AI- and Data-Driven Support
For FINTECH

Beng Chin OOI

www.comp.nus.edu.sg/~ooibc
Contents

• AI and QoS

• GIT for Data: Facilitating AI Services without Data Sharing
  o Our Software Stack

• Financial News Analytics
  o An Example on using AI/ML
MAS Fintech Challenges

- Integrated cross platform payments
- Automated insurance claims
- Mobile banking for visually impaired
- Seamless season parking fee payments
- Aggregated financial statements manager
- Smart portfolio management
- Cashless schools
- Engaging investor education
- Robot-advisor 2.0
- Secure digital authentication
- Common industry APIs
- Data-driven investment solutions
- Fraudulent behaviour detection and prediction
The Reality of Exploiting AI

The actual implementation of the ML algorithm is usually less than 5% lines of code in a real, non-trivial application.

The main effort (i.e. those 95% LOC) is spent on:
- Data cleaning & annotation
- Data extraction, transformation, loading
- Data integration & pruning
- Parameter tuning
- Model training & deployment
  … …

This blurs the line between DB and “non-DB” processing, and calls for better integration.

These are what we have been doing!
The BIG Data Analytics Pipeline*

AI
In the nutshell

Training → AI → Prediction
Preparation and QoS

Data Cleansing → Training → AI → Prediction → QoS Monitoring
The Workflow

Data Collections

- Data Cleansing
- Training
- AI
- Prediction

Query Interactions

- QoS Monitoring
Data Collections

- Original Data
- Data Cleansing
- Training Data
- AI Model
- Prediction History
- QoS Statistics

Query Interactions

1. Data Collections
2. Training
3. AI
4. Prediction
5. QoS Monitoring
Scaling AI

How to manage data for multi-user/application?

Original Data
Training Data
AI Model
Prediction History
QoS Statistics

Data Collections

Query Interactions

Data Cleansing
Training
AI
Prediction
QoS Monitoring

How to coordinate collaborative data cleansing?
How to support machine/deep learning under privacy concern?
How to preform timely quality control for online prediction?
ForkCloud

GIT FOR DATA
Data and Learning

- Due to PDPA/GDPR, data is not widely available
- Even if it is stored in the cloud, data cannot be seen by external AI experts and developers
- Without data, models cannot be well designed or trained
- No Free Lunch Theorem [1997]
  - there is no useful model/algorithm for all the data distributions
Big Data Analytics in Banking

- Risk Analytics
- Customer Experience
- Operations Optimization

Scattered Data
No Single View of the Customer
Fraud Identification
Governance
Risk Analytics

Banking Data

• Customer basic information
• Customer asset status
• Customer transaction history
• Customer trading pattern (e.g., trading accounts, trading frequency, ...)
• Customer credit history
• .......
Risk Analytics

Banking Data

- Customer basic information
- Customer asset status
- Customer transaction history
- Customer trading pattern (e.g., trading accounts, trading frequency, ...)
- Customer credit history
- .......

External Information

Third-party Analytics
Facilitating AI Services

Git for data

- Data-driven services to facilitate development of domain-specific AI solutions
- Data cleansing, curation and integration services
- Data versioning to facilitate collaboration, data provenance and improvement of quality data

Applications / Target market

- “Git” for data – cleaning and curation, privacy-preserving crowdsourcing for service, data protection and provenance, data immutability
- AI-as-a-Service in the cloud
- Tools for data scientists and business analysts to conduct analytics on sensitive data
The ForkCloud Stack

Access Control

AI Service

AI Development

Data Cleansing

Data Store

Open Access to Individual Tools

Automatic ML/DL

Deep Learning

Data Cleansing & Integration

Crowdsourcing

Cohort Analysis

Data Access Control / User Permissions / Sandbox

ForkBase
Git for Data

Third-party/Crowdsourced AI Development
Software Stack and Architecture

ForkBase

Data Owner

Data Processing Vendor

Analytics Provider

Access Control

CDAS
Crowdsourcing

CohAna
Cohort Analysis

PANDA
AI as a Service

Rafiki
Automatic Machine/Deep Learning Service

Apache SINGA
Deep Learning

DICE
Data Cleansing & Integration

ForkBase
Financial News Analytics

For fintech applications
World Stock Market Performance
MSCI All Country World Index with Selected Headlines from 2012

- "Germany's Economy Shrinks, Recession Fears Loom in Europe"
- "Euro Zone Edges Closer to Recession"
- "Oil Falls on Sputtering Economic Growth"
- "Spanish Crisis Deepens"
- "Manufacturing Downturn Spreads Gloom across Asia, Europe"
- "Wheat Soars after Russian Crop Failure"
- "Europe Poses Global Recession Threat"
- "China Slowdown Could Spur Global Recession"
- "Greek Economy Shrinks Further"
- "Moody's Downgrades Global Banks"
- "Syrian Violence Escalates into Civil War"
- "UN Urges Action to Avert Global Food Crisis"
- "IMF: Global Recession Risk Grows"
- "World Stocks Set to End Year with Gains of 15%"
- "Japan Falls into Recession"
- "Oil Spikes on Iran Fears"
- "Higher Oil Prices Will Sap World Growth"
- "More Bad News for the UK Economy"
- "The Bottom Line: Central Bankers Are Worried"
- "World Bank Sees Long Crisis Effect"
- "EU Budget Talks Collapse"

MSCI ALL COUNTRY WORLD INDEX (NET)
Annualized returns as of December 31, 2012

- 1 Year: 16.13%
- 3 Years: 6.63%
- 5 Years: -1.16%
- 10 Years: 8.11%

Source: MSCI.
In US dollars. Index is not available for direct investment. Performance does not reflect the expenses associated with management of an actual portfolio. Past performance is not a guarantee of future results.
UK foreign office says it's exploring a 'sanctions regime' with the EU against Saudi Arabia

CNBC • today

- Trump refuses to sanction Saudi crown prince over death of Khashoggi
  Daily Mail • today

- Trump ally says Saudi sanctions must be in budget, as lawmakers rip Trump's defense of crown prince
  Washington Post • 3 hours ago

- US Senate Demands Trump Explore Saudi Sanctions Over Khashoggi Murder
  Sputnik International • 3 hours ago
Risk

IsoRay (ISR) Downgraded by Maxim Group to Hold
Fairfield Current • 3 days ago

Cracker Barrel Old Country Store (CBRL) Lowered to Hold at Maxim Group
Baseball Daily News • today

Maxim Group Downgrades IsoRay, Inc. (ISR) to Hold
StreetInsider.com • 11 days ago
Opportunity

Huawei Mate Pro 20: What would a real mate do?
Irish Examiner • yesterday

- Huawei set to launch Mate 20 Pro in India on November 27
  Times Now • today

Huawei Mate 20 Pro with a 5G-ready chipset to be launched in India on 27 November
Firstpost • one hour ago
Opportunity

Unilever to acquire 75% of Italian personal care business Equilibra
WebWire • 4 months ago

Unilever To Acquire 75% Stake In Italian Personal Care Firm Equilibra | ESM Magazine
ESM - The European Supermarket Magazine • 4 months ago

How Unilever Could Benefit From Acquiring Equilibra
Forbes Now • 4 months ago
Over 30,000 news publishers indexed by newsapi.org
Many articles

- China says certain countries trying to force views on others at recently-concluded Apec summit
  The Straits Times • yesterday

- China’s former trade negotiator questions tariff strategy, Latest World News
  The New Paper • yesterday

- US renews unfair trade accusations against China
  Channel NewsAsia • 4 hours ago

- China turns focus on US with claim its Apec amendment was vetoed
  South China Morning Post • today
Few articles

Why the Lantau Tomorrow Vision needs a rebranding
South China Morning Post · 15 days ago · Opinion

Stop Lantau reclamation, say dolphin activists, after mega-bridge marine park deemed 'useless'
Hong Kong Free Press · 28 days ago

Sentek Marine Named Singapore’s Biggest Bunker Supplier by Volume in 2017
Ship & Bunker · 10 months ago
<table>
<thead>
<tr>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>- News deduplication</td>
</tr>
<tr>
<td>- Knowledge graph construction</td>
</tr>
<tr>
<td>- Risk/opportunity discovery</td>
</tr>
<tr>
<td>- Stock price/index prediction</td>
</tr>
</tbody>
</table>
Knowledge graph if few articles

Clustering if many articles

Risk/opportunity prediction
News deduplication by clustering

- **Document Clustering**
  - *Extract representation of the article/document*
    - TF.IDF, word2vector -> Document vector, etc.
  - *Apply clustering algorithms*
    - K-means, DBSCAN, etc.

- **Challenges**
  - *Too many articles are generated every day* -> Big data
  - *Clustering algorithms are expensive*

- **Hierarchical clustering**
  - *Identify articles of the same company*
  - *Apply clustering algorithms over the articles about the same company*
News deduplication by clustering

Hierarchical clustering
Knowledge Graph

• Knowledge
  • Relation tuples, \((A, \text{ relation}, B)\).
    • Tim Cook, CEO of Apple, has attended the meeting. \(\rightarrow (\text{Tim Cook, CEO, Apple})\)
    • Foxconn that makes Apple’s iPhones... \(\rightarrow (\text{Foxconn, Supplier, Apple})\)

• Application
  • News extrapolation
    • \((\text{Person A, CEO, Company B}), (\text{Company B, Supplier, Company C})\)
    • If the big company C has risks, then the small company B is likely to have troubles
  • Stock price/index analysis
    • The drop of price for Company C would affect that of Company B
Knowledge Graph Construction

• **Subtasks**
  
  • **Named Entity Recognition**
    
    • Locate and classify *named entity* mentioned in unstructured text into pre-defined categories, such as person, organizations, locations...
    
    • e.g., [John]_{person} bought 500 shares of [Apple Inc.]_{organization} in [2015]_{time}.
  
  • **Entity Linking**
    
    • Link the mention into the entity in the knowledge base.
    
    • e.g., ”Paris is the capital of France” → Paris is referring to this entity.
  
  • **Relation Prediction**
    
    • Classify the semantic relationship between entities
    
    • e.g., "Yesterday, Foo Inc. announced their acquisition of Bar Corp." → (Foo Inc., acquire, Bar Corp)
Knowledge Graph Construction

• NLP is challenging
  • Ambiguous words
    • e.g., Apple & apple
  • Uppercase words impacts
    • e.g. A and B Jointly Developed ... "B Jointly Developed" can be mistakenly recognized as one entity
  • Multiple relations in one sentence
    • e.g., Mary married with a guy, who is the chairman of Bar Corp.
Knowledge Extraction --- Deep Learning

• **Supervised learning**
  • *Select the interested relations*
    • *Key Person (e.g., employee, founder, CEO, chairman, head, etc)*
    • *Related Company (e.g., supplier, partner, competitor, etc)*
    • *Location*
  • *Select relevant entities*
    • *PERSON, LOCATION, ORGANIZATION*
  • *Train neural networks that map (sentence, entities) -> relation*
    • *Bill Gates is the founder of Microsoft. -> (Bill Gates, founder, Microsoft)*
    • *Mondrian takes 6.6% stake in Sheng Siong (Mondrian, shareholder, Sheng Siong)
Knowledge Extraction --- DL Data Flow
Knowledge Extraction --- Distant Supervision

• **Steps**
  
  • Find text (sentences) discussing the entities of a relation
    - *Zuckerberg says he won't step down as Facebook CEO*
    - *Apple’s Tim Cook says tech regulation ‘inevitable’...*
      - *not talking about the CEO relation ➔ noisy sample*
  
  • Learn over the noisy text to correlate
    - *the descriptions, e.g., step down, CEO, Facebook*
    - *the relation, CEO*
  
  • Infer the relation of entities in new sentences
    - *Microsoft CEO Satya Nadella sells 30 percent of common stock*
      - *Returns: Satya Nadella, CEO, Microsoft*
Risk/opportunity detection

- **Sentiment analysis**
  - Too general

- **Risk**
  - Rating downgrade
  - Sanctions
  - More

- **Opportunity**
  - New market
  - New product
  - More

<table>
<thead>
<tr>
<th>text</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unilever to increase investment in Greece</td>
<td>New market/opportunity</td>
</tr>
<tr>
<td>Japan cryptocurrency exchange plans Europe, Asia sales hubs</td>
<td>New market/opportunity</td>
</tr>
<tr>
<td>Trump administration unveils sanctions aimed at starving North Korea of resources</td>
<td>Sanctions/risk</td>
</tr>
<tr>
<td>Why Did Maxim Group Downgrade Domino’ Pizza?</td>
<td>Rating downgrade/risk</td>
</tr>
</tbody>
</table>
Risk/opportunity detection

• **Document classification**
  
  • *Annotate the articles as training data*
    
    • *Expensive*
    
    • *Inconsistency between annotators*
  
  • *Data cleaning and pre-processing is important*
    
    • *Cluster articles for fast annotate*
    
    • *Rule-based annotation*
    
    • ...
  
  • *Deep learning models for classification*
    
    • *CNN, RNN*
    
    • *BERT (bidirectional encoder representations from transformers)*
Stock Prediction

• Traditional solutions for stock prediction are based on time-series models.
• With the recent success of deep neural networks in modeling sequential data, deep learning has become a promising choice for stock prediction.
## Related Work

<table>
<thead>
<tr>
<th>Topic</th>
<th>Type of Data</th>
<th>Number of stocks</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Dual-Stage Attention-Based Recurrent Neural Network for Time Series Prediction – Yao et al (2017)</td>
<td>Index Prediction</td>
<td>100</td>
</tr>
</tbody>
</table>
Optimized Trade Execution

The goal: is to sell (respectively, buy) V shares of a given stock within a fixed time period (or horizon) H, in a manner that maximizes the revenue received (respectively, minimizes the capital spent).

Trade-offs: such as between the speed of execution and the prices obtained for the shares.

Virtually every strategy's profitability will depend on how well it is implemented.
Methodology

Try all possible actions!
Experiment results

*Figure 3.* Expected cost under S&L and RL: adding private variables T and I decreases costs
Fraud Detection

Deep Learning Opportunity

- Current models can only catch ~70% of all fraud cases
- Traditional ML models view transactions atomically
- Often missed fraud transactions are part of a series
- Capturing correlation across many features
Summary

• AI has provided great improvement in prediction on a variety of applications, including high stack applications

• However, unlike games, machines cannot learn by playing out every possibility, and also, there are just too many external factors

• AutoML, NAS, Model Verification have reduced the barrier for adoption, but a lot more needs to be done

• Data is a new oil, and data quality determines its value but, sensitive data is problematic to handle

• GIT for Data for data provenance and governance, and supporting AI development without disclosing the data
Thanks!