CSE6242 / CX4242:

Data & Visual Analytics

Duen Horng (Polo) Chau
Associate Professor, College of Computing
Associate Director, MS Analytics
Machine Learning Area Leader, College of Computing
Georgia Tech
Google “Polo Chau” (only one in the world)

POLO CHAU

Associate Director, MS in Analytics
Assistant Professor, School of Computational Science & Engineering

College of Computing
Georgia Tech

Admin: Carolyn Young    Financial Manager: Arlene Washington
polo@gatech.edu    www.cc.gatech.edu/~dchau
Office: Klaus 1324    404-385-7682
Google Scholar    YouTube videos

LinkedIN profile    Follow @PoloChau

POSITIONS

May 2014 -
Associate Director
MS in Analytics, Georgia Tech

Aug 2012 -
Assistant Professor
School of Computational Science & Engineering, Georgia Tech

Dec 2012 - Dec 2015
Adjunct Assistant Professor
School of Interactive Computing, Georgia Tech

EDUCATION

Aug 2012
Ph.D. Machine Learning    Carnegie Mellon University

Research Group & GitHub

I'm expanding my Polo Club of Data Science research group to work on human-centered AI! Apply to our CS/CSE PhD program. Strong interest and experience in visualization, HCI, deep learning, machine learning, or data mining are big pluses!

Students (see more)
Robert Pienta, CSE PhD
Minsuk (Bryan) Kahng, CS PhD
Shang-Tse Chen, CS PhD
Fred Hohman, CSE PhD
Nilakash Das, CS PhD
Madhuri Shanbhogue, MS CS
Dezhi (Andy) Fang, CS UG
Siwei (Bob) Li, CS UG
Joon Kim, CS UG
Matthew Keezer, MS CS
Prasenjeet Biswal, MS CS
Varum Bezzam, MS CS
How to address Polo?

**Grammatically correct**

Prof. Chau

Dr. Chau

**Grammatically incorrect, but popular**

Prof. Polo

Dr. Polo
Course Registration

This class room seats 300. If you are on the waitlist, please wait for seats to released (some students typically “drop” after today).

- As of 3pm today
  - **CSE 6242 A**
    - 186/202 seats filled
    - 81/250 waitlist slots taken
  - **CX 4242 A**
    - 50/68 seats filled
    - 4/100 waitlist slots taken
- **CSE 6242 Q (distance-learning)**: 6 students
Course TAs  Be very very nice to them!

Neetha Ravishankar
Jennifer Ma
Mansi Mathur
Arathi Arivayutham
Vineet Vinayak Pasupulety
Siddharth Gulati

Office hours and locations (TBD) on course homepage poloclub.gatech.edu/cse6242
Scalable. Interactive. Interpretable.

At Georgia Tech, we innovate at the intersection of data mining and human-computer interaction (HCI) to synthesize scalable, interactive, and interpretable tools that amplify human's ability to understand and interact with billion-scale data and machine learning models. Our focus application areas include cybersecurity (e.g., fraud detection, malware detection, and adversarial machine learning), health, and social good.

Machine Learning Visualization & Interpretation

Interpretable deep learning and machine Learning through interactive visualization, with application in adversarial machine learning.
Scalable. Interactive. Interpretable.

At Georgia Tech, we innovate at the intersection of data mining and human-computer interaction (HCI) to synthesize scalable, interactive, and interpretable tools that amplify human's ability to understand and interact with billion-scale data and machine learning models. Our focus application areas include cybersecurity (e.g., fraud detection, malware detection, and adversarial machine learning), health, and social good.

Machine Learning Visualization & Interpretation

Interpretable deep learning and machine Learning through interactive visualization, with application in adversarial machine learning.
We work with (really) **large** data.
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**
Both develop methods for making sense of network data

<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>
How to do that?

**COMPUTATION**

Automatic

Summarization, clustering, classification

>Millions of nodes

**INTERACTIVE VIS**
# How to do that?

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

<table>
<thead>
<tr>
<th><strong>COMPUTATION</strong></th>
<th><strong>INTERACTIVE VIS</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Network Diagram" /></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>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>Interaction, visualization</td>
<td>Thousands of nodes</td>
</tr>
</tbody>
</table>

>Millions of nodes

Thousands of nodes
Our Approach for Big Data Analytics

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

ActiVis
Visual Exploration of Facebook Deep Neural Network Models

SHIELD
Fast, practical defense for deep learning

Cyber Security

Cyber MoneyBall
Predicting Cyber Threats with Virtual Security Products

MARCO
Fake Review Detection

Large Graph Mining & Visualization

MMap
Easy billion-scale graph computation on a PC using virtual memory

Apolo
Explore million-node graphs in real time

Social Good & Health

DeepPop
Deep Learning on Satellite Imagery for Population Estimation

Firebird
Predicting Fire Risk in Atlanta
Logistics

Course homepage  poloclub.gatech.edu/cse6242/
All assignments, slides posted here

Discussion, Q&A, find teammates
Piazza: link available on canvas.gatech.edu

Assignment Submission  Canvas
(Use Piazza for discussion)

Make sure you’re at the right Piazza! (CSE-6242-O01, CSE-6242-OAN have their Piazza forums too)
Course Homepage
For syllabus, HWs, projects, datasets, etc.

Google “cse6242”
poloclub.gatech.edu/cse6242/
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, 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.
Important to join Piazza because...

- Polo will announce events related to this class and data science in general
- Distinguished lectures
- Seminars
- Hackathons (free food, prizes)
- Company recruitment events (free food, swag)
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?
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 massive amounts of information that human beings create every single day.
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

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”.

Collection
Cleaning
Integration
Analysis
Visualization
Presentation
Dissemination

Can skip some

Can go back (two-way street)

- **Data types** inform **visualization** design
- **Data size** informs choice of **algorithms**
- **Visualization** motivates more **data cleaning**
- **Visualization** challenges algorithm assumptions
  e.g., user finds that results don’t make sense
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] In-class pop quizzes
  - Each quiz is worth 1% course grade
- No exams
Policies

On website; we go through them now

Grading, plagiarism, collaboration, late submission, and the “warning” 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. HCl): 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.
PASSAGE: A Travel Safety Assistant With Safe Path Recommendations For Pedestrians

Matthew Garvey  
College of Computing  
Georgia Institute of Technology  
Atlanta, GA 30332, USA  
mgarvey6@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, our group has developed a mobile application, PASSAGE, which uses location and crime data to find "safe paths" for users in Atlanta.

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.”

“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 Polo expects from you

• Actively participate throughout the course!
• Ask questions during class and on Piazza
• Help out whenever you can, e.g., help answer questions on Piazza
• Polo reserves last few minutes of every class for Q&A
FREE After-class Coffee 🍵

• After class, Polo randomly selects 5 students (+2 volunteers) for FREE after-class coffee

• Polo’s treat. You can order coffee, tea, pastries — whatever you want

• Very casual — you can ask me ANYTHING

• Will try doing this at least once a week, starting next week!