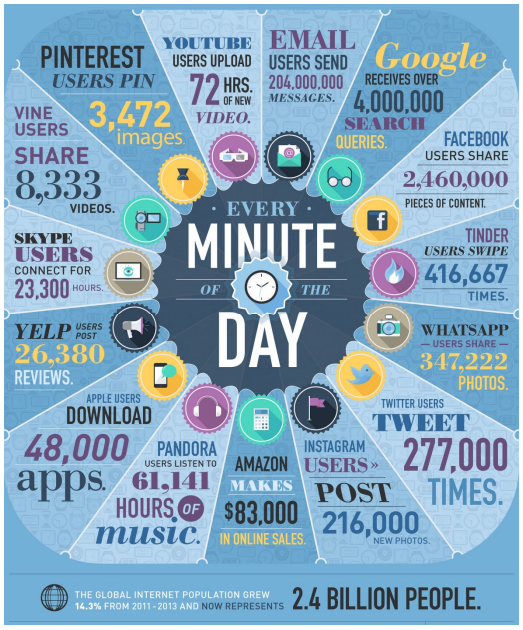
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

BC3402

Weeks 10 to 11

Assoc Prof Goh Kim Huat
Nanyang Business School

1

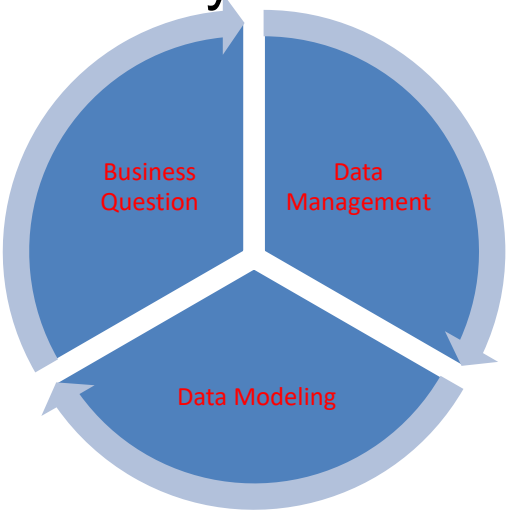


Make a guess…… What percentage of the world's data was created in the last 2 years?

2

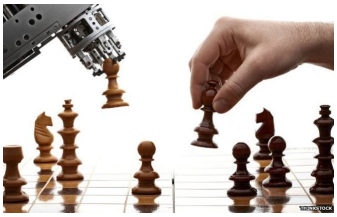
3

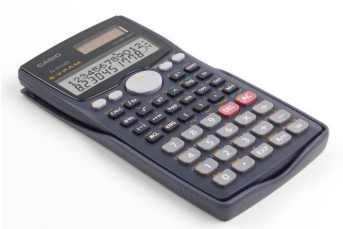
Business Analytics



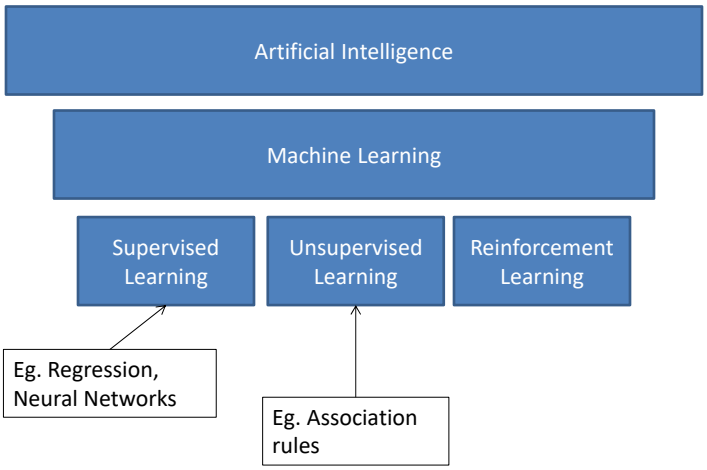
5

Which of the following is artificial intelligence?





Semantic Tree



Supervised vs. Unsupervised
Alice in Wonderland

"Would you tell me, please, which way I ought to go from here?"
"That depends a good deal on where you want to get to."
"I don't much care where –"
"Then it doesn't matter which way you go."
- Lewis Carroll

8

Supervised Learning

- Goal: Predict a single "target" or "outcome" variable
- Training data, where target value is known
- Score to data where value is not known
- Uses:
  - Explaining
  - Predicting

9

Unsupervised Learning

- Goal: no specific “target” or “outcome”
- There is no target (outcome) variable to predict or classify
- Broad, fuzzy goals NOT specific “targets”
- Multiple purposes:
  - Segment data into meaningful segments
  - detect patterns
  - explore the data
  - etc.....

Supervised: Classification

- Goal: Predict categorical target (outcome) variable
- Examples: Purchase/no purchase, fraud/no fraud, creditworthy/not creditworthy…
- Each row is a case (customer, tax return, applicant)
- Each column is a variable
- Target variable is often binary (yes/no)

Supervised: Prediction

- Goal: Predict numerical target (outcome) variable
- Examples: sales, revenue, performance
- As in classification:
- Each row is a case (customer, tax return, applicant)
- Each column is a variable
- Taken together, classification and prediction constitute "predictive analytics"

Unsupervised: Association Rules

- Goal: Produce rules that define "what goes with what"
- Example: "If X was purchased, Y was also purchased"
- Rows are transactions
- Used in recommender systems – "Our records show you bought X, you may also like Y"
- Also called "affinity analysis"

Unsupervised: Data Reduction

- Distillation of complex/large data into simpler/smaller data
- Reducing the number of variables/columns (e.g., principal components)
- Reducing the number of records/rows (e.g., clustering)

Unsupervised: Data Visualization

- Graphs and plots of data
- Histograms, boxplots, bar charts, scatterplots
- Especially useful to examine relationships between pairs of variables

Data Exploration

- Data sets are typically large, complex & messy
- Need to review the data to help refine the task
- Use techniques of Reduction and Visualization

Before we start: Data Scoping

- Unit of analysis problems
- Datafication
- Prediction of success problems

Unit of Analysis

18

Datafication: Leveraging on Existing Data

27

Matthew Fontaine Maury: Pathfinder of the sea

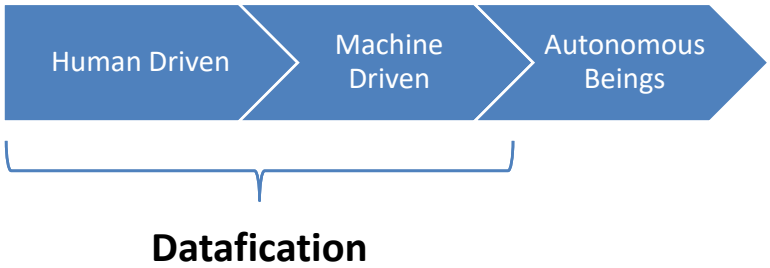
28

Three Phases

- Consumer
- Business

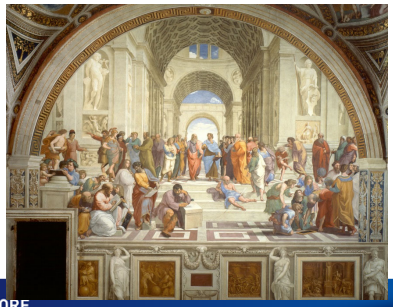
- Image recognition
- Video analytics
- Text analytics

- Drones
- Cars
- Machinery



30

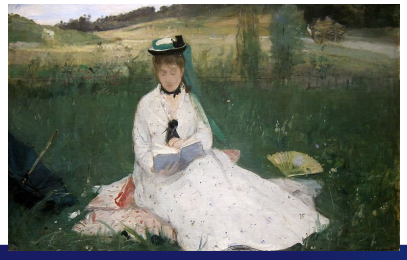
Renaissance
[1400 – 1600]



NANYANG TECHNOLOGICAL UNIVERSITY | SINGAPORE

31

(Post) Impressionism
[1865 – post 1900s]



NANYANG TECHNOLOGICAL UNIVERSITY | SINGAPORE

32

Concept of Error: Survey

LTA survey findings on MRT commuter graciousness greeted with scepticism

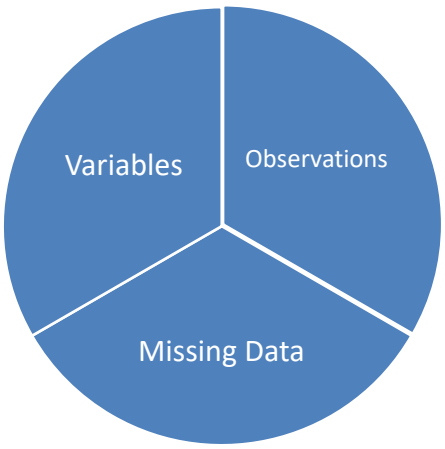
Published on Aug 22, 2013 6:50 AM

(http:// (javas www. cript:v straitstiqid(0);) mes.co

Will you give up your seat to someone who needs it more? **94% Yes**
Will you queue up and give way to aligning passenger? **98% Yes**
Will you move in for others to board the train? **96% Yes**

33

Predicting Success of Projects



Maccabitech
- 3V, 4V , 5V?
- NDA?

34

SUPERVISED LEARNING

TECHNIQUES TO EXPLAIN AND TO PREDICT

35

The Hierarchy

Level 4 – Prescriptive Analytics

Suggest what you should be doing after weighing the possible states

Level 3 – Predictive Analytics

What could happen? Different possible states

Level 2 – Diagnostic Analytics

Explaining what is happening

Level 1 – Descriptive analytics

What happened

36

Uses of Supervised Learning Tools

Explanation

- Give reasons why?
- Policy-making/ decision-making/ Justification
- Business/ consumer insights
- “More” scientific

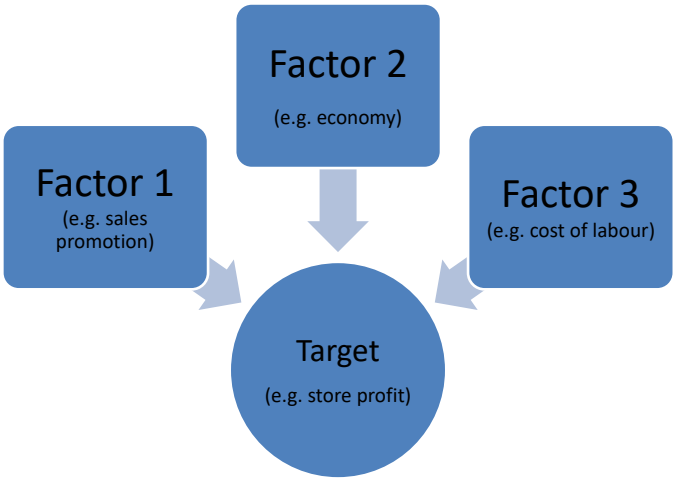
Prediction

- If X, then Y
- More application in operations
- May not care about the underlying cause
- May be “less” scientific

37

38

Regression (To Explain)



39

Traffic Statistics (Singapore)

	2010	2011	2012
Speeding Violations	205,623	225,550	244,806
Red-Running	17,185	17,492	17,705
All traffic Violations	304,472	316,214	327,503

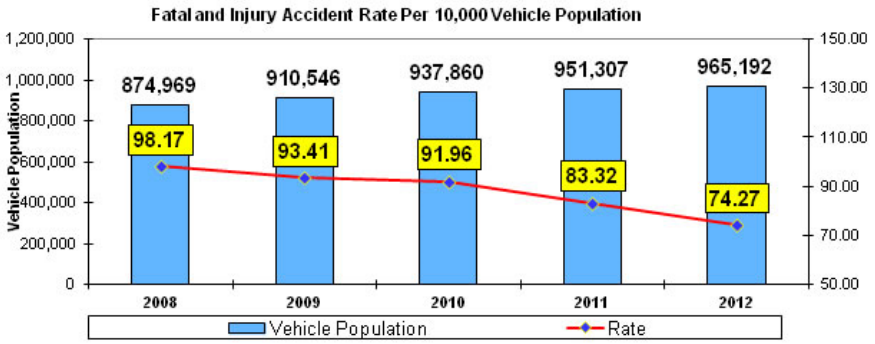
Are Singapore drivers becoming more reckless?

What is the "target" for this model to answer the above question?

http://www.sgtrafficlaws.com/

40

Traffic Statistics (Singapore)



Are Singapore drivers becoming more reckless?

What is the "target" for this model to answer the above question?

41

Proxies & Partial Out The Reasons

- Target measure is never perfect. Proxies used
  - Accuracy of proxy
- Need other factors to partial out the alternative explanations
- More factors greater fidelity

"Holding XYZ constant"
here are my conclusions

42

Questions to (Transaction) Data

CASES OF SUPERVISED LEARNING

45

ATM

Banks get a peek into customers' whims and needs

Big Data analytics provides unprecedented insights, prompts change

BY AMIT ROY CHOUDHURY

PRINT | EMAIL THIS ARTICLE

Real-time reach: Banks can engage customers based on where they are and what they are doing. - FILE PHOTO

[SINGAPORE] How often should cash at ATMs be replenished? Which products should specific customers be offered?

Thanks to Big Data analytics, Singapore-based banks are taking the guesswork out of such questions, while gaining deeper insights into customer preferences.

David Gledhill, managing director and head of group technology and operations at DBS, said that the bank used Big Data analytics to improve cash availability for its islandwide ATM service while saving on costs.

Logic + Creativity + Statistics

46

Three key questions

- Is there “something” that we need to predict or explain to achieve our overall objective
- What is the target for the model
- Can you see the target in the dataset

48

A Different Type of Transaction…

MARINA BAY Sands SINGAPORE

Panopticon

51

Detecting Fraud

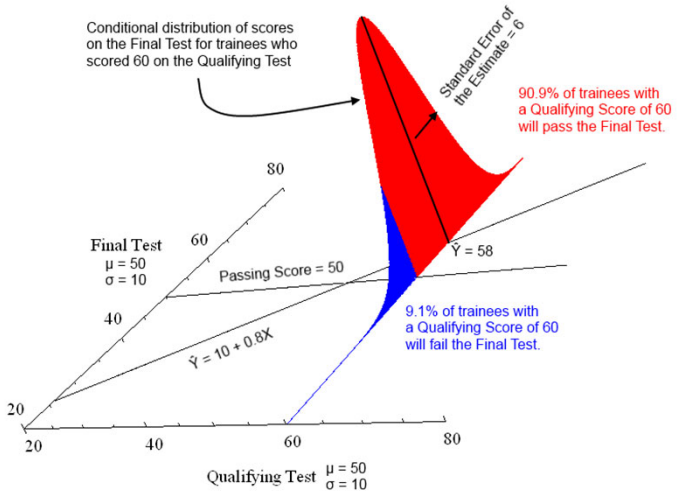
d	*P*(*d*)
1	30.1%
2	17.6%
3	12.5%
4	9.7%
5	7.9%
6	6.7%
7	5.8%
8	5.1%
9	4.6%

$$P(d) = \log_{10}(1+1/d)$$

Genius you are vs. the genius you have: Elizabeth Gilbert

Logic + Creativity + Statistics + Luck

52



It is about predicting a distribution …

53

Supervised Learning: Predicting Transaction Distribution

- Distribution of multi-dimensional behavior (ATM)
- Distribution of known probability functions (Casino)
- Distribution of natural numbers – outliers detection (Fraud detection)