HUST

ĐẠI HỌC BÁCH KHOA HÀ NỘI

HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

ONE LOVE. ONE FUTURE.

ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Introduction to
Machine Learning and Data Mining

IT3190

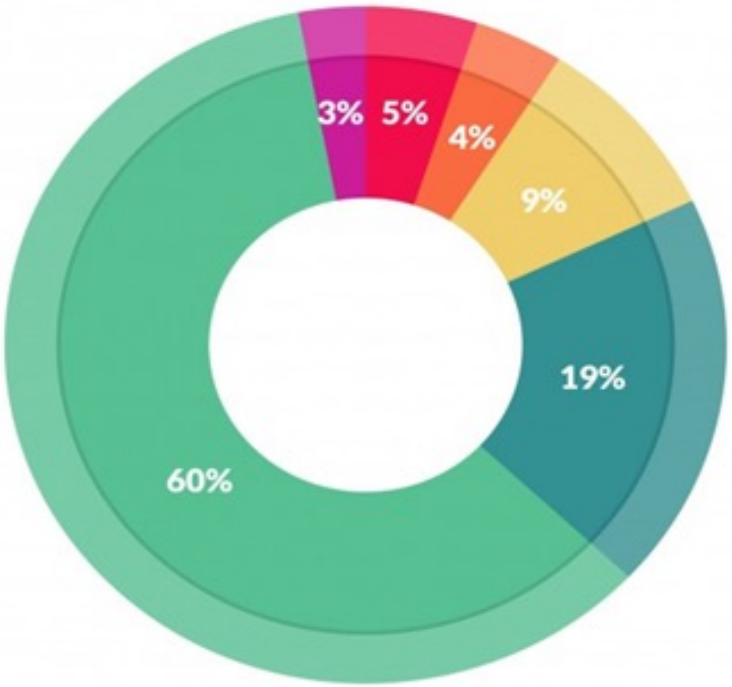
Lecture: Data crawling and pre-processing

ONE LOVE. ONE FUTURE.

Contents

- Lecture 1: Introduction to Machine Learning & Data Mining
- **Lecture 2: Data crawling and pre-processing**
- Lecture 3: Linear regression
- Lecture 4+5: Clustering
- Lecture 6: Decision tree and Random forest
- Lecture 7: Neural networks
- Lecture 8: Support vector machines
- Lecture 9: Performance evaluation
- Lecture 10: Probabilistic models
- Lecture 11: Basics of data mining
- Lecture 12: Association rule mining
- Lecture 13: Regularization and advanced topics

Time budget



CrowdFlower Inc., 2016

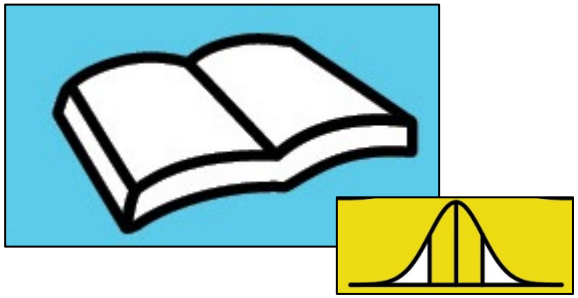
- What data scientists spend the most time doing?
  - **Collecting data: 19%**
  - **Cleaning and organizing data: 60%**
  - Building training datasets: 3%
  - Data mining: 9%
  - Refining algorithms: 4%
  - Others: 5%

Why?

- Why preprocess the data?
  - Convenient in storage, query data
  - Machine learning models usually work with structured data: matrices, vectors, arrays, etc.
  - Machine learning usually works well if there is a suitable representation of the data

Input

Problems to be solved



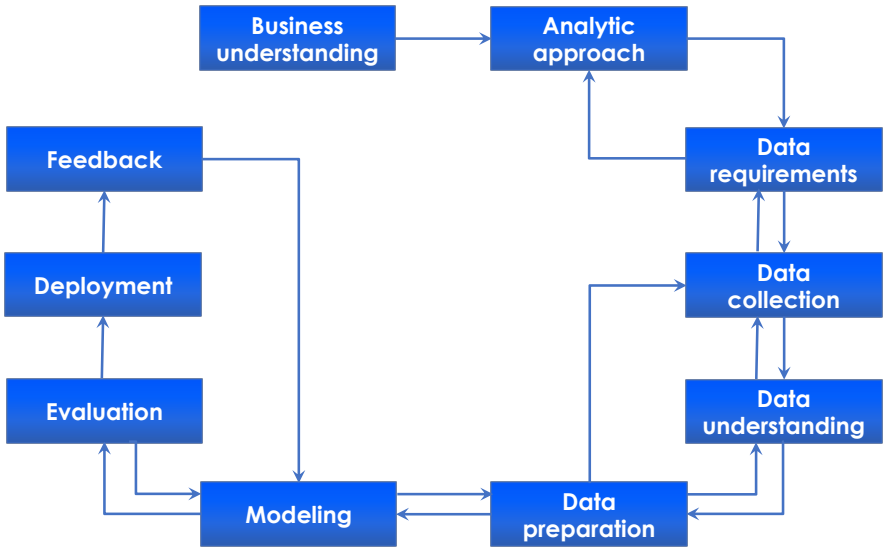
Output

Numeric data - matrix, vector

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

How?

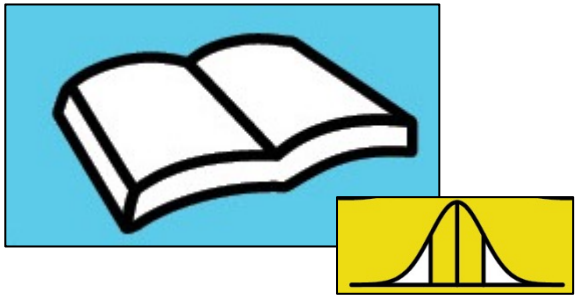
- Data collection
  - Sampling
  - Method: crawling, logging, scraping
- Data processing
  - Noise filtering, cleaning, digitizing…



Data collection

Input

Problems to be solved



Output

Data samples

	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					

Fundamentals :: Sampling

- **WHAT** – Take a small, generalized set of samples to represent the field to be learned
- **WHY** – can't learn the whole thing due to time and computational power limitations
- **HOW** – Collect samples from real life, or web data sources, databases…

"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

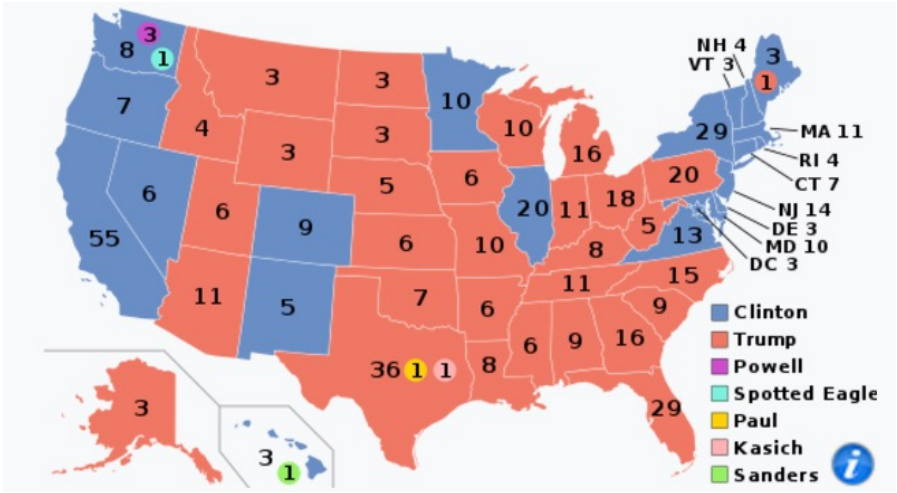
- Variety – the sample set is diverse enough to cover all contexts of the field/domain.
- Bias – data needs to be generalised, not biased towards a small part of the field.

"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.

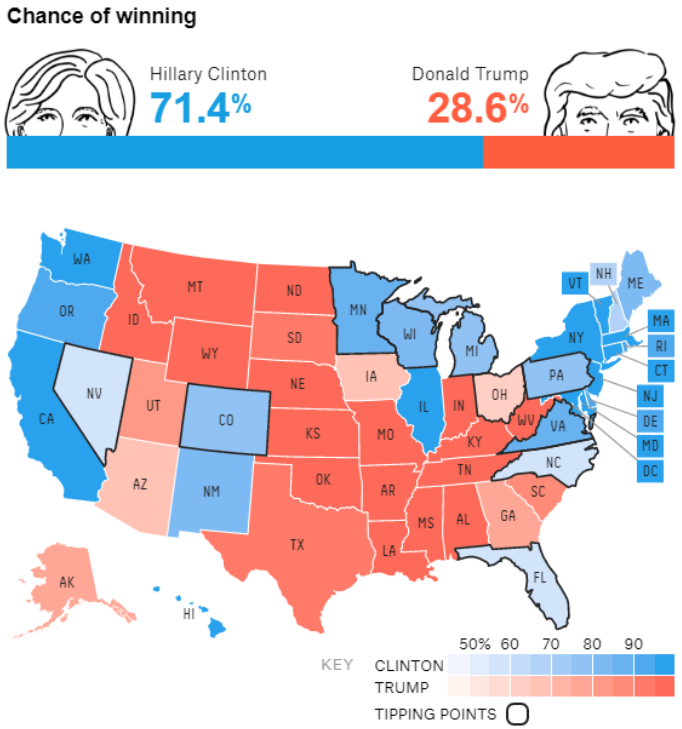
https://www.coursera.org/learn/inferential-statistics-intro

Fundamentals :: Sampling :: How

- Variety – samples vary enough to reflect reality?



Actual results



Electoral votes	
Hillary Clinton	302.2
Donald Trump	235.0

Popular vote	
Hillary Clinton	48.5%
Donald Trump	44.9%

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

Techniques

- **Crowd-sourcing:** Conduct surveys
- **Logging**: record user interaction history, product access...
- **Scrapping:** Search data sources on websites, download, extract, filter data

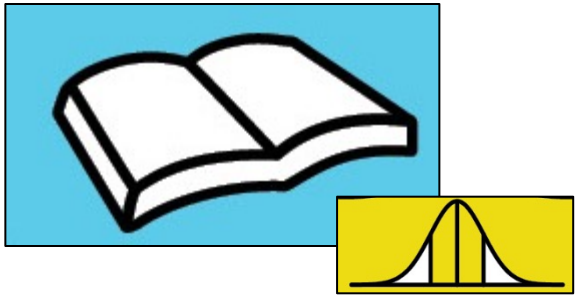
Techniques :: Scrapping :: DEMO

- **Objective:** Data for the problem of text classification – newspaper articles.
- **DEMO**: Newspaper crawling system

DEMO

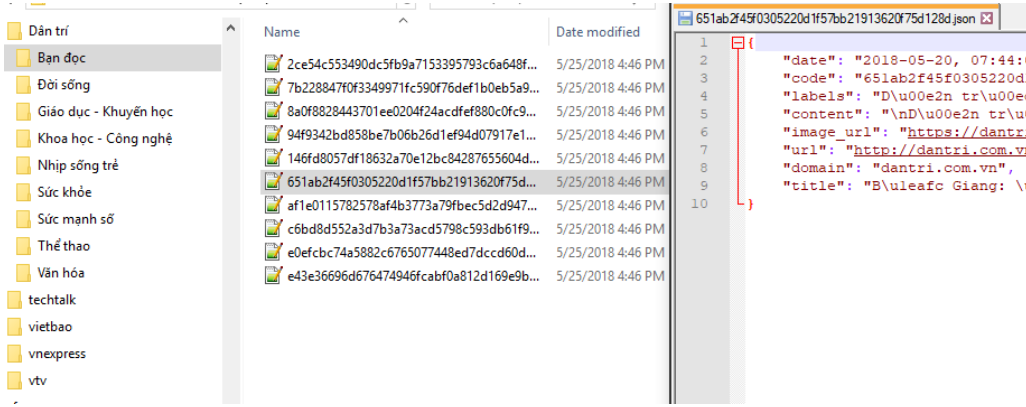
Input

Problem: classifying newspaper articles



Output

Sample data: newspaper articles and labels



DEMO :: Steps

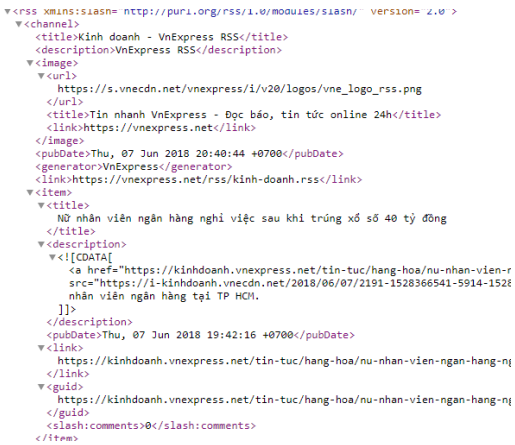
Rss

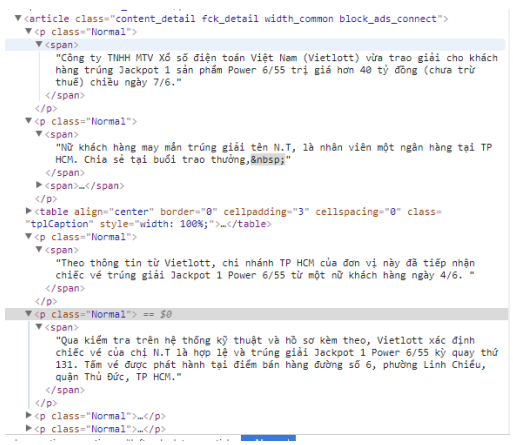
Item

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





DEMO :: Sample

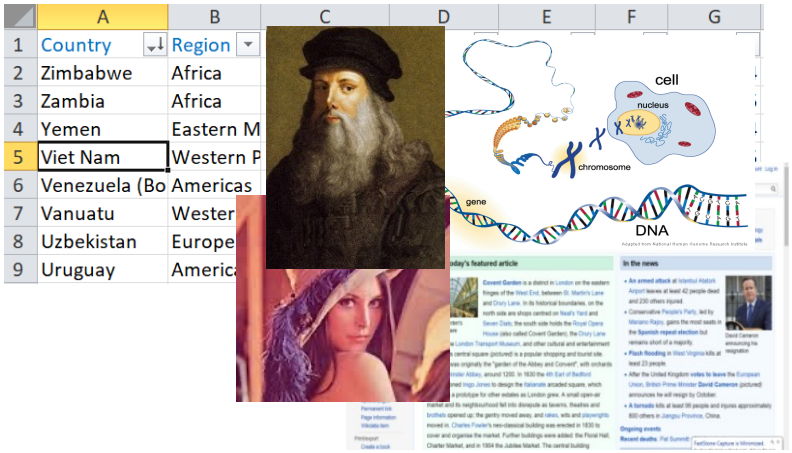
JSON

- 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 quâ
- image_url : "https://dantricdn.com/zoom/80_50/2018/5/20/7-15267765177174980230 80.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 cơi nới thùng rầm rập chạy qua mặt cảnh sát giao thông?"

Data preprocessing

Input

Raw data samples (text, image, audio...)



Output

Digital data for ML/AI model(s)

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

Fundamentals :: Data “rawness”

Completeness
(đầy đủ)

Each collected sample should have all the required attribute

Integrity
(trung thực)

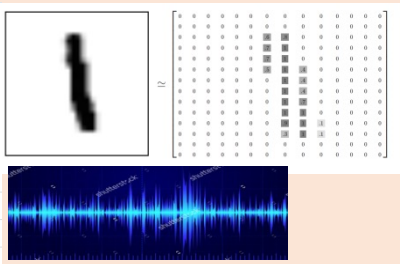
- Official collection source, ensure the sample contains the correct value in reality
- Jan. 1 as *everyone’s* birthday? – *intentional (systematic) noises*

Homogeneity
(đồng nhất)

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

Structures
(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

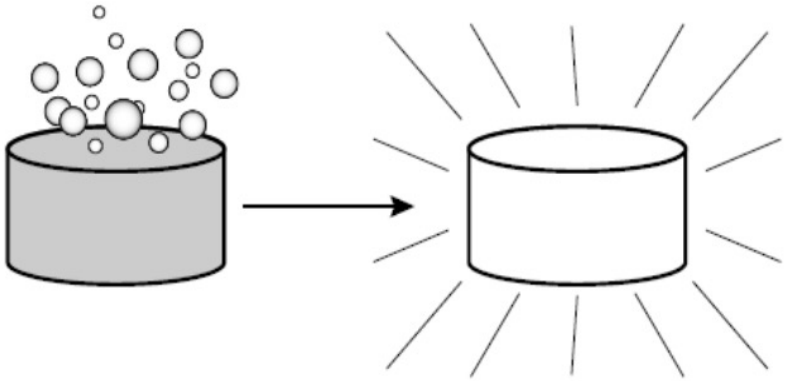
Cleaning

Integrating

Transforming

Techniques :: Cleaning

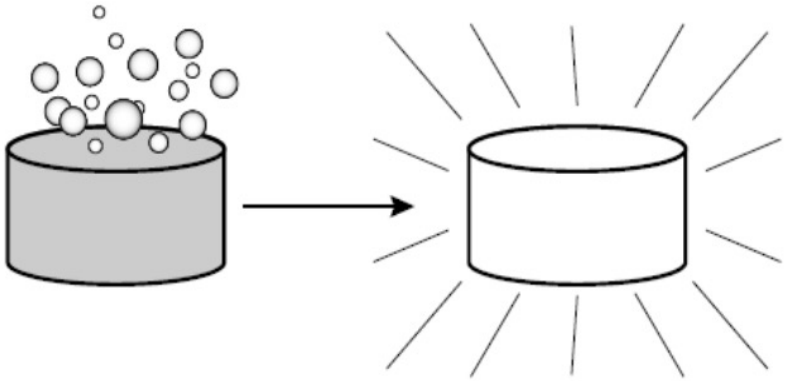
- Tính đầy đủ + trung thực



- Data samples should be collected from reliable sources. Reflect the problem to be solved.
- Eliminate (outliers) noise: remove some data samples that are significantly different from other samples
- A data sample may be empty (missing, incomplete), a suitable strategy is needed:
  - Ignored, not included in the analysis?
  - Add missing fields to the data sample?

Techniques :: Cleaning

- Missing values

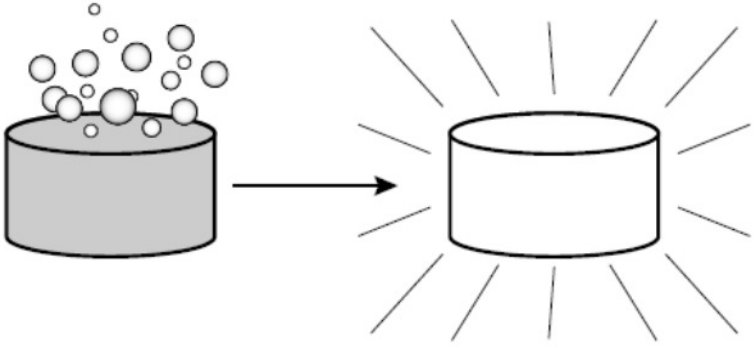


1. Fill in the missing value manually
2. Use a global constant
3. Use and “average” value
4. Use average value for all samples belonging to the same class/group
5. Use the most probable value (regression, bayesian inference)

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 (cont.)

- **Homogeneity**



Different data presentation, units of measure, metrics etc.

Examples:

Rating “1, 2, 3” & “A, B, C”;

Age = 42 & Birthday = 03/08/2020

Techniques :: Integrating w/ some Transforming

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-chuye
    "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, …

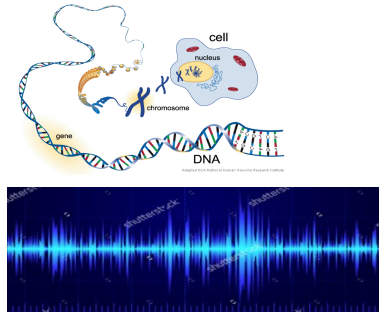


image credits: wikipedia, shutterstock, CNN

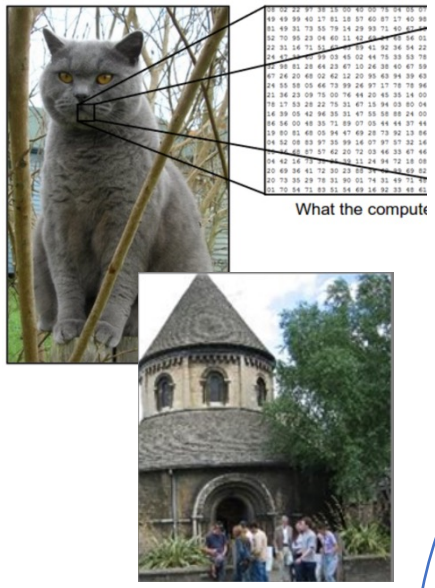
Techniques :: Transforming

Semantics?

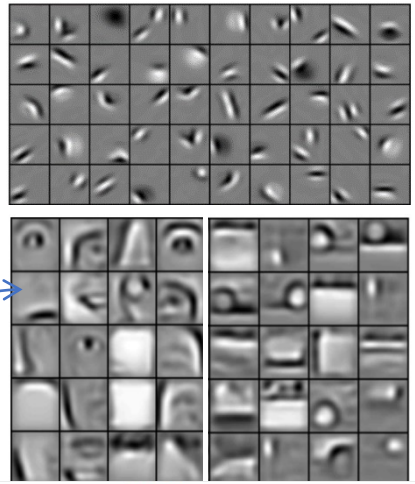
Extract semantic features, normalize

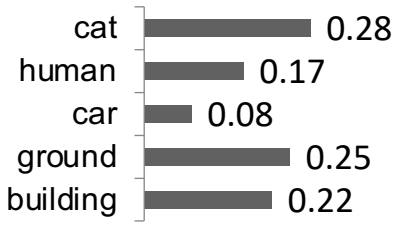
Semantics example: visual data

Low-level semantics
(raw pixels)

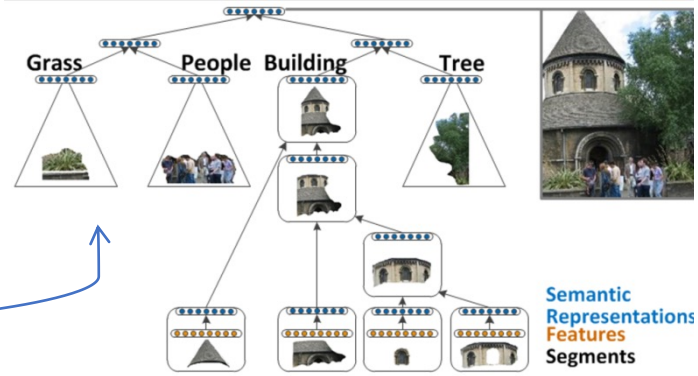


Mid-/High-level semantics
(e.g. human-interpretable features)





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



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

Minimum semantic level to understand:

- Text classification
- Emotional analysis
- AI Chatbot (various semantic levels)

Image credits: CS231n, Stanford University; Lee et al, 2009; Socher et al, 2011

Techniques :: Transforming (cont.)

- Objective: to extract semantic features.

```
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

One-hot encoding

$1 = [1\,0\,0\,0\,0]$

$3 = [0\,0\,1\,0\,0]$

...

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

- For specific field, type of data, using different semantic feature extraction techniques (text data, images, ...)

… and *standardize*

- *Feature discretization* (rời rạc hoá): Some attributes are more efficient when grouped values.
- *Feature normalization* (chuẩn hóa): normalize attribute values to the same domain, easy to calculate.

Techniques :: Transforming (cont.)

- Data reduction:
  - Helps reduce the size of the data and, at the same time, preserve the core semantics of the data.
  - Helps speed up the process of learning or knowledge discovery.
- Some strategies:
  - *Feature selection:* redundant attributes or dimensions can also be eliminated
  - *Dimensional reduction:* use some algorithms (eg. PCA, ICA) to transform the original data into a less dimensional space.
  - *Abstraction:* raw data values are replaced by abstract concepts.

Techniques :: Transforming example & demo

DEMO

Input
Raw sample: json text


```
{
    "image_url": "https://i-kinhdoanh.vnecdn.net/2018/
    "url": "https://kinhdoanh.vnexpress.net/tin-tuc/eb
    "title": "Sacombank n\u00e2ng c\ulea5p nhi\uleclu
    "code": "db274d03b9a61aa16d70c7fd68929d799058b866",
    "domain": "kinhdoanh.vnexpress.net",
    "date": "2018-05-25, 17:00:00+07:00",
    "content": "\nHi\ulec7n th\ulebb qu\uledlc t\ulebf
    "labels": "vnexpress/Kinh doanh/Ebank\u00a0/Kinh d
}
```


Output
Numeric sample


```
(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
```

DEMO :: Steps

Tokenize

Dictionary

Data Input (tfidf-Vector)

```
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 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 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', ',', 'thẻ', 'thanh toán', 'và', 'thẻ', 'trả', 'trước', '.', 'Các', 'sản
 phẩm', 'này', 'có', 'tiện ích', 'chung', 'như', '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', '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 hợp', 'Sacombank', 'Visa', ')',
 'lên', 'các', 'máy', 'POS', 'NFC.', 'Ngoài ra', ',', 'người', 'dùng', 'còn', 'có
 thể', 'chi tiêu', 'thông qua', 'tính năng', 'quét', 'mã', 'QR', 'trên', 'ứng
```

```
{'dân_trí': 6928, 'sở': 17869, 'gd': 7729, 'đt': 23214, 'tỉnh': 2
8, 'sgddt': 17039, 'vp': 21572, 'chấn_chỉnh': 4971, 'tiếp_thị': 1
6, 'giáo_dục': 7955, 'chỉ_đạo': 5092, 'tuyệt_đối': 20254, 'phép':
o': 16194, 'mua_bán': 12653, 'dụng_cụ': 7191, 'học_tập': 9557, 'g
63, 'tổ_chức': 20928, 'ngành': 13667, 'tham_gia': 18129, 'giới_th
ua': 12651, 'phát_hành': 15346, 'tham_khảo': 18130, 'phụ_huynh':
ng': 14805, 'lành_mạnh': 11553, 'chương_trình': 4935, 'phổ_thông'
ai_sót': 16816, 'báo_cáo': 3493, 'hướng': 9359, 'sơ': 17704, 'đề_
'cán_bộ': 5693, 'chuyên_viên': 4681, 'đồ_dùng': 24003, 'công_khai
g': 15421, 'ngăn_chặn': 13743, 'báo': 3490, 'thông_tin': 18676, '
5492, 'chư_păh': 4929, 'tờ': 20984, 'giấy': 8066, 'thông_báo': 18
'thị': 18993, '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, 'nghiê
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
học': 17142, 'đại_học': 23619, 'cornell': 5477, 'đồng_nghiệp': 24
4520, 'vta': 21588, 'ventral': 21329, 'tegmental': 18076, 'area':
8922, 'tín_hiệu': 20537, 'giúp': 7983, 'chim_sẻ': 4528, 'vằn': 21
```

```
(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)    0.16543131176963556
(0, 18993)    0.04356540203990621
(0, 18676)    0.03464691377452836
(0, 18659)    0.050391345485324875
```

DEMO :: Exercise

- **Exercise:** Calculate vector representation of text with small dataset.
- **Data:** 2 articles from Dantri site.
- **Request:**
  - Use the word separator module.
  - Build a dictionary from 2 documents
  - Use a list of stopwords filter stopwords.
  - Convert 2 documents into 2 tf.idf vectors

Summary

- The data in a field before entering the machine learning system must be collected and represented in a structured form with some characteristics: completeness, integrity, homogeneity, well-defined structure.
- The data collected for the learning process is a small set, but it should reflect all aspects of the problem to be solved.
- Raw data, after collection and preprocessing, must retain the full range of semantic features – features that affect problem solving.
- Data science is a broad field, in addition to using applied tools, mastering the basics is important.

HUST

THANK YOU !