11-631 Data Science Seminar

Eric Nyberg

Carnegie Mellon University

Fall 2019

Outline for Today

Meet and Greet

Data Science

Seminar and Capstone Details

Meet and Greet

What is (Computational) Data Science?

Write down five - ten phrases describing data science

What is Data Science? (Anthony)

Data Extraction
Data Integration
Learning Models
Prediction
Analysis
Error

Research Idea
Design
Participation
People
Experiments
Products

What is Data Science? (Matthias)

Complexity

Patterns

Prediction

Clusters

Domain problem

Interdisciplinary

Visualization dimension

Semantic interpretation

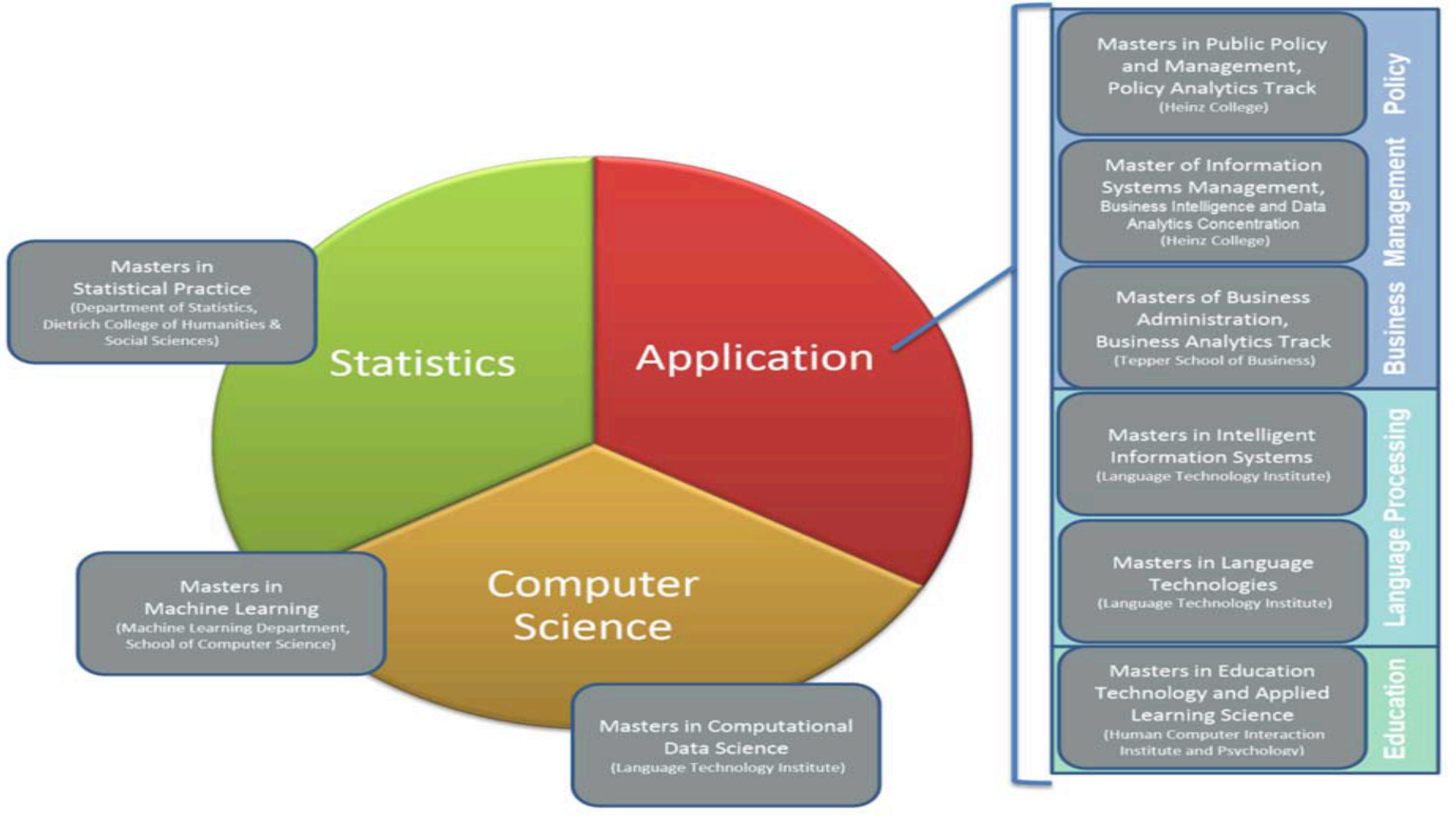
Data Science Employer Survey

40 total companies replied

American Express, Apple, AT&T, Bain & Co, Barclays, Caterpillar, Citadel, Collected, Comprehend Systems, Conversant Labs, Diamond Kinetics, Doctor.com, Google, Green Hills Software, Groupon, IBM, Immunetrics, kWantera, Liberty Mutual, Lockheed Martin, Microsoft, Pittsburgh Equity Partners, Ricoh Innovations Corporation, Rocket Fuel, Salesforce, Sandia National Laboratories, Shoefitr, Splice Machine, TE Connectivity

50 different job titles

Data Science Survey


Masters in Statistical Practice
(Department of Statistics, Dietrich College of Humanities & Social Sciences)
Statistics
Application
Masters in Machine Learning
(Machine Learning Department, School of Computer Science)
Computer Science
Masters in Computational Data Science
(Language Technology Institute)
Masters in Public Policy and Management, Policy Analytics Track
(Heinz College)
Master of Information Systems Management, Business Intelligence and Data Analytics Concentration
(Heinz College)
Masters of Business Administration, Business Analytics Track
(Tepper School of Business)
Masters in Intelligent Information Systems
(Language Technology Institute)
Masters in Language Technologies
(Language Technology Institute)
Masters in Education Technology and Applied Learning Science
(Human Computer Interaction Institute and Psychology)
Policy
Business Management
Language Processing
Education

Course Learning Outcomes

Undergrad = apply textbooks

Master = apply literature

PhD = extend literature

Lifelong Critical Analysis Skills

Communication skills

Capstone Process

Fall: 11-631 Data Science Seminar (12 units)

- If you want your own project, you need to do much of the Spring steps now
- Convince faculty mentor to supervise and confirm with MCDS committee

Spring: 11-634 Capstone Planning Seminar (12 units)

- Student teams formed, weekly meeting with advisor
- Literature survey & discussion
- Data survey & exploratory analysis
- Start system/model building
- Report & Presentation on proposed work in fall

Fall: 11-632 Capstone Course (24 units)

- Resume work, weekly meeting with advisor
- Midterm milestone presentations
- Final project presentation
- Deliverable: Workshop-level project paper

MCDS Capstone Learning Outcomes

After your capstone project, you should be able to …

… identify tradeoffs among data science techniques (analytics, systems and/or human-centered) and contrast design alternatives, within the context of specific data science application domains.

… survey, interpret and comparatively criticize state of the art research talks and papers, with emphasis on constructive improvements.

.... organize, execute, report on, and present a real world data science project in collaboration with other researchers/programmers.

Class Format

Class has one section which meets twice per week

Semester falls into three phases

- First half with student presentations
- Second half with “surprise papers” and related work drafting
- Final phase reviewing 2nd year work

During student presentations, we will meet during both sessions

For surprise paper sessions, we will meet **only during the earlier slot**

All schedule details will be posted on Piazza

Syllabus

Weekly

- Read assigned paper(s)
- Submit analyses

First Semester Half

- Rank paper list
- Matched into teams of 3
- Research & Present paper
- Class discussion

Second Semester Half

- Read “Surprise” paper in class
- Write & submit related work paragraph
- Class discussion
- Pick one session, research papers, and write detailed survey

Semester End

- Review 2nd year Capstone draft reports
- Attend Capstone final presentation

Weekly Paper Analysis

Read paper & submit analysis (Google Form):

- Summarize paper in three sentences
- Three positive things about the paper
- Three negative things about the paper
- Three questions you would like to ask authors

Submission deadline: each **Wednesday 6 pm**

Presentation papers

1. **Visualizing and understanding recurrent networks.** Karpathy, A., Justin J., and Li F.-F.
2. **The Knowledge Accelerator: Big Picture Thinking in Small Pieces**. Hahn, N., Chang, J. C., Kim, J., Kittur, A.
3. **Enriching word vectors with subword information**. Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov
4. **Understanding deep learning requires rethinking generalization**. Zhang, Chiyuan, et al.
5. **Men also like shopping: Reducing gender bias amplification using corpus-level constraints.** Zhao, J., T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang
6. **Stress Test Evaluation for Natural Language Inference** A. Naik, A. Ravichander, N. Sadeh, C. Rose, G. Neubig.
7. **Scaling Distributed Machine Learning with the Parameter Server**. Mu Li, D. G. Andersen, J. W. Park, A. J. Smola, A. Ahmed, V. Josifovski, J. Long, E. J. Shekita, and B.-Y. Su.
8. **Automatic Database Management System Tuning Through Large-scale Machine Learning**. Dana Van Aken, Andrew Pavlo, Geoffrey J. Gordon, and Bohan Zhang
9. **Snorkel: Rapid Training Data Creation with Weak Supervision**. A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, C. Re
10. **Accelerating innovation through analogy mining**. Hope, T., Chan, J., Kittur, A., & Shahaf, D.

Presentation Guide

Format

- 15 minutes presentation (5 min each)
- 10 minutes moderated discussion
- Use ⅓ of time to summarize paper
- Use remaining ⅔ to discuss related work *and its connection*
- Presentation grade will consist of ⅓ and ⅔, respectively
- Focus is on your analysis and discussion of the paper and your research around it
- Peer review by two students in class

Research

- Work through paper
- Look up what you do not understand
- Look & read related work
- Ask for someone to review slides
- Review & finalize slides
- Rehearse/practice presentation

Paper Title		
	Praise	**Suggestions for Improvement**
General presentation		
Speaker #1		
Speaker #2		

Researching Related Work

?

Researching Related Work

Sources

- Mainstream search engines
- Specialized: Google Scholar, arxiv, etc.
- Author websites
- Conference websites
- Journal websites
- ...

Research Process

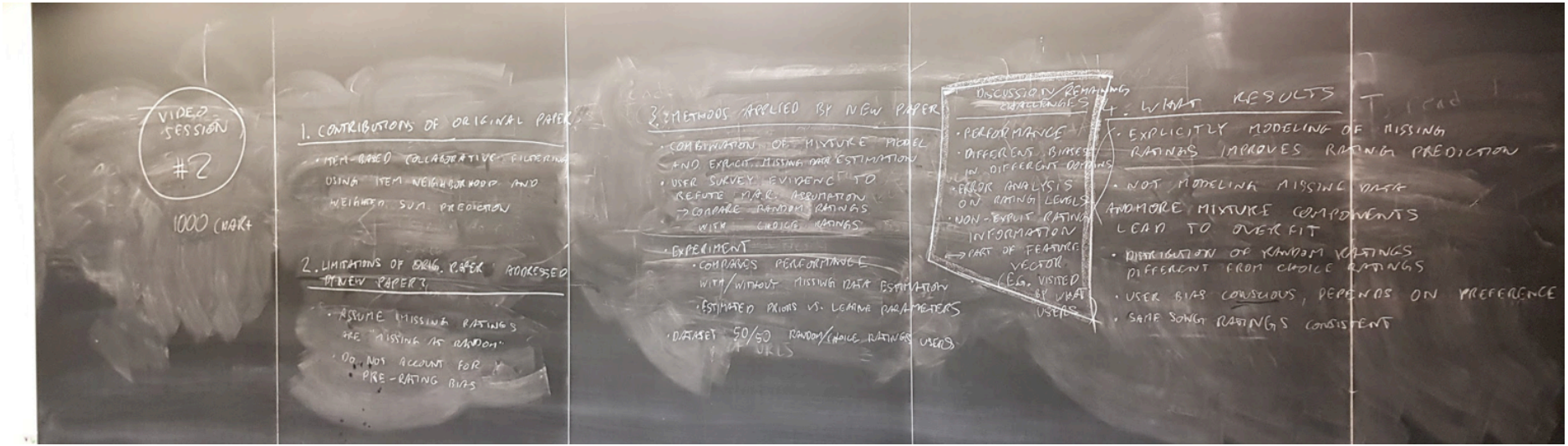
- Do searches & pool results
- Do skimming passes, look at abstracts & conclusions
- Follow incoming & outgoing citations
- Look up authors
  - Look up their students, advisors, lab colleagues, etc.
- Look up conferences/workshops
  - Topical tracks, sessions, etc.
- Iterate until pool converges, then focus on papers that you can read given available time
- Keep chronological overview

Related Work Survey

- **Guiding Questions:** Given focal paper P and related paper R:
  - What original contributions does P make?
  - What limitations or open questions of P did R address?
  - What methods did R apply to overcome these limitations?
  - What results/insights were obtained in R?
  - Discussion: What problems/questions remain?

- This pattern can be found everywhere in academic writing
- Doing it well is a skill. As MCDS graduates, you will be expected to be proficient working with cutting edge academic literature
- Introduction in Data Science Seminar, practice in Capstone Planning Seminar, demonstrate in Capstone Project

Related Work Survey



Anthony Tomasic, Matthias Grabmair, Eric Nyberg

Course Assessment

Presentation (30%)

Participation & Attendance (10%)

Submitted Analyses (30%) (weekly & in-class)

Related Work Survey (30%)

AIV Policy

For the **presentation**, you share work with your teammate.

For the **survey**, you may only work by yourself.

Researched material **must be referenced**!

For your **weekly analyses**, your own work only. Do not use the internet or other sources.

No Laptop Policy

No laptops in class except when instructed.
Bring paper and pen.

No excessive cellphone use.

Absences & Punctuality

One permitted absence for the semester.

Additional absences need to be approved ahead of time.

Interviews are no valid absence justification.

Be punctual! Repeatedly coming late will result in a grade penalty (*it happened!*).

Any Questions?

ACTION ITEMS:

Read syllabus

Choose preferences by Thursday

Watch presentation tutorial video