Data Visualization

Yong-Yeol (YY) Ahn
yyahn@iu.edu

Info East 130 (M) / 109 (W)
MW 4:00pm–5:15pm.
Office hours: W 9:15am-10am

Assistant Instructors
Elise Jing (jingy@iu.edu)
Jaehyuk Park (jp70@iu.edu)
Harika Putti (haputti@iu.edu)

COMMUNICATION

Announcements and other communication will be through Slack:

https://iu-dviz-course.slack.com

Slack allows both private (group) communications and public fora. The number one recommendation from the previous students for the new students was to learn to use Slack and get used to it as soon as possible.

Please join as soon as possible. You can join with your IU email adddress by visiting the following link.

https://join.slack.com/t/iu-dviz-course/signup

COURSE DESCRIPTION

From TV news to cutting-edge scientific papers, from a home office to the largest companies in the world, data visualization is extensively used to reveal patterns in data and to tell stories. More and more data is collected, and more and more decisions are made through data analysis. Data visualization is indispensible for understanding data, and thus is an essential skill for every knowledge worker. This course is an introduction to basic statistical data analysis and visualization. We will learn fundamentals of data visualization in the context of perception, integrity, design, statistics, types of

data, and visualization techniques. The hands-on exercises using the Python stack aim to equip you with practical data visualization skills and they will be an integral part of the course.

Relationships with S637 Information Visualization (IVMOOC): Compared with S637, this course is geared more towards producing fundamental statistical visualizations and exploratory data analysis, using the Python data science and visualization stack. Therefore, this course may be more suitable for students who pursue their careers in research, developement, and data analysis, and those who will directly analyze complex datasets.

COURSE OBJECTIVES

By the end of the course, you are expected to be able to understand, explain, and manipulate basic types of data, analyze them by applying basic exploratory visualization techniques, and create explanatory visualizations. You will also be able to evaluate the effectiveness of data visualizations based on the principles of human perception, design, types of data, and visualization techniques.

PREREQUISITES

This course's residential version is open to advanced undergraduate students as well as graduate students; online version of this course is open to graduate students. Because producing visualizations using Python data & visualization stack is an integral part of the course, it is required to have good understanding and working knowledge of programming, as well as working knowledge of using open-source libraries. It is also recommended to have basic understanding of mathematics, statistics, and Web (HTML, CSS, Javascript, and JSON).

For the undergraduate students, the prerequisites are:

· Programming foundations (I210 & I211, or equivalent).

The following courses are recommended:

· I123: Data Fluency

· I300: HCI/Interaction Design

· I308: Information Representation

· I360: Web Design

For self-assessment, visit the following link: http://bit.ly/dvizselfassess. Contact the instructor if you are uncertain about your background.

EXPECTATIONS AND REQUIREMENTS

(All sections) The final assessment will be mainly based on a mid-term exam and a final course project. The choice of your project topic can be guided by the instructors but you have freedom to choose. You are required to submit a final paper that contains detailed explanation of the visualization process and results as well as the visualization artifact itself (e.g. a visualization tool or a webpage) depending on the nature of your project.

(Residential course) You are expected to attend every class and engage in class discussions. You are not allowed to use your phone or computer during the class unless explicitly asked to do so. You are expected to read assigned reading materials prior to the class meetings if there is any. At the beginning of most class meetings, there will be an *in-class quiz* based on the assigned readings and materials from the previous classes. You are expected to complete all weekly assignments.

(Online) You are expected to complete all course modules and assignments. You are also expected to engage in discussions on Canvas and Slack.

BOOKS AND KEY MATERIALS

There is no required textbook, but we will mainly use materials from the following books:

1. Fundamentals of Data Visualization by Claus O. Wilke (book preview available online at `https://serialmentor.com/dataviz/`)

2. The Visual Display of Quantitative Information (2nd ed.) by E.R. Tufte: one of the foundational book on visualization. It contains a rich set of historical visualization, thoughtful discussion on visualization principles.

See also Visualization books on my wiki (`http://yyahnwiki.appspot.com/Information_visualization`).

If you are still in the process of learning Python basics, the following books and websites may be helpful for you:

Python and data analysis

1. `https://docs.python.org/3/`: Python 3 Official Documentation

2. `http://www.diveintopython3.net/index.html`: Dive Into Python by Mark Pilgrim

3. `http://work.thaslwanter.at/Stats/html/`: An introduction to statistics (with Python) by Thomas Haslwanter (available online): this book uses Python to explain basic statistics. It also contains a succinct tutorial for Python and data visualization using Python.

4. `http://www.learnpython.org`: A web-based interactive tutorial

5. `http://ipython.rossant.net`: Learning IPython for Interactive Computing and Data Visualization by Cyrille Rossant: Introduction to IPython as well as lots of advanced analysis

POLICIES

1. *Disabilities.* Every attempt will be made to accommodate qualified students with disabilities (e.g. mental health, learning, chronic health, physical, hearing, vision, neurological, etc.). You must have established your eligibility for support services through Disability Services for Students.

Note that services are confidential, may take time to put into place, and are not retroactive. Captions and alternate media for print materials may take three or more weeks to get produced. Please contact Disability Services for Students at `http://disabilityservices.indiana.edu` or 812-855-7578 as soon as possible if accommodations are needed. The office is located on the third floor, west tower, of the Wells Library (Room W302). Walk-ins are welcome 8 AM to 5 PM, Monday through Friday. You can also locate a variety of campus resources for students and visitors who need assistance at `http://www.iu.edu/~ada/index.shtml`.

2. *Sexual misconduct and Title IX.* As your instructor, one of my responsibilities is to create a positive learning environment for all students. Title IX and IU's Sexual Misconduct Policy prohibit sexual misconduct in any form, including sexual harassment, sexual assault, stalking, and dating and domestic violence. If you have experienced sexual misconduct, or know someone who has, the University can help. If you are seeking help and would like to speak to someone confidentially, you can make an appointment with:

   a) The Sexual Assault Crisis Services (SACS) at (812) 855-8900 (counseling services)

   b) Confidential Victim Advocates (CVA) at (812) 856-2469 (advocacy and advice services)

   c) IU Health Center at (812) 855-4011 (health and medical services)

   It is also important that you know that Title IX and University policy require me to share any information brought to my attention about potential sexual misconduct, with the campus Deputy Title IX Coordinator or IU's Title IX Coordinator. In that event, those individuals will work to ensure that appropriate measures are taken and resources are made available. Protecting student privacy is of utmost concern, and information will only be shared with those that need to know to ensure the University can respond and assist. I encourage you to visit *stopsexualviolence.iu.edu* to learn more.

3. *No electronics—laptops, tablets, and smartphones—may be used in class*, unless the usage is specifically requested by the instructors (e.g. Top Hat). It has been shown that using laptops in class hurts learning, *even if* you are using it to take notes. If you must have electronics due to a special reason, please obtain permission beforehand.

4. *Be honest.* Your assignments and papers should be your own work. If you find useful resources for your assignments, share them and cite them. If your friends helped you, acknowledge them. You should feel free to discuss both online and offline, but do not show your code directly. Any cases of academic misconduct (cheating, fabrication, plagiarism, etc) will be reported to the School and the Dean of Students, following the standard procedure. *Cheating is not cool.*

5. *You have the responsibility of backing up all your data and code.* Always use at least Box, Dropbox, or Google Drive. Ideally, learn version control systems and use `https://github.iu.edu` or `https://github.com`. Loss of data, code, or papers due to various reasons (e.g. malfunction of your laptop) is not an acceptable excuse for delayed or missing submission.

6. *Inform your excused absences prior to class.* Please contact the instructor until the previous day for an excused absence.

7. *Late assignments.* There will be 10% late penalty for the late assignments.

8. If you have any mental issues, don't hesistate to contact me or IU's Counseling and Psychological Services, which provides free counseling sessions.

GRADING

· Attendance, Quiz, and Participation: 20%

· Assignments: 30%

· Mid-term: 20%

· Final project: 30%

COURSE SCHEDULE

Week 1 (8/20-): Why visualization? | Overview of Python visualization libraries

· J. Heer et al. A Tour through the Visualization Zoo. `https://queue.acm.org/detail.cfm?id=1805128`
· J. VanderPlas, The Python Visualization Landscape. `https://youtu.be/FytuB8nFHPQ`

Week 2 (8/27-): History and integrity | Plotting basics with the Snow's map

· E.R. Tufte, The Visual Display of Quantitative Information, Ch. 1–2.
· C.O. Wilke, Fundamentals of Data Visualization Ch. 1 (`https://serialmentor.com/dataviz/introduction.html`).

Week 3 (9/3-): Labor day | Web, Declarative vs. Procedural visualization

Week 4 (9/10-): Perception | Perception experiments

· C.G. Healey, Perception in Visualization, `https://www.csc2.ncsu.edu/faculty/healey/PP/index.html`
· B. Wong, Color Coding, Nature Methods (2010).
· B. Wong, Avoiding color, Nature Methods (2011).
· C.O. Wilke, Fundamentals of Data Visualization Ch. 4 Color scales (`https://serialmentor.com/dataviz/color-basics.html`).
· C.O. Wilke, Fundamentals of Data Visualization Ch. 15 Common pitfalls of color use (`https://serialmentor.com/dataviz/color-pitfalls.html`).

Week 5 (9/17-): Design | Colors, Vector, and Bitmaps

· B. Wong, Gestalt Principles I & II, Nature Methods (2010).

· E.R. Tufte, The Visual Display of Quantitative Information, Ch. 4.
· S. Bateman et al., Useful Junk? The Effects of Visual Embellishment on Comprehension and Memorability of Charts, CHI'10.
· C.O. Wilke, Fundamentals of Data Visualization Ch. 18–21 (`https://serialmentor.com/dataviz/optimize-data-signal.html`).

Week 6 (9/24-): Data Types and 1-D data | Tidy data and 1D data

· H. Wickham, Tidy Data, Journal of Statistical Software, `https://vita.had.co.nz/papers/tidy-data.pdf`
· C.O. Wilke, Fundamentals of Data Visualization Ch. 14 (`https://serialmentor.com/dataviz/overlapping-points.html`).

Week 7 (10/1-): Histogram and Boxplot | Histogram and CDF

· C.O. Wilke, Fundamentals of Data Visualization Ch. 6–7 (`https://serialmentor.com/dataviz/overlapping-points.html`).

Week 8 (10/8-): Estimation | Estimation

· C.O. Wilke, Fundamentals of Data Visualization Ch. 8–9 (`https://serialmentor.com/dataviz/overlapping-points.html`).

Week 9 (10/15-): Logscale and Beyond 1-D | Logscale and CCDF

· Khan Academy: Logarithmic scale with Vi Hart (`https://www.khanacademy.org/math/algebra2/exponential-and-logarithmic-functions/logarithmic-scale`).

Week 10 (10/22-): High-dimensional data | High-dimensional data

· C.O. Wilke, Fundamentals of Data Visualization Ch. 11 (`https://serialmentor.com/dataviz/visualizing-associations.html`).
· 3Blue1Brown, Eigenvectors and eigenvalues `https://www.youtube.com/watch?v=PFDu9oVAE-g`.
· Victor Powell, PCA `http://setosa.io/ev/principal-component-analysis/`.
· L. van der Maaten & G. Hinton, Visualizing data using t-SNE, JMLR 2008 `http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf`.

Week 11 (10/29-): Mid-term

Week 12 (11/5-): Maps | Maps

· Vsauce, What does earth look like? `https://youtu.be/2lR7s1Y6Zig`

· Vox, Why all world maps are wrong `https://youtu.be/kIID5FDi2JQ`

Week 13 (11/12-): Text and Networks | Text and networks

· J. Harris, Word clouds considered harmful, `http://www.niemanlab.org/2011/10/word-clouds-considered-harmful/`.
· The Observatory of Economic Complexity, `https://atlas.media.mit.edu/en/profile/country/usa/`.

Week 14 (11/19-): Thanksgiving break

Week 15 (11/26-): Project Hacks

Week 16 (12/3-): Final project presentation