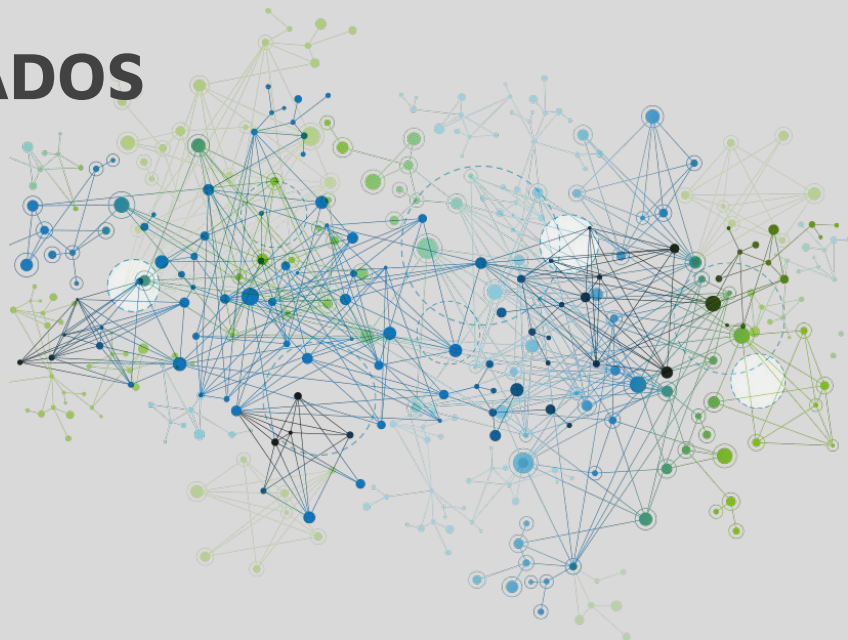


**DA ANÁLISE EXPLORATÓRIA AO  
APRENDIZADO DE MÁQUINA:**

**ESTATÍSTICOS NA ÁREA DE DADOS**





## **Anderson Neisse**

Estatístico

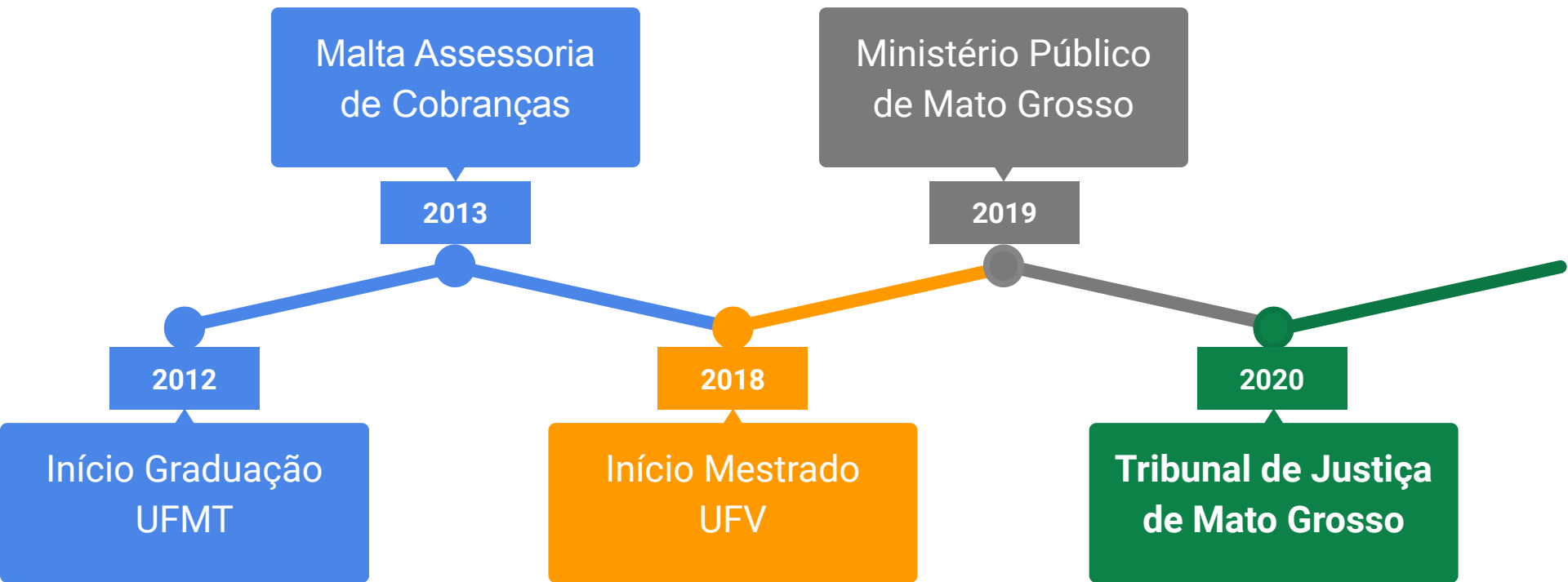
Mestre em Estatística Aplicada e Biometria

Co-fundador do DataSciMT

[aneisse.com](http://aneisse.com)

**Gerente de Estatística e Gestão Estratégica**

**Tribunal de Justiça de Mato Grosso**



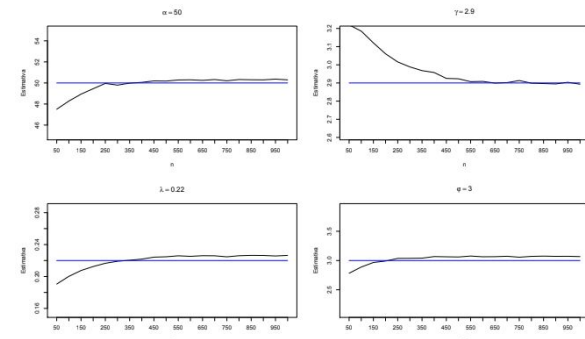


# Tópico do TCC

## IMPLEMENTAÇÃO E ESTUDO DE SIMULAÇÃO DAS DISTRIBUIÇÕES KUMARASWAMY WEIBULL E KUMARASWAMY LOG-LOGÍSTICA

- Distribuições da Família Kumaraswamy;
- Avaliação dos estimadores dos parâmetros;
- Simulação de Monte Carlo;

Figura 7: Estimativas médias para 20 tamanhos amostrais da KumLL.

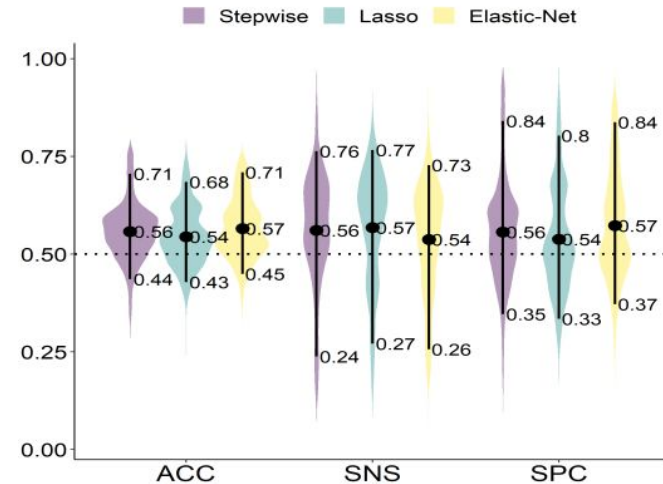




# Dissertação

## SÍNDROME DA FADIGA CRÔNICA E ABSENTEÍSMO: ESTUDO DE TRABALHADORES EM TURNOS COMPARANDO STEPWISE E ELASTIC-NET

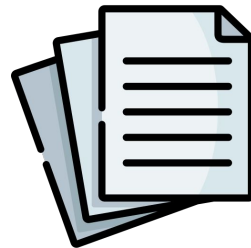
- Dados de saúde (Trabalhadores de minas);
- Stepwise, Lasso, Ridge, Elastic-Net;
- Bootstrap para validação cruzada;





## Alguns artigos

- AMMI and GGE Biplot for genotype environment interaction: a medoid-based hierarchical cluster analysis approach for high-dimensional data;
- Aplicação de componentes principais e análise fatorial a dados criminais de 26 estados dos EUA;
- Inventário de ninhos de abelhas sociais sem ferrão em uma área antropizada;



# DataSciMT

## Grupo de discussão de Data Science no estado de Mato Grosso



V Semana da Estatística - 10 Anos de Curso (Dia #1)  
1,9 mil visualizações • Transmitido há 2 anos



1º Data Science Talks - Carreiras (Dia 3)  
186 visualizações • Transmitido há 2 anos



Entrevista com Mario Filho  
209 visualizações • Transmitido há 2 anos



V Semana da Estatística - 10 Anos de Curso (Dia #2)  
1,9 mil visualizações • Transmitido há 2 anos



V Semana da Estatística - 10 Anos de Curso (Dia #4)  
1,4 mil visualizações • Transmitido há 2 anos



1º Data Science Talks - Papéis e Processos (Dia 1)  
257 visualizações • Transmitido há 2 anos



V Semana da Estatística - 10 Anos de Curso (Dia #5)  
1,2 mil visualizações • Transmitido há 2 anos



1º Data Science Talks - Ferramentas e Skills / Segurança da Informação (Dia 2)  
190 visualizações • Transmitido há 2 anos



V Semana da Estatística - 10 Anos de Curso (Dia #3)  
1,5 mil visualizações • Transmitido há 2 anos



1º Meetup online - Lucas Bianchi (Covid-19 em dados)  
279 visualizações • Transmitido há 2 anos



# Tribunal de Justiça de Mato Grosso

- Validar, consolidar e enviar dados estatísticos destinados ao CNJ;
- Subsidiar a tomada de decisões da administração do Poder Judiciário;
- Divulgar dados a respeito do desempenho e atividades do Poder Judiciário;
- Auxiliar nas atividades de planejamento, análises, projetos e controle estatístico;
- Atuar em medidas de saneamento e correção dos dados, sempre que necessário;
- Atuar na frente de inovação com dados aprimorando a Gestão e a Prestação à Sociedade;







# O estatístico e o mercado de dados



# A Ciência de Dados é a disciplina de tornar dados úteis.

- Cassie Kozyrkov  
Chief Decision Scientist (Google)



# Data Science

**Analytics**

Get inspired

**Statistics**

Decide wisely

**ML/AI**

Make recipe



@quaesita

[bit.ly/quaesita\\_datasci](https://bit.ly/quaesita_datasci)



# Funções nas carreiras de dados

- Estrategista de Dados
- **Engenheiro de Dados**
- **Analista de BI**
- **Analista de Dados**
- **Cientista de Dados** (Cientista de Negócio?)
- Engenheiro de ML

**E o Estatístico?**

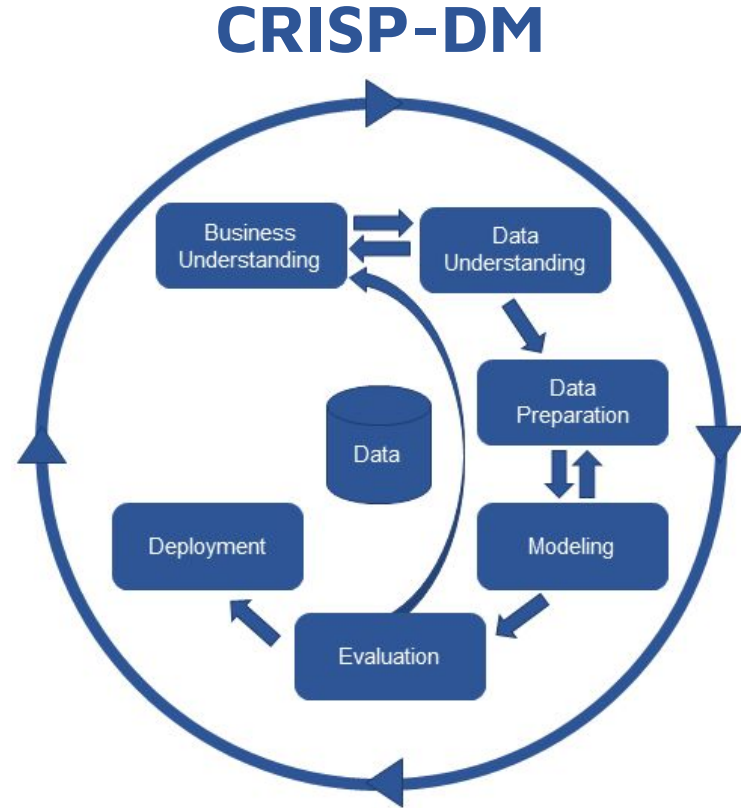


|   | Data Analyst       | Machine Learning Engineer | Data Engineer      | Data Scientist     |
|---|--------------------|---------------------------|--------------------|--------------------|
| Programming Tools                         | Very important     | Very important            | Very important     | Very important     |
| Data Visualization and Communication      | Very important     | Somewhat important        | Somewhat important | Very important     |
| Data Intuition                            | Somewhat important | Very important            | Somewhat important | Very important     |
| Statistics                                | Somewhat important | Very important            | Somewhat important | Very important     |
| Data Wrangling                            | Not that important | Not that important        | Very important     | Very important     |
| Machine Learning                          | Not that important | Very important            | Not that important | Very important     |
| Software Engineering                      | Not that important | Somewhat important        | Very important     | Somewhat important |
| Multivariable Calculus and Linear Algebra | Not that important | Very important            | Not that important | Somewhat important |

Legend:  
● Not that important  
● Somewhat important  
● Very important

# Outras Skills

- Computação em nuvem;
- Conhecimento de Negócio;
- Solução de problemas;
- Tomada de decisão;
- Processo de trabalho de DS;





**E ferramentas?**

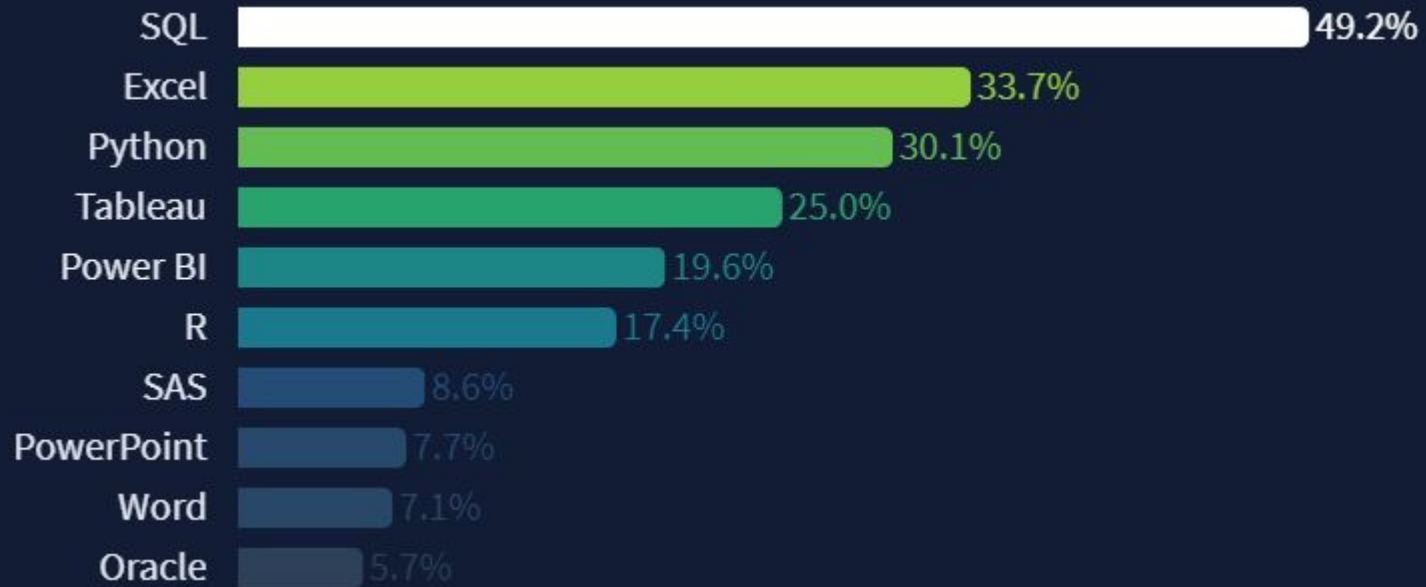


# E ferramentas?

| Data Pipeline    |   |               | Data Exploration + Modeling       |                | Deployment                        |                                   | Production App                     |  |
|------------------|---|---------------|-----------------------------------|----------------|-----------------------------------|-----------------------------------|------------------------------------|--|
| Data Sources     | ETL   | Data Process  | Data Exploration                  | Model Building | Model Mgt.                        | Model Deploy                      |                                    |  |
| Databases        | Version Control   |               | Viz + Analysis                    | Drag and Drop  | VC + Collab                       | API + Scheduler                   |                                    |  |
| <br><br><br>     | <br><br>  | <br><br>      | <br>                              | <br>           | <br>                              | <br><br><br>                      |                                    |  |
| SaaS Apps        | Cloud ELT/ETL   | Data WHouse   | Traditional BI/Analytics Tools    |                | Function Specific BI/Analytics    |                                   | Reporting                          |  |
| <br><br><br><br> | <br><br><b>Data Preparation</b><br><br><b>SQL Models Mgt.</b><br> | <br><br><br>  | <br><br><br>                      |                | <b>Prod Analytics</b><br><br><br> | <b>Infrastructure</b><br><br><br> |                                    |  |
| Files            | Schedulers  | Orchestration | Horizontal - Nextgen BI/Analytics |                |                                   |                                   |                                    |  |
|                  |   |               |                                   |                |                                   |                                   | <b>Data Tooling Marketmap 2020</b> |  |

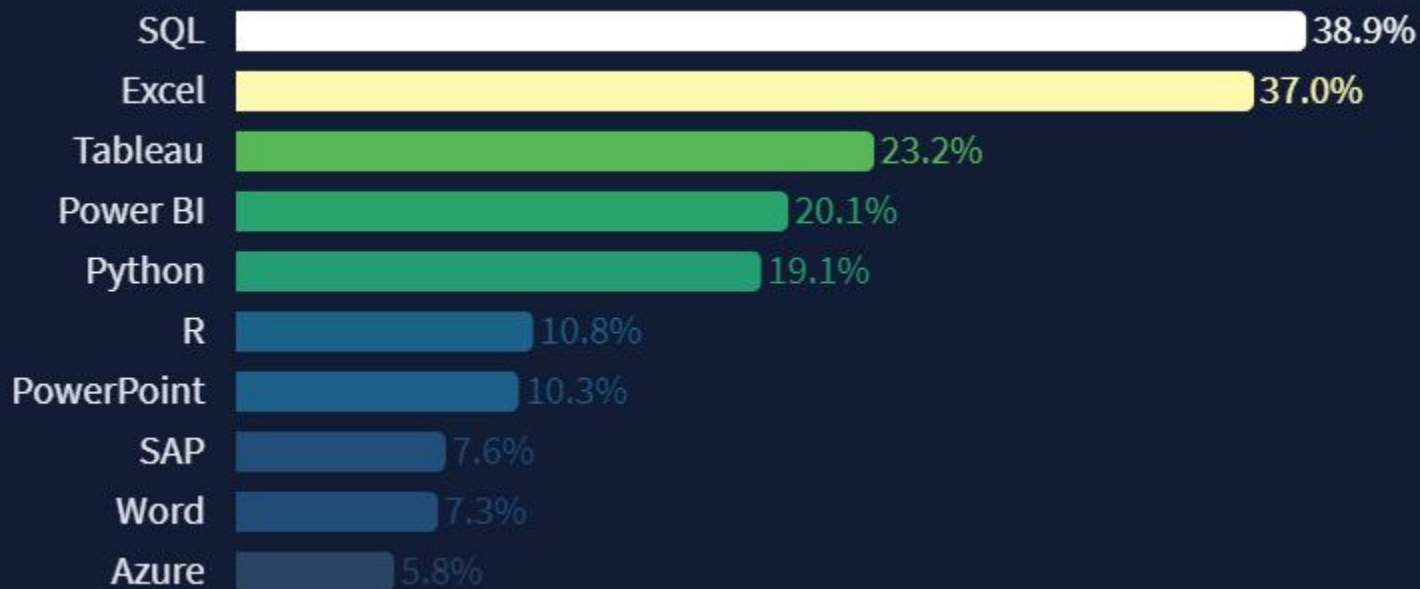
# As mais solicitadas...

## Analista de Dados



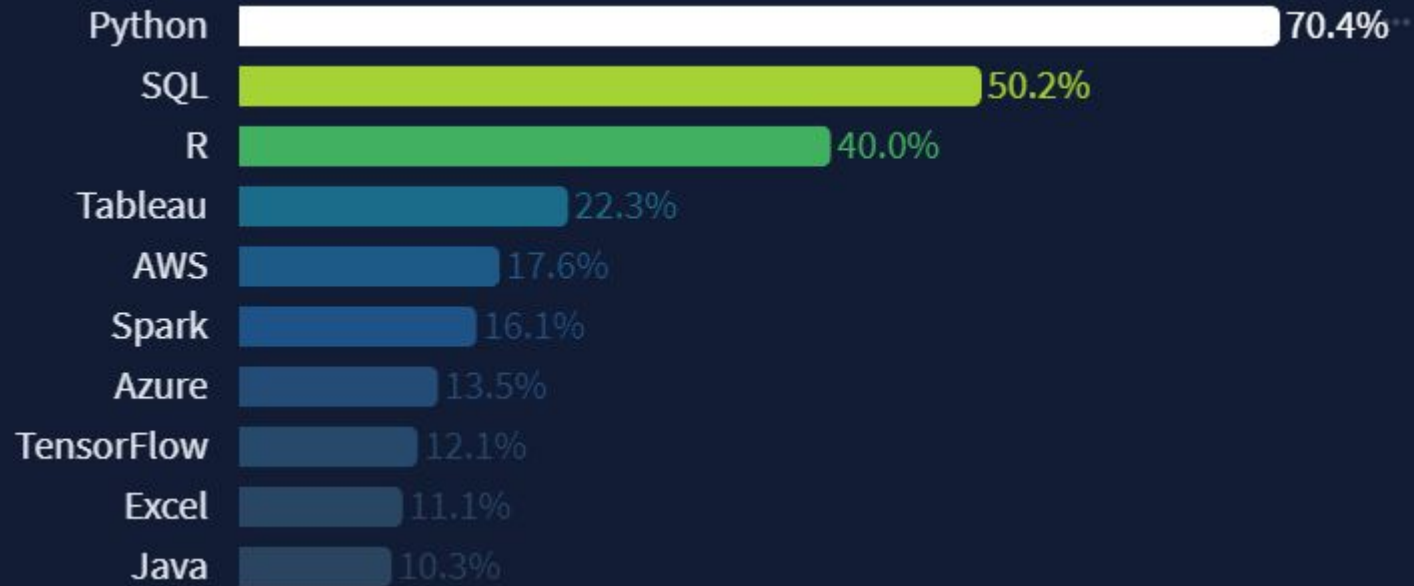
# As mais solicitadas...

## Analista de BI



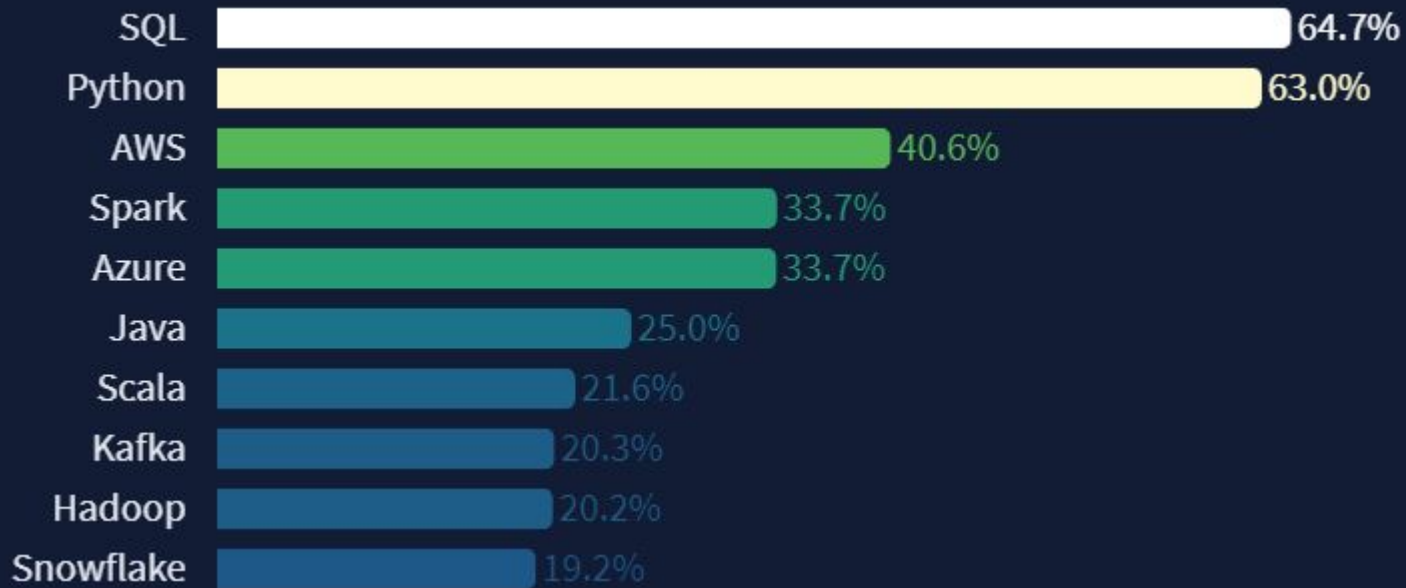
# As mais solicitadas...

## Cientista de Dados



# As mais solicitadas...

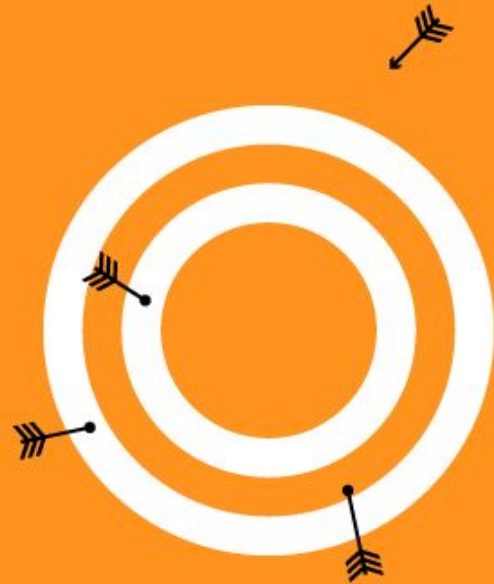
## Engenheiro de Dados

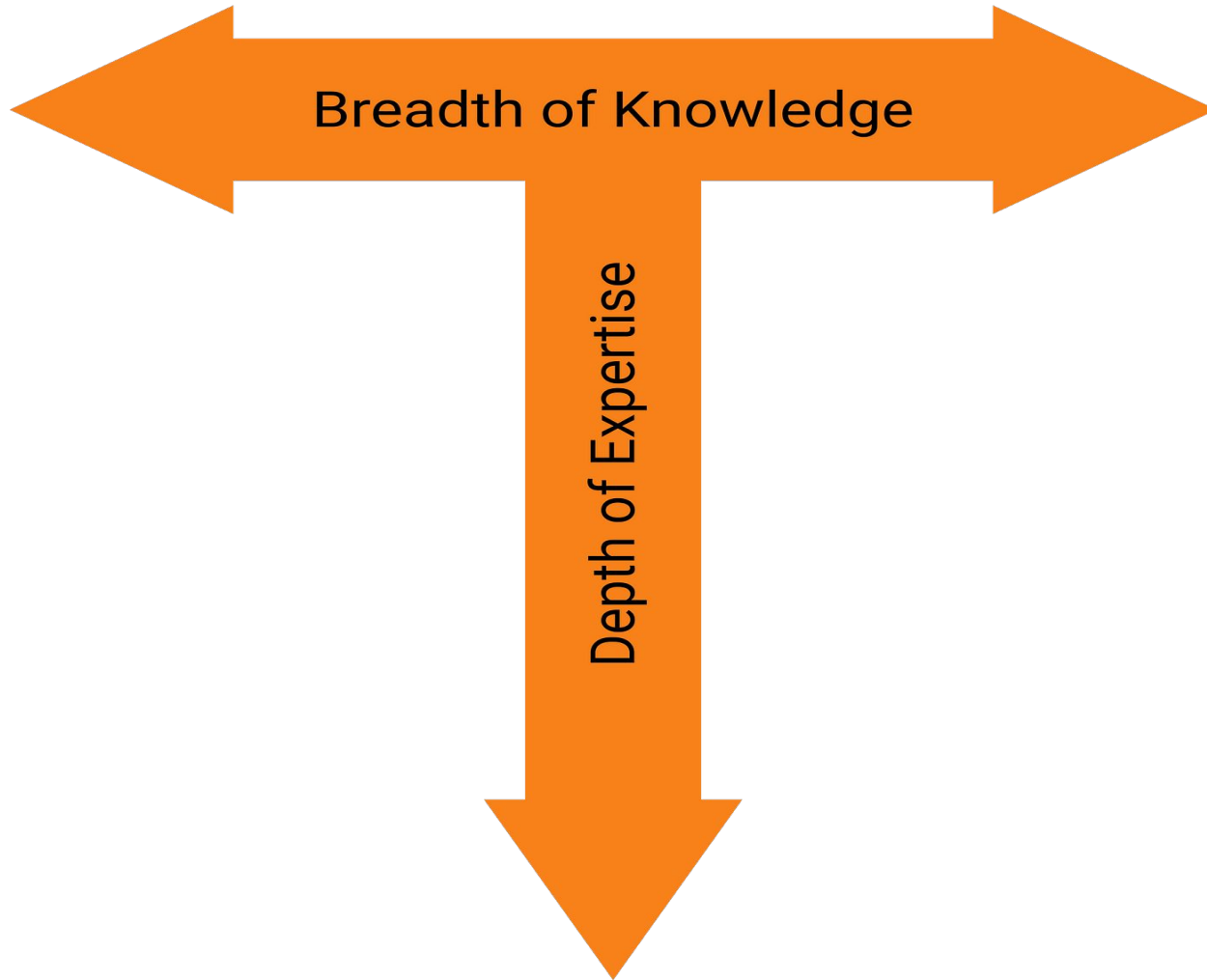


Specialists

vs

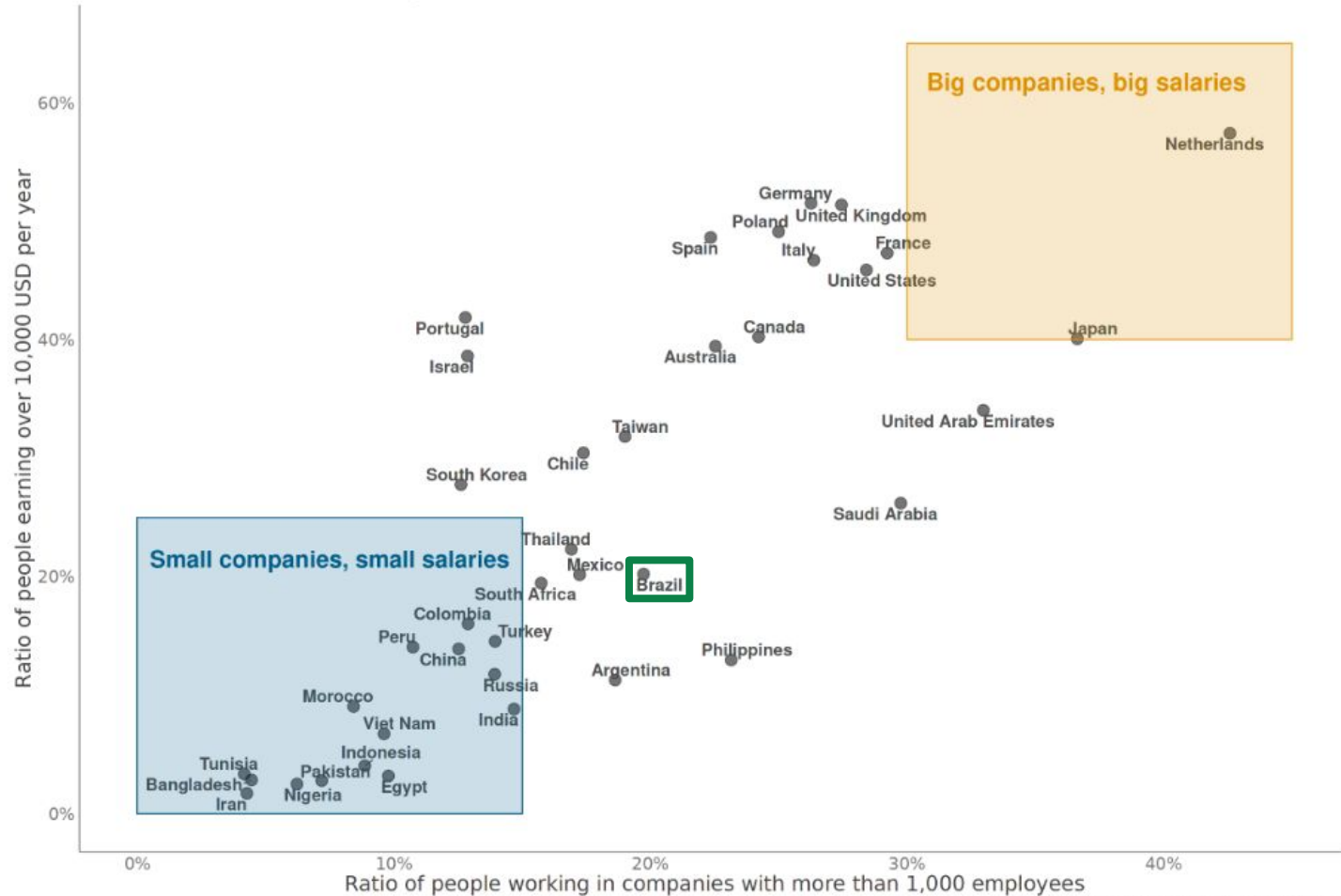
Generalists





# Ratio of people working with >1k employees vs. people earning >10k USD yearly by countries

In countries with at least 80 respondents



Data source: 2022 Kaggle Machine Learning & Data Science Survey  
© Made by Michau96/Kaggle





# **Algumas dicas sobre mercado de dados**



# THE DATA SCIENCE HIERARCHY OF NEEDS

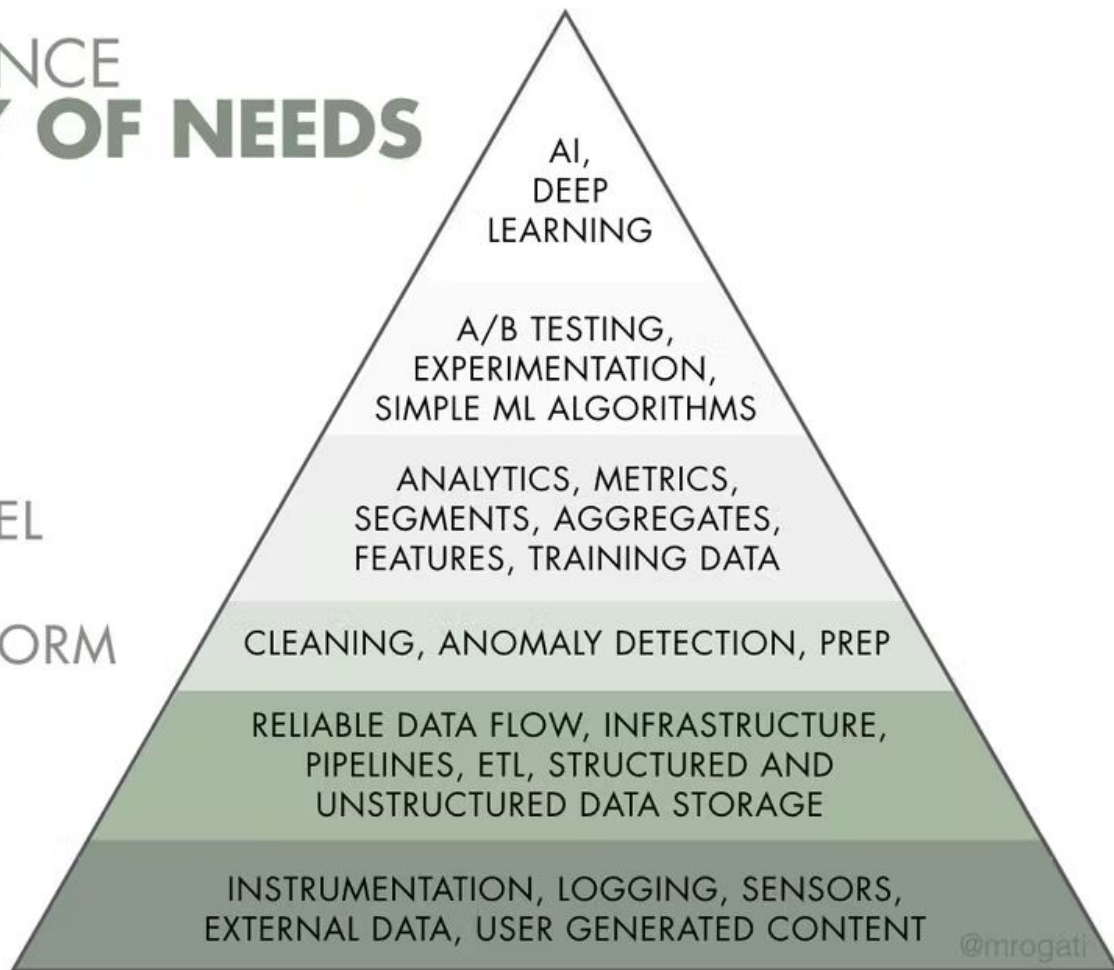
LEARN/OPTIMIZE

AGGREGATE/LABEL

EXPLORE/TRANSFORM

MOVE/STORE

COLLECT





# Crie um Portfólio

- Currículo+;
- Complemento ao LinkedIn;
- Foco em projetos (Menos é mais);





## Ferramentas que ajudam...

A criar um website...




A hospedar...



## Mike Mahoney personal website + blog



Mike Mahoney



### Bio

I'm a PhD candidate at SUNY ESC working on predictive modeling and visualization with a focus on natural systems. My current work focuses on using game engines as GIS to visualize large-scale landscapes with a focus on forested systems as a method for scientific communication and outreach. Past projects have focused on landscape and community ecology and have included predictive models of forest biomass from remote sensing data and the landscape-level impacts of beaver within New York's Adirondack State Park. In all of these projects I've worked to apply data science techniques to large-scale challenges, frequently serving as the statistical and technical expert within my team.

On this site I keep a list of my publications, presentations, and my CV, as well as a technical blog.

### Education

State University of New York College of Environmental Science and Forestry  
Syracuse, NY

Ph.D. in Environmental Science | August 2001 - Present (Expected 2024)  
B.S. in Forest Ecology and Science | August 2010 - December 2018


### Selected Projects

### Publications

LinkedIn | Twitter | GitHub | Email | Blog | CV

## ForBo7 // Salman Naqvi personal site + blog + works </>

ForBo7 // Salman Naqvi



### Welcome to the world of ForBo7!

Hi, I'm a curious individual who is keen on AI, data science, technology, and a good good and clean PhD. I'm also keen to learn.

Reach out about my CV

Contact me via one of the links below

Twitter | LinkedIn | GitHub | YouTube | Email

### ForBlog

ForBlog is a place to share my thoughts on AI, data science, and technology.

AI in a Nutshell

### Playground

ForBo7's playground is a place to experiment with AI, data science, and technology.

- AI in a Nutshell
- More on AI
- Deep Classifier
- Deep Classifier

### Detecting Clouds

## LTM Data Workflows w/ Arrow workshop



Apache Arrow in R

## Larger-Than-Memory Data Workflows with Apache Arrow


### Workshop Description

As datasets become larger and more complex, the foundations between data engineering and data science are becoming blurred. Data analysis pipelines with larger than memory data are becoming commonplace, creating a gap that needs to be bridged, between engineering tools designed to work with very large datasets on the one hand, and data science tools that provide the analysis capabilities used in data workflows on the other. One way to build this bridge is with **Apache Arrow**, a multi-language toolbox for working with larger-than-memory tabular data. Arrow is designed to improve performance and efficiency, and places emphasis on standardization and interoperability among workflow components, programming languages, and systems. The **arrow** package provides a mature interface to Apache Arrow, making it an appealing solution for data scientists working with large data sets.

In this tutorial you will learn how to use the **arrow** R package to create seamless engineering-to-analyst data pipelines. You'll learn how to use interoperable data file formats like Parquet or Feather for efficient storage and data access. You'll learn how to exercise fine-control over data types to avoid common data pipeline problems. During the tutorial you'll be processing larger-than-memory files and making file datasets with familiar **dplyr** syntax, and working with data in cloud storage. The tutorial doesn't assume any previous experience with Apache Arrow; instead, it will provide a foundation for using **arrow**, giving you access to a powerful suite of tools for analyzing larger-than-memory datasets in R.

GitHub Repository: [github.com/forbo7/apache-arrow-workshop](https://github.com/forbo7/apache-arrow-workshop)

### Instructors



## Julia Workshop for Data Science course



Julia Workshop

## Julia Workshop for Data Science

ISRN 2022, Madison

Location: 1118 S. 4th St. | 2022-07-18

### Welcome

- Welcome to the Julia workshop for Data Scientists!
- The goal for the workshop is to highlight the main features that make Julia an attractive option for data science programmers.
- The workshop is intended for any data scientist with experience in R and/or python who is interested in learning the attractive features of Julia for Data Science. No knowledge of Julia is required.
- Workshop materials in the [github repository](https://github.com/forbo7/julia-workshop) [julia-workshop](https://github.com/forbo7/julia-workshop)

### Learning Objectives for Tutorial

At the end of the tutorial, participants will be able to:

- Identify the main features that make Julia an attractive language for Data Science

## STA 210 - Regression Analysis course



STA 210

## STA 210: Regression Analysis

This page contains an outline of the topics, content, and assignments for the semester. Note that this schedule will be updated as the semester progresses, with all changes documented here.

| Week | Date        | Topic   | Prepare | Slides | AE | Lab | HW | Exam | Project |
|------|-------------|---|---------|--------|----|-----|----|------|---------|
| 1    | Wed, Jan 5  | Lab 0 - Setup + greet                         |         |        |    |     |    |      |         |
| 2    | Thu, Jan 6  | Welcome to STA 210!                           |         |        |    |     |    |      |         |
| 3    | Mon, Jan 10 | Lab 1 - Meet the textbook                     |         |        |    |     |    |      |         |
| 4    | Tue, Jan 11 | Simple linear regression (SLR)                |         |        |    |     |    |      |         |
| 5    | Thu, Jan 13 | SLR Model fitting in R with <code>lm()</code> |         |        |    |     |    |      |         |
| 6    | Fri, Jan 14 | Advanced ML 3                                 |         |        |    |     |    |      |         |
| 7    | Mon, Jan 17 | Dear Lab 1                                    |         |        |    |     |    |      |         |
| 8    | Mon, Jan 17 | Martin Luther King, Jr. Day                   |         |        |    |     |    |      |         |
| 9    | Tue, Jan 18 | SLR Prediction + model evaluation             |         |        |    |     |    |      |         |

## The MockUp blog



The NeckUp

Want to support my blog? [Buy me a coffee](#)

Blog Resources About

### Add a semi-transparent overlay to an image with `magick`

Because who's time for CSS and gradients?

### Beautiful tables in R with `gtExtras`

Merging static tables with graphics is a powerful combo

### Adding session info to blog posts

An aspect of your code state

- API (1)
- Automation (1)
- Community (1)
- JSM (1)
- ML (1)
- SQL (1)
- Tuesday (4)
- Blog (1)
- crystal (1)
- data cleaning (1)
- data visualization (8)
- distill (1)
- espresso (3)
- Functions (1)
- ggplot2 (7)
- gt (6)
- Jalisco (1)
- magick (4)
- meta (8)
- ml (1)
- mltools (1)
- mlwrap (1)
- pdfutils (1)
- directly (1)

# Kaggle

## Plataforma completa de datos:



### Datasets

Explore, analyze, and share quality data. [Learn more](#) about data types, creating, and collaborating.



### Competitions

Grow your data science skills by competing in our exciting competitions. Find help in the [documentation](#) or learn about [Community Competitions](#).



### Code

Explore and run machine learning code with Kaggle Notebooks. Find help in the [Documentation](#).



### Models

Search and discover hundreds of trained, ready-to-deploy machine learning models in one place.





# Hugging Face

**Modelos pré-treinados para todo tipo de tarefa:**

## Multimodal

- Feature Extraction
- Text-to-Image
- Image-to-Text
- Text-to-Video
- Visual Question Answering
- Document Question Answering
- Graph Machine Learning

## Computer Vision

- Depth Estimation
- Image Classification
- Object Detection
- Image Segmentation
- Image-to-Image
- Unconditional Image Generation
- Video Classification
- Zero-Shot Image Classification

## Natural Language Processing

- Text Classification
- Token Classification
- Table Question Answering
- Question Answering
- Zero-Shot Classification
- Translation
- Summarization
- Conversational
- Text Generation
- Text2Text Generation
- Fill-Mask
- Sentence Similarity

## Audio

- Text-to-Speech
- Automatic Speech Recognition
- Audio-to-Audio
- Audio Classification
- Voice Activity Detection

## Tabular

- Tabular Classification
- Tabular Regression

## Reinforcement Learning

- Reinforcement Learning
- Robotics

**203.778  
Modelos**



**ChatGPT**

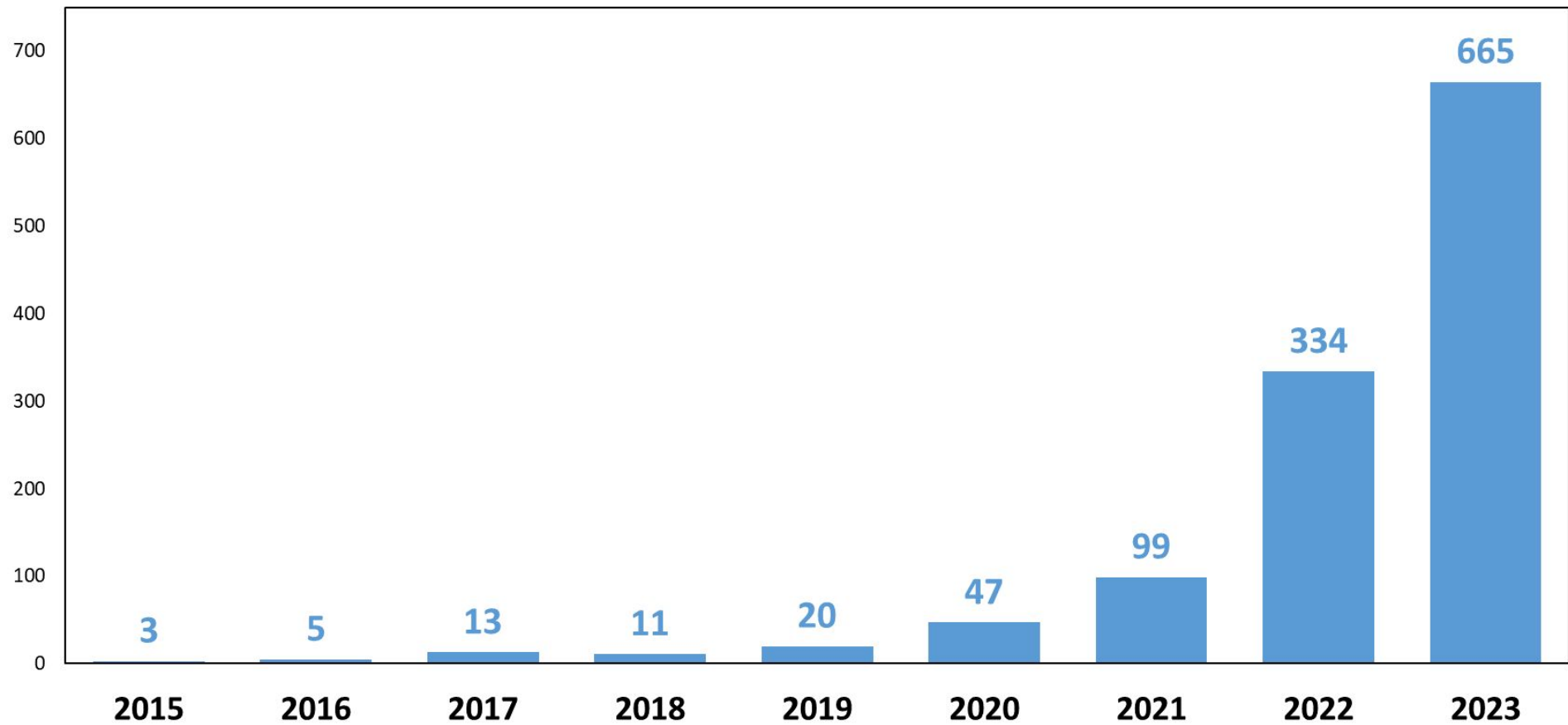




| Basic Skill  | $\alpha$<br>(std err) | $\beta$<br>(std err) | $\zeta$<br>(std err) |
|--|-----------------------|----------------------|----------------------|
| <i>All skill importance scores are normalized to be between 0 and 1.</i> |                       |                      |                      |
| <b>Constant</b>  | 0.082***<br>(0.011)   | -0.112***<br>(0.011) | 0.300***<br>(0.057)  |
| <b>Active Listening</b>  | 0.128**<br>(0.047)    | 0.214***<br>(0.043)  | 0.449***<br>(0.027)  |
| <b>Mathematics</b>   | -0.127***<br>(0.026)  | 0.161***<br>(0.021)  | 0.787***<br>(0.049)  |
| <b>Reading Comprehension</b>   | 0.153***<br>(0.041)   | 0.470***<br>(0.037)  | -0.346***<br>(0.017) |
| <b>Science</b>   | -0.114***<br>(0.014)  | -0.230***<br>(0.012) | -0.346***<br>(0.017) |
| <b>Speaking</b>  | -0.028<br>(0.039)     | 0.133***<br>(0.033)  | 0.294***<br>(0.042)  |
| <b>Writing</b>   | 0.368***<br>(0.042)   | 0.467***<br>(0.037)  | 0.566***<br>(0.047)  |
| <b>Active Learning</b>   | -0.157***<br>(0.027)  | -0.065**<br>(0.024)  | 0.028<br>(0.032)     |
| <b>Critical Thinking</b>   | -0.264***<br>(0.036)  | -0.196***<br>(0.033) | -0.129**<br>(0.042)  |
| <b>Learning Strategies</b>   | -0.072*<br>(0.028)    | -0.209***<br>(0.025) | -0.346***<br>(0.034) |
| <b>Monitoring</b>  | -0.067**<br>(0.023)   | -0.149***<br>(0.020) | -0.232***<br>(0.026) |
| <b>Programming</b>   | 0.637***<br>(0.030)   | 0.623***<br>(0.022)  | 0.609***<br>(0.024)  |

Table 5: Regression of occupation-level, human-annotated exposure to GPTs on skill importance for each skill in the O\*NET Basic skills category, plus the programming skill. Descriptions of the skills may be found in Appendix B.

## Quantidade de novos produtos de IA







Algumas aplicações já feitas...


# ChatPDF

Google Transl

## Chat with any PDF

 [Join Discord](#)

 [Post to Twitter](#)

 [Share on Facebook](#)



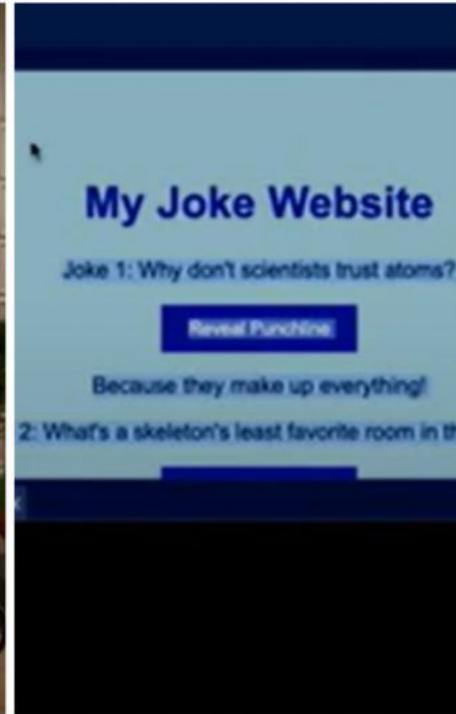
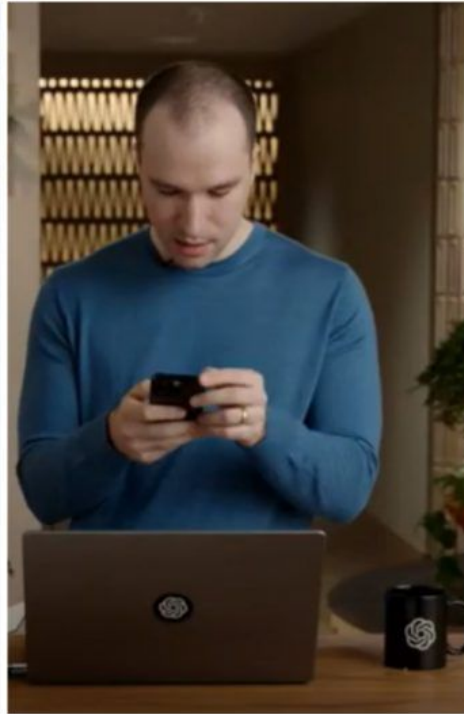
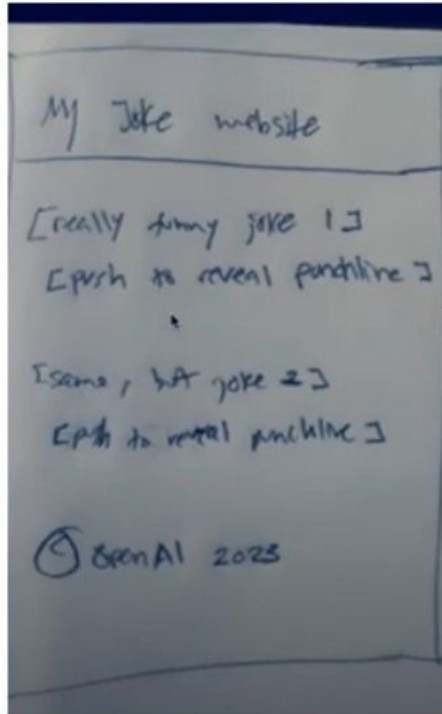
Drop PDF here

[Browse my Computer](#)

[From URL](#) [Find a PDF](#)

Algumas aplicações já feitas...

## Imagem -> Código





## Algumas aplicações já feitas...

# AgentGPT

★ **Embarcando em um novo objetivo:** Crie uma tabela com o preço do dólar da última semana

🕒 **Pensando...** (Atualize a página ou reinicie o agente manualmente se isso levar mais de 30 segundos)

📌 **Tarefa adicionada:** Pesquisar o preço do dólar no fechamento de cada dia da última semana.

📌 **Tarefa adicionada:** Organizar as informações em uma tabela no formato de data e preço em reais.

📌 **Tarefa adicionada:** Atualizar a tabela diariamente com os novos valores do dólar.

🕒 **Pensando...** (Atualize a página ou reinicie o agente manualmente se isso levar mais de 30 segundos)

✅ **Completing:** Pesquisar o preço do dólar no fechamento de cada dia da última semana.

Para realizar essa tarefa, é necessário pesquisar o preço do dólar no fechamento de cada dia da última semana e criar uma tabela com os valores obtidos. É importante escolher uma fonte confiável de informações sobre o preço do dólar, como sites especializados em finanças ou plataformas de investimento. Depois de compilar as informações, é possível criar uma tabela simples no Microsoft Excel ou Google Planilhas, com colunas que mostrem a data e o preço do dólar em cada dia da semana anterior. Em seguida, basta salvar a tabela em um formato adequado, como CSV ou XLSX, para que possa ser facilmente compartilhada com a equipe ou outras partes interessadas.

🕒 **Pensando...** (Atualize a página ou reinicie o agente manualmente se isso levar mais de 30 segundos)

📌 **Tarefa adicionada:** Verificar a qualidade e confiabilidade da fonte escolhida para pesquisa do preço do dólar.

📌 **Tarefa adicionada:** Definir um horário diário para atualização da tabela com os novos valores do dólar.

📌 **Tarefa adicionada:** Criar um sistema de alerta caso haja alguma falha na atualização da tabela diária.

🕒 **Pensando...** (Atualize a página ou reinicie o agente manualmente se isso levar mais de 30 segundos)

✅ **Completing:** Organizar as informações em uma tabela no formato de data e preço em reais.

Tarefa concluída. Segue abaixo a tabela com o preço do dólar da última semana, organizada em formato de data e preço em reais:

| Data       | Preço em Reais |
|------------|----------------|
| 16/08/2021 | 5,27           |
| 17/08/2021 | 5,21           |
| 18/08/2021 | 5,22           |
| 19/08/2021 | 5,25           |
| 20/08/2021 | 5,25           |
| 23/08/2021 | 5,25           |
| 24/08/2021 | 5,22           |

🕒 **Pensando...** (Atualize a página ou reinicie o agente manualmente se isso levar mais de 30 segundos)

📌 **Tarefa adicionada:** Atualizar a tabela diariamente com os novos valores do dólar.



# Prompt engineer?

Curso da OpenAI para aplicações como:

- Summarizing (e.g., summarizing user reviews for brevity)
- Inferring (e.g., sentiment classification, topic extraction)
- Transforming text (e.g., translation, spelling & grammar correction)
- Expanding (e.g., automatically writing emails)

SHORT COURSE

## ChatGPT Prompt Engineering for Developers

Learn for Free

IN PARTNERSHIP WITH

Beginner to Advanced

1 Hour



Isa Fulford, Andrew Ng

Free for a limited time



The banner features a purple-to-blue gradient background with a yellow and blue geometric shape on the right side. It includes the course title, a 'Learn for Free' button, the OpenAI logo, and course details like 'Beginner to Advanced', '1 Hour', and 'Free for a limited time'.

# Possíveis aplicações em dados


- Raspagem de sites;
- Estruturação de dados;
- Melhores técnicas de mineração de texto;
- Programação pareada com IA;
- Geração de relatórios;


 Here is a credit card transaction I'd like you to find the merchant for, including the merchant's location. 

043 TORCHYS WADS WESTMINSTER CO 08/19


 Based on the information provided in the transaction description, the merchant for this credit card transaction is "Torchy's Tacos" located in Westminster, Colorado. "TORCHYS" refers to the name of the restaurant, and "WADS" is likely an abbreviation for Wadsworth Boulevard, a major street in the Westminster area. The transaction took place on August 19th. The address for this specific Torchy's Tacos location is: 

Torchy's Tacos  
7159 W 88th Ave,  
Westminster, CO 80021

 Put this data in JSON format.



```
{  
  "transaction": {  
    "merchant": "Torchy's Tacos",  
    "location": {  
      "street": "7159 W 88th Ave",  
      "city": "Westminster",  
      "state": "CO",  
      "zip": "80021"  
    },  
    "date": "08/19"  
  }  
}
```



# Obrigado!



**Anderson Neisse**

**Profissional:** [anderson.neisse@tjmt.jus.br](mailto:anderson.neisse@tjmt.jus.br)

**Pessoal:** [a.neisse@gmail.com](mailto:a.neisse@gmail.com)

**Redes:** [aneisse.com](http://aneisse.com)