CERN openlab

Big Data in Healthcare

Known Unknowns and Lessons Learnt in Science

Futuro Stage - Taghi Aliyev

29/09/2017

Overview

- ‘Big Data’
  - Definitions, understanding and initial feeling
- Common pitfalls in face of multi-dimensional, complex data sets
- Importance of Statistics
  - Ambiguity and Trickery
- Platforms and Personal research

Big Data

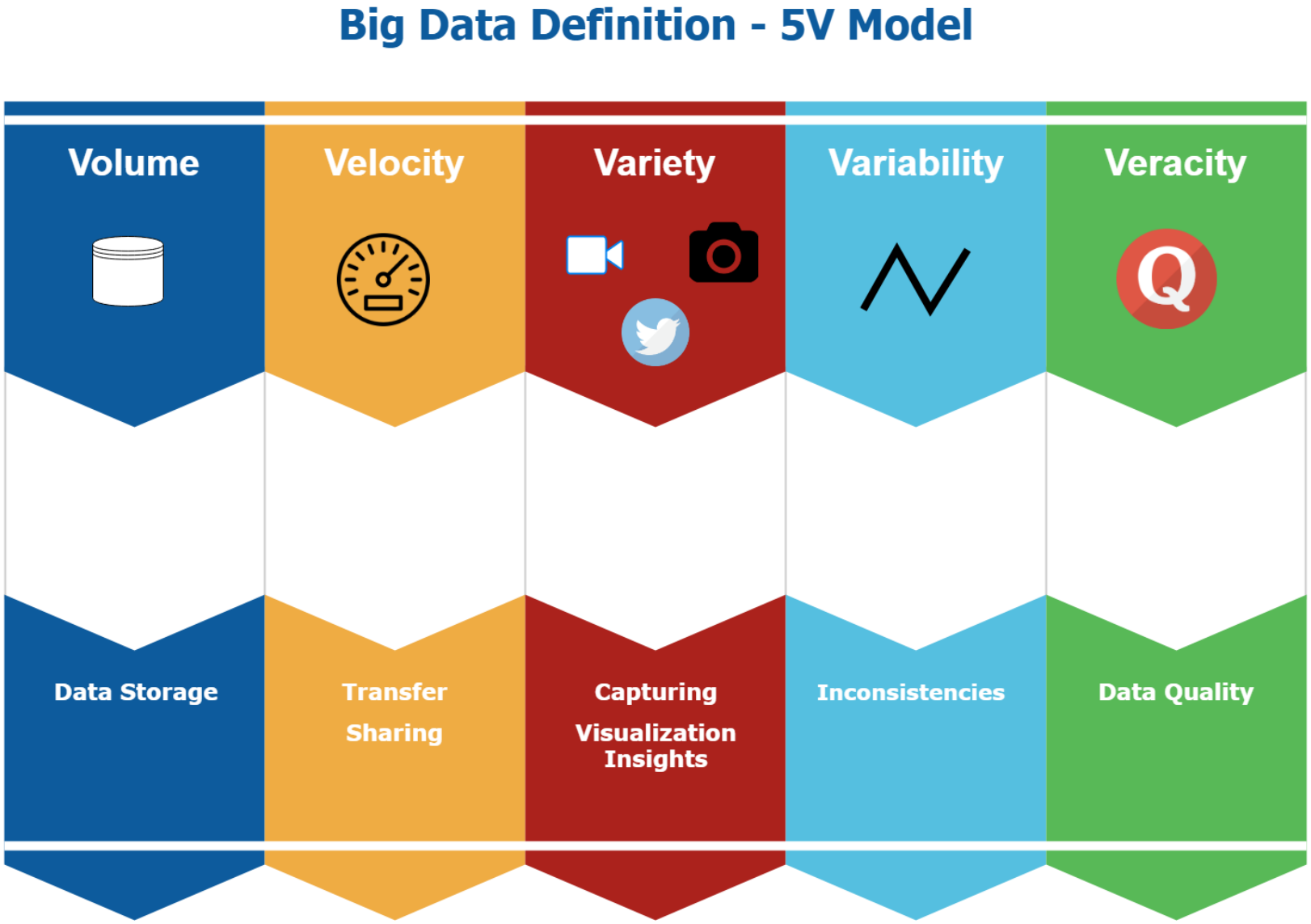
Definitions, definitions, definitions

- What is not ‘Big Data’?
  - ‘File is too big to send over email!’
  - ‘My OS can not handle such large files!’
  - ‘I can not invert this mid-sized matrix on my personal laptop!’

- What is ‘Big Data’ then?
  - Commonly used definition based on 5-V model
  - Spans over problems, challenges of different branches of Computer Science and Information Technologies

Big Data

5-V Definition



Big Data

How is it used and how does it look in Healthcare

- Personalized healthcare
  - Not the most successful ones
  - Best successes so far in Research and Public Health
- Advances in technology → a lot of data coming in
- Most not well structured, not suitable with current techniques
- Limitation of old approaches and intuition

Common Pitfalls

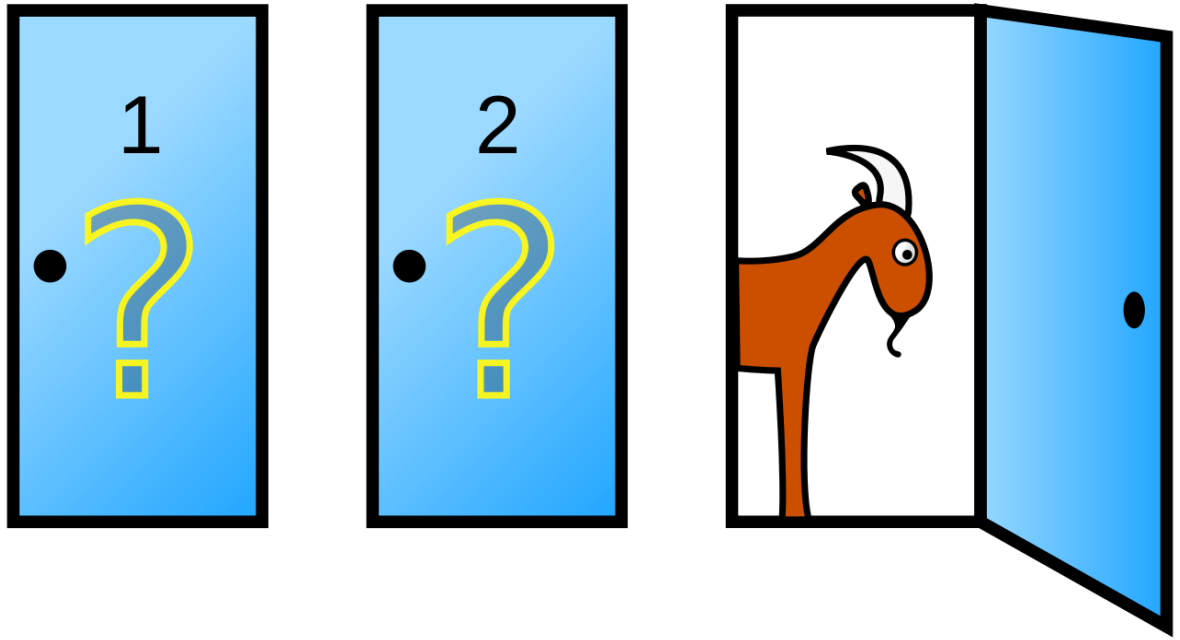
Do not always just trust the intuition

- Confirmation Bias
- Simplifying cases to fit to past experiences
- Decision-making problems
- Especially, in the chance based games
  - Poker, betting
  - Game theoretical decision-strategy-based games

An example of the pitfall

Three Doors/Monty Hall Problem

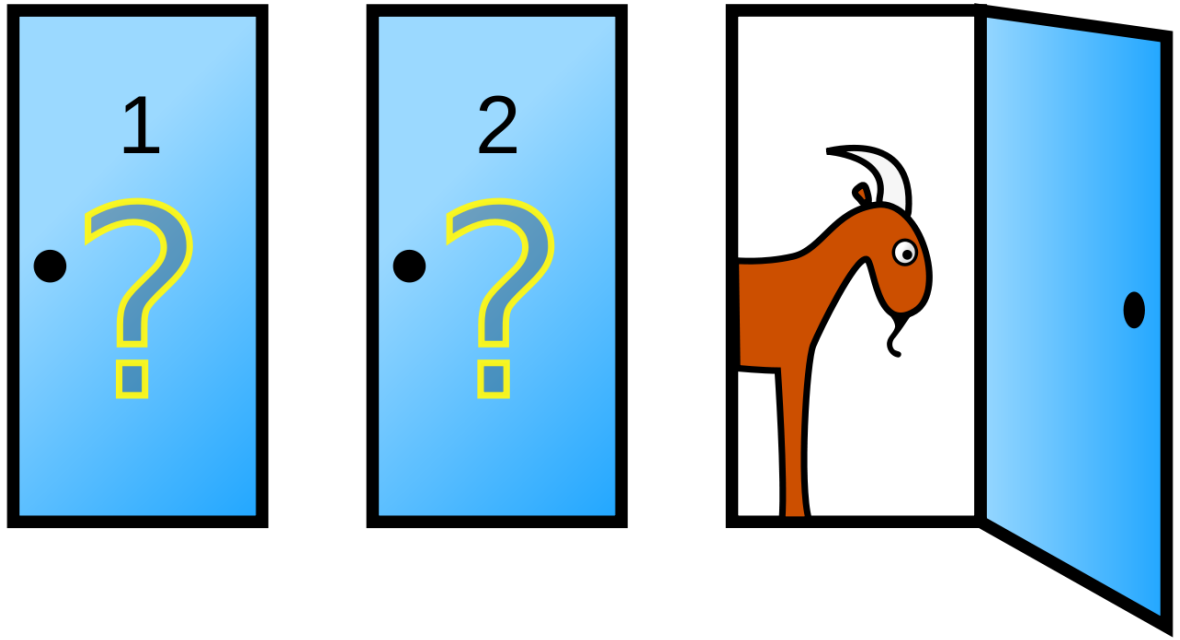
- One of the more famous examples of human intuition's pitfall in Probability Theory
- Player chooses a door
- Host opens a door where prize is not
- Should player switch the doors?
- Most of initial 10,000 believed 'no'
  - Including 1,000 PhDs



An example of the pitfall

Three Doors/Monty Hall Problem

- One of the more famous examples of human intuition's pitfall in Probability Theory
- Player chooses a door
- Host opens a door where prize is not
- Should player switch the doors?
- Most of initial 10,000 believed 'no'
  - Including 1,000 PhDs
- Answer is 'Yes'
  - You win with 2/3 chance



An example of the pitfall

- Prisoner’s Dilemma
- Classic example from Game Theory
- Assume, you and a friend are captured by the police. You are in different rooms with no communication possible between two. Police gives you following offer:
  - If you and a friend both betray each other, each serves 2 years in prison
  - If one betrays, betraying person is free and the other one is imprisoned for 3 years
  - If both are silent, both get 1 year of prison
  - What do you do?

An example of the pitfall

- Prisoner’s Dilemma
- Classic example from Game Theory
- “Rational” player not optimum
- Actual optimum strategy in iterated version is to co-operate
  - Long-term reward/punishment is minimal on average

Prisoner's dilemma payoff matrix

A \ B	B stays silent	B betrays
A stays silent	-1 / -1	0 / -3
A betrays	-3 / 0	-2 / -2

Importance of Statistics

- Human intuition has pitfalls
  - Especially with larger, more complex data sets
- In today's world, huge emphasis on model selection/approximation
- Models usually assume a distribution
  - To fit the given data points
- Even if distribution fits well, does it make sense?

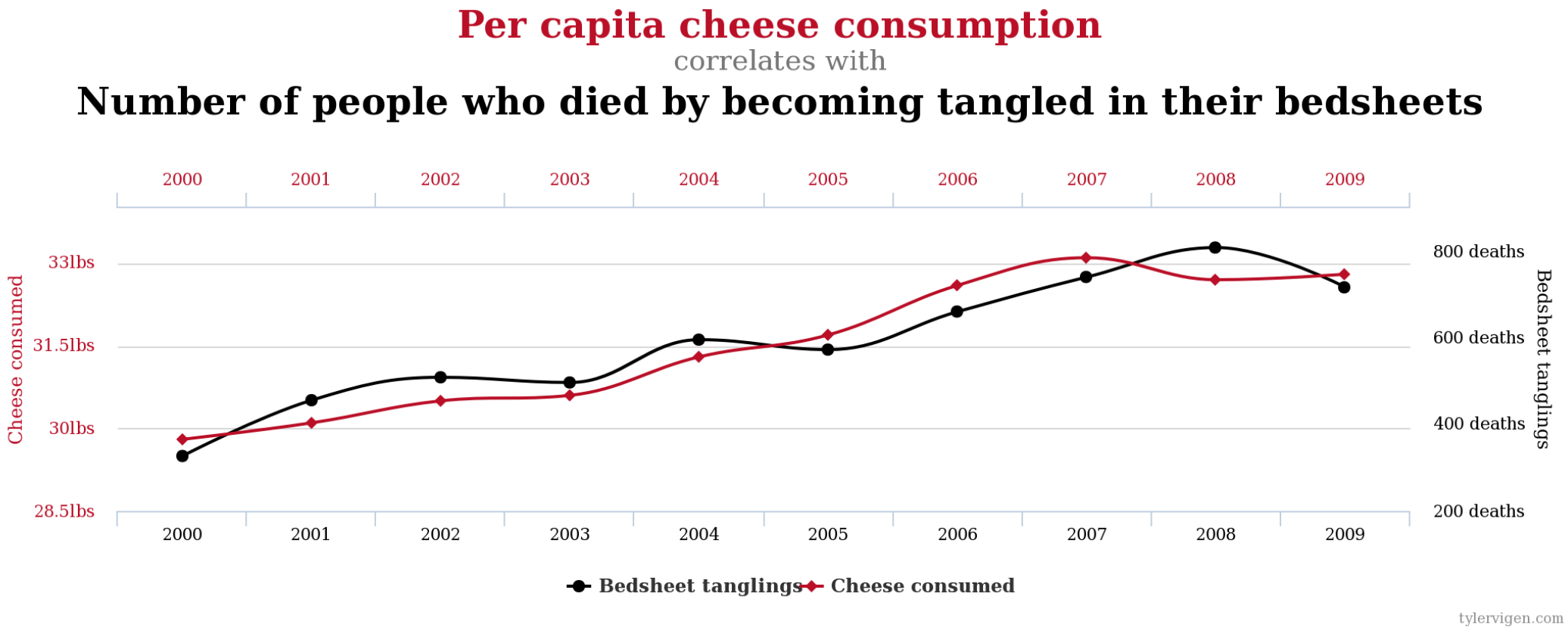
Importance of Statistics

- In Mathematics and Physics, models need to be explainable
- Common phrase in Machine Learning:
  - “We should just trust the machine if it works, as long as they are doing better than humans”
  - Many publications of such kind
- We should focus on generating explainable, understandable and good fitting models!
  - Not the black boxes
- 2-way street
  - Machines helping to overcome pitfalls
  - Human verifying results

Examples

Weird Correlations. Are these images good examples?

Idea is to tell that just good distributions/models are not enough, humans need to verify the actual use and meaning of it.



Example - 2

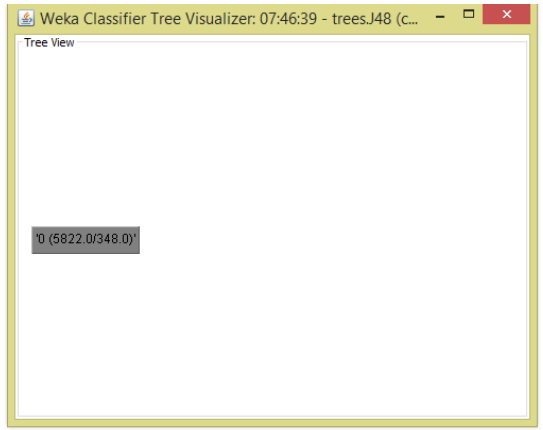
Famous Caravan Insurance Problem

- Initially a Data Mining Challenge in 2000
- Given a data set about customers, find who will take an insurance?
- Huge imbalance in data set
  - Very few actual cases
- Model outputting ‘No’ at all times has ~95% accuracy
- High accuracy, however completely wrong and irrelevant model
  - Which is figured out after inspection of the model and other metrics

Caravan example

- Further investigation of the model finds out the issues
- Similar behavior in healthcare at times
  - Mood prediction algorithms based on signals from phone
  - Rare disease prediction

```
Size of the tree :      1


Time taken to build model: 0.31 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        5471               93.9711 %
Incorrectly Classified Instances       351                6.0289 %
Kappa statistic                         -0.001
Mean absolute error                      0.1127
Root mean squared error                  0.2382
Relative absolute error                100.1452 %
Root relative squared error            100.4729 %
Total Number of Instances             5822

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.999     1          0.94      0.999     0.969      0.497    0
                 0         0.001      0         0         0          0.497    1
Weighted Avg.    0.94      0.94       0.884     0.94      0.911      0.497

=== Confusion Matrix ===

    a    b   <-- classified as
 5471    3 |    a = 0
  348    0 |    b = 1
```



Many analysts different results

Team	Analytic Approach	OR
12	Zero–inflated Poisson regression	0.89
17	Bayesian logistic regression	0.96
15	Hierarchical log–linear modeling	1.02
10	Multilevel regression and logistic regression	1.03
18	Hierarchical Bayes model	1.10
31	Logistic regression	1.12
1	Ordinary least squares with robust standard errors, logistic regression	1.18
4	Spearman correlation	1.21
14	Weighted least squares regression with referee fixed–effects and clustered standard errors	1.21
11	Multiple linear regression	1.25
30	Clustered robust binomial logistic regression	1.28
6	Linear Probability Model	1.28
26	Three–level hierarchical generalized linear modeling with Poisson sampling	1.30
3	Multilevel Binomial Logistic Regression using bayesian inference	1.31
23	Mixed model logistic regression	1.31
16	Hierarchical Poisson Regression	1.32
2	Linear probability model, logistic regression	1.34
5	Generalized linear mixed models	1.38
24	Multilevel logistic regression	1.38
28	Mixed effects logistic regression	1.38
32	Generalized linear models for binary data	1.39
8	Negative binomial regression with a log link analysis	1.39
20	Cross–classified multilevel negative binomial model	1.40
13	Poisson Multi–level modeling	1.41
25	Multilevel logistic binomial regression	1.42
9	Generalized linear mixed effects models with a logit link function	1.48
7	Dirichlet process Bayesian clustering	1.71
21	Tobit regression	2.88
27	Poisson regression	2.93

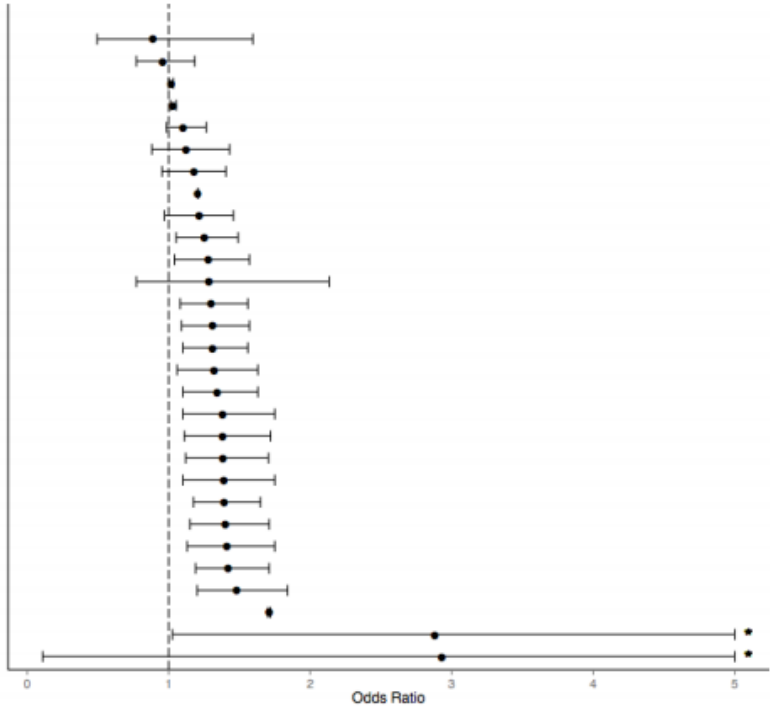
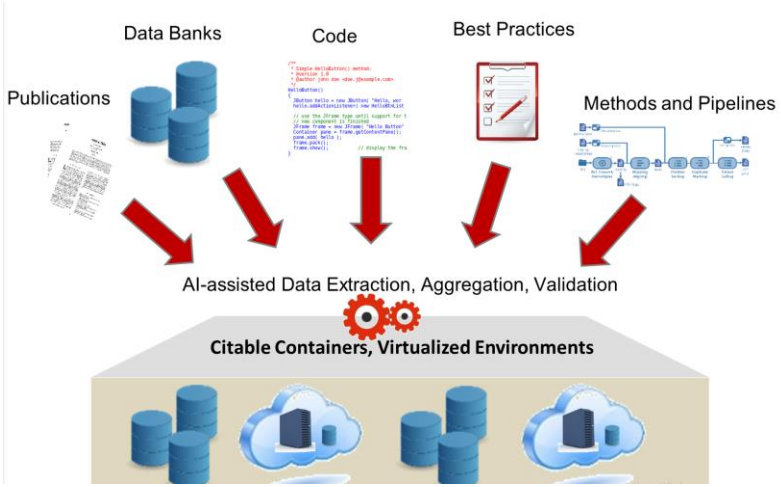


Figure 1. Point estimates and 95% confidence intervals for analysis teams for the primary research question: Are soccer referees more likely to give red cards to dark skin toned players than light skin toned players? Note that the asterisks correspond to a truncated upper bound for Team 21 (11.47) and Team 27 (78.66) to increase the interpretability of this plot.

Platform

The Importance and Goal

- Large-scale collaborative research platform
- A powerful ecosystem for researchers
  - To negotiate and challenge the shared values and the value chain of a given field
- Providing tools for humans to collaborate and work together with the machines
- Focused on ‘why’, rather than ‘how’
- Work and Collaborate with us!



Platforms

CoreTeam/Nucleus!

Alberto Di Meglio

Marco Manca

Erik Biessen

Mario Falchi

Taghi Aliyev

Thank You

taghi.aliyev@cern.ch

Twitter: @TaghiAliyev