TAUS 2020

Powering Language Data

DATA FOR MT IMPROVEMENT

Accelerate the growth of your business by boosting the performance of your machine translation

DATA FOR BUSINESS INTELLIGENCE

Eliminate the guesswork by tracking and benchmarking your translation productivity and quality

Data for MT Improvement

Low-resource Languages Data

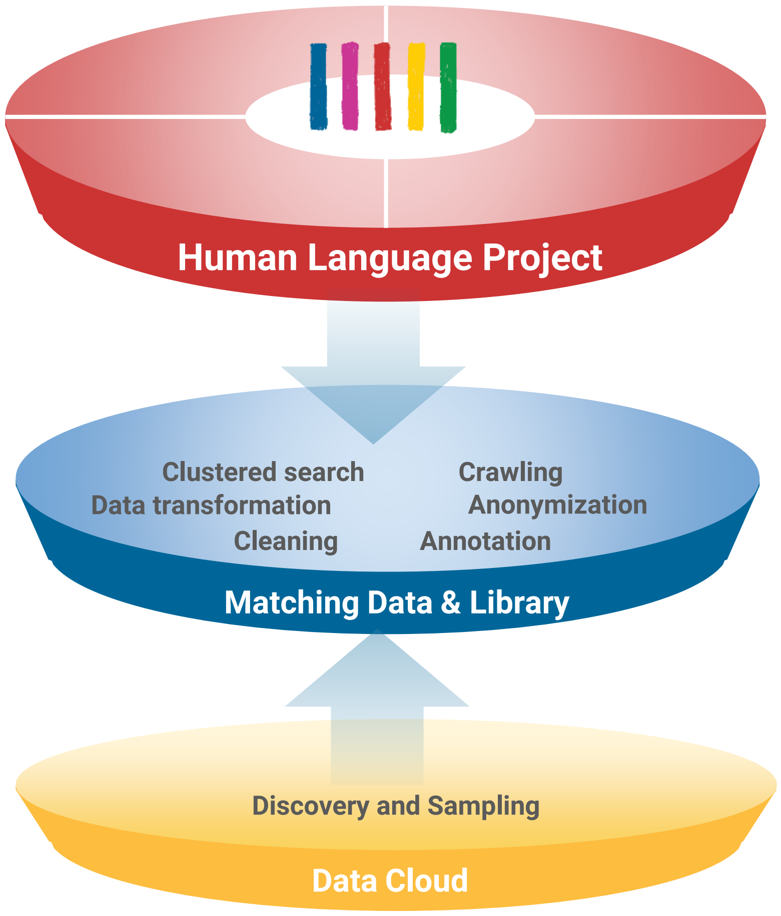
Data sets created by native speakers of rare languages.

High-fidelity Data

Matched corpora, based on user-provided query data.

Data Pool

Parallel data repository with data from member uploads.

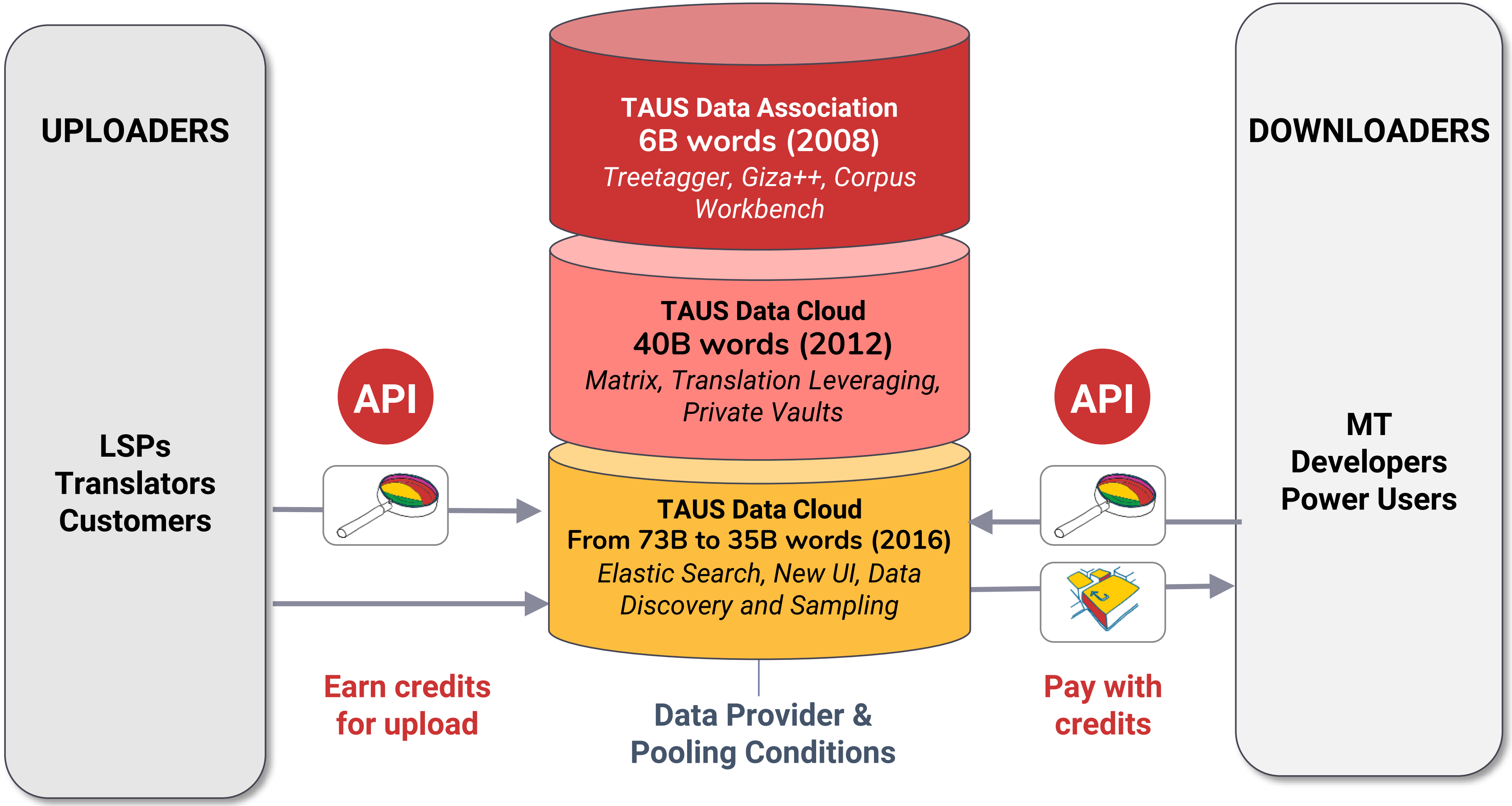


- Focus on the long-tail locales
- Crowdsourcing approach
- Available as a service or ready-made corpora
- TAUS owned data

- Segment-level matching
- Data selection based on relevance to the query data
- Includes cleaning and anonymization
- Available as a service or ready-made corpora
- Governed by DPPC

- Document-level search
- User-provided meta data
- Reciprocal upload/download or data purchase
- Governed by DPPC

Data Cloud 2008-2016



UPLOADERS

LSPs
Translators
Customers

API

Earn credits
for upload

TAUS Data Association
6B words (2008)
Treetagger, Giza++, Corpus Workbench

TAUS Data Cloud
40B words (2012)
Matrix, Translation Leveraging, Private Vaults

TAUS Data Cloud
From 73B to 35B words (2016)
Elastic Search, New UI, Data Discovery and Sampling

Data Provider &
Pooling Conditions

API

Pay with
credits

DOWNLOADERS

MT
Developers
Power Users

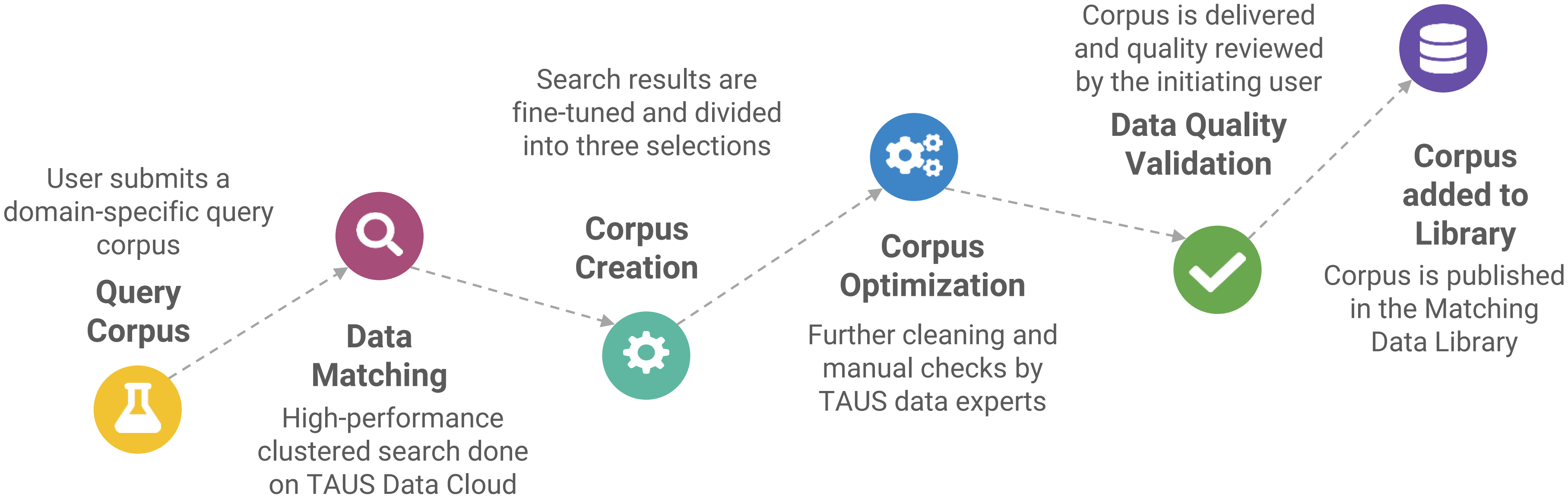
Explore the Data Cloud

Matching Data

Sophisticated clustered search technology to create high-fidelity data sets.

Creation of a Customized Corpus

From Matching Data to Matching Data Library



User submits a domain-specific query corpus

Query Corpus

Data Matching

High-performance clustered search done on TAUS Data Cloud

Search results are fine-tuned and divided into three selections

Corpus Creation

Corpus Optimization

Further cleaning and manual checks by TAUS data experts

Corpus is delivered and quality reviewed by the initiating user

Data Quality Validation

Corpus added to Library

Corpus is published in the Matching Data Library

Visit the Matching Data website

Matching Data Applications

TAUS Data Cloud

The largest industry-shared language data repository, with more than 35 billion words in 600 language pairs.

Crawled data

Existing resources crawled from the internet, such as ParaCrawl. TAUS can also undertake targeted crawling.

Legacy TMs

Your own legacy TM databases can be uploaded to a dedicated private server where Matching Data is performed.

Matching Data Library

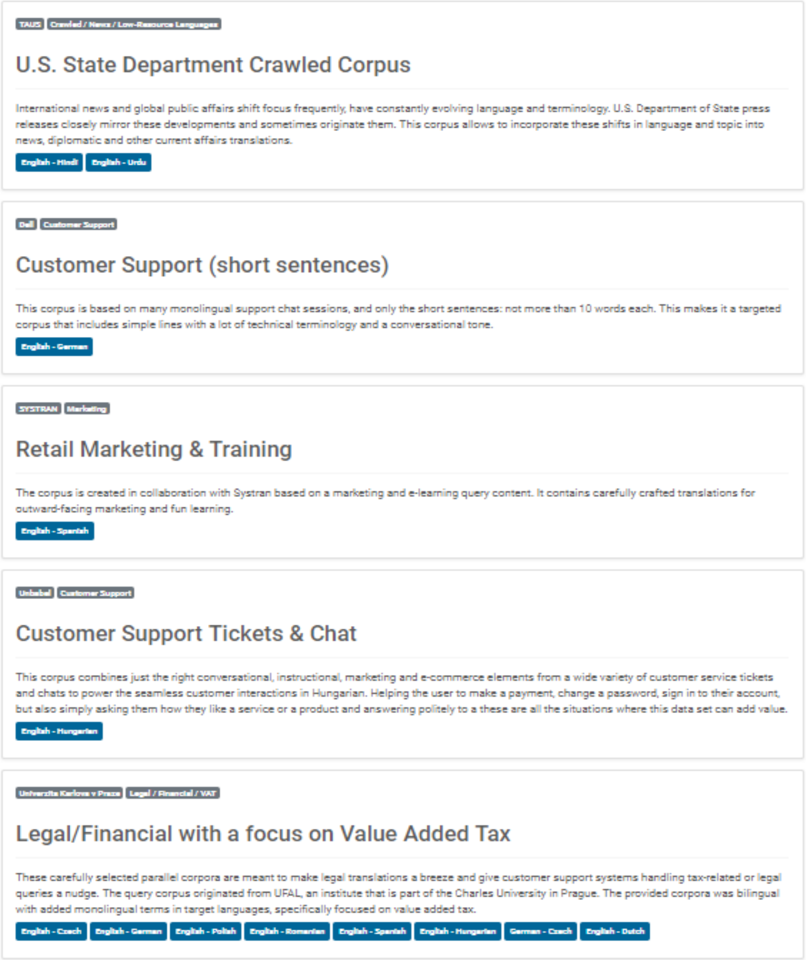
Titles in Library: 45

Domains:

- U.S. State department Crawled
- Customer Support Short Sentences
- Retail Marketing & Training
- Customer Support Tickets & Chat
- Legal/Financial (VAT focus)
- E-commerce
- Medical/Pharmaceutical
- Customer Support
- Colloquial

Languages:

EN-ES, EN-PT(BR), EN-ZH(CN), EN-KO, EN-JA, EN-NL, EN, DE, EN-RU, FR-NL, DE-PL, DE-IT, EN-IT, EN-CZ, EN-Pl, EN-RO, DE-CZ, EN-HU, EN-HI, EN-UR (PK)



TAUS | Crawled / News / Low-Resource Languages

U.S. State Department Crawled Corpus

International news and global public affairs shift focus frequently, have constantly evolving language and terminology. U.S. Department of State press releases closely mirror these developments and sometimes originate them. This corpus allows to incorporate these shifts in language and topic into news, diplomatic and other current affairs translations.

English - Hindi | English - Urdu

Dell | Customer Support

Customer Support (short sentences)

This corpus is based on many monolingual support chat sessions, and only the short sentences: not more than 10 words each. This makes it a targeted corpus that includes simple lines with a lot of technical terminology and a conversational tone.

English - German

SYSTRAN | Marketing

Retail Marketing & Training

The corpus is created in collaboration with Systran based on a marketing and e-learning query content. It contains carefully crafted translations for outward-facing marketing and fun learning.

English - Spanish

Unbabel | Customer Support

Customer Support Tickets & Chat

This corpus combines just the right conversational, instructional, marketing and e-commerce elements from a wide variety of customer service tickets and chats to power the seamless customer interactions in Hungarian. Helping the user to make a payment, change a password, sign in to their account, but also simply asking them how they like a service or a product and answering politely to a these are all the situations where this data set can add value.

English - Hungarian

Univerzita Karlova v Praze | Legal / Financial / VAT

Legal/Financial with a focus on Value Added Tax

These carefully selected parallel corpora are meant to make legal translations a breeze and give customer support systems handling tax-related or legal queries a nudge. The query corpus originated from UFAL, an institute that is part of the Charles University in Prague. The provided corpora was bilingual with added monolingual terms in target languages, specifically focused on value added tax.

English - Czech | English - German | English - Polish | English - Romanian | English - Spanish | English - Hungarian | German - Czech | English - Dutch

Check for the latest corpora in the Matching Data Library

Human Language Project

To be truly global, today's content needs to reach the users in high-growth markets, in their local language. However, there is often not enough language data available in these languages to make that possible.

By engaging communities of native speakers to build data sets in the scarce language pairs and a variety of high-demand domains, TAUS aims to bridge that gap.

Focus:

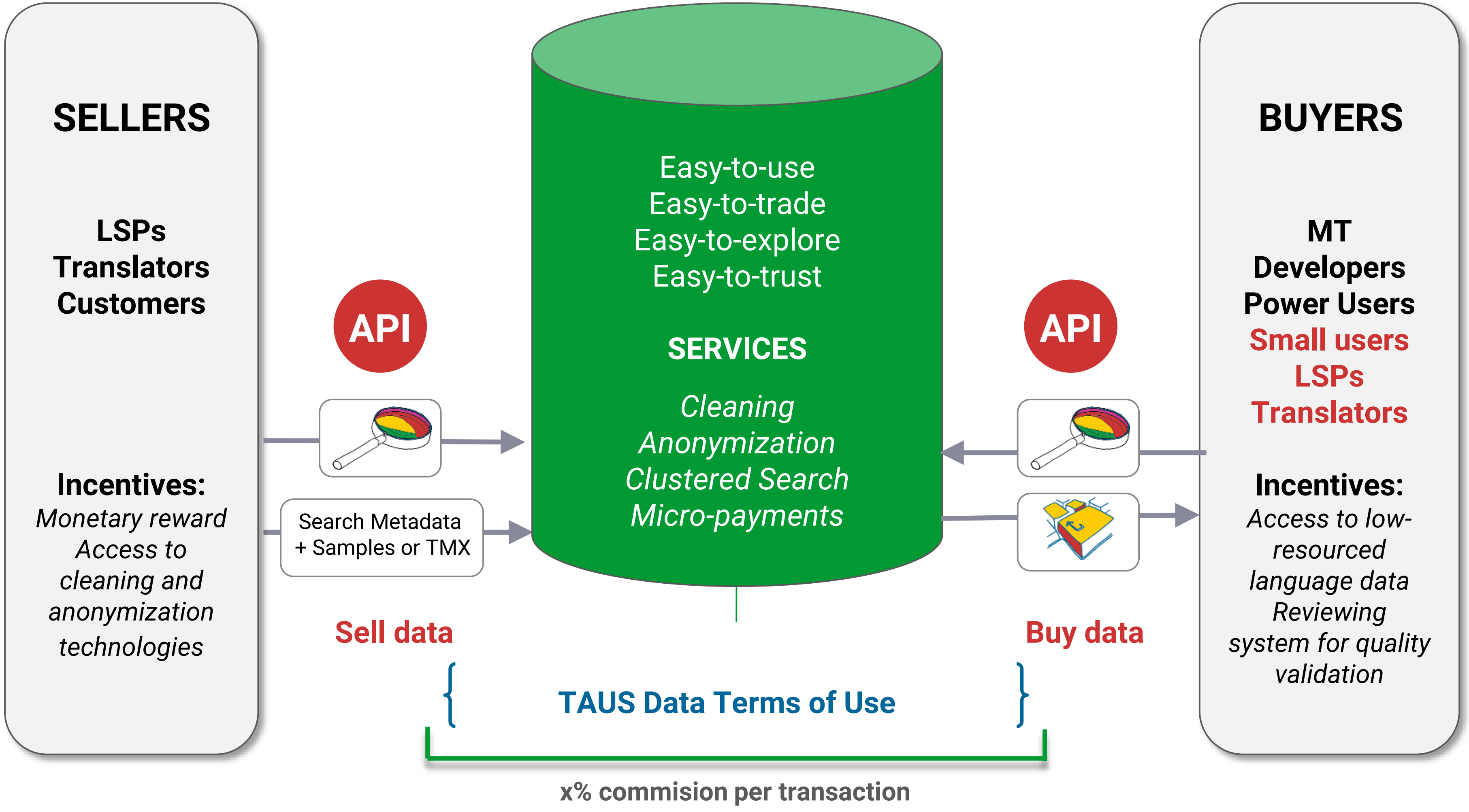
- Indian languages
- African languages

Read more about this project

CEF Data Marketplace – 2020

Funded by European Commission. Consortium: Translated, FBK Trento and TAUS

Data Marketplace will establish a secure and continuous supply of language data for all languages and domains, by encouraging the data owners to contribute their translation data.



SELLERS

LSPs
Translators
Customers

Incentives:
Monetary reward
Access to cleaning and anonymization technologies

API

Search Metadata
+ Samples or TMX

Sell data

Easy-to-use
Easy-to-trade
Easy-to-explore
Easy-to-trust

SERVICES

Cleaning
Anonymization
Clustered Search
Micro-payments

API

Buy data

BUYERS

MT
Developers
Power Users
Small users
LSPs
Translators

Incentives:
Access to low-resourced language data
Reviewing system for quality validation

TAUS Data Terms of Use

x% commision per transaction

Data Legal Framework

Data Provider & Pooling Conditions

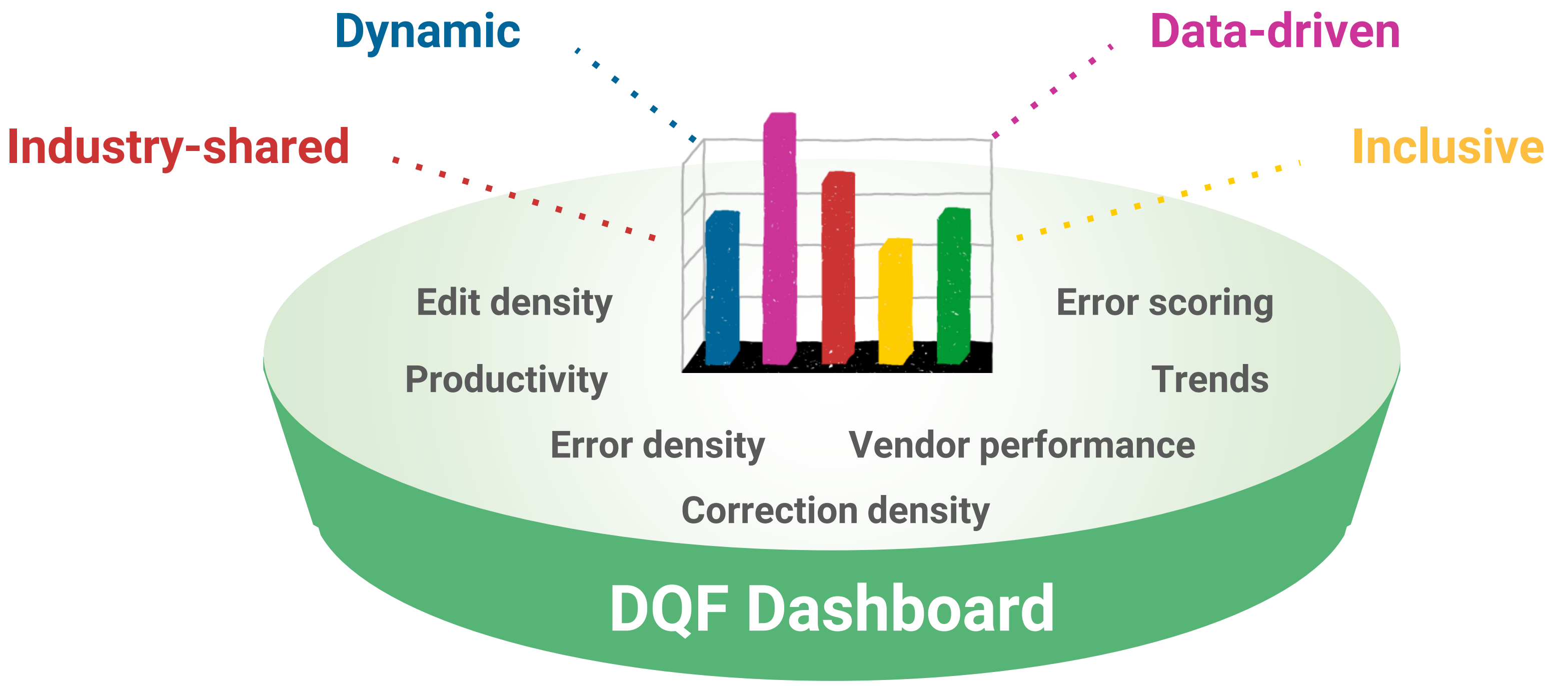
- Free to use translations
- Free to develop derivative work
- Copyright remains with data owner
- Agreed by all users and members
- Other users can develop derivative work and use data to increase efficiency and improve translation quality
- TAUS Data owns the IP to the infrastructure

Currently being reviewed and updated by the TAUS Partner Committee and Baker & McKenzie.

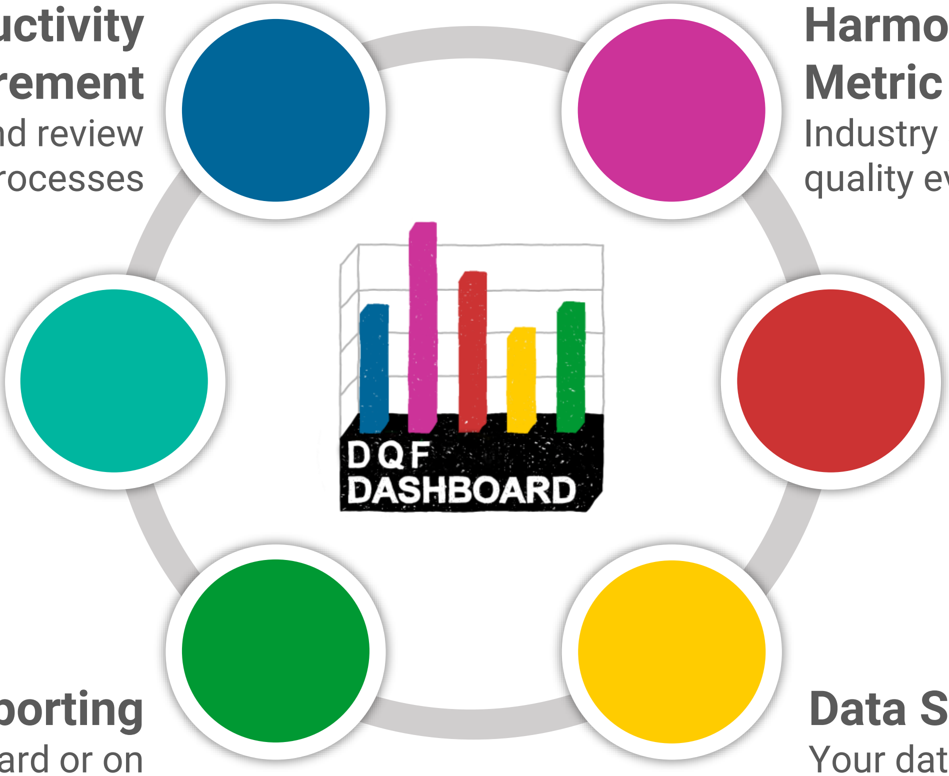
Data for Business Intelligence

DQF Productivity & Quality Data

- Standard measurements
- Industry benchmarks
- Integrated solution
- REST API (send-metrics-only available soon)
- Governed by DQF Terms of Use



DQF Dashboard Features



Quality and Productivity Measurement
For translation and review processes

Harmonized DQF-MQM Metric
Industry shared metric for quality evaluation

Seamless Integration
With CAT tools and TMS through an open API

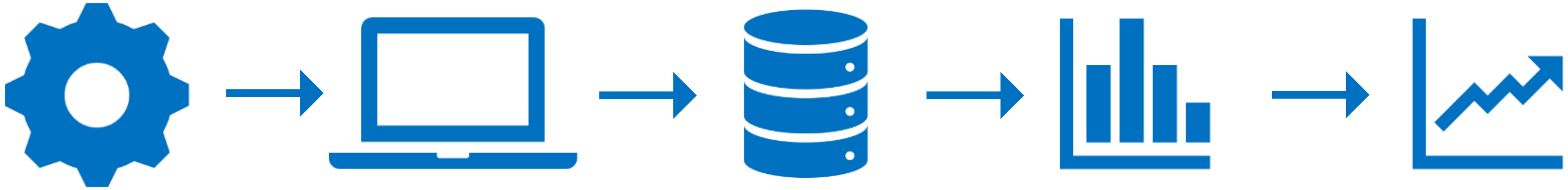
Benchmarking
Between your own projects or against other industry users

Real-time Reporting
On the DQF Dashboard or on your own dashboard with the Data Connector

Data Safety Guarantee
Your data is protected by a solid legal framework and is not shared with anyone

Take a look at the DQF Dashboard

How DQF Works



Connectivity
Open API & plugins

Integration
Most common
CAT tools & TMS

Data collection
Translation & review
processes

Visualization
Interactive
dashboard

Performance tracking
Reports: project,
benchmark and trend

DQF Data

PRODUCTIVITY DATA

- Source and target segment content
- Word count, character count
- Segment origin
- MT engine
- TM match rate
- Time spent

QUALITY DATA

- Annotated/corrected target content
- corrections
- Error category and severity
- Type of review (correction/error annotation/combined)
- Pass/fail threshold
- Penalty points per severity level
- Review sampling percentage

PROJECT-LEVEL INFORMATION

- CAT/TMS name
- File name
- Source and target languages
- Project creation/update date and time
- User email
- Project labels
- Vendor groups
- Sector/Content Type/Process/Quality Level (used as filters on the Dashboard)

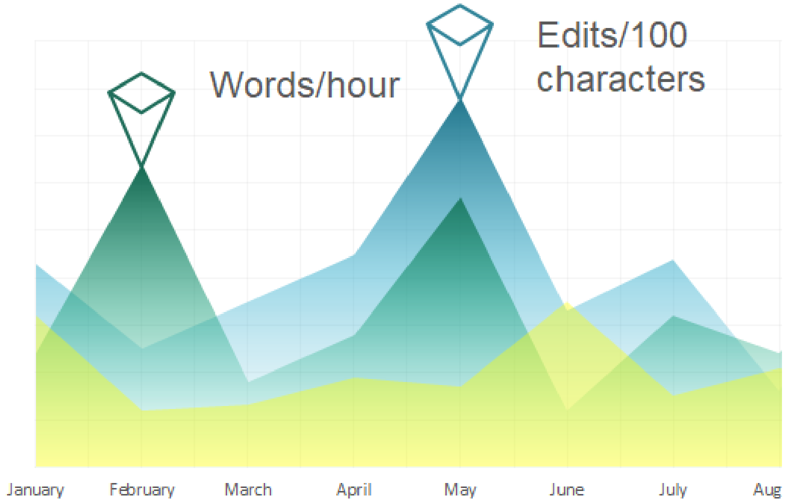
Unique Business Intelligence

Quantify translation effort

Compare efficiency of your resources

Establish language-specific quality baseline

Verify suitability of your payment model



Easy-to-Integrate

Thanks to open APIs, you can plug into all our data sources within seconds.

Matching Data API

Allows systems to identify and pull domain-specific matching data sets directly from the TAUS Matching Data.

Data Cloud API

Allows programs to use the translation memory sharing, data pooling and TAUS segment search features of the Data Cloud.

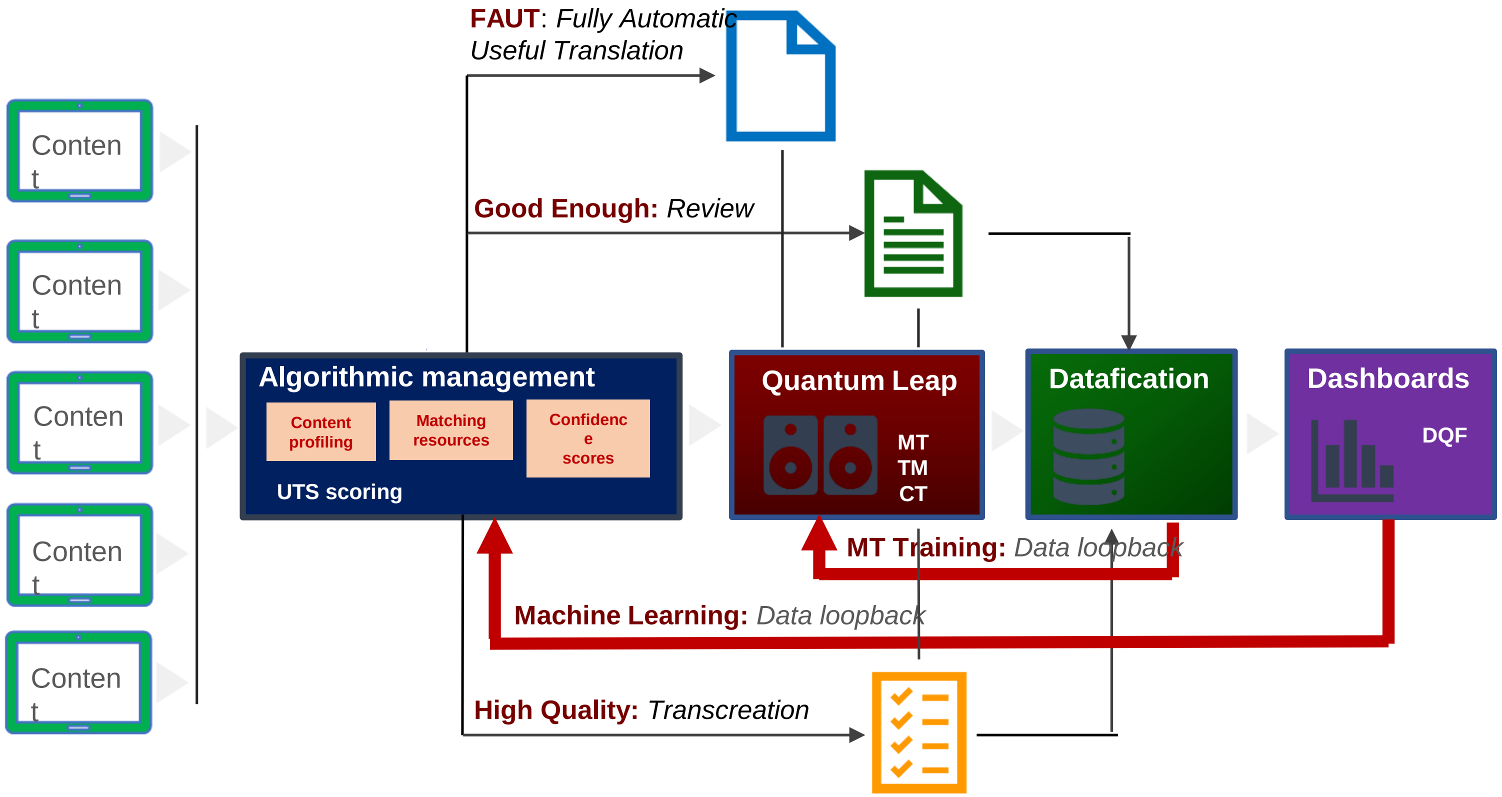
DQF API

Allows users to measure the translation productivity and quality in any CAT/TMS and access the analytics on the DQF Dashboard.

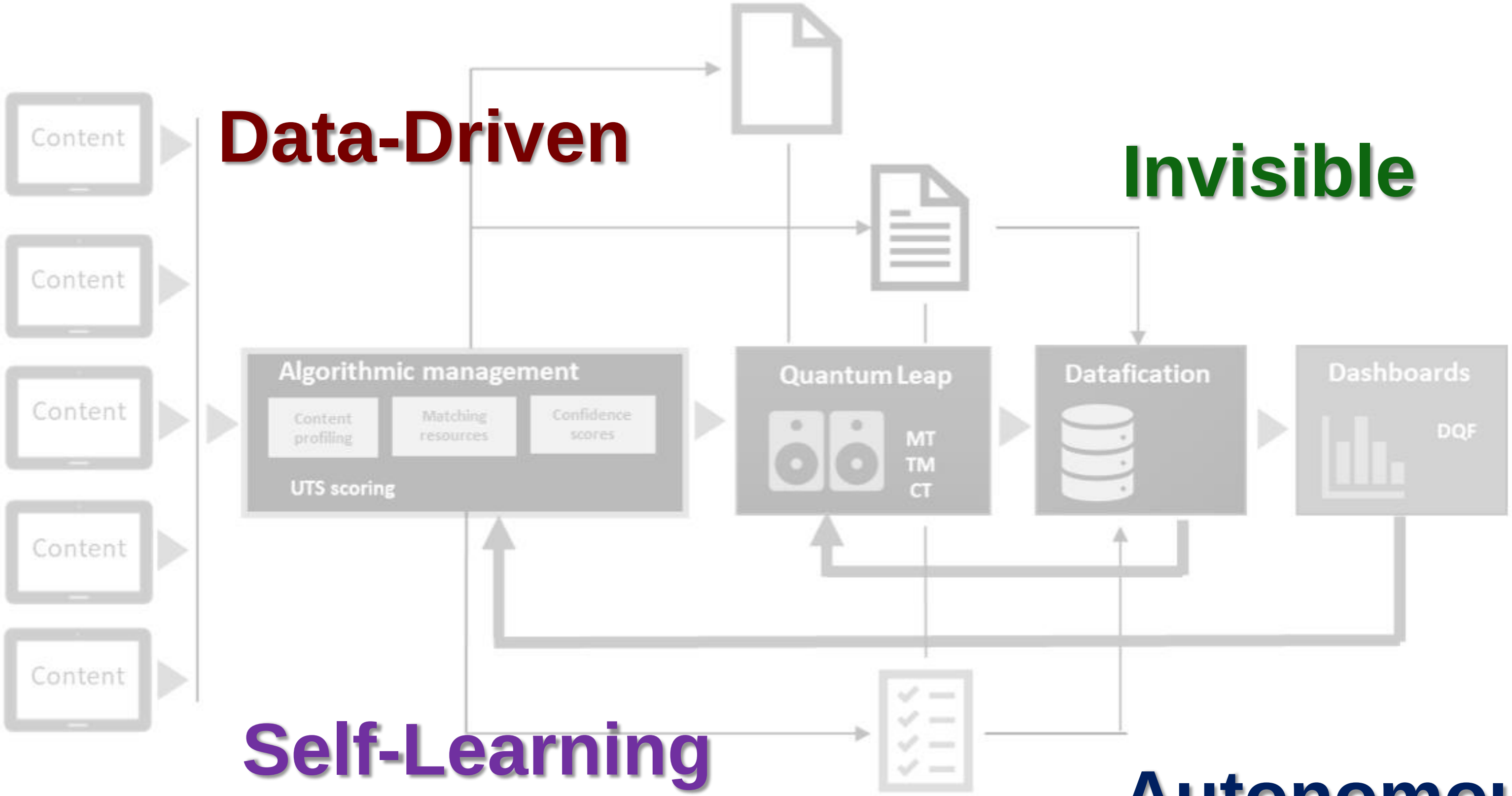
Data Connector API

Allows users and developers to extract user or organization project data as well as the aggregated industry data from the TAUS DQF server and enhance the internal reporting on their own dashboards.

The *Modern* Translation Pipeline



The *Modern* Translation Pipeline …



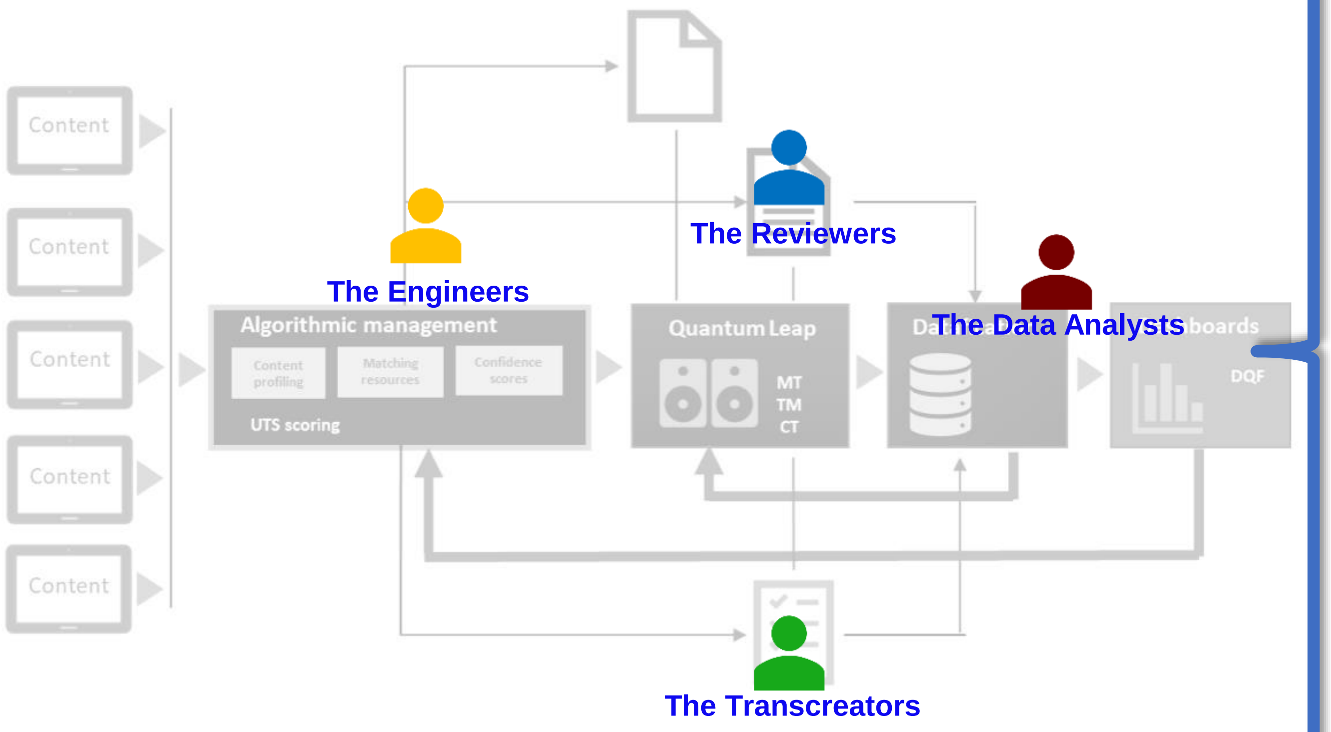
Data-Driven

Invisible

Self-Learning

Autonomous

… and where are the jobs?!



A search for global multilingual talents to undertake a plethora of new value-added services:

- **Brand Validation**
- **Multilingual Digital Marketing**
- **eDiscovery**
- **Usability Testing**
- **Campaign Management**
- **Content Curation**
- **Community Development**

MT Thousands Time Bigger than HT

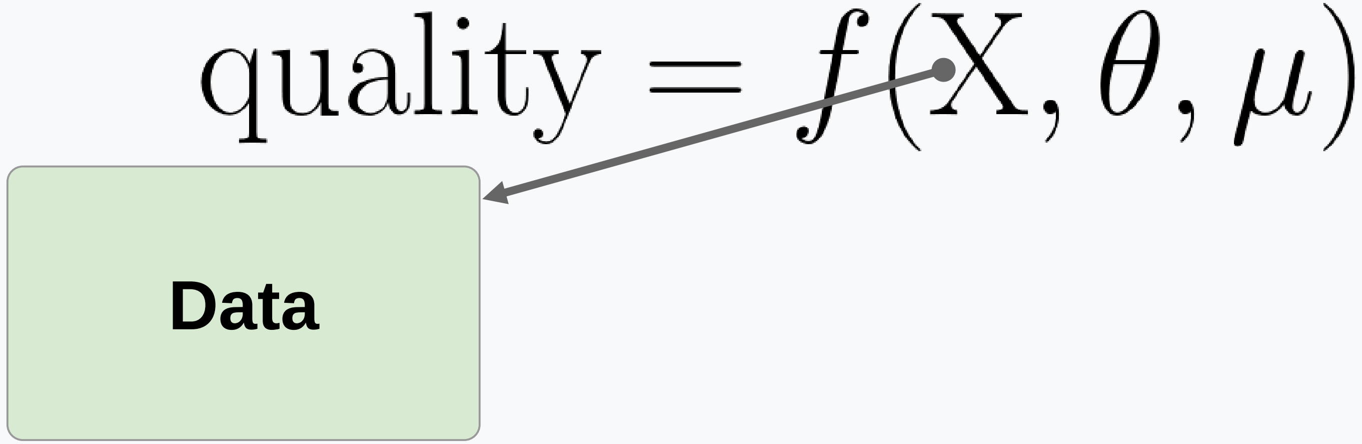
300 trillion words

vs.

200 billion words translated by professionals

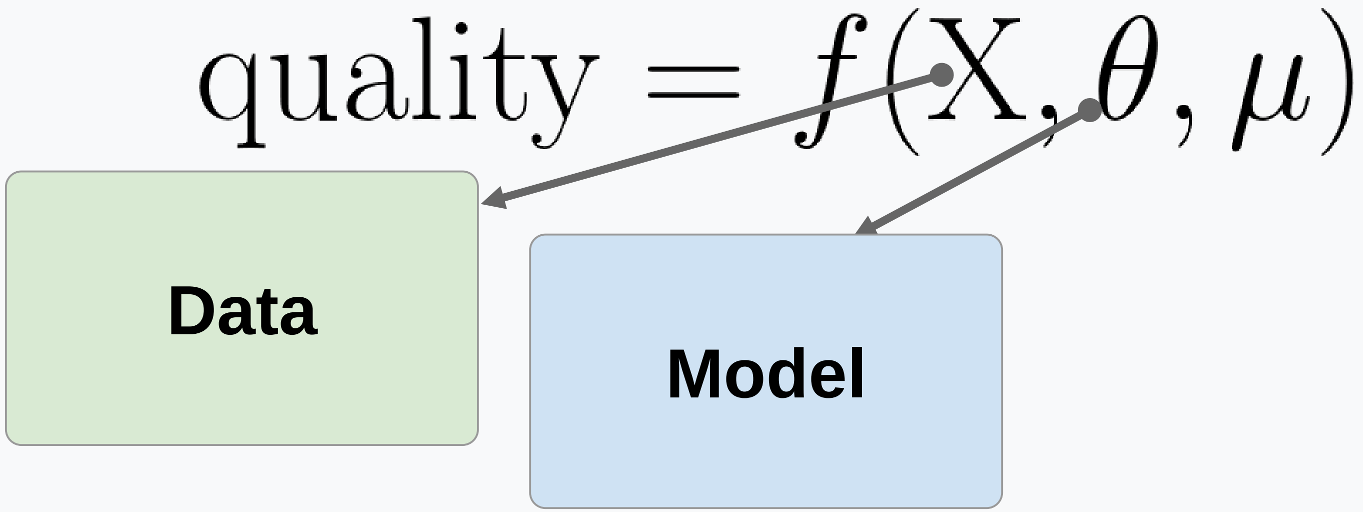
MT is a Simple Sum of Algorithms and Data

$$\mathrm{quality} = f(\mathrm{X}, \theta, \mu)$$



$$\text{quality} = f(\mathrm{X}, \theta, \mu)$$

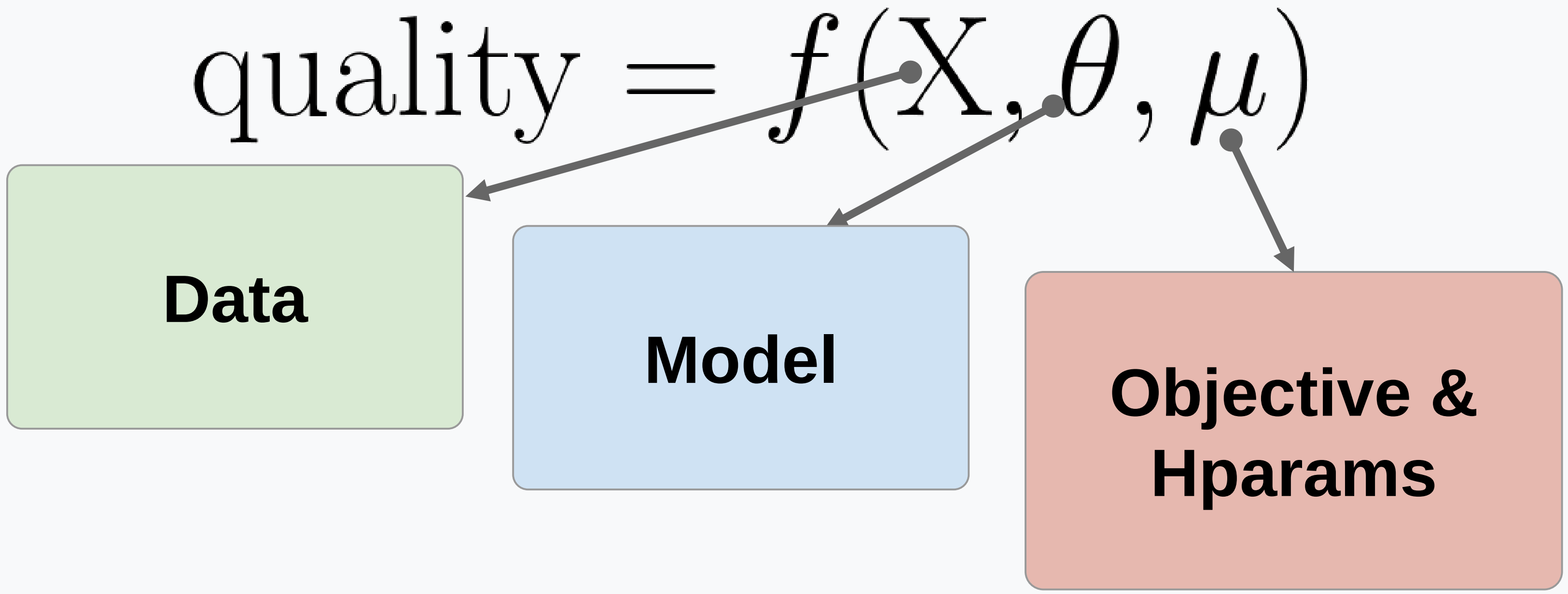
Data



$$\text{quality} = f(X, \theta, \mu)$$

Data

Model

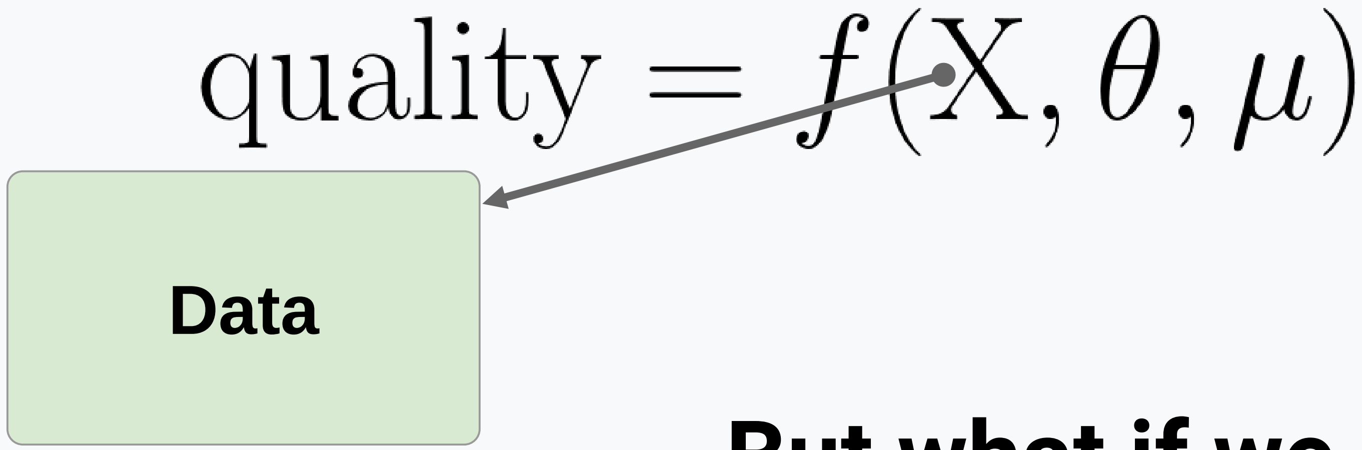


$$\text{quality} = f(X, \theta, \mu)$$

Data

Model

Objective & Hparams



$$\text{quality} = f(\mathrm{X}, \theta, \mu)$$

Data

But what if we don’t have enough current quality data?

TAUS

the language data network