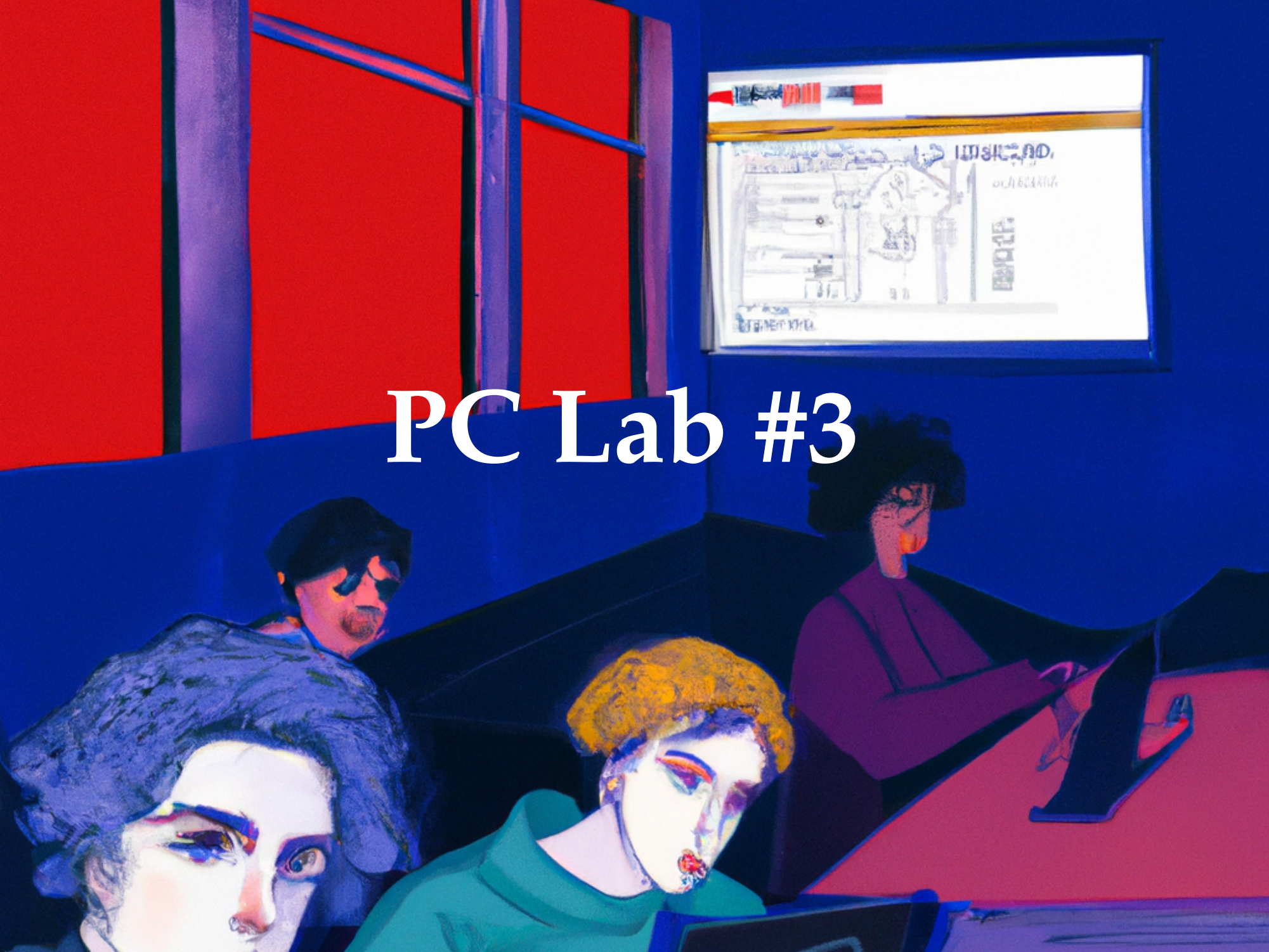
20598 – Finance with Big Data

PC Lab #3: Creating a Factor from Text Data (Week 4)

Clément Mazet-Sonilhac
clement.mazetsonilhac@unibocconi.it

Department of Finance, Bocconi University


PC Lab #3

PC Labs Grading

- PC Labs solutions are submitted as Jupyter Notebooks, via email
  - Email title : PCLab#3 - Group X - Name1 Name2 Name3
  - Your Jupyter Notebook starts in the same way (same .ipynb name)
  - Tell me (again) how long did it take

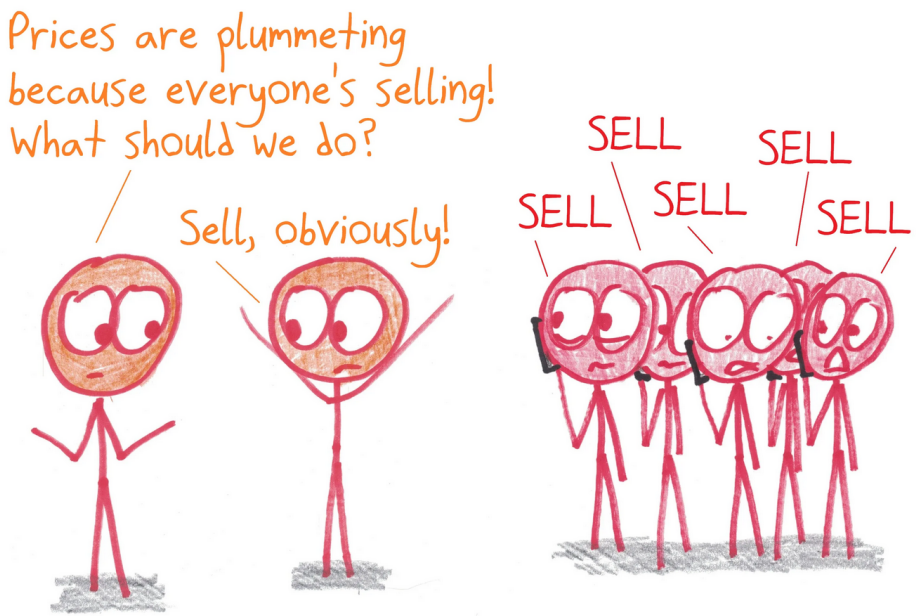
- PC Labs grade will depend on :
  - Your ability to submit it before the deadline **(Friday, midnight)**
  - The quality of your code (comments, readability, use of functions, etc.)
  - The structure of the Jupyter Notebook: well organized, explain what you are doing and why
  - Your ability to complete the tasks and innovate
  - You should maybe produce less, but more *useful* outputs

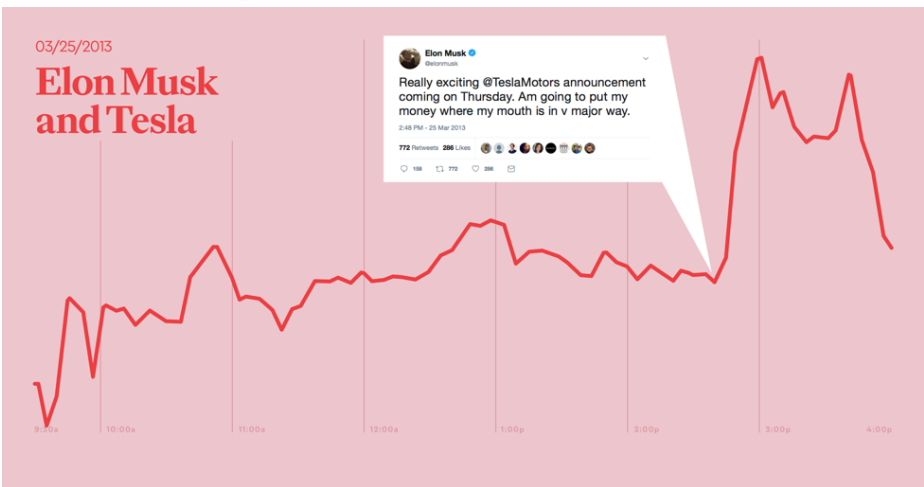

StockCats
@StockCats
Follow
EMERGENCY ALERTS
now
Emergency Alert
THE S&P 500 HAS GONE NEGATIVE ON THE DAY.
SEEK IMMEDIATE SHELTER. THIS IS NOT A DRILL.
Slide for more
1:52 PM - 16 Jan 2018
Michael B
The ma
compa
$4,095
The ma
500 co
$4,092
Downtown Josh Brown
@ReformedBroker
Follow
this person knocked $3 billion off the mar
value of Snapchat with a tweet, just in cas
you're curious about what inning humanit
in.
Kylie Jenner @KylieJenner
last night i had cereal with milk for the first time. life changing.
9:50 PM - 18 Sep 2018
The_Real_Fly
@The_Real_Fly
Follow
TRADERS AT OPEN VS CLOSE TODAY
4:56 PM - 29 Oct 2018

Goals

- Manipulate and visualize financial Tweets
- Clean text data
- Perform Tweets' sentiment analysis
- Compute a measure of media attention

Big picture context



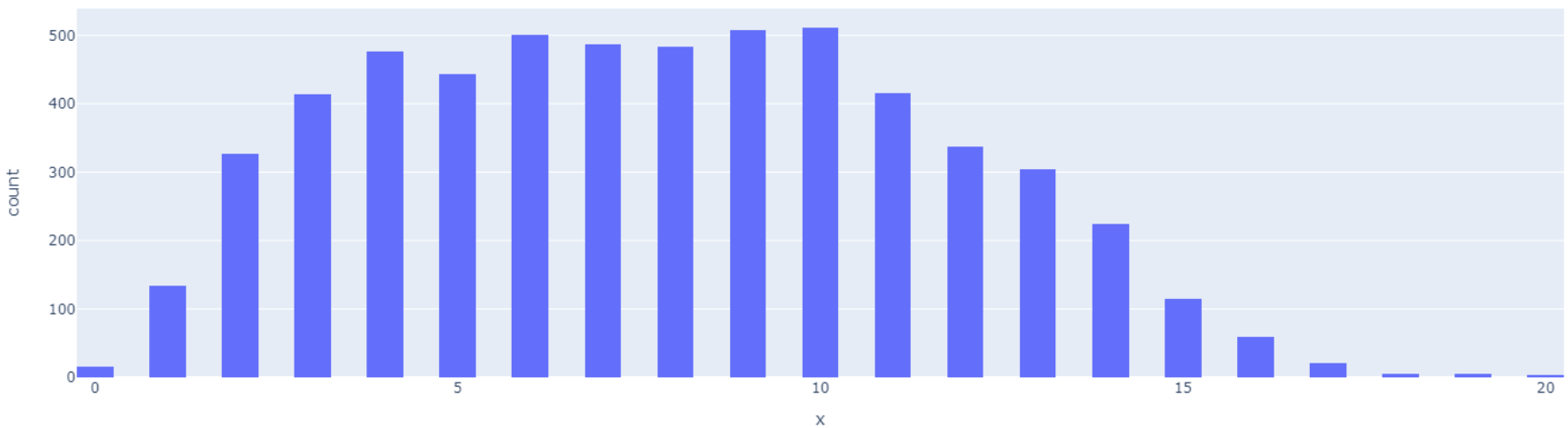


- You've just been hired by a sophisticated hedge-fund
- The hedge-fund manager is interested in Twitter's predictive power
- He asks you to perform sentiment analysis on a sample of recent financial tweets
- ... and to build a firm-level measure of media attention: that may be a great factor idea !

Task #1: Basic manipulation and descriptive statistics

- Import the `Data_PCLab3_Twitter_Stock_Sentiment.csv` data and describe the sample (data available on BBoard)

- How many tweets, how many words per tweets, distribution of number of words per tweets, average sentiment, etc.



Task #2 : Cleaning and Visualization

(You might know better than me)

- Usual cleaning steps:
  - cleaning URLs, mentions, hashtags, emojis
  - tokenization, lemmatization, stopword removal
  - tweet-specific text preprocessing with **ekphrasis** (normalizes hashtags, elongated words, emoticons).
- Plot a word cloud for text with positive and negative sentiment separately
- What is the number of unique words ?

Task #3 : Sentiment analysis

- Tweet sentiment is provided (1 vs. 0) — Should you trust it ?

Task #3 : Sentiment analysis

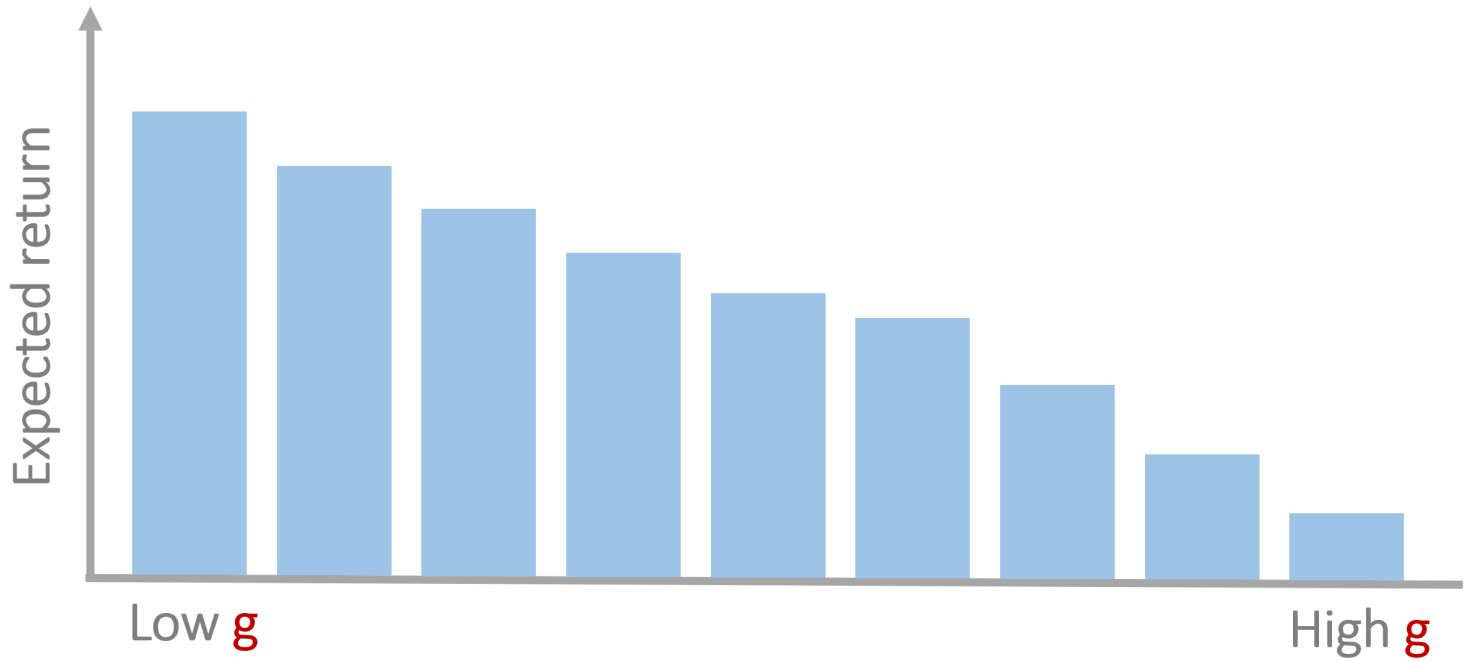
- Tweet sentiment is provided (1 vs. 0) — Should you trust it ?
- To produce your own sentiment analysis, use either :
  - Pretrained transformer (e.g., Twitter-RoBERTa, FinBERT)
  - GPT-style LLM zero-shot/few-shot classification
- What is the performance of these methods on financial tweets ? I.e., compare it to the original classification.

Task #4 : Measuring media attention (1)

- Use the list of tickers gathered during last PC Lab (see the web-scrapping part) to compute the number of tweets about each stock
  - e.g., `AAPL: 36 tweets, 12 negative, 24 positive`
- Rank the stocks by their amount of total media attention, or, alternatively : positive and negative media attention, level of disagreement (dispersion), etc.
- Be creative !

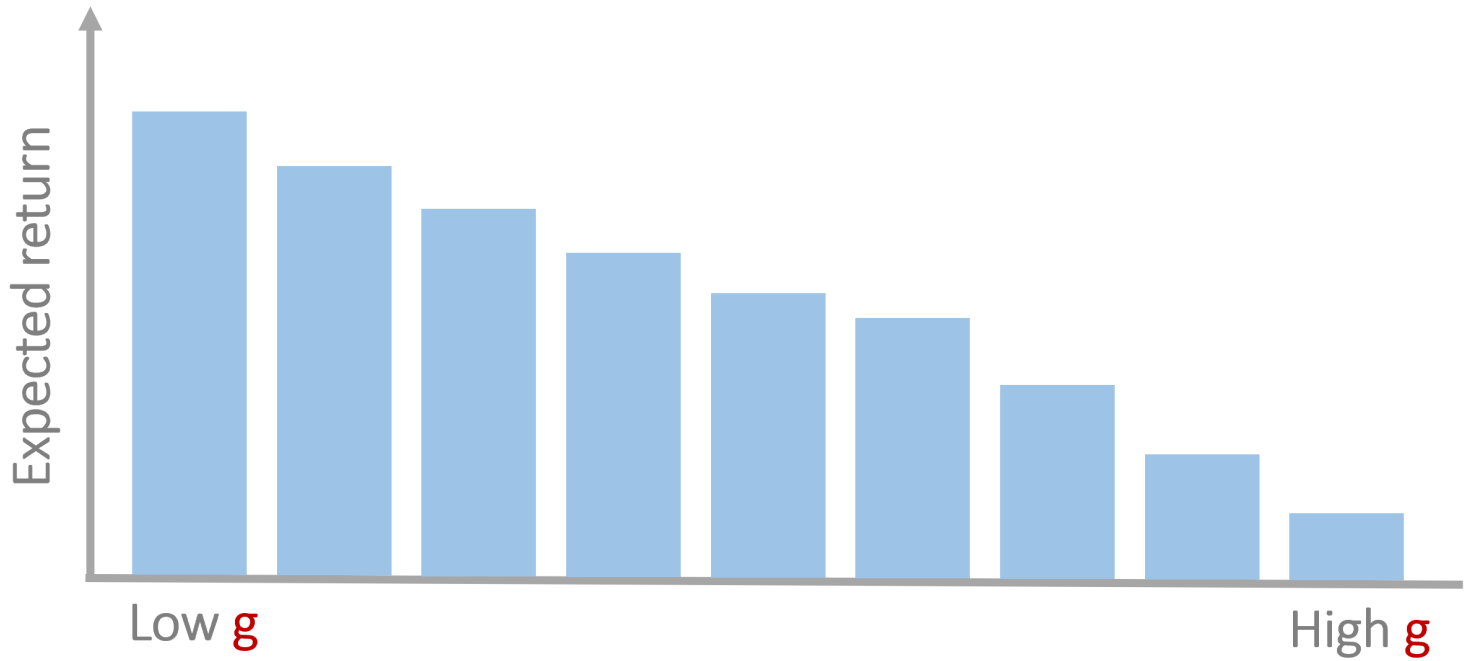
Task #4 : Measuring media attention (2)

- Create 10 portfolios based on your preferred measure of media attention
- Do you see a correlation between media attention **g** and stock returns ?



Task #4 : Measuring media attention (2)

- Create 10 portfolios based on your preferred measure of media attention
- Do you see a correlation between media attention **g** and stock returns ?



- If yes, could Twitter attention is likely to be a good factor ?
- Optional : same, but use FF-5 factors to purge returns from what a traditional model predicts, and plot media attention vs. erros

Task #4 : Measuring media attention (3) — Optional

- Download the Fama–French 5 factors (Mkt-RF, SMB, HML, RMW, CMA, and RF) from Ken French's Data Library
- Regress the stock's excess returns on these five factors
- Save the residuals from this regression. These represent the component of returns unexplained by traditional risk factors (abnormal returns).
- Redo same plot (see previous slide), but residuals vs. media attention.

Packages you may need (Outdated)

- Among others: `wordcloud`, `nltk.stem`, `nltk.corpus`, `nltk.tokenize`, `gensim`, `tensorflow`, `string.punctuation`, `sklearn`, `etc.`