



Sculpting Data for Machine Learning

Rishabh Misra


 [@rishabh_misra_](https://twitter.com/@rishabh_misra_)

ML Engineer, Twitter

Jigyasa Grover

 [@jigyasa_grover](https://twitter.com/@jigyasa_grover)

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**

A hand in a white glove holding a wand that emits a glowing, colorful trail of light against a dark purple background. The wand is held diagonally, and the light trail curves around it, creating a magical effect. The background is dark with some faint, colorful particles.

Machine Learning is the future?

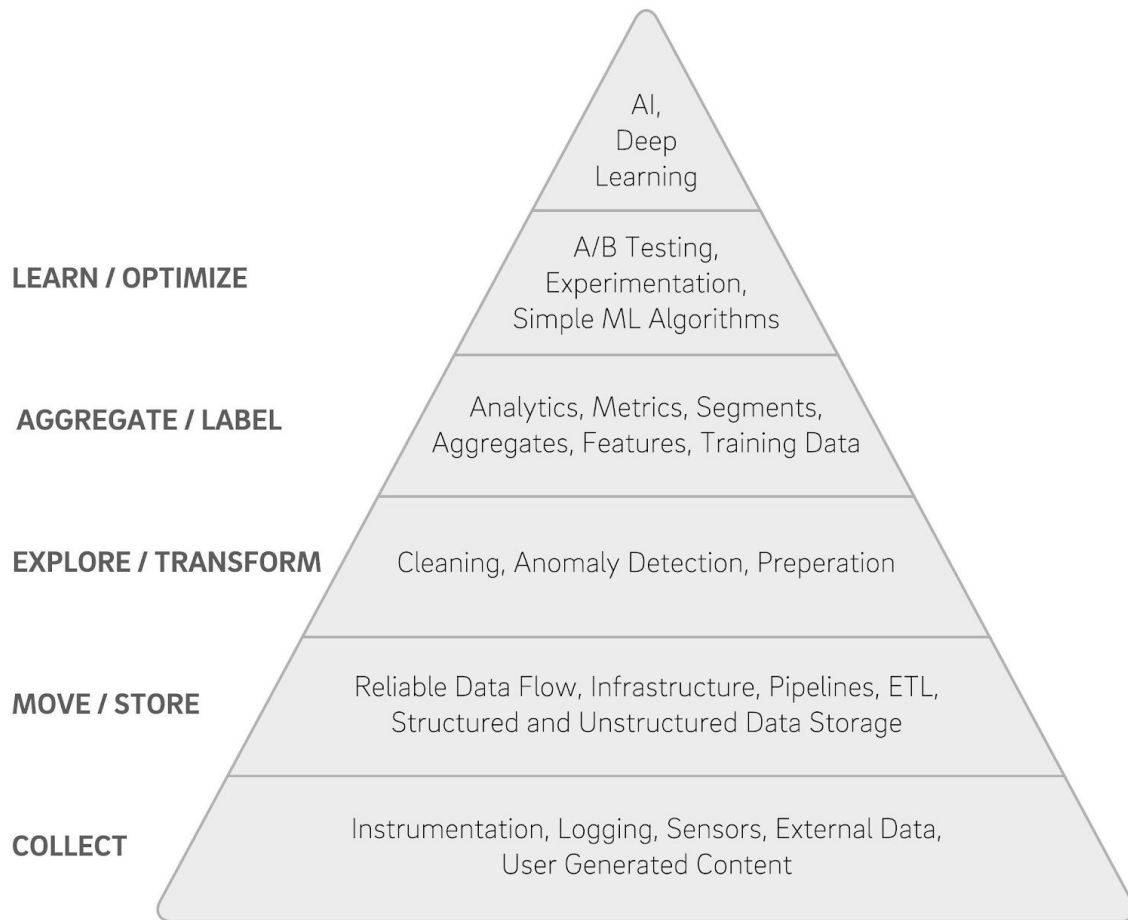
duh.


Is it magic ?

A top-down view of a person's hands holding a silver fork and a silver knife over a white, empty plate. The person is wearing a red and black checkered shirt. The background is a yellow and white checkered tablecloth. A white dotted line forms a rectangular frame around the plate and the text.

**Modern
approaches are
data-hungry.**

AI Hierarchy of Needs.



A person wearing a white cap and blue jeans is feeding two small brown goats with bottles in a barn. The goats are standing on a bed of wood shavings. The person is holding two bottles, one for each goat. The scene is set in a barn with wooden walls.

**Feeding the
right format of
data matters.**

A glass of water splashing, with a background of binary code (0s and 1s) and a blue color scheme. The text is overlaid on the image.

**“More data beats clever algorithms,
but better data beats more data.”**

- Peter Norvig



Data Curation.

A person wearing a red long-sleeved shirt, olive green cargo pants, and blue sneakers is walking a brown dog on a city street. The dog is wearing a red vest with Chinese characters and a small cartoon character on it. The background shows a blurred city street with cars and trees. The text "Situational Analysis." is overlaid in the center of the image, enclosed in a white dotted rectangular border.

Situational Analysis.



Clothing Fit Dataset.

A photograph of two young boys. The boy on the left has a playful, slightly pouting face with his eyes squinted. The boy on the right has a more serious, slightly pouting face. The image is overlaid with a semi-transparent dark blue filter. A white dotted rectangular box is centered over the text.

Sarcasm Dataset.



WALL STREET JOURNAL
© 1996 Dow Jones & Company
All Rights Reserved.
MONDAY, JUNE 17, 1996
Markets Sur...

FINANCIAL TIMES
Lucy Kellaway
Today's surveys
Vietnam; Devon
and Cornwall

What's News



DEMAIN
DANS
INITIATIVES
EMPLOI
Les premiers pas
dans la vie active

ANNONCES CLASSÉES
de la page V à la page XII

Le Monde
INITIATIVES

Horst Stern: Reise durch ostdeutsche Landschaften (Seite 6)

DIE ZEIT
WOCHENZEITUNG FÜR POLITIK · WIRTSCHAFT · KUNST · KULTUR

MONDE. Syrie-Turquie
Tension entre
Damas et
Ankara



La Turquie et la Syrie
ont procédé à des
mouvements de
troupes le long de leur
frontière commune.
Cette crise survient
après la signature d'un accord militaire
entre Tel-Aviv et Ankara, qui ne ménage
pas ses critiques à l'égard du régime d'As-
sad (ci-dessus à gauche). Page 6

FRANCE. Deux-Sèvres
Le Marais
poitevin vidé
de ses
Des...

Libération

Frankfurter Allgemeine
ZEITUNG FÜR DEUTSCHLAND

Herald Tribune
INTERNATIONAL
PUBLISHED WITH THE NEW YORK TIMES AND THE WASHINGTON POST

Paris, Friday, June 21, 1996

Russie : un général arbitre du second tou
La Tribune
QUOTIDIEN ■ MARDI 18 JUN 1996

Paris CAC 40 + 0,06 %	New York DJ - 0,08 %
Dollar Paris 5,135 FF	

POUR ÉVIT

peans Face

Yeltsin Dismisses Hard-I

A person wearing a red long-sleeved shirt, olive green cargo pants, and blue sneakers is walking a brown dog on a city street. The dog is wearing a red vest with Chinese characters and a small cartoon character on it. The background shows a blurred city street with cars and trees. The text "Situational Analysis." is overlaid in white, bold font, enclosed in a white dotted rectangular border.

Situational Analysis.

A soccer ball with a white and black pattern is positioned on a green grass field. In the background, a soccer goal is visible, and the scene is set against a clear blue sky with some trees. The text "Guided Search." is overlaid on the image within a white dotted rectangular border.

Guided Search.

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



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

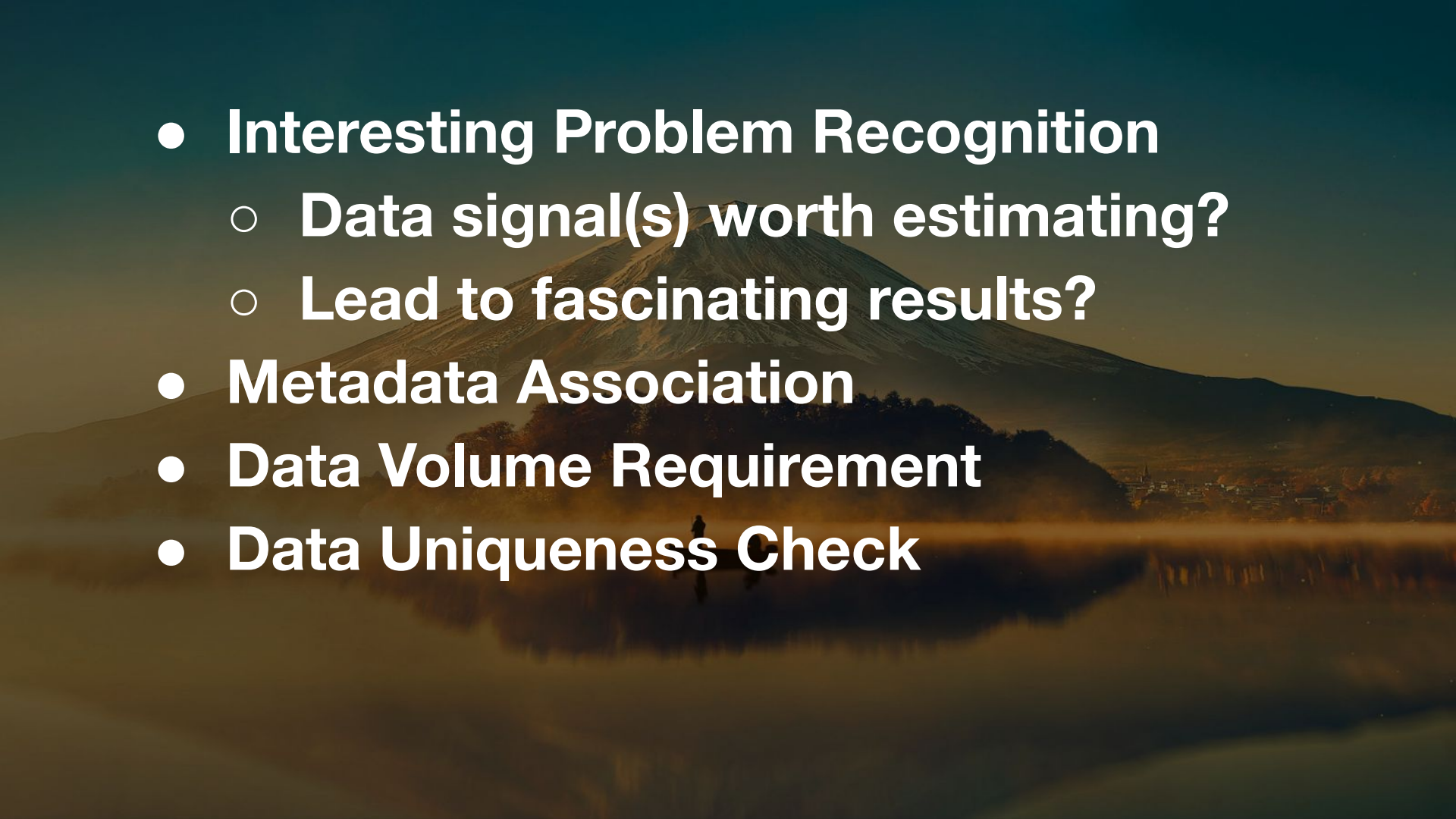


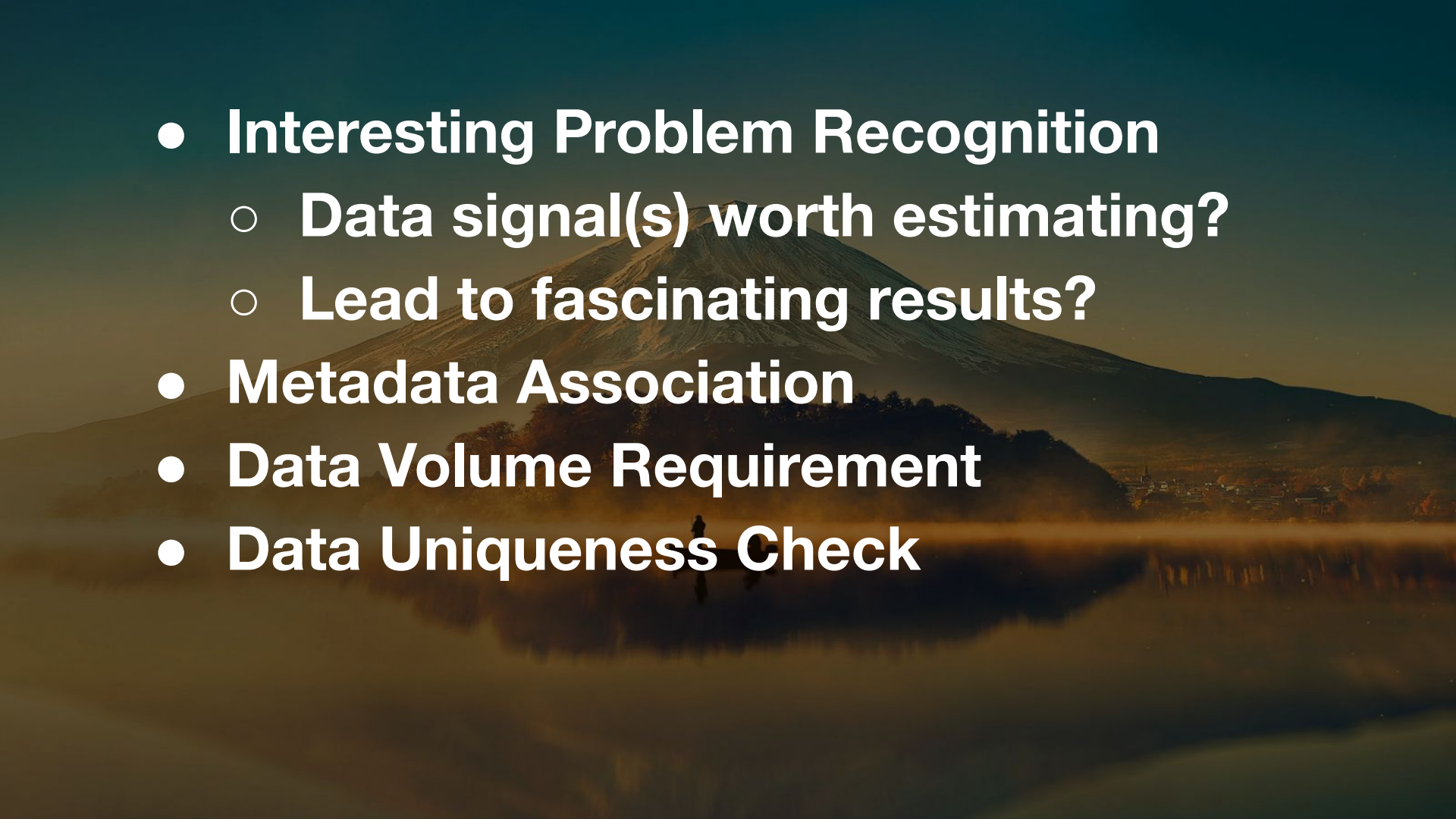
- **Formal Problem Definition**
- **Essential Data Signal Determination**
- **Data Volume Requirement**





Unguided Search.

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

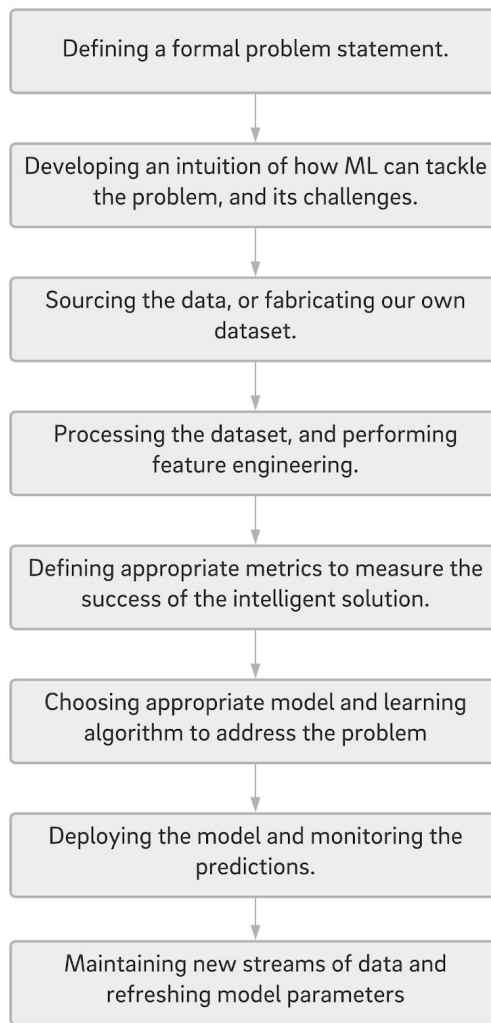
- 
- **Interesting Problem Recognition**
 - **Data signal(s) worth estimating?**
 - **Lead to fascinating results?**
 - **Metadata Association**
 - **Data Volume Requirement**
 - **Data Uniqueness Check**



Bottom Line.



Workflow.





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.

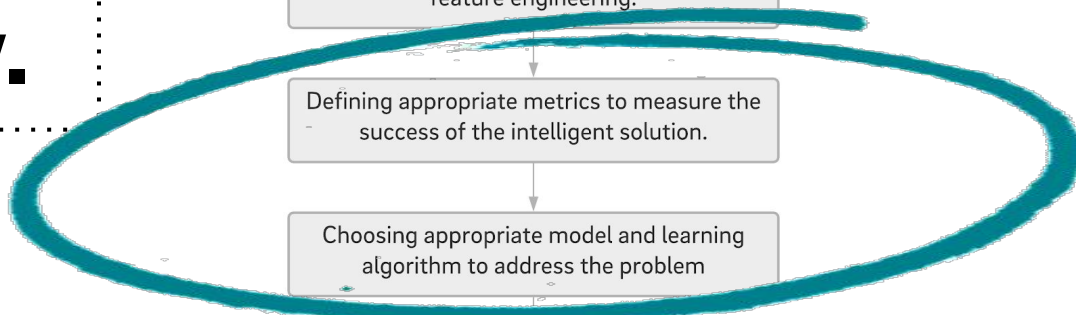
Processing the dataset, and performing feature engineering.

Defining appropriate metrics to measure the success of the intelligent solution.

Choosing appropriate model and learning algorithm to address the problem

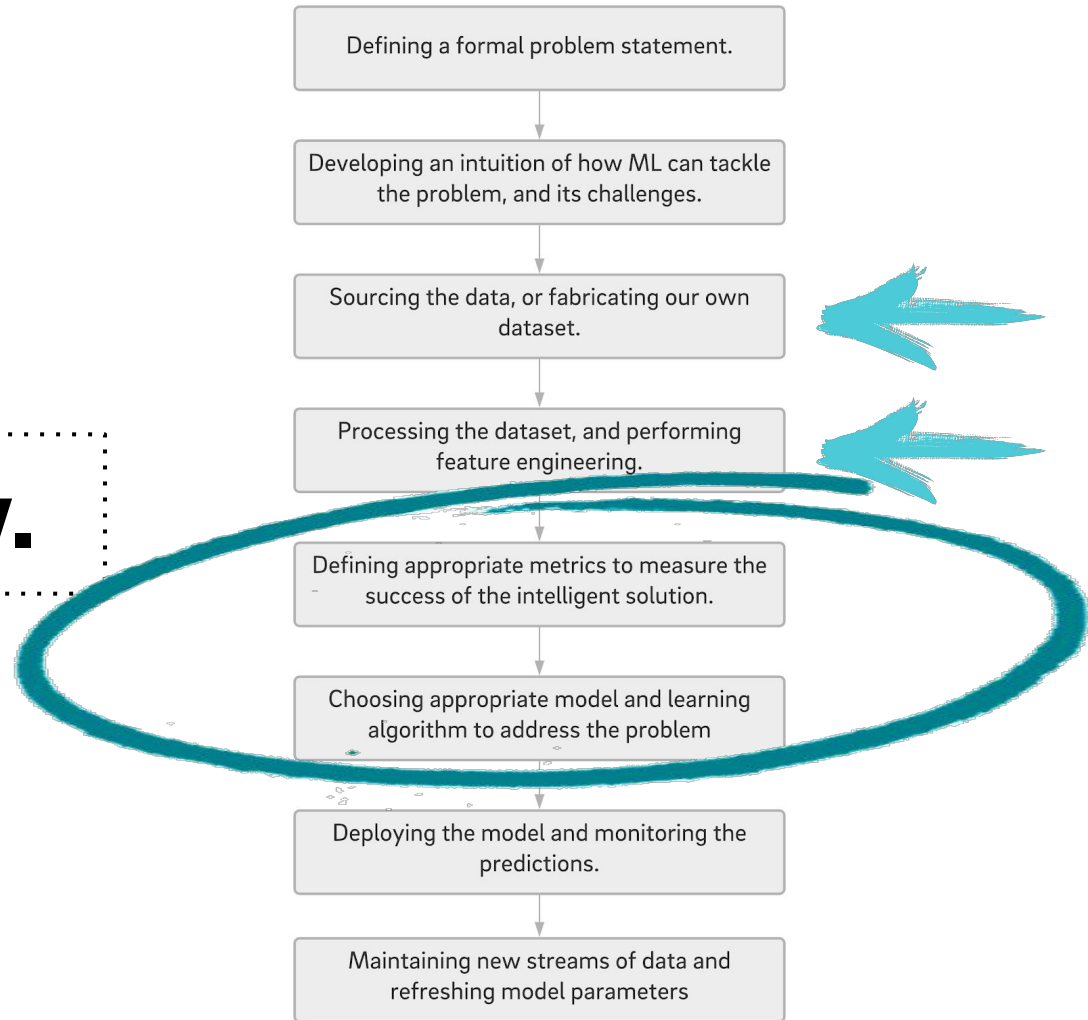
Deploying the model and monitoring the predictions.

Maintaining new streams of data and refreshing model parameters





Workflow.





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



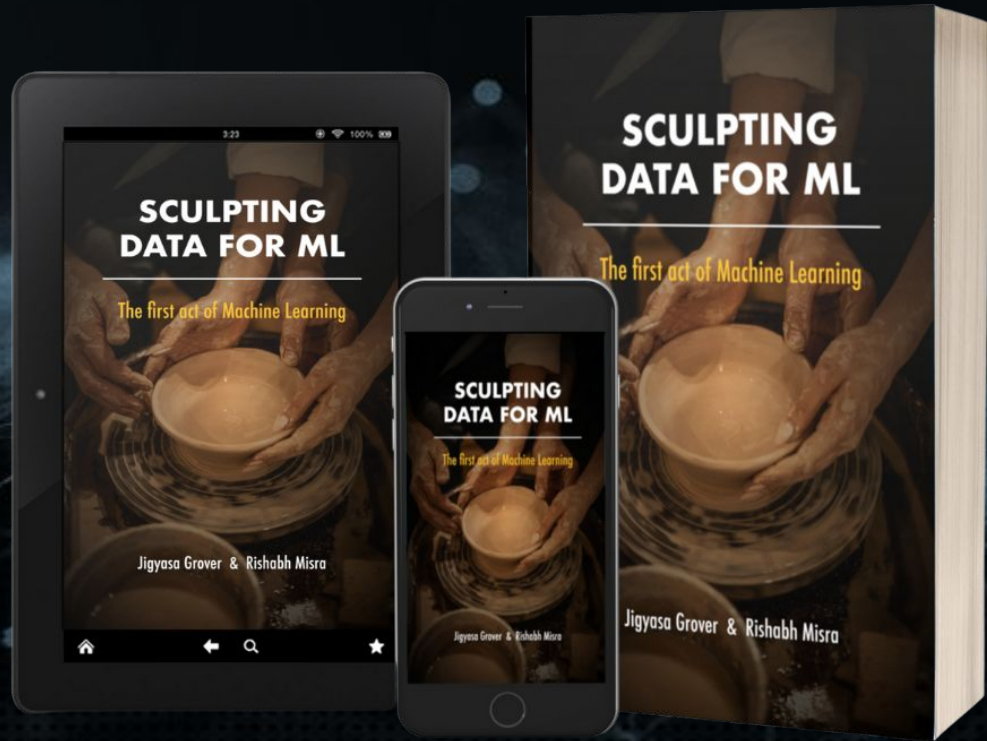
49



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

Thank
you



@rishabh_misra

rishabhmisra.github.io



@jigyasa_grover

jigyasa-grover.github.io