Big Data Analytics 2021/2022

Instructors:

LUCA PAPPALARDO

FOSCA GIANNOTTI

Tutor:

Giuliano Cornacchia

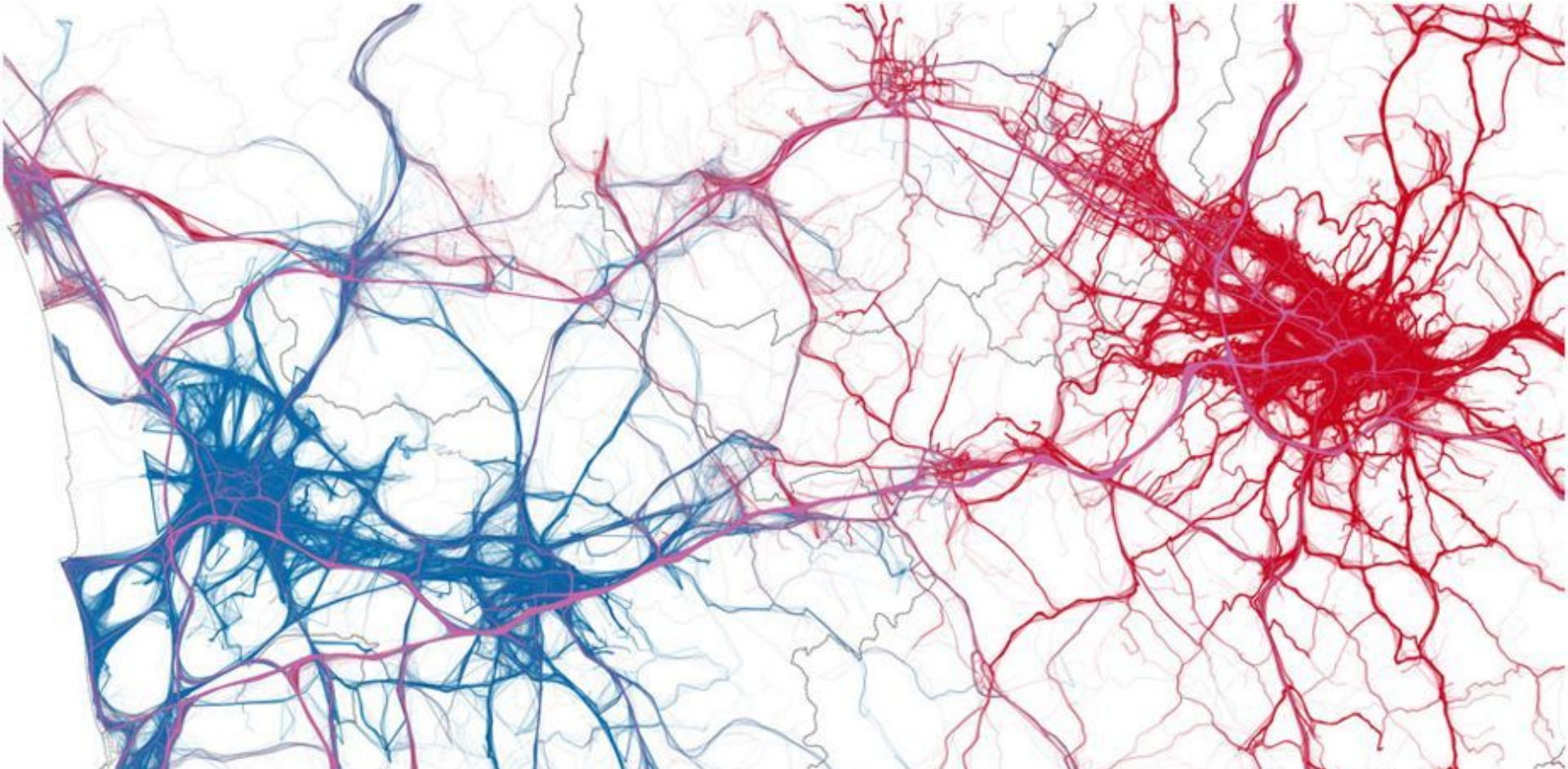
Consiglio Nazionale
delle Ricerche

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

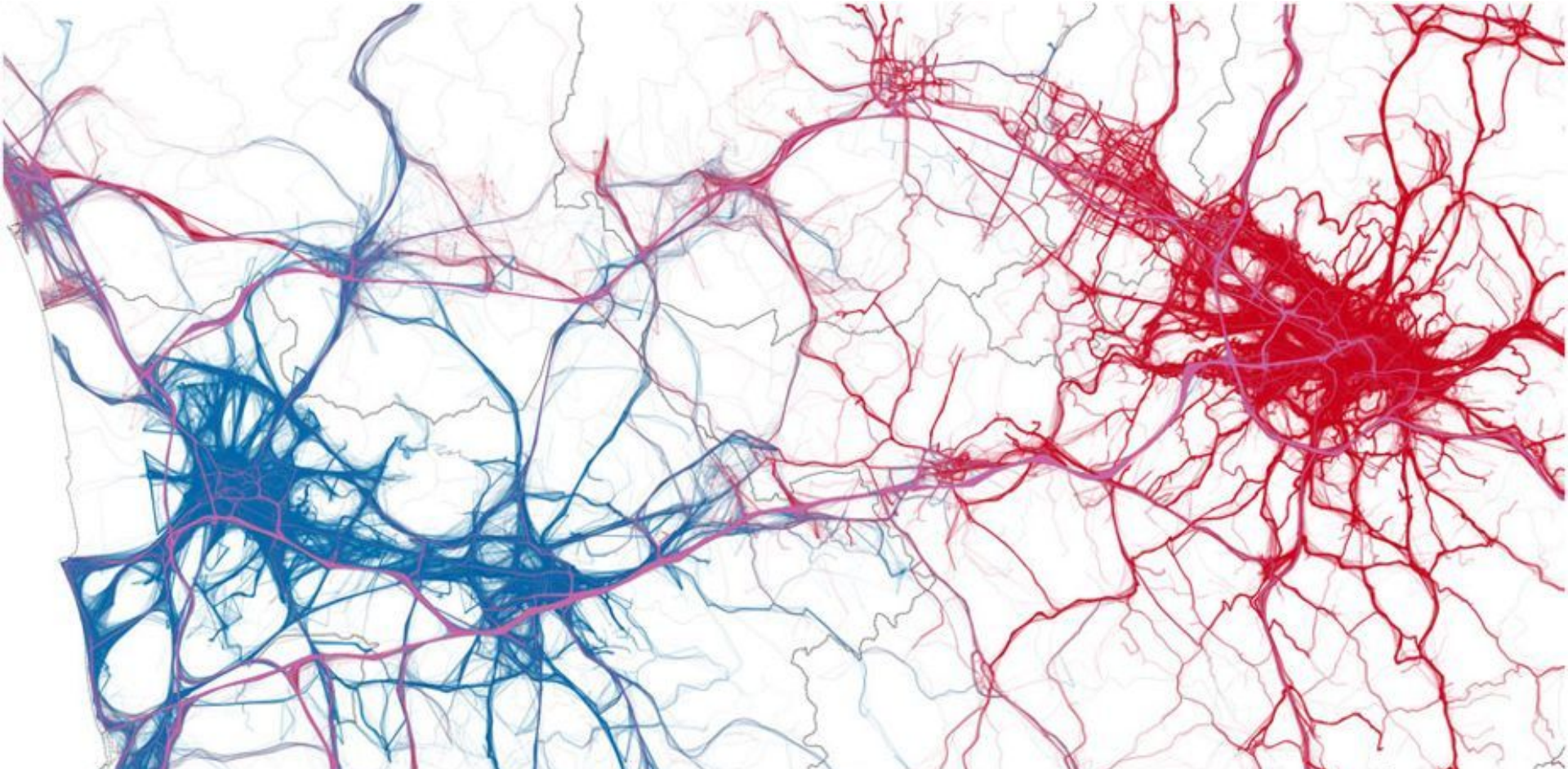
Important info

- when you come to the classroom, remember to frame the QR code of the place where you sit
- this is to confirm your presence in view of the contact tracing required by the health protocols
- failing to do so reduces your priority in the subsequent assignment of face-to-face places for this course


2005
Luca Bruno / AP
2013
NBC NEWS
Michael Sohn / AP

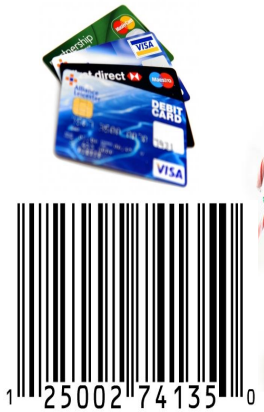


Private vehicles traveling in Tuscany (on-board GPS devices)

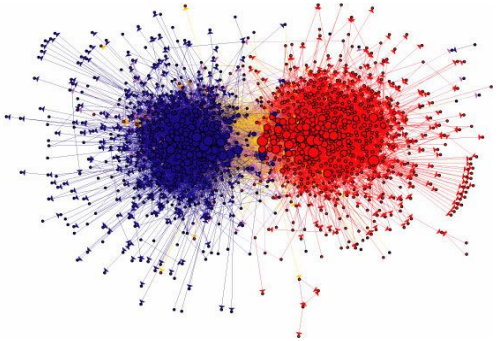


Digital Footprints of Human Activities

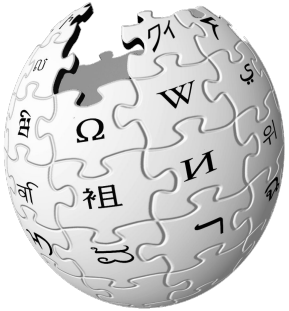
Shopping patterns



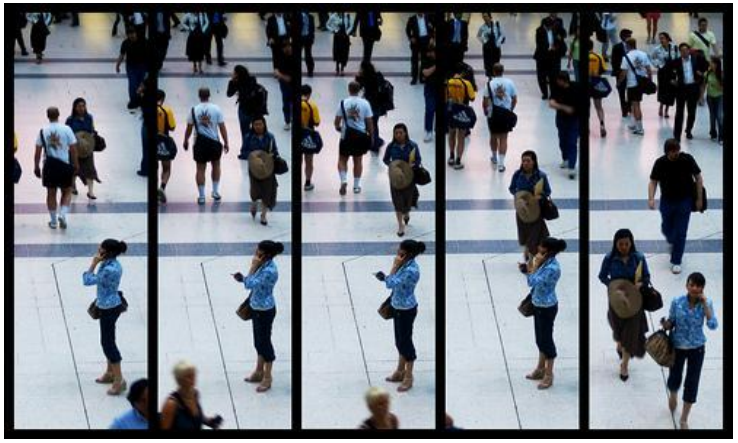
Social Ties



Opinions



Movements



The **V**s of Big Data

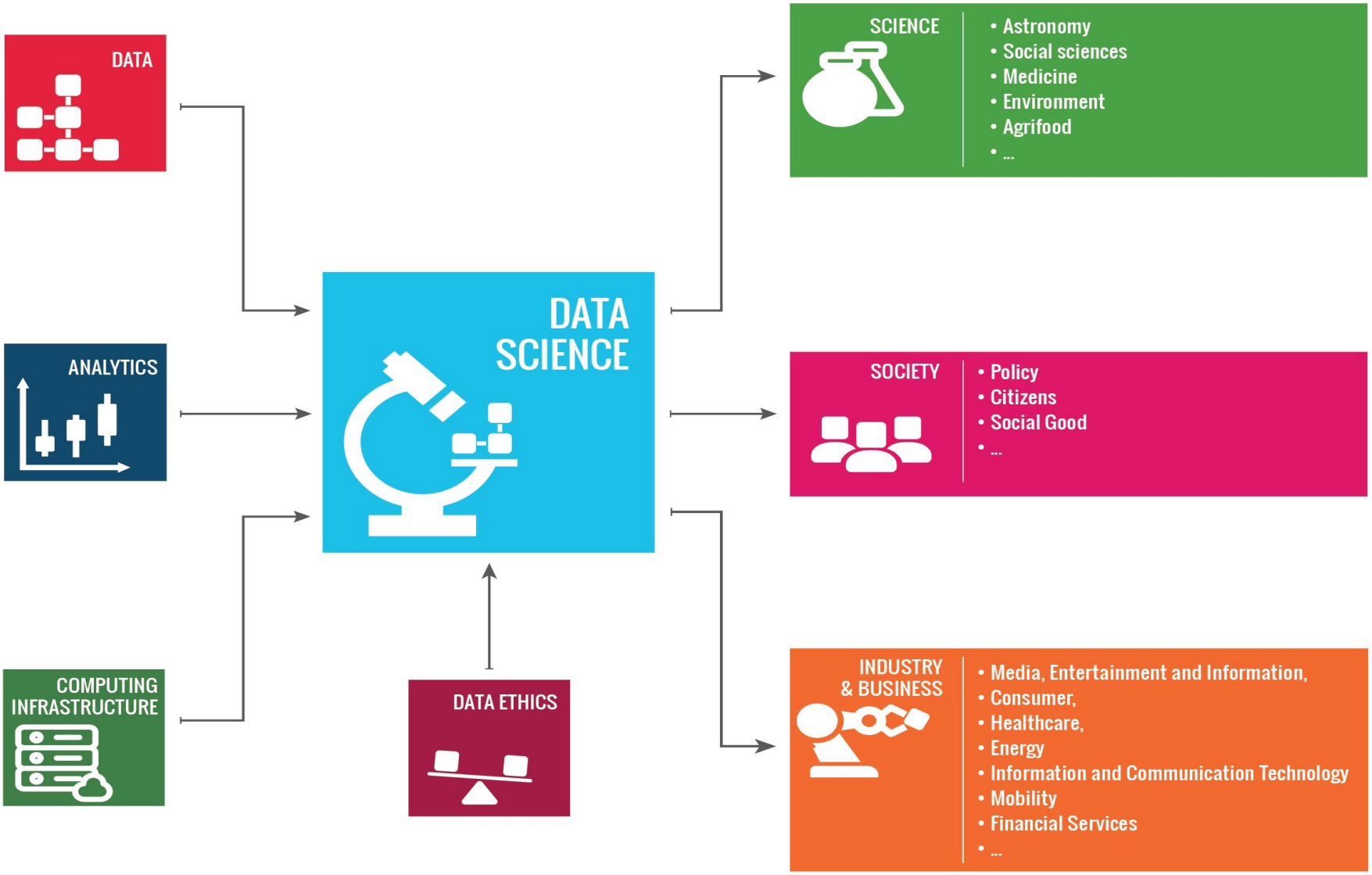
Volume: the incredible amounts of data generated each second

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


DATA
ANALYTICS
COMPUTING INFRASTRUCTURE
DATA SCIENCE
DATA ETHICS
SCIENCE
• Astronomy
• Social sciences
• Medicine
• Environment
• Agrifood
• ...
SOCIETY
• Policy
• Citizens
• Social Good
• ...
INDUSTRY & BUSINESS
• Media, Entertainment and Information,
• Consumer,
• Healthcare,
• Energy
• Information and Communication Technology
• Mobility
• Financial Services
• ...

WORLD ECONOMIC FORUM

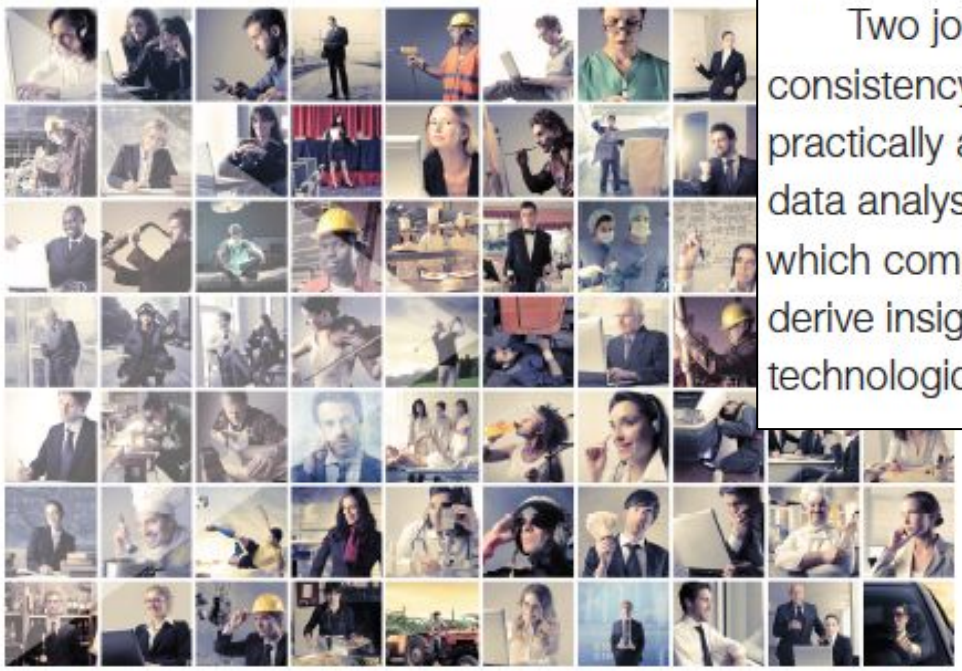
COMMITTED TO IMPROVING THE STATE OF THE WORLD

Global Challenge Insight Report

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

WORLD ECONOMIC FORUM

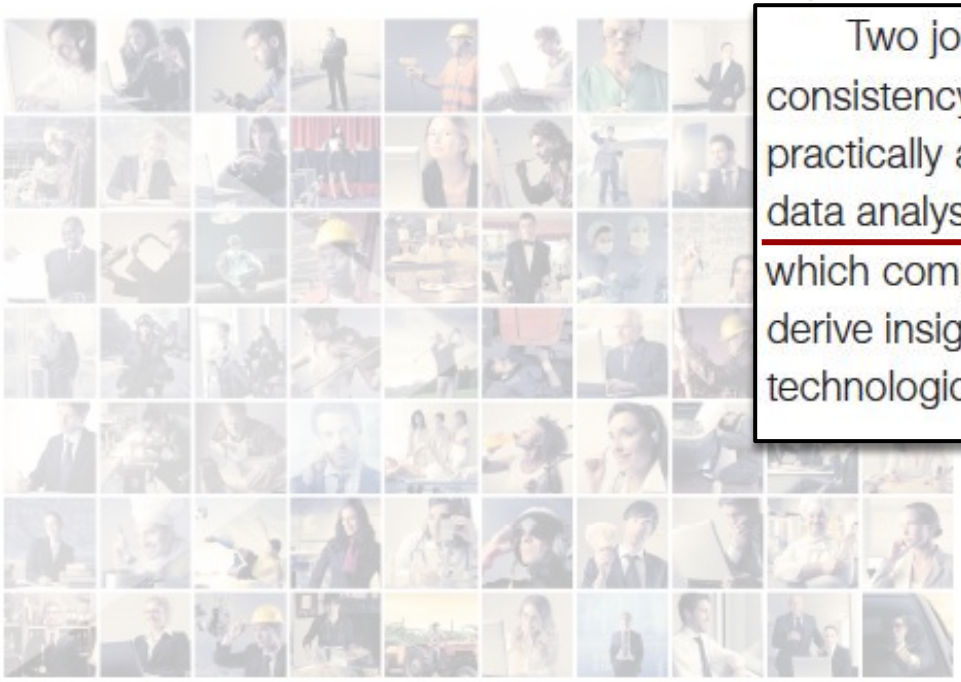
COMMITTED TO IMPROVING THE STATE OF THE WORLD

Global Challenge Insight Report

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

Harvard Business Review

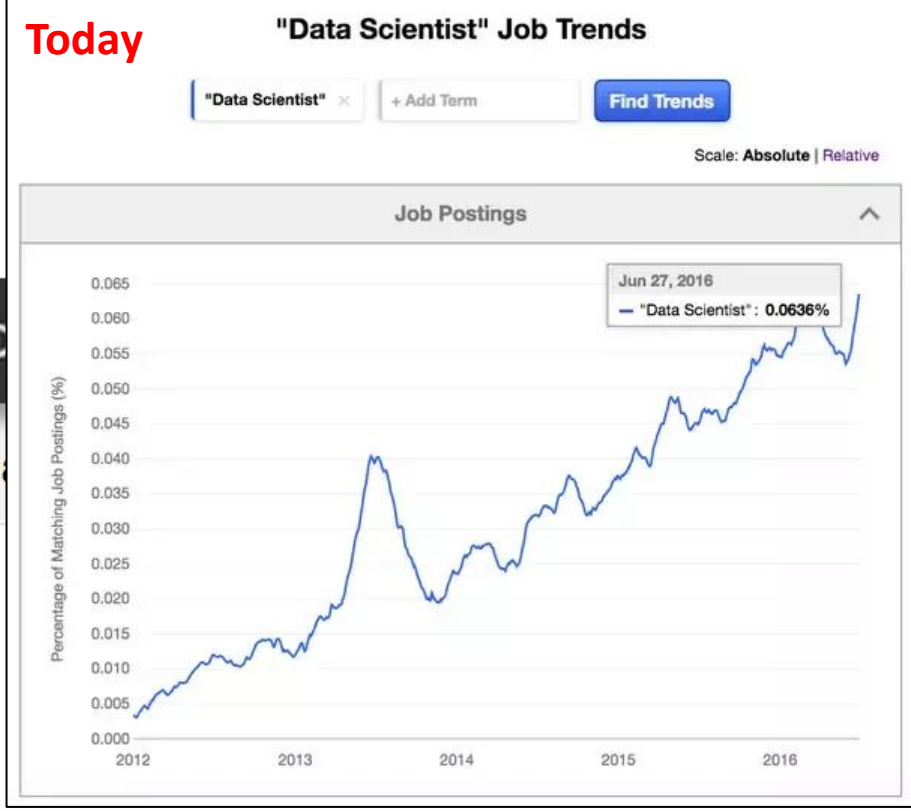
THE MAGAZINE BLOGS VID

Guest Subscribe today

THE MAGAZINE

October 2012

Today



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 and develop a project

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/February*: 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, NN, 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

César A. Hidalgo @cesifoti · Sep 12

For most of their education students are evaluated using tests & homework. We are all familiar with the process. The student is asked to do some work; they turn it in, and get a grade (eg a B+,B, A, etc.) 2/N

2 60 809

César A. Hidalgo @cesifoti · Sep 12

But when they make it to grad school (or to their first job) it is quite different. Once they turn in work, they are not given a grade. They are given feedback and asked to do the work again. Sometimes several times. 3/N

8 123 1.3K

César A. Hidalgo @cesifoti · Sep 12

Many students struggle enormously with this change. They are not trained to do the same thing over and over until they get it right. They are trained to do better than average on their first try and to forget about what they did as soon as they turn their work in. 4/N

César A. Hidalgo @cesifoti · Sep 12

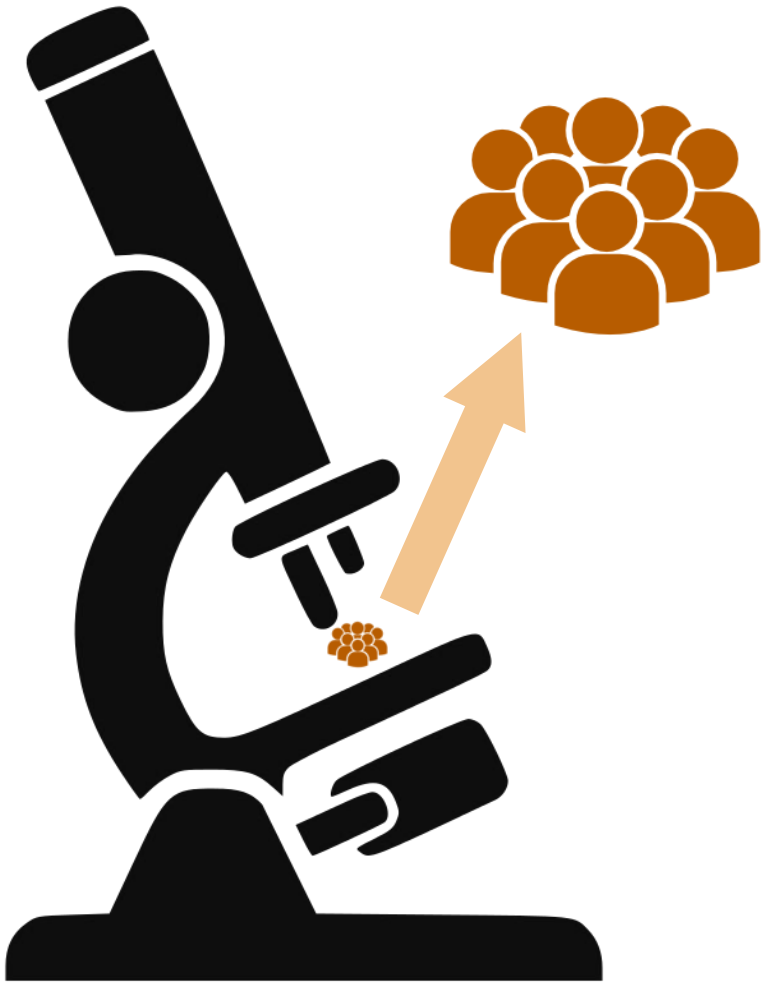
So students show up with a first draft of a paper and tell you “I am done!” So you give them feedback and ask them to keep working on it. And you do this again, and again, and tension can begin to build. 5/N

6 99 1.2K

César A. Hidalgo @cesifoti · Sep 12

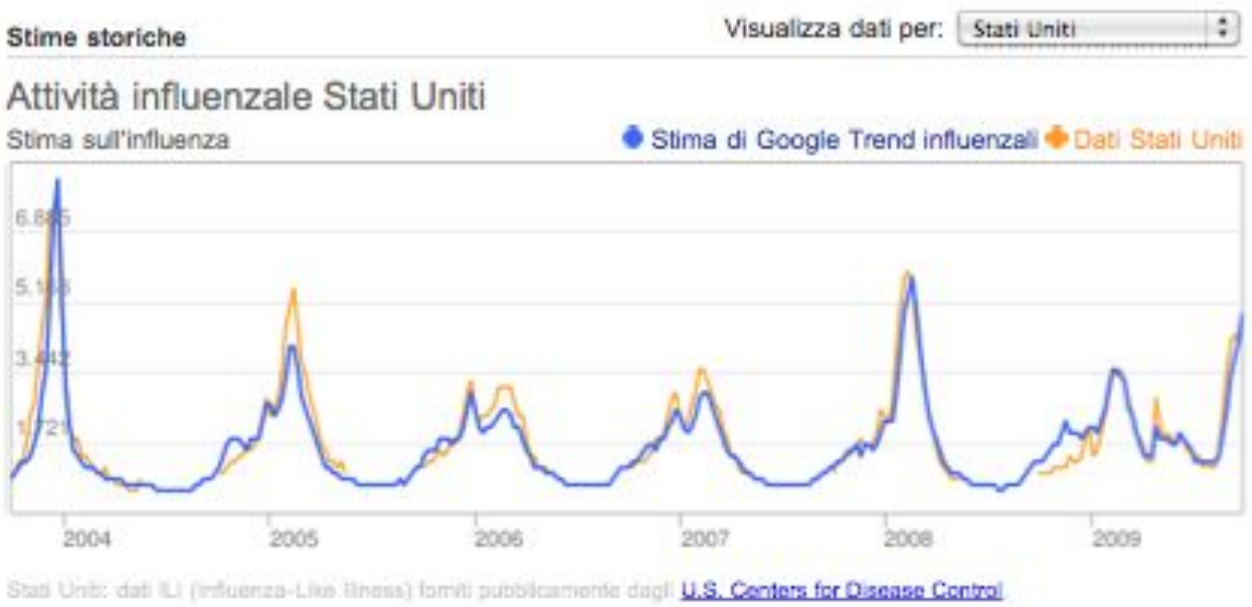
It is a tough aha moment that helps them realize work is always unfinished. You are there to help improve the work, not to grade it. And that only by iterating with others they will manage to make a sword out of the lump of steel. 8/N

Big Data:
the social microscope



Data Science for #SocialGood

Measuring health and well-being



Google

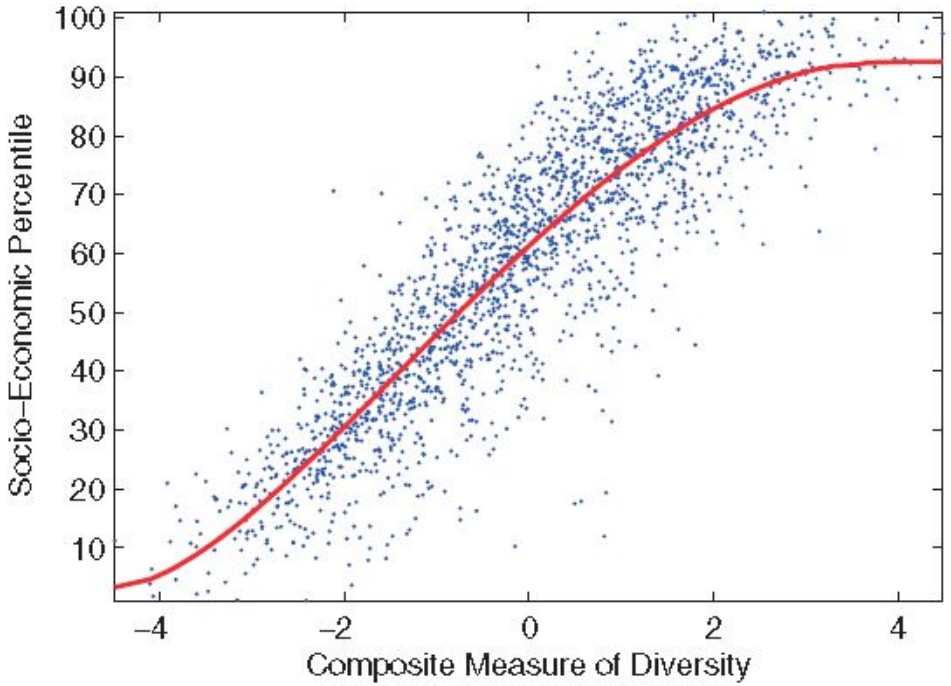
Detecting influenza epidemics using search engine query data

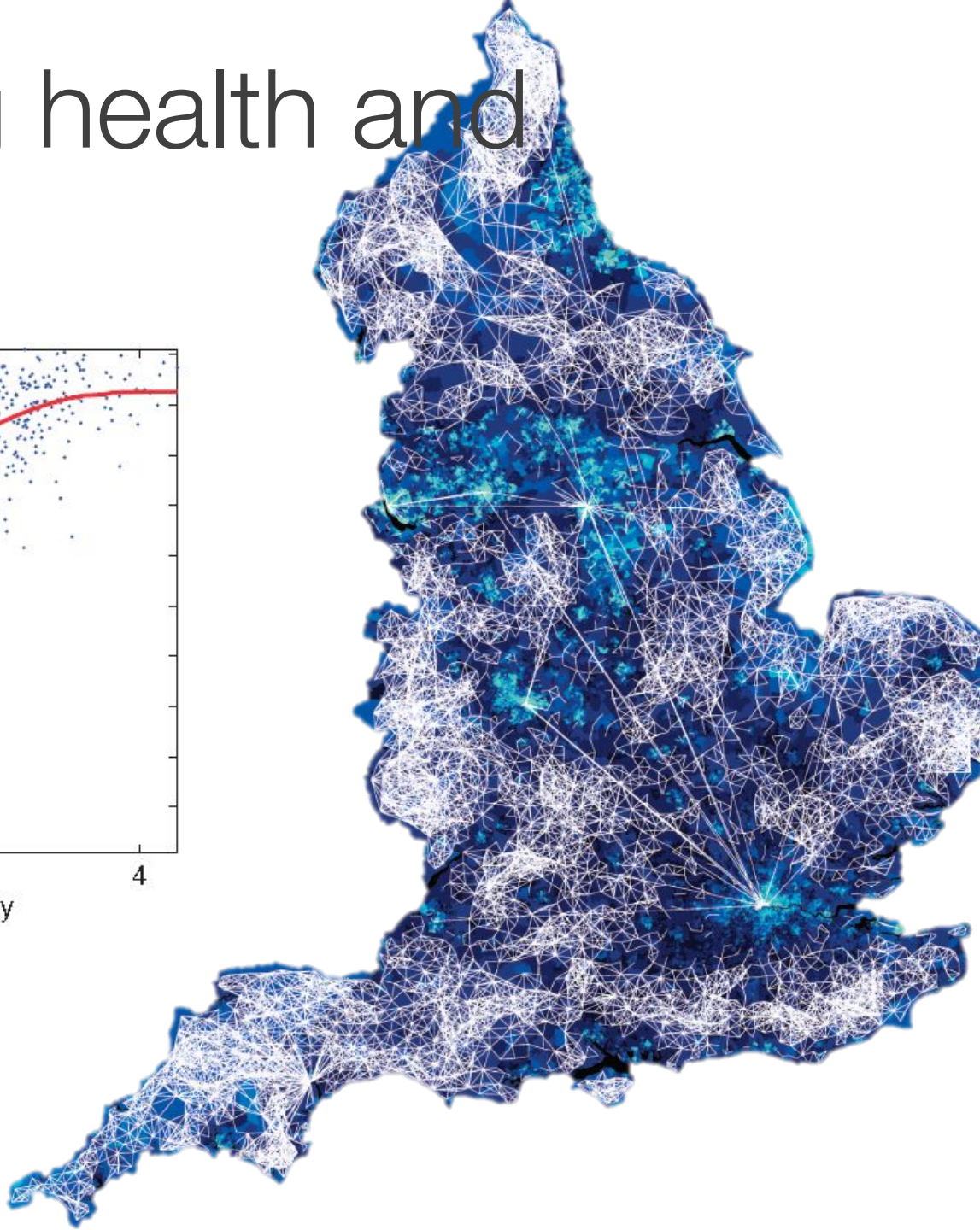
Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

[1]*Google Inc.* [2]*Centers for Disease Control and Prevention*

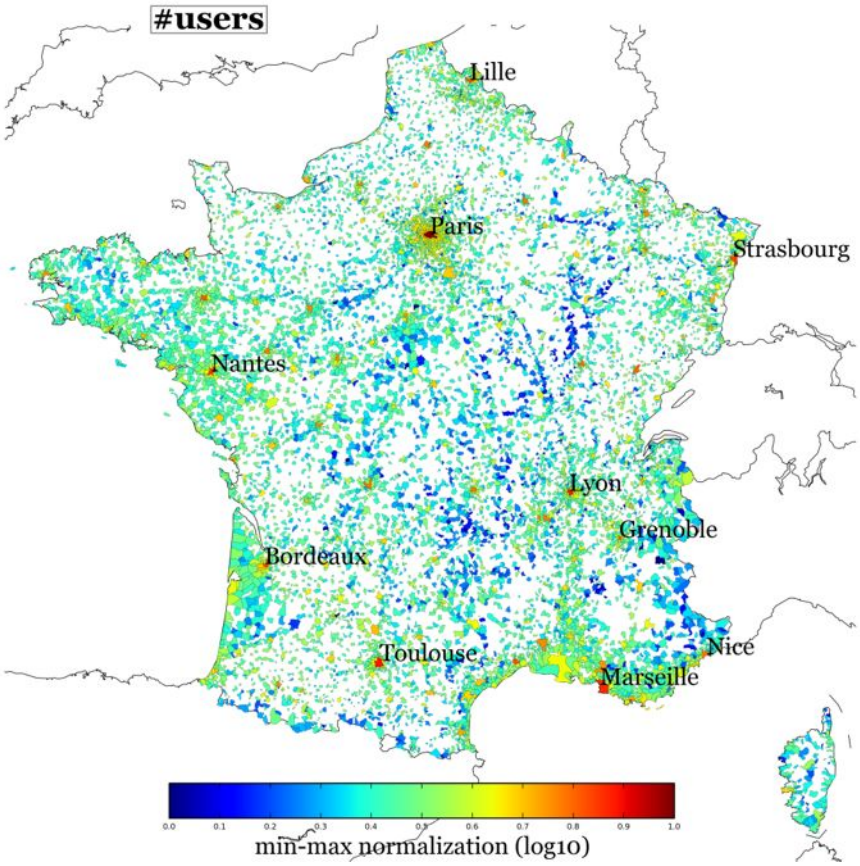
Nature 457, 1012-1014 (2009)

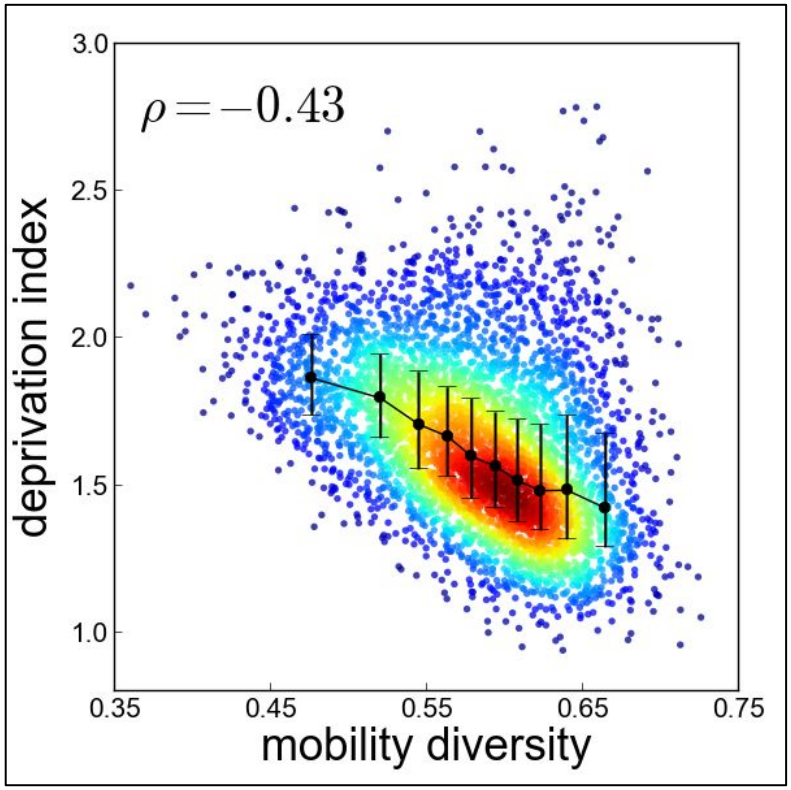
Measuring health and well-being



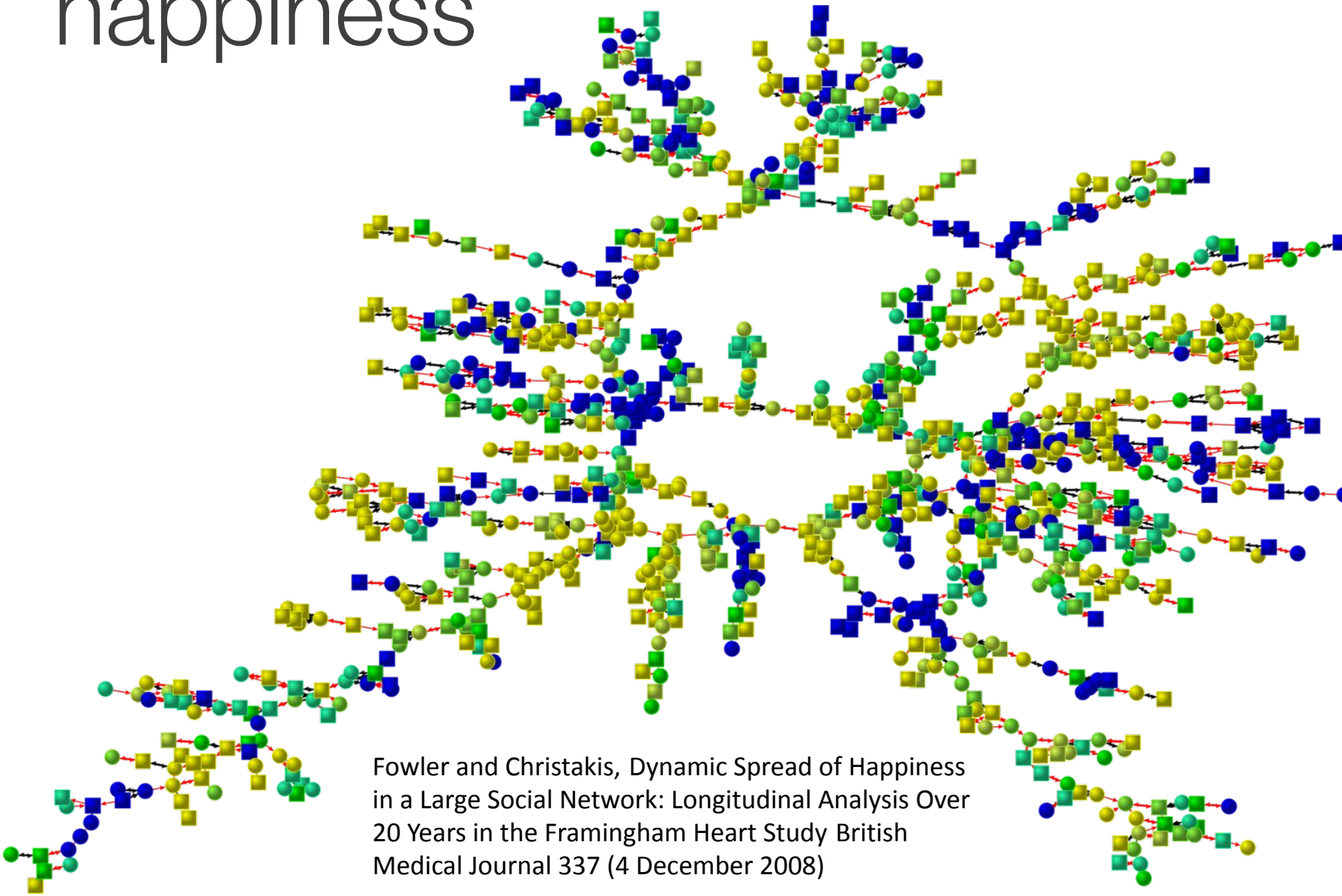


Measuring health and well-being





Measuring the spread of happiness



Fowler and Christakis, Dynamic Spread of Happiness in a Large Social Network: Longitudinal Analysis Over 20 Years in the Framingham Heart Study British Medical Journal 337 (4 December 2008)

Health

- 1M customers
- 7K items
- 10 years

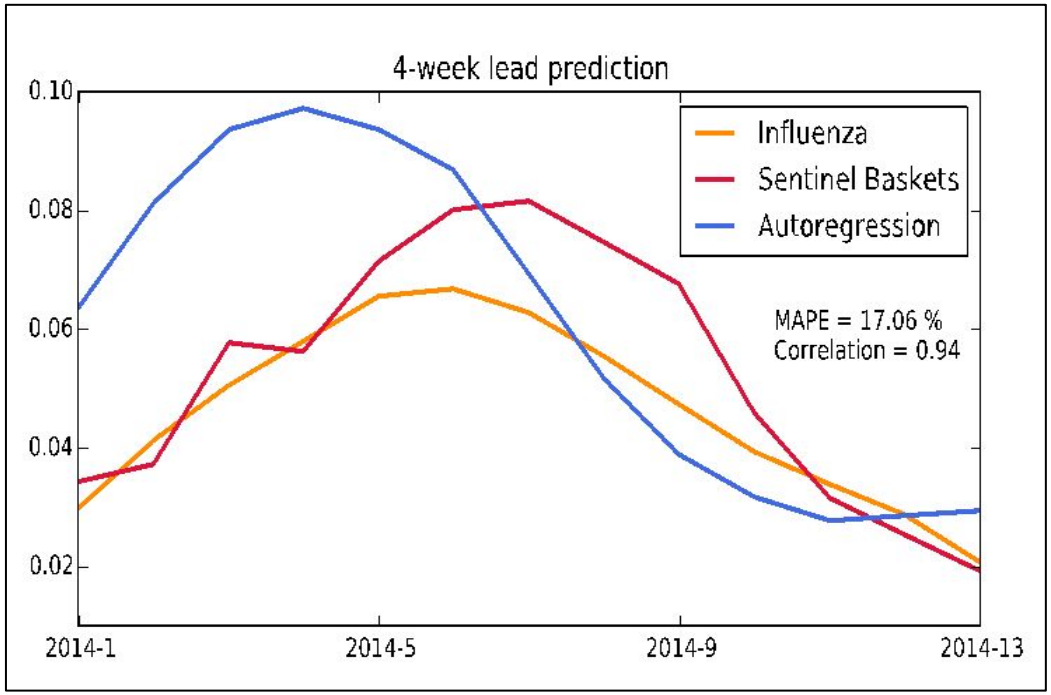
Health

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

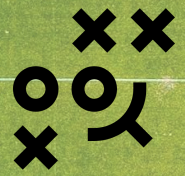


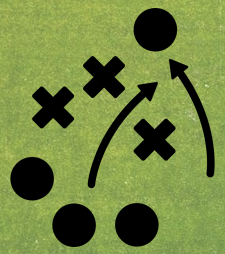
from 50% error
to 20% error

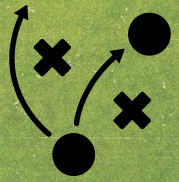
in 4-week
prediction

Soccer Analytics

when Data Science takes the field

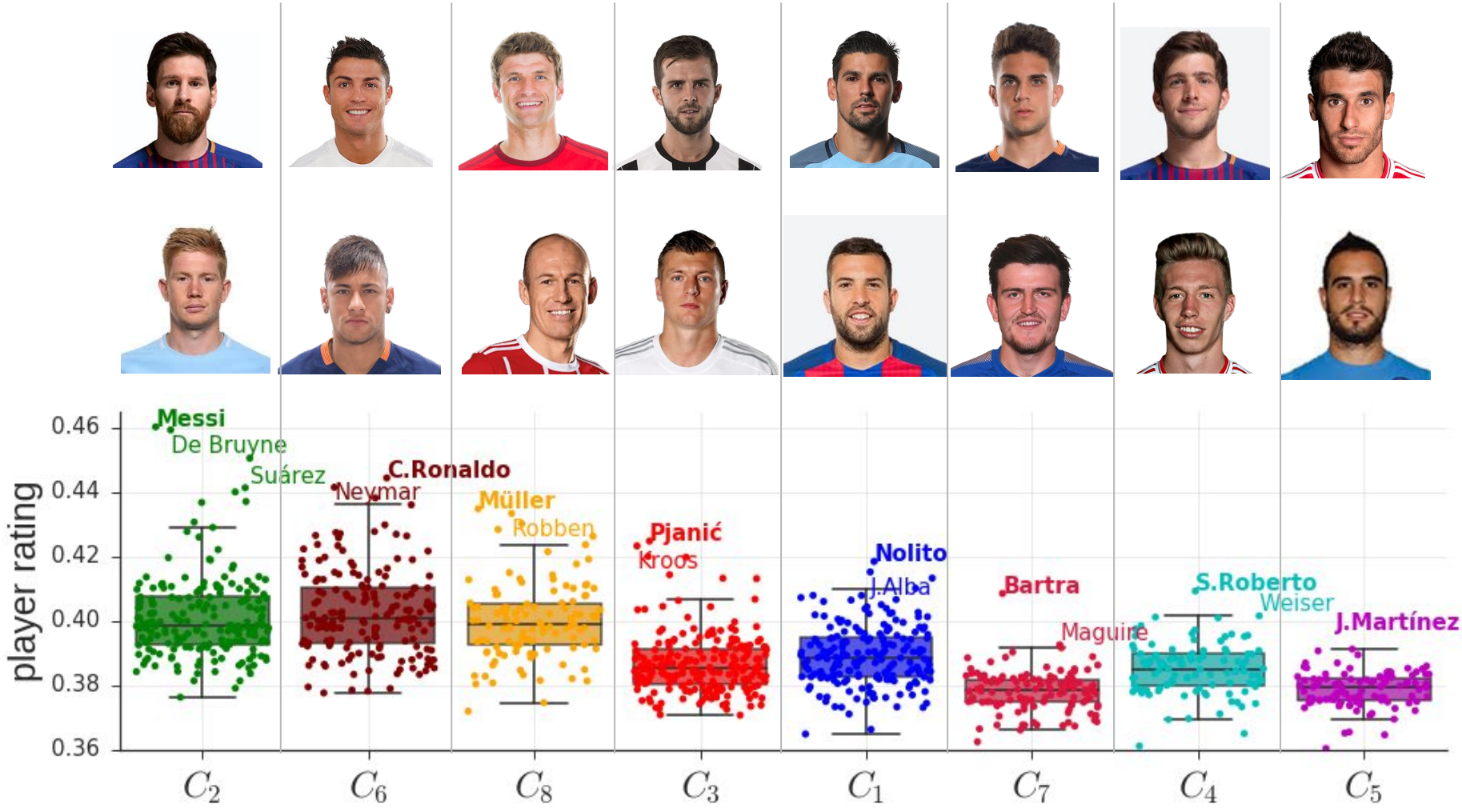




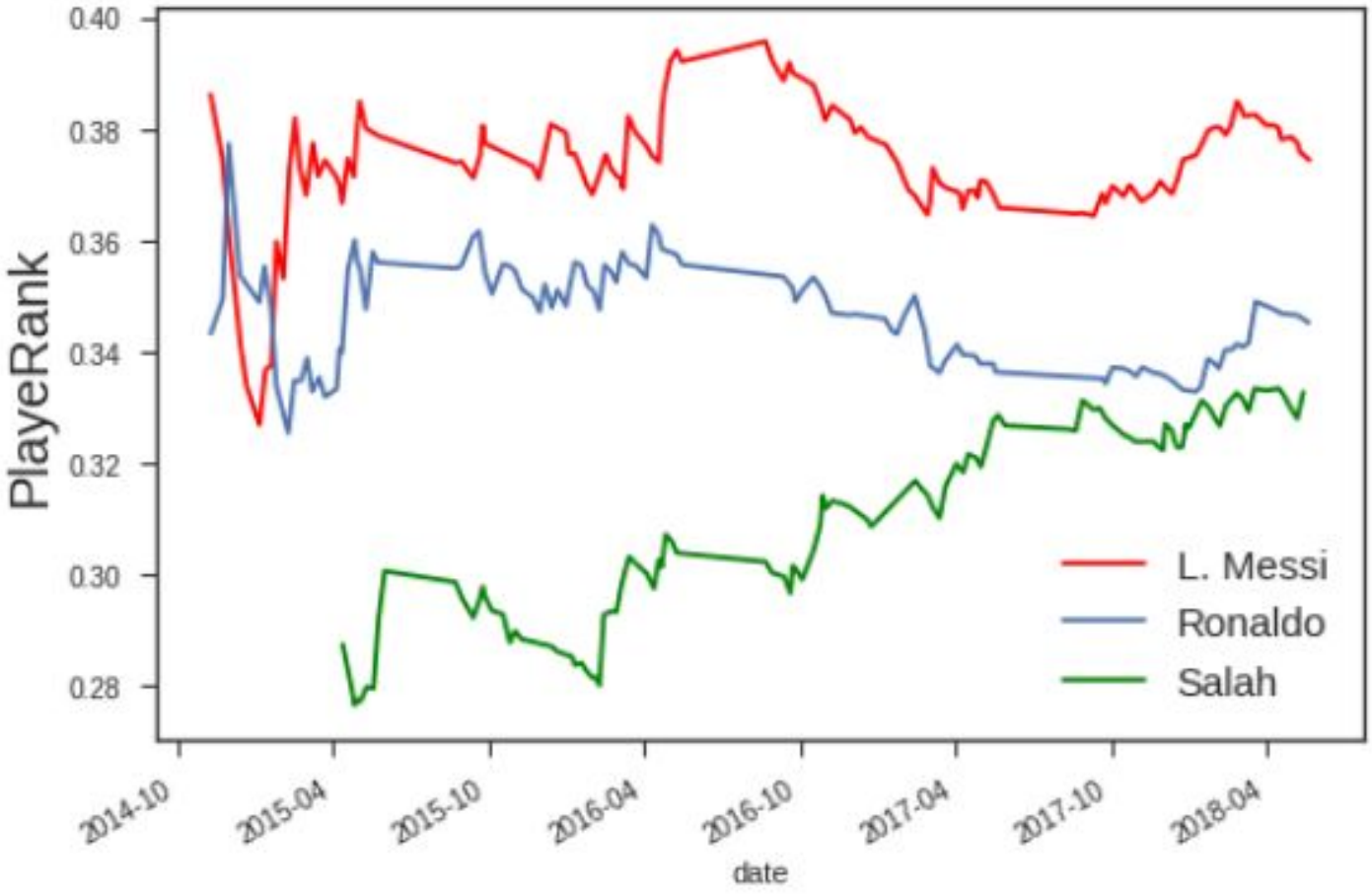


Intercepts	Crosses	Takeons	Clearances	Tackles	Shots	Passes
1	4	8	0	15	5	47

Best players in the dataset

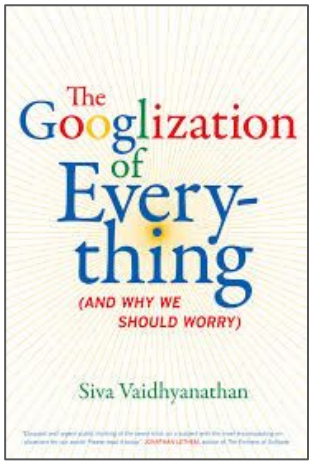


Evolution of players



L'algoritmo definitivo

Pedro Domingos



The Googlization of everything

Siva Vaidhyanathan

SPECIAL | 07 JULY 2021

Computational social science

The availability of big data has greatly expanded opportunities to study society and human behaviour through the prism of computational analyses. The resulting field is known as computational social science and is defined by its interdisciplinary approaches. However,... show more

Special home Panel at Networks 2021

https://www.nature.com/collections/cadaddgige/