E6895 Advanced Big Data Analytics and AI Lecture 1:

*Introduction of Advanced Big Data and AI*

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

January 21, 2022
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>
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
https://www.youtube.com/watch?v=BV8qFeZxZPE
Course Grading

- Task: 45%
  - Teamwork: 1 - 2 students per team
  - Choose a task from 60 potential tasks
  - Language Requirement: C/C++, Java, JavaScript, Python, Perl
  - 3 milestones (45%): Presentation, Slides, Report and Source Code

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

- Research Study: 15%
  - 3 research paper presentations related to Advanced AI: Slides

- Class Participation: 10%
  - Attendance
  - Discussion (Asking/Answering Questions)

Task Sign-Up Spreadsheet is available until midnight 1/28
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.
## 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>01/21/22</td>
<td>1</td>
<td>Introduction of Advanced Big Data and AI</td>
<td></td>
</tr>
<tr>
<td>01/28/22</td>
<td>2</td>
<td>Big Data Foundations</td>
<td></td>
</tr>
<tr>
<td>02/04/22</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>02/11/22</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>02/18/22</td>
<td>5</td>
<td>Massive Data Analysis</td>
<td>Full-Brain AI (I) &amp; Green Earth (I)</td>
</tr>
<tr>
<td>02/25/22</td>
<td>6</td>
<td>Machine Reasoning</td>
<td>Financial Advisor (I) &amp; Healthy Life (I)</td>
</tr>
<tr>
<td>03/04/22</td>
<td>7</td>
<td></td>
<td>Advanced AI Study</td>
</tr>
<tr>
<td>03/11/22</td>
<td>8</td>
<td></td>
<td>Advanced AI Study</td>
</tr>
<tr>
<td>03/18/22</td>
<td></td>
<td>SPRING BREAK</td>
<td>Advanced AI Study</td>
</tr>
<tr>
<td>03/25/22</td>
<td>9</td>
<td>Advanced AI Platform</td>
<td>Advanced AI Study</td>
</tr>
<tr>
<td>04/01/22</td>
<td>10</td>
<td>Social &amp; Cognitive Analytics</td>
<td>Advanced AI Study</td>
</tr>
<tr>
<td>04/08/22</td>
<td>11</td>
<td></td>
<td>Full-Brain AI (III) &amp; Green Earth (III)</td>
</tr>
<tr>
<td>04/15/22</td>
<td>12</td>
<td></td>
<td>Financial Advisor (III) &amp; Healthy Life (III)</td>
</tr>
<tr>
<td>04/22/22</td>
<td>13</td>
<td>Advanced Artificial Intelligence</td>
<td>Advanced AI Study</td>
</tr>
<tr>
<td>04/29/22</td>
<td>14</td>
<td>Advanced Artificial Intelligence</td>
<td>Advanced AI Study</td>
</tr>
<tr>
<td>05/06/22</td>
<td>15</td>
<td></td>
<td>Final Project Workshop</td>
</tr>
</tbody>
</table>

Presentation Schedule shall be adjusted by on the distribution of tasks.
Contact Information and TAs

- **Professor Lin:**
  - **Office Hours and Location:**
    - Friday 9:30pm – 10:00pm (lecture room) or by appointment (500 Fifth Ave., Suite 2420, New York, NY 10110)
  - **Contact:** c.lin@columbia.edu

- **TAs:**
  - Wei Zhang (wz2363)
  - TBD

---

**Special Request — thanks:**

- If you may not take the class, please do not sign up the task spreadsheet.
- Please remove your name from the task sign-up sheet immediately when you drop the class.
- Please drop the class as early as possible, if you are not planning to take the class.
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

- **Machine Reasoning:**
  - Bayesian Networks
  - Game Theory Tools
  - Multimodal Analysis Platform

1. Graph Database Technologies
2. Network Analytics Technologies
3. Machine Learning Technologies
4. Machine Reasoning Technologies
Reference Advanced AI + Big Data Platform

- Terabyte-sized native GraphDB, supports trillion of vertices and edges
- ACID-compliant and distributed Graph database and analytics
- Asynchronous job scheduling (both Autonomous ML and GraphDB)
- Scalable, distributed Analytics, modular and expandable through plugins
- Cluster, Replication and High-Availability with disaster recovery
- Error and event Logging, Monitoring, Backup and Recovery
Task Area 1: Full-Brain Machines
Area 1 ‘Cognitive Machine’ Tasks List:

A1: Deep Video Understanding (Visual + Knowledge) — Face Recognition, Feeling Recognition, and Interaction
A2: Deep Video Understanding (Language + Knowledge) — Speech Recognition, Gesture Recognition, and Feeling Recognition
A3: Deep Video Understanding — Event and Story Understanding
A4: Humanized Conversation — Personality-Based Conversations
A5: Autonomous Robot Learning of Physical Environment
A6: Autonomous Task Learning via Mimicking
A7: Digital Human - Creation and Facial Expression
A8: Digital Human - Action
A9: Digital Human - Text-to-Audio, Lip Sync, and Audio-to-Text
A10: Human and Digital Human Interactions
A11: Feeling and Art Recognition
A12: Creative Writing & Story Telling
A13: Knowledge Learning & Construction
A14: Dreams — Simulating Brain functions while sleeping
A15: Self-Consciousness, Ethics, and Morality
Robot Cognition

- **Sensor Information**
  - Audio
  - Visual
  - Infra-red

- **Artificial Empathy**
  - Motion
  - Expression

- **Emotion Perception**
  - Understanding

- **Motor Mimicry**

- **Emotion Contagion**

- **Multi-Modality**
  - Audio
  - Visual
  - Infra-red

- **Robot**
  - Head Tactile Sensor
  - Microphones
  - Inertial Unit
  - Joint Position Sensors
  - Hand Tactile Sensor
  - Infra-red
  - Video Cameras
  - Force Sensitive Resistors
Emotion and Cheers
How Robot cheers you up

A complete system combining these for video understanding

- Visual Recognition
- Speech Recognition
- Knowledge Graph
- Face Recognition
- Emotion Recognition
- Speaker Identification
- Relationship Inference
- Event and Action Understanding

Potentially Target at NIST Deep Video Understanding 2022
A4. Humanized Conversation with Personality

Description:
- Virtual Agents are progressing fast and entering people’s life. However, the voice presented by the agents are mostly ‘flat’ — like machines.
- The first step to make virtual agents to be like human is to add the “personality” aspect in conversation.

Goal:
- Create Personality-based Speaking Model Text for Conversation

Advanced Goal:
- Modify the Speech Tones to reflect Personality
A5. Autonomous Learning of Physical Environments

Description:

- Simultaneous Localization and Mapping (SLAM) refers to the problem of incrementally building the map of a previously unseen environment while at the same time locating the robot on it.
- Active localization was proven that picking actions to minimize the localization’s uncertainty would result in a better localization than using a passive approach.
- Active SLAM augments this approach to the SLAM problem, and it can be defined as the paradigm of controlling a robot which is performing SLAM to reduce the uncertainty of its localization and the map’s representation.

Goal:

- Robot Awareness of Physical Environments
- Robot Action with Environments
A6. Autonomous Learning of Tasks via Mimicking

**Description:**
- Machine learning to act based on actions of human
- Watch how human activity in an environment and then learn how to behave by itself.

**Goal:**
- Observation and Action Extraction
- Reinforcement Learning to correct own actions
A7 - A10. Digital Human
A11. Feeling and Art Recognition

• **Background**
  - Let machine to feel and appreciate arts like human

• **Project Goal**
  - A team will work on the subjective machine feeling of visual information
  - Allow machines to interpret arts.
A12. Creative Writing and Story Telling

• **Background**
  - Overwhelming real-time information on media.
  - Automatic writing and telling a story based a set of news articles.

• **Project Goal**
  - A team will design and implement a platform that conducts data mining on various related media of a field.
  - Using NLP to summarize key text information.
  - Using visualization to create charts and graphs.
  - Automatically create descriptions
A13. Knowledge Learning and Construction

The boy said:

“Airplane”

“Grandma”

“Grandma is in Taiwan”

“Auntie is also in Taiwan”

“I like grandma”

“I like grandpa”

“I like grandma and grandpa”

Image Source: http://wonderforgood.com/category/visual-storytelling/
Memory with Knowledge Graphs

Graph platform works like the human mind, connecting the dots when comprehending.

Image Source: http://wonderforgood.com/category/visual-storytelling/
A14. Dreams Simulating Brain Functions while Sleeping

• **Background**
  • When human sleeps, our brain works on ‘storing’ the massive information we see, hear, and learn during the day time into ‘storage’
  • Our brain would later on organize (in a bizarre way) to create dreams.

• **Project Goal**
  • A team will work on simulating how brain functions during sleep.
  • Create ‘dreams’,

---

*Why We Sleep*
Matthew Walker, PhD
"A book on a mission... recommended for nighttime reading in the most pragmatic sense." — The New York Times Book Review
A15. Self-consciousness, Ethics and Morality

• **Background**
  • Consciousness is how robots know its own existence
  • Can robot have self-identification?

• **Project Goal**
  • Simulates empathy
  • Simulates ethics and morality
Task Area 2: Financial Advisors

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)
Area 2 ‘Finance Advisor’ Tasks List:

B1: Market Intelligence — Constructing Financial Knowledge Graphs
B2: Market Intelligence — Company Environmental, Societal, and Governance Performance
B3: Market Intelligence — Event Linkage and Impact Prediction
B4: Market Intelligence — Alpha Generation from Alternative Sources
B5: Advance KYC — Customer Profiling based on Personality, Needs, and Value
B6: Advanced KYC — Customer Behavior Prediction
B7: Investment Strategy — AI Trader (Foreign Exchange)
B8: Investment Strategy — AI Trader (Stock Markets)
B9: Investment Strategy — Automatic Dynamic Asset Allocation
B10: Customer Interaction — Customer Communication Strategies
B11: Customer Interaction — Insurance Product Sales & Marketing Strategy
B12: Automatic Story Telling for Marketing
B13: Automatic Market Competition Analysis
B14: Automatic Consumer Sales Leads Finding
B15: Human Capital Growth Recommendations
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
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)

- **Rebalance**
  - 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></td>
</tr>
<tr>
<td>• Use Machine Learning and Deep Learning, based on historical economic numbers, find out how factors impact financial markets.</td>
<td></td>
<td>• Explainability of ‘what ifs’ to customer to the customer.</td>
<td></td>
</tr>
</tbody>
</table>

**Data**
- Product Data
- Market Data
- Historical Economic Data
- Industry-related Data

---

**Data**
- Customer Data
- Interaction Data
B1 - B4: Market Intelligence

Impact score of news on each stock

News to be analyzed

News Ranking List
- Apple Beats Sales Estimates; Shares Rise
- Apple Earnings Show Growing Immunity to Smartphone Market
- T-Mobile and Sprint CEOs State Case for Merger at FCC
- Sprint, T-Mobile Three Big Takeaways
- Should T-Mobile and Sprint Be Allowed to Merge? Not If You Want an App

News Article:

The leaders of T-Mobile US Inc. and Sprint Corp. appeared at the Federal Communications Commission on Tuesday to begin laying the groundwork for their proposed $26.5 billion merger, according to an agency official who spoke on condition of anonymity because the meeting wasn’t public.

T-Mobile’s John Legere and Sprint’s Marcelo Claure met with FCC officials and laid out much the same case that the companies have presented in public since announcing T-Mobile’s proposed purchase of its smaller rival on April 29, said the official.

Meetings had been scheduled with Commissioner Michael O’Rielly, a Republican, and Commissioner Jessica

The need for market intelligence is crucial in understanding the impact of news on stock prices.
Knowledge Graphs
Example of Building and Utilizing 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.
B5 - B6: Advanced KYC — Customer Profiling and Behavior Prediction
Anita avatars are earning: $2,503.26

- ANITA-324658
  PER $22,630 EARN: $-467.51

- ANITA-253758
  PER $30,178 EARN: $1,106.20

- ANITA-247917
  PER $31,809 EARN: $350.48

- ANITA-428339
  PER $39,494 EARN: $620.17

- ANITA-164762
  PER $29,395 EARN: $-17.07

- ANITA-450214
  PER $36,088 EARN: $178.12

- ANITA-247502
  PER $46,253 EARN: $318.35

- ANITA-267139
  PER $21,287 EARN: $44.81

- ANITA-544716
  PER $46,442 EARN: $166.03

- ANITA-418870
  PER $28,764 EARN: $21.32

- ANITA-432722
  PER $24,712 EARN: $132.59

- ANITA-208134
  PER $16,576 EARN: $49.76
Anita 267139
-- an Adventurous AI Trader

Specialized at: EUR-USD

Knowledgeable of: Oil, Gold and Twitter

Strategy Learning Frequency at: 2.0 hours

Original: $1,000.00, Current: $1,404.50, Performance: Gain $404.50

Activities

<table>
<thead>
<tr>
<th>Time</th>
<th>Action</th>
<th>Cash</th>
<th>Unit</th>
<th>Balance</th>
</tr>
</thead>
<tbody>
<tr>
<td>2017-10-12 13:45:05</td>
<td>Sell 50,000</td>
<td>$1,404.50</td>
<td>0</td>
<td>$1,404.50</td>
</tr>
<tr>
<td>2017-10-12 12:57:25</td>
<td>Buy 100,000</td>
<td>-$57,792.00</td>
<td>50,000</td>
<td>$1,388.50</td>
</tr>
<tr>
<td>2017-10-12 11:18:10</td>
<td>Sell 100,000</td>
<td>$60,577.00</td>
<td>-50,000</td>
<td>$1,372.00</td>
</tr>
<tr>
<td>2017-10-12 11:11:55</td>
<td>Buy 100,000</td>
<td>-$57,822.00</td>
<td>50,000</td>
<td>$1,366.00</td>
</tr>
<tr>
<td>2017-10-12 09:08:05</td>
<td>Sell 100,000</td>
<td>$60,586.00</td>
<td>-50,000</td>
<td>$1,310.00</td>
</tr>
<tr>
<td>2017-10-12 08:34:40</td>
<td>Buy 100,000</td>
<td>-$57,935.00</td>
<td>50,000</td>
<td>$1,267.50</td>
</tr>
</tbody>
</table>
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
B12. Automatic Story Telling for Marketing

• **Background**
  • Using Raw Materials in an organization to create marketing materials

• **Project Goal**
  • A team will design and implement a platform that uses data in an organization
  • Automatically organize the information on a particular topic
  • Using visualization to create charts and graphs.
  • Manually or automatically create descriptions
  • Creates a video to tell story
B13. Automatic Market Competition Analysis

• **Background**
  • Automatic searching internet to find competitor’s information

• **Project Goal**
  • Automatic extraction of key information
  • Automatic compare key products and services of the company
  • Finding financial performance if those are available.

- **Project Goal**
  - Using Public Raw Materials on Social Media to find potential customers

“Booked my trip to @Overviewproject summit in NYC. 5 days of peace, love, and documents.”

“Have to make a last second business trip to NY.”

“Heading to NYC”

List of NYC-bound travelers

Trait Profiles

Coming to NYC? Follow us for personalized tips on fun bars, broadway shows, and even free kayaking: http://bit.ly/1ap1ub6

“Thanks for the tip!”

“Thanks for sharing with me. I will check it out.”

“Tweets

© CY Lin 2022, Columbia University
B15. Human Capital Growth Recommendation

• **Project Goal**
  • Automatic analyzing a person’s personality and goals
  • Analyzing similar successful people from public datasources, e.g., LinkedIn.
  • Creating Knowledge Graphs that makes successful on goals
  • Suggesting what to learn to be competitive
Task Area 3: Healthy Life
Area 3 ‘Healthy Life’ Tasks List:

C1: Precision Health — Gene and Protein Analysis of Network, Pathway, and Biomarkers
C2: Large-Scale System for Human Genome Analysis
C3: Secure Patient Data
C4: Medical Image Analysis
C5: Drugable Targets for Precision Medicine
C6: Virus Mutations and Function Prediction
C7: Microbe and Disease Knowledge Graph
C8: Disease Symptoms Knowledge Graphs
C9: Virtual Doctor
C10: Knowledge Graphs for Gene Interaction and Disease Similarity
C11: Biomedical Knowledge Construction and Extraction
C12: Generating Gene Therapy
C13: Molecular Drug Synthesis
C14: Protein Interaction Predictor
C15: Aging Impacts
Life is composed of graph of atoms
Central Dogma of Biology
The Emergence of Digital Biology
AI Tools power Digital Biology
C1: Precision Health - Multiple Omics

• **Background**
  • Utilizing whole genome information can provide valuable information to patients

• **Goal**
  • Study open source whole genome data and explore their impact on disease prediction.
C2. Large-Scale System for Human Genome Analysis

- **Background**
  - Size of whole genome data is very large

- **Goal**
  - Study is needed on big data systems for genomic analysis and other comic analysis
C3. Secure Patient Data

• **Background**
  • Health record privacy is always a concern in practice

• **Goal**
  • Utilize novel technologies such as encrypted-domain data mining may be helpful in establishing a system acceptable in practice.
C4. Medical Image Analysis

- Flow of whole process

  Raw Pap Smear Image → Preprocessing: Cell Segmentation → Status Assessment → Decision

- Best medical image DL model: U-Net

- K-means for image clustering

- U-Net Cell Segmentation

- Characteristics of cancer cells

<table>
<thead>
<tr>
<th>Normal</th>
<th>Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Normal Cells" /></td>
<td><img src="image2" alt="Cancer Cells" /></td>
</tr>
<tr>
<td>Large, variably shaped nuclei</td>
<td>Many dividing cells; Disorganized arrangement</td>
</tr>
<tr>
<td>Variation in size and shape</td>
<td>Loss of normal features</td>
</tr>
</tbody>
</table>
C5. Drugable Targets for Precision Medicine

• Background
  • Next-generation medicine will be based on personal genome data, proteome data, and pathway prediction.
  • It’s a continuous challenging problem to explore the appropriate drugs for diseases

• Goal
  • Utilize Knowledge graph of disease and drugs.
  • Use the pathway analysis of patients to identify key variants
  • Analyze the potential drug targets
C6. Mutations and Function Prediction

• Background
  • We have been monitoring COVID-19 worldwide mutations since Feb 2020.
  • More than 5,000,000 virus strains have been sequenced.
  • Continuous monitoring of large-scale data become more and more challenging.

• Goal
  • Keep exploring key algorithms for virus mutation classifications.
  • Use Protein function prediction tools to estimate the mutated virus impact.
C7. Microbe and Disease Knowledge Graph

• **Background**
  • Microbes are tiny living things that are found all around us and are too small to be seen by the naked eye. They live in water, soil, air, and in human body, which is also called microorganisms. The most common types are bacteria, viruses and fungi.
  • Researches indicate that microbes and human health have strong correlations.

• **Goal**
  • Find the similarity of microbes and similarity of diseases.
  • Build the correlation network of microbes and diseases to help diagnose potential health conditions.
C8. Disease Symptoms Knowledge Graph

- **Background**
  - Clinical symptoms are usually used as clues to identify diseases.
  - A disease could have multiple symptoms in every course of disease, and multiple diseases could have similar or the same symptoms. The above reasons enhance the difficulty on identify one’s disease.

- **Goal**
  - Collect the symptoms of the diseases and find the similarity either/both between the symptoms or/and between the diseases
  - Predict one’s disease by using symptoms, course of disease, and other related information.
C9. Virtual Doctor

• Background
  • There is a wide variety of medical information on the internet and it is sometimes difficult to obtain the specific piece of information. The chatbot is designed for the general public with the specific questions, and the medical chatbot would be able to answer the questions based on the knowledge graph from the database. By sorting the relationships between diseases and genes with the knowledge graph, the general public can access the information more systematically and efficiently.

• Goal
  • Since acquiring some medical knowledge gradually becomes a must for the general public, there is a great potential in developing a medical chatbot. The medical chatbot will assist the general public in understanding the relationships between diseases and genes in order to improve general public’s medical knowledge and understanding.
C10. Knowledge Graph of Gene Interaction and Disease Similarity

• **Background**
  - Understanding the genetic networks and their associations in diseases is one of the important objectives of biological researchers. The knowledge graph serves as a powerful tool to investigate this topic.

• **Goal**
  - Construct and visualize knowledge graphs demonstrating associations among genes based on disease similarity.

The top 20 targets related to APOE based on a computation of similar target-to-disease connections and overall association scores. Clicking on a related target bubble will show you a list of the diseases used to calculate the similarity between the two targets. For more information on this visualisation and how we compute similarity scores, please read our documentation on target and disease similarity.

Source: Open Targets
C11. Biomedical Knowledge Construction and Extraction

• Background
  • In now days, it is common for patients to access biomedical information on the internet to learn medical information, make decisions, and verify doctors’ advices. However, the medical text is often written for medical experts and thus a burden for the general public to understand. With the development of biomedical text simplification, patients without adequate literacy can easily comprehend the medical content and take necessary steps in time.

• Goal
  • To enable the general public to understand the medical, biological, and genetic information, computationally simplifying the biomedical text plays an important role.
C12. Generating Gene or Immuno Therapy

• **Background**
  - CRISPR, allowing to precisely edit the genome of cells by inducing double-stranded DNA (dsDNA) breaks at specific loci, is both an efficient and cost-effective technological tool.
  - But how to design perfect sgRNA for detect cell DNA and without any off-target is challenging
  - We need use outsourcing data to make a deep learning algorithm to solve this problem.

• **Goal**
  - Use TDC open source data to predict what kind of sgRNA have ability to edit or repair cell DNA. And design an auto encoder or GAN to generate template sequence.

---

**Basic Considerations**

- **GC Content**
  - 40%-80%

- **Length**
  - 17-24 Base Pairs

- **Potential Off-Target Effects**
  - 17-24 Base Pairs

• **Background**
  • Generate small molecular by deep learning will not be hard. In fact, these molecular are hard to be manufacture or synthesis. We need to design a algorithm to simulate chemical reaction and predict molecular synthesis feasibility.

• **Goal**
  • Use open source data to predict the molecular can be produce or not, and try to simulate the synthesis processes by molecular properties.
C14. Protein Interaction Prediction

**Background**
- Protein-protein interactions (PPIs) are useful for understanding signaling cascades, predicting protein function, associating proteins with disease and fathoming drug mechanism of action.
- Currently, only ~10% of human PPIs may be known, and about one-third of human proteins have no known interactions.

**Goal**
- Predicting Protein-Protein interactions from gene sequences or protein structures
C15. AI Exploration and Understanding of Aging

• **Background**
  • Aging is a major impact on human. Recent studies have been giving more and more information on how aging functions and whether it’s possible to try to delay or even reverse some functions.

• **Goal**
  • Study the mechanisms causing aging in the gene and protein level.
  • Use the protein structure prediction, protein-protein interaction, protein-drug binding tools to explore...
Task Area 4: Green Earth and Advanced Topics
Area 4 ‘Green Earth’ Tasks List:

D1: Distributed Solar Power Load Forecasting and Predictive Maintenance
D2: Distributed Wind Power Load Forecasting and Predictive Maintenance
D3: Power Flow Optimization
D4: Smart Grid Pricing Strategy
D5: Cybersecurity of Smart Grid
D6: Stimulating Crop Growth
D7: Electronic Car Sensing and Predictive Maintenance
D8: Autonomous Driving
D9: Smart City of Connected Cars
D10: Social Policy Monitoring
D11: International Relationships and Policy Monitoring
D12: Mobile Cognition
D13: AI Chip Design
D14: Visual Exploration in Immersive Environment
D15: Computer Vision Enhanced Immersive Environment
AI + Big Data Makes Smart Grid Possible

- Energy Forecasting
- Price Forecasting
- Risk Analysis
- Revenue Management

Market Operators

- Wide area control
- State Estimation
- Stability enhancements
- Real-time Monitoring
- Congestion Management
- Visualization

Transmission Operators

- End Users
  - Demand Response
  - Load Forecasting
  - Real-time Energy Management
  - Customer Behaviors

- Generation Control Center
  - Generation Planning
  - Economic Load Dispatch

- Distribution Operators
  - Real-time operations
  - Equipment health monitoring
  - Asset management
  - Planning
  - Forecasting

- Aggregators/other Businesses
  - Flexibility Assessment
  - Risk Analysis
  - Forecasting
Key factors contribute to the complexity of the electric grid

➢ The variability and intermittency of renewable generation.
➢ Decreased frequency response capability and decreasing system inertia.
➢ Changing load patterns and unpredictability.
➢ The need to manage vastly increasing number of endpoints.
➢ Growing cyber attack risks to the electric grid.

Department of Energy. Smart Grid System Report, Nov 2018
Power Quality Analysis

➢ Power system modeling
➢ Power system classifying
➢ Point of event occurrence into one unified frame
➢ Equipment sensitivity to the event disturbance
➢ Power quality events detection and characterizing
  ➢ E.g. sag, swell, outage, harmonic, notch, flicker, impulse, etc.

Voltage Waveform ➞ Fourier and Wavelet-transform Based Feature Extraction ➞ Fuzzy Neural Network Decision Making ➞ Detection and Classification of a Disturbance

Detection and Classification ➞ Characterization ➞ Equipment Sensitivity Study ➞ Event Location Prediction
Predictive Maintenance

➢ Apply machine learning to historical power system data to reduce operating costs and failure risk
➢ Avoid or minimize the downtimes and reduce associated costs
➢ Optimize the periodic maintenance operations.
➢ Health indicator by machine learning
   ➢ Classification – health indicator predicts what is the probability of failure in the future.
   ➢ Regression approach – health indicator predicts how much time is left before the next failure.

Flowchart:

1. Historical power system data
2. Feature extraction
3. Machine learning algorithms
4. Failure Prediction
5. Health indicator

Conceptual diagram:

- Perception
- Recognition
- Sensors
- Representation
- Memory
- Comprehension
- Strategy
D1. Distributed Solar Power Load Forecasting and Maintenance

• **Background**
  - Situation of Solar Power Plants varies and are time dependent. Power companies need good prediction on conditions and exclude anomalies in short time.

• **Task Goal**
  - Predict Solar Power Generations based on weather data
  - Anomaly Detection of Solar Power Plants
  - Predictive Maintenance of solar power plants.
D2. Distributed Wind Power Load Forecasting and Maintenance

• **Background**
  • Situation of Wind Power Plants varies and are time dependent. Power companies need good prediction on conditions and exclude anomalies in short time.

• **Task Goal**
  • Predict WindPower Generations based on weather data
  • Anomaly Detection of Wind Power Plants
  • Predictive Maintenance of solar power plants.
D3. Power Flow Optimization

• Background
  • Transmission is key to a Low-Cost Decarbonized US Grid

• Task Goal
  • Study the optimal stragglers for power flow
  • Simulate various scenarios

https://www.greentechmedia.com/articles/read/study-transmission-is-the-key-to-a-low-cost-decarbonized-u.s-grid
D4. Smart Grid Pricing Strategy

• **Background**
  • Pricing strategy can be a way to optimize consumer behavior
  • After more and more cars and IoT devices rely on power, it’s critical to influence customer behavior to optimize use of power grid

• **Task Goal**
  • Implement methodologies that can help change customer behavior
  • Game theory is a possible solution.
  • Other solutions should be also considered.

D5. Cybersecurity of Power Grid

• **Background**
  • Power Infrastructure is among the most obvious target of cyber attacks.

• **Task Goal**
  • Study NIST Framework for Critical Infrastructure Cybersecurity
  • Propose Solutions that can help reduce the attach risks
D6. Stimulating Crop Growth

• **Background**
  - Machine learning in image recognition crop growth status and crop management strategy

• **Task Goal**
  - Establish the ideal growth model of crops
  - By using the image recognition find the crop growth status. Mark the unwanted growth. Give suggestion of the location to do the pruning.
  - Using climate and soil data to give suggestion for irrigation and fertilization
D7. Electronic Car Predictive Maintenance

• **Background**
  • Car Fixing and Predictive Maintenance are important issues in the automobile industry
  • Pure electronic car is relatively new

• **Task Goal**
  • Model Knowledge Graphs of the functioning of subsystems in an electronic car
  • Studying the sensors available in novel cars
  • Detection Car Problems from sensors
  • Prediction of maintenance requirements based on sensor signals
  • Incorporating other information such as environmental and demographical patterns into consideration.
D8. Autonomous Driving

• **Background**
  • Autonomous Driving is becoming mature
  • Autonomous Driving has to consider the complex situations in the road.

• **Task Goal**
  • Explore and experiment on autonomous driving technologies
  • Utilizing sensors to come up with optimal strategies to drive the car.
  • Build a Game Theory and Bayesian Network model to consider the complex behaviors on the road.
D9. Smart City of Connected Cars

• **Background**
  - Cars are being connected with all kinds of systems in a city

• **Task Goal**
  - Identify possible connections via connected cars
  - Exploring the usefulness and the impact of such connections
  - Simulating such impact of connected cars in a smart city
  - Security requirements for connected cars
D10. Social Policy Monitoring

• **Background**
  • Social Issue and Policies have been impacting people’s life

• **Task Goal**
  • Information Mining from Social Media to analyze the impact of social policy.
  • Analyze the effectiveness of policy making
D10. International Relations and Policy Monitoring

• **Background**
  • Relationships between countries have been a major issue toward world policy changes

• **Task Goal**
  • Large-Scale Data Mining of international relationship evolutions
  • Visualize and create early alert of relationship changes
D12. Mobile Cognition in Complex Scenarios

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

Acceleration Computer Vision and Audio Recognition on Mobile Device
Example: Mobile Cognition in complex scenario
D13. AI Chips

• **Background**
  • Hardware AI Chips design is getting more and more popular

• **Task Goal**
  • Explore the functions and roadmaps of AI Chips
D14. Visual Exploration 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.
D15. 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.