

ΟΙΚΟΝΟΜΙΚΟ  
ΠΑΝΕΠΙΣΤΗΜΙΟ  
ΑΘΗΝΩΝ



ATHENS UNIVERSITY  
OF ECONOMICS  
AND BUSINESS

# Elements of Statistics and Probability

*LECTURE 1 – Introduction & Motivation*

Xanthi Pedeli

*Assistant Professor, [xpedeli@aueb.gr](mailto:xpedeli@aueb.gr)  
Department of Statistics, AUEB*

*Notes by Ioannis Ntzoufras, Professor  
Department of Statistics, AUEB*

# DATA SCIENCE



# Course details



## Lecture Schedule

1	Tuesday, 12 September 2023	Motivation & Introduction to Uncertainty and Probability Theory
2	Friday, 15 September 2023	Introduction to R
3	Monday, 18 September 2023	Introduction to R
4	Tuesday, 19 September 2023	Descriptive Statistics
5	Tuesday, 26 September 2023	Basics of Estimation, Hypothesis Testing and Regression using R
6	Friday, 29 September 2023	Exams

# Course details

## Exams

- Written Exams (80%)
- Assignment (20%)

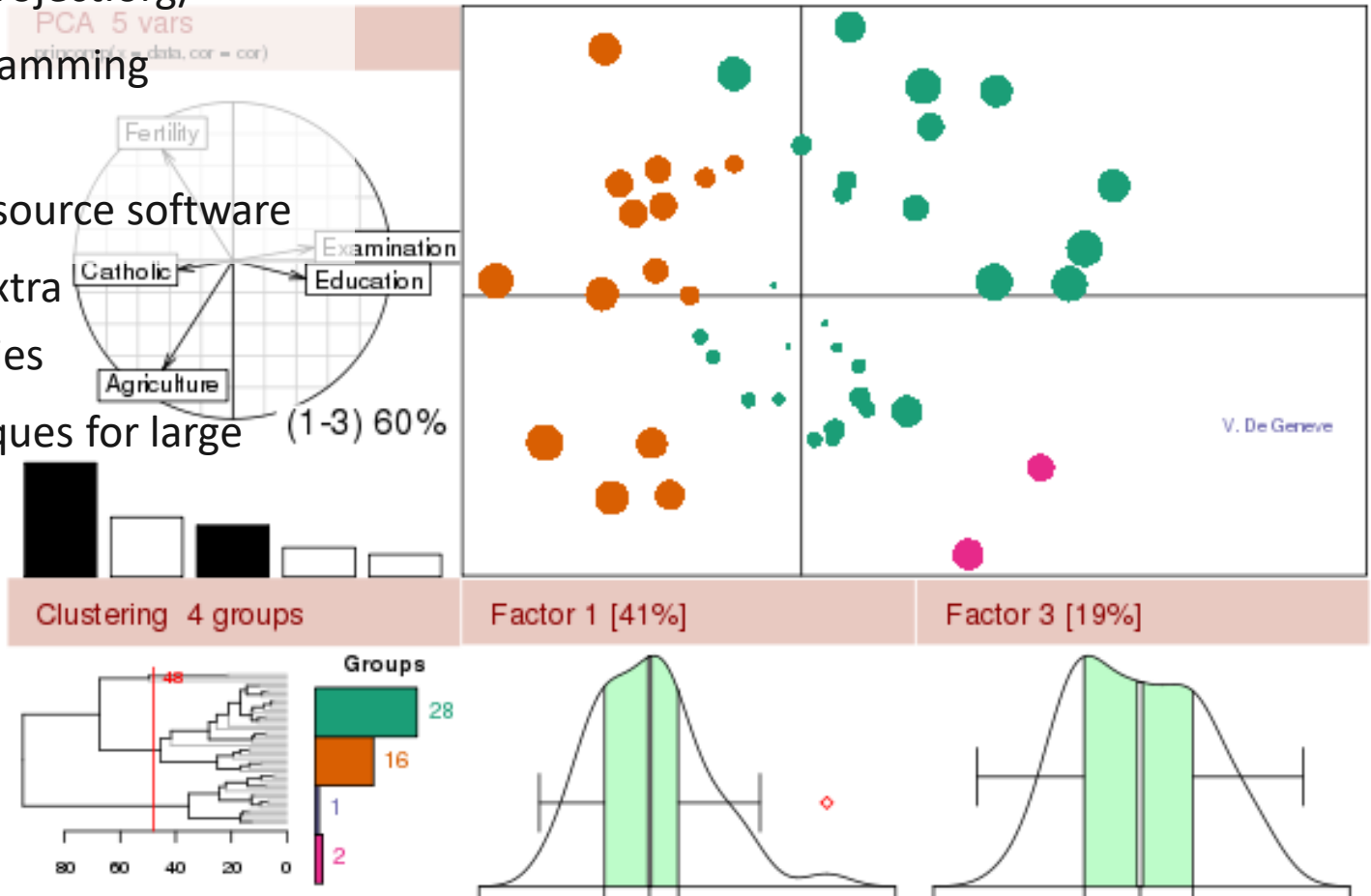
## e-class

<https://eclass.aueb.gr/courses/INF297/>

# R for modern data analytics

## The R Project for Statistical Computing

- <http://www.r-project.org/>
- Statistical programming language
- Free and open-source software
- Thousands of extra packages/libraries
- Modern techniques for large data



# Indicative bibliography

## Basic Statistics and probability:

- Diez, D., Barr, C., & Cetinkaya-Rundel, M. (2019). *OpenIntro statistics* (Fourth Edition). Free Open Book; available at <https://www.dbooks.org/openintro-statistics-1943450072/>

## Regression

- Faraway, J. (2002). *Practical regression and ANOVA using R*; available at <http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf>
- Fox J. & Weisberg H.S. (2018). *An R Companion to Applied Regression*. 3rd edition. SAGE Publications Inc.

## An all around classic (and free):

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer; available at <http://www-bcf.usc.edu/~gareth/ISL/>

# Indicative bibliography

## **R related books and material:**

- Crawley M.J. (2014). Statistics: An Introduction Using R. 2nd Edition. Wiley-Blackwell.
- Forte R.M. (2015). Mastering Predictive Analytics with R Paperback. Packt Publishing
- Miller J.D. & Forte R.M. (2017). Mastering Predictive Analytics with R - Second Edition: Machine learning techniques for advanced models. Packt Publishing

# Indicative bibliography

## R related books and material:

- Καρλής Δ. & Ντζούφρας Ι. (2015). *Εισαγωγή στον προγραμματισμό και στη στατιστική ανάλυση με R*. [ηλεκτρ. βιβλ.] Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. Διαθέσιμο στο: <http://hdl.handle.net/11419/2601>
- Φωκιανός Κ & Χαραλάμπους Χ. (2010). *Εισαγωγή στην R – Πρόχειρες σημειώσεις*. Πανεπιστημιακές σημειώσεις. 2η έκδοση. Τμήμα Μαθηματικών & Στατιστικής. Πανεπιστήμιο Κύπρου, url: <http://cran.r-project.org/doc/contrib/mainfokianoscharalambous.pdf>.
- Φουσκάκης Δ. (2013). *Ανάλυση Δεδομένων με Χρήση της R*. Εκδόσεις Τσότρας. Αθήνα. (Κωδικός Βιβλίου στον Εύδοξο: 33134029).
- Πετράκος Γ. (2016). *Εφαρμογές της Θεωρίας πιθανοτήτων με τη χρήση της R*. Εκδόσεις Τσότρας.

# 1.1. Introduction and motivation

- What is Statistics?
- What is Data Analytics?
- What is Data Science?



# 1.1. Introduction and motivation

## Definition of statistics



### DEFINITION 1

*“**Statistics** is a branch of mathematics dealing with the collection, organization, analysis, interpretation and presentation of data”*

*Source: Wikipedia via*

*Dodge, Y. (2006) The Oxford Dictionary of Statistical Terms, OUP.*

# 1.1. Introduction and motivation

## Definition of statistics



### DEFINITION 2

“The science that quantifies uncertainty”

*Source: The cartoon Guide to statistics*

# 1.1. Introduction and motivation

## Definition of statistics



### DEFINITION 3

“Statistics is a science, not a branch of mathematics, but uses mathematical models as essential tools”

*- John Tukey*

*Source: American Statistical Association*

# 1.1. Introduction and motivation

## *What is a statistician?*



- Statistician
- Data analyst
- Data Scientist
- Statistical programmer
- Statistical analyst
- Sampling expert/manager
- Data analysis expert
- Data modeler
- Computational statistician
- Statistical Analyst
- Prediction expert
- Biostatistician
- Sports modeler
- Psychometrician
- Behavioral analyst
- Market analyst
- Econometrician
- Statistical consultants
- Actuary
- Risk manager

# 1.1. Introduction and motivation

## *What is a statistician?*



- A statistician is a quantitative scientist that analyses data.
- A statistician implements **quantitative methods** to finally deduce **inference** about a problem **in another field of science** (e.g., medicine, sociology, epidemiology, psychology, economics etc.)

# 1.1. Introduction and motivation

## *What is a statistician?*



- A statistician is not talking about statistical facts in another science (he extracts them from data)
- What is the difference with data scientist?
- **Data science** is a broader field including **special skills from other fields** such as informatics and operational research.

# 1.1. Introduction and motivation

## What is Data Analytics?



**Analytics** is the use of:

- data,
- information technology,
- statistical analysis,
- quantitative methods, and
- mathematical or computer-based models

to help scientists or managers to gain **improved insight** about their research and make better, **fact-based decisions**.

Analytics often favors **data visualization** to communicate insight.

# 1.1. Introduction and motivation

## What is Data Science?



- **Data science** is an umbrella term that encompasses data analytics, data mining, machine learning, and several other related disciplines.
- While a data scientist is expected to forecast the future based on past patterns, data analysts extract meaningful insights from various data sources.
- A data scientist creates questions while a data analyst finds answers to the existing set of questions.

*Source: [www.simplilearn.com](http://www.simplilearn.com)*



# 1.1. Introduction and motivation

## What is Data Science?



### Data Scientists

- Data scientists **solve complex data problems** by employing deep expertise in some scientific discipline.
- It is generally expected that **data scientists be able to work with various elements** of mathematics, statistics and computer science, although expertise in these subjects is not required.
- However, a data scientist is most likely to be an **expert in only one or two of these disciplines** and proficient in another two or three.

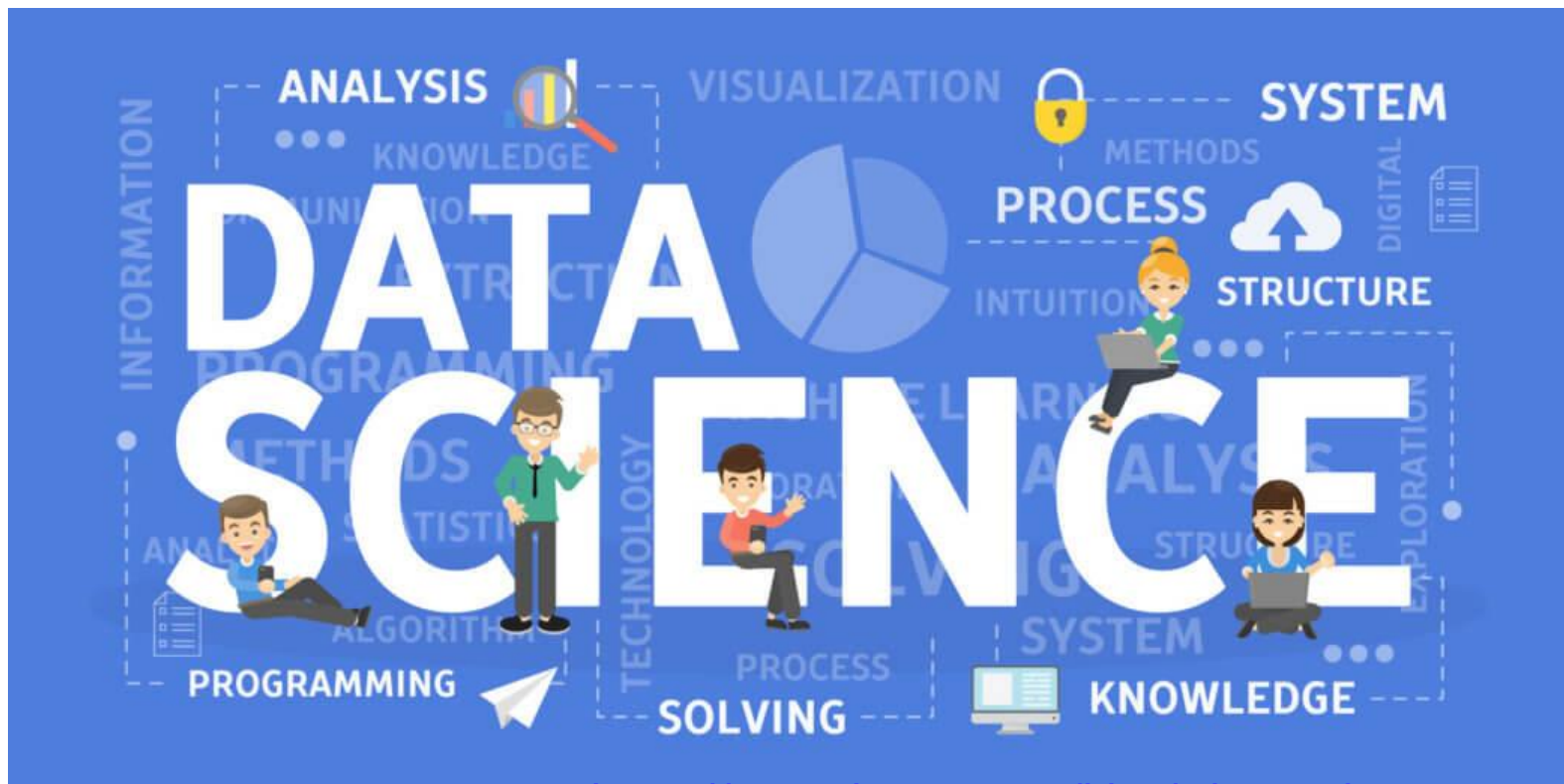


# 1.1. Introduction and motivation

## What is Data Science?

### Data Scientists

- Therefore, data science is **practiced as a team**, where the members of the team have a variety of expertise.



# 1.1. Introduction and motivation

## What is Data Science?



### Data Science

- The key word is science.
- The subject is **not restricted to only big data**, although the fact that data is scaling up makes big data an important aspect of data science.

# 1.1. Introduction and motivation



## Data Science VS Big Data VS Data Analytics

DATA IS GROWING FASTER THAN EVER BEFORE.



Each person-  
**1.7 megabytes**  
created



### WHAT ARE THEY?



**Data Science** is a field that comprises of everything that related to data cleansing, preparation, and analysis.



**Big Data** is something that can be used to analyze insights which can lead to better decision and strategic business moves.



**Data Analytics** Involves automating insights into a certain dataset as well as supposes the usage of queries and data aggregation procedures.

### WHERE ARE THEY USED?

Data Science algorithms are used in industries like:



Big Data is used in industries like:



Data Analytics is used in industries like:



## WHAT ARE THE SKILLS REQUIRED?



### DATA SCIENTIST

- In-depth knowledge in SAS and/or R
- Python coding
- Hadoop platform
- SQL database/coding
- Working with unstructured data

### BIG DATA SPECIALIST

- Analytical skills
- Creativity
- Mathematics and
- Statistical skills
- Computer science
- Business skills

### DATA ANALYST

- Programming skills
- Statistical skills
- Mathematics
- Machine learning skills
- Data wrangling skills
- Communication and Data Visualization skills
- Data Intuition

### DATA SCIENTIST

**\$113,436**  
per year.

### BIG DATA SPECIALIST

**\$62,066**  
per year.


### DATA ANALYST

**\$60,476**  
per year.


# 1.1. Introduction and motivation

## Data Science vs. Big Data vs. Data Analytics

### WHAT ARE THEY?



**Data Science** is a field that comprises of everything that related to data cleansing, preparation, and analysis.



**Big Data** is something that can be used to analyze insights which can lead to better decision and strategic business moves.

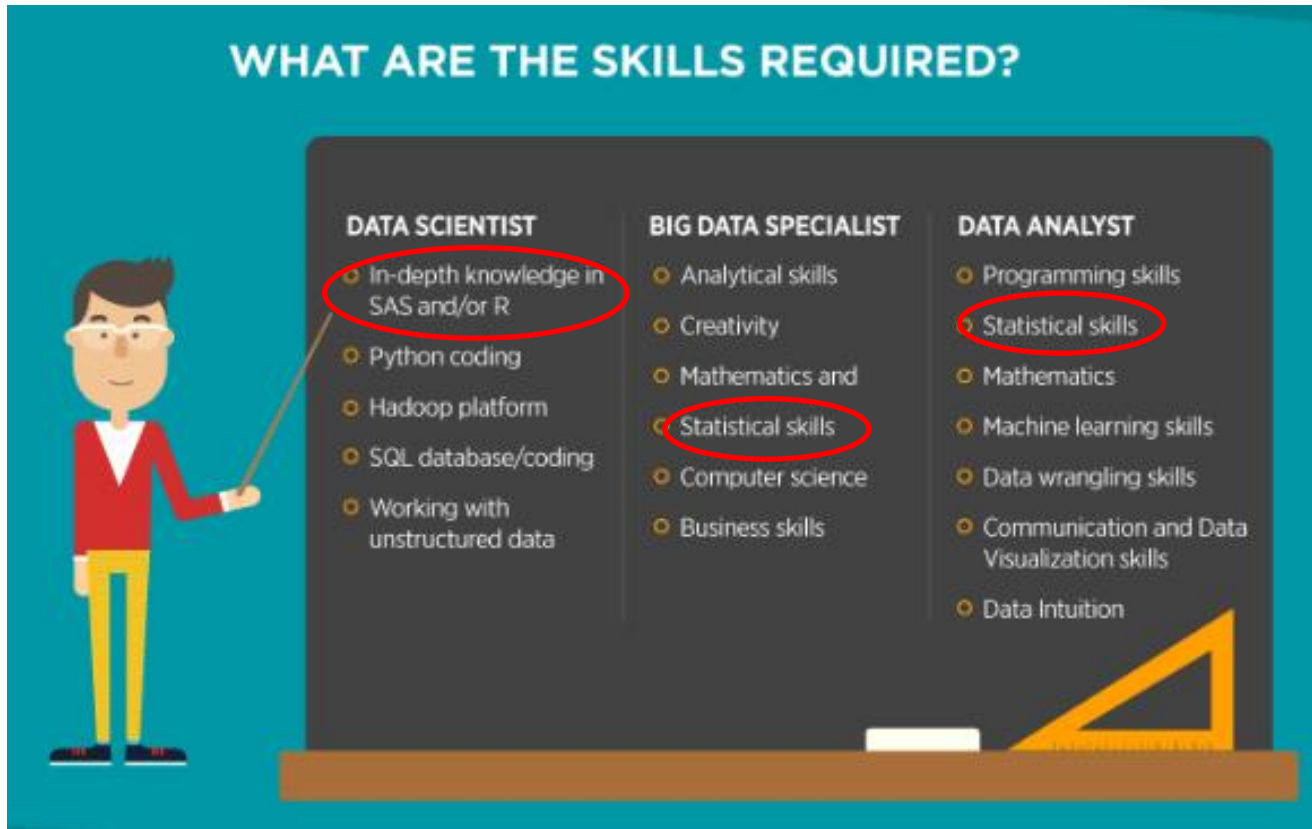


**Data Analytics** Involves automating insights into a certain dataset as well as supposes the usage of queries and data aggregation procedures.

# 1.1. Introduction and motivation

## Data Science vs. Big Data vs. Data Analytics

### WHAT ARE THE SKILLS REQUIRED?



DATA SCIENTIST	BIG DATA SPECIALIST	DATA ANALYST
<ul style="list-style-type: none"><li>In-depth knowledge in SAS and/or R</li><li>Python coding</li><li>Hadoop platform</li><li>SQL database/coding</li><li>Working with unstructured data</li></ul>	<ul style="list-style-type: none"><li>Analytical skills</li><li>Creativity</li><li>Mathematics and</li><li>Statistical skills</li><li>Computer science</li><li>Business skills</li></ul>	<ul style="list-style-type: none"><li>Programming skills</li><li>Statistical skills</li><li>Mathematics</li><li>Machine learning skills</li><li>Data wrangling skills</li><li>Communication and Data Visualization skills</li><li>Data Intuition</li></ul>

# 1.1. Introduction and motivation

## Math Meets Programming: A Quick History



- In many ways, **data science** is the result of a **merger between** two fields that have been around for decades: **statistics and computer science**.
- Statisticians, of course, have been crunching numbers for centuries. But the dawn of computer science in the mid 20th century provided statisticians with a new tool for analyzing data faster than had previously been possible.

<https://www.dataquest.io/blog/what-is-data-science/>

# 1.1. Introduction and motivation

## Math Meets Programming: A Quick History



- **1960s:** John W. Tukey were theorizing about how computers could revolutionize the field, but their **impact at the time was minimal** — they were simply too slow and too expensive.
- **1980s:** the rise of personal computers made **digital data collection** possible, and companies started collecting what they could.

<https://www.dataquest.io/blog/what-is-data-science/>



# 1.1. Introduction and motivation

## Math Meets Programming: A Quick History



- **1990s:** Some were successfully making use of that data to design marketing strategies. **Analyzing these new digital data sets** required both the statistics knowledge of a statistician and the programming skills of a computer scientist.
- **2000s:** Thanks in part to the advent of the internet, many companies had **access to mountains of data**. At the same time, computer processing power had advanced to the point that **complex analyses of huge data sets was possible**, and more advanced techniques like predictive analytics with machine learning were coming into reach.

# 1.1. Introduction and motivation

## Math Meets Programming: A Quick History

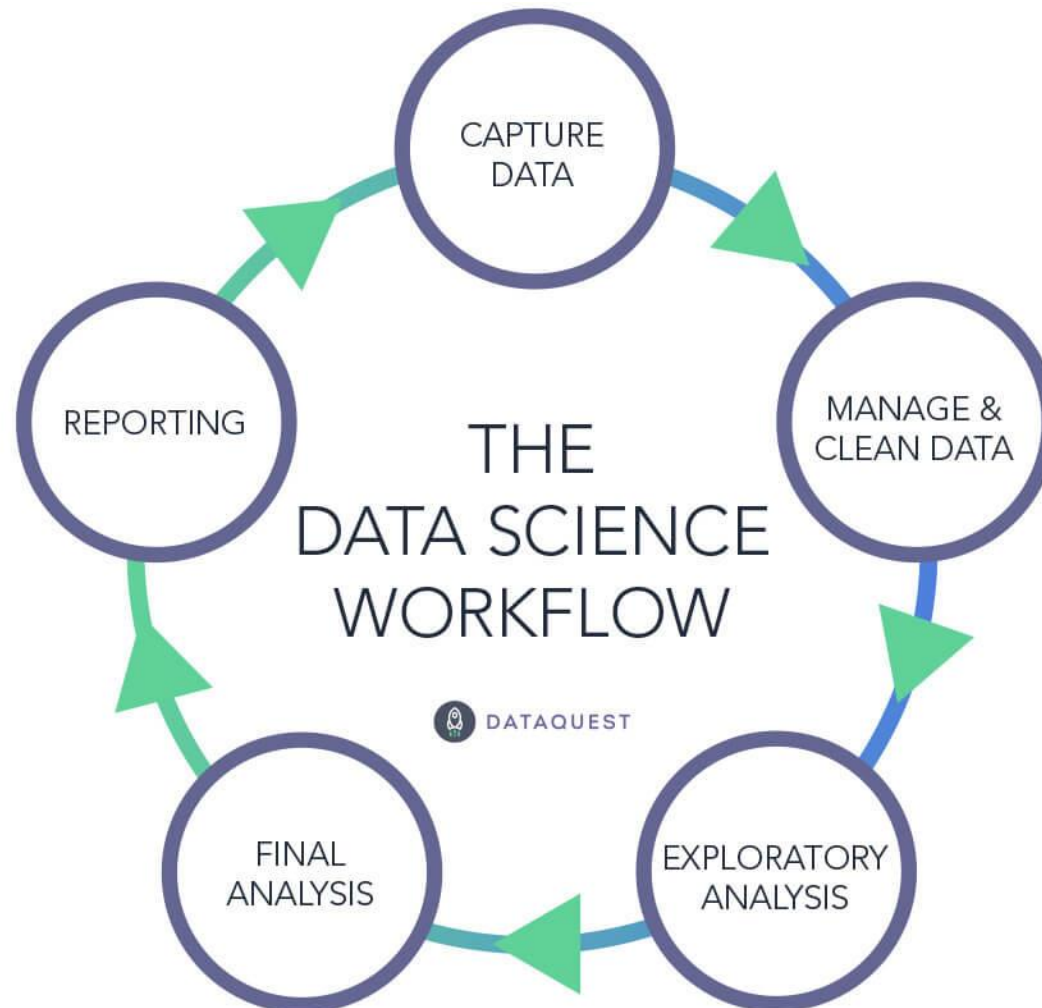


- Both business and academia began to recognize the value of having experts with the programming skills required to collect, manipulate, and analyze digital data *and* the statistics skills required to select the type of analysis needed to accurately answer questions and gain meaningful insights.
- “Data Science,” a term that had been around for decades by that point, became the mainstream phrase of choice to describe this confluence of skills.

<https://www.dataquest.io/blog/what-is-data-science/>

# 1.1. Introduction and motivation

## What Do Data Scientists Do?



# 1.1. Introduction and motivation

## What Do Data Scientists Do?



- 1. Capture data.** Pulling the data from a company database, scraping it from a website, accessing an API, etc.
- 2. Manage data.** Properly storing the data and almost always cleaning the data.
- 3. Exploratory Analysis.** Performing different analyses and visualizing the data in various ways to look for patterns, questions, and opportunities for deeper study.

# 1.1. Introduction and motivation

## What Do Data Scientists Do?



- 4. Final Analysis.** Digging deeper into the data to answer specific business questions and fine-tuning predictive models for the most accurate results.
- 5. Reporting.** Presenting the results of analysis to management, which might include writing a report, producing visualizations, and making recommendations based on the results of analysis. Reporting might also mean plugging the results of analysis into a data product or dashboard so that other team members or clients can easily access it.

# 1.1. Introduction and motivation

## *How much data do we use?*



### 1 How much data is generated every minute?

Source: Domo

 **41,666,667**

messages shared  
by WhatsApp users

 **1,388,889**

video / voice calls made  
by people worldwide

 **404,444**

hours of video streamed  
by Netflix users

 **347,222**

stories posted by Instagram users

 **150,000**

messages shared by Facebook users

 **147,000**

photos shared by Facebook users

Source: <https://financesonline.com/how-much-data-is-created-every-day/>

# 1.1. Introduction and motivation

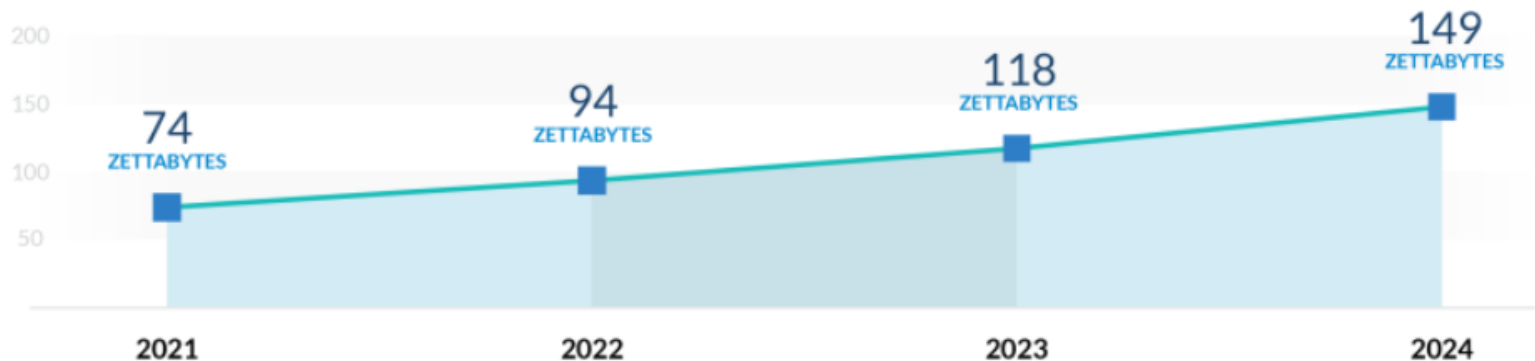
*How much data do we use?*



ΟΠΑ  
ΑΥΕΒ

## 2 Estimated Data Consumption from 2021 to 2024

Source: IDC / Statista



Source: <https://financesonline.com/how-much-data-is-created-every-day/>

# 1.1. Introduction and motivation

## *How much data do we use?*

### 3 Data Growth in 2021

Sources: TechJury, Internet Live Stats, Cisco, PurpleSec

 **2** TRILLION

searches on Google by the end of 2021

 **1.134** TRILLION MB

volume of data created every day

 **3,026,626**

emails sent every second, 67% of which are spam

 **278,108** PETABYTES

global IP data per month by the end of 2021

 **230,000**

new malware versions created every day

 **82%**

share of video in total global internet traffic at the end of 2021

Source: <https://financesonline.com/how-much-data-is-created-every-day/>



# 1.1. Introduction and motivation

## *Data scientists in the 21<sup>st</sup> century*



- For the year 2022, Glassdor named Data Scientist as the 3rd most desired job in the United States with 10,071 openings and a median base data scientist salary of \$120,000 with a job satisfaction rate of 4.1/5 (1st Enterprise Architect, 2nd Full Stack Engineer)

### Common Skill Sets

- |                    |                               |
|--------------------|-------------------------------|
| ✓ Machine Learning | ✓ Statistics                  |
| ✓ Python           | ✓ Natural Language Processing |
| ✓ Hadoop SPARK     | ✓ Algorithms                  |
| ✓ SQL              | ✓ Programming Languages       |

# 1.1. Introduction and motivation

## Visualizing the skills of a Data Scientist



### MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21st century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

#### MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants



#### PROGRAMMING & DATABASE

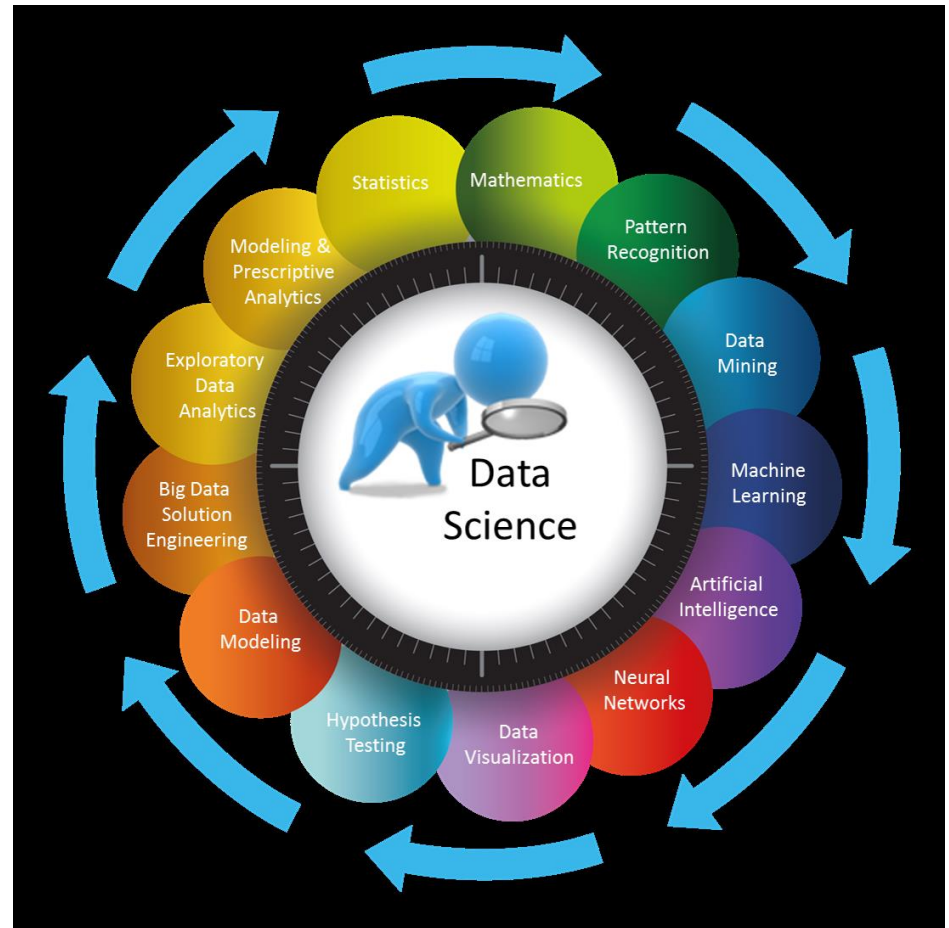
- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing packages, e.g., R
- ☆ Databases: SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

#### DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

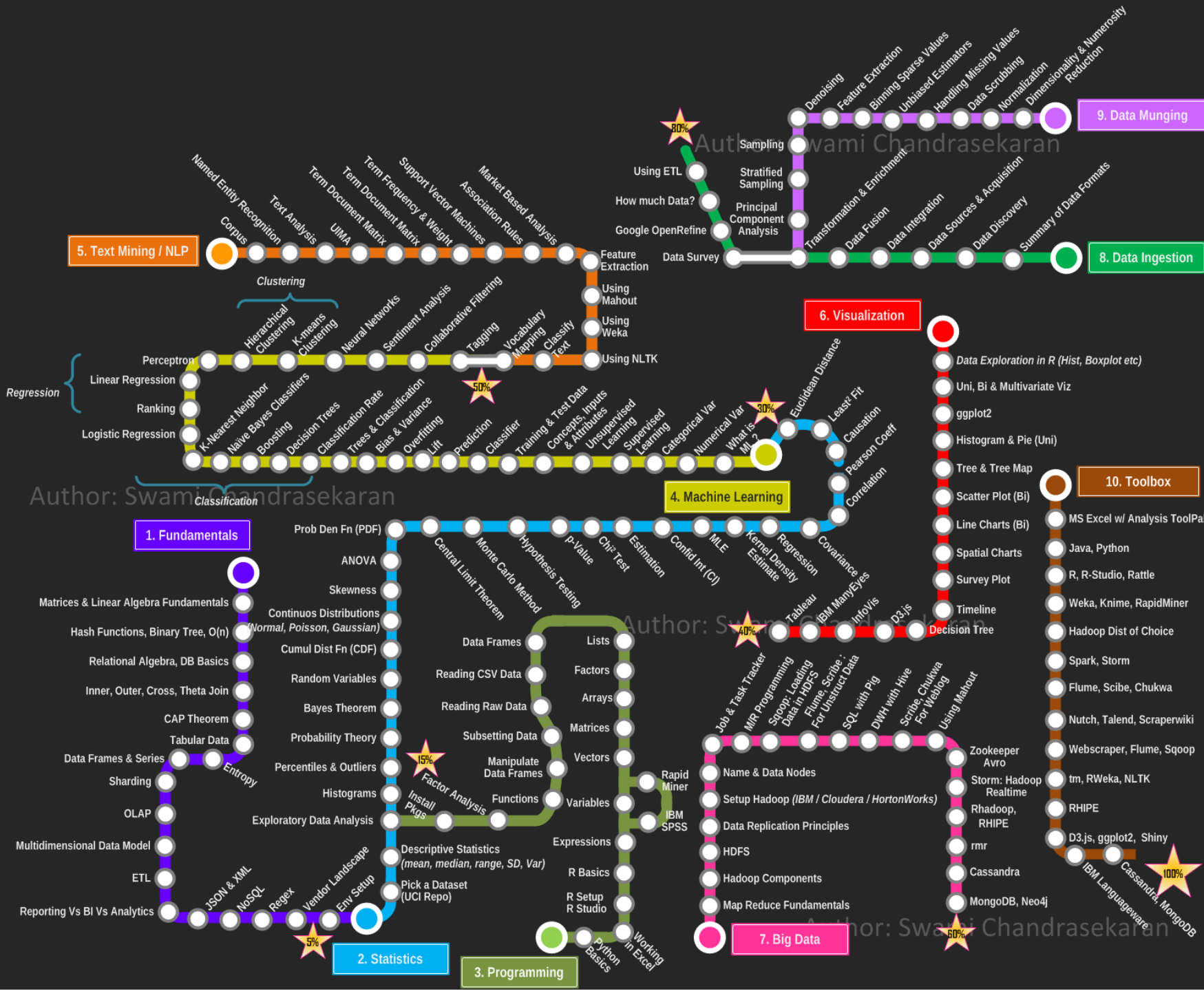
#### COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau





PA  
AUEB



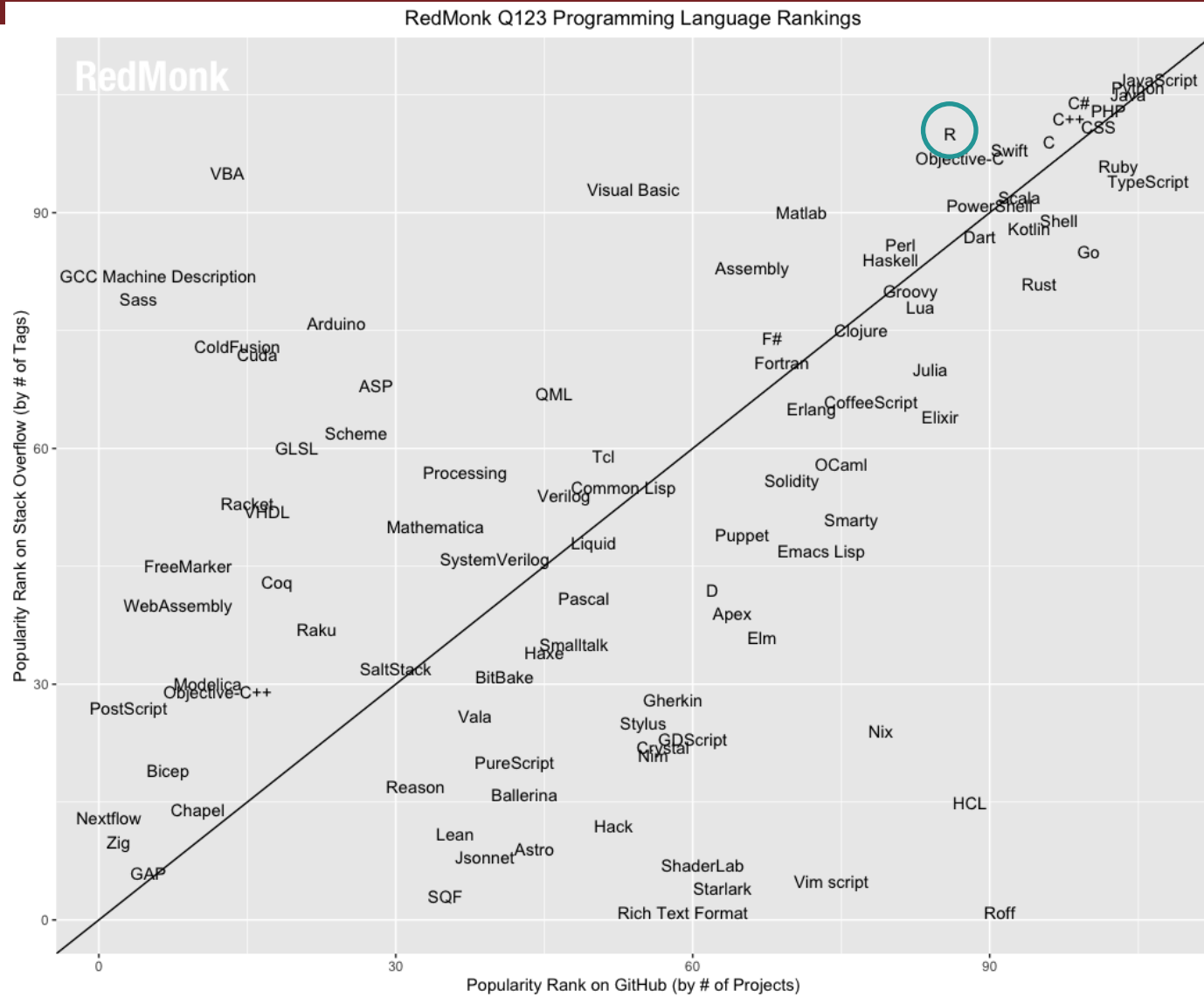
Author: Swami Chandrasekaran

Author: Swami Chandrasekaran

Author: Swami Chandrasekaran

# 1.1. Introduction and motivation

## *R in the 21<sup>st</sup> century*



# 1.1. Introduction and motivation

## Data analysis can be fun



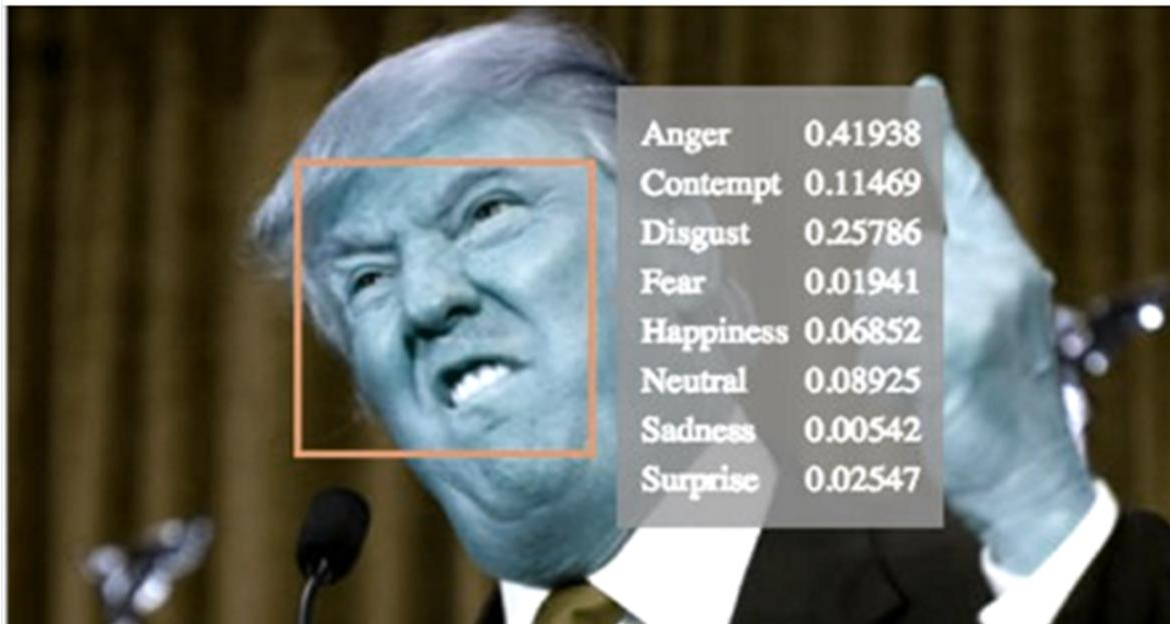
- Statistical data analysis can be implemented in any kind of problem
- This can make our job fun (sometimes) or boring (other times)
- A course on data analysis may include everything (all about statistics)
- Some intriguing examples follow

# 1.1. Introduction and motivation

## Data analysis can be fun

 **R bloggers**  
Yesterday at 6:42am

Election 2016: Tracking Emotions with R and Python



Election 2016: Tracking Emotions with R and Python

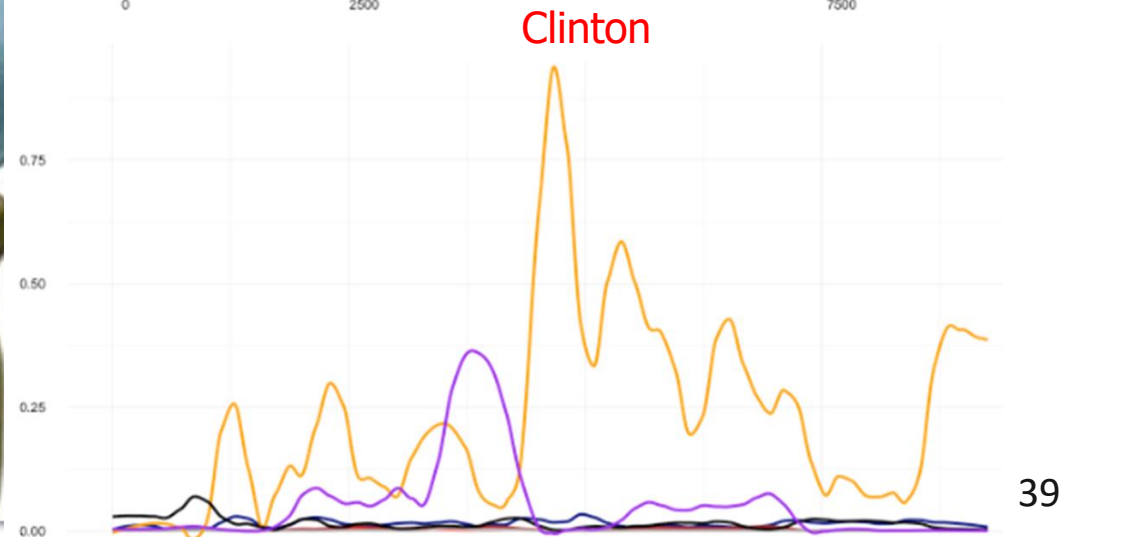
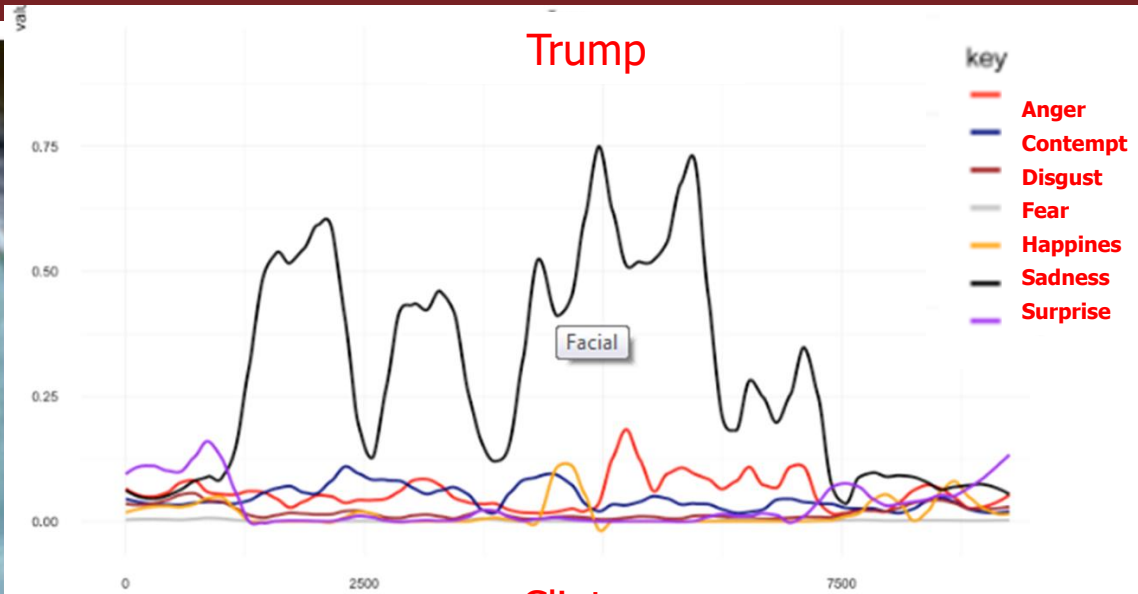
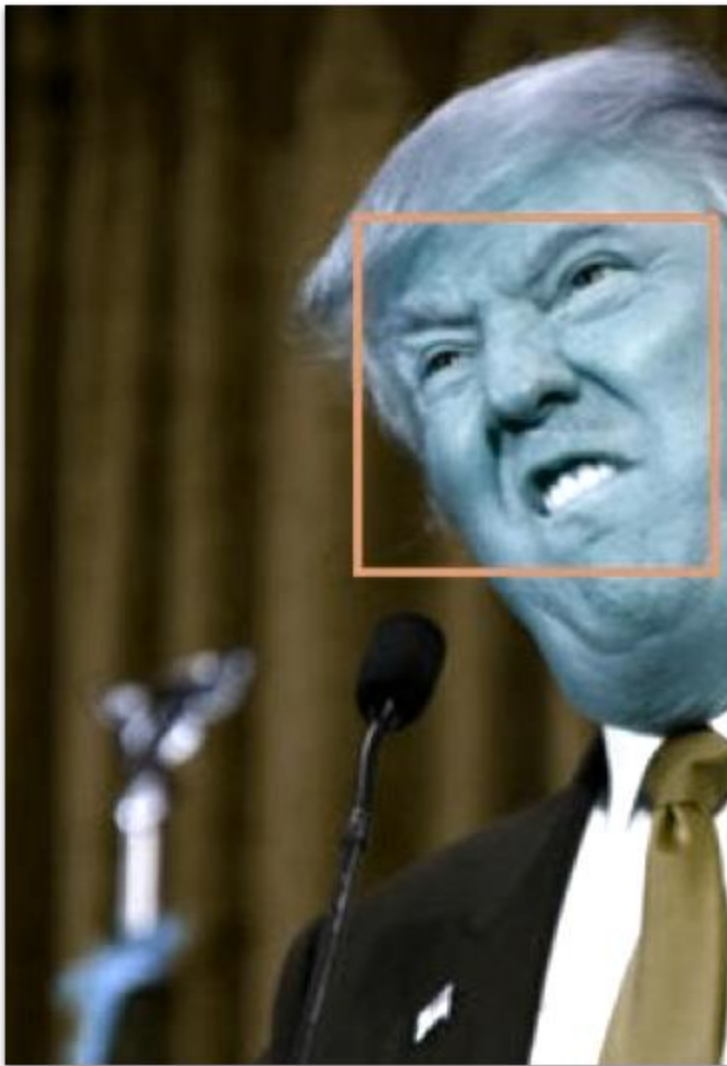
Temperament has been a key issue in the 2016 presidential election between Hillary Clinton and Donald Trump, and an issue highlighted in the series of three debates that concluded this week. Quanti...

R-BLOGGERS



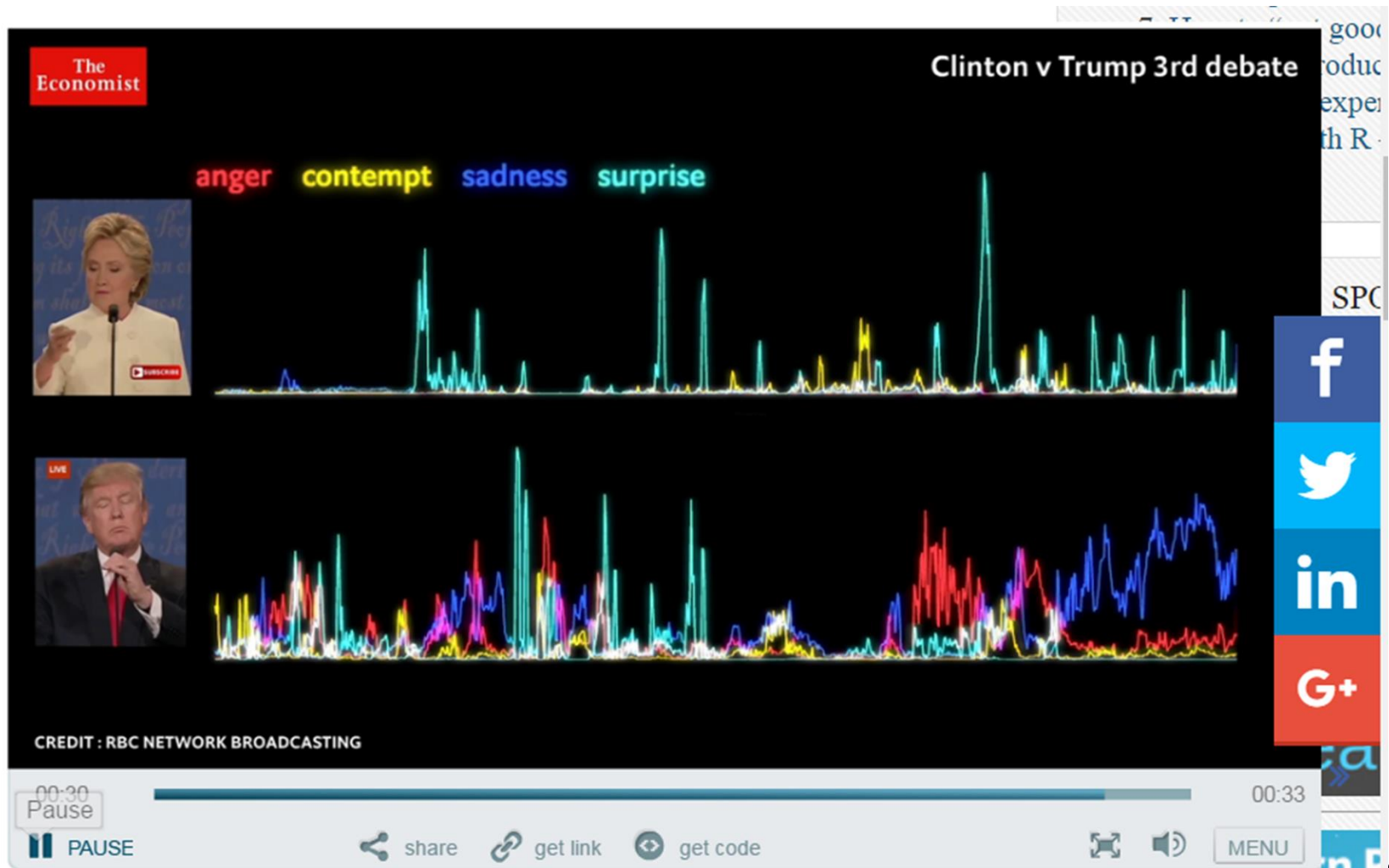
# 1.1. Introduction and motivation

## Data analysis can be fun



# 1.1. Introduction and motivation

## Data analysis can be fun



<https://www.r-bloggers.com/election-2016-tracking-emotions-with-r-and-python/>



# 1.1. Introduction and motivation

## Data analysis can be fun



ΟΠΑ  
ΑΔΕΒ

### Bayesian hierarchical model for the prediction of football results

Gianluca Baio<sup>1,2\*</sup>

Marta A. Blangiardo<sup>3</sup>

### Publishable Stuff

Rasmus Bååth's Research Blog

About | CV | Research | Blog | Archives

Search

JUL 21ST, 2013

## Modeling Match Results in La Liga Using a Hierarchical Bayesian Poisson Model: Part one.

*This is a slightly modified version of my submission to the [UseR 2013 Data Analysis Contest](#) which I had the fortune of winning :) The purpose of the contest was to do something interesting with a dataset consisting of the match results from the last five seasons of La Liga, the premium Spanish football (aka soccer) league. In total there were 1900 rows in the dataset each with information regarding which was the home and away team, what these teams scored and what season it was. I decided to develop a Bayesian model of the distribution of the end scores. Here we go...*

## why spain will win...

...maybe? **Dr Ian Hale**, senior lecturer in statistics at the University of Salford, discusses how mathematical models of football matches are used in the gambling industry – and sportingly puts his neck on the line by supplying his own predictions for the World Cup 2010

PREDICTING football results is a rapidly growing area of academic interest. Economists use models to assess the efficiency of betting markets, operational researchers use models to experiment with the various effects of tournament design, and statisticians showcase their proficiency with

advanced statistical techniques by modelling the intricacies of football data. It is not, of course, just academics who are mining the archives of football scores. Bookmakers live and breathe football prediction models – as do the more committed flutters. Mistakes cost money

and jobs, whilst finding a small advantage can carry great rewards.

**BETTING MARKETS**  
In academia, the most common application of football forecasting models is to test for betting market efficiency. The Efficient Markets Hypothesis

(EMH) is a cornerstone of financial theory and, in its simplest form, states that an investor should not be able to consistently obtain returns above the average. Finding a forecasting model of football that can generate better-than-average – or even positive – returns usually results in a publication for the academic as an example of a violation of the EMH, but the proprietary nature of the models means that the published ones rarely (if ever) represent the very best models, and even less often generate positive returns consistently.

S4 Science 4 All  
Quality Popular Science

Home Write About



By Lê Nguyễn Hoàng

PhD Student in Applied Maths at Polytechnique of Montreal.  
Engineer of the Ecole Polytechnique, France. (X2007)

### A Model of Football Games

PRINT

Abstract | Outline | Posted: April 4th, 2013 | Last Modified: February 28th, 2014 | Tags: Mathematics, Modeling, Optimization, Probability, Soccer, Statistics | Views: 3272 | No Comments »

Support the author by sharing: [Share](#) [Tweet](#) [15](#) [+1](#) [45](#) [Share](#) [61](#) [repost](#)

While the 2014 World Cup in Brazil is approaching, controversy has been raised about the FIFA ranking which plays a central role in the drawing of the first stage groups. In particular, it's puzzling to see that [this ranking \(in October 2014\)](#) features Columbia and Belgium at 4th and 5th position, while world-cup winners Brazil and France are 11th and 21st... Does this ranking even make sense?

« [Panel discussion on data science and "small data"](#) | [Main](#) | [The ACM 2013 Mining Big Data Cam "Un-Conference"](#) »

October 16, 2013

## Fantasy Football Modeling with R

Boris Chen, a data scientist for the *New York Times*, has been running since August a [weekly blog with statistical analysis of NFL players](#), as fodder for Fantasy Football players around the country. Here's how he describes what he does:

# 1.1. Introduction and motivation

## Data analysis can be fun



### CHANCE

A Magazine for People Interested in the Analysis of Data

Articles

Columns

Editor's Letter

Letters to the Editor

S

Speaking Stats to Justice: Expert Testimony in a Guatemalan Human Rights Trial Based on Statistical Sampling

Bond. James Bond. A Statistical Look at Cinema's Most Famous Spy

Road Crashes and the Next U.S. Presidential Election

[• Articles](#)

*By Donald A. Redelmeier and Robert J. Tibshirani*

Does Banning Hand-Held Cell Phone Use While Driving Reduce Collisions?

Statistical Modeling of Sleep

[• Articles](#)

*By James E. Slaven, Michael E. Andrew, Anna Mnatsakanova, John M. Violanti, Cecil M. Burchfiel, and Bryan J. Vila*

# 1.1. Introduction and motivation

## Data analysis can be fun



**significance**  
statistics making sense

WEB EXCLUSIVES MAG

**The World Cup group stage: predictions through the betting markets**  
Dominic Cortis  
**Published:** Jun 02, 2014  
SPORTS

**House prices: statistics, politics behaviour**  
Oz Flanagan  
**Published:** May 30, 2014  
FINANCE & THE ECONOMY

**How well do FIFA's ratings predict World Cup success?**  
Ray Stefani  
**Published:** May 28, 2014  
SPORTS

**How to measure democracy**  
Andrew McCulloch  
**Published:** May 21, 2014  
SOCIAL SCIENCES

**significance**  
statistics making sense

WEB EXCLUSIVES MAGAZINE

HOME WEB EXCLUSIVE ARTICLE

**Is the UK shunned at Eurovision?**  
Gianluca Baio & Marta Blangiardo

It's that time of the year again. One of the biggest events in Europe's (and the world's) cultural calendar, the Eurovision song contest is legendary.



1.1. I

# POPULAR SCIENCE

Login/Register | Newsletter | Subscribe

GADGETS | CARS | S

GALLERIES /// VIDEOS /// BLOGS ///



## Cold Hard Facts

# Eurovision statistics: full predictions

This is part three of a series of posts describing a predictive model. The full set of posts can be found [here](#).

## Keeping score

Last time round, I compared my model's original predictions to the actual results. I managed to predict eight of the ten qualifiers correctly, which is about what random guessing would give, and seemed to compare well with human predictions (which have the benefit of knowing what

the critically acclaimed Solomon Northup masterpiece that major prize among



Social Oscars screenshot The Social Oscars is one of several statistical models now at work predicting this year's Oscar winners. Screenshot from the Social Oscars by Brandwatch and The Credits

**ESCInsight** Home of "The Unofficial" ESC

Home About Us Contact Us Tour the Site

Articles Headlines Podcast

## Our Statistical Analysis Of The Eurovision 2014 Semi-Final Draw

Posted by Ewan Spence on Jan 20th, 2014 in Articles, Editorial, Sabremetrics | 6 comments



Just like that, we have the draw for the semi-finals of this year's Eurovision Song Contest.

Delegations are away to book hotel rooms, look at the cost of flights, and lock in all their travel plans for two weeks in Copenhagen. Latvia will be tightening the purse strings as they'll be asked to arrive for the first day of rehearsals, while the 'we're not quite sure we can afford it' Slovenia can stay at home for three extra days until they are needed during the fourth day of rehearsals.

But we're not going to look at the logistics just yet. No, it's time to decide who is going to qualify from the semi-finals, even though we have only heard three songs (and the chances are two will have a significant remix, and the 'Cheesecake' will go off by May to be replaced by 'Danish Pastry').

Rather than go for an emotional response or a gut feel, we've taken the information from today's draw, fed it into a very complicated spreadsheet of historical data and trends, and come up with a statistical prediction for the ten qualifiers out of each semi-final.

