

# Thu thập và tiền xử lý dữ liệu

Thân Quang Khoát  
Nguyễn Minh Phương + Lê Minh Hoà  
Bài giảng của DSLab  
Viện nghiên cứu cao cấp về Toán (VIASM)

# Data Scientist

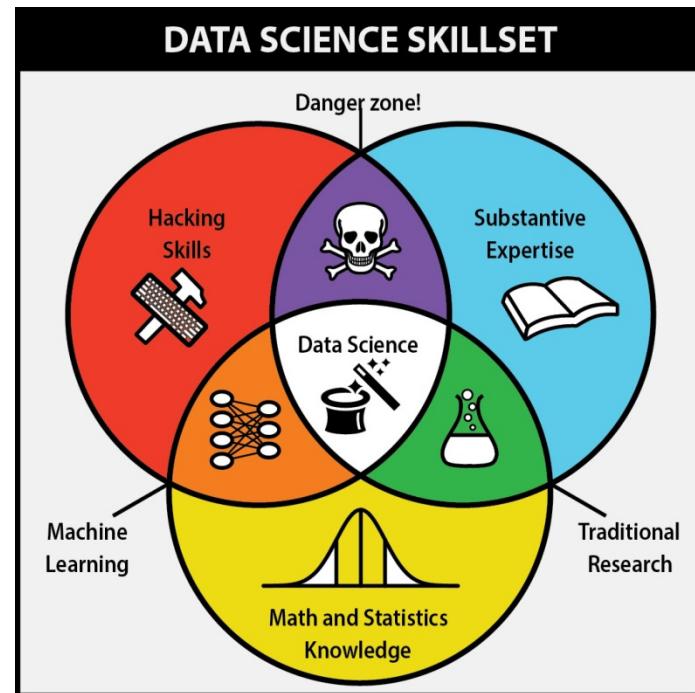
Data Scientist làm gì ?

**Statistics & Machine**

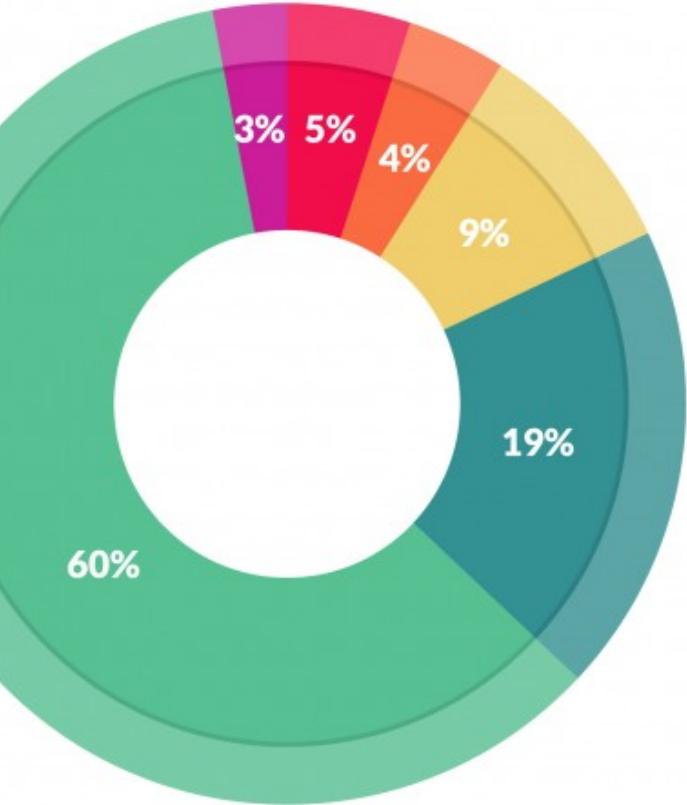
**Learning knowledge** – thu thập dữ liệu, xử lý, liên kết các thành phần dữ liệu rời rạc tạo nên ý nghĩa.

**Technology competency** – công cụ để hiện thực hóa các giải pháp, kiến thức.

**Domain expertise** – người hiểu các vấn đề có thể xảy ra, và yếu tố ảnh hưởng.



<http://berkeleysciencereview.com/how-to-become-a-data-scientist-before-you-graduate/>



*CrowdFlower Inc., 2016*

## What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

# Quỹ thời gian.

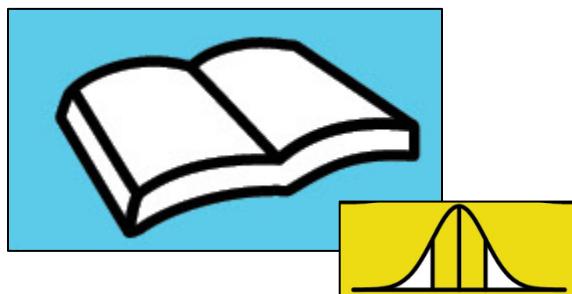
# WHY?

Tiền xử lý để làm gì ??

- Các mô hình Học máy chỉ làm việc với dữ liệu ma trận hoặc vector
- Mô hình xác suất/ mô hình học máy đạt hiệu quả
- Dễ lưu trữ / truy vấn.

## Input

Vấn đề cần giải quyết của  
lĩnh vực



## Output

Dữ liệu số - ma trận vector

$$x^{(n)} = \begin{bmatrix} -0.0920 \\ 3.4931 \\ -1.8493 \\ \dots \\ \dots \\ -0.2010 \\ -1.3079 \end{bmatrix}$$

$$\mathcal{D} = \begin{bmatrix} -x^{(1)} & - \\ -x^{(2)} & - \\ \dots & \\ -x^{(n)} & - \\ \dots & \end{bmatrix}$$

# Content (i.e. HOW?)

---

- **Thu thập dữ liệu**
  - *Lấy mẫu (sampling)*
  - *Kỹ thuật: crawling, logging, scraping*
- **Xử lý dữ liệu**
  - *Dữ liệu cần lọc nhiễu, số hóa.*
  - *Kỹ thuật – làm sạch, số hóa, lưu trữ.*

# Thu thập dữ liệu

**Input**  
Vấn đề cần giải quyết



**Output**  
Mẫu dữ liệu

A screenshot of a Wikipedia page titled "Cavendish Laboratory". It features a table of data comparing countries based on population and under-15 age groups, and several biological illustrations including a cell diagram, a DNA helix, and a gene structure.

# Fundamentals :: Sampling

- **WHAT** – lấy tập mẫu nhỏ, phổ biến để đại diện cho lĩnh vực cần học.
- **HOW** – thu thập các mẫu từ thực tế, hoặc các nguồn chứa dữ liệu web, database, ..
- **WHY** – không thể học toàn bộ. Giới hạn về thời gian và khả năng tính toán

*"One or more small spoon(s) can be enough to assess whether the soup is good or not."*



<https://www.coursera.org/learn/inferential-statistics-intro>

# Fundamentals :: Sampling :: HOW

- **Variety** – tập mẫu thu được đủ đa dạng để phủ hết các ngũ cẩm của lĩnh vực.
- **Sample size** – tổng số mẫu, thực thể,.. thu được
- **Biases** – dữ liệu cần tổng quát, không bị sai lệch, thiên vị về 1 bộ phận nhỏ nào đó của lĩnh vực.

*"One or more small spoon(s) can be enough to assess whether the soup is good or not."*

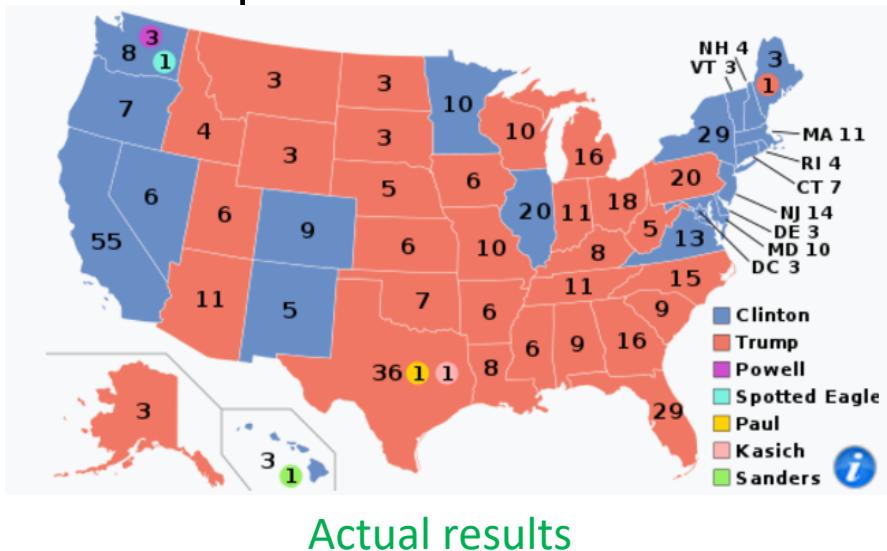
*Remember to stir to avoid tasting biases.*



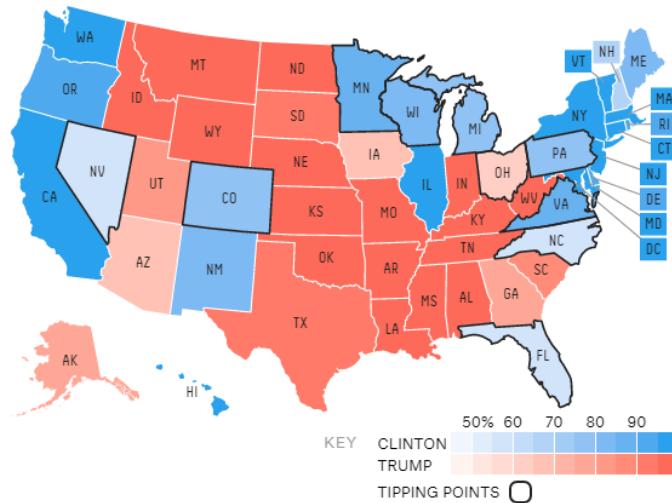
<https://www.coursera.org/learn/inferential-statistics-intro>

# Fundamentals :: Sampling :: HOW

- **Variety** – các mẫu đủ đa dạng để phản ánh khách quan ?



<https://projects.fivethirtyeight.com/2016-election-forecast/>  
<http://edition.cnn.com/election/results/president>  
Image credit: Wikipedia, FiveThirtyEight



## Electoral votes

Hillary Clinton 302.2

Donald Trump 235.0

232

306

## Popular vote

Hillary Clinton 48.5%

Donald Trump 44.9%

48.2%

46.1%

# Techniques

---

- **Crowd-sourcing:** Survey – *thực hiện các khảo sát*
- **Logging** vd: lưu lại lịch sử tương tác của người dùng, truy cập sản phẩm,...
- **Scraping** tìm kiếm nguồn dữ liệu trên các website, tải về bóc tách, lọc ...

## Techniques :: Scrapping :: DEMO

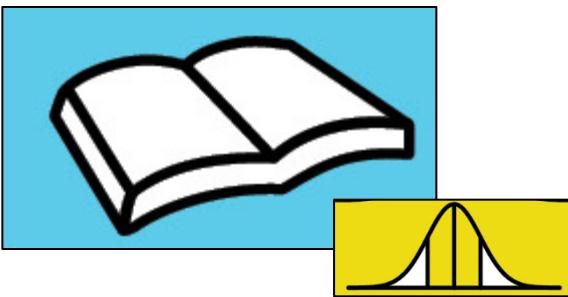
---

- Mục tiêu: Dữ liệu cho bài toán phân loại văn bản – miền báo chí.
- DEMO: Hệ thống crawl dữ liệu báo

# DEMO

## Input

Vấn đề phân loại văn  
bản báo chí



## Output

Mẫu dữ liệu báo chí và  
nhân tương ứng

The image shows a Windows file explorer window on the left displaying a folder named 'Dân trí' containing various files. On the right, a JSON file is being viewed in a code editor. The JSON file contains a list of documents with metadata such as date, code, labels, content, url, domain, and title. A red box highlights the first document in the list.

Name	Date modified
2ce54c553499dc5fb9a71533995793c6a648f...	5/25/2018 4:46 PM
7b2288470f3349971fc590f76de1b0e6b5a9...	5/25/2018 4:46 PM
8a0f8828443701ee020424acdefef880c0fc9...	5/25/2018 4:46 PM
949342b0d858be7b06b26d1e94d07917e...	5/25/2018 4:46 PM
146fd8057df18632a70e12bc424287655604d...	5/25/2018 4:46 PM
651ab245f0305220d1f57bb21913620f75d128d.json	5/25/2018 4:46 PM
a1f0115782578af4b3773a79fbcc5d2d94f...	5/25/2018 4:46 PM
c6bd8d1552a3d7b3a73acd5798c593db61f9...	5/25/2018 4:46 PM
e0efcbc74a5882c6765077448ed7dccd60d...	5/25/2018 4:46 PM
e43e36696d67647494fcabf0a812d169e9b...	5/25/2018 4:46 PM

# DEMO :: Steps

Rss

## Kênh do VnExpress cung cấp

Trang chủ



Thời sự



Thế giới



Kinh doanh



Startup



Giải trí



Thể thao



Pháp luật



Giáo dục



Item

```
<rss xmlns:sisasn="http://purl.org/rss/1.0/modules/sisasn/" version="2.0">
  <channel>
    <title>Kinh doanh - VnExpress RSS</title>
    <description>VnExpress RSS</description>
    <image>
      <url>
        https://s.vnecdn.net/vnexpress/i/v20/logos/vne_logo_rss.png
      </url>
      <title>Tin nhanh VnExpress - Đọc báo, tin tức online 24h</title>
      <link>https://vnexpress.net/link</link>
    </image>
    <pubDate>Thu, 07 Jun 2018 20:40:44 +0700</pubDate>
    <generator>VnExpress</generator>
    <link>https://vnexpress.net/rss/kinh-doanh.rss</link>
  </channel>
</rss>
```

Content

```
<article class="content_detail fck_detail width_common block_ads_connect">
  <p class="Normal">
    <span>
      Công ty TNHH MTV Xổ số điện toán Việt Nam (Vietlott) vừa trao giải cho khách hàng trúng Jackpot 1 sản phẩm Power 6/55 trị giá hơn 40 tỷ đồng (chưa trừ thuế) chiều ngày 7/6.
    </span>
  </p>
  <p class="Normal">
    <span>
      "Nữ khách hàng may mắn trúng giải tên N.T, là nhân viên một ngân hàng tại TP HCM. Chị sẽ tái buổi trao thưởng,"&nbsp;
    </span>
  </p>
  <table align="center" border="0" cellpadding="3" cellspacing="0" class="tblCaption" style="width: 100%;"></table>
  <p class="Normal">
    <span>
      Theo thông tin từ Vietlott, chi nhánh TP HCM của đơn vị này đã tiếp nhận chiếc vé trúng giải Jackpot 1 Power 6/55 từ một nữ khách hàng ngày 4/6.
    </span>
  </p>
  <p class="Normal"> == $0
    <span>
      "Qua kiểm tra trên hệ thống kỹ thuật và hồ sơ kèm theo, Vietlott xác định chiếc vé của chị N.T là hợp lệ và trúng giải Jackpot 1 Power 6/55 kỳ quay thứ 131. Tấm vé được phát hành tại điểm bán hàng đường số 6, phường Linh Chiểu, quận Thủ Đức, TP HCM."
    </span>
  </p>
  <p class="Normal">...</p>
  <p class="Normal">...</p>
</article>
```

# DEMO :: Sample

---

```
JSON
  └─ date : "2018-05-20, 07:44:00-07:00"
  └─ code : "651ab2f45f0305220d1f57bb21913620f75d128d"
  └─ labels : "Dân trí/Bạn đọc"
  └─ content : "Dân trí Sau khi Bí thư Tỉnh ủy Bắc Giang yêu cầu dẹp tan nạn xe quá tải trong năm 2018, Phòng CSGT Công an tỉnh Bắc Giang đã tổ chức ra quân
  └─ image_url : "https://dantricdn.com/zoom/80_50/2018/5/20/7-1526776517717498023080.png"
  └─ url : "http://dantri.com.vn/ban-doc/bac-giang-doan-xe-coi-noi-thung-ram-rap-chayqua-mat-canh-sat-giao-thong-20180520074415778.htm"
  └─ domain : "dantri.com.vn"
  └─ title : "Bắc Giang: Đoàn xe coi nón thùng rầm rập chạy qua mặt cảnh sát giao thông?"
```

# Data preprocessing

## Input

Mẫu dữ liệu thô (text, ảnh, audio, ...)



Dữ liệu số theo từng ML/AI model(s)

$$x^{(n)} = \begin{bmatrix} -0.0920 \\ 3.4931 \\ -1.8493 \\ \dots \\ \dots \\ -0.2010 \\ -1.3079 \end{bmatrix}$$
$$\mathcal{D} = \begin{bmatrix} -x^{(1)} \\ -x^{(2)} \\ \dots \\ -x^{(n)} \end{bmatrix}$$

# Fundamentals :: Data “rawness”

## Completeness (đầy đủ)

Từng mẫu thu thập nên đầy đủ thông tin các trường thuộc tính cần thiết

## Integrity (rõ ràng)

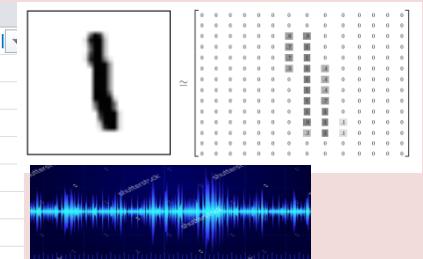
- Nguồn thu thập chính thống, đảm bảo mẫu thu được chứa giá trị chính xác trên thực tế.
- Jan. 1 as everyone's birthday? – *intentional (systematic) noises*

## Homogeneity (đồng nhất)

- Rating “1, 2, 3” & “A, B, C”; or Age = “42” & Birthday = “03/07/2010” (*inconsistency*)
- Heterogenous data sources / schemas

## Structures & Semantics (cấu trúc & ý nghĩa)

C	D	E	F
Population	Under15	Over60	Fertil
13724	40.24	5.68	3.64
14075	46.73	3.95	5.77
23852	40.72	4.54	4.35
90796	22.87	9.32	1.79
29955	28.84	9.17	2.44
247	37.37	6.02	3.46
28541	28.9	6.38	2.38
3395	22.05	18.59	2.07



Jiawei Han, Data Mining: Concepts & Techniques VI(RSM)

Image credit: [http://colah.github.io/posts/2014-10-Visualizing-MNIST\\_shutterstock](http://colah.github.io/posts/2014-10-Visualizing-MNIST_shutterstock)

Vietnam Institute for  
Advanced Study in Mathematics

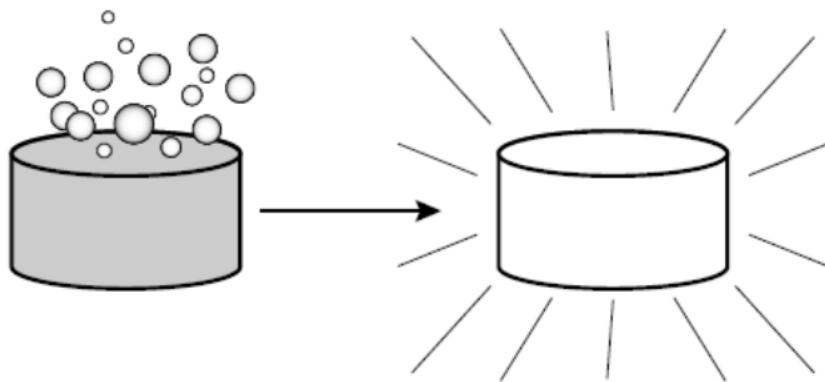
# Techniques

---

Cleaning  
Integrating  
Transforming

# Techniques :: Cleaning

## Tính đầy đủ



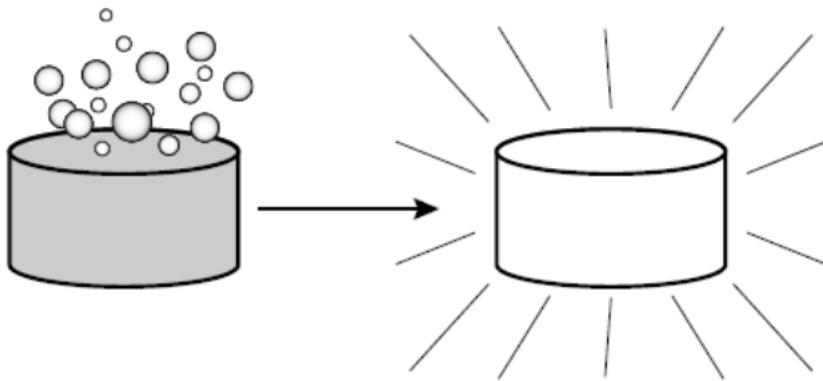
Mẫu dữ liệu thu thập đôi khi không thể đầy đủ, cần có chiến lược phù hợp:

- Bỏ qua, không đưa vào dữ liệu học.
- Bổ sung các trường còn thiếu cho mẫu:
  - Bằng tay
  - Tự động (heuristic)



# Techniques :: Cleaning (cont.)

Tính rõ ràng, ít nhiễu

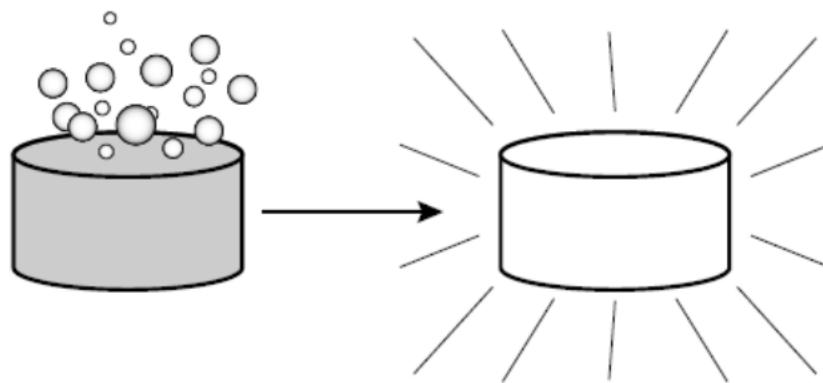


- random noises
  - systematic noises (technical)
  - intentional noises
- collect/synthesize and label *negative data*

- Thủ công, tùy thuộc vào từng loại nhiễu, cần tổng hợp và loại bỏ.
- Một số phương pháp tự động sử dụng ML
  - Làm mịn dữ liệu sử dụng regression
  - Loại bỏ một số ngoại lệ: vd.: Clustering, ...

# Techniques :: Cleaning (cont.)

## Tính đồng nhất



Các mẫu dữ liệu cần có tính đồng nhất về cách biểu diễn, ký hiệu.

Vd: Rating “1, 2, 3” & “A, B, C”; Age = “42” & Birthday = “03/07/2010”

tính **đồng nhất** của dữ liệu

# Techniques :: Integrating w/ some Transforming

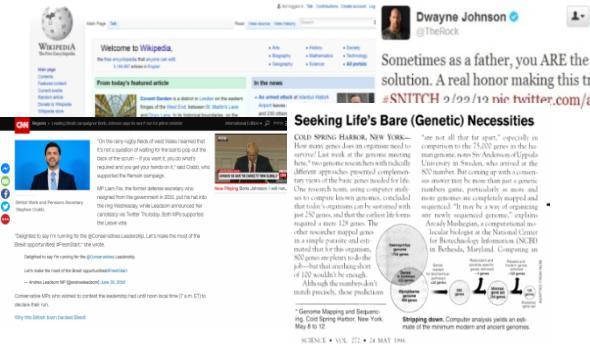
Structured – relational (table-like)

	A	B	C	D	E	F	G
1	Country	Region	Population	Under15	Over60	Fertil	LifeExp
2	Zimbabwe	Africa	13724	40.24	5.68	3.64	54
3	Zambia	Africa	14075	46.73	3.95	5.77	55
4	Yemen	Eastern M	23852	40.72	4.54	4.35	64
5	Viet Nam	Western P	90796	22.87	9.32	1.79	75
6	Venezuela	(Bo Americas	29955	28.84	9.17	2.44	75
7	Vanuatu	Western P	247	37.37	6.02	3.46	72
8	Uzbekistan	Europe	28541	28.9	6.38	2.38	68
9	Uruguay	Americas	3395	22.05	18.59	2.07	77

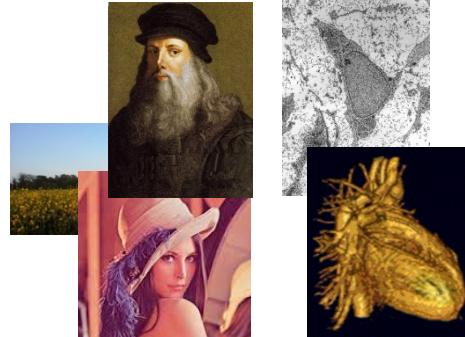
Un-structured  
Semi-structured – XML, JSON, HDF

```
{
  "code": "1473a6fd39d1d8fa48654aac9d8cc2754232",
  "title": "[Updating] Câu chuyện xuyên mưa về :",
  "url": "http://techtalk.vn/updating-cau-chuyen",
  "labels": "techtalk/Cong nghe",
  "content": "Vào chiều tối ngày 09/12/2016 vừa",
  "image_url": "",
  "date": "2016-12-10T03:51:10Z"
}
```

texts in websites, emails, articles, tweets



2D/3D images, videos + meta



spectrograms, DNAs, ...

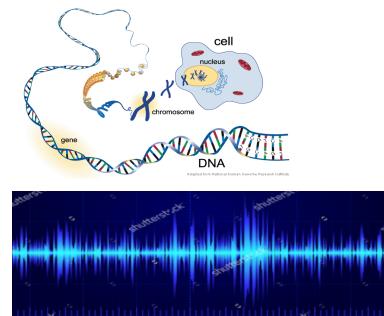


image credits: wikipedia, shutterstock, CNN

Vietnam Institute for  
Advanced Study in Mathematics

# *Extra – Math!*

represent data by language of math

Data points

Vector, Matrix

# Techniques :: Transforming

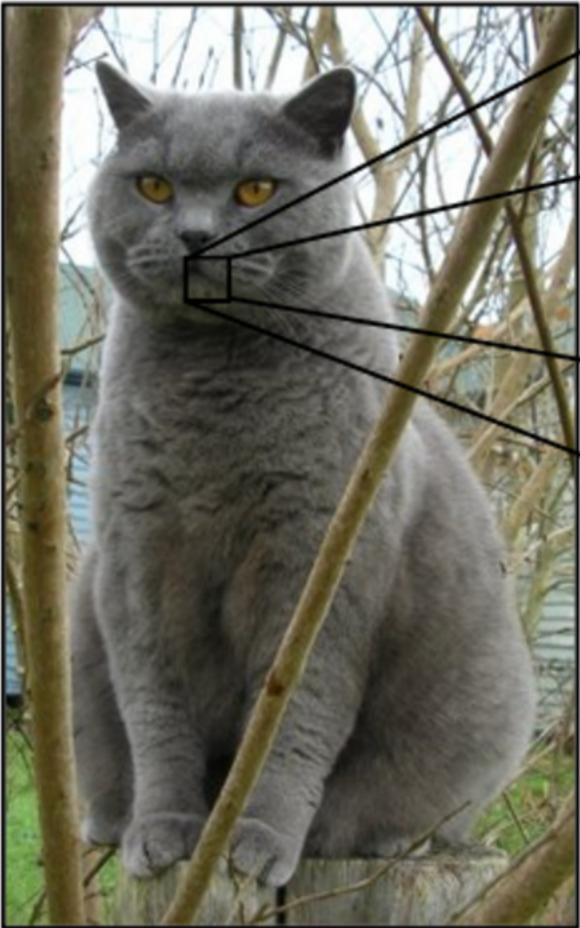
---

Semantics?

Trích xuất các đặc trưng ngữ nghĩa, chuẩn hóa

# Semantics

## example: visual data



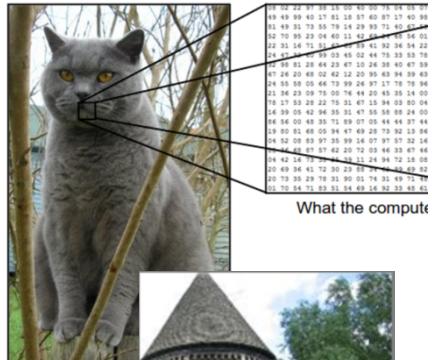
08	02	22	97	38	15	00	40	00	75	04	05	07	78	52	12	50	77	91	66
49	49	99	40	17	81	18	57	60	87	17	40	98	43	69	48	61	56	62	00
81	49	31	73	55	79	14	29	93	71	40	67	55	88	30	03	49	13	36	65
52	70	95	23	04	60	11	42	69	21	68	56	01	32	56	71	37	02	36	91
22	31	16	71	51	67	03	89	41	92	36	54	22	40	40	28	66	33	13	80
24	47	34	00	99	03	45	02	44	75	33	53	78	36	84	20	35	17	12	50
32	98	81	28	64	23	67	10	26	38	40	67	59	54	70	66	18	38	64	70
67	26	20	68	02	62	12	20	95	63	94	39	63	08	40	91	66	49	94	21
24	55	58	05	66	73	99	26	97	17	78	78	96	83	14	88	34	89	63	72
21	36	23	09	75	00	76	44	20	45	35	14	00	61	33	97	34	31	33	95
78	17	53	28	22	75	31	67	15	94	03	80	04	62	16	14	09	53	56	92
16	39	05	42	96	35	31	47	55	58	88	24	00	17	54	24	36	29	85	57
86	56	00	48	35	71	89	07	05	44	44	37	44	60	21	58	51	54	17	58
19	80	81	68	05	94	47	69	28	73	92	13	86	52	17	77	04	89	55	40
04	52	08	83	97	35	99	16	07	97	57	32	16	26	26	79	33	27	98	66
03	46	68	87	57	62	20	72	03	46	33	67	46	55	12	32	63	93	53	69
04	42	16	73	38	84	39	11	24	94	72	18	08	46	29	32	40	62	76	36
20	69	36	41	72	30	23	88	31	62	99	69	82	67	59	85	74	04	36	16
20	73	35	29	78	31	90	01	74	31	49	71	48	04	81	16	23	57	05	54
01	70	54	71	83	51	54	69	16	92	33	48	61	43	52	01	89	31	67	48

What the computer sees

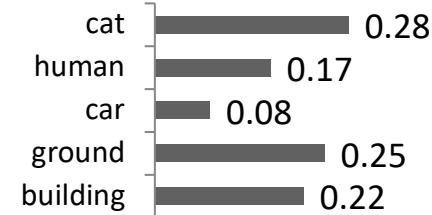
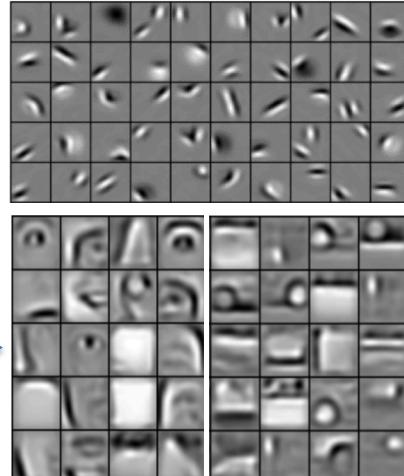
# Semantics

## example: visual data

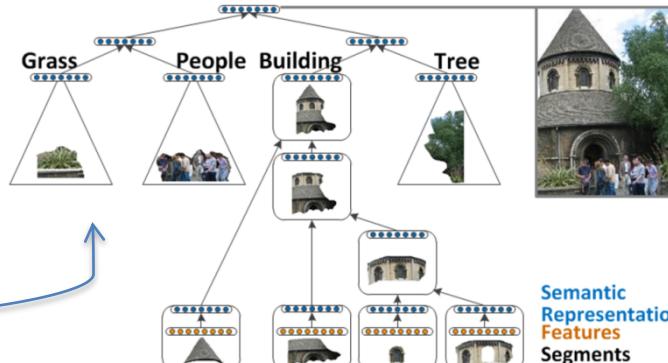
Low-level semantics  
(raw features)



Mid-/High-level semantics  
(e.g. human-interpretable features)



cat → not on → car  
people ← behind ← building  
car → is → red

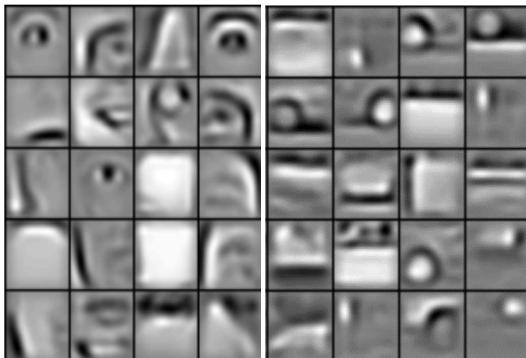


Mức ngữ nghĩa tối thiểu để có thể hiểu:  

- Phân loại văn bản
- Phân tích cảm xúc
- AI Chatbot (nhiều mức ngữ nghĩa khác nhau)

# Techniques :: Transforming (cont.)

Mục tiêu: trích xuất các **đặc trưng ngữ nghĩa** của vấn đề.



- Từng lĩnh vực cụ thể, từng loại dữ liệu sử dụng các kỹ thuật xuất đặc trưng ngữ nghĩa khác nhau (dữ liệu text, hình ảnh, ...)

B	C	D	E	F	G	
Region	P	Populat	Under1	Over60	Fertil	LifeExp
Africa		-0.416	0.748	-0.483	0.299	54
Africa		-0.403	1.464	-0.850	1.881	55
Eastern M		-0.060	0.801	-0.725	0.826	64
Western P		2.287	-1.169	0.289	-1.075	75
Americas		0.154	-0.511	0.257	-0.592	75
Western P		-0.888	0.431	-0.411	0.165	72
Europe		0.104	-0.504	-0.334	-0.637	68
Americas		-0.778	-1.260	2.256	-0.867	77

One-hot encoding

$$\begin{aligned}1 &= [1 \ 0 \ 0 \ 0 \ 0] \\3 &= [0 \ 0 \ 1 \ 0 \ 0]\end{aligned}$$

...

$$\frac{x - \bar{x}}{s}$$

## Techniques :: Transforming example & demo

---

Transforming text data

# DEMO

---

## Input

Mẫu dữ liệu thô: json text

```
{  
    "image_url": "https://i-kinhdoanh.vnecdn.net/2018/1",  
    "url": "https://kinhdoanh.vnexpress.net/tin-tuc/eb",  
    "title": "Sacombank n\u00e2n c\u00f3m t\u00f3n t\u00f3n v\u00e0o t\u00f3n t\u00f3n",  
    "code": "db274d03b9a61aa16d70c7fd68929d799058b866",  
    "domain": "kinhdoanh.vnexpress.net",  
    "date": "2018-05-25, 17:00:00+07:00",  
    "content": "\nHi\u00e1c\u00e1n th\u00e1nh ph\u00e1t tri\u00e8n kinh doanh v\u00e0 t\u00f3n t\u00f3n",  
    "labels": "vnexpress/Kinh doanh/Ebank\u00a0/Kinh doanh",  
}
```

## Output

Dữ liệu số theo từng ML/AI model(s)

```
{'content': [206, 207, 186, 208, 209, 210, 26, 211, 207, 212, 7, 207, 148, 94, 207, 213, 214, 23, 26, 215, 28, 10, 216, 217, 89, 148, 7, 58, 218, 219, 7, 220, 221, 7, 64, 222, 110, 223, 224, 225, 114, 226, 227, 228, 207, 229, 23, 207, 45, 230, 214, 7, 148, 41, 231, 232, 233, 234, 235, 7, 236, 227, 237, 34, 238, 239, 7, 240, 241, 242, 243, 224, 244, 245, 207, 181, 148, 246, 7, 247, 38, 248, 249, 106, 250, 251, 207, 66, 251, 252, 10, 253, 254, 255, 256, 3, 167, 257, 208, 209, 5, 258, 26, 259, 260, 261, 178, 7, 21, 158, 262, 181, 230, 263, 264, 265, 266, 267, 38, 254, 268, 23, 208, 25, 269, 114, 270, 225, 2, 271, 245, 207, 58, 106, 267, 246, 94, 135, 225, 136, 272, 69, 273, 274, 19, 100, 275, 276, 207, 23, 277, 278, 279, 280, 7, 47, 281, 110, 282, 7, 208, 283, 284, 285, 286, 287, 69, 278, 279, 280, 63, 288, 207, 94, 10, 230, 270, 47, 289, 290, 291, 292, 235, 47, 63, 293, 207, 23, 245, 207, 262, 64, 294, 266, 295, 237, 105, 296, 40, 47, 289, 290, 7, 144, 297, 298, 299, 300, 301, 234, 302, 94, 292, 303, 304, 23, 305, 266, 295, 237, 306, 145, 63, 86, 111, 40, 148, 251, 207, 19, 307, 258, 260, 7, 266, 267, 66, 255, 256, 23, 178, 7, 0, 167, 308, 207, 213, 214, 309, 208, 209, 105, 26, 310, 311, 276, 312, 278, 279, 94, 231, 223, 224, 313, 314, 23, 315], 'label': 0}
```

# DEMO :: Steps

Tokenize

Dictionary

Data Input

Hiện thẻ quốc tế Sacombank Visa gồm các dòng thẻ tín dụng, thẻ thanh toán và thẻ trả trước. Các sản phẩm này có tiện ích chung nhu thanh toán, rút tiền khắp thế giới, mua sắm trực tuyến, nhận giảm giá đến 50% tại hàng trăm điểm chấp nhận thẻ liên kết. Thẻ hỗ trợ chi tiêu trước, thanh toán sau miễn lãi tối đa 55 ngày, tích lũy điểm thường để đổi quà, mua hàng trả góp lãi suất 0%...

Chủ thẻ có thể thanh toán nhanh chóng, thuận tiện trên phạm vi toàn cầu bằng cách chạm thẻ hoặc chạm điện thoại có cài ứng dụng Samsung Pay (đồng thời tích

['Hiện', 'thẻ', 'quốc tế', 'Sacombank', 'Visa', 'gồm', 'các', 'dòng', 'thẻ', 'tín dụng', ',', 'thẻ', 'thanh toán', 'và', 'thẻ', 'trả', 'trước', ',', 'các', 'sản phẩm', 'này', 'có', 'tiện ích', 'chung', 'như', 'thanh toán', ',', 'rút tiền', '50%', '%', 'tai', 'hàng', 'trảm', 'diễn', 'chấp nhận', 'thẻ', 'liên kết', ',', 'tối đa', '55', 'ngày', ',', 'tích lũy', 'diễn', 'thường', 'để', 'đổi', 'qua', ',', 'mua hàng', 'trả góp', 'lãi suất', '0%', '...', 'Chí', 'thẻ', 'có thẻ', 'thanh toán', 'nhanh chóng', ',', 'thuận tiện', 'trên', 'phản vị', 'toàn cầu', 'bằng', 'cách', 'chạm', 'thẻ', 'hoặc', 'chạm', 'điện thoại', 'có', 'cài', 'ứng dụng', 'Samsung', 'Pay', '(', 'đồng thời', 'tích hợp', 'Sacombank', 'Visa', ')', 'lên', 'các', 'máy', 'POS', 'NFC.', 'Ngoài ra', ',', 'người', 'dùng', 'còn', 'có thẻ', 'chi tiêu', 'thông qua', 'tinh năng', 'quét', 'mã', 'QR', 'trên', 'ứng

```
{'ngân hàng': 0, 'ngoại thương': 1, 'việt nam': 2, '(: 3, 'vietcombank': 4, ')': 5, 'cánh báo': 6, ':': 7, 'gắn': 8, 'dây': 9, 'cô': 10, 'tình trạng': 11, 'một số': 12, 'doanh nghiệp': 13, 'chuyển': 14, 'tiền': 15, 'cho': 16, 'đổi tác': 17, 'nuôi ngoài': 18, 'không': 19, 'dùng': 20, 'người': 21, 'thu hưởng': 22, '.'': 23, 'nguyên nhân': 24, 'là': 25, 'các': 26, 'đơn vị': 27, 'này': 28, 'bi': 29, 'hacker': 30, 'xâm nhập': 31, 'trái phép': 32, 'email': 33, 'để': 34, 'thay đổi': 35, 'thông tin': 36, 'hướng': 37, 'trên': 38, 'chung tu': 39, 'giao dịch': 40, 'sau': 41, 'đó': 42, 'cũng': 43, 'yêu cầu': 44, 'hỗ trợ': 45, 'đôi': 46, 'tù': 47, 'tuy nhiên': 48, 'biết': 49, 'khoa năng': 50, 'được tiền': 51, 'đôi với': 52, 'hack': 53, 'rất': 54, 'tháp': 55, 'đo': 56, 'thường': 57, 'rút tiền': 58, 'ra': 59, 'khỏi': 60, 'tài khoản': 61, 'ngay': 62, 'khi': 63, 'nhận': 64, 'được': 65, 'hoặc': 66, 'thứ tục': 67, 'phúc tạp': 68, 'của': 69, 'hình thức': 70, 'lừa đảo': 71, 'phổ biến': 72, 'bao gồm': 73, '!!!': 74, 'sử': 75, 'nội': 76, 'dung': 77, 'hợp đồng': 78, 'ký': 79, 'qua': 80, 'giả mạo': 81, 'hóa': 82, 'đom': 83, 'chèn': 84, 'giả': 85, 'khách hàng': 86, 'tối': 87, 'thị trường': 88, 'như': 89, 'trung quốc': 90, 'hong kong': 91, 'malaysia': 92, 'mỹ': 93, 'vâ': 94, 'đặc biệt': 95, 'anh': 96, 'ngân chán': 97, 'rửa ro': 98, 'đã': 99, 'cần': 100, 'chỉ ý': 101, 'đ': 102, 'dấu hiệu': 103, 'trong': 104, 'với': 105, 'bảng': 106, 'thứ': 107, 'nhất': 108, 'liê quan': 109, 'đến': 110, 'thu hiện': 111, 'thông báo': 112, 'giao': 113, 'hang': 114, 'thường lượng': 115, '.'': 116, 'đều': 117, 'lâm': 118, 'trong kh': 119, 'bên': 120, 'xuất khẩu': 121, 'nhập khẩu': 122, 'xác nhận': 123, 'liên lạc': 124, 'khác': 125, 'hai': 126, 'đổi tượng': 127, 'hướng': 128, 'chỉ yêu': 129, 'vua': 130, 'nhô': 131, 'công ty': 132, 'tinh': 133, 'bảo mật': 134, 'an toàn': 135, 'vui': 136, 'tự': 137, 'tự': 138, 'tự': 139, 'tự': 140, 'tự': 141, 'tự': 142, 'tự': 143, 'tự': 144, 'tự': 145, 'tự': 146, 'tự': 147, 'tự': 148, 'tự': 149, 'tự': 150, 'tự': 151, 'tự': 152, 'tự': 153, 'tự': 154, 'tự': 155, 'tự': 156, 'tự': 157, 'tự': 158, 'tự': 159, 'tự': 160, 'tự': 161, 'tự': 162, 'tự': 163, 'tự': 164, 'tự': 165, 'tự': 166, 'tự': 167, 'tự': 168, 'tự': 169, 'tự': 170, 'tự': 171, 'tự': 172, 'tự': 173, 'tự': 174, 'tự': 175, 'tự': 176, 'tự': 177, 'tự': 178, 'tự': 179, 'tự': 180, 'tự': 181, 'tự': 182, 'tự': 183, 'tự': 184, 'tự': 185, 'tự': 186, 'tự': 187, 'tự': 188, 'tự': 189, 'tự': 190, 'tự': 191, 'tự': 192, 'tự': 193, 'tự': 194, 'tự': 195, 'tự': 196, 'tự': 197, 'tự': 198, 'tự': 199, 'tự': 200, 'tự': 201, 'tự': 202, 'tự': 203, 'tự': 204, 'tự': 205, 'tự': 206, 'tự': 207, 'tự': 208, 'tự': 209, 'tự': 210, 'tự': 211, 'tự': 212, 'tự': 213, 'tự': 214, 'tự': 215, 'tự': 216, 'tự': 217, 'tự': 218, 'tự': 219, 'tự': 220, 'tự': 221, 'tự': 222, 'tự': 223, 'tự': 224, 'tự': 225, 'tự': 226, 'tự': 227, 'tự': 228, 'tự': 207, 'tự': 229, 'tự': 23, 'tự': 207, 'tự': 45, 'tự': 230, 'tự': 214, 'tự': 218, 'tự': 219, 'tự': 220, 'tự': 221, 'tự': 222, 'tự': 223, 'tự': 224, 'tự': 225, 'tự': 226, 'tự': 227, 'tự': 228, 'tự': 229, 'tự': 230, 'tự': 227, 'tự': 231, 'tự': 232, 'tự': 233, 'tự': 234, 'tự': 235, 'tự': 236, 'tự': 237, 'tự': 237, 'tự': 238, 'tự': 239, 'tự': 240, 'tự': 241, 'tự': 242, 'tự': 243, 'tự': 244, 'tự': 245, 'tự': 207, 'tự': 181, 'tự': 148, 'tự': 246, 'tự': 7, 'tự': 247, 'tự': 38, 'tự': 248, 'tự': 249, 'tự': 106, 'tự': 250, 'tự': 251, 'tự': 207, 'tự': 251, 'tự': 252, 'tự': 10, 'tự': 253, 'tự': 254, 'tự': 255, 'tự': 256, 'tự': 3, 'tự': 167, 'tự': 257, 'tự': 208, 'tự': 209, 'tự': 5, 'tự': 258, 'tự': 26, 'tự': 259, 'tự': 260, 'tự': 261, 'tự': 178, 'tự': 7, 'tự': 21, 'tự': 158, 'tự': 262, 'tự': 181, 'tự': 230, 'tự': 263, 'tự': 264, 'tự': 265, 'tự': 266, 'tự': 267, 'tự': 38, 'tự': 254, 'tự': 268, 'tự': 23, 'tự': 208, 'tự': 25, 'tự': 269, 'tự': 114, 'tự': 270, 'tự': 225, 'tự': 2, 'tự': 271, 'tự': 245, 'tự': 207, 'tự': 58, 'tự': 106, 'tự': 267, 'tự': 246, 'tự': 94, 'tự': 135, 'tự': 225, 'tự': 136, 'tự': 272, 'tự': 69, 'tự': 273, 'tự': 274, 'tự': 19, 'tự': 100, 'tự': 275, 'tự': 276, 'tự': 207, 'tự': 23, 'tự': 277, 'tự': 278, 'tự': 279, 'tự': 280, 'tự': 7, 'tự': 47, 'tự': 281, 'tự': 110, 'tự': 282, 'tự': 7, 'tự': 208, 'tự': 283, 'tự': 284, 'tự': 285, 'tự': 286, 'tự': 287, 'tự': 69, 'tự': 278, 'tự': 279, 'tự': 280, 'tự': 63, 'tự': 288, 'tự': 207, 'tự': 94, 'tự': 10, 'tự': 230, 'tự': 270, 'tự': 47, 'tự': 289, 'tự': 290, 'tự': 291, 'tự': 292, 'tự': 235, 'tự': 47, 'tự': 63, 'tự': 293, 'tự': 207, 'tự': 23, 'tự': 245, 'tự': 262, 'tự': 64, 'tự': 254, 'tự': 266, 'tự': 295, 'tự': 237, 'tự': 105, 'tự': 296, 'tự': 40, 'tự': 47, 'tự': 289, 'tự': 290, 'tự': 7, 'tự': 144, 'tự': 297, 'tự': 298, 'tự': 299, 'tự': 300, 'tự': 301, 'tự': 234, 'tự': 302, 'tự': 94, 'tự': 292, 'tự': 303, 'tự': 304, 'tự': 23, 'tự': 305, 'tự': 266, 'tự': 295, 'tự': 237, 'tự': 306, 'tự': 145, 'tự': 63, 'tự': 86, 'tự': 111, 'tự': 40, 'tự': 148, 'tự': 251, 'tự': 207, 'tự': 19, 'tự': 307, 'tự': 258, 'tự': 260, 'tự': 7, 'tự': 266, 'tự': 267, 'tự': 66, 'tự': 255, 'tự': 256, 'tự': 23, 'tự': 178, 'tự': 7, 'tự': 0, 'tự': 167, 'tự': 308, 'tự': 207, 'tự': 213, 'tự': 214, 'tự': 309, 'tự': 206, 'tự': 209, 'tự': 105, 'tự': 26, 'tự': 310, 'tự': 311, 'tự': 276, 'tự': 312, 'tự': 278, 'tự': 279, 'tự': 94, 'tự': 231, 'tự': 223, 'tự': 224, 'tự': 313, 'tự': 314, 'tự': 23, 'tự': 315], 'label': 0}
```

# Summary

## (Take-home messages)

---

- Dữ liệu trong một lĩnh vực trước khi vào hệ thống học máy phải được thu thập và biểu diễn thành dạng cấu trúc với một số đặc tính: đầy đủ, ít nhiễu, nhất quán, có cấu trúc xác định.
- Dữ liệu thu thập cho quá trình học là tập nhỏ, tuy vậy cần phản ánh đầy đủ các mặt vấn đề cần giải quyết.
- Dữ liệu thô sau khi thu thập và tiền xử lý phải giữ được sự đầy đủ các đặc trưng ngữ nghĩa – các đặc trưng ảnh hưởng đến khả năng giải quyết vấn đề.
- Khoa học dữ liệu là một lĩnh vực rộng, ngoài việc sử dụng công cụ áp dụng, nắm vững được các kiến thức cơ bản là điều quan trọng.