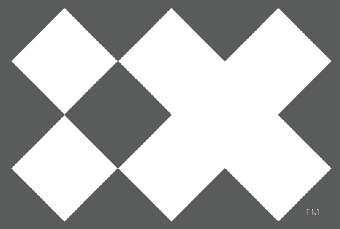
Watson Assistance Best Practices

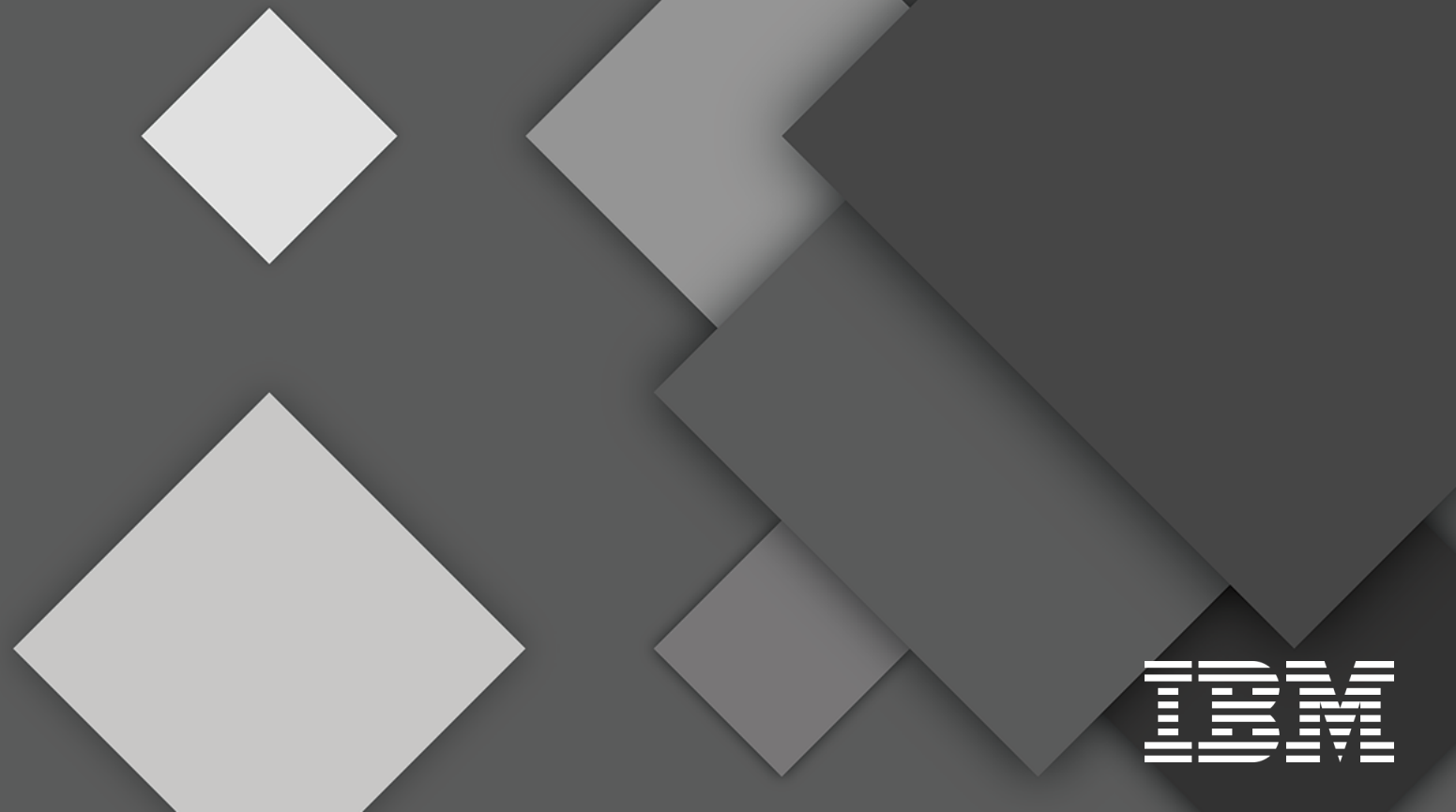
02.26.2021

- Ivan Portilla
- ivanp@us.ibm.com
- github.com/jiportilla/ic

IBM

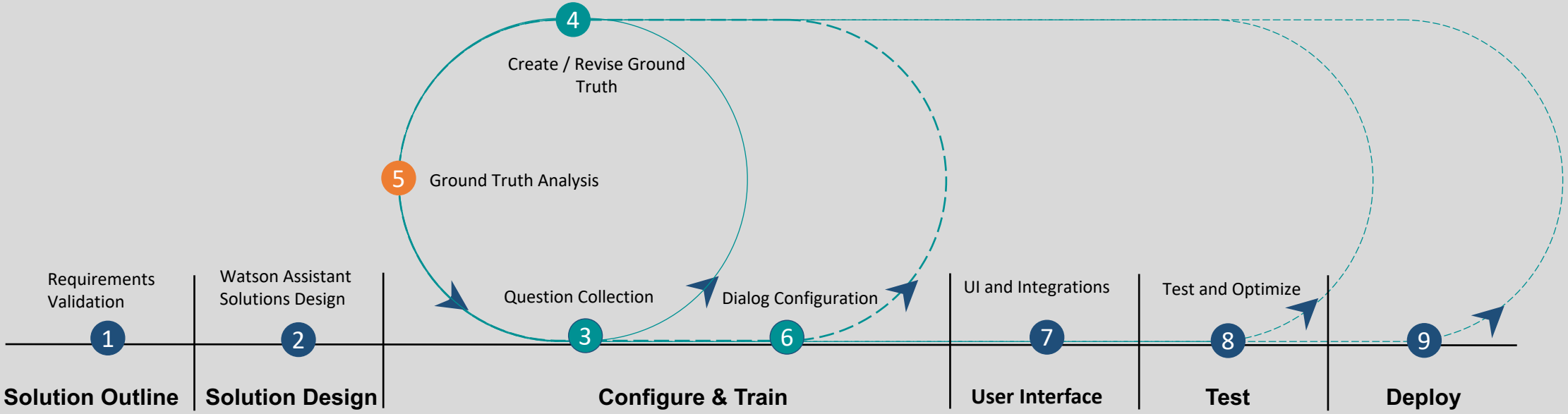
Improvement Framework





Watson Assistant Implementation Cycle

Ground Truth Analysis



1. Validate User Scenario, Use Cases and Requirements, and Watson Assistant Technology Pattern
2. Define Watson solution design
3. Collect representative questions in “voice” of end users that will be used to teach and test Watson Assistant
4. Group representative questions into intent classes that will be used to create Ground truth
5. **Evaluate the Ground truth**
6. Configure the Dialog component for the conversational flow you wish to have with your end users. Question and answering, chit chat, off topic, disambiguation, etc.
7. Configure a UI to access Watson Assistant and any integrations as needed for the solution
8. Evaluate performance, gather new questions, revise ground truth and dialog. Repeat.
9. Deploy Watson MVP to Pilot

Agenda – Improvement Framework

1. Establishing a Baseline
2. Coverage
3. Effectiveness
4. AI, ML , Supervised Learning
5. Advanced Analytics

Establishing a Baseline

Establishing a Baseline – Business KPIs

Business Key Processes Indicators (KPIs) are “King”

If it costs more to staff a team to to maintain a virtual assistant than the assistant saves in human agent support cost, then there’s a problem…

Cost

Revenue

Engagement

IBM

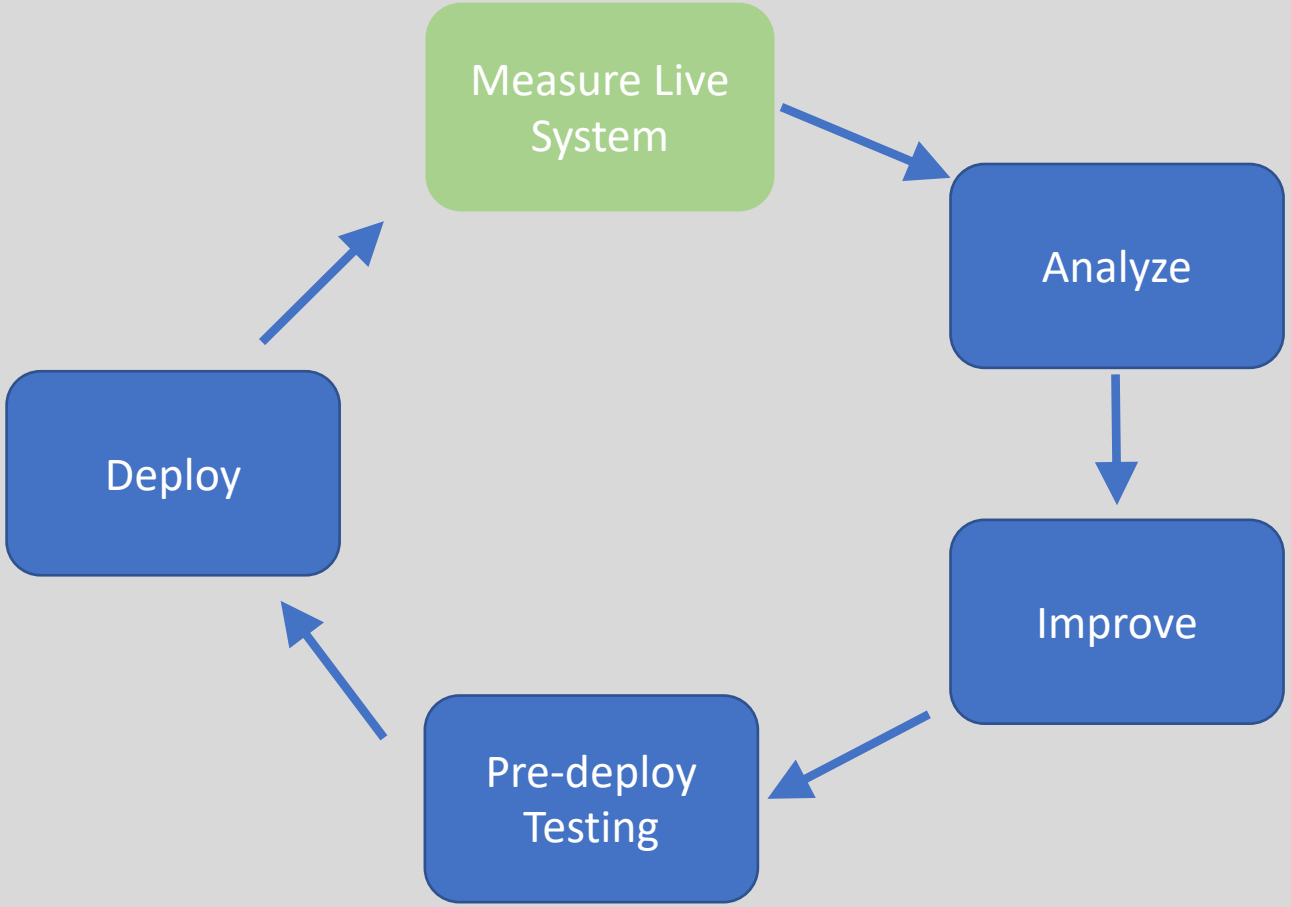
Establishing a Baseline – Why?

Provides an understanding of **performance**

Allows you to **prioritize** your improvement effort

Makes improvement as **efficiently** as possible

Establishing a Baseline – Process Phases



Coverage

Coverage is the percentage of the total conversations or messages your assistant attempts to engage

$$\text{Coverage} = \frac{\text{Questions answered}}{\text{Total Number of Questions}} \times 100$$

Coverage

Coverage is the percentage of the total conversations or messages your assistant attempts to engage

*Coverage is a view of the **range** and **depth** of subject matter your assistant is trained on*

*Coverage can be measured by **conversation** or by **message***

*The intent **confidence thresholds** you set directly impact coverage*

*Coverage can be measured live in **production** and **offline** with test sets or historic logs*

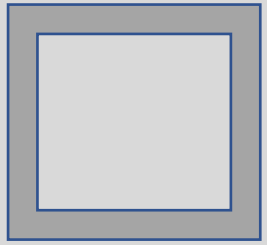
Coverage

Q1

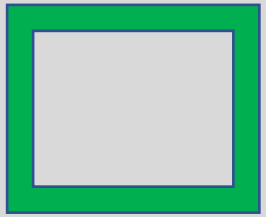
Q2

Watson Assistant

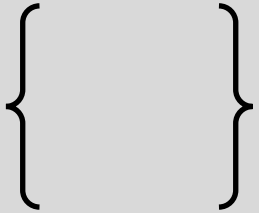
Intent 1



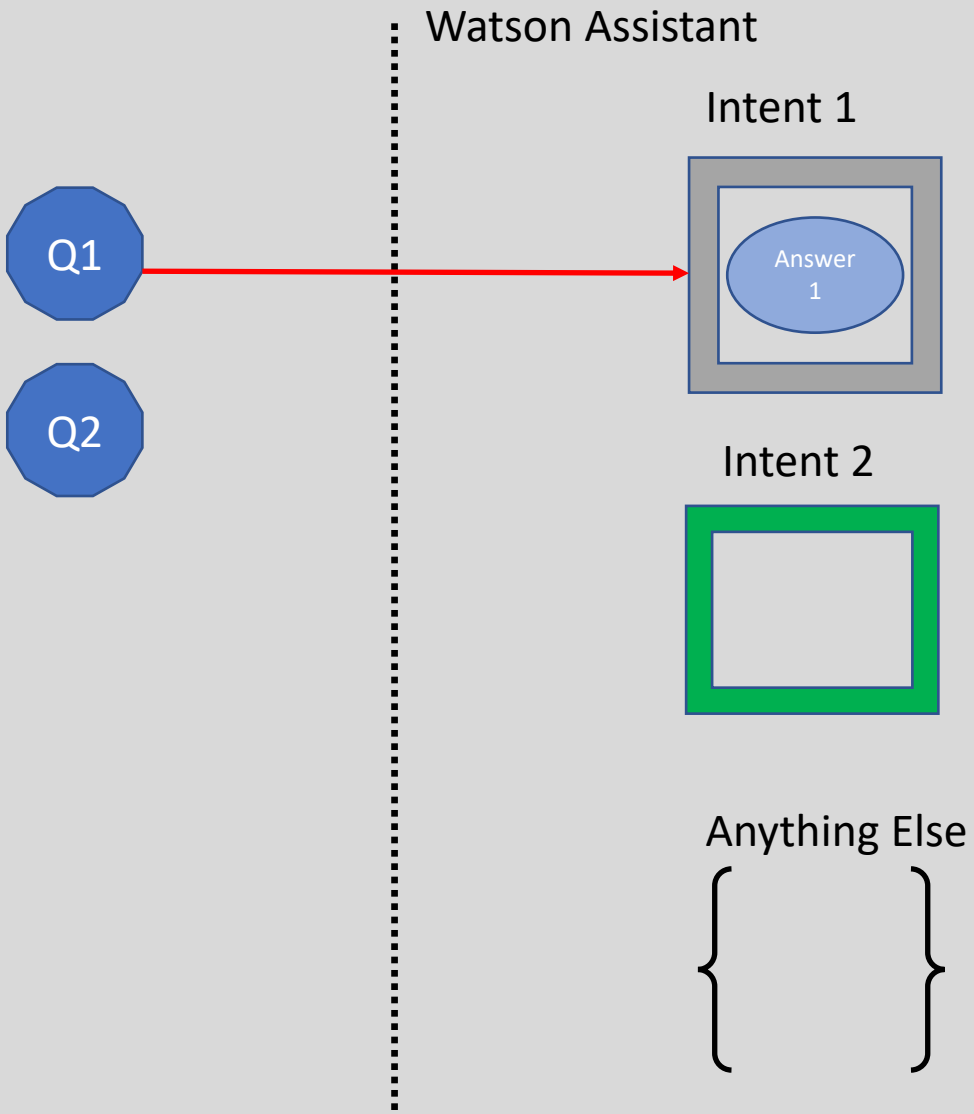
Intent 2



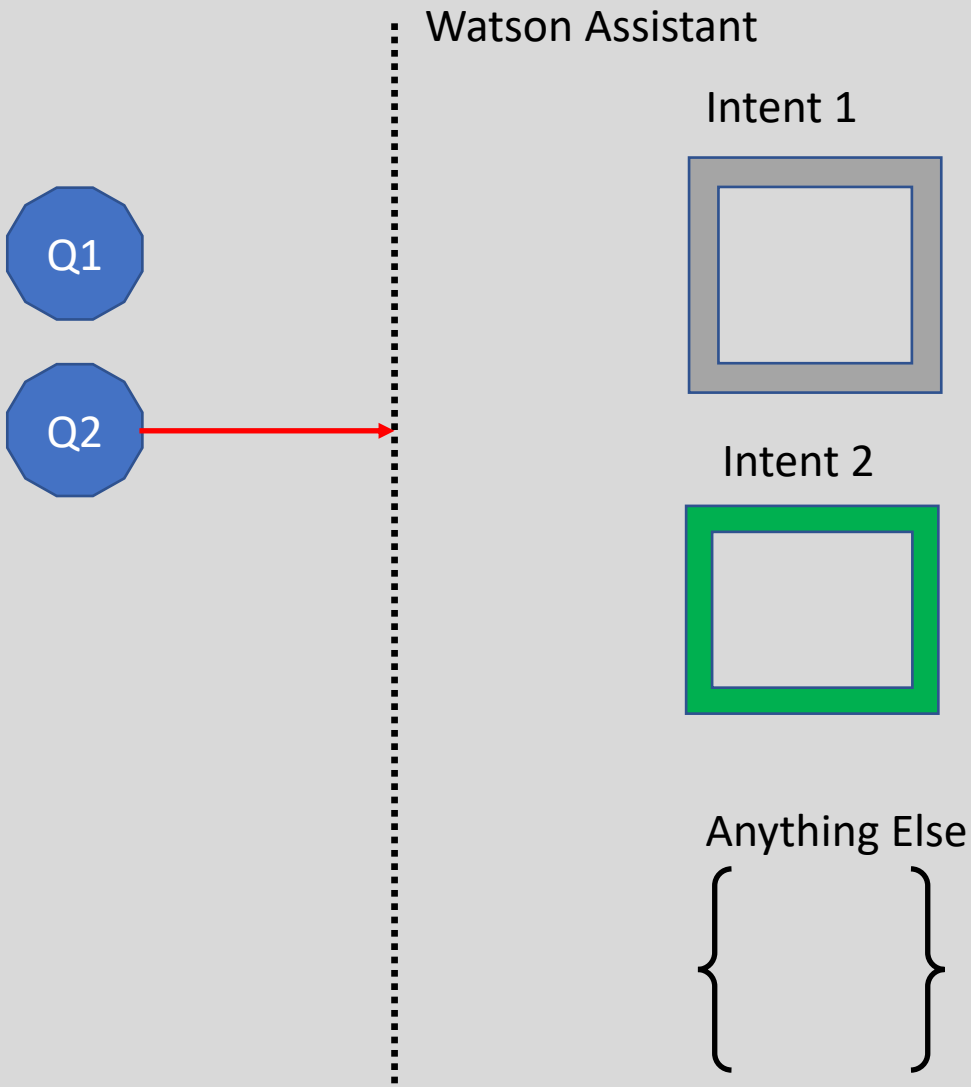
Anything Else



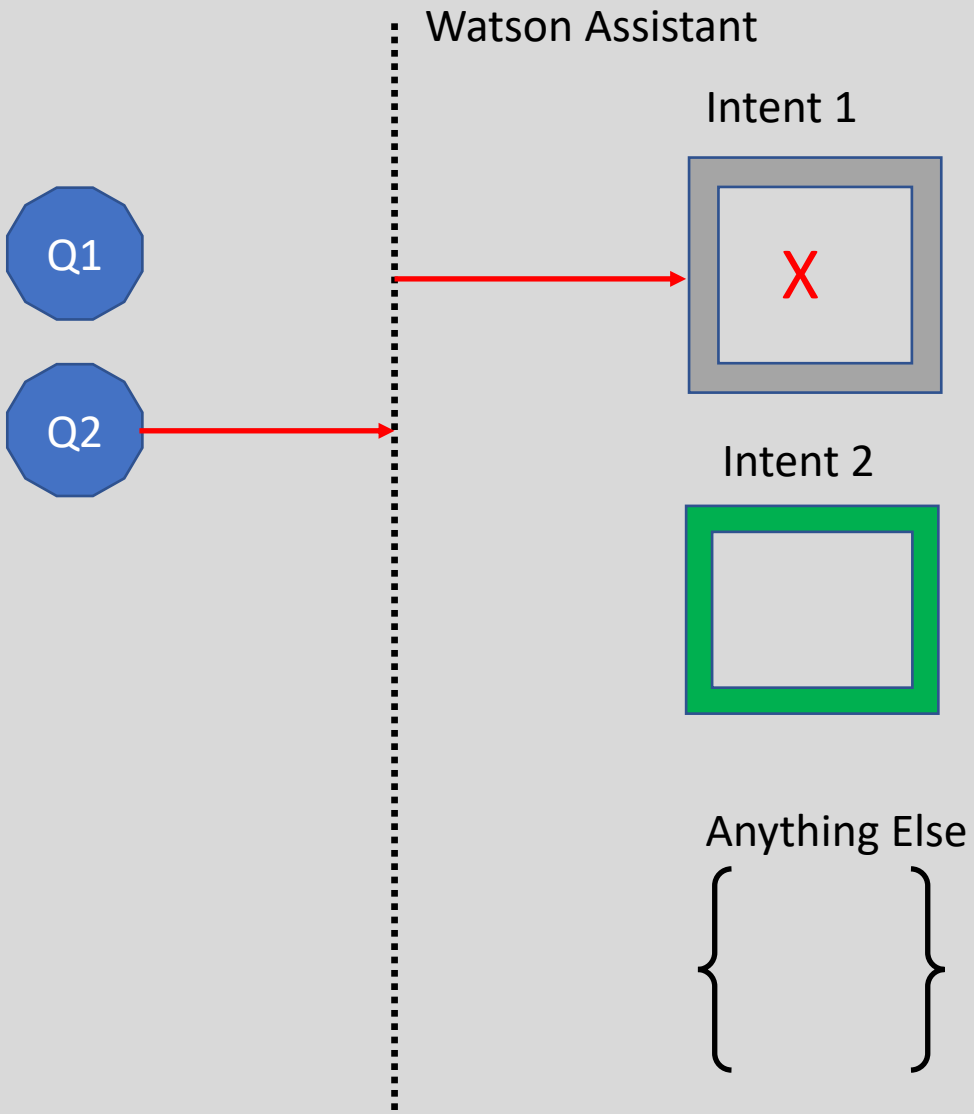
Coverage


Watson Assistant
Intent 1
Q1
Answer 1
Q2
Intent 2
Anything Else

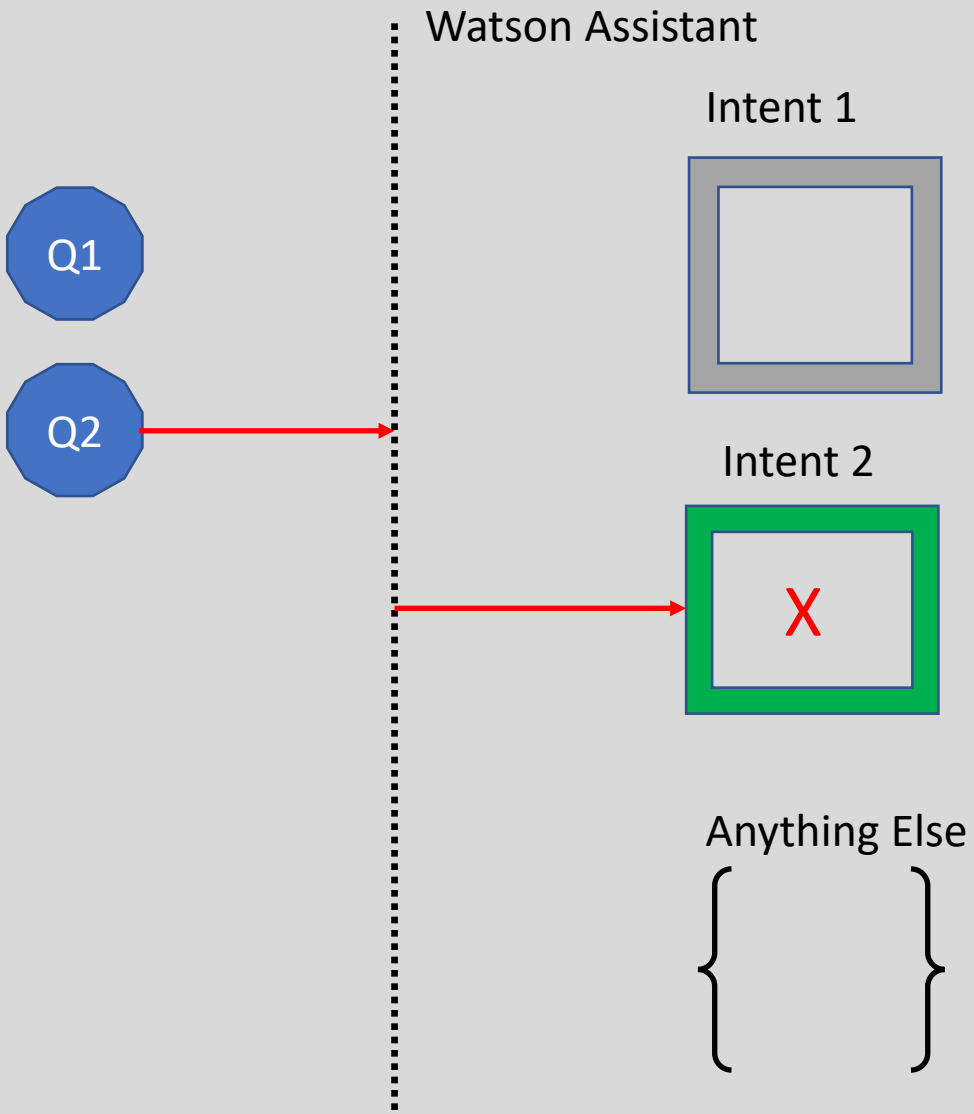
Coverage



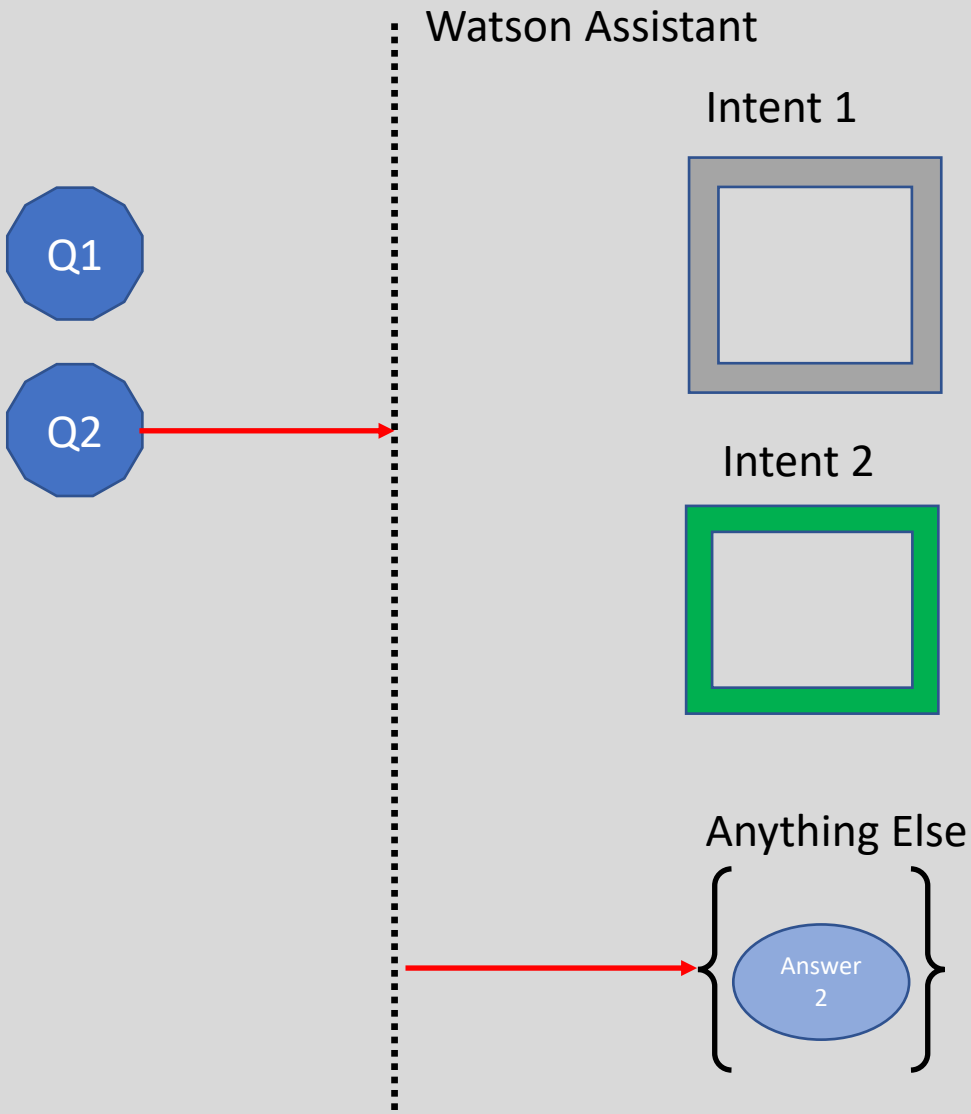
Coverage



Coverage



Coverage


Watson Assistant
Intent 1
Q1
Q2
Intent 2
Anything Else
Answer
2

Effectiveness - Defined

Effectiveness is the quality of the experiences your assistant provided during the conversations and messages it did engage

*Effectiveness can be measured **live** in production with metrics dashboards or **off-line** using labeled test sets or training data.*

Measurements of effectiveness include:

- *Conversation containment*
- *Conversation success*
- *Intent confidence*
- *Sentiment analysis*
- *Precision*
- *Explicit user feedback*

Effectiveness

Service Desk Assistant

Hello

Hello, how can I help you?

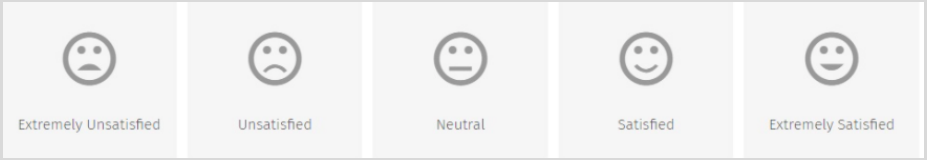
I need to cancel my cell phone contract

Sorry to hear that, may I ask you why?

Coverage in my area is very bad

We are very sorry, but you must call our customer service at 1-800-123-4567

How was your experience today?

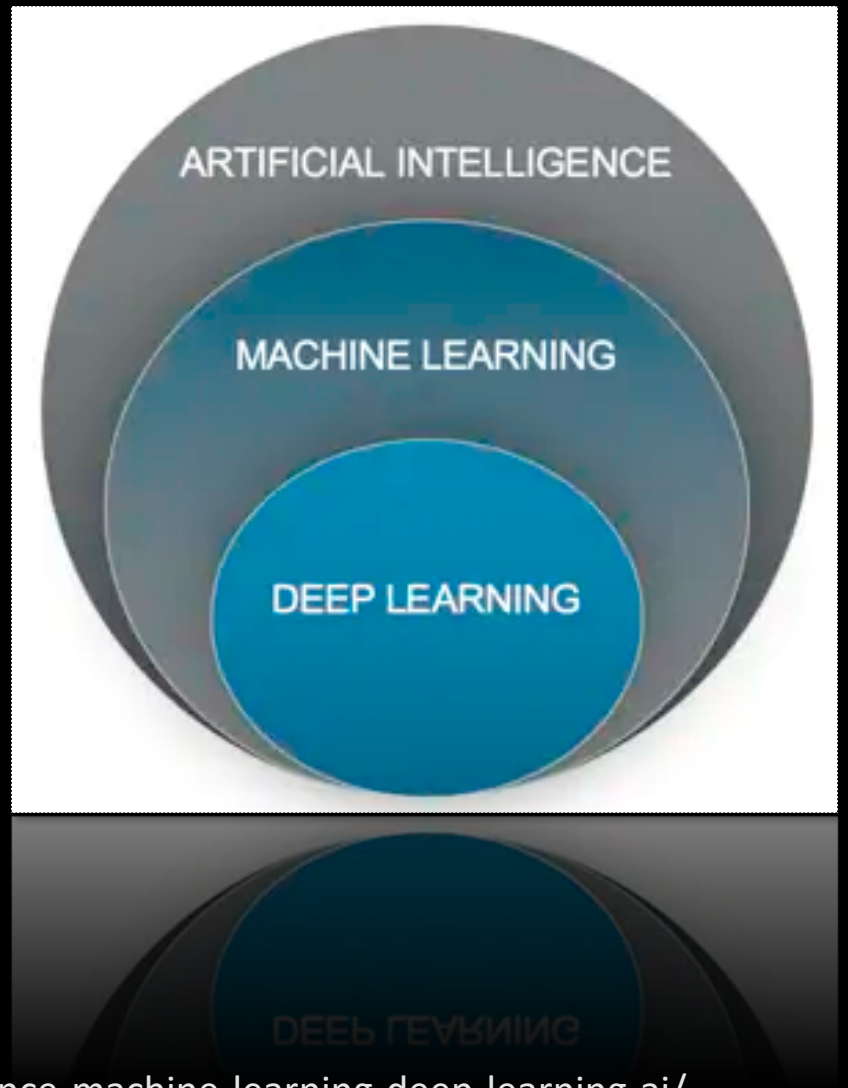


Agenda – Improvement Framework

1. Establishing a Baseline
2. Coverage
3. Effectiveness
4. AI, ML , Supervised Learning
5. Advanced Analytics

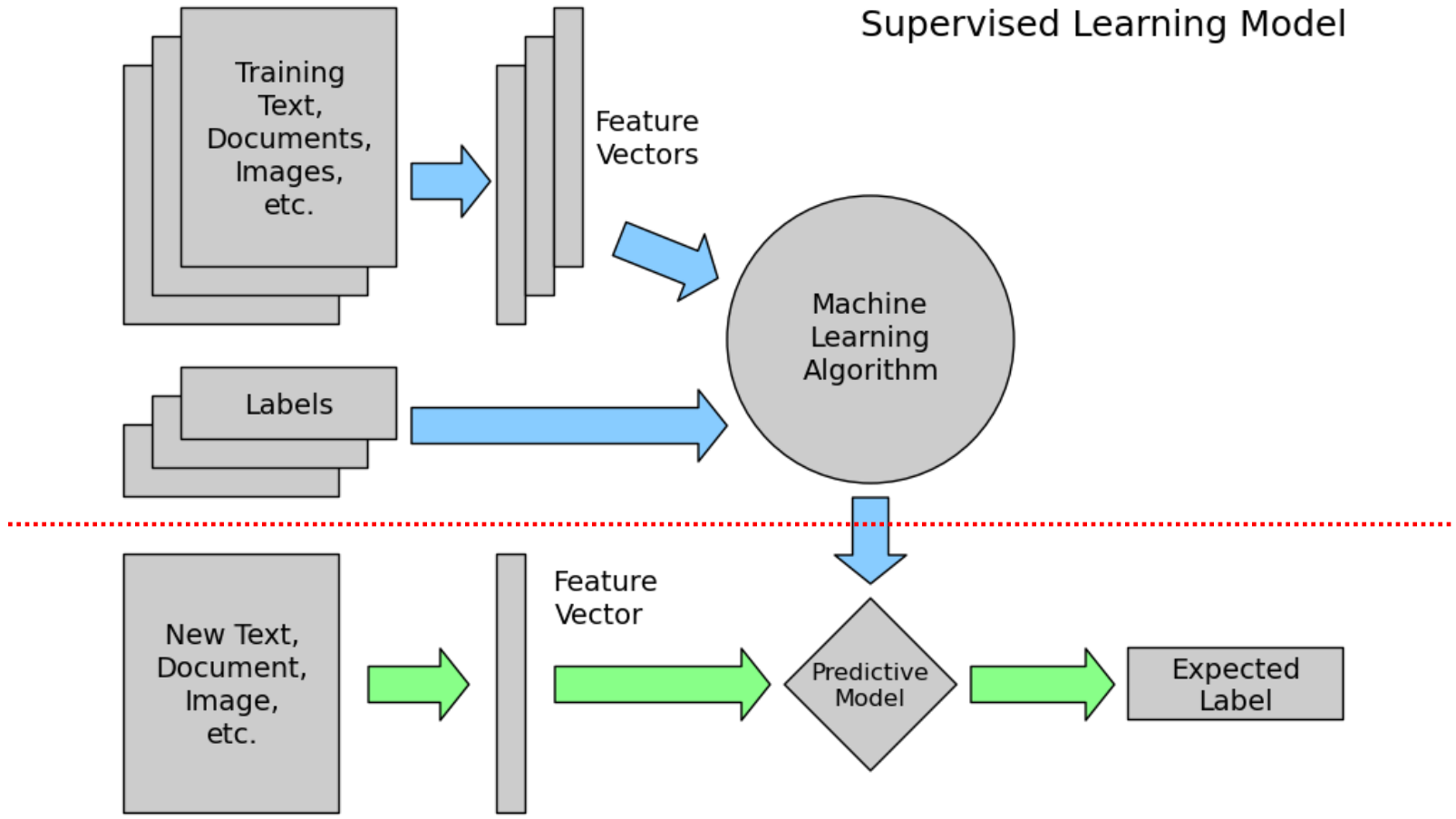
Recap: Definitions

- AI: Intelligence demonstrated by machines
- ML: Set of algorithms that allow computers to learn from (big) data
- DL: Set of learning techniques focused on models and neural networks



https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

Supervised Machine Learning



Compare test data & training data

Test Data Evaluation

	Intent	% of Train	% of Test	Absolute Difference %	Train Examples	Test Examples
0	General_Greetings	15.080000	63.830000	48.750000	30	30
1	Help	4.020000	17.020000	13.000000	8	8
2	Cancel	3.520000	14.890000	11.380000	7	7

Distribution Mismatch Color Code

Red - Severe

Blue - Caution

Green - Good

Data Distribution Divergence Test vs Train 76.0%

Ideally the Test and Training Data distributions should be similar. The following metrics can help identify gaps between Test Set and Training Set:

1. The distribution of User Examples per Intent for the Test Data should be comparable to the Training Data

2. Average length of User Examples for Test and Training Data should be comparable

3. The vocabulary and phrasing of utterances in the Test Data should be comparable to the Training Data

If your test data comprises of examples labelled from your logs, & the training data comprises of examples created by human subject matter experts, there may be discrepancies between what the virtual assistant designers thought the end users would type and the way they actually type in production. Thus, if you find discrepancies in this section, you might want to consider changing your design to more closely resemble the way end users use your system.

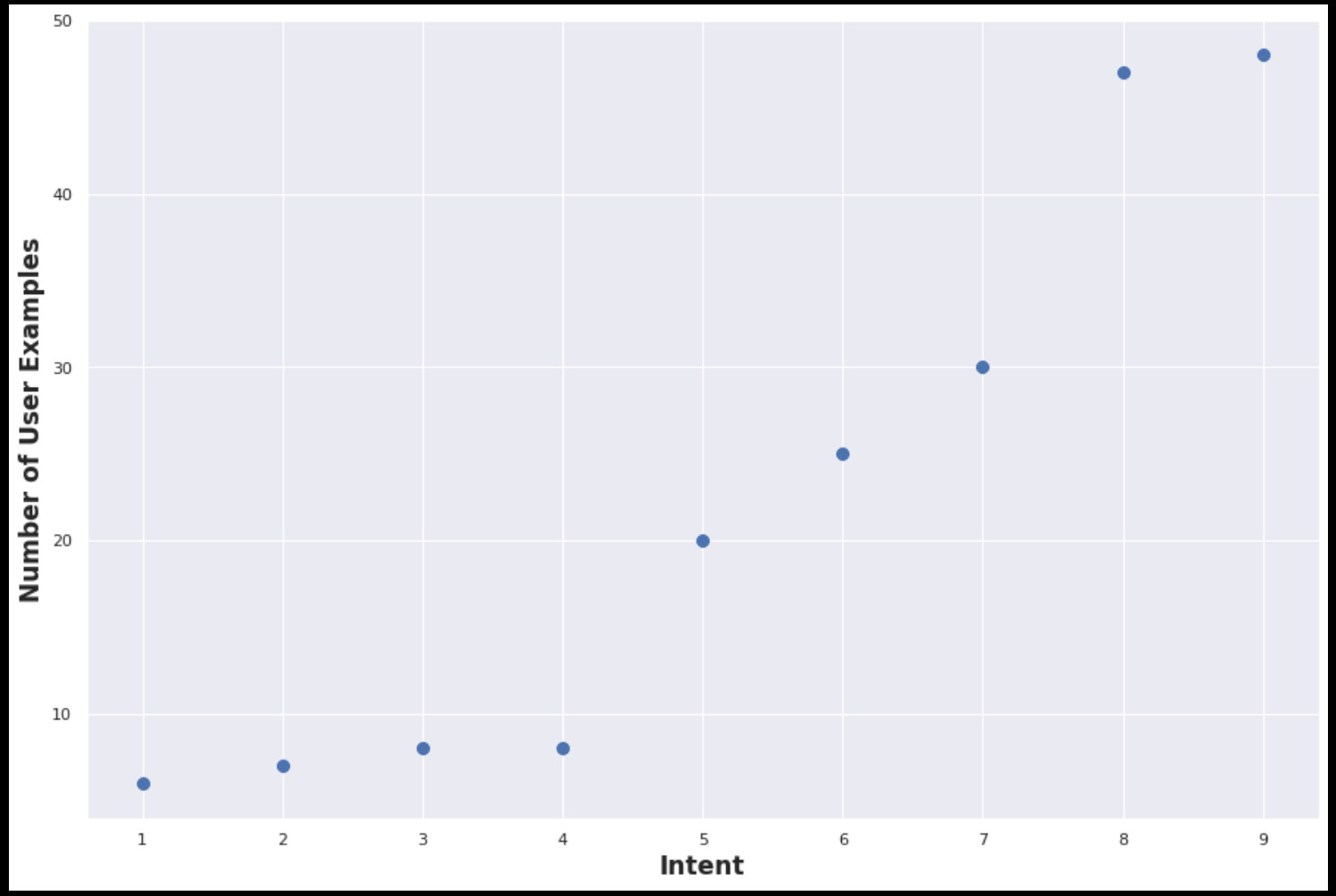
Advanced Analysis

Class Imbalance Analysis

Size largest Intent <= 2 * Smallest Intent ?

Sorted Distribution of User Examples per Intent

	Intent	Number of User Examples
1	Goodbye	6
2	Cancel	7
3	Help	8
4	Thanks	8
5	Customer_Care_Appointments	20
6	Customer_Care_Store_Location	25
7	General_Greetings	30
8	General_Connect_to_Agent	47
9	Customer_Care_Store_Hours	48



Advanced Analysis

Remediation for Class Imbalance

Size largest Intent <= 2 * Smallest Intent ?

Sorted Distribution of User Examples per Intent

	Intent	Number of User Examples
1	Goodbye	6
2	Cancel	7
3	Help	8
4	Thanks	8
5	Customer_Care_Appointments	20
6	Customer_Care_Store_Location	25
7	General_Greetings	30
8	General_Connect_to_Agent	47
9	Customer_Care_Store_Hours	48

Class imbalance will not always lead to lower accuracy. All intents (classes) thus need not have the same number of examples.

1. For intents like updateBankAccount and addNewAccountHolder where the semantics difference between them is subtler, the number of examples per intent needs to be somewhat balanced else the classifier might favor the intent with the higher number of examples.
2. For intents like greetings that are semantically distinct from other intents like updateBankAccount, it may be okay for it to have fewer examples per intent and still be easy for the intent detector to classify.

If during testing it seems like intent classification accuracy is lower than expected, we advise you to re-examine this distribution analysis.

With regard to sorted distribution of examples per intent, if the sorted number of user examples varies a lot across different intents, it can be a potential source of bias for intent detection. Large imbalances in general should be avoided. This can potentially lead to lower accuracy. If your graph displays this characteristic, this might be a source of error.

https://cloud.ibm.com/docs/assistant?topic=assistant-intent-recommendations#intent-recommendations-get-example-recommendations

Resources

https://github.com/watson-developer-cloud/assistant-improve-recommendations-notebook

https://dataplatform.cloud.ibm.com/exchange/public/entry/view/133dfc4cd1480bbe4eaa78d3f635e568
https://medium.com/ibm-watson/announcing-dialog-skill-analysis-for-watson-assistant-83cdfb968178?
https://github.com/watson-developer-cloud/assistant-improve-recommendations-notebook

https://github.com/jiportilla/assistant-dialog-skill-analysis
https://dataplatform.cloud.ibm.com/exchange/public/entry/view/4d77701840fcb2f21587e39fdb887049