

Abstract geometric lines in the top-left corner of the page, consisting of several thin, black, overlapping lines that form a complex, non-representational shape.

INTRODUCTION TO INTELLECTUAL DATA ANALYSIS

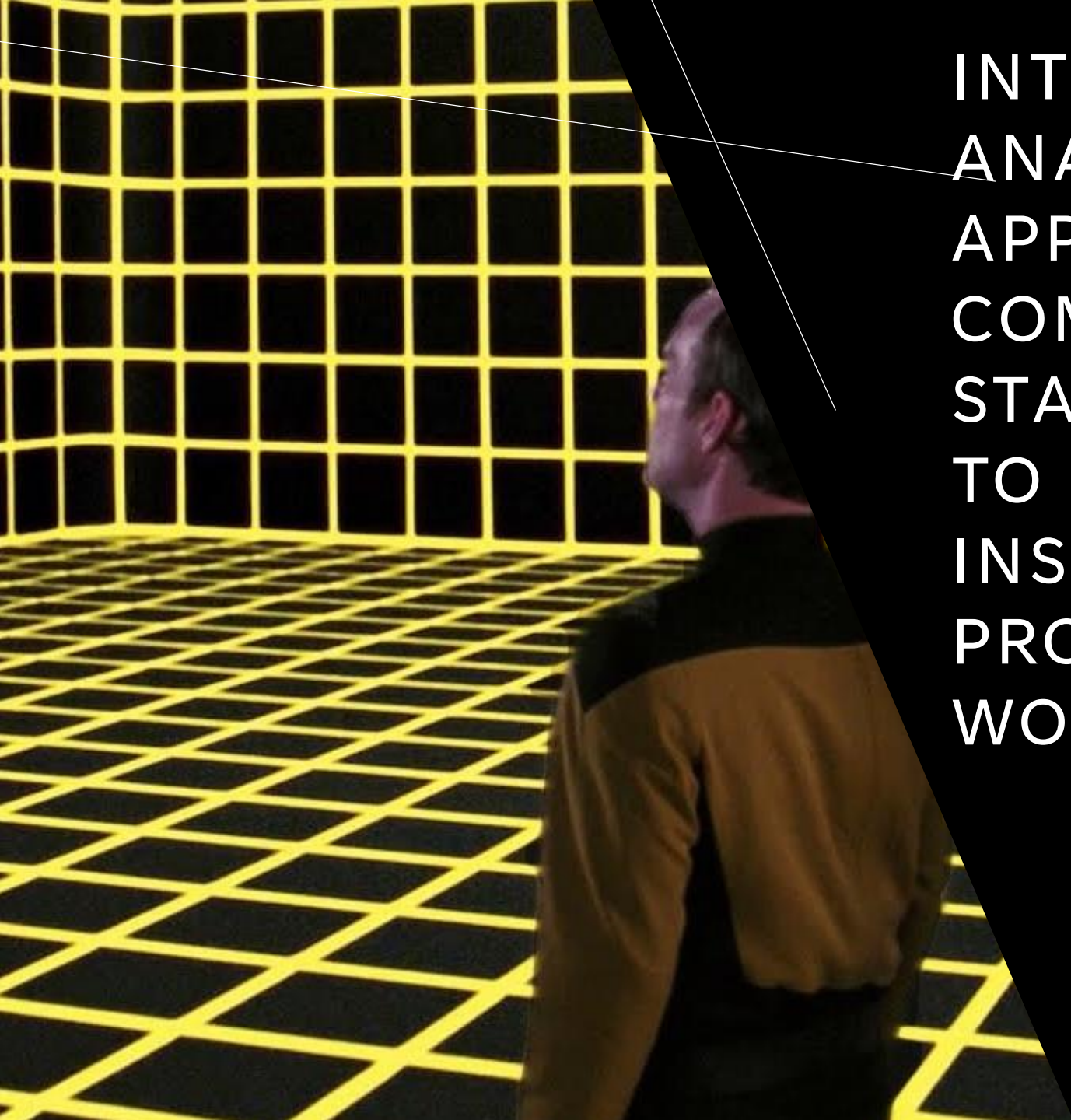
AGENDA

What is Data Analysis ?


What is Data Analysis not
?

Some examples

Conclusion



INTELLECTUAL DATA
ANALYSIS IS THE
APPLICATION OF
COMPUTATIONAL AND
STATISTICAL TECHNIQUES
TO ADDRESS OR GAIN
INSIGHT INTO SOME
PROBLEM IN THE REAL
WORLD.

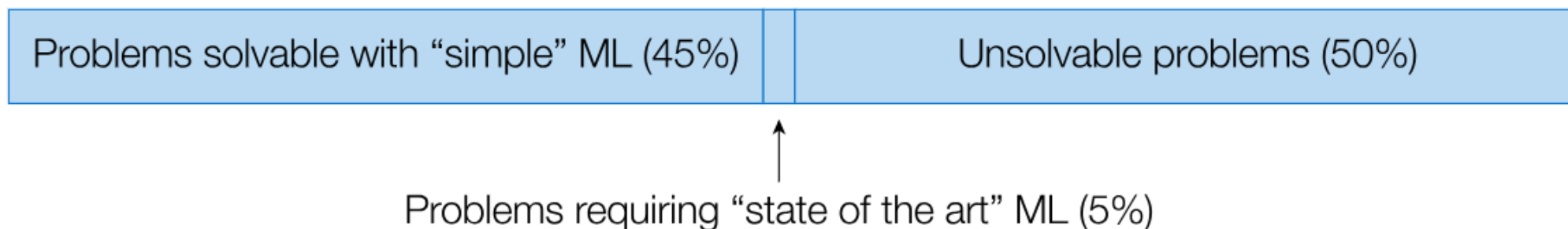
- 
- Data science = statistics + data collection + data preprocessing + machine learning + visualization + business insights + scientific hypotheses + big data + (etc)
 - While data analysts and data scientists both work with data, the main difference lies in what they do with it.
 - Data analysts examine large datasets to identify trends, develop charts, and create visual presentations that help businesses make more strategic decisions.
 - Data scientists, on the other hand, design and construct new processes for data modeling and production using prototypes, algorithms, predictive models, and custom analysis.



IT IS NOT (JUST)
ML

- Making good machine learning based predictions can be an important part of data science, but the truly hard elements of data science involve also involve collecting the data to begin with, defining the problem you're trying to solve (and frequently, re-defining it many times based upon improved understanding of the problem over time), and then interpreting and understanding the results, and knowing what actions to take based upon this. Kaggle contests decidedly do not address almost any of these points.
- The reality is that for many data science problems, simple machine learning algorithms suffice to attain sufficiently good performance.

Universe of machine learning problems

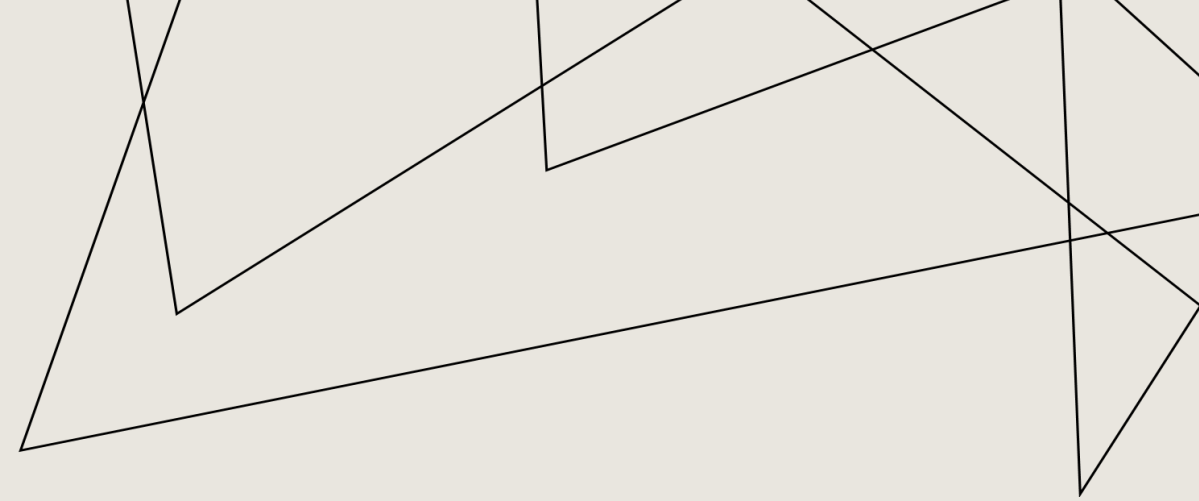


A (rhetorical) breakdown of the universe of machine learning problems



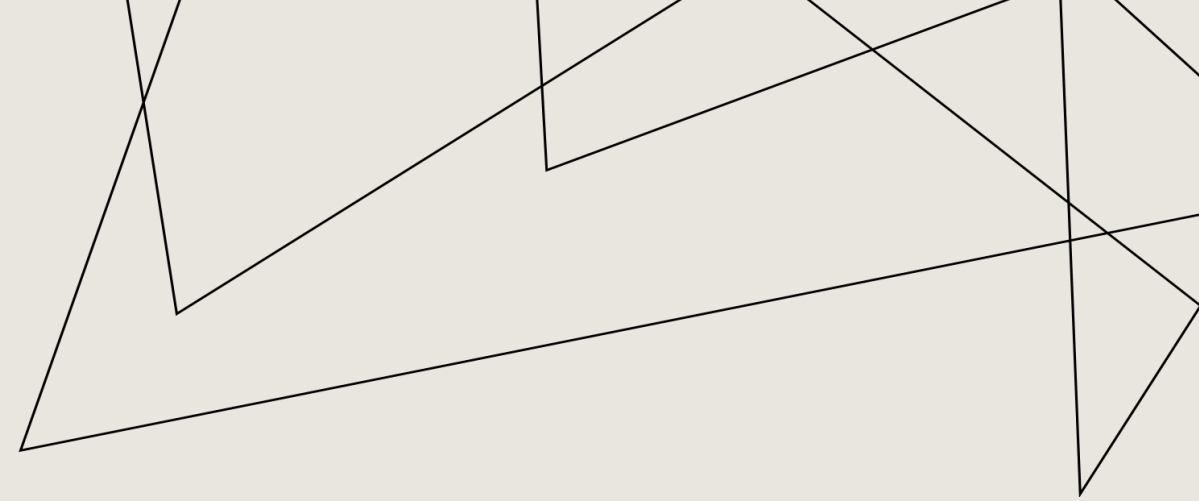
IT IS NOT (JUST)
STATISTICS

- Historically, the academic field of statistics has tended more towards the theoretical aspects of data analysis than the practical aspects.
- From a historical context, data science has evolved from computer science as much as it has from statistics: topics like data scraping, and data processing more generally, are core to data science, typically are steeped more in the historical context of computer science, and are unlikely to appear in many statistics courses.

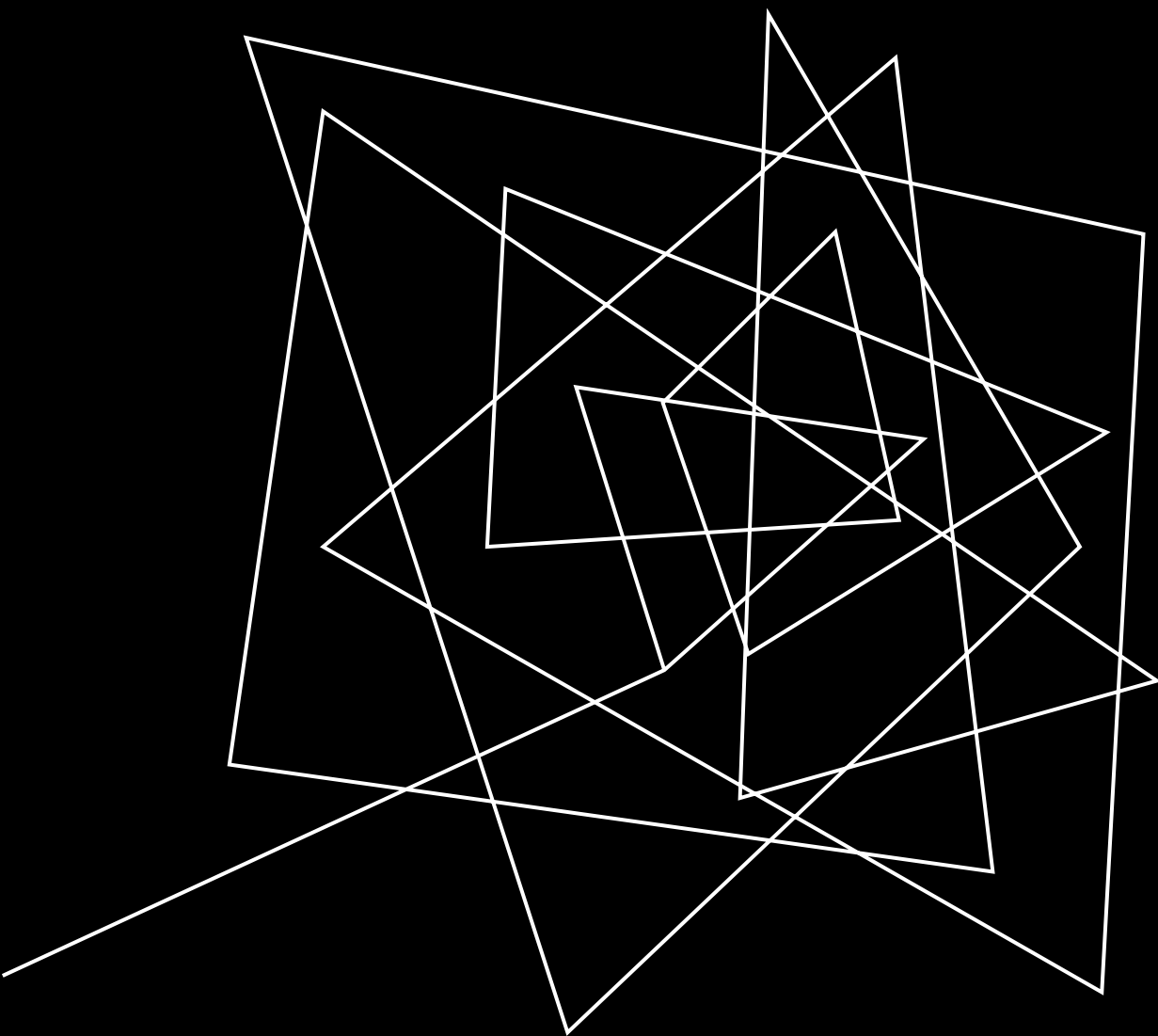




IT IS NOT (JUST)
BIG DATA



- Lastly, there is still a contingent that equates data science with the rise of big data. And while it's absolutely true that some data science work really does use vast amounts of data to build models or gain insights, this is frankly the exception rather than the rule. Most data science can work just fine using the (right set of) data that still fits into memory on a single machine. It's useful to know the techniques needed to address big data challenges, but don't create more work for yourself if you don't have to.



SOME EXAMPLES

EXAMPLE 1

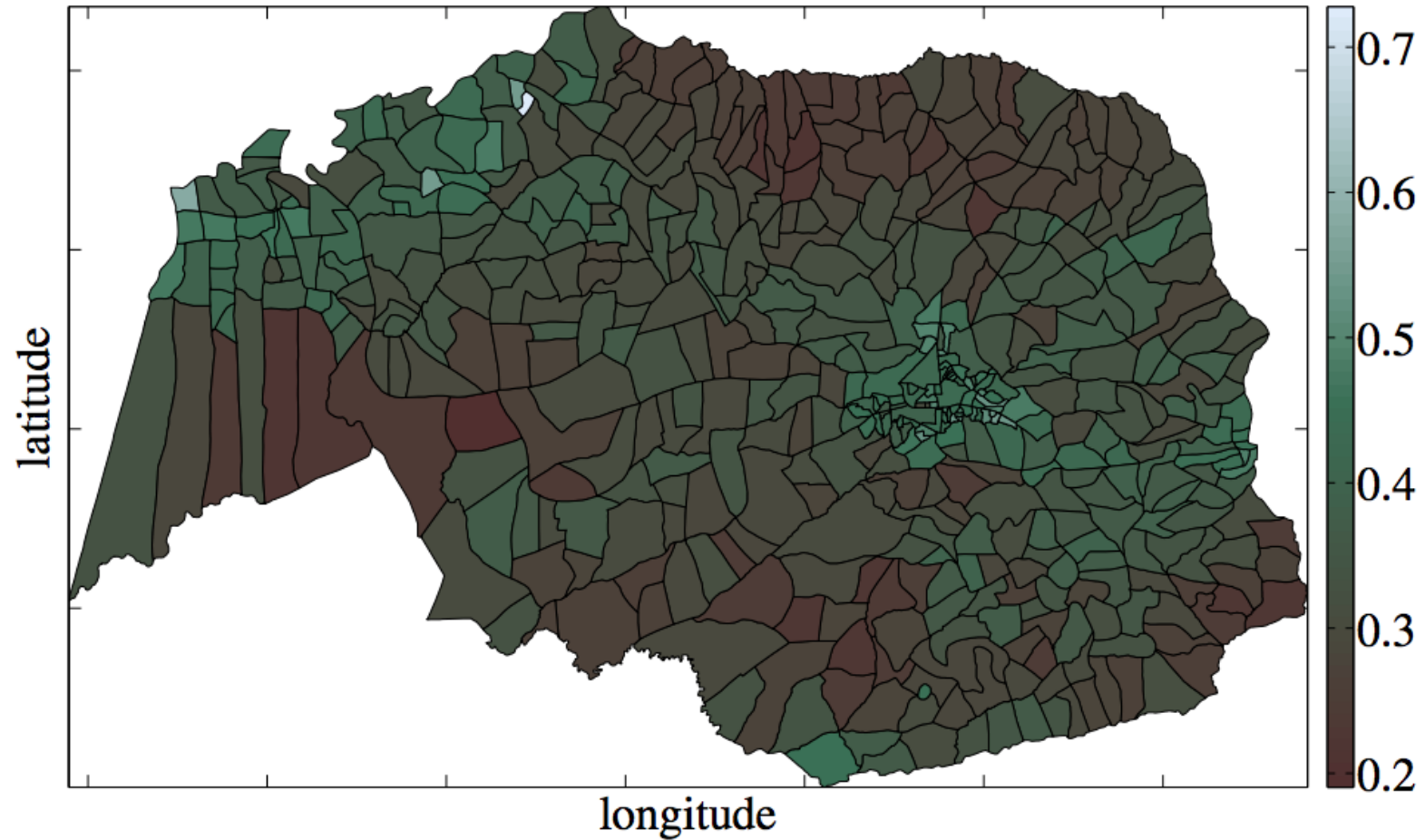
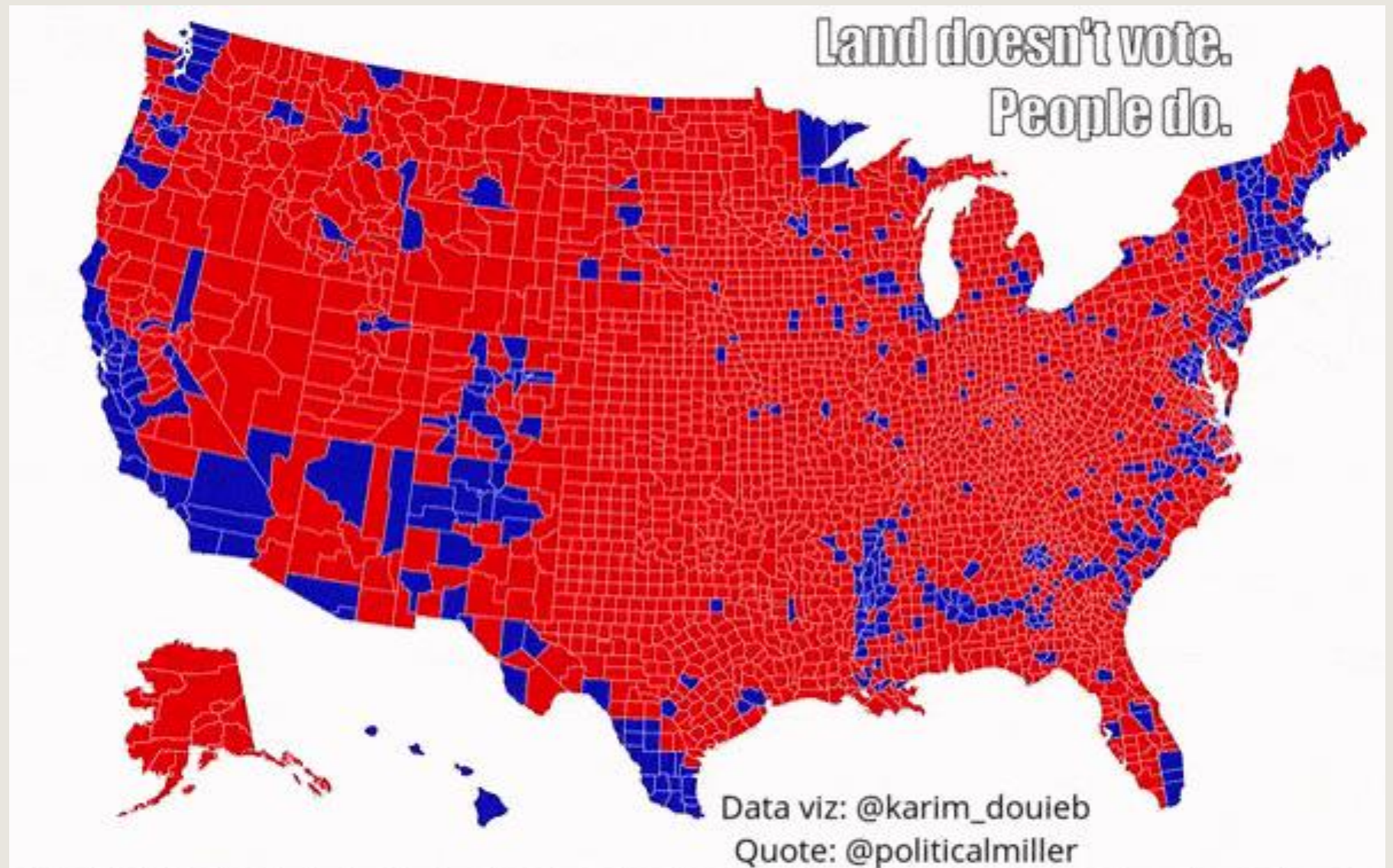
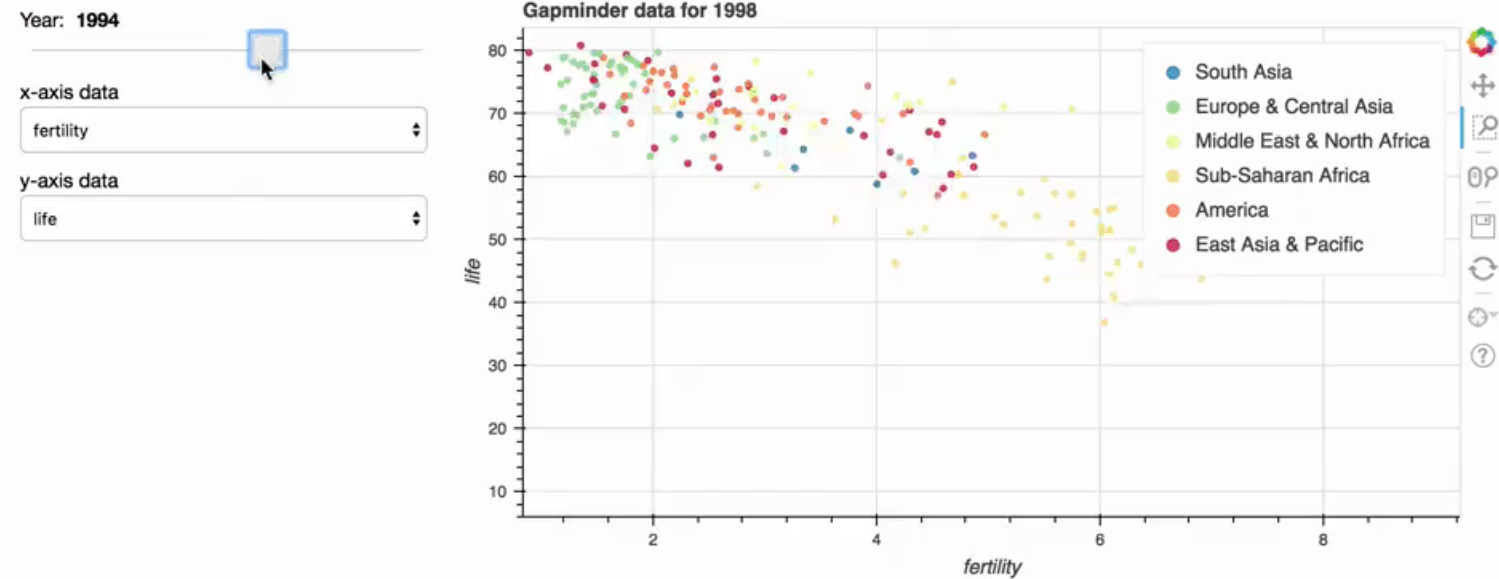


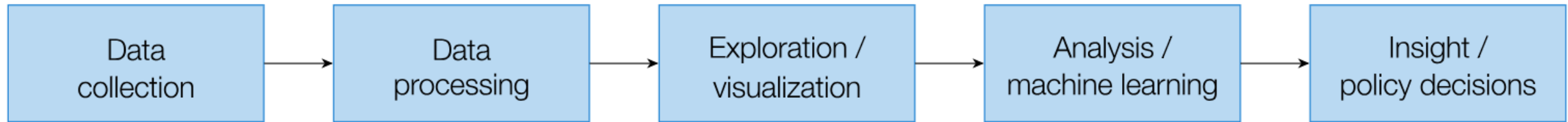
Figure 12: Estimated proportion of metal roofs in villages in the region of interest.

EXAMPLE 2

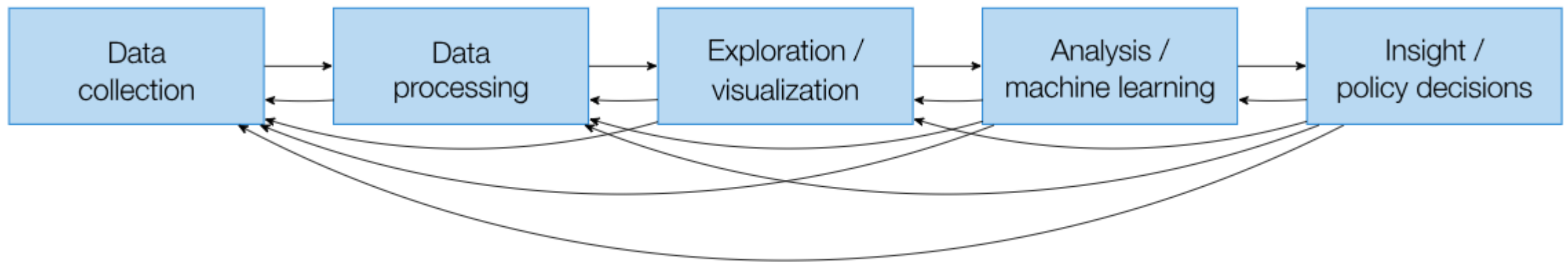


EXAMPLE 3





The data science pipeline



A more accurate data science pipeline



Q&A

A series of white, thin, overlapping geometric lines on a black background, forming a complex, abstract shape on the left side of the slide.

THANK YOU