Making Claims with Data

Unit 4

Nepal Data Literacy Program, 2019

Organized by

THE WORLD BANK
IBRD • IDA

Supported by

UKaid
from the British people

Objectives of the Unit

- Get acquainted with data and underlying context
- Become familiar with group characteristics of data
- Learn what can and cannot be implied by data
- How all of this leads to better decision-making

Image source: Cookerly.com

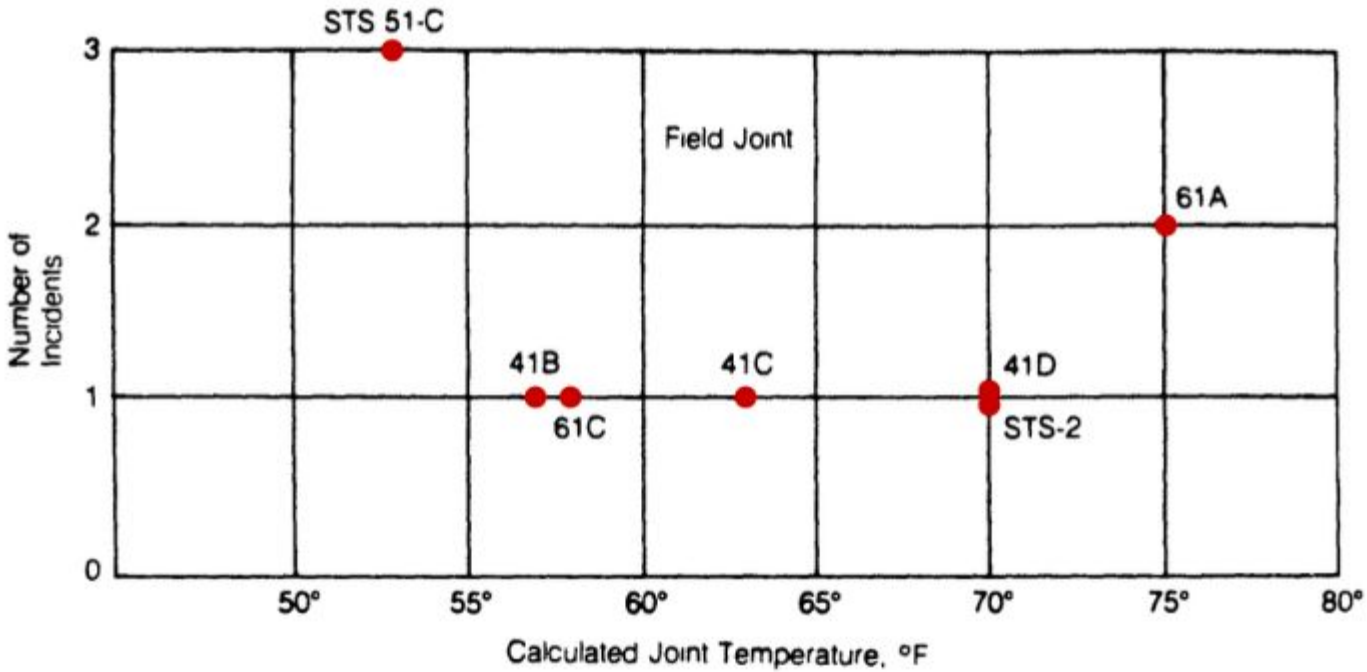
Module 1

Getting Familiar With the Context

Credit:http://bit.ly/2FvFhzU

Number of O-Ring incidents vs. Joint Temperature

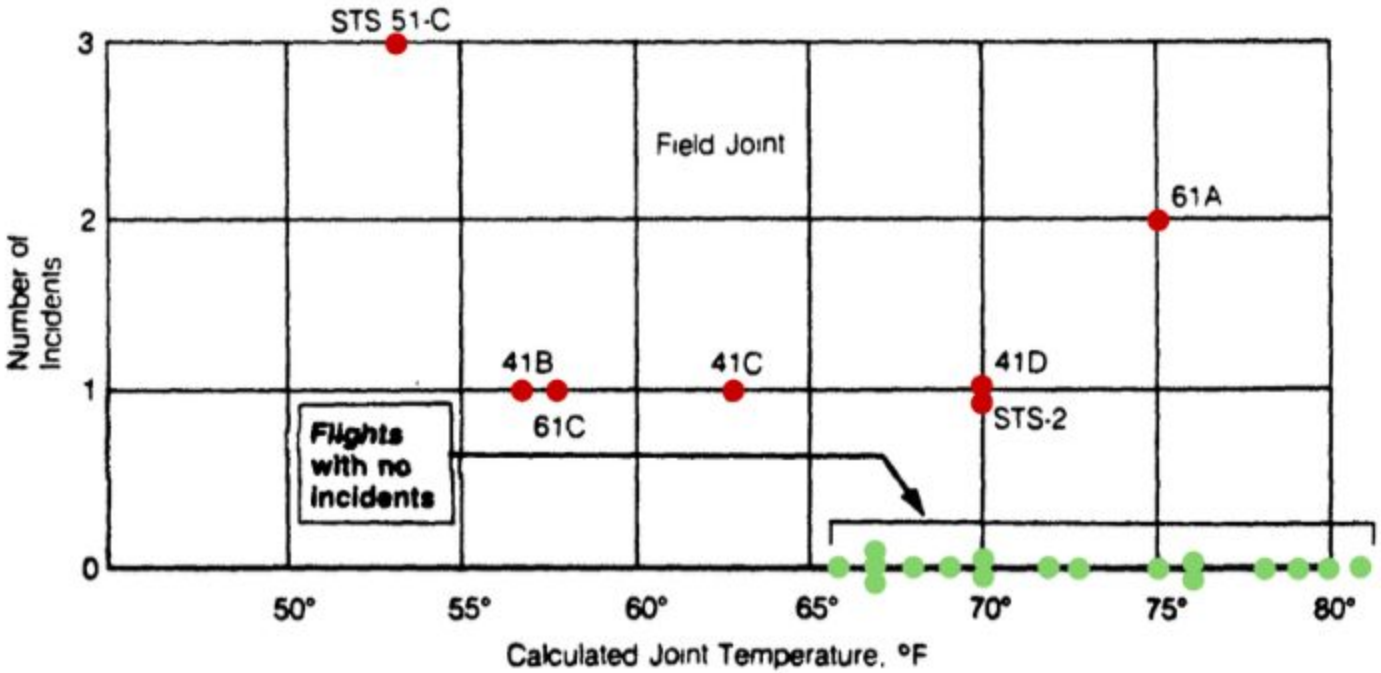
(Incidents when O-Rings failed)



Credit: http://bit.ly/2Xdr1Xd

Number of O-Ring incidents vs. Joint Temperature

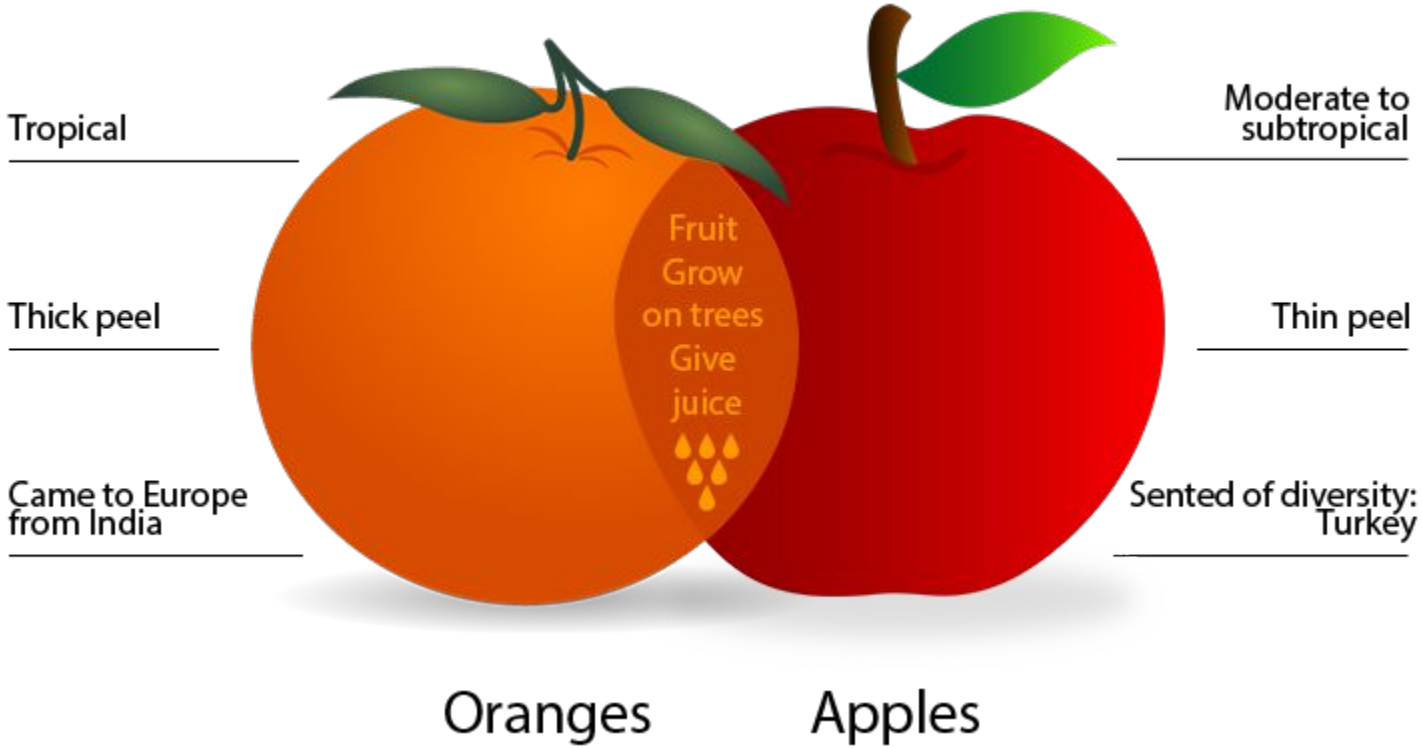
(failures AND successes)



Credit:http://bit.ly/2FvFhzU

“A **careful analysis** of the flight history of O-ring performance would have revealed the **correlation** of O-ring damage in low temperature”

Credit:http://bit.ly/2FvFhzU



Credit: http://bit.ly/2XFfb7y

Question of Interest ?

Internal Data

External Data

Credit: http://tiny.cc/0r676y, http://tiny.cc/7awe8y

What is a Claim ?

- Claim = Statistical Inquiry

How old am
I ?

How old are
students in
our class ?

1 Response

N Responses

NON STATISTICAL

STATISTICAL

Credit: http://tiny.cc/l5iw7y

What is a Good Claim ?

- Clear and Precise

Men are taller than women

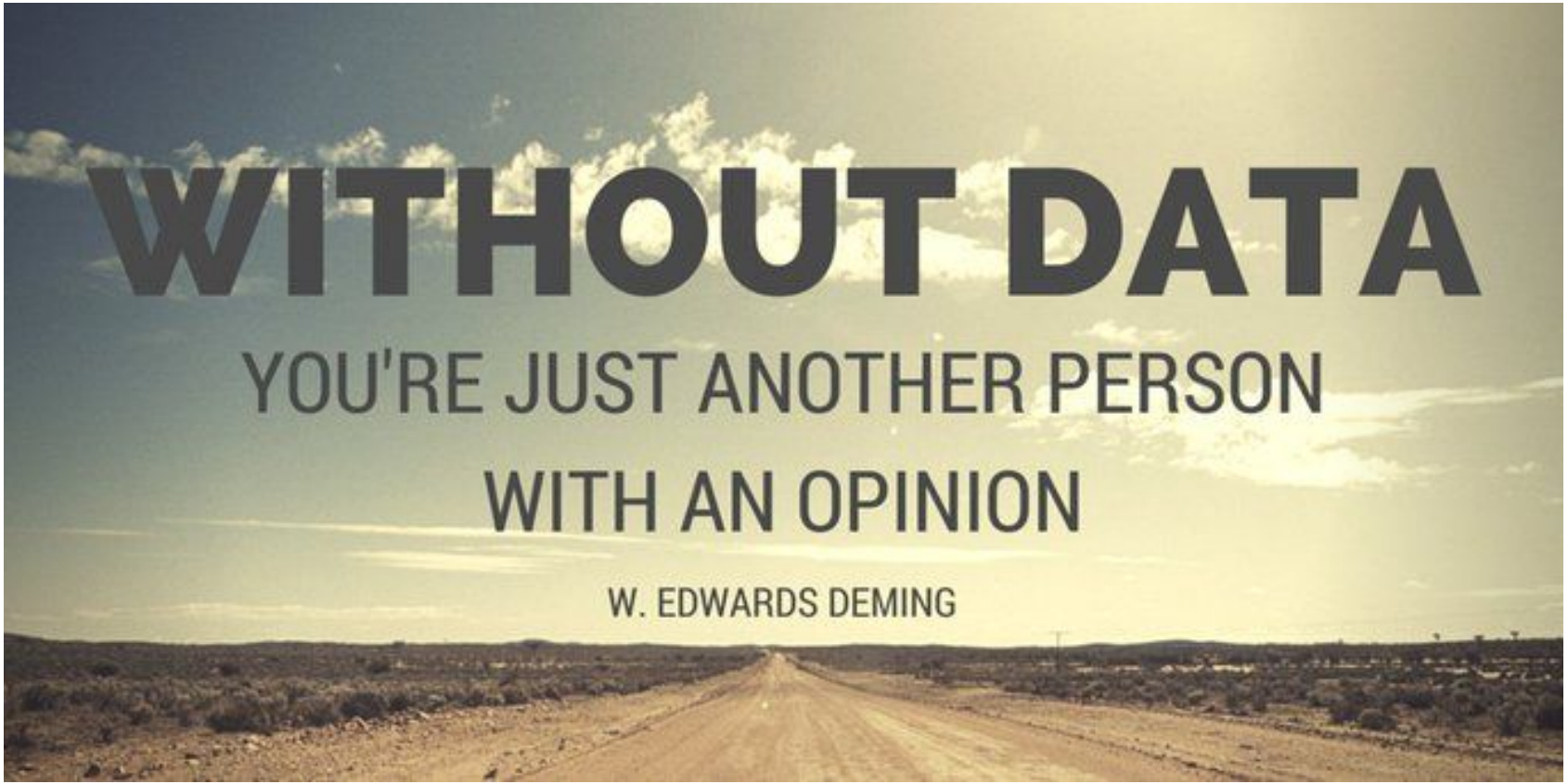
Men are generally taller than women

On an average, men are 10 cm taller than women

GENERIC → SPECIFIC

Credit: http://tiny.cc/v1iw7y

A good claim has to be grounded in data



Credit: http://tiny.cc/l5iw7y

Group of Measurements

2 YEARS

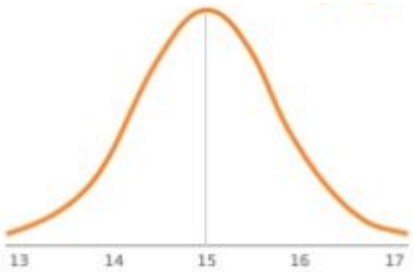
18 YEARS

80 YEARS

VARIABILITY

Credit: http://tiny.cc/u8iw7y, http://tiny.cc/pkmf8y

Group of Measurements



Variability

LESS

MORE

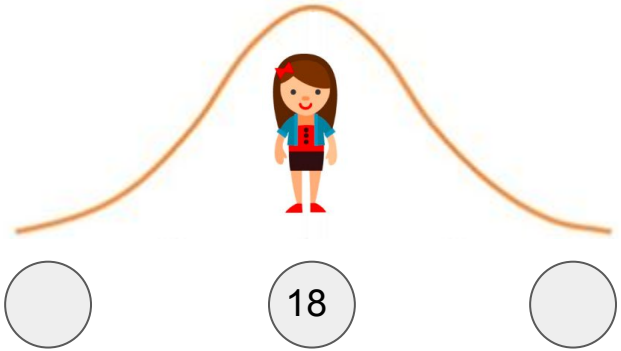
How “Spread Out” a group of measurements are

Credit: http://tiny.cc/u8iw7y

Measures of Variability

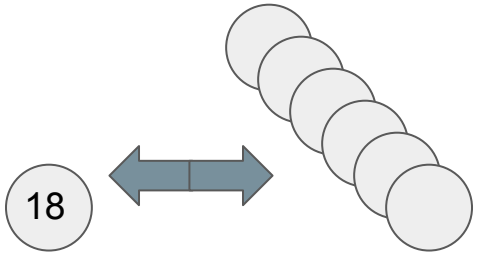
1

Describes the distribution



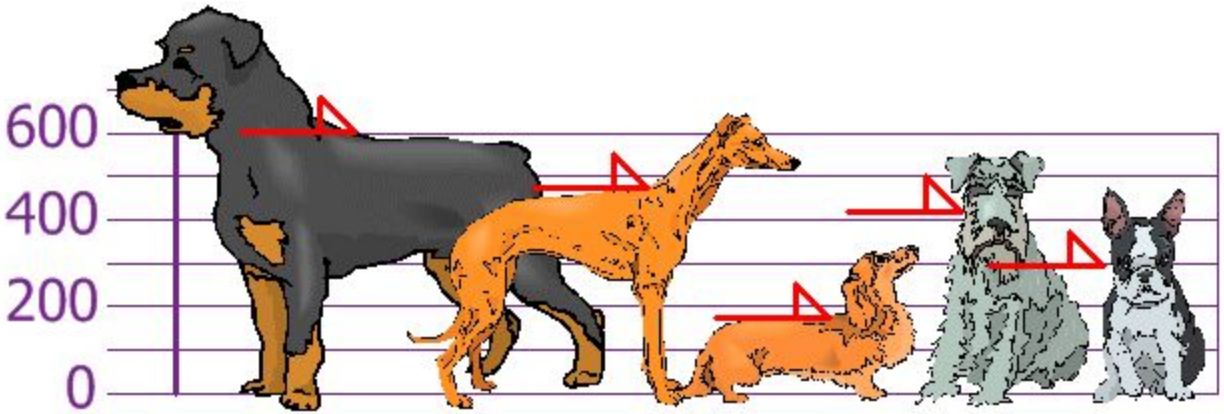
2

Measures how well an individual score represents the distribution



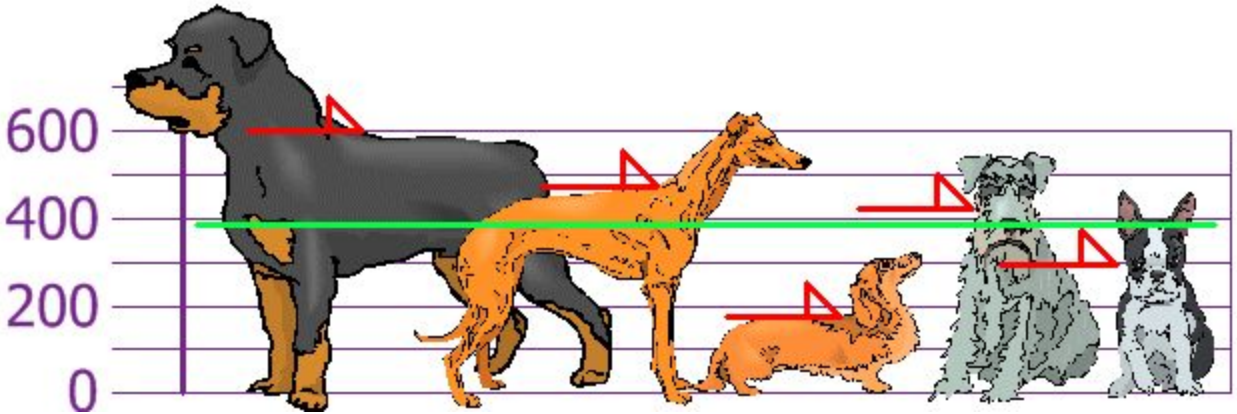
Credit: http://tiny.cc/u8iw7y

Measures of Variability - Variance



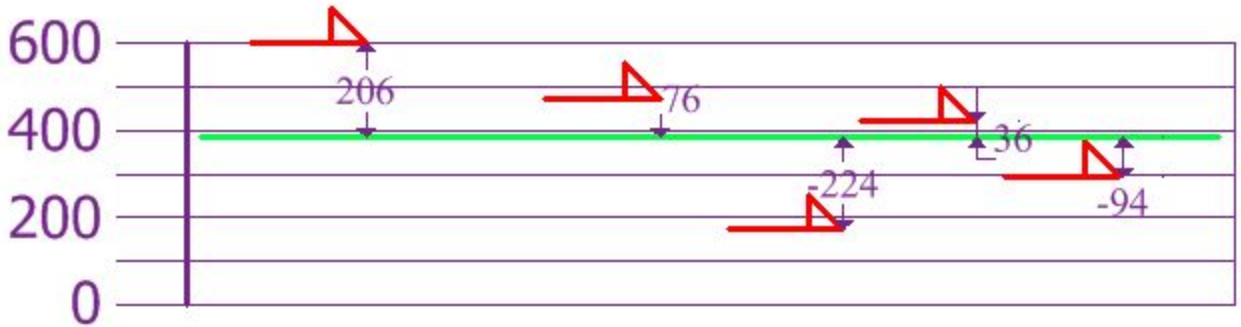
Credit: http://bit.ly/2JbLge3

Measures of Variability



Credit: http://bit.ly/2JbLge3

Measures of Variability



Credit: http://bit.ly/2JbLge3

Measures of Variability

$$= \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5}$$

$$= \frac{42436 + 5776 + 50176 + 1296 + 8836}{5}$$

$$= \frac{108520}{5}$$

$$= 21704$$

Credit: http://bit.ly/2JbLge3

Measures of Variability

$$= \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5}$$

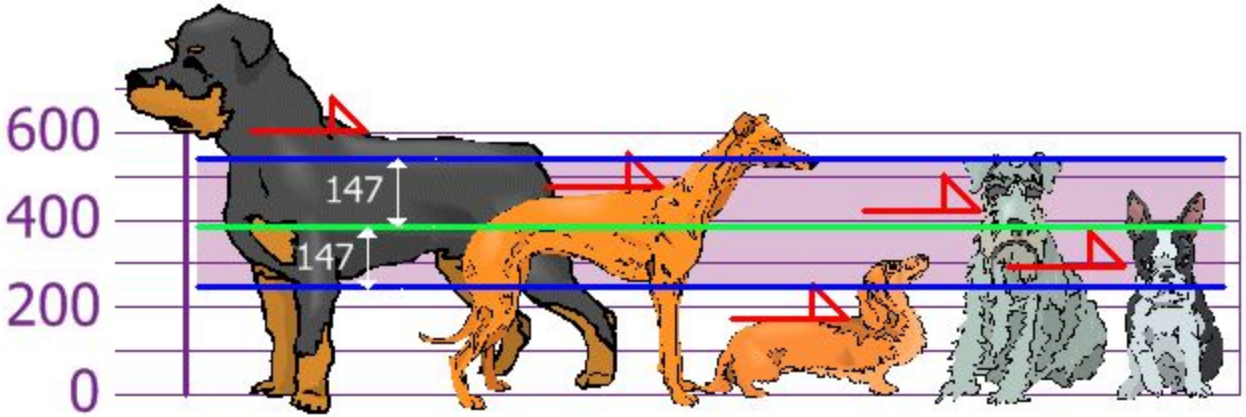
$$= \frac{42436 + 5776 + 50176 + 1296 + 8836}{5}$$

$$= \frac{108520}{5}$$

$$= 21704$$

Credit: http://bit.ly/2JbLge3

Measures of Variability - Std Dev



= $\sqrt{21704}$

= 147.32...

= **147** *(to the nearest mm)*

Credit: http://bit.ly/2JbLge3

What is a Good Claim ?

- About a group
- A statistical statement - something that varies within the group
- Uses language that recognizes that variability

What is a Good Claim ?

- Phrased with Precision - we use:
  - Words describing proportions - like: some, most
  - Percentages
  - Proportions

Credit: http://tiny.cc/l5iw7y

What is a Good Claim ?

“Most Nepalese live in a city”

- About a group (or groups) → Nepalese is a group
- A statistical statement → Living in a city is a measure that changes
- Uses language that recognizes variability → Use of Qualifier “Most“

Is that a CLAIM ?

What is a Good Claim ?

“Most Nepalese live in a city”

- About a group (or groups) → Nepalese is a group
- A statistical statement → Living in a city is a measure that changes
- Uses language that recognizes variability → Use of Qualifier “Most“

Is that a CLAIM ? ✓

Is that CLAIM precise ?

What is a Good Claim ?

“ Pokhara is the capital of Gandaki ”

Is that a CLAIM ?

What is a Good Claim ?

“ Pokhara is the capital of Gandaki ”

Is that a CLAIM ? ✕

It does not anticipate variability in the data. It is just a fact which one can verify by looking up on the map

What is a Good Claim ?

“ Gandaki is bigger than Karnali ”

Is that a CLAIM ?

What is a Good Claim ?

“ Gandaki is bigger than Karnali ”

Is that a CLAIM ? ✓

The group in this claim could be all the people in Gandaki or Karnali, and the thing that varies is which city they live in

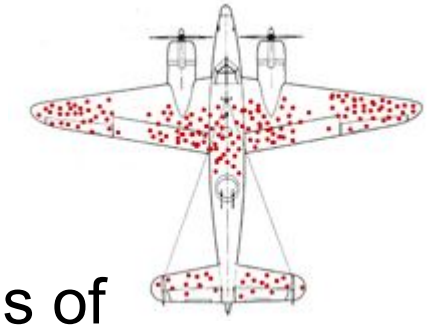
What is a Good Claim ?

“ Gandaki is bigger than Karnali ”

Is that a CLAIM ? ✓

Is that CLAIM precise ?

A Word of Caution …



- One must also consider the data in terms of possible biases that could influence or even distort its characteristics and content
- Data never contains absolute truths, only relative truths that offer one a more or less useful view of a problem
- Always be aware of the truthfulness of data and apply critical reasoning as part of your analysis of it

Credit: http://tiny.cc/tmwe8y

Thank You

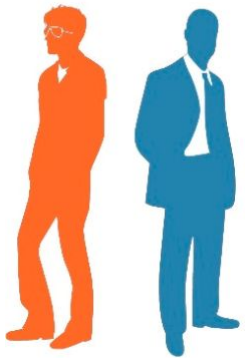
Credit: http://bit.ly/2N4wniG

Module 2

Assess Validity of Claims

What is Validity ?

- Appropriateness of inferences drawn from data
- Implies purpose for which inferences are drawn

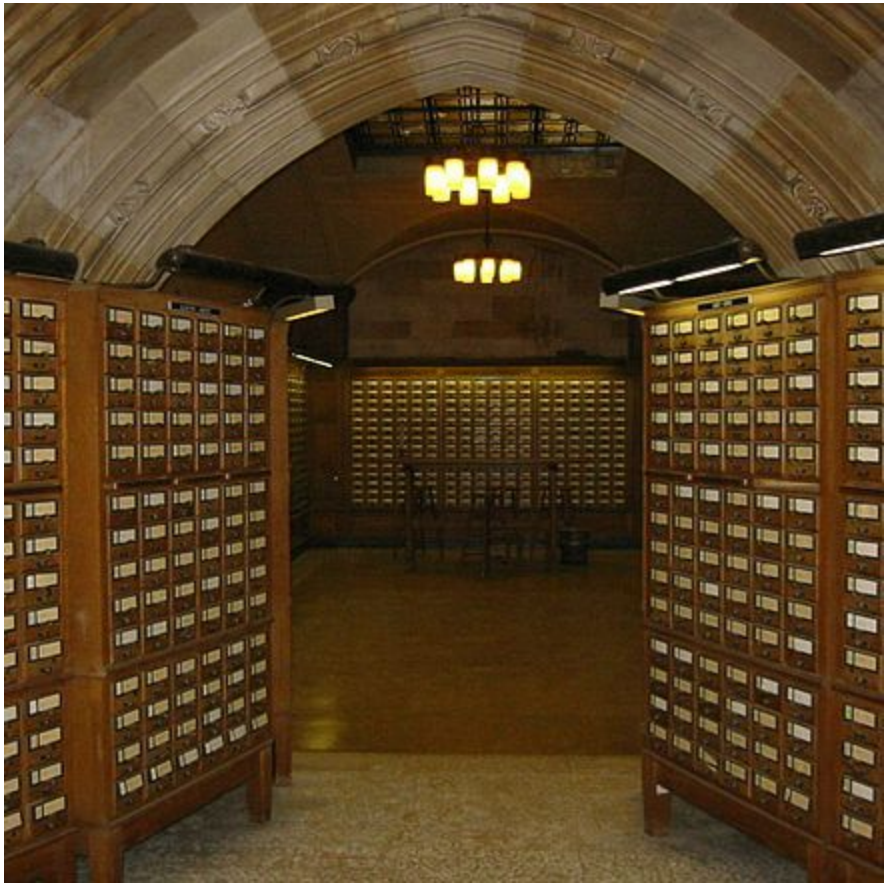


EXAMPLE
use of information
posted at Facebook
and other
social media sites
by employers
to make hiring decisions
(i.e., predictions about
future job performance)

Credit: http://tiny.cc/44xe8y, http://tiny.cc/wr8x7y, http://bit.ly/2XaYuwz

How to Assess a Claim ?

- Understand the data - available values
- Understand the data - unavailable or missing values
- Understand the data - relationships
- Understand the data - visualizations



Credit: http://bit.ly/32b4eu6

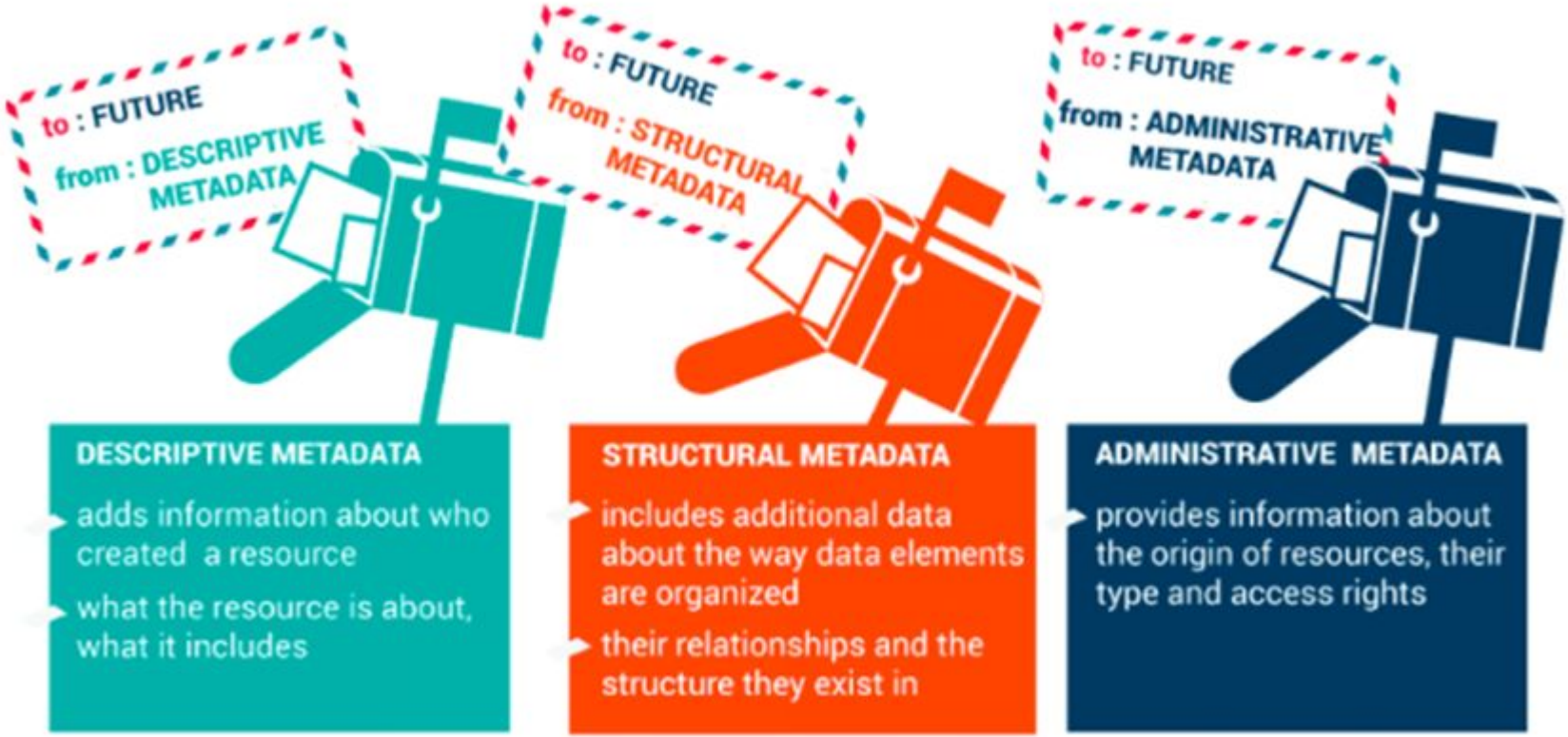
Meta Data

Metadata, you see, is really a love note – it might be to yourself, but in fact it's a love note to the person after you, or the machine after you, where you've saved someone that amount of time to find something by telling them what this thing is.

- Jason Scott

Credit: http://tiny.cc/viwe8y

Meta Data - Types



Credit: http://tiny.cc/6jwe8y

Metadata - Example

DATA ACCESS AND LICENSING

This dataset is classified as **Public** under the Access to Information Classification Policy. Users inside and outside the Bank can access this dataset.

This dataset is licensed under **CC-BY 4.0**

RELATED LINKS

- WDI products page
- Data Updates and Errata
- Database Archives
- International Debt Statistics

AVAILABLE FORMATS

- Online Tables

SHARE METADATA

The information on this page (the dataset metadata) is also available in these formats.

PRINT EMAIL JSON RDF

TWEET SHARE SHARE +

Back

The primary World Bank collection of development indicators, compiled from officially-recognized international sources. It presents the most current and accurate global development data available, and includes national, regional and global estimates

Overview | Data & Resources | Visualization | Additional Information | Citations

Acronym:	WDI
Type:	Time Series
Topics:	Agriculture and Food Security, Climate Change, Economic Growth, Education, Energy and Extractives, Environment and Natural Resources, Financial Sector Development, Gender, Health, Nutrition and Population, Macroeconomic Vulnerability and Debt, Poverty, Private Sector Development, Public Sector Management, Social Development, Social Protection and Labor, Trade, Urban Development
Economy Coverage:	High Income, IBRD, IDA, Low Income, Lower Middle Income, Upper Middle Income
Languages Supported:	English, Arabic, Chinese, French, Spanish
Number of Economies:	217
Geographical Coverage:	World, East Asia & Pacific, American Samoa, Australia, Brunei Darussalam, Cambodia, China More...
External Contact Email:	Data@worldbank.org
Access Options:	Query Tool, API Documentation, Download
Temporal Coverage:	1960 - 2018
Update Frequency:	Quarterly
Release Date:	June 11, 2010
Harvest Source:	Indicators API
Harvest Source ID:	2
Last Updated:	June 28, 2019

Credit: http://bit.ly/2XcwtEI

Metadata - Example

Name

- API_SH.DYN.NMRT_DS2_en_csv_v2_3824
- Metadata_Country_API_SH.DYN.NMRT_DS2_en_csv_v2_3824
- Metadata_Indicator_API_SH.DYN.NMRT_DS2_en_csv_v2_3824

	A	B	C	D	E
1	Country Code	Region	IncomeGroup	SpecialNotes	TableName
2	ABW	Latin America & Caribbean	High income		Aruba
3	AFG	South Asia	Low income		Afghanistan
4	AGO	Sub-Saharan Africa	Lower middle income		Angola
5	ALB	Europe & Central Asia	Upper middle income		Albania
6	AND	Europe & Central Asia	High income		Andorra
7	ARB			Arab World aggregate. Arab World is composed of members of the League of Arab States.	Arab World
8	ARE	Middle East & North Africa	High income		United Arab Emirates
9	ARG	Latin America & Caribbean	Upper middle income		Argentina
10	ARM	Europe & Central Asia	Upper middle income		Armenia
11	ASM	East Asia & Pacific	Upper middle income		American Samoa
12	ATG	Latin America & Caribbean	High income		Antigua and Barbuda
13	AUS	East Asia & Pacific	High income	Fiscal year end: June 30; reporting period for national accounts data: FY.	Australia
14	AUT	Europe & Central Asia	High income	A simple multiplier is used to convert the national currencies of EMU members to euros. The following irrevocable	Austria
15	AZE	Europe & Central Asia	Upper middle income		Azerbaijan
16	BDI	Sub-Saharan Africa	Low income		Burundi
17	BEL	Europe & Central Asia	High income	A simple multiplier is used to convert the national currencies of EMU members to euros. The following irrevocable	Belgium
18	BEN	Sub-Saharan Africa	Low income		Benin
19	BFA	Sub-Saharan Africa	Low income		Burkina Faso
20	BGD	South Asia	Lower middle income	Fiscal year end: June 30; reporting period for national accounts data: FY.	Bangladesh
21	BGR	Europe & Central Asia	Upper middle income		Bulgaria
22	BHR	Middle East & North Africa	High income		Bahrain
23	BHS	Latin America & Caribbean	High income		Bahamas, The
24	BIH	Europe & Central Asia	Upper middle income		Bosnia and Herzegovina
25	BLR	Europe & Central Asia	Upper middle income	Data before 2015 were adjusted to reflect the new denomination effective from July 1, 2016 (BYN), a decrease of 1	Belarus

Credit: http://bit.ly/2XcwtEI

How to Assess a Claim ?

- Understand the data

Secondary Data

Primary Data

Credit: http://bit.ly/2Xif2Yj

How to Assess a Claim ?

- Secondary Data - Sources

Internal Data	External Data

Credit: http://bit.ly/2Xif2Yj

How to Assess a Claim ?

- Secondary Data

Advantages	Disadvantages

Credit: http://bit.ly/2Xif2Yj

Secondary Data - Extending Limits

- Identify Potential Gaps
- Plug External Data
- Feature Engineering
- Discover New Insights

Identify Potential Gaps

नेपालमा हेलिकोप्टर दुर्घटनाको इतिहास

मिति (सन्)	हेलिकोप्टर कम्पनी	दुर्घटना स्थल	मृतक संख्या
१९७९	भीमीआईपी	लाङ्टाङ्	६
१९९३	हिमालयन हेलि	लाङ्टाङ्	०
१९९६	नेपाल एयरवेज	सोताङ्	०
१९९७	कर्णाली	बुपेन छोलिङ्	१
१९९७	गोरखा एयरलाइन्स	कालिकोट	०
१९९८	भीमीआईपी	दिपायल	०
१९९८	एसियन एयरलाइन्स	मूल सर्क	३
१९९९	कर्णाली एयर	लिसंखु, सिन्धुपाल्चोक	०
१९९९	मनकामना एयरवेज	रामेछाप	०
२००१	एयर अनन्य	मिमी	०
२००१	फिस्टल एयर	रारा ताल	४

No Geospatial Information

Credit: http://bit.ly/2K3mXBI

1

नेपालमा हेलिकोप्टर दुर्घटनाको इतिहास

मिति (सन्)	हेलिकोप्टर कम्पनी	दुर्घटना स्थल	मृतक संख्या
१९७९	भीमीआईपी	लाङ्टाङ्	६
१९९३	हिमालयन हेलि	लाङ्टाङ्	०
१९९६	नेपाल एयरवेज	सोताङ्	०
१९९७	कर्णाली	थुप्टेन छोलिङ्	१
१९९७	गोरखा एयरलाइन्स	कालिकोट	०
१९९८	भीमीआईपी	दिपायल	०
१९९८	एसियन एयरलाइन्स	मूल खर्क	३
१९९९	कर्णाली एयर	लिसंखु, सिन्धुपाल्चोक	०
१९९९	मनकामना एयरवेज	रामेछाप	०
२००१	एयर अनन्य	मिमी	०
२००१	फिस्टल एयर	रारा ताल	४

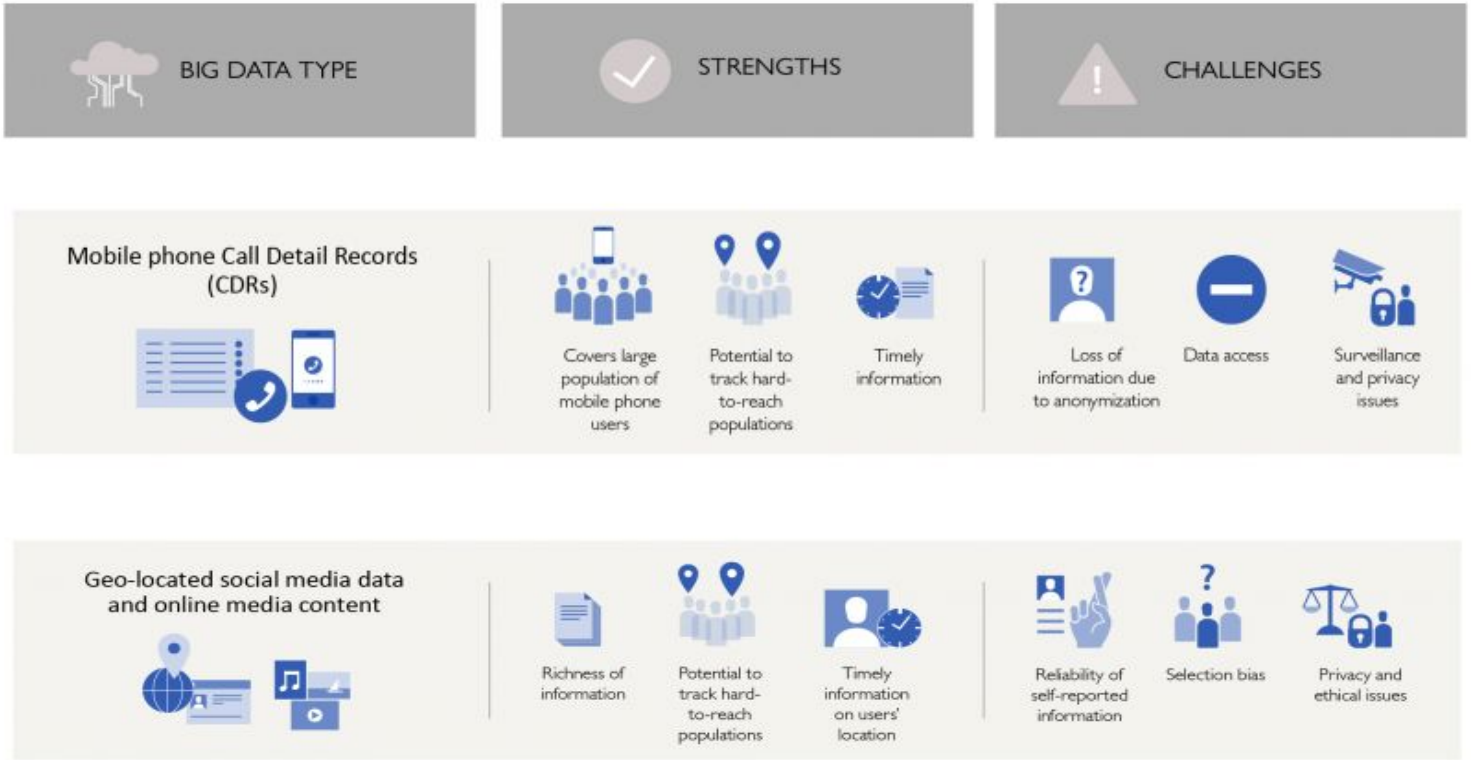
2

S.N.	Date of Accident	A/C Reg. No.	Type of A/C	Operator/Owner	Place of Accident	Fatality	Survival	Remarks
1	27/12/1979	9N RAE	Allutte-III	VVIP	Langtang	6	0	
2	27/04/1993	9N ACK	Bell-206	Himalayan Helicopter	Langtang	0		Closed operation
3	24/01/1996	9N ADM	MI-17	Nepal Airways	Sotang	0	3	Closed operation
4	30/09/1997	9N AEC	AS-350	Karnali Air	Thupten Choling	1	4	Closed operation
5	13/12/1997	9N ADT	MI-17	Gorkha Airlines	Kalikot	0		Closed operation
6	04/01/1998	9N RAL	Bell-206	VVIP Flight	Dipayal			
7	24/10/1998	9N ACY	AS-350B	Asian Airlines	Mul Khark	3	0	Closed operation
8	30/04/1999	9N AEJ	AS-350BA	Karnali Air	Lisunkhu, Sindhupalchowk	0		Closed operation
9	31/05/1999	9N ADI	AS-350B2	Manakamana Airways	Ramechhap	0		Closed operation
10	11/09/2001	9N ADK	MI-17	Air Ananya	Mimi	0	5	Renamed as Shree Airlines
11	12/11/2001	9N AFP	AS-350B	Fishtail Air	Rara Lake, Mugu	4	2	
12	12/05/2002	9N AGE	AS 350B2	Karnali Air	Makalu Base Camp	0	1	Closed operation
13	30/09/2002	9N ACU	MI-17 (MI8-MTV)	Asian Airlines	Sholumkhumbu*	0	None	Closed operation

+ district
+ lat-long
(based on
location)

Credit: http://bit.ly/2K3mXBI

Plug External Data



Credit: http://tiny.cc/xiye8y

Plug External Data

Credit: http://tiny.cc/xiye8y

Feature Engineering

- How GDP and Population can be combined to give GDP Per Capita
- How profit and cost can be combined to calculate revenue

How to Assess a Claim ?

- Primary Data

Advantages	Disadvantages

Credit: http://bit.ly/2Xif2Yj

How to Assess a Claim ?

- Primary Data - Types

Survey

Observation

Credit: http://bit.ly/2Xif2Yj

How to Assess a Claim ?

- Primary Data - Survey Techniques - Comparison

	Remote Survey (Mail, Internet etc)	Telephone Survey	In-Person Survey (Mall Intercept, etc)
Cost	LEAST	MODERATE	MOST
Ability to ask complex questions	LITTLE (Self-administered format must be short and simple)	SOME (Interviewer can probe and elaborate)	MUCH (Interviwer can show visuals, probe, establish rapport)
Opportunity for Interviewer to bias results	NONE (Form is completed without interviewer)	SOME (Because of voice inflection of interviwer)	SIGNIFICANT (Voice and Facial expression of interviewer)
Anonymity for Respondent	COMPLETE (No signature is needed)	SOME (Because of telephone contact)	LITTLE (Because of face-to-face contact)

Credit: http://bit.ly/2Xif2Yj

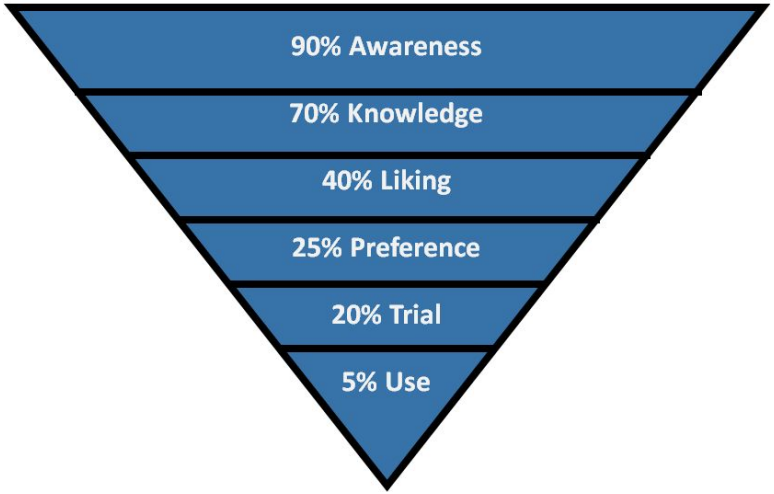
How to Assess a Claim ?

- Bias

Credit: http://bit.ly/2Xif2Yj

How to Assess a Claim ?

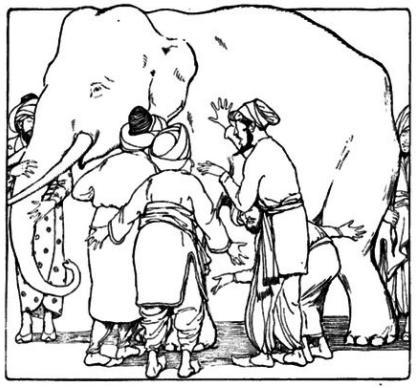
- Carryover Effect
- Practise Effect
- Fatigue Effect



Credit: http://tiny.cc/rhye8y, http://bit.ly/2QWOSnv

Blind Men and the Elephant

- None of them were wrong, but none of them were right either
- One can come across the some problem when one is looking at data



Credit: http://tiny.cc/mkjy7y

How to Assess a Claim ?

- Understand the data - available values
- Understand the data - unavailable or missing values

Credit: http://bit.ly/2XYeALe

Missing Values

- Missing Completely at Random (MCAR)
  – The Good
- Missing At Random (MAR)
- Missing Not At Random (MNAR)

Complete data	
Age	IQ score
25	133
26	121
29	91
30	105
30	110
31	98
44	118
46	93
48	141
51	104
51	116
54	97

Incomplete data	
Age	IQ score
25	
26	121
29	91
30	
30	110
31	
44	118
46	93
48	
51	
51	116
54	

Credit: http://tiny.cc/6ewe8y

Missing Values

- Missing Completely at Random (MCAR)
- Missing At Random (MAR) – The Bad
- Missing Not At Random (MNAR)

Complete data	
Age	IQ score
25	133
26	121
29	91
30	105
30	110
31	98
44	118
46	93
48	141
51	104
51	116
54	97

Incomplete data	
Age	IQ score
25	
26	
29	
30	
30	
31	
44	118
46	93
48	141
51	104
51	116
54	97

Credit: http://tiny.cc/6ewe8y

Missing Values

- Missing Completely at Random (MCAR)
- Missing At Random (MAR)
- Missing Not At Random (MNAR) – The Ugly

Complete data	
Age	IQ score
25	133
26	121
29	91
30	105
30	110
31	98
44	118
46	93
48	141
51	104
51	116
54	97

Incomplete data	
Age	IQ score
25	133
26	121
29	
30	
30	110
31	
44	118
46	
48	141
51	
51	116
54	

Credit: http://tiny.cc/6ewe8y, http://tiny.cc/lhwe8y

How to Assess a Claim ?

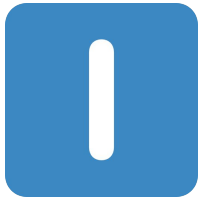
- Understand the data - available values
- Understand the data - unavailable or missing values
- Understand the data - relationships

Credit: http://bit.ly/2XYeALe

$$y = m x + c$$

Independent vs Dependent Variable

- Intentionally manipulated
- Controlled
- Vary at known rate
- Cause



- Intentionally left alone
- Measured
- Vary at unknown rate
- Effect

Credit: http://tiny.cc/y5xe8y, http://tiny.cc/0v8x7y

Control Variable

Factors or conditions that are kept the same (unchanged) in an experiment

Credit: http://bit.ly/2MSzESI

Examples

“ Does changing the color of light ” affect the growth rate of plants?

I

D

Credit: http://tiny.cc/w6xe8y, http://bit.ly/2Rik1ly

Examples

“ Does changing the color of light affect the growth rate of plants? ”

I

D

Credit: http://tiny.cc/w6xe8y, http://bit.ly/2Rik1ly

Examples

“ Does changing the color of light affect the growth rate of plants? ”

- Same type / size of plant
- Same amount of water
- Same soil
- Same exposure to light
- etc…

Credit: http://tiny.cc/w6xe8y, http://bit.ly/2Rik1ly

Examples

“ Does the size of a parachute affect the time it takes a hippo to free fall to the ground? ”

I

D

Credit: http://tiny.cc/w6xe8y, http://bit.ly/2MSzESI

Examples

“ Does the size of a parachute affect the time it takes a hippo to free fall to the ground? ”

I

D

Credit: http://tiny.cc/w6xe8y, http://bit.ly/2MSzESI

Examples

“ Does the size of a parachute affect the time it takes a hippo to free fall to the ground? ”

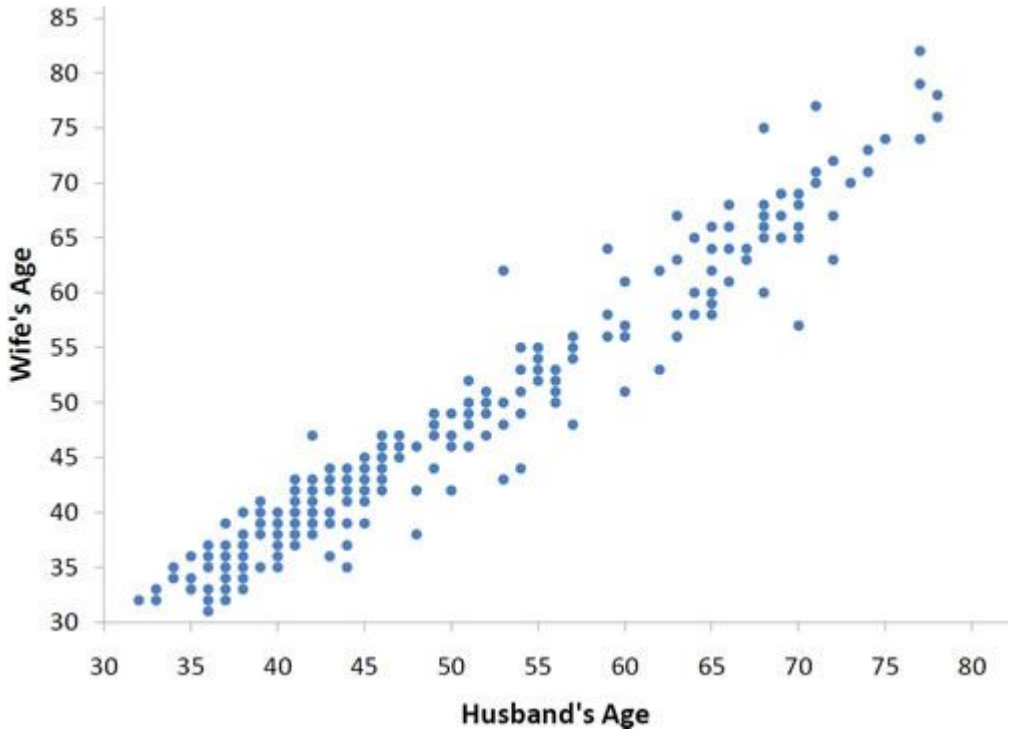
- Same hippo
- Same drop height
- Same parachute fabric
- Same length of strings
- etc…

Credit: http://tiny.cc/w6xe8y, http://bit.ly/2MSzESI

Correlation vs Causation

- Correlation -

  A mutual connection or relationship between two or more things



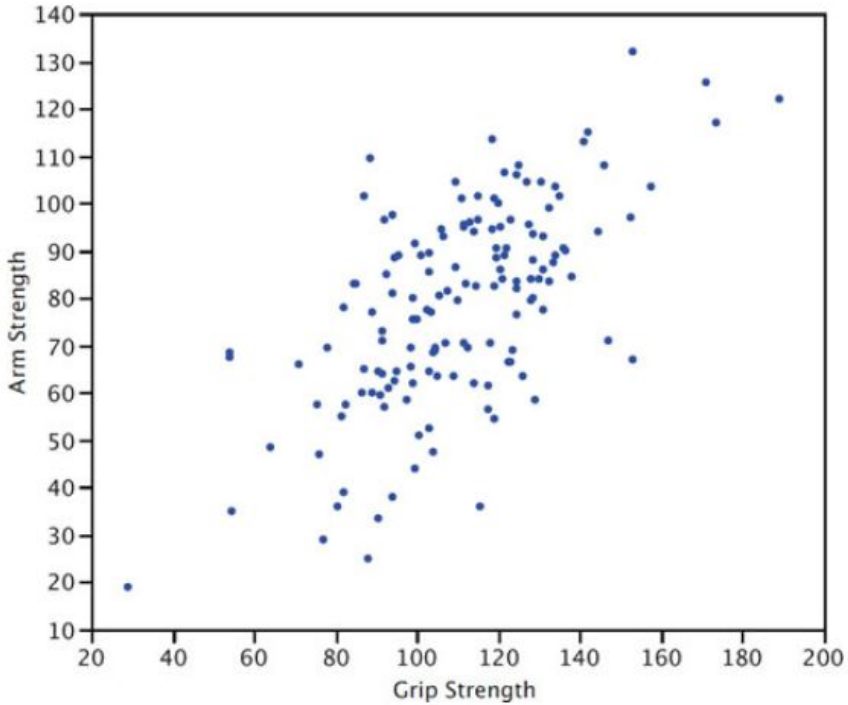
Relationship is almost LINEAR

Credit: http://bit.ly/2MSb8Re

Correlation vs Causation

- Correlation -

  A mutual connection or relationship between two or more things

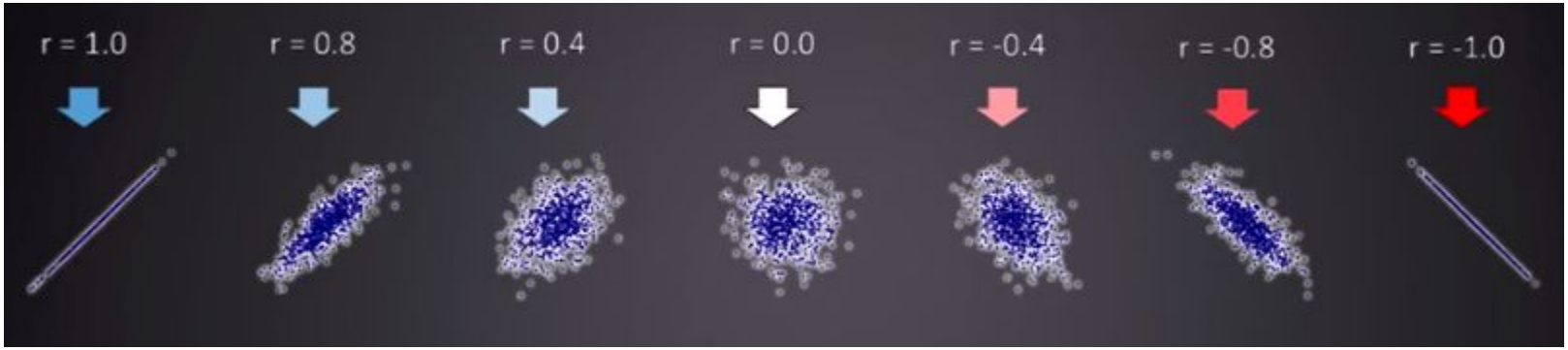


Credit: http://bit.ly/2MSb8Re

Correlation vs Causation

- Correlation Coefficient (r)

  Degree to which there is a linear relationship between the variables

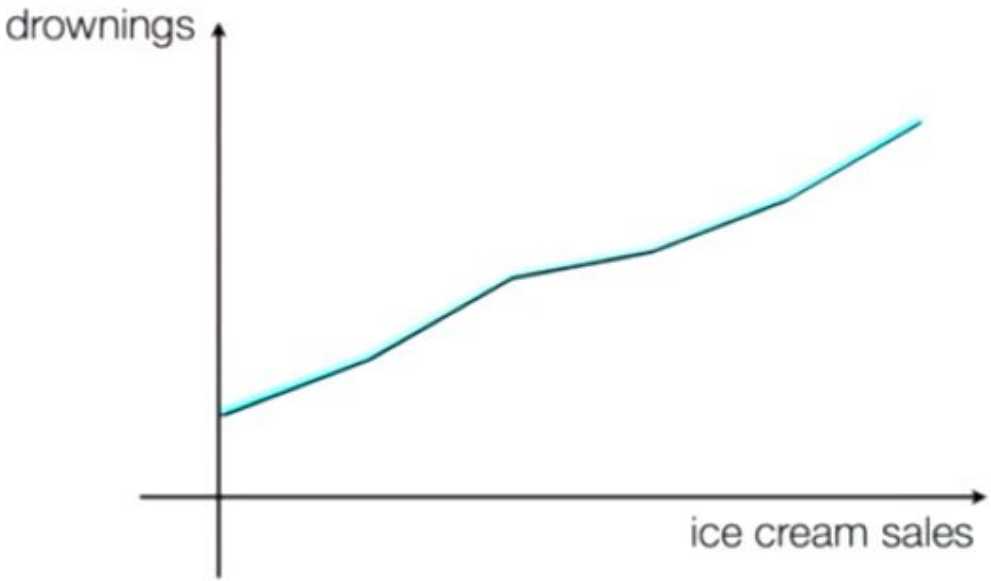


Credit: http://bit.ly/2MSb8Re

Correlation vs Causation

- Causation

  One event or state is the result of the occurrence of another event or state

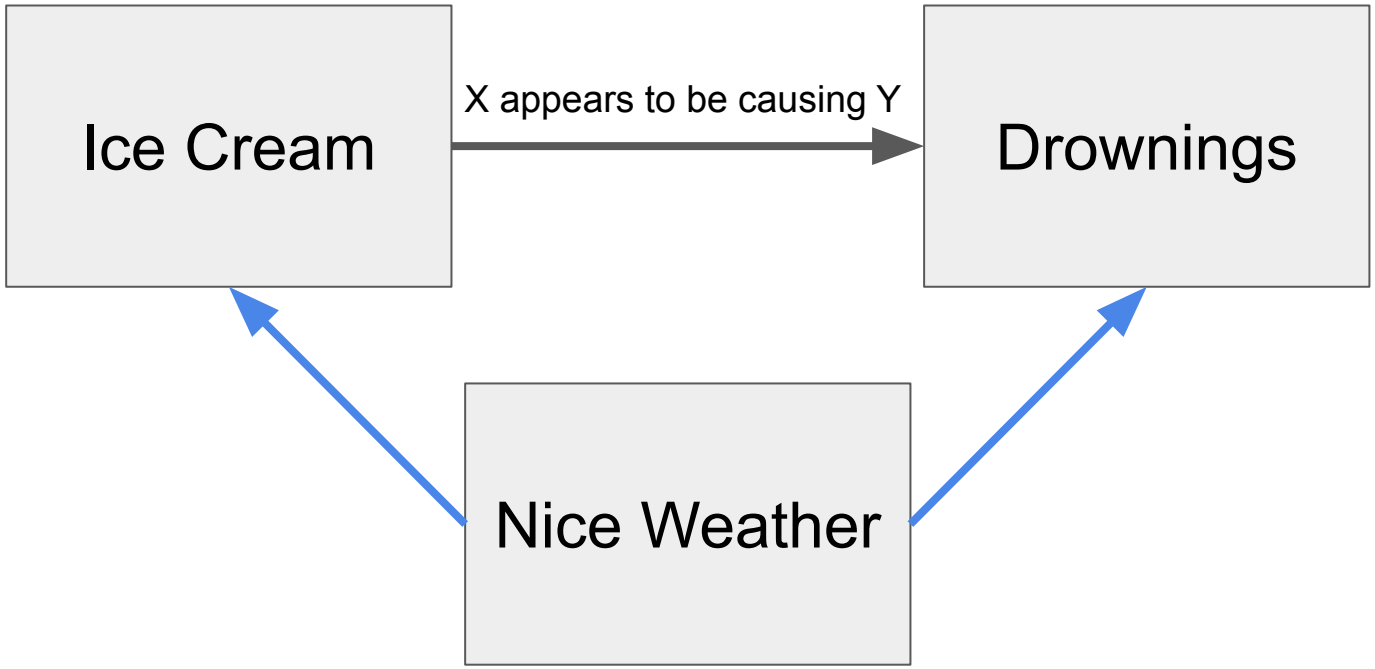


Credit: http://bit.ly/2NqFcUg

Correlation vs Causation

- Causation

  One event or state is the result of the occurrence of another event or state

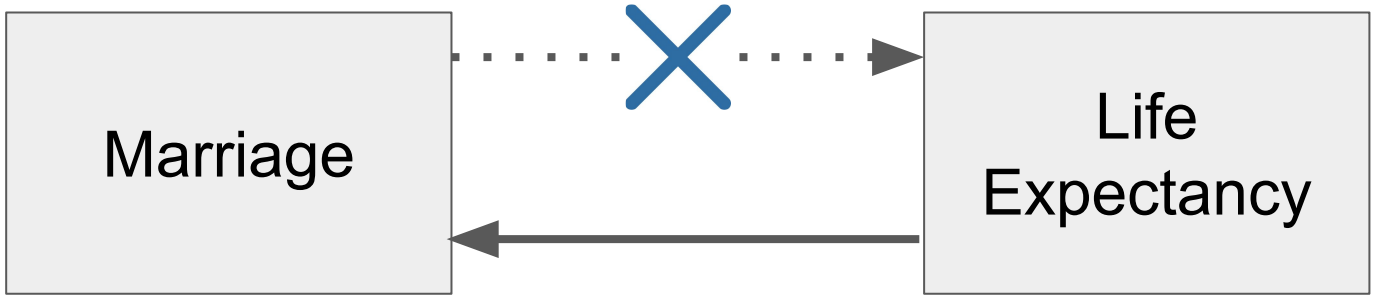


Credit: http://bit.ly/2NqFcUg

Correlation vs Causation

- Causation

  One event or state is the result of the occurrence of another event or state

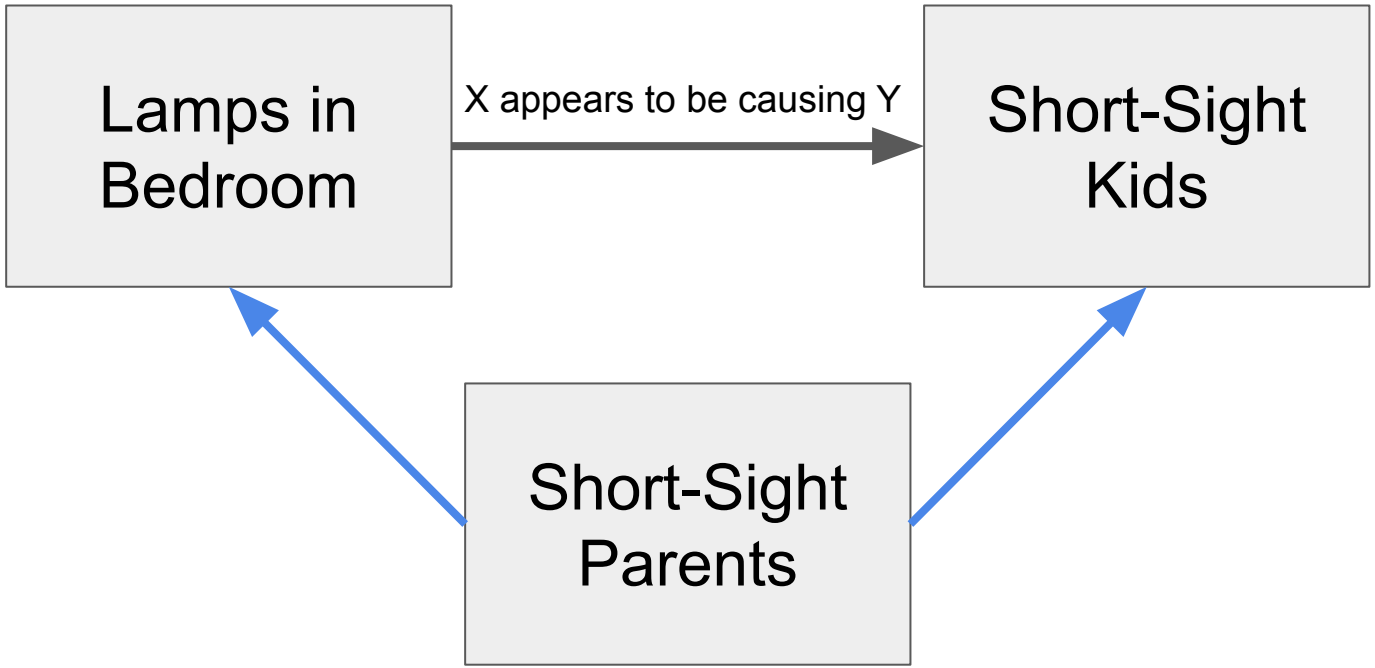


Credit: http://bit.ly/2NqFcUg

Correlation vs Causation

- Causation

  One event or state is the result of the occurrence of another event or state

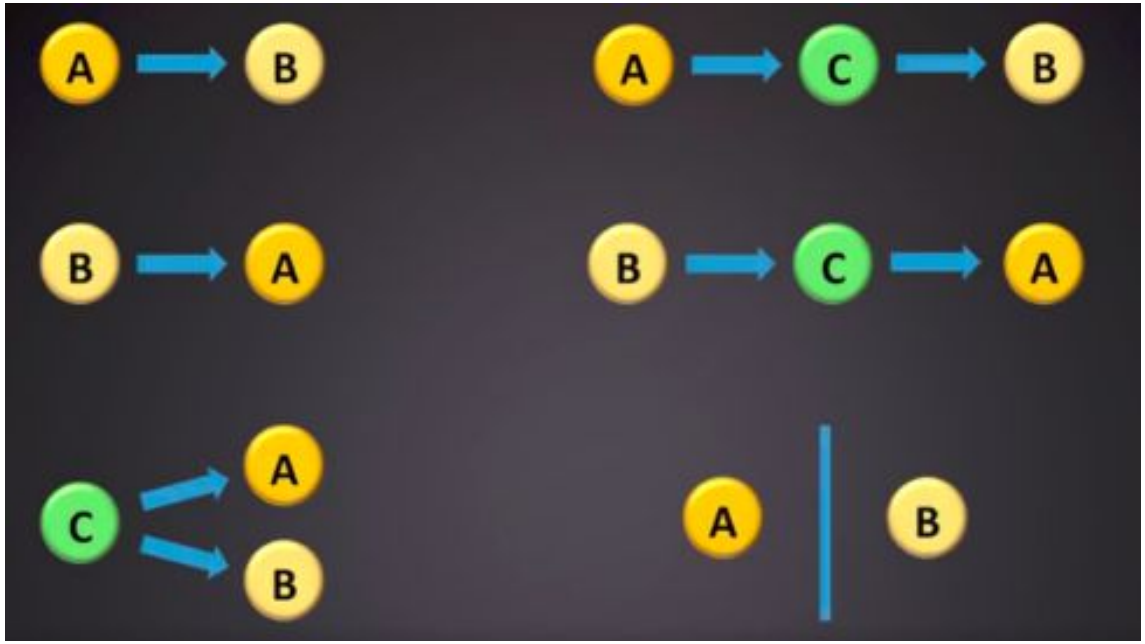


Credit: http://bit.ly/2NqFcUg

Correlation vs Causation

- Causation

  One event or state is the result of the occurrence of another event or state



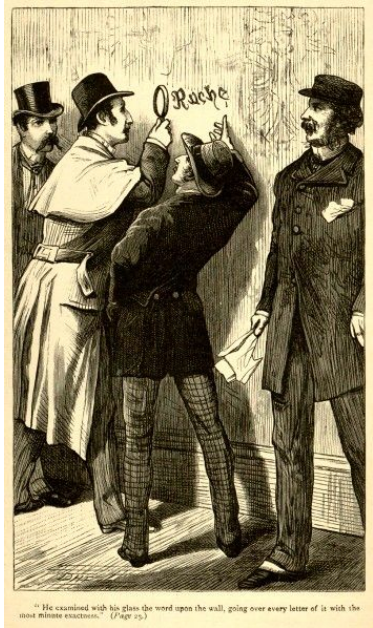
Credit: http://bit.ly/2MSb8Re

How to Assess a Claim ?

- Understand the data - available values
- Understand the data - unavailable or missing values
- Understand the data - relationships
- Understand the data - visualizations

Red Herring

- Something that misleads or distracts from a relevant or important question
- May be intentional, or unintentional



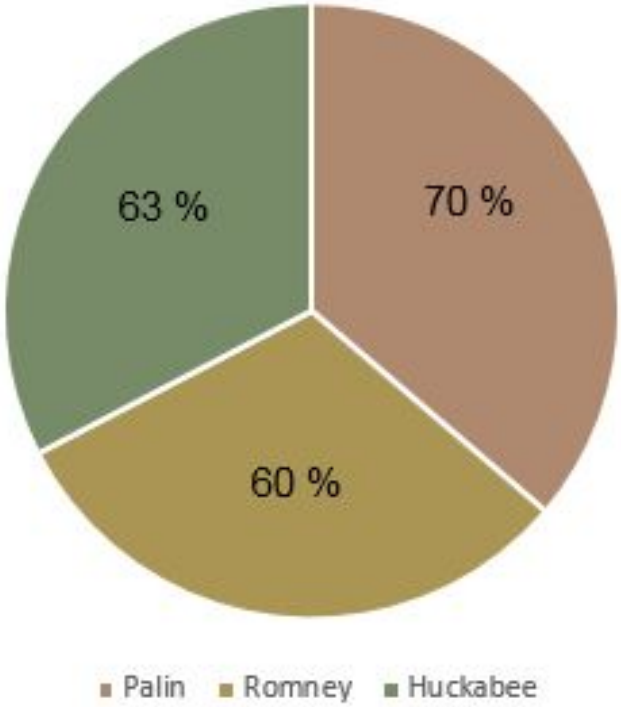
"He examined with his glass the word upon the wall, going over every letter of it with the most minute exactness." (Page 25.)

Credit: http://tiny.cc/v9xe8y, http://tiny.cc/9a9x7y, http://tiny.cc/tb9x7y

Enlighten, Confuse or Deceive ?

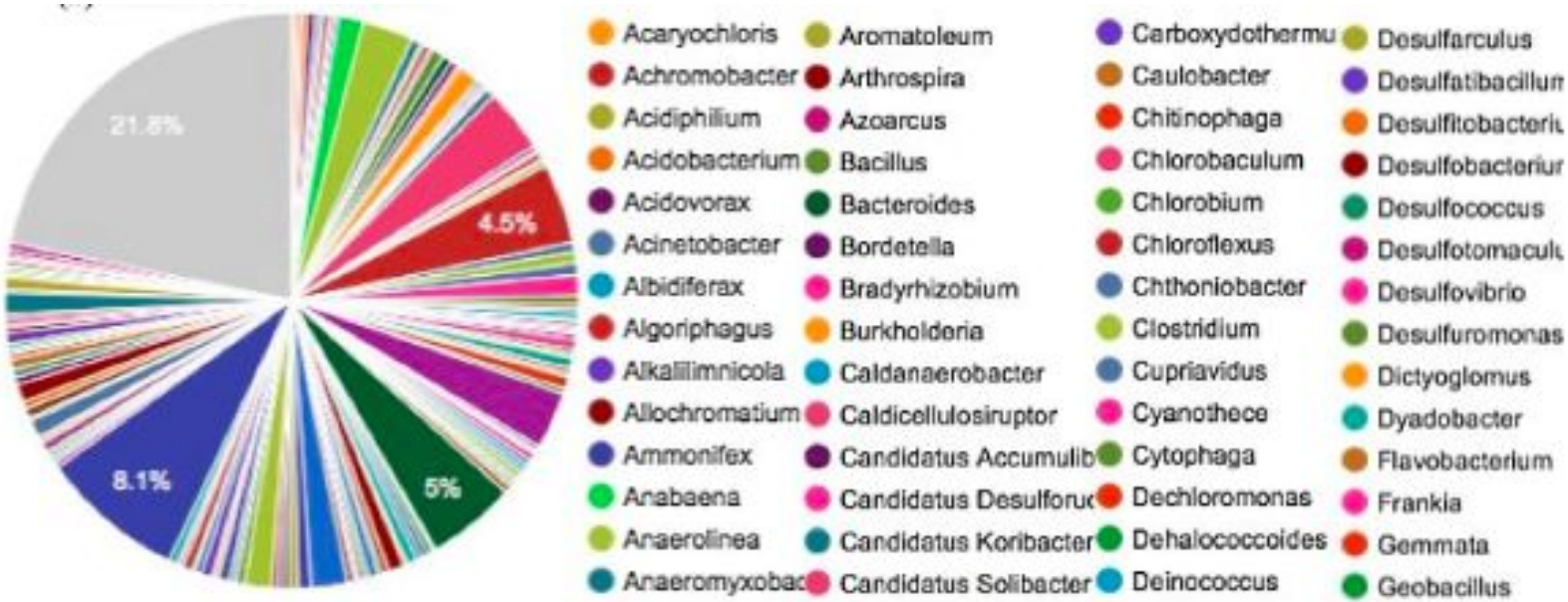
- Errors
- Overwhelming
- Misleading

Errors



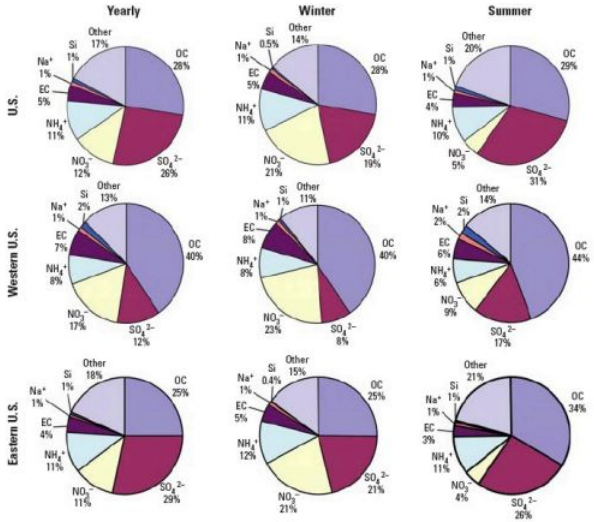
Credit: http://bit.ly/2IolBQ7

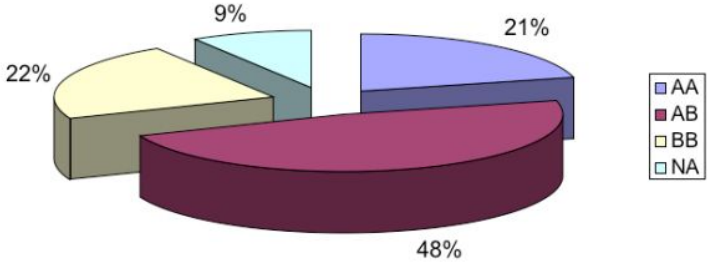
Overwhelming



Credit: http://tiny.cc/fkdy7y

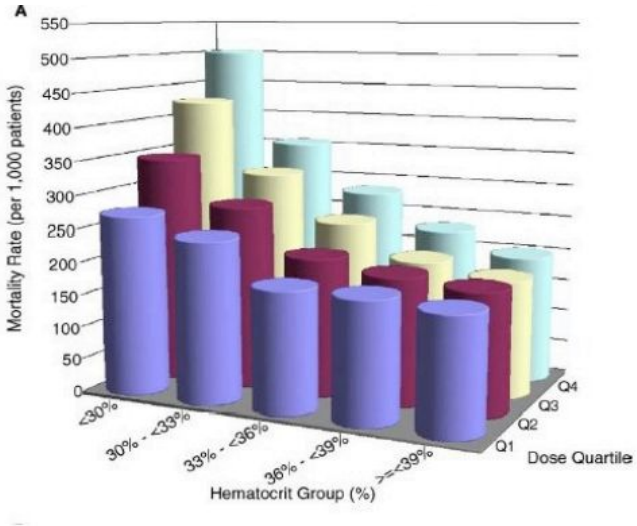
Overwhelming

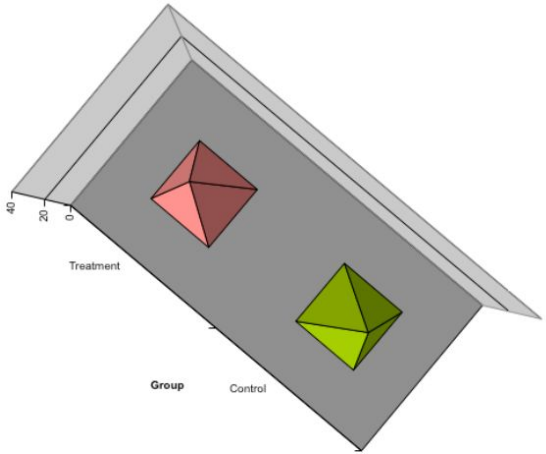




Credit: http://bit.ly/2ZwNsn6

Overwhelming

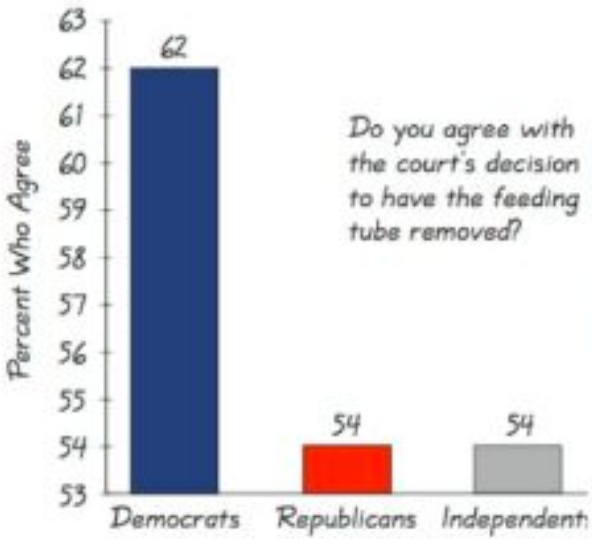


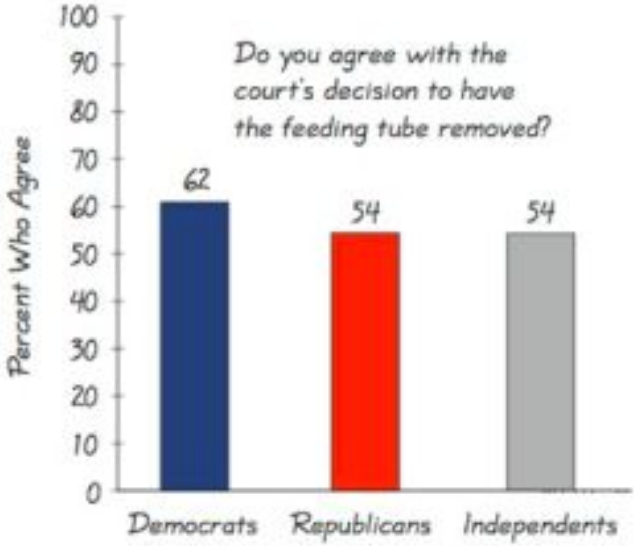


Credit: http://bit.ly/2ZwNsn6

Misleading

- Non-Zero Axis

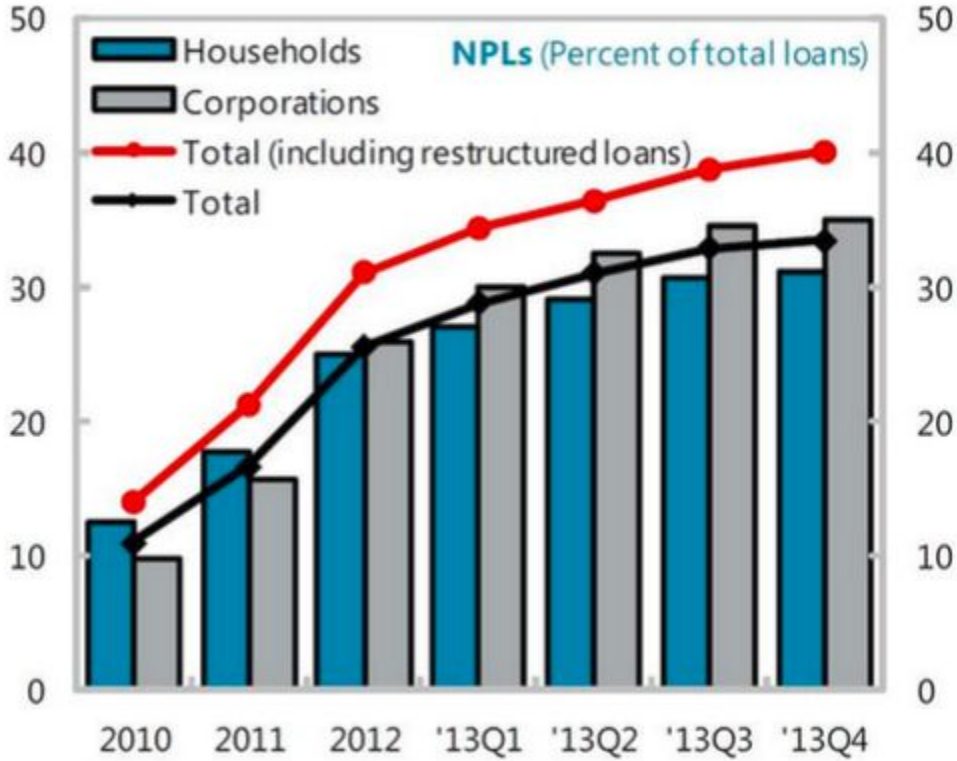




Credit: http://bit.ly/2x01GjQ

Misleading

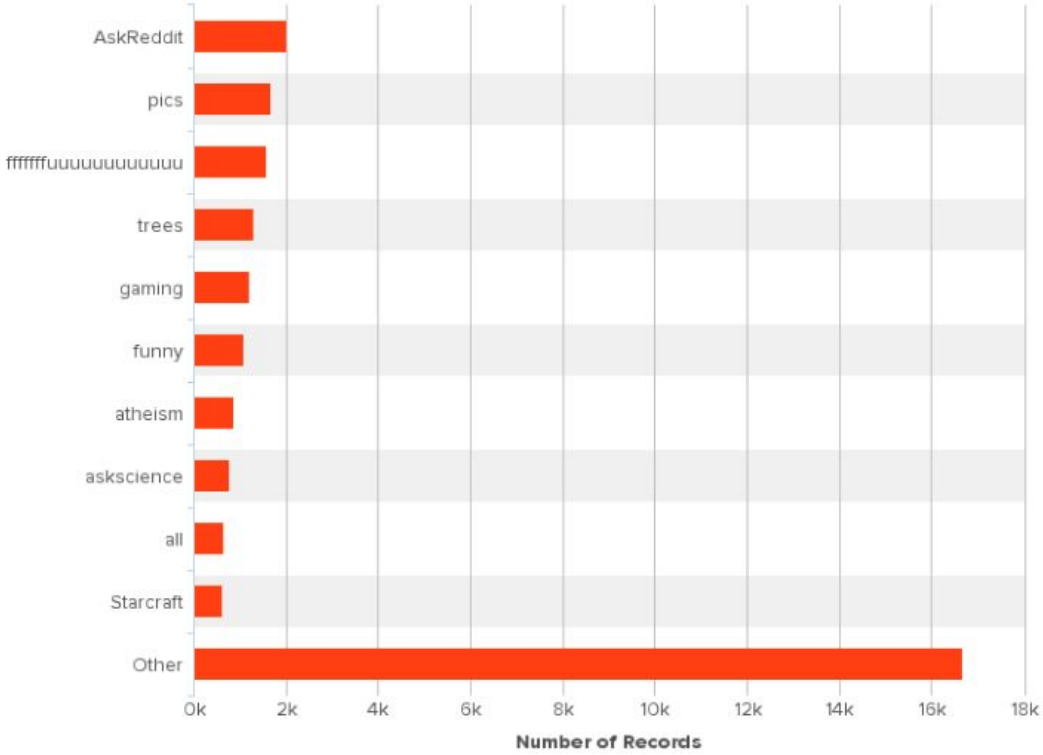
- Inconsistent Scales



Credit: http://tiny.cc/fkdy7y

Misleading

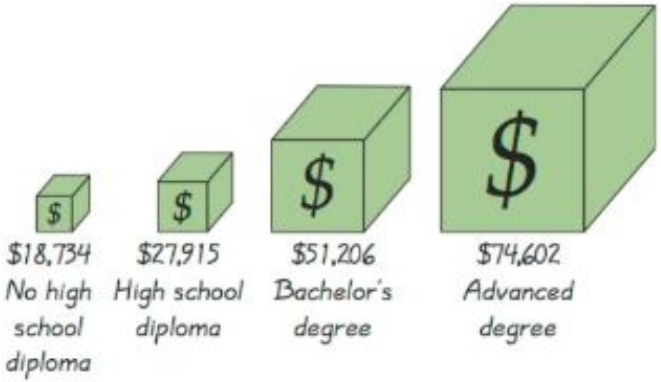
- “Others”



Credit: http://bit.ly/2KYkaJK

Misleading

- Pictographs



Credit: http://bit.ly/2x01GjQ

Thank You

Credit: http://bit.ly/2N4wniG