SWE404/DMT413 BIG DATA ANALYTICS

Lecture I: Introduction

Lecturer: Dr. Yang Lu Email: luyang@xmu.edu.my Office: A I - 432 Office hour: TBD

Lecturer Information

Yang Lu (Jason), Ph.D.

- Experience:
 - 2020~Present: Assistant Professor, School of Electrical and Computer Engineering, Xiamen University Malaysia.
 - 2019~Present: Assistant Professor, Department of Computer Science, School of Informatics, Xiamen University.
 - 2018~2019: Research Intern, Tencent.
 - 2014~2019: Ph.D. student, Department of Computer Science, Hong Kong Baptist University.
 - 2012~2014: Master student, Department of Software Engineering, University of Macau.
 - 2008~2012: Bachelor student, Department of Software Engineering, University of Macau.
- Research interests:
 - Artificial intelligence, machine learning.
- Personal website:
 - https://jasonyanglu.github.io/







Consultation Hours

- Office location: AI-432
- Consultation Hours:
 - TBD
- For other time slot:
 - Send email to confirm my presence first.
 - Or take your luck to knock the door (I will be in my office for most of the time in this semester).







Class Rules

- You can do anything except:
 - Make noises (chatting, singing...)
 - Catch others' eyes (dancing, push-ups...)
 - Eat food with strong smell (instant noodles, durian...)
- Feel free to interrupt me if you have questions (no need to raise hand).
- According to the university policy, taking attendance is needed.
 - Important: you are required to have an 80% attendance to be able to seat for the final exam.







Course Assessment

- Temporary according to the situation:
 - Final exam: 50%
 - Assignment: 20%, individually
 - Project: 30%, 2-3 members per group, report and presentation are required.
- Important: cheating and plagiarism will get no marks.







4

Textbook

- No textbook used on this course.
- Links of necessary materials will be shown in the slides.







What Is Big Data?

The term "big data" refers to data that is so large, fast or complex that it's difficult or impossible to process using traditional methods.







Image source: https://mapreducer.files.wordpress.com/2015/04/3values.png



Volume of Big Data

- Organizations collect data from a variety of sources, including
 - business transactions
 - smart (IoT) devices
 - industrial equipment
 - videos
 - social media
 - • •
- In the past, storing it would have been a problem but cheaper storage on platforms like data lakes and Hadoop have eased the burden.









Image source: <u>https://healthcaresocialmedia.files.wordpress.com/2019/03/1_zphbpoloawhbumudpaz7qq.jpeg?w=600</u>

Velocity of Big Data

- With the growth in the Internet of Things, data streams in to businesses at an unprecedented speed and must be handled in a timely manner.
- Big data techniques are driving the need to deal with these torrents of data in near-real time.







Image source: https://www.dataversity.net/wp-content/uploads/2015/03/ig.png



Variety of Big Data

- Data comes in all types of formats
 - Structured: quantitative data, fits neatly within fixed fields and columns in relational databases and spreadsheets
 - Numeric data
 - Database
 - ...
 - Unstructured: qualitative data, cannot be processed and analyzed using conventional tools and methods.
 - Text documents
 - Videos
 - Financial transactions
 - ...

Structured data

Unstructured data







Image source: https://learn.g2.com/structured-vs-unstructured-data



Big Data Market Worldwide Segment Revenue Forecast 2011-2026

Big Data Market Forecast Worldwide from 2011 to 2026, by segment (in billion U.S. dollars)





Image source: https://www.forbes.com/sites/louiscolumbus/2018/05/23/10-charts-that-will-change-your-perspective-of-big-datas-growth/#4f7acfd72926

Big Data Landscape 2012

Vertical Apps PREDICTIVE POLICING Deloomreach. Log Data Apps splunk: loggly & sumologic Data As factual. CNIP MARKI Weder Are	Ad/Media Apps Collective (C) Collective (C) Collective (C) Recorded future Media Science TURN Data XC Collective (C) Recorded future Data XC Collective (C) Recorded future Collective (C) Recorded future Recorded future	Business Intelligence ORACLE Hyperion Business Objects Rimetrics Microsoft Business Intelligence Microsoft Business Intelligence MicroSoft Business Intelligence MicroSoft Business Intelligence MicroSoft Business Intelligence MicroSoft Business Intelligence MicroSoft Business Intelligence	Analytics and Visualization + a b l e a u Palantir Palantir Metamarkets Feradata Aster SSAS TIBCO SSAS TIBCO Palanter Contrifuge Palanter Contribute Jose Angel Datameer Datameer Diatfora alteryx Visually Ayata
Analytics Infrastructure Hortonworks Cloudera EMC ² © GREENPLUM. N NETEZZA DRTRSTRX: EASOL	Operational Infrastructure Couchbase 10gen FERADATA HADAPT FERADATA HADAPT FERACOTTA VoltDB MarkLogic INFORMATICA	Infrastructure As A Service Construction A Services Construction Con	Structured Databases ORACLE: Mysol SQL Server SQL Server DB2. SYBASE SYBASE
Copyright © 2012 Dave Feinleib	Techno Techno dave@vco		Cassandra blogs.forbes.com/davefeinleil



Big Data Landscape 2016 (Version 3.0)

Image source: https://mattturck.com/big-data-landscape-2016-v18-final/

DATA & AI LANDSCAPE 2019								
INFRASTRUCTURE	ANALYTICS & MACHINE INTELL		APPLICATIONS - ENTERP	RISE				
HADOOP NN PREMISE cloudera withowski MARR Pivotal IBM InioSphere' jethro Jethro	DATA ANALYST PLATFORMS Microsoft @ pentoho alteryx @ Dotter Microsoft @ pentoho alteryx @ Dotter Microsoft @ pentoho ATTIV/O Datameer inconfa. inter/ena Microsofta Sistu septenced @ serburk	Action of the second s	MARKETING - B28 RADIUS App Aniz Control Control Cont	CUSTOMER EXPERIENCE / SERVICE patrice" & HEDALLIA & Surviverung Gimine Generation Contraction Carrier (Contraction) Generation Carrier (Contraction) Contraction Carrier (Contraction) Con				
Nesol DATABASES CRACLE Human Alars CracLe Human Alars Construct For Tables Construct	BI PLATFORMS USUAL ZATION USUAL ZATION US	MACHINE LEARNING MACHINE LEARNING Martine Ma	REGISCHE RECOMPLIANCE FINANCE AND AND A RPA AUTOMATION &	SECURITY CTANUM & CYLLANCE CASCAR Statement (Character Cost Control Control Control Control Control Control NVECTRA LAND From Statement (Control Control Co				
DATA TRANSFORMATION -talend @pender atteryx @transca @transform MR Paxtat @transform WR Paxtat @transfor	COMPUTER VISION Morecon Accord Computer Vision Computer Vision Compute	SPEECH & NLP	EDUCATION REAL ESTATE ELIMINAN DI REAL Sectors Celorer Celore	TRY TRANCE- FINANCE-				
STORAGE WYS O WYS O WYS O WYS O WYS O WYS O WYS O WYS O Goode Cloud II Microsoft TRM STOR Microsoft Microsoft TRM STOR Microsoft Microsoft Microsoft TRM ST	SEARCH CONCLUSION CONCLUSICONCLUS CONCLUSION CONCLUSION CONCLUSION CONCLUSION	A Artoble RESCI Web MAPR cloudera E.	Interview Color Constraints Interview Color Constrain	AGRYCUITURE AGRYCUITURE Sontaar Aurora				
FRAMEWORKS Sport Sport Sport Standards	OPEN S STREAMING & MESSAGING Sport On III (Sport On IIII (Sport On III (Sport On III (Sport On III	SOURCE AI/MACHINE LEARNING / DEEP LEARNING Plensorfilow Arm Mark Constraints / DEEP LEARNING Plensorfilow Arm Mark Constraints / Deep Learning / Original Constraints / Deep Learning	Caffe Minimut Correct Matt DMSUM Feedback Correct Matt DMSUM Feedback Correct Matt Correct Matt	RING WISUALIZATION matpl:/stilb Genution Genution CollABOBRATION Matpl:/stilb M				
HEALTH WALFORC Protocolusion + fitbit GARMIN MANNAL & ECONOMIC DATA Bloomberg () THOMSON REUTERS DESCRET © Thingwork Intervention Intervention Mannal & ECONOMIC DATA Bloomberg () THOMSON REUTERS Mannal & ECONOMIC DATA Bloomberg () THOMSON REUTERS Bloomberg () THOMSON R	DATA SOURCES & APIs AR/SPACE/SEA AR/SPACE/	Control Intelligence Foursquare Consolution Sinsdeview Sensel 10 Consolution Sensel 10 Consolution Sensel 10 Same Graphic Sensel 10	OTHER OTAGOV D MAGENET Ladride 2000 CRUX Constant Jagonicio	DATA RESOURCES INDES & SCHOOLS State				
July 16, 2019 - FINAL 2019 VERSION	D Matt Turck (@mattturck), Lisa Xu (@lisaxu9:	2), & FirstMark (@firstmarkcap) mattturck	com/data2019	EARLY STAGE VENTURE CAPITAL				

Image source: https://www.zdnet.com/article/the-data-and-ai-market-landscape-2019-the-next-wave-of-hybrid-emerges/

What Is Analytics

- The importance of big data doesn't revolve around how much data you have, but what you do with it.
- Analytics is the scientific process of transforming data into insight for making better decisions, offering new opportunities for a competitive advantage.







Types of Analytics

Predictive Analytics

- Predicting the future based on historical patterns.
- What could happen?

Descriptive Analytics

- Mining historical data to provide business insights.
- What has happened?

Prescriptive Analytics

- Enabling smart decisions based on data.
- What should we do?







Job Market for Analytics











Image source: https://analyticsindiamag.com/study-analytics-and-data-science-jobs-in-india-2019-by-great-learning-aim/

Salary

Popular Employer Salaries for Data Scientist

Booz, Allen, and Hamilton	\$81k
Microsoft Corp	\$123k
Amazon.com Inc	\$119k
Facebook Inc	\$121k
International Business Machines (IBM) Corp.	\$101k
Ford Motor Company	\$95k
Apple Computer, Inc	\$125k
Google, Inc.	\$140k







Data source: https://www.payscale.com/research/US/Job=Data Scientist/Salary

Top 10 Applications of Big Data Across Industries

- Banking and Securities
 - Trade analytics used in high-frequency trading, pre-trade decisionsupport analytics, sentiment measurement.
- Communications, Media and Entertainment
 - Create content for different target audiences, recommend content on demand, measure content performance.
- Healthcare Providers
 - Evidence-based medicine test, faster identification, tracking the spread of chronic disease.
- Education
 - Track study activity, measure teacher's effectiveness, correct astray students.
- Manufacturing and Natural Resources
 - Geospatial data analysis, seismic interpretation.





- Government
 - Energy exploration, financial market analysis, fraud detection, health-related research, and environmental protection.
- Insurance
 - provide customer insights for transparent and simpler products, fraud detection.
- Retail and Wholesale Trade
 - Optimized staffing, reduced fraud, timely analysis of inventory.
- Transportation
 - Traffic control, route planning, intelligent transport systems, congestion management.
- Energy and Utilities
 - Faster meter readers, better asset and workforce management.



Case I: Personalized Recommendation

- Recommend based on your profile:
 - Demographic characteristics, historical order, recent clicks....





2-day shipping

Customers also viewed these products











WWE 2K19, 2K, PlayStation 4

\$59.99

2-day shipping

UFC 3, Electronic Arts, PlayStation Sony Playstation 4 DualShock 4 Controller, Black, 711719504290 **** 1100 \$46.96 2-day shipping

2-day shipping

19

Image source: https://www.yuspify.com/blog/examples-of-personalized-recommendations-in-ecommerce/

2-day shipping

Case 2: Social Network Analysis by Call Detail Records (CDR)

- Construct a weighed graph to model social network, representing how close two subscribers are to each other.
 - Predict the relationship between subscribers.
- Usage: market campaign, churn prediction, prevent phone fraud,...





Visualization for a sample of the Telecom social network

Source	Destination	Weight
963-9*****14	963-9******22	0.0425
963-9*****31	963-9*****62	0.0496
963-9*****94	963-9*****11	0.0272
963-9*****34	963-9*****78	0.0335

Sample of Telecom social network data



Reference: Al-Molhem, N.R., Rahal, Y. & Dakkak, M. Social network analysis in Telecom data. J Big Data 6, 99 (2019). https://doi.org/10.1186/s40537-019-0264-6

Case 3: Tracing Information Diffusion with Data Visualization

- Social media, like Twitter, has been increasingly used for exchanging information, opinions and emotions about events that are happening across the world.
- Three major characteristics of diffusion processes in social media:
 - temporal trend
 - social-spatial extent
 - community response of a topic of interest





Reference: N. Cao, Y. Lin, X. Sun, D. Lazer, S. Liu and H. Qu, "Whisper: Tracing the Spatiotemporal Process of Information Diffusion in Real Time," in IEEE Transactions on Visualization and Computer Graphics, vol. 18, no. 12, pp. 2649-2658, Dec. 2012.

Case 4: Financial Analytics

- Expected outcome:
 - Rejected transactions
 - Real time alerts
 - Real time dashboard
 - Automated learning and improvement
 - Audit trails and analytics









Image source: https://www.slideshare.net/HadoopSummit/big-data-application-architectures-fraud-detection

Case 5: Detecting Cyber Attack

- Using big data technology to cope with such substantial amounts of data allows us to quickly perform the following analysis to assess alerts and determine if they are suspicious or benign.
 - Risk scoring
 - Graph analysis
 - Time series analysis
 - Determine causal relationships between entities







Case 6: Smart Traffic Management

- Use sensors and traffic signals to monitor, control and respond to traffic conditions.
- The aim of smart traffic management systems:
 - Reduce day-to-day congestion by improving traffic flow
 - Prioritize traffic according to real-time changes in traffic conditions
 - Reduce pollution by limiting traffic jams
 - Prioritize buses entering intersections
 - Improve traffic incident response time









Image source: https://ars.els-cdn.com/content/image/1-s2.0-S0386111218300293-gr1.jpg

Case 7:Video Surveillance



Pedestrian attributes recognition











Abandoned object alert



Pedestrian tracking





25

Reference

Fan, Quanfu & Gabbur, Prasad & Pankanti, S.. (2013). Relative Attributes for Large-Scale Abandoned Object Detection. 2736-2743. Xia, Yu & Li, Ju & Zhou, L.-f. (2016). A two-stage visual tracking algorithm using dual-template. International Journal of Advanced Robotic Systems. Wang, Xiao & Zheng, Shaofei & Yang, Rui & Bin, Luo & Tang, Jin. (2019). Pedestrian Attribute Recognition: A Survey.

Big Data Tools









- No. I market leader in analytics.
 - The largest independent vendor in the business intelligence market.
 - The industry standard for Clinical Data Analysis.
- Integrated platform for end to end solutions.
 - SAS provides an integrated set of software products and services and integrated technologies for information management, advanced analytics and reporting.
- Business solutions across domains and industries.
 - Unmatched domain specific industry focused analytics solutions.









Used in 60,000+ companies in over 135 countries R

- R is a language and environment for statistical computing and graphics.
- R provides a wide variety of statistical and graphical techniques, and is highly extensible.











Image source: https://data-flair.training/blogs/using-r-for-data-science/



Hadoop

- Hadoop is the most popular big data ecosystem.
- Hadoop is highly scalable, that is designed to accommodate computation ranging from a single server to a cluster of thousands of machines.







计算机科学系

outer Science Department of Xiamen University



Image source: https://subscription.packtpub.com/book/application_development/9781788995092/1/ch011v11sec14/overview-of-the-hadoop-ecosystem

Python

- Python is an interpreted, high-level, general-purpose programming language.
- One of the most popular programming language in recent years.
- Ten areas that uses Python most frequently:
 - Web Development
 - Game Development
 - Machine Learning and Artificial Intelligence
 - Data Science and Data Visualization
 - Desktop GUI



- Web Scraping Applications
- Business Applications
- Audio and Video Applications
- CAD Applications
- Embedded Applications







Tableau

- Tableau is a data visualization tool that is widely used for business intelligence.
- Create interactive graphs and charts in the form of dashboards and worksheets to gain business insights.







Computer Science Department of Xiamen University





Image source: https://www.tableau.com/learn/articles/data-visualization

Skills in demand in 2019



ML algorithm

2%

2%

1%

1%

0%

Segmentation Case and Case and Case and Decision reserve and



Data visualization





Distributed framework



Image source: https://analyticsindiamag.com/study-analytics-and-data-science-jobs-in-india-2019-by-great-learning-aim/

Thank you!

- Any question?
- Don't hesitate to send email to me for asking questions and discussion. 😳

Acknowledgement: Thankfully acknowledge slide contents shared by Prof. Yongxuan Lai





