Inteligencia Artificial para el Análisis de Datos.

Retos, Estrategias y Aplicaciones

Un enfoque práctico para proyectos de prototipado e innovación

Un poco sobre mi...

Roger Galindo

- **Físico por entrenamiento**
  - Lic. Física
  - Maestría en Ciencias - Física
  - PhD. Experimental High Energy Physics
- **Científico de datos de profesión**
  - Head of Data and Engineering
    - Demetria Colombia SAS
  - Google Cloud Certifications
    - Data Engineer
    - ML Engineer

... Hay una convocatoria

El CTO llama a una reunión y lanza la frase:
“...Queremos identificar temas y priorizar respuestas de emails...”



¿Qué perfiles necesito?

Front-End

Data Engineering

Data Analytics

Data Scientist

Infrastructure

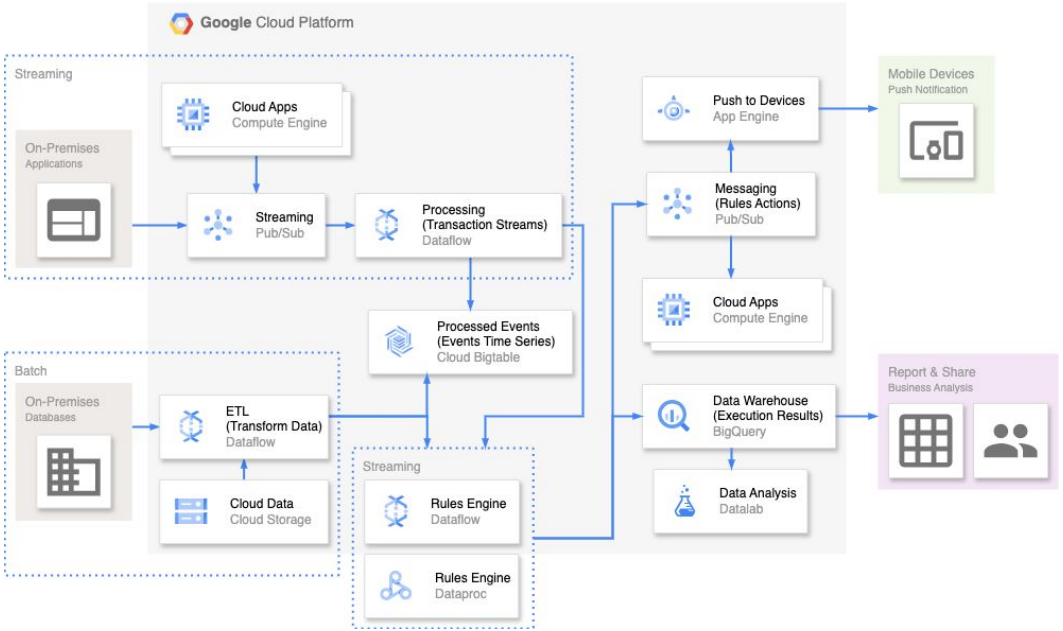
Devops

QA

Security

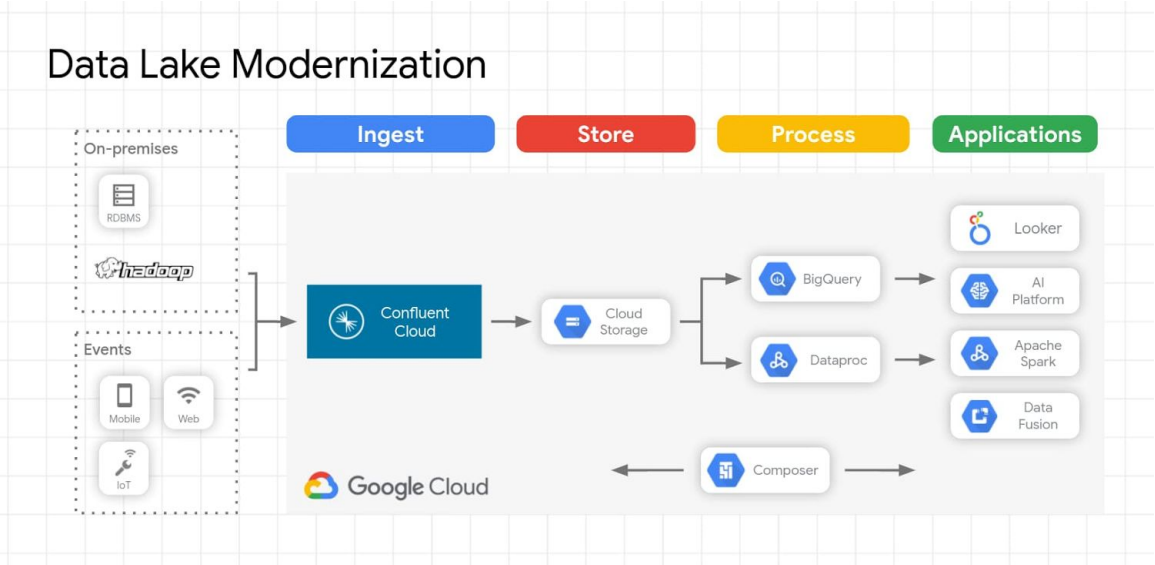
Back-End

¿Cómo lo voy a hacer?



- ¿Cuáles son mis fuentes de datos?
- ¿Cuál va a ser el flujo de datos?
- ¿Qué modelos de IA voy a implementar?
- ¿Cuál es el sistema de respuesta y escalamiento?
- ¿Cuál va a ser el almacenamiento y monitoreo?
- ¿Cómo va a ser la retroalimentación y la mejora?

¿Cómo lo voy a hacer? - Data Lake

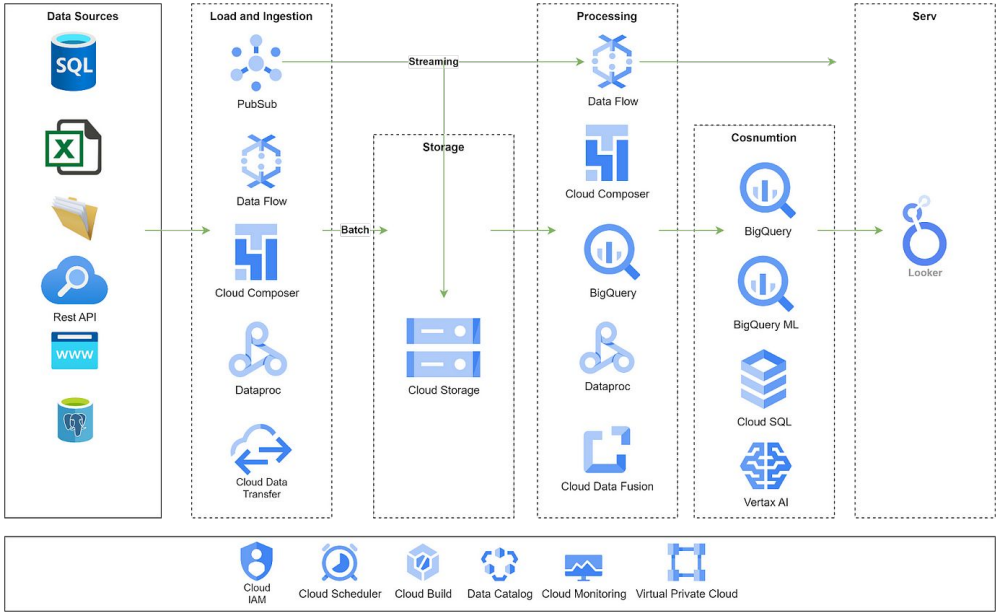


- **¿Cuáles son mis fuentes de datos?**
  - API correos
    - gmail
    - outlook
  - ETL
    - Kafka
    - Dataflow

Data Ingestion.

- Data Lake (layers)
  - Bronze
  - Silver
  - Gold

¿Cómo lo voy a hacer? - ETL

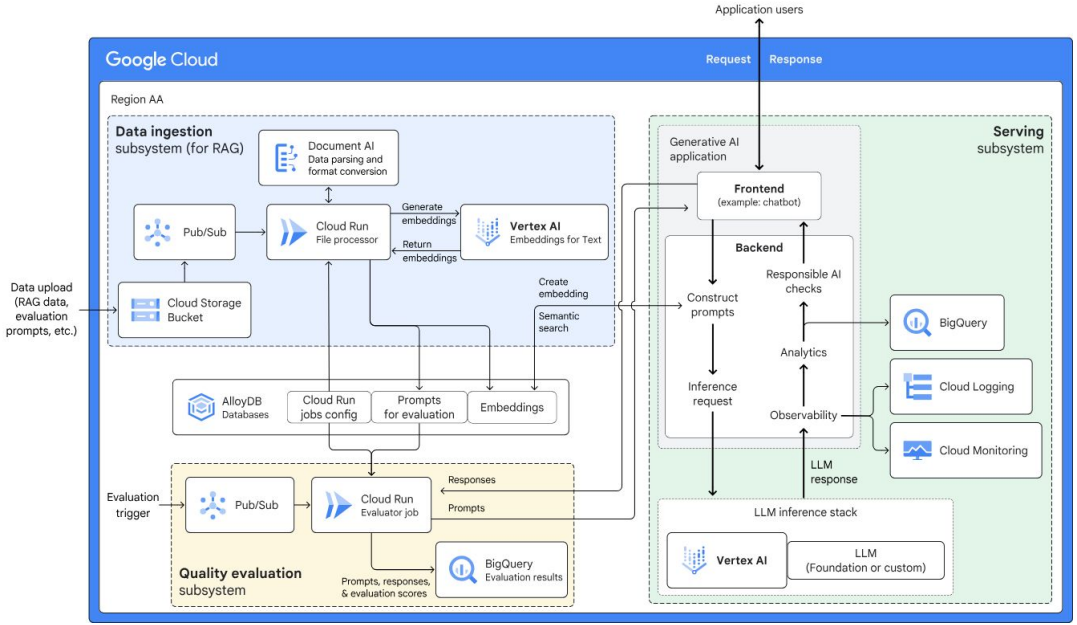


- **¿Cuál va a ser el flujo de datos?**
  - Limpieza de texto
  - Tokenización
  - Análisis de lenguaje

Data Preprocessing.

- Batch preprocessing
  - Cloud scheduler + Cloud Functions
- Bulk preprocessing
  - Apache Airflow
- Streaming preprocessing
  - Cloud Pub/Sub + Dataflow

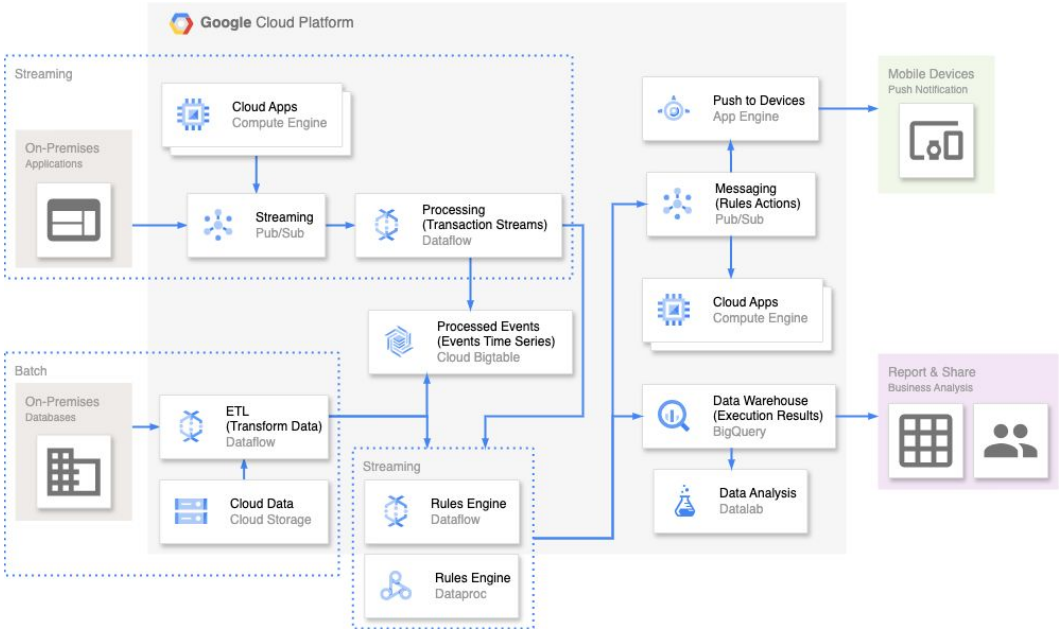
¿Cómo lo voy a hacer? - IA



- **¿Qué modelos de IA voy a implementar?**
  - Clasificador temático
  - Clasificador de prioridad
  - Generador de respuestas

AI Modeling.

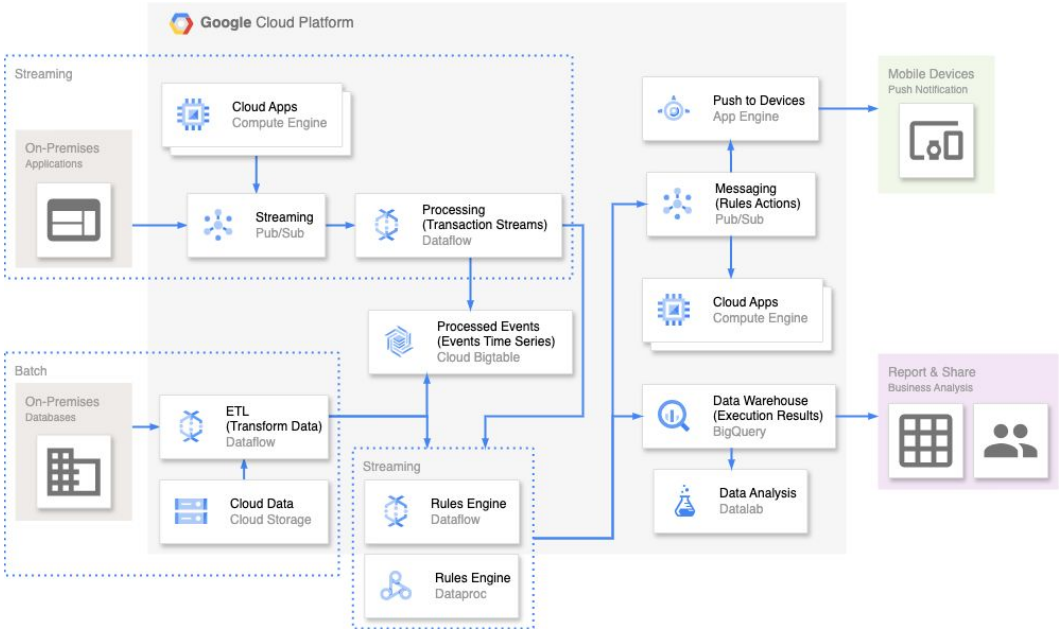
¿Cómo lo voy a hacer? - El MVP



- **¿Cuál es el sistema de respuesta y escalamiento?**
  - Respuesta automático
    - Chatbot
  - Escalamiento a un humano

Model Usage

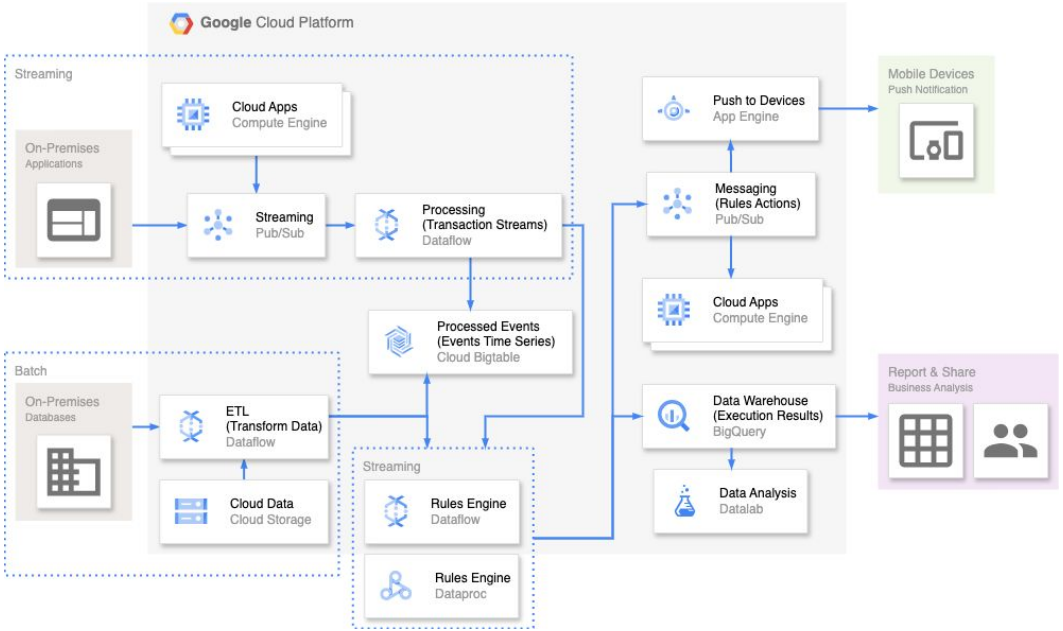
¿Cómo lo voy a hacer? - Desempeño del MVP



- **¿Cuál va a ser el almacenamiento y monitoreo?**
  - Bases de datos
  - Dashboards

Model Monitoring

¿Cómo lo voy a hacer? - Evolución del MVP



- ¿Cómo va a ser la retroalimentación y la mejora?
  - Datos etiquetados
  - Entrenamiento Continuo

MLOps

… Durante el proyecto

El PM llama a una reunión (Planeación del sprint) y lanza la pregunta:

“…¿Cómo vamos?…”





Pruebas controladas.

Entender la capacidades del producto



Simulación del entorno.

Entender el cómo se comportaria el producto en producción

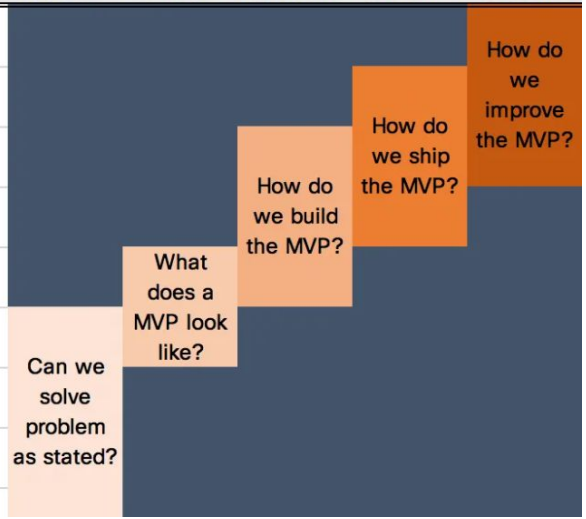


Puesta en producción.

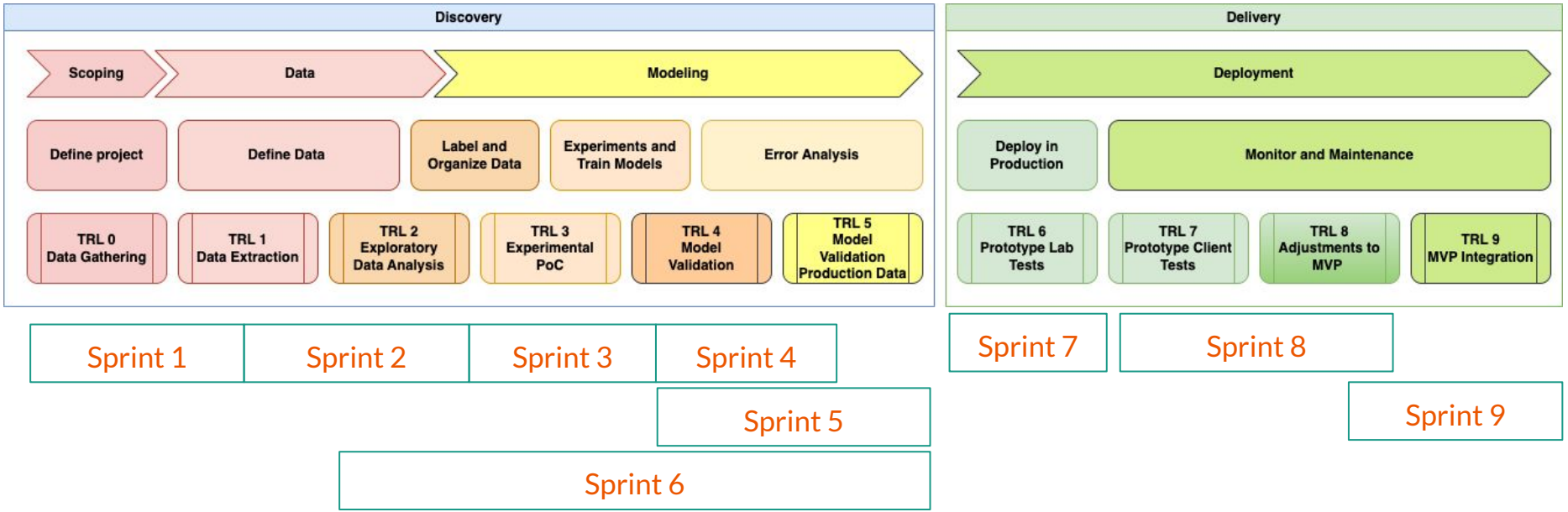
Monitoreo del comportamiento en producción

TRL - Technical Readiness Level

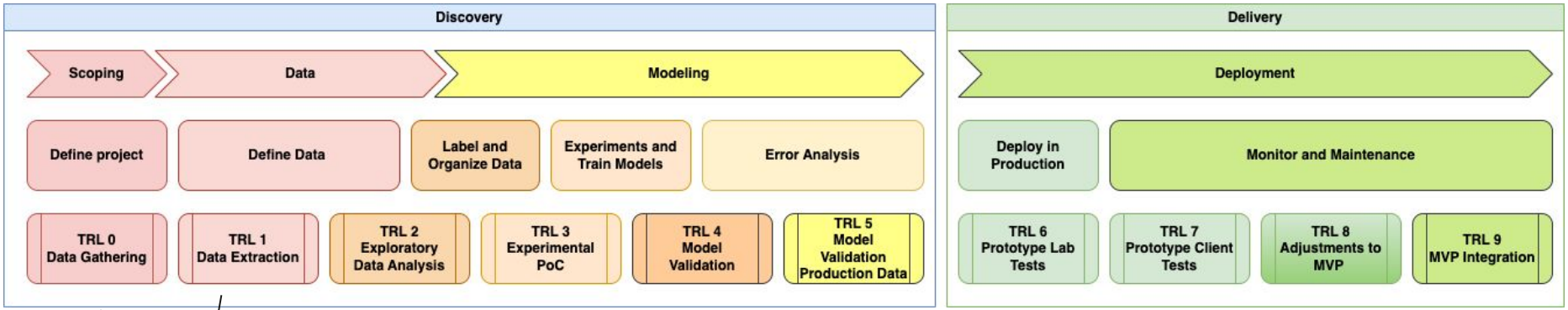
	Index	Technical Readiness Level	Product Readiness Levels	Data Science Readiness Levels
Delivery	9	System Proven in Operational Environment	System Running in Production	System Proven in Operational Environment
	8	System Complete & Qualified	QA Passed & Ready for Staging Environment	System Compete & Qualified
	7	Prototype Demonstrated in Operational Environment	Beta version in limited Staging Environment	Prototype Demonstrated in Operational Environment
	6	Technology Demonstrated in Relevant Environment	In QA	Algorithm Integrated in Development Environment
	5	Technology Validated in Relevant Environment	Tests passing in Development Environment	Algorithm Validated against Production Data
Discovery	4	Technology Validated in Lab	Tests passing on Development Machines	Algorithm Validated against Sample Data
	3	Experimental Proof of Concept	Tests cases written with full coverage	Experimental Proof of Concept
	2	Technology Concept Formulated	Technology Concept Formulated	Data Explored, Described, and Documented
	1	Basic Principles Observed	Need or shortcoming identified	Algorithm Design & Development



Más acerca de TRL - Desde la Ideación al MVP

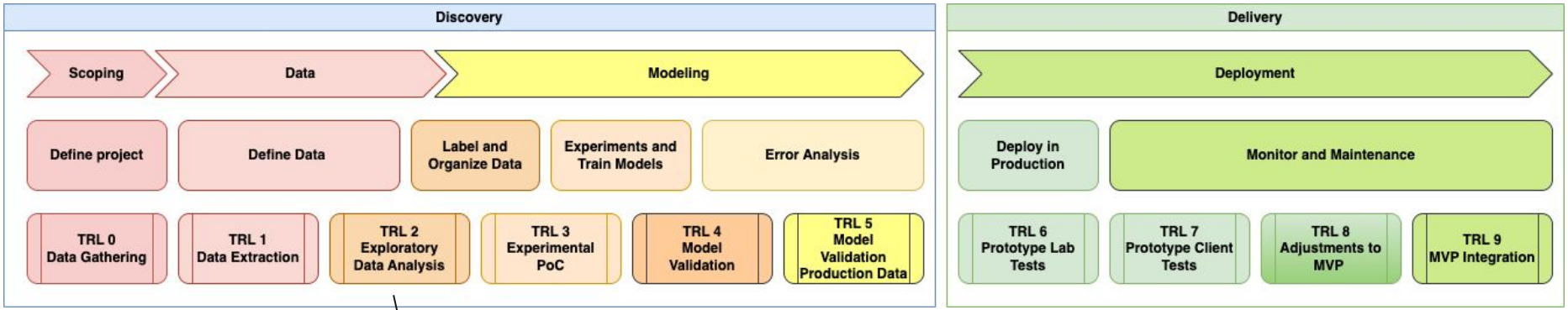


Más acerca de TRL - Desde la Ideación al MVP



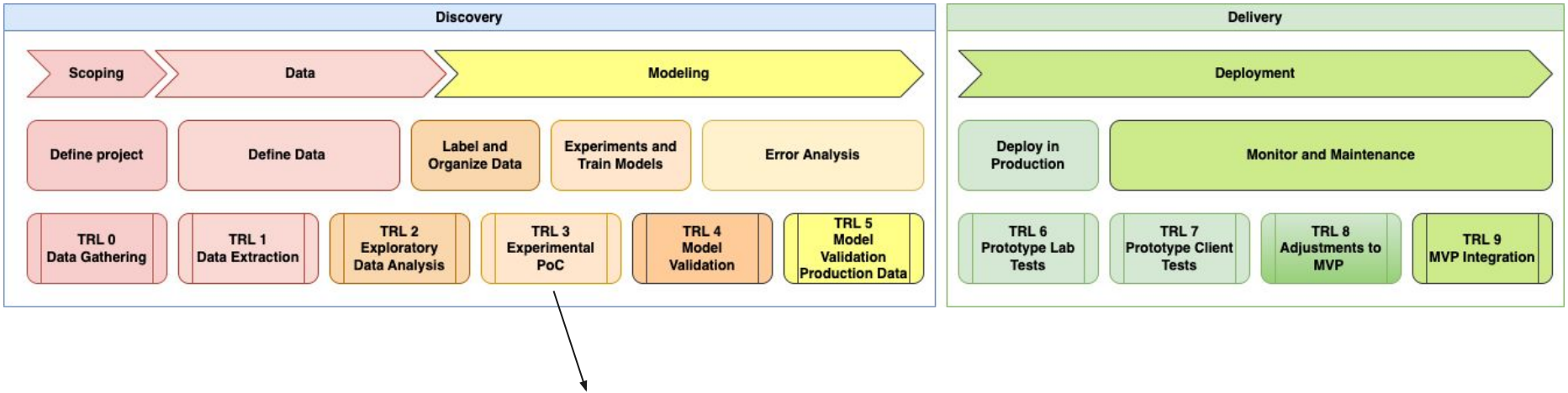
- ¿Tenemos los datos?
- ¿Cuál es la solución ideal?
- ¿Cómo los capturamos?

Más acerca de TRL - Desde la Ideación al MVP



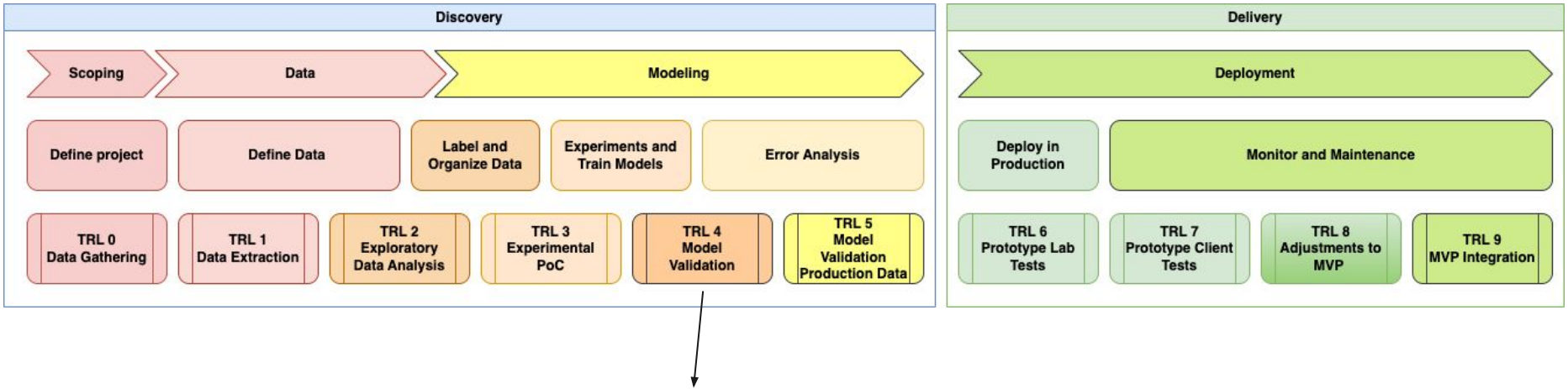
- SQL, Jupyter notebooks
- Feature Engineering
- ¿Cómo los capturamos?

Más acerca de TRL - Desde la Ideación al MVP



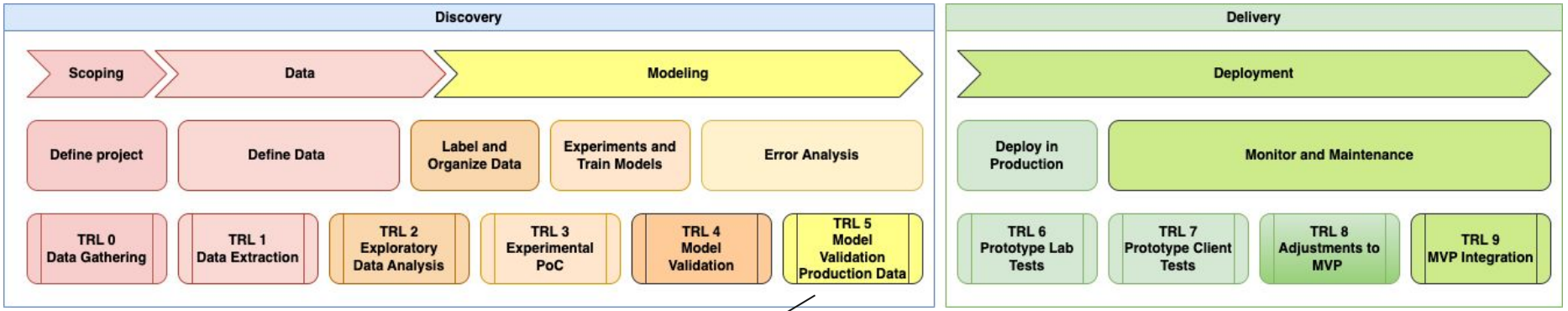
- Experimentación de modelos
- Definición de pipelines

Más acerca de TRL - Desde la Ideación al MVP



- Cross validation
- Model performance

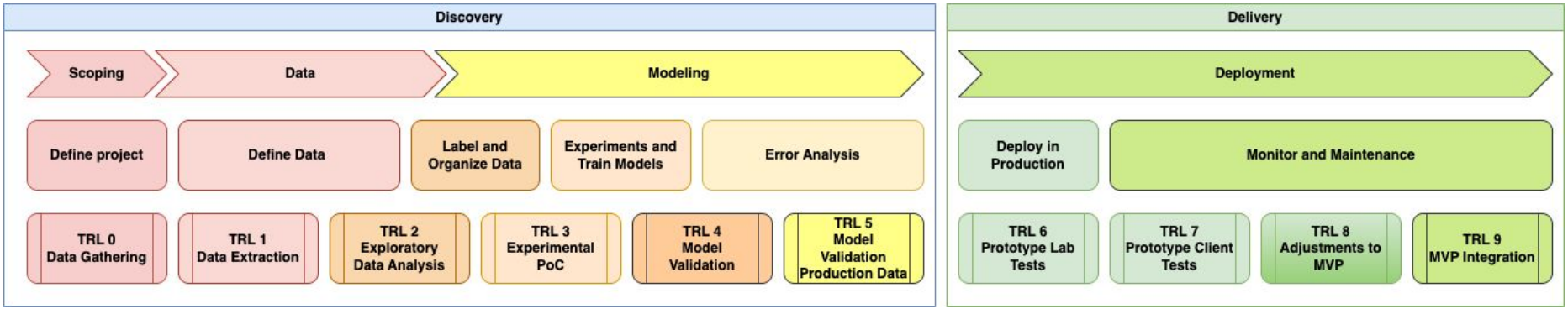
Más acerca de TRL - Desde la Ideación al MVP



- Extraer datos de producción y comportamiento del modelo
- **Detección de data drift**



Más acerca de TRL - Desde la Ideación al MVP



¿Cómo monitorear el proyecto?

TRL como épica

- Historias como metas a realizar
  - Tareas como entregables específicos
    - Subtareas si es necesario

Tipo	Clave	Resumen	Estado	Persona asignada
	SMGC-1	TRL0 - Data Gathering	FINALIZADA	
	SMGC-2	TRL1 - Data Extraction	FINALIZADA	
	SMGC-3	TRL2 - Exploratory Data Analysis	FINALIZADA	
	SMGC-4	TRL3 - Experimental PoC	FINALIZADA	
	SMGC-5	TRL4 - Model Validation Test Data	FINALIZADA	
	SMGC-6	TRL5 - Model Validation Production Data	FINALIZADA	
	SMGC-7	TRL6 - Prototype Lab Tests	EN CURSO	
	SMGC-14	Bean by Bean performance	EN CURSO	
	SMGC-15	Average sample vs Average sessions performance	EN CURSO	
	SMGC-8	TRL7 - Prototype Client Tests	EN CURSO	
	SMGC-11	Carcafé Performance	FINALIZADA	RG Roger Galindo
	SMGC-12	Cooperativa Manizales Performance	FINALIZADA	RG Roger Galindo
	SMGC-16	Almacafé - Soacha Pilot	EN CURSO	RG Roger Galindo
	SMGC-13	Cenicafé Performance	EN CURSO	RG Roger Galindo
	SMGC-9	TRL8 - Adjustments to MVP	EN CURSO	
	SMGC-10	TRL9 - MVP Integration	EN CURSO	

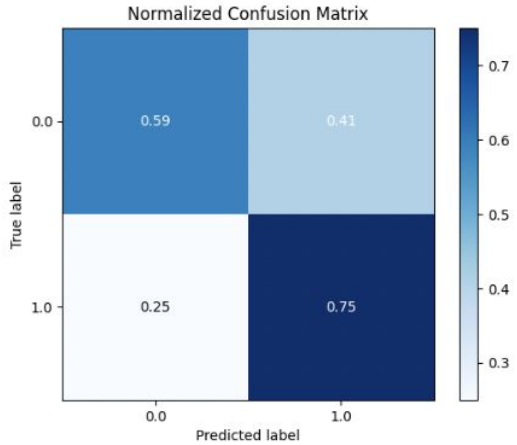
… Durante la entrega del proyecto

El cliente llama a una reunión (aceptación del producto) y lanza la pregunta:

“…¿Cuál es el desempeño?…”



… ¿Se resolvió el problema? …



Típicamente se presentan resultados basados en el performance del modelo

- TRL 4 - TRL 5.

¿Éstos números nos dicen si se resolvió el problema?

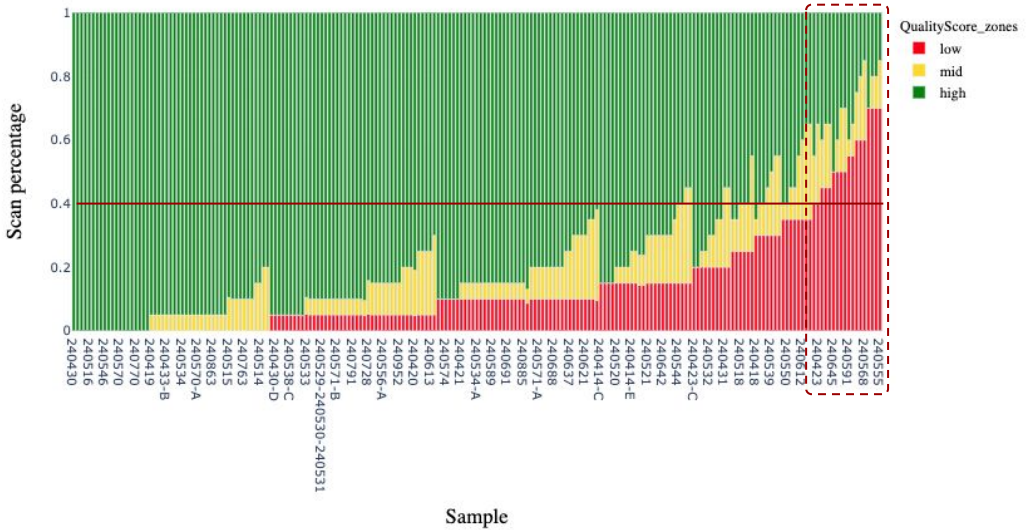
Un caso específico - Café con buenos sabores

Using more than 200 samples, all of them evaluated as Good.

Can we identify the presence of odd flavors?

Setting a threshold, where the number of beans with low probability is higher than the 40% of the beans analyzed per sample, we found 18 samples with potential issues

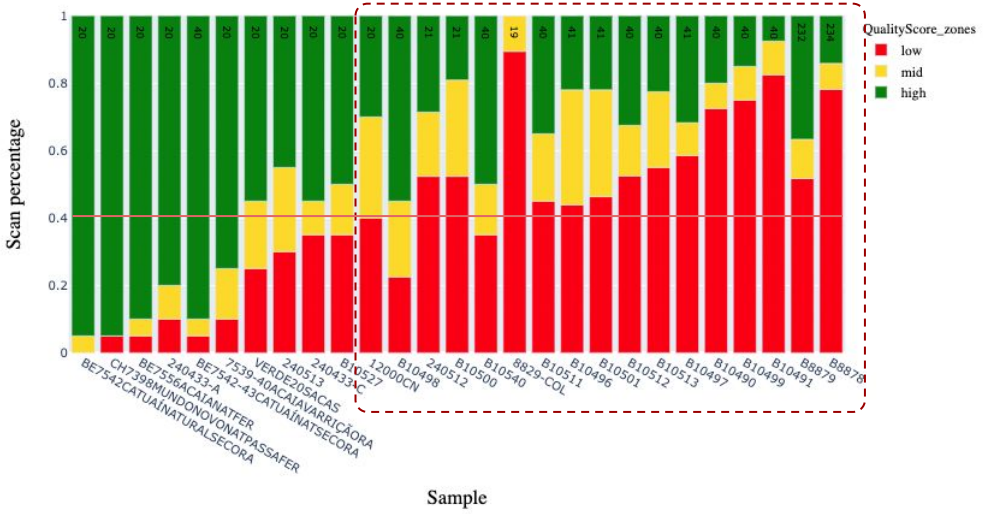
Percentage of predictions in probability zones, 209/209 samples



Un caso específico - Café con malos sabores

After analyzing samples with reported defects, We can identify a group of the worst samples that shows a big proportion of beans with low quality.

Percentage of predictions in probability zones, 27/27 samples



… ¿Se resolvió el problema? …

Descripción de la evaluación.

“Nuestra herramienta demuestra la capacidad de detectar correos … y priorizar…”

- Identificando correos de solicitud de … en un 85%
- Priorización de intervención humana en un 65%
    - reducción de gastos de operación…

Beneficios.

1. Detección temprana de …
2. Optimización de asignación de …

Riesgos de NO usarlo.

1. Lentitud en la correcta asignación…
2. Generación de inconformidad en el…

Gracias!!!