DIS

Machine learning *in the wild*

Tales from machine learning after college

2nd Edition

DIS 11/06/2023
Duarte O.Carmo
duarteocarmo.com - @duarteocarmo

DIS

Hello! I'm Duarte.

/du-art/ - it's Portuguese

ML/Software Engineer & contractor

From Portugal, based in Copenhagen, Denmark

I like **running**, and **writing** on my blog

Past: Strategy, Product Mgmt., New Ventures, Mgmt. Consulting

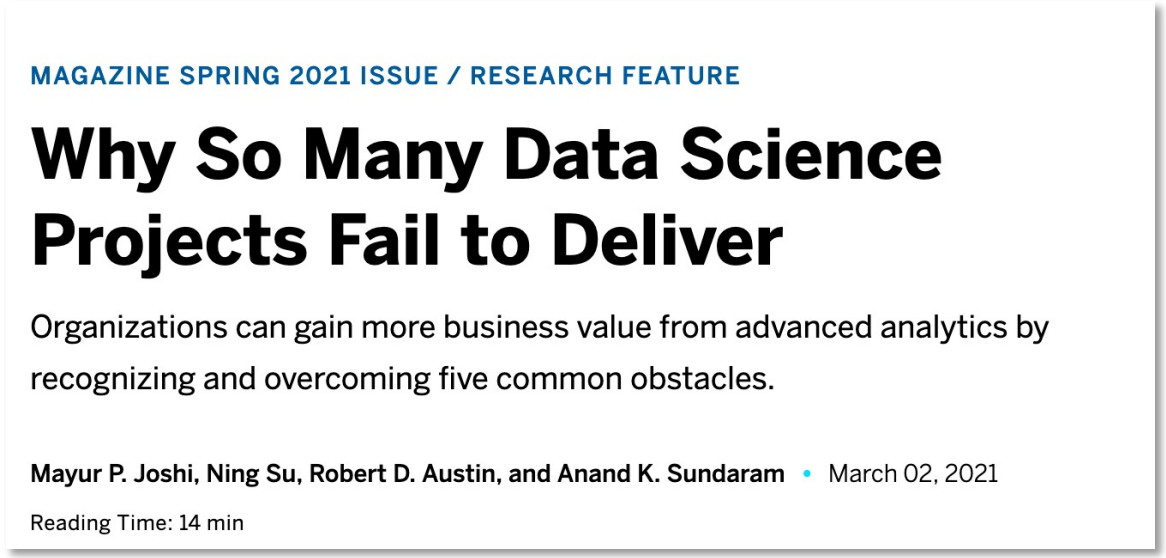
Now: I help companies solve **tough** problems end-to-end

DIS

Today, we'll talk about machine learning from what I've seen out there

- How (I think) ML engineers should work
- 3 example problems from the wild
- "MLOps"
- Learning

- *Opinions*
- *Experiences*



MAGAZINE SPRING 2021 ISSUE / RESEARCH FEATURE

Why So Many Data Science Projects Fail to Deliver

Organizations can gain more business value from advanced analytics by recognizing and overcoming five common obstacles.

Mayur P. Joshi, Ning Su, Robert D. Austin, and Anand K. Sundaram • March 02, 2021

Reading Time: 14 min

DIS

1 | How I work

DIS

“We need a model”

(you probably don’t)

DIS

Don’t start with models, start with people

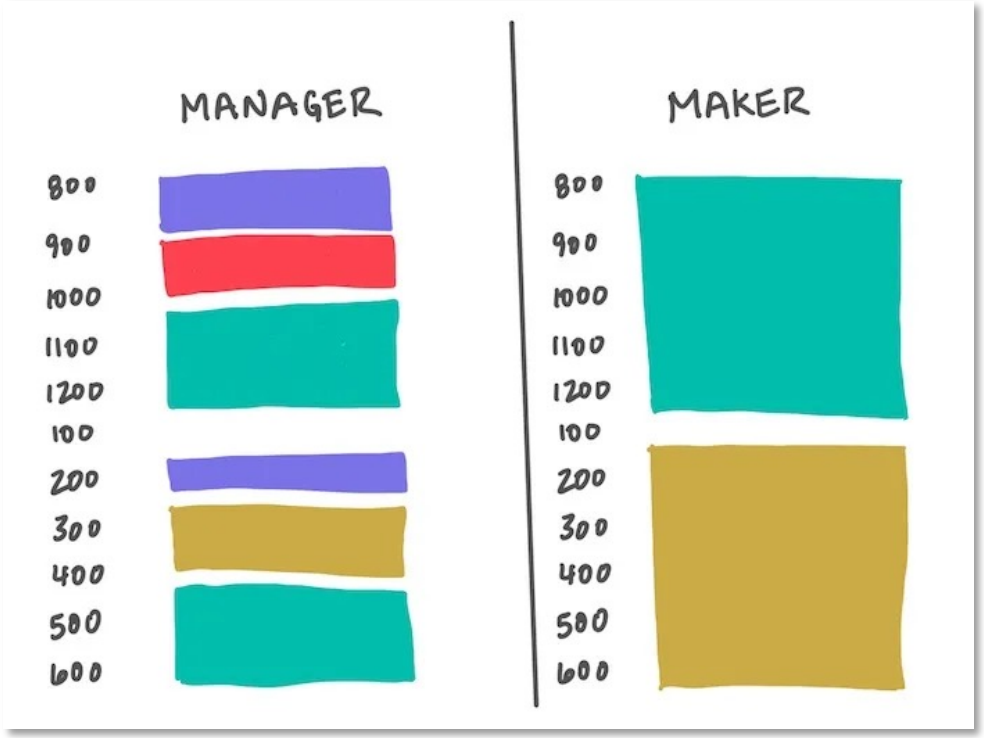
- Define the business goal, and the success metric
- This is real world (bad) data - not Kaggle: cr*p in, cr*p out
- Start with heuristics, and increase complexity as needed
- Put it out there as fast as possible, then iterate

DIS

Go end-to-end, early!


Don’t build in the basement

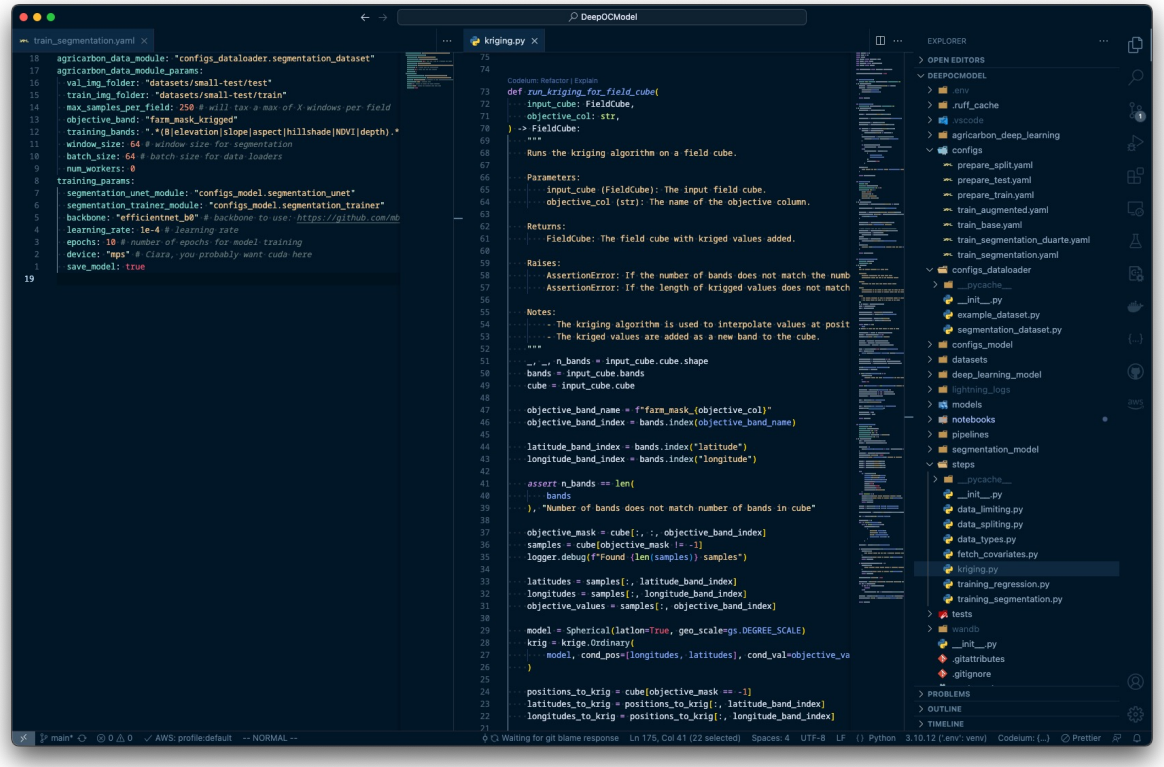
You are makers at heart – and should treat your schedules like it



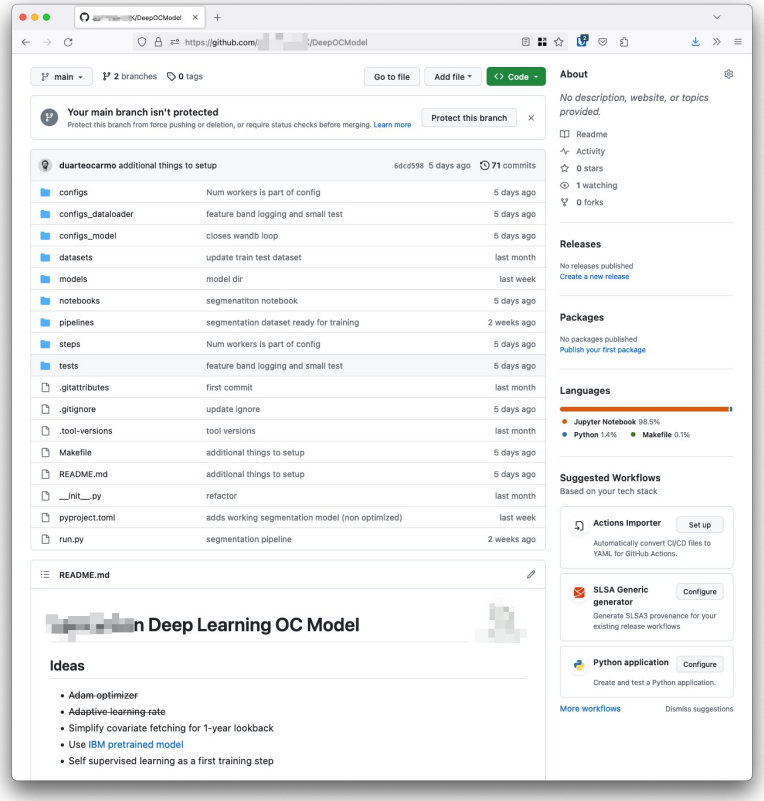
- Minimize time in meetings and double down on communication
- Fridays = no meetings
- We are on an emerging tech field, studying is important
- We are builders of things, disruptions are not welcome

DIS

Tools are irrelevant, until you have to use them every day

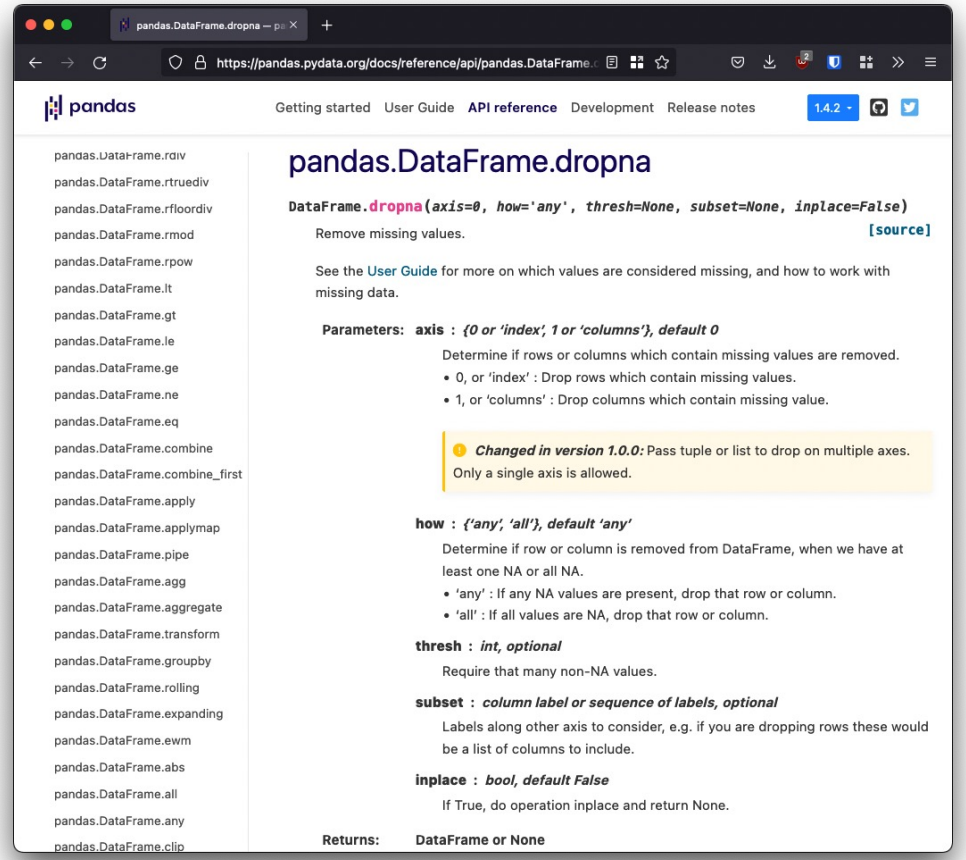


VSCode - (I actually use NeoVim)



GitHub, some people prefer GitLab

But there’s quite nothing like reading



- What does it do?
- Options?
- Default behaviors
- Maybe I can re-use this
- **It *actually* sticks**

Use AI to code

(But not at the expense of your brain)

DIS

Google Cloud

DEVELOPER'S CHEAT SHEET

Created by the Google Developer Relations Team
Maintained at https://4words.dev

Feedback? @pvergadia @GoogleCloudTech

OMPUTE

ud Functions	Event-driven serverless functions
Engine	Managed app platform
ud Run	Serverless for containerized applications
ernetes Engine (GKE)	Managed Kubernetes/containers
mpute Engine	VMs, GPUs, TPUs, Disks
e Metal Solution	Hardware for specialized workloads
emptible VMs	Short-lived compute instances
elded VMs	Hardened VMs
e-tenant Nodes	Dedicated physical servers
ware Engine	VMware on Compute Engine

TORAGE

ud Filestore	Managed NFS server
ud Storage	Multi-class multi-region object storage
sistent Disk	Block storage for VMs
al SSD	VM locally attached SSDs

ATABASE

ud Bigtable	Petabyte-scale, low-latency, non-relational
ud Firestore	Serverless NoSQL document DB
ud Memorystore	Managed Redis and Memcached
ud Spanner	Horizontally scalable relational DB
ud SQL	Managed MySQL,PostgreSQL,SQL Server
abase Migration Service	Migrate to Cloud SQL
Insights	SQL Inspector

ATA ANALYTICS

Query	Data warehouse/analytics
Query BI Engine	In-memory analytics engine
Query ML	BigQuery model training/serving
Query GIS	BigQuery geospatial functions/support
Query DTS	Automated data ingestion service
nnected Sheets	Spreadsheet interface for (big)data
ud Composer	Managed workflow orchestration service
a Fusion	Graphically manage data pipelines
aflow	Stream/batch data processing
aprep by Trifacta	Visual data wrangling
aproc	Managed Spark and Hadoop
astream	Change data capture/replication service
/Sub	Global real-time messaging
a Catalog	Metadata management service
a Studio	Collaborative data exploration/dashboarding
ker	Enterprise BI and Analytics
lic Datasets	Hosted data in BigQuery/GCS

YBRID AND MULTI-CLOUD

hos	Enterprise hybrid/multi-cloud platform
hos Clusters	Hybrid/on-prem Kubernetes Engine
hos Config Management	Policy and security automation
hos Service Mesh	Managed service mesh (Istio)
ud Run for Anthos	Servlerless development for Anthos
Marketplace for Anthos	Pre-configured containerized apps
rate for Anthos	Migrate VMs to Kubernetes Engine

Vertex Explainable AI	Understand ML model predictions
Vertex AI Feature Store	Managed ML feature repository
Vertex ML Metadata	Artifact, lineage, and execution tracking
Vertex AI Model Monitoring	Monitor models for skew/drift
Vertex AI Tensorboard	Managed TensorBoard for ML-experiment Visualization
Vertex AI Vizier	Black-box hyperparameter tuning
Cloud Speech-To-Text API	Convert audio to text
Cloud Talent Solutions API	Job search with ML
Cloud Text-To-Speech API	Convert text to audio
Cloud TPU	Hardware acceleration for ML
Cloud Translation API	Language detection and translation
Cloud Video Intelligence API	Scene-level video annotation
Cloud Vision API	Image recognition and classification
Contact Center AI	AI in your contact center
Dialogflow	Create conversational interfaces
Document AI	Analyze, classify, search documents
Recommendations AI	Create custom recommendations
Vision Product Search	Visual search for products

NETWORKING

Carrier Peering	Peer through a carrier
Direct Peering	Peer with GCP
Dedicated Interconnect	Dedicated private network connection
Partner Interconnect	Connect on-prem network to VPC
Cloud Armor	DDoS protection and WAF
Cloud CDN	Content delivery network
Cloud DNS	Programmable DNS serving
Cloud Load Balancing	Multi-region load distribution/balancing
Cloud NAT	Network address translation service
Cloud Router	VPC/on-prem network route exchange (BGP)
Cloud VPN (HA)	VPN (Virtual private network connection)
Network Service Tiers	Price vs performance tiering
Network Telemetry	Network telemetry service
Traffic Director	Service mesh traffic management
Google Cloud Service Mesh	Service-aware network management
Virtual Private Cloud	Software defined networking
VPC Service Controls	Security perimeters for API-based services
Network Intelligence Center	Network monitoring and topology

IDENTITY AND SECURITY

Access Transparency	Audit cloud provider access
Assured Workloads	Workload compliance controls
Binary Authorization	Kubernetes deploy-time security
Certificate Authority Service	Managed private CAs
Cloud Asset Inventory	All assets, one place
Cloud Audit Logs	Audit trails for GCP
Cloud DLP	Classify and redact sensitive data
Cloud HSM	Hardware security module service
Cloud EKM	External keys you control
Cloud IAM	Resource access control
Cloud Identity	Manage users, devices & apps
Cloud Identity-Aware Proxy	Identity-based app access
Cloud KMS	Hosted key management service
Cloud Resource Manager	Cloud project metadata management
Security Command Center	Security management & data risk platform
Cloud Security Scanner	App engine security scanner
Confidential Computing	Encrypt data in-use
Context-aware Access	End-user attribute-based access control
Event Threat Detection	Scans for suspicious activity
Managed Service for Microsoft Active Directory	Managed Microsoft Active Directory
Secret Manager	Store and manage secrets
Security Key Enforcement	Two-step key verification
Shielded VMs	Hardened VMs
Titan Security Key	Two-factor authentication (2FA) device
VPC Service Controls	VPC data constraints
Chronicle	Find threats from security telemetry
VirusTotal	Research/hunt for Malware
Risk Manager	Evaluate organization's security posture
reCPATCHA Enterprise	Protection against bot/spam/abuse
BeyondCorp Enterprise	Zero trust secure access
Access Context Manager	Fine-grained, attribute based access-control
Web Security Scanner	Identifies web-app security vulnerabilities

OPERATIONS & MONITORING

Cloud Debugger	Live production debugging
Error Reporting	App error reporting

Cloud Tasks	Asynchronous task execution
Cloud Workflows	HTTP services orchestration
Pub/Sub	Global real-time messaging

API PLATFORM AND ECOSYSTEMS

API Analytics	API metrics
API Monetization	Monetize APIs
Apigee API Platform	Develop, secure, monitor APIs
API Gateway	Fully managed API Gateway
Apigee Hybrid	Manage hybrid/multi-cloud API environments
Apigee Sense	API protection from attacks
Cloud Endpoints	Cloud API gateway
Developer Portal	API management portal
Marketplace	Partner & open source marketplace
AppSheet	No-code App creation

INTERNET OF THINGS (IOT)

Cloud IoT Core	Manage devices, ingest data

GAMING

Google Cloud Game Servers	Orchestrate Agones clusters

HEALTHCARE

Cloud Healthcare API	Healthcare system GCP interoperability
Apigee Healthcare APIx	Healthcare system GCP interoperability
Healthcare Natural Language AI	Real-time insights from media-text
Cloud Life Sciences	Manage, process, transform biomedical-data

RETAIL

Vision Product Search	Visual search for products
Recommendations AI	Create custom recommendations
Visual Inspection AI	Train/deploy models to detect defects

MANAGEMENT TOOLS

VM Manager	Manage OS VM Fleets
Cloud APIs	APIs for cloud services
Cloud Billing API	Programmatically manage GCP billing
Cloud Billing	Billing and cost management tools
Cloud Console	Web-based management console
Cloud Deployment Manager	Templated infrastructure deployment
Cloud Mobile App	iOS/Android GCP manager app
Private Catalog	Internal Solutions Catalog

DEVELOPER TOOLS

Cloud Code for IntelliJ	IntelliJ GCP tools
Cloud Code for VS Code	VS Code GCP tools
Cloud Code	Cloud native IDE extensions
Cloud Tools for Eclipse	Eclipse GCP tools
Cloud Tools for Visual Studio	Visual Studio GCP tools
Gradle App Engine Plugin	Gradle App Engine plugin
Maven App Engine Plugin	Maven App Engine plugin
Cloud SDK	CLI for GCP
Cloud Shell	Browser-based terminal/CLI

MIGRATION TO GOOGLE CLOUD

BigQuery Data Transfer Service	Bulk import analytics data
Cloud Data Transfer	Data migration tools/CLI
Google Transfer Appliance	Rentable data transport box
Storage Transfer Service	Online/on-premises data transfer
Migrate for Anthos	Migrate VMs to GKE containers
Migrate for Compute Engine	Compute Engine migration tools
Migrate from Amazon Redshift	Migrate from Redshift to BigQuery
Migrate from Teradata	Migrate from Teradata to BigQuery
Cloud Foundation Toolkit	Infrastructure as Code templates
KF	Cloud Foundry to Kubernetes

GOOGLE MAPS PLATFORM

Directions API	Get directions between locations
Distance Matrix API	Multi-origin/destination travel times
Geocoding API	Convert address to/from coordinates
Geolocation API	Derive location without GPS
Maps Embed API	Display iframe embedded maps
Maps JavaScript API	Dynamic web maps
Maps SDK for Android	Maps for Android apps
Maps SDK for iOS	Maps for iOS apps
Maps Static API	Display static map images
Maps SDK for Unity	Unity SDK for games

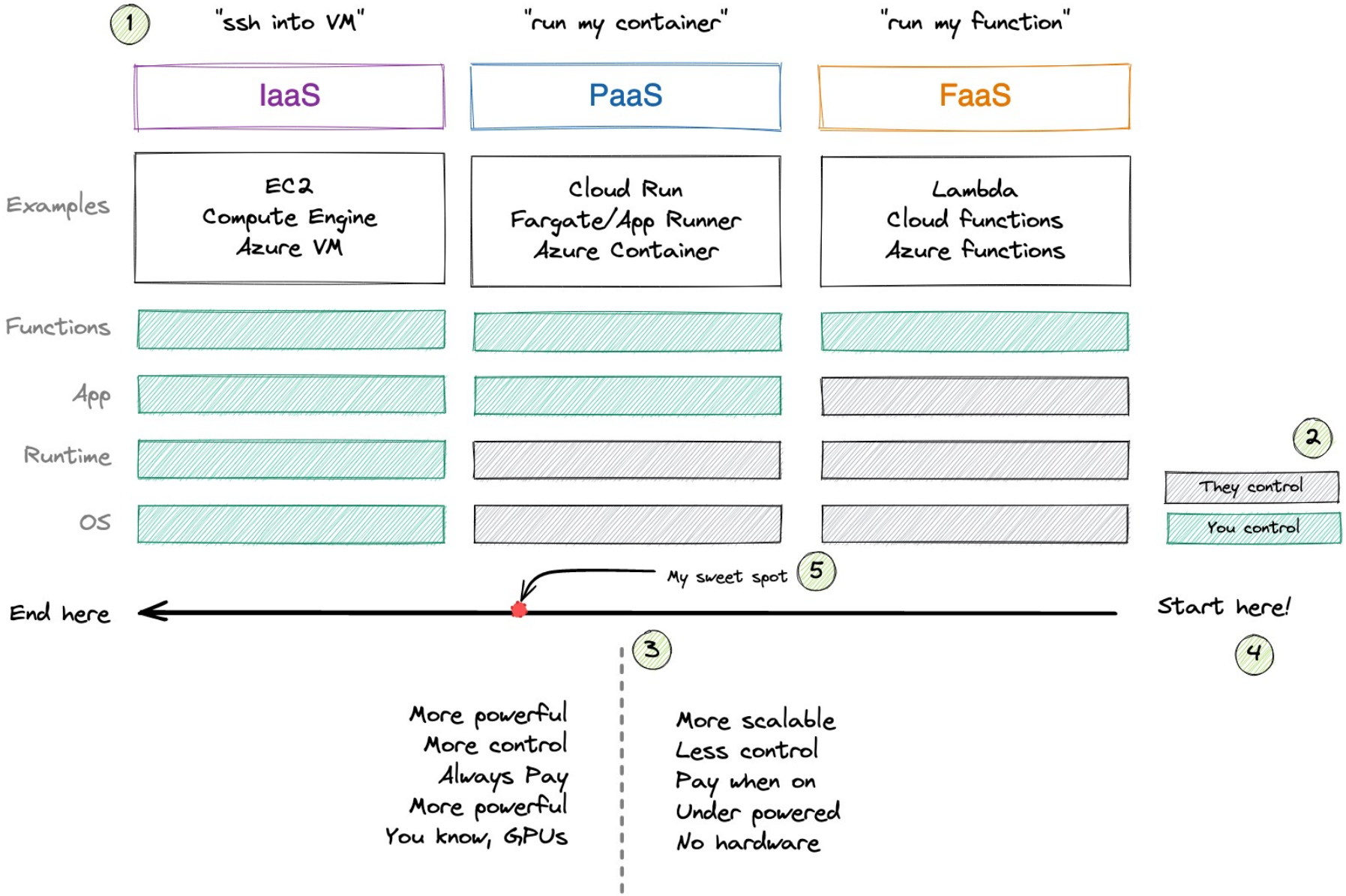
Cloud Search	Unified search for enterprise
Docs API	Create and edit documents
Drive Activity API	Retrieve Google Drive activity
Drive API	Read and write files
Drive Picker	Drive file selection widget
Email Markup	Interactive email using schema.org
Google Workspace Add-ons	Extend Google Workspace apps
Google Workspace Marketplace	Storefront for integrated applications
Gmail API	Enhance Gmail
Google Chats API	Conversational bots in chat
People API	Manage user's Contacts
Sheets API	Read and write spreadsheets
Slides API	Create and edit presentations
Task API	Search, read & update Tasks
Vault API	Manage your organization's eDiscovery

MOBILE (FIREBASE)

Cloud Firestore	Document store and sync
Cloud Functions for Firebase	Event-driven serverless applications
Cloud Storage for Firebase	Object storage and serving
Crashlytics	Crash reporting and analytics
Firebase A/B Testing	Create A/B test experiments
Firebase App Distribution	Trusted tester early access
Firebase Authentication	Drop-in authentication
Firebase Cloud Messaging	Send device notifications
Firebase Dynamic Links	Link to app content
Firebase Extensions	Pre-packaged development solutions
Firebase Hosting	Web hosting with CDN/SSL
Firebase In-App Messaging	Send in-app contextual messages
Firebase Performance Monitoring	App/web performance monitoring
Firebase Predictions	Predict user targeting
Firebase Realtime Database	Real-time data synchronization
Firebase Remote Config	Remotely configure installed apps
Firebase Test Lab	Mobile testing device farm
Google Analytics for Firebase	Mobile app analytics
ML Kit for Firebase	ML APIs for mobile

ADDITIONAL RESOURCES

Google Cloud Home Page	cloud.google.com
Google Cloud Blog	cloud.google.com/blog
Google Cloud Platform Podcast	gcppodcast.com
Kubernetes Podcast from Google	kubernetespodcast.com
Google Cloud Reader	podcasts.google.com
Google Cloud Open Source	opensource.google/projects/list/cloud
GCP Medium Publication	medium.com/google-cloud
Apigee Blog	apigee.com/about/blog
Firebase Blog	firebase.googleblog.com
Google Workspace Developers Blog	gsuite-developers.googleblog.com
Google Workspace GitHub	github.com/gsuitedevs
Google Workspace Twitter	twitter.com/gsuitedevs
Google Cloud Certifications	cloud.google.com/certification
Google Cloud System Status	status.cloud.google.com
Google Cloud Training	cloud.google.com/training
Google Developers Blog	developers.googleblog.com
Google Maps Platform Blog	mapsplatform.googleblog.com
Google Open Source Blog	opensource.googleblog.com
Google Security Blog	security.googleblog.com
Kaggle Home Page	www.kaggle.com
Kubernetes Blog	kubernetes.io/blog
Regions and Network Map	cloud.google.com/about/locations
DORA - Software & Delivery Research	cloud.google.com/devops
Cloud Security Podcast	cloud.withgoogle.com/cloudsecurity/podcast
GCP Sketchnote	goo.gle/gcpsketchnote
Google Cloud Solutions Library	cloud.google.com/solutions
Google Workspace Solutions Gallery	developers.google.com/gsuite/solutions
Google Cloud Support Hub	cloud.google.com/support-hub
GCP Pricing	cloud.google.com/pricing
GCP Pricing Calculator	cloud.google.com/products/calculator
Qwiklabs Home Page	www.qwiklabs.com
Codelabs Home Page	codelabs.developers.google.com
Reddit - www.reddit.com/	
Googlecloud	/r/googlecloud
AppEngine	/r/AppEngine
BigQuery	/r/bigquery
Dataflow	/r/dataflow
Firebase	/r/firebase


1
"ssh into VM"
"run my container"
"run my function"
IaaS
PaaS
FaaS
Examples
EC2
Compute Engine
Azure VM
Cloud Run
Fargate/App Runner
Azure Container
Lambda
Cloud functions
Azure functions
Functions
App
Runtime
OS
2
They control
You control
My sweet spot
5
End here
Start here!
3
4
More powerful
More control
Always Pay
More powerful
You know, GPUs
More scalable
Less control
Pay when on
Under powered
No hardware

2 | Problems

2.1 | Machine learning in Space

Geospatial data: All data generated from observing our planet in space

duarteocarmo.com - @duarteocarmo

Images: RGB bands from the Sentinel-2

Temperature: From multi/hyper spectral satellites

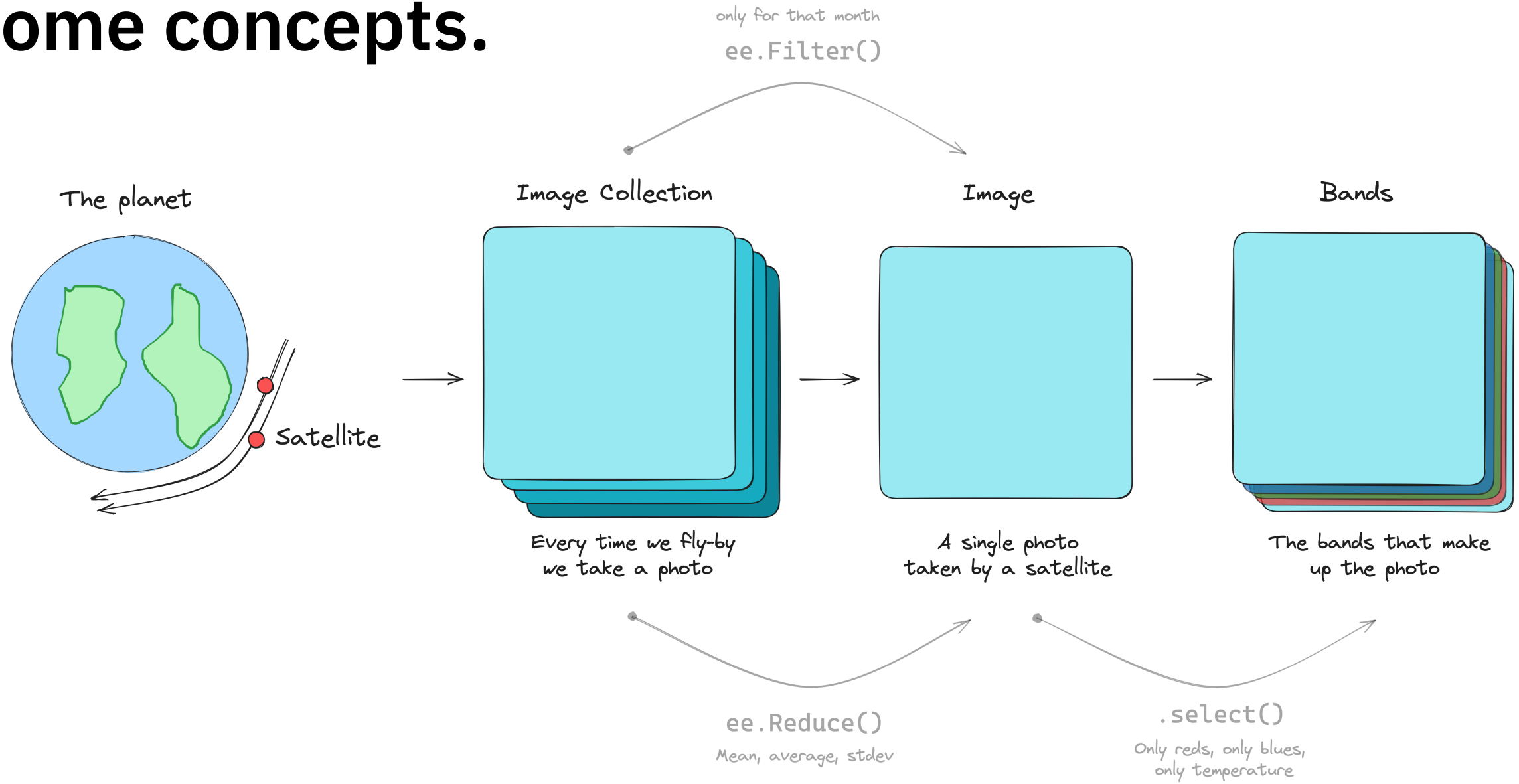
Vegetation: NDVI (vegetation index)

Landcover: Is this a city? Farm?

Precipitation: How much rainfall?

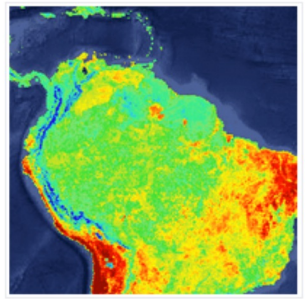
...

Some concepts.



DIS

MOD11A2.061 Terra Land Surface Temperature and Emissivity 8-Day Global 1km



Dataset Availability

2000-02-18T00:00:00Z–2023-07-04T00:00:00

For when is this data available?

Dataset Provider

NASA LP DAAC at the USGS EROS Center

Earth Engine Snippet

Image Collection or Image?

```
ee.ImageCollection("MODIS/061/MOD11A2")
```

Tags

8-day emissivity global lst mod11a2 modis nasa surface-temperature terra usgs

Description of all bands

Description | Bands | Terms of Use | Citations | DOIs

Resolution
1000 meters

What is the resolution?

Bands

Band details

Name	Units	Min	Max	Scale	Offset	Description
LST_Day_1km	K	7500	65535	0.02		Day land surface temperature
QC_Day						Daytime LST quality indicators
⊕ Bitmask for QC_Day						
Day_view_time	h	0	240	0.1		Local time of day observation
Day_view_angl	deg	0	130		-65	View zenith angle of day observation
LST_Night_1km	K	7500	65635	0.02		Night land surface temperature
QC_Night						Nighttime LST quality indicators
⊕ Bitmask for QC_Night						
Night_view_time	h	0	240	0.1		Local time of night observation

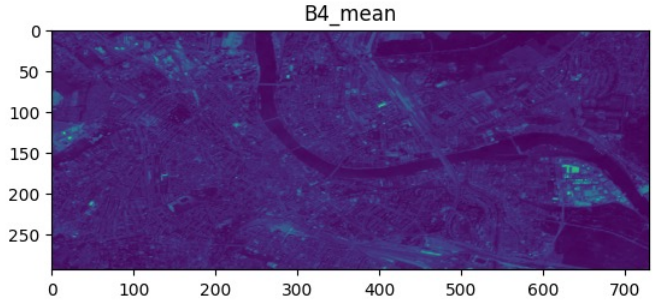
Tabular/Timeseries | Using Google Earth Engine to extract rainfall data as a time series

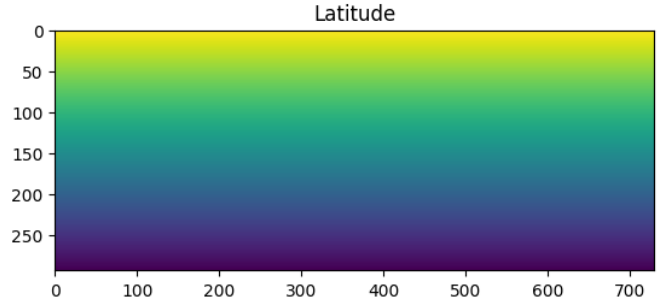
	country	long	lat	maincause	displaced	date	0_30d_t0_rainfall_m	1_30d_t1_rainfall_m	2_30d_t2_rainfall_m	3_30d_t3_rainfall_m
4967	India	79.226600	17.344400	Monsoonal Rain	2100	2020-10-16 00:00:00	0.619084	1.037469	0.242630	0.302843
4964	Vietnam	107.833000	15.854800	Heavy Rain	900000	2020-10-06 00:00:00	NaN	NaN	NaN	NaN
4963	Mozambique	38.127600	-14.581300	Heavy Rain	4000	2020-10-02 00:00:00	NaN	NaN	NaN	NaN
4966	Togo	0.777334	9.710160	Heavy Rain	16000	2020-09-15 00:00:00	0.099871	0.075639	0.060188	0.513882
4965	Nigeria	6.258020	7.774120	Heavy Rain	25000	2020-09-15 00:00:00	4.037949	1.330742	4.710331	3.221587
4962	USA	-86.579600	32.650800	Tropical Storm Sally	0	2020-09-15 00:00:00	0.017025	0.019261	0.060149	0.044176
4961	India	95.037700	27.848600	Monsoonal Rain	0	2020-09-13 00:00:00	NaN	NaN	NaN	NaN
4950	Afghanistan	67.720900	34.916000	Torrential Rain	0	2020-08-25 00:00:00	0.843814	0.881901	0.289106	0.776603
4951	Pakistan	68.215700	26.957300	Monsoonal Rain	1300	2020-08-24 00:00:00	0.730026	2.174792	0.494846	0.542660
4960	USA	-70.941700	19.564200	Tropical Storm Laura	600000	2020-08-22 00:00:00	0.391254	0.358152	0.596996	0.543690
4954	Haiti	-70.904600	18.907900	Tropical Storm Laura	0	2020-08-21 00:00:00	0.493560	0.363616	0.596426	0.601292
4959	India	77.469800	24.709500	Monsoonal Rain	60	2020-08-21 00:00:00	0.195524	0.190712	0.053476	0.149555
4953	Kenya	37.202100	2.750120	Dam Release and Heavy Rain	5000	2020-08-20 00:00:00	NaN	NaN	NaN	NaN
4955	Uganda	30.333400	0.385265	Torrential Rain	0	2020-08-19 00:00:00	0.006394	0.006034	0.010361	0.001781
4952	Chad	14.646100	10.199000	Heavy Rain	38000	2020-08-10 00:00:00	0.609008	0.030559	0.008041	0.000344

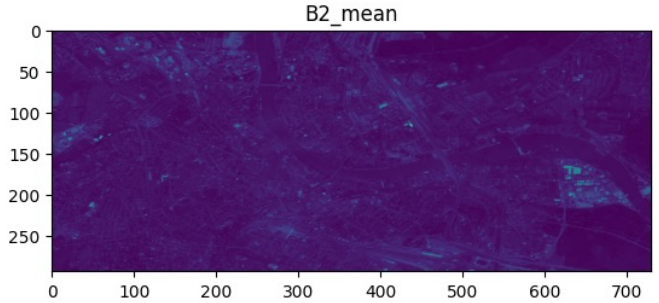
Rainfall (30d periods from observation date)

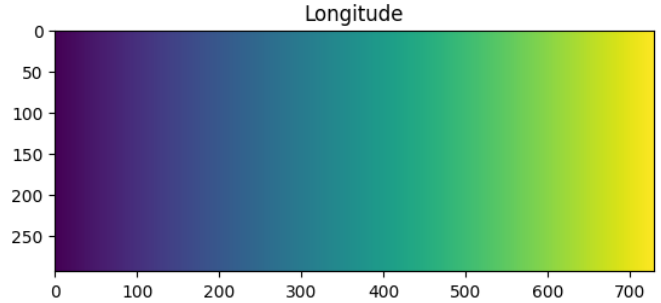
DIS

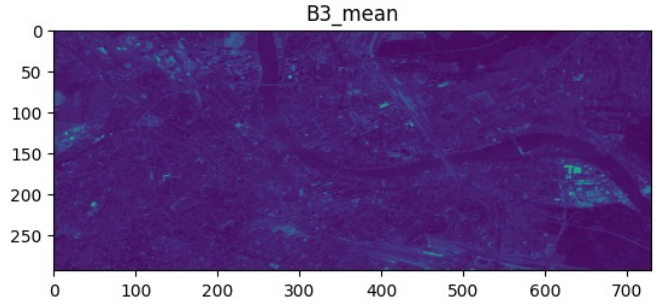
Computer Vision | From coordinates to images for Computer Vision applications

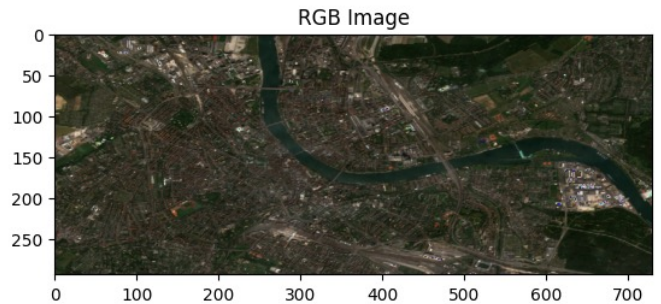












- ~275x750 images
- 1 pixel = 15 meters
- Lat and Long images
- RGB reconstructions

DIS

2.2 | Recommendations with embeddings

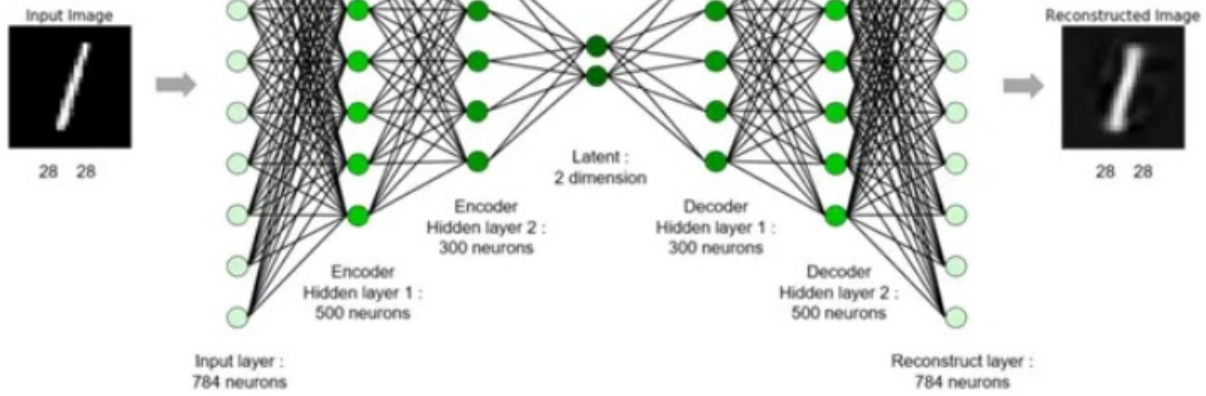
Helping sales teams find their ideal customers

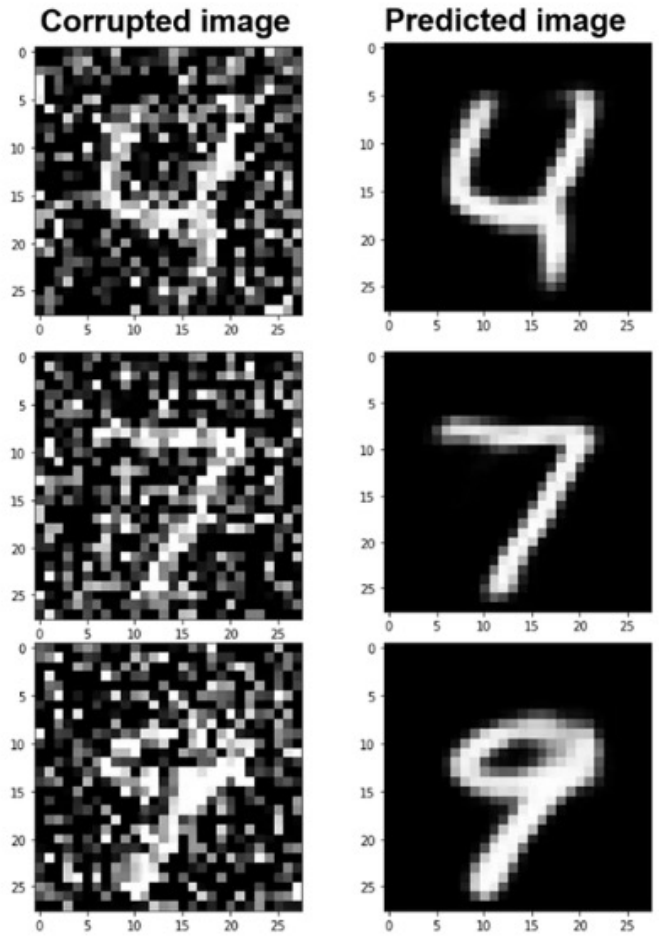
- Lead qualification is manual
- Lots of time spent qualifying
- How can we support this process?

Company Name	Description	Potential Customer?
Novo Nordisk	The Novo nordisk foun..	✅
Facebook	A social media..	❌
Budweiser	We are a bever..	✅
Nike	World leader in..	✅
Google	At Google, we're..	❌
...	...	...

DIS

First, a quick introduction to embeddings

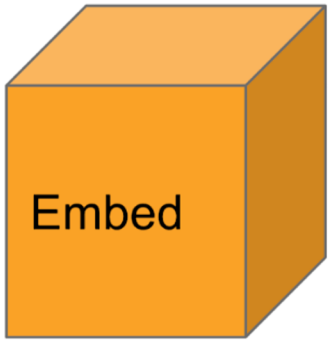




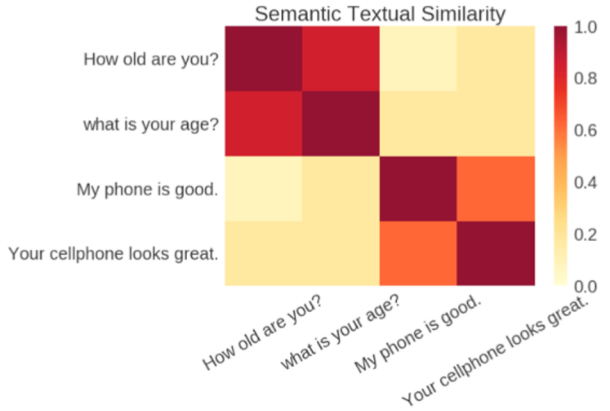
DIS

There are a lot of ways to use embeddings in real-world ML problems

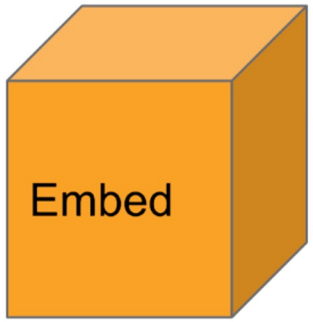
"How old are you?"
"What is your age?"
"My phone is good."
...



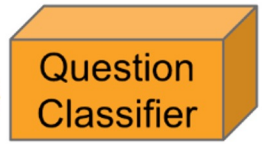
[0.3, 0.2, …]
[0.2, 0.1, …]
[0.9, 0.6, …]
...



"How old are you?"
"What is your age?"
"My phone is good."
...



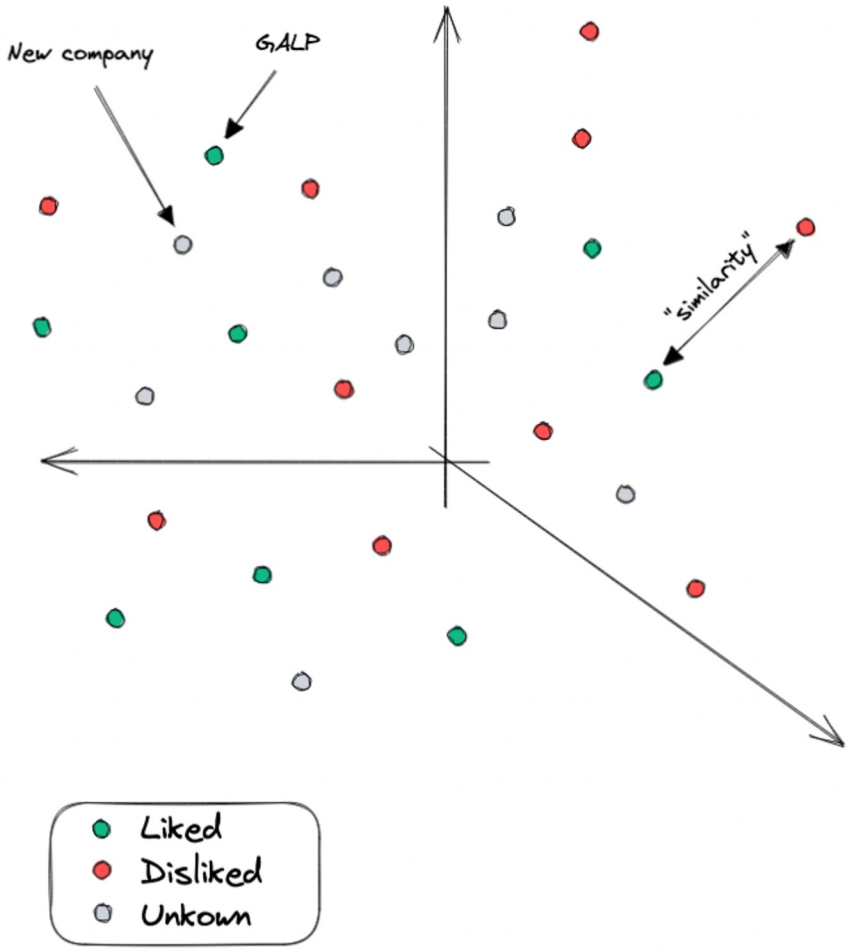
[0.3, 0.2, …]
[0.2, 0.1, …]
[0.9, 0.6, …]
...



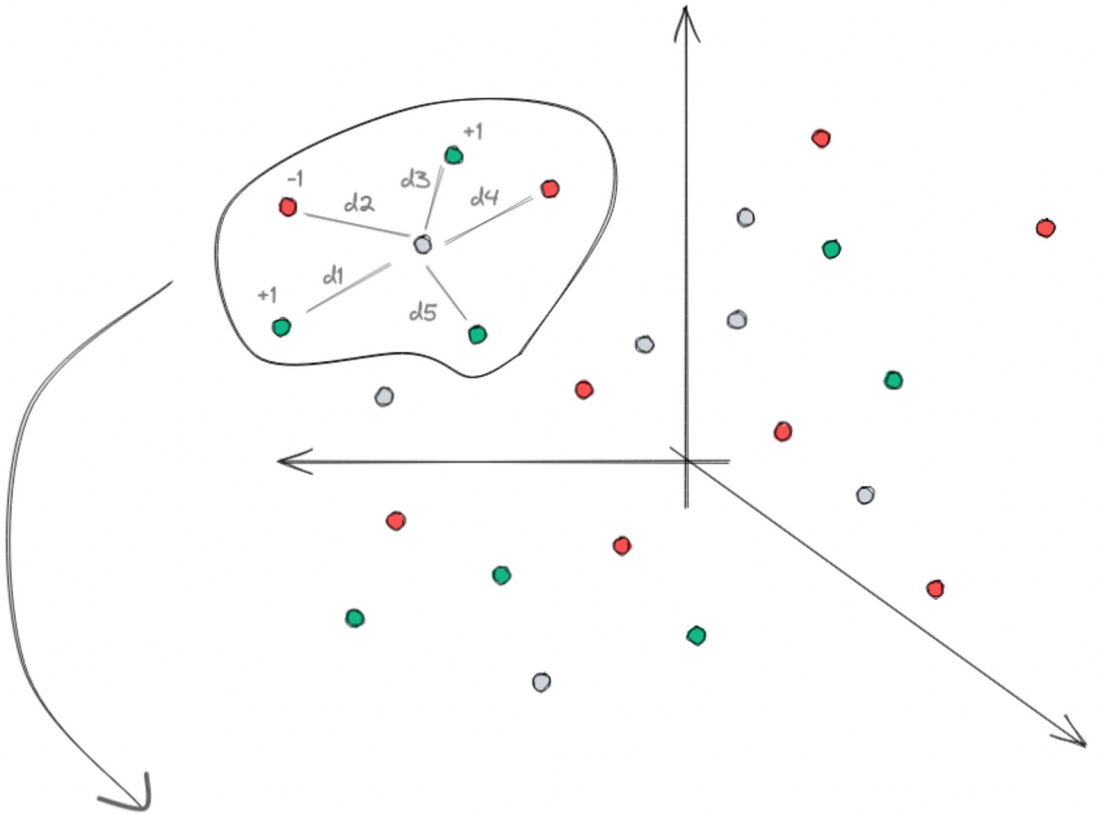
Confidence is a question
(96%) "How old are you?"
(98%) "What is your age?"
(7%) "My phone is good."
...

DIS

1 Companies in Space



2 Scoring a new company




```
score = np.average(neighbour_score, weights=distances)
```

Bonus, find what is wrong on this formula..

DIS

KNN can more *contextual* than a traditional binary classifier

- Embeddings can be re-used - if well chosen
- Recommendations are interpretable
- We can weigh different factors
- Multilingual with unbalanced data
- Generally easier to deploy

Company Name	Most similar	Rec Score
Budweiser	[... N ...]	0.879
Olx	[... N ...]	0.789
Unbabel	[... N ...]	0.678
Novo Nordisk	[... N ...]	0.001
LA Firefighters	[... N ...]	−0.995
...	...	...

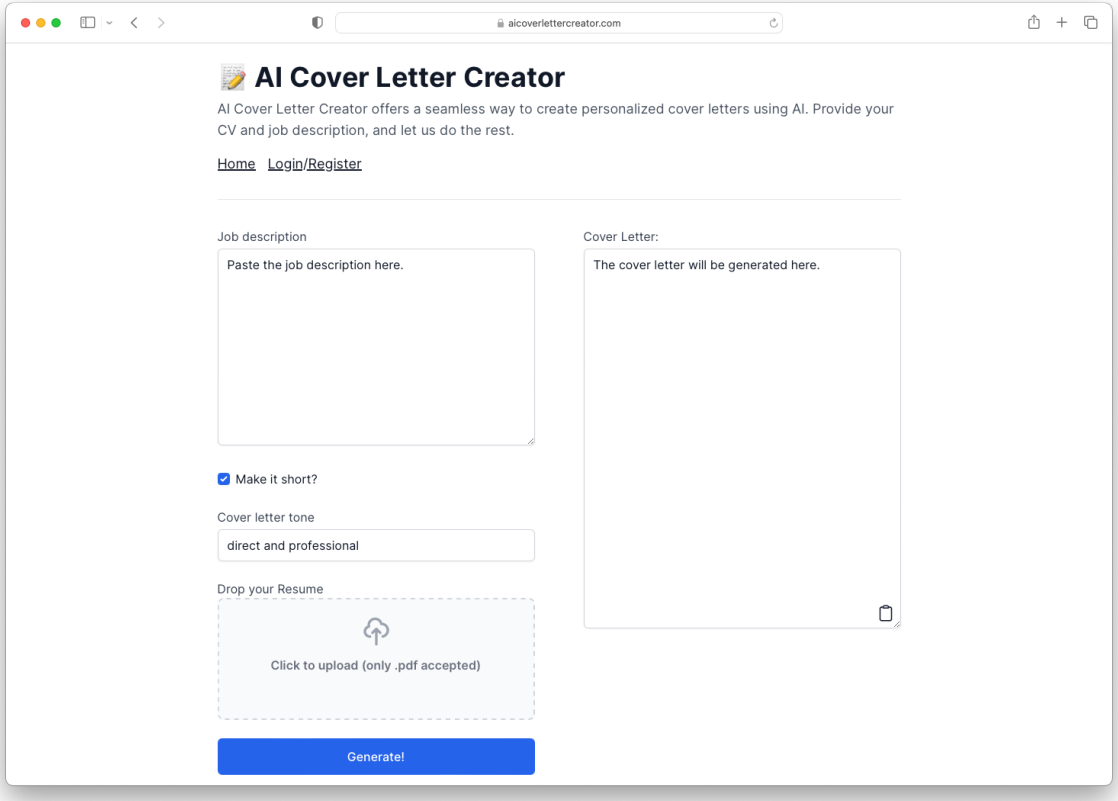
Most similar neighbors

Company score

DIS

2.3 | Using generative AI

It’s becoming simpler to build cool apps fast – but generative is not for everything!



aicoverlettercreator.com

- Django for the web interface
- Some Javascript for interactions
- OpenAI models
- Some cool prompts

DIS

Code Blame 22 lines (18 loc) · 839 Bytes Raw

```
Act like an assistant helping the user write personalized cover letter.

The cover letter should respect the following instructions:

Tone and style:
- Use a {{ tone }} tone.
- Do not use cliche sentences like "to express my interest", "I am hereby expressing".
- Be concise and direct.

Format:
- Do not use template tags.
- The cover letter should be easy to read.
{% if short == True %}- Use 2 paragraphs this is extremely important.
- Use 150 words or less for the cover letter.
{% else %}- Use 3 paragraphs this is extremely important.
- Use 250 words or less for the cover letter.
{% endif %}

Content:
- Be honest and truthful, do not make things up.
- Start with a strong, original, and punchy opening statement to catch the attention of recruiters.
- Use information about the user to tailor the cover letter for the job poster.
```

The system prompt

Code Blame Raw

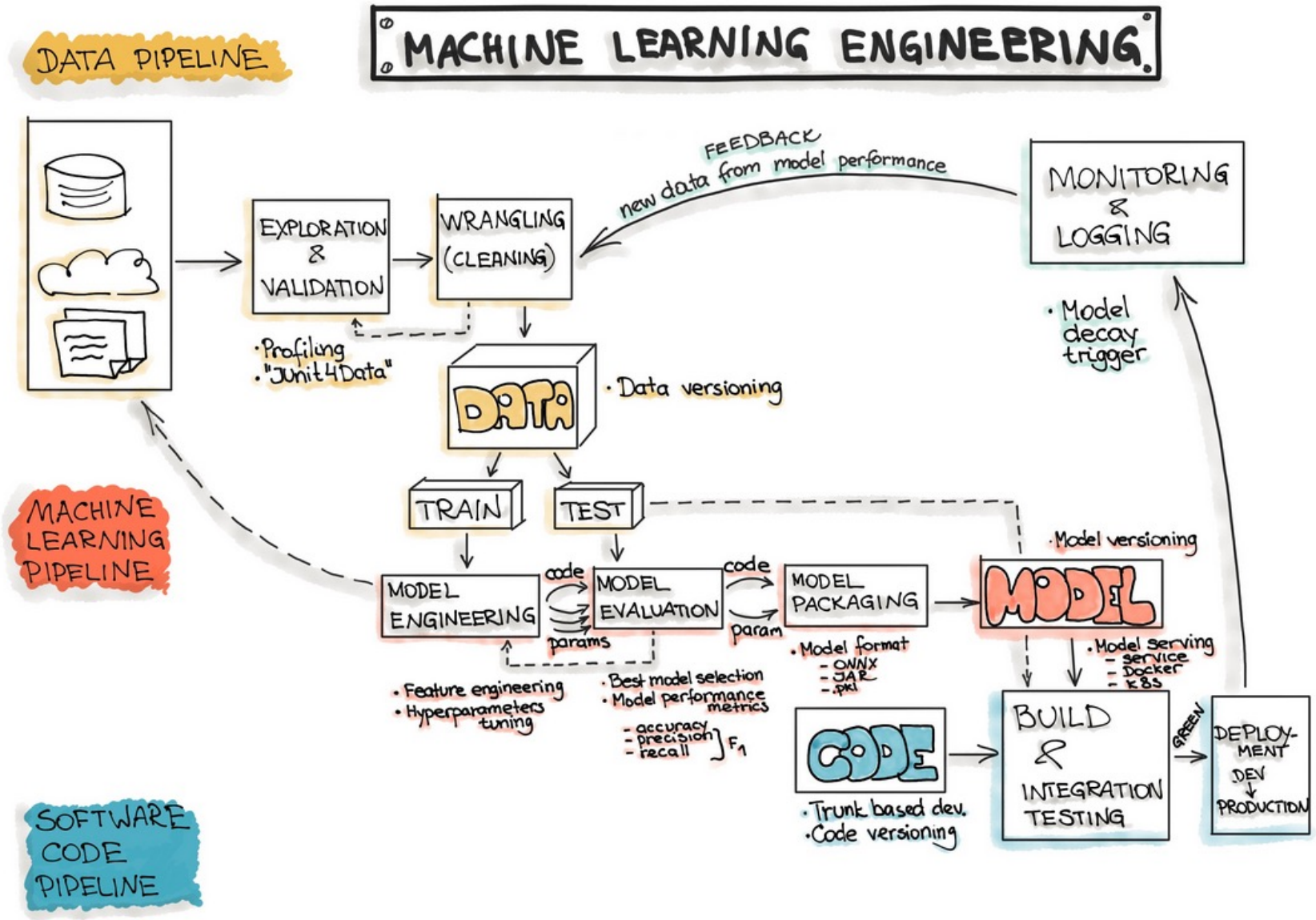
```
Information about the user:
-----
{{ resume_text }}
-----

Information about the job description:
-----
{{ job_description }}
-----

Personalized cover letter:
---
```

The user prompt

3 | “MLOps”

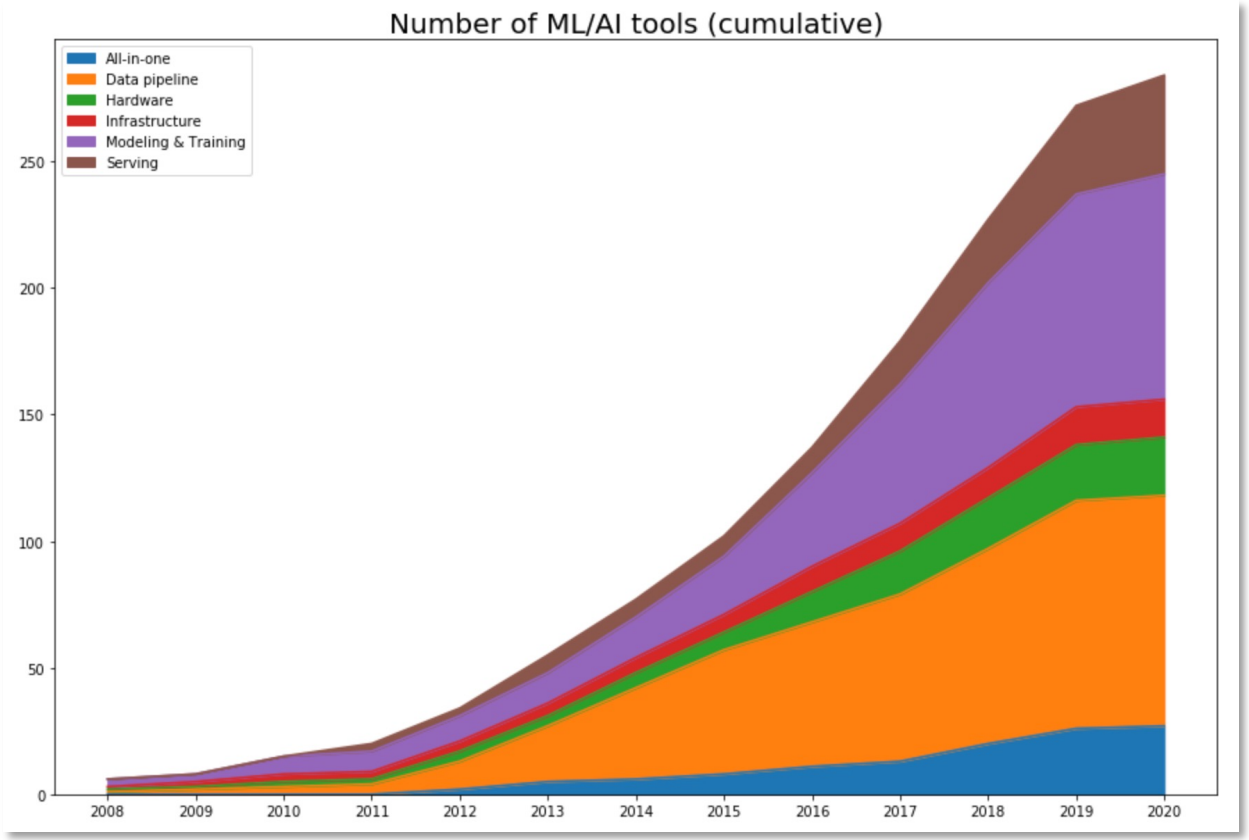


Credits: ml-ops.org

DIS

MLOps is not about adopting tools, it's about delivering value

- Gold Rush Age
- FOMO
- Spam emails
- Focus on tools
- 22% have put a model in production
- The real problem: Providing value.



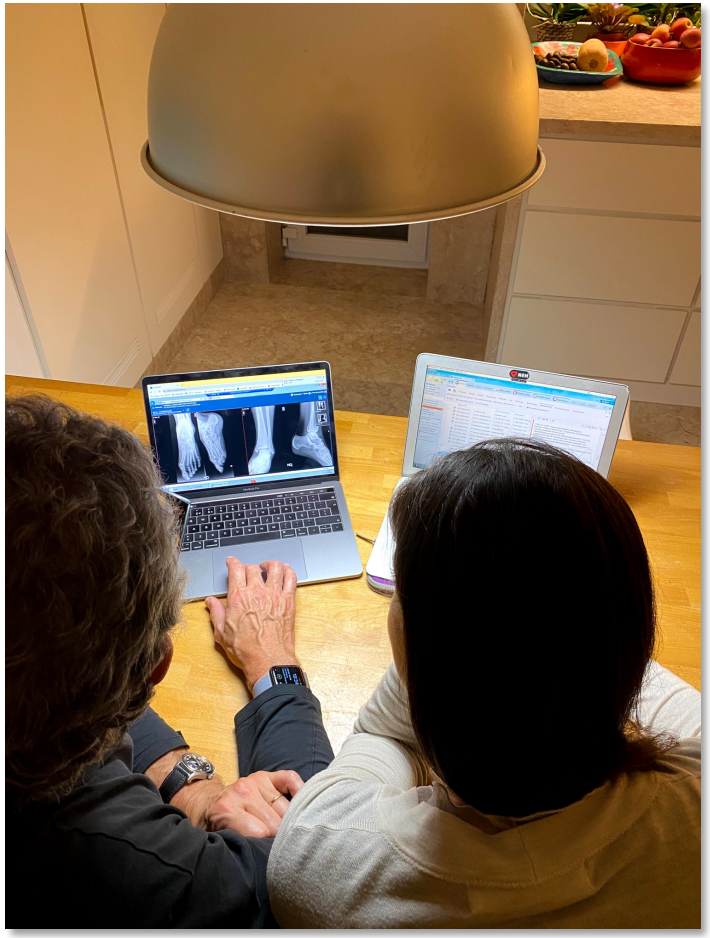
Credits: huyenchip.com

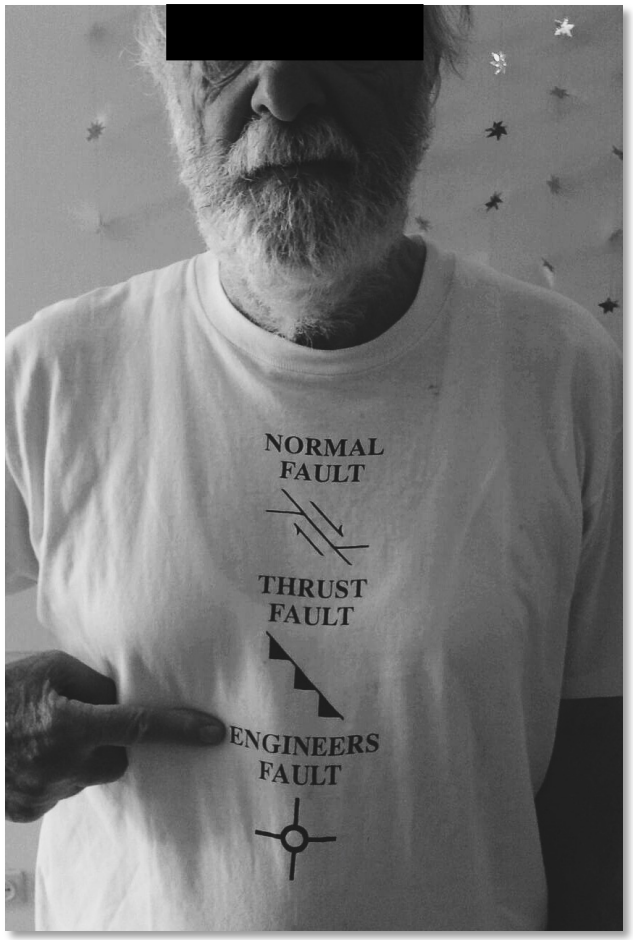
DIS

4 | Always be learning

ML is our craft

DIS



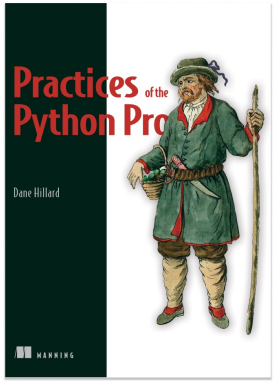

NORMAL
FAULT
THRUST
FAULT
ENGINEERS
FAULT

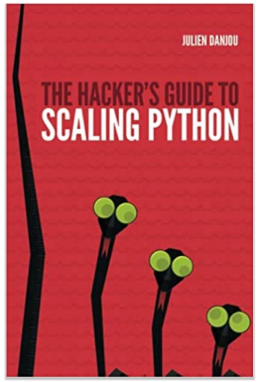

EINSTEIN
TINHA
RAZÃO?
Testando a Teoria da Relatividade Geral
Clifford M.Will
Portugal
Madeira
NOVO
ATLAS
GEOGRÁFICO

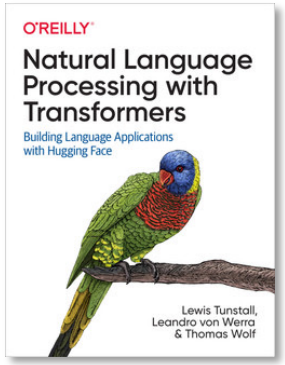
DIS

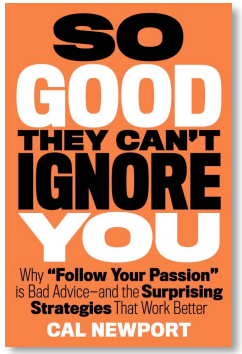
We should be masters of our craft

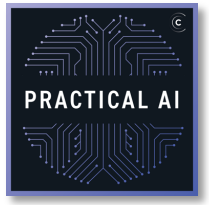
- Study
- Stay up-to-date
- Lean regularly
- Build things
- Give back and write











DIS

An OCD list of resources

Books

Practices of the Python Pro

Hacker's guide to scaling Python

Designing Data-Intensive Applications

Serious Python

Tutorials

Flask Mega-tutorial

RealPython

Stack Abuse

Kaggle + GitHub

YouTube

CodingTech

Sentdex

Abhishek Thakur

MLOPs Community

Podcasts

Talk Python to Me

Python Bytes

Podcast.__init__

Practical AI

News

PyCoder's Weekly

Medium

Awesome Python Weekly

Reddit RSS

...

DIS

Thank you, questions?

(reach out anytime via email: me@duarteocarmo.com)

DIS