Inspire…Educate…Transform.

Data Science: Big Picture

Also, introduction to CPEE program

Dr. K. V Dakshinamurthy

President, INSOFE

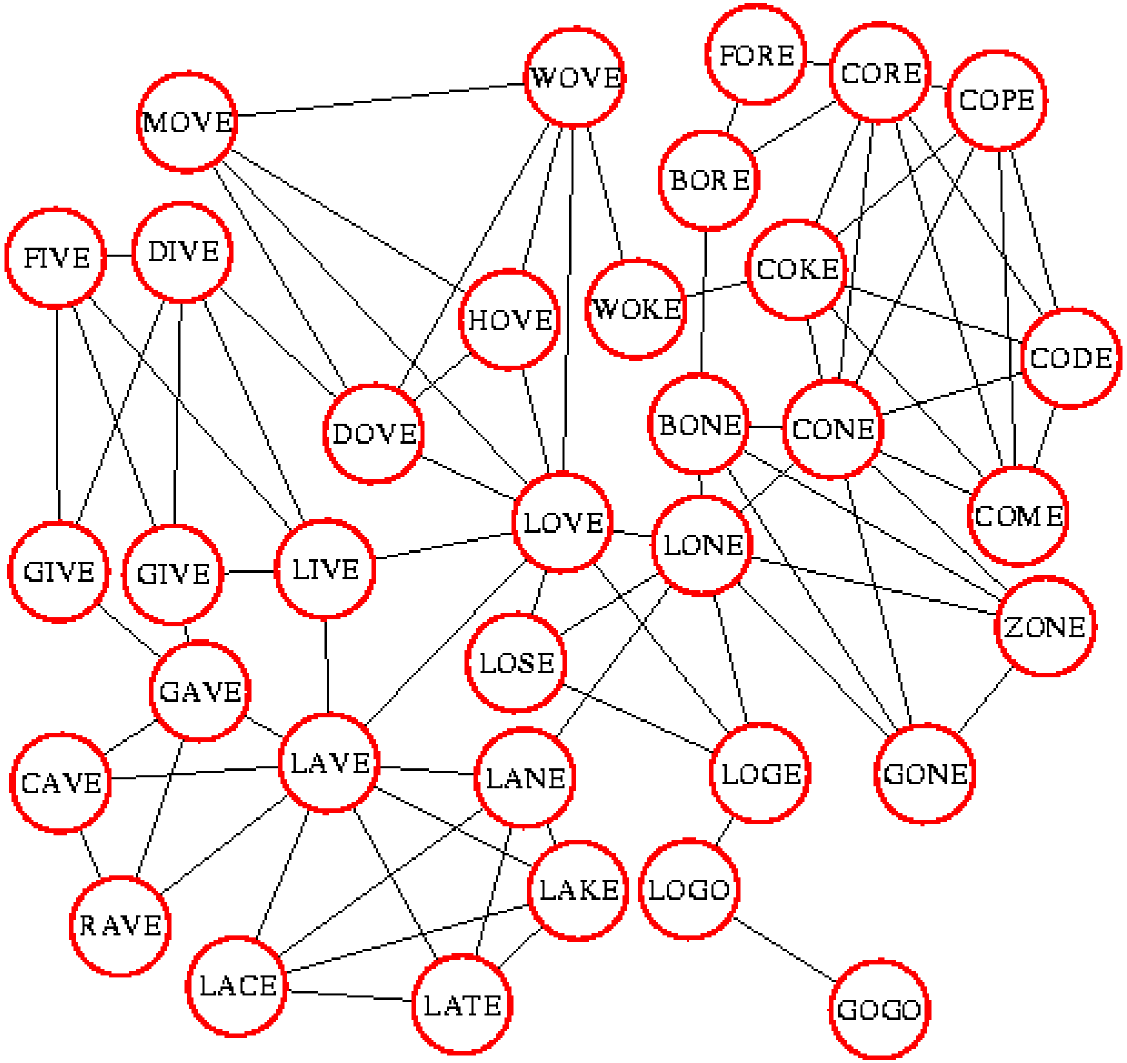
May 31, 2015

THE OUTPUT

Multiple forms

- Rules
  - –If x1, x2, x3 then Y
- Equations
  - –Y=f(x)

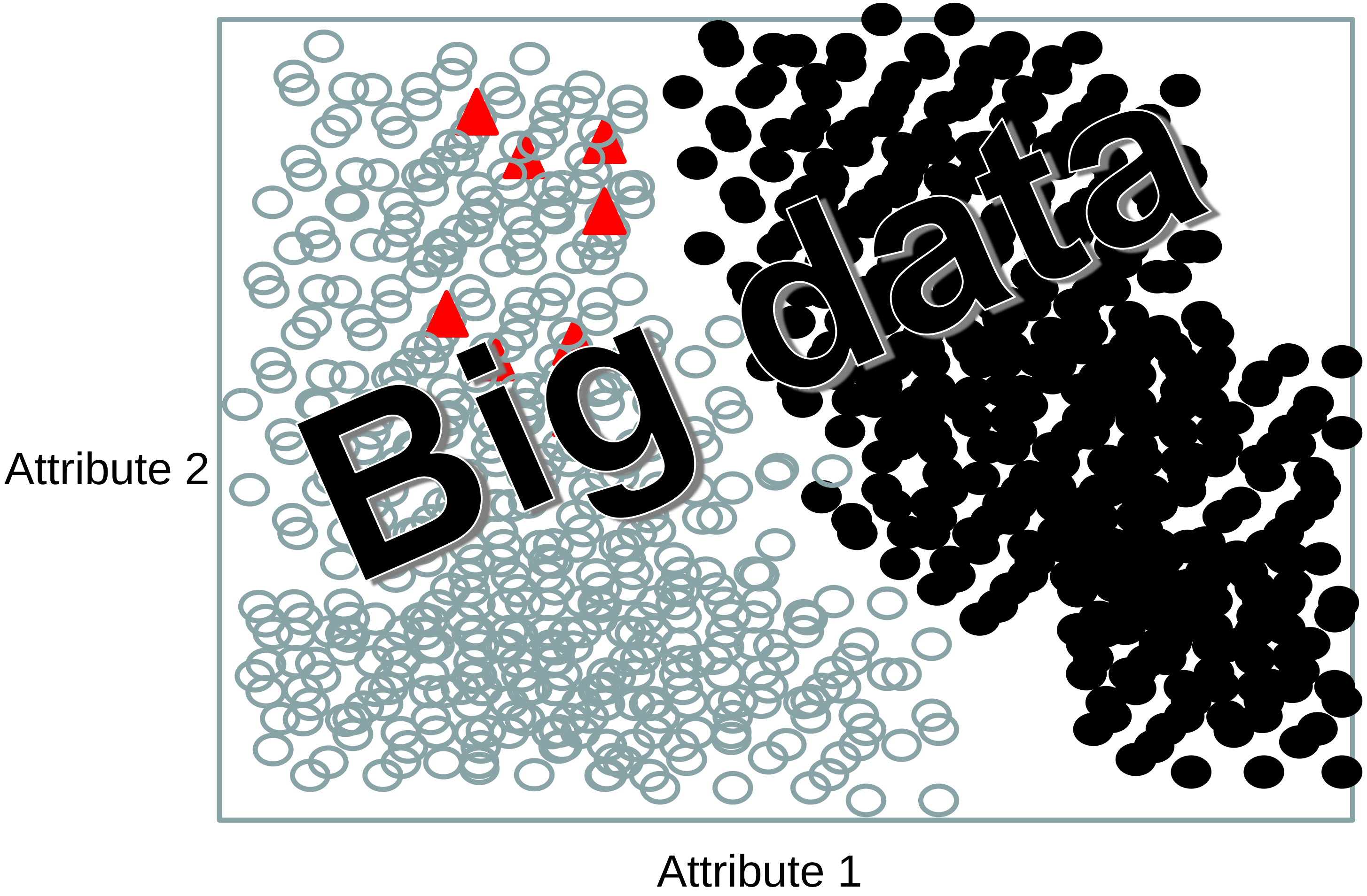
Graphs


WOVE
FORE
CORE
COPE
MOVE
BORE
FIVE
DIVE
COKE
HOVE
WOKE
CODE
DOVE
BONE
CONE
COME
LOVE
LONE
GIVE
GIVE
LIVE
ZONE
LOSE
GAVE
LAVE
CAVE
LANE
LOGE
GONE
LAKE
LOGO
RAVE
LACE
LATE
GOGO

Output

- Similarities
- Blackbox

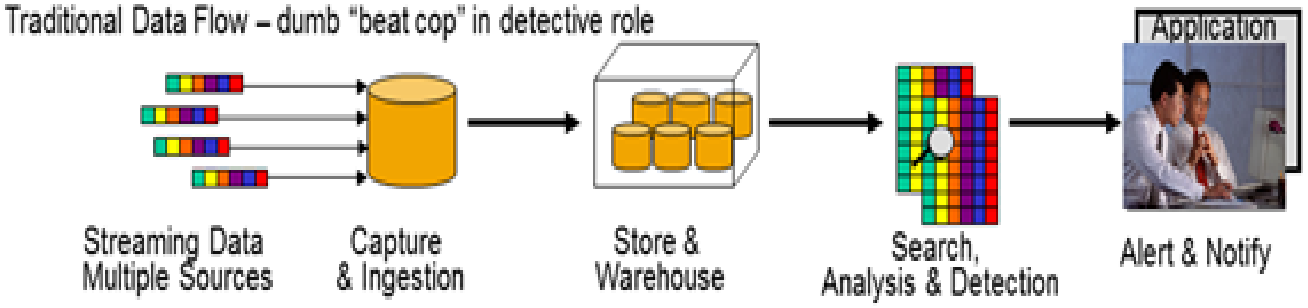
TYPES OF CHALLENGES

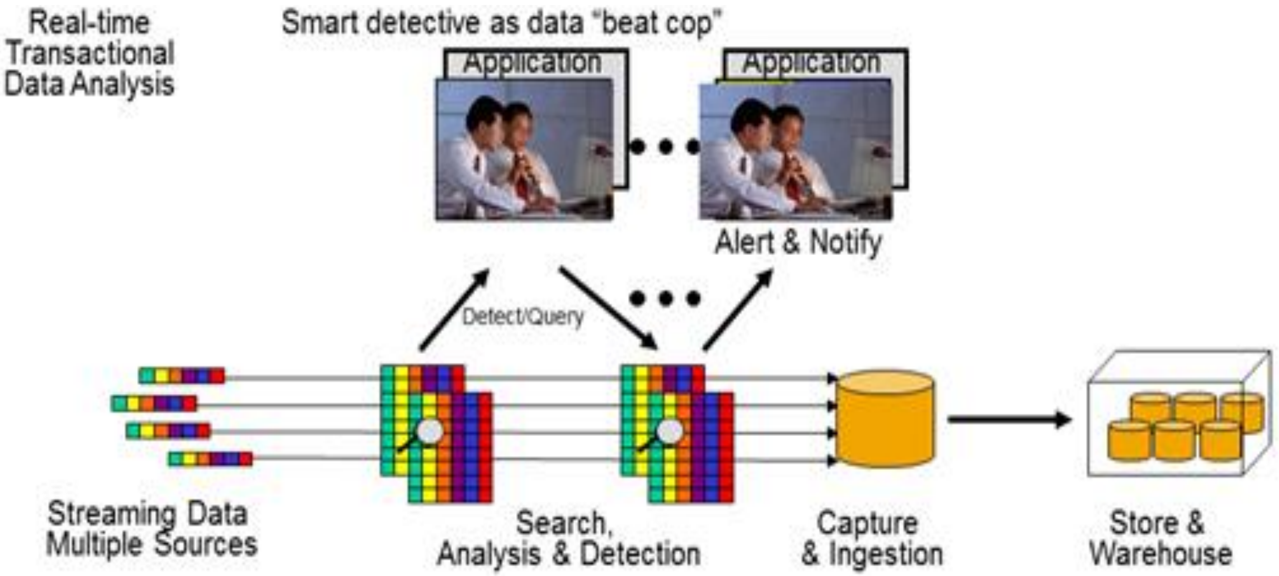

Big data
Attribute 2
Attribute 1

Velocity

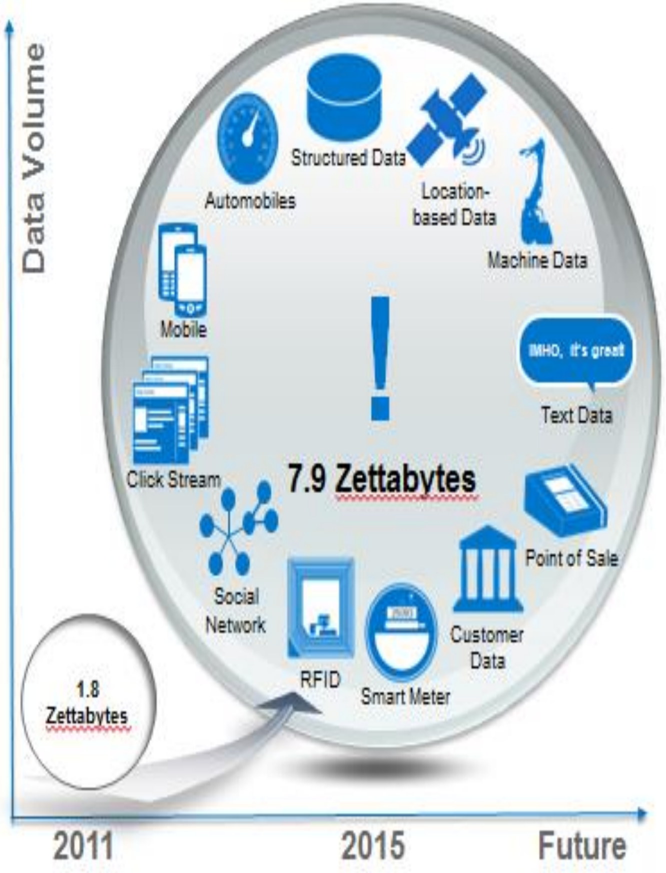
- Every minute, 100 hours of video on YouTube, 200 million emails and 300,000 tweets.

12


Traditional Data Flow – dumb "beat cop" in detective role
Streaming Data
Multiple Sources
Capture
& Ingestion
Store &
Warehouse
Search,
Analysis & Detection
Application
Alert & Notify


Real-time
Transactional
Data Analysis
Smart detective as data "beat cop"
Application
Application
Alert & Notify
Detect/Query
Streaming Data
Multiple Sources
Search,
Analysis & Detection
Capture
& Ingestion
Store &
Warehouse

Variety



- 1 Terabyte = 1024 Gigabytes
- 1 Petabyte = 1024 Terabytes
- 1 Exabyte = 1024 Petabytes
- 1 Zettabyte = 1024 Exabytes

Variability

- Understanding unstructured information is complex
  - My in-laws are as sweet as Nazis
  - Police is tracking terrorists with bombs
  - He eats, shoots and leaves
  - Great
    - "Delicious muesli from the @imaginarycafe- what a great way to start the day!
    - "Greatly disappointed that my local Imaginary Cafe have stopped stocking BLTs."
    - "Had to wait in line for 45 minutes at the Imaginary Cafe today. Great, well there's my lunchbreak gone…"

Veracity

- Lots of cleaning and noise removal is needed.

16

Value

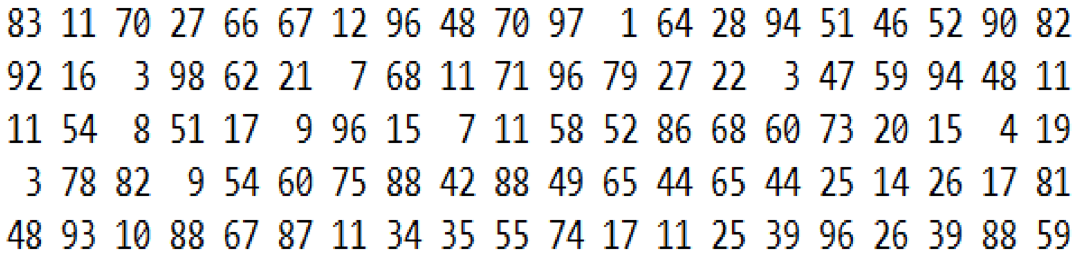
- Healthcare related big data efforts “could account for $300 billion to $450 billion in reduced health-care spending
- Data on its own is worthless. The value lies in rigorous analysis of accurate data for valuable insights

Visualization

- Do visualizations really make a big difference?

17

How many numbers are less than 10



```
83 11 70 27 66 67 12 96 48 70 97  1 64 28 94 51 46 52 90 82
92 16  3 98 62 21  7 68 11 71 96 79 27 22  3 47 59 94 48 11
11 54  8 51 17  9 96 15  7 11 58 52 86 68 60 73 20 15  4 19
 3 78 82  9 54 60 75 88 42 88 49 65 44 65 44 25 14 26 17 81
48 93 10 88 67 87 11 34 35 55 74 17 11 25 39 96 26 39 88 59
```

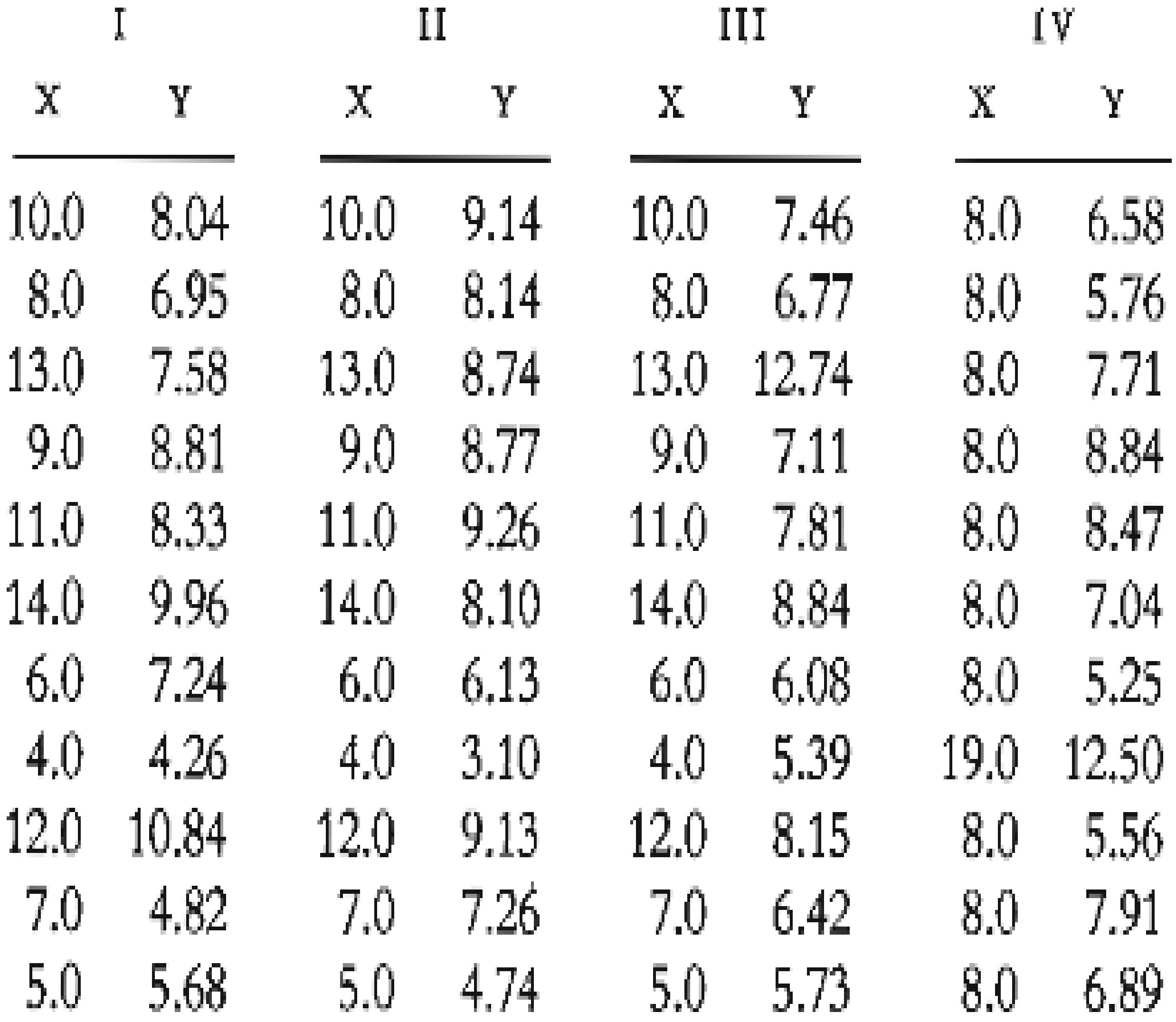
You have 8 seconds

18

How many numbers are less than 10

83 11 70 27 66 67 12 96 48 70 97 **1** 64 28 94 51 46
52 90 82 92 16 **3** 98 62 21 **7** 68 11 71 96 79 27 22 **3**
47 59 94 48 11 11 54 **8** 51 17 **9** 96 15 **7** 11 58 52 86
68 60 73 20 15 **4** 19 **3** 78 82 **9** 54 60 75 88 42 88 49
65 44 65 44 25 14 26 17 81 48 93 10 88 67 87 11 34
35 55 74 17 11 25 39 96 26 39 88 59

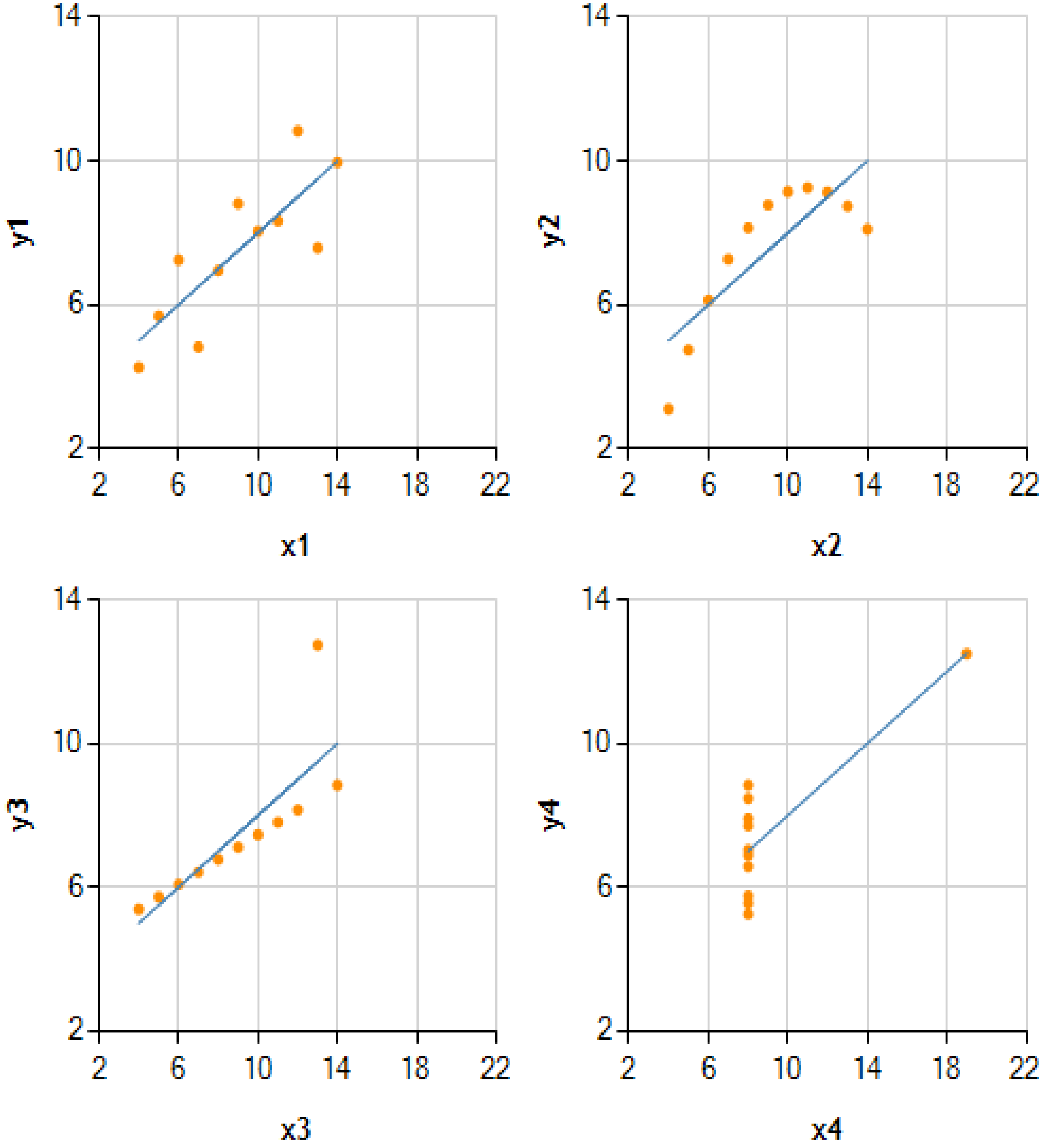
You have 4 seconds



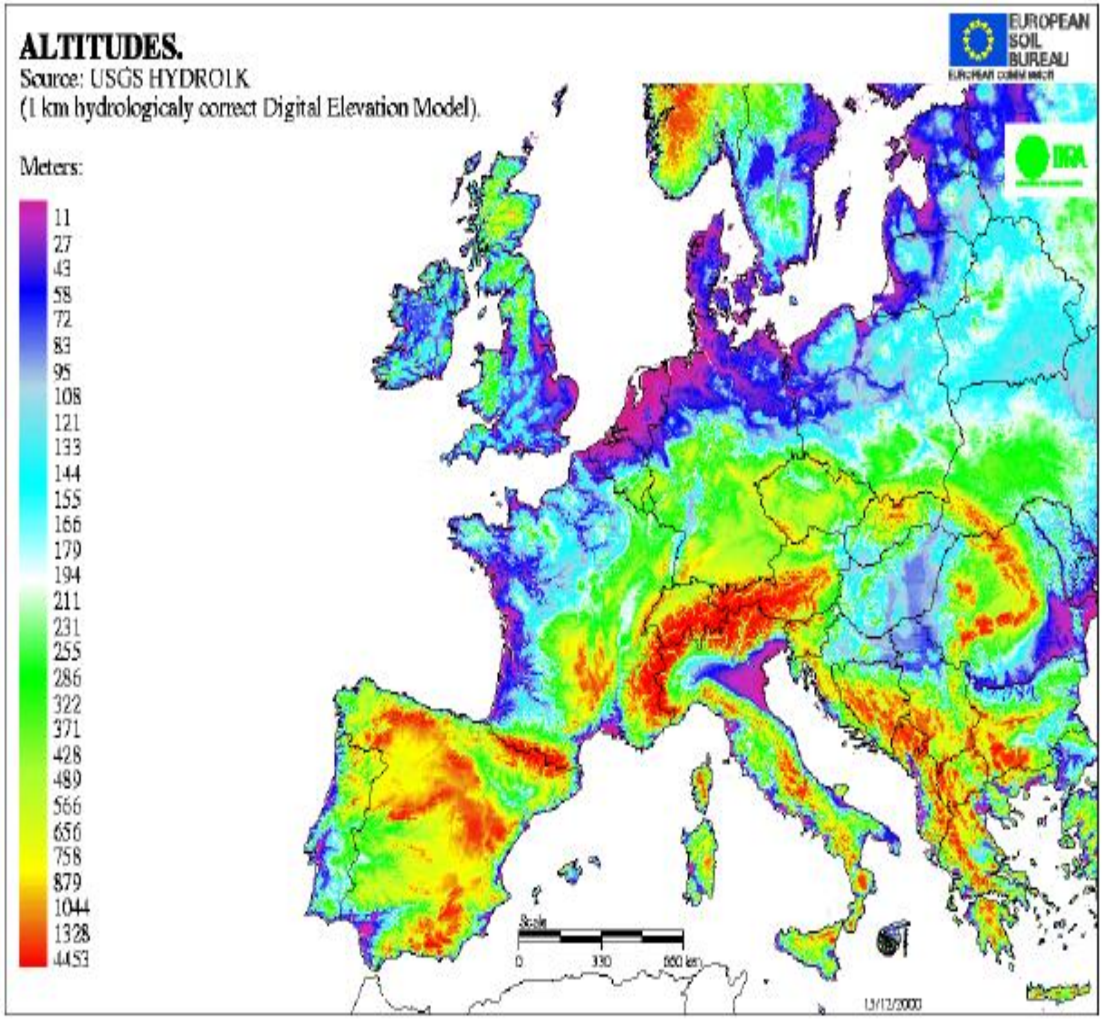
I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

- Mean of x: [9 9 9 9]
- Variance of x: [11 11 11 11]
- Mean of y: [7.5 7.5 7.5 7.5]
- Variance of y: [4.127 4.128 4.123 4.123]
- Correlation of x-y: 0.816 0.816 0.816 0.817

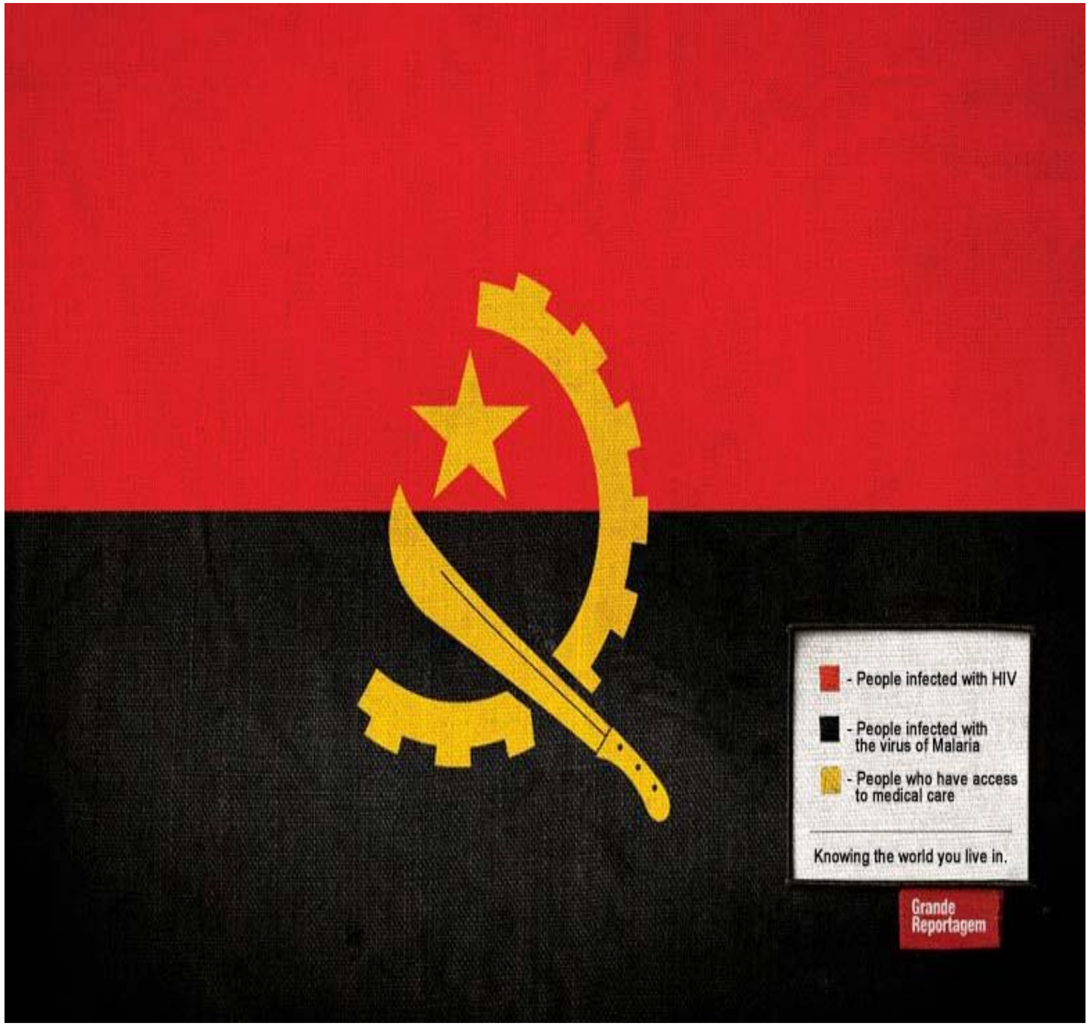
- Equation of regression line for
  - 1: $Y = 3 + 0.5X$; r2: 0.67
  - 2: $Y = 3 + 0.5X$; r2: 0.67
  - 3: $Y = 3 + 0.5X$; r2: 0.67
  - 4: $Y = 3 + 0.5X$; r2: 0.67


14
10
6
2
y1
2
6
10
14
18
22
x1
14
10
6
2
y2
2
6
10
14
18
22
x2
14
10
6
2
y3
2
6
10
14
18
22
x3
14
10
6
2
y4
2
6
10
14
18
22
x4

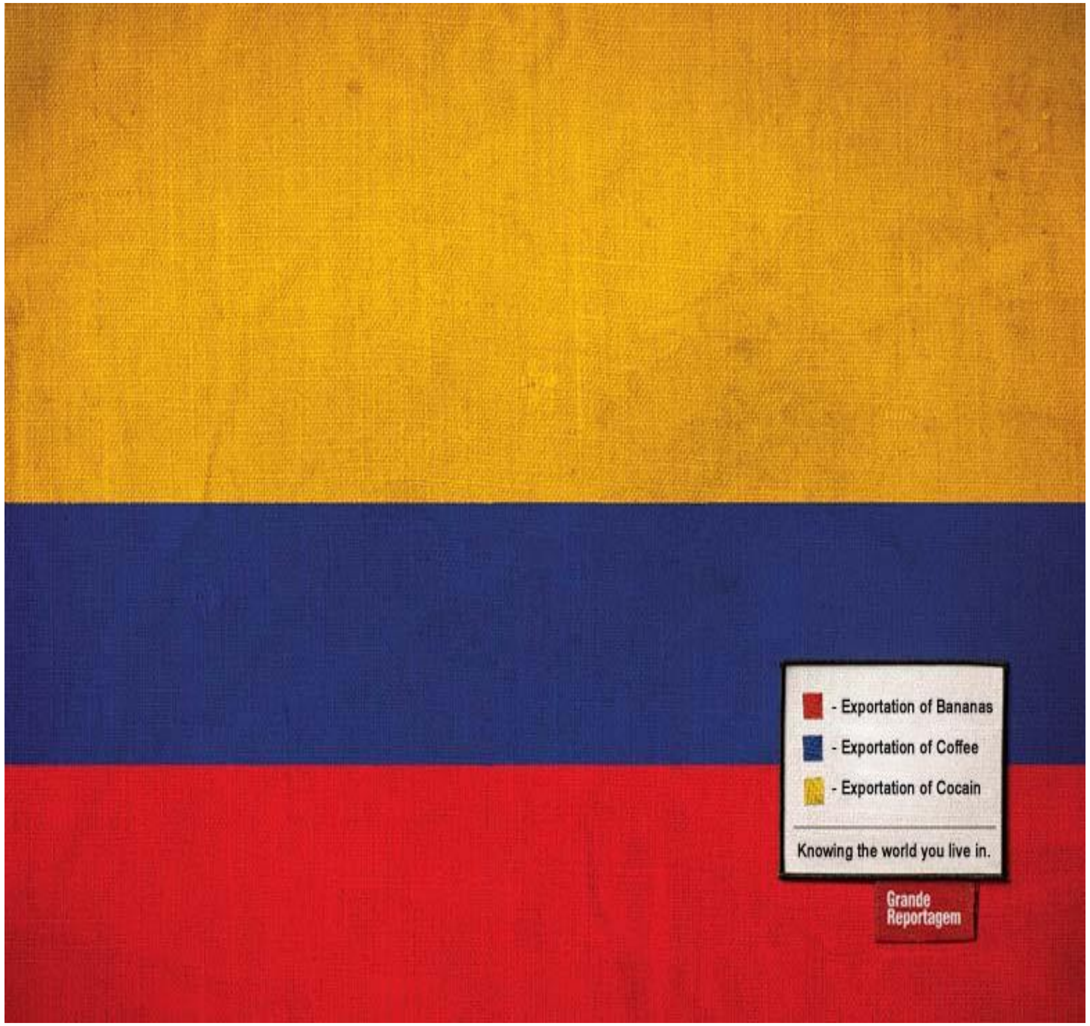
Some bad visualizations



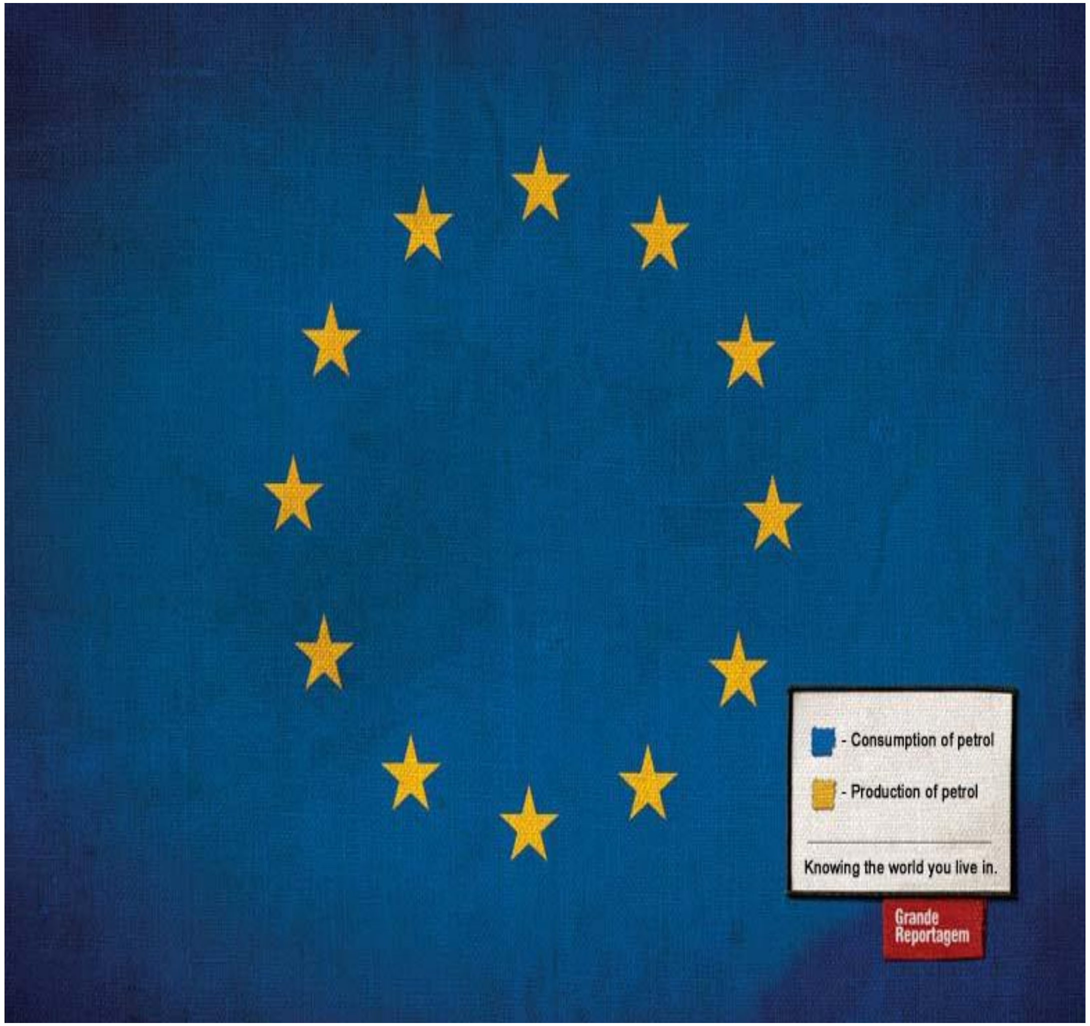
Great visualization An example


🟥 - People infected with HIV
⬛ - People infected with the virus of Malaria
🟨 - People who have access to medical care
Knowing the world you live in.
Grande Reportagem

Angola

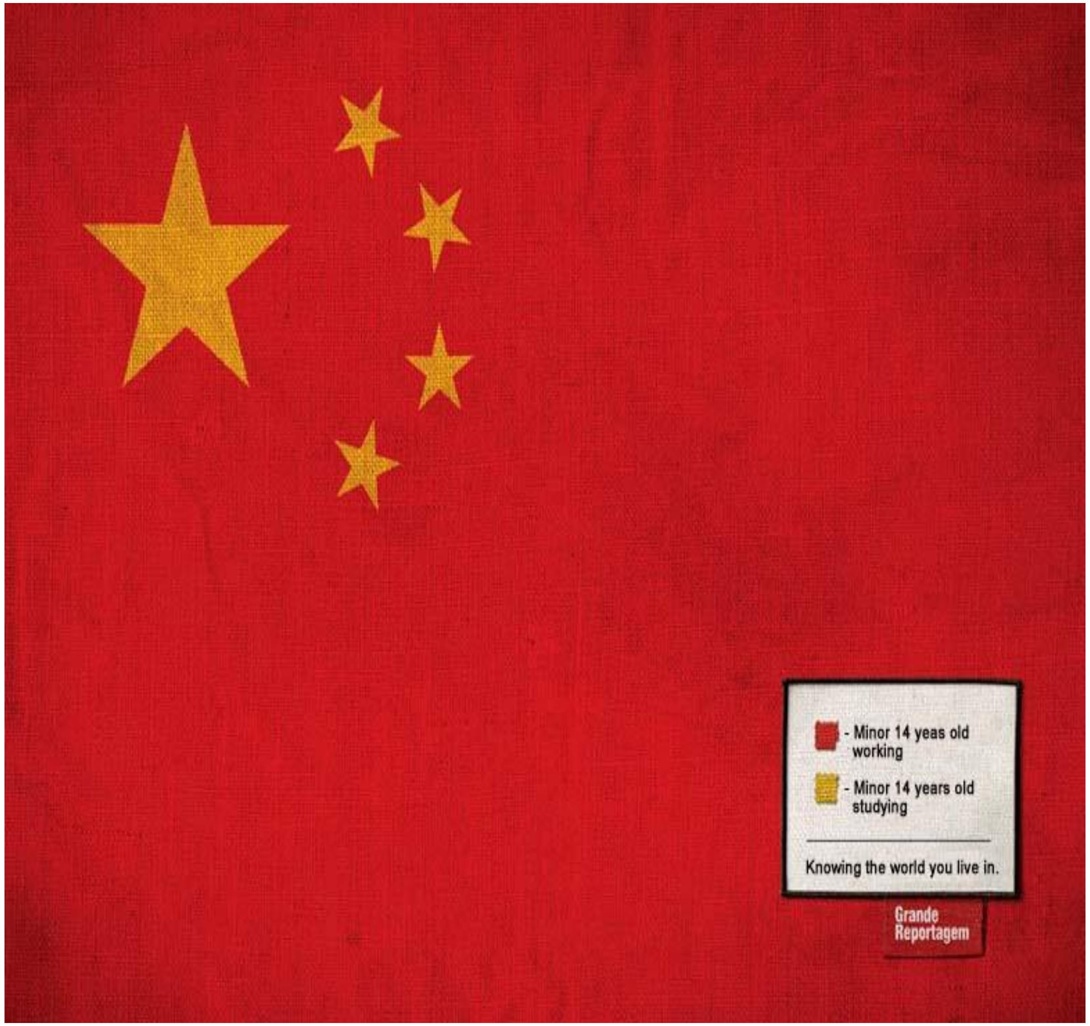

🟥 - Exportation of Bananas
🟦 - Exportation of Coffee
🟨 - Exportation of Cocain
Knowing the world you live in.
Grande Reportagem

Columbia

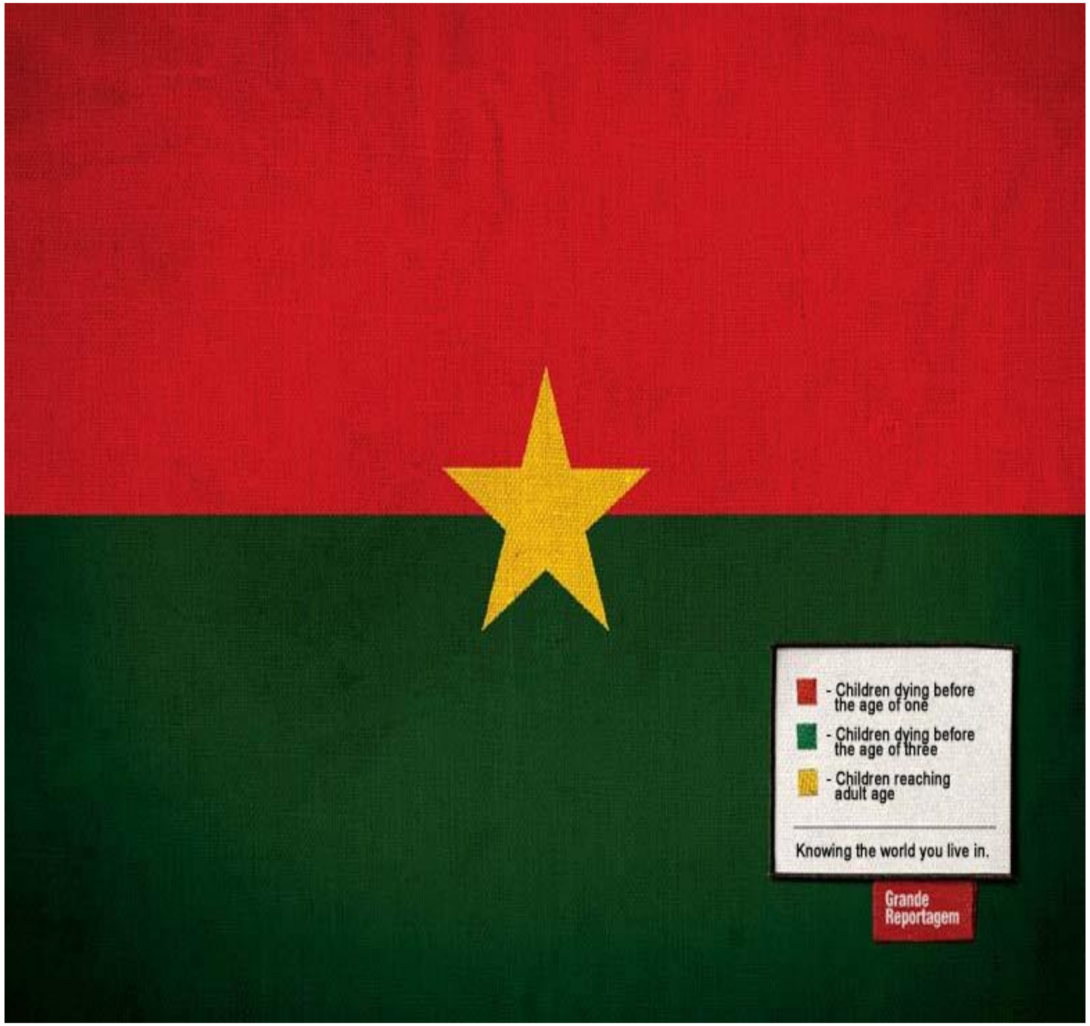


🟦 - Consumption of petrol
🟨 - Production of petrol
Knowing the world you live in.
Grande Reportagem

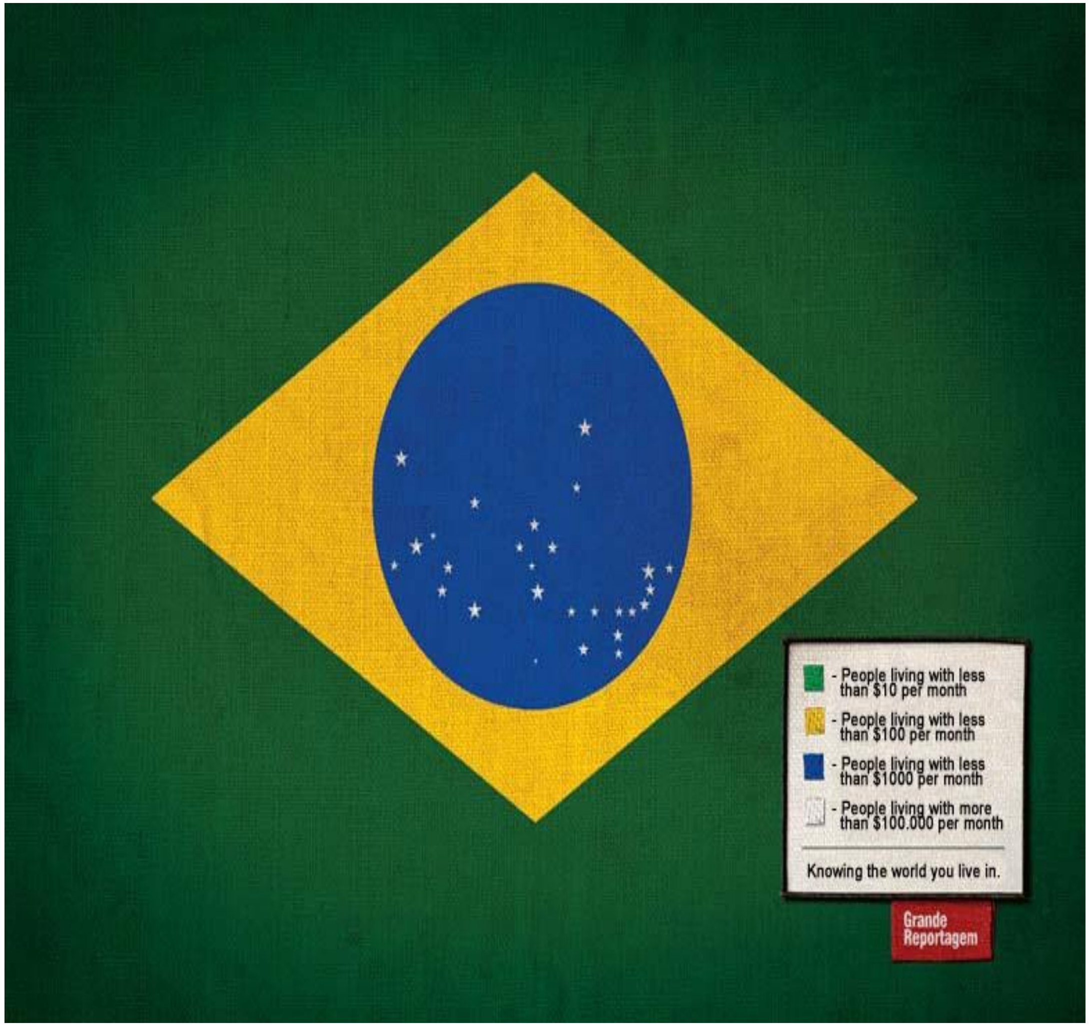
Europe


🟥 - Minor 14 yeas old working
🟨 - Minor 14 years old studying
Knowing the world you live in.
Grande Reportagem

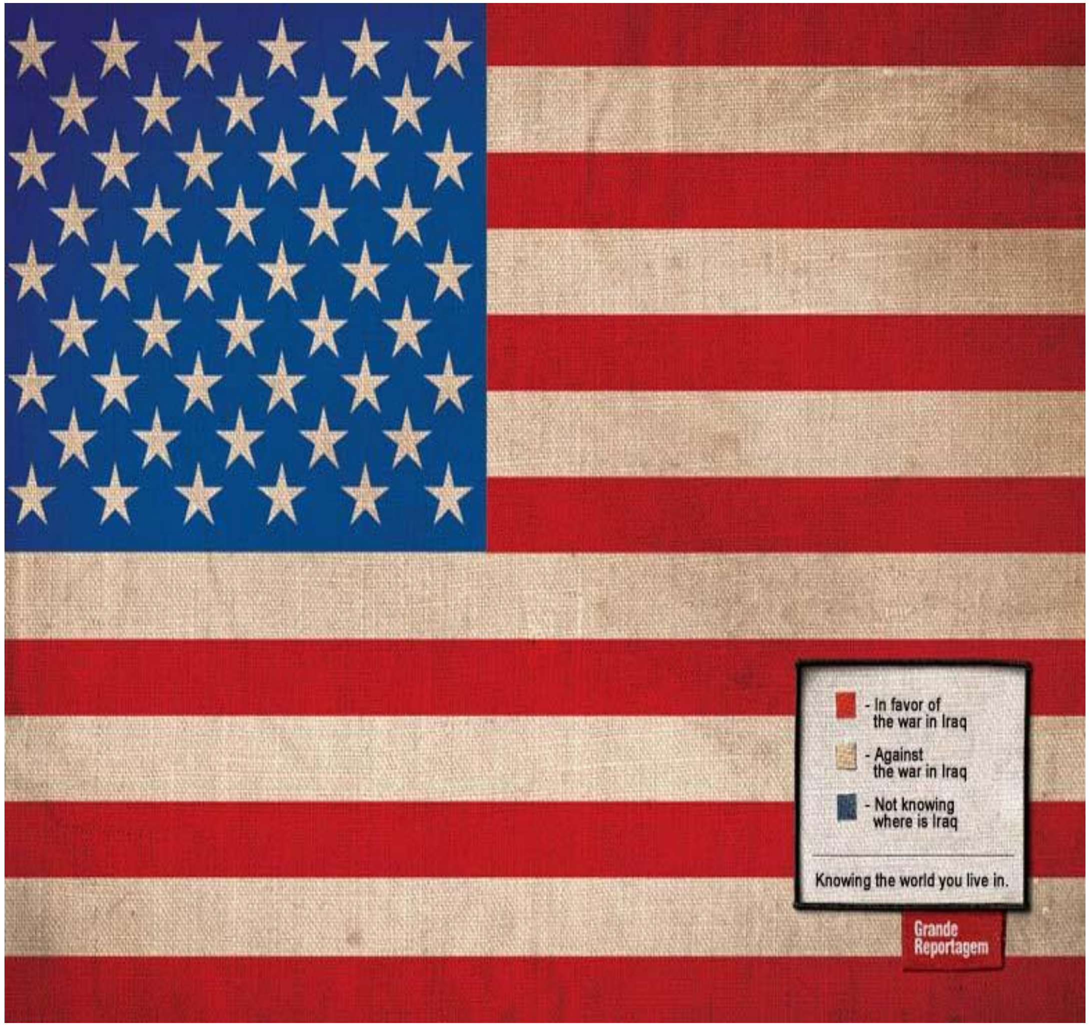
China


🟥 - Children dying before the age of one
🟩 - Children dying before the age of three
🟨 - Children reaching adult age
Knowing the world you live in.
Grande Reportagem

Burkina_Faso


🟩 - People living with less than $10 per month
🟨 - People living with less than $100 per month
🟦 - People living with less than $1000 per month
⬜ - People living with more than $100.000 per month
Knowing the world you live in.
Grande Reportagem

Brazil


🟥 - In favor of the war in Iraq
▨ - Against the war in Iraq
🟦 - Not knowing where is Iraq
Knowing the world you live in.
Grande Reportagem

USA

7 Vs.

- Volume
- Velocity
- Variety
- Variability
- Veracity
- Visualization
- Value

SOLUTION ARCHITECTURE

ROI understanding

- Business problem
- Current approach
- Advantages of data way of thinking and what problem you are solving
  - Fraud detection

Feature engineering

- Can I add, transform existing attributes to generate new attributes

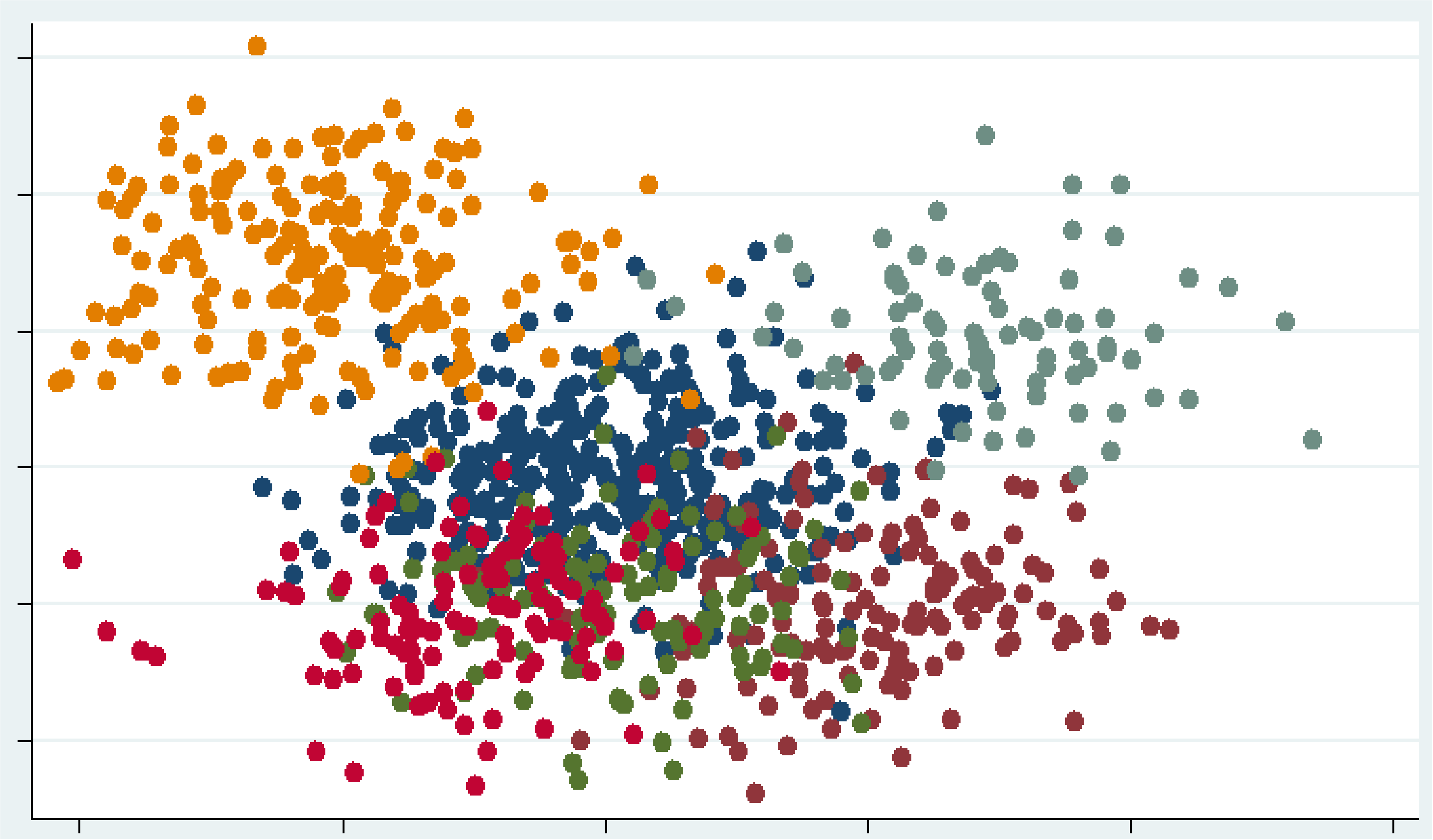
Getting the data into a structured form

Known/Easy to measure	Known/Easy to measure	Known/Easy to measure	Difficult to measure

Sharpening the data

- Missing values
- Type conversions (based on domain)
- Retaining important attributes

Data exploration



Model building

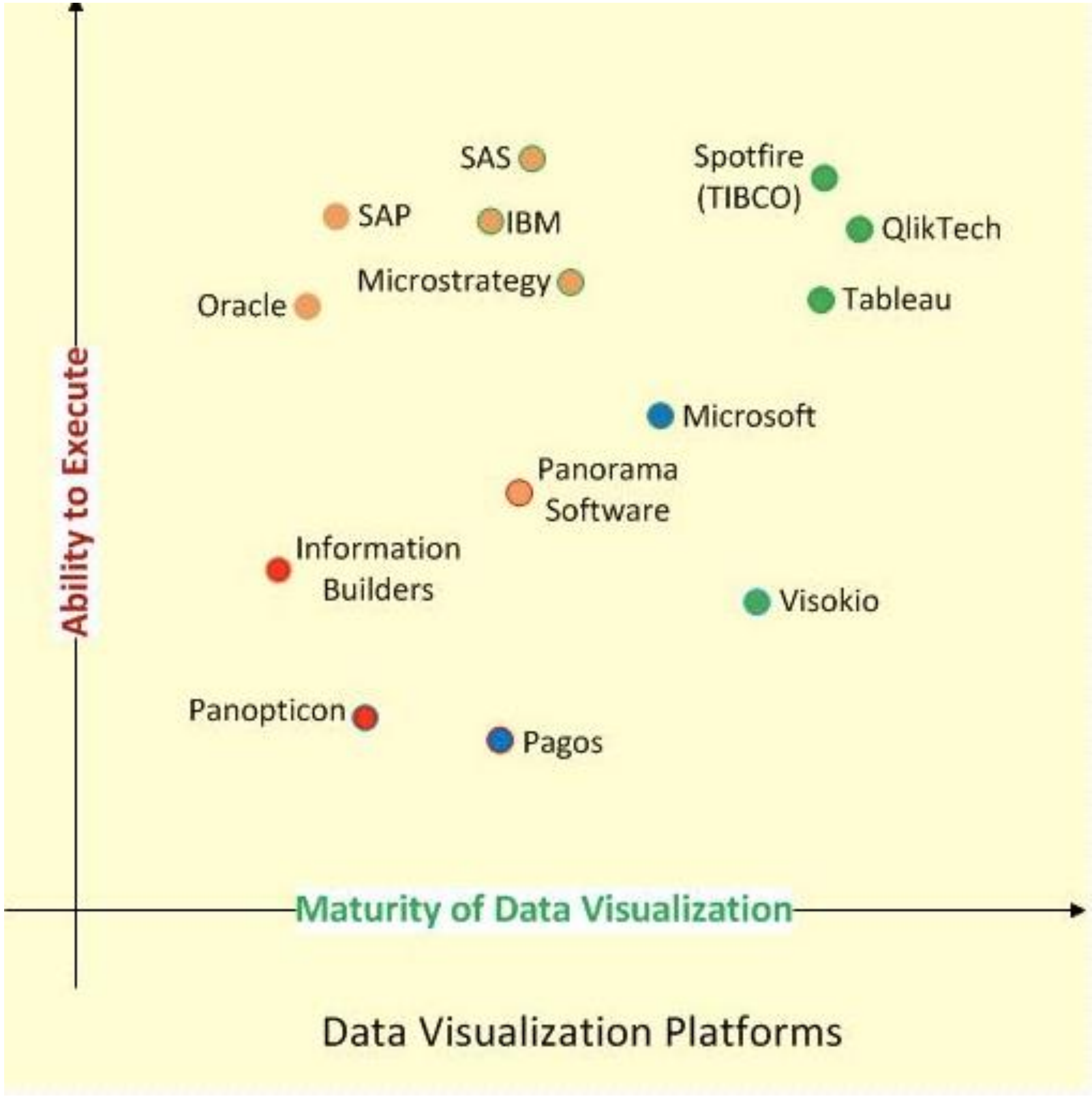
Story telling

ECO SYSTEM

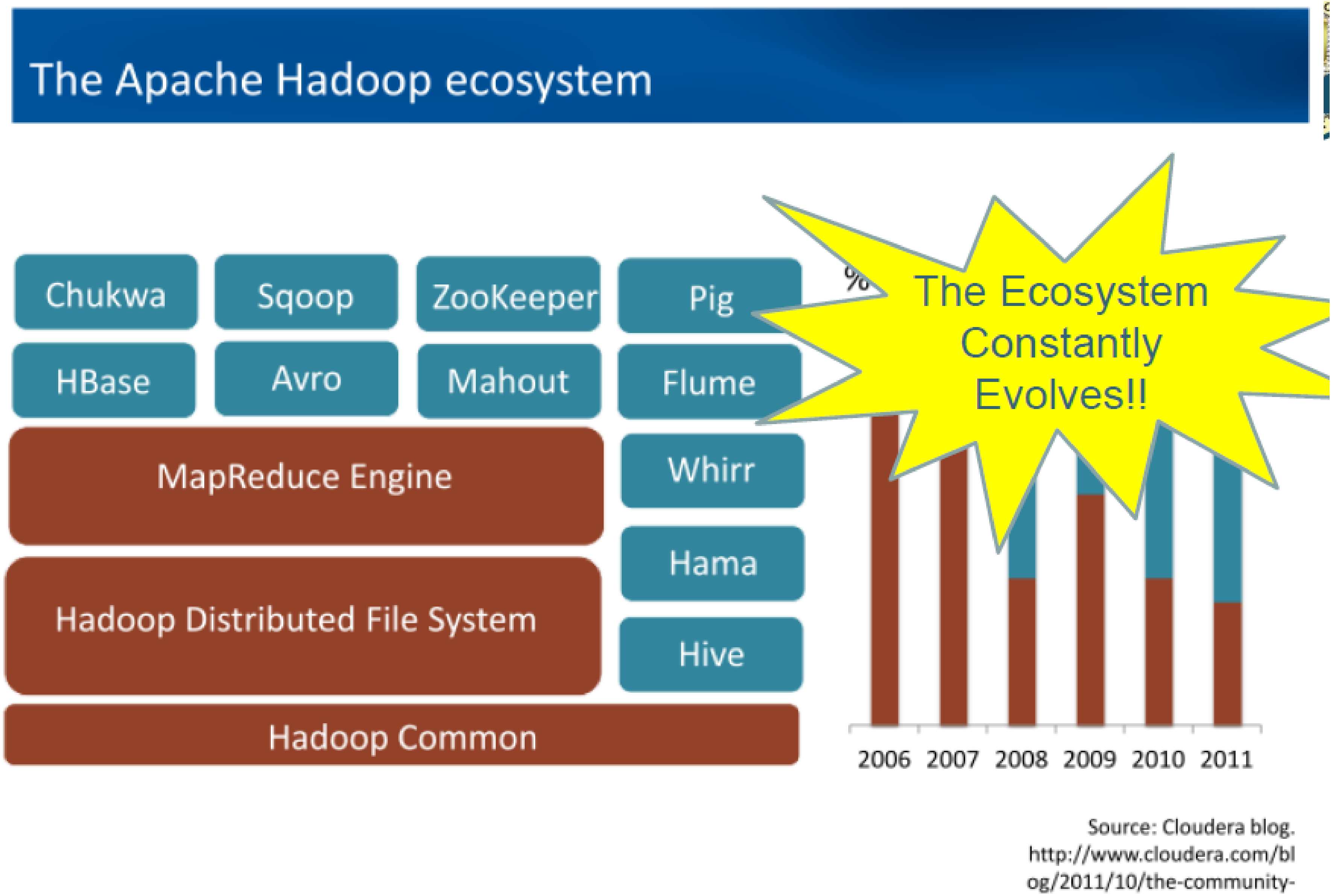
Data Science Environment

SPLUS
STATA
SPSS
SAS
MINITAB
R
JMP
MATLAB

Visualization tools



What to implement



The Apache Hadoop ecosystem

Chukwa

Sqoop

ZooKeeper

Pig

HBase

Avro

Mahout

Flume

MapReduce Engine

Whirr

Hama

Hadoop Distributed File System

Hive

Hadoop Common

The Ecosystem Constantly Evolves!!

Source: Cloudera blog. http://www.cloudera.com/blog/2011/10/the-community-

Learn enough domain

- Learn Enough Domain and look accessible in the meeting
  - http://www.ibm.com/analytics/us/en/solutions/index.html

Define the problem

- Goal
- Assumptions
- Process
- Business use

A definition

- We will identify customers who are likely to buy in the next campaign (or)

A better definition

- *We shall identify most likely target customers for a new campaign based on similar campaigns of the past.* **We assume that demographic and sales habits define behavior towards the campaign.** We will use demographic and sales data of the past campaigns to unearth relationship between the known characteristics of a customer to her reaction to a specific campaign.

Business use

- The customer can reduce the number of contacts while not compromising on revenues
- There are more complex cases where the business use is not obvious.
  - Telecom use
  - Clinical trials

ERROR METRICS

A rare disease

- 1 in 100,000 get it

- The model is “No body has it”
  - What is the accuracy?
    - 99.999

Quality of the analysis

- Types of errors in classification

	Predicted positive	Predicted negative
Actual positive	TP: 500	FN: 400
Actual negative	FP: 100	TN: 9000

Error metrics: Classification

- Accuracy is the percent of the predictions that were correct?
    - The "accuracy" is (9,000+500) out of 10,000 = 95%

More on error measures

- **Sensitivity**, *true positive rate*, or the **recall rate** measures the proportion of actual positives which are correctly identified as such

$$Recall = Sensitivity = P(\hat{Y} = 1 | Y = 1)$$

Precision

- Precision is how many of my predicted positives are actually positives

- $Precision = P(Y=1|\hat{Y}=1)$

- In some business cases, both precision and recall may be important. Then people use F1 statistic defined by

- $F1\ Statistic = Metric = \frac{2PR}{P+R}$

Specificity

- Measures the proportion of negatives which are correctly identified as such (e.g. the percentage of healthy people who are correctly identified as not having the condition, sometimes called the *true negative rate*).

- $Specificity = P(\hat{Y} = 0 | Y = 0)$

Summary

	P	
A	**63**	**37**
	28	72

TPR=Sensitivity=Recall= 0.63

FPR=1-specifificty= 0.28

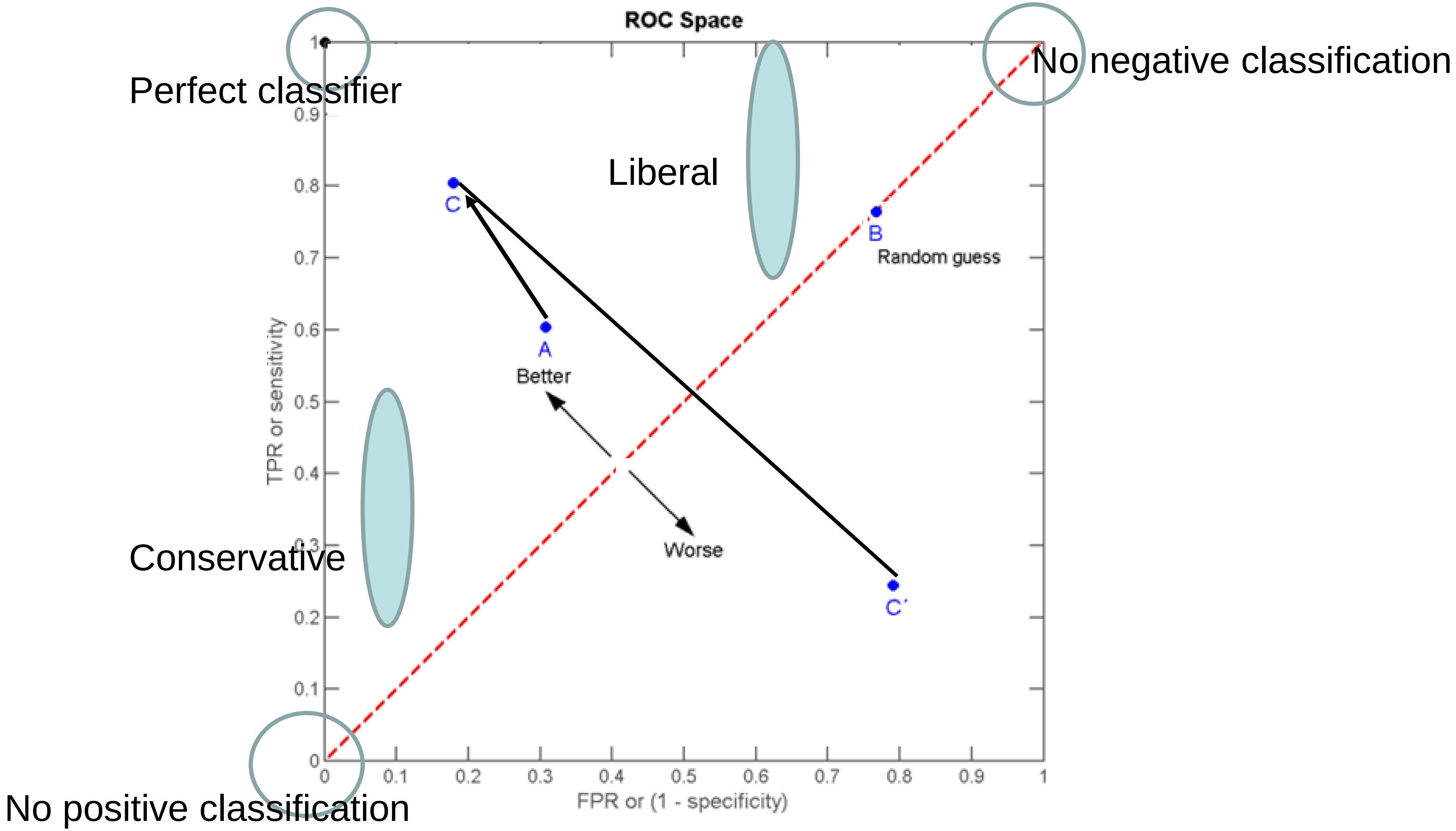
Precision= 0.69

F1= 0.66

Accuracy= 0.68

False positive rate = Percentage of negatives incorrectly classified

Receiver operating characteristic curve



Reference

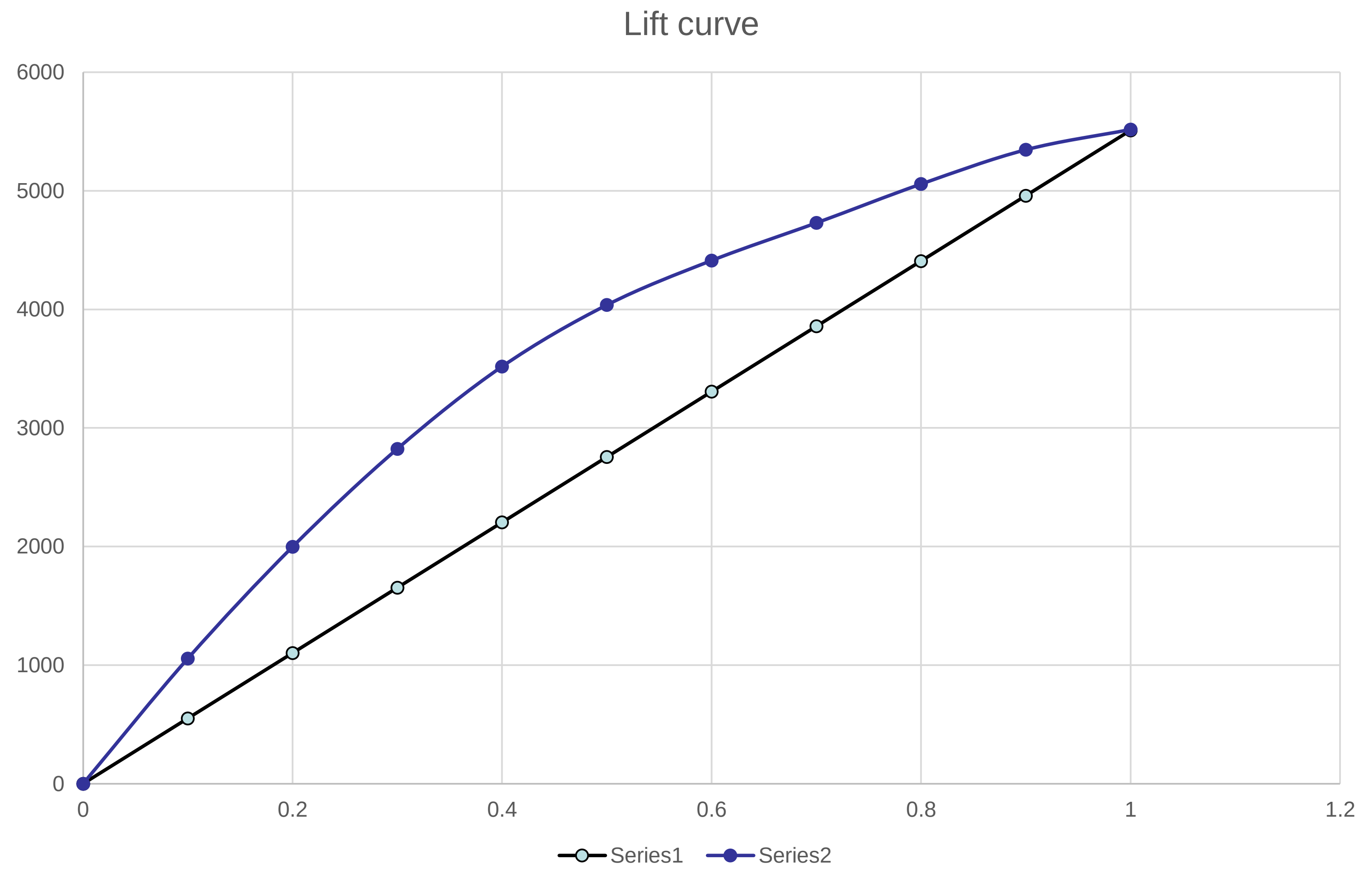
https://cours.etsmtl.ca/sys828/REFS/A1/Fawcett_PRL2006.pdf

Lift curves: Another goodness metric

- An analysis is done to identify potential customers. Around 47% of potentials are customers.
- Later it is applied on real data.
- The potentials are sorted in the order of the probability of becoming a customer based on the model

Bins	Random	Model
1162	551	1056
1162	551	942
1162	551	826
1162	551	694
1162	551	519
1162	551	375
1162	551	317
1162	551	328
1162	551	289
1162	551	171
11620	5517	5517
	0.474785	

Bins	Random	Model
0	0	0
10%	551	1056
20%	1102	1998
30.0%	1653	2824
40.0%	2204	3518
50.0%	2755	4037
60.0%	3306	4412
70.0%	3857	4729
80.0%	4408	5057
90.0%	4959	5346
100.0%	5510	5517

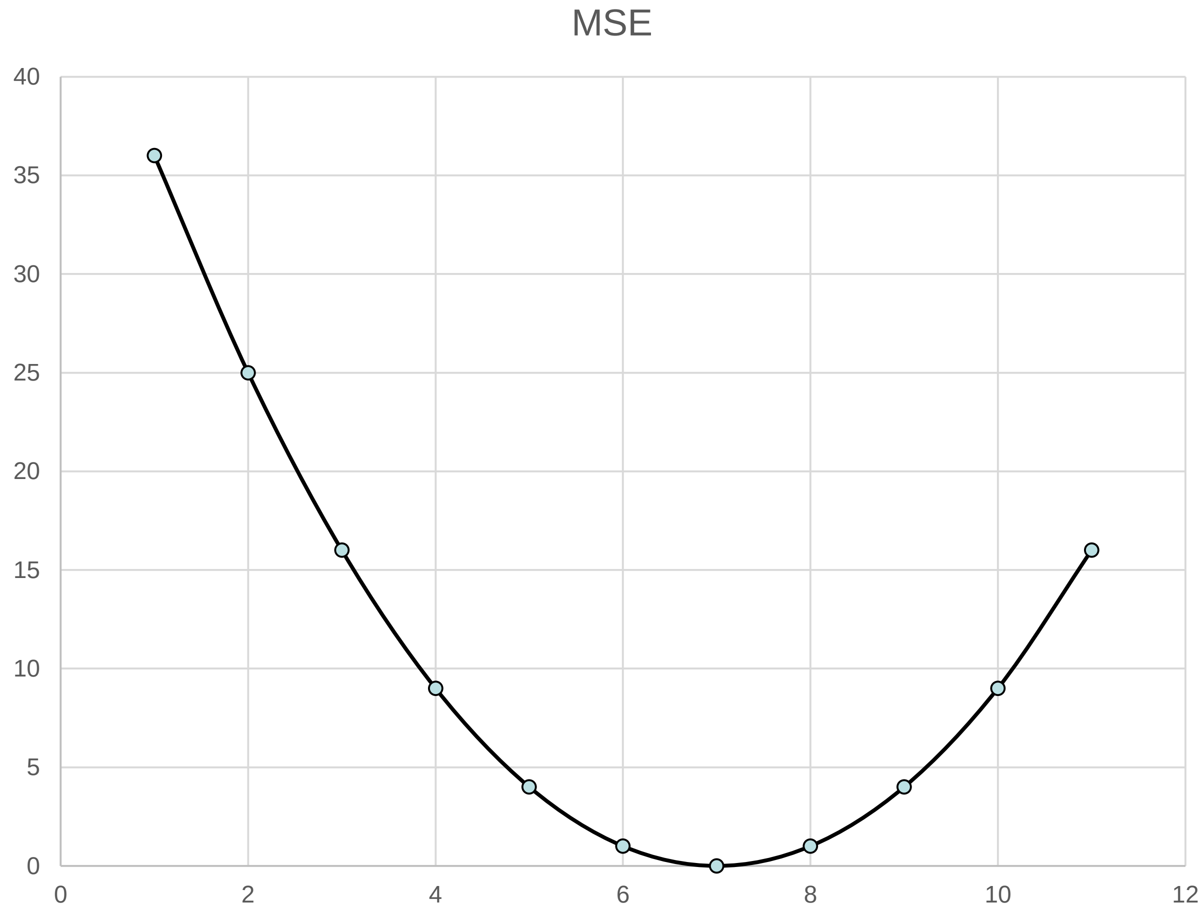

Lift curve
6000
5000
4000
3000
2000
1000
0
0
0.2
0.4
0.6
0.8
1
1.2
Series1
Series2

Multi-class

- One versus all metrics
- Any two class metrics
- Your own metrics

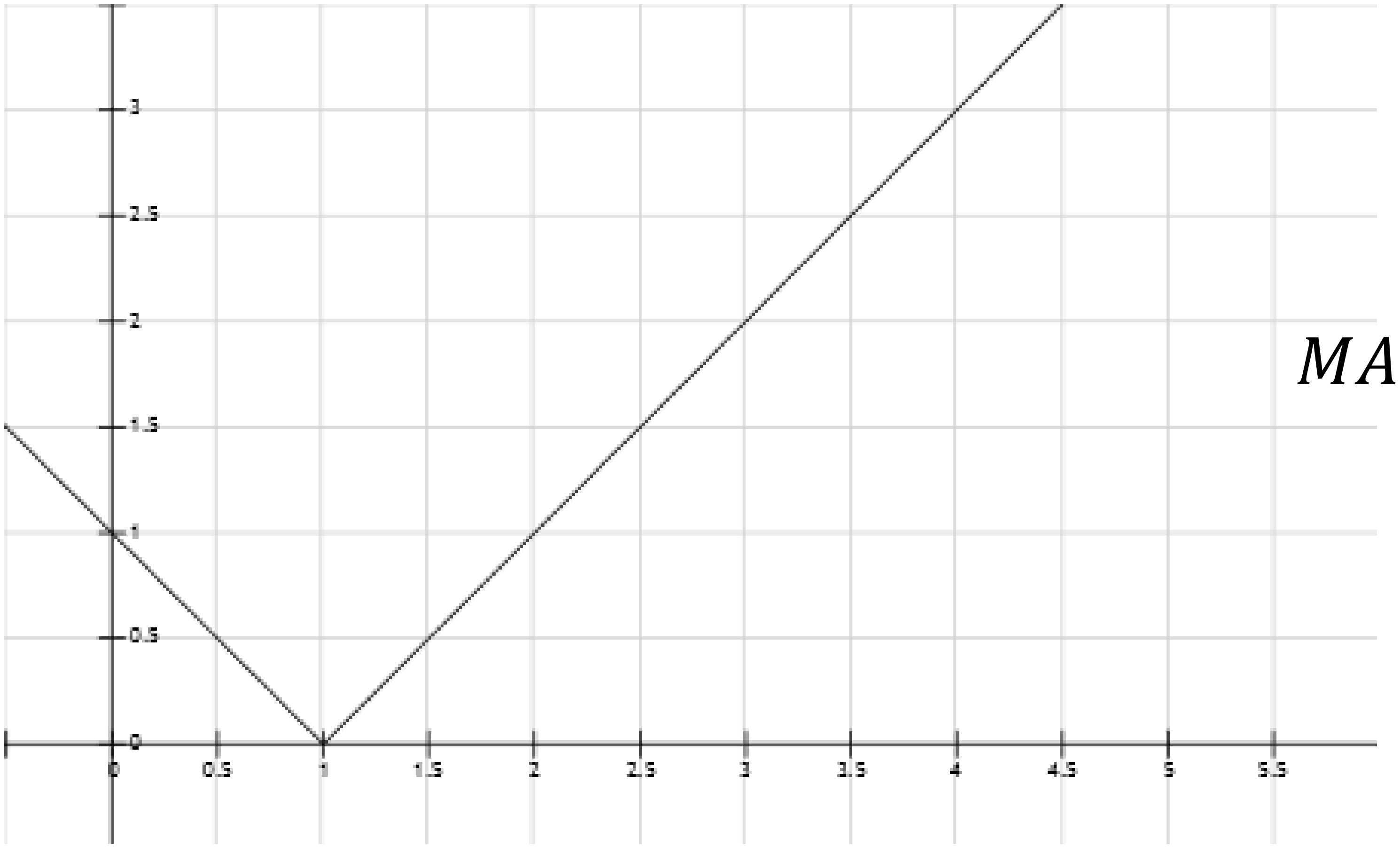
FORECASTING

MSE(Mean square error)



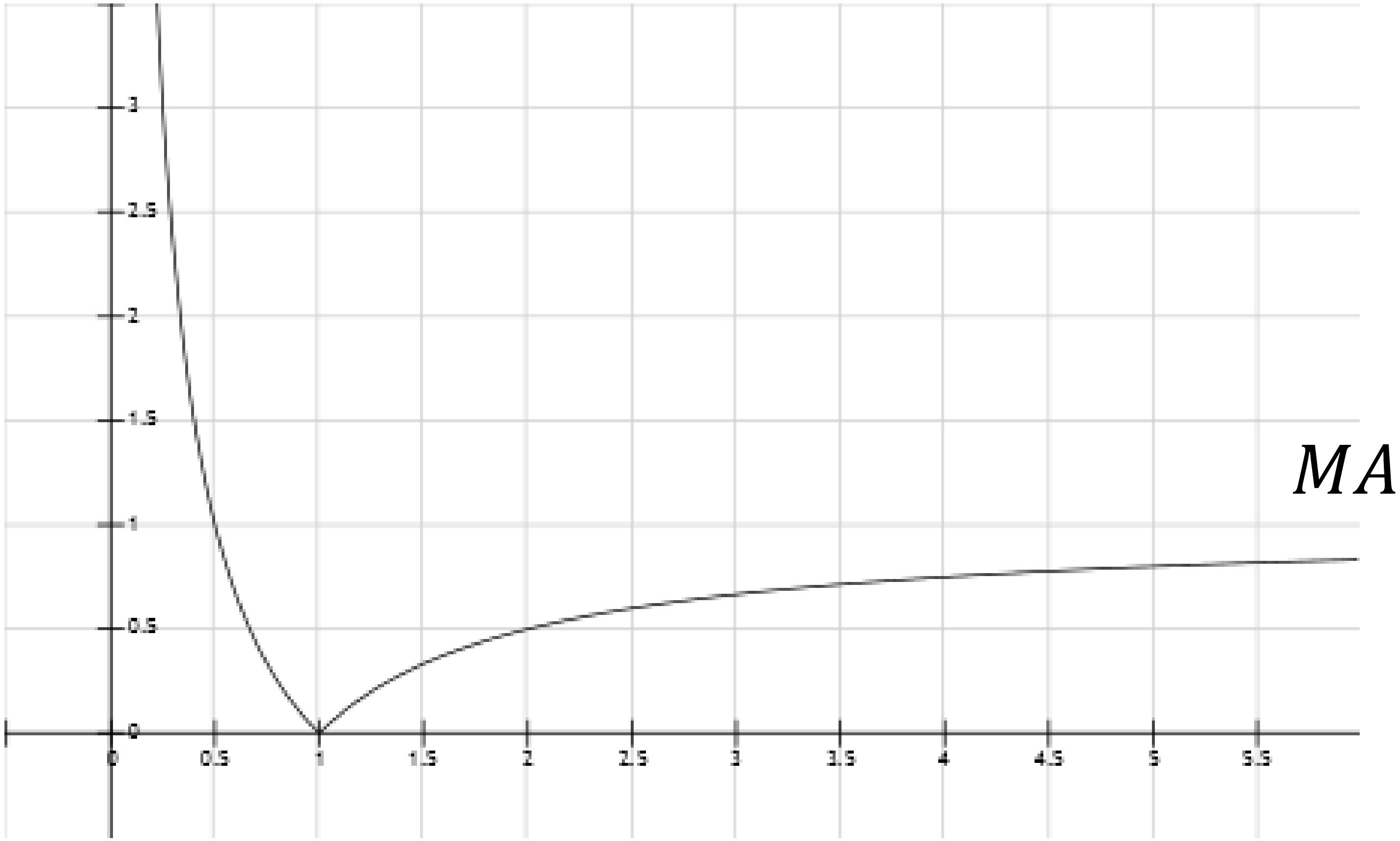
$$MSE = \frac{\sum_{i=1}^{n}(P_i - A_i)^2}{n}$$

MAE (Mean absolute error)



$$MAE = \frac{\sum_{i=1}^{n} |P_i - A_i|}{n}$$

MAPE(Mean absolute percentage error)



$$MAPE = \frac{\sum_{i=1}^{n} \frac{|P_i - A_i|}{A_i}}{n}$$

NMSE (Normalized Mean Square error)

$$NMSE = \frac{MSE\ of\ developed\ model}{MSE\ of\ naive\ model}$$

International School of Engineering

Plot 63/A, 1st Floor, Road # 13, Film Nagar, Jubilee Hills, Hyderabad - 500 033

For Individuals:	+91-9502334561/63 or 040-65743991
For Corporates:	+91-9618483483
Web:	http://www.insofe.edu.in
Facebook:	https://www.facebook.com/insofe
Twitter:	https://twitter.com/Insofeedu
YouTube:	http://www.youtube.com/InsofeVideos
SlideShare:	http://www.slideshare.net/INSOFE
LinkedIn:	http://www.linkedin.com/company/international-school-of-engineering

This presentation may contain references to findings of various reports available in the public domain. INSOFE makes no representation as to their accuracy or that the organization subscribes to those findings.