

Big Data Analytics 2020/2021

LUCA PAPPALARDO

FOSCA GIANNOTTI



Consiglio Nazionale
delle Ricerche

<http://didawiki.di.unipi.it/doku.php/bigdataanalytics/bda/start>

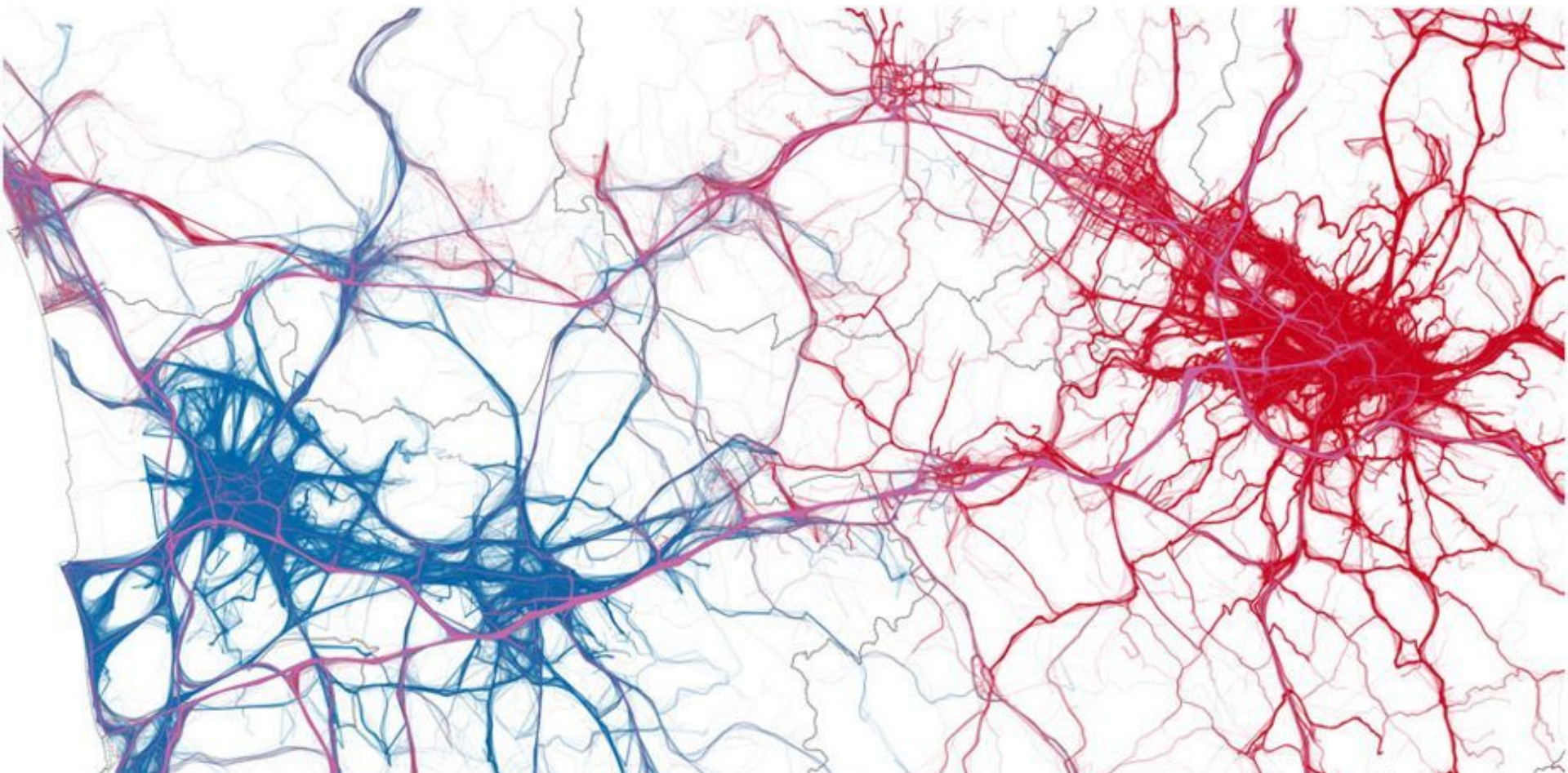
DIPARTIMENTO DI INFORMATICA - Università di Pisa







Private vehicles traveling in Tuscany (on-board GPS devices)

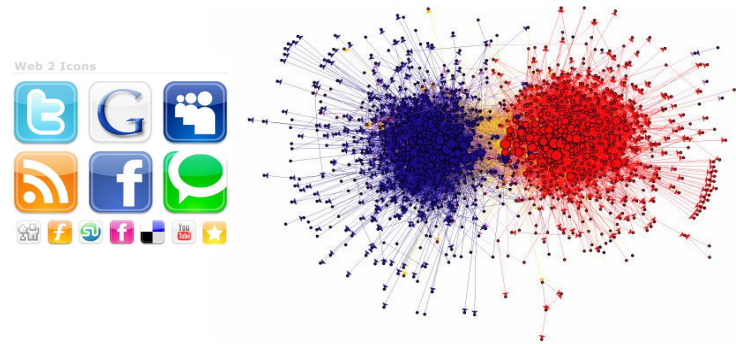


Digital Footprints of Human Activities

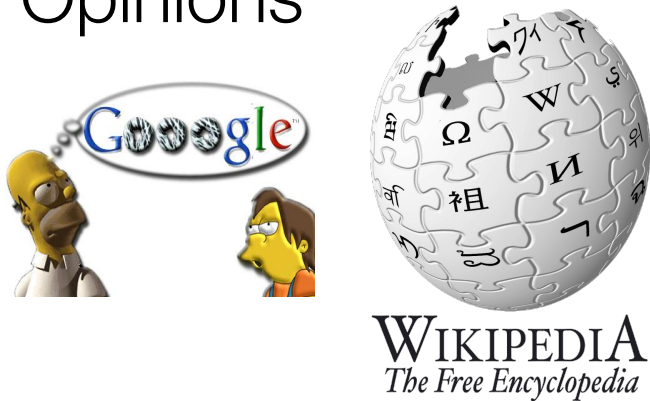
Shopping patterns



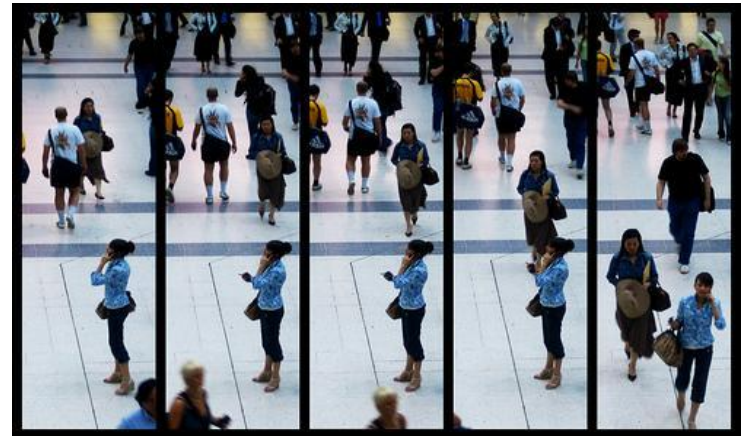
Social Ties



Opinions



Movements



The Vs of Big Data

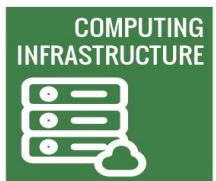
Volume: the incredible amounts of data generated each second

Velocity: speed at which vast amounts of data are being generated, collected and analyzed.

Variety: the different types of data we can now use

Veracity: quality or trustworthiness of the data

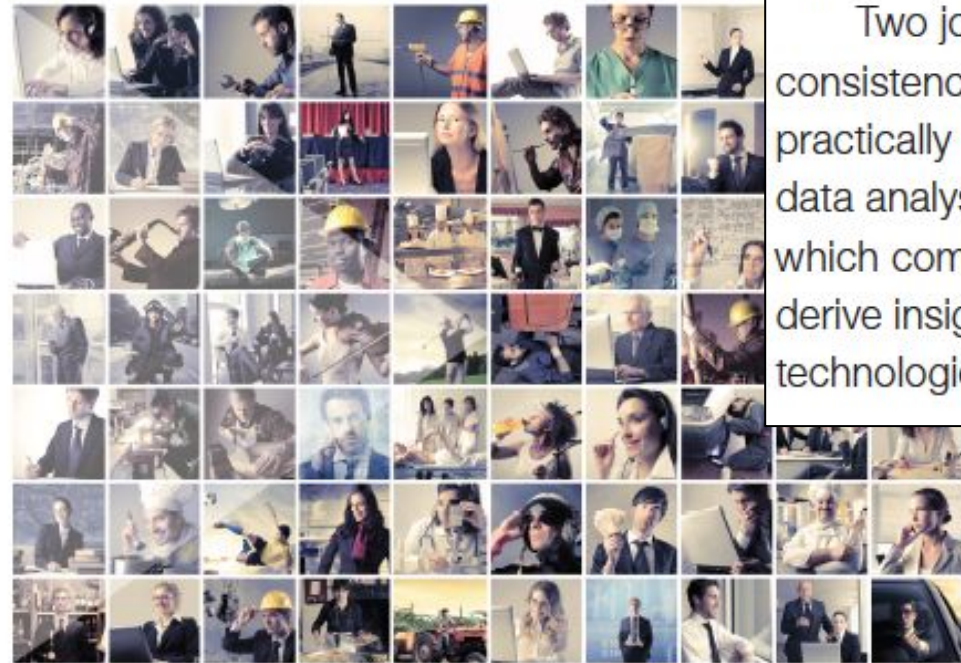
Value: the worth of the data being extracted



The Future of Jobs

Employment, Skills and
Workforce Strategy for the
Fourth Industrial Revolution

January 2016



New and Emerging Roles

Our research also explicitly asked respondents about new and emerging job categories and functions that they expect to become critically important to their industry by the year 2020, and where within their global operations they would expect to locate such roles.

Two job types stand out due to the frequency and consistency with which they were mentioned across practically all industries and geographies. The first are data analysts, as already frequently mentioned above, which companies expect will help them make sense and derive insights from the torrent of data generated by the technological disruptions referenced above. The second

http://www3.weforum.org/docs/WEF_Future_of_Jobs.pdf

The Future of Jobs

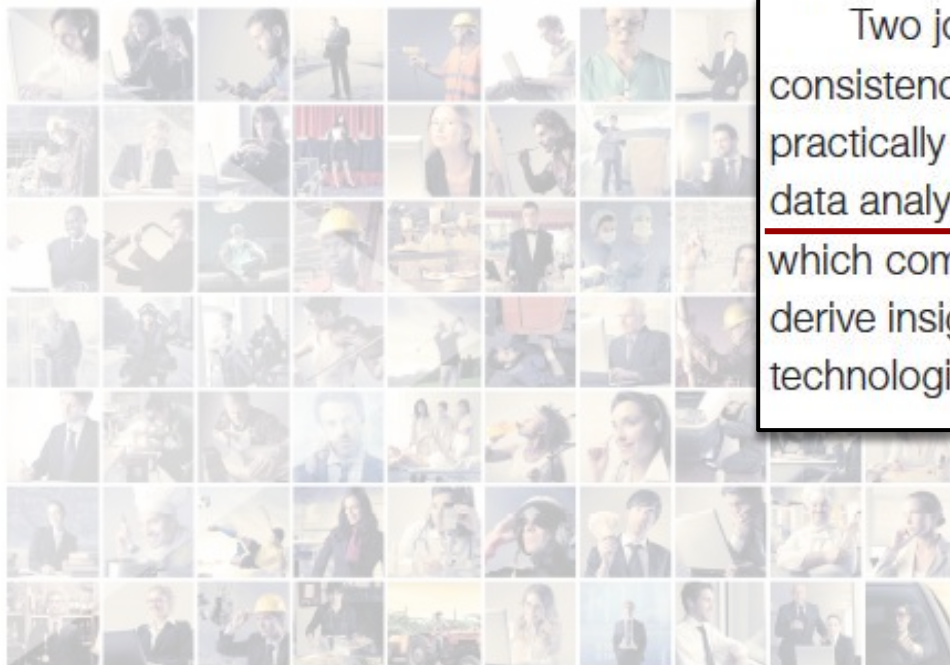
Employment, Skills and Workforce Strategy for the Fourth Industrial Revolution

January 2016

New and Emerging Roles

Our research also explicitly asked respondents about new and emerging job categories and functions that they expect to become critically important to their industry by the year 2020, and where within their global operations they would expect to locate such roles.

Two job types stand out due to the frequency and consistency with which they were mentioned across practically all industries and geographies. The first are data analysts, as already frequently mentioned above, which companies expect will help them make sense and derive insights from the torrent of data generated by the technological disruptions referenced above. The second



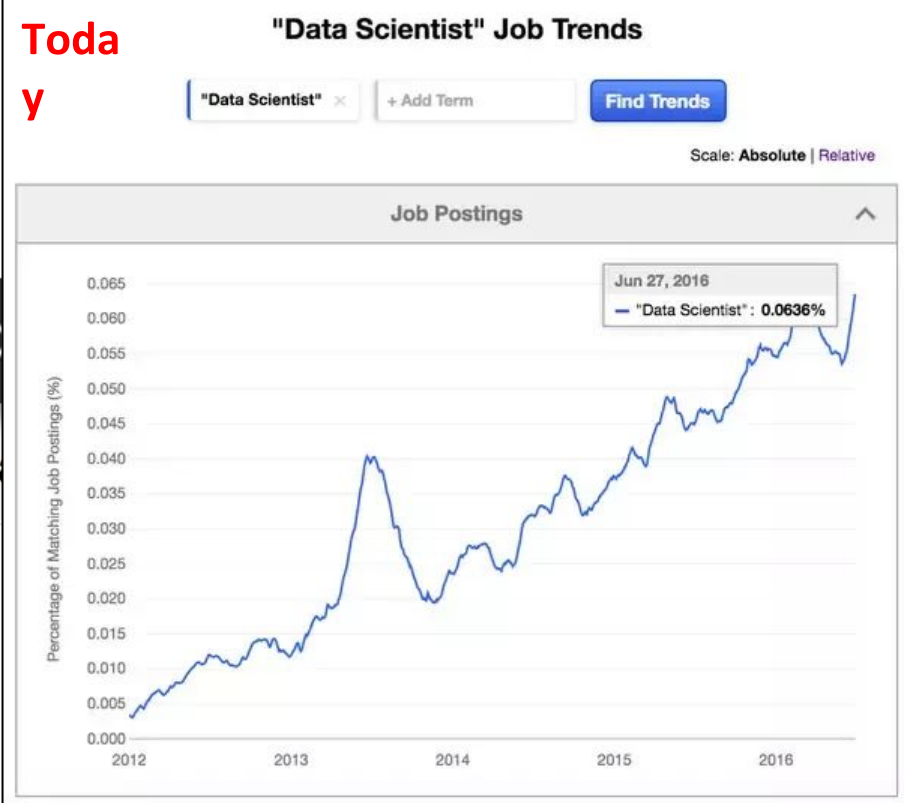


Guest

Subscribe today

THE MAGAZINE

October 2012



ARTICLE PREVIEW To read the full article, **sign-in** or **register**. HBR subscribers, click **here** to **for FREE access »**

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

Goals of this Course

- It is an introduction to the emergent field of Big Data Analytics and Social Mining
- It aims at analyzing big data from multiple sources to the purpose of discovering the patterns of human behavior

Module 1: Technologies

1. Python for Data Science
2. The Jupyter Notebook: developing open-source and reproducible data science
3. MongoDB: fast querying and aggregation in NoSQL databases
4. Scikit-learn: machine learning with Python
5. GeoPandas and scikit-mobility: analyze geo-spatial data with Python
6. Keras: deep learning in Python

Module 2: Case Studies

Sports Analytics: What is possible to observe with IoT data? Sensor data in sports. Predicting injuries and evaluating performance.

Model Construction and Validation

Mobility analysis: What is possible to observe with mobile phone and GPS data? Analysis of human mobility at individual and collective levels. Mobility Data mining methods in a nutshell.

Data preparation, Model construction and Validation.

Well-being: Can we measure the well-being and happiness of people through Big Data?

Quantification. Data preparation, Model Construction and Validation

Module 3: laboratory for interactive project development

- Create teams of “data analysts”
 - Choose a dataset among those proposed
1. *October*: 1st Mid Term (Data Understanding and Project Formulation)
 2. *November*: 2nd Mid Term (Model implementation and evaluation)
 3. *December*: 3rd Mid Term (Model interpretation and explanation)
 4. *January*: Exam (Final Project results)

1st Mid Term

Data Understanding and Project Formulation

- 20 minutes presentation (+10 minutes questions)
- Demonstrate that you properly explore the chosen dataset
- Propose a predictive task
- Upload the report, code and presentation 2 days before

2nd Mid Term

Model Implementation and Evaluation

- 20 minutes presentation (+10 minutes questions)
- Show the predictive models you built (e.g., decision tree, SVM, etc.)
- Show how you evaluate their goodness (baseline models, accuracy, precision, etc.)
- Upload, report, code and presentation 2 days before

3rd Mid Term

Model Interpretation and Explanation

- 20 minutes presentation (+10 minutes questions)
- Show how to interpret your model to gain knowledge (feature importance, decision rules, local explanations)
- Upload, report, code and presentation 2 days before

Exam

Final project results

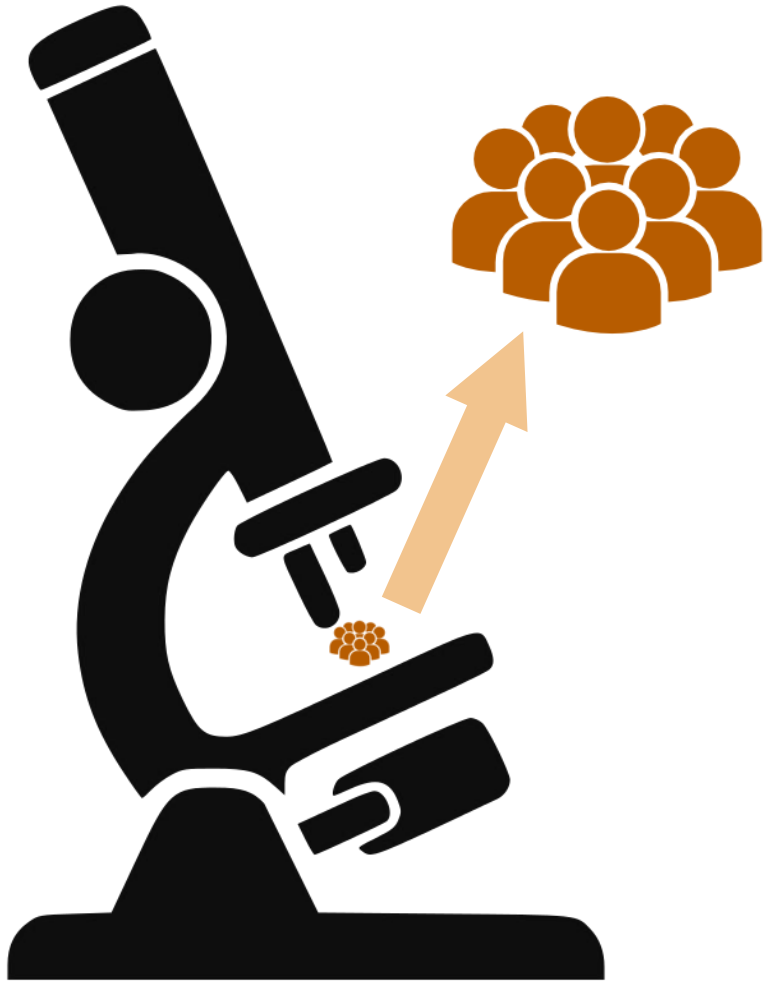
- 20 minutes presentation (+10 minutes questions)
- We provide you a few instances, you must run live you code on these instances to obtain predictions and interpretations/explanations
- Upload the final report, code and presentation 2 days before the exam

Evaluation criteria

- we evaluate the overall quality of the project at the exam
- each student will read a paper related to what they are developing and present it in a presentation (to be done before the end of the course)
- on the basis of evaluation of the project and the evaluation of the paper presentation we assign the final grade to each student

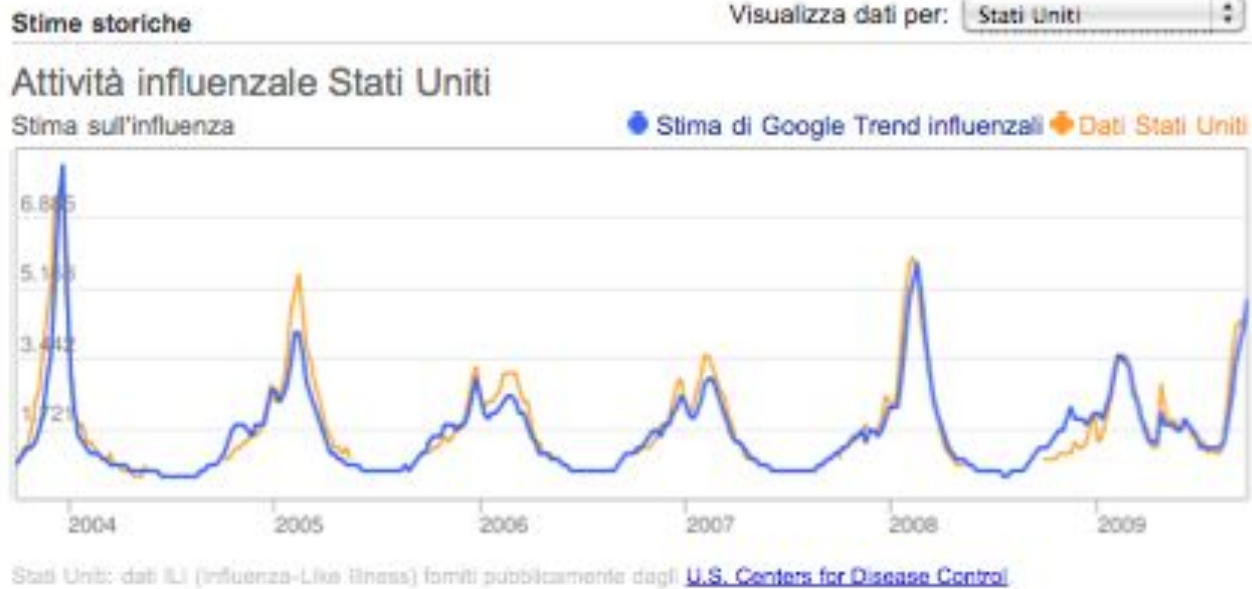
	Feature Construction	Model Construction	Validation	Interpretation and story telling
Required (all) Grade range: 18-24	Study of existing features, correlation analysis, selection of the interesting ones, transformation, construction of useful features	Select a modeling task appropriate to the analytical objective that the student has proposed	Provide a discussion on the base of objective measures of the methods about performance: SSE, Accuracy, ROC, Lift, Support, Confidence,	Discuss the achievements w.r.t the potential usage of the model and discuss the potential improvements
Advanced (all the required plus at least one) Grade range: 24-28	Integration with external (new) sources	Combine several models or adopt more advanced to archive better explanation or better performances (for example combine clustering with pattern mining or do ensemble methods or sophisticated classification as multilabel classification or cascade)	also consider possible domain dependent cost function	also discuss the potential improvements w.r.t a comparison w.r.t a quantitative baseline
Challenging (Advanced plus at least one) Grade range: 28-30L	Invention of new features	Compare your results with those obtained with other models and algorithms (e.g. compare a decision tree model with SVM-CNN trained on the same dataset)	Also discuss w.r.t a ground truth obtained by a null model or a human generated labelling , or other “true” source	also discuss the potential improvements w.r.t a comparison w.r.t a quantitative baseline

Big Data: the social microscope



Data Science
for #SocialGood

Measuring health and well-being



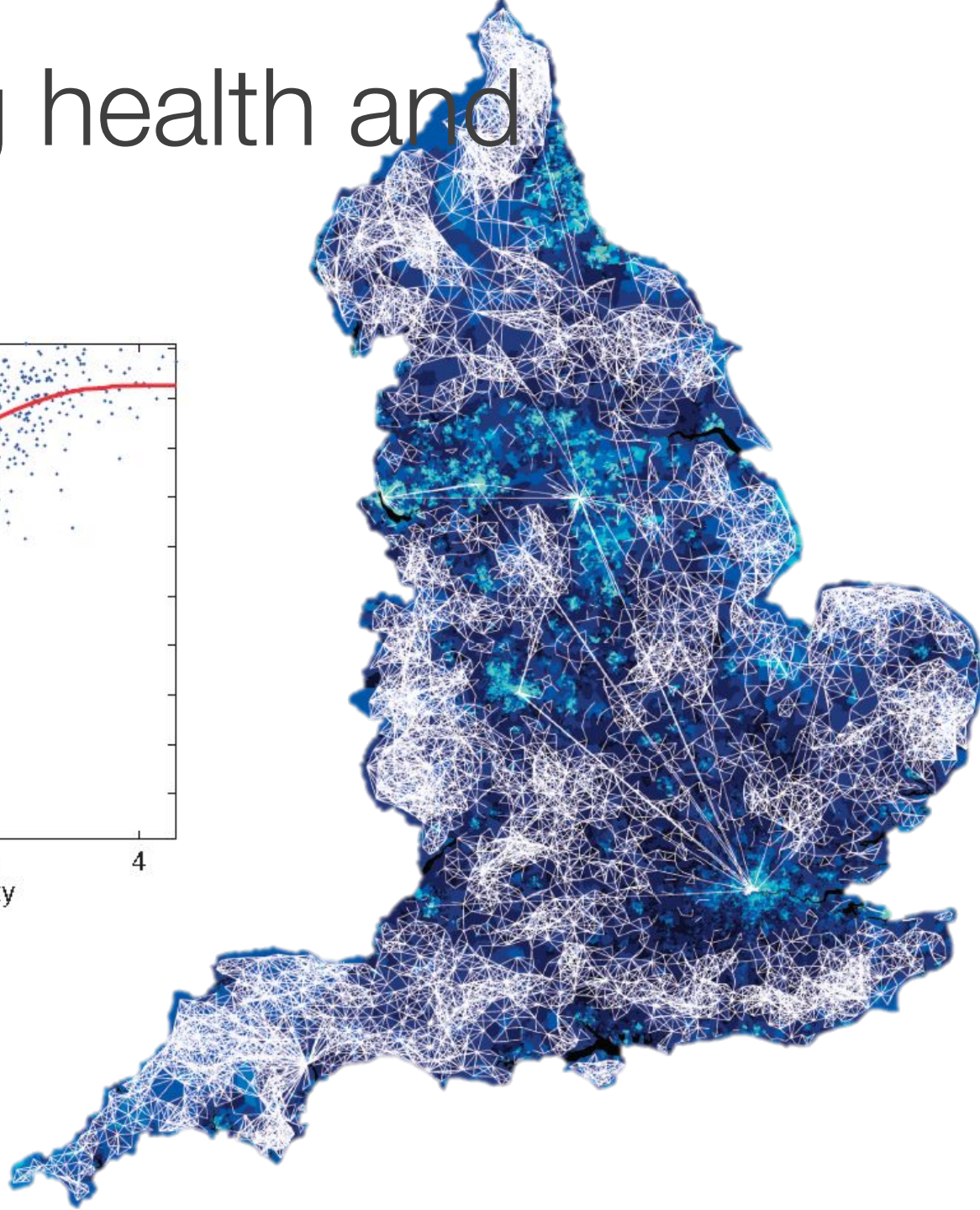
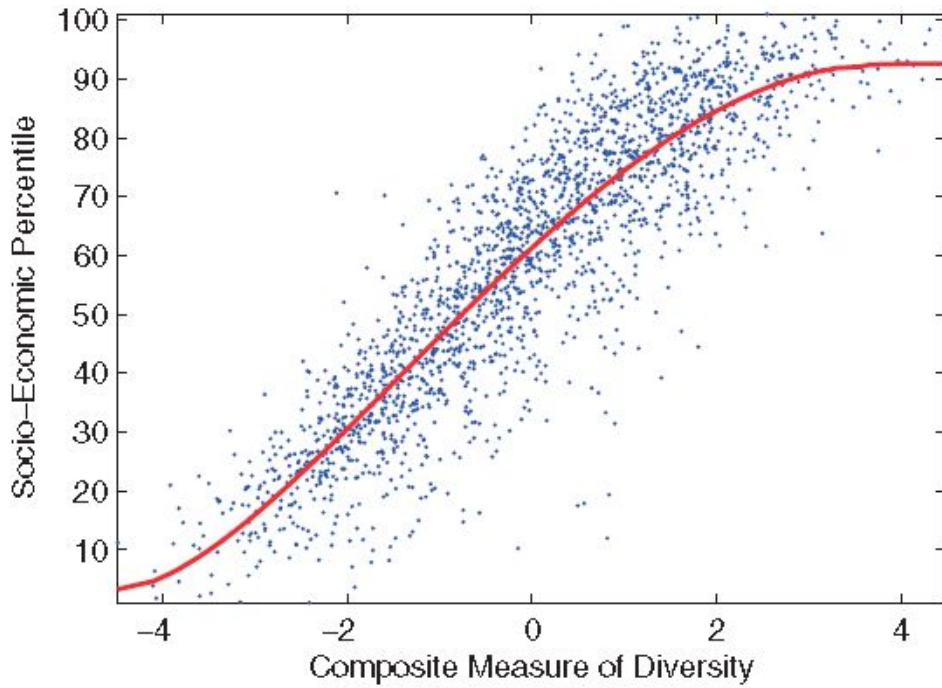
Detecting influenza epidemics using search engine query data

Jeremy Ginsberg¹, Matthew H. Mohebbi¹, Rajan S. Patel¹, Lynnette Brammer², Mark S. Smolinski¹ & Larry Brilliant¹

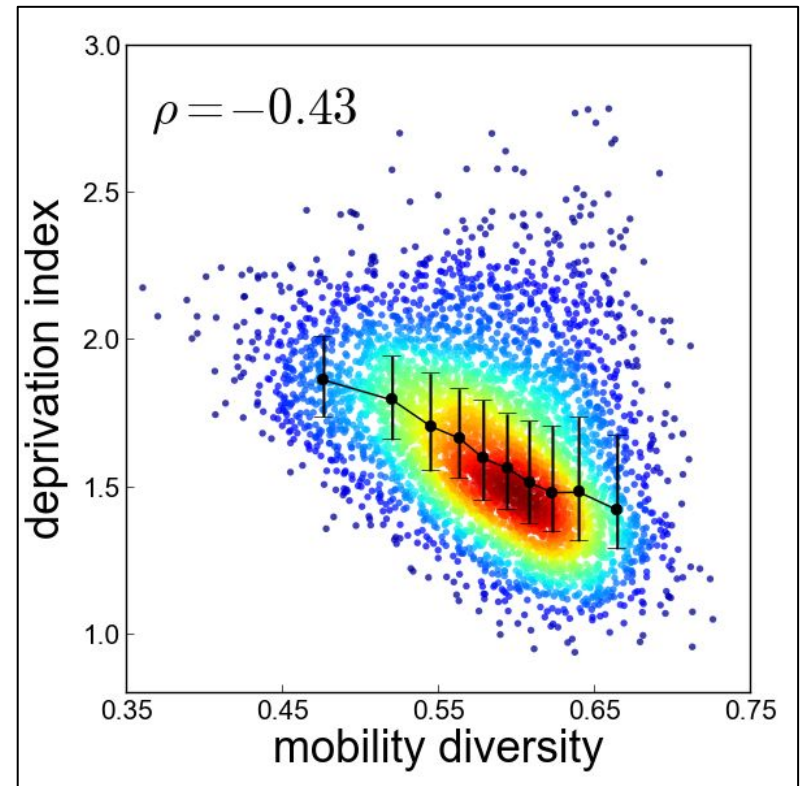
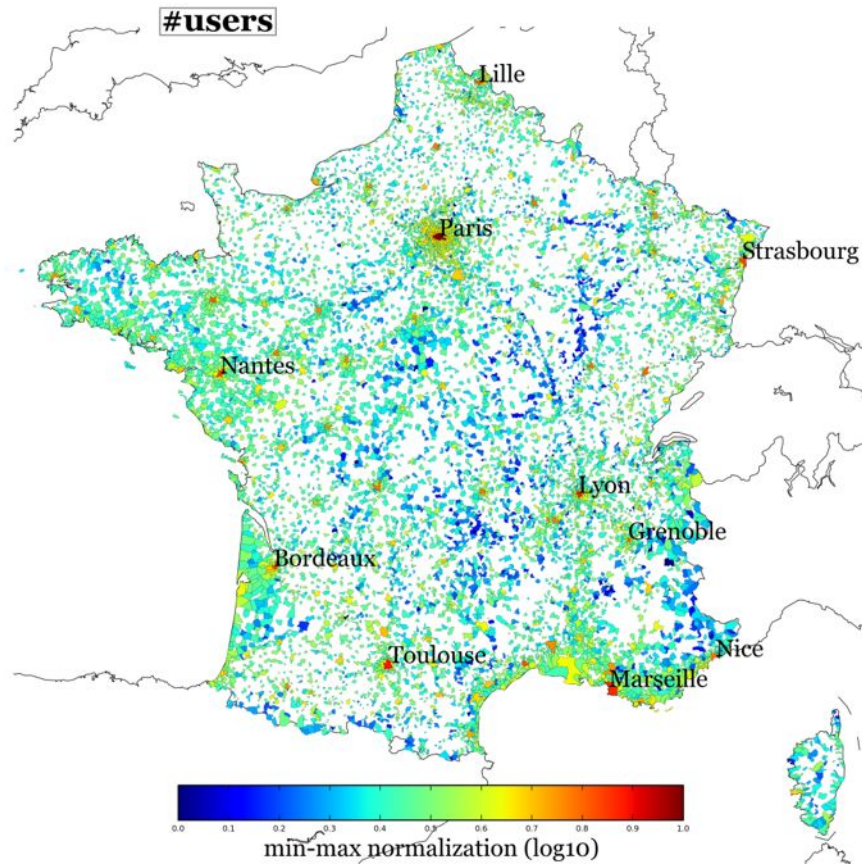
¹Google Inc. ²Centers for Disease Control and Prevention

Nature 457, 1012-1014 (2009)

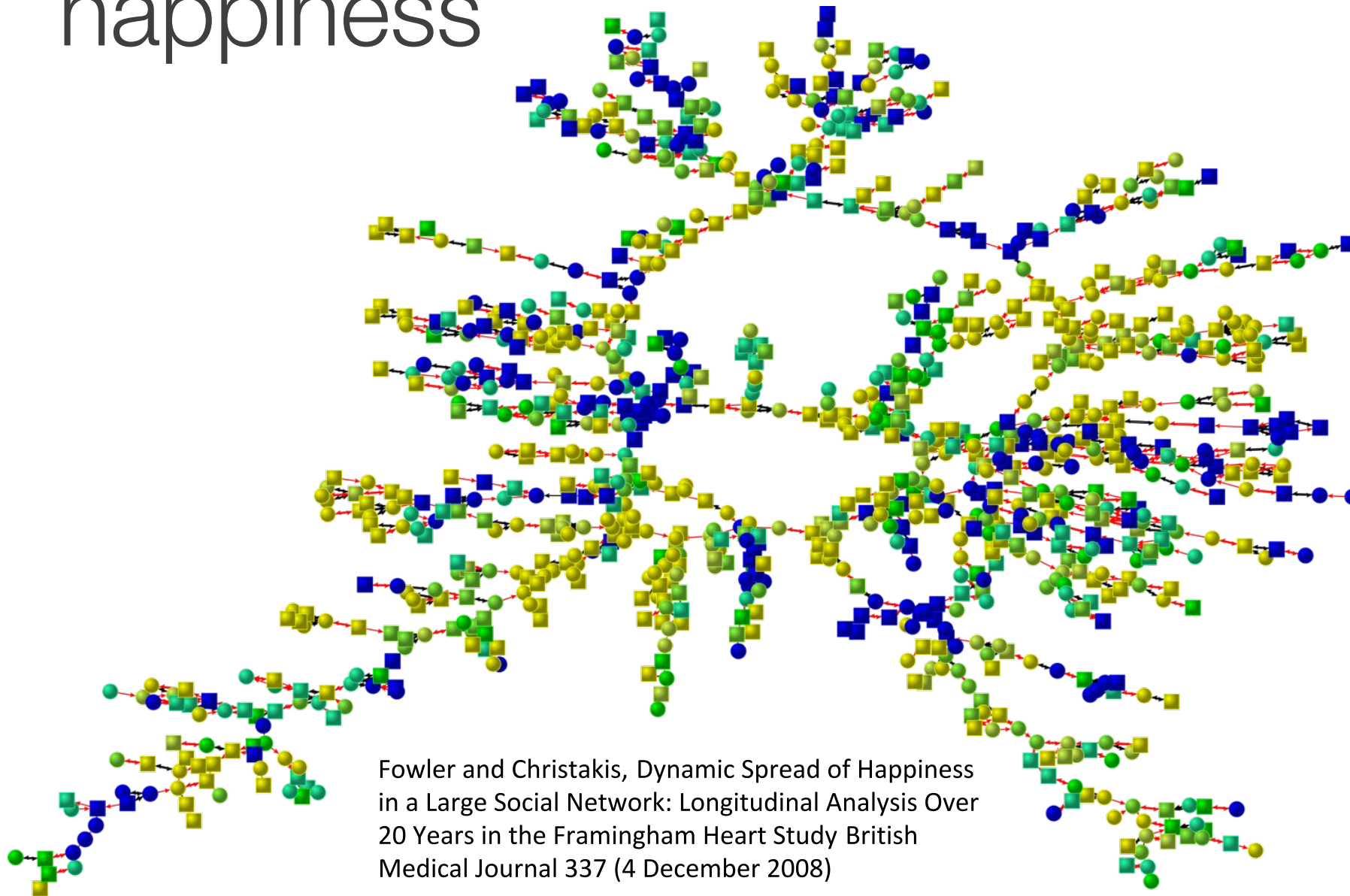
Measuring health and well-being



Measuring health and well-being



Measuring the spread of happiness



Health



- 1M customers
- 7K items
- 10 years



Health



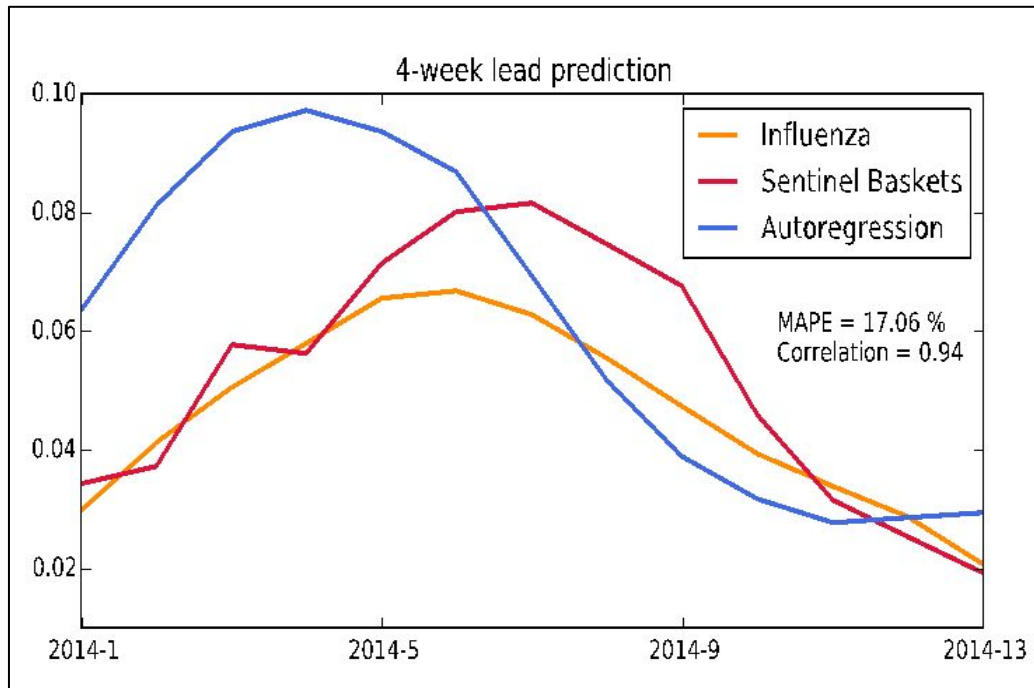
How to use [#BigData](#) to
[#forecast](#) the spread of
[#influenza](#)?

Health



- Identify products with trends similar to flu trends
- Identify [#sentinels](#): baskets of people buying during the peaks
- Predict weekly values of flu
extended regression model

Health

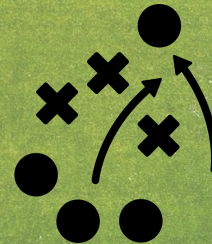
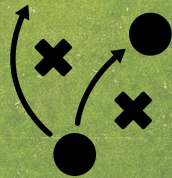
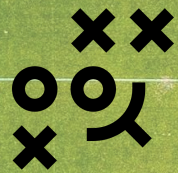


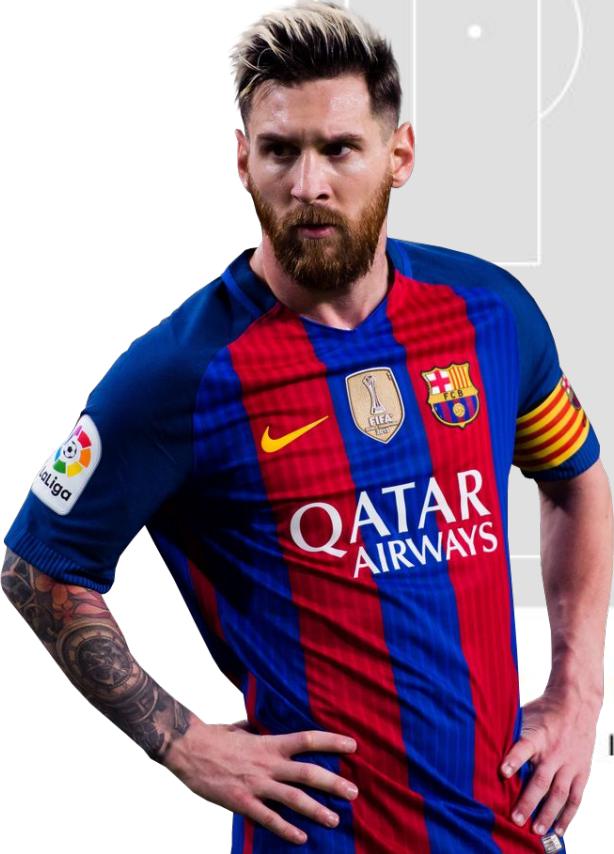
from 50% error
to 20% error

in 4-week
prediction

Soccer Analytics

when Data Science takes the field





Intercepts

1



Crosses

4



Takeons

8



Clearances

0



Tackles

15



Shots

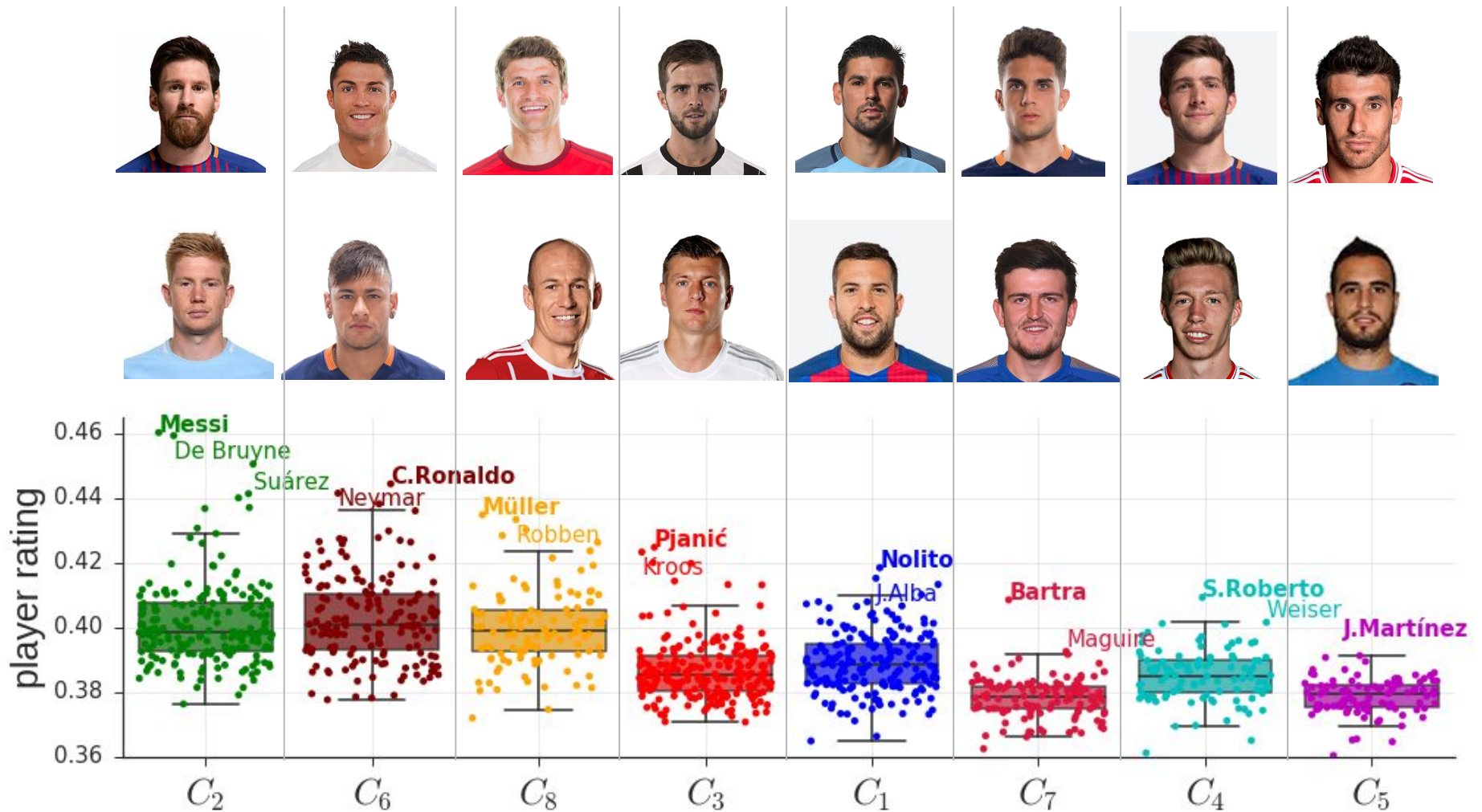
5



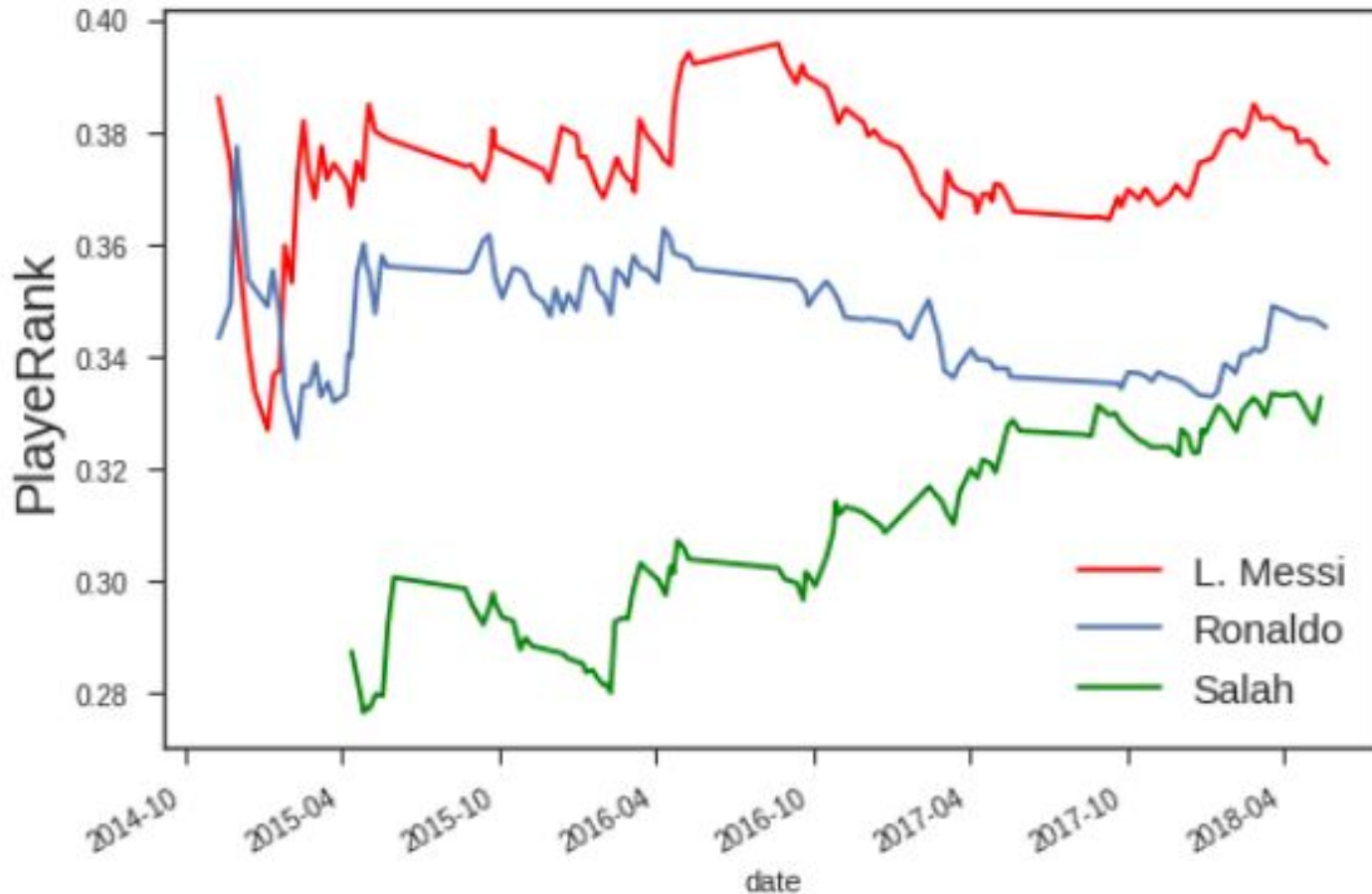
Passes

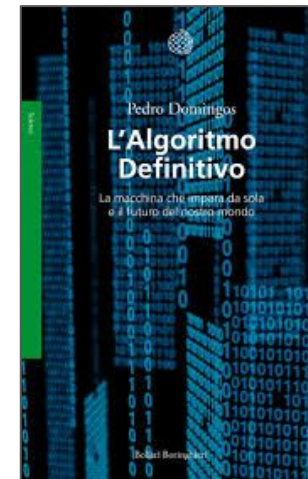
47

Best players in the dataset



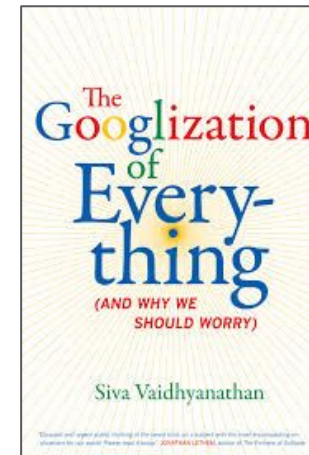
Evolution of players





L'algoritmo definitivo

Pedro Domingos



The Googlization of everything

Siva Vaidhyanathan