

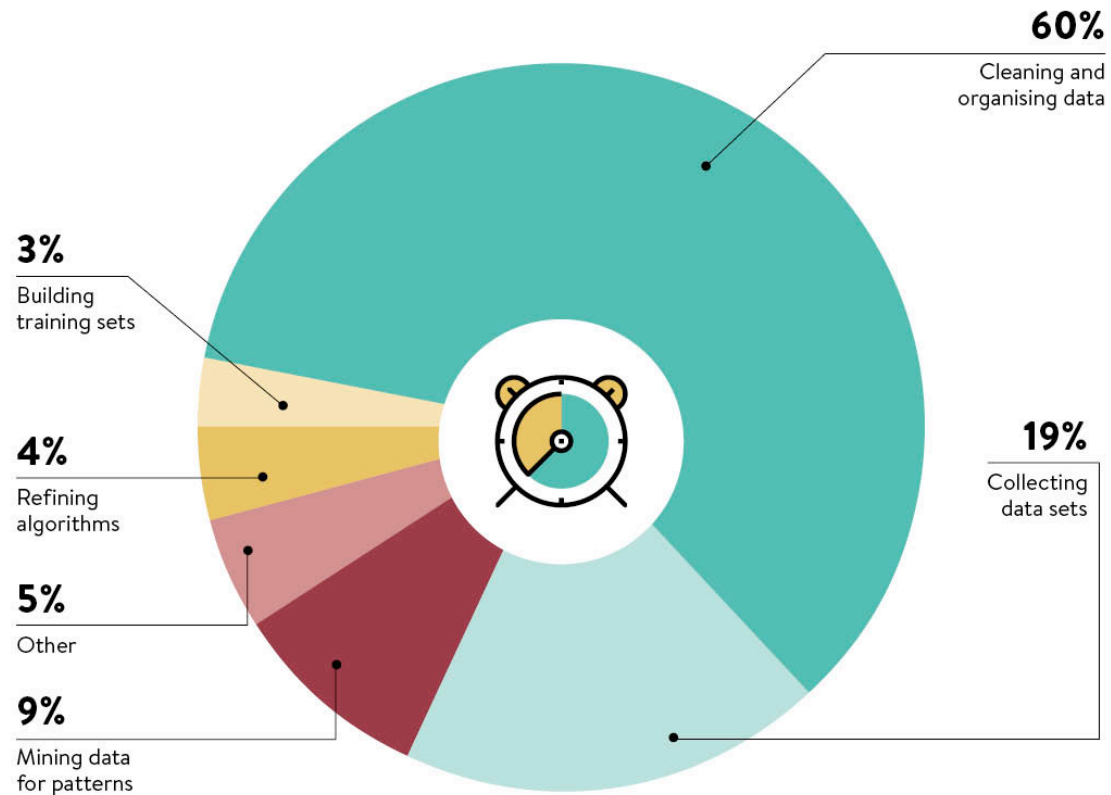
Data crawling and Pre-processing



Hoai Thuan TRAN
Gia Dinh University

Time budget

WHAT DATA SCIENTISTS SPEND THE MOST TIME DOING



Source: CrowdFlower 2016

Data crawling and Pre-processing

Why Pre-processing?

- Convenient for storing and querying
- Machine learning models often work with **structured data**: matrices, vectors, strings,...
- It works well if there is **an appropriate data representation**.

Input

The problem needs to be solved



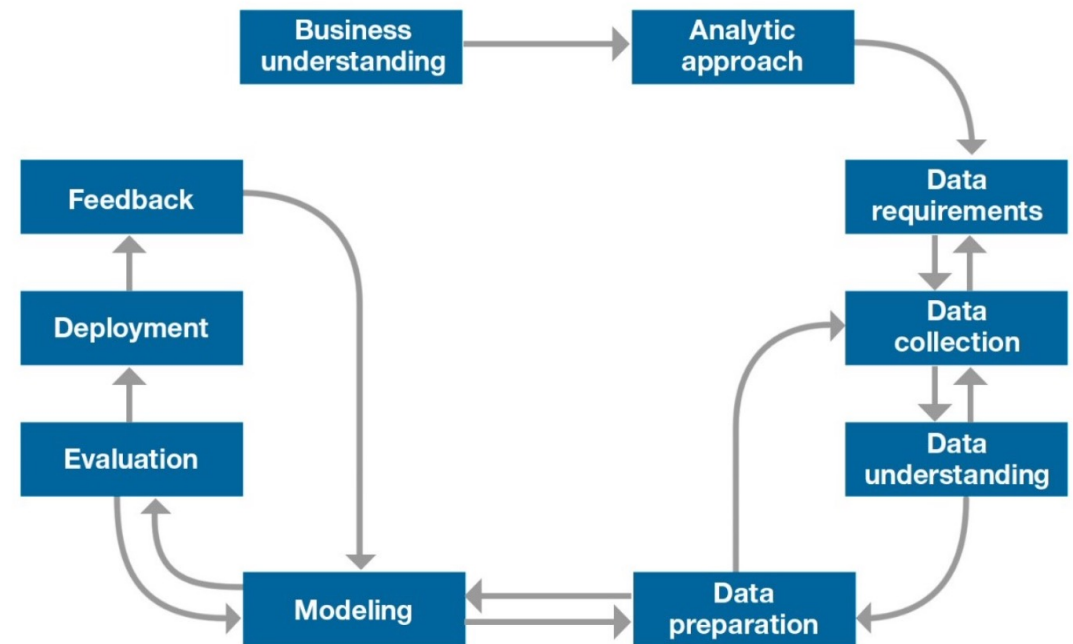
Output

Matrices, vectors,...

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

How?

- Data collection:
 - Sampling
 - Crawling, logging, scarping
- Data processing:
 - Noise filtering, cleaning, digitizing,...



Data crawling and Pre-processing

Data collection

Input


The problem needs to be solved



Output

Data sample

	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				
8	Uzbekistan	Europe					
9	Uruguay	Americas					



Data crawling and Pre-processing

Fundamentals : Sampling

- **WHAT** – Take a small, popular sample set to represent the field to be studied.
- **WHY** – It's impossible to learn everything. Time and computing capabilities are limited.
- **HOW** – Collect samples from reality, or data sources: web, databases,...

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



Fundamentals : Sampling

- **Variety**: The sample set should be cover all contexts of the field.
- **Bias**: Data needs to be general, not biased towards a small part of the field.

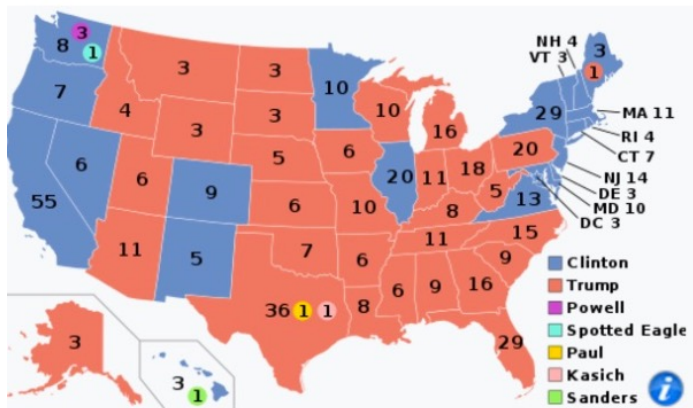
One or more small spoon(s) can be enough to assess whether the soup is good or not."

Remember to stir to avoid tasting biases.



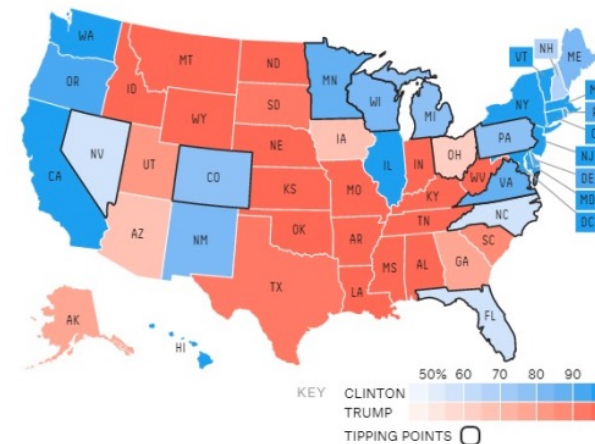
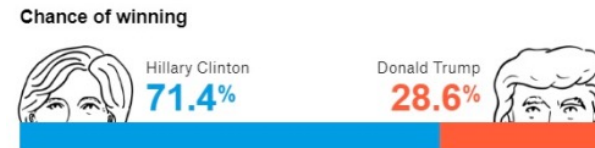
Example

- **Variety**: Are the samples diverse enough?



Actual results

<https://edition.cnn.com/election/results/president>



Electoral votes

Hillary Clinton	302 . 2
Donald Trump	235 . 0

Popular vote

Hillary Clinton	48 . 5%
Donald Trump	44 . 9%

Techniques

- **Crowd-sourcing:** Survey
- **Logging:** save user interaction history, retrieve access products,...
- **Scrapping:** Search for data sources on websites.

DEMO

- Hệ thống crawl dữ liệu từ VNExpress

DEMO

Input

Vấn đề: phân loại văn bản, báo chí



Output

Mẫu dữ liệu: báo chí và nhãn tương ứng

Name	Date modified	
2ce54c553490dc5fb9a7153395793c6a648f...	5/25/2018 4:46 PM	1
7b228847f0f3349971fc590f76def1b0eb5a9...	5/25/2018 4:46 PM	2
8a0f8828443701ee0204f24acdfe880c0fc9...	5/25/2018 4:46 PM	3
94f9342bd858be7b06b26d1ef94d07917e1...	5/25/2018 4:46 PM	4
146fd8057df18632a70e12bc84287655604d...	5/25/2018 4:46 PM	5
651ab2f45f0305220d1f57bb21913620f75d...	5/25/2018 4:46 PM	6
af1e0115782578af4b3773a79f9bec5d2d947...	5/25/2018 4:46 PM	7
c6bd8d552a3d7b3a73acd5798c593db61f9...	5/25/2018 4:46 PM	8
e0efcbc74a5882c6765077448ed7dcd60d...	5/25/2018 4:46 PM	9
e43e36696d676474946fcabf0a812d169e9b...	5/25/2018 4:46 PM	10

```
1  "date": "2018-05-20, 07:44:10"
2  "code": "651ab2f45f0305220d1f57bb21913620f75d128d"
3  "labels": "D\\u00e2n tr\\u00e2n"
4  "content": "\nD\\u00e2n tr\\u00e2n"
5  "image_url": "https://dantri.com.vn"
6  "url": "http://dantri.com.vn"
7  "domain": "dantri.com.vn",
8  "title": "B\\u00e2n Giang: \\\n"
```

DEMO: Steps

Rss

Item

Content


Kênh do VnExpress cung cấp

Trang chủ	RSS 
Thời sự	RSS 
Thế giới	RSS 
Kinh doanh	RSS 
Startup	RSS 
Giải trí	RSS 
Thể thao	RSS 
Pháp luật	RSS 
Giáo dục	RSS 

```
<rss xmlns:siasn="http://purl.org/rss/1.0/modules/siasn/ version="2.0"
  ><channel>
    <title>Kinh doanh - VnExpress RSS</title>
    <description>VnExpress RSS</description>
    <image>
      <url>
        https://s.vnecdn.net/vnexpress/1/v20/logos/vne_logo_rss.png
      </url>
      <title>Tin nhanh VnExpress - Đọc báo, tin tức online 24h</title>
      <link>https://vnexpress.net/link>
    </image>
    <pubDate>Thu, 07 Jun 2018 20:40:44 +0700</pubDate>
    <generator>VnExpress</generator>
    <link>https://vnexpress.net/rss/kinh-doanh.rss</link>
  </item>
  <item>
    NH nhân viên ngân hàng nghỉ việc sau khi trúng xổ số 40 tỷ đồng
    </title>
    <description>
      <![CDATA[
        <a href="https://kinhdoanh.vnexpress.net/tin-tuc/hang-hoa/nu-nhan-vien-r
          <br>src="https://i.kinhdoanh.vnecdn.net/2018/06/07/2191-1528366541-5914-1522
          <br>nhan-vien-ngan-hang-tai-TP-HCM.
        ]]>
      </description>
    <pubDate>Thu, 07 Jun 2018 19:42:16 +0700</pubDate>
    <link>
      https://kinhdoanh.vnexpress.net/tin-tuc/hang-hoa/nu-nhan-vien-ngan-hang-n
    </link>
    <guid>
      https://kinhdoanh.vnexpress.net/tin-tuc/hang-hoa/nu-nhan-vien-ngan-hang-n
    </guid>
    <slash:comments>0</slash:comments>
  </item>
```

```
<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>  
            "Hũ khách hàng may mắn trúng giải tên N.T., là nhân viên một ngân hàng tại TP HCM. Chĩa s tai buổi trao thưởng,&nbsp;&nbsp;&nbsp;&nbsp;"  
        </span>  
        <span><br/></span>  
    </p>  
    <table align="center" border="0" cellpadding="3" cellspacing="0" class="tplCaption" style="width: 100%;><br/></table>  
    <p class="Normal">  
        <span>  
            "Theo thông tin từ Vietlott, chỉ 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  
  
        "Quả kiến tra trên hệ thống kỹ thuật và hồ sơ kèm theo, Vietlott xác định chính xác của chủ hũ N.T là hợp lệ và trúng giải Jackpot 1 Power 6/55 kỳ quay thứ 131. Tầm về được phát hành tại điểm bán hàng đường số 6, phường Linh Chiểu, quận Thủ Đức, TP HCM."  
    </p>  
    <p class="Normal"></p>  
    <p class="Normal"></p>
```

DEMO: Sample

 JSON

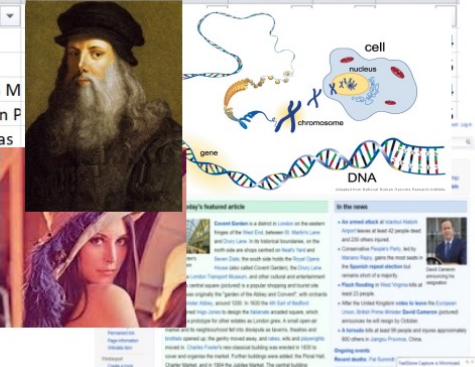
```
... ■ date : "2018-05-20, 07:44:00-07:00"
... ■ code : "651ab2f45f0305220d1f57bb21913620f75d128d"
... ■ labels : "Dân trí/Bạn đọc"
... ■ content : " Dân trí Sau khi Bí thư Tỉnh ủy Bắc Giang yêu cầu dẹp tan nạn xe quá tải trong năm 2018, Phòng CSGT Công an tỉnh Bắc Giang
... ■ image_url : "https://dantrcdn.com/zoom/80_50/2018/5/20/7-1526776517717498023080.png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

Data Pre-processing

Input

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

	A	B	C	D	E	F	G
1	Country	Region					
2	Zimbabwe	Africa					
3	Zambia	Africa					
4	Yemen	Eastern M					
5	Viet Nam	Western P					
6	Venezuela (Bo	Americas					
7	Vanuatu	Wester					
8	Uzbekistan	Europe					
9	Uruguay	America					



Output

Dữ liệu số theo từng ML/AI model(s)

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

Fundamentals : Data “rawness”

Completeness (đầy đủ)

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

Integrity (trung thực)

Nguồn thu thập chính thống, đảm bảo mẫu thu được chứa giá trị chính xác thực tế.

Homogeneity (đồng nhất)

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

Structures (cấu trúc)

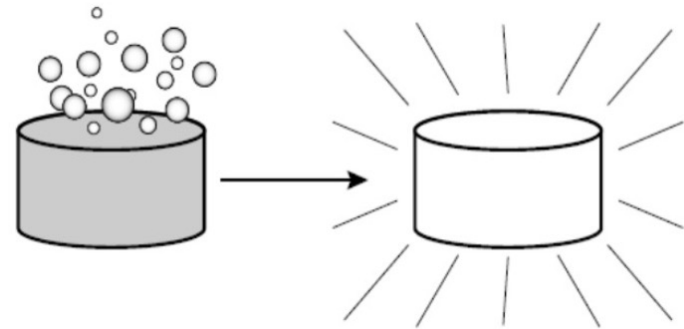
C	D	E	F
Population	Under15	Over60	Fertil
13724	40.24	5.68	3.64
14075	46.73	3.95	5.77
23852	40.72	4.54	4.35
90796	22.87	9.32	1.79
29955	28.84	9.17	2.44
247	37.37	6.02	3.46
28541	28.9	6.38	2.38
3395	22.05	18.59	2.07



Techniques : Cleaning

■ Tính đầy đủ + trung thực

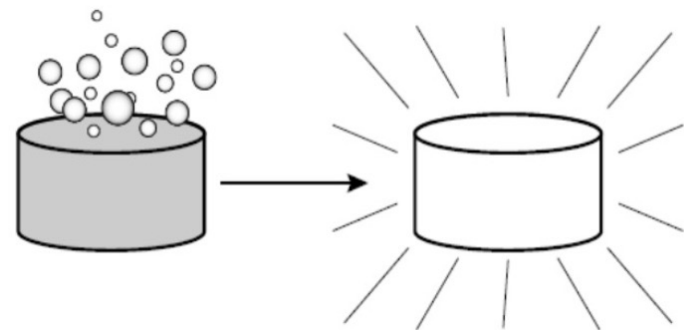
- Mẫu dữ liệu cần được thu thập từ các **nguồn đáng tin cậy**. Phản ánh vấn đề cần giải quyết.
- Loại bỏ **nhiều** (ngoại lai): bỏ vài mẫu dữ liệu mà có khác biệt lớn với các mẫu khác.
- Một mẫu dữ liệu có thể bị trống (thiếu, chưa đầy đủ), cần có chiến lược phù hợp:
 - Bỏ qua, không đưa vào phân tích?
 - Bổ sung các trường còn thiếu cho mẫu?



Techniques : Cleaning

■ Điền giá trị thiếu

- Điền lại giá trị bằng tay
- Gán cho giá trị nhãn đặc biệt hay ngoài khoảng biểu diễn.
- Gán giá trị trung bình cho nó.
- Gán giá trị trung bình của các mẫu khác thuộc cùng lớp đó.
- Tìm giá trị có xác suất lớn nhất điền vào chỗ bị mất (hồi quy, suy diễn Bayes,...).



A1	A2	A3	A4	A5	A6	A7	A8	y
?	3.683	?	-0.634	1	0.409	7	30	5
?	?	60	1.573	0	0.639	7	30	5
?	3.096	67	0.249	0	0.089	?	80	3
2.887	3.870	68	-1.347	?	1.276	?	60	5
2.731	3.945	79	1.967	1	2.487	?	100	4

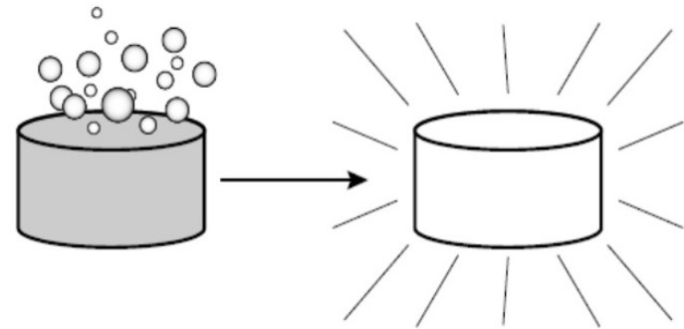
Techniques : Cleaning

■ Tính đồng nhất

- Các mẫu dữ liệu cần có tính đồng nhất về cách biểu diễn, ký hiệu
- Ví dụ không đồng nhất:

Rating “1, 2, 3” & “A, B, C”;

Age = 42 & Birthday = 03/08/2020



Techniques : Integrating

Structured – relational (table-like)

Un-structured

	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

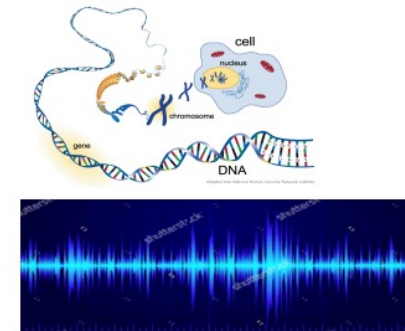
```
{
  "code": "1473a6fd39d1d8fa48654aac9d8cc2754232",
  "title": "[Updating] Câu chuyện xuyên mưa về :",
  "url": "http://techtalk.vn/updating-cau-chuyen",
  "labels": "techtalk/Cong nghe",
  "content": "Vào chiều tối ngày 09/12/2016 vừa",
  "image_url": "",
  "date": "2016-12-10T03:51:10Z"
}
```

texts in websites, emails, articles, tweets

2D/3D images, videos + meta

spectrograms, DNAs, ...

The collage illustrates diverse data types: a Wikipedia 'Welcome' page, a tweet from Dwayne Johnson (@TheRock) with the text 'Sometimes as a father, you ARE the only solution. A real honor making this true story. @SNITCH 9/99/19 via twitter.com/aJhoF6d', a news article titled 'Seeking Life's Bare (Genetic) Necessities' from Scientific American, and a diagram of a cell showing the nucleus, DNA, and chromosomes.



Introduction to Machine Learning and Data Mining

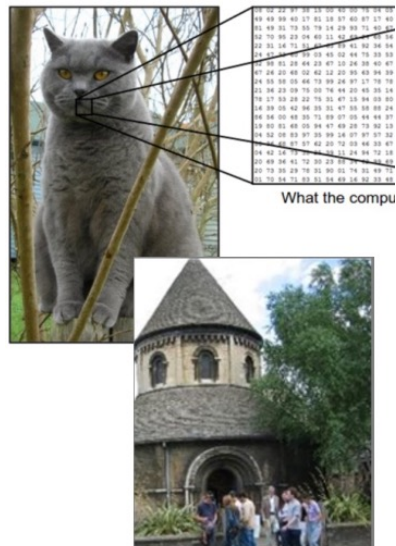
Techniques : Transforming

- **Semantics?**

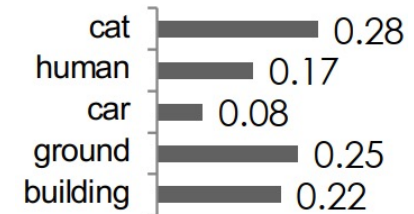
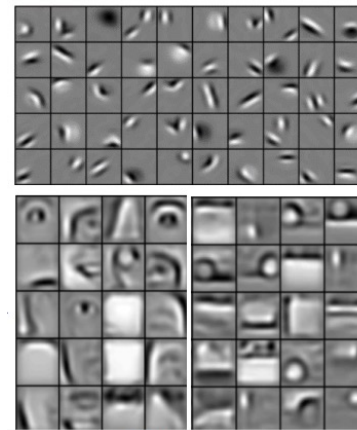
Trích xuất các đặc trưng ngữ nghĩa, chuẩn hóa

Semantics example: visual data

Low-level semantics (raw pixels)



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

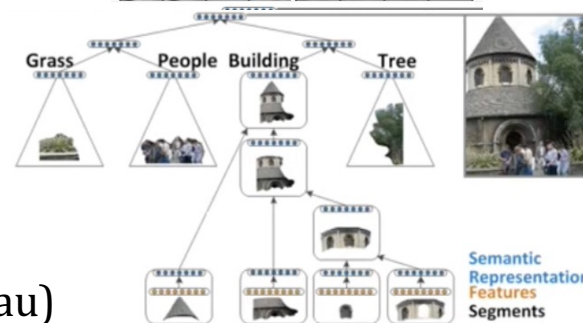


cat → not on → car

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)



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

Data crawling and Pre-processing

Techniques : Transforming

- Mục tiêu: trích xuất các đặc trưng ngữ nghĩa.

USD điều_chỉnh trái chiều , vàng SJC quay đầu tăng

```
(0, 24506) 0.2077168092100841
(0, 23857) 0.34468369118902636
(0, 22309) 0.31713411814089415
(0, 21894) 0.3025597601047669
(0, 21265) 0.2449372095782497
(0, 20409) 0.3276089788346888
(0, 17739) 0.515839529548281
(0, 16499) 0.33820735665113805
(0, 4648) 0.3132633187744836
```

B	C	D	E	F	G
Region	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

- 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, ...)
- *Feature discretization* (rời rạc hoá): một số thuộc tính tỏ ra hiệu quả hơn khi được gom nhóm các giá trị.
- *Feature normalization*: chuẩn hóa giá trị thuộc tính, về cùng một miền giá trị, dễ dàng trong tính toán.

Techniques : Transforming

- Giảm kích cỡ:
 - Giúp giảm kích thước của dữ liệu và đồng thời giữ được ngữ nghĩa cốt lõi của dữ liệu.
 - Giúp tăng tốc quá trình học hoặc khai phá tri thức.
- Vài chiến lược:
 - **Lựa chọn đặc trưng** (feature selection): các thuộc tính không liên quan, dư thừa hoặc các chiều cũng có thể xóa hay loại bỏ.
 - **Giảm chiều** (dimension reduction): dùng một số thuật toán (ví dụ PCA, ICA, LDA,...) để biến đổi dữ liệu ban đầu về không gian có ít chiều hơn.
 - **Trừu tượng hoá**: các giá trị dữ liệu thô được thay thế bằng các khái niệm trừu tượng.

DEMO

- Transforming text data

DEMO

Input

Mẫu dữ liệu thô: json text

```
{  
  "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"  
}
```

Output

Dữ liệu số theo từng ML/AI model(s)

(0, 24003)	0.08875917745394017
(0, 23874)	0.08543368833593054
(0, 23214)	0.06269100273800875
(0, 23085)	0.10941900286727153
(0, 22547)	0.047792971979914244
(0, 22446)	0.05082334424962779
(0, 21910)	0.08271656588481778
(0, 21905)	0.06404674731000018
(0, 21779)	0.11899134180006703
(0, 21572)	0.08401328893873479

DEMO: Steps

Tokenize

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 khắp thẻ 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 thưởng 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', 'và', 'thẻ', 'thanh toán', 'và', 'thẻ', 'trả', 'trước', 'và', 'các', 'sản phẩm', 'này', 'có', 'tiện ích', 'chung', 'như', 'thanh toán', 'và', 'rút tiền', 'khắp', 'thế giới', 'và', 'mua sắm', 'trực tuyến', 'và', 'nhận', 'giảm giá', 'đến', '50%', 'tại', 'hàng', 'trăm', 'điểm', 'chấp nhận', 'thẻ', 'liên kết', 'và', 'thẻ', 'hỗ trợ', 'chi tiêu', 'trước', 'và', 'thanh toán', 'sau', 'miễn', 'lãi', 'tối đa', '55', 'ngày', 'và', 'tích lũy', 'điểm', 'thưởng', 'để', 'đổi', 'quà', 'và', 'mua hàng', 'trả góp', 'lãi suất', '0%', 'và', 'chủ', 'thẻ', 'có thể', 'thanh toán', 'nhanh chóng', 'và', 'thuận tiện', 'trên', 'phạm vi', 'toàn cầu', 'bằng', 'cách', 'chạm', 'thẻ', 'hoặc', 'chạm', 'điện thoại', 'có', 'cài', 'ứng dụng', 'Samsung', 'Pay', 'và', 'đồng thời', 'tích hợp', 'Sacombank', 'Visa', 'và', 'lên', 'các', 'máy', 'POS', 'NFC', 'Ngoài ra', 'và', 'người', 'dùng', 'còn', 'có thể', 'chi tiêu', 'thông qua', 'tính năng', 'quét', 'mã', 'QR', 'trên', 'ứng

Dictionary

Data Input (tfidf-Vector)

```
{'dân_trí': 6928, 'sở': 17869, 'gd': 7729, 'dt': 23214, 'tỉnh': 218, 'sgdtd': 17039, 'vp': 21572, 'chấn_chính': 4971, 'tiếp_thị': 16, 'giáo_dục': 7955, 'chỉ_đạo': 5092, 'tuyệt_đối': 20254, 'phép': 16194, 'mua_bán': 12653, 'dụng_cụ': 7191, 'học_tập': 9557, 'g': 63, 'tổ_chức': 20928, 'ngành': 13667, 'tham_gia': 18129, 'giới_th': 12651, 'phát_hành': 15346, 'tham_khảo': 18130, 'phụ_huynh': 14805, 'lành_mạnh': 11553, 'chương_trình': 4935, 'phổ_thông': 16816, 'bảo_cáo': 3493, 'hướng': 9359, 'sơ': 17704, 'đề': 5693, 'chuyên_viên': 4681, 'dồ_dùng': 24003, 'công_khai': 15421, 'ngăn_chặn': 13743, 'báo': 3490, 'thông_tin': 18676, '5492, 'chư_păn': 4929, 'tờ': 20984, 'giấy': 8066, 'thông_báo': 1818993, 'nga': 13400, 'hiệu_trưởng': 8753, 'hôm': 9267, 'xả': 004, 'chim': 4524, 'non': 14434, 'học': 9534, 'hót': 9259, 'bảo_đ': 50, 'địa_phương': 23924, 'đặc_điểm': 23836, 'loài': 11400, 'nghiên': 12940, 'nơron': 14632, 'thần_kinh': 18881, 'trách_nhiệm': 19790, 'ông_bố': 5853, 'ấn_bản': 24292, '09': 168, '12': 348, 'tạp_chí': 132, 'trúc': 19889, 'não_bộ': 14521, 'thí_nghiệm': 18628, 'tiến_s': 17142, 'đại_học': 23619, 'cornell': 5477, 'đồng_nghiệp': 24
```

```
(0, 24003) 0.08875917745394017  
(0, 23874) 0.08543368833593054  
(0, 23214) 0.06269100273800875  
(0, 23085) 0.10941900286727153  
(0, 22547) 0.047792971979914244  
(0, 22446) 0.05082334424962779  
(0, 21910) 0.08271656588481778  
(0, 21905) 0.06404674731000018  
(0, 21779) 0.11899134180006703  
(0, 21572) 0.08401328893873479  
(0, 20984) 0.0603014300399073  
(0, 20928) 0.03425727291794896  
(0, 20851) 0.04139691505815508  
(0, 20796) 0.06515117203347312  
(0, 20272) 0.09576360104259622  
(0, 20254) 0.21906274633402326  
(0, 19934) 0.09329205643046397  
(0, 19928) 0.0815770967825164  
(0, 19410) 0.06593571705754445  
(0, 19370) 0.03950424960970291  
(0, 19345) 0.16543313176963556  
(0, 18993) 0.04356540203990621
```

Exercise

- Bài tập: Tính vector biểu diễn của văn bản với bộ dữ liệu nhỏ.
- Dữ liệu: 2 bài báo từ trang dân trí
- Yêu cầu:
 - Sử dụng module tách từ.
 - Build tập từ điển từ 2 văn bản
 - Sử dụng stopwords lọc từ dừng.
 - Chuyển hoá 2 văn bản thành 2 vector tfidf

Summary

- 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.