

DefinedCrowd

**Challenges Around Collecting and Training
Language
and Multimodal Related Datasets to Support HCI**

*Stephen Rauch, VP of Product
srauch@definedcrowd.com*

Emotion AI Summit, Boston, MA

September 6, 2018

Industry 4.0 and Artificial Intelligence (AI) Revolution



Artificial Intelligence Applications



Call Centers



**Self-driven cars /
public transportation**



Shopping Attendants



**Medical diagnosis and
surgeries**



Border Control



**Education and learning
experiences**



Taking care of elderly

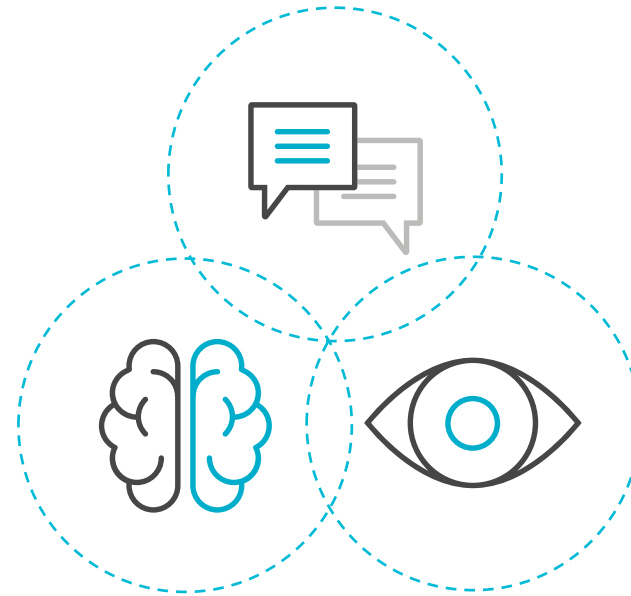


House Tasks

Artificial Intelligence: Mimicking the Human Brain

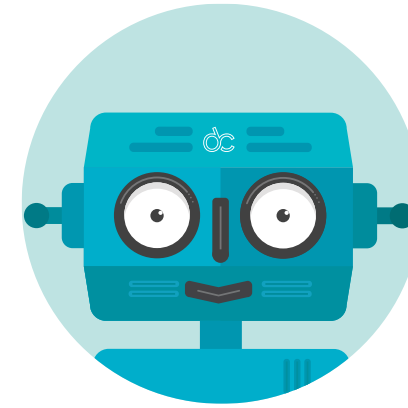


communication



reasoning

vision



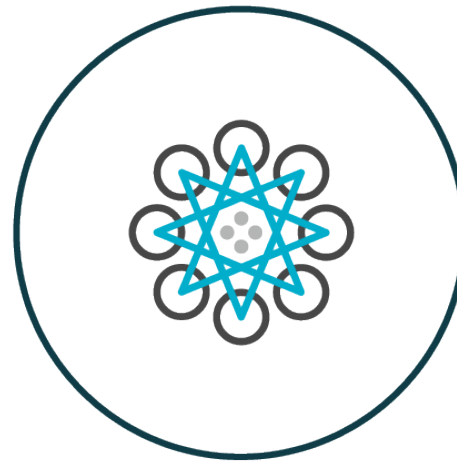
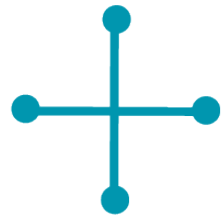
“It Takes a Village To Raise an AI. Just like a human child, every advanced AI system needs to be “trained” before it can interact properly and respond appropriately in different situations. Kids need parents, teachers, books and experiences. AI need data sets, machine learning systems and actual human beings to interact with.”

– How Crowds of Humans Are Making AI Systems Scary-Smart, Rob Salkowitz –
<https://futurism.media/how-crowds-of-humans-are-making-ai-systems-scary-smart>

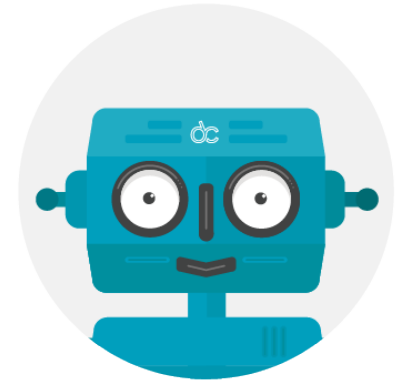
What does it take to build Smart AI Systems?



High Quality
Training Data

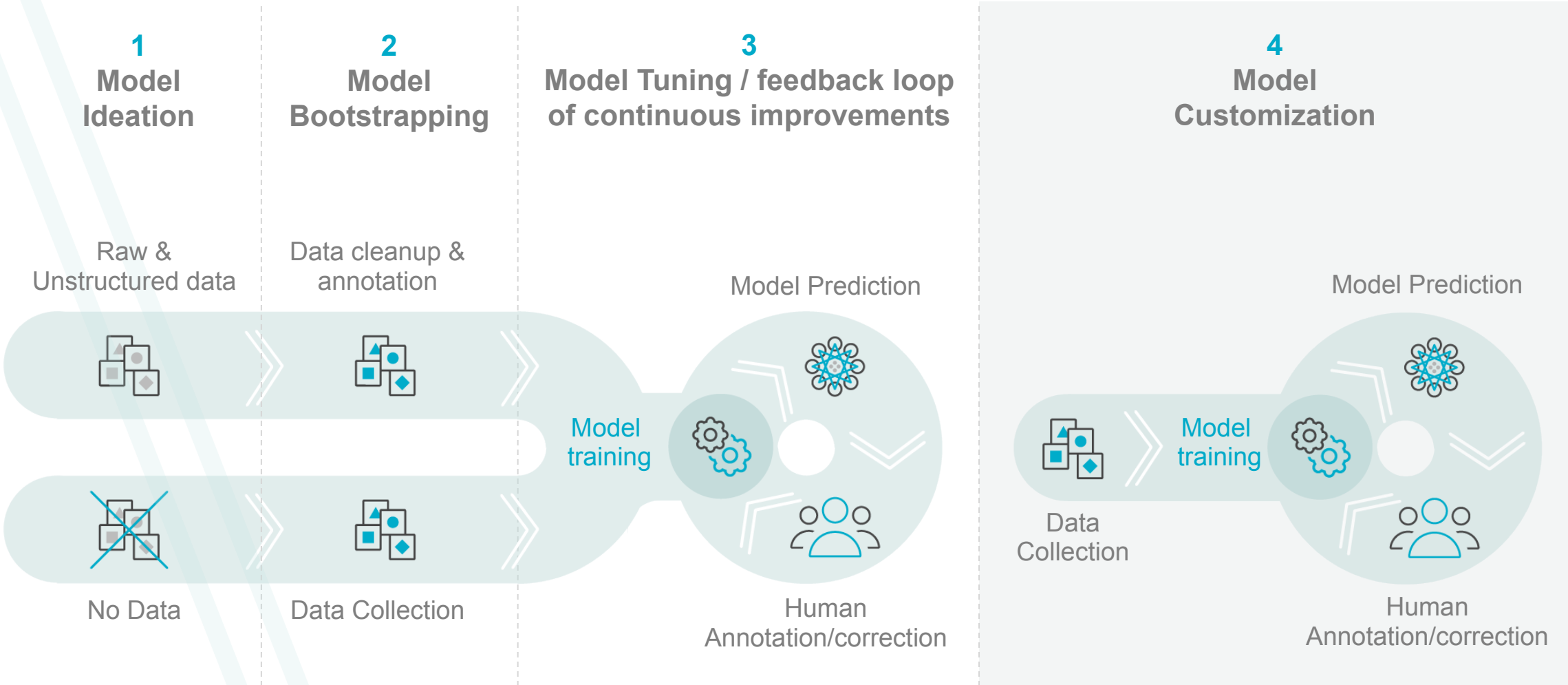


Machine Learning
Algorithms

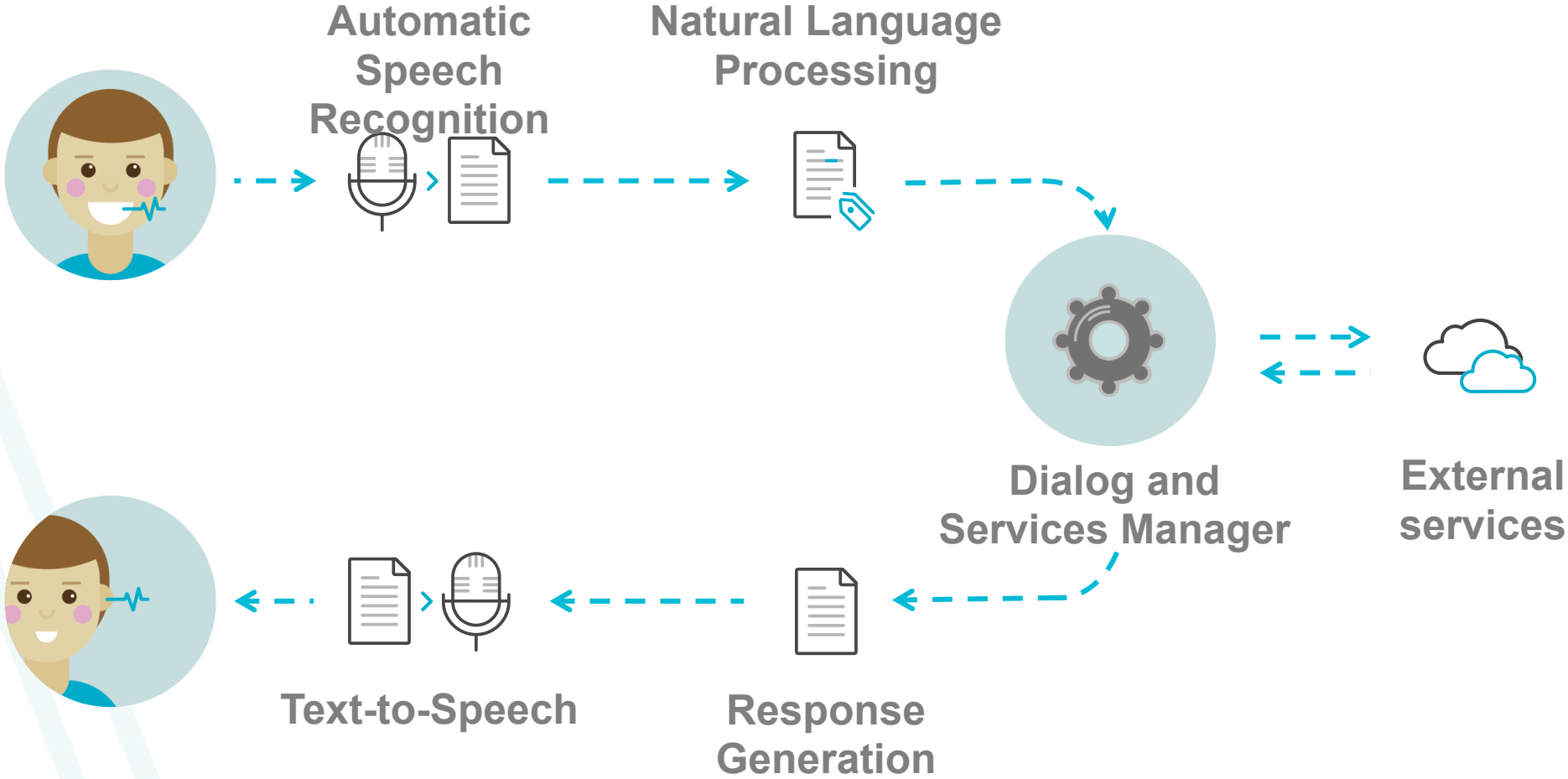


AI

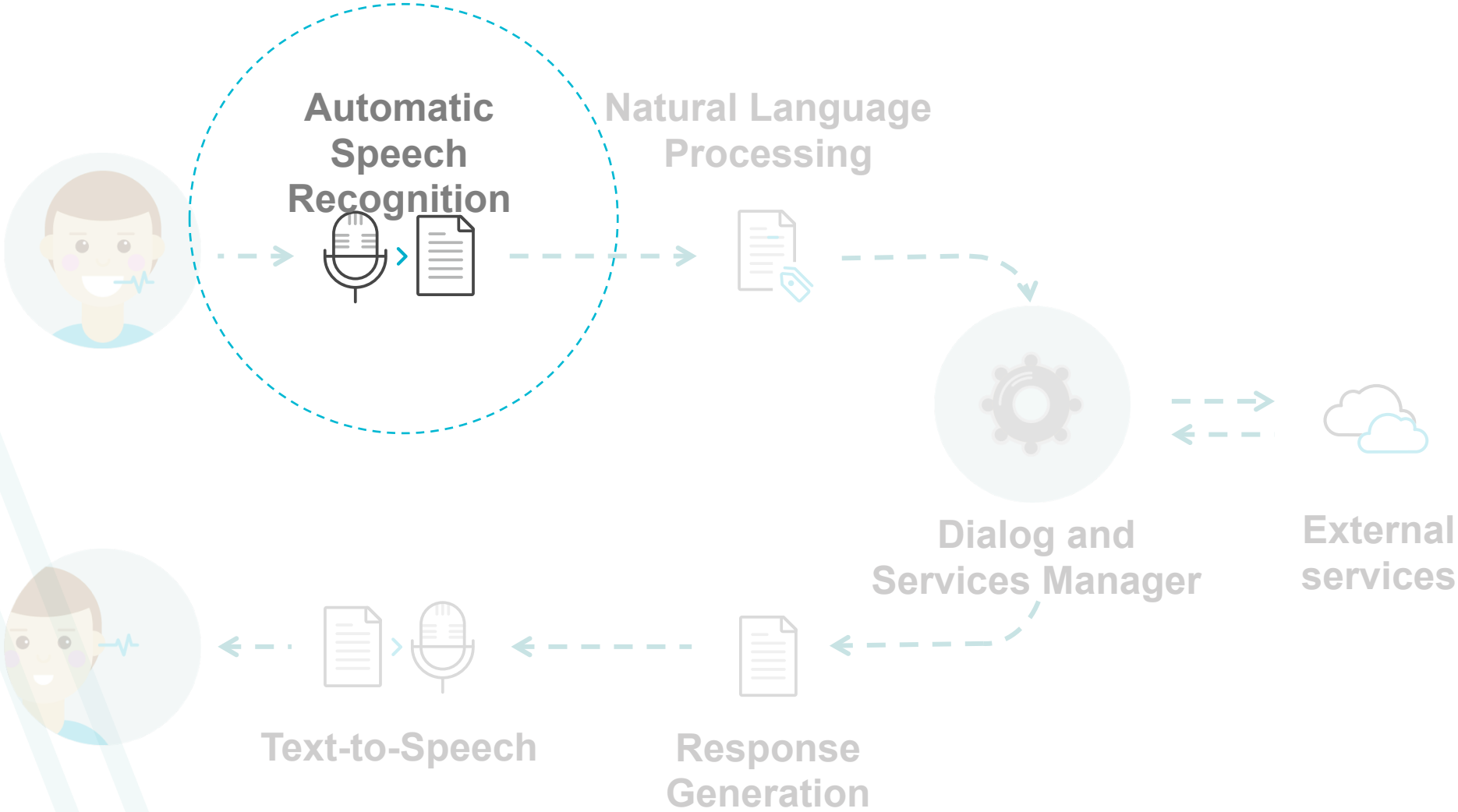
Four Stages in the Machine Learning Lifecycle



Example: Dialog System



Example: Dialog System



Use case:

IBM Client in Canada needs to transcribe Call Center conversations per regulation

Problem #1 Call Center data are narrowband (8 kHz, 8 bits)

Problem #2 There is no French Canadian ASR model available
but Watson has a French from France model

Preliminary Results

50% WER (Word Error Rate) in ASR (Automatic Speech Recognition) output

Word error rate can then be computed

as:

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}$$

Where:

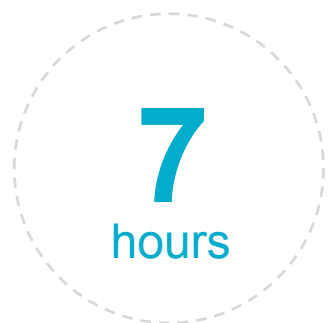
- S is the number of substitutions;
- D is the number of deletions;
- I is the number of insertions;
- C is the number of correct words;
- N is the number of words in the reference ($N=S+D+C$)

IBM Watson “Speech to Text” Service

The screenshot shows the IBM Watson Speech to Text API documentation page. The browser address bar displays the URL: `https://www.ibm.com/watson/developercloud/assistant/api/v1/node.html?node`. The page features a dark header with the 'Watson' logo and two buttons: 'Get Started Free' and 'Sign in to IBM Cloud'. A left sidebar contains a navigation menu with 'Speech to Text' highlighted, and sub-items including 'API reference', 'Introduction', 'API explorer', 'Authentication', 'Error handling', 'Data labels', 'Data collection', 'WebSockets', 'Models', 'List models', 'Get a model', 'Sessionless', 'Recognize audio', 'Sessions', and 'Create a session'. The main content area is titled 'Introduction' and contains the following text: 'The IBM® Speech to Text service provides an API that uses IBM's speech-recognition capabilities to produce transcripts of spoken audio. The service can transcribe speech from various languages and audio formats. In addition to basic transcription, the service can produce detailed information about many aspects of the audio. For most languages, the service supports two sampling rates, broadband and narrowband. It returns all JSON response content in the UTF-8 character set. For more information about the service, see the [IBM® Cloud documentation](#).' Below this is a section for 'API USAGE GUIDELINES' with two bullet points: 'Audio formats: The service accepts audio in many formats (MIME types). See [Audio formats](#).' and 'HTTP interfaces: The service provides three HTTP interfaces for speech recognition. The sessionless interface includes a single synchronous method. The session-based interface includes...'. On the right side, there are tabs for 'Curl', 'Node', 'Java', and 'Python'. The 'Curl' tab is active, showing the 'API Endpoint' as `https://stream.watsonplatform.net/speech-to-text/api`. A note below states: 'Services on IBM Cloud Dedicated might not use this endpoint. Check your endpoint URL by clicking the service instance on the Dashboard.'

Towards a “Custom” Acoustic Model

Data “Step by Step”



7
hours

of data collection

Step 1 - Refine Domain & Data Preparation

- Client Product Analysis for Customer Service
- Scenario Creation and Goal Setting

Step 2 - Scripts Generation

Step 3 - Voice Collection

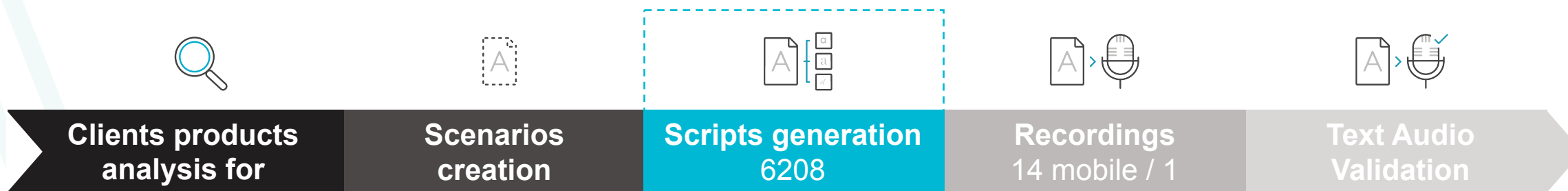
- Sourcing of (15) French Canadian speakers
- Recordings in mobile and desktop
(distribution of acoustic environments)

Step 4 – Data Validation

- Text Audio Validation

Canadian French Data Workflow

Crowd Sourced Scripts Generation



Imaginez que vous êtes en communication avec l'assistance téléphonique et que vous voulez savoir comment utiliser votre appareil, par exemple: "Comment dois-je procéder pour déverrouiller mon smartphone ?"

Écrivez trois phrases:

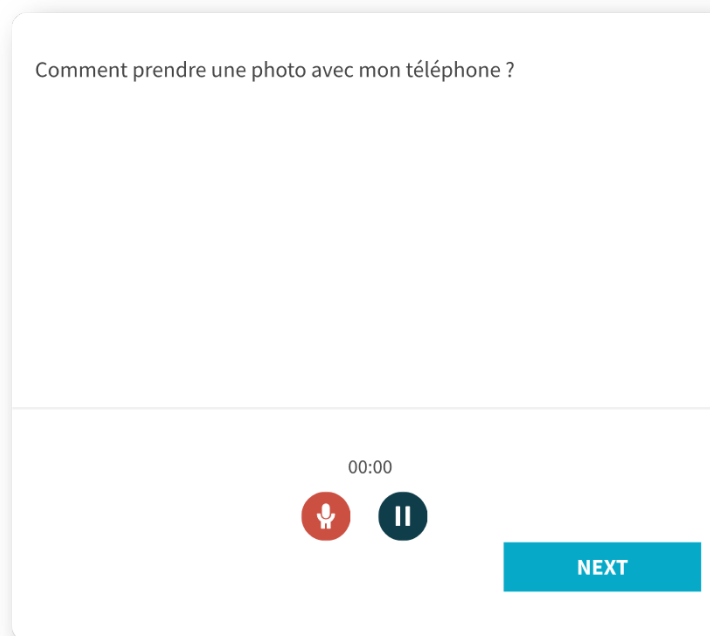
3. Comment puis-je me connecter à Internet ?
2. Comment prendre une photo avec mon téléphone ?
1. Pourriez-vous m'indiquer où insérer ma carte SIM ?

NEXT

***Sourcing French Canadian speakers**
15 contributors

Canadian French Data Workflow

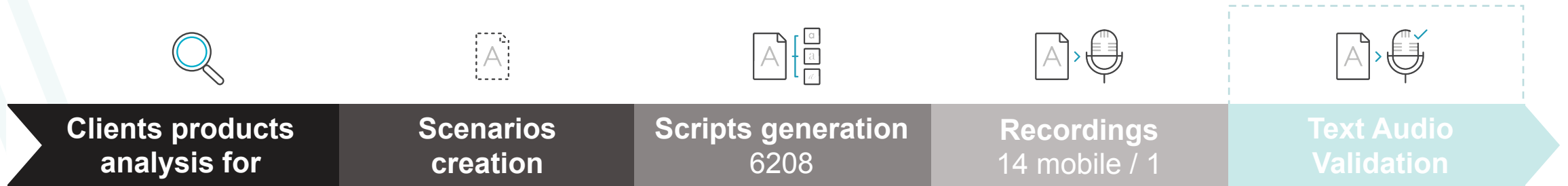
Crowd Sourced Recordings Collection



***Sourcing French Canadian speakers**
15 contributors

Canadian French Data Workflow

Crowd Sourced Text Audio Validation



Comment prendre une photo avec mon téléphone ?

▶ 0:00 / 0:07

Does the text match the audio?

Yes

No

***Sourcing French Canadian speakers**
15 contributors

IBM Watson “Speech to Text” Customization

The screenshot shows a web browser window displaying the IBM Watson developer cloud documentation. The URL in the address bar is <https://www.ibm.com/watson/developercloud/assistant/api/v1/node.html?node>. The page features a dark blue header with the IBM Cloud logo, navigation links for 'Catalog' and 'Docs', and 'Log in' and 'Sign up' buttons. A search bar contains the text 'speech-to-text x' and the placeholder 'Search documentation'. The main content area is titled 'The customization interface' and includes a 'Table of contents' sidebar on the right. The sidebar lists items such as 'Language model custo', 'Acoustic model custom', 'Using acoustic an customization tog', 'Language support', and 'Usage notes for custom'. A 'FEEDBACK' button is visible on the right side of the sidebar. The main text describes the customization interface and provides links to 'Input features', 'Output features', and 'Parameter summary'.

IBM Cloud Catalog Docs Log in Sign up

Demos Additional resources

HOW TO

- Making a recognition request
- Audio formats
- Input features
- Output features
- Parameter summary
- The WebSocket interface
- The HTTP REST interface
- The asynchronous HTTP interface
- The customization interface**
- Language model customization

speech-to-text x Search documentation

IBM Cloud Docs / Speech to Text

The customization interface

Last Updated: 2018-05-14 | [Edit in GitHub](#)

The Speech to Text service offers a customization interface that you can use to augment its speech recognition capabilities. You can use customization to improve the accuracy of speech recognition requests by customizing a base model for your domain and audio. Customization is available for only some languages and at different levels of support for different languages; see [Language support for customization](#).

Speech recognition works the same either with or without a custom model. When you use a custom model for speech recognition, you can use all of the input and output parameters that are normally available with a recognition request. For more information, see [Input features](#), [Output features](#), and the [Parameter summary](#).

Language model customization

The service was developed with a broad, general audience in mind. The service's base vocabulary contains many words

Table of contents

- Language model custo
- Acoustic model custom
- Using acoustic an customization tog
- Language support
- Usage notes for custom

FEEDBACK

IBM Watson “Speech to Text” Customization Results

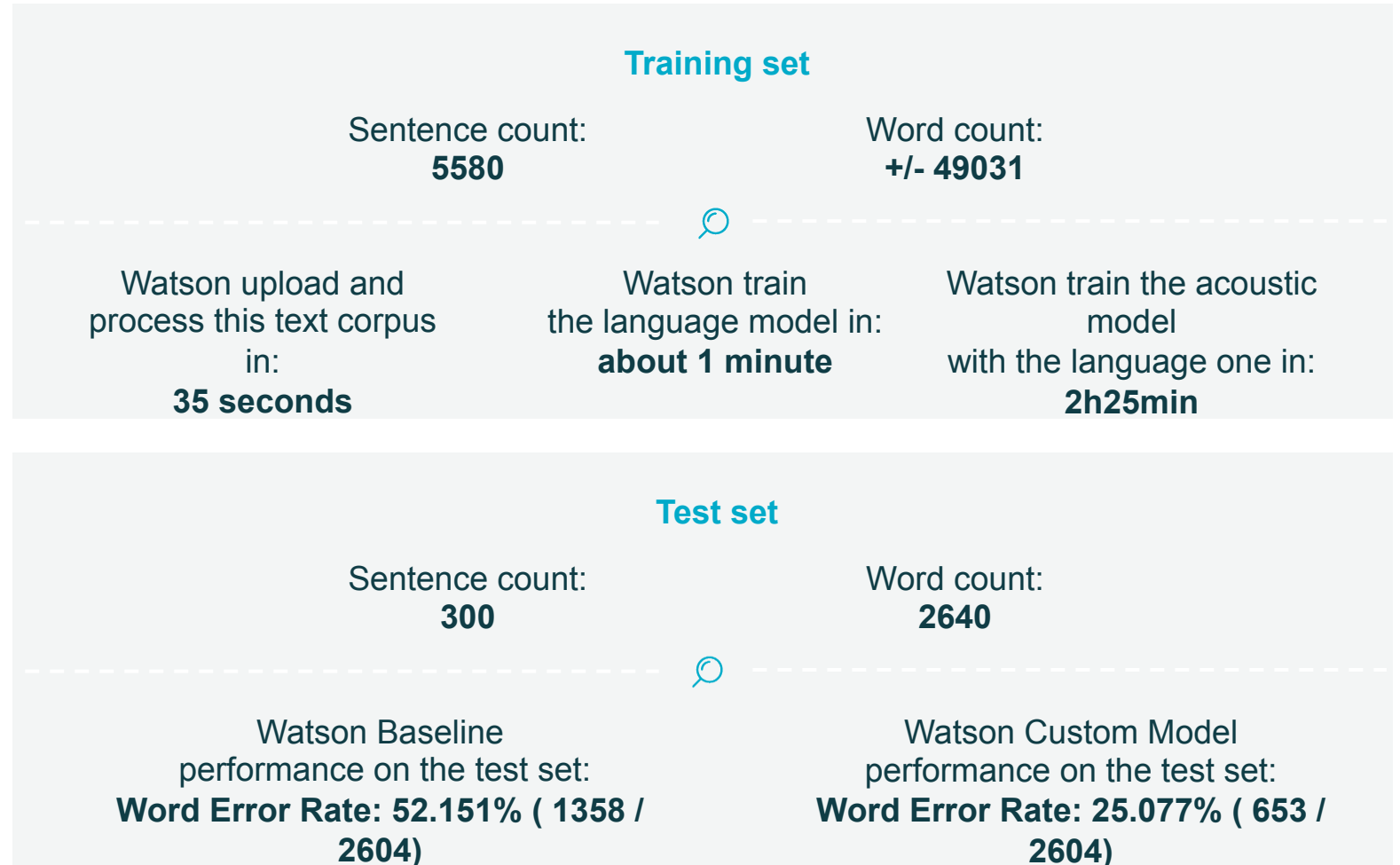


Total length of all the uploaded audio

358,08 mins

- roughly 6 hours -

Watson analyze and accept the audio as valid in:
roughly 6 minutes

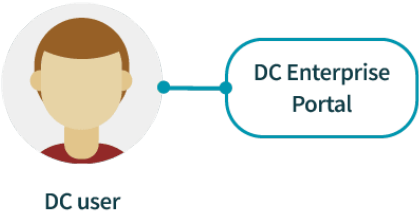
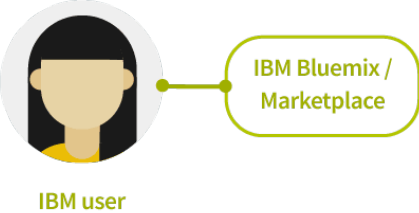


End-to-End “Text to Speech” Service – Scripted Speech

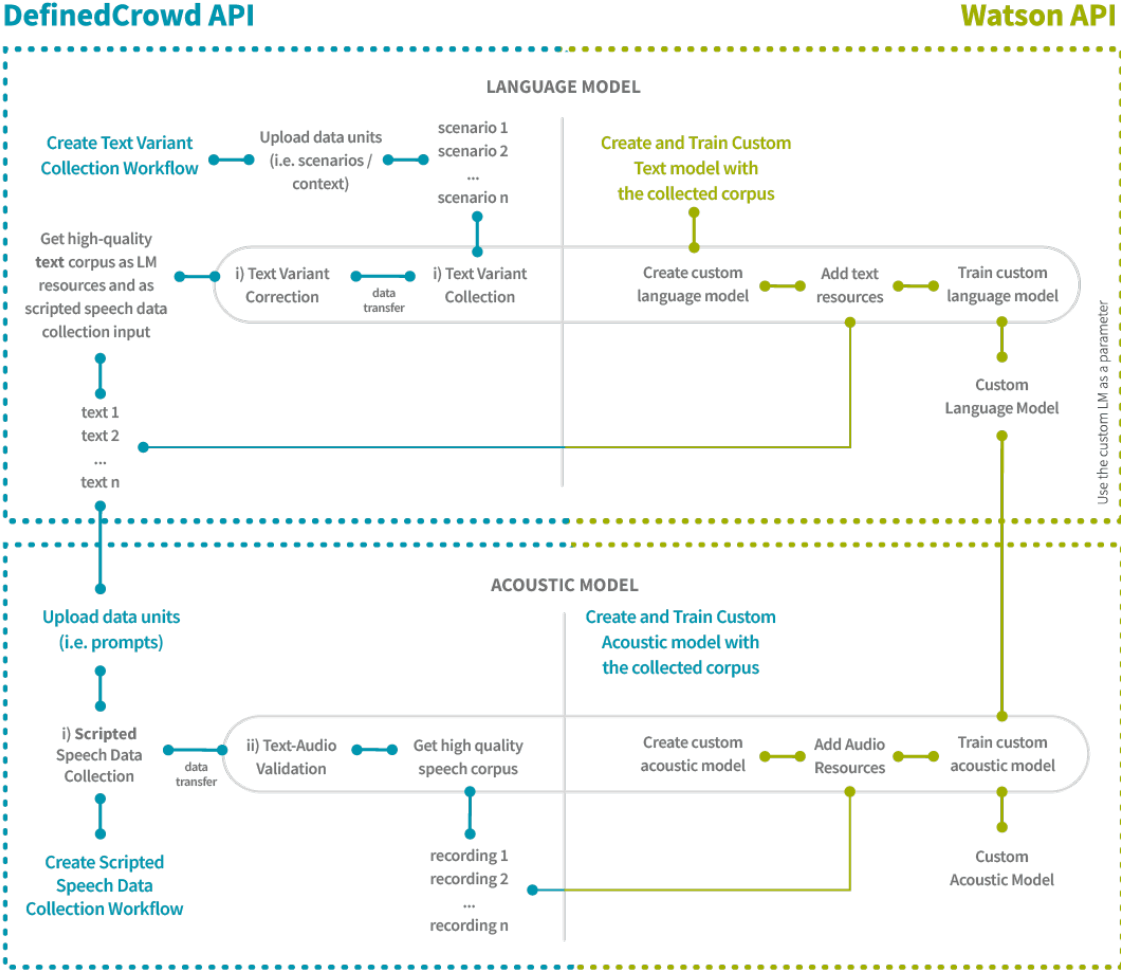
1 Access service
(via market place/bluemix catalog or DC enterprise portal)

2 Configure data and model settings

3 Get trained custom model



E2E Speech to Text Service



Summary: Challenges

Wide Domain



- domain too wide or unrelated to problem
- lack of up front test cases.



Poor Quality



- 15-20% of data is garbage



Slow Speed



- slow delivery time.
- No control over the process, often resulting in >6 months of redefining and adapting internal tools.



Hard to Scale



- new ML techniques require large amounts of training data that cannot be source in-house.
- cannot be source in-house In their market offer.
- not customer's core business.



Thank you

Request a trial on:

<https://enterprise.definedcrowd.com/en-us/account/requestdemo/>

Or email us at:

[**sales@definedcrowd.com**](mailto:sales@definedcrowd.com)

SEATTLE | LISBON | PORTO | TOKYO

