CSE6242 / CX4242: Data & Visual Analytics

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

Legal name: Duen Horng Chau

Associate Director, MS in Analytics
Assistant Professor, School of Computational Science & Engineering
Adjunct Assistant Professor, School of Interactive Computing
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 (h-index: 22)  
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 -
Adjunct Assistant Professor
School of Interactive Computing, Georgia Tech

EDUCATION

Aug 2012
Ph.D. Machine Learning  
Carnegie Mellon University
Thesis: Data Mining Meets HCI: Making Sense of Large Graphs

NIH MD2K Center of Excellence, Co-PI
NIH MD2K Center of Excellence for Mobile Sensor Data-to-Knowledge
Associate Director
Visualization at Georgia Tech
IDEA workshop at KDD  
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. I plan to increase seats to accommodate students on the waitlist. If you are on the waitlist, you will very likely officially get a seat

- As of 2pm today (Jan 10, 2017)
  - CSE 6242 A
    - 160/160 seats filled
    - 65/100 waitlist slots taken
  - CX 4242 A
    - 44/44 seats filled
    - 14/30 waitlist slots taken
- (Distance-learning CSE 6242 Q: 9 students)
TAs (Be very very nice to them!)

Meghna - Head TA
Bhanu Verma
Fred Hohman
Kiran Sudhir
Varun Bezzam
Chirag Tailor

Office hours and locations (TBD) on course homepage poloclub.gatech.edu/cse6242
We work with (really) large data.
Internet
50 Billion Web Pages
Facebook
1.2 Billion Users
Citation Network
250 Million Articles
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

Large Networks We 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>
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, 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?

**COMPUTATION**

Automatic

Summarization, clustering, classification

> Millions of nodes

**INTERACTIVE VIS**
How to do that?

**COMPUTATION**

- Automatic
- Summarization, clustering, classification

**INTERACTIVE VIS**

- 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="network_diagram.png" 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><img src="image1.png" alt="Diagram" /></td>
<td><img src="image2.png" alt="Diagram" /></td>
</tr>
<tr>
<td><strong>COMPUTATION</strong></td>
<td><strong>INTERACTIVE VIS</strong></td>
</tr>
<tr>
<td><strong>Automatic</strong></td>
<td><strong>User-driven; iterative</strong></td>
</tr>
<tr>
<td><strong>Summarization, clustering, classification</strong></td>
<td><strong>Interaction, visualization</strong></td>
</tr>
<tr>
<td><strong>Interaction, visualization</strong></td>
<td><strong>Thousands of nodes</strong></td>
</tr>
</tbody>
</table>
Our Approach for Big Data Analytics

**DATA MINING** | **HCI**
---|---
Automatic | User-driven; iterative
Summarization, clustering, classification | Interaction, visualization
>Millions of items | Thousands of items

Our research combines the **Best of Both Worlds**
“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.)
Recently received $1.2 Million NSF award

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

Polo’s primary application area:

Cyber Security
Polonium

Patented with Symantec
Finds malware from 37 billion file relationships
Serving 120 million users worldwide
Published at SDM’11
NetProbe
Auction Fraud Detection on eBay
Insider Trading Detection
with Securities and Exchange Commission (SEC)
MARCO

Detecting Fake Yelp Reviews

Best papers of SDM 2014
(top data mining conference)
Latent Gesture

CS undergraduate students: Prem Saravanan, Samuel Clarke

won prestigious Astronaut Scholarship
Logistics

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

Discussion, Q&A, find teammates
Piazza: goo.gl/b0TkEP
Or https://piazza.com/class/ixpgu1xccuo47d

Assignment Submission
T-Square
(Use Piazza for discussion)

Make sure you’re at the right Piazza!
(CSE 6242 O has its Piazza too)
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
- Total minutes spent on Facebook each month: 50 million
- Data sent and received by mobile internet users: 700 billion
- Products ordered on Amazon per second: 1.3 exabytes
- Items

Sources: CNN, Fox News, MapReduce, Reddit, Groupon, 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.

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

("big data" is buzz word, so is "IoT" - Internet of Things)

- Many companies’ businesses are based on big data (Google, Facebook, Amazon, Apple, Symantec, LinkedIn, and many more)
- Web search
  - Rank webpages (PageRank algorithm)
  - Predict what you’re going to type
- Advertisement (e.g., on Facebook)
  - Infer users’ interest; show relevant ads
  - Infer what you like, based on what your friends like
- Recommendation systems (e.g., Netflix, Pandora, Amazon)
- Online education
- Health IT: patient records (EMR)
- Bio and Chemical modeling:
  - Finance
  - Cybersecuruity
- Internet of Things (IoT)
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.
Analytics Building Blocks
Collection
Cleaning
Integration
Analysis
Visualization
Presentation
Dissemination
Building blocks, not “steps”

- Can skip some
- Can go back (two-way street)
- Examples
  - Data types inform visualization design
  - Data informs choice of algorithms
  - Visualization informs data cleaning (dirty data)
  - Visualization informs algorithm design (user finds that results don’t make sense)
Schedule

- Collection
- Cleaning
- Integration
- Analysis
- Visualization
- Presentation
- Dissemination
Course Goals

• Learn visual and computation techniques and tools, for typical data types

• Learn how to complement each kind of methods

• Work on real data & problem

• Learn practical know-how (useful for jobs, research)

• Gain breath of knowledge
Schedule

See course homepage
poloclub.gatech.edu/cse6242/
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
- No exams
Are You Ready to Take this Course?

- Require **a lot of programming**
- Needs to learn new languages quickly (e.g., Javascript, Scala)
- HW2 (D3 data vis) is most demanding
- Javascript + CSS + HTML
- You need to be prepared to **learn many things** in short amount of time
- **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

- [http://poloclub.gatech.edu/cse6242/2016fall/](http://poloclub.gatech.edu/cse6242/2016fall/)
- [http://poloclub.gatech.edu/cse6242/2016spring/](http://poloclub.gatech.edu/cse6242/2016spring/)
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.
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 utilizes crime data to find "safe paths" for users.

Figure 1: Paths recommended by PASSAGE

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Firebird: Predicting Fire Risk and Prioritizing Fire Inspections in Atlanta

Michael Madaio  
Carnegie Mellon University  
Pittsburgh, PA, USA  
mmadaio@cs.cmu.edu

Shang-Tse Chen  
Georgia Tech  
Atlanta, GA, USA  
schen351@gatech.edu

Oliver L. Haimson  
University of California, Irvine  
Irvine, CA, USA  
ohaimson@uci.edu

Wenwen Zhang  
Georgia Tech  
Atlanta, GA, USA  
wzhang300@gatech.edu

Xiang Cheng  
Emory University  
Atlanta, GA, USA  
xcheng7@emory.edu

Matthew Hinds-Aldrich  
Atlanta Fire Rescue Dept.  
Atlanta, GA, USA  
mhinds-aldrich@atlantaga.gov

Duen Horng Chau  
Georgia Tech  
Atlanta, GA, USA  
polo@gatech.edu

Bistra Dilkina  
Georgia Tech  
Atlanta, GA, USA  
bdilkina@cc.gatech.edu

ABSTRACT

The Atlanta Fire Rescue Department (AFRD), like many municipal fire departments, actively works to reduce fire risk by inspecting commercial properties for potential hazards and fire code violations. However, AFRD’s fire inspection practices relied on tradition and intuition, with no existing data-driven process for prioritizing fire inspections or identifying new properties requiring inspection. In collaboration with AFRD, we developed the Firebird framework to help municipal fire departments identify and prioritize commercial property fire inspections, using machine learning, geocoding, and information visualization. Firebird computes fire risk scores for over 5,000 buildings in the city,
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 5-10min of every class for Q&A
FREE After-class Coffee

- After each 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.