BRAVE NEW WORLD OF DATA...

ADMIRÁVEL MUNDO NOVO DE DADOS...

Francisco Louzada
CER-CEMEAI
ICMC-USP

In this conference I discuss the opportunities that the huge amount of data offers us, in the light of the counterpoint between the ultra-modernity of data capture mechanisms and the statistical methodologies.

And how we may contribute to the training of a new professional required to act within it efficiently.

CONCERN OF STATISTICS → **PRIMARY DATA ANALYSIS**

- DATA ARE COLLECTED WITH A PARTICULAR PURPOSE (CONJECTURE)
- WHICH CAN BE TRANSLATED INTO HYPOTHESES TO BE TESTED

EXAMPLES

Expected Survival Loss due to various causes (years)

To whom people pray?

What are the most violent cities in the State of São Paulo, Brazil?

The production process is under control?

Expected Survival Loss due to various causes (years)

Cause	Years
Being Single (Male).....	- 9.6
Being Left Handed.....	- 9.0
Smoke (Male).....	- 6.3
Being Single (Female).....	- 4.4
Smoke (Female).....	- 2.2
30% + overweight.....	- 3.6
20% - underweight.....	- 2.5
Pap Test.....	4.0
Airbags in car.....	3.2

Source: Rao, C.R. (1997). Statistics And Truth: Putting Chance To Work, 2ed. Wspc.

9 in 10 adults say they pray.
For whom they pray?
(sample of 2550 people)

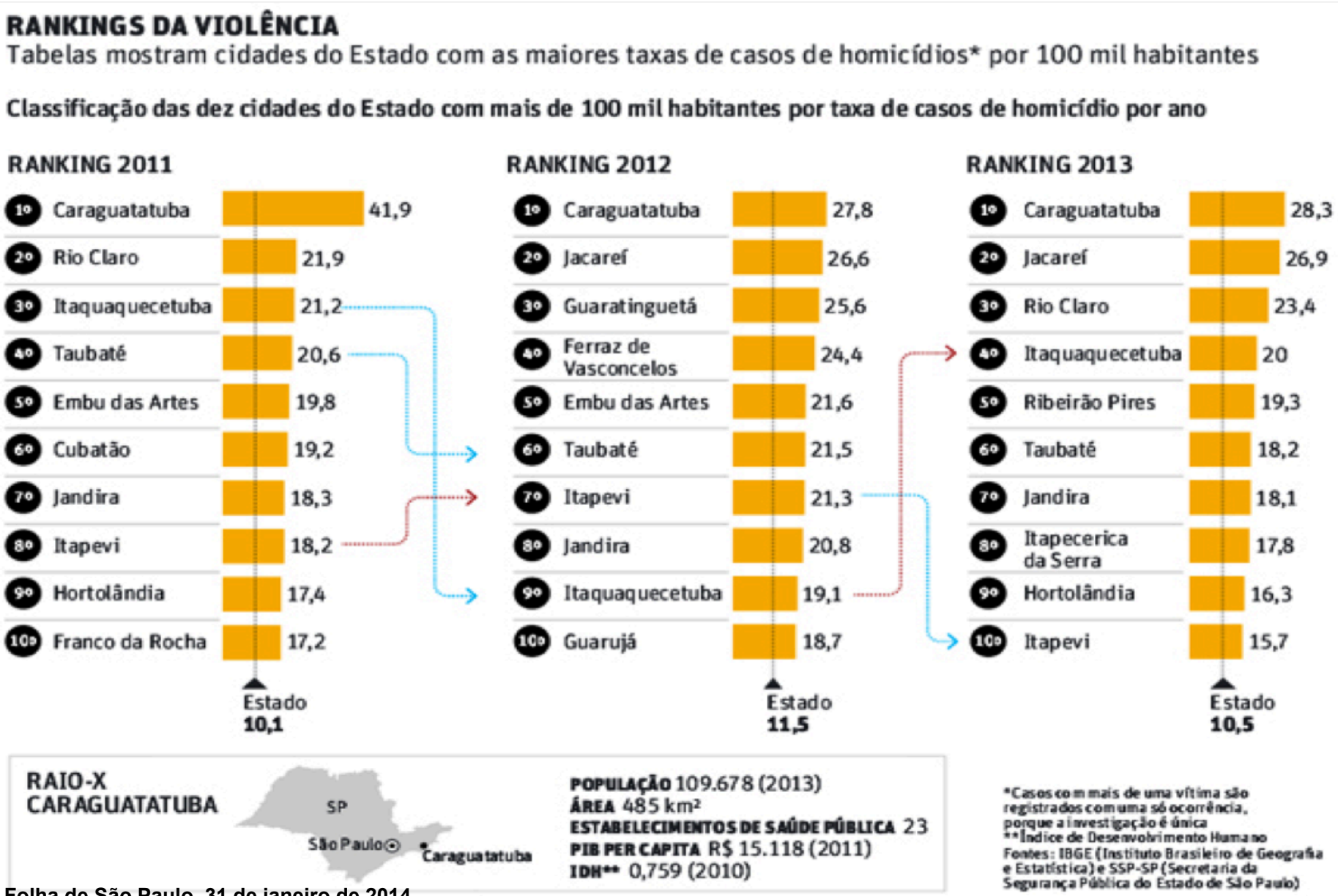
For Family

Children the World

Peace in the world

Work Colleagues

The most violent cities in the State of São Paulo, Brazil.



Folha de São Paulo, 31 de janeiro de 2014
http://www1.folha.uol.com.br/cotidiano/2014/01/1404357-caraguatatuba-e-a-cidade-mais-violenta-de-sp-pelo-3-ano-consecutivo.shtml

CEP ONLINE
Statistical Control Process System

Home | CER | Quality | Contac

Generated on dia 28-04-2014
at 17:21:48

Print

Type of data:
Quantitative data

Type of graphic:
S and X-bar chart

Type of condition:
without specification

If the process is under statistical control, set the specification limits for the assessment of its capability and the level of confidence of the confidence intervals for each index.

LSL:

USL:

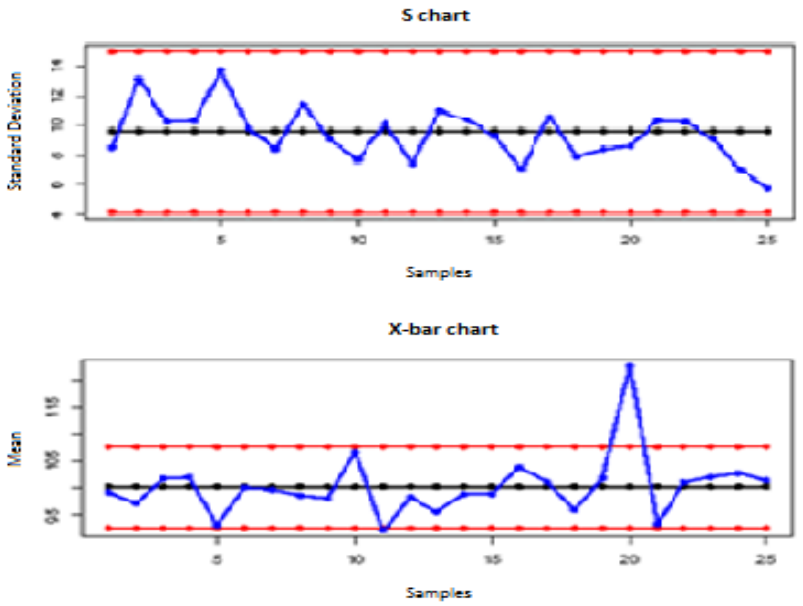
Alpha*:

Calculate

* Set the Alpha value in decimal number and use '.' as decimal separator.

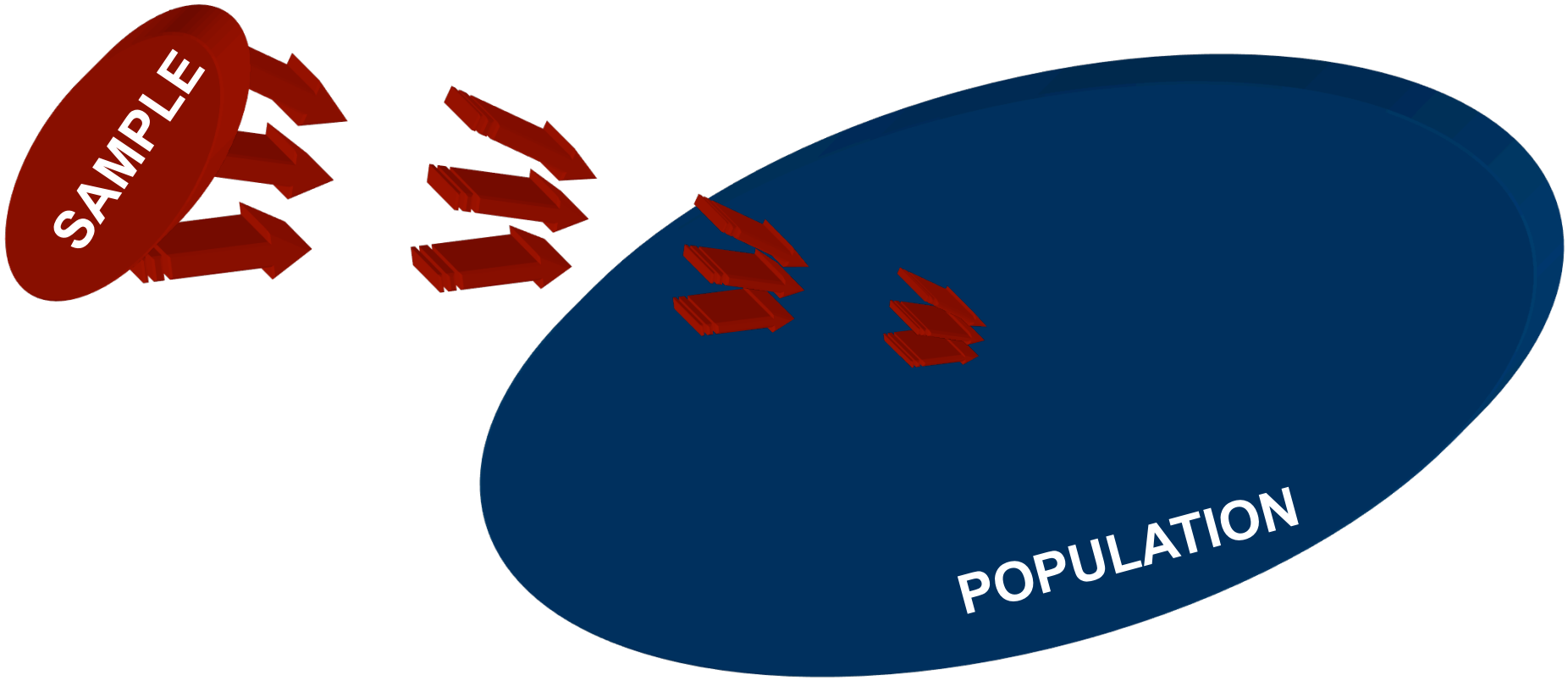
Caution: The instant analysis generated by the CEP online system is presented below. These are some statistical techniques for monitoring the inserted process.

Aspects of quality of the chosen graphics:



Data: weight (in grams) of chocolate bars produced by a certain food company. In this case, 25 samples with 15 observations (bars) each, taken from the production line of the company in question were observed.

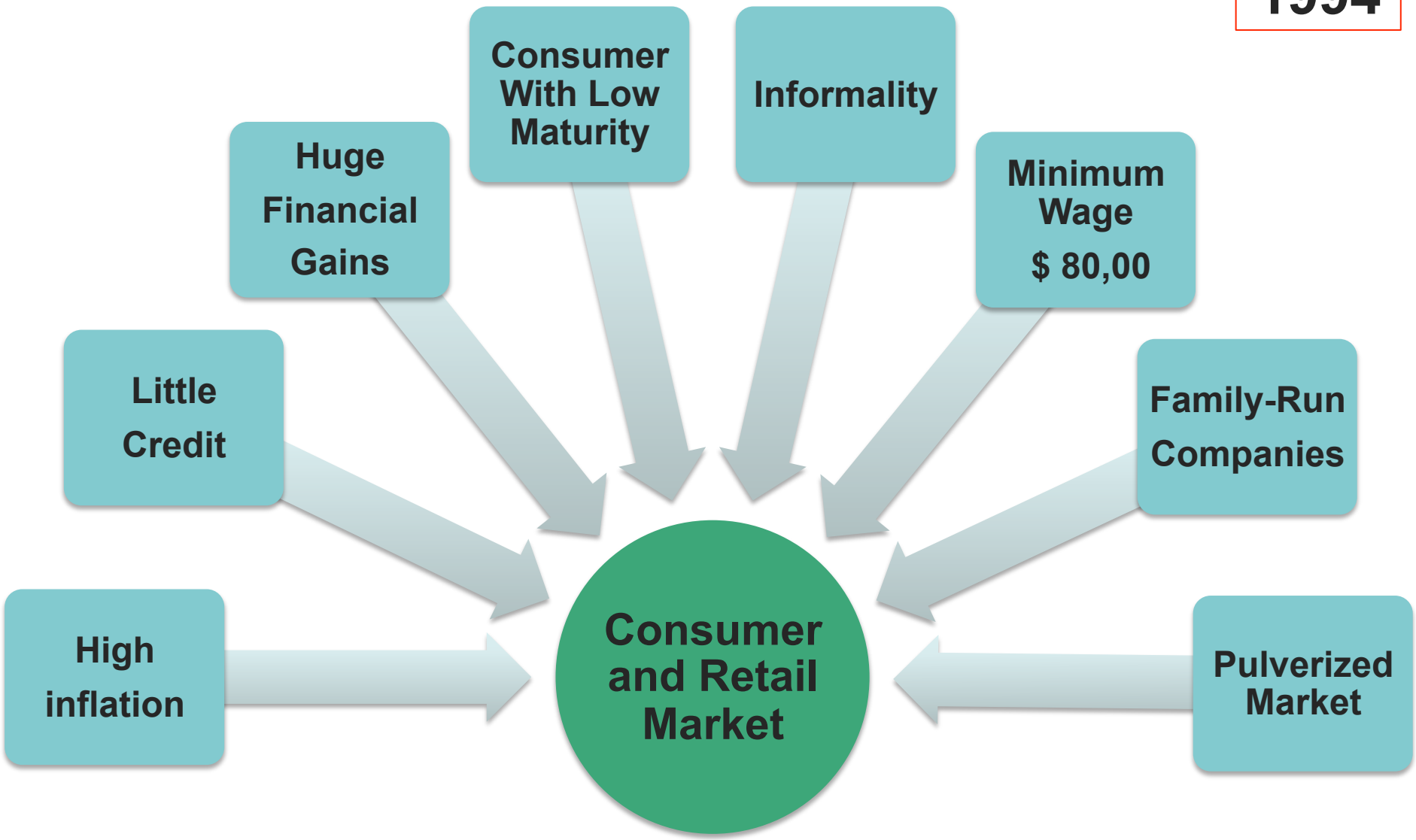
The common feature of the presented examples



NEW INDUSTRIAL ORDER

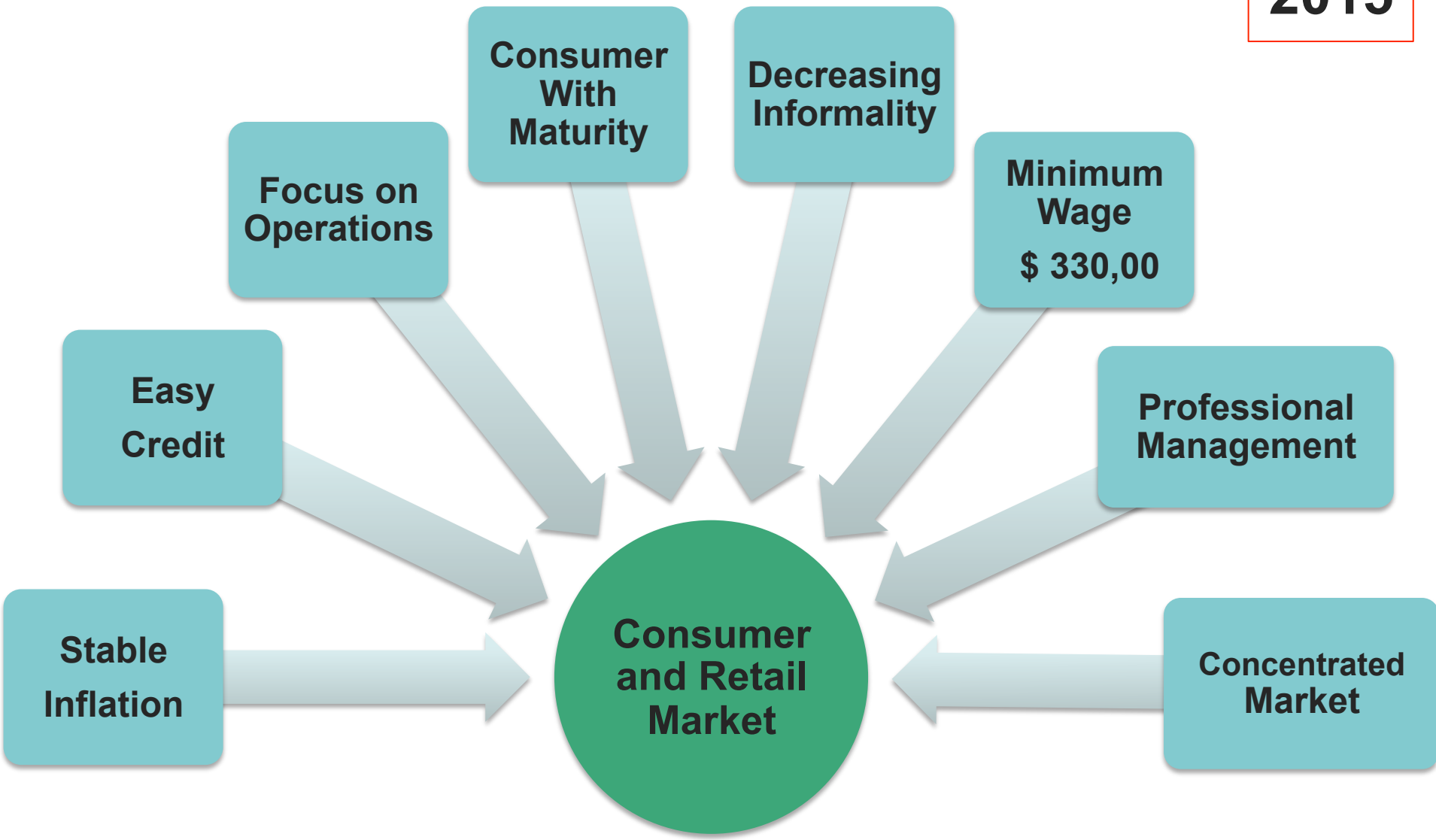
(Since 1994)

1994



MERCADO DE CONSUMO E VAREJO

2015



MERCADO DE CONSUMO E VAREJO

TRADITIONAL FOOD SUPPLY CHAIN
(CADEIA DE ABASTECIMENTO TRADICIONAL)

TODAY SUPPLY CHAIN
(CADEIA DE ABASTECIMENTO ATUAL)

E-Commerce Sales

- E-Commerce in Brazil in 2002 – US$ 0.5 million
- E-Commerce in Brazil in 2013 – US$ 10 billion
- E-Commerce worldwide in 2015 – US$ 1.7 trillion
- From 2010 to 2011: increasing of 48%
- 30 million of e-customers
- Average ticket R$ 400,00 (US$ 160,00)

A new order of consumption of goods, products

Before

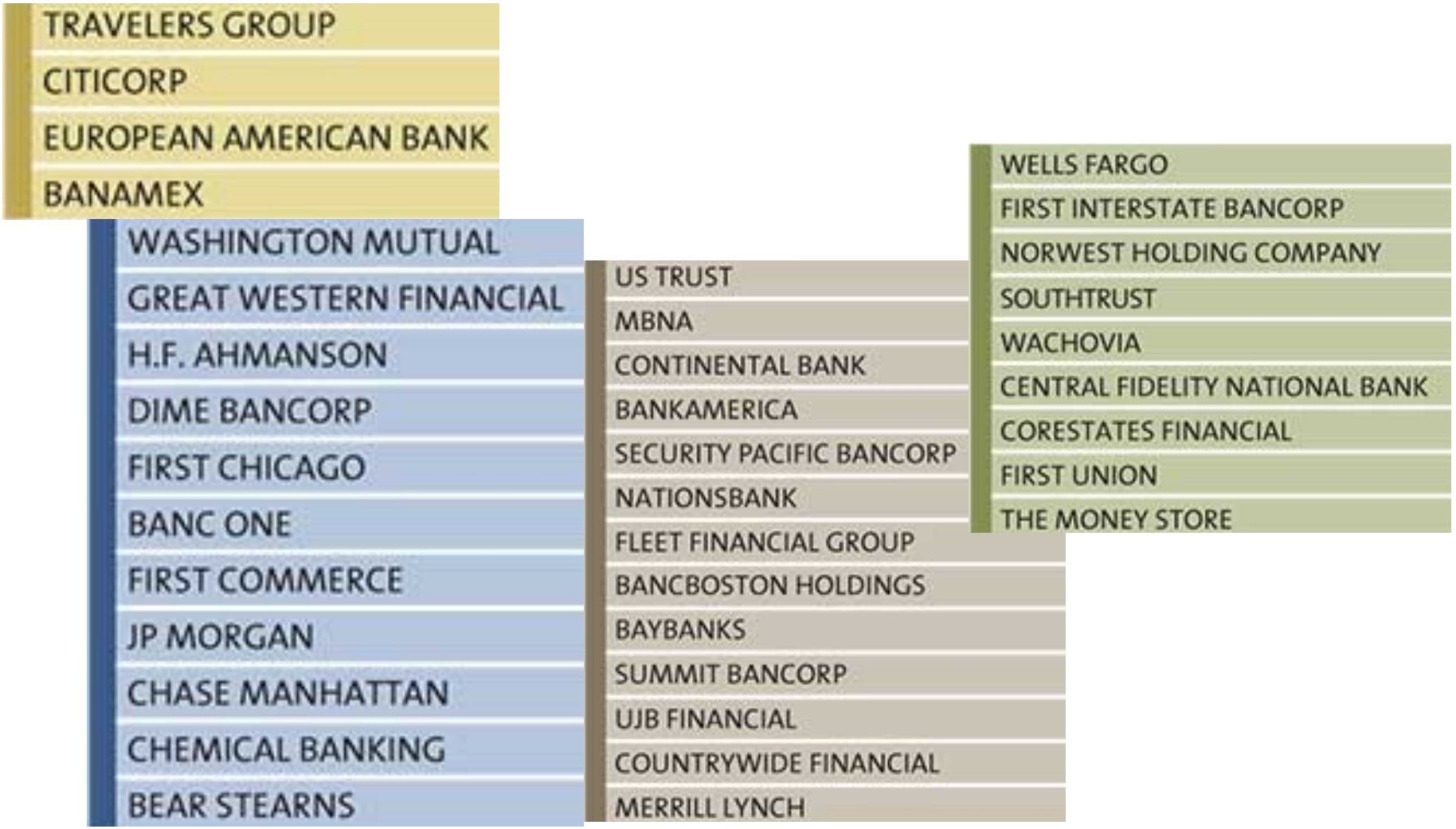
- 1980 – Several Companies
- 1983 – 90% of Media controled by 50 companies
- 1996 – 37 Banks began merging

After

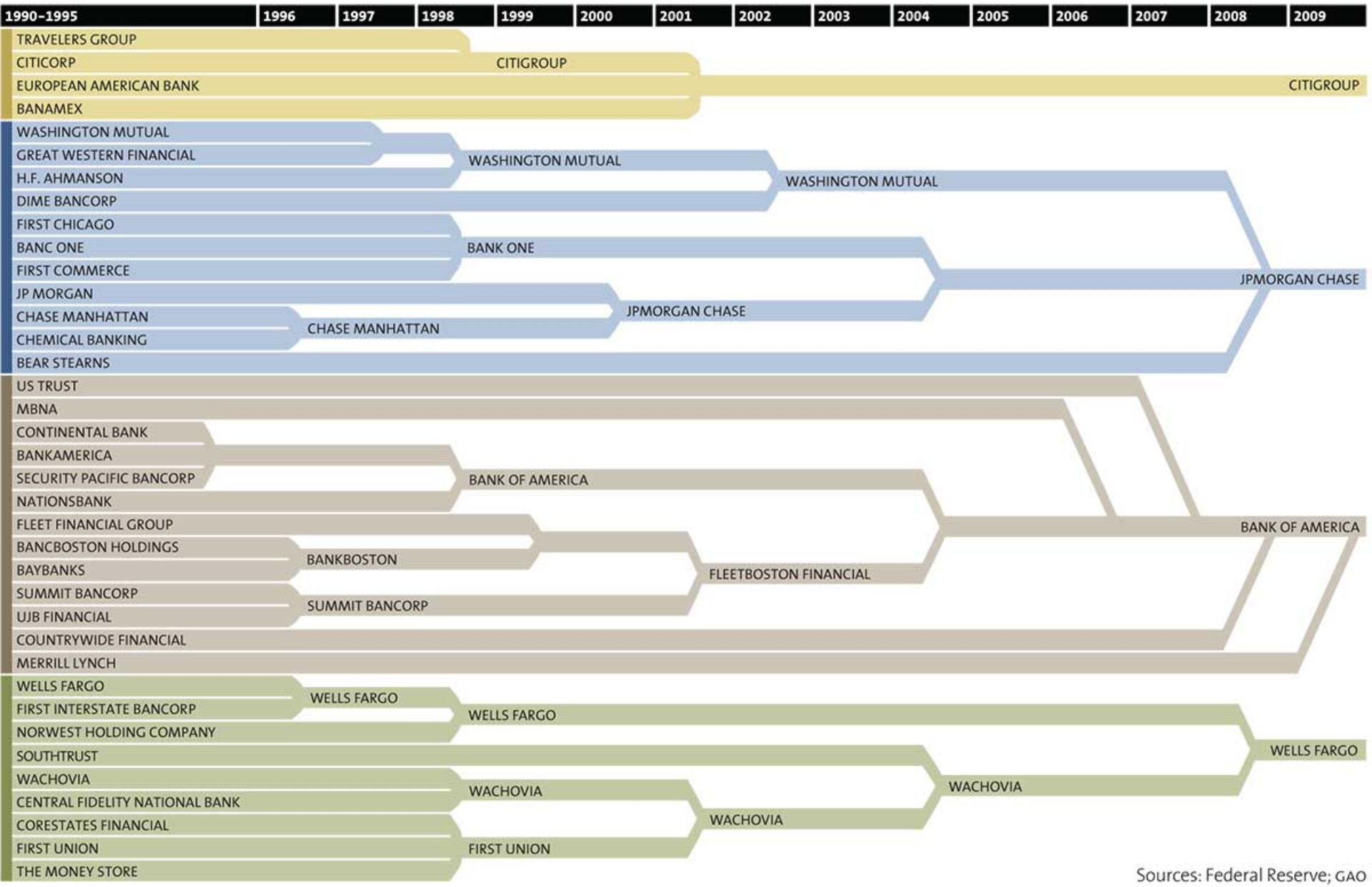
- 2009 – 10 control practically Everything
- 2001 – 90% of Media controlled by 06 companies
- 2009 – of 4 Banks (USA)

ALL FOR A FEW
(TUDO POR POUCAS!)

In 1995, 37 BANKS (USA)



Banks in USA start to merge and nowadays…



Sources: Federal Reserve; GAO

Bank merger boom in Brazil

Instituição compradora	Instituição comprada	Data	Instituição compradora	Instituição comprada	Data
Bradesco	BCN	Jun./98	Unibanco	Dibens	Jun./98
Bradesco	Bco. Crédito Real MG	Dez./97	Unibanco	Credibanco	Jun./00
Bradesco	Pontual (p/ BCN)	Dez./99	Unibanco	Bandeirantes	Dez./00
Bradesco	Baneb	Dez./99	Unibanco	BNL	Jun./04
Bradesco	Boavista	Dez./00	Santander	Bco. Geral do Comércio	Dez./97
Bradesco	Mercantil de SP	Mar./02	Santander	Noroeste	Dez./97
Bradesco	Bilbao Vizcaya	Jun./03	Santander	Bozano, Simonsen	Jun./98
Itaú	Banerj	Dez./96	Santander	Meridional	Set./00
Itaú	Bemge	Dez./98	Santander	Banespa	Mar./01
Itaú	Banestado	Dez./00	ABN Amro	Bco. Real	Jun./99
Itaú	BEG	Mar./02	ABN Amro	Sudameris	Dez./03
Itaú	BBA Creditanstalt	Set./02	HSBC	Bamerindus	Jun./98
			HSBC	Lloyds Bank	Mar./04

The product becomes less important than the experience of being bought!

A new consumer ...

New technologies ...

New behaviors ...

New RETAIL!

(Um Novo VAREJO!)

PROBLEM!!!! (TODAY SITUATION)

Huge masses of data!

QUOTE

"The data sets expand to fill the available space for storage"

"Os conjuntos de dados expandem para preencher o espaço disponível para estocagem"

(Cyril N. Parkinson, 1955)

THE SIZE OF THE DATA BASES

It is estimated that the amount of data in the world doubles every 20 month!

Nowadays, databases with Giga, Terabytes, Petabytes are common!

EXAMPLES

- *Up to 2003, it is believed that 5 exabites data were produced.*
- *Currently 5 exabites are produced every 2 days*

- *In the 2nd quarter of 2013:*
- *2.7 billion people online (40% of world population)*
- *6.8 billion of mobile phones*
- *Facebook: 1.1 billion of active users per month*
- *Twiter: 288 million of active users per month*
- *Youtube: 6 billion hours of videos watched per month*
- *Google: 360 million of active users*
- *Linkedn: 200 million of users, 2 new users every second*

Fonte: http://prezi.com/hyavodg6h0nb/big-data/

EXAMPLES IN BRAZIL

SERASA – Daily Transactions from a base of over 150 million CPFs

UOL – 50 terabytes of data daily!

Caixa Económica – Database with over 70 Terabytes

WHO/WHAT IS RESPONSIBLE FOR THIS?

Automatic Data Collection

electronic instrumentation

Online Transactions

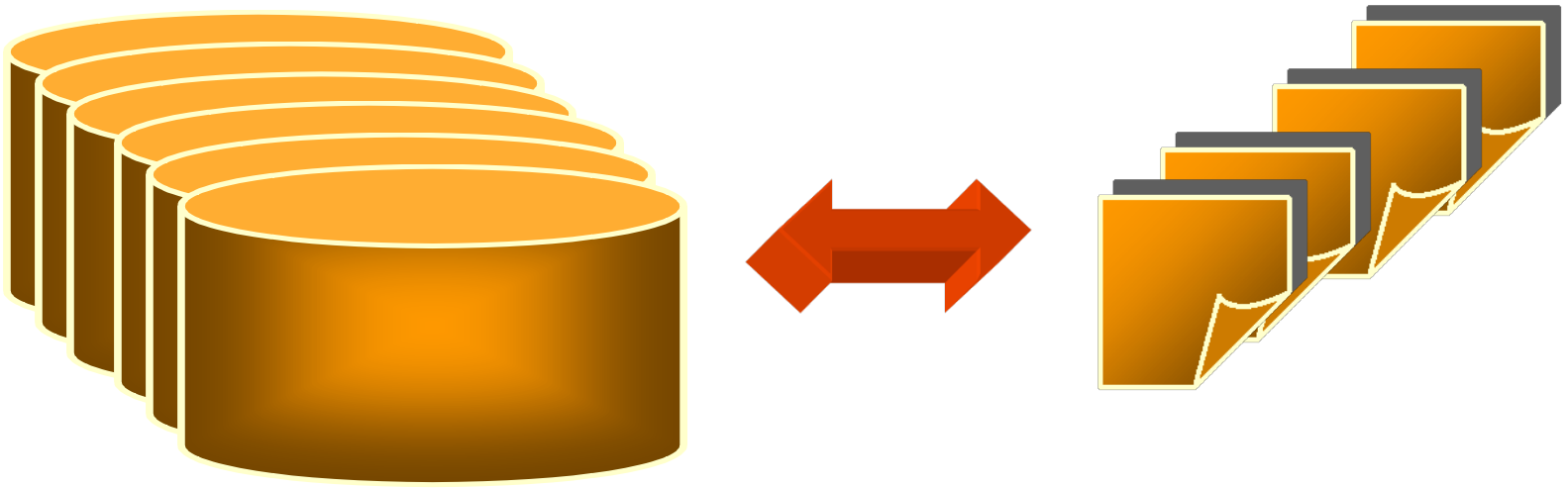
Historical data collected over many years

Multiple of Bite

Name	Symbol	Multiple
byte	B	2^{0}
kibibyte	KiB	2^{10}
mebibyte	MiB	2^{20}
gibibyte	GiB	2^{30}
tebibyte	TiB	2^{40}
pebibyte	PiB	2^{50}
exbibyte	EiB	2^{60}
zebibyte	ZiB	2^{70}
yobibyte	YiB	2^{80}

CURIOSITIES

- 1 Kilobyte = 2^{10} bytes ≅ 1/2 sheet A4
- 1 Gigabyte = 2^{30} bytes ≅ 55 sheet A4
- 1 Terabyte = 2^{40} bytes ≅ 51 Km sheet A4



QUESTIONS

- *What can we do with this large amount of data?*
- *How to identify and use information hidden in the data?*
- *How can we reverse this information on benefits (useful knowledge)?*

DATA MINING / BIG DATA

CAUTION: How to Search?

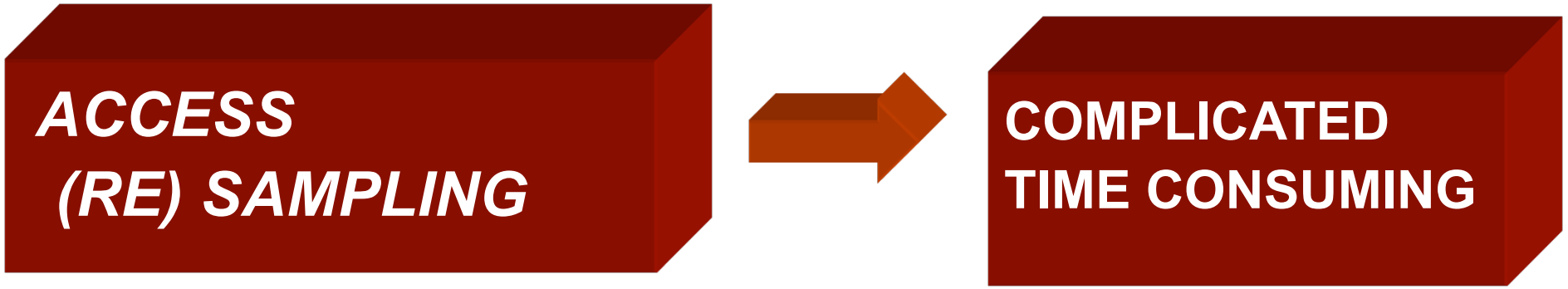
Statistical procedures should be properly adapted to this new reality.

PROBLEMS WITH USUAL STATISTICAL TOOLS FOR LARGE DATA SETS

- LARGE DATA SETS
- FIXING ERRORS
- CONTAMINATED DATA
- MISSING DATA
- DATA i i d
- NONSTATIONARITY
- NON-NUMERIC VARIABLES

THE SIZE OF THE DATASETS

MULTIPLE DATA FILES
(physically separated)



THE SIZE OF THE DATASETS

CONSEQUENCE

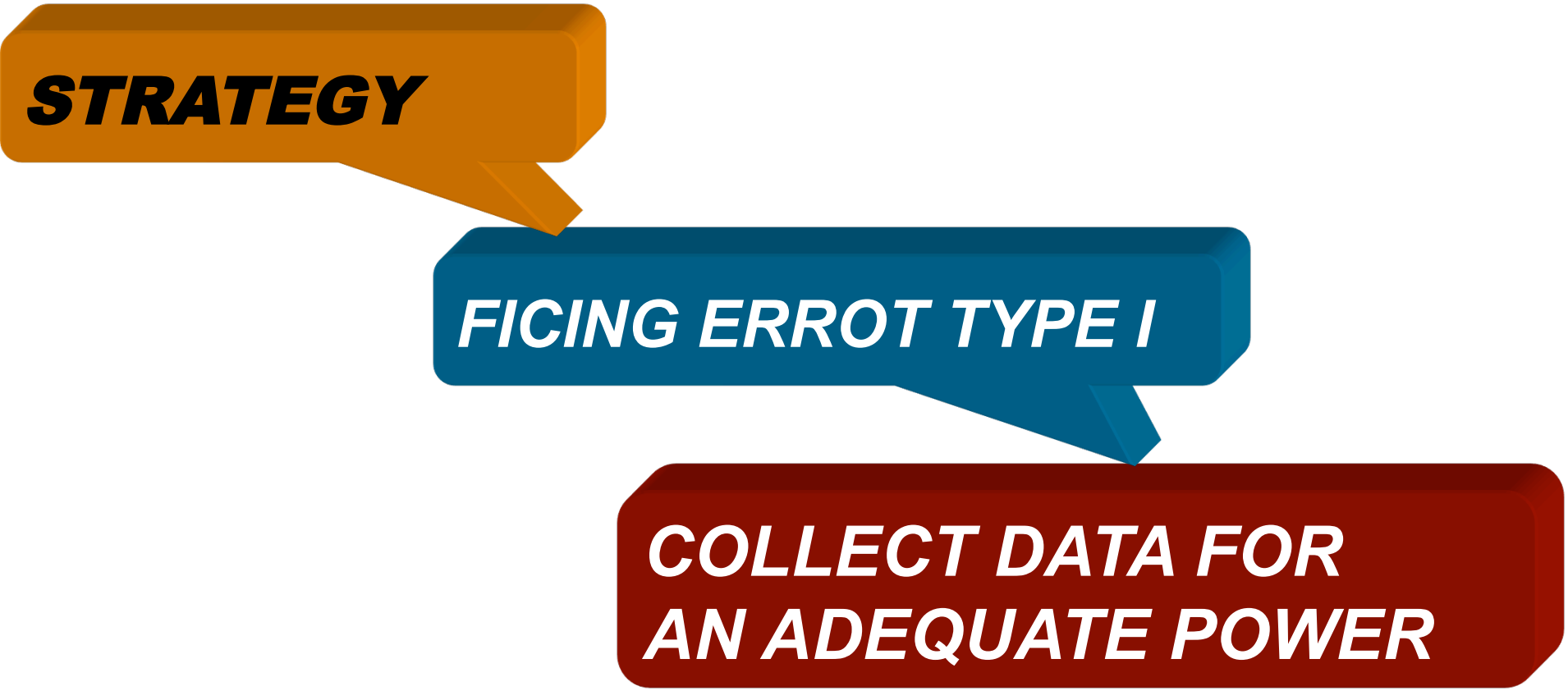
Direct statistical methods may not be applicable.

THE SIZE OF THE DATASETS

NEED FOR VARIANTS With stratification

Adaptive Methods Sequential Techniques

Fixing Errors



Fixing Errors

THIS STRATEGY IS QUESTIONABLE IN DATABASES EXTREMELY LARGE!

RESULTS ALWAYS INDICATE STRONG EVIDENCE, EVEN FOR SMALL ANY EFFECTS.

Fixing Errors

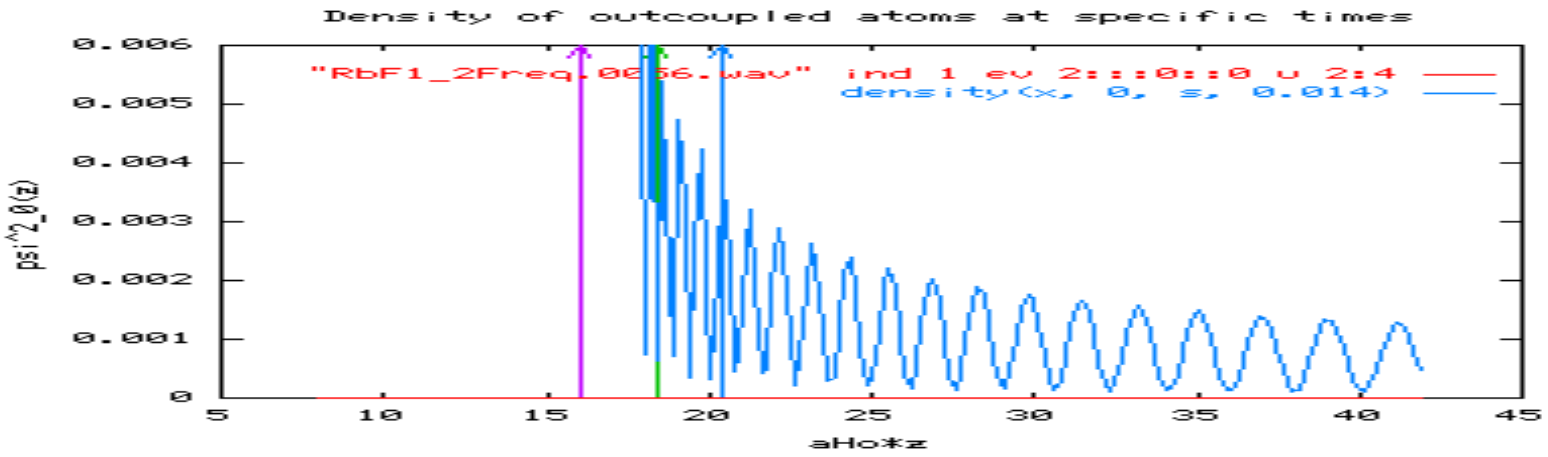
IMPORTANCE OF SUBJECTIVE SIGNIFICANCE

Answers to questions about the importance and/or value of the effect should be more decisive

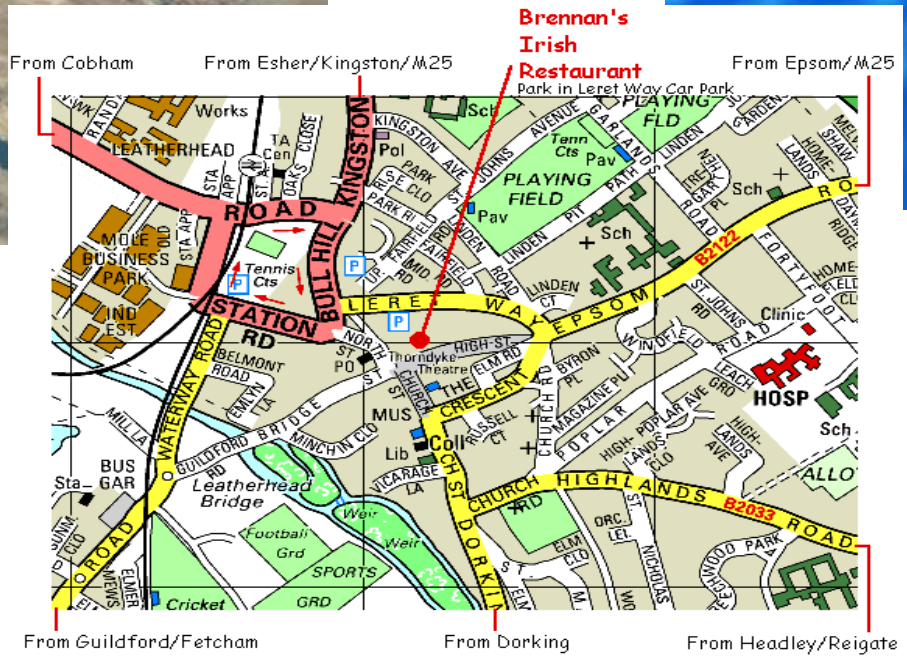
CONTAMINATED DATA / MISSING DATA

REAL TIME ANALYSIS

- DATABASE IS CONSTANTLY MODIFIED



NON-NUMERIC VARIABLES



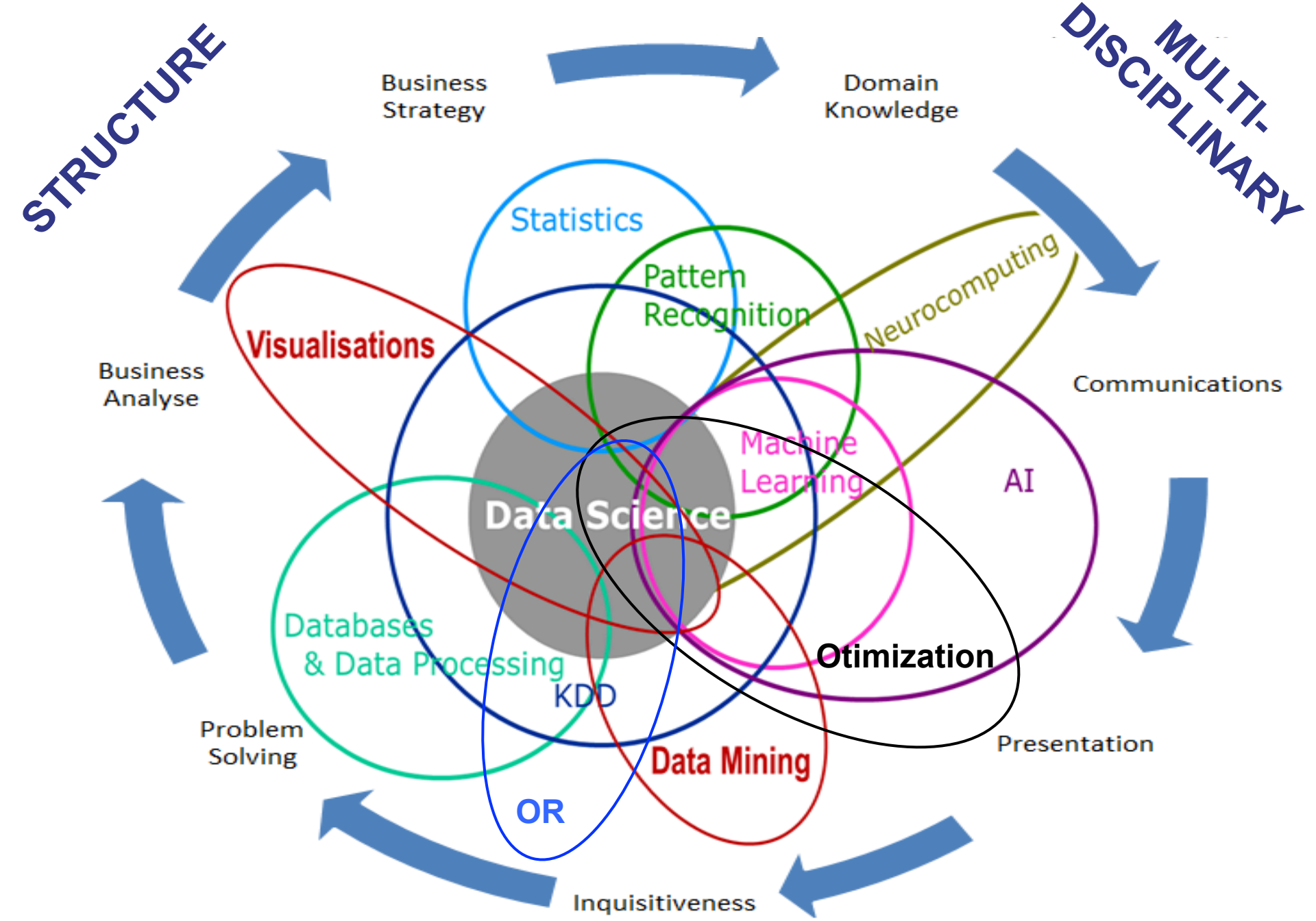
The Need...

A new professional

with multiple skills

"DATA SCIENTIST"

Statistical Inference	Data Base	Pattern Recognition / Artificial intelligence
Regression Modeling	Algorithm / Structured Programming	Graph Theory
Statistical Data Mining	Operational Research	Complex Networks
Sampling/ Design of Experiments	Data Visualization	Optimization


STRUCTURE
MULTI-DISCIPLINARY
Business Strategy
Domain Knowledge
Business Analyse
Communications
Problem Solving
Presentation
Inquisitiveness
Statistics
Pattern Recognition
Neurocomputing
Visualisations
Machine Learning
AI
Data Science
Databases & Data Processing
KDD
Otimization
Data Mining
OR

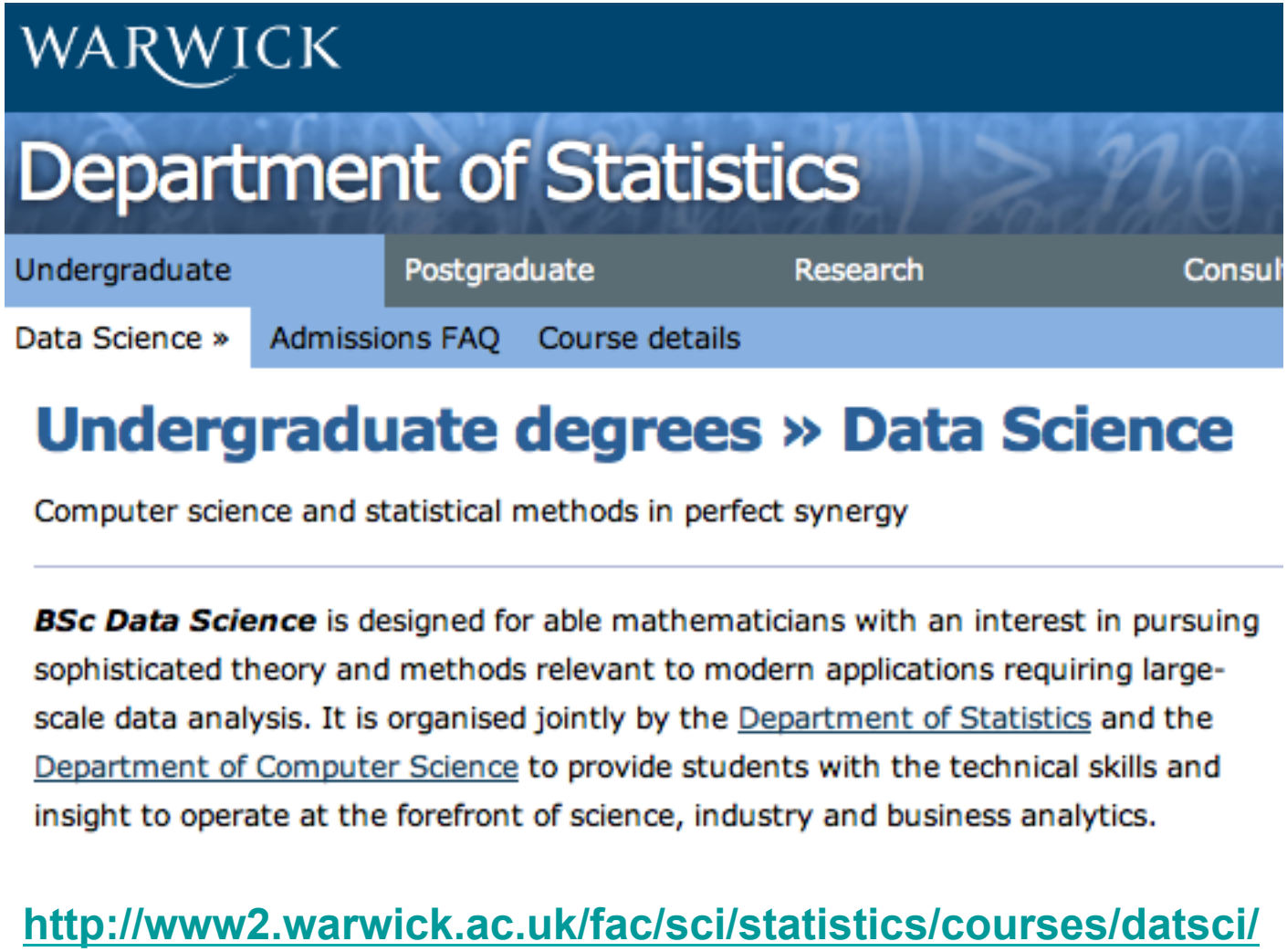
Market Need: ***Data Analysts, Statisticians, Mathematicians with Diversified Training who can work effectively in a multidisciplinary environment.***

(Necessidade do Mercado: Analistas de Dados, Estatísticos, Matemáticos com Formação Diversificada que possam trabalhar com eficiência em ambiente multidisciplinar)

DATA SCIENTISTS

In the world

DATA SCIENTIST in the world – Undergraduate



http://www2.warwick.ac.uk/fac/sci/statistics/courses/datsci/

DATA SCIENTIST in the world – Master

Master Program Overview

Skills acquired in this master program

- Seize the opportunity of data-driven value creation within an organization
- Recognize appropriate statistical methodologies and optimization techniques for complex problems
- Work with database management systems and distributed processing in a cloud computing environment
- Gain experience analyzing Big Data from the Internet of Things (industrial sensor data), the Internet of People (social and location data) and business transaction data
- Communicate data analysis results effectively with data presentation and aesthetic charting skills
- Work in a data-driven, heterogenous and research-oriented environment

- Course offerings and syllabus
- Program faculty

Join this master program

Apply for 2014-15 master programs

Masters begin September 2014

- Admission Criteria
- Application Deadlines

- All Masters Brochure [pdf]
- Data Science Brochure [pdf]
- Academic Calendar [pdf]

Master Degree in Data Science awarded jointly by:

barcelona gse graduate school of economics

upf. Universitat Pompeu Fabra Barcelona

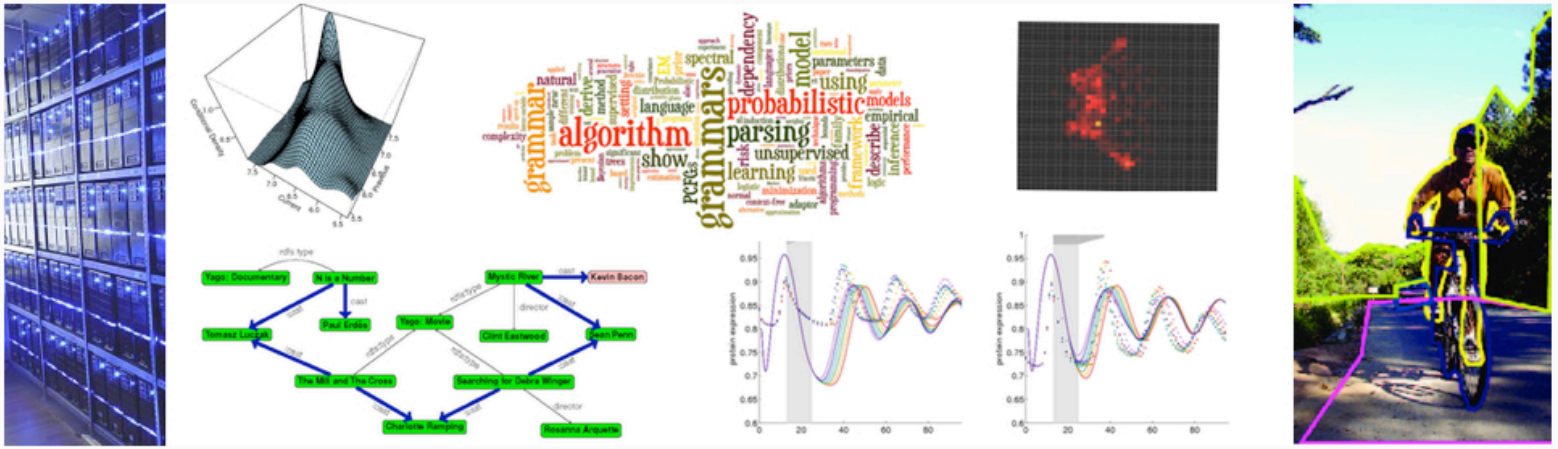
UAB Universitat Autònoma de Barcelona

http://www.barcelonagse.eu/master-data-science.html

DATA SCIENTIST in the world – PhD

DATA SCIENCE
Centre for Doctoral Training

Home About People Research The Course Partners Contact
Apply



What is data science?

Data science is the study of computational methods for extracting knowledge from data.

Read more

The Course

We have a four year PhD programme combining advanced coursework and independent PhD-level research.

Read more

Research

Check out some examples of the wide-ranging research in data science currently being performed by Centre faculty.

Read more

The People

Our team of over 45 academic staff carry out exciting research in all areas of data science.

Read more

http://datascience.inf.ed.ac.uk

AND HERE IN BRAZIL?

ARE WE PREPARED TO MEET THIS DEMAND?

WHAT WE HAVE MADE TO MEET THIS DEMAND?

INEVITABLY WE NEED TO THINK OF WAYS TO MEET THIS DEMAND, THAT IN MANY CASES HAVE BEEN ATTENDED (INEFFICIENTLY) BY PROFESSIONALS OF OTHER AREAS OF KNOWLEDGE!

(INEVITAVELMENTE PRECISAMOS PENSAR EM FORMAS DE SUPRIR ESTA DEMANDA, QUE, EM MUITOS CASOS TEM SIDO ATENDIDA (INEFICIENTEMENTE) POR PROFISSIONAIS DE OUTRAS ÁREAS DO CONHECIMENTO!)

SOME INITIAL POSSIBILITIES...

ACTION 1: UNDERGRATUATE COURSE - EMPHASIS "DATA SCIENCE"

IDEA

Proposition of an emphasis on the disciplines of Statistics, Applied Mathematics and Computing.

COURSES

1. STATISTICAL INFERENCE
2. REGRESSION MODELING
3. STATISTICAL DATA MINING
4. DATA BASE
5. ALGORITHM / STRUCTURED PROGRAMMING
6. ARTIFICIAL INTELIGENCE
7. DATA VISUALIZATION
8. OTIMIZATION
9. COMPLEX NETWORKS

ACTION 2: PROFESSIONAL MASTER– MODULE "DATA SCIENCE"

IDEA

- Proposition a module in our Professional Master Program MECAI.

COURSES

1. STATISTICAL INFERENCE
2. REGRESSION MODELING
3. STATISTICAL DATA MINING
4. DATA BASE
5. ALGORITHM / STRUCTURED PROGRAMMING
6. ARTIFICIAL INTELIGENCE
7. DATA VISUALIZATION
8. OTIMIZATION
9. COMPLEX NETWORKS

ACTION 3: UNDERGRADUATE COURSE IN “DATA SCIENCE”

ACTION 4: MASTER/PhD in “DATA SCIENCE”

EDUCATION

PROFISSIONAL MASTER PROGRAM

Professional Master Program

2014 (August)

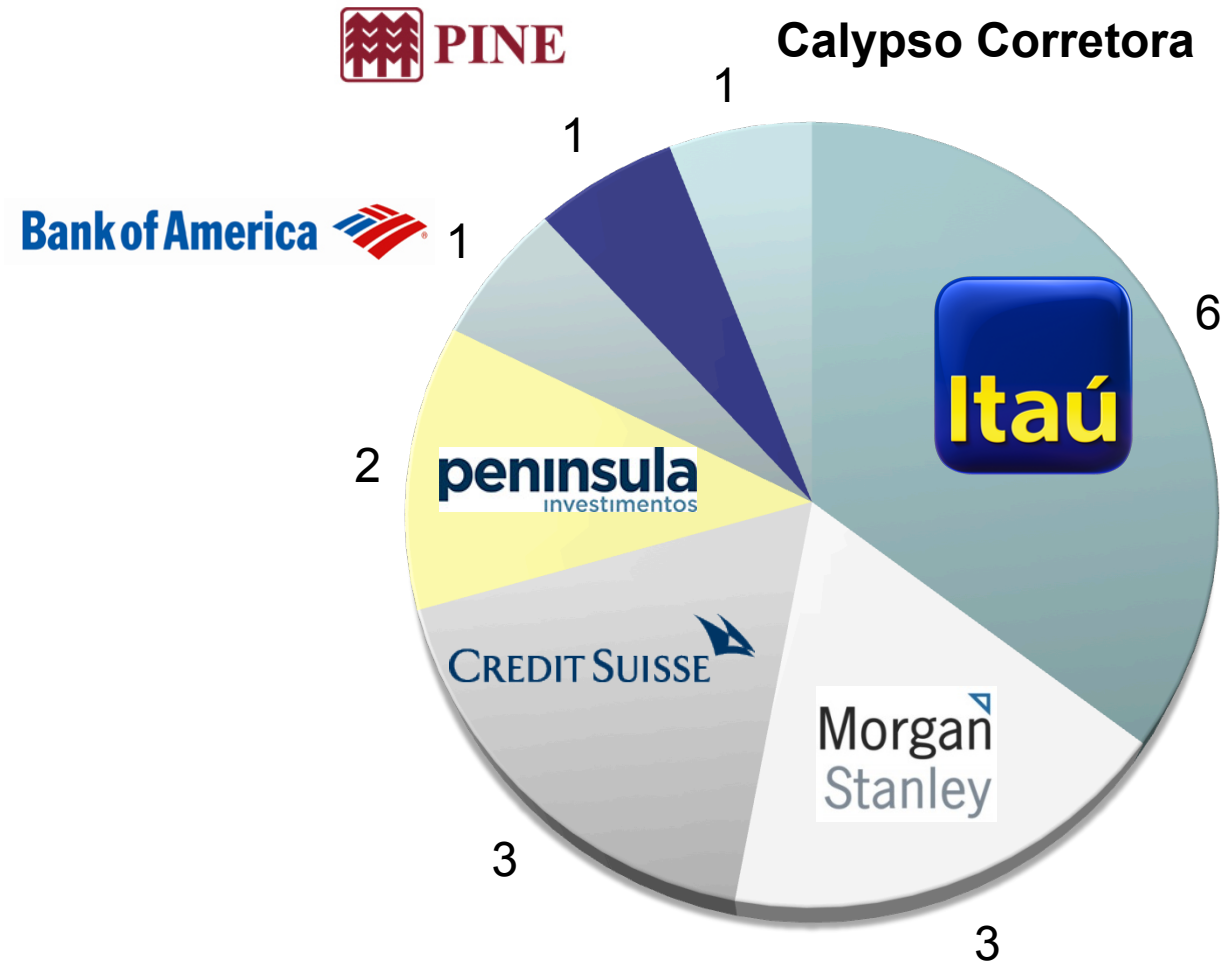
- Professional Master Program on Mathematics, Statistics and Computing Applied to Industry (Mestrado Profissional em Matemática Aplicada, Estatística e Computação Aplicadas à Indústria - MECAI)

- FISRT MODULE: FINANCE
- SECOND MODULE: DATA SCIENCE

  (from August 2015)

STUDENT PROFILE

Module Finance



BUT WE ALSO NEED SPECIALIZED LECTURES

- HIRING DATA SCIENTISTS
- MATHEMATICIAN/STATISTICIAN WITH STRONG COMPUTER TRAINING AND VARIOUS OTHER ABILITIES.

BUT THIS IS SUBJECT TO ANOTHER TALK!

Quote

"The difficulty lies not so much in developing new ideas as in escaping from old ones."

(John Maynard Keynes)

"A dificuldade não está tanto no desenvolvimento de novas idéias mas como escapar da antigos."

(John Maynard Keynes)

THANKS!
(MUITO OBRIGADO!)

www.mwstat.com/franciscolouzada