AI SWEDEN

Data Readiness Lab

Databeredskapsverkstaden

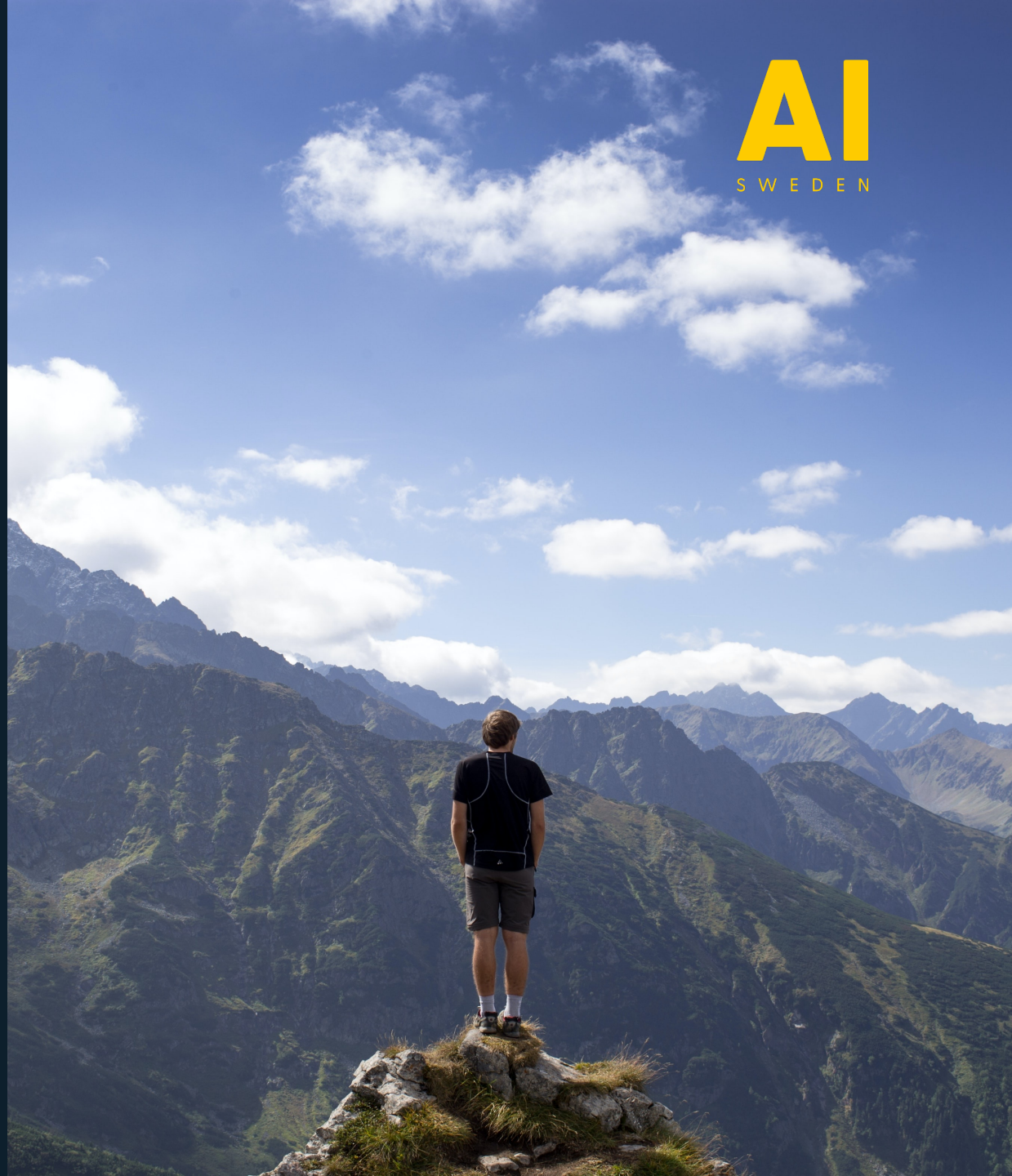
Referensgruppmöte (2022-06-08)

Background

- October 2021 – October 2023
  Funded by VINNOVA

- Data Readiness is
  - an overlooked problem
  - often poorly understood
  - a huge bottleneck for the application of AI (Swedish Language Models)

- Project focuses on
  - the public sector
  - text data



Project Partners

AI SWEDEN

EKONOMISTYRNINGSVERKET

SCB

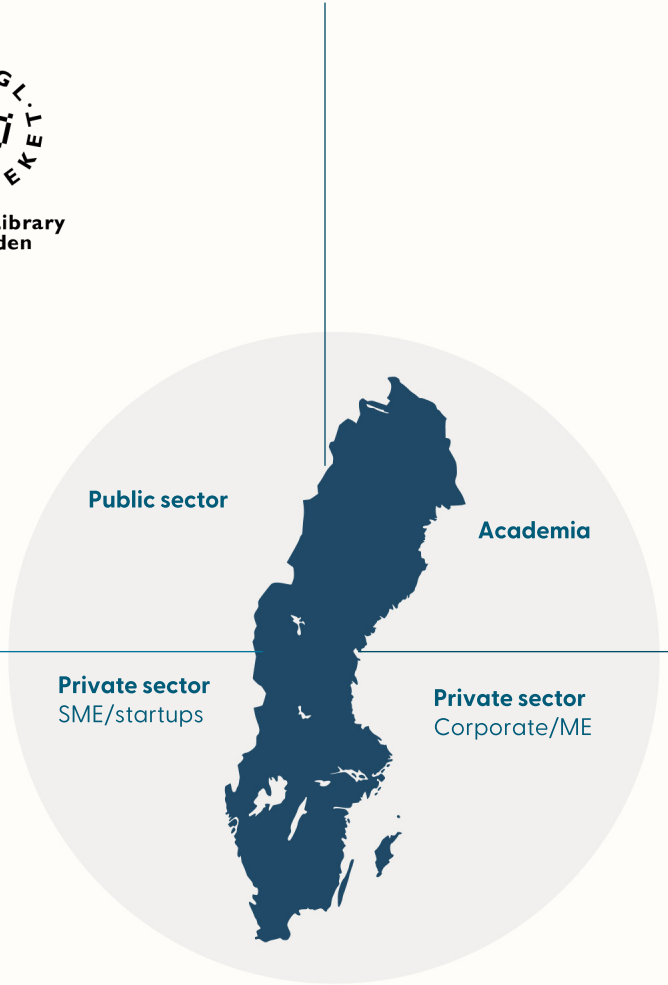
National Library of Sweden

ARBETSFÖRMEDLINGEN
SWEDISH PUBLIC EMPLOYMENT SERVICE

STRÄNGNÄS KOMMUN

Sveriges Kommuner och Regioner

Polisen



Peltarion

Gavagai

Challenges


Evaluation
Legal demands (e.g. GDPR)
Data
annotated
Applications
Fine-tuning (Unsupervised)
Fine-tuning (Supervised)

Challenges

You want to apply a language model, but it turns out the data..

- .. is hard to get access to
- .. is not in the format you expected
- .. does not meet your expectations w.r.t. quantity
- .. does not meet your expectations w.r.t. quality (noise, missing values, ..)

Data Readiness

- Availability
- Validity
- Utility

Challenges

You want to apply a language model, but it turns out the data..

- .. annotation quality is insufficient (=> poor model quality)
- .. annotation takes too much time (= data is not enough)

Annotation

- Quality
- Efficiency
- Reusability

Challenges

You want to apply a language model, but it turns out the data..

- .. contains personal information

Anonymization

- Hands-on Application

Challenges

You want to apply a language model, but it turns out the data..

- the later you find out, the worse it is!

Evaluation

- Data Quality Assessment
- Test Data

Challenges

Evaluation
- Test Data
- Quality Assessment

Anonymization
- Hands-on Application

Data Readiness
- Availability
- Validity
- Utility

Annotation
- Quality
- Efficiency
- Reusability

Project Goal

AI SWEDEN

> Att i stor skala möjliggöra ökad databeredskap hos behovsägare genom att tillhandahålla verktyg, ramverk och resurser.

Project Goal

We share
- problems
- solutions

.. but not the data!

Data Readiness Lab

Sveriges Kommuner och Regioner

Strängnäs kommun

Arbetsförmedlingen Swedish Public Employment Service

Ekonomistyrningsverket

Polisen

Peltarion

Gavagai

Kungl. Biblioteket National Library of Sweden

SCB

Work packages & Case Studies

WORK PACKAGE
AP 1 - Tools, Frameworks, Guidelines
AP 2 - Knowledge sharing, Education, Communication

Needs ↑ ↓ Application

CASE STUDY
CS 1 - Strängnäs
CS 2 - SKR
CS 3 - SCB
CS 4 - Polismyndigheten
CS 5 - ESV

Data Readiness

Annotation

Anonymization

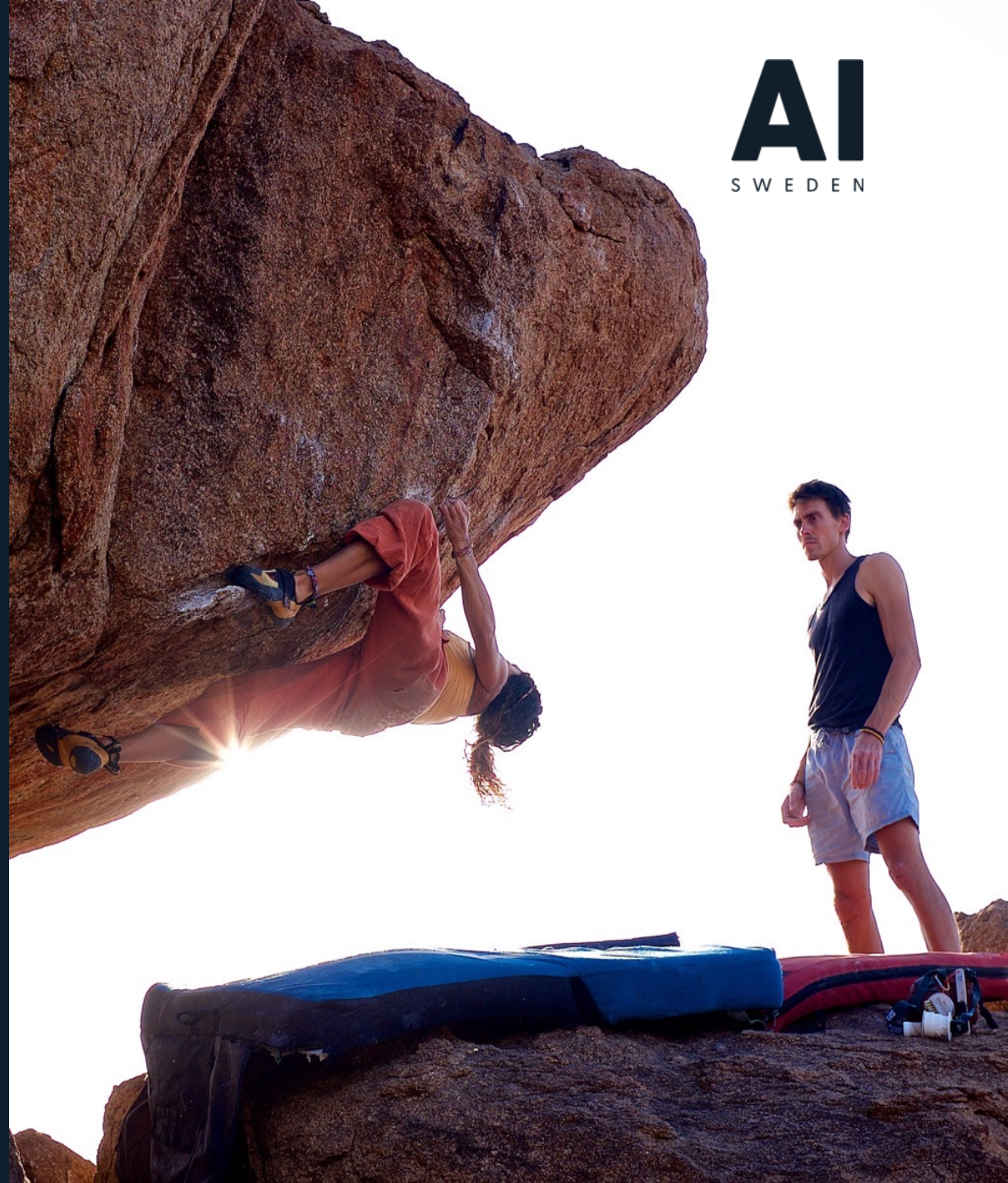
Evaluation

Data Readiness

First Step: Data Readiness Assessment

- Method: “We need to Talk About Data: The Importance of Data Readiness in Natural Language Processing”
  (Olsson & Sahlgren - 2021)

- Application to project partners



Data Readiness

AI SWEDEN

Goal

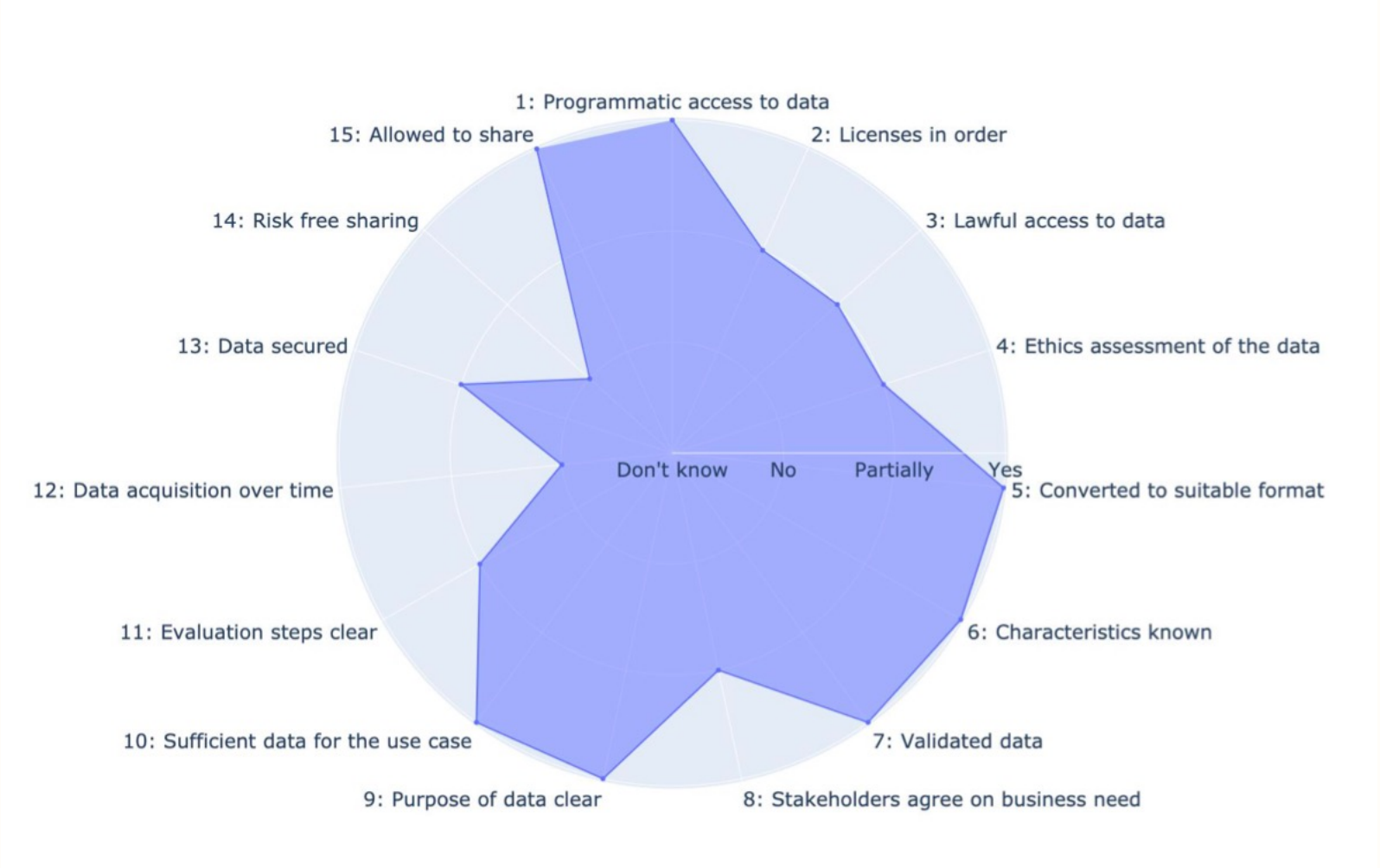
Band	Level	Meaning	The state of data
A	*A-1*	*Utility*	Ready for analysis with respect to given business objectives, hypotheses, questions, or context.
	...		
B	*B-1*	*Validity*	Ready to define the candidate questions and hypotheses. Includes exploratory analysis, data characterization, entity disambiguation, and de-duplication, etc.
	...		
C	*C-1*	*Accessibility*	Ready to be loaded in analysis software. Includes programmatic access, format conversions, and legal aspects, etc.
	...		
	C-n	*Hearsay data*	*"I'm sure AI can solve this problem. We have lots of data!"*

Start

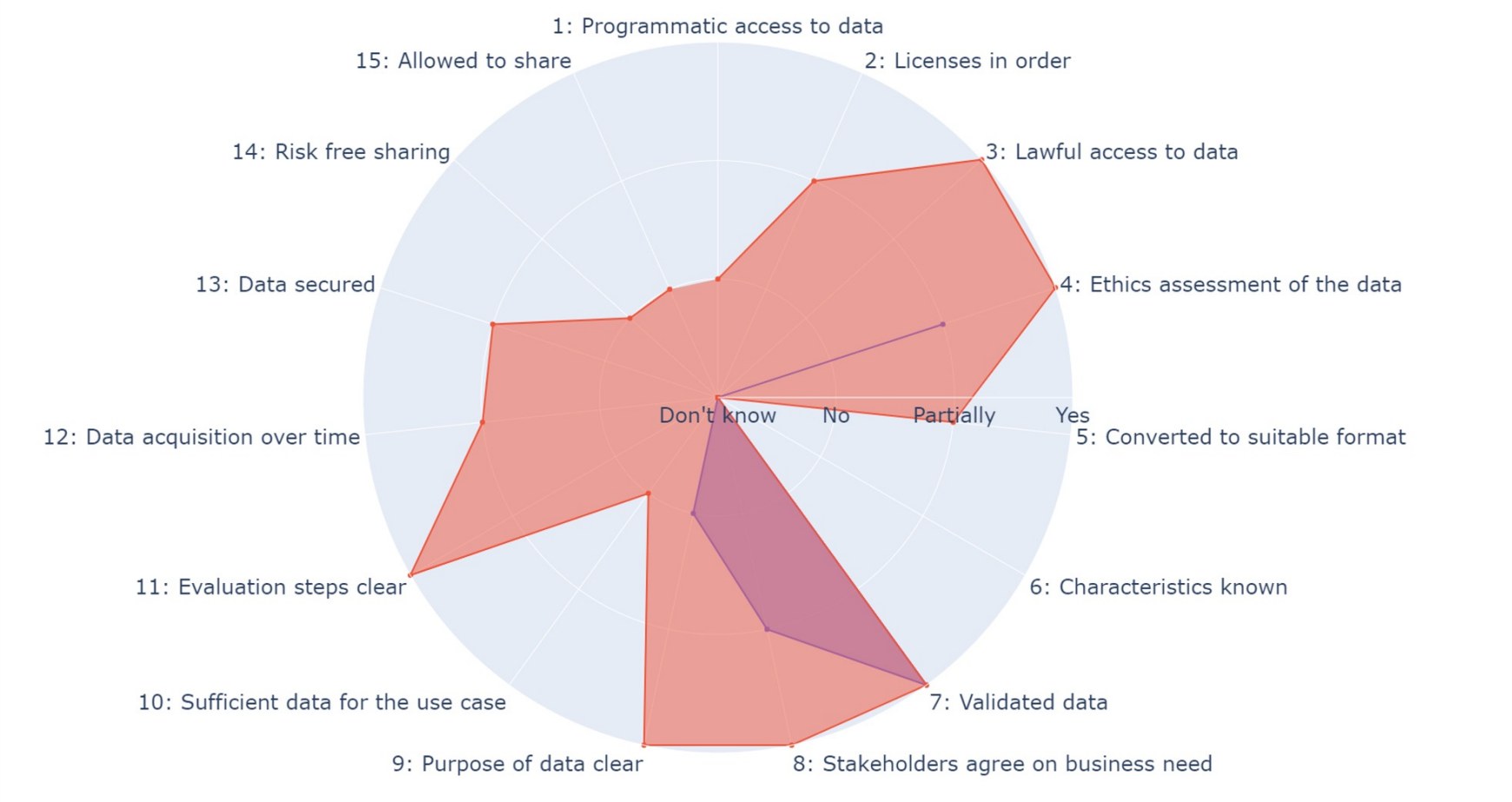
Lawrence (2017) "Data Readiness Levels".

Gavagai

Data Readiness

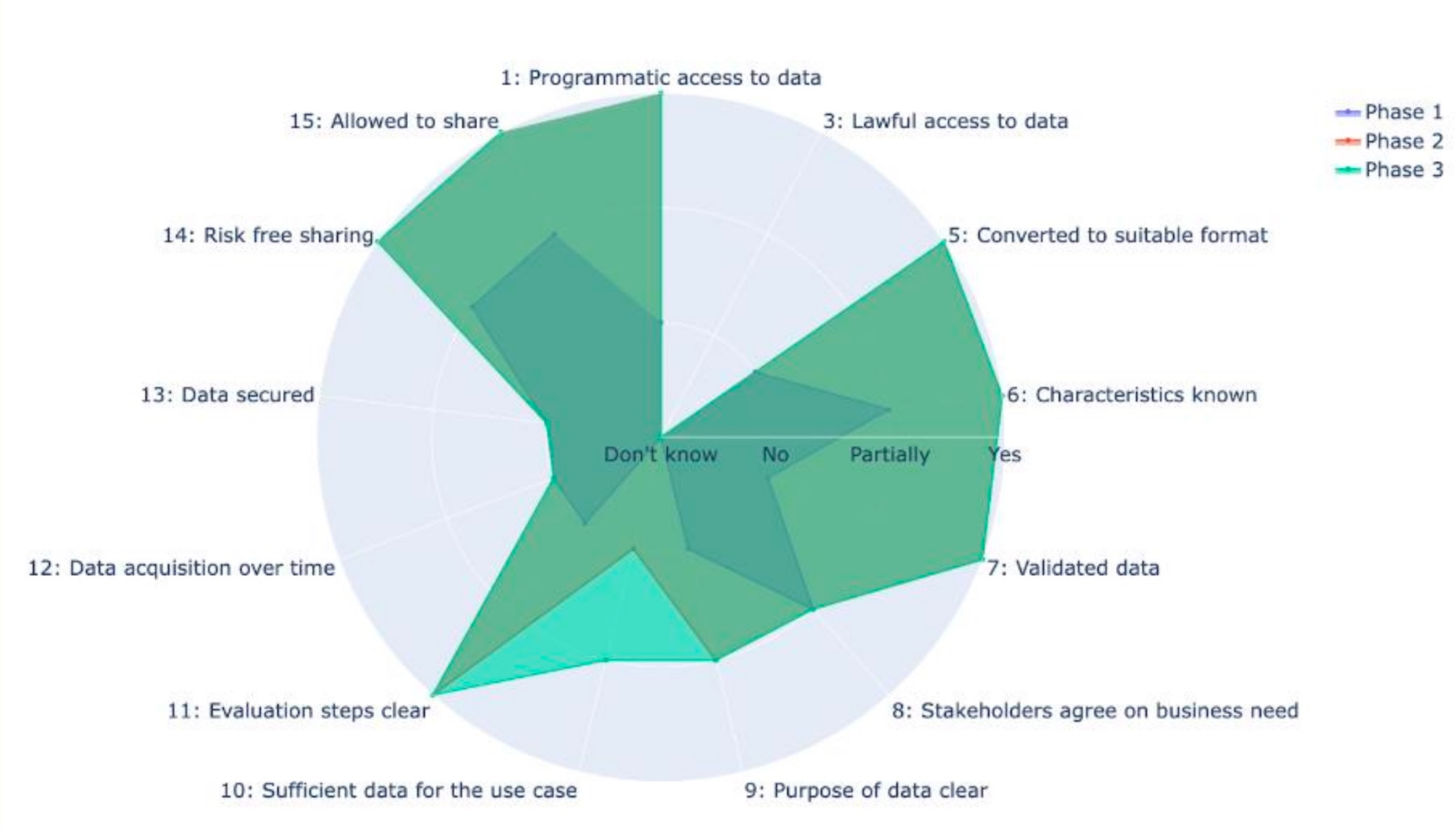


Data Readiness



STRÄNGNÄS KOMMUN

Data Readiness



Data Readiness

First Step: Data Readiness Assessment

- Method: “We need to Talk About Data: The Importance of Data Readiness in Natural Language Processing” (Olsson & Sahlgren - 2021)
- Application to project partners

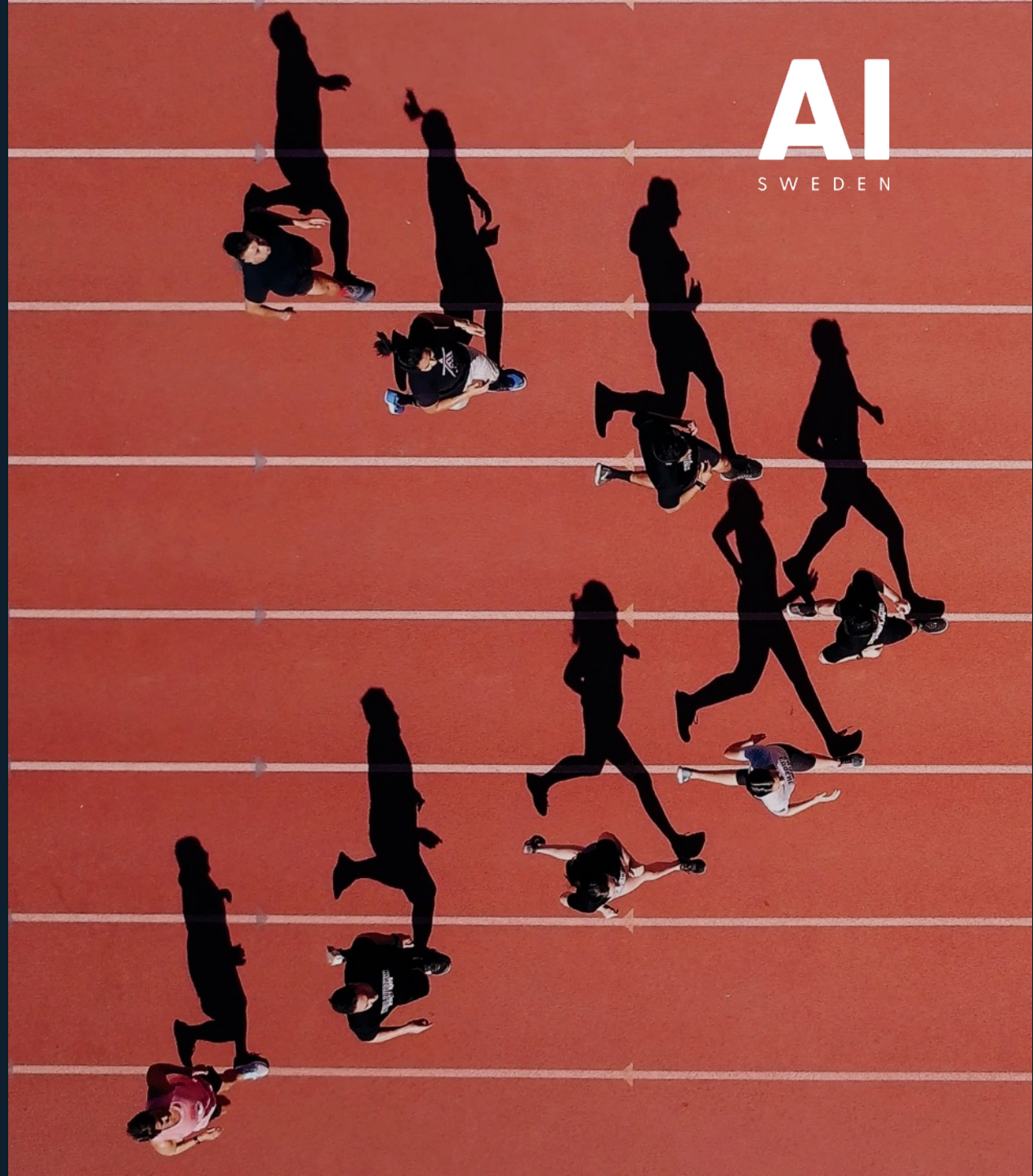
Next Steps:

- Identification of major needs
- Creation of guidelines and tools for improved data readiness

Annotation

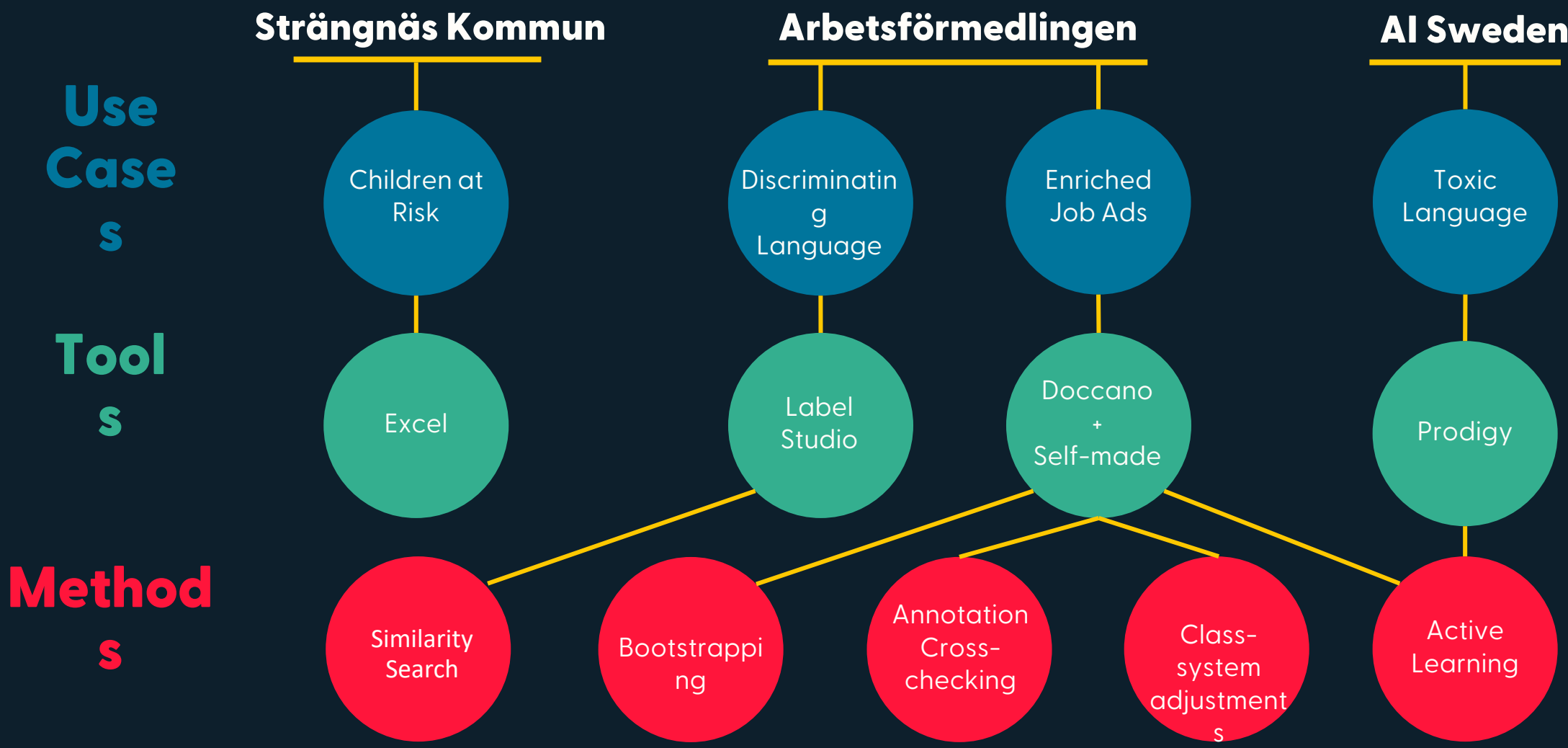
First Step: Audit of Tools & Methods

- Project Partners: Use Cases & Experiences
- Feature Comparison of Text Annotation Tools (Olsson, https://nlp-data-readiness.readthedocs.io)
- Job Ad Annotation at Arbetsförmedlingen (Stollenwerk, Fastlund, Nyqvist, Öhman – 2021)



Annotation

- Project Partners: Use Cases & Experiences



Annotation

- Feature Comparison of Text Annotation Tools (Olsson, https://nlp-data-readiness.readthedocs.io)

Tool	Prodigy	Label studio	Doccano
Website	https://prodi.gy/	https://labelstud.io/	https://doc
Tagline	"Radically efficient machine teaching. An annotation tool powered by active learning."	"Open Source Data Labeling Tool — Simplicity built-in, no overcomplicated UIs — Supports different datatypes — Visually configurable from top to the bottom"	"Text annc Just create data and s You can b hours."
SaaS	No	No	No
Self hosted	Yes, local server with web interface for annotation.	Yes	Yes
Active Learning	Yes, NER, text classification, dependecy parsing, pos tagging	No, but it can probably be implemented.	No, but it c implement
Multiple annotators for same data	Yes	No, but maybe.	Yes

Annotation

- Job Ad Annotation at Arbetsförmedlingen
  (Stollenwerk, Fastlund, Nyqvist, Öhman – 2021)

METHOD	PURPOSE
Bootstrapping	Acceleration
Active Learning	Acceleration
Annotation Cross-checking	Quality Assurance
Class system Adjustments	Quality Assurance

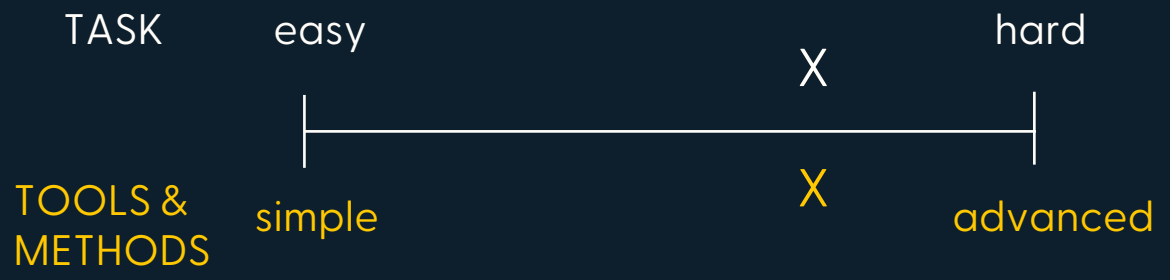
Annotation

First Step: Audit of Tools & Methods

- Project Partners: Use Cases & Experiences
- Feature Comparison of Text Annotation Tools (Olsson, https://nlp-data-readiness.readthedocs.io)
- Job Ad Annotation at Arbetsförmedlingen (Stollenwerk, Fastlund, Nyqvist, Öhman – 2021)

Next Steps:

- Feature Comparison of Text Annotation Tools: Update
- Creation of Guidelines



AI SWEDEN

Contact:

felix.stollenwerk@ai.se