CSE6242 / CX4242: 

Data & Visual Analytics

Duen Horng (Polo) Chau
Assistant Professor
Associate Director, MS Analytics
Georgia Tech
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

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

Students (see more)
Robert Pienta, CSE PhD
Minsuk (Brian) Kahng, CS PhD
Shang-Tse Chen, CS PhD
Fred Hohman, CSE PhD
Nilaksh Das, CSE 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 Bezzal, 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. *Almost all physical seats have been filled.* If you are on the waitlist, please wait for seats to released (some students typically “drop” after today).

- As of 3pm today (Jan 9, 2018)
  - CSE 6242 A
    - 217/220 seats filled
    - 2/65 waitlist slots taken
  - CX 4242 A
    - 78/80 seats filled
    - 0/50 waitlist slots taken
  - CSE 6242 Q (distance-learning): 9 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.
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
Data

Insights
How to do that?

**COMPUTATION**

+ **HUMAN INTUITION**
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,</td>
<td>Interaction, visualization</td>
</tr>
<tr>
<td>clustering,</td>
<td></td>
</tr>
<tr>
<td>classification</td>
<td></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,</td>
<td></td>
</tr>
<tr>
<td>clustering,</td>
<td></td>
</tr>
<tr>
<td>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>COMPUTATION</th>
<th>INTERACTIVE VIS</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="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?

**COMPUTATION**

- Automatic
- Summarization, clustering, classification

**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 Human-Computer Interaction</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.)
Machine Learning + Visualization

Recently received $1.2 Million NSF award

http://www.scs.gatech.edu/news/522401/12m-nsf-award-helps-consumers-enter-age-big-data


Apolo: Explore million-node graphs in real time

The cost structure of sensemaking
245 citations 8 versions
Carina: Million-node Graph Exploration in Web Browser [www’17]
Find **co-directors** who made at least **two films** together, starring the same **actor**.

```
MATCH (d1:director)--(f1:film),
     (d1)--(f2:film), (d1)--(f3:film),
     (f1)--(d2:director)--(f2),
     (d2)--(f3),
     (f1)--(a:actor)--(f2), (a)--(f3)
WHERE f1.decade = 1990 AND d1 <> d2
RETURN d1, d2, f1, f2, f3, a
```
ActiVis
Visualization & Interpretation of Deep Learning Models
Deployed on ML platform of \textit{facebook}

A. Model Architecture
B. Neuron Activation
C. Instance Selection

ActiVis: Visual Exploration of Industry-Scale Deep Neural Network Models.
Polo’s primary application area:

Cyber Security
Polonium & AESOP

Patented with Symantec
Finds malware from 37 billion file relationships
Serving 120 million users worldwide
Published at SDM’11, KDD’14
NetProbe

Auction Fraud Detection on eBay

THE WALL STREET JOURNAL.

CNNMoney.com

MSNBC

KDKA 02
MARCO

Detecting Fake Yelp Reviews

Best papers of SDM 2014
(top data mining conference)
Insider Trading Detection with Securities and Exchange Commission (SEC)

6-node Clique: each edge is an electrical company

Triangle: company A: biotech; company B: medical supplies

Chain: each edge is an electrical utilities company

Each edge above corresponds to an Electrical Utilities Company

Bush tax cut May 23, 2003

Quant Meltdown Aug 8, 2007

Logistics

Course homepage: [polo.club.gatech.edu/cse6242/](polo.club.gatech.edu/cse6242/)
All assignments, slides posted here

Discussion, Q&A, find teammates:
Piazza: [goo.gl/cGvHeE](goo.gl/cGvHeE)
Or [piazza.com/gatech/spring2018/cse6242aqcx4242a](piazza.com/gatech/spring2018/cse6242aqcx4242a)

Make sure you’re at the right Piazza! (CSE-6242-O01, CSE-6242-OAN have their Piazza forums too)

Assignment Submission
T-Square
(Use Piazza for discussion)
Course Homepage

For syllabus, HWs, projects, datasets, etc.

Google “cse6242”

poloclub.gatech.edu/cse6242/2018spring
Join Piazza ASAP

googl1/cGvHeE

Announcements and Discussion

We use Piazza for announcements and discussion.


Double check that you are joining the right Piazza!

When you have questions about class, homework, project, etc., post your questions there. Our teaching help answer them quickly. You can also use Piazza to find project teammates.

T-square will only be used for submission of assignments and projects.

While we welcome everyone to share their experiences in tackling issues and helping each other out, please may affect the learning experience of your fellow classmates.
Important to join Piazza because...

**The fastest way** to get help with homework assignments is to post your questions on Piazza. If you prefer to address your questions 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).
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?
THE WORLD OF DATA

- 2.9 million emails sent every second
- 375 megabytes consumed by households each day
- 20 hours of video uploaded to YouTube every minute
- 24 petabytes of data per day processed by Google
- 50 million tweets per day
- 700 billion total minutes spent on Facebook each month
- 1.3 exabytes of data sent and received by mobile internet users per second
- 72.9 items ordered on Amazon per second

Sources: Cisco, comScore, MapR, Kundra Partnership Group, Twitter, YouTube

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 (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”

| Collection | Can skip some |
| Cleaning   |               |
| Integration|               |
| Analysis   | Can go back (two-way street) |
| Visualization | • Data types inform visualization design |
|             | • Data size informs choice of algorithms |
| Presentation | • Visualization motivates more data cleaning |
| Dissemination | • 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 **breath** 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

Collaborating on homework
Late submission policy
While collaboration is allowed for homework assignments, each student must write up their own answers. All GT students must observe the honor code. Any suspected plagiarism and academic misconduct will be reported and directly handled by the Office of Student Integrity (OSI).
WARNING
You’ll be writing a lot of code

Q: Is it OK to copy and use code found on the web?
A: No

Q: Why?
A: Here’s why…
WARNING: Do not plagiarize!

• Using code as reference does not mean copying and modifying it.

• To use example code as reference, you should go over it, understand what it is doing, and then try to accomplish what it is trying to do using your own code. And it’s a good practice to cite the sources (e.g., as part of your code comments).

• The analogy is like how you would write an essay or a speech. You can get inspirations from others, but you should use your own words, otherwise it will be considered plagiarism. Plagiarism can lead to severe consequences.

• http://www.plagiarism.org/plagiarism-101/what-is-plagiarism/
Late Submissions Policy

- **Homework**: each student has **4 slip days** total. No questions asked.
- **Project**: each team has **3 slip days** total. No questions asked. Slip days may not be used on in-class activities (e.g., proposal presentation, poster presentation, etc.).
- To use slip days, **specify the number of days you have used in the textbox on T-Square** (when you submit your work).
- **Each slip day equals 24 hours**. E.g., if a submission is late for 30 hours, that counts as 2 slip days.
- After all slip days are used up, **5% deduction for every 24 hours of delay**. (e.g., 5 points for a 100-point homework).
- We **will not consider late submission of any missing parts** of an homework assignment or project deliverable. To make sure you have submitted everything, download your submitted files to double check.
- We **will not accept** any submission that is more than 7 days late.
- No penalties for medical reasons or emergencies. And should they arise, you **must** contact the Dean of Students office. Doctor's notes, medical documentation, explanation of emergencies, etc. should be submitted to the Dean’s office. After their office receives the information, we will notify me on your behalf.
Distance Learning Sections (Q & Q3)

A standard 3-day lag applies to all homework and project deliverables. For project presentation, a group that has DL student member can choose to:

1. Present in class without 3-day lag; or
2. Submit a video presentation with 3-day lag (e.g., screen capture)
Are You Ready to Take this Course?

- Require **a lot of programming**
- You need to **learn many new things in short amount of time**
- HW2 (D3 data vis) is most demanding: Javascript + CSS + HTML
- **Very common in industry**
Are You Ready to Take this Course?

The **best way** to find out is to check out previous semester’s homework assignments

- poloclub.gatech.edu/cse6242/2017fall/
- poloclub.gatech.edu/cse6242/2017spring/
- poloclub.gatech.edu/cse6242/2016fall/
- poloclub.gatech.edu/cse6242/2016spring/
Prerequisites & Expectation
For both CSE 6242 (grad) and CX 4242 (undergrad)

Students are expected to complete **significant** programming assignments (homework, project) that may involve higher-level languages or scripting (e.g., Java, R, Matlab, Python, C++, etc.).

Some assignments may involve web programming and D3 (e.g., Javascript, CSS).

**You are expected to quickly learn many new things.** For example, an assignment on Hadoop programming may require you to learn some basic Java and Scala quickly, which should not be too challenging if you already know another high-level language like Python or C++. **Please make sure you are comfortable with this.**

Please take a look at the assignments (homework and project) of the previous offerings of this course, which will give you some idea about the difficulty level of the assignments.

Basic linear algebra, probability knowledge is expected.

e.g., http://poloclub.gatech.edu/cse6242/2017fall/
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
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, a group has developed a mobile application, PASSAGE, that uses crime data to find "safe paths" for users in Atlanta.

Authors
Matthew Garvey
Meghna Natraj
Nilaksh Das
Bhanu Verma
Jiaxing Su

ACM SIGCHI Beachside Conference on Human-Computer Interaction, May 2023, Orlando, Florida, USA

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!