructions

1. Reframe in your mind the issue and why it might be relevant/important, ask questions (and possibly answer them ….)
2. What are possible actions your can take in your role and specifically what (micro) decision/choices do they translate to?
3. How do you measure ‘better’ decisions?
4. How can the data you have (or should have) help you make those decisions better?
5. Can you evaluate your strategy BEFORE implementing it?
6. What are the alternative baseline strategies you should compare to?

Case Example:
Direct Mailing campaign

You are working for your favorite non-profit organization. They have been running bi-annual mailing campaigns soliciting monetary donations for the past 5 years based on a database of existing donors. Historically they sent a letter to everybody. You are in charge of running this years mailing campaign. You are wondering if you can do better …

What makes for a good Project Problem?

- You can take some action
- You have some hope that a predictive model can add some value
- You are likely to be able to demonstrate business value by simulating a decision

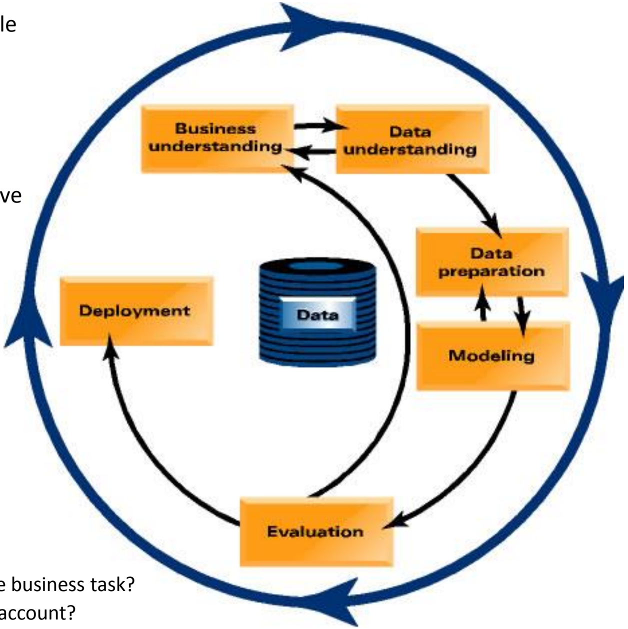
Bad Project ideas …

- Flight delay
- Box office return
- NYC traffic accidents
- Stock returns

Putting it to the test: Proposal Evaluation

- What exactly is the business problem to be solved and the action to be taken?
- What business entity does an instance/example correspond to?
- Is a target variable defined?
  - Supervised vs. unsupervised
- If so, is it define precisely?
  - Think about the values it can take
- Are the attributes defined precisely?
  - Think about the values they can take
- Will modeling this target variable solve/improve the stated business problem?

- Will it be reasonably possible to get values for attributes and put them into a single table?
- Will it be reasonably possible to get values for the target variable (for training) and put them into the table?
  - How exactly does one acquire values for the target variable? Is there any cost involved? If so, is it taken into account?



- Is holdout data used?
  - cross-validation is one technique
- Is there a plan for domain-knowledge validation?
- Is the evaluation setup and metric appropriate for the business task?
  - E.g., are business costs/benefits taken into account?
  - For classification, how is threshold chosen?
  - Are probability estimates used directly?
  - Is ranking more appropriate (e.g., for a fixed budget)?
- Will deployment as planned actually (best) address stated business problem?

- Is the choice of model appropriate for the choice of target variable?
  - Classification, regression
- Does the model/modeling technique meet the other requirements of the task?
  - Accuracy, comprehensibility, speed of learning, speed of application, amount of data required, type of data, missing values, fit with knowledge of problem (e.g., definitely non-linear)
  - See chart on next slide
- Should various models be tried and compared (in evaluation)?

Things Students struggle with most

- Recognizing a predictive modeling problem
- How good is good?
- Value of good baselines
- Translating a model into an action
- Precise language (‘accuracy’)
- Data preparation for project is always late …

Thank You!

dstillery