

The background of the image is a dark, cluttered workshop. It features various tools, equipment, and materials, including what appears to be a workbench, a vise, and some hanging tools. The lighting is dim, creating a moody and industrial atmosphere. The text is overlaid on this background.

Intro to Data Science and Machine Learning

@amitkaps | @bargava

A person is holding a lit sparkler, with bright sparks emanating from the tip. The person's face is partially visible in the background, and they are wearing a dark, textured garment. The overall scene is dimly lit, with the primary light source being the sparkler.

Welcome

A black and white composite image featuring two men. The left side shows a man with short dark hair, smiling broadly, wearing a dark shirt. The right side shows a man with short dark hair and glasses, wearing a patterned shirt and a light-colored shawl draped over his shoulders. The word "Facilitators" is centered in a large, bold, white sans-serif font across the middle of the image.

Facilitators



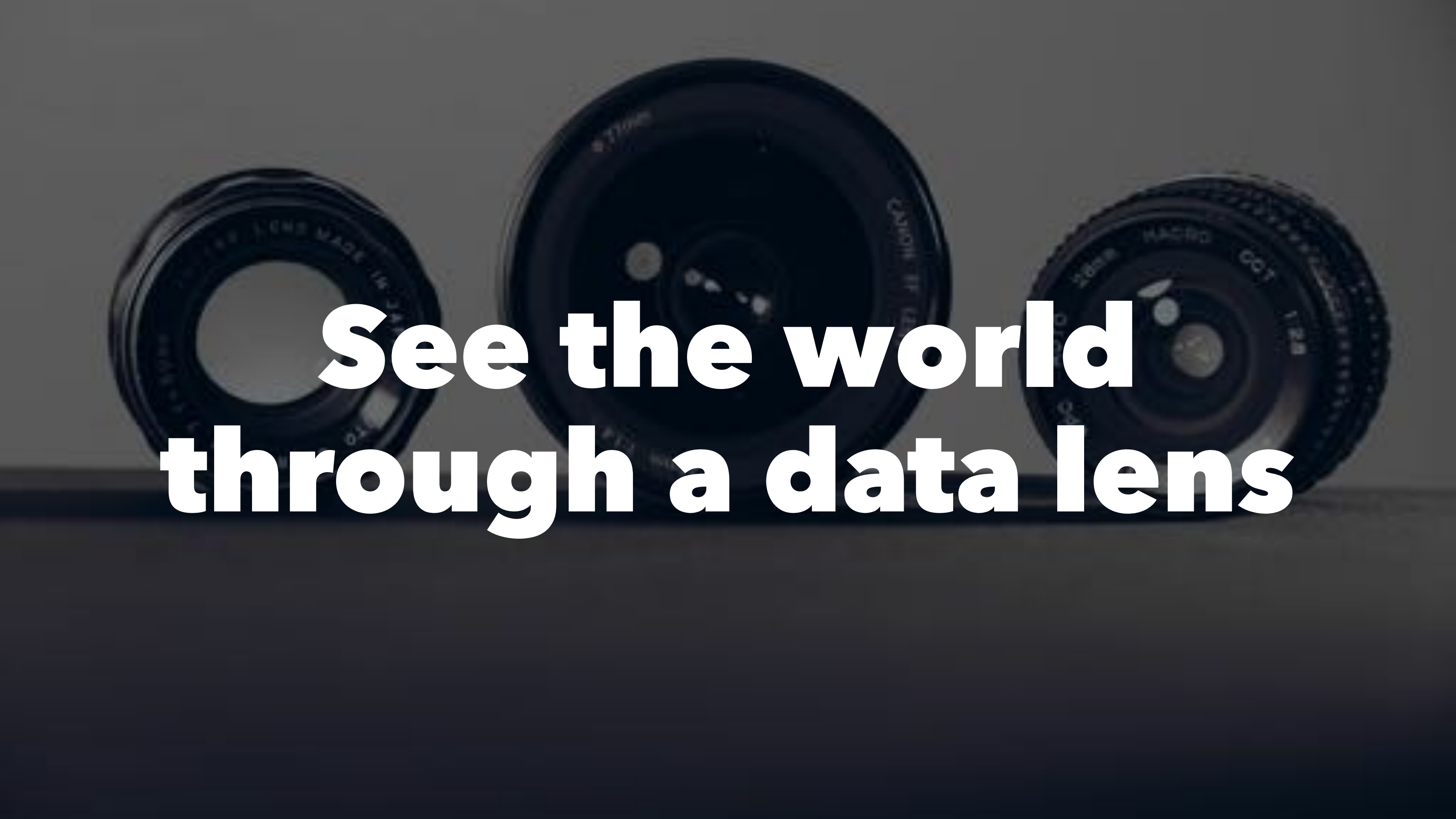
Amit

@amitkaps



Bargava

@bargava



**See the world
through a data lens**

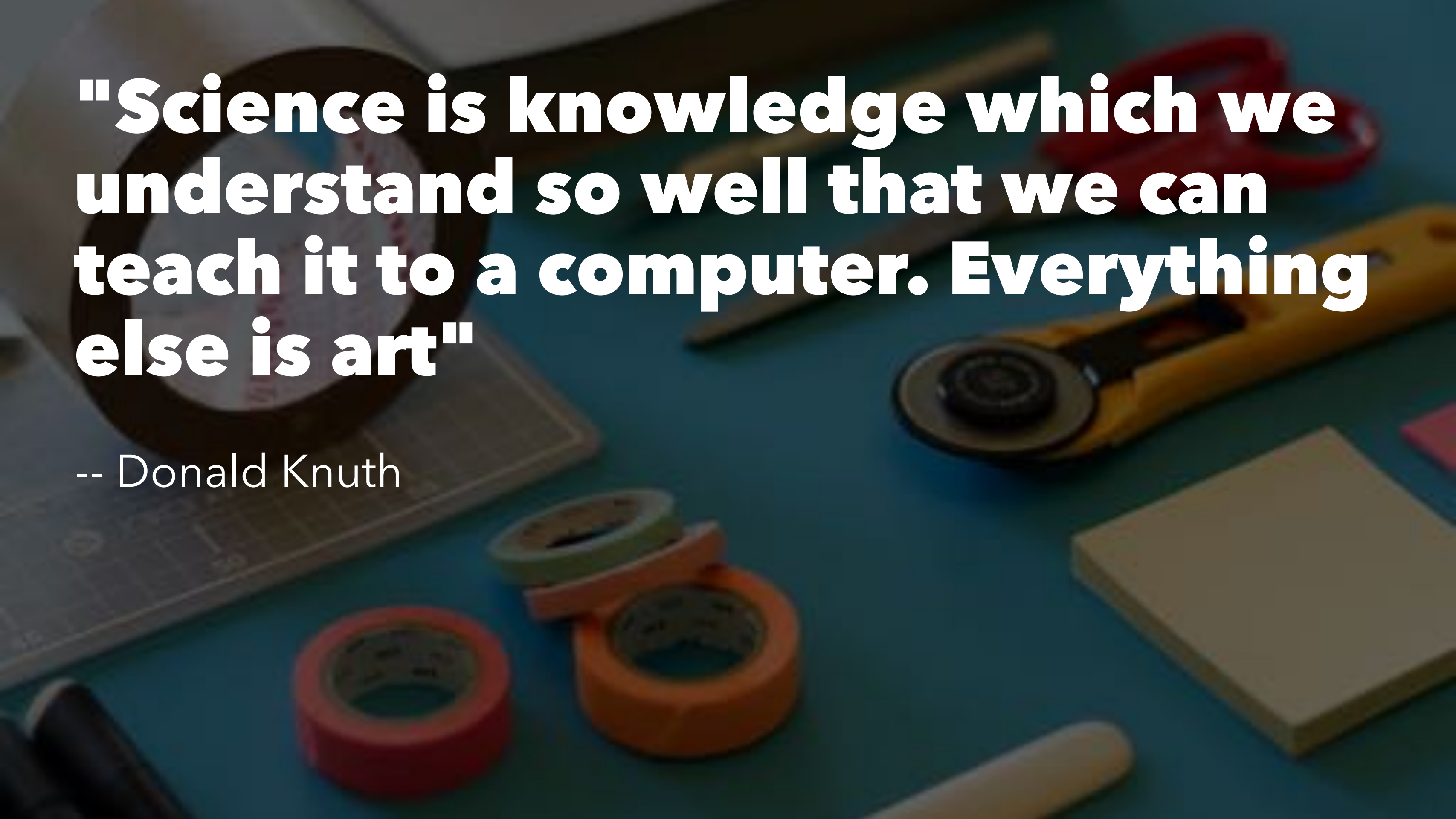


**"Data is just a clue to the end
truth"**

-- Josh Smith



Data Driven Decisions



"Science is knowledge which we understand so well that we can teach it to a computer. Everything else is art"

-- Donald Knuth



Data Science is an Art

A person is playing a violin, with their hand visible on the bow. The background is dark and out of focus, showing some papers or documents. The text "Hypothesis Driven Approach" is overlaid in white, bold, sans-serif font.

Hypothesis Driven Approach



Frame

"An approximate answer to the right problem is worth a good deal"



Acquire

"80% perspiration, 10% great idea, 10% great output"



Refine

"All data is messy."



Explore

**"I don't know, what I don't
know."**

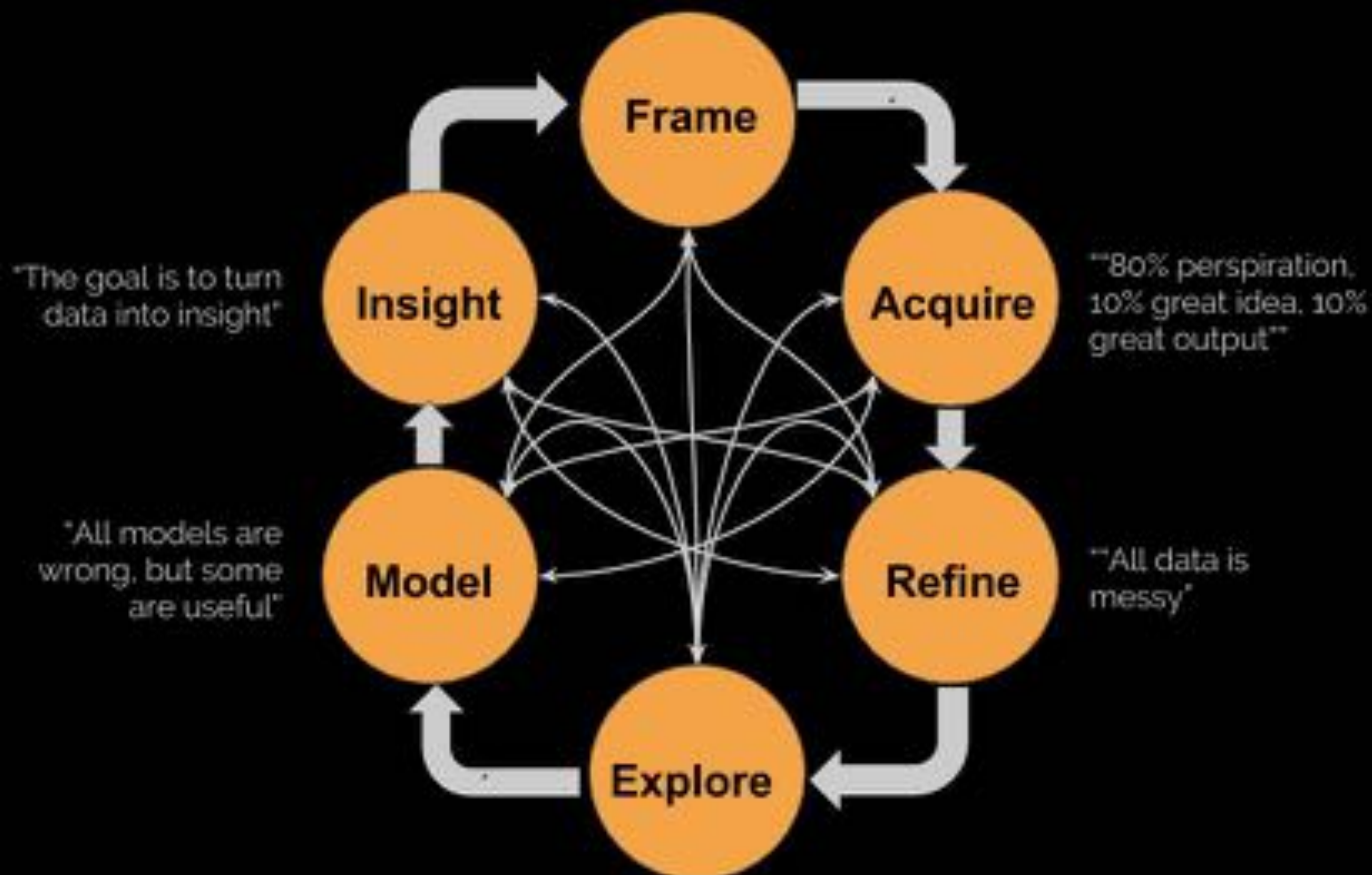
Model

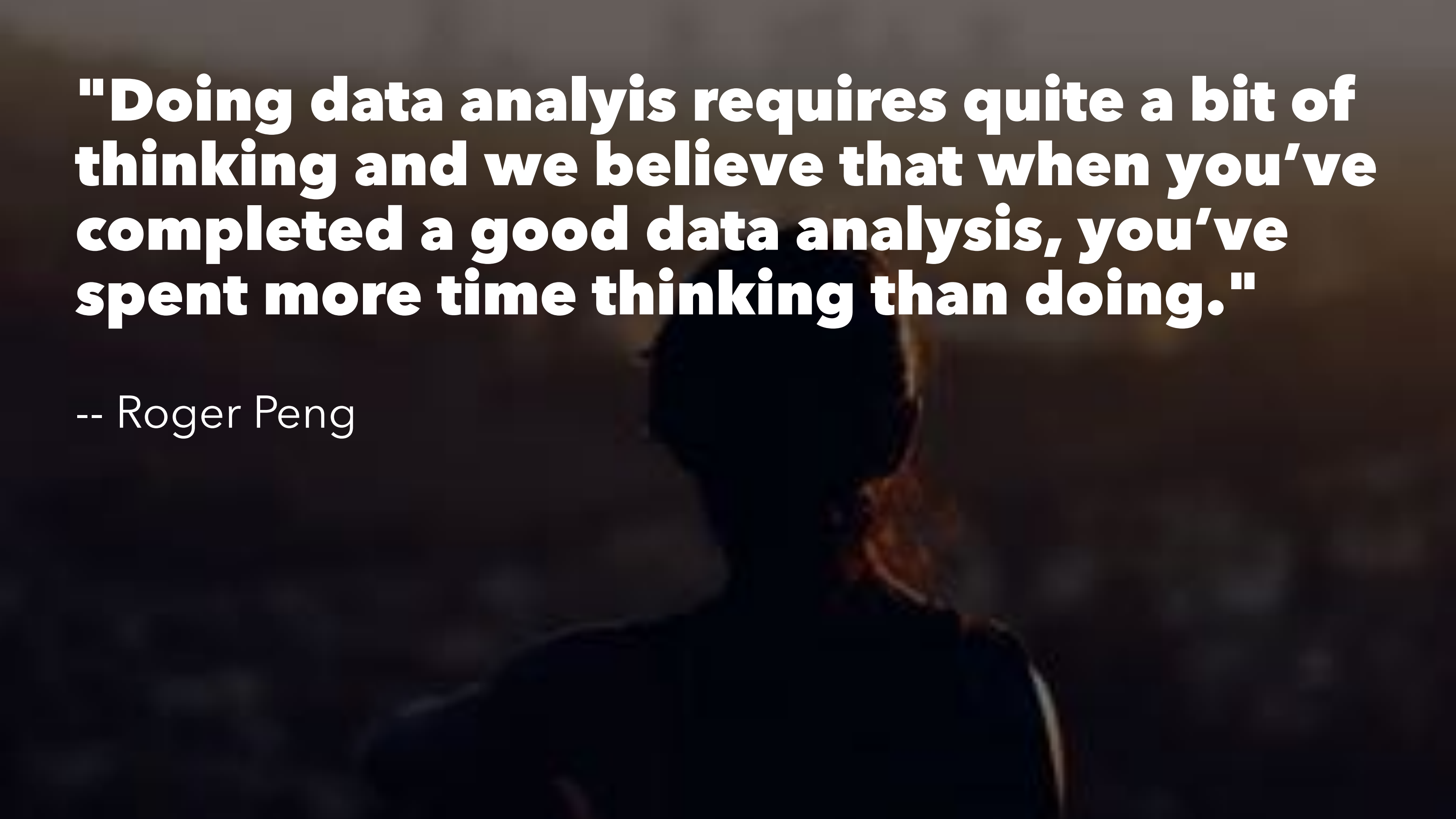
"All models are wrong, but some are useful"

A rustic log cabin with a chimney, nestled in a dense forest of tall trees. The scene is dimly lit, suggesting dusk or dawn, with a soft glow from the cabin's interior. The text is overlaid in white, bold font.

Insight

**"The goal is to turn data into
insight"**



A person in a dark suit is seen from the side, looking out at a city at night. The city lights are visible in the background, creating a bokeh effect. The person's face is in shadow, and they appear to be looking towards the right side of the frame.

"Doing data analysis requires quite a bit of thinking and we believe that when you've completed a good data analysis, you've spent more time thinking than doing."

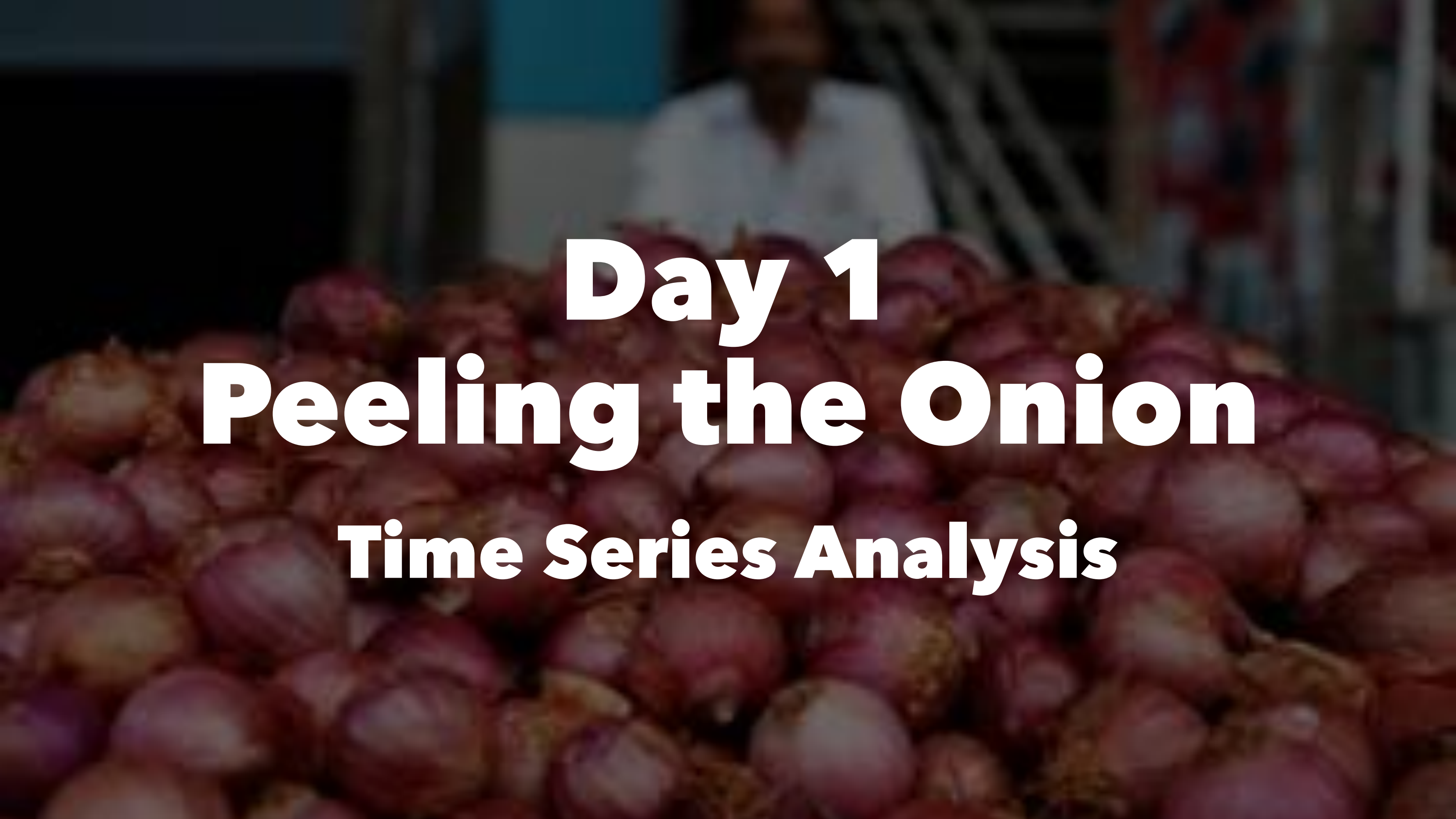
-- Roger Peng

A collection of various tools including axes, hammers, wrenches, and gloves, arranged on a dark wooden surface. The tools are scattered across the frame, with some axes and hammers having wooden handles and others being metal. There are also wrenches and a pair of gloves visible. The background is a dark, textured wooden surface.

Python Data Stack

A stack of old, worn books with the text 'Case Studies' overlaid in a large, white, bold font. The books are stacked horizontally, and the text is centered across the middle of the image. The background is a dark, muted color, and the overall tone is academic and professional.

Case Studies



Day 1

Peeling the Onion

Time Series Analysis

Day 2

Grocery

**Market Basket Analysis /
Collaborative Filter**

Day 2

Bank Marketing

**Random Forest and Gradient
Boosting**

Day 3

DataTau


Text Analytics

A close-up photograph of two hands, one appearing to be an adult's and the other a child's, gently holding a small, dark, textured object. The hands are positioned over a light-colored, textured surface, possibly sand or a workbench. The background is blurred, showing hints of red and blue. The overall tone is warm and focused on the tactile experience.

Learning Approach



Do the Exercises

A photograph of two men in an office environment. The man on the left, wearing glasses and a light-colored shirt, is pointing his right index finger towards a computer monitor. The man on the right, with a beard and wearing a light-colored shirt, is looking at the monitor. They are both seated at a desk. On the desk, there is a yellow coffee cup with a black lid, a white computer mouse, and some papers. The background shows a wall with various sticky notes and a whiteboard. The text "Pair up & Learn" is overlaid in the center of the image in a large, white, bold font.

Pair up & Learn



Call for Help



Enjoy the workshop

Github Repo

<https://github.com/amitkaps/machine-learning>

Exercise

1. Time Series Exercise

"Predict the number of tickets that will be raised in the next week"

- **Frame:** What to forecast? At what horizon? At what level?
- **Acquire, Refine, Explore:** Do EDA to understand the trend and pattern within the data
- **Models:** Mean Model, Linear Trend, Random Walk, Simple Moving Average, Exp Smoothing, Decomposition, ARIMA
- **Insight:** Share the insight through a datavis of the models

2. Text Analytics Exercise

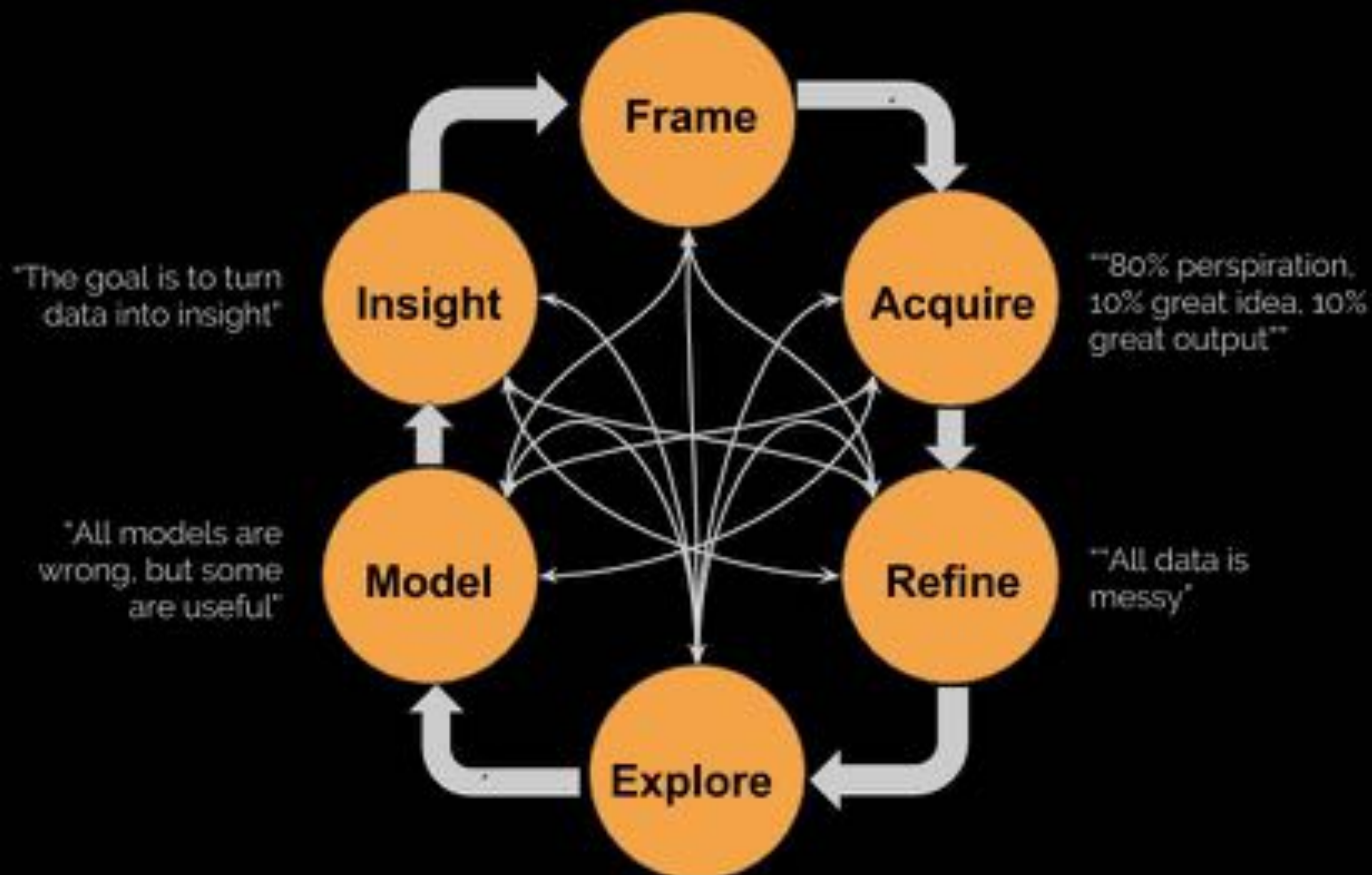
"Identify the entity, features and topics in the 'Comments Data' or 'Twitter Data' within your text data"

- **Frame:** What are the comments you are trying to understand?
- **Acquire, Refine, Explore:** Do Wordcloud, Lemmatization, Part of Speech Analysis, and Entity Chunking
- **Models:** TF-IDF, Topic Modelling, Sentiment Analysis
- **Insight:** Share the insight through word cloud and topic

Feedback

<https://amitkaps.typeform.com/to/i6wl2E>

Recap



Frame

- **Toy Problems**
- **Simple Problems**
- Complex Problems
- Business Problems
- Research Problems



Acquire

- **Scraping** (structured, unstructured)
- **Files** (csv, xls, json, xml, pdf, ...)
- Database (sqlite, ...)
- APIs
- Streaming

Refine



- Data Cleaning (inconsistent, missing, ...)
- **Data Refining** (derive, parse, merge, filter, convert, ...)
- **Data Transformations** (group by, pivot, aggregate, sample, summarise, ...)

Explore

- **Simple Vis**
- Multi Dimensional Vis
- Geographic Vis
- Large Data Vis (Bin - Summarise - Smooth)
- Interactive Vis

Model - Supervised Learning

- *Continuous*: Regression - **Linear**, Polynomial, Tree Based Methods - CART, **Random Forest**, Gradient Boosting Machines
- *Classification* - **Logistics Regression**, Tree, KNN, SVM, Naive-Bayes, Bayesian Network

Model - UnSupervised Learning

- *Continuous*: Clustering & Dimensionality Reduction like PCA, SVD, MDS, K-means
- *Categorical*: Association Analysis

Model - Advanced /

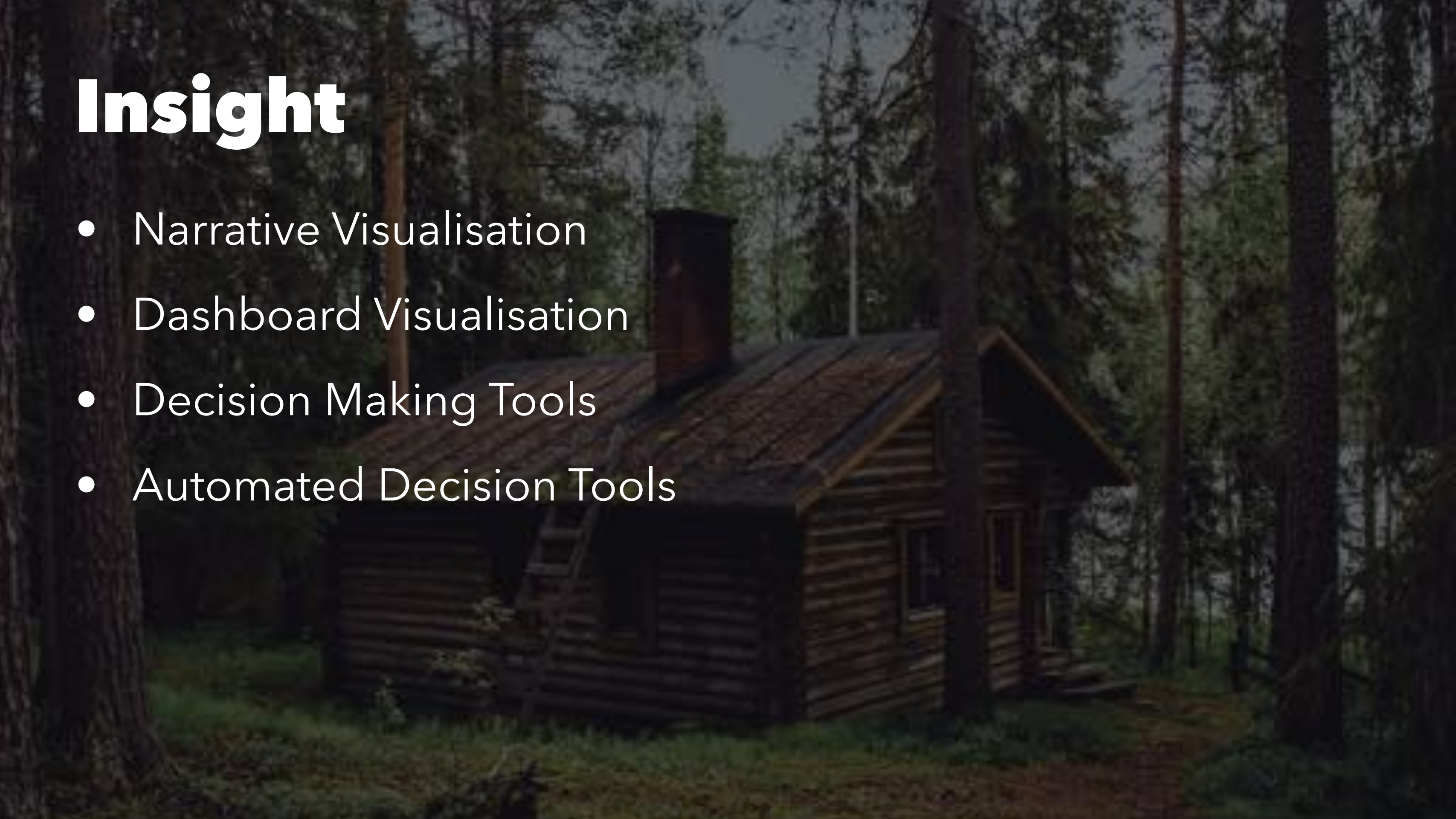
- **Time Series**
- **Text Analytics**
- Network / Graph Analytics
- Optimization

Model - Specialized

- Reinforcement Learning
- Online Learning
- Deep Learning
- Other Applications: Image, Speech

Insight

- Narrative Visualisation
- Dashboard Visualisation
- Decision Making Tools
- Automated Decision Tools

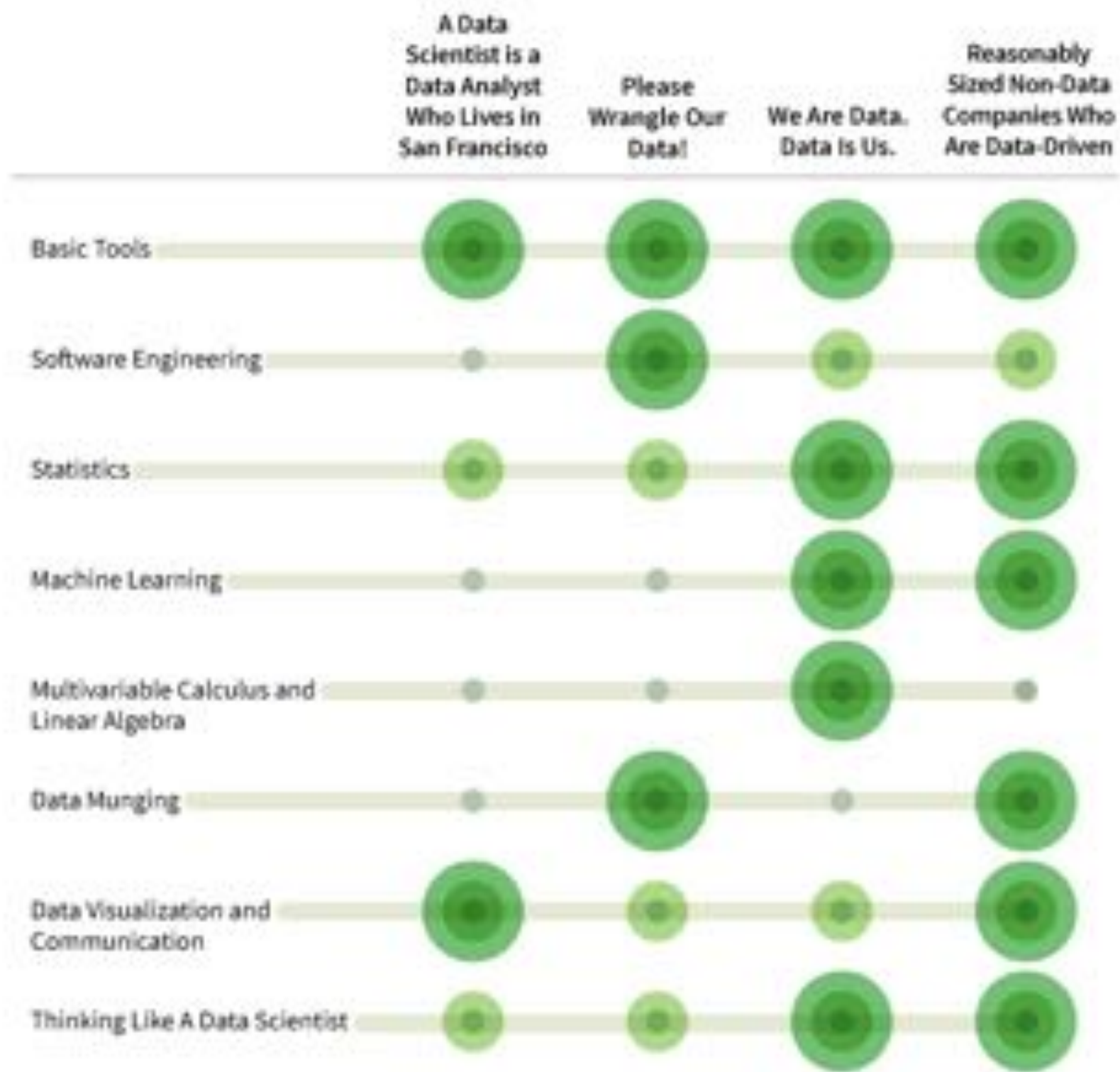


PyData Stack

- **Acquire / Refine:** Pandas, BeautifulSoup, Selenium, Requests, SQL Alchemy, Numpy, Blaze
- **Explore:** Matplotlib, Seaborn, Bokeh, Plotly, Vega, Folium
- **Model:** Scikit-Learn, StatsModels, SciPy, Gensim, Keras, Tensor Flow, PySpark
- **Insight:** Django, Flask

skills





Very important



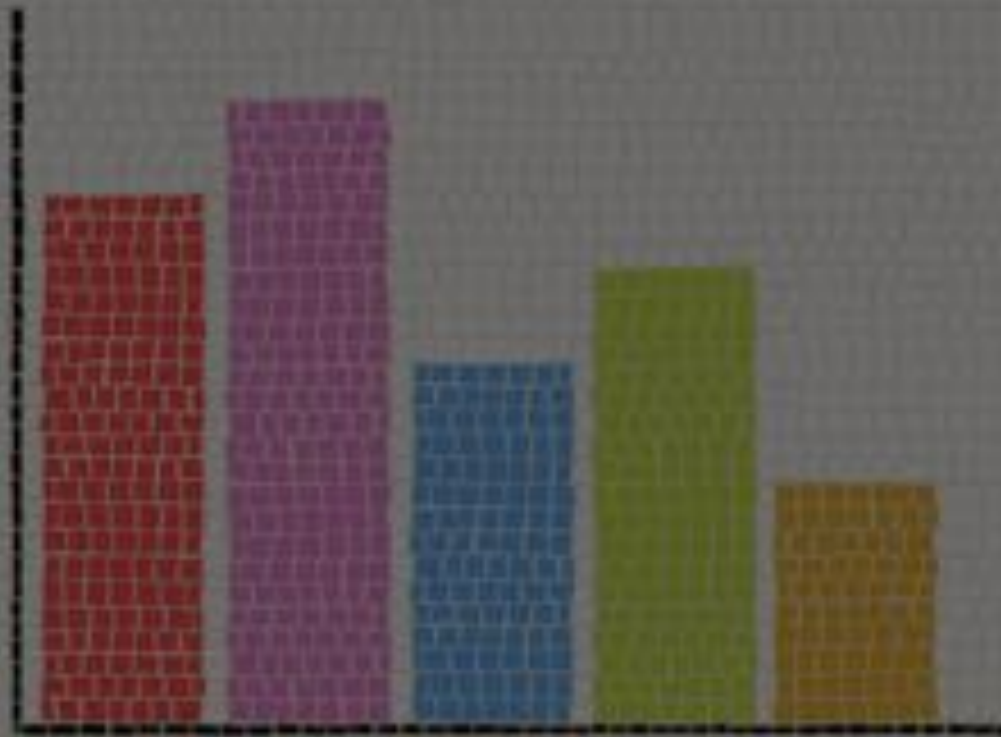
Somewhat important



Not that important

The Art of Data Science

A Guide for Anyone Who Works with Data



Roger D. Peng & Elizabeth Matsui

Data Wrangling with Pandas, NumPy, and IPython

Python for Data Analysis



O'REILLY®

Wes McKinney

O'REILLY®

2nd Edition

Think Stats

EXPLORATORY
DATA ANALYSIS



Allen B. Downey

Books

The Art of Data Science

A Guide for Anyone Who Works with Data



Roger D. Peng & Elizabeth Matsui

Data Wrangling with Pandas, NumPy, and IPython

Python for Data Analysis



O'REILLY*

Wes McKinney

O'REILLY*

2nd Edition

Think Stats

EXPLORATORY
DATA ANALYSIS



Allen B. Downey

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

Resources - Statistical Learning

- One of the good books on statistical learning is ISLR -> [An Introduction to Statistical Learning with Application in R](#)
- You can find all the ISLR code in python at this github repo - <https://github.com/JWarmenhoven/ISLR-python>

Resources - Time Series

- Forecasting: Principle and Text
- Statistical forecasting: Notes on regression and time series analysis Case

Resources - Text Analytics

- Natural Language Processing with Python

Online Course

- Harvard Data Science Course - [CS 109 Course](#) (It is structured in similar way to the approach we shared)
- Data Science Specialisation - [JHU Data Science](#) (It is a good course, though the material is coded in R)
- Many more on Coursera & Udacity...





Speak to Us!



Thank you
@amitkaps | @bargava