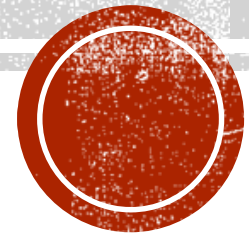


# **INTRODUCTION DATA SCIENCE & BIG DATA**

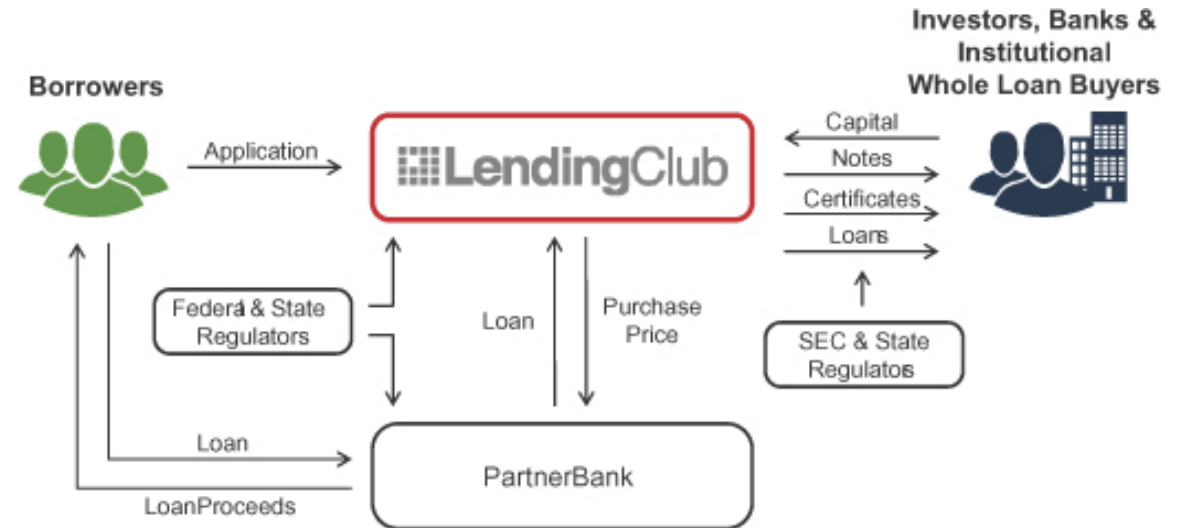
Sonpvh



# OUTLIER

1. Introduction: Data Science Applications
2. History
3. Data science
4. Outlier of course

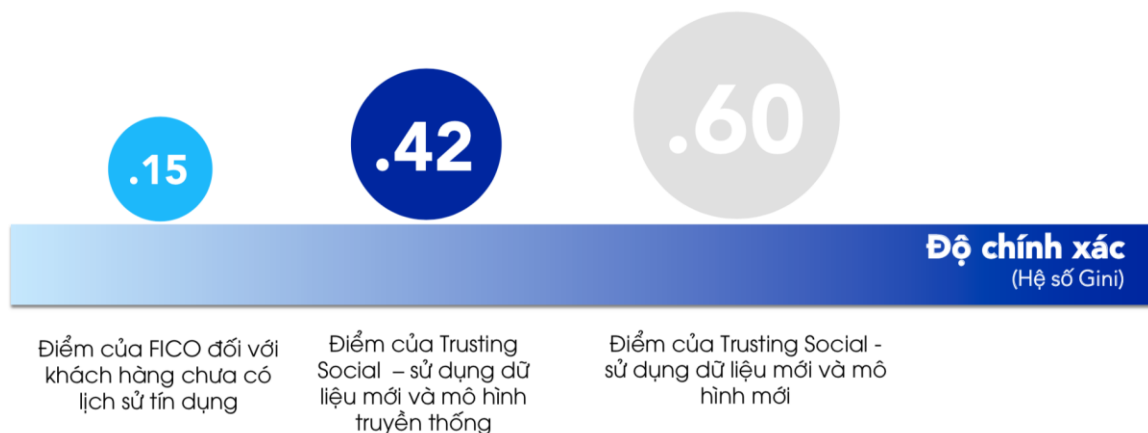
# 1. DATA SCIENCE APPLICATIONS



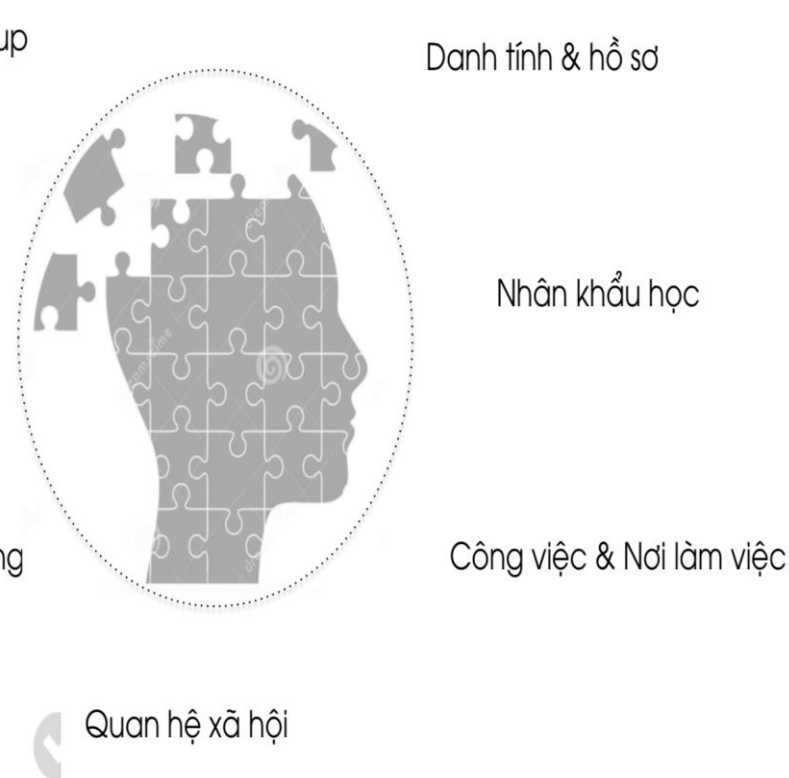
- President: Scott Sanborn
- Founded: 2006
- Valuing the company: 8.5 bn

# 1. DATA SCIENCE APPLICATIONS

trustingsocial



Nguyen Nguyen  
CEO





# 1. DATA SCIENCE APPLICATIONS



Everything is a Recommendation



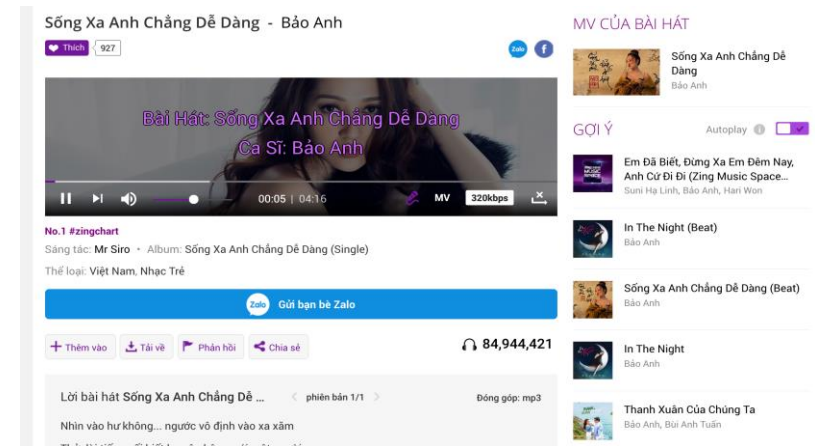
Over 80% of what people watch comes from our recommendations

Recommendations are driven by Machine Learning

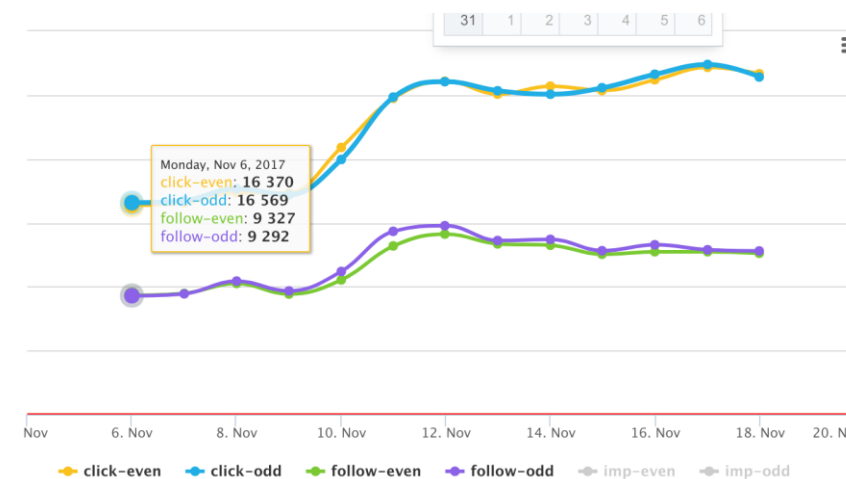
NETFLIX

# 1. DATA SCIENCE APPLICATIONS

- Netflix: 2/3 of the movies watched are recommended
- Google News: recommendations generate 38% more clickthrough
- Amazon: 35% sales from recommendations
- Choicestream: 28% of the people would buy more music if they found what they liked.



ZingMp3: >30% traffic



ZOA: improve >30% total click and follow

[4] Xavier 2014

# THE AGE OF DISCOVERY

- **Chris Anderson in “The Long Tail”**
  - *“We are leaving the age of information and entering the age of [recommendation](#)”*
- **CNN Money, “The race to create a 'smart' Google”:**
  - *“The Web, they say, is leaving the era of search and entering one of [discovery](#). What's the difference? Search is what you do when you're looking for something. Discovery is when something wonderful that you didn't know existed, or didn't know how to ask for, finds you.”*

# THE PERSONAL EXPERIENCES

## MAN

From people: friends, co-workers, family, acquaintances, anything person to person.

Number of Recommendations

117

TOP TEN ALBUM RECOMMENDATIONS



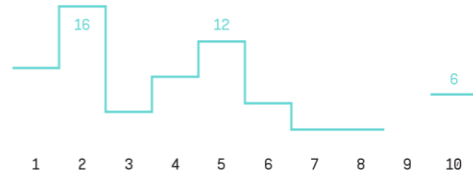
## MEDIA

From media sources: social media, music blogs, musicians, live shows, TV, movies.

Number of Recommendations

785

TOP TEN ALBUM RECOMMENDATIONS



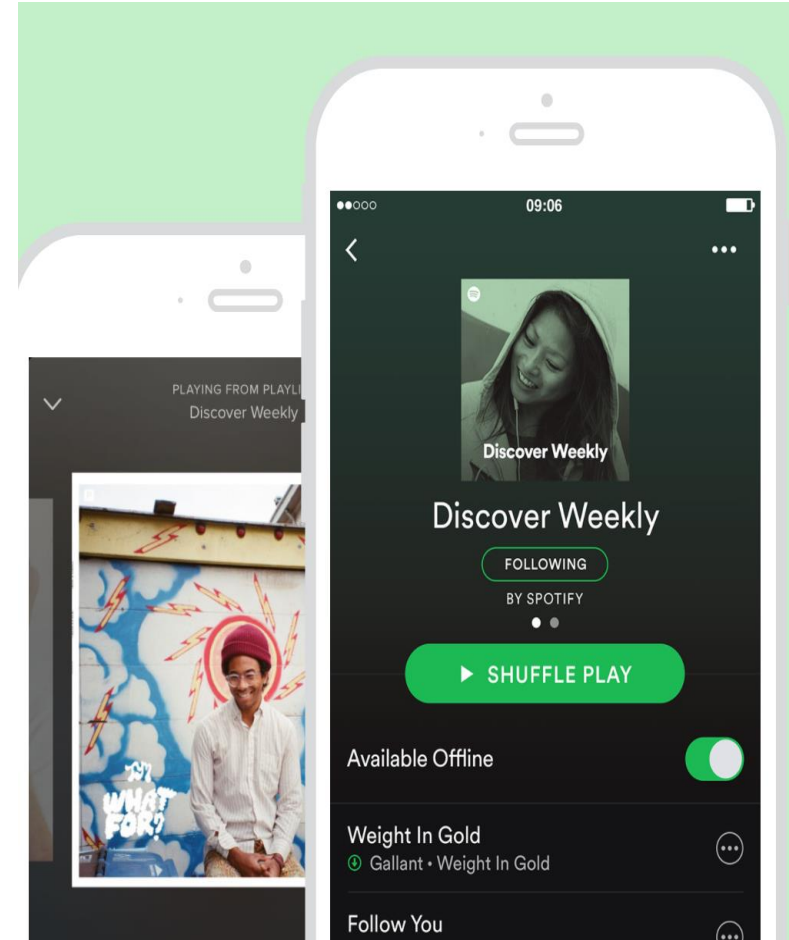
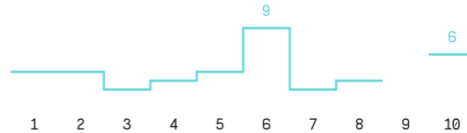
## MACHINE

From algorithms: personalized playlists on Spotify like Release Radar.

Number of Recommendations

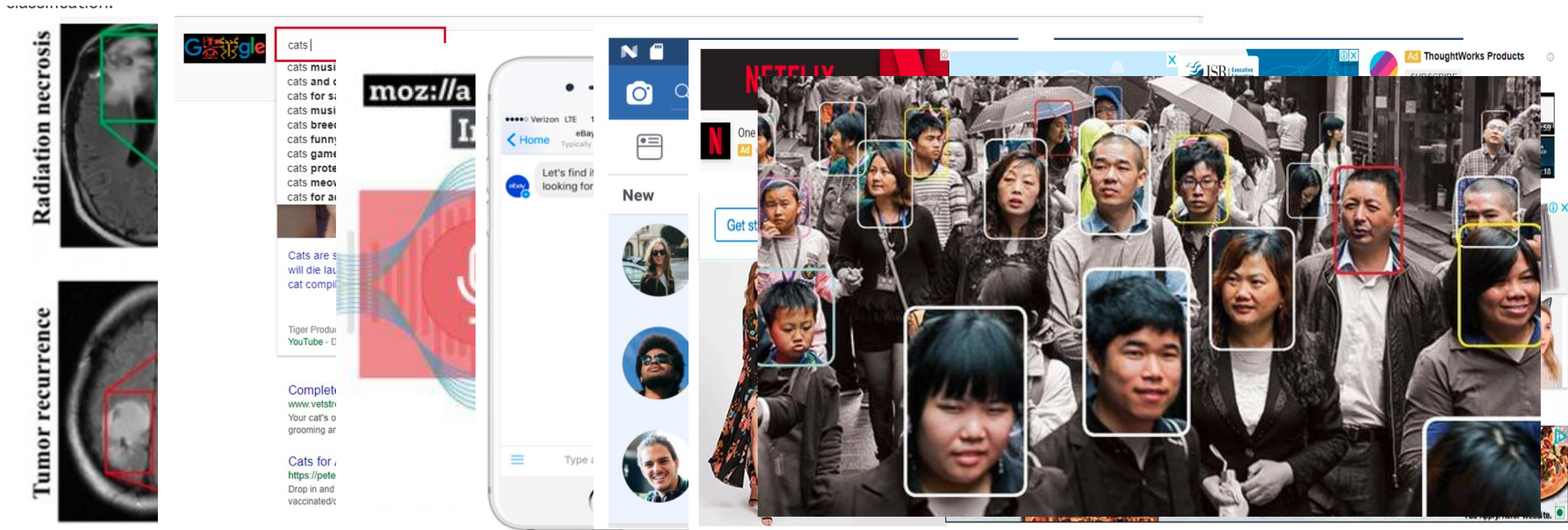
1,363\*

TOP TEN ALBUM RECOMMENDATIONS












# 1. DATA SCIENCE APPLICATIONS

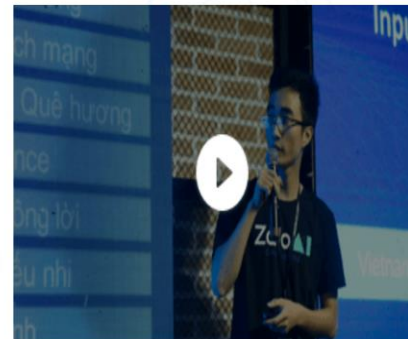


# 1. DATA SCIENCE APPLICATIONS

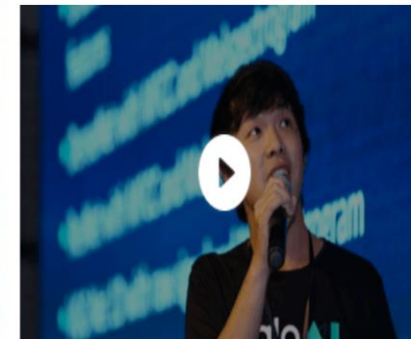
	<b>Handwritten Digit Recognizer</b> Learn computer vision fundamentals with the famous MNIST data <a href="#">Getting Started</a> · Ongoing · 📁 tabular data, image data, multiclass classification, object identification	Knowledge 2,577 teams
	<b>Titanic: Machine Learning from Disaster</b> Start here! Predict survival on the Titanic and get familiar with ML basics <a href="#">Getting Started</a> · Ongoing · 📁 tutorial, tabular data, binary classification	Knowledge 10,433 teams
	<b>House Prices: Advanced Regression Techniques</b> Predict sales prices and practice feature engineering, RFs, and gradient boosting <a href="#">Getting Started</a> · Ongoing · 📁 tabular data, regression	Knowledge 4,322 teams
	<b>ImageNet Object Localization Challenge</b> Identify the objects in images <a href="#">Research</a> · 11 years to go · 📁 image data, object detection	Knowledge 36 teams
	<b>Predict Future Sales</b> Final project for "How to win a data science competition" Coursera course <a href="#">Playground</a> · 9 months to go	Kudos 2,630 teams
	<b>iNaturalist 2019 at FGVC6</b> Fine-grained classification spanning a thousand species <a href="#">Research</a> · 2 months to go	Kudos 1 team
	<b>iWildCam 2019 - FGVC6</b> Categorize animals in the wild <a href="#">Playground</a> · 2 months to go · 📁 image data, multiclass classification	Kudos 27 teams



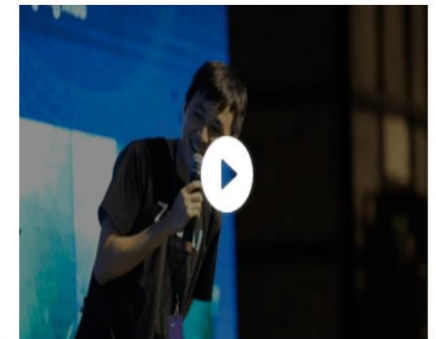
Music Genre Classification  
[Nguyen Ba Dung](#)



Voice Gender Classification  
[Team VietAI](#)



Landmark Identification  
[Team Phoenixxx](#)



# 1. DATA SCIENCE APPLICATIONS

“Data scientist is  
the sexiest job  
of the 21st century.”

Harvard Business Review





## 2. HISTORY - STATISTICS



- 1763 – Thomas Bayes – English statistician

$$P(A | B) = \frac{P(B | A) P(A)}{P(B)}$$

Bayes theorem

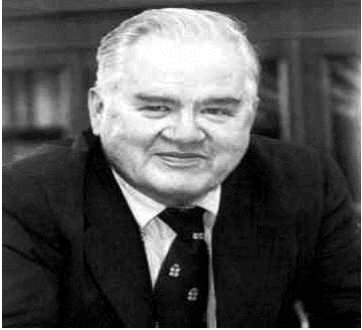


- 1763 – Carl Friedrich Gauss (1809) (1821) & Legendre (1805)

Regression – Method of least squares – predict the movement of planet

[10] – regression analysis

## 2. HISTORY - STATISTIC



- 1962 - John W. Tukey – US mathematician  
“The Future of data analytics” - “I have come to feel that my central interest is in **data analysis**... **Data analysis**, and the parts of statistics ...”



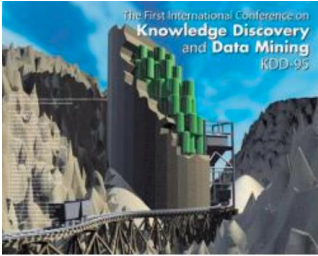
- 1976 - Peter Naur – Danish Computer Scientist  
“Datalogy, the science of data and of data processes and its place in education” -  
“**Data Science** - The science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences.”



- 1977 The International Association for Statistical Computing  
“It is the mission of the IASC to **link** traditional **statistical methodology**, **modern computer technology**, and **the knowledge of domain** experts in order to convert data into information and knowledge.”



# 2. HISTORY - STATISTICS



- 1989 – KDD - SIGKDD Conference on Knowledge Discovery and Data Mining  
First conference about data mining

- 1994 – Business week “Databased Marketing”

Companies are **collecting mountains of information about you**, crunching it to **predict how likely you are to buy a product**, and using that knowledge to **craft a marketing message precisely calibrated** to get you to do so...

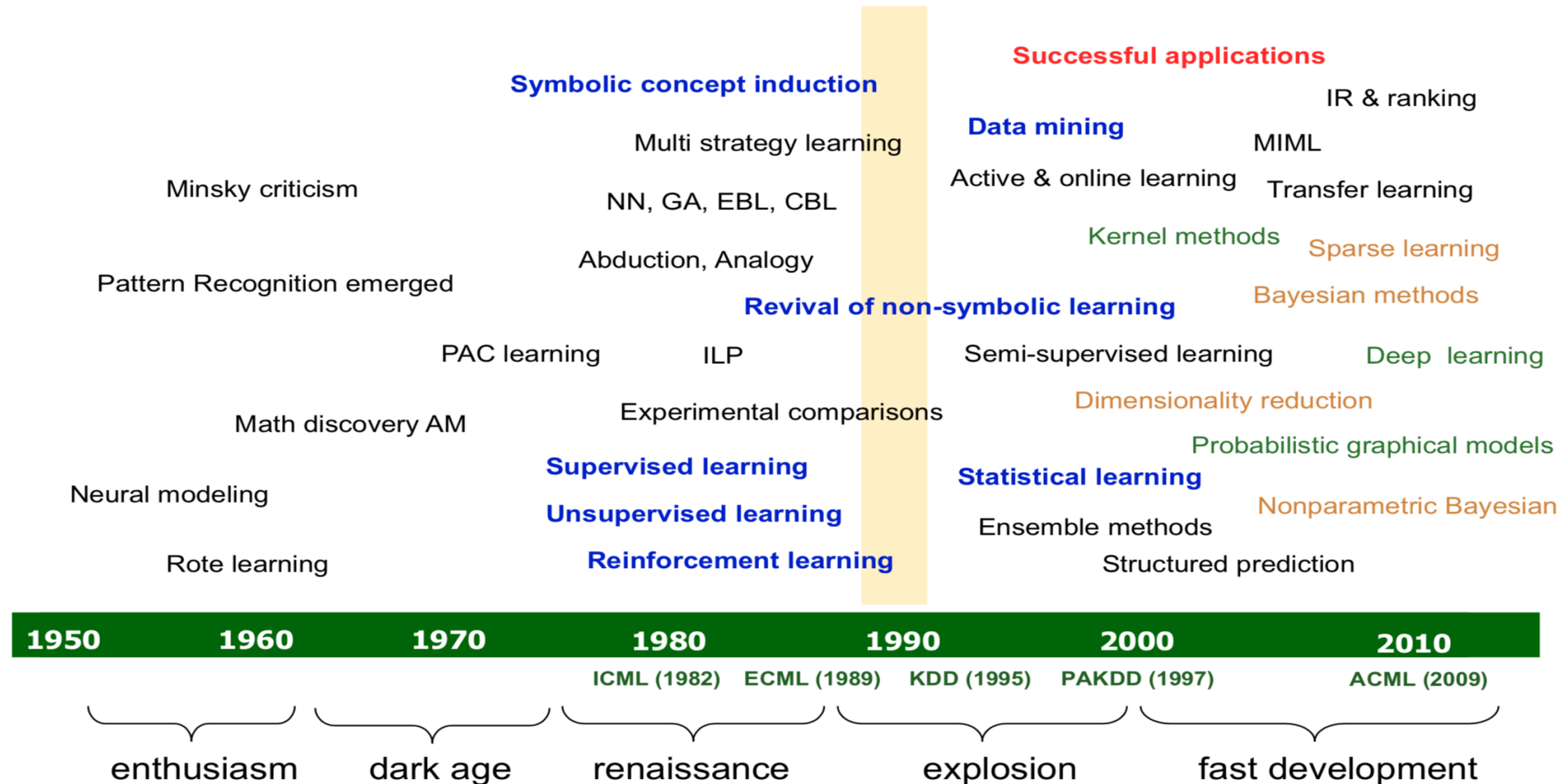


- 1997 – Professor C. F. Jeff Wu - University of Michigan  
calls for **statistics** to be renamed **data science** and **statisticians** to be renamed **data scientists**.



- 1999 - Prof. Moshe Zviran  
“ Conventional statistical methods work well with small data sets. Today's databases, however, can **involve millions of rows and scores of columns of data** ... “

# 2. HISTORY — DATA MINING

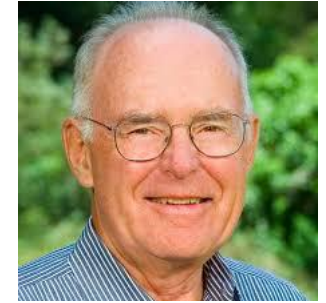


# 2. THE HISTORY - COMPUTATION

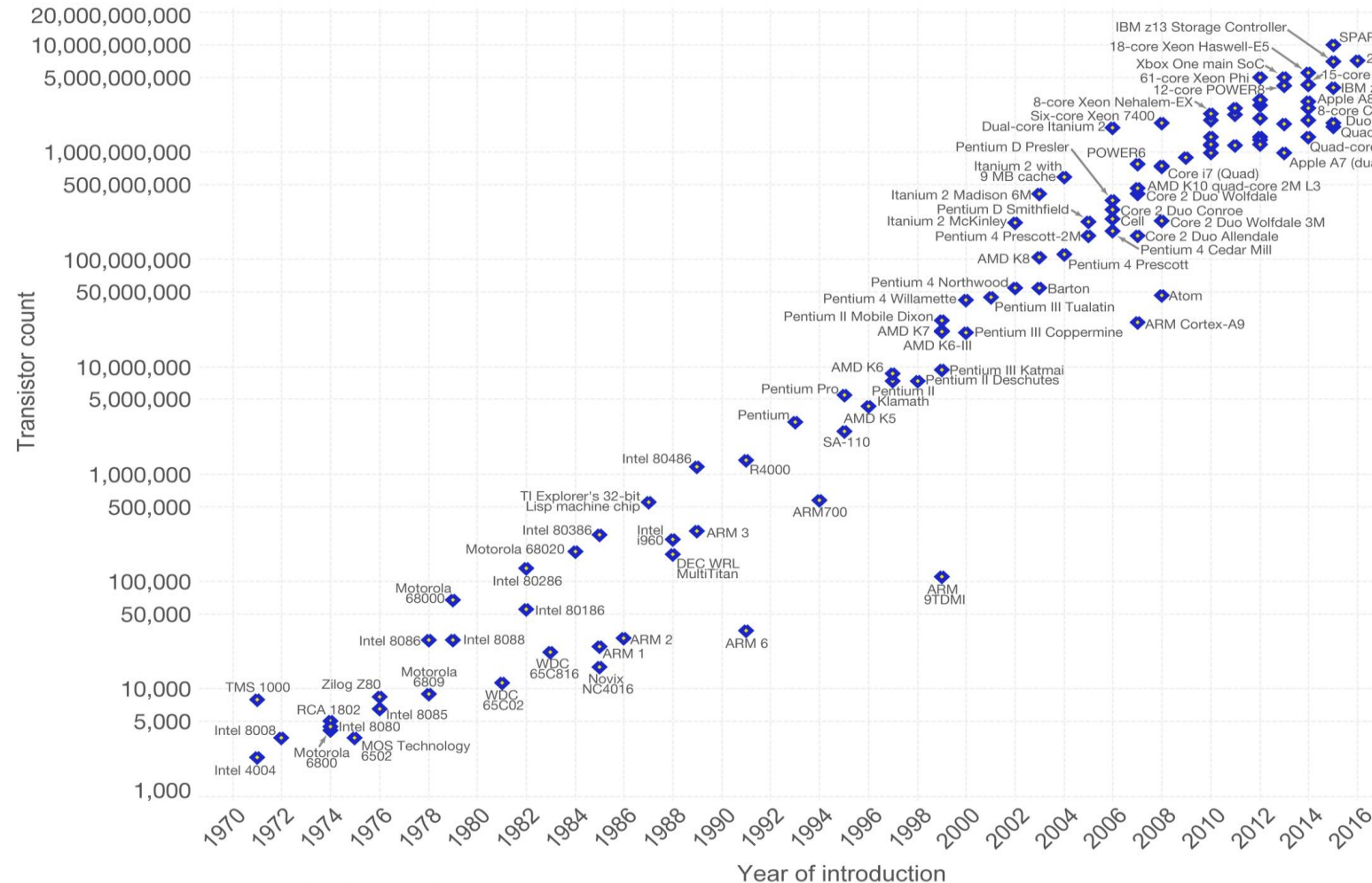
## Moore's Law – The number of transistors on integrated circuit chips (1971-2016)

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important as other aspects of technological progress – such as processing speed or the price of electronic products – are strongly linked to Moore's law.

Our World  
in Data



Gordon Earle Moore  
US Businessman



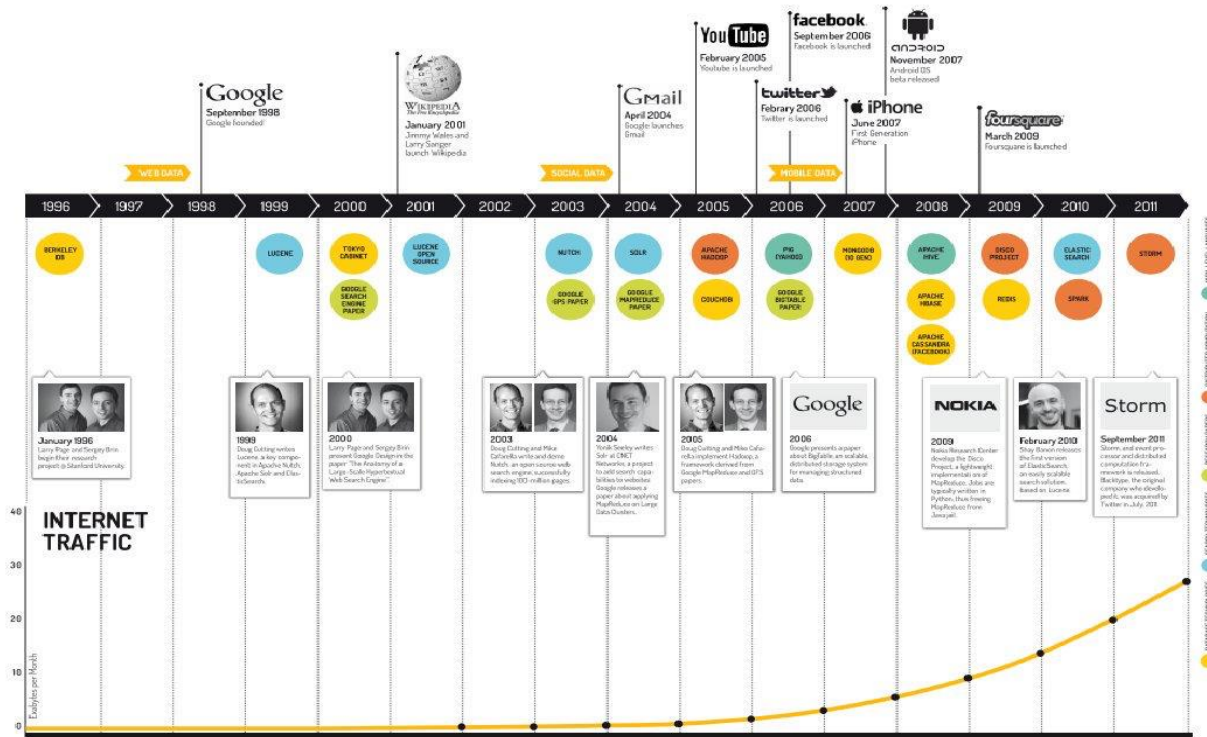
Data source: Wikipedia ([https://en.wikipedia.org/wiki/Transistor\\_count](https://en.wikipedia.org/wiki/Transistor_count))

The data visualization is available at [OurWorldinData.org](https://www.ourworldindata.org). There you find more visualizations and research on this topic.

Licensed under CC-BY-SA by the author Max Roser.

# 2. HISTORY – BIG DATA

## BIG DATA A BRIEF HISTORY



90% OF THE AVAILABLE DATA HAS BEEN  
CREATED IN THE LAST TWO YEARS





## 2. HISTORY – BIG DATA

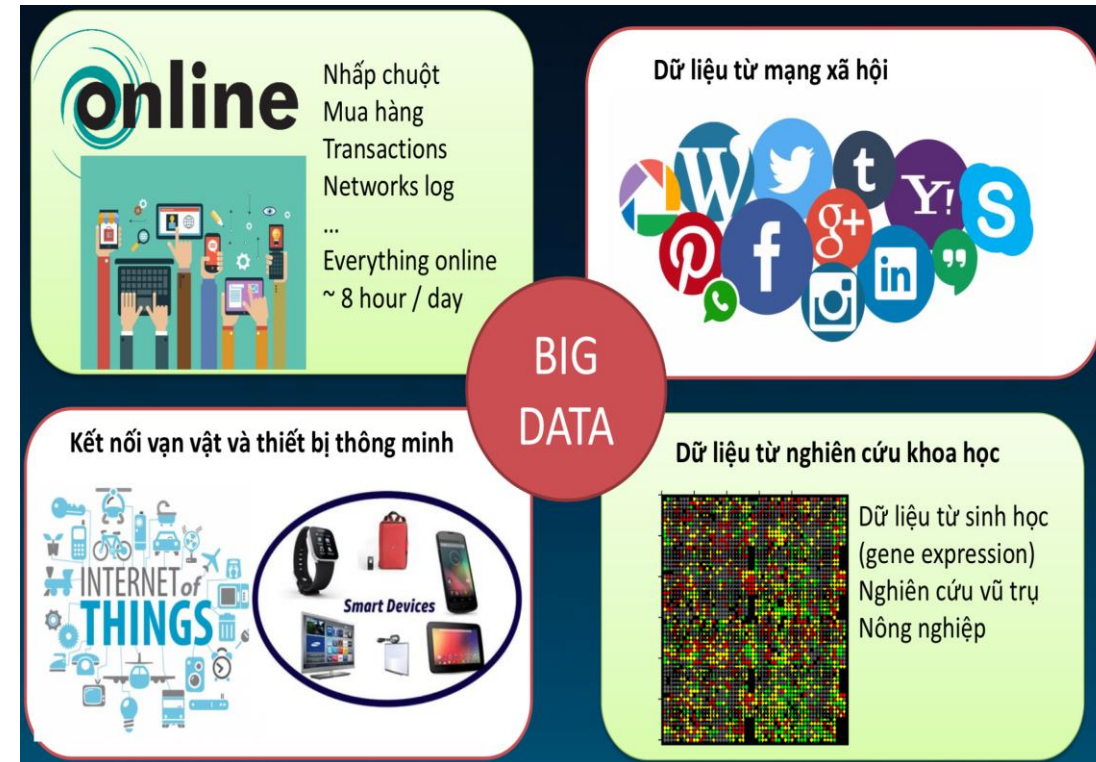


“The average person today processes more data in a single day than a person in the 1500’s did in an entire life time”

Src: [12]

### Every 60 seconds

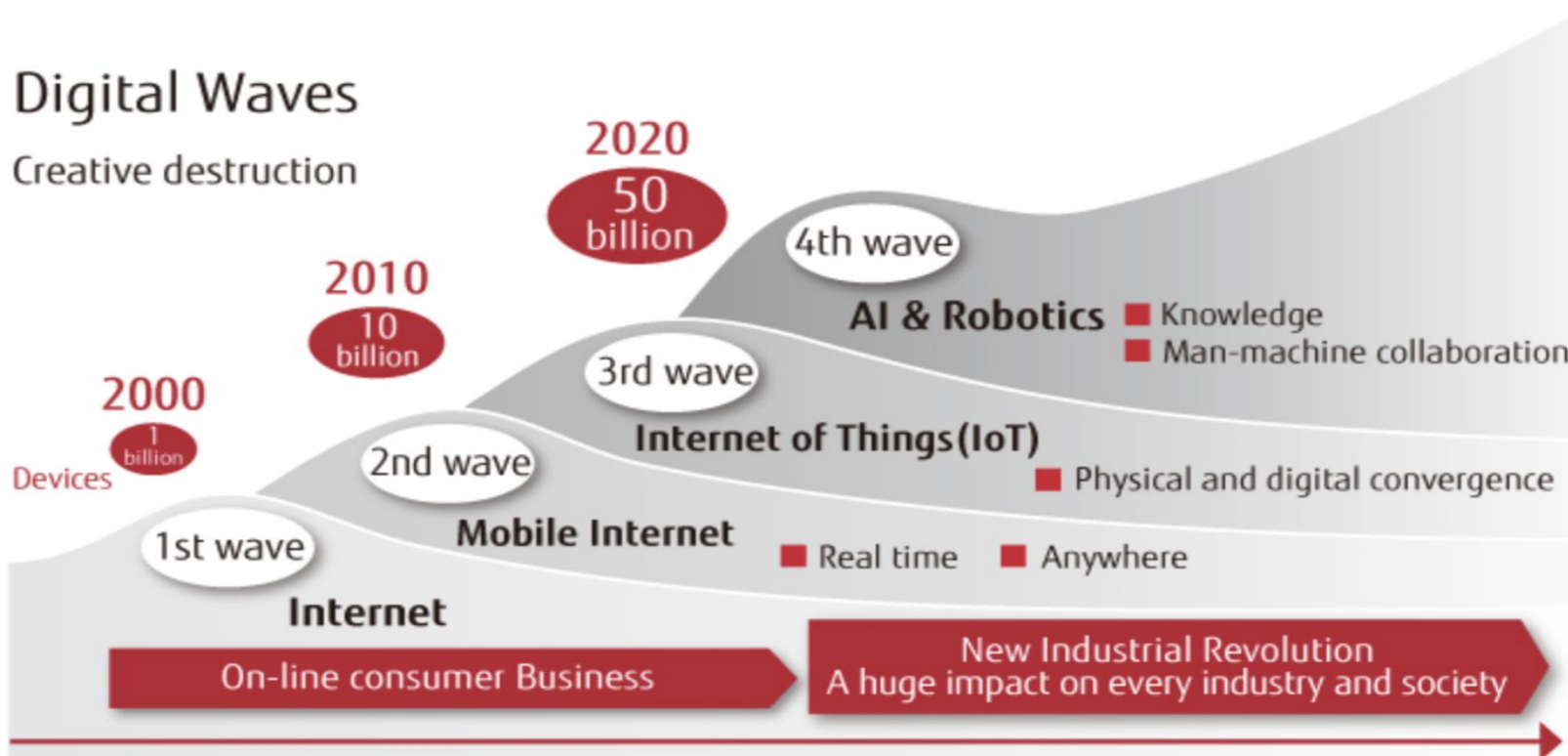
- 98,000+ tweets
- 695,000 status updates
- 11 million instant messages
- 698,445 Google searches
- 168 million+ emails sent
- 1,820TB of data created
- 217 new mobile web users



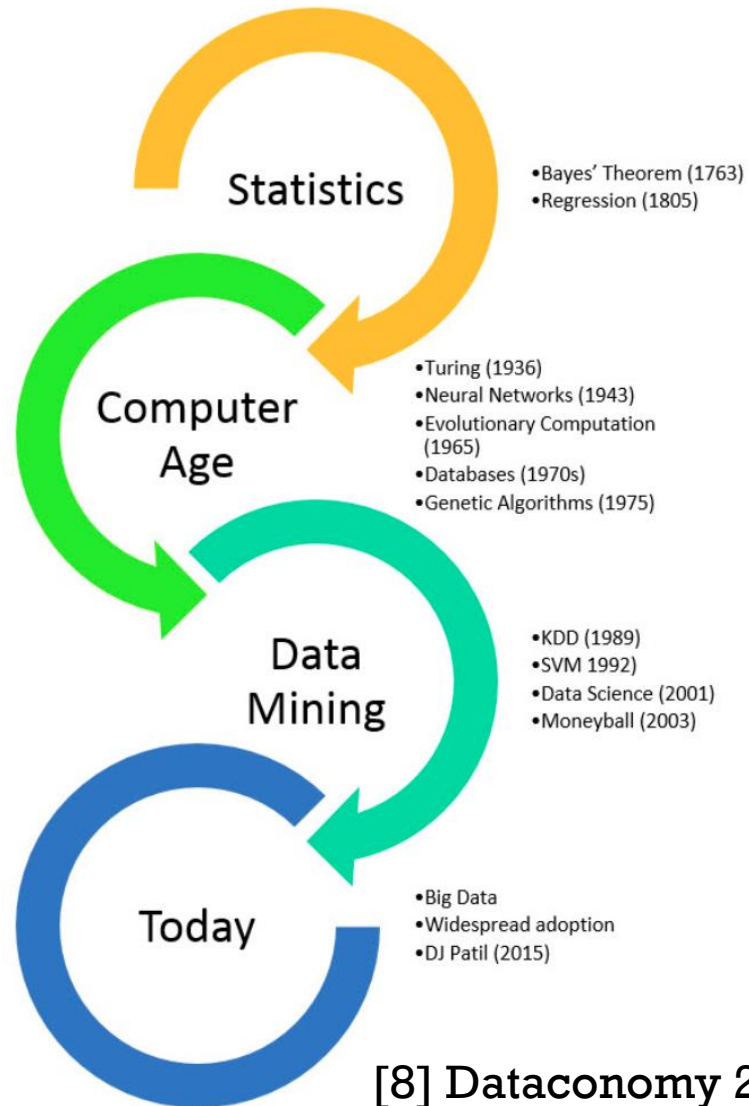
Src: [13]



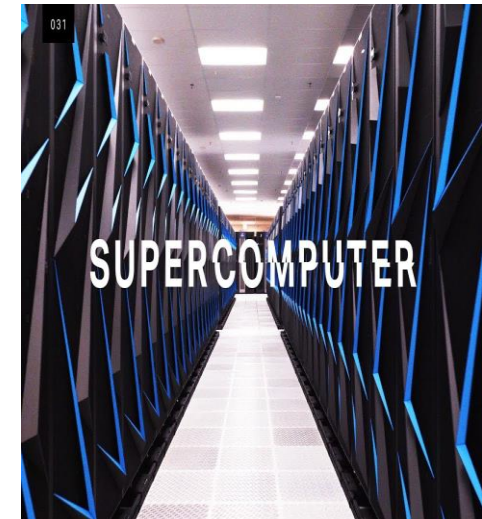
## 2. HISTORY— APPLICATION - DIGITAL WAVES



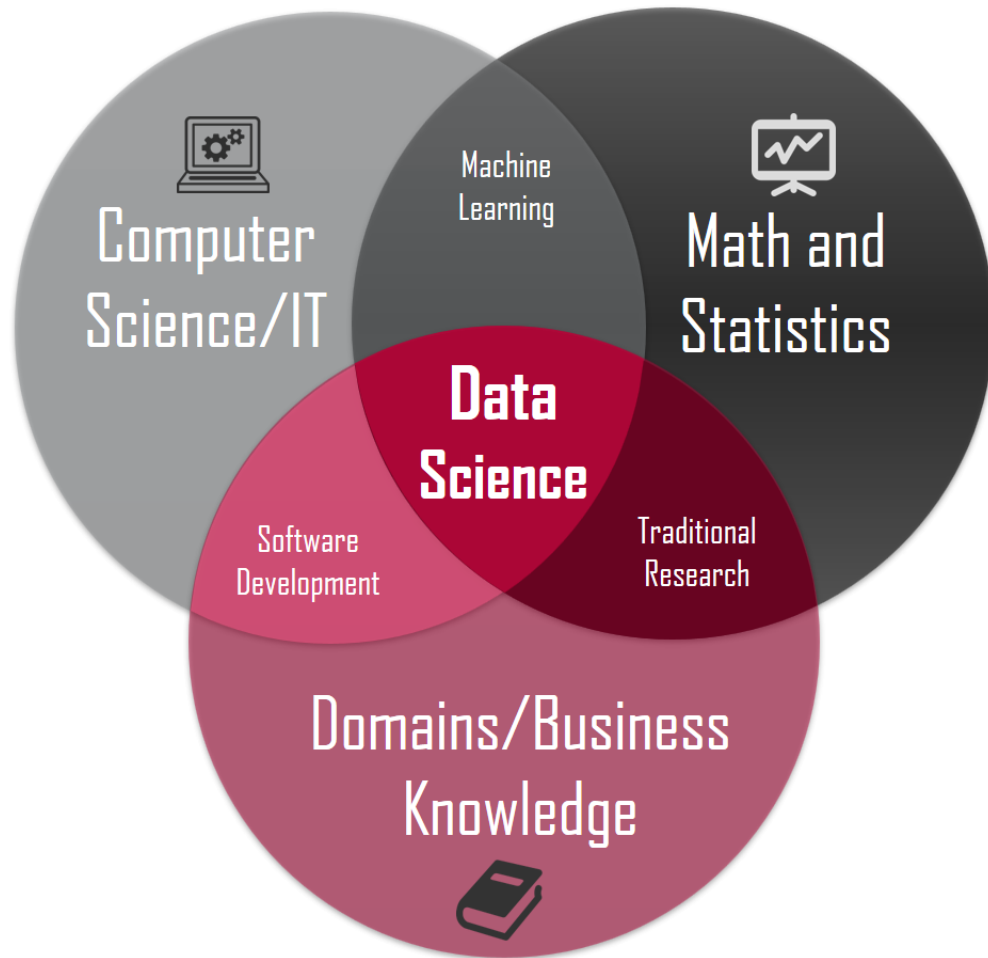
# 2. HISTORY



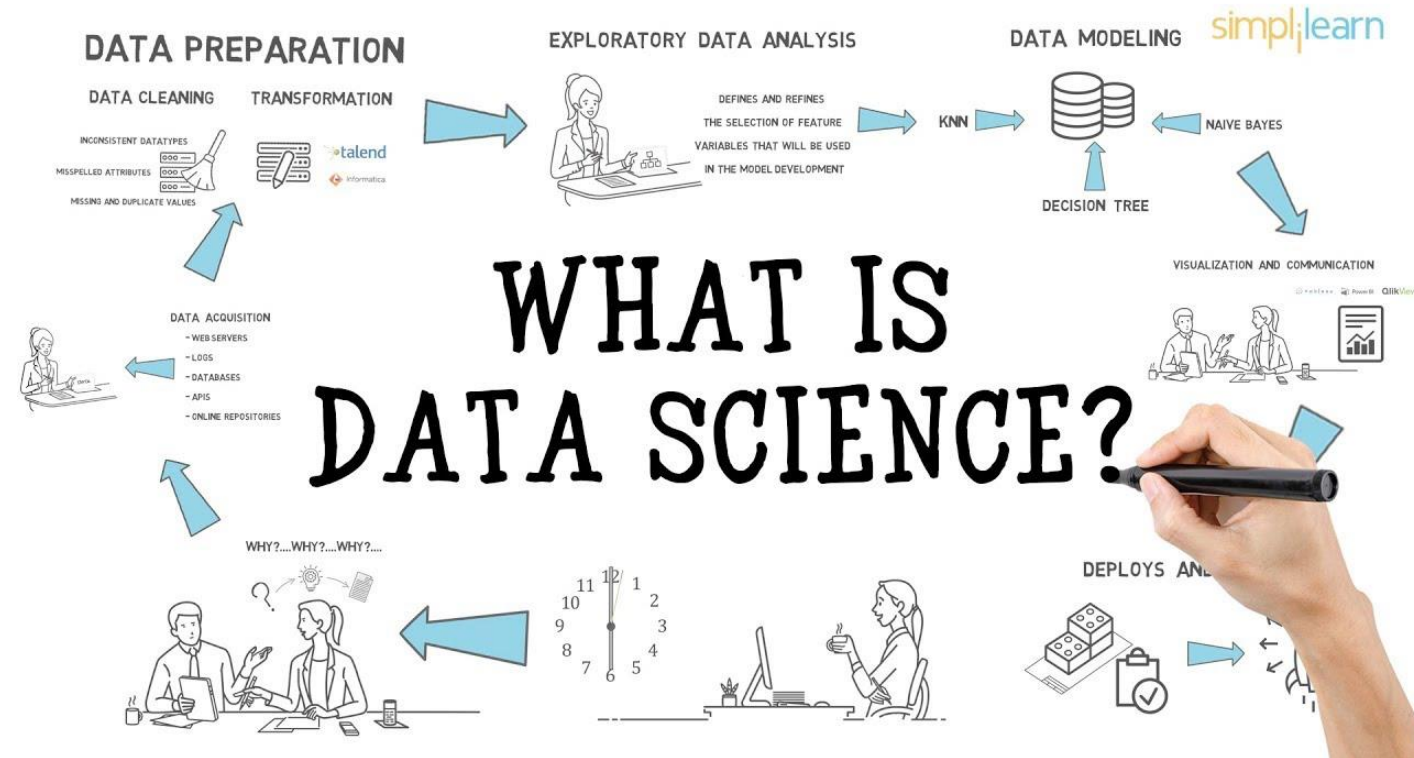
[8] Dataconomy 2016



# 3. DATA SCIENCE & BIG DATA



[17] Towards Data Science 2018



[18] SimpliLearn



# 3. DATA SCIENCE & BIG DATA

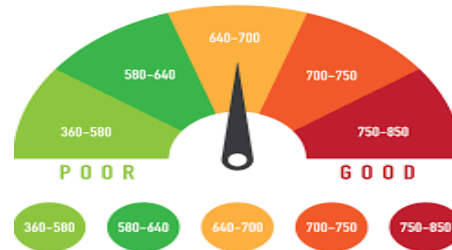
Lịch sử tín dụng của user



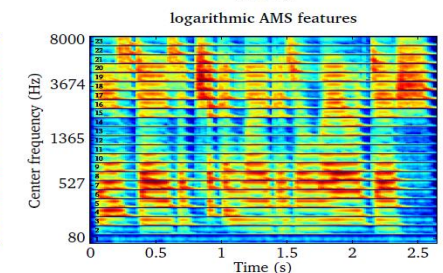
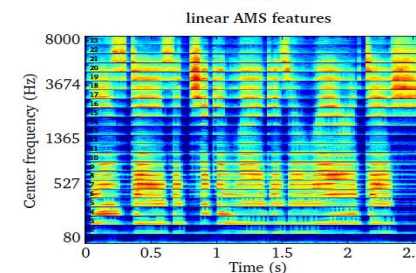
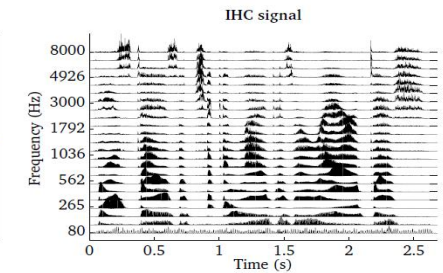
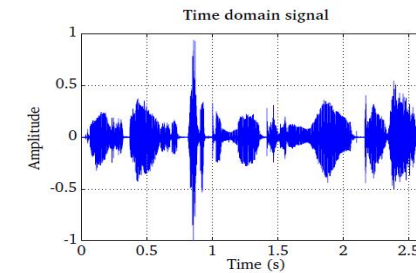
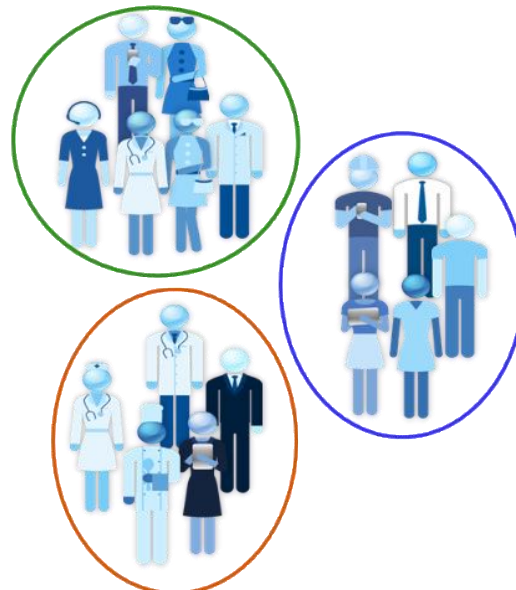
Lịch sử của gói tín dụng



Thông tin khách hàng



Credit scoring



# TYPE OF DATA

## Structural data

Sales

Edit Design

Import Data

Filter

Sort

Add

Delete

More

Search Data

	Date	T Region	T Product Category	T Product	T Customer Name	Sales	Cost		
1	06 June, 2014	West	Grocery	Fruits and Vegetables	Vincent Herbert	\$1,682.39	\$200.05		
2	08 June, 2014	East	Furniture	Clocks	John Britto	\$272.34	\$14.58		
3	11 June, 2014	West	Grocery	Fruits and Vegetables	David Flashing	\$2,970.27	\$1,635.85		
4	13 June, 2014	East	Stationery	File Labels	Maxwell Schwartz	\$190.05	\$90.85		
5	16 June, 2014	West	Grocery	Fruits and Vegetables	Lele Donovan	\$5,342.57	\$1,929.65		
6	18 June, 2014	East	Stationery	Art Supplies	Susan Juliet	\$45.31	\$12.93		
7	20 June, 2014	East	Grocery	Fruits and Vegetables	Carl Lewis	\$2,974.81	\$986.08		
8	22 June, 2014	East	Stationery	Specialty Envelopes	Pete Zachariah	\$455.08	\$195.66		
9	23 June, 2014	West	Grocery	Fruits and Vegetables	Andy Roddick	\$3,928.38	\$1,386.98		
10	25 June, 2014	West	Stationery	Copy Paper	Venus Powell	\$409.51	\$40.92		
11	27 June, 2014	East	Stationery	Computer Paper	Pete Zachariah	\$27.69	\$9.51		
12	28 June, 2014	East	Grocery	Fruits and Vegetables	Hilary Holden	\$955.88	\$573.23		
13	29 June, 2014	West	Stationery	Highlighters	Joseph Aaron	\$37.48	\$0.13		
14	29 June, 2014	East	Stationery	Standard Labels	Patrick O'Brill	\$225.84	\$99.76		

6

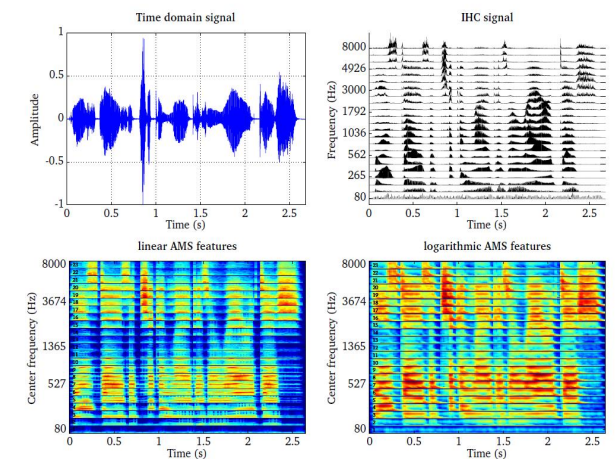
Rows: 755

2

## Unstructured data

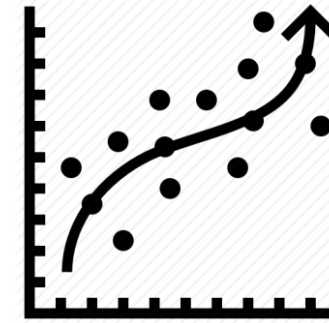
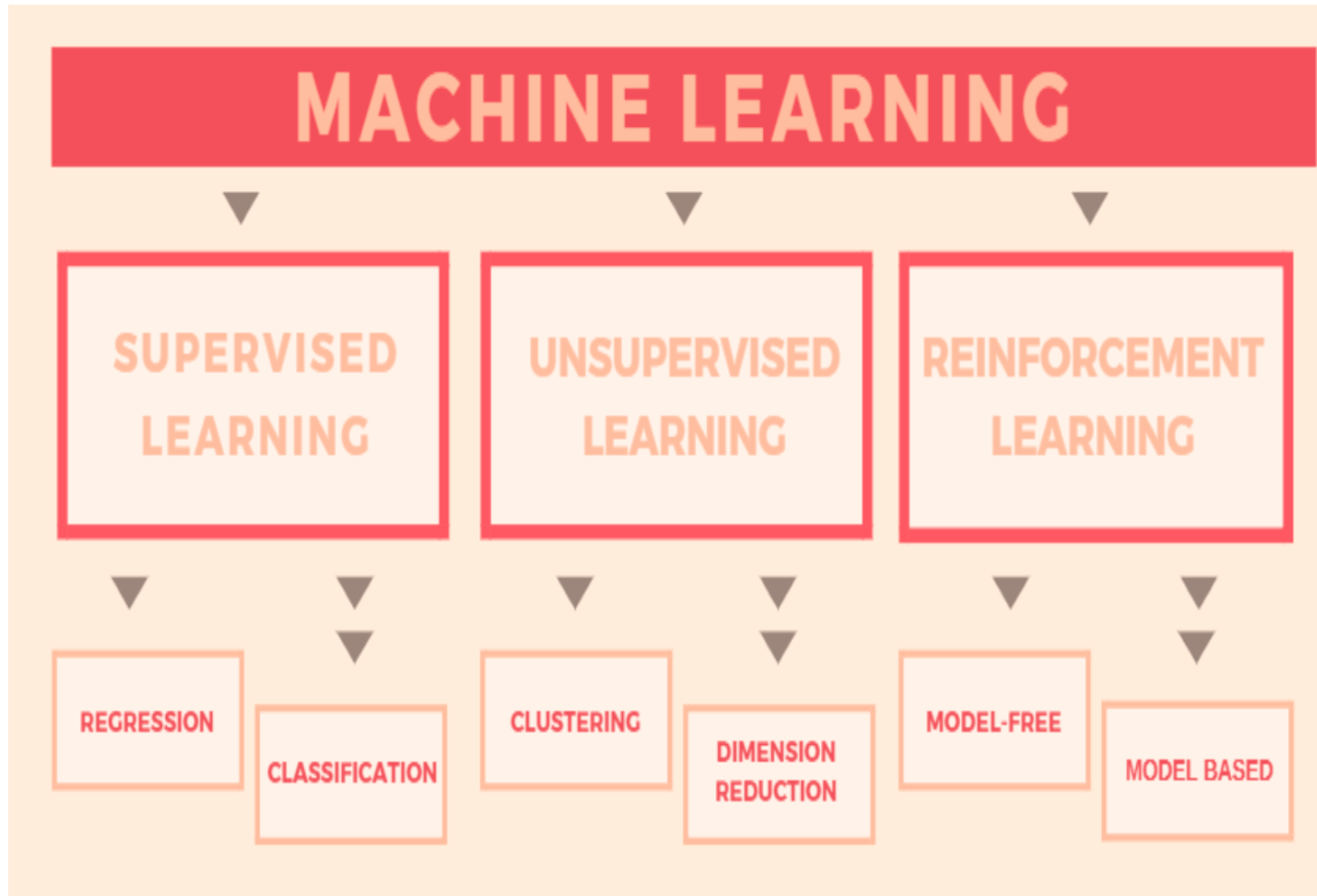


and ...

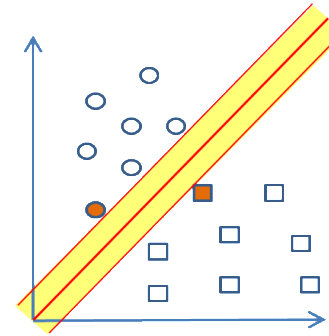




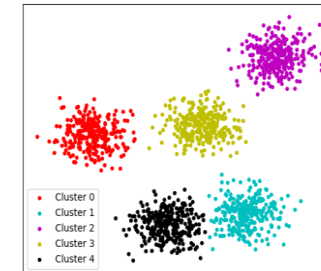
# TYPE OF MACHINE LEARNING



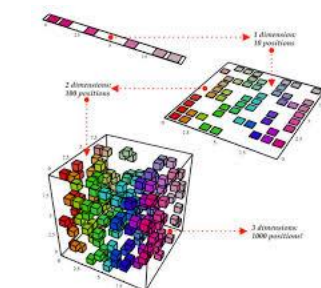
**Regression**  
Income prediction  
Credit scoring



**Classification**  
Bad user detection  
Fraud detection

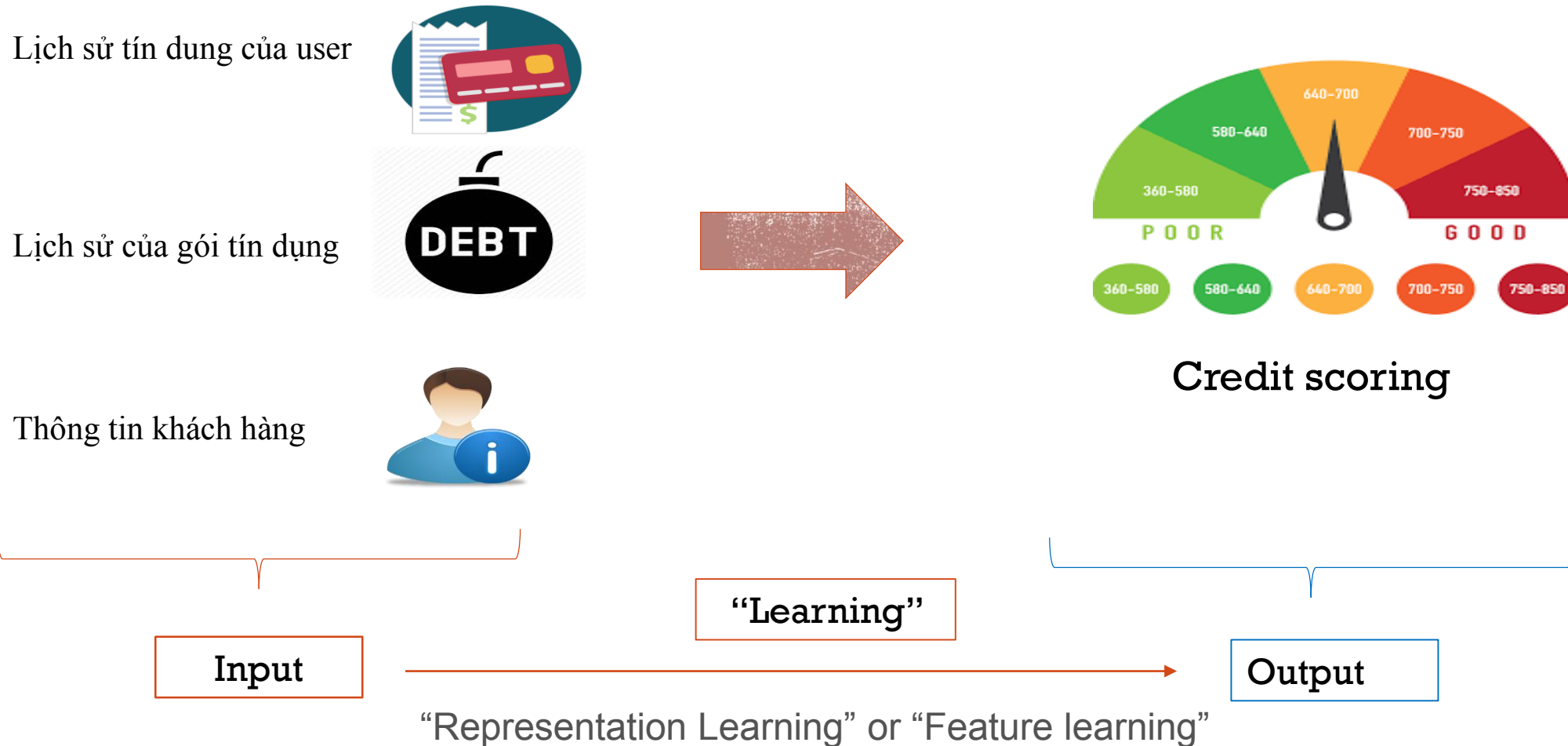


**Clustering**  
Topic modeling

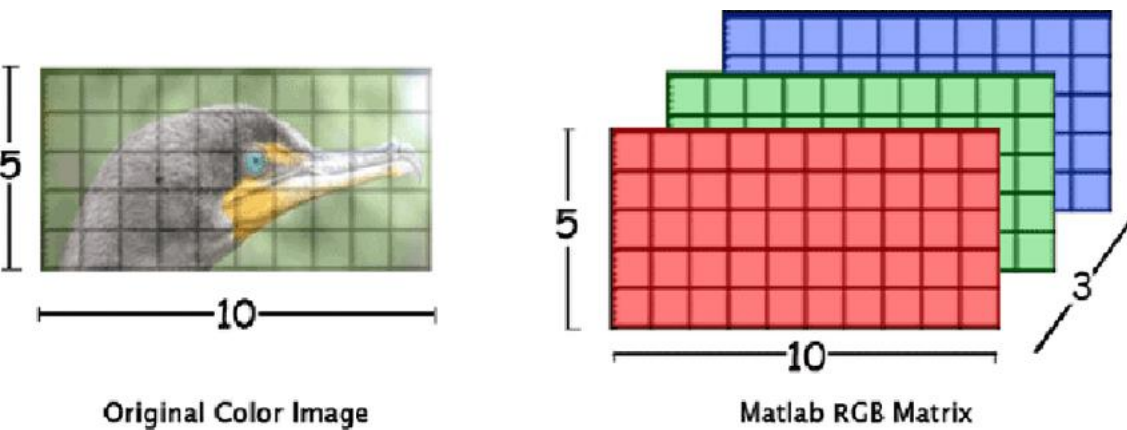


**Dimension reduction**  
TSNE  
PCA

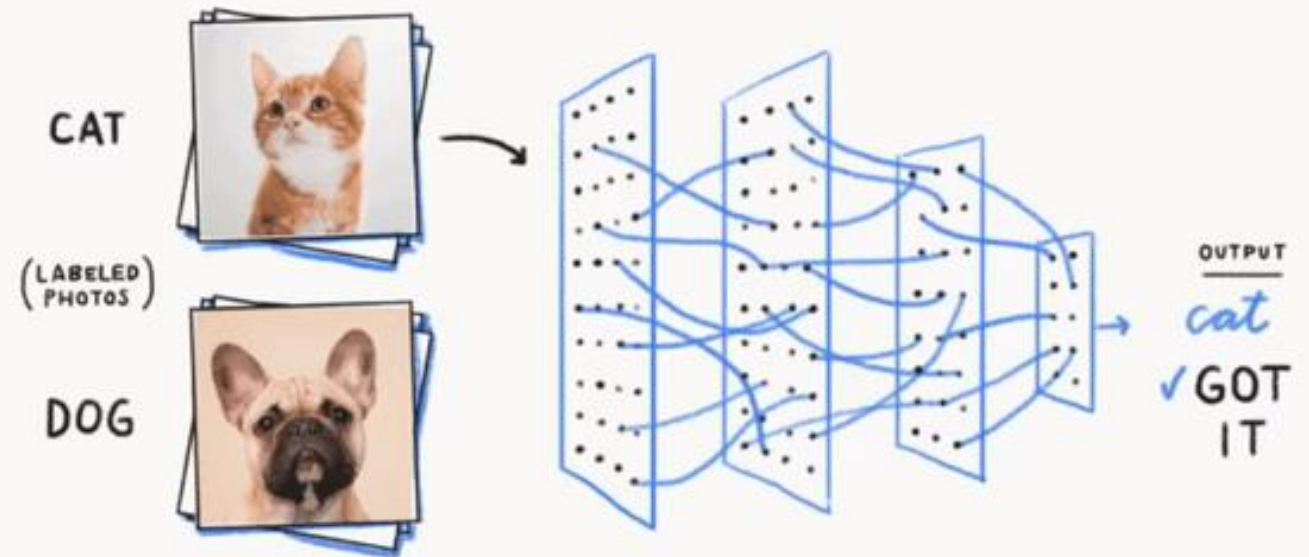
# LEARNING PROCESSING



# LEARNING PROCESSING



A Neural Network is a **function** that can learn



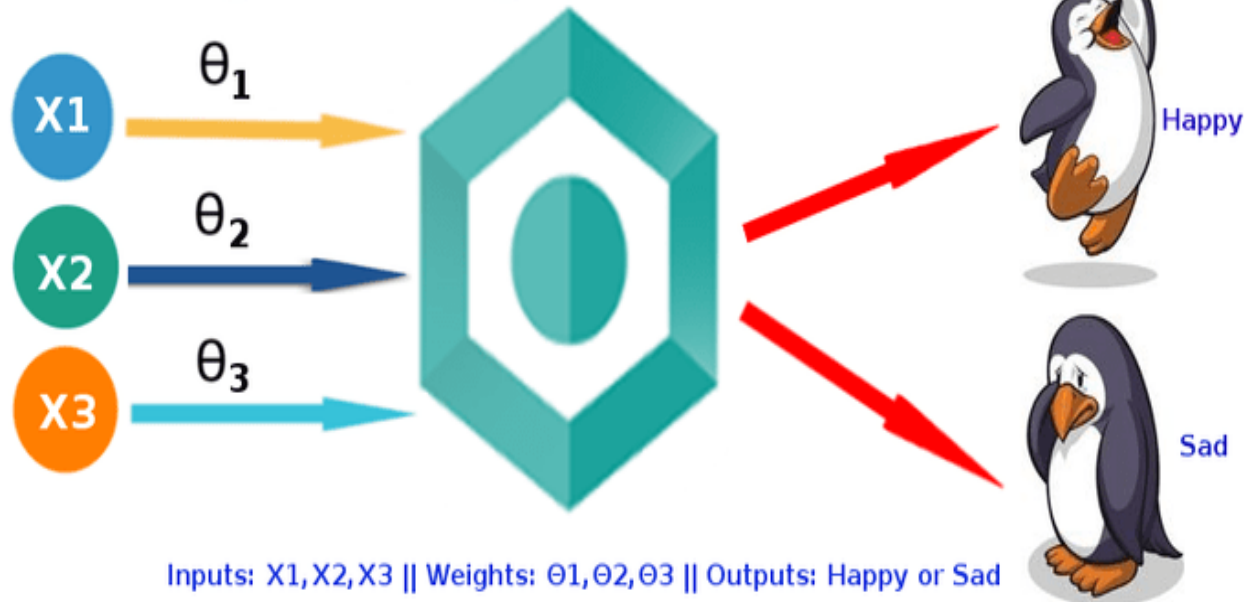
Input

"Learning"

Output

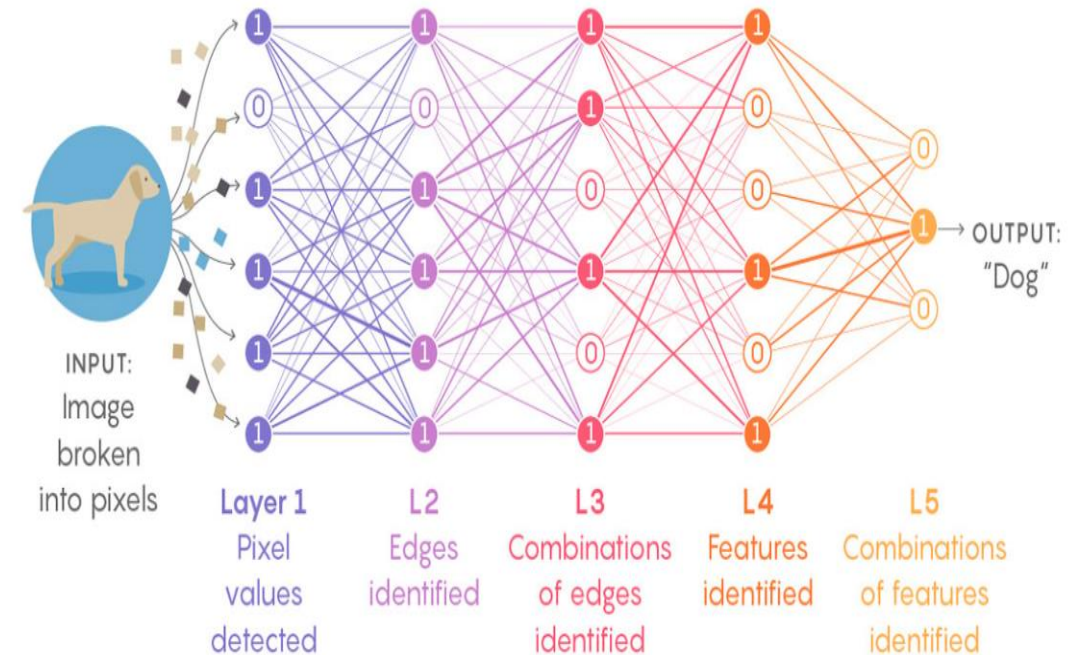
# SALLOW LEARNING VS DEEP LEARNING

## Logistic Regression Model



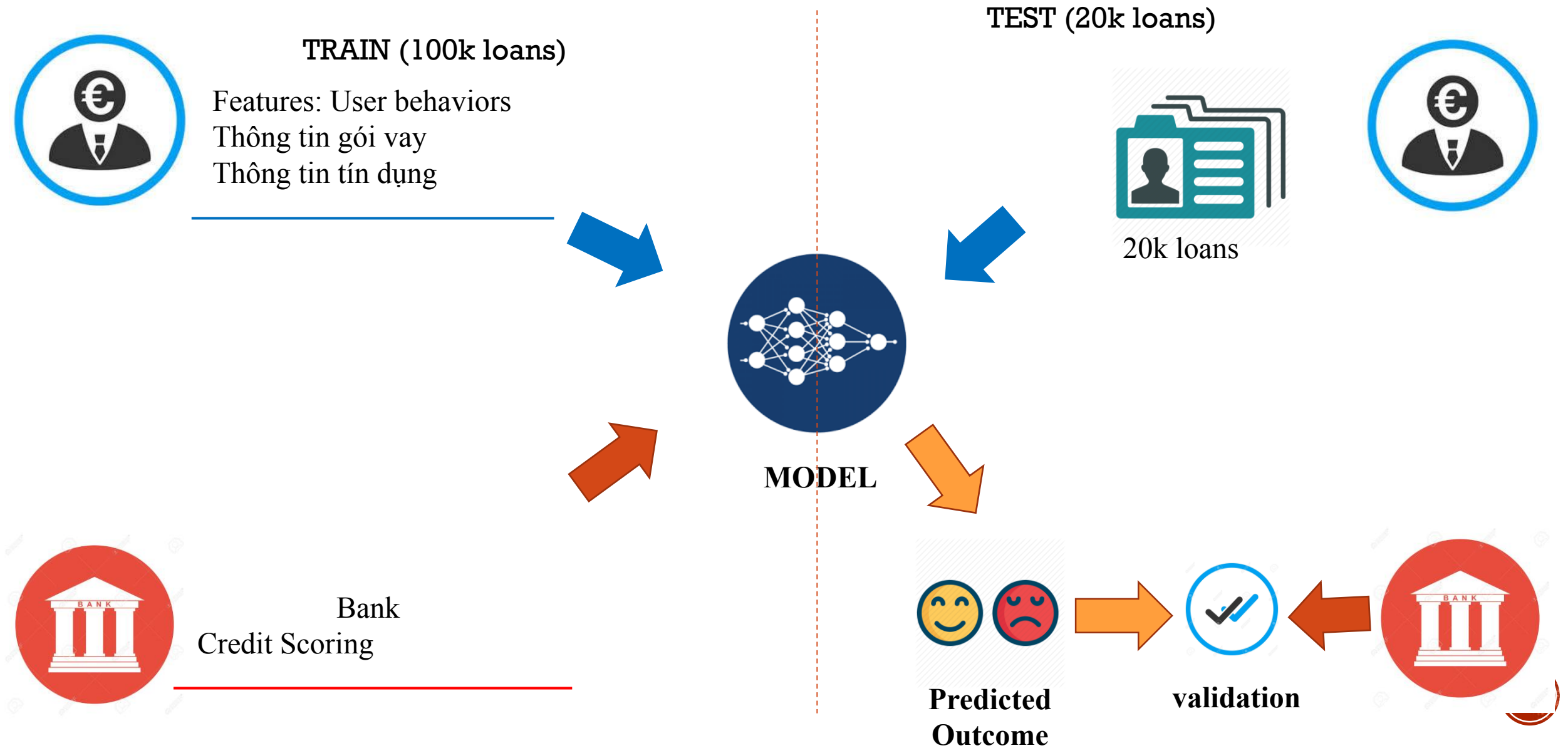
@dataaspirant.com

“Feature Engineering” or “Feature Selection”



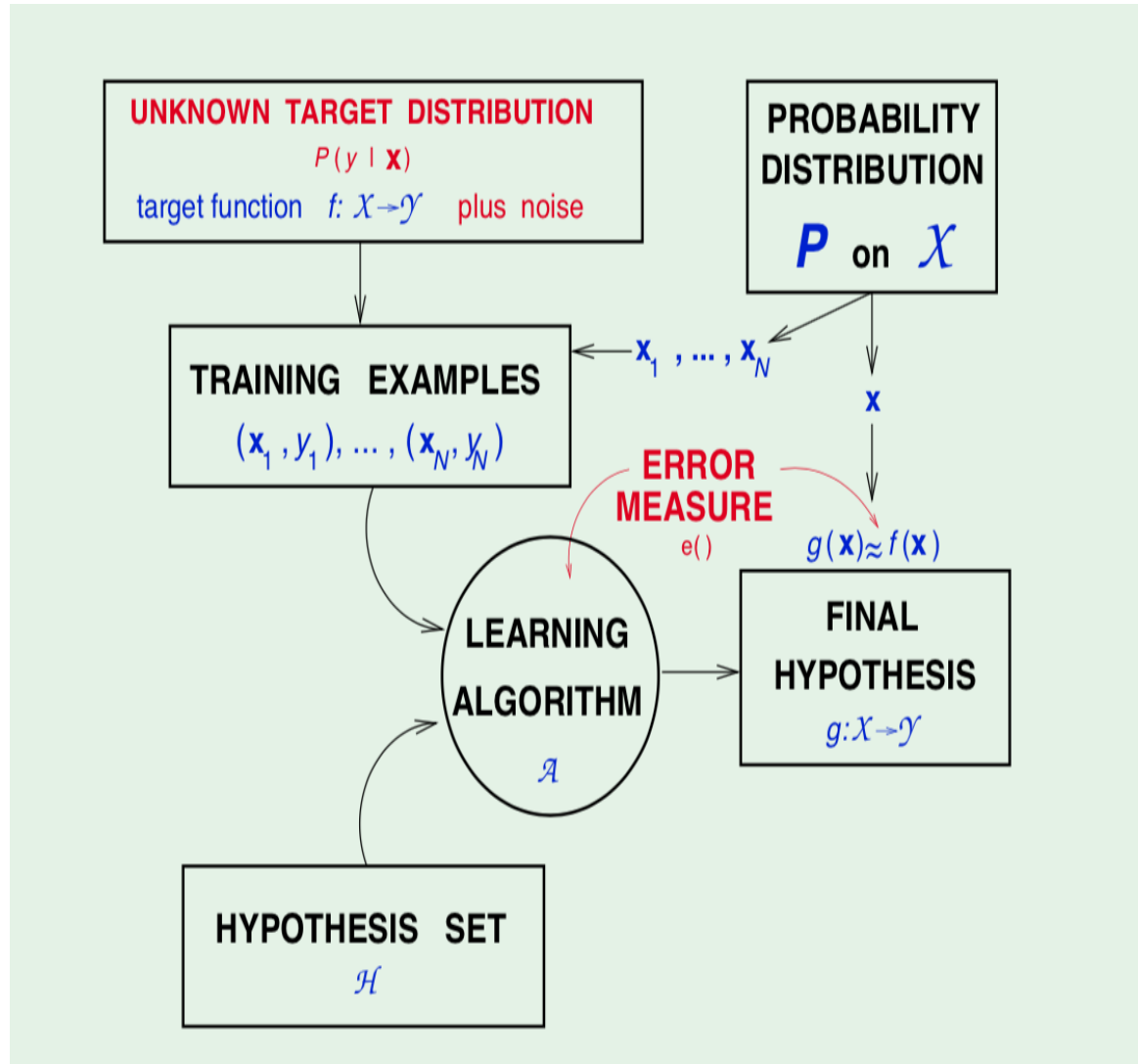
Deep learning

# LEARNING PROCESSING



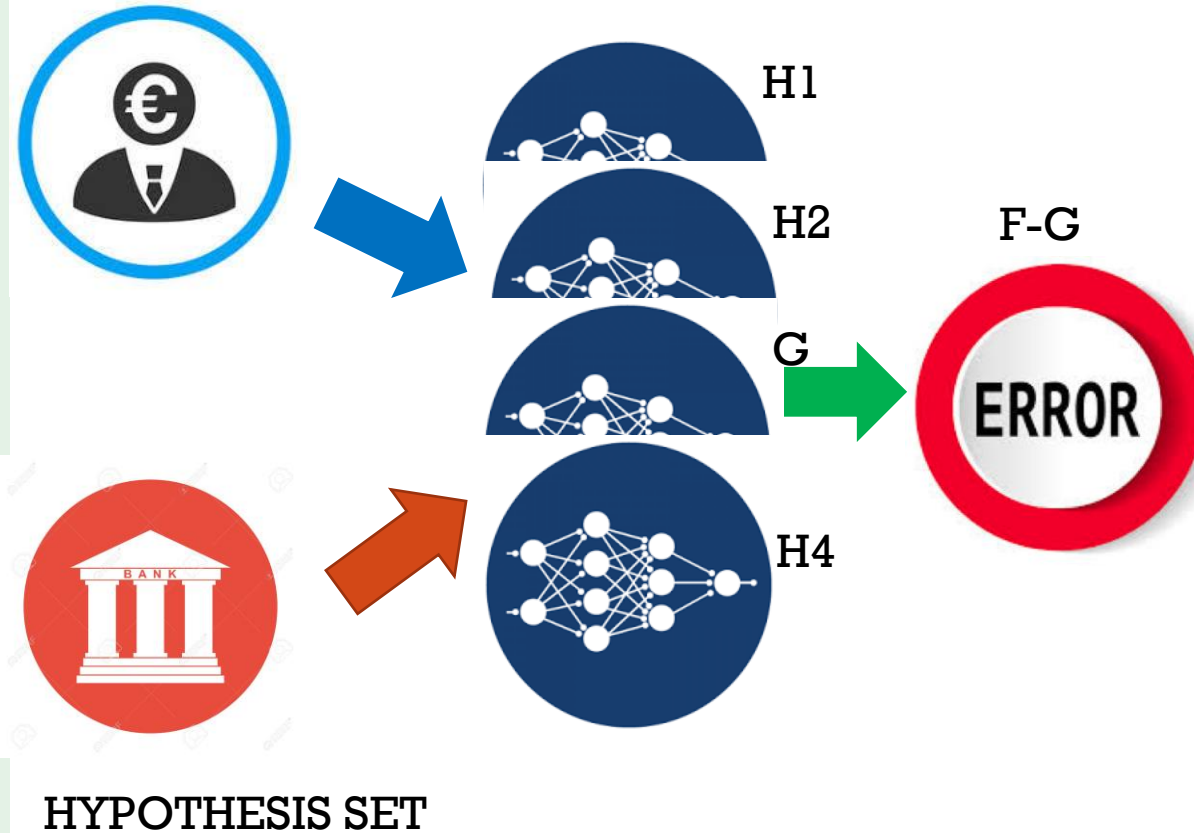


# LEARNING PROCESSING



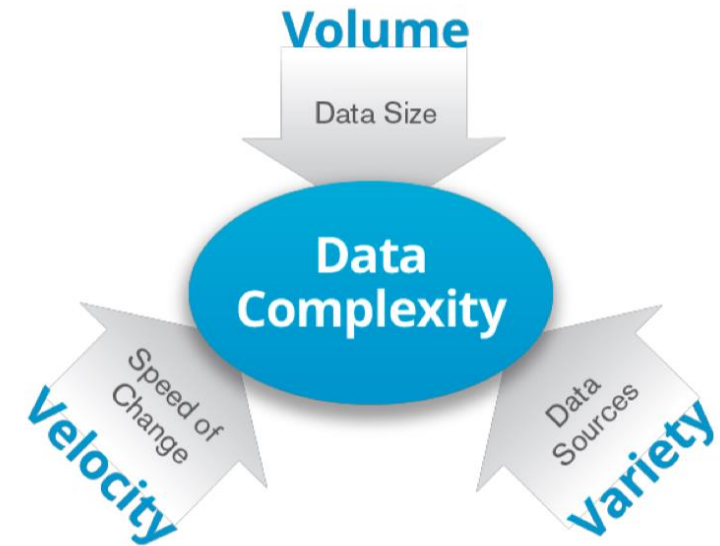
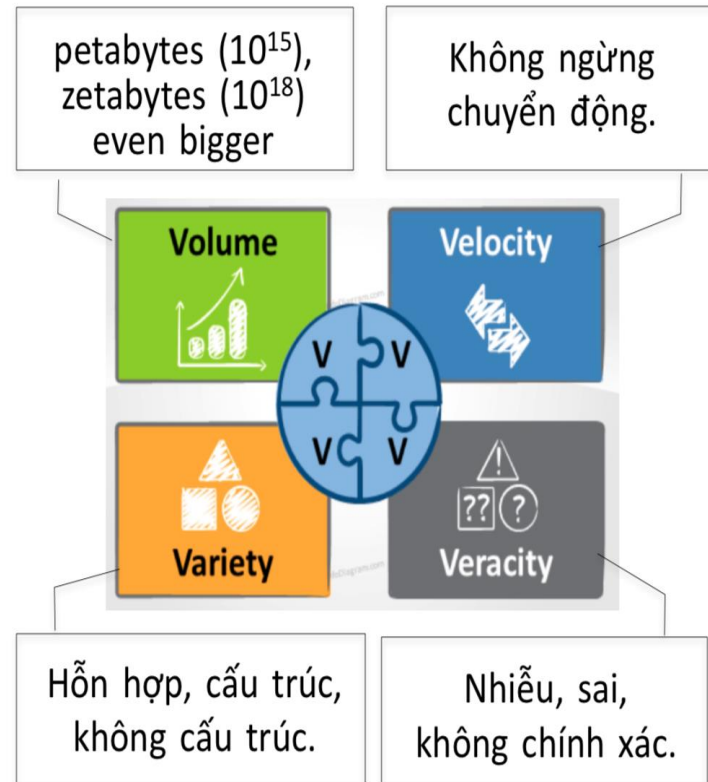
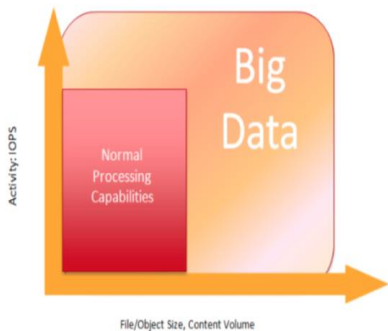
TRAIN (100k loans)

TEST



# 1. BIG DATA – WHAT IS BIG DATA?

Dữ liệu lớn nói về các tập **dữ liệu rất lớn** và/hoặc **rất phức tạp**, vượt quá khả năng xử lý của các kỹ thuật IT truyền thống (View 1).

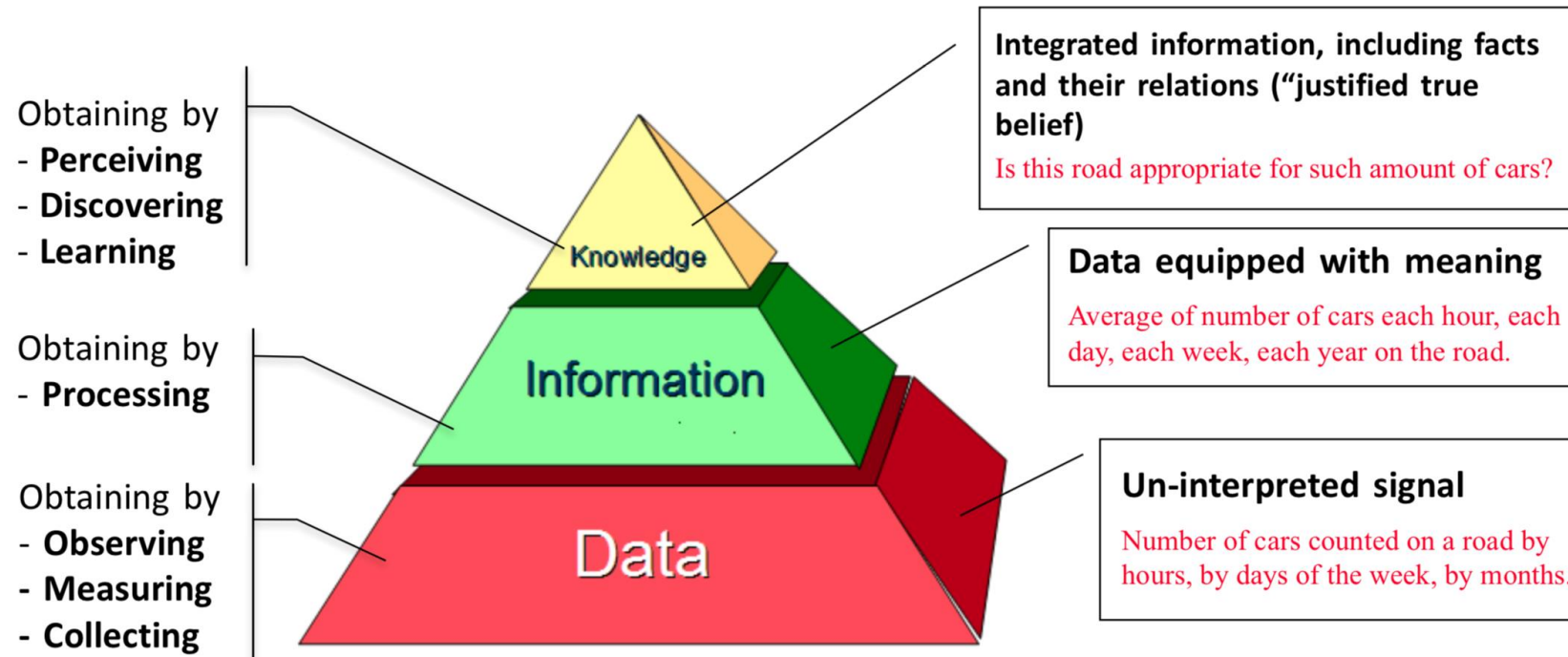


*“Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.” - Gartner*

# 1. BIG DATA – VALUE OF BIG DATA ANALYTICS

*“Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable **enhanced insight, decision making, and process automation.**” - Gartner*

Src: [5]



Src: [1]

# 3. COURSE SCHEMA

1. Introduction (1<sup>st</sup> days)
2. The learning problems [Caltech, Microsoft (bitshop)] (2<sup>nd</sup> day)
3. Exploratory Data Analysis – Data visualization [R] (2<sup>nd</sup> day)
4. Bias – variance trade-off. [Caltech] (3<sup>rd</sup> day)
5. Overfitting vs Underfitting [Caltech, Stanford] (3<sup>rd</sup> day)
6. Learning curve (3<sup>rd</sup> day)
7. Running model [R] (3<sup>rd</sup> day)
8. Cross Validation [Caltech, Stanford] (4<sup>rd</sup> day)
9. Regularization (4<sup>rd</sup> day)
10. Tuning [R] (4<sup>rd</sup> day)
11. Learning Principal [Caltech] (5<sup>rd</sup> day)
12. Evaluation [sonpvh] (5<sup>rd</sup> day) [R]
13. **Summary**



- 31/3: outlier + 5 presentation
- 6/4: feedback (thầy Phú) + code R (sơn)
- 13/4: full code R (sơn)

# REFERENCES:

1. [https://en.wikipedia.org/wiki/Lending\\_Club](https://en.wikipedia.org/wiki/Lending_Club)
2. <https://trustingsocial.com/about>
3. <https://towardsdatascience.com/prototyping-a-recommender-system-step-by-step-part-1-knn-item-based-collaborative-filtering-637969614ea>
4. <https://www.slideshare.net/xamat/recommender-systems-machine-learning-summer-school-2014-cmu>
5. <https://www.edureka.co/blog/data-science-applications/>
6. <https://www.kaggle.com/competitions>
7. <https://challenge.zalo.ai>
8. <https://dataconomy.com/2016/06/history-data-mining/>
9. <https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#458efabc55cf>
10. [https://simple.wikipedia.org/wiki/Regression\\_analysis](https://simple.wikipedia.org/wiki/Regression_analysis)
11. Hồ Tú Bảo, Khoa học dữ liệu và cách mạng công nghiệp lần thứ 4
12. Smolan and Erwit, The human face of big data, 2013
13. Đình Phùng, phương pháp và công nghệ dữ liệu lớn, 2017
14. Fujitsu Journal, How digital technology will transform the world, 1.2016
15. NTNU, Introduction to big data
16. [https://courses.edx.org/asset-v1:ColumbiaX+CSMM.101x+1T2017+type@asset+block@AI\\_edx\\_ml\\_5.intro.pdf](https://courses.edx.org/asset-v1:ColumbiaX+CSMM.101x+1T2017+type@asset+block@AI_edx_ml_5.intro.pdf)

# REFERENCES:

17. <https://towardsdatascience.com/introduction-to-statistics-e9d72d818745>
18. <https://www.youtube.com/watch?v=X3paOmcrTjQ>