Gestión de Empresas de Base Tecnológica

Machine Learning - Operations

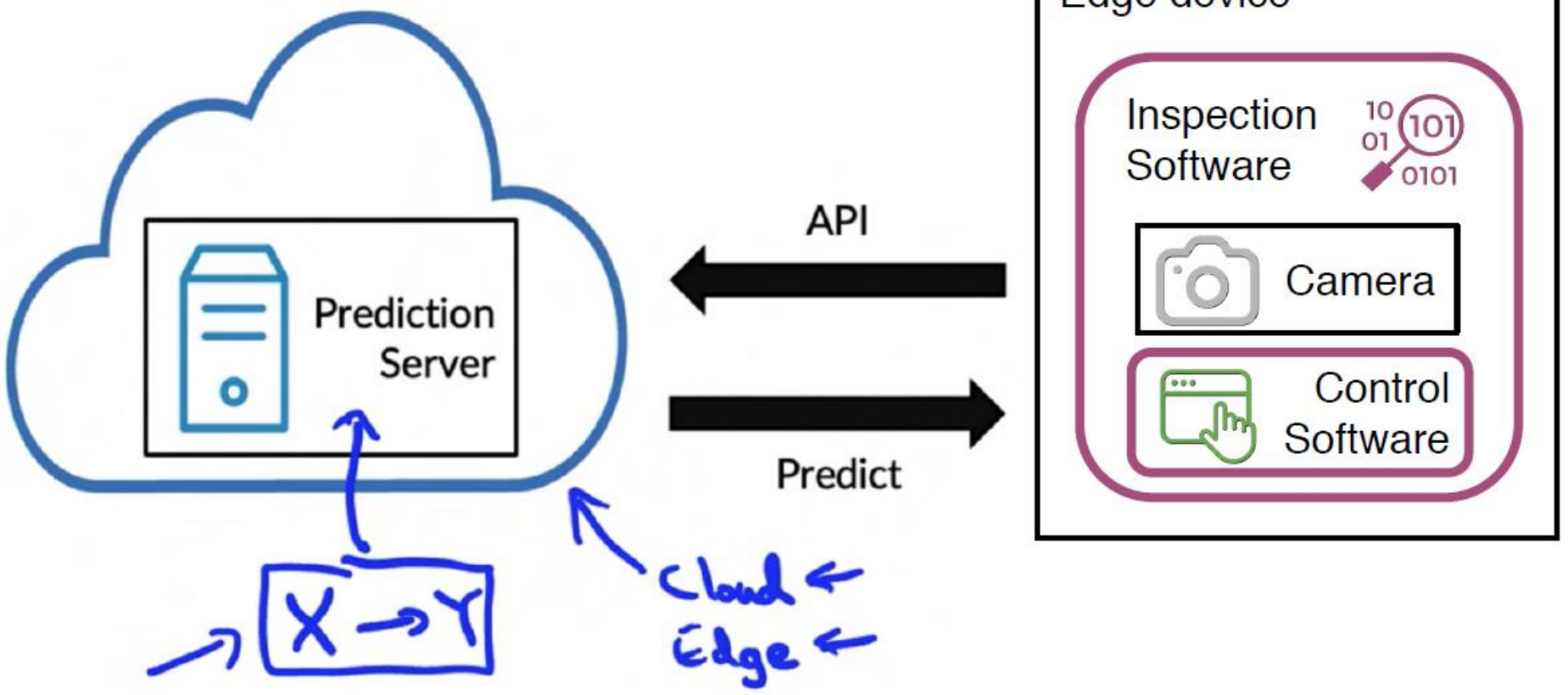
Máster en Ingeniería Informática
Universidad Complutense de Madrid, 2022-2023

Humberto Martínez Silva

Basado en el curso “Introduction to Machine Learning in Production” de COURSERA (*deeplearning.ai)*

Deployment example

Photo from camera

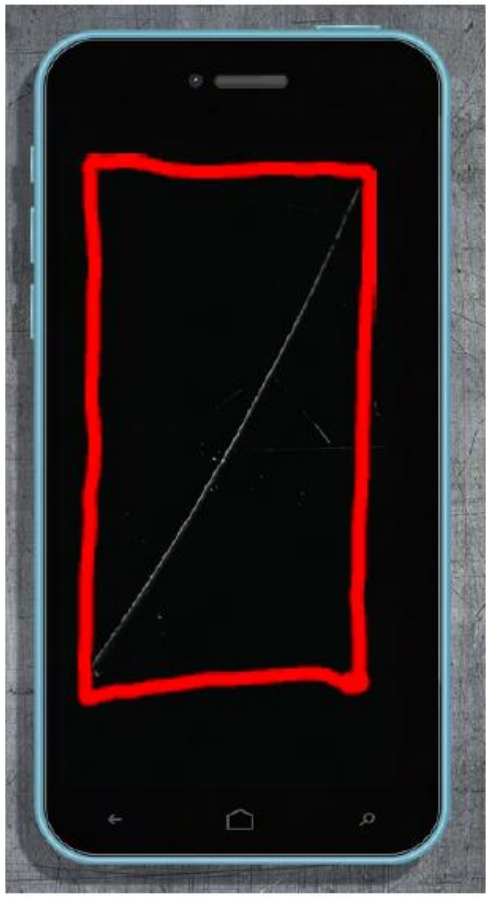


¿Cuál de los siguientes puntos constituye una ventaja de tener el servicio de predicción en *cloud* en comparación con tenerlo en un servicio *edge?*

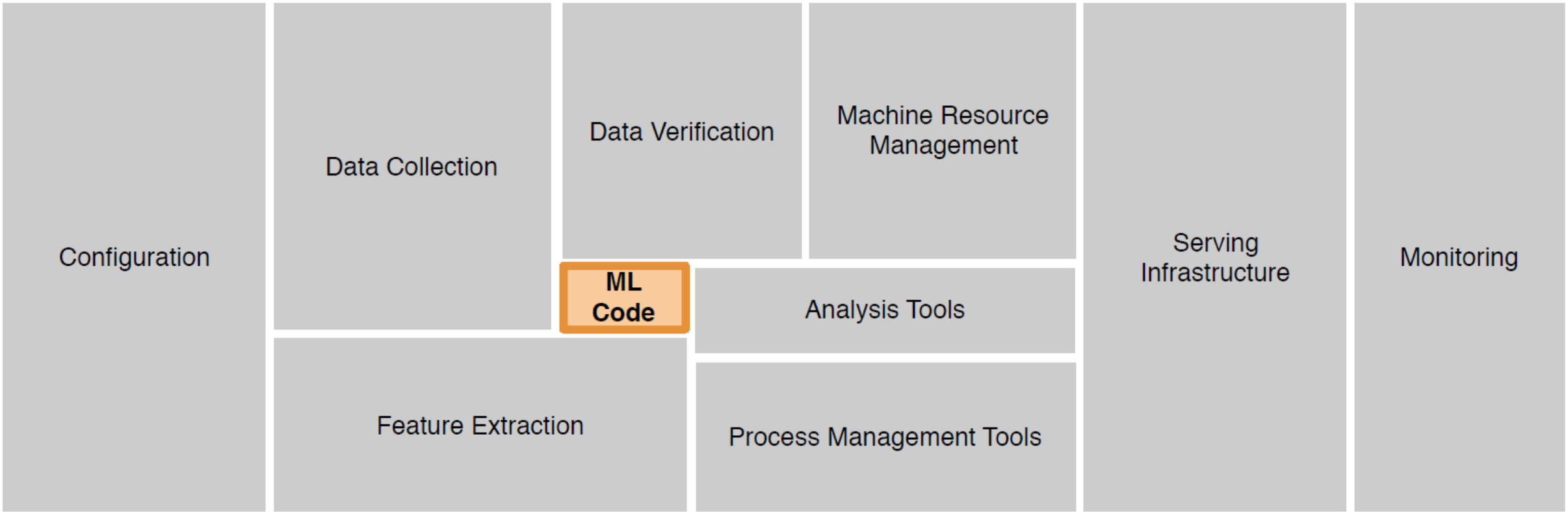
- ❑ Más potencia computacional disponible.
- ❑ Se necesita menos ancho de banda de red.
- ❑ Puede funcionar incluso si la conexión de red tiene problemas.
- ❑ Menor latencia.

Ref: deeplearning.ai

Visual inspection example

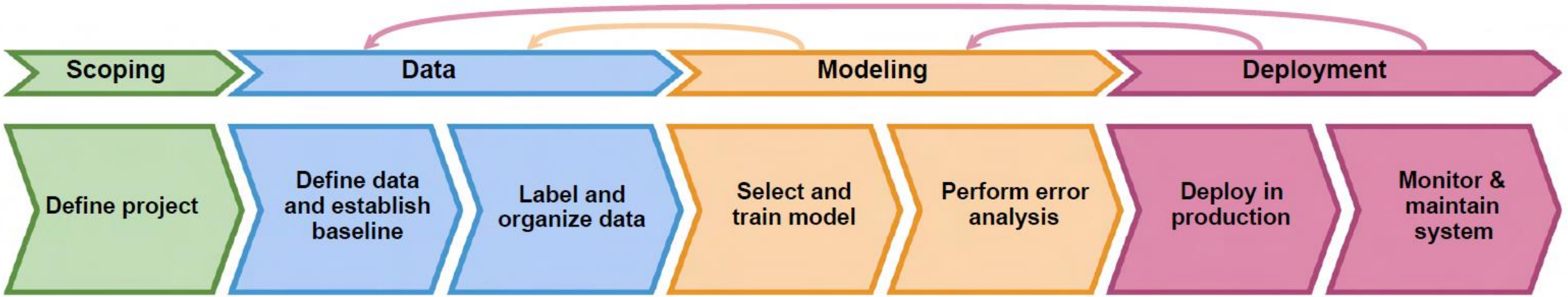


The requirements surrounding ML infrastructure



[D. Sculley et. al. NIPS 2015: Hidden Technical Debt in Machine Learning Systems]

The ML project lifecycle



X → Y

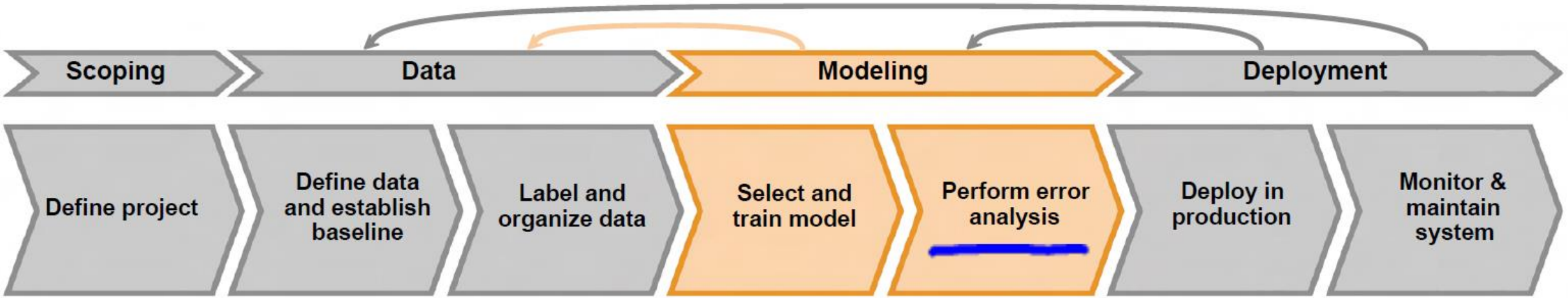
Software engineering issues

Checklist of questions

- Realtime or Batch
- Cloud vs. Edge/Browser
- Compute resources (CPU/GPU/memory)
- Latency, throughput (QPS)
- Logging
- Security and privacy

500ms, 1000 QPS

Speech recognition: Modeling stage



Research/Academia

Product Team

Code (algorithm/model)

Hyperparameters

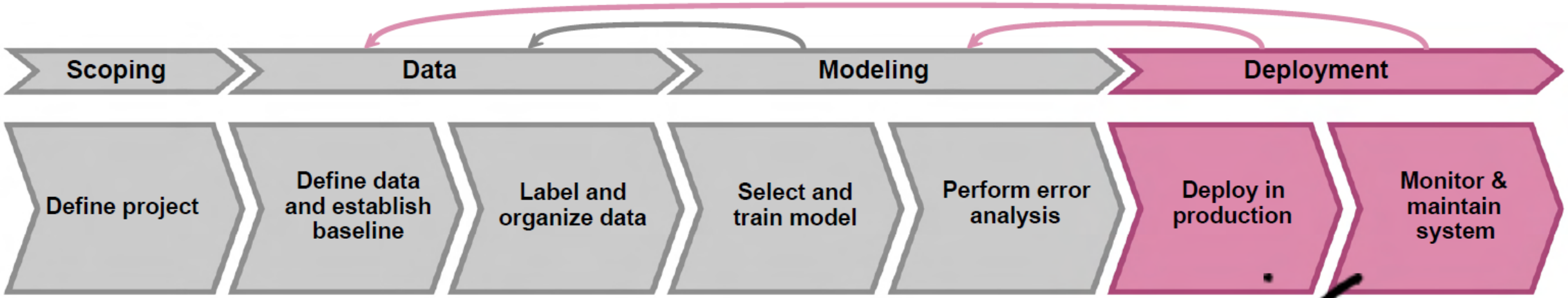
Data

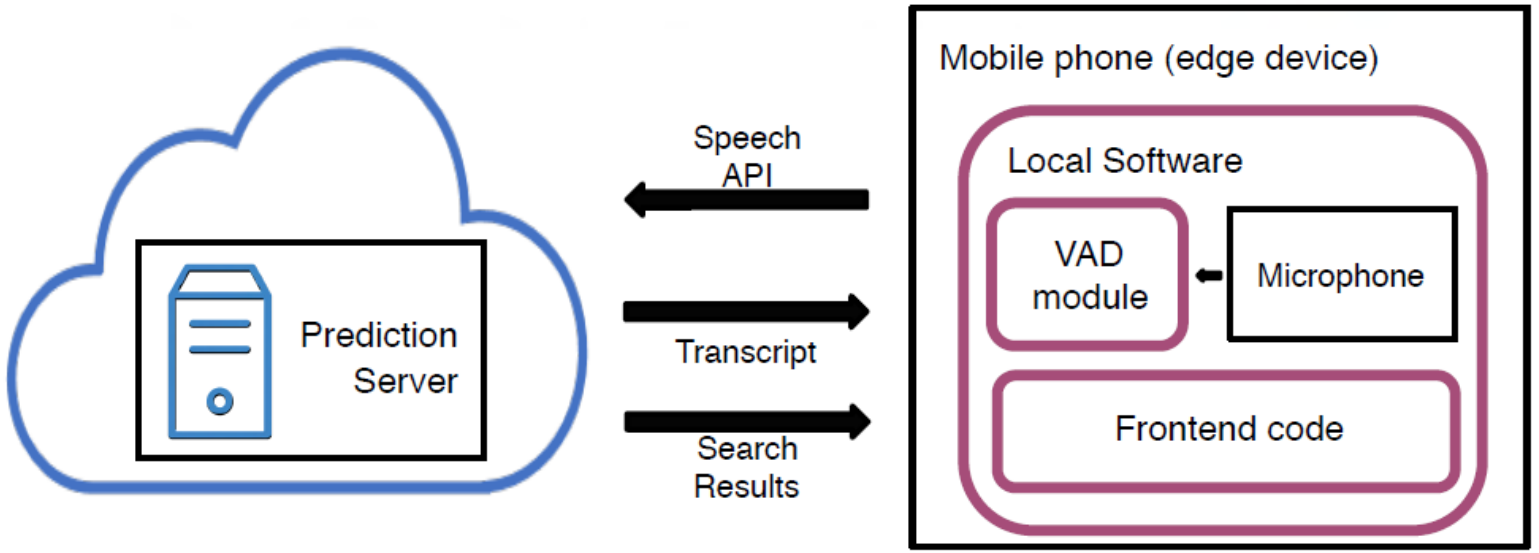
ML Model

X → Y

ML System = Code + Data

Speech recognition: Deployment stage





Common deployment cases

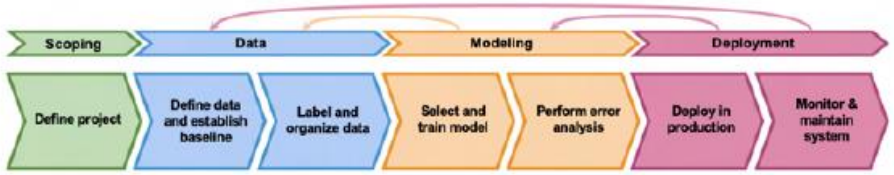
1. New product/capability
2. Automate/assist with manual task
3. Replace previous ML system

Key ideas:

- Gradual ramp up with monitoring
- Rollback

Visual inspection example

shadow mode



Human ✓

ML ✓

Human ✗

ML ✗

Human ✗

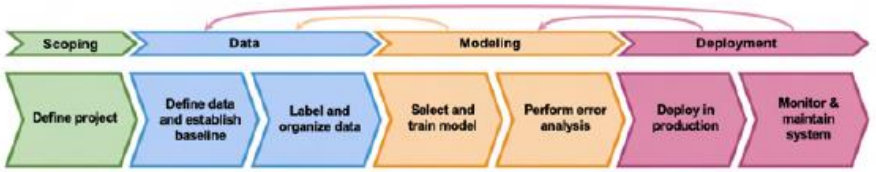
ML ✓

ML system shadows the human and runs in parallel.

ML system's output not used for any decisions during this phase.

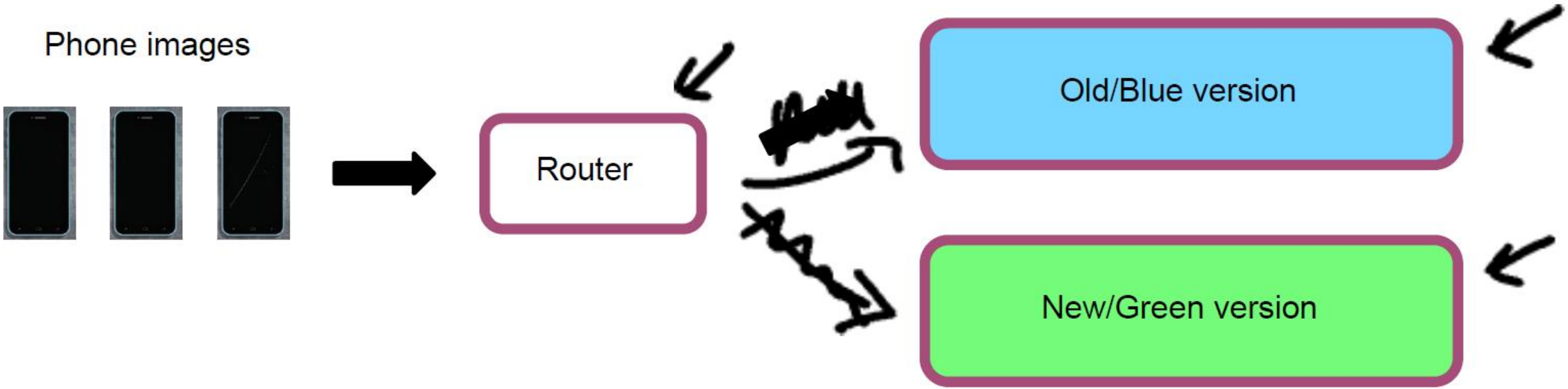
Sample outputs and verify predictions of ML system.

Canary deployment



- Roll out to small fraction (say 5%) of traffic initially.
- Monitor system and ramp up traffic gradually.

Blue green deployment



Easy way to enable rollback

Degrees of automation

Human

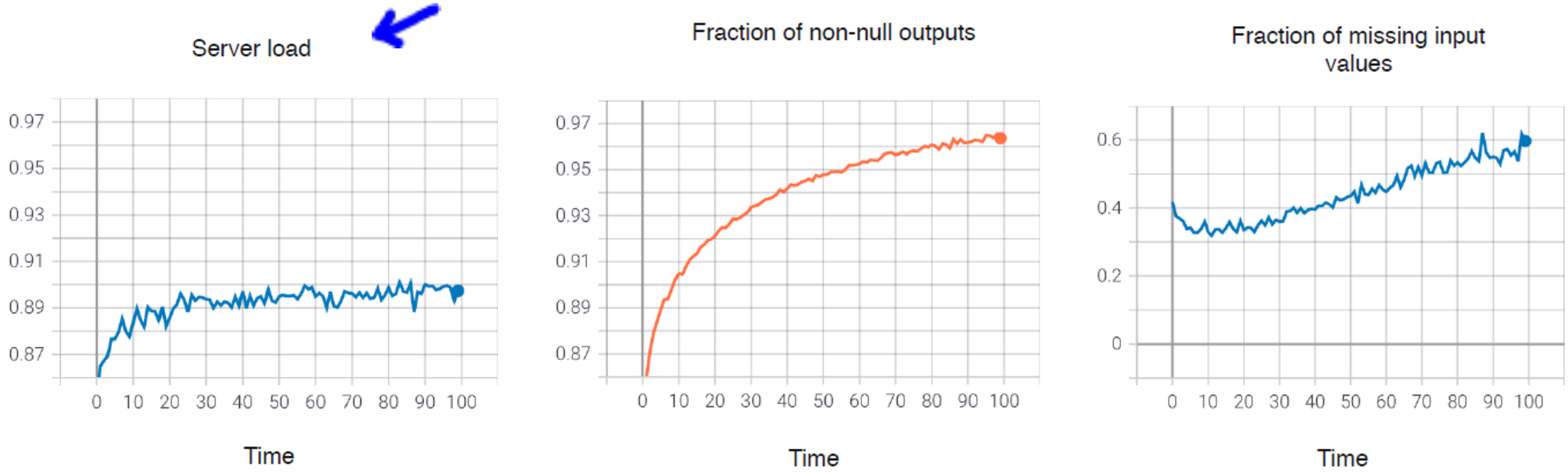
AI

Human only → Shadow mode → AI assistance → Partial automation → Full automation

human in the loop

You can choose to stop before getting to full automation.

Monitoring dashboard

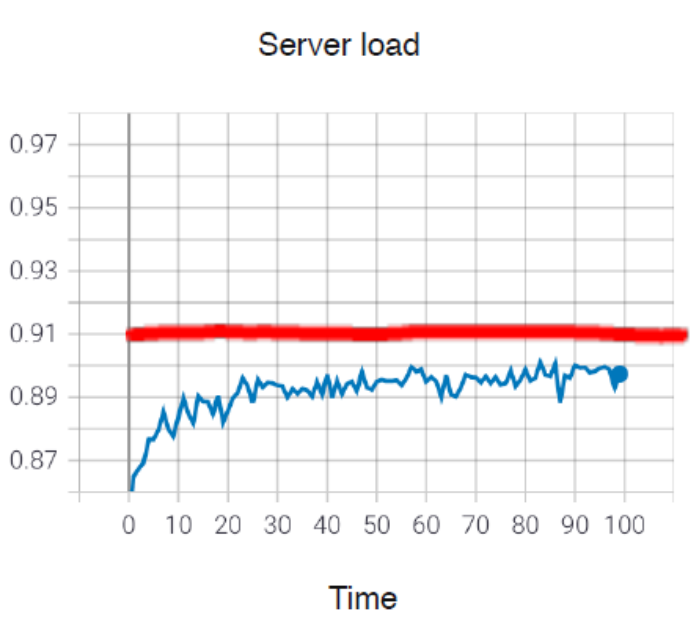


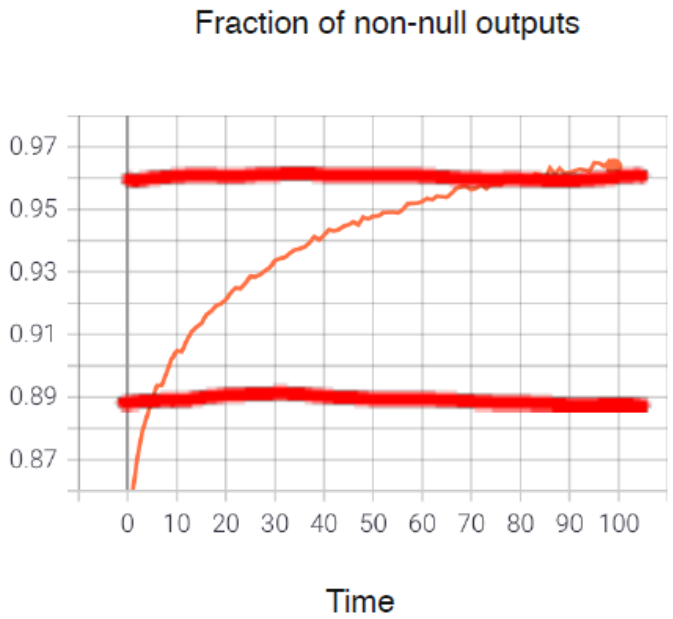
- Brainstorm the things that could go wrong.
- Brainstorm a few statistics/metrics that will detect the problem.
- It is ok to use many metrics initially and gradually remove the ones you find not useful.

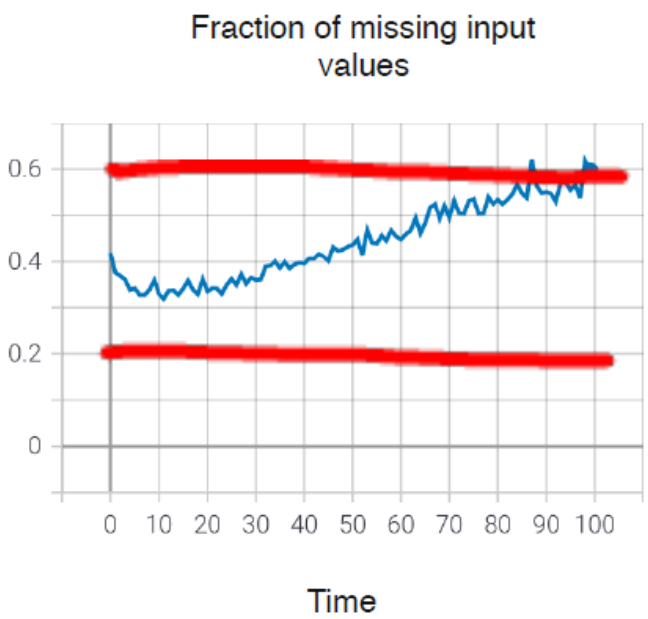
Examples of metrics to track

Software metrics:	Memory, compute, latency, throughput, server load
Input metrics: 	Avg input length Avg input volume Num missing values Avg image brightness
Output metrics: 	# times return " " (null) # times user redoes search # times user switches to typing CTR

Monitoring dashboard

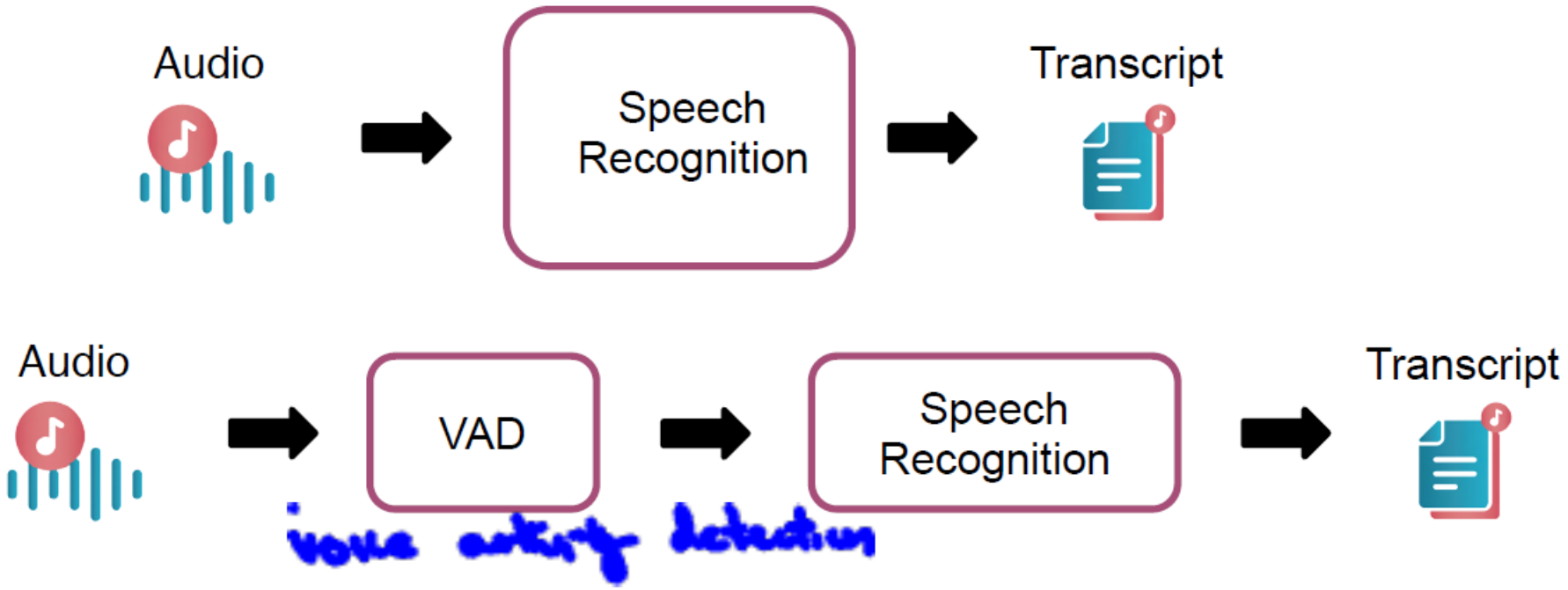






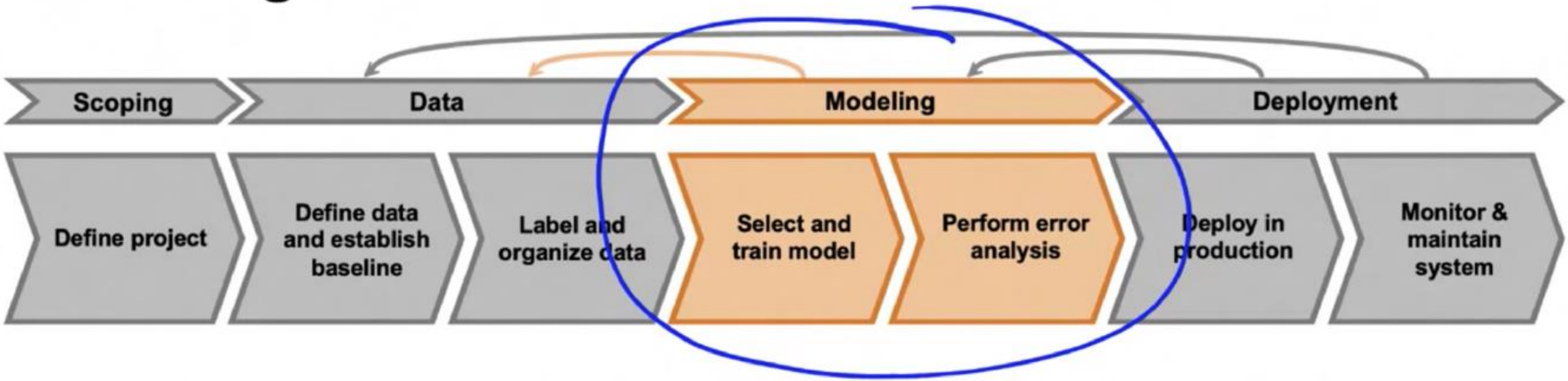
- Set thresholds for alarms
- Adapt metrics and thresholds over time

Speech recognition example



Some cellphones might have VAD clip audio differently, leading to degraded performance

Modeling



Model-centric AI development

Data-centric AI development

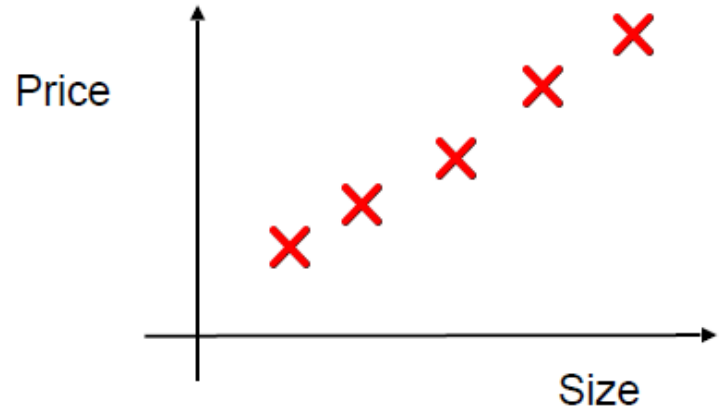
AI system = Code + Data

(algorithm/model)

Challenges in model development

1. Doing well on training set (usually measured by average training error).

2. Doing well on dev/test sets.



3. Doing well on business metrics/project goals.

Performance on key slices of the dataset

Example: ML for loan approval

Make sure not to discriminate by ethnicity, gender, location, language or other protected attributes.

Example: Product recommendations from retailers

Be careful to treat fairly all major user, retailer, and product categories.

Supongamos que tenemos un algoritmo que diagnostica enfermedades a partir de radiografías médicas y logra una alta precisión en el *test set*. Teniendo en cuenta esa sola afirmación, marque todas las respuestas correctas.

- ❑ El algoritmo lo hace bien incluso en enfermedades menos conocidas.
- ❑ Sus diagnósticos son aproximadamente igual de precisos en todos los géneros y etnias, por lo que estamos seguros de que no está sesgado.
- ❑ El sistema se puede implantar de forma segura en un entorno de atención médica real.
- ❑ Ninguna de las respuestas anteriores es correcta.

Ref: deeplearning.ai

Establishing a baseline

Speech recognition example

Type	Accuracy
Clear Speech	94%
Car Noise	89%
People Noise	87%
Low Bandwidth	70%

Establishing a baseline level of performance

Speech recognition example:

Type	Accuracy	Human level performance	
Clear Speech	94%	95%	1%
Car Noise	89%	93%	4%
People Noise	87%	89%	2%
Low Bandwidth	70%	70%	~0%

HLP

Ways to establish a baseline

- Human level performance (HLP)
- Literature search for state-of-the-art/open source
- Older system

Baseline gives an estimate of the irreducible error / Bayes error and indicates what might be possible.

Prioritizing what to work on

Type	Accuracy	Human level performance	Gap to HLP
Clean Speech	94%	95%	1%
Car Noise	89%	93%	4%
People Noise	87%	89%	2%
Low Bandwidth	70%	70%	0%

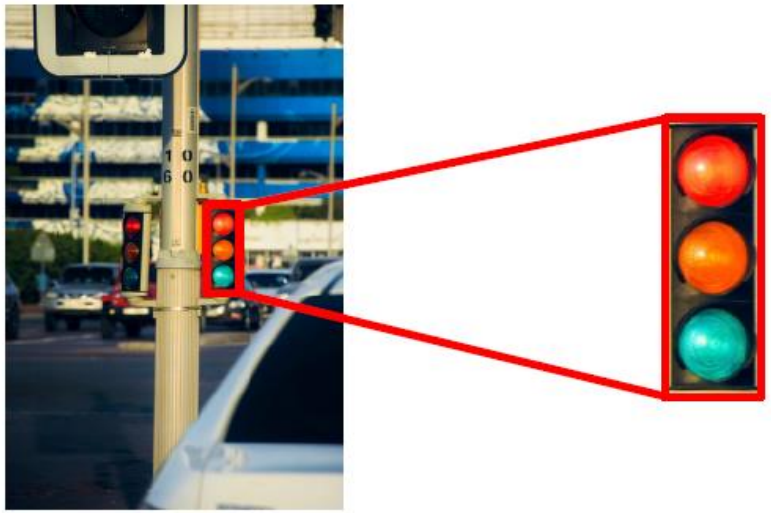
Prioritizing what to work on

Type	Accuracy	Human level performance	Gap to HLP	% of data
Clean Speech	94%	95%	1%	60% → 0.6%
Car Noise	89%	93%	4%	4% → 0.16%
People Noise	87%	89%	2%	30% → 0.6%
Low Bandwidth	70%	70%	0%	6% → ~0%

Why use HLP to benchmark?

People are very good on unstructured data tasks

Criteria: Can a human, given the same data, perform the task?



Prioritizing what to work on

Decide on most important categories to work on based on:

- How much room for improvement there is.
- How frequently that category appears.
- How easy is to improve accuracy in that category.
- How important it is to improve in that category.

Estamos trabajando sobre un problema de inspección visual de coches.
Supongamos que incluso un inspector humano que mira una imagen no puede decir si hay un rasguño. Sin embargo, si el mismo inspector mirara el coche directamente (en lugar de una imagen), entonces sí podría decirlo.
¿Qué harías?

- ❑ Obtener muchos más datos para entrenar el algoritmo.

- ❑ Intentar mejorar el sistema de imágenes (cámara / iluminación / resolución) para mejorar la calidad de las imágenes.

- ❑ Intentar mejorar la consistencia del etiquetado.

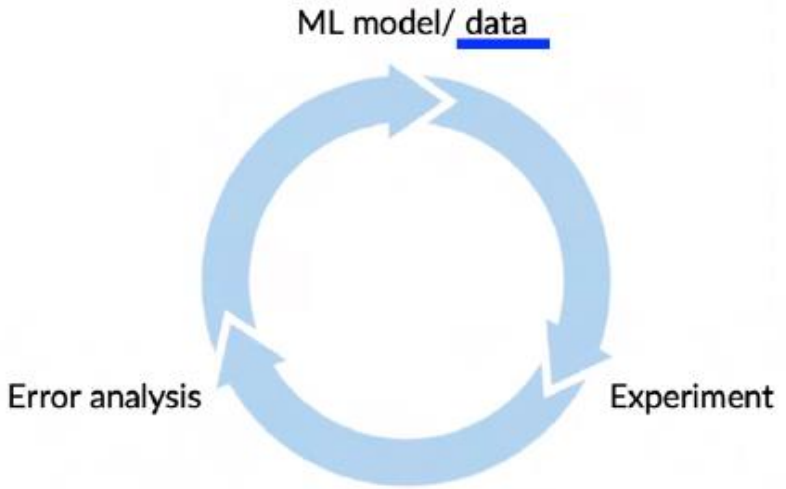
- ❑ Contratar mejores etiquetadores.

Adding data

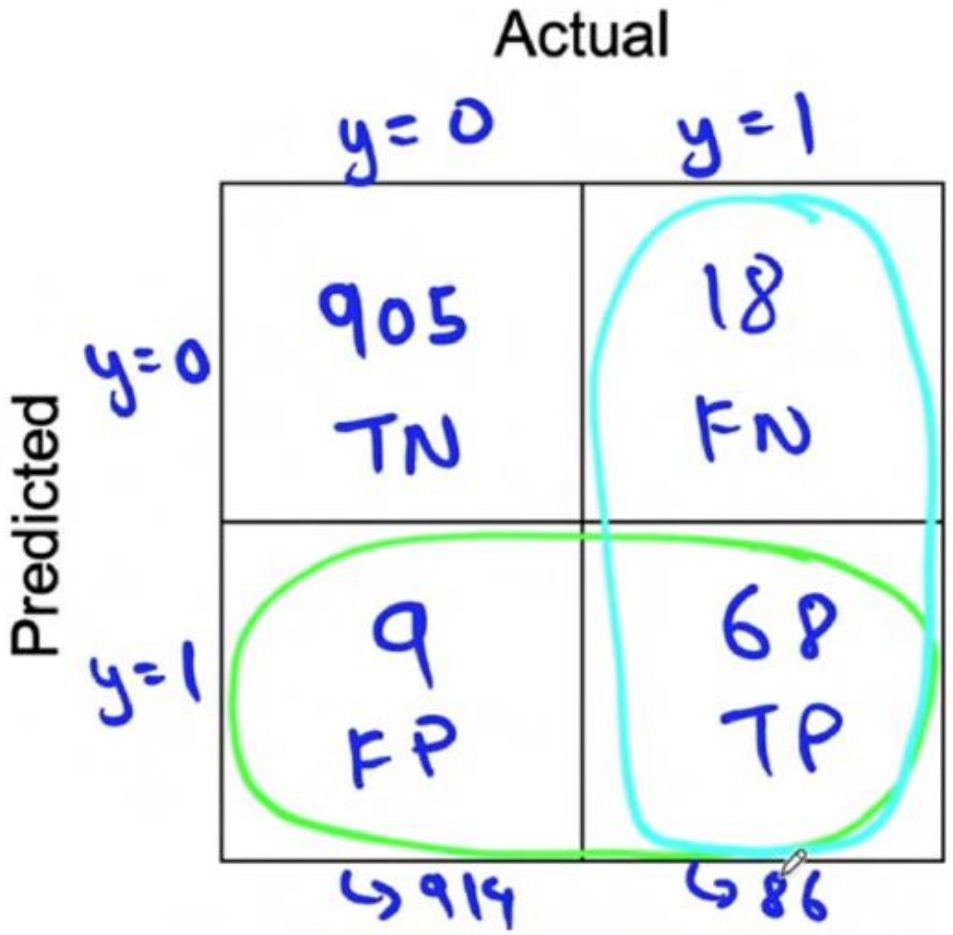
For categories you want to prioritize:

- Collect more data (or improve label accuracy)
- Use data augmentation to get more data

Type	Accuracy	Human level performance	Gap to HLP	% of data
Clean Speech	94%	95%	1%	60%
Car Noise	84%	93%	4%	40%
People Noise	87%	84%	2%	30%
Low Bandwidth	70%	70%	0%	6%



Confusion matrix: precision and recall



TN: True Negative

TP: True Positive

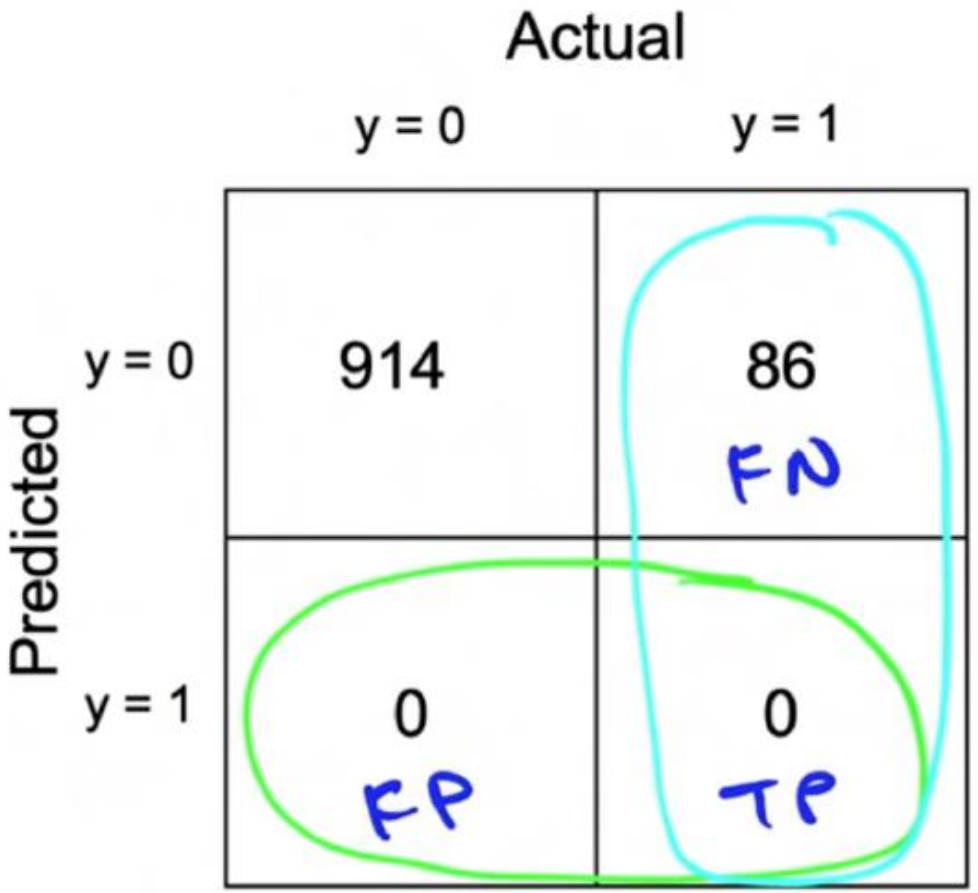
FN: False Negative

FP: False Positive

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{68}{68+9} = 88.3\%$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{68}{68+18} = 79.1\%$$

What happens with print("0")?



$$\text{Precision} = \frac{TP}{TP + FP} = \frac{0}{0+0}$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{0}{0+86} = 0\%$$

Combining precision and recall – F_1 score

	Precision (P)	Recall (R)	F_1
Model 1	88.3	79.1	83.4 %
Model 2	97.0	7.3	13.6 %

$$F_1 = \frac{2}{\frac{1}{P} + \frac{1}{R}}$$

Multi-class metrics

Classes: Scratch, Dent, Pit mark, Discoloration

Defect Type	Precision	Recall	F_1
Scratch	82.1%	99.2%	89.8%
Dent	92.1%	99.5%	95.7%
Pit mark	85.3%	98.7%	91.5%
Discoloration	72.1%	97%	82.7%

Estás trabajando en un algoritmo ML de clasificación binaria que detecta si un paciente tiene una enfermedad específica. En su conjunto de datos, el 98% de los ejemplos de entrenamiento (pacientes) no tienen la enfermedad, por lo que el conjunto de datos está muy *sesgado (no balanceado)*. ¿Qué métrica te daría mayor confianza de que se ha construido un sistema útil y no trivial? (Seleccione uno).

- ❑ Precision
- ❑ Accuracy
- ❑ Recall
- ❑ F1 score

Ref: deeplearning.ai

En el problema anterior, con un 98% de ejemplos negativos, si tu algoritmo es print("1") (es decir, dice que todos tienen la enfermedad). ¿Cuál de estas afirmaciones es cierta?

- ❑ El algoritmo consigue 100% recall.
- ❑ El algoritmo consigue 100% precision.
- ❑ El algoritmo consigue 0% precision.
- ❑ El algoritmo consigue 0% recall.

Ref: deeplearning.ai

Auditing framework

Check for accuracy, fairness and bias.

1. Brainstorm the ways the system might go wrong.
   - Performance on subsets of data (e.g., ethnicity, gender).
   - Prevalence of specific errors/outputs (e.g., FP, FN).
   - Performance on rare classes.

2. Establish metrics to assess performance against these issues on appropriate slices of data.

3. Get business/product owner buy-in.

Data-centric AI development

Model-centric view

Collect what data you can, and develop a model good enough to deal with the noise in the data.

Hold the data fixed and iteratively improve the code/ model.

Data-centric view

The consistency of the data is paramount. Use tools to improve the data quality; this will allow multiple models to do well.

Hold the code fixed and iteratively improve the data.

Data augmentation

Goal:

Create realistic examples that (i) the algorithm does poorly on, but (ii) humans (or other baseline) do well on

Checklist:

- Does it sound realistic?
- Is the X → Y mapping clear? (e.g., can humans recognize speech?)
- Is the algorithm currently doing poorly on it?

Image example

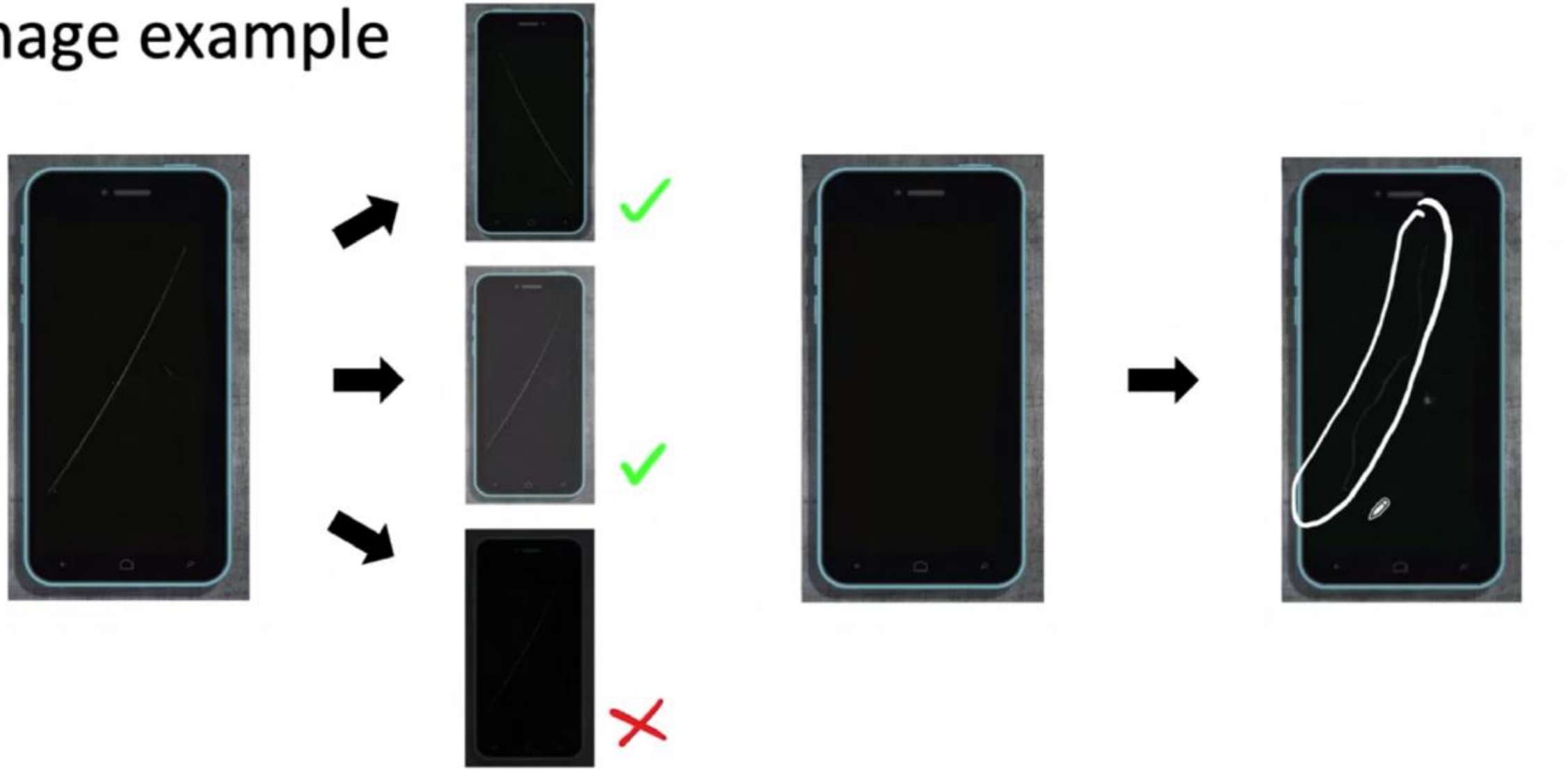


Photo OCR counterexample

1

high accuracy

I

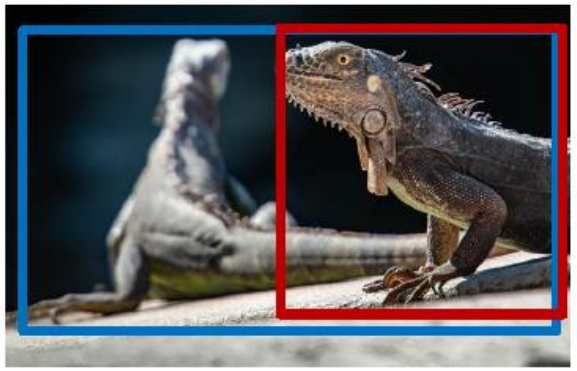
421

low accuracy

1? I?

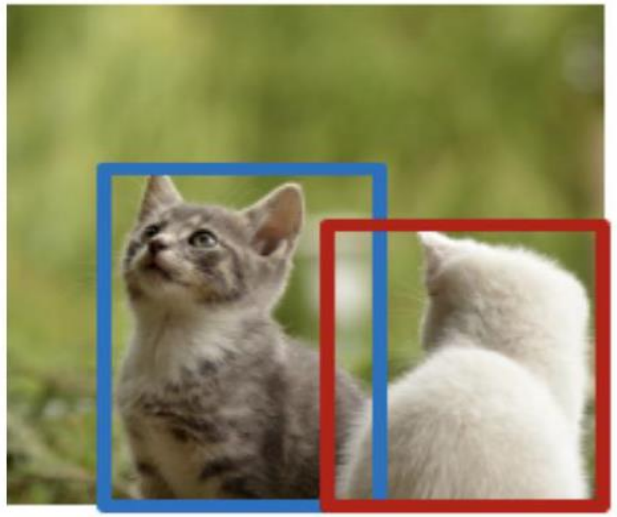
Adding a lot of new "I"s may skew the dataset and hurt performance

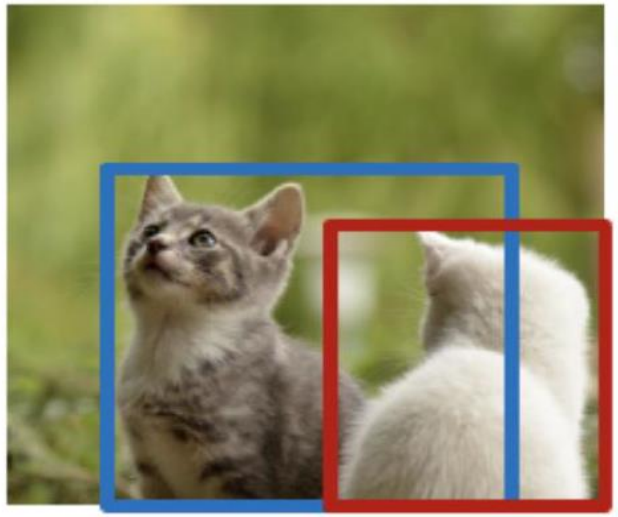
Iguana detection example



Labeling instructions: "Use bounding boxes to indicate the position of iguanas"

Estás construyendo un sistema para detectar gatos. Los diferentes etiquetadores etiquetan de la siguiente manera. ¿A qué puede deberse?





- ❑ No se está realizando supervisión sobre el trabajo de los etiquetadores.
- ❑ Se está etiquetando correctamente, no es posible aumentar más el performance del algoritmo.
- ❑ Instrucciones ambiguas para el etiquetado.
- ❑ Los etiquetadores no confían el uno en el otro.

Major types of data problems

	Unstructured	Structured
Small data	Manufacturing visual inspection from 100 training examples	Housing price prediction based on square footage, etc. from 50 training examples
Big data	Speech recognition from 50 million training examples	Online shopping recommendations for 1 million users

≤ 10,000

Clean labels are critical.

> 10,000

Emphasis on data process.

Humans can label data.

Harder to obtain more data.

Data augmentation.

Unstructured vs. structured data

Unstructured data

- May or may not have huge collection of unlabeled examples x.
- Humans can label more data.
- Data augmentation more likely to be helpful.

Structured data

- May be more difficult to obtain more data.
- Human labeling may not be possible (with some exceptions).

Small data vs. big data

Small data

- Clean labels are critical.
- Can manually look through dataset and fix labels.
- Can get all the labelers to talk to each other.

Big data

- Emphasis data process.

¿Con cuál de estas afirmaciones está de acuerdo con respecto a los problemas de datos estructurados frente a los no estructurados?

- En general, es más fácil para los humanos etiquetar *datos estructurados* y es más fácil aplicar el aumento de datos en *datos no estructurados*.

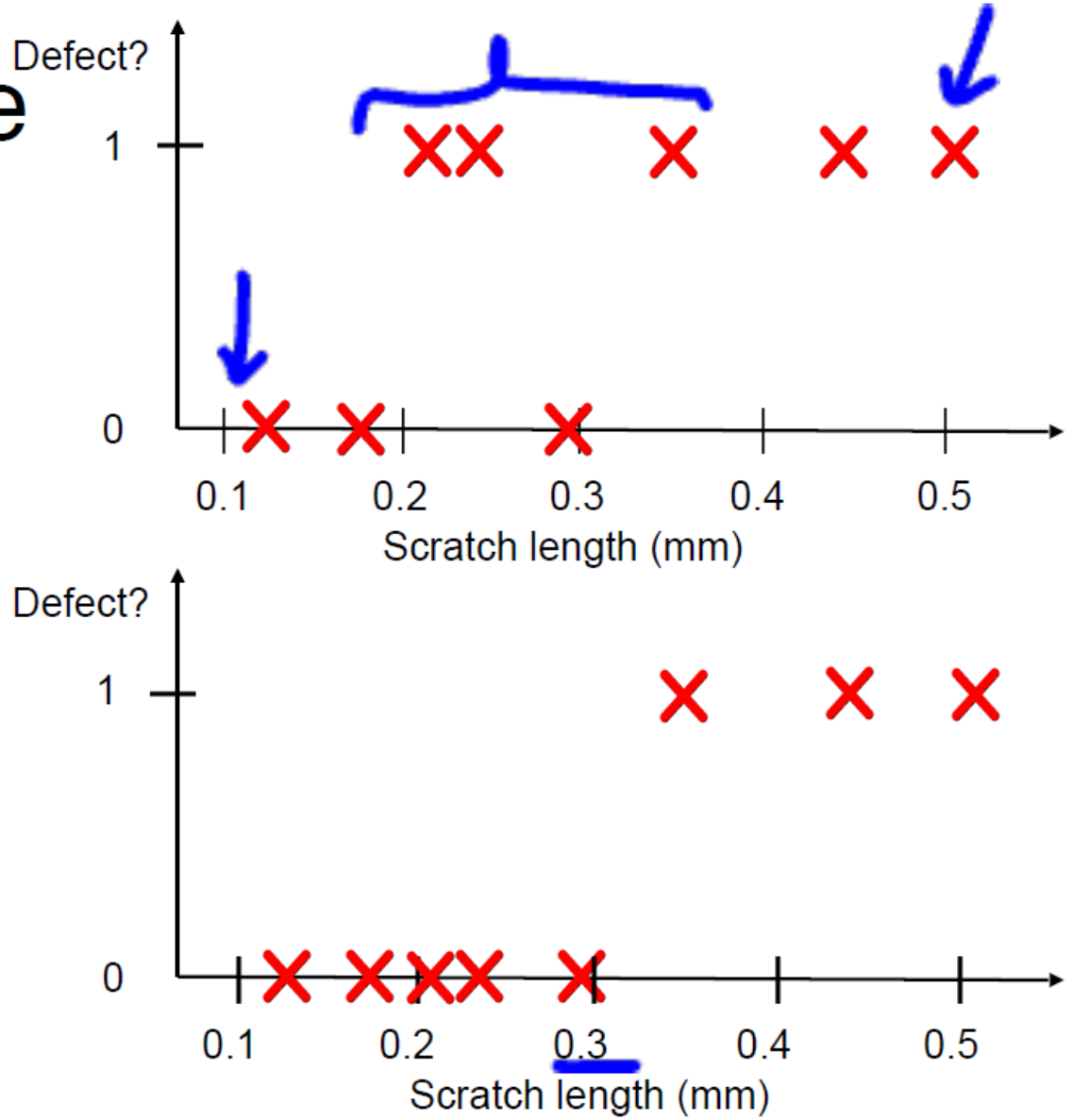
- En general, es más fácil para los humanos etiquetar *datos no estructurados* y es más fácil aplicar el aumento de datos en *datos no estructurados*.

- En general, es más fácil para los humanos etiquetar *datos no estructurados* y es más fácil aplicar el aumento de datos en *datos estructurados*.

- En general, es más fácil para los humanos etiquetar *datos estructurados* y es más fácil aplicar el aumento de datos en *datos estructurados*.

Ref: deeplearning.ai

Phone defect example



Improving label consistency

- Have multiple labelers label same example.
- When there is disagreement, have MLE, subject matter expert (SME) and/or labelers discuss definition of y to reach agreement.
- If labelers believe that x doesn't contain enough information, consider changing x.
- Iterate until it is hard to significantly increase agreement.

Examples

- Standardize labels

  "Um, nearest gas station"

  "Umm, nearest gas station"

  ”Nearest gas station [unintelligible]”

  ➡ "Um, nearest gas station"

- Merge classes

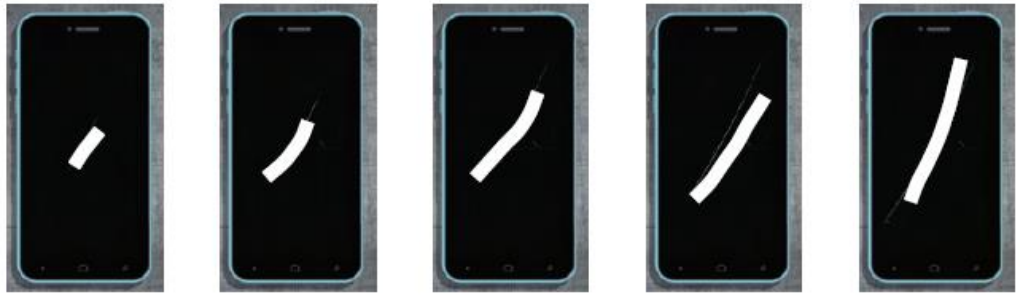
Deep scratch

Shallow scratch

➡ Scratch

Have a class/label to capture uncertainty

- Defect: 0 or 1



Alternative: 0, Borderline, 1

- Unintelligible audio

“nearest go”

“nearest grocery”

“nearest [unintelligible]”

Why measure HLP?

- Estimate Bayes error / irreducible error to help with error analysis and prioritization.

99%

Ground Truth Label	Inspector
1	1 ✓
1	0 ✗
1	1 ✓
0	0 ✓
0	0 ✓
0	1 ✗

↑ Human²

66.7% accuracy

Raising HLP

When the ground truth label is externally defined, HLP gives an estimate for Bayes error / irreducible error.

But often ground truth is just another human label.

Ground Truth Label	Inspector
1	1
1	0
1	1
0	0
0	0
0	1

Estamos construyendo un sistema de inspección visual. El análisis de errores *(error analysis)* encuentra:

Type of defect	Accuracy	HLP	% of data
Scratch	95%	98%	50%
Discoloration	90%	90%	50%

Crees que podría valer la pena verificar la consistencia del *labeling* tanto en los defectos de rayado (*scratch*) como en los defectos de decoloración (*discoloration*). Si tuvieras que elegir uno para empezar, ¿cuál elegirías?

- Es más prometedor verificar (y potencialmente mejorar) la consistencia del *labeling* en los defectos de rayado (*scratch*) que los defectos de decoloración (*discoloration*), ya que el HLP es más alto en los defectos de rayado *(scratch)* y, por lo tanto, es más razonable esperar una alta consistencia.

- Es más prometedor verificar (y potencialmente mejorar) la consistencia del *labeling* en defectos de *discoloration* que en defectos de rayado *(scratch)*. Dado que el HLP es más bajo en la decoloración, es posible que haya instrucciones de etiquetado ambiguas que estén afectando al HLP.

Ref: deeplearning.ai

Estamos construyendo un sistema de inspección visual. El análisis de errores *(error analysis)* encuentra:

Type of defect	Accuracy	HLP	% of data
Scratch	95%	98%	50%
Discoloration	90%	90%	50%

Si el *HLP* no es mejorable, ¿en cuál de los dos tipos de defectos priorizaríamos seguir mejorando?

- Es más prometedor enfocarse en los defectos de rayado (*scratch*) que en los defectos de decoloración (*discoloration*).

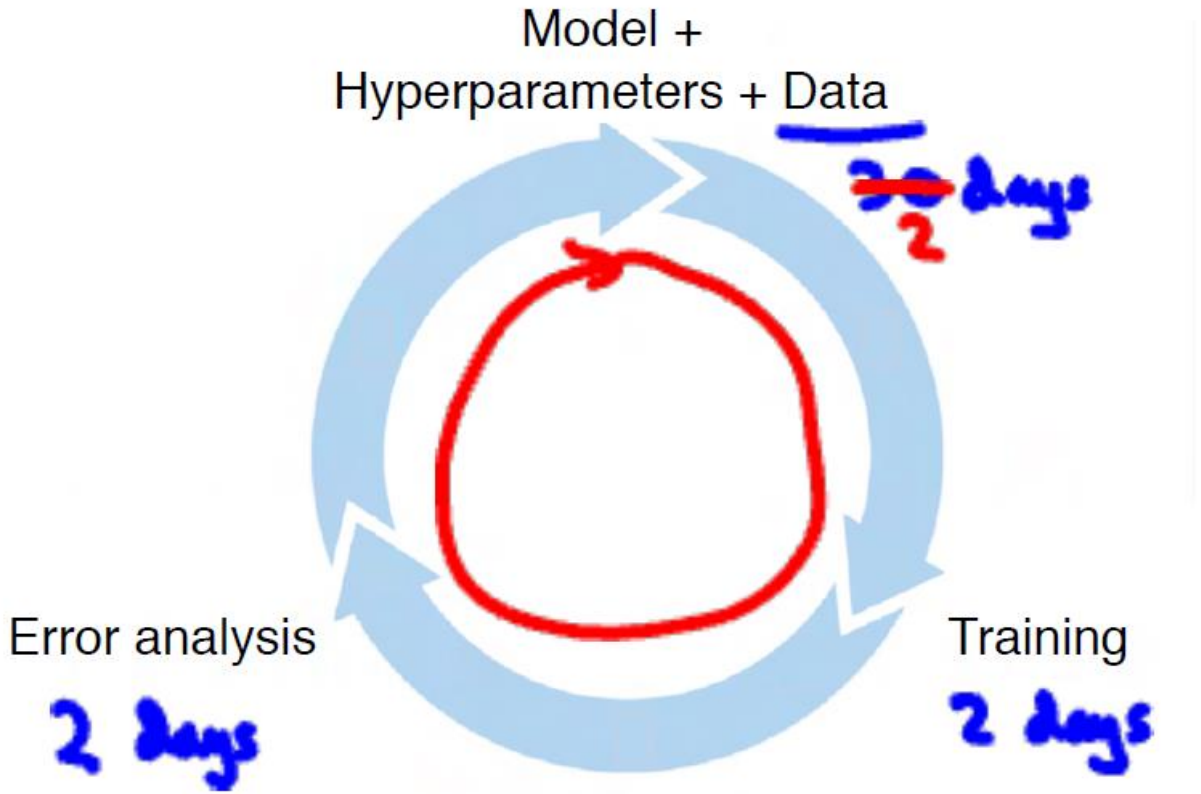
- Es más prometedor enfocarse en los defectos de decoloración (*discoloration*) que en los defectos de rayado (*scratch*).

Ref: deeplearning.ai

Algunos etiquetadores traducen los testimonios del juicio como "..." (como en, "*Um... él le asaltó*") mientras que otros lo hacen con obviando el “titubeo” (“*Él le asaltó*”). El *Human level performance* (HLP) se mide de acuerdo con lo bien que un traductor está de acuerdo con otro. Trabajas con el equipo y consigues que todos prescindan del “titubeo”. ¿Qué efecto tendrá esto en HLP?

- HLP disminuirá.
- HLP seguirá siendo el mismo.
- HLP aumentará.

How long should you spend obtaining data?



- Get into this iteration loop as quickly possible.
- Instead of asking: How long it would take to obtain m examples?
  Ask: How much data can we obtain in k days.
- Exception: If you have worked on the problem before and from experience you know you need m examples.

Inventory data

Brainstorm list of data sources ( speech recognition)

Source	Amount	Cost	
Owned	100h	$0	✓
Crowdsourced – Reading	1000h	$10000	
Pay for labels	100h	$6000	
Purchase data	1000h	$10000	✓

Other factors: Data quality, privacy, regulatory constraints

Es siempre muy recomendable tomarse todo el tiempo necesario para construir primero el *dataset* correcto, recopilar todos los datos posibles antes de iniciar el aprendizaje..

- ❑ True
- ❑ False

Está considerando construir un sistema de reconocimiento de voz, con la innovación de usar un micrófono de muy bajo coste que acaba de ser inventado. Supongamos que un humano, dado un clip de audio grabado desde este nuevo micrófono, a menudo no puede decir lo que se dijo (porque el micrófono tiene demasiado ruido). ¿Qué puede concluir de esto?

- ❑ Existe una alta probabilidad de que no sea técnicamente posible que un algoritmo de aprendizaje logre una alta precisión en esta tarea.
- ❑ Este problema será técnicamente factible sólo si podemos adquirir más datos para el entrenamiento.
- ❑ Deberíamos conseguir múltiples etiquetadores para transcribir cada clip de audio y así reducir el ruido.
- ❑ Este problema es técnicamente factible sólo si un humano escucha el sonido original con sus propios oídos (no a través del sistema de micrófono) y puede descifrar lo que se dijo.

Ref: deeplearning.ai