BUSN 20800: Big Data

Lecture 1: Data Exploration and Visualization

Dacheng Xiu

University of Chicago Booth School of Business

This is a class about Inference at Scale

We're here to make sure you have the tools for making good decisions based on large and complicated data.

A mix of practice and principles and a hands-on experience

- Solid understanding of essential statistical principles
- Concrete analysis ability and best-practice guidelines

What is in a name?

Statistics, Machine Learning, and Data Science

There are many labels for what we do... The similarities are much bigger than any distinctions.

"Statistics means the practice or science of collecting and analyzing numerical data in large quantities."

"Data Science means the practice of liberating and creating meaning from data using scientific methods."

D. Donoho "50 Years of Data Science"

Arthur Samuel coined the term Machine Learning in his 1959 paper: "Some studies in machine learning using the game of checkers".

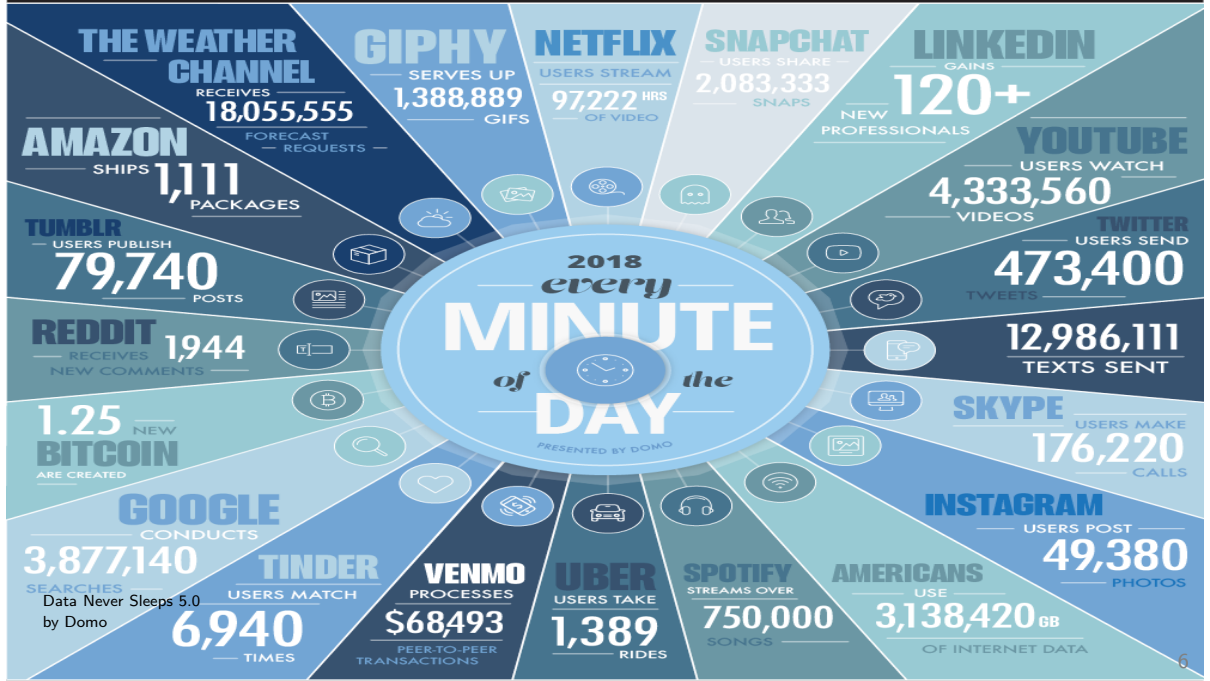
Combine expertise from statistics on how to extract information from data with computational ideas that enable efficient implementation on large data sets.

Machine learning is to big data as human learning is to life experience.

The Big Data Opportunity

New Technologies have made available vast quantities of digital data

- According to a report by IBM in 2017, 90% of all data in existence was created in the past TWO YEARS!
- "There were 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days" — Eric Schmidt, CEO of Google



Data Never Sleeps 5.0
by Domo

A Paradigm Shift

There is a paradigm shift from machine *programming* to machine *learning*.

- In conventional programming, tell computer what to do, breaking big problems into many small, precisely defined tasks
- Learn (estimate) from observational data, instead of requiring pre-specified logic, for decision making and problem solving

A Paradigm Shift From Machine *Programming* to Machine *Learning*

Is this a valid email address? `dacheng.xiu@chicagobooth.edu`

Conventional Programming

`IF` upper/lowercase letters or digits

`AND` "@" followed by a valid domain

`AND NOT` special characters "!#$..."

`AND` ...

`THEN` valid

`ELSE` invalid

A Paradigm Shift From Machine *Programming* to Machine *Learning*

Is this a valid email address? `dacheng.xiu@chicagobooth.edu`

Conventional Programming

`IF` upper/lowercase letters or digits

`AND` "@" followed by a valid domain

`AND NOT` special characters "!#$..."

`AND` ...

`THEN` valid

`ELSE` invalid

Machine Learning

1. (Big) data on valid/invalid email addresses

Y/N	Address
0	jaime@lannister
1	hound@clegane.com
0	john.snow@GO†.edu
...	...

+

2. Stats Model: $\mathrm{Y/N} = b_0 + b_1(@) + b_2(!\#\$\&\%?{*}) \ldots$

=

3. Estimated probability of valid email

Machine Learning Rocks!

Machine Learning Successes

Machine Learning Rocks!

Machine Learning Successes

1997 Deep Blue beats Kasparov

Machine Learning Rocks!

Machine Learning Successes

1997 Deep Blue beats Kasparov

2009 Google self-driving car

Machine Learning Rocks!

Machine Learning Successes

1997 Deep Blue beats Kasparov

2009 Google self-driving car

2011 Watson wins Jeopardy!

Machine Learning Rocks!

Machine Learning Successes

1997 Deep Blue beats Kasparov

2009 Google self-driving car

2011 Watson wins Jeopardy!

2012 Microsoft translates English to Chinese in real time

Machine Learning Rocks!

Machine Learning Successes

1997 Deep Blue beats Kasparov

2009 Google self-driving car

2011 Watson wins Jeopardy!

2012 Microsoft translates English to Chinese in real time

2016 AlphaGo beats Lee Sedol;

Machine Learning Rocks!

Machine Learning Successes

1997 Deep Blue beats Kasparov

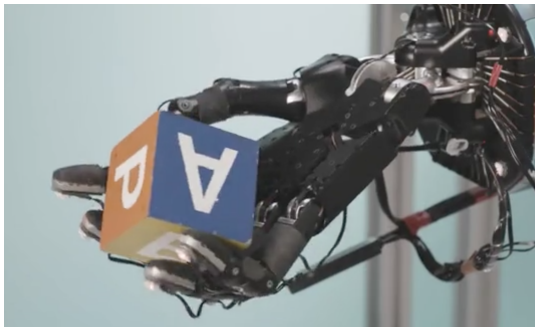
2009 Google self-driving car

2011 Watson wins Jeopardy!

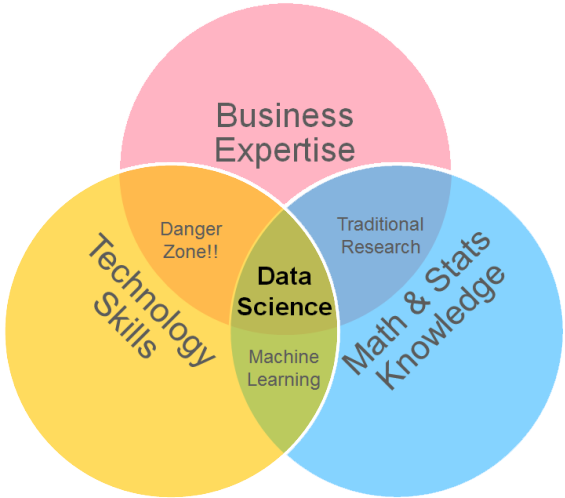
2012 Microsoft translates English to Chinese in real time

2016 AlphaGo beats Lee Sedol;

2018 OpenAI's "Dactyl" learns to manipulate objects



ML and Data Science



Today's Data Science Movement

According to a report by the US Bureau of Labour Statistics, the rise of data science needs will create roughly 11.5 million job openings by 2026.

"There will be a shortage of talent necessary for organization to take advantage of big data. By 2018, the United States alone could face a shortage of 140 000-190 000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions."
Big data: The next frontier for innovation, competition, and productivity (McKinsey Report 2011)

"Statistical thinking not only helps make scientific discoveries, but it quantifies the reliability, reproducibility and general uncertainty associated with these discoveries. Because one can easily be fooled by complicated biases and patterns arising by chance, and because statistics has matured around making discoveries from data, statistical thinking will be integral to Big Data challenges."
Discovery with Data: Leveraging Statistics with Computer Science to Transform Science and Society (ASA White Paper 2014)

Today's Data Science Movement

Google's Chief Economist Hal Varian on Statistics and Data

POSTED TO QUOTES, STATISTICS | NATHAN YAU

> *The ability to take data – to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it.*
>
> *I think statisticians are part of it, but it's just a part. You also want to be able to visualize the data, communicate the data, and utilize it effectively. But I do think those skills – of being able to access, understand, and communicate the insights you get from data analysis – are going to be extremely important. Managers need to be able to access and understand the data themselves.*

Danger Zone?

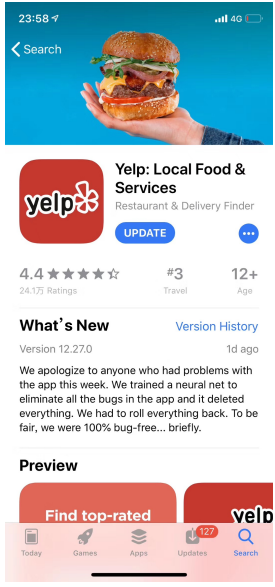
Data science is the umbrella term for inference in a world that is messier than in your old statistic textbook.

BUSINESS

What Went Wrong With Zillow? A Real-Estate Algorithm Derailed Its Big Bet

The company had staked its future growth on its digital home-flipping business, but getting the algorithm right proved difficult

By Will Parker and Konrad Putzier
Nov. 17, 2021 11:29 am ET



What does ‘Big’ Data mean?

Big in both the number of observations (size ‘N’)
and in the number of variables (dimension ‘P’).

In these high dimensional settings, you cannot:
Look at each individual variable and make a decision (*t-tests*).
Choose amongst a small set of candidate models (*F-test*).
Plot every variable to look for interactions or transformations.

Paradigms

A big aspect of Big Data is 'pattern discovery' or 'data mining'
Good DM is about inferring useful signal at massive scale.

Our goal is to summarize really high dimensional data in such a way that you can relate it to structural models of interest.

$\Rightarrow$ Unsupervised learning. AKA: "exploratory data analysis"

We also want to predict! If things don't change too much...

$\Rightarrow$ Supervised learning. AKA: "predictive analytics"

Big Data is focused on actionable knowledge extraction from very large datasets (integral in business and industrial applications).

- Infer patterns from complex high dimensional data.
- Simplicity and scalability of algorithms is essential.
- We keep an eye on both *useful* and *true*.
- The end product is a *decision*.

Course Overview

subject to change...

[1] Data: Computing, plotting, and principles.

[2] Regression: A grand overview, linear and logistic.

[3] Model Selection: penalties, information criteria, cross-validation

[4] Classification: Multinomials, KNN, sensitivity/specificity

[5] Clustering: Mixture models, k-means, and association rules.

Midterm!

[6] Factors: Latent variables, PCA

[7] Trees: CART and random forests, ensembles

[8] Neural Networks: deep learning

[9] Text Data: topic models, sentiment prediction

We'll be working with real data analysis examples

- Mining client information: Who buys your stuff, what do they pay, what do they think of your new product?
- Online behavior tracking: Who is on what websites, what do they buy, how do/can we affect behavior?
- Collaborative filtering: predict preferences from people who do what you do; space-time recommender engines.
- Text mining: Connect blogs/emails/news to sentiment, beliefs, or intent. Parsing unstructured data.
- Big covariates: mining data to predict asset returns; using unstructured alternative data.

Many are applicable to marketing and finance, but we're far more general.

Data Visualization

Data visualization is an essential part of exploratory data analysis (EDA)

Some guidelines:

- Statistics: reducing dimension of your data to a few rich variables for comparison. Can be just picking two features to scatterplot, or can involve more complicated projections.
- Design: effective communication – with shapes, space, and color – for a given set of variable observations.
- Language: making it easy to move from Stats to Design.

They're all interconnected, but we'll focus on statistics.

Python

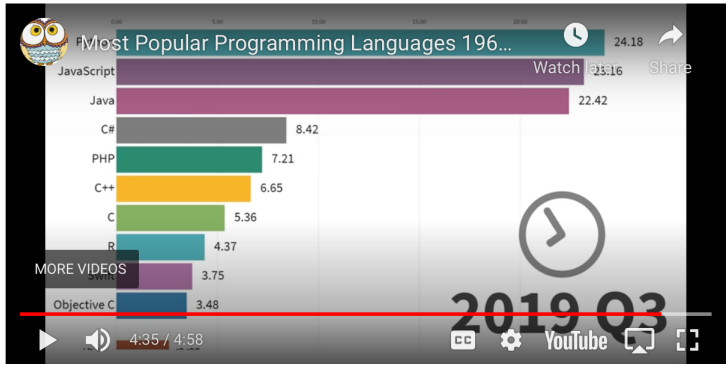
Python is a high-level, interpreted, general-purpose, object-oriented language.

- high-level: focus on readability
- interpreted: no need for a compiler
- general-purpose: website, database management, web scraping, gluing
- object-oriented: as opposed to procedural programing

Python ranks among the most popular and fastest-growing languages in the world.

```
In [1]: from IPython.display import YouTubeVideo
        YouTubeVideo("0g847HVwRSI", start=250, width=600)
```

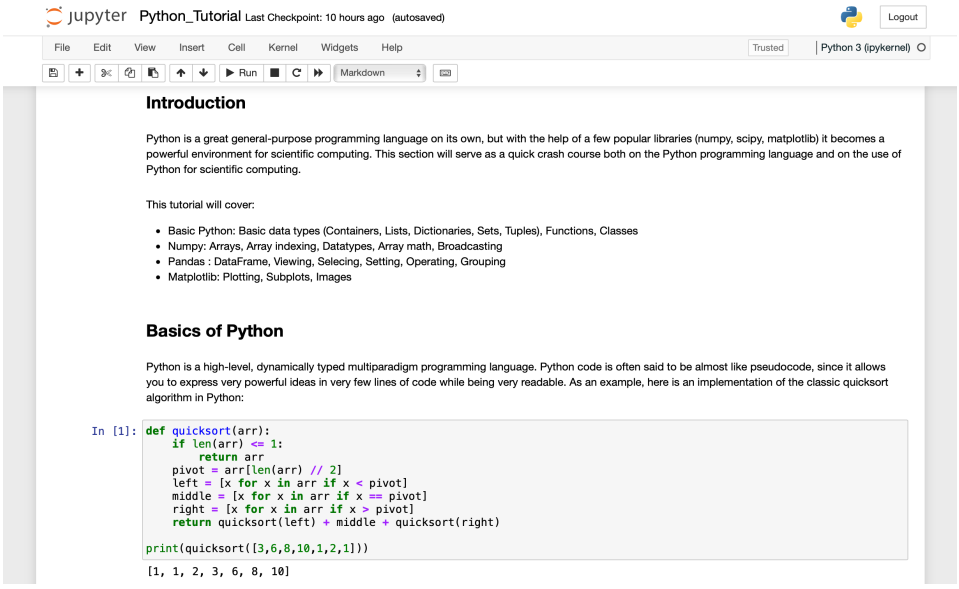
Out[1]:



R vs Python

- R is mainly used for statistical analysis while Python provides a more general approach to data science
- The primary objective of R is Data analysis and Statistics whereas the primary objective of Python is Deployment and Production
- R users mainly consists of Scholars and R&D professionals while Python users are mostly Programmers and Developers
- Both R and Python can handle huge size of database
- R can be used on the R Studio IDE while Python can be used on Spyder and Ipython Notebook IDEs
- R consists various packages and libraries like tidyverse, ggplot2, whereas Python consists packages and libraries like pandas, numpy, scikit-learn, TensorFlow, seaborn

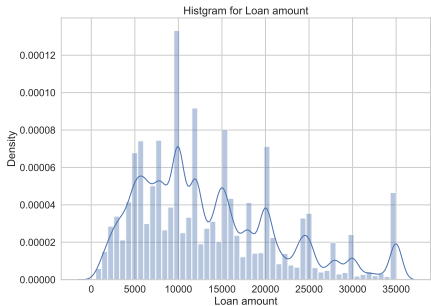
Jupyter Notebook

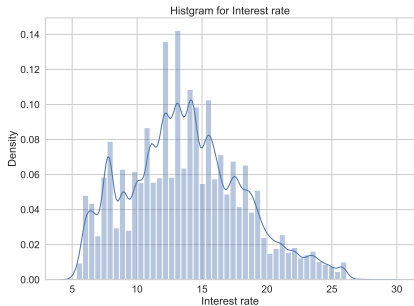


Jupyter Hub

- We use Jupyter Hub for this class:
  `https://jupyter-class.chicagobooth.edu/`
- Use your Cnet ID and password to log on
- Each of you has a folder. You can use shared data in this folder:
  `/classes/2080001_spr2022/` and run codes on the Booth server.
- You do not even need have Python installed on your personal computer.

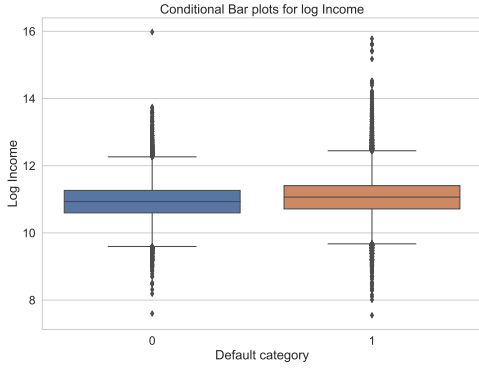
The simple histogram for continuous variables





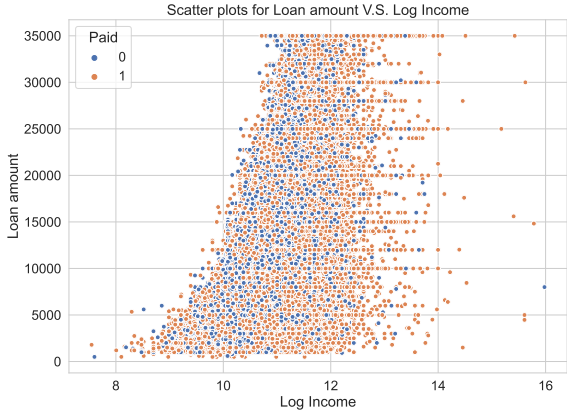
Data is binned and plotted bar height is the count in each bin.

Boxplots: summarizing conditional distributions



The box is the Interquartile Range (IQR; i.e., 25^{th} to 75^{th} %), with the median marked in the middle. The whiskers extend to the most extreme point which is no more than 1.5 times the IQR width from the box.

Use scatterplots to compare variables.



Scatterplots are a fundamental unit of statistics.

Scatterplot are more informative than correlations. The latter only reflect linear relationship and are easily influenced by outliers.

If you're able to find and compare meaningful low-dimensional summary statistics, then you are winning the DM game.

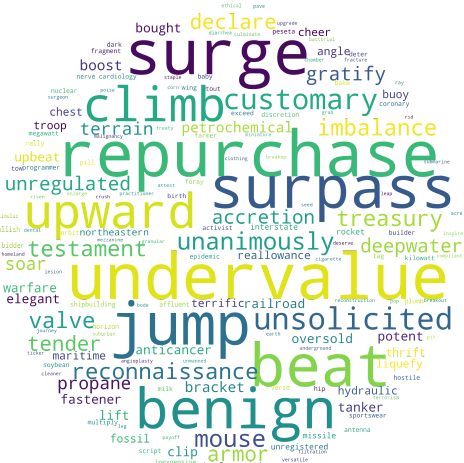
- Humans are good at comparing a few variables.
- If we can put it in a picture, we can build intuition.
- Prediction is easy in low dimensions.

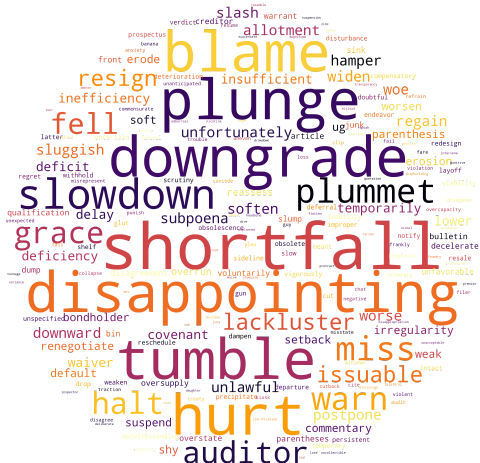
The key to good graphics is to reduce high-dimensional data to a few very informative summary statistics, then plot them.

We'll focus on info visualization throughout this course.

How to find important variables / features from many?

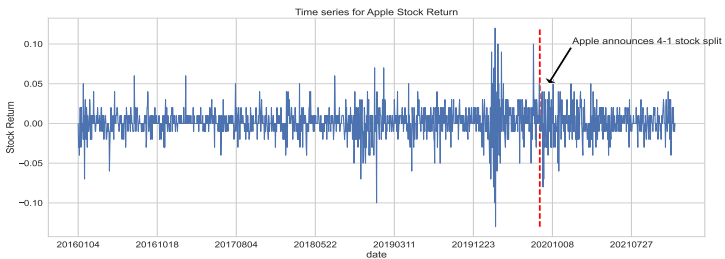
A simple strategy for **feature selection** is **correlation screening**.



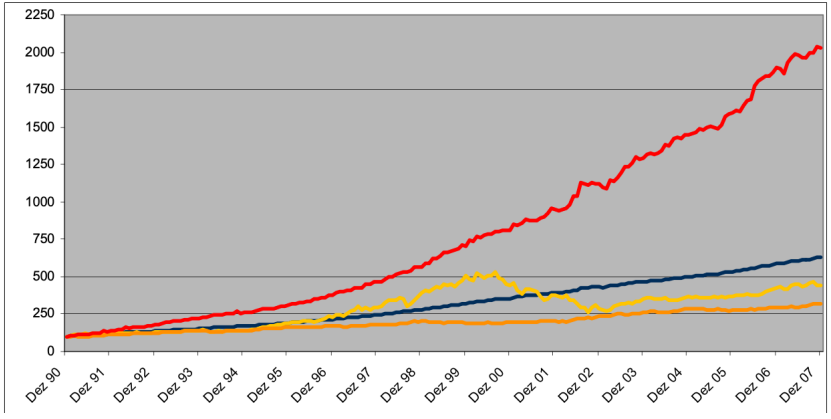


Source: "Predicting Returns with Text Data", Ke, Kelly, and Xiu (2021). 39,673 words, millions of articles.

Plot time series to show the trend of data.

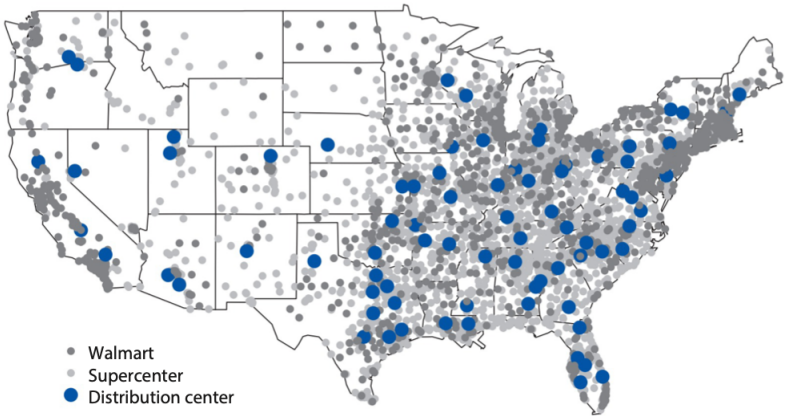


Would you like to invest in this new product (red)?



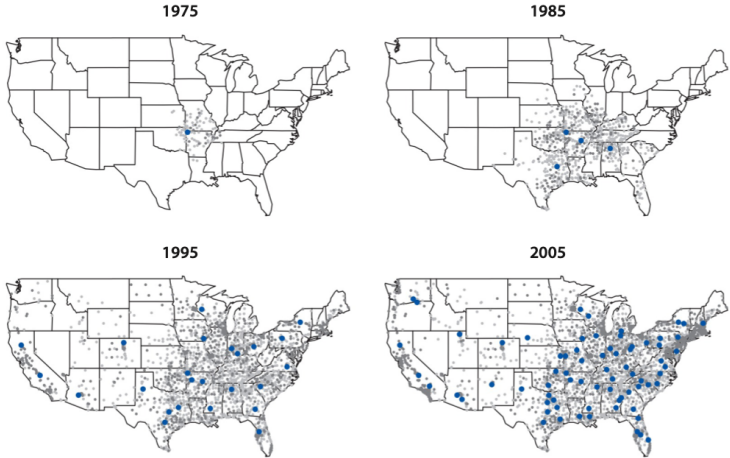
S&P 100 (yellow), Lehman Bond Index (orange), FS (blue), Simulated 3x FS (red)

Plot spatial data on a map.



Source: "Quantitative Social Science: An Introduction", Kosuke Imai

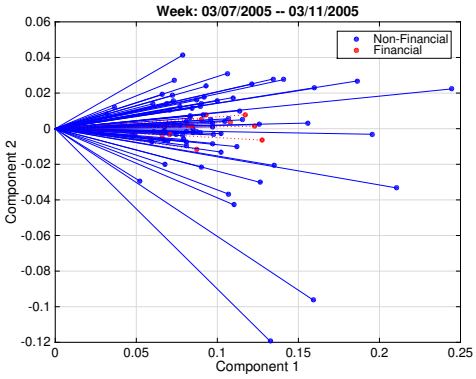
Visualizing spatial-temporal data already tells a story.

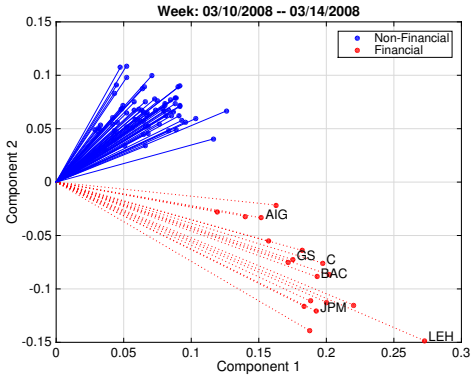


Source: "Quantitative Social Science: An Introduction", Kosuke Imai

A **Biplot** is constructed by the **singular value decomposition (SVD)** to obtain a low-rank approximation to the data

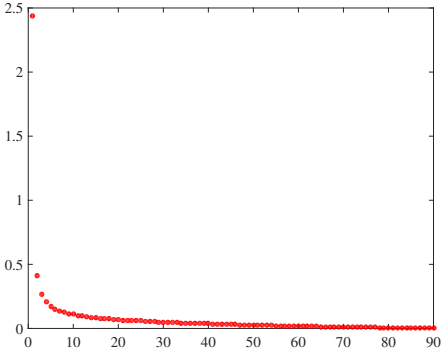
$$X \approx UDV^{\intercal}, \quad \text{where } X \text{ is } N \times P,\ U \text{ is } N \times 2,\ D \text{ is } 2 \times 2,\ V \text{ is } T \times 2.$$

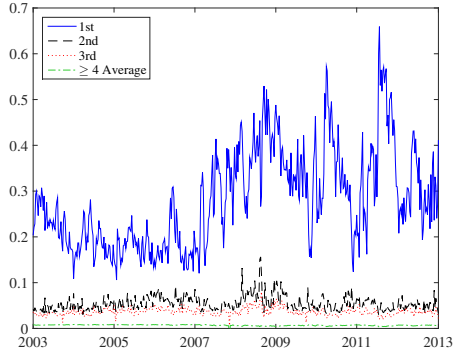




Source: "Principal Component Analysis of High-Frequency Data", Aït-Sahalia and Xiu (2019), JASA Vol 114, No 525, 287-303.

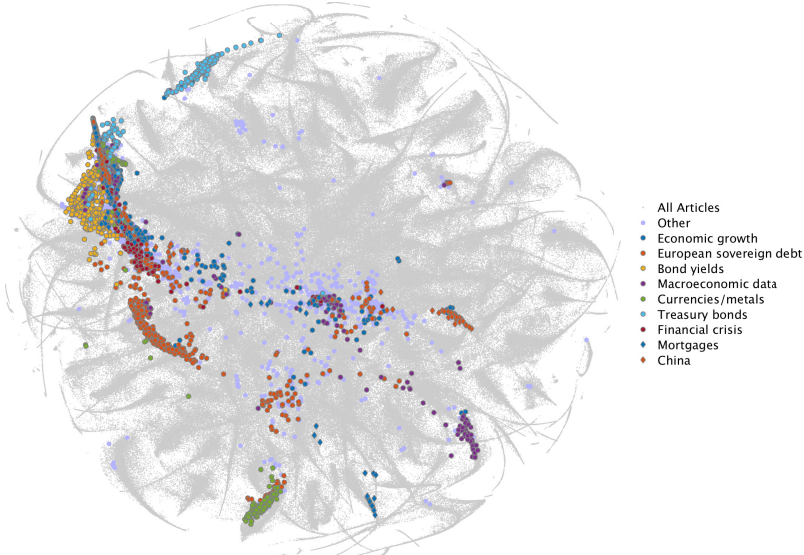
A **scree plot** is a graph of eigenvalues against the corresponding PC number.



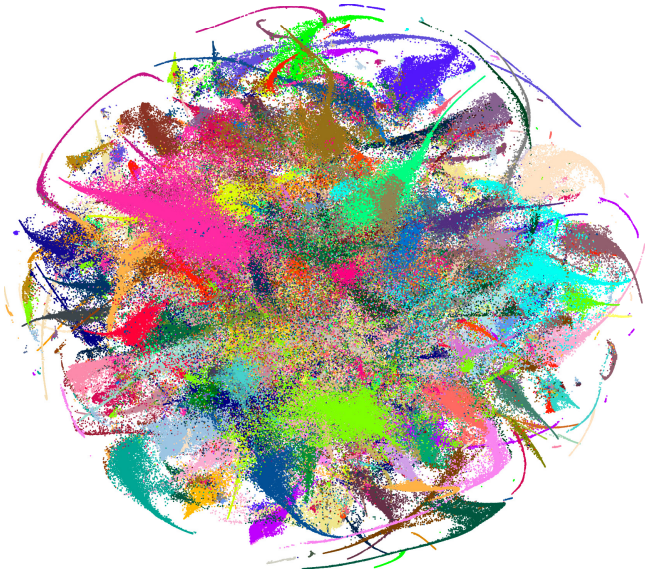


Source: "Principal Component Analysis of High-Frequency Data", Aït-Sahalia and Xiu (2019), JASA Vol 114, No 525, 287-303.

t-SNE visualizes **nonlinear dimension reduction** Ref: van der Maaten and Hinton (2008, JMLR).



Source: "Business News and Business Cycles", Bybee, Kelly, Manela, and Xiu (2021)

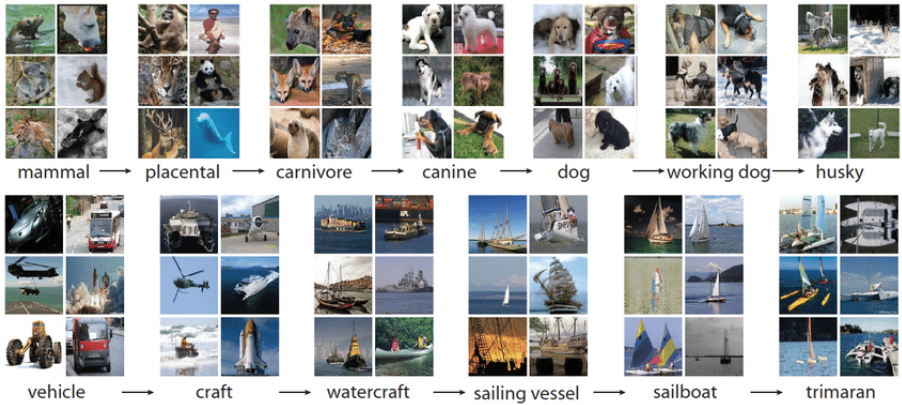


Source: "Business News and Business Cycles", Bybee, Kelly, Manela, and Xiu (2021)

Text as data


great
people
president
country
thank
big
time
border
news
fake
good
job
medium
trump
want
democrats
election
states
american
america
bad
like
u.s
year
state
work
crime
new
united
today
deal
win
china
vote
know
come
say
wall
take
republican
military
hillary
day
go
look
security
clinton
russia
campaign
democrat
honor
mexico
well
trade
house
fbi
billion
collusion
story
need
high
get
think
tax
far
leave
law
strong
korea
world
hunt
long
dollar
obama
happen
number
watch
history
report
cut
love
witch
talk
give
hard
pay
economy
north
mueller
way
fact
thing
help
total
money
massive
illegal
build

Visualizing Images



ImageNet (`https://image-net.org/`): The data that transformed AI research—and possibly the world