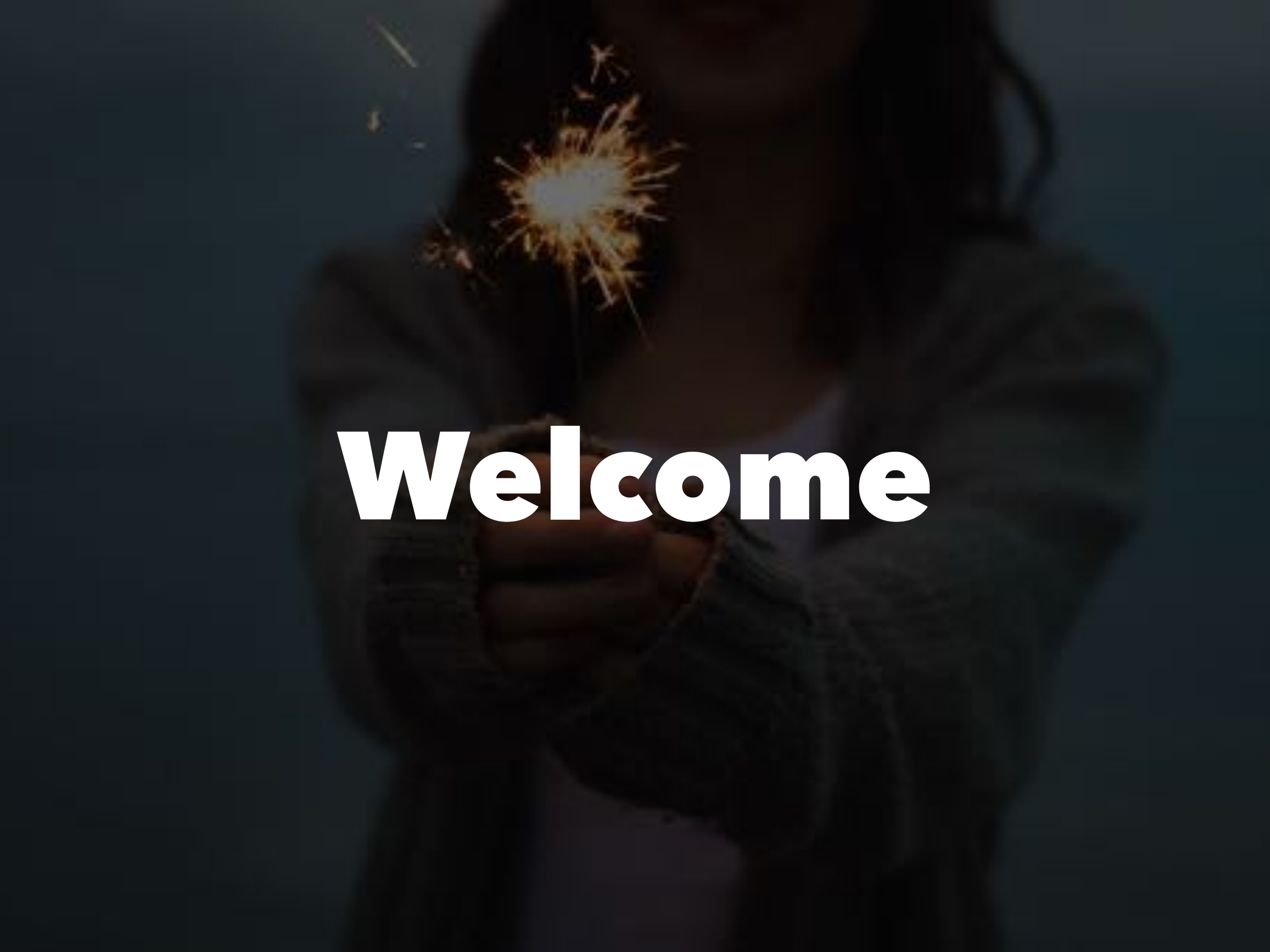
Intro to Data Science in Python Workshop

@unnati_xyz | @hasgeek


Welcome

Facilitators


Amit
@amitkaps


Nischal
@nischalhp


Raghotham
@raghothams

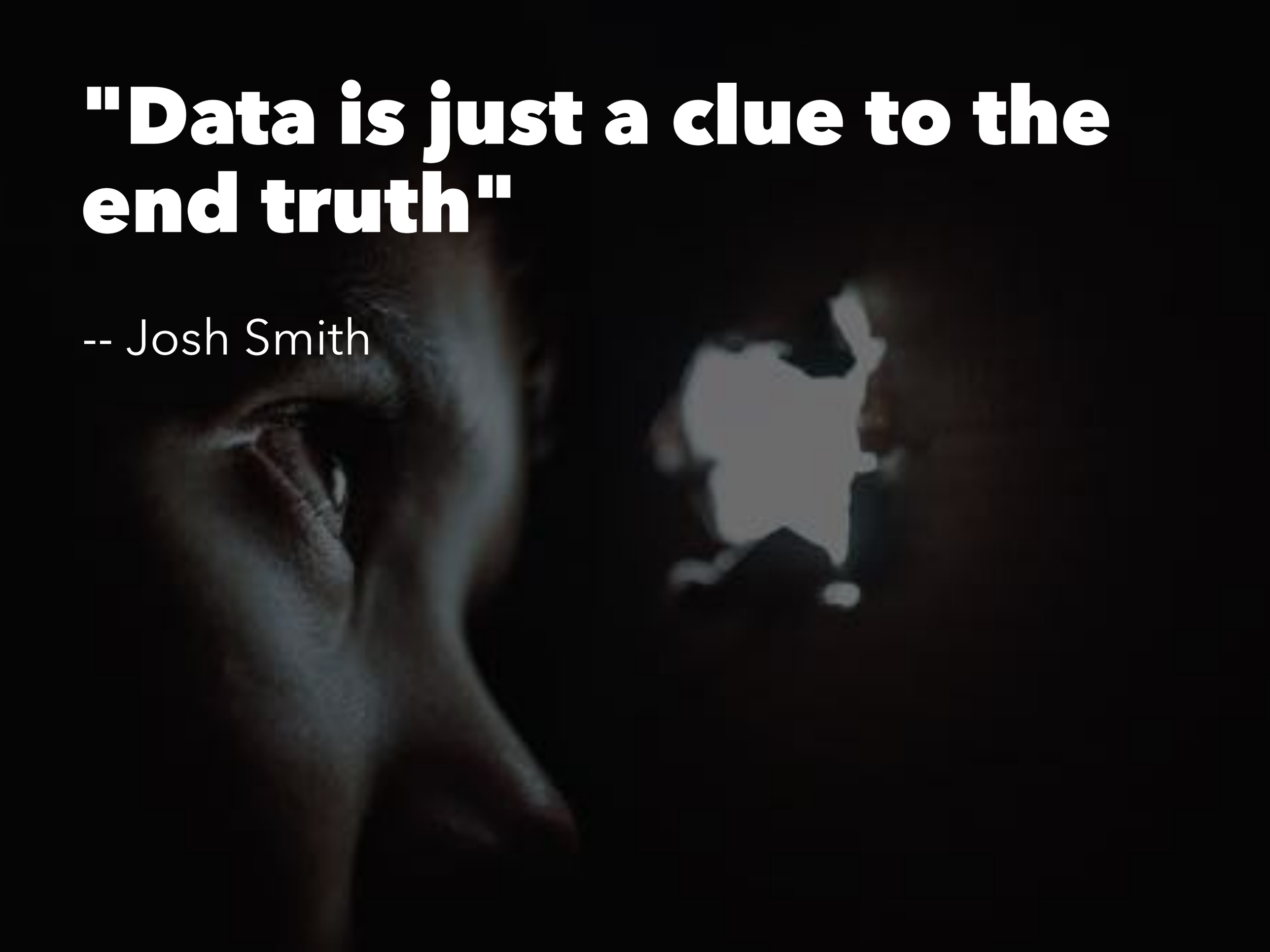

Bargava
@bargava


Shrayas
@shrayasr

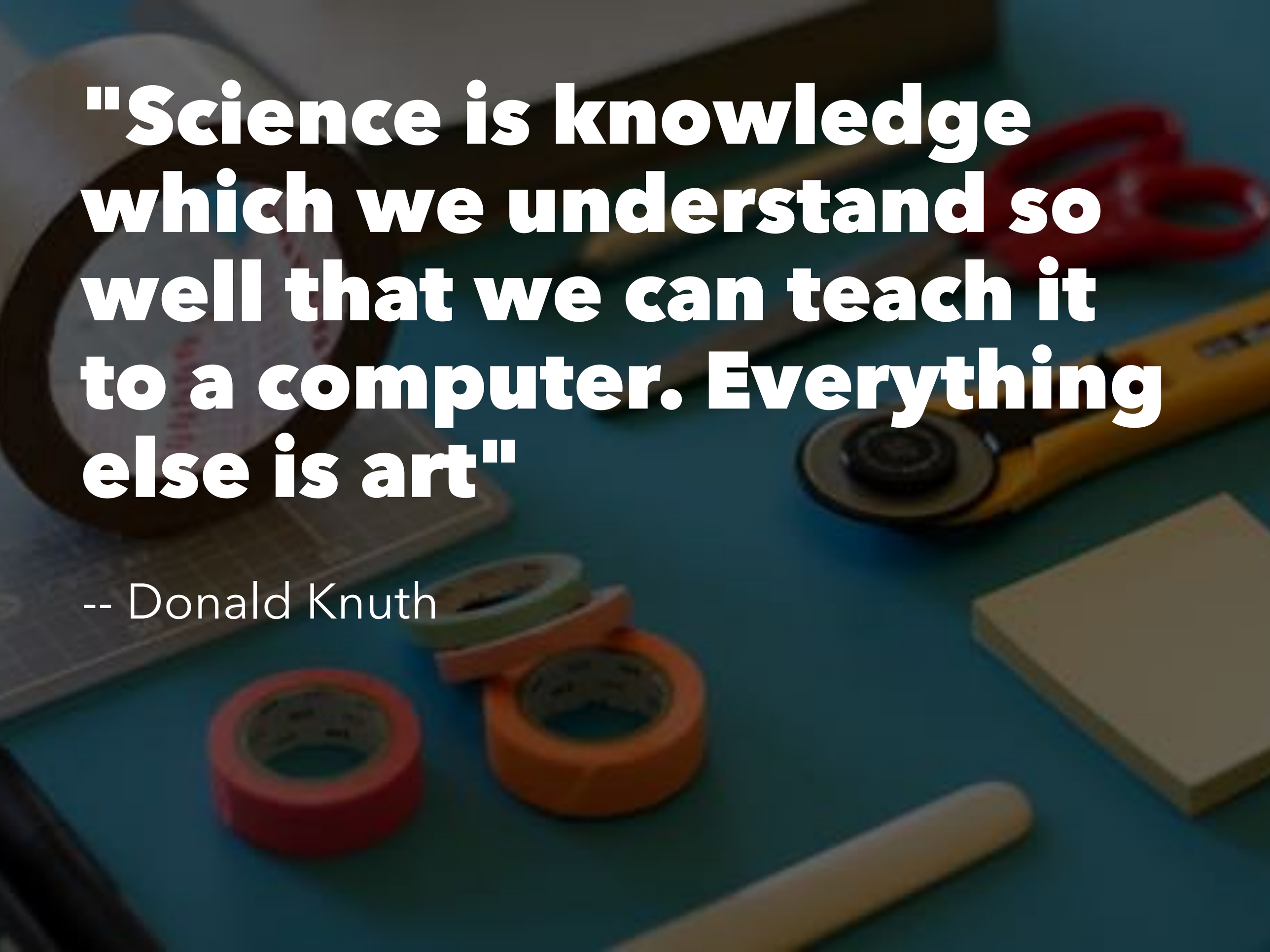
See the world through a data lens

"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

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"

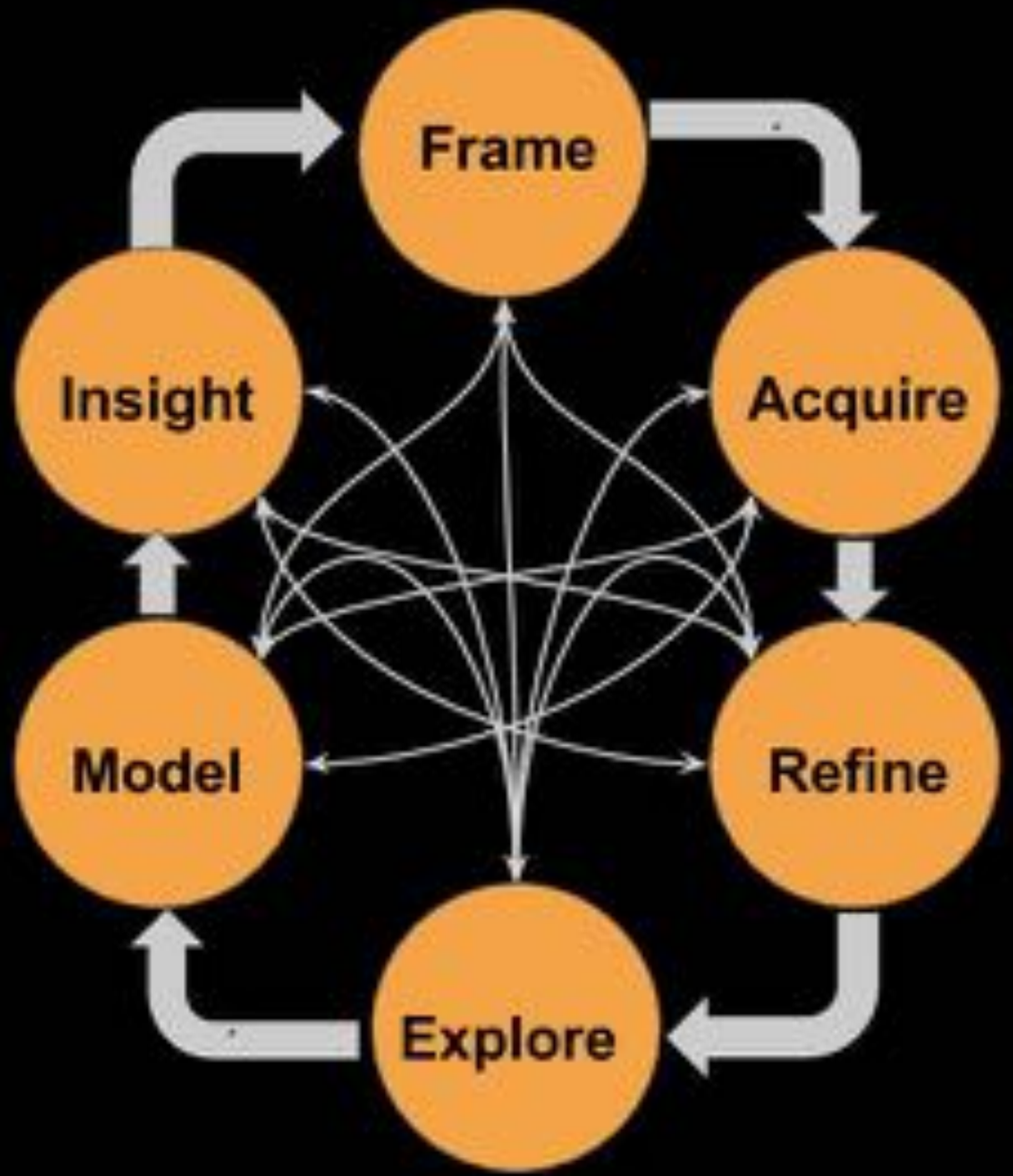
Insight

"The goal is to turn data into insight"

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

"The goal is to turn data into insight"

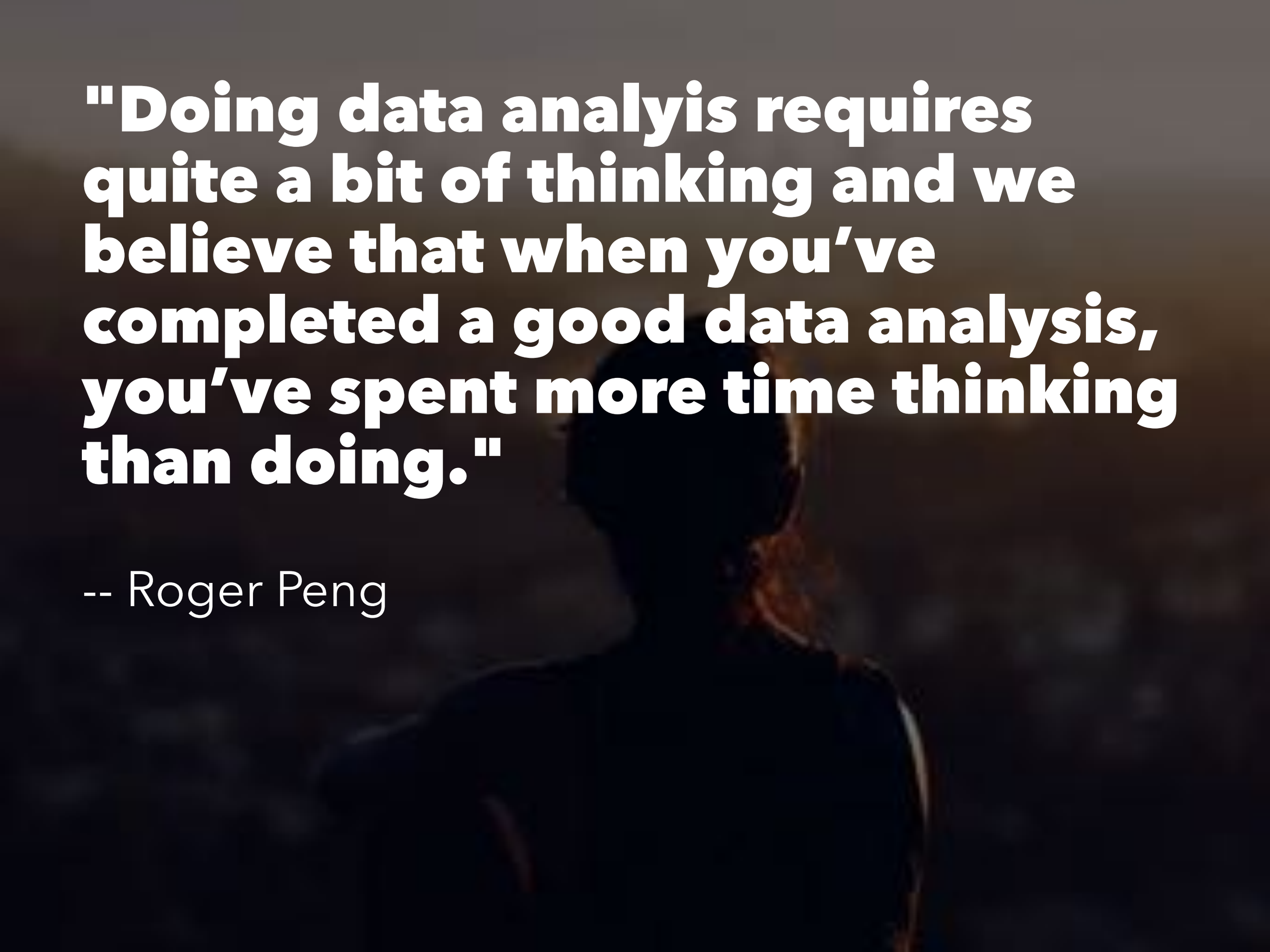
"All models are wrong, but some are useful"



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

"All data is messy"

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



"Doing data analyis 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

Python Data Stack

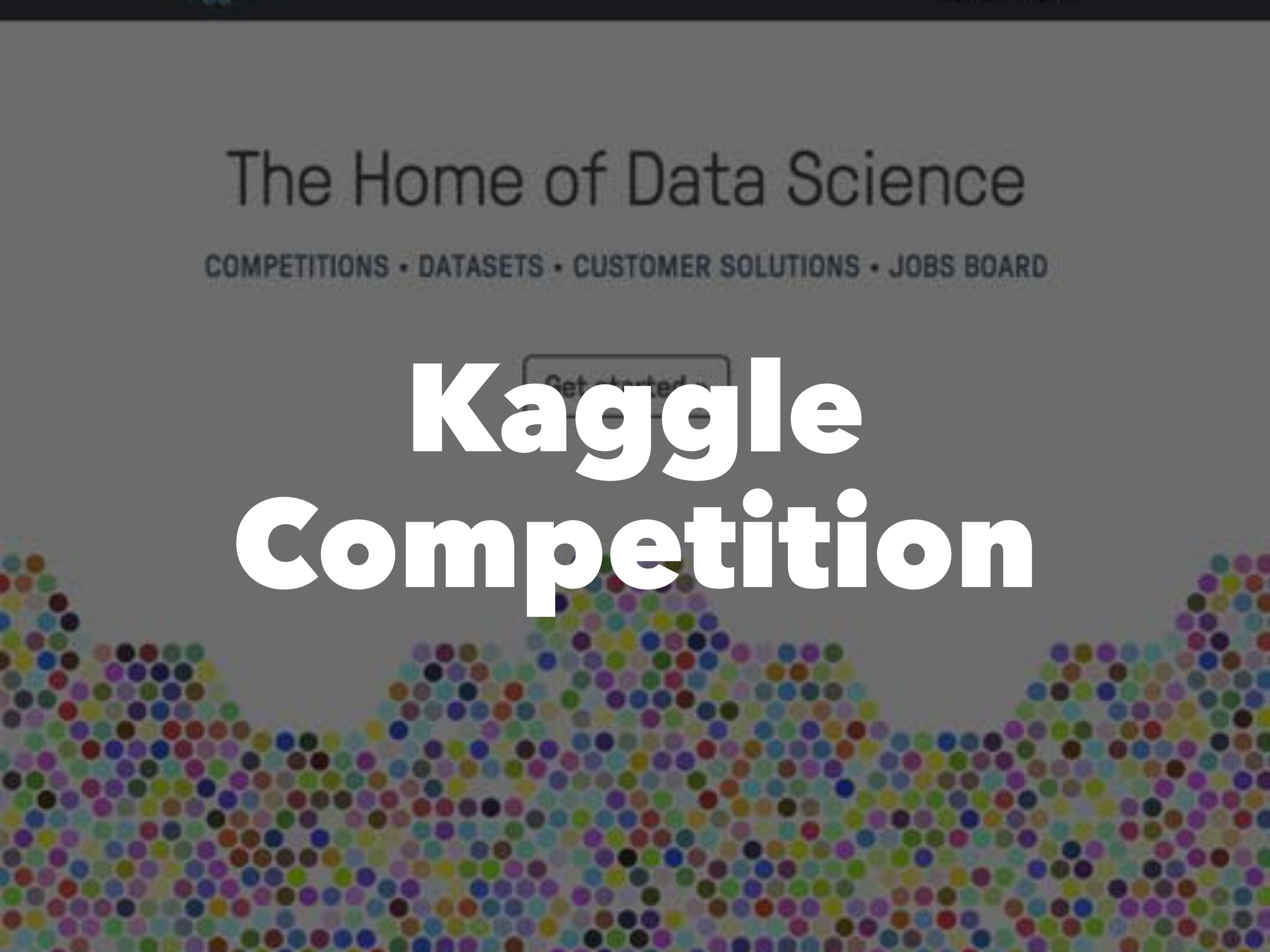
Case Studies

Peeling the Onion

Hard Disk Failure

Shining Diamonds

Red Wine Selection


The Home of Data Science
COMPETITIONS • DATASETS • CUSTOMER SOLUTIONS • JOBS BOARD
Kaggle
Competition

Learning Approach

Do the Exercises

Pair up & Learn

Call for Help

Schedule

Session 0 (0830 - 0930)

- Installation

Session 1 (0930 - 1115)

- Overview of Data Science
- Data Science Process
- How to use Jupyter Notebook
- Intro to Data Structures in Python

Session 2 (1135 - 1300)

- Case 1: Peeling the Onion
- Exploring Onion Price and Quantity

Session 3 (1400 - 1530)

- Fun visualization exercise
- Communicating the Onion Insights
- Acquiring the Onion data (Web Scraping)

Session 4 (1550 - 1730)

- Case 2: Hard Disk Failure
- Explore & build intuition around Failure Rates

Optional Advanced Session (1730 - 1830)

- Working with SQL to Acquire and Refine Hard Disk Data
- Office Hours

Food and Hydration

0830 - 0930: Breakfast

1115 - 1135: Tea Break

1300 - 1400: Lunch

1530 - 1550: Tea Break

Session 5 (0930 - 1115)

- Reflections
- Intro to Machine Learning
- Case 3: Shining Diamonds
- Linear Regression

Session 6 (1135 - 1300)

- Case 4: Red Wine
- Logistics Regression for Classification

Session 7 (1400 - 1530)

- Fun Demo : Music Visualization
- Case 5: Entering a Kaggle Competition

Session 8 (1550 - 1730)

- Kaggle Competition (contd.)
- Recap and Way Forward
- Feedback

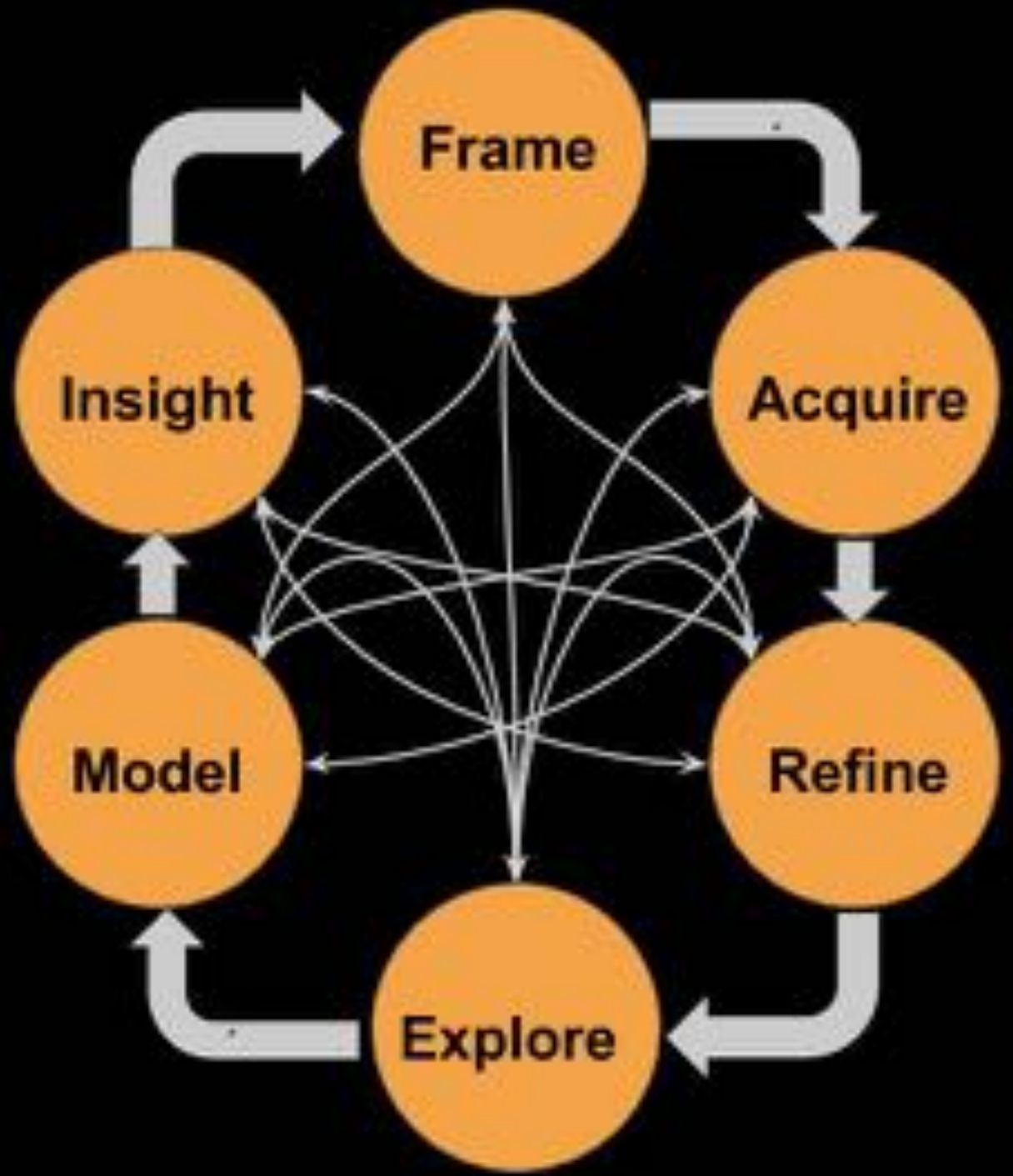
Enjoy the workshop

Recap

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

"The goal is to turn data into insight"

"All models are wrong, but some are useful"



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

"All data is messy"

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

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 / Specialized

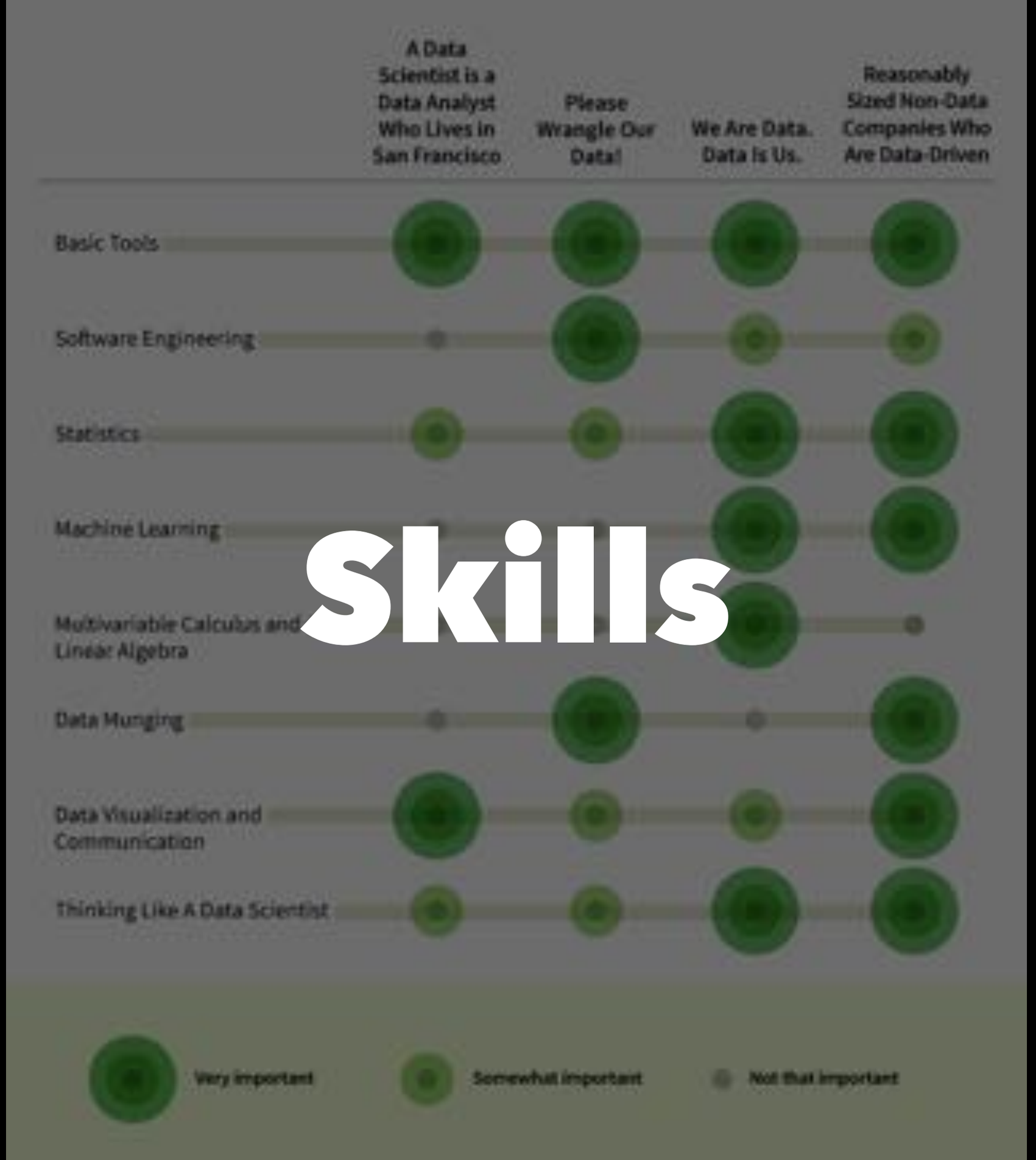
- Network / Graph Analytics
- Optimization
- Reinforcement Learning
- Online Learning
- Deep Learning
- Applications: Time Series, Text, Image, Speech

Insight

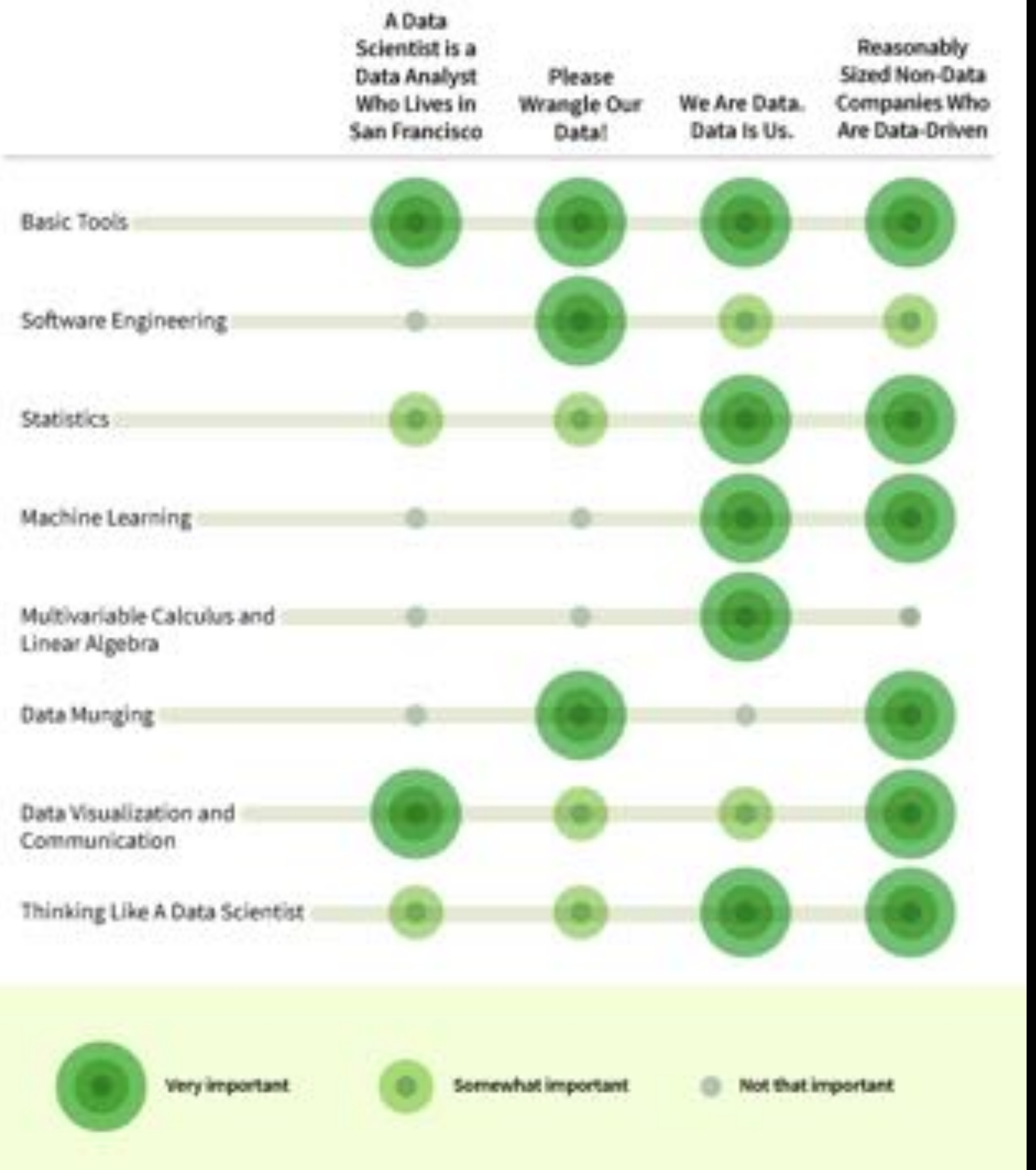
- Narrative Visualisation
- Dashboard Visualisation
- Decision Making Tools
- Automated Decision Tools

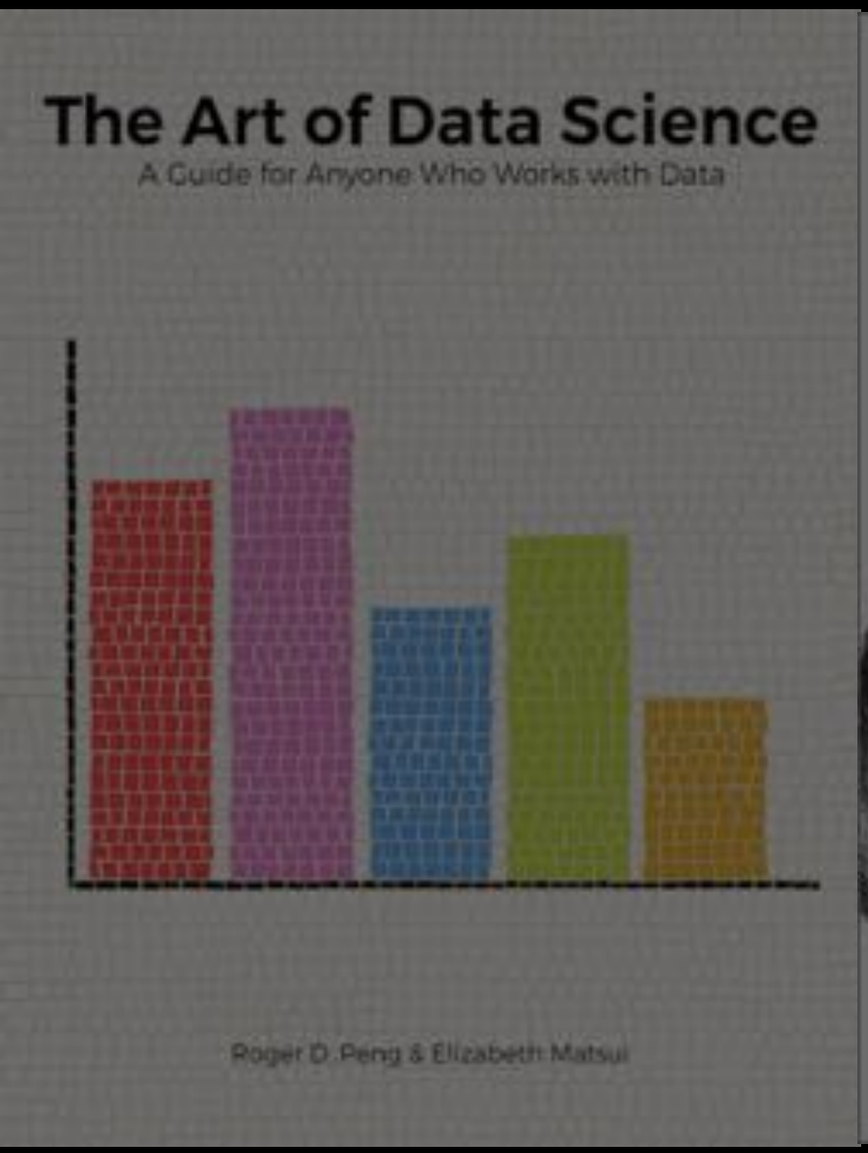
PyData Stack

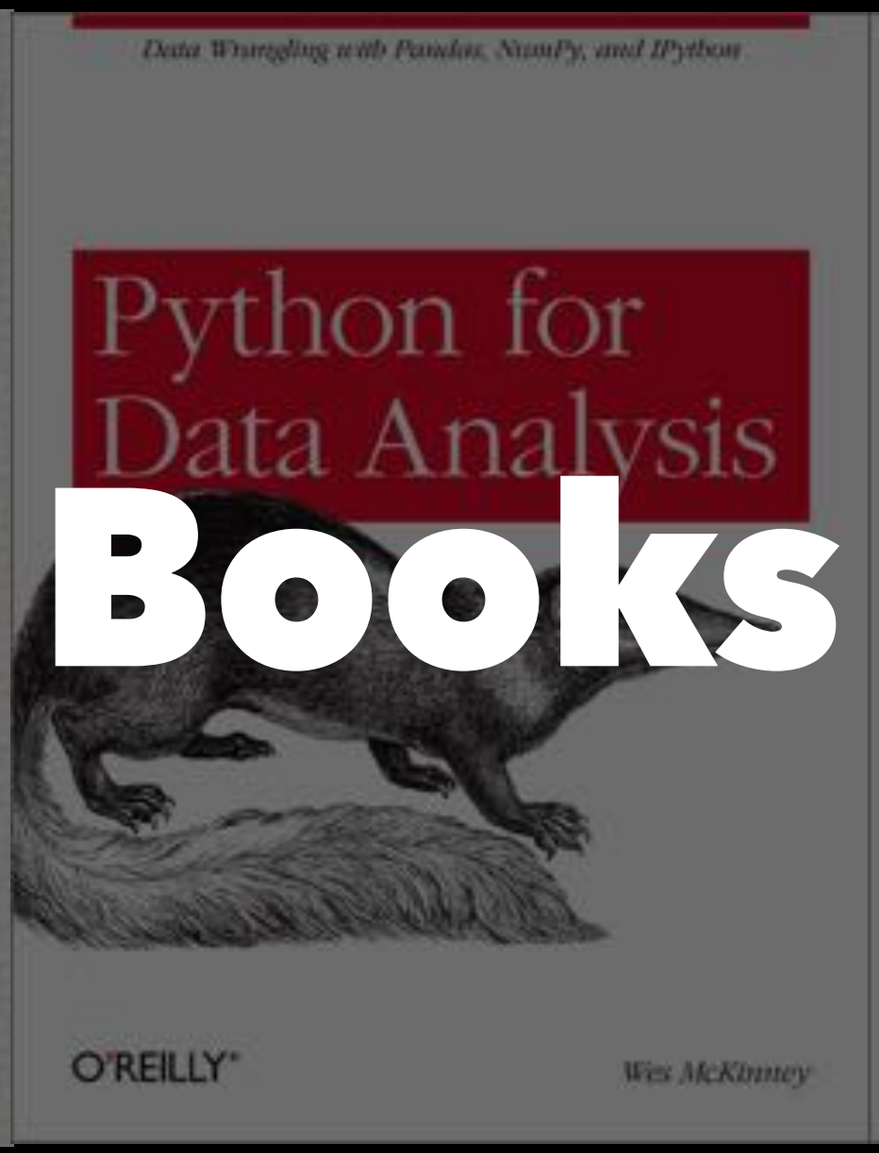
- **Acquire / Refine**: Pandas, Beautiful Soup, Selenium, Requests, SQL Alchemy, Numpy, Blaze
- **Explore**: MatPlotLib, Seaborn, Bokeh, Plotly, Vega, Folium
- **Model**: Scikit-Learn, StatsModels, SciPy, Gensim, Keras, Tensor Flow, PySpark
- **Insight**: Django, Flask


A Data Scientist is a Data Analyst Who Lives in San Francisco
Please Wrangle Our Data!
We Are Data. Data Is Us.
Reasonably Sized Non-Data Companies Who Are Data-Driven
Basic Tools
Software Engineering
Statistics
Machine Learning
Multivariable Calculus and Linear Algebra
Data Munging
Data Visualization and Communication
Thinking Like A Data Scientist
Very important
Somewhat important
Not that important

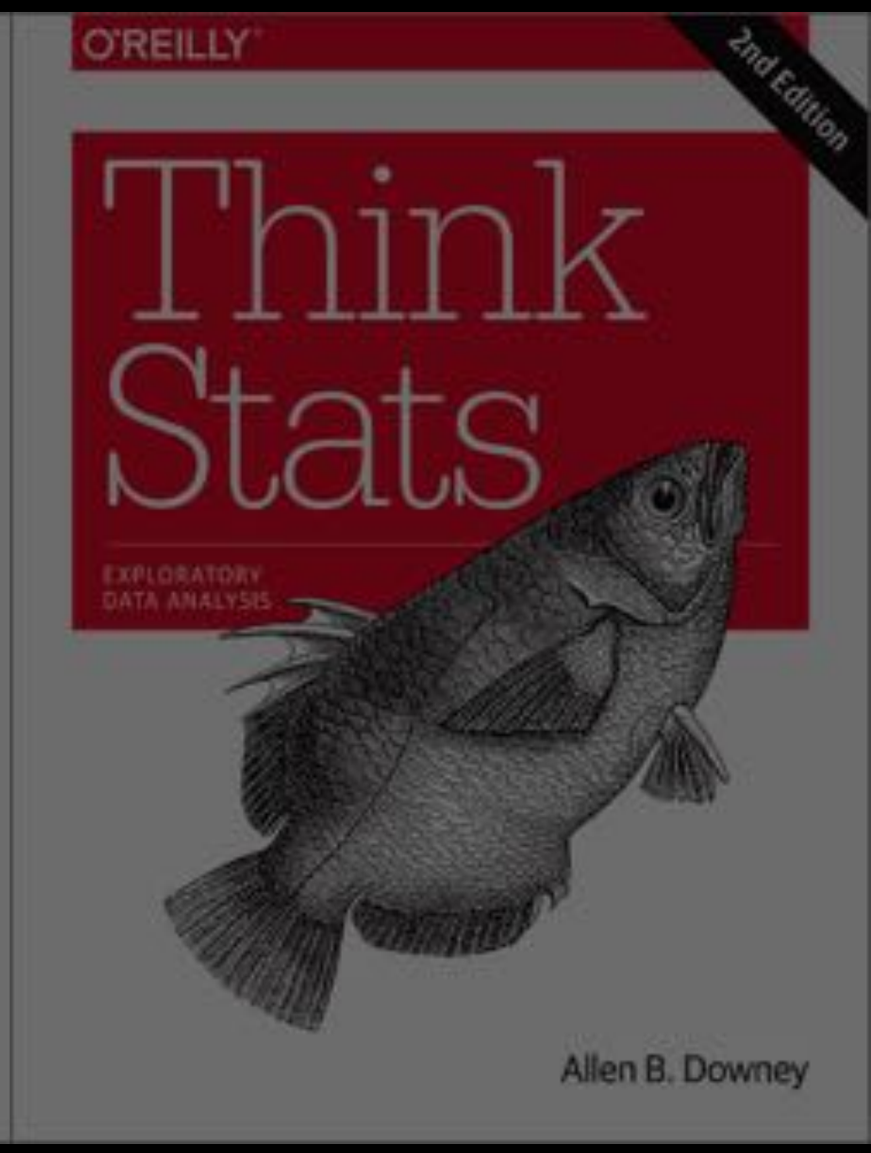
skills


A Data Scientist is a Data Analyst Who Lives in San Francisco
Please Wrangle Our Data!
We Are Data. Data Is Us.
Reasonably Sized Non-Data Companies Who Are Data-Driven
Basic Tools
Software Engineering
Statistics
Machine Learning
Multivariable Calculus and Linear Algebra
Data Munging
Data Visualization and Communication
Thinking Like A Data Scientist
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

Books


O'REILLY
2nd Edition
Think Stats
EXPLORATORY DATA ANALYSIS
Allen B. Downey

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®

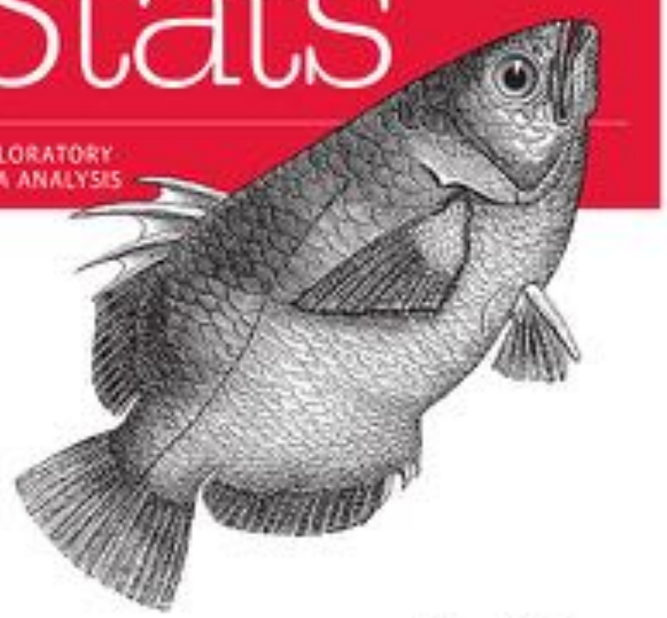
Wes McKinney

O'REILLY®

2nd Edition

Think Stats

EXPLORATORY DATA ANALYSIS



Allen B. Downey

Online Course

- Harvard Data Science Course - CS 109 Course
- Structured in similar way to the approach we shared
- Many more on Coursera & Udacity...

Upcoming Workshops

- Advanced Data Science (Machine Learning, Statistics)
- Data Science at Scale (Spark)
- Visualisation (Multi-Dimensional, Geographic, Large Data)
- Deep Learning (Text, Speech, Image, Video)

Speak to Us!

Custom Workshops

Data Science Consulting

Thank you

unnati.xyz/workshop-feedback

has geek