HEC PARIS

The more you know, the more you dare®

LAW IN THE AGE OF BIG DATA

David Restrepo A
Law and Tax Department
HEC PARIS

Cristina Golomoz
Centre for Socio-Legal Studies
UNIVERSITY OF OXFORD

affiliated to
CCI PARIS ILE-DE-FRANCE

WHY BIG DATA MATTERS?

$300 billion
potential annual value to US health care—more than double the total annual health care spending in Spain

$600 billion
potential annual consumer surplus from using personal location data globally

€250 billion
potential annual value to Europe’s public sector administration—more than GDP of Greece

60% potential increase in retailers’ operating margins possible with big data

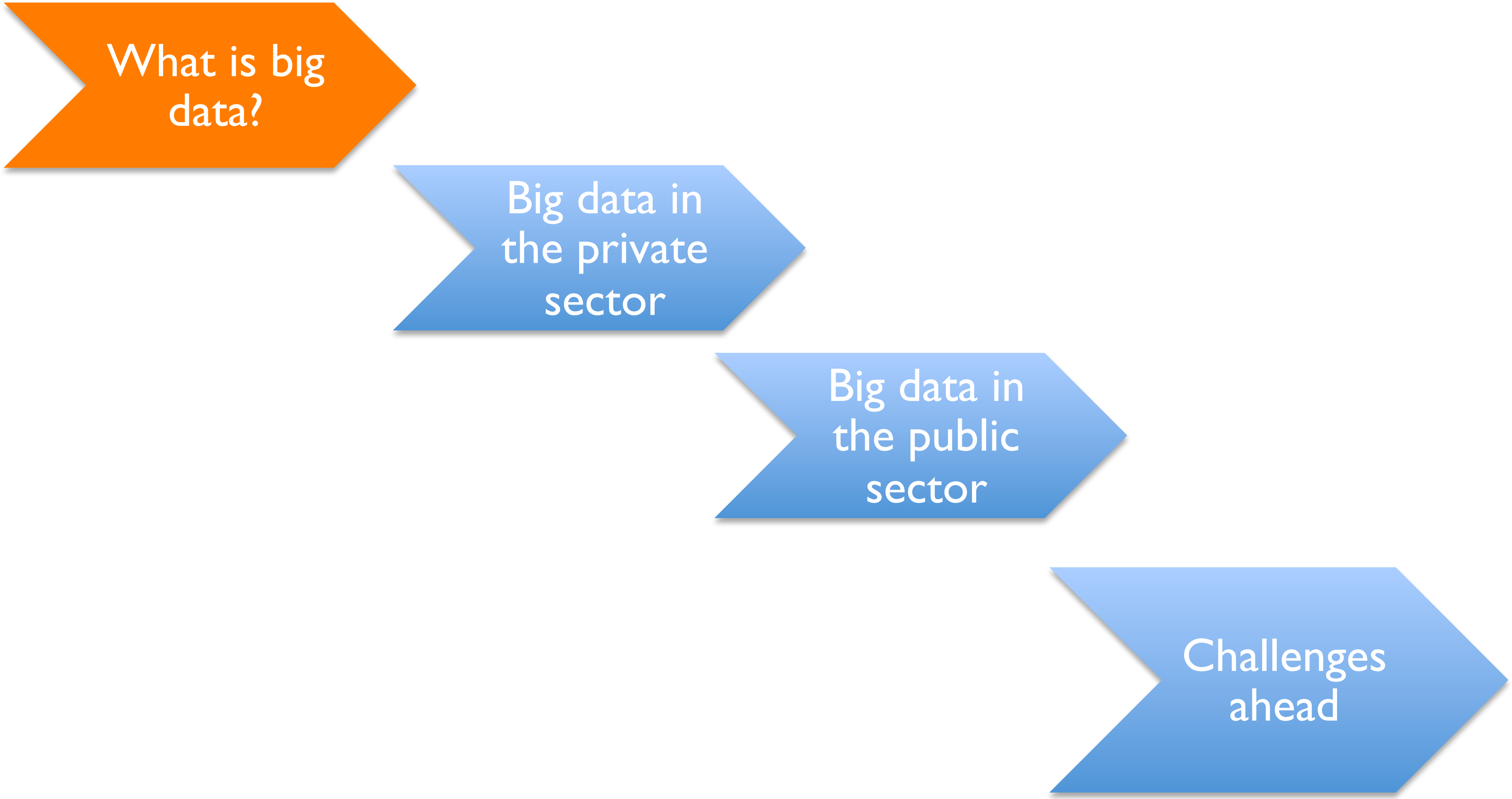
140,000–190,000
more deep analytical talent positions, and

1.5 million
more data-savvy managers needed to take full advantage of big data in the United States

McKinsey Global Institute

McKinsey&Company

HEC
PARIS



What is big data?

Big data in the private sector

Big data in the public sector

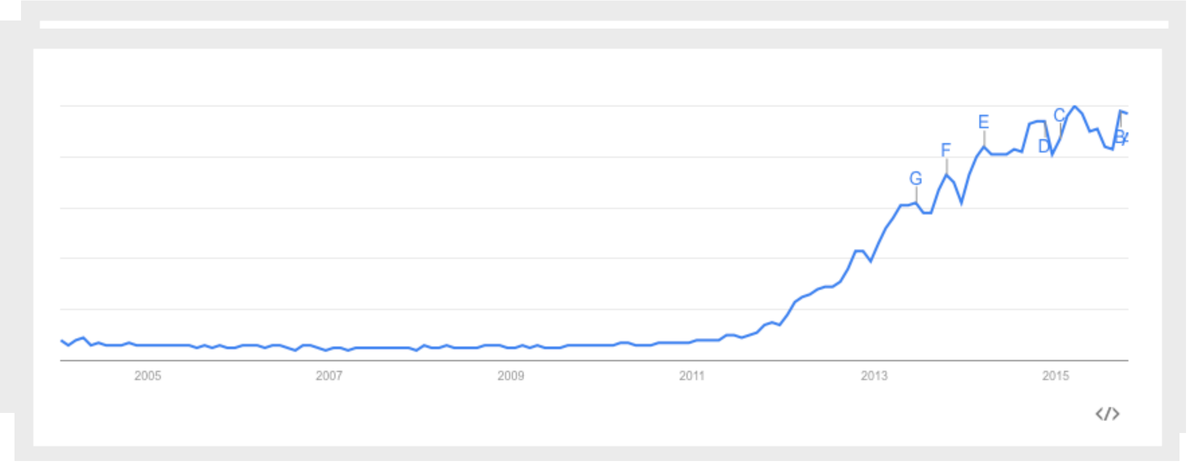
Challenges ahead

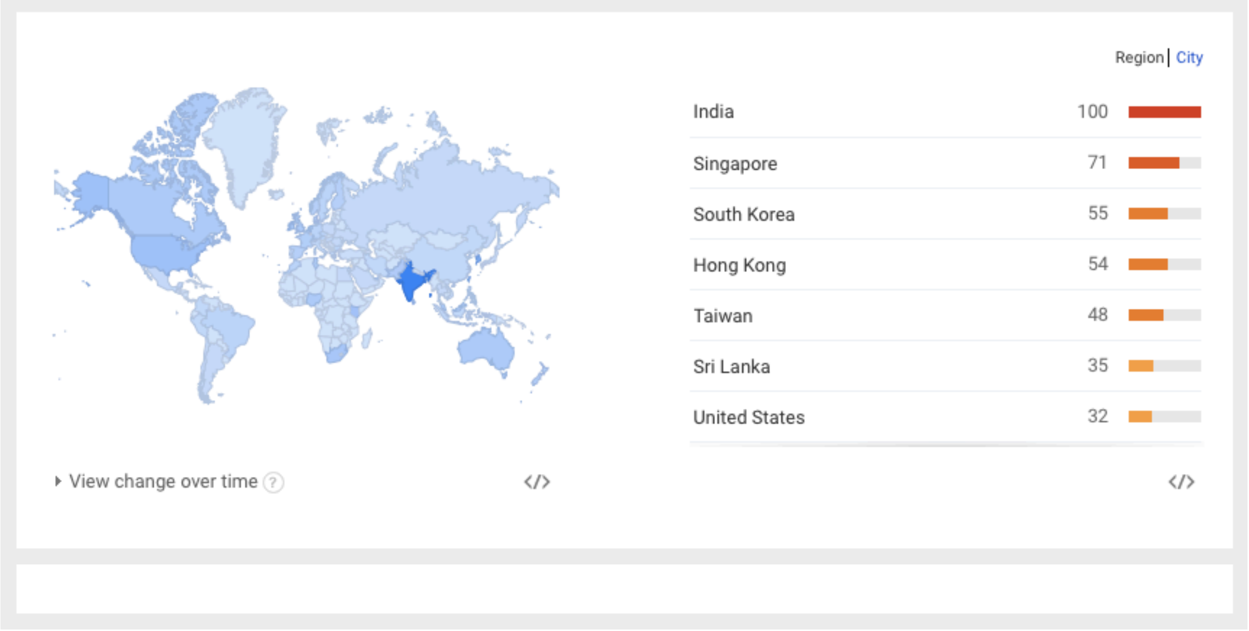
HEC PARIS

INTRODUCTION

WHAT IS BIG DATA?

GOOGLE TREND
"BIG DATA"





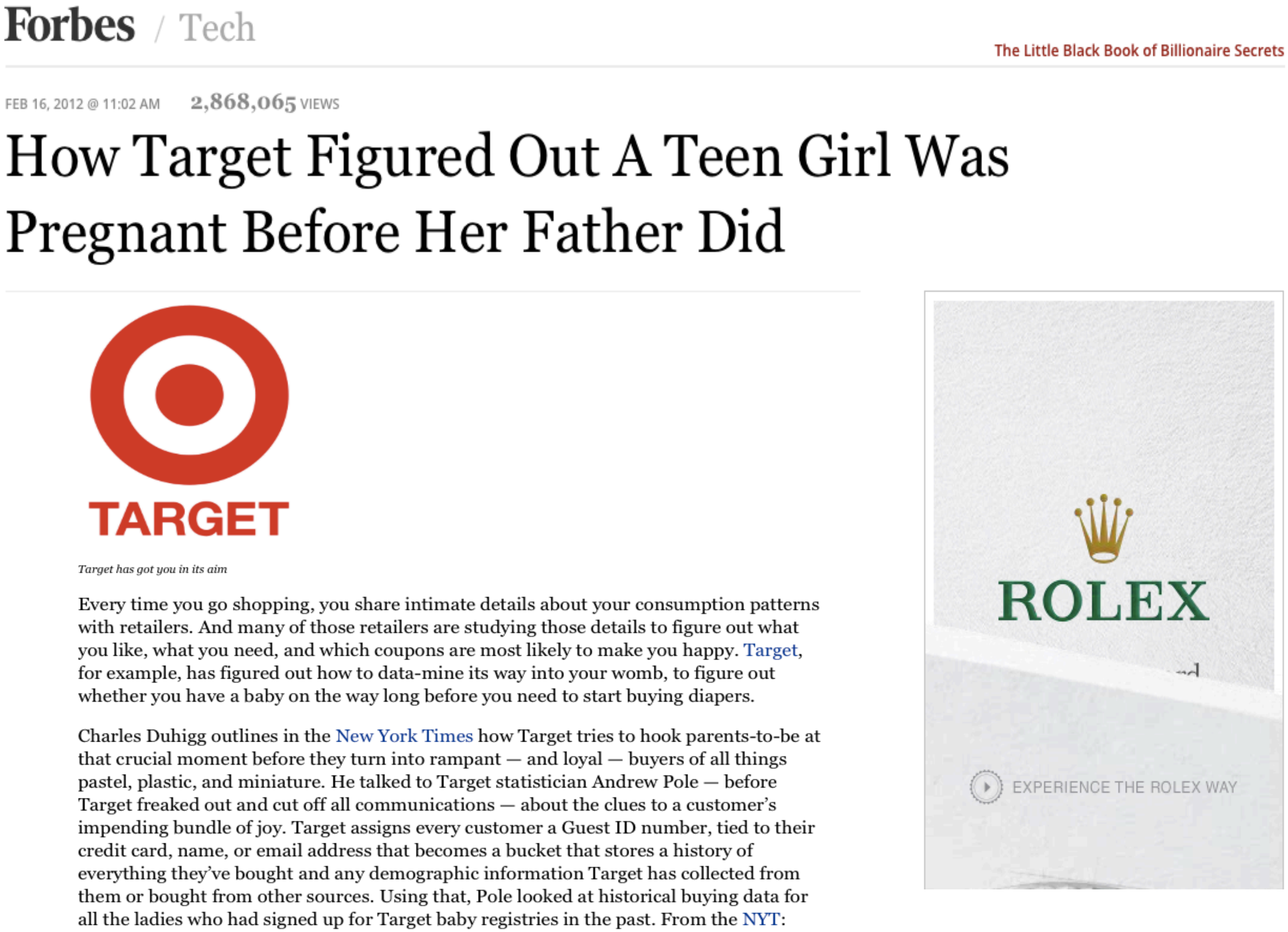
Region | City

India	100	
Singapore	71	
South Korea	55	
Hong Kong	54	
Taiwan	48	
Sri Lanka	35	
United States	32	

BIG DATA IS OUT THERE!

BIG DATA
USING SMART BIG DATA ANALYTICS AND METRICS TO MAKE BETTER DECISIONS AND IMPROVE PERFORMANCE
BERNARD MARR
'Data and analytics power everything that we do. This book is the go-to-guide on data for 2015.' – Henrik von Scheel, Advisory Board Member at Google
WILEY
J.-C. Cointot
Y. Eychenne
La Révolution
BIG DATA
A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK
VIKTOR MAYER-SCHÖNBERGER
KENNETH CUKIER
BIG DATA
A Revolution That Will Transform How We Live, Work, and Think
BIG DATA
La révolution des données est en marche
Viktor Mayer-Schönberger
Kenneth Cukier
Robert Laffont

TARGET & PURCHASING HABITS



Forbes / Tech

FEB 16, 2012 @ 11:02 AM 2,868,065 VIEWS

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Target has got you in its aim

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

Charles Duhigg outlines in the New York Times how Target tries to hook parents-to-be at that crucial moment before they turn into rampant — and loyal — buyers of all things pastel, plastic, and miniature. He talked to Target statistician Andrew Pole — before Target freaked out and cut off all communications — about the clues to a customer's impending bundle of joy. Target assigns every customer a Guest ID number, tied to their credit card, name, or email address that becomes a bucket that stores a history of everything they've bought and any demographic information Target has collected from them or bought from other sources. Using that, Pole looked at historical buying data for all the ladies who had signed up for Target baby registries in the past. From the NYT:

TAKING ABOUT BIG DATA…

BIG (terabytes – 1024 GB)

- Is not about the size; it about what you do with it (sample size is not important here!)

Data

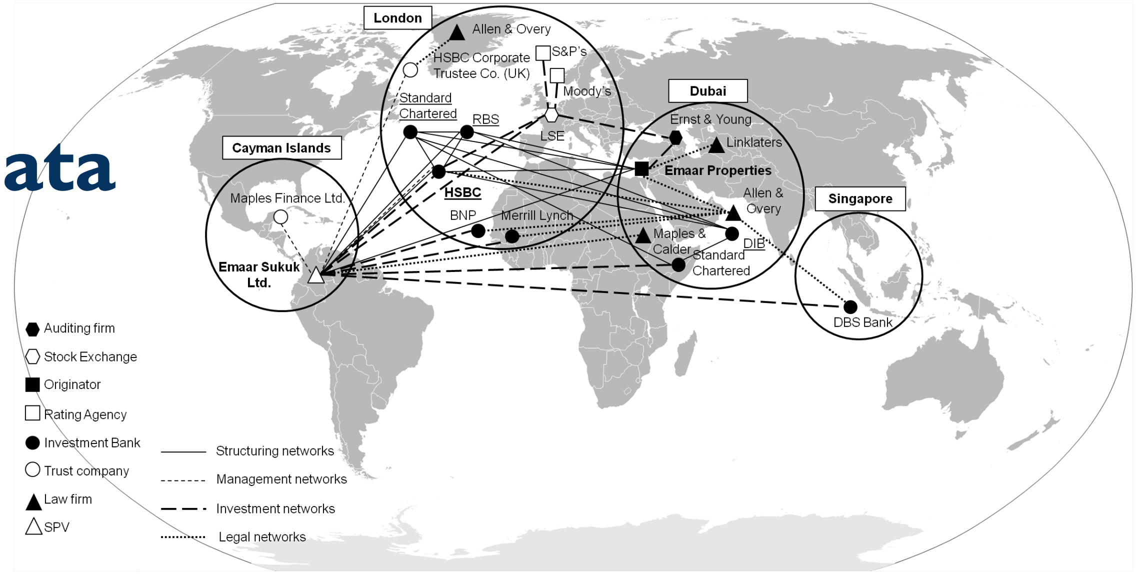
- Structured (ex. Sale records, cash machines)
- Unstructured (Facebook, pictures, videos, GPS locations etc.)

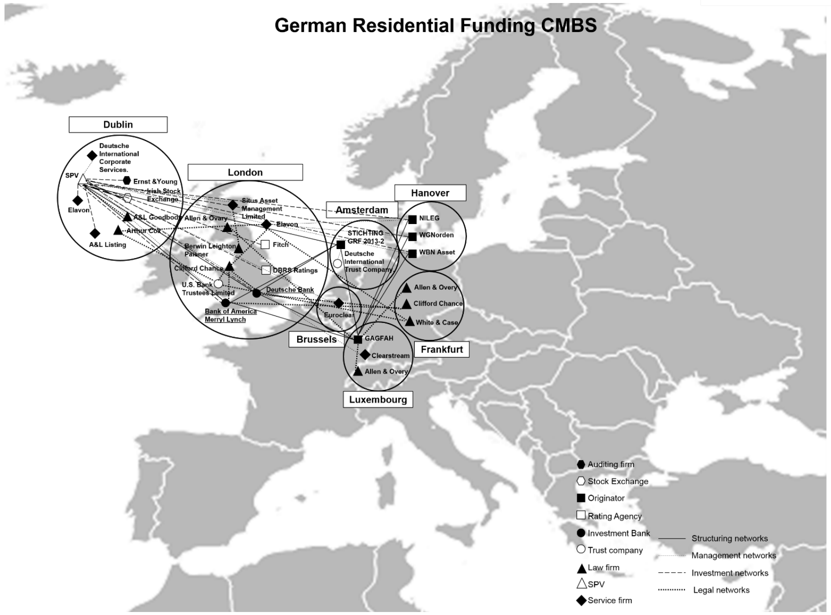
STRUCTURED LEGAL DATA

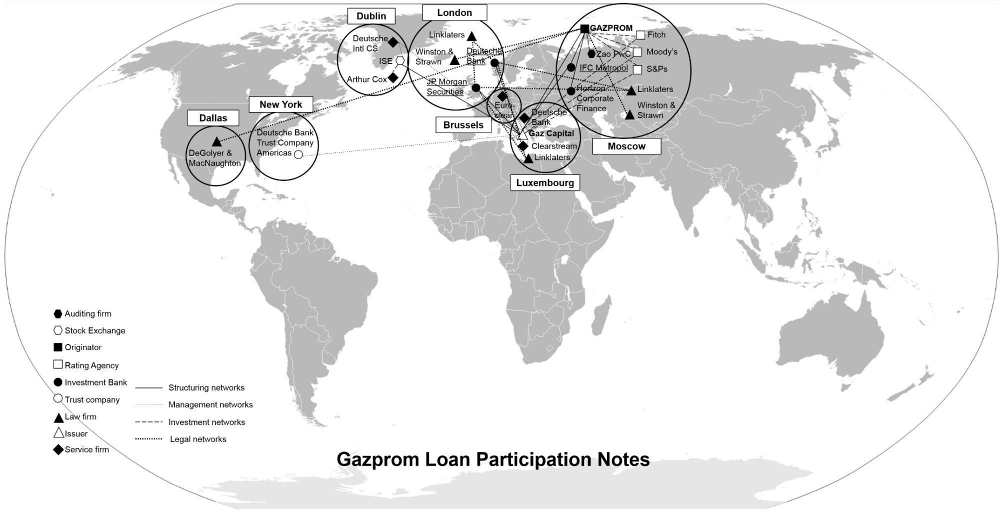
Indicator	Issuer(s)	Year of creation	Countries covered
WGI Rule of Law Indicator	World Bank	1996	215
CPIA Property Rights and Rule-based Governance Rating	World Bank	1977	81
Doing Business	World Bank	2004	189
Rule of Law Index	World Justice Project	2010	99
Global Rights Index	International Trade Union Confederation	2014	139
Realization of Children's Rights Index	Humanium	2011	190
Sub-indicator “Institutions”- Global Competitiveness Index	World Economic Forum	2005	144
Freedom in the World	Freedom House	1972	195
Freedom of the Press	Freedom House	1980	197
World Press Freedom Index	Reporters Sans Frontières	2002	180
Bertelsmann Transformation Index –Rule of Law	Bertelsmann Foundation	2003	129
Bertelsmann Transformation Index –Property Rights	Bertelsmann Foundation	2003	129
Index of Economic Freedom - Property Rights	Heritage Foundation	1995	178
Global Integrity Index – Anticorruption & Rule of Law	Global Integrity	2004	100
CIRI Human Rights Data - Freedom of Speech	David L. Cingranelli, David L. Richards, K. Chad Clay	1981	202
CIRI Human Rights Data - Independence of the Judiciary			
Democracy Index	Economist Intelligence Unit	2006	167
Global Business Rule of Law Dashboard	U.S. Chamber of Commerce	2013	80
S&P's Sovereign Credit Rating – Political Score	Standard & Poor	-	129
Financial Secrecy Index	Tax Justice Network	2009	82
Investment Across Borders	World Bank	2010	87
Sustainable Governance Indicators - Democracy	Bertelsmann Foundation	2009	41

Structured legal data

Law firms & in-house counsel







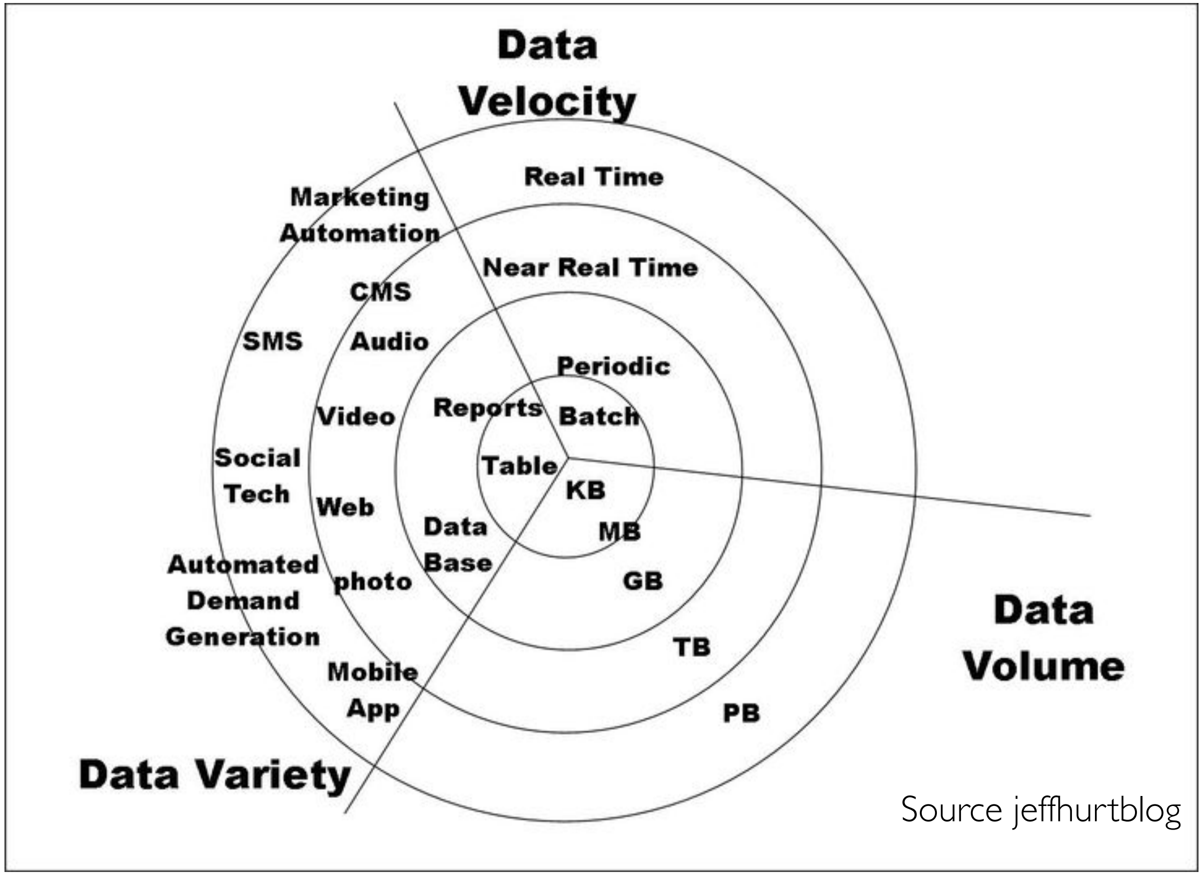
David Bassens & Marion Berzin - Vrije Universiteit Brussel

Cartography of global finance space

STRUCTURED LEGAL DATA

	Incorporation		Financing			Choice of Jurisdiction			Bankrp.
	WGI- S	DB-R	FSI-S	GCI-S	DB-R	DB-R	RLI-S	GCI-S	DB
France	1,40	35	41	4,7	80	7	0,68	4,4	44
Germany	1,62	19	59	4,3	71	5	80	4,9	19
England & Wales	1,67	11	40	5	10	57	72	5,4	8
Netherl.	1,81	30	50	4,4	113	30	80	5,6	6
Belgium	1,40	32	45	4,1	16	16	0,68	4,2	7

3Vs OF BIG DATA

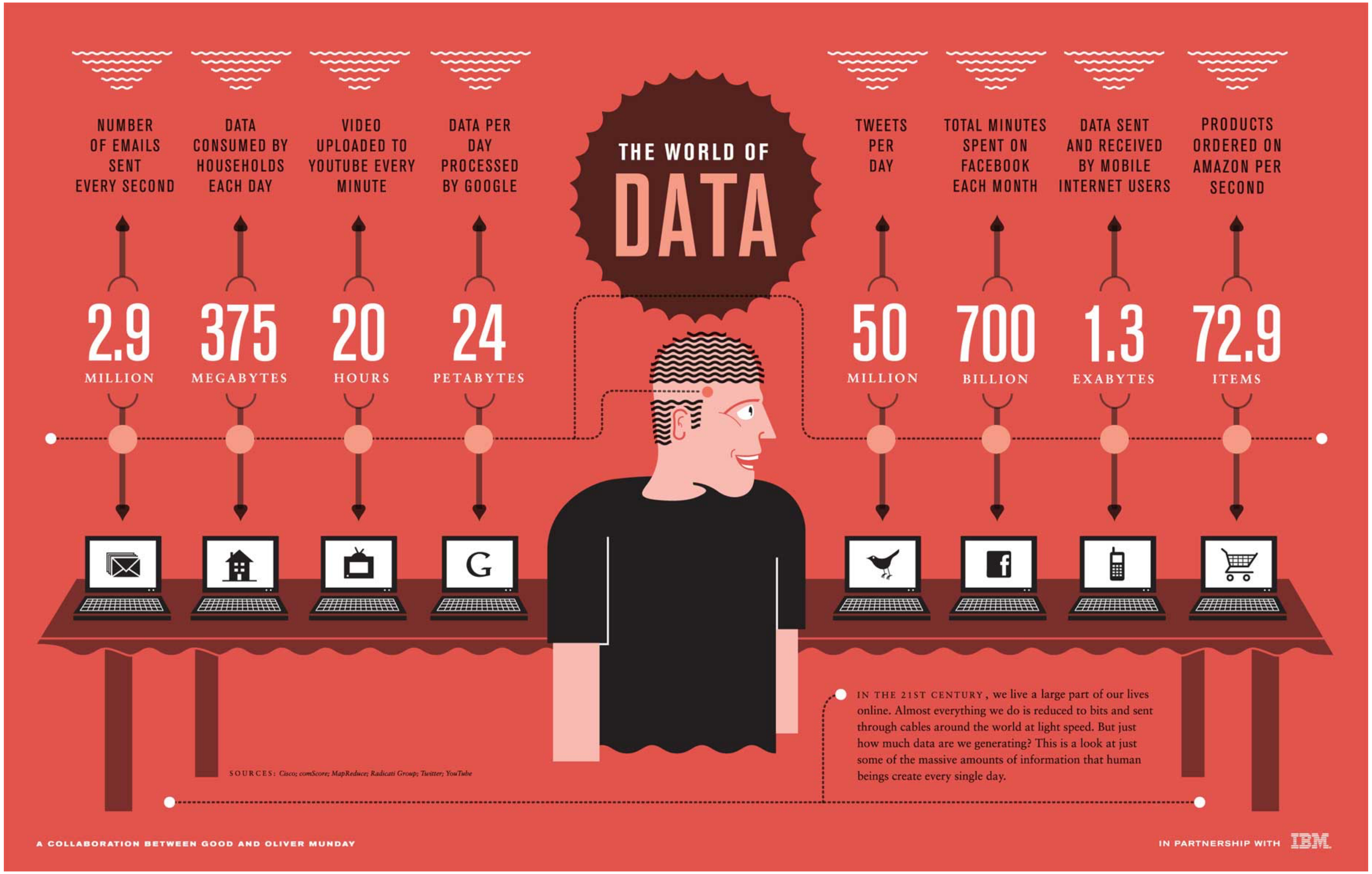


Source jeffhurtblog

“Big data are high volume, high velocity, and / or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization.”

Svetlana Sicular Gartner

(I) DATA VOLUME


THE WORLD OF DATA
NUMBER OF EMAILS SENT EVERY SECOND
2.9 MILLION
DATA CONSUMED BY HOUSEHOLDS EACH DAY
375 MEGABYTES
VIDEO UPLOADED TO YOUTUBE EVERY MINUTE
20 HOURS
DATA PER DAY PROCESSED BY GOOGLE
24 PETABYTES
TWEETS PER DAY
50 MILLION
TOTAL MINUTES SPENT ON FACEBOOK EACH MONTH
700 BILLION
DATA SENT AND RECEIVED BY MOBILE INTERNET USERS
1.3 EXABYTES
PRODUCTS ORDERED ON AMAZON PER SECOND
72.9 ITEMS
G

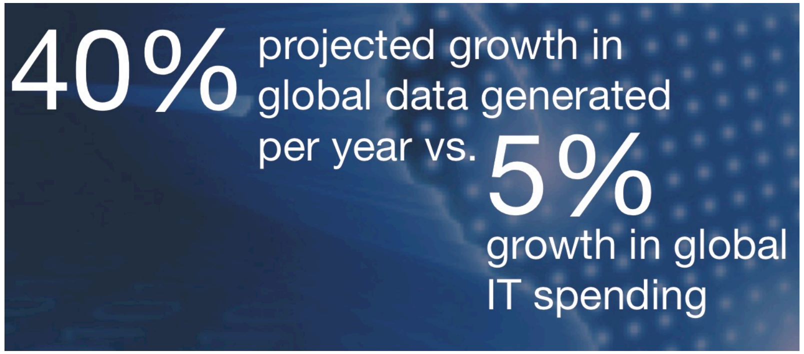
IN THE 21ST CENTURY, we live a large part of our lives online. Almost everything we do is reduced to bits and sent through cables around the world at light speed. But just how much data are we generating? This is a look at just some of the massive amounts of information that human beings create every single day.

SOURCES: *Cisco; comScore; MapReduce; Radicati Group; Twitter; YouTube*

A COLLABORATION BETWEEN GOOD AND OLIVER MUNDAY

IN PARTNERSHIP WITH IBM

(I) DATA VOLUME



40% projected growth in global data generated per year vs. 5% growth in global IT spending

235 terabytes data collected by the US Library of Congress by April 2011

15 out of 17 sectors in the United States have more data stored per company than the US Library of Congress

Name	Symbol	Equal to
Kilobyte	KB	1024 B
Megabyte	MB	1024 KB
Gigabyte	GB	1024 MB
Terabyte	TB	1024 GB
Peabyte	PB	1024 TB
Exabyte	EB	1024 PB
Zettabyte	ZB	1024 EB
Yottabyte	YB	1024 ZB

Humankind has stored more than 295 billion gigabytes (or 295 exabytes) of data since 1986

University of Southern California 2011

(I) DATA VOLUME

Consider that 90% of the world's data has been produced in just the last two years.This explosion of information is known as “Big Data,”

Peter Ebbs – HEC Paris

2015 - 8000 Exabytes – 500 billion 16GB Ipads

Computer that put the first man on the moon - 64kbytes

“Storage is not the problem but how we can extract information from it”

(2) DATA VARIETY

STRUCTURED DATA

+

UNSTRUCTURED DATA (INTERACTION)

CHARACTERIZING DATA

- Pressure to collect more data
  - Traditional data has a life spam
  - Company customer data vs. big data giants
- **Anti-depression** *(68 min v. 17 min)*

Shopping/medicine/ etc. through GPS phone location and usage.

- **Manufacturers**

"All of a sudden, we have a whole new way of making money that doesn't rest on a customer throwing something out and buying new; you can fix it before it fails and get paid for that."

Michael Porter Harvard Business School (WSJ)

CAESARS ENTERTAINMENT CORP

Analyzes health-insurance claim data for its 65,000 employees and their covered family members.

- How employees use medical services?
- Number of emergency-room visits?
- Whether they choose a generic or brand-name drug?

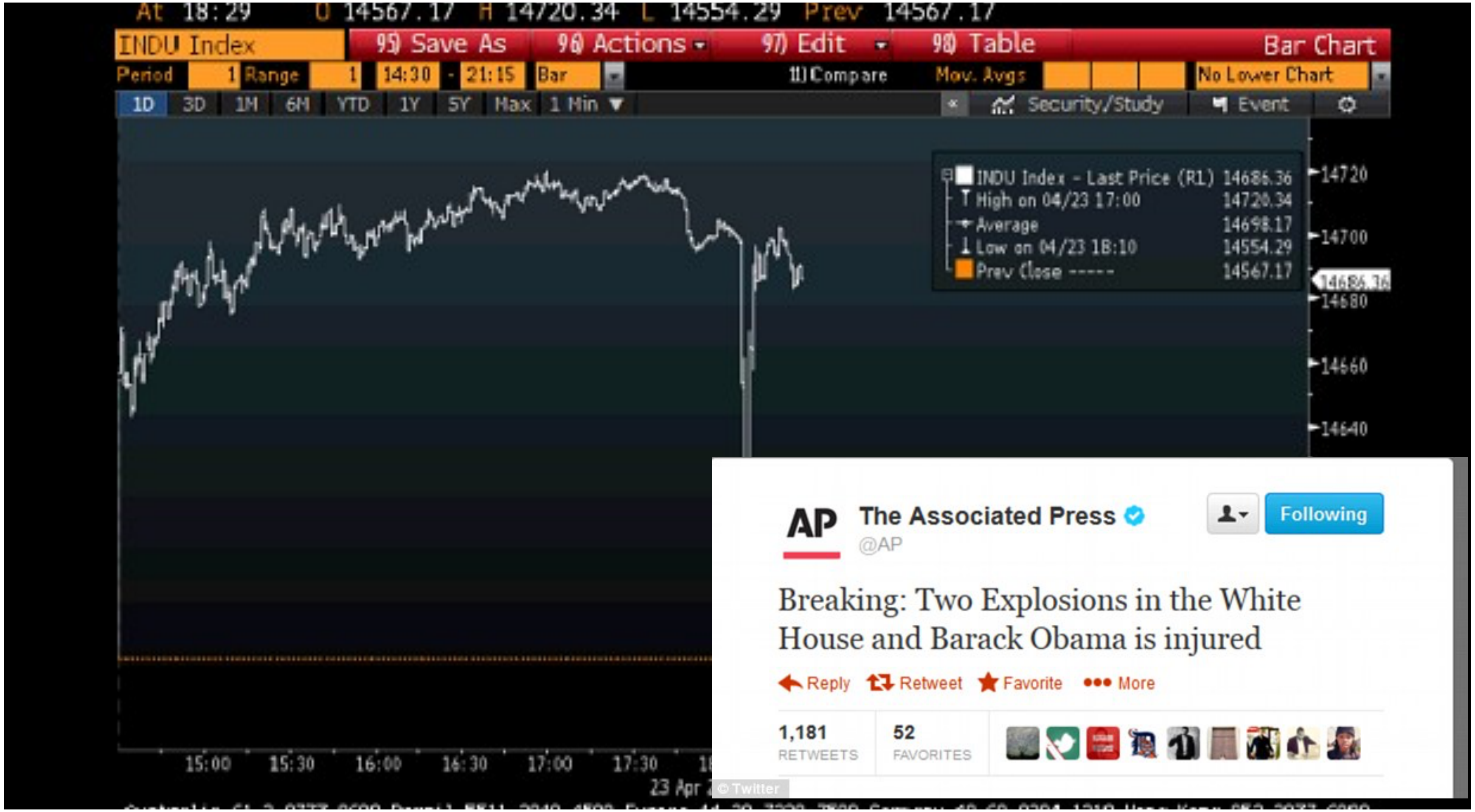
In 2010 in Philadelphia only about 11% of emergencies were being treated at less-expensive urgent-care facilities, versus 34% across all of Caesars. The Harrah's team launched a campaign to remind employees of the high cost of ER visits and provided a list of alternative facilities. Two years later, 17% of emergencies were going to urgent care, and the percentage of individuals making multiple ER visits fell to 30% from 40%.

(3) DATA VELOCITY

- Analysis of streaming data to enable decisions within fractions of second

- Real time data

DOW JONES & DATA



AP The Associated Press @AP

Following

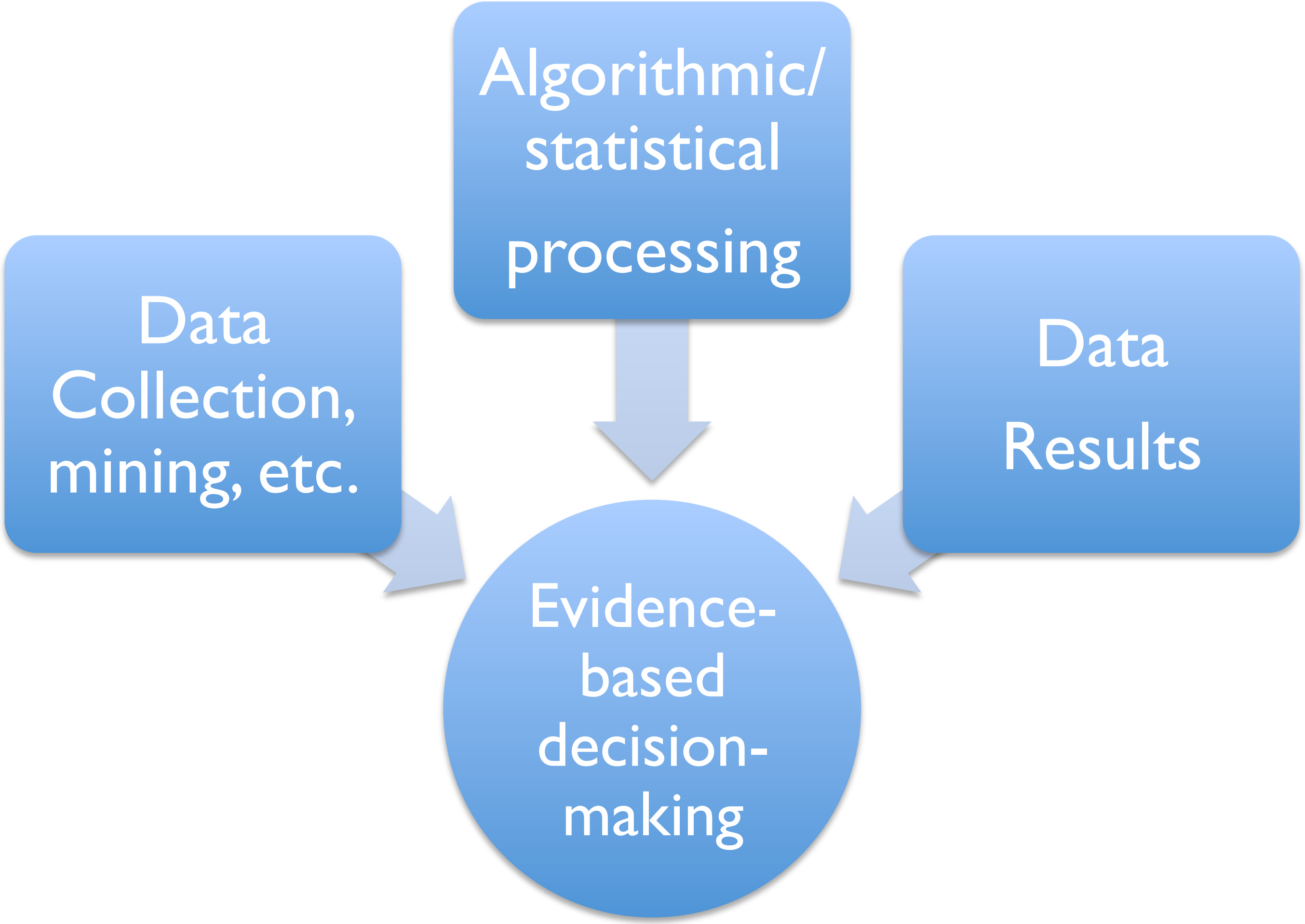
Breaking: Two Explosions in the White House and Barack Obama is injured

1,181 RETWEETS 52 FAVORITES

HEC
PARIS

BIG DATA IN THE PRIVATE SECTOR

Data and decision-making



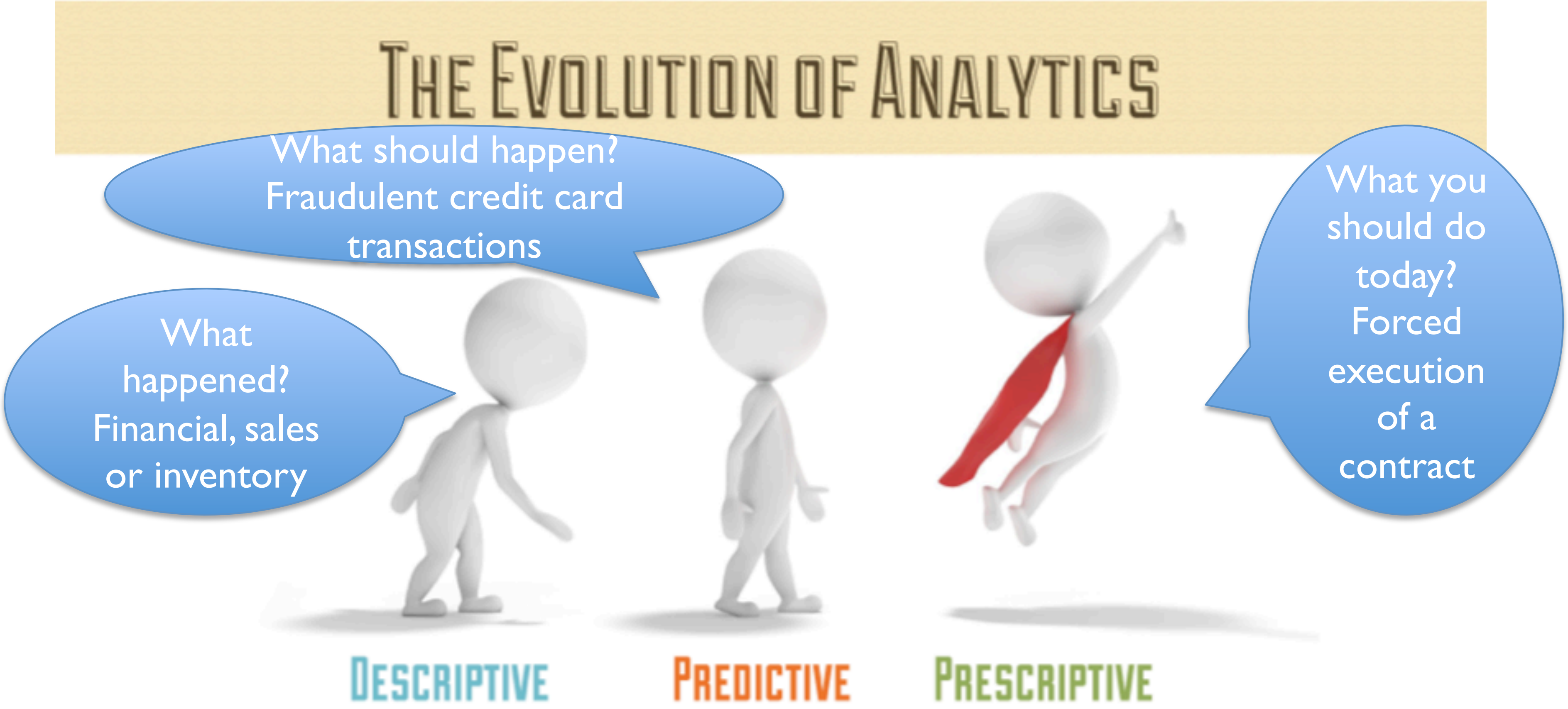
Algorithmic/ statistical processing

Data Collection, mining, etc.

Data Results

Evidence-based decision-making

VS. BIG DATA

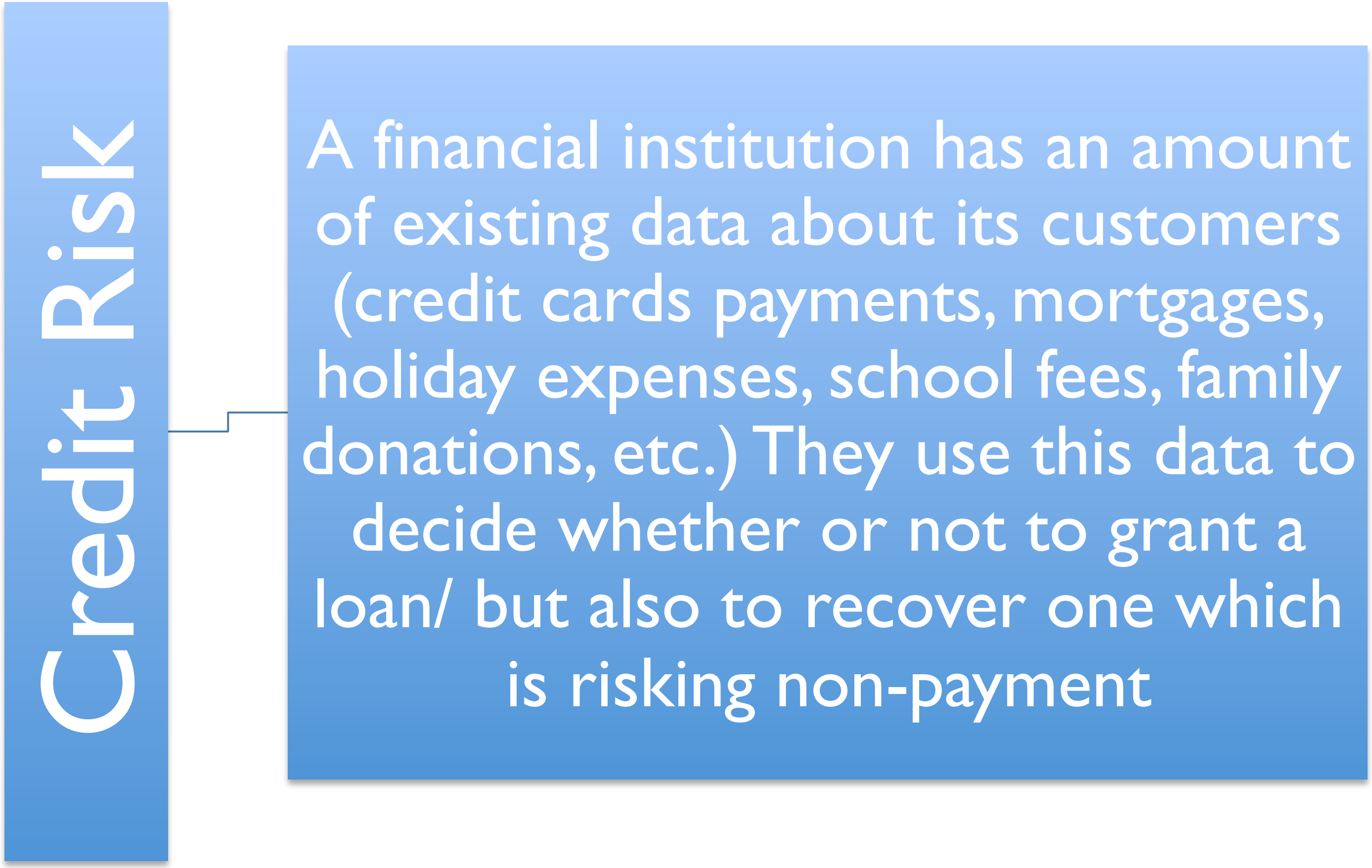


What happened? Financial, sales or inventory

What should happen? Fraudulent credit card transactions

What you should do today? Forced execution of a contract

DATA-MINING IN FINANCIAL AND BANKING LAW



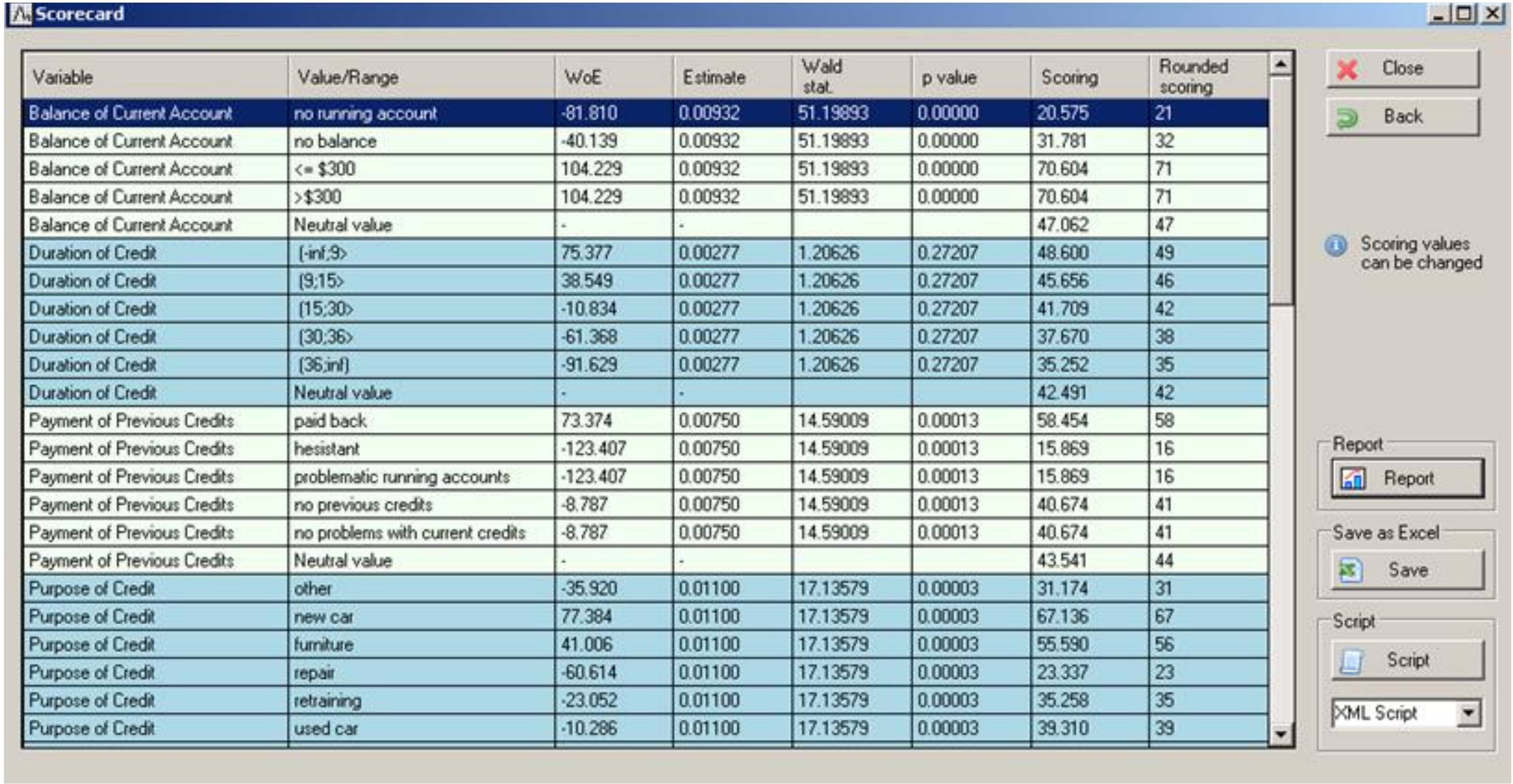
Credit Risk

A financial institution has an amount of existing data about its customers (credit cards payments, mortgages, holiday expenses, school fees, family donations, etc.) They use this data to decide whether or not to grant a loan/ but also to recover one which is risking non-payment

HEC PARIS

By statsoft/STATISTICA

CREDIT SCORECARD



Scorecard

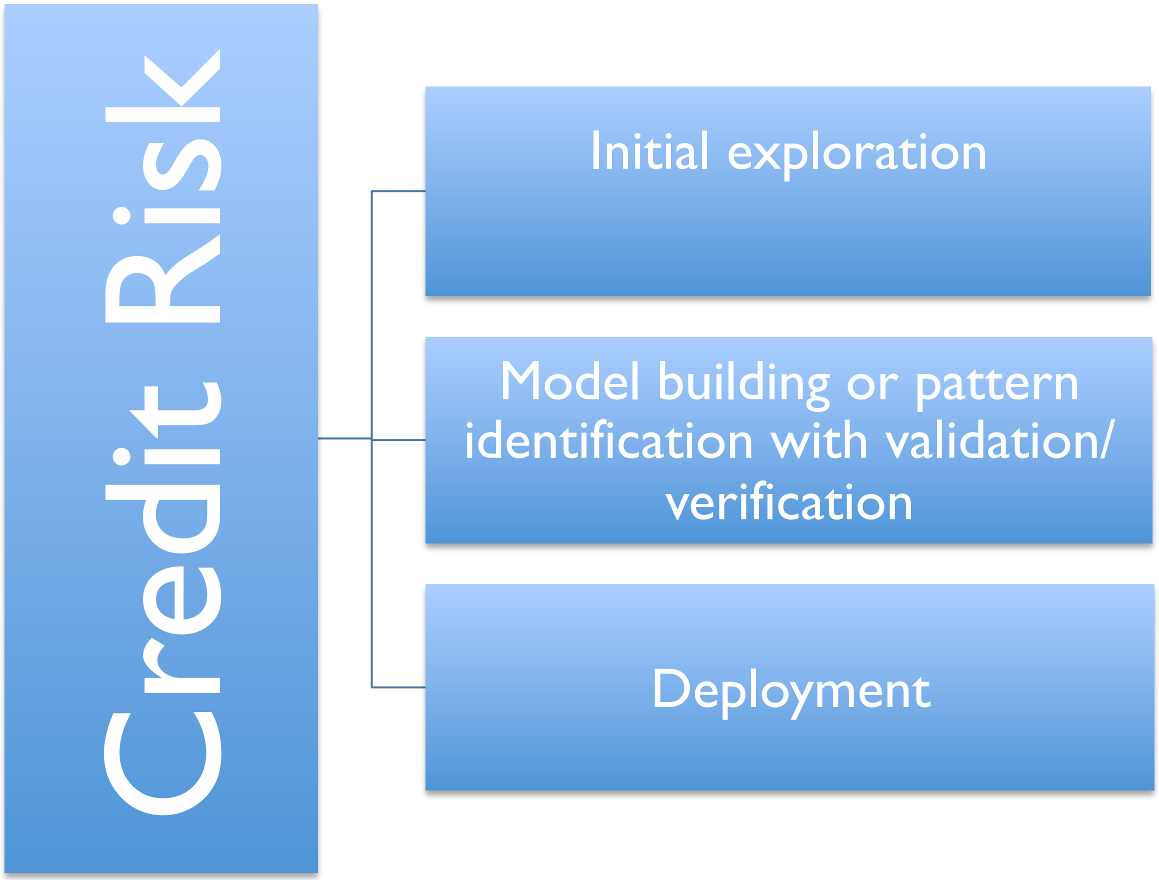
Variable	Value/Range	WoE	Estimate	Wald stat.	p value	Scoring	Rounded scoring
Balance of Current Account	no running account	-81.810	0.00932	51.19893	0.00000	20.575	21
Balance of Current Account	no balance	-40.139	0.00932	51.19893	0.00000	31.781	32
Balance of Current Account	<= $300	104.229	0.00932	51.19893	0.00000	70.604	71
Balance of Current Account	>$300	104.229	0.00932	51.19893	0.00000	70.604	71
Balance of Current Account	Neutral value	-	-			47.062	47
Duration of Credit	(-inf;9>	75.377	0.00277	1.20626	0.27207	48.600	49
Duration of Credit	(9;15>	38.549	0.00277	1.20626	0.27207	45.656	46
Duration of Credit	(15;30>	-10.834	0.00277	1.20626	0.27207	41.709	42
Duration of Credit	(30;36>	-61.368	0.00277	1.20626	0.27207	37.670	38
Duration of Credit	(36;inf)	-91.629	0.00277	1.20626	0.27207	35.252	35
Duration of Credit	Neutral value	-	-			42.491	42
Payment of Previous Credits	paid back	73.374	0.00750	14.59009	0.00013	58.454	58
Payment of Previous Credits	hesistant	-123.407	0.00750	14.59009	0.00013	15.869	16
Payment of Previous Credits	problematic running accounts	-123.407	0.00750	14.59009	0.00013	15.869	16
Payment of Previous Credits	no previous credits	-8.787	0.00750	14.59009	0.00013	40.674	41
Payment of Previous Credits	no problems with current credits	-8.787	0.00750	14.59009	0.00013	40.674	41
Payment of Previous Credits	Neutral value	-	-			43.541	44
Purpose of Credit	other	-35.920	0.01100	17.13579	0.00003	31.174	31
Purpose of Credit	new car	77.384	0.01100	17.13579	0.00003	67.136	67
Purpose of Credit	furniture	41.006	0.01100	17.13579	0.00003	55.590	56
Purpose of Credit	repair	-60.614	0.01100	17.13579	0.00003	23.337	23
Purpose of Credit	retraining	-23.052	0.01100	17.13579	0.00003	35.258	35
Purpose of Credit	used car	-10.286	0.01100	17.13579	0.00003	39.310	39

Scoring values can be changed

Classic model

Source Statsoft

DATA-MINING IN FINANCIAL AND BANKING LAW



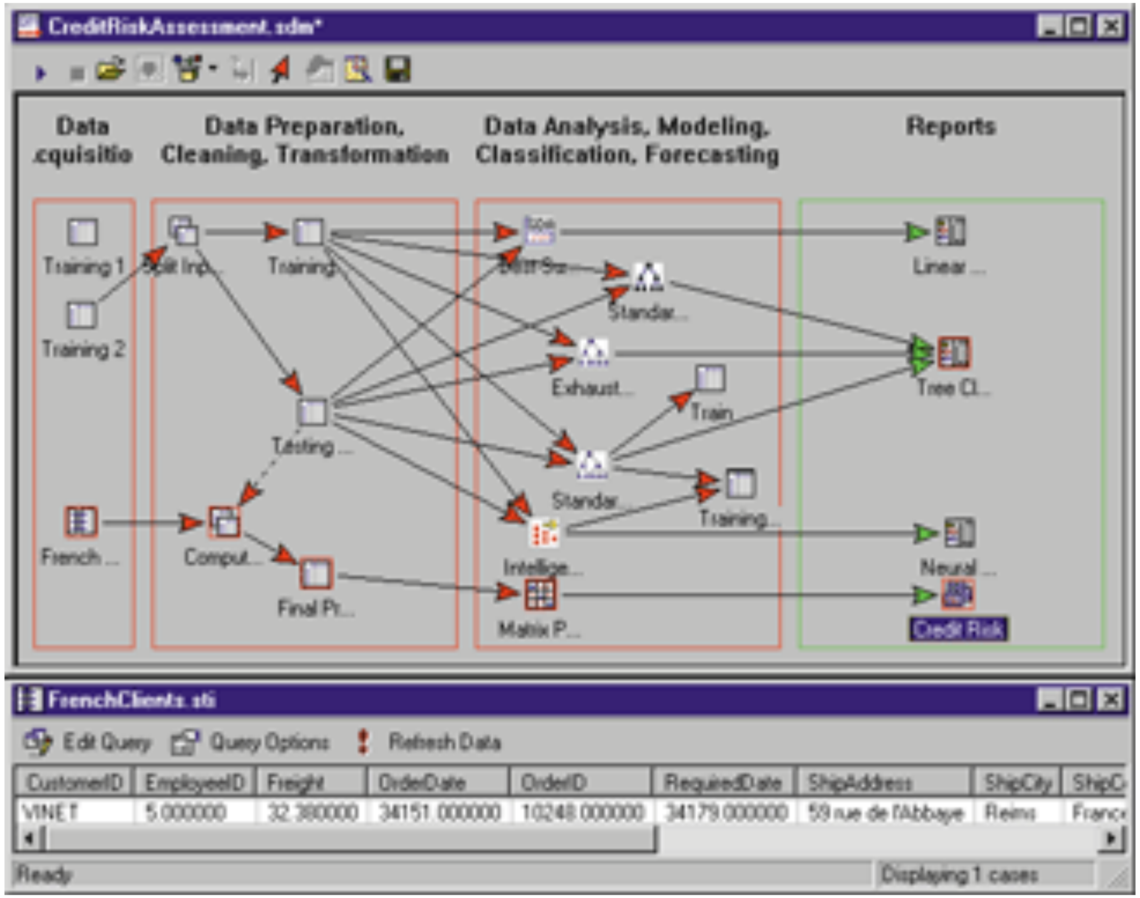
Credit Risk

Initial exploration

Model building or pattern identification with validation/ verification

Deployment

EX. OF CREDIT RISK ASSESSMENT WITH STATISTICA



CustomerID	EmployeeID	Freight	OrderDate	OrderID	RequiredDate	ShipAddress	ShipCity	ShipC
VINET	5.000000	32.380000	34151.000000	10248.000000	34179.000000	59 rue de l'Abbaye	Reims	Franc

Ready

Displaying 1 cases

Source Statsoft

PERFORMANCE ASPECT

Predict the future payment behavior of existing debtors in order to identify/isolate bad customers to direct more attention and assistance to them, thereby reducing the likelihood that these debtors will later become a problem.

Example

- **Behavioral scoring.** Scoring models that evaluate the risk levels of existing debtors.

BAD DEBT MANAGEMENT

Select optimal collections policies in order to minimize the cost of administering collections or maximizing the amount recovered from a delinquent's account.

Scoring models for collection decisions:
Scoring models that determine when actions should be taken on the accounts of delinquents and which of several alternative collection techniques might be more appropriate and successful.

BIG DATA

Thus, the overall objective of credit scoring is not only to determine whether the applicant is credit worthy, but also…

to attract quality credit applicants who can subsequently be retained and controlled while maintaining an overall profitable portfolio.

BIG DATA & ALGORITHMIC LAW ENFORCEMENT

“RIGHT TO BE FORGOTTEN” *(C-131/12)*

The Spanish court referred the case to the Court of Justice of the European Union asking:

(a) whether the EU’s 1995 Data Protection Directive applied to search engines such as Google?

(b) whether EU law (the Directive) applied to Google Spain, given that the company’s data processing server was in the United States?

(c) whether an individual has the right to request that his or her personal data be removed from accessibility via a search engine (the ‘right to be forgotten’)?

ECJ RULING

a) On the territoriality of EU rules : Even if the physical server of a company processing data is located outside Europe, EU rules apply to search engine operators.

b) On the applicability of EU data protection rules to a search engine : Search engines are controllers of personal data

c) On the "Right to be Forgotten" : Individuals have the right - under certain conditions - to ask search engines to remove links with personal information about them. This applies where the information is inaccurate, inadequate, irrelevant or excessive.

OPTIONS FOR THE OWNER

Youtube Content ID

Big Data & algorithmic enforcement

The system notifies the alleged owner, he can:

(1) mute audio that matches their music;

(2) (block a whole video from being viewed;

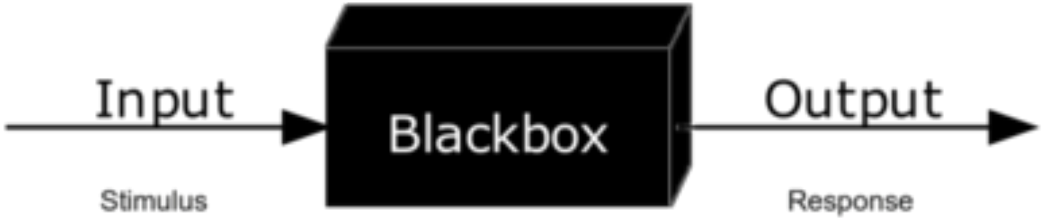
(3) monetize the video by running ads against it; or

(4) track the video's viewership statistics

THE USER CAN

1) acknowledge the claim;
2) if the claim is for a piece of music in the video, the user can choose to remove the song without having to edit and reload the video;
3) the user can swap out the allegedly infringing song with a free-to- use song;
4) the user can share revenue with the copyright owner; or
5) dispute the claim.

YOUTUBE BLACK BOX EXPERIMENT



Accountability in Algorithmic Enforcement: ***Lessons from Copyright Enforcement by Online Intermediaries***
Maayan Perel+ & Niva Elkin-Koren

VIACOM INTERNATIONAL, INC. V. YOUTUBE, INC., NO. 07 CIV. 2103

- Viacom filed a US$1 billion lawsuit against Google and YouTube
  - Alleged copyright infringement by allowing users to upload and view copyrighted material owned by Viacom - SpongeBob SquarePants and The Daily Show, -1.5 billion times.
- The judge refused to force YouTube to provide Viacom with the computer source code which controls both the YouTube.com search function and Google's internet search tool "Google.com."

> *The search code is the product of over a thousand person-years of work and there is no dispute that its secrecy is of enormous commercial value.*

UNITED STATES DISTRICT COURT SOUTHERN DISTRICT OF NEW YORK

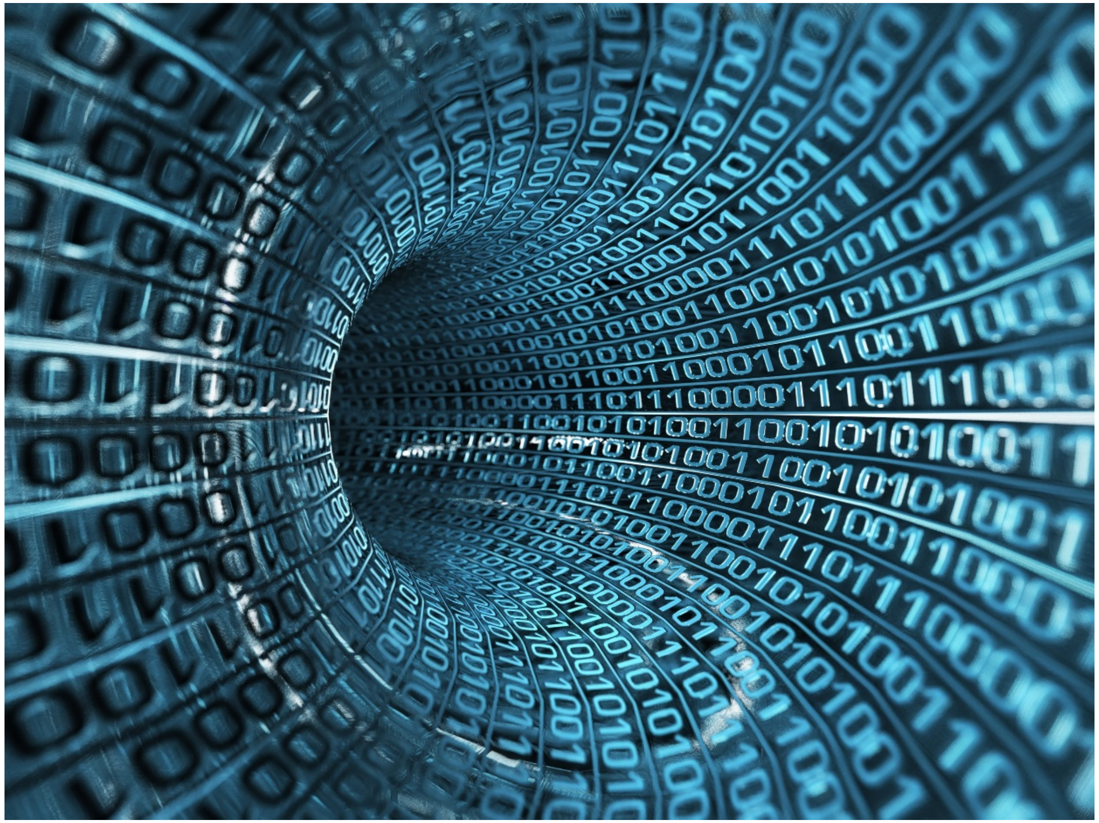
HEC PARIS

BIG DATA IN THE PUBLIC SECTOR

DATA.
IT'S ALL AROUND US

Transport for London (TfL) collects card data from **8 million** trips per day

2.5 EXABYTES PER DAY



That’s big by anyone’s standards.
But it's the velocity, variety and volume of data that has merited the new term.

WHAT CAN IT DO?

Big Data Analytics:

Capable of uncovering hidden patterns, unknown correlations, formerly impossible insights into societal problems.

More coordinated and integrated understanding of social activity and of society.

WHY IS IT IMPORTANT FOR GOVERNMENT?

More efficient, saves money, identifies fraud, and helps public institutions better serve the citizens.

Reduce administrative costs by 15% to 20%

150 billion to EUR 300 billion new value

UK public sector: GBP 2 billion in fraud detection and generate GBP 4 billion

(OECD, 2013: 329)

DO GOVERNMENTS HAVE DATA?

One of the most data intensive sectors

The public sector has lots of data about the public, including very personal data about income, employment type, health, lifestyle, etc.

www.data.gov.uk → the world's leading data portal, featuring over 10,000 datasets

WHY SHOULD I CARE?

Already

'Game-Changing' in Law Enforcement
Criminal Justice
Taxation

Soon to come: Law-making & Law Practice

Policy-Making:
what governments do.

CASE STUDY 1

HMRC AND TAX ECTION

'Connect' Computer System

Costed £45m, launched in the summer of 2010

Aim: Help find undeclared tax

How does it work?

1. Data from banks, local councils, land registry, popular online marketplaces such as eBay or Gumtree, and even social media like Facebook and Twitter.

2. It matches its findings against the information the taxpayer has provided through their tax return.

3. It hunts for income discrepancies, which can then prompt a tax investigation.

HMRC AND TAX COLLECTION

A success?

Yes.

£3bn extra tax has been clawed as a result of Connect since its launch in 2008

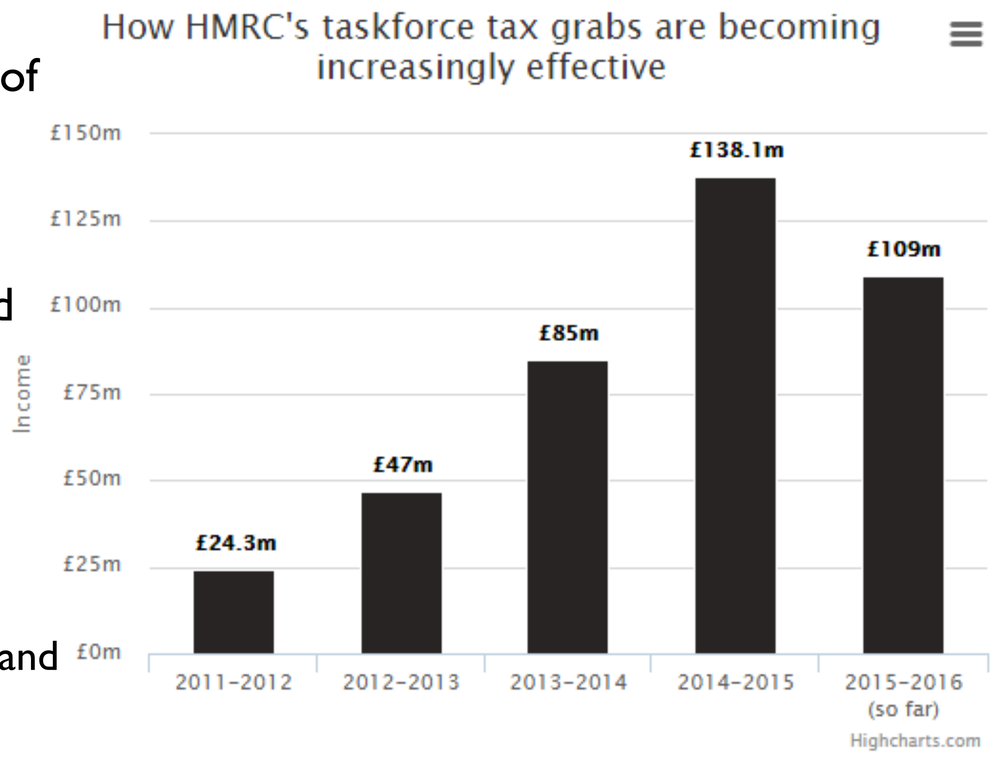
Time for investigations has been reduced significantly.

Investigators are more effectively deployed

Source: BBC, June 2015

But many have doubts

- Fears about security
- Privacy: so much information residing in the hand of the state



Source: The Telegraph, 25 Oct 2015

CASE STUDY 2
HOME OFFICE, EAST MIDLANDS, UK

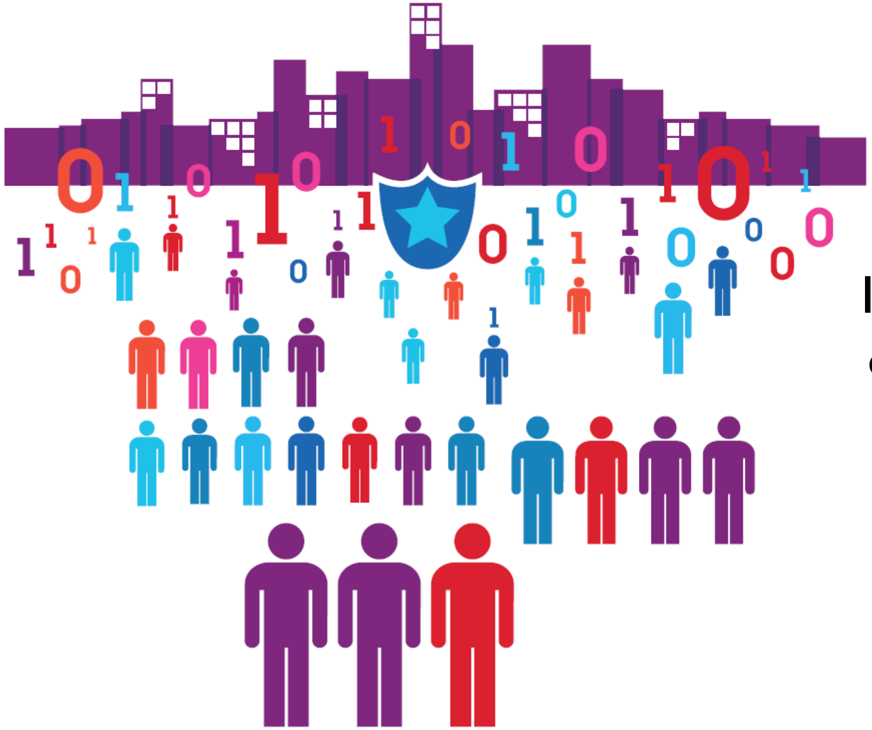
Home Office

Predictive policing:
Forecasting Crime Locations 'Hotspots'

How does it work?

- The software collects info about certain types of crime, e.g., burglary
- Algorithms forecast 'hot-spots' where the probability pf crime will be greater.
- This informs police decisions about which areas to visit on foot patrol.

HOME OFFICE, EAST MIDLANDS, UK



A success?

Its projections were accurate in 78% of cases, compared to 51% accuracy using traditional techniques.

Source: Houses of Parliament, Postnote 470 July 2014

CASE STUDY 3
CENTERS FOR MEDICARE AND MEDICAID SERVICES, US

A collaboration with IBM since 2010

'Fraud Prevention System'

Aim: to ensure that correct payments are made to legitimate providers.

CENTERS FOR MEDICARE AND MEDICAID SERVICES, US

How does it work?

- It collects data from various healthcare providers.
- It identifies suspicious billing patterns by healthcare providers: e.g., providers bill.

Does it work?

- Saved $3 for every $1 invested in the first year.
- It prevented or identified an estimate of $115 million payments
- Generated 536 new investigations and augmented info for 511 pre-existing investigations

Source: US Department of Health and Human Services

HEC PARIS

LAW AND BIG DATA: CHALLENGES AND OPPORTUNITIES

CHALLENGES & OPPORTUNITIES

- Privacy concerns: Invasive?
- Ownership of data: What jurisdiction?
- Expensive and inaccessible to some Governments
- Need for ‘data-literate’ civil servants
- Technocratic … but undemocratic?
- Legal Risk Assessment
- Get your data right!

DOSOMETHING.ORG

How many views make a YouTube video a success?

- Dosomething.org launches a campaign to donate sports equipment to youth in need.
- Post video featuring youtube celebrities
- An get the most views ever for a dosomething.com posted video and the second higher for a donation campaign using youtube only.
- 1.5 millions views but..
- 0€ in donations!

Wrong metric!

Cristina Golomoz
Centre for Socio-Legal Studies
Department of Law
University of Oxford
cristina.golomoz@law.ox.ac.uk

David Restrepo A
Assistant Professor
Law and Tax Department
HEC Paris
Restrepo-amariles@hec.fr