Using social media data to enhance predictive models in finance

Nicholas Mitchell

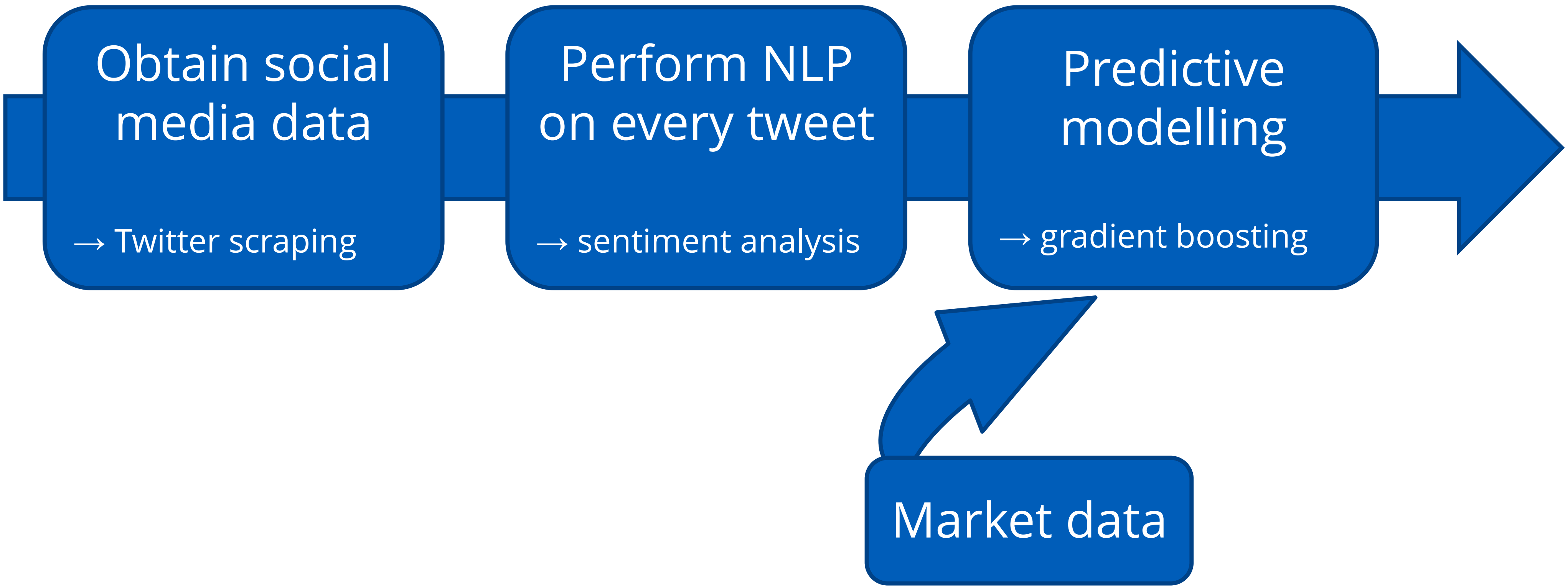
Munich, September 2016

We aimed to answer the question(s):

Can social media data improve predictive models of financial markets?

...by how much?

The Roadmap


Obtain social media data
→ Twitter scraping
Perform NLP on every tweet
→ sentiment analysis
Predictive modelling
→ gradient boosting
Market data

Tapping the *Twitterverse*

1. Purchase them
    - Every tweet since Twitter inception
    - Enterprise solution
    - Expensive

2. Official API
    - Free
    - Fast
    - Limited to last 10 days of data

3. Unofficial API… Scraping
    - Free
    - (currently) slow
    - (currently) unlimited data*

Scraping

- Twitter's advanced search:
  - Choose a search term
  - Select a date range
  - Tweets are displayed

- Page loads - starting with youngest tweets
- Scroll downwards → backwards in time
- Dynamically loading page

- Save all HTML code from Browser – WYSINAYG!

Search

Parsing

- Take raw HTML → parse using Xpath-Element-trees
- Create a one-tweet-per-line CSV file
- In addition to tweet text, extract useful meta-data:
  - Date
  - Number of likes
  - Number of retweets
  - Unique tweet ID
  - Username & ID
  - Geo-coordinates

Number of search terms:	13
Date range:	14.01.2013 – 11.09.2015
Total timeline:	982 days (695 weekdays)
Total tweets obtained:	2, 350, 217

Cleaning the tweets

- Required to facilitate accurate sen

- Tweets had to go from:

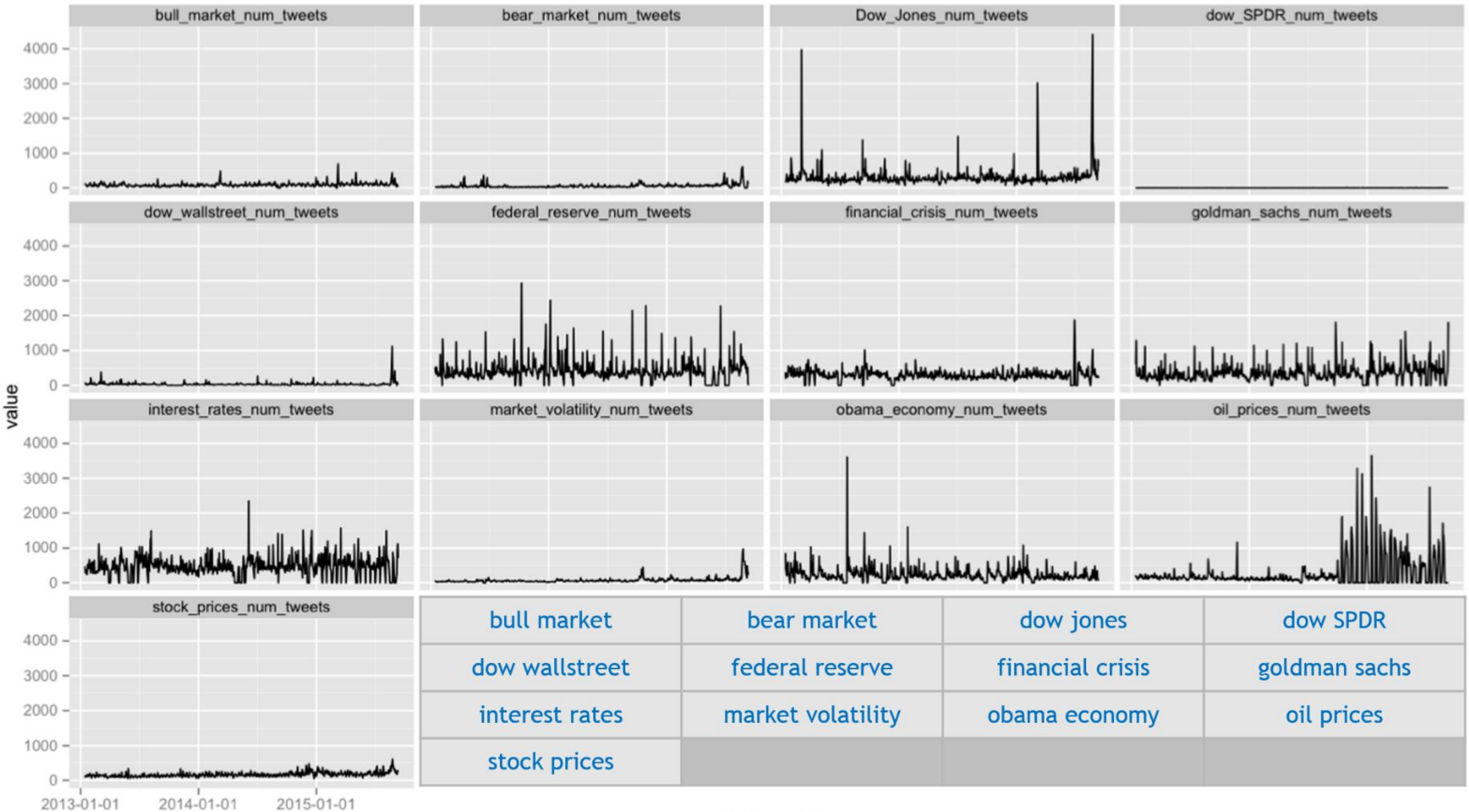
"""I wonder what people think abou
""Death Cross"" now? :)^M #trendfo

- To this:

I wonder what people think about t
Cross now? :) trendfollowing

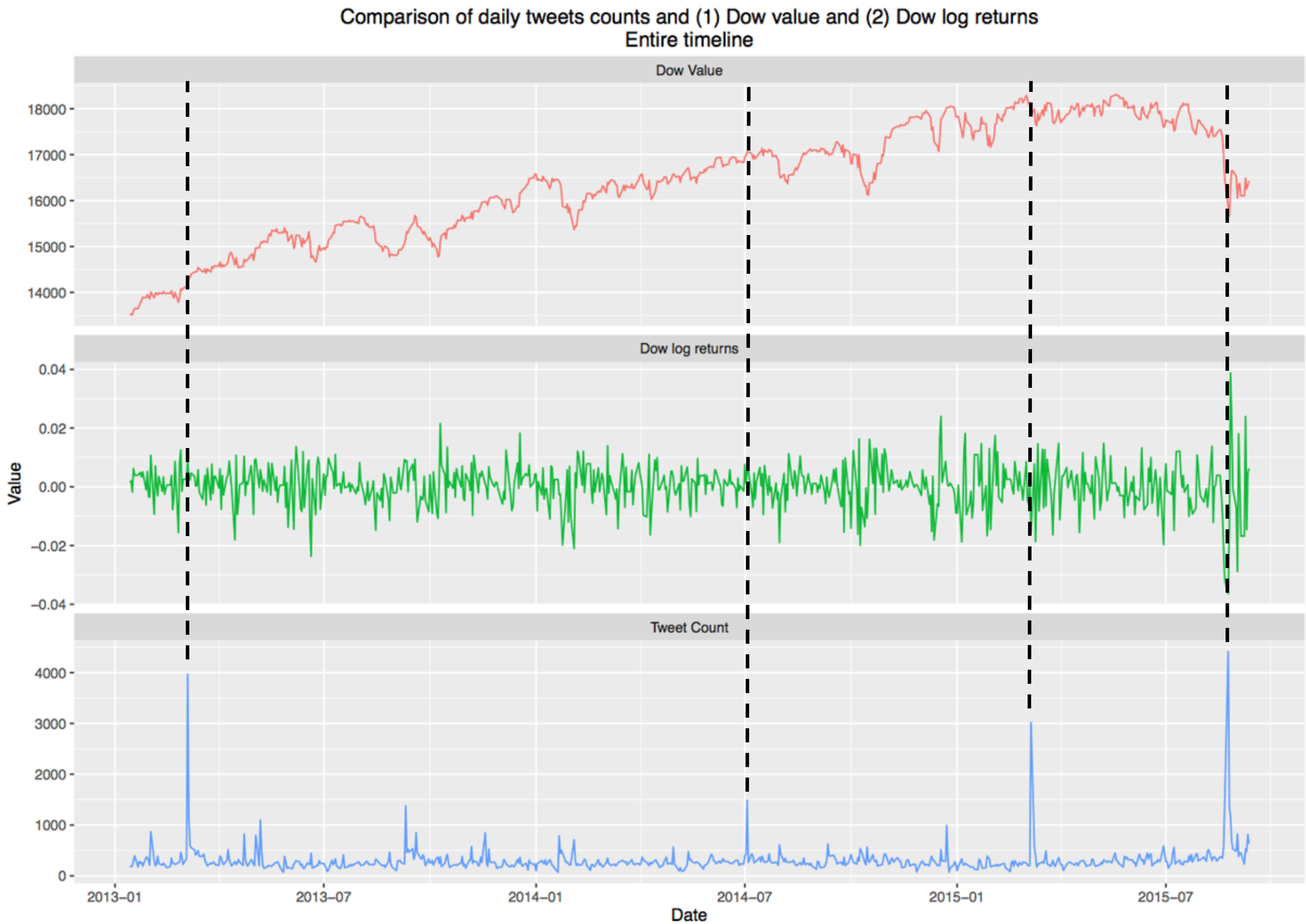
→ RegEx using Perl engine
→ hexadecimal char definitions
→ using an ASCII table

Breakdown of search terms and tweet count

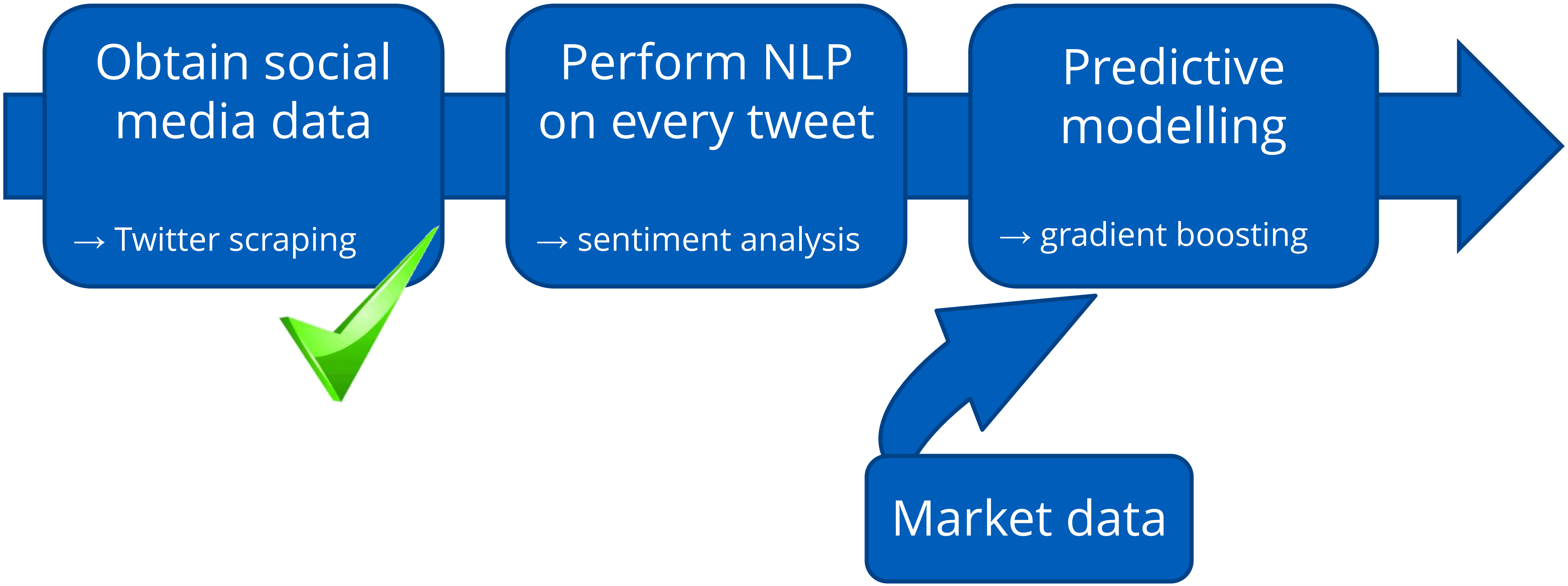


bull market	bear market	dow jones	dow SPDR
dow wallstreet	federal reserve	financial crisis	goldman sachs
interest rates	market volatility	obama economy	oil prices
stock prices			

Inspecting the Twitter data (1)



The Roadmap



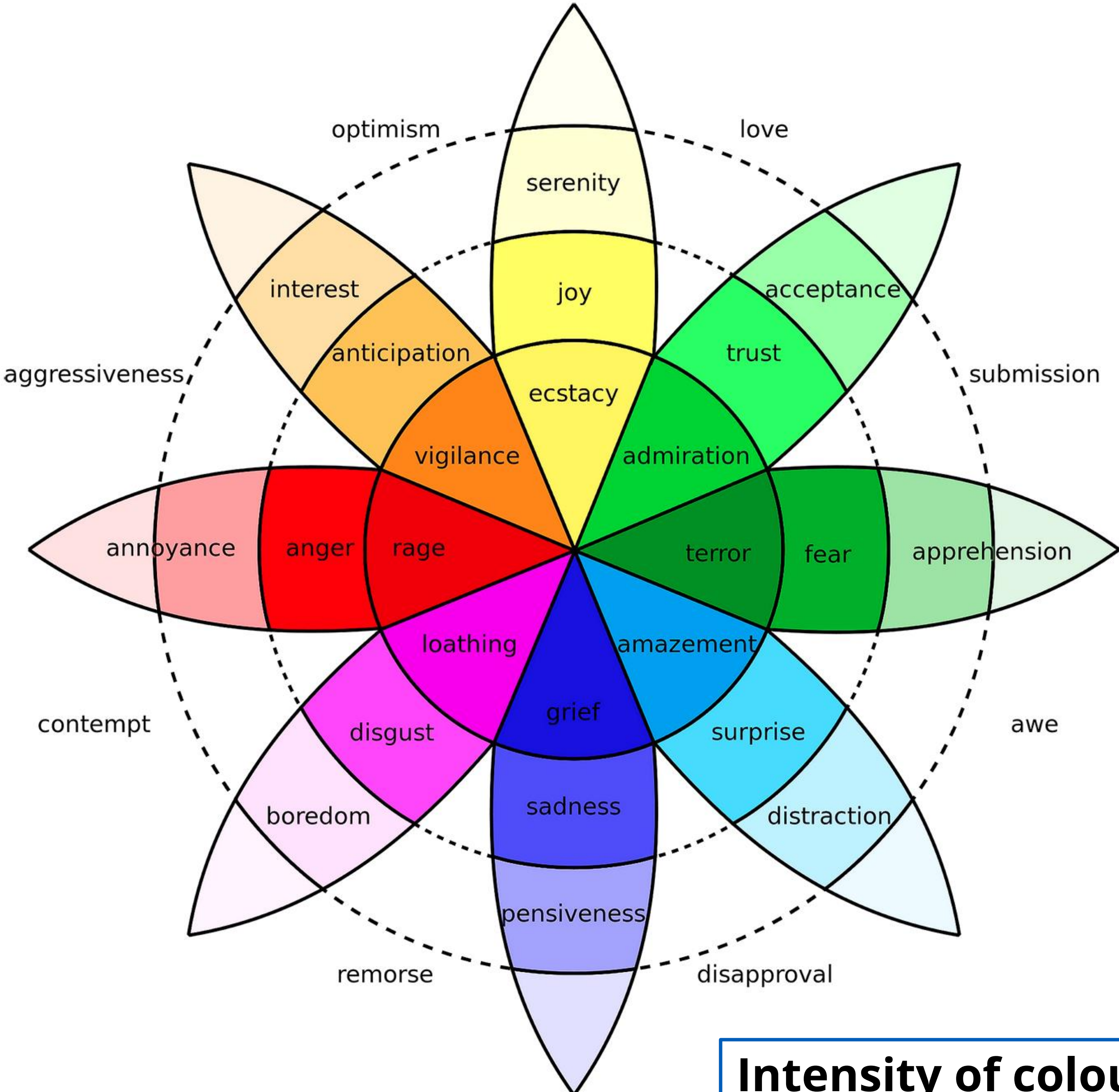
Obtain social media data
→ Twitter scraping
Perform NLP on every tweet
→ sentiment analysis
Predictive modelling
→ gradient boosting
Market data

Sentiment Analysis (1)

- 5 differing models
  - Each focussed on short texts / social media data
  - Employ diverse linguistic approaches

- Example approaches:
  - Grammar based – scored word lists for nouns, verbs, adjectives, etc.
  - Informal text – scores for smileys, slang and profanity
  - Pure word list, create by a community/mechanical Turk

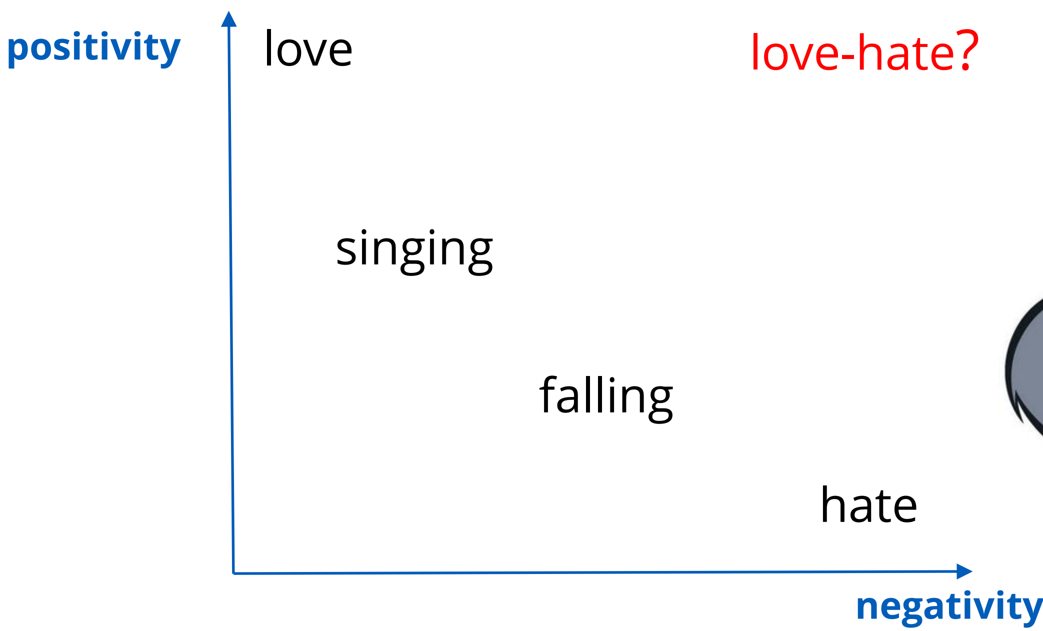
→ example: Plutchik's *Wheel of Emotion*



Intensity of colour ≈ intensity of emotion

Sentiment Analysis (2)

- Basic example: score each word on two dimensions:



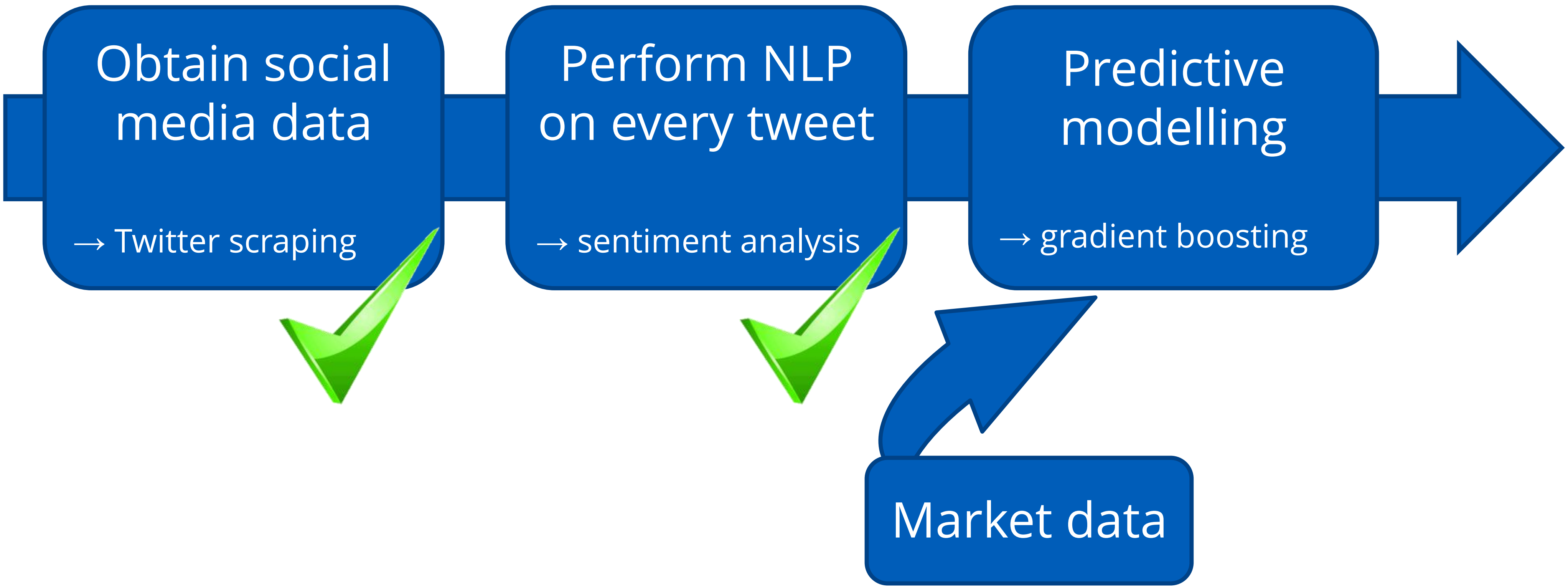
- Example:

Rupert loves safe investments, but Niko loves risky OTM options

(1,1) (4,1) (3,2) (1,1) (1,1) (1,1) (4,1) (1,3) (1,1) (2,1)

$\sum = (19, 13)$

The Roadmap


Obtain social media data
→ Twitter scraping
Perform NLP on every tweet
→ sentiment analysis
Predictive modelling
→ gradient boosting
Market data

Market data

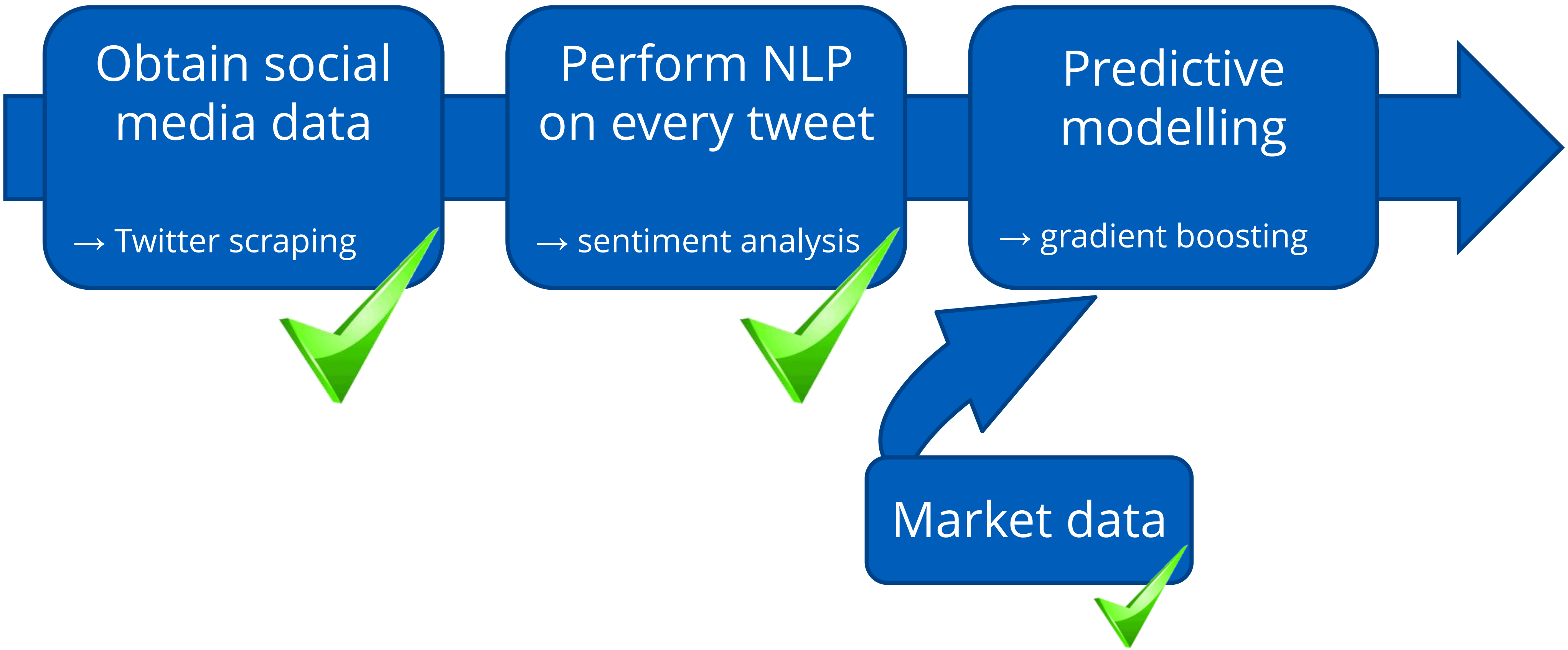
Commodities	Currency pairs	Fixed income
Gold spot Gold 3M Copper spot Copper 3M Oil (WTI) Natural gas	USD-AUD USD-CAD USD-EUR USD-GBP USD-JPY	U.S. Zero-coupon 1Y U.S. Zero-coupon 2Y U.S. Zero-coupon 5Y U.S. Zero-coupon 10Y U.S. Zero-coupon 15Y U.S. Zero-coupon 20Y
Indices	**Volatility indicators**	**ETF**
DAX Dow Jones FTSE100 Nikkei 225 S&P500 Shanghai SE	VIX (S&P500) Gold spread Copper spread	MSCI Emerging Markets

Datasets for comparison

		Σ
• Traditional market data		
– Small selection	$\rightarrow$ $\text{traditional}_{small}$	6
– Everything	$\rightarrow$ $\text{traditional}_{large}$	36
• Sentiment data only		
– Aggregated	$\rightarrow$ sentiment_{small}	22
– Individual	$\rightarrow$ sentiment_{large}	100
• Market + social media data	$\rightarrow$ combined	142

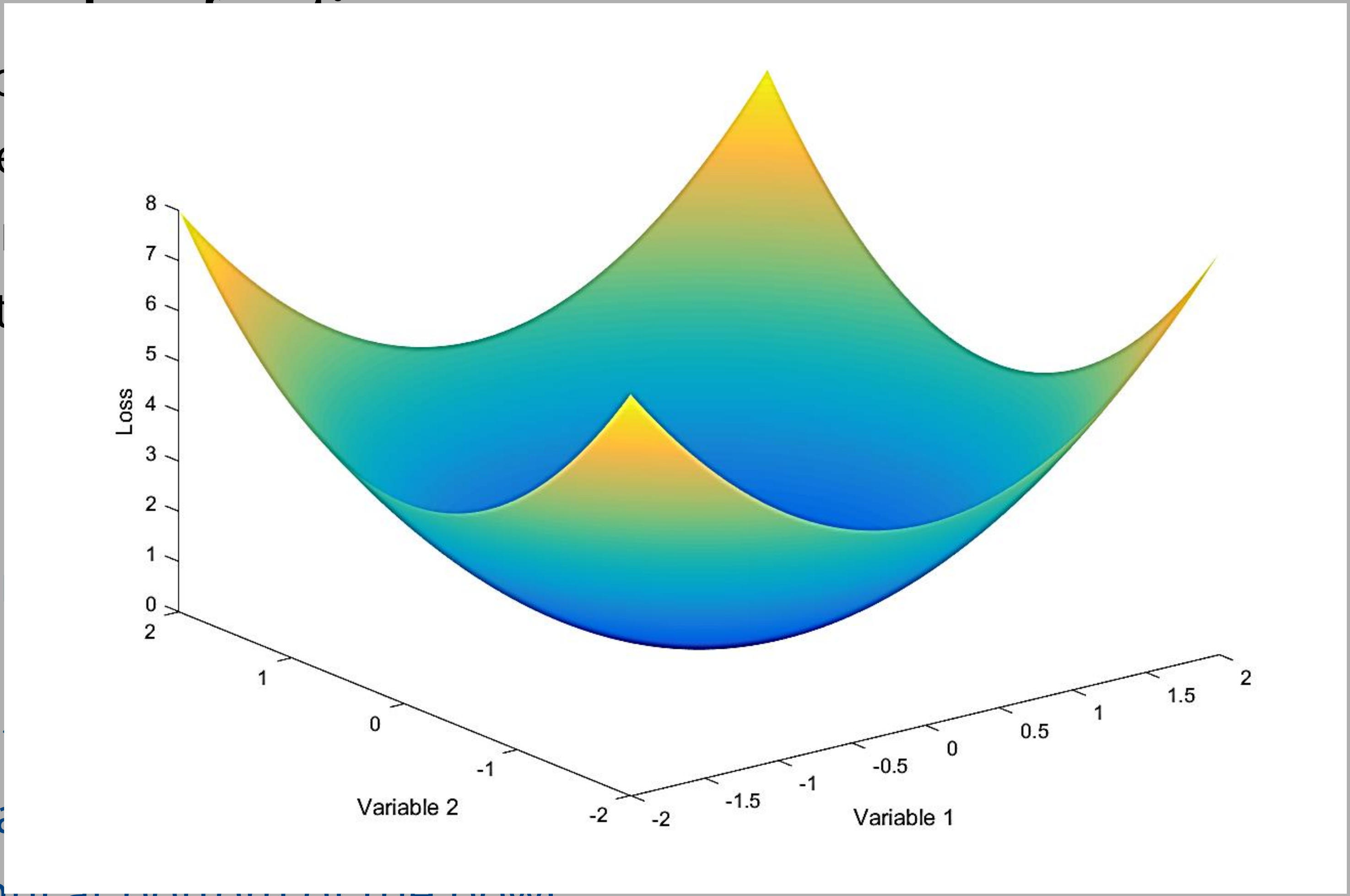
$\rightarrow$ **"Correlation cut-off" used, reducing datasets for final model**

The Roadmap


Obtain social media data
→ Twitter scraping
Perform NLP on every tweet
→ sentiment analysis
Predictive modelling
→ gradient boosting
Market data

Modelling: Gradient Boosting

- Defin
  - co
- Conve
- Conti
- Boost



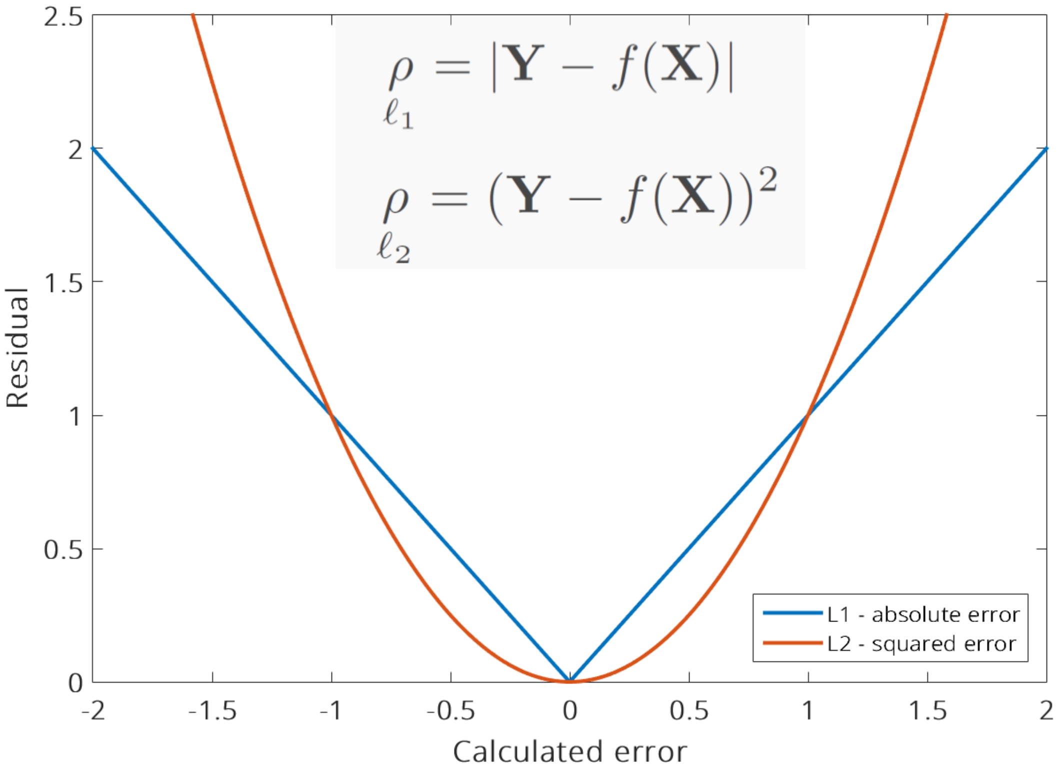
- Fit so
- Meas
- Upda
- Repea
- …. Until at bottom of the bowl

Gradient Boosting: theory in a nutshell

$$f^* := \underset{f(\cdot)}{argmin}\ \mathbb{E}_{\mathbf{Y},\mathbf{X}}[\rho(\mathbf{Y}, f(\mathbf{X}))]$$

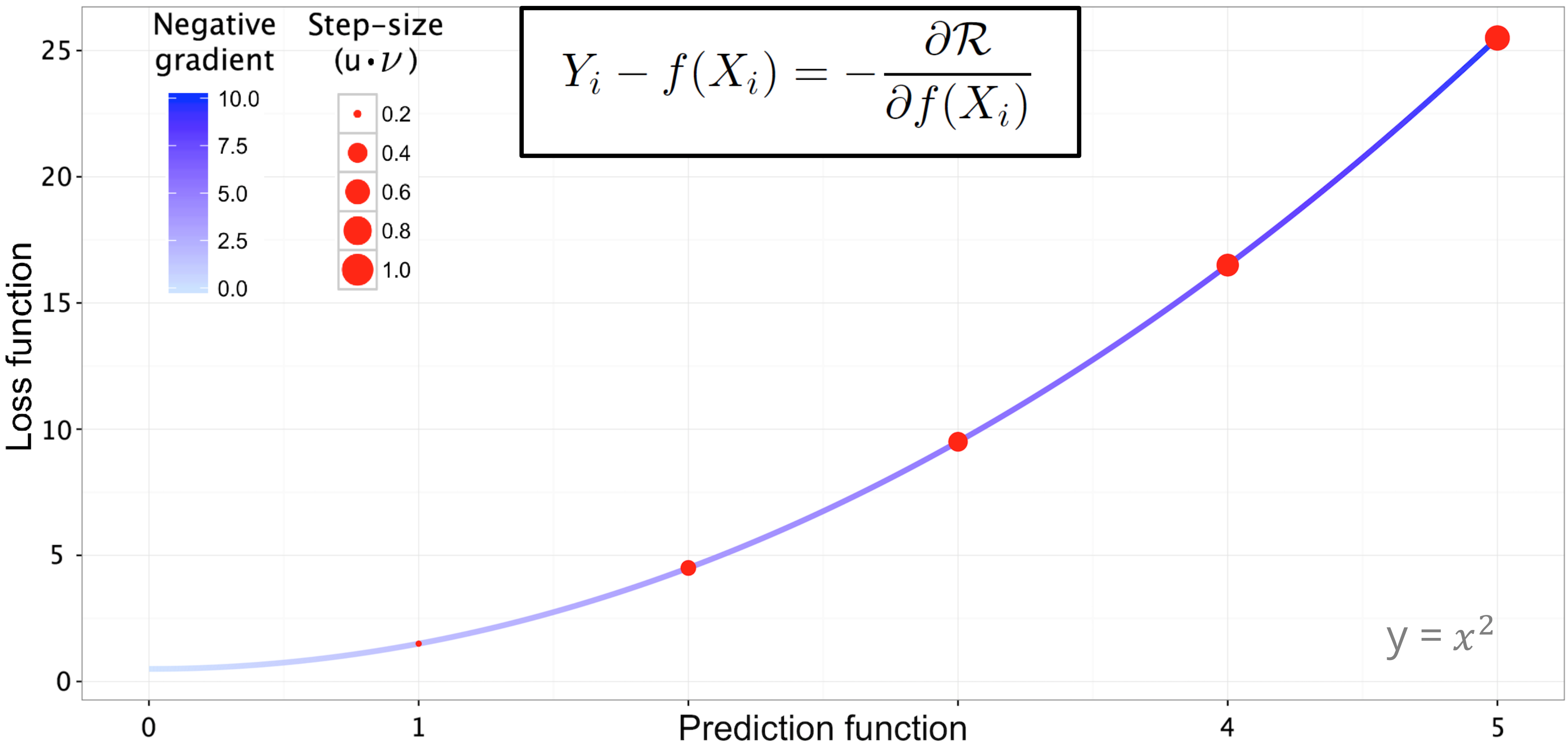
$$\mathcal{R} = \frac{1}{n}\sum_{i=1}^{n} \rho(Y_i, f(X_i))$$

$$\mathcal{R} = \frac{1}{n}\sum_{i=1}^{n} (Y_i - f(X_i))^2$$



$$\frac{\partial \mathcal{R}}{\partial f(X_i)} = \frac{\partial}{\partial f(X_i)}\left(\sum_{i=1}^{n} \rho(Y_i, f(X_i))\right) = \frac{\partial}{\partial f(X_i)}\left(\rho(Y_i, f(X_i))\right) = f(X_i) - Y_i$$

Adaptive descent

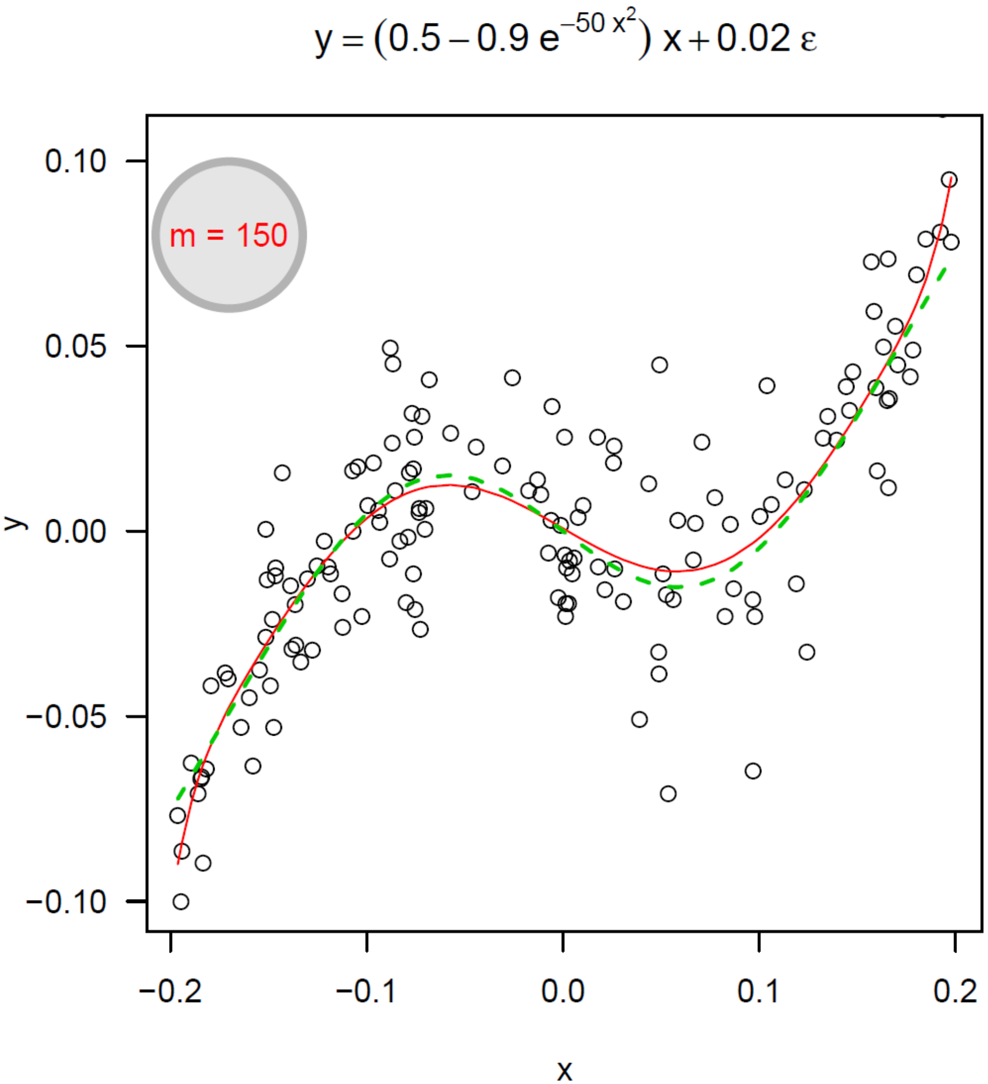


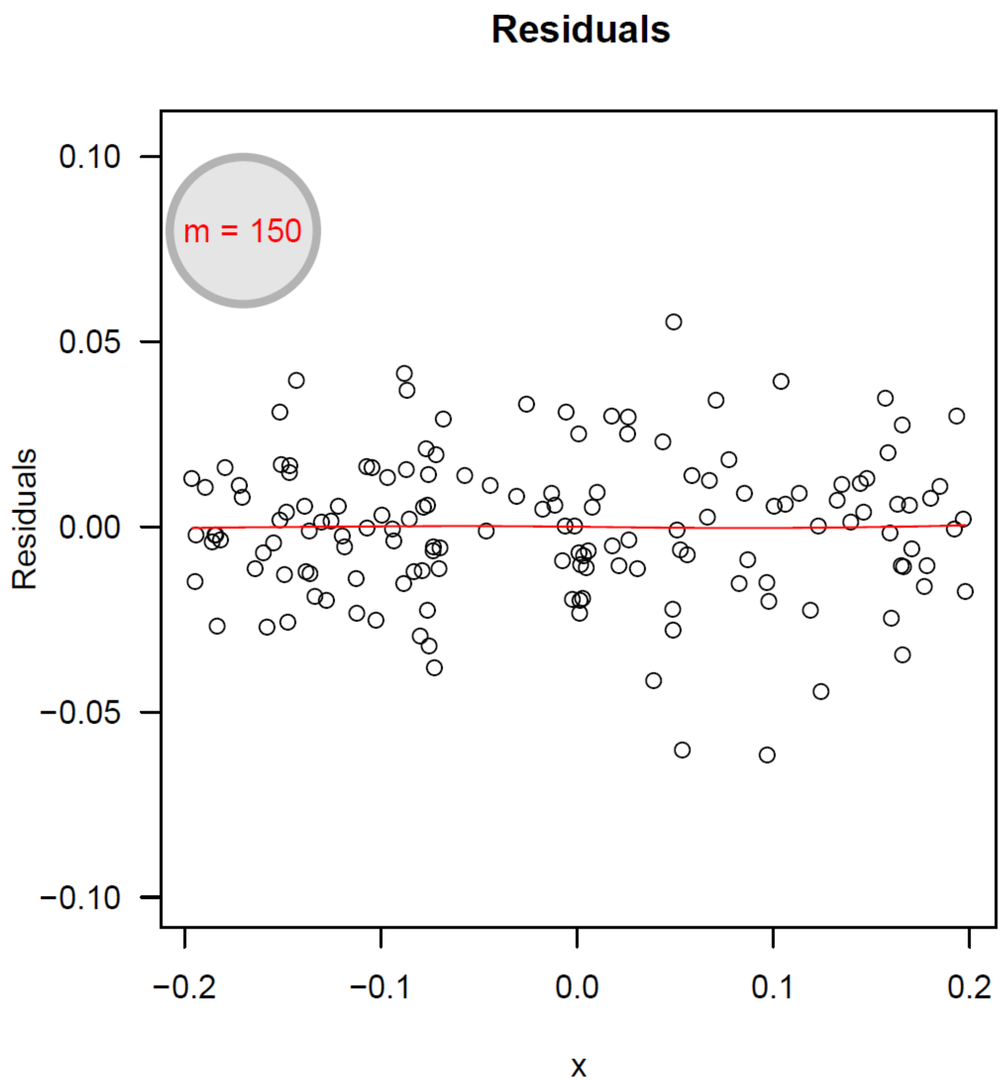
Iteratively improving our estimation

- Iteration counter: m
- Learning rate: v
- Total base-learners: n

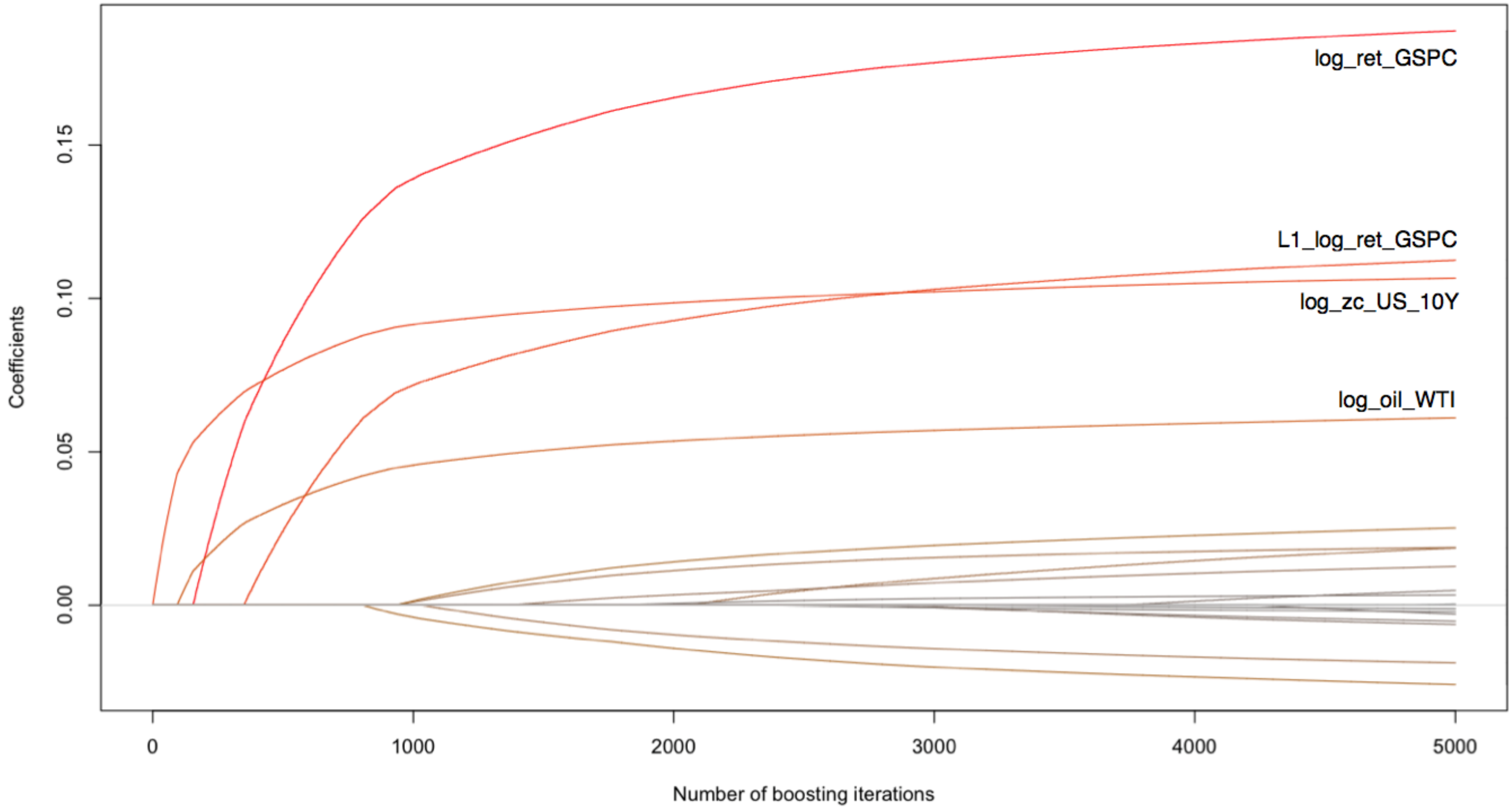
$$\hat{f}_{(1)}^{[m]} = \hat{f}_{(1)}^{[m-1]} + v \cdot -\frac{\partial}{\partial f_{(1)}}(\hat{f}_{(1)}^{[m-1]})$$

Example iteration – function approximation





Evolution of coefficients



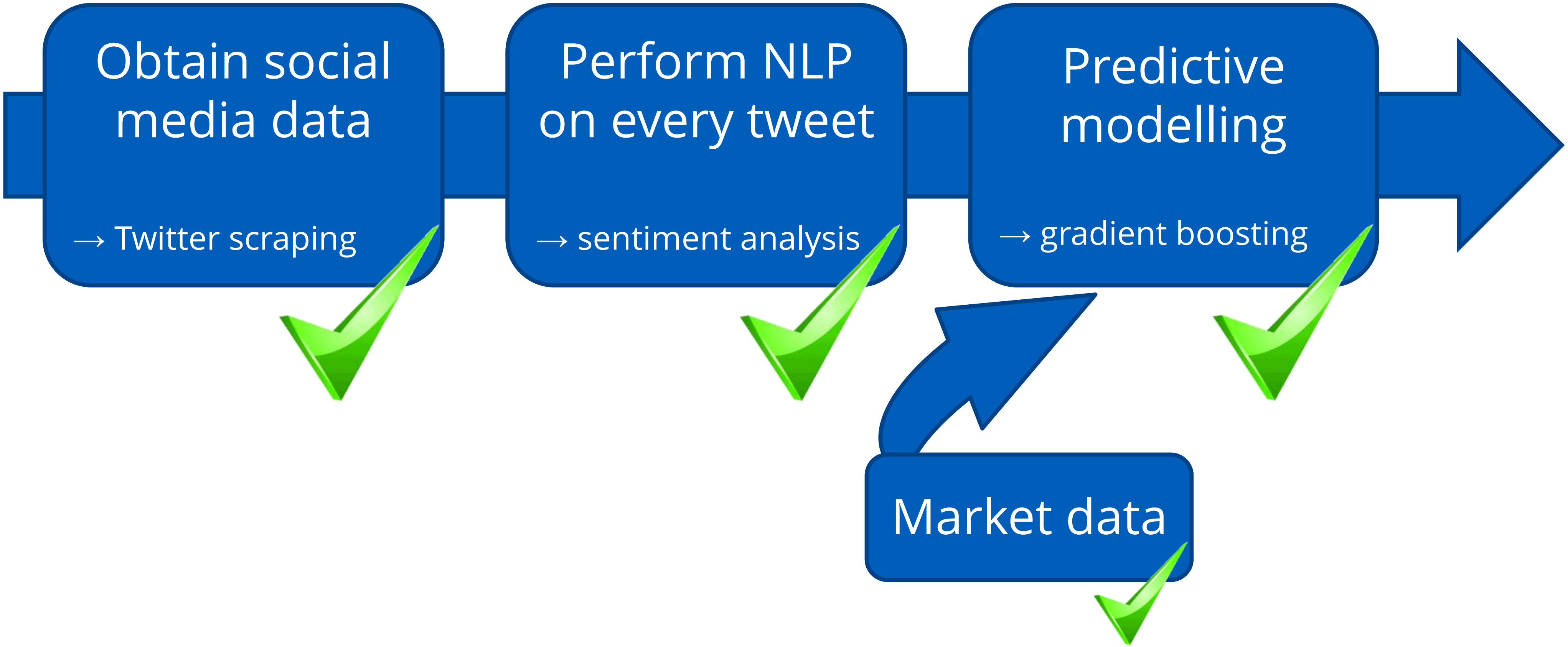
S&P500

Lagged S&P500

US Gov. bonds

Crude oil

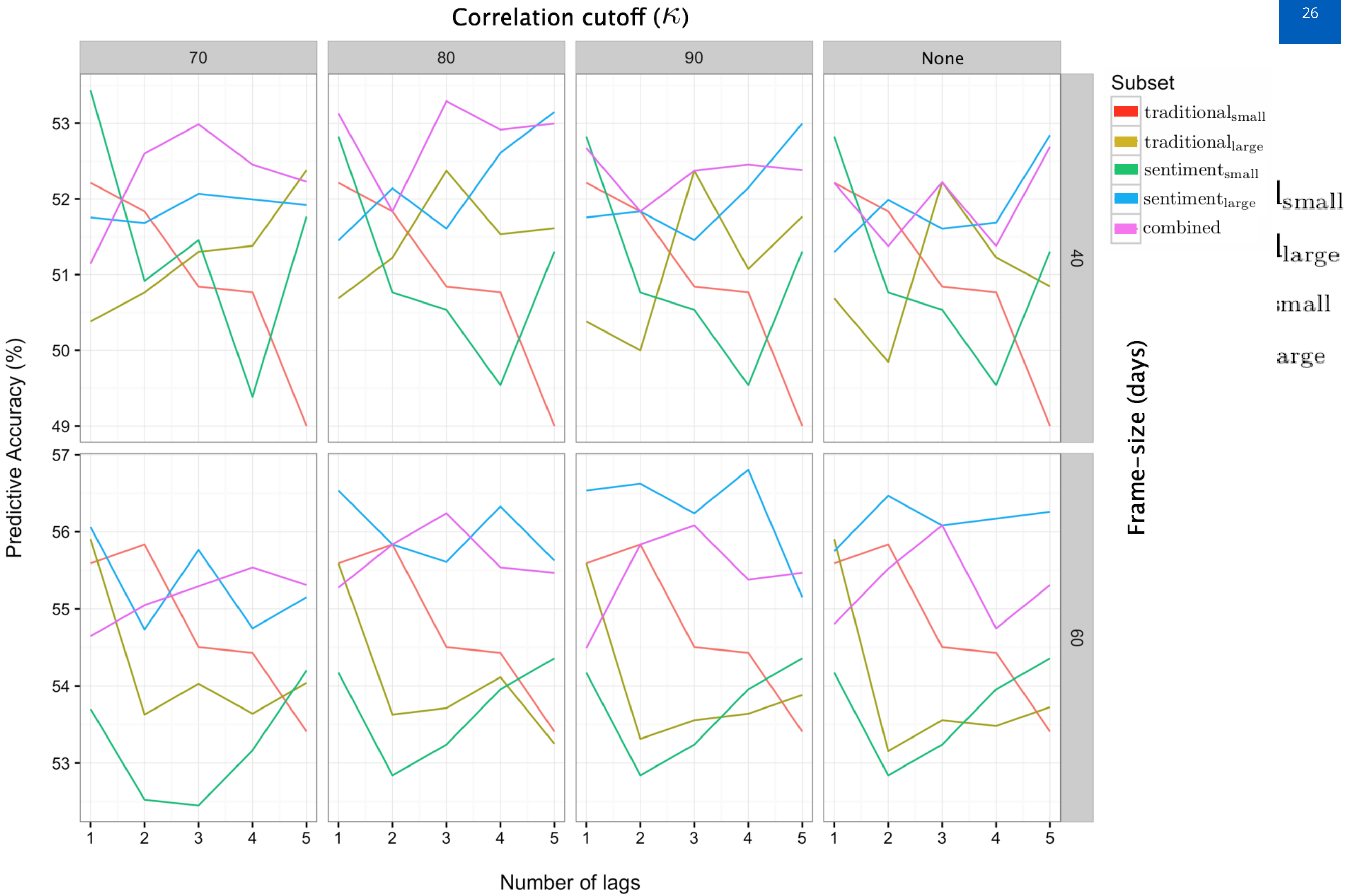
The Roadmap - master



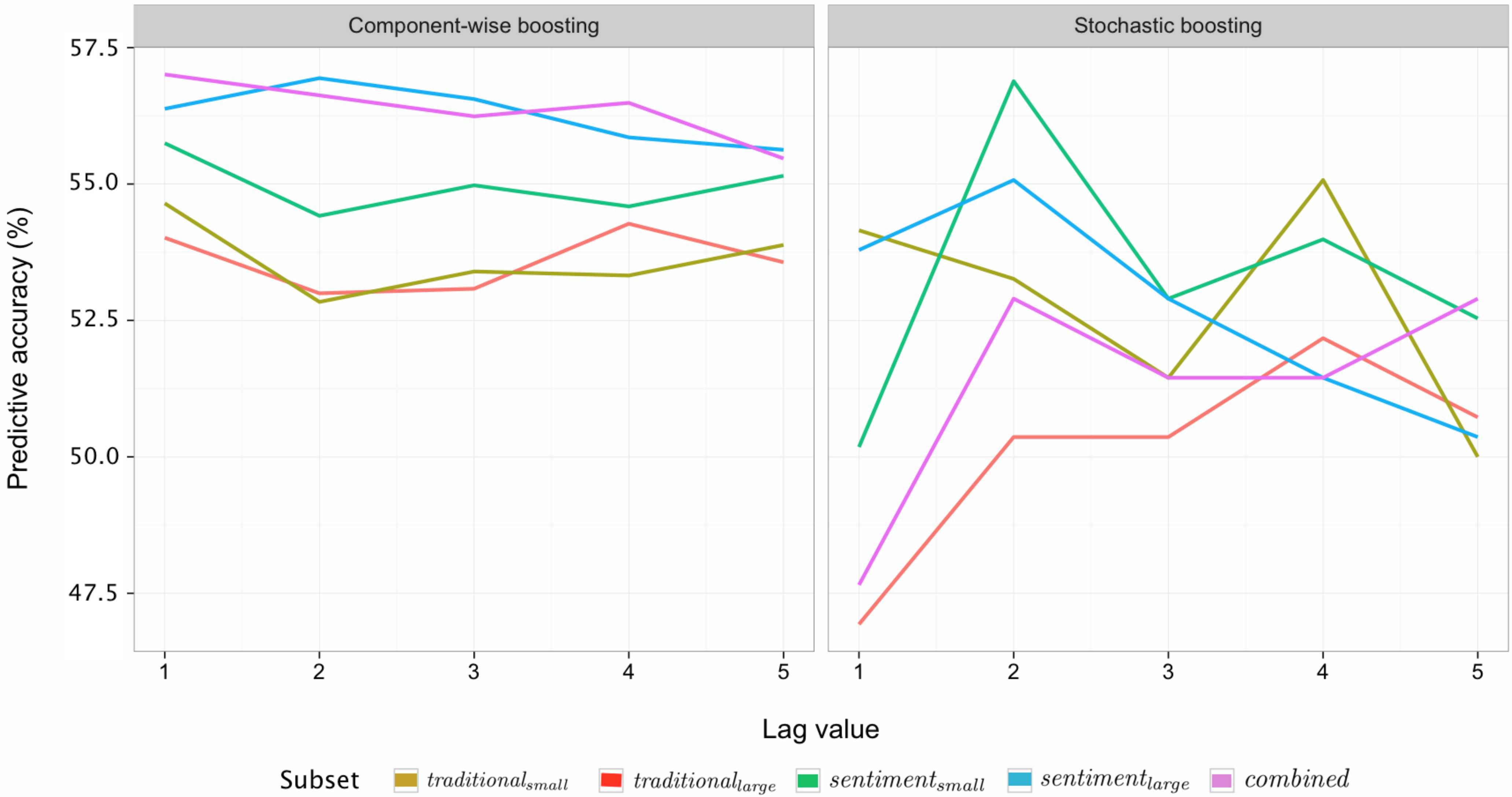
Reminder of the datasets

		Σ
• Traditional market data		
• small selection	$\rightarrow$ $traditional_{small}$	6
• everything	$\rightarrow$ $traditional_{large}$	36
• Sentiment data only		
• Aggregated	$\rightarrow$ $sentiment_{small}$	22
• Individual	$\rightarrow$ $sentiment_{large}$	100
• Market + social media data	$\rightarrow$ combined	142

$\rightarrow$ **Correlation measures taken reduce datasets for final model**

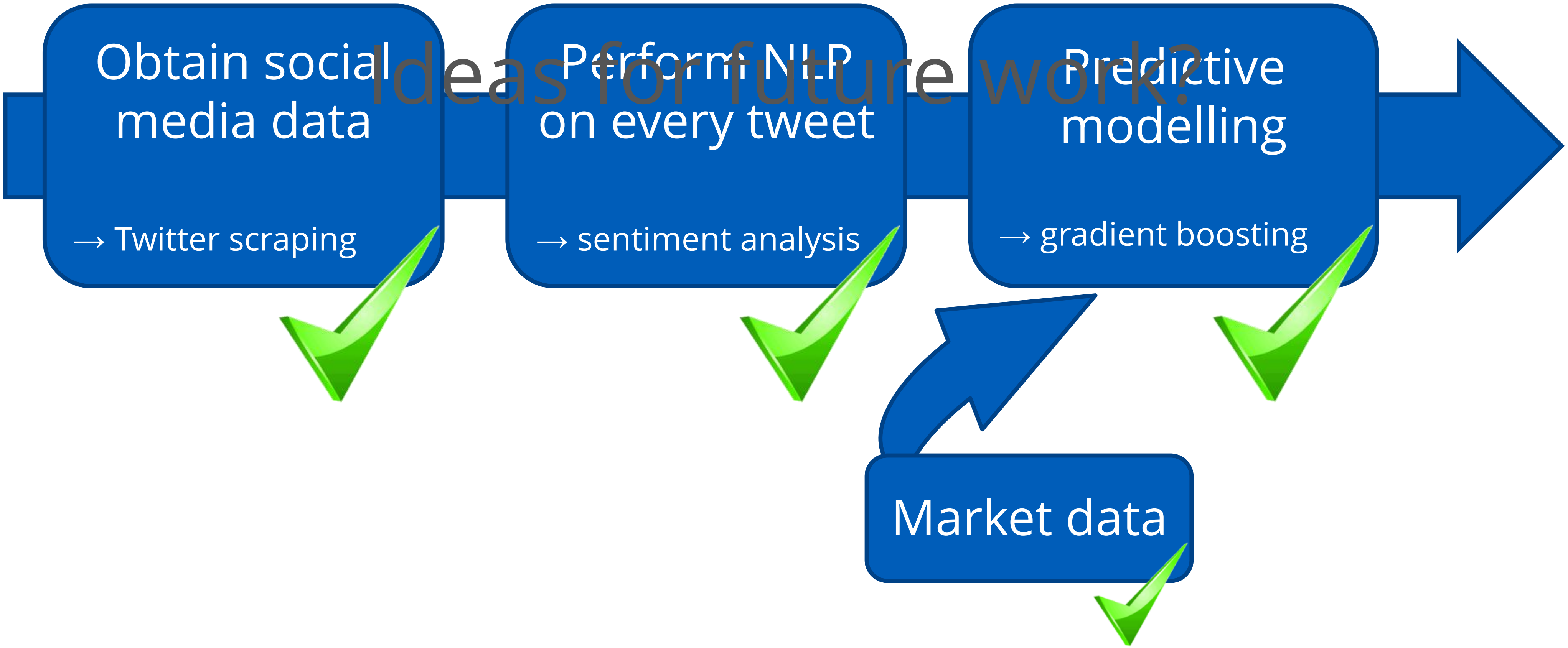

Correlation cutoff (κ)
70
80
90
None
Subset
$\text{traditional}_{\text{small}}$
$\text{traditional}_{\text{large}}$
$\text{sentiment}_{\text{small}}$
$\text{sentiment}_{\text{large}}$
combined
40
60
Frame-size (days)
Predictive Accuracy (%)
Number of lags
49
50
51
52
53
54
55
56
57
1
2
3
4
5

Stochastic Gradient Boosting - comparison



The Roadmap

Questions?



Ideas for future work?