BIG DIVE

DATA SCIENCE & ANALYTICS

# A SOFT INTRO TO APPLIED DATA SCIENCE

CHRISTIAN RACCA

Organized by

top ix torino piemonte internet exchange

Designed for

INTESA SANPAOLO

In collaboration with

aizoon technology consulting

ISI Foundation

TODO

# BEFORE STARTING

**Please,**
**wear your**
**badge**

**SSID bigdive**
**PWD wifi4divers**

**09:00**

**17:30**

# BEFORE STARTING

**Coffe machine & kitchen available in house**

**Please notify delay or absence**

**We'll share all the teaching materials on GIT**

# BEFORE STARTING

**Remote Desktop to VM**

**Slack Channel**

**Git Repo**

# PROGRAM

| | DAY 1<br>(09 ott) | DAY 2<br>(10 ott) |
|---|---|---|
| 09:30 - 11:00 | *Soft intro to applied data science* | *Python basics* |
| 11:00 - 11:15 | *Break* | *Break* |
| 11:15 - 12:30 | *ISI lecture* | *Python basics* |
| 12:30 - 13:30 | *Aizoon lecture* | |
| 13:30 - 14:30 | *Lunch break* | *Lunch break* |
| 14:30 - 15:30 | *Data Viz intro* | *Python basics* |
| 15:30 - 16:30 | *Data Dev intro* | *HTML & CSS essentials* |

BIG DIVE

(Disclaimer)

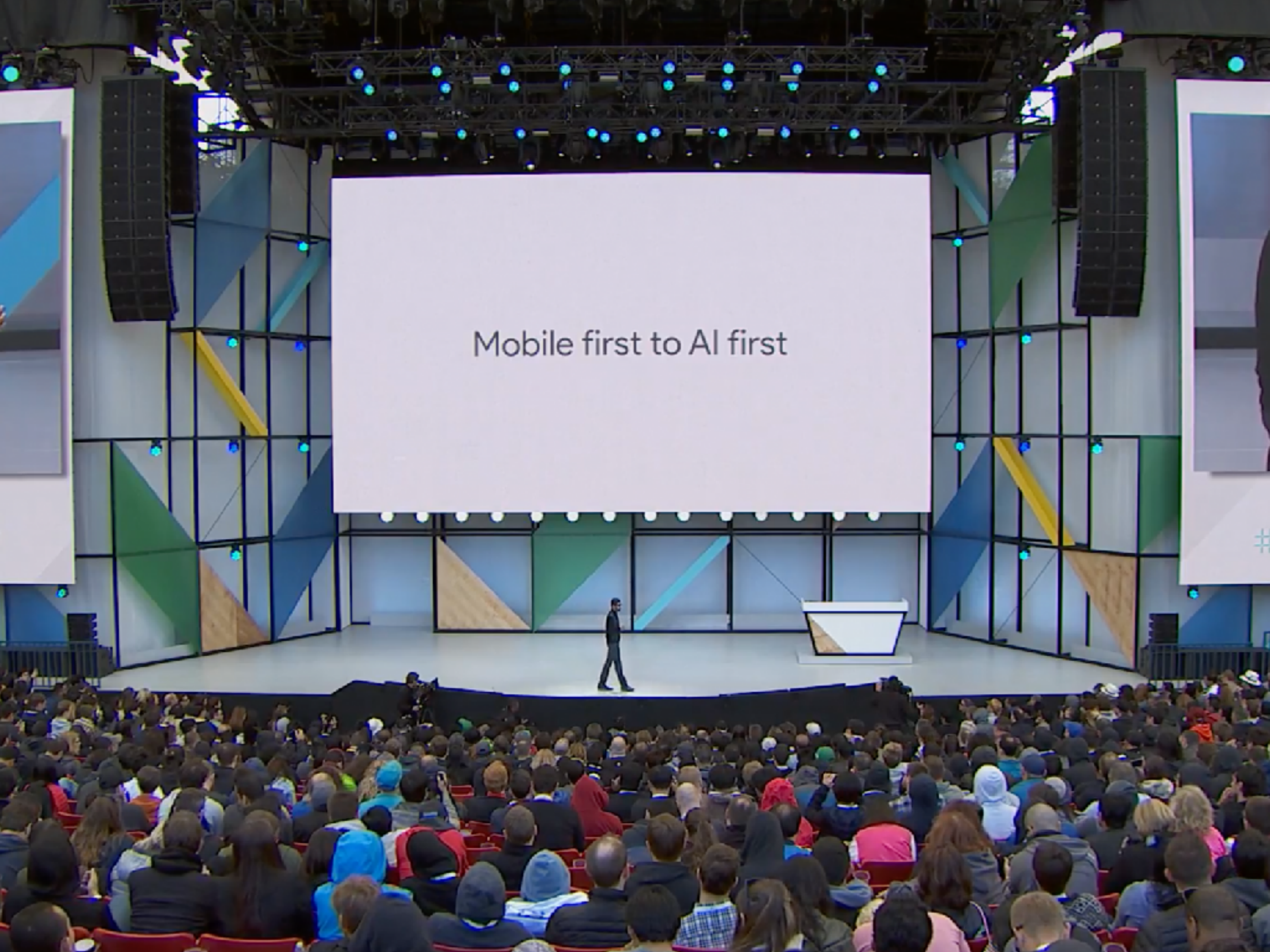
I AM NOT A DATA SCIENTIST

# (BIG) DATA OPPORTUNITIES BEYOND THE BUZZWORD

# THIS IS BOOKING.COM

| Tipologia camera | Per | Il prezzo di oggi | Opzioni disponibili | Scegli le camere | Conferma la tua prenotazione |
|---|---|---|---|---|---|
| Camera Doppia con Letti Singoli con Bagno Privato Esterno — Il nostro consiglio per te<br>Super richiesta!<br>Sclo 1 camera rimasta sul nostro sito!<br>2 letti singoli<br>vista città · TV a schermo piatto · aria condizionata · insonorizzazione · bagno privato · WIFI gratis<br>• vista • TV • scrivania • ferro e asse da stiro<br>• riscaldamento<br>• disponibilità di camere comunicanti<br>• ingresso indipendente<br>• parquet o pavimento in legno<br>• stand appendiabiti • doccia • asciugacapelli<br>• accappatoio • WC • bagno privato<br>• vasca o doccia • bidet • carta igienica<br>• bollitore tè / macchina caffè • frigorifero<br>• bollitore elettrico • macchina da caffè<br>• zona pranzo all'aperto<br>• piani superiori accessibili solo tramite scale<br>• libri, DVD o musica per bambini<br>**Include**: 10 % di IVA, Colazione.<br>**Non include**: 2.00 € di Tassa di soggiorno per persona a notte . | [redacted] | [redacted] | Colazione eccezionale inclusa nel prezzo<br>• **Prezzo basso** – Nessun rimborso<br>✔ **NON SERVE ALCUN PAGAMENTO ANTICIPATO – Paga in struttura** | 0 | Seleziona<br>Conferma immediata<br>I data scientist di Booking.com hanno fatto i conti: 1 altra persona sta guardando questa pagina.<br>**I prezzi potrebbero aumentare. Garantisci oggi la tua prenotazione.** |
| **Bel colpo!** Questo è il prezzo più basso che hai visto per le tue date a Milano.<br>**1 notte ( gio 21 set - ven 22 set )** | | | | | |
| | [redacted] | | Colazione eccezionale inclusa nel prezzo<br>• **Prezzo basso** – Nessun rimborso<br>✔ **NON SERVE ALCUN PAGAMENTO ANTICIPATO – Paga in struttura** | 0 | |

BIG DIVE


Mobile first to AI first

# YES, BIG DATA IS A "BUZZWORD"

## THE TERM AMBIGUITY

**quantitative**
**marketing**
**technology**
**research**

VS

**qualitative**
**innovation**
**process**
**business**

# WHAT'S "NEW" ABOUT DATA

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

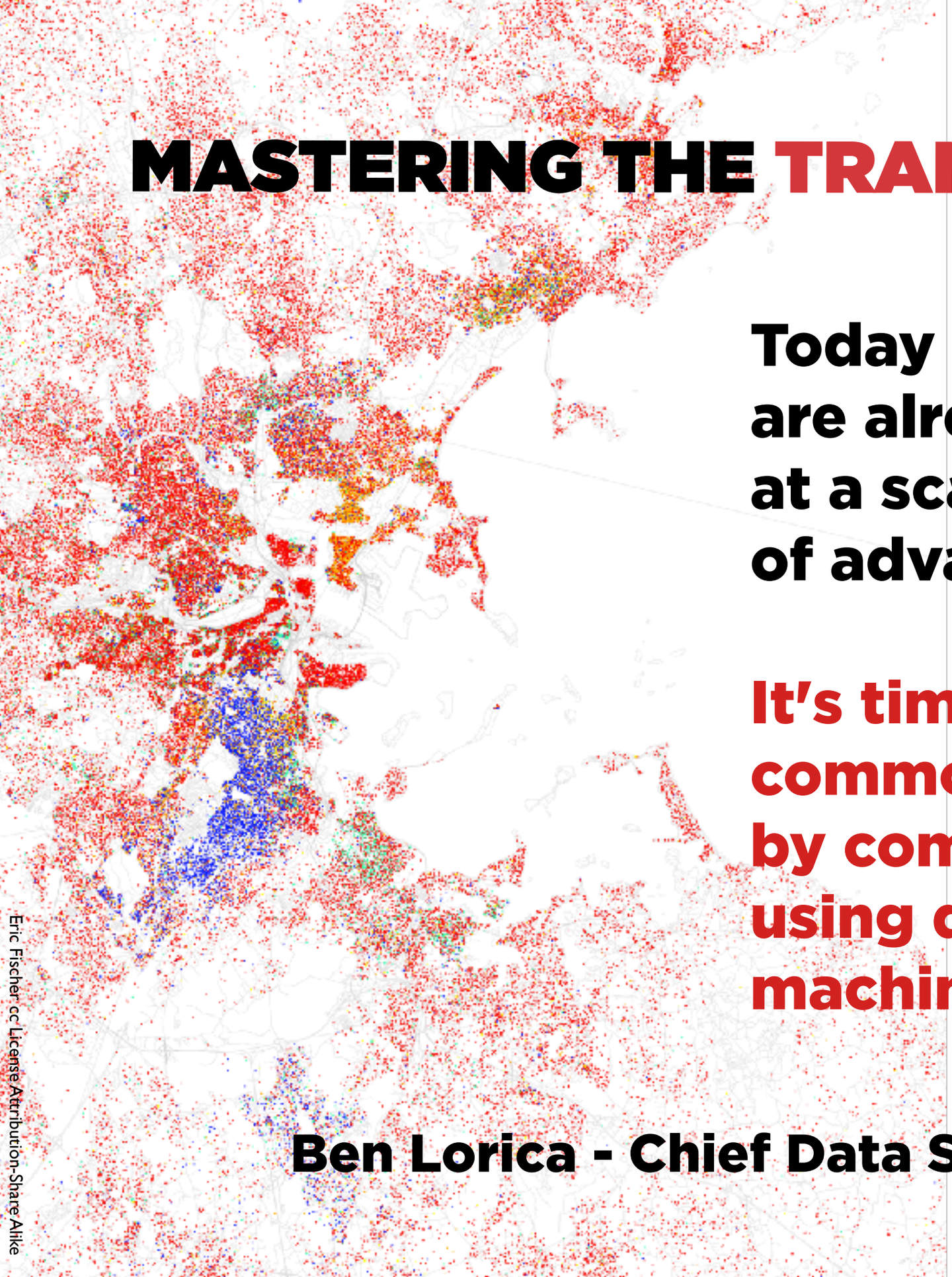
**INFRASTRUCTURE AS A COMMODITY**
/ Cloud
/ HPC & HPN
/ Frameworks


**SKILLS TO EXTRACT INFORMATION ARE NOW MORE ACCESSIBLE**

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

# 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.



# MASTERING THE TRANSITION

**Today many organizations are already collecting data at a scale that justifies use of advanced analytic tools...**

**It's time to focus at the common challenges faced by companies interested in using data science and machine learning.**

**Ben Lorica - Chief Data Scientist at O' Reilly Media**

# COMMON OPEN CHALLENGES

/ THE DATA

/ THE SKILLS

/ FROM PROTOTYPE TO...

/ THE RESULTS INTERPRETATION AND
THE EXPLAINABILITY ISSUE

/ "GREY ZONES" IN DATA EXPLOITATION

/ THE PURSUIT OF INNOVATION

# DATA METADATA FEATURES

## CHALLENGE #1

# DATA REMAINS THE STARTING POINT

## *Volume*

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

# 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}***

# 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.***

## *Data augmentation*

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

**TO RELEVANT DATA**

# 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.***

# 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.***

# DATA QUALITY

| Characteristic | Metric |
|---|---|
| Traceability | Track of creation |
| | Track of updates |
| Currentness | Percentage of current rows |
| | Delay in publication |
| Expiration | Delay after expiration |
| Completeness | Percentage of complete cells |
| | Percentage of complete rows |
| Compliance | Percentage of standardized columns |
| | eGMS Compliance |
| | Five star Open Data |
| Understandability | Percentage of columns with metadata |
| | Percentage of columns in comprehensible format |
| Accuracy | Percentage of accurate cells |
| | Accuracy in aggregation |



*Thanks to* Nexa Center *for Internet & Society*

# DATA QUALITY

***Before***

| Name | Gender | Street | House # | Zip code | City | State | D.O.B |
|---|---|---|---|---|---|---|---|
| John Doe | Male | 60th street | 45 | | New York | New York | 08/12/64 |
| Jane Doe | Female | Jonathan | 36 | 10023 | Poughkeepsy | NY | 21-dec-1954 |

***After***

| Name | Gender | Street | House # | Zip code | City | State | D.O.B |
|---|---|---|---|---|---|---|---|
| John Doe | Male | **E 60th St** | 45 | **10022** | New York | **NY** | 08/12/64 |
| Jane Doe | Female | **Jonathan St** | 36 | 10023 | **Poughkeepsie** | NY | **12/21/54** |

● Completeness ● Accuracy ● Conformity ○ Consistency ● Uniqueness

***Source: http://bit.ly/2xPtNSG***

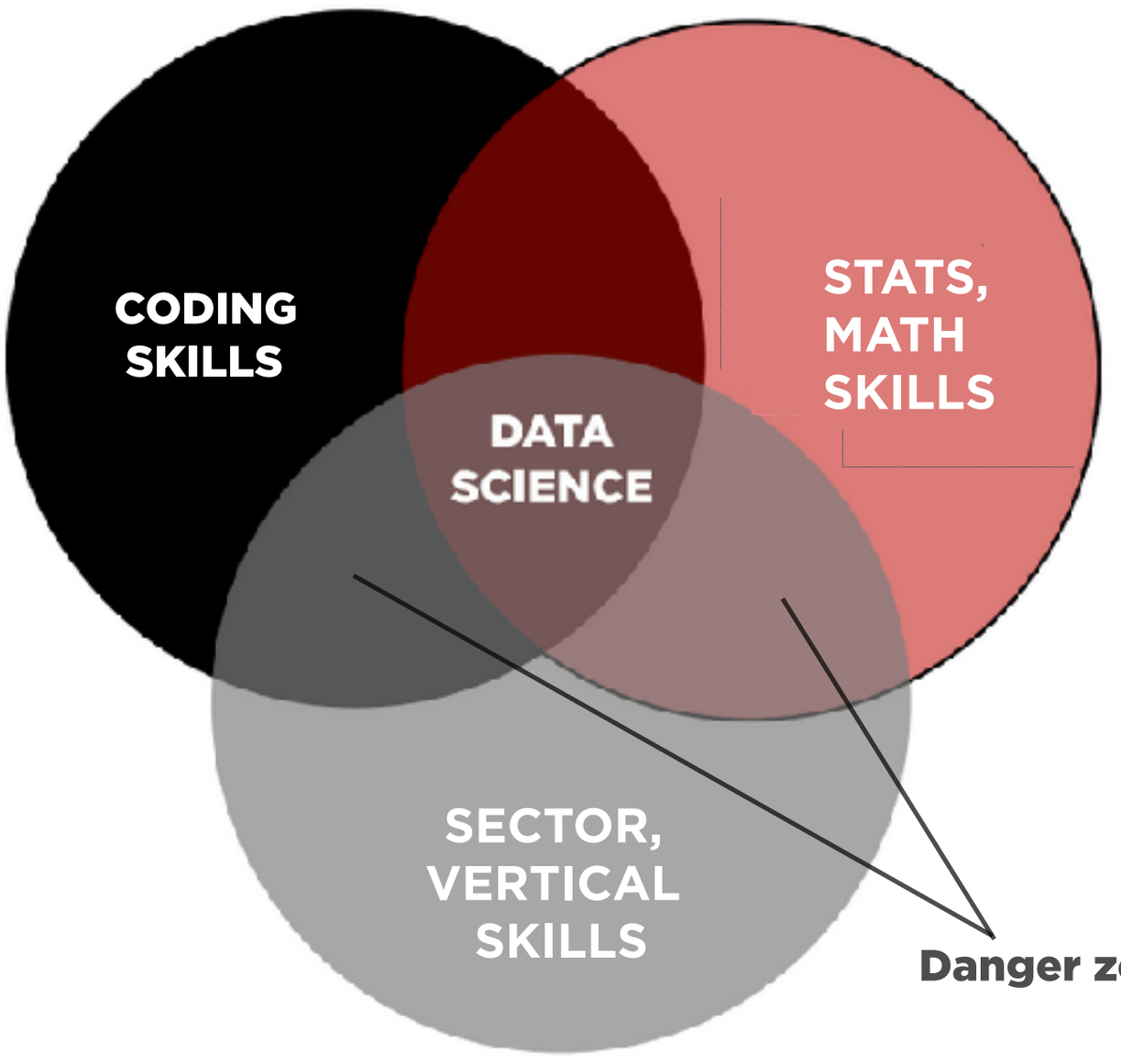
BIG DIVE

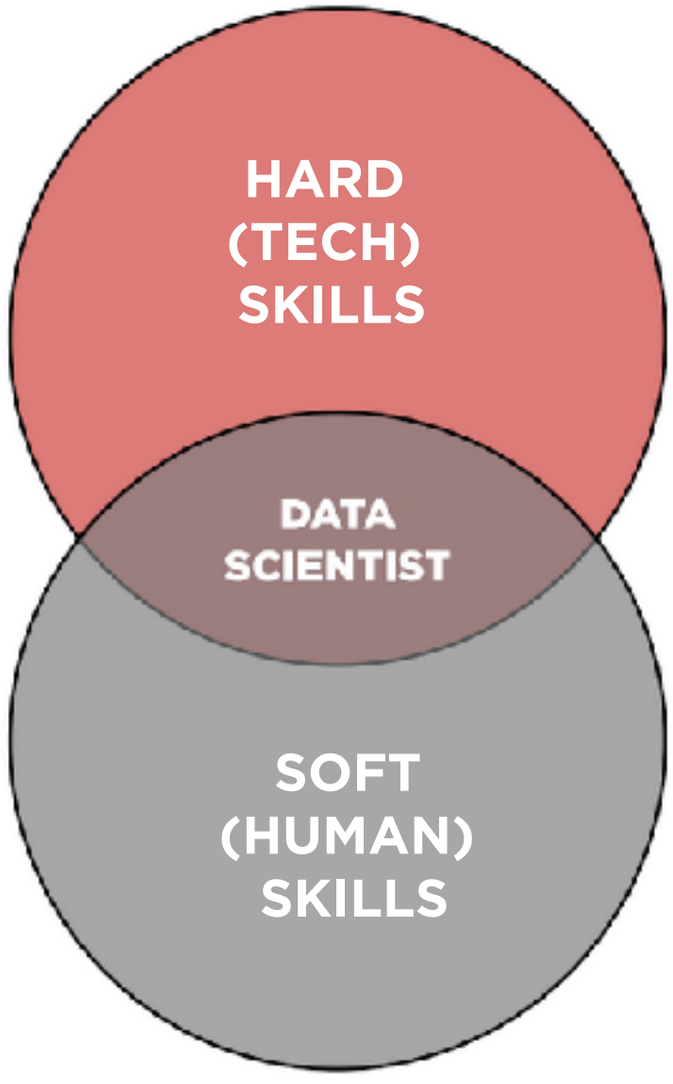
# THE DATA TEAM

## CHALLENGE #2

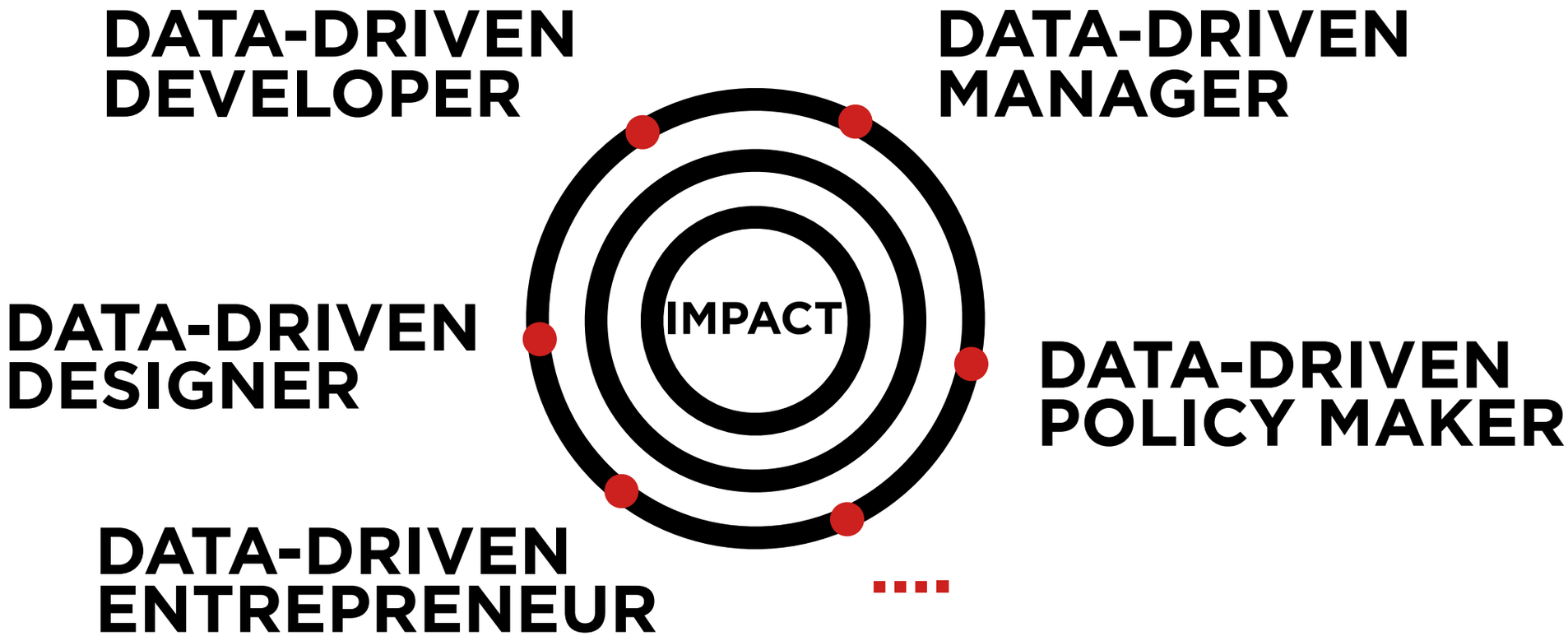
# LOOKING FOR UNICORNS

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



# THE DATA TEAM



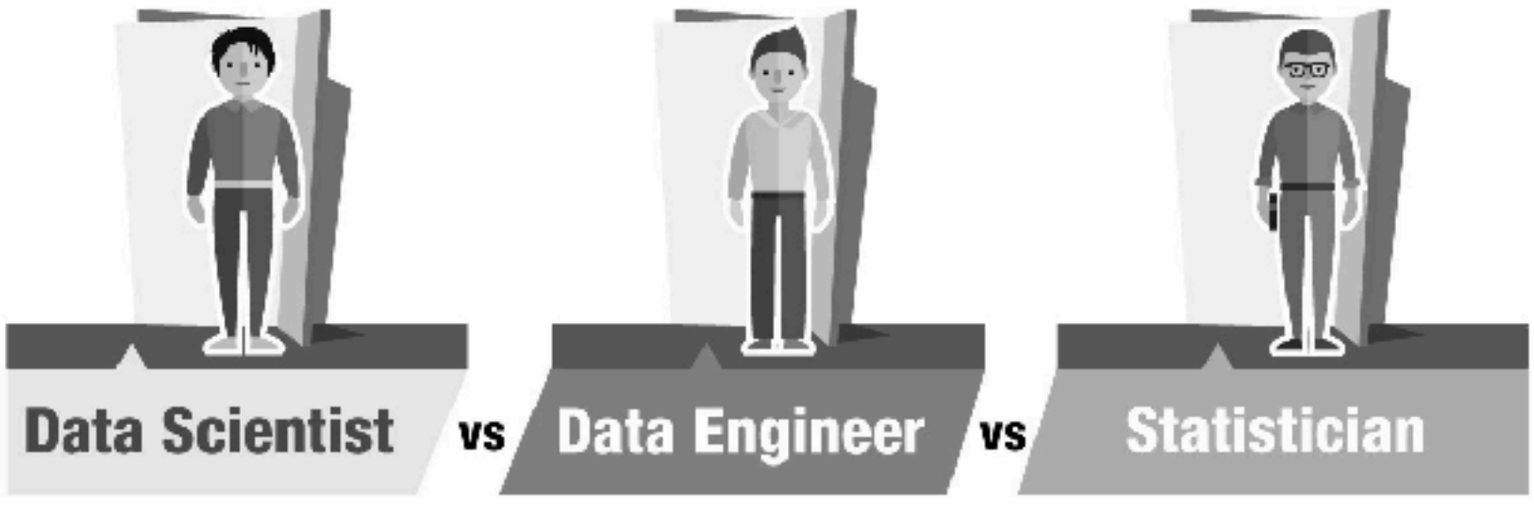

# SINGLE VS TEAM

# FROM PROTOTYPE TO PRODUCTION

## CHALLENGE #3

# COMPARISON

## — Jobs in Data Science —




| | Data Scientist | Data Engineer | Statistician |
|---|---|---|---|
| | These people use their analytical and technical capabilities to extract meaning insights from data. | These people ensure uninterrupted flow of data between servers and applications. They are responsible for data architecture. | These people understand statistics theoretically and apply them to real life problems. |
| Responsibilities | Develop and plan required analytic projects in response to business needs.<br><br>Contribute to data mining architectures, modeling standards, reporting, and data analysis methodologies.<br><br>Collaborate with stakeholders to integrate data mining results with existing systems.<br><br>Monitor data mining system performance and implement efficiency improvements. | Design, construct, install, test and maintain highly scalable data management systems<br><br>Improve data foundational procedures, guidelines and standards<br><br>Integrate new data management technologies and software engineering tools into existing structures<br><br>Create custom software components (e.g. specialized UDFs) and analytics applications | Apply statistical theories and methods to solve practical problems of various industries<br><br>Determine methods for finding or collecting data<br><br>Design surveys or experiments or opinion polls to collect data<br><br>Analyze, interpret & undertake data analysis<br><br>Report conclusions from their analyses |
| Skills | Programming, Mathematics, Business Understanding, Statistics, Data Visualization, Machine Learning, Attention to detail | Database design, Production coding, Data collection, data warehousing, Data transformation, Work diligently with data | Technical and Analytics Skills, Mathematics, Operational Research, Writing skills, Ability to Analyze, Model and interpret data, Flair of explaining difficult concepts in simple manner |

BIG DIVE

Original infographics from Analytics Vidhya

# A BABEL OF (CODING) LANGUAGES

***PRODUCTION***

**JAVA, C, C++, ...**

***REFACTORING***

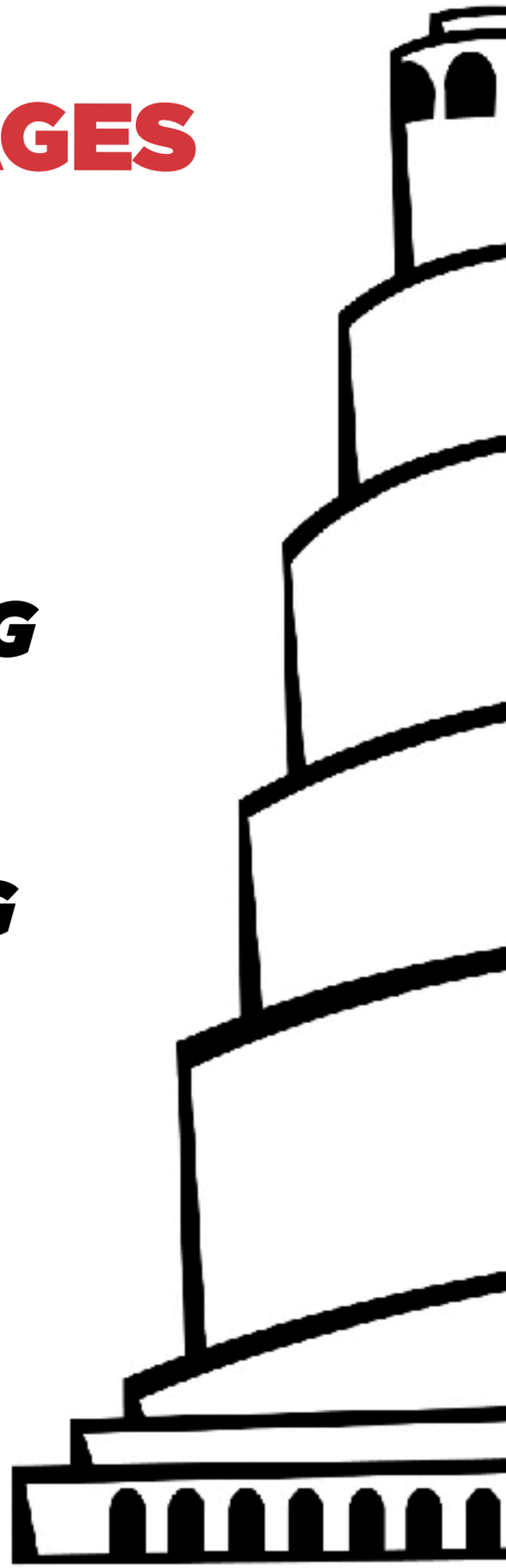
***-***

***DATA ENGINEERING***

**[ SCALA, ...]**

***DATA-DRIVEN PROTOTYPE***

**PYTHON, R, D3.JS**



BIG DIVE

# RESULTS INTERPRETATION & EXPLAINABILITY

## CHALLENGE #4

# THE EXPLAINABILITY ISSUE

**LOW EXPLAINABILITY**

***Deep learning***

***Machine learning***

***Inferential statistics***

**HIGH EXPLAINABILITY**

# MACHINE LEARNING IS NOT PERFECT

## BAD LEARNING CAUSES:

/ Errors in the training set

# GARBAGE-IN GARBAGE-OUT

**MACHINE learns ONLY through the training data**
**(no prejudices, no additional elaboration)**

***Training set:***
***2+2 = 5***
***2+2 = 5***
***2+2 = 5***
***2+2 = 5***
***2+2 = 5***
***2+2 = 4***

**>>> MACHINE SAYS THAT 2+2 = 5**

# MACHINE LEARNING IS NOT PERFECT

## BAD LEARNING CAUSES:

/ Errors in the training set

# MACHINE LEARNING IS NOT PERFECT

## BAD LEARNING CAUSES:

/ Errors in the training set

/ Errors in generalization

# MACHINE LEARNING IS NOT PERFECT

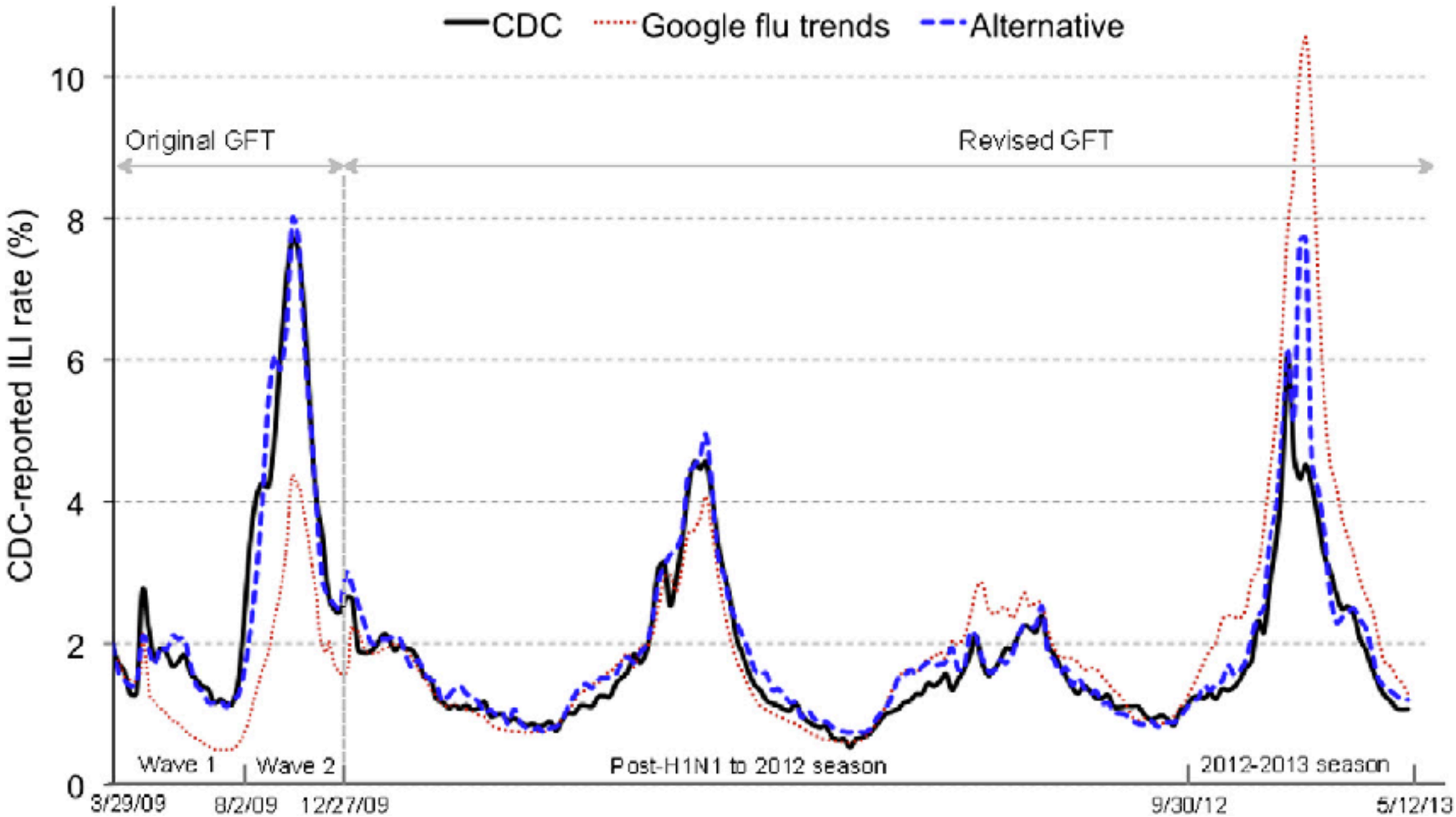
## BAD LEARNING CAUSES:

/ Errors in the training set

/ Errors in generalization

/ Missing update in the model

# ABOUT CORRELATION: GOOGLE FLU




BIG DIVE

# MACHINE LEARNING IS NOT PERFECT

## BAD LEARNING CAUSES:

/ Errors in the training set

/ Errors in generalization

/ Missing update in the model

# MACHINE LEARNING IS NOT PERFECT

## BAD LEARNING CAUSES:

/ Errors in the training set

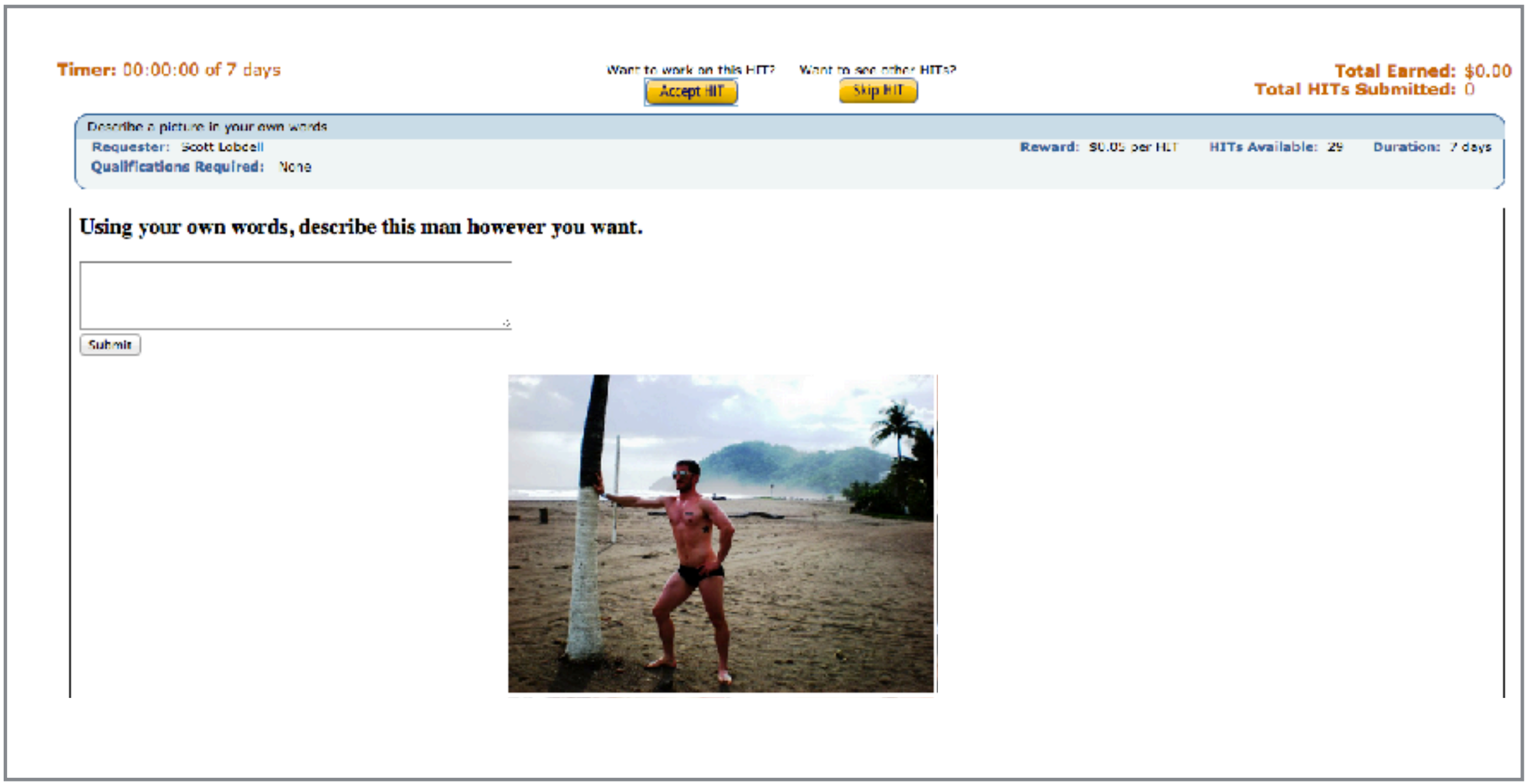
/ Errors in generalization

/ Missing update in the model

/ Human errors

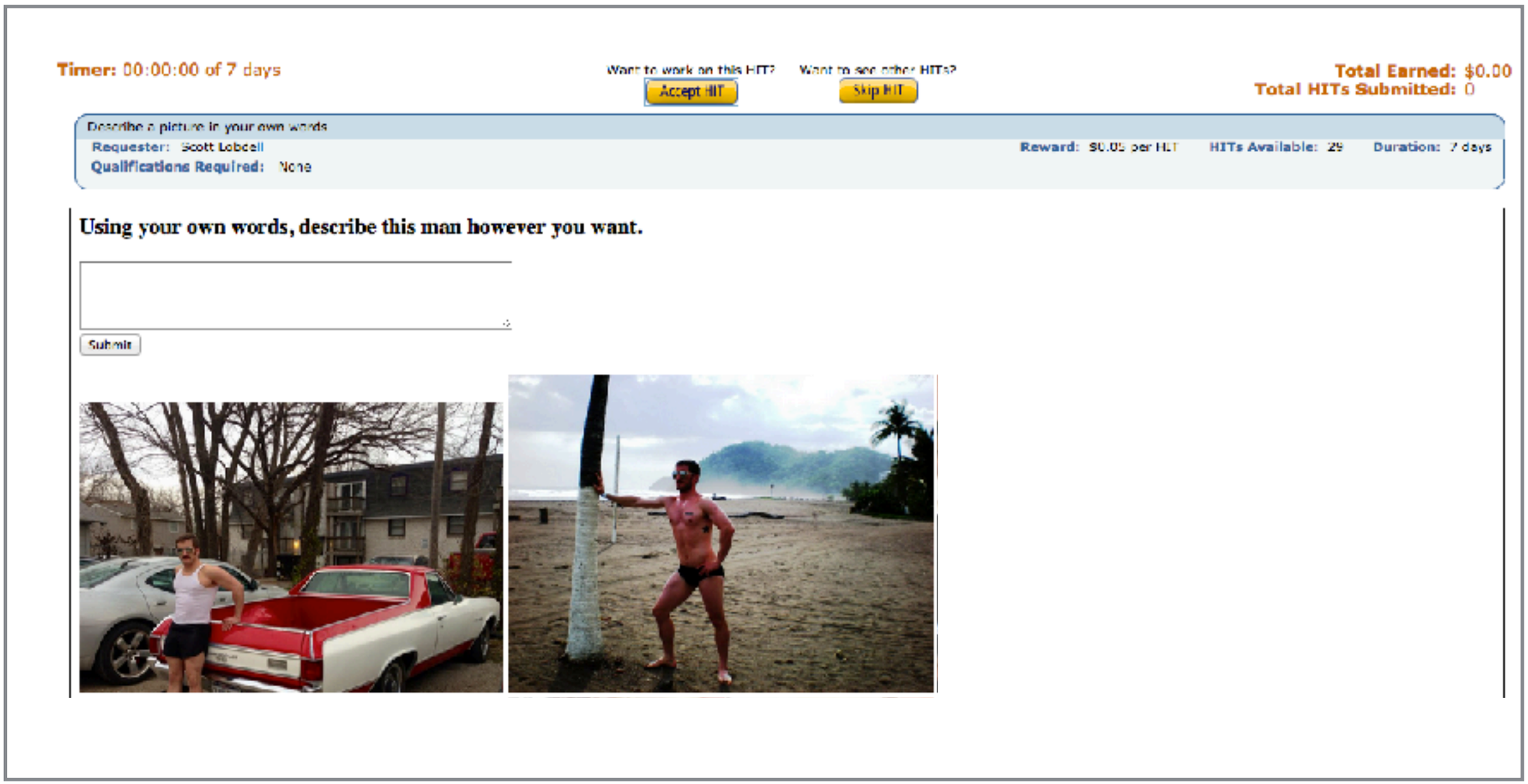
# MACHINE LEARNING IS NOT PERFECT

## One of the biggest error in creating a training dataset is assuming people think and behave in a common-similar way.



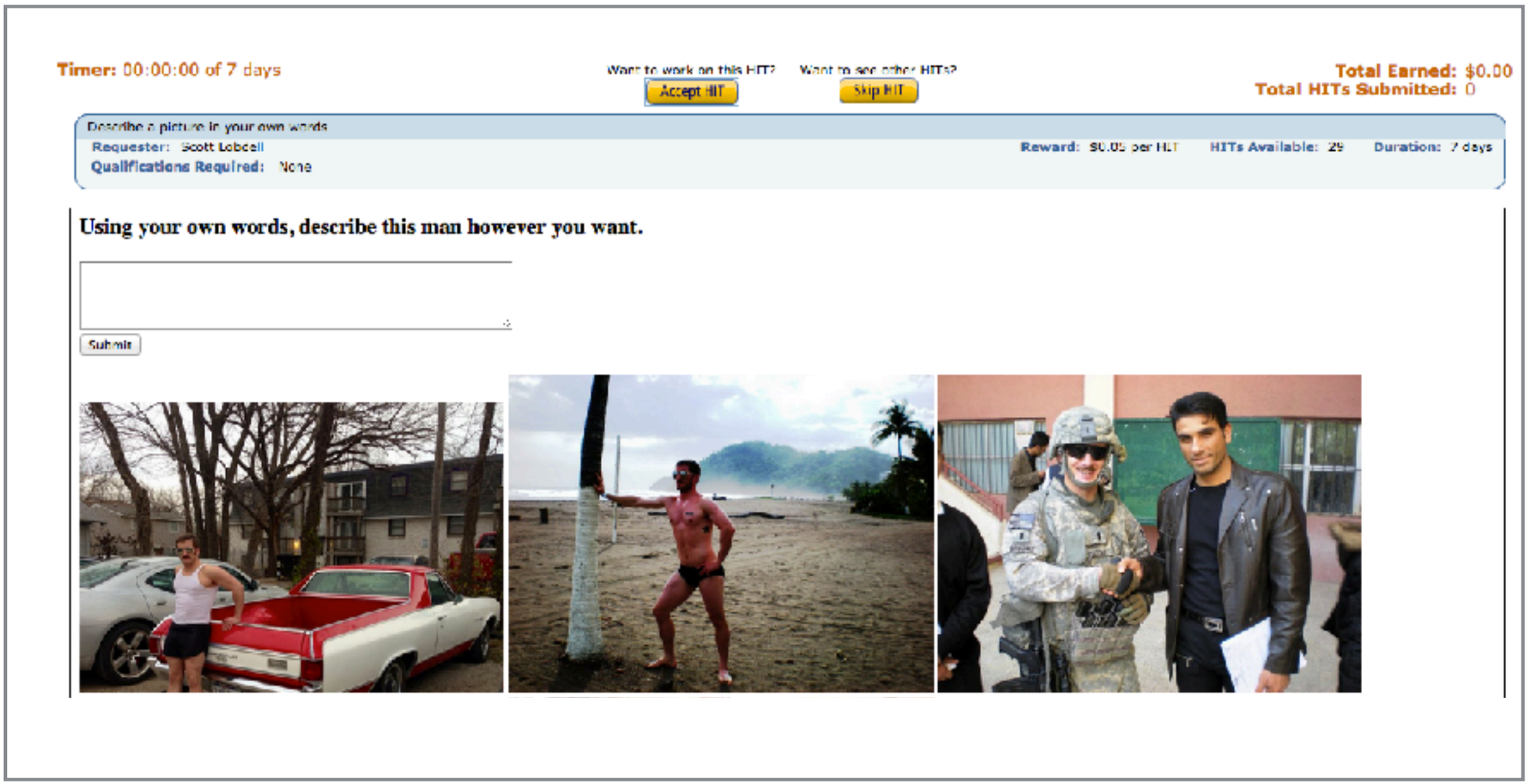
# MACHINE LEARNING IS NOT PERFECT

## One of the biggest error in creating a training dataset is assuming people think and behave in a common-similar way.



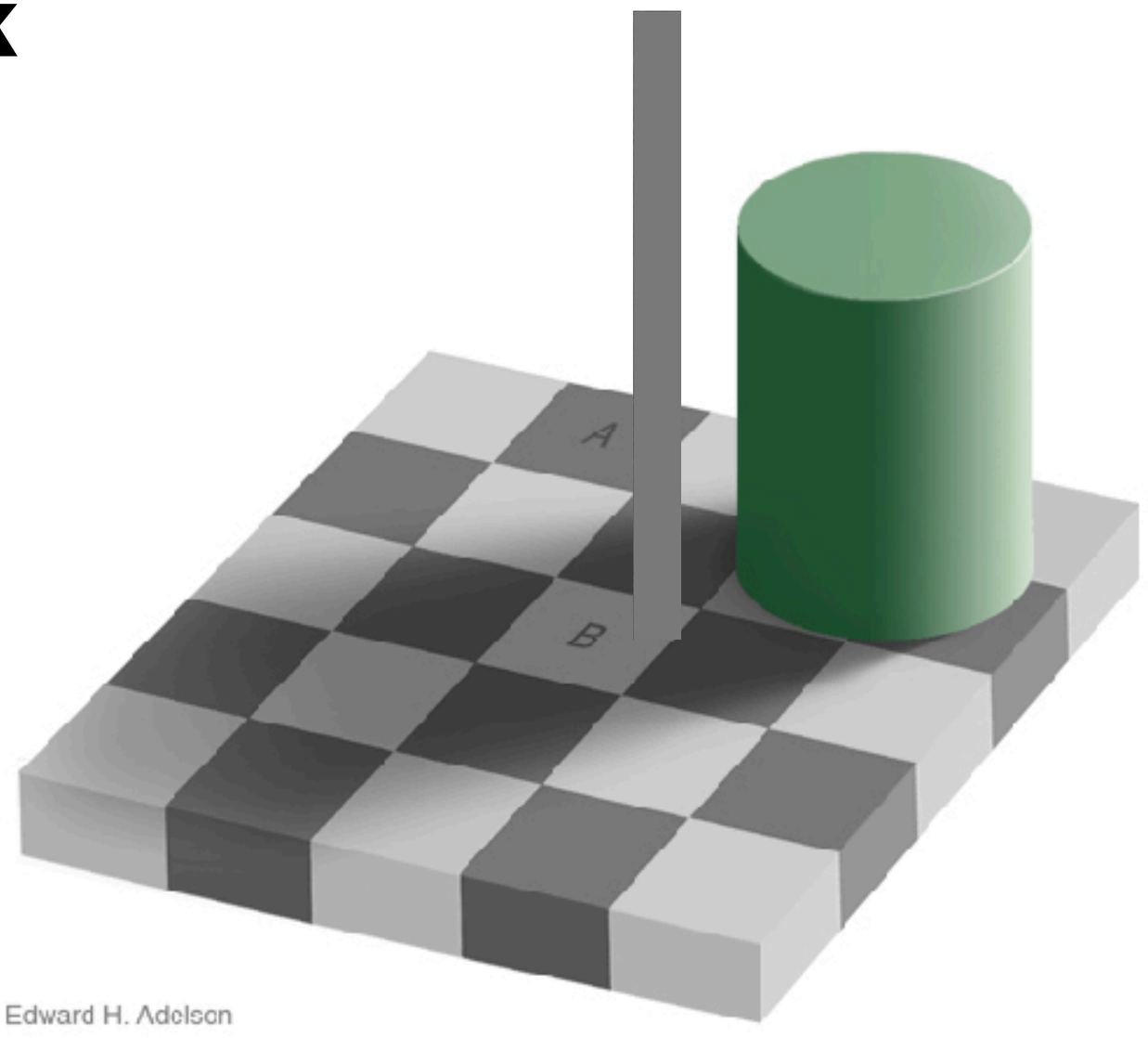
# MACHINE LEARNING IS NOT PERFECT

## One of the biggest error in creating a training dataset is assuming people think and behave in a common-similar way.



# MACHINE LEARNING IS NOT PERFECT

## OPTICAL ILLUSIONS TRICK HUMAN MIND: do A & B squares have the same color ?




Edward H. Adelson

# GREY ZONES IN DATA EXPLOITATION

## CHALLENGE #5

# THE DARK SIDE OF BIG DATA

/ DATA DEMOCRACY

/ CORRELATION DOES NOT MEAN CAUSATION

/ BUBBLE FILTERS

/ BIAS

/ PRIVACY & LEGAL ISSUES

/ HUMAN RIGHTS

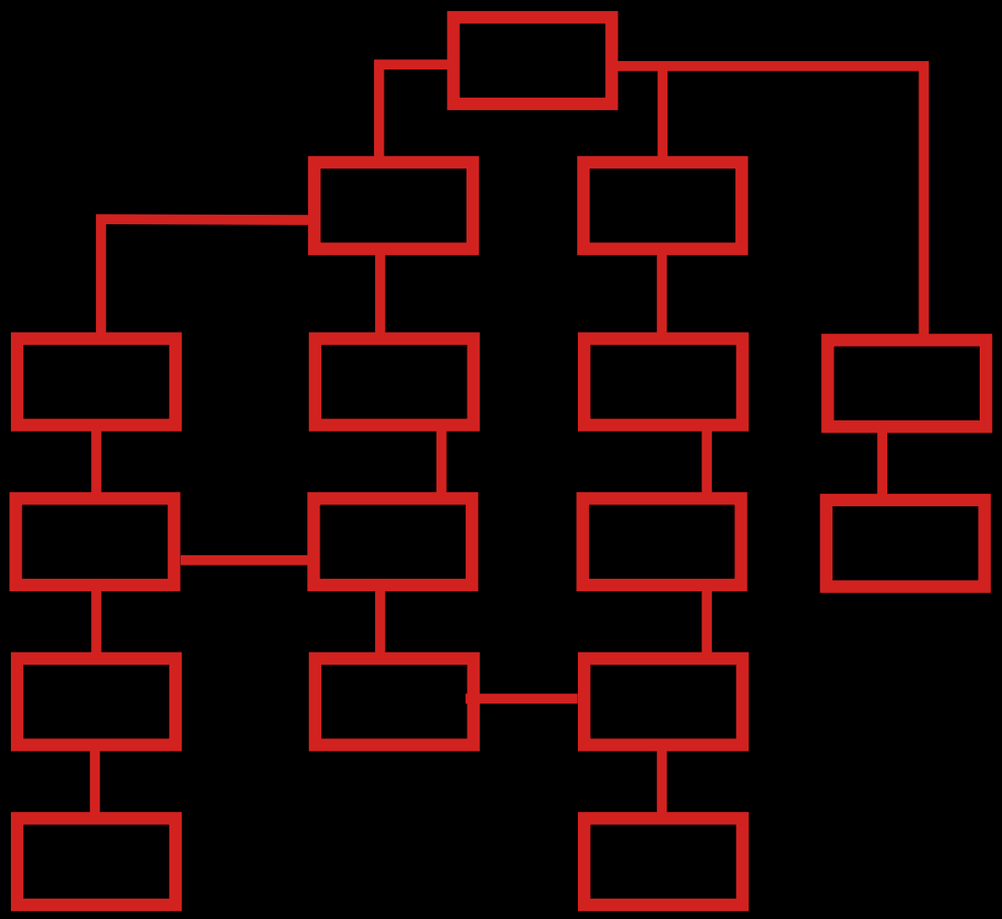
/ ...

# GRID NATIVES
# VS
# COMPLEX NATIVES

## CHALLENGE #6

# HOW TO BECOME A DATA DRIVEN COMPANY ?

# THE GRID VS THE COMPLEXITY



| GRID "NATIVES" | COMPLEX "NATIVES" |
| --- | --- |
| **HAVE LOW INTERCONNECTION** | ***Have high interconnection*** |
| **DON'T LIKE CHANGES & "CHAOS"** | ***Change a lot & like chaos (edge)*** |
| **GET OLD AND NEED CONTINUOUS FIXING** | ***Get old but also evolve/ adapt (auto-repair)*** |
| **ARE HIGHLY EXPOSED TO BIG "CRISIS" (black swans)** | ***Prefer failing (small mistakes) fast & soon*** |
| **YOU NEED A MAP TO UNDERSTAND THEM** | ***You need dynamic data to understand them*** |

# GRID “NATIVES”

**SOFTWARE VENDORS**

**BANKS**

**HARDWARE VENDORS**

**PORTALS**

**...**

# COMPLEX “NATIVES”

***service (experience) vendors***

***mobile payment platforms***

***cloud vendors***

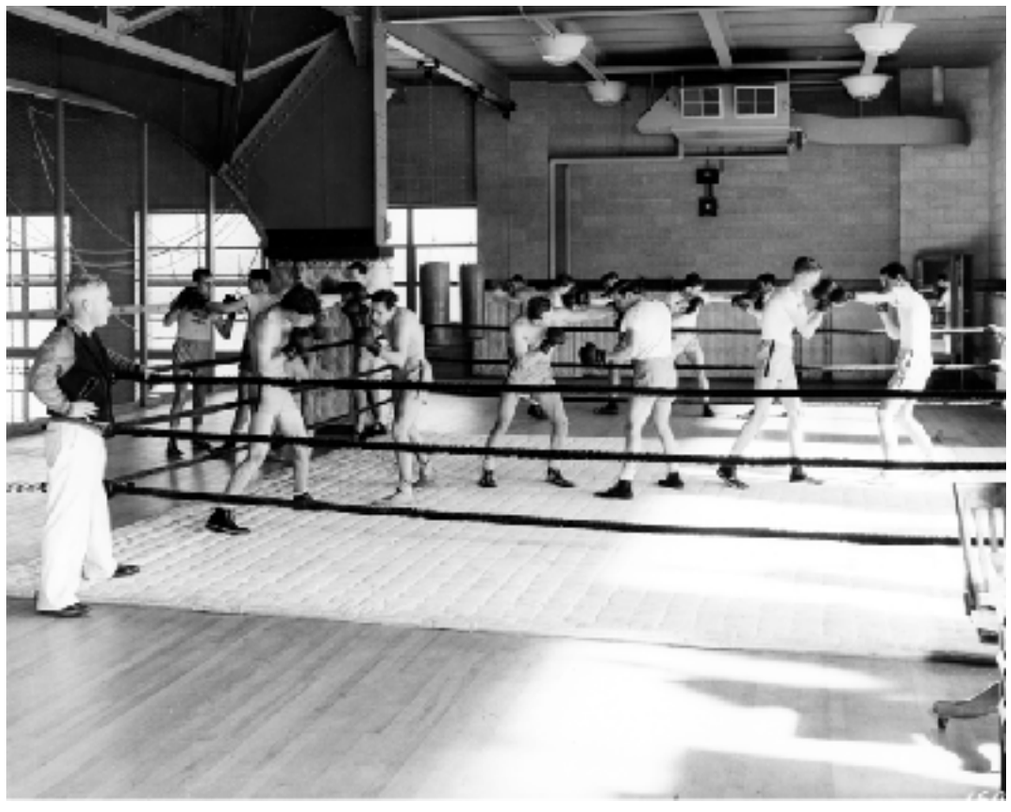
***social networks***

**...**

BIG DIVE

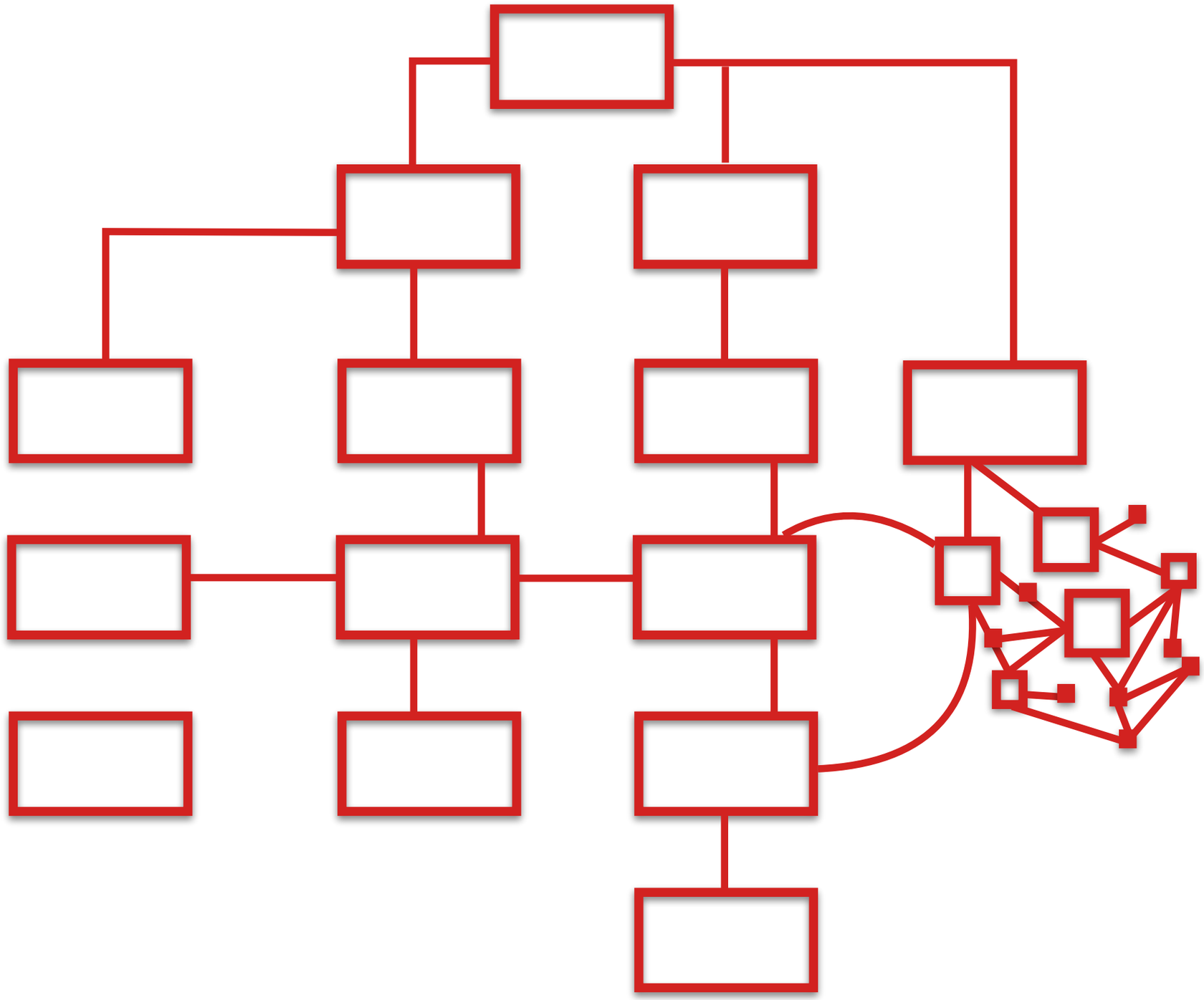
# THE BIG DIVE CREDO

## WE DON'T NEED A UNIVERSITY

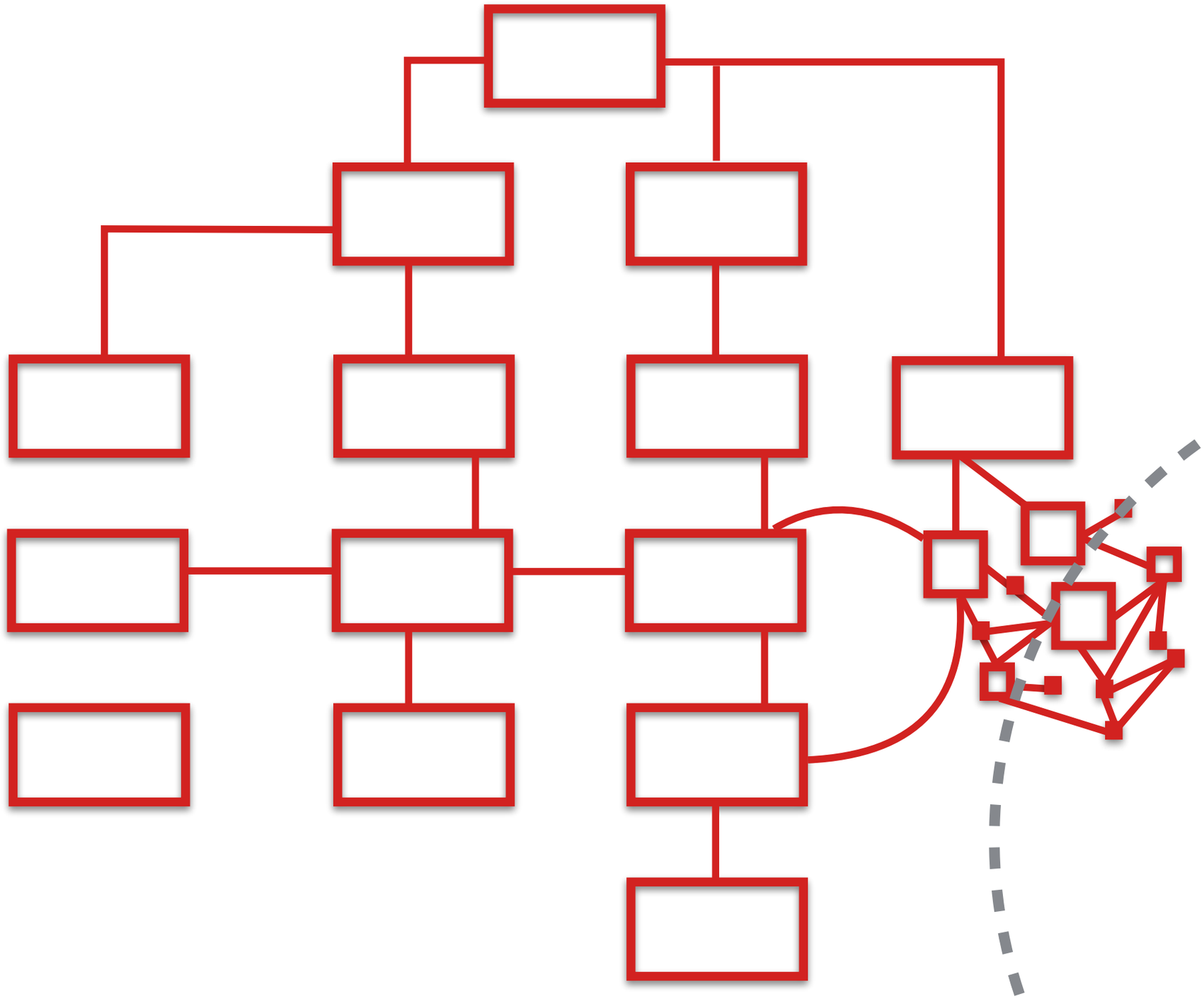


## WE NEED STREET FIGHTING

# MAKING THE GRID *MORE COMPLEX*



# IN VS OUT

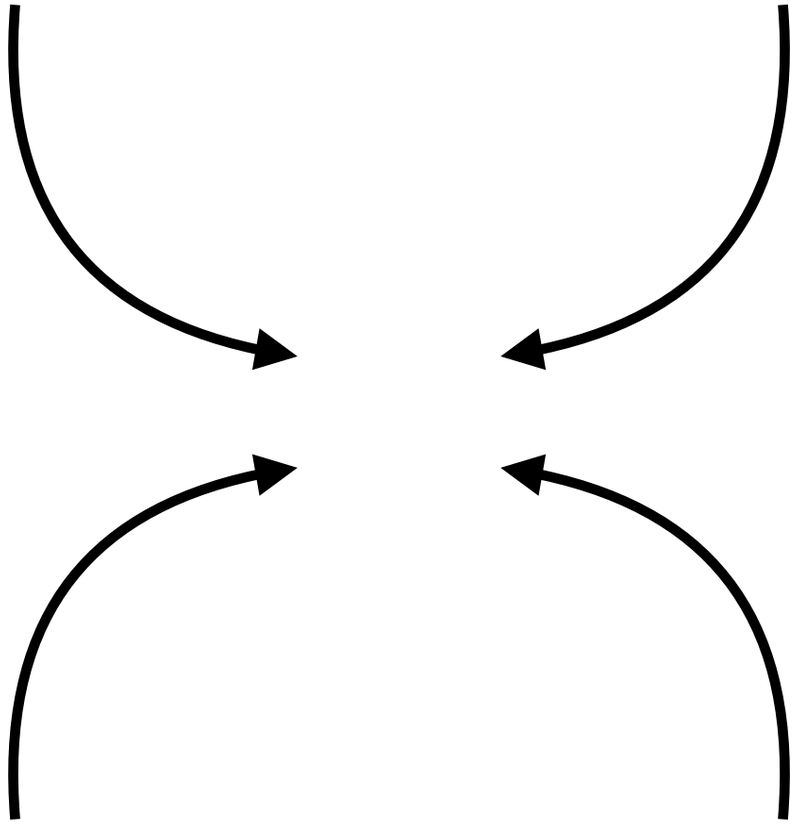


OUT OF THE BUILDING

# MASTERING THE TRANSFORMATION

**DATA & TOOLS**

***Open Approach***



***Complexity & Interconnection***

**SOFT & HARD SKILLS**

# LESSONS LEARNED ABOUT DATA

1. **DATA is a core asset that it is advisable to develop internally (at the companies).**

2. **DON'T FOCUS ON TOOLS, FOCUS ON SCIENCE & INNOVATION PROCESSES.**

3. **GO OPEN (source)! Open Community is the new R&D department.**

4. **Approach must be SYSTEMIC, LEARN-BY-DOING, DYNAMIC, VALUE-BASED.**

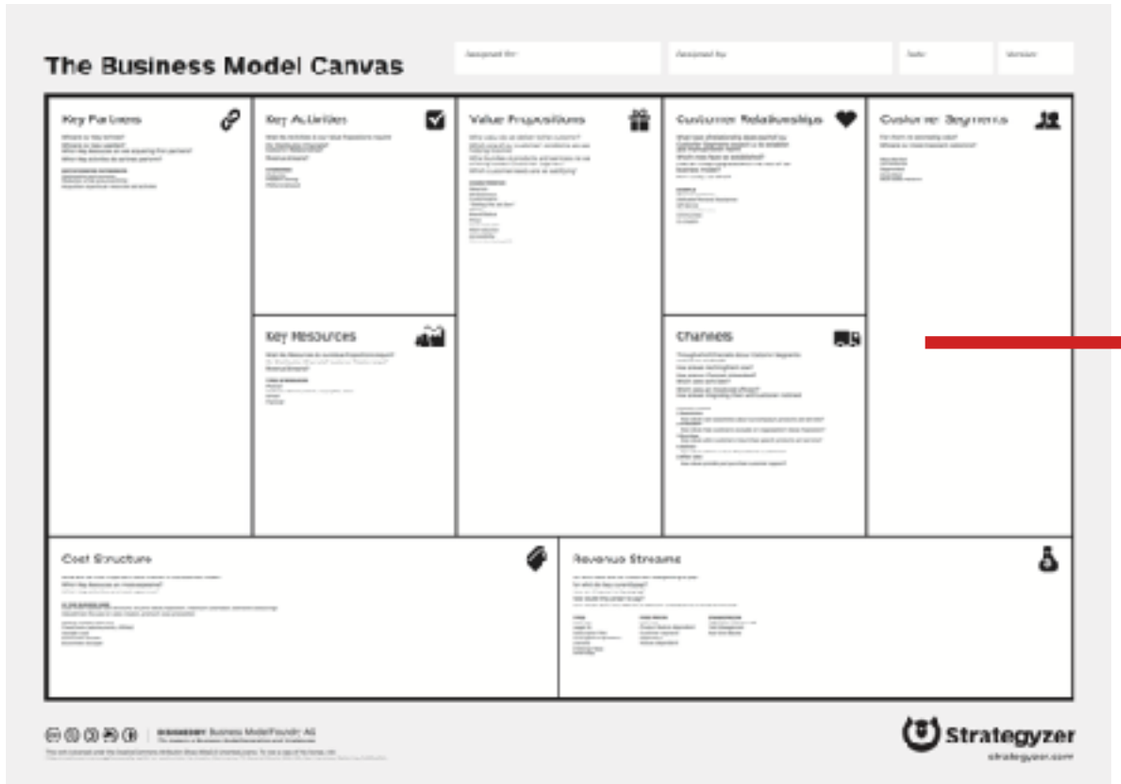
5. **DECENTRALIZATION is the new mantra.**

6. **NEVER SEPARATE DATA FROM CONTEXT & SOURCES.**

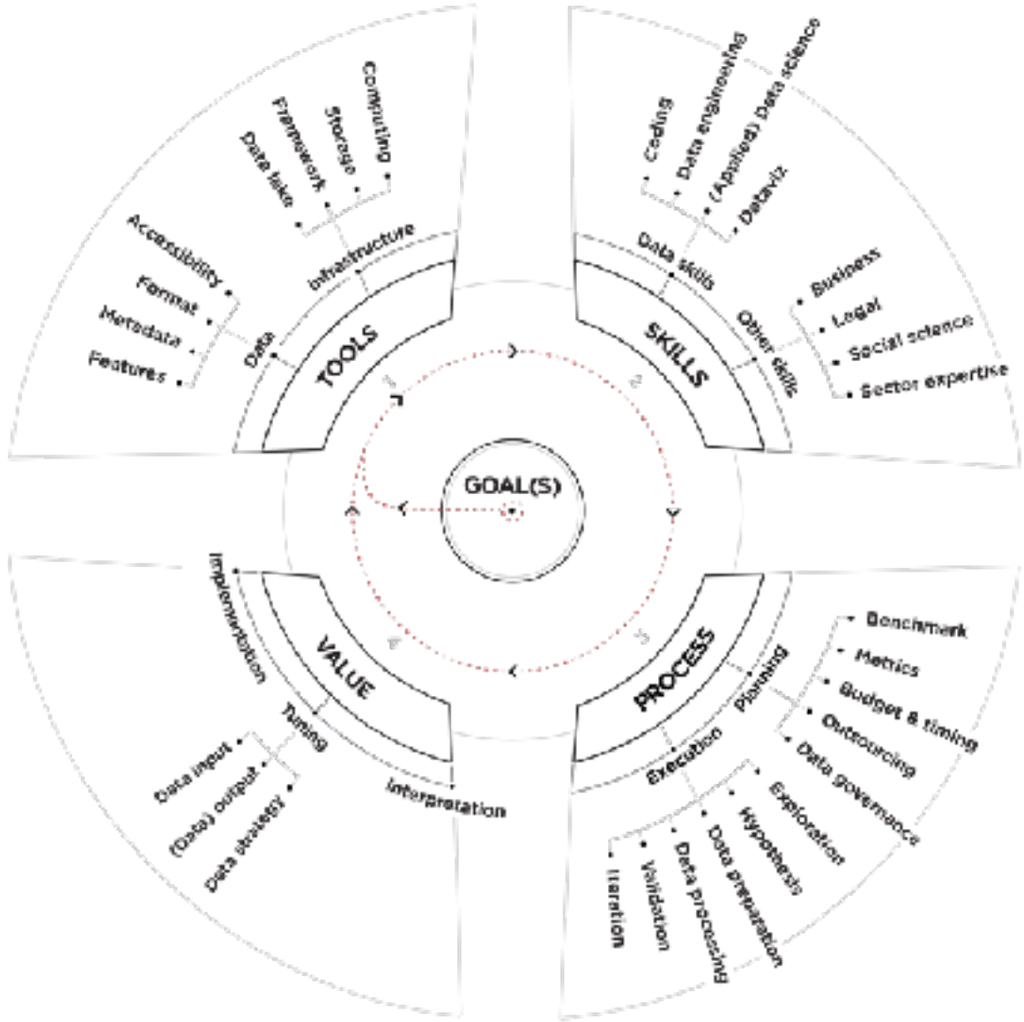
# LEVERAGING (BIG) DATA OPPORTUNITIES REQUIRES A PROPER METHOD

# THE DATA RING

# THE CANVAS APPROACH




The inspiring precursor




The Data Ring

BIG DIVE

# WHY A DATA 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.

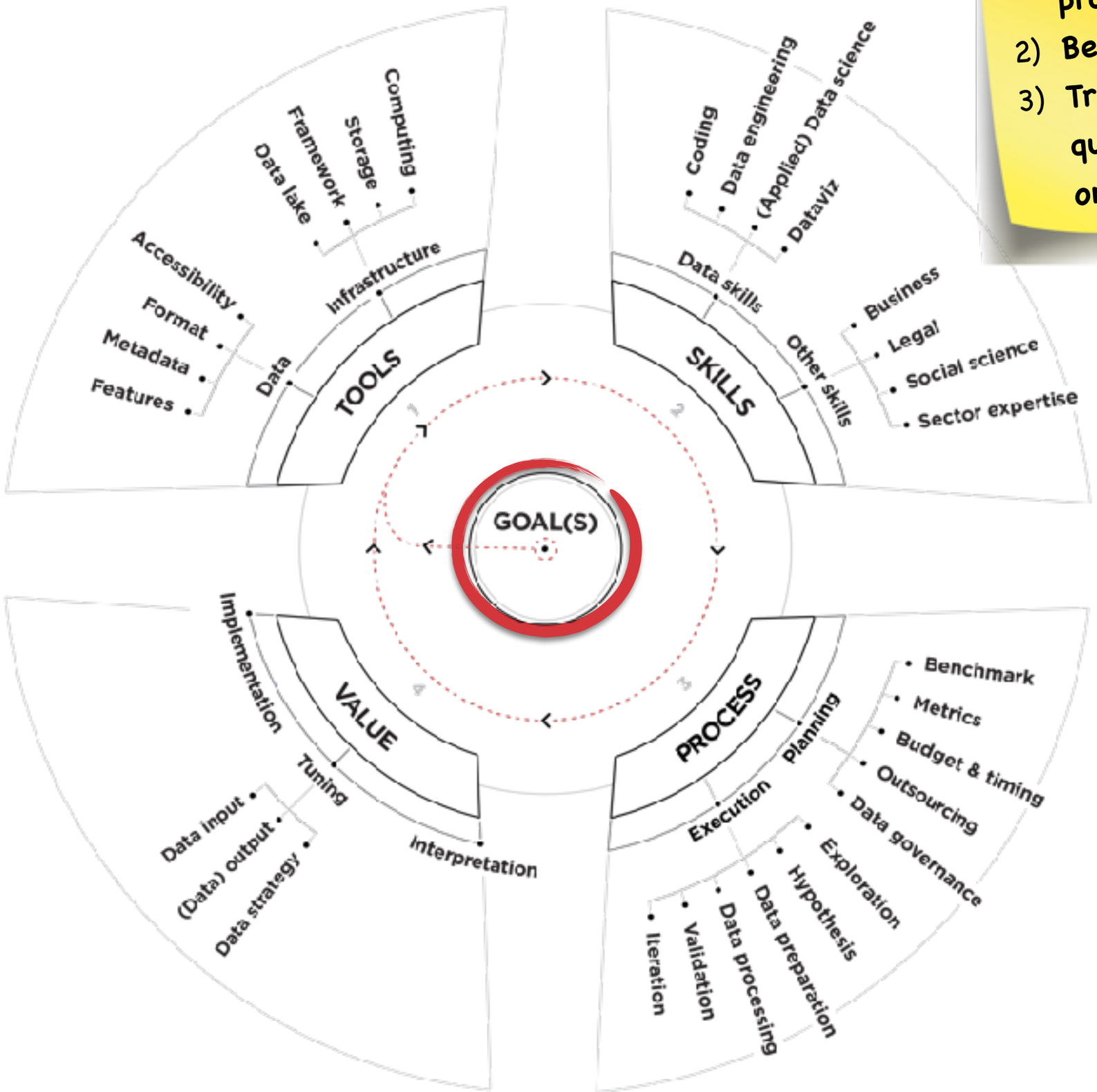
/ It provides for a complete representation of the process that can be explained to third parties too.

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

# DATA RING CANVAS


BIG DIVE

# GOAL(S) SETTING




BEST PRACTICES:

1) Start from the problem-setting
2) Be specific!
3) Try to be quantitative (not only qualitative)

BIG DIVE

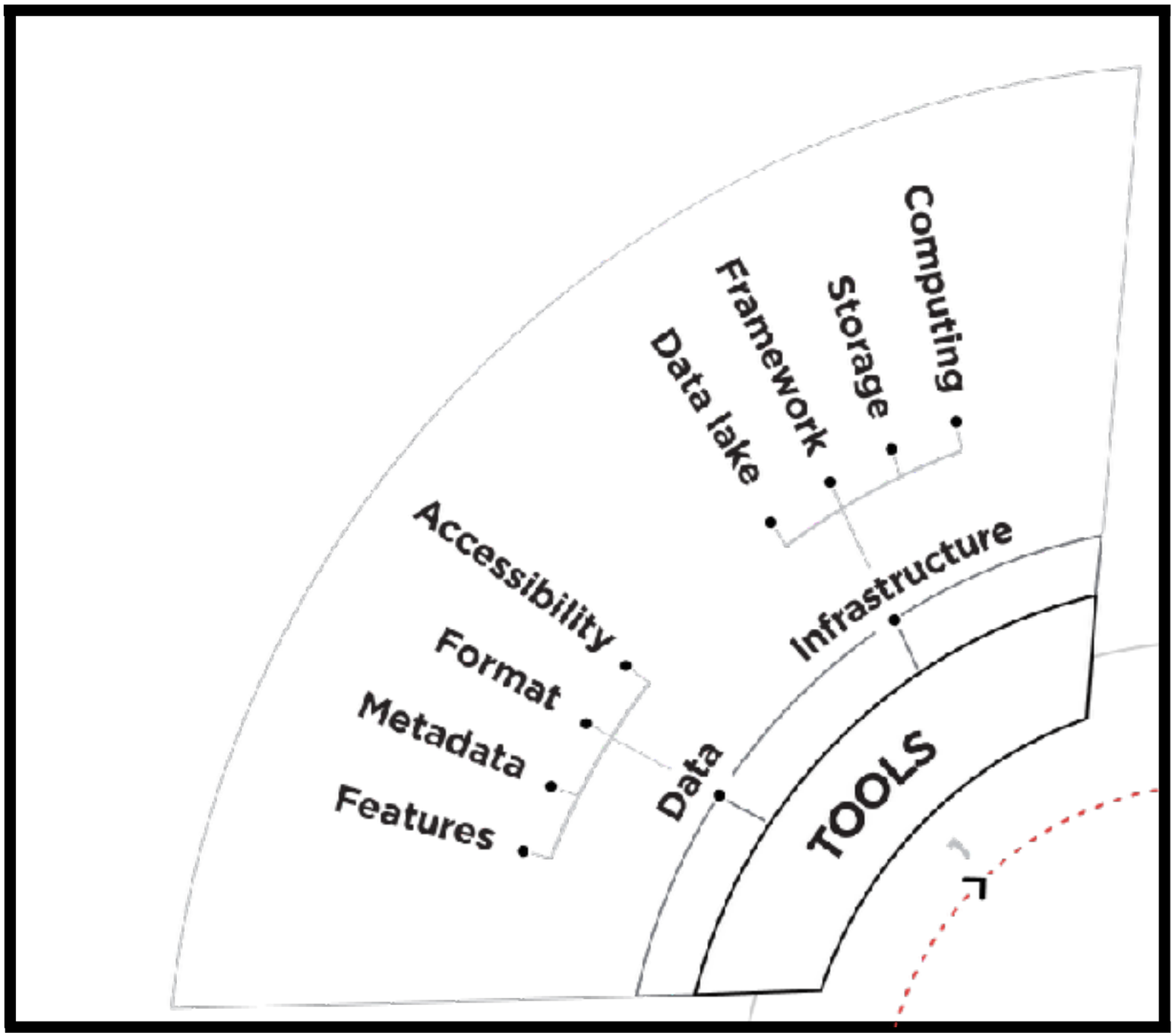
# GOAL(S) SETTING


**RISKS**

- ✓ LACK OF FOCUS
- ✓ TORTURED GOALS
- ✓ NOT PRAGMATIC GOALS

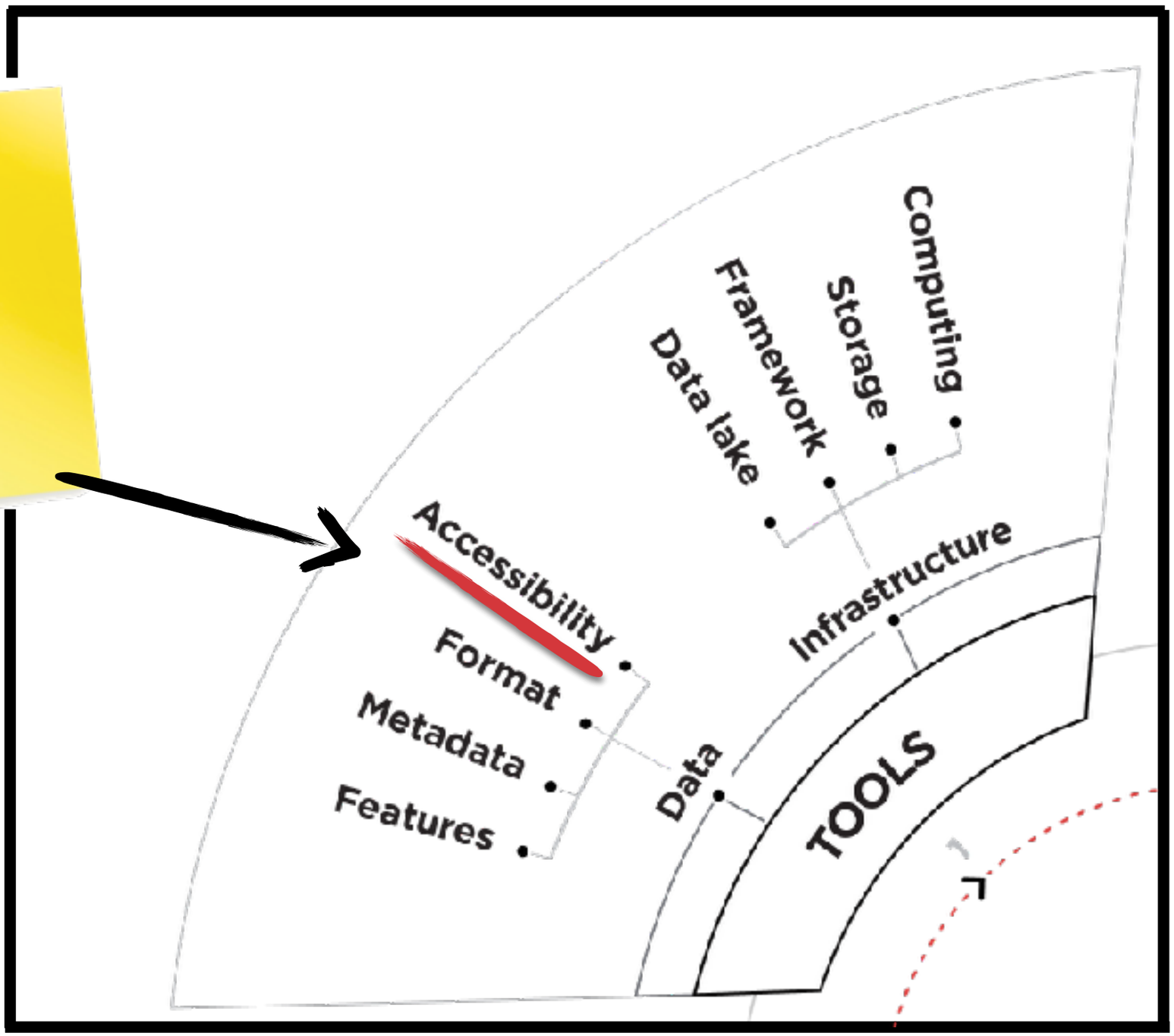
BIG DIVE

# TOOLS

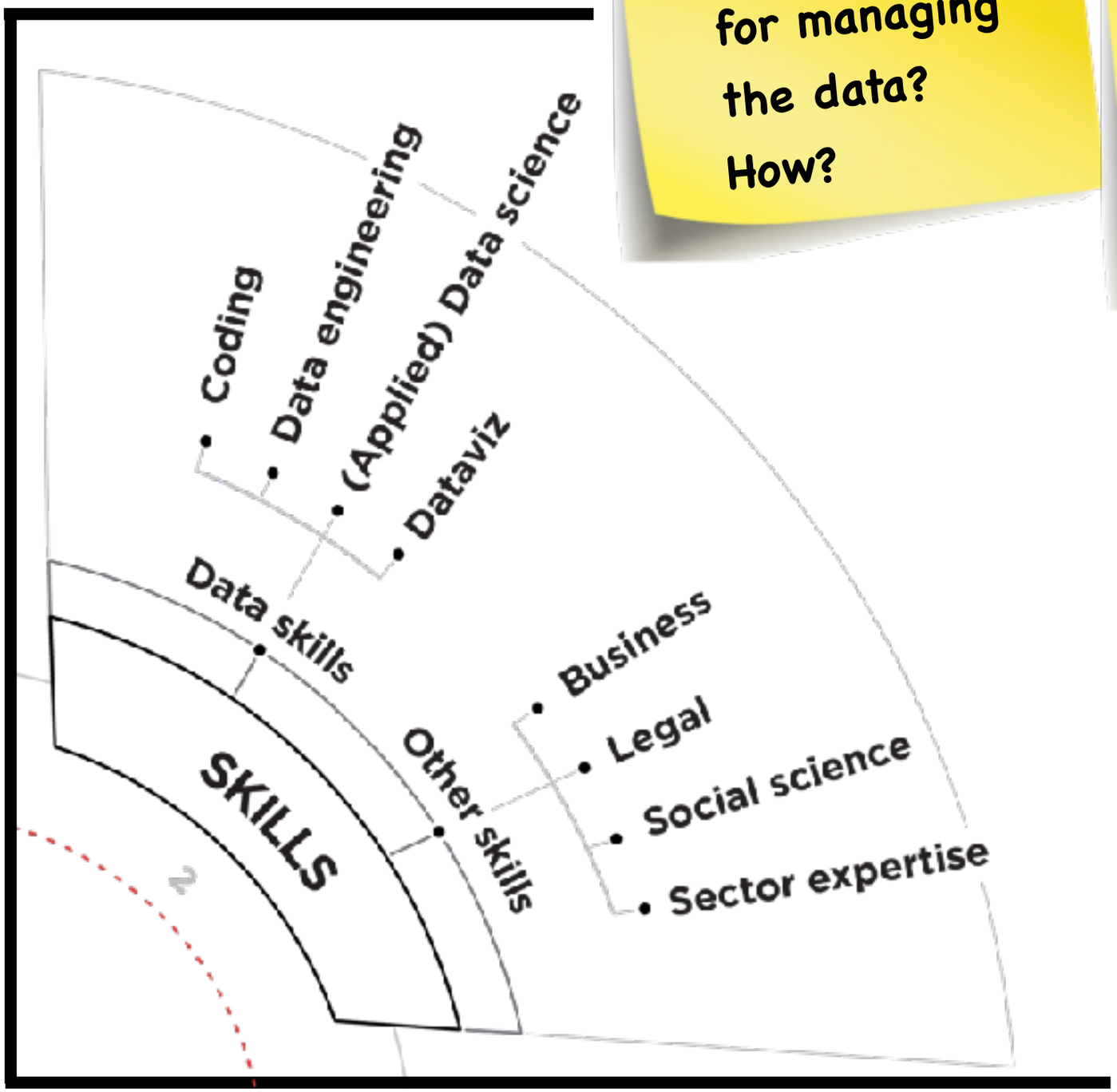

Computing
Storage
Framework
Data lake
Infrastructure
Accessibility
Format
Metadata
Features
Data
TOOLS

# TOOLS

- ✓ Legal barriers
- ✓ Strategic barriers
- ✓ Technological barriers



# SKILLS



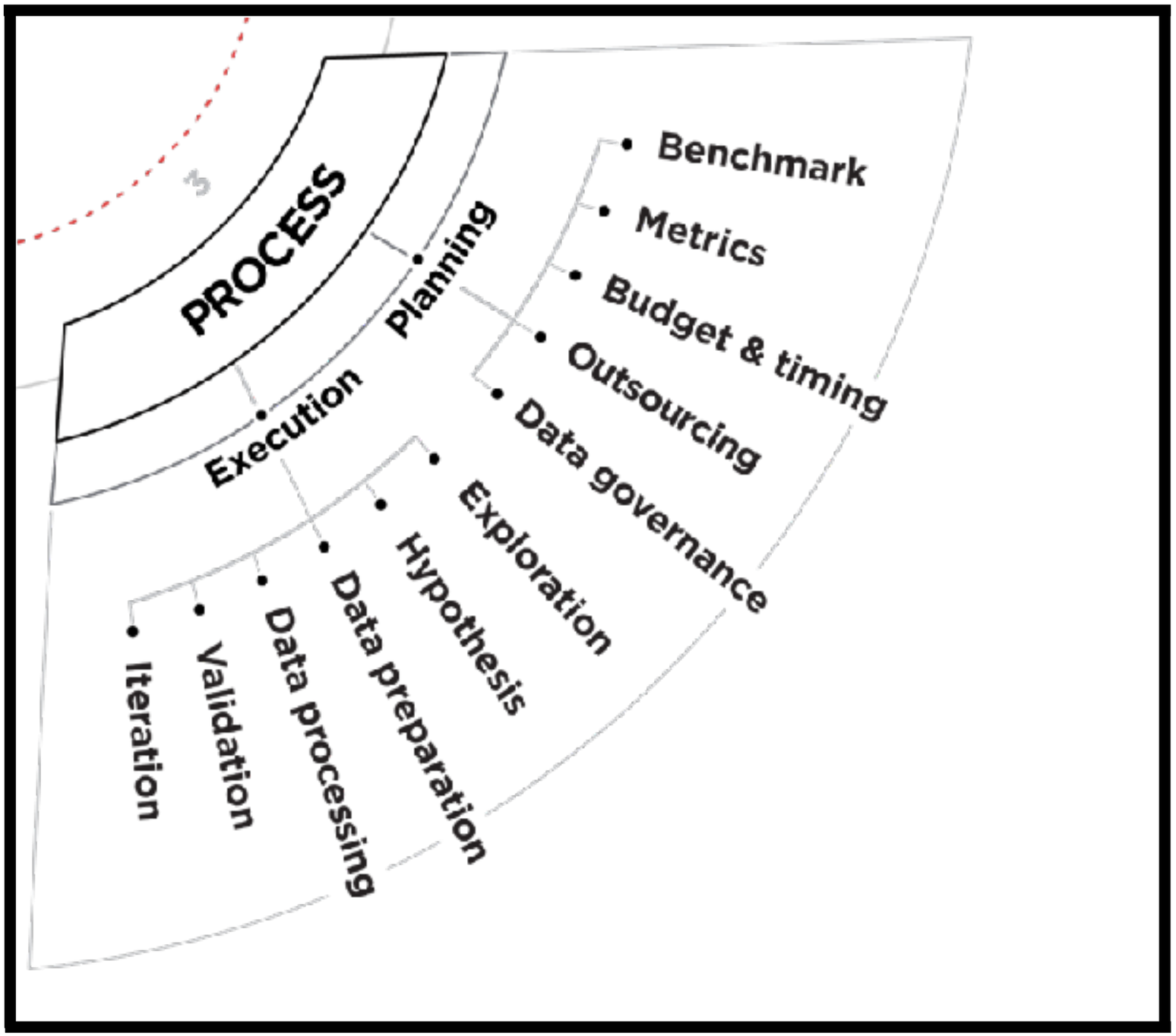

Who is responsible for managing the data? How?

How do you ensure scientific validation?

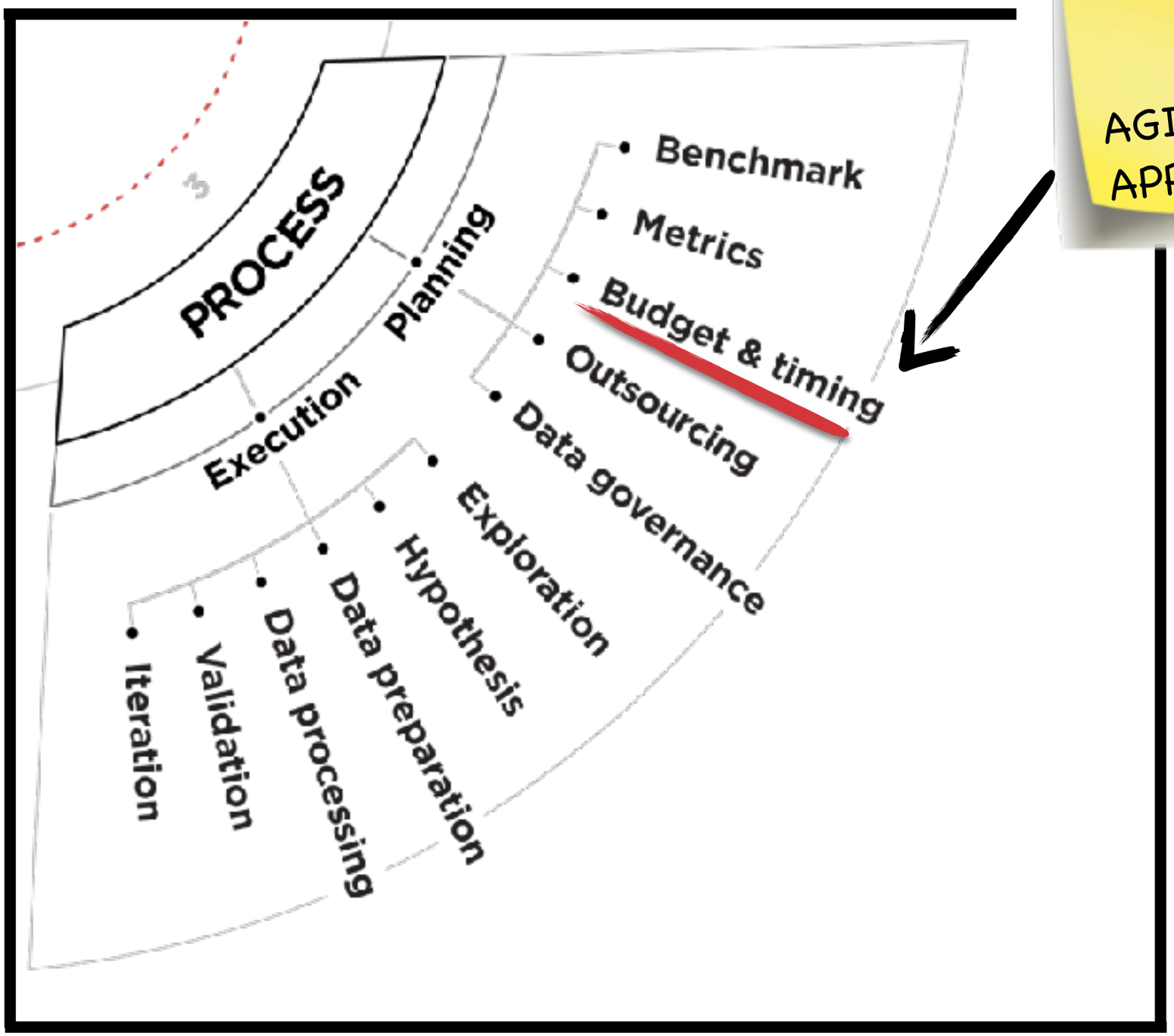
Which are your recruiting channels?

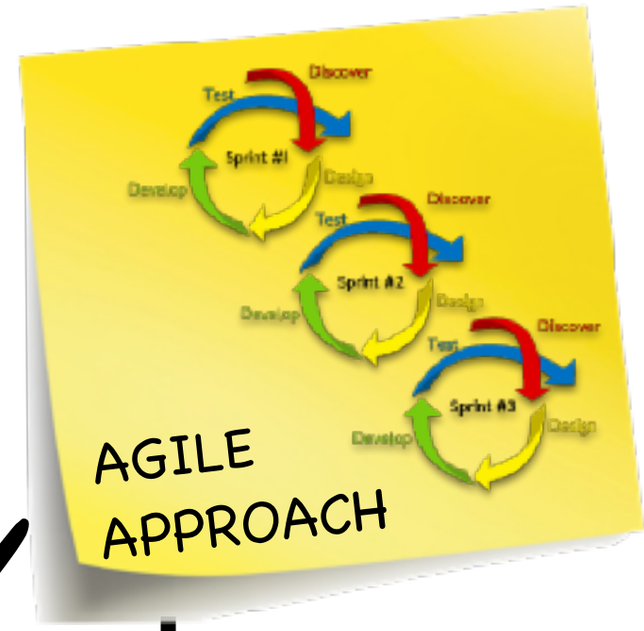
Is there a collaboration between the data team and other business units?

# PROCESS

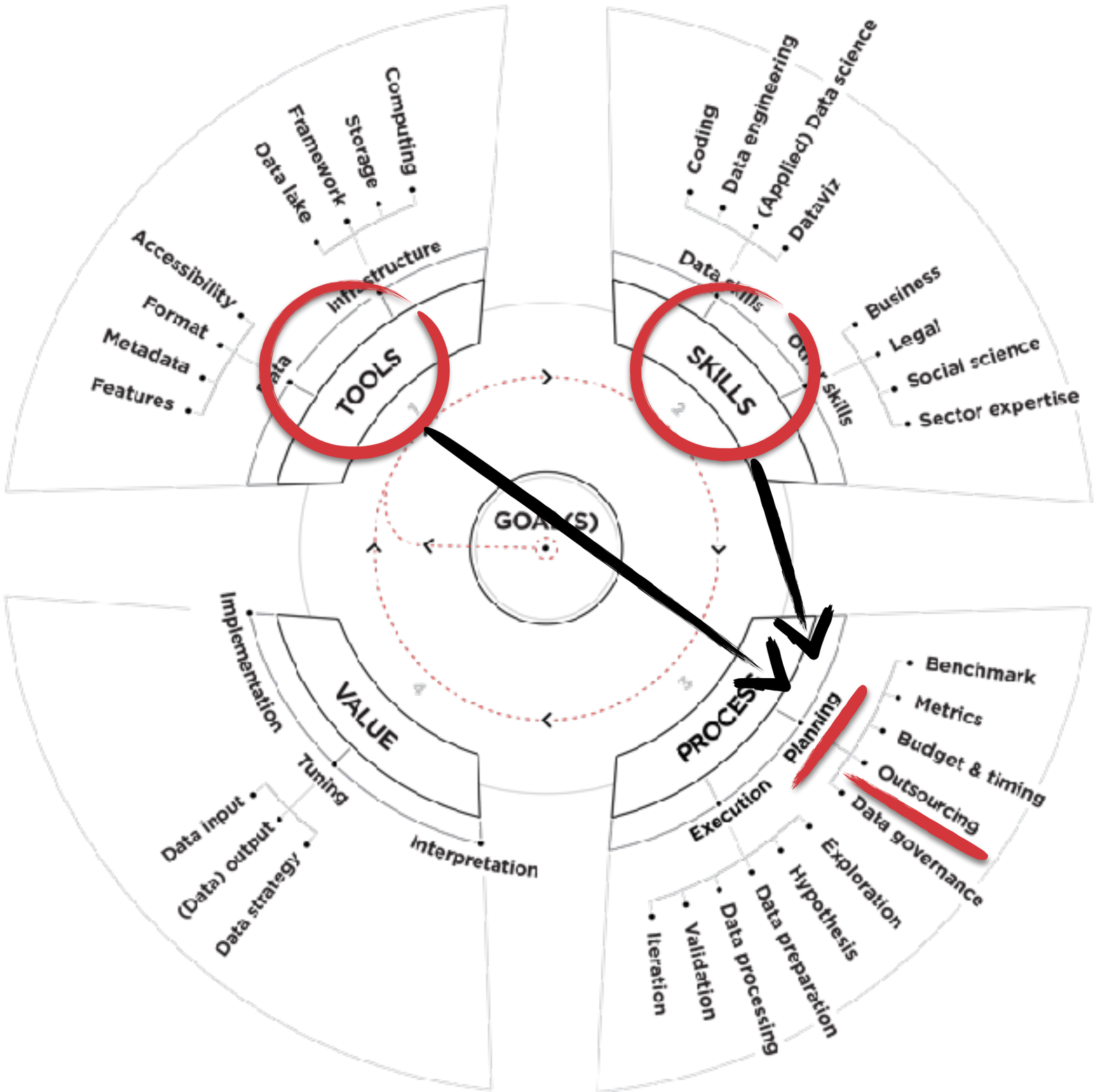

PROCESS
Planning
Execution
Benchmark
Metrics
Budget & timing
Outsourcing
Data governance
Exploration
Hypothesis
Data preparation
Data processing
Validation
Iteration

# PROCESS


PROCESS
Planning
Execution
Benchmark
Metrics
Budget & timing
Outsourcing
Data governance
Exploration
Hypothesis
Data preparation
Data processing
Validation
Iteration

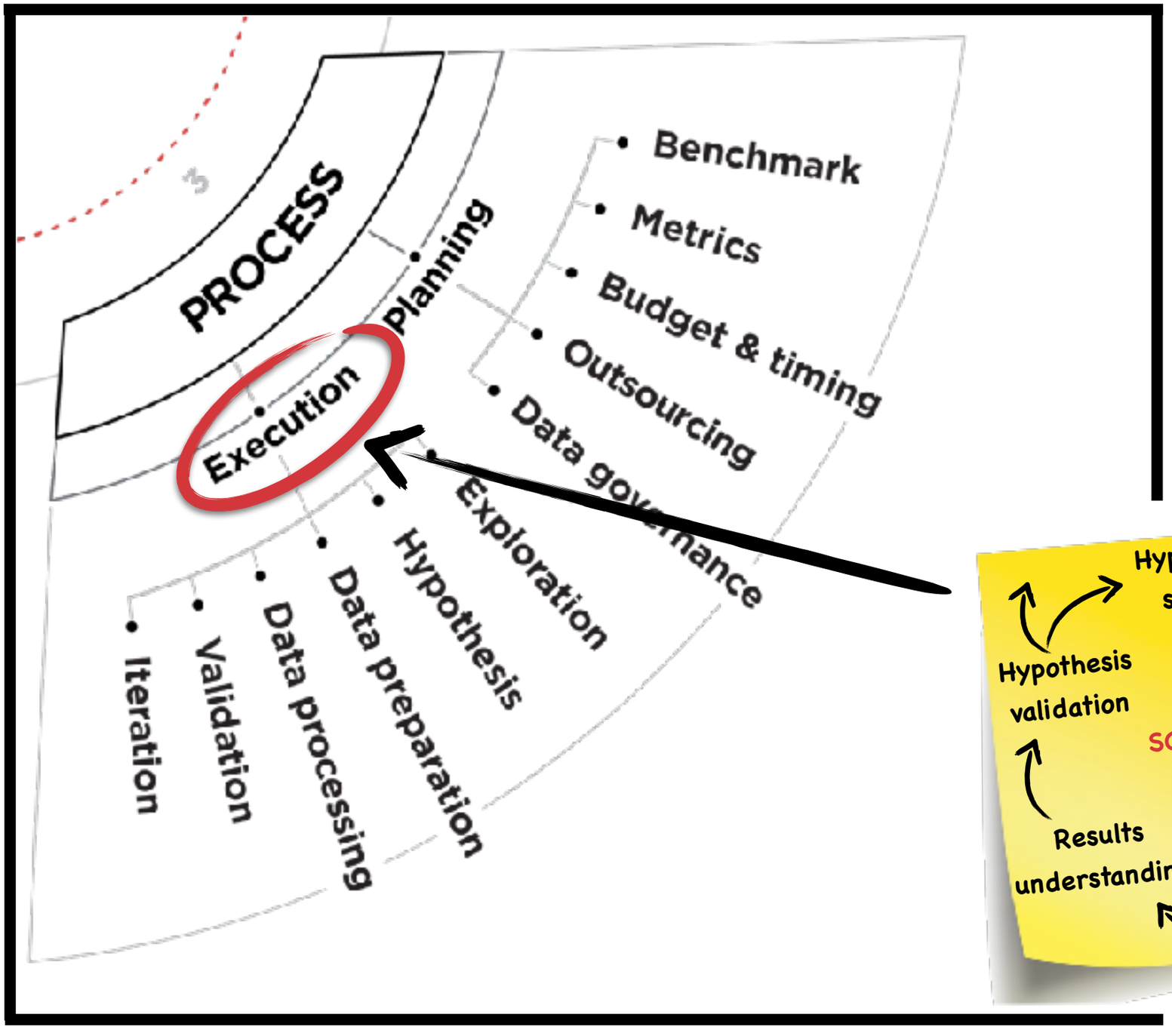

Discover
Test
Sprint #1
Develop
Design
Discover
Test
Sprint #2
Develop
Design
Discover
Test
Sprint #3
Develop
Design
AGILE
APPROACH

# PROCESS

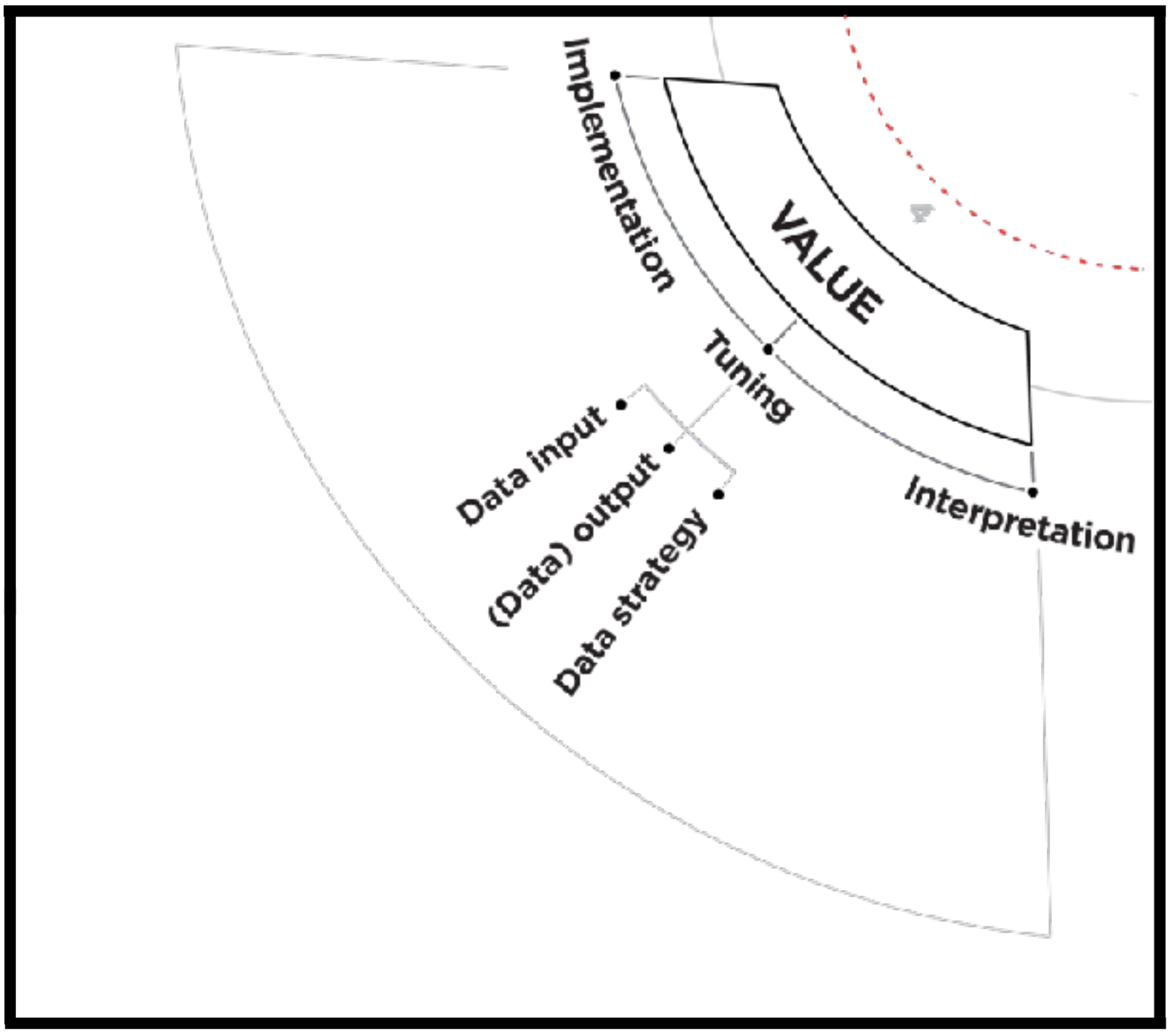



BIG DIVE

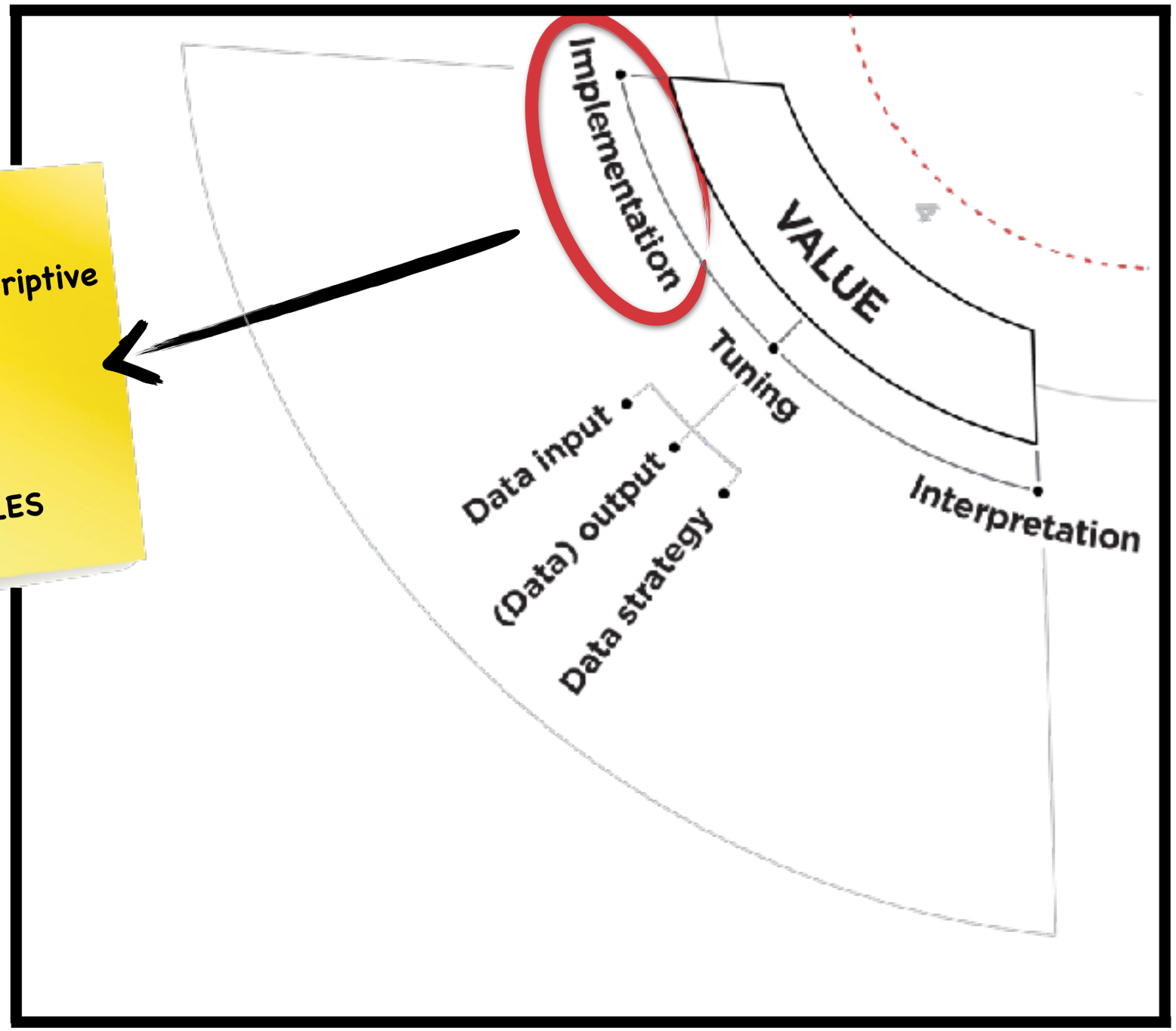
# PROCESS


PROCESS
Planning
Execution
Benchmark
Metrics
Budget & timing
Outsourcing
Data gover
Exploration
Hypothesis
Data preparation
Data processing
Validation
Iteration
Hypothesis setting
Cleaning, exploring Data
Running data science tools
Results understanding
Hypothesis validation
SCIENTIFIC METHOD

# VALUE


Implementation
VALUE
Tuning
Data input
(Data) output
Data strategy
Interpretation

# VALUE




E.g.

✓ Predictive/prescriptive ANALYTICS

✓ AUTOMATION

✓ BUSINESS RULES

# The Data Ring Canvas

| Project name: | Designed by: | Date: | Version: |
|---|---|---|---|
| | | | |

# DOWNLOAD, USE, COMMENT, REFINE IT (CC LICENSE)

- www.dataring.eu-

# The Data Ring Canvas

| Project name: | Designed by: | Date: | Version: |
|---|---|---|---|
| | | | |

# Q & A

christian.racca@top-ix.org

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

@top_ix
@bigdive_eu

# THANKS!