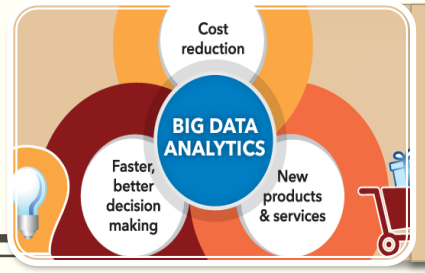
ภาษาการเขียนโปรแกรม

Week9

02-212-213

อ.ธิดาวรรณ คล้ายศรี

บทที่3



Big Data Analytics

- Data Analytics Life Cycle
- Big data concepts & Hadoop ecosystem

3.1 Data Analytics Lifecycle (วงจรชีวิตการวิเคราะห์ข้อมูล)

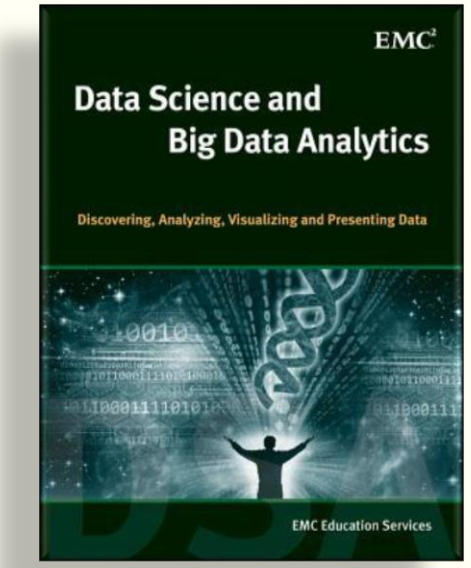
- 3.1.1 Describing Problem (กำหนดปัญหา)
- 3.1.2 Identifying Data Sources (ระบุแหล่งข้อมูล)
- 3.1.3 Preparing Data (การเตรียมความพร้อมข้อมูล)
- 3.1.4 Planing Model (วางแผนโมเดล)
- 3.1.5 Building Model (สร้างโมเดล)
- 3.1.6 Visaulising Results (นำเสนอข้อมูลวิเคราะห์)

3.2 Big Data Concepts (แนวความคิดในเรื่อง big data)

3.3 Introductory Hadoop and Its Ecosystem (บทนำและวงชีวิตของฮาดูพ)

Source of the content:

Majority of the slides: "**Data Science and Big Data Analytic**" text book by EMC Education Services



Data Introductory

แนะนำเกี่ยวกับข้อมูล

What is data?

- กลุ่ม-ระดมสมอง ตอบคำถามว่า “ข้อมูลคือ อะไร”
- อาจเป็น set of recorded facts, numbers, events ที่อาจไม่มีความหมาย
- Data จะไม่มีความหมายถ้าหากไม่นำมาทำการเชื่อมโยง
- Data ที่จัดเก็บในคอมพิวเตอร์ เป็น 0 หรือ 1

Data vs. Information

Data

- Raw facts
  - Raw data - Not yet been processed to reveal the meaning
- Building blocks of information
- **Data management**
  - Generation, storage, and retrieval of data

Information

- Produced by processing data
- Reveals the meaning of data
- Enables **knowledge** creation
- Should be accurate, relevant, and timely to enable good decision making

✓ 30082017 → Data

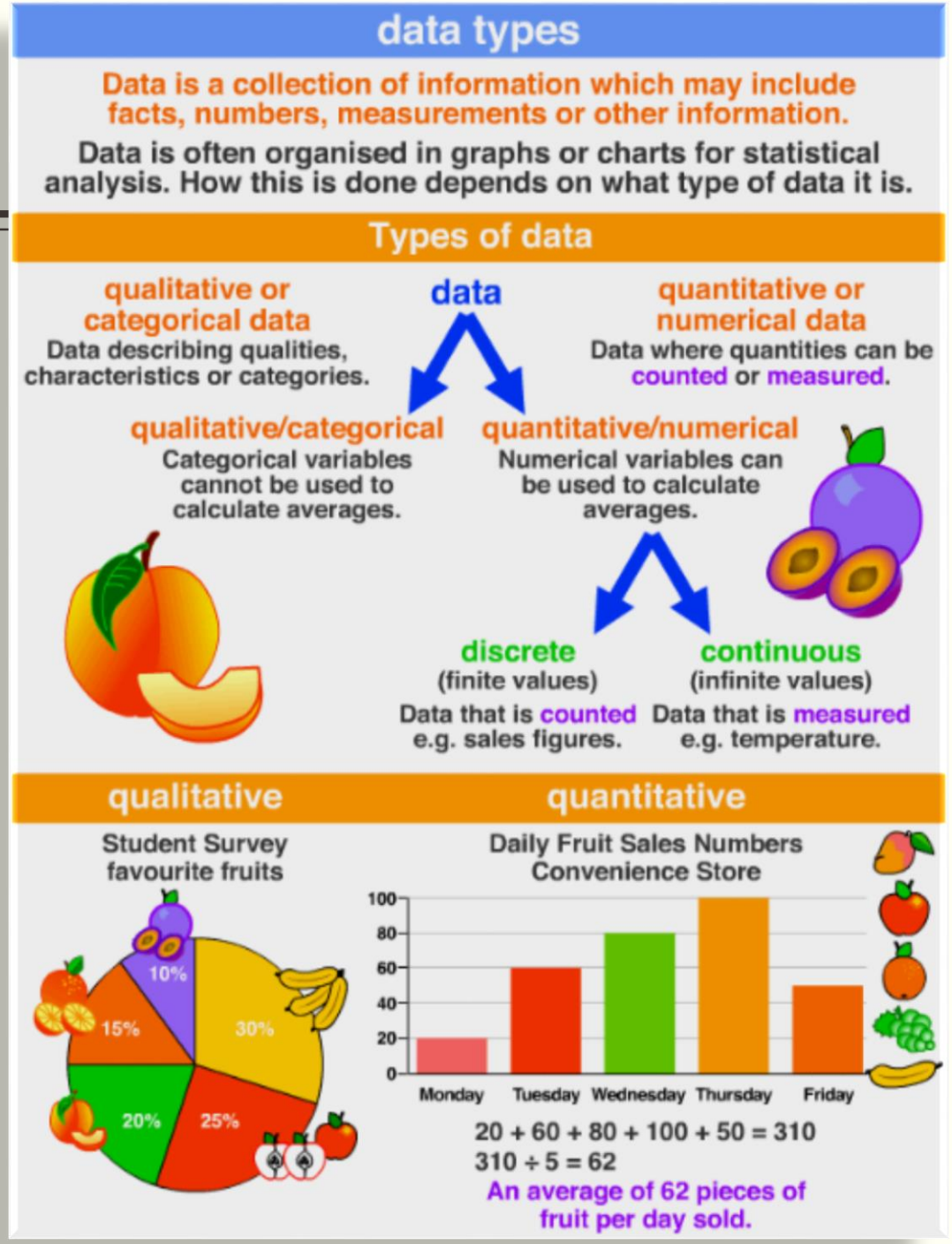
✓ 30/08/2017 → วันที่วันนี้ → Information

✓ 30,082,017 → จำนวนยอดคลิกของสมาชิกเว็บไซต์ Ebay.co.uk ครั้ง/1นาที

✓ £30,082,017 → ยอดขายสินค้า Ebay.co.uk ในครึ่งวันทำการเว็บไซต์

Types of Data

- Text
- Numeric
- Boolean/ logical
- Images
- Audio
- Video



[Content: Jenny Eather]
http://www.amathsdictionaryforkids.com/

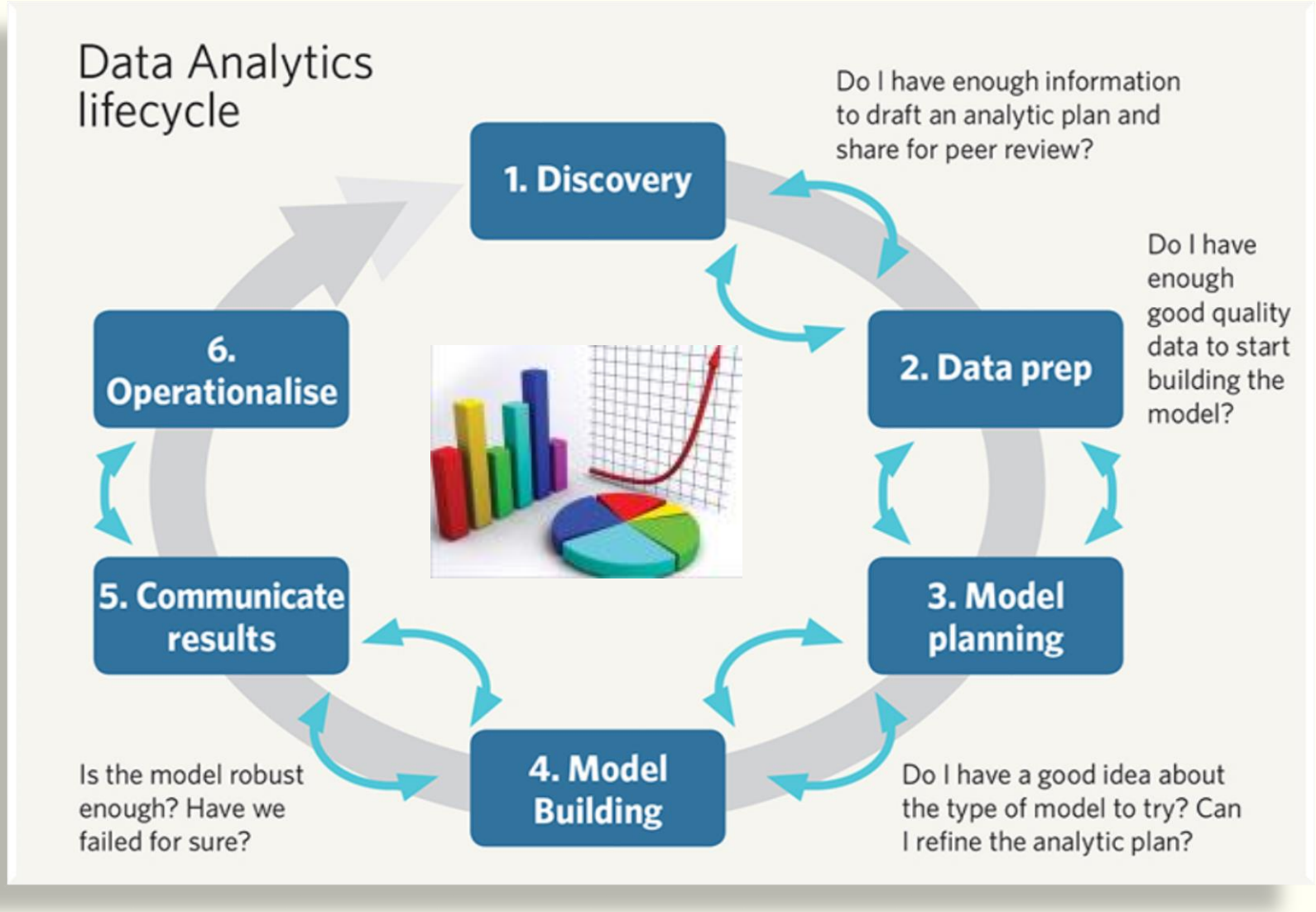
Data Integrity Definition & Guidance

Term	Data Integrity
Definition	The extent to which all data are • Complete • Consistent • Accurate throughout the data lifecycle.
Expectation or Guidance	Data integrity arrangements must ensure that the • Accuracy • Completeness • Content meaning of data is retained throughout the data lifecycle.

Part 1: 3.1 Data Analytics Life Cycle

วงจรชีวิตของการวิเคราะห์ข้อมูล

6 Processed of Data Analytics (กระบวนการวิเคราะห์ข้อมูล)



[By EMC/Dell]

Phase 1. Data Discovery

1.1 Learning business domain/ problem → Data scientist

1.2 Resources -ทีมทำการประเมิน Resources ต่าง ๆ

1.3 Describing problem = กำหนดปัญหา/วัตถุประสงค์เพื่อนำไปเขียนสมมุติฐานเบื้องต้น

1.4 Identifying Key Stakeholders=ระบุผู้ทำนำงานไปใช้/กลุ่มร่วมธุรกิจ,

ระบุความเสี่ยง/ความสำเร็จของโปรเจ็ค

1.5 Interviewing the Analytic Sponsor = เก็บ Requirements

1.6 Developing Initial Hypotheses = สมมุติฐานเบื้องต้นเพื่อนำ data มาทดสอบให้ได้ผลลัพธ์

1.7 Identifying potential data sources = ระบุแหล่งข้อมูลนำเข้า

- Data
- Tools/technology
- Systems
- Team/ people
- Project management

Question:

Do I have **enough information** to do draft an analytic plan and share for peer review?

Phase 2. Data Preparation (การเตรียมความพร้อมของข้อมูล)

2.1 เตรียม Analytic sandbox/work space ที่จะใช้ในการวิเคราะห์ข้อมูล

2.2 ทำความเข้าใจกับ Data

2.3 Data Conditioning = ตระเตรียม Data ซึ่งอาจต้องทำกระบวนการ cleaning data, normalizing data หรือทำการแปลงข้อมูลก่อนให้พร้อมที่จะนำไปวิเคราะห์

2.4 Survey and Visualize data

2.5 Toolsได้แก่ Hadoop, Apine Miner, OpenRefine, Data Wrangler

Question:

Do I have good **enough good quality data** to start building the model?

Phase 3. Model Planning

3.1 Data Exploration and Variable Selection

- Trying to understand data = ทำการอธิบายข้อมูล-ปฏิบัติการ/ใช้ค่าสถิติของข้อมูล เพื่อทำความเข้าใจข้อมูล
- คัดเลือกตัวแปรที่จะนำมาใช้สร้างโมเดล

3.2 Model Selection

- Identifying/ordering potential models

= ระบุโมเดลต่างๆ ที่จะนำมาใช้

3.3 Tools สำหรับวางแผนโมเดลได้แก่

R, SQL analysis services, SAS

Market Sector	Analytic Techniques/Methods Used
Consumer Packaged Goods	Multiple linear regression, automatic relevance determination (ARD), and decision tree
Retail Banking	Multiple regression
Retail Business	Logistic regression, ARD, decision tree
Wireless Telecom	Neural network, decision tree, hierarchical neurofuzzy systems, rule evolver, logistic regression

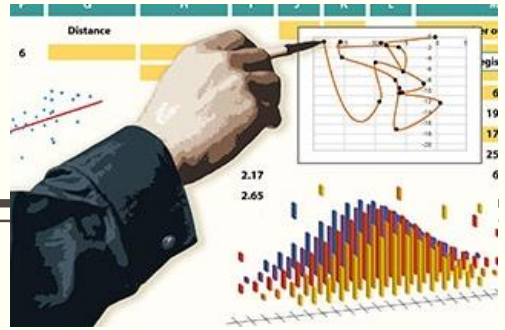
3.4 Preparing training/testing data sets

Question:

Do I have good idea about the **type of models** to try?

Can I refine the **analytic plan**?

Phase 4. Model Building



- สร้างโมเดล จาก Training set แล้วทำการ Fitting model
- ทดสอบโมเดลกับ Test set แล้วประเมินผลประสิทธิภาพการทำงานของโมเดลว่า Robust หรือไม่
- คำนึงถึง Model Over fitting/Under fitting
- Tools ที่ใช้สร้างโมเดลได้แก่ R and PL/R, SAS enterprise miner, SPSS modeller, Matlab, WEKA, Python, SQL, Octave, Alpine miner

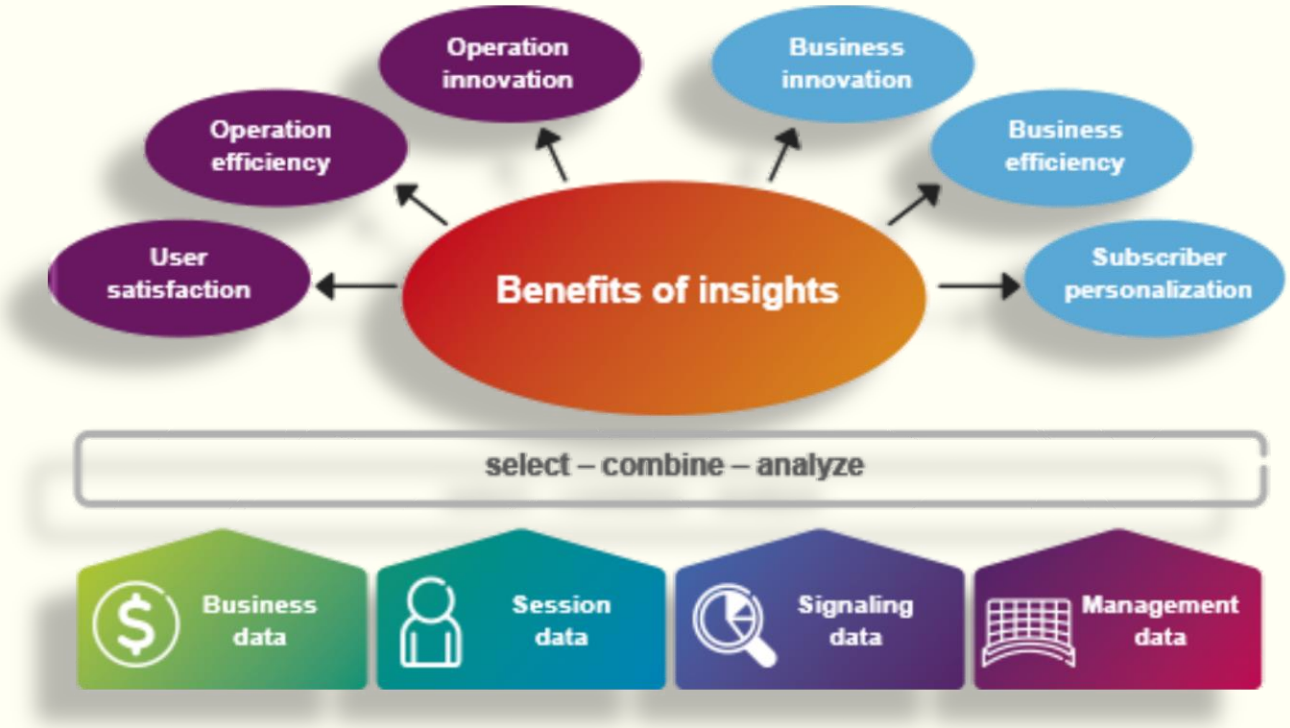
Question:

Is the model robust enough?

Have we failed for sure?

Phase 5. Communicate Results

- นำผลลัพธ์ที่ได้จากโมเดลที่รันดูว่าประสบความสำเร็จหรือล้มเหลว อาจต้องกลับไปตั้งสมมุติฐานใหม่
- Data science team จัดทำเอกสารว่าได้ findings/outcomes/knowledge/trend อะไรจากการวิเคราะห์ข้อมูล
- จะได้ business values ออกมา



Phase 6. Operationalize

- ถ่ายทอด นำเสนอความสำเร็จของโปรเจ็ควิเคราะห์ข้อมูลให้กับ Stakeholders/ sponsors ได้ทราบ
- ได้แก่ insights/ findings ต่าง ๆ
- ทำการติดตั้ง นำไปใช้งานจริงในธุรกิจตามที่วางแผนไว้
- สร้าง business values

Mini project- งานกลุ่ม

- ให้กลุ่มนศ. รับ Data set แล้วทำ case study Data Analysis โดยทำการสร้างโมเดลเพื่อพยากรณ์ข้อมูล
- โดยให้ดำเนินงานตาม Data Analytic Life Cycle ทั้ง 6 phase พร้อมส่งรายงานกลุ่ม