BIG DIVE

THE DATA RING: A CANVAS FOR DATA PROJECT

top ix torino piemonte internet exchange

Designed for AMLD

top ix | torino piemonte internet exchange

27-01-2018

BIG DIVE

CHRISTIAN RACCA

STEFANIA DELPRETE

TO DATA BEERS

top IX

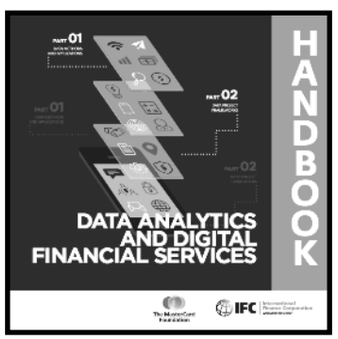
@pendolare

pendolare_digitale

pendolaredigitale

pendolare

astrastefania



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

WHAT'S "NEW" ABOUT DATA

DATA AVAILABILITY
/ Exponential growth
/ Machine VS human
/ Structured VS un-structured

INFRASTRUCTURE AS A COMMODITY
/ Cloud
/ HPC & HPN
/ Frameworks

SKILLS TO EXTRACT INFORMATION ARE NOW MORE ACCESSIBLE

CULTURE & APPROACH
/ Complexity science
/ Network thinking
/ Open Innovation

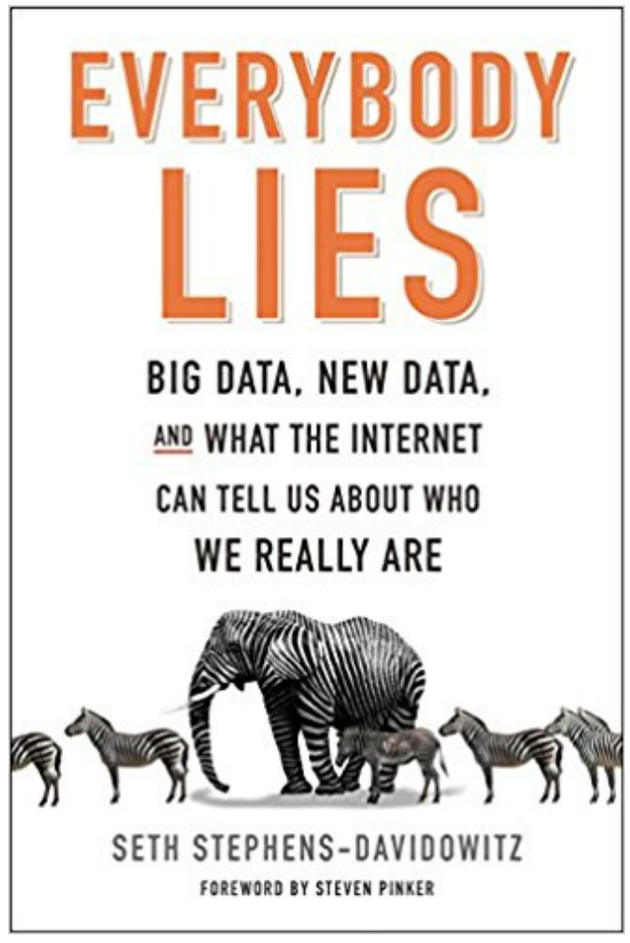
SUPERPOWERS ENABLED BY BIG DATA

i. Offering up **new type** and sources of Data. (e.g. Social Network)

ii. BIG DATA allows us to finally see **what people really want** and really do. (e.g. Google Search bar)

iii. Allowing us to **zoom in** on small subsets of people.

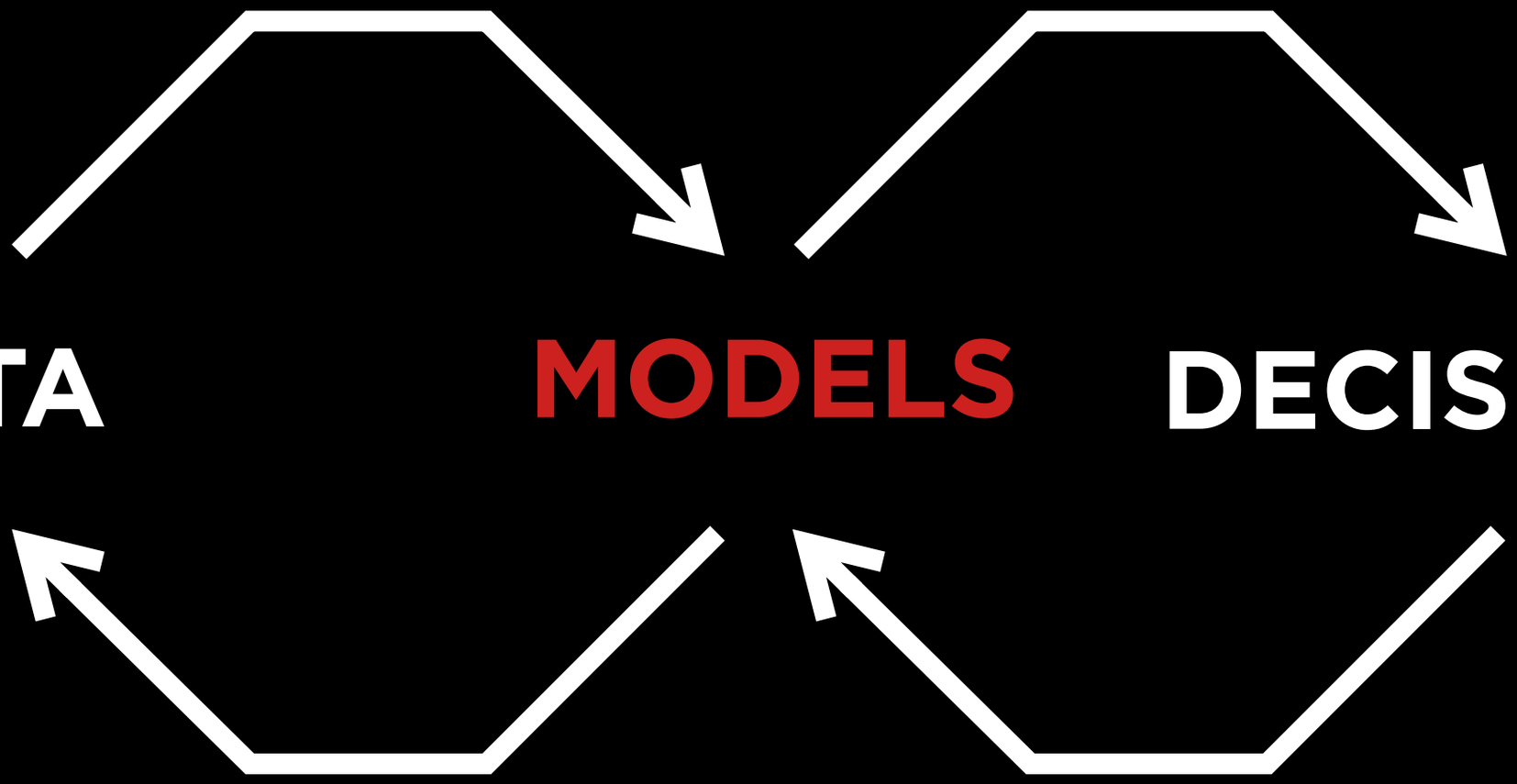
iv. Allowing proper "**scientific**" **experiments** on a large scale.



BIG DATA + ML = The NEW STACK

Big Data technologies are used to handle core data engineering challenges, and machine learning is used to extract value from the data.

THE MANTRA: FROM DATA TO IMPACT



... AND BACK

COMMON OPEN CHALLENGES

/ THE DATA

/ THE SKILLS

/ FROM PROTOTYPE TO...

/ THE RESULTS INTERPRETATION AND
THE EXPLAINABILITY ISSUE

/ "GREY ZONES" IN DATA EXPLOITATION

DATA METADATA FEATURES

CHALLENGE #1

DATA REMAINS THE STARTING POINT

Volume

The effective amount of <u>usable</u> data. No a-priori objective parameters. On field validation is required.

Metadata

"Data" that provides information about other data. {Descriptive, Structural, Administrative}

Features Selection

Refers to the process of extracting useful information (or features) from existing data.

ABOUT FEATURES...

FROM SOURCE DATA

Features “reduction”

Noisy or redundant data makes it more difficult to discover meaningful patterns.

High-dimensional dataset requires more complex models/algorithms and more computational power.

VS

Data augmentation

Enriching existing data with open data or through third-party data providers.

TO RELEVANT DATA

ABOUT FEATURES...

FROM SOURCE DATA

Proxy data

Quantitative, high correlation... but still a proxy!

VS

Direct* data

Sometime qualitative and difficult to grab.

What gets chosen is usually whatever is easiest to quantify, rather than the fairest.

TO RELEVANT DATA

BIG DIVE

**** the opposite of Proxy***

DATA REMAINS THE STARTING POINT

Volume

The effective amount of usable data. No a-priori objective parameters. On field validation is required.

Metadata

"Data" that provides information about other data. {Descriptive, Structural, Administrative}

Features Selection

Refers to the process of extracting useful information (or features) from existing data.

Data Quality

Traceability, expiration, completeness, currentness compliance, understandability, accuracy.

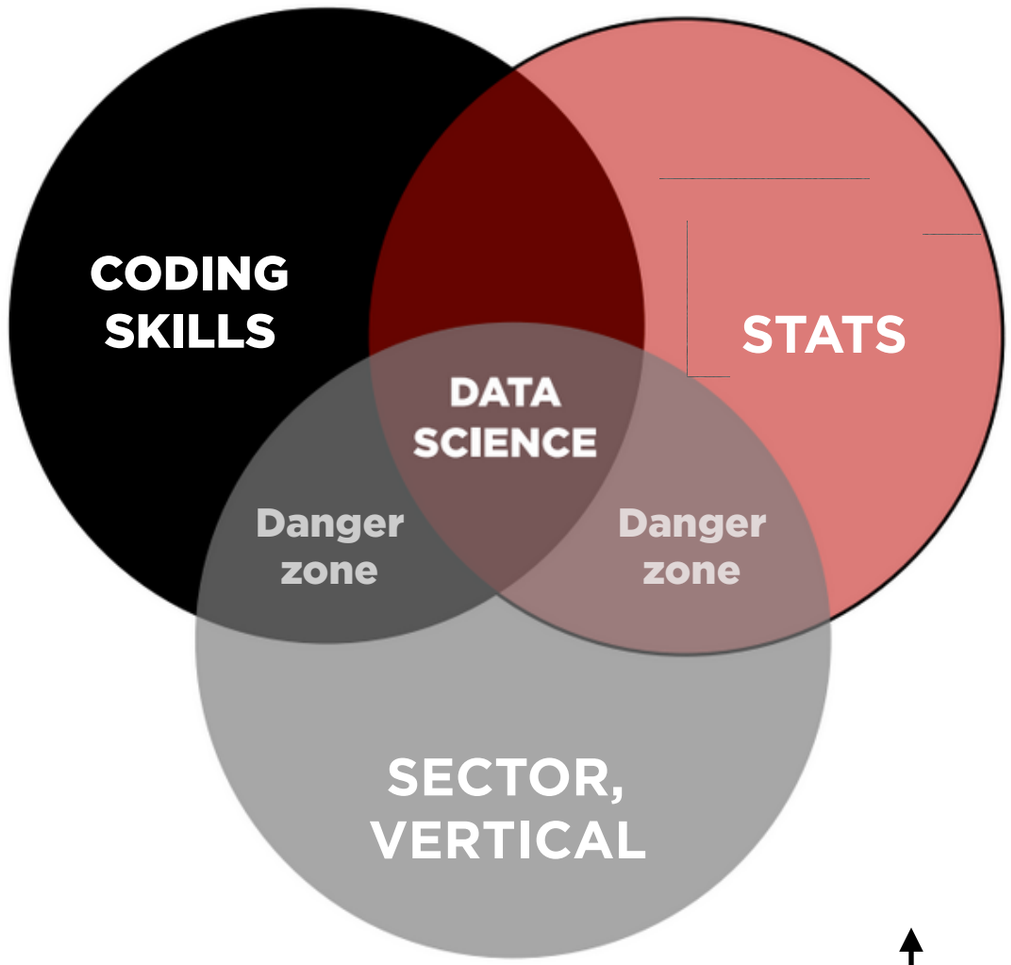
THE DATA TEAM

CHALLENGE #2

LOOKING FOR UNICORNS

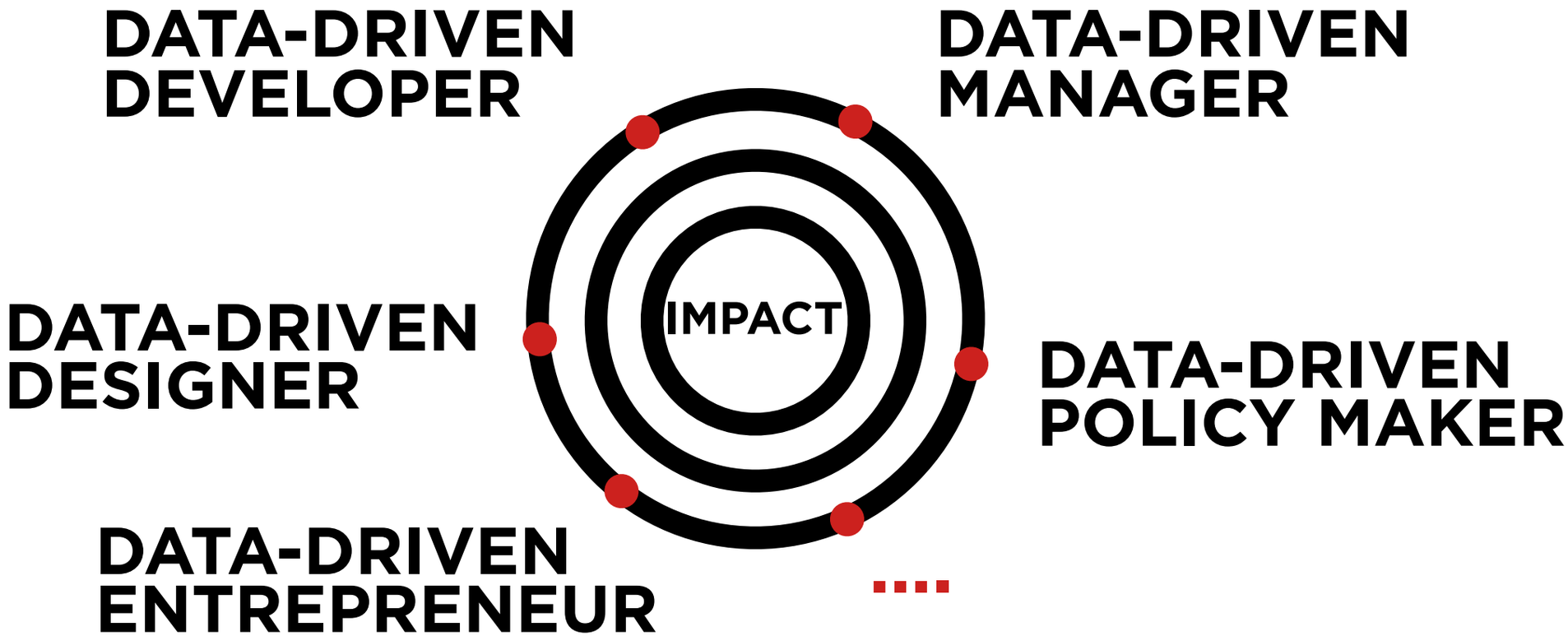
WHAT PEOPLE LOOK FOR IN A DATA SCIENTIST IS SOMEONE WHO CAN COME UP WITH A PROBLEM THAT CAN IMPROVE THE BUSINESS, DESIGN AN EXPERIMENT TO COLLECT RELEVANT DATA TO ANSWER THAT PROBLEM, CLEAN THE DATA TO GET TO THE RELEVANT INFORMATION, ANALYZE THIS DATA AND DRAW CONCLUSIONS.

- Priya Gupta -



Re-arrangement of the THE DATA SCIENCE VENN DIAGRAM by Drew Conway

THE DATA TEAM



SINGLE VS TEAM

FROM PROTOTYPE TO PRODUCTION

CHALLENGE #3

A BABEL OF (CODING) LANGUAGES

PRODUCTION

JAVA, C, C++, ...

REFACTORING

-

DATA ENGINEERING

[SCALA, ...]

DATA-DRIVEN PROTOTYPE

PYTHON, R, D3.JS

RESULTS INTERPRETATION & EXPLAINABILITY

CHALLENGE #4

THE EXPLAINABILITY ISSUE

LOW EXPLAINABILITY

Deep learning

Machine learning

Inferential statistics

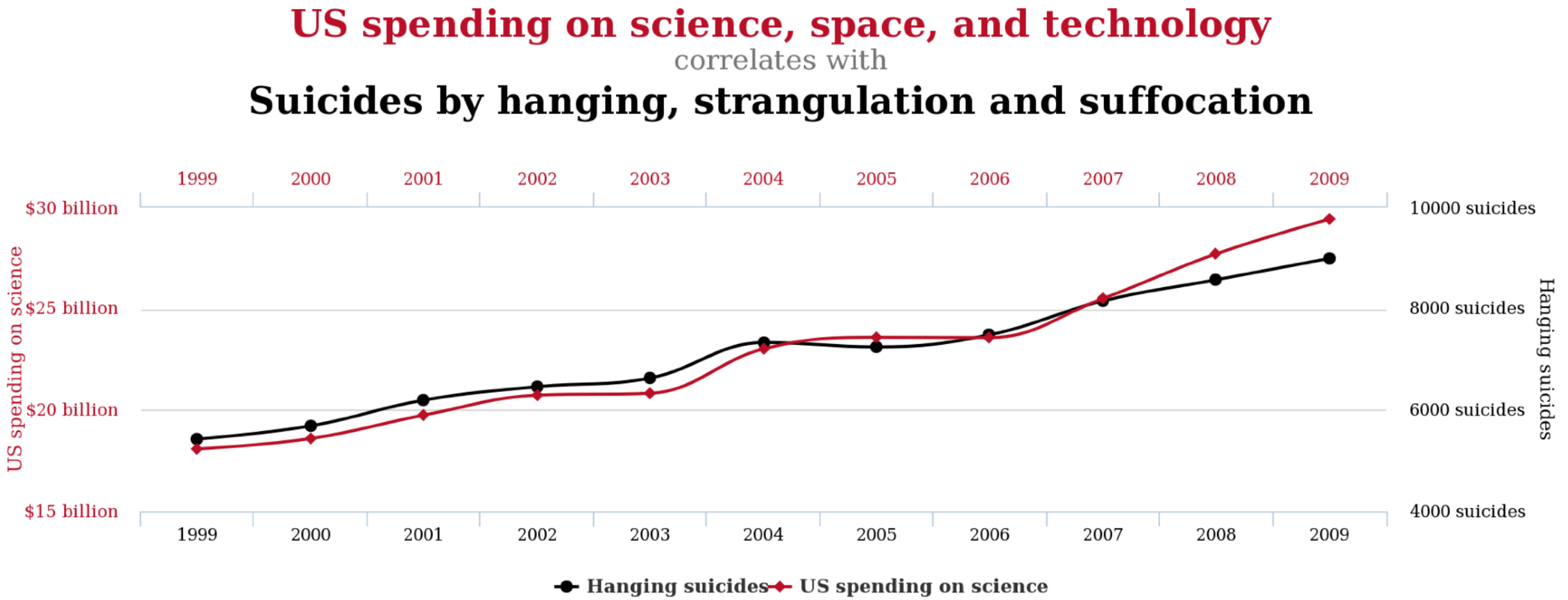
HIGH EXPLAINABILITY

GREY ZONES IN DATA EXPLOITATION

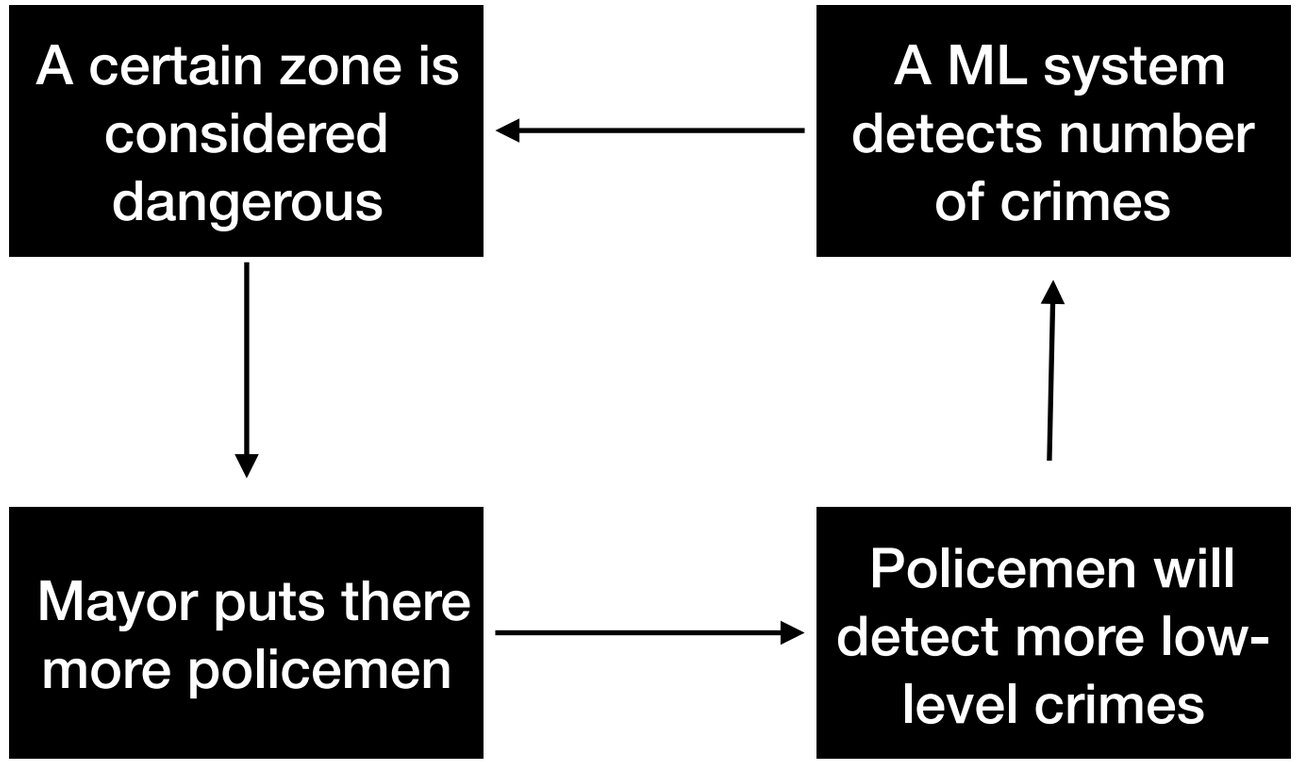
(A.K.A. THE DARK SIDE OF BIG DATA)

CHALLENGE #5

ALL MIGHT BE CORRELATED



BUBBLES & NEGATIVE LOOPS



DATA DEMOCRACY VS BIG DATA OLIGARCHY

MANAGING TWITTER FIREHOSE

Budget

≃ 25 K
Daily license

≃ 250 Milioni
Tweets / day

Tech needs

936 CPU core

30 Hadoop Nodes

400 TB Storage

Bandwidth Peak
260 Mbit/s

ENTRY-BARRIERS TO "MANAGE" BIG DATA ARE ALWAYS TOO HIGH.

... IN → ... OUT

MACHINE learns ONLY through the training data (no additional elaboration, no context, ...)

GARBAGE IN → GARBAGE OUT

Training set:

2+2 = 5

2+2 = 5

2+2 = 5

2+2 = 5

2+2 = 4

>>> MACHINE SAYS THAT 2+2 = 5

… IN → … OUT

MACHINE learns ONLY through the training data (no additional elaboration, no context, …)

BIAS IN ⟶ BIAS OUT

Training set:
A man from Country X did not return a loan
A woman from Country X did not return a loan
A ….

>>> MACHINE DECISION: DENY LOANS TO PEOPLE FROM COUNTRY X

THE PRIVACY ISSUE

THE PURPOSE SPECIFICATION PRINCIPLE FAILS TO KEEP PACE WITH DEVELOPMENTS IN TECHNOLOGY AND SERENDIPITY.

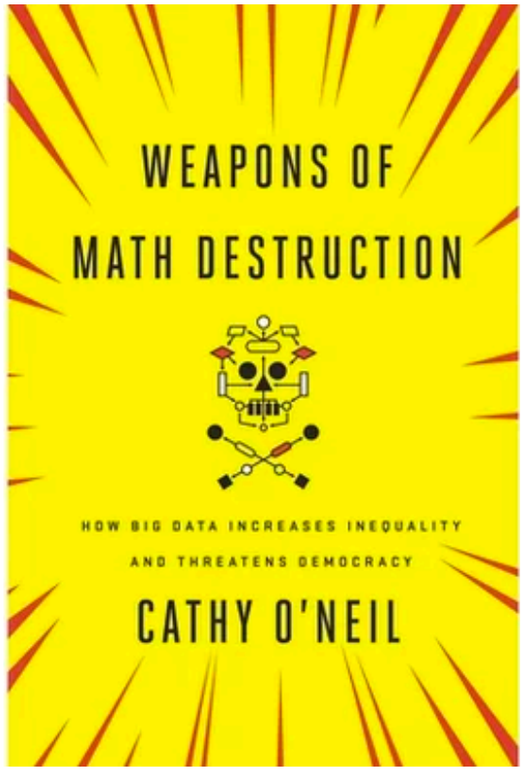
THE ETHIC PROBLEM

i. Gender discrimination

ii. Race discrimination

iii. Class discrimination

Rich people are likely to be evaluated by humans while poor people are likely to be evaluated by machines.



SOME NICE-TO-HAVE PRACTICES

/ An "Hippocratic Oath" for the Data Scientists is advisable

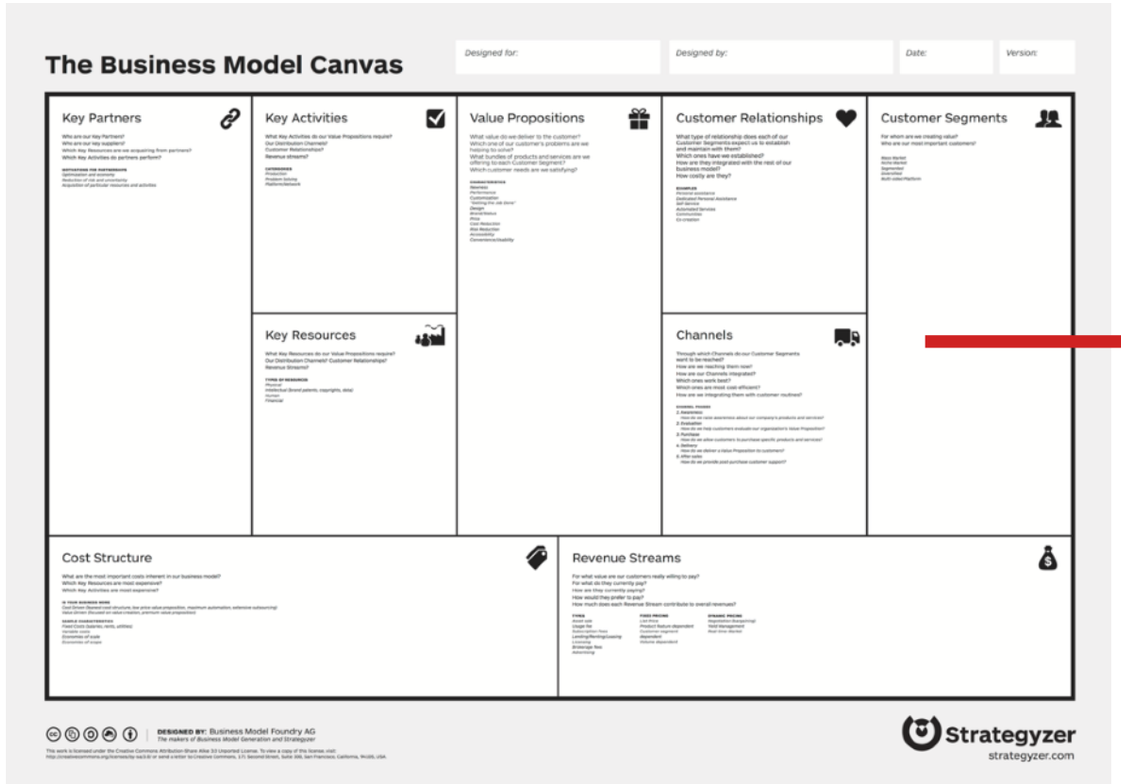
/ Don't forget that results of ML algorithms often affect real people

/ Try to avoid negative loops & hardcoded prejudice

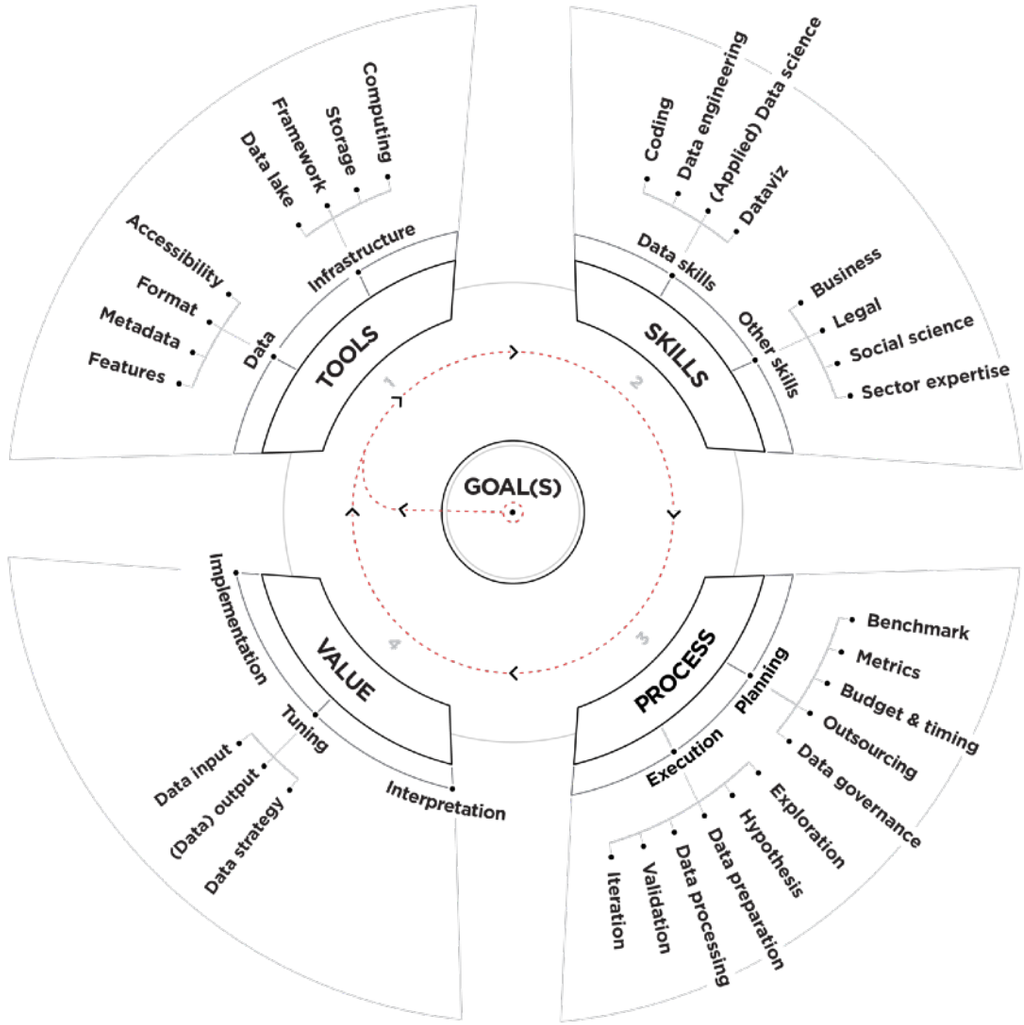
/ Model improvements & updates by design

LEVERAGING (BIG) DATA OPPORTUNITIES REQUIRES A PROPER METHOD

THE CANVAS APPROACH



The inspiring precursor



The Data Ring

WHY A CANVAS ?

/ It forces the project owner to state crystal clear the value proposition of the project.

/ It is an analytical tool, devoted to self-diagnosis and to define and respect an internal strategy.

/ It provides for a complete representation of the process that can be explained to third parties too.

/ It is not a “static shot” but it evolves through time according to project evolution.

/ It is not the solution but it helps to reduce failure risk.

THE DATA RING

THE DATA RING

http://dataring.eu/

BIG DIVE

The Data Ring Canvas

Project name:	Designed by:	Date:	Version:

BIG DIVE

KIND OF PROBLEMS YOU CAN AVOID

i. Discovering massive lack of data or bad quality when it's too late.

ii. Being stopped by tech lock-in, or legal constraints.

iii. Creating un-effective P.o.C.

iv. Developing data-tools that can't be deployed in-production.

v. Defining "ex-post" the generated impact.

vi. Underestimating skills, training needs and resource.

RESOURCES

https://github.com/pendolare/AMLD

FEEDBACK

http://bit.ly/bigdive-amld

WE ARE HIRING !

Open position: Python Full Stack Developer

christian.racca@top-ix.org
stefania.delprete@top-ix.org

www.top-ix.org
www.bigdive.eu

@top_ix
@bigdive_eu

THANKS!