

ORACLE

# Machine Learning Entities

---

# Program agenda

---

- 1 Introduction
- 2 Machine learning entities
- 3 ML entity lifecycle
- 4 ML entity design
- 5 ML entity data sourcing
- 6 Best Practices

# Program agenda

---

- 1 **Introduction**
- 2 Machine learning entities
- 3 ML entity lifecycle
- 4 ML entity design
- 5 ML entity data sourcing
- 6 Best Practices

# Introduction

Which entity information do you spot?

---

I want to expense a wild water ride of \$75 issued by Rafting 4 Life

I had a bacon sandwich for \$16 at Sandy's Beachhouse

On June 12 2021 I paid \$23 at Mike's Parking City for car cleaning

I want to return these shoes because they don't fit

# Introduction

Which entity information do you spot?

I want to expense a **wild water ride** of **\$75** issued by **Rafting 4 Life**

expense item

currency

merchant

I had a **bacon sandwich** for **\$16** at **Sandy's Beachhouse**

expense item

currency

merchant

On **June 12 2021** I paid **\$23** at **Mike's Parking City** for **car cleaning**

date

currency

merchant

expense item

I want to return these **shoes** because **they don't fit**

product

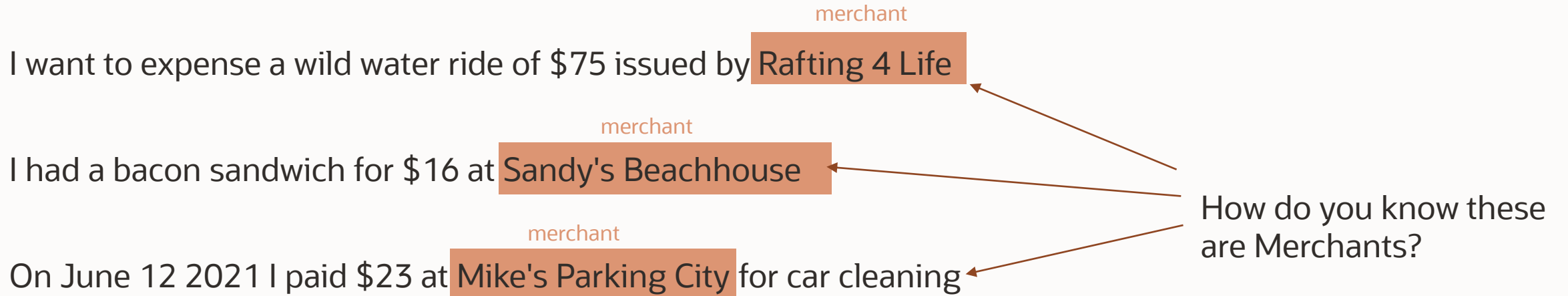
reason

Why did you recognize these values as entities?



# Introduction

Which entity information do you spot?



# Introduction

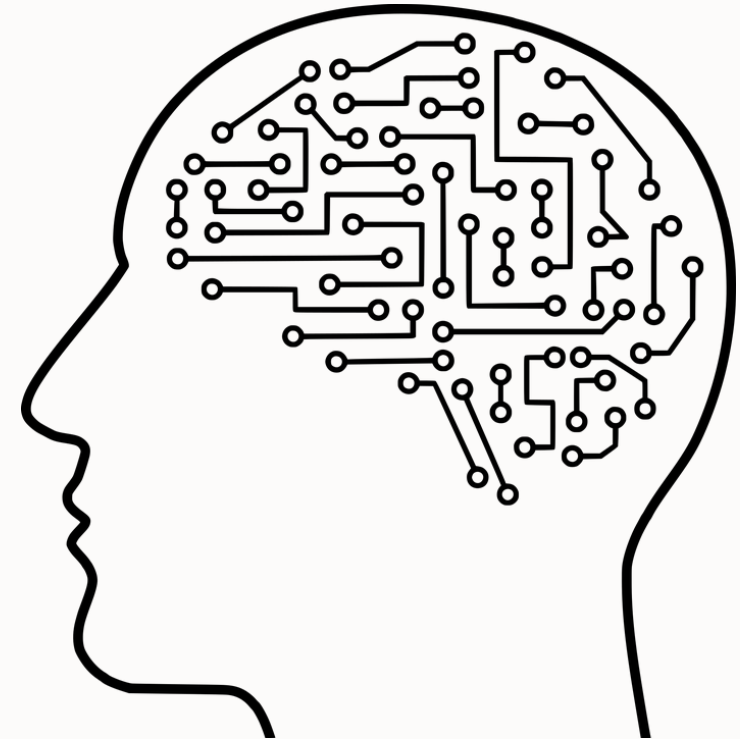
---

How do you know *Sandy's Beachhouse* and *Mike's Parking* are merchants?

Our natural understanding of language allows us to infer from context.

How can we mimic this ability in a conversational AI platform?

- List of Values can hold many Merchant names.
  - But can it hold every possible value?



# Program agenda

---

- 1 Introduction
- 2 **Machine learning entities**
- 3 ML entity lifecycle
- 4 ML entity design
- 5 ML entity data sourcing
- 6 Best Practices



”

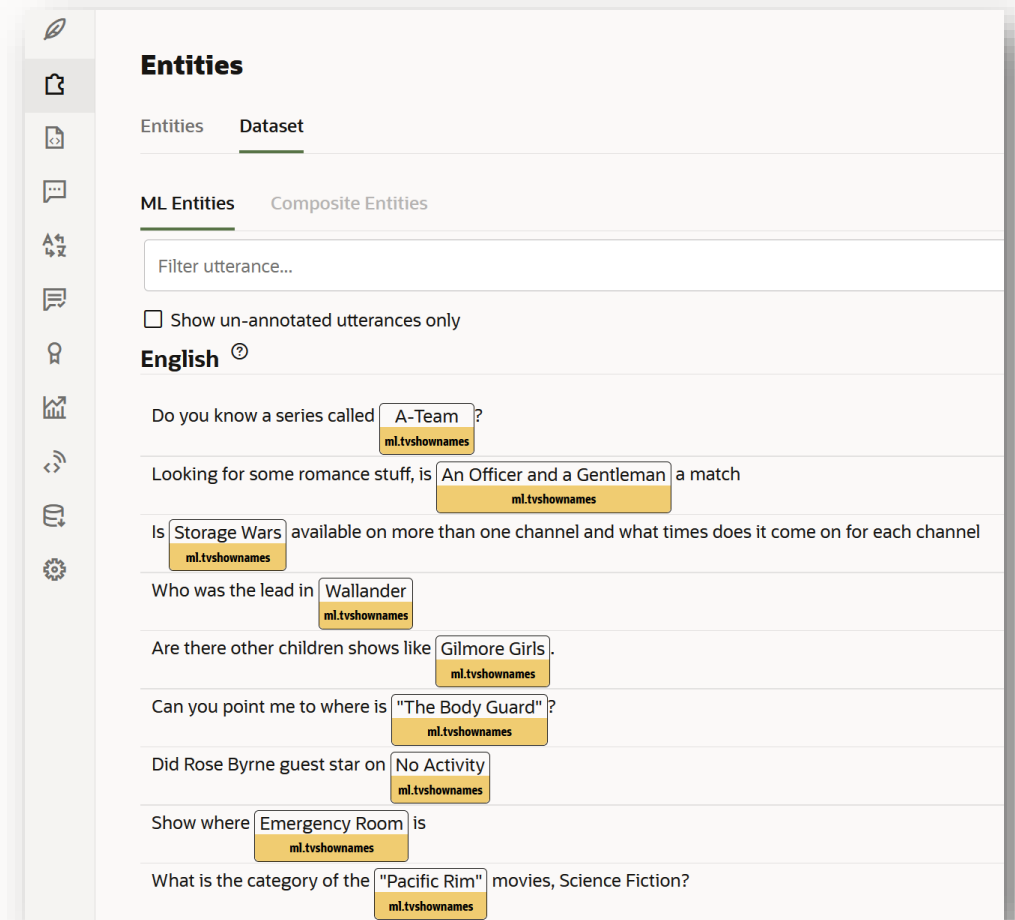
In some cases, we do not know all the possible values or the entity patterns either. Machine Learning entities solves this problem.

# Machine learning entities

ML entities extract known and unknown values from an infinite list of options

The machine model is trained with utterances that are annotated for the ML entities it contains

The model does not learn from the values themselves, but instead by how and where these values are used in an utterance.



The screenshot displays a software interface for managing machine learning entities. On the left is a vertical sidebar with icons for home, entities, dataset, filter, settings, and other functions. The main area is titled "Entities" and has two tabs: "Entities" and "Dataset". Under "Entities", there are sub-tabs for "ML Entities" and "Composite Entities". A search bar labeled "Filter utterance..." is present. A checkbox option "Show un-annotated utterances only" is visible. The interface is set to "English". Below, a list of utterances is shown, each with one or more entities highlighted in yellow boxes. Each entity box contains the entity name and the label "ml.tvshownames".

Utterance	Entity
Do you know a series called A-Team ?	A-Team (ml.tvshownames)
Looking for some romance stuff, is An Officer and a Gentleman a match	An Officer and a Gentleman (ml.tvshownames)
Is Storage Wars available on more than one channel and what times does it come on for each channel	Storage Wars (ml.tvshownames)
Who was the lead in Wallander	Wallander (ml.tvshownames)
Are there other children shows like Gilmore Girls	Gilmore Girls (ml.tvshownames)
Can you point me to where is "The Body Guard"?	"The Body Guard" (ml.tvshownames)
Did Rose Byrne guest star on No Activity	No Activity (ml.tvshownames)
Show where Emergency Room is	Emergency Room (ml.tvshownames)
What is the category of the "Pacific Rim" movies, Science Fiction?	"Pacific Rim" (ml.tvshownames)

## Machine learning entities

---

” ML entities are a complement to the existing set, not a magical solution for every use case

Sometimes an entity with a specific purpose is the best solution.

- E.g. If you have a finite list of values, use Value List entity
- E.g. If the value follows specific pattern, use regular expression entity

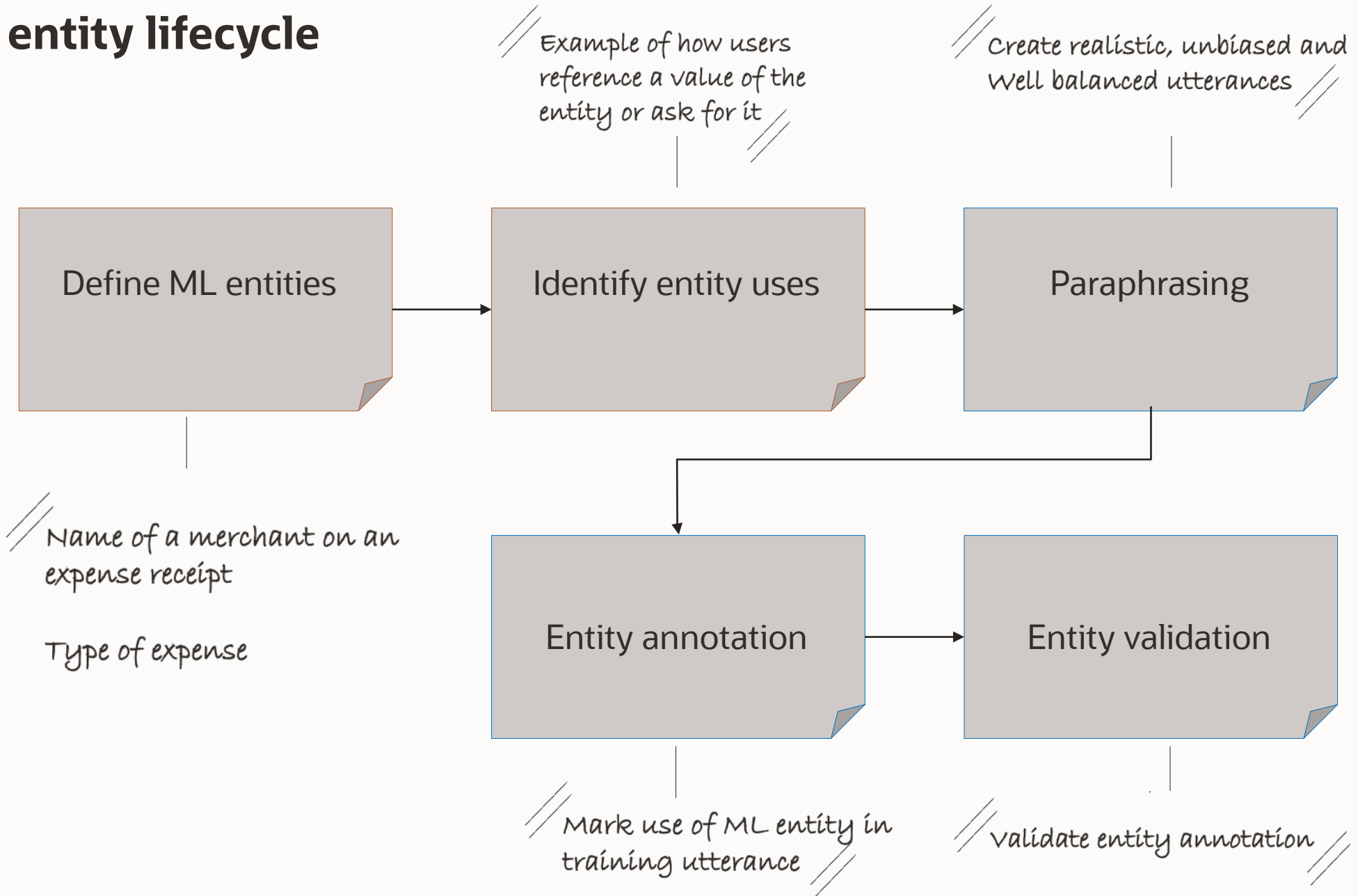
ML entities solve hard challenges, but also require more effort to build a good training data set.

# Program agenda

---

- 1 Introduction
- 2 Machine learning entities
- 3 **ML entity lifecycle**
- 4 ML entity design
- 5 ML entity data sourcing
- 6 Best Practices

# ML entity lifecycle



# Program agenda

---

- 1 Introduction
- 2 Machine learning entities
- 3 ML entity lifecycle
- 4 **ML entity design**
- 5 ML entity data sourcing
- 6 Best Practices

# ML entity design

## Define ML entities

---

An ML entity extracts generalized values, which is why it is important to clearly define and describe it

- When defining entity uses it is important to understand what it is for
- Entity definitions should be clear and distinguishable

## Examples

- Merchant
  - The provider of a product or service for which an expense report is being submitted
- ExpenseItem
  - The product name or service description for which an expense is being submitted

# ML entity design

## Identify entity use cases

---

Use cases communicated to people creating the training utterances.

Examples for a *merchant* in an expense report

- Create expenses
  - "Refund me a \$ 12 sandwich I ate at Johnny London's diner"
  - "I bought a ball pen for 7 US\$ at the Winchester Writer Store"
- View expenses
  - "Show me the list of Walgreens expenses I filed last week "
  - "List my recent Apple expenses"
  - "Have my most recent Whole-Mart groceries been reimbursed?"



# Program agenda

---

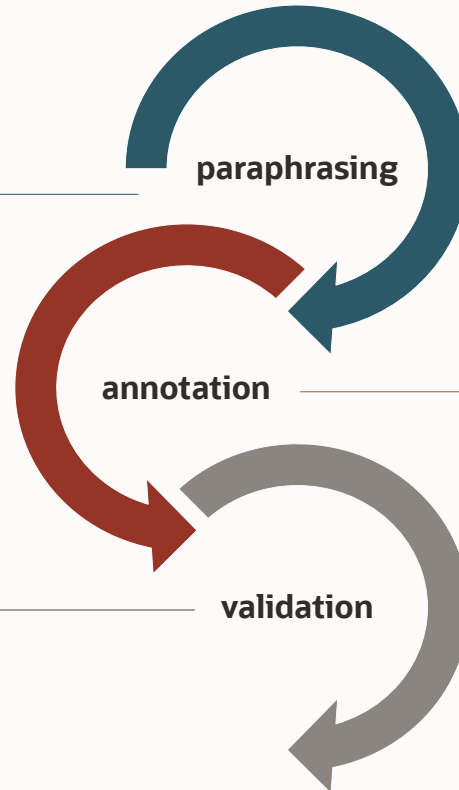
- 1 Introduction
- 2 Machine learning entities
- 3 ML entity lifecycle
- 4 ML entity design
- 5 **ML entity data sourcing**
- 6 Best Practices

# ML entity data sourcing

## Paraphrasing, Annotation and Validation

Crowd sourcing of sample utterances for a use case. An utterance may or may not contain ML entity values

Use to implement a "4 eyes principle" to avoid falsely classified entities



Process in which a crowd worker tells the NLP model where in a sample utterance a particular entity value is located

# ML entity data sourcing with ODA data manufacturing

Data Manufacturing is native feature of Oracle Digital Assistant skills

Allows you to create paraphrasing, annotation and validation jobs for

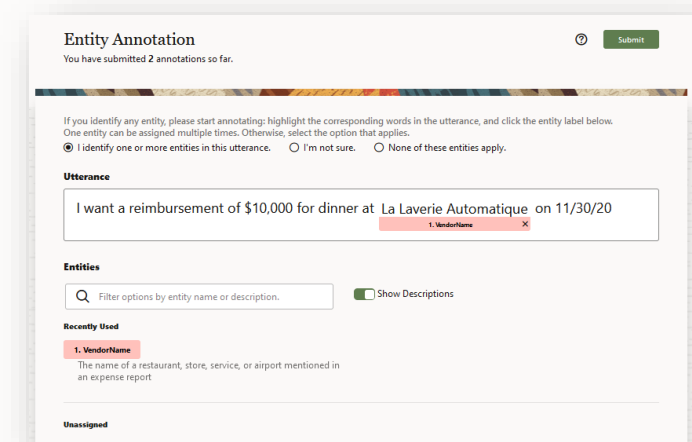
- Intents
- ML Entities

A "job" results in a URL you share with crowd workers

- URL displays form for crowd workers to complete their task

At the end of each job

- You can review the collected data and download them for editing



The screenshot shows the 'Entity Annotation' interface. At the top, it says 'Entity Annotation' and 'You have submitted 2 annotations so far.' with a 'Submit' button. Below this, there is a section for 'Utterance' with the text: 'I want a reimbursement of \$10,000 for dinner at La Laverie Automatique on 11/30/20'. The words 'La Laverie Automatique' are highlighted in red, and a red box labeled '1. VendorName' is positioned below them. Below the utterance, there is a section for 'Entities' with a search bar and a 'Show Descriptions' toggle. Under 'Recently Used', there is a list item for '1. VendorName' with a description: 'The name of a restaurant, store, service, or airport mentioned in an expense report'. At the bottom, there is a section for 'Unassigned'.

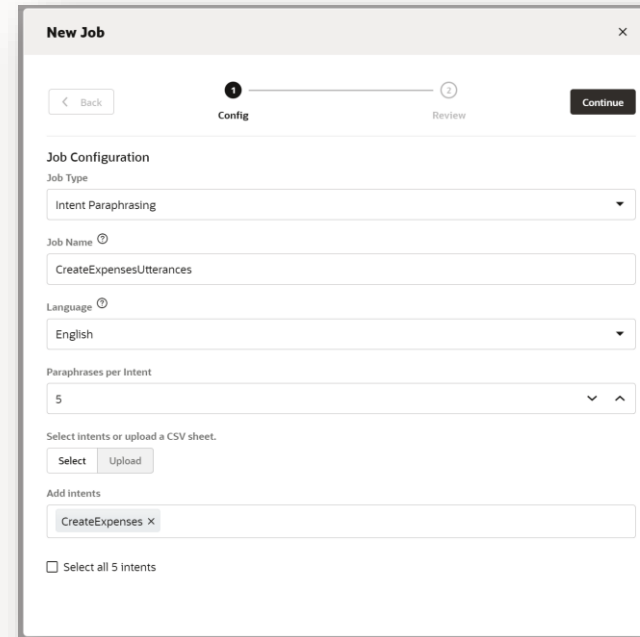
# ML entity data manufacturing

## Paraphrasing

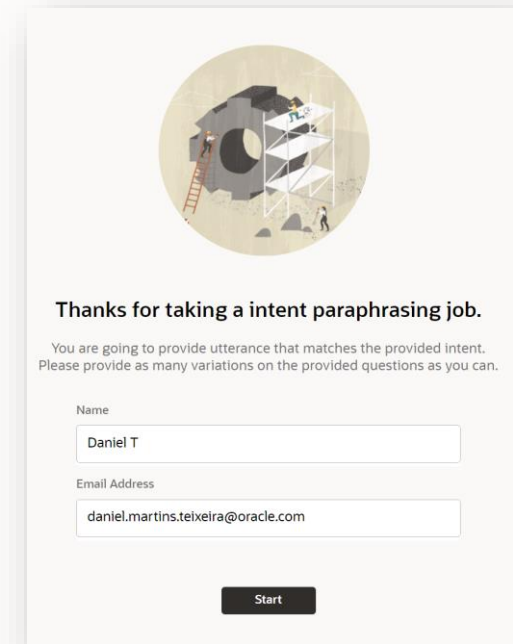
With paraphrasing we can ask the crowd to provide sample utterances

This allows you to gather real life data that will improve the model

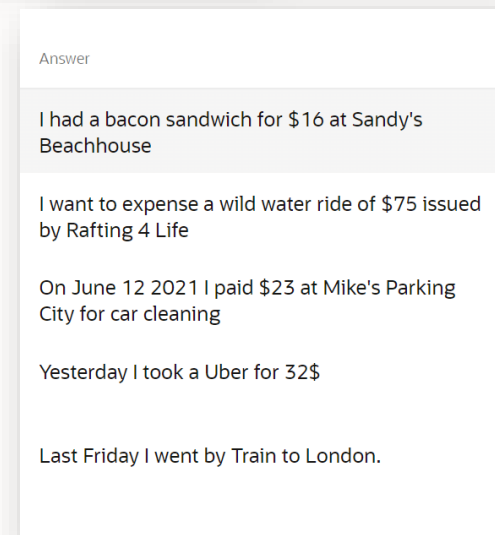
Paraphrasing adds variety that enriches the quality of gathered utterances



The screenshot shows a 'New Job' configuration window with a progress bar at the top indicating the 'Config' step. The 'Job Configuration' section includes a 'Job Type' dropdown set to 'Intent Paraphrasing', a 'Job Name' field containing 'CreateExpensesUtterances', and a 'Language' dropdown set to 'English'. The 'Paraphrases per Intent' is set to 5. Below this, there are 'Select' and 'Upload' buttons for selecting intents or uploading a CSV sheet. An 'Add intents' section shows a tag for 'CreateExpenses' with a close button. At the bottom, there is a checkbox for 'Select all 5 intents'.



The screenshot shows a confirmation screen with a circular illustration of a gear and a person. The text reads: 'Thanks for taking a intent paraphrasing job. You are going to provide utterance that matches the provided intent. Please provide as many variations on the provided questions as you can.' Below this, there are input fields for 'Name' (containing 'Daniel T') and 'Email Address' (containing 'daniel.martins.teixeira@oracle.com'). A 'Start' button is at the bottom.



The screenshot shows an 'Answer' section with five generated paraphrases:

- I had a bacon sandwich for \$16 at Sandy's Beachhouse
- I want to expense a wild water ride of \$75 issued by Rafting 4 Life
- On June 12 2021 I paid \$23 at Mike's Parking City for car cleaning
- Yesterday I took a Uber for 32\$
- Last Friday I went by Train to London.

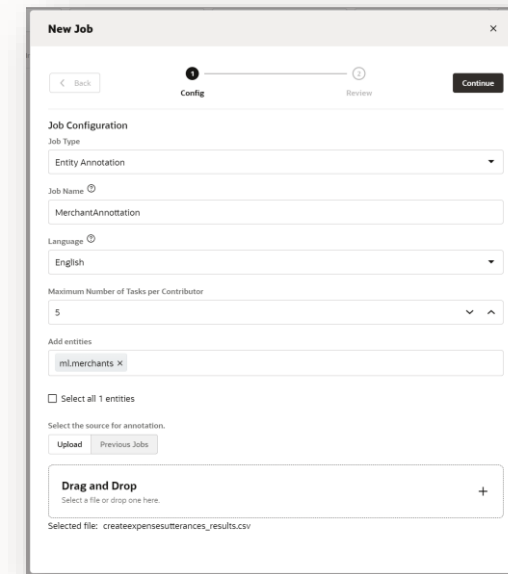
# ML entity data manufacturing

## Annotation

Once we have the utterances, we need to Annotate the entities

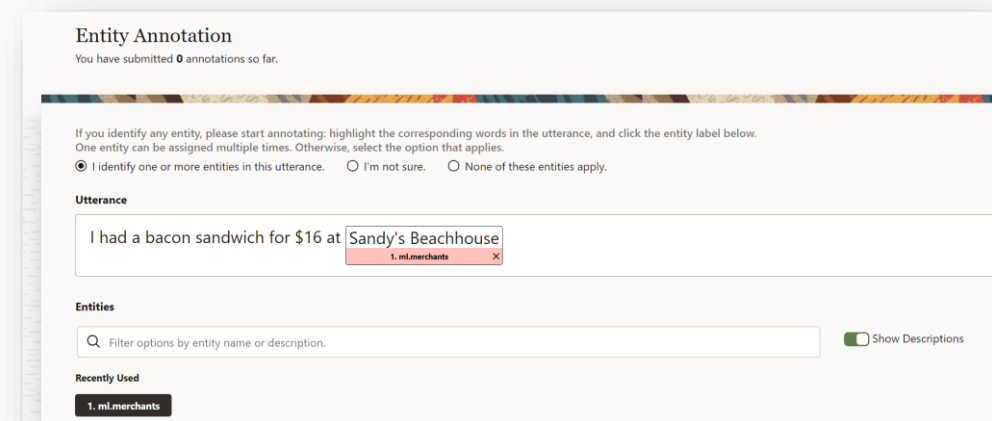
That means “labeling” them in the utterance.

We can also choose to not annotate. This helps to train the model for those cases where no entity is present.



The 'New Job' configuration interface includes a progress bar with 'Config' and 'Review' steps, and a 'Continue' button. The 'Job Configuration' section contains the following fields:

- Job Type: Entity Annotation
- Job Name: MerchantAnnotation
- Language: English
- Maximum Number of Tasks per Contributor: 5
- Add entities: ml.merchants X
- Select all 1 entities:
- Select the source for annotation: Upload Previous Jobs
- Drag and Drop: Select a file or drop one here. Selected file: createexpensesutterances\_results.csv



The 'Entity Annotation' interface shows the following components:

- Header: Entity Annotation, You have submitted 0 annotations so far.
- Instructions: If you identify any entity, please start annotating: highlight the corresponding words in the utterance, and click the entity label below. One entity can be assigned multiple times. Otherwise, select the option that applies.
- Options:  I identify one or more entities in this utterance.  I'm not sure.  None of these entities apply.
- Utterance: I had a bacon sandwich for \$16 at Sandy's Beachhouse (with a red box around 'Sandy's Beachhouse' and a label '1. ml.merchants X' below it).
- Entities: Search bar with 'Filter options by entity name or description.' and a 'Show Descriptions' toggle.
- Recently Used: 1. ml.merchants

# ML entity data manufacturing

## Validation

The last step is the validation of the previous annotations

This final step is the gatekeeping for quality control and that we are not inserting bad data into the model

**New Job**

Back **1** Config Review Continue

**Job Configuration**

Job Type  
Entity Validation

Job Name  
ValidateMerchants

Language  
English

Maximum Number of Tasks per Contributor  
5

Select the source for validation.  
Upload Previous Jobs

Add Jobs  
MerchantAnnotation X

Select all 1 jobs

**Entity Validation**  
You have submitted 0 validation(s) so far.

Examine whether the entity labels assigned on the utterance are correct or not.  
**Utterance**

I had a bacon sandwich for \$16 at Sandy's Beachhouse  
ml.merchants

Correct Incorrect Not Sure

This phrase has been annotated correctly because:  
All entities have been identified.  
Each term has been annotated with the correct entity.  
Each term has been selected completely.

**Entities**

Entities Dataset

**ML Entities** Composite Entities

Filter utterance...

Show un-annotated utterances only

**English**

I had a bacon sandwich for \$16 at Sandy's Beachhouse  
ml.merchants

On June 12 2021 I paid \$23 at Mike's Parking City for car cleaning  
ml.merchants

Yesterday I took a Uber for 32\$  
ml.merchants

I want to expense a wild water ride of \$75 issued by Rafting 4 Life  
ml.merchants

Page 1 (1-4 items) |< < 1 > >|

# Program agenda

---

- 1 Introduction
- 2 Machine learning entities
- 3 ML entity lifecycle
- 4 ML entity design
- 5 ML entity data sourcing
- 6 **Best Practices**

## Best practices

---

### ” ML models rely on good training data

Ensure each ML entity has a similar number of value occurrences in sample utterances

Make sure training utterances vary in the structure and wording of a message

If a ML entity value can be provided in different formats, ensure examples for those values are balanced too

Provide negative examples to prevent false positives

- utterances that don't contain entity values



## Best practices

---

1000

This is the recommended number for the total amount of utterances

## Best practices

---

80  
—  
20

A split of 80/20 is recommended as the ratio between training data and test data in machine learning

When creating utterances for testing

- Use utterances with the same structure as training utterances
- Use utterances with a slightly different structure from utterance in training
- Use entity values that are used in the training utterance
- Use entity values that are unknown to the ML entity model

## Best practices

---

” A model is only as good as the data it is trained on

Accuracy of the ML model is improved

- for user messages having the same or a similar structure as the training structures
- for user message containing entity values that were used for training

Larger data sets with many variations work best for user messages containing entity values that were not included in the training data set

If a wide variety of values is expected to be used by users, then you need to train the model with many unique entity values instead of a few that are used repeatedly

## Best practices

---

” A model is only as good as the data it is trained on

If users may enter 2- or 3-word long entity values, then these must also be represented in the training data

Consider special characters in the training data if they are expected in entity values in user messages

If you are using multiple ML entities, be sure to provide the same number of values in the training data

Always add negative example to training utterances

- Sentence that does not contain an entity value

---

”

ML entities may not be telling  
the truth

---

false positives

false negatives

split words

common words

under or over prediction

---

## Alternative data sourcing

---

” In a perfect world, you would use real data to create training and test utterances. But when it's not there, you can mimic real data until you can collect real data in a production environment.

### Synthesizing (generating)

Identify representative messages and use a generation tool to create multiples of them

Generates large data sets with ease

- Low cost
- Large data sets from a few seed sentences

ORACLE