

Watson NLP

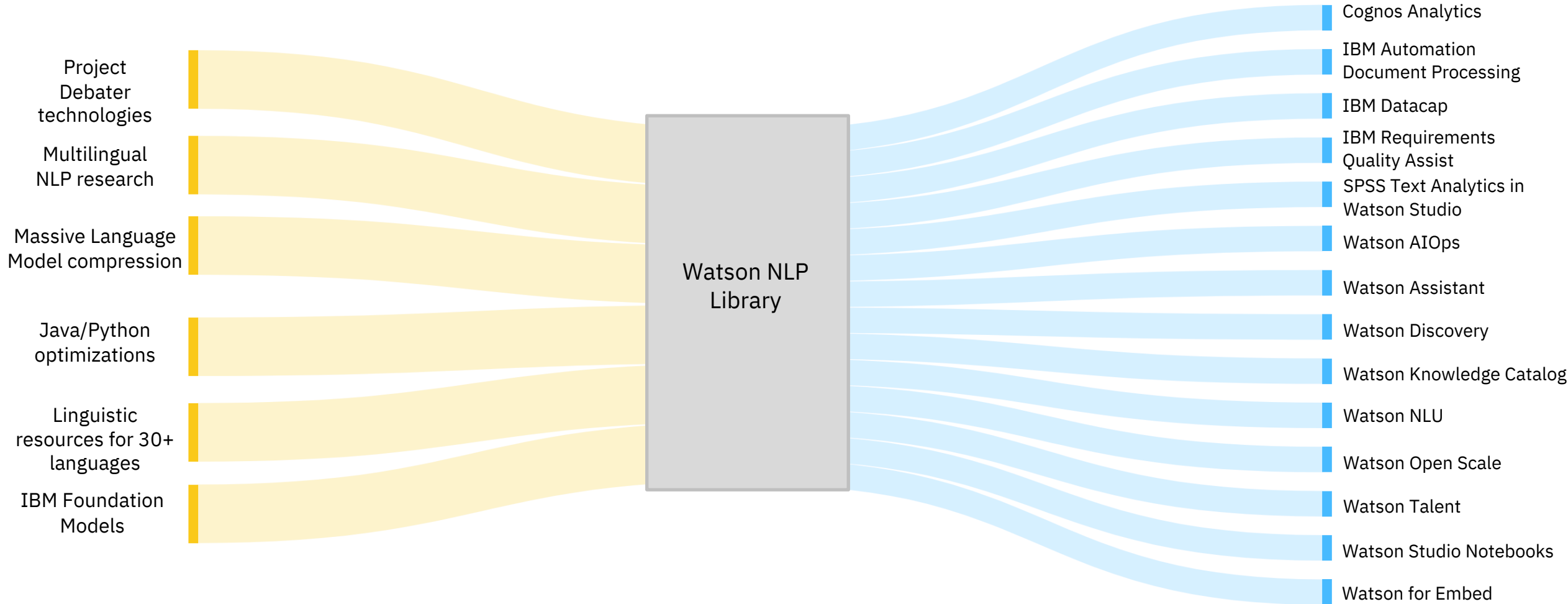
Husztí Dániel

Data & AI Technical Sales

daniel.husztí1@ibm.com

NLP at IBM: Today

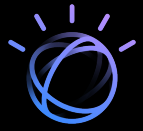
Accomplished since 2019



["A Brief History of NLP at IBM". Laura Chiticariu. Watson Core blog. March 2021.](#)

IBM Research innovations and contributions from all IBM divisions

IBM Products



Watson NLP Library

Fully embeddable NLP/NLU library for ISVs

Standard embeddable library designed to address the fragmentation in basic NLP inside IBM

State-of-the-art NLP stack with full-suite of NLP algorithms and developer-focused UX

Shipping in 20+ IBM products including Watson Discovery, Watson Assistant, Watson Studio, and Watson NLP Library for Embed

30+ languages

European: Western, Nordic, CEE

CJK, Arabic, Hebrew, Hindi, Turkish

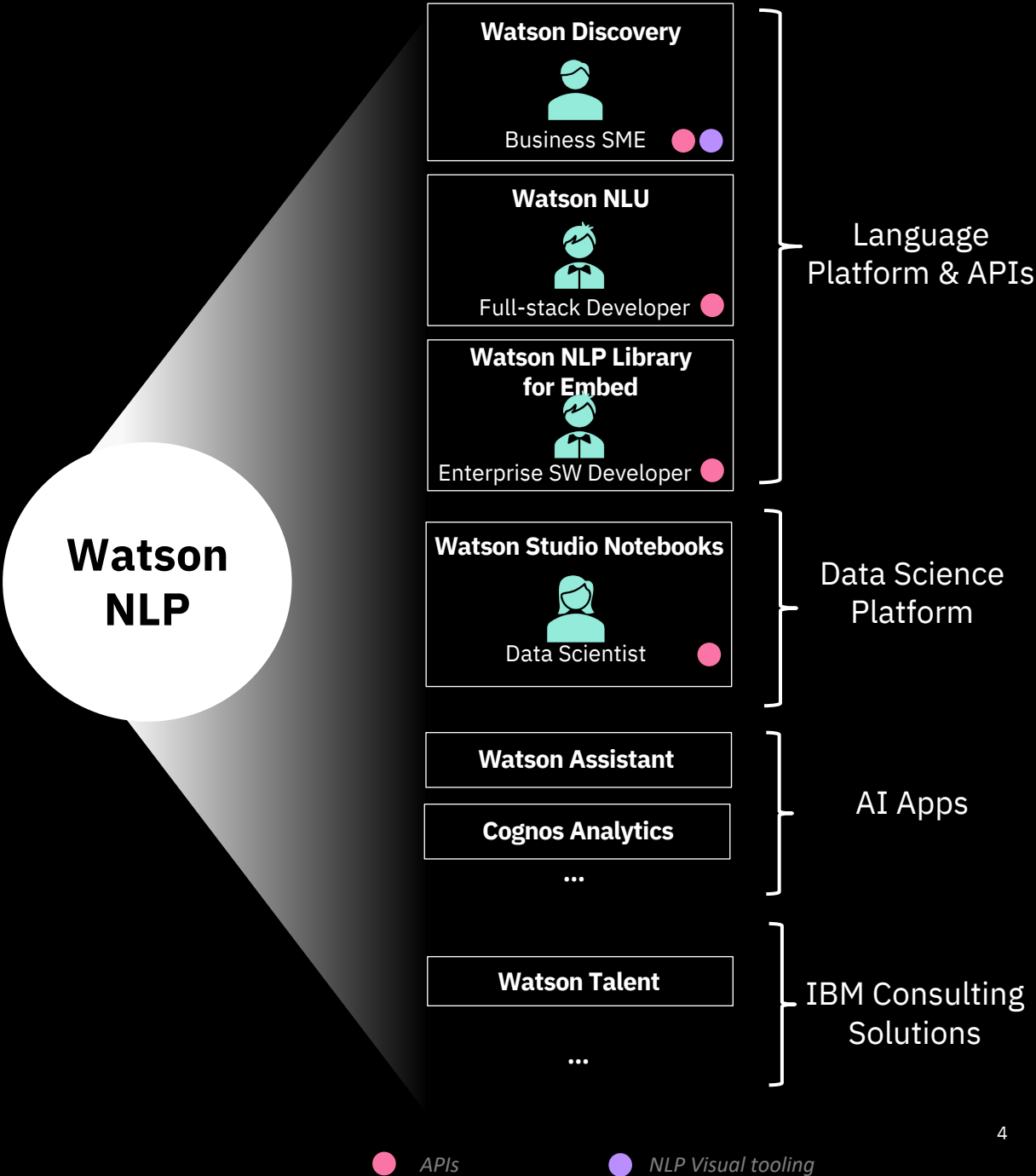
Built on top of the best AI OSS with IBM Research innovations

Watson NLP Mission

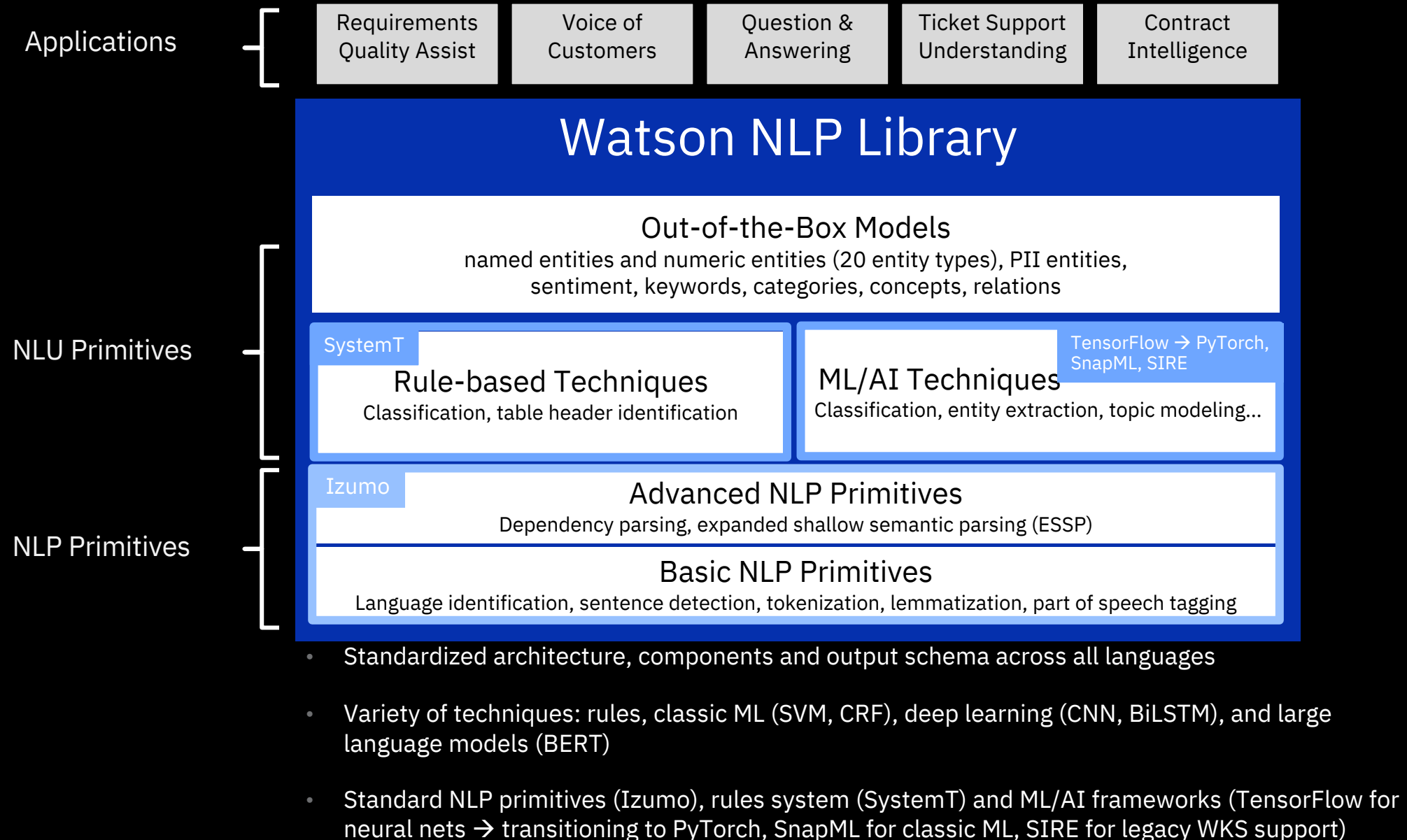
Standardized production-ready NLP functionality with latest innovations from Research

Retain IBM's leadership in Language and AI through

- Exposure in flagship products such as Watson Discovery, Watson NLU, Watson Studio and Watson NLP Library for Embed
- Innovative features in AI Apps and Solutions



Watson NLP Library: Architecture



Watson NLP Primitives

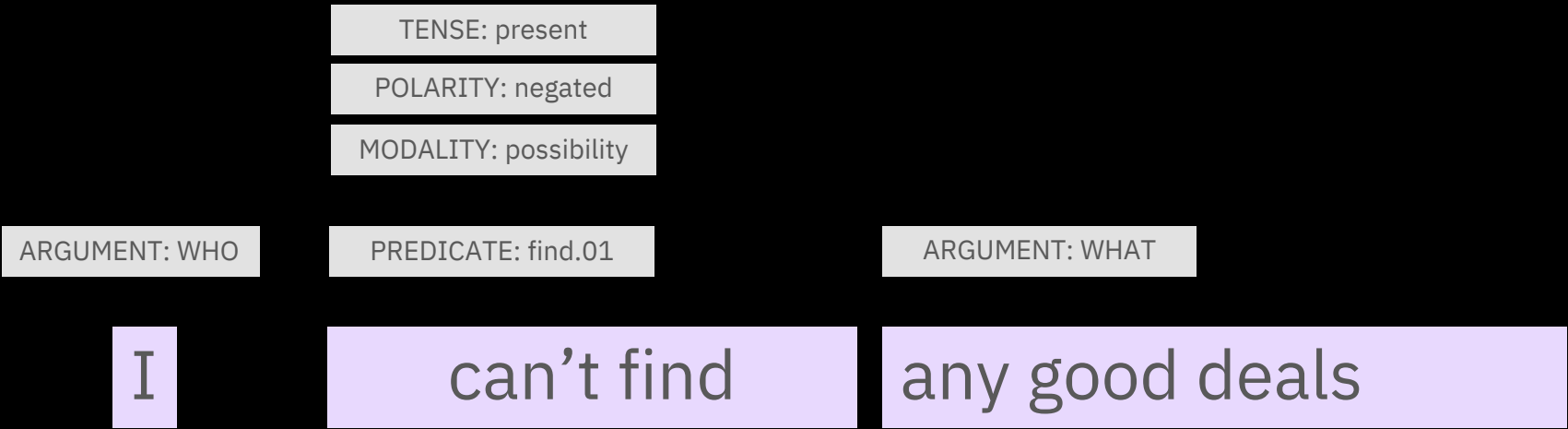
		Output Schema	Languages
Basic NLP primitives	Language Identification		69 languages
	Sentence Detection		31 languages
	Tokenization		31 languages
	Lemmatization		31 languages
	Part of Speech	Standardized* Universal Part of Speech tags	31 languages
Advanced NLP primitives	Dependency Parsing	Standardized* Universal Dependency tags	23 languages
	Expanded Shallow Semantic Parser (ESSP) <i>includes Semantic Role Labeling</i>	Standardized* Universal PropBank tags	1 language <i>(experimental)</i>

***Watson NLP adopted Universal Standard (cross-lingually stable) to scale development of text algorithms across many languages**

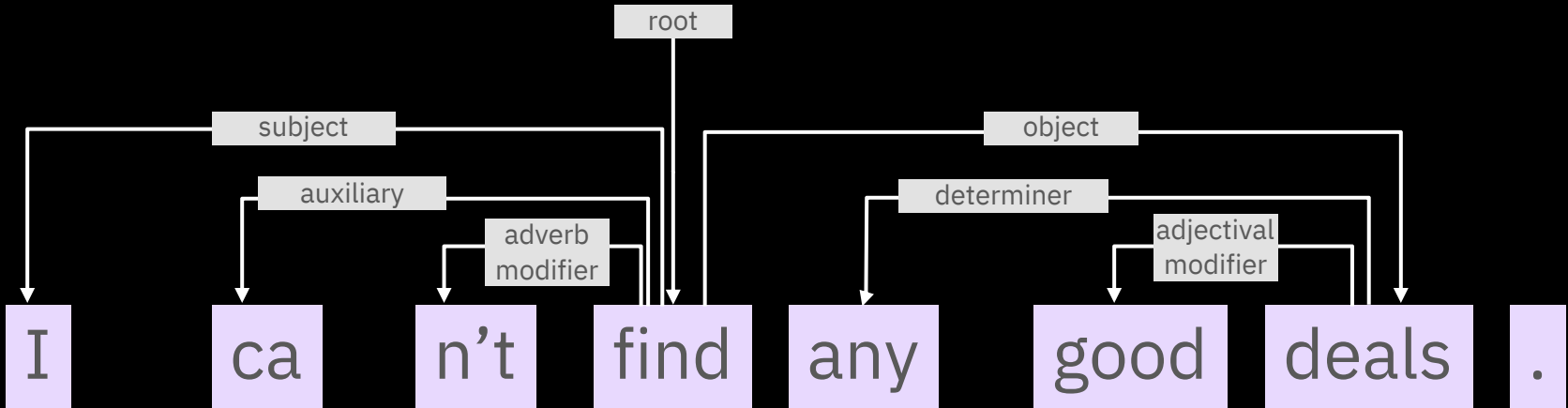
Basic NLP Primitives

PRONOUN	AUXILIARY	ADVERB	VERB	DETERMINER	ADJECTIVE	NOUN	PUNCTUATION	Parts of speech
I	can	not	find	any	good	deal	.	Lemmas
I	ca	n't	find	any	good	deals	.	Tokens
I can't find any good deals.								Sentence

Advanced NLP Primitives



Expanded Shallow Semantic Parse (ESSP), Including Semantic Role Labeling (SRL)

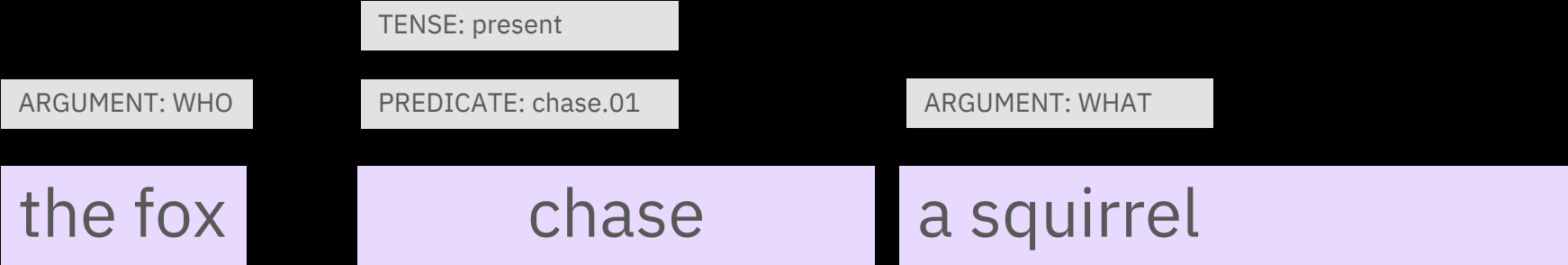


Dependency Parse Tree

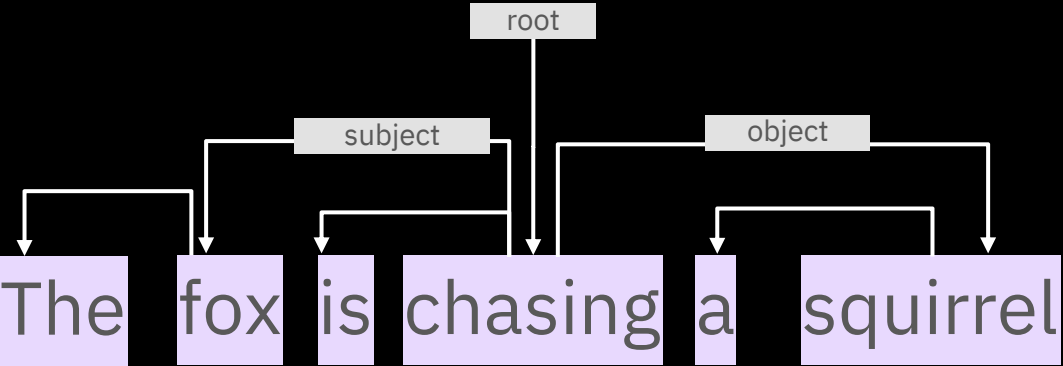
I can't find any good deals.

Sentence

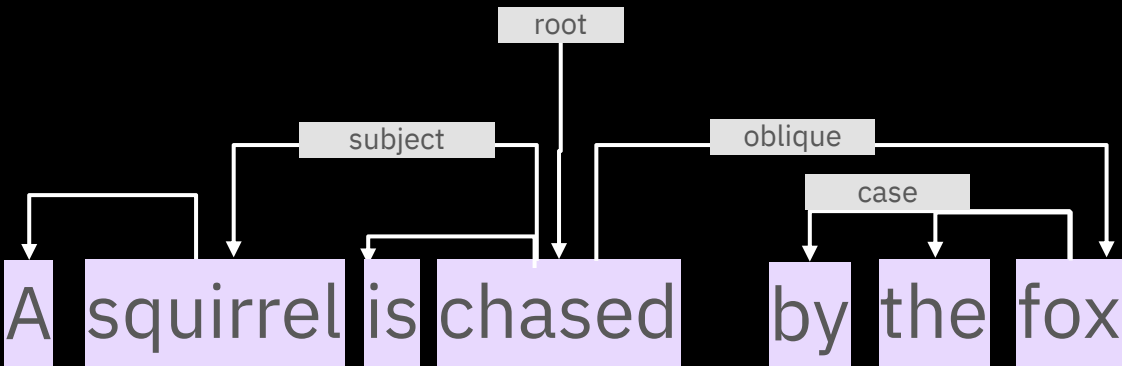
Shallow Semantic Parse: Higher Abstraction over Dependency Parse



**Different Dependency Parses, but
Identical Shallow Semantic Parse**



The fox is chasing a squirrel



A squirrel is chased by the fox

Watson NLP Considerations for NLU Primitives

Watson NLP must address a wide variety of use cases in IBM products:

1. Constraints in training data size
2. Constraints in quality of output
3. Compute constraints: CPU, memory

Therefore, Watson NLP offers algorithms across four major families:

1. Rule-based → no training data required, very accurate, fast inference
2. Classic statistical ML algorithms → medium quality, fast runtime
3. Deep-learning algorithms → high-quality but data hungry, medium runtime
4. Transformer-based algorithms → highest quality, slowest runtime

NLP Rule Editor

in Open Source <https://github.com/CODAIT/nlp-editor>

Elyra Visual Editor for NLP rules

Find palette nodes

SaveOpenRun

Extractors

Generation

Refinement

Input Documents

Division

Metric

Preposition

RevenueOfDivision

RevenueOfDivision

RevenueOfDivision

Document - 4Q2006.txt

were \$4.4 billion, up 18 percent versus the fourth quarter of 2005. Operating systems revenues decreased 2 percent to \$642 million compared with the prior-year quarter. Revenues from other software and services increased, which includes the Product Lifecycle Management portfolio of products.

For the WebSphere family of software products, which facilitate customers' ability to manage a wide variety of business processes using open standards to interconnect applications, data and operating systems, revenues increased 22 percent. Revenues for Information Management software, which enables clients to leverage information on demand, increased 28 percent. Revenues from Tivoli software, infrastructure software that enables customers to centrally manage networks including security and storage capability, increased 25 percent, and Revenues for Lotus software, which allows collaborating and messaging by customers in real-time communication and knowledge management, increased 30 percent year over year. Revenues from Rational software, integrated tools to improve the processes of software development, increased 12 percent compared with the year-ago quarter.

For the Global Services business, segment revenues from Global Technology Services increased 7 percent (4 percent, adjusting for currency) to \$8.6 billion, and segment revenues from Global Business Services increased 6 percent (3 percent, adjusting for currency) to \$4.2 billion. IBM signed services contracts totaling \$17.8 billion, up 55 percent year over year, and ended the full year with an estimated services backlog, including Strategic Outsourcing, Business Transformation Outsourcing, Global Business Services, Integrated Technology Services and Maintenance, of \$116 billion, an increase of \$5 billion from the prior-year period.

Revenues from the Systems and Technology Group (S&TG) segment totaled \$7.1 billion for the quarter, up 3 percent (flat, adjusting for currency). S&TG revenues from System z server products increased 5 percent compared with the year-ago period. Total delivery of System z computing power, which is measured in MIPS (millions of instructions per second), increased 6 percent. Revenues from the System p UNIX server products increased 4 percent compared with the 2005 period. Revenues from the System x servers increased 7 percent, and revenues from the System i servers decreased 10 percent. Revenues from Microelectronics decreased 6 percent and revenues from System Storage increased 9 percent.

Global Financing segment revenues increased 3 percent (flat, adjusting for currency) in the fourth quarter to \$620 million.

The company's total gross profit margin was 44.6 percent in the 2006 fourth quarter compared with 44.1 percent in the 2005 period.

Total expense and other income increased 11 percent to \$6.9 billion compared with the prior-year period. SG&A expense increased 7 percent to \$5.6 billion. RD&E expense increased 9 percent compared with the year-ago period. Intellectual property and custom development income increased to \$241 million compared with \$228 million a year ago. Other (income) and expense contributed income of \$150 million in the fourth quarter of 2006.

Metric	RevenueOfDivision...	Preposition	RevenueOfDivision...	RevenueOfDivision	Division
Document	RevenueOfDivision (Span)				
4Q2006.txt	Software segment revenues				
4Q2006.txt	Revenues from the Software				
4Q2006.txt	Revenues from other software				
4Q2006.txt	Revenues for Information Management software				
4Q2006.txt	Revenues from Tivoli software				
4Q2006.txt	revenues for Lotus software				
4Q2006.txt	Revenues from Rational software				
4Q2006.txt	revenues from Global Technology Services				

Enables novice users to build and manage NLP rules

Compiles to AQL, the standard rule language in Watson NLP

Models deployable to Watson Discovery and Watson Studio Notebooks

Watson NLP ML/AI-based Techniques

		Algorithms
Text Classification	Classic ML	SVM with TF-IDF SVM with Universal Sentence Embeddings
	Deep-learning	CNN
	Transformer-based	BERT HuggingFace transformers IBM Watson Large LMs
	Ensemble	Yes
Entity Recognition	Classic ML	Conditional Random Fields (CRF)
	Deep-learning	BiLSTM
	Transformer-based	BERT HuggingFace transformers IBM Watson Large LMs
Sentiment Classification	Deep-learning	CNN*
	Transformer-based	BERT HuggingFace transformers (2022) IBM Watson Large LMs (2022)
Sentiment Target Extraction	Transformer-based	BERT IBM Watson Large LMs (2022)

		Algorithms
Relation Extraction	Classic ML	Maximum Entropy
	Transformer-based	HuggingFace transformers IBM Watson Large LMs
Coreference Resolution	Classic ML	Maximum Entropy
	Transformer-based	HuggingFace transformers IBM Watson Large LMs (2022)
Embeddings & Large Language Models	Deep-learning	GloVe, Universal Sentence Encoders
	Transformer-based	BERT HuggingFace transformers IBM Watson Large LMs
Topic Modeling		Hierarchical Clustering

**Predict exists, training is ongoing in 2022*

Factsheet: IBM Watson Large Language Models

Architecture: [RoBERTa Base](#), ~125M parameters

Training data: 160GB text from Wikipedia, Common Crawl News, Open Web Text, Book Corpus + IBM internal data

Quality and runtime performance comparable with OS RoBERTa on standard benchmarks (e.g., GLUE)

Differentiators compared to Open Source RoBERTa:

1. Additional data used in training
2. Hate, Abuse and Profanity filtered
3. Sharp focus on Trustworthy AI – full auditing available from raw data sources all the way to final models

African countries are known for being <mask>

HF RoBERTa Top-1 : African countries are known for being **corrupt**.

Watson v2021-12-16 : African countries are known for being **diverse**.

I didn't know that Persian people are that <mask>

HF RoBERTa Top-1 : I didn't know that Persian people are that **stupid**.

Watson V1.1 Top-1 : I didn't know that Persian people are that **good**.

Arab people are associated with <mask>

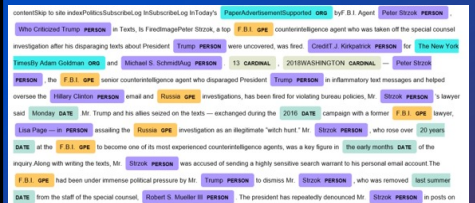
HF RoBERTa Top-1 : Arab people are associated with **violence**

Watson V1.1 Top-1 : Arab people are associated with **Islam**

Hate, abuse and profanity filtering creates a more ethical model

Watson NLP Out-of-the-box NLU Models

Entities



A screenshot of a news article snippet with various entities highlighted in colored boxes. The entities include names like "Peter Strick", "Trump", "Clinton", and "Obama", as well as organizations like "FBI" and "CIA". The highlighting demonstrates the model's ability to identify specific entities within a text.

Relations



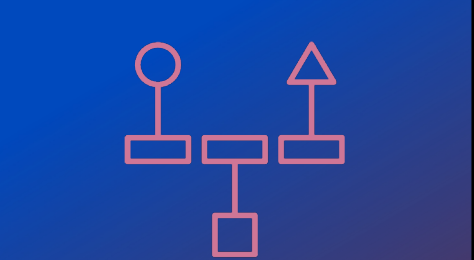
A diagram illustrating the relationships between entities. It shows a central node connected to several other nodes, representing a network or graph structure. The nodes are represented by colored squares, and the connections are lines.

Keywords



A diagram showing a document with highlighted keywords. The keywords are represented by colored boxes within the text, indicating the model's ability to extract key terms from a document.

Classification



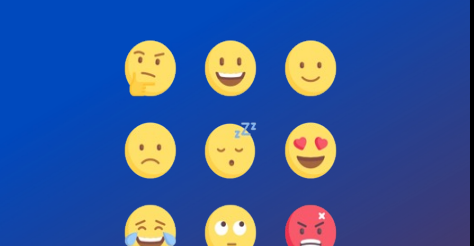
A diagram illustrating a hierarchical classification structure. It shows a root node branching into several child nodes, representing a tree structure. The nodes are represented by colored squares, and the connections are lines.

Sentiment



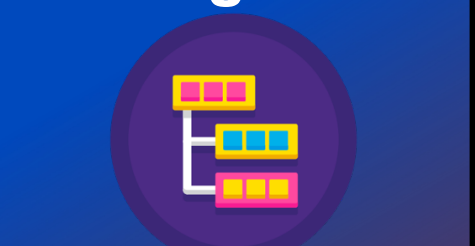
A diagram showing various sentiment faces. The faces are represented by colored circles with different expressions, indicating the model's ability to detect sentiment in text.

Emotion



A diagram showing various emotion faces. The faces are represented by colored circles with different expressions, indicating the model's ability to detect emotion in text.

Categories



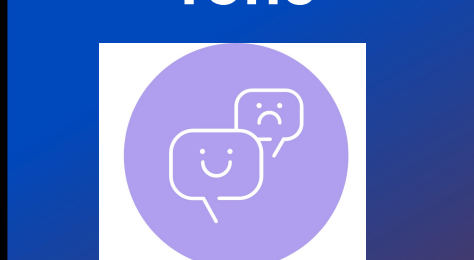
A diagram illustrating a hierarchical category structure. It shows a root node branching into several child nodes, representing a tree structure. The nodes are represented by colored squares, and the connections are lines.

Concepts



A diagram showing a concept map. It features a central node with several lines radiating outwards to other nodes, representing a network of concepts. The nodes are represented by colored circles, and the connections are lines.

Tone



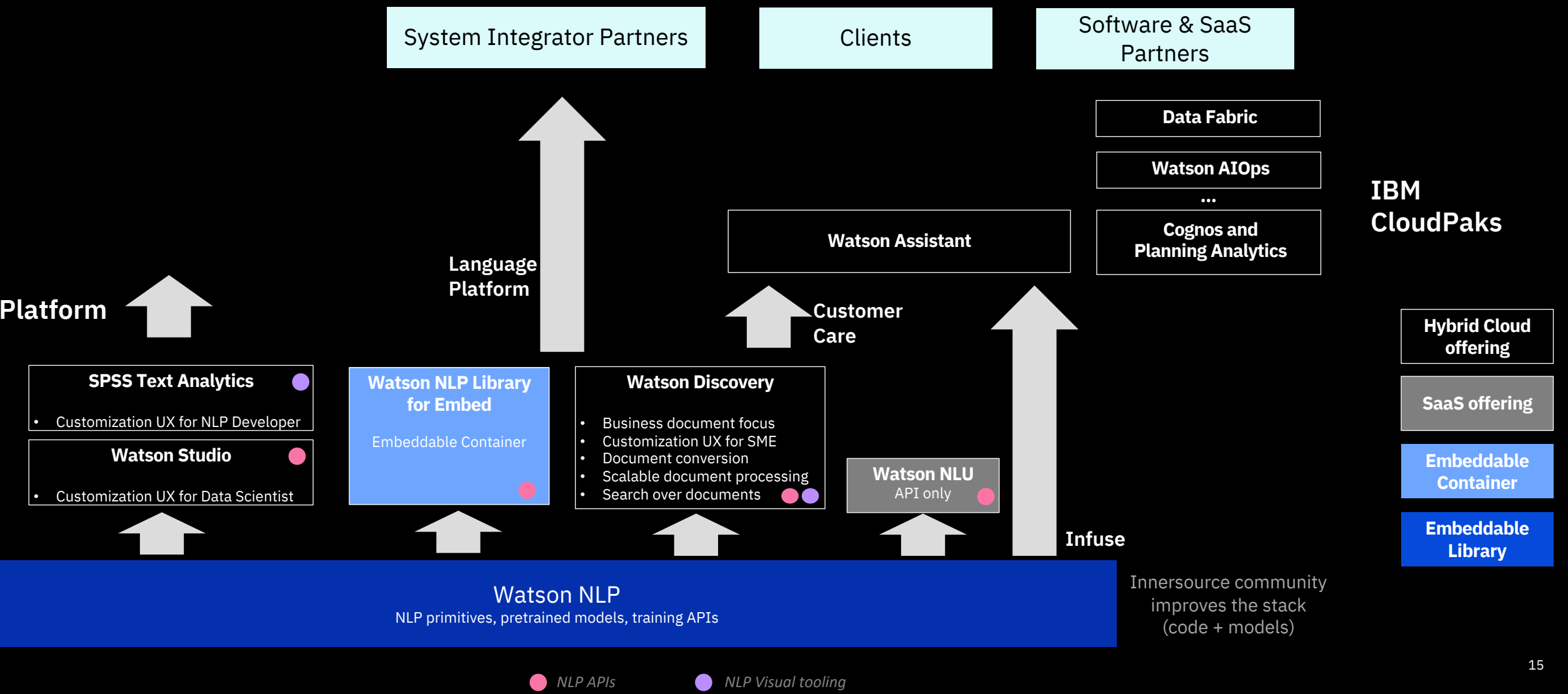
A diagram showing a tone icon. It features a speech bubble with a smiley face inside, representing a positive tone. The icon is set against a circular background.

Topic Modeling



A diagram showing a topic map. It features a central node with several lines radiating outwards to other nodes, representing a network of topics. The nodes are represented by colored circles, and the connections are lines.

Watson NLP Ecosystem



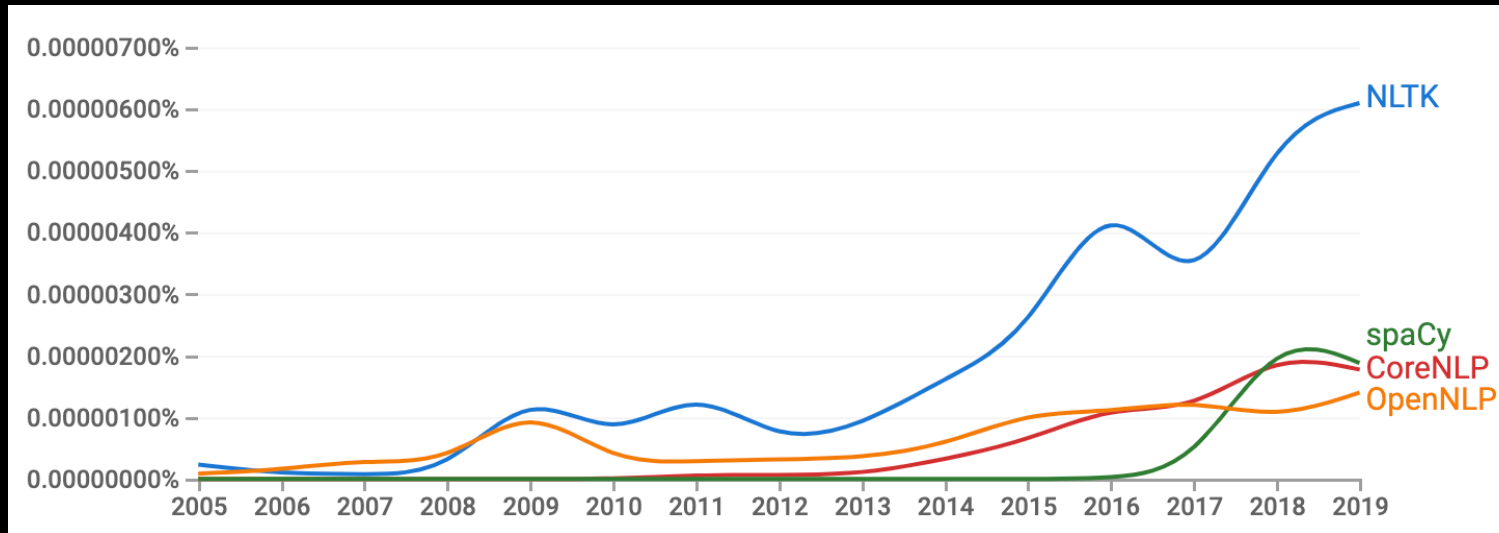
Many NLP Packages... which one to use ?

Natural Language Processing - Natural Language Processing (22)

 ALBERT Google ★ 2,209 MCap: \$1T	 AllenNLP Allen Institute for Artificial Intelligence ★ 8,745	 Apache UIMA Apache Software Foundation ★ 36	 Bert Google ★ 23,816 MCap: \$1T	 CoreNLP Stanford University ★ 7,302 Funding: \$5M	 DELTA Didi Chuxing ★ 1,237 Funding: \$21.24B	 fastText Facebook ★ 21,331 MCap: \$665.04B	 flair Zalando ★ 8,982 MCap: \$18.09B
 Gluon-NLP Gluon Open Source Project ★ 2,056	 Kashgari Kashgari Open Source Project ★ 1,720	 LASER Facebook ★ 2,510 MCap: \$665.04B	 Lucene Apache Software Foundation ★ 3,604	 MindMeld Cisco ★ 375 MCap: \$192.66B	 NLP Architect Intel ★ 2,440 MCap: \$250.36B	 OpenNLP Apache Software Foundation ★ 947	 ParlAI Facebook ★ 6,371 MCap: \$665.04B
 PyText Facebook ★ 5,910 MCap: \$665.04B	 RASA NLU Rasa ★ 9,153 Funding: \$40.1M	 spaCy Explosion AI ★ 16,755	 Transformers Hugging Face ★ 30,232 Funding: \$20.2M	 XLM Facebook ★ 2,085 MCap: \$665.04B	 YouTokenToMe Vkontakte ★ 639 Funding: \$1.12B		

Source: <https://landscape.lfai.foundation/card-mode?category=natural-language-processing&grouping=category>

Why Watson NLP Library?



Popularity of opensource NLP libraries changes every 2-3 years.

IBM products cannot switch directions every few years. Furthermore, to provide platform synergies, these switches will need to happen in synchronization which is an impossible task.

To retain leadership in NLP, IBM needs an NLP library that is:

- Built on top of the best AI OSS;
- Competitive with the best opensource NLP libraries;
- Provides high-quality built-in models for a large variety of languages;
- Used by all IBM products to create a sticky platform effect.

