Introduction to data science

Patrick Shafto

Department of Math and Computer Science

Plan for today

- Introduce myself
- Review syllabus
- Introduce yourselves
- Homework

About me

- History
  - PhD from Northeastern University
  - Postdoc at MIT
  - Faculty position at University of Louisville
  - Faculty in Math and Computer Science at Rutgers
- Self-taught programmer. I speak: Python, Matlab, R, C, LaTeX, some LISP, others…

About me

- My research / expertise
  - Intersection between human and machine learning
  - Historically probabilistic models, markov chain monte carlo, but also gaussian processes, deep GPs, kernel methods, manifolds, topology, groups, recommender systems, active learning, machine teaching, etc
  - shaftolab.com

Patrick Shafto

Henry Rutgers Term Chair in Data Science & Associate Professor of Mathematics and Computer Science, Rutgers University - Newark

I am Henry Rutgers Term Chair in Data Science and Associate Professor in the Department of Mathematics and Computer Science at Rutgers - Newark. I am also affiliated with the Institute for Data Science, Learning and Applications (I-DSLA) and have appointments in Psychology, Rutgers Business School, and the Center for Molecular and Behavioral Neuroscience (CMBN) at Rutgers.

See my Curriculum Vitae.

Funding

- DARPA XAI: (2017-2021)
- NSF Science of Learning collaborative: (2016-2018)
- NSF EAGER MAKER (CISE): (2016-2018)
- NSF Cyber-human systems (CISE): Perception and Augmented reality (2015-2018)
- NSF INSPIRE (CISE, SBE): Human algorithm interaction (2015-2018)
- NSF REESE, CAREER award: Investigating the implications of social reasoning for learning from teaching (2012-2017)
- DARPA XData: Developing tools to facilitate analysis of very large data sets (2012-2015)

Open source software projects

- BayesDB: A Bayesian database table, lets users query the probable implications of their data as easily as a SQL database lets them query the data itself.
- CrossCat: A domain-general Bayesian method for analyzing heterogenous, high-dimensional data.

In the works...

Lu, C-K. Yang, S.C-H., & Shafto, P. (accepted). Standing wave decomposition Gaussian Process. Physical Review E. arXiv

Yang, S.C-H., Vong, W.K., Yu, Y. & Shafto, P. (under review). A unifying computational framework for teaching and active learning. arXiv code

Gweon, H., Shafto, P. & Schulz, L.E. (accepted). Too much information? Prior knowledge and the cost of information in learning and teaching. Developmental Science.

Bass, I., Gopnik, A., Hanson, M., Ramarajan, D., Shafto, P., Wellman, H., & Bonawitz, E.B. (under review). Children's developing theory of mind and pedagogical evidence selection.

2018

Yang, S.C-H., Yu, Y., Givchi, A., Wang, P., Vong, W.K., & Shafto, P. (2018). Optimal cooperative inference. Proceedings of the 21st international conference on Artificial Intelligence and Statistics (AISTATS). arXiv

What is data science?

- More on this in monday's class

Plan for today

- Introduce myself
- Review syllabus
- Introduce yourselves
- Homework

Syllabus

- Will be available on blackboard
- I have given you a paper copy

Four basic parts of this course

- Ingesting and cleaning manipulating data
- Asking and answering questions
- Data analysis (statistics and machine learning)
- Visualization

- Plus other important stuff: what is data science, programming for data science, collaboration and community, tools for big data, …

The most important part of this course

- Learning to ask and answer questions

My philosophy…

- The key thing that I can teach you is resourcefulness
- Learning is one thing, learning to learn is much more powerful
- We will be leveraging lots of materials that are freely available on the web to get us started

Homework

- Two parts
    - Part 1: Due Mondays will be an assignment
    - Part 2: Due Wednesdays will be a written critique of 2 other students, together with a 0-3 grade
- The goals are twofold
    - Do data science
    - Think critically & carefully about someone else work with the goal of making your work better

QUANTITATIVE LITERACY VALUE RUBRIC

for more information, please contact value@aacu.org

Association of American Colleges and Universities

Definition

Quantitative Literacy (QL) – also known as Numeracy or Quantitative Reasoning (QR) – is a "habit of mind," competency, and comfort in working with numerical data. Individuals with strong QL skills possess the ability to reason and solve quantitative problems from a wide array of authentic contexts and everyday life situations. They understand and can create sophisticated arguments supported by quantitative evidence and they can clearly communicate those arguments in a variety of formats (using words, tables, graphs, mathematical equations, etc., as appropriate).

Evaluators are encouraged to assign a zero to any work sample or collection of work that does not meet benchmark (cell one) level performance.

	Capstone 4	**Milestones** 3	2	1
Interpretation *Ability to explain information presented in mathematical forms (e.g., equations, graphs, diagrams, tables, words)*	Provides accurate explanations of information presented in mathematical forms. Makes appropriate inferences based on that information. *For example, accurately explains the trend data shown in a graph and makes reasonable predictions regarding what the data suggest about future events.*	Provides accurate explanations of information presented in mathematical forms. *For instance, accurately explains the trend data shown in a graph.*	Provides somewhat accurate explanations of information presented in mathematical forms, but occasionally makes minor errors related to computations or units. *For instance, accurately explains trend data shown in a graph, but may miscalculate the slope of the trend line.*	Attempts to explain information presented in mathematical forms, but draws incorrect conclusions about what the information means. *For example, attempts to explain the trend data shown in a graph, but will frequently misinterpret the nature of that trend, perhaps by confusing positive and negative trends.*
Representation *Ability to convert relevant information into various mathematical forms (e.g., equations, graphs, diagrams, tables, words)*	Skillfully converts relevant information into an insightful mathematical portrayal in a way that contributes to a further or deeper understanding.	Competently converts relevant information into an appropriate and desired mathematical portrayal.	Completes conversion of information but resulting mathematical portrayal is only partially appropriate or accurate.	Completes conversion of information but resulting mathematical portrayal is inappropriate or inaccurate.
Calculation	Calculations attempted are essentially all successful and sufficiently comprehensive to solve the problem. Calculations are also presented elegantly (clearly, concisely, etc.)	Calculations attempted are essentially all successful and sufficiently comprehensive to solve the problem.	Calculations attempted are either unsuccessful or represent only a portion of the calculations required to comprehensively solve the problem.	Calculations are attempted but are both unsuccessful and are not comprehensive.
Application / Analysis *Ability to make judgments and draw appropriate conclusions based on the quantitative analysis of data, while recognizing the limits of this analysis*	Uses the quantitative analysis of data as the basis for deep and thoughtful judgments, drawing insightful, carefully qualified conclusions from this work.	Uses the quantitative analysis of data as the basis for competent judgments, drawing reasonable and appropriately qualified conclusions from this work.	Uses the quantitative analysis of data as the basis for workmanlike (without inspiration or nuance, ordinary) judgments, drawing plausible conclusions from this work.	Uses the quantitative analysis of data as the basis for tentative, basic judgments, although is hesitant or uncertain about drawing conclusions from this work.
Assumptions *Ability to make and evaluate important assumptions in estimation, modeling, and data analysis*	Explicitly describes assumptions and provides compelling rationale for why each assumption is appropriate. Shows awareness that confidence in final conclusions is limited by the accuracy of the assumptions.	Explicitly describes assumptions and provides compelling rationale for why assumptions are appropriate.	Explicitly describes assumptions.	Attempts to describe assumptions.
Communication *Expressing quantitative evidence in support of the argument or purpose of the work (in terms of what evidence is used and how it is formatted, presented, and contextualized)*	Uses quantitative information in connection with the argument or purpose of the work, presents it in an effective format, and explicates it with consistently high quality.	Uses quantitative information in connection with the argument or purpose of the work, though data may be presented in a less than completely effective format or some parts of the explication may be uneven.	Uses quantitative information, but does not effectively connect it to the argument or purpose of the work.	Presents an argument for which quantitative evidence is pertinent, but does not provide adequate explicit numerical support. (May use quasi-quantitative words such as "many," "few," "increasing," "small," and the like in place of actual quantities.)

	Capstone 4
Interpretation *Ability to explain information presented in mathematical forms (e.g., equations, graphs, diagrams, tables, words)*	Provides accurate explanations of information presented in mathematical forms. Makes appropriate inferences based on that information. *For example, accurately explains the trend data shown in a graph and makes reasonable predictions regarding what the data suggest about future events.*
Representation *Ability to convert relevant information into various mathematical forms (e.g., equations, graphs, diagrams, tables, words)*	Skillfully converts relevant information into an insightful mathematical portrayal in a way that contributes to a further or deeper understanding.
Calculation	Calculations attempted are essentially all successful and sufficiently comprehensive to solve the problem. Calculations are also presented elegantly (clearly, concisely, etc.)
Application / Analysis *Ability to make judgments and draw appropriate conclusions based on the quantitative analysis of data, while recognizing the limits of this analysis*	Uses the quantitative analysis of data as the basis for deep and thoughtful judgments, drawing insightful, carefully qualified conclusions from this work.
Assumptions *Ability to make and evaluate important assumptions in estimation, modeling, and data analysis*	Explicitly describes assumptions and provides compelling rationale for why each assumption is appropriate. Shows awareness that confidence in final conclusions is limited by the accuracy of the assumptions.
Communication *Expressing quantitative evidence in support of the argument or purpose of the work (in terms of what evidence is used and how it is formatted, presented, and contextualized)*	Uses quantitative information in connection with the argument or purpose of the work, presents it in an effective format, and explicates it with consistently high quality.

Homework

- Full list of homework assignments and dates will be posted on blackboard by Monday.
- We will review these on Monday

Homework

- Hand in your homework on time!

Homework

- Hand in your homework on time!
- Hand in your homework on time!

Homework

- Hand in your homework on time!
- Hand in your homework on time!
- Hand in your homework on time!

Homework

- Hand in your homework on time!
- Hand in your homework on time!
- Hand in your homework on time!
- Hand in your homework on time!

Homework

- Hand in your homework on time!
- Hand in your homework on time!
- Hand in your homework on time!
- Hand in your homework on time!
- Sunday by midnight for Monday's class. Tuesday by midnight for Wednesday's class.

Plan for today

- Introduce myself
- Review syllabus
- Introduce yourselves
- Homework

About you

- What program?
- What year?
- Computer programming? What language?
- Do you have a laptop available?
- Why are you taking the course?

Plan for today

- Introduce myself
- Review syllabus
- Introduce yourselves
- Homework

Python

- Is a general purpose language.
- Is a high-level language.
- Supports multiple paradigms (object-oriented, imperative, functional, etc).
- Is free, open-source software.

How to find help

- Google “Python blah” where blah is a command you are interested in.
    - Look for pages from source forge. These are often good.
- Google “Python tutorial” and copy their examples.
- Google “jupyter notebook examples” and see what is out there.
- Explore one of the innumerable courses on python. e.g. this one. or this one. or this one. or this one. (all obtained by typing “intro python course” into google.)
- Last resort is to ask me. Why? Because the goal is for you to learn to program. The only way to do that is to learn how to learn to program.

Data science in Python

- Python can be used at the command-line.
    - e.g. on a mac, open “Terminal”
- Also, using other methods.
    - We will use Jupiter (iPython) notebooks: http://jupyter.org/
    - For those who do not have python already, recommend **anaconda** (from continuum)
    - Instructions here: http://jupyter.readthedocs.io/en/latest/install.html

Data science in Python

- Which version of python?
    - Download the newest. 3.x

Homework

- Install Jupyter and Python
- Demo that it is working by designing a tutorial that covers the topics of strings, lists, sorting, and dicts in a Jupyter notebook.
- Your goals are: to cover core competencies in creating and manipulating these variables
- You may not: copy tutorials.
- You may: Reference existing tutorials / talk to friends, but you should produce interestingly novel examples.
- You must: Cite any sources that you reference. Hand in your own work.
- You will: be graded based on a 0-3 scale. We will talk more about this on Monday.
- Please submit the notebook after exporting to html format

Homework Bonus

Go online and find your favorite three examples of data science

http://nbviewer.jupyter.org/github/buzzfeednews/2014-08-st-louis-county-segregation/blob/master/notebooks/segregation-analysis.ipynb

https://nbviewer.jupyter.org/github/brianckeegan/Bechdel/blob/master/Bechdel_test.ipynb

Fun stuff to get started

- Python wiki page
- Join the Python community
- Read about “What is data science”?
- Read about machine learning & artificial intelligence
- Think about: How is data science different from machine learning/statistics/AI?