ML in Production

Part 3: Data Definition and Baseline

Ramin Toosi

Why is data definition hard?

Why is data definition hard?

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

Why is data definition hard?

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

Why is data definition hard?

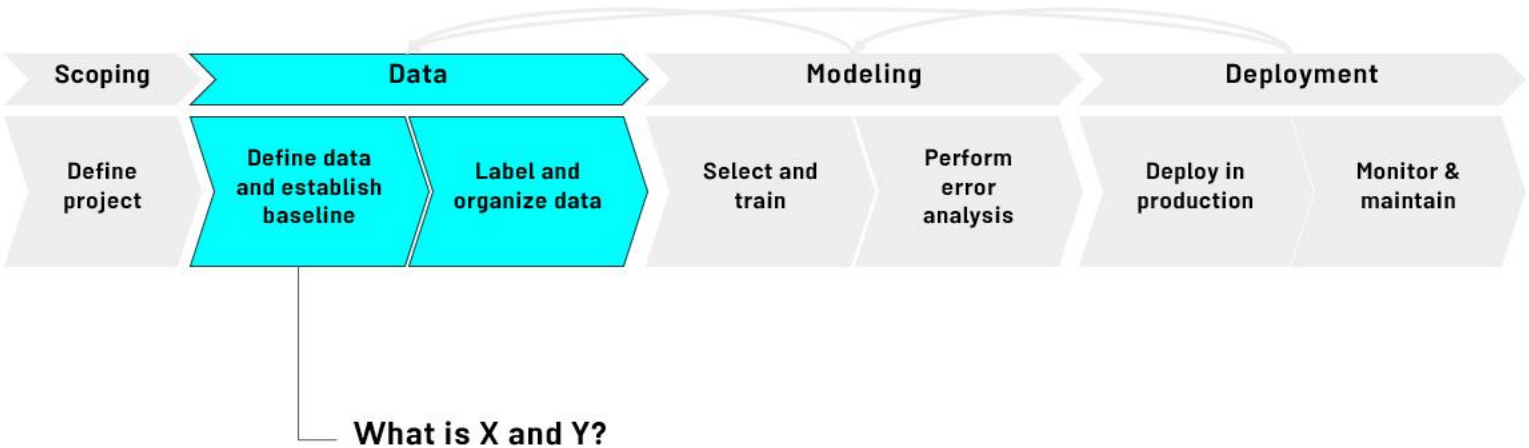
Labeling instructions: "Use bounding boxes to indicate the position of iguanas"

Why is data definition hard?

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

Phone defect detection

Data stage


Scoping
Data
Modeling
Deployment
Define project
Define data and establish baseline
Label and organize data
Select and train
Perform error analysis
Deploy in production
Monitor & maintain
What is X and Y?

User ID merge example

	Job Board (website)	Resume chat (app)
Email	ramin@avir.co.com	rtoosi@mci.ir
First Name	Ramin	Ramin
Last Name	Toosi	Tsi
Address	Vanak	?
State	Tehran	?
Zip	9393	9393

User ID merge example

	Job Board (website)	Resume chat (app)
Email	ramin@avir.co.com	rtoosi@mci.ir
First Name	Ramin	Ramin
Last Name	Toosi	Tsi
Address	Vanak	?
State	Tehran	?
Zip	9393	9393

- Bot detection?
- How a human decides?

Data definition questions

- What is the input x?
  - Lightning? Contrast? Resolution?
  - What features need to be included?
- What is the target label y?
  - How can we ensure labelers give consistent labels?

Data definition questions

- What is the input x?
  - Lightning? Contrast? Resolution?
  - What features need to be included?
- What is the target label y?
  - How can we ensure labelers give consistent labels?

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

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	< 10,000
Big data	Speech recognition from 50 million training examples	Online shopping recommendations for 1 million users	> 10,000

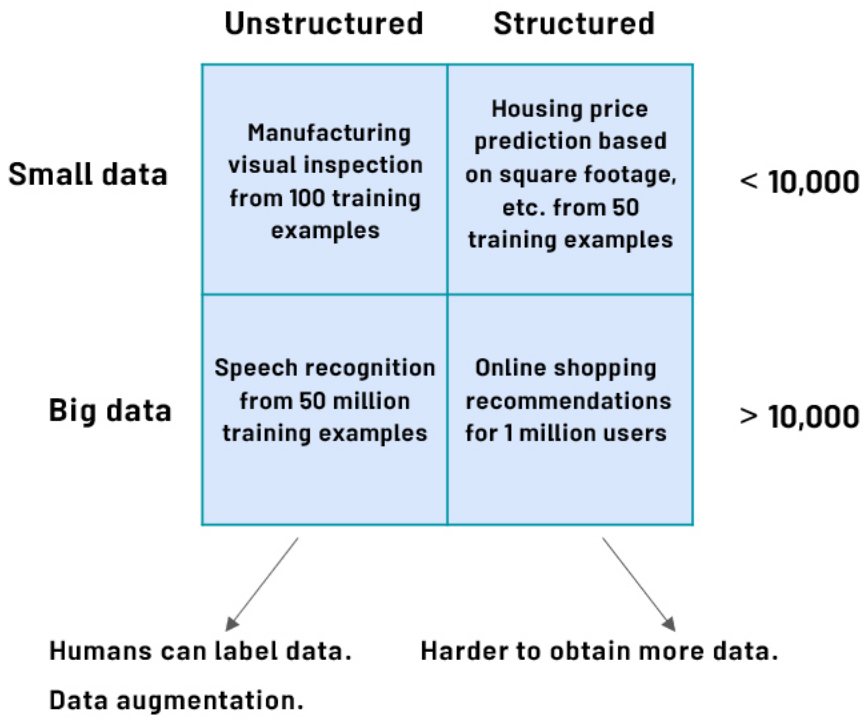
Major types of data problems

	Unstructured	Structured	
Small data	Manufacturing visual inspection from 100 training examples	Housing price prediction based on square footage, etc. from 50 training examples	< 10,000
Big data	Speech recognition from 50 million training examples	Online shopping recommendations for 1 million users	> 10,000

Humans can label data.

Data augmentation.

Major types of data problems



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	< 10,000 Clean labels are critical.
Big data	Speech recognition from 50 million training examples	Online shopping recommendations for 1 million users	> 10,000

Humans can label data.

Data augmentation.

Harder to obtain more data.

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	< 10,000 Clean labels are critical.
Big data	Speech recognition from 50 million training examples	Online shopping recommendations for 1 million users	> 10,000 Emphasis on data process.

Humans can label data.

Data augmentation.

Harder to obtain more data.

Unstructured vs. structured data

- Unstructured data

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.

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**

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.

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.

Major types of data problems

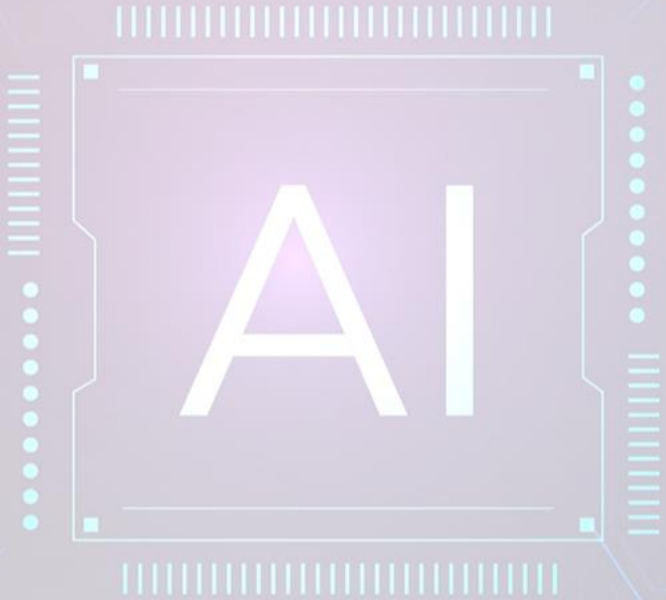
	Unstructured	Structured	
Small data	Manufacturing visual inspection from 100 training examples	Housing price prediction based on square footage, etc. from 50 training examples	< 10,000 Clean labels are critical.
Big data	Speech recognition from 50 million training examples	Online shopping recommendations for 1 million users	> 10,000 Emphasis on data process.

Humans can label data.

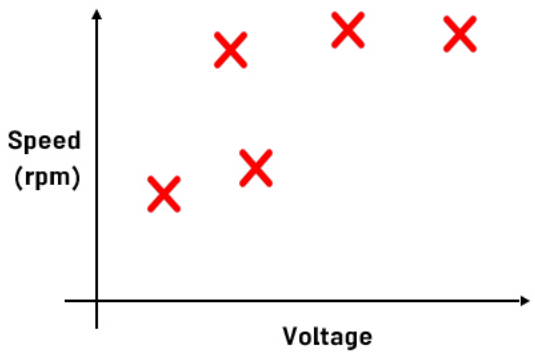
Data augmentation.

Harder to obtain more data.

label consistency

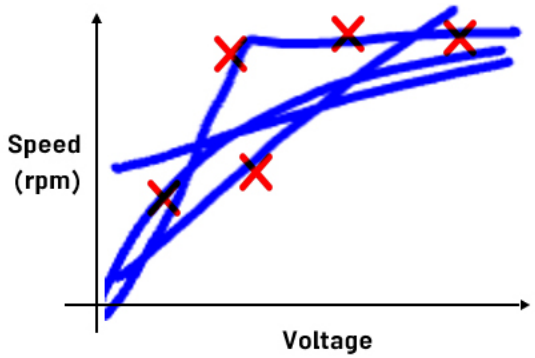


Why label consistency is important



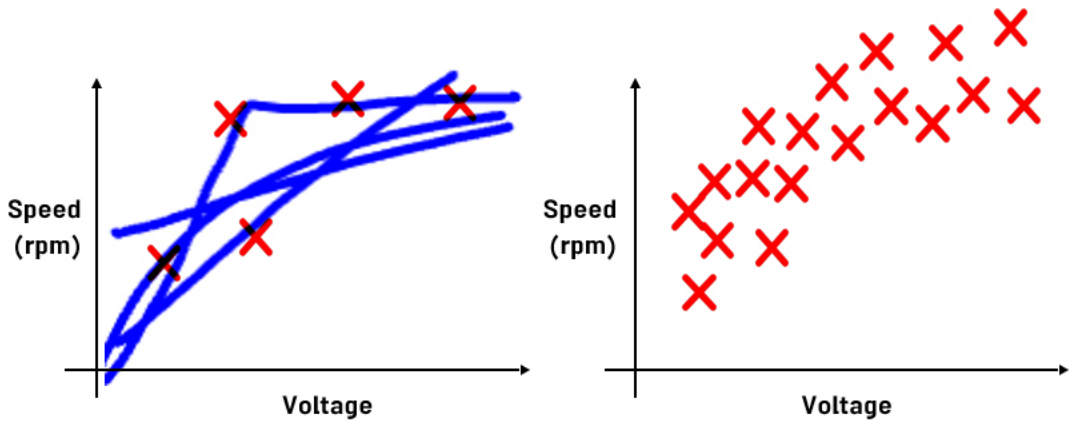
- Small data
- Noisy labels

Why label consistency is important



- Small data
- Noisy labels

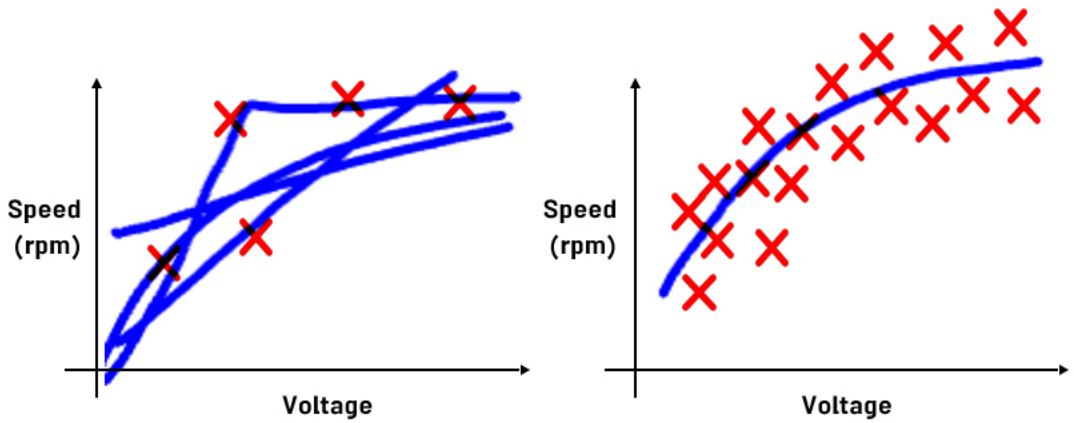
Why label consistency is important



- Small data
- Noisy labels

- Big data
- Noisy labels

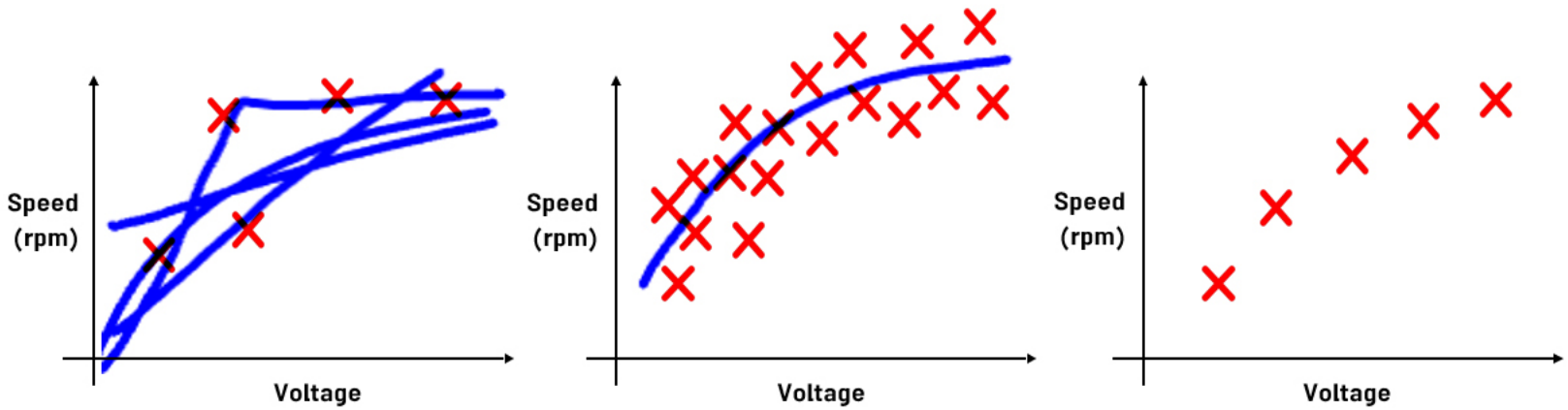
Why label consistency is important



- Small data
- Noisy labels

- Big data
- Noisy labels

Why label consistency is important

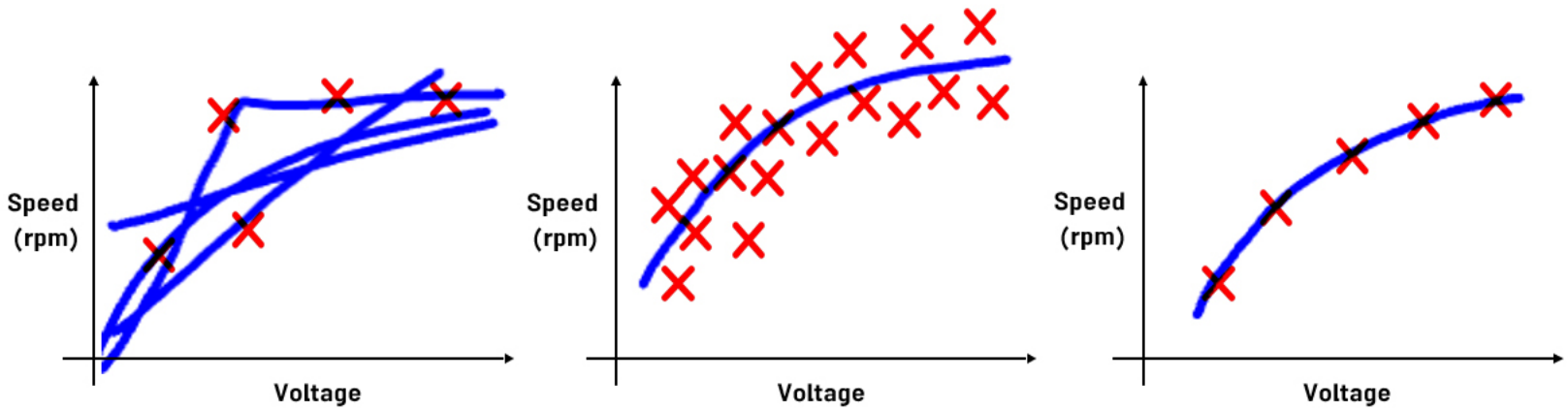


- Small data
- Noisy labels

- Big data
- Noisy labels

- Small data
- Clean (consistent) labels

Why label consistency is important



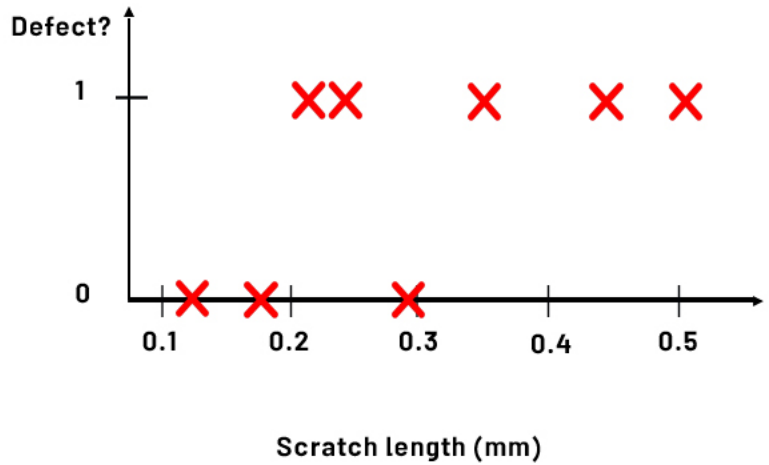
- Small data
- Noisy labels

- Big data
- Noisy labels

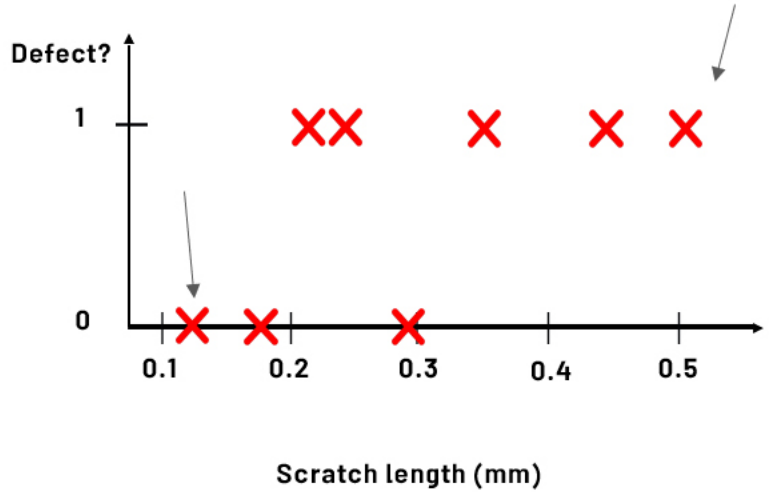
- Small data
- Clean (consistent) labels

Phone defect example

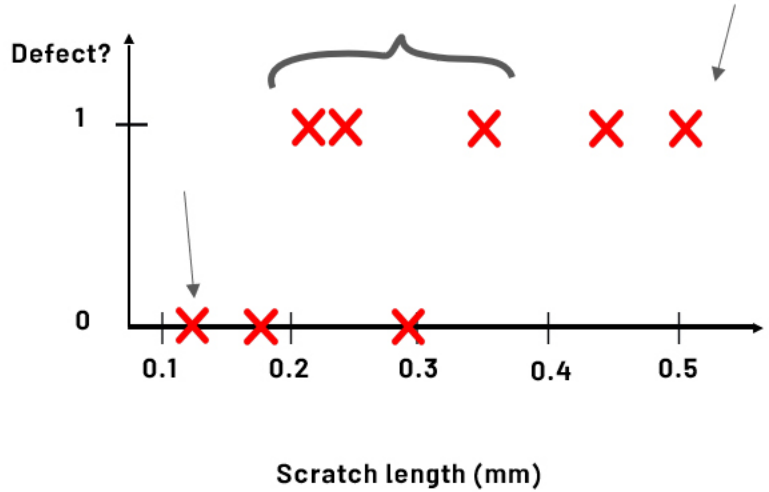
Phone defect example



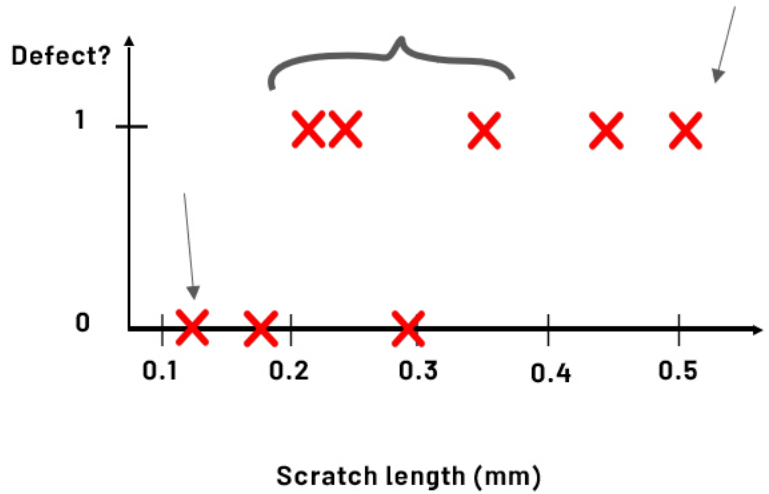
Phone defect example

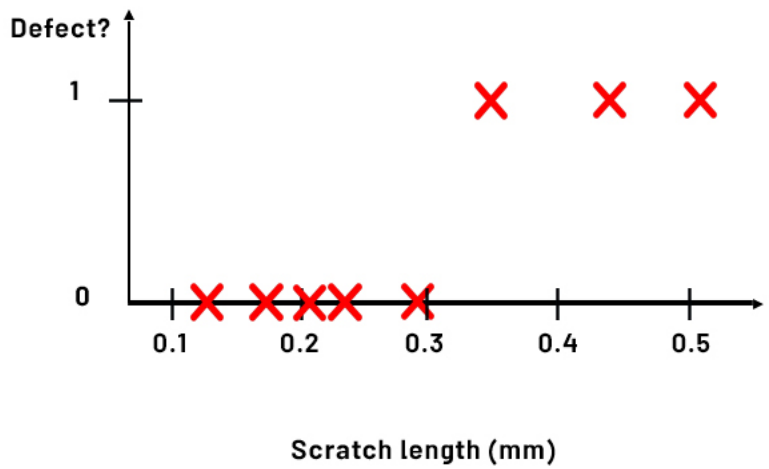


Phone defect example



Phone defect example





Big data problems can have small data challenges too

- Problems with a large dataset but where there's a long tail of rare events in the input will have small data challenges too.
  - Web search
  - Self-driving cars
  - Product recommendation systems

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]"

Examples

- **Standardize labels**
  - “Um, nearest gas station”
  - “Umm, nearest gas station”
  - “Nearest gas station [unintelligible]”

→ "Um, nearest gas station"

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

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

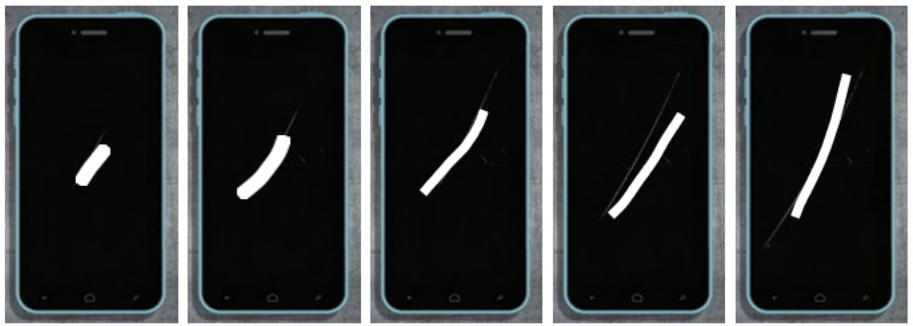
➡ Scratch

Have a class/label to capture uncertainty

- Defect: 0 or 1

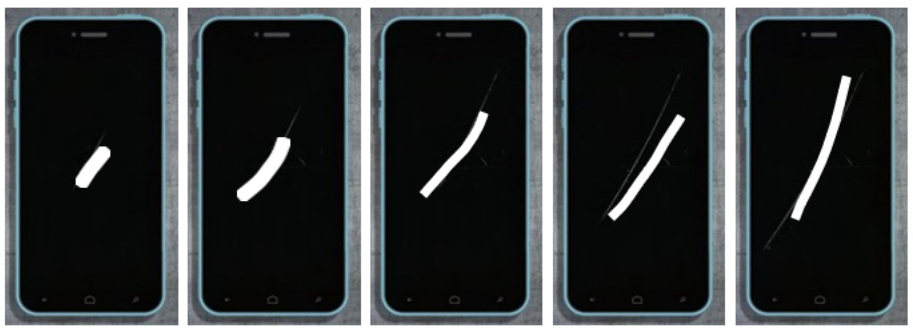
Have a class/label to capture uncertainty

- Defect: 0 or 1



Have a class/label to capture uncertainty

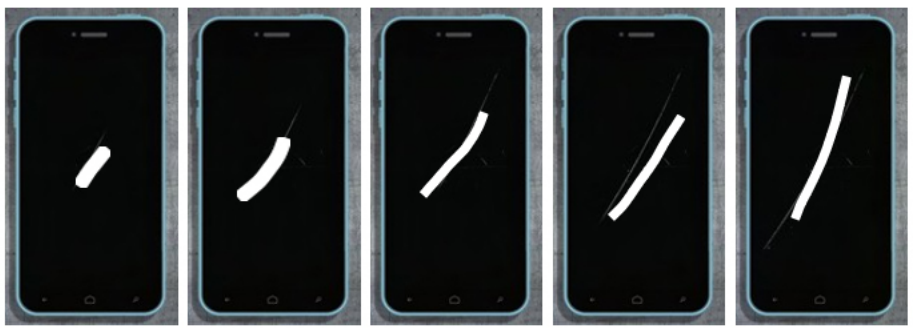
- Defect: 0 or 1



Alternative: 0, Borderline, 1

Have a class/label to capture uncertainty

- Defect: 0 or 1

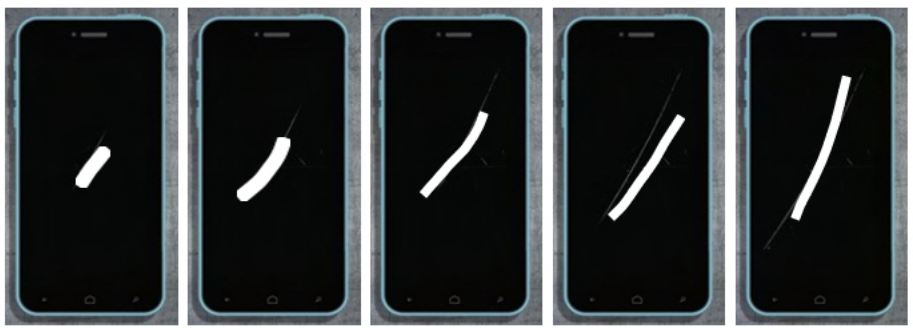


Alternative: 0, Borderline, 1

- Unintelligible audio
  - “nearest go”
  - “nearest grocery”

Have a class/label to capture uncertainty

- **Defect: 0 or 1**



Alternative: 0, Borderline, 1

- **Unintelligible audio**
  - "nearest go"
  - "nearest grocery"

"nearest [unintelligible]"

Small data vs. big data (unstructured data)

Small data

- Usually small number of labelers.
- Can ask labelers to discuss specific labels.

Small data vs. big data (unstructured data)

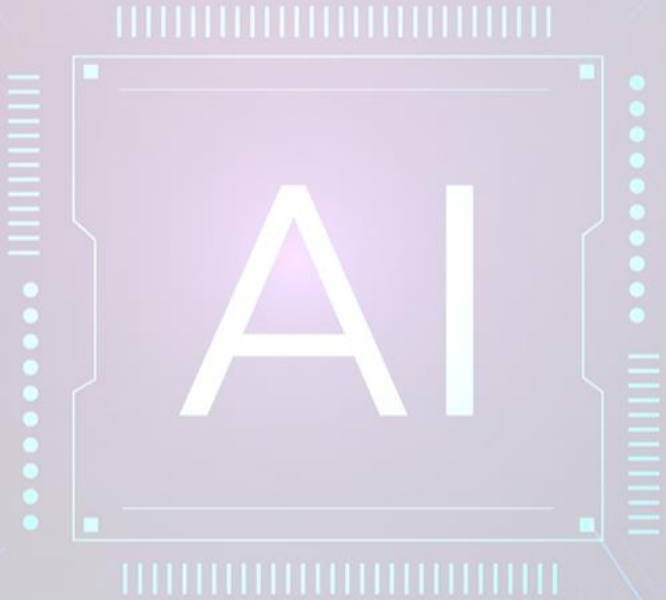
Small data

- Usually small number of labelers.
- Can ask labelers to discuss specific labels.

Big data

- Get to consistent definition with a small group.
- Then send labeling instructions to labelers.
- Can consider having multiple labelers label every example and using voting or consensus labels to increase accuracy.

Human Level Performance



Why measure HLP?

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

Why measure HLP?

- **Estimate Bayes error / irreducible error to help with error analysis and prioritization.**
  - **Let's develop a 99% accurate model!**

Why measure HLP?

- **Estimate Bayes error / irreducible error to help with error analysis and prioritization.**
  - **Let's develop a 99% accurate model!**

Ground Truth	Inspector

Why measure HLP?

- **Estimate Bayes error / irreducible error to help with error analysis and prioritization.**
  - **Let's develop a 99% accurate model!**

Ground Truth	Inspector
1	
1	
1	
0	
0	
0	

Why measure HLP?

- **Estimate Bayes error / irreducible error to help with error analysis and prioritization.**
  - **Let's develop a 99% accurate model!**

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

Why measure HLP?

- **Estimate Bayes error / irreducible error to help with error analysis and prioritization.**
  - Let's develop a 99% accurate model!

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

  - HLP is 66.7% !
  - Irreducible error or inspector agreement?

Other uses of HLP

- In academia, establish and beat a respectable benchmark to support publication.

Other uses of HLP

- In academia, establish and beat a respectable benchmark to support publication.
- Business or product owner asks for 99% accuracy. HLP helps establish a more reasonable target.

Other uses of HLP

- In academia, establish and beat a respectable benchmark to support publication.
- Business or product owner asks for 99% accuracy. HLP helps establish a more reasonable target.
- "Prove" the ML system is superior to humans doing the job and thus the business or product owner should adopt it.

Other uses of HLP

- In academia, establish and beat a respectable benchmark to support publication.
- Business or product owner asks for 99% accuracy. HLP helps establish a more reasonable target.
- "Prove" the ML system is superior to humans doing the job and thus the business or product owner should adopt it.

The problem with beating HLP as a “proof” of ML “superiority”

- "Um... nearest gas station"
- "Um, nearest gas station"

The problem with beating HLP as a "proof" of ML "superiority"

- "Um... nearest gas station"
- "Um, nearest gas station"

70% of labels

The problem with beating HLP as a "proof" of ML "superiority"

- "Um... nearest gas station" — 70% of labels
- "Um, nearest gas station" — 30% of labels

The problem with beating HLP as a "proof" of ML "superiority"

- "Um... nearest gas station" — 70% of labels
- "Um, nearest gas station" — 30% of labels
- Two random labelers agree:
- ML agrees with humans:

The problem with beating HLP as a "proof" of ML "superiority"

- "Um... nearest gas station" — 70% of labels
- "Um, nearest gas station" — 30% of labels
- Two random labelers agree: — 0.7 x 0.7 + 0.3 x 0.3 = 0.58
- ML agrees with humans:

The problem with beating HLP as a "proof" of ML "superiority"

"Um... nearest gas station"	70% of labels
"Um, nearest gas station"	30% of labels
Two random labelers agree:	0.7 x 0.7 + 0.3 x 0.3 = 0.58
ML agrees with humans:	0.7

The problem with beating HLP as a "proof" of ML "superiority"

"Um... nearest gas station"	70% of labels
"Um, nearest gas station"	30% of labels
Two random labelers agree:	0.7 x 0.7 + 0.3 x 0.3 = 0.58
ML agrees with humans:	0.7

- The 12% better performance is not important for anything! This can also mask more significant errors ML may be making.

The problem with beating HLP as a "proof" of ML "superiority"

- "Um... nearest gas station" — 70% of labels
- "Um, nearest gas station" — 30% of labels
- Two random labelers agree: — 0.7 x 0.7 + 0.3 x 0.3 = 0.58
- ML agrees with humans: — 0.7
- The 12% better performance is not important for anything! This can also mask more significant errors ML may be making.

Try to raise HLP !

Raising HLP

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

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.

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	Inspector
1	1
1	0
1	1
0	0
0	0
0	1

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	Inspector
1	1
1	0
1	1
0	0
0	0
0	1

66.7%

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.

Scratch Length	Ground Truth	Inspector
0.7	1	1
0.2	1	0
0.5	1	1
0.2	0	0
0.1	0	0
0.1	0	1

66.7%

Raising HLP

- When the label y comes from a human label, HLP << 100% may indicate ambiguous labeling instructions.
- Improving label consistency will raise HLP.
- This makes it harder for ML to beat HLP. But the more consistent labels will raise ML performance, which is ultimately likely to benefit the actual application performance.

HLP on structured data

- Structured data problems are less likely to involve human labelers, thus HLP is less frequently used.

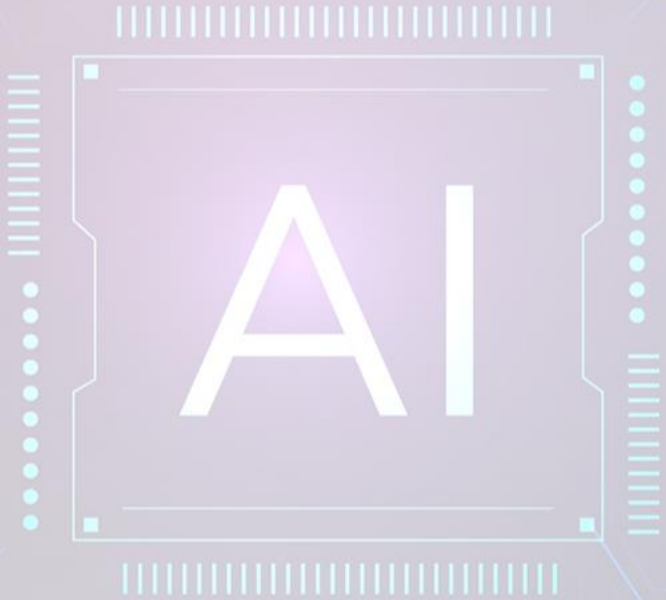
HLP on structured data

- Structured data problems are less likely to involve human labelers, thus HLP is less frequently used.
- Some exceptions:
  - User ID merging: Same person?
  - Based on network traffic, is the computer hacked?
  - Is the transaction fraudulent?
  - Spam account? Bot?
  - From GPS, what is the mode of transportation – on foot, bike, car, bus?

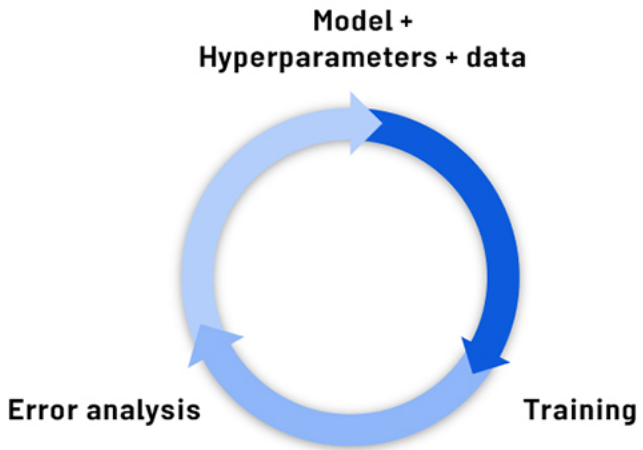
Key point on HLP

- HLP is important
- What is possible?
- HLP could be affected by labeling inconsistency
  - Cleaner data
- Improve labeling consistency

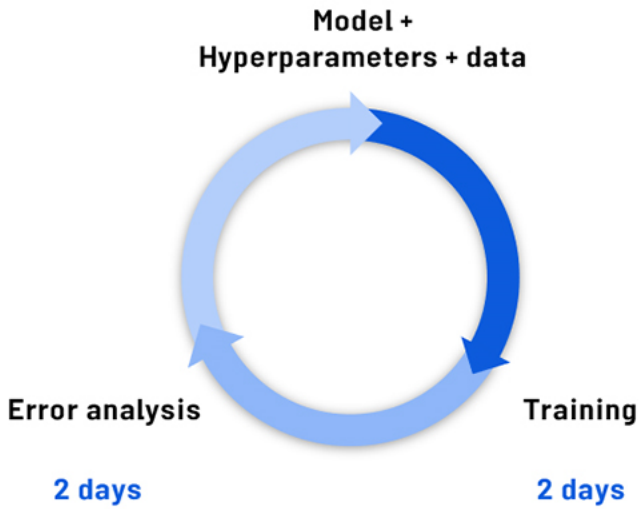
Obtaining data



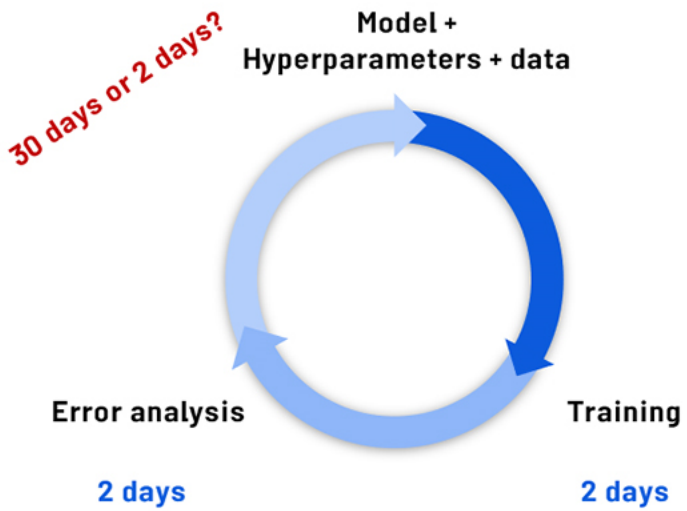
How long should you spend obtaining data?



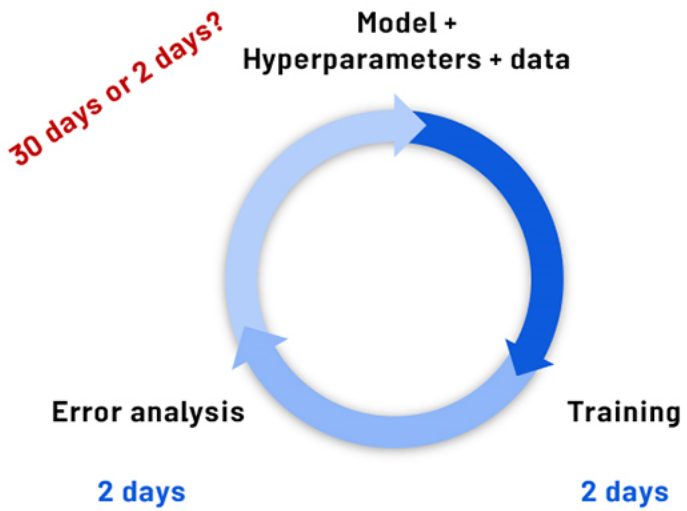
How long should you spend obtaining data?



How long should you spend obtaining data?

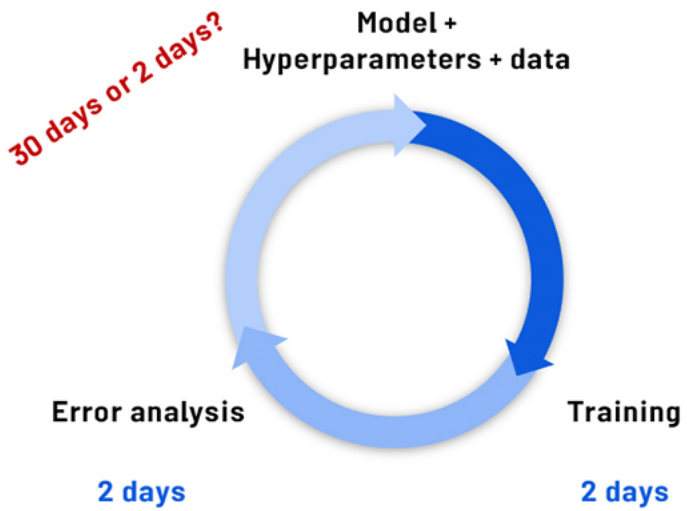


How long should you spend obtaining data?



- Get into this iteration loop as quickly possible.

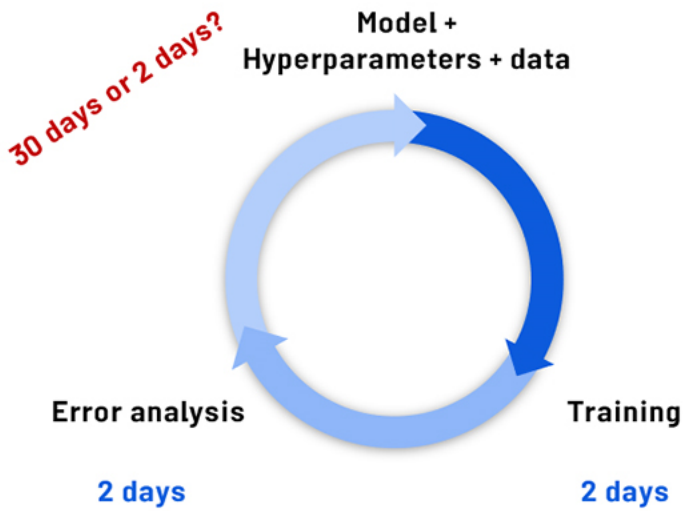
How long should you spend obtaining data?



- 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.

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)

Inventory data

- Brainstorm list of data sources ( speech recognition)

Source	Amount	Cost

Inventory data

- Brainstorm list of data sources ( speech recognition)

Source	Amount	Cost
Owned		
Crowdsourced - Reading		
Pay for labels		
Purchase data		

Inventory data

- Brainstorm list of data sources ( speech recognition)

Source	Amount	Cost
Owned	100h	
Crowdsourced - Reading	1000h	
Pay for labels	100h	
Purchase data	1000h	

Inventory data

- Brainstorm list of data sources ( speech recognition)

Source	Amount	Cost
Owned	100h	0
Crowdsourced - Reading	1000h	60$
Pay for labels	100h	30$
Purchase data	1000h	50$

Inventory data

- Brainstorm list of data sources ( speech recognition)

Source	Amount	Cost
Owned	100h	0
Crowdsourced - Reading	1000h	60
Pay for labels	100h	30
Purchase data	1000h	50

Other factors: Data quality, privacy, regulatory constraints

Inventory data

- Brainstorm list of data sources ( speech recognition)

Source	Amount	Cost	Time
Owned	100h	0	0
Crowdsourced - Reading	1000h	60	14 days
Pay for labels	100h	30	8 days
Purchase data	1000h	50	1 day

Other factors: Data quality, privacy, regulatory constraints

Labeling data

- Options: In-house vs. outsourced vs. crowdsourced
- Having MLEs label data is expensive. But doing this for just a few days is usually fine.
- Who is qualified to label?
  - Speech recognition – any reasonably fluent speaker
  - Factory inspection, medical image diagnosis – SME (subject matter expert)
  - Recommender systems – maybe impossible to label well

- Don't increase data by more than 10x at a time

Data pipeline

	Job Board (website)	Resume chat (app)
Email	ramin@avir.co.com	rtoosi@mci.ir
First Name	Ramin	Ramin
Last Name	Toosi	Tsi
Address	Vanak	?
State	Tehran	?
Zip	9393	9393

Data pipeline

	Job Board (website)	Resume chat (app)
Email	ramin@avir.co.com	rtoosi@mci.ir
First Name	Ramin	Ramin
Last Name	Toosi	Tsi
Address	Vanak	?
State	Tehran	?
Zip	9393	9393

Raw data

Data pipeline

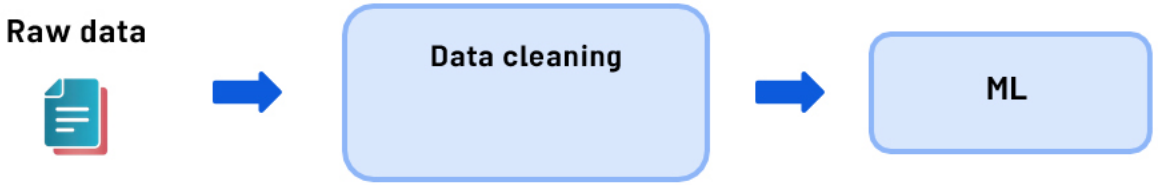
	Job Board (website)	Resume chat (app)
Email	ramin@avir.co.com	rtoosi@mci.ir
First Name	Ramin	Ramin
Last Name	Toosi	Tsi
Address	Vanak	?
State	Tehran	?
Zip	9393	9393

Raw data

Data cleaning

Data pipeline

	Job Board (website)	Resume chat (app)
Email	ramin@avir.co.com	rtoosi@mci.ir
First Name	Ramin	Ramin
Last Name	Toosi	Tsi
Address	Vanak	?
State	Tehran	?
Zip	9393	9393



Data pipeline

	Job Board (website)	Resume chat (app)
Email	ramin@avir.co.com	rtoosi@mci.ir
First Name	Ramin	Ramin
Last Name	Toosi	Tsi
Address	Vanak	?
State	Tehran	?
Zip	9393	9393

Data pipeline

Development

Data

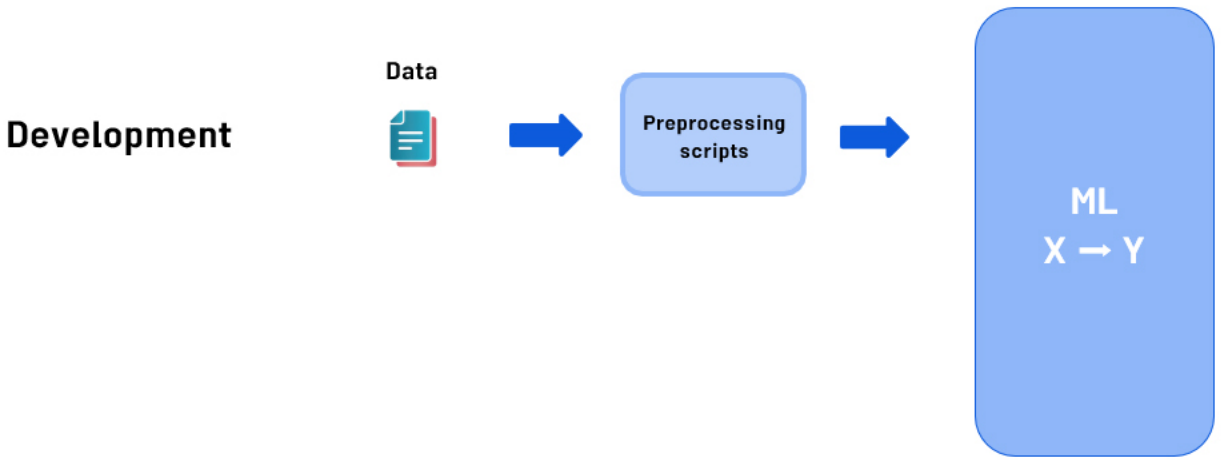
Data pipeline

Development

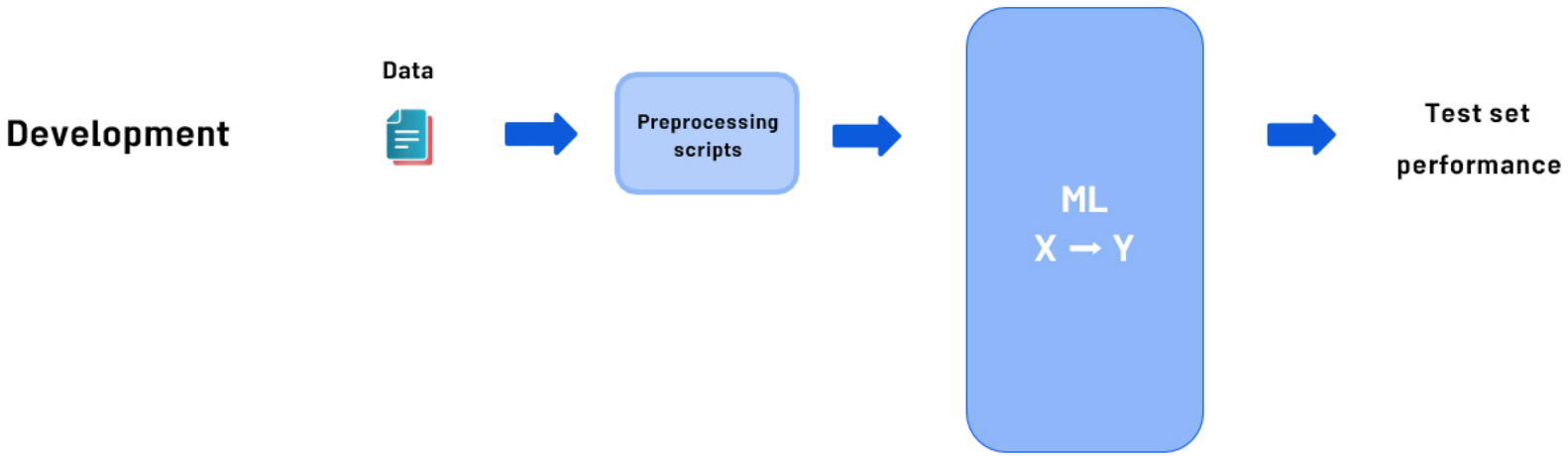
Data

Preprocessing scripts

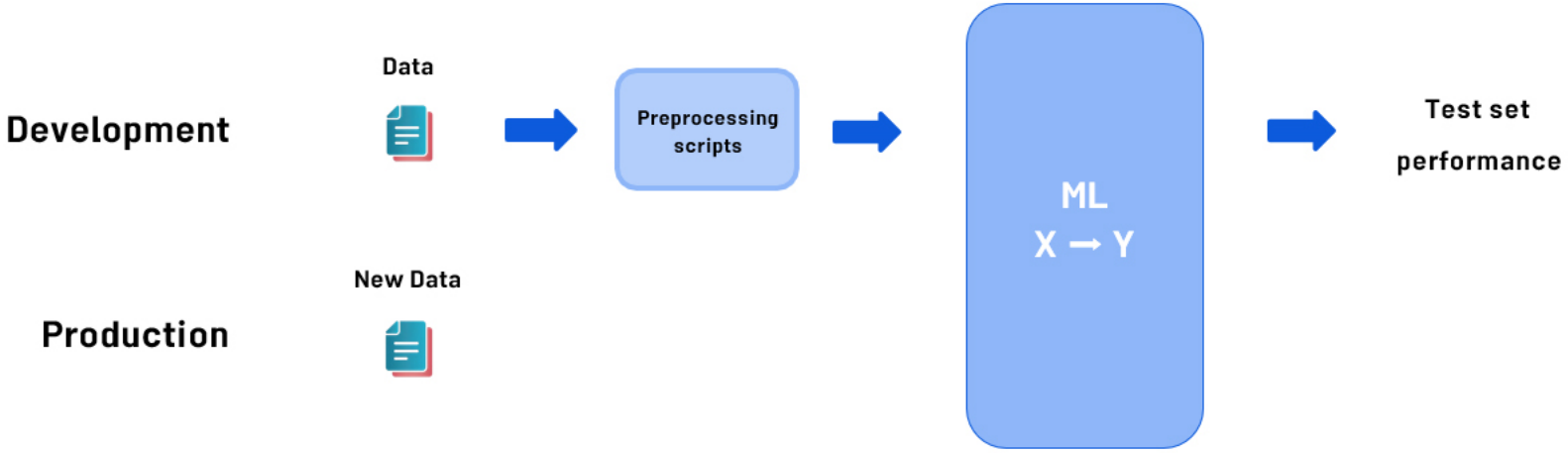
Data pipeline



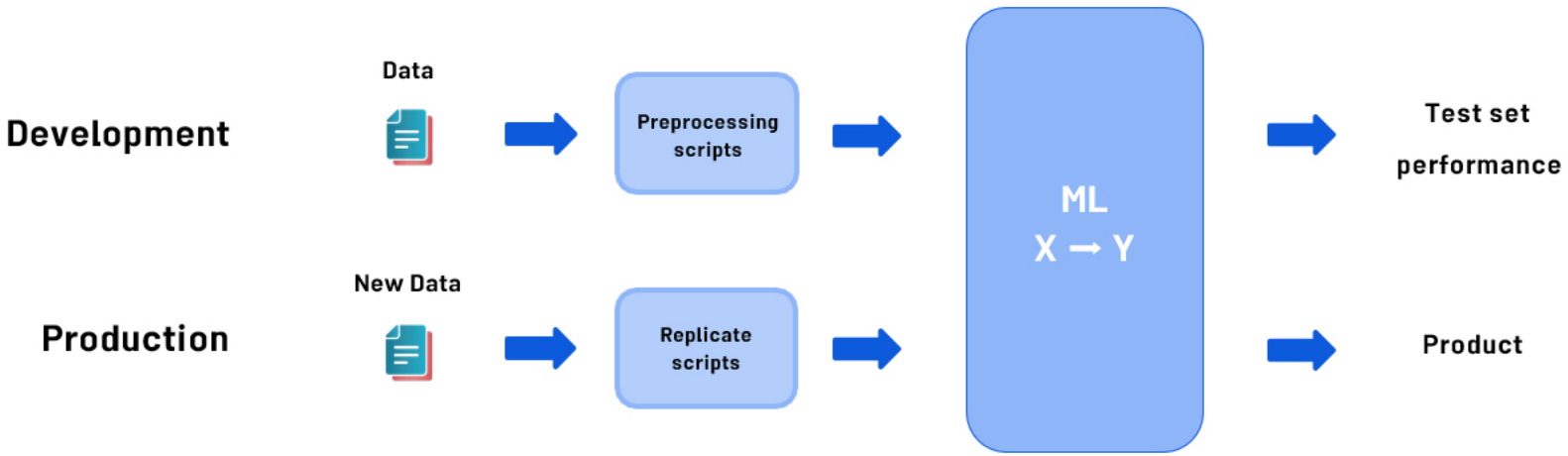
Data pipeline



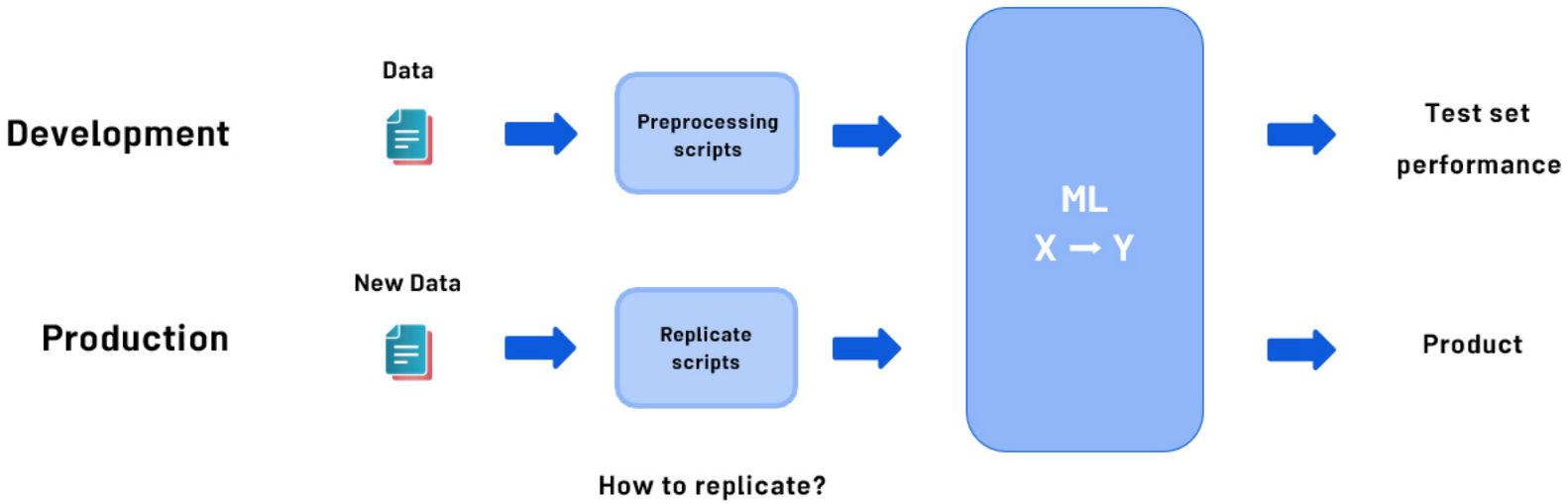
Data pipeline



Data pipeline



Data pipeline



POC and Production phases

- **POC (proof-of-concept):**
  - Goal is to decide if the application is workable and worth deploying.
  - Focus on getting the prototype to work!
  - It's ok if data pre-processing is manual. But take extensive notes/comments.

POC and Production phases

- **POC (proof-of-concept):**
  - Goal is to decide if the application is workable and worth deploying.
  - Focus on getting the prototype to work!
  - It's ok if data pre-processing is manual. But take extensive notes/comments.
- **Production phase:**
  - After project utility is established, use more sophisticated tools to make sure the data pipeline is replicable.
  - E.g., TensorFlow Transform, Apache Beam, Airflow,...

A more complex data pipeline example

- Task: Predict if someone is looking for a job.

  (x = user data, y = looking for a job?)

A more complex data pipeline example

- Task: Predict if someone is looking for a job.

  (x = user data, y = looking for a job?)

Spam dataset

A more complex data pipeline example

- Task: Predict if someone is looking for a job.

  (x = user data, y = looking for a job?)

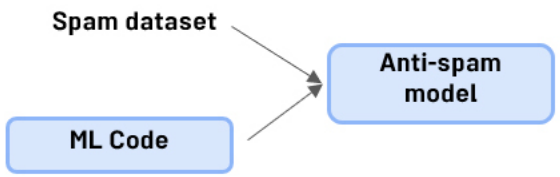
Spam dataset

ML Code

A more complex data pipeline example

- Task: Predict if someone is looking for a job.

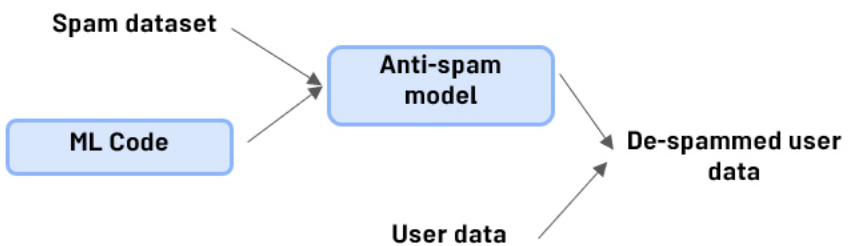
  (x = user data, y = looking for a job?)



A more complex data pipeline example

- Task: Predict if someone is looking for a job.

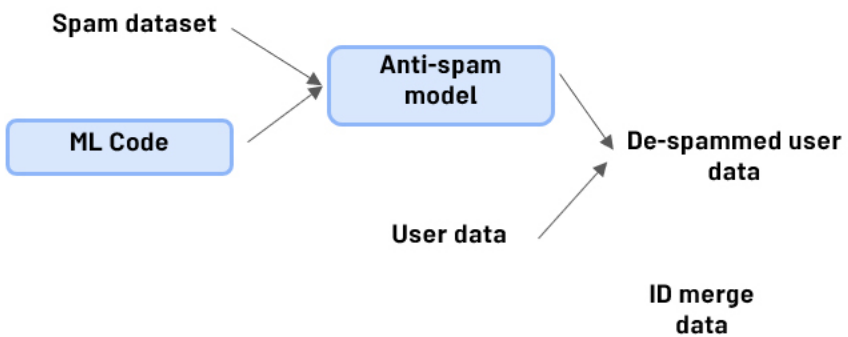
  (x = user data, y = looking for a job?)



A more complex data pipeline example

- Task: Predict if someone is looking for a job.

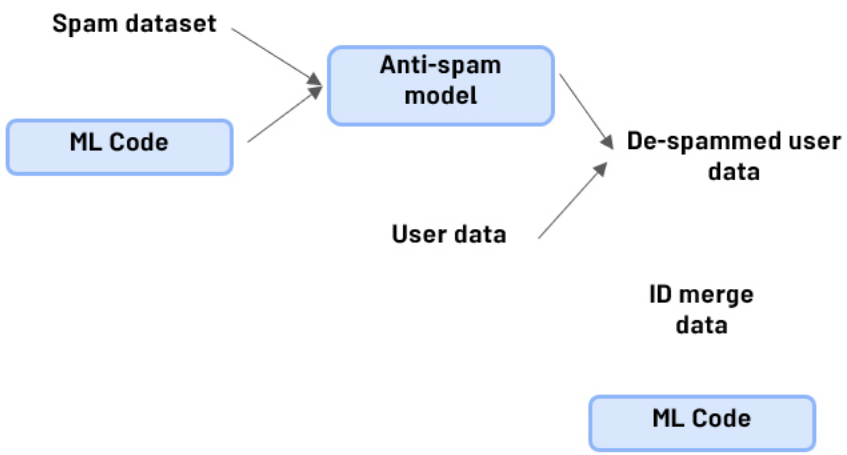
  (x = user data, y = looking for a job?)



A more complex data pipeline example

- Task: Predict if someone is looking for a job.

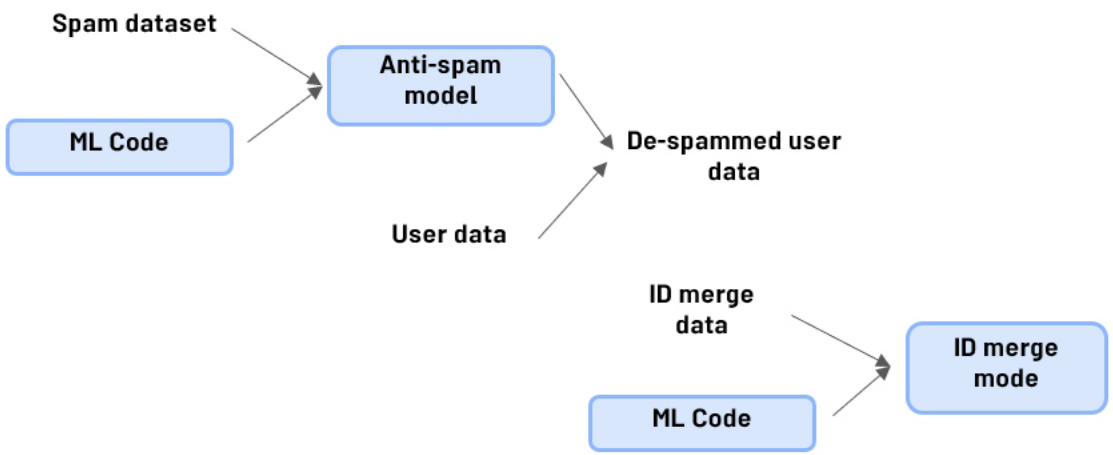
  (x = user data, y = looking for a job?)



A more complex data pipeline example

- Task: Predict if someone is looking for a job.

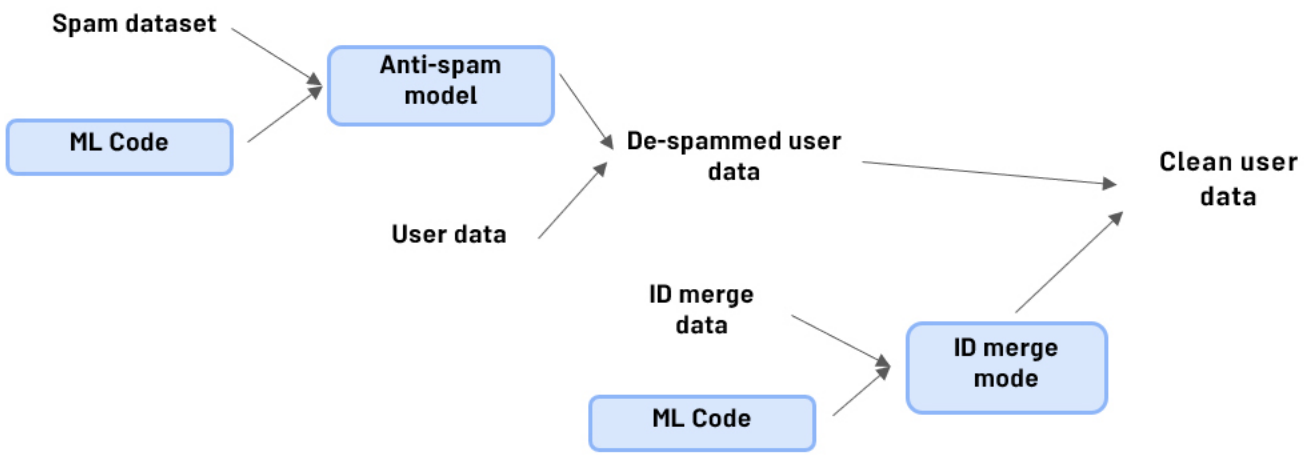
  (x = user data, y = looking for a job?)



A more complex data pipeline example

- Task: Predict if someone is looking for a job.

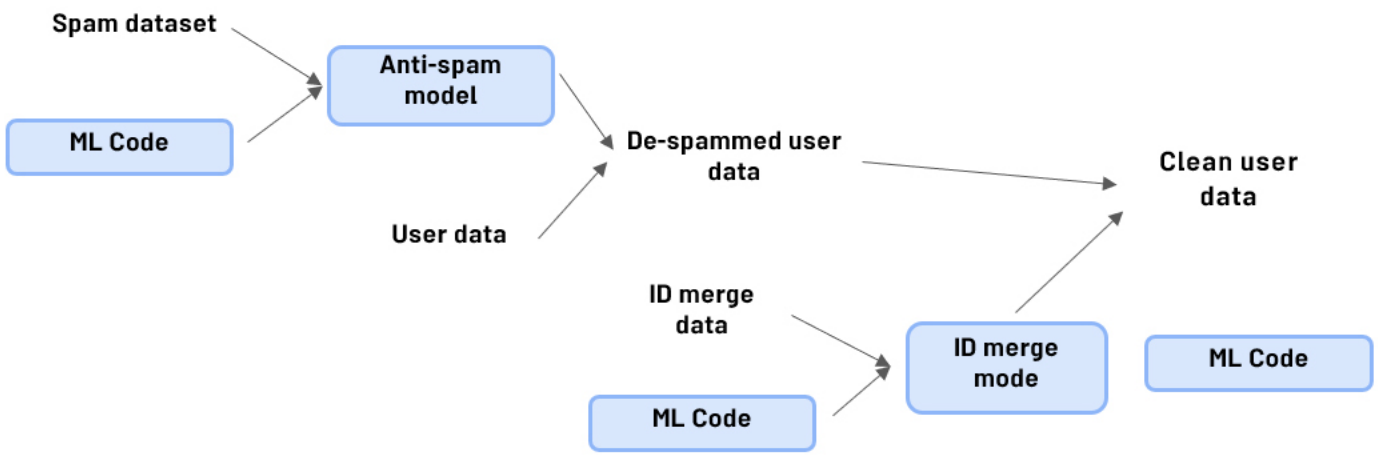
  (x = user data, y = looking for a job?)



A more complex data pipeline example

- Task: Predict if someone is looking for a job.

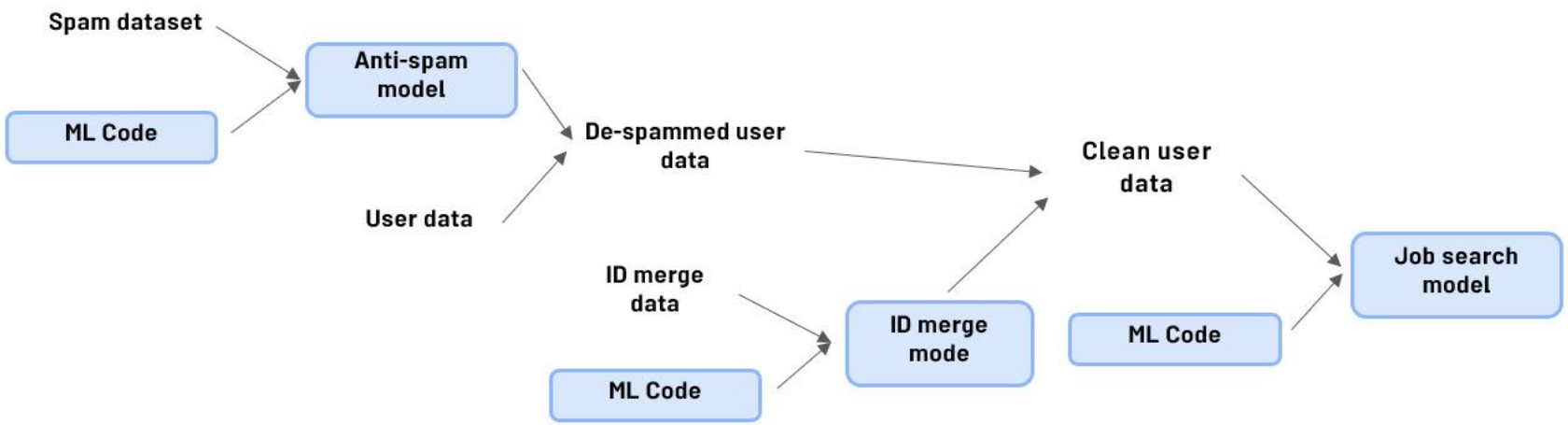
  (x = user data, y = looking for a job?)



A more complex data pipeline example

- **Task: Predict if someone is looking for a job.**

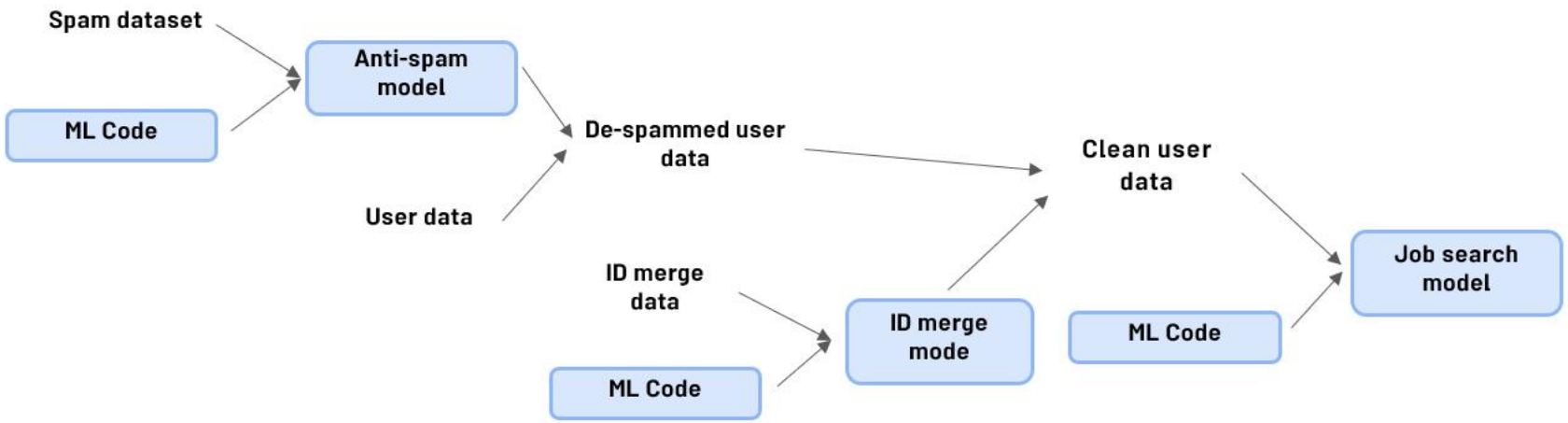
  (x = user data, y = looking for a job?)



A more complex data pipeline example

- Task: Predict if someone is looking for a job.

  (x = user data, y = looking for a job?)



Keep track of data provenance and lineage

where it comes from sequence of steps

Meta-data

- **Examples:**
  - Manufacturing visual inspection: Time, factory, line # camera settings, phone model, inspector ID,....
  - Speech recognition: Device type, labeler ID, VAD model ID,....

Meta-data

Examples:

- Manufacturing visual inspection: Time, factory, line # camera settings, phone model, inspector ID,....
- Speech recognition: Device type, labeler ID, VAD model ID,....

Useful for:

- Error analysis. Spotting unexpected effects.
- Keeping track of data provenance.

Balanced train/dev/test splits in small data problems

- Visual inspection example: 100 examples, 30 positive (defective)
  - Train/dev/test:
  - Random split:
  - Want:
- No need to worry about this with large datasets – a random split will be representative.

Balanced train/dev/test splits in small data problems

- Visual inspection example: 100 examples, 30 positive (defective)
  - Train/dev/test: 60% / 20% / 20%
  - Random split:
  - Want:
- No need to worry about this with large datasets – a random split will be representative.

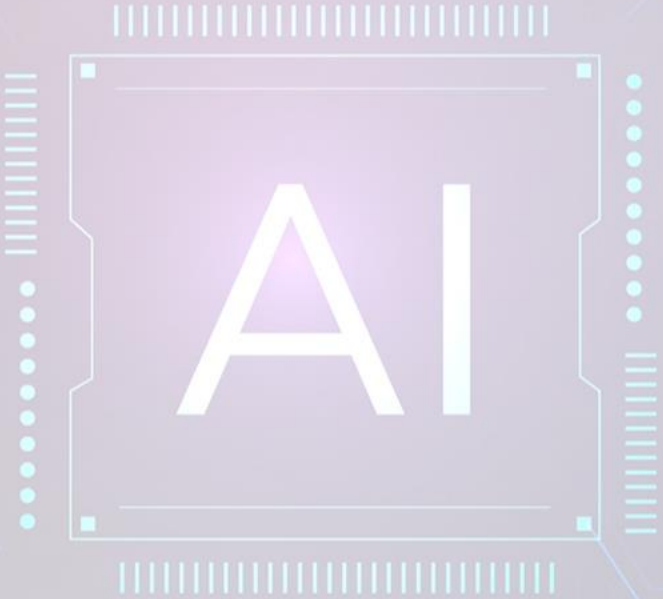
Balanced train/dev/test splits in small data problems

- Visual inspection example: 100 examples, 30 positive (defective)
  - Train/dev/test: 60% / 20% / 20%
  - Random split: 21 / 2 / 7 35% / 10% / 35% Positive examples
  - Want:
- No need to worry about this with large datasets – a random split will be representative.

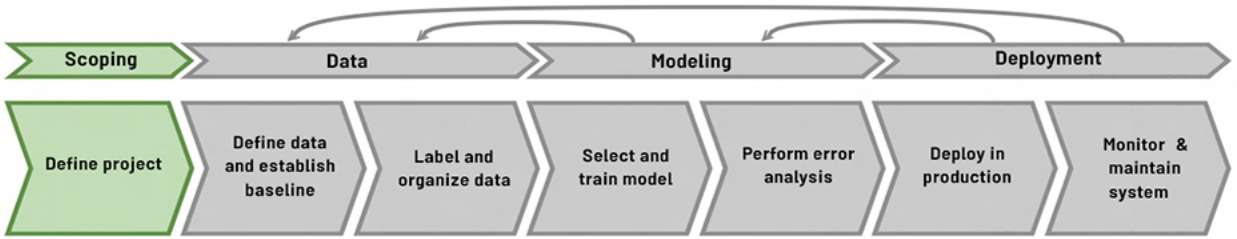
Balanced train/dev/test splits in small data problems

- Visual inspection example: 100 examples, 30 positive (defective)
  - Train/dev/test: 60% / 20% / 20%
  - Random split: 21 / 2 / 7 35% / 10% / 35% Positive examples
  - Want: 18 / 6 / 6 30% / 30% / 30%
- No need to worry about this with large datasets – a random split will be representative.

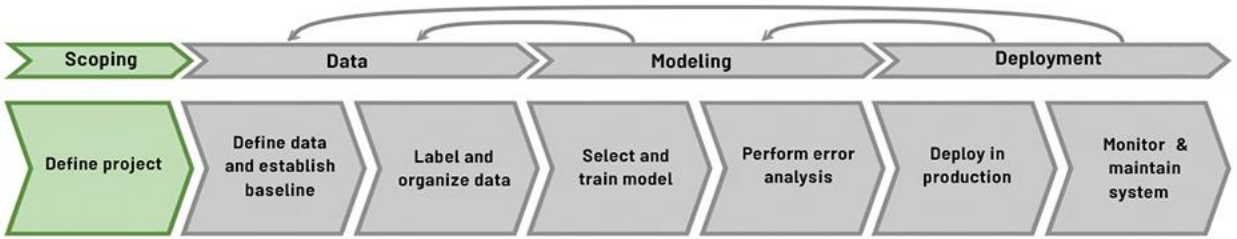
Scoping



Scoping example

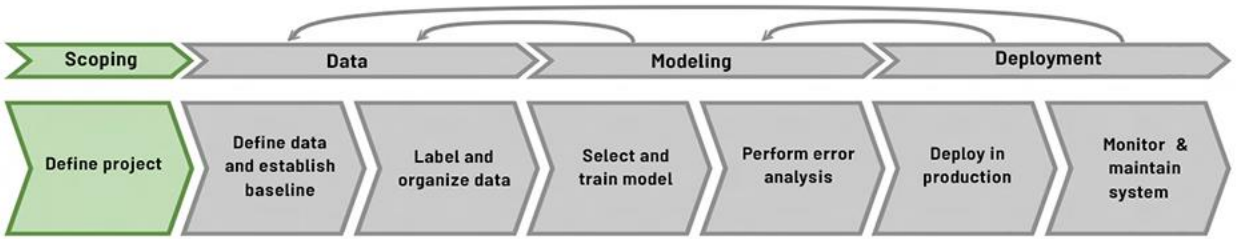


Scoping example



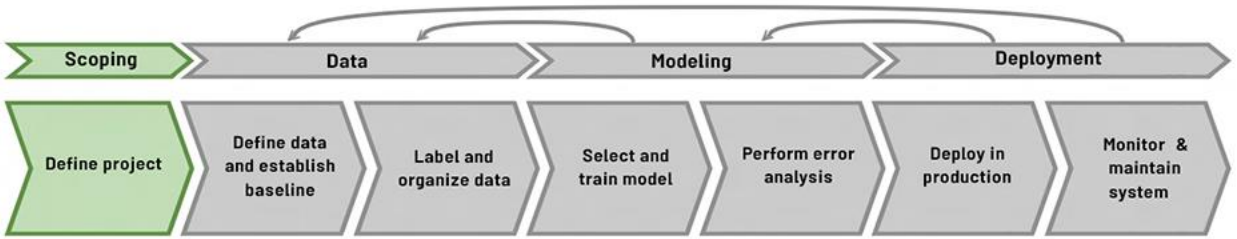
- Ecommerce retailer looking to increase sales

Scoping example



- Ecommerce retailer looking to increase sales
  - Better recommender system
  - Better search
  - Improve catalog data
  - Inventory management
  - Price optimization

Scoping example



- **Ecommerce retailer looking to increase sales**
  - Better recommender system
  - Better search
  - Improve catalog data
  - Inventory management
  - Price optimization

- **Questions**
  - What projects should we work on?
  - What are the metrics for success?
  - What are the resources (data, time,people) needed?

Scoping process

Identify a business problem
(not an AI problem)

Scoping process

Identify a business problem (not an AI problem)

- **What are the top 3 things you wish were working better?**
  - Increase conversion
  - Reduce inventory
  - Increase margin (profit per item)

Scoping process

Identify a business problem
(not an AI problem)

⬇

Brainstorm AI solutions

- **What are the top 3 things you wish were working better?**
  - Increase conversion
  - Reduce inventory
  - Increase margin (profit per item)

Scoping process

Identify a business problem (not an AI problem)

⬇

Brainstorm AI solutions

⬇

Assess the feasibility and value of potential solutions

- **What are the top 3 things you wish were working better?**
  - Increase conversion
  - Reduce inventory
  - Increase margin (profit per item)

Scoping process

- Identify a business problem (not an AI problem)
- Brainstorm AI solutions
- Assess the feasibility and value of potential solutions
- Determine milestones

- **What are the top 3 things you wish were working better?**
  - Increase conversion
  - Reduce inventory
  - Increase margin (profit per item)

Scoping process

- Identify a business problem (not an AI problem)
- Brainstorm AI solutions
- Assess the feasibility and value of potential solutions
- Determine milestones
- Budget for resources

- **What are the top 3 things you wish were working better?**
  - Increase conversion
  - Reduce inventory
  - Increase margin (profit per item)

Separating problem identification from solution

Problem	Solution

Separating problem identification from solution

Problem	Solution
Increase conversion	Search, recommendations

Separating problem identification from solution

Problem	Solution
Increase conversion	Search, recommendations
Reduce inventory	

Separating problem identification from solution

Problem	Solution
Increase conversion	Search, recommendations
Reduce inventory	Demand prediction, marketing

Separating problem identification from solution

Problem	Solution
Increase conversion	Search, recommendations
Reduce inventory	Demand prediction, marketing
Increase margin (profit per item)	

Separating problem identification from solution

Problem	Solution
Increase conversion	Search, recommendations
Reduce inventory	Demand prediction, marketing
Increase margin (profit per item)	Optimizing what to sell (e.g., merchandising), recommend bundles

What to achieve

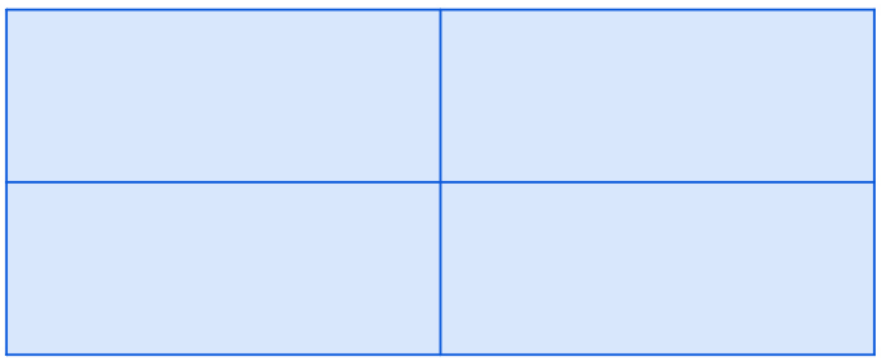
How to achieve

Feasibility: Is this project technically feasible?

- Use external benchmark (literature, other company, competitor)

Feasibility: Is this project technically feasible?

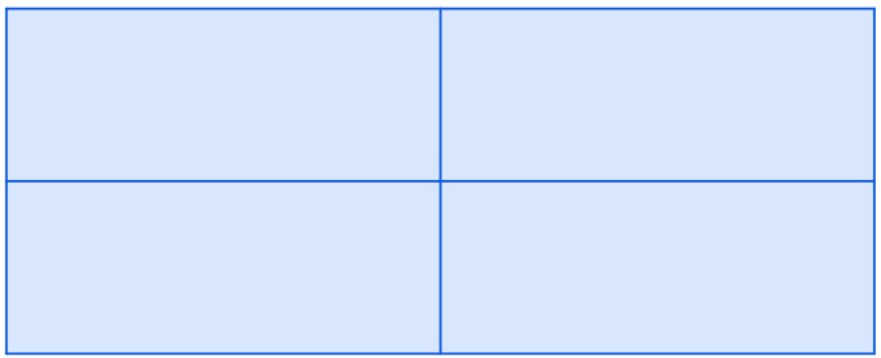
- Use external benchmark (literature, other company, competitor)



Feasibility: Is this project technically feasible?

- Use external benchmark (literature, other company, competitor)

Unstructured
(e.g., speech, images)



Feasibility: Is this project technically feasible?

- Use external benchmark (literature, other company, competitor)

Unstructured (e.g., speech, images)	Structured (e.g., transactions, records)

Feasibility: Is this project technically feasible?

- Use external benchmark (literature, other company, competitor)

	Unstructured (e.g., speech, images)	Structured (e.g., transactions, records)
New		

Feasibility: Is this project technically feasible?

- Use external benchmark (literature, other company, competitor)

	Unstructured (e.g., speech, images)	Structured (e.g., transactions, records)
New		
Existing		

Feasibility: Is this project technically feasible?

- Use external benchmark (literature, other company, competitor)

	Unstructured (e.g., speech, images)	Structured (e.g., transactions, records)
New	HLP	
Existing		

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

Feasibility: Is this project technically feasible?

- Use external benchmark (literature, other company, competitor)

	Unstructured (e.g., speech, images)	Structured (e.g., transactions, records)
New	HLP	Are predictive feature available?
Existing		

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

Feasibility: Is this project technically feasible?

- Use external benchmark (literature, other company, competitor)

	Unstructured (e.g., speech, images)	Structured (e.g., transactions, records)
New	HLP	Are predictive feature available?
Existing	HLP History of project	

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

Feasibility: Is this project technically feasible?

- Use external benchmark (literature, other company, competitor)

	Unstructured (e.g., speech, images)	Structured (e.g., transactions, records)
New	HLP	Are predictive feature available?
Existing	HLP History of project	New predictive features? History of project

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

Why use HLP to benchmark?

- People are very good on unstructured data tasks
- Criteria: Can a human, given the same data, perform the task?

Why use HLP to benchmark?

- People are very good on unstructured data tasks
- Criteria: Can a human, given the same data, perform the task?

Why use HLP to benchmark?

- People are very good on unstructured data tasks
- Criteria: Can a human, given the same data, perform the task?

Do we have features that are predictive?

- Given past purchases, predict future purchases
- Given weather, predict shopping mall foot traffic
- Given DNA info, predict heart disease
- Given social media chatter, predict demand for a clothing style
- Given stock price history, predict future price

Do we have features that are predictive?

-  Given past purchases, predict future purchases
- Given weather, predict shopping mall foot traffic
- Given DNA info, predict heart disease
- Given social media chatter, predict demand for a clothing style
- Given stock price history, predict future price

Do we have features that are predictive?

-  Given past purchases, predict future purchases ✓
- Given weather, predict shopping mall foot traffic
- Given DNA info, predict heart disease
- Given social media chatter, predict demand for a clothing style
- Given stock price history, predict future price

Do we have features that are predictive?

-  Given past purchases, predict future purchases 
-  Given weather, predict shopping mall foot traffic
- Given DNA info, predict heart disease
- Given social media chatter, predict demand for a clothing style
- Given stock price history, predict future price

Do we have features that are predictive?

-  Given past purchases, predict future purchases 
-  Given weather, predict shopping mall foot traffic 
- Given DNA info, predict heart disease
- Given social media chatter, predict demand for a clothing style
- Given stock price history, predict future price

Do we have features that are predictive?

-  Given past purchases, predict future purchases ✓
-  Given weather, predict shopping mall foot traffic ✓
-  Given DNA info, predict heart disease
- Given social media chatter, predict demand for a clothing style
- Given stock price history, predict future price

Do we have features that are predictive?

-  Given past purchases, predict future purchases ✓
-  Given weather, predict shopping mall foot traffic ✓
-  Given DNA info, predict heart disease ?
- Given social media chatter, predict demand for a clothing style
- Given stock price history, predict future price

Do we have features that are predictive?

- Given past purchases, predict future purchases ✓
- Given weather, predict shopping mall foot traffic ✓
- Given DNA info, predict heart disease ?
- Given social media chatter, predict demand for a clothing style
- Given stock price history, predict future price

Do we have features that are predictive?

-  Given past purchases, predict future purchases ✓
-  Given weather, predict shopping mall foot traffic ✓
-  Given DNA info, predict heart disease ?
-  Given social media chatter, predict demand for a clothing style ?
- Given stock price history, predict future price

Do we have features that are predictive?

- Given past purchases, predict future purchases ✓
- Given weather, predict shopping mall foot traffic ✓
- Given DNA info, predict heart disease ?
- Given social media chatter, predict demand for a clothing style ?
- Given stock price history, predict future price

Do we have features that are predictive?

- Given past purchases, predict future purchases ✓
- Given weather, predict shopping mall foot traffic ✓
- Given DNA info, predict heart disease ?
- Given social media chatter, predict demand for a clothing style ?
- Given stock price history, predict future price ✗

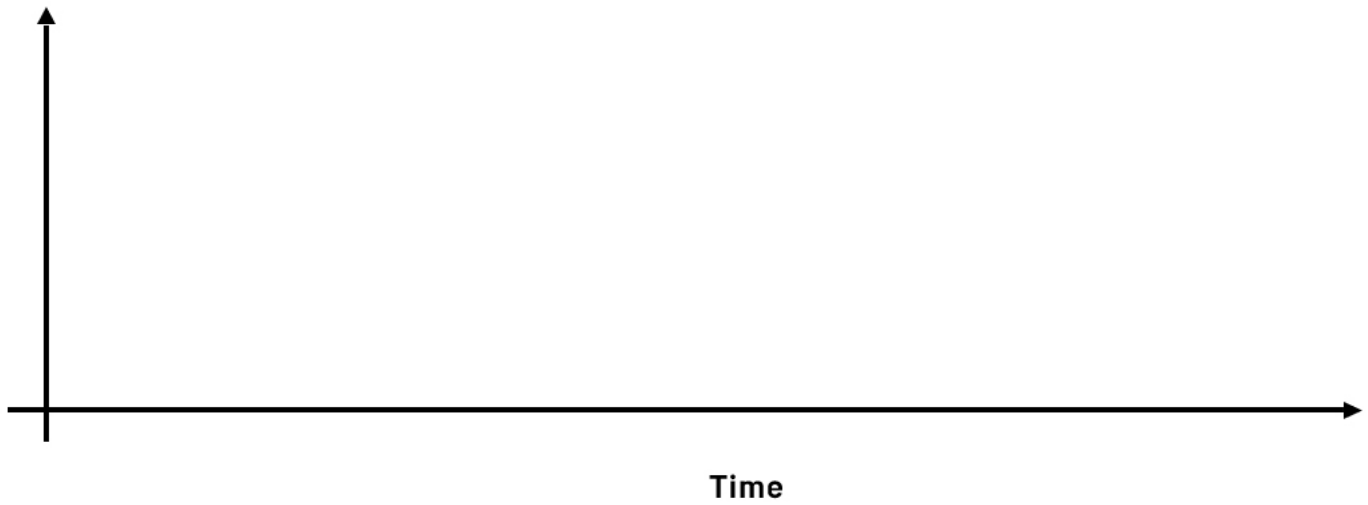
Feasibility: Is this project technically feasible?

- Use external benchmark (literature, other company, competitor)

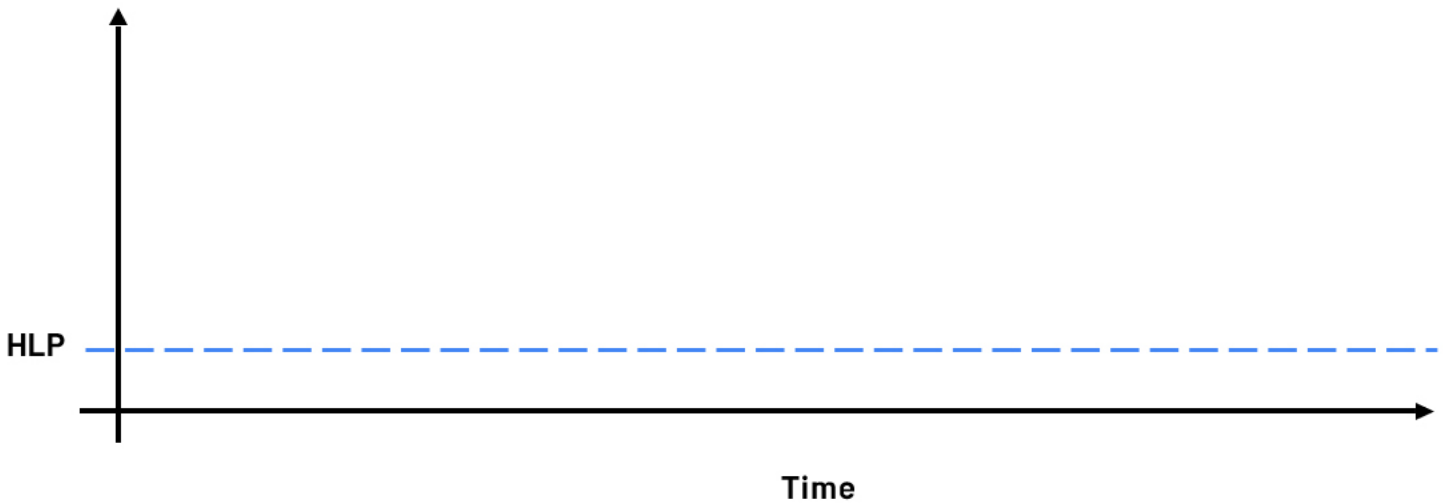
	Unstructured (e.g., speech, images)	Structured (e.g., transactions, records)
New	HLP	Are predictive feature available?
Existing	HLP History of project	New predictive features? History of project

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

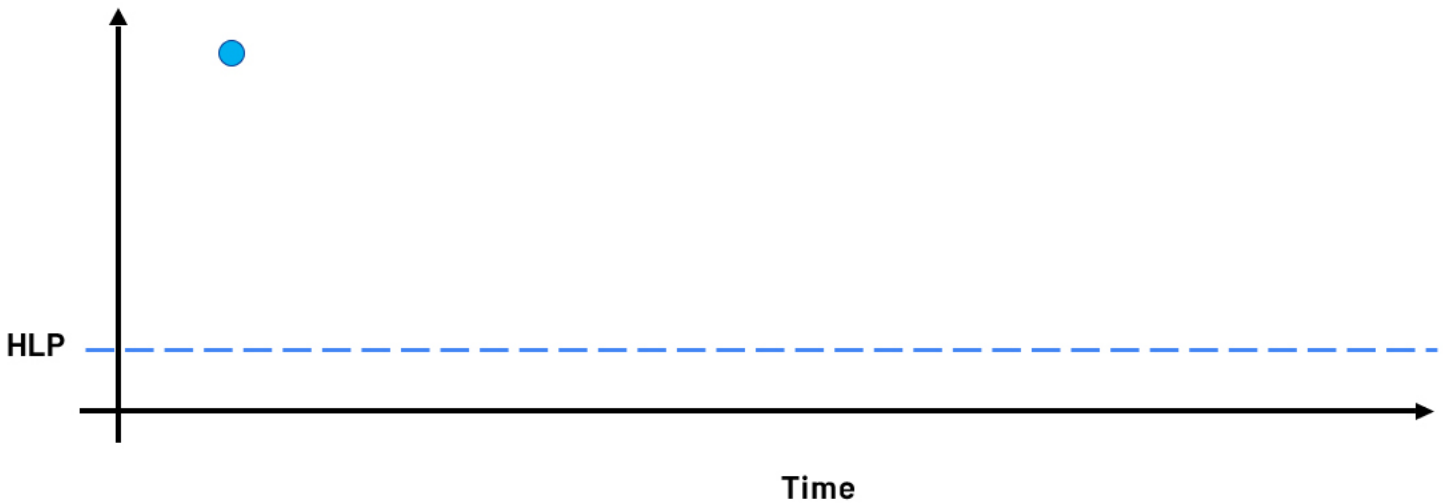
History of project



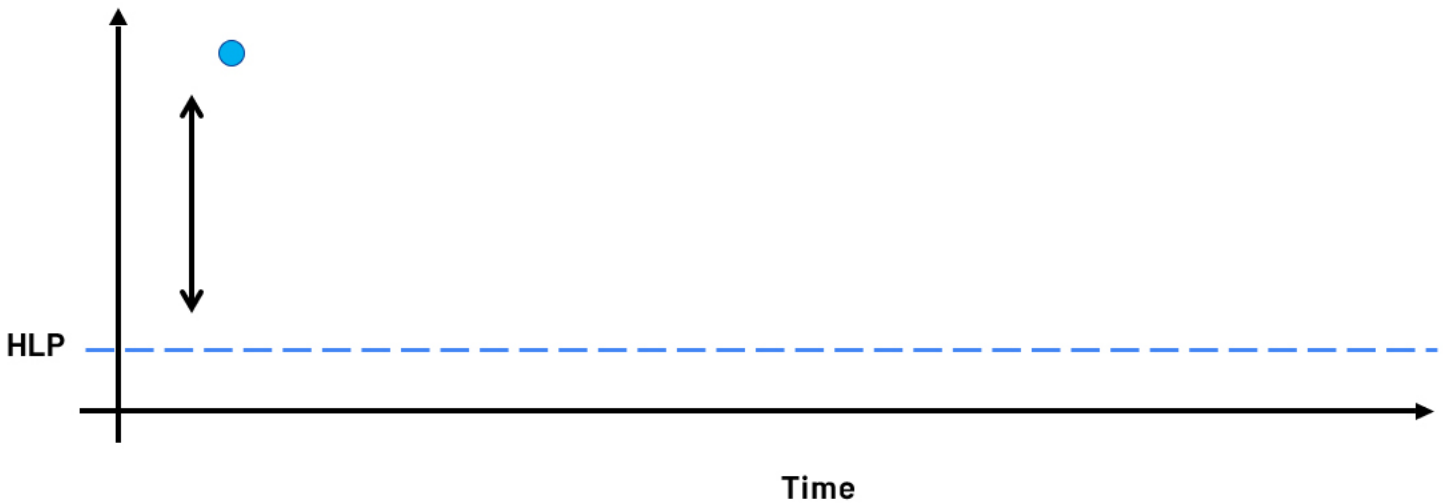
History of project



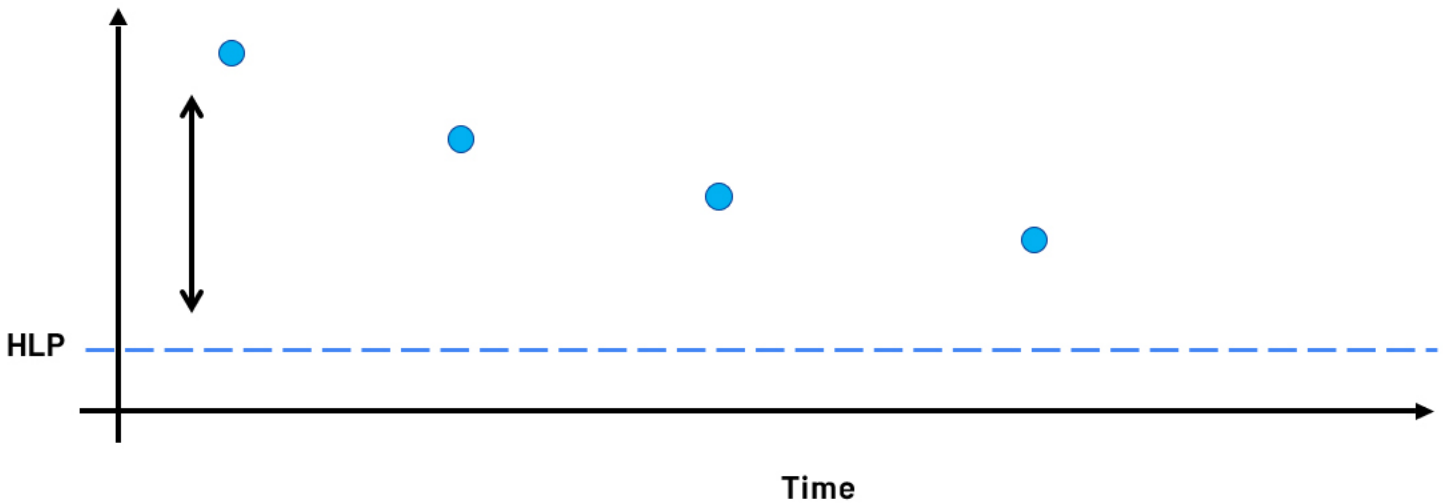
History of project



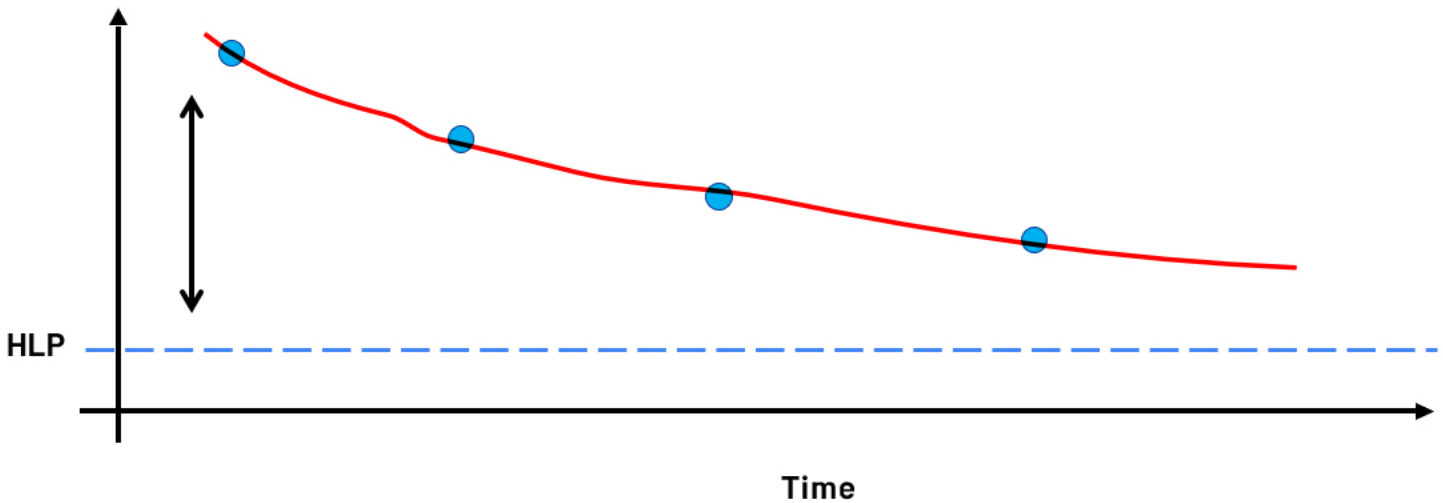
History of project



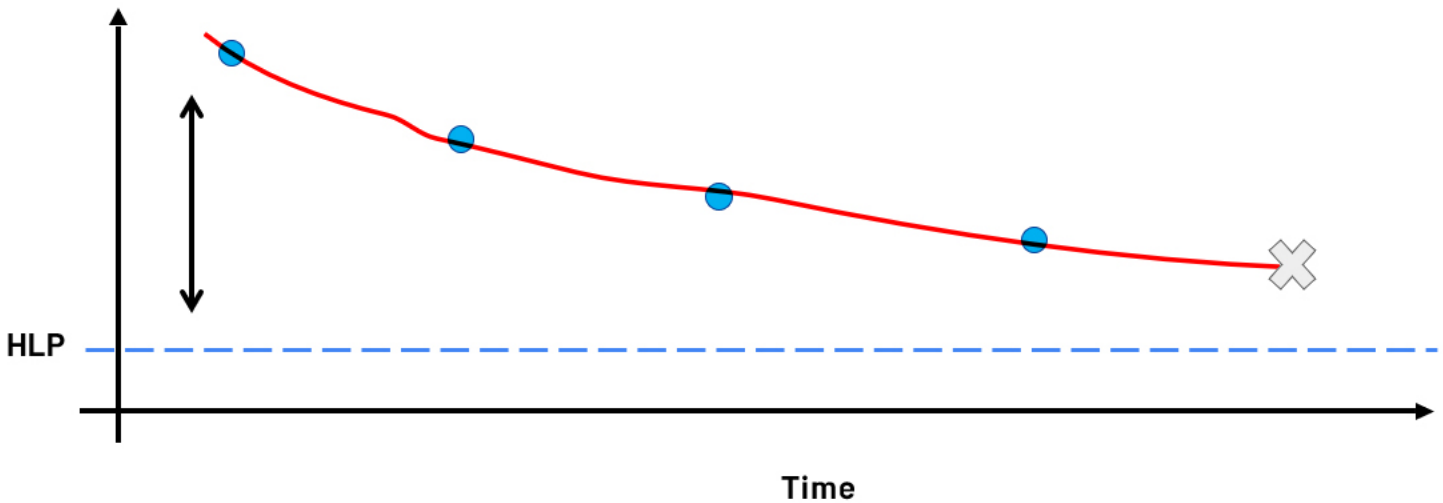
History of project



History of project



History of project



Diligence on value

MLE
metrics

Diligence on value

MLE metrics

Business metrics

Diligence on value

MLE metrics

Business metrics

Word-level accuracy

Diligence on value

MLE metrics

Business metrics

Word-level accuracy

Query-level accuracy

Diligence on value

MLE metrics

Business metrics

Word-level accuracy

Query-level accuracy

Search result quality

Diligence on value

MLE metrics

Business metrics

Word-level accuracy

Query-level accuracy

Search result quality

User engagement

Diligence on value

MLE metrics

Business metrics

Word-level accuracy

Query-level accuracy

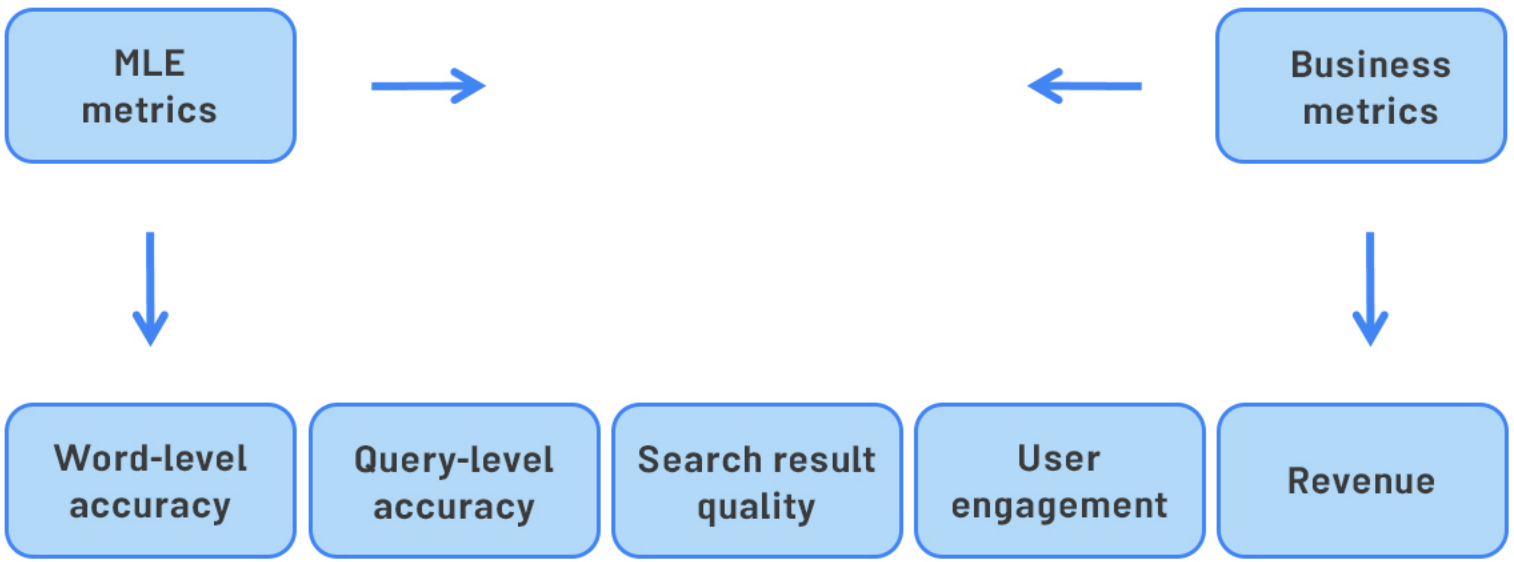
Search result quality

User engagement

Revenue

Diligence on value

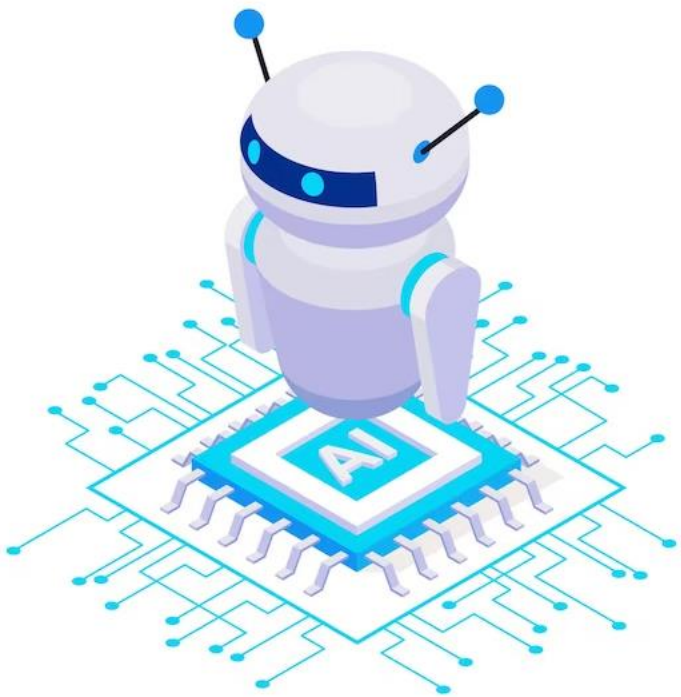
Diligence on value



Have technical and business teams try to agree on metrics that both are comfortable with.

Ethical considerations

- Is this project creating net positive societal value?
- Is this project reasonably fair and free from bias?
- Have any ethical concerns been openly aired and debated?



Milestones

- **Key specifications:**
  - ML metrics (accuracy, precision/recall, etc.)
  - Software metrics (latency, throughput, etc. given compute resources)
  - Business metrics (revenue, etc.)
  - Resources needed (data, personnel, help from other teams)
  - Timeline

If unsure, consider benchmarking to other projects, or building a POC (Proof of Concept) first.