Jaime Pabon

jpabonr

Digital Product Management Powered by Machine Learning

Case Study: Predicting User Retention From Early User Behavior

Jaime Pabon

jpabonr

- 13+ years experience digital marketing and digital product mgmt.
- Head of Strategy, The Online Project Dubai (digital agency)
- Ex-founder tech startup
- Industrial Engineer
- Machine Learning Engineer

This is *not* a presentation about the technical aspects of machine learning!

It’s about how machine learning can be applied to digital product strategy.

Show of Hands

Can you explain the very basic idea behind machine learning to a friend?

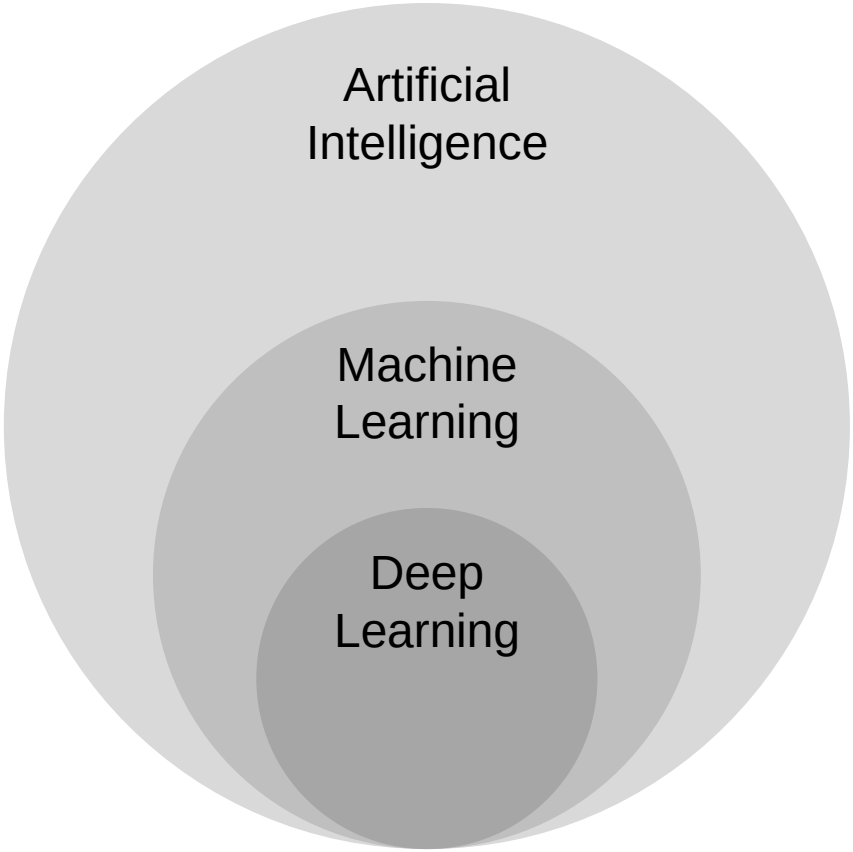
Image: Clipart Panda

Machine Learning: The “Very Basic Idea”

A computational technique that uses algorithms to find patterns in data.

ML systems *learn* to complete tasks by ‘seeing’ examples (not by following instructions).

For the full definition go to Wikipedia…



Applied Machine Learning

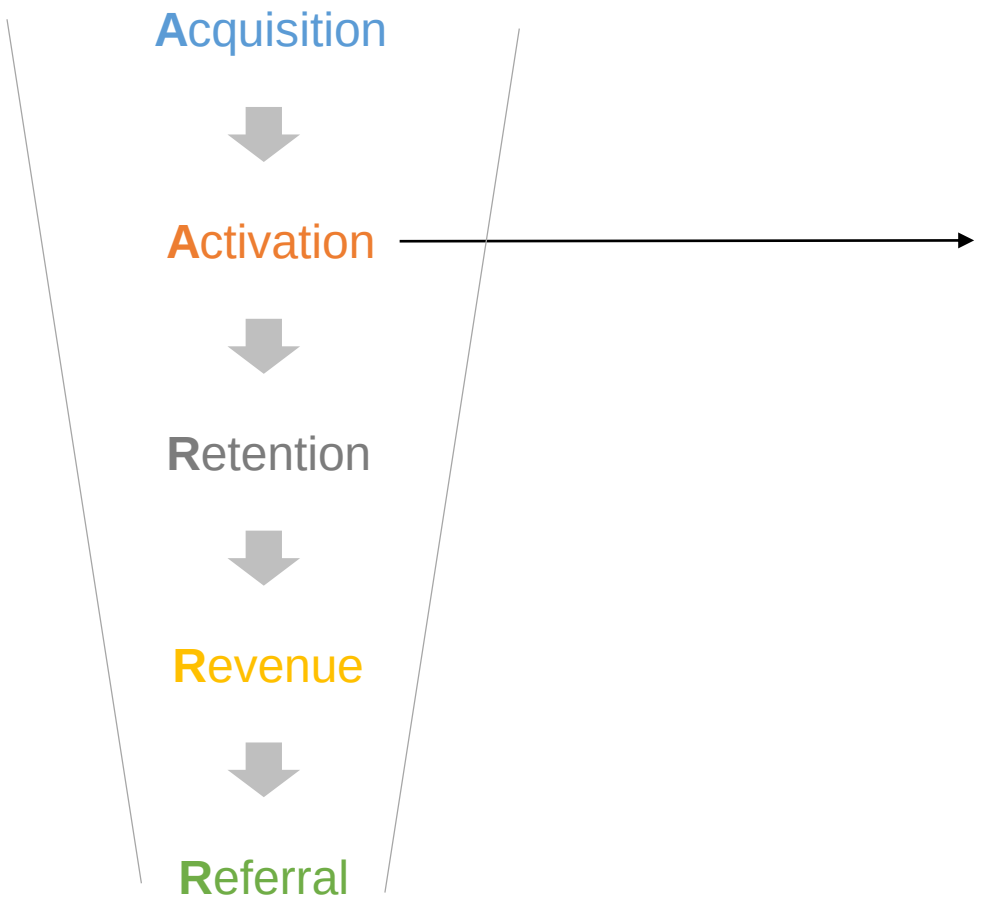
Machine Learning in The Industry

Decision-making

Product

Operations

The AARRR Metrics Model



The key user action that triggers the product's delivery of value.

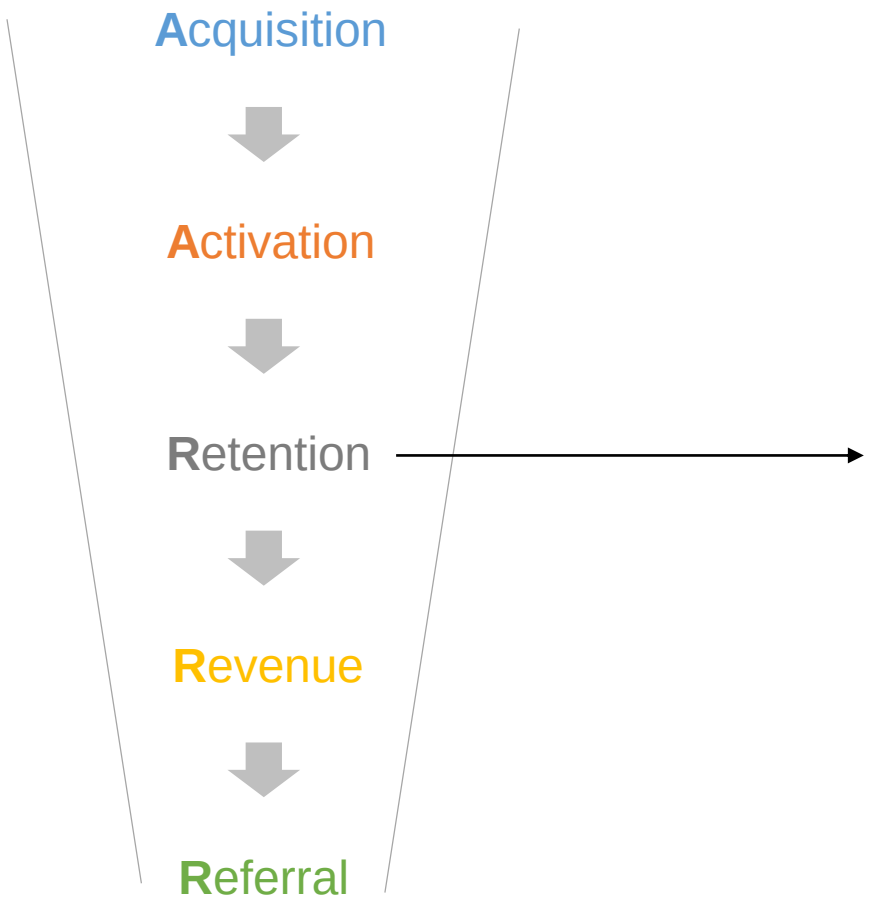
aka

"the north star"
"The ONE metric that matters"

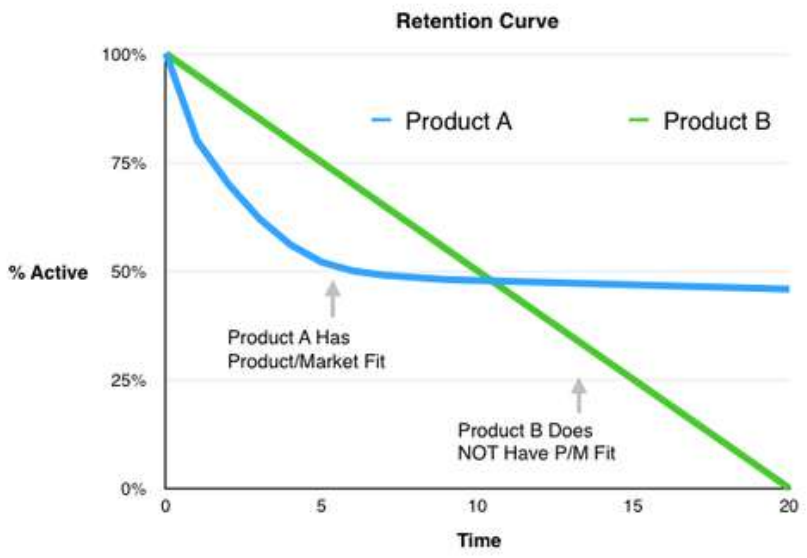
e.g.

Facebook – Connect with *7 friends in 10 days*

User Retention is a Big Deal

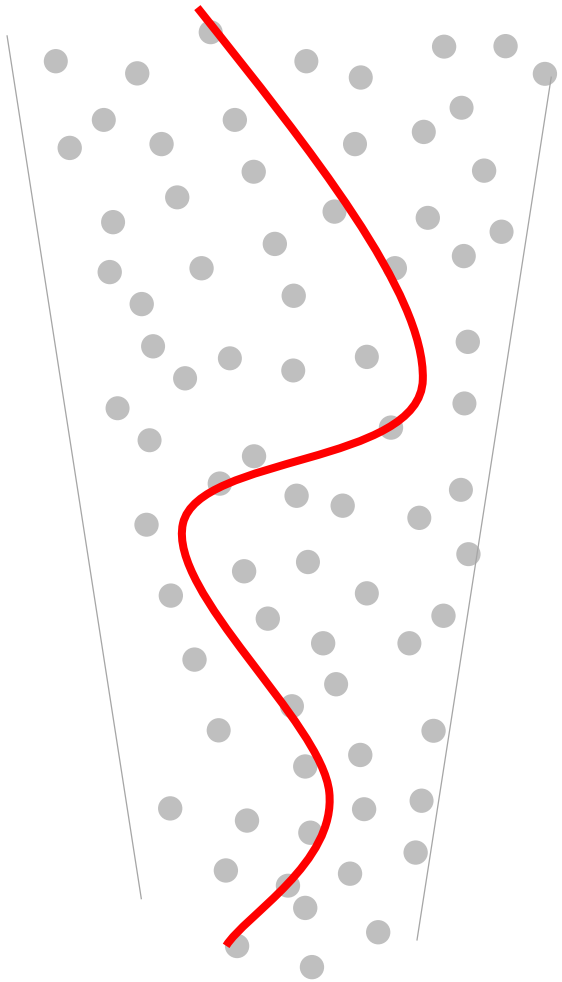


User retention determines Product-Market Fit



Source: Brian Balfour

The Big Idea

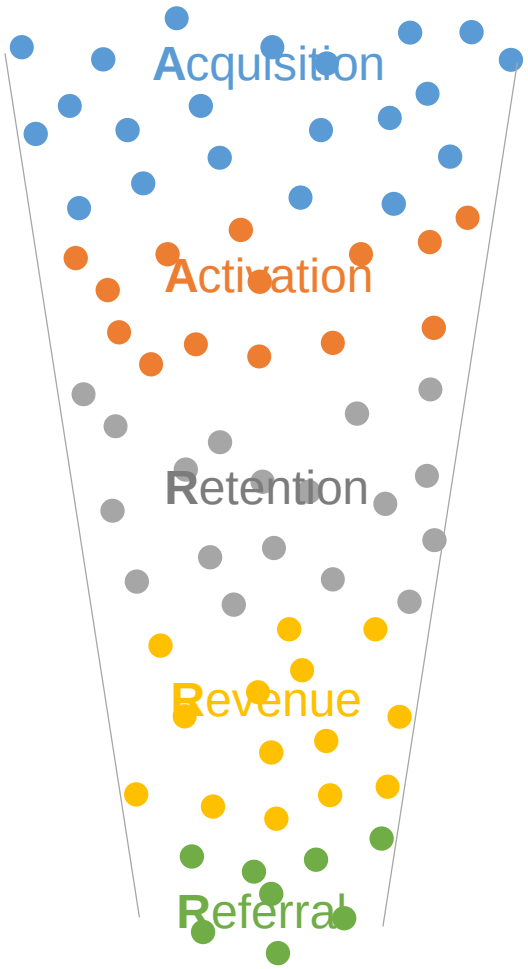


There's a hidden pattern that cuts through the funnel

This pattern can be discovered with machine learning.

Highly efficient product and growth strategies exploit this pattern.

Data Greedy!



Common product management knowledge says it's not about tracking *all* the data, but the *right* data.

But…

To identify the right data we first need "all" the data.

Correlation vs Causation

Use ML to find correlations and formulate causation hypotheses

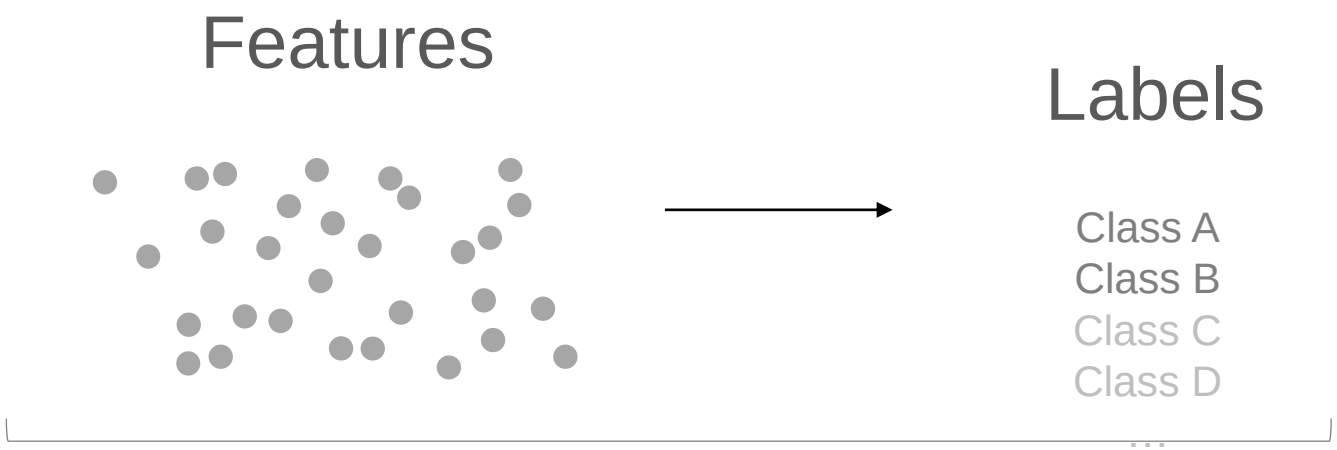
→

Validate causation with A/B Testing

Benefits of Predicting Long-Term User Retention From Early User Behavior

1. No need to wait: early decision-making leading to faster prod. Iteration
2. Improved efficiency by focusing on leading indicators
3. Customized UX for users with negative retention predictions
4. Improved accuracy of LTV estimation (retention, revenue)
5. Improved customer acquisition budgeting (retention, CPA)

Supervised Learning – Classification: Learning to Mapping Features (Inputs) to Labels



An algorithm learns the mapping function

Data vs. Features (Feature Engineering)

A log in the DB

User action

Potential feature

- Intensity
- Frequency
- Time
- Sequence
- Interpretation
- Absence of

This is where domain knowledge and creative thinking matters.

Case Study

Predicting User Retention From Early User Behavior

CLOSETREMIX

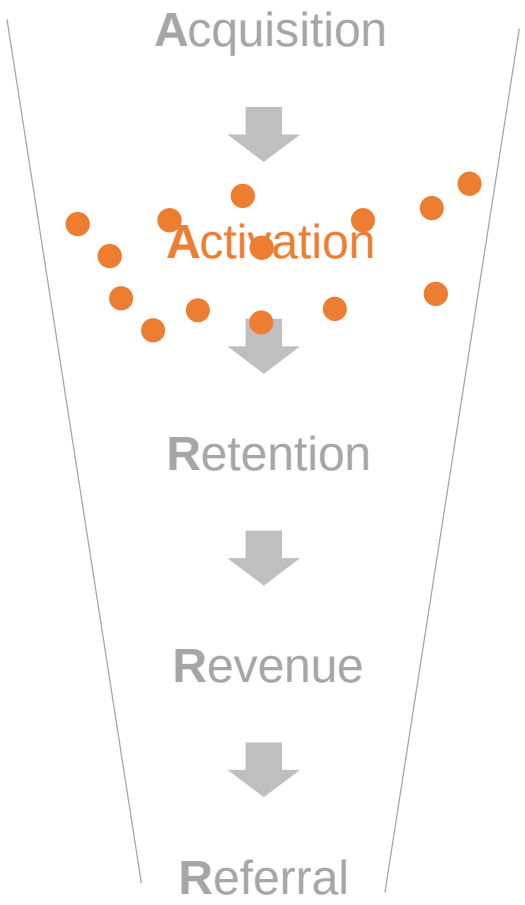
CLOSETREMIX

Mobile social network where young women upload their closets to receive outfit ideas from friends.*

* CLOSETREMIX operated in 2014 – 2015. What follows is a post-mortem analysis.

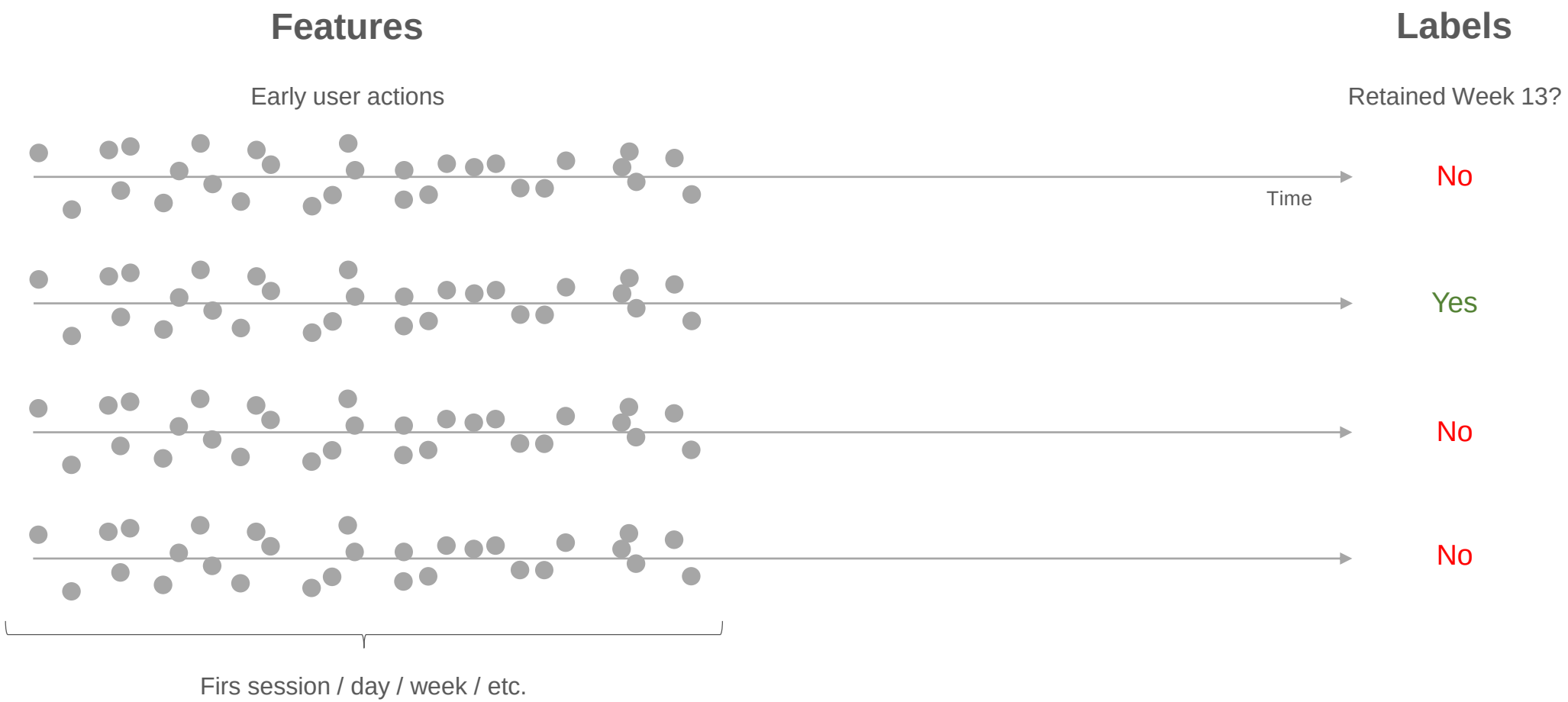
 jpabonr

Challenge: To Figure Out The Activation Metric



What early user action puts him/her on the path of long-term retention?

Supervised ML > Two-Class Classification Problem

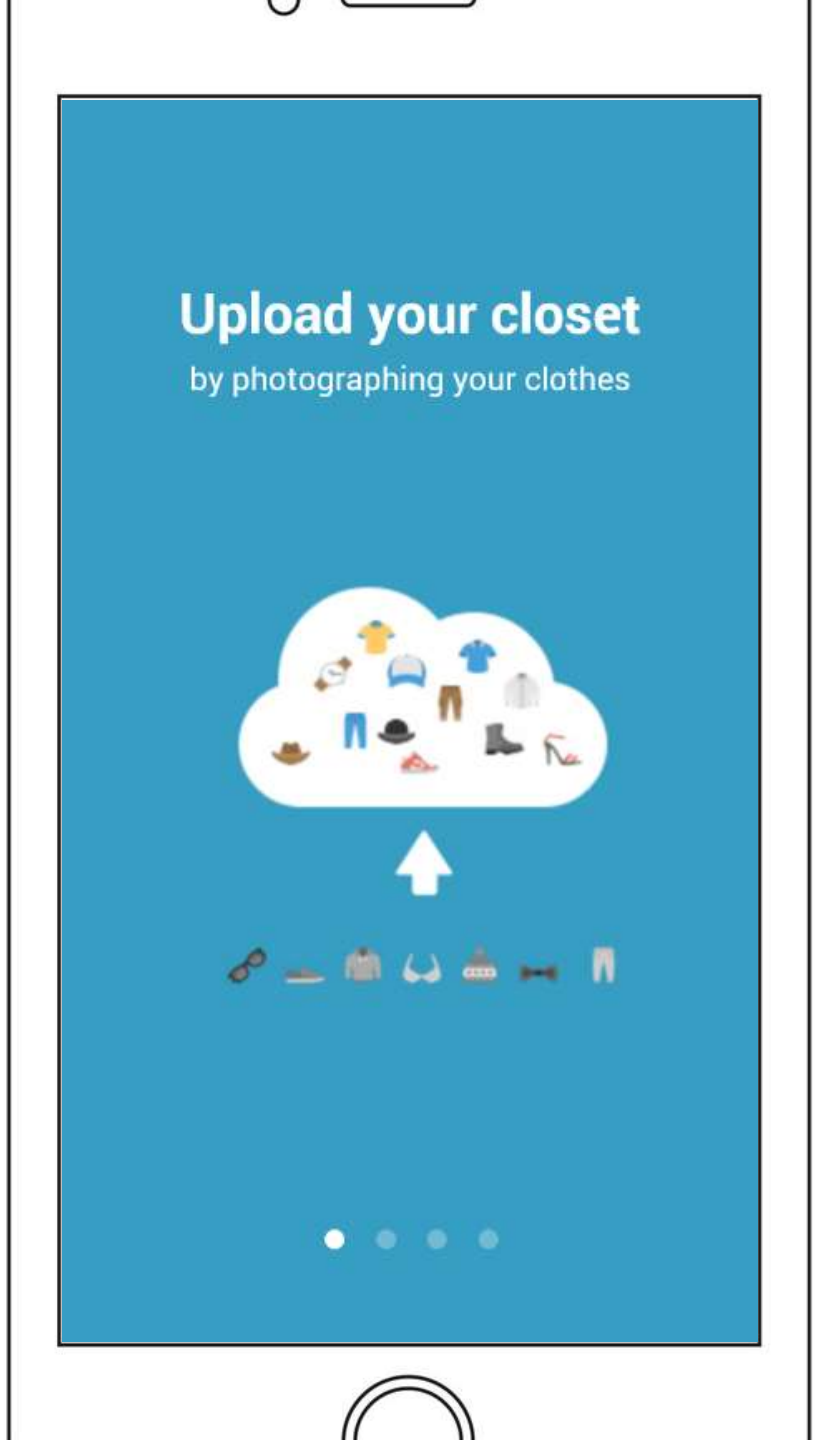


Early User Behavior: User Onboarding

User onboarding plays a big role in early user actions. It's then useful to review how CLOSETREMIX user onboarding used to be.

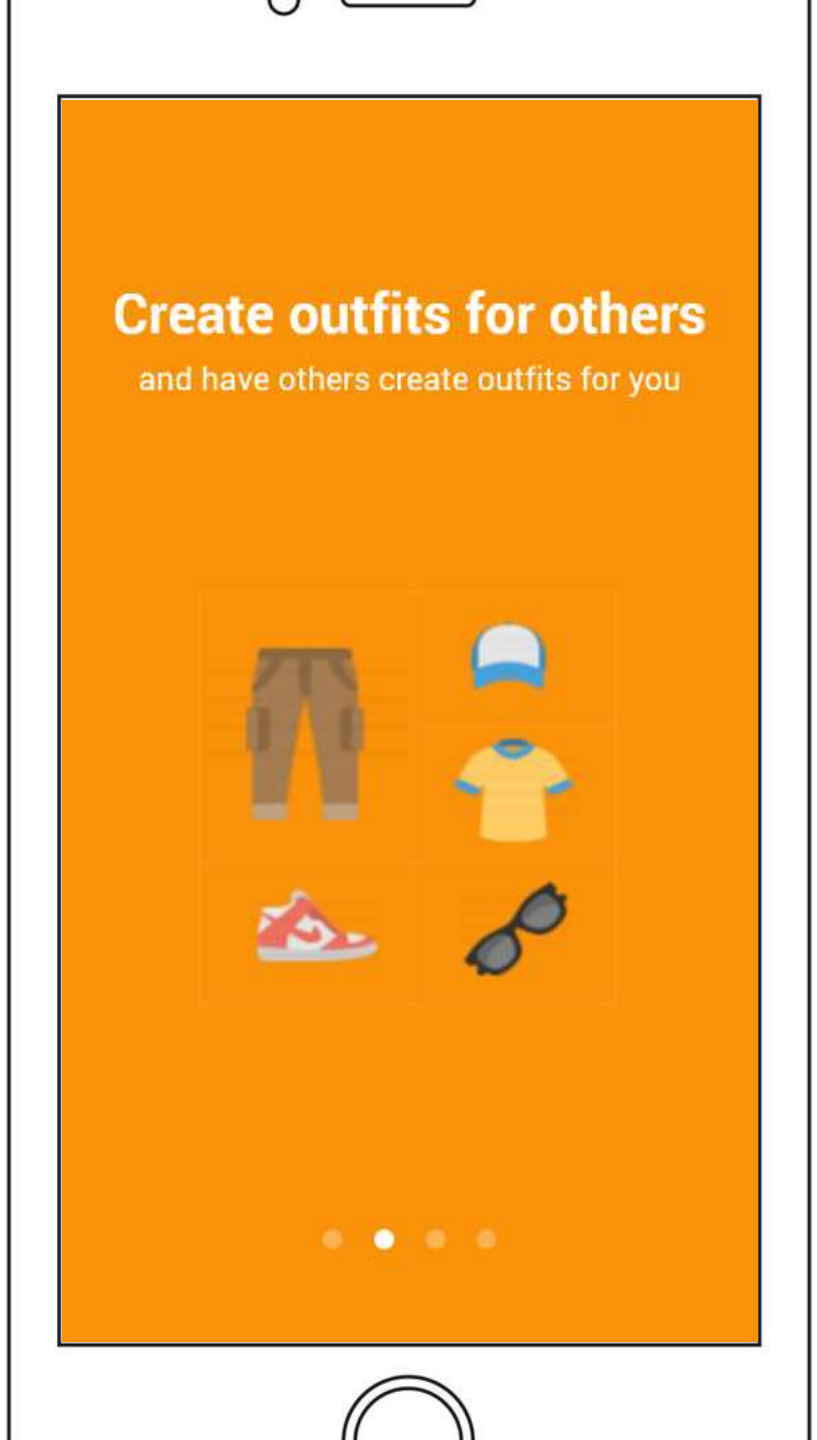
User onboarding

Slide 1



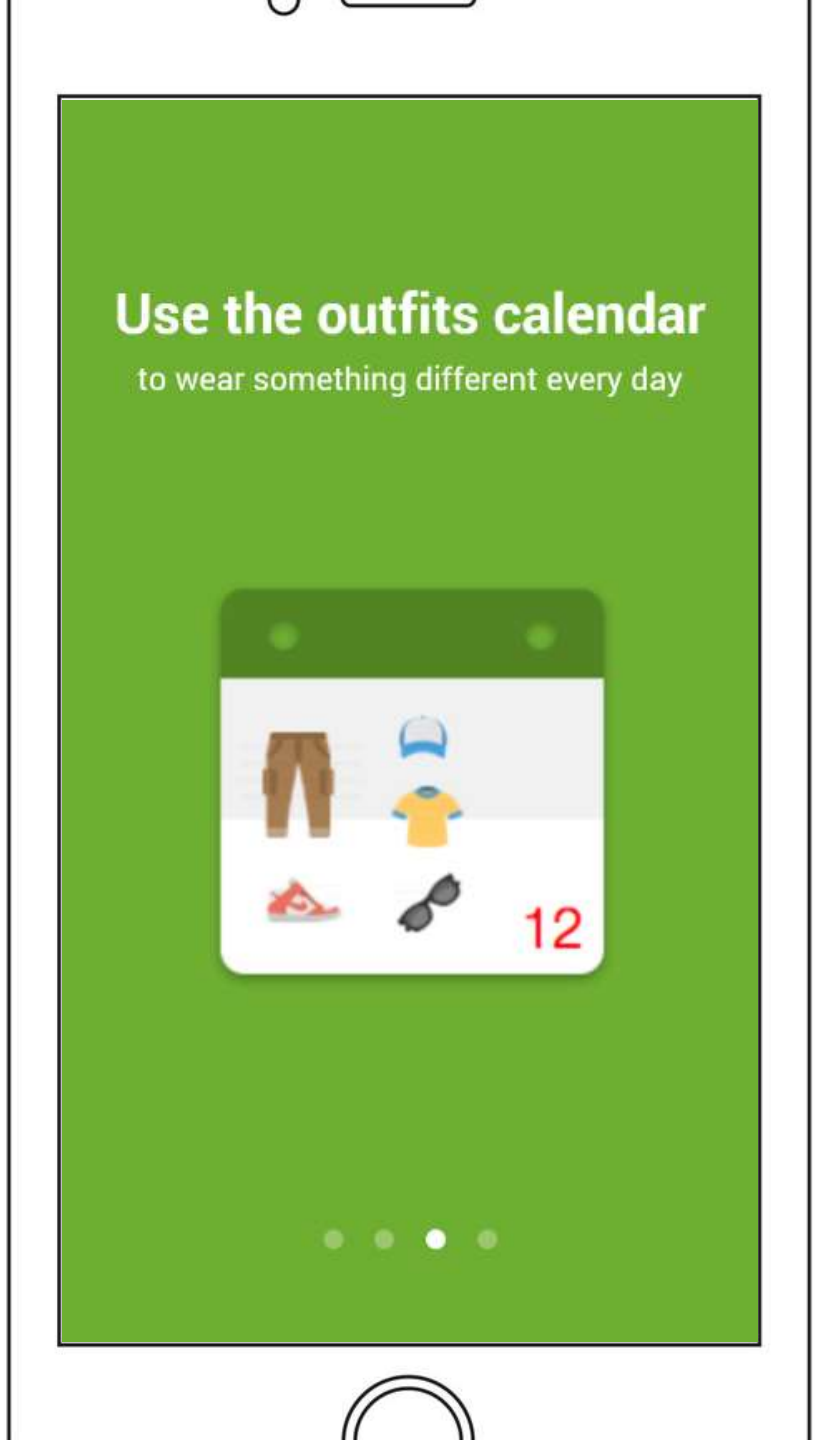
User onboarding

Slide 2



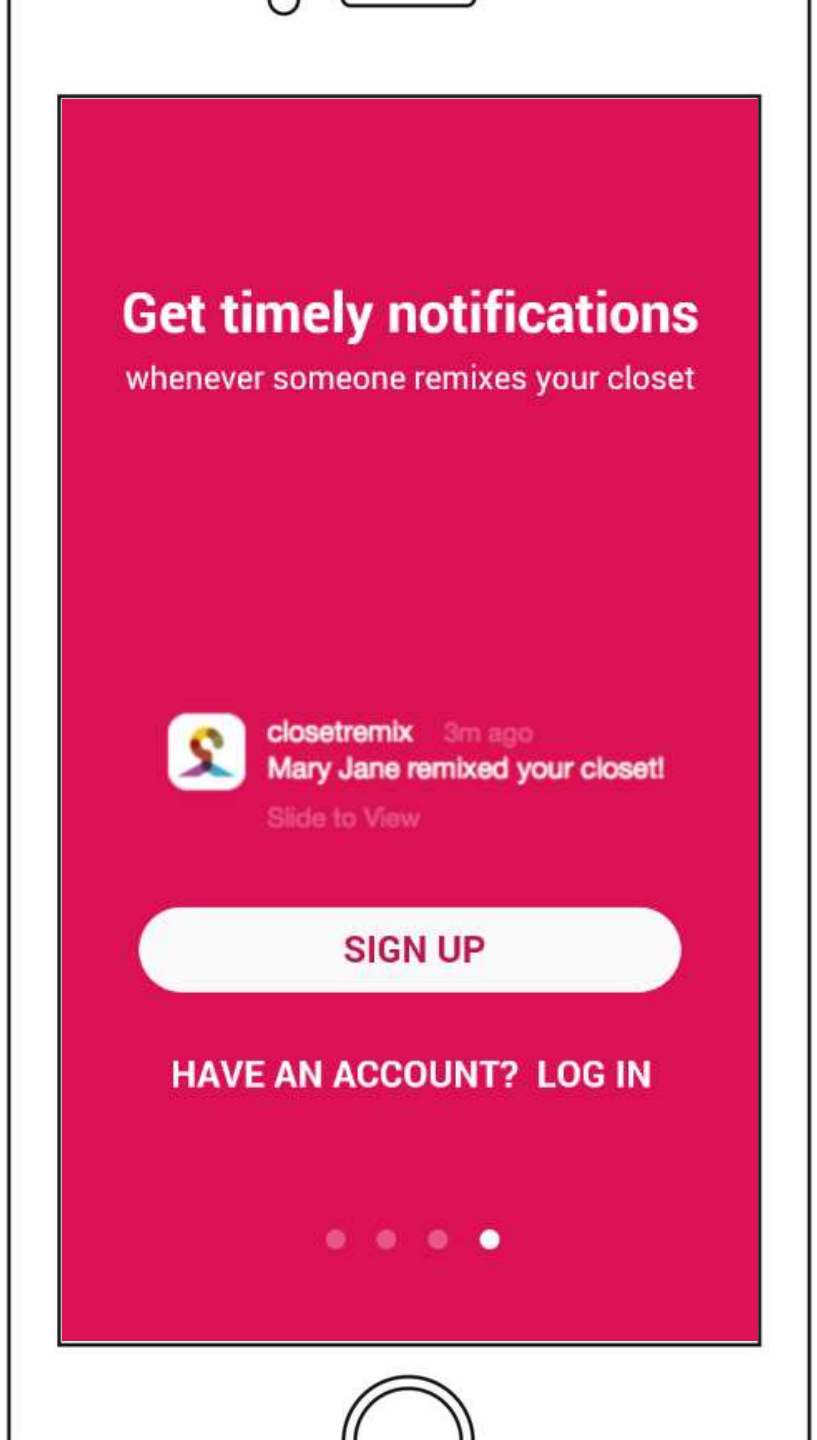
User onboarding

Slide 3



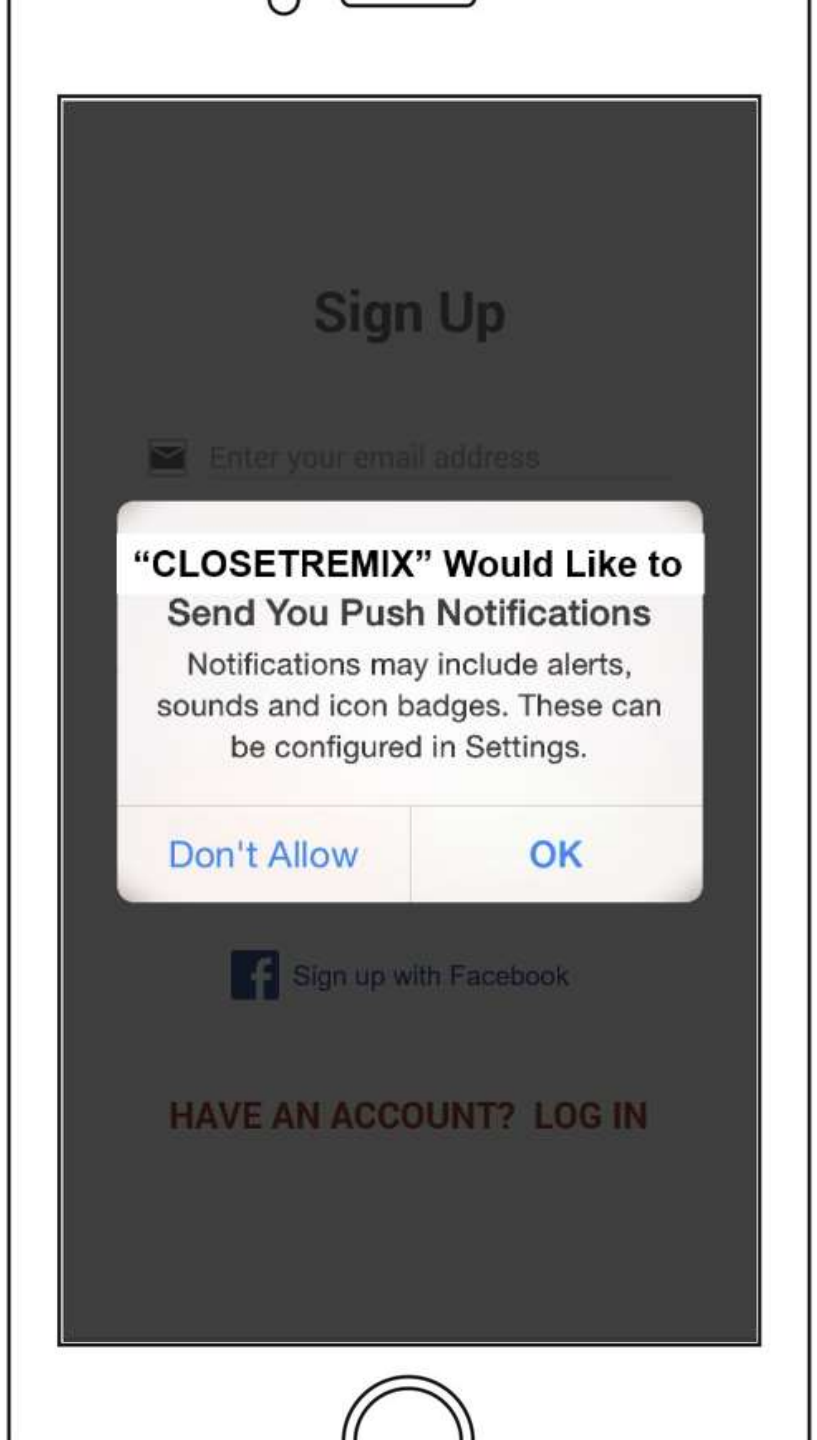
User onboarding

Slide 4



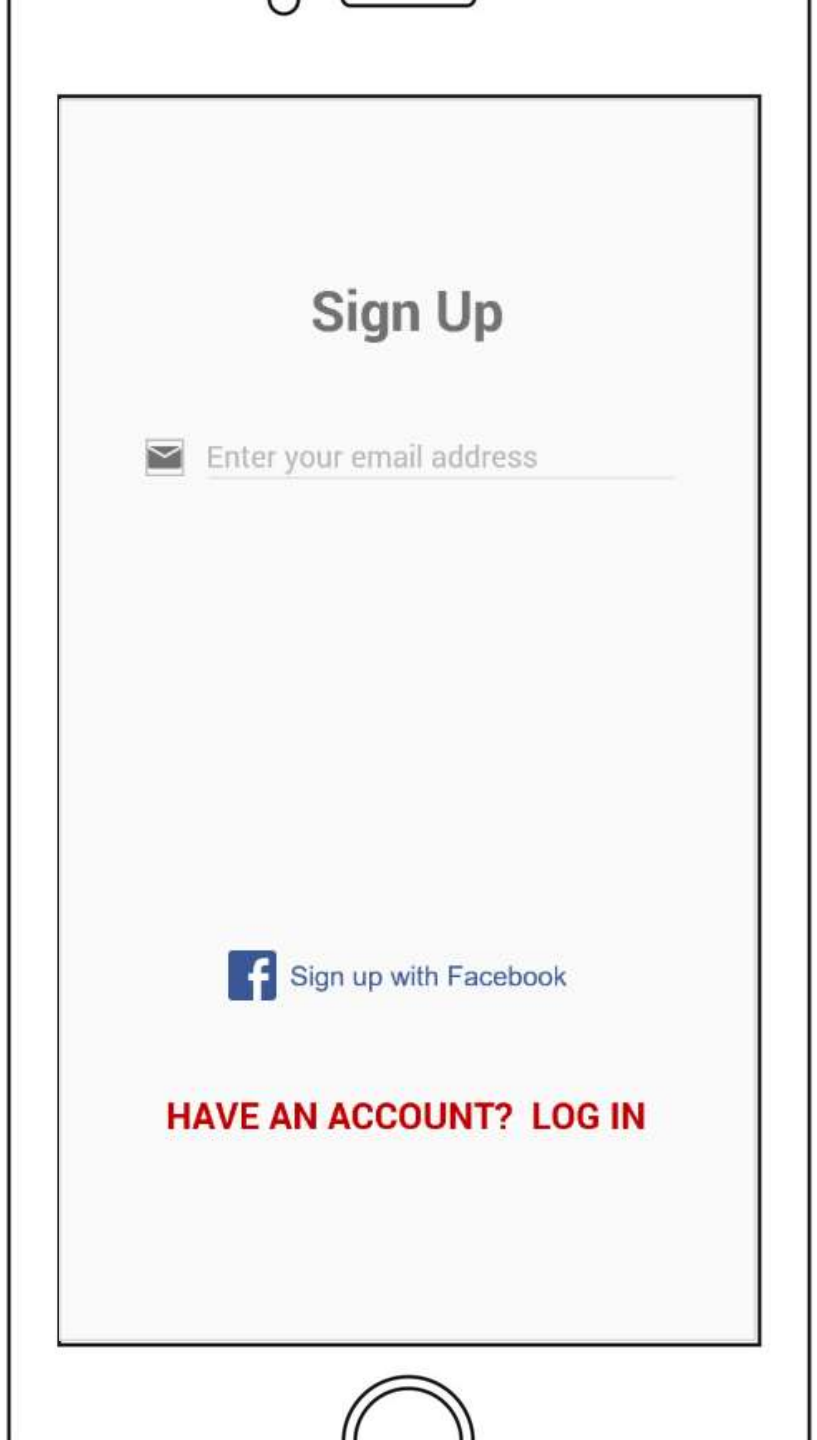
User onboarding

iOS Dialog

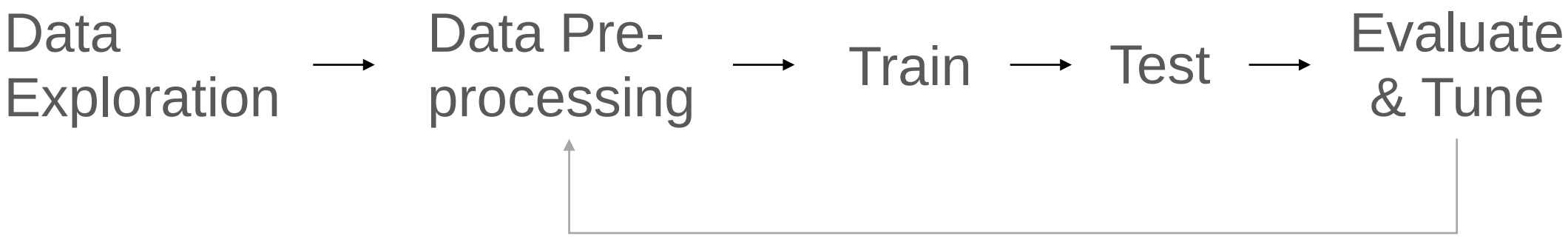


User onboarding

Sign Up



The (Simplified) Machine Learning Process



Model Implementation – Python and Scikit Learn

Model 3: Gaussian Naive Bayes (GNB) Classifier

Model 2: Support Vector Machines (SVM) Classifier

Model 1: Random Forest (RF) Classifier

```
from sklearn.ensemble import RandomForestClassifier

param_n_estimators = range(1, 11)

beta = 10 # # Beta value for F-score that makes strong empahsis on recall.

# Initiates/clears all the stored grid search performance scores
f10_scores_grid_search = np.array([])
recall_scores_grid_search = np.array([])
precision_scores_grid_search = np.array([])
accuracy_scores_grid_search = np.array([])

# Conducts a grid search to opitmize 'n_estimators'
for param in param_n_estimators:

    # Initiates/clears the performance scores
    f10_scores_all_splits = np.array([])
    recall_scores_all_splits = np.array([])
    precision_scores_all_splits = np.array([])
    accuracy_scores_all_splits = np.array([])

    # Instantiates stratified shuffle
    sss = StratifiedShuffleSplit(n_splits=10, test_size=.2, random_state=0)

    # Loops through all the stratified random splits and trains/tests a RF algorithm
    for train_index, test_index in sss.split(features_final_compressed, labels_final):

        # Generates training feature and label sets
        X_train, y_train = features_final_compressed[train_index], labels_final[train_index]
```

Data Exploration: Understanding The Features

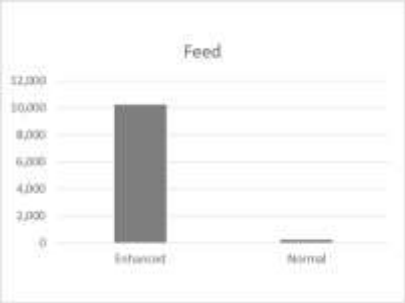
Feature	Values
Sign Up Day-Of-Week	Monday, Tuesday, …, Sunday
Push Notification	On, Off
Email Notification	On, Off
Sign Up Method	Email, Facebook
Onboarding Invite	Yes, No
Items In Closet	Integer
Outfits In Closet	Integer
Privacy	Public, Private
Feed	Normal, Enhanced
Outfits Oneself	Integer
Outfits Other	Integer
Following	Integer
Followers	Integer
Days 1st item	Same Day, Next Day, 2-7, 8-30, 31-60, 61+, Never
Day-Of-Week 1st Item	Monday, Tuesday, …, Sunday
Days 30th item	Same Day, Next Day, 2-7, 8-30, 31-60, 61+, Never
Days 1st Outfit Others	Same Day, Next Day, 2-7, 8-30, 31-60, 61+, Never

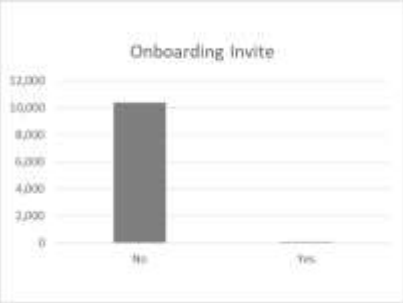
→

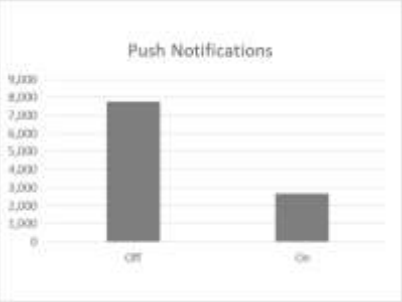
6 categorical Variables

11 Continuous Variables

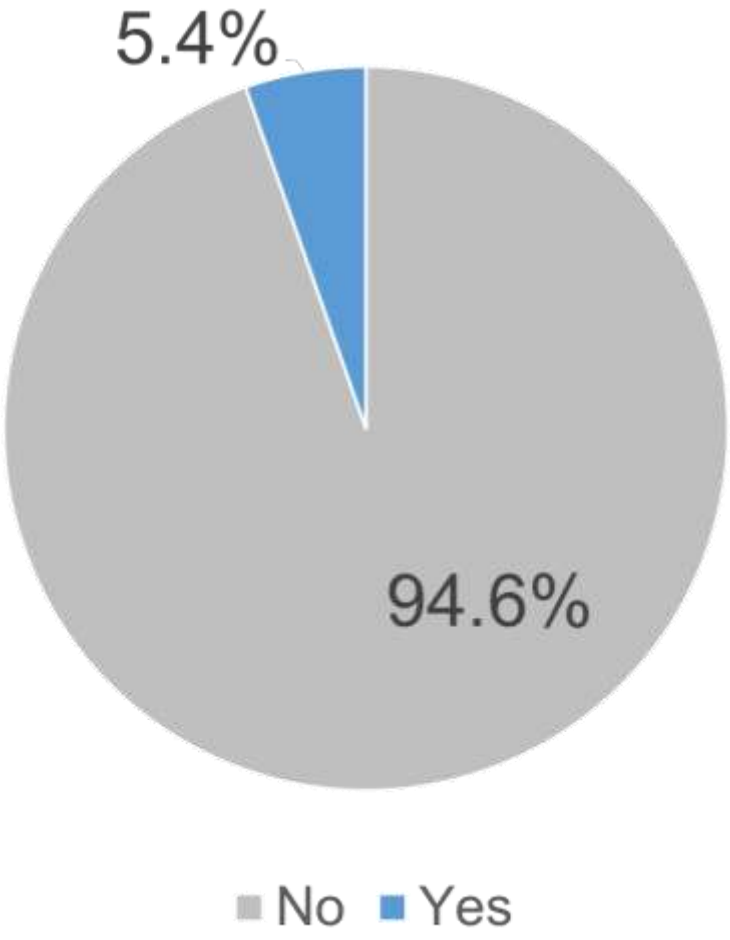
Data Exploration: Visualizations Are Key







Data Exploration: Labels – User Retention Week 13



A case of skewed labels
(class imbalance)

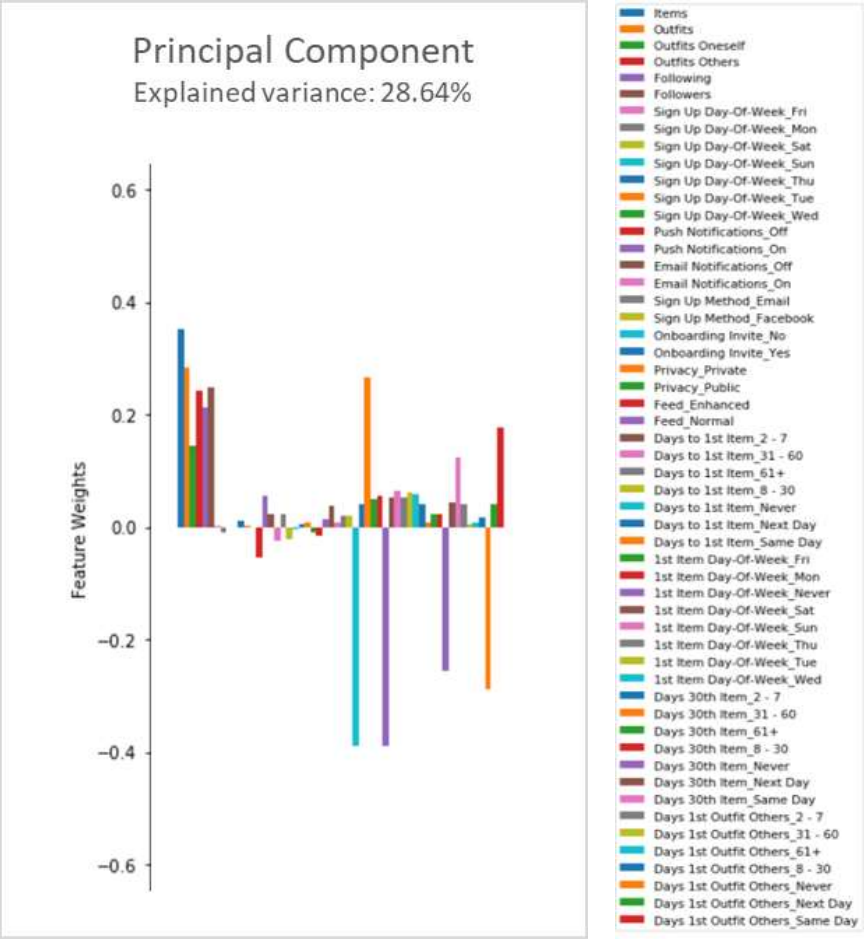
↓

Compensate with oversampling

↓

Model Evaluation: F10-score to
minimize false negatives

Data Pre-processing: Principal Latent Variable - PCA



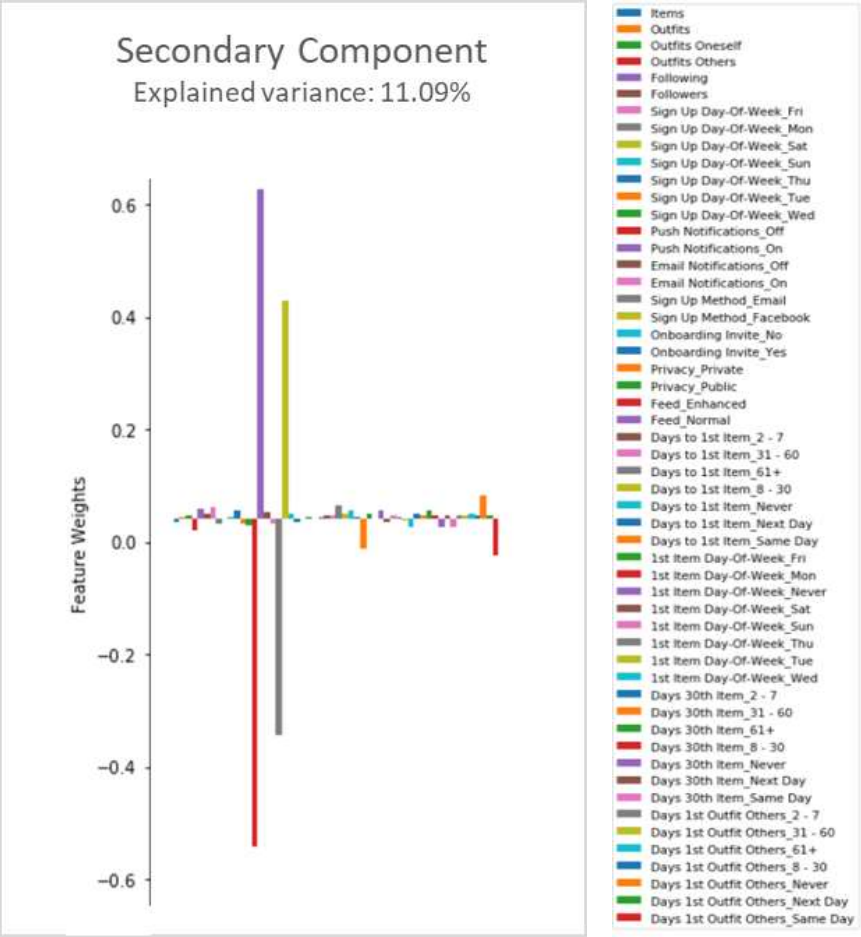
Moderate-to-high positive correlation (feature weight >= 0.1) with:

- 'Days to 1st Item: Same Day'
- 'Items'
- 'Days Outfits Others: Same Day'
- 'Days to 30th Item: Same Day'
- 'Outfits'
- 'Outfits Others'
- 'Followers'
- 'Following'

Moderate-to-high negative correlation (feature weight <= -0.1) with:

- '1st Item Day-Of-Week: Never'
- 'Days to 1st Item: Never'
- 'Days 1st Outfit Others: Never'
- 'Days 30th Outfit Others: Never'
- 'Push notifications: Off'

Data Pre-processing: Secondary Latent Variable - PCA



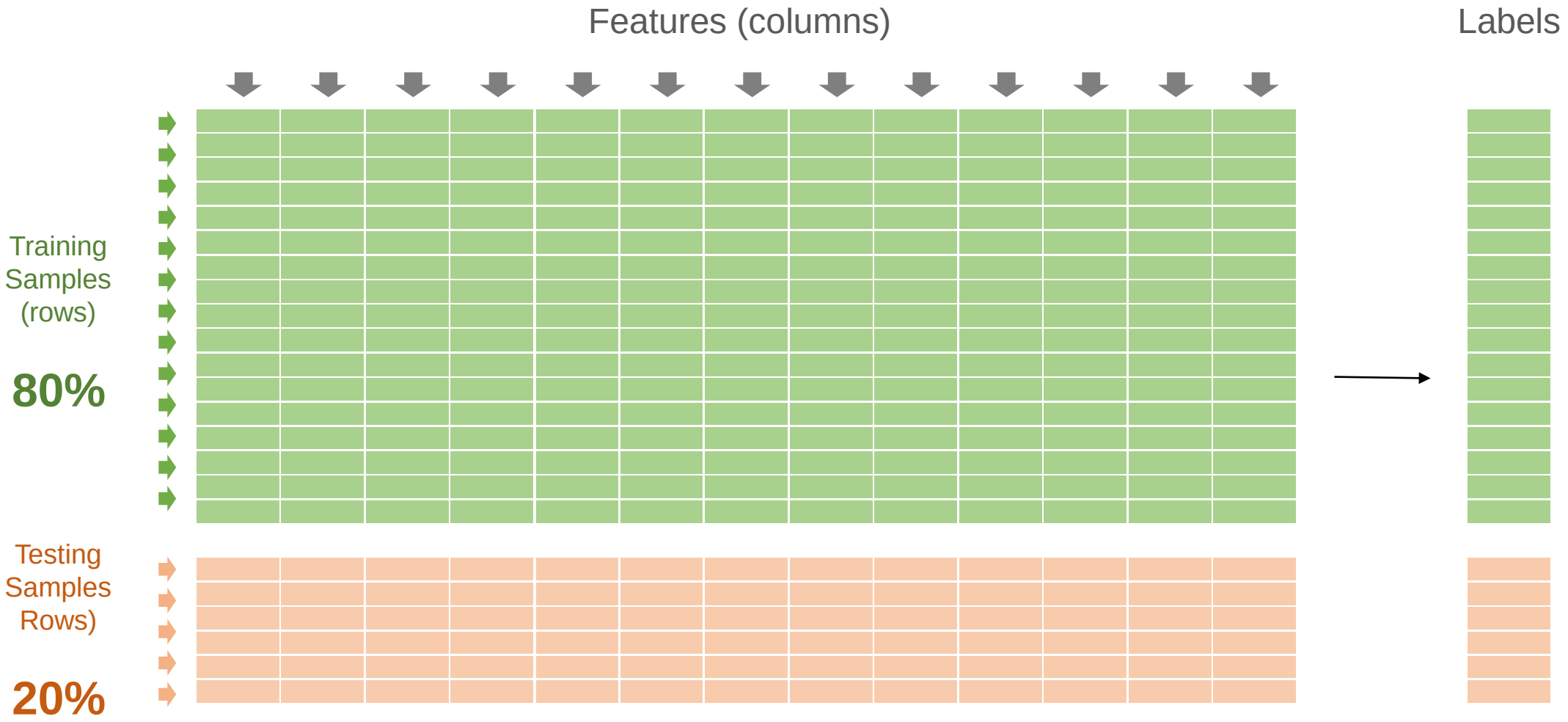
Moderate-to-high positive correlation (feature weight >= 0.1) with:

- 'Push Notifications: On'
- 'Sign Up Method: Facebook'

Moderate-to-high negative correlation (feature weight <= -0.1) with:

- 'Push Notifications: Off'
- 'Sign Up Method: Email'

Splitting the Data – 10,427 Samples



After going through the entire ML process…

Model Evaluation & Selection

Best performing model!

Primary model evaluation metric

	Naïve Predictor	SVM	Random Forest	AdaBoost	Naïve Bayes	MLP
F10-score	0.00	0.7566	0.4269	0.7377	0.7186	0.6244
Recall	0.00%	77.86%	42.95%	75.98%	74.46%	64.57%
Precision	100.00%	19.82%	26.82%	18.91%	16.02%	14.55%
Accuracy	94.62%	81.87%	90.63%	81.19%	0.78%	71.42%

For All The Technical Details Check Out The White Paper

Digital Product Strategy Powered by Machine Learning

Predicting User Retention from Early User Behavior

Jaime Pabon, @jpabonr

2017

I. Definition

Overview

In the context of the development of new software based products (e.g. web apps) the main objective is to achieve product market fit[1], which makes reference to the scenario where a product has long term adoption by a consumer segment in the market. Getting to product market fit typically requires continuous product iteration informed by implicit and explicit consumer feedback. The preferred metric to evaluate long-term product adoption is user retention[2], reason why it's closely monitored by product strategists/managers. Though user retention is influenced by a myriad of factors, most of these can be linked back to the quality of the user experience delivered by the product in terms of utility and usability.

A key aspect of the analysis involved in maximizing user retention is to identify the specific actions taken by a new user at the early stages of using the product that strongly correlate with (and possibly have a causal effect on) their long-term adoption. With this understanding the product development efforts can be aimed at delivering an optimized user experience that leads to a bigger proportion of new users engaging in the "right" actions (a proportion known as the user activation rate) that ultimately result in an uplift in user retention. The technique to effectively conduct such analysis can be greatly enhanced with the adoption of machine learning models. Such models would empower product managers,

[1] https://en.wikipedia.org/wiki/Product-market_fit

[2] http://www.commerce.com/blog/product-market-fit/

http://bit.ly/predict_retention

Next Steps: UX-related Experiments

1. A/B test Facebook vs email sign up
2. A/B test the tactics (copy, timing, frequency) in onboarding to prime and prompt users to enable push notifications
3. A/B test moments to prompt users to enable push notifications throughout the UX.
4. A/B test user onboarding process that includes creating an outfit for others
   a) Monitor impact of previous test on proportion of users who upload their closet

Final Reflection: Prod. Mgmt. is Increasingly Challenging

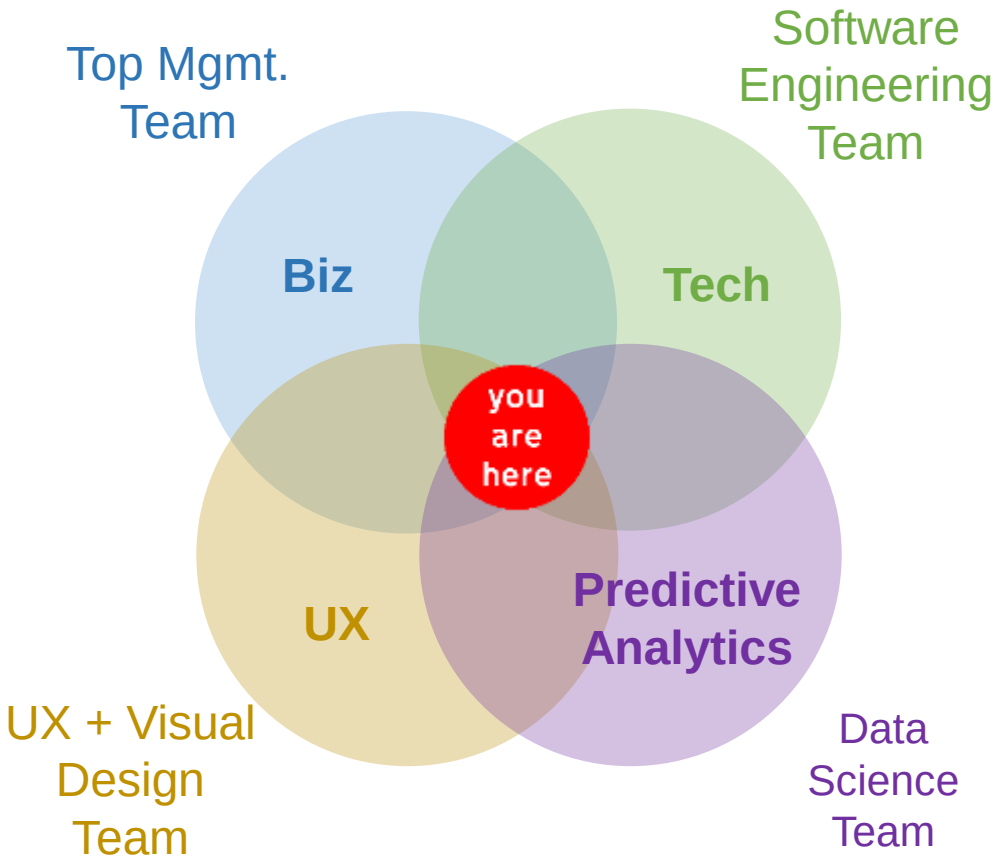


Image: Wired Magazine

> **“We’re still in the early phases of trying to figure out how on earth to organize the work of AI products.**

Andrew Ng

Venture capitalist, ex-Baidu chief scientist, Google Brain Team founder, Coursera co-founder, Stanford adjunct professor.

Jaime Pabon

jpabonr

Thank You!