Data Science as a Team Sport

Gabriela de Queiroz
@gdequeiroz | linktr.ee/gdq

slides: **bit.ly/eusr20**

Image: www.freepik.com/vectors/people

Hi, I'm Gabriela de Queiroz

Sr. Machine Learning Manager, IBM

- Founder of **R-Ladies** (rladies.org)
- Founder of **AI Inclusive** (ai-inclusive.org)
- Member of the **R Foundation** (r-project.org)

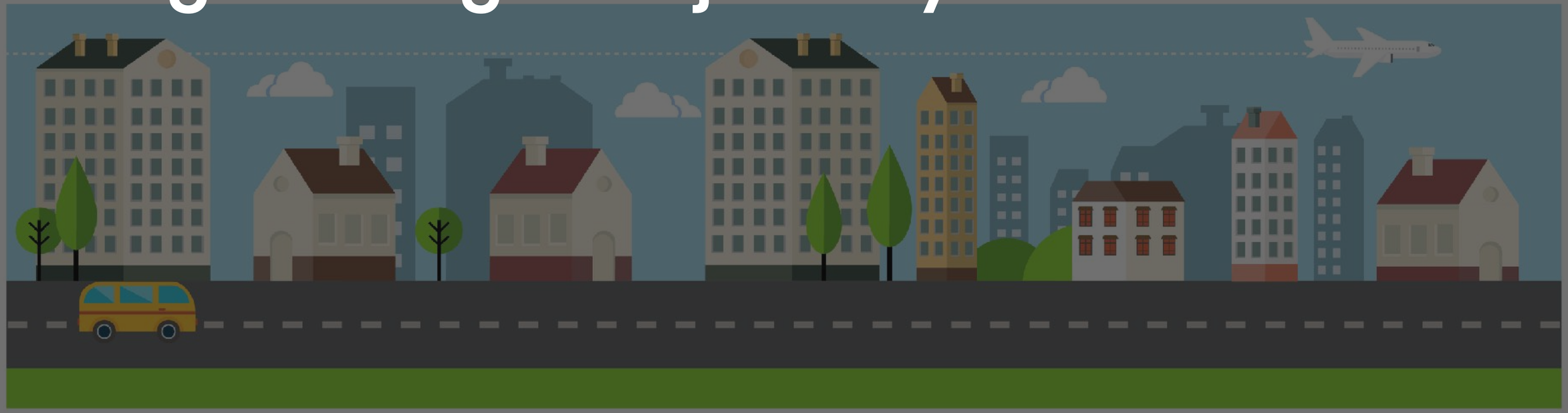
- B.S. in Statistics
- MSc. in Epidemiology
- MSc. in Statistics

Data Scientist + Developer Advocate + Open Source Developer + Manager +

Statistician + Epidemiologist + Community Builder + Mentor + Speaker + Educator

slides: bit.ly/eusr20

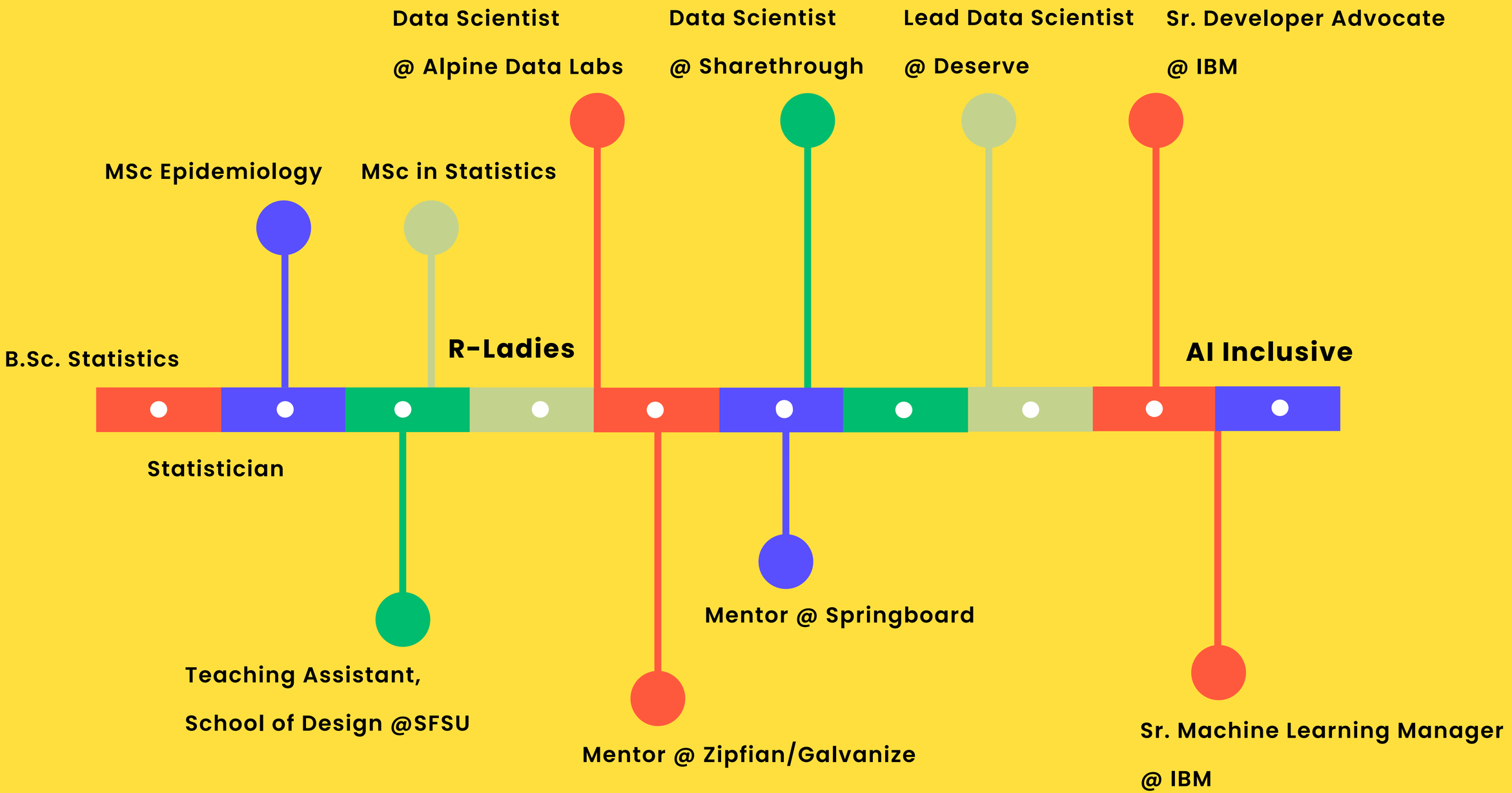
Let’s go through the journey



Population: ~200 million

Official Language: Portuguese

Population: ~6 million


Data Scientist
@ Alpine Data Labs
Data Scientist
@ Sharethrough
Lead Data Scientist
@ Deserve
Sr. Developer Advocate
@ IBM
MSc Epidemiology
MSc in Statistics
B.Sc. Statistics
R-Ladies
AI Inclusive
Statistician
Teaching Assistant,
School of Design @SFSU
Mentor @ Springboard
Mentor @ Zipfian/Galvanize
Sr. Machine Learning Manager
@ IBM

R- ladies

It was founded in October 2012.

The idea was to give back to the community and create a place where people would feel comfortable, safe and welcome.

A place where people could ask questions, learn together and share.

31 OCT

Wednesday, October 31, 2012

Introduction to R (beginners and pre-beginners)

Hosted by
Gabriela de Queiroz

Details

Hello R-ladies!

The first meetup will take place on October 31st at the Google office in San Francisco.

For this first meetup, we'll do an introduction to R. We'll go over the following topics:

installing R setting up an R environment (RStudio) basic commands (open files, simple dataset manipulation, simple plots, etc) loading packages the help function and how to read its output

All you need is your laptop and charger.

We look forward to seeing you!

R-Ladies

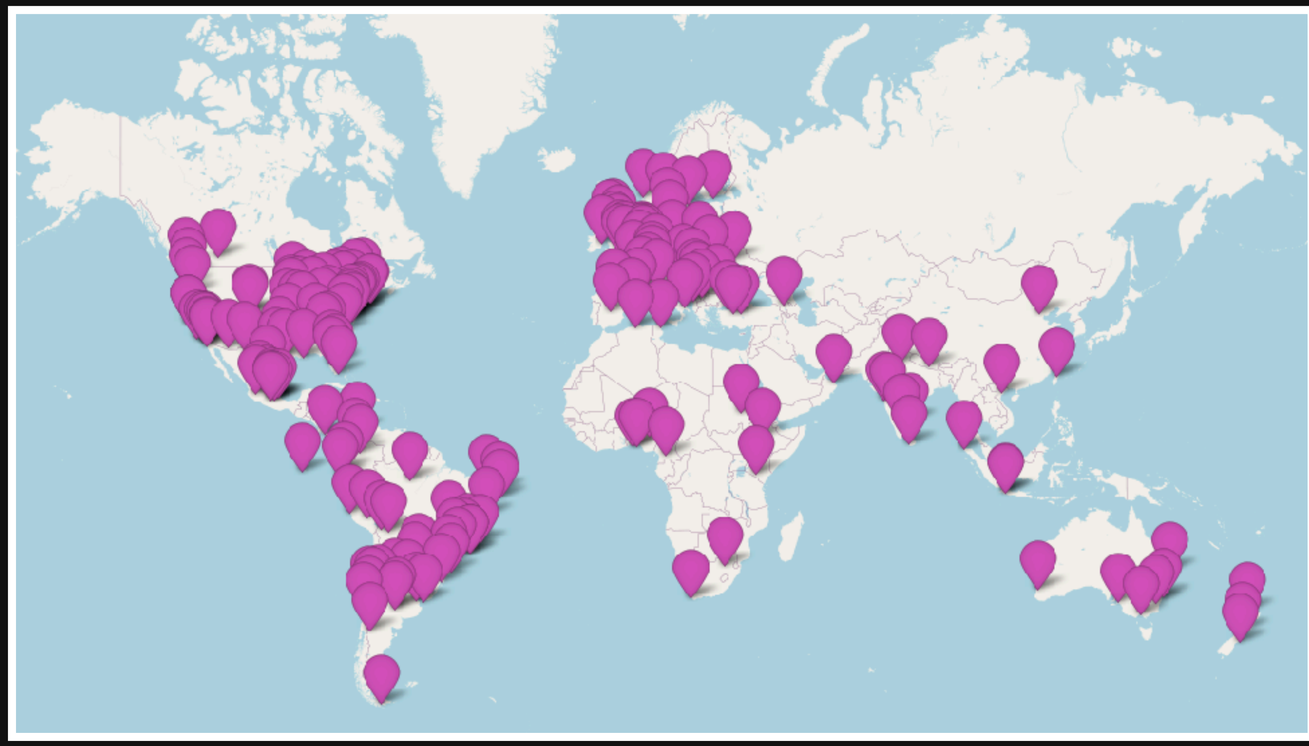
rladies.org

Worldwide organization that promotes diversity in the R community via meetups and mentorship in a friendly and safe environment

54
R-Ladies Countries

193
R-Ladies Cities

71618
R-Ladies members on meetup.com



AI Inclusive

> Together, we are building a community to make **AI** more **inclusive** to everyone.

Mission: Increase the representation and participation of minority groups in Artificial Intelligence

- Website: ai-inclusive.org
- Twitter: bit.ly/ai-inclusive-twitter
- Facebook: bit.ly/ai-inclusive-facebook
- Instagram: bit.ly/ai-inclusive-instagram
- Youtube: bit.ly/ai-inclusive-youtube

If you want to start a chapter, send us an email: info@ai-inclusive.org

Researcher/Statistician

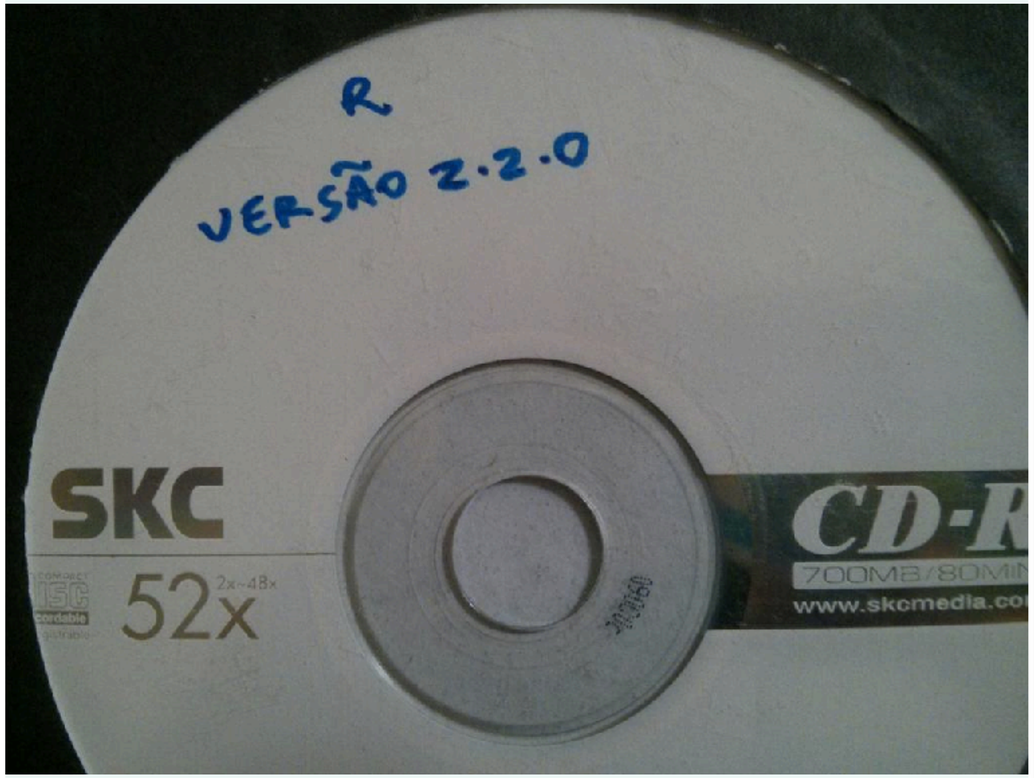
Undergraduate

- STATE UNIVERSITY
- PUBLIC AND FREE UNIVERSITY
- **BACHELOR IN STATISTICS**

Grad School

- SCIENTIFIC INSTITUTION FOR RESEARCH
- PUBLIC AND FREE
- **MSC. IN EPIDEMIOLOGY**

❤️ First exposure to R ❤️



R 2.2.0 (October, 2005)

R 3.1.0 (April, 2014)
R 3.0.3 (March, 2014)
R 3.0.2 (September, 2013)
R 3.0.1 (May, 2013)
R 3.0.0 (April, 2013)
R 2.15.3 (March, 2013)
R 2.15.2 (October, 2012)
R 2.15.1 (June, 2012)
R 2.15.0 (March, 2012)
R 2.14.2 (February, 2012)
R 2.14.1 (December, 2011)
R 2.14.0 (November, 2011)
R 2.13.2 (September, 2011)
R 2.13.1 (July, 2011)
R 2.13.0 (April, 2011)
R 2.12.2 (February, 2011)
R 2.12.1 (December, 2010)
R 2.12.0 (October, 2010)
R 2.11.1 (May, 2010)
R 2.11.0 (April, 2010)
R 2.10.1 (December, 2009)
R 2.10.0 (October, 2009)
R 2.9.2 (August, 2009)
R 2.9.1 (June, 2009)
R 2.9.0 (April, 2009)
R 2.8.1 (December, 2008)
R 2.8.0 (October, 2008)
R 2.7.2 (August, 2008)
R 2.7.1 (June, 2008)
R 2.7.0 (April, 2008)
R 2.6.2 (February, 2008)
R 2.6.1 (November, 2007)
R 2.6.0 (October, 2007)
R 2.5.1 (July, 2007)
R 2.5.0 (April, 2007)
R 2.4.1 (December, 2006)
R 2.4.0 (October, 2006)
R 2.3.1 (June, 2006)
R 2.3.0 (April, 2006)
R 2.2.1 (December, 2005)
R 2.2.0 (October, 2005)

R 4.0.3 (October, 2020)
R 4.0.2 (June, 2020)
R 4.0.1 (June, 2020)
R 4.0.0 (April, 2020)
R 3.6.3 (February, 2020)
R 3.6.2 (December, 2019)
R 3.6.1 (July, 2019)
R 3.6.0 (April, 2019)
R 3.5.3 (March, 2019)
R 3.5.2 (December, 2018)
R 3.5.1 (July, 2018)
R 3.5.0 (April, 2018)
R 3.4.4 (March, 2018)
R 3.4.3 (November, 2017)
R 3.4.2 (September, 2017)
R 3.4.1 (June, 2017)
R 3.4.0 (April, 2017)
R 3.3.3 (March, 2017)
R 3.3.2 (October, 2016)
R 3.3.1 (June, 2016)
R 3.3.0 (April, 2016)
R 3.2.5 (April, 2016)
R 3.2.4 (March, 2016)
R 3.2.3 (December, 2015)
R 3.2.2 (August, 2015)
R 3.2.1 (June, 2015)
R 3.2.0 (April, 2015)
R 3.1.3 (March, 2015)
R 3.1.2 (October, 2014)
R 3.1.1 (July, 2014)

The R Inferno

Patrick Burns[1]

30th April 2011

Contents

List of Figures

List of Tables

1 Falling into the Floating Point Trap

2 Growing Objects

3 Failing to Vectorize
3.1 Subscripting
3.2 Vectorized if
3.3 Vectorization impossible

4 Over-Vectorizing

g Functions
ction
city
tency

bal Assignments

n Object Orientation
hods
generic functions
methods
inheritance
hods

An Introduction to R

Notes on R: A Programming Environment for Data Analysis and Graphics
Version 2.5.1 (2007-06-27)

W. N. Venables, D. M. Smith
and the R Development Core Team

1 Introduction and preliminaries
1.1 The R environment
1.2 Related software and documentation
1.3 R and statistics
1.4 R and the window system
1.5 Using R interactively
1.6 An introductory session
1.7 Getting help with functions and features
1.8 R commands, case sensitivity, etc.
1.9 Recall and correction of previous commands
1.10 Executing commands from or diverting output to a file
1.11 Data permanency and removing objects

2 Simple manipulations; numbers and vectors
2.1 Vectors and assignment
2.2 Vector arithmetic
2.3 Generating regular sequences
2.4 Logical vectors
2.5 Missing values
2.6 Character vectors
2.7 Index vectors; selecting and modifying subsets of a data set
2.8 Other types of objects

3 Objects, their modes and attributes
3.1 Intrinsic attributes: mode and length
3.2 Changing the length of an object
3.3 Getting and setting attributes
3.4 The class of an object

4 Ordered and unordered factors

HEI

HEALTH EFFECTS INSTITUTE

STATEMENT

Synopsis of Research Report 171

Multicity Study of Air Pollution and Mortality in Latin America (the ESCALA Study)

BACKGROUND

For nearly two decades, scientists seeking to understand the role that air pollution might play in population health effects have relied heavily on epidemiologic studies known as time-series studies. Time-series studies use information on daily changes in air pollutant concentrations and daily counts of mortality and morbidity. Although initially conducted at the individual city level, coordinated analyses across many cities have recently emerged as the tool of choice for developing more reliable and comparable estimates of the short-term effects of air pollution on health in regions around the world. HEI has a long-standing interest in these coordinated analyses; it has funded studies such as the National Morbidity, Mortality, and Air Pollution Study; Air Pollution and Health: A European and North American Approach; and Public Health and Air Pollution in Asia.

The present study, referred to hereafter by its Spanish acronym ESCALA (Estudio de Salud y Contaminación del Aire en Latinoamérica), was initiated to address underlying data and methodologic limitations in the epidemiologic literature on the health effects of air pollution in Latin America that had been identified in a 2005 review by the Pan American Health Organization. The William and Flora Hewlett Foundation, which has a strong interest in understanding air pollution and health in Latin America, provided HEI with supplemental support to address gaps in the evidence necessary to inform regulatory decisions, and in the process to build a network of health experts capable of carrying out research on air pollution in the future. The multicenter study was led by Dr. Isabelle Romieu, then at the Instituto Nacional de Salud Pública in México, in collaboration with Dr. Nelson Gouveia in Brazil and Dr. Luis Cifuentes in Chile.

APPROACH

The primary objective of the ESCALA study was to estimate the effect of daily exposures to PM_{10} (particulate matter ≤ 10 µm in aerodynamic diameter) and to ozone on daily mortality from several causes (all natural causes, cardiopulmonary disease, respiratory disease, cardiovascular disease, cerebrovascular–stroke, and chronic obstructive pulmonary disease) and for several age groups (all-age, ≥ 65 years, < 1 year, 1–4 years, 1–14 years) in nine Latin American cities, and for the region as a whole, using a common analytic framework. The nine cities were Monterrey, Toluca, and Mexico City in México; Rio de Janeiro, São Paulo, and Porto Alegre in Brazil; and Santiago, Concepción, and Temuco in Chile. Of these, three cities (Porto Alegre, Concepción, and Temuco) were excluded from the ozone analyses because of the lack of adequate ozone monitoring data.

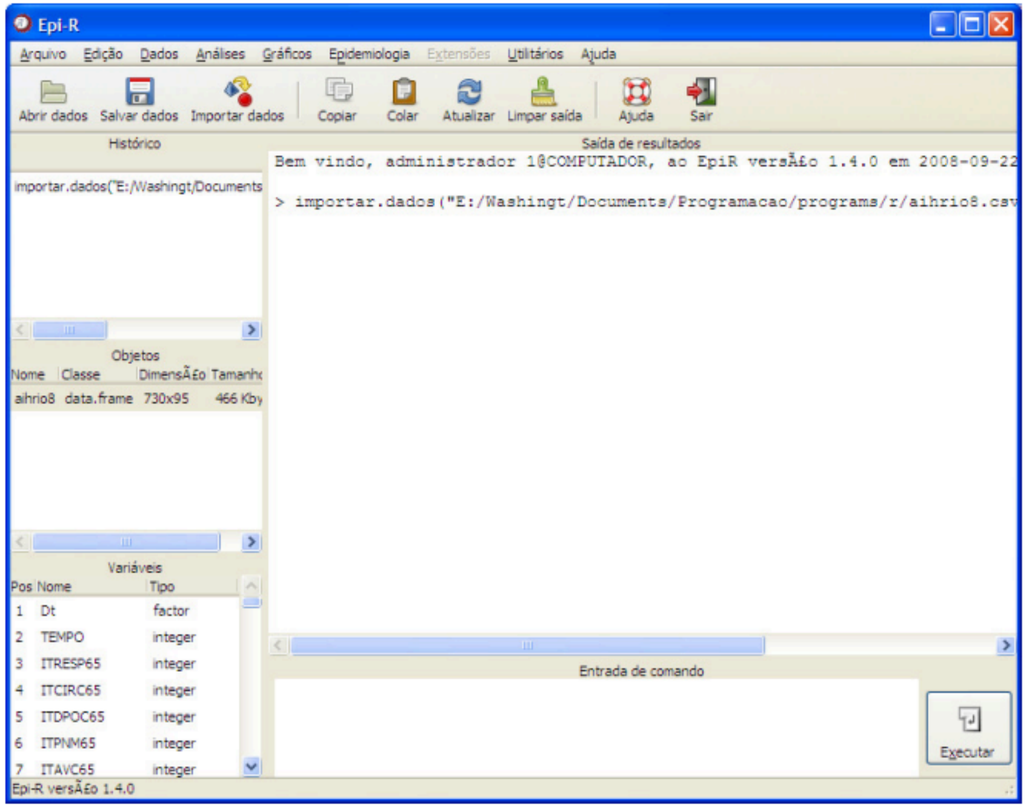
In the first stage of the analyses, the investigators estimated the percentage change in the risk of mortality per 10-µg/m^3 increase in PM_{10} or ozone for each combination of age group and cause of death for the individual cities in each country. They followed a common protocol for fitting the widely used Poisson regression models to the air pollution and mortality time-series data in each city while controlling for other factors that might also explain the temporal patterns of mortality (e.g., temperature, humidity, season, day-of-the-week, holidays). The investigators also carried out analyses to test the sensitivity of the results to various details of the models. Ultimately, the final models used in the individual cities were chosen to fit specific patterns of mortality in those cities.

With the individual city data, the investigators also explored two-pollutant models, in which PM_{10} results were controlled for the presence of ozone and vice versa; whether the association of ozone with mortality differed by warm and cold season;

This Statement, prepared by the Health Effects Institute, summarizes a research project funded by HEI and conducted by Dr. Isabelle Romieu, at the Instituto Nacional de Salud Pública, Morelos, México, and colleagues. The complete report, *Multicity Study of Air Pollution and Mortality in Latin America (the ESCALA Study)* (© 2012 Health Effects Institute), can be obtained from HEI or our Web site (see last page).

Epi-R

A graphic user interface oriented to epidemiological data analysis



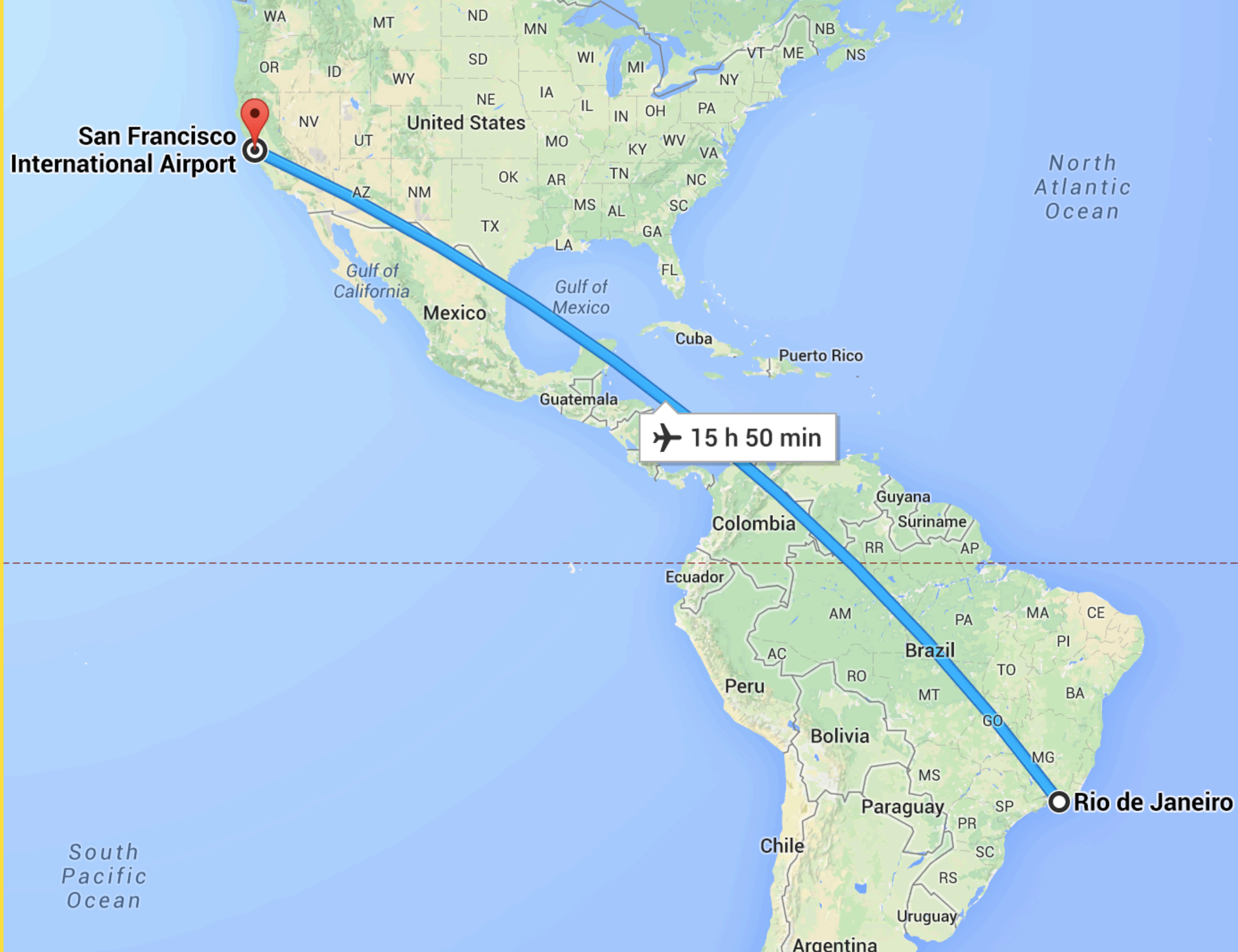
slides: bit.ly/eusr20

2012

Rio de Janeiro

››››

San Francisco

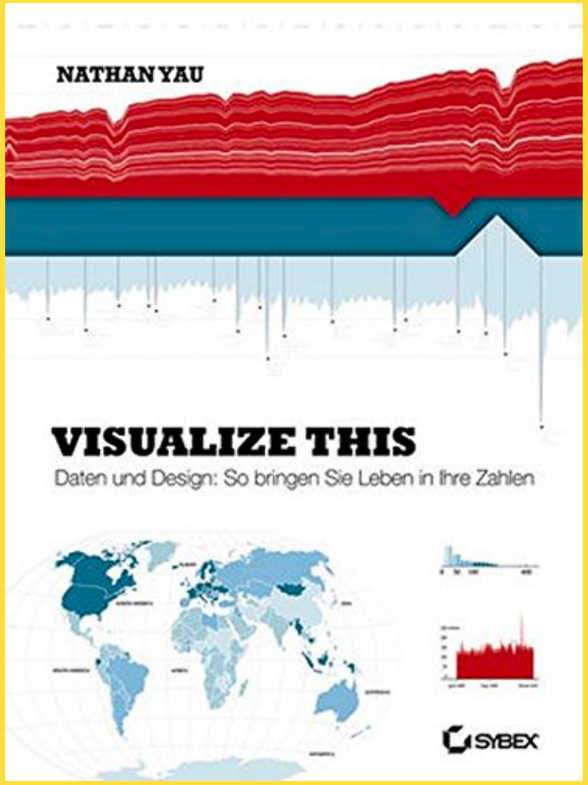


My name is Pino Trogu and I teach a Data Visualization class in the Fall at San Francisco State University. I come from a design background and I have played a little bit with coding. This semester I will be using Visualize This by Nathan Yau as the class textbook. The majority of examples are done using R and I am looking for TAs to help me with the R exercises described in the book and maybe with a little bit of javascript.

The class is really an introduction to data visualization with emphasis on best design practices.

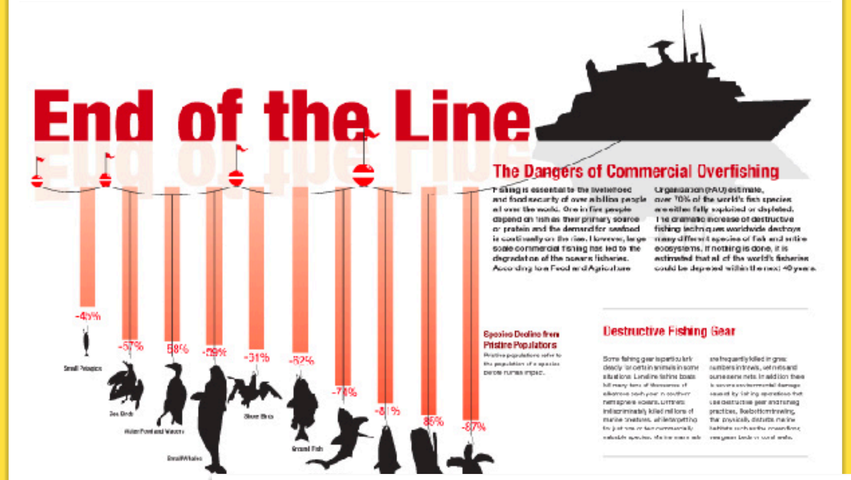
This Fall I teach two sections on M/W afternoons and evenings.

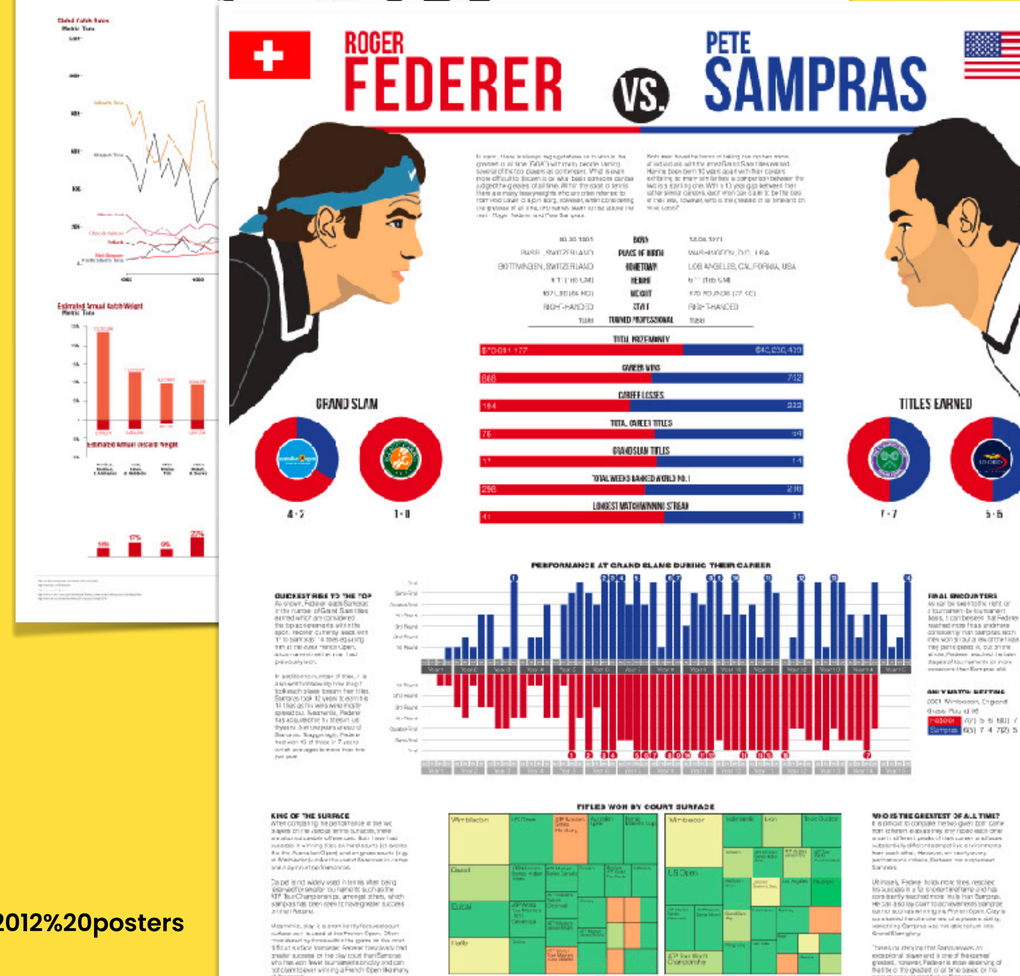
R seems great as a base for graphics that can then be finalized with Illustrator for print-based publication. Last year we worked a bit with D3 which I think has a steeper learning curve, and is more geared to interaction and animation. Of course anyone with some knowledge of D3 and javascript in general would be great for this also.



Teaching Assistant

School of Design @SFSU





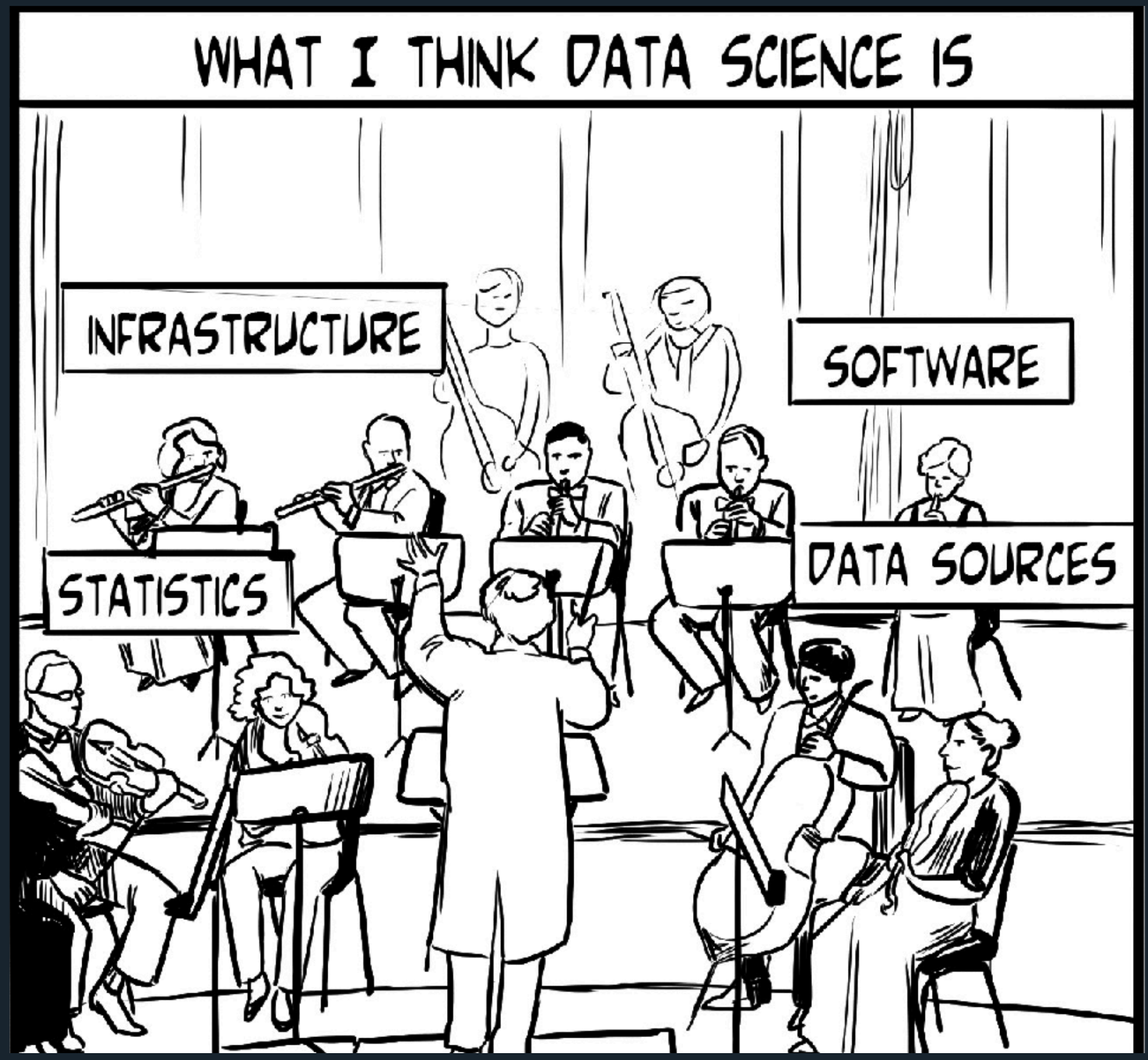
You can find more:
http://523informationdesign.blogspot.com/search/label/2012%20posters

The Data Science Career

What is Data Science?

slides: bit.ly/eusr20

What is Data Science?



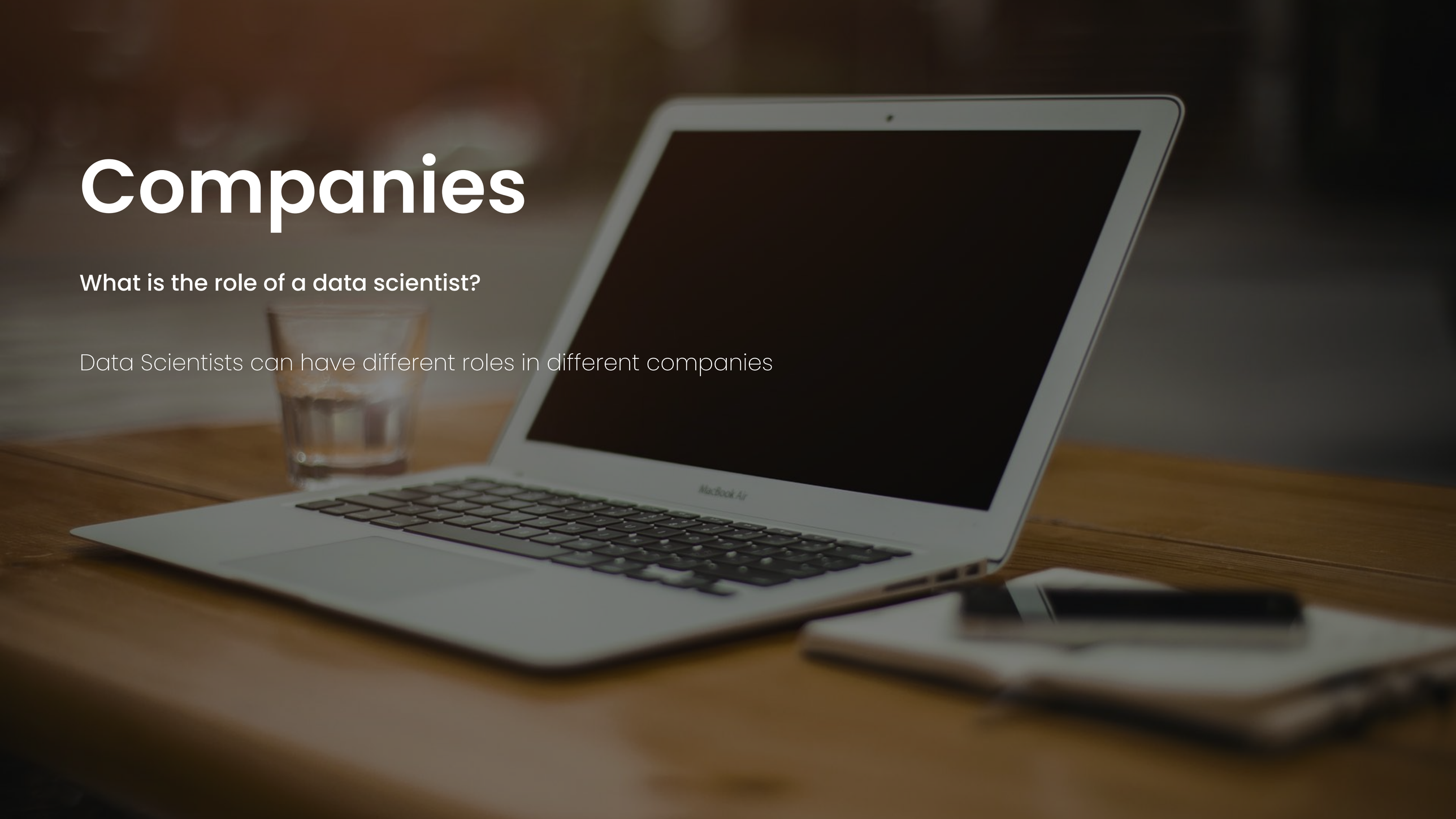
slides: bit.ly/eusr20


Data Scientists are adaptable and
flexible professionals

Companies

What is the role of a data scientist?

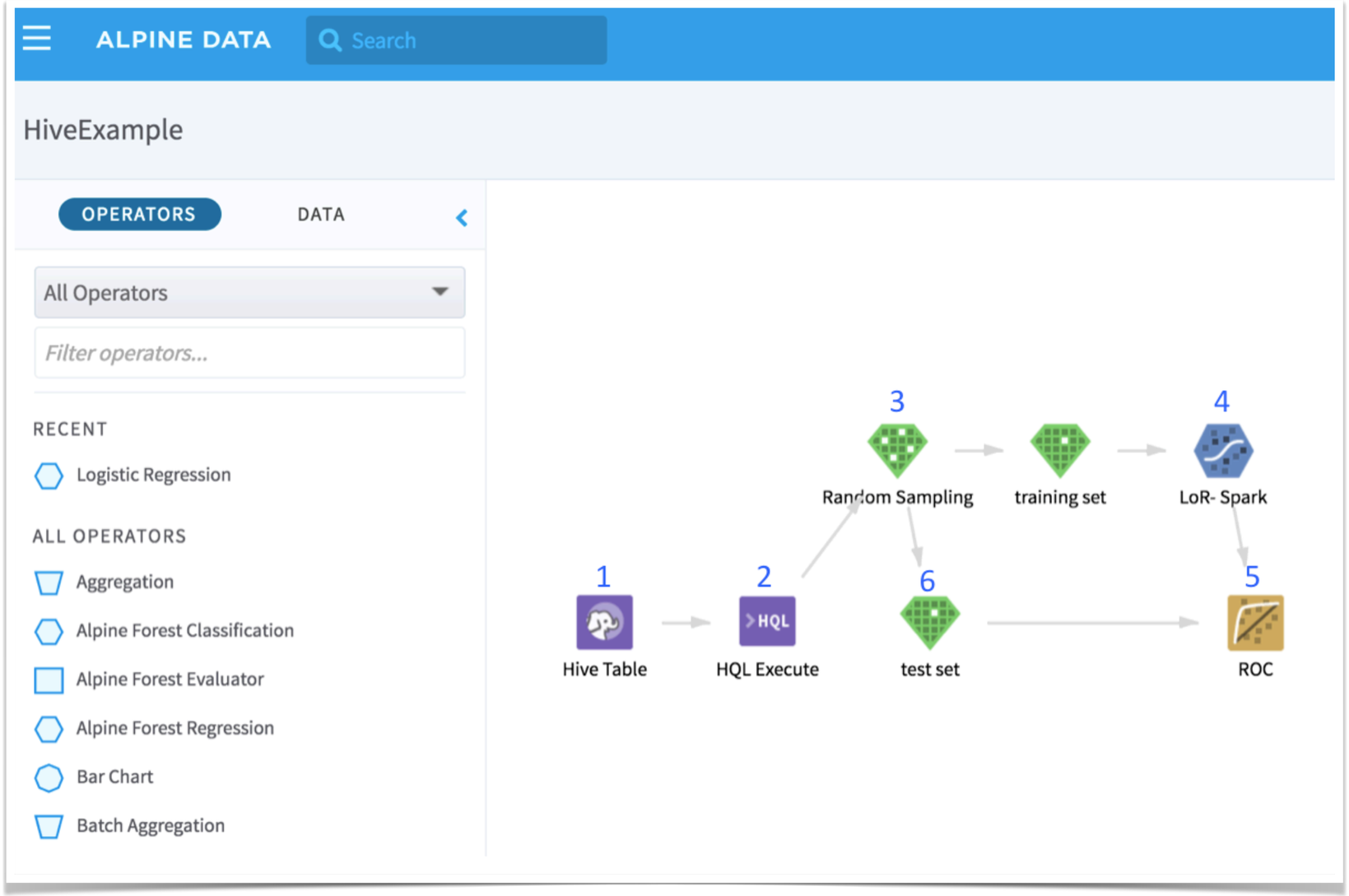
Data Scientists can have different roles in different companies



Advanced Analytics Platform for Big Data

ALPINE DATA™

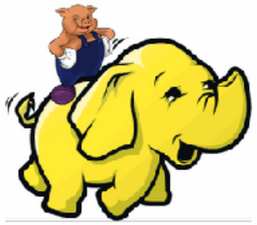
Empowers business users to define and participate in data science projects and gives data scientists the tools they need to create value from data



Company Numbers

- **40** Employees
- **10** Engineers
- **5** Data Scientist

Tool

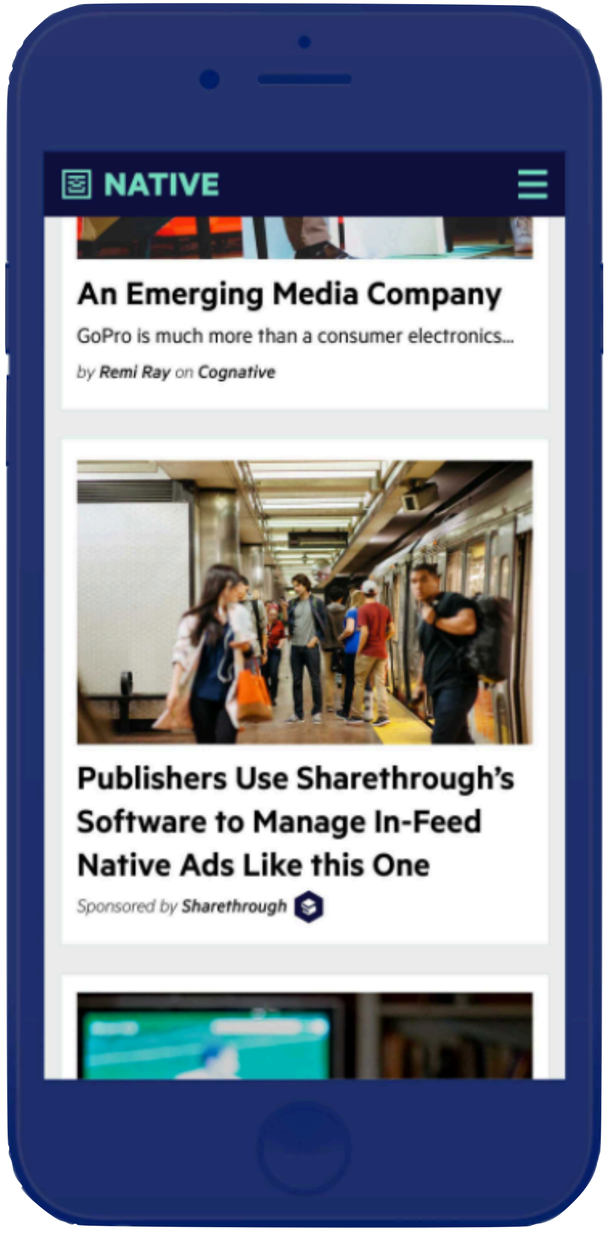


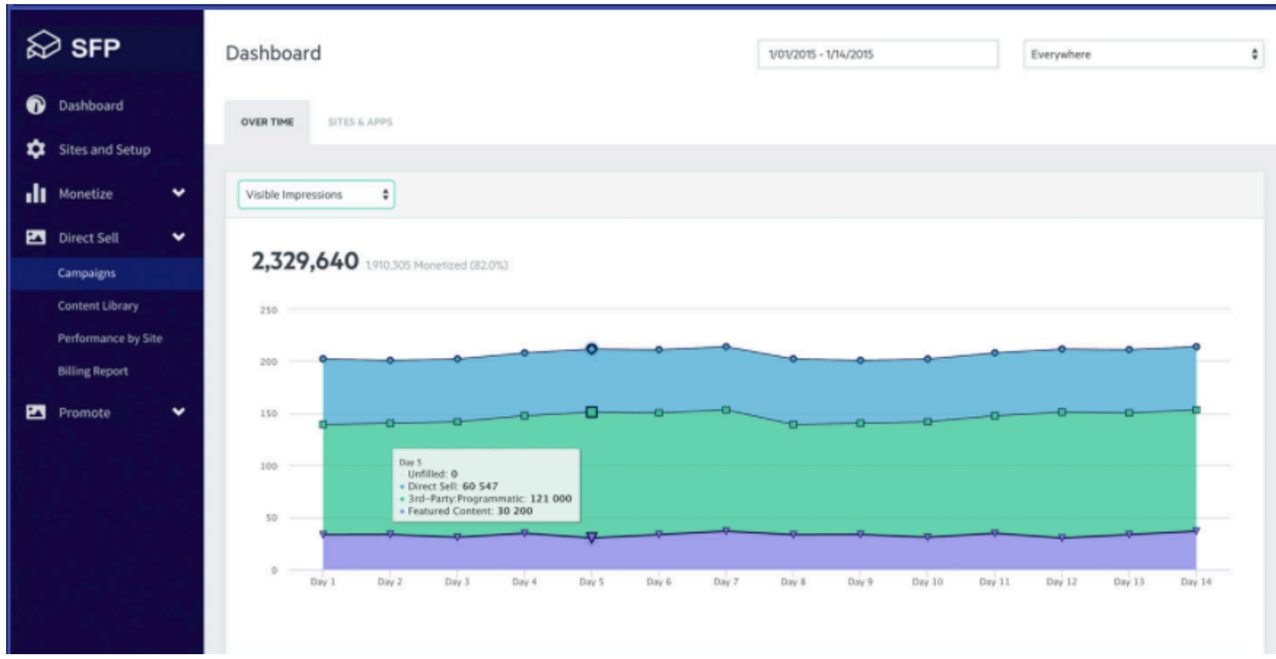
Key Responsibilities:

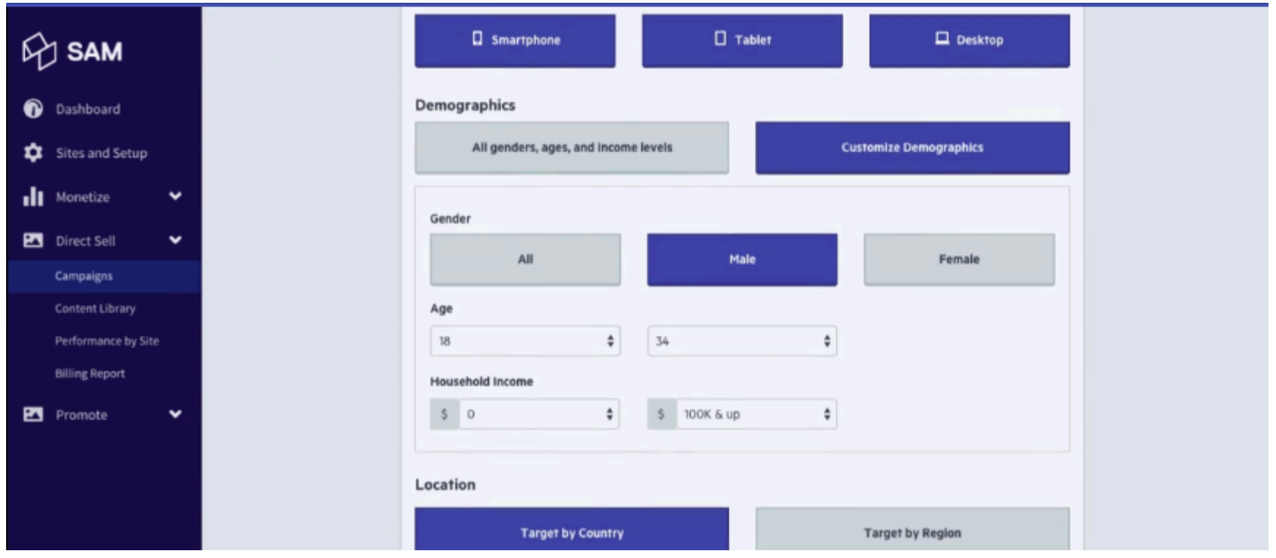
- Consultant
- Write Documentation
- Train Customers
- Prioritization

Native Advertising software for publishers, app developers & advertisers.

sharethrough







Company Numbers

- **150** Employees
- **20** Engineers
- **1** Data Scientist

Tool

Key Responsibilities:

Infrastructure

Models / Optimization

Experimental Design

Education

Credit Card designed for Generation Z

(INCLUDING INTERNATIONAL STUDENTS)

Through machine learning and alternative data, Deserve is helping millennials and Gen Z's, the next wave of credit owners, gain financial independence through access to fair credit products.

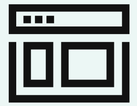
Company Numbers

- **35** Employees
- **7** Engineers
- **4** Data Scientist

Tool

Key Responsibilities:

Infrastructure



Dashboard

R packages

Models

Open Source @ IBM

CODAIT

Center for Open Source Data
and AI Technologies

Open Source, Data & AI Technologies

Company Numbers

- **+300k** Employees
- **?** Engineers
- **?** Data Scientist

Tool

python™

docker

R

Key Responsibilities:

Senior Open Source Developer - DL/ML/AI Developer and Advocate

Key Responsibilities:

1) Write great code in key open source communities
2) Democratize AI by building tools, launching new open source projects, and improving existing ones
3) Gain eminence in the community by socializing your work, and speaking at events
4) Work with offering managers and product teams on applications
5) Guide and mentor clients to become self-sufficient open source developers
6) Be authentic; mentor others, and be open to mentoring by others
7) Read and comment on more code than you write; fix bugs, test cases, and documentation
8) Etc ...

Sr. Machine Learning Manager

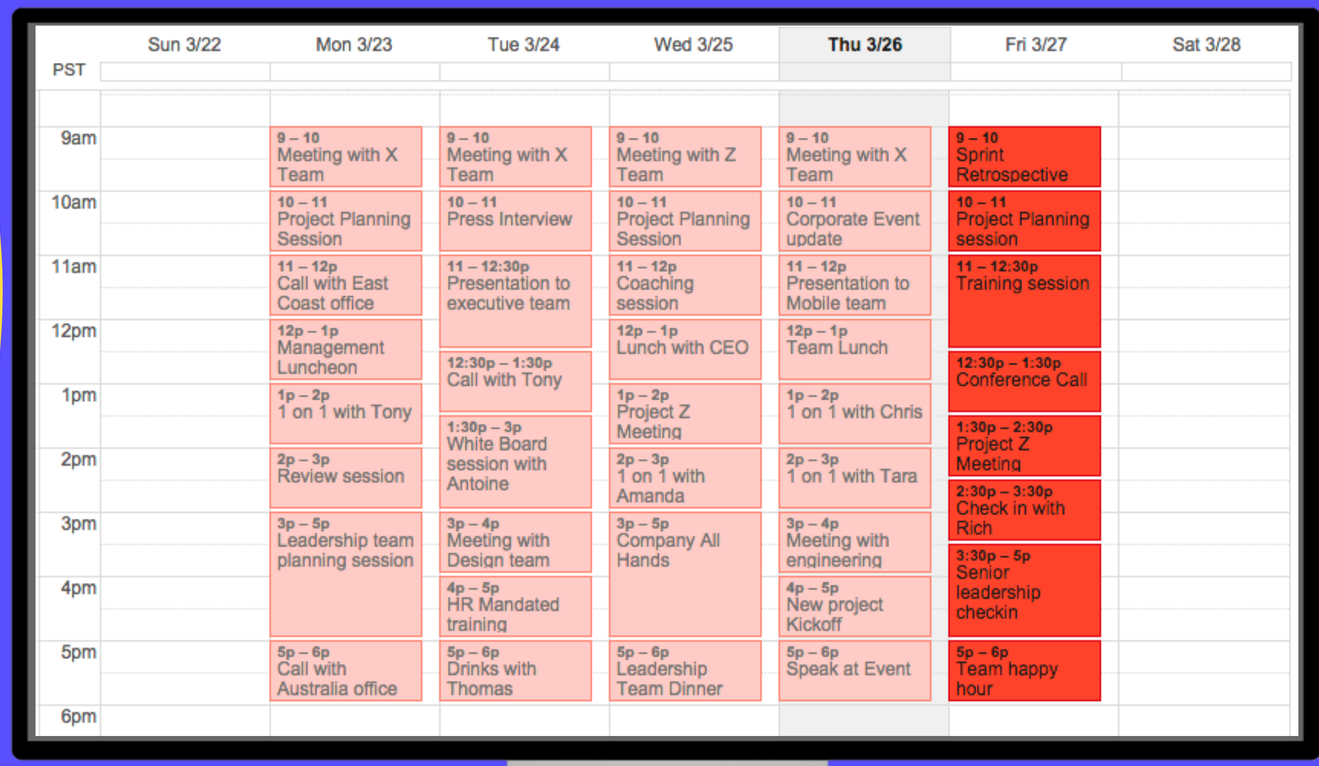
Responsibilities: (MANY!)

MACHINE LEARNING TEAM

10 open source developers

- Data Scientists
- Software Engineers

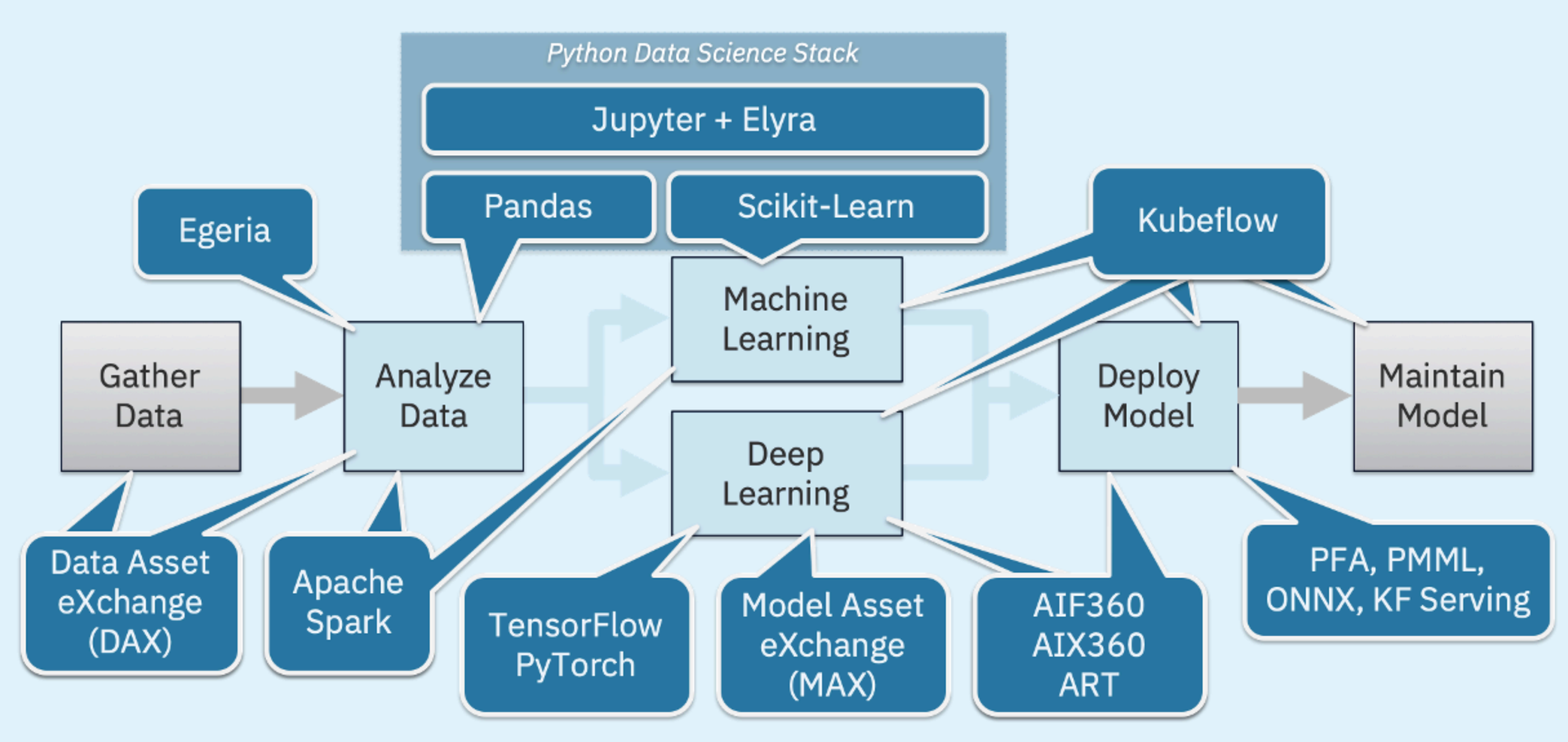
>> Backgrounds in Stats, Math, Engineering, Computer Science, Finance.



PST	Sun 3/22	Mon 3/23	Tue 3/24	Wed 3/25	Thu 3/26	Fri 3/27	Sat 3/28
9am		9 – 10 Meeting with X Team	9 – 10 Meeting with X Team	9 – 10 Meeting with Z Team	9 – 10 Meeting with X Team	9 – 10 Sprint Retrospective	
10am		10 – 11 Project Planning Session	10 – 11 Press Interview	10 – 11 Project Planning Session	10 – 11 Corporate Event update	10 – 11 Project Planning session	
11am		11 – 12p Call with East Coast office	11 – 12:30p Presentation to executive team	11 – 12p Coaching session	11 – 12p Presentation to Mobile team	11 – 12:30p Training session	
12pm		12p – 1p Management Luncheon	12:30p – 1:30p Call with Tony	12p – 1p Lunch with CEO	12p – 1p Team Lunch	12:30p – 1:30p Conference Call	
1pm		1p – 2p 1 on 1 with Tony	1:30p – 3p White Board session with Antoine	1p – 2p Project Z Meeting	1p – 2p 1 on 1 with Chris	1:30p – 2:30p Project Z Meeting	
2pm		2p – 3p Review session		2p – 3p 1 on 1 with Amanda	2p – 3p 1 on 1 with Tara	2:30p – 3:30p Check in with Rich	
3pm		3p – 5p Leadership team planning session	3p – 4p Meeting with Design team	3p – 5p Company All Hands	3p – 4p Meeting with engineering	3:30p – 5p Senior leadership checkin	
4pm			4p – 5p HR Mandated training		4p – 5p New project Kickoff		
5pm		5p – 6p Call with Australia office	5p – 6p Drinks with Thomas	5p – 6p Leadership Team Dinner	5p – 6p Speak at Event	5p – 6p Team happy hour	
6pm							

We build tools to make AI accessible and available to everybody

(codait.org)



We contribute to and advocate for the **open-source** technologies

codait.org

What is Open Source?

- The term open source refers to something people can modify and share because its design is publicly accessible

- Open source software (OSS) is software with source code that anyone can inspect, modify, and enhance user problems and help you achieve your business goals.

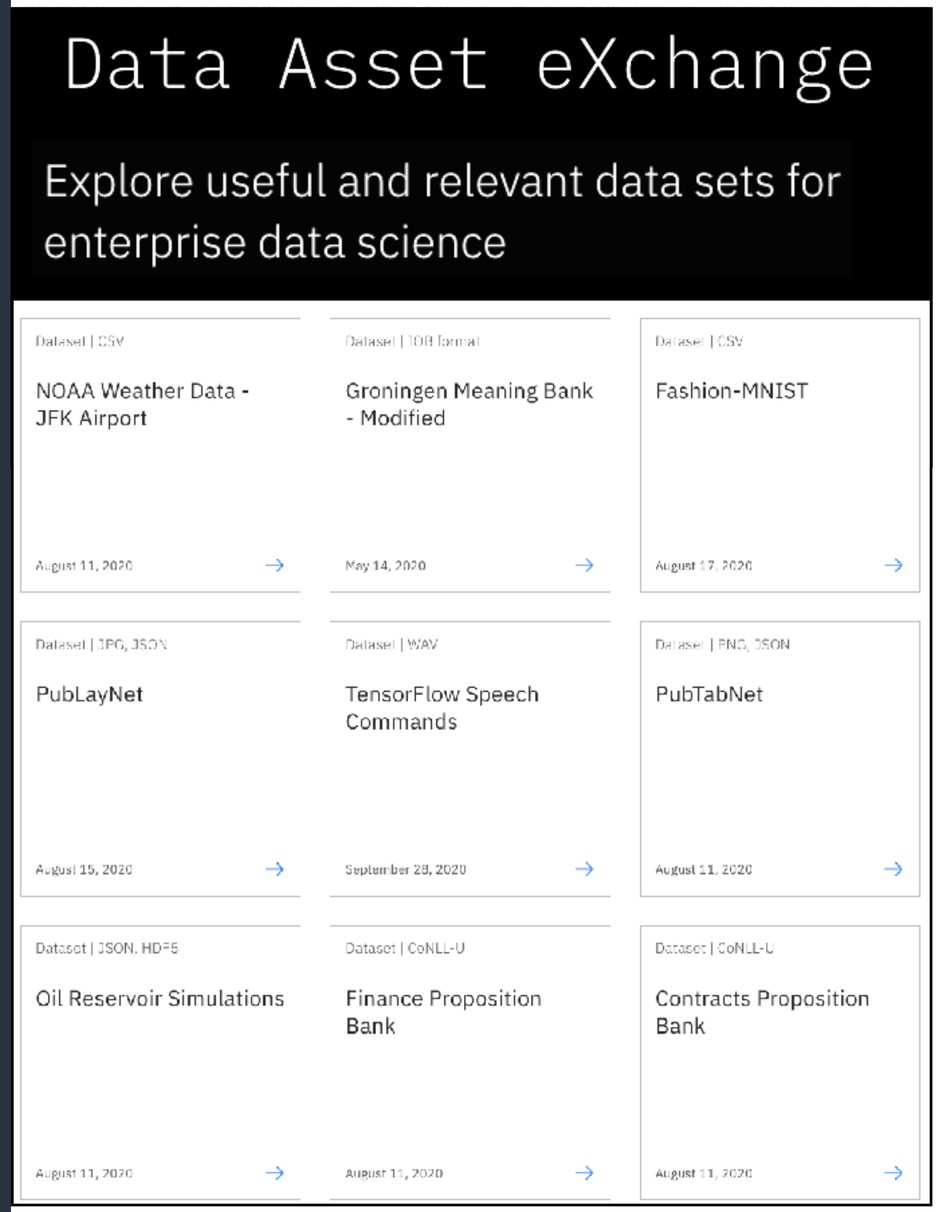
https://opensource.com/resources/what-open-source

Open Source @ IBM

Some Projects

Data Asset eXchange (DAX)

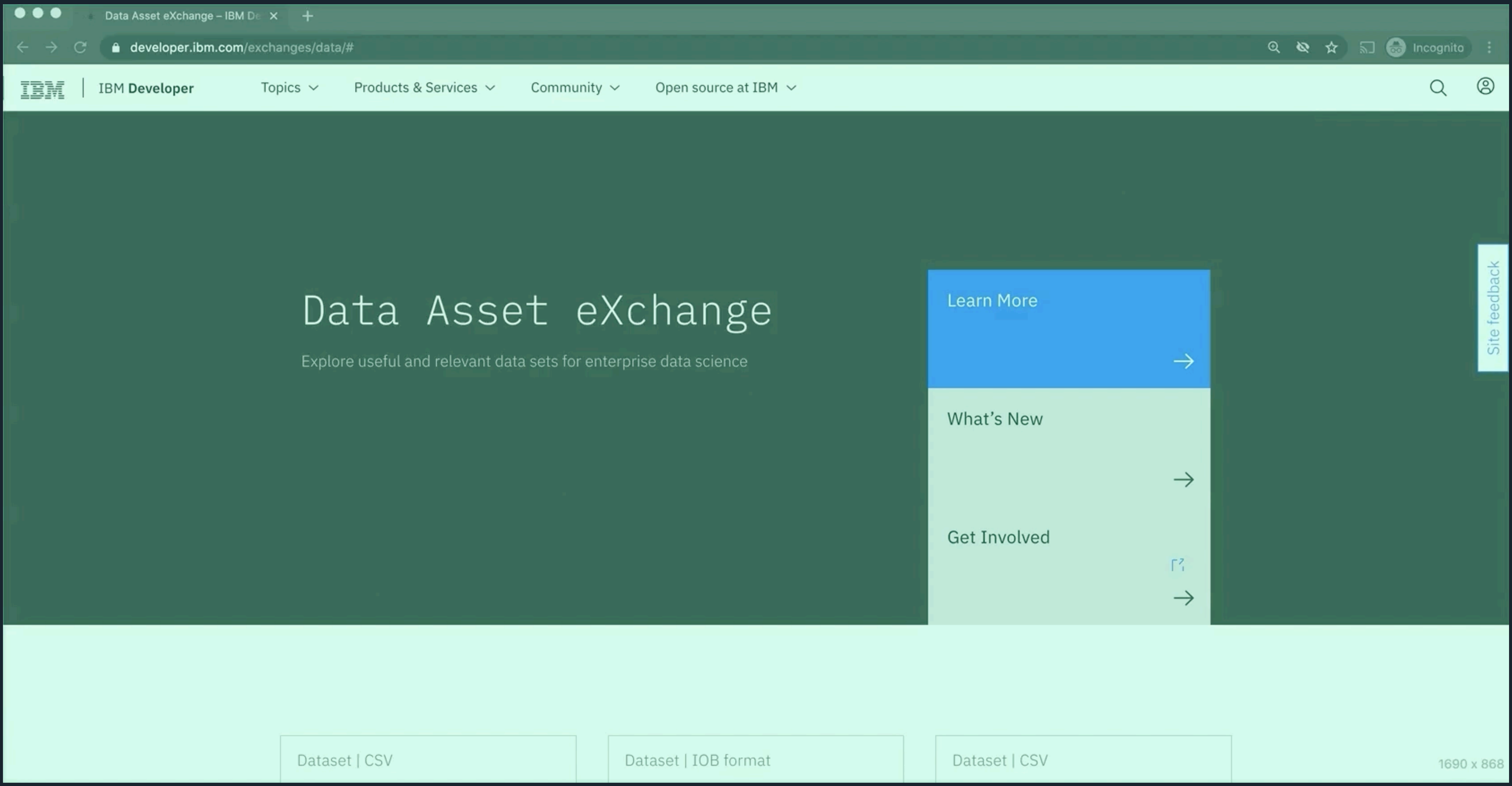
Website:
ibm.biz/data-exchange



Data Asset eXchange (DAX)

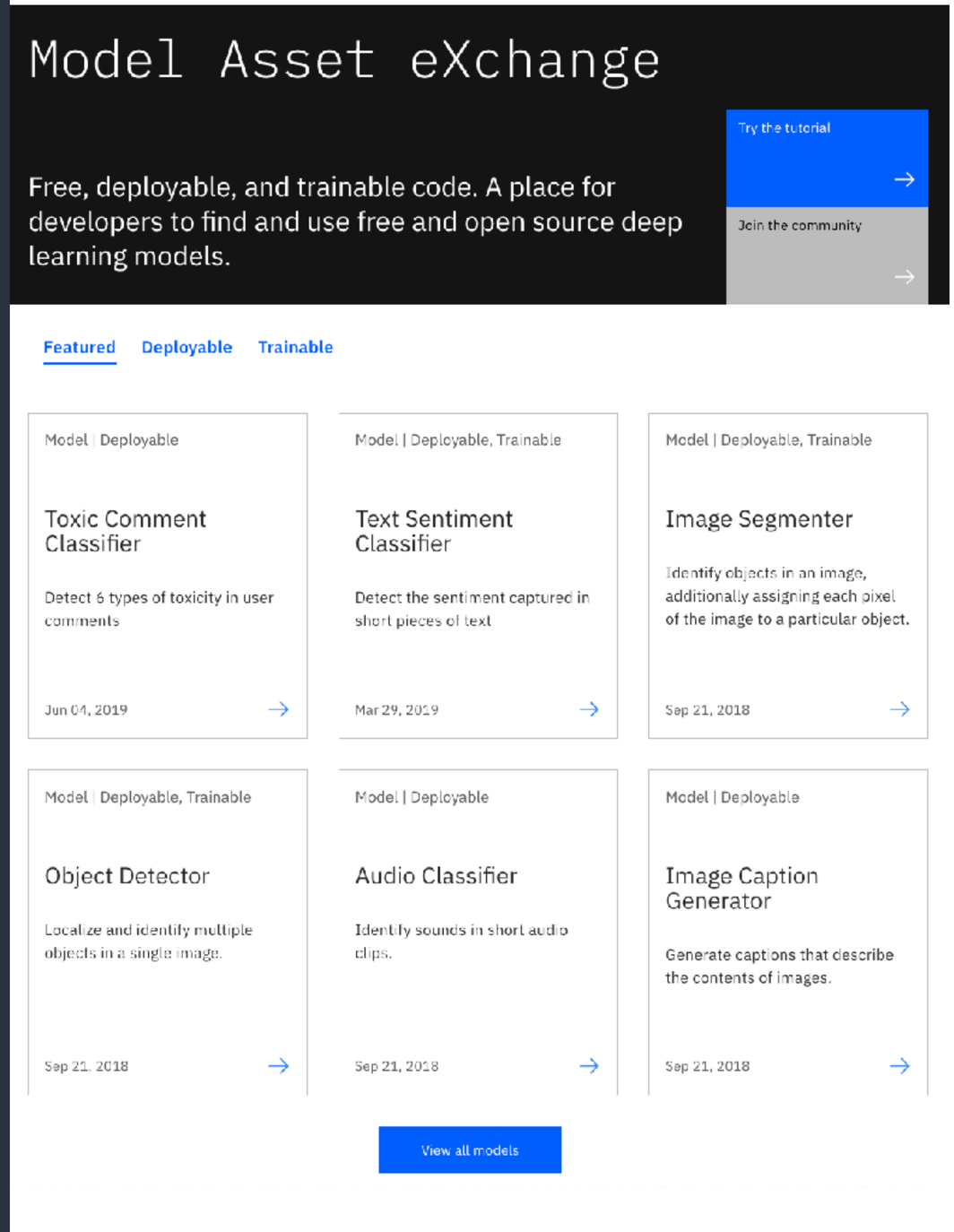
- Curated repository for **open datasets** from IBM Research and third-parties
- Published under data **friendly licenses**
- **Standardized** dataset **formats** and **metadata**
- All datasets include notebooks
  - Data ingesting
  - Data exploration
  - Data analysis

ibm.biz/data-exchange


Data Asset eXchange – IBM De
developer.ibm.com/exchanges/data/#
Incognito
IBM Developer
Topics
Products & Services
Community
Open source at IBM
Data Asset eXchange
Explore useful and relevant data sets for enterprise data science
Learn More
What's New
Get Involved
Site feedback
Dataset | CSV
Dataset | IOB format
Dataset | CSV
1690 x 868

Model Asset eXchange (MAX)

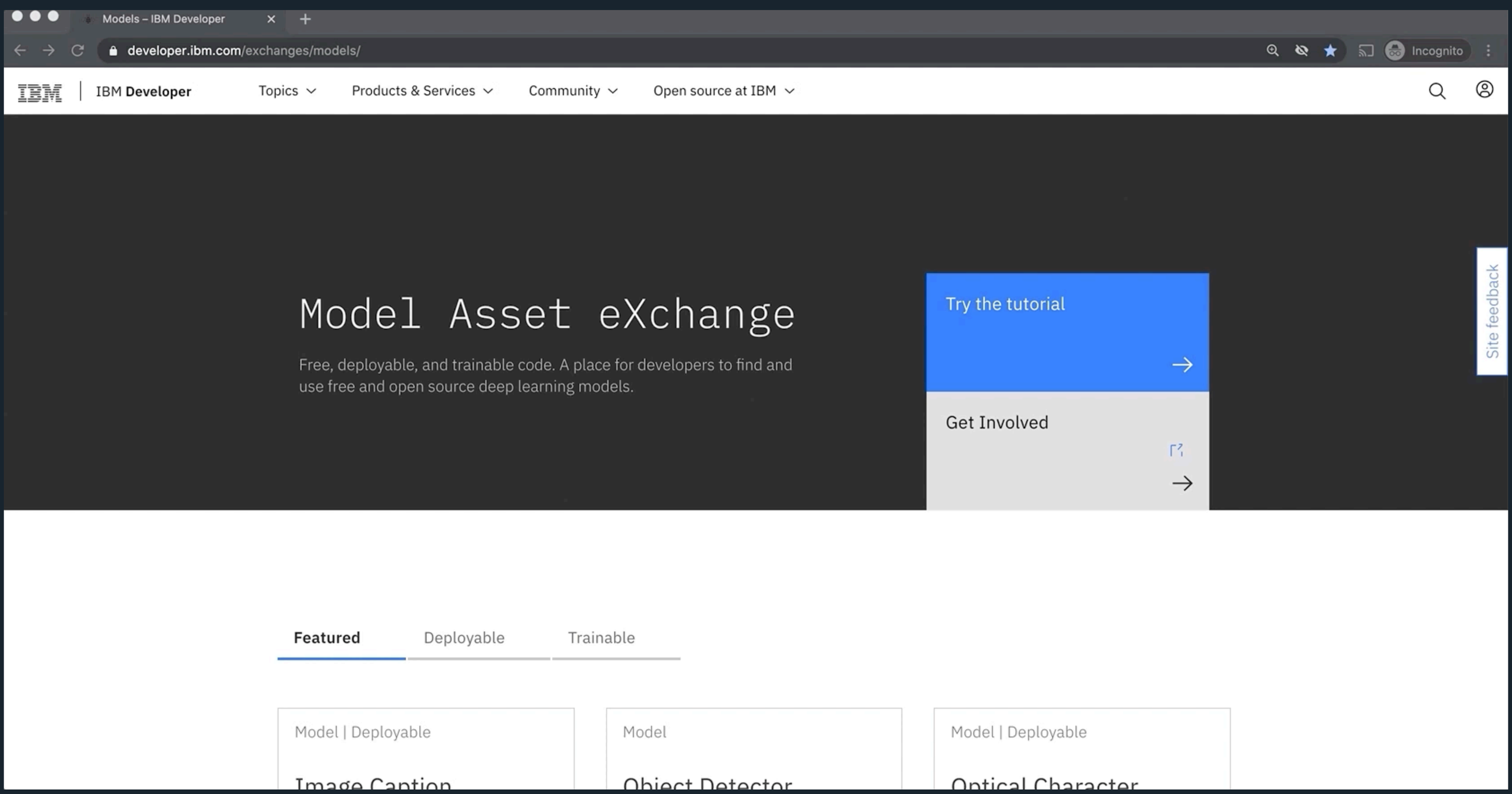
Website:
ibm.biz/model-exchange

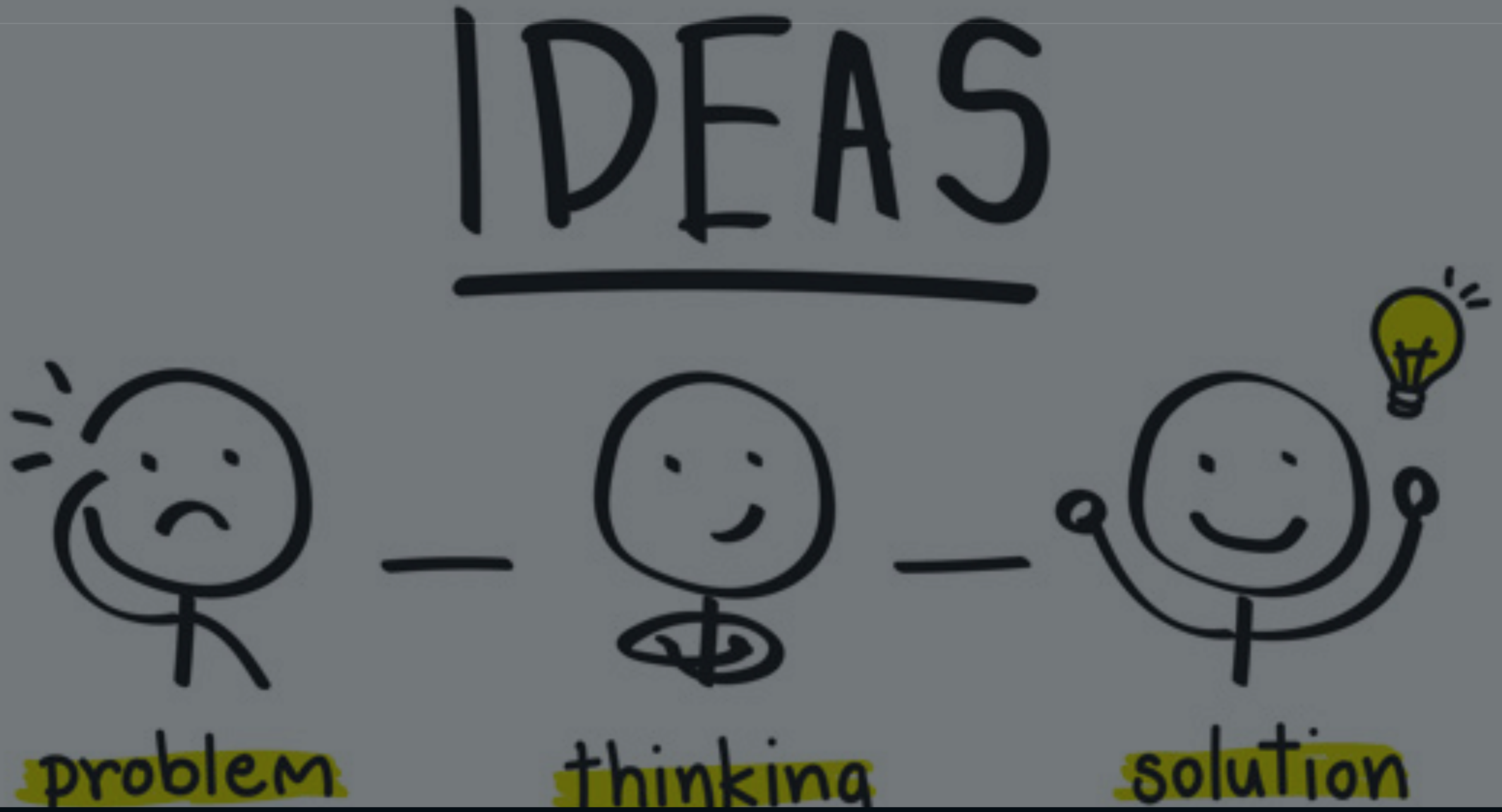


Model Asset eXchange (MAX)

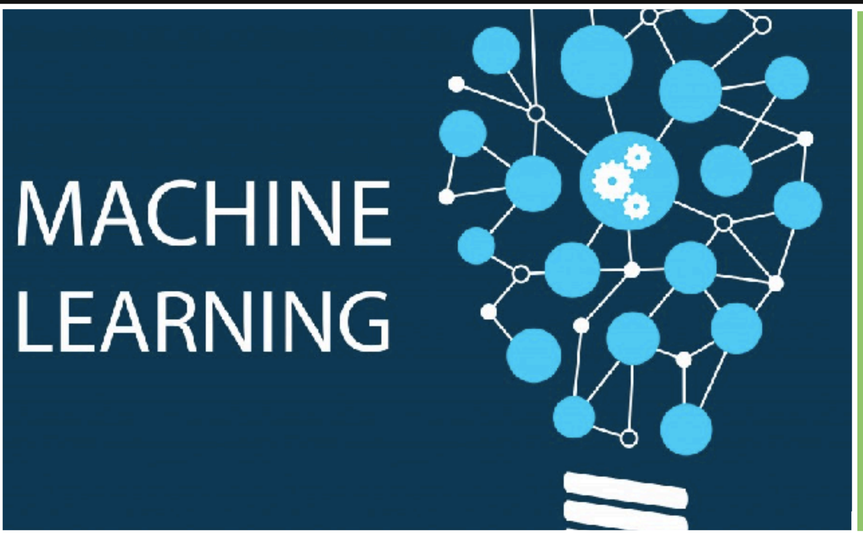
- A place for developers/data scientists to find and use **free** and **open source** deep learning models
- Wide variety of domains (text, audio, image, etc)
- Multiple deep learning frameworks (TensorFlow, PyTorch, Keras)
- Trainable and Deployable versions

ibm.biz/model-exchange


Models – IBM Developer
developer.ibm.com/exchanges/models/
Incognito
IBM Developer
Topics
Products & Services
Community
Open source at IBM
Model Asset eXchange
Free, deployable, and trainable code. A place for developers to find and use free and open source deep learning models.
Try the tutorial
Get Involved
Site feedback
Featured
Deployable
Trainable
Model | Deployable
Image Caption
Model
Object Detector
Model | Deployable
Optical Character

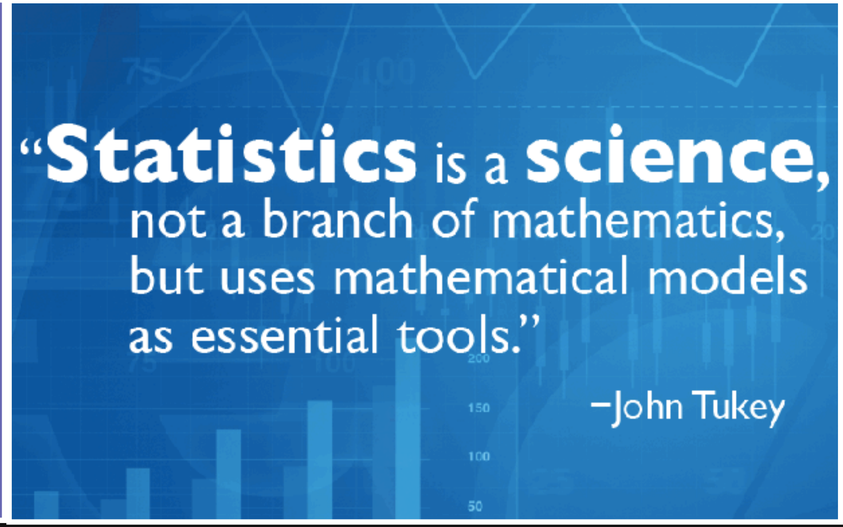


IMPORTANT SKILLS FOR A DATA SCIENTIST



Machine Learning

Statistics



Programming

Communication

Curiosity
(keep asking why)

Flexibility

Critical Thinking

Ethics

Be yourself

Thank you!

slides: **bit.ly/eusr20**

ai-inclusive.org

Follow us:

bit.ly/ai-inclusive-instagram

Resources on AI, DS, ML
Events, Free Tickets and much more