

Data Science Processes and Tools

Dr. Sethavidh Gertphol

Outline

- * Data Gathering
- * Data Inspection and Cleaning
- * Data Transformation and Integration
- * Data Analysis or Data Modeling
- * Model Evaluation
- * Data Visualization
- * Data Dissemination

เริ่มต้นจากการตั้งคำถาม

- * เราต้องการวิเคราะห์หรือสร้าง โมเดลอธิบายและทำนายอะไร เพื่อ**สร้างคุณค่าเพิ่ม**ให้กับองค์กร
 - * ปัจจัยใดที่เพิ่มความเสี่ยงที่นิสิตจะเรียนไม่จบภายใน 4 ปี
 - * ทำนายตั้งแต่ นิสิตจบปี 1 ได้ไหม
- * เราต้องการข้อมูล ใดเพื่อสร้าง โมเดลที่ต้องการ
 - * การลงทะเบียนและเกรดของนิสิต วิชาที่มีปัญหา
 - * ข้อมูลการเข้าเรียนและการทำกิจกรรม
 - * ข้อมูล โรงเรียนเก่าและเกรดที่ได้ คะแนนสอบเข้า
 - * ข้อมูลทางประชากรศาสตร์
- * อย่าให้ข้อมูลที่มีจำกัดเป้าหมายของการวิเคราะห์ ให้หาแหล่งข้อมูลเพิ่ม
- * ควร**ศึกษาวิเคราะห์ข้อมูลที่มีเบื้องต้น**ก่อนที่จะเริ่มตั้งคำถามอย่างจริงจัง

Data Gathering

- * การเก็บรวบรวมข้อมูลเป็นส่วนประกอบหนึ่งของการทำวิจัยในทุกๆ ศาสตร์
- * วิธีในการรวบรวมข้อมูลมีได้หลากหลาย
 - * Manual: Interviews (direct), Questionnaires (indirect), Case Studies, Existing Documents
 - * API e.g. Google API, Facebook API, Youtube API, Twitter API
 - * Web scraping e.g. webhose.io, import.io
 - * Sensors (IoT cabbage)
- * เป้าหมาย
 - * การรวบรวมข้อมูลที่มีคุณภาพเพียงพอที่สามารถวิเคราะห์เพื่อตอบโจทย์ปัญหาที่ตั้งไว้ได้

Manual: Questionnaires

- * วัตถุประสงค์

- * ต้องการความรวดเร็วและเรียบง่ายในการเก็บรวบรวมข้อมูลจากคนจำนวนมากและจำนวนคำถามจำนวนมาก

- * ข้อดี

- * สามารถทำได้โดยไม่ต้องระบุตัวตน
- * ค่าใช้จ่ายต่ำ
- * ง่ายในการเปรียบเทียบและวิเคราะห์
- * ได้ข้อมูลจำนวนมาก

Manual: Existing Documents

- * วัตถุประสงค์

- * เพื่อนำข้อมูลในอดีตในรูปแบบต่างๆ เช่น ข้อมูลในฐานข้อมูล รายงานการประชุม เอกสารบันทึกการเข้างานเอกสารบันทึกทางการเงิน เป็นต้น มารวบรวมกันเป็นชุดข้อมูลและจึงนำมาดำเนินการต่อไป

- * ข้อดี

- * สามารถนำข้อมูลมาใช้ได้ทันที
- * ไม่เสียเวลา
- * ข้อมูลในองค์กรมีจำนวนมาก

API

- * Application Programming Interface

- * “It is a set of clearly defined methods of **communication** between various **software components**.”

- * Data Format

- * JSON - JavaScript Object Notation

- * XML - EXtensible Markup Language

- * YAML is a human friendly data serialization standard for all programming languages

- * <http://yaml.org/>

JSON

- * Developer Friendly format
- * The most famous use as API format
- * Key-Value Pair format
 - * Separate key and value with :
 - * Group property by { }

```
1 {  
2   "title": "Age Gate",  
3   "type": "object",  
4   "properties": {  
5     "firstName": {  
6       "type": "string"  
7     },  
8     "knownValue": {  
9       "type": "boolean"  
10    },  
11    "age": {  
12      "description": "Age in years",  
13      "type": "integer",  
14      "minimum": 18  
15    },  
16  },  
17  "required": ["firstName", "lastName"]  
18 }
```


XML

- * Standard Generalized Markup Language
- * Markup and Content
- * A tag is a markup construct that begins with < and ends with >
- * Start-tag, such as <section>;
- * End-tag, such as </section>;
- * Empty-element tag, such as <line-break />.

```
<?xml version="1.0" encoding="UTF-8"?>  
<note>  
  <to>Tove</to>  
  <from>Jani</from>  
  <heading>Reminder</heading>  
  <body>Don't forget me this weekend!</body>  
</note>
```

Source: <https://www.w3schools.com/xml/default.asp>

API: YouTube API

“The YouTube Data API lets you incorporate functions normally executed on the YouTube website into your own website or application. The lists below identify the different types of resources that you can retrieve using the API. The API also supports methods to insert, update, or delete many of these resources.”

<https://developers.google.com/youtube/v3/docs/>

id

Ks-_Mh1QhMc

EXECUTE

part

snippet,contentDetails,statistics

Load in APIs Explorer

200 (OK)

```
/**
 * API response
 */
{
  "kind": "youtube#videoListResponse",
  "etag": "\"DuHzAJ-eQIiCip7p4ldoVcVA0eY/wfExypbRoPnT9CYQdAoV4BNUuSE\"",
  "pageInfo": {
    "totalResults": 1,
    "resultsPerPage": 1
  },
  "items": [
    {
      "kind": "youtube#video",
      "etag": "\"DuHzAJ-eQIiCip7p4ldoVcVA0eY/Gd21ddV4xfqnnqeMcjp0vXcG1TA\"",
      "id": "Ks-_Mh1QhMc",
      "snippet": {
        "publishedAt": "2012-10-01T15:27:35.000Z",
        "channelId": "UCAuUUnT6oDeKwE6v1NGQxug",
        "title": "Your body language may shape who you are | Amy Cuddy",
```

API: Facebook API

Access Token: CAACEdEose0cBAHFJKxh9taNU6jKWIJrQzjUj90XxXtQYX0P5do3o0cpvndxaFL! X Debug ⇌ Get Access Token ⇌ Get App Token

Graph API FQL Query

GET → /v2.2/cocacola Submit

Learn more about the Graph API syntax.

Node: cocacola

Search for a field

```
{
  "id": "40796308305",
  "about": "The Coca-Cola Facebook Page is a collection of your stories showing how people from around the world have helped make Coke into what it is today.",
  "can_post": true,
  "category": "Food/beverages",
  "checkins": 13591,
  "cover": {
    "cover_id": "10152297032458306",
    "offset_x": 0,
    "offset_y": 0,
    "source": "https://fbcdn-sphotos-f-a.akamaihd.net/hphotos-ak-prn2/v/t1.0-9/s720x720/625442_10152297032458306_574021701_n.jpg?oh=8f881acd51650ae8c66a6be3123e188b&oe=552738E9&__gda__=1429403392_b8d599146136ece911a37360a3a128a2",
    "id": "10152297032458306"
  },
  "description": "Created in 1886 in Atlanta, Georgia, by Dr. John S. Pemberton, Coca-Cola was first offered as a fountain beverage at Jacob's Pharmacy by mixing Coca-Cola syrup with carbonated water. \n\nCoca-Cola was patented in 1887, registered as a trademark in 1893 and by 1895 it was being sold in every state and territory in the United States. In 1899, The Coca-Cola Company began franchised bottling operations in the United States. \n\nCoca-Cola might owe its origins to the United States, but its popularity has made it truly universal. Today, you can find Coca-Cola in virtually every part of the world.\n\nCoca-Cola Page House Rules: http://CokeURL.com/q28a",
  "founded": "1886",
  "has_added_app": false,
  "is_community_page": false,
  "is_published": true,
  "likes": 92685430,
  "link": "https://www.facebook.com/coca-cola",
  "name": "Coca-Cola",
  "parking": {
    "lot": 0,
    "street": 0,
    "valet": 0
  },
  "talking_about_count": 1000000
}
```

Source: http://vitalflux.com/wp-content/uploads/2015/01/facebook_graph_api_explorer_cocacola.png

Web Scrapping (สกัดข้อมูลจากหน้าเว็บ)

- * ข้อมูลอยู่บนหน้าเว็บแต่ไม่มี API ให้ดึงข้อมูล

- * สามารถเขียน โปรแกรมสกัดข้อมูลออกมาได้

- * Tools

 - * BeautifulSoup Python Library

 - * Perl

 - * Awk & Sed

- * ถ้ามี API ให้ใช้ API ก่อน

- * Scrape อย่างมีจริยธรรมด้วย






E-Labsheet | 01418112 Fundamental Programming Concept Exam. (for Exam

View/Grade Assignments

[All Assignments](#)

Assignment: 601 Midterm Set A → PPNHM - ค่าจอตกรอีกแล้ว ☒ ช้อนอัตโนมัติหลัง save

Number of submitted assignments: 188 out of 228 student(s)

-  (Score: 0/0) [[view all submissions](#)]
Submitted at Oct 08 2017 12:05:19 from 10.3.13.83. Result: FAILED ❌ [-----]
-  (Score: 0/0) [[view all submissions](#)]
Submitted at Oct 08 2017 12:05:09 from 10.3.13.30. Result: FAILED ❌ [PPPPPPPP-]
-  (Score: 0/0) [[view all submissions](#)]
Submitted at Oct 08 2017 12:04:55 from 10.3.13.107. Result: FAILED ❌ [P--P-PPPP]
-  (Score: 0/0) [[view all submissions](#)]
Submitted at Oct 08 2017 12:04:49 from 10.3.13.67. Result: FAILED ❌ [P--PPPP-PP]
-  (Score: 0/0) [[view all submissions](#)]
Submitted at Oct 08 2017 12:04:21 from 10.3.13.62. Result: FAILED ❌ [SSSSSSSSSS]

Log

- * Log คือไฟล์ที่บันทึกการทำงานของระบบคอมพิวเตอร์หรือบันทึกข้อมูลจากเซ็นเซอร์ (Internet of Things)
- * มีรูปแบบที่แน่นอนเพราะสร้างโดยระบบคอมพิวเตอร์
- * แต่มักไม่อยู่ในรูปแบบที่นำมาใช้ได้ทันที
- * ต้องเขียน โปรแกรมสรุปข้อมูลจาก Log ก่อน

Log Example

```
(b5810402542/10.3.4.74)10.3.4.47 LOGIN: [2016-12-17 11:21:30] (b5810601804/10.3.4.47)10.3.20.204 SUBMIT: id: 3480338, task-id: 9071, sect-id: 503 [2016-12-17 11:21:25] (b5910450131/10.3.20.204)10.3.4.66 SUBMIT: id: 3480327, task-id: 9063, sect-id: 503 [2016-12-17 11:21:06] (b5710450588/10.3.4.66)10.3.20.48 SUBMIT: id: 3480325, task-id: 9069, sect-id: 503 [2016-12-17 11:21:01] (b5910450255/10.3.20.48)10.3.20.47 SUBMIT: id: 3480324, task-id: 9071, sect-id: 503 [2016-12-17 11:21:00] (b5910450247/10.3.20.47)10.3.4.97 SUBMIT: id: 3480323, task-id: 9063, sect-id: 503 [2016-12-17 11:20:56] (b5910400185/10.3.4.97)10.3.20.202 SUBMIT: id: 3480322, task-id: 9067, sect-id: 503 [2016-12-17 11:20:54] (b5610450802/10.3.20.202)10.3.20.117 SUBMIT: id: 3480321, task-id: 9065, sect-id: 503 [2016-12-17 11:20:52] (b5910406469/10.3.20.117)10.3.20.48 SUBMIT: id: 3480320, task-id: 9069, sect-id: 503 [2016-12-17 11:20:51] (b5910450255/10.3.20.48)10.3.4.97 SUBMIT: id: 3480319, task-id: 9063, sect-id: 503 [2016-12-17 11:20:50] (b5910400185/10.3.4.97)10.3.4.52 SUBMIT: id: 3480318, task-id: 9069, sect-id: 503 [2016-12-17 11:20:46] (b5810400574/10.3.4.52)10.3.4.85 SUBMIT: id: 3480316, task-id: 9065, sect-id: 503 [2016-12-17 11:20:46] (b5510405686/10.3.4.85)10.3.20.117 SUBMIT: id: 3480315, task-id: 9069, sect-id: 503 [2016-12-17 11:20:42] (b5910406469/10.3.20.117)10.3.4.27 SUBMIT: id: 3480314, task-id: 9069, sect-id: 503 [2016-12-17 11:20:40] (b5910402668/10.3.4.27)10.3.20.202 SUBMIT: id: 3480313, task-id: 9067, sect-id: 503 [2016-12-17 11:20:38] (b5610450802/10.3.20.202)10.3.4.97 SUBMIT: id: 3480312, task-id: 9063, sect-id: 503 [2016-12-17 11:20:37] (b5910400185/10.3.4.97)10.3.4.57 SUBMIT: id: 3480310, task-id: 9063, sect-id: 503 [2016-12-17 11:20:28] (b5910406221/10.3.4.57)10.3.20.42 SUBMIT: id: 3480309, task-id: 9069, sect-id: 503 [2016-12-17 11:20:24] (b5710450944/10.3.20.42)10.3.4.68 SUBMIT: id: 3480307, task-id: 9069, sect-id: 503 [2016-12-17 11:20:17] (b5910401131/10.3.4.68)10.3.4.102 SUBMIT: id: 3480306, task-id: 9069, sect-id: 503 [2016-12-17 11:20:10] (b5910406213/10.3.4.102)10.3.4.106 SUBMIT: id: 3480305, task-id: 9065, sect-id: 503 [2016-12-17 11:20:08] (b5910406086/10.3.4.106)10.3.20.88 SUBMIT: id: 3480297, task-id: 9069, sect-id: 503 [2016-12-17 11:19:38] (b5910450026/10.3.20.88)10.3.20.94 SUBMIT: id: 3480296, task-id: 9069, sect-id: 503 [2016-12-17 11:19:30] (b5710401943/10.3.20.94)10.3.20.85 SUBMIT: id: 3480294, task-id: 9069, sect-id: 503 [2016-12-17 11:19:28] (b5910406345/10.3.20.85)10.3.20.119 SUBMIT: id: 3480292, task-id: 9067, sect-id: 503 [2016-12-17 11:19:16] (b5910406477/10.3.20.119)10.3.4.47 SUBMIT: id: 3480289, task-id: 9069, sect-id: 503 [2016-12-17 11:18:55]
```


Data Inspection

- * ทำความเข้าใจข้อมูลเบื้องต้น โดยเฉพาะข้อมูลที่เราไม่ได้เก็บมาเอง
 - * จำนวนแถวและคอลัมน์ทั้งหมดของข้อมูล
 - * ข้อมูลแต่ละคอลัมน์หมายถึงอะไร?
 - * หน่วยของข้อมูล (ล้าน พันล้าน)
 - * ปกติแล้วจะมีเอกสารกำกับ ลองอ่านและทำความเข้าใจ
- * การตรวจทานความถูกต้องและความสมบูรณ์ของข้อมูล เช่น
 - * ชื่อสิ่งเดียวกันแต่มีการเก็บข้อมูลที่แตกต่างกันใน (กรุงเทพมหานคร, กรุงเทพฯ, กทม)
 - * ข้อมูลที่ขาดหายไป
 - * อาจเป็นช่องว่าง หรืออาจบันทึกเลข 0 หรืออาจบันทึก NA หรือ ...

Data Cleansing

- * กำจัดข้อมูลที่ไม่จำเป็นต่อการวิเคราะห์
 - * คำอธิบายข้อมูล
 - * ข้อมูลซ้ำ
 - * ข้อมูลที่ไม่ครบถ้วน
- * เติมข้อมูลที่ขาดหายไปหรือผิดพลาด (ถ้าทำได้)
 - * ชื่อที่ต่างกันแต่หมายถึงสิ่งเดียวกัน
 - * ค่าที่ขาดหายไปอาจแทนได้ด้วยค่าเฉลี่ย
 - * เทคนิคอื่นที่ใช้เติมข้อมูลที่ขาดหาย (data imputation)
- * เพื่อให้ข้อมูลมีคุณภาพที่เหมาะสมก่อนนำไปประมวลผล

Data Transformation

- * การเปลี่ยนข้อมูลจากรูปแบบหนึ่งไปเป็นอีกรูปแบบหนึ่งให้เหมาะสมกับการวิเคราะห์
- * ปรับรูปแบบข้อมูลจากหลายแหล่งให้เหมือนกัน
- * ปรับรูปแบบตาราง
- * ปรับภาพสีให้เป็น **gray scale** ที่มีความละเอียดเท่ากัน
- * ปรับความละเอียดของช่วงเวลา
- * รวมข้อมูลจากหลายตารางเข้าด้วยกัน
- * การปรับเปลี่ยนมักถูกกำหนดโดยวิธีการวิเคราะห์ข้อมูล

world_bank.csv

คำอธิบายข้อมูล ต้องตัดทิ้ง

	A	B	H	I	J	K
1	Data Source	World Development				
2	Last Updated Date	7/22/2016				
3	Country Name	Country Code	1963	1964	1965	1966
81	Gabon	GAB	2858765931	2988967100	3238047761	3384019102
82	United Kingdom	GBR	7.85E+11	8.25E+11	8.48E+11	8.65E+11
83	Georgia	GEO			7010305063	7538575805
84	Ghana	GHA	7877147681	8051179672	8161400230	7813864110

ข้อมูลหาย ทำไงดี?

olympic.csv

ตัดรหัสประเทศที่ซ้ำ ตัด () [] ทิ้ง

	A	B	C	D	E	F
1		No Summer	01 !	02 !	03 !	Total
50	Ghana (GHA) [GHA]	13	0	1	3	4
51	Great Britain (GBR) [GBR] [Z]	27	236	272	272	780
52	Greece (GRE) [Z]	27	30	42	39	111
53	Grenada (GRN)	8	1	0	0	1

ควรแยกรหัสประเทศออกมาเป็นอีกคอลัมน์ (เหมือนในไฟล์ world_bank.csv)

ประเทศที่มีใน

olympics แต่ไม่

มีใน world_bank

และกลับกัน?

merge?

Data Integration

- * การนำข้อมูลจากหลากหลายที่มาแล้วนำมารวมกันเพื่อให้เกิดเป็นชุดข้อมูลใหม่
- * ETL tools - Extract Transform Load
- * งานตั้งแต่ gathering ถึง integration นั้นจุกจิกและใช้เวลานานที่สุดในกระบวนการของ Data Science



Data Analysis

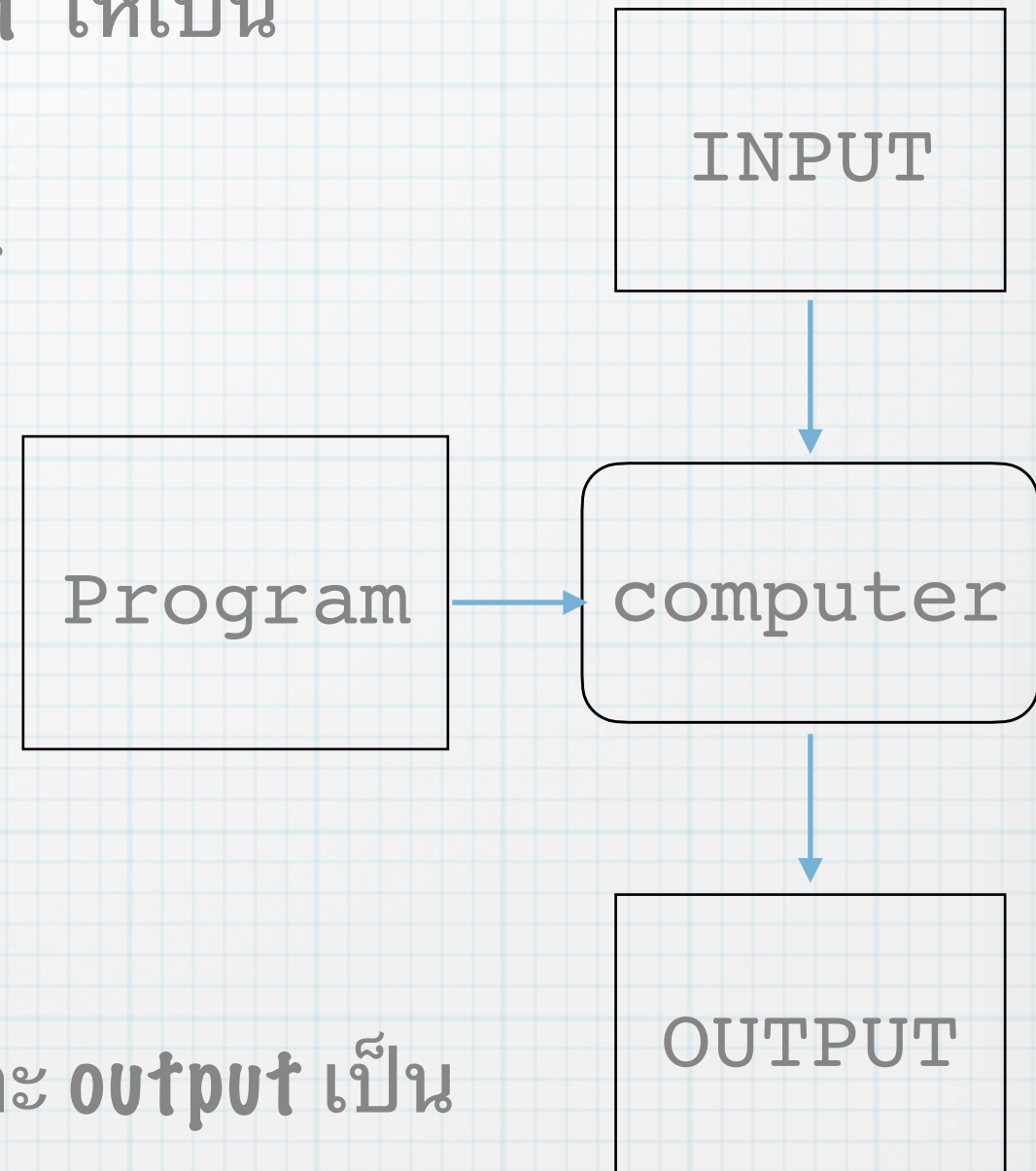
- * การวิเคราะห์มีหลายรูปแบบ
- * **Descriptive:** เพื่อให้ทราบสถานการณ์ตั้งแต่อดีตถึงปัจจุบัน
 - * มักใช้สถิติเชิงพรรณนา (Descriptive Statistics) ในการวิเคราะห์
 - * อาจไม่จำเป็นต้องตั้งโจทย์ไว้ก่อน
- * **Diagnostic:** เพื่อให้เข้าใจสาเหตุของสิ่งที่เกิดขึ้นจนถึงปัจจุบัน
 - * วิเคราะห์ในหลายแง่มุมจนเกิด insight
 - * อาจใช้การตั้งสมมติฐานว่าสิ่งที่เกิดขึ้นนั้นอาจเกิดจากอะไร
 - * แล้วใช้สถิติเชิงอนุมาน (inferential statistics) ในการทดสอบสมมติฐาน

การวิเคราะห์ข้อมูล

- * **Predictive:** ทำนายอนาคต
 - * ใช้วิธีการสร้างโมเดลแทนข้อมูลในอดีต
 - * โมเดลสามารถคำนวณผลลัพธ์ในอนาคตได้
 - * วิธีการสร้างโมเดลมีทั้งกระบวนการทางสถิติและ Machine Learning
- * **Prescriptive:** หาการกระทำที่ได้ผลดีที่สุด
 - * ใช้เทคนิคด้าน **optimization** ในการหาลำดับการทำงานที่ดีที่สุด

Data Modeling

- * โปรแกรมคือชุดคำสั่งในการเปลี่ยน **input** ให้เป็น **output**
 - * แสดงความสัมพันธ์ระหว่าง **input** กับ **output**
- * **โมเดล**ก็คล้ายกับ โปรแกรม
- * ข้อแตกต่างคือ
 - * โปรแกรมนั้นถูกสร้างขึ้น โดย โปรแกรมเมอร์
 - * โมเดลนั้น**ถูกอุปนัย**ขึ้นจากข้อมูล
- * ทางสถิติจะเรียก **input** เป็น ตัวแปรต้น และ **output** เป็น ตัวแปรตาม



ตัวแปรต้นและตัวแปรตาม

- * ค่าของตัวแปรตามจะขึ้นกับตัวแปรต้น
- * ศาสตร์ต่างกันอาจจะเรียกชื่อตัวแปรต้นตัวแปรตามต่างกัน

ตัวแปรต้น	ตัวแปรตาม	ศาสตร์
Input	Output	Computer Science
Independent variable	Dependent variable	คณิตศาสตร์, สถิติ
Regressor, control variable, explanatory	Response variable, outcome variable	สถิติ
Feature, attribute	Label, target attribute	Machine Learning

โมเดล กับ โปรแกรม

- * เราสร้างโปรแกรมเมื่อเข้าใจกฎเกณฑ์ในการเปลี่ยน input เป็น output อย่างชัดเจน

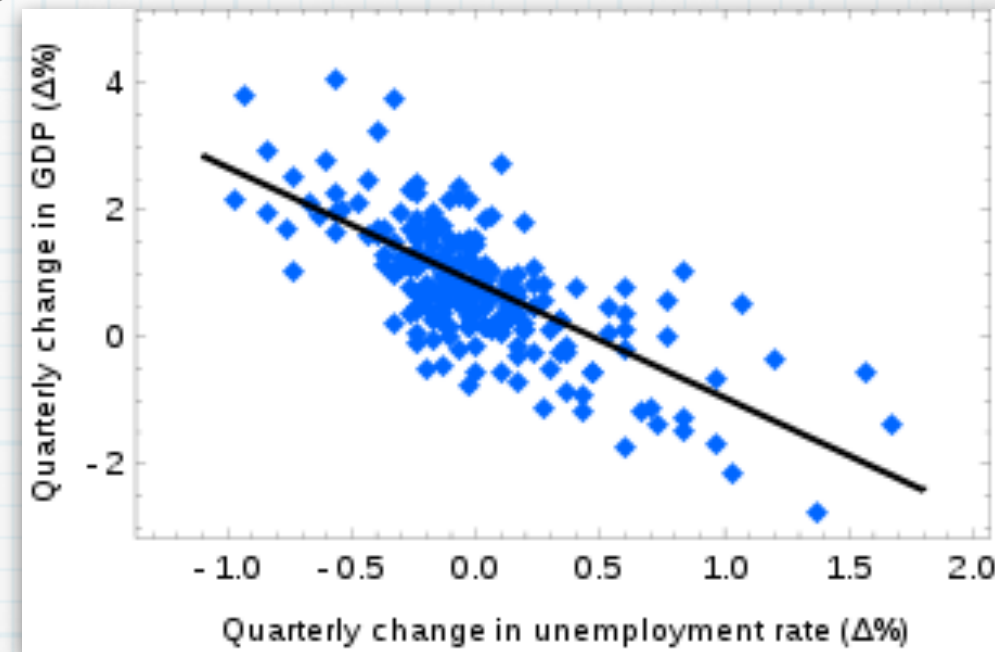
- * ค่าจอดรถ: 15 นาทีแรกฟรี ต่อไปชั่วโมงละ 20 บาท
เศษของชั่วโมงคิดเป็นหนึ่งชั่วโมง สูงสุด 8 ชั่วโมง

- * เราสร้างโมเดลในกรณีที่กฎเกณฑ์ในการแปลงข้อมูลเข้าเป็นข้อมูลออกนั้นซับซ้อนไม่ชัดเจน หรือมีความไม่แน่นอนเข้ามาเกี่ยวข้อง

- * การเปลี่ยนแปลงจำนวนผู้ว่างงานส่งผลต่อ GDP อย่างไร

- * สร้างโมเดลจากข้อมูลการเปลี่ยนแปลงของ GDP กับ การเปลี่ยนแปลงจำนวนคนว่างงานในแต่ละไตรมาส

- * $\% \text{Change GDP} = 0.789 - 1.654 * (\text{Change Unemployment Rate})$



https://en.wikipedia.org/wiki/Okun%27s_law

ประเภทของ โมเดล

* Regression

- * หาความสัมพันธ์ระหว่างตัวแปรต้นกับตัวแปรตาม
- * ตัวแปรตามต้องเป็นประเภทตัวเลข (numeric)
- * ตัวแปรต้นอาจมีมากกว่าหนึ่งตัว แต่ไม่จำเป็นต้องเป็นประเภท numeric ทุกตัว

* Classification

- * โมเดลใช้ทำนายว่าตัวอย่างจัดอยู่ในประเภทใด
- * ตัวแปรตามต้องเป็นประเภท category

* Regression กับ Classification ถือเป็น Supervised Learning คือตัวอย่างต้องมีผลเฉลย (ตัวแปรตาม)

* Clustering

- * ใช้จัดข้อมูลให้เป็นกลุ่ม โดยตัวอย่างในกลุ่มเดียวกันจะมีความ“เหมือนกัน” มากกว่าตัวอย่างนอกกลุ่ม
- * จัดเป็น Unsupervised Learning คือตัวอย่างไม่มีผลเฉลย (ตัวแปรตาม)

Evaluation

- * หลังจากวิเคราะห์ข้อมูลแล้ว ควรตรวจสอบประเมินผลการวิเคราะห์ด้วยว่าถูกต้องแม่นยำแค่ไหน
- * **Descriptive:** ตรวจสอบผลลัพธ์ของการวิเคราะห์ด้วยเครื่องมือหรือวิธีการอื่น
- * **Diagnostic:** วิเคราะห์ว่า **insight** ที่ได้นั้นมั่นใจได้แค่ไหน มีวงกว้างแค่ไหน และมีผลกระทบแค่ไหน
- * **Predictive:** ประเมินว่า โมเดลที่สร้างขึ้นมามีความแม่นยำถูกต้องมากน้อยแค่ไหน
- * **Prescriptive:** ใช้มาตรวัดตรวจสอบว่าการกระทำให้ผลลัพธ์ที่ต้องการนั้นดีขึ้นจริงแค่ไหน

Model Evaluation for Supervised Learning

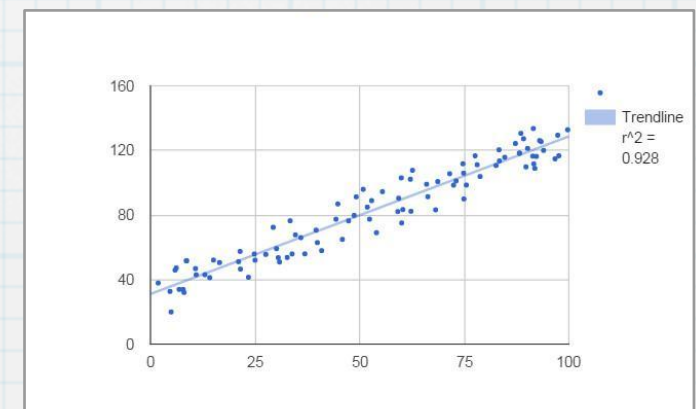
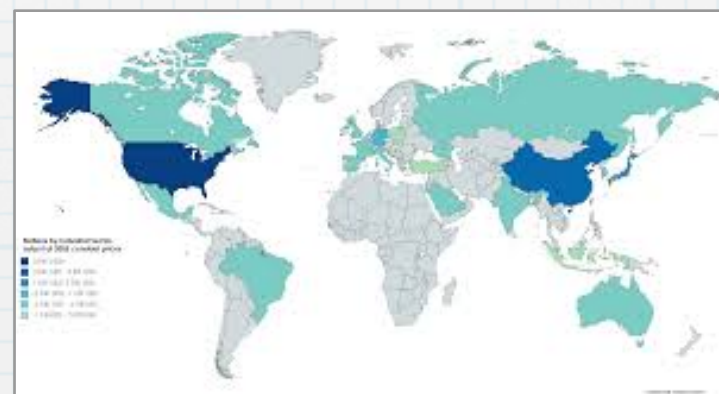
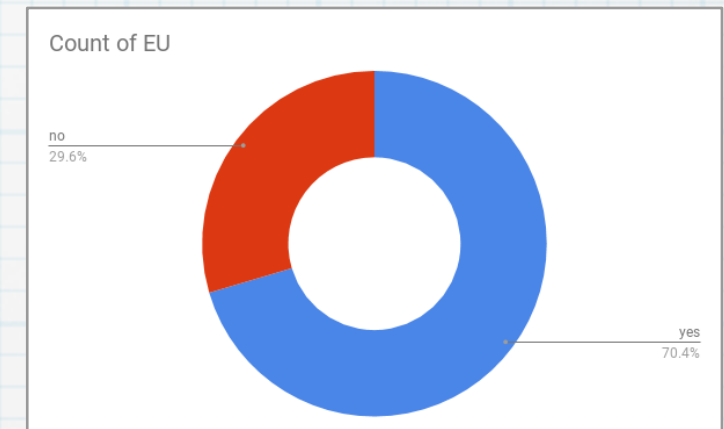
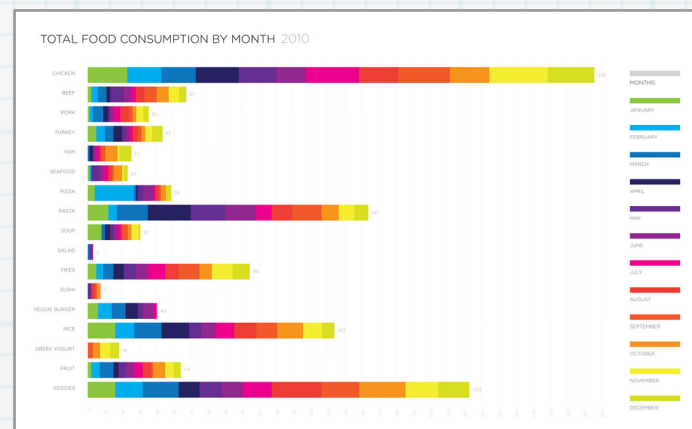
- * การสร้าง โมเดลจากข้อมูลเพื่ออธิบายสิ่งใดนั้นเป็นวิธีการทาง **อุปนัย (induction)** เพื่อ**เหมารวม (generalization)** สิ่งนั้นทั้งหมด จากตัวอย่างที่มี
- * ไม่สามารถ "พิสูจน์" ได้ว่า โมเดลที่ได้นั้นถูกต้อง
- * **Confidence level** คือค่าที่ระบุความมั่นใจว่าสมมติฐานที่สรุปมานั้น เป็นผลที่มาจากข้อมูล โดย**ไม่ใช่ความบังเอิญ**
- * **Model Evaluation** คือ การวัดประสิทธิภาพ โมเดลด้วยเทคนิคต่างๆ เพื่อทำให้มั่นใจว่า โมเดลสามารถทำงานได้ดีเมื่อนำไปใช้กับข้อมูล ในอนาคต
- * ปัญหาคือเราไม่มีข้อมูลในอนาคตมาทดสอบ
- * ใช้เทคนิคกระบวนการทดสอบที่ "จำลอง" ข้อมูล ในอนาคต

Data Visualization

- * เทคนิคการนำเสนอด้วยรูปภาพ แผนภาพหรือภาพเคลื่อนไหวเพื่อแสดงถึงข้อมูลที่ต้องการสื่อ

- * Tools

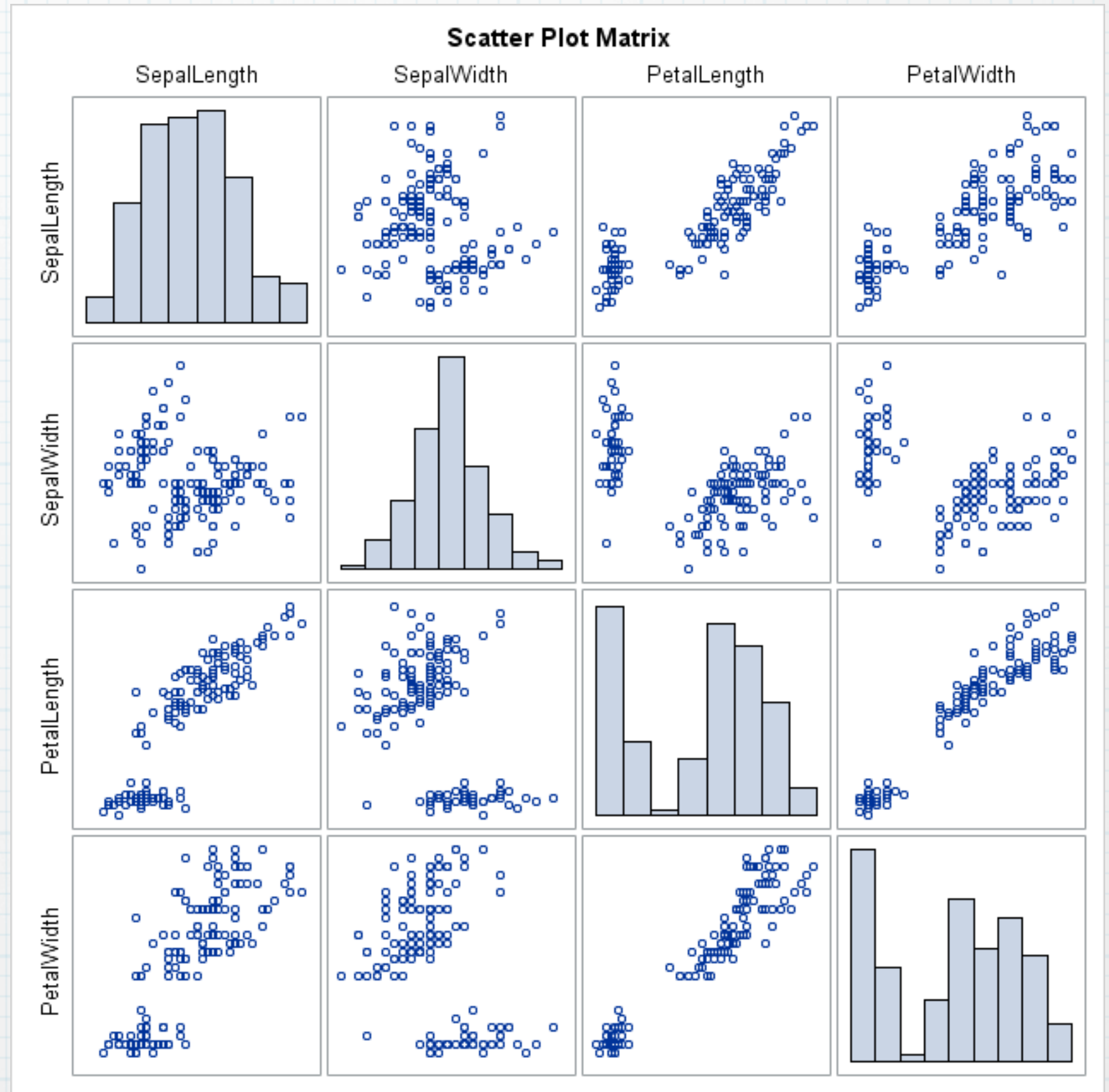
- * Google Sheet
- * Microsoft Excel
- * Tableau Public
- * PowerBI



รูปแบบการทำ Data Visualization

- * มีสองแบบหลัก
- * **Exploratory Visualization**
 - * มักทำระหว่างการสร้าง โมเดล
 - * เพื่อให้ค้นหารูปแบบความสัมพันธ์ระหว่างข้อมูลง่ายขึ้น
- * **Explanatory Visualization**
 - * ทำเมื่อสร้าง โมเดลเสร็จแล้ว
 - * ใช้เพื่ออธิบายและสื่อสารกับผู้รับ

Exploratory Visualization



<http://proc-x.com/2011/08/visualizing-correlations-between-variables-in-sas/>

Explanatory: นโยบายเลียนแบบคริสต์เชียว

Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.

Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.

Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zones. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Chiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre. Pour mieux faire juger à l'oeil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mohilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.

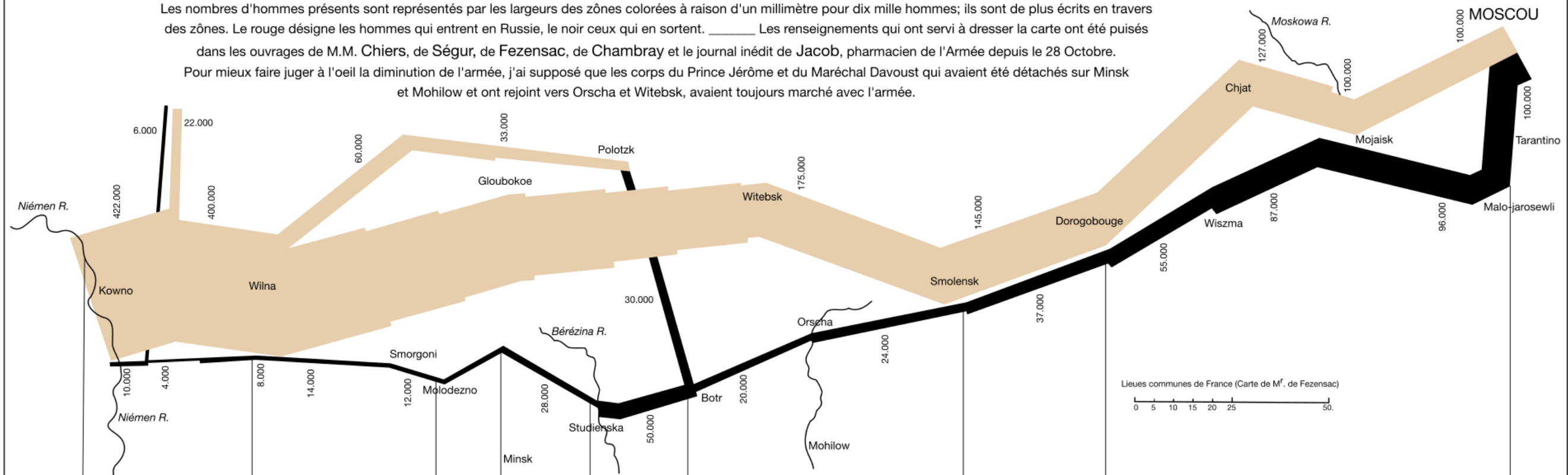
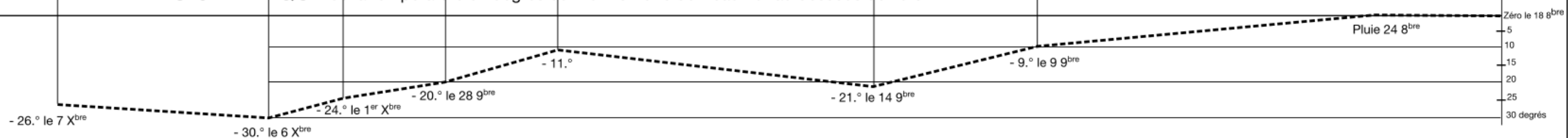


TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.



Autog. par Regnier, 8. Pas. S^{te} Marie S^t G^{er} à Paris.

[Vectorization CC-BY-SA martingrandjean.ch 2014]

Imp. Lith. Regnier et Dourdet.

Source: By Martin Grandjean - Grandjean Martin, Historical Data Visualization: Minard's map vectorized and revisited, 2014, online: <http://www.martingrandjean.ch/historical-data-visualization-minard-map/>, CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=32985771>

Data Dissemination

- * การเผยแพร่ข้อมูลสู่สาธารณะเพื่อนำไปใช้ให้เกิดประโยชน์ต่อไป เช่น
 - * เผยแพร่ข้อมูลที่ใช้วิเคราะห์
 - * การเผยแพร่ขึ้น Cloud ทำเป็น Report Services
 - * การสร้างคลิปวิดีโอเพื่อเผยแพร่
 - * การเผยแพร่กระบวนการแต่ละขั้นตอน
 - * การให้บริการการใช้งาน โมเดลผ่าน Web Service

References

- * Hoeller, S. (2018). 9 countries that ceased to exist in the 20th century. [online] Business Insider. Available at: <http://www.businessinsider.com/countries-that-no-longer-exist> [Accessed 24 May 2018].
- * Saedsayad.com. (2018). Model Evaluation. [online] Available at: http://www.saedsayad.com/model_evaluation_r.htm [Accessed 24 May 2018].
- * En.wikipedia.org. (2018). Application programming interface. [online] Available at: https://en.wikipedia.org/wiki/Application_programming_interface [Accessed 24 May 2018].
- * Sandoval, K. (2018). What Data Formats Should My API Support? | Nordic APIs I. [online] Nordic APIs. Available at: <https://nordicapis.com/what-data-formats-should-my-api-support/> [Accessed 24 May 2018].
- * En.wikipedia.org. (2018). JSON. [online] Available at: <https://en.wikipedia.org/wiki/JSON> [Accessed 24 May 2018].
- * En.wikipedia.org. (2018). XML. [online] Available at: <https://en.wikipedia.org/wiki/XML> [Accessed 24 May 2018].