Data Analytics

How and why we should be aware about the limits of inferring human behavior from data

Prof Jorge Sanz

Week 9 (class on March 18)

TECHNOLOGY

How Trustworthy Is Big Data?

February 1, 2018

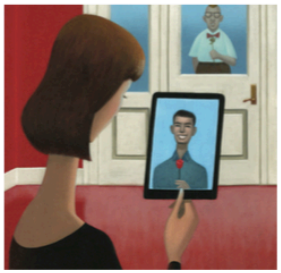
Timo Elliott

Vice President, Global Innovation Evangelist at SAP

Deloitte Review

Issue 21 | July 2017

Complimentary article reprint



Predictably inaccurate

The prevalence and perils of bad big data

By John Lucker, Susan K. Hogan, and Trevor Bischoff

Illustration by Jon Krause

Topics

PART 1 – Lesson 8

- What is Data Analytics ?
- Why does Data Analytics matters to this course ?
- Understanding why some ideas underlying Data Analytics used enterprises and society need to be put under careful scrutiny
  - The bright side and the dark side of Analytics innovations
- Is Data from individuals a trustable footprint of their behavior ?
  - Following some Nobel Prizes and behavioral / cognitive scientists

PART 2 – Lesson 9

- How the impact of Data Analytics is augmented with *Big Data*
- Data Analytics from Ethics, Legal, Business and Society perspectives
- Challenge yourself
  - Personal: *Do I know the potential issues or damage from some "great ideas" ?*
  - Business consequences: *What may be illegal in these ideas ?*
  - Societal consequences: *What are the implications in society my individual posture ?*
- Conclusions and Take-away
- Recommended reading and more literature

Data Analytics - Definition

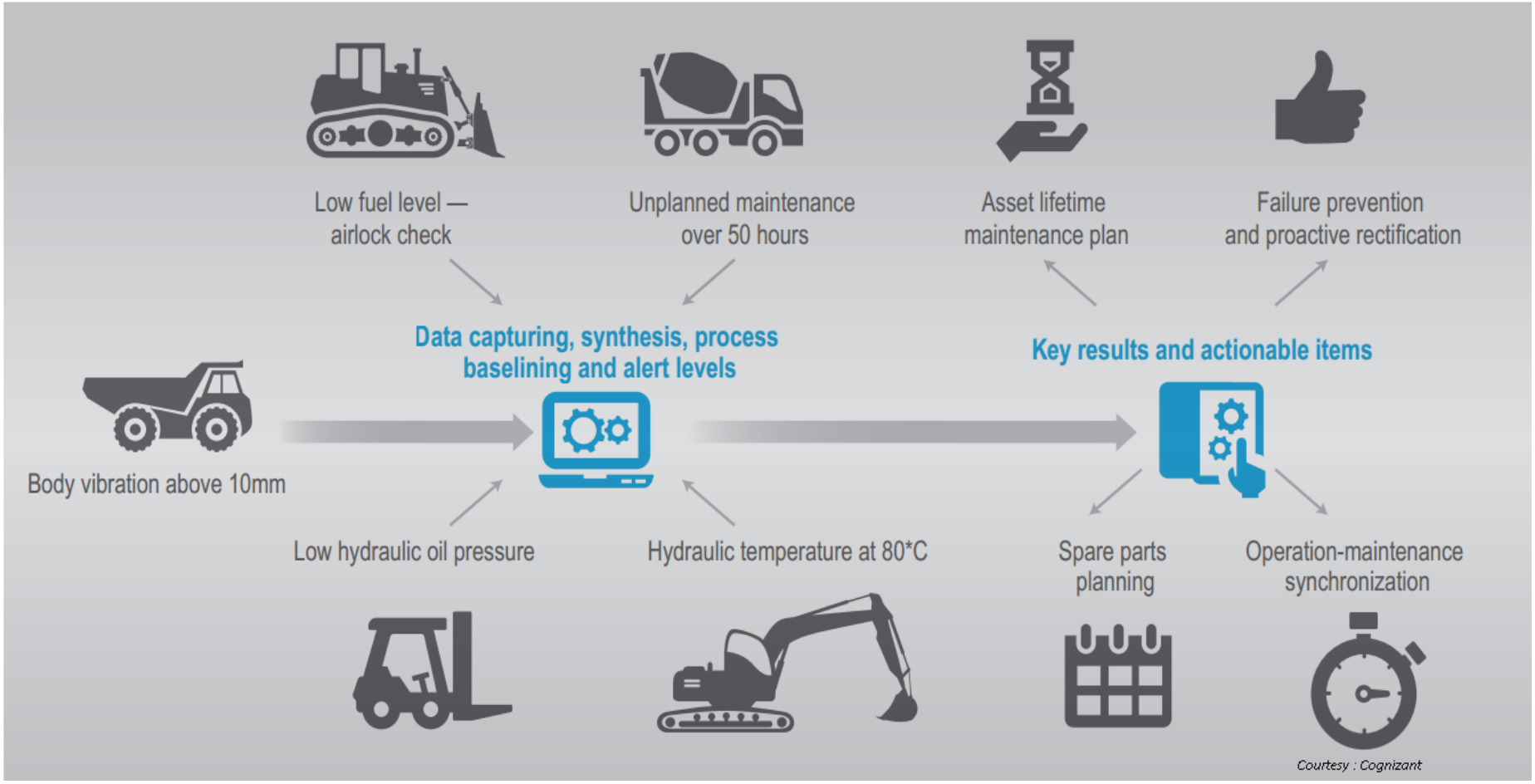
- The analysis of data collected on any system with the goal of summarizing or revealing something important and/or actionable about it

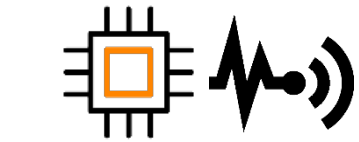
- You will hear a number of related terms, such as Business Analytics, Data Science, Business Intelligence …
  - Technically, these are all **different** …
  - … But you can safely use one or the other for our core discussion at play here

Used in enterprises with Data coming from …

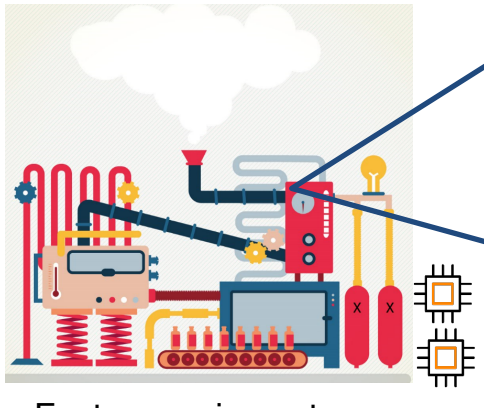
- … systems made of different mechanical and electronic components
  - An ATM Machine; the lamination / printing equipment for consumer packaging industry
- … observations of complex processes like in Manufacturing …
  - The blooming and rolling process for steel beams; the extruding process for manufacturing tires, etc.
- … systems where humans are an essential part of and matter to their functioning and performance:
  - Examples: an enterprise; this class; lending processes or regulatory compliance processes in a bank; the economy of Singapore, etc.

Industry 4.0 – Monetizing Data with Analytics





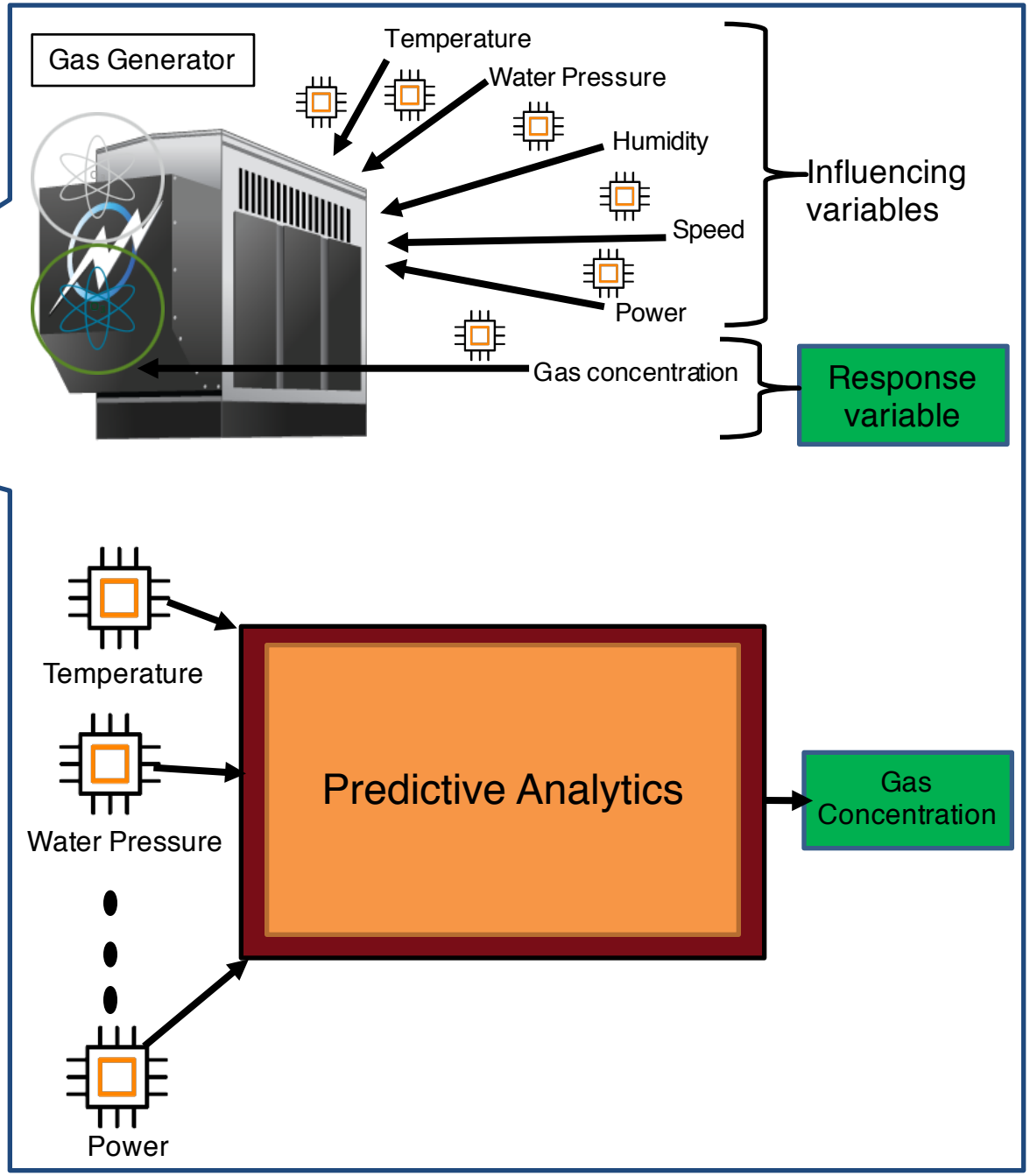
Sensor data from the equipment



Factory equipment

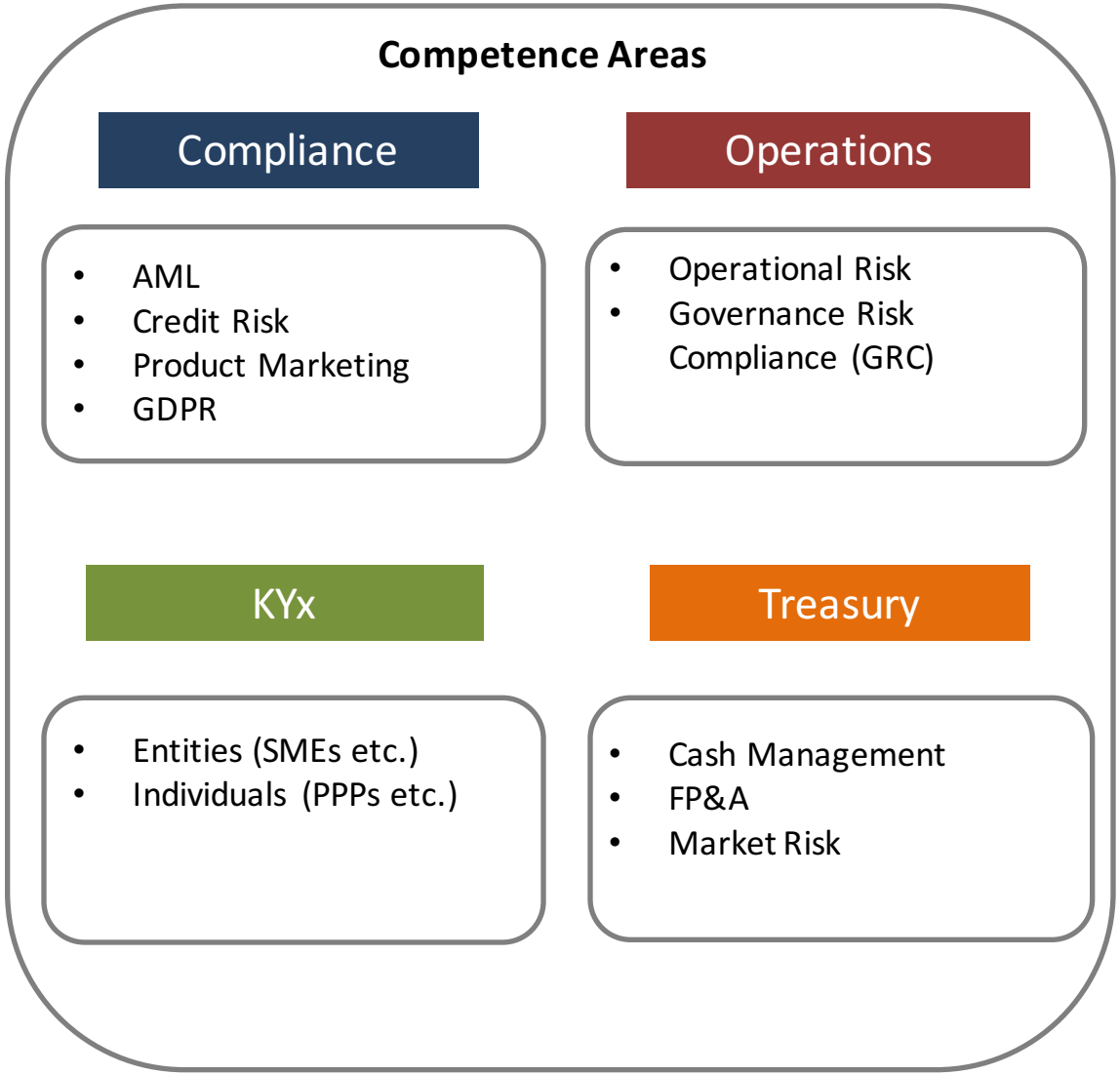
- 200 sensors on every equipment on the shop floor
- Sensors emit data at every 500 millisecond interval

Predictive Analytics being performed to understand the underlying the data patterns and to predict the abnormality of gas concentration in equipment.



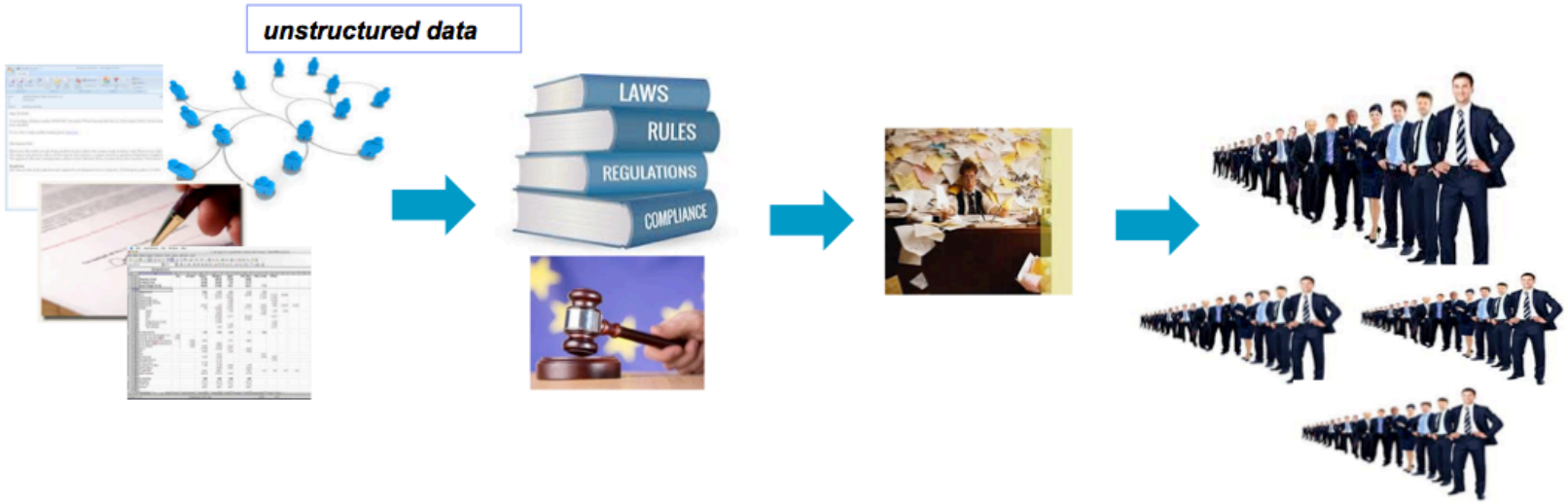
Analytics for Predictive Maintenance of some Systems

Analytics in Finance Competences and Processes



Analytics in the *Law-to-Report* value-chain

The unsustainable approach to hire more and more people …



Analytics in a Legal-Tech Platform for GRC

1	2	3	4	5
				
Vertical application – to improve collaboration, research and development, and innovation adaption through linked and interoperable legal financial documents.	That will allow to design legal financial service knowledge organization system. This will enable information systems to do more with legal financial knowledge and financial big data.	Query to linked and interoperable legal financial documents and knowledge bases. Extract complex financial information that can be returned, for instance, in a table format, which can be further incorporated into another complex financial knowledge search engine.	The inference techniques used in the legal financial knowledge bases will create, among others, two added values – (a) help to discover possible inconsistencies in the integrated financial data, and (b) discover new relationships between legal financial knowledge/rules and financial data.	An API Economy for financial regulation can be further deployed at a full scale. This will be extremely useful to apply legal analytics and regulatory measures in order to create new financial values. These will be cloud-deployed and exploited to create even richer and customized applications.

Data Analytics has varying depth …

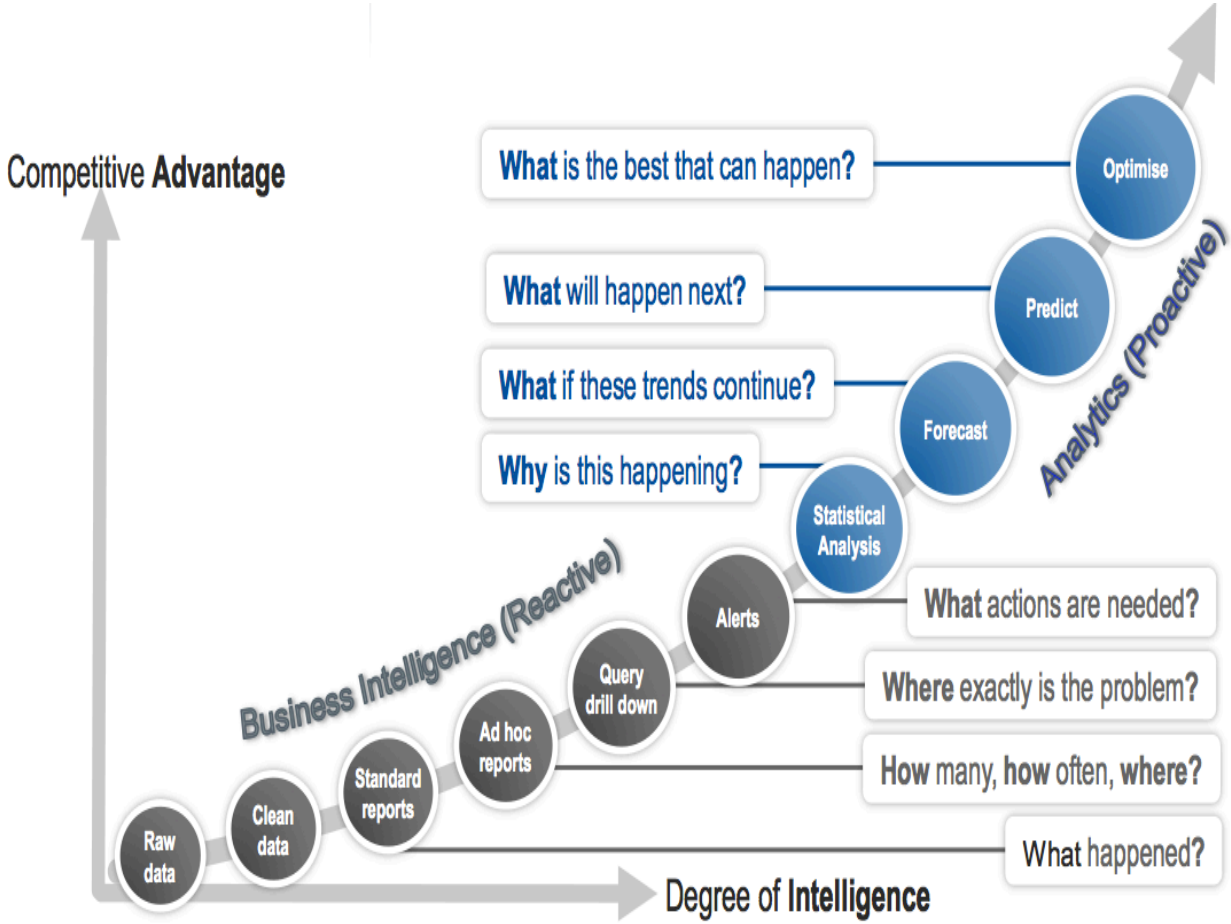
- It encompasses distinct levels of sophistication and different goals, ranging from *reporting* to *prescribing*

➔ Increasingly used for achieving automation in decision-making

Data Analytics: Business and Technical Chains

Analytics
Business Value Chain

Analytics
Technical Value Chain



Organized and Disorganized Complexity

- **Disorganized Complexity:** Structures made of disconnected parts (from a few to billions), operating in random fashion with their interaction being chaotic and predictable – not based on intentional engineering but described through probability and chance computations

- **Organized complexity:** Structures whose character depends not only on the properties of the individual elements of which they are composed, and the relative frequency with which they occur, but also on the manner in which the individual elements are connected with each other

Warren Weaver and many others 60 years after him

Making your life a bit simpler …

- **Disorganized Complexity:**
  - Most IoT examples …
- **Organized complexity:**
  - Enterprises, like true socio-technical systems, are the most prominent form of ***organized complexity***
  - Social Sciences address phenomena of organized complexity

What happens when analysis for *disorganized complexity phenomena* is used for *phenomena of organized complexity* ?

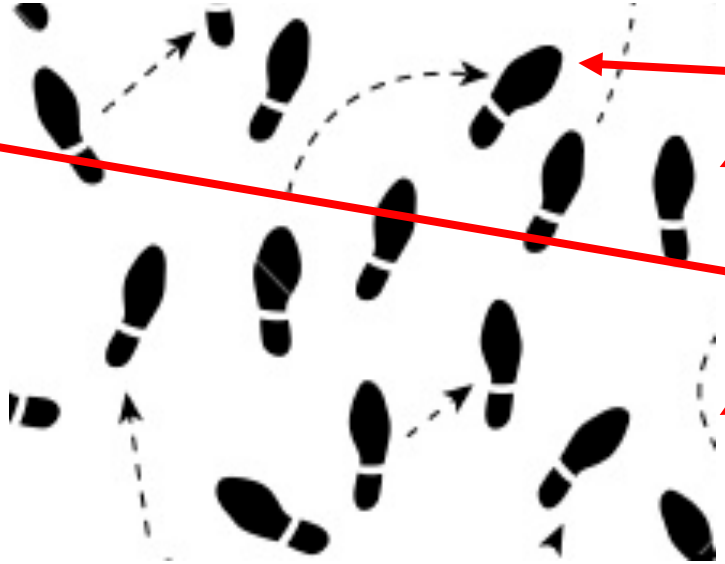
I will try to be as "undiplomatic" as I could … ☺

Story: Jorge being profiled by a "data scientist" …

#Jorge says:
I wish I could walk crazy like this … but every day

#Jorge says soon after:
I love this Vodka … I wish every day would be like today !!!

BB



Data …
(picture from a blog written from Jorge)

Analytics carried out on image …

Conclusion on "Jorge's *behavior*" …

"Jorge may be drunk too often"

Actions and emerging Consequences

Jorge's credit score may drop to 459

Let's sell Jorge more Smirnoff

Real-life Jorge's behavior …

Jorge was in Tango dancing practices this month
He drunk Vodka at his birthday party

Let's follow Von Hayek (Nobel Prize 1974)...

"*Unlike the position that exists in the physical sciences, in economics and other* ***disciplines that deal with essentially complex phenomena****, the aspects of the events to be accounted for about which* ***we can get quantitative data are necessarily limited and may not include the important ones****.*

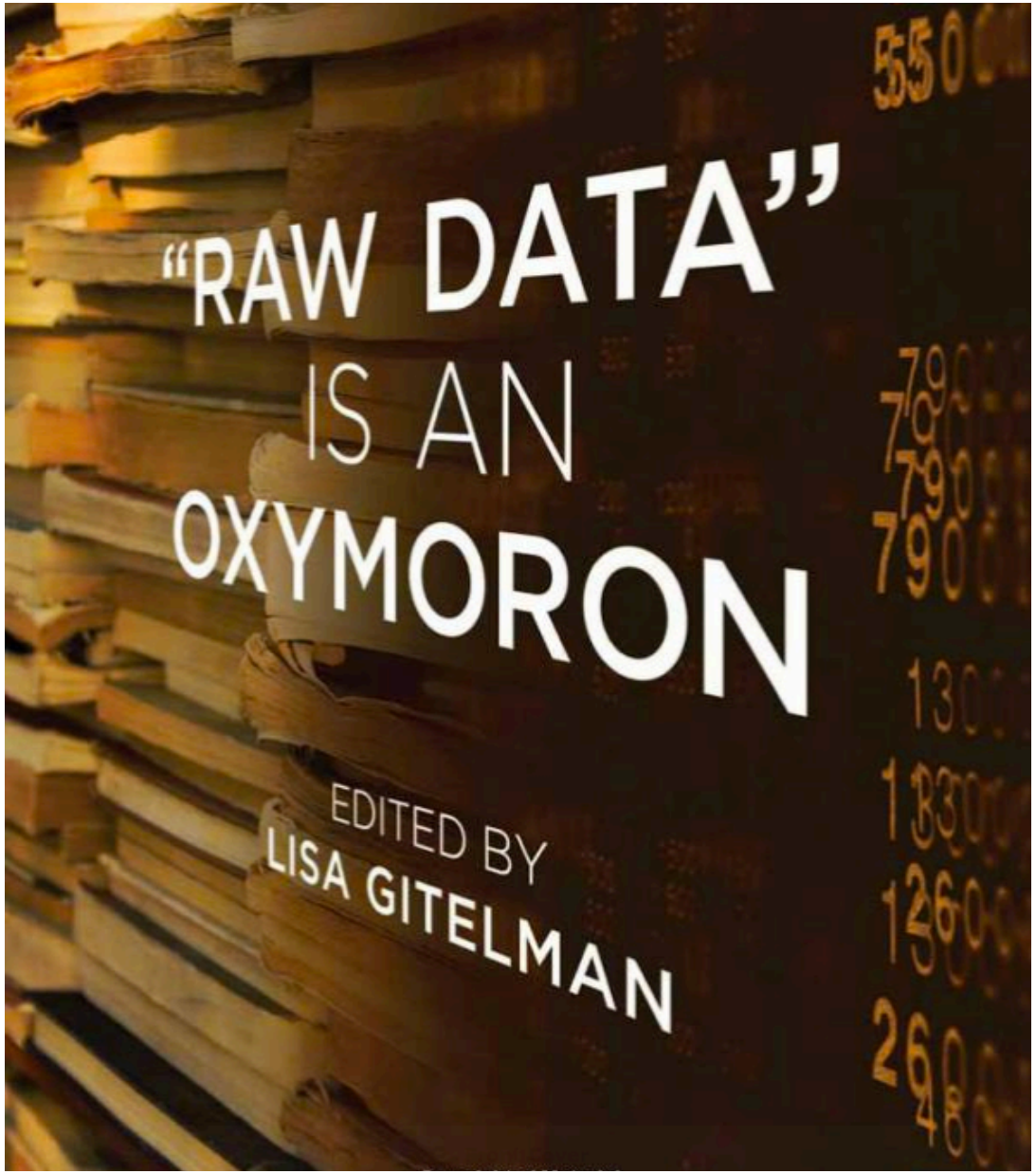
While in the physical sciences it is generally assumed, probably with good reason, that <u>any important factor which determines the observed events will itself be directly observable and measurable</u>, in the study of such complex phenomena as the market, which depend on the actions of many individuals, all the circumstances that will determine the outcome of a process ***will hardly ever be fully known or measurable****"*.

"*Scientism and the Study of Society*", *Economica*, vol. IX, no. 35, August 1942, reprinted in *The Counter-Revolution of Science*, Glencoe, Ill., 1952, p. 15 of this reprint.

Von Hayek, **The Pretense of Knowledge**, 1974 [dissertation - Nobel Prize in Economics]
https://www.nobelprize.org/prizes/economic-sciences/1974/hayek/lecture/

You may have probably heard, and you will definitely hear a lot more in your studies, the term …

“Raw Data”



MIT Press, 2013

https://www.goodreads.com/book/show/15926284-raw-data-is-an-oxymoron

Raw Data is an Oxymoron in Social Phenomena …

Because there is nothing in social systems that may be measurable without making hard choices about *what to observe* (and thus, most likely, also affecting the subject of the observation) …

But instead, there seems to be a "magical" equation implicitly in many Data Analytics pursuits:

Data → Information → Knowledge → Intention

Challenge

Can we device an algorithm that infers the actual intention of what a business person may be implying when writing an email ?

Does the attempt make any sense ?

Let's follow Von Hayek again …

"It seems to me that this failure of the economists […] is closely connected with their propensity to **imitate as closely as possible the procedures of the brilliantly successful physical sciences** - an attempt that in our field may lead to outright error.

… It is an approach which has come to be described as the *scientistic* attitude - an attitude which, as I defined it some thirty years ago, is decidedly unscientific in the true sense of the word, since it involves a mechanical and uncritical application of habits of thought to fields different from those in which they have been formed."

Von Hayek, **The Pretense of Knowledge**, 1974 [dissertation - Nobel Prize in Economics]
https://www.nobelprize.org/prizes/economic-sciences/1974/hayek/lecture/

13 JAN 2003 RESEARCH & IDEAS

The Subconscious Mind of the Consumer (And How To Reach It)

by Manda Mahoney

Harvard Business School professor Gerald Zaltman says that 95 percent of our purchase decision making takes place in the subconscious mind. But how does a marketer reach the subconscious? Zaltman explains in this Q&A.

Harvard Business School professor Gerald Zaltman's latest book, How Customers Think: Essential Insights into the Mind of the Market, delves into the subconscious mind of the consumer—the place where most purchasing decisions are made. The question: How can marketers understand unconscious consumer thinking? HBS Working Knowledge staffer Manda Mahoney questioned Zaltman about the new book, published by Harvard Business School Publishing.

Mahoney: You state that 95 percent of all cognition occurs in the subconscious mind. How can marketers begin to understand behaviors and attitudes of which customers themselves are not aware?

Zaltman: There are several helpful approaches. One is to double check stated beliefs with actual behavior. For example, many consumers report handling competing brands and

People use less information than they think to make up their minds

Nadav Klein and Ed O'Brien

PNAS December 26, 2018 115 (52) 13222-13227; published ahead of print December 10, 2018 https://doi.org/10.1073/pnas.1805327115

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved November 2, 2018 (received for review March 27, 2018)

Significance

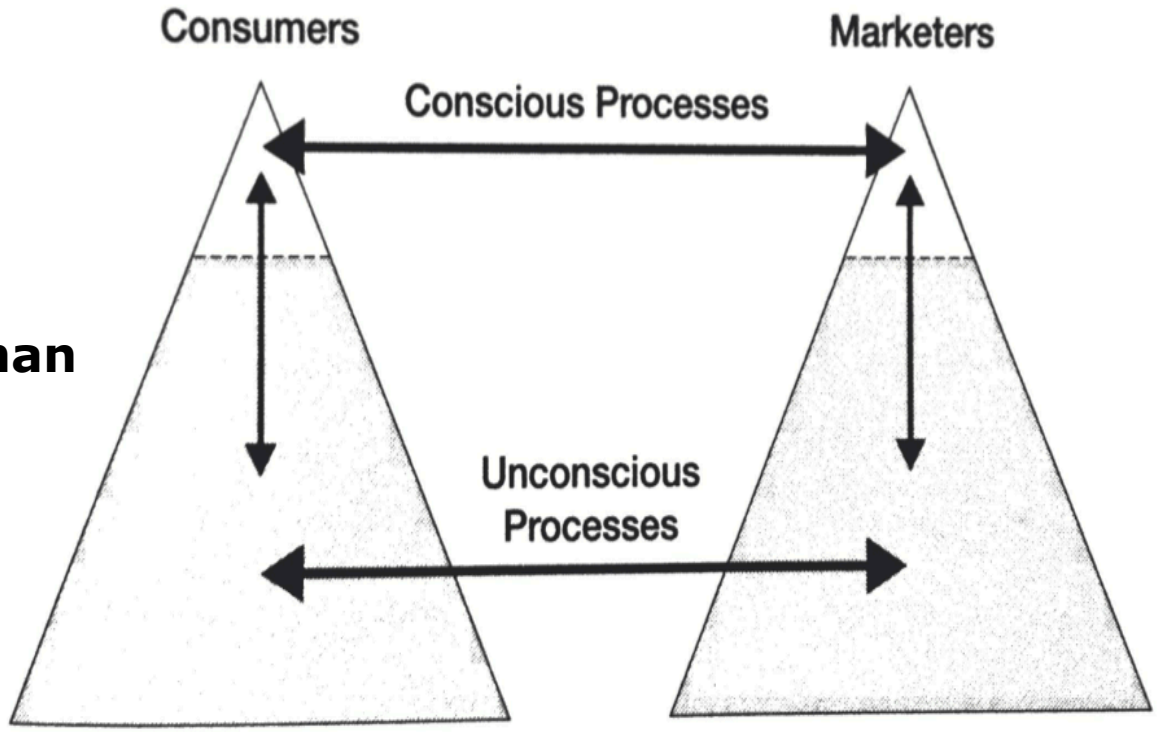
People readily categorize things as good or bad, a welcome adaptation that enables action and reduces information overload. The present research reveals an unforeseen consequence: People do not fully appreciate this immediacy of judgment, instead assuming that they and others will consider more information before forming conclusions than they and others actually do. This discrepancy in perceived versus actual information use reveals a general psychological bias that bears particular relevance in today's information age. Presumably, one hopes that easy access to abundant information fosters uniformly more-informed opinions and perspectives. The present research suggests mere access is not enough: Even after paying costs to acquire and share ever-more information, people then stop short and do not incorporate it into their judgments.

Abstract

A world where information is abundant promises unprecedented opportunities for information exchange. Seven studies suggest these opportunities work better in theory than in practice: People fail to anticipate how quickly minds change, believing that they and others will evaluate more evidence before making up their minds than they and others

The mind and process of the market

80% of the human processes are *Unconscious*...



*Can **data footprints** left by people be the omnipresent source for 'knowing' their behaviors ?*

No surprise: Customers are rightly saying …

“You don’t know me”

Intolerance of mass-market, impersonalized approaches

“You’re not connecting with me”

Demand for interaction on channel of choice

“You make it too hard”

Expectations for immediate results

Chief Marketing Officers Survey, Institute of Business Value, IBM Corporation, 2014

What is it going wrong ?

The fact that customers have *intentionality and choice* means that we cannot "know them" by ***just describing them***

- **Proposal to discuss**: It would be great if those advocates of "knowing" customers by observing their social network interactions and putting people under an "IT microscope" could just get the point and stop it …
- **Rationale:** The social risk of inferencing from digital footprints and related systems interactions …
  - We may generate entirely *theatrical customer* behaviors
  - Examples: In many blogs where people should post a picture, we already see masks or caricatures; people document what they want others to believe about them (not necessarily what they are); etc.

What is it going wrong ?

But who would go that far in such a *digital pretense*, right ? …

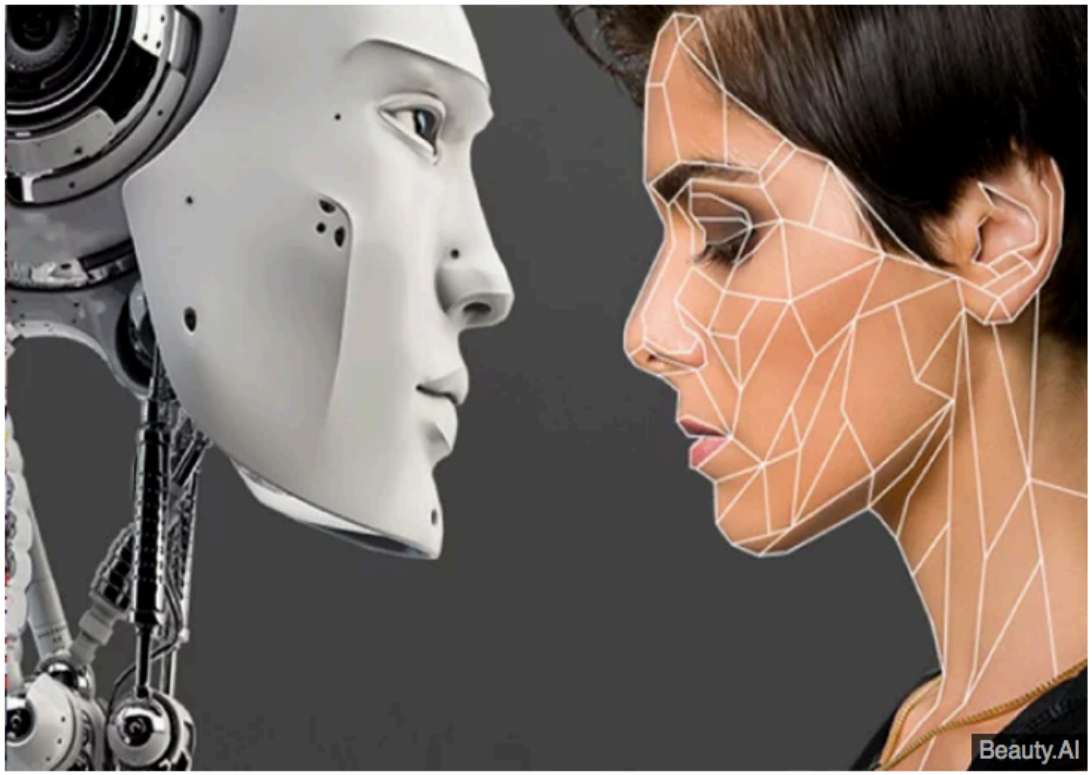
How *Big Data* may seed *"questionable practices"*

… Credit Scoring Perils

- *Favorable credit rating is necessary to purchase a home or car, to start a new business, to seek higher education, or to pursue other important goals. Also necessary to gain access to employment, rental housing, and essential services such as insurance*
- *However, Americans have very little control over how they are scored and have even less ability to contest inaccurate, biased, or unfair assessments of their credit*
- *Traditional credit-scoring tools raise longstanding concerns of accuracy and unfairness. The* ***recent advent of new "big-data" credit-scoring products heightens these concerns***
- *The credit-scoring industry has experienced a recent explosion of start-ups that take an "all data is credit data" approach, combining conventional credit information with thousands of data points mined from consumers' offline and online activities*
  - *Big-data scoring tools base credit decisions on where people shop, the purchases they make, their online social media networks, and various other factors that are not intuitively related to creditworthiness*
- *Credit-scoring tools that integrate thousands of data points, most of which are collected without consumer knowledge, create serious problems of transparency*
  - *Recent studies have also questioned the accuracy of the data used by these tools, in some cases identifying serious flaws that have a substantial bearing on lending decisions*
- *Discriminatory scoring may not be intentional; instead, sophisticated algorithms may combine facially neutral data points and treat them as proxies for immutable characteristics such as race or gender, thereby circumventing existing non-discrimination laws and systematically denying credit access to certain groups*
- *Laws are insufficient to respond to the challenges posed by credit scoring in the era of big-data*

Double Bias in Data Analytics

- Roughly 6,000 people from more than 100 countries submitted photos for an automated algorithm, created by Youth Laboratories, to judge
- Strange to judge beauty – an extremely subjective attribute – with objective algorithms
- It turned out that out of 44 winners, nearly all were white, a handful were Asian, and only one had dark skin
- Only a hand-full of people were not white, so the training of the analytics techniques had severe biased population



"The idea that you could come up with a culturally neutral, racially neutral conception of beauty is simply mind-boggling" …

Prof B. Harcourt, Univ. of Columbia

False Positives

The untold story in data analytics classification

- *Target* looked at data from women who had baby registries with the store, merged that with demographics data, and identified a set of 25 products that, when a combination of them was purchased, would indicate an increased probability that the woman was pregnant

- Even better, depending on the type of products purchased, Target was able to identify the pregnancy's trimester, and thus to predict the baby's approximate due date. This in turn was very helpful so they can market specific products depending on the due-date

- They used these ideas to send coupons to clients in a more customized manner

One of the targeted women was a high-school minor whose family did not know about any potential pregnancy …

What is the lesson-learnt in this real-life case ?

- *Anything illegal ?*
- *Questionable ethics ?*
- *Broader societal issues ?*

Data Analytics *without Consent*

Information derived / inferred without the knowledge of individuals

Have people "consented" to this practice?

Capabilities to "infer" personal information about people

- Breaching the privacy right of an individual who may never have explicitly provided or consented that any entity would be authorized to infer

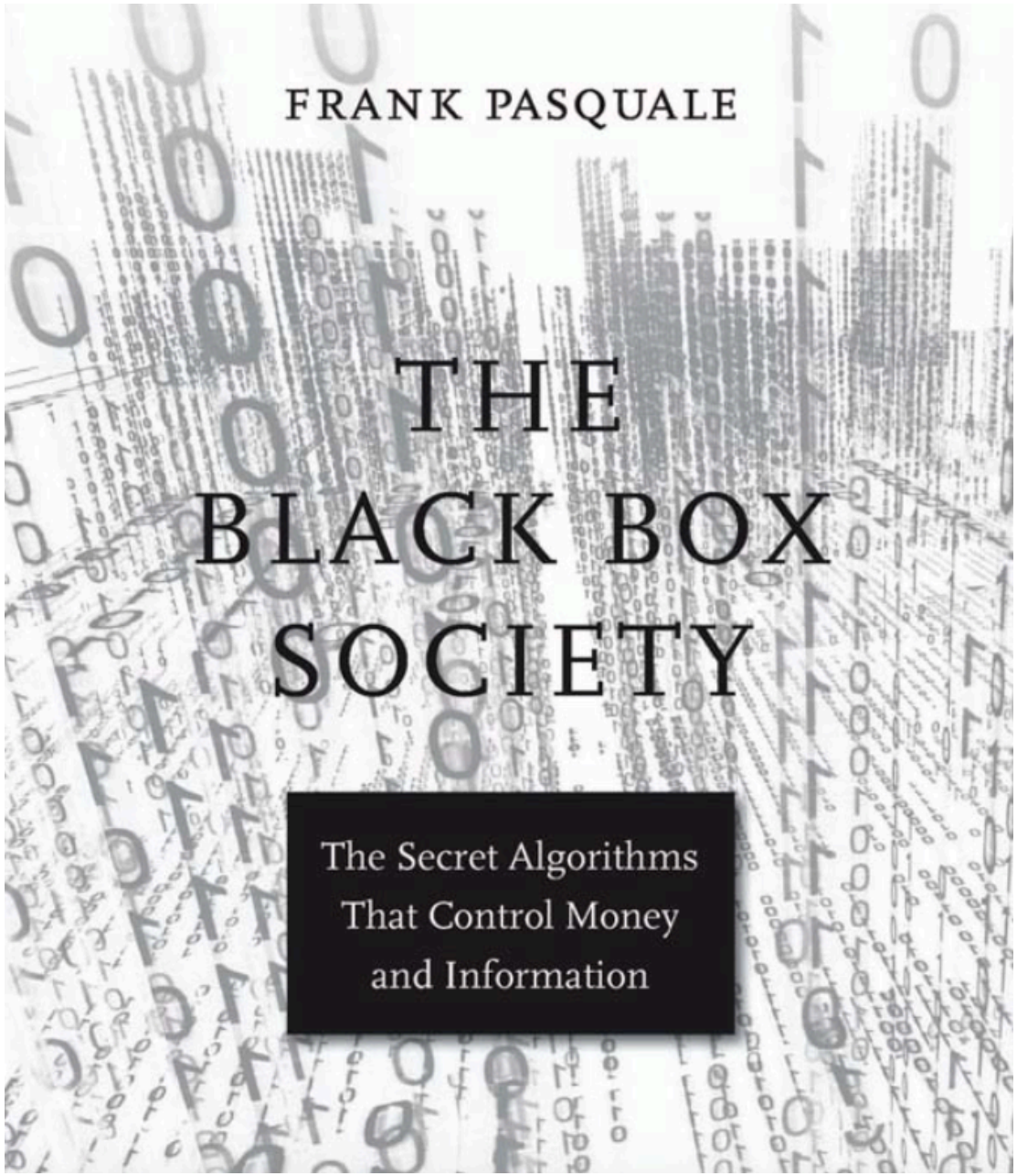
Source of picture from http://www.mirror.co.uk/news/uk-news/june-birthdays-happiest-november-milestones-5792686

Example:

Joe (wearing grey shirt) attended a birthday party and posted the following picture on his Facebook page with a comment *on 28 Sept*:

"What a wonderful party yesterday night. Happy Birthday to Rachel Wilson ! Wishing you lots of happiness and kudos for getting that promotion tomorrow!"

So much we can infer about this picture and text, right?



Harvard University Press, 2015

https://www.goodreads.com/book/show/21878126-the-black-box-society

European Economic and Social Committee

The ethics of Big Data: Balancing economic benefits and ethical questions of Big Data in the EU policy context

Study

Abstract

This study, carried out to support the activities of the EESC, explores the ethical dimensions of Big Data in an attempt to balance them with the need for economic growth within the EU. In the first part of the study an in-depth review of the available literature was carried out, to highlight ethical issues connected with Big Data. Five actions were devised as tools to strike the balance described above.

The second phase of the study involved interviews with a number of key stakeholders and conducting a survey that acquired information on the general knowledge of the issues connected to the use of Big Data. Feedback on the proposed balancing actions was also sought and taken into consideration in the final analysis. Attitudes as emerged from interviews and survey most often ranged from concerned to worried, while benefits of Big Data were seldom discussed by the respondents. Benefits are, nevertheless, intrinsic to Big Data, as well as risks, and they are discussed more broadly throughout the study.

2017


Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector
MAS
Monetary Authority of Singapore

2018

Laws on Personal Data are important …

… But paraphrasing Bill Clinton's US Presidential campaign in 1991 …

It is the algorithm, stupid !

Analyzing Personality through Social Media Profile Picture Choice - 2016

"We concluded that each personality trait has a specific type of profile picture posting. Users that are either high in openness or neuroticism post less photos of people and when these are present, they tend not to express positive emotions. The difference between the groups is in the aesthetic quality of the photos, higher for openness and lower for neuroticism …"

(a) Extraverted.

(b) Conscientious.

☹

Figure 1: Example Twitter profile pictures for users scoring high in a personality trait.

Understanding Personality through Social Media - 2013

(a) Top correlated unigram words for Thinking

(b) Top correlated unigram words for Feeling

(c) Top correlated bigram phrases for Introversion

(d) Top correlated bigram phrases for Extroversion

Figure 1: Correlations between n-gram words and phrases, and the personality traits

☹

"For instance, extroverts tend to use hash-tag and phrases like "so proud", "so excited", and "can't wait". People who like to use emoticon are more likely to be Sensing and Feeling personality type. Moreover, we investigate the predictive power of individual features and combined features in our analysis. With the concatenation of all the features we extracted, we can predict the personality traits with an average AUC of 0.661."

Social Theories are *self-fulfilling*

The nature of the liaison between a theory and practice in *social domains* (for example, enterprises and organizations) explains that the effects of a theory, unlike in other sciences, **become self-fulfilling**

A physical sciences theory, right or wrong, does not change the behavior of the subject being studied. ***But managers adapt their behaviors to doctrines and people adapt their behaviors to what you measure them for …***

Example: "*A theory that assumes that people can behave opportunistically and draws its conclusions for managing people based on that assumption can induce managerial actions that are likely to enhance opportunistic behavior among people*" (Ghoshal & Moran, 1996)

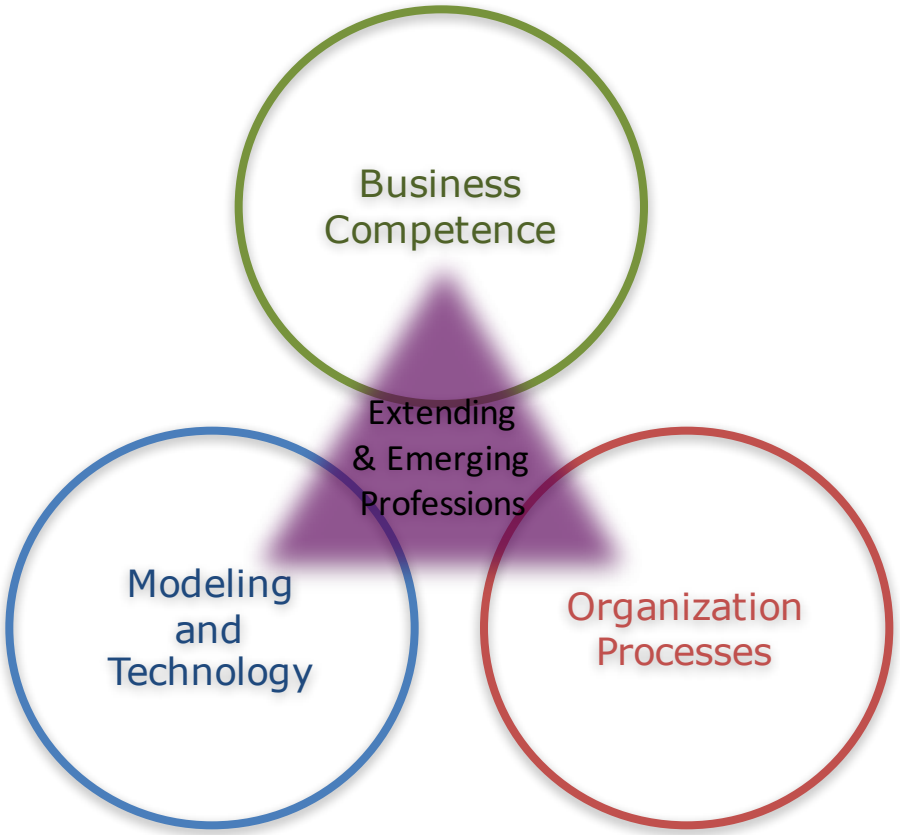
Automated Decisions - *GDPR says NO !*

The data subject should have the right ***not to be subject to a decision****, which may include a measure, evaluating personal aspects relating to him or her which is* ***based solely on automated processing*** *and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices* ***without any human intervention****.*

Such processing includes 'profiling' that consists of any form of automated processing of personal data evaluating the personal aspects relating to a natural person, in particular to analyse or predict aspects concerning the data subject's performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, where it produces legal effects concerning him or her or similarly significantly affects him or her.

However, decision-making based on such processing, including profiling, should be allowed where expressly authorised by Union or Member State law to which the controller is subject, including for fraud and tax-evasion monitoring and prevention purposes conducted in accordance with the regulations, standards and recommendations of Union institutions or national oversight bodies and to ensure the security and reliability of a service provided by the controller, or necessary for the entering or performance of a contract between the data subject and a controller, or when the data subject has given his or her explicit consent. In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision.

Knowledge Areas building Business Analytics



Business Competence

- Finance, accounting, marketing, supply chain, HR, channels, IT, customer relationship …
- Industry-specific competences: underwriting, fraud, claim life-cycle, product design, wealth management, traffic …

Organization Processes and Behavior

- The design and transformation of work processes
  - ✓ Decision-making processes
  - ✓ Strategy processes
  - ✓ Operational processes
- Human processes and motivations; how information may guide better behavior

Modeling and Technology

- Stochastic Models, Operations Research (and related tools: R, SAS, SPSS, …)
- IoT and data generation sources (eg: GPS locator, Surveillance cameras, ATMs, etc)
- Systems in support of Cognitive and Information processes (Watson, HANA, etc)

By judiciously linking data to behavioral provenance, we leverage digital footprints as a way to understand social systems better

Concerned with
Organized Complexity

Concerned with
Unorganized Complexity

An honorable (and generous) reconciliation proposed

Summary and Take-Away, I

- Data Analytics is a source of economic and social value and thus, very exciting
- Large quantities and variety of data bring bigger even opportunities - more excitement !
- Unorganized complexity systems are generally to benefit from Data Analytics *"safely"*
  - ✓ Most IoT in manufacturing and other industries
- However, systems involving individuals are a whole different matter and Data Analytics could lead to very ***questionable practices***
- Injudicious applications of Data Analytics can be ***embarrassing***, ***harassing*** and even ***severely damaging*** to human beings
  - ✓ We saw a number of examples in class
  - ✓ We learned some deep warnings from Nobel Prize winners

Summary and Take-Away, II

- The reasoning framework is common throughout IS1103, i.e.:
  - ✓ *Is it LEGAL or ILLEGAL ?*
  - ✓ *Is it aligned to my values or not ?* (ethical)
  - ✓ *Is it likely to have an impact on our society and seed undesirable or desirable effects ?*
- Before judging any enterprise intervention as unethical, we analyze the potential degree of flawed analytics carried out under "*the pretense for scientistic, data-driven attitude*"
  - ✓ Data are not *raw* in any socio-technical system; observations are designed and metrics chosen as part of a belief (biased, on top)
  - ✓ We question the validity of what we are told and demand for evidence (analytics may be *silly wrong* before *intentionally bad*)
- Emerging recommendation and guidance frameworks for assisting enterprises to assess a suitable balance between data monetization and negative side-effects on people
  - ✓ Should all organizations explicitly state their respect for individuals with a commitment of ***not putting people under a digital scope*** *beyond the oversimplification of "asking for consent"* ?

Otherwise …

"We have indeed at the moment little cause for pride: as a profession we have made a mess of things"

Von Hayek, **The Pretense of Knowledge**, 1974 [dissertation - Nobel Prize in Economics]

Reading and Literature

Asadi, S, I., Breidbach , F, C., Davern, M., Shanks , G., “Ethical Implications of Big Data Analytics”, *Twenty-Fourth European Conference on Information Systems (ECIS*), 2016.

Buchanan, E., "Considering the ethics of big data research: A case of Twitter and ISIS/ISIL", *PLoS One,* vol. 12, no. 12, 2017.

DeLallo, L. 2018, “AI business applications and ethical issues in focus at O’Reilly AI conference” , *Mckinsey Digital*, https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/digital-blog/ai-business-applications-and-ethical-issues-in-focus

Erickson, L, C., Harris, E, N. & Lee, M,M. 2018, “It’s Time to Talk About Data Ethics”, *Oracle+DataScience.com*, https://www.datascience.com/blog/data-ethics-for-data-scientists

European Economic and Social Committee ,“The ethics of Big Data: Balancing economic benefits and ethical questions of Big Data in the EU policy context”, https://www.eesc.europa.eu/resources/docs/qe-02-17-159-en-n.pdf

Fung, K., “The Ethics Conversation We’re Not Having About Data” , *Harvard Business Review.,* 2015.

Light, B., Mitchell, P. & Wikström, P., "Big Data, Method and the Ethics of Location: A Case Study of a Hookup App for Men Who Have Sex with Men", *Social Media + Society,* vol. 4, no. 2, 2018.

Loukides, M., Mason,H. & Patil, DJ. 2018, “Case studies in data ethics“, O’Reilly Media, Inc. https://www.oreilly.com/ideas/case-studies-in-data-ethics

Martin, K. “Ethical Issues in the Big Data Industry”, MIS Quarterly Executive, 2015.

Nersessian, D. 2018, "The law and ethics of big data analytics: A new role for international human rights in the search for global standards", *Business Horizons,* [Online], vol. 61, no. 6, pp. 845-854.

Newell S., Marabelli, M. “Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of ‘datification’”, Journal of Strategic Information Systems, 24, 2015.

O’Leary, E, D. 2016, “Ethics for Big Data and Analytics”, *University of Southern California,* http://raw.rutgers.edu/docs/wcars/38wcars/Presentations/Dan%20-%20Ethics%20for%20Big%20Data%20and%20Analytics.pdf

Pauleen, D.J., Rooney, D. & Intezari, A., "Big data, little wisdom: trouble brewing? Ethical implications for the information systems discipline", *Social Epistemology,* vol. 31, no. 4, pp. 400-416, 2017.

Trustees of Princeton University, 2019, “Dialogues on AI and Ethics”, https://aiethics.princeton.edu/case-studies/case-study-pdfs/