

Machine Learning Basics

อ. ปรัชญ์ ปิยะวงศ์วิศาล

Pratch Piyawongwisal

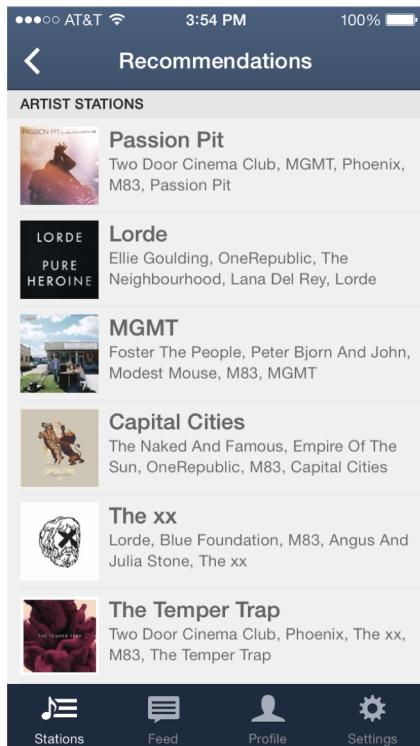
Today

- Machine Learning
 - Supervised vs Unsupervised Learning
 - Classification vs Regression
- Lab: Supervised Learning

What is Machine Learning (ML) ?

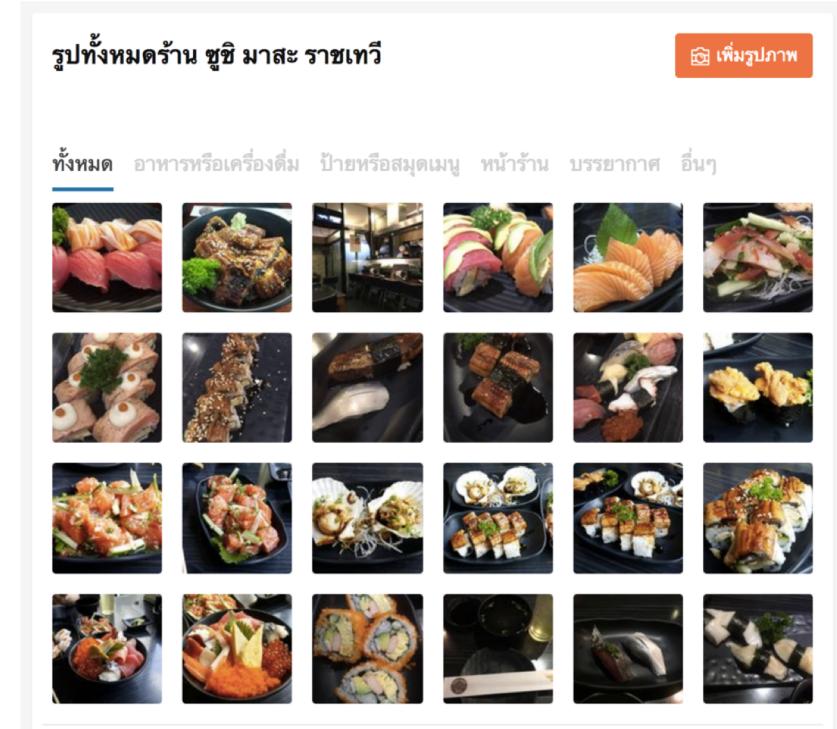
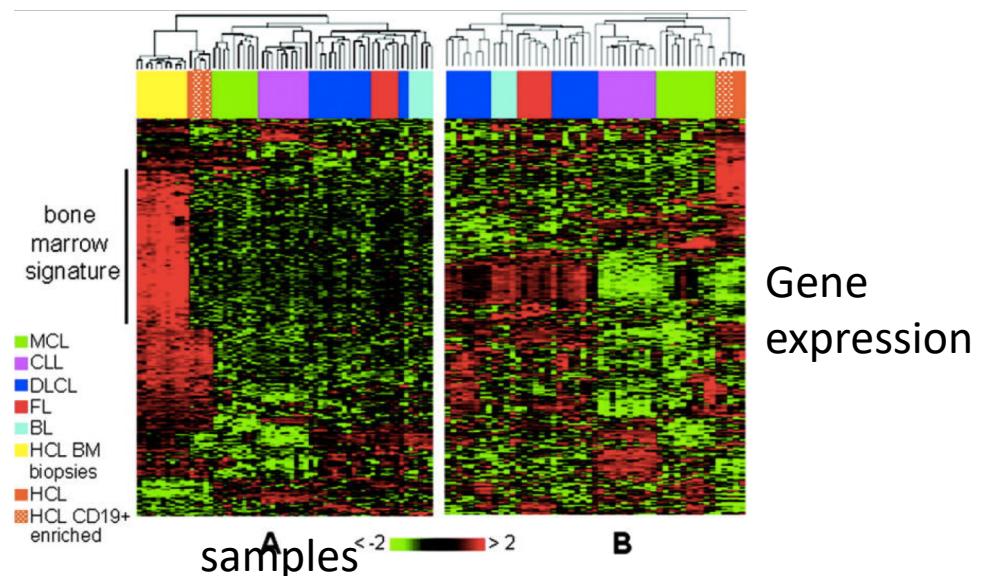
- The science of getting machines to “**learn**” from data and make predictions **without being explicitly programmed**
- การเรียนรู้ของเครื่อง คือการสอนให้เครื่องแก้ปัญหา โดยให้เครื่องเรียนรู้จากข้อมูลเอง แทนการเขียนโปรแกรมแก้ปัญหาตรงๆ
- เครื่องจะหา **pattern** ในข้อมูลเพื่อนำไปใช้ทำนายคำตอบที่ถูกต้องแม่นยำ

Machine Learning เป็นสิ่งใกล้ตัว



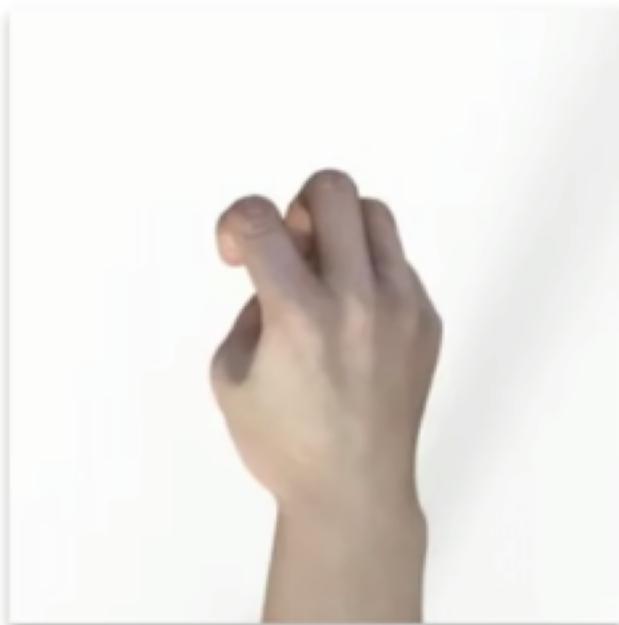
A screenshot of the Amazon.com website. The main heading is "amazon.com" with the Amazon smile logo. To the right, the text "Recommended for You" is displayed. Below this, a message says: "Amazon.com has new recommendations for you based on items you purchased or told us you own." Three book covers are shown with "LOOK INSIDE!" arrows:

- Google Apps Deciphered: Compute in the Cloud to Streamline Your Desktop
- Google Apps Administrator Guide: A Private-Label Web Workspace
- Googlepedia: The Ultimate Google Resource (3rd Edition)



Example 1

- สมมติว่าเราจะเขียนเกม **rock-paper-scissors** โดยใช้ภาพถ่ายจากมือถือ



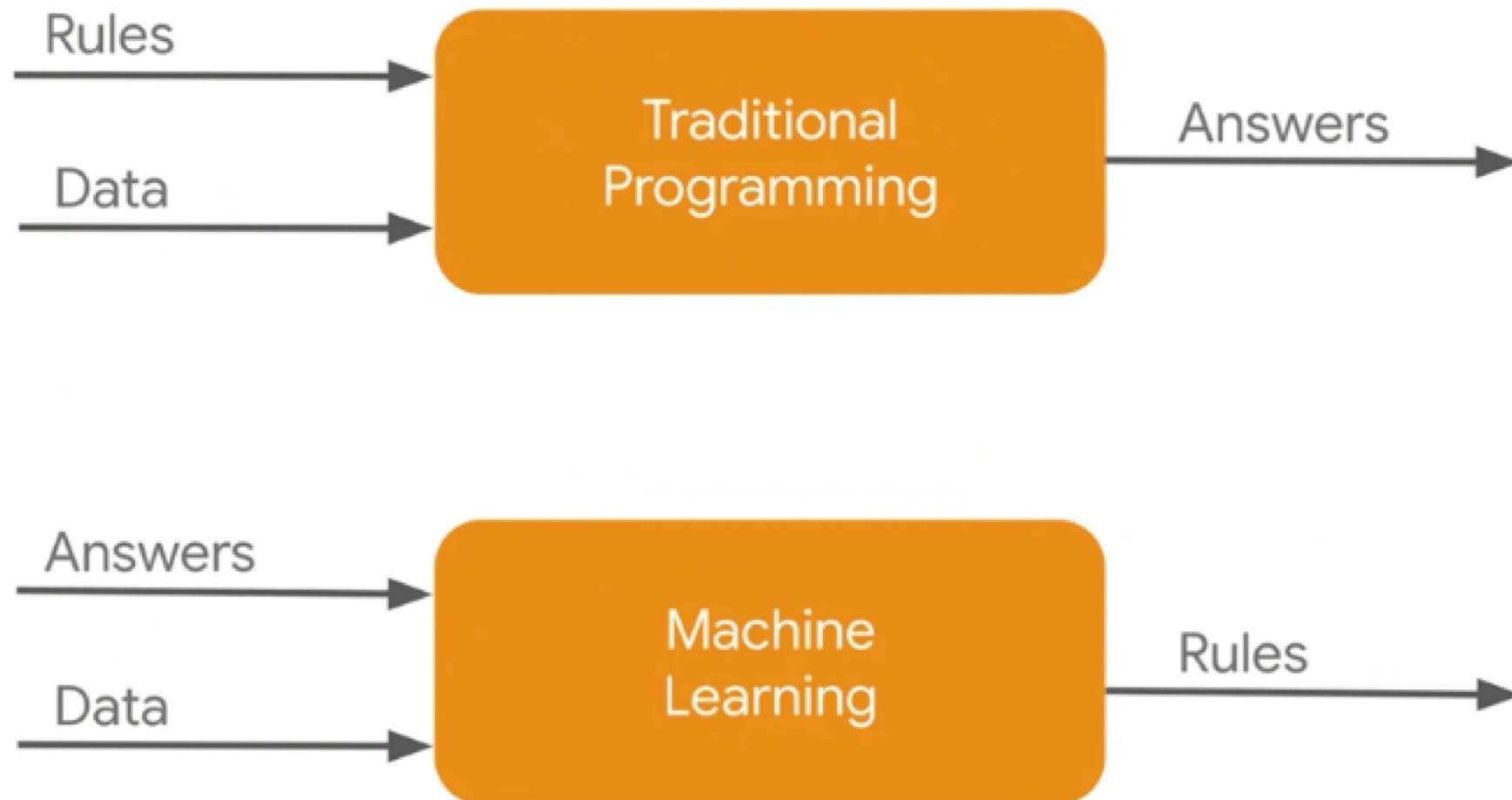
Example 2

- เขียนโปรแกรมเพื่อคัดกรองอีเมล spam ออกจากอีเมลปกติ (ham)



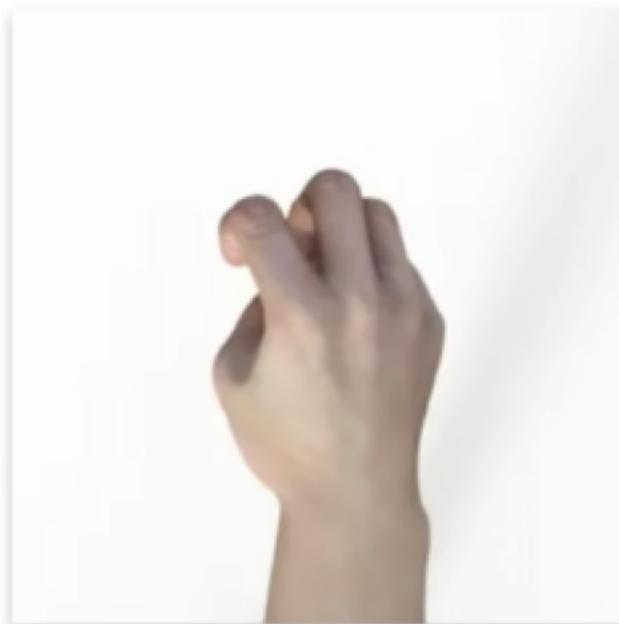
A screenshot of a Gmail inbox interface. The top navigation bar includes links for Gmail, Calendar, Documents, Reader, Web, and more. A search bar contains the query "in:spam". Below the search bar, a message titled "Spam Imperial Tortilla Sandwiches - To serve, cut each roll in half" is listed. A toolbar above the message includes options for Delete forever, Not spam, Move to, Labels, More actions, and Refresh. A link "Delete all spam messages now (messages that have been in Spam)" is visible. The left sidebar shows navigation links for Compose Mail, Inbox, Sent Mail, Drafts, and Spam (595), with the Spam link being highlighted. Under the Spam link, there are links for [Imap]/Deleted Items and [Imap]/Drafts, followed by a "15 more" link. The main content area displays a list of 15 spam messages, each with a checkbox, a star icon, and a preview. The first message is "Free Viagra Sample" with the preview "Get the real pills for free - Erectile Dysfun". The second message is "Try Viagra4Free" with the preview "Age is no longer a barrier for me in bed". The third message is "VIAGRA (c) Official Vend." with the preview "User steel.tree Brand 84% off Sale - Havi". The fourth message is "WorldWinner Player Servi." with the preview "Play Bejeweled 2 online - Compete again". The fifth message is "FTD Exclusive Offer" with the preview "Valentine's Day Roses from \$19.99 - Vale". The sixth message is "Viagra Sample" with the preview "Viagra for \$0 - Free Cialis http://theirwinter".

Source: <https://towardsdatascience.com/applied-text-classification-on-email-spam-filtering-part-1-1861e1a83246>



Data and Labels

- ในการสอนเครื่อง เราจะให้มนุษย์กำหนด **label** เฉลย สำหรับข้อมูลแต่ละชิ้น



Label = ROCK



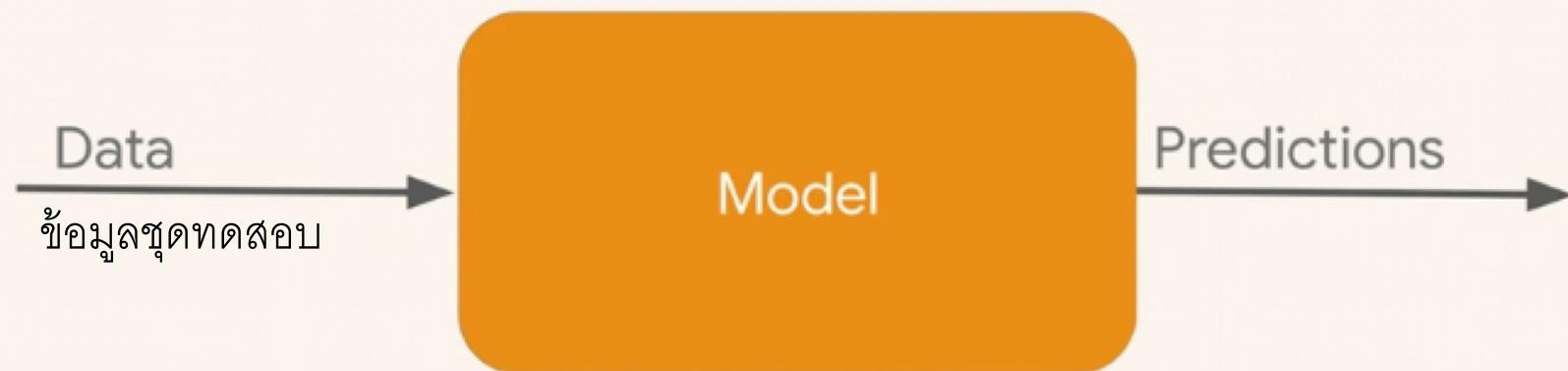
Label = PAPER



Label = SCISSORS



Training Phase



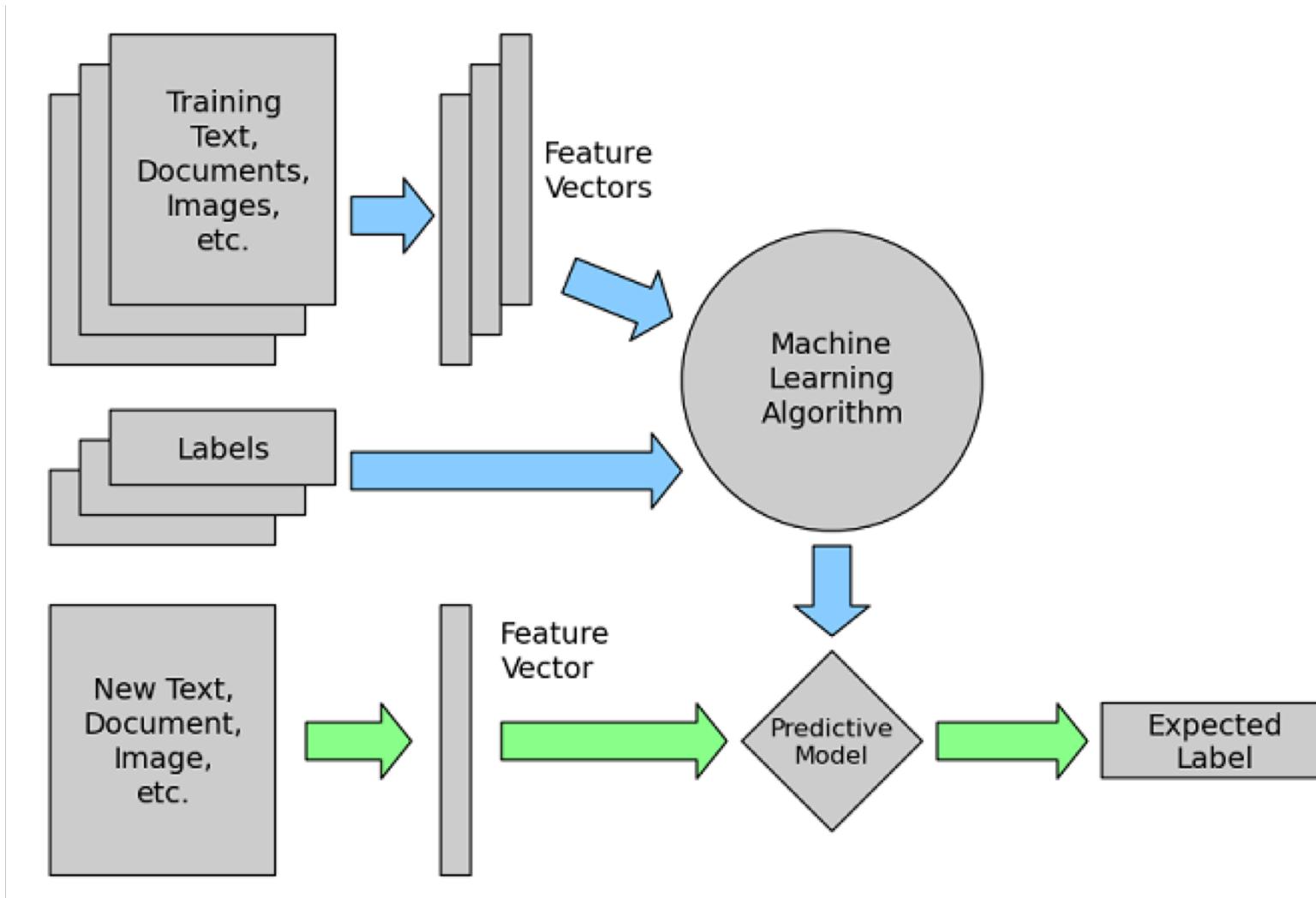
Inference Phase

เช่น Linear Regression,
SVM, Neural Networks
(*ยกเว้น k-NN ไม่มี Model*)

* inference = การอนุมาน

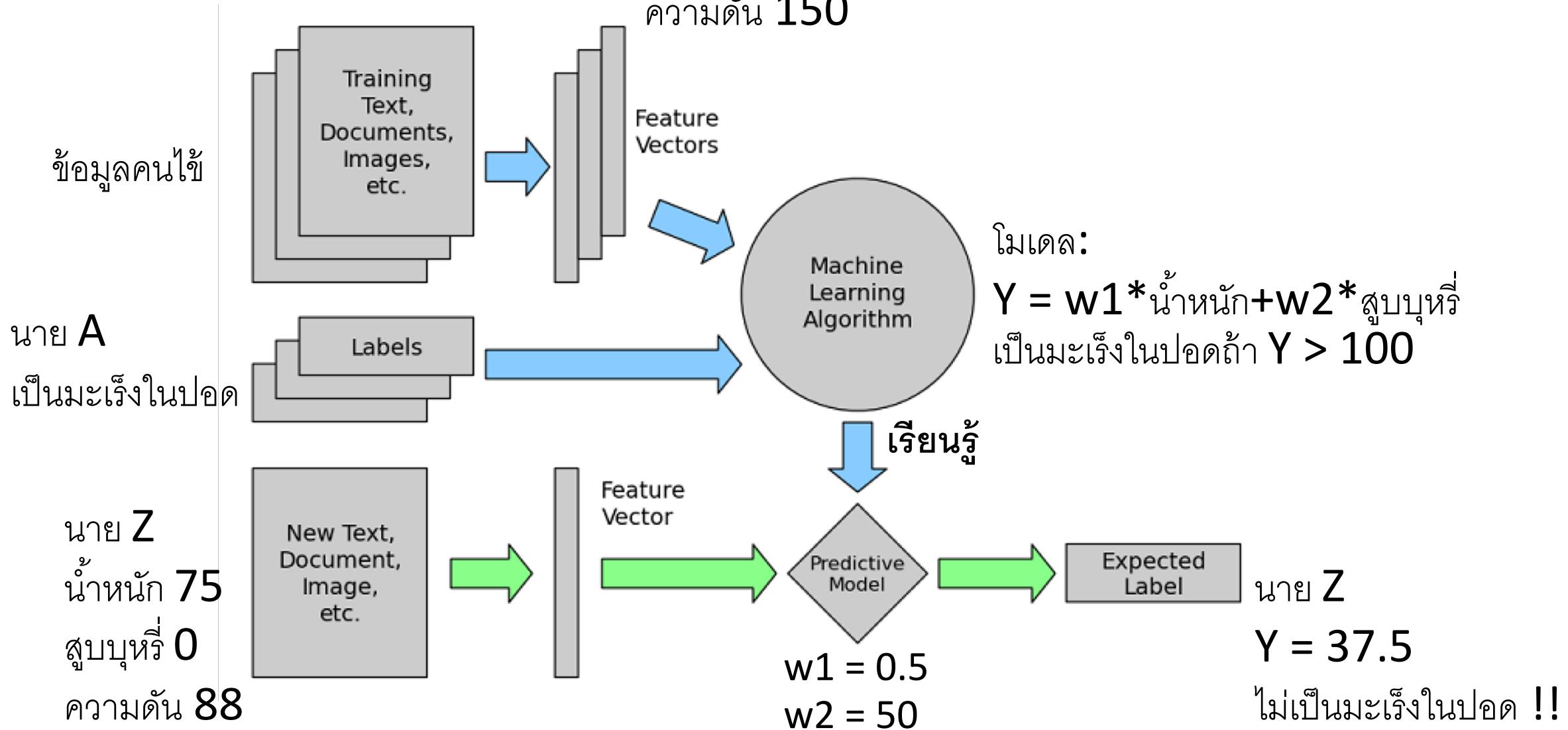
Source: Machine Learning Zero to Hero (Google I/O'19)

How ML works?



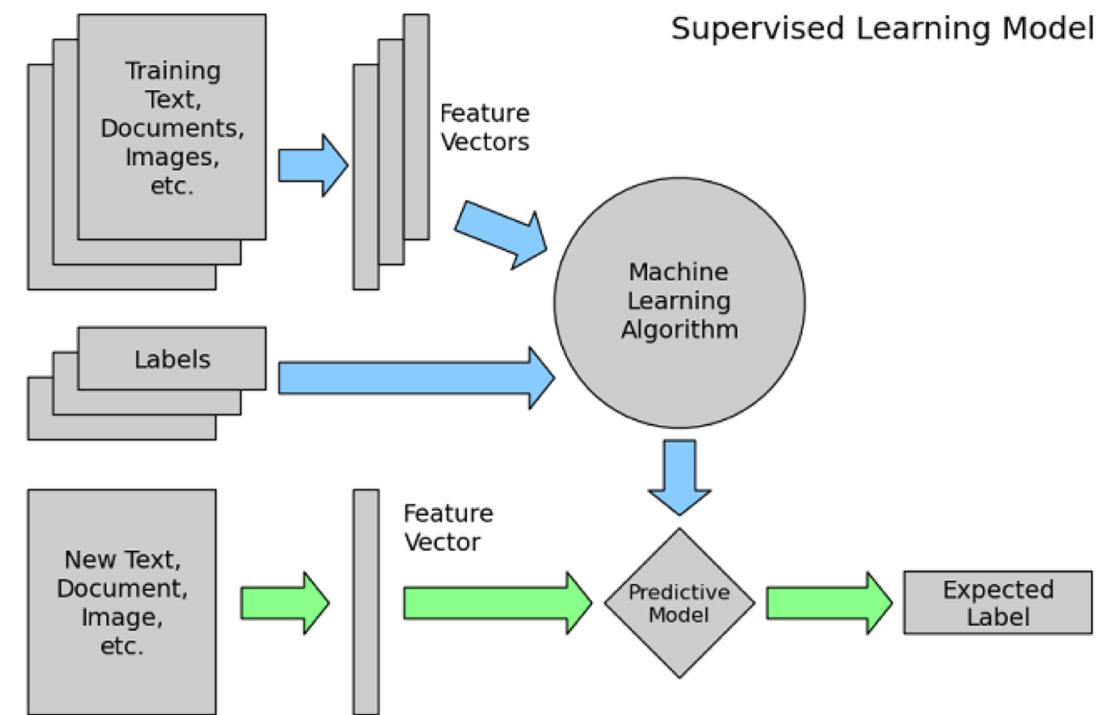
Source: https://ogrissel.github.io/scikit-learn.org/sklearn-tutorial/tutorial/text_analytics/general_concepts.html

Example



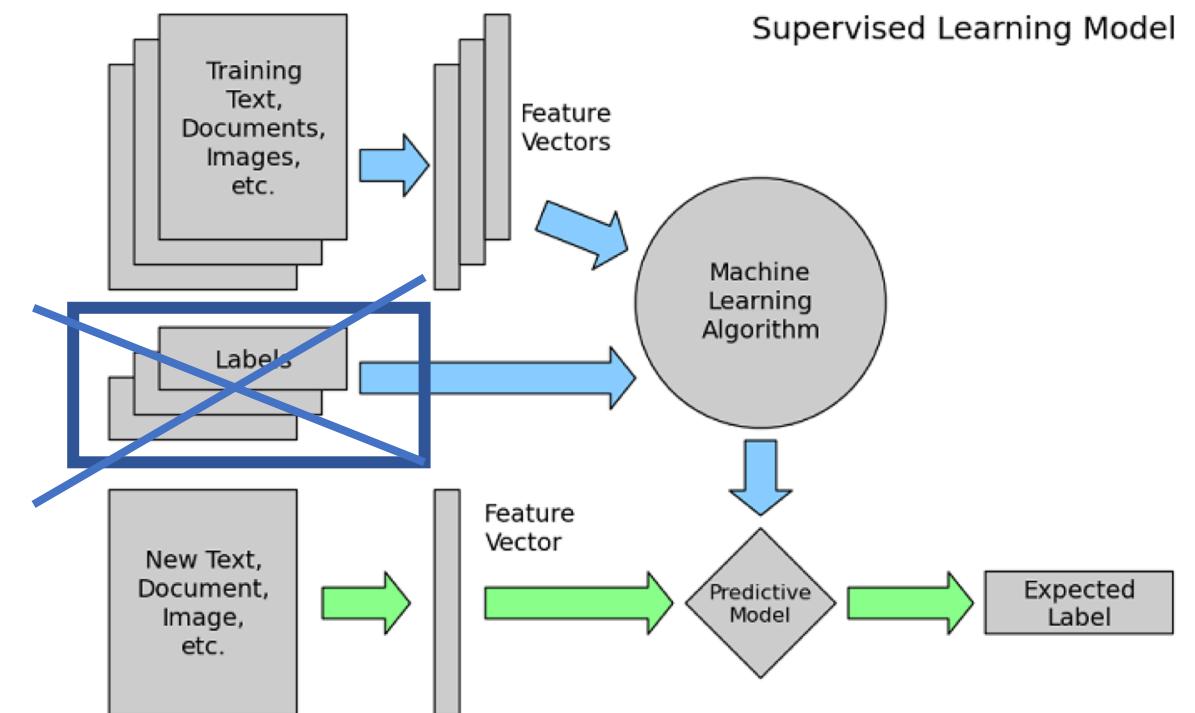
Must-know Terminology

- Training set ข้อมูลชุดสอน
- Test set ข้อมูลชุดทดสอบ
- Feature คุณลักษณะเด่น
- Target สิ่งที่จะทำนาย
- Class Label ชนิดที่จำแนก
- Model แบบจำลองคณิตศาสตร์
- Predictor ตัวทำนาย
- Classifier ตัวจำแนกชนิด
- Training Error ค่าคาดเคลื่อนในการฝึก
- Testing Error ค่าคาดเคลื่อนในการทดสอบ



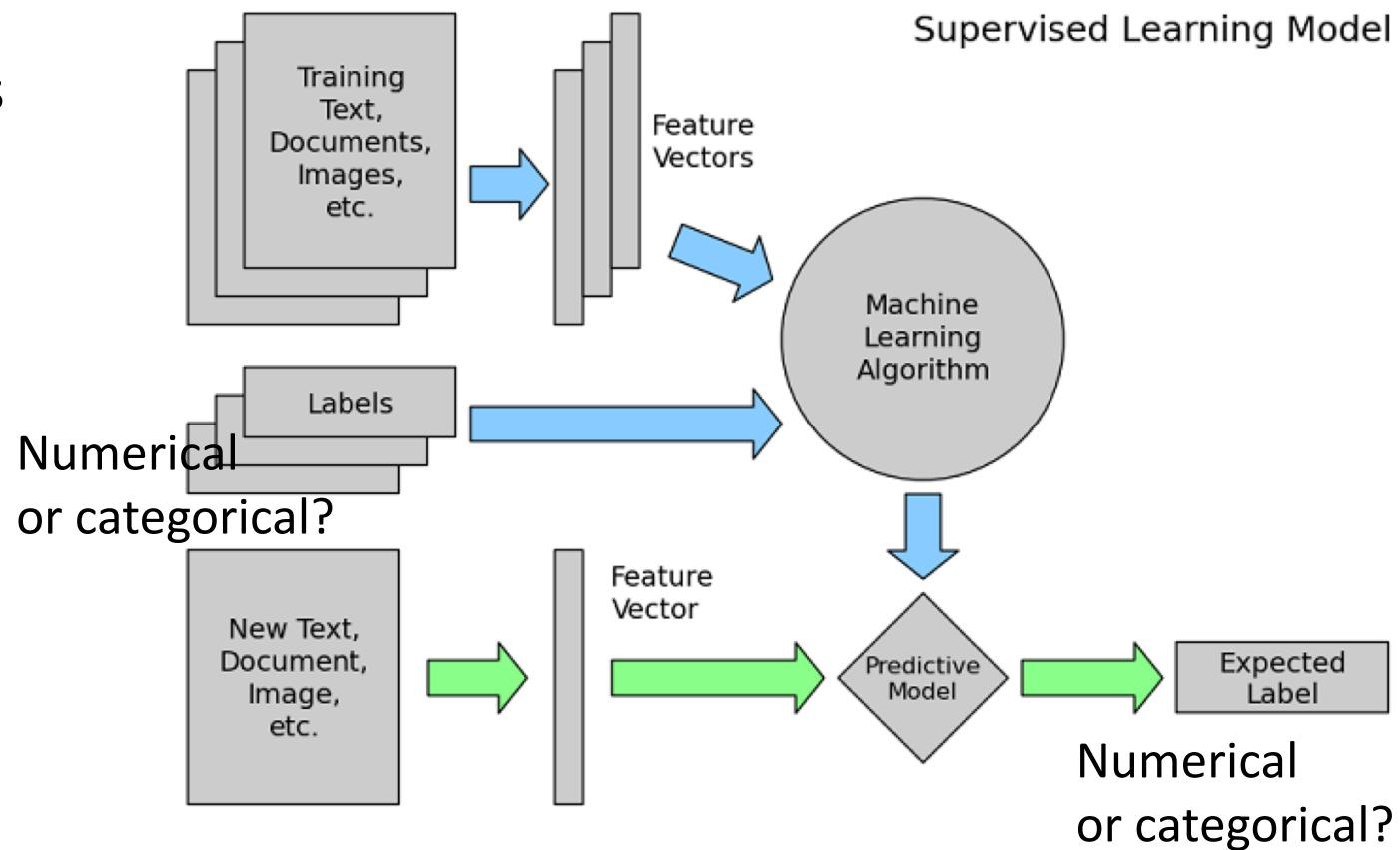
2 Types of ML Systems

- Supervised Learning
 - Trained **with** human supervision
 - Training set has **labels**
- Unsupervised Learning
 - Trained **without** human supervision
 - No class labels

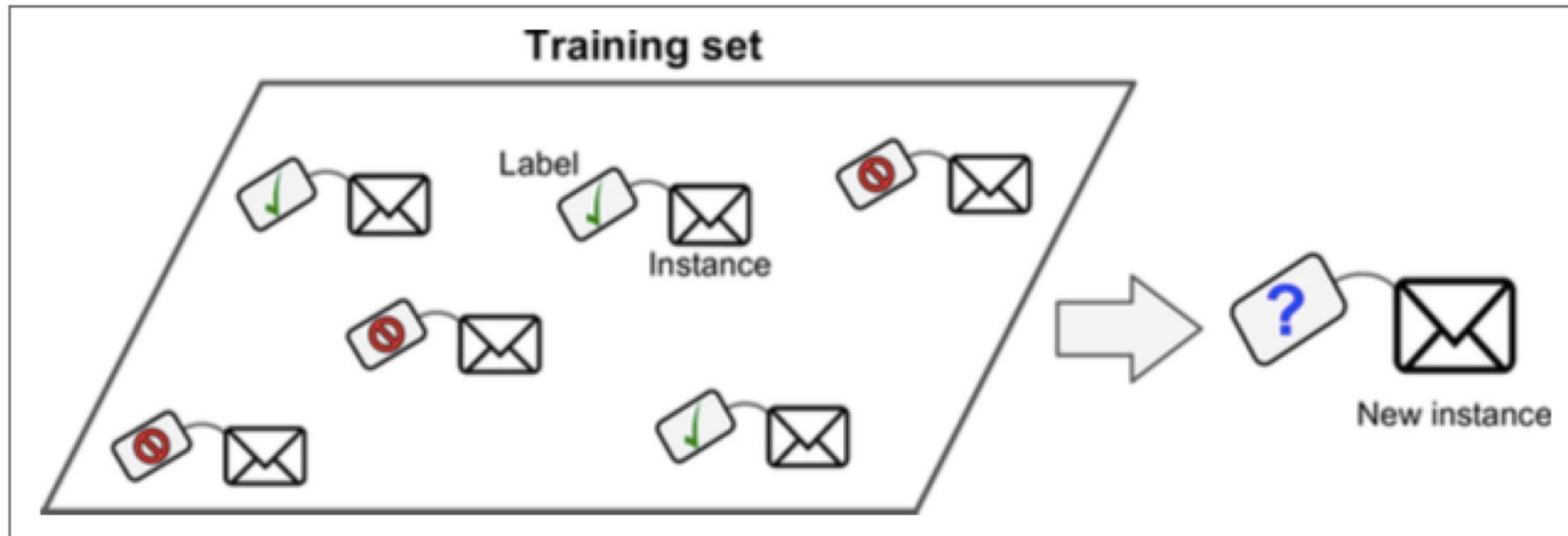


Common Supervised Learning Tasks

- Classification
 - Predicts class labels/categories
 - Example:
 - cancer/no cancer
 - husky/malamute/chiba/akita
 - 1/2/3/4/5/6/7/8/9/0
- Regression
 - Predicts continuous values
 - Example:
 - House pricing
 - Temperature



Example – Spam Filtering



Is this Classification or Regression?

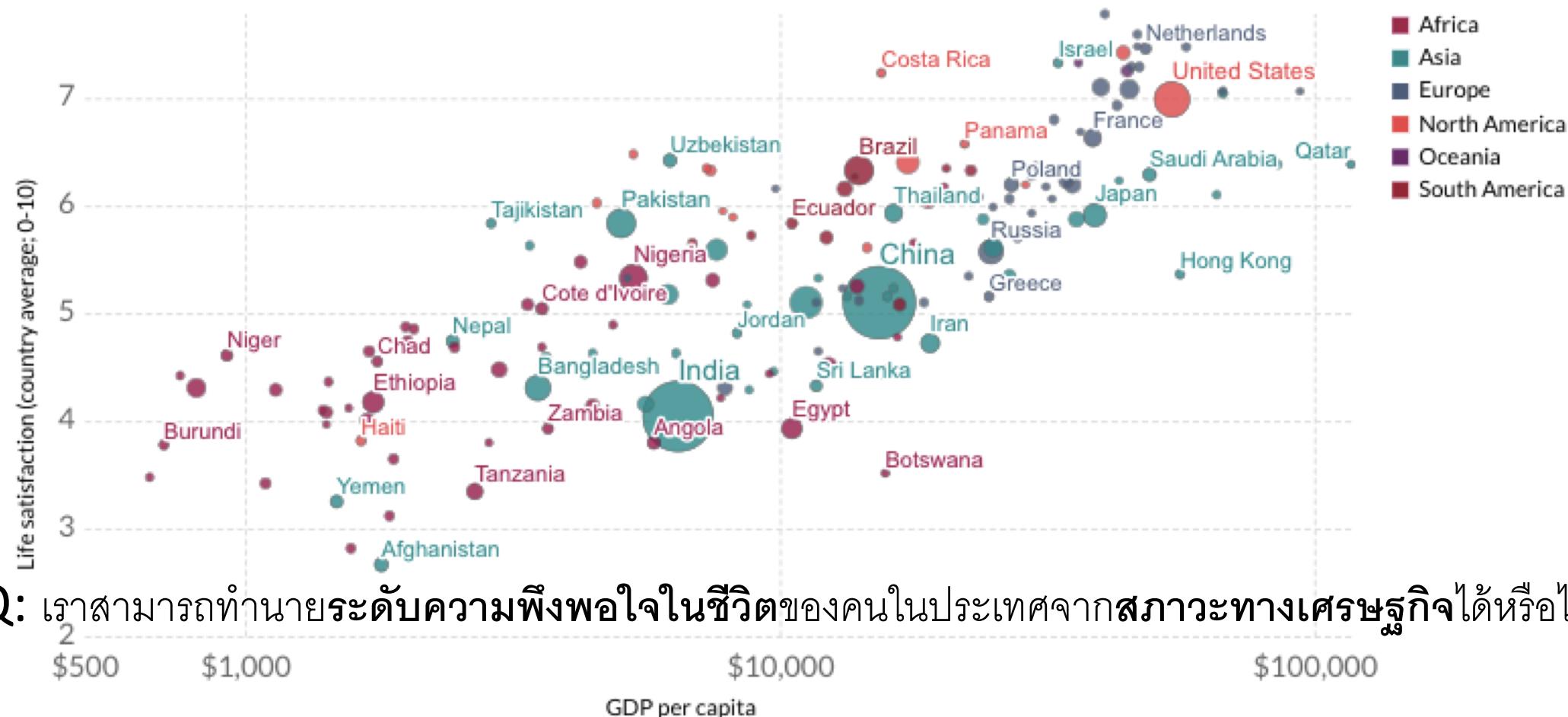
Supervised Learning Algorithms

- k-Nearest Neighbors (k-NN)
- Linear Regression
- Logistic Regression
- Support Vector Machines (SVM)
- Decision Trees
- Neural Network

ML เหนาะกับงานประเกทได

- ปัญหาที่ต้องแก้ด้วยโปรแกรมที่มีกฎและเงื่อนไขจำนวนมาก
 - e.g. rock paper scissors, handwritten digit recognition
- ปัญหาที่มีความซับซ้อนสูงและไม่สามารถหาคำตอบได้ด้วยวิธีการเดิมๆ
 - e.g. stock market prices, medical diagnosis
- ปัญหาที่ขึ้นกับสภาพแวดล้อมที่มีการเปลี่ยนแปลงอยู่ตลอดเวลา
 - e.g. spam filtering
- ปัญหาการสั่งเคราะห์สารสนเทศที่เป็นประโยชน์จากการข้อมูลจำนวนมากมหาศาล
 - e.g. gene expression data

Lab: Predicting Life Satisfaction by GDP per capita



Q: เรายสามารถทำนายระดับความพึงพอใจในชีวิตของคนในประเทศจากสภาวะทางเศรษฐกิจได้หรือไม่ ?

Lab: Predicting Life Satisfaction by GDP per capita

Q: เราสามารถทำนาย ระดับความพึงพอใจในชีวิต ของคนในประเทศจากสภาวะทางเศรษฐกิจได้หรือไม่ ?

- OECD Better Life Index (ดัชนีชี้วัดความเป็นอยู่ที่ดี)
 - จัดทำขึ้นโดย OECD (องค์การเพื่อความร่วมมือและการพัฒนาทางเศรษฐกิจ)
 - ทำ questionnaire ตามประชาชัตติในประเทศต่างๆ
 - วัด 11 มิติของความเป็นอยู่ที่ดี
 - ที่อยู่อาศัย, รายได้, การจ้างงาน, ความสัมพันธ์ทางสังคม, การศึกษา, สิ่งแวดล้อม, การบริหารจัดการของสถาบันต่างๆ, สุขภาพ, ความพึงพอใจโดยทั่วไป, ความมั่นคงปลอดภัย, และสมดุลระหว่างงานกับครอบครัว
 - ดาวน์โหลดข้อมูลได้จาก: <https://stats.oecd.org/index.aspx?DataSetCode=BLI>
 - กด Export to CSV

Lab: Predicting Life Satisfaction by GDP per capita

- OECD BLI Data (After some processing)

Indicator	Air pollution	Assault rate	Consultation on rule-making	Dwellings without basic facilities	Educational attainment	Employees working very long hours	Employment rate	Homicide rate	Household net adjusted disposable income	Household net financial wealth	... unemployment rate	Long-term rate	Personal earnings
Country													
Australia	13.0	2.1	10.5	1.1	76.0	14.02	72.0	0.8	31588.0	47657.0	...	1.08	50449.0
Austria	27.0	3.4	7.1	1.0	83.0	7.61	72.0	0.4	31173.0	49887.0	...	1.19	45199.0
Belgium	21.0	6.6	4.5	2.0	72.0	4.57	62.0	1.1	28307.0	83876.0	...	3.88	48082.0
Brazil	18.0	7.9	4.0	6.7	45.0	10.41	67.0	25.5	11664.0	6844.0	...	1.97	17177.0
Canada	15.0	1.3	10.5	0.2	89.0	3.94	72.0	1.5	29365.0	67913.0	...	0.90	46911.0
Chile	46.0	6.9	2.0	9.4	57.0	15.42	62.0	4.4	14533.0	17733.0	...	1.59	22101.0
Czech Republic	16.0	2.8	6.8	0.9	92.0	6.98	68.0	0.8	18404.0	17299.0	...	3.12	20338.0
Denmark	15.0	3.9	7.0	0.9	78.0	2.03	73.0	0.3	26491.0	44488.0	...	1.78	48347.0
Estonia	9.0	5.5	3.3	8.1	90.0	3.30	68.0	4.8	15167.0	7680.0	...	3.82	18944.0
Finland	15.0	2.4	9.0	0.6	85.0	3.58	69.0	1.4	27927.0	18761.0	...	1.73	40060.0
France	12.0	5.0	3.5	0.5	73.0	8.15	64.0	0.6	28799.0	48741.0	...	3.99	40242.0

Lab: Predicting Life Satisfaction by GDP per capita

Q: เราสามารถทำนายระดับความพึงพอใจในชีวิตของคนในประเทศจาก สภาวะทางเศรษฐกิจ ได้หรือไม่ ?

- **GDP (Gross Domestic Product)**

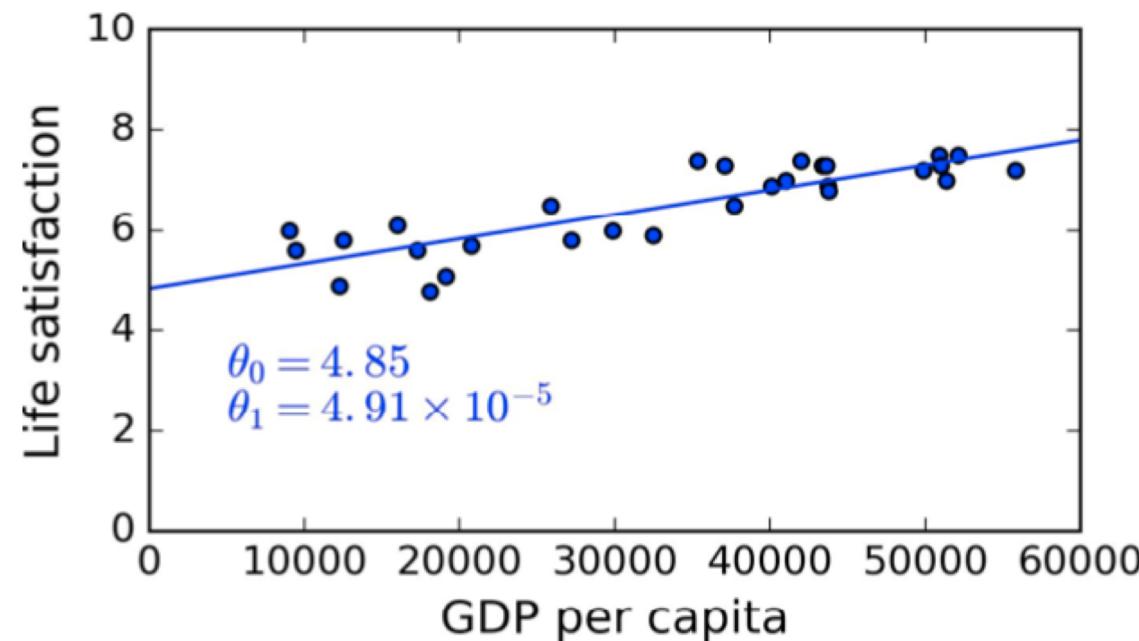
- เป็นตัวชี้วัดอย่างหนึ่งต่อสภาวะทางเศรษฐกิจของประเทศ
- $\text{GDP per capita} = \text{GDP} \text{ ต่อหัว}$
- ดาวน์โหลดข้อมูลได้จาก [IMF: http://goo.gl/j1mWfE](http://goo.gl/j1mWfE)

Country	Subject Descriptor	Units	Scale	Country/Series-specific Notes	2015
Afghanistan	Gross domestic product per capita, current prices	U.S. dollars	Units	i	599.994
Albania	Gross domestic product per capita, current prices	U.S. dollars	Units	i	3,995.383
Algeria	Gross domestic product per capita, current prices	U.S. dollars	Units	i	4,318.135
Angola	Gross domestic product per capita, current prices	U.S. dollars	Units	i	4,100.315
Antigua and Barbuda	Gross domestic product per capita, current prices	U.S. dollars	Units	i	14,414.302
Argentina	Gross domestic product per capita, current prices	U.S. dollars	Units	i	13,588.846

Lab: Predicting Life Satisfaction by GDP per capita

- ใช้ Scikit-Learn ในการสร้างโมเดลเพื่อทำนาย Life Satisfaction จาก GDP per capita
 - `sklearn.linear_model.LinearRegression()`

$$life_satisfaction = \theta_0 + \theta_1 \times GDP_per_capita$$



1. นำเข้าข้อมูลจากไฟล์ .csv ด้วย Pandas

```
import pandas as pd
```

```
lifesat = pd.read_csv('lifesat_gdp.csv')
```

ลองดูข้อมูลด้วยคำสั่ง

- lifesat
- lifesat.info()
- lifesat.describe()
- lifesat ["GDP per capita"]
- lifesat.plot(kind='scatter', x="GDP per capita", y='Life satisfaction')

2. สร้างตัวแปร X, y เพื่อเก็บ feature, target

```
X = np.c_[lifesat["GDP per capita"]]      # feature  
y = np.c_[lifesat["Life satisfaction"]]    # target
```

คำสั่ง `np.c_[]` จะแปลงข้อมูลให้อยู่ในรูป `numpy array` ขนาด $n \times 1$ (colums)

3. ทำการสร้างและ train model

```
# Select a linear model
```

```
model = sklearn.linear_model.LinearRegression()
```

```
# Train the model
```

```
model.fit(X, y)
```

3. นำ model ที่ผ่านการ train แล้วไปใช้ทำนายค่า

```
# Make a prediction for Cyprus
```

```
X_new = [[22587]] # Cyprus' GDP per capita
```

```
print(model.predict(X_new))
```

4. วิเคราะห์ผลและการทำนาย

```
country_stats.plot(kind='scatter', x="GDP per capita", y='Life satisfaction')
```

```
plt.plot(X, model.predict(X), color='k')
```

Next week

- Supervised Learning Algorithm
 - K-NN