# The Reasonable Effectiveness of Data

Achim Ruopp, Director of Data Cloud

# Agenda

- Introduction
- Data
  - Collect
  - Combine
  - Select
- Grey Box Testing
- Lessons Learned

# ModernMT Project

Horizon 2020 Innovation Action
3M € funding
3 years: 2015-2017

Goal:

deliver a large-scale commercial online **machine translation** service based on a new open-source distributed architecture.


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 645487.

# ModernMT Consortium

## Business



## Research


# Modern MT in a nutshell

**Zero** training **time**

Manages **context**

**Learns** from users

**Scales** with data and users

For the first three points see Marcello Frederico's talk:
Machine Translation Adaptation from Translation Memories in ModernMT

# Introduction
## The Unreasonable Effectiveness of Data

- 2009 article  by Alon Halevy, Peter Norvig, and Fernando Pereira
- Large sets of unlabeled data
  - “invariably, simple models and a lot of data trump more elaborate models based on less data.”
  - Translation “a natural task done every day”
  - “a threshold of sufficient data”
- Representational model: “For many tasks, words and word combinations provide all the representational machinery we need to learn from text.
- Solving the “semantic interpretation problem” based on text

# Introduction
## The Next Challenge: Data Efficiency

- Kamran and Sima'an, 2015
  - "blind concatenation of all available training data may shift translation probabilities away from the domain that the user is interested in"
- Eetemadi et al, 2015
  - "We now find ourselves, however, the victims of our own success, in that it has become increasingly difficult to train on such large sets of data, due to limitations in memory, processing power, and ultimately, speed"
  - "training data has a wide quality spectrum. A variety of methods for data cleaning and data selection have been developed to address these issues."

Amir Kamran and Khalil Sima'an, **Technical report, DatAptor STW project number 12271**, University of Amsterdam, 2015

Sauleh Eetemadi, William Lewis, Kristina Toutanova, and Hayder Radha. 2015. **Survey of data-selection methods in statistical machine translation.** *Machine Translation* 29, 3-4

# Collect
## Translated's MyMemory


- The world's largest Translation Memory
- Seeded originally with translation data from public organizations and the web
- Search
  - Edit search results
  - CAT Tool plug-ins
- Users can upload TMs
  - Private or publicly shared
- TM matching
  - No mass downloading
- September 2016: 1.491.231.338 human contributions
- 100s of millions of words per ModernMT language pair

# Collect
## TAUS Data Cloud


- Largest industry-shared repository of translation data
- A neutral and secure repository platform for
  - Sharing/pooling translation data based on a reciprocity model
  - Searching domain-specific or general data
  - Leveraging Translation Data
- Solid legal framework established by 45 founding members
- Addresses the shortage of available in-domain parallel data from the industry
- September 2016: 72,476,886,904 words in the repository
- 10s to 100s of millions of words per ModernMT language pair

# Collect

## The Web – Crawling it is hard

- The Web is large - even the so-called Surface or Indexable Web
- The Web is messy
- The Web is constantly in flux
- Not many organizations crawl the entire indexable web
  - Google - about 49 billion web pages in index (Source: http://www.worldwidewebsize.com/)
  - Microsoft - about 20 billion web pages in index (Source: http://www.worldwidewebsize.com/)
- Other crawls are focused crawls on a subset with certain criteria/goals
  - Still hard for the same reasons

# Collect

## CommonCrawl to the Rescue

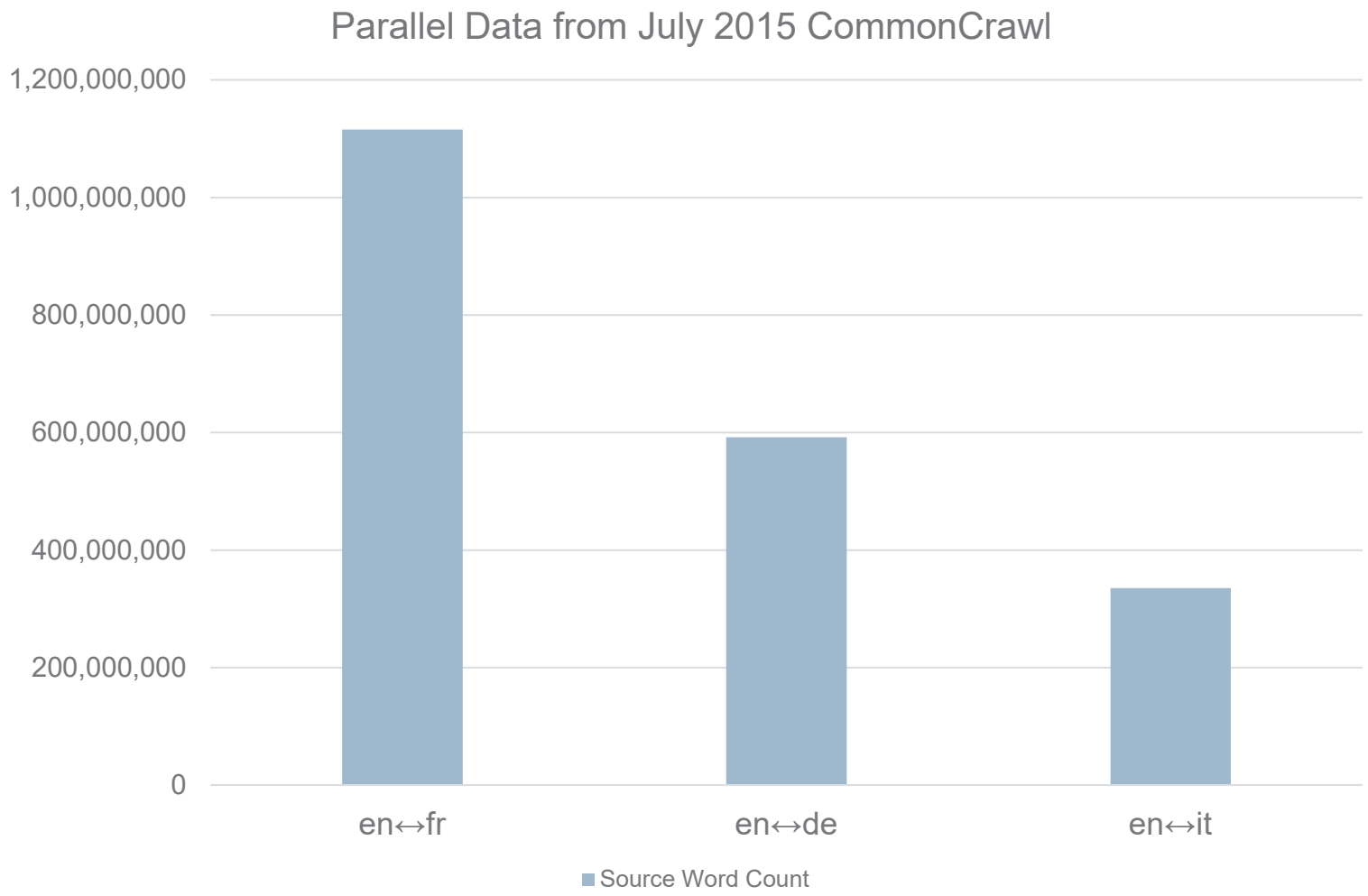
- commoncrawl.org
  - “CommonCrawl is a 501(c)(3) non-profit organization dedicated to providing a copy of the internet to internet researchers, companies and individuals at no cost for the purpose of research and analysis.”
- On average 1.5 billion unique URLs per crawl
- A very good resource for sourcing bilingual and monolingual data for machine translation purposes
  - Prototype developed by academic developers in 2012/2013 showed potential to mine parallel corpora with millions of source words

# Collect

## CommonCrawl to the Rescue

- Implemented data collection pipeline based on prototype techniques
- Collecting monolingual and bilingual data
- Open sourced at https://github.com/ModernMT/DataCollection
- We are making the indices of parallel pages we discover available
  - Saves running half of the data collection pipeline
  - Each user still has to download their own data
    - Avoids potential copyright issues

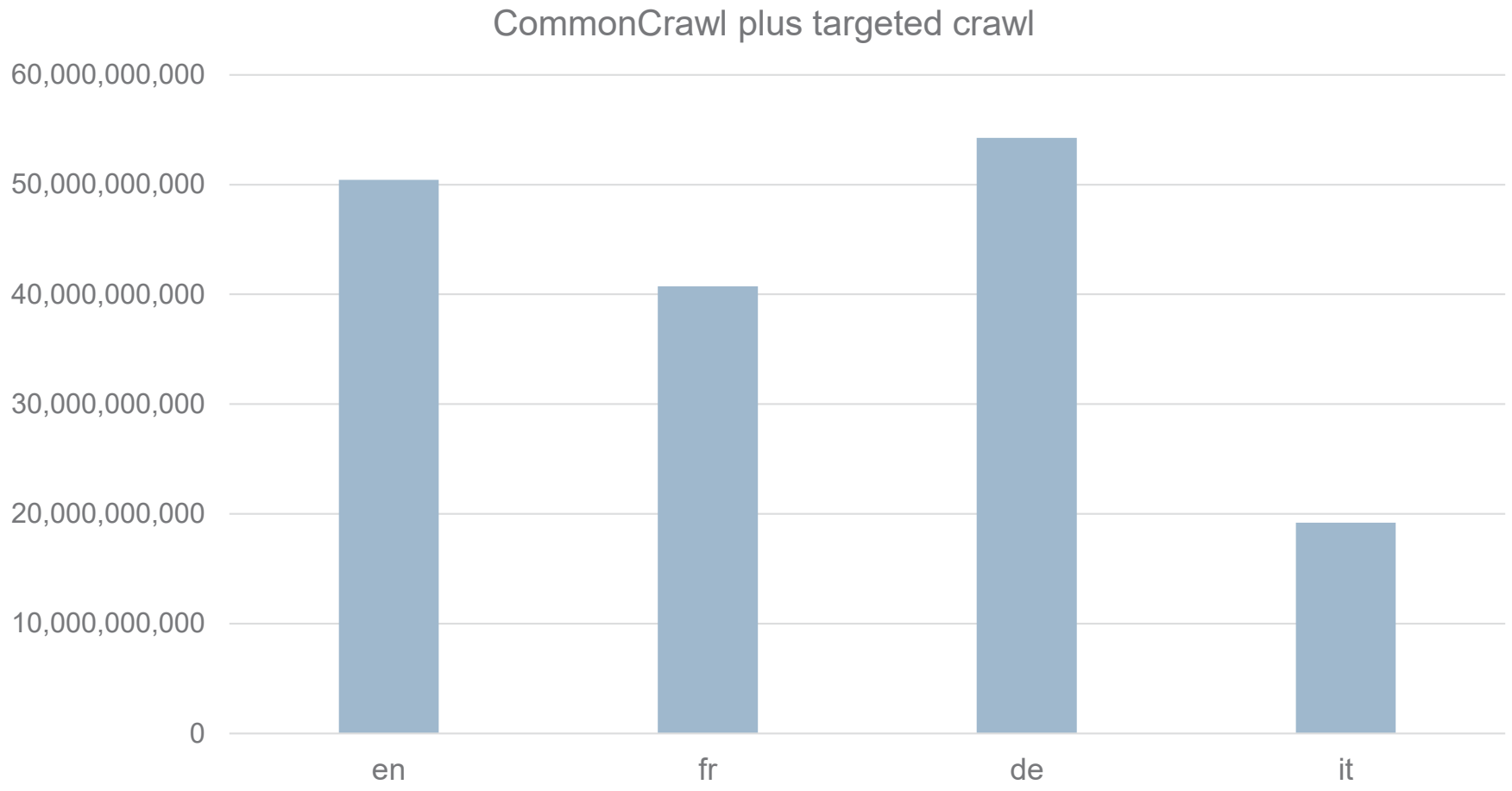
# Collect
## CommonCrawl Parallel Data




- Original source language not detectable – data can be used in both directions
- Not deduplicated – separated by registered web domains

# Collect

## CommonCrawl Monolingual Data




- Deduplicated
- More raw English, French and Italian data available

# Combine

- As plain text corpora
  - TMX files are converted to plain text
- “Document” concept
  - One TMX, one document
  - One site (web domain), one document
- Uniform pre-processing and post-processing for all data
- Repository meta-data data not unified/combined
  - Ontologies for domain/content type too different
  - Not available for web crawled content
  - Is it even useful?

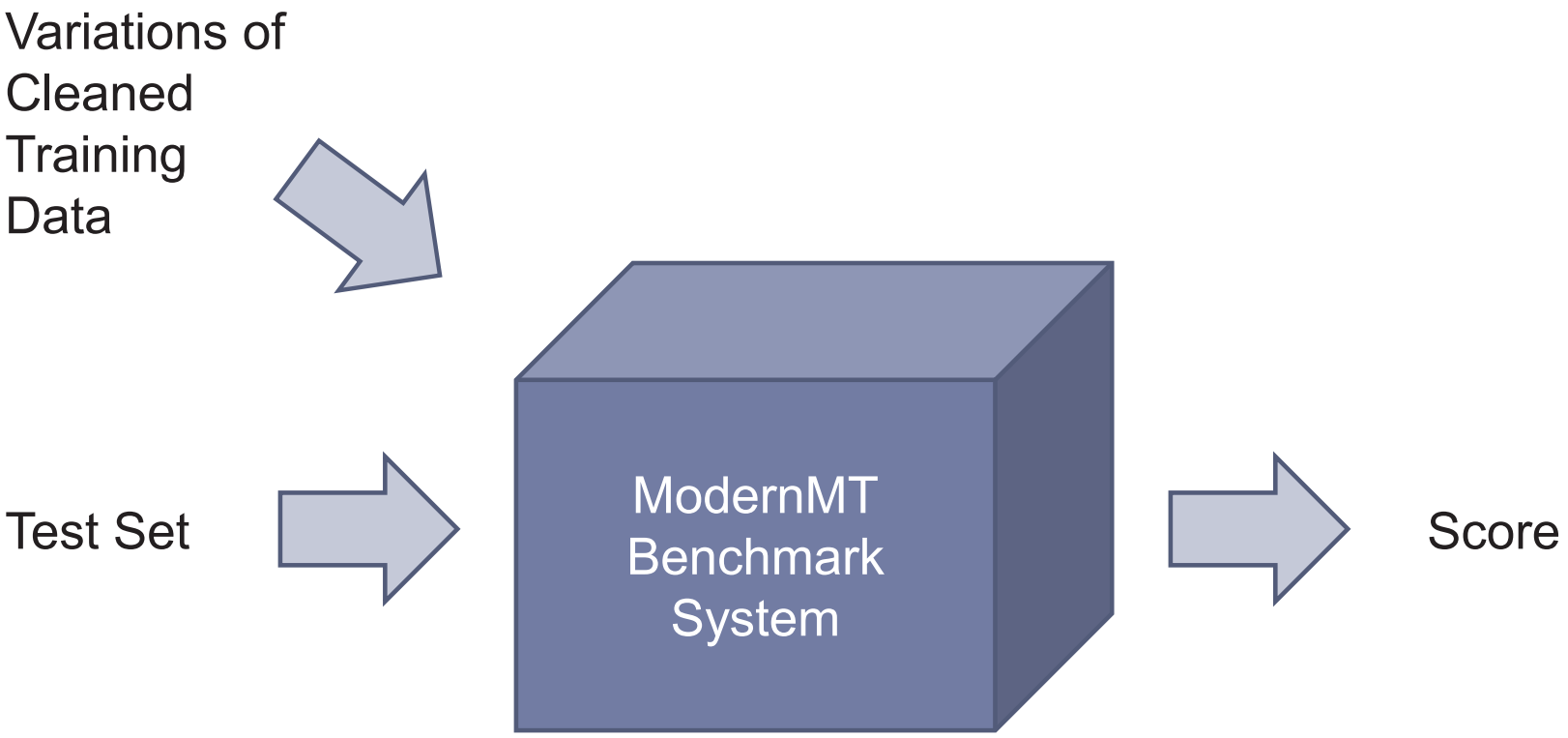
# Select
## Segment Level Data Cleaning

- Clean data from all sources to a uniform level
- Some quality indicators
  - Empty or identical source/target
  - Mismatches in number of sentences
  - Very lengthy segments
  - No alphabetical characters in segments
  - Inconsistency in tags or dates, etc.
  - Big difference in segment length between source and target
  - Language identification

# Select
## Document Level

- Context Analyzer
  - Training data “Document” selection via cosine similarity between user-provided translation input and training data
  - Proven to be competitive with more complex training data selection methods for domain adaptation

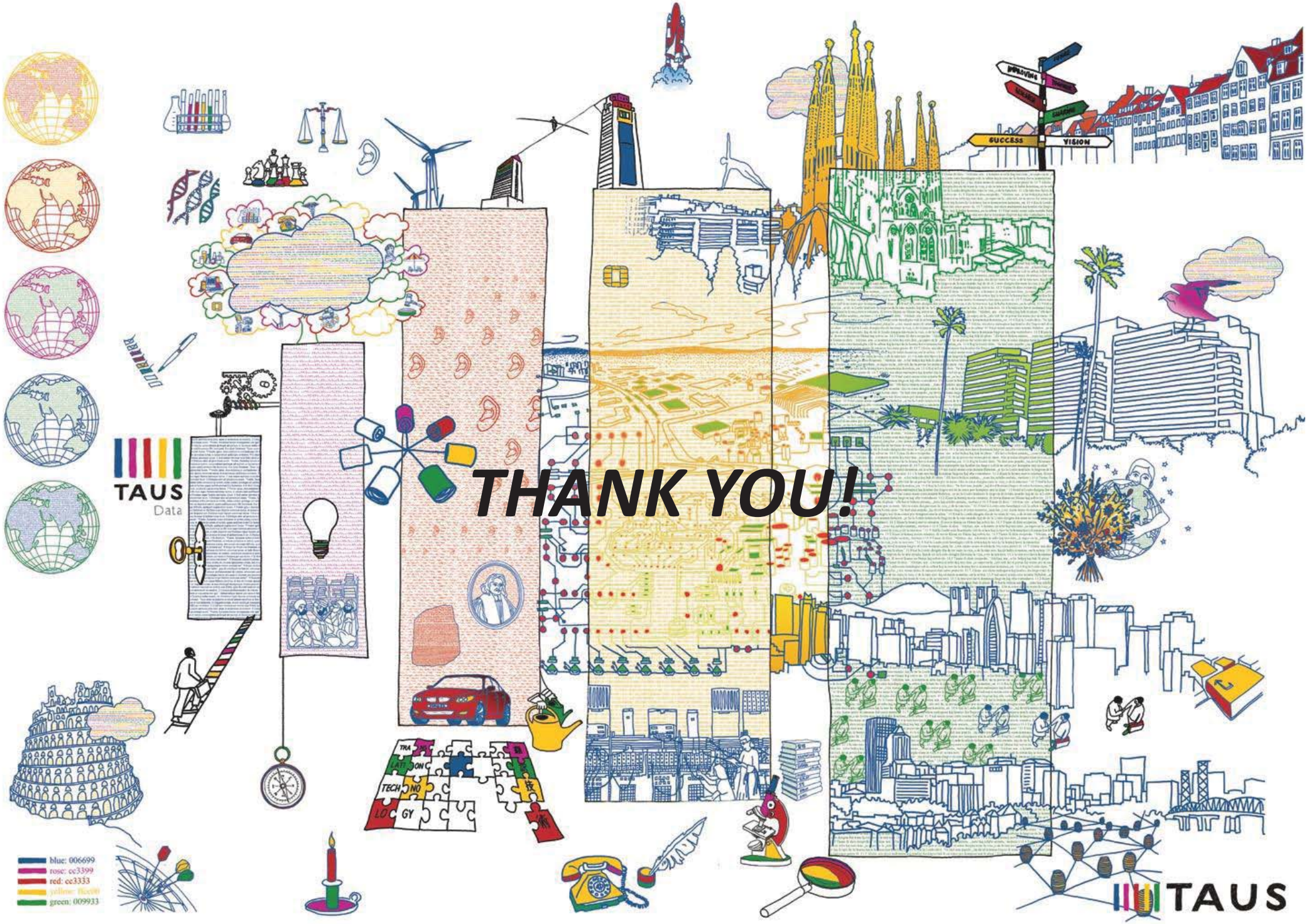
# Grey Box Testing



# Lessons learned so far

- CommonCrawl is a great source for MT training data – please donate!
- Some common sense cleaning of web data is necessary
  - More complex cleaning has diminishing returns
- Web data helps with translating text with wide ranging topics/tone
  - You might not need it if you have focused domains/topics with data available
- Web data does not help to translate items out-of-vocabulary named entities/abbreviations specific to the text you are translating

⇒ Test set design/compilation is crucial


THANK YOU!
TAUS
Data
SUCCESS
VISION
TAUS