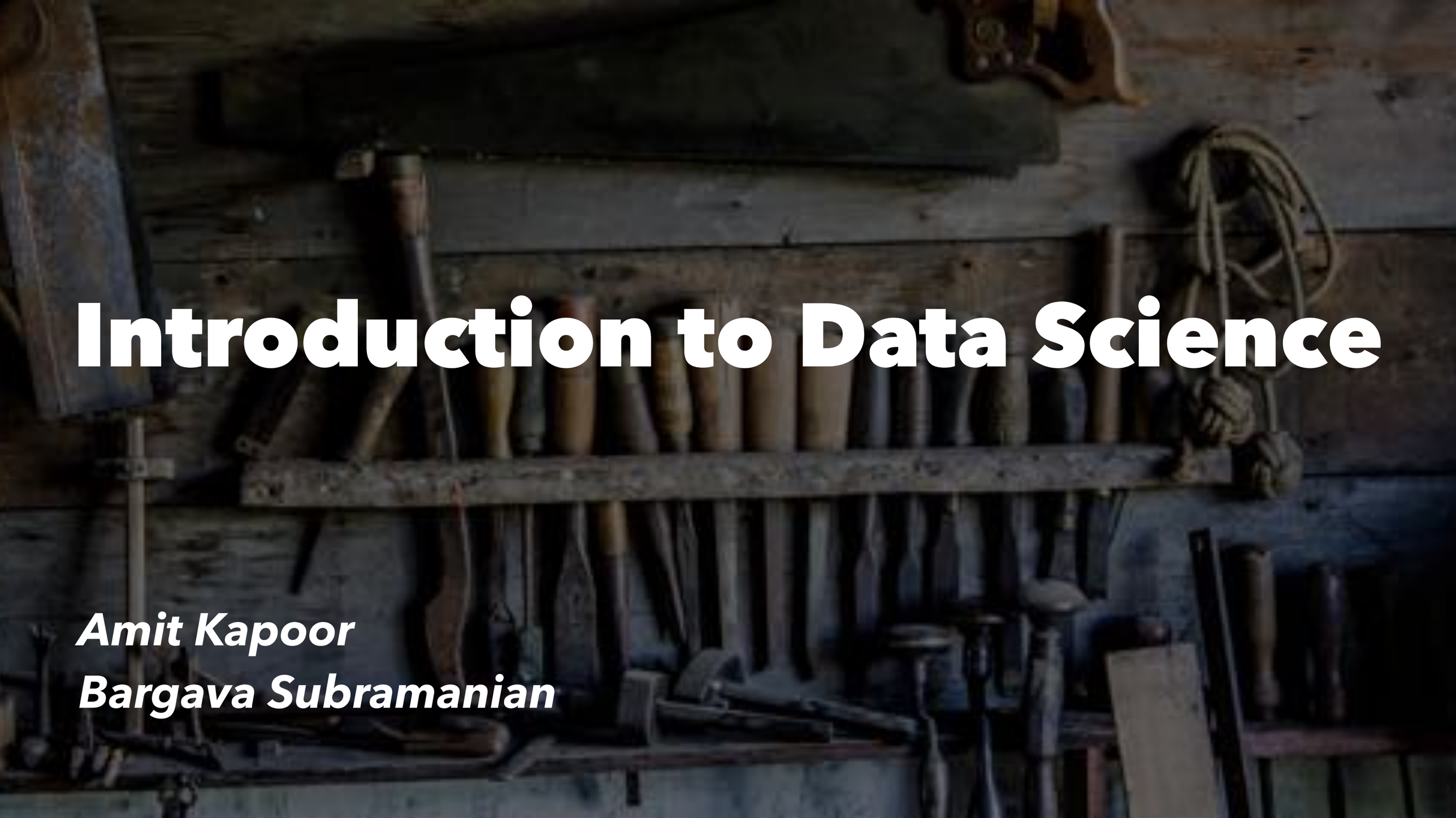
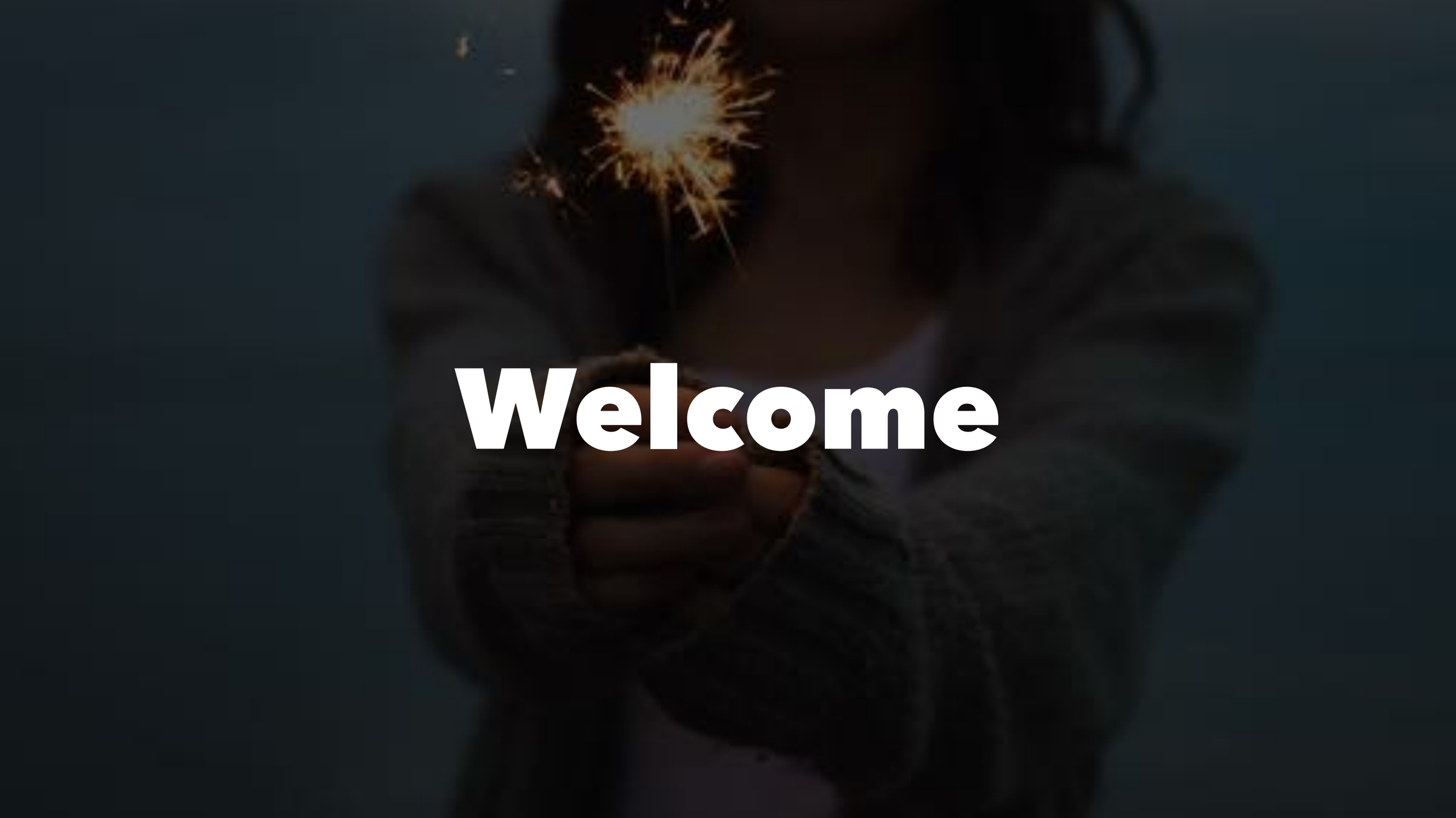


Introduction to Data Science

Amit Kapoor
Bargava Subramanian


Welcome


Amit Kapoor
@amitkaps


Bargava
@bargava

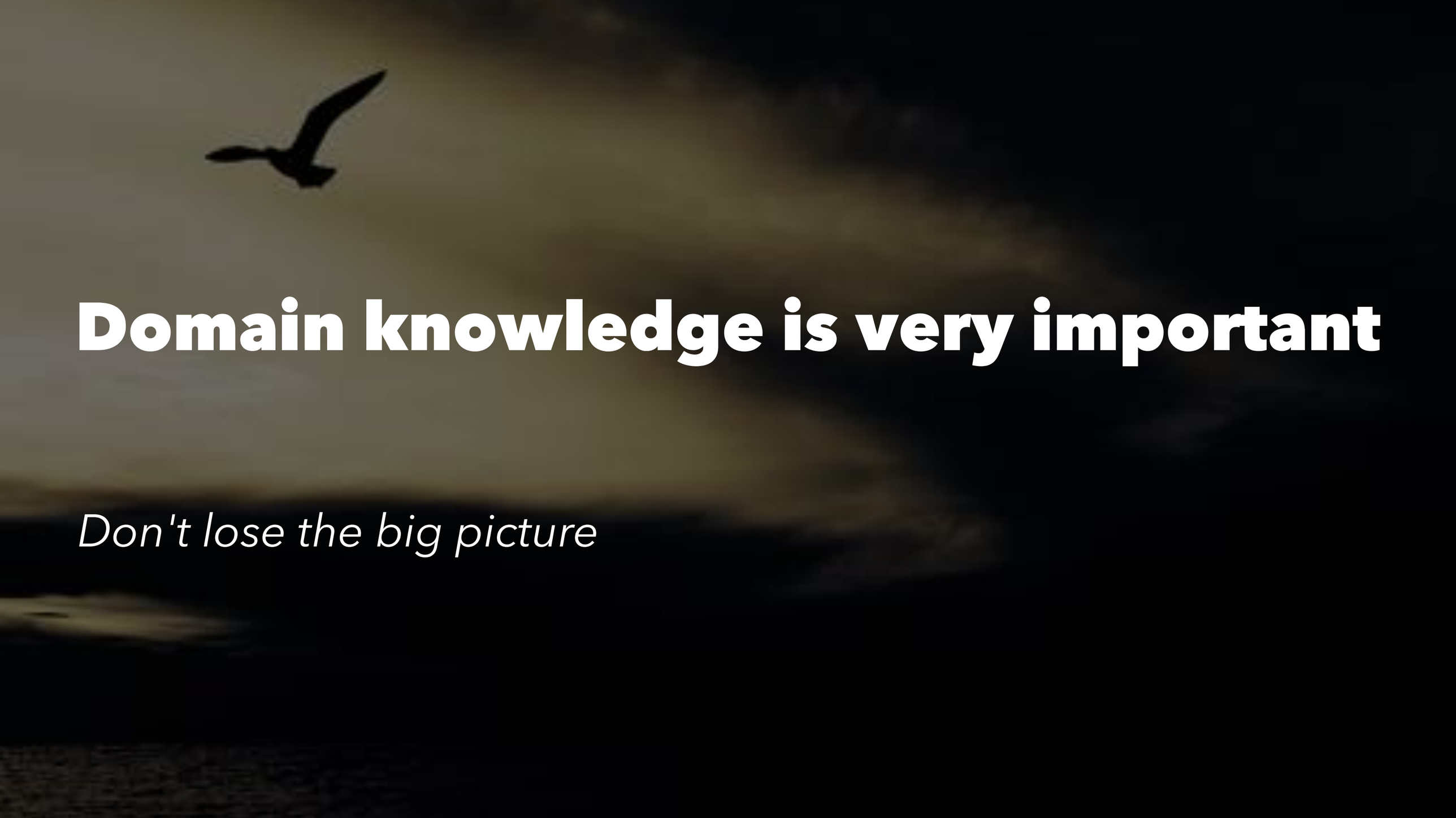
Agenda


5. Text Mining / NLP
Corpus
Text Analytics
Tokenization
Term Document Matrix
Term Frequency & Weight
Support Vector Machines
Association Rules
Market Basket Analysis
Feature Extraction
Using Mahout
Using Weka
Using NLTK
Clustering
Hierarchical Clustering
K-means Clustering
Neural Networks
Sentiment Analysis
Collaborative Filtering
Perceptron
Regression
Linear Regression
Ranking
Logistic Regression
K-Nearest Neighbor
Naive Bayes Classifiers
Boosting
Decision Trees
Classification Rule
Trees & Classification
Bias & Variance
Overfitting
SVM
Prediction
Clustering
Training & Test Data
Concepts, Inputs & Attributes
Unsupervised Learning
Supervised Learning
Categorical Var
Numerical Var
Euclidean Distance
Linear Fit
Classification
Pattern Clust
Clustering
4. Machine Learning
Author: Swaminathan Sundrasekaran
1. Fundamentals
Matrices & Linear Algebra Fundamentals
Hash Functions, Binary Tree, O(n)
Relational Algebra, DB Basics
Inner, Outer, Cross, Theta Join
CAP Theorem
Tabular Data
Data Frames & Series
Latency
Sharding
OLAP
Prob Den Fn (PDF)
ANOVA
Skewness
Continuous Distributions (Normal, Poisson, Gaussian)
Cumul Dist Fn (CDF)
Random Variables
Bayes Theorem
Probability Theory
Percentiles & Outliers
Histograms
Exploratory Data Analysis
Central Limit Theorem
Monte Carlo Method
Hypothesis Testing
P-Value
Chi² Test
Estimation
Confidence Int (CI)
MLE
Linear Discriminant Analysis
Regression
Correlation
Data Frames
Reading CSV Data
Reading Raw Data
Subsetting Data
Manipulate Data Frames
Functions
Lists
Factors
Arrays
Matrices
Vectors
Variables
Tableau
IBM Manufacture
Informix
SQL
Decision Tree
Job & Task Tracker
Map Programming
Sqoop: Loading Data in HDFS
Flume: Buffer For Unstructured Data
SQL with Pig
DBM with Java
Avro: Chunks For Memory
Using Mahout
Name & Data Nodes
Setup Hadoop (IBM / Cloudera / HortonWorks)
Data Replication Principles
Zookeeper Avro
Storm: Hadoop Realtime
Rhadoop, RHIPE
Rapid Miner
IBM SPSS
6. Visualization
Data Exploration in R (Hist, Boxplot etc)
Uni, Bi & Multivariate Viz
ggplot2
Histogram & Pie (Uni)
Tree & Tree Map
Scatter Plot (Bi)
Line Charts (Bi)
Spatial Charts
Survey Plot
Timeline
10. Toolbox
MS Excel w/ Analysis ToolPak
Java, Python
R, R-Studio, Rattle
Weka, Knime, RapidMiner
Hadoop Dist of Choice
Spark, Storm
Flume, Scribe, Chukwa
Hulch, Talend, Scraperwiki
Webscraper, Flume, Sqoop
Im, IRWeka, NLTK
RHIPE
Using Mahout
7. Data Ingestion
Using ETL
How much Data?
Google OpenRefine
Data Survey
Stratified Sampling
Principal Component Analysis
Transformation & Enrichment
Data Fusion
Data Integration
Data Sources & Acquisition
Data Discovery
Summary of Data Formats

Be Curious | Have Fun

How many birds remain on the tree?

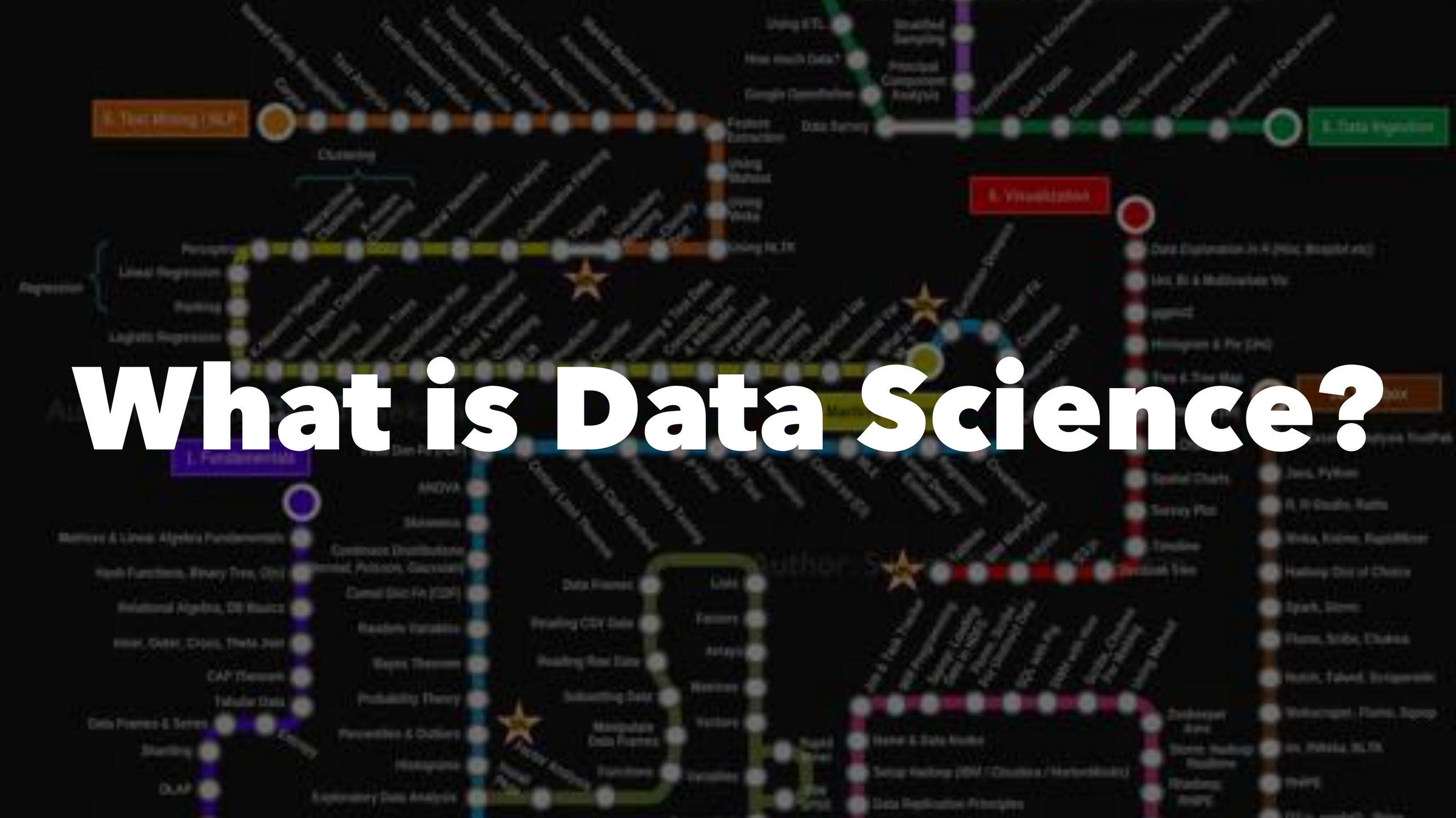
- P(Hunter to hit target) = 0.2
- Number of birds n = 150
- Shots = 3
- Birds hit = 1

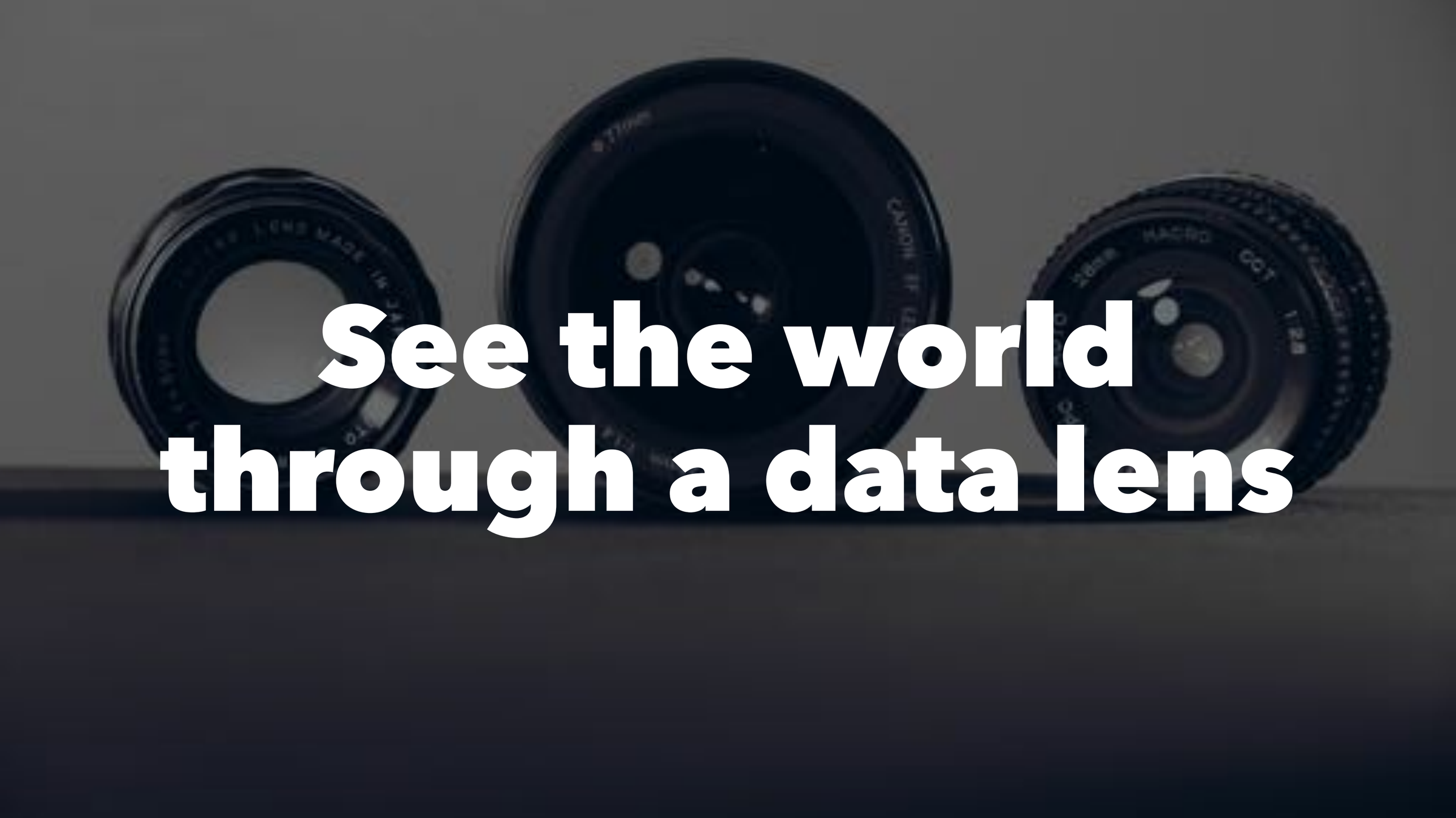


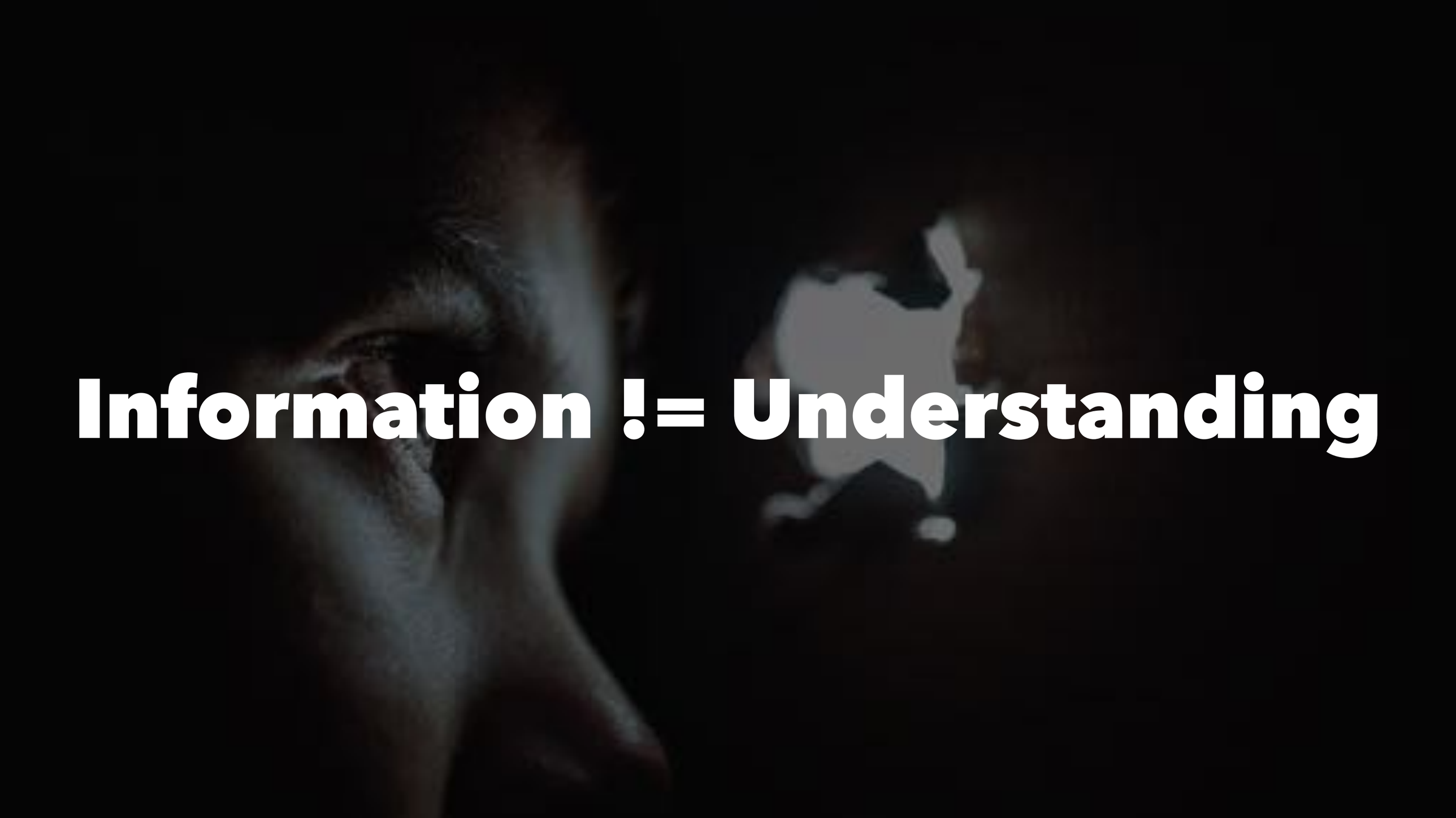
Domain knowledge is very important

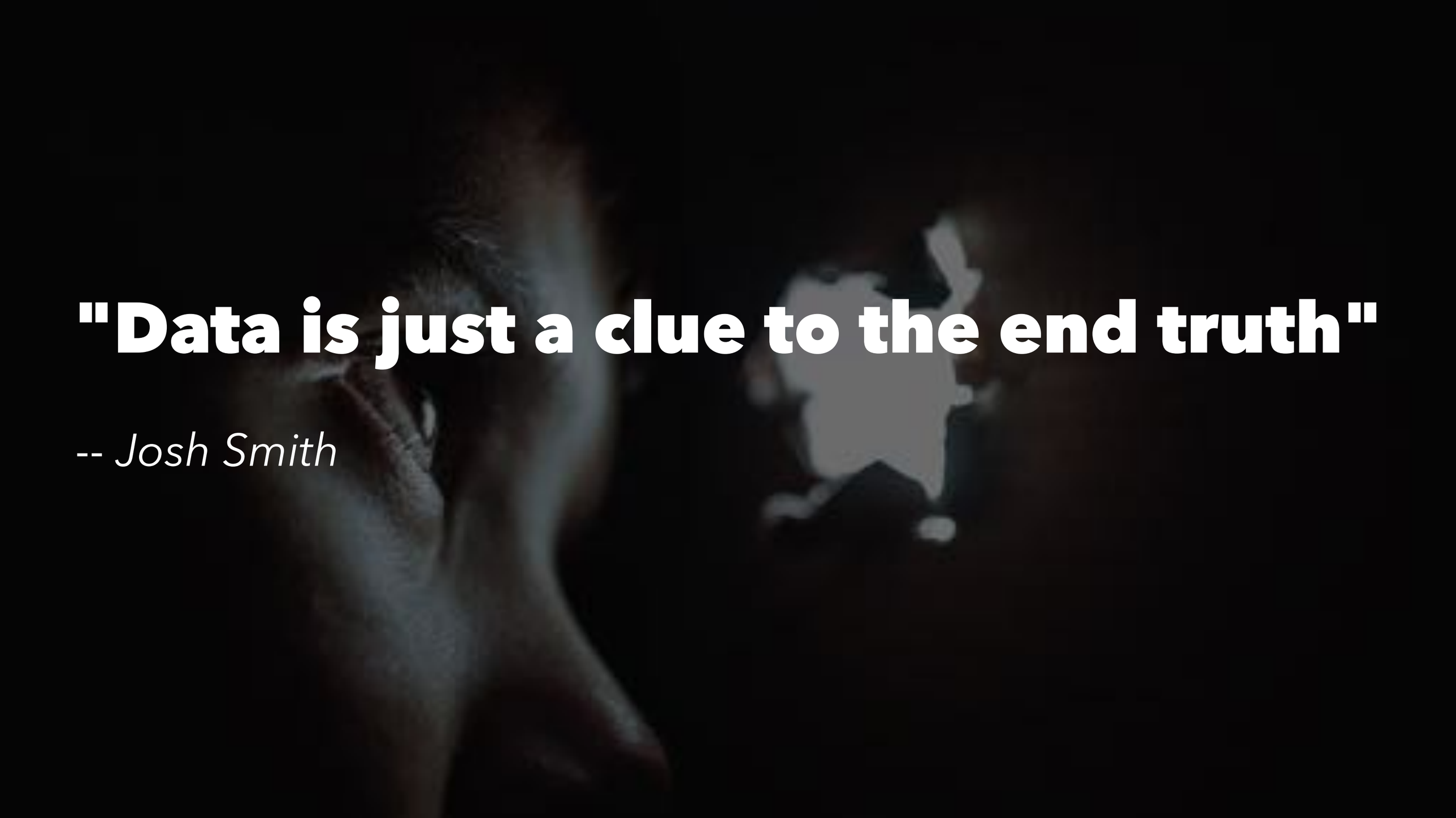
Don't lose the big picture

What is Data Science?




See the world
through a data lens


Information != Understanding


"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"

Frame

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

Types of Questions

- Descriptive
- Exploratory
- Inferential
- Predictive
- Causal
- Mechanistic

Descriptive

How many people signed up for today's event?

Exploratory

How many people turned up from each company ?

Inferential

How many managers signed up and amongst them how many turned up? Is there a correlation with the company they come from?

Predictive

Given that Ms. X has come for the previous workshop, will she turn up for today's workshop?

Causal

Why did more people from a particular company turn up for today's workshop?

Mechanistic

If more managers from a particular company turn up, how many more employees will turn up for the workshop?

Acquire

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

Acquire

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

Refine

"All data is messy."

Refine

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

Explore

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

Explore

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

Model

"All models are wrong, but some are useful"

Model - Supervised Learning

- *Continuous: Regression*
- *Discrete: Classification*

Model - Supervised Learning

- *Continuous: Regression* :
  What will be the annual revenue for 2017 ?

Model - Supervised Learning

- *Discrete: Classification* :
  Will company XYZ buy from us?

Model - Supervised Learning

- *Continuous: Regression*
- *Discrete: Classification*

Algorithms:

Linear Regression, Logistic Regression, CART, Random Forest, Gradient Boosting Machines, K-Nearest Neighbor, Support Vector Machines, Naive-Bayes, Bayesian Networks

Model - Unsupervised Learning

- *Cluster Analysis*
- *Dimensionality Reduction*

Model - Unsupervised Learning

- *Cluster Analysis* :
  If we segment our customers into three types, what would they look like?

Model - Unsupervised Learning

- *Dimensionality Reduction* :
  Data is too huge to load into memory. Is there a better representation of the data?

Model - Unsupervised Learning

Cluster Analysis: K-means, DBSCAN

Dimensionality Reduction: Principal Component Analysis, Singular Value Decomposition, MDS

Model - Advanced / Specialized

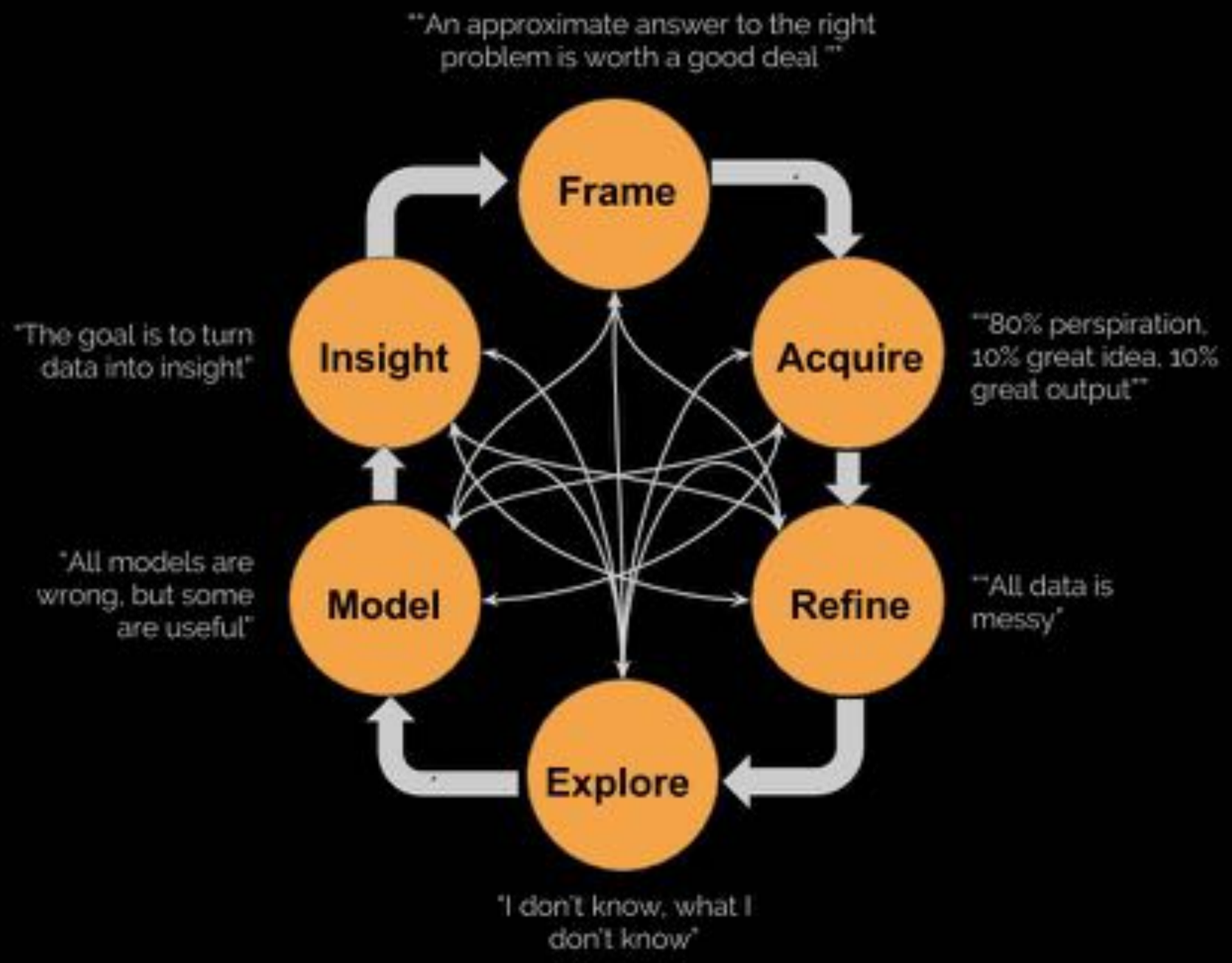
- Deep Learning
- Network / Graph Analytics
- Optimization
- Reinforcement Learning
- Online Learning
- Applications: Time Series, Text, Image, Speech

Insight

"The goal is to turn data into insight"

Insight

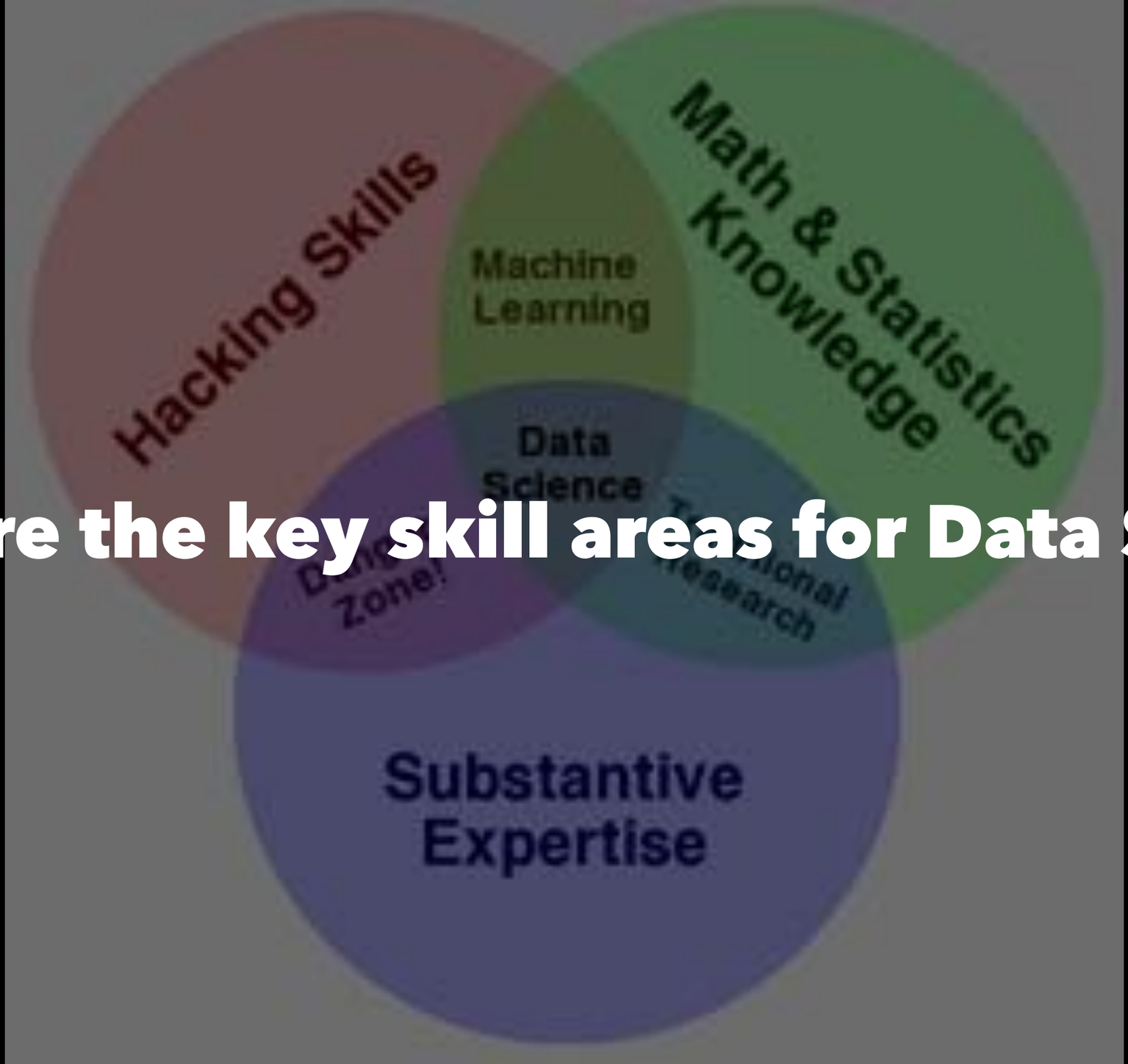
- Narrative Visualisation
- Dashboard Visualisation
- Decision Making Tools
- Automated Decision Tools

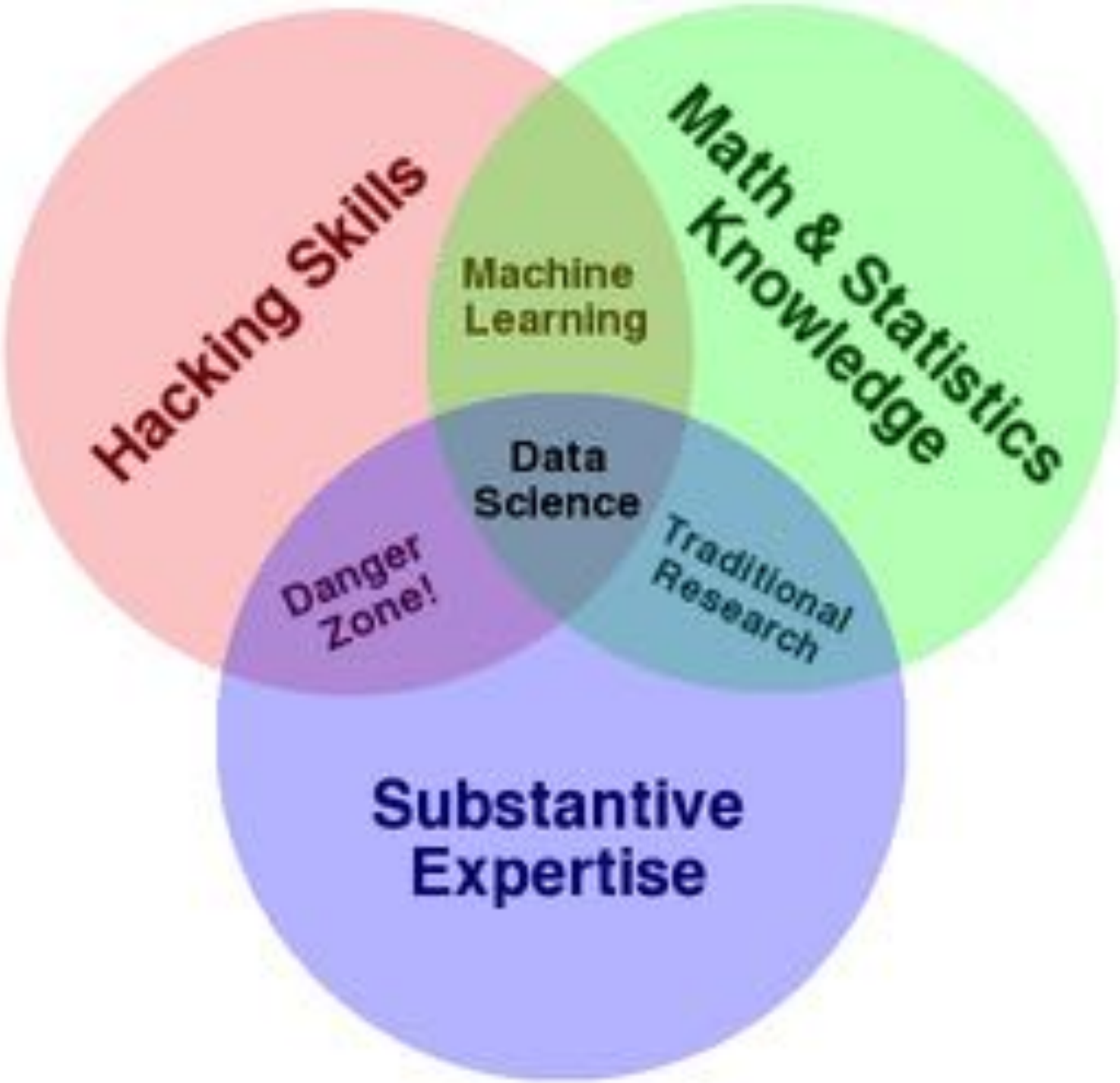

"An approximate answer to the right problem is worth a good deal"
Frame
Acquire
"80% perspiration, 10% great idea, 10% great output"
Refine
"All data is messy"
Explore
"I don't know, what I don't know"
Model
"All models are wrong, but some are useful"
Insight
"The goal is to turn data into insight"

"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

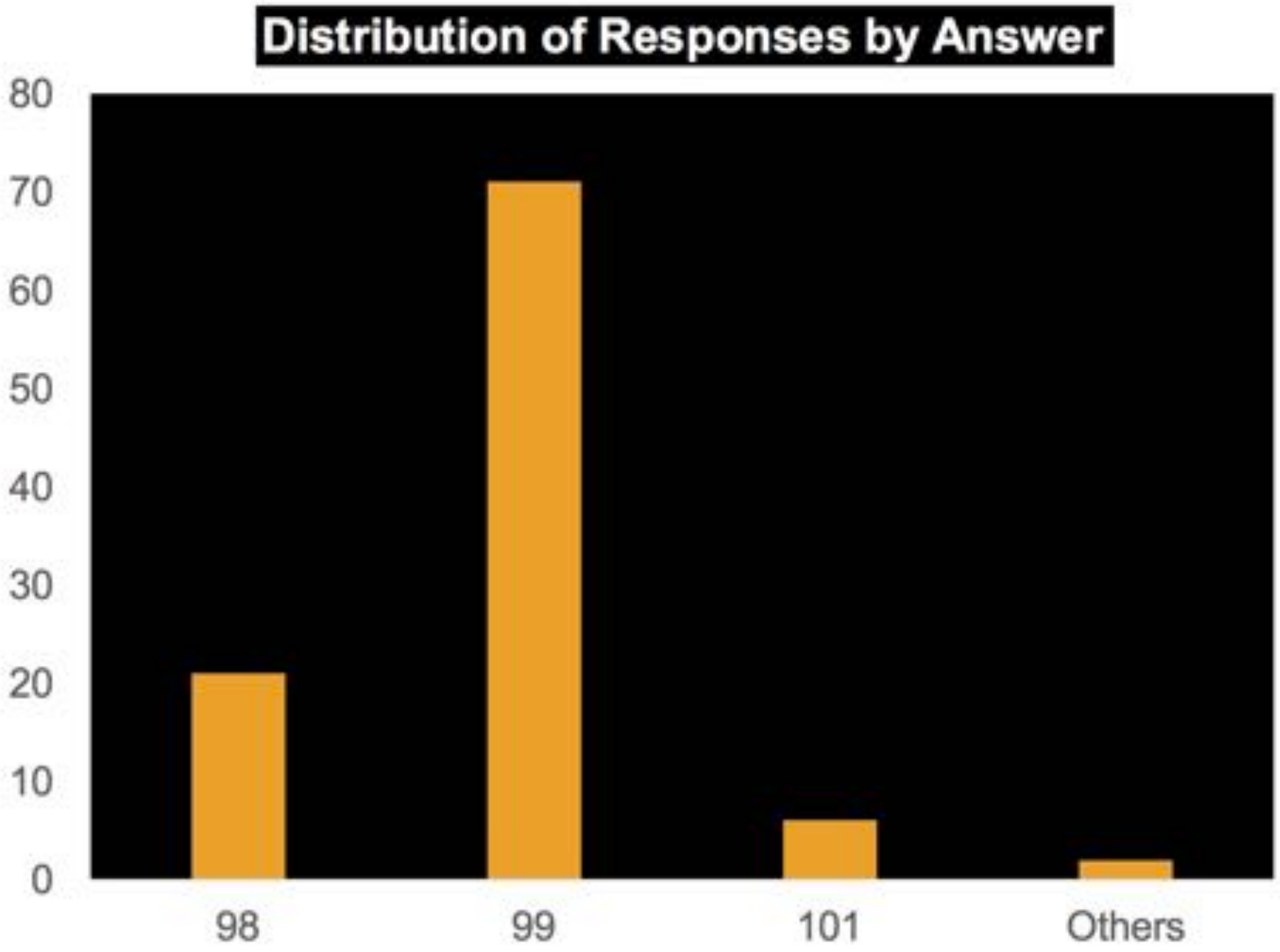
What are the key skill areas for Data Science




Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise

SOLVE THE PUZZLE

3	2	=	7
5	4	=	23
7	6	=	47
9	8	=	79
10	9	=	?


Distribution of Responses by Answer
80
70
60
50
40
30
20
10
0
98
99
101
Others

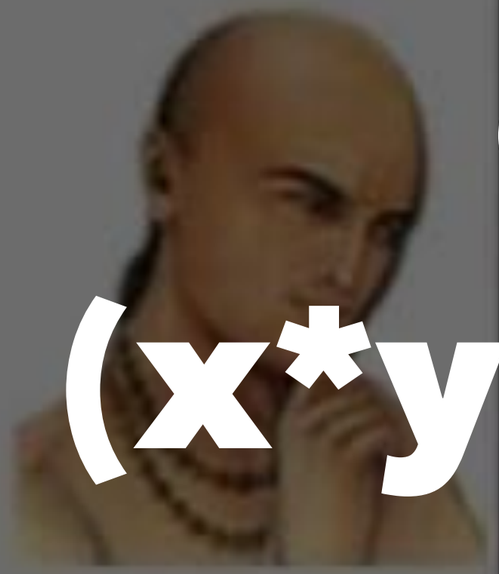
SOLVE THE PUZZLE

3	2	=	7
5	4	=	23
7	6	=	47
9	8	=	79
10	9	=	?

99

(x*y + odd counter)

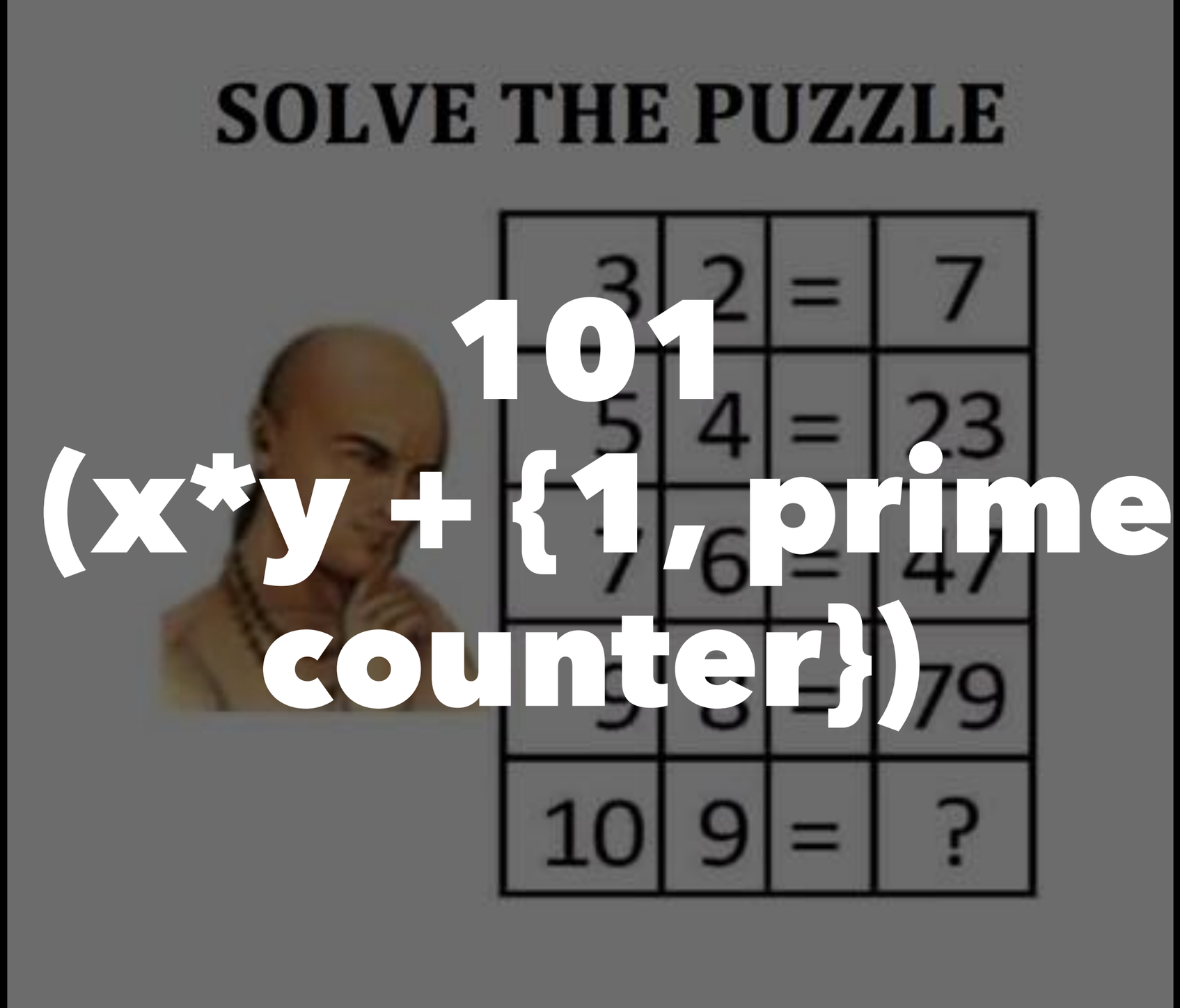
SOLVE THE PUZZLE



3	2	=	7
5	4	=	23
7	6	=	47
9	8	=	79
10	9	=	?

98

(x*y + y - 1)

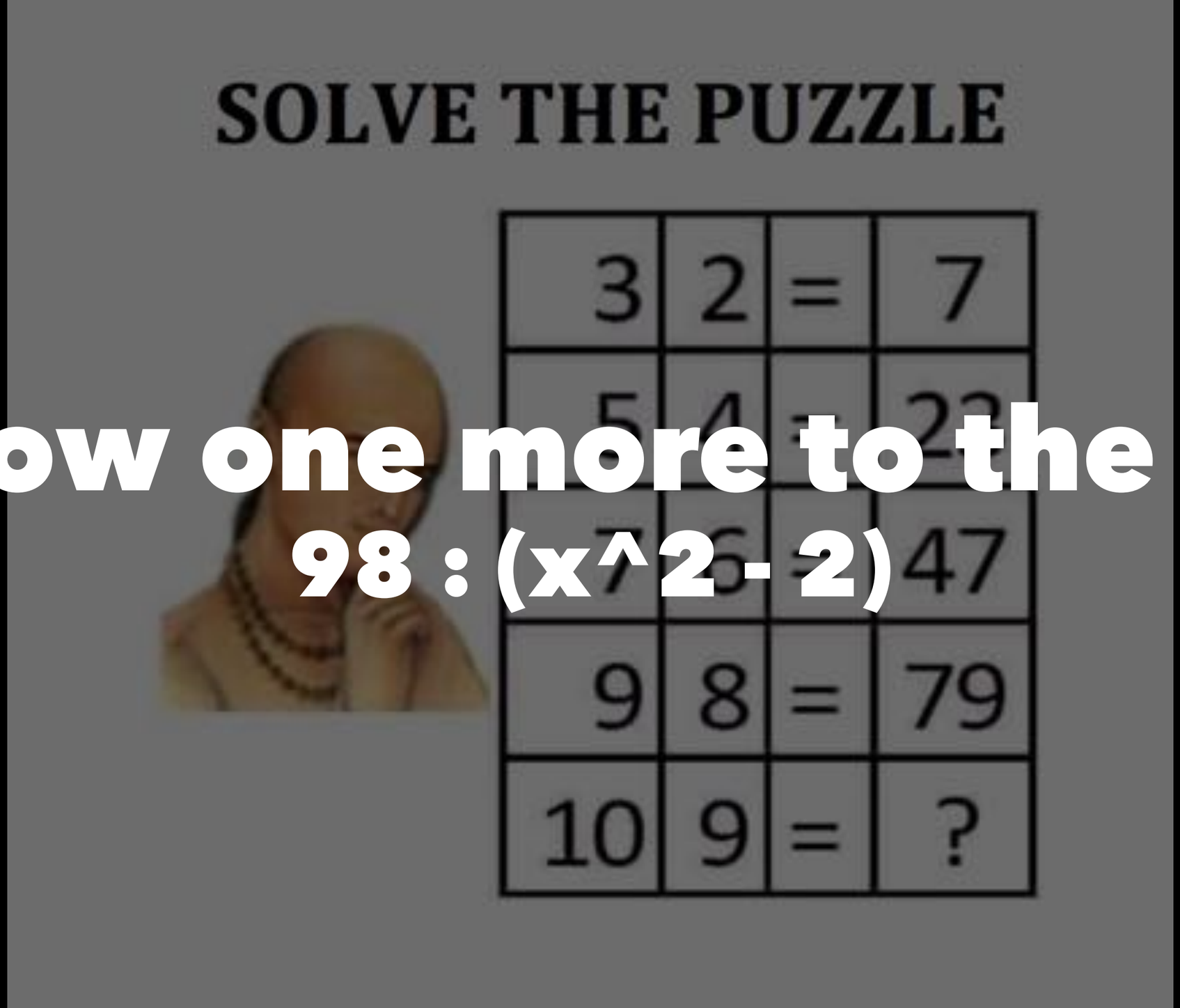


101
(x*y + {1, prime counter})


Stakeholder/Client wants an answer
Wisdom of crowds?

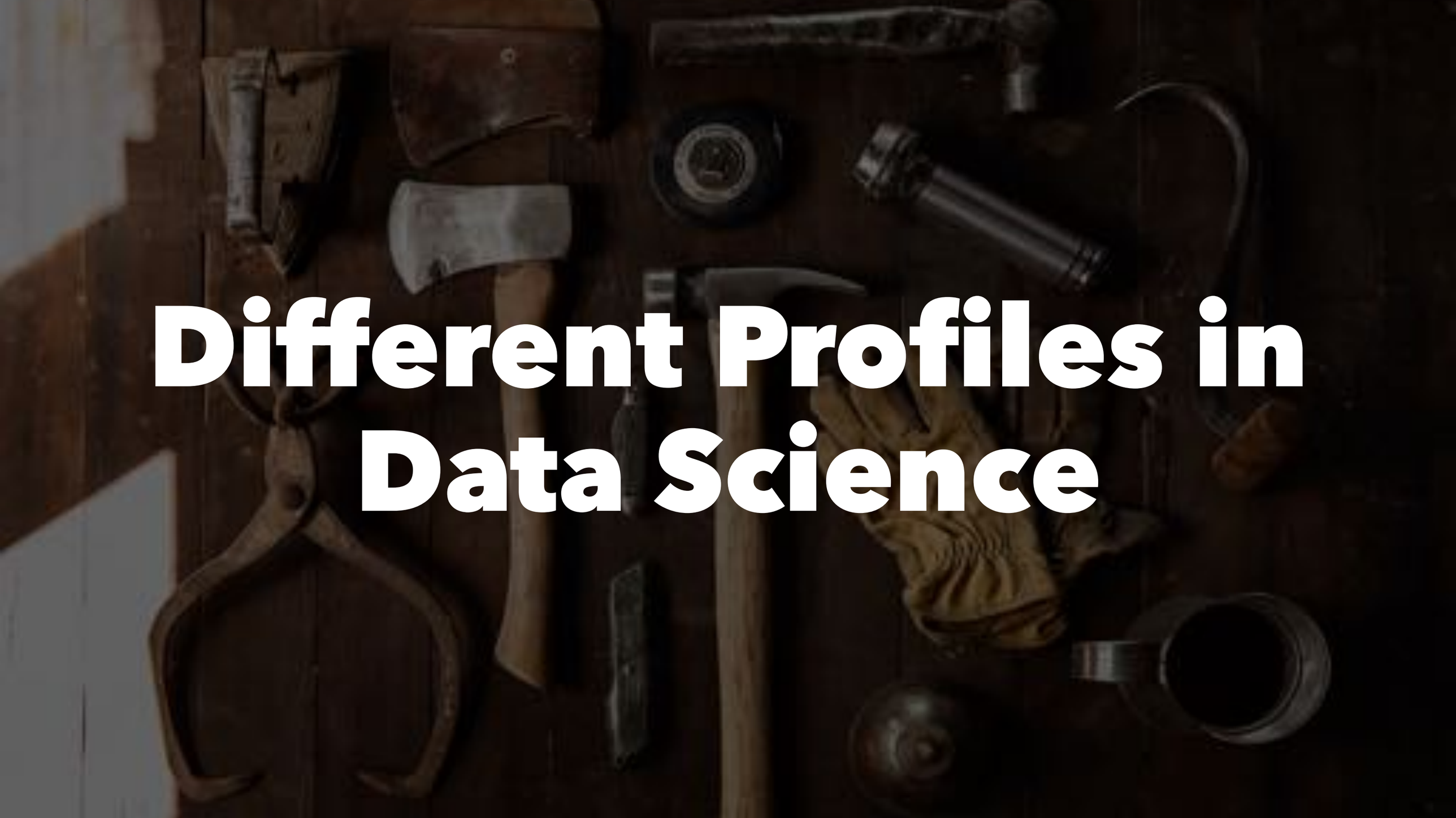
Throw one more to the fire

98 : (x^2- 2)



Occam's Razor

Problem solving principle: Amongst competing hypotheses, the one with the fewest assumptions should be selected



Different Profiles in Data Science

Different Profiles in Data Science

- Data Analyst
- Data Engineer
- Data Visualization
- Data SME
- Data Scientist

Data Analyst

A budding/junior data scientist. Supports EDA and data wrangling.

Typical skills:

- Good with Excel.
- Basic knowledge of R/Python/SQL

Data Engineer

Builds and supports the data pipeline. Data Architect.

Typical skills:

- SQL
- Spark
- Hadoop/Cassandra
- Data Orchestration(Eg: Luigi)

Data Visualization

Builds visualization

Typical skills:

- Tableau/Qlik/D3.js
- Basics of HTML/JavaScript

Data SME

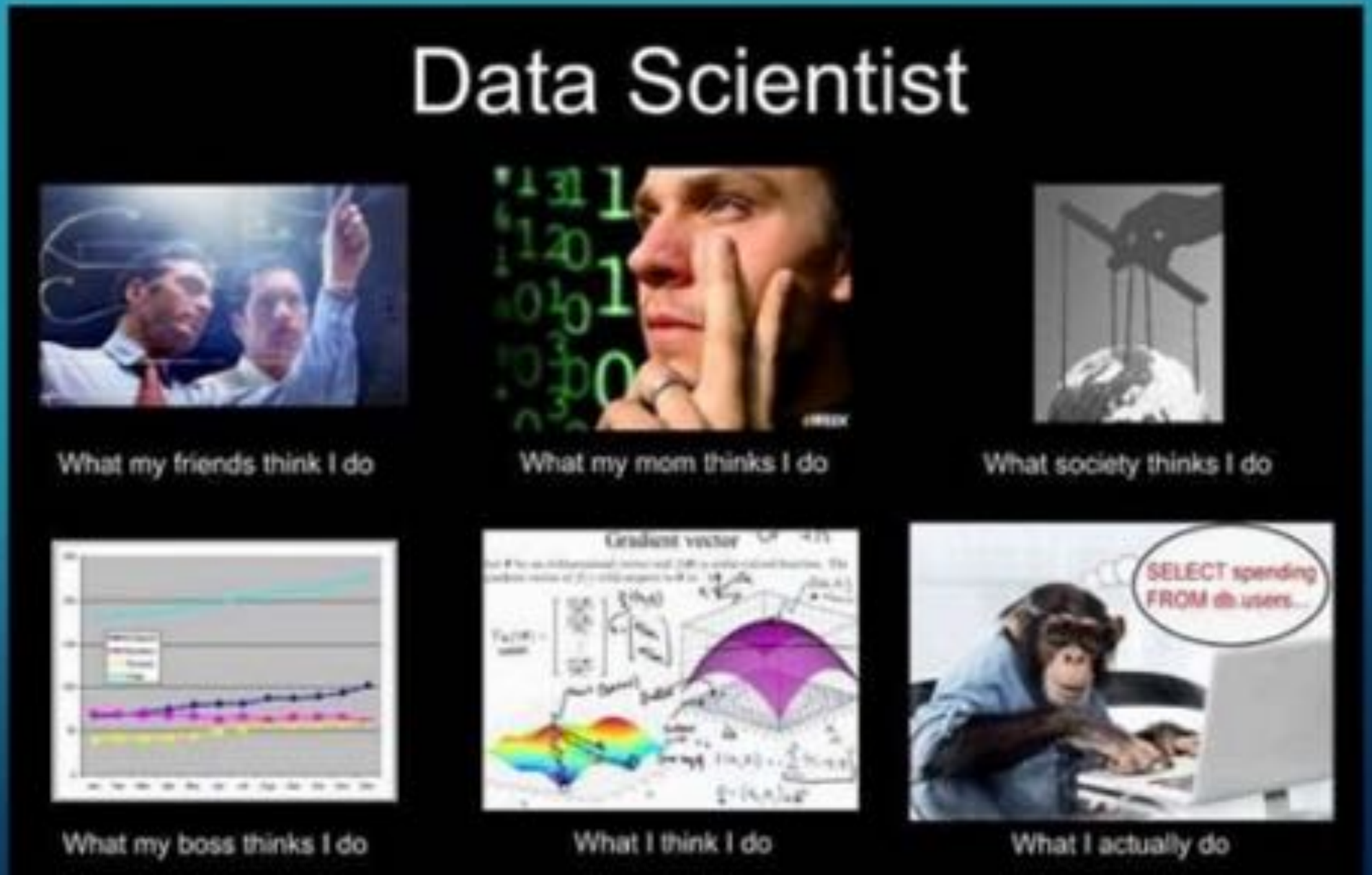
The data guru. Understands business impact of each of the attribute stored in the system.

Typical skills:

-Domain knowledge + system architecture

Data Scientist

SO WHAT DOES A DATA SCIENTIST DO?



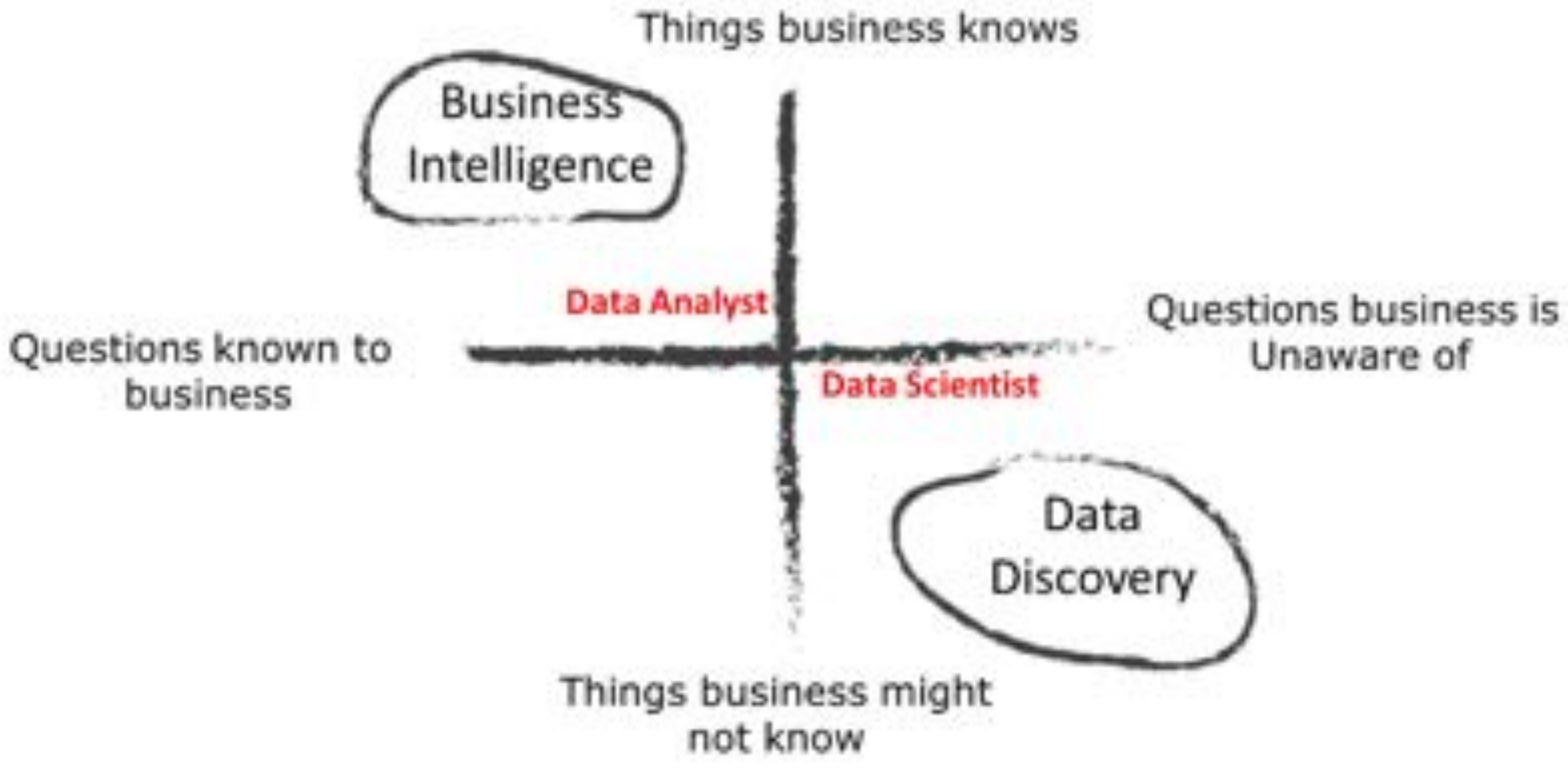
Data Scientist

Builds models and find insights.

Typical skills:

- R
- Python
- Spark
- Machine Learning
- Math/Statistics

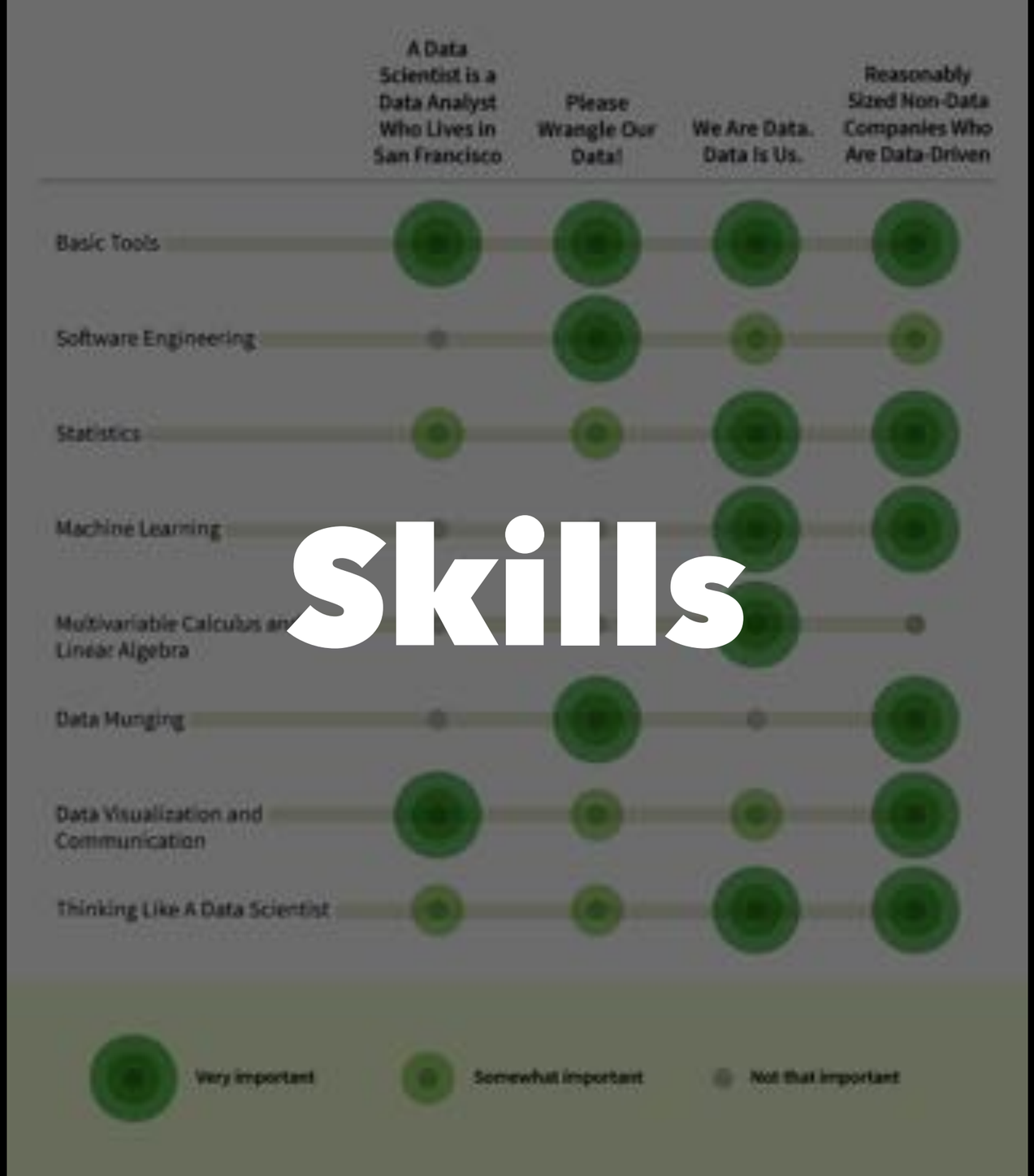
Data Analyst
Vs
Data Scientist?

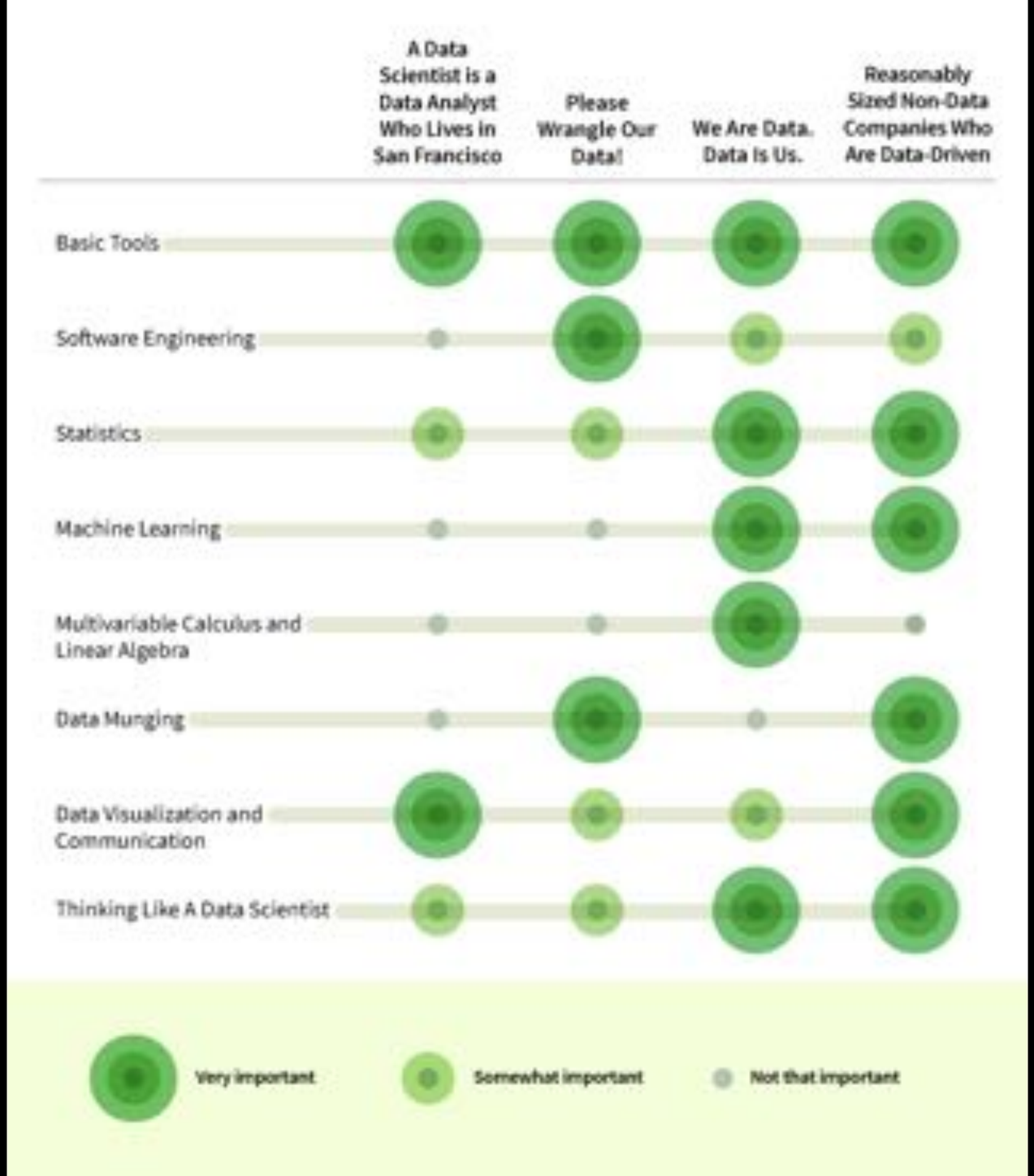

Things business knows
Business Intelligence
Data Analyst
Questions known to business
Questions business is Unaware of
Data Scientist
Data Discovery
Things business might not know


Chicago Daily Tribune
DEWEY DEFEATS TRUMAN

Selection Bias !

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


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

R Stack

- **Acquire**: `rvest, XML, jsonlite, httr, RSQLite, RPostgreSQL, readxl, haven, readr, data.table`
- **Refine**: `dplyr, tidyr, lubridate, stringr`
- **Explore**: `graphics, ggplot2, ggvis, ggmap, map, vcd, rgl, htmlwidgets, leaflet, choroplethr, plotly`
- **Model**: `stats, caret, ranger, glmnet, xgboost, party, mxnet, forecast`
- **Insight**: `OpenCPU, Rserve, shiny, RMarkdown, knitr`

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 day in the life of a Data Scientist

$\sum_{k=1}^{n} p_k$ $\quad y=\varphi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}dt$ $\quad S(\alpha,T)=\frac{2}{\pi}\int_0^{\infty}\frac{\sin^2 t}{t}dt$ $\quad P(\eta_\infty<x)=F(x)$ $\quad \lim_{n\to\infty}\binom{2n}{n}$

$W_k=\binom{n}{k}p^k(1-p)^{n-k}$ $\quad P(\eta<y|\xi=x)=\sup_{y'<y,\,y'\in\mathbb{Q}} P(\eta<y'|\xi=x)$

$S_n=A_n\cup\pi A_n$ $\quad f(t|y)=\frac{2e^{\frac{y^2}{t}}}{\sqrt{2\pi}}\int_{-\frac{y}{\sqrt{t}}}^{+\infty}\frac{e^{-\frac{u^2}{t}}du}{\left(1-\frac{y^2}{u^2}\right)^{\frac{1}{2}}}$ $\quad \Delta N=\sum_{k=1}^{N}\frac{\xi_k}{k}$

$|A_n|=\frac{n!}{2}\left|\int_{|x|>A} f(x)\log_2\frac{1}{f(x)}dx\right|<\varepsilon$ $\quad g^{-1}\cdot g=e$ $\quad \zeta=\sqrt{\frac{2n}{\nu_n}}\left(\frac{\eta_{2n}}{\sqrt{2n}}+\frac{\eta_{2n}-\eta_{2n}}{\sqrt{2n}}\right)$

$\sum_{k=0}^{+\infty} e^{-\frac{k^2\pi^2}{2t}}=H(t)$ $\quad \prod_{k\le b}$; $\bigcup_{i=1}^{n-1}H_i$; $\bigcap_{n=0}^{\infty}X_n$ $\quad H_r(x)=\frac{G_r(x)}{1+G_r(x)}$

$\int_{-\infty}^{+\infty} dG_k(x)\ge\frac{1}{2}$ $\quad f_n(t)=\frac{2^n t^{n-1}e^{-2t}}{(n-1)!}$ $\quad U_n^{+}(0)=\binom{2n}{n}-\binom{2n}{n-c}$

$f_{n-r}(t)=\int_0^t f_n(u)f_r(t-u)du=\frac{2^{n+1}t^n e^{-2t}}{n!}$ $\quad \lim_{t\to 0}(\ldots)=0$ $\quad \lim_{n\to\infty}\frac{\zeta(n)}{n}=P_k$ $\quad R=\int_{-\infty}^{\infty}\varphi(t)dt$ $\quad \left|\frac{\sin tx}{tx}[\varphi(t)e^{-itx}+\varphi(-t)e^{\ldots}\right.$

$\log\varphi(t)=i\gamma t-c|t|^{\alpha}\left[1+i\beta\frac{t}{|t|}\omega(t,\alpha)\right]$ $\quad B(v)=\sum_{k=1}^{r}\varphi^{*}(k_k v)$ $\quad c_{iv}=\sum_{j=1}^{n}a_{ij}b_{jv}$ $\quad \lim_{n\to\infty}P\left(\frac{\zeta_{n+1}-k(n)-\log\frac{1}{q}}{\sqrt{\frac{1-q}{q}}}\right)$ $\quad C_n(\alpha)\ge\frac{n!}{\prod_{k=1}^{r}n_k(\alpha)!}$ $\quad \lim_{m}\varphi(t)=\varphi\left(c\left(\frac{m}{m}\right)t\right)$

$\int_{-\infty}^{\infty}e^{-\frac{u^2}{2}}du=F(x)\left(\frac{1}{\sqrt{2\pi}}\right)^{-1}$ $\quad |\varphi_\xi(t)|=\left|\int_{-\infty}^{\infty}e^{itx}dF(x)\right|\le\int_{-\infty}^{\infty}e^{-vx}dF(x)=\varphi_\xi(iv)$ $\quad q_k(\alpha)=\frac{p_k^{\alpha}}{\sum_{j=1}^{r}p_j^{\alpha}}$ $\quad P(\Pi_2=$

$\Gamma_m=\Gamma_r\Gamma_{m-r}$ $\quad g^{-1}Ng=\{g^{-1}ng\,|\,n\in N\}$ $\quad Q=F^{-1}(q)$

$|X\cup Y|=|X|+|Y|-|X\cap Y|$ $\quad \lim_{n\to\infty}\frac{1}{\sqrt{n}}k_n\left(\frac{x}{\sqrt{n}}\right)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$ $\quad P_n(k_k)=\frac{C_n^{(k)}}{\ldots}$ $\quad P\left(\lim_{n\to\infty}\sup\frac{|h_n|}{\sqrt{2n\log\log n}}\le 1\right)=1$ $\quad \varphi(t)=1-\sqrt{1-e^{2vt}}$

$f: X\to X\cap W$

$\chi(A)=\int_A\chi(\omega)dP$ $\quad \ell'(\alpha)=-\log 2\left(\frac{\sum_{k=1}^{r}p_k^{\alpha}\log_2\frac{1}{p_k}}{\sum_{k=1}^{r}p_k^{\alpha}}-\left(\frac{\sum_{k=1}^{r}p_k^{\alpha}\log_2\frac{1}{p_k}}{\sum_{k=1}^{r}p_k^{\alpha}}\right)^2\right)$ $\quad fg(u_i)=f\left(\sum_{j=1}^{\dim V_k}a_{ji}v_j\right)=\sum_{j=1}^{\dim V_k}a_{ji}\left(\sum_{k=1}^{\dim W_k}b_{kj}w_k\right)$ $\quad \frac{\binom{2k}{k}}{2^{2k}}\approx\frac{1}{\sqrt{\pi k}}$

$q\left(e^{-x}\sqrt{\frac{1-q}{nq}}-1\right)=x\sqrt{\frac{q(1-q)}{n}}+o\left(\frac{1}{n}\right)$ $\quad \prod_{k=1}^{r}\left[g_k\left(\frac{t}{\sqrt{n}}\right)\right]^{N_k\alpha_k}=e^{-\frac{t^2}{2}}$ $\quad P_{jk}^{(m)}=\sum_{t=0}^{\infty}P_{jt}^{(r)}P_{tk}^{(m-r)}$ $\quad \frac{1}{2\pi}\int_{-\infty}^{\infty}\mathrm{Re}\left\{\varphi(t)\frac{e^{itx}-e^{-itx}}{it}\right\}dt$ $\quad P(|\alpha_n|\ge t)\le\frac{C_\varphi}{\log N}$

$\liminf_{N\to\infty}\int_{-\infty}^{+\infty}f_N(x)^{\alpha}dx\ge\int_{-\infty}^{+\infty}f(x)^{\alpha}dx$ $\quad M(|\delta_j-1|^3)=\int_{-\infty}^{\infty}|x-1|^3e^{-x}dx$ $\quad \lim_{N\to\infty}\int_{-A}^{+A}f_N(x)\log_2\frac{1}{f_N(x)}dx=\int_{-A}^{A}f(x)\log_2\frac{1}{f(x)}dx$ $\quad N\varepsilon_n-\varepsilon_k=\binom{2n}{n+k_c}=\binom{2n}{n-}$

$$\log \varphi(u) = -\frac{\sigma^2 u^2}{2}$$

$$y = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt$$

$$S(\alpha, T) = \frac{2}{\pi} \int_0^{\pi} \frac{\sin \alpha t}{t}\, dt$$

$$P(\eta_\infty < x) = F(x)$$

$$\binom{n}{k} p^k (1-p)^{n-k}$$

$$P(\eta < y \mid \xi = x) = \sup_{y' < y, y' \in Q} P(\eta < y' \mid \xi = x)$$

$$\left| \log_2 \frac{1}{f(x)}\, dx \right| < \varepsilon$$

$$g^{-1} \cdot g = e$$

$$\zeta = \sqrt{\frac{2\lambda_n}{\nu_n}} \left(\frac{\eta_{2n}}{\sqrt{2\lambda_n}} + \frac{\eta_{2n} - \eta_{2n}}{\sqrt{2\lambda_n}} \right)$$

$$f(t \mid y) = \frac{2e^{\frac{x^2}{2}}}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \frac{e}{(1}$$

$$e^{-\frac{\lambda^2 \sigma^2}{2}} = H(\lambda)$$

$$\prod_{k \le b}; \quad \bigcup_{i=1}^{n-1} H_i; \quad \bigcap_{n=0}^{\infty} X_n$$

$$f_n(t) = \frac{2^{-\frac{1}{2}} n^{-1} e^{-2t}}{(n-1)!}$$

$$H_r(x) = \frac{G_r(x)}{1 + G_r(x)}$$

$$(t-u)\, du = \frac{2^{n+1} t^n e^{-2t}}{n!}$$

$$\lim_{t \to \infty} (\ldots) = 0$$

$$\lim_{n \to +\infty} \frac{\xi(n)}{n} = P_k$$

$$R = \int_{-\infty}^{\infty} \varphi(t)\, dt$$

$$C_{iv} = \sum_{j=1}^{n} a_{ij} b_{jv}$$

$$B(u) = \sum_{k=1}^{r} \varphi^*(b_k v)$$

$$\lim_{n \to \infty} P\left(\frac{\xi_{n+1} - k(n) - \log \frac{1}{q}}{\sqrt{\frac{1-q}{q}}} \right)$$

$$C_n(\alpha) \ge \frac{n!}{\prod_{k=1}^{r} n_k(\alpha)!}$$

$$|\psi_\xi(t)| = \left| \int_{-\infty}^{\infty} e^{itx}\, dF(x) \right| \le \int_{-\infty}^{\infty} e^{-vx}\, dF(x) = \varphi_\xi(iv)$$

$$g^{-1} N g = \{ g^{-1} n g \mid n \in N \}$$

$$Q = F^{-1}(q)$$

$$\lim_{n \to \infty} \frac{1}{\sqrt{n}} k_n\left(\frac{x}{\sqrt{n}}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

$$P_n(k) = \ldots$$

$$P\left(\lim_{n \to \infty} \sup \frac{|h_n|}{\sqrt{2n \log \log n}} \le 1 \right)$$

$$fg(u_i) = f\left(\sum_{j=1}^{\dim V_k} a_{ji} v_j \right) = \sum_{j=1}^{\dim V_k} a_{ji} \left(\sum_{k=1}^{\dim W_k} b_{kj} \ldots \right.$$

$$P_{jk}^{(m)} = \sum_{r=0}^{\infty} P_{jr}^{(r)} P_{rk}^{(m-r)}$$

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{Re} \left\{ \varphi(t) \frac{e^{itx}}{\ldots} \right.$$

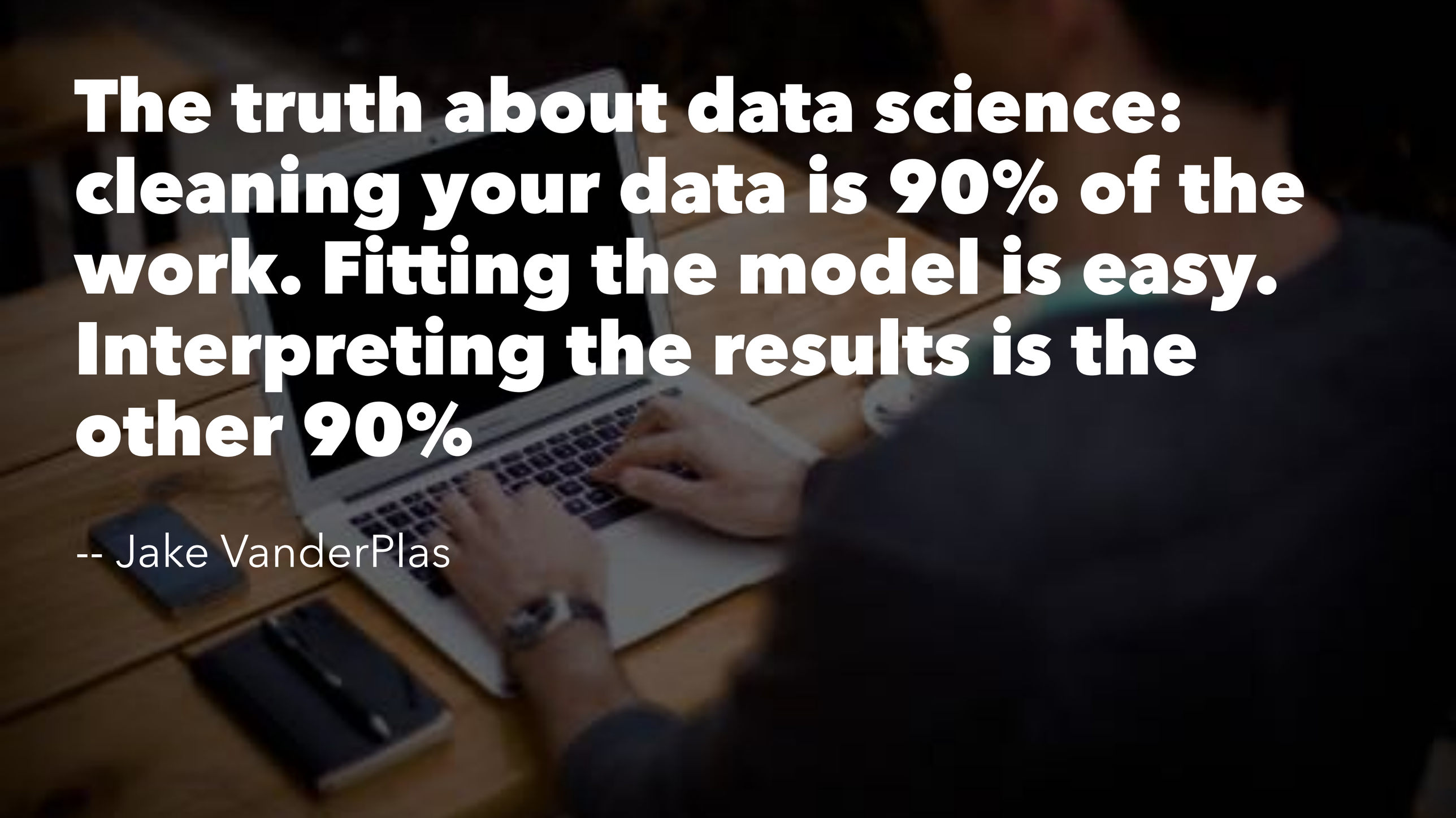
$$\prod_{k=1}^{r} \left[g_k \left(\frac{t}{\sqrt{n}} \right) \right]^{N_k \alpha_k} = e^{-\frac{t^2}{2}}$$

$$\lim_{N \to \infty} \int_{-A}^{A} f_N(x) \log_2 \frac{1}{f_N(x)}\, dx = \int_{-A}^{A} f(x) \log_2 \frac{1}{f(x)}$$

$$M(|\delta_j - 1|^3) = \int_0^{\infty} |x-1|^3 e^{-x}\, dx$$

$$\det(M') = \det(M) + \det(M'') = \det(M)$$

$$h(x,y) = \frac{1}{2\pi} \left[\sqrt{2}\, e^{-\frac{x^2}{2}} - e^{-x} \right]$$



The truth about data science: cleaning your data is 90% of the work. Fitting the model is easy. Interpreting the results is the other 90%

-- Jake VanderPlas

Key Challenges for a Data Scientist

- Data Cleaning/wrangling
- Feature Engineering
- Hyperparameter Optimization
- Insights


Questions?

Thank you