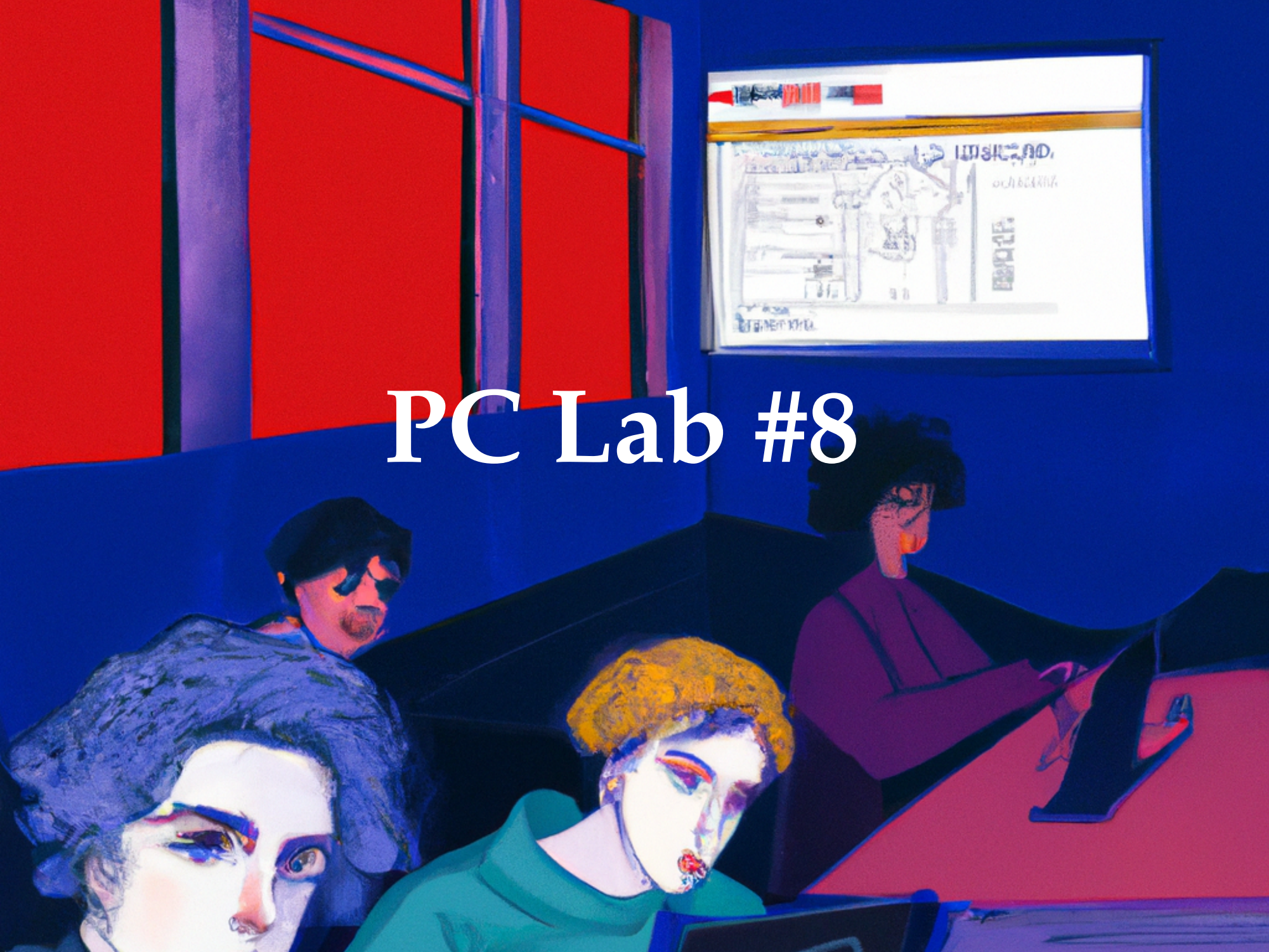
20598 – Finance with Big Data

PC Lab #8: Measuring Startup Innovation (Week 9)

Clément Mazet-Sonilhac
clement.mazetsonilhac@unibocconi.it

Department of Finance, Bocconi University


PC Lab #8

PC Labs Grading

- PC Labs solutions are submitted as Jupyter Notebooks, via email
  - Email title : PCLab#8 - Group X - Name1 Name2 Name3
  - Your Jupyter Notebook starts in the same way (same .ipynb name)
  - Tell me (again) how long did it take

Reminder: Andrea Andolfatto (andrea.andolfatto@phd.unibocconi.it) should be cc'ed to your email.

PC Labs Grading

- PC Labs grade will depend on :
  - Your ability to submit it before the deadline
  - 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


Who is the next
Facebook?

Goals

- Manipulate and visualize patent data
- Use unsupervised learning methods to classify patents
- Implement a measure of textual similarity

Big picture context

- You've been hired as a data scientist by **Sequoia Capital** (WhatsApp, Zoom, Airbnb, etc.)
- The fund is constantly looking for the *needle in the haystack*
- You are responsible to identify breakthrough patents (see Bonelli, 2022)

Kelly et al., 2021

Measuring Technological Innovation over the Long Run†

By Bryan Kelly, Dimitris Papanikolaou, Amit Seru, and Matt Taddy*

We use textual analysis of high-dimensional data from patent documents to create new indicators of technological innovation. We identify important patents based on textual similarity of a given patent to previous and subsequent work: these patents are distinct from previous work but related to subsequent innovations. Our importance indicators correlate with existing measures of patent quality but also provide complementary information. We identify breakthrough innovations as the most important patents—those in the right tail of our measure—and construct time series indices of technological change at the aggregate and sectoral levels. Our technology indices capture the evolution of technological waves over a long time span (1840 to the present) and cover innovation by private and public firms as well as nonprofit organizations and the US government. Advances in electricity and transportation drive the index in the 1880s, chemicals and electricity in the 1920s and 1930s, and computers and communication in the post-1980s. (*JEL* C43, N71, N72, O31, O33, O34)

Task #1 : Basic manipulation and descriptive statistics

- Import the `PatentAbstract_PCLab8.csv` data from here (Link)
- Usual cleaning steps : remove punctuation, stopwords, short words, etc.
- Describe the sample :
  - How many patents do you have ? How many unique words ?
  - Frequency of letters (Lewand's law) and average nb. of words per abstract
  - Most common words?
- **If the data is too large (especially for the following tasks) don't hesitate to consider only a sub-sample.**

Task #2 : Create patents' technological class

- In the provided data, you don't have the date nor the technological class of the patent
- Consider that all patents are from the same (couple of) year(s)
- Use `Gensim`'s LDA package to perform Topic Modeling to extract the hidden topics (technological classification) from the patents.You want to assign each patent to a technological class.
- What are the number of classes you end up with ? What are the most frequent words in each class ? Does it make sense ?

Task #3 : identify breakthrough patents

- Create a measure of semantic similarity (inspired by Kelly et al. (2021) or pick your own method)
- Alternatively, you could project the patents in a bi/tri-dimensional vector space and tag the outliers
- Identify, for each technological class, the outliers (i.e., breakthrough patents)
- Find a way to graphically illustrate your findings
- What do you think about this idea to measure breakthrough innovation. Ideally, what would you do ?

No **optional** task?

- Yes, because they are all optional. You already have 5 (very good) grades!
- This last PC Lab is for fun, you can do it entirely or not, submit it or not. In any case, **try to learn something new!**