The Devil is in the Data

{

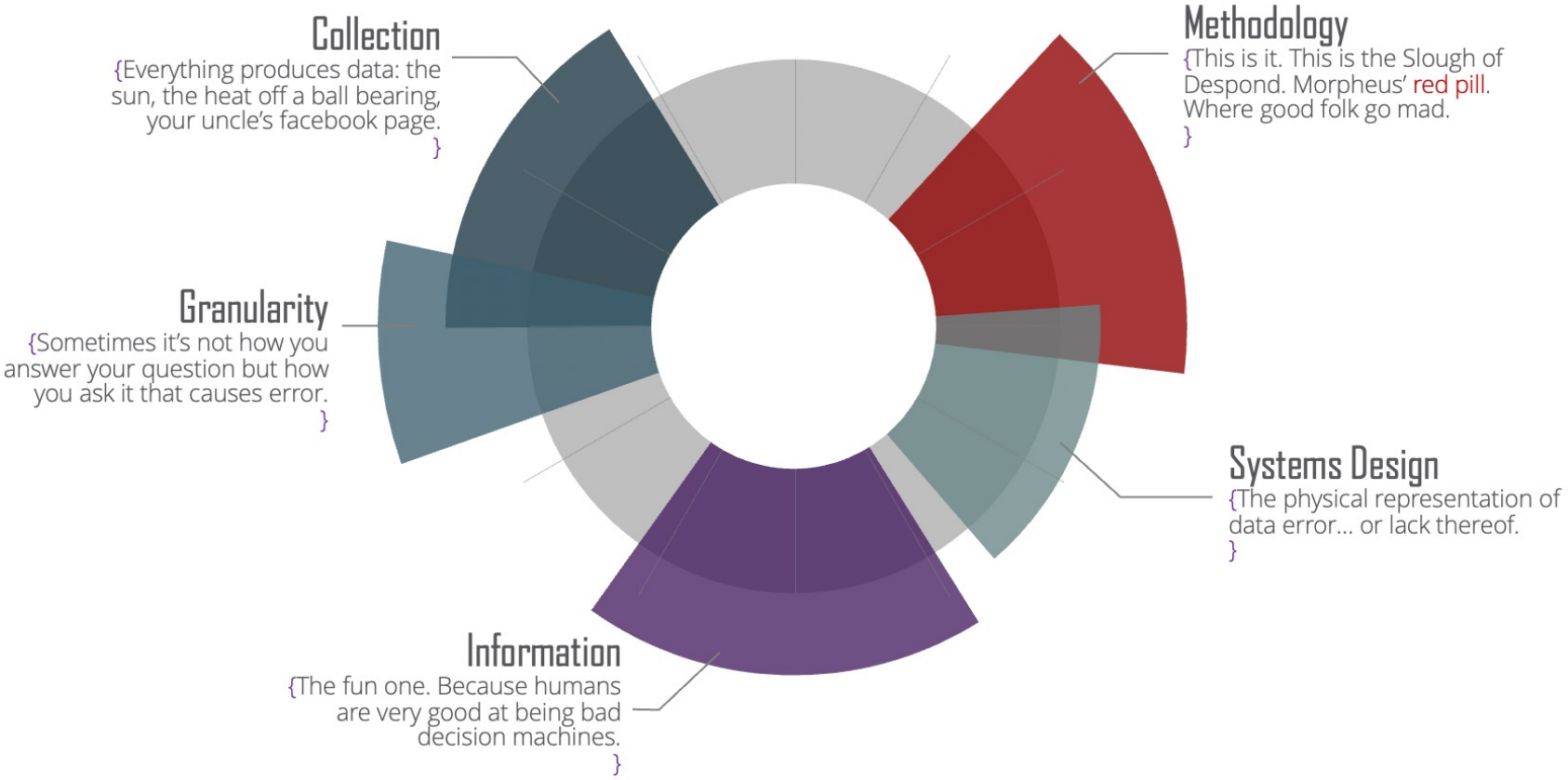
[**KU-SOB**, 8 April 2022

EduardoErazo

]

}

Sources of Error



Human
Collection

Priming:
{Psychological bias produced by exposure to a stimulus that can influence your decisions
}

Something else:
> Likert Scales

Q: How happy are you with this presentation?

A: Pretty Happy → Satisfied → MEGA-LOATHED IT!!

	Treatment A	Treatment B
Small Stones	*93% (81/87)*	87% (234/270)
Large Stones	*73% (192/263)*	69% (55/80)
Total	78% (273/350)	*83% (289/350)*

Br Med J (Clin Res Ed) 1986;292:879

Muddy Data

Granularity

Case Study:
CLAMAZONE.COM

By the #:
{400k emp
25% YoY Personnel Increase
20% Yearly Turnover
}

The ask:
Can you help us fix retention?

Muddy Data

This Error:

{Wayyy more common than you think.

This is closely related to “operations research.”
}

[Reference: Malcolm Gladwell & Howard Moskowitz]

Chopping Data

{Categorizing continuous scales.

[High Performers = 8-10 sales/day
Mid Performers = 4-7 sales/day
Low Performers = 1-3 sales/day
]

}

$$\Delta \text{ of “low mid” and “high low”} < \Delta \text{ of “low mid” and “high mid”}$$

QRM_ Research Methodology

DATA

Number → Contain Information

...some more than others.

4th Place

Finish: 32min

Correlation

Pearson R: Interval or Ratio
Spearman's Rho: Ordinal
Kendall's Tau: Ordinal (& non-parametric)

OR

Somer's Delta: Ordinal but between independent and dependent
Goodman & Kushkal's Gama: Ordinal, less error when ties
Shepherd's Pi: Trimmed ordinal
Biweight Midcorrelations: Don't trim, leave outliers
Polychorics: Use with latent variables (tetra- if dichotomous)

QRM: Assumptions_ Methodology

>>Chi Square: non-parametric statistical test to compare groups

(where letter data ...becomes number data!
..and grown-ups cry)

{Assumption 1:

[Independence |Back to COLLECTION|]

Assumption 2:

[Deviation distributions of Observed x Expected in normal form]

Check: Degrees of Freedom

|Cochran (1954): ef > 1 & < 20% of ef > 5|

or

|Camili & Hopkins (1978): sample > 20 for (2x2 table) with E ≥ 1 ...or E≥2, in ≤ 2 cells|

or

|Wickens (1989): n>count of cells x4 ...or 5|

}

QRM: Assumptions_ Methodology

{Example 2: ANOVA

[-needs homoscedasticity. So run Levene's test. If that fails, maybe try Brown-Forsyth. If THAT fails, lower your alpha by .025, but only if results are moderate (Keppel, 1991). What is a "moderate" figure again? But, also *SHOULD* run a post hoc like a Tukey, or Bonferroni... which really depends on your scales + collection]

}

{Example 3: Principal Component Analysis_ Factor Reduction

[-When running a principal components analysis and you're trying to do a factor reduction. You should request orthogonal rotation when you suspect that the data are related. Maybe prove that with a correlation? Idk. Maybe just think through it and then you can run it both ways on the rotation and see what the eigen values do. But also eigenvalues are unreliable so then maybe also request a scree plot. And wait what KIND of orthogonal rotation did you use again? Promax or varimax? Becausifyouusedpromax]

}

Systems Design_ Infrastructure

ETL: Many many many pieces of code all working together to move data around.

{"Data are missing!"

[What if...

An ETL script fails from a source system?

>Let's make sure the server is running.

Wait... what if the ETL didn't fail but we had a bottleneck?

>Resize the server for more bandwidth.

Or maybe there was a character it didn't recognize?

>Add a filter, or convert all, or exception.

Is your SCD strategy appropriate?

>Check that you have Type 0,1,2 replacement. (or 6?)

Maybe someone typo-ed?

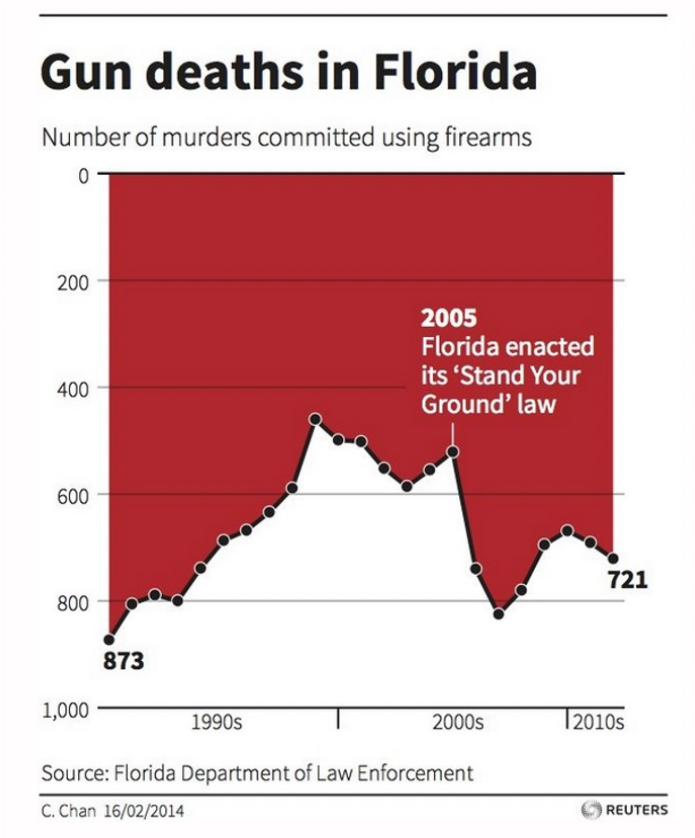
>Make sure we have data validation or micro-corrections.

Wait! No! No, it was it a holiday? Ahhh,yeah! Phew!

]

}


Information
Granularity
Collection
Methodology
Systems
Design


Gun deaths in Florida
Number of murders committed using firearms
0
200
400
600
800
1,000
2005
Florida enacted its 'Stand Your Ground' law
873
721
1990s
2000s
2010s
Source: Florida Department of Law Enforcement
C. Chan 16/02/2014
REUTERS

{There are lies,
damned lies,
and statistics.
}

[Mark Twain]