

# Cloud Based Python Microservices

## *ML in Production*



Andy McMahon

Head of Data Science & Machine Learning  
andy@streamba.net

STREAMBA

V O R 

# STREAMBA

Integrate data from across the supply chain.

Perform predictive and prescriptive analytics in real time.

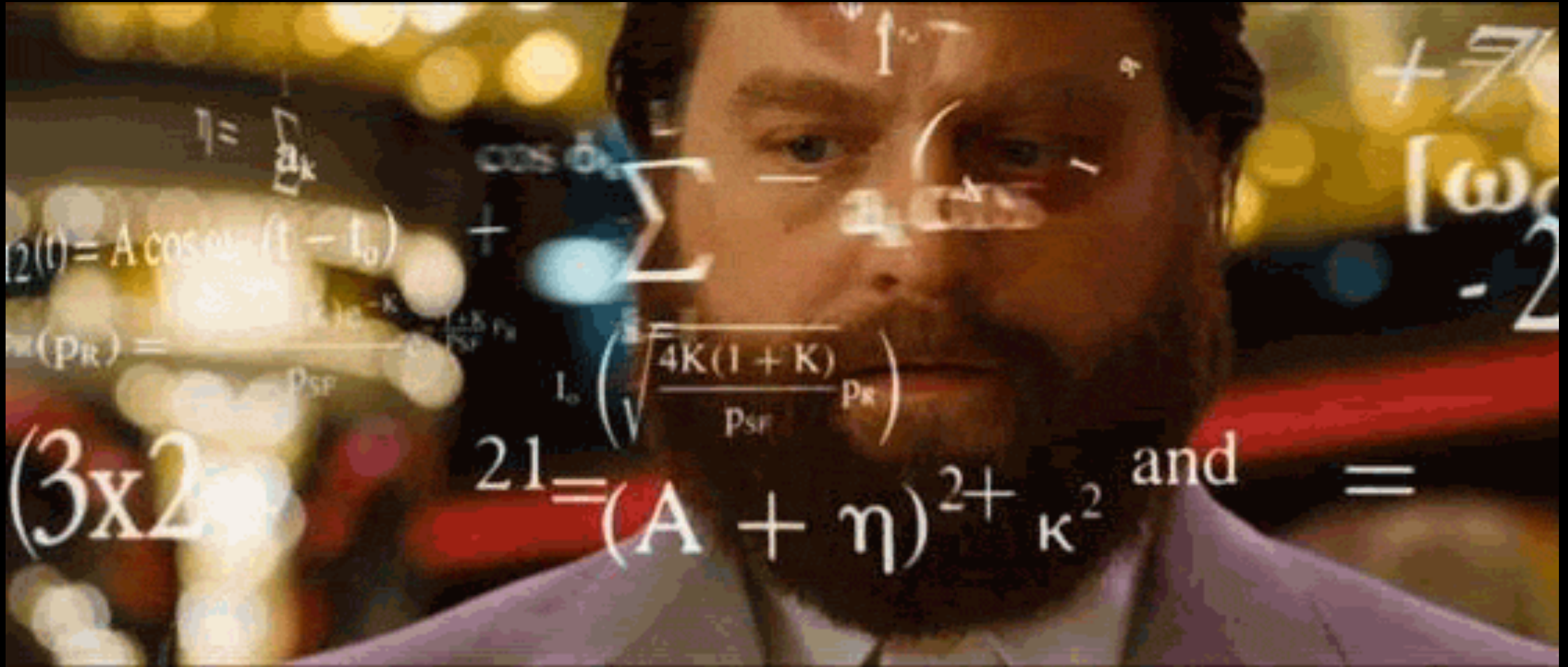
V O R 

Drive efficiency through smart suggestions and visibility.

Augment humans with computational and artificial intelligence capabilities.

STREAMBA

ML is hard





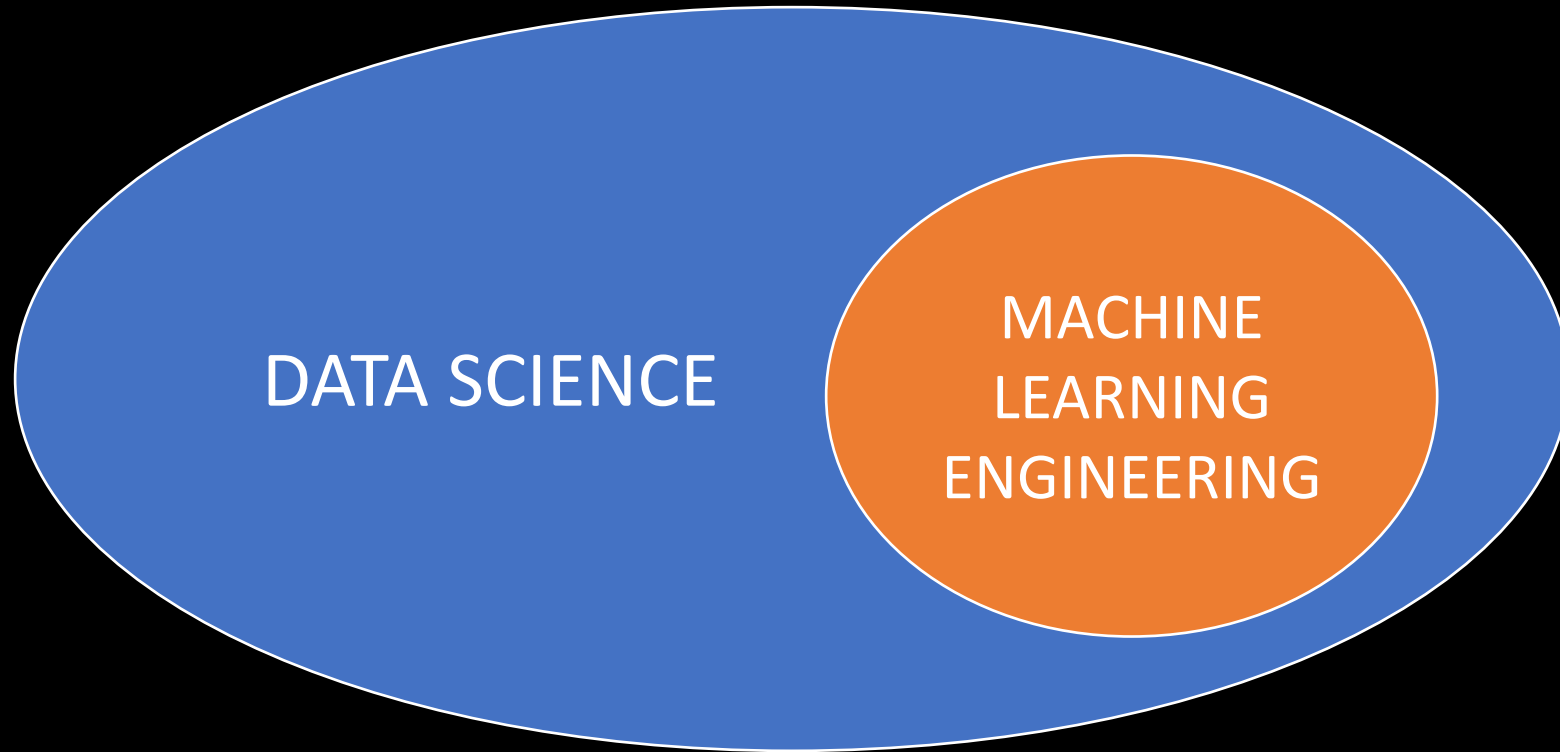
JORGE CHAM © 2017

DRIVE-BY MENTORING.

WWW.PHDCOMICS.COM

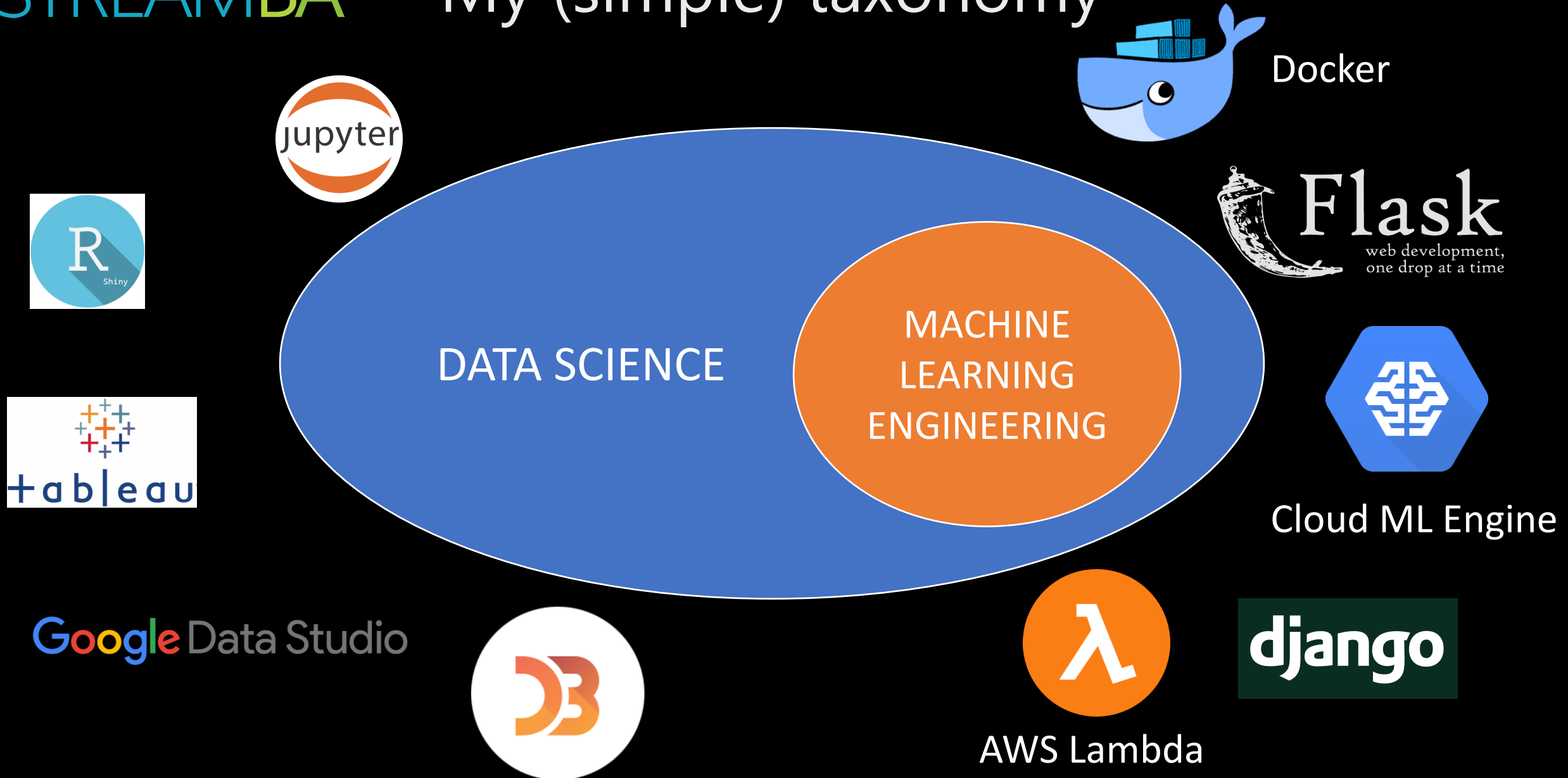
STREAMBA

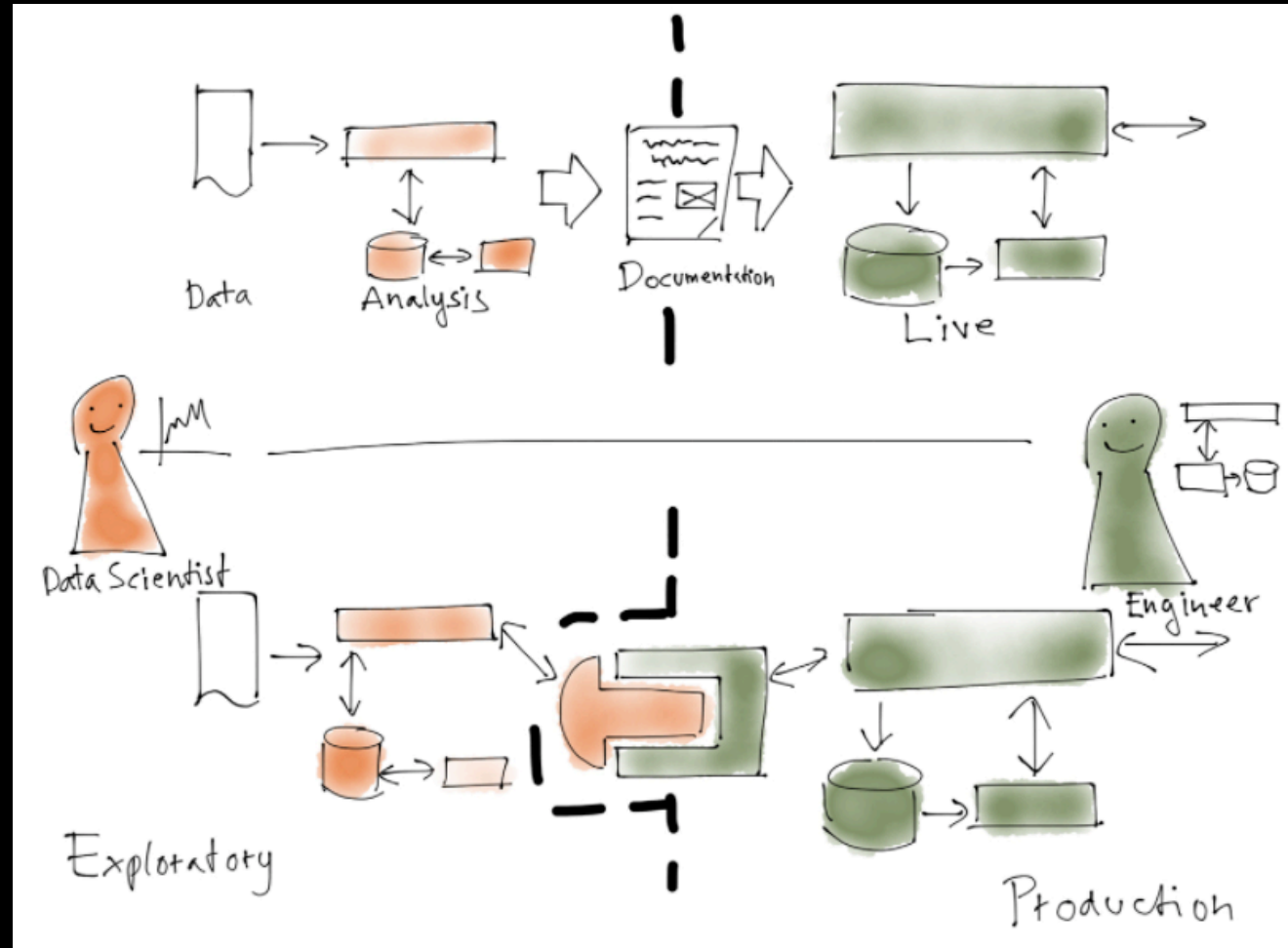
# My (simple) taxonomy



STREAMBA

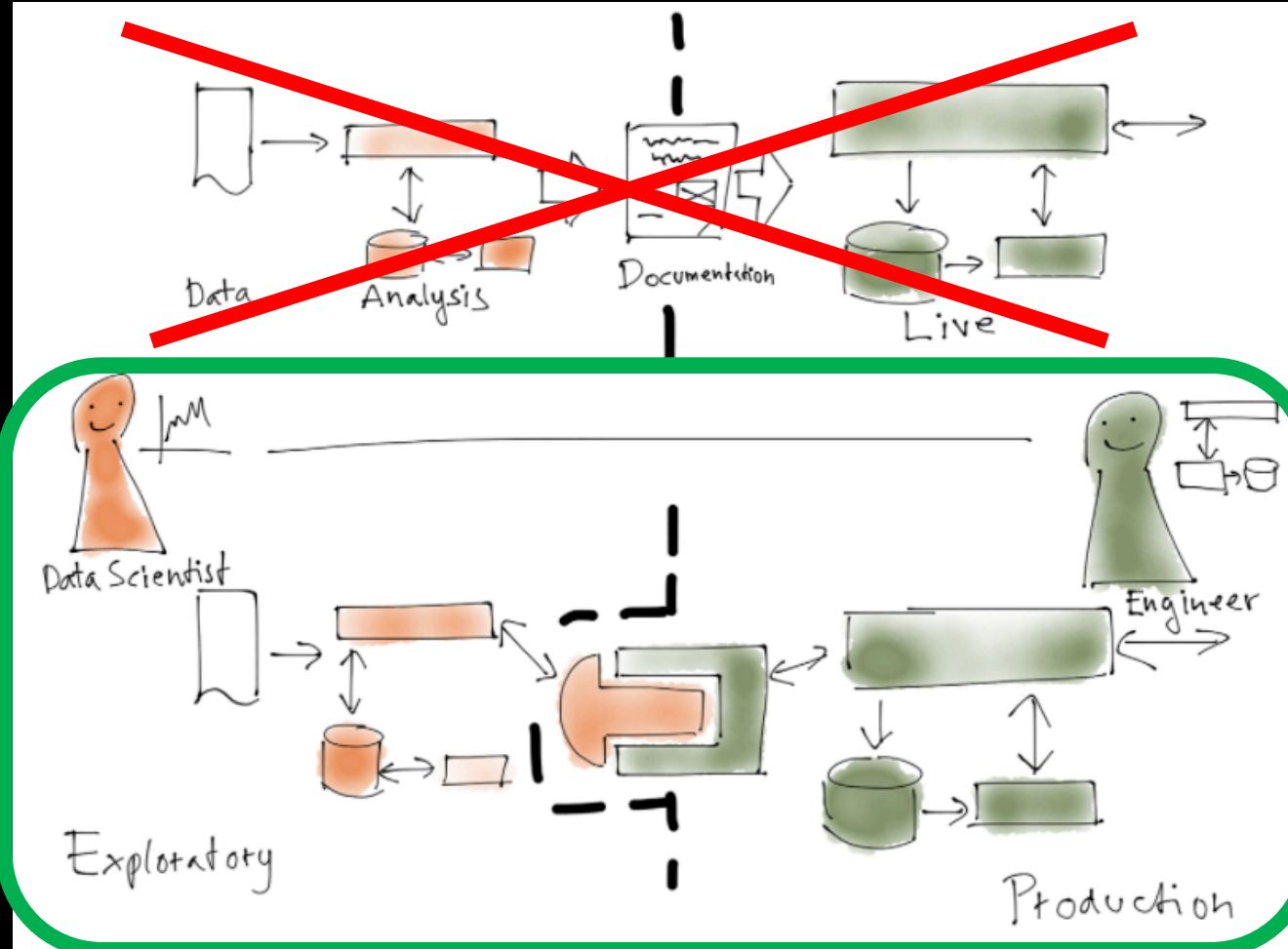
# My (simple) taxonomy







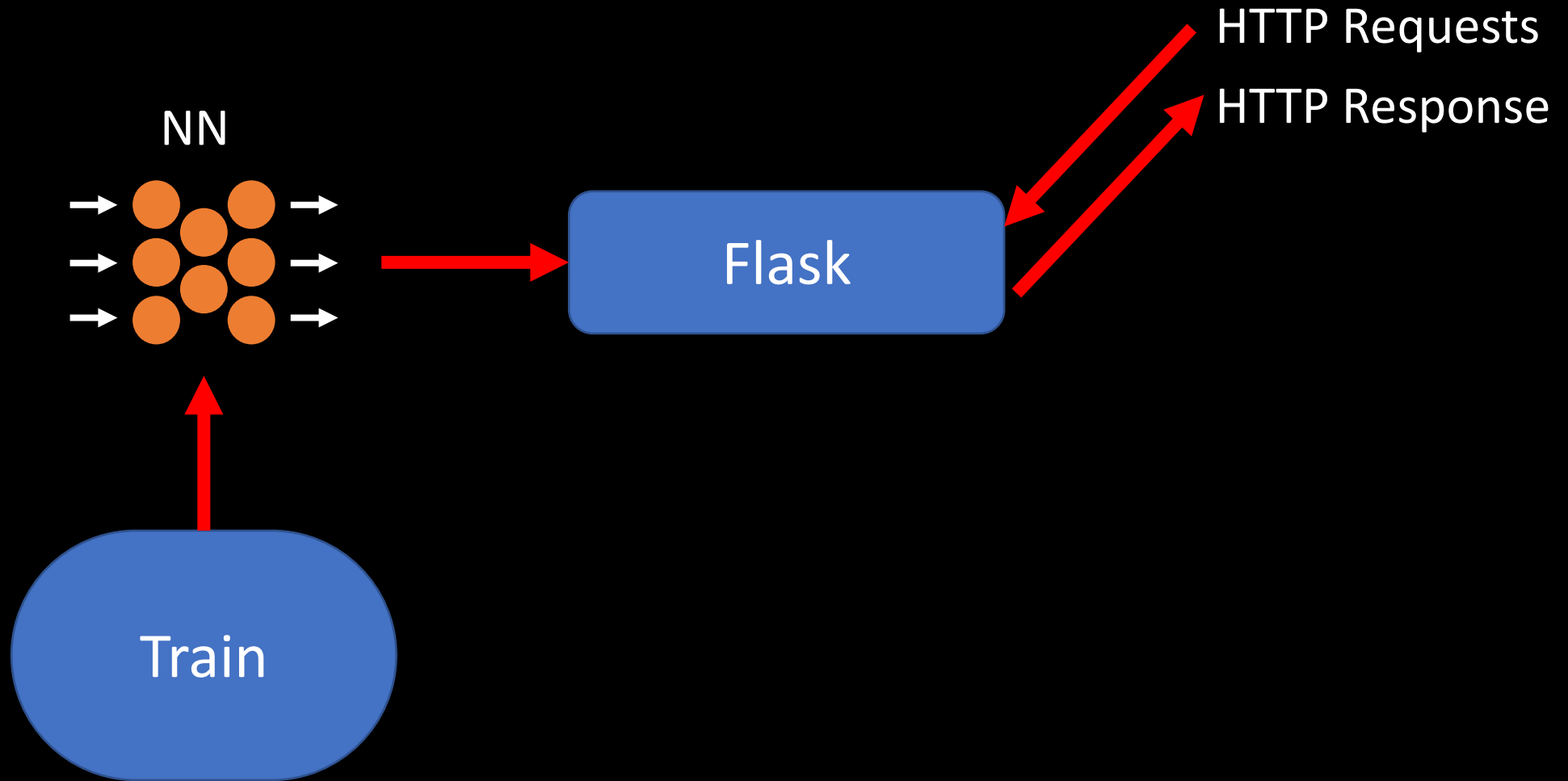
*Cross-collaboration,  
no siloes.*

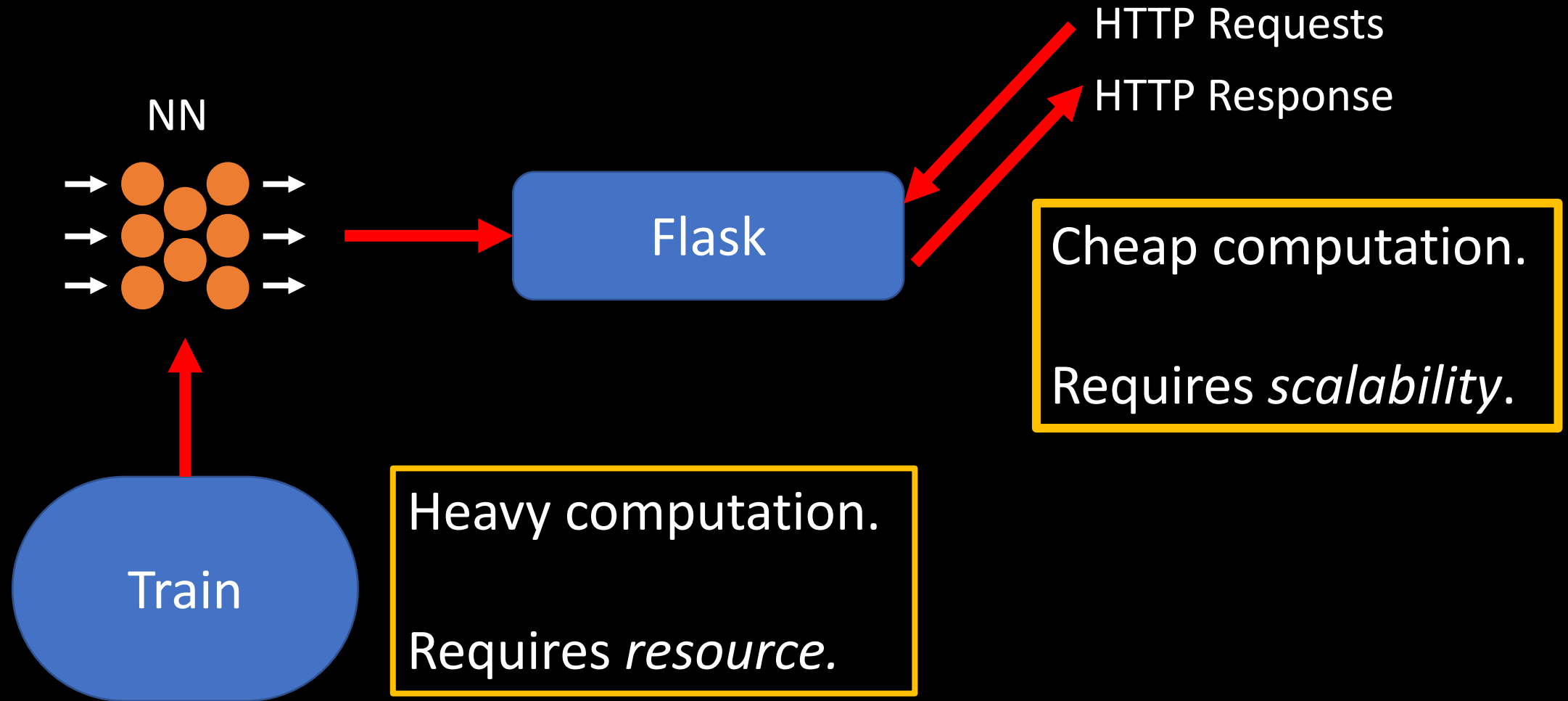


*Data team own  
their products.*

*Faster and more  
robust to change.*

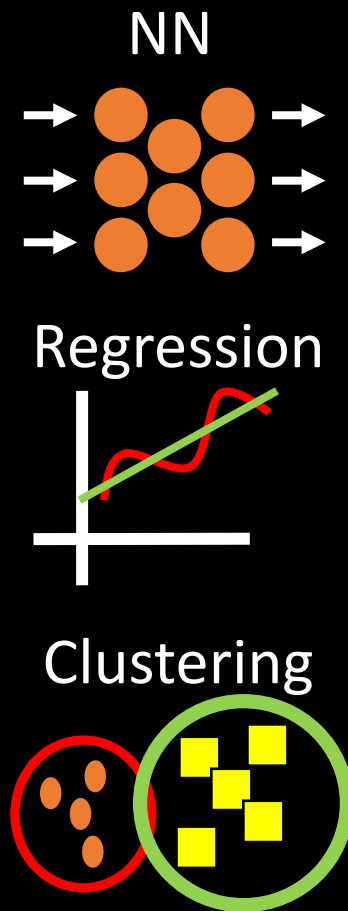
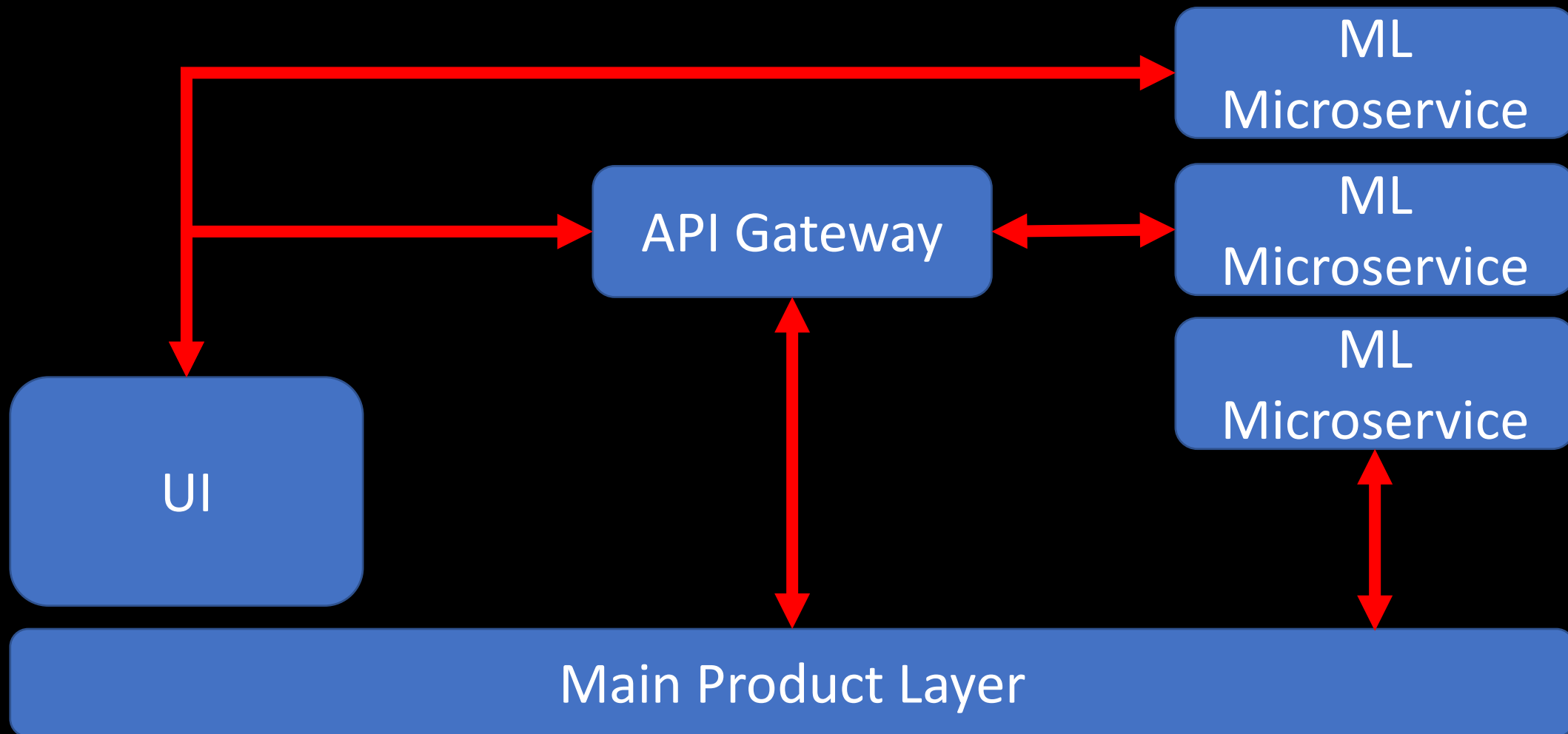
*Contracts between  
ML services and  
main product(s).*





# STREAMBA

## ML Microservices



STREAMBA

# ML as a Service

Train model

Save model

Load model

Serve model

Train model

Locally,  
Cloud,  
On Premises  
data centre.


Save model

Cloud (S3,  
Google Cloud  
Storage, Azure  
Storage).

Load model

Serve model

Cloud (Google  
App Engine,  
Cloud ML  
Engine, AWS  
Lambda, EC2).




The screenshot shows the Google Cloud website's product page for the Cloud Machine Learning Engine. At the top is the Google Cloud logo and a navigation bar with links: Why Google, Products (highlighted), Solutions, Launcher, Pricing, Security, Customers, and Documentation. Below the navigation bar, the main heading is 'CLOUD MACHINE LEARNING ENGINE' with the subtitle 'Machine Learning on any data, of any size'. A blue button with a brain icon and the text 'TRY IT FREE' is positioned below the subtitle. Further down, the section 'Managed Scalable Machine Learning' contains a paragraph describing the service as a managed platform for building and training machine learning models using TensorFlow, and serving predictions at scale. It mentions integration with other Google Cloud services like Dataflow, Storage, and BigQuery.

Google Cloud

Why Google Products Solutions Launcher Pricing Security Customers Documentation

## CLOUD MACHINE LEARNING ENGINE

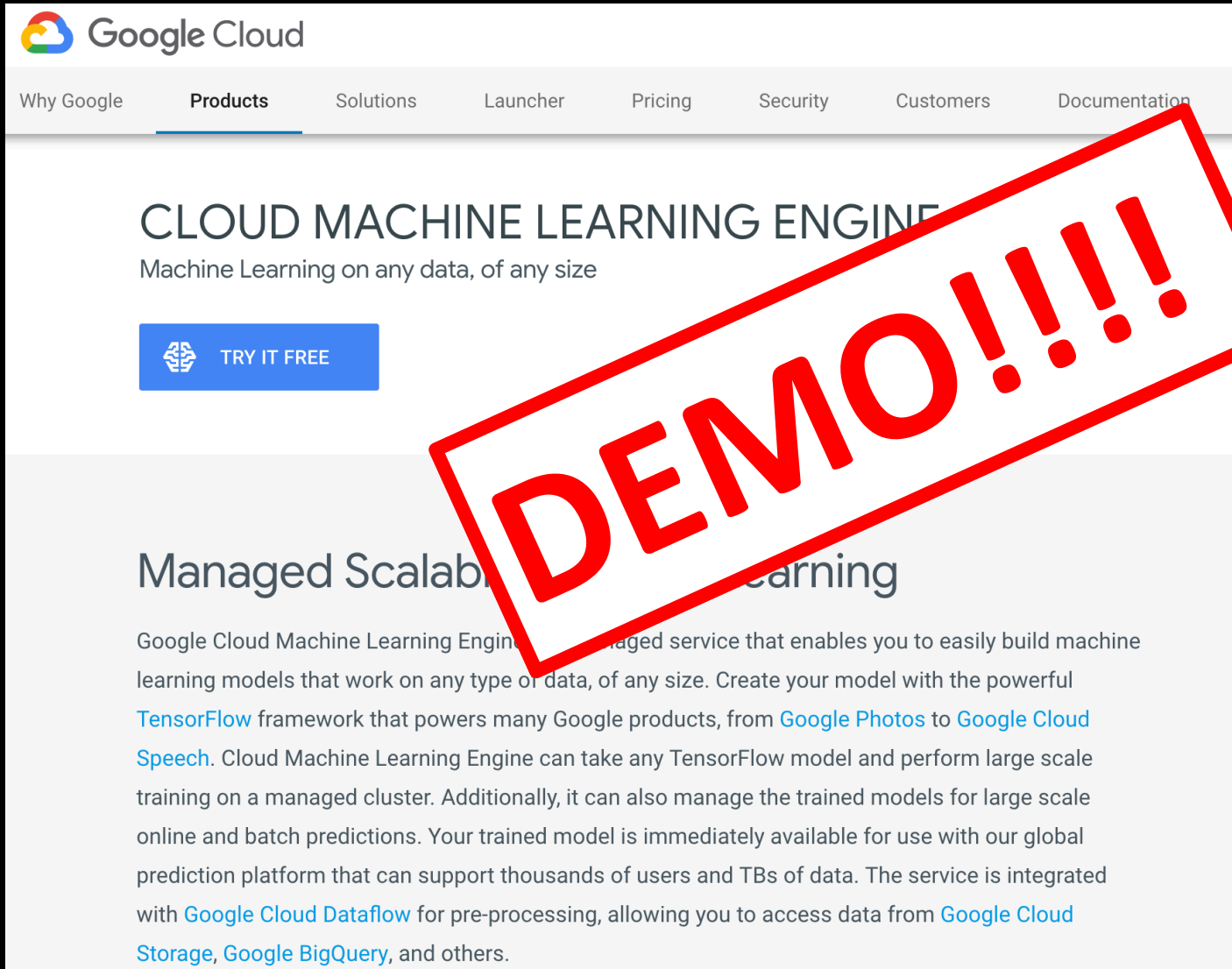
Machine Learning on any data, of any size

 TRY IT FREE

### Managed Scalable Machine Learning

Google Cloud Machine Learning Engine is a managed service that enables you to easily build machine learning models that work on any type of data, of any size. Create your model with the powerful [TensorFlow](#) framework that powers many Google products, from [Google Photos](#) to [Google Cloud Speech](#). Cloud Machine Learning Engine can take any TensorFlow model and perform large scale training on a managed cluster. Additionally, it can also manage the trained models for large scale online and batch predictions. Your trained model is immediately available for use with our global prediction platform that can support thousands of users and TBs of data. The service is integrated with [Google Cloud Dataflow](#) for pre-processing, allowing you to access data from [Google Cloud Storage](#), [Google BigQuery](#), and others.

- Managed clusters set up for Tensorflow jobs.
- Can train model on cluster / locally.
- Can serve predictions via an endpoint managed by CloudML. Model loaded into Google Cloud Storage to be read.




Google Cloud

Why Google Products Solutions Launcher Pricing Security Customers Documentation

## CLOUD MACHINE LEARNING ENGINE

Machine Learning on any data, of any size

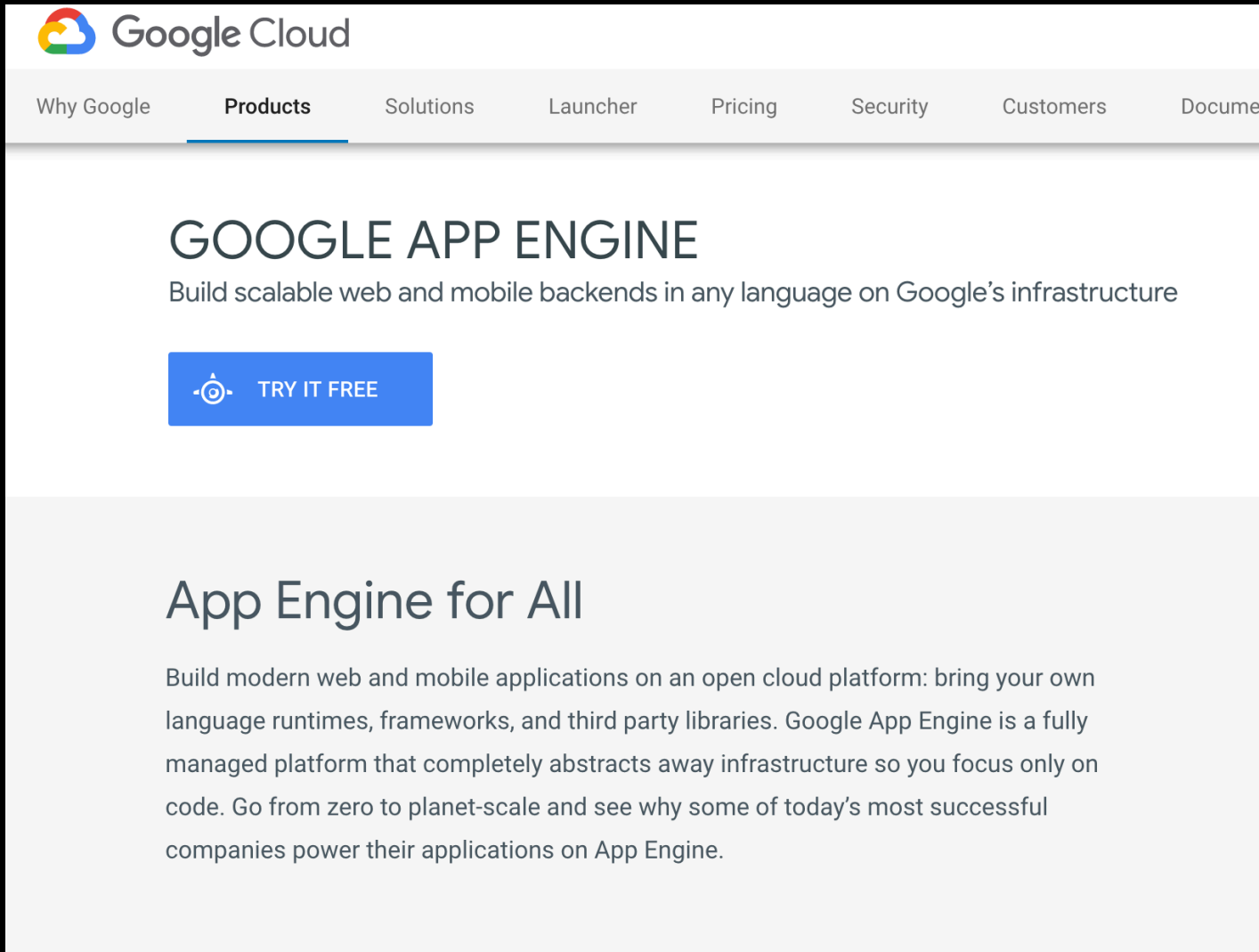
 TRY IT FREE

### Managed Scalable Machine Learning

Google Cloud Machine Learning Engine is a managed service that enables you to easily build machine learning models that work on any type of data, of any size. Create your model with the powerful [TensorFlow](#) framework that powers many Google products, from [Google Photos](#) to [Google Cloud Speech](#). Cloud Machine Learning Engine can take any TensorFlow model and perform large scale training on a managed cluster. Additionally, it can also manage the trained models for large scale online and batch predictions. Your trained model is immediately available for use with our global prediction platform that can support thousands of users and TBs of data. The service is integrated with [Google Cloud Dataflow](#) for pre-processing, allowing you to access data from [Google Cloud Storage](#), [Google BigQuery](#), and others.

- Managed clusters set up for Tensorflow jobs.
- Can train model on cluster / locally.
- Can serve predictions via an endpoint managed by CloudML. Model loaded into Google Cloud Storage to be read.






The screenshot shows the Google Cloud App Engine landing page. At the top is the Google Cloud logo. Below it is a navigation bar with links: Why Google, Products (highlighted), Solutions, Launcher, Pricing, Security, Customers, and Documents. The main heading is 'GOOGLE APP ENGINE' with the subtext 'Build scalable web and mobile backends in any language on Google's infrastructure'. A blue button with a play icon and the text 'TRY IT FREE' is present. Below this is a section titled 'App Engine for All' with a paragraph describing the platform as an open cloud platform for building modern web and mobile applications, highlighting its managed nature and scalability.

Google Cloud

Why Google Products Solutions Launcher Pricing Security Customers Documents

## GOOGLE APP ENGINE

Build scalable web and mobile backends in any language on Google's infrastructure

 TRY IT FREE

### App Engine for All

Build modern web and mobile applications on an open cloud platform: bring your own language runtimes, frameworks, and third party libraries. Google App Engine is a fully managed platform that completely abstracts away infrastructure so you focus only on code. Go from zero to planet-scale and see why some of today's most successful companies power their applications on App Engine.

- Managed VM infrastructure.
- Supported languages and Docker containers.
- Built to scale.
- Very easy to use but you may want even more control (Compute Engine / EC2).



The image shows a screenshot of the Google Cloud App Engine landing page. At the top, the Google Cloud logo is visible, followed by a navigation bar with links: 'Why Google', 'Products' (which is underlined), 'Solutions', 'Launcher', 'Pricing', 'Security', 'Customers', and 'Partners'. Below the navigation bar, the heading 'GOOGLE APP ENGINE' is displayed, followed by the subtext 'Build scalable web and mobile backends in any language'. A blue button with a play icon and the text 'TRY IT FREE' is present. A large, red, tilted rectangular stamp with the word 'DEMO!!!!' in bold, white, sans-serif font is overlaid diagonally across the center of the page. Below the stamp, the heading 'App Engine for' is visible, followed by a paragraph: 'Build modern web and mobile applications on an open cloud platform: bring your own language runtimes, frameworks, and third party libraries. Google App Engine is a fully managed platform that completely abstracts away infrastructure so you focus only on code. Go from zero to planet-scale and see why some of today's most successful companies power their applications on App Engine.'

- Managed VM infrastructure.
- Supported languages and Docker containers.
- Built to scale.
- Very easy to use but you may want even more control (Compute Engine / EC2).

- It is important you know how to share your ML work.
- Learn the tools – just pick one and start playing.
- GCP has options specifically for Tensorflow/Scikit-learn and more general options. AWS and Azure have great tools too.
- Don't fall into the trap of thinking "that's the engineer's job", it's *your job!*