Escreva sua pergunta aqui: https://app.sli.do/event/k1narrnf

MAC0459/MAC5865 - Tópicos em Ciência e Engenharia de Dados

Aula 03

Sejam bem-vindas, sejam bem-vindos!

Entre no link https://app.sli.do/event/k1narrnf ou e faça suas perguntas da aula.

R. Hirata Jr.

Escreva sua pergunta aqui: https://app.sli.do/event/k1narrnf

Objetivos de hoje

- Ao final da aula de hoje você deve:
  - Conhecer alguns tipos de dados e as dificuldades de tratá-los
  - Conhecer a discussão Big Data vs Small Data
  - Conhecer o seu grupo

“Onde a gente poderia procurar mais sobre como fazer uma boa pergunta?”

Escreva sua pergunta aqui: https://app.sli.do/event/k1narrnf

Relembrando a aula passada

Simple pipeline – Data Science Method

1. Pose question
2. Get the data
3. Explore the data
4. Model the data
5. Report results

Escreva sua pergunta aqui: https://app.sli.do/event/k1narrnf

Scientific questions

- To call in the statistician after the experiment is done may be no more than asking him/her to perform a postmortem examination: she/he may be able to say what the experiment died of.

Sir Ronald Fisher

Escreva sua pergunta aqui: https://app.sli.do/event/k1narrnf

Scientific questions

Good experiments are designed

Hypothesis vs data driven science

- HD science:
  - given a problem, what available data will help us answer it?
- Data driven science:
  - given data, what interesting problems can we apply it to?

Escreva sua pergunta aqui: https://app.sli.do/event/k1narrnf

Learning to ask questions

- Computer scientists students are not used to ask questions, why?
- Good data scientists develop an inherent curiosity about the world around them and have wide-ranging interests.

Simple pipeline – Scientific Method

1. Pose a question
2. Formulate a hypothesis
3. Formulate an experiment
4. Observe (data collecting)
5. Analyse the results
6. Go back to step 2 if the hypothesis is not correct/supported
7. Report results

Example of application

1. Pose a question
   - Is lettuce mostly composed by water?
2. Formulate a hypothesis
   - Lettuce leaves have about 95% of water
3. Formulate an experiment

Experimental protocol

1. Food dehydrator
2. Lettuce (origin etc)
3. Clean and dry leaves
4. Weight the leaves
5. Put it to dry for 60 minutes

https://www.westonsupply.com/Weston-Stainless-Steel-Food-Dehydrator-p/74-1001-w.htm

Escreva sua pergunta aqui: https://app.sli.do/event/k1narrnf

Experimental protocol

4. Weight the dried leaves
5. Take the difference of weights and check if it is 95% of the original weight
6. Go back to step 2 if the hypothesis is not correct/supported
7. Report results

https://www.westonsupply.com/Weston-Stainless-Steel-Food-Dehydrator-p/74-1001-w.htm

Escreva sua pergunta aqui: https://app.sli.do/event/k1narrnf

What is Science?

- “We absolutely must leave room for doubt or there is no progress and there is no learning. There is no learning without having to pose a question. And a question requires doubt. People search for certainty. But there is no certainty. People are terrified — how can you live and not know? It is not odd at all. You only think you know, as a matter of fact. And most of your actions are based on incomplete knowledge and you really don’t know what it is all about, or what the purpose of the world is, or know a great deal of other things. It is possible to live and not know.” Feynman

Escreva sua pergunta aqui: https://app.sli.do/event/k1narrnf

What is Science?

- In our example, did we left room for doubt?
- The lemma of our disciple should be:

De omnibus dubitandum

Learning to ask questions

- The baseball encyclopedia
- The Internet Movie Database (IMDb)
- Google Ngrams
- New York Taxi Records

Escreva sua pergunta aqui: https://app.sli.do/event/k1narrnf

Pronto, pode acordar!

INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
IME 50
UNIVERSIDADE DE SÃO PAULO

Escreva sua pergunta aqui: https://app.sli.do/event/k1narrnf

Você está presente?

Properties of Data

- Structured vs Unstructured Data
- Quantitative vs Categorical Data
- Big Data vs Small Data
- Classification vs Regression

Types of Data

- Treatment - any condition that is applied to the subjects being measured
- Treatment level - different versions, aspects, of a treatment
- Block - group of subjects that share certain characteristics

Types of Data

Twenty four subjects - 12 males and 12 females – consumer reaction – 1 to 10

Age of subjects	Sports	coupe	Four-door	sedan
“treatments”	Male	Female	Male	Female
21-44	8	7	6	7
	7	6	8	5
45-64	7	6	6	8
	7	5	7	8
65+	4	3	7	9
	6	5	9	8

Types of Data

- Three types of fertilizers (treatments) are being tested.
- Three fields (blocks) of equal size are being used.

Productivity in bushels

Field	Gro-fast	Fertilizer King's Formula 6	Greenway
A	126	137	119
B	84	89	87
C	113	121	124

Escreva sua pergunta aqui: **https://app.sli.do/event/k1narrnf**

Types of Data

- ## Dependent and independent samples

Interest

Date	Bank 1	Bank 2	Bank 3
01/15	9.6	10.1	9.8
03/10	9.4	9.9	9.8
07/08	9.3	9.6	9.5
10/01	10.6	11.0	10.4

hours

Brand A	Brand B
852	810
829	801
864	835
843	807
832	819

An Introduction to Mathematical Statistics and its Applications - Larsen & Marx

Types of Data

- Similar and dissimilar units

Property	Living area	Price
P1	2860	210.500
P2	3210	219.900
P3	2350	146.000
P4	5340	359.500
P5	7234	467.300

An Introduction to Mathematical Statistics and its Applications - Larsen & Marx

Types of Data

- Quantitative and qualitative measurements

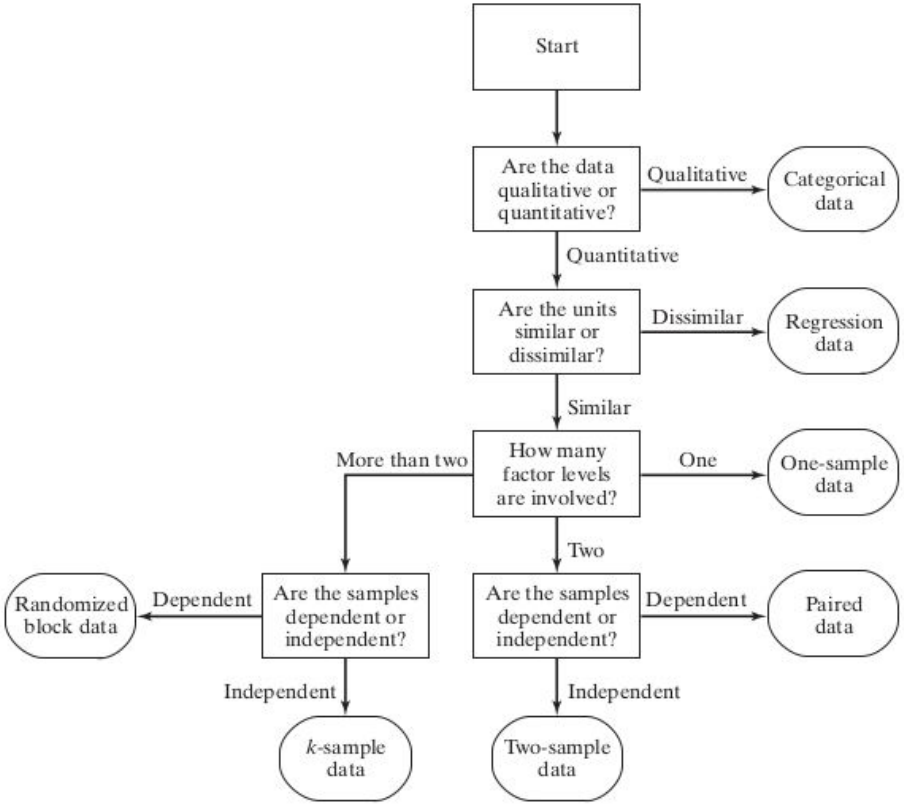
Property	Living area	Price	Class
P1	2860	210.500	Apartment
P2	3210	219.900	Condo
P3	2350	146.000	Bungalow
P4	5340	359.500	Apartment
P5	7234	467.300	Castle

An Introduction to Mathematical Statistics and its Applications - Larsen & Marx

Types of Data

- Experimental data
  - One can design and perform an experiment
- Observational data
  - Data is drawn from a sample of a population

Flowchart for classifying data



An Introduction to Mathematical Statistics and its Applications
Larsen & Marx

Big Data – really?

- The best accepted definition for Big Data (Ling Liu):
  - A dataset, or set of datasets that are beyond the ability of legacy approaches to manage at an acceptable level of quality and/or
  - That exceeds the capacity of conventional systems (hardware and/or software) to process within an acceptable elapsed time

Big Data – really?

- The definition is subjective and evolving
  - As technology advances over time, the size of datasets increase.
  - The definition is varying by sector, depending on
    - **Software tools used**
    - **Domain of application**

Big Datasets – characteristics (Liu)

- Huge in Volume
- Distributed
- Dynamic (Velocity)
- Heterogeneous (Variety)
  - Many agents access/update data
- Noisy (Veracity)
  - Inherent
  - Unintentional
  - Malicious
- Unstructured / semi-structured
  - No database schema

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

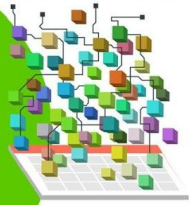
Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day



6 BILLION PEOPLE
have cell phones

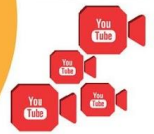
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month



30 BILLION PIECES OF CONTENT
are shared on Facebook every month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

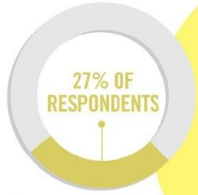
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions



in one survey were unsure of how much of their data was inaccurate

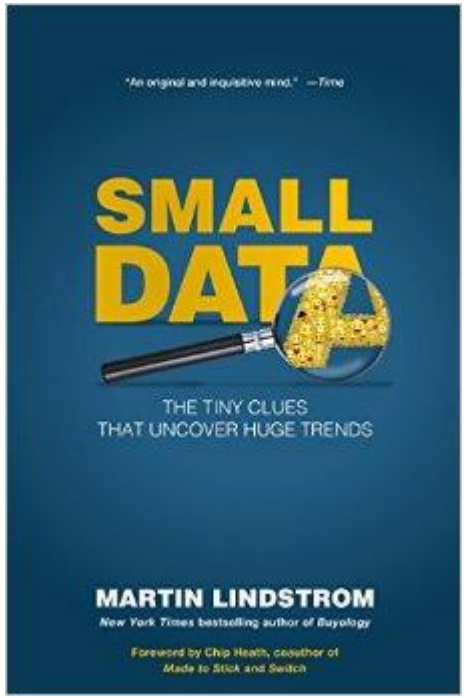
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

Escreva sua pergunta aqui: **https://app.sli.do/event/k1narrnf**

Small Data

- Small Data (Martin Lindstrom):
  - The Tiny Clues That Uncover Huge Trends
  - https://www.marketingjournal.org/small-data-big-impact-an-interview-with-martin-lindstrom/
  - Several cases
    - **Lego case: “instant gratification, lacking the patience or the attention to engage with complex building projects.”**
    - **“Big data studies suggested that future generations would lose interest in LEGO.”!**



INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
IME 50
UNIVERSIDADE DE SÃO PAULO

Escreva sua pergunta aqui: **https://app.sli.do/event/k1narrnf**

Obrigado!