NLP Use Case Database Project

A bottom-up qualitative NLP success/failure database

Etzard Stolte, PhD, Roche
Bikalpa Neupane, PhD, Takeda
Birthe Nielsen, PhD, Pistoia Alliance

NLPdatabase@pistoiaalliance.org

Collaborate to Innovate Conference Boston 2022

Agenda

- The team
- The problem
- The project
- What has been achieved to date
- What is planned
- Get involved

NLP Use Case Database

Project manager: Birthe Nielsen
nlpdatabase@pistoiaalliance.org

Problem Statement:

Pharma companies apply NLP methods in hopes of automation and insight generation. Although NLP algorithms have matured quite a bit during the past years, practical value for most NLP pilots tends to be poor, and very few NLP-driven projects are seen through to production. Exceptions are typically topics w/ good metadata quality, large training sets and willing business colleagues to verify results, and a serendipitous combination of technical expertise and suitable use cases.

Value Proposition:

This type of knowledge is of value to share in a pre-competitive manner among Pistoia Alliance members. A simple database could contain characterization of use case, data characterization, pipelines & algorithms used, quality criteria, outcomes, comments, etc. The value for participating members would be a reference database to help narrow down successful use case scenarios, less experimentation, and more successes.

Project member roles: Global Knowledge Management, Director of Advanced Technologies, Head of Text Analytics, Literature-based Scientific Discovery, Global Information and Analysis, NLP Lead/Engineer, Data Scientist

Project champion:

Etzard Stolte, Roche ; Bikalpa Neupane, Takeda.

Steering committee (Currently forming):

Roche Takeda

$5-15K ensures a seat on the Steering Committee

Project deliverables:

- A bottom-up qualitative NLP Success Failure Database
- Agreed annotations of NLP use case methods and success/failure criteria
- Collaborative insight into why NLP use cases may fail or succeed with an industry –wide view

Q1 2022
Roundtable Discussion

Q2 2022
Presentation of Use Cases

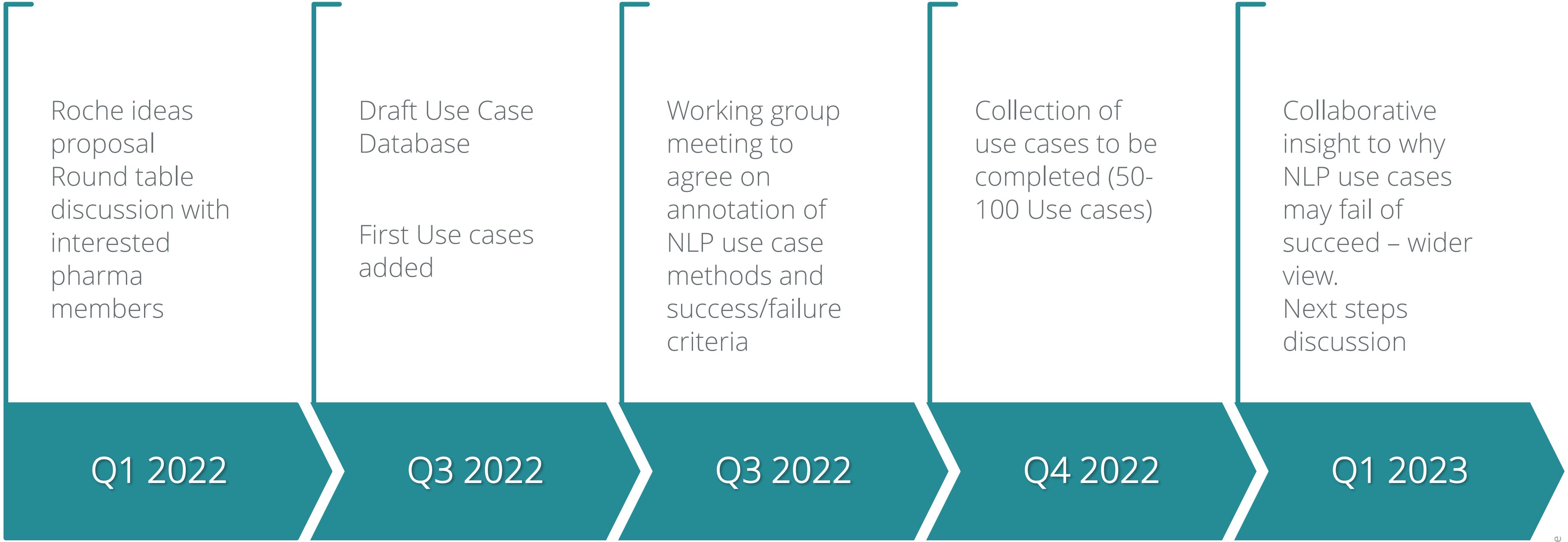
Q3 2022
Draft Database Created, First Use Cases Added

Q4 2022
Populate 50-100 Use Cases

2023
Standardize nomenclature, continue populating, analyze success prediction criteria

https://drive.google.com/file/d/1AkmetDVgCH92zBJ4HZjuVhSgn5H1XqZU/view?usp=sharing

Project timeline till date



Q1 2022

Roche ideas proposal
Round table discussion with interested pharma members

Q3 2022

Draft Use Case Database

First Use cases added

Q3 2022

Working group meeting to agree on annotation of NLP use case methods and success/failure criteria

Q4 2022

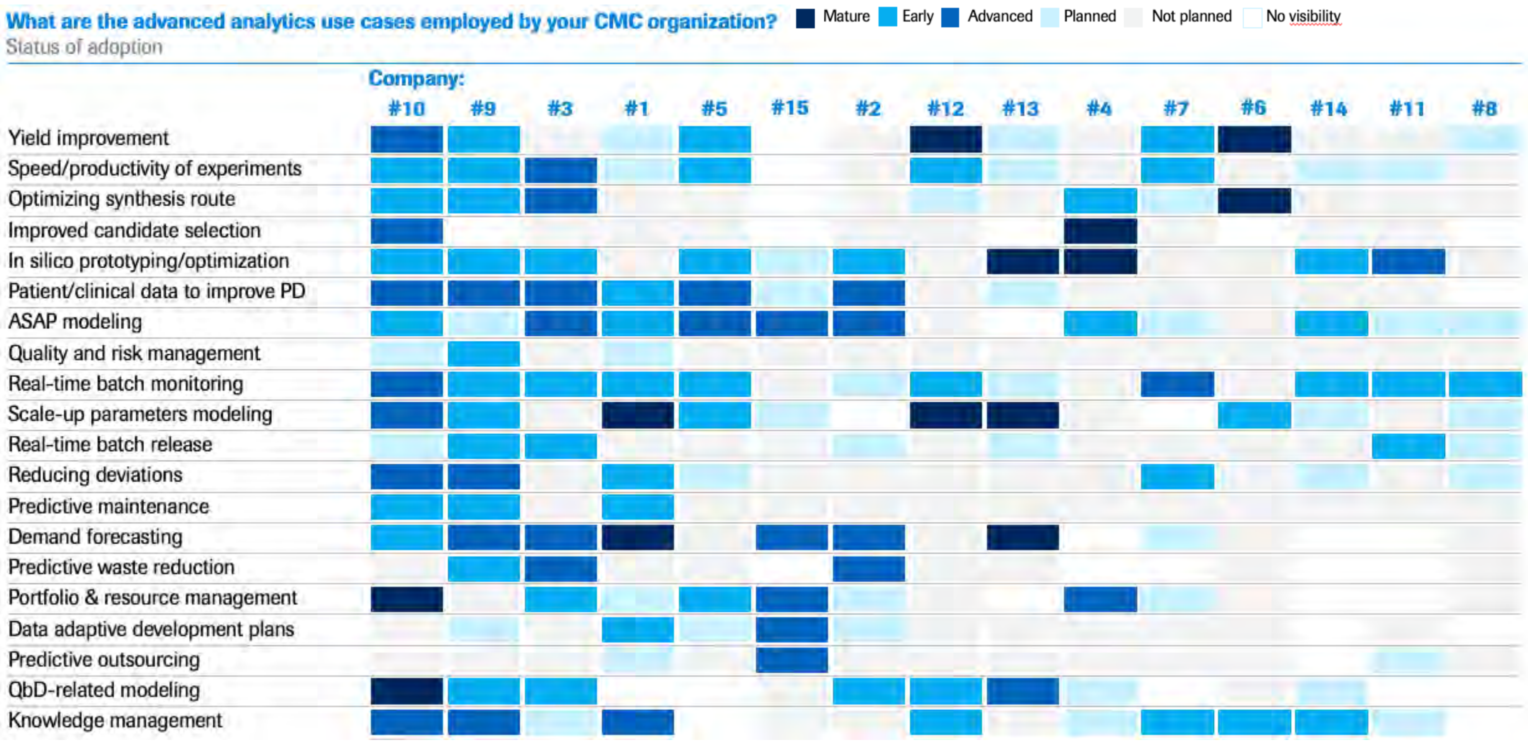
Collection of use cases to be completed (50-100 Use cases)

Q1 2023

Collaborative insight to why NLP use cases may fail of succeed – wider view.
Next steps discussion

Machine Learning (NLP, NER) uses cases in Manufacturing theory

Wide range of use cases



What are the advanced analytics use cases employed by your CMC organization?

Status of adoption

Mature Early Advanced Planned Not planned No visibility

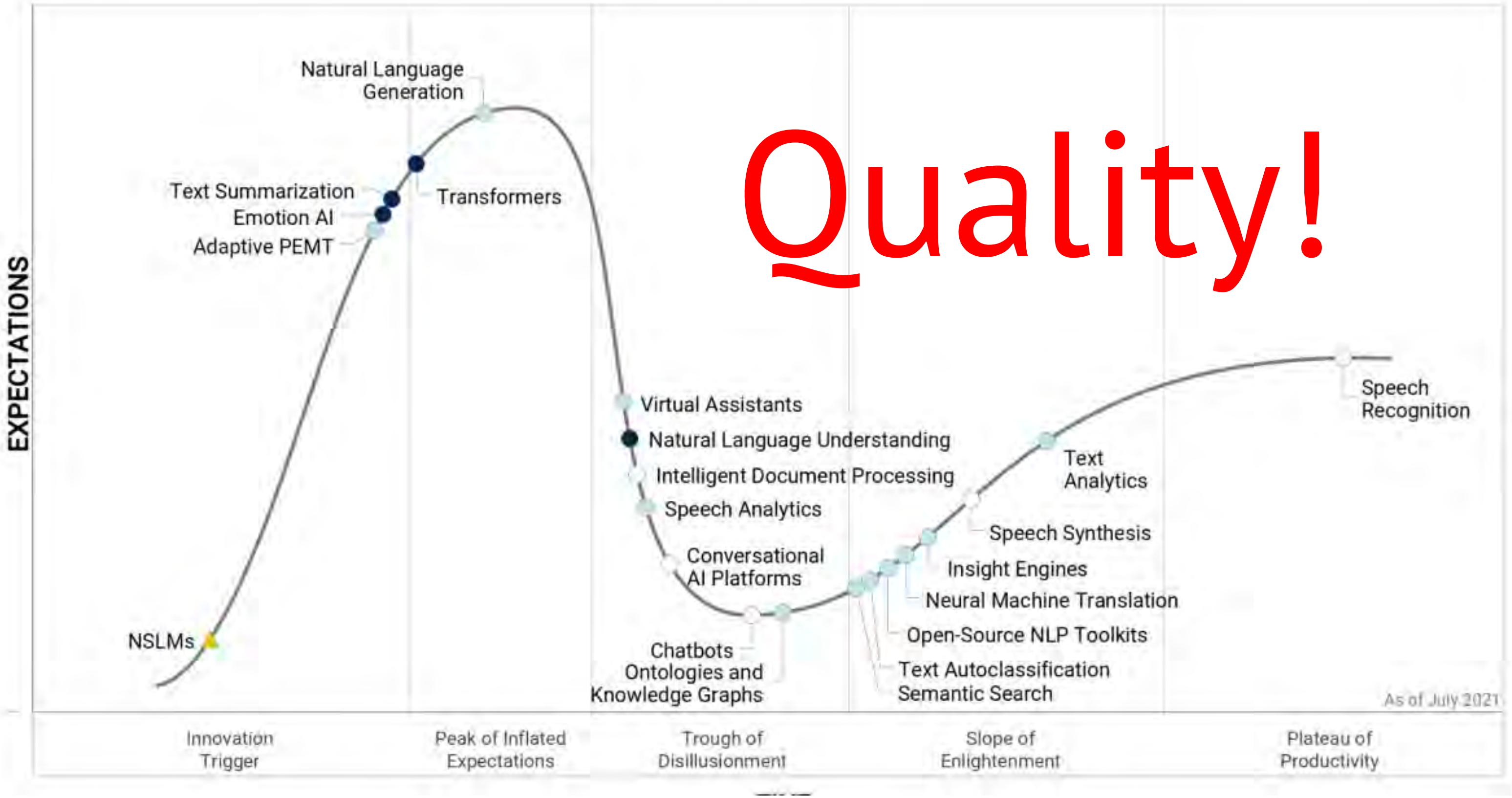
Company:

	#10	#9	#3	#1	#5	#15	#2	#12	#13	#4	#7	#6	#14	#11	#8
Yield improvement	Advanced	Early	Not planned	Planned	Early	No visibility	Not planned	Mature	Planned	Not planned	Early	Mature	Not planned	Not planned	Planned
Speed/productivity of experiments	Early	Early	Advanced	Planned	Early	No visibility	Not planned	Early	Planned	Not planned	Early	Not planned	Planned	Planned	Not planned
Optimizing synthesis route	Early	Early	Advanced	Not planned	Not planned	No visibility	Not planned	Planned	Not planned	Early	Planned	Mature	Not planned	Not planned	Not planned
Improved candidate selection	Advanced	No visibility	Not planned	Not planned	Not planned	Not planned	Not planned	Not planned	No visibility	Mature	Not planned	No visibility	Not planned	Not planned	No visibility
In silico prototyping/optimization	Early	Early	Early	Not planned	Early	Planned	Early	Not planned	Mature	Mature	Not planned	Not planned	Early	Advanced	Not planned
Patient/clinical data to improve PD	Advanced	Advanced	Advanced	Early	Advanced	Planned	Advanced	Not planned	Planned	Not planned	Not planned	Not planned	Not planned	Not planned	No visibility
ASAP modeling	Early	Planned	Advanced	Early	Advanced	Advanced	Advanced	Not planned	No visibility	Early	Planned	Not planned	Early	Planned	Planned
Quality and risk management	Planned	Early	Not planned	Planned	Not planned	Not planned	Not planned	Not planned	Not planned	Not planned	Not planned	Not planned	Not planned	Not planned	Not planned
Real-time batch monitoring	Advanced	Early	Early	Early	Early	Not planned	Planned	Early	Planned	Not planned	Advanced	Not planned	Early	Early	Early
Scale-up parameters modeling	Advanced	Early	Not planned	Mature	Early	Planned	No visibility	Mature	Mature	Not planned	No visibility	Early	Planned	Not planned	Planned
Real-time batch release	Planned	Early	Early	Not planned	Not planned	Not planned	Planned	Not planned	Planned	Not planned	Not planned	Not planned	Not planned	Early	Planned
Reducing deviations	Advanced	Advanced	Not planned	Early	Planned	Not planned	Not planned	Not planned	Not planned	Not planned	Early	Not planned	Planned	Not planned	Planned
Predictive maintenance	Early	Early	Not planned	Early	Not planned	Not planned	Not planned	Not planned	Not planned	Not planned	Not planned	Not planned	Not planned	Not planned	Not planned
Demand forecasting	Early	Advanced	Advanced	Mature	Not planned	Advanced	Advanced	Not planned	Mature	No visibility	Planned	Not planned	No visibility	No visibility	Not planned
Predictive waste reduction	Not planned	Early	Advanced	Not planned	Not planned	No visibility	Advanced	Not planned	Not planned	No visibility	Not planned	Not planned	No visibility	No visibility	Not planned
Portfolio & resource management	Mature	Not planned	Early	Planned	Early	Advanced	Planned	Not planned	No visibility	Advanced	Planned	Not planned	No visibility	No visibility	Not planned
Data adaptive development plans	Not planned	Planned	Not planned	Early	Planned	Advanced	Planned	Not planned	Not planned	Not planned	Not planned	Not planned	No visibility	No visibility	No visibility
Predictive outsourcing	Not planned	Not planned	Not planned	Planned	Not planned	Advanced	Not planned	Not planned	Not planned	Not planned	Not planned	No visibility	Not planned	Planned	Not planned
QbD-related modeling	Mature	Early	Early	No visibility	No visibility	Not planned	Early	Early	Advanced	Planned	No visibility	Not planned	Planned	No visibility	No visibility
Knowledge management	Advanced	Advanced	Planned	Advanced	No visibility	Not planned	Not planned	Early	Not planned	Planned	Early	Early	Early	Planned	No visibility

Hype Cycle for AI, ML, NLP

reality

Gartner, 2020: real world applications of AI/ML/NLP struggle with quality



Quality!

Natural Language Processing (NLP)

reality

Challenges ? Hell yes - it simply does not deliver the quality we need in Pharma

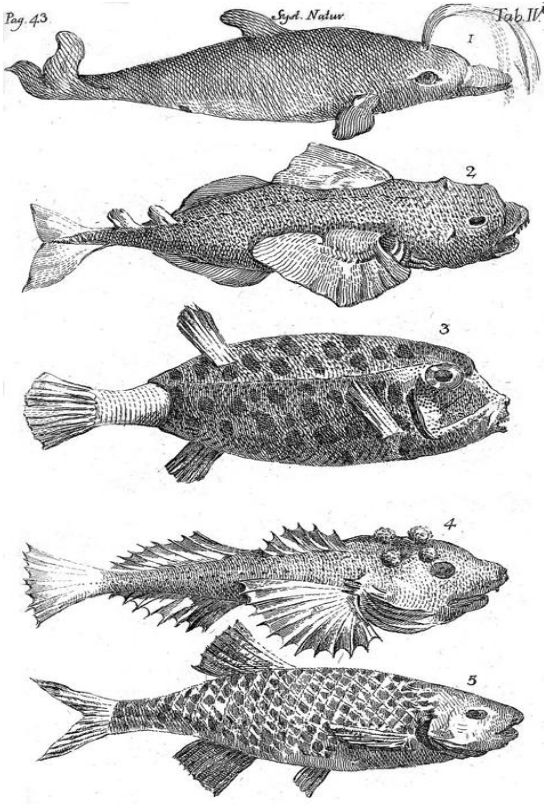
- Pharma is moving beyond buzzwords to value generation
- NLP is more of an art, then engineering
- Observations
  - Hard to determine up front, if a use-case will work
  - "successful" pipelines do not work for slightly different data
  - Manual curation, training sets are exceedingly work-intensive
  - Success seems very use case specific
  - Expensive to find talent that actually delivers the quality we need

Pistoia Alliance - NLP Expert Group

Value Proposition

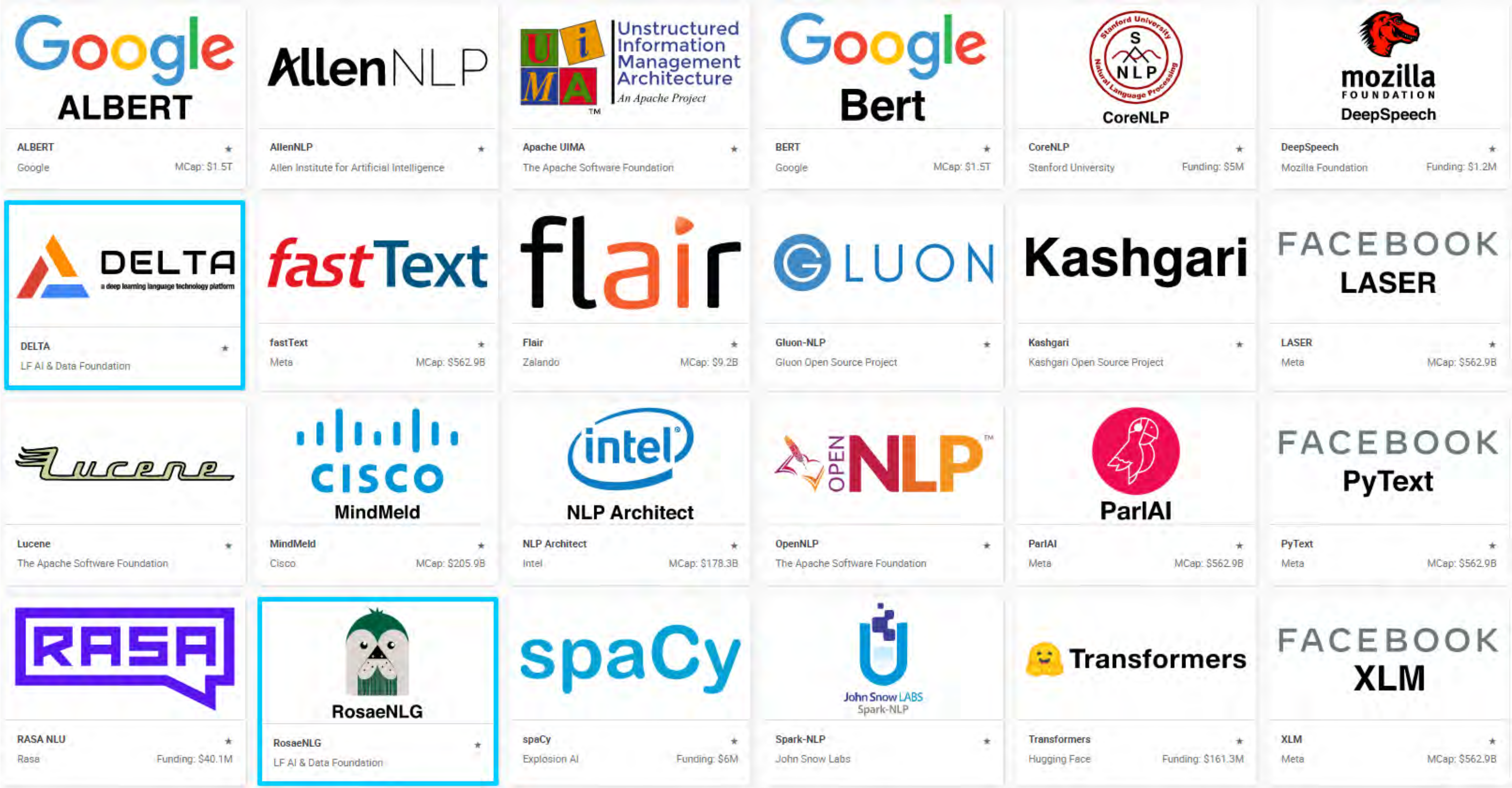
- Share best-practice – get an idea what works & what does not work
- Build database of use cases to be shared in this group; incl. failures

- Obvious challenges to address
  – Can we come up with a categorization that makes sense?
  – If NLP is an art, can we structure use-cases in a meaningful way?
  – Is a phenotypic classification sufficient (aka like Linne)?
  – Can we deliver on the effort to identify and enter your use cases iteratively into some database?



Systema naturæ sistens regna tria naturæ, in classes et ordines, genera et species redacta tabulisque æneis illustrate, Systema Naturae, 1768, Carl von Linné

Natural Language Processing Landscape



ALBERT Google MCap: $1.5T	AllenNLP Allen Institute for Artificial Intelligence	Apache UIMA The Apache Software Foundation	BERT Google MCap: $1.5T	CoreNLP Stanford University Funding: $5M	DeepSpeech Mozilla Foundation Funding: $1.2M
DELTA LF AI & Data Foundation	fastText Meta MCap: $562.9B	Flair Zalando MCap: $9.2B	Gluon-NLP Gluon Open Source Project	Kashgari Kashgari Open Source Project	LASER Meta MCap: $562.9B
Lucene The Apache Software Foundation	MindMeld Cisco MCap: $205.9B	NLP Architect Intel MCap: $178.3B	OpenNLP The Apache Software Foundation	ParlAI Meta MCap: $562.9B	PyText Meta MCap: $562.9B
RASA NLU Rasa Funding: $40.1M	RosaeNLG LF AI & Data Foundation	spaCy Explosion AI Funding: $6M	Spark-NLP John Snow Labs	Transformers Hugging Face Funding: $161.3M	XLM Meta MCap: $562.9B

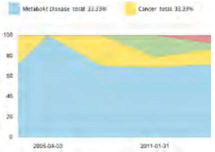
Source: https://landscape.lfai.foundation/card-mode?category=natural-language-processing&grouping=category

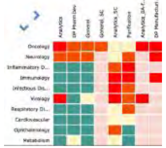
Meaning Based Computing

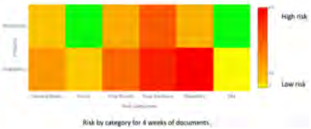
use cases

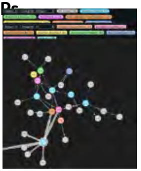
Use Cases types and examples

	Concepts	Trend Detection	Quality Controls	Risk Prediction	Cause Analysis
Type	*Automatically extract and correlate concepts using semantic tools*	*Interactively explore emerging concepts / trends*	*Monitor operational & content quality measures hidden in un-structured information*	*Real-time classifiers to sense risks across sources, sites & organizations*	*Extract relationships among concepts in a Knowledge Graph*
Examples	• find documents/experts • Automatically add quality tags to lab analysis • Show potentially recurring deviations for new TrackWise entries	• sense patterns in health authority records • early warning signs for batch or release records • off-label use of products	• detect mis-filed documents in submission structures • flag spurious relationships in manufacturing logs	• sentiment analysis of HA ePARs with respect to submission success • emerging issues in batch records	• understand cause / effect relationships in manufacturing deviations • visualise relationships / dependencies inside SOPs









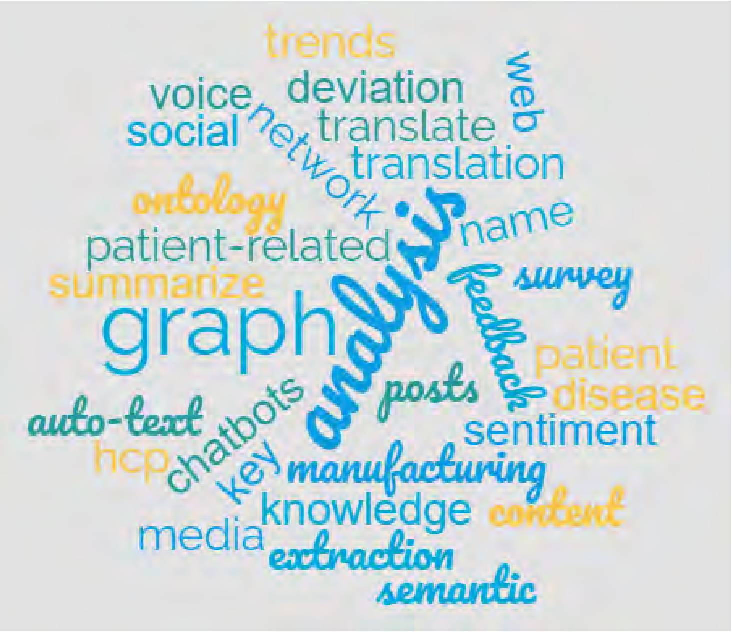
The Use Case Database

Contributions till date:

Pfizer
Roche
Pistoia Alliance
Takeda
VERTEX
CSL Behring
Biotherapies for Life™
MERCK
BAYER

Where our current database stands today

- Named-entity recognition (NER)
- Concept Extraction (Taxonomy extraction, concept tagging, document classification)
- Text-Classification
- Unsupervised topic modeling
- Extractive auto-summarization
- Similarity Search
- Chatbots/Virtual Assistants



Database content

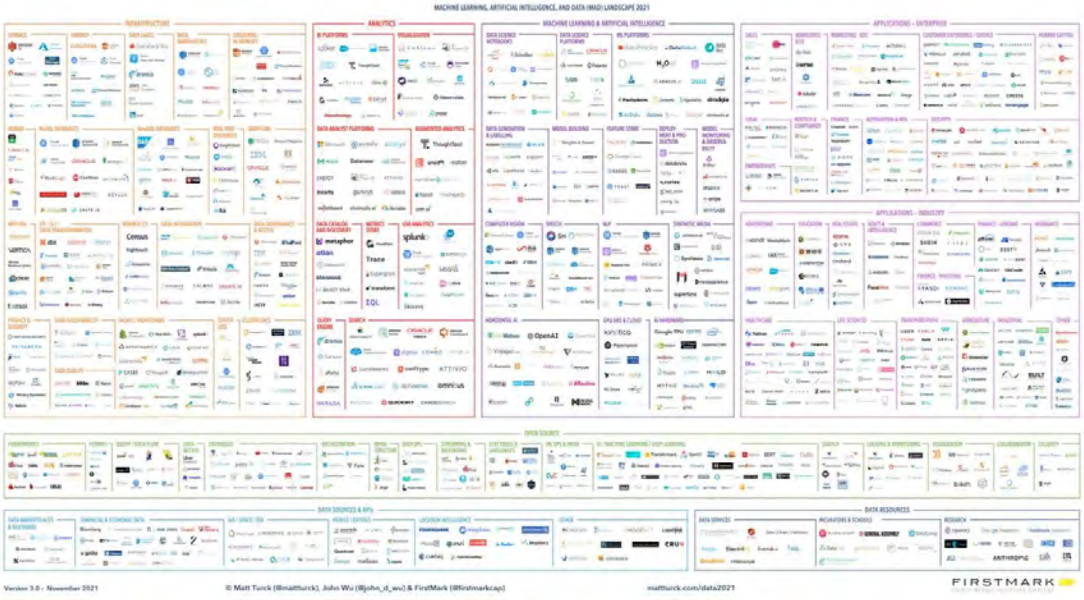
- Use case
  - Deployment type
  - Project level
  - Company
  - Language
  - etc
- Data
  - Description
  - Size
  - Issues
  - Preparation
  - FTE involved/team
  - etc
- Algorithm
  - Summary
  - Version
  - Library
  - etc
- Outcome
  - Learnings
  - Success achieved
  - etc
- Discussions
  - Effort/ benefit ratio
  - Non-NLP approach
  - Success
  - Support
- Outlook
  - Build vs buy recommendation
  - Other applications of technology
  - etc

Pistoia Alliance - NLP Expert Group

worry

What do we **NOT** want to do

- Encyclopedia of NLP methods
- Another NLP discussion group
- Vendor messages, vision documents, general statements
- Evangelism of a particular NLP school of thought or technology
- ???


MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, AND DATA (MAD) LANDSCAPE 2021
INFRASTRUCTURE
ANALYTICS
MACHINE LEARNING & ARTIFICIAL INTELLIGENCE
APPLICATIONS – ENTERPRISE
APPLICATIONS – INDUSTRY
OPEN SOURCE
DATA SOURCES & APIs
DATA RESOURCES
Version 3.0 - November 2021
© Matt Turck (@mattturck), John Wu (@john_d_wu) & FirstMark (@firstmarkcap)
mattturck.com/data2021
FIRSTMARK

https://mattturck.com/, November 2021 – Machine Learning & Data Landscape

NLP Build vs Buy

- ❑ *Language is hard – ambiguous,* Challenges because human expression is diverse, often ambiguous, affected by age, demographics, socio-economics, medium/channel & regional attributes of the author.
- ❑ Building a POC NLP model is trivial. However, gathering domain specific dataset, training large models and deploying them reliability (serving thousands of requests at low latency) is extremely hard.
- ❑ Running Experiments in Parallel and Serving Models is computationally expensive
- ❑ Market Barriers – Lack of Clean and accurate data - Lack of human-experts;
- ❑ Compliance is not trivial (PII, PHI, GDPR, FERPAA) – Customizability, AI Explainability, and Fairness are daunting challenges to fulfill without a large, in-house data science team
- ❑ Head Count + Maintenance Costs + Infrastructure Costs ➔ ROI ?

What if you were to purchase, checklist

❑ Return on Investment (ROI)
❑ Domain Fit Accuracy of the results
❑ Production Readiness - Scalability and Performance, Deployment Options
❑ Interoperability and Fit with existing software stack
❑ Integration with existing ML pipeline
❑ Flexibility - NLP custom models
❑ Usability
❑ Industry Leaders – Look at who is doing the real NLP work
❑ Compare and contrast regular use cases

We are eager to have you….

More information here:

- https://www.pistoiaalliance.org/projects/current-projects/natural-language-processing-use-case-database/

Get in touch:

- nlpdatabase@pistoiaalliance.org

Questions

nlpdatabase@pistoiaalliance.org