Using Machine Learning To Increase Returns For Retail Investors

Prepared For:
Robinhood Executive Team

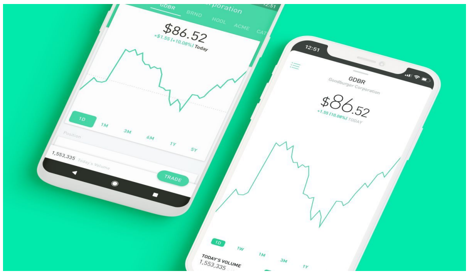
Kevin Fu, Minjie Xu, Ricardo Jenez
W201 Fall 2020 Group Pink
December 12th, 2020

[Kevin] Good evening Robinhood executive team,
My name is Kevin, and I’m here with my colleagues Ricardo and Minjie.

We’re here to talk to you tonight about using machine learning to increase returns for our Robinhood customers.
We’ve developed a machine learning algorithm that we think has the potential to help both the company as well as our customers, and we would love to present our proposal on how we can test this tool with real-time data.
Our goal is that after hearing our proposed **Research Design**, you will be interested in incrementally rolling out our tool on the Robinhood app.
Let’s get started. NEXT SLIDE

Retail Investors Gaining On The Well Heeled

Stock Brokers (1980s)

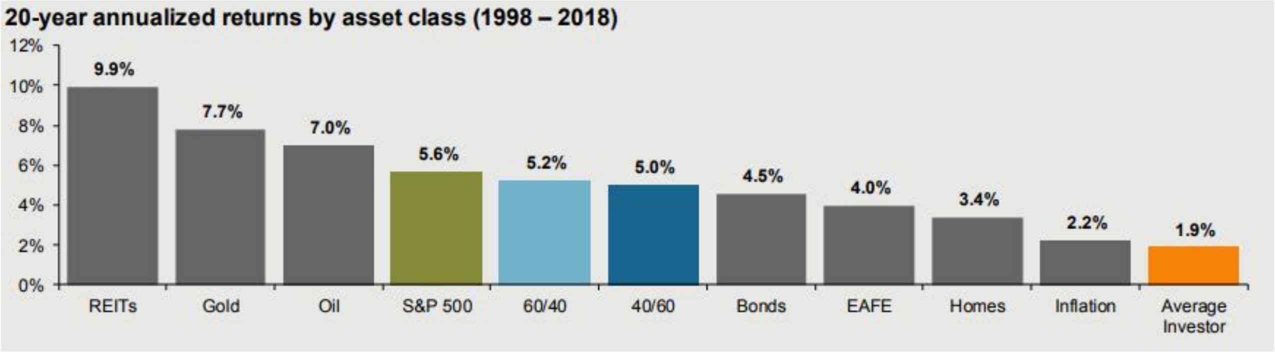


Robinhood (2013-present)

[Kevin]
Now than ever before: Retail investors HAVE more tools at their disposal when it comes to investing in Equities.

40 years ago, the retail investor had to **Read** the newspaper to just get stock quotes, call up their Broker, and pay an enormous commission - just to buy 100 shares of Apple.
Today, the retail investor can just pick up their smartphone, **inhale** a ton of quality research at their fingertips, and buy those 100 shares instantaneously - without **any** commissions.
NEXT SLIDE

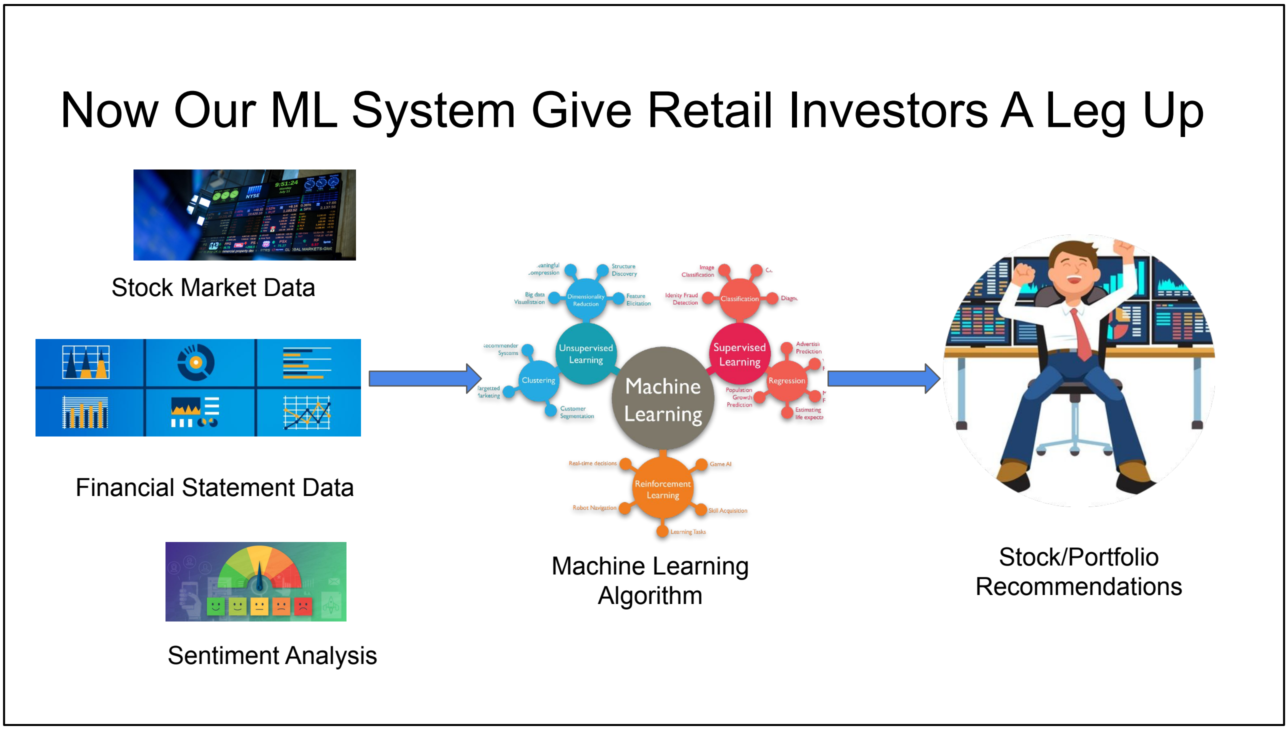
But Retail Investors Continue to Lag Behind...



Source: JP Morgan

[Kevin]
However, there is 1 **glaring** problem.

The average investor struggles to generate returns on a consistent basis, lagging behind **both** the broad market indices, as well as inflation.
As we can see here, it looks like the average investor makes around a 1.9% return per year! Over the past 20 years or so.
We have a solution for this problem - a solution that will benefit both Robinhood and our investors.
NEXT SLIDE



Now Our ML System Give Retail Investors A Leg Up

Stock Market Data

Financial Statement Data

Sentiment Analysis

Machine Learning Algorithm

Stock/Portfolio Recommendations

We believe that we can build a model based upon fundamental, technical and sentiment analysis that gives us an ability to deliver to our users the same experience as those that have paid investment advice. We intend to use liquid stocks to allow us to propose specific trading actions that each investor can easily implement. Stocktwits and CRSP will be our main source of data to the platform. We will use a sophisticated machine learning model that we will detail in this study that will allow us to provide an excess return vs. the SPY. We think giving our customers the benefit of an automated system for deep analysis will provide them better returns and increase their interest in our platform and will be an enticement for new customers on the platform.

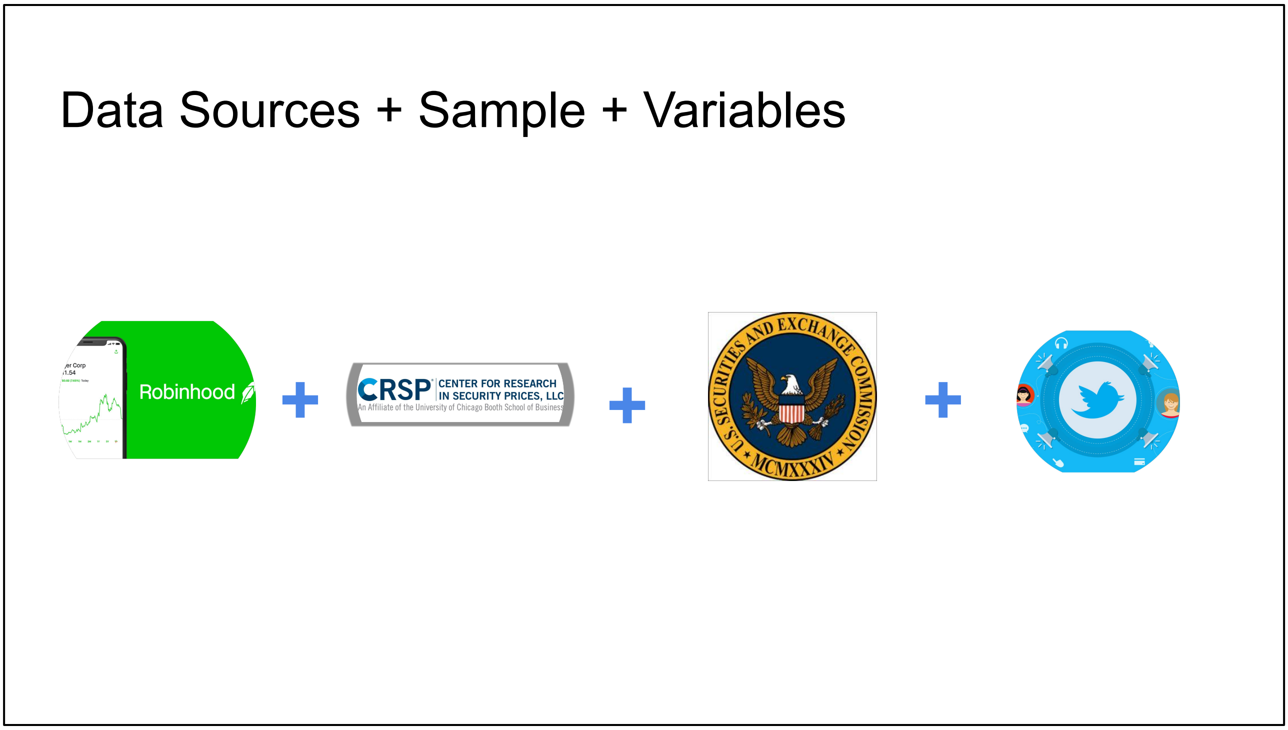
We expect to use already available sources of financial, stock and sentiment data:

- Data from the Robinhood trading platform.
  - We currently provide real-time market index data and stock ticker data[trades], financial news information,
  - From the real-time market data, our ML models will generate comprehensive technical indicators (e.g., price and volume chart/patterns, quant strategies) to train the portfolio.
  - Securities and Exchange Commission (SEC) reports (annual and quarterly financial statements, earnings reports, etc.) to our customers.

    - We will also look at our customer stock portfolios, current and past, to evaluate our tool's performance and help train our model in the later stages of this project.
- Current and historical data (stock trades, fundamental analysis) from the Center on Research of Stock Prices (CRSP/Compustat Merged Database (CCM)) to build our data model and create training and test data sets. This data source accounts for survivorship bias on returns.
- For real-time and historical stock and market sentiment analysis from Twitter and services like Stock Twits (https://stocktwits.com). Both offer excellent APIs for developers to gather specific information to train and test our system. These go back on over 10-12 years, but will become more important as we move forward.

Stock Market Data
https://www.researchgate.net/publication/336664897_Optical_Interconnection_Networks_for_High_Performance_Systems/figures?lo=1&utm_source=google&utm_medium=organic

https://www.dreamstime.com/stock-image-stock-investing-scale-decision-buy-sell-hold-image18323231
https://www.researchgate.net/publication/336664897_Optical_Interconnection_Networks_for_High_Performance_Systems/figures?lo=1



Data Sources + Sample + Variables

Robinhood
+
CRSP CENTER FOR RESEARCH IN SECURITY PRICES, LLC
An Affiliate of the University of Chicago Booth School of Business
+
U.S. SECURITIES AND EXCHANGE COMMISSION
MCMXXXIV
+

Data: **Robinhood(Portfolio), SEC (CRSP) (Fundamental), Exchange Data (Technical), StockTwits/Twitter(Sentiment)**

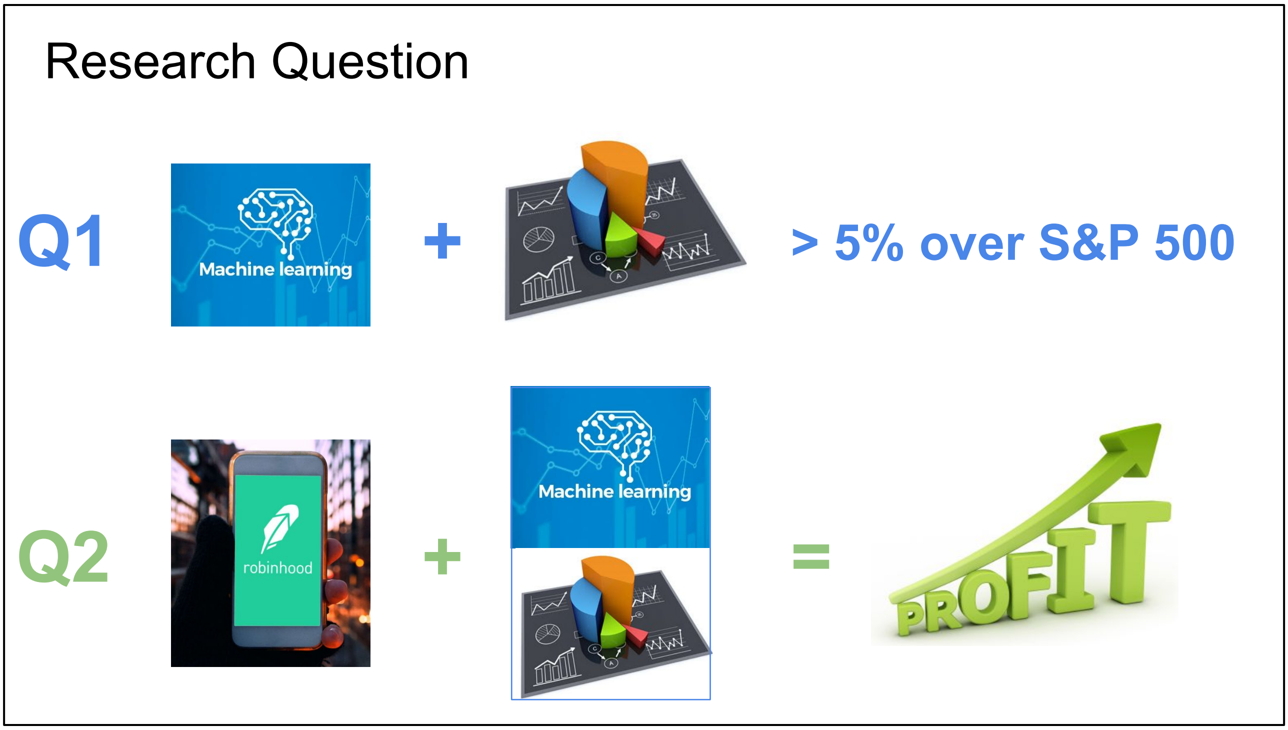
We are bringing new date onto our platform that includes CRSP, SEC and Twitter date. Good news is that we have real time trading data coming into our Robinhood platform already.

The historical data we are bringing in includes fundamental data that includes information about growth and value (Return on Assets, and other fundamental measures) and fundamental company valuation from CRSP. SEC data will also bring in fundamental value and risk information. Stocktwits and twitter will provide sentiment data in terms of how the overall market feels about individual stocks and the market in general. We are able to bring fundamental, technical and sentiment analysis together in one data set.

Data Bias. Need to be careful of survivor bias in the data (accounted for in CRSP) and need to be sure that we introduce information with appropriate timing as when information would be delivered in the real world.

Sample size. Out of the universe of 6100 or so stocks listed on the NYSE and Nasdaq, we are selecting the 1000 with ample liquidity (> $20-80M of trade volume) to put in our training data set for our ML model.

- The ML will then look at different stock portfolio sizes (5, 10, 20 stocks - small, medium, large) at any given time that it and a trading pattern of (< 5, < 20, <100 trades per year) believes will generate the most alpha for the next 12 months. With these different size considerations we can better match our customers propensity to different size portfolios and different trading patterns.
- Roughly 60% of robinhood customers hold mid-cap ($2B+) or larger companies. Therefore, in a further effort to manage risk, we can filter out small cap companies, leaving the larger, more established companies within our ML algorithm’s universe.
- This portfolio must be easily purchasable by Robinhood’s retail customers.



Research Question

Q1 + > 5% over S&P 500

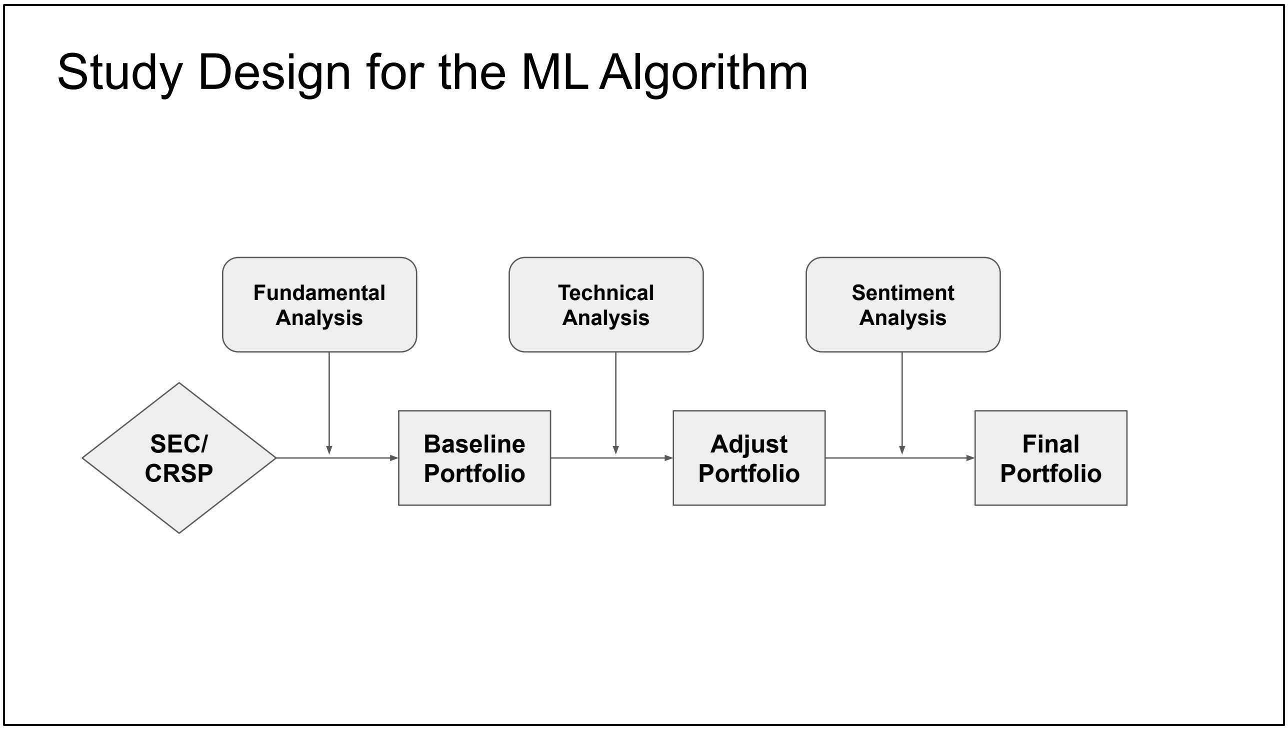
Q2 + =

Research Question: Can a ML-backed stock portfolio generate outperformance versus the S&P 500 index over the long-term?

Hypothesis: Based on a hybrid of fundamental, technical, and sentiment analysis, the ML portfolio will outperform the S&P 500 index (9.8% average annualized return) by more than 5% annually.

Sub-question: Can the ML-backed portfolio provide effective advisory to Robinhood users in terms of affecting their trading performances and returns?

Hypothesis: On average, Robinhood users who used our service will outperform their past performances and a control group.



Study Design for the ML Algorithm

Fundamental Analysis

Technical Analysis

Sentiment Analysis

SEC/ CRSP

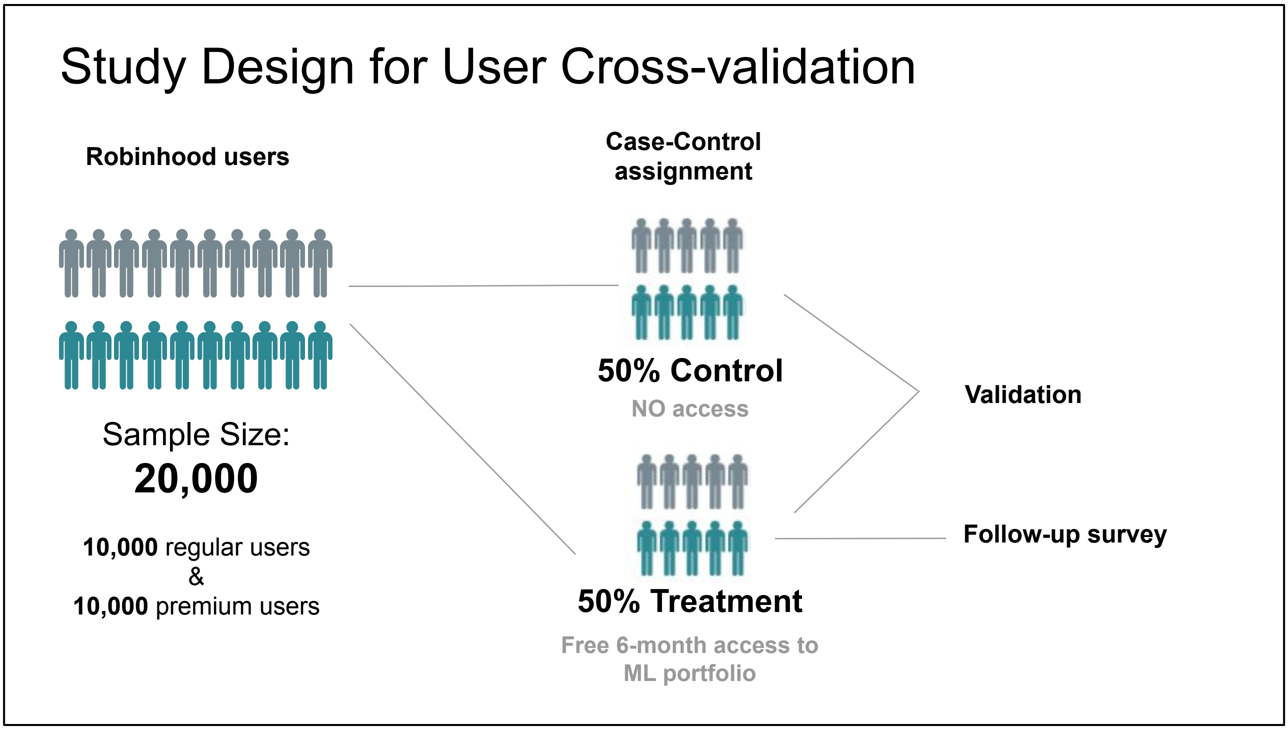
Baseline Portfolio

Adjust Portfolio

Final Portfolio

To achieve 5% outperformance per annum vs. the S&P 500

- Initially, we will use fundamental analysis to create a baseline portfolio that contains 20 stocks in equal weight.
- Then we bring technical analysis to further improve the algorithm. During the step, our algorithm will auto adjust the shares of the portfolio to avoid major sell-off or market crash while buying dip/rebound.
- We will use historical time-series data sets to train the models. We will gather the past 20 years of stock market data to backtest the hypotheses for retrospective validation. This approach will allow us to validate that we can provide 5% more above the historical S&P 500 based upon historical market trends and price activity.
- Last, we will incorporate real-time market data and sentiment analysis to further improve the adjusted portfolio to see if it can beat the S&P 500 by at least 5% annualized return over six-month, while predicting future market trend and portfolio performance.



Study Design for User Cross-validation

Robinhood users

Sample Size:
20,000

10,000 regular users
&
10,000 premium users

Case-Control assignment

50% Control

NO access

50% Treatment

Free 6-month access to ML portfolio

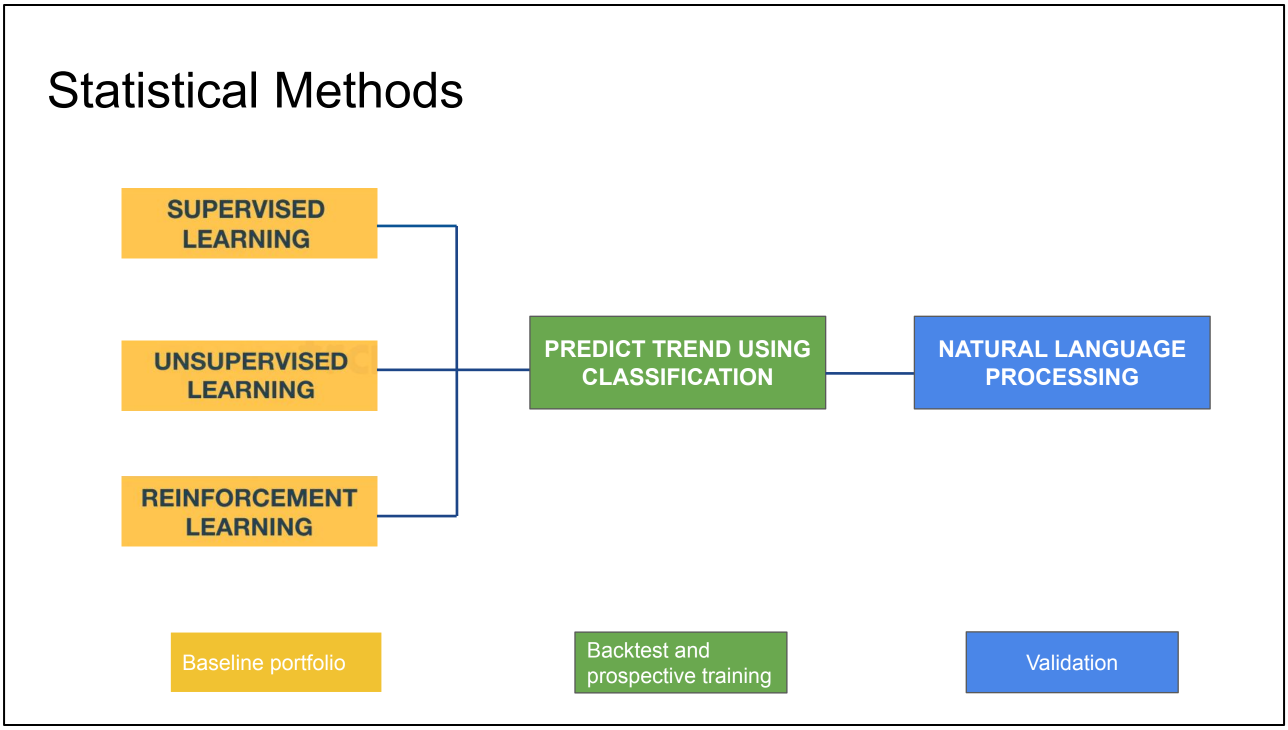
Validation

Follow-up survey

To test the effectiveness and validity of the ML-backed portfolio with Robinhood users, we propose a quasi-experiment study by inviting Robinhood users to beta test our ML-backed service from a pool of 20,000 Robinhood users. We will randomly select 10,000 regular users and 10,000 premium users (who paid $5 monthly subscription fee), the first 5000 regular users and 5000 premium users will get free access to our portfolio and service for 6-month, and become our treatment group, while the rest of 10,000 users will be the control group.

After 6 months, we will send a short follow-up survey to the treatment group to ask 1) their trading behavior and/or performance changes, 2) their expected and actual returns with/without our service, 3) their willingness to use and/or pay for our service in future, and 4) other feedback if any.

Furthermore, all 20,000 users’ trading behavior and performance will be analyzed and compared between different groups: 1) before vs. after, 2) treatment vs. control, and 3) regular vs. premium users.



Statistical Methods

SUPERVISED LEARNING

UNSUPERVISED LEARNING

REINFORCEMENT LEARNING

PREDICT TREND USING CLASSIFICATION

NATURAL LANGUAGE PROCESSING

Baseline portfolio

Backtest and prospective training

Validation

We will apply a batch of mainstream machine learning models used in trading to train our ML-backed system. In the first block, we use supervised learning models, unsupervised learning models and reinforcement learning models to create the baseline profolio. Then we use classification to backtest our strategies and do prospective training on our model. Last, we will also use natural language processing for sentiment analysis to adjust the portfolio.

Potential Risks

Bank of America
Goldman Sachs
citibank
Morgan Stanley
WELLS FARGO
JPMorganChase

Hindsight Bias

Competition

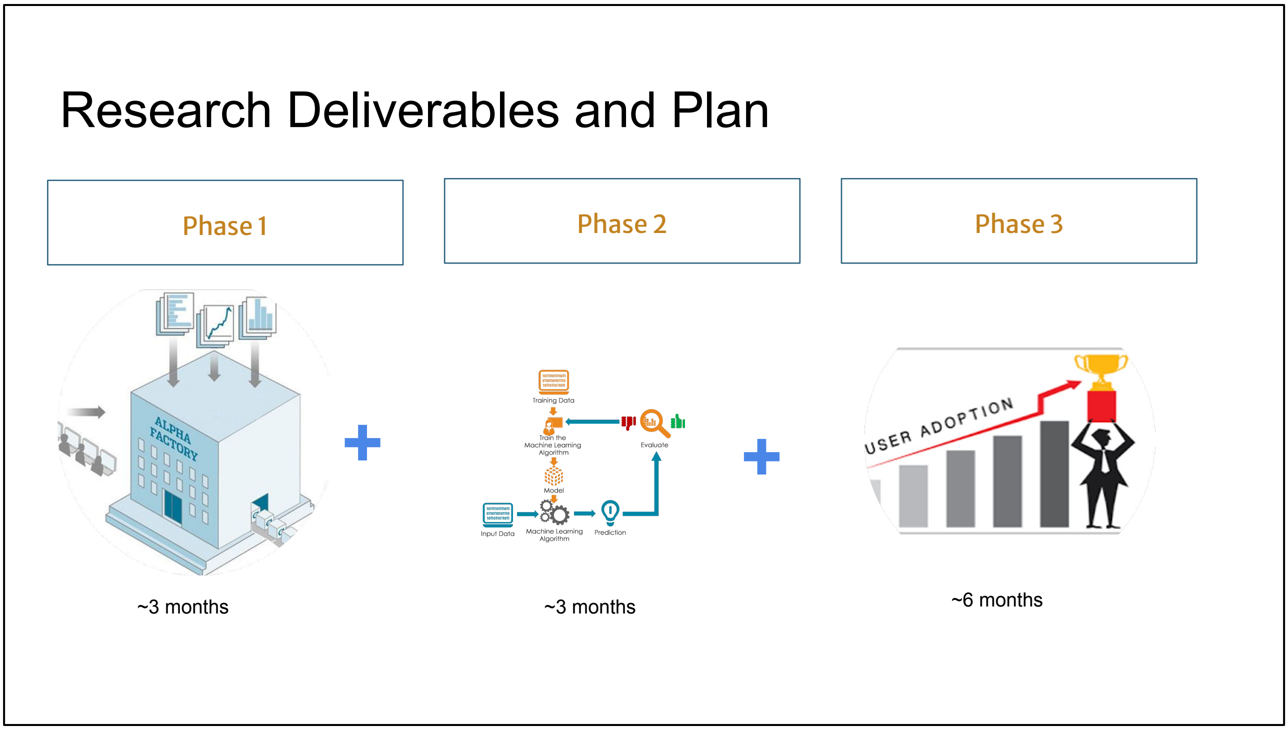
[Kevin]
So, Of course, we have to consider any potential risks.
Hindsight Bias is a risk. We have to rely on analyzing past data to predict future data points.

- **Our response is:** We can use a **Monte Carlo scenario analysis**, in which we artificially create scenarios given a **limited data set**. For example, we can pretend that its 2008, and only feed our algorithm historical prices and SEC filings from 2008 and before. We will withhold all data from 2009 and onwards, and see how our algorithm performs. This process will be repeatable for any given time period, so that this way, our algorithm can learn from an infinite number of hypothetical scenarios.

Another risk is **Heightened competition**. Many of our competitors (Morgan Stanley, Goldman Sachs) have substantially greater financial resources than we do.

- **Our response to that is:** Data is king. We have a first-mover advantage in terms of providing a ML-backed solution for the everyday, retail investor. This means that if we implement this today, our algorithm will have a longer track record and more experience in terms of working with real-time data, as opposed to our newer competitors who enter the market later.

NEXT SLIDE



Research Deliverables and Plan

Phase 1

Phase 2

Phase 3

~3 months

~3 months

~6 months

http://www.christiernan.com/how-to-measure-your-users-salesforce-adoption-rate/

We are excited about bringing this project to you and will transformative for Robinhood. BUT it is important for us to tell you how long it is going to take. The first 3 months will be bringing the data in from external sources. Good news is that we already have live trading and a data pipeline serving our customers today. We will only need add fundamental historical data from CRSP and SEC, and flow in sentiment through APIs from Stocktwits an twitter. We already have a robust data infrastructure and tools so we think we can get this done in a fairly quick time period.
The next phase is that we will get our training and test sets together and start implementing our machine learning pipeline. Good news again, we already have a good ML system in place from other projects so we can use these tools to build and run our models. Again turn around will be 3 months.

We will spend six months validating to be sure we hit our target return with data and some look at customer portfolios. We think we have a good shot at doing this in this time period.
At the end of twelve months, we will have a prototype system, a full report on what we found and a plan for moving this system into production and a financial plan as to what the economic returns will be.. This is really exciting and we think relatively low

cost for us to be able to bring a huge potential value to the company.

- Phase 1: Data gathering/collection/cleansing - 3 months
    a. From Twitter/Stocktwits/CRSP/SEC.gov/any public data source(s)
    b. Develop model for the training sets and test sets from the data
- Phase 2: Experimental phase/testing - 3 months
    a. Program our ML algorithm and train our models on our training sets and validate on our test sets
    b. Run models on a unrestricted set of historical data to see if it does better than the market (backtests)
- Phase 3: Model validation - 6 months
    a. If we do not reach the hurdle of outperformance by 5% per annum, repeat Phase 1-2
    b. We can annualize our returns if there are time constraints
- Phase 4: Present our findings to the product team and senior leadership at Robinhood
    a. Write up the final report with insights and recommendations.
    b. Document lessons learned
    c. Project plan and budget for next project to get to a production system
    d. Financial outline of the expected benefits to RobinHood and our customers

Conclusion

[Kevin]
In conclusion, we think that we have a really neat idea on our hands.
Future studies might include analyzing if our algorithm can **profitably** Alter its strategy given different risk parameters.
It would be very useful if our algorithm can be customized **based on the different risk appetites of individual investors**, as opposed to the one-size-fits-all strategy we're initially working on.

Thank you very much for your time today. We look forward to working with you on this exciting, new business opportunity!

Bibliography

Articles:

- https://www.roche.com/research_and_development/what_we_are_working_on/oncology/cancer-immunotherapy/measuring-patient-outcome.htm

Images:

- https://aquinahealth.com/wp-content/uploads/2018/07/Pulse_71118.jpeg
- https://advantagecaredtc.org/wp-content/uploads/2018/08/doctor-patient-relationship_advantage-care-health-centers.png
- https://cdn.jotfor.ms/form-templates/screenshots/legacy/325x400_202091509530851/new-patient-enrollment-form.png
- https://thumbs.dreamstime.com/b/doctor-cartoon-yes-no-signs-info-graphic-design-108383874.jpg
- https://app.biorender.com/
- https://www.google.com/finance
- http://www.stockspin.com/building-good-portfolio/
- https://www.thekickassentrepreneur.com/profit-average-small-business/
- https://www.dnaindia.com/personal-finance/report-how-retail-investors-can-ride-the-stock-market-2570786
- https://medium.com/@tomyuz/a-sentiment-analysis-approach-to-predicting-stock-returns-d5ca8b75a42