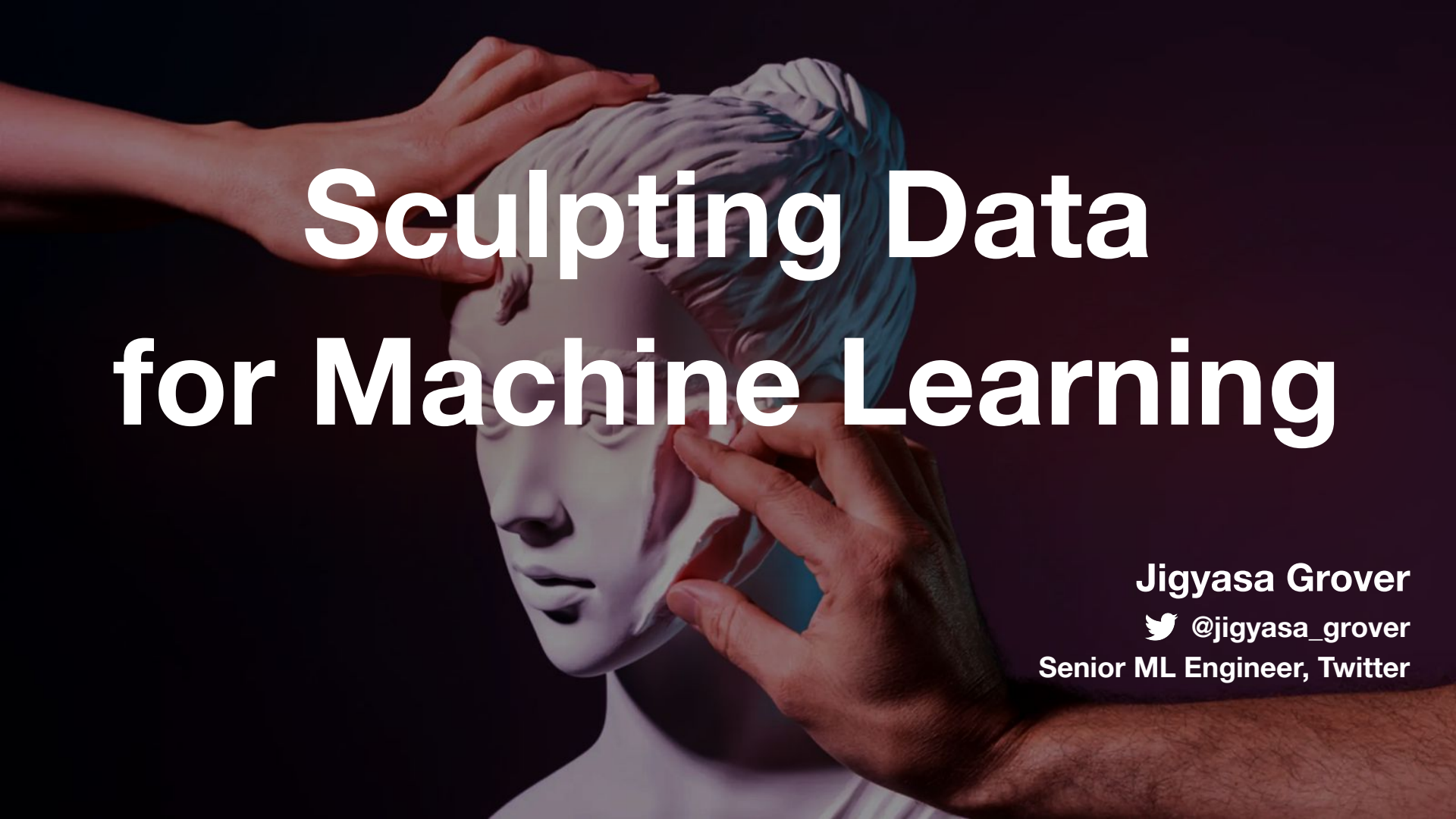


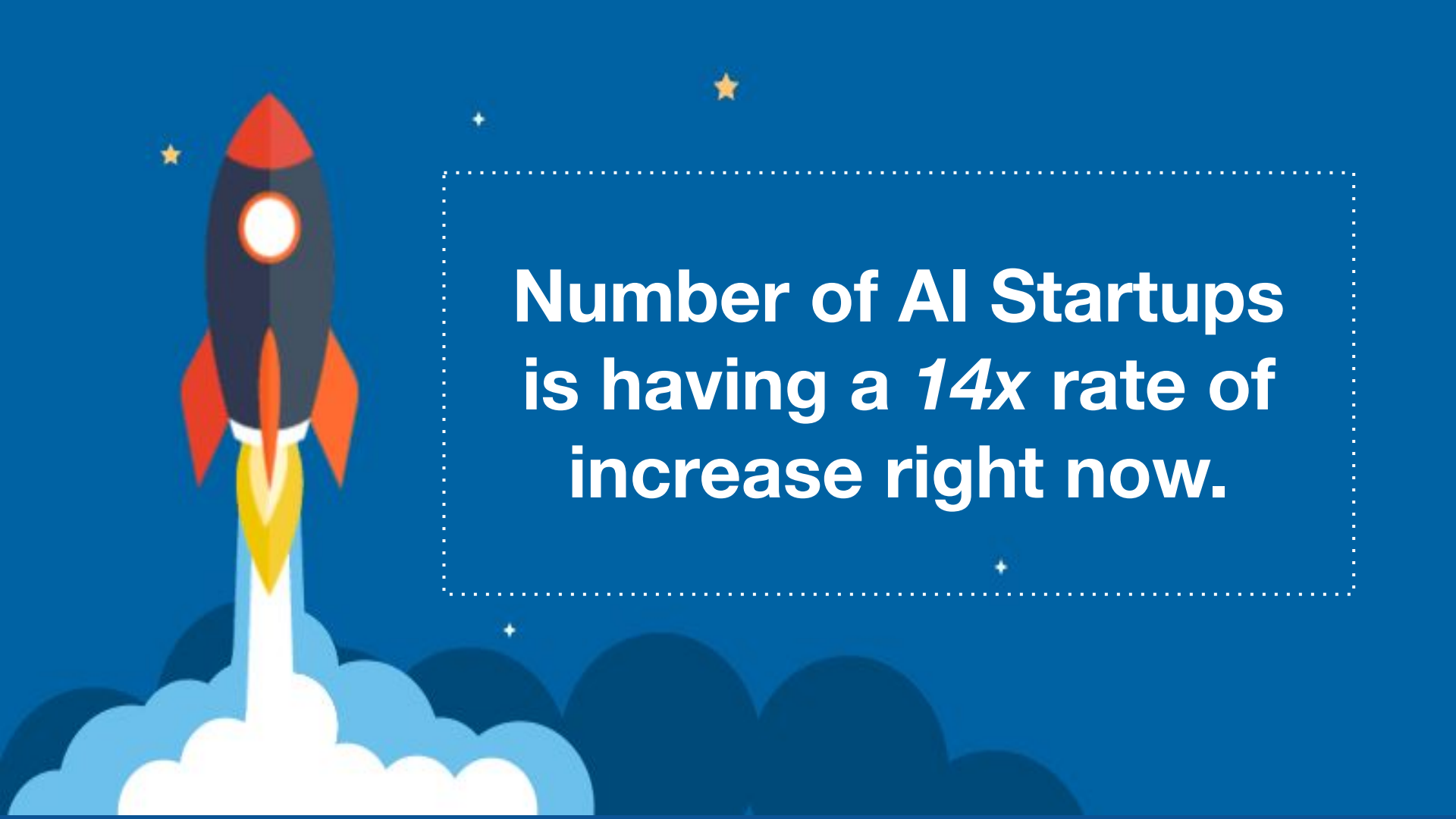
Sculpting Data for Machine Learning

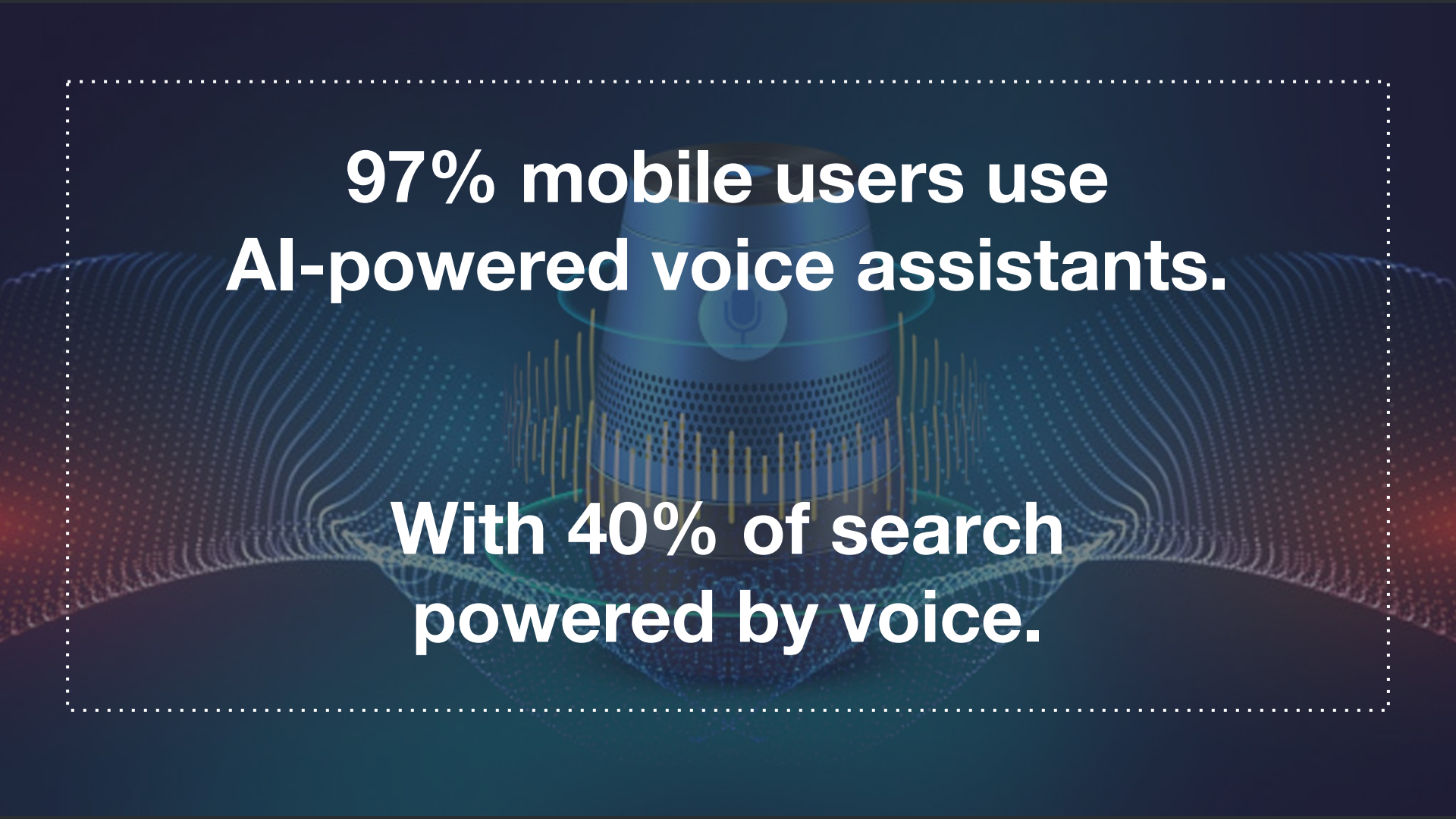
Jigyasa Grover
@jigyasa_grover
Senior ML Engineer, Twitter

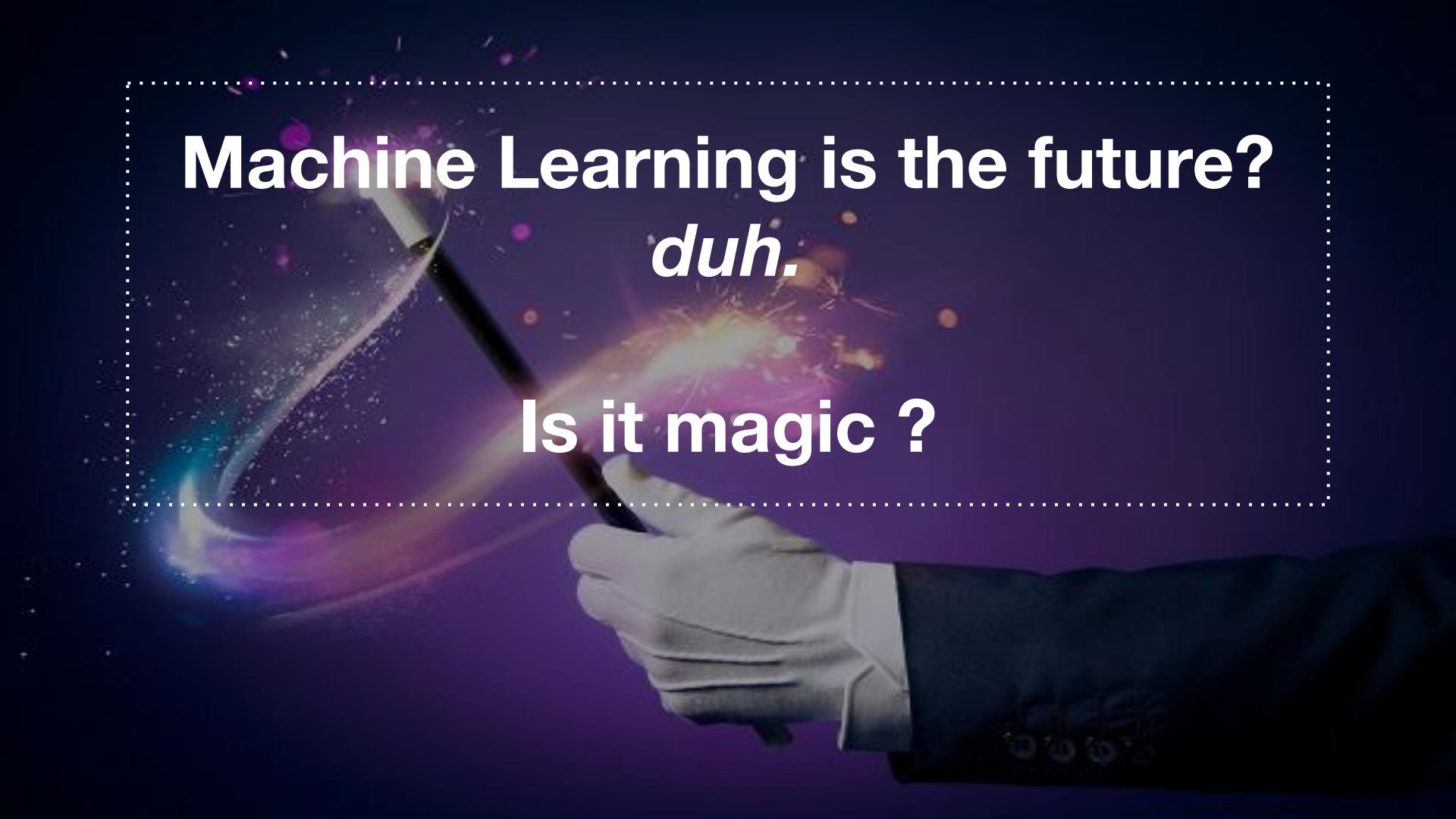

People You May Know
Your Weekend Album From Lake Tahoe
Top Movie Picks For You
Top Trends for you
Tag Your Friend X
Move this screenshot to archive
Inspired by your shopping list


Rapid Bloom and Dominance
of Machine Learning

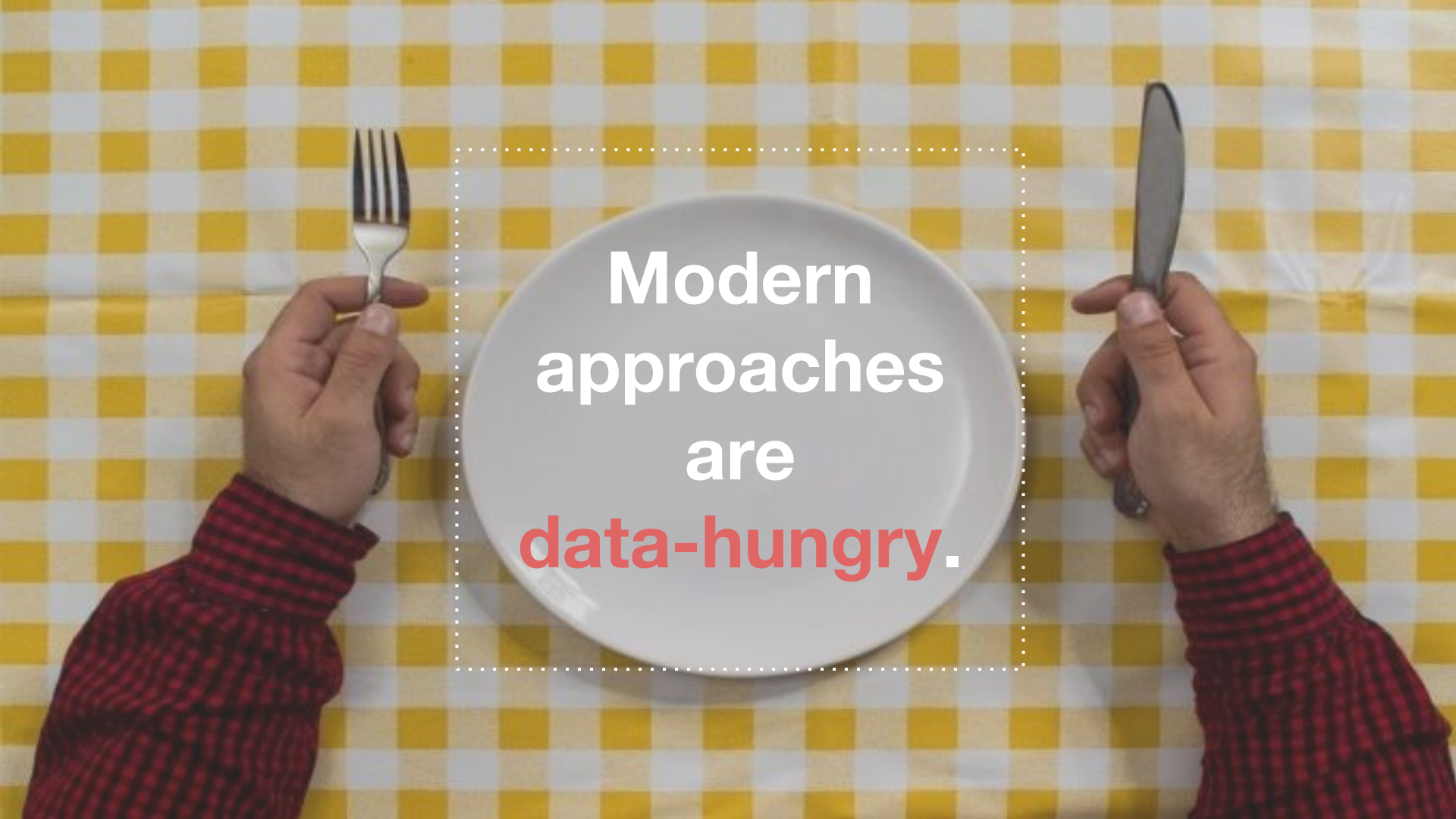

$28.5 Billion allocated to ML
in 2019 Q1 worldwide.
It is expected to grow to
$40 Billion by 2025 from just
$1.3 Billion in 2016.


Number of AI Startups is having a *14x* rate of increase right now.

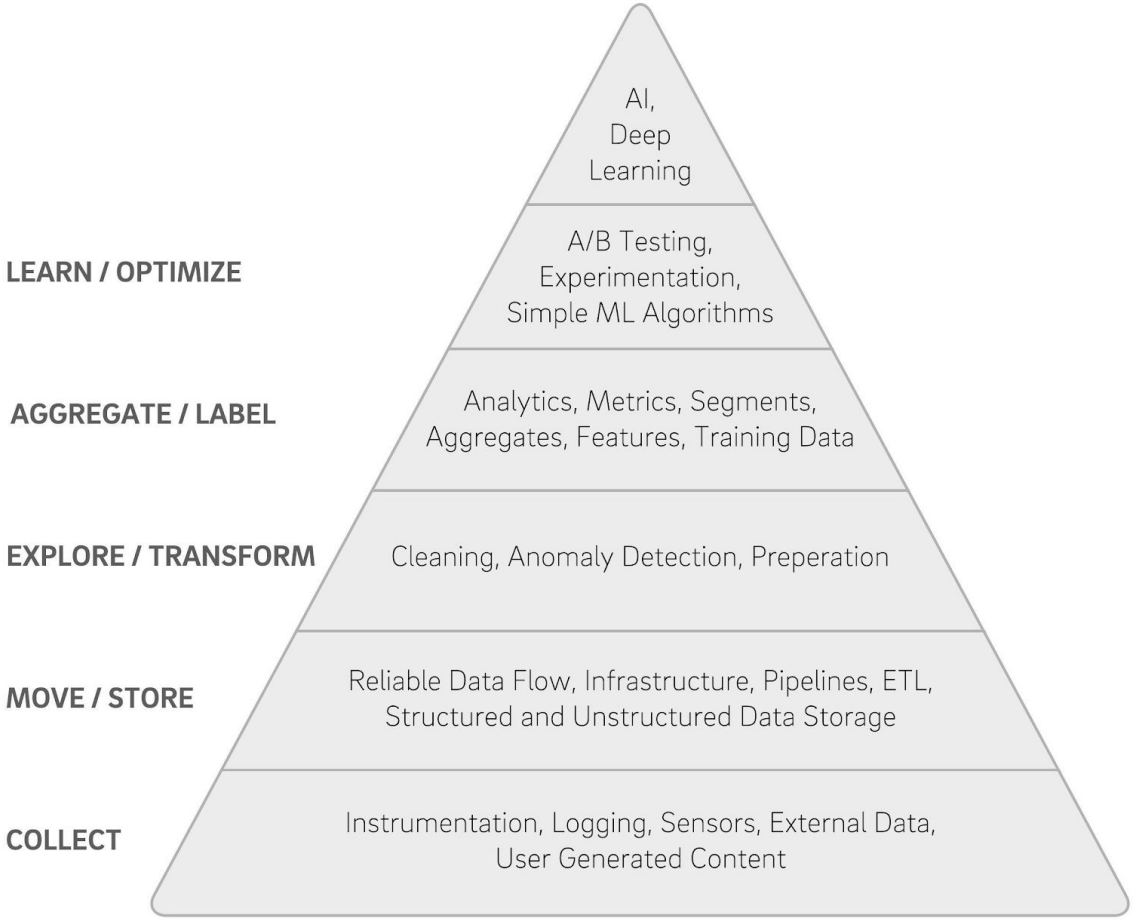

97% mobile users use
AI-powered voice assistants.
With 40% of search
powered by voice.

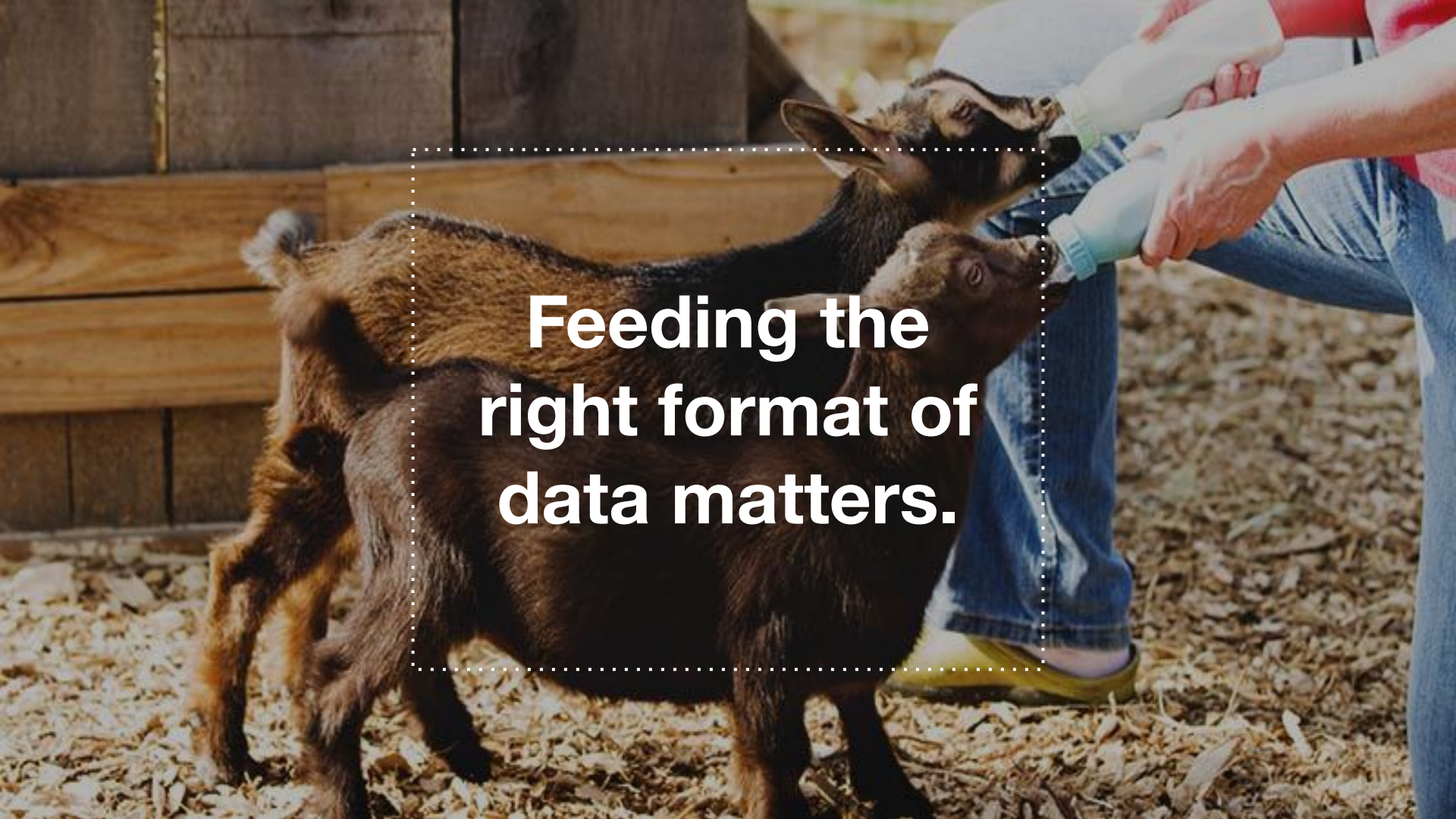

Machine Learning is the future?
duh.
Is it magic ?


Well-Calibrated Data
Sophisticated Algorithms
Efficient Computation


Modern approaches are data-hungry.

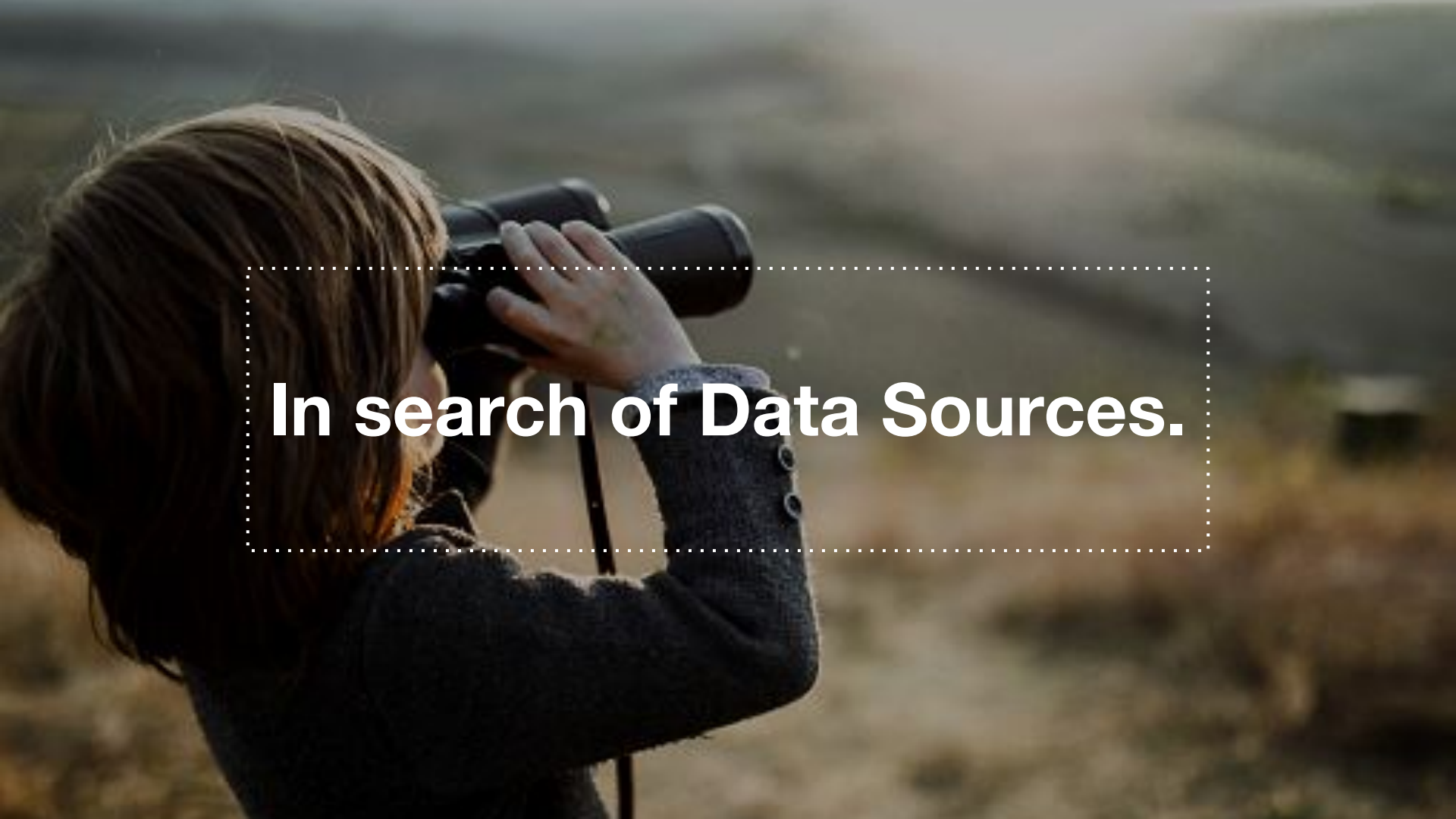
AI Hierarchy of Needs.



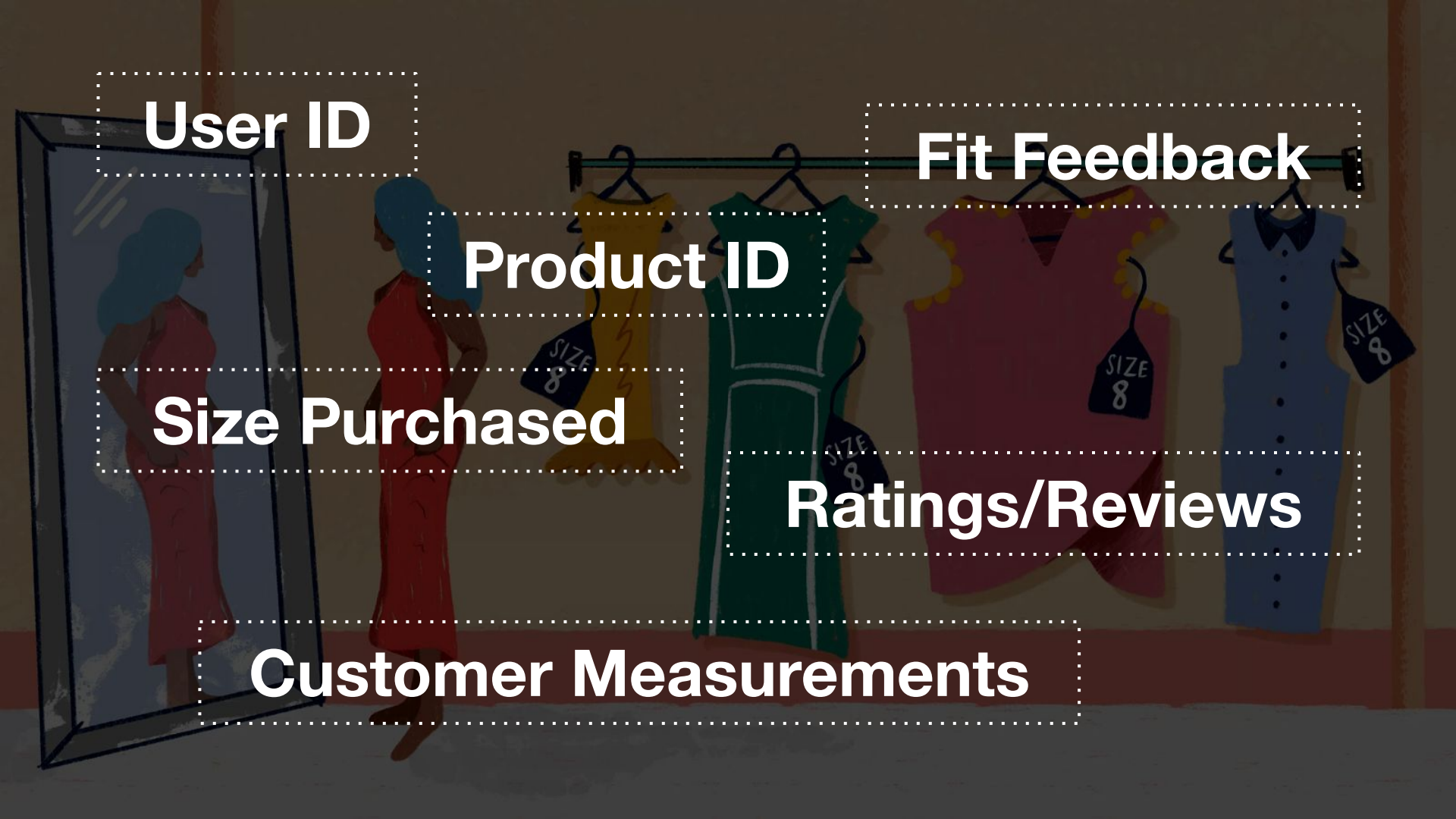

Feeding the
right format of
data matters.

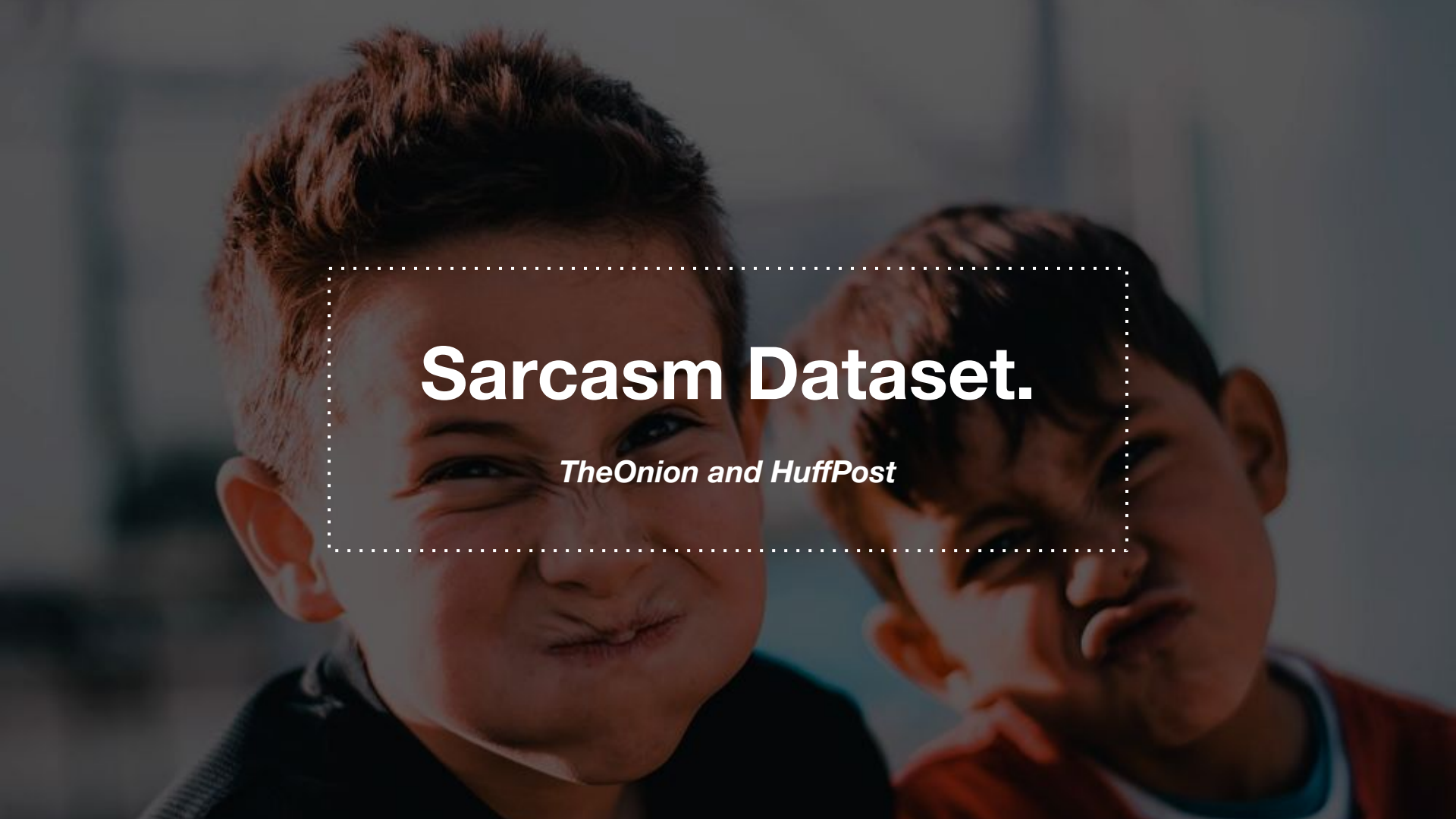

“More data beats clever algorithms, but better data beats more data.”
- Peter Norvig


Data Curation.


In search of Data Sources.


Clothing Fit Dataset.
ModCloth


User ID
Fit Feedback
Product ID
Size Purchased
Ratings/Reviews
Customer Measurements

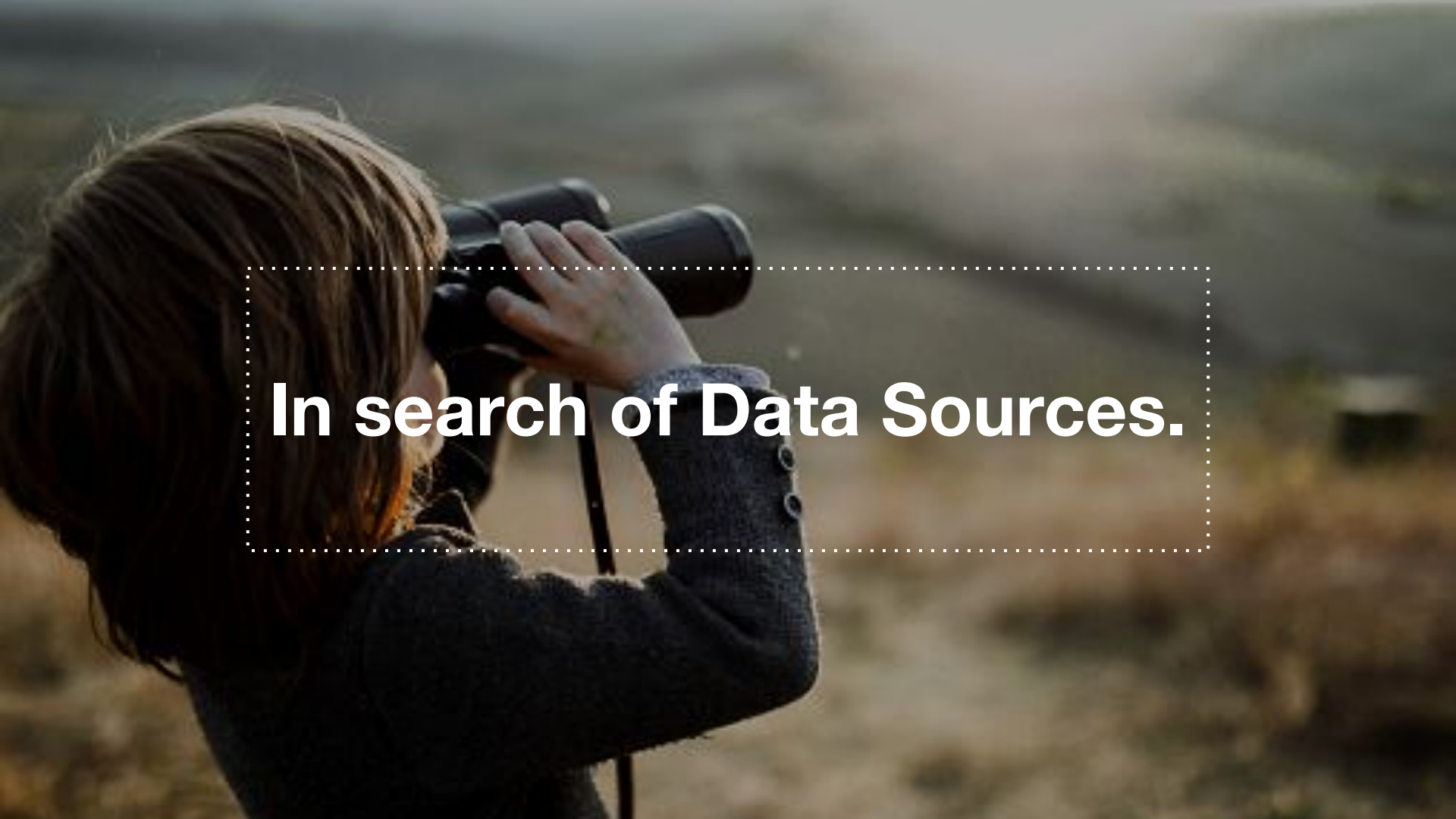

Sarcasm Dataset.
TheOnion and HuffPost

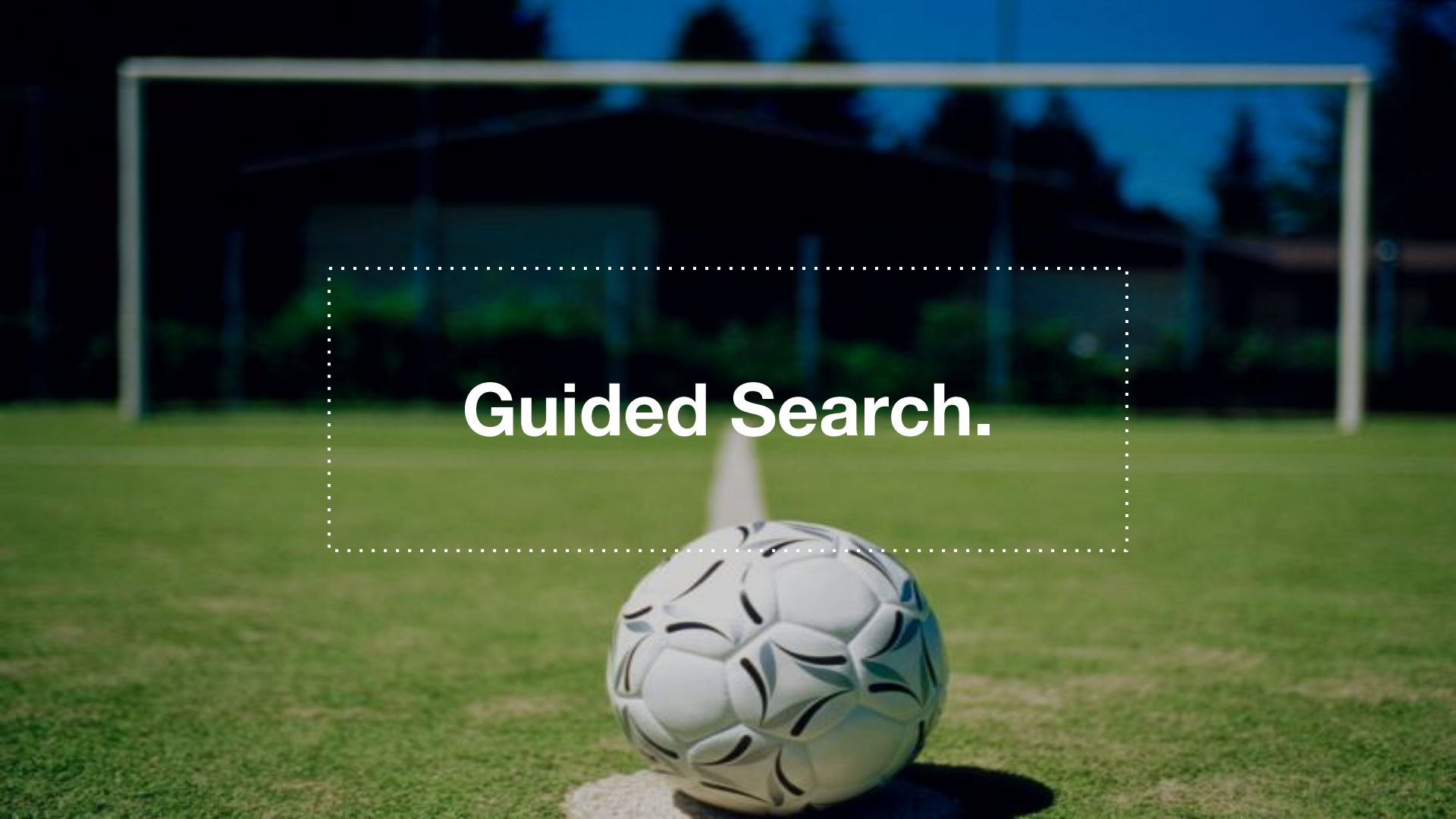

Sarcastic News
Non-Sarcastic/Real News

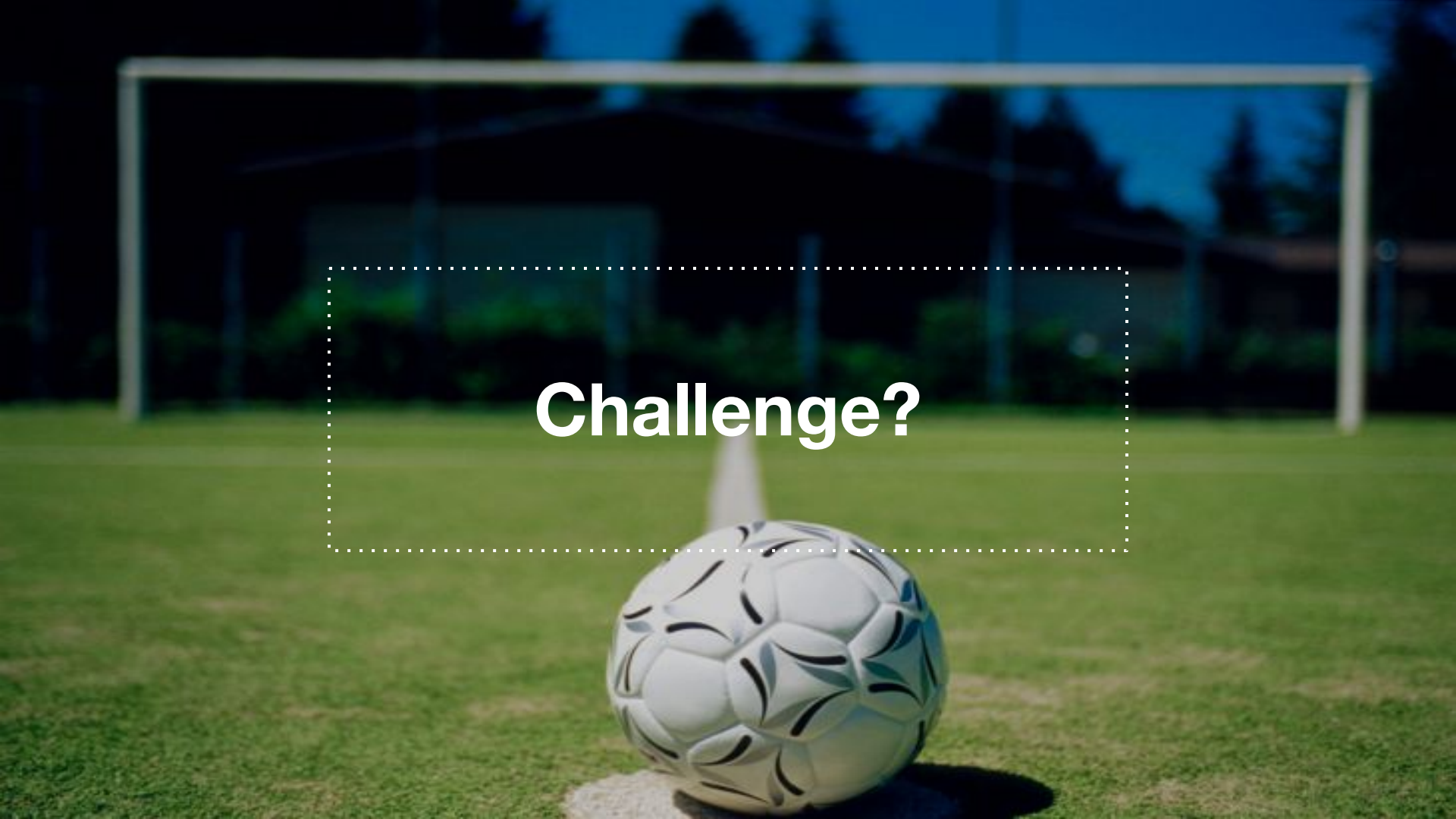
News Category Dataset.

HuffPost


Headline
Preview Text
Tags
Author
Date Published
WALL STREET JOURNAL
FINANCIAL TIMES
Libération
DIE ZEIT
WOCHENZEITUNG FÜR POLITIK · WIRTSCHAFT · HANDEL UND KULTUR
Frankfurter Allgemeine
ZEITUNG FÜR DEUTSCHLAND
Herald Tribune
PUBLISHED WITH THE NEW YORK TIMES AND THE WASHINGTON POST
Paris, Friday, June 21, 1996
Yeltsin Dismisses Hard-I
La Tribune
Russie : un général arbitre du second tou


In search of Data Sources.


Guided Search.


Challenge?

Approach.

- **Formal Problem Definition**

Approach.

- **Formal Problem Definition**

To address the *Size Recommendation* problem, we define the problem as:

“Given a user and a product with different sizes, recommend a product that fits user best.”

Approach.

- Formal Problem Definition
- **Essential Data Signal Determination**

Approach.

- Formal Problem Definition
- **Essential Data Signal Determination**

In case of size recommendation problem, it would be *UserID*, *ProductID*, *size purchased*, and *fit feedback*.

We can collect other data signals like *product category* or *price*; however, they are not of absolute necessity.

Approach.

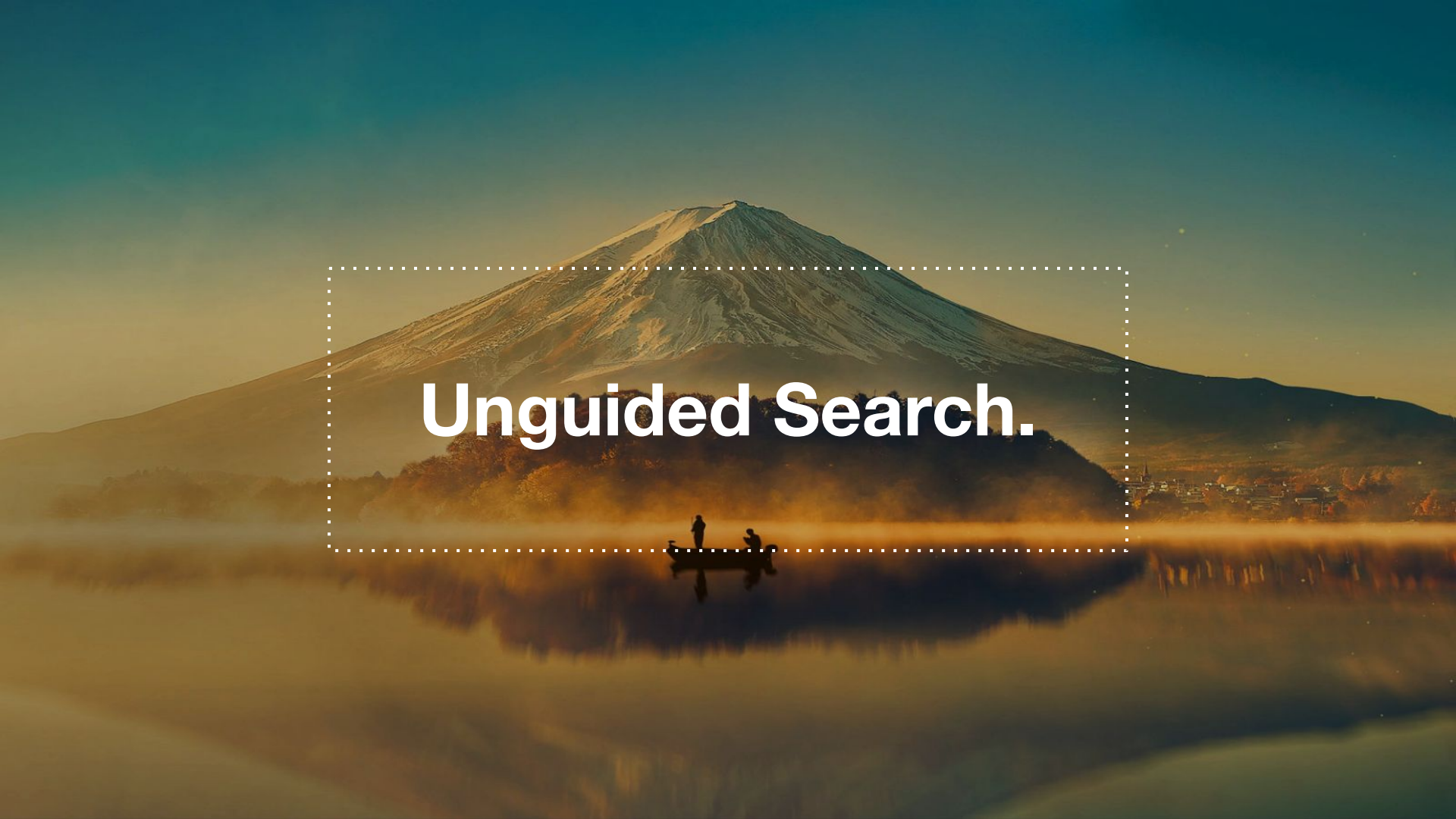
- Formal Problem Definition
- Essential Data Signal Determination
- **Data Volume Requirement**

Approach.

- Formal Problem Definition
- Essential Data Signal Determination
- **Data Volume Requirement**

Data sources should contain enough historical data to construct a sufficiently large dataset.

Machine Learning models might not yield good results when trained on underpowered dataset.


Unguided Search.


Challenge?

Approach.

- Interesting Problem Recognition
  - Data signal(s) worth estimating?
  - Lead to fascinating results?

Approach.

- Interesting Problem Recognition
  - Data signal(s) worth estimating?
  - Lead to fascinating results?

Training a classifier on the News Category dataset could help identify any prose's writing style.

It can also help in tagging uncategorized news articles.

Approach.

- Interesting Problem Recognition
  - Data signal(s) worth estimating?
  - Lead to fascinating results?
- **Metadata Association**

Approach.

- Interesting Problem Recognition
  - Data signal(s) worth estimating?
  - Lead to fascinating results?
- **Metadata Association**

Explore to see how you can collect more data around the essential signals.

Is enough side-information present in the web source? Can we join data from other sources?

Approach.

- Interesting Problem Recognition
  - Data signal(s) worth estimating?
  - Lead to fascinating results?
- Metadata Association
- **Data Volume Requirement**

Approach.

- Interesting Problem Recognition
  - Data signal(s) worth estimating?
  - Lead to fascinating results?
- Metadata Association
- **Data Volume Requirement**

Relaxed requirements here since we are not bounded by a specific problem definition.

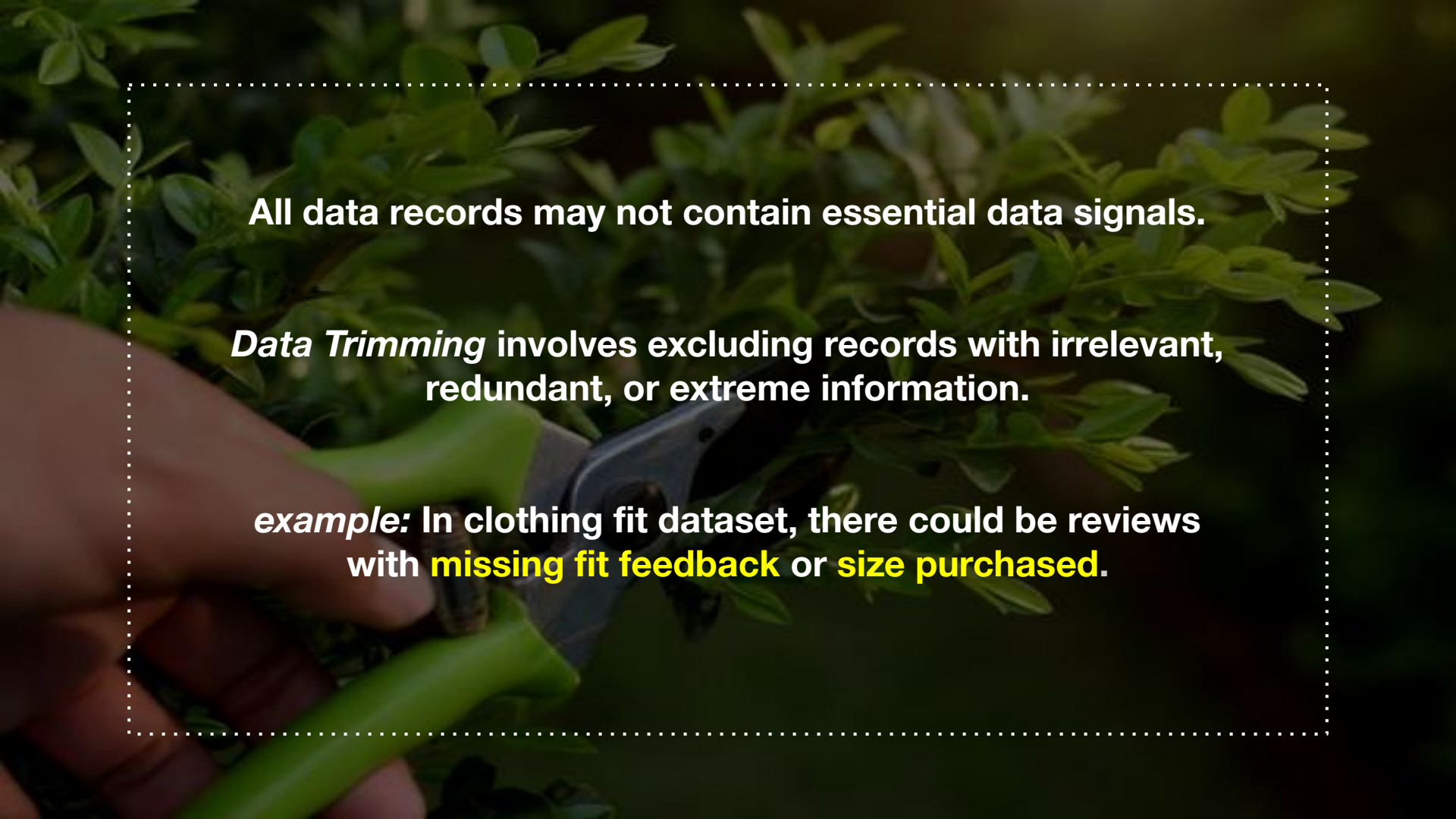
We could propose new problem statements based on data availability.

Approach.

- Interesting Problem Recognition
  - Data signal(s) worth estimating?
  - Lead to fascinating results?
- Metadata Association
- Data Volume Requirement
- **Data Uniqueness Check**


Dataset Preparation.


Data Trimming.



All data records may not contain essential data signals.

Data Trimming involves excluding records with irrelevant, redundant, or extreme information.

example: In clothing fit dataset, there could be reviews with missing fit feedback or size purchased.


Data Integration.

Vertical Integration

Sarcasm Detection dataset: TheOnion + HuffPost.

To create a quality dataset, we fuse the records collected by appending them vertically.

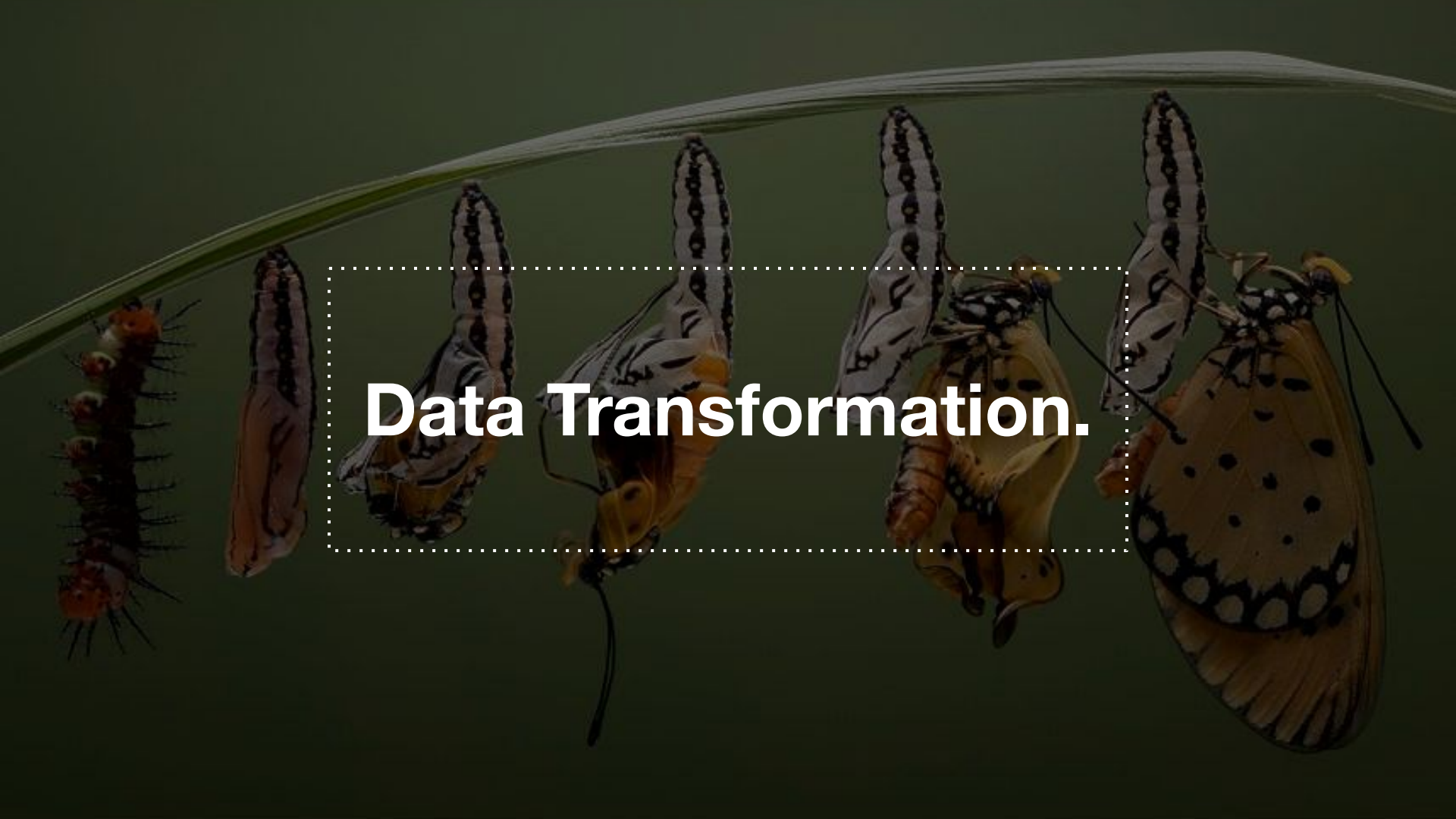
Since attributes in both the sets are similar, namely `article_link`, `headline`, and `is_sarcastic` label, it is easy to connect them vertically to form a dataset.

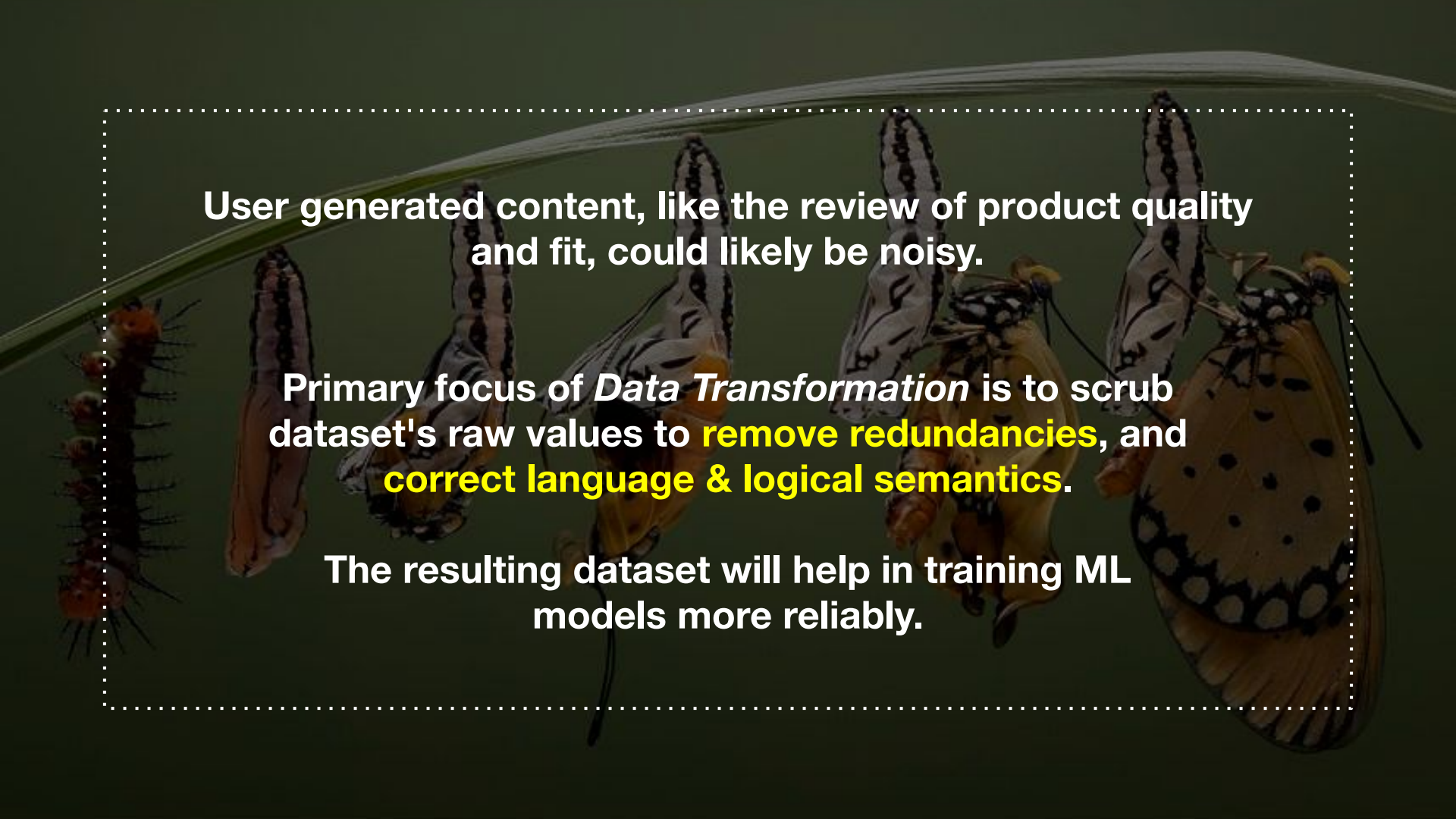
Horizontal Integration

Example: We may want to build a dataset regarding different cities globally.

Highly probable that one source may not have all the information. Attributes like `state`, `country`, and `population` might be available from one source, whereas `demographics` and `culture` from another.

In such cases, we need to perform the integration of data horizontally by joining based on a key.

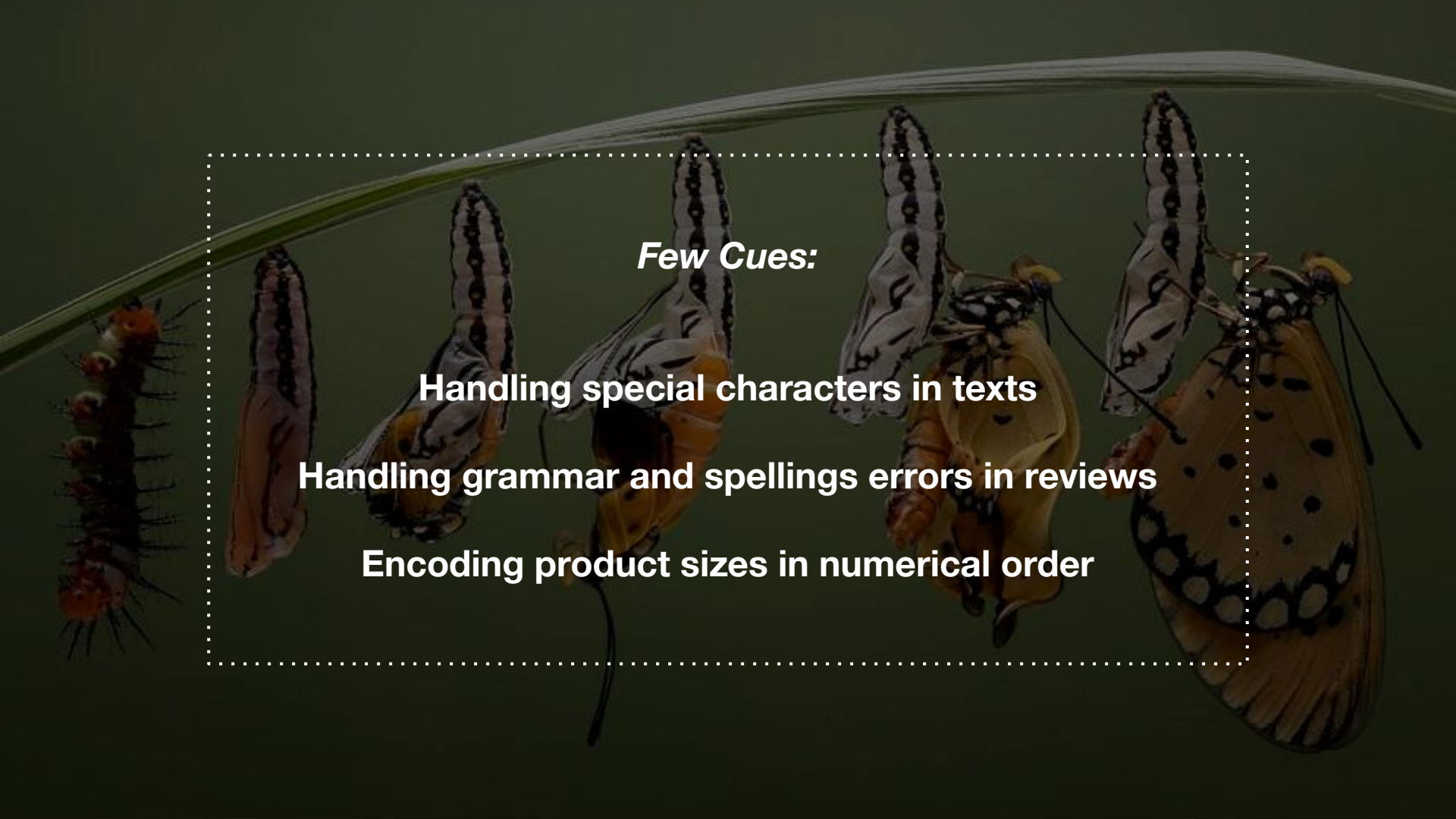

Data Transformation.



User generated content, like the review of product quality and fit, could likely be noisy.

Primary focus of *Data Transformation* is to scrub dataset's raw values to remove redundancies, and correct language & logical semantics.

The resulting dataset will help in training ML models more reliably.



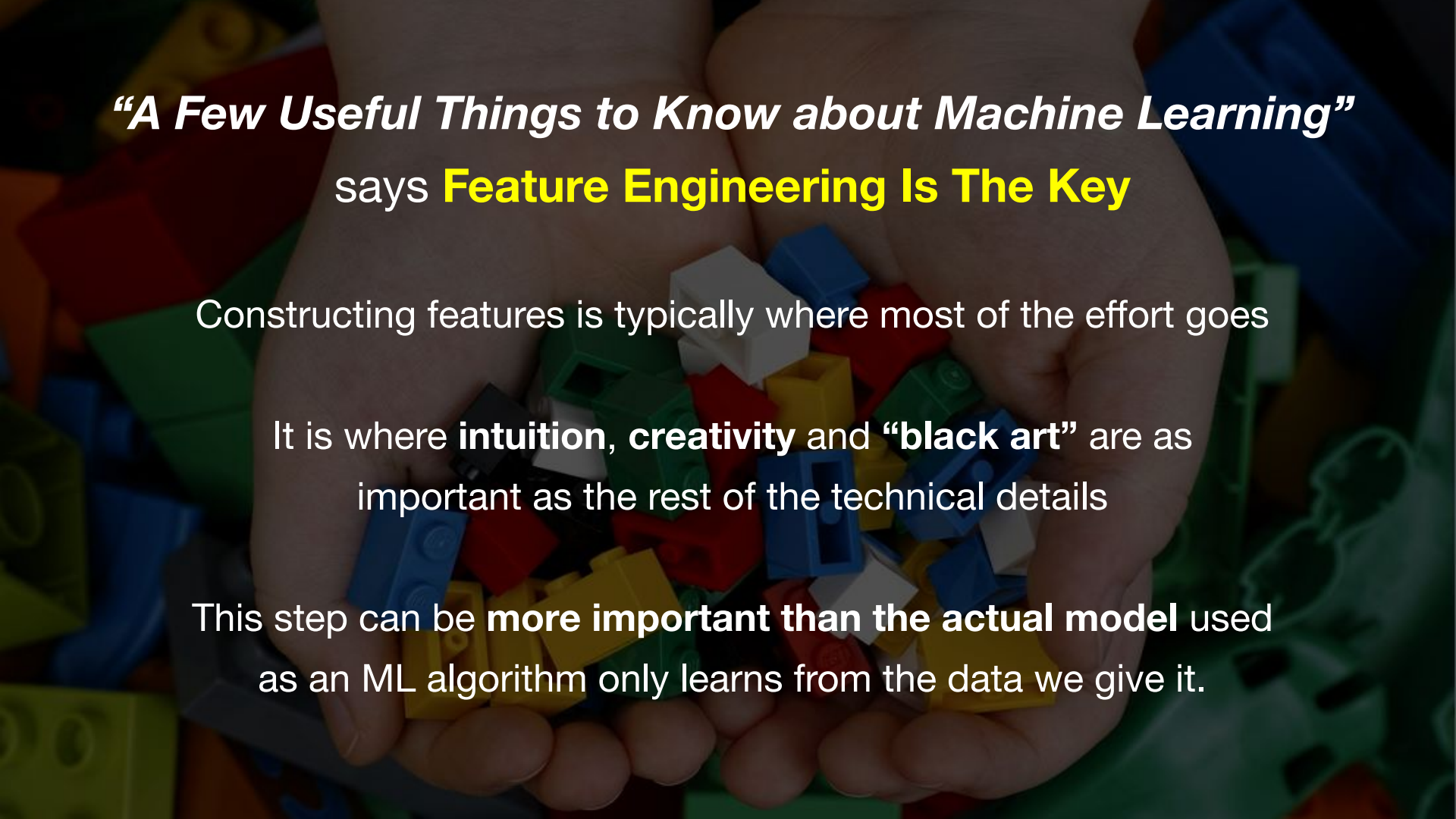
Few Cues:

Handling special characters in texts

Handling grammar and spellings errors in reviews

Encoding product sizes in numerical order


Feature Engineering.



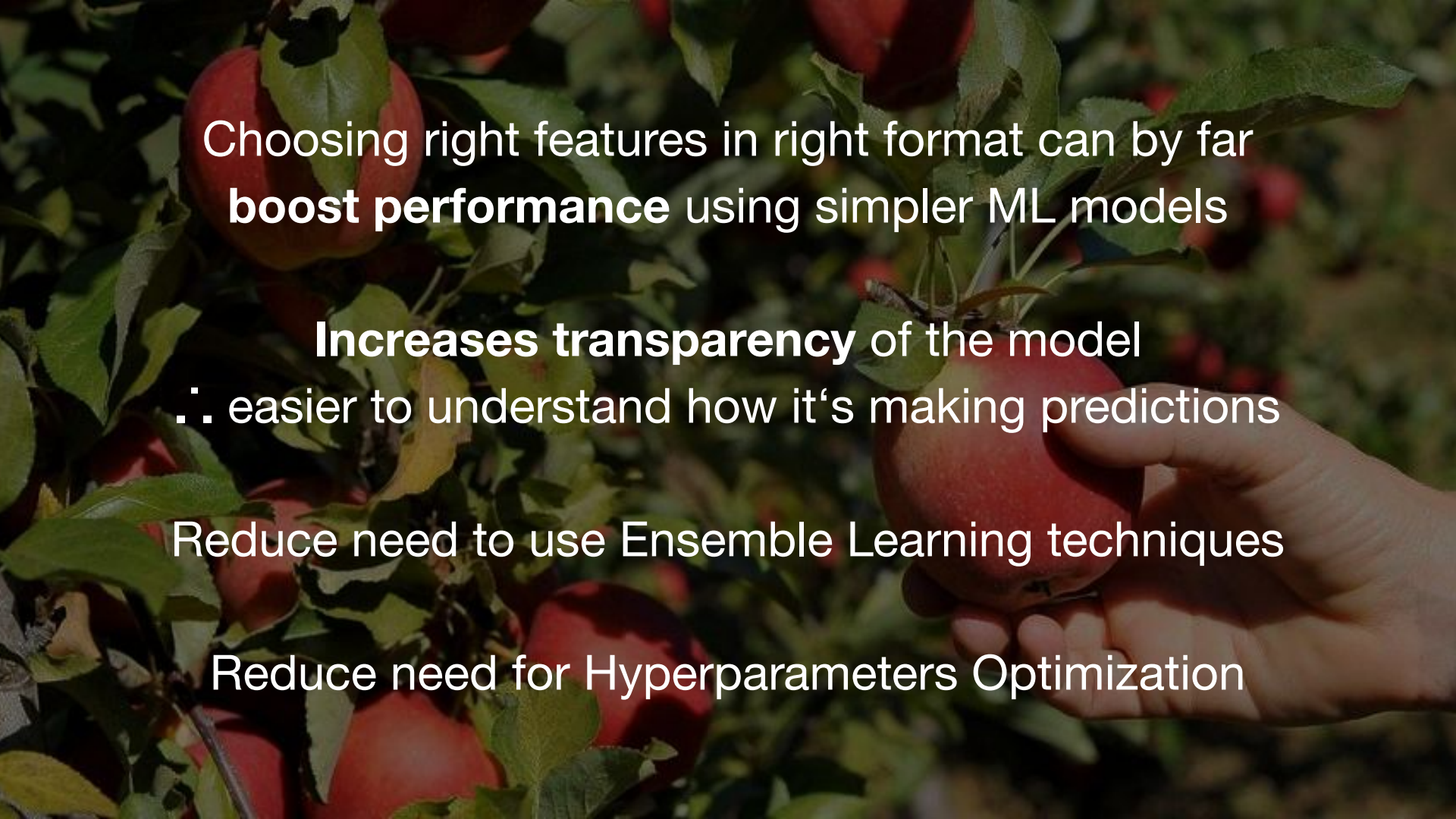
“A Few Useful Things to Know about Machine Learning”

says **Feature Engineering Is The Key**

Constructing features is typically where most of the effort goes

It is where **intuition**, **creativity** and **“black art”** are as important as the rest of the technical details

This step can be **more important than the actual model** used as an ML algorithm only learns from the data we give it.



Choosing right features in right format can by far **boost performance** using simpler ML models

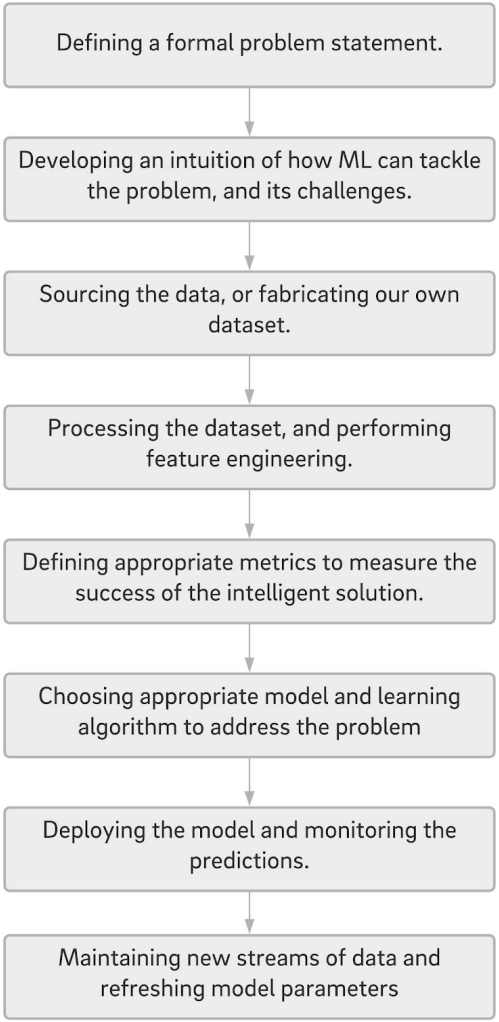
Increases transparency of the model
∴ easier to understand how it‘s making predictions

Reduce need to use Ensemble Learning techniques

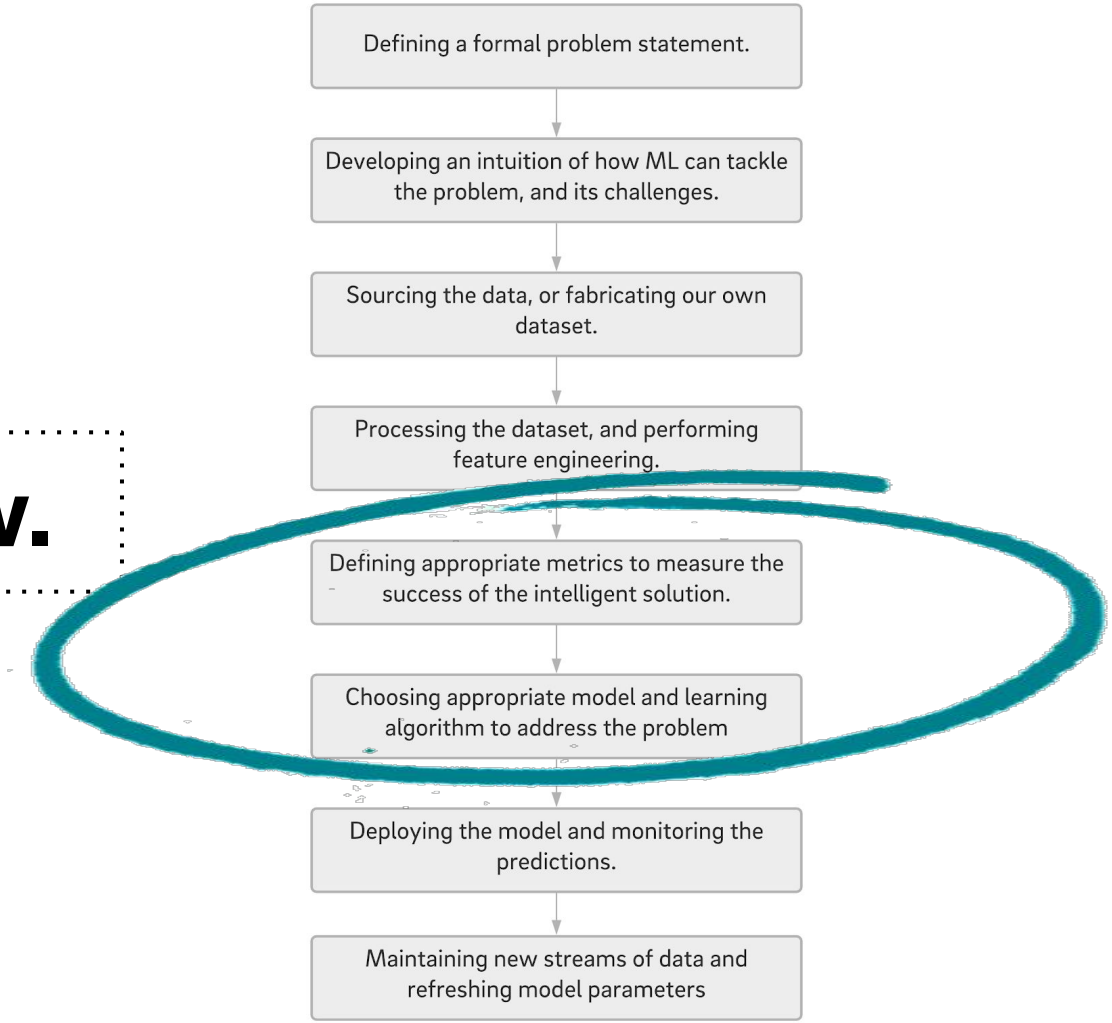
Reduce need for Hyperparameters Optimization


Bottom Line.

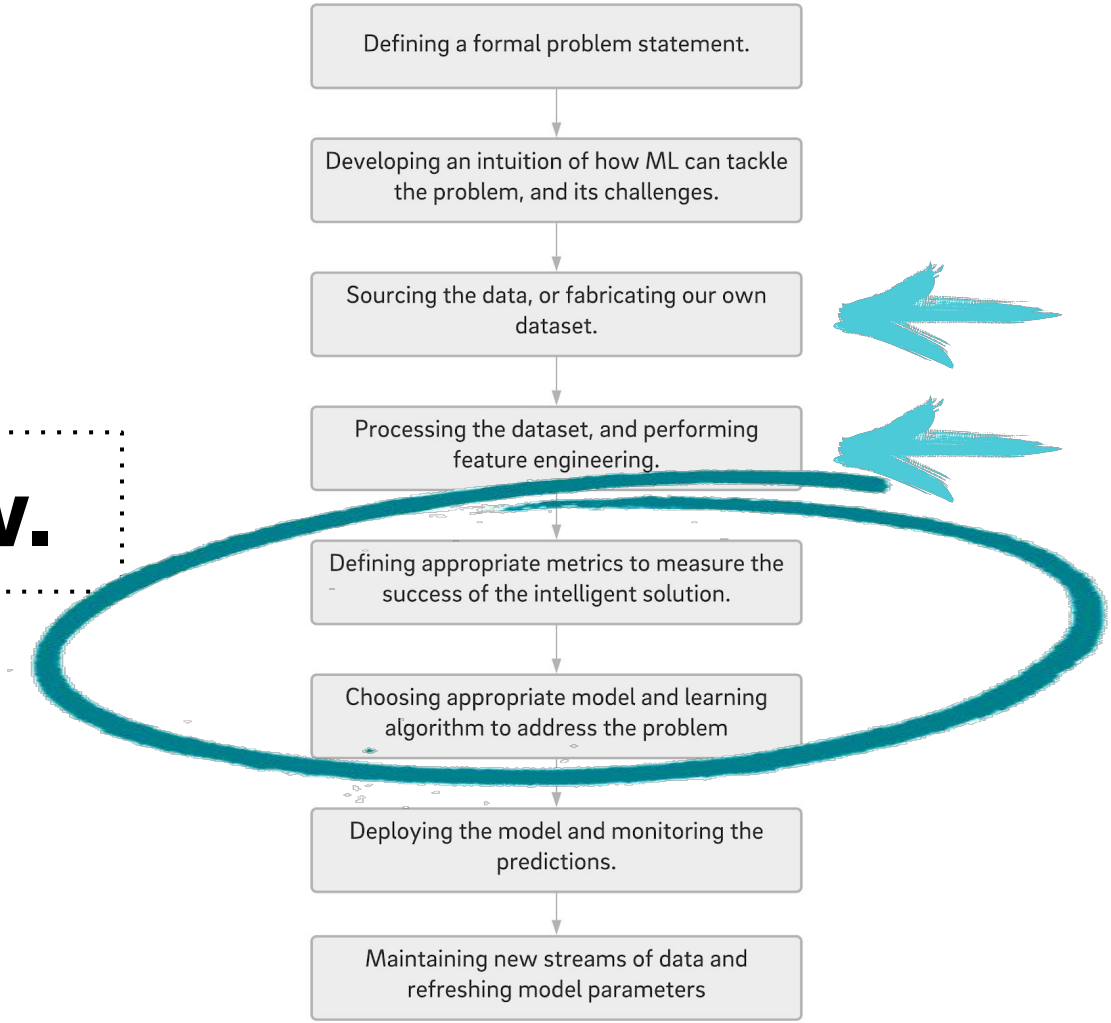
Workflow.


Defining a formal problem statement.
Developing an intuition of how ML can tackle the problem, and its challenges.
Sourcing the data, or fabricating our own dataset.
Processing the dataset, and performing feature engineering.
Defining appropriate metrics to measure the success of the intelligent solution.
Choosing appropriate model and learning algorithm to address the problem
Deploying the model and monitoring the predictions.
Maintaining new streams of data and refreshing model parameters

Workflow.


Defining a formal problem statement.
Developing an intuition of how ML can tackle the problem, and its challenges.
Sourcing the data, or fabricating our own dataset.
Processing the dataset, and performing feature engineering.
Defining appropriate metrics to measure the success of the intelligent solution.
Choosing appropriate model and learning algorithm to address the problem
Deploying the model and monitoring the predictions.
Maintaining new streams of data and refreshing model parameters

Workflow.


Defining a formal problem statement.
Developing an intuition of how ML can tackle the problem, and its challenges.
Sourcing the data, or fabricating our own dataset.
Processing the dataset, and performing feature engineering.
Defining appropriate metrics to measure the success of the intelligent solution.
Choosing appropriate model and learning algorithm to address the problem
Deploying the model and monitoring the predictions.
Maintaining new streams of data and refreshing model parameters

Andrew Ng
@AndrewYNg

Would love your feedback on this: AI Systems = Code (model/algorithm) + Data. Most academic benchmarks/competitions hold the Data fixed, and let teams work on the Code. Thinking of organizing something where we hold the Code fixed, and ask teams to work on the Data. (1/2)

1:12 PM · May 24, 2021 · Twitter Web App

402 Retweets **91** Quote Tweets **3,498** Likes



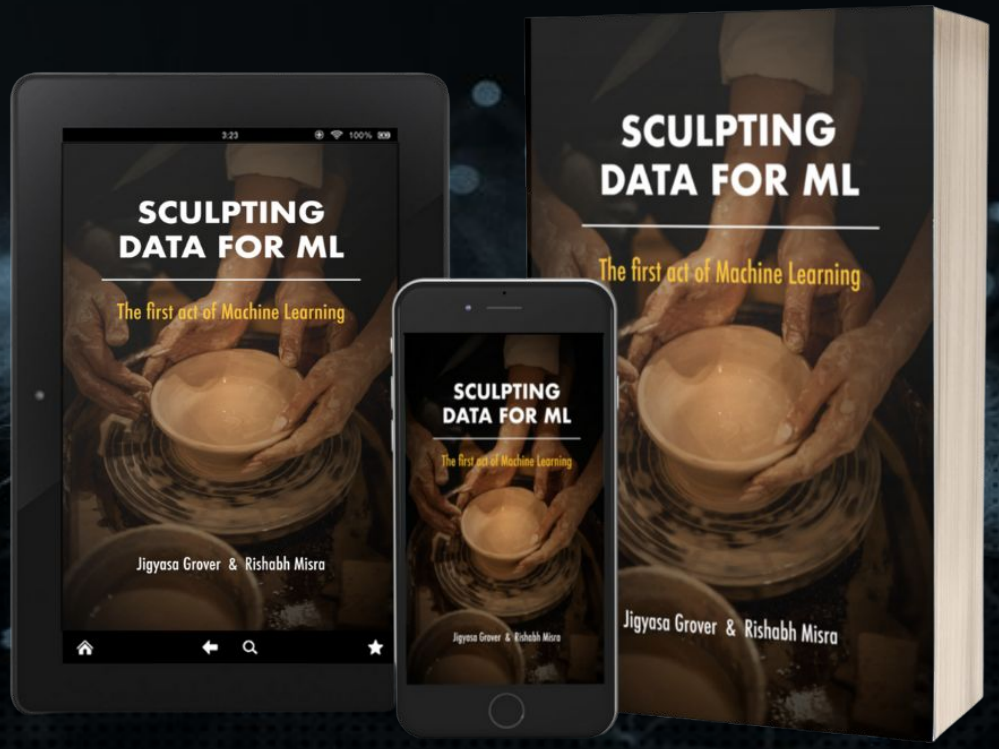

Tweet your reply

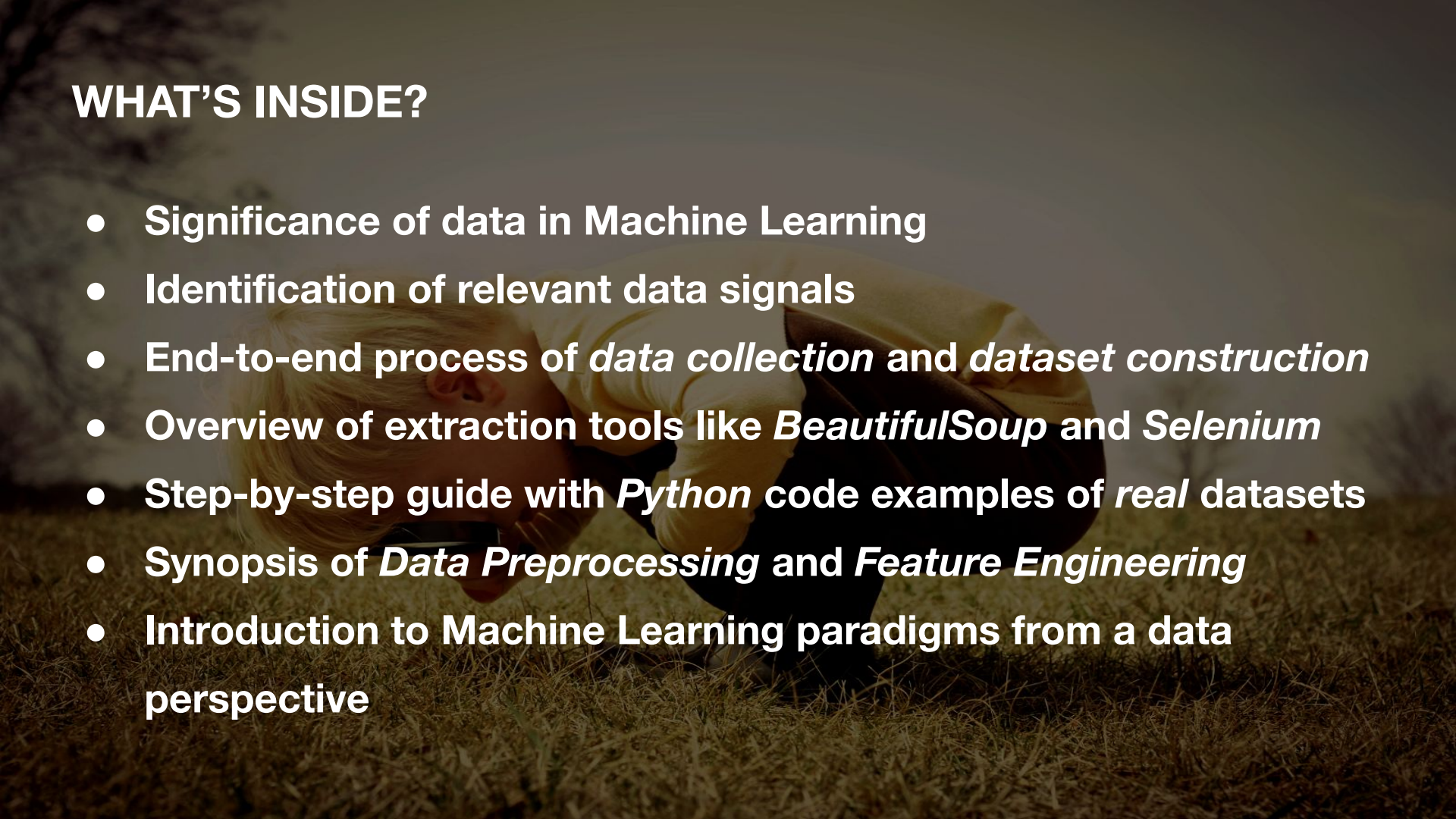
Reply

Andrew Ng  @AndrewYNg · May 24
Replying to @AndrewYNg
Hoping this will more closely reflect ML application practice, and also spur innovative research on data-centric AI development. What do you think? (2/2)

 102  49  1K 

amzn.com/B08RN47C5T





WHAT'S INSIDE?

- Significance of data in Machine Learning
- Identification of relevant data signals
- End-to-end process of *data collection* and *dataset construction*
- Overview of extraction tools like *BeautifulSoup* and *Selenium*
- Step-by-step guide with *Python* code examples of *real* datasets
- Synopsis of *Data Preprocessing* and *Feature Engineering*
- Introduction to Machine Learning paradigms from a data perspective

Endorsed by leading ML experts.
Read forewords by:

Julian McAuley
Associate Professor at UC San Diego

Laurence Moroney
Lead Artificial Intelligence Advocate at Google

Mengting Wan
Senior Applied Scientist at Microsoft

@jigyasa_grover
linkedin.com/in/jigyasa-grover