Big Data & Data Science: Economic applications

José García Montalvo

Master of Data Science
Course: "Economics for the Era of Big Data"
October 13, 2015

Summary

- Introduction
- Some preliminary comments on data science and big data
- Big data and economics
- My experience with big data and economics:
  - Finding the real price of housing in Spain
  - Breaking the world 100 times in small pieces
  - Looking at infrastructures and companies
  - Men, women and scoring
  - Electoral predictions
  - Big data for financial services and marketing
- Danger of *big data:* confidentiality and correlation
- Concluding remarks

Introduction

- "In God we trust; all other must bring data", Edwards Deming
- "Without data you are just one more person with an opinion" (anónimo)
- "We are drowning in information but starved for knowledge." John Naisbitt

Introduction

- "Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it." Dan Ariely

Introduction: Target's legend

- Excerpt from "Predictive Analytics" by Eric Siegel:

  In 2010, I invited an expert at Target, Andrew Pole, to keynote at Predictive Analytics World, a conference for which I serve as program chair. Pole manages dozens of analytics professionals who run various predictive analytics (PA) projects at Target. In October of that year, Pole delivered a stellar keynote on a wide range of PA deployments at Target. He took the stage and dynamically engaged the audience, revealing detailed examples, interesting stories, and meaningful business results that left the audience clearly enthused.

Introduction: Target's legend

- Excerpt from "Predictive Analytics" by Eric Siegel:

  Toward the end, Pole describes a project to predict customer pregnancy. Given that there's a tremendous sales opportunity when a family prepares for a newborn, you can see the marketing potential.

- Featured in the New York Times (Charles Duhigg,"How Companies Learn Your Secrets", 16-2-12) and FT

Introduction: Target's legend

- Andrew Pole had just started working as a statistician for Target in 2002, when two colleagues from the marketing department stopped by his desk to ask an odd question: "If we wanted to figure out if a customer is pregnant, even if she didn't want us to know, can you do that? "
- Pole has a master's degree in statistics and another in economics
- "We knew that if we could identify them in their second trimester, there's a good chance we could capture them for years," Pole told me. "As soon as we get them buying diapers from us, they're going to start buying everything else too.

Introduction: Target's legend

- Pole was working for Target's Guest Marketing Analytics department": he was a statistician and a mathematician
- Pole was asked to find unique moments of the life of costumers such that consumption habits could be flexible enough to be attracted by the department store as a loyal client -> divorce, moving to a new house, birth of a child,… are some of those moments
- … however when the child is born it is already too late: they needed to "attract" the costumer before that moment

Introduction: Target's legend

- In those particular situation costumers are more vulnerable to marketing
- They identify 25 products (calcium supplements, magnesium, zinc, unscented soap, large cotton bags, etc.) that pregnant women buy during the first 20 weeks as predictors of pregnancy
- They used "habit looping" algorithm

Introduction: the case of Amazon

- Amazon used dozen of literary critics to suggest titles to its costumers until 2001 -> "Amazon voice" was considered by the WSJ as the most influential critic of the US
- At some point Jeff Bezos questioned if there was not better way to offer client specific recommendations: using the history of products bought by a costumer plus similarities with other costumer plus Linden's "item by item" algorithm the automatic recommendation system beat big time the human critics

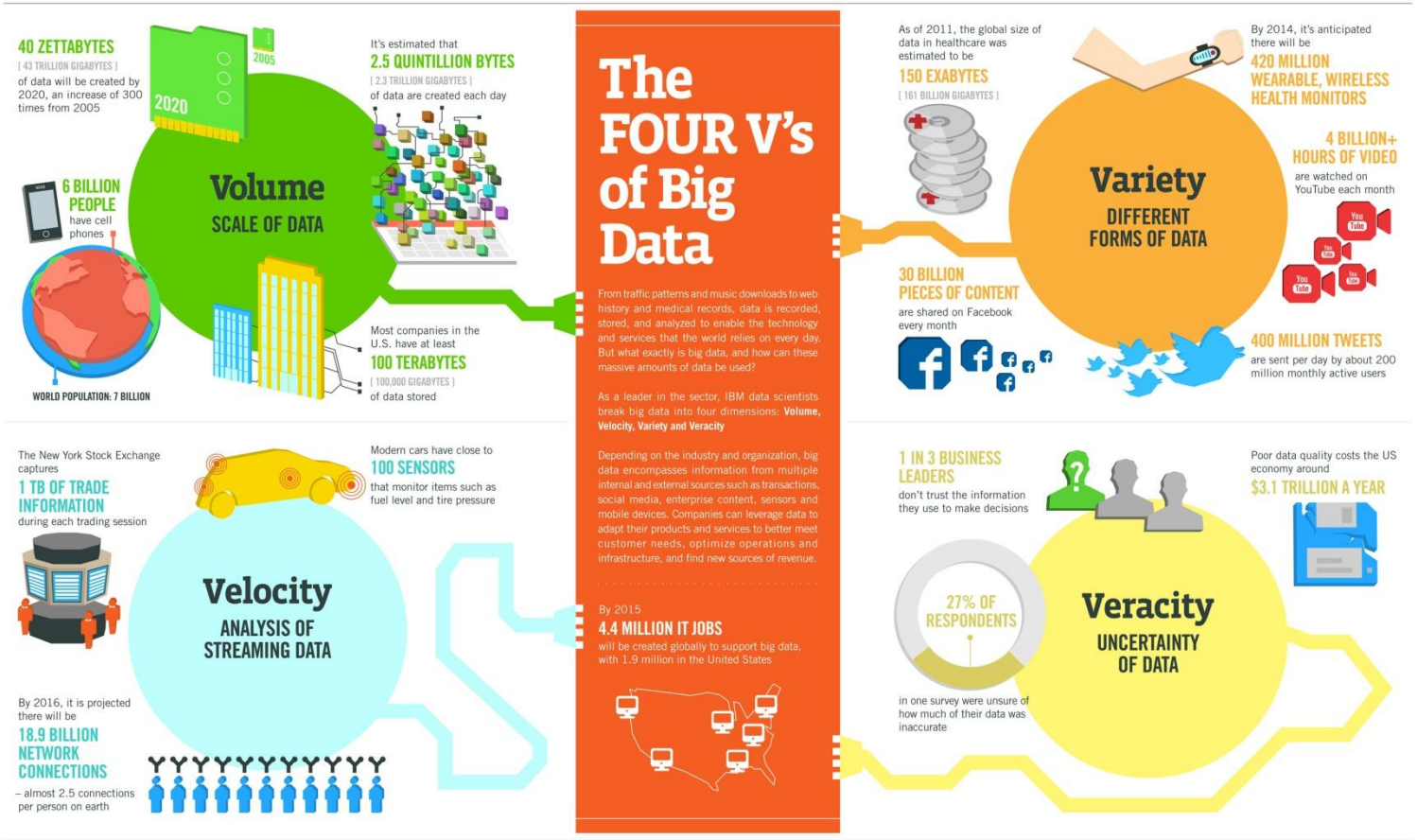
Introduction: the case of Amazon

- "Amazon voice" o "machine learning"? Human critics or algorithms? -> Algorithm won clearly and all critics were fired
- Today 1/3 of the sale of Amazon come from the system of personalized recommendations
- Linden's item by item algorithm has been adopted by many digital shops including Netflix

Data Science

- During the last millenium science has been an empiric endeavor (description of natural phenomenon)
- During the last centuries science opened up to model, formulations and generalizations
- In recent decades computational science and simulation of complex phenomenon
- Currently eScience:
  - Unifies theory, experiments and simulation
  - Massive data capture using specific software or generating them by simulation
  - Knowledge and information is stored in computers

Big Data



barcelona gse
graduate school of economics

Big data

- Massive amount of information: “the sample is the population” -> anything is potentially useful
- Huge data format heterogeneity: sensors, GPS locations, clicks, logs of servers, emails, images, voice, etc.
- Reutilization of data gather/constructed for other purposes and merge of databases from different sources
- Low level of signal to noise ratios
- Explanations do not pretend to be causal but merely predictive -> causality is irrelevant, only correlation matters: Amazon does not really care why costumers who bought a toaster also buy the book Redeployment… they only know that recommending that book works…

Big data

- What are the most relevant characteristics of big data for economics?
  - Reutilization
  - Merge of data from different sources

Big data

Multiples of bytes v·d·e				
SI decimal prefixes		Binary usage	IEC binary prefixes	
Name (Symbol)	Value		Name (Symbol)	Value
kilobyte (kB)	10^3	2^{10}	kibibyte (KiB)	2^{10}
megabyte (MB)	10^6	2^{20}	mebibyte (MiB)	2^{20}
gigabyte (GB)	10^9	2^{30}	gibibyte (GiB)	2^{30}
terabyte (TB)	10^{12}	2^{40}	tebibyte (TiB)	2^{40}
petabyte (PB)	10^{15}	2^{50}	pebibyte (PiB)	2^{50}
exabyte (EB)	10^{18}	2^{60}	exbibyte (EiB)	2^{60}
zettabyte (ZB)	10^{21}	2^{70}	zebibyte (ZiB)	2^{70}
yottabyte (YB)	10^{24}	2^{80}	yobibyte (YiB)	2^{80}

Processor or Virtual Storage	Disk Storage
· 1 Bit = Binary Digit	· 1 Bit = Binary Digit
· 8 Bits = 1 Byte	· 8 Bits = 1 Byte
· 1024 Bytes = 1 Kilobyte	· 1000 Bytes = 1 Kilobyte
· 1024 Kilobytes = 1 Megabyte	· 1000 Kilobytes = 1 Megabyte
· 1024 Megabytes = 1 Gigabyte	· 1000 Megabytes = 1 Gigabyte
· 1024 Gigabytes = 1 Terabyte	· 1000 Gigabytes = 1 Terabyte
· 1024 Terabytes = 1 Petabyte	· 1000 Terabytes = 1 Petabyte
· 1024 Petabytes = 1 Exabyte	· 1000 Petabytes = 1 Exabyte
· 1024 Exabytes = 1 Zettabyte	· 1000 Exabytes = 1 Zettabyte
· 1024 Zettabytes = 1 Yottabyte	· 1000 Zettabytes = 1 Yottabyte
· 1024 Yottabytes = 1 Brontobyte	· 1000 Yottabytes = 1 Brontobyte
· 1024 Brontobytes = 1 Geopbyte	· 1000 Brontobytes = 1 Geopbyte

barcelona gse
graduate school of economics

Big data misconceptions

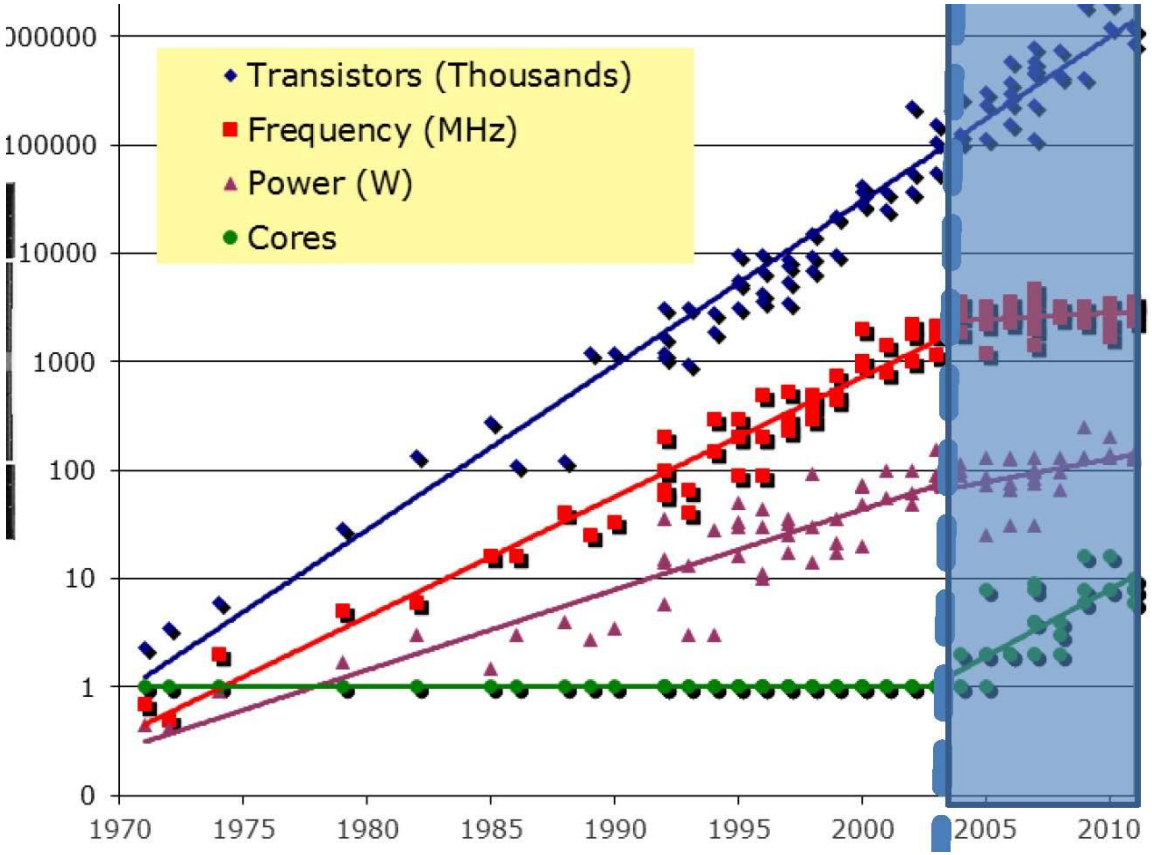
- Two questions:
  - What are the largest data generating devices/places?
  - What is the basic problem of big data infrastructure in our days?

Big data misconceptions

- Some basic facts, and misconceptions, about big data:
  - Currently the largest data producers are not internet or the NSA but sensors and large scientific infrastructures: the Large Hadron Collider of CERN produces 600 TB/sec using its 15 million sensors and, even after filtering of these data, it need 25 PB/year
  - Currently the basic problem is not computation power but the need of storage to accumulate the large amount of information produced and the energy needed to move information from processor to storage devices and back

Big data: the problem of storage

Moore's law + memory wall + energy wall



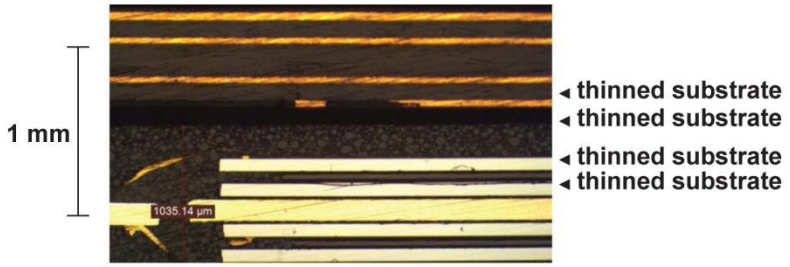
barcelona gse
graduate school of economics

Big data: the problem of storage

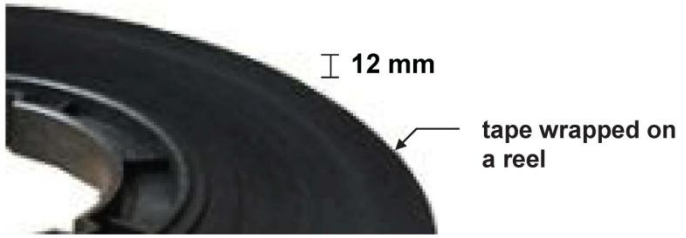
HDD (3 TB 3.5" Drive)
-- Areal Density 730 Gbit/in²
-- Media Density 2.4 Tb/in³
-- Component Density 126 GB/in³

NAND (0.5 TB 2.5" Form Factor Drive)
-- Areal Density 550 Gbit/in²
-- Media Density 6.7 Tb/in³
-- Component Density 121 GB/in³

NAND (0.5 TB Gum Stick Form Factor Drive)
-- Areal Density 550 Gbit/in²
-- Media Density 6.7 Tb/in³
-- Component Density 714 GB/in³



TAPE (1.5 TB LTO5 Cartridge)
-- Areal Density 1.2 Gbit/in²
-- Media Density 0.7Tb/in³
-- Component Density 106 GB/in³

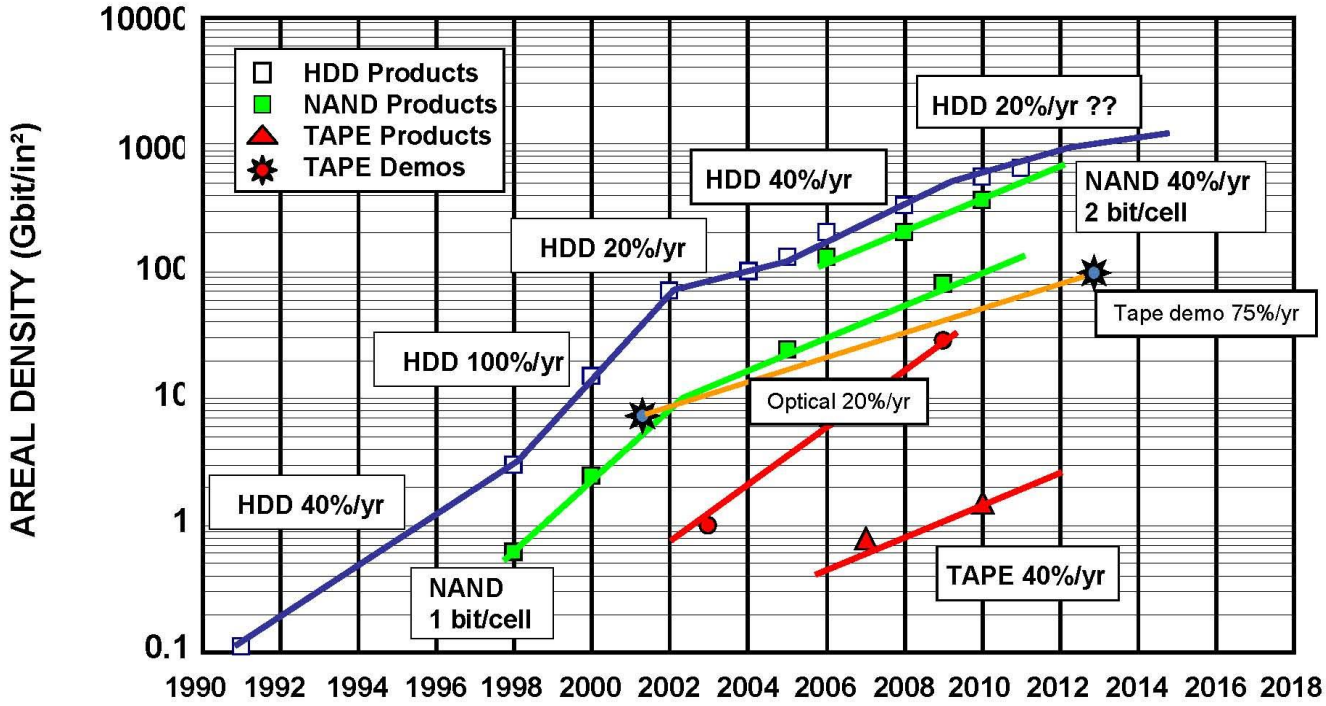


NAND: flash memory, memory card and solid state disks

barcelona gse
graduate school of economics

Big data: the problem of storage

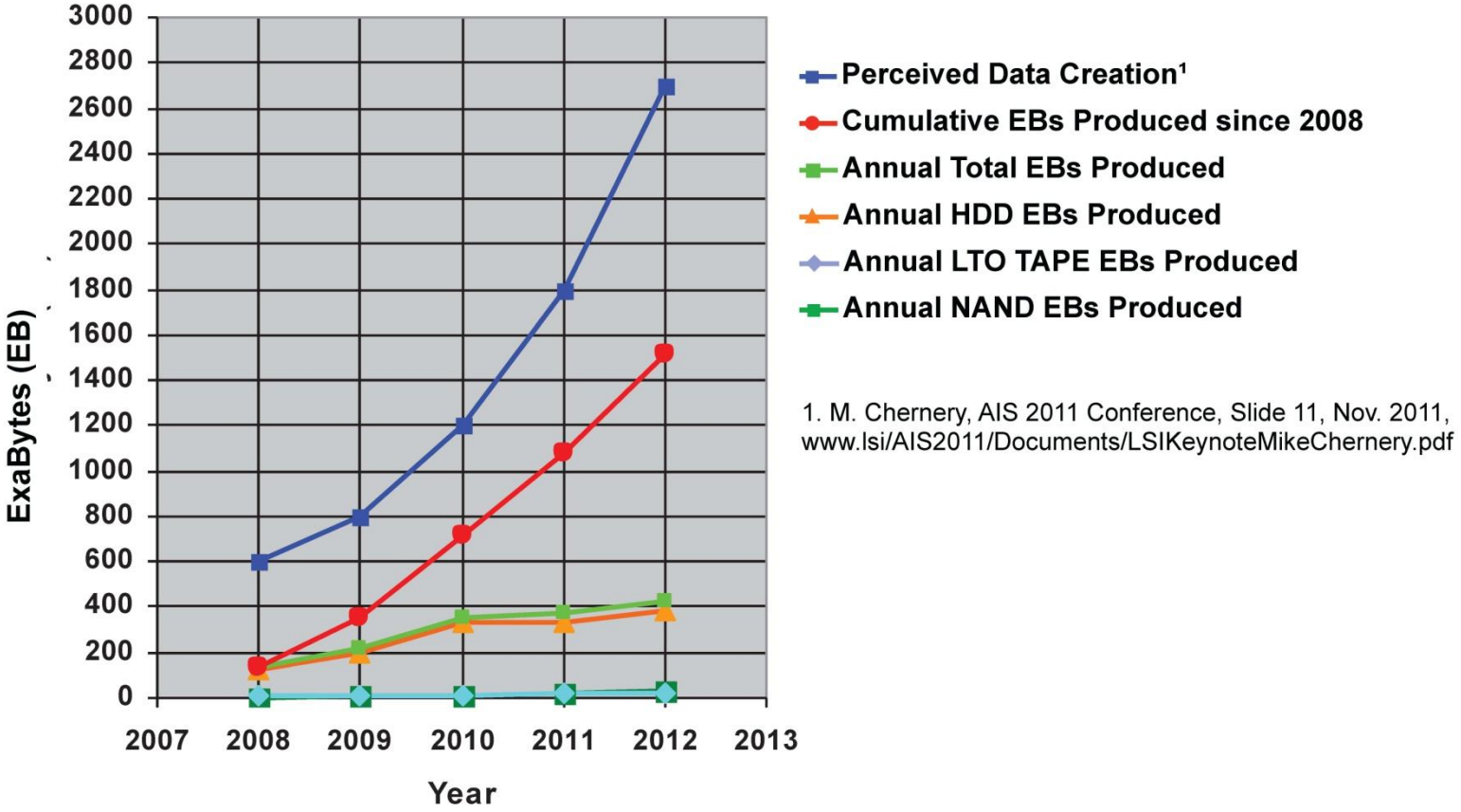
Petabyte demand for memory increases at a greater rate than areal density improvements (technology metric that allows to manufacture more bits per units of area)



Source: Decad, Fontana,Hetzler – IBM Journal

barcelona gse
graduate school of economics

Big data: the problem of storage



1. M. Chernery, AIS 2011 Conference, Slide 11, Nov. 2011, www.lsi/AIS2011/Documents/LSIKeynoteMikeChernery.pdf

barcelona gse
graduate school of economics

Big data

- More basic facts:
  - Need to move from a model centered around a computer to a model centered in the data with parallel massive computation (many cores and accelerators) and memory persistence – new architectures to mitigate the issue of heat and energy consumption (why bitcoin mining takes place in Iceland?)
  - Need to move to distributed computing (scalable and parallel computing) and new types of data generate the need for new tool that can work with non relational databases (non-SQL) like Map Reduce

Big data

- More basic facts:
  - New solutions in data science have reduced significantly the cost of complex processes like genome sequencing or micro-segmentation
  - Demand of statisticians, mathematicians and computer science graduates who know economics/business has increase drastically

Big data and econometrics

- Important differences between the econometrics before and after big data:
  - Supervised learning (inputs and outputs) uses classification methods, decision trees and neural networks when we use to call that regression
  - Unsupervised learning (only inputs) by shrinkage or dimensionality reduction while we use to call that non parametric estimation
  - Traditional econometrics stop explaining ridge regressions long time ago… to go back to it
  - Stata/Mata/Gauss versus Hadoop/MapReduce/Pig/Mahout /OpenRefine/ Hive/Hbase/ZooKeeper and R
  - Importance of over-fitting and cross validation

Big data and econometrics

- Important differences between the econometrics before big data and after (classical econometrics):
  - Large dimensionality: k>>n: need to regularize and use methods to reduce the dimensionality of the models like LASSO (least absolute shrinkage and Selection Operator any other that penalizes the number of parameters)
  - The path towards causality (using randomized experiments or natural experiments) has taken a detour towards machine learning techniques where prediction is emphasized over causality
  - Big data techniques are subject to the Lucas critique: social reputation and scoring: credit card fraud and testing of 99 cents operations

Big data and economics

- The availability of increasingly larger databases, in many cases geocoded, that merge information of diverse origin make economics a discipline more and more scientific:
  - The Billion Prices Project: real time estimation of the evolution of prices using millions of prices of on-line business. The project shows that official inflation is quite similar to on-line calculations for countries like Brazil, Chile, Venezuela or Colombia but not for Argentina (accumulated difference between 2007 and 2011 of 65%)
  - What can we do with information on 24 million credits?
  - The STAR experiment and its current effects
  - Choi and Varian (2014): use big data techniques to improve prediction models – Example: AR(1) model for weekly unemployment subsidies using as transfer function Google Trends for words like jobs, welfare or unemployment

Big data and economics

- The Billion Prices Project (MIT)
  - They use the stability or change in the HTML tags used to construct the web pages of online department stores to identify changes of prices over time
  - Using this principle their software identifies the information relevant for the product and its price
  - The URL of the page that indexes the products is used to classify them by categories

Big data and economics: house prices

- House prices in Spain:
  - Appraisals
  - Ask prices (they can be found crawling Internet)
  - Closing prices
  - “Official prices”
  - Registry prices
  - Notary prices

Big data and economics: house prices

- Montalvo and Raya (2012), "Imaginary prices…" Increasing the appraisal price to produce a mortgage:
  - Finally I got data to show what it was theoretically obvious
  - Appraisal values set in function of the financial needs of the clients and not to represent the value of the house

Big data and economics: house prices

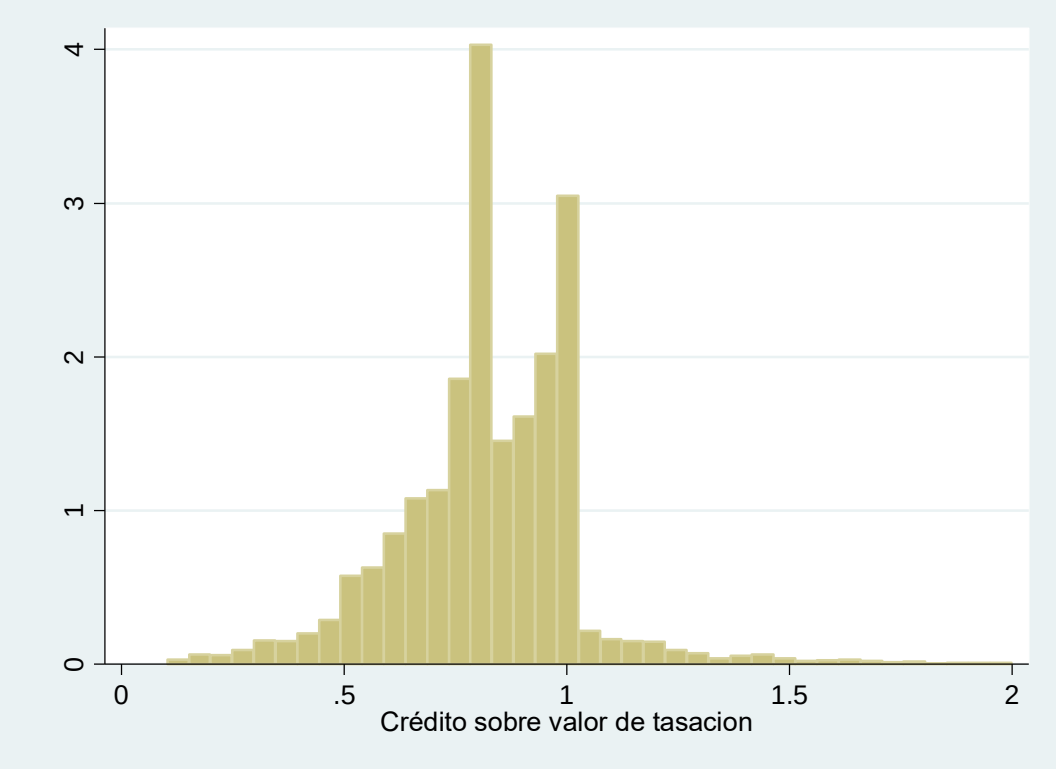
- Montalvo and Raya (2012), "Imaginary prices…" merge four datasets from very different sources:
  - Housing intermediary: market prices
  - Financial institution: appraisal prices, amount of the mortgage
  - Ask prices: robots in Internet
  - Official Registry of Real Estate Properties: amount of mortgage, registry price (reported in the official ownership document)
  - General Directory of Real Estate Properties: unique id number

Big data and economics: house prices

- The damaging role of appraisal firms and its conflict of interest with banks and savings and loans
- Equity withdrawals very limited in Spain: well, not really… you get it upfront
- The condescending view of the regulator: countercyclical buffer, no securitization, strict regulation of out of balance operations, large consolidation perimeter… but huge problems of incentives of appraisal companies and banks

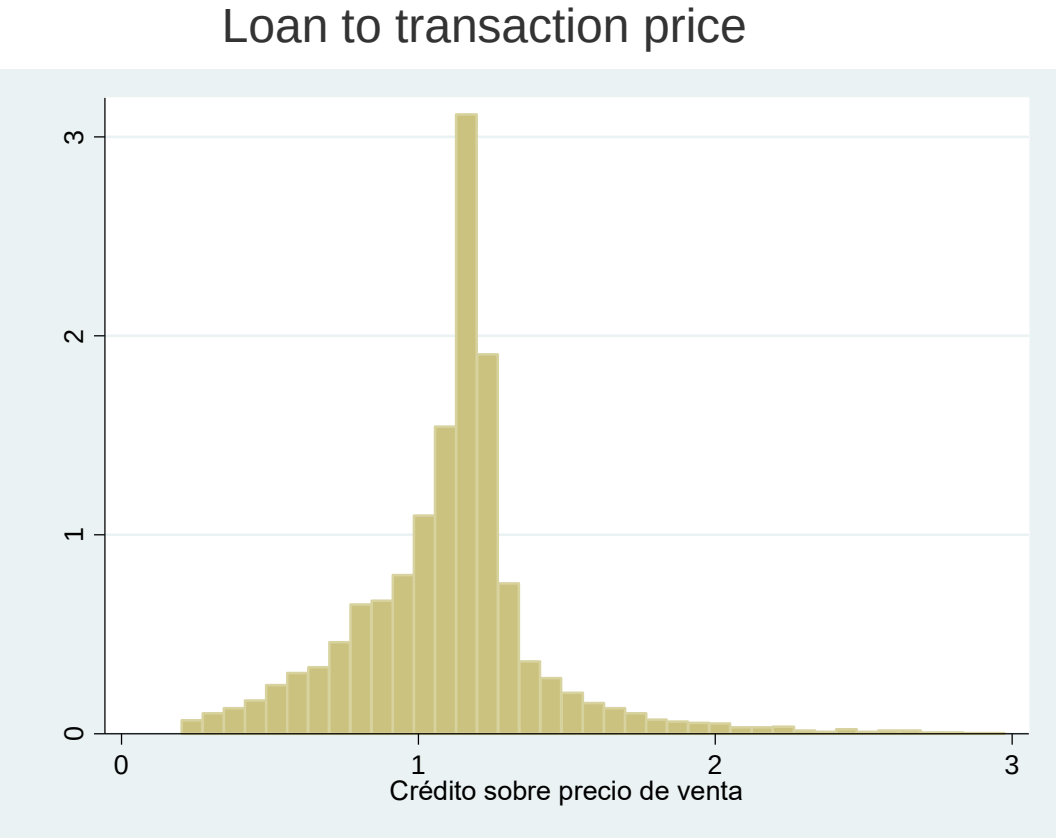
Big data and economics: house prices

Loan to official value (appraisal)



barcelona gse
graduate school of economics

Big data and economics: house prices



barcelona gse
graduate school of economics

Big data and economics: house prices

barcelona gse
graduate school of economics

Big data and economics: geo-diversity

- What is the effect of ethnic diversity on economic growth and development? slides
- What's the impact of infrastructures on firm location? Paper

barcelona gse
graduate school of economics

Big data and economics: millions of loans

- Did banks reduce their standards to originate credits? Is it better more capital or more women as risk officers? Paper

Big data and economics: predicting elections

- How to predict a difficult election: from simple time series baseline models to complex integrated models with census post-stratification and Bayesian logit models updated using thousands of polls. Paper

Big data and marketing

- Changes in the buyers journey: need to understand what the buyer does 60 to 80% of the time before contacting a company representative
- Digital body language: visits to the web, time spend, emails interactions, interaction with social media, behavior after watching a video, banner or report; searching before and after; etc.
- Chief Marketing Officer (CMO) and Chief Information Officer (CIO) have to be very close and also in constant interaction with the sales department

Big data and marketing

- Technologies for marketing
  - MAS: marketing automation software (ExactTarget, Marketo, Eloqua, etc.)
  - Business Inteligence Databases (IBM, Oracle, SAP)
  - CRM: Customer Relationship Management (SAP, NetSuite, Salesforce, Oracle, etc.)
  - CMS: Content management system (Abode, OpenText, Oracle, etc.)
  - Platforms blog: WordPress, Moveable type, etc.
  - DMP: Data management platforms (Abode, BlueKai, CoreAudience, Krux, Lotame,etc.)
  - Analytic tools: web analytics, charbeat, google analytics, mint, etc.
  - SMM: social media management software (Abode social, buddy Media, Web trends, HootSuite, etc.)

Big data and marketing

- Technologies for marketing:
  - Predictive lead scoring vendors: FlipTop, Infer, KXEN (SAP), Lattice Engines (all of them use big data techniques)
  - Call- center software
  - SEM platforms (Search Engine Management Platforms) – to manage, automatize and optimize marketing in search pages and pay-for-click campaigns
  - DPS: Demand Side Platforms – allows the marketing department and its agencies to manage in real time the auctions for advertising (RTB: real time bidding) simultaneously in several online advertising markets

Big data and financial services

- After what has happened during the financial crisis financial institutions need to gain their lost reputation
- The EBA, IMF, ECB, etc. insist that the European Banking sector has an important profitability problem (low ROEs) -> need new business model
- Increasing competition from new non-bank actors in the financial intermediation business is eroding large parts of the value chain of banking products

1. Recovering clients' trust

- Could the banking industry do like Amazon and recommend individualized products to its clients? -> banks swim in a huge sea of very relevant data which opens the door to adapt products to the needs of each client (instead of inventing product that then they promote across all types of clients)
- Objective: improve access of families of low income to financial product at a cost that is reasonable for their income profile, ability to pay and risk aversion of the costumers

1. Recovering clients' trust

- In many countries, including the US, there is a high proportion of clients that, either for having no credit history or a short credit history, cannot access to banking services
- In the US these potential costumers end up in payday loan services paying very high interest rates and having a low maturity product

2. New business models

- The reduced profitability of banking, the increase in regulation and the high level of leverage of the economy requires efficiency improvement in the banking sector
- The future of their business model can be based on big data (large databases created by banks): reducing inefficiencies, increasing products costumization and costumers satisfaction using analytics and big data
- Big data is also critical to confront increasing demand of information for regulation purposes

3. How to confront new competitors

- The financial disintermediation is affecting the profits of financial institutions
- Until recently these competitors attacked basically the payment instruments chain link (cryptocurrencies, mobile payments, complementary currencies, etc.) but the are moving fast to other parts of the value chain (peer to peer loans, personal loans, etc.)

Big data and banks: applications

- Basic applications
  - Optimization of the relationship with clients
  - Improvement in the financial functions
  - Risk reduction
  - Compliance with new regulations

1. Big data and banking: scoring

- FICO (Fair Isaac Corporation) and internal models
- Behavioral models for long time customers and concessional models (based on demographics and few more observation) for recent customers or even no clients

FICO

- Factors to calculate the credit score (the exact formula is a secret). Approximate weights:
  - 35% — Payment History – Late payments on bills, such as a mortgage, credit card or automobile loan, can cause a consumer's FICO score to drop.
  - 30% — Credit Utilization – The ratio of current revolving debt (such as credit card balances) to the total available revolving credit (credit limits).
  - 15% — Length of Credit History – As consumer's credit history ages, assuming they pay their bills, it can have a positive impact on their FICO score.
  - 10% — Types of Credit Used (installment, revolving, consumer finance) – Consumers can benefit by having a history of managing different types of credit.
  - 10% — Recent search for credit and/or amount of credit obtained recently – Multiple credit inquiries for a consumer seeking to open new credit, such as credit cards, retail store accounts, and personal loans, can hurt an individual's score.

1. Big data and banking: scoring

- Testimony of J. P. Morgan before the House of Representative (1912): "The most important factor to ge a credit is not wealth but reputation. A man who is not trustable would not obtain from me any money eve if he owns all the wealth of Christianity
- Reputation -> social reputation -> social credit score -> social status online
- Basic sources to measure social reputation: Facebook, Twitter, LinkedIn, etc.

1. Big data and banking: scoring examples

- Neo Finance (Palo Alto) specializes in loan targeted to young people who want to buy a car but do not have a lengthy credit history -> this would imply to pay very high interest rates:
  - Neo Finance uses the number and quality of the connections of the loan applicant in LinkedIn. They look specially for links to workers in the same company to predict stability of job in the future and income
  - They also use the contacts in other companies to estimate the probability of finding a job conditional on being fired of her company to estimate the time to finding a job after being dismissed -> objective estimate job stability

1. Big data and banking: scoring examples

- Keditech
  - They use the location of the residence of friends and their jobs to calculate a credit score. If friends have delinquent credits this reduces the probability of getting a credit approved.
  - The algorithm generates a decision in 8 seconds (average) and produces a delinquency rate less than 10%
- Lenddo generates a social capital online score between 0 and 1000 using the number of followers in Facebook, their characteristics (demographics, residence, jobs, etc.), their education degrees, their employer and credit history, and the credit history of their friends. Therefore if a friend stops paying bills or loans my score is affected -> similar idea to microcredits in developing countries

1. Big data and banking: scoring examples

- Experian (Extended View) y Equifax (Vantage Score) are open to use social reputation information coming from Internet.
- FICO has shown no interest in this information
- Some analyst call these new techniques to calculate the score as the reinvention of the wheel of credit scores: "it's the Wild West like the early days of FICO" Pete Fader
- Banks have high quality costumers' data and experience that it is difficult that can be beaten by social reputation indices (very noisy) at least currently

1. Big data and banking

- BBVA and credit card terminals: http://mwcimpact.com/
- Importance for new business and credit
- MasterCard and Spending Pulse: real time data on consumption in different commercial activities
- VISA produces high frequency predictions using economic surveys
- Moody's Analytics forecasts each month the employment in the private sector using information on 500.000 companies that use their payroll software

1. Big data and banking: examples scoring

- Khandani et al (2012) increase the number of variables on transactions of the clientes and credit scores generated by agencies (Experia, etc.) -> improves forecast reaching 85% of right predictions and 6-25% saving on total losses

2. Big data and banking: credit cards

- Reason for big data used to detect credit card fraud is simple: it saves millions of euros to a bank – it takes advantage of one of the V (velocity) of big data
- Size of data is huge: employers data, applications to jobs, loans, etc., death lists, incarcerated, black lists, IRS, etc. as well as patters that could be used to analyze the geographical location of payments, characteristics of the business and similar businesses, etc.

2. Big data and banking: credit cards

- Four approaches:
  - Based on rules (known patterns)
  - Detection of anomalies or outliers (unknown patterns)
  - Predictive analysis searching for complex patterns
  - Hybrid models

3. Other utilities

- Peer to peer lending: RateSetter, Zopa, Lending Club, etc.
- Platforms like Social finance, CommonBond o Upstart for students debt -> applicants accept that those platforms get any information needed to score them using data of employers or online social networks
- Car insurance and the rise of sensors

The dangers of big data

- Big data provides very useful tools to manage business in an uncertain environment with increasing regulation and mistrust from costumers in the banking industry -> however, it is not sure that any big data project will generate a successful strategy:
  - Decreasing returns to the accumulation of information
  - Data are not informative if they are not properly analyzed
  - Need to calculate the cost benefit of any big data project (return on investment, ROI)

The dangers of big data

- A huge amount of data cannot overcome the foundations of statistics, the influence of measurement errors or the dangers of spurious correlations
- You need technical knowledge but also be open to evaluate constantly the predictive ability of your model and adjust it if there is lost of precision: the experience of Google Trends with Google Flu Trends is a good precautionary tale

Carefull with algorithms!

- Small area flu forecast (using "Google Flu Trends"): big data and algorithms have limitations -> in the last three years the models have over-estimated the flue by 50%
- Princeton versus Facebook
  - Princeton study: Facebook will loose 80% of its users (epidemic model on the number of times that the word "Facebook" was searched in Google) and analogy with MySpace
  - Facebook technologists strike back: "Using the same principles that correlation implies causality, and the same analogy of Princeton's paper we find that the university of Princeton will disappear soon"

The dangers of big data

- Data privacy and reutilization: new consent clauses
- The possibility that mistakes in the capture, merge or cleaning of data may generate negative effects for citizens by the application of big data to specific problems. For example, what happen if a company calculates a wrong credit score for a citizen using big data techniques and, that leads to the denial of a credit? -> NCLC (2014) (National Consumer Law Center)

The dangers of big data

- The issue of anonymized data -> several papers have shown how to find out the name of the anonymous person to whom the data refer to -> recurrent topic in the latest meeting of the American Statistical Association

UNIVERSITAT POMPEU FABRA