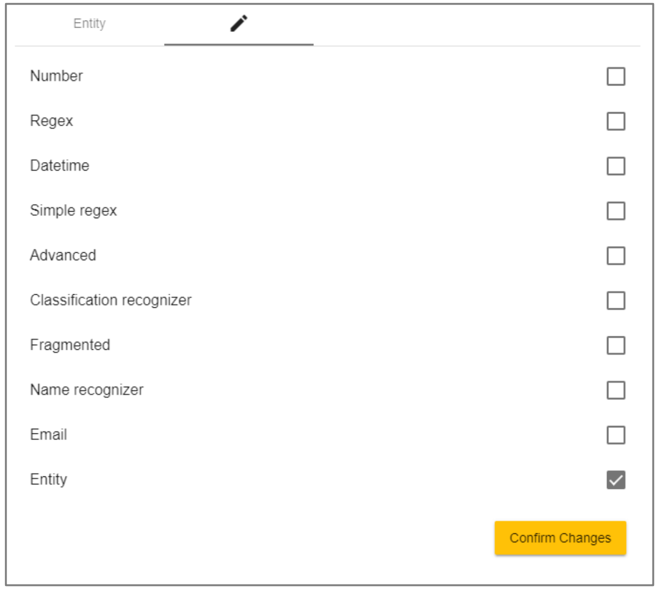
NATURAL LANGUAGE UNDERSTANDING WITH SAGA

INTRODUCTION

WHAT IS SAGA?

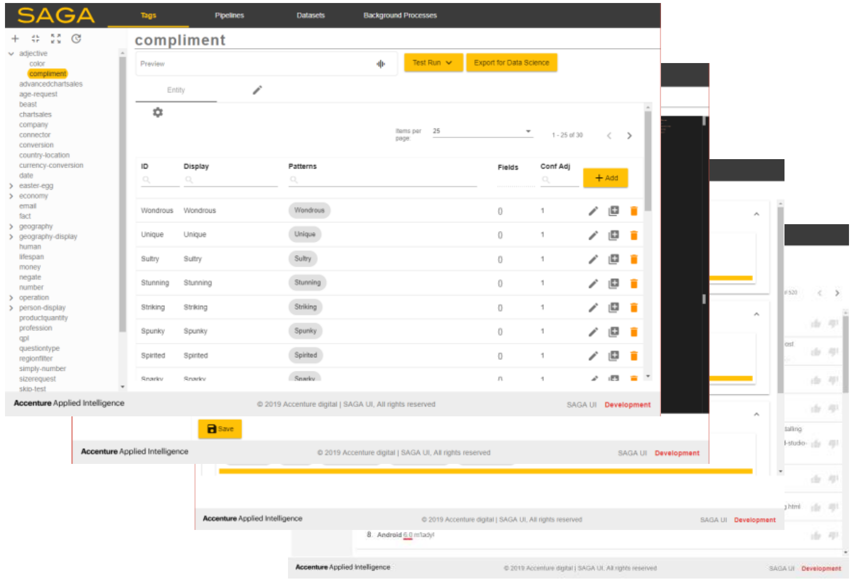
An Accenture asset for *maintainable & scalable* Natural Language Understanding



Entity

Number	□
Regex	□
Datetime	□
Simple regex	□
Advanced	□
Classification recognizer	□
Fragmented	□
Name recognizer	□
Email	□
Entity	☑

Confirm Changes



SAGA — Tags | Pipelines | Datasets | Background Processes

adjective
color
compliment
advancedchartsales
age-request
beast
chartsales
company
connector
conversion
country-location
currency-conversion
date
easter-egg
economy
email
fact
geography
geography-display
human
lifespan
money
negate
number
operation
person-display
productquantity
profession
qpl
questiontype
regionfilter
simply-number
sizerequest
skip-test

compliment

Preview | Test Run | Export for Data Science

Entity

Items per page: 25 | 1 - 25 of 30

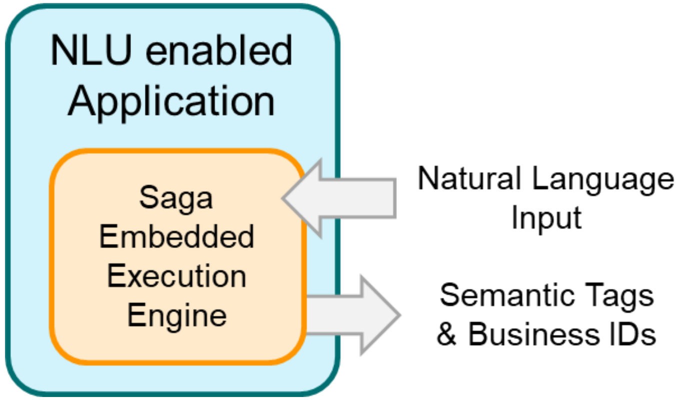
ID	Display	Patterns	Fields	Conf Adj
Wondrous	Wondrous	Wondrous	0	1
Unique	Unique	Unique	0	1
Sultry	Sultry	Sultry	0	1
Stunning	Stunning	Stunning	0	1
Striking	Striking	Striking	0	1
Spunky	Spunky	Spunky	0	1
Spirited	Spirited	Spirited	0	1
Sparky	Sparky	Sparky	0	1

+ Add

Accenture Applied Intelligence © 2019 Accenture digital | SAGA UI, All rights reserved SAGA UI Development

Save

Accenture Applied Intelligence © 2019 Accenture digital | SAGA UI, All rights reserved SAGA UI Development

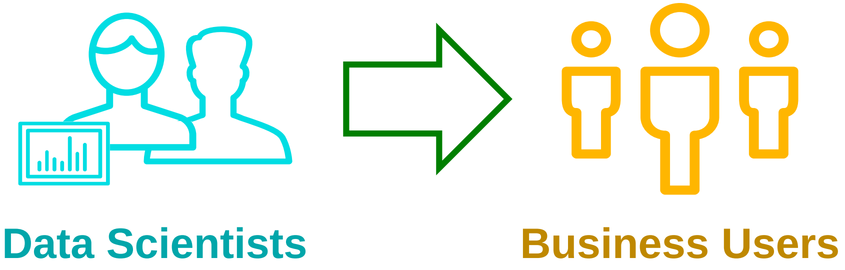


Pre-Built, Pre-Tested Language Algorithms

Business-Friendly UIs for Language Modeling

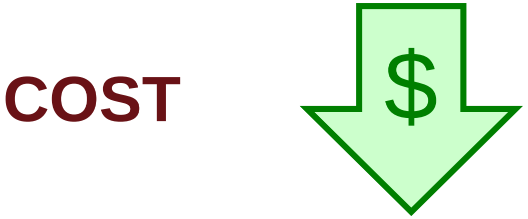
Easy Integration into Business Applications

SAGA – PRIMARY BUSINESS BENEFITS



New Language Models can be created by the business

- Import your own data
- Choose, configure, test & evaluate language models
- Especially good for complex, business or industry specific domains



Reduce cost

- Create and test new models much more quickly
- User interfaces help manage, coordinate & automate the process
- 90% of the work does not require programming

Maintain & Improve

- Designed for on-going maintenance
- Maintenance can be managed by the business

WHY AND HOW IS SAGA…

… scalable?

Megabytes / second of text processed

- *Embedded library for NLP execution*
  - *No client/server call for every transaction*

Can be run on-premises

No license limitations per application

- *Any amount of hardware*
- *Any amount of content*

Can handle large and complex language models

- *Builds the pipelines for you*
- *Designed to handle multiple models from multiple teams applied to the same content*
- *Dictionaries & advanced patterns scaled to millions of entries*

… maintainable?

Business-Friendly User Interfaces

- *Dictionary & Pattern Maintenance*
- *Easy-to-use search & markup interface*
- *Interfaces for manual training [FUTURE]*

Built-In Testing and Evaluation

- *ML Training & Evaluation*
- *Imports and Manages test & training datasets*
- *Automated retraining and retesting when language dependencies change [FUTURE]*

Pre-Built Language Models

Weakly Supervised Training

It manages the algorithms & resource data for you

WHAT IT DOES AND DOES NOT DO

Saga does well

Text Extraction

Semantic Tagging

- *Large pattern extraction (phrases, clauses)*

Text Classification (sentences, paragraphs sections)

Ambiguity Resolution

- *Multiple, ambiguous models can be applied to the same text*
- *A built-in confidence model allows for choosing the most likely interpretation*

Tagging to Business Objects & Business ID's

- *Saga provides a standard import format to ingest taxonomy & entity lists*

Extraction of Knowledge Graph Relationships

When there is a lack of Training Data

When data is too small for Machine Learning

Saga does *not* do

Unsupervised Clustering [possible future extension]

- *Recommend: Do this with post-processing / external analysis*

Ingestion, Document Processing

- *Recommend: Use Aspire or other ingestion / data prep s/w*

Post-Processing Business Rules

- *Recommend: Implement post-processing in the application using Saga standard outputs*

Data Science, inventing or testing brand-new algorithms

- *Saga has 'export for data science' for this purpose*

End-To-End NLP Application

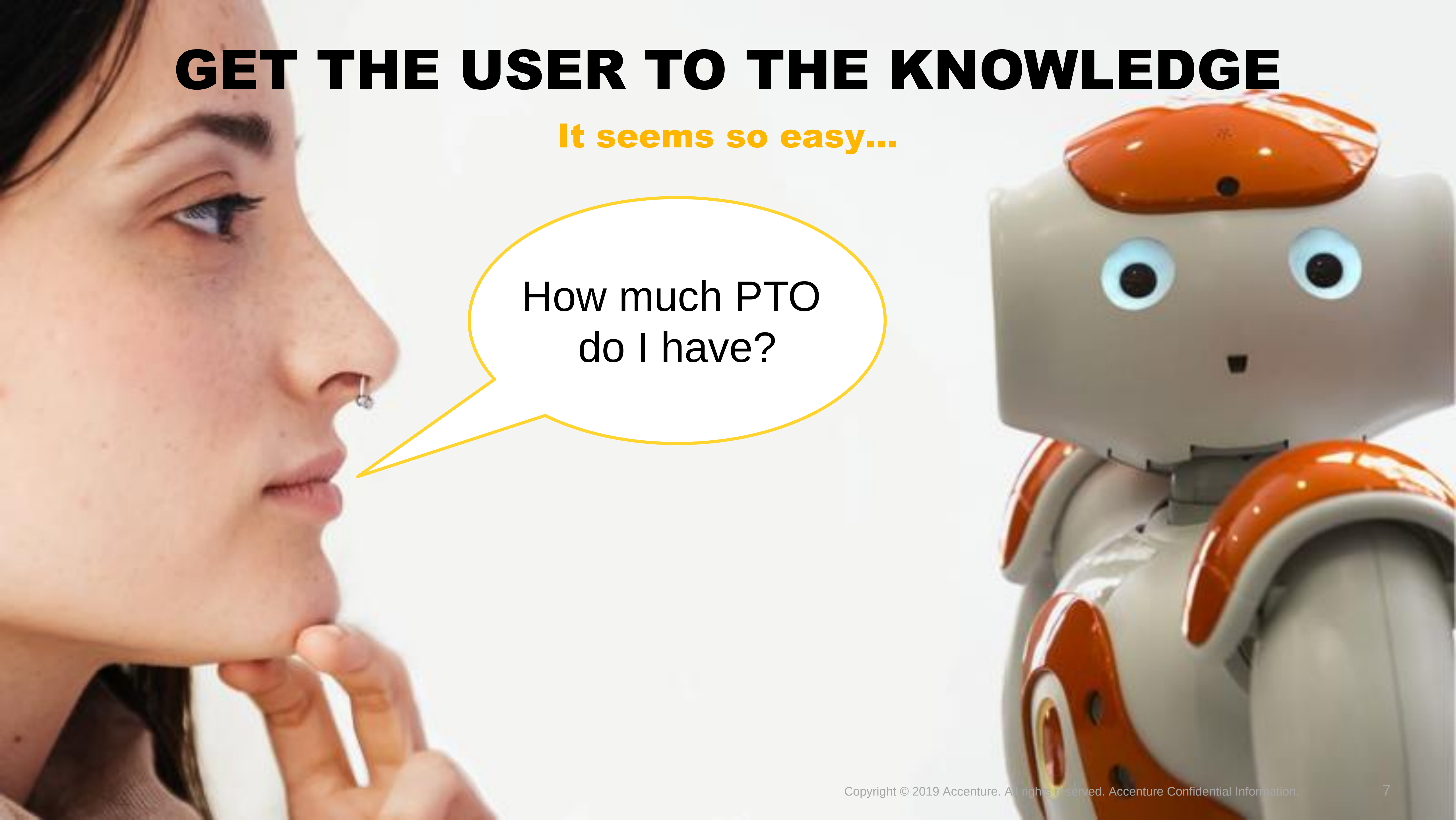
Chatbot Dialog Flow

NATURAL LANGUAGE UNDERSTANDING WITH SAGA

APPLICATIONS

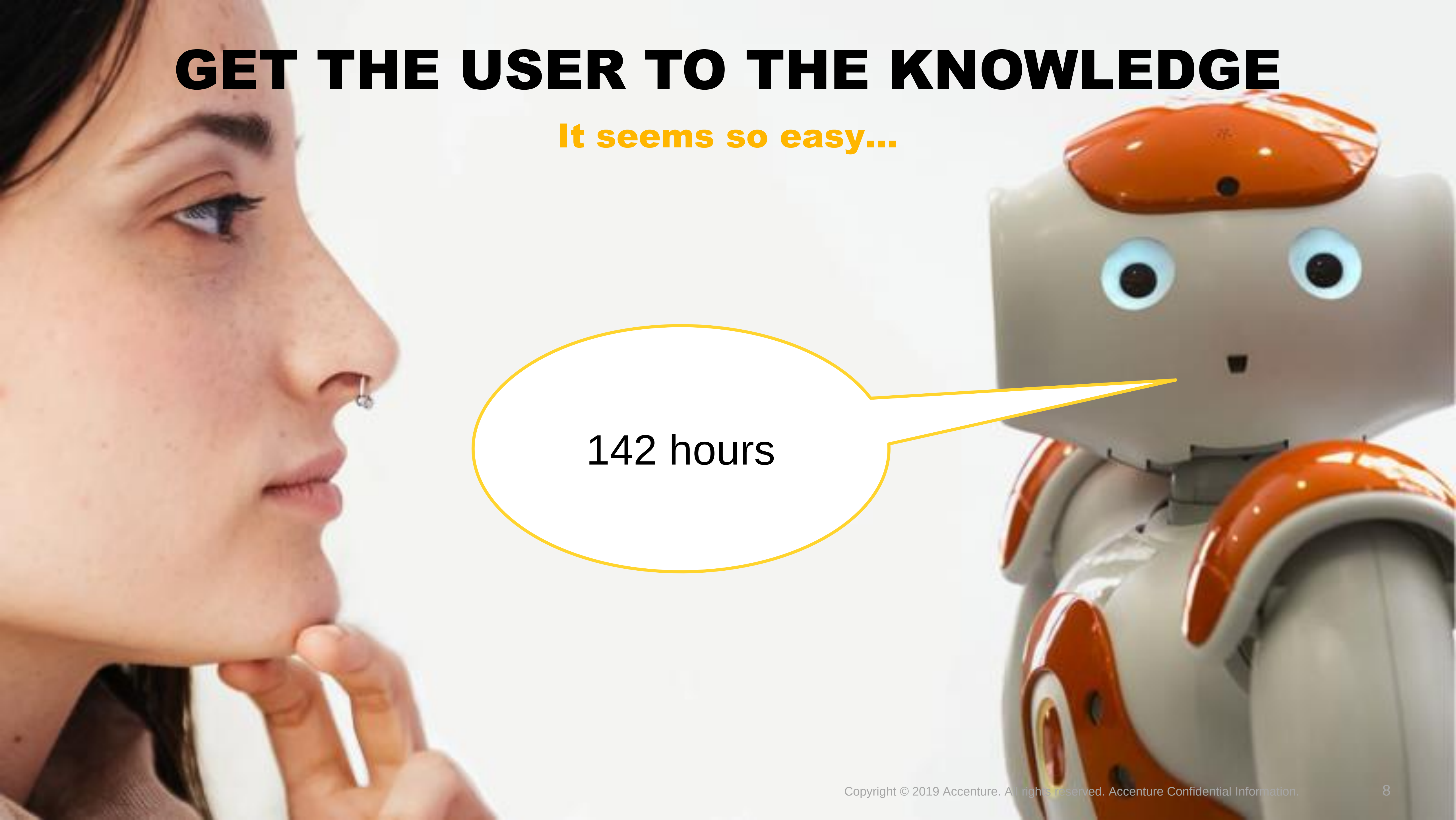
GET THE USER TO THE KNOWLEDGE

It seems so easy...



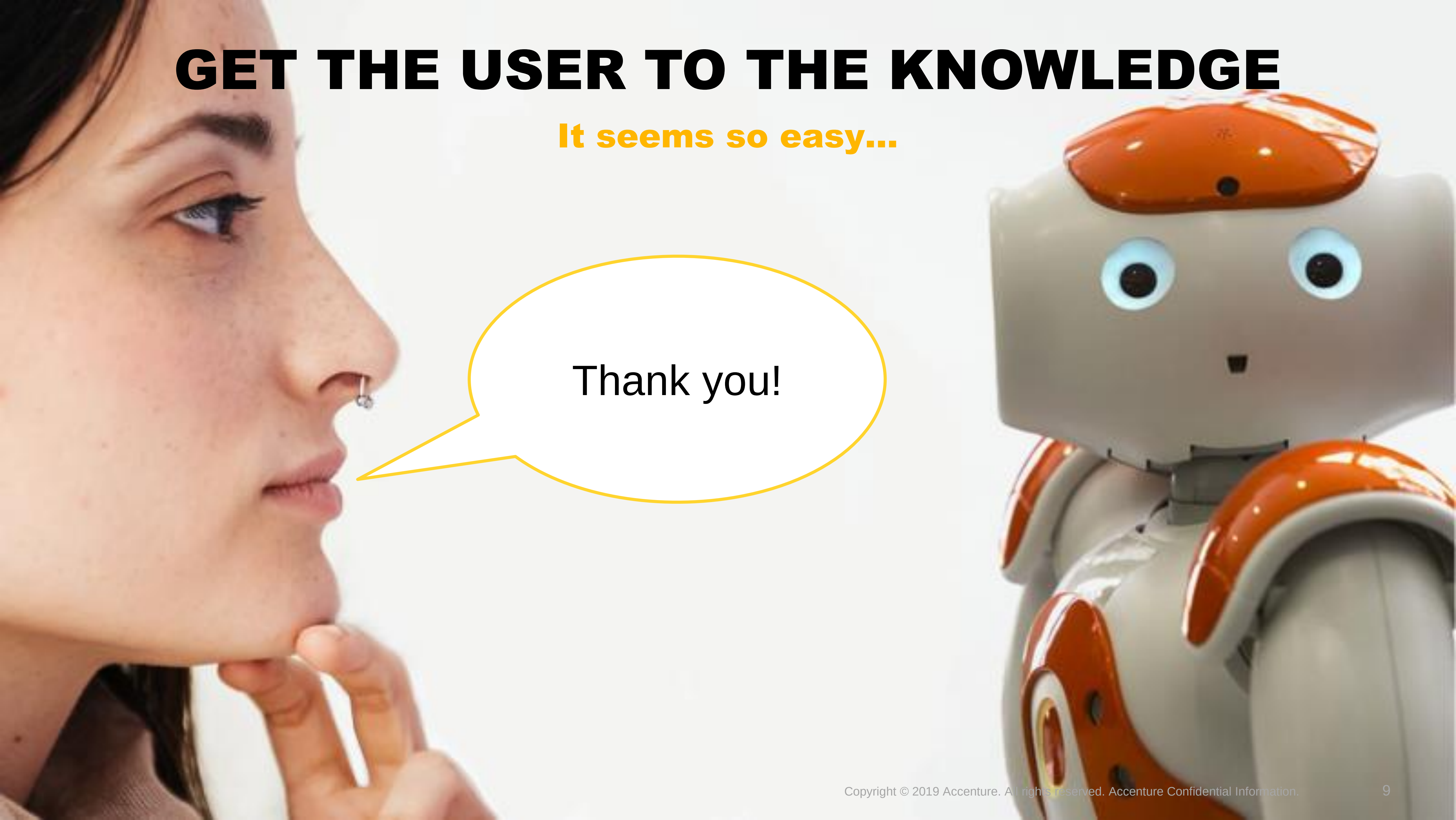
GET THE USER TO THE KNOWLEDGE

It seems so easy…



GET THE USER TO THE KNOWLEDGE

It seems so easy…



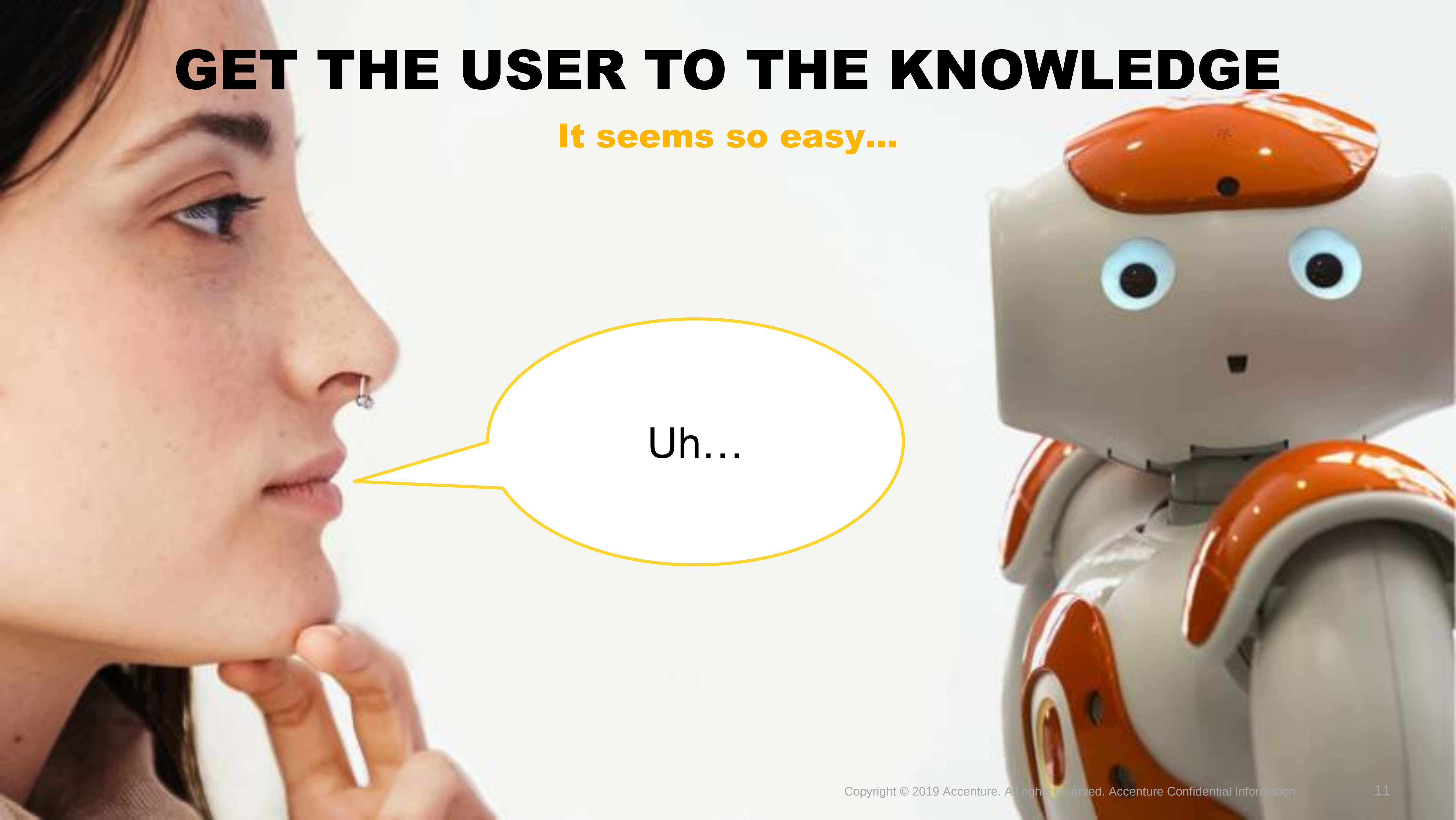
Thank you!

GET THE USER TO THE KNOWLEDGE

It seems so easy...

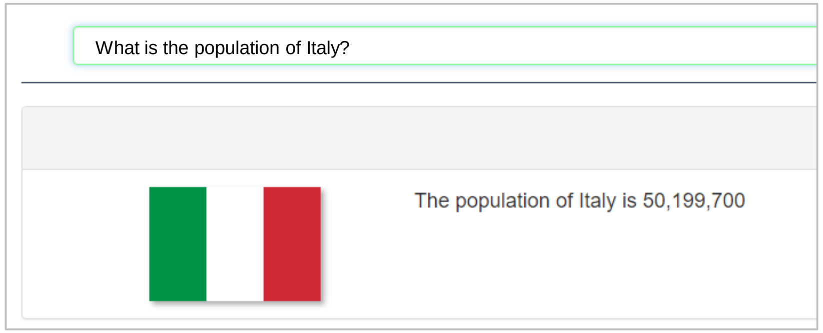
GET THE USER TO THE KNOWLEDGE

It seems so easy…



SAGA APPLICATIONS

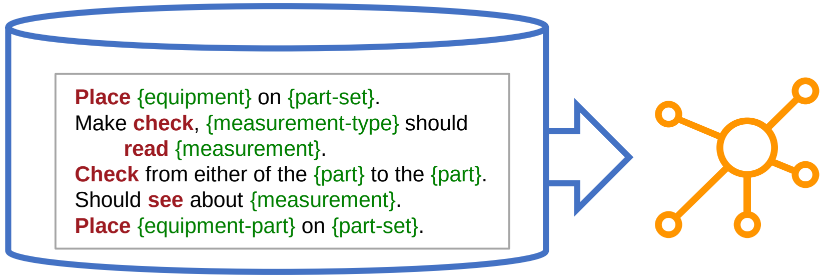
Question / Answer



What is the population of Italy?

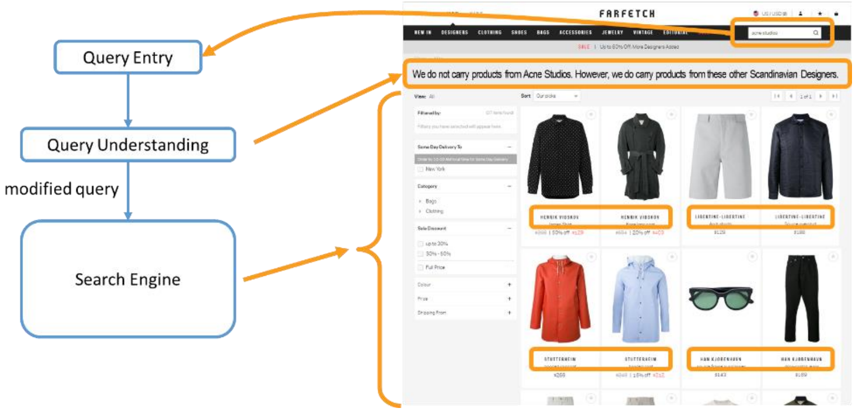
The population of Italy is 50,199,700

Building Knowledge Graphs



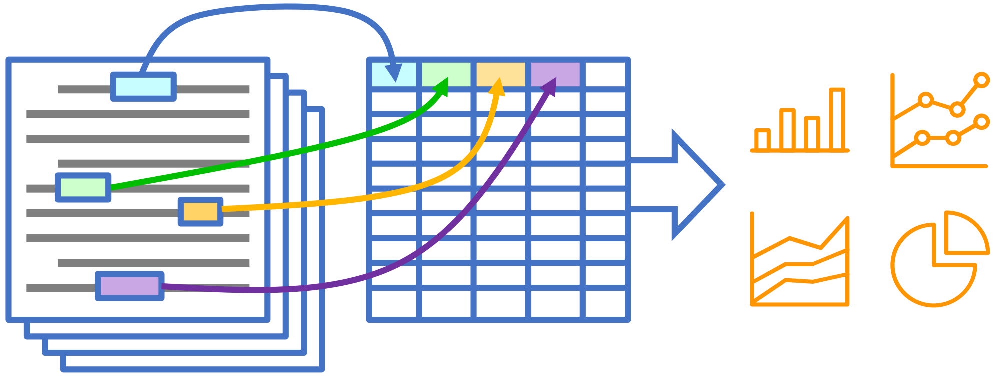
Place {equipment} on {part-set}.
Make **check**, {measurement-type} should **read** {measurement}.
Check from either of the {part} to the {part}.
Should **see** about {measurement}.
Place {equipment-part} on {part-set}.

Semantic Search



We do not carry products from Acne Studios. However, we do carry products from these other Scandinavian Designers.

Analytics on Unstructured Content



LARGE EUROPEAN BANK

Natural Language Data Analytics

- Business users unable to accurately locate and extract critical business data from Data Lake for self-service analytics.
- Custom self-service analytics never get done because people don’t know how to get the data and don’t have time to learn it. The “cognitive load” is too high.
- Ambiguity in requests (e.g. ‘Barcelona’ is both a city and a province) require multiple lookups and knowledge of data peculiarities.

Self-service data requests In Natural Language

Ingest and identification of of business entities

Resolution of entities to business IDs

Output mapped to knowledge graph of business data

ACCENTURE SOLUTION

Saga for NLP / NLU coupled with a knowledge graph of business data.

Saga identifies the user intent and identifies business categories and entities and resolves them to actionable business IDs.

Saga also identifies areas of ambiguity and available alternatives.

Post-processing chatbot asks for user help to resolve ambiguity.

Solution then leverages knowledge graph to create appropriate SQL, verifies SQL with user and then delivers the data for self-service analytics.

OIL AND GAS SUPPLIER

Daily Drilling Reports

Daily Drilling Reports (DDRs) contain a summary of information about what happens every day when drilling a well.

- Mud Loss (and how much), stuck pipe, equipment used, soil composition tests, angle of drill, depth, mud pressure, equipment dropped in the hole

This information is unstructured text and therefore not-usable for standard predictive analytics. Facts, metrics and entities must be extracted from this content (using NLP) and normalized before it can be used.

Extract Drilling Metrics and Entities from DDRs

Analytics to guide drilling operations

Automatically Extract Best Practices & correlate to outcomes

Identify potential problems before they occur

ACCENTURE SOLUTION

Saga, a light-weight NLP/NLU Library coupled with machine learning is used to identify critical drilling problems.

Extraction of equipment and metadata will allow for best practices to be identified and correlated to outcomes.

Drilling procedures and behavior can be compared across rigs and drilling teams. Automated suggestions for improvements from well-to-well comparisons can be provided.

RECRUITING

Automatically Match Jobs to Candidates

- Large recruiting companies need to quickly fill candidates for new jobs.
- Candidates must be filled within 4 hours.
- Recruiters are typically entry-level college graduates with little real-world experience.
- Recruiters are not search or candidate sourcing experts.
- Even the largest recruiting companies only fill a small percentage (5%) of the jobs they get – so there is no lack of opportunity.

Automatically process job descriptions & résumés / CVs

Semantic analysis on jobs and skills, freshess, experience

Increase fill rate by 6%

Reduce time to fill by 25%

ACCENTURE SOLUTION

Ingest and process jobs and résumés (CVs) with NLP / NLU processing to determine skills, job titles, companies, overall capabilities, legal requirements, education, skill freshness, skill experience, etc.

Create matching algorithms to automatically recommend jobs for candidates, candidates for jobs as well as finding similar candidates and jobs.

Use past hiring information to perform machine learning and to test and score and continuously improve the algorithms to optimize accuracy.

Use NLP to extract and handle complex "reports to" and "managed by" relationships.

Use NLP to extract and handle legally-binding requirements.

CONSUMER ELECTRONICS

Automatically Answer Customer Support IMs

- Large consumer electronics firm receives support questions over a very large number of intents (1000+)
- This can include data such as device type, software app, etc.
- The input is large and extremely dirty
- The customer does not have training data
- Chatbots are not scaled to the volume and variety of this input

Explore current logs for language modeling

Normalize language so it can be processed

Machine learning to determine intents

NLP extraction to extract key metadata

ACCENTURE SOLUTION

Use assisted training to identify patterns which indicate the desired intents.

Identify key terms and phrases which indicate each intent.

Refine process with both automatic and manual steps to scale to a large number of intents.

Use multi-model ambiguity resolution to combine all models together.

Extraction of key entities (products, features, applications, services, etc.) to aid in classification and answer handling.

FUTURE: Use immediate robot feedback to help control interaction.

BUSINESS INSURANCE

Learn More About Customers for Accurate Pricing

Commercial insurance rates for business customers is based on the customer’s industry.

- High-risk industries will have higher insurance rates than low-risk industries.

Most of the businesses are small businesses that may only be represented by a web site or Facebook page.

The process for determining the industry for a customer requires manual research and is prone to error.

Use Aspire to fetch Customer data from the internet

Normalize language so it can be processed

Extract key indicators from unstructured content

Machine learning to determine SIC code

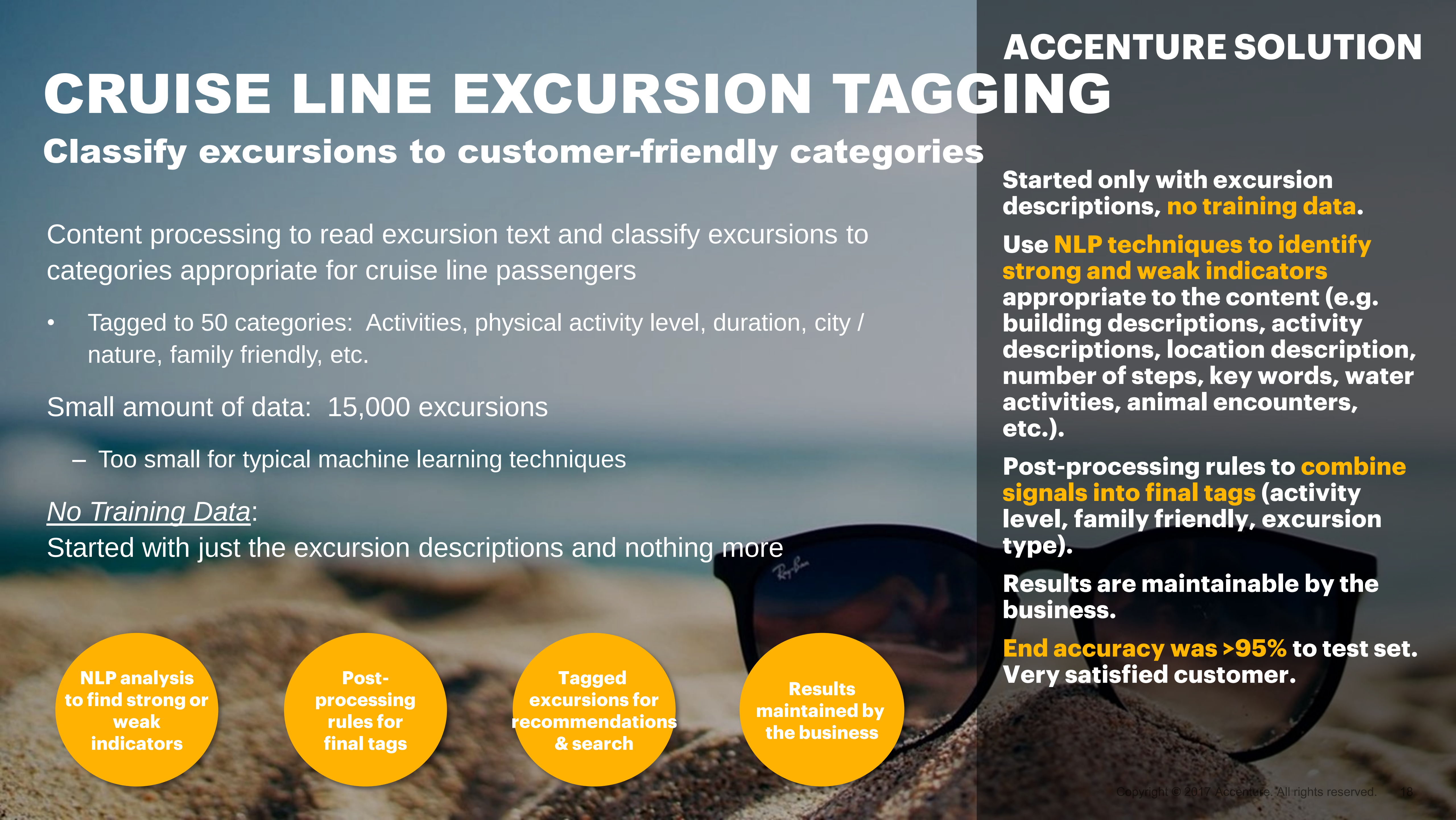
ACCENTURE SOLUTION

Use Aspire to download the company’s web site and Facebook pages from the internet.

Use Saga to cleanse the text and extract key industry indicators (e.g. retail vs distributor vs manufacturer).

High-quality text processing is required to achieve 90+ accuracy rates.

Use machine learning to classify businesses to any of 3000 industry codes. Use both specific (context and syntax sensitive) classification rules along with machine learning (general understanding) rules.



CRUISE LINE EXCURSION TAGGING

Classify excursions to customer-friendly categories

Content processing to read excursion text and classify excursions to categories appropriate for cruise line passengers

- Tagged to 50 categories: Activities, physical activity level, duration, city / nature, family friendly, etc.

Small amount of data: 15,000 excursions

- Too small for typical machine learning techniques

No Training Data:
Started with just the excursion descriptions and nothing more

NLP analysis to find strong or weak indicators

Post-processing rules for final tags

Tagged excursions for recommendations & search

Results maintained by the business

ACCENTURE SOLUTION

Started only with excursion descriptions, no training data.

Use NLP techniques to identify strong and weak indicators appropriate to the content (e.g. building descriptions, activity descriptions, location description, number of steps, key words, water activities, animal encounters, etc.).

Post-processing rules to combine signals into final tags (activity level, family friendly, excursion type).

Results are maintainable by the business.

End accuracy was >95% to test set. Very satisfied customer.

MORE USE CASES

Just the ones we've encountered so far

eCommerce – *Increase sales*

- Intelligent, targeted response for queries

Pharmacovigilance – *Reduce / eliminate manual effort*

- Extract entities from ADRs

Customer Support – *Decrease cost, opportunity for upsell*

- Automatically process many requests

News Analysis – *Identify bus. opportunities quickly, Increase revenue*

Vendor Contract Analysis – *Increase revenue*

- Identify vendors who are in breech of contract

Lien / Loan Contract Fact Extraction – *Replace manual process*

- Extract Loan information for marketing and analytics

Key Use Cases

New business Insights

Leverage unstructured content for analysis

Extract machine-readable knowledge from unstructured content

Improve human-computer interface for mobile employees

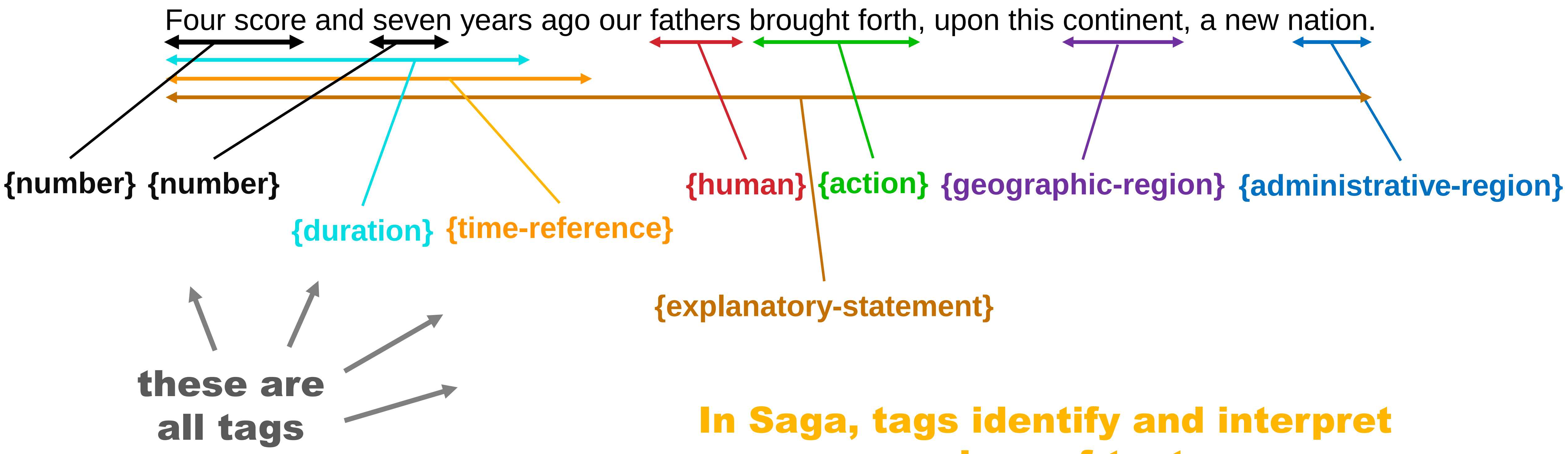
Learn more about your customers

NATURAL LANGUAGE UNDERSTANDING WITH SAGA

CONCEPTS & TERMINOLOGY

IT’S ALL ABOUT THE TAGS…

Semantic tags are the Organizing Structure for all of Saga



Four score and seven years ago our fathers brought forth, upon this continent, a new nation.

{number} {number}

{duration} {time-reference}

{human} {action} {geographic-region} {administrative-region}

{explanatory-statement}

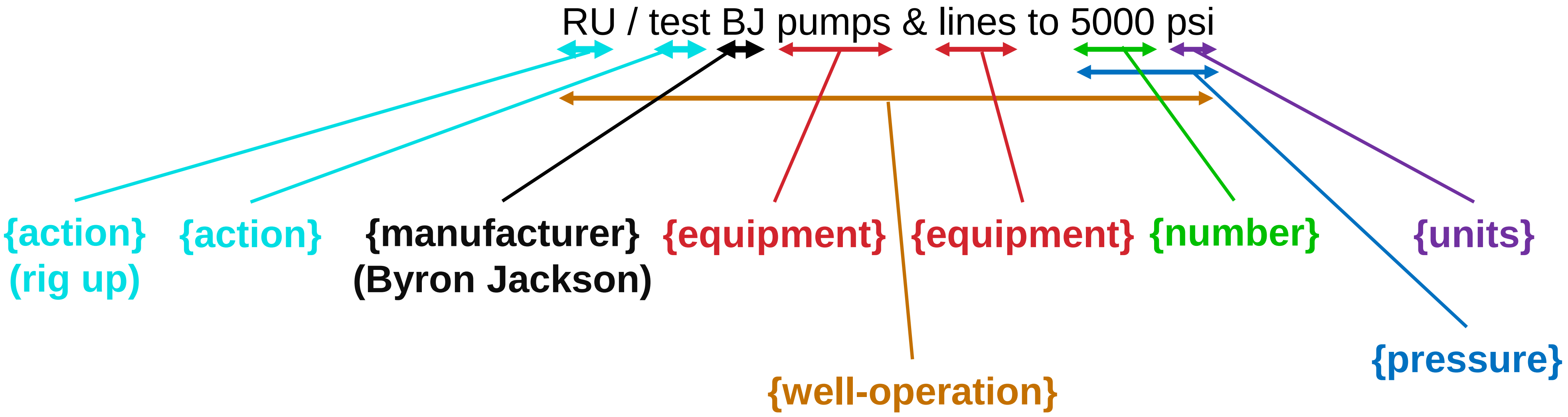
these are all tags

In Saga, tags identify and interpret regions of text.

By convention, tags are shown in {curly-braces}

TAGS ARE OFTEN APPLICATION SPECIFIC

Example from a daily drilling report

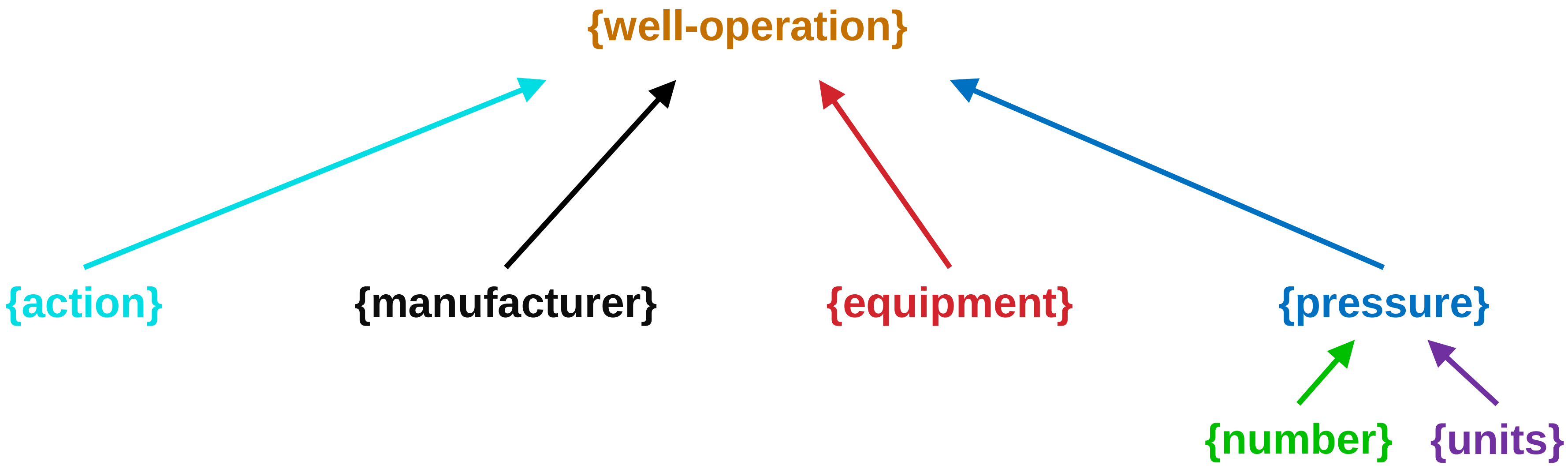


Different domains have different language and different meanings

Saga is designed to create new tags for domain-specific text

SAGA TRACKS DEPENDENCIES

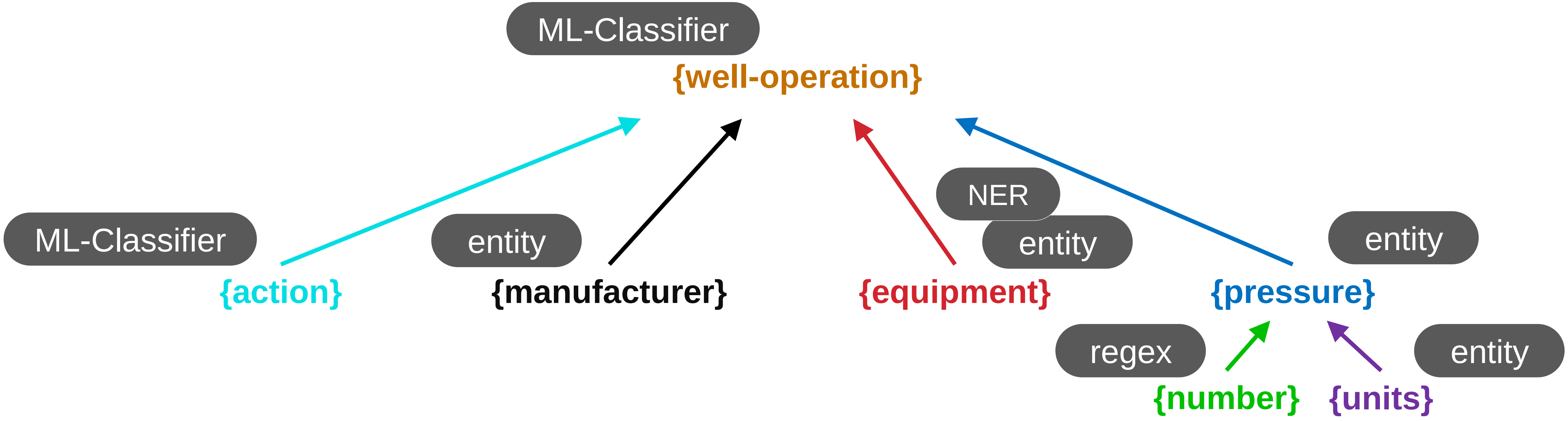
Understanding is built up from the bottom



Dependencies are configured as part of language modeling.

TAGS ARE TIED TO "RECOGNIZERS"

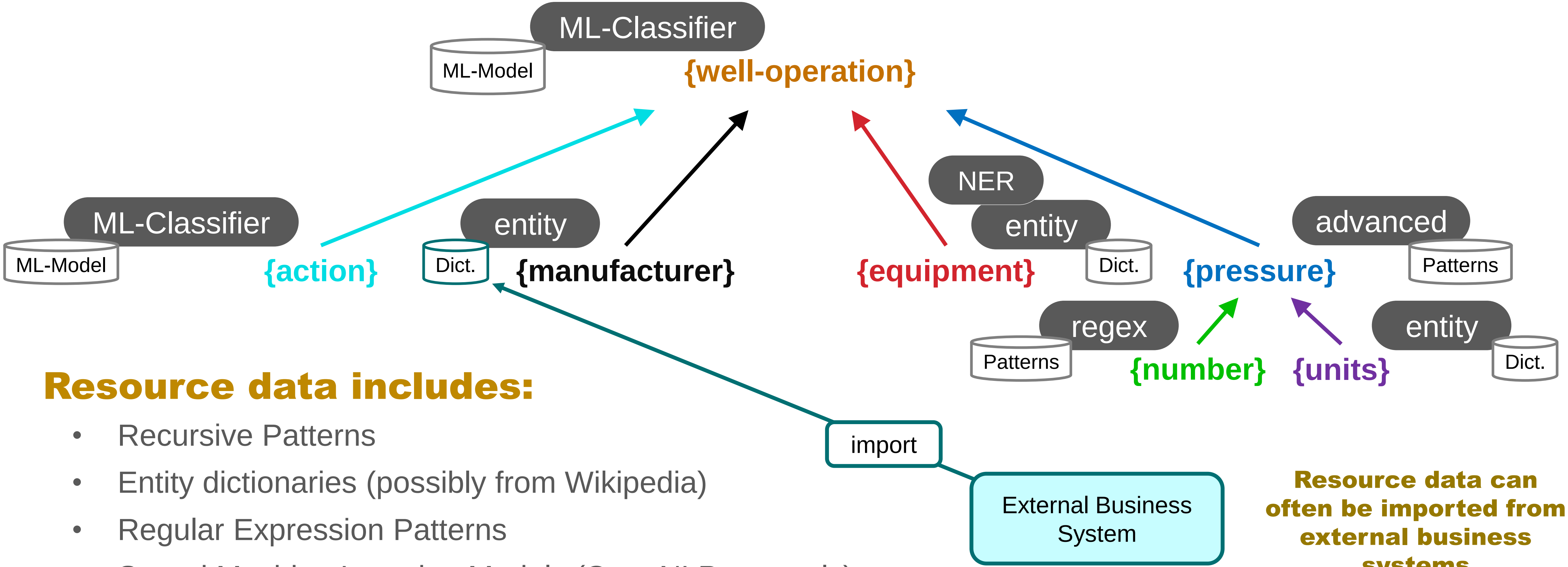
"Recognizer" = Algorithm needed to implement the tag



- **Recognizer algorithms are pre-packaged.**
  - New ones can be plugged in as needed
- **Business users choose the best recognizer(s) for each tag.**
- **Tags can have multiple recognizers.**

RECOGNIZERS HAVE RESOURCE DATA

All resource data is managed entirely by Saga



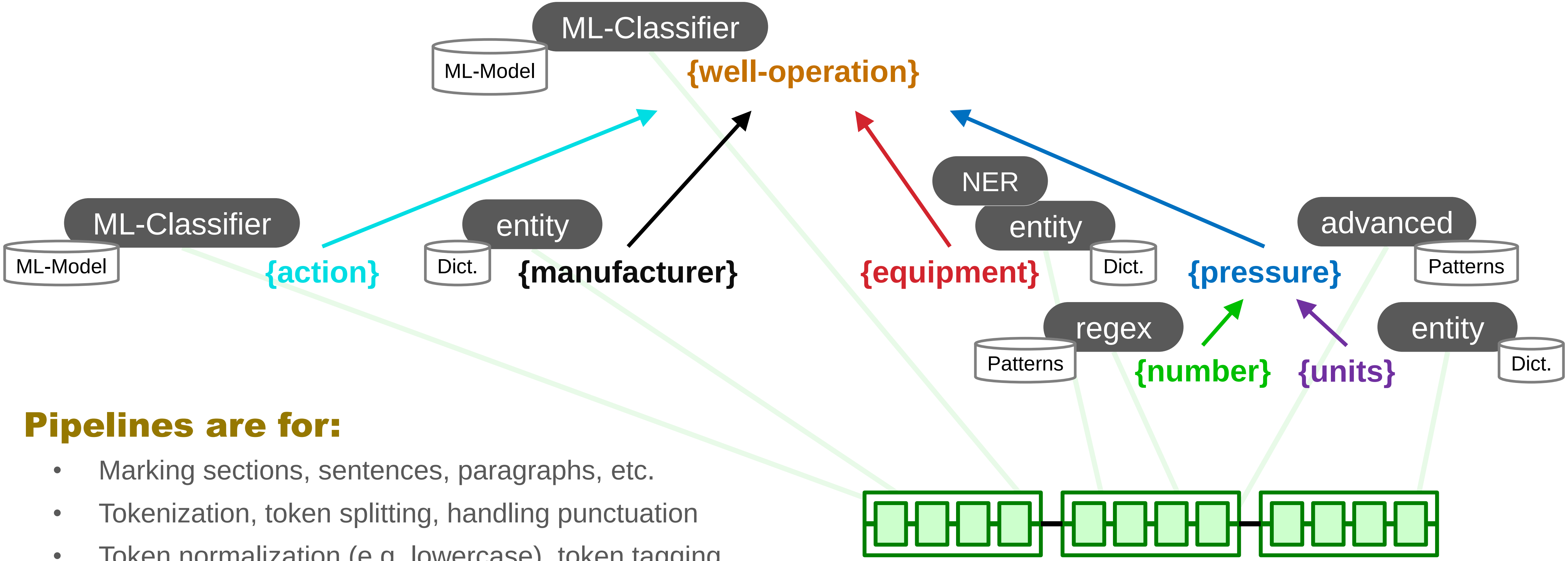
Resource data includes:

- Recursive Patterns
- Entity dictionaries (possibly from Wikipedia)
- Regular Expression Patterns
- Stored Machine Learning Models (OpenNLP currently)
- Configuration parameters

Resource data can often be imported from external business systems

RECOGNIZERS SHARE TEXT PIPELINES

All pipelines are managed and automatically created by Saga

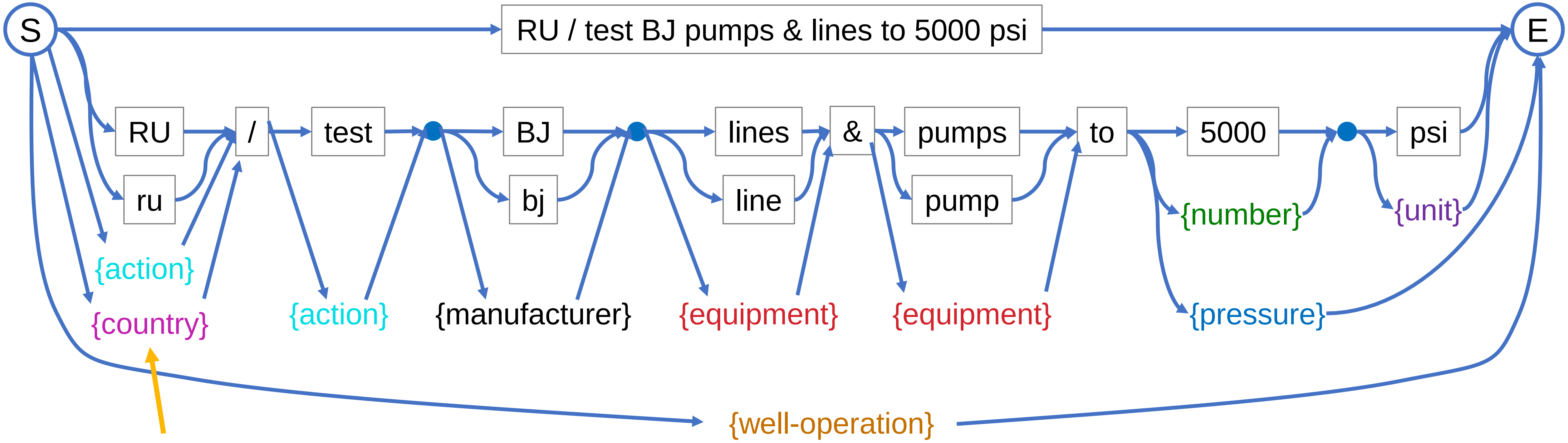


Pipelines are for:

- Marking sections, sentences, paragraphs, etc.
- Tokenization, token splitting, handling punctuation
- Token normalization (e.g. lowercase), token tagging
- Lemmatization (reducing variants)

SAGA CREATES INTERPRETATION GRAPHS

Interpretation graphs allow for the expression of ambiguity



The interpretation graph allows for ambiguity of interpretation to be gracefully represented

Tag names are used internally and output by Saga to external applications

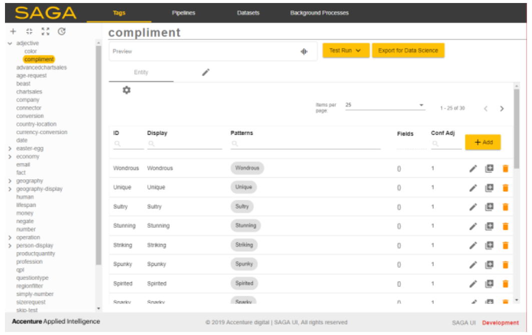
Interpretation graph represents text understanding internally.

NATURAL LANGUAGE UNDERSTANDING WITH SAGA

USER INTERFACE DETAILS

SAGA: A SYSTEM FOR MANAGING NLU

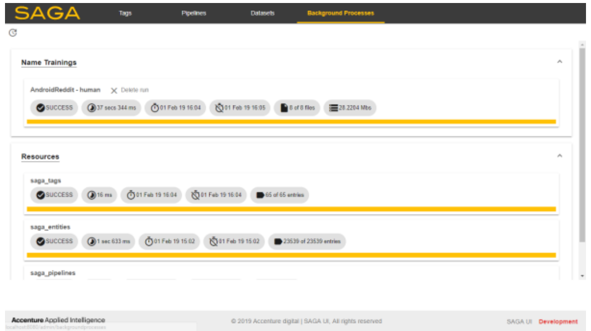
Provides Components to handle End-to-End NLU



compliment

ID	Display	Patterns	Fields	Conf Adj
Wondrous	Wondrous	Wondrous	0	1
Unique	Unique	Unique	0	1
Sultry	Sultry	Sultry	0	1
Stunning	Stunning	Stunning	0	1
Striking	Striking	Striking	0	1
Spunky	Spunky	Spunky	0	1
Spirited	Spirited	Spirited	0	1
Sparky	Sparky	Sparky	0	1

Importing & Editing Entity Databases


SAGA
Tags
Pipelines
Datasets
Background Processes
Delete run
Accenture Applied Intelligence
© 2019 Accenture digital | SAGA UI, All rights reserved
SAGA UI
Development

Name Trainings

AndroidReddit - human

SUCCESS | 37 secs 344 ms | 01 Feb 19 16:04 | 01 Feb 19 16:05 | 8 of 8 files | 28.2204 Mbs

Resources

saga_tags

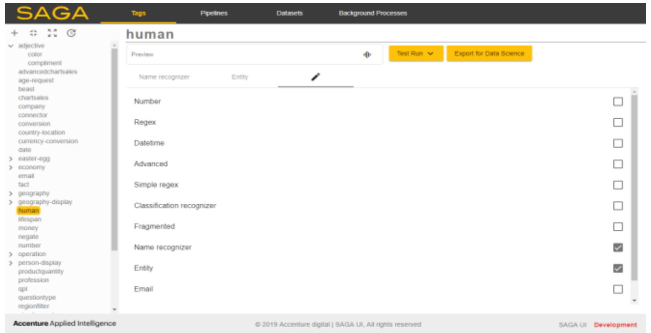
SUCCESS | 16 ms | 01 Feb 19 16:04 | 01 Feb 19 16:04 | 65 of 65 entries

saga_entities

SUCCESS | 1 sec 633 ms | 01 Feb 19 15:02 | 01 Feb 19 15:02 | 23539 of 23539 entries

saga_pipelines

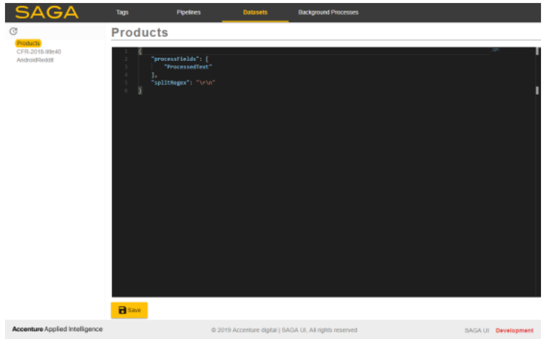
Machine Learning Training & Evaluation



human

- Number ☐
- Regex ☐
- Datetime ☐
- Advanced ☐
- Simple regex ☐
- Classification recognizer ☐
- Fragmented ☐
- Name recognizer ☑
- Entity ☑
- Email ☐

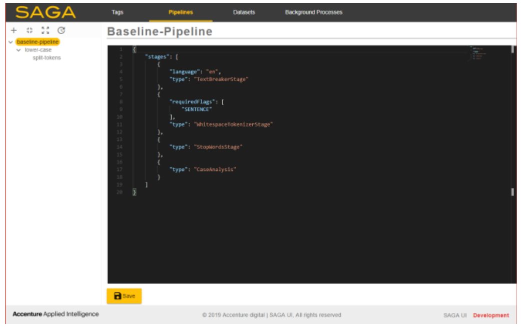
Installing, Configuring & Testing Recognizers



Products

```
{
    "processFields": [
        "ProcessedText"
    ],
    "splitRegex": "\r\n"
}
```

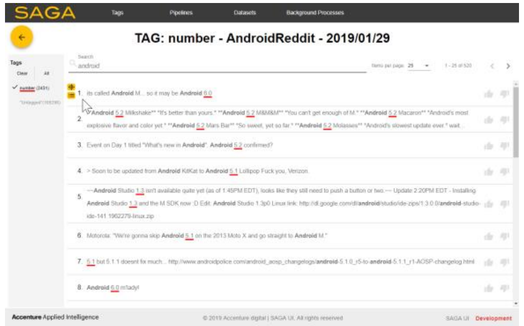
Register Training Data



Baseline-Pipeline

```
{
    "stages": [
        {
            "language": "en",
            "type": "TextBreakerStage"
        },
        {
            "requiredFlags": [
                "SENTENCE"
            ],
            "type": "WhitespaceTokenizerStage"
        },
        {
            "type": "StopWordsStage"
        },
        {
            "type": "CaseAnalysis"
        }
    ]
}
```

Creating & Testing Text Processing Pipelines



TAG: number - AndroidReddit - 2019/01/29

1. its called Android M... so it may be Android 6.0
2. **Android 5.2 Milkshake** "It's better than yours." **Android 5.2 M&M&M** "You can't get enough of M." **Android 5.2 Macaron** "Android's most explosive flavor and color yet." **Android 5.2 Mars Bar** "So sweet, yet so far." **Android 5.2 Molasses** "Android's slowest update ever." wait...
3. Event on Day 1 titled "What's new in Android". Android 5.2 confirmed?
4. > Soon to be updated from Android KitKat to Android 5.1 Lollipop Fuck you, Verizon.
5. --- Android Studio 1.3 isn't available quite yet (as of 1:45PM EDT), looks like they still need to push a button or two.--- Update 2:20PM EDT - Installing Android Studio 1.3 and the M SDK now :D Edit: Android Studio 1.3p0 Linux link: http://dl.google.com/dl/android/studio/ide-zips/1.3.0.0/android-studio-ide-141.1962279-linux.zip
6. Motorola: "We're gonna skip Android 5.1 on the 2013 Moto X and go straight to Android M."
7. 5.1 but 5.1.1 doesnt fix much... http://www.androidpolice.com/android_aosp_changelogs/android-5.1.0_r5-to-android-5.1.1_r1-AOSP-changelog.html
8. Android 6.0 m'lady!

Supervised Evaluation, Regression Testing & Training

SEMANTIC TAGS IN SAGA

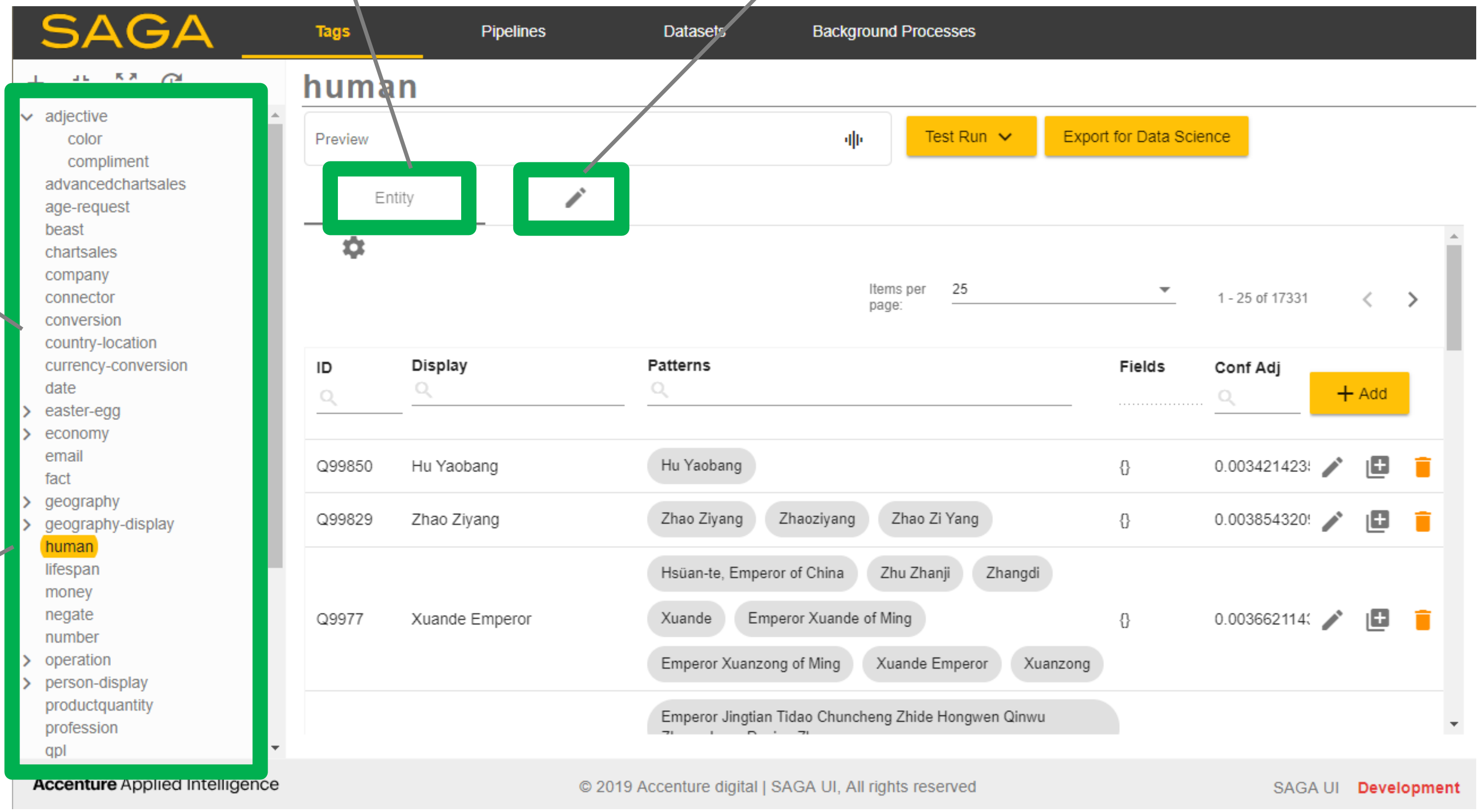
All Functionality is Organized by Semantic Tags

Semantic tags identify semantic understanding for extracted entities and classifications / intents

Tags can have multiple ***recognizers***, algorithms which identify when the semantic tag occurs in the content

Click the ***edit*** icon to add new recognizers to a tag

Tags are organized by ***semantic hierarchy***; sub-tags are specialized instances of higher-level tags



- adjective
  - color
  - compliment
- advancedchartsales
- age-request
- beast
- chartsales
- company
- connector
- conversion
- country-location
- currency-conversion
- date
- easter-egg
- economy
- email
- fact
- geography
- geography-display
- human
- lifespan
- money
- negate
- number
- operation
- person-display
- productquantity
- profession
- qpl

human

Items per page: 25

1 - 25 of 17331

ID	Display	Patterns	Fields	Conf Adj
Q99850	Hu Yaobang	Hu Yaobang	{}	0.003421423!
Q99829	Zhao Ziyang	Zhao Ziyang Zhaoziyang Zhao Zi Yang	{}	0.003854320!
Q9977	Xuande Emperor	Hsüan-te, Emperor of China Zhu Zhanji Zhangdi Xuande Emperor Xuande of Ming Emperor Xuanzong of Ming Xuande Emperor Xuanzong	{}	0.003662114;
		Emperor Jingtian Tidao Chuncheng Zhide Hongwen Qinwu		

Accenture Applied Intelligence

SAGA UI **Development**

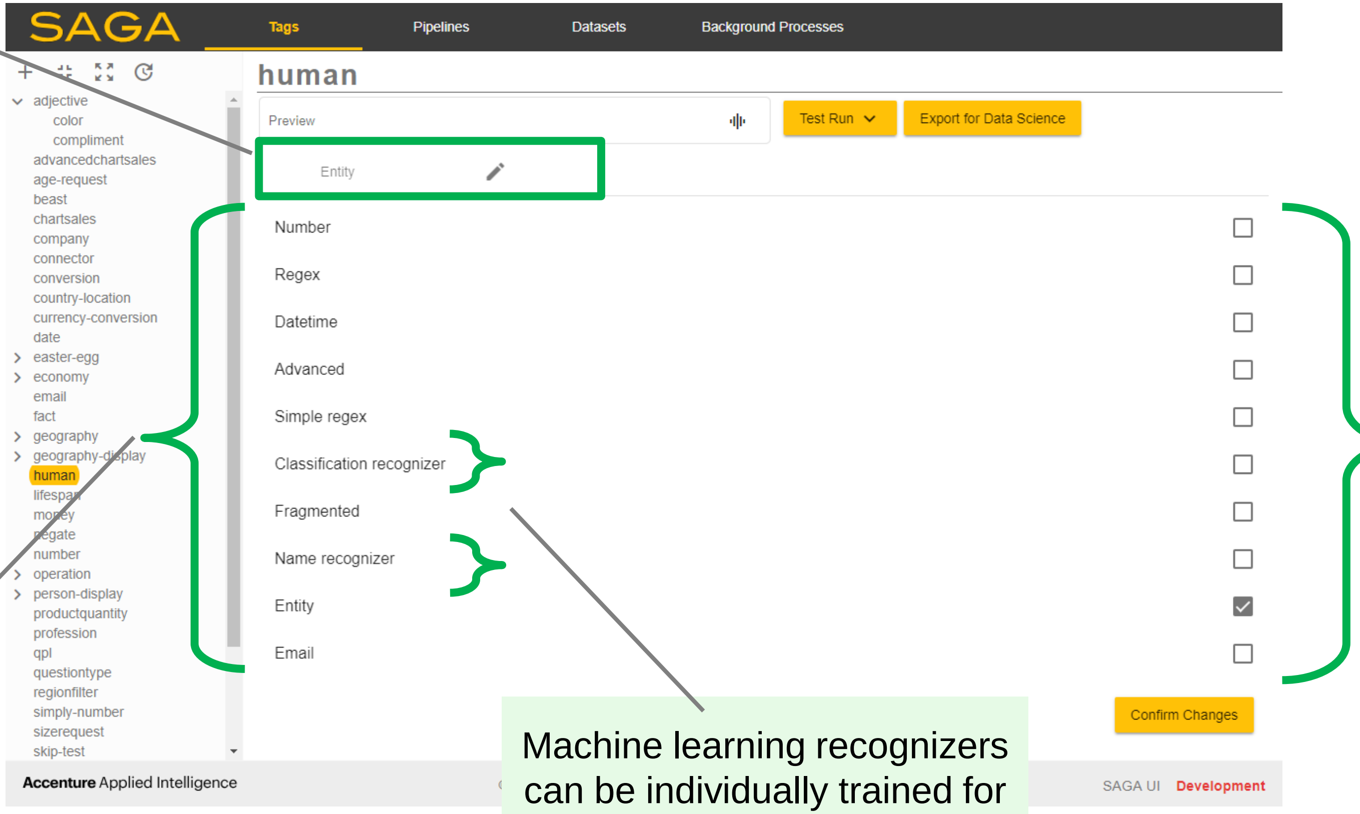
CHOOSE RECOGNIZERS

Saga ships with Out-Of-The-Box Recognizers (but you can also plug-in your own)

Selected recognizers show up as tabs inside the tag

Select the recognizers you want to use for the tag (one or more)

Only 10 recognizers today
Ultimately, we expect there will be 100’s (including industry-specific recognizers)



adjective
color
compliment
advancedchartsales
age-request
beast
chartsales
company
connector
conversion
country-location
currency-conversion
date
easter-egg
economy
email
fact
geography
geography-display
human
lifespan
money
negate
number
operation
person-display
productquantity
profession
qpl
questiontype
regionfilter
simply-number
sizerequest
skip-test

human

Preview

Number □
Regex □
Datetime □
Advanced □
Simple regex □
Classification recognizer □
Fragmented □
Name recognizer □
Entity ☑
Email □

Machine learning recognizers can be individually trained for each tag

RECOGNIZERS

Pattern-Based

Pre-Packaged:

- ***Number***: *1, 1.4, 1.4e100, first, second, 2nd, iii*
- ***Datetime***: *December 12, 1/4/2019, June 1998, 2019-01-01, 8:30, 20140910, June 2nd*
- ***E-Mail***: *paul.e.nelson@accenture.com*

Regular Expressions:

- ***Cross-Token Regex***: *Slower, more comprehensive, all variations across multiple tokens*
- ***Simple Regex***: *Faster, must match within a single token*

Dictionary Based Entity Recognizer: *Scaled to very large dictionaries (millions of items)*

Advanced Patterns: *Recursive nested patterns of tokens and other tags*

Fragment Patterns: *Matches sets of items (tokens and other tags) within specified proximity*

Machine-Learning Based

Named Entity Recognizer

- Machine-Learning Entity Recognizer
  - OpenNLP: Perceptron & MaxEnt algorithms
- Uses pattern data as training input
- Pre-trained English & Spanish person recognizers

Text Classifier

- Machine-Learning Based classifier
  - OpenNLP: MaxEnt, Naïve Bayes, Perceptron
  - Bag Of Words with configurable n-gram word sequences
- Configurable to sentence, text-block or other text-breaker boundary

MORE TO COME

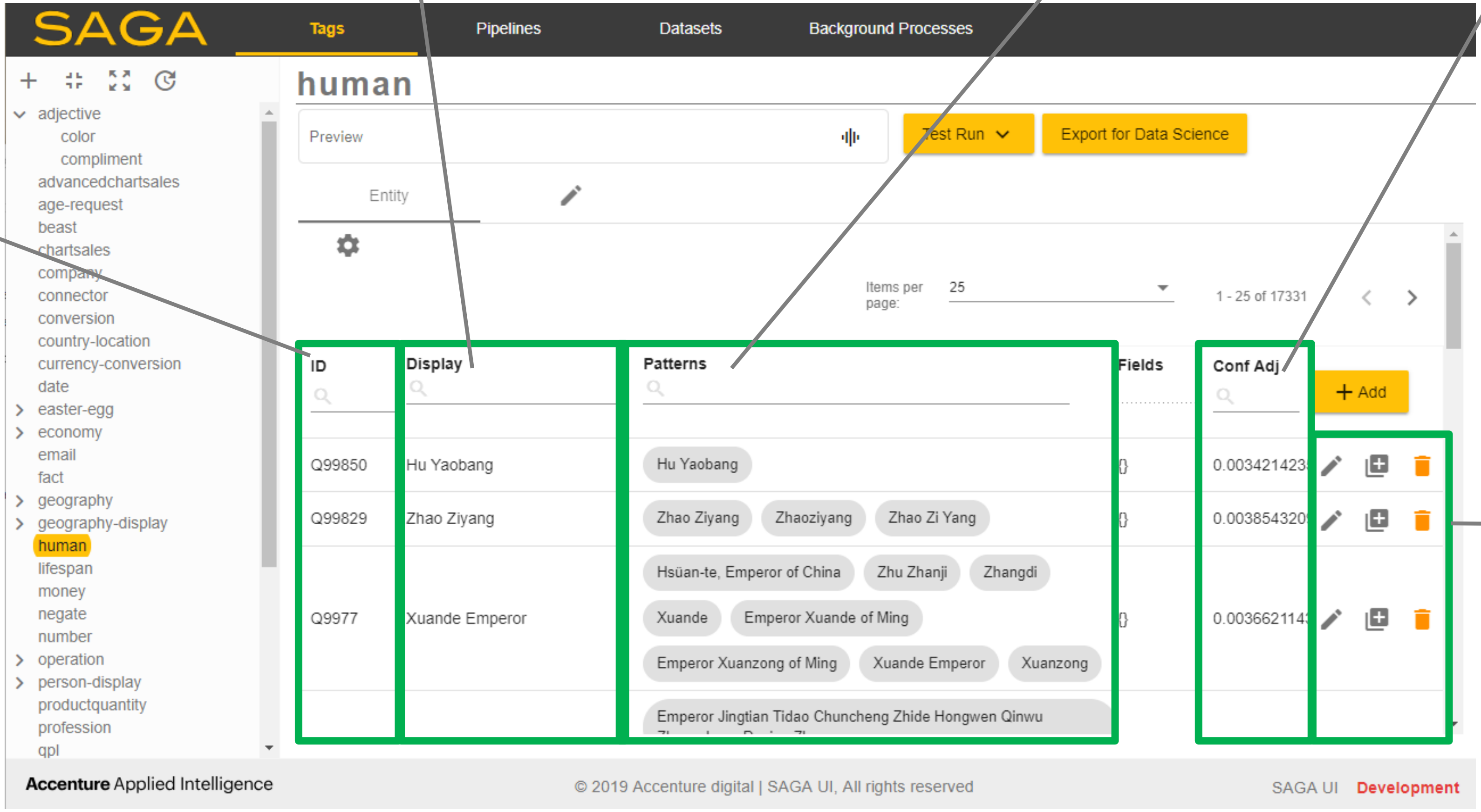
DICTIONARY ENTITY EDITOR

Entities have business identifiers (keys into business systems) to link the Natural Language Output to business objects

The user-friendly or canonical name of the entity to display to the user

Each entity can have multiple dictionary patterns which identify the entity in the text.
Patterns can be ambiguous across entities (same pattern for multiple entities)

Confidence helps to disambiguate one entity from another (FUTURE: Confidence based on context will be available)



- adjective
  - color
  - compliment
- advancedchartsales
- age-request
- beast
- chartsales
- company
- connector
- conversion
- country-location
- currency-conversion
- date
- easter-egg
- economy
- email
- fact
- geography
- geography-display
- human
- lifespan
- money
- negate
- number
- operation
- person-display
- productquantity
- profession
- qpl

human

Items per page: 25

1 - 25 of 17331

ID	Display	Patterns	Fields	Conf Adj
Q99850	Hu Yaobang	Hu Yaobang	{}	0.003421423
Q99829	Zhao Ziyang	Zhao Ziyang Zhaoziyang Zhao Zi Yang	{}	0.003854320
Q9977	Xuande Emperor	Hsüan-te, Emperor of China Zhu Zhanji Zhangdi Xuande Emperor Xuande of Ming Emperor Xuanzong of Ming Xuande Emperor Xuanzong	{}	0.003662114
		Emperor Jingtian Tidao Chuncheng Zhide Hongwen Qinwu		

Accenture Applied Intelligence

© 2019 Accenture digital | SAGA UI, All rights reserved

SAGA UI **Development**

Editing Controls

Entities are business objects of interest for the application

ADVANCED PATTERNS

Recursive and Nested Pattern Parser

Enter text at any time for an immediate preview. Saga builds the new pipeline on-the-fly and shows you the results

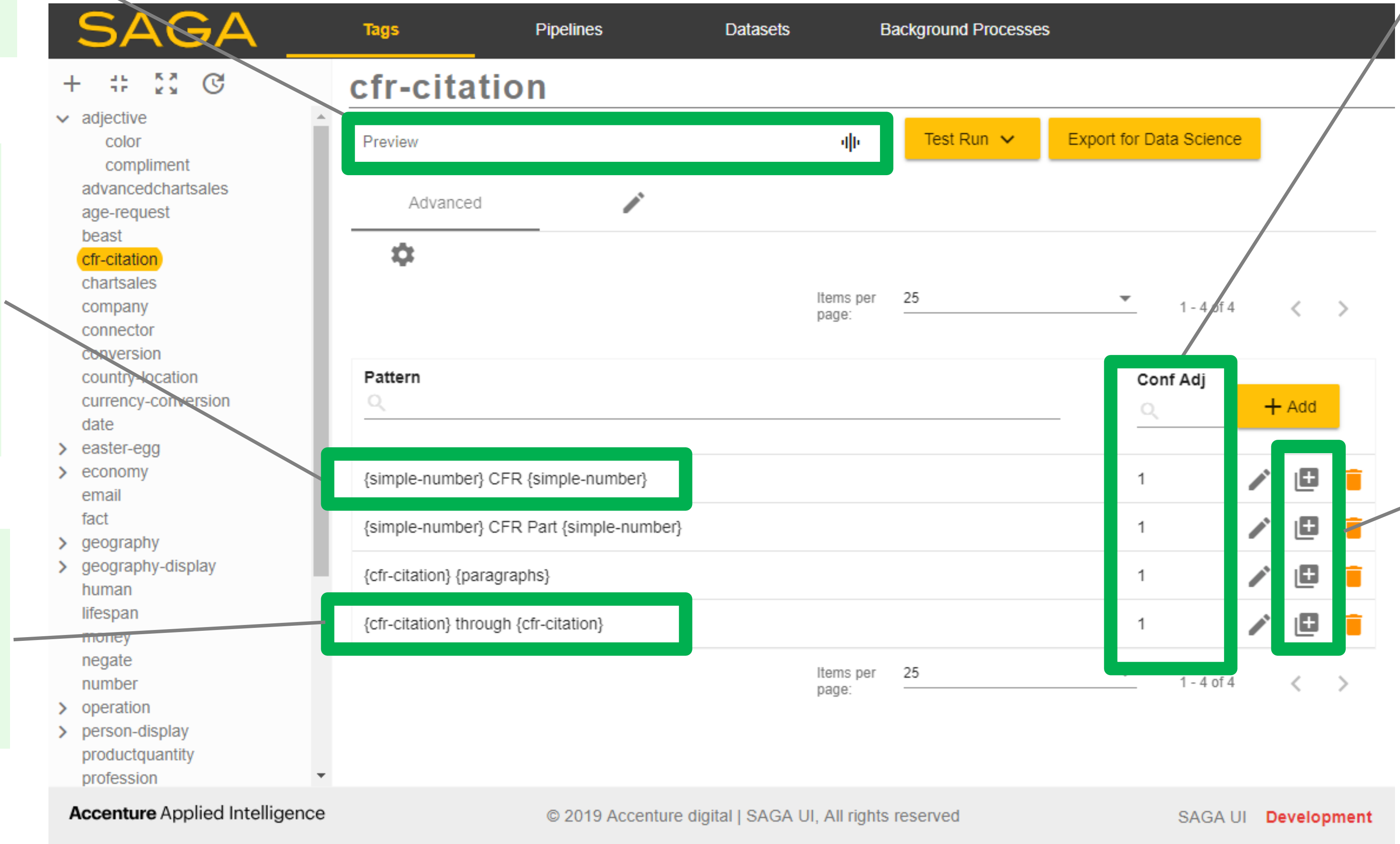
Advanced patterns can include references to other tags as {tag}.

Pipeline processing for dependency tags will be included automatically

Advanced patterns can reference themselves to create nested and recursive patterns

Use Confidence Adjustment to boost or reduce the resulting pattern.

Special editing control to copy and modify a pattern



cfr-citation

Pattern	Conf Adj
{simple-number} CFR {simple-number}	1
{simple-number} CFR Part {simple-number}	1
{cfr-citation} {paragraphs}	1
{cfr-citation} through {cfr-citation}	1

Accenture Applied Intelligence

© 2019 Accenture digital | SAGA UI, All rights reserved

SAGA UI **Development**

MACHINE LEARNING RECOGNIZERS

Weakly Superviced Training: Use Pattern Output as Training Data

Machine learning semantic analysis appear as recognizers, just like any other

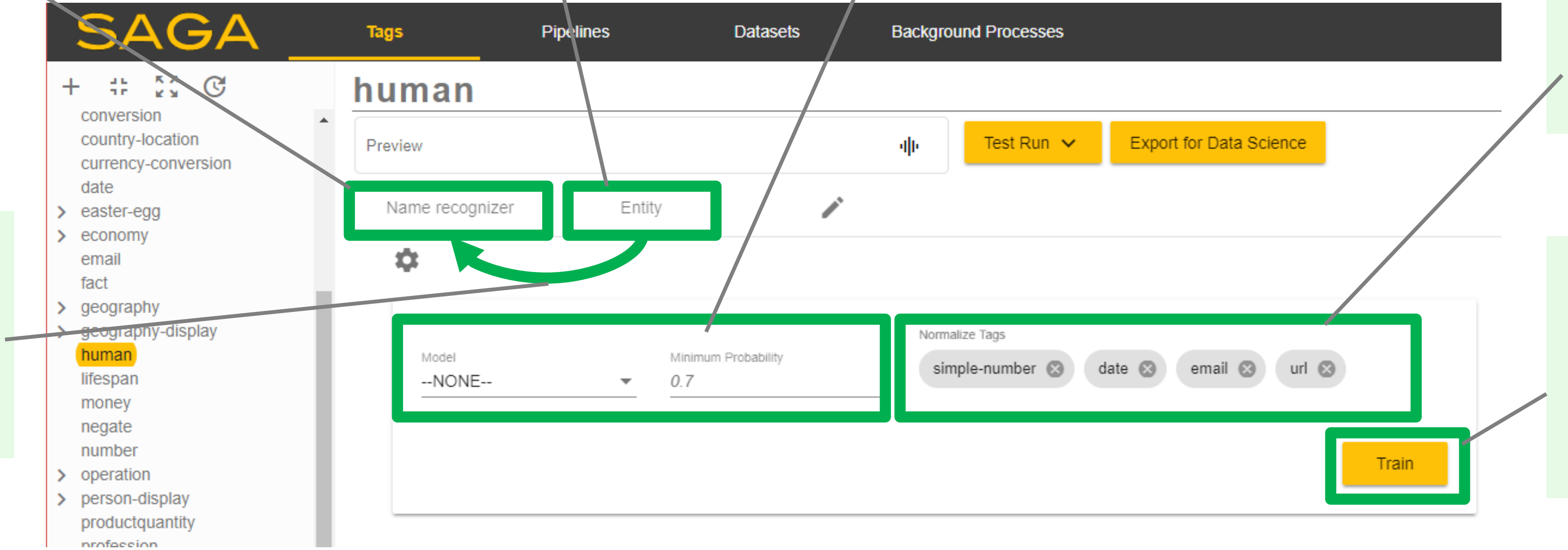
Note that both pattern-based recognizers and ML recognizers can be used simultaneously

Past trained models and pre-trained models can be selected to be run in production with a probability threshold.

Specified tags will be processed to cleanse & normalize the input to machine learning

Pattern-based tags are used, automatically, as training input for machine learning

Click to re-train the ML Model. Training runs are queued as background tasks



Model: --NONE--

Minimum Probability: 0.7

Normalize Tags: simple-number, date, email, url

MACHINE LEARNING: TRAINING RUNS

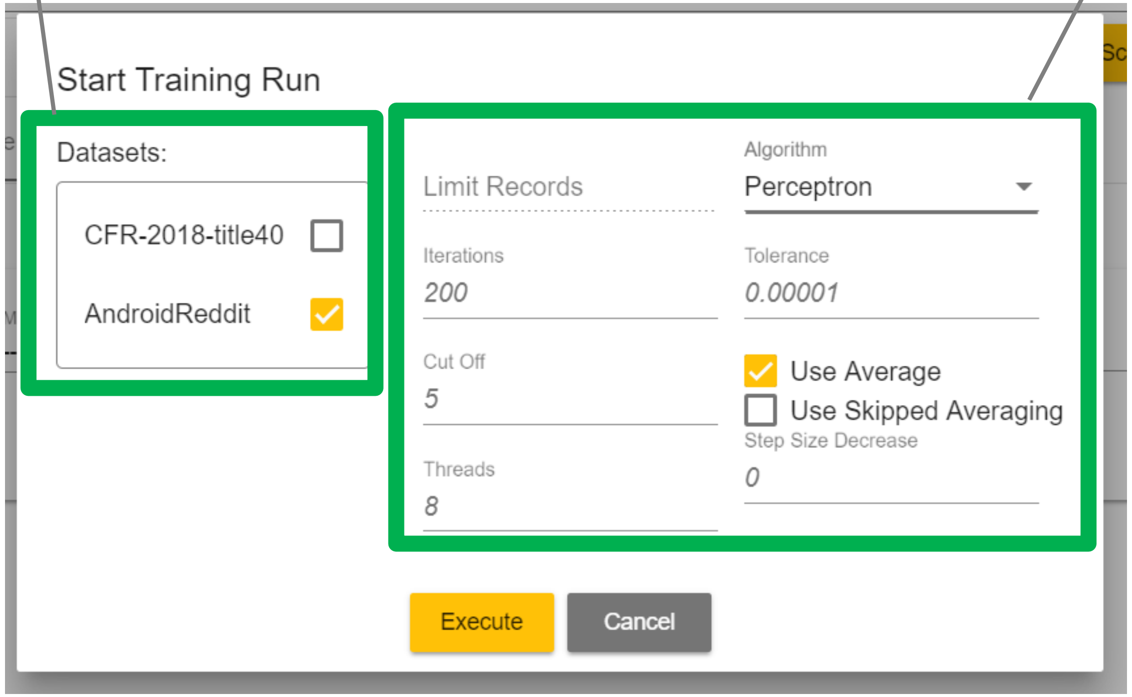
Algorithms from OpenNLP

Choose dataset to run training over

Specify algorithm and training parameters

Specify types of Word N-Grams used for classification

ML – Named Entity Recognizer



Start Training Run

Datasets:

CFR-2018-title40 □

AndroidReddit ☑

Limit Records

Algorithm: Perceptron

Iterations: 200

Tolerance: 0.00001

Cut Off: 5

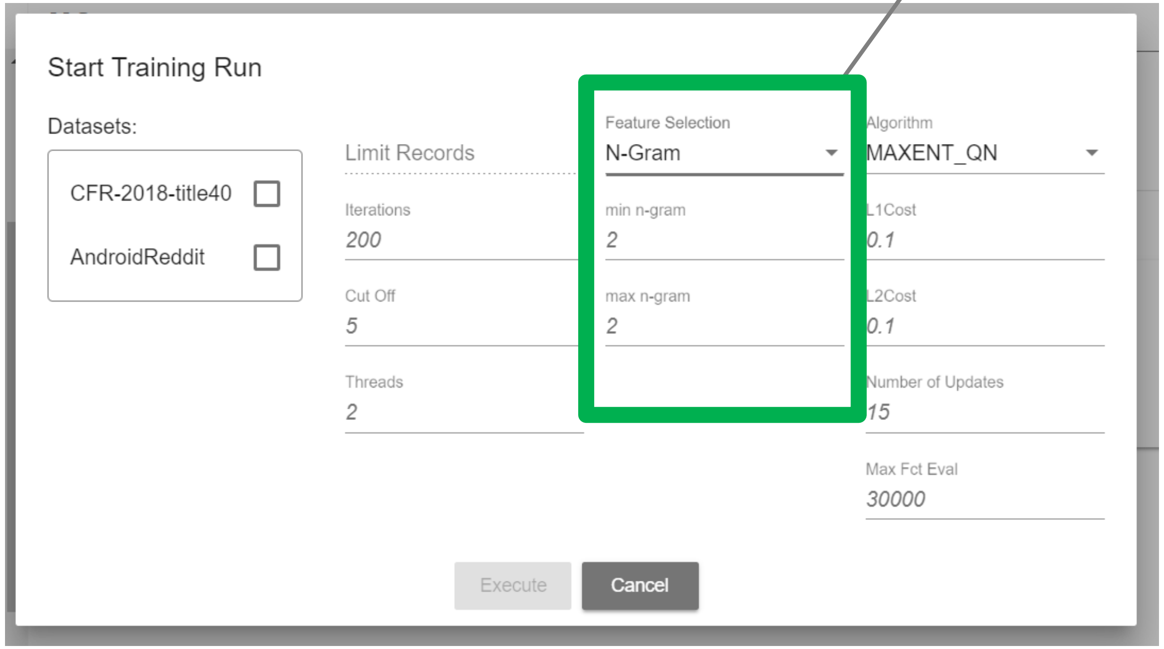
☑ Use Average

□ Use Skipped Averaging

Step Size Decrease: 0

Threads: 8

ML – Text Classifier



Start Training Run

Datasets:

CFR-2018-title40 □

AndroidReddit □

Limit Records

Iterations: 200

Cut Off: 5

Threads: 2

Feature Selection: N-Gram

min n-gram: 2

max n-gram: 2

Algorithm: MAXENT_QN

L1Cost: 0.1

L2Cost: 0.1

Number of Updates: 15

Max Fct Eval: 30000

COMMON CONTROLS FOR ALL TAGS

Integrated Testing and Data Science Exports

Type in some text to **Preview** the output of the tag.

Saga will dynamically construct the pipeline necessary to process the tag and show a detailed parse output.

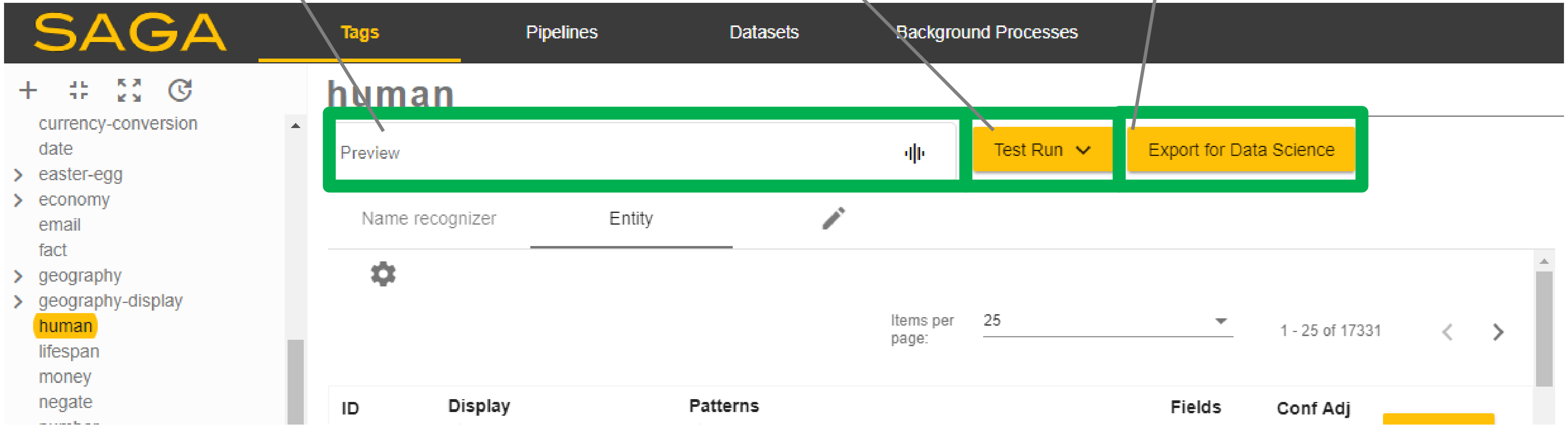
Click **Test Run** to run an entire data set of sample data through this tag (and all of its dependency tags). Results will be shown in a search interface with markup.

Test runs are spooled to the background as a background task

Export for Data Science will run all of the text through the pipeline and export it to JSON lines files so it can be analyzed off-line by a data scientist.

New models created by the data scientist can then be plugged back into Saga as a new, configurable recognizer.

Data Science Exports are spooled to the background.



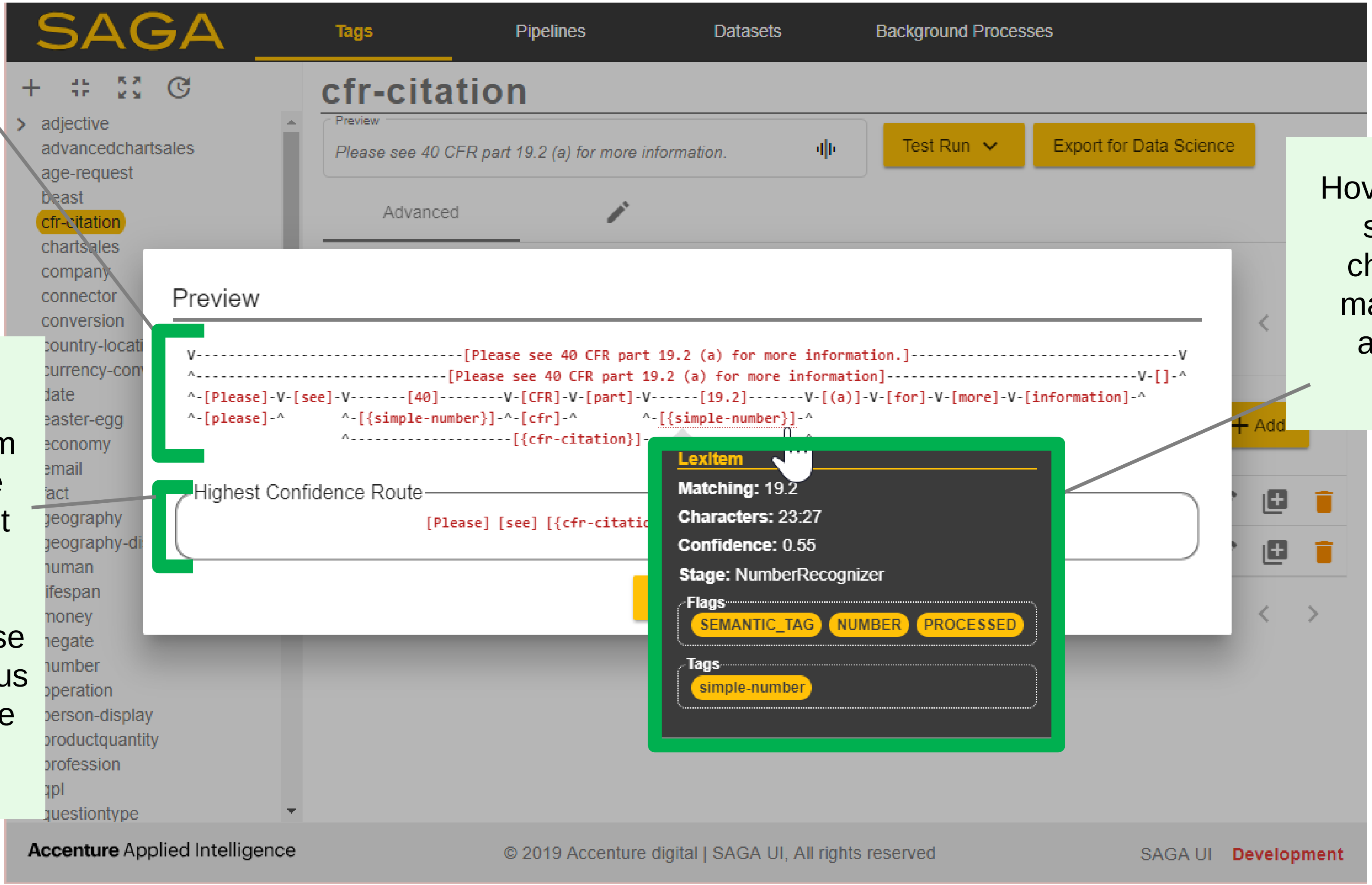
SHOW PREVIEW

Enter any text and immediately see how Saga interprets it

The interpretation graph shows how the text is interpreted by Saga every step of the way

The “highest confidence route” shows the path from the start to the end of the text which has the highest average confidence.

This is often used to choose between multiple ambiguous interpretations of the same content



cfr-citation

Preview: Please see 40 CFR part 19.2 (a) for more information.

Preview

```
V-------------------------------[Please see 40 CFR part 19.2 (a) for more information.]---------------------------------V
^-----------------------------[Please see 40 CFR part 19.2 (a) for more information]----------------------------V-[]-^
^-[Please]-V-[see]-V-------[40]--------V-[CFR]-V-[part]-V------[19.2]-------V-[(a)]-V-[for]-V-[more]-V-[information]-^
^-[please]-^       ^-[{simple-number}]-^-[cfr]-^         ^-[{simple-number}]-^
                   ^-------------------[{cfr-citation}]-                        -^
```

Highest Confidence Route

[Please] [see] [{cfr-citatic

LexItem

Matching: 19.2

Characters: 23:27

Confidence: 0.55

Stage: NumberRecognizer

Flags: SEMANTIC_TAG NUMBER PROCESSED

Tags: simple-number

Accenture Applied Intelligence

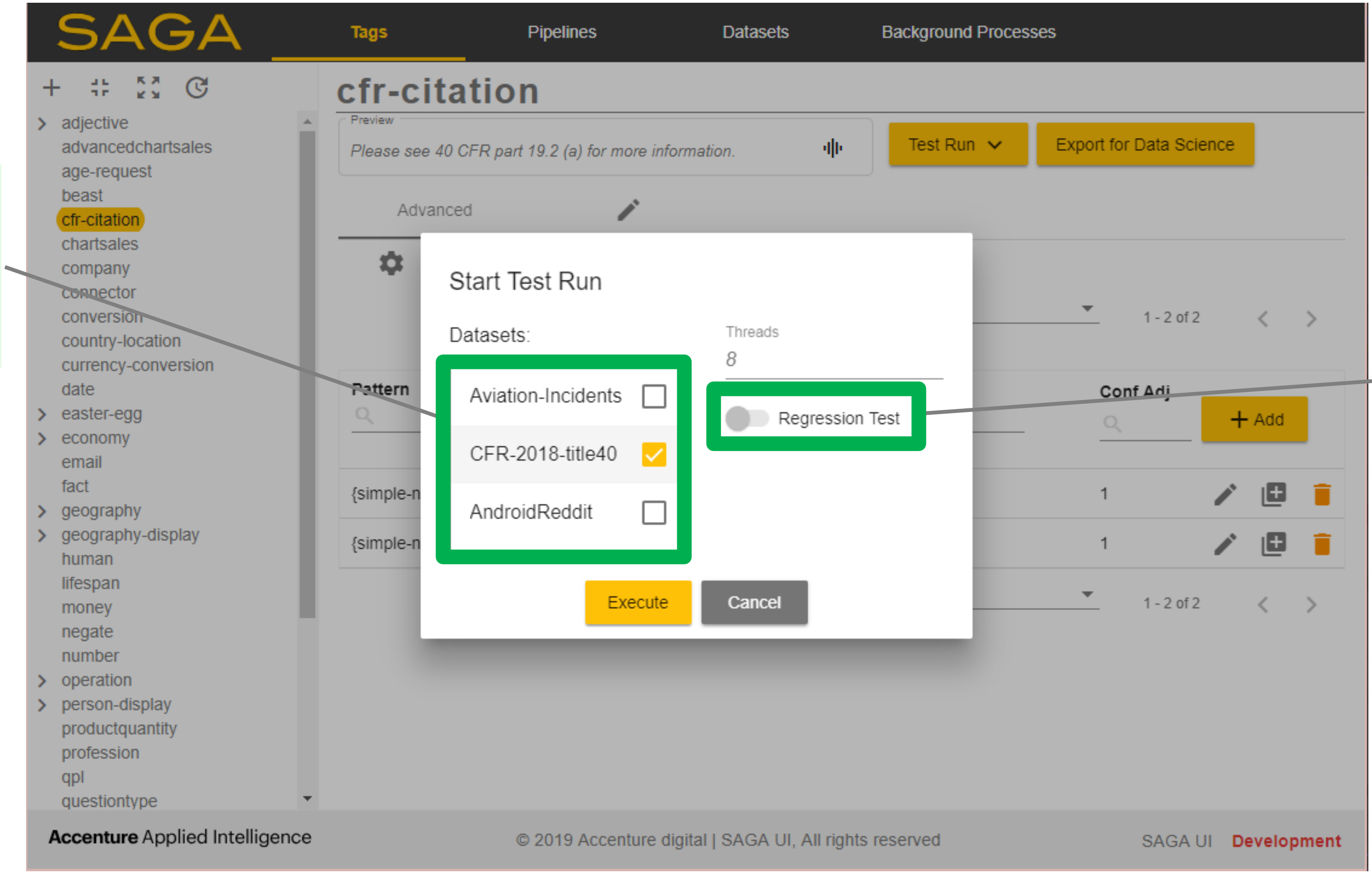
© 2019 Accenture digital | SAGA UI, All rights reserved

Hover over any lexical item shows matching text, character positions, the matching pattern (where appropriate) flags and semantic tags

STARTING A TEST RUN

Test your tag against a large amount of sample data.

Choose data set to process for the selected tag



Preview: *Please see 40 CFR part 19.2 (a) for more information.*

Start Test Run

Datasets:

- Aviation-Incidents ☐
- CFR-2018-title40 ☑
- AndroidReddit ☐

Threads: 8

○ Regression Test

Pattern	Conf Adj
{simple-n	1
{simple-n	1

Regression testing will compare accuracy to prior run

(FUTURE RELEASE)

BACKGROUND TASKS

Multiple Long-Running Tasks

Test Runs

– *Run large-scale content through language model*

Export Runs

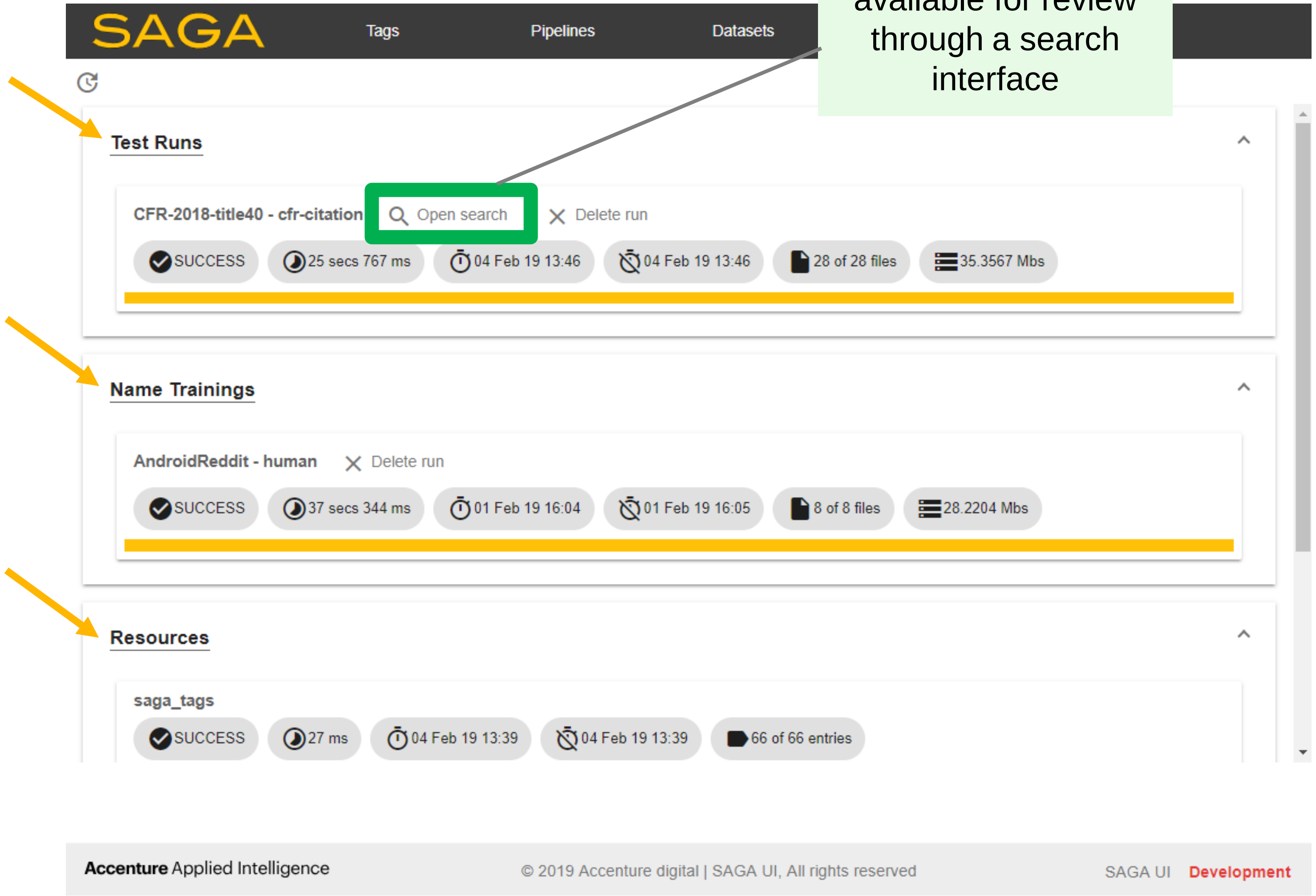
– *Export language-modeled content for external data science*

Training Runs

– *Run machine learning on dataset*

Resource Loading

– *Load large resources (large dictionaries), typically on startup*



Test run results are available for review through a search interface

Test Runs

CFR-2018-title40 - cfr-citation — Open search — ✕ Delete run

SUCCESS | 25 secs 767 ms | 04 Feb 19 13:46 | 04 Feb 19 13:46 | 28 of 28 files | 35.3567 Mbs

Name Trainings

AndroidReddit - human — ✕ Delete run

SUCCESS | 37 secs 344 ms | 01 Feb 19 16:04 | 01 Feb 19 16:05 | 8 of 8 files | 28.2204 Mbs

Resources

saga_tags

SUCCESS | 27 ms | 04 Feb 19 13:39 | 04 Feb 19 13:39 | 66 of 66 entries

Accenture Applied Intelligence

© 2019 Accenture digital | SAGA UI, All rights reserved

SAGA UI **Development**

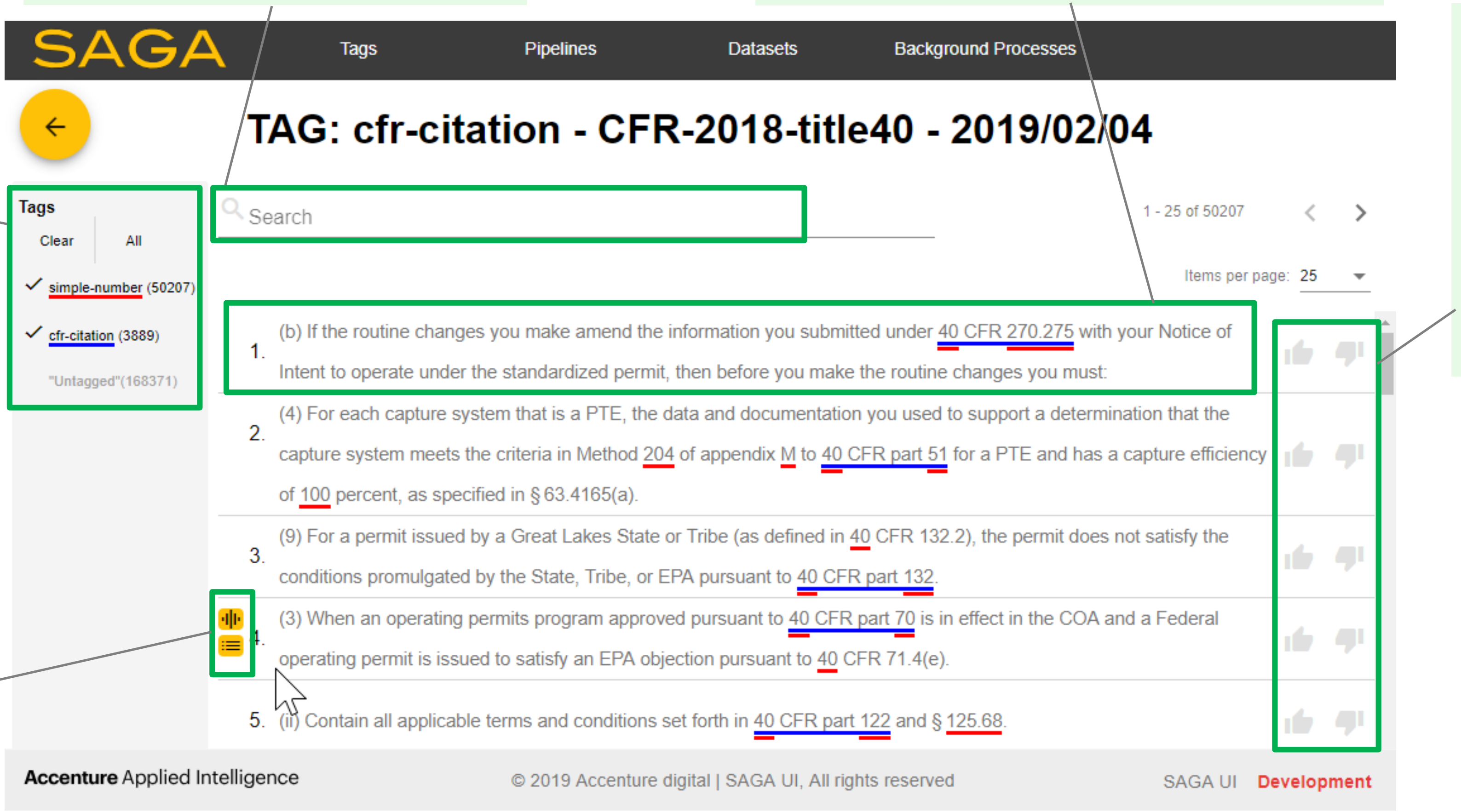
TEST RUN EVALUATION USER INTERFACE

Available after doing a test run on a tag

Generic text search box allows user to easily find and check records of interest

Color-coded tags are underlined in the sample text

All tags identified in the text are represented here as filters.



TAG: cfr-citation - CFR-2018-title40 - 2019/02/04

Tags

✓ simple-number (50207)

✓ cfr-citation (3889)

"Untagged"(168371)

1 - 25 of 50207

Items per page: 25

1. (b) If the routine changes you make amend the information you submitted under 40 CFR 270.275 with your Notice of Intent to operate under the standardized permit, then before you make the routine changes you must:
2. (4) For each capture system that is a PTE, the data and documentation you used to support a determination that the capture system meets the criteria in Method 204 of appendix M to 40 CFR part 51 for a PTE and has a capture efficiency of 100 percent, as specified in § 63.4165(a).
3. (9) For a permit issued by a Great Lakes State or Tribe (as defined in 40 CFR 132.2), the permit does not satisfy the conditions promulgated by the State, Tribe, or EPA pursuant to 40 CFR part 132.
4. (3) When an operating permits program approved pursuant to 40 CFR part 70 is in effect in the COA and a Federal operating permit is issued to satisfy an EPA objection pursuant to 40 CFR 71.4(e).
5. (ii) Contain all applicable terms and conditions set forth in 40 CFR part 122 and § 125.68.

Accenture Applied Intelligence

© 2019 Accenture digital | SAGA UI, All rights reserved

FUTURE: Buttons used for identifying correct and incorrect understanding.

Will be used for regression testing and manual training

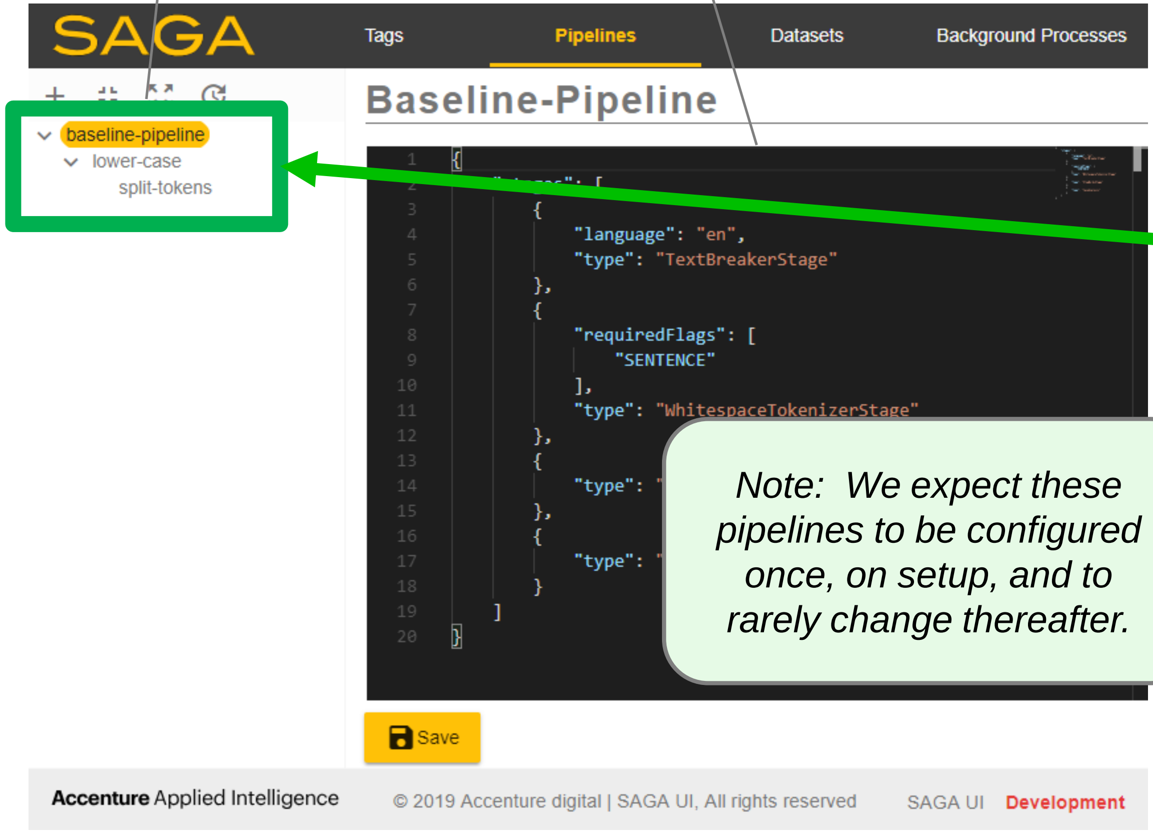
Special buttons to show the interpretation graph & complete metadata associated with a record

LOW-LEVEL TEXT PROCESSING

Multiple Pipelines can be Configured

Multiple, dependent chained pipelines can be chained together

Pipeline configured with JSON.
FUTURE: Graphical UI



Baseline-Pipeline

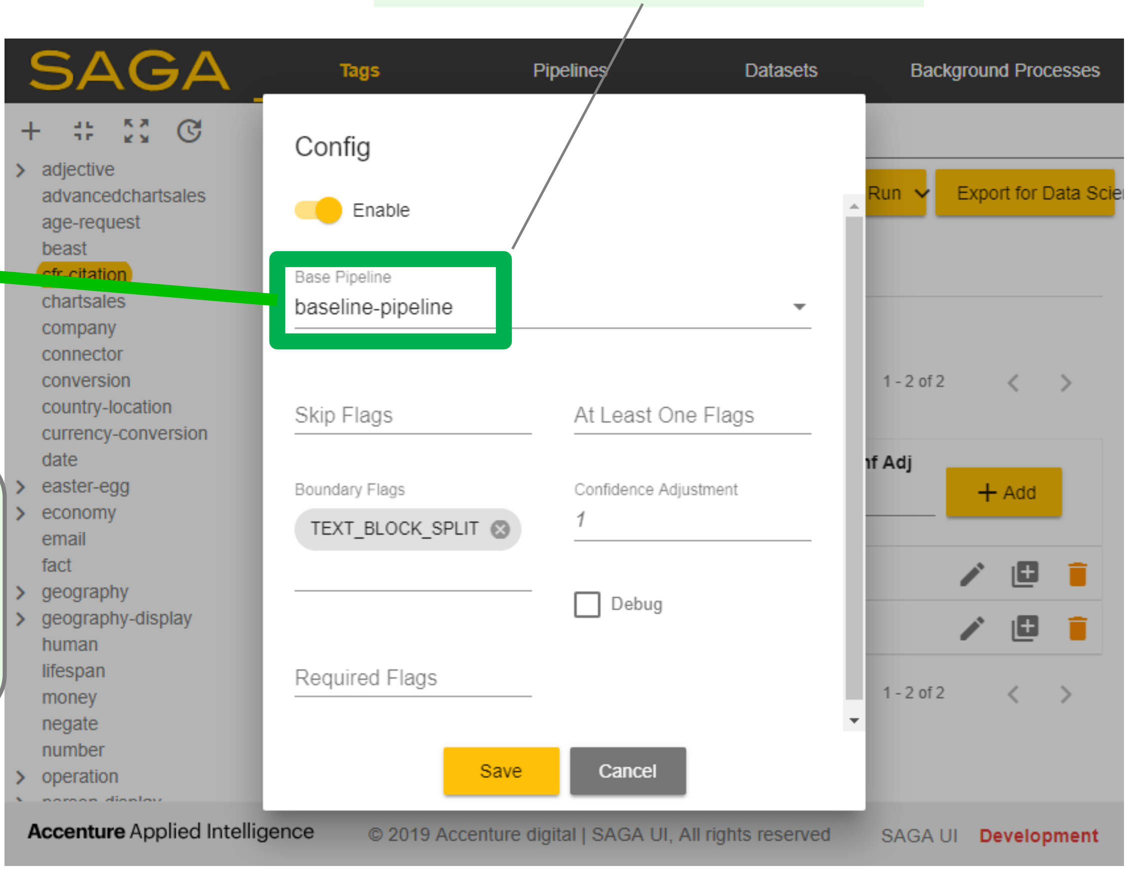
```
{
    "stages": [
        {
            "language": "en",
            "type": "TextBreakerStage"
        },
        {
            "requiredFlags": [
                "SENTENCE"
            ],
            "type": "WhitespaceTokenizerStage"
        },
        {
            "type": "
        },
        {
            "type": "
        }
    ]
}
```

Accenture Applied Intelligence © 2019 Accenture digital | SAGA UI, All rights reserved SAGA UI Development

Note: We expect these pipelines to be configured once, on setup, and to rarely change thereafter.

Each tag can tap into any pipeline

Every tag can tap into any pipeline to specify what stream of content it wishes to use



Config

Enable

Base Pipeline: baseline-pipeline

Skip Flags

At Least One Flags

Boundary Flags: TEXT_BLOCK_SPLIT

Confidence Adjustment: 1

□ Debug

Required Flags

Accenture Applied Intelligence © 2019 Accenture digital | SAGA UI, All rights reserved SAGA UI Development

LOW LEVEL TEXT PROCESSING

Currently Available

Text Breaker → *Divide up text into sentences & paragraphs*

Sentence Breaker → *ML method to identify sentence breaks [Open NLP]*

White-space tokenization → *Split text on white space*

Case analysis / lower case → *Analyze case & create lower case alternative*

Character Change Splitter → *Split tokens on punctuation or character changes (numbers, upper/lower case, etc.)*

Advanced Splitter → *Split off punctuation at the beginnings and endings of words, sentences, etc. (e.g. parenthesis, quotes, periods, etc.)*

Stop Words Tagger → *Tag small function words (articles, prepositions, small functional verbs, small adverbs, interrogatives, etc.) so they can be optionally skipped by other processors*

Lemmatizers → *Reduce words to root words, backed by Wiktionary database*

- Languages currently available: English, Spanish

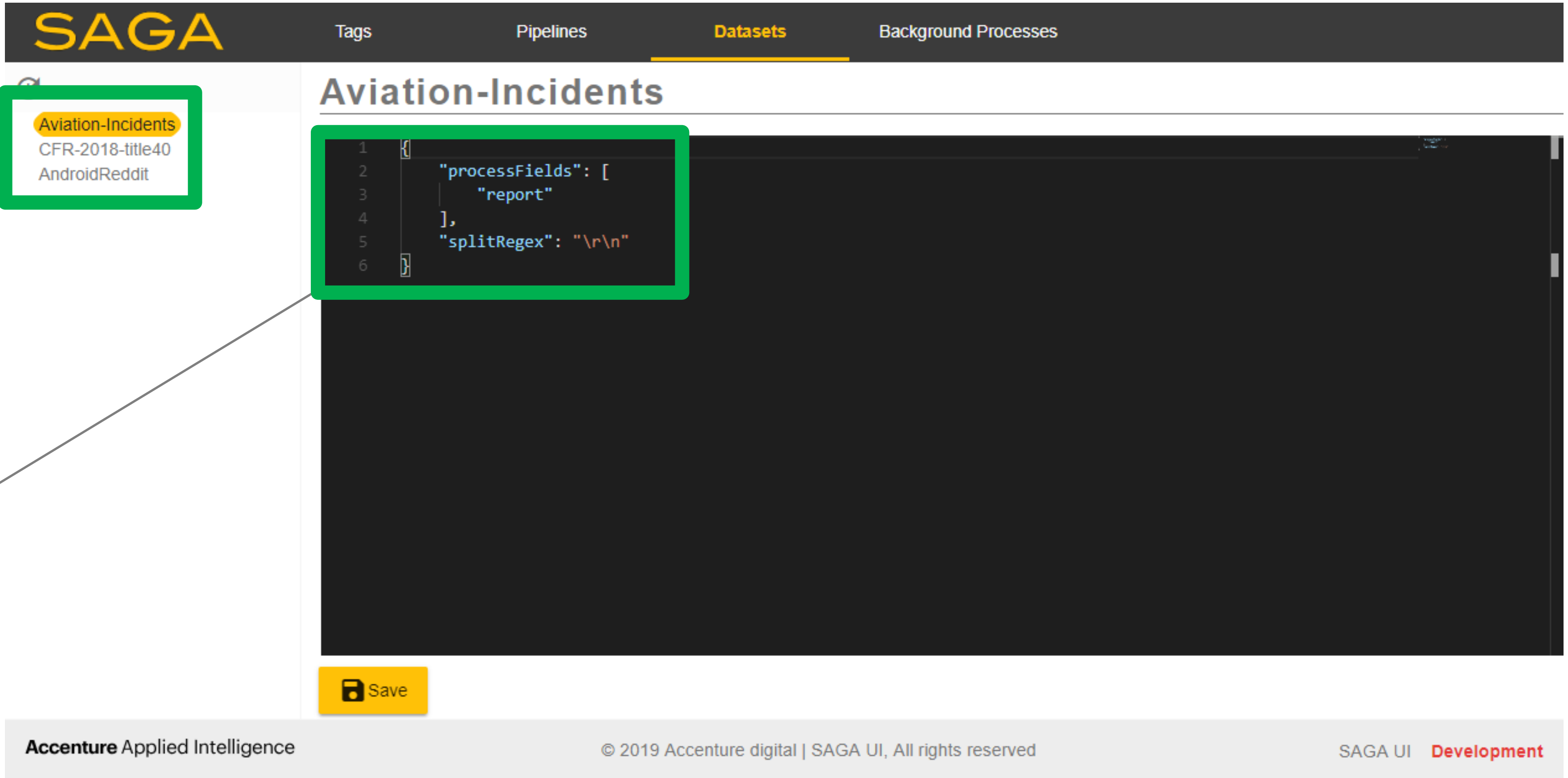
MANY MORE TO COME

DATASETS – FOR TRAINING & TESTING

Datasets are loaded automatically

Datasets are loaded into a special directory in Saga and are automatically available

Dataset configuration specifies what fields to process with NLP and how best to split large text blocks



```
{
    "processFields": [
        "report"
    ],
    "splitRegex": "\r\n"
}
```