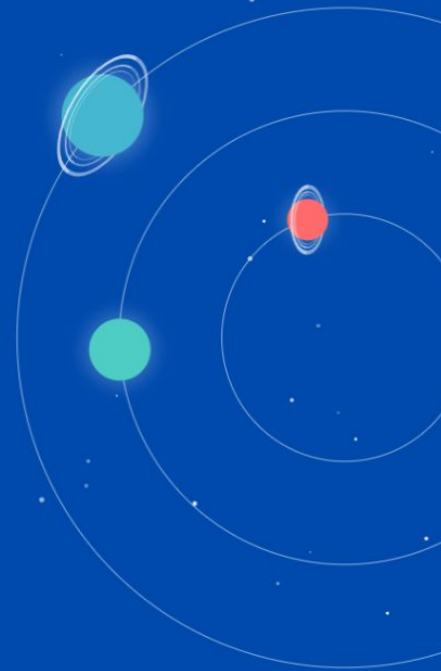


# Modeling Stock Returns with Emojis

Roman Paolucci  
Founder, Quant Guild



QUANT GUILD



# ★ About Me

## ***Professional Experience***

- ★ Quantitative Researcher/Trader
- ★ Data Scientist
- ★ Software Engineer
- ★ Technical Writer
- ★ Founder, Quant Guild

## ***Education***

- ★ *A lot of Math*

## ***Hobbies & Interests***

- ★ Brazilian Jiu-Jitsu
- ★ Bodybuilding
- ★ Golf
- ★ Fishing



# ★ Navigating the Quant Space

After assessing your compatibility for the role (**NOT** a trivial task!)  
Position yourself based on a combination of your interests and natural skillset

★ **I was always a better empiricist but had a preference for theory**

## My Two-Pronged Approach to Positioning Myself as a Quant

### Quant Research/Trading

- *A ton of empirical work*
- *Some theory in asset pricing for strat analysis*
- *Equity/Equity Derivatives Focus*

### Derivatives Pricing

- *A ton of theory (:) )*
- *Specifically Options and Extrapolating Exotic Prices*
- *Space for ML/AI in Current Literature*

I was particularly interested in roles in QR/QT, then it came time to build and enhance the necessary skills and engage in projects in the space

# ★ Navigating the Quant Space

After assessing your compatibility for the role (**NOT** a trivial task!)  
Position yourself based on a combination of your interests and natural skillset

★ **I was always a better empiricist but had a preference for theory**

## My Two-Pronged Approach to Positioning Myself as a Quant

### Quant Research/Trading

- *A ton of empirical work*
- *Some theory in asset pricing for strat analysis*
- *Equity/Equity Derivatives Focus*

### Derivatives Pricing

- *A ton of theory (:) )*
- *Specifically Options and Extrapolating Exotic Prices*
- *Space for ML/AI in Current Literature*

**This talk focuses on a project in this space from an academic perspective, research in an academic and industrial settings are somewhat equivalent but their outcomes vary (roughly open-source publication or for-profit implementation)**

# ★ Academic vs. Industrial Research

We'll focus on this idea of **quant trading** by observing my academic paper

## Academic Research

- ★ *\*Difficult\** to get data
- ★ PhD necessary for tenured positions
- ★ Typically open source
- ★ Some industrial collaboration
- ★ Publish or perish culture

## Industrial or Private Research

- ★ *\*Easier\** to get data
- ★ PhD **NOT** necessary
- ★ Not usually open source  
(*sometimes with academic collaboration*)

### How Many Words is a Picture Worth? Using Emojis from Social Media to Predict Future Stock Returns

57 Pages • Posted: 27 Apr 2023 • Last revised: 21 Jan 2024

Date Written: December 11, 2023

#### Abstract

Using a new and comprehensive sample of more than 87 million Twitter posts referencing Russell 3000 firms between 2012 and 2022, we introduce a novel, unsupervised method of scoring the sentiment of emojis. Our method generates point-in-time dictionaries that map individual emojis to the contextual sentiment of recent tweets that contain them. In out-of-sample tests, we find that even controlling for the sentiment extracted from words, news, and corporate events, emoji sentiment correctly predicts future firm-level stock returns. Importantly, we show a newly emergent generation of Twitter users drive emoji-based return predictability, while more experienced users better predict returns using words. Understanding the sentiment of emojis has become increasingly important as individuals and market professionals continue to adopt these new forms of communication.

**Keywords:** Twitter, emojis, social media sentiment

**Remark/Opinion:** Let me save you some time, **DO NOT** get a Phd. because you want to get into QR, many Bachelors/MS holders go into QR. . .pursue a PhD because you have a passion for research (*passion does not equal "wants lots of money"*) - those 5+ years will otherwise be quite painful!

# ★ Academic vs. Industrial Research

We'll focus on this idea of **quant trading** by observing my academic paper

## Academic Research

- ★ *\*Difficult\** to get data
- ★ PhD necessary for tenured positions
- ★ Typically open source
- ★ Some industrial collaboration
- ★ Publish or perish culture

**We will be focusing on academic research in the field of finance – there really is no such notion of qualitative research when it comes to finance research, this is the so called “quant research”, there are varying degrees of “quant” but I digress**

### How Many Words is a Picture Worth? Using Emojis from Social Media to Predict Future Stock Returns

57 Pages • Posted: 27 Apr 2023 • Last revised: 21 Jan 2024

Date Written: December 11, 2023

#### Abstract

Using a new and comprehensive sample of more than 87 million Twitter posts referencing Russell 3000 firms between 2012 and 2022, we introduce a novel, unsupervised method of scoring the sentiment of emojis. Our method generates point-in-time dictionaries that map individual emojis to the contextual sentiment of recent tweets that contain them. In out-of-sample tests, we find that even controlling for the sentiment extracted from words, news, and corporate events, emoji sentiment correctly predicts future firm-level stock returns. Importantly, we show a newly emergent generation of Twitter users drive emoji-based return predictability, while more experienced users better predict returns using words. Understanding the sentiment of emojis has become increasingly important as individuals and market professionals continue to adopt these new forms of communication.

**Keywords:** Twitter, emojis, social media sentiment

# ★ Academic Research Questions About Emojis 🐮🐻

After working on social feeds and observing new meanings of emojis (i.e. "TO THE MOON 🚀🌙") I had several questions I wanted to answer. . .

## ★ Does the meaning of words change over time?

- *Not a novel idea, language velocity is well studied*

## ★ Is there a way to capture the meaning at a specific point in time?

- *Not necessarily novel, but certainly relevant for backtesting*

## ★ Can we apply this methodology to emojis?

- *Several models had done this, but not in a time varying capacity*

## ★ Do emojis offer unique information in cross-sectional return "prediction" ?

- *Now we're getting somewhere in terms of novel, time-variant emoji distributions to model cross-sectional returns*

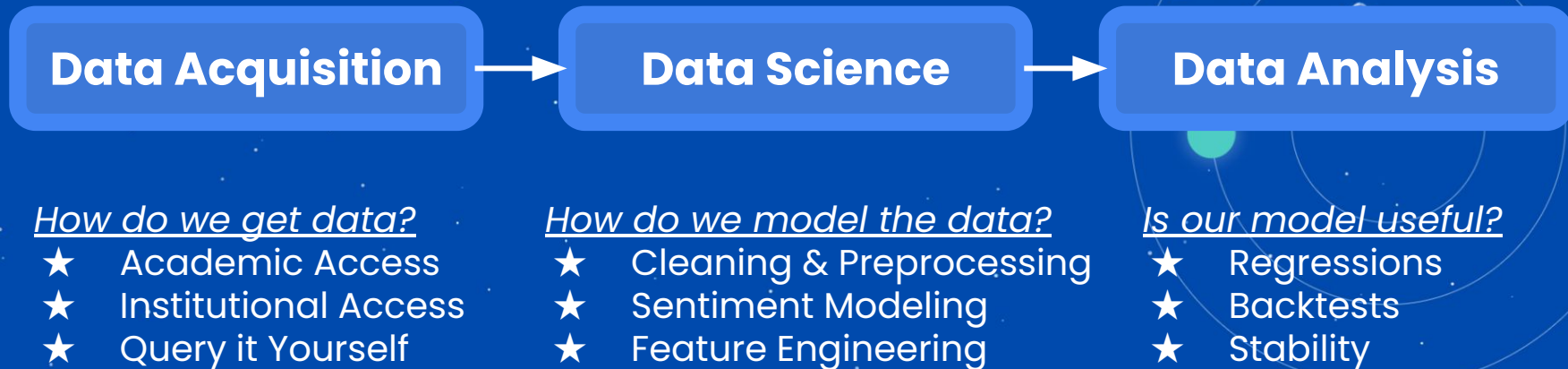




# ★ Process Map of Academic Research

I've always been better suited to empirical research from my background as an engineer – I can develop data pipelines and implement APIs quite quickly...

**Effectively, there are 3 stages to this research process**





# ★ Process Map of Academic Research

I've always been better suited to empirical research from my background as an engineer – I can develop data pipelines and implement APIs quite quickly...

**Effectively, there are 3 stages to quant research process**

## Data Acquisition

How do we get data?

- ★ Academic Access
- ★ Institutional Access
- ★ Query it Yourself

## Data Science

How do we model the data?

- ★ Cleaning & Preprocessing
- ★ Sentiment Modeling
- ★ Feature Engineering

## Data Analysis

Is our model useful?

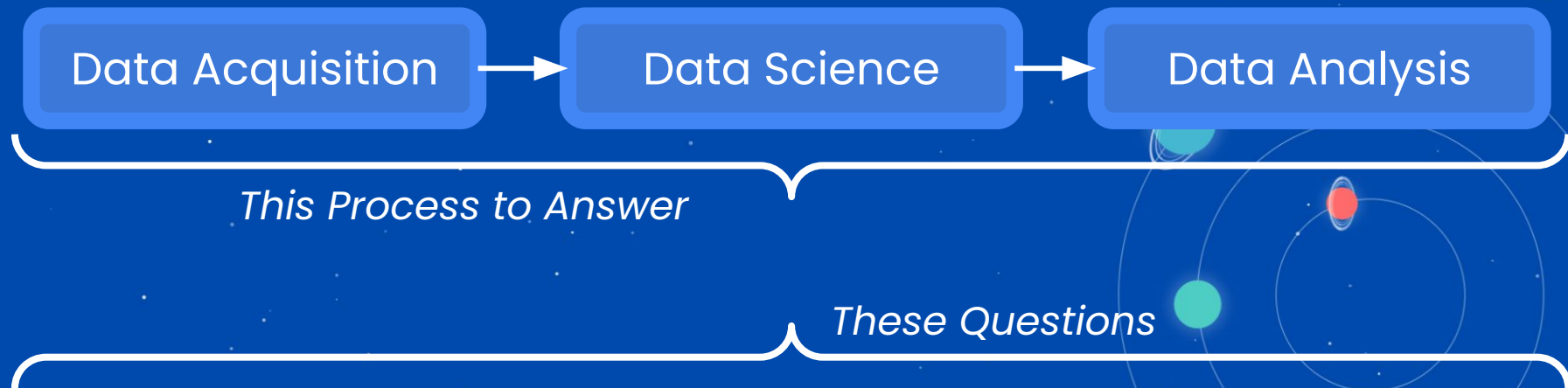
- ★ Regressions
- ★ Backtests
- ★ Stability

**95%+ of the time is spent here, it is the most important part. You must enjoy data work!**

**The fun part!**



# Let's Walk Through This Research Process



- ★ *Does the meaning of words change over time?*
- ★ *Is there a way to capture the meaning at a specific point in time?*
- ★ *Can we apply this methodology to emojis?*
- ★ *Do emojis offer unique information in cross-sectional return “prediction” ?*

# ★ Research Process: Data Acquisition

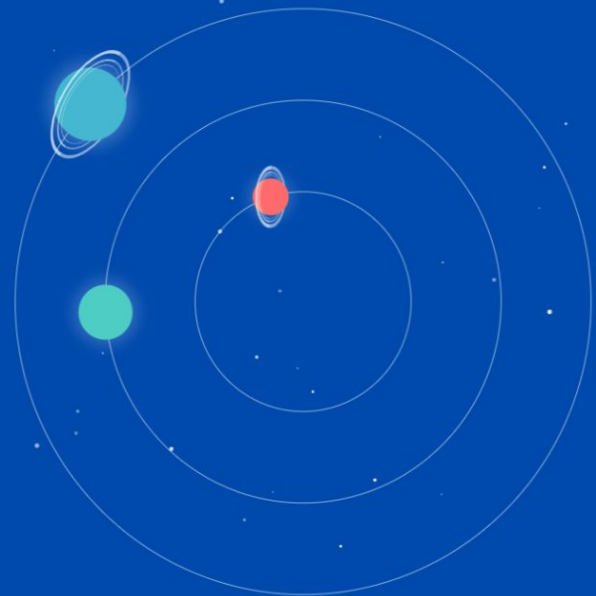
**The data we need is a function of our research questions**

- ★ *Does the meaning of words change over time?*
- ★ *Is there a way to capture the meaning at a specific point in time?*
- ★ *Can we apply this methodology to emojis?*
- ★ *Do emojis offer unique information in cross-sectional return “prediction” ?*

## **Data Required**

- ★ *Equity Universe (SPX, R3000, ...)*
- ★ *Pricing Data (Returns)*
- ★ *Social Text Data (Tweets)*

Okay, so where do we get it??? This depends on the scope of your project.



# ★ Research Process: Data Acquisition

The data we need is a function of our research questions

- ★ Does the meaning of words change over time?
- ★ Is there a way to capture the meaning at a specific point in time?
- ★ Can we apply this methodology to emojis?
- ★ Do emojis offer unique information in cross-sectional return “prediction” ?

## Data Required

- ★ ~~Equity Universe (SPX, **R3000**, ...)~~
- ★ ~~Pricing Data (**Returns**)~~
- ★ **Social Text Data (Tweets) ??????????**

Our scope was **academic**...

Academic Access

Institutional Access

Query it Yourself



# ★ Research Process: Data Acquisition

The data we need is a function of our research questions

- ★ Does the meaning of words change over time?
- ★ Is there a way to capture the meaning at a specific point in time?
- ★ Can we apply this methodology to emojis?
- ★ Do emojis offer unique information in cross-sectional return "prediction" ?

## Data Required

★ ~~Equity Universe (SPX, **R3000**, ...)~~

★ ~~Pricing Data (**Returns**)~~

★ **Social Text Data (Tweets)**

*Time to pull up the ol' bootstraps...*

Way too \$\$\$

Academic Access

Institutional Access

Query it Yourself

Doesn't Exist

# ★ Research Process: Data Acquisition

The data we need is a function of our research questions

- ★ Does the meaning of words change over time?
- ★ Is there a way to capture the meaning at a specific point in time?
- ★ Can we apply this methodology to emojis?
- ★ Do emojis offer unique information in cross-sectional return “prediction” ?

## Data Required

★ ~~Equity Universe (SPX, **R3000**, ...)~~

★ ~~Pricing Data (**Returns**)~~

★ **Social Text Data (Tweets)**

😬 Time to go to work...



Query it Yourself



# Research Process: Data Acquisition

1. First, we need our equity universe (easily accessible with academic access)
2. Then we query data based on documents (tweets) containing tickers (\$AAPL)
3. This took quite some time to query - we ended up with over 87 million tweets!

## Data Acquisition Pipeline (Built from Scratch)

Equity Universe

Tickers in R3000  
for Period Rebalanced  
Annually (This is Changing)

Firms July - July 20XX



twitter 

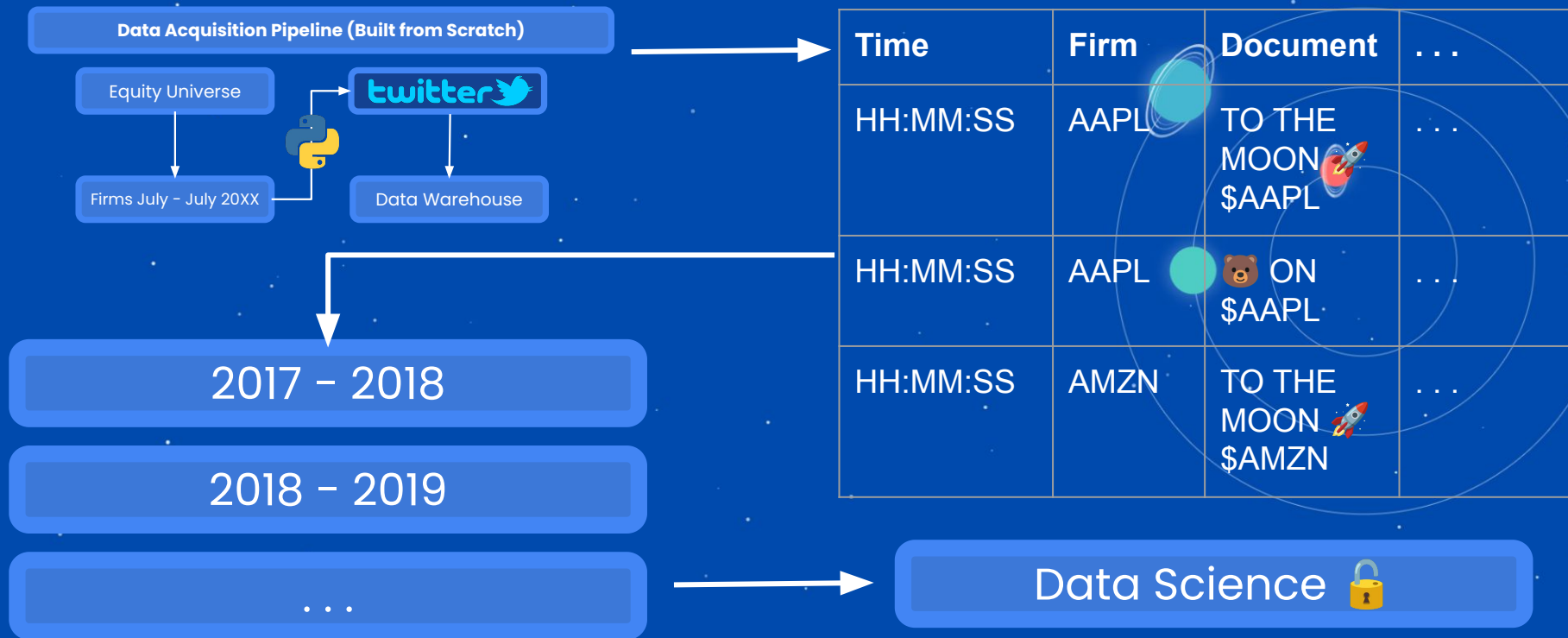
Twitter's Academic API  
- All tweets for the period  
by \$ (e.g. \$AAPL)

Data Warehouse



# ★ Research Process: Data Acquisition

It's always easier to speak in terms of a database...



# ★ Research Process: Data Science

## Ah yes, the data science step...

*How do we manage stop words, capitalization, punctuation?*

*How do we handle foreign languages?*

*How do we handle an imbalance in data across firms?*

*How do we handle outliers in any capacity?*

*How do we extract signal (sentiment) from text?*

*How do we handle multiple measures of sentiment?*

*How do we handle documents with multiple firms?*

*How do we handle firms without documents?*

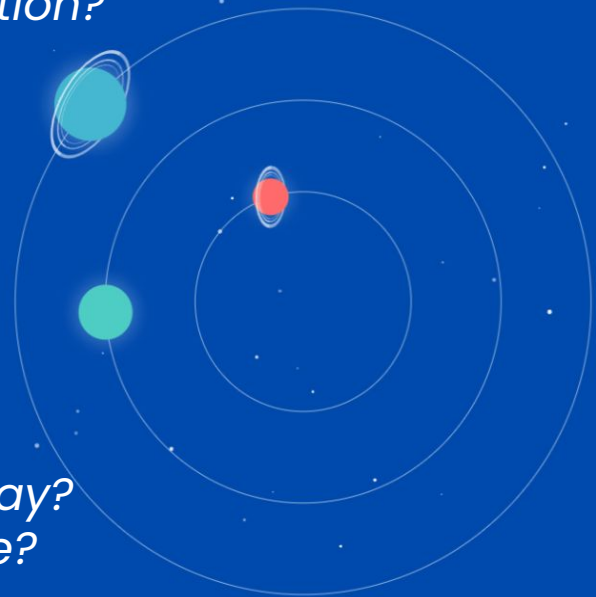
*How do we handle documents without sentiment?*

*How do we handle high variance sentiment on a firm-day?*

*How do we determine if our sentiment model is effective?*

....

We could spend several hours discussing our methodology – let's focus on one of the primary contributions: **point-in-time sentiment modeling**



# ★ Research Process: Data Science

## Words in a social context (especially online) change over time...

There are many approaches to sentiment modeling (essentially scoring text) Dictionaries, ML, LLMs. . .  
How do we assign meaningful scores to text without imposing our own bias?  
For example, how should our sentiment model score these emojis?



Tweet: "\$AAPL 📈"

Model

Score: .75 (positive)

We need to build this to extract a score for analysis (regressions, backtests, etc. . .)

We want the scores to make sense intuitively, but also account for variation over time. . .

# ★ Research Process: Data Science

To develop a measure of point-in-time sentiment we...

Time	Firm	Document	...
HH:MM:SS	AAPL	TO THE MOON 🚀 \$AAPL	...
HH:MM:SS	AAPL	🐻 ON \$AAPL	...
HH:MM:SS	AMZN	TO THE MOON 🚀 \$AMZN	...

- ★ Applied a **general** sentiment model
  - Something we consider to be stable over time (time invariant)
- ★ Took the average sentiment for the document containing the word of interest
- ★ By the CLT this distribution would be normal at that **point-in-time!**
  - Extremely useful for testing, confidence intervals, and probabilities even though the distribution changes over time!

Effectively, we are trying to develop an *unsupervised* way to assign a score to words or tokens like emojis so we can account for changes over time...

# ★ Research Process: Data Science

## Example: Point-In-Time Sentiment Model

### Tweet Data

Lots of Tweets in 2000 with 🚀

“BILLIONAIRES GOING TO THE MOON? 🚀”

...

Lots of Tweets in 2018 with 🚀

“\$AAPL IS GREAT TO THE MOON 🚀”

...

### Point-In-Time Sentiment Model

General Sentiment Model

Pos: 0.0, Neg: 0.0

...

General Sentiment Model

Pos: 1.0, Neg: 0.0

...

2000 Average 🚀

Pos: 0.0, Neg: 0.0

2018 Average 🚀



Pos: 1.0, Neg: 0.0

Effectively, applying a **general** sentiment model extracts common words like good, bad, great, awful, etc. . . to determine the overall tone *at that point in time* as those words are likely more time invariant than other words or *tokens* like emojis - we can see in this example this model effectively captures the changing meaning of the rocket emoji!



# ★ Research Process: Data Science

## What we find in the data applying this methodology to emojis...

Anecdotal examples of capturing time variant dynamics

 - Positive  
 - Negative



 - Positive  
 - Negative

 - Positive  
 - Positive

 - Positive  
 - Negative



After accounting for other data science steps... we now have sufficient data for our analysis phase!

Time	Firm	Document	Score
HH:MM:SS	AAPL	TO THE MOON  \$AAPL	.75
HH:MM:SS	AAPL	 ON \$AAPL	-.5

These make sense intuitively in a social sense, and we didn't have to impose that they were positive/negative explicitly! Cool!



# ★ Research Process: Data Science

## What we find in the data applying this methodology to emojis...

Anecdotal examples of capturing time variant dynamics

 - Positive  
 - Negative

 - Positive  
 - Negative

 - Positive  
 - Positive

 - Positive  
 - Negative

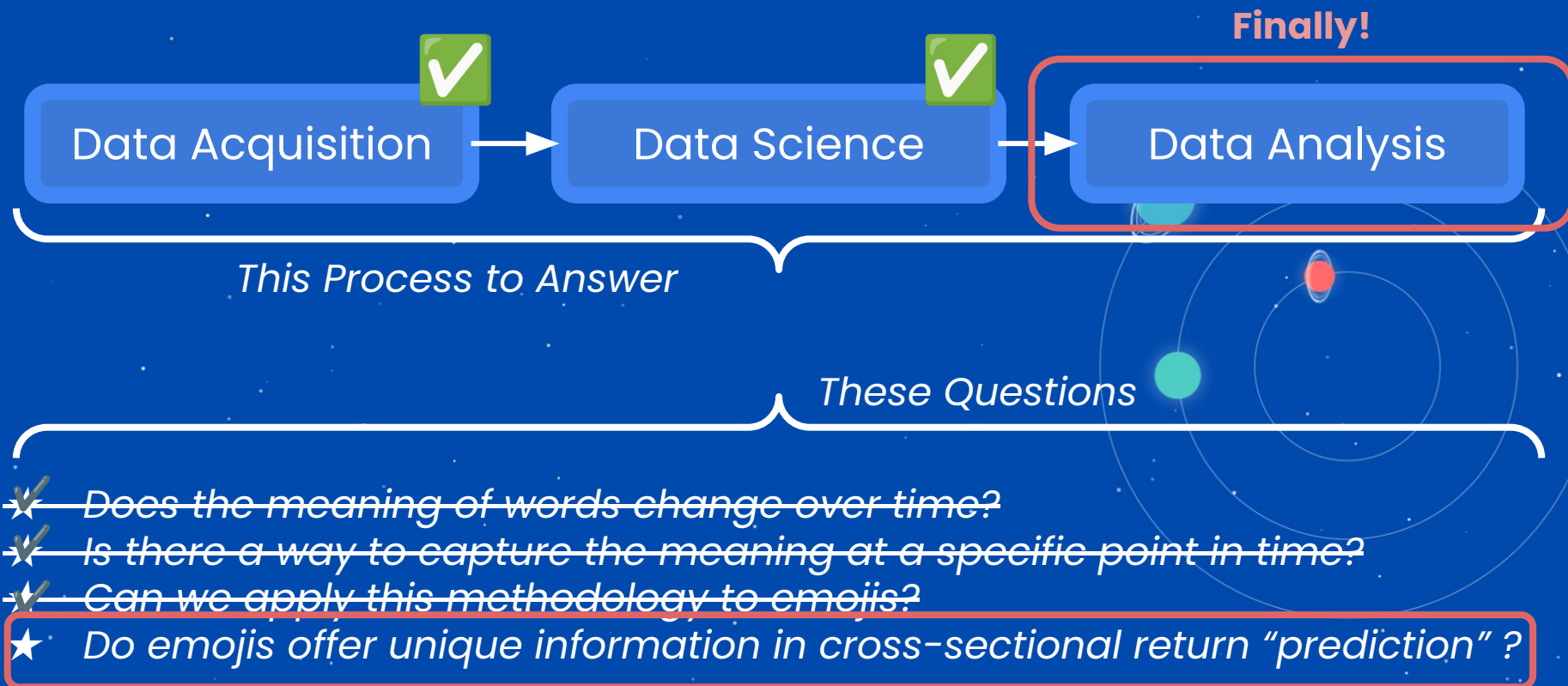
**So...Is It Useful?**

Data Analysis 

These make sense intuitively in a social sense, and we didn't have to impose that they were positive/negative explicitly! Cool!



# ★ Research Process: Data Analysis



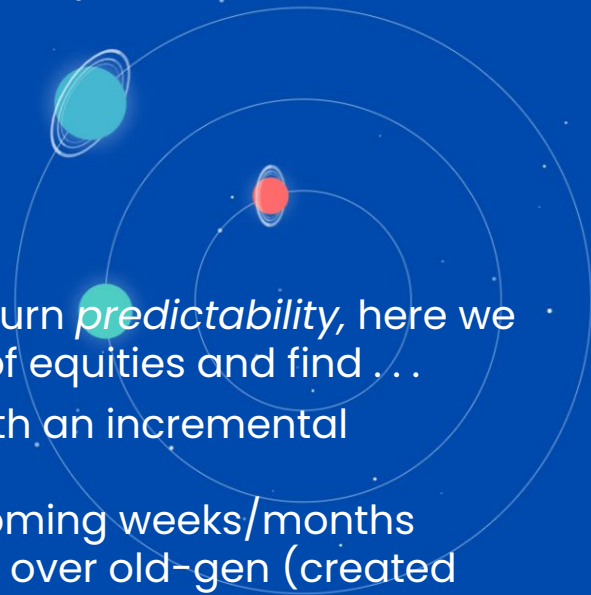
# ★ Research Process: Data Analysis

## We Conduct a Regression Analysis Controlling for...

- ★ Established Sentiment Measures in the Literature (Word-Based)
- ★ Past Returns
- ★ Firm Size
- ★ Book-to-Market Ratio
- ★ Time Fixed Effects
- ★ News Sentiment and Corporate Disclosure Dates

And we observe emojis 📈💰 still offer statistically significant return *predictability*, here we observe the cumulative abnormal returns in the cross-section of equities and find ...

- ★ A one standard deviation change in emoji is associated with an incremental annualized return of more than 5%
- ★ There is no evidence of price movement reversing in the coming weeks/months
- ★ New-gen accounts (created 2020+) drive this emoji-alpha over old-gen (created before 2020) accounts where alpha is found in words not emojis



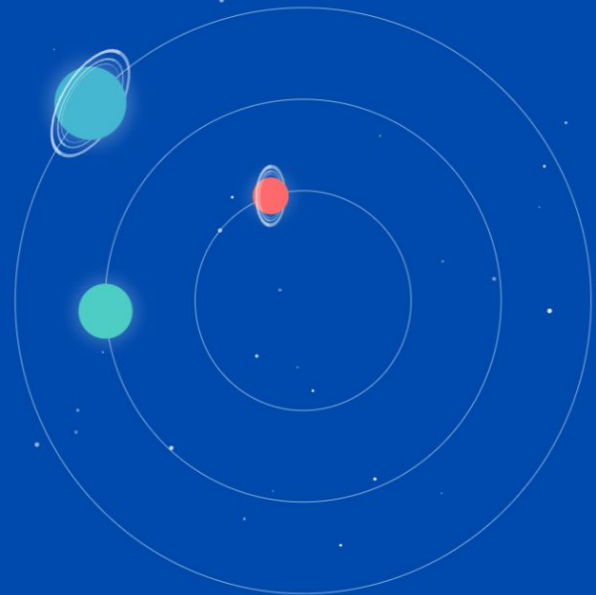
# ★ Research Process: Data Analysis

**Yeah, yeah, yeah, cool, cool, cool – but is it tradable?**

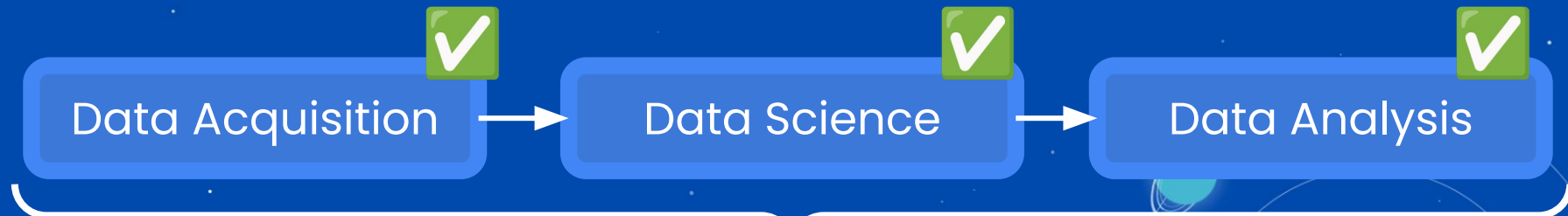
**Long Answer:** We have a lot of considerations and optimizations to make to determine if this is tradable

- Infrastructure
- Execution speed
- Transaction costs
- Data access
- Pipeline efficiency (everything was done offline)
- ...

**Short Answer:** Yes, and I've seen strategies like this deployed at several firms in varying capacities...



# ★ Research Process: Completed!



*This Process to Answer*

*These Questions*

- ~~★ Does the meaning of words change over time?~~
- ~~★ Is there a way to capture the meaning at a specific point in time?~~
- ~~★ Can we apply this methodology to emojis?~~
- ~~★ Do emojis offer unique information in cross-sectional return "prediction"?~~

# ★ Research Process: Completed!

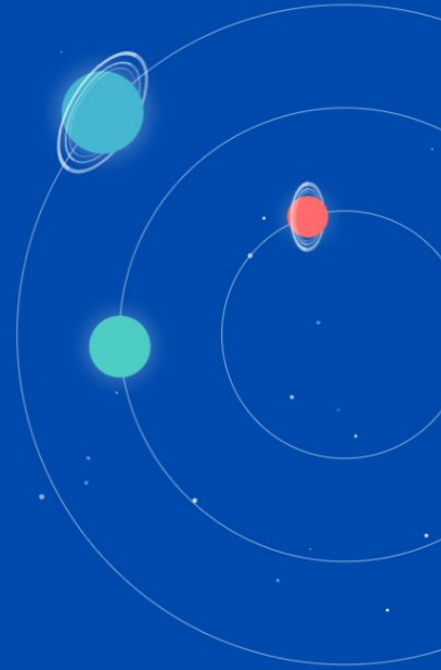


- ~~★ Does the meaning of words change over time?~~
- ~~★ Is there a way to capture the meaning at a specific point in time?~~
- ~~★ Can we apply this methodology to emojis?~~
- ~~★ Do emojis offer unique information in cross-sectional return "prediction"?~~

**Remark:** 98%+ of the questions you answer may yield unsatisfying, inconclusive, misleading, and the similar results – this is the nature of the industry! You may conduct a project for years+ and scrap it altogether but this is **NOT** a sunk cost, you are always developing as a researcher and that is the goal.

# Thank You!

## Questions?



QUANT GUILD

