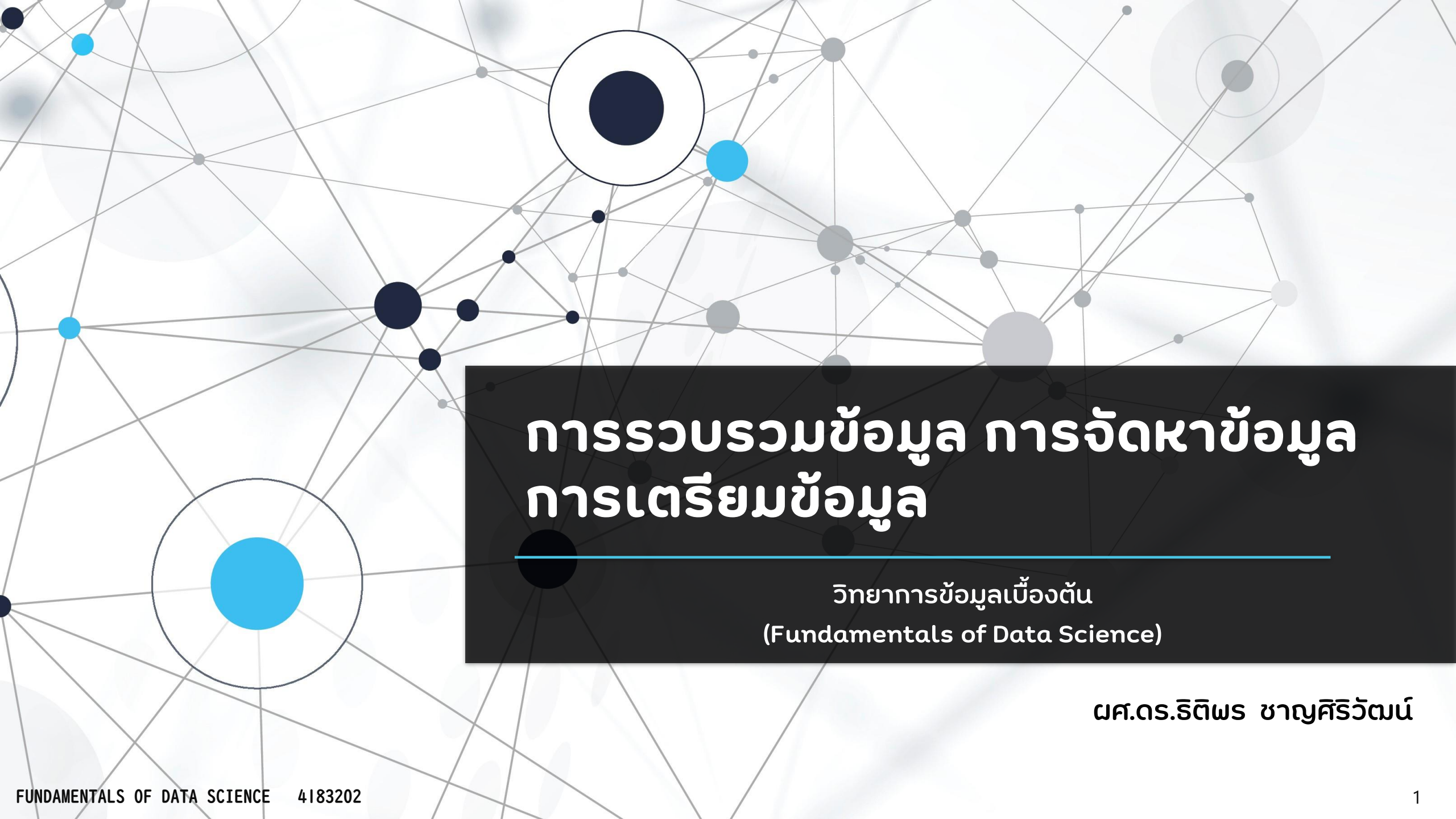


การรวบรวมข้อมูล การจัดหาข้อมูล การเตรียมข้อมูล

วิทยาการข้อมูลเบื้องต้น

(Fundamentals of Data Science)

ผศ.ดร.ธิติพร ชาญศิริวัฒน์

DATA SOURCE

DATA REPRESENTATION

DATA CLEANSING

DATA SOURCE

Questionnaires

- Paper-based questionnaires
- Electronic-based questionnaires
- Online questionnaires

Web Servers

Server software, or hardware dedicated to running said software, that can satisfy World Wide Web client requests.

Web Services

A service offered by an electronic device to another electronic device, communicating with each other via the World Wide Web

DATA SOURCE

Database

An organized collection of data, generally stored and accessed electronically from a computer system

Logs

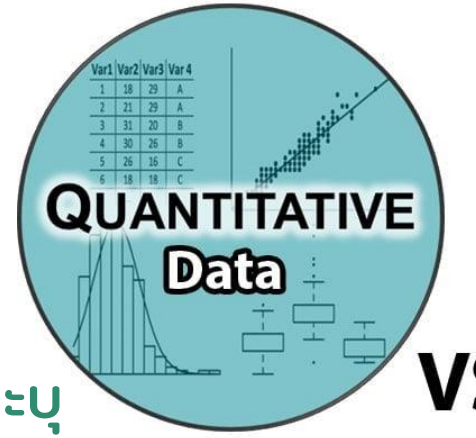
- Records of events.
- In computer, for example, a file that records either events that occur in an operating system or other software runs, or messages between different users of a communication software.

Online Repositories

- A repository is a central place in which an aggregation of data is kept and maintained in an organized way, usually in computer storage.
- An online repository is a digital library or archive which is accessible via the internet.

TYPES OF DATA

ข้อมูลเชิงปริมาณ หรือ ข้อมูลเชิงตัวเลข (Numerical data)



VS

ข้อมูลเชิงคุณภาพ หรือ ข้อมูลเชิงกลุ่ม (Categorical data)

ข้อมูลในลักษณะตัวเลขต่างๆ ระบุถึงความมากน้อยในเชิงปริมาณข้อมูลด้วยค่าต่างๆ

- Discrete data (ข้อมูลแบบไม่ต่อเนื่อง): ข้อมูลที่นับได้และมีค่าไม่ต่อเนื่อง มักอยู่ในรูปของจำนวนเต็ม ไม่สามารถแบ่งย่อยออกเป็นส่วนเล็กๆ ได้ เช่น จำนวนพนักงานในแผนก 60 คน
- Continuous data (ข้อมูลแบบต่อเนื่อง): ข้อมูลที่วัดค่าได้ภายในช่วงที่กำหนด และข้อมูลเป็นค่าที่ต่อเนื่อง เช่น ส่วนสูง อุณหภูมิ น้ำหนัก

ไม่สามารถวัดค่าด้วยอุปกรณ์ มักอยู่ในรูปแบบข้อความ หรือตัวเลขที่ไม่มีความหมายเชิงปริมาณ

- Nominal data (ข้อมูลนามบัญญัติ): จำแนกประเภทของคุณลักษณะหรือสิ่งของต่างๆ เช่น กรุ๊ปเลือด A, B, C, O, AB
- Ordinal data (ข้อมูลเชิงอันดับ): เรียงอันดับสิ่งต่างๆ ตามลักษณะหนึ่งๆ เช่น ระดับของลูกค้า Platinum, Gold, Silver

การรวบรวมข้อมูล (DATA COLLECTION)

- สัมภาษณ์ สังเกต ทดลอง แบบสอบถาม
- ฐานข้อมูลในองค์กร
- แหล่งข้อมูลภายนอก
- ชุดข้อมูล (Data Set) มาตรฐาน
- ฯลฯ

DATA PREPARATION

กระบวนการนำข้อมูลดิบที่รวบรวมมาจัดเตรียมให้อยู่ในรูปแบบที่พร้อมสำหรับการวิเคราะห์

- การสกัดข้อมูล (Data extraction)
- การรวมข้อมูล (Data integration)
- การแปลงข้อมูล (Data transformation)
- การลดรูปข้อมูล (Data reduction)
- การทำความสะอาดข้อมูล (Data cleansing)

การสกัดข้อมูล (DATA EXTRACTION)

เป็นการคัดข้อมูลที่ไม่เกี่ยวข้องออกไปจากกลุ่มตัวอย่าง
สกัดเอาเฉพาะบาง Field หรือบาง Column ที่ไม่เกี่ยวข้องออกไปจากข้อมูล
การคัดเลือกเฉพาะข้อมูลที่เราสนใจจะนำมาวิเคราะห์

- รหัสนักศึกษา
- ชื่อนักศึกษา
- ระดับชั้นปี
- สาขาวิชา
- คณะ
- จำนวนหน่วยกิตที่เรียน
- เกรดเฉลี่ยสะสม
- อาจารย์ที่ปรึกษา
- ที่อยู่
- เบอร์โทรศัพท์

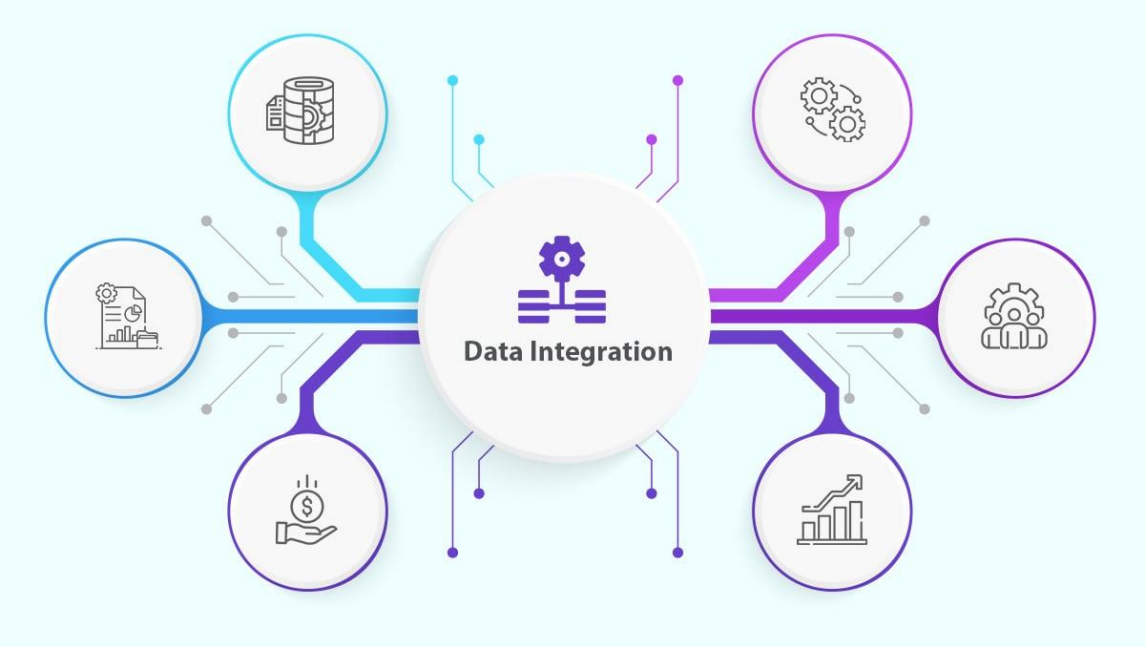
ทำนายแนวโน้ม
การพ้นสภาพของนักศึกษา

- รหัสนักศึกษา
- ชื่อนักศึกษา
- ระดับชั้นปี
- สาขาวิชา
- คณะ
- จำนวนหน่วยกิตที่เรียน
- เกรดเฉลี่ยสะสม

การรวมข้อมูล (DATA INTEGRATION)

เป็นขั้นตอนการรวมข้อมูลจากแหล่งข้อมูล ซึ่งมีข้อมูลหลากหลายมารวมไว้ที่เดียวกัน

- ปัญหาความไม่สอดคล้องกันของข้อมูล
- ประเภทของของมูลแตกต่างกัน
- รูปแบบการจัดเก็บข้อมูลที่แตกต่างกัน



การรวมข้อมูล (DATA INTEGRATION)

แหล่งข้อมูลที่ 1

รหัสนักศึกษา	Assignment Score	Final Score	ผลการเรียน
1010	A	99	Pass
1012	B	80	Fail
1013	B	75	Pass

ต.ย. ปัญหาจากจำนวนคุณลักษณะหรือจำนวนฟิลด์แตกต่างกัน

แหล่งข้อมูลที่ 2

รหัสนักศึกษา	ชื่อ-นามสกุล	คะแนนเก็บ (เต็ม 30 คะแนน)	คะแนนสอบ (เต็ม 70 คะแนน)	ผลการเรียน
115	John Doe	30	68	ผ่าน
116	James Johnson	25	67	ไม่ผ่าน
117	Maxim Gates	20	25	ไม่ผ่าน

การรวมข้อมูล (DATA INTEGRATION)

แหล่งข้อมูลที่ 1

รหัสนักศึกษา	Assignment Score	Final Score	ผลการเรียน
1010	A	99	Pass
1012	B	80	Fail
1013	B	75	Pass

ต.ย. ปัญหาจากการใช้รหัสแทนค่าข้อมูลแตกต่างกัน

แหล่งข้อมูลที่ 2

รหัสนักศึกษา	ชื่อ-นามสกุล	คะแนนเก็บ (เต็ม 30 คะแนน)	คะแนนสอบ (เต็ม 70 คะแนน)	ผลการเรียน
115	John Doe	30	68	ผ่าน
116	James Johnson	25	67	ไม่ผ่าน
117	Maxim Gates	20	25	ไม่ผ่าน

การรวมข้อมูล (DATA INTEGRATION)

แหล่งข้อมูลที่ 1

รหัสนักศึกษา	Assignment Score	Final Score	ผลการเรียน
1010	A	99	Pass
1012	B	80	Fail
1013	B	75	Pass

ต.ย. ปัญหาจากการกำหนดรูปแบบการจัดเก็บข้อมูลแตกต่างกัน

แหล่งข้อมูลที่ 2

รหัสนักศึกษา	ชื่อ-นามสกุล	คะแนนเก็บ (เต็ม 30 คะแนน)	คะแนนสอบ (เต็ม 70 คะแนน)	ผลการเรียน
115	John Doe	30	68	ผ่าน
116	James Johnson	25	67	ไม่ผ่าน
117	Maxim Gates	20	25	ไม่ผ่าน

การรวมข้อมูล (DATA INTEGRATION)

แหล่งข้อมูลที่ 1

รหัสนักศึกษา	Assignment Score	Final Score	ผลการเรียน
1010	A	99	Pass
1012	B	80	Fail
1013	B	75	Pass

ต.ย. ปัญหาจากการกำหนดมาตรฐานข้อมูลแตกต่างกัน

แหล่งข้อมูลที่ 2

รหัสนักศึกษา	ชื่อ-นามสกุล	คะแนนเก็บ (เต็ม 30 คะแนน)	คะแนนสอบ (เต็ม 70 คะแนน)	ผลการเรียน
115	John Doe	30	68	ผ่าน
116	James Johnson	25	67	ไม่ผ่าน
117	Maxim Gates	20	25	ไม่ผ่าน

[No Title]

การแปลงข้อมูล (DATA TRANSFORMATION)

เป็นขั้นตอนการแปลงข้อมูลให้มีความเหมาะสม มีมาตรฐานเดียวกัน

- เช่นการแปลงข้อมูล จาก เซนติเมตร เป็น เมตร
- จัดเก็บข้อมูลให้อยู่ในรูปแบบมาตรฐาน
- การปรับเปรียบประเภทของข้อมูล

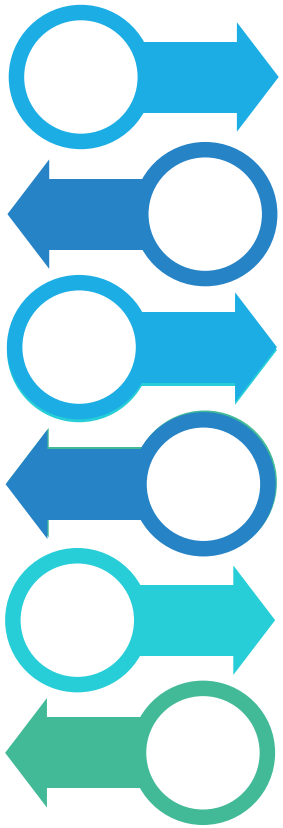
การลดรูปข้อมูล (DATA REDUCTION)

เป็นขั้นตอนการลดมิติข้อมูล เพื่อใช้ในเป็นตัวแทนข้อมูลทั้งหมดสำหรับการวิเคราะห์

ภาควิชา	ชื่อภาควิชา (ลดรูป)
วิทยาการคอมพิวเตอร์	CS
ธรณีวิทยา	GEOL
เคมี	CHEM
คณิตศาสตร์	MATH
ชีววิทยา	BIO

การทำความสะอาดข้อมูล (DATA CLEANSING)

เป็นขั้นตอนการตรวจสอบ แก้ไขข้อมูล โดยการคัดข้อมูลที่ไม่มีคุณภาพ หรือปรับปรุงข้อมูล เพื่อให้ได้ข้อมูลที่มีคุณภาพสำหรับการนำไปวิเคราะห์



- ค่าข้อมูลที่เป็นค่าว่าง ขาดหาย (Missing Values)
- ข้อมูลรบกวน (Noisy Data)
- ข้อมูลไม่สอดคล้อง (Inconsistency Data)
- ข้อมูลขาดรูปแบบ (Lack of uniformity Data)
- ข้อมูลสะกดผิด (Misspellings Data)
- ข้อมูลซ้ำ (Duplicate Data)

ค่าข้อมูลที่เป็นค่าว่าง ขาดหาย (MISSING VALUES)

รหัสนักศึกษา	คะแนนเก็บ (เต็ม 30 คะแนน)	คะแนนสอบ (เต็ม 70 คะแนน)	ผลการเรียน
111		69	ผ่าน
112	35	665	ผ่าน
113	20	1	ไม่ผ่าน
114	25	67	ผ่าน
115	30	68	ผ่าน
116	25	67	ไม่ผ่าน
117	10	40	ไม่ผ่าน
118	ขส	65	ผ่าน

ข้อมูลรบกวน (NOISY DATA)

รหัสนักศึกษา	คะแนนเก็บ (เต็ม 30 คะแนน)	คะแนนสอบ (เต็ม 70 คะแนน)	ผลการเรียน
111		69	ผ่าน
112	35	665	ผ่าน
113	20	1	ไม่ผ่าน
114	25	67	ผ่าน
115	30	68	ผ่าน
116	25	67	ไม่ผ่าน
117	20	25	ไม่ผ่าน
118	ขส	65	ผ่าน

Outliers
ค่าผิดปกติ

ข้อมูลรบกวน (NOISY DATA)

Error Values
ค่าผิดพลาด

รหัสนักศึกษา	ภาควิชา	คะแนนเก็บ (เต็ม 30 คะแนน)	คะแนนสอบ (เต็ม 70 คะแนน)	ผลการเรียน
111	Comp Sci		69	ผ่าน
112	Stat	35	665	ผ่าน
113	Maths	20	1	ไม่ผ่าน
114	maths	25	67	ผ่าน
115	Mathematics	30	68	ผ่าน
116	Computer Sc	25	67	ไม่ผ่าน
117	stat	20	25	ไม่ผ่าน
118	COMP	ขส	65	ผ่าน

การจัดการข้อมูล MISSING VALUES/ NOISY DATA

แทนที่ด้วย

- ค่า เช่น N/A หรือ none หรือ null
- ค่าน้อยที่สุด ในกรณีที่แอตทริบิวต์เป็นตัวเลข (numeric)
- ค่ามากที่สุด ในกรณีที่แอตทริบิวต์เป็นตัวเลข (numeric)
- ค่าเฉลี่ย (mean/average) ในกรณีที่แอตทริบิวต์เป็นตัวเลข (numeric)
- ค่าฐานนิยม (mode) ในกรณีที่แอตทริบิวต์เป็นกลุ่ม (nominal)
- ค่า 0 เช่น จำนวนบุตร
- ค่าที่ระบุเอง เช่น ไม่ระบุ

ข้อมูลไม่สอดคล้อง (INCONSISTENCY DATA)

รหัสนักศึกษา	คะแนนเก็บ (เต็ม 30 คะแนน)	คะแนนสอบ (เต็ม 70 คะแนน)	ผลการเรียน
111		69	ผ่าน
112	35	665	ผ่าน
113	20	1	ไม่ผ่าน
114	25	67	ผ่าน
115	30	68	ผ่าน
116	25	67	ไม่ผ่าน
117	20	25	ไม่ผ่าน
118	ขส	65	ผ่าน

วิธีจัดการวิธี INCONSISTENCY DATA

รหัสนักศึกษา	คะแนนเก็บ (เต็ม 30 คะแนน)	คะแนนสอบ (เต็ม 70 คะแนน)	ผลการเรียน
113	20	1	ไม่ผ่าน
117	20	25	ไม่ผ่าน
114	25	67	ผ่าน
116	25	67	ไม่ผ่าน
115	30	68	ผ่าน
112	35	665	ผ่าน
118	ขส	65	ผ่าน
111		69	ผ่าน

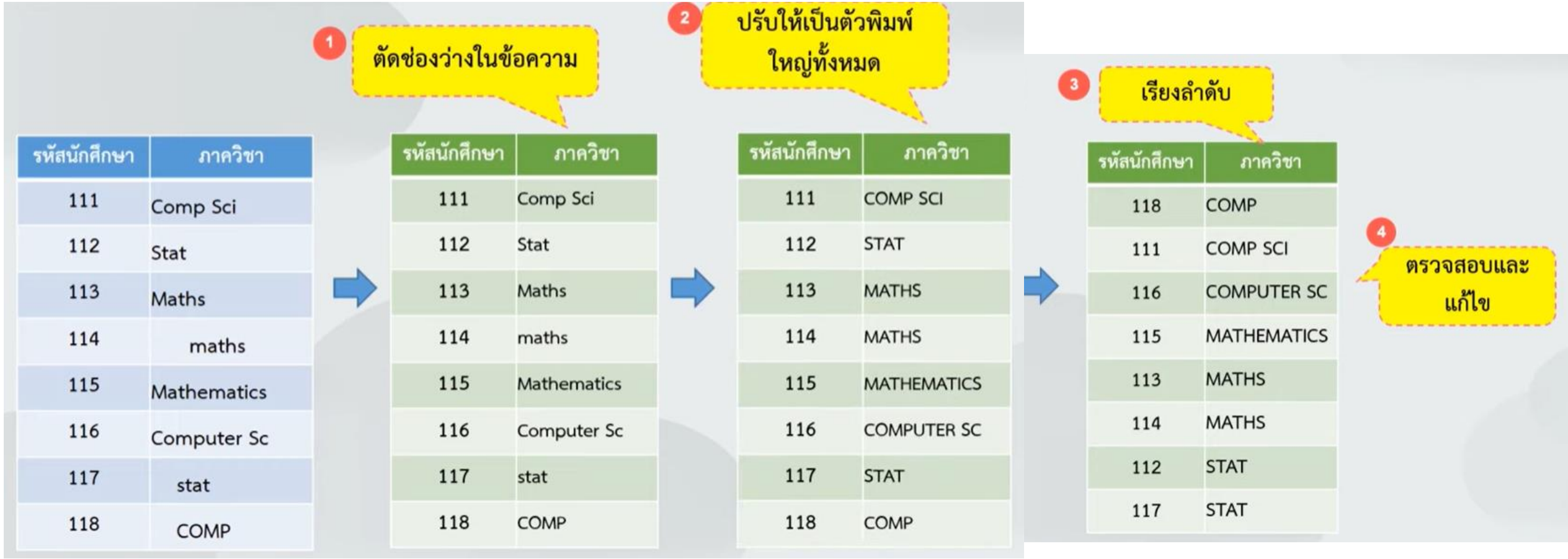
ข้อมูลขาดรูปแบบ (LACK OF UNIFORMITY DATA)

รหัสนักศึกษา	ภาควิชา
111	Comp Sci
112	Stat
113	Maths
114	maths
115	Mathematics
116	Computer Sc
117	stat
118	COMP

ข้อมูลสะกดผิด (MISSPELLINGS DATA)

รหัสนักศึกษา	ภาควิชา	ลงทะเบียนเข้าอบรม
111	Comp Sci	ลงทะเบียน
112	Stat	ไม่ลงทะเบียน
113	Maths	ไม่ลงทะเบียน
114	maths	ลงทะเบียน
115	Mathematics	ลงทะเบียน
116	Computer Sc	ไม่ลงทะเบียน
117	Statics	ลงทะเบียน
111	Comp	ลงทะเบียน
118	Maht	ลงทะเบียน

การแก้ปัญหา MISSPELLINGS DATA



รหัสนักศึกษา	ภาควิชา
111	Comp Sci
112	Stat
113	Maths
114	maths
115	Mathematics
116	Computer Sc
117	stat
118	COMP

รหัสนักศึกษา	ภาควิชา
111	Comp Sci
112	Stat
113	Maths
114	maths
115	Mathematics
116	Computer Sc
117	stat
118	COMP

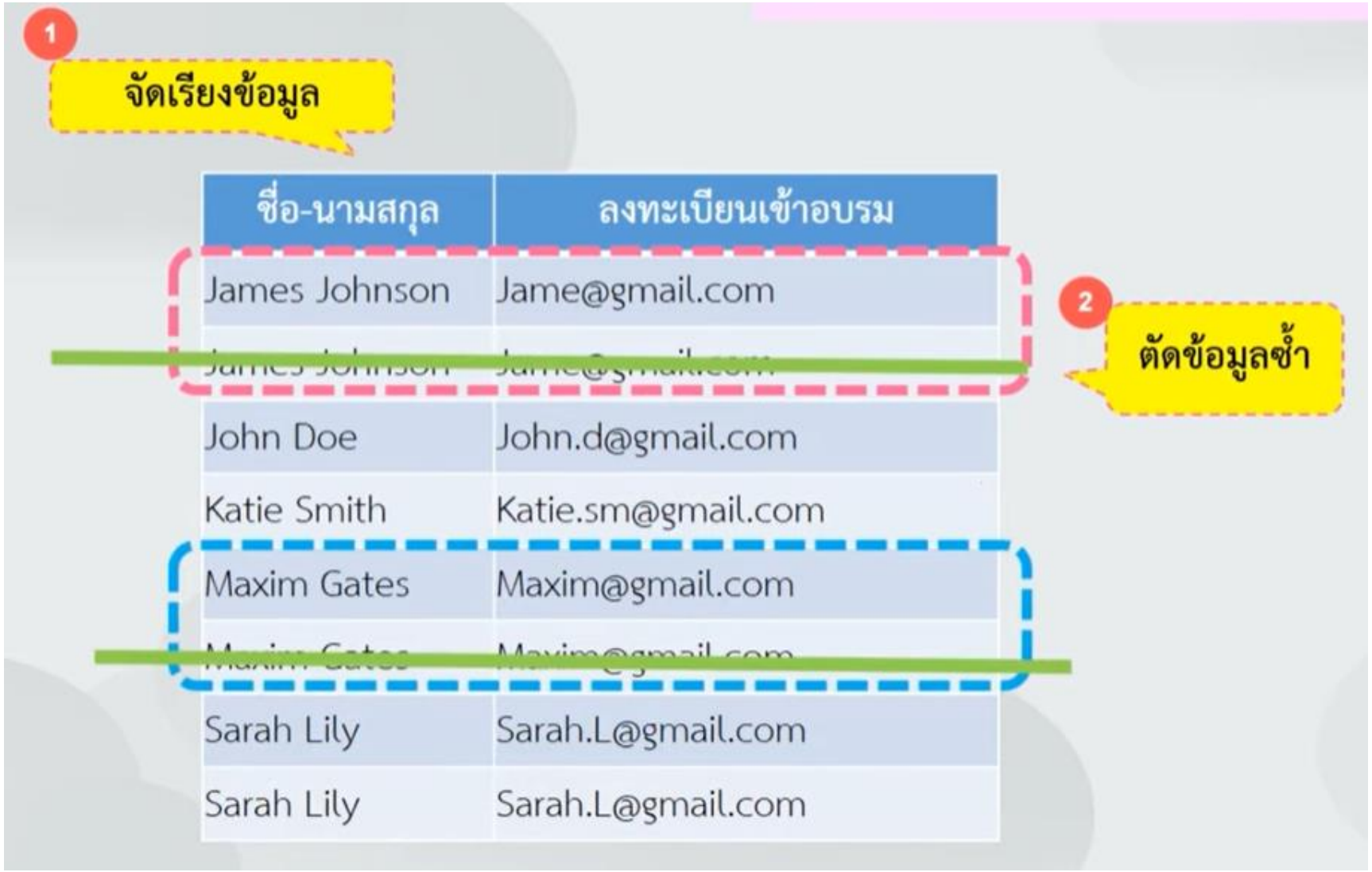
รหัสนักศึกษา	ภาควิชา
111	COMP SCI
112	STAT
113	MATHS
114	MATHS
115	MATHEMATICS
116	COMPUTER SC
117	STAT
118	COMP

รหัสนักศึกษา	ภาควิชา
118	COMP
111	COMP SCI
116	COMPUTER SC
115	MATHEMATICS
113	MATHS
114	MATHS
112	STAT
117	STAT

ข้อมูลซ้ำ (DUPLICATE DATA)

ชื่อ-นามสกุล	ลงทะเบียนเข้าอบรม
James Johnson	Jame@gmail.com
Maxim Gates	Maxim@gmail.com
Sarah Lily	Sarah.L@gmail.com
Katie Smith	Katie.sm@gmail.com
Sarah Lily	Sarah.L@gmail.com
John Doe	John.d@gmail.com
James Johnson	Jame@gmail.com
Maxim Gates	Maxim@gmail.com

การแก้ปัญหา DUPLICATE DATA




END