



Bighead

Airbnb's end-to-end machine learning platform

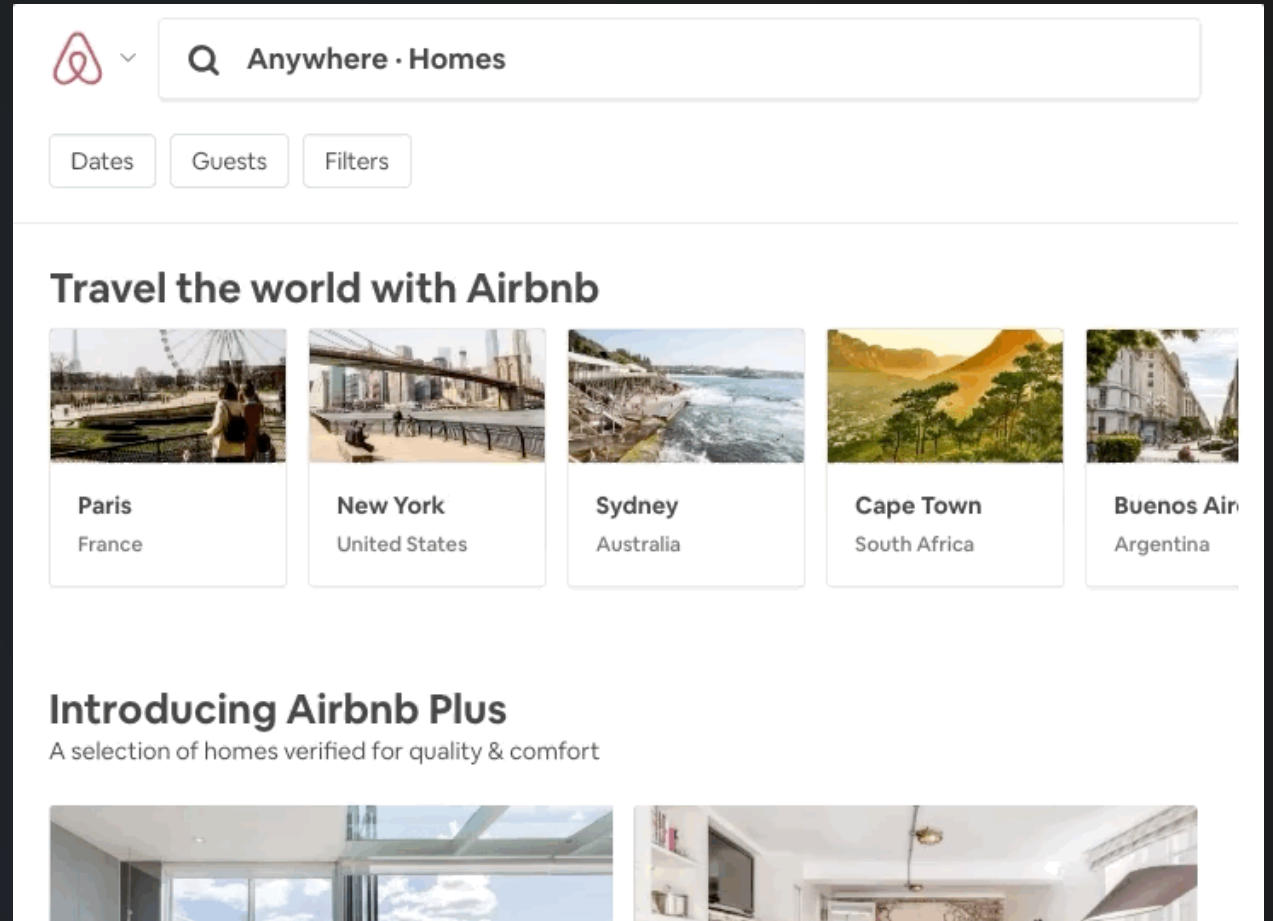
Nikhil Simha



Background

Product

A global travel community of amazing hosts and cool places.





ML @ Airbnb

Q Hawaii Big Island Piano Tuning, Kaaupahu Home...

Dates Guests Home type Price Instant Book Trip type More

Over 1,700 guest reviews for The Big Island homes, with an average of 4.8 out of 5 stars

58 homes



ENTIRE CABIN · 1 BED
Hawaiian Tropical Hideaway Cottage
\$125 per night · Free cancellation
★★★★★ 138 · Superhost



ENTIRE LOFT · 2 BEDS
Cedar House Loft with Hot Tub
\$105 per night · Free cancellation
★★★★★ 133 · Superhost



ENTIRE BUNGALOW · 1 BED
Cozy Ohana Cottage in Hawaii
\$85 per night · Free cancellation
★★★★★ 82 · Superhost



ENTIRE HOUSE · 2 BEDS
Hale Keawe Iki- A Lovely Home in the heart of Hawaii
\$124 per night · Free cancellation
★★★★★ 47 · Superhost

Search Ranking

\$125 per night
★★★★★ 138

Dates
10/17/2018 → 10/20/2018

Guests
2 guests

\$125 x 3 nights	\$375
Cleaning fee ?	\$75
Service fee ?	\$64
Total	\$514

Pay less up front and pay the rest near check-in.
No extra fee. [Learn more](#)

Book

You won't be charged yet

This is a rare find.
Ed's place is usually booked.



Smart Pricing



Risk scoring

Every Airbnb reservation is scored for risk before it's confirmed. We use predictive analytics and machine learning to instantly evaluate hundreds of signals that help us flag and investigate suspicious activity before it happens.

Fraud detection

Team



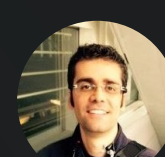
Atul Kale
Engineering Manager



Andrew Hoh
Product Manager



Aaron Siegel
Director, Data Platform



Bahador Nooraei
Engineer



Varant Zanoian
Engineer



Alfredo Luque
Engineer



Evgeny Shapiro
Engineer



Mani Bushan
Engineer



John Park
Engineer



Nikhil Simha
Engineer



Conglei Shi
Engineer



David Zhang
Engineer



Patrick Yoon
Engineer



Xiaohan Zeng
Engineer



Andrew Cheong
Engineer



Why?



THIS IS YOUR MACHINE LEARNING SYSTEM?

YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.

WHAT IF THE ANSWERS ARE WRONG?

JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.





Four part problem

Data

Modeling

Predictions

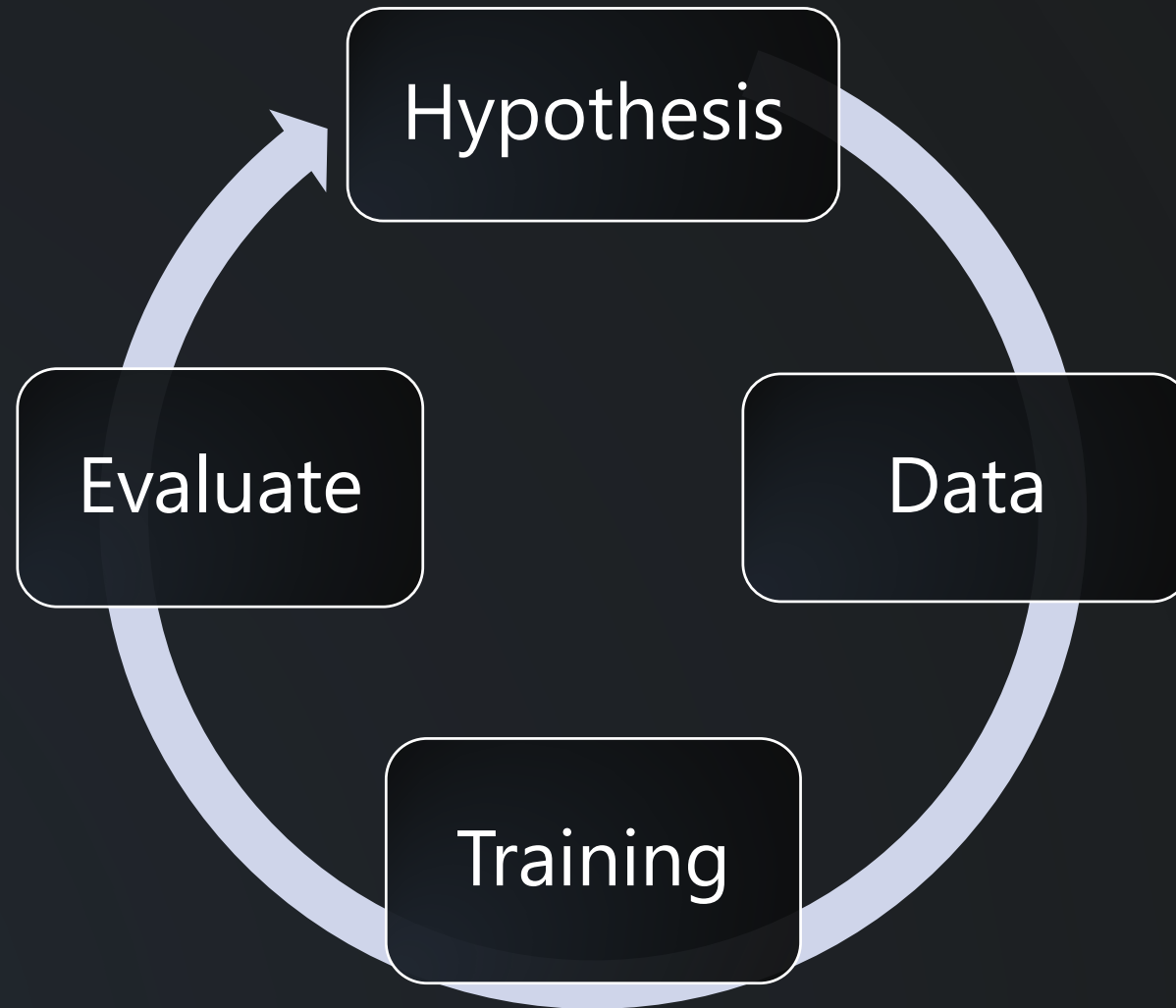
Scientist tooling



Machine learning lifecycle

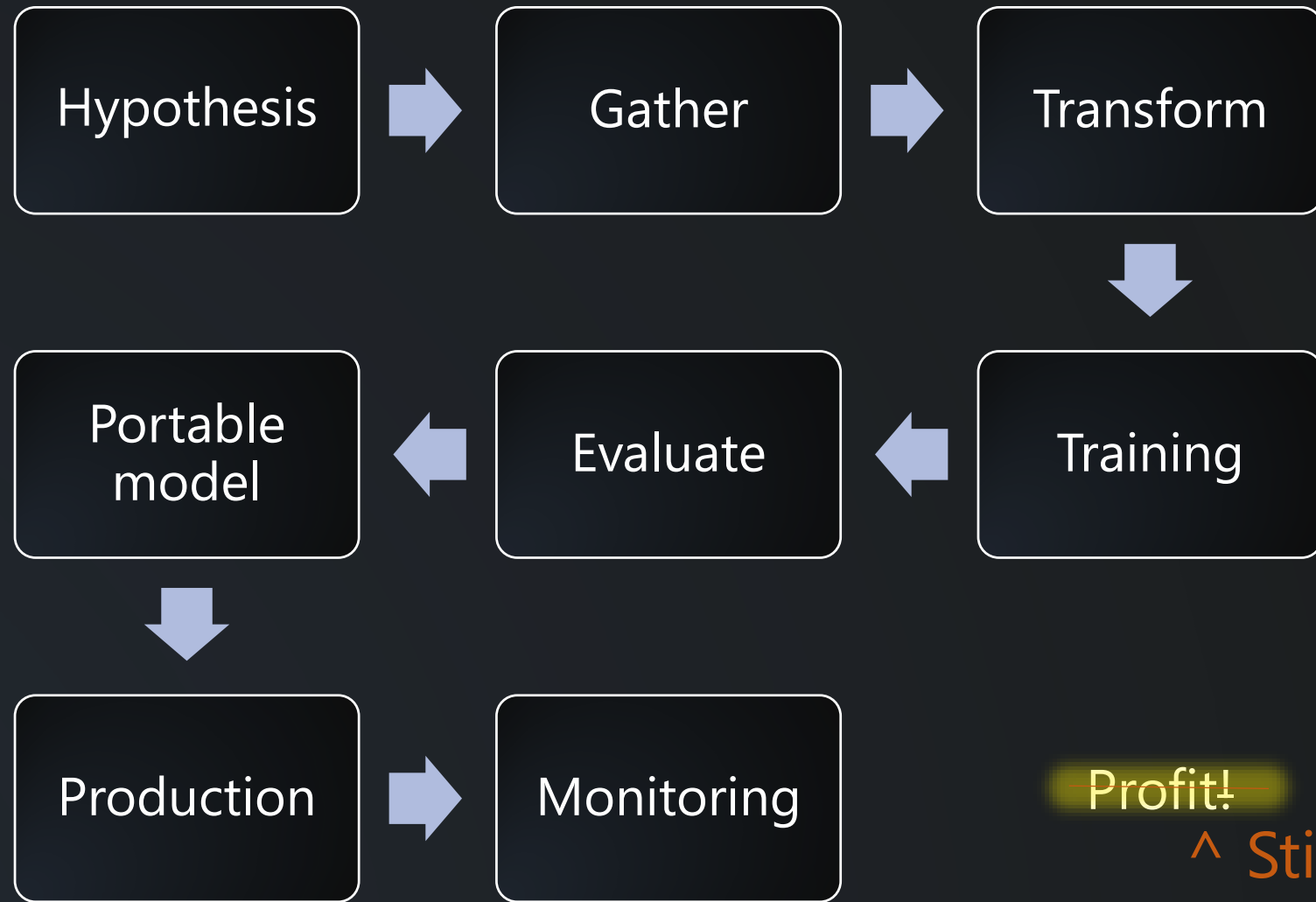


Profit! —
^ Lie





Machine learning lifecycle – For Realz



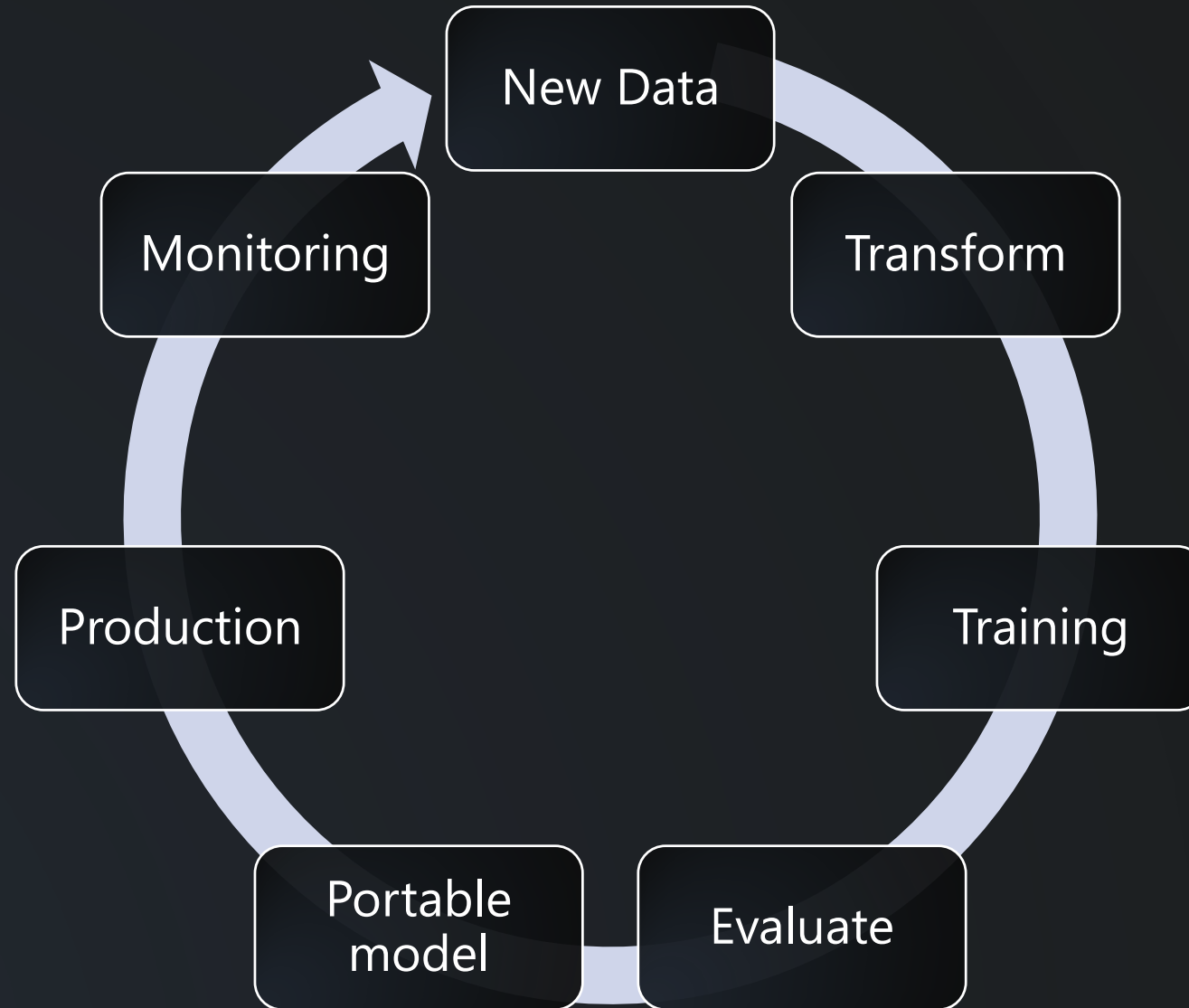
~~Profit!~~
^ Still a Lie



Everything changes

- Features
- Model
- Data





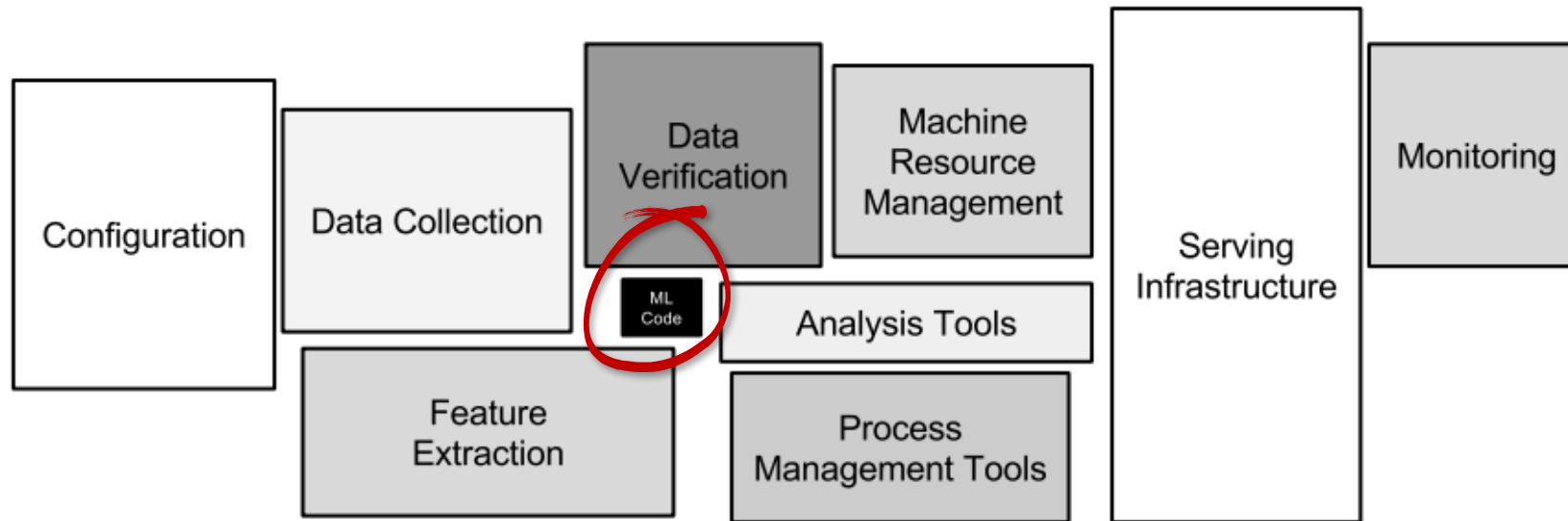


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

"We recognize that a mature system might end up being (at most) 5% machine learning code and (at least) 95% glue code"

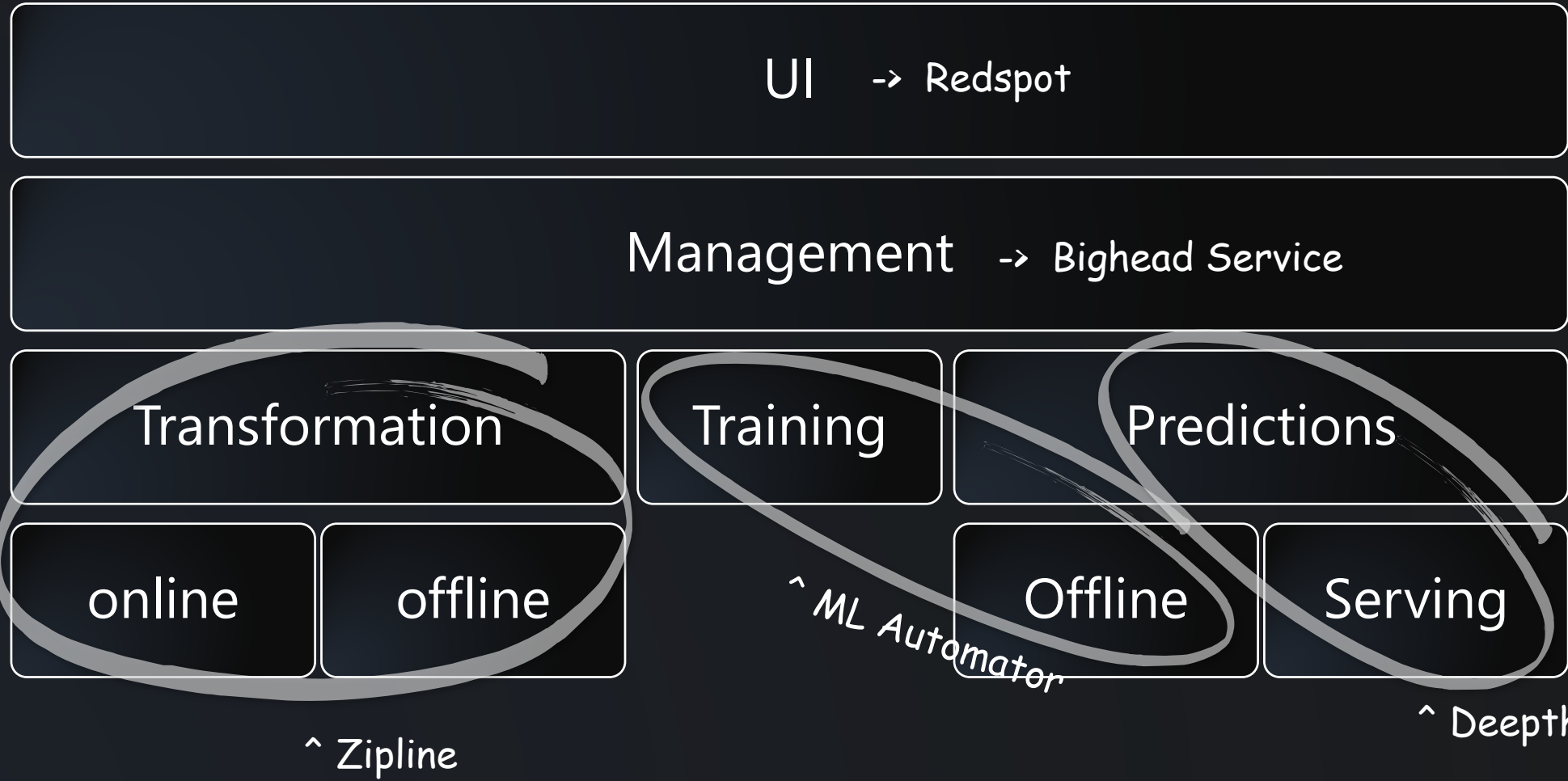


A principled approach



Necessary conditions

Versatility	Many models, many data sources
Consistency	Same model and data
Scalability	Transformation + Training + Serving
Reliability	





Feature Engineering

"Coming up with features is difficult, time-consuming, requires expert knowledge. 'Applied machine learning' is basically feature engineering."— Prof. Andrew Ng.

Take a guess: In your machine learning project, how much time will you typically spend on data preparation and transformation?

✓ More than half of the project time



Correct: you'll spend the majority of time on a machine learning project constructing data sets and transforming data.

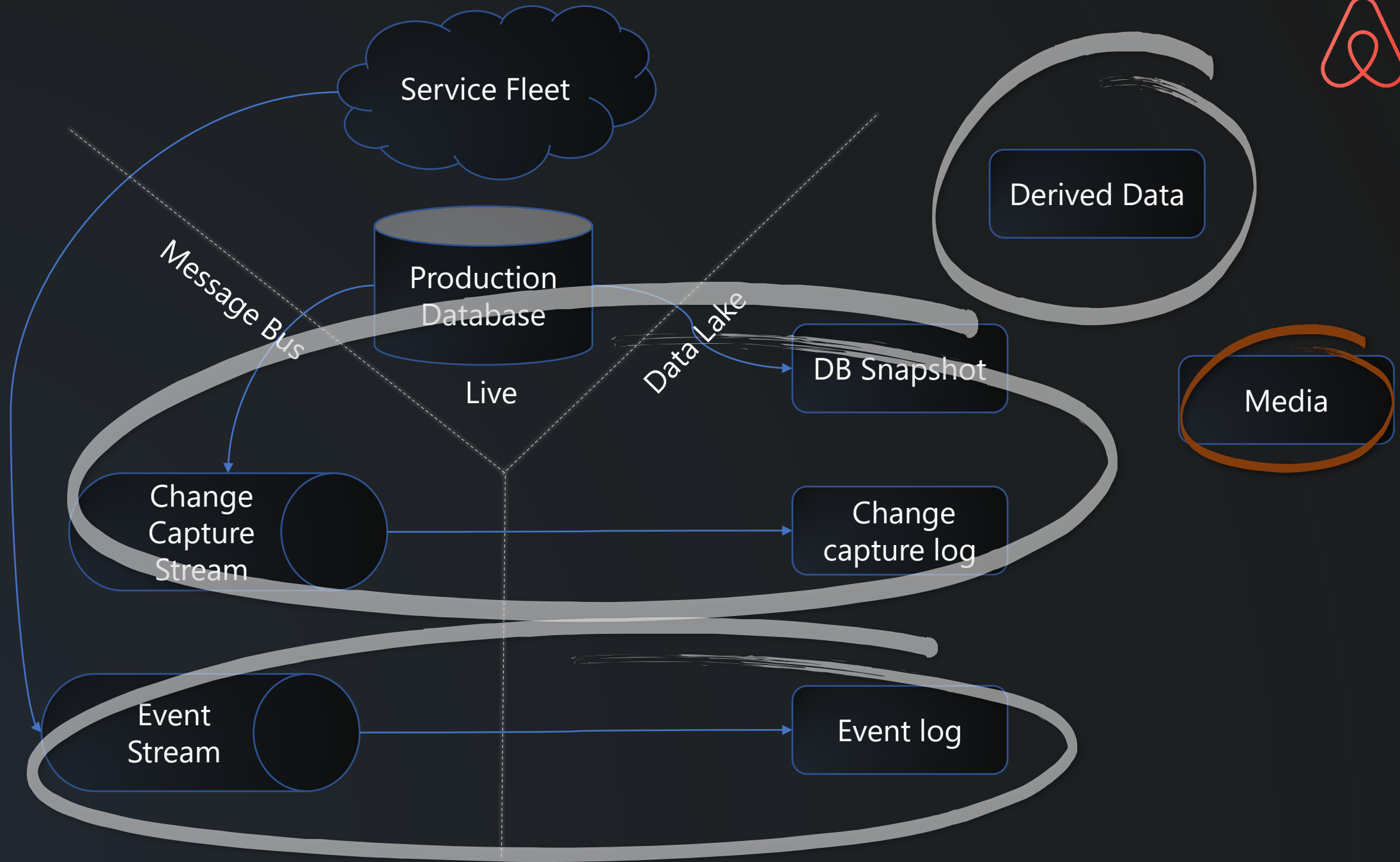
Correct answer.

Less than half of the project time





Data Architecture of modern web companies





An example

- Predict likelihood to book when a user views an experience
- Example Feature:
 - Sum(bookings) in last 7 days

Paris' Best Kept Secrets Tour

History experience
Hosted by [Olivier, Charles & Fabien](#)



 [Paris](#)

 4 hours total

 Equipment

 Offered in English

About your host, Olivier, Charles & Fabien

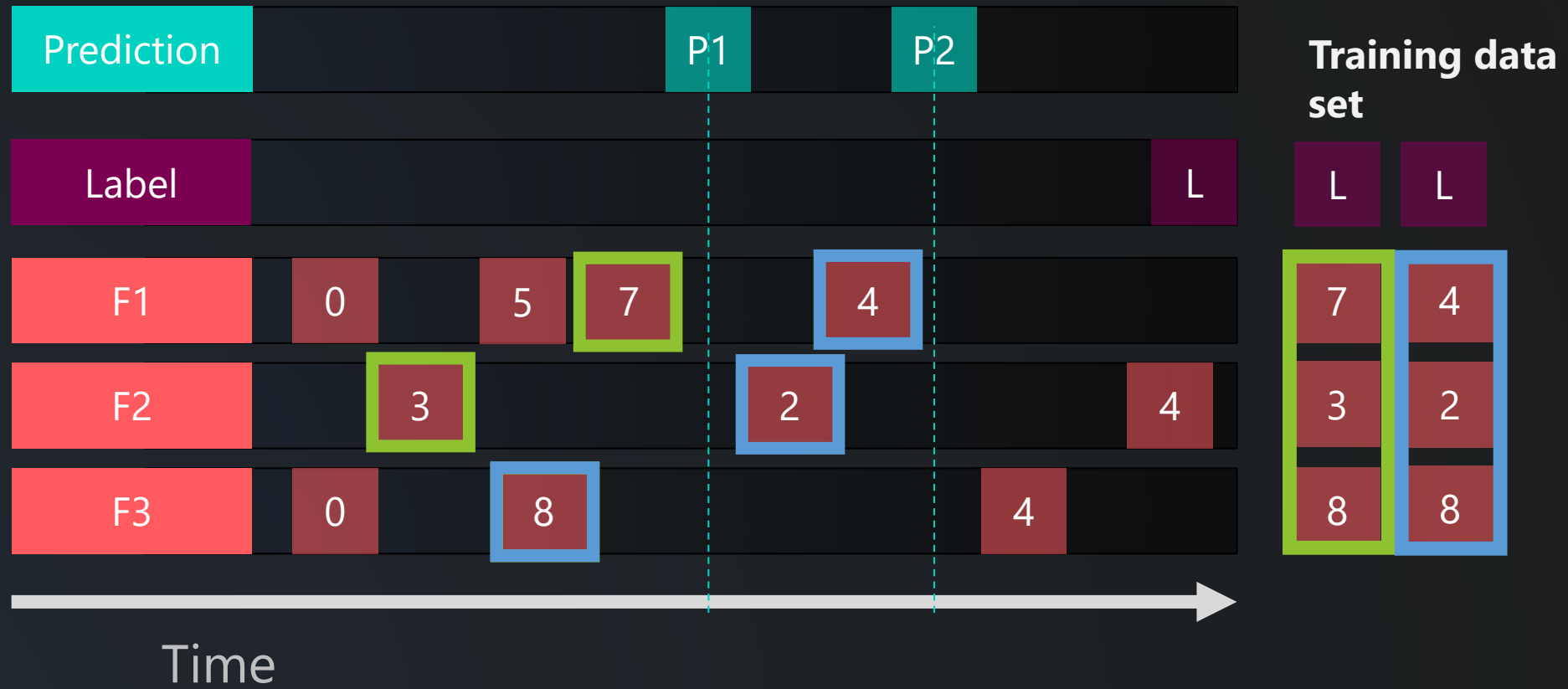
Inhabited by a passion for history, I gave myself 4 years of intense research, particularly at the National Archives in Paris. I was fortunate to be able to work hand in hand with some of the greatest historians on a subject that I am deeply passionate about: the mistakes in the French and in Paris' history.

What we'll do

[Save to list](#) 



Life of a feature





Concepts

Sources	Bookings Table
Keys	User id
Feature Sets	Count(booking_id, window="7d")
Training Sets	search event log



Feature Set Example

```
source: {  
  type: hive  
  query: ""  
    SELECT  
      , guest as user  
      , ts  
      , value as value  
      1 as count  
    FROM bookings  
    WHERE  
      ds BETWEEN '{{ start_date }}' AND '{{ end_date }}'  
  dependencies: [bookings]  
}
```

```
features: {  
  count: {  
    doc: "Total bookings."  
    column: count  
    operation: sum  
    windows: ["7d", "30d", "180d", "1y"]  
  },  
  avg_price: {  
    doc: "Average booking value."  
    column: value  
    operation: avg  
    windows: ["7d", "30d", "180d", "1y"]  
  }  
}
```



Training Set Example

```
from zipline.training_set import TrainingSet

features = [
    "user_bookings_past_7d",
    "user_bookings_past_30d",
    "listing_booking_requests_past_7d",
    "listing_booking_requests_past_30d"

query = '''
SELECT
    id_guest as user,
    id_listing as listing,
    ts as ts
FROM some_table
WHERE ds BETWEEN '{{ start_date }}' AND '{{ end_date }}'
'''

team = 'ml_infra'
name = 'source_test_3'
start_date = '2018-07-01'

training_set = TrainingSet(team, name, features, query, start_date)
```



Output Example

User provides keys			Zipline fills in features	
user	listing	time	bookings_sum_7d	bookings_sum_14d
123	567	2018-01-01 ...	1	2
234	678	2018-01-01 ...	4	4
345	789	2018-01-02 ...	0	1



Guarantees

Online-offline Consistency

Point-in-time accuracy

Efficient backfills

Shared Repository



Techniques

Monoids

Massive Parallelism

Interval Trees

Data compression

Tetris algorithm

Backfill efficiency

Lambda architecture

Data freshness

And More



Big Picture

