INSTITUTO DE
MATEMÁTICA
E ESTATÍSTICA
IME
50
UNIVERSIDADE
DE SÃO PAULO

MAC0459/MAC5865 - Tópicos em Ciência e Engenharia de Dados

Aula 02

Sejam bem-vindas, sejam bem-vindos!

Entre no link https://app.sli.do/event/kgwf8yjm ou e faça suas perguntas da aula.

R. Hirata Jr.

Objetivos de hoje

- Ao final da aula de hoje você deve:
  - Conhecer alguns datasets abertos
  - Conhecer algumas perguntas simples que podem ser feitas para os datasets
  - Conhecer algumas perguntas interessantes que podem ser feitas para os datasets
  - Conhecer os tipos de dados

Relembrando a semana passada

Simple pipeline – Scientific Method

1. Pose a question
2. Formulate a hypothesis
3. Formulate an experiment
4. Observe (data collecting)
5. Analyse the results
6. Go back to step 2 if the hypothesis is not correct/supported
7. Report results

Simple pipeline – Engineering Method

1. Define a problem
2. Specify requirements
3. Brainstorm, evaluate, choose solution
4. Develop a prototype solution
5. Tests solution
6. Go back to step 3 if the results, or data, etc in case the solution does not meet requirements

Simple pipeline – Data Science Method

1. Pose question
2. Get the data
3. Explore the data
4. Model the data
5. Report results

Repeatability vs Reproducibility

- Repeatability
    - variability caused by the measurement device
- Reproducibility
    - variability caused by different labs/operators

CS, DS, Real Science

- “Computer scientists, by nature, don’t respect data”
- Real scientists will only use as many significant digits as the worst precision of any measurement in the process.

Hypothesis vs data driven science

- HD science - ask specific questions of the world and then generating the specific data to confirm, or deny it
- Data driven science - a new paradigm to model the world

Escreva sua pergunta aqui: https://app.sli.do/event/kgwf8yjm

Learning to ask questions

- Computer scientists students are not used to ask questions, why?
- Good data scientists develop an inherent curiosity about the world around them and have wide-ranging interests.

What is Science?

- “We absolutely must leave room for doubt or there is no progress and there is no learning. There is no learning without having to pose a question. And a question requires doubt. People search for certainty. But there is no certainty. People are terrified — how can you live and not know? It is not odd at all. You only think you know, as a matter of fact. And most of your actions are based on incomplete knowledge and you really don’t know what it is all about, or what the purpose of the world is, or know a great deal of other things. It is possible to live and not know.” Feynman

What is Science?

- In our example, did we left room for doubt?
- The lemma of our disciple should be:

De omnibus dubitandum

Pronto, pode acordar!

Você está presente?

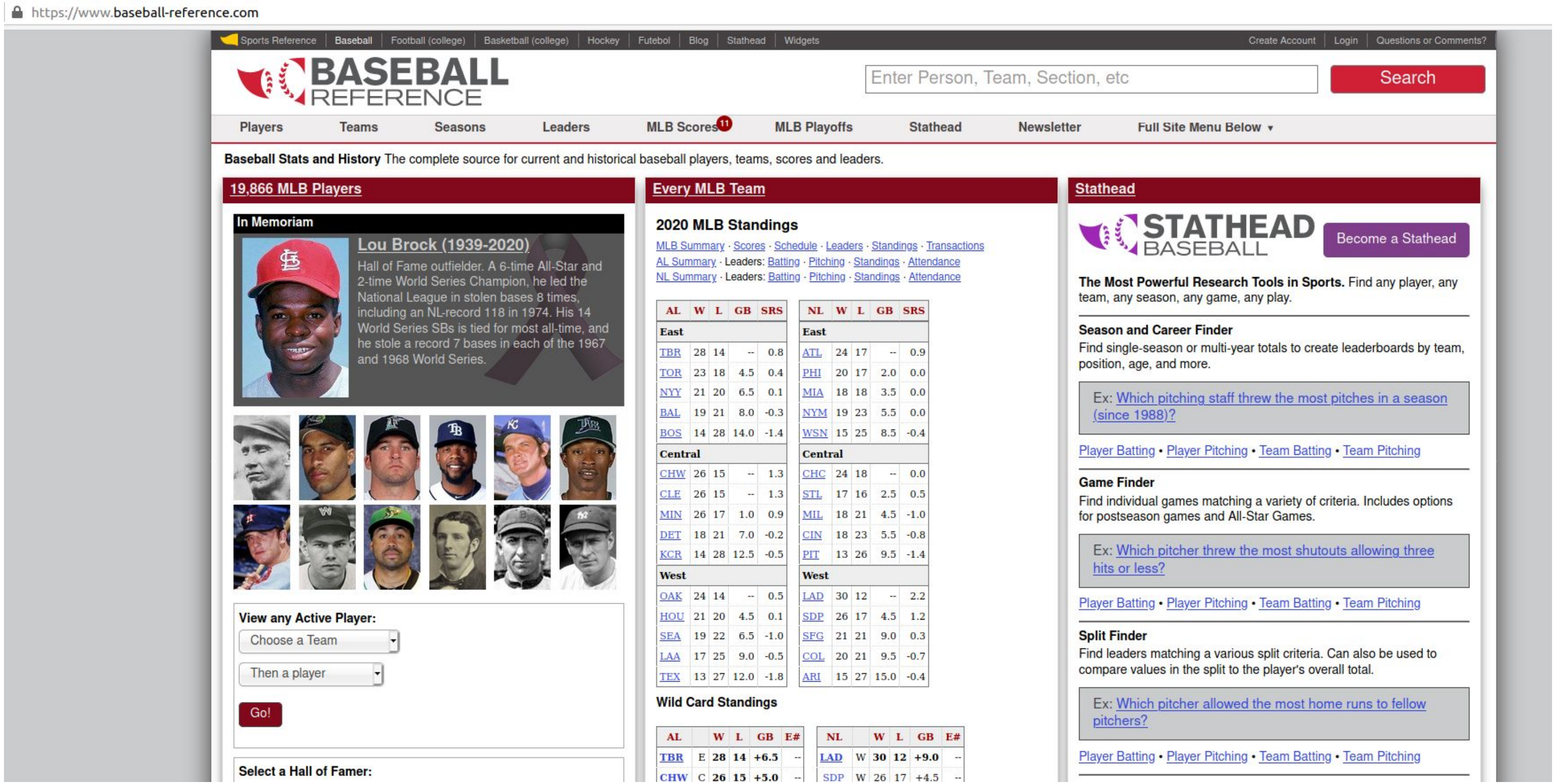
Learning to ask questions

- The baseball encyclopedia
- The Internet Movie Database (IMDb)
- Google Ngrams
- New York Taxi Records

Escreva sua pergunta aqui: https://app.sli.do/event/kgwf8yjm

Learning to ask questions

- The baseball encyclopedia



Baseball Stats and History The complete source for current and historical baseball players, teams, scores and leaders.

19,866 MLB Players

In Memoriam

Lou Brock (1939-2020)

Hall of Fame outfielder. A 6-time All-Star and 2-time World Series Champion, he led the National League in stolen bases 8 times, including an NL-record 118 in 1974. His 14 World Series SBs is tied for most all-time, and he stole a record 7 bases in each of the 1967 and 1968 World Series.

View any Active Player:

Choose a Team ▾

Then a player ▾

Go!

Select a Hall of Famer:

Every MLB Team

2020 MLB Standings

MLB Summary · Scores · Schedule · Leaders · Standings · Transactions
AL Summary · Leaders: Batting · Pitching · Standings · Attendance
NL Summary · Leaders: Batting · Pitching · Standings · Attendance

AL	W	L	GB	SRS
East				
TBR	28	14	--	0.8
TOR	23	18	4.5	0.4
NYY	21	20	6.5	0.1
BAL	19	21	8.0	-0.3
BOS	14	28	14.0	-1.4
Central				
CHW	26	15	--	1.3
CLE	26	15	--	1.3
MIN	26	17	1.0	0.9
DET	18	21	7.0	-0.2
KCR	14	28	12.5	-0.5
West				
OAK	24	14	--	0.5
HOU	21	20	4.5	0.1
SEA	19	22	6.5	-1.0
LAA	17	25	9.0	-0.5
TEX	13	27	12.0	-1.8

NL	W	L	GB	SRS
East				
ATL	24	17	--	0.9
PHI	20	17	2.0	0.0
MIA	18	18	3.5	0.0
NYM	19	23	5.5	0.0
WSN	15	25	8.5	-0.4
Central				
CHC	24	18	--	0.0
STL	17	16	2.5	0.5
MIL	18	21	4.5	-1.0
CIN	18	23	5.5	-0.8
PIT	13	26	9.5	-1.4
West				
LAD	30	12	--	2.2
SDP	26	17	4.5	1.2
SFG	21	21	9.0	0.3
COL	20	21	9.5	-0.7
ARI	15	27	15.0	-0.4

Wild Card Standings

AL		W	L	GB	E#
TBR	E	**28**	**14**	**+6.5**	--
CHW	C	**26**	**15**	**+5.0**	--

NL		W	L	GB	E#
LAD	W	**30**	**12**	**+9.0**	--
SDP	W	26	17	+4.5	--

Stathead

STATHEAD BASEBALL

Become a Stathead

The Most Powerful Research Tools in Sports. Find any player, any team, any season, any game, any play.

Season and Career Finder

Find single-season or multi-year totals to create leaderboards by team, position, age, and more.

Ex: Which pitching staff threw the most pitches in a season (since 1988)?

Player Batting • Player Pitching • Team Batting • Team Pitching

Game Finder

Find individual games matching a variety of criteria. Includes options for postseason games and All-Star Games.

Ex: Which pitcher threw the most shutouts allowing three hits or less?

Player Batting • Player Pitching • Team Batting • Team Pitching

Split Finder

Find leaders matching a various split criteria. Can also be used to compare values in the split to the player's overall total.

Ex: Which pitcher allowed the most home runs to fellow pitchers?

Player Batting • Player Pitching • Team Batting • Team Pitching

The baseball encyclopedia

- Obvious questions:
  - How can we best measure an individual player's skill or value?
  - How fairly do trades between teams generally work out?
  - What is the general trajectory of player's performance level as they mature and age?

The baseball encyclopedia

- Demographic and Social issues
  - Do left-handed people have shorter lifespans than right-handers?
  - How often do people return to live in the same place where were born?
  - Do player salaries generally reflect past, present or future performance?
  - To what extent have heights and weights been increasing in the population at large?

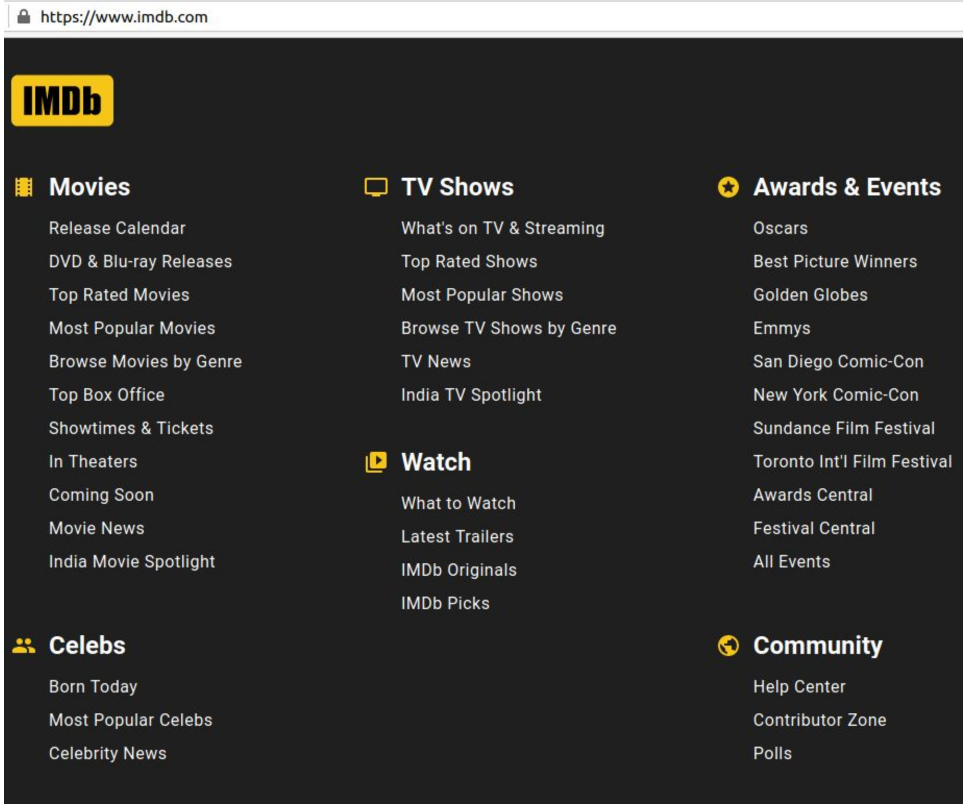
The baseball encyclopedia

- Identifiers and reference tags (metadata) are as interesting, or more, than the statistical records.
- Statistical proxy: use data you have to substitute for the one you really want.
- The data set of your dreams likely does not exist.
- A good data scientist is a pragmatist!

Escreva sua pergunta aqui: https://app.sli.do/event/kgwf8yjm

Learning to ask questions

- The Internet Movie Database (IMDb)



The Internet Movie Database (IMDb)

- Some obvious questions:
  - Which actors appeared in most films? Earned the most money? Appeared in the lowest rated films? Had the longest career or the shortest lifespan?
  - What was the highest rated film each year, or the best in each genre? Which movies lost the most money, had the highest-powered casts, or got the least favorable reviews.

The Internet Movie Database (IMDb)

- Larger scale questions
  - How well does movie gross correlate with viewer ratings and awards? Do customers instinctively flock to trash, or is virtue on the part of the creative team properly rewarded?
  - How do Hollywood movies compare to Bollywood movie, in terms of ratings, budget, and gross? Are American movies better received than foreign films, and how does this differ between between US and non-US reviewers?

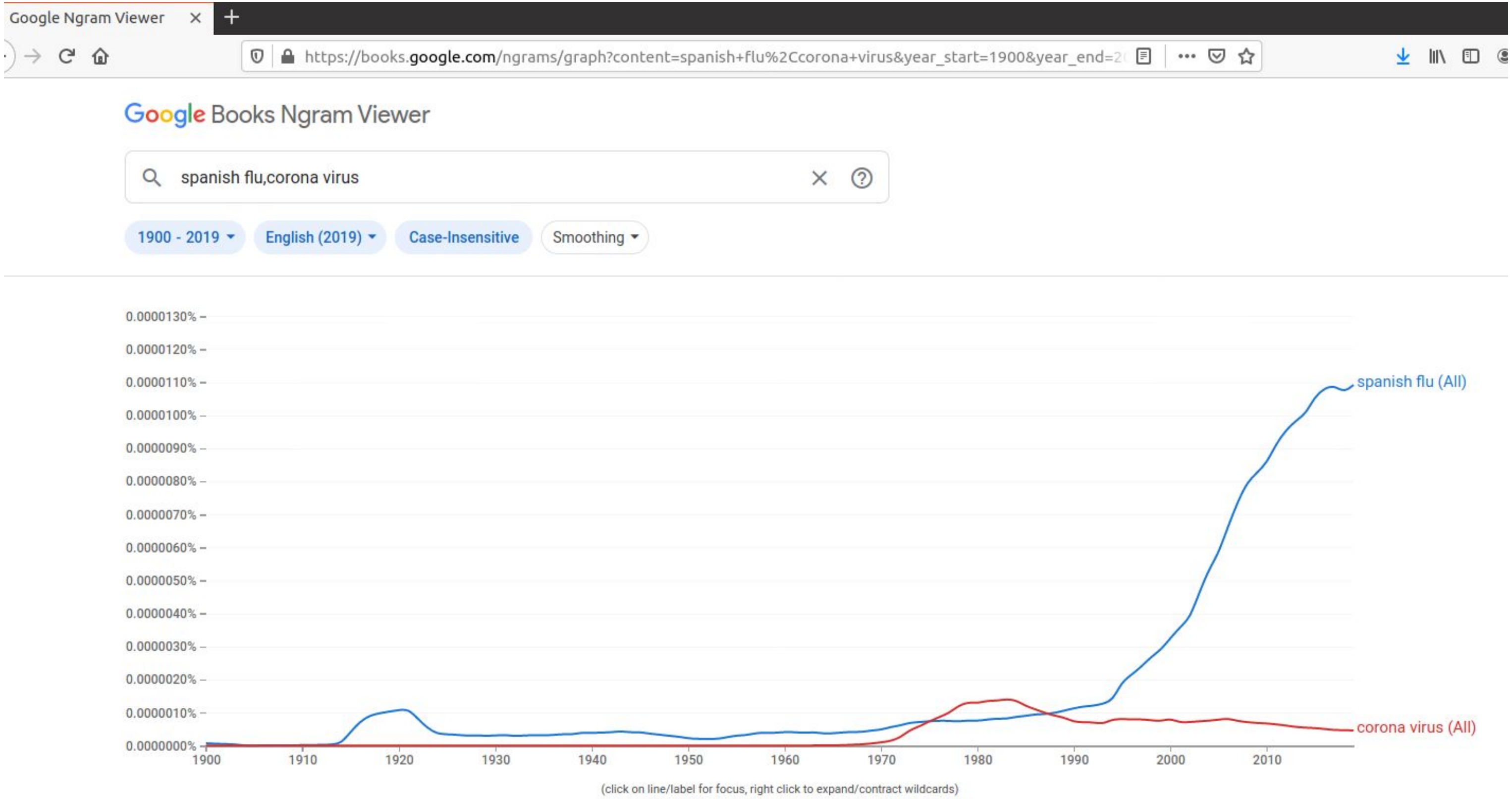
The Internet Movie Database (IMDb)

- Larger scale questions
  - What is the age distribution of actors and actresses in films? How much younger is the actress playing the wife, on average, than the actor playing the husband? Has this disparity been increasing or decreasing with time?
  - Live fast, die young, and leave a good-looking corpse? Do movie stars live longer or shorter lives than bit players, or compared to the general public?

Escreva sua pergunta aqui: https://app.sli.do/event/kgwf8yjm

Learning to ask questions

- Google Ngram Viewer



Google Ngram Viewer

https://books.google.com/ngrams/graph?content=spanish+flu%2Ccorona+virus&year_start=1900&year_end=2

Google Books Ngram Viewer

spanish flu,corona virus

1900 - 2019 · English (2019) · Case-Insensitive · Smoothing

(click on line/label for focus, right click to expand/contract wildcards)

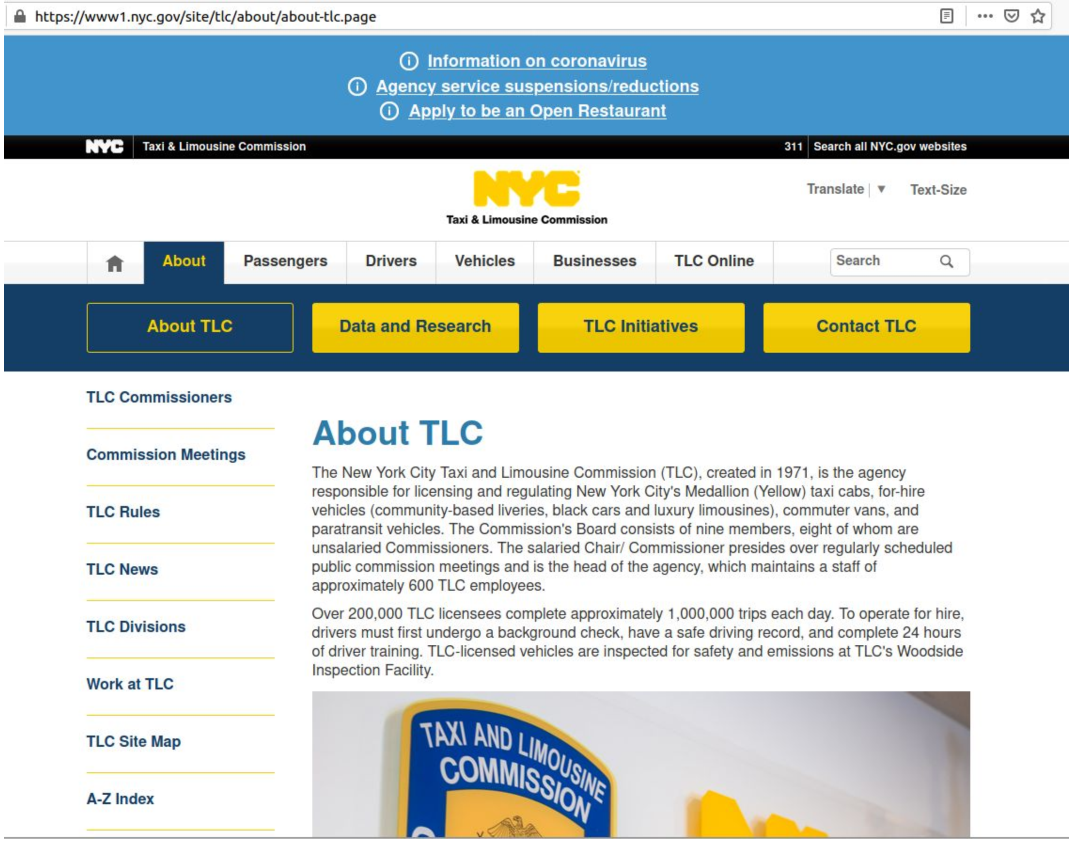
Google Ngram Viewer

- Some interesting questions:
  - How often new words emerge and get popular? Do these words tend to stay in common usage, or rapidly fade away? Can we detect when words change meaning over time? Like the transition of gay from happy to

Escreva sua pergunta aqui: https://app.sli.do/event/kgwf8yjm

Learning to ask questions

New York Taxi Records



https://www1.nyc.gov/site/tlc/about/about-tlc.page

Information on coronavirus
Agency service suspensions/reductions
Apply to be an Open Restaurant

About TLC

The New York City Taxi and Limousine Commission (TLC), created in 1971, is the agency responsible for licensing and regulating New York City's Medallion (Yellow) taxi cabs, for-hire vehicles (community-based liveries, black cars and luxury limousines), commuter vans, and paratransit vehicles. The Commission's Board consists of nine members, eight of whom are unsalaried Commissioners. The salaried Chair/ Commissioner presides over regularly scheduled public commission meetings and is the head of the agency, which maintains a staff of approximately 600 TLC employees.

Over 200,000 TLC licensees complete approximately 1,000,000 trips each day. To operate for hire, drivers must first undergo a background check, have a safe driving record, and complete 24 hours of driver training. TLC-licensed vehicles are inspected for safety and emissions at TLC's Woodside Inspection Facility.

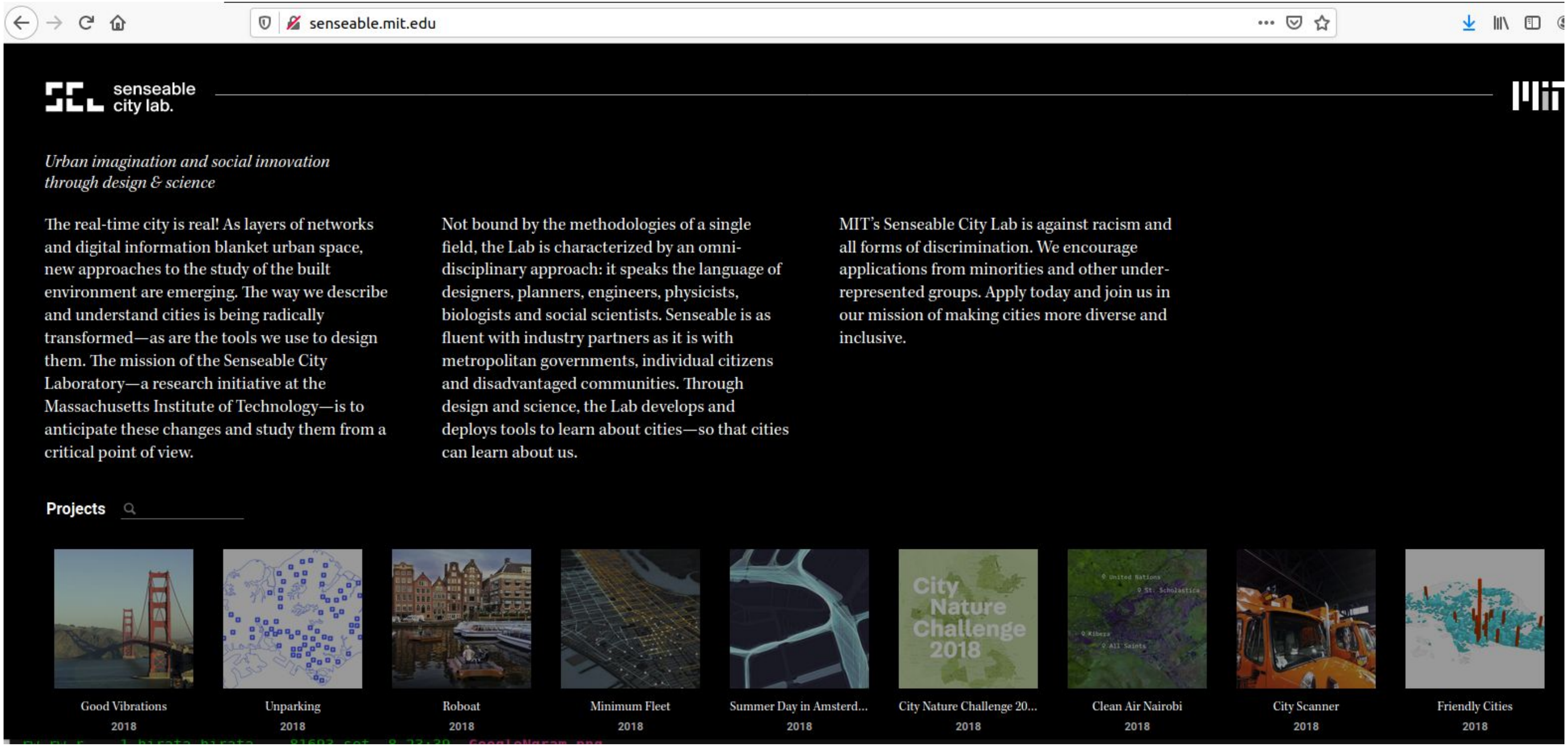
Escreva sua pergunta aqui: https://app.sli.do/event/kgwf8yjm

New York Taxi Records

- Some interesting questions:
  - How much money do drivers make each night, on average? What is the distribution? Do drivers make more money on sunny days, or rainy days?
  - Where are the best spots in the city for drivers to cruise, in order to pick up profitable fares? How does this vary at different times of the day?
  - How much are drivers tipped, and why? Do faster drivers get tipped

Escreva sua pergunta aqui: https://app.sli.do/event/kgwf8yjm

New York Taxi Records



senseable.mit.edu

senseable city lab.

Urban imagination and social innovation through design & science

The real-time city is real! As layers of networks and digital information blanket urban space, new approaches to the study of the built environment are emerging. The way we describe and understand cities is being radically transformed—as are the tools we use to design them. The mission of the Senseable City Laboratory—a research initiative at the Massachusetts Institute of Technology—is to anticipate these changes and study them from a critical point of view.

Not bound by the methodologies of a single field, the Lab is characterized by an omni-disciplinary approach: it speaks the language of designers, planners, engineers, physicists, biologists and social scientists. Senseable is as fluent with industry partners as it is with metropolitan governments, individual citizens and disadvantaged communities. Through design and science, the Lab develops and deploys tools to learn about cities—so that cities can learn about us.

MIT's Senseable City Lab is against racism and all forms of discrimination. We encourage applications from minorities and other under-represented groups. Apply today and join us in our mission of making cities more diverse and inclusive.

Projects

Good Vibrations 2018

Unparking 2018

Roboat 2018

Minimum Fleet 2018

Summer Day in Amsterd... 2018

City Nature Challenge 20... 2018

Clean Air Nairobi 2018

City Scanner 2018

Friendly Cities 2018

```
rw-rw-r--   1 hirata hirata    81693 set  8 23:39  GoogleNgram.png
```

Obrigado!