Capstone: Systems for Interactive Data Analysis

Spring 2019

Instructor: Azza Abouzied azza@nyu.edu Meetings: Once a week @ A2 177. Time TBD

Course Description

Research typically involves:

- **Deeply understanding a problem**: who faces a certain challenge and why?
- **(Re)-Defining the problem**: problems can seem tricky from one perspective. Solving problems is often about changing your perspective. This is usually the most time-consuming aspect of research: decomposing what was perceived as hard, too broad, or too complex into an elegant and tractable problem.
- **Understanding the space of solutions**: Why certain solutions don't work? Perhaps they incorrectly characterized the problem? What techniques did others attempt that we can apply to the current situation? Why should these techniques work or not work?
- **Solving the problem and evaluating the solution**: This should follow naturally if you got the previous steps right!

The capstone research seminar and courses will follow this research progression and so most of your time will be spent on tackling the first three steps with the last capstone semester dedicated to the last step.

The space of problems we are interested in will be **limited to interactive data analysis**. This is my area of expertise but it is also a rich field that brings together many research communities together: data systems (DS), machine learning (ML) and human computer interaction (HCI).

This space focuses on how can we help users, from the data-science, python-hacking, stats-shooting experts to your grandfather[1], understand their data and its features as well as articulate their analysis needs from fun exploration to constructing robust predictive models. Thus, we need to build interactive data-analysis tools that work with the strengths and limitations of human perception and cognition but are powerful and efficient.

As your understanding of this space evolves, you will eventually be able to articulate a clear, well-defined research problem with any emphasis that interests you: The project can be around explaining ML models, cleaning messy data, scraping forums, generating stories, to tools for huddling around data.

With that end goal in sight, in the first two semesters you will read at least one research paper a week, critique it and occasionally discuss/present the paper to the research group.

Grading

Capstone Seminar
Research Critiques & Presentations (peer reviewed) 80%
Capstone Seminar Write-up (Lit survey heavy) 20%

Capstone Course 1
Focused Research Readings & Presentations (peer reviewed) 40%

[1] It is usually harder to explain tech stuff to your grandfather than your Android-slinging grandmother.

Capstone Proposal & Project Report 60%

Capstone Course 2
Project Implementation 40%
Project Evaluation 40%
Project Write-up (in research paper form) 20%

If you miss a reading or presentation, you will receive no grade for that assignment and there will be no room for making up the lost marks. In the first semester, all students read the same paper and critique it every week. By the end of the seminar, you would have defined an area of interest. For example, *"I'm interested in helping people analyze personal, multimodal data or explore the knowledge graph", etc.* In capstone course 1, you can focus your readings on your area of interest.

Note that you are co-graded on the seminar and the final capstone report and presentation by a secondary advisor.

Readings by Theme

This is a starting list that can change over the course of the semester.

General

- Jeffrey Heer. **Agency plus automation: Designing artificial intelligence into interactive systems**. In: *Proceedings of the National Academy of Sciences* 116.6 (2019), pp. 1844–1850. eprint: https://www.pnas.org/content/116/6/1844.full.pdf. URL: https://www.pnas.org/content/116/6/1844.
- Kristi Morton et al. **Support the Data Enthusiast: Challenges for Next-generation Data-analysis Systems**. In: *Proc. VLDB Endow.* 7.6 (Feb. 2014), pp. 453–456. URL: http://dx.doi.org/10.14778/2732279.2732282.
- Jeffrey Heer and Sean Kandel. **Interactive Analysis of Big Data**. In: *XRDS* 19.1 (Sept. 2012), pp. 50–54. URL: http://doi.acm.org/10.1145/2331042.2331058.

Visualization

- D. Moritz et al. **Formalizing Visualization Design Knowledge as Constraints: Actionable and Extensible Models in Draco**. In: *IEEE Transactions on Visualization and Computer Graphics* 25.1 (Jan. 2019), pp. 438–448. URL: https://idl.cs.washington.edu/files/2019-Draco-InfoVis.pdf.
- Bahador Saket et al. **Beyond Heuristics: Learning Visualization Design**. 2018. eprint: arXiv:1807.06641. URL: https://arxiv.org/pdf/1807.06641.pdf.
- Arvind Satyanarayan, Dominik Moritz, et al. **Vega-Lite: A Grammar of Interactive Graphics**. In: *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)* (2017). URL: http://idl.cs.washington.edu/papers/vega-lite.
- Zhicheng Liu and Jeffrey Heer. **The Effects of Interactive Latency on Exploratory Visual Analysis**. In: *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)* (2014). URL: http://idl.cs.washington.edu/papers/latency.
- Arvind Satyanarayan, Kanit Wongsuphasawat, and Jeffrey Heer. **Declarative Interaction Design for Data Visualization**. In: *ACM User Interface Software & Technology (UIST)*. 2014. URL: http://idl.cs.washington.edu/papers/reactive-vega.
- Manasi Vartak et al. **SeeDB: Automatically Generating Query Visualizations**. In: *Proc. VLDB Endow.* 7.13 (Aug. 2014), pp. 1581–1584. URL: http://dx.doi.org/10.14778/2733004.2733035.

Example-Driven Interfaces & Mixed Initiative User Interfaces

- Maeda F. Hanafi et al. **SEER: Auto-Generating Information Extraction Rules from User-Specified Examples**. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. Denver, Colorado, USA, 2017, pp. 6672–6682. URL: http://doi.acm.org/10.1145/3025453.3025540.
- Jeffrey Heer, Joseph M Hellerstein, and Sean Kandel. **Predictive Interaction for Data Transformation**. In: *CIDR*. 2015. URL: https://idl.cs.washington.edu/files/2015-PredictiveInteraction-CIDR.pdf.
- Mikaël Mayer et al. **User Interaction Models for Disambiguation in Programming by Example**. In: *Proceedings of the 28th Annual ACM Symposium on User Interface Software Technology*. UIST '15. Charlotte, NC, USA, 2015, pp. 291–301. URL: http://doi.acm.org/10.1145/2807442.2807459.

- Azza Abouzied, Joseph Hellerstein, and Avi Silberschatz. **DataPlay: Interactive Tweaking and Example-driven Correction of Graphical Database Queries**. In: *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology*. UIST '12. Cambridge, Massachusetts, USA, 2012, pp. 207–218. URL: http://doi.acm.org/10.1145/2380116.2380144.
- Sumit Gulwani. **Automating String Processing in Spreadsheets Using Input-output Examples**. In: *SIGPLAN Not.* 46.1 (Jan. 2011), pp. 317–330. URL: http://doi.acm.org/10.1145/1925844.1926423.
- Sean Kandel et al. **Wrangler: Interactive Visual Specification of Data Transformation Scripts**. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '11. Vancouver, BC, Canada, 2011, pp. 3363–3372. URL: http://doi.acm.org/10.1145/1978942.1979444.
- Eric Horvitz. **Principles of Mixed-initiative User Interfaces**. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '99. Pittsburgh, Pennsylvania, USA, 1999, pp. 159–166. URL: http://doi.acm.org/10.1145/302979.303030.

Machine Learning & Interpretability

- Chris Olah et al. **The Building Blocks of Interpretability**. In: *Distill* (2018). URL: https://distill.pub/2018/building-blocks.
- K. Wongsuphasawat et al. **Visualizing Dataflow Graphs of Deep Learning Models in TensorFlow**. In: *IEEE Transactions on Visualization and Computer Graphics* 24.1 (Jan. 2018), pp. 1–12.
- Shan Carter and Michael Nielsen. **Using Artificial Intelligence to Augment Human Intelligence**. In: *Distill* (2017). URL: https://distill.pub/2017/aia.
- Michael Nielsen. **Reinventing Explanation**. 2014. URL: http://michaelnielsen.org/reinventing_explanation/.
- Jason Chuang et al. **Interpretation and Trust: Designing Model-driven Visualizations for Text Analysis**. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '12. Austin, Texas, USA, 2012, pp. 443–452. URL: http://doi.acm.org/10.1145/2207676.2207738.

Debugging

- Jane Hoffswell, Arvind Satyanarayan, and Jeffrey Heer. **Augmenting Code with In Situ Visualizations to Aid Program Understanding**. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Montreal QC, Canada, 2018, 532:1–532:12. URL: http://doi.acm.org/10.1145/3173574.3174106.
- Daniel W. Barowy, Dimitar Gochev, and Emery D. Berger. **CheckCell: Data Debugging for Spreadsheets**. In: *Proceedings of the 2014 ACM International Conference on Object Oriented Programming Systems Languages & Applications*. OOPSLA '14. Portland, Oregon, USA, 2014, pp. 507–523. URL: http://doi.acm.org/10.1145/2660193.2660207.
- Emma Tosch and Emery D. Berger. **SurveyMan: Programming and Automatically Debugging Surveys**. In: *Proceedings of the 2014 ACM International Conference on Object Oriented Programming Systems Languages & Applications*. OOPSLA '14. Portland, Oregon, USA, 2014, pp. 197–211. URL: http://doi.acm.org/10.1145/2660193.2660206.

Graphs

- Jane Hoffswell, Alan Borning, and Jeffrey Heer. **SetCoLa: High-Level Constraints for Graph Layout**. In: *Computer Graphics Forum (Proc. EuroVis)*. 2018. URL: http://idl.cs.washington.edu/files/2018-SetCoLa-EuroVis.pdf.
- Jeffrey Heer and Adam Perer. **Orion: A System for Modeling, Transformation and Visualization of Multidimensional Heterogeneous Networks**. In: *Information Visualization* 13.2 (Apr. 2014), pp. 111–133. URL: https://doi.org/10.1177/1473871612462152.

Uncertainty

- Dominik Moritz and Danyel Fisher. **What Users Don'T Expect About Exploratory Data Analysis on Approximate Query Processing Systems**. In: *Proceedings of the 2Nd Workshop on Human-In-the-Loop Data Analytics*. HILDA'17. Chicago, IL, USA, 2017, 9:1–9:4. URL: http://doi.acm.org/10.1145/3077257.3077258.
- Dominik Moritz, Danyel Fisher, et al. **Trust, but Verify: Optimistic Visualizations of Approximate Queries for Exploring Big Data**. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. Denver, Colorado, USA, 2017, pp. 2904–2915. URL: http://doi.acm.org/10.1145/3025453.3025456.
- Bolin Ding et al. **Sample + Seek: Approximating Aggregates with Distribution Precision Guarantee**. In: *Proceedings of the 2016 International Conference on Management of Data*. SIGMOD '16. San Francisco, California, USA, 2016, pp. 679–694. URL: http://doi.acm.org/10.1145/2882903.2915249.

Time

- Miro Mannino and Azza Abouzied. **Expressive Time Series Querying with Hand-Drawn Scale-Free Sketches**. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Montreal QC, Canada, 2018, 388:1–388:13. URL: http://doi.acm.org/10.1145/3173574.3173962.
- Christian Holz and Steven Feiner. **Relaxed Selection Techniques for Querying Time-series Graphs**. In: *Proceedings of the 22Nd Annual ACM Symposium on User Interface Software and Technology*. UIST '09. Victoria, BC, Canada, 2009, pp. 213–222. URL: http://doi.acm.org/10.1145/1622176.1622217.

Miscellaneous, Unusual but Cool

- Matthew Conlen and Jeffrey Heer. **Idyll: A Markup Language for Authoring and Publishing Interactive Articles on the Web**. In: *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. UIST '18. Berlin, Germany, 2018, pp. 977–989. URL: http://doi.acm.org/10.1145/3242587.3242600.
- Kelly Mack et al. **Characterizing Scalability Issues in Spreadsheet Software Using Online Forums**. In: *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI EA '18. Montreal QC, Canada, 2018, CS04:1–CS04:9. URL: http://doi.acm.org/10.1145/3170427.3174359.
- Haoci Zhang, Thibault Sellam, and Eugene Wu. **Mining Precision Interfaces From Query Logs**. In: *CoRR* abs/1712.00078 (2017). URL: http://arxiv.org/abs/1712.00078.
- Xiong Zhang and Philip J. Guo. **DS.Js: Turn Any Webpage into an Example-Centric Live Programming Environment for Learning Data Science**. In: *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*. UIST '17. Québec City, QC, Canada, 2017, pp. 691–702. URL: http://doi.acm.org/10.1145/3126594.3126663.
- S. Alspaugh et al. **Analyzing Log Analysis: An Empirical Study of User Log Mining**. In: *Proceedings of the 28th USENIX Conference on Large Installation System Administration*. LISA'14. Seattle, WA, 2014, pp. 53–68. URL: http://dl.acm.org/citation.cfm?id=2717491.2717495.
- Sumit Gulwani and Mark Marron. **NLyze: Interactive Programming by Natural Language for Spreadsheet Data Analysis and Manipulation**. In: *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*. SIGMOD '14. Snowbird, Utah, USA, 2014, pp. 803–814. URL: http://doi.acm.org/10.1145/2588555.2612177.
- Kuang Chen, Akshay Kannan, et al. **Shreddr: Pipelined Paper Digitization for Low-resource Organizations**. In: *Proceedings of the 2Nd ACM Symposium on Computing for Development*. ACM DEV '12. Atlanta, Georgia, 2012, 3:1–3:10. URL: http://doi.acm.org/10.1145/2160601.2160605.
- Mathias Eitz, James Hays, and Marc Alexa. **How Do Humans Sketch Objects?** In: *ACM Trans. Graph.* 31.4 (July 2012), 44:1–44:10. URL: http://doi.acm.org/10.1145/2185520.2185540.
- Vladimir G. Kim et al. **Exploring Collections of 3D Models Using Fuzzy Correspondences**. In: *ACM Trans. Graph.* 31.4 (July 2012), 54:1–54:11. URL: http://doi.acm.org/10.1145/2185520.2185550.
- Kuang Chen, Harr Chen, et al. **Usher: Improving Data Quality with Dynamic Forms**. In: *IEEE Trans. on Knowl. and Data Eng.* 23.8 (Aug. 2011), pp. 1138–1153. URL: http://dx.doi.org/10.1109/TKDE.2011.31.

In addition to course readings, we will also be having tutorials on data visualization and UI design.

Academic Integrity: As set forth in NYU Abu Dhabi's Academic Integrity Policy, the relationship between students and faculty at NYU Abu Dhabi is defined by a shared commitment to academic excellence and is grounded in an expectation of fairness, honesty, and respect, which are essential to maintaining the integrity of the community. Every student who enrolls and everyone who accepts an appointment as a member of the faculty or staff at NYU Abu Dhabi agrees to abide by the expectation of academic honesty.

The full policies and procedures relating to Academic Integrity may be found on the NYUAD Student Portal.