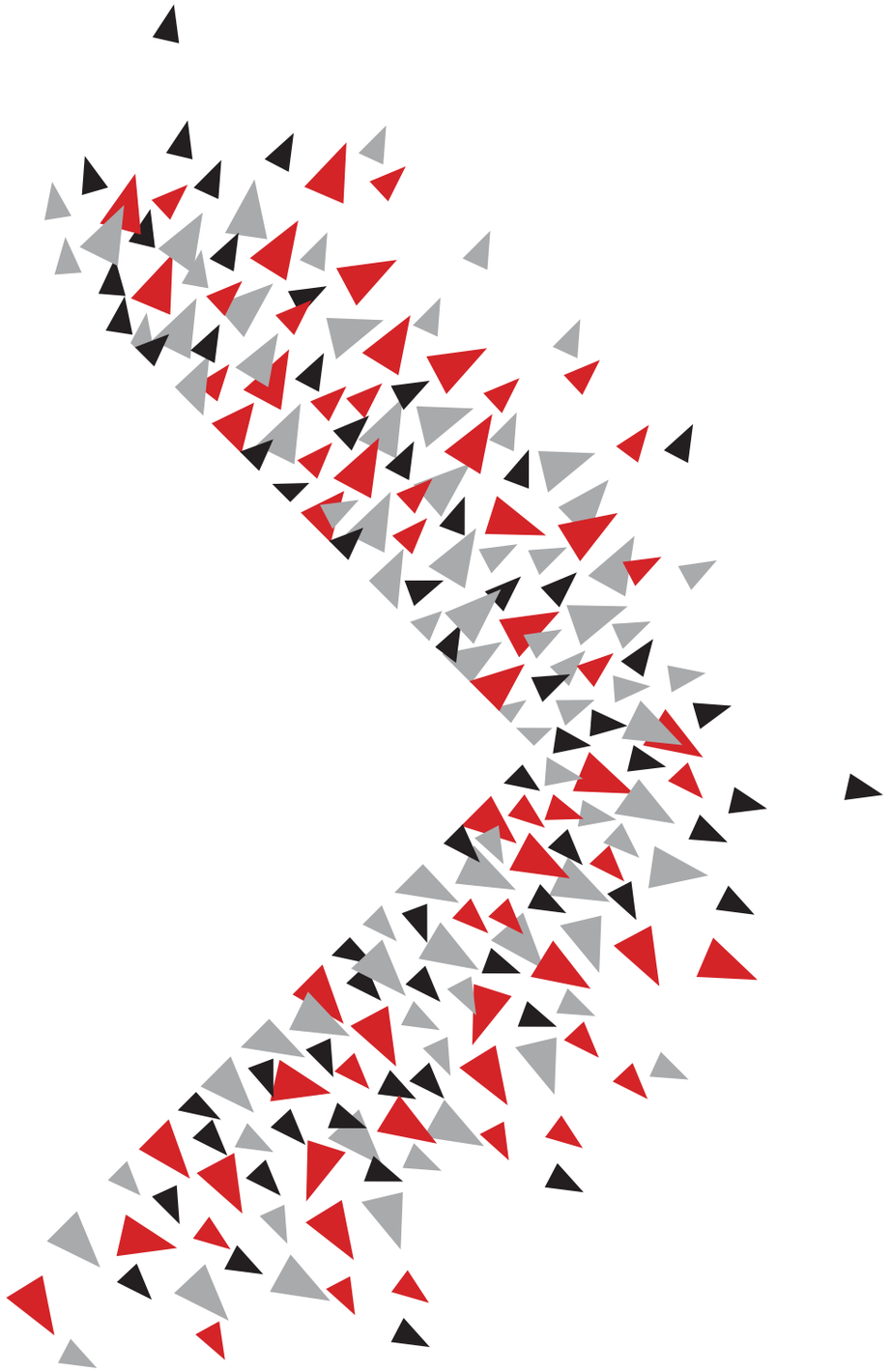
BIG DIVE

TECH. CUSTOM EDITION

A project by TOP-IX
designed for Intesa Sanpaolo



www.bigdive.eu / @bigdive_eu / #bigdive

06102016

WELCOME!

BIG DIVE IS A HANDS-ON INTENSIVE TRAINING PROGRAM AIMED AT BOOSTING THE TECH SKILLS IN ORDER TO EXTRACT VALUE FROM DATA AND TO GENERATE IMPACT.

CHRISTIAN
RACCA

@pendolare

TO DATA BEERS

dati.piemonte.it

top IX

BIG DIVE

PIEMONTE
VISUAL CONTEST

top ix

torino
piemonte
internet
exchange

THE CONSORTIUM

NON PROFIT CONSORTIUM

MISSION

/ MANAGE IX IN THE NORTH-WEST OF ITALY

/ BOOST "INNOVATION" & GENERATE LOCAL IMPACT BY LEVERAGING INFRASTRUCTURE ASSETS

16 EMPLOYEES

6 EXTERNAL COLLABORATORS

80+ MEMBERS

PUBLIC AND PRIVATE PARTICIPATION

CORE ACTIVITIES

TO MANAGE, DEVELOP AND SCALE INTERNATIONALLY THE

NAP (NEUTRAL ACCESS POINT)

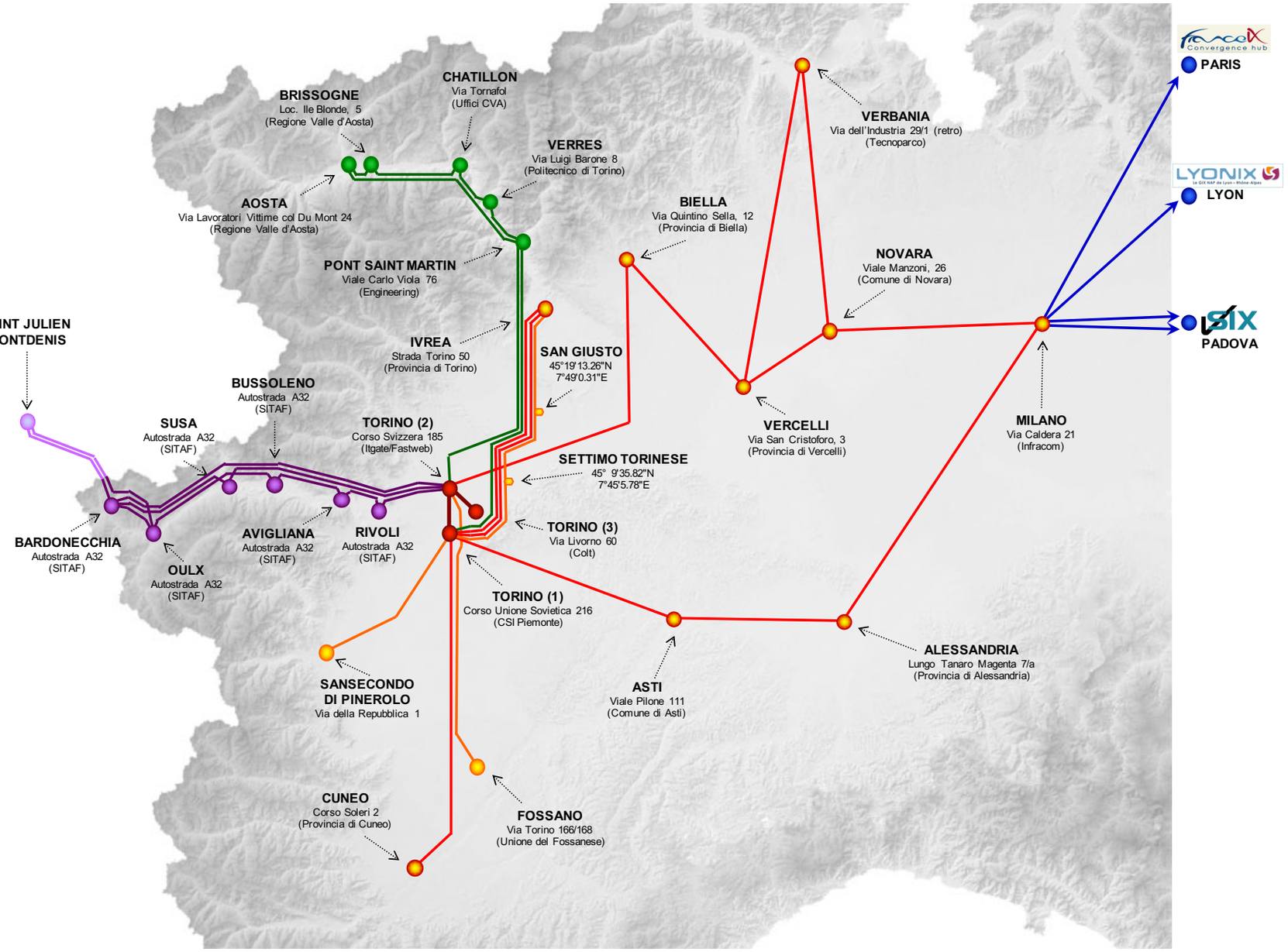
MODEL AIMED AT INTERNET TRAFFIC EXCHANGE

DEVELOPMENT PROGRAM

IS A TOOL TO FOSTER INNOVATION, ENTREPRENEURSHIP AND KNOWLEDGE SHARING IN ORDER TO GENERATE SOCIAL AND ECONOMIC IMPACT

NEUTRAL INFRASTRUCTURE

TOP-IX BACKBONE 30 MAIN NODES



top IX | DEVELOPMENT PROGRAM

EDUCATION

Hands-on training for data-scientists, software developers, city-makers.

CORPORATE INNOVATION

Open innovation projects focused on companies.

START-UP

Infrastructural support and prototype development.

FUNDED PROJECTS

EU **funded projects** and other grants.

SOCIAL INNOVATION

Tools & actions to support" **evidence based policy-making**" and sustainable growth.

PUBLIC POLICY

Supporting **public policies** concerning digital tech.

IX + **IAAS** + **CLOUD**

DATA

top ix | DP - EDUCATION

EDUCATION

Hands-on training
for data-scientists,
software developers,
city-makers.

HACKADEMY BIG DIVE PLEXUM

BIG DIVE

HACKING DEVELOPMENT,
VISUALIZATION & SCIENCE

top ix torino piemonte internet exchange

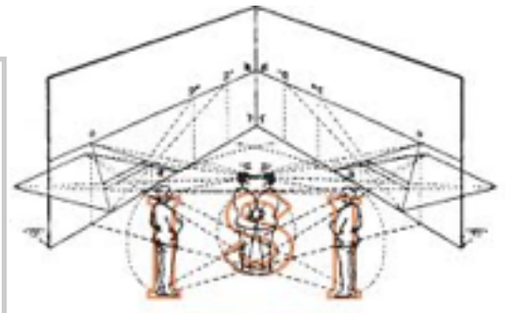
AXANT

Management
+ coding

Coding

Science

Viz



INSTITUTE
FOR SCIENTIFIC INTERCHANGE
FOUNDATION

TODO

BIG DIVE

TECH. CUSTOM EDITION

top ix torino piemonte internet exchange

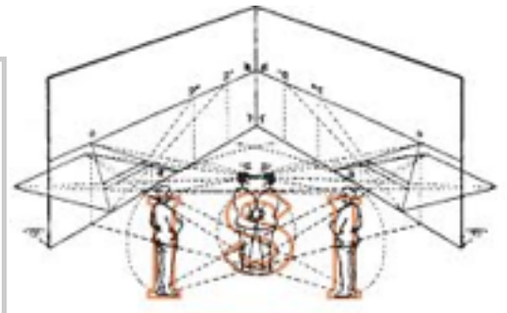
Management
+ coding

Science

Science

Viz

aizoOn® AUSTRALIA EUROPE USA
TECHNOLOGY CONSULTING



INSTITUTE
FOR SCIENTIFIC INTERCHANGE
FOUNDATION

TODO

BIG DIVE HISTORY

2012

4 weeks
20 divers
8 countries

2013

5 weeks
15 divers
5 countries

2014

5 weeks
15 divers
6 countries

2015

5 weeks
19 divers
6 countries

2015

Custom
Executive
#finance

2016

5 weeks
22 divers
5 countries

+ other 1-day events for corporates

2016

Technical
Custom

YOU!

YOUR COURSE PROGRAM

October 2016

Mondar	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
					1	2
3	4	5	6 Kick-off VERSIONING + PYTHON BASICS	7 PYTHON BASICS DATAVIZ THEORY & INTRO	8	9
10	11	12	13 DATAVIZ PRACTICE	14 DATAVIZ PRACTICE	15	16
17	18	19	20 DATAVIZ PRACTICE	21 DATAVIZ CLOSING SCIENTIFIC PYTHON	22	23
24	25	26	27 STATISTICS THEORY DATA ANALYSIS	28 MAPREDUCE & MrJob STATISTICS THEORY	29	30

TOP-IX
TODO
Aizoon
ISI

YOUR COURSE PROGRAM

November 2016

Mondar	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
31	1	2	3 MAPREDUCE & AWS MAPREDUCE & MONGODB	4 MAPREDUCE & MONGODB SPARK DEMO	5	6
7	8	9	10 MACHINE LEARNING THEORY & PRACTICE	11 MACHINE LEARNING PRACTICE	12	13
14	15	16	17 MACHINE LEARNING EXERCISE	18 MACHINE LEARNING DEV TOOLS FOR NET SCIENCE	19	20
21	22	23	24 DEV TOOLS FOR NET SCIENCE NET SCIENCE	25 NET SCIENCE NET SCIENCE EXERCISE WRAP-UP	26	27

TOP-IX
TODO
Alzoon
ISI

YOUR COURSE PROGRAM

November 2016

Mondar	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
31	1	2	3 MAPREDUCE & AWS MAPREDUCE & MONGODB	4 MAPREDUCE & MONGODB SPARK DEMO	5	6
7	8	9	10 MACHINE LEARNING THEORY & PRACTICE	11 MACHINE LEARNING PRACTICE	12	13
14	15	16	17 MACHINE LEARNING EXERCISE	18 MACHINE LEARNING DEV TOOLS FOR NET SCIENCE	19	20
21	22	23	24 DEV TOOLS FOR NET SCIENCE NET SCIENCE	25 NET SCIENCE NET SCIENCE EXERCISE WRAP-UP	26	27

+ **FINAL PROJ. (T.B.D.)**

DAILY SCHEDULE

09:15 - 09:30	GATHERING
09:30 - 11:00	LESSON & EXERCISE
11:00 - 11:30	BREAK
11:30 - 13:30	LESSON & EXERCISE
13:30 - 14:30	LUNCH BREAK
14:30 - 16:30	LESSON & EXERCISE

TEACHERS

#datascience
ANDRÈ PANISSON
(ISI)

#datascience
MICHELE TIZZONI
(ISI)

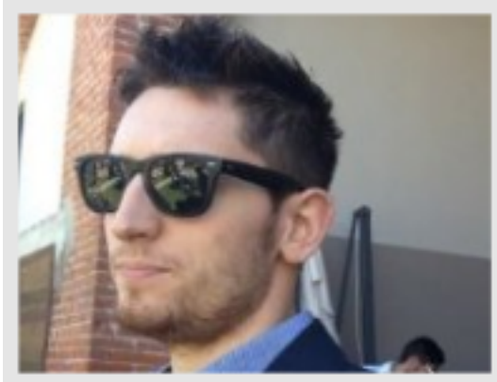
#datascience
LAETITIA GAUVIN
(ISI)

#dataviz
FABIO FRANCHINO
(TODO)

#datascience
PAOLO BAJARDI
(AIZOON)

#datascience
ALAN PEROTTI
(AIZOON)

#coding
ALEX COMUNIAN
(TOP-IX)



#coding
ANDREA BECCARIS
(TOP-IX)

BIG DIVE X FACTOR

YOU!

/ WHO ARE YOU ?
/ YOUR GOALS ?
/ YOUR “SUPER”-POWERS ?

THE VENUE

ACCELERATOR OF KNOWLEDGE AND ENTREPRENEURSHIP FOR SOCIAL IMPACT

Rinascimenti Sociali

PRACTICAL INFORMATION

/ VENUE OPENING HOURS (9:00 AM - 6:30 PM)

/ KITCHEN AVAILABLE

/ COFFEE MACHINE (COINS NEEDED) + WATER (FREE)

/ INTERNAL COMMUNICATION TOOL

/ GITHUB REPOSITORY

/ VIRTUAL DESKTOPS

/ HOMEWORK + EVALUATION

FIRST ASSIGNMENT

PLEASE SEND AN EMAIL TO INFO.BIGDIVE@TOP-IX.ORG

FROM THE ADDRESS YOU WANT TO USE FOR THE COURSE AND PUT INSIDE YOUR GITHUB ACCOUNT

FINAL RECOMMENDATIONS

/ THIS IS A HANDS-ON COURSE (DO NOT EXPECT TOO MUCH THEORY CONTEXT)

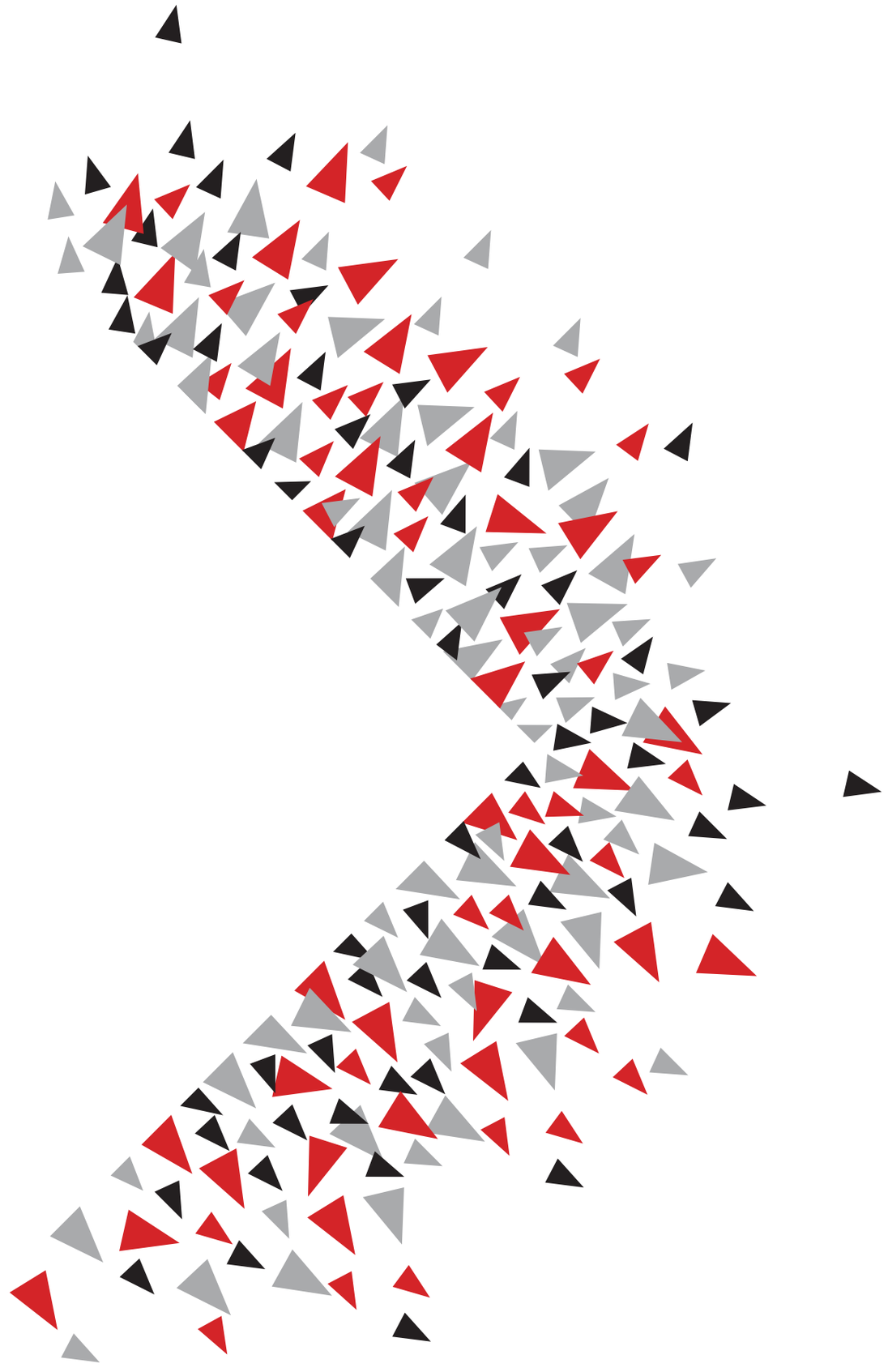
/ BE PREPARED TO OVERLOAD (TRACKING YOUR PROGRESS)

/ PLEASE RESPECT LESSON TIME AND SHARED SPACES

/ DO NOT HESITATE TO ASK IF YOU HAVE PROBLEMS

/ TRY TO HAVE FUN!

Q&A?



FROM BIG DATA TO IMPACT

DISCLAIMER: I AM NOT A DATA SCIENTIST !

BIG DATA = BUZZWORD

L'AMBIGUITÀ DEL TERMINE

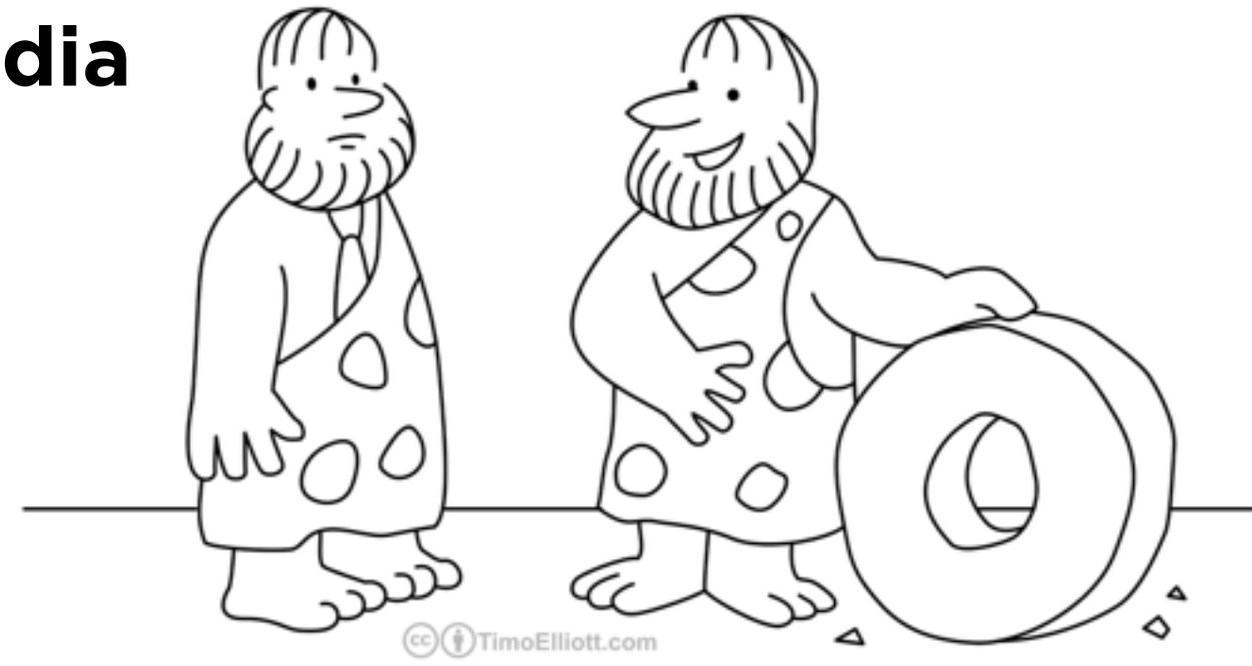
quantitativo
marketing
tecnologia
ricerca

VS

qualitativo
innovazione
processo
business

FALSI MITI

1. Big Data is Only About Massive Data Volume
2. Big Data Means Hadoop
3. Big Data Means Unstructured Data
4. Big Data is for Social Media
5. NoSQL means No SQL



"It does look similar—but this one is powered by Hadoop"

SOURCE: mashable

BIG DATA OLTRE LA BUZZWORD

PERCHÉ OGGI SI PARLA DI BIG DATA?

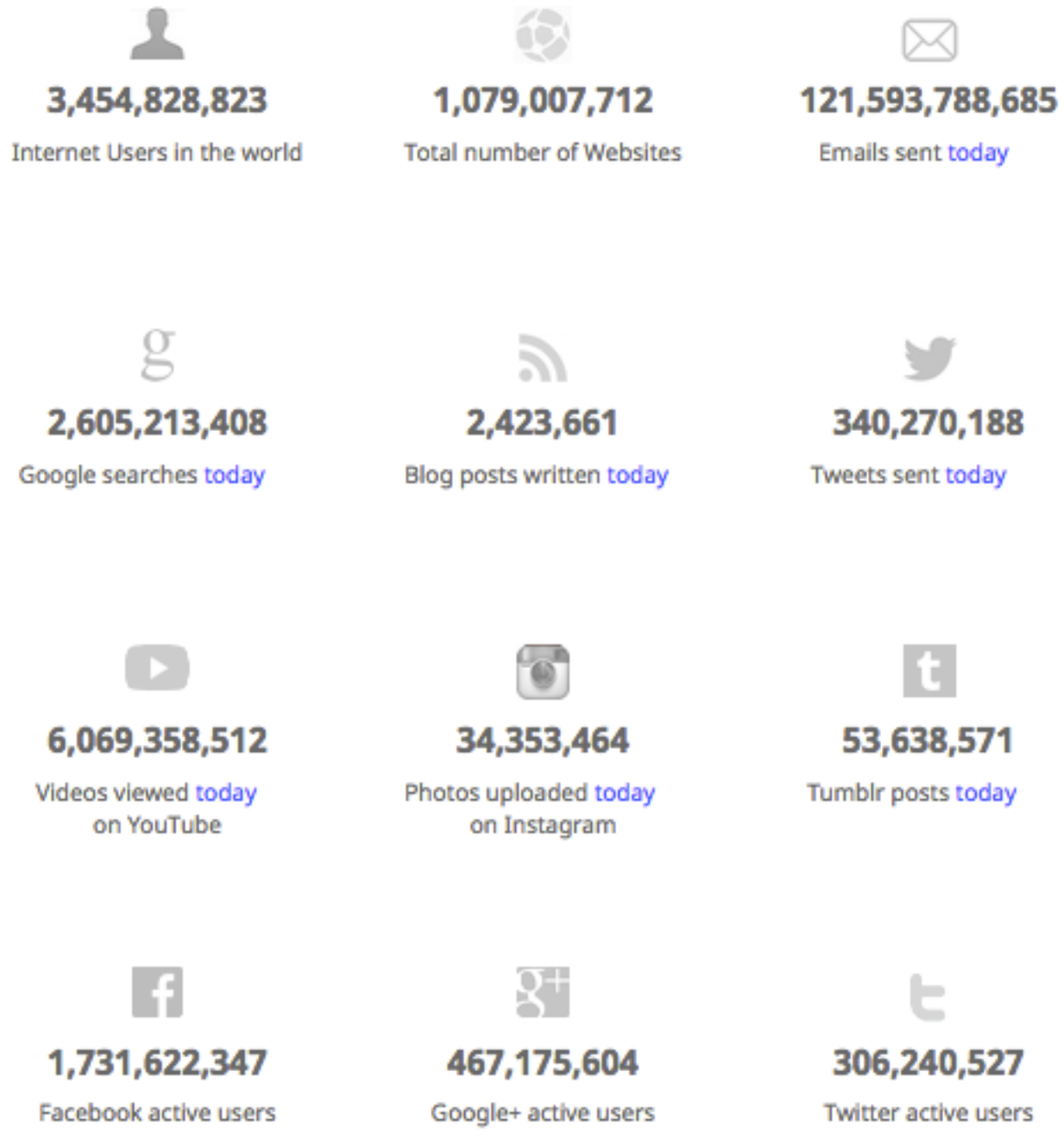
/ disponibilità di dati

/ disponibilità di strumenti

/ approccio - cultura

DATI

RACCOGLIAMO DATI AD UN RITMO ESPONENZIALE



Global Internet traffic in 2019 will be 66 times the volume of the entire global Internet in 2005 (Zettabyte era Cisco report)

I DATI NON SONO TUTTI UGUALI

20%

dati strutturati

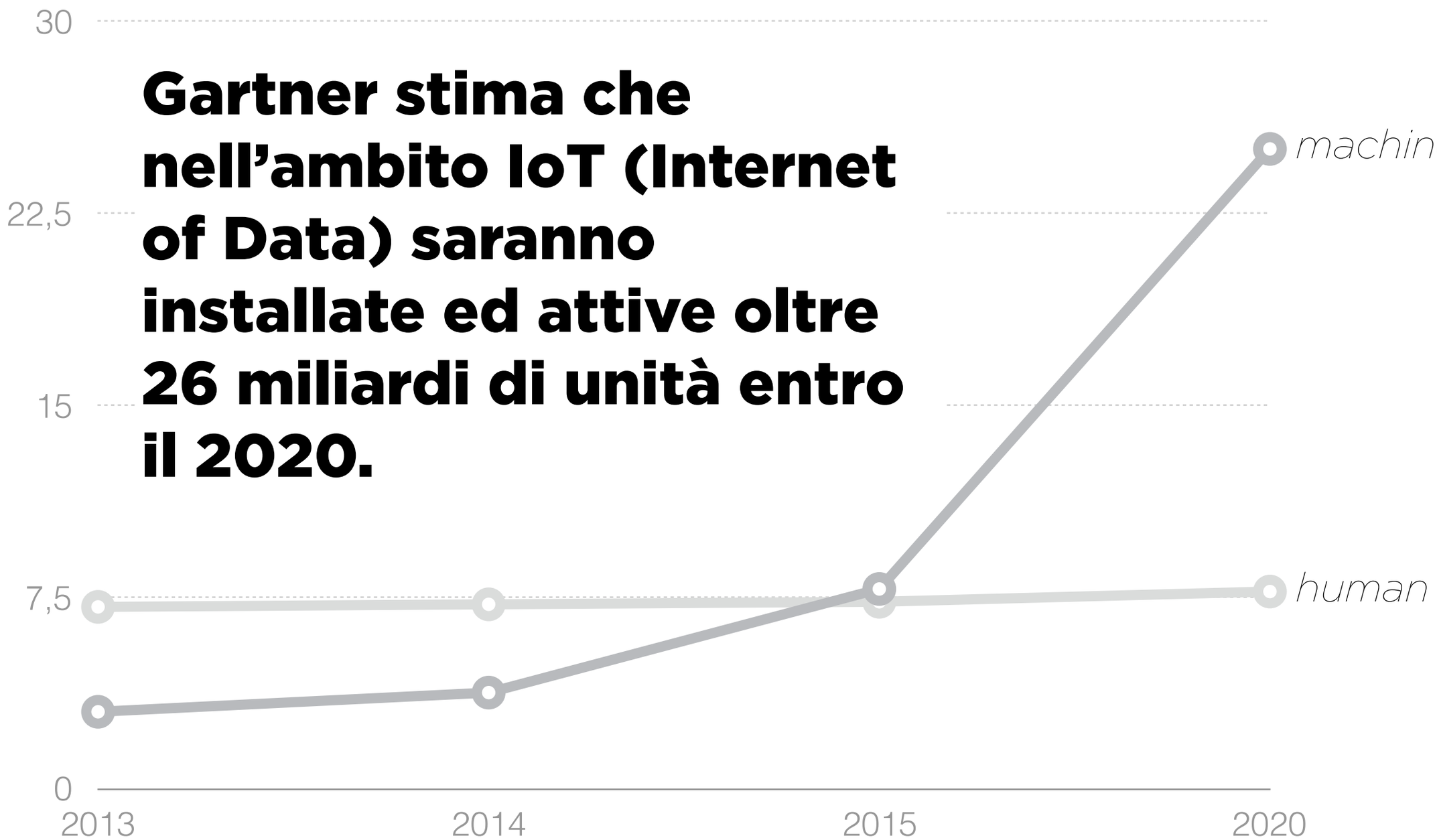
80%

dati non strutturati

- **Semistructured data: weblog, machine log, etc.**
- **Unstructured text data: blog, e-mail, commenti, etc.**
- **Binary data: foto, immagini, audio, video, etc.**

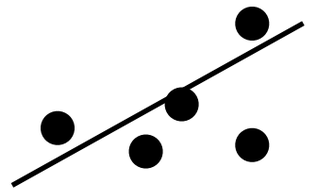
MACHINE-GENERATED DATA VS HUMAN-GENERATED DATA

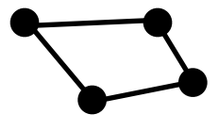
Gartner stima che nell'ambito IoT (Internet of Data) saranno installate ed attive oltre 26 miliardi di unità entro il 2020.

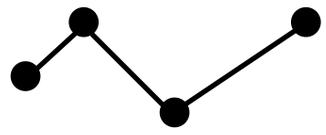


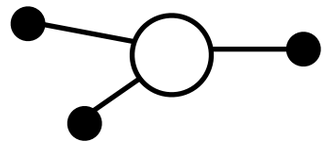
STRUMENTI

LE TECNICHE PER ESTRARRE CONOSCENZA DAI DATI SONO OGGI PIÙ ACCESSIBILI

A. MACHINE LEARNING 

B. NETWORK SCIENCE 

C. DATA VIZ 

D. SEMANTICA / NLP 

INFRASTRUTTURA A SUPPORTO

/ CLOUD COMPUTING

/ HPC - HIGH PERFORMANCE COMPUTING

/ HPN - HIGH PERFORMANCE NETWORKS

/ DATA FRAMEWORKS (HADOOP, SPARK, ...)

APPROCCIO - CULTURA

SHARING ECONOMY

La “sharing economy” è un nuovo modello economico, capace di rispondere alle sfide sociali e di promuovere forme di consumo più consapevoli basate sul riuso invece che sull’acquisto e sull’accesso piuttosto che sulla proprietà.

CONDIVISIONE

PEER-TO-PEER

TECNOLOGIA

“SCIENZA” DELLA COMPLESSITÀ

Peculiarità di un sistema complesso:

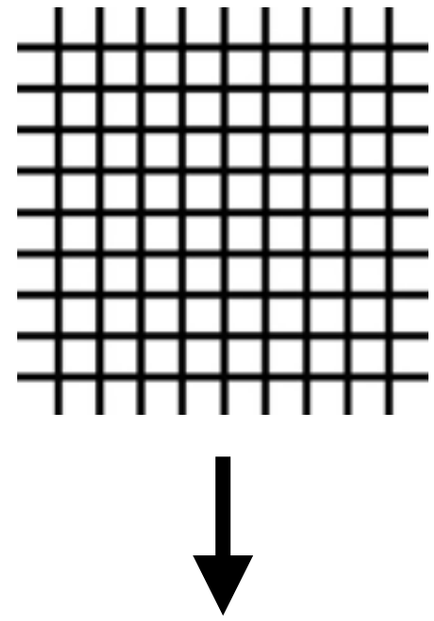
/ CARATTERISTICHE EMERGENTI

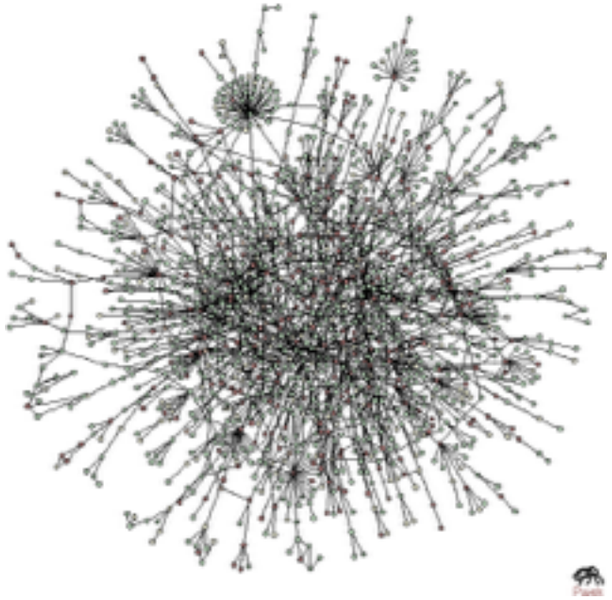
/ AUTO-ORGANIZZAZIONE

/ CONNETTIVITÀ

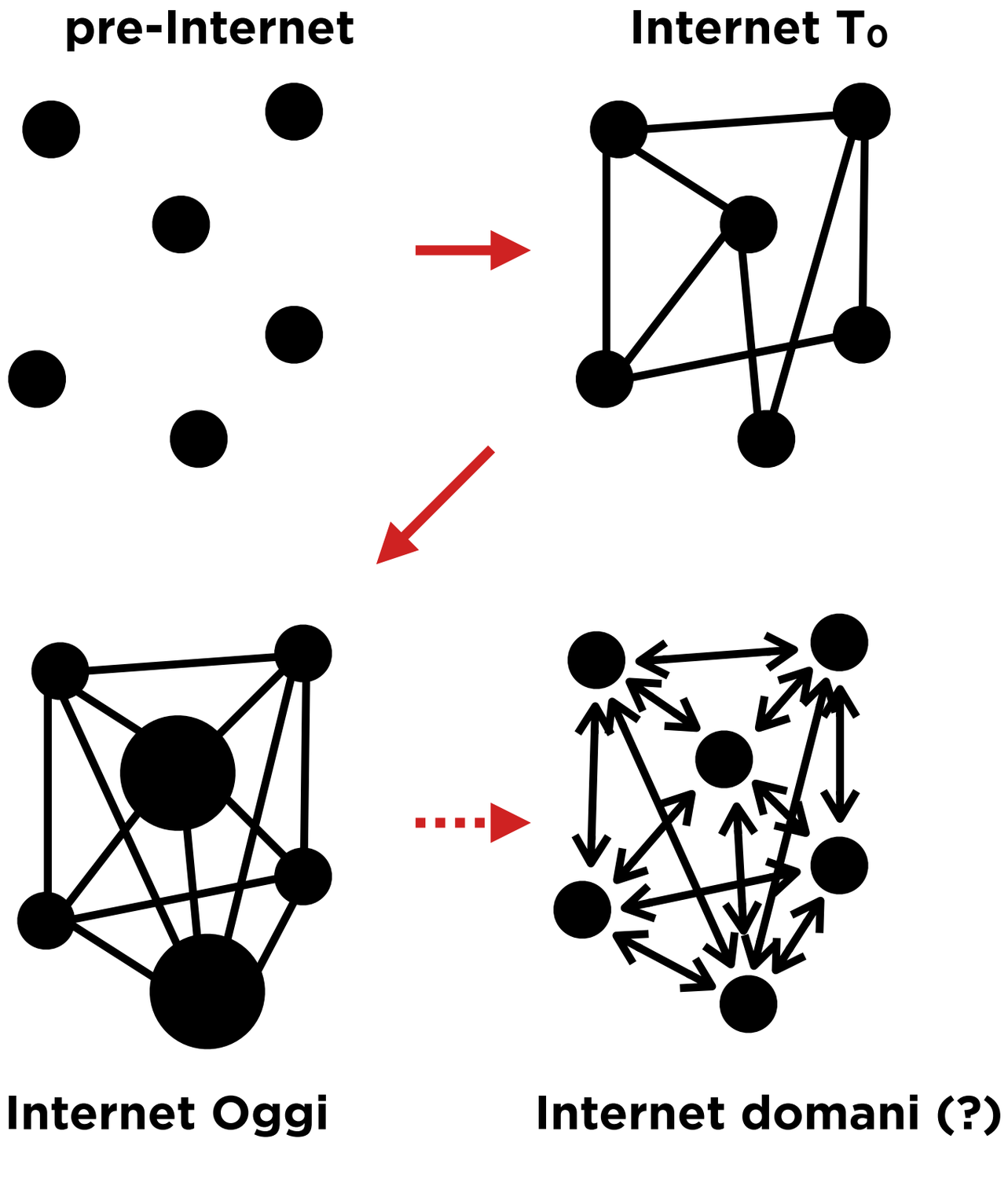
/ ADATTAMENTO

/ NON LINEARITÀ





NETWORK THINKING & SOCIETY



OPEN-INNOVATION

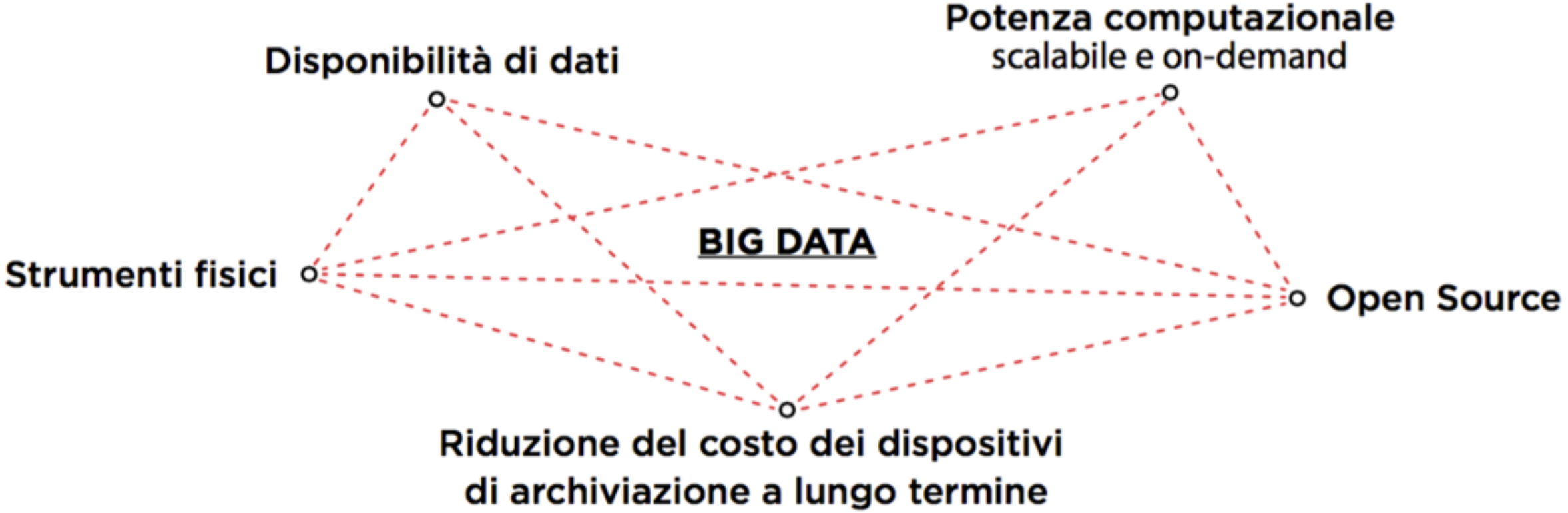
La ricerca e l'attivazione di nuovi paradigmi che spingono verso una apertura nella ricerca di innovazione oltre i "confini" aziendali

APERTURA

CONTAMINAZIONE

CROSS-DISCIPLINARIETÀ

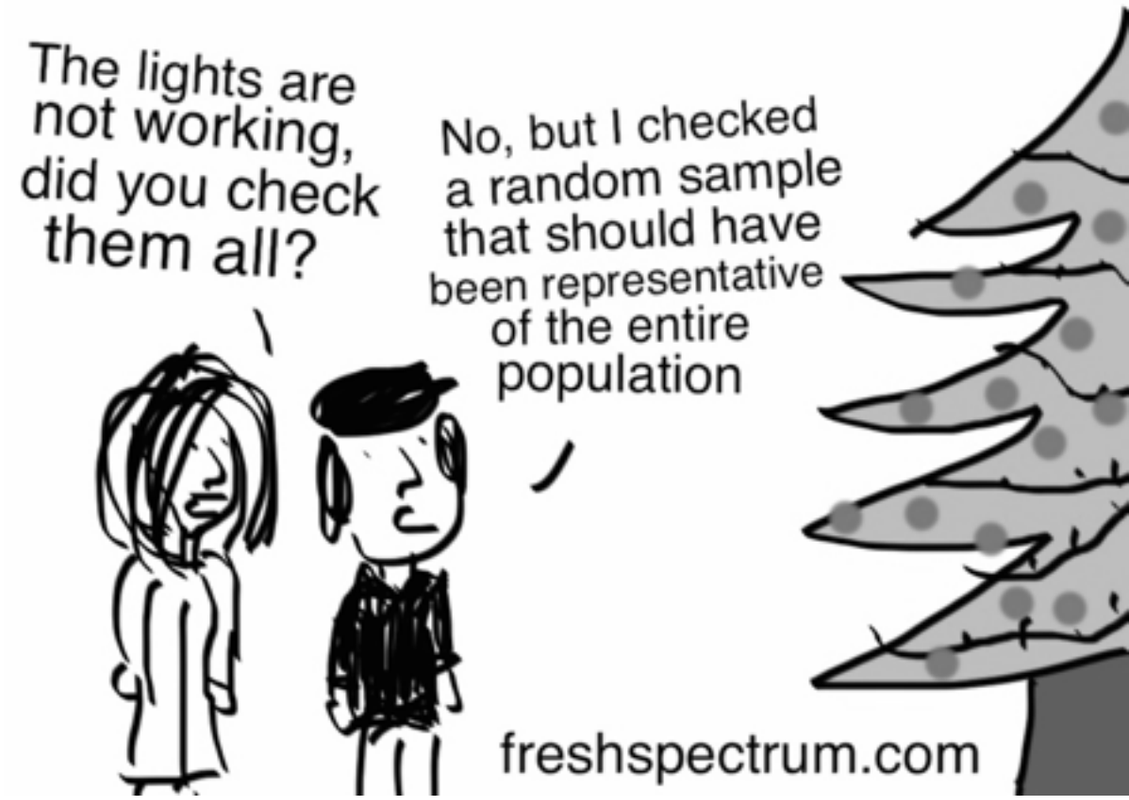
BIG DATA COME CONVERGENZA DI FATTORI



UN FENOMENO DI CONVERGENZA CHE ABILITA UN CAMBIO DI PARADIGMA (O VICEVERSA ?)

CAMPIONAMENTO VS “ALL-DATA”

Conviviamo con il Campionamento. (Censimento, teoria dei segnali…)



BIG DATA ≠ ALL-DATA

Maggior consapevolezza dell’informazione che scartiamo !

PROPENSIONE ALL'ERRORE & BLACK BOXES

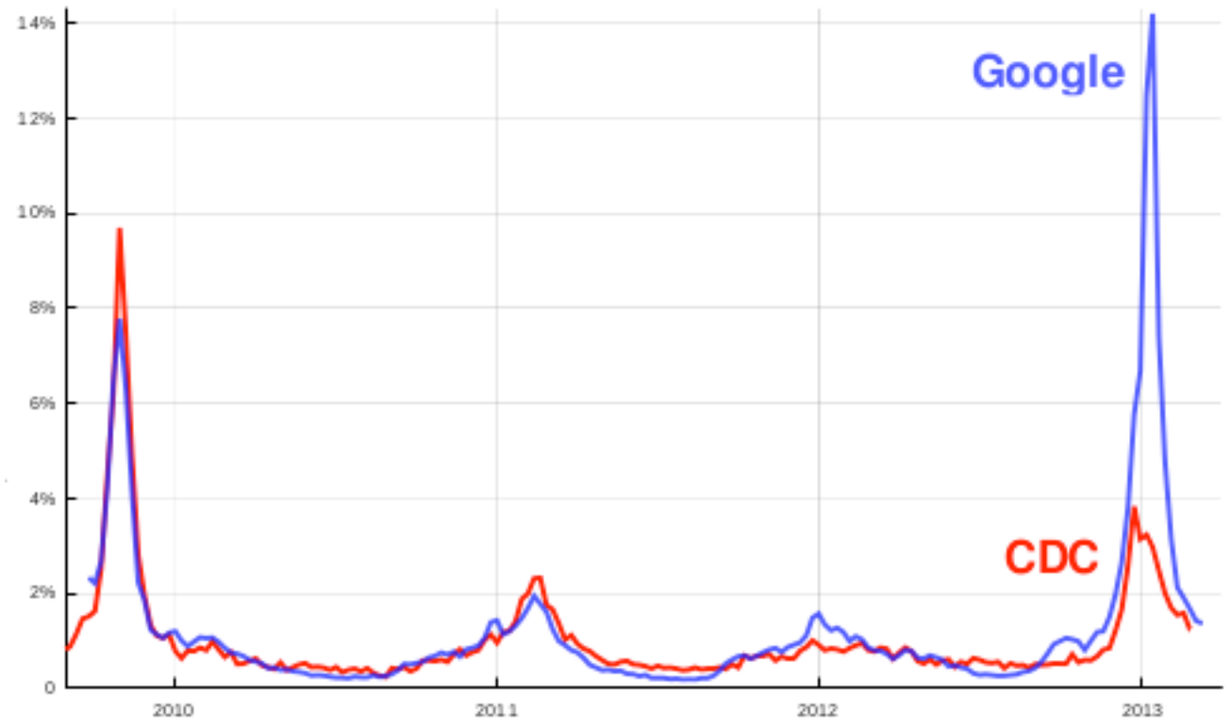
"The goal of ML is never to make "perfect" guesses, because ML deals in domains where there is no such thing. The goal is to make guesses that are good enough to be useful. "

George E. P. Box

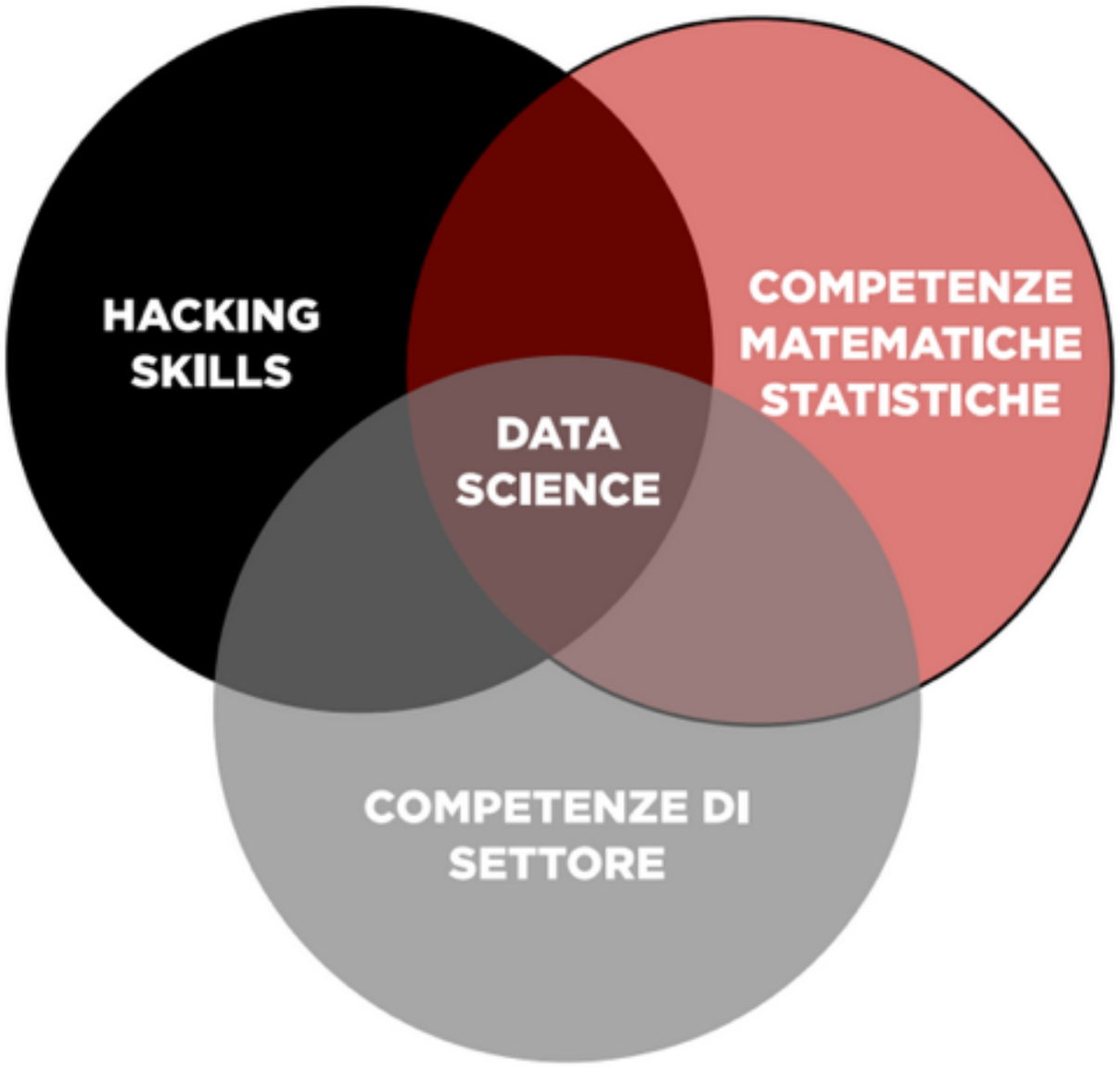
CAUSALITÀ VS CORRELAZIONE

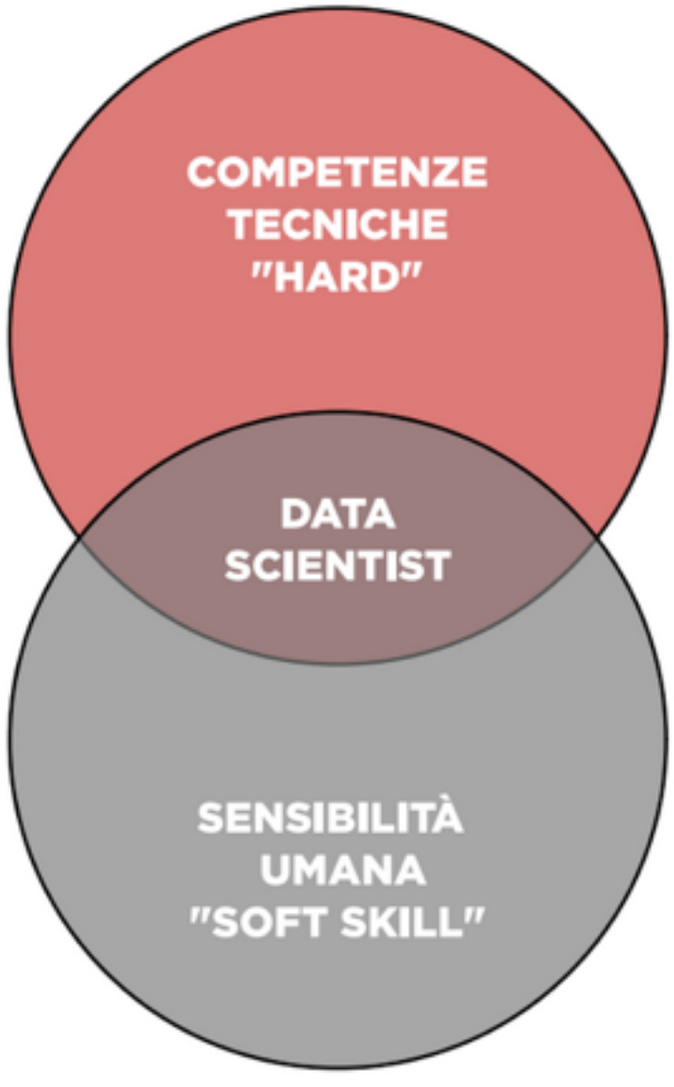
In 2008, researchers from Google explored this potential, claiming that they could "nowcast" the flu based on people's searches. The essential idea, published in a paper in Nature

... then, GFT failed—and failed spectacularly—missing at the peak of the 2013 flu season by 140 percent.



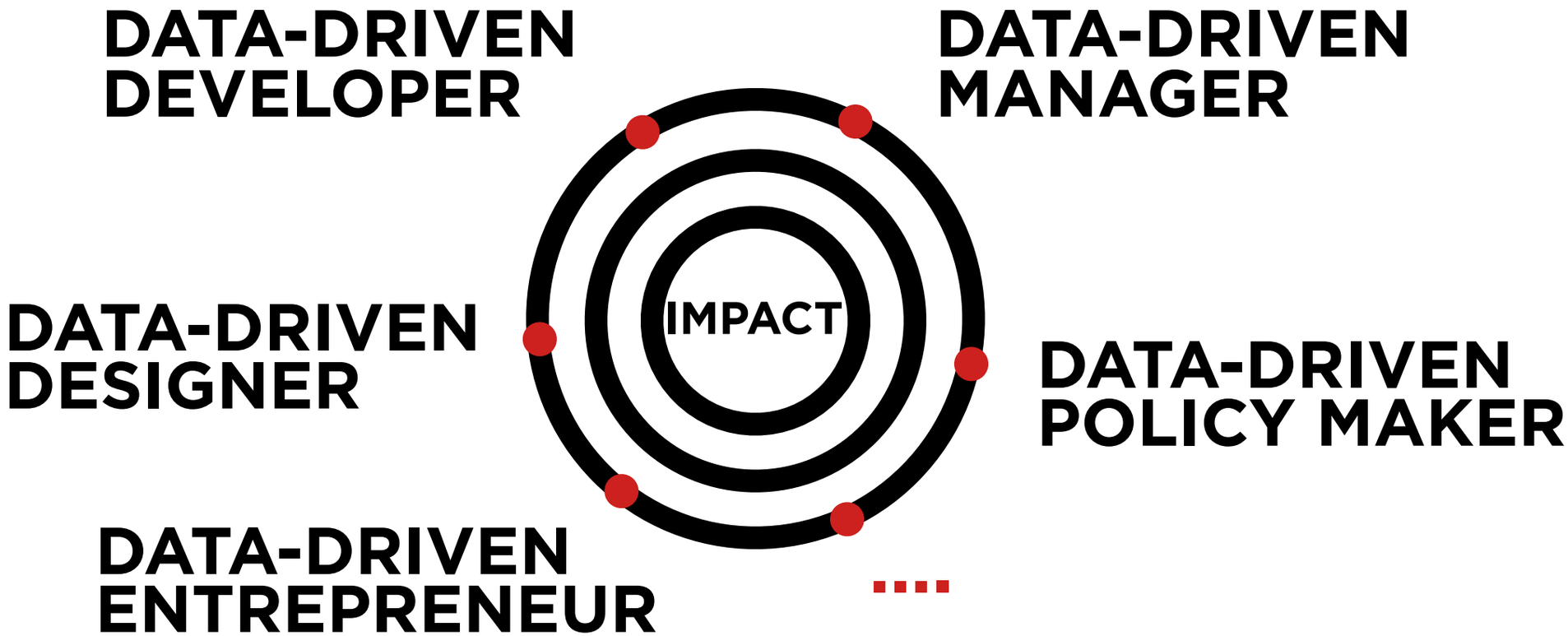
IL RUOLO DEL DATA SCIENTIST





Rielaborazione del diagramma di Venn di Drew Conway

NON SOLO DATA SCIENTIST



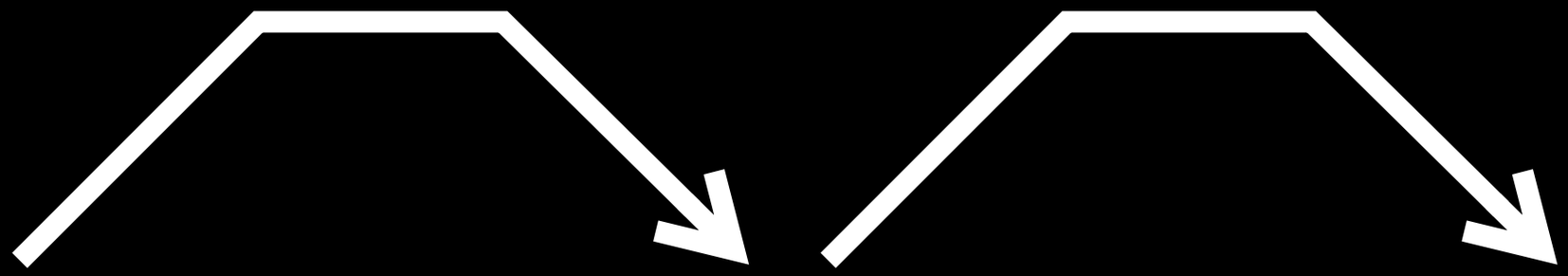
SINGOLO VS TEAM

BIG DATA IN PRACTICE

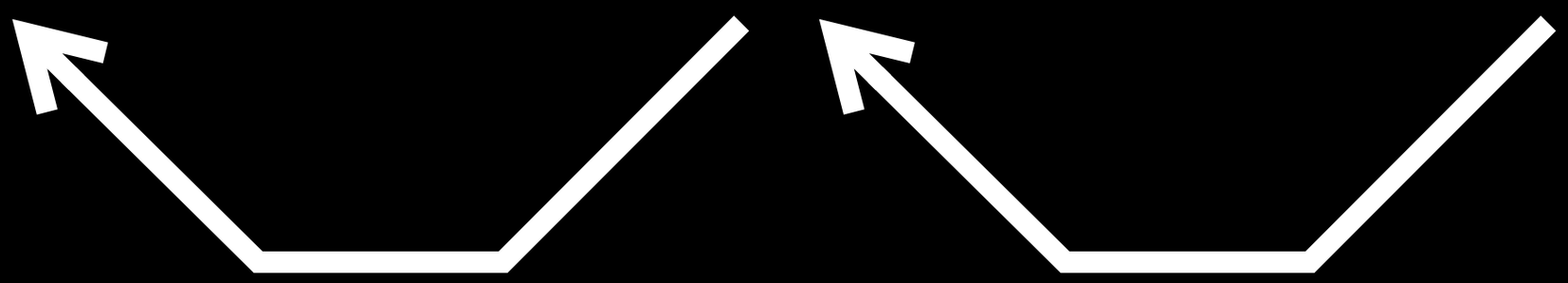
"Any sufficiently advanced technology
is indistinguishable from magic"

Clarke's Third Law

CREARE VALORE - IMPATTO



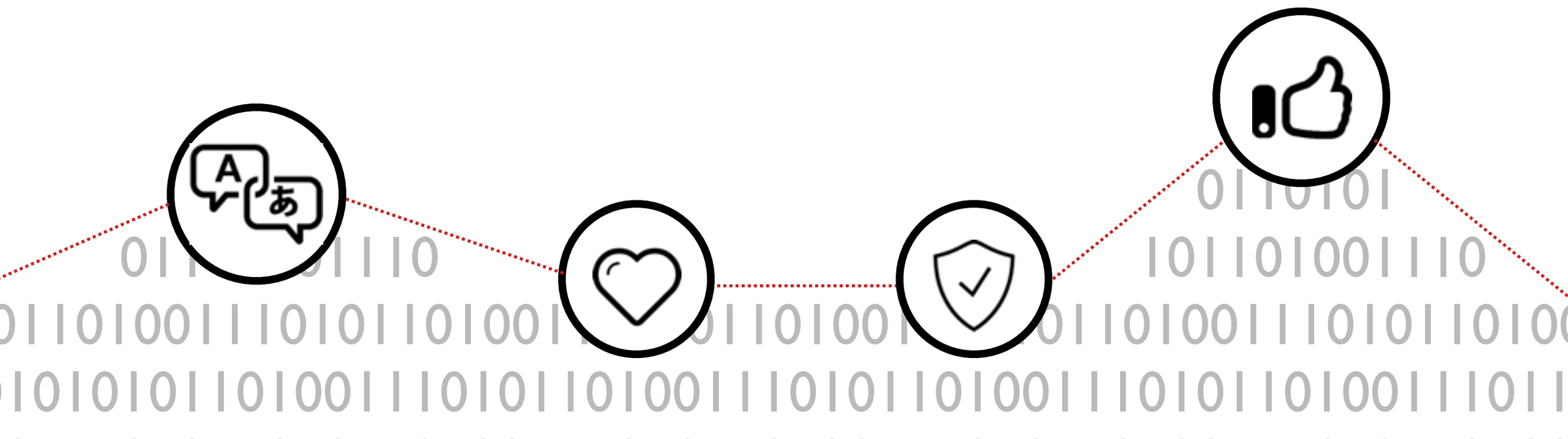
DATI MODELLI DECISIONI



DAI DATI ALL'IMPATTO E RITORNO

ROADMAP TO IMPACT

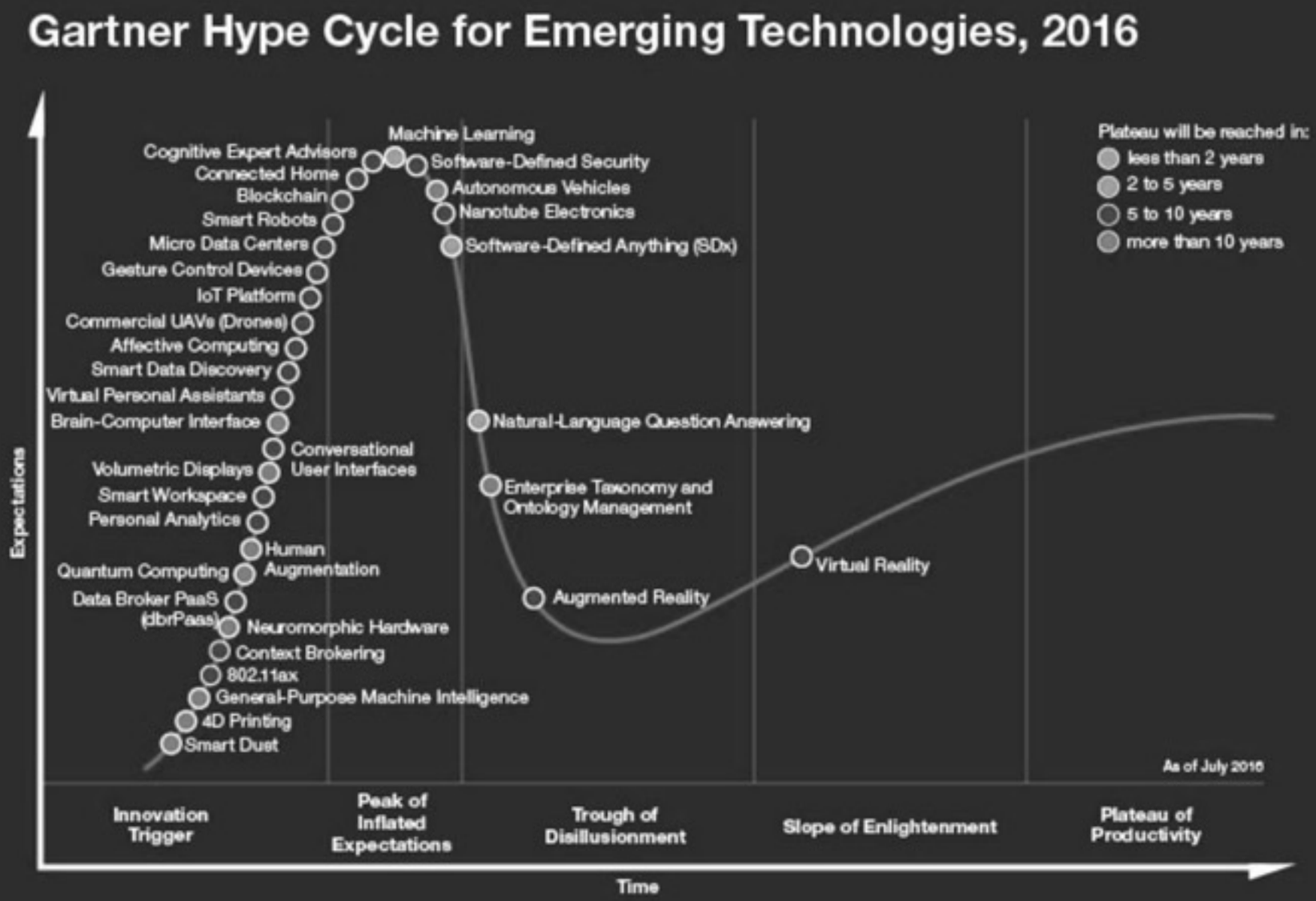
1. Collecting data
2. Enabling data access (e.g. Data Lake)
3. Identifying trends and patterns
4. ...
5. Enabling D-D-D making



TOWARDS ARTIFICIAL DATA INTELLIGENCE

probably WE DON'T need an ORACLE but a human-like agent

TOWARDS ARTIFICIAL DATA INTELLIGENCE

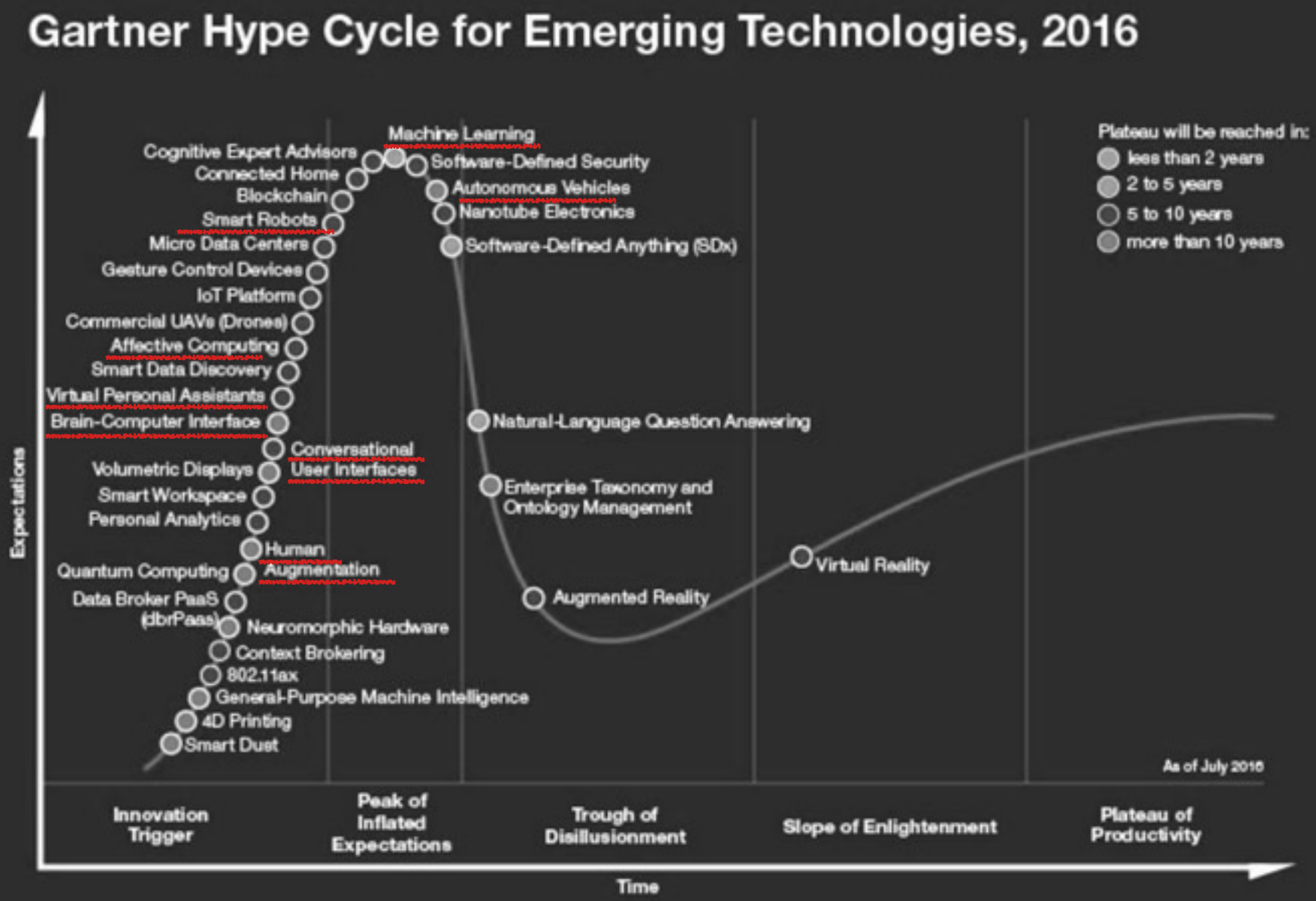


gartner.com/SmarterWithGartner

Source: Gartner

Gartner.

TOWARDS ARTIFICIAL DATA INTELLIGENCE



gartner.com/SmarterWithGartner

Source: Gartner

Gartner.

THANKS TO BIG DATA
WE ARE NOW ABLE TO
TRAIN
ALGORITHMS
LIKE NEVER BEFORE

SELF-DRIVING CARS

AUTOMATIC TEXT GENERATION

READ

"Google AI system has already eaten 2,865 novels to improve language understanding" (may 2015)

WRITE

"A team from Future University Hakodate in Japan built an AI program that wrote a novel called 'The Day a Computer Writes a Novel'."

GAME PLAYING

In March 2016 Google's artificial intelligence program AlphaGo dominated its match with South Korean world Go champion Lee Sedol, winning with a 4-1 score.

COMPLEXITY (# of configurations)

CHESS ~ 10^{50} (brute force)

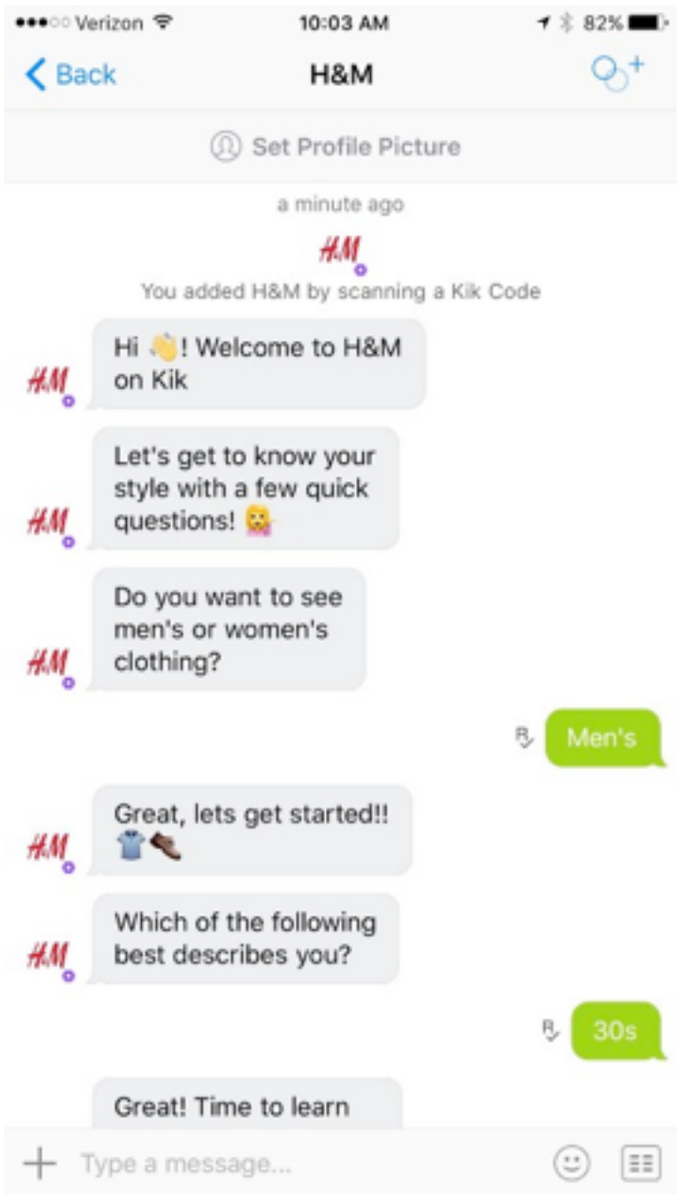
GO (grid 19x19) ~ 10^{170} (learning from experience)

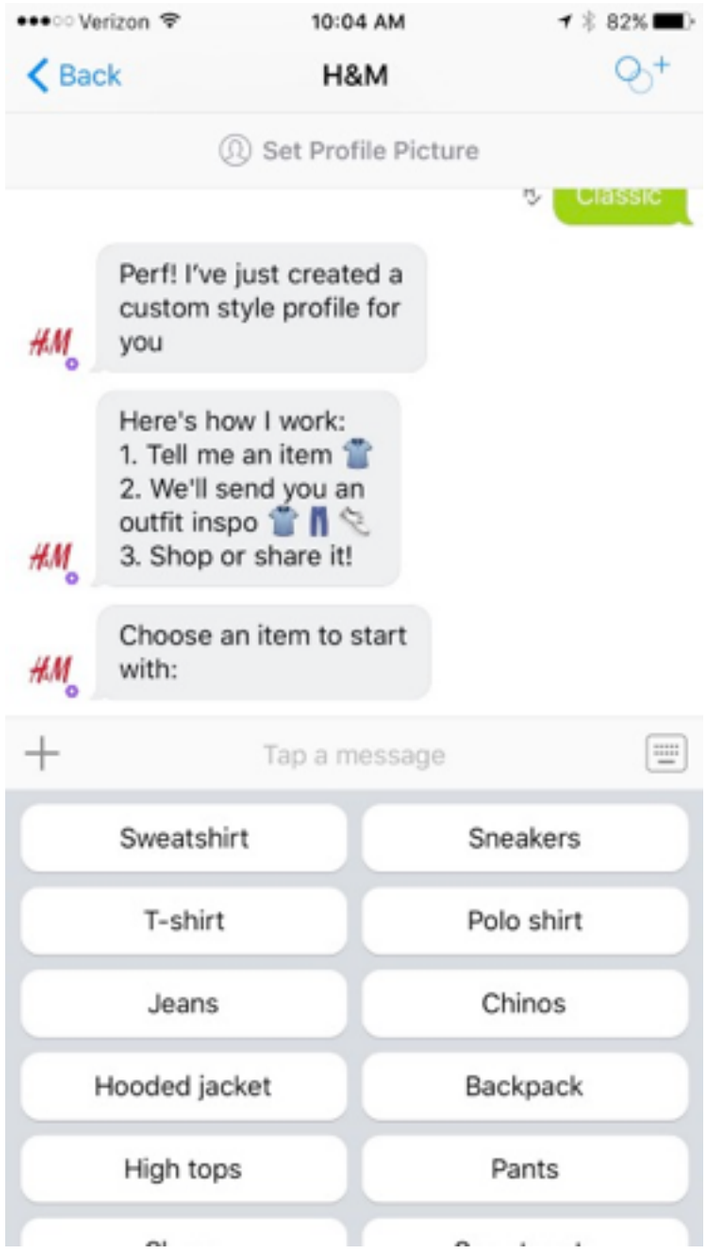
IMAGE (PATTERN) RECOGNITION

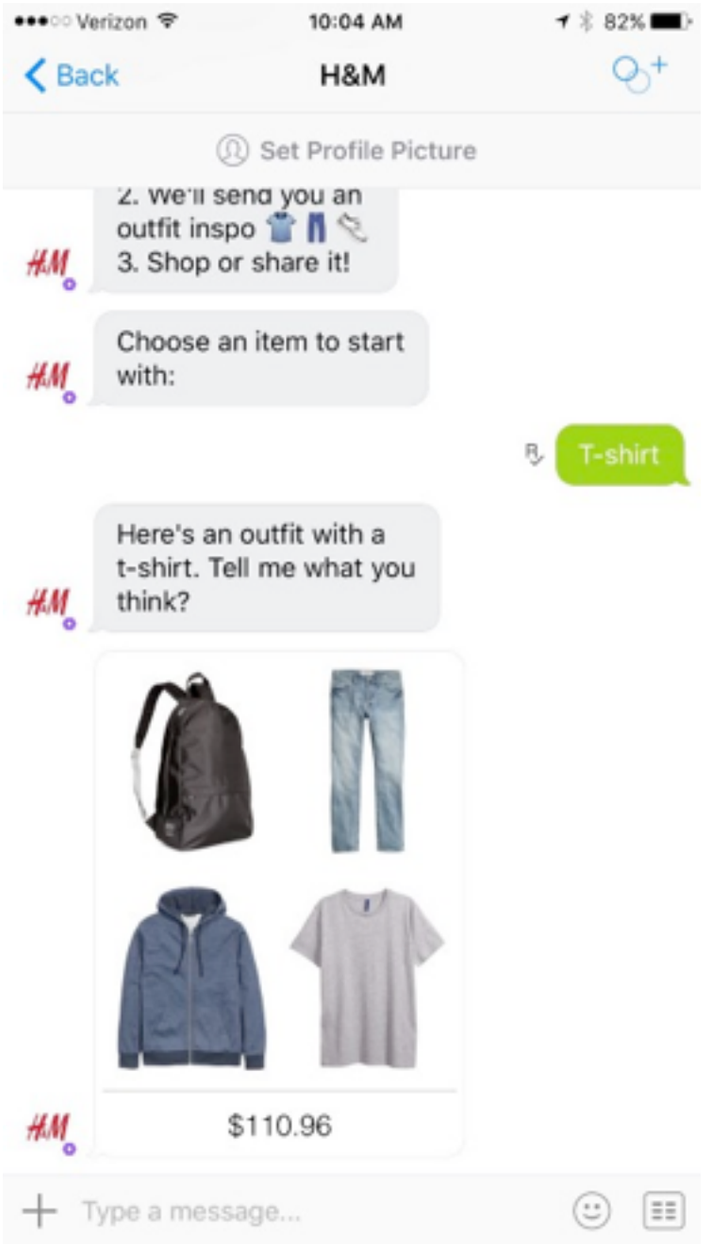
At ImageNet Large Scale Visual Recognition Challenge 2015 Microsoft "machine" had beat the human benchmark of 5.1% errors with a 4.94%

Is this a Pomeranian or a Shih Tzu ?

HUMAN EMULATION (AKA VIRTUAL BOTS)

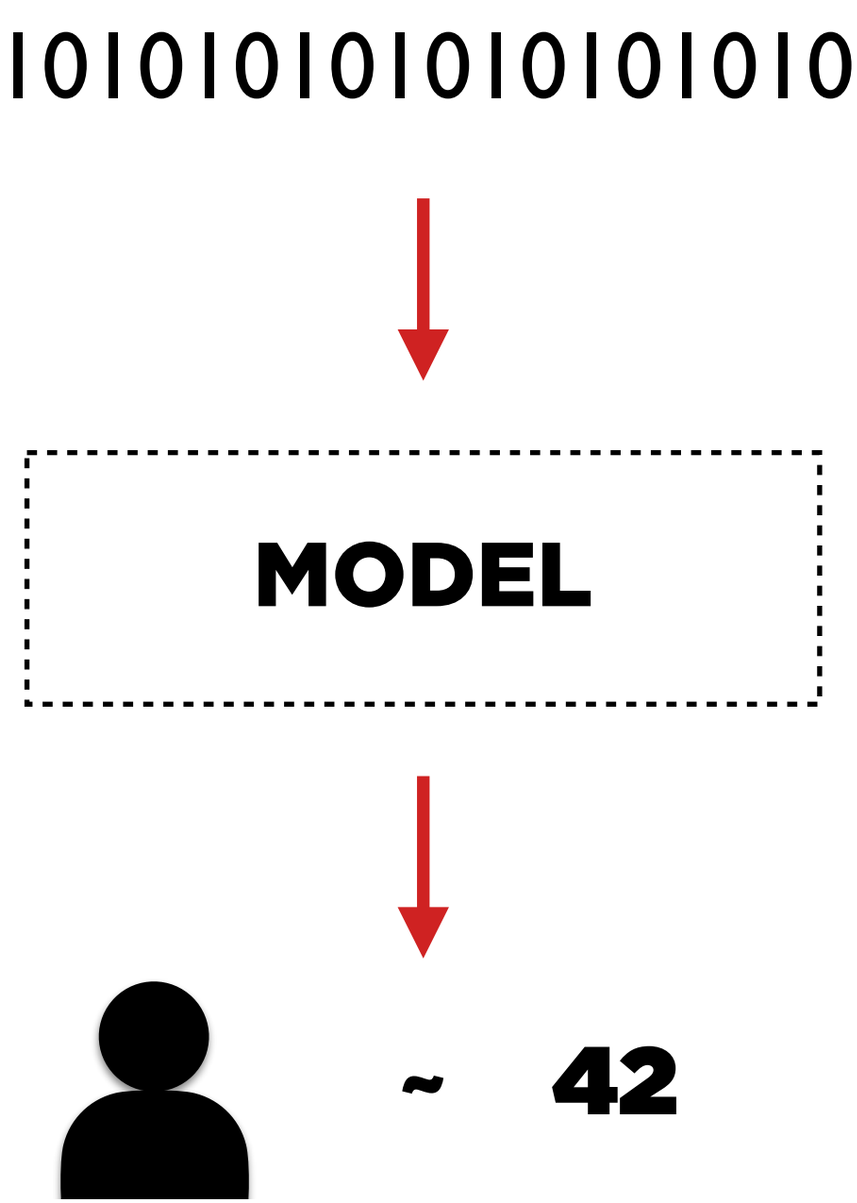




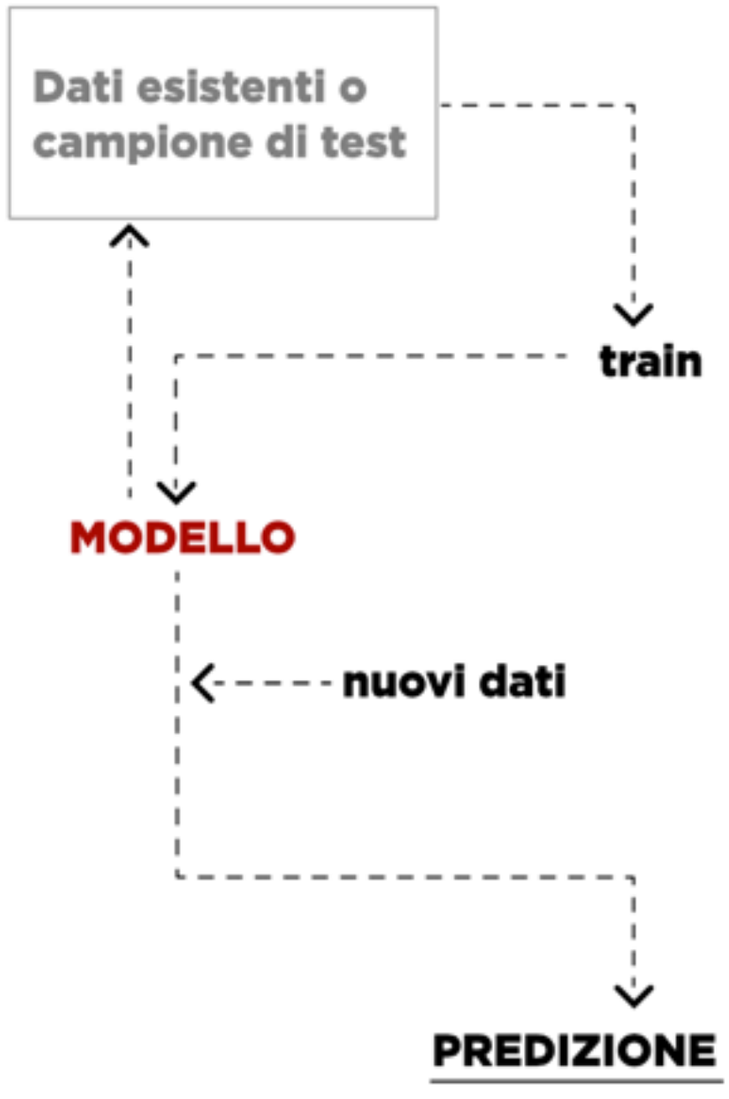


AUTOMATIC XXX SCORING

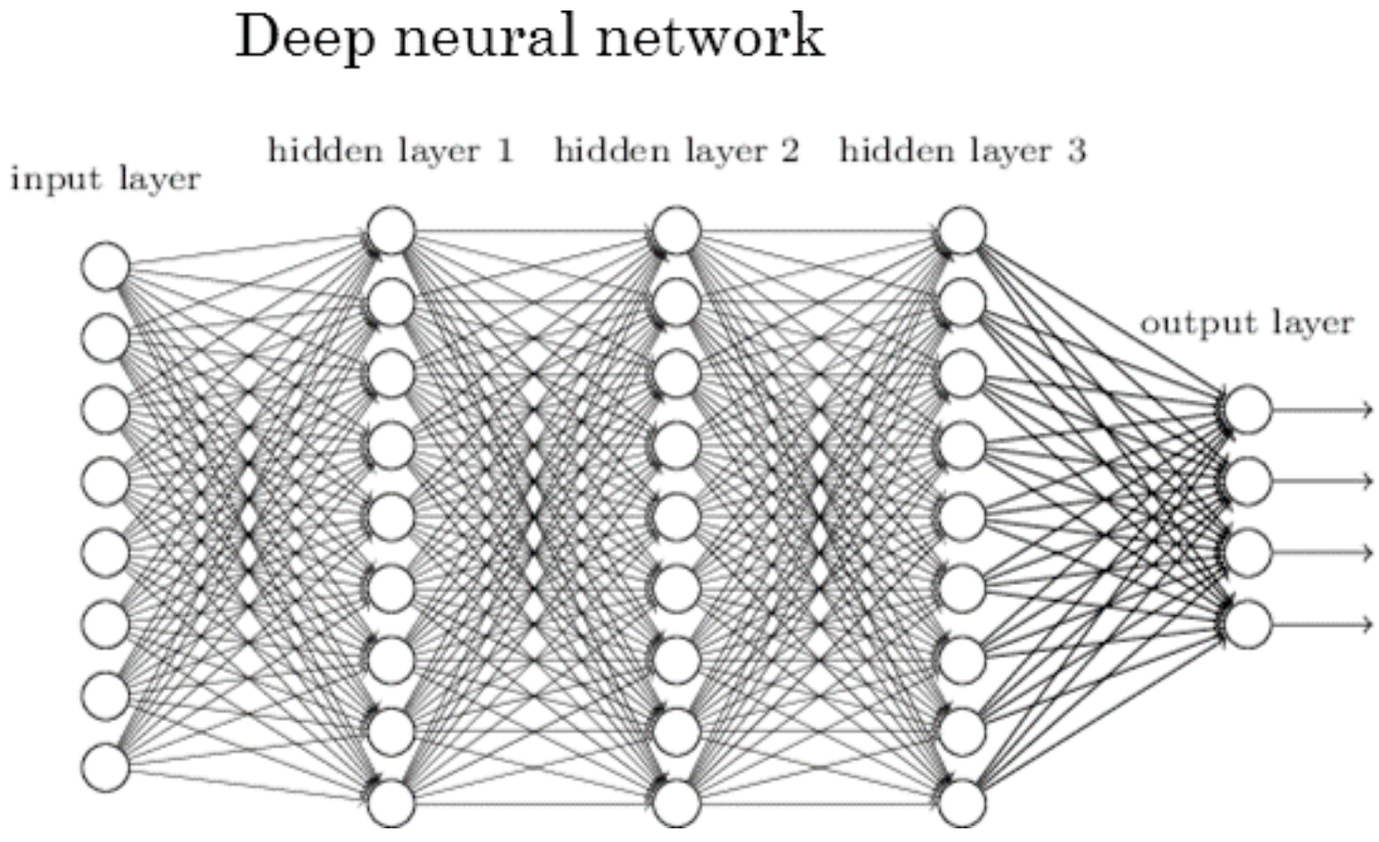
In the US, some 72% of CVs are never seen by human eyes. Computer programs flip through them...



MACHINE LEARNING



DEEP LEARNING



REMARK: YOU CAN'T ISOLATE DATA FROM ITS SOURCES AND MOTIVATIONS

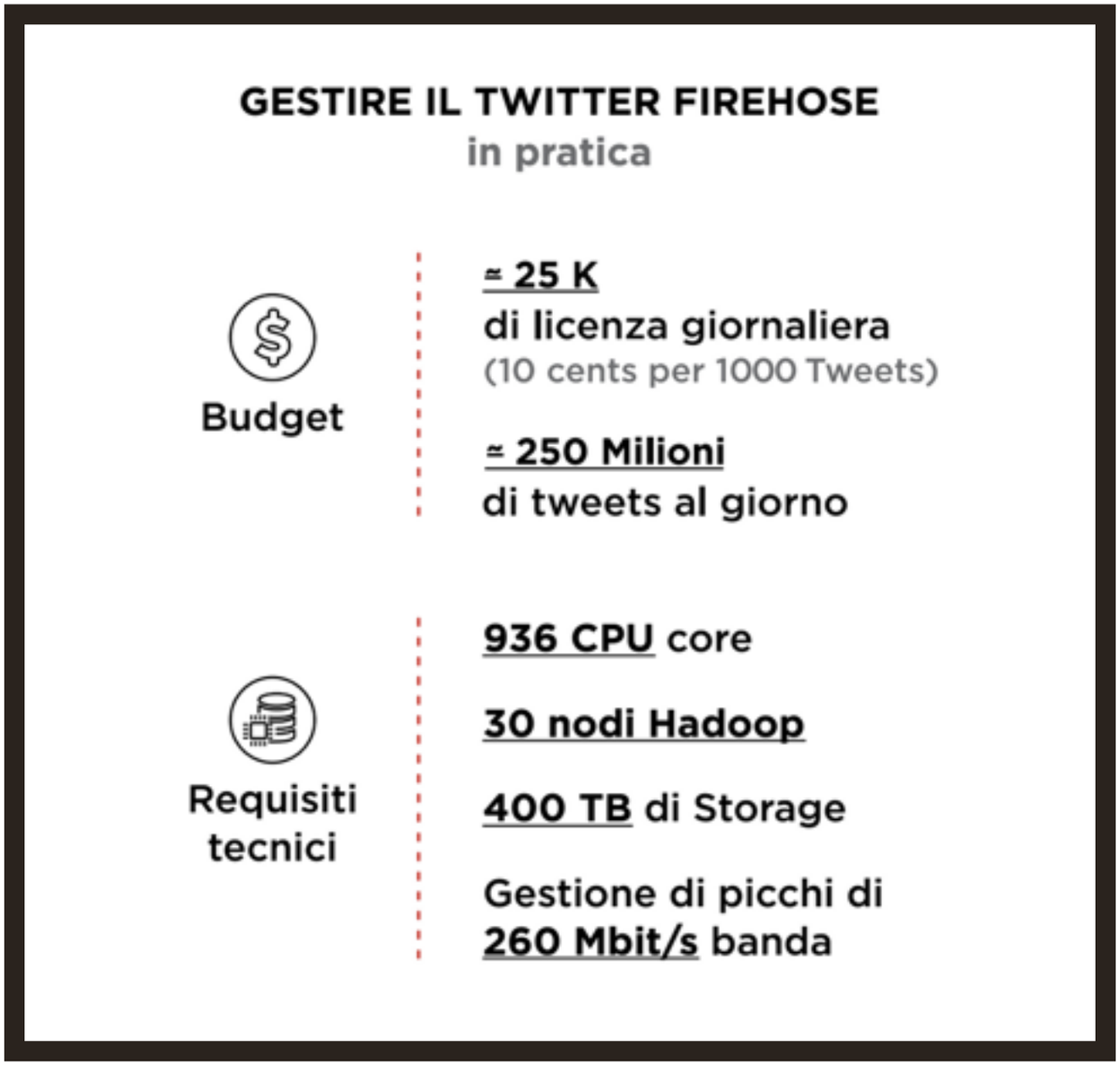
" ... in addition to looking at data in the traditional way, we must now consider political and social structures, and how people learn from and influence each other; we must consider how ideas flow through social networks, what motivates people to contribute to discussions, and the consequences of engagement."

Mindjet data scientist Anna Gordon

RIF. EU CoeGSS Project - www.coegss.eu

THE DARK SIDE OF BIG DATA

DEMOCRAZIA DEI DATI VS OLIGARCHIA DEI BIG DATA

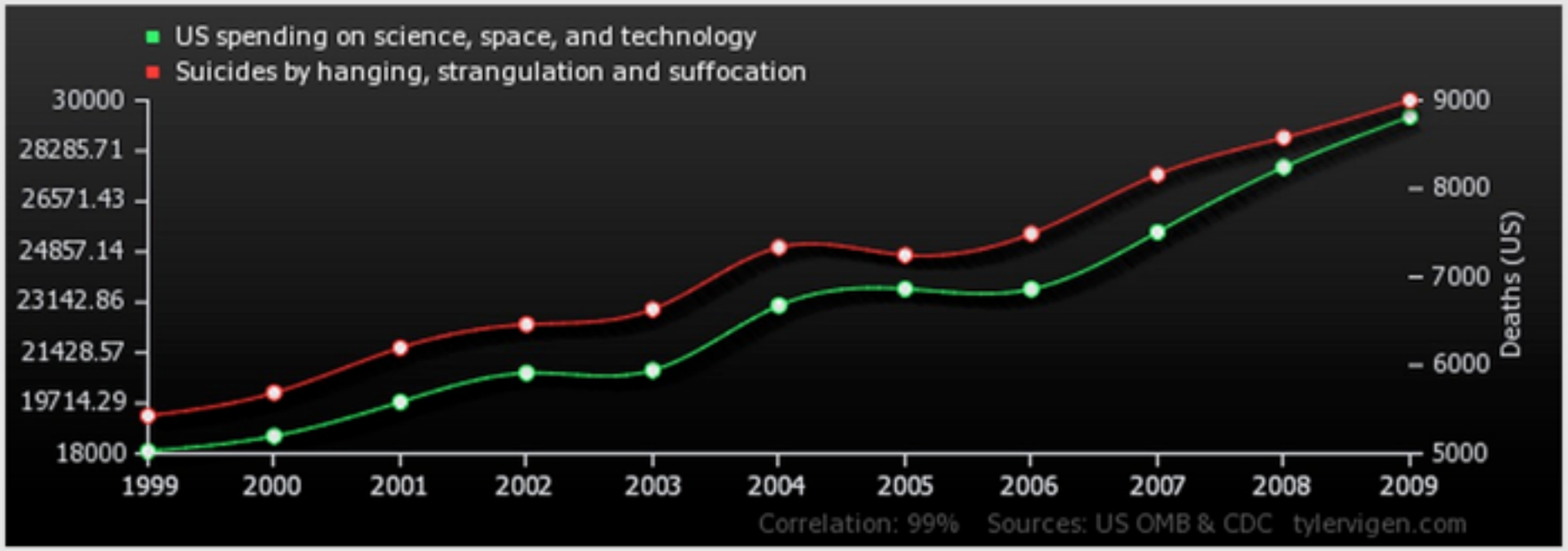


TUTTO È CORRELATO ?

US spending on science, space, and technology

correlates with

Suicides by hanging, strangulation and suffocation



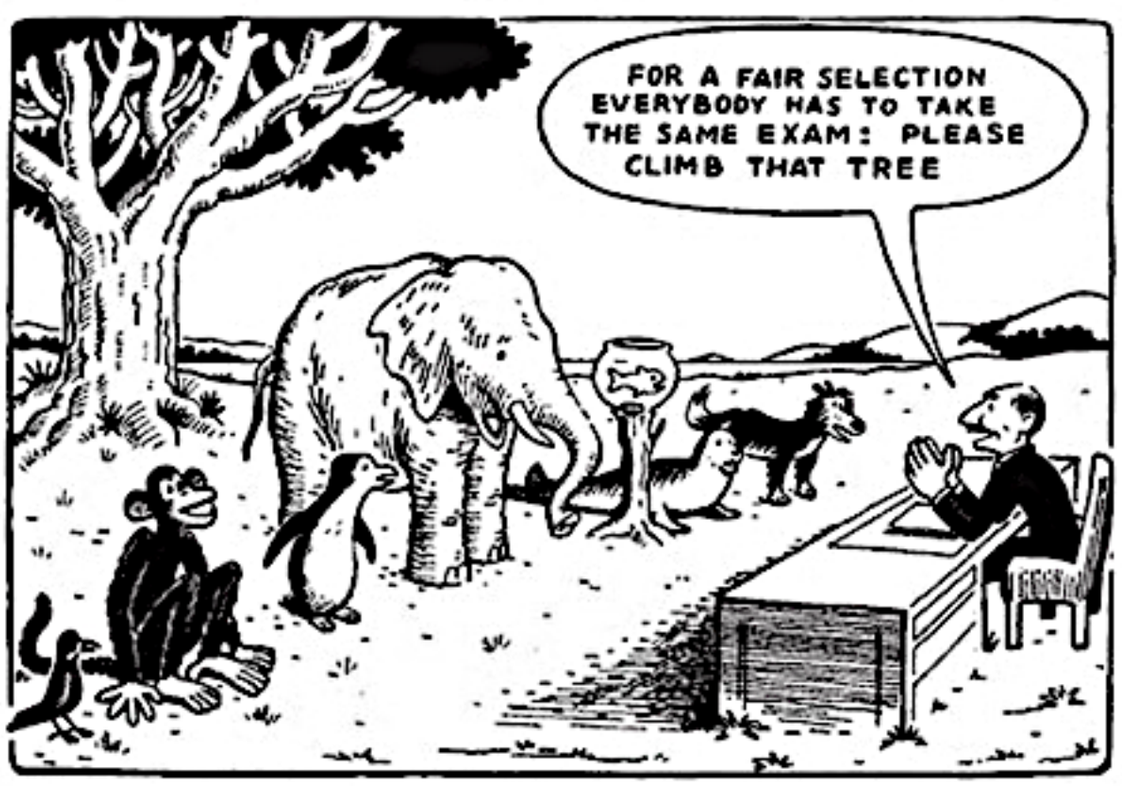
	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
US spending on science, space, and technology *Millions of todays dollars (US OMB)*	18,079	18,594	19,753	20,734	20,831	23,029	23,597	23,584	25,525	27,731	29,449
Suicides by hanging, strangulation and suffocation *Deaths (US) (CDC)*	5,427	5,688	6,198	6,462	6,635	7,336	7,248	7,491	8,161	8,578	9,000

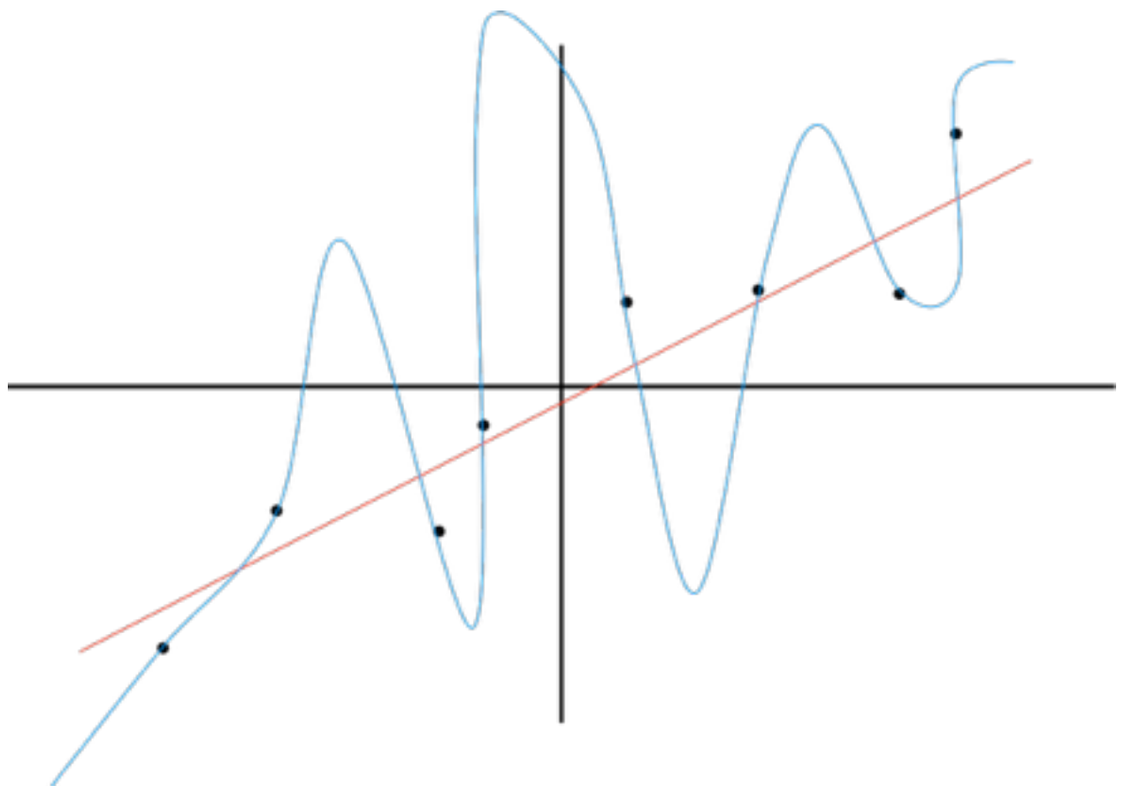
Correlation: 0.992082

Permalink - Mark as interesting (5,147) - Not interesting (2,370)

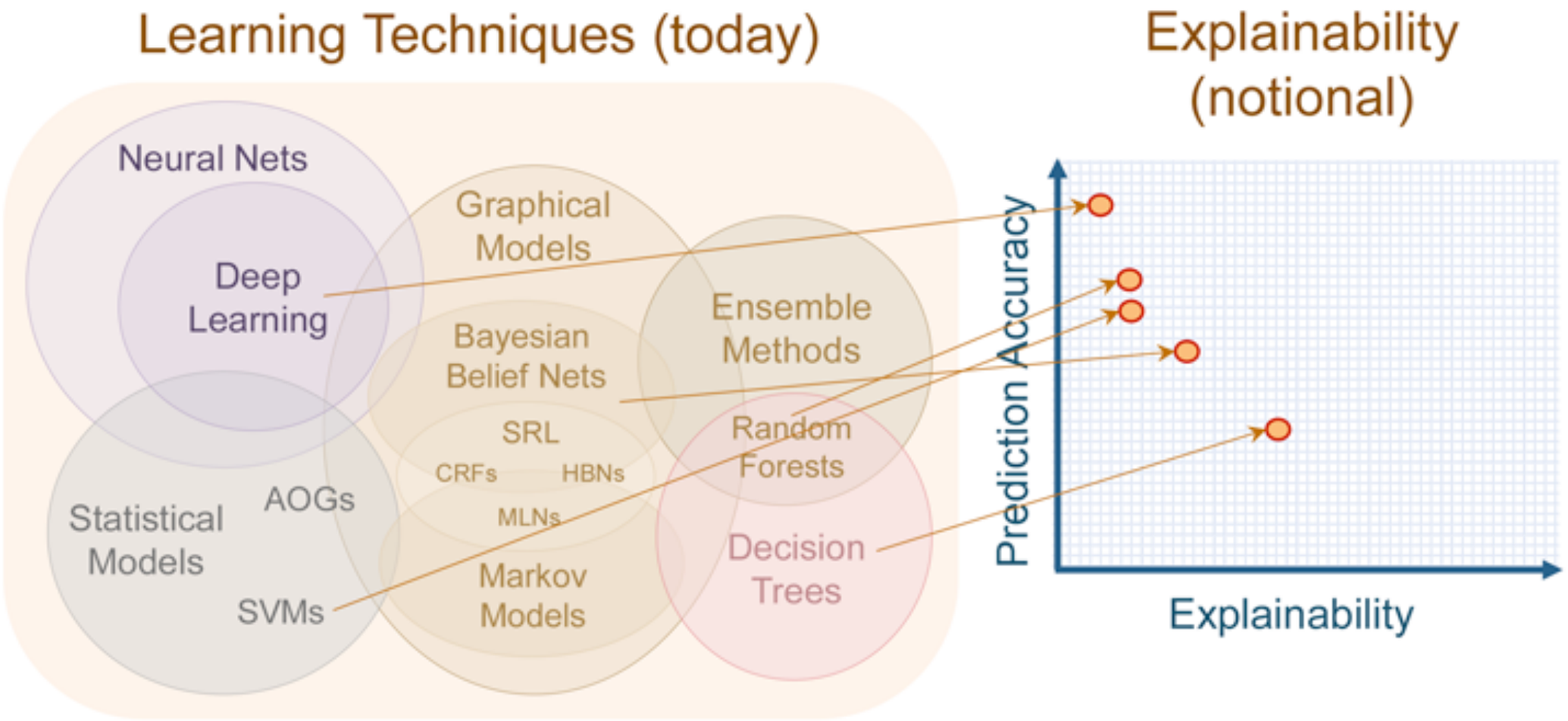
BOLLE, FILTRI E DETERMINISMO

BIAS & OVERFITTING



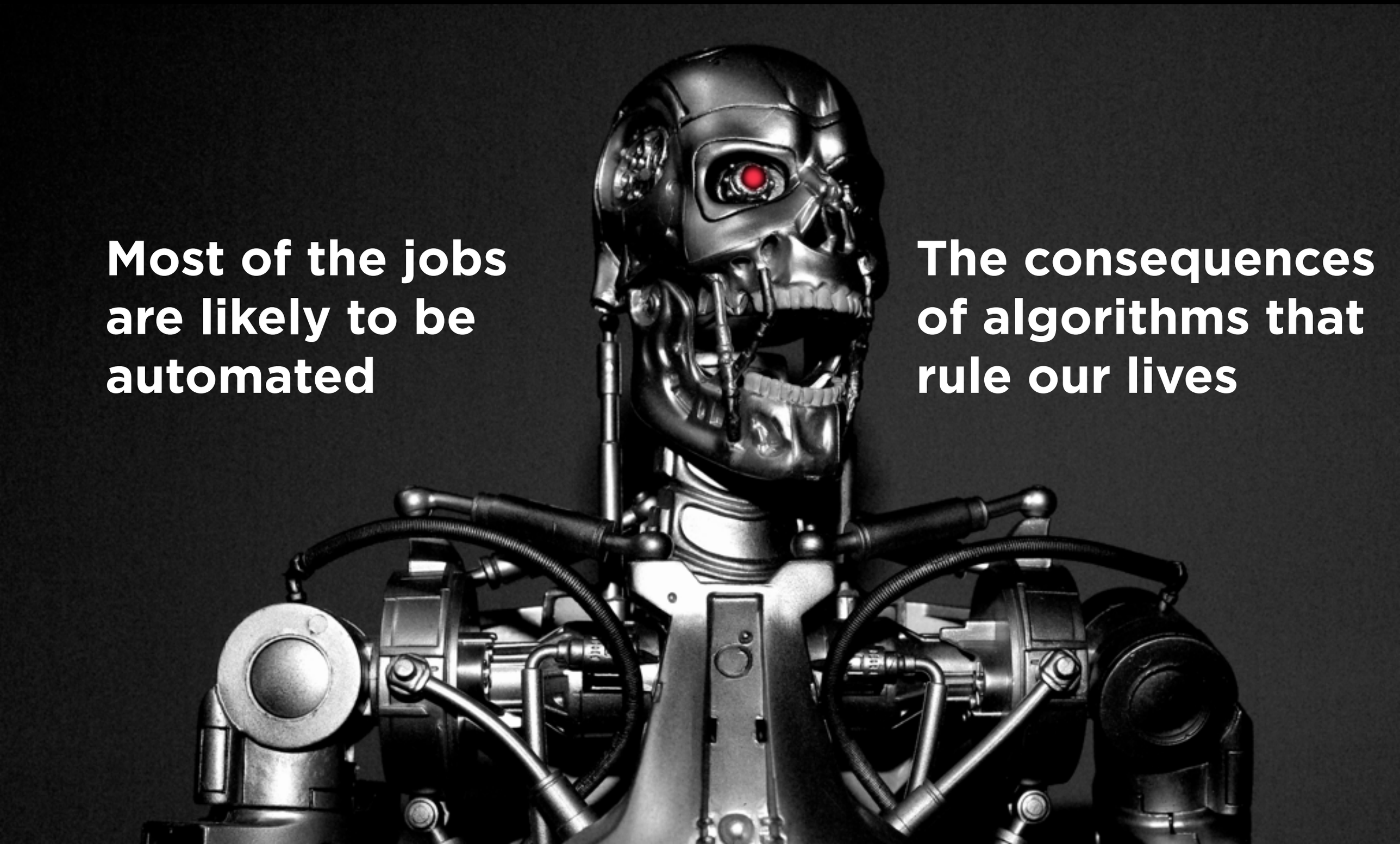


EXPLAINABILITY VS ACCURACY



http://nautil.us/issue/40/learning/is-artificial-intelligence-permanently-inscrutable

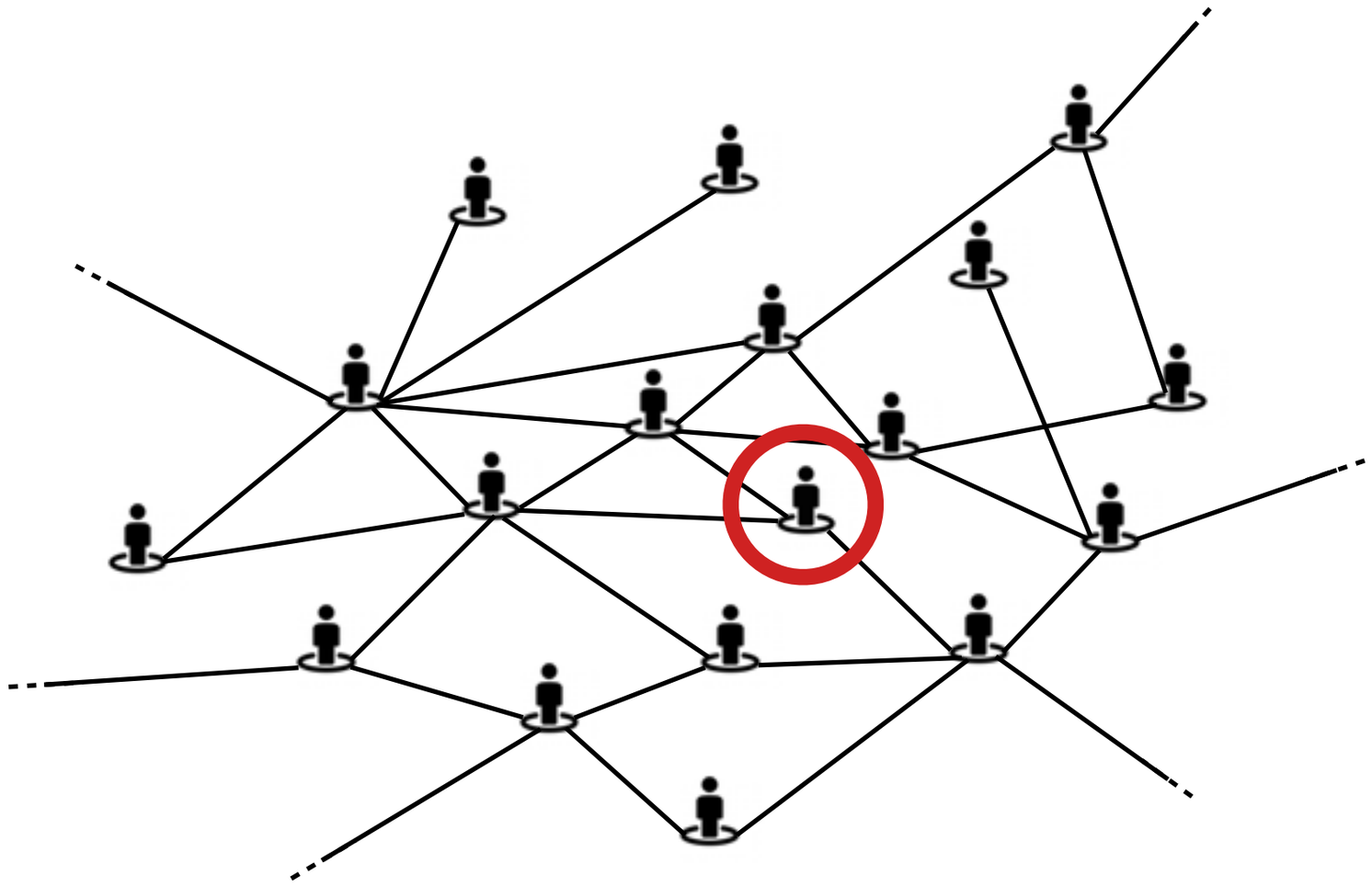
THE TERMINATOR EFFECT



Most of the jobs are likely to be automated

The consequences of algorithms that rule our lives

DATI ANONIMI E DISCRIMINAZIONE



HUMAN RIGHTS ARE NOT OPTIONAL

On August 4, 2006, AOL Research, released a compressed text file containing twenty million search keywords for over 650,000 users over a 3-month period.

The New York Times was able to locate an individual from the released and anonymized search records by cross referencing them with phonebook listings.

FINAL RECOMENDATIONS

1. **DATA is the core ASSET, consider BUILDING YOUR OWN DATASET.**

2. **DON'T FOCUS ON TOOLS, FOCUS ON SCIENCE.**

3. **RIGHT NOW the only way to "master" complexity is throughout data (having input from various theoretical perspectives & from a range of empirical approaches).**

4. **GO OPEN (source), engage the crowd as your R&D department.**

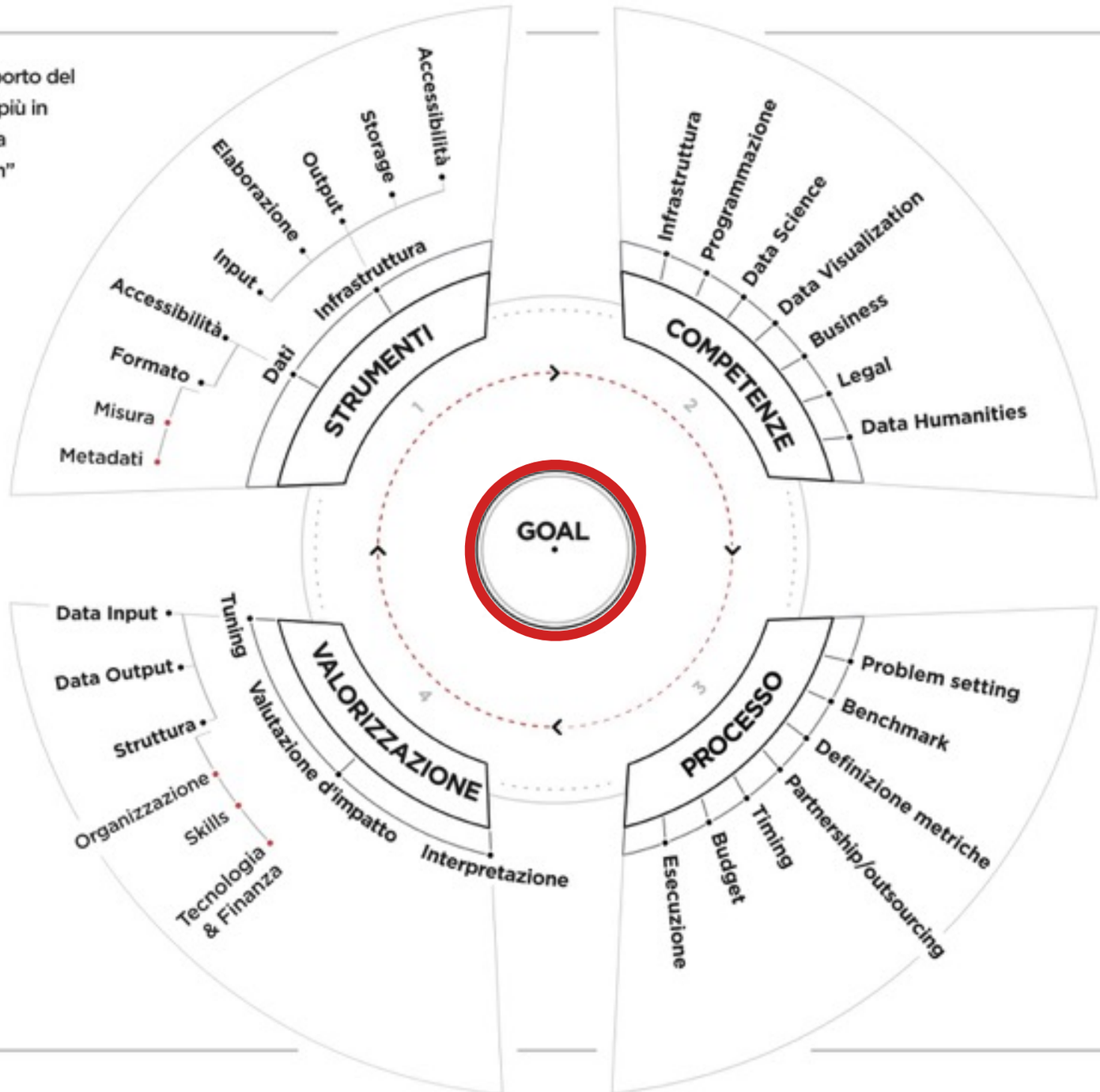
5. **DECENTRALIZATION might be the new mantra.**

6. **IMPLEMENT A SYSTEMIC, LEARN-BY-DOING, DYNAMIC, VALUE-BASED APPROACH...**

DATA RING CANVAS

DATA RING

Un “data-canvas” a supporto del manager, dei decisori, e più in generale di chi approccia un progetto “data-driven”



COMPOSIZIONE DATA RING

GOAL
Definizione degli obiettivi da raggiungere

4 FASI PRINCIPALI
Sezioni in cui si articola il Data Ring

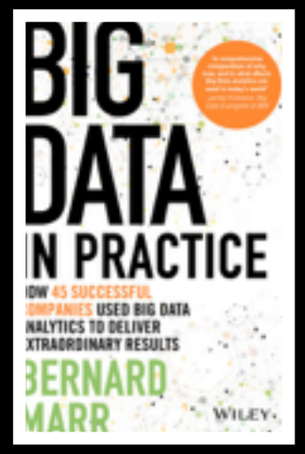
SOTTOFASI
Composizione delle fasi principali

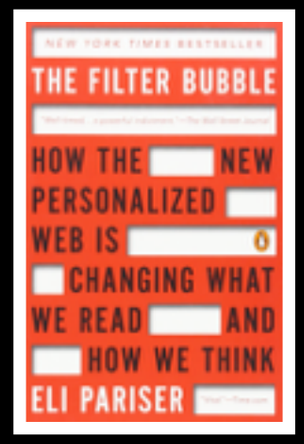
ELEMENTI DELLE SOTTOFASI

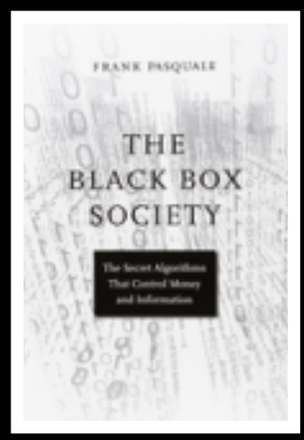
PROGETTO DI: Leonardo Camiciotti, Christian Racca
Grafica a cura di Laura Pippinato

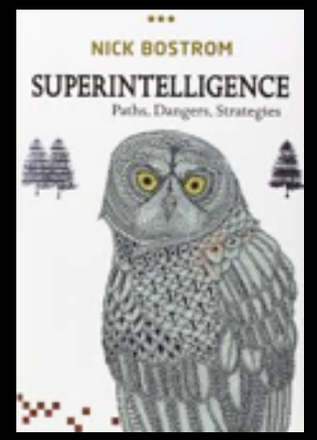
www.dataring.eu

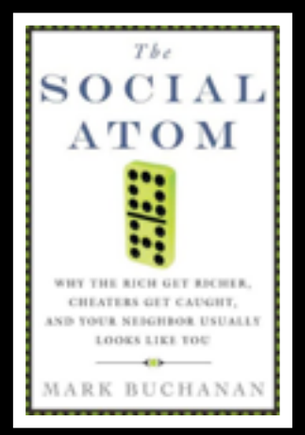
BOOK REFERENCES

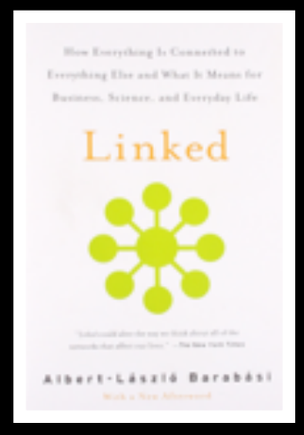












THANKS!