E6895 Advanced Big Data Analytics Lecture 1:

*Advanced Big Data Analytics to Build a Brain*

Ching-Yung Lin, Ph.D.
Adjunct Professor, Depts. of Electrical Engineering and Computer Science

January 18, 2018
Definition and Characteristics of Big Data

“Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.” -- Gartner

which was derived from:

“While enterprises struggle to consolidate systems and collapse redundant databases to enable greater operational, analytical, and collaborative consistencies, changing economic conditions have made this job more difficult. E-commerce, in particular, has exploded data management challenges along three dimensions: volumes, velocity and variety. In 2001/02, IT organizations much compile a variety of approaches to have at their disposal for dealing each.” – Doug Laney
Scalability — Scale Up & Scale Out

- **Scale out**
  - Use more resources to distribute workload in parallel
  - Higher data access latency is typically incurred

- **Scale up**
  - Efficiently use the resources
  - Architecture-aware algorithm design

Example: Resource utilization for a large production cluster at Twitter data center

For independent data ==> scale up may not have obvious advantage than scale out
For linked data ==> utilizing scale up as much as possible before scale out

Techniques towards Big Data

- Massive Parallelism
- Huge Data Volumes Storage
- Data Distribution
- High-Speed Networks
- High-Performance Computing
- Task and Thread Management
- Data Mining and Analytics
- Data Retrieval
- Machine Learning
- Data Visualization

Techniques exist for years to decades. Why is Big Data hot now?
## Contrasting Approaches to Traditional Scenarios

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Typical Scenario</th>
<th>Big Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Application development</td>
<td>Applications that take advantage of massive parallelism developed by specialized developers skilled in high-performance computing, performance optimization, and code tuning</td>
<td>A simplified application execution model encompassing a distributed file system, application programming model, distributed database, and program scheduling is packaged within Hadoop, an open source framework for reliable, scalable, distributed, and parallel computing</td>
</tr>
<tr>
<td>Platform</td>
<td>Uses high-cost massively parallel processing (MPP) computers, utilizing high-bandwidth networks, and massive I/O devices</td>
<td>Innovative methods of creating scalable and yet elastic virtualized platforms take advantage of clusters of commodity hardware components (either cycle harvesting from local resources or through cloud-based utility computing services) coupled with open source tools and technology</td>
</tr>
<tr>
<td>Data management</td>
<td>Limited to file-based or relational database management systems (RDBMS) using standard row-oriented data layouts</td>
<td>Alternate models for data management (often referred to as NoSQL or “Not Only SQL”) provide a variety of methods for managing information to best suit specific business process needs, such as in-memory data management (for rapid access), columnar layouts to speed query response, and graph databases (for social network analytics)</td>
</tr>
<tr>
<td>Resources</td>
<td>Requires large capital investment in purchasing high-end hardware to be installed and managed in-house</td>
<td>The ability to deploy systems like Hadoop on virtualized platforms allows small and medium businesses to utilize cloud-based environments that, from both a cost accounting and a practical perspective, are much friendlier to the bottom line</td>
</tr>
</tbody>
</table>

“Big Data Analytics”, David Loshin, 2013
Why Big Data now?

- More data are being collected and stored
- Open source code
- Commodity hardware / Cloud

- High-Volume
- High-Velocity
- High-Variety

→ Artificial Intelligence
Agenda:

• Introduction of IBM System G
• Answering the Questions raised by FINRA
• Large-Scale Graph Computing:
  • System G Graph Database
  • System G Graph Analytics
• Demo of System G Graph Tools
• Relationship Extraction
• Machine Reasoning
• Discussion
2011

THINK

Who is Stoker?
$1,000

Who is Bram Stoker?
$17,973

WHO IS BRAM STOKER?
$5,600

$24,000

$77,147

$21,600
A 45-year-old male bonobo grooms a 21-year-old male. To properly focus on his grooming, the aging chimp extends his hands farther from his face than younger bonobos. Photo by Heungjin Ryu/Cell Press
Network Graph is the way we remember, we associate, and we reason.
## Course Outline

<table>
<thead>
<tr>
<th>Class Date</th>
<th>Class Number</th>
<th>Lecture Topics</th>
<th>Student Presentations</th>
</tr>
</thead>
<tbody>
<tr>
<td>01/18/18</td>
<td>1</td>
<td>Advanced Big Data Analytics to Build a Brain</td>
<td></td>
</tr>
<tr>
<td>01/25/18</td>
<td>2</td>
<td>Foundations toward Human-Like Artificial Intelligence</td>
<td></td>
</tr>
<tr>
<td>02/01/18</td>
<td>3</td>
<td>Next-Generation Hardwares (I)</td>
<td>Cognitive Robot (I)</td>
</tr>
<tr>
<td>02/08/18</td>
<td>4</td>
<td>Social Machines</td>
<td>Finance Robo-Advisor (I)</td>
</tr>
<tr>
<td>02/15/18</td>
<td>5</td>
<td>New Chips for AI &amp; Big Data</td>
<td>Knowledge Graphs (I)</td>
</tr>
<tr>
<td>02/22/18</td>
<td>6</td>
<td>Ethics Machines (I)</td>
<td>Advanced Topics (I)</td>
</tr>
<tr>
<td>03/01/18</td>
<td>7</td>
<td>Ethics Machines (II)</td>
<td>Cognitive Robot (II)</td>
</tr>
<tr>
<td>03/08/18</td>
<td>8</td>
<td>Feeling Machines (I)</td>
<td>Finance Robo-Advisor (II)</td>
</tr>
<tr>
<td>03/22/18</td>
<td>9</td>
<td>Feeling Machines (II)</td>
<td>Knowledge Graphs (II)</td>
</tr>
<tr>
<td>03/29/18</td>
<td>10</td>
<td>Arts-Aware Machines</td>
<td>Advanced Topics (II)</td>
</tr>
<tr>
<td>04/05/18</td>
<td>11</td>
<td>Reasoning Machines (I)</td>
<td>Cognitive Robot (III)</td>
</tr>
<tr>
<td>04/12/18</td>
<td>12</td>
<td>Reasoning Machines (II)</td>
<td>Finance Robo-Advisor (III)</td>
</tr>
<tr>
<td>04/19/18</td>
<td>13</td>
<td>Consciousness Machines</td>
<td>Knowledge Graphs (III)</td>
</tr>
<tr>
<td>04/26/18</td>
<td>14</td>
<td>Dreaming Machines</td>
<td>Advanced Topics (III)</td>
</tr>
<tr>
<td>05/10/18</td>
<td>15</td>
<td></td>
<td>Final Project Presentations</td>
</tr>
</tbody>
</table>
Course Grading

- **Task (3 milestones): 60%**
  - Language Requirement: C/C++, Java, JavaScript, Python, Perl
  - Presentation, Report and Source Code

- **Final Project: 30%**
  - Teamwork: 1 - 2 students per team
  - Final Report (paper, up to 12 pages)
  - Workshop Presentation (Oral, Demo) and Video
  - Open Source

- **Class Participation: 10%**
  - Attendance
  - Discussion (Asking/Answering Questions)
Contact Information and TAs

- Professor Lin:
  - Office Hours:
    Thursday 9:30pm – 10:00pm (lecture room)
  - Contact: c.lin@columbia.edu (the same as <cl300>)
  - Telephone: 914-396-6924

- TAs:
  - Vishal Anand
  - TBD
Course Information

- Website:
  
  http://www.ee.columbia.edu/~cylin/course/bigdata/

- Textbook:
  
  -- None, but reference book(s) and/or articles/papers will be provided each lecture.
Part 1

THE BASICS OF HUMAN LIFE

1. Are Human Brains Unique?.......................... 7
2. Would a Chimp Make a Good Date?............... 28

Part 2

NAVIGATING THE SOCIAL WORLD

3. Big Brains and Expanding Social Relationships 39
4. The Moral Compass Within.......................... 113
5. I Feel Your Pain........................................ 158

Part 3

THE GLORY OF BEING HUMAN

6. What's Up with the Arts?.............................. 203
7. We All Act Like Dualists: The Converter Function 246
8. Is Anybody There?..................................... 276

Part 4

BEYOND CURRENT CONSTRAINTS

9. Who Needs Flesh?...................................... 335
Afterword................................................. 386
Notes...................................................... 391
Index..................................................... 432
Reference Book

"[A] compelling read... It lingers like a sublime night vision."  
— Entertainment Weekly

The Mind at Night

The New Science of How and Why We Dream

Andrea Rock

1. Rockettes, EEGs, and Banana Cream Pie
2. The Anti-Freud
3. Experiments of Nature
4. The Lesson of the Spiny Anteater
5. Rerunning the Maze
6. Nocturnal Therapy
7. The Ultimate Spin Doctor
8. Creative Chaos
9. Altered States
10. Consciousness and Beyond
Reference Foundation

- **Graph Middleware:**
  - Parallel Prog. Lib.
  - Power Optimization
  - GPU Optimization

- **Graph Analytics:**
  - Topological Analysis
  - Matching and Search
  - Path and Flow

- **Spatiotemporal Analytics:**
  - Spatiotemporal Mining
  - Spatiotemporal Indexing

- **Graph Database:**
  - Native Store
  - GBase

- **Graph Visualization:**
  - Multivariate Graph
  - Dynamic Graph
  - Big Graph

- **Machine Learning:**
  - Deep Learning Tools
  - Visual and Text Sentiment Tools
  - Anomaly Detection Tools

- **Mobile Cognition:**
  - iOS Cognition Tools
  - Robot Cognition Tools

- **Graph Database Technologies**
- **Network Analytics Technologies**
- **Machine Learning Technologies**
- **Mobile Cognition Technologies**
- **Machine Reasoning Technologies**

1. Graph Database Technologies
2. Network Analytics Technologies
3. Machine Learning Technologies
4. Machine Reasoning Technologies

Brain sections:
- Sensing & Observation
- Reasoning & Strategy
- Memory
- Perception & Representation
- Judgment
5 Key Big Data Use Case Categories

Big Data Exploration
Find, visualize, understand all big data to improve decision making

Enhanced 360° View of the Customer
Extend existing customer views (MDM, CRM, etc) by incorporating additional internal and external information sources

Security/Intelligence Extension
Lower risk, detect fraud and monitor cyber security in real-time

Operations Analysis
Analyze a variety of machine data for improved business results

Data Warehouse Augmentation
Integrate big data and data warehouse capabilities to increase operational efficiency
1. Expertise Location
2. Recommendation
3. Commerce
4. Financial Analysis
5. Social Media Monitoring
6. Telco Customer Analysis
7. Healthcare Analysis
8. Data Exploration and Visualization
9. Personalized Search
10. Predictive Anomaly Detection
11. Fraud Detection
12. Cybersecurity
13. Sensor Monitoring
14. Cellular Network Monitoring
15. Cloud Monitoring
17. Traffic Navigation
18. Data Curation
19. Genomic Medicine
20. Image and Video Understanding
21. Brain Network Analysis
22. Near Earth Object Analysis
23. Cognitive Robot
24. Complex Scene detection for Autonomous Driving
Task Area 1: Cognitive Robot
Area 1 Tasks List:

A1: Text Recognition (English)
A2: Text Recognition (Chinese)
A3: General Object Recognition
A4: Vehicle Object Recognition
A5: Object Tracking
A6: Face Recognition
A7: Facial Expression Recognition
A8: Emotion Recognition
A9: Gesture Recognition
A10: Audio-Visual Event Detection (Public Area)
A11: Audio-Visual Event Detection (Home)
A12: Speech Recognition (English)
A13: Speech Recognition (Chinese)
A14: Robot-Human Interaction (Conversation)
A15: Robot-Human Interaction (Physical)
Example - Generic Object Recognition running natively on mobile device (iPhone)
Demo: Detecting Cars and Pedestrians in complex scenario
Large Video Event Ontology Browsing, Search and Tagging (EventNet Demo)
Task Area 2: AI for Finance and Wealth Management

Market Data Analysis and Investment Targets
Advanced Dynamic ‘Know Your Customer’
Optimized Personalized Investment Strategy
Bank-Customer Interaction Strategy

High End Customers (Private Bank / Special Investment Services)

Targeted Customers (Consumer Bank Services): $15K - $1M
(Customer #: 30M ~ 50M in China)

General Public (Consumer Bank Services)
(Customer #: > 1B in China)
What is Robo-Advisor?

Robo-Advisor is a new type of wealth management service. Based on the risk level and investment goals provided by the investor, and it uses a series of ‘smart algorithm’ to calculate the optimal investment suggestions.

Robo-advisors directly managed about $19 billion as of December 2014. By 2020 the global assets under management of robo-advisers is forecast to grow to an estimated US$255B.

Features:
- Strongly depend on technology, algorithm and financial theory
- Distributed investment, maximum long-term return
- Personalized portfolio allocation.

Example: Harry Markowitz Theory
Example: Wealthfront——low entry requirement, low fee

- On Sept 2015, the total asset is $2.6B.
- The estimated value of the company is $1B as in 2015.

- Low entry requirement: min investment value USD $500.
- Low fee:
  - Zero annual fee, if account is lower than $10K.
  - 0.25% fee is charged for the part of asset amount that is larger than $10K.
  - No agent fee
  - Based on Wealthfront, in average, each user only needs to pay 0.12% fee.

Ditch your mutual fund
Upgrade to Institutional Class Asset Management

Outperformed Morningstar 5 Star Mutual Funds

<table>
<thead>
<tr>
<th>Morningstar 5 Star Equity Funds</th>
<th>managers on wealthfront</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disclosure</td>
<td>26.5%</td>
</tr>
<tr>
<td></td>
<td>30.0%</td>
</tr>
</tbody>
</table>

Lower Fees than Actively Managed Mutual Funds

<table>
<thead>
<tr>
<th>Actively Managed Mutual Funds</th>
<th>managers on wealthfront</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disclosure</td>
<td>2.17%</td>
</tr>
<tr>
<td></td>
<td>1.30%</td>
</tr>
</tbody>
</table>

Expenses are represented as an asset-weighted average percentage of managed assets for mutual funds and a simple average for managers on Wealthfront. Learn More
Typical Steps of Robo-Advisory

Most of the robo-advisor platform is built based on the modern investment portfolio theory, using Exchange Trade Funds (ETFs) to build portfolio.

Customer Profiling
- design questionnaire;
- Score Risk Capacity and Risk Willingness based on the answers of the questionnaire.

Construct Portfolio
- portfolio strategy;
- type analysis;
- optimum allocation;

Tracing Portfolio
- Monte Carlo Simulation
- Judge whether the goal is achieved
- Suggest adjustments;

Receiving Benefits
- Saving tax through the loss to compensate the gains;
- outcome is highly related to the income;
- Investment income tax (not applicable in China);
- set tolerance level to avoid over adjustment;

Based on a survey of Wells Fargo, in US, there is only 16% of population in their 20s and 30s are willing to interact with investment consultants. The remaining people prefer to use these types of AI consultant.
### Four Steps to use Big Data Cognitive Analysis for Robo-Advisor

<table>
<thead>
<tr>
<th>Investment Market Analysis</th>
<th>Dynamically Know Your Customer</th>
<th>Optimized Personalized Investment Strategy</th>
<th>Precise Bank-Customer Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Analyze the market performance of various kinds of funds</td>
<td>• Customer Profiling, e.g., based on IPQ (Individual Profile Questionnaire), Feedback, Risk Capacity and Risk Willingness</td>
<td>• Strategy computation and optimization based on personal history</td>
<td>• Create and predict customer interaction strategy, including when, method, content to interact with customer – to achieve max customer and bank benefit.</td>
</tr>
<tr>
<td>• Analyze domestic and international financial and economic changes and how they may impact CPI, PPI, or GDP.</td>
<td>• Understand what the customer really wants based on their past behaviors interacting with bank</td>
<td>• Demonstrate / Simulate ‘what ifs’ when the portfolio has different allocation.</td>
<td>Data</td>
</tr>
<tr>
<td>• Use Machine Learning and Deep Learning, based on historical economic numbers, find out how factors impact financial markets.</td>
<td>Data</td>
<td>• Explainability of ‘what ifs’ to customer to the customer.</td>
<td>• Customer Data</td>
</tr>
<tr>
<td><strong>Data</strong></td>
<td>• Customer Data</td>
<td><strong>Data</strong></td>
<td>• Interaction Data</td>
</tr>
<tr>
<td>• Product Data</td>
<td>• Behavior Data / Interaction Data</td>
<td>• Market Data</td>
<td></td>
</tr>
</tbody>
</table>
Potentially Analyzing Social Media for Market Intelligence

- Live Monitoring: Monitoring real-time tweets on keyword
- Trend Monitoring: Analyzing trend of conversations based on hashtags
- Multimedia Monitoring: Recognizing visual content and analyzing visual sentiments
- Geo Monitoring: Monitoring the places that people are sending out tweets
- Link Exploration: Visualizing relationships, discussion sequences, and graphs
- Impact Prediction: Analyzing conversations and predicting their impact to business
- Story Detection: Detecting live developing stories on social media and their evolution
- Person Analytics: Analyzing a person's personality, trustworthiness, etc.
- Target Discovery: Inspecting potential users for bot detection, marketing, or influencing
- Forensic Analytics: Analyzing retweet sequences and displaying anomalies

Images in tweets that belong to one story
Text of the newest tweet in this story
Example: Major Wealth Management

Hundreds of products/campaigns
Combinations with incompatibilities

How much of each product/campaign?

Telesales, Mail, email, Office, etc…
Done through which channel?

Nightly batch run, select over 1.2M

Experts doing what-if to improve process

To which customers?
Several millions of customers

When?
Select actions for the next days

Hundreds of products/campaigns
Combinations with incompatibilities

How much of each product/campaign?

Telesales, Mail, email, Office, etc…
Done through which channel?

Nightly batch run, select over 1.2M

Experts doing what-if to improve process

To which customers?
Several millions of customers

When?
Select actions for the next days
Area 2 Tasks List:

B1 – B5. Market Intelligence Analysis (US and China) and Investment Targets

- B1. Finance company knowledge graph (Company Information – maybe only one specific sector) Different knowledge graph from same data source using different algorithms

- B2. Finance events linkage map (Finance news – maybe only one specific sector)

- B3. Bayesian network for core financial events (such as M&A, interest rate change, competitor actions, natural disasters)

- B4. News sentiment (classify news as facts/expectations) + Time decay of news sentiment

- B5. Alpha generation from alternative data sources (estimate unemployment from tweets, image recognition on ships)
Area 2 Tasks List:

**B6 – B8. Advanced Dynamic “Know Your Customer”:**
- B6. Financial product classification
- B7. Customer profiling
- B8. Match product and customer

**B9 – B12. Optimized Personal Investment Strategy:**
- B9. Quantitative/Qualitative Strategy for a single instrument based investment objective
- B10. Asset allocation strategies based investment objective/risk level
- B11. Compare and back-test different strategies
- B12. Investment portfolio recommendation system

**B13 – B15. Precise Bank-Customer Interaction:**
- B13. Analyze customer behaviors and transactions to predict loan/mortgage/new credit card needs
- B14. Credit default risk analysis
- B15. Customer relationship management (customer loyalty, precision promotion strategies, contact timing and channel)
Task Area 3: Knowledge Graphs
Area 3 Tasks List:

C1. Data Acquisition for Building “Columbia” Knowledge Graphs
C2. APIs and Representation for “Columbia” Knowledge Graphs
C3. Mapping Real-World Applications into Graph Query Languages
C4. Key Primitives for Graph Query Languages
C5. GPU-based Graph Analytics (I)
C6. GPU-based Graph Analytics (II)
C7. Graph Workflow System
C8. Graph Analytics with Deep Learning
C9. High-Performance Dynamic Graph Database
C10. Python for High-Performance Graph Analytics
C11. Relational Database to Graph Database Migration
C12. Exploration Environment for Graph Analytics
C13. Visual Query and Authoring Tool for Graph Database
C14. Visualization of Large Dynamic Network
C15. Visual Exploration of Large Graph in Immersive Environment
Evolution of Data Size and Complexity — Graphs as Example

N.T. Bliss, Confronting the Challenges of Graphs and Networks, Lincoln Laboratory Journal, 2013

<table>
<thead>
<tr>
<th>Classical Era</th>
<th>Modern Era</th>
</tr>
</thead>
<tbody>
<tr>
<td>1700-1800</td>
<td>2000-2010</td>
</tr>
<tr>
<td>Seven Bridges of Königsberg</td>
<td>2011</td>
</tr>
<tr>
<td>Zachary's Karate Club</td>
<td>2005</td>
</tr>
<tr>
<td>1903</td>
<td>2009</td>
</tr>
<tr>
<td>Les Misérables characters</td>
<td>2010</td>
</tr>
<tr>
<td>1999</td>
<td>2011</td>
</tr>
<tr>
<td>Internet topology</td>
<td>Web proxy logs</td>
</tr>
<tr>
<td>2001</td>
<td>2012</td>
</tr>
<tr>
<td>129,164 nodes, 330,870 edges</td>
<td>2013</td>
</tr>
<tr>
<td>2013</td>
<td>2014</td>
</tr>
</tbody>
</table>

Social graph, Relation graph
- Facebook: 1.23 billion users (2013/12)
- WhatsApp: 1 billion users (2015/08)
- Our Twitter graph: 120M nodes, 2B edges (2012-14)
- Our Bitcoin graph: 70M nodes, 200M edges (2014)

Brain neural network
- 100 billion neurons
- 100 trillion connections

Seven bridge problem in 1735

AIDS partner graph

US road map
- 129,164 nodes, 330,870 edges

Hamilton graph

N.T. Bliss, Confronting the Challenges of Graphs and Networks, Lincoln Laboratory Journal, 2013
The size of graphs used by Graph 500

No. of edges

\[ \log_2(m) \]

1 trillion edges

1 billion edges

\[ \log_2(n) \]

No. of nodes

1 billion nodes

1 trillion nodes

Graph500 (Toy)

Graph500 (Small)

Graph500 (Mini)

Graph500 (Medium)

Graph500 (Large)

Graph500 (Huge)

USA Road Network

Human Brain Project

Symbolic Network

USA-road-d.NY.gr

USA-road-d.LKS.gr

USA-road-d.USA.gr

Twitter (tweets/day)

CJMMJPO

OPEFT

USJMMJPO

OPEFT

FEHFT

FEHFT

Symbolic Network

64\textsuperscript{3}PBE/FUXPSL

5XJUUFS\textbackslash UXFFUTEBZ

/PPGOPEFT

/PPGFEHFT
C1 & C2: Building “Columbia” Knowledge Graph

• **Background**
  
  • For artificial intelligence and better search, many search companies have created a knowledge graph.
  
  • However there are few knowledge graphs in the public domain.

• **Project Goal**
  
  • A team will create knowledge graphs in several application domains (e.g., Finance, Medical, etc) by crawling public web pages, news, Twitter, Wikipedia, etc.
  
  • A team will need to design the way of efficiently crawling data set, store them in a limited space, and quickly searching for required data set with the indexing functionality.
C3 & C4. Graph Query Language

• **Background**
  • A variety of graph query languages have been invented. For example SPARQL, PGQL, Cypher, Gremlin. Each has strengths and weaknesses. Yet they all strive to make it easier for people to generate and quickly try graph queries without understanding the intricacies of procedural graph API’s.

• **Task Goals:**
  • Select a common set of use cases and design a graph query language that suits this set of use cases better than popular existing graph query languages. Provide an implementation of the graph query language you design. Justify your design.
  • Provide an innovative alternative graph query modality implementation. Ex. voice query, free-hand pen, gesture, textual natural language query, etc
  • Implement an existing graph query language on top of an existing graph API that lacks such an implementation. Demonstrate a compelling use of this implementation.

• **Steps:**
  • Demonstrate their tools working with the knowledge graph created by C1 & C2.
  • Require the use of an underlying graph query language or query engine. Where those do not already exist, the Pregel or graph pattern matching engine produced by another class project should be used. If that is not available, the projects here are expected to implement a minimal engine that that is sufficient to support queries that would be valuable to a partner project and support a compelling demo.
Background: The latest research paper from Supercomputing 2016 [Pan 2016] shows that 2 billion edges are processed on 3 GPUs of NVIDIA K40 with 16 GB memory by achieving around 300 GTEPS (Billion traversed edges per second) with 3 GPGPUs in a single node (24 bytes per edge). Due to hardware improvements and algorithm innovation, the performance is much improved over prior work.

However, this solution is not sufficient to reach trillion-scale graph processing. Though the number of GPGPUs was increased up to 6 GPGPUs in a single node, the throughput is not improved since the problem size cannot be increased due to GPU memory size being bounded by the 16GB of each GPU card.

This memory limitation is a significant problem when it is required that we handle data sets 1000 times larger (i.e. a trillion scale graph) when offloading graph processing to GPGPUs. To process trillion-scale data, we would need 50 TB of memory, if we assume we would need the same 50 bytes per edge of our previous efforts. Moreover, if we consider a graph where each vertex and edge has a set of properties evolving through time, we would need even more memory space to store them.

GPU-based Graph Analytics (cont’d)

• **Approach**: In order to work with trillion-scale graph, without purchasing thousands GPGPUs, a solution is to leverage the DRAM and SSD of the memory hierarchy.

• A team will build a scalable GPU-based processing framework that uses CPU resources to decompose an entire graph into a chunk of graphs that would fit with a GPU memory, with the CPU then continuously sending small chunks of graph to GPUs to perform analytics.

• If this chunk decomposition and data transfer on CPUs can completely overlap with accelerated graph analytics on GPUs, we can leverage enough memory capacity on the CPU side by utilizing DRAM and SSD.

• A team will tackle the challenging research problem of how we can partition the entire graph into suitable partitions that balance loads and fit in GPU memory. In order to build a highly scalable runtime of graph analytics on GPUs, we propose to further conduct fundamental research on optimizing the followings:  .
C7. Graph Workflow System

• **Background**
  - Over the years, graph analytics have been used for structural analysis.
  - However, in the typical customer projects, graph analytics are used for calculating a set of features for machine learning that are eventually used for classification and prediction.
  - Graph analytics and machine learning now come into play together.

• **Project Goal**
  - In this project a team will design and implement a workflow system that allows developers to specify a set of graph analytics and required machine learning methods.
  - A team might use an existing tool like TensorFlow, but they decide the underlying tool at the design phase.
C8. Graph Analytics with Deep Learning

• Background
  • Over the years, graph analytics have been used for structural analysis
  • Deep Learning has become one of the most powerful techs in Machine Learning.
  • Deep Walk/Node2Vec has proposed to deal with labeling problem in graph.

• Project Goal
  • In this project a team will leverage CNN or word2Vec technologies to propose a new method on how to analysis graph from a new aspect of view.
  • A team might use an existing tool like TensorFlow or Word2Vec, but they decide the underlying tool at the design phase.

(a) Input: Karate Graph

(b) Output: Representation
C9. High-Performance Dynamic Graph Database

- **Background**
  - Graph structure is usually dynamically changed over time, and also properties on vertices and edges tend to have time-series data set.
  - For this real requirements, no graph database fulfills this requirement

- **Project Goal**
  - A team will design and implement a highly scalable graph database that handles dynamic graphs efficiently by allowing vertices and edges to have time-series data set.
  - Since time-series data set tends to be huge, a team considers how to store those data set in a secondary storage like SSD, however bringing the high speed for reading.
C10. Python Library for High-Performance Graph Analytics and Graphical Models

• **Background**
  • When you need to process big graph, a python-based implementation totally does not work, and then you need to switch to C++ based implementation or distributed processing framework like Spark.
  • However, Python provides easy-to-use interface to developers, so there is always tradeoff between productivity and performance.

• **Project Goal**
  • A team will design and implement a set of building blocks that are fundamentally required for graph analytics in C++.
  • And then exposing fine-grained Python interface to developers, but behind the scene, high-performance C++ implementation runs instead.
  • The goal of this project is to build a Python interface and demonstrate whether they can handle billion-scale graph data.

![Building Blocks for Graph Analytics (C++)](image)
C11. Relational Databases or Big Data Databases to Graph Database Migration

• **Background**
  • Most legacy data today resides in relational databases, yet many newer use cases are better served by graph databases.

• **Goal**
  • Build and demonstrate a tool for creating a graph database. Where necessary provide a compelling user interface to allow an administrator to provide hinting. This conversion should consider schema constraints and indexes.
  • Extra credit: provide a tool that converts SQL queries in to efficient graph queries compatible with graphs created with the above tool.
C12. Exploration Environment of Data for Graph Analytics

• **Background:**
  • A key component of many big data and machine learning systems is a compelling productive user interface.

• **Task Goals:**
  • Provide a more compelling help or tutorial interface to RStudio, Jupyter, or similar platform
  • Provide a new compelling graph visualization that is compatible with RStudio, Jupyter or similar platform.
  • Integrate the work of one of the graph query language or SQL query language class projects in to RStudio, Jupyter or similar.
  • Provide a graph editor that interfaces with RStudio, Jupyter or similar and persists to a popular or class project graph database.
  • Provide a query visualizer that supports queries in a popular graph query language or a class project graph query language.
  • Provide a graph schema visualization to facilitate exploration of unfamiliar graphs. This should be integrated with a RStudio, Jupyter or similar as well as with graphs produced by other class projects.
  • Invent your own improvement to platforms like RStudio and Jupyter.
C13. Visual Query and Authoring Tool for Graphs

• **Background:**
  • Query is a major part of any database, and graph tools provides power query languages to conduct queries.
  • However, for some complex searching, it is hard for a general audience to write the queries.

• **Task Goal:**
  • A team will design and implement a web based graph authoring tool to draw graphs. (Step one)
  • Based on the authoring tool, they will build a visual query system, which allows users to draw graph patterns, and the system will convert it into queries.
  • The system will connect to graph database and graphical models, conduct queries, and render the query results.
C14. Visualization of Large Dynamic Network

• **Background**
  - Graphs are widely used to model real world applications, and in most of cases, graphs are dynamic: new nodes can be added and new connections can be formed between nodes over time.
  - Analysis of the dynamic aspect of network is valuable and yet challenging, especially on large scale, such as transaction network.

• **Task Goal**
  - A team will design and implement an interactive visualization system to visualize large dynamic network.
  - The system will be web based application (webgl technology is preferred), which connects to graph database to retrieve the large scale network data.
  - A novel visualization design will be proposed for users to explore the dynamic network.
  - Stream based graph analytics (such as community detection), will be implemented to enhance the system capability.
C15. Visual Exploration of Large Graph in Immersive Environment

• **Background**
  - Augment Reality is now becoming more popular and more and more devices have been available in the market.
  - So far, less research and few systems are available for exploring networks in such condition.

• **Task Goal**
  - A team will design and implement an augment reality application based on HoloLens.
  - The application will provide 3D visualization of large geo-spatio graphs.
  - Gesture based interaction will be provided for users to explore the graphs.
Task Area 4: Advanced Topics
Area 4 Tasks List:

D1: Computer Vision Enhanced Immersive Environment
D2: Mobile Vision and Behavior Analysis on mobile devices
D3: Medical data association with B cell receptor repertoire for cancer research
D4: Technical Analysis of Cryptocurrencies
D5 - D7: Design and Implementation of Novel Cryptocurrency based on Graphs
D8: Explainable ML: Visualization of Training Process of Deep Learning
D9: Explainable ML: Visual Analytics of Interactive Machine Learning
D10: Autonomous Learning: from Text to Vision
D11: Autonomous Learning: from Vision and Text to Knowledge
D12: Machine Reasoning with Large-Scale Bayesian Networks
D13: Strategic Planning with Game Theoretic Machines
D14: Mapping Suitable Applications on Quantum Computing
D15: Fundamental Software Framework on Quantum Computing
D1. Computer Vision Enhanced Immersive Environment

• **Background**
  • Augment Reality is now becoming more popular and more and more devices have been available in the market.
  • Computer Vision techniques, such as objection recognition, can further enhance the intelligence and improve the capability of what can be achieved.

• **Project Goal**
  • A team will design and implement an augment reality application based on HoloLens.
  • Some computer vision techniques will be implemented, such as object recognition and OCR.
  • The team encouraged to bring out any interesting usage scenarios on how these techniques can seamlessly enhance user experience of HoloLens.
D2. Computer Vision and Behavior Analysis on Mobile devices

- How to trade off between algorithmic complexity and performance?
- How to utilize hardware effectively

Acceleration CNN on Local Device

Apple A9X SoC, 12-core GPU
D3: Medical data association with B cell receptor repertoire for cancer research (with Harvard University)

Background:
- Human body has approximately $10^{10}$-$10^{11}$ B cells which targets potential any antigen that may be good or bad for human health
- The binding specificity of B cell receptors to antigens is determined by two short amino acid sequences

Task Goals:
- A team will design a Deep Learning (DL) model to get the patterns of B cell receptors in thousands of cancer patients
- First, the team need to find out the best DL architecture for separating different tumor types
- Second, the model should be extended to other clinical factors such as Overall Survival (OS)
D4. Technical Analysis of Crypto Currencies

• **Background**
  • Crypto Currencies have become very popular in the last few years.

• **Task Goal**
  • A thorough technical experiment and analysis of the security, scalability, usability, etc on existing and emerging cryptocurrencies
D5-D7. Design and Implementation of Novel Cryptocurrency based on Graphs

• **Background**
  • A common issue of the existing cryptocurrency is its scalability

• **Task Goal**
  • A novel GraphCoin or BlockGraph that are based on Graph properties may provide novel insight and advantage to the existing cryptocurrencies and BlockChain Technologies
  • One comparable example is the use of graph as Tangle to create IOTA cryptocurrency.

• **Background**
  
  • Machine learning methods have been proved to be very powerful for certain tasks, such as computer vision, and NLP.
  
  • However, in most cases, such as learning a neural network, it is still a black box about the model and the model itself is not explainable.

• **Task Goal**
  
  • A team will get familiar with a popular deep learning framework, such as TensorFlow and Pytorch.
  
  • They will design and implement a visualization system to visualize the training process of CNN (or RNN), based on any major deep learning framework.

- **Background:**
  - Machine learning methods have been proved to be very powerful for certain tasks, such as computer vision, and NLP.
  - Analyzing the training results (e.g., false negative, false positive) can help better fine tune the machine learning model.

- **Task Goal:**
  - A team will get familiar with a popular deep learning framework, such as TensorFlow and Pytorch.
  - They will design a visual analytics system to specifically analyze how the errors happened from the training result and interactively fine tune learning models.
D10 & D11. Autonomous Learning

Image Source: http://wonderforgood.com/category/visual-storytelling/
D12. Machine Reasoning with Large-Scale Bayesian Networks

- **Description:**
  - Probabilistic graphical models are a key element of modern machine learning. They can be used to compactly represent and solve problems in probabilistic inference (e.g. using Bayesian Networks).

- **Goal:**
  - Implement fast probabilistic inference in a graph database using e.g. expectation propagation (EP) and by developing an efficient data structure for conditional densities (both continuous and discrete).

Probabilistic understanding what might have caused an alarm to go off.

D13. Strategic Planning with Game Theoretic Machines

**Description:** Probabilistic graphical models can also be used to express and solve problems in game theory, using the Influence Diagram (ID) representation (an extension of Bayesian Networks to allow for decisions and utilities).

**Goal:** Implement a “game theory database” by allowing games to be specified and solved with graphs (Single-Agent IDs, Multi-Agent IDs, Networks of IDs). Games should support (at least) discrete conditional densities like in the figure.

![Influence Diagram](image)

Deciding on whether or not to bring an umbrella, based on the (potentially incorrect) forecast.

www.eecs.harvard.edu/~gal/tutorial4perPage.pdf

What Can A Quantum Computer Do Better?

Quantum computing will solve a class of problems that are unsolvable today, opening up a new realm of applications.

- Searching Big Data
- Designing Better Drugs & New Materials
- Machine Learning
- Cryptography