Data & Visual Analytics

Duen Horng (Polo) Chau
Associate Professor, College of Computing
Associate Director, MS Analytics
Georgia Tech

Mahdi Roozbahani
Lecturer, Computational Science & Engineering, Georgia Tech
Founder of Filio, a visual asset management platform
Course TAs  Be very very nice to them!

Sam Stentz (Co-Head TA)
Dhaval Desai (Co-Head TA)
Soo Hyung Park
Edmund Chen
Riya Bakhtiani
Jianyuan Lu
The course focuses on working with big data.

(Also the focus of Polo’s research group)
poloclub.gatech.edu

Scalable. Interactive. Interpretable.

At Georgia Tech, we innovate scalable, interactive, and interpretable tools that amplify human's ability to understand and interact with billion-scale data and machine learning models. Our current research thrusts: human-centered AI (interpretable, fair, safe AI; adversarial ML); large graph visualization and mining; cybersecurity; and social good (health, energy).

Current Polo Club members and recent alum. See all alum.
Internet
50 Billion Web Pages
Facebook
2 Billion Users
Many More

Twitter
Who-follows-whom (500 million users)

Amazon
Who-buys-what (120 million users)

AT&T Cellphone Network
Who-calls-whom (100 million users)

Protein-protein interactions
200 million possible interactions in human genome

"Big Data" Analyzed

<table>
<thead>
<tr>
<th>Graph</th>
<th>Nodes</th>
<th>Edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>YahooWeb</td>
<td>1.4 Billion</td>
<td>6 Billion</td>
</tr>
<tr>
<td>Symantec Machine-File Graph</td>
<td>1 Billion</td>
<td>37 Billion</td>
</tr>
<tr>
<td>Twitter</td>
<td>104 Million</td>
<td>3.7 Billion</td>
</tr>
<tr>
<td>Phone call network</td>
<td>30 Million</td>
<td>260 Million</td>
</tr>
</tbody>
</table>

We also work with small data. Small data also needs love.
7±2

Number of **items** an average human holds in **working memory**

*George Miller, 1956*
How to do that?

**COMPUTATION**

**+**

**HUMAN INTUITION**
Or, to ride the AI wave...

**ARTIFICIAL INTELLIGENCE**

**+**

**HUMAN INTELLIGENCE**
How to do that?

<table>
<thead>
<tr>
<th>COMPUTATION</th>
<th>INTERACTIVE VIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic</td>
<td>User-driven; iterative</td>
</tr>
<tr>
<td>Summarization, clustering, classification</td>
<td>Interaction, visualization</td>
</tr>
<tr>
<td>&gt;Millions of nodes</td>
<td>Thousands of nodes</td>
</tr>
</tbody>
</table>

Both develop methods for making sense of network data
How to do that?

<table>
<thead>
<tr>
<th>COMPUTATION</th>
<th>INTERACTIVE VIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic</td>
<td></td>
</tr>
<tr>
<td>Summarization, clustering, classification</td>
<td></td>
</tr>
<tr>
<td>&gt;Millions of nodes</td>
<td></td>
</tr>
</tbody>
</table>
How to do that?

<table>
<thead>
<tr>
<th>COMPUTATION</th>
<th>INTERACTIVE VIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic</td>
<td></td>
</tr>
<tr>
<td>Summarization, clustering, classification</td>
<td></td>
</tr>
<tr>
<td>&gt;Millions of nodes</td>
<td></td>
</tr>
</tbody>
</table>
How to do that?

<table>
<thead>
<tr>
<th>COMPUTATION</th>
<th>INTERACTIVE VIS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>User-driven; iterative</td>
</tr>
<tr>
<td></td>
<td>Interaction, visualization</td>
</tr>
<tr>
<td></td>
<td>Thousands of nodes</td>
</tr>
</tbody>
</table>
**How to do that?**

<table>
<thead>
<tr>
<th><strong>COMPUTATION</strong></th>
<th><strong>INTERACTIVE VIS</strong></th>
</tr>
</thead>
</table>
| ![Diagram](image)

- User-driven; iterative
- Interaction, visualization
- Thousands of nodes
How to do that?

<table>
<thead>
<tr>
<th>COMPUTATION</th>
<th>INTERACTIVE VIS</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Diagram of a network" /></td>
<td><img src="image2.png" alt="Diagram of a network" /></td>
</tr>
</tbody>
</table>

**COMPUTATION**
- Automatic
- User-driven; iterative
- Interaction, visualization

**INTERACTIVE VIS**
- User-driven; iterative
- Interaction, visualization
- Thousands of nodes
Our Approach for Big Data Analytics

<table>
<thead>
<tr>
<th>DATA MINING</th>
<th>HCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic</td>
<td>User-driven; iterative</td>
</tr>
<tr>
<td>Summarization, clustering, classification</td>
<td>Interaction, visualization</td>
</tr>
<tr>
<td>&gt;Millions of items</td>
<td>Thousands of items</td>
</tr>
</tbody>
</table>

Our research combines the **Best of Both Worlds**
Our mission & vision:

**Scalable, interactive, usable**

tools for big data analytics
“Computers are incredibly fast, accurate, and stupid.

Human beings are incredibly slow, inaccurate, and brilliant.

Together they are powerful beyond imagination.”

(Einstein might or might not have said this.)
## Logistics

<table>
<thead>
<tr>
<th><strong>Course website</strong> (policies, syllabus, schedule, etc.)</th>
<th><a href="https://poloclub.github.io/cse6242-2021fall-campus/">https://poloclub.github.io/cse6242-2021fall-campus/</a> (link also available on Canvas)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Discussion, Q&amp;A, find teammates</strong></td>
<td><strong>Piazza</strong> (link/tab available on Canvas)</td>
</tr>
<tr>
<td></td>
<td>Make sure you’re in the right Piazza! (CSE-6242-O01, CSE-6242-OAN have their Piazza forums too)</td>
</tr>
<tr>
<td><strong>Assignment Submission</strong></td>
<td><strong>Canvas/Gradescope</strong></td>
</tr>
</tbody>
</table>
Course Homepage

For syllabus, schedule, projects, datasets, etc.

If you Google “cse6242”, you will see many matches. Make sure you click the correct site!
Join Piazza ASAP via canvas.gatech.edu

Announcements and Discussion

We use Piazza for all announcements and discussion. Everyone must join this class's Piazza (link available on Canvas). Double check that you are joining the correct Piazza! There are multiple concurrent course sections with the same name and course number taking place, e.g., online for OMSA and OMSCS, and campus for Atlanta-based students.

The fastest way to get help with homework assignments is to post your questions on Piazza. That way, not only our TAs and instructor can help, your peers can too.

If you prefer that your question addresses to only our TAs and the instructor, you can use the private post feature (i.e., check the "Individual Students(s) / Instructors(s)" radio box).

While we welcome everyone to share their experiences in tackling issues and helping each other out, but please do not post your answers, as that may affect the learning experience of your fellow classmates.

For special cases such as failed submissions due to system errors, missing grades, failed file uploads, emergencies that prevent you from submitting, personal issues, you can contact the staff using a private Piazza post.

Canvas will be used for submission of assignments and projects, but not for announcements or discussion.
Important to join Piazza because...

- We will announce events related to this class and data science in general
  - Distinguished lectures, seminars
  - Hackathons
  - Company recruitment events
Since course is remote, add your photo

If you need help cropping headshot photo into square, use **Magic Crop** ([https://poloclub.github.io/magic-crop/](https://poloclub.github.io/magic-crop/))
Course Goals
What is **Data & Visual Analytics?**
What is Data & Visual Analytics?

No formal definition!
What is **Data & Visual Analytics**?

No formal definition!

**Polo’s definition:**
the *interdisciplinary* science of combining *computation techniques* and *interactive visualization* to transform and model data to aid *discovery, decision making, etc.*
What are the “ingredients”?
What are the “ingredients”?

Need to worry (a lot) about: storage, complex system design, scalability of algorithms, visualization techniques, interaction techniques, statistical tests, etc.

Wasn’t this complex before this big data era. Why?
THE WORLD OF DATA

- NUMBER OF EMAILS SENT EVERY SECOND: 2.9 MILLION
- DATA CONSUMED BY HOUSEHOLDS EACH DAY: 375 MEGABYTES
- VIDEO UPLOADED TO YOUTUBE EVERY MINUTE: 20 HOURS
- DATA PER DAY PROCESSED BY GOOGLE: 24 PETABYTES
- TWEETS PER DAY: 50 MILLION
- TOTAL MINUTES SPENT ON FACEBOOK EACH MONTH: 700 BILLION
- DATA SENT AND RECEIVED BY MOBILE INTERNET USERS: 1.3 EXABYTES
- PRODUCTS ORDERED ON AMAZON PER SECOND: 72.9 ITEMS

IN THE 21ST CENTURY, we live a large part of our lives online. Almost everything we do is reduced to bits and sent through cables around the world at light speed. But just how much data are we generating? This is a look at just some of the amazing numbers of data that human beings create every single day.

http://spanning.com/blog/choosing-between-storage-based-and-unlimited-storage-for-cloud-data-backup/
What is **big data**? Why care?

Many businesses are based on big data.

**Search engines**: rank webpages, predict what you’re going to type

**Advertisement**: infer what you like, based on what your friends like; show relevant ads

**E-commerce**: recommends movies/products (e.g., Netflix, Amazon)

**Health IT**: patient records (EMR)

**Finance**
Good news! Many jobs!

Most companies are looking for “data scientists”

The data scientist role is critical for organizations looking to extract insight from information assets for ‘big data’ initiatives and requires a broad combination of skills that may be fulfilled better as a team

- Gartner (http://www.gartner.com/it-glossary/data-scientist)

Breadth of knowledge is important. This course helps you learn some important skills.
Course Schedule
(Analytics Building Blocks)

Collection
Cleaning
Integration
Analysis
Visualization
Presentation
Dissemination
Building blocks. Not Rigid “Steps”.

<table>
<thead>
<tr>
<th>Collection</th>
<th>Can skip some</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cleaning</td>
<td>Can go back (two-way street)</td>
</tr>
<tr>
<td>Integration</td>
<td></td>
</tr>
<tr>
<td>Analysis</td>
<td>• Data types inform visualization design</td>
</tr>
<tr>
<td>Visualization</td>
<td>• Data size informs choice of algorithms</td>
</tr>
<tr>
<td>Presentation</td>
<td>• Visualization motivates more data cleaning</td>
</tr>
<tr>
<td>Dissemination</td>
<td>• Visualization challenges algorithm assumptions</td>
</tr>
<tr>
<td></td>
<td>e.g., user finds that results don’t make sense</td>
</tr>
</tbody>
</table>
Course Goals

• Learn **visual** and **computation** techniques and use them in **complementary** ways

• Gain a **breadth** of knowledge

• Learn **practical** know-how by working on **real data & problems**
Grading

- [50%] 4 homework assignments
  - End-to-end analysis
  - Techniques (computation and vis)
  - “Big data” tools, e.g., Hadoop, Spark, etc.
- [50%] Group project — 4 to 6 people
- [Bonus points] Pop quizzes
  - Multiple over semester; ~10min each during DVA Live (except Q students)
  - 1% course grade each; lowest score dropped
- No Exams
Policies. Very Important!
(on course website)

Grading, plagiarism, collaboration, late submission, and the “warnings” about the difficulty this course
From Previous Classes…

• Class projects turned into papers at top conferences (KDD, IUI, etc.)
• Projects as portfolio pieces on CV
• Increased job and internship opportunities
• Former students sent me “thank you” notes
Aurigo: An Interactive Tour Planner for Personalized Itineraries

Alexandre Yahi; Antoine Chassang; Louis Raynaud; Hugo Duthil; Duen Horng (Polo) Chau
Georgia Institute of Technology
{alexandre.yahi, antoine.chassang, l.raynaud, hduthil, polo}@gatech.edu

ABSTRACT
Planning personalized tour itineraries is a complex and challenging task for both humans and computers. Doing it manually is time-consuming; approaching it as an optimization problem is computationally NP hard. We present Aurigo, a tour planning system combining a recommendation algorithm with interactive visualization to create personalized itineraries. This hybrid approach enables Aurigo to take into account both quantitative and qualitative preferences of the user. We conducted a within-subject study with 10 participants, which demonstrated that Aurigo helped them find points of interest quickly. Most participants chose Aurigo over Google Maps as their preferred tools to create personalized itineraries. Aurigo may be integrated into review websites or social networks, to leverage their databases of reviews and ratings and provide better itinerary recommendations.

Author Keywords
User Interfaces; Visualization; Recommendation; Tour itinerary planning

ACM Classification Keywords
(e.g. HCI): User interfaces
ISPARK: Interactive Visual Analytics for Fire Incidents and Station Placement

Subhajit Das, Andrea McCarter, Joe Minieri, Nandita Damaraju, Sriram Padmanabhan, Duen Horng (Polo) Chau
Georgia Tech
Atlanta, GA, USA
{das, andream, jminieri, nandita, sriramp, polo}@gatech.edu

ABSTRACT
In support of helping to reduce the response time of firefighters, and thus deaths, injuries, and property loss due to fires, we introduce ISPARK. The ISPARK system determines where fire stations should be located, analyzes the primary causes of fires, the existing infrastructure, and response times, by using visualizations which show the GIS mapping of fire stations on a dashboard. Incidents and response times are shown as additional layers, with clustering of fire incidents to determine predicted fire station locations, forecasting of fire incidents using regression, causal infrastructure, and personnel analysis, creating an interactive, multi-faceted method for locating fire stations. A comparison of urban and rural fire incident response times is another dimension of this study. We demonstrate ISPARK’s usage and benefits using a publicly available dataset describing 300,000 fire incidents in the states of Massachusetts and Maine. ISPARK is generalizable to other geographic areas.

Figure 1: Screenshot of ISPARK showing actual (pink) and predicted (green) fire station locations in Maine determined by our approach, using coordinates with actual driving distances from fire stations to actual fire incidents. Fire incidents are shown as small yellow dots. ISPARK reduces the average...
PASSAGE: A Travel Safety Assistant With Safe Path Recommendations For Pedestrians

Matthew Garvey
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332, USA
mgarvey@gatech.edu

Meghna Natraj
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332, USA
mnatraj@gatech.edu

Nilaksh Das
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332, USA
nilakshdas@gatech.edu

Bhanu Verma
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332, USA
bhanuverma@gatech.edu

Jiaxing Su
College of Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA
jiaxingsu@gatech.edu

Abstract
Atlanta has consistently ranked as one of the most dangerous cities in America with over 2.5 million crime events recorded within the past six years. People who commute by walking are highly susceptible to crime here. To address this problem, we have developed a mobile application, PASSAGE, using crime data to find "safe paths" through Atlanta. The user inputs their starting and ending locations, and the application recommends safe paths. Users can also add other users as friends to share paths and safety tips.

Figure 1: Paths recommended by PASSAGE
“I feel like the concepts from your class are like a rite of passage for an aspiring data scientist. Assignments lead to a feelings of accomplishment and truly progressing in my area of passion.”

“…thank you for the materials taught in DVA. As it was perfectly aligned with the what employers are looking out for. It made less challenging for me to secure this new job [Business Intelligence engineer at Amazon] in this competitive job market.”

“I really get more intuition about how to deal with data with some powerful tools in HW3 [uses AWS]. That feeling is beyond description for me.”

“I would like to say thank you for your class! Thanks to the skills I got from the class and the project, I got the offer.”
What we expect from you

• Actively participate throughout the course!

• If you need help, let us know — the earlier you let us know, the more help we can offer

• Help your fellow classmates out, e.g., help answer questions on Piazza

• Share your ideas! Ideas for improving learning experiences, let us know