CS5344:Big Data Analytics
Lesson 1: Introduction

https://canvas.nus.edu.sg/courses/38824

Anthony Tung

Department of Computer Science

anthony@comp.nus.edu.sg

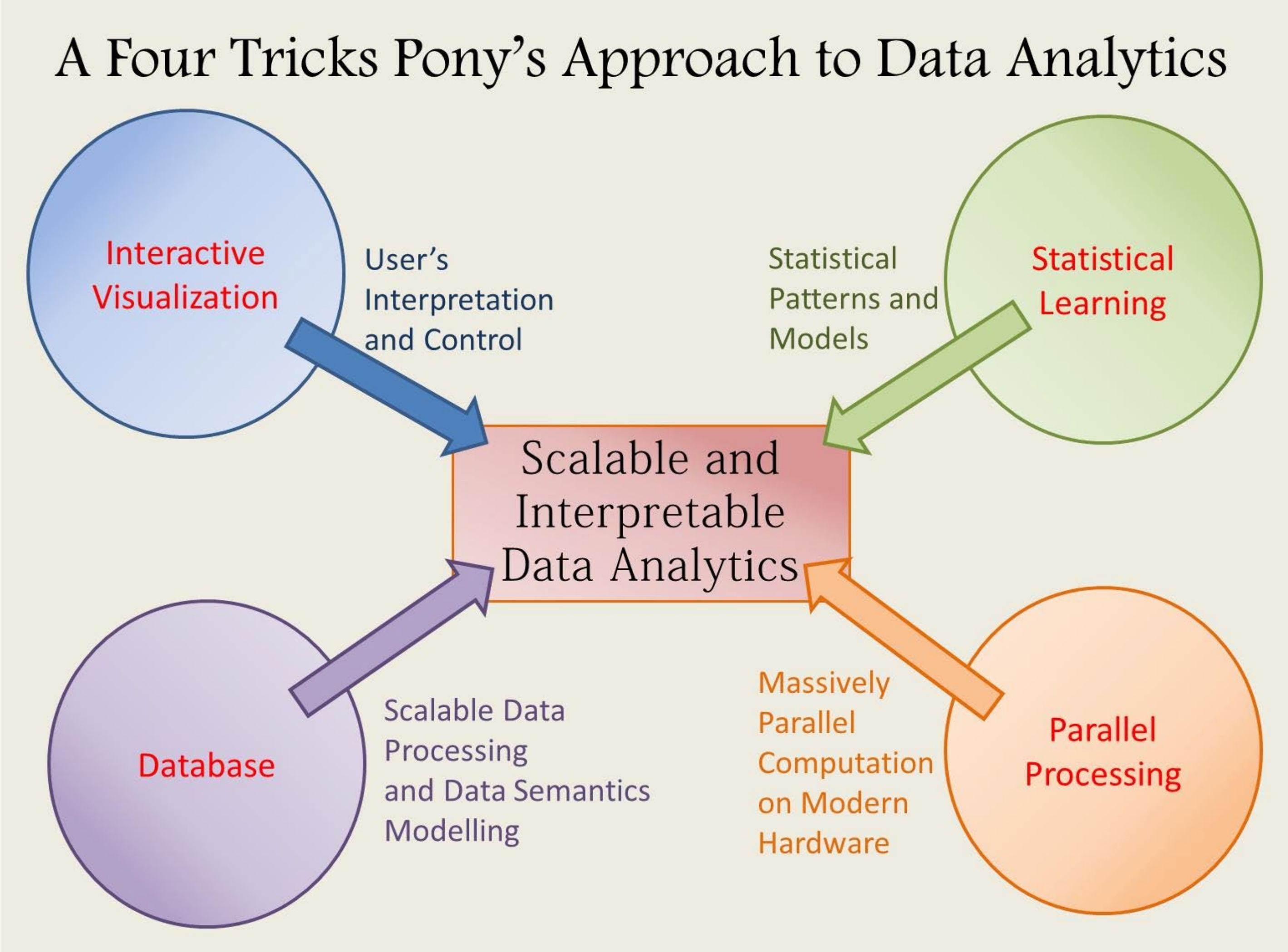
Right Infringements on NUS Course Materials

All course participants (including permitted guest students) who have access to the course materials on LumiNUS or any approved platforms by NUS for delivery of NUS modules are not allowed to re-distribute the contents in any forms to third parties without the explicit consent from the module instructors or authorized NUS officials

Class Information

- **Lecturer: Anthony Tung**
  - Email: anthony@comp.nus.edu.sg
- **Tutors:**
  - Wu Shengqiong swu@u.nus.edu
  - Xu Danni dannixu@u.nus.edu
- **Lectures on Monday 1830 – 2030 Physical F2F only**
- **Office hours/Project Consultation: Monday 2035 - 2200hrs**
- **Course website: https://canvas.nus.edu.sg/courses/61715**
- **Reference text(Do NOT need to buy)**
  - Mining of Massive Datasets by J. Leskovec, A. Rajaraman and J.D. Ullman (available online: http:///www.mmds.org)
  - Introduction to Data Mining (Second Edition) by Anuj Karpatne , Michael Steinbach, Pang-Ning Tan, Vipin Kumar

About Myself



A Four Tricks Pony's Approach to Data Analytics

Interactive Visualization

User's Interpretation and Control

Statistical Learning

Statistical Patterns and Models

Scalable and Interpretable Data Analytics

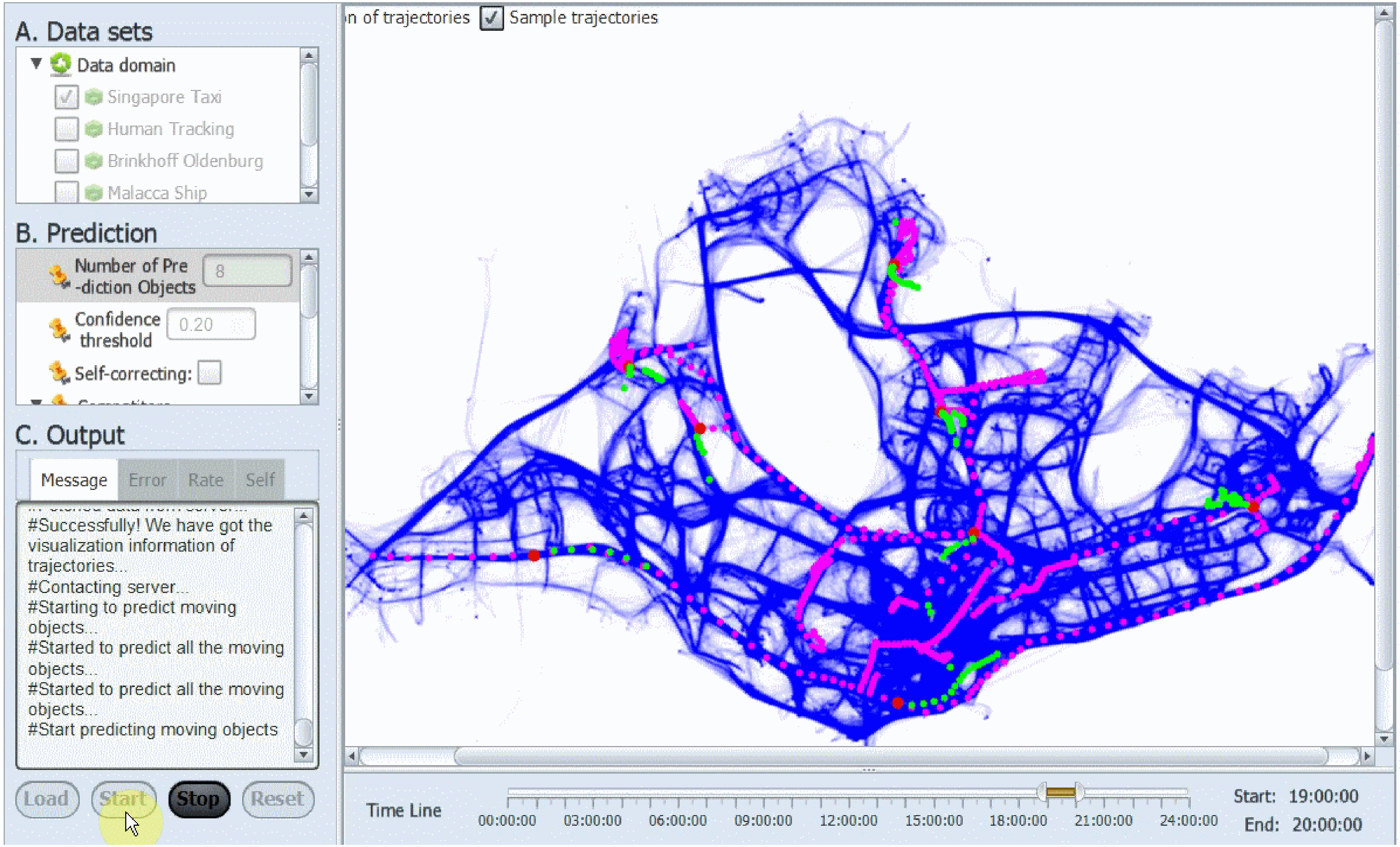
Database

Scalable Data Processing and Data Semantics Modelling

Parallel Processing

Massively Parallel Computation on Modern Hardware

LAMP and GENIE Example:“Semi-Lazy” Path Prediction



A. Data sets

▼ Data domain

- ☑ Singapore Taxi
- □ Human Tracking
- □ Brinkhoff Oldenburg
- □ Malacca Ship

B. Prediction

- Number of Pre -diction Objects: 8
- Confidence threshold: 0.20
- Self-correcting: □

C. Output

#Successfully! We have got the visualization information of trajectories...
#Contacting server...
#Starting to predict moving objects...
#Started to predict all the moving objects...
#Started to predict all the moving objects...
#Start predicting moving objects

☑ Sample trajectories

Start: 19:00:00
End: 20:00:00

What is Big Data?

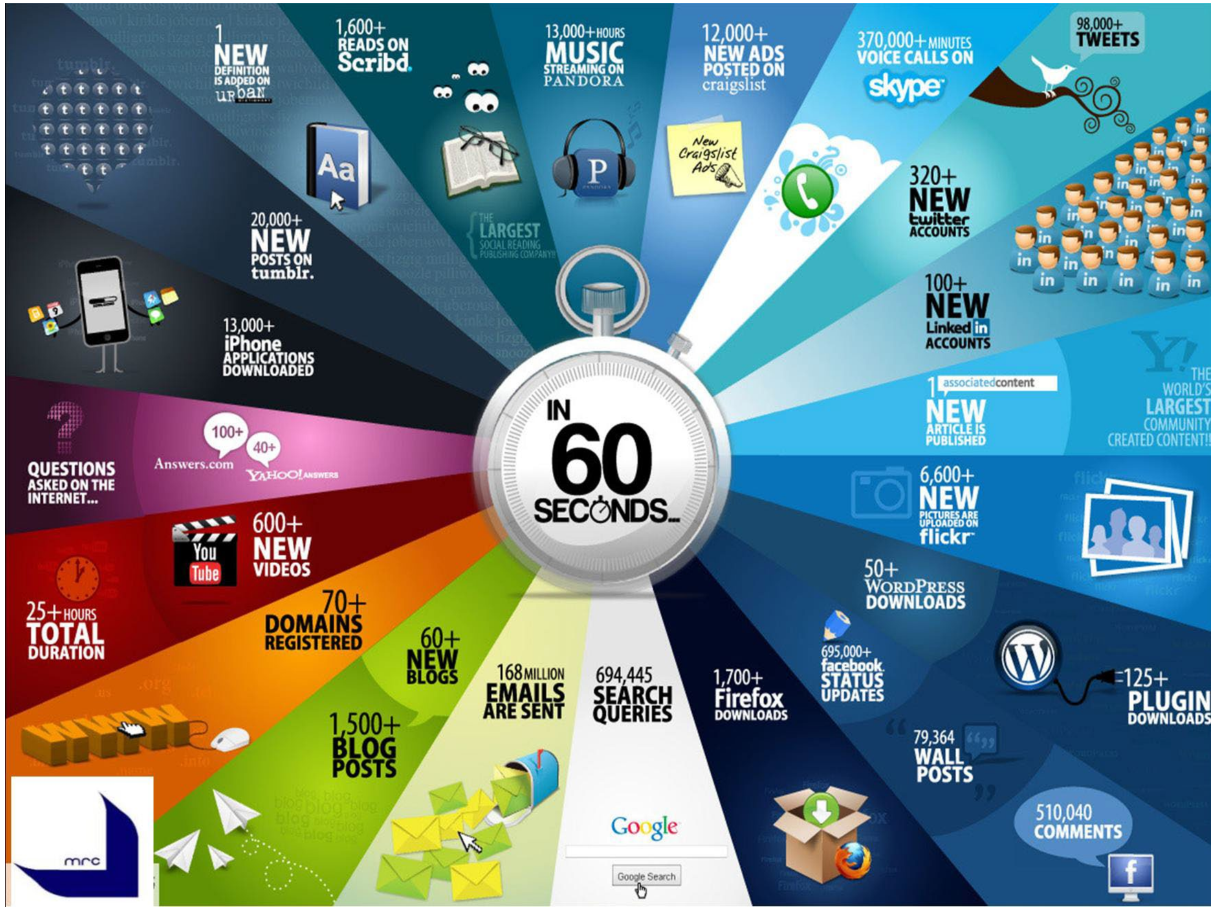
- **Gartner's Definition**

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

- **Information assets characterized by 3Vs**
  - **High-volume (Terabytes → Zettabytes)**
  - **High-velocity (Batch → Streaming data)**
  - **High-variety (Structured → Semistructured & unstructured)**

Data becomes BIG when the volume, velocity or variety EXCEEDS the abilities of our IT systems to ingest, store, analyze and process it to derive actionable intelligence in a TIMELY manner.

Volume: How Much Data?



IN 60 SECONDS...
1 NEW DEFINITION IS ADDED ON urban
1,600+ READS ON Scribd.
THE LARGEST SOCIAL READING PUBLISHING COMPANY!!
13,000+ HOURS MUSIC STREAMING ON PANDORA
12,000+ NEW ADS POSTED ON craigslist
New Craigslist Ads
370,000+ MINUTES VOICE CALLS ON skype
98,000+ TWEETS
320+ NEW twitter ACCOUNTS
100+ NEW LinkedIn ACCOUNTS
1 associatedcontent NEW ARTICLE IS PUBLISHED
Y! THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!!
6,600+ NEW PICTURES ARE UPLOADED ON flickr
50+ WORDPRESS DOWNLOADS
125+ PLUGIN DOWNLOADS
695,000+ facebook STATUS UPDATES
79,364 WALL POSTS
510,040 COMMENTS
1,700+ Firefox DOWNLOADS
694,445 SEARCH QUERIES
Google
Google Search
168 MILLION EMAILS ARE SENT
1,500+ BLOG POSTS
60+ NEW BLOGS
70+ DOMAINS REGISTERED
25+ HOURS TOTAL DURATION
600+ NEW VIDEOS
You Tube
QUESTIONS ASKED ON THE INTERNET...
100+ Answers.com
40+ YAHOO! ANSWERS
13,000+ iPhone APPLICATIONS DOWNLOADED
20,000+ NEW POSTS ON tumblr.
mrc

- Amount of data we create every day, every minute
- 90% of the data in the world today has been created in one year alone
- Data comes from everywhere e.g. sensors gather climate data, posts to social media, digital pictures and videos, purchase transaction records, cell phone GPS signals etc.

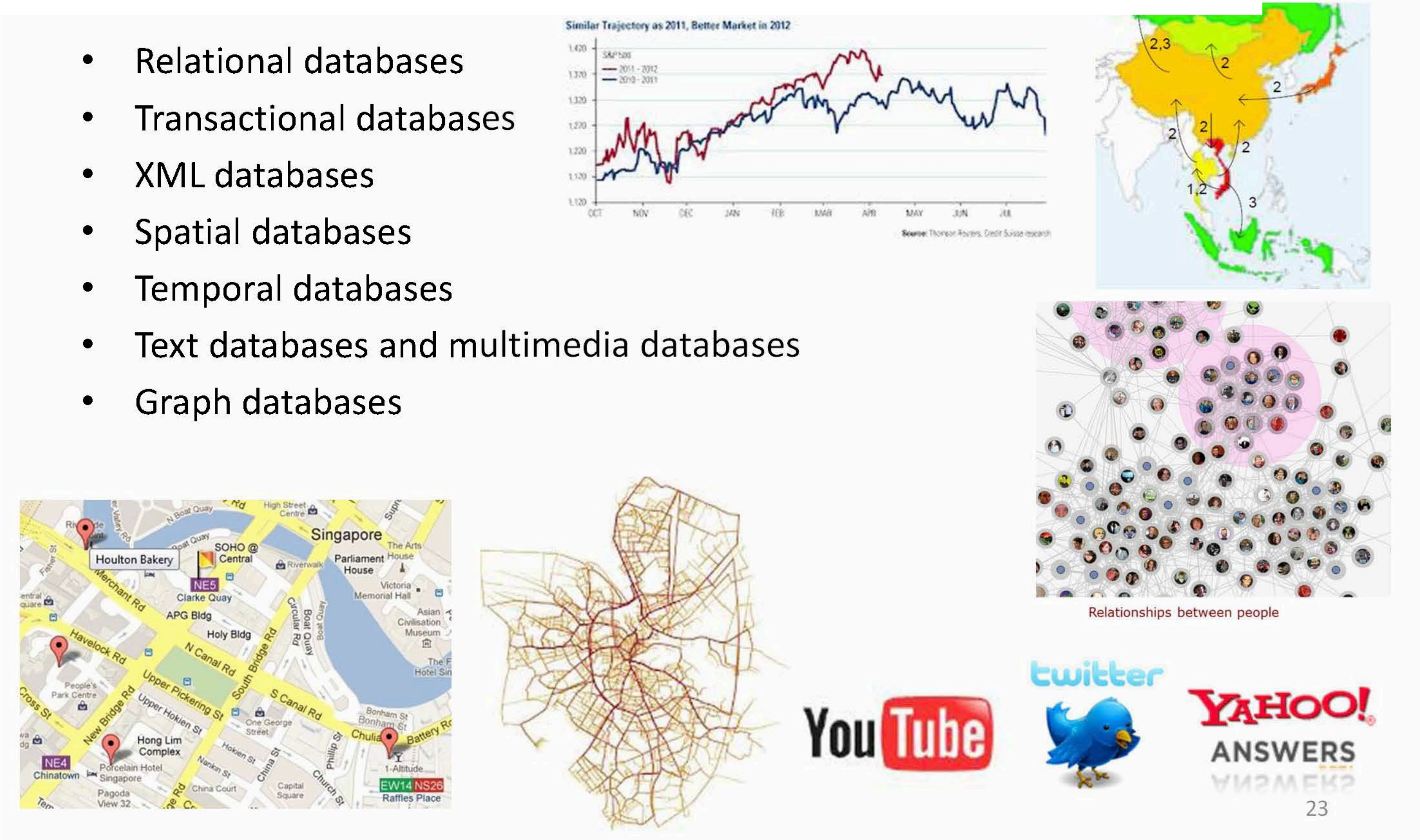
Volume: How Much Data?

- **Facebook**
  - >250 billion photos (>600 petabyte)
  - 6 billion messages per day (5-10 terabyte)
  - >1500 million users (2 trillion connections?)
- **Sloan Digital Sky Survey**
  - 35% of the sky mapped
  - >1 billion objects classified
  - 100 terabyte of data available

Velocity: At What Speed?

- Pace at which data flows in from sources
- Bursts of activities
- Real-time analysis
  - Late decisions → Missed opportunities

Variety: What Kind of Data?



- Relational databases
- Transactional databases
- XML databases
- Spatial databases
- Temporal databases
- Text databases and multimedia databases
- Graph databases

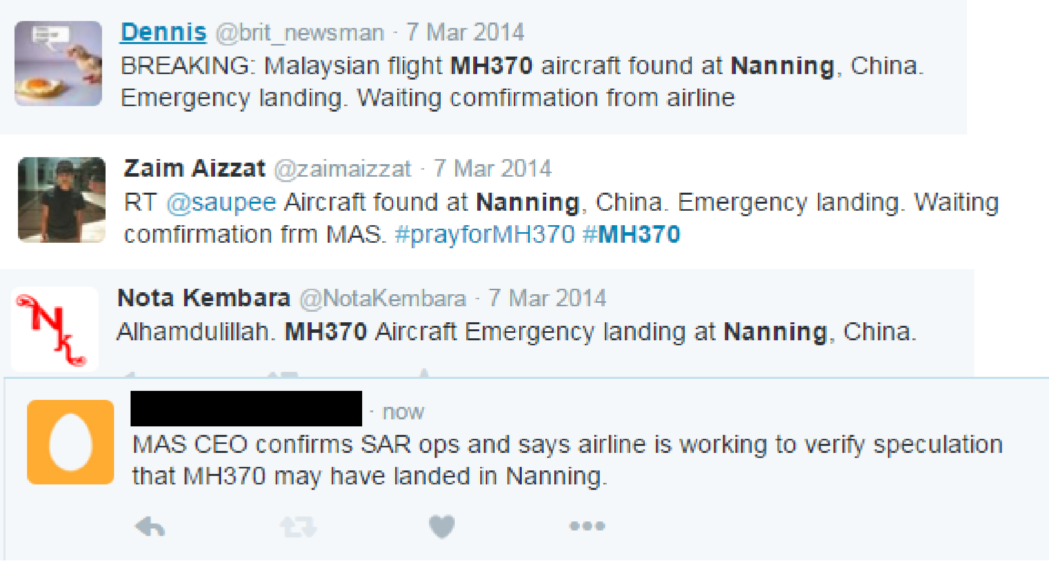
Relationships between people

23

Do not fit into a data warehouse, into neat tables of columns and rows.
Better place in Hadoop Distributed File System (HDFS) or in non-relational NoSQL databases.

Fourth V - Veracity

- How accurate or trustworthy is the data?
- Bias, inconsistencies, half truth
- Reliability of data source



Dennis @brit_newsman · 7 Mar 2014
BREAKING: Malaysian flight **MH370** aircraft found at **Nanning**, China. Emergency landing. Waiting comfirmation from airline

Zaim Aizzat @zaimaizzat · 7 Mar 2014
RT @saupee Aircraft found at **Nanning**, China. Emergency landing. Waiting comfirmation frm MAS. #prayforMH370 #**MH370**

Nota Kembara @NotaKembara · 7 Mar 2014
Alhamdulillah. **MH370** Aircraft Emergency landing at **Nanning**, China.

[redacted] · now
MAS CEO confirms SAR ops and says airline is working to verify speculation that MH370 may have landed in Nanning.

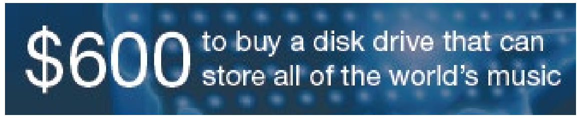
Why Big Data?

- Can collect cheaply, due to automation

$5 million vs $500

Price of fastest supercomputer in 1975 and iPhone with comparable performance

- Can store cheaply, due to falling media prices
- Can create Value



$600 to buy a disk drive that can store all of the world's music

  - Turn 12 terabytes of tweets created each day into improved product sentiment analysis
  - Convert 350 billion meter readings to better predict power consumption
  - Find communication patterns of successful projects in emails
  - Analyze elevator logs to predict vacated real estate
  - Scrutinize 5 million trade events created each day to identify potential fraud (time-sensitive, sometimes 2 minutes is too late)
  - Monitor 100's of live video feeds from surveillance cameras to target points of interest (new insights when you link and analyse different data types together)

Why Big Data?

Data contains Value and Knowledge

POPULAR SCIENCE
THE FUTURE NOW
THE CONTROL CENTERS
Using Data to Feed the World, Solve Cold Cases, Battle Malware, Predict Our Fate
OFFICER ALGORITHM
Can a Crime Be Prevented Before It Begins?
NEW WAYS OF SEEING
A Gallery of Extraordinary Infographics
PLUS
Juan Enriquez Reprograms Life
James Gleick Unsplits the Bit
AND
Lawrence Weschler Questions the Cloud
SPECIAL ISSUE
DATA IS POWER
HOW INFORMATION IS DRIVING THE FUTURE

Big Data Analytics

- **From raw data to actionable information**
  - Complex process of examining large and varied datasets to uncover information (hidden patterns, unknown correlations, market trends, customer preferences) that can help organizations make informed business decisions
- **Data needs to be**
  - **Stored**
  - **Managed**
  - **and ANALYZED**

Discover - Do we really know what we have?

Explore - How do different data relate to each other?

Iterative - What are the actual relationships?

Big Data Analytics

≈ Data Mining ≈ Data Science

- **Discover patterns and models that are**
  - **Valid:** hold on new data with some certainty
  - **Useful:** should be possible to act on the item
  - **Unexpected:** non-obvious to the system
  - **Understandable:** humans should be able to interpret the pattern

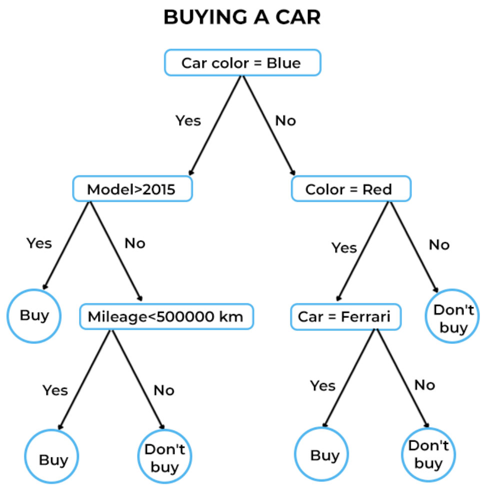
Data Analytics Tasks

Descriptive methods

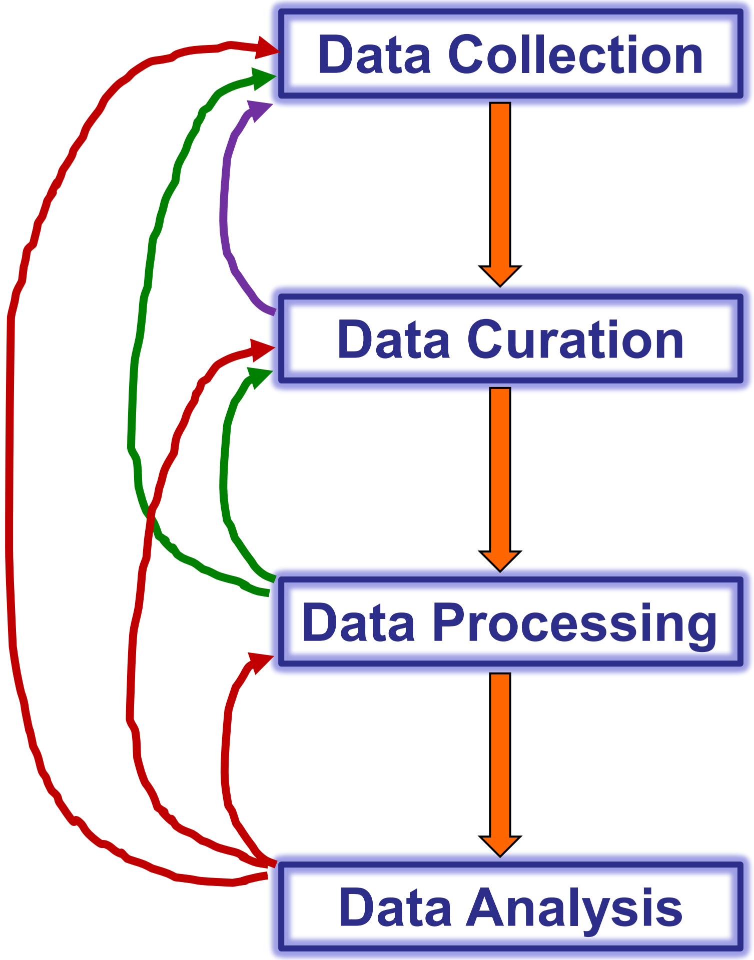
- Find human-interpretable patterns that describe the data
- Example: **Clustering**

Predictive methods

- Use some variables to predict unknown or future values of other variables
- Example: **Classification**



Data Analytics Pipeline



Acquire data from different sources

Clean, format, integrate with other datasets, store in database

Run queries (aggregate), plot graphs

Examine trends and anomalies, understand results

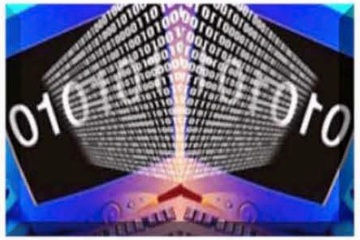
Big Data Applications

Smarter Healthcare

Multi-channel

Finance

Log Analysis



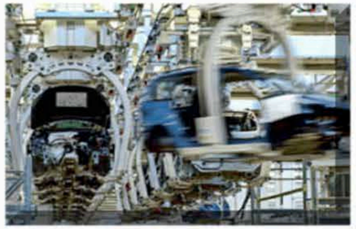
Homeland Security

Traffic Control

Telecom

Search Quality

Manufacturing

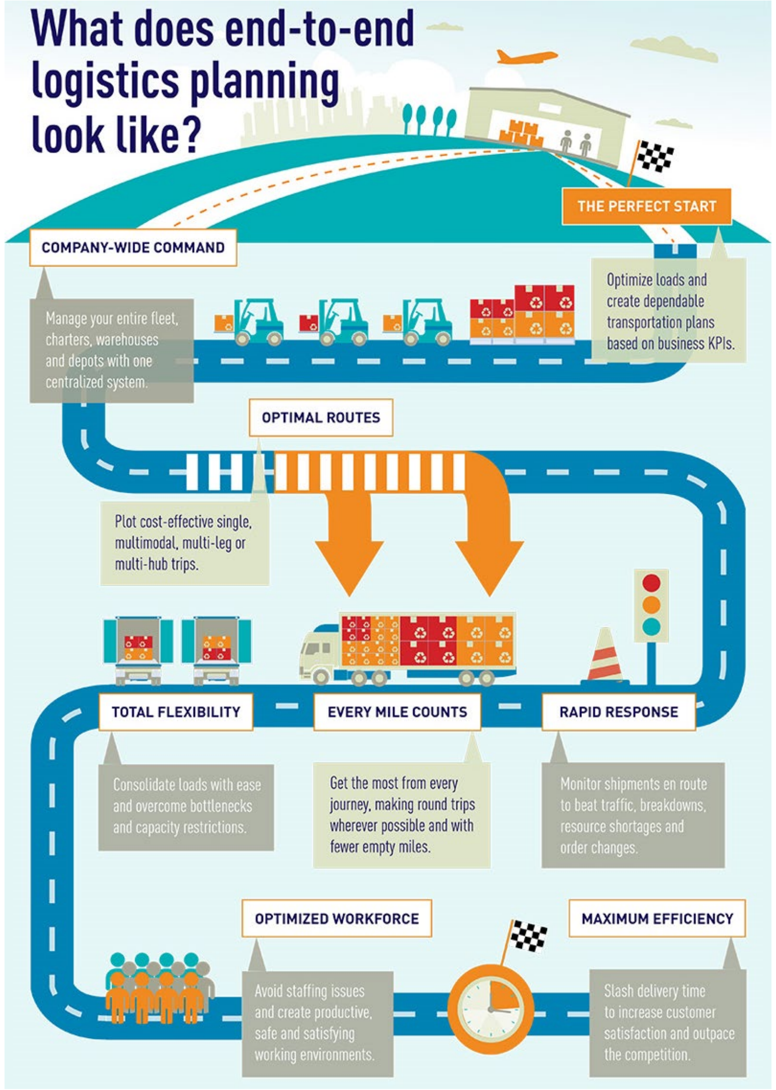


Trading Analytics

Fraud and Risk

Retail: Churn, NBO

Big Data Applications(I): Logistic



What does end-to-end logistics planning look like?

THE PERFECT START

Optimize loads and create dependable transportation plans based on business KPIs.

COMPANY-WIDE COMMAND

Manage your entire fleet, charters, warehouses and depots with one centralized system.

OPTIMAL ROUTES

Plot cost-effective single, multimodal, multi-leg or multi-hub trips.

TOTAL FLEXIBILITY

Consolidate loads with ease and overcome bottlenecks and capacity restrictions.

EVERY MILE COUNTS

Get the most from every journey, making round trips wherever possible and with fewer empty miles.

RAPID RESPONSE

Monitor shipments en route to beat traffic, breakdowns, resource shortages and order changes.

OPTIMIZED WORKFORCE

Avoid staffing issues and create productive, safe and satisfying working environments.

MAXIMUM EFFICIENCY

Slash delivery time to increase customer satisfaction and outpace the competition.

Supplier → Manufacture
→Distributor→Customer

Transport capacity

The delivery time is affected by traffic and weather conditions.

Storage capacity and price

Accuracy of the plan

Big Data Applications(II): Transportation

Early Warning of Human Crowds Based on Query Data from Baidu Map: Analysis Based on Shanghai Stampede

Jingbo Zhou, Hongbin Pei and Haishan Wu*

Baidu Research – Big Data Lab, Beijing, China

[Media Report: MIT Technology Review, Wall Street Journal, South China Morning Post]

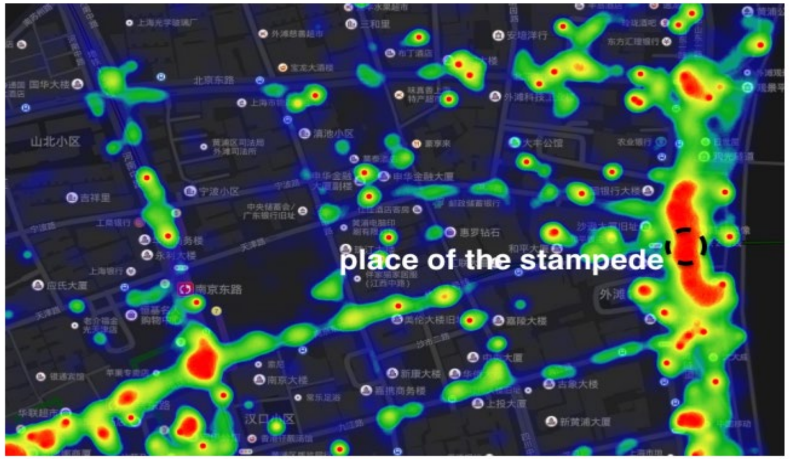


Figure 2: Human population density between 23:00-24:00 on Dec. 31th 2014.

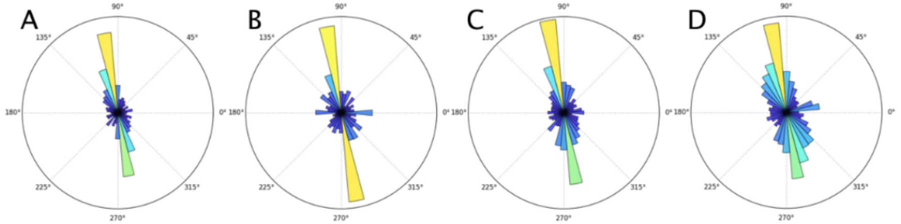


Figure 5: Human flow direction distribution in Chenyi Square (the specific disaster area of 2014 Shanghai Stampede) from 22:00 to 24:00 in: A – a common weekend (Aug. 23th 2014); B – the eve of the Mid-Autumn Festival (Sept. 7th 2014); C – the China's National Day (Oct. 1st 2014) and D – New Year's Eve of 2014

Integration of transportation data

- Multiple sources: car, taxi, bus, pedestrian, sensor
- Multiple organizations: telecom corporation, taxi company, bus company, government
- Data sharing and integrating

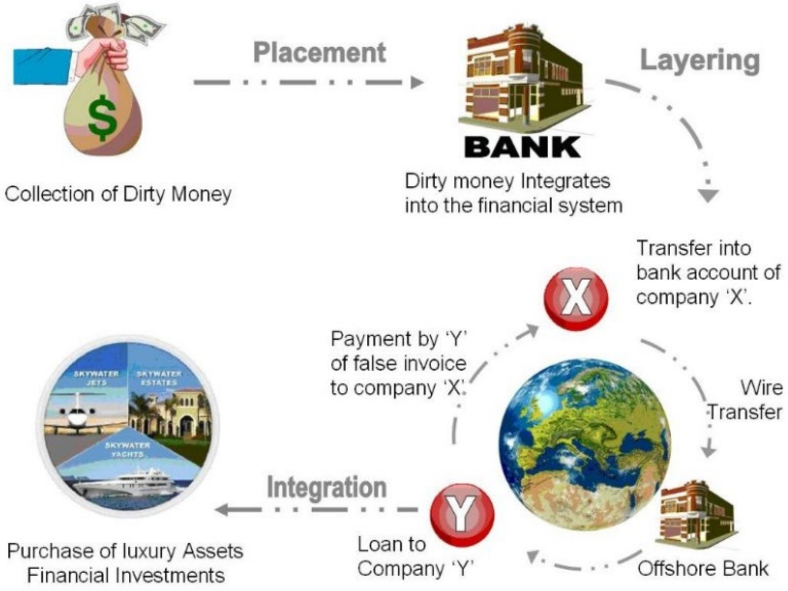
Transportation planning

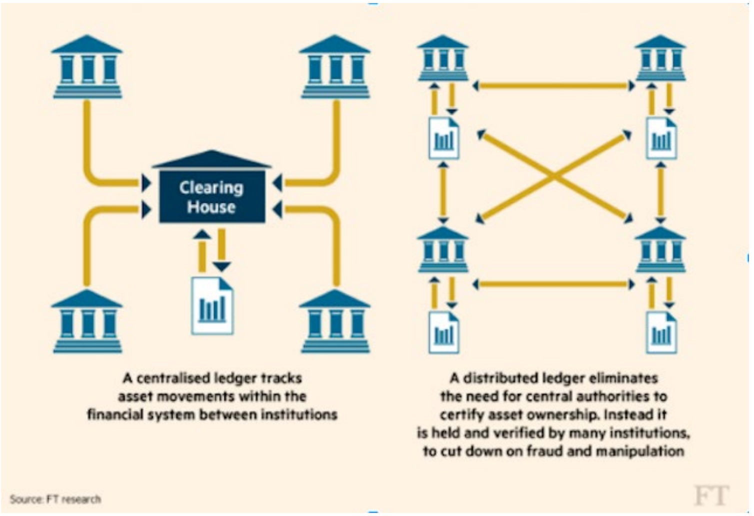
- Construction of new roads
- Location of transport junction
- Answer “what-if” questions

Transportation management

- Prevent traffic jam
- Optimize traffic lights
- Direct human crowds
- Identify bottlenecks

Big Data Applications(III): Finance


Placement
Layering
BANK
Collection of Dirty Money
Dirty money Integrates into the financial system
Transfer into bank account of company 'X'.
X
Payment by 'Y' of false invoice to company 'X'.
Wire Transfer
Integration
Y
Purchase of luxury Assets Financial Investments
Loan to Company 'Y'
Offshore Bank



A centralised ledger tracks asset movements within the financial system between institutions

A distributed ledger eliminates the need for central authorities to certify asset ownership. Instead it is held and verified by many institutions, to cut down on fraud and manipulation

Source: FT research

- Finance Prediction/Policy

  Cash flow

Abnormity Detection

Fraud

Money laundering

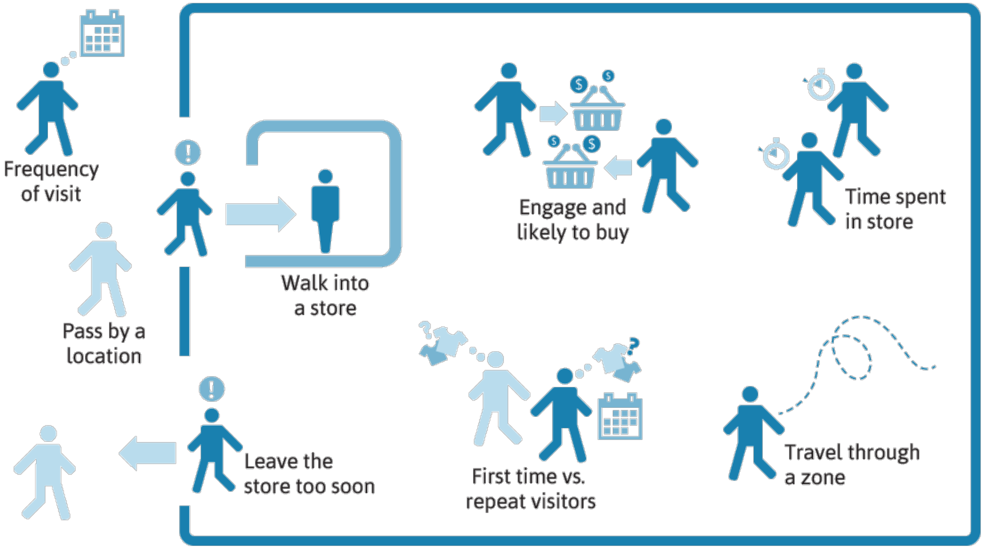
Tax evasion

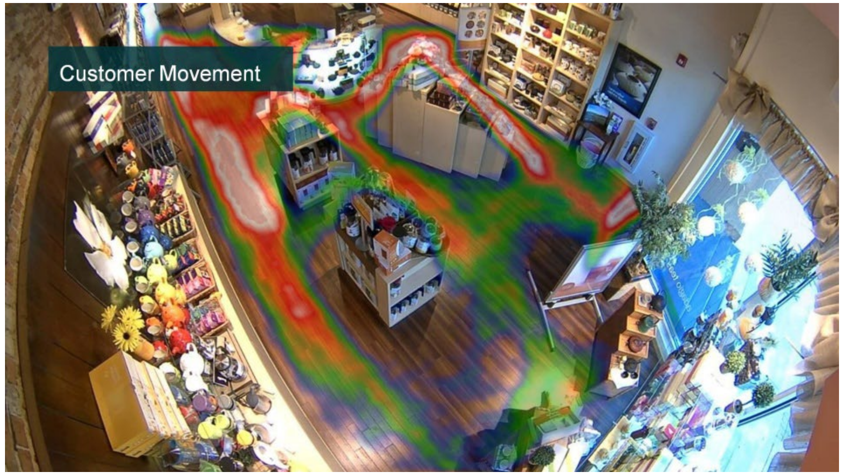
Fintech (financial technology)

Blockchain

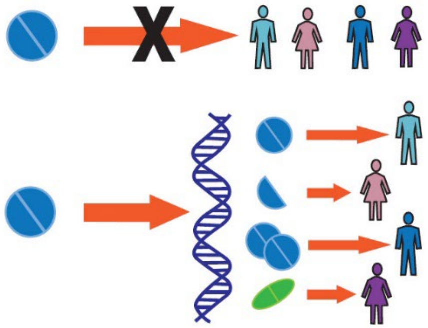
P2P loan

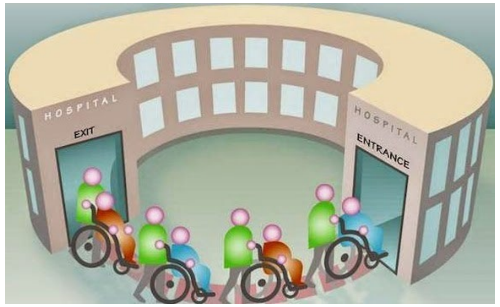
Big Data Applications(IV): Retail Analytics

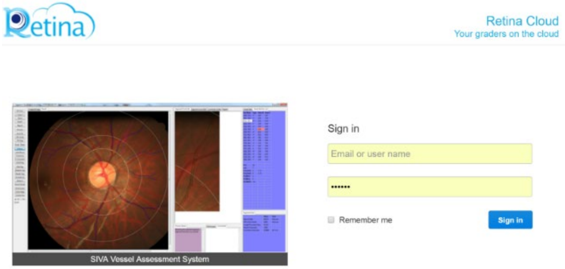




Big Data Applications(V): Medical







Retina

Retina Cloud
Your graders on the cloud

Sign in

Email or user name

••••••

□ Remember me

Sign in

SIVA Vessel Assessment System

Medicine control

- Drug allergy
- Drug and Poison Analysis

Personal Medicine

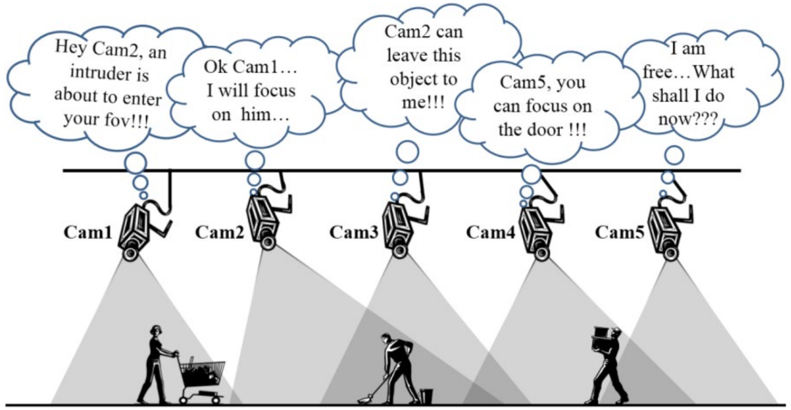
Hospital/Clinic management

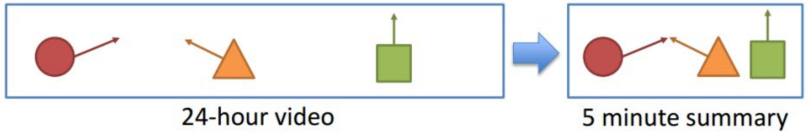
- Medical record
- Probability of re-hospitalization

Doctor on the Cloud: Retinal-scan analysis

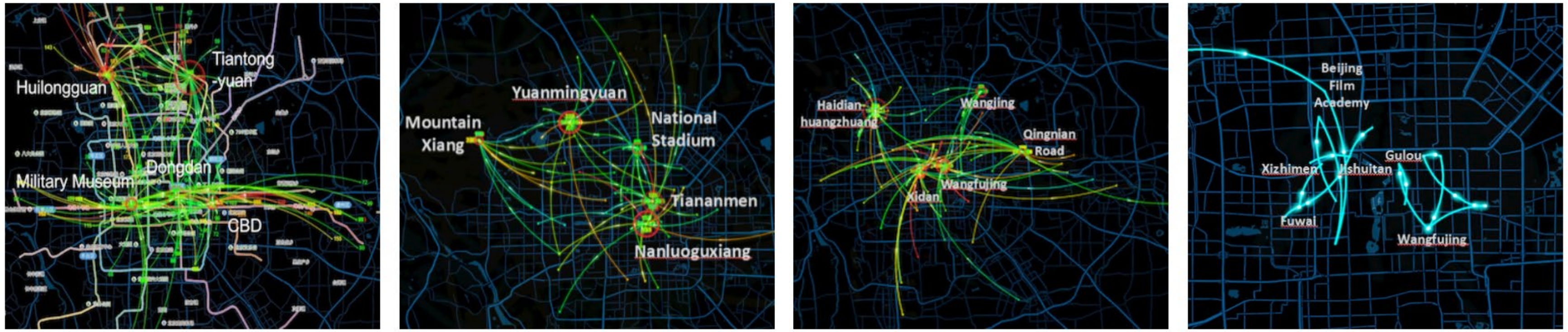
- https://retinacloud.d1.comp.nus.edu.sg/users/sign_in

Big Data Applications(VI): Security


Hey Cam2, an intruder is about to enter your fov!!!
Ok Cam1… I will focus on him…
Cam2 can leave this object to me!!!
Cam5, you can focus on the door !!!
I am free…What shall I do now???
Cam1
Cam2
Cam3
Cam4
Cam5

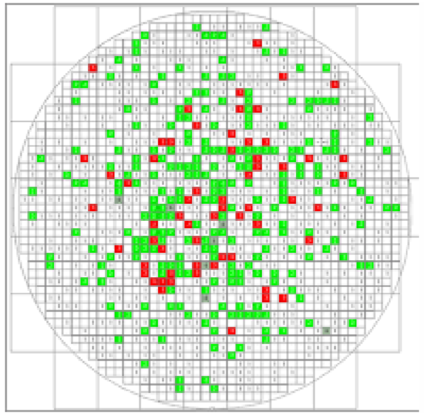


- **Monitoring**
  - CCTV
  - IC cards
- **Facial recognition to detect strangers**
- **Exit passageway monitoring**
- **Crime analysis**



(a) all passengers (b) visitors (c) shoppers (d) thieves

Big Data Applications(VII): Manufacturing



High returns products

- Wafer
- Petroleum

Manufacturing data

- Product imaging
- Machine sensors
- Machine repaired logs

Usage/Applications

- Predictive Maintenance
  - Which machine affected the quality?
  - Which part of the machine needs to be repaired?
  - How to fully utilize the machines?
- Product quality control

Course Focus

- **Handle data that cannot fit in main memory**
  - Scalability of algorithms
  - Cluster computing architecture
- **Real world problems**
  - **Market basket analysis**
    - Finding **frequent itemsets**
  - **Customer segmentation**
    - **Clustering** large high dimensional data
  - **Recommender engines**
    - **Similarity Search**

Course Focus

- Tools and Techniques
  - Hadoop ecosystem
    - Open source framework for **distributed processing** of large datasets
    - **Hadoop Distributed File System (HDFS)** for reliability and availability
    - **MapReduce, a Data-parallel programming model** to operate on large amounts of data
  - Apache Spark
    - Unified engine for distributed data processing
    - Fast in-memory processing and iterative processing with RDDs (Resilient Distributed Datasets)
  - Search engine technology
    - **Google’s PageRank**, link-spam detection, hubs and authorities

Assessment – 100% CA

- **Team-based Project (100%) --------- Max 2 members per team**
  - Project Proposal (20%)
    - *Proposal Presentation (10%)*
    - *Proposal Writeup(10%) ----------- 2 pages*
  - Project Updates (10%)
  - Final Project Presentation (20%)
  - Final Project Report(30%) ----------- 8 pages
  - Active Participation(20%)

You are reminded ***Plagiarism*** *is a very* ***SERIOUS*** *offence, and disciplinary action (including possibility of expulsion from the university) will be taken against any individual or team found plagiarizing.*

Timetable(Approximate)

Week No.	Date	Topic	Comments
1	12th Aug	Introduction/Data	
2	19th Aug	Hadoop /MapReduce	
3	26th Aug	*Similarity Search	
4	2nd Sep	*Frequent Items/Association Rules	Project Grouping Finalized 2nd Sep 23:59
5	9th Sep	*Clustering & Anomaly Detection(I)	
6	16th Sep	*Clustering & Anomaly Detection(II)	
	BREAK		
7	30th Sep	Project Proposal(F2F Presentation)	
8	7th Oct	*Classification/Regression I	
9	14th Oct	*Classification/Regression II	
10	21th Oct	Project Update (F2F Presentation)	
11	28th Oct	*Graph Mining I	
12	4th Nov	*Graph Mining II	
13	11th Nov	Project Presentation (F2F Presentation)	Project Report, 17th Nov. 23:59

*** Office Hour/Project Consultation Available from 20:35 to 22:00hrs**

Desiderata of a Good Project

- **Innovation**
  - New Applications
  - New Algorithms
  - New Ways of looking at old problems
- **Complexity**
  - Application Complexity
  - Algorithm Complexity
  - Data Complexity
- **Technical Depth**
  - None trivial implementation
  - Thorough experiments and analysis