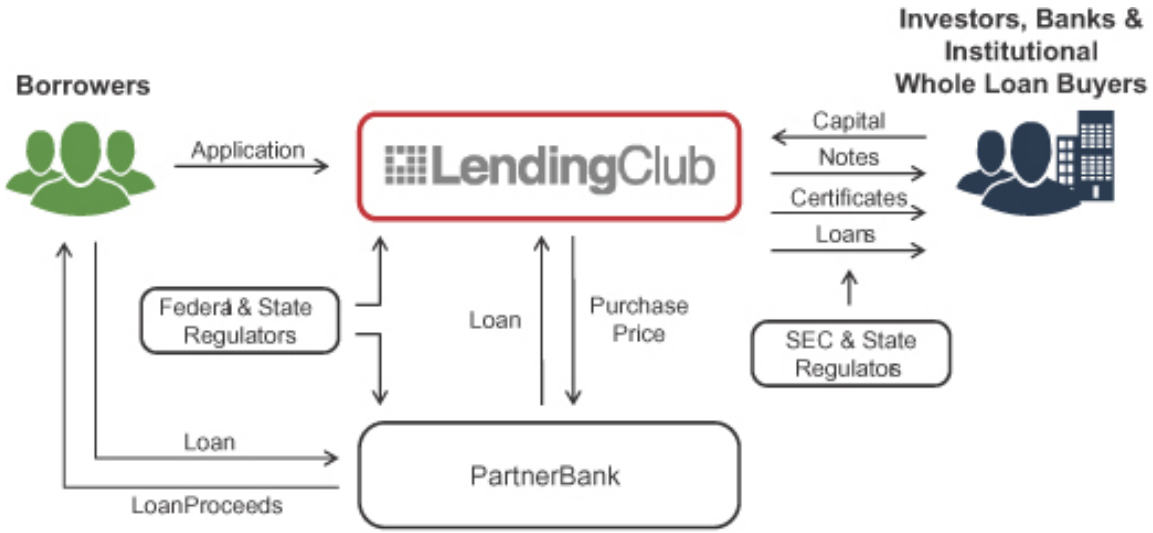
INTRODUCTION DATA SCIENCE & BIG DATA

Sonpvh

OUTLIER

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

1. DATA SCIENCE APPLICATIONS

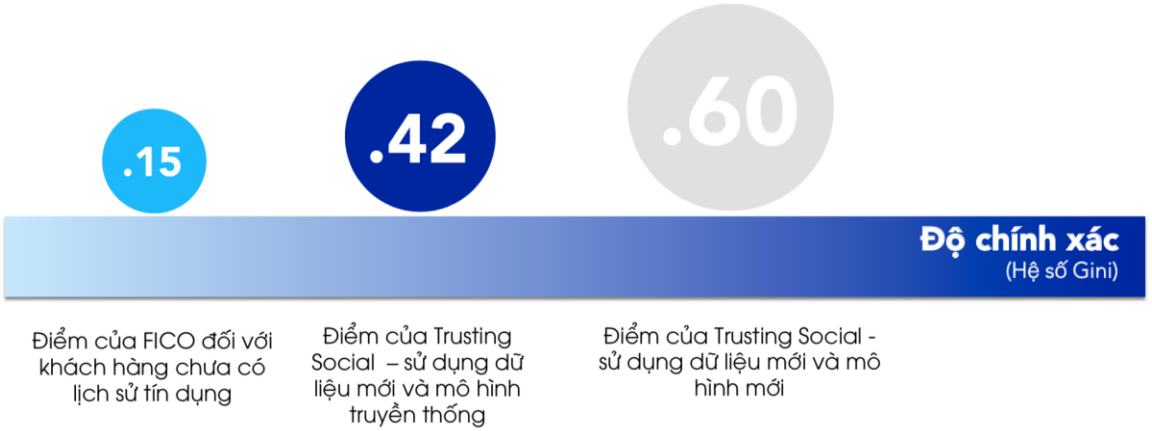


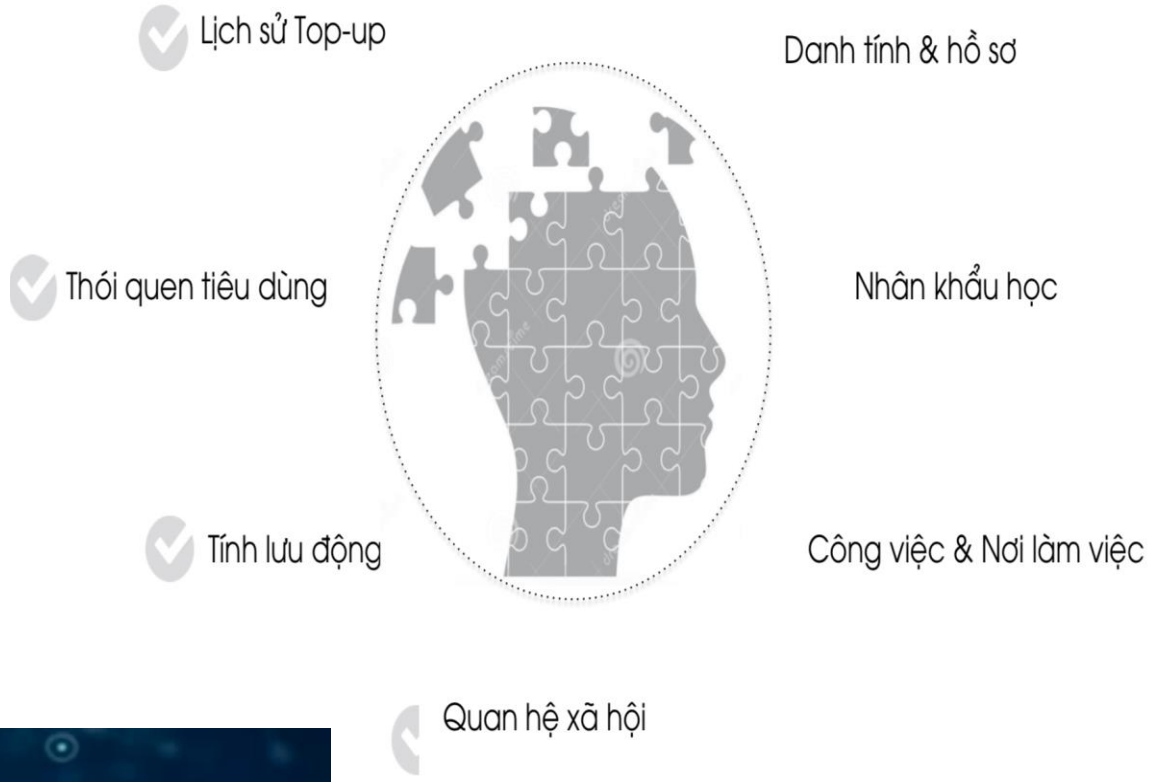
- President: Scott Sanborn
- Founded: 2006
- Valuing the company: 8.5 bn

[1]: wiki

1. DATA SCIENCE APPLICATIONS

trustingsocial





Nguyen Nguyen
CEO

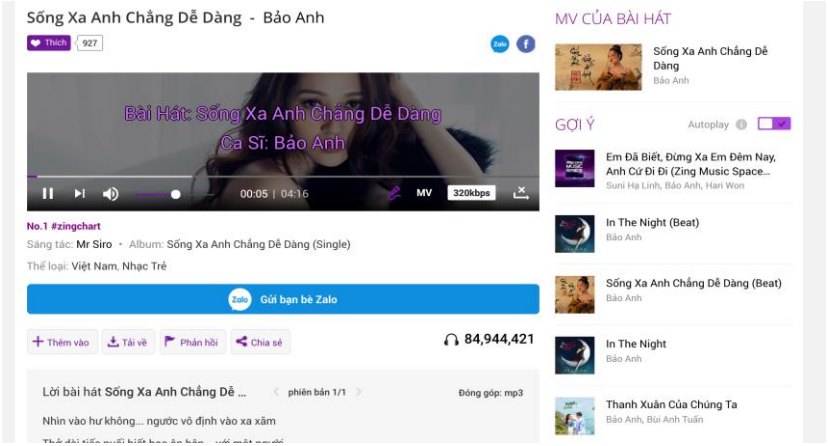
[2]: Trusting social

1. DATA SCIENCE APPLICATIONS

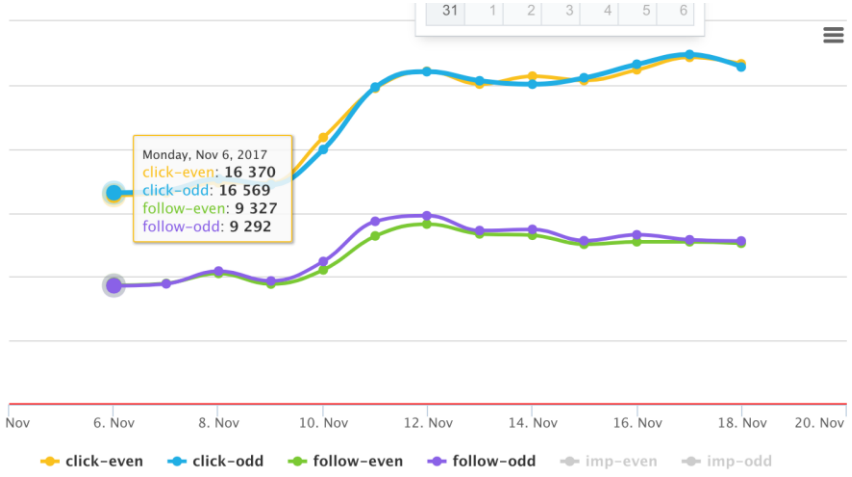
[3]: toward data science

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."*

[4] Xavier 2014

THE PERSONAL EXPERIENCES

MAN

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

Number of Recommendations

117

TOP TEN ALBUM RECOMMENDATIONS

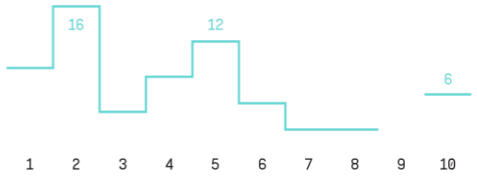
MEDIA

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

Number of Recommendations

785

TOP TEN ALBUM RECOMMENDATIONS



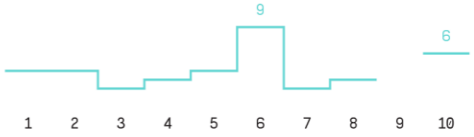
MACHINE

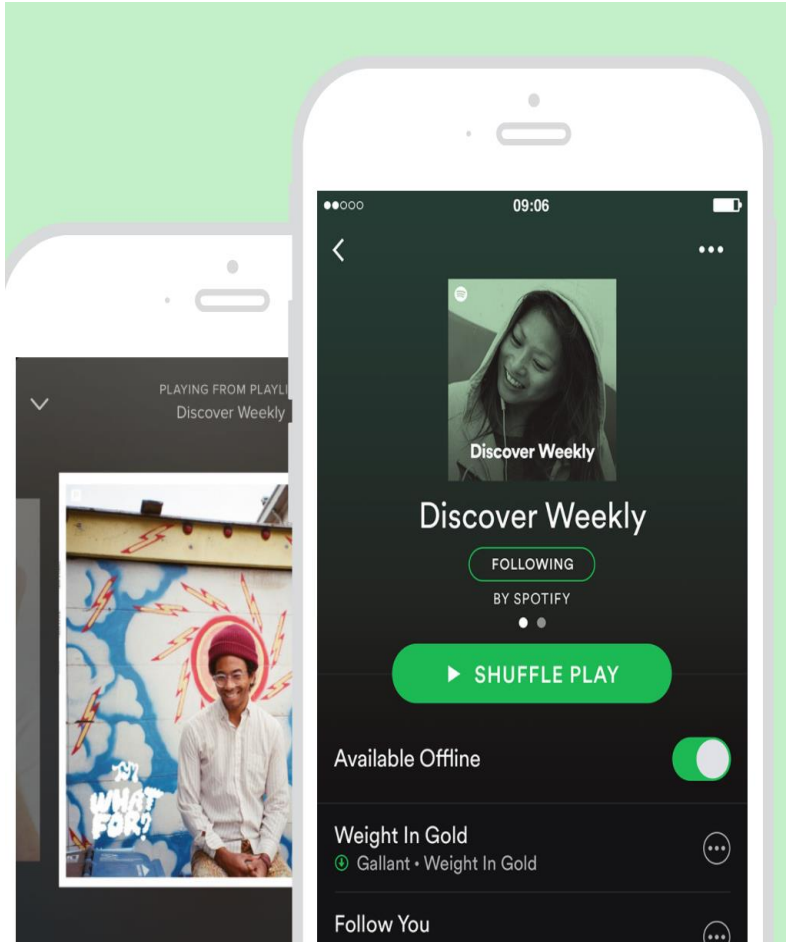
From algorithms: personalized playlists on Spotify like Release Radar.

Number of Recommendations

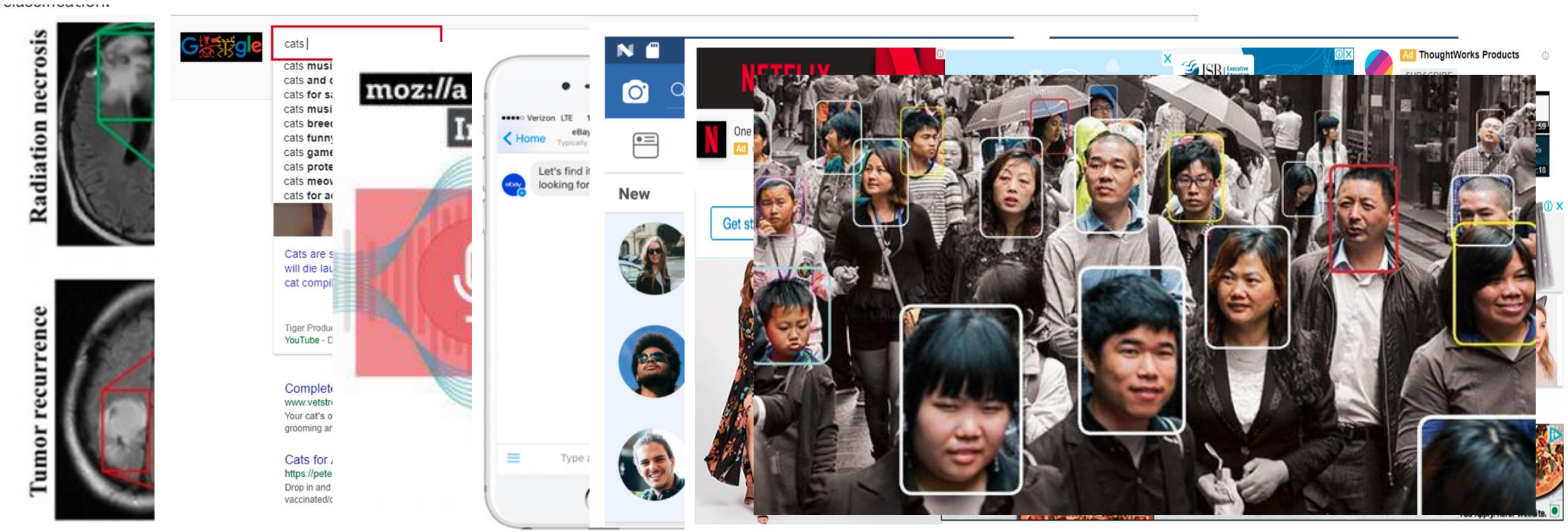
1,363*

TOP TEN ALBUM RECOMMENDATIONS



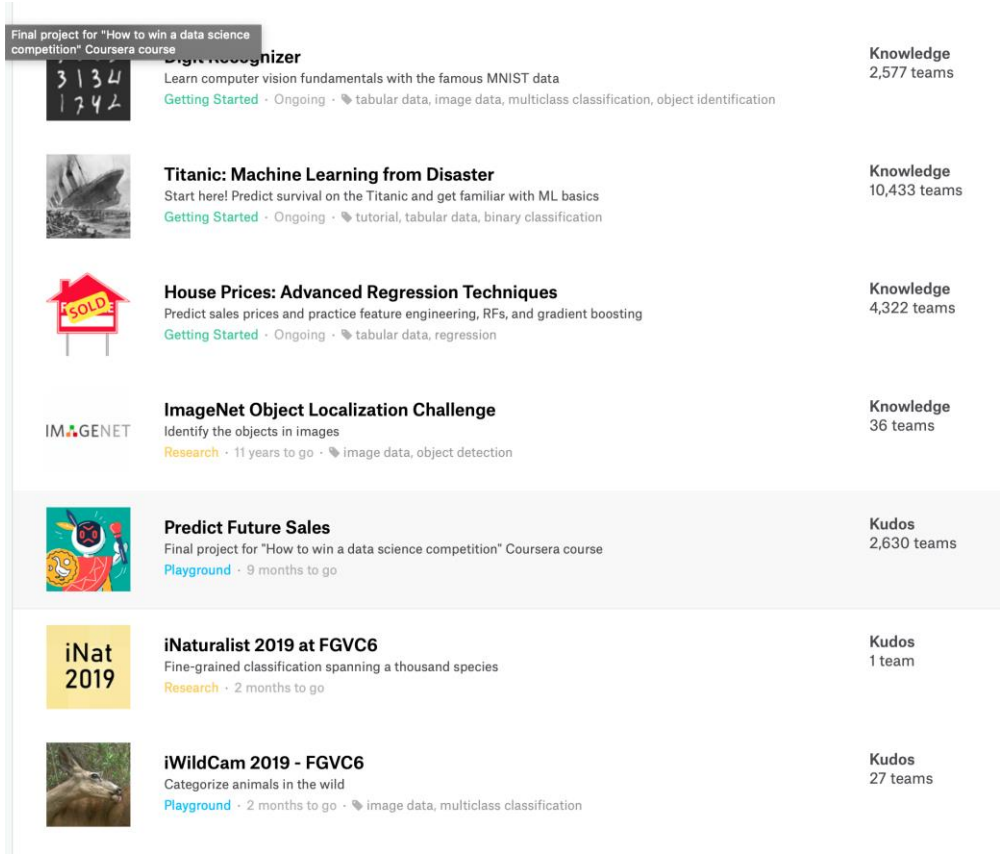


1. DATA SCIENCE APPLICATIONS

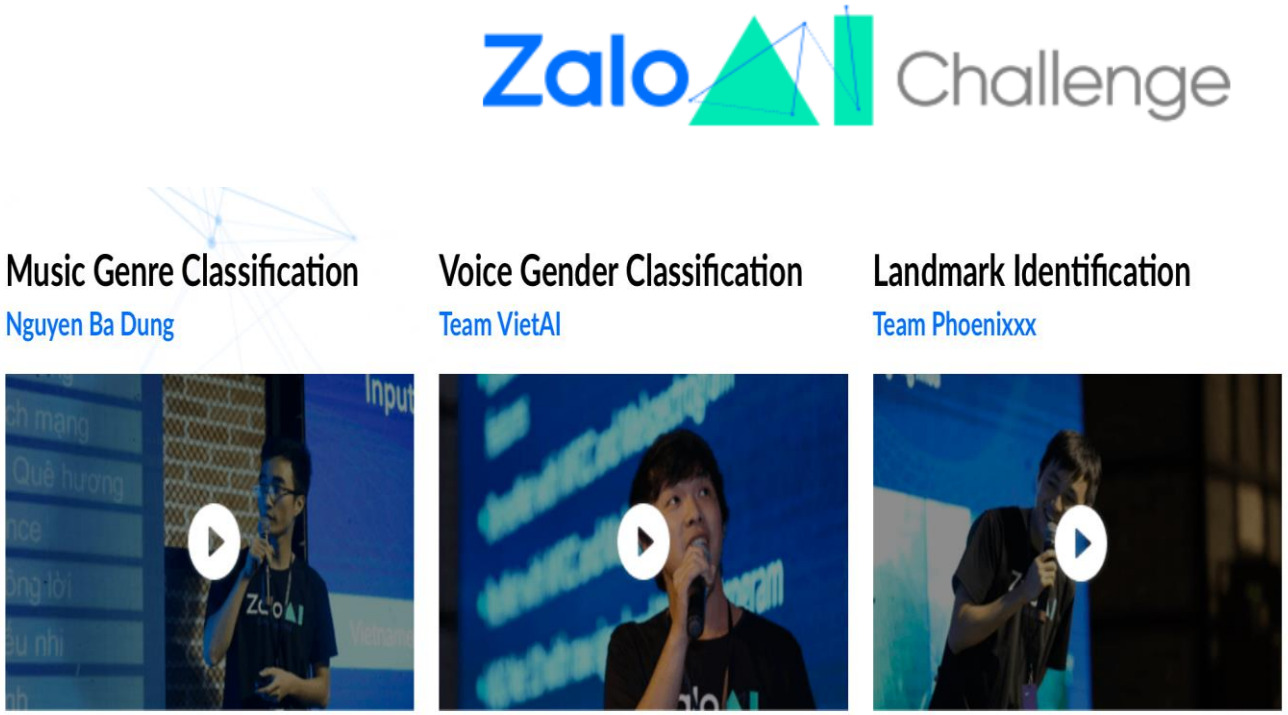


[5] edureka 2019

1. DATA SCIENCE APPLICATIONS



[6] Kaggle



[7] Zalo AI

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

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

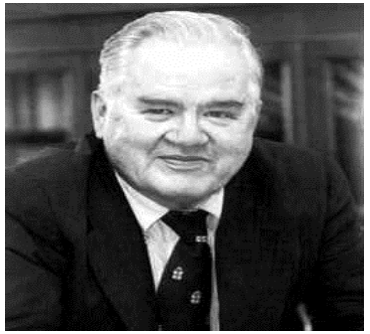
Bayes theorem

- 1763 – Carl Friedrich Gauss (1809) (1821) & Lengendre (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

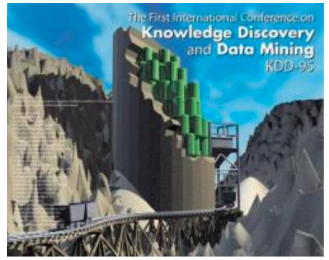
"Datalogy, the science of data and of data processes and its place in education" - "Data Science - The science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences."

- 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."

[9] Gil Press 2013

2. HISTORY - STATISTICS



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

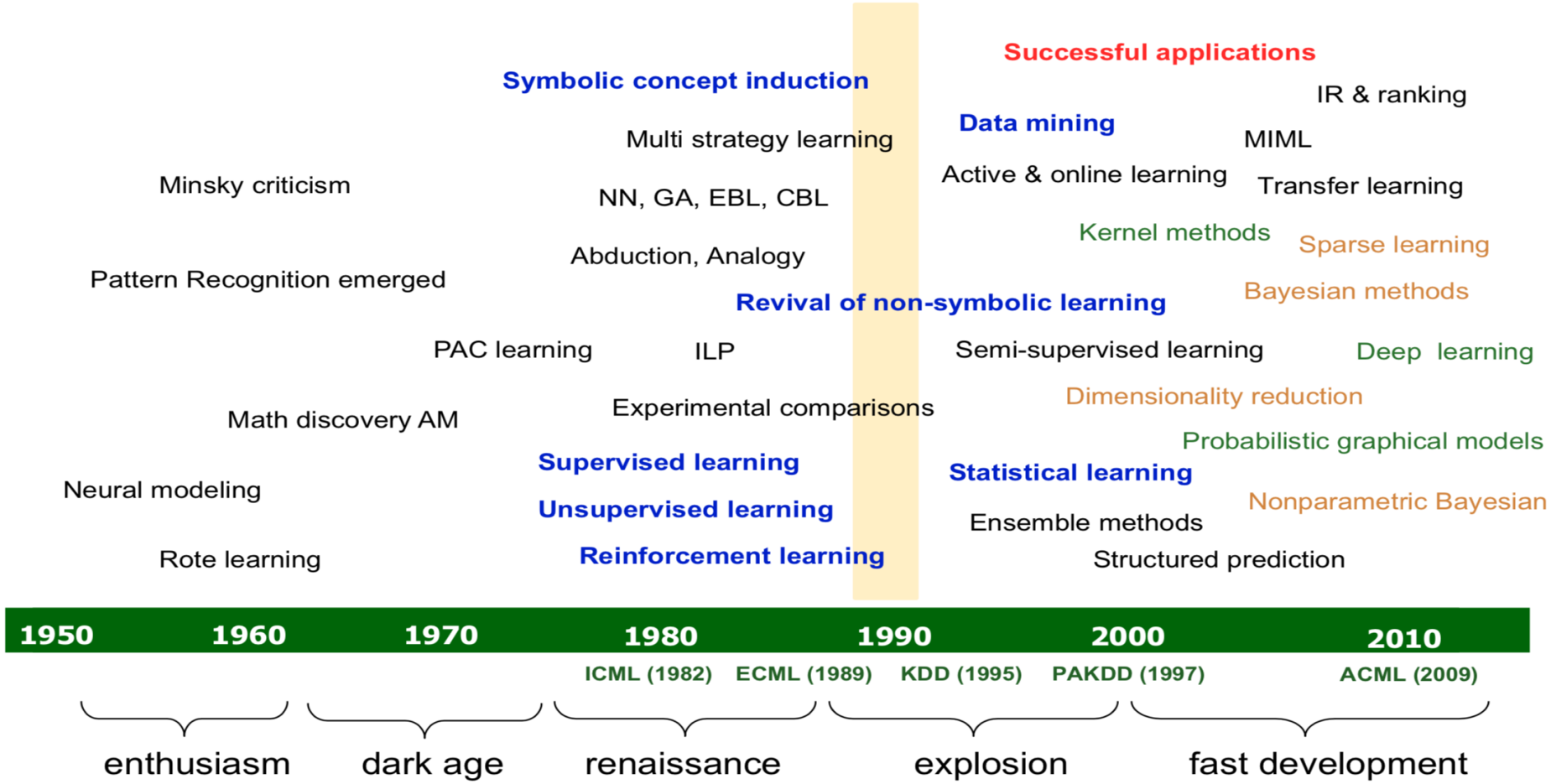
- 1994 – Business week "Databased Marketing"
Companies are **collecting mountains of information about you**, crunching it to **predict how likely you are to buy a product**, and using that knowledge to **craft a marketing message precisely calibrated** to get you to do so…

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

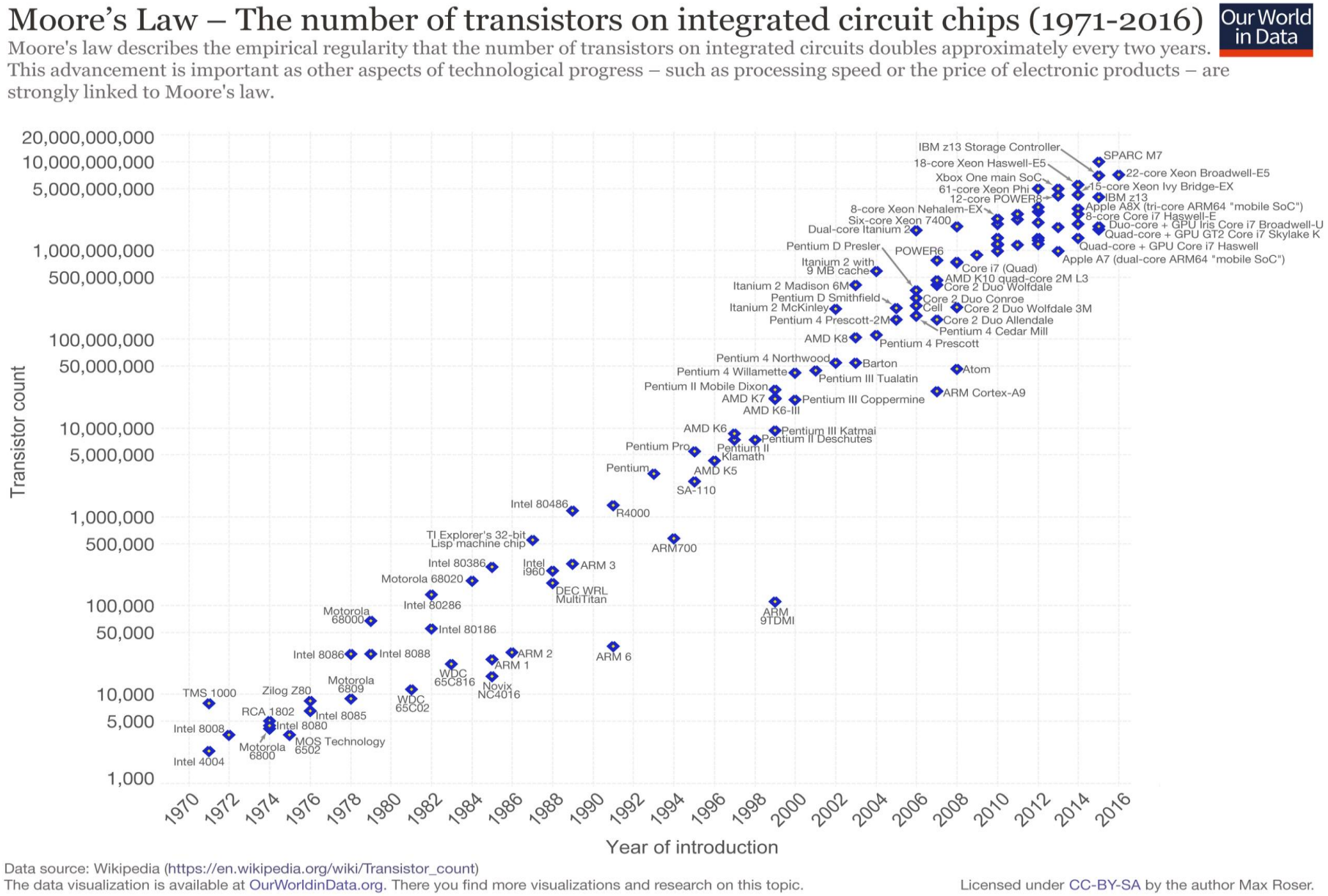
- 1999 - Prof. Moshe Zviran
" Conventional statistical methods work well with small data sets. Today's databases, however, can **involve millions of rows and scores of columns of data** … "

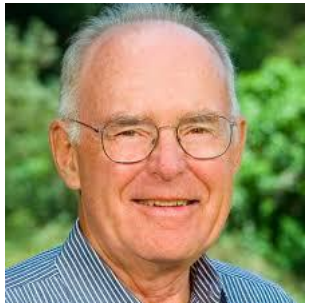
[9] Gil Press 2013

2. HISTORY – DATA MINING



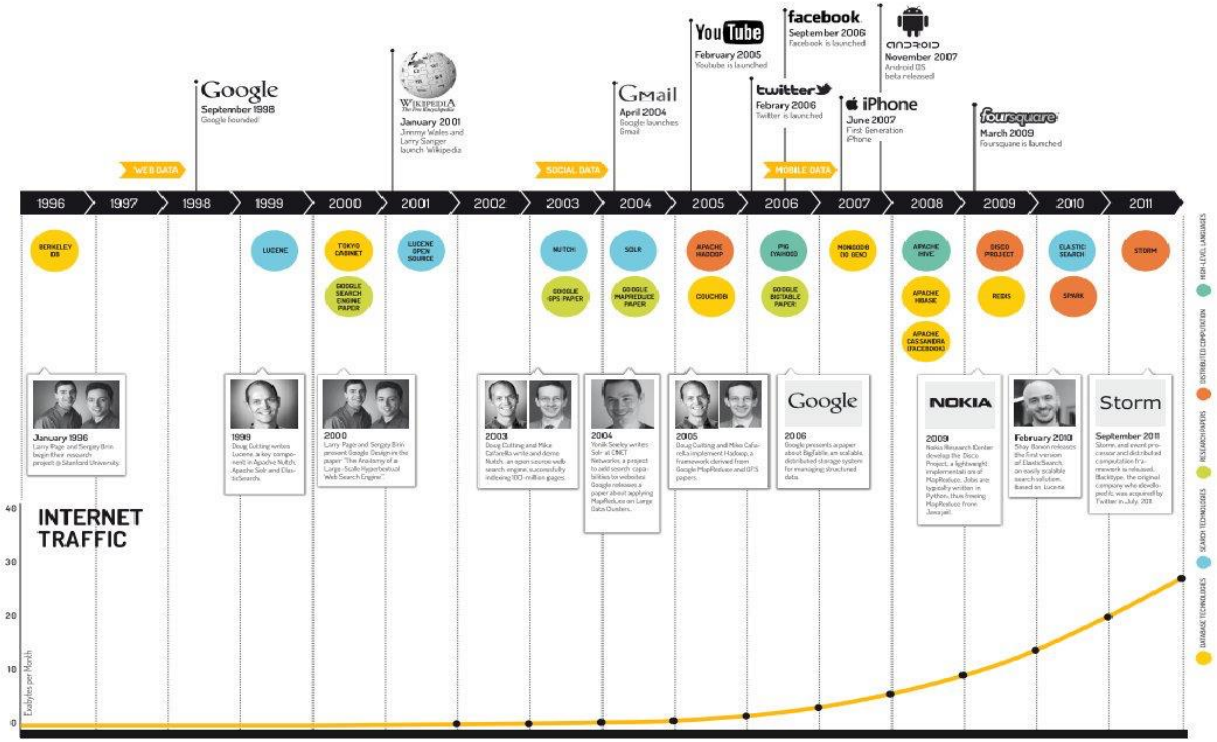
2. THE HISTORY - COMPUTATION



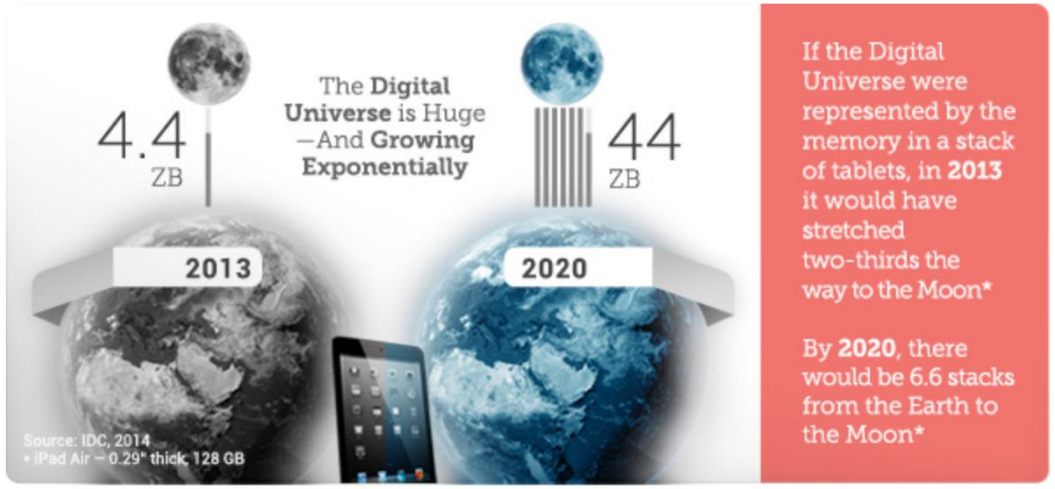


Gordon Earle Moore
US Businessman

2. HISTORY – BIG DATA



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

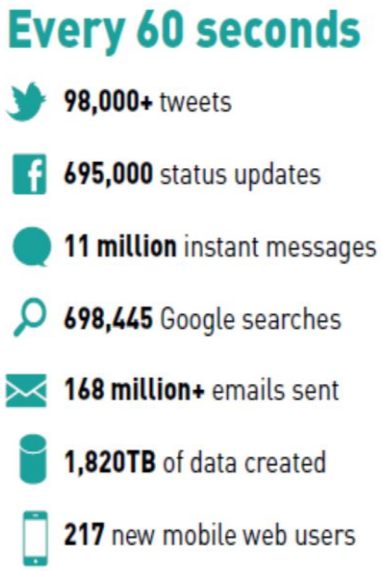


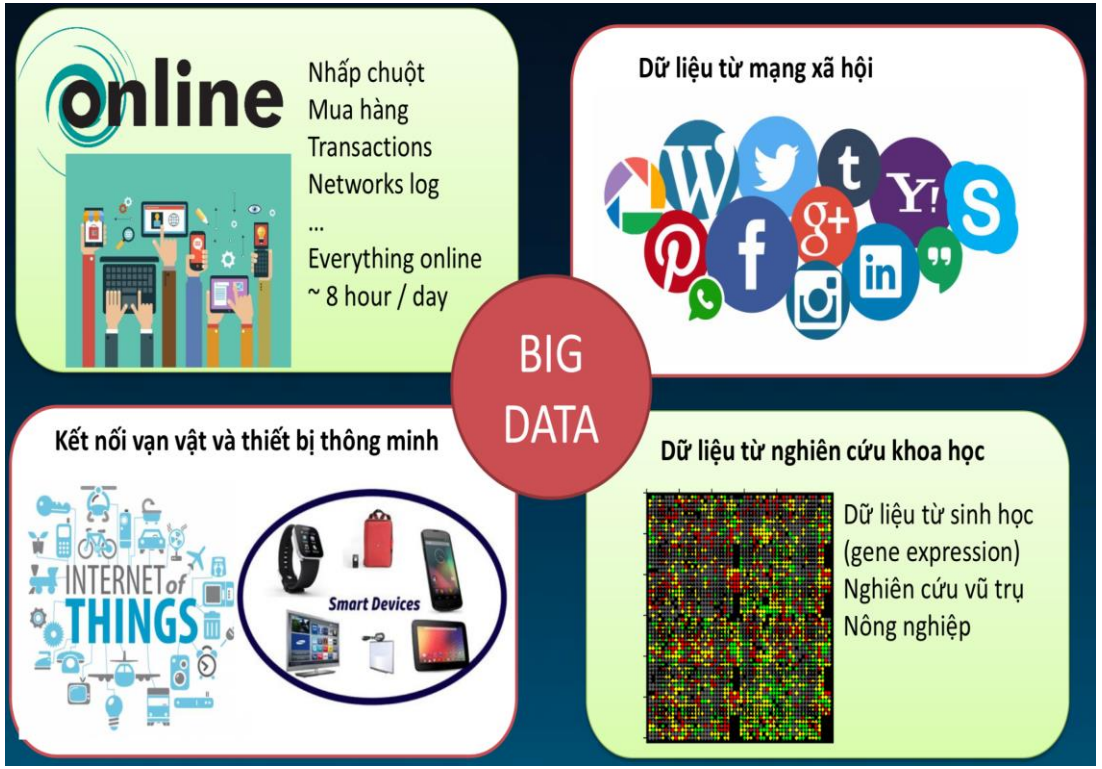
[11] Bigdata - 2016

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"

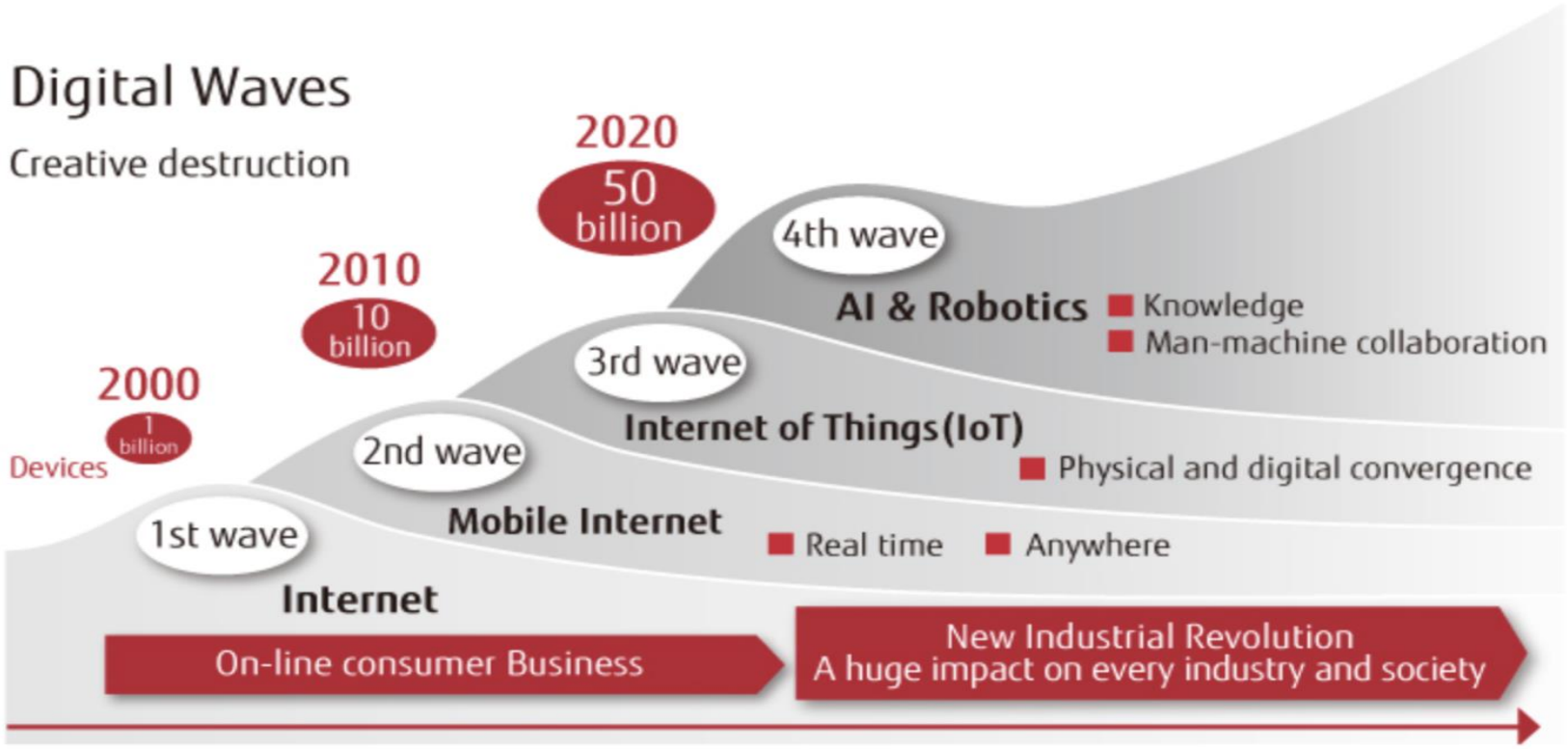
Src: [12]





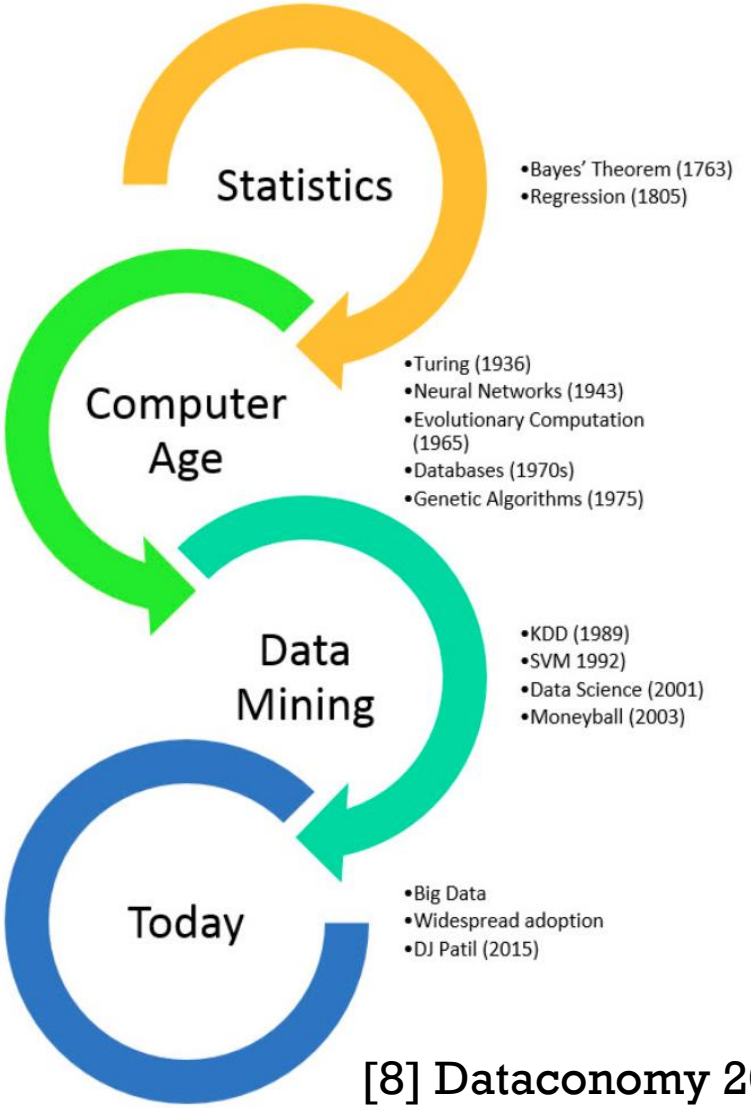
Src: [13]

2. HISTORY– APPLICATION - DIGITAL WAVES

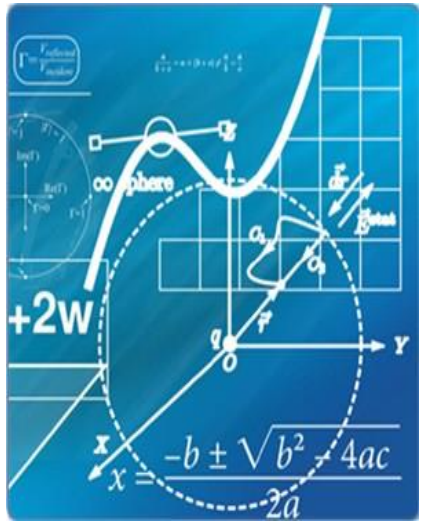


Src: [14]

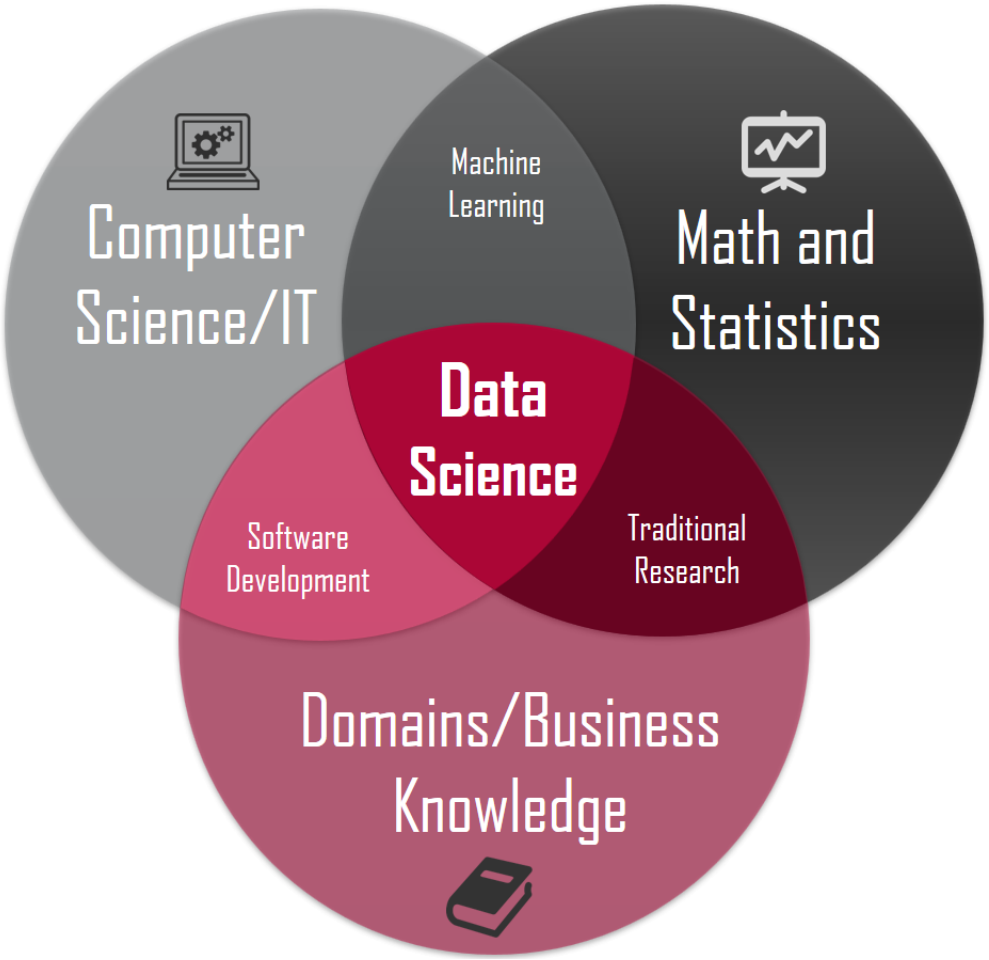
2. HISTORY

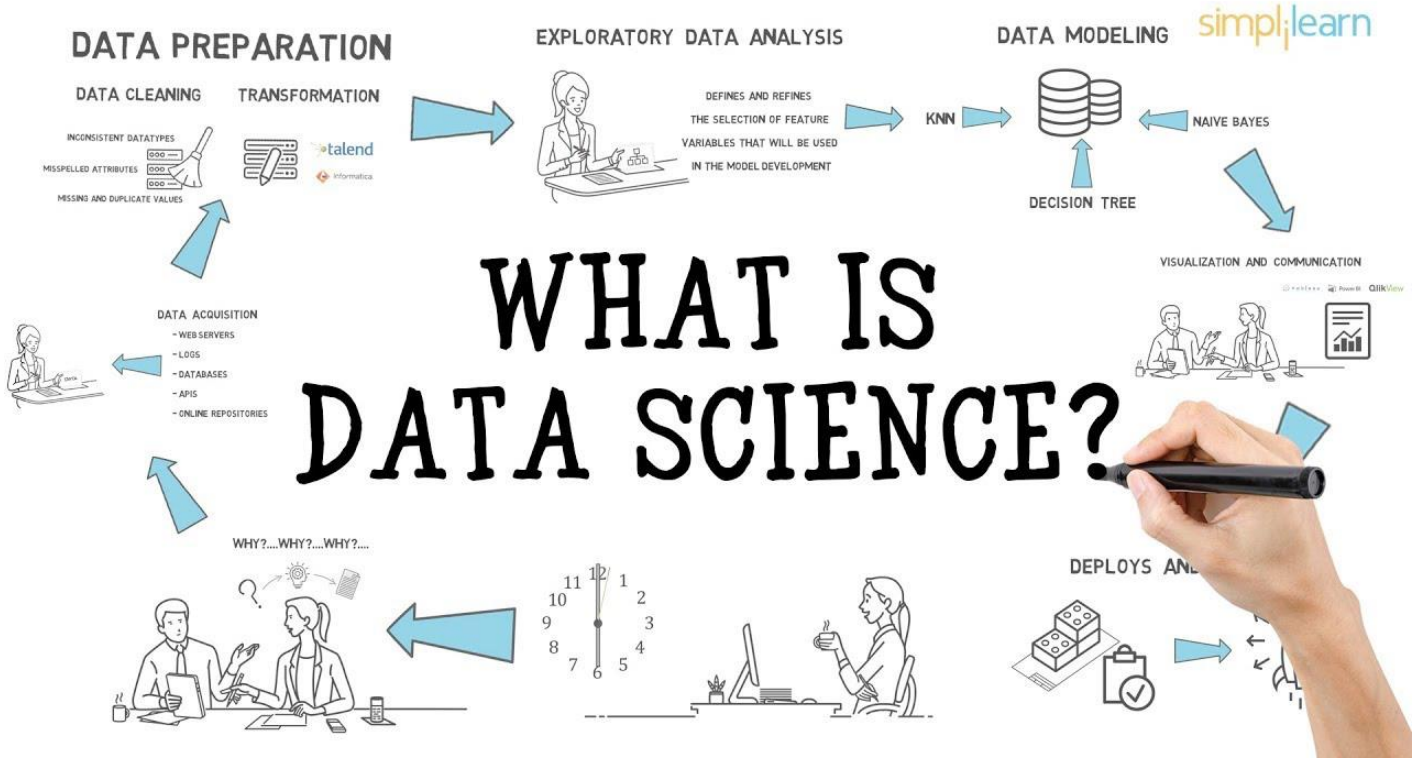


[8] Dataconomy 2016



3. DATA SCIENCE & BIG DATA





[17] Towards Data Science 2018

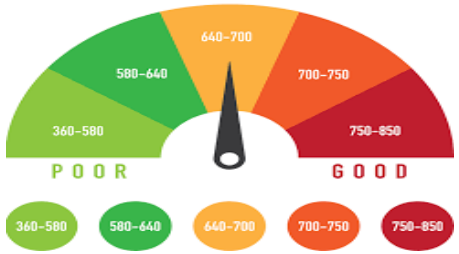
[18] SimpliLearn

3. DATA SCIENCE & BIG DATA

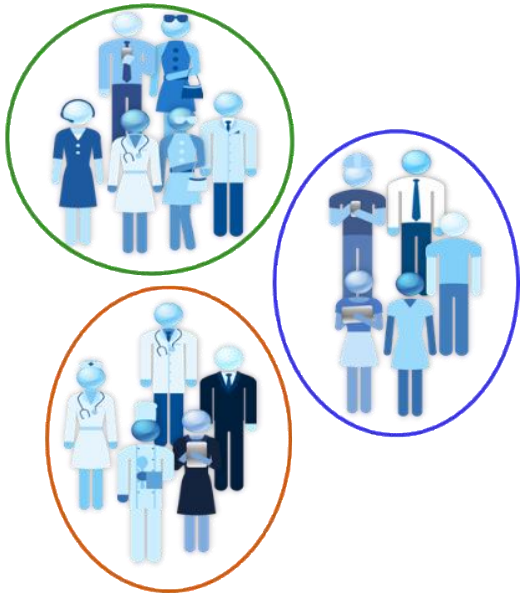
Lịch sử tín dung của user

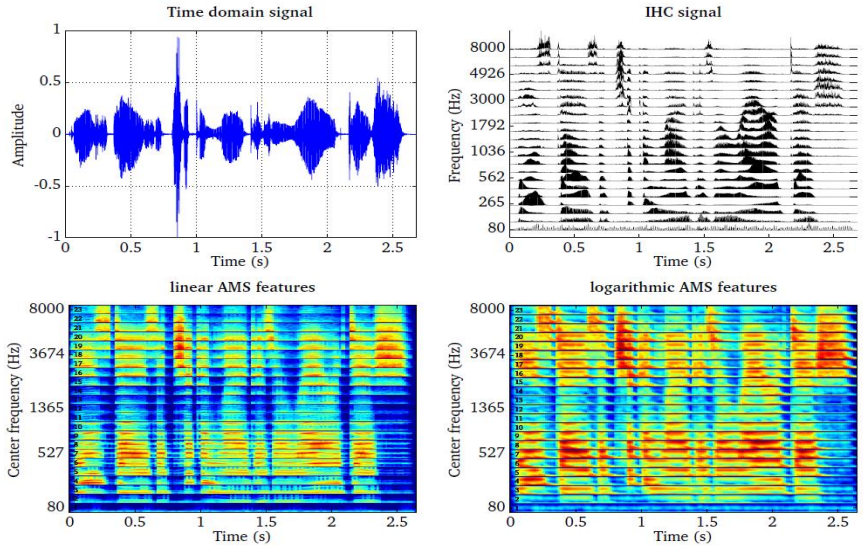
Lịch sử của gói tín dụng

Thông tin khách hàng



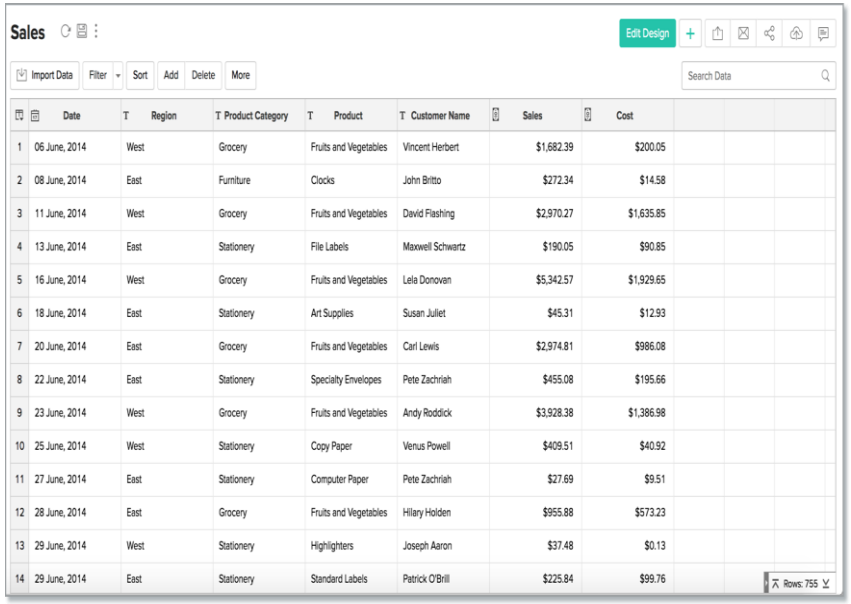
Credit scoring





TYPE OF DATA

Structural data



Sales

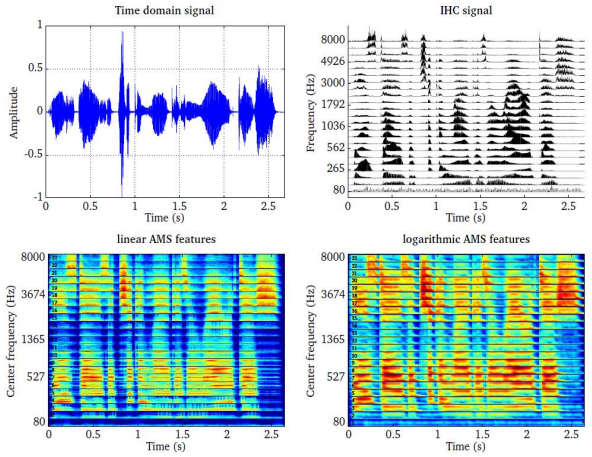
Edit Design | Import Data | Filter | Sort | Add | Delete | More | Search Data

	Date	Region	Product Category	Product	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	Lela 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 Zachriah	$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 Zachriah	$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

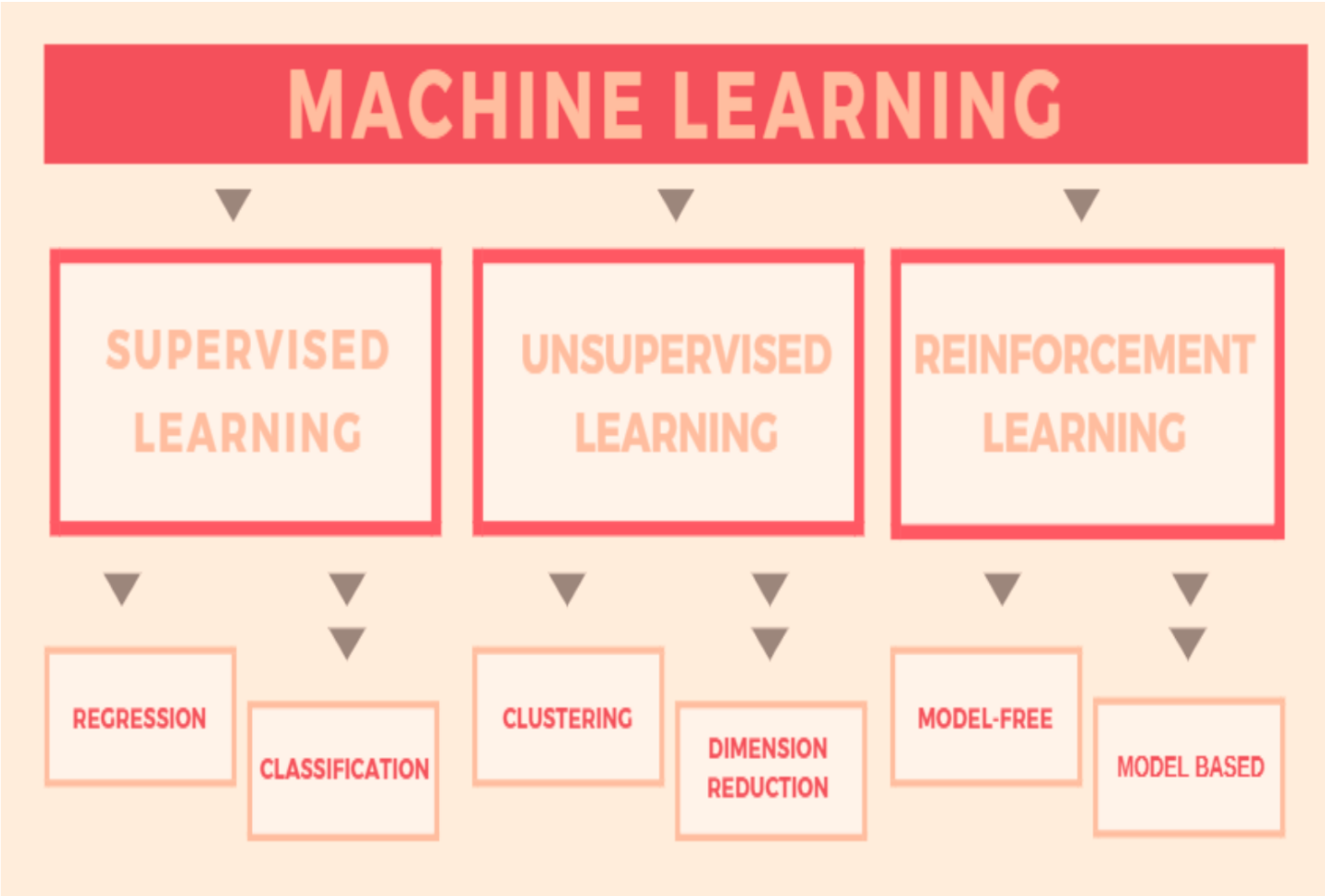
Rows: 755

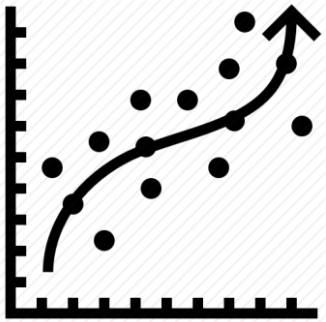
Unstructured data

and …

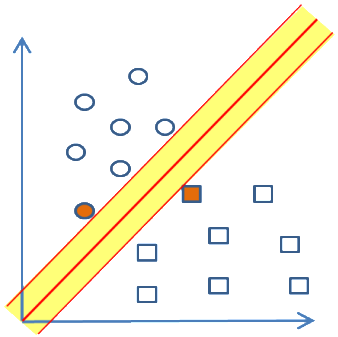


TYPE OF MACHINE LEARNING

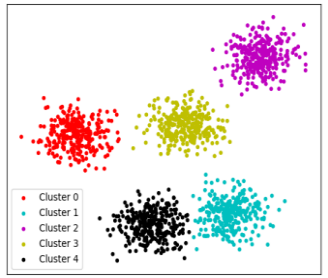




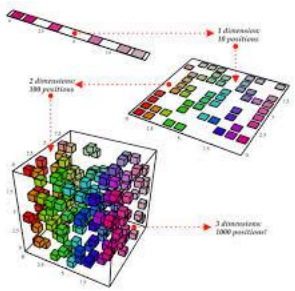
Regression
Income prediction
Credit scoring



Classification
Bad user detection
Fraud detection



Clustering
Topic modeling



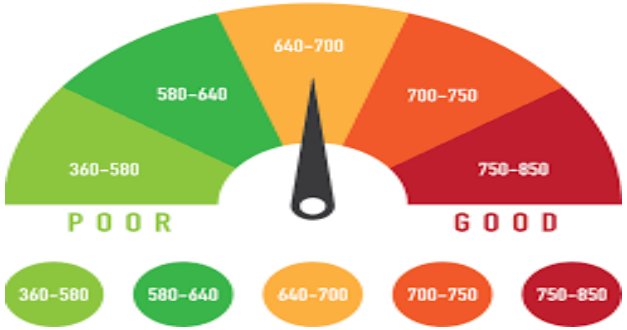
Dimension reduction
TSNE
PCA

LEARNING PROCESSING

Lịch sử tín dung của user

Lịch sử của gói tín dụng

Thông tin khách hàng



Credit scoring

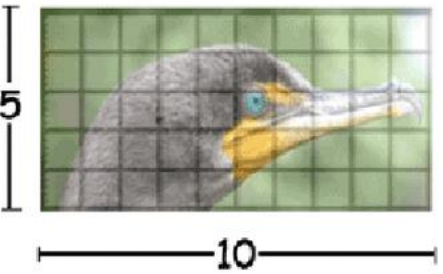
Input

"Learning"

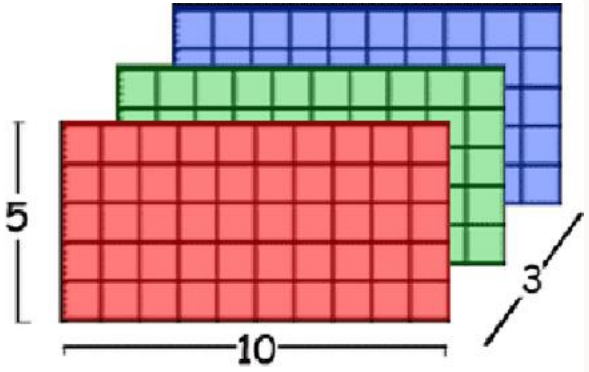
Output

"Representation Learning" or "Feature learning"

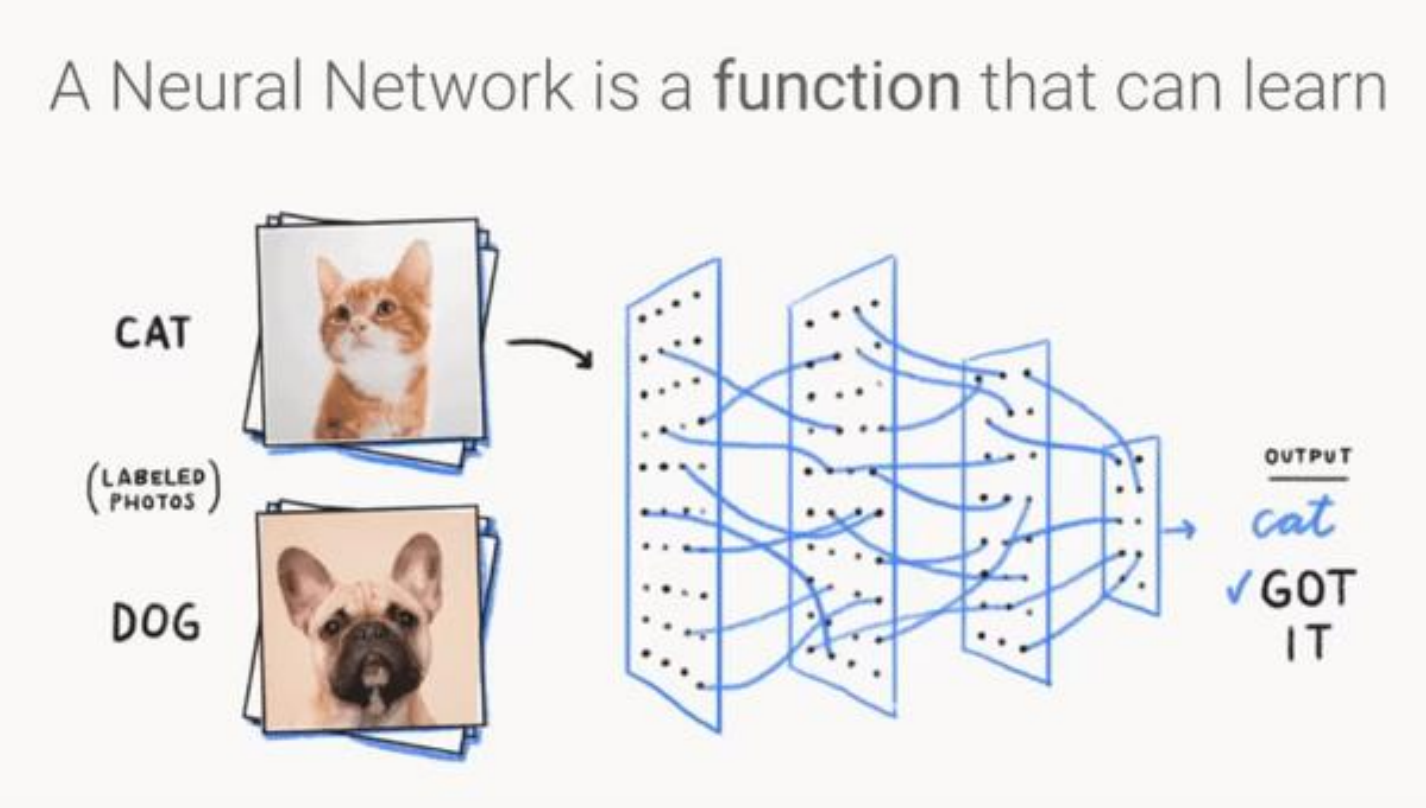
LEARNING PROCESSING



Original Color Image

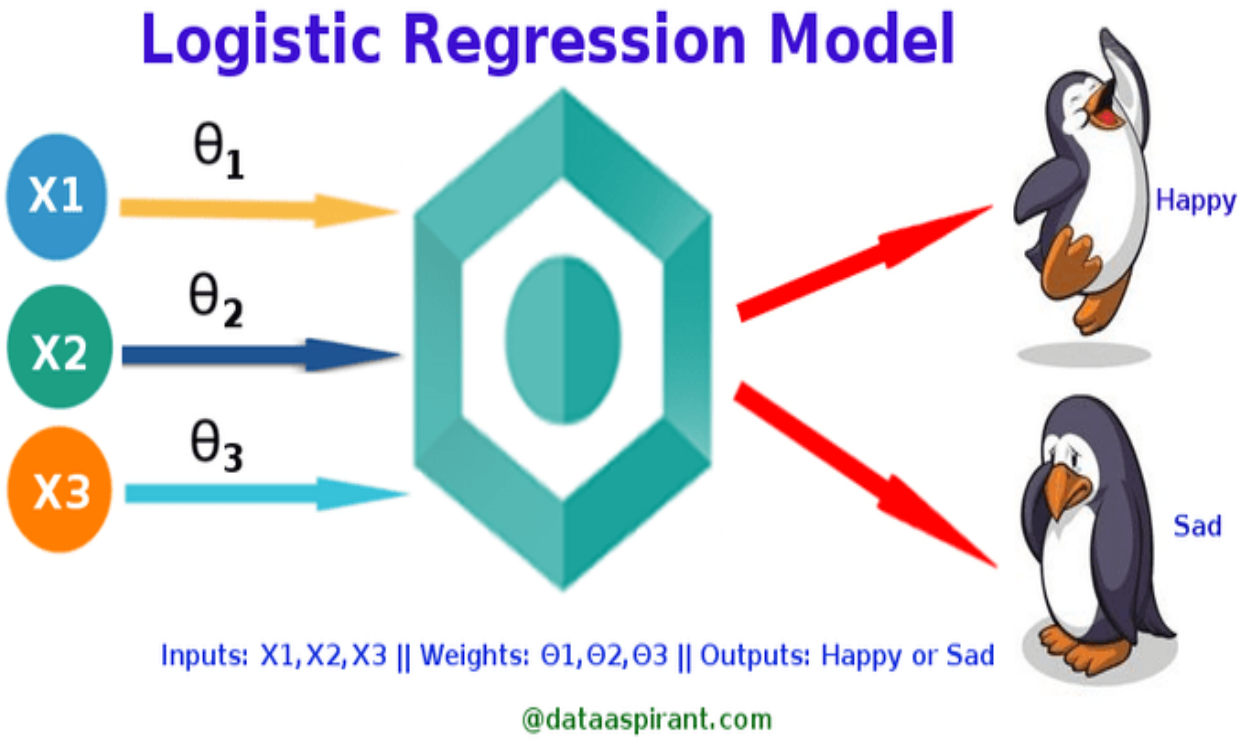


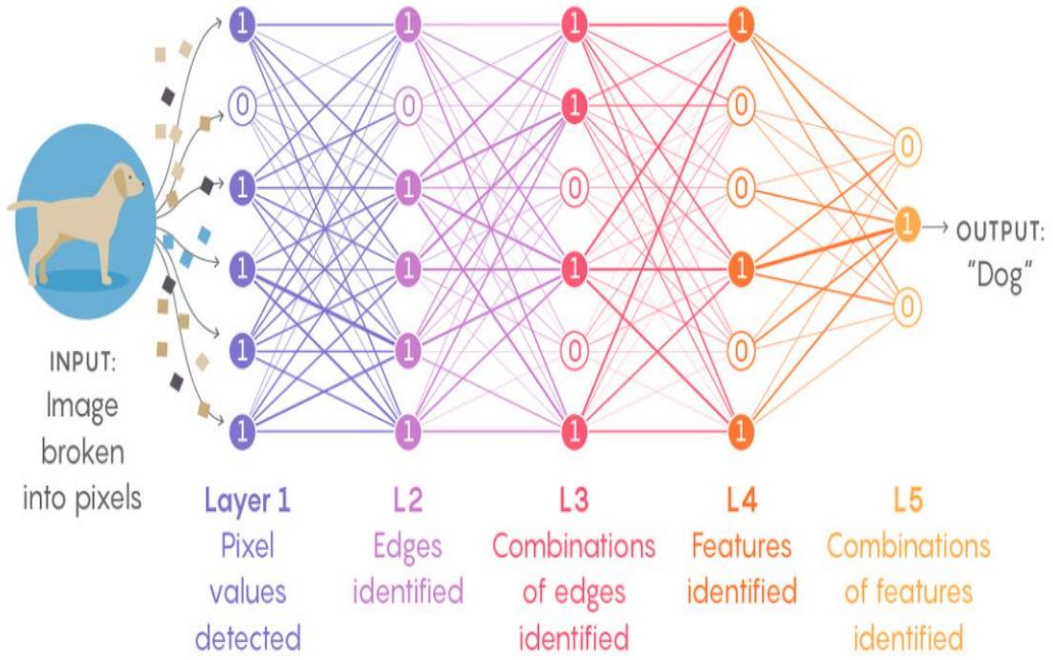
Matlab RGB Matrix



Input → "Learning" → Output

SALLOW LEARNING VS DEEP LEARNING





"Feature Engineering" or "Feature Selection"

Deep learning

LEARNING PROCESSING

TRAIN (100k loans)

Features: User behaviors
Thông tin gói vay
Thông tin tín dụng

TEST (20k loans)

20k loans

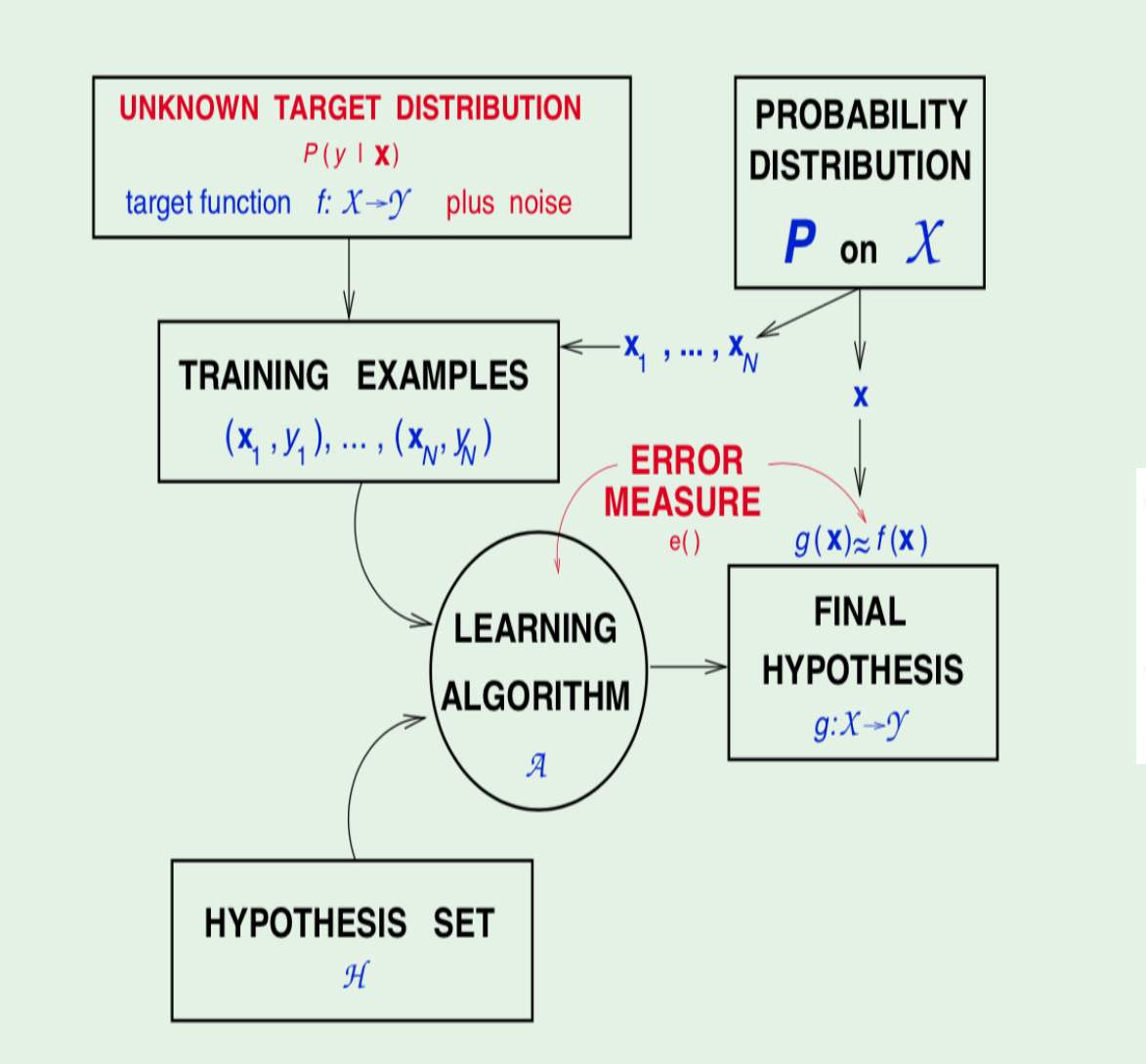
MODEL

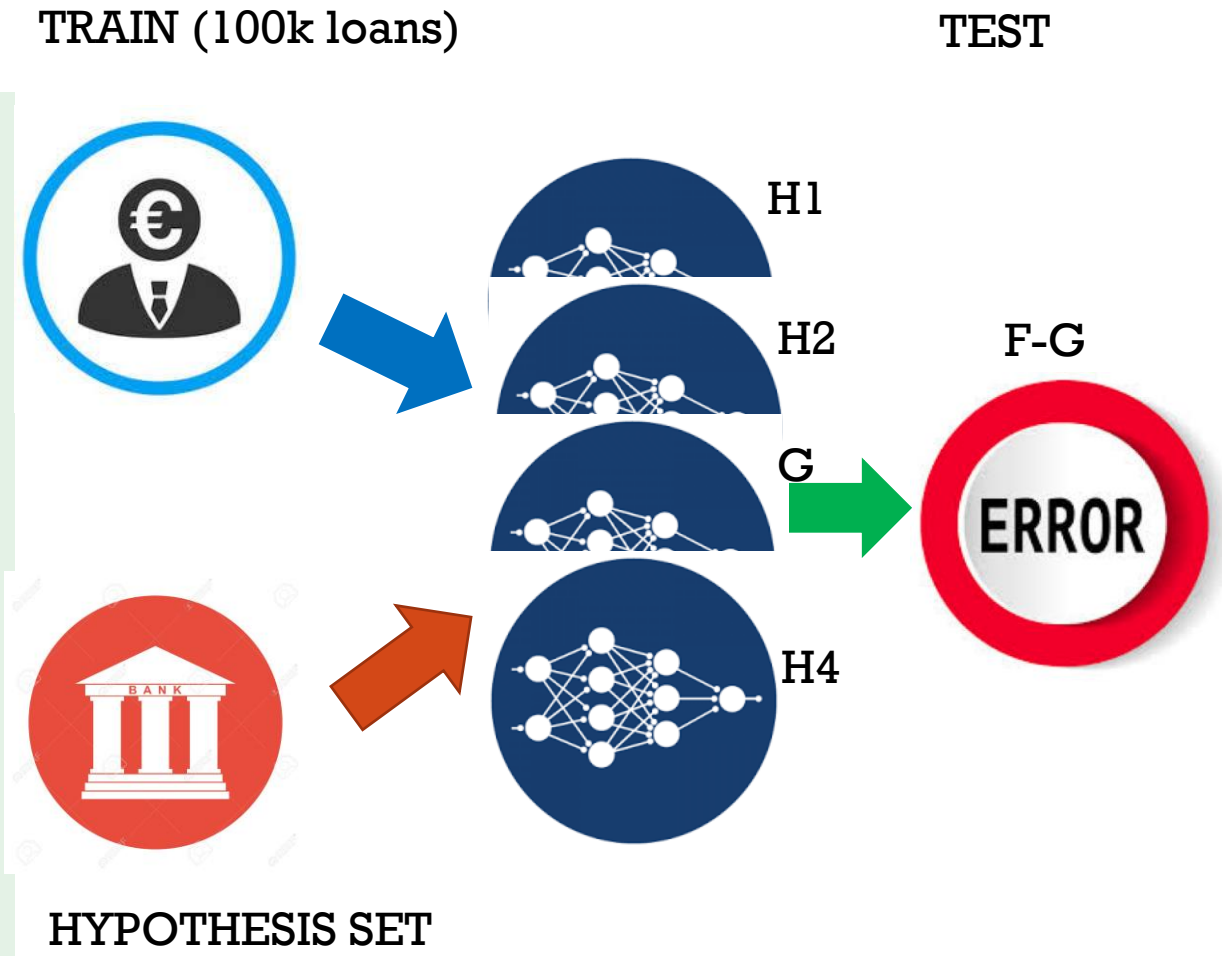
Bank
Credit Scoring

Predicted Outcome

validation

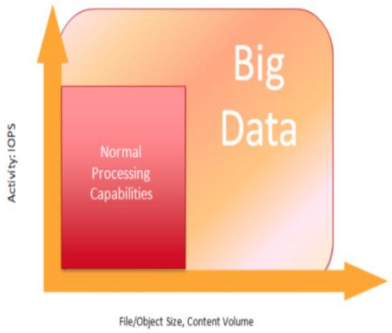
LEARNING PROCESSING

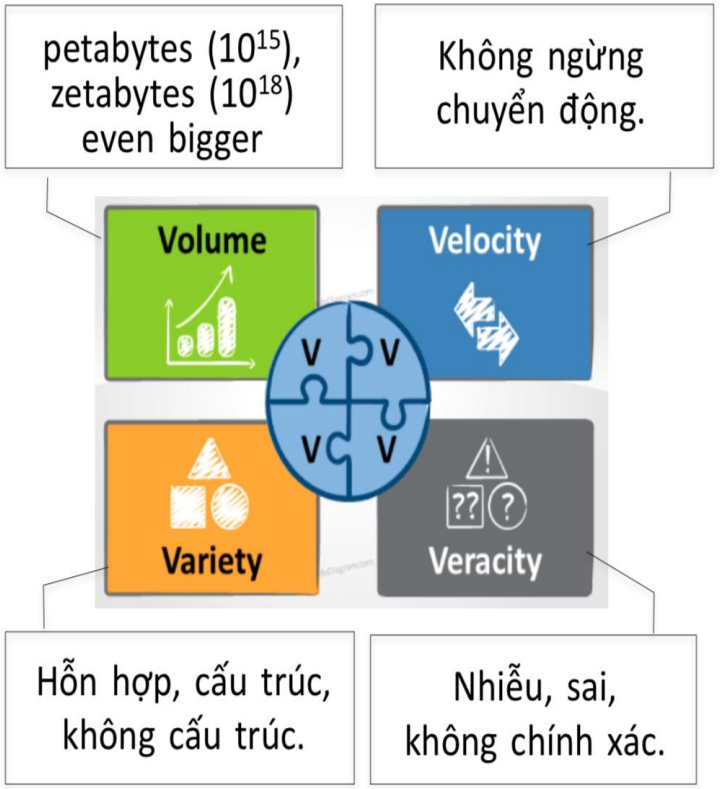


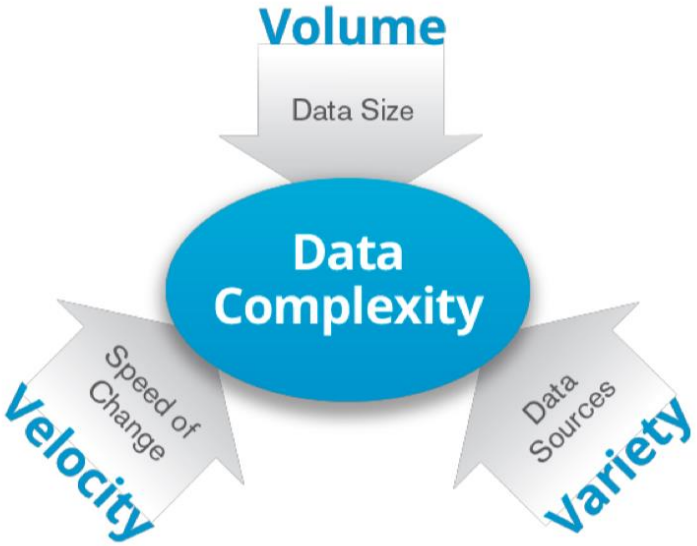


1. BIG DATA – WHAT IS BIG DATA?

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

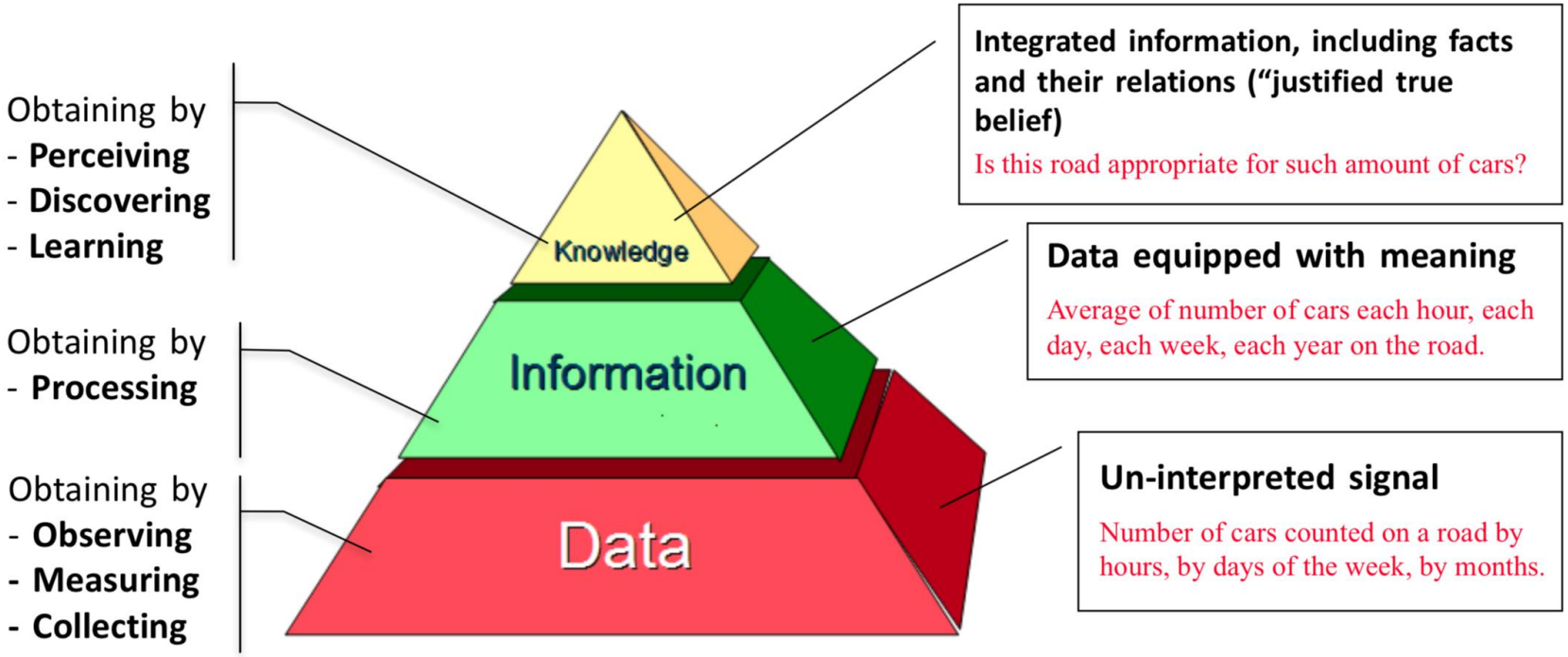
Src: [1]

Src: [5]

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 (1st days)
2. The learning problems [Caltech, Microsoft (bitshop)] (2nd day)
3. Exploratory Data Analysis – Data visualization [R] (2nd day)
4. Bias – variance trade-off. [Caltech] (3rd day)
5. Overfitting vs Underfitting [Caltech, Stanford] (3rd day)
6. Learning curve (3rd day)
7. Running model [R] (3rd day)
8. Cross Validation [Caltech, Stanford] (4rd day)
9. Regularization (4rd day)
10. Tuning [R] (4rd day)
11. Learning Principal [Caltech] (5rd day)
12. Evaluation [sonpvh] (5rd 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
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
11. Hồ Tú Bảo, Khoa học dữ lieu và cách mạng công nghiệp lần thứ 4
12. Smolan and Erwitt, The human face of big data, 2013
13. Đình Phùng, phương pháp và công nghệ dữ lieu 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_5.1intro.pdf

REFERENCES:

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