TÉCNICO LISBOA

DADOS e sua influência no desenvolvimento da ESTATÍSTICA

João A. Branco

Departamento de Matemática, IST-UL

Escola de Inverno de Matemática

Lisboa, 2 e 3 de Fevereiro de 2015

Antes de começar

Algumas perguntas que se podem fazer

- O que é a Estatística? Para que serve e a quem serve?
- Qual a relação da Estatística com as outras ciências e a Matemática em particular?
- Como nasceu e qual o seu percurso?
- Qual a sua importância actual e qual o seu futuro?

A preocupação central da estatística é a extracção de informação a partir de dados

Qual o objectivo deste Minicurso?

Salientar que a Estatística é de grande utilidade na resolução de uma enorme diversidade de problemas e que a natureza dos dados influencia decisivamente o seu desenvolvimento.

TÉCNICO LISBOA

Sumário

Parte I - Dados

Parte II - Modern Data

Parte I - Dados

The Compeling Need:

> "*Data! Data! Data!*" he cried impatiently. "*I can't make bricks without clay.*"
>
> So said Sherlock Holmes to Dr.Watson in The Adventure of the Copper Beeches (Conan Doyle, 1892)

Dados são o alimento da Estatística

Sem dados não há Estatística!

Parte I - Dados

Afinal o que são Dados?

São o resultado da observação ou medição de características de objectos, indivíduos ou fenómenos.

João, 25 anos, 168cm de altura, solteiro, motorista, salário baixo,

são dados relativos a esta pessoa.

Não confundir Dados com as formas geralmente usadas para os descrever (números, palavras, figuras, gráficos e imagens):

25 anos - transporta informação sobre a idade do João
25 - como número isolado não transporta informação

Parte I - Dados

Consideram-se geralmente 3 fases no tratamento de Dados:

Recolha - motivada por razões objectivas (resolver um problema ou explicar um fenómeno), ou por mera rotina sem a preocupação de que os Dados venham a ser analisados (pode seguir-se a via observacional ou a via experimental)

Anterior à recolha - fase de reflexão e planeamento

Posterior à recolha - consiste na análise dos Dados

Parte I - Dados

Três exemplos históricos:

Primeiros sinais de Primavera - Robert Marsham iniciou o registo dos primeiros sinais em 1736 e a sua família continuou até 1947 (mais de 200 anos).

Mais de 400 esqueletos - Numa expedição a Naqada (Egipto - finais do séc. XIX) foram recolhidos mais de 400 esqueletos. Em Londres Karl Pearson mediu 48 variáveis nos respectivos crânios (um projecto de estudo da raça humana).

Um observatório privado - Miguel Franzini, político e homem de ciência mandou construir um pequeno observatório junto à sua residência onde observou (1816 - 1855) condições atmosféricas por considerar que a informação seria útil para a caracterização do clima local.

Análise de Dados (AD):

A Análise de Dados preconizada por John Tukey (1962) surge como uma manifestação de insatisfação com o uso sistemático de métodos de Inferência Estatística introduzidos por Ronald Fisher (1922). Estes métodos caracterizam-se por uma certa rigidez e apresentam limitações, pois assumem a existência de um único modelo que explica os dados, modelo esse que assenta em hipóteses convenientes que nem sempre se verificam na prática.

O que é então a Análise de Dados?

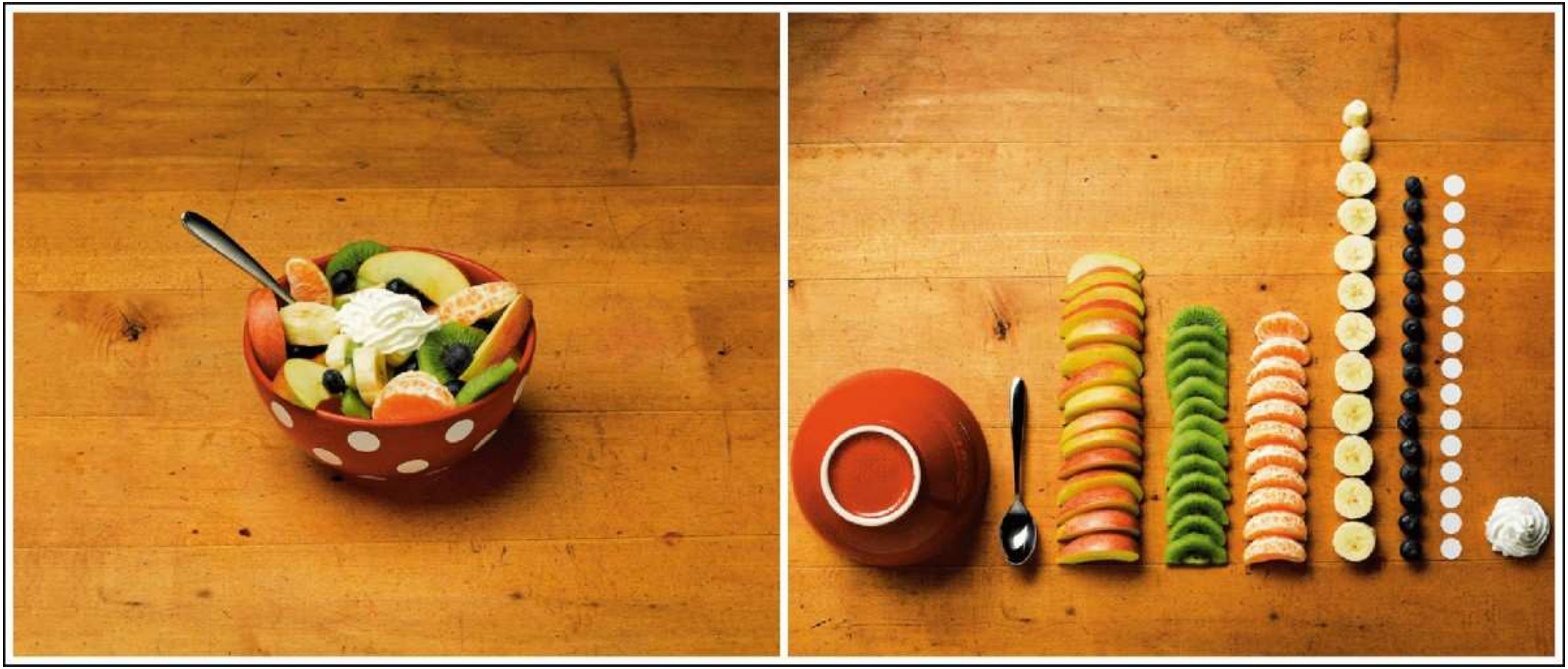
Eis a definição de John Tukey:

> *All in all, I have come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.*
>
> *One should do everything sensible, using all means, arguments and tools, to answer as better as possible the question that we think only the data can help to answer properly.*

Parte I - Dados

A AD é uma abordagem muito flexível que requer imaginação e persistência ao sugerir a procura de vários caminhos, muitos dos quais podem não ser adequados.

Contudo, há caminhos que parecendo obrigatórios não fazem qualquer sentido.

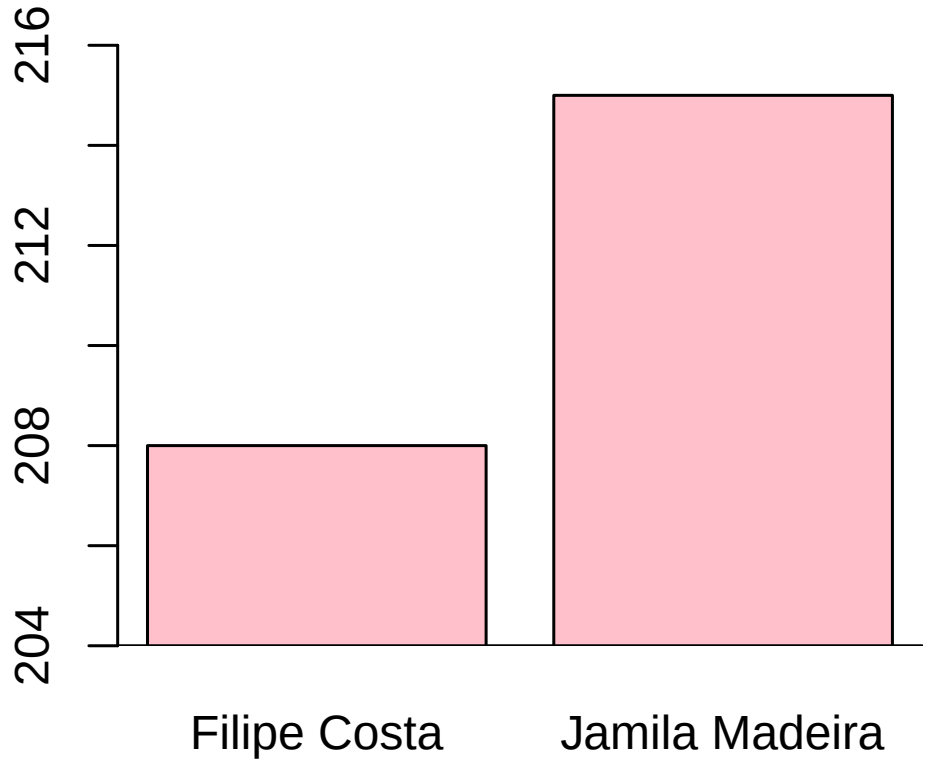



请勿乱扔杂物

Parte I - Dados

A flexibilidade da AD também não permite abusos, manipulações e tortura!

Ninguém apanha a Jamila
(Sondagem, Expresso, 22/6/2002)



Parte I - Dados

Dados fabricados (intencionalmente?)

Greve dos professores (17.10.2006) - Público

14 Sindicatos ?! 85% de adesão

Ministério da Educação ?! 39% de adesão

(É indispensável perguntar quem produziu a estatística que nos é apresentada)

Parte I - Dados

"*O senhor primeiro-ministro já nos habituou a torcer... ou melhor, a torturar as estatísticas.*" — Jerónimo de Sousa, no dia 4 de Junho de 2011, no Parlamento, em relação às afirmações de José Sócrates, que conseguiu ver sinais positivos nos 10.8 por cento de desemprego.

Tipo de Dados

Uma AD deve começar por identificar o tipo de Dados pois isso reduz o leque de métodos estatísticos disponíveis.

A classificação habitual consiste em atribuir aos Dados o nome das variáveis que lhe deram origem:

Quantitativos: contínuos (peso)
discretos (número de acidentes)

Qualitativos: ordinais (salário: baixo, médio, alto)
nominais (estado civil)

Parte I - Dados

Outras designações para Dados

Muitas vezes a designação atribuída a Dados assenta no nome da área de trabalho em que os Dados foram recolhidos, do próprio processo de recolha ou de alguma característica específica importante.

Ex: Dados experimentais, observacionais, univariados, multivariados, categorizados, composicionais, longitudinais, retrospectivos, prospectivos, de sobrevivência, sequenciais, de séries temporais, de funções, espaciais, de sondagens, econométricos, das ciências sociais e outros.

Parte I - Dados

Métodos estatísticos

Para além dos métodos gerais de Estatística descritiva e de inferência estatística (estimação e testes de hipóteses) os métodos de regressão e correlação, de estatística não paramétrica, análise de variância e delineamento experimental são requeridos por muitos utilizadores. Paralelamente a estes métodos fundamentais existe uma grande abundância de métodos especiais concebidos para tratamento dos vários tipos de dados mencionados.

Ex: Métodos para dados multivariados (Componentes Principais, Análise Factorial, Análise de clusters, Análise de correspondências, e outros).

Métodos para aAnálise de sobrevivência (Kaplan Meier, riscos proporcionais, e outros)

Parte I - Dados

Fala-se ainda de dados omissos, reais, artificiais, simulados, forjados,. . . e também de

Poucos Dados

Será preferível ter Muitos Dados a Poucos Dados?

Parte I - Dados

Nem sempre, pois aumentando a informação (dimensão da amostra) pode sempre chegar-se ao ponto de rejeitar uma dada hipótese, H_0. Porém o que é estatisticamente significante pode não ser cientificamente significativo.

Dois tratamentos podem ser estatisticamente diferentes (diferença significante estatisticamente) mas a diferença que levou à rejeição de H_0 pode não corresponder a um efeito real/palpável na saúde dos doentes (diferença clinicamente não significativa/não importante).

Branco, J.A. e Pires, A.M., Poucos dados não é derrota e muitos dados não é vitória. Em Ferrão, M.E., Nunes, C. e Braumann, C.A. (editores), Estatística Ciência Interdisciplinar: Actas do XIV Congresso Anual da SPE, Edições SPE, Lisboa, pp. 257–268, 2007.

Parte I - Dados

À medida que novos problemas, trazendo novos dados, vão surgindo, a Estatística responde com novos métodos.

Os Dados são, por isso, o motor de desenvolvimento da Estatística.

FIM

Part II - Modern Data

The Data Age

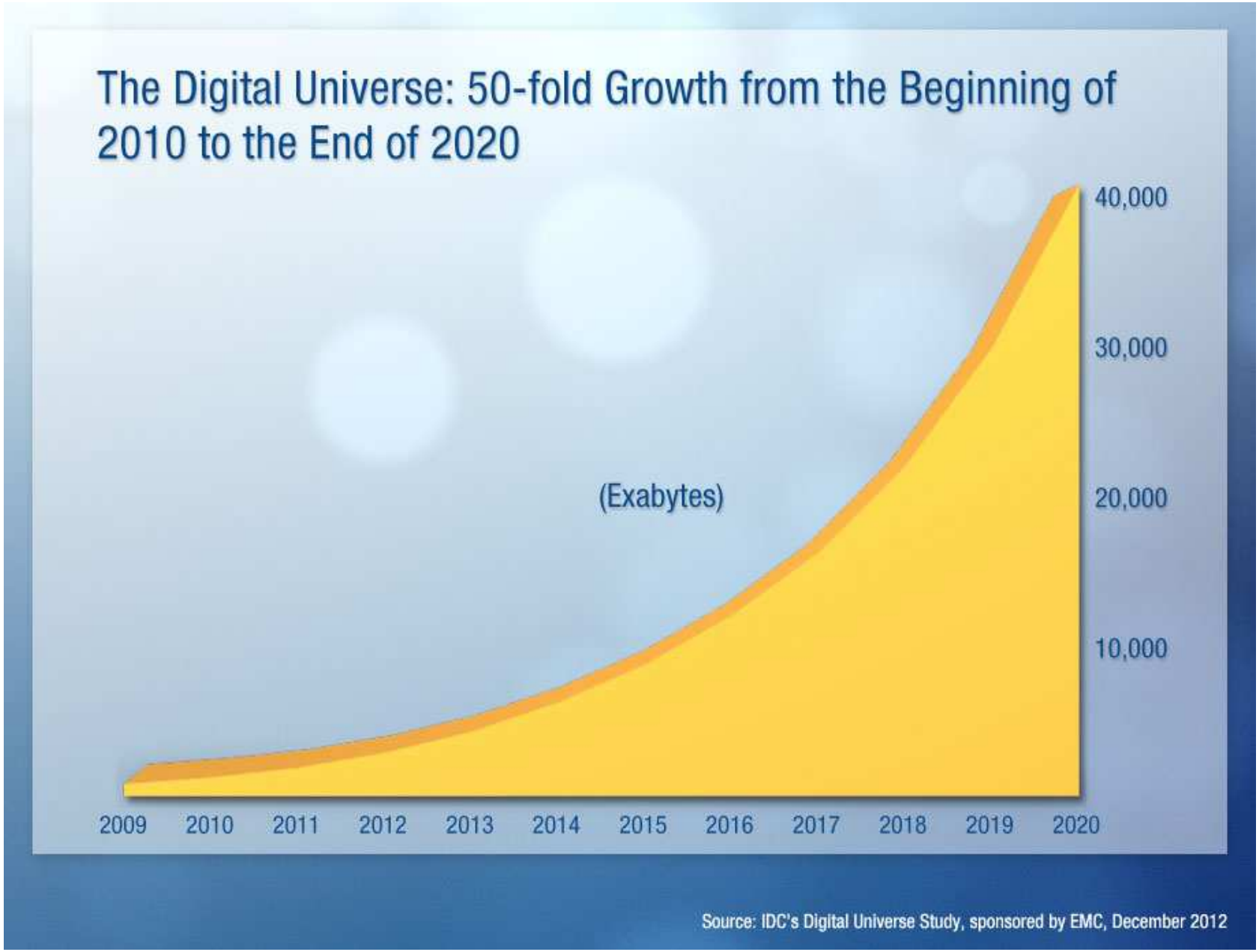
- Conventional statistics has privileged small to moderate data sets (preferably homogeneous, sanitized data). Hand, D.J. (1995). *A Handbook of Small Data Sets*.

- New technology allows automatic/quick measurement. The production of data in every field is enormous (tsunami of data). Data has become a factor of production (it is a new oil/soil).

  Switzerland new business — going from banks storing money to bunkers storing Data.

- Decisions will increasingly be based on data and results of DA rather than on experience and intuition.

How Data is growing



What does new data look like?

- Unstructured, non-homogeneous, unsanitized, incomplete, messy and large. These data come in great volumes, at great speed and different forms: numerical, images, text, maps and networks.
- tiny (10^2 bytes), small (10^4), medium (10^6), large (10^8), huge (10^{10}), monster (10^{12}) (Huber, 1994, classification).
- Abello, J. et al, editors (2002). *Handbook of Massive Data Sets.*
- More recently one has BIG DATA, big n, big p ($n > p$, $n < p$):
  - Astronomy: SDSS (200 GB per night)
  - FICO: protects 2.1 billion active accounts world wide
  - Genomics
  - Internet text and documents
  - Social networks data

Part II - Modern Data

- How can these data be analysed?
- Who is analyzing these modern data?
- What are statisticians (Bayesians and frequentists) doing?
- **What am I doing?**

Part II - Modern Data

It is hoped that statistics can help in the analysis of modern data.

However a multitude of datasets have a very large number of observations (n) and/or variables (p), and that raises difficulties because traditional statistical methods were not designed to deal with large datasets:

- computational difficulties (storage, handling, processing, ... using available computer systems)
- difficulties with the analysis (how to do the analysis of data when $p \gg n$?)

Definition of Mahalanobis distance (MD):

For a univariate sample, $(x_1, \ldots, x_n)$, with mean $\overline{x}$ and variance s^2:

$$\text{MD to the center} \qquad D^2_{x,\overline{x}} = \frac{(x - \overline{x})^2}{s^2}$$

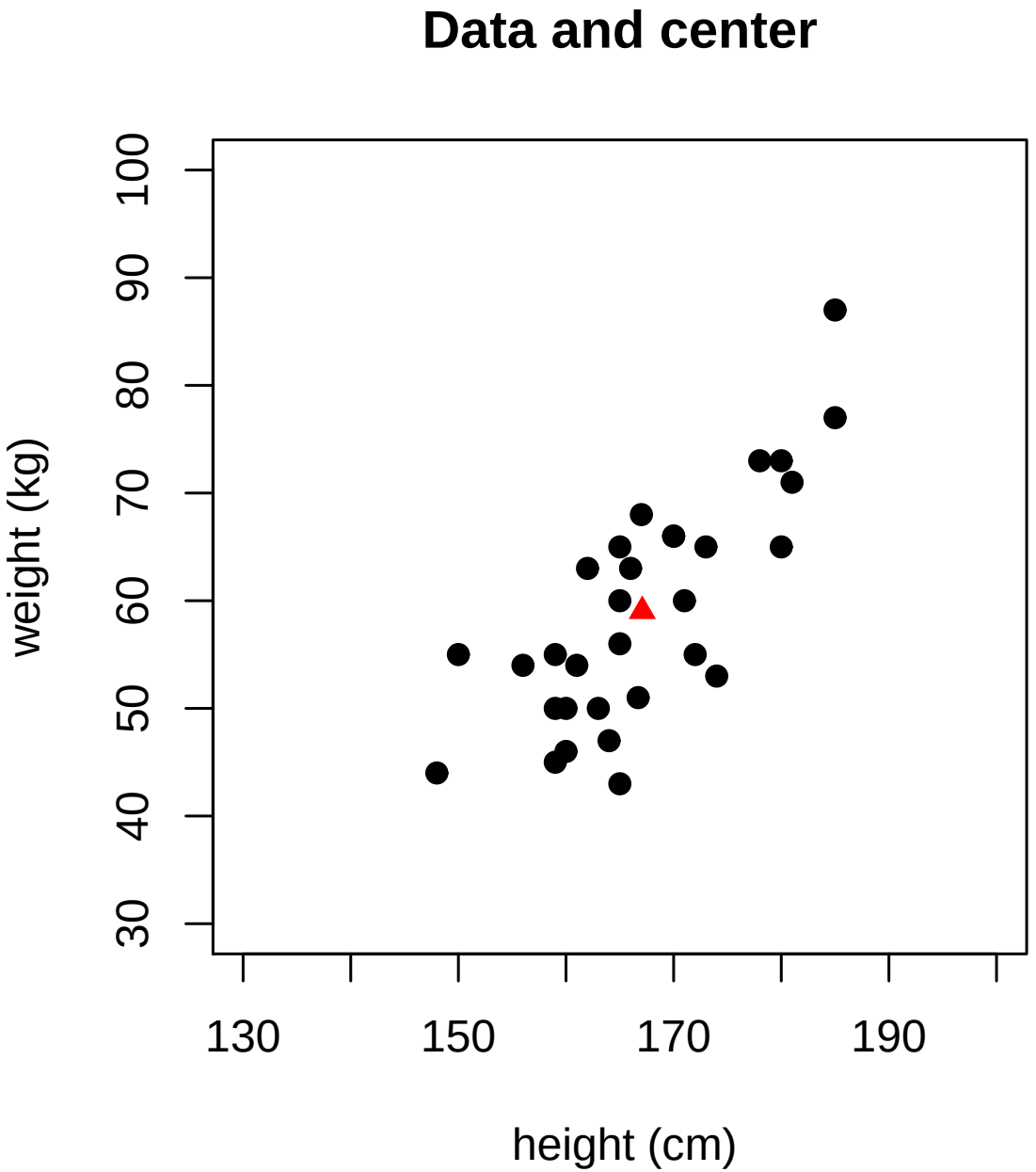
$$\text{MD between 2 points} \qquad D^2_{x_i,x_j} = \frac{(x_i - x_j)^2}{s^2}$$

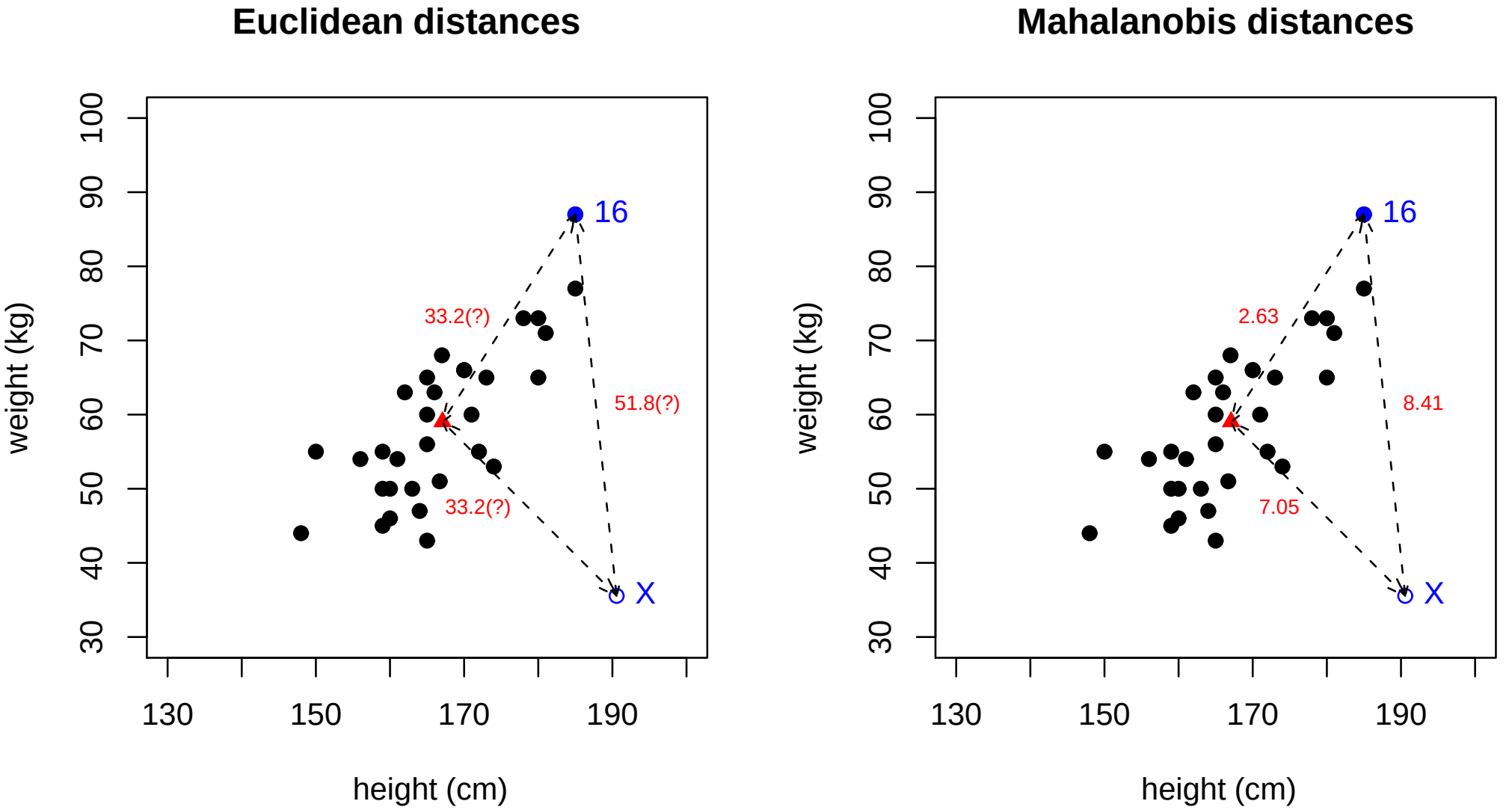
For a multivariate sample, $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, with mean vector $\overline{\mathbf{x}}$ and covariance matrix $\mathbf{S}$:

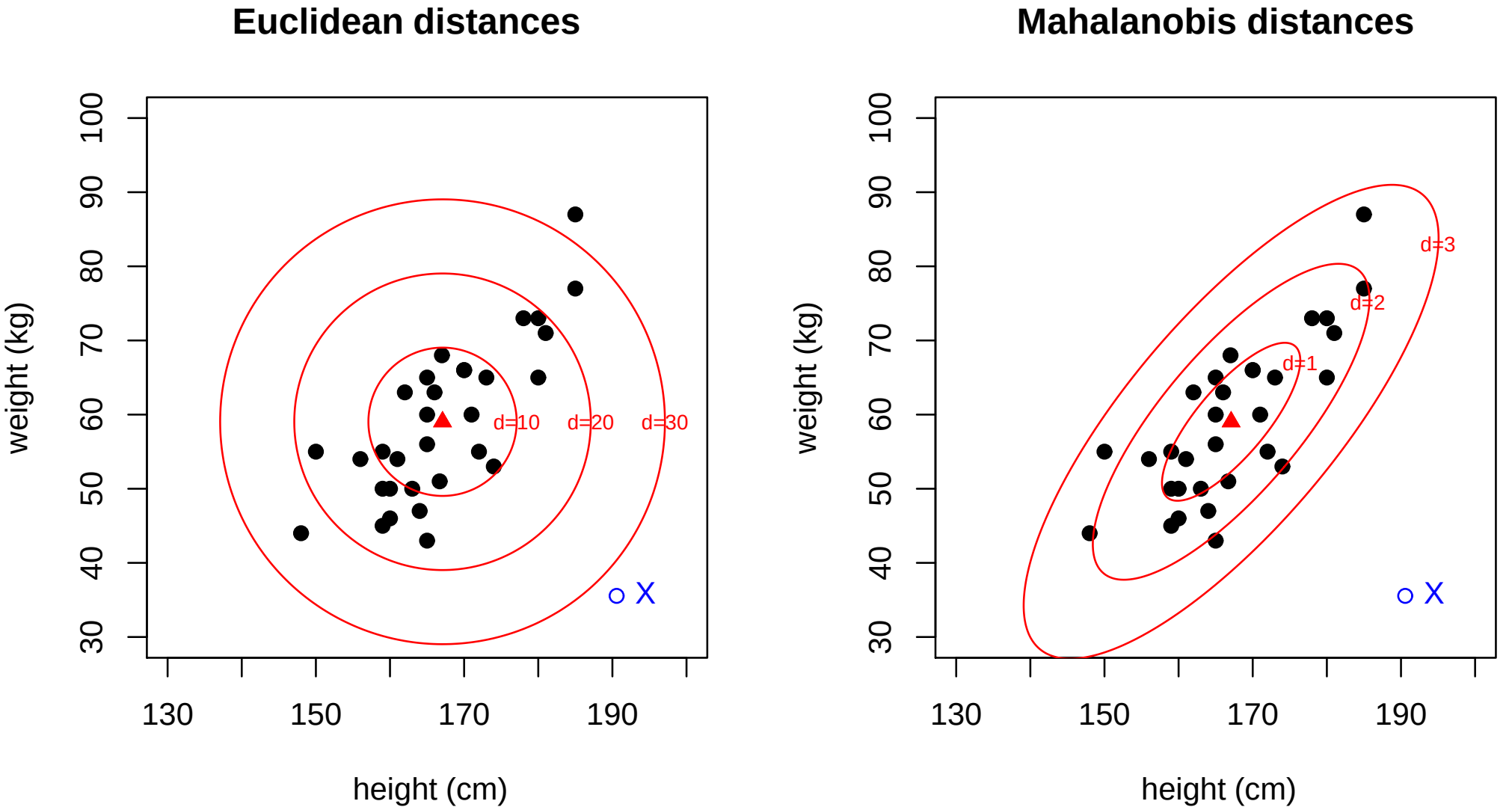
$$\text{MD to the center} \qquad D^2_{\mathbf{x},\overline{\mathbf{x}}} = (\mathbf{x} - \overline{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x} - \overline{\mathbf{x}})$$

$$\text{MD between 2 points} \qquad D^2_{\mathbf{x}_i,\mathbf{x}_j} = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j)$$

MD versus ED:


Data and center
weight (kg)
30
40
50
60
70
80
90
100
130
150
170
190
height (cm)


Euclidean distances
weight (kg)
height (cm)
30
40
50
60
70
80
90
100
130
150
170
190
16
X
33.2(?)
51.8(?)
33.2(?)
Mahalanobis distances
weight (kg)
height (cm)
16
X
2.63
8.41
7.05


Euclidean distances
weight (kg)
height (cm)
d=10
d=20
d=30
X
Mahalanobis distances
weight (kg)
height (cm)
d=1
d=2
d=3
X

Theorem: For any data matrix referring to n observations and $p \geq n-1$ variables, such that $\mathbf{S}$ has the maximum possible rank, equal to $n-1$, we have that

(i) $$D^2_{\mathbf{x}_i,\overline{\mathbf{x}}} = (\mathbf{x}_i - \overline{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x}_i - \overline{\mathbf{x}}) = \frac{(n-1)^2}{n}$$ for all $i = 1, \ldots, n$

(ii) $$D^2_{\mathbf{x}_i,\mathbf{x}_j} = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j) = 2\,(n-1)$$ for all i and j, with $i \neq j$

This means that:

Any data set of n observations measured in $p \geq n-1$ variables,

is transformed by $\mathbf{z}_i = (\mathbf{S}^{-})^{1/2}(\mathbf{x}_i - \overline{\mathbf{x}})$ into a set such that:

- all points are at the same Euclidean distance from the center,
- the distance between the two points of any pair of points are all the same.

Part II - Modern Data

And the implications are:

1. It is impossible to distinguish outliers from non outlier observations
2. It is impossible to detect any kind of deviations from symmetric structures
3. It is not possible to distinguish between linear and non linear structures.

Part II - Modern Data

These datasets are attracting a number of people

(data scientists – data science)

Who are they?

What is data science?

According to Provost and Fawcett (2013):[1]

- Data science is a set of fundamental principles that support and guide the extraction of information and knowledge from data.
- Closely related to data mining (but based on a much smaller and more concise set of fundamental principles).
- A large portion of what has traditionally been studied within the field of statistics is fundamental to data science.
- Methods for visualizing data are vital.
- Often, intuition, creativity, common sense, and knowledge of a particular application must be brought to bear.

[1] F. Provost and T. Fawcett, Data science and its relationship to big data and data-driven decision making. *Big Data*. March 2013, **1(1)**: 51–59.

Part II - Modern Data

Companies are very fond of finding FAST ALGORITHMS to deal with the ever growing sets of data.

Fast cars, fast women, fast algorithms. . . what more could a man want?

(Joe Mattis)

Crowdsourcing DA

- New York, Chicago, Seattle are making city data publicly available.
- Various platforms are making data available for scientific, industrial and education challenges/competitions

Logic: opening a problem to millions of people will achieve a better result than could be achieved through traditional research.

Popular competitions/platforms

- Biomag 2012 (decoding brain activity)
- KDD (since 1997)
- Datafest (UCLA Dep. of Stat., combines stats with speed)
- INNOCENTIVE
- CrowdANALYTIX
- TUNEDIT
- Kaggle

Kaggle competitions

91 competitions since 2010

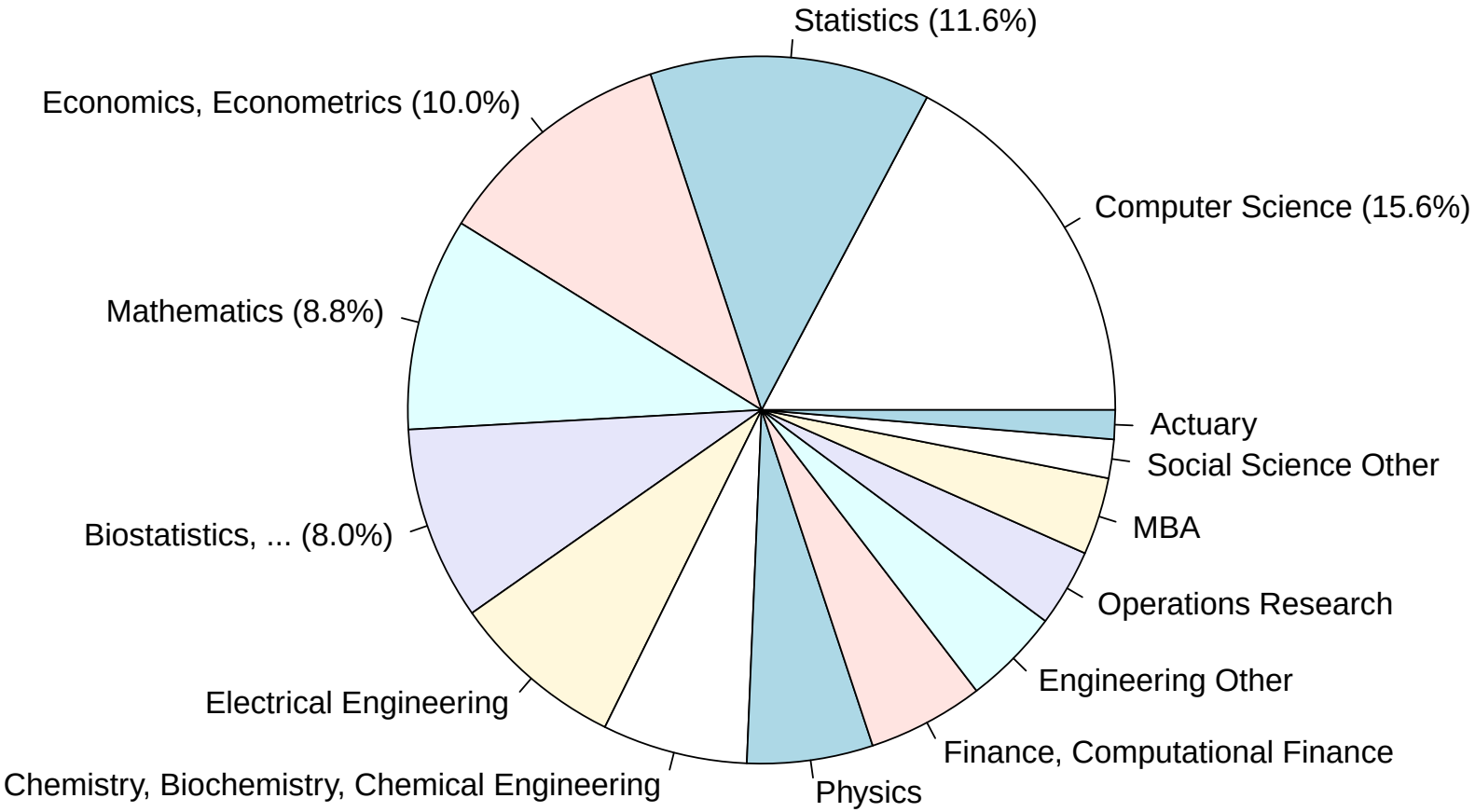
15447 teams

87530 analysts (data scientists)

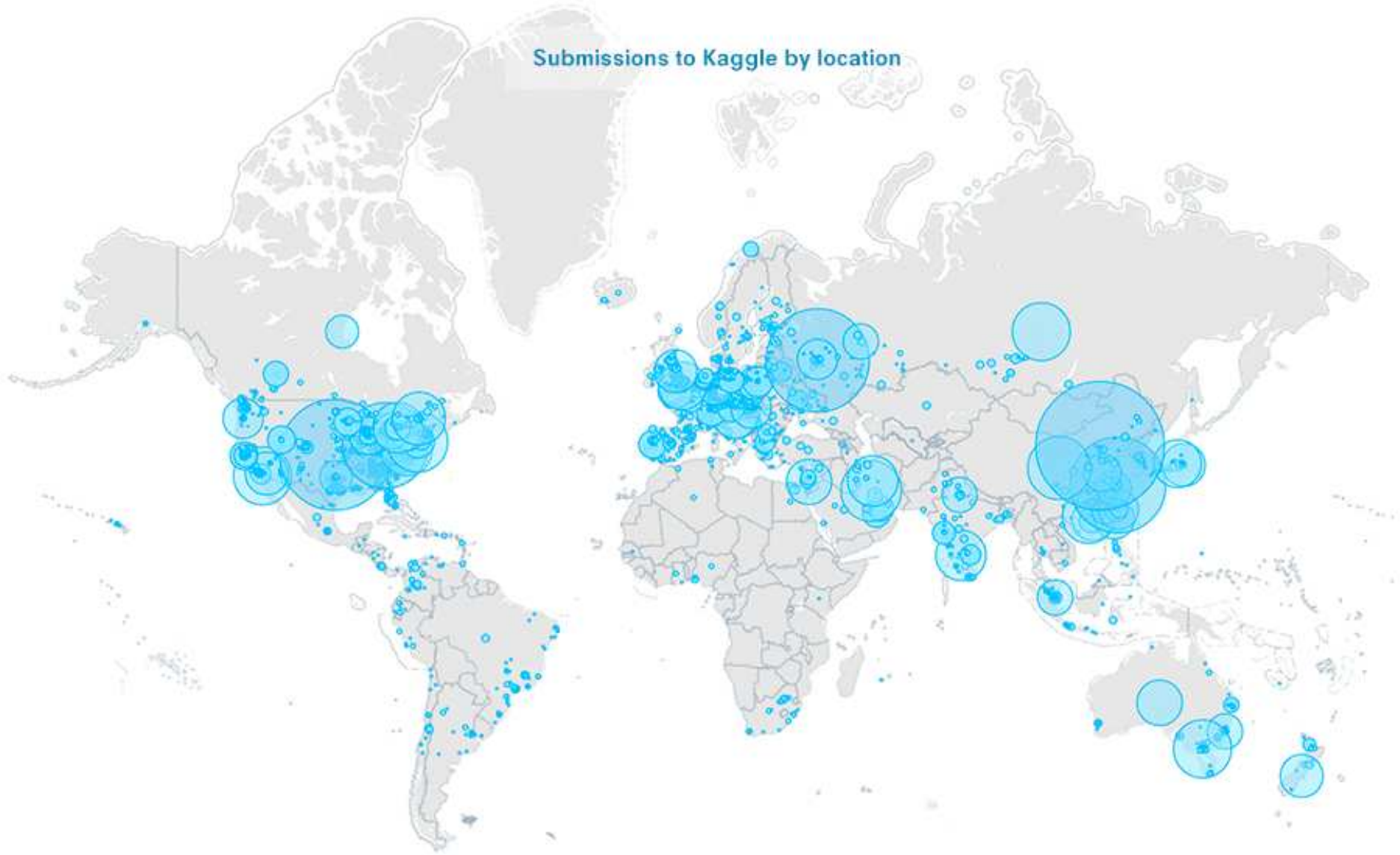
over 200000 entries

Total reward $ 4060897 (in public competitions)

Skillbase of kaggle analysts



Part II - Modern Data



Heritage Health Prize (kaggle competition)

Goal: Identify patients who will be admitted to a hospital within the next year, using historical claims data.

> *More than 71 million individuals in the United States are admitted to hospitals each year, according to the latest survey from the American Hospital Association.*
>
> *Studies have concluded that in 2006 well over $30 billion was spent on unnecessary hospital admissions.*
>
> *Can we identify earlier those most at risk and ensure they get the treatment they need?*

Data – nature and volume

$$252\,\text{MB} \begin{cases} 113000 & \text{patients} \\ 14 & \text{data fields} \\ 2668990 & \text{claims} \end{cases}$$

Some hypothetical claims (the number of claims varies from patient to patient):

```
78816124,321261,152610,23317,Y3,Laboratory,Independent Lab,18,,0- 1 month,INFEC4,0,PL,0
42097174,8889271,306649,33303,Y3,Diagnostic Imaging,Office,17,,0- 1 month,ARTHSPIN,0,RAD,0
91795962,1513798,877893,26051,Y2,Internal,Office,50,,11-12 months,ARTHSPIN,1-2,EM,0
33609291,3180459,818497,23610,Y1,Other,Office,59,,6- 7 months,GYNECA,1-2,EM,0
82096246,8307544,708111,20893,Y3,General Practice,Office,21,,2- 3 months,COPD,1-2,EM,0
63105048,4107701,164823,36452,Y1,Laboratory,Independent Lab,14,,6- 7 months,METAB3,0,PL,0
79790903,,,,Y3,,Inpatient Hospital,0,,,SKNAUT,0,PL,0
```

Part II - Modern Data

- Tentative predictions of days in hospital (next year), $\hat{y}_i$, are evaluated for 30% of the patients (test set).
- Kaggle discloses the accuracy of the prediction

$$\sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n}}$$

- Accuracy of the final prediction is based on the remaining 70% of the data (validation set).

Galaxy Zoo - The Galaxy Challenge

Goal: to analyze the JPG images of galaxies to find automated metrics that reproduce the probability distributions derived from human classifications.

Data – nature and volume

$$2\,\text{GB} \begin{cases} 141553 & \text{images} \\ 37 & \text{response variables} \\ 424^2 & \text{candidate explanatory variables} \end{cases}$$

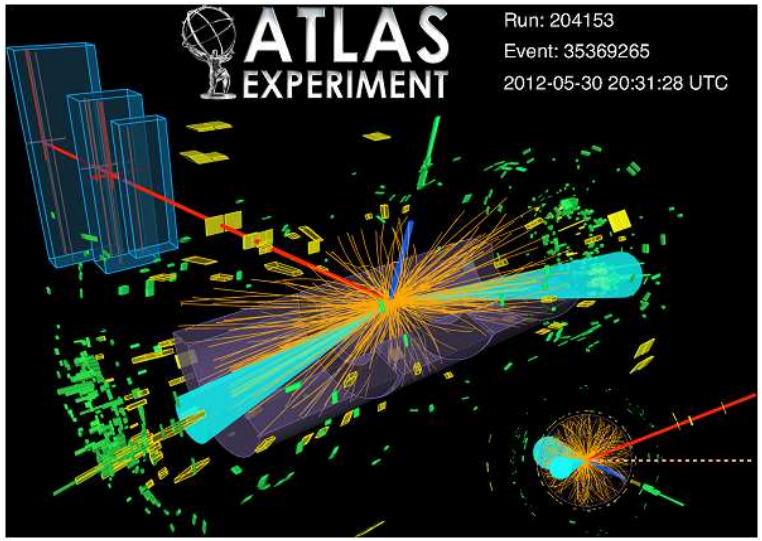
Some "data points":

Higgs Boson Machine Learning Challenge

Goal: The ATLAS experiment has recently observed a signal of the Higgs boson decaying into two tau particles, but this decay is a small signal buried in background noise. Aim is to classify events into "tau tau decay of a Higgs boson" versus "background."

Data – nature and volume

50 MB —- 750000 data points —- 32 variables

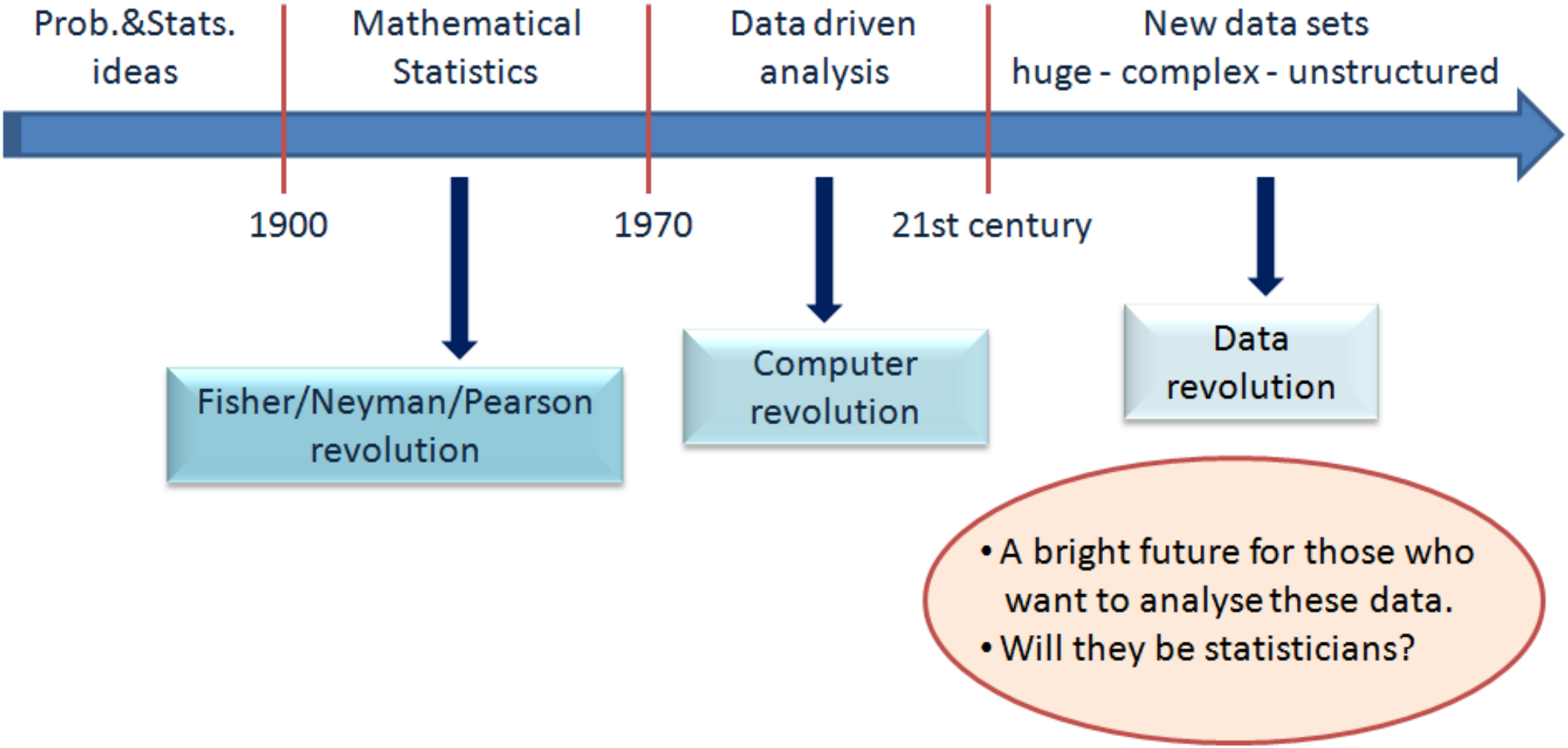


Final remarks

- Data is being produced incessantly (from many sources and in great volumes).
- Problems are like we have never seen before (in number and complexity).

If Statistics is seen as a way

- to use data to solve problems, the future seems very bright!
- to apply well known methods to familiar problems, the future is a time of lost opportunities for Statistics and statisticians!!


Prob.&Stats. ideas
Mathematical Statistics
Data driven analysis
New data sets huge - complex - unstructured
1900
1970
21st century
Fisher/Neyman/Pearson revolution
Computer revolution
Data revolution
• A bright future for those who want to analyse these data.
• Will they be statisticians?