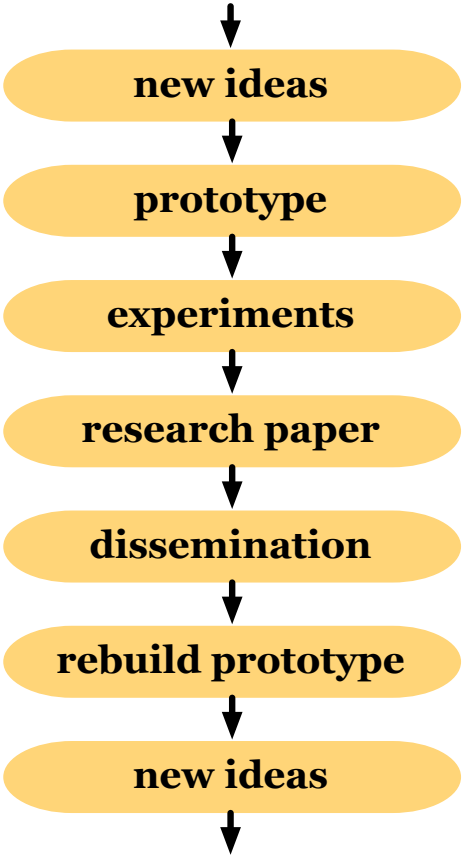
Open Source Tools for Statistical Machine Translation

Philipp Koehn, University of Edinburgh

28 February 2008

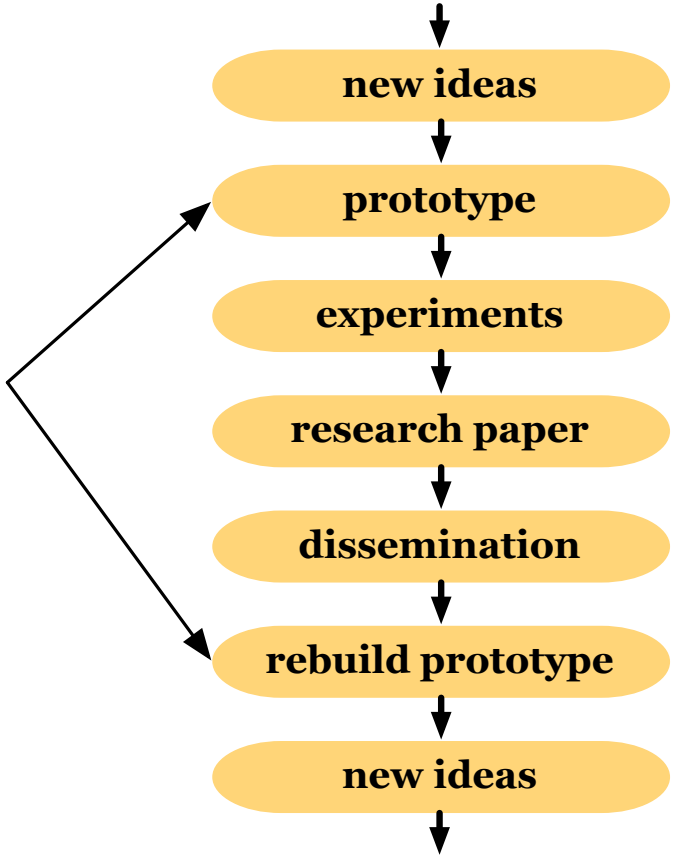
Research Process



Research Process

SMT is increasingly
a big systems field

building prototypes
requires huge efforts



Requirements for Building MT Systems

- **Data resources**
  - **parallel** corpora (translated texts)
  - **monolingual** corpora, especially for output language
- **Support tools**
  - basic **corpus preparation**: tokenization, sentence alignment
  - **linguistic** tools: tagger, parsers, morphology, semantic processing
- **MT tools**
  - word alignment, **training**
  - **decoding** (translation engine)
  - tuning (optimization)
  - re-ranking, incl. posterior methods

Who will do MT Research?

- If MT research requires the development of **many resources**
  - who will be able to do relevant research?
  - who will be able to deploy the technology?
- A **few** big labs?

Google™ IBM Microsoft®

- ... or a **broad network** of academic and commercial institutions?

MT is diverse

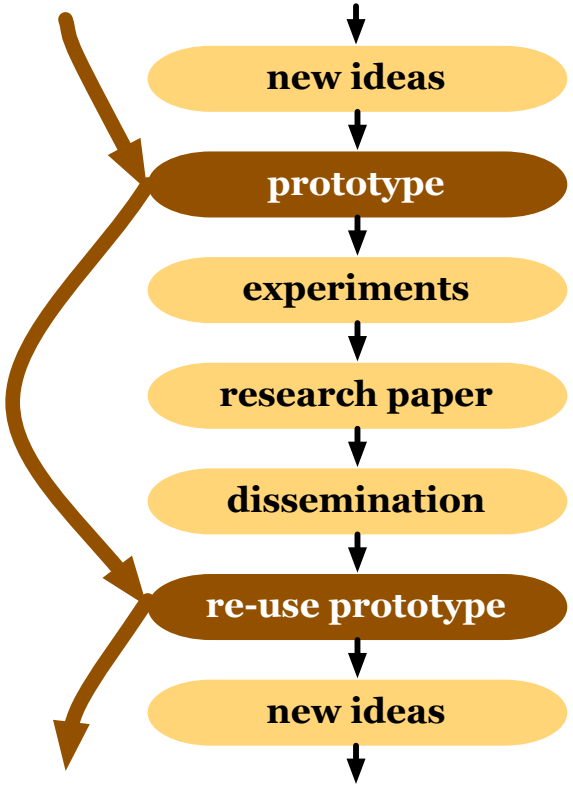
- Many different **stakeholders**
  - academic researchers
  - commercial developers
  - multi-lingual or trans-lingual content providers
  - end users of online translation services
  - human translation service providers
- Many different **language pairs**
  - few languages with rich resources: English, Spanish, German, Chinese, ...
  - many second tier languages: Czech, Danish, Greek, ...
  - many under-resourced languages: Gaelic, Basque, ...

Open Research

SMT is increasingly
a big systems field

building prototypes
requires huge efforts

**sharing of resources
reduces duplication
of efforts**



Making Open Research Work

- Non-restrictive **licensing**
- Active **development**
  - working high-quality prototype
  - ongoing development
  - open to contributions
- **Support** and dissemination
  - support by email, web sites, documentation
  - offering tutorials and courses

EuroMatrix: **Open Research**

- Open source **statistical MT system**
- Open source rule-based system
- **Parallel corpora**
- **Dissemination** activities
  - MT Marathon
  - Evaluation campaign and workshops
  - Online platform

Moses: Open Source SMT

- **Open source** statistical machine translation system
  - state-of-the-art **phrase-based** approach
  - full SMT system: training, tuning, decoding
  - incorporates research on **factored translation models**

- Additional features
  - **confusion network decoding**
  - support for **very large models** through **memory-efficient** data structures
  - multiple language models, translation tables for domain adaptation
  - minimum Bayes risk decoding

Collaboration Beyond EuroMatrix

- **Active development** centered at U Edinburgh, but also
  - Charles University
  - ITC-irst, Italy
  - University of Maryland, USA
- Development also **supported by**
  - EC-funded **TC-STAR** project
  - Johns Hopkins Summer Workshop 2006
  - **US** funding agencies DARPA, NSF

Web Site

- URL: `http://www.statmt.org/moses/`
- **Download**
  - compiled **binaries** for Unix and Windows
  - current **source code** from SVN repository
- **Documentation**
  - **introduction** to statistical MT methods
  - step-by-step **tutorial** on training, decoding, factored models
  - step-by-step instructions on how to build a baseline system
  - descriptions of all features
  - automatically generated **code documentation**
  - **mailing lists** for users and developers

Widely Used

- **Web site** gets **3000 visits** per month
- **Mailing list** distributes **100 emails** per month
- **Academic** uses
  - de-facto **benchmark** for new MT methods
  - **starting point** for most new research groups
  - half of IWSLT submissions used Moses
- **Commercial** uses
  - explored by many machine translation **developers** (incl. Systran)
  - **systems built** for second tier languages (e.g. Swedish, Danish)

Online Demos

- English to Czech
  - provided by Charles University
  - hosted at `https://blackbird.ms.mff.cuni.cz/cgi-bin/bojar/mt_cgi.pl`
- German, Spanish, French to English and back
  - provided by Edinburgh University
  - hosted at `http://demo.statmt.org/webtrans/`
- Outside parties have also created demos
  - Finnish to English, Swedish and back
  - English to Russian

Online Demos

MT Marathon

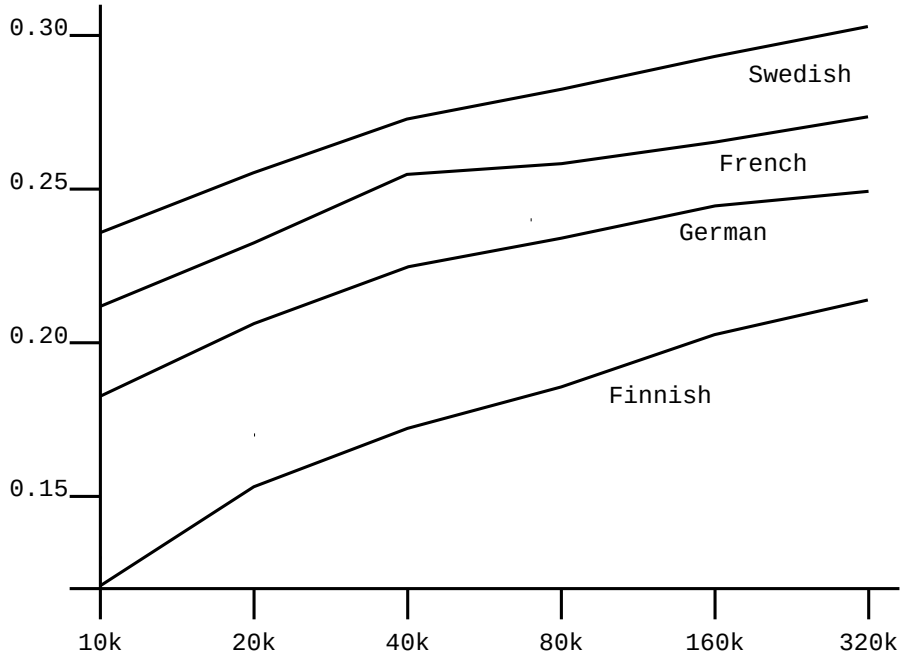
- First MT Marathon: April 2007, Edinburgh
  - one-week intense class with hands-on experience
  - research showcase with talks from leading researchers
- Second MT Marathon: **12-20 May 2008, Berlin**
  - one-week **intense class** with hands-on experience
  - **research showcase** with talks from leading researchers
  - **open source** convention
  - evaluation workshop
  - Translingual Europe **conference**

The Matrix

- `http://matrix.statmt.org/`
- Listing of **available resources**
  - parallel and monolingual **corpora**
  - **tools** and **systems**
  - can be augmented and **edited by users**
- Online **evaluation campaign**
  - developers can **upload their translations** of standard test sets
  - **reference performance** for all language pairs of official EU languages
- Note: currently functional, but still working on some features

Parallel Data: the Bottleneck?

- **More data, better performance** with statistical systems



[from Koehn, 2003: Europarl]

- Where do we get more translated texts from?

Parallel Corpora

- **Europarl**: proceedings of the European Parliament
  - Release of v3 in September 2007
  - 30-40 million words per language, all 11 official languages of EU-15
- **News Commentary**: from http://www.project-syndicate.com/
  - used in ACL WMT 2007 Shared Task
  - 1-2 million words in English, French, Spanish, German, Czech, Arabic, ...
- **Other** corpus projects
  - Acquis Communitaire: includes all 23 languages of EU-25 (JRC)
  - CzEng corpus build by Charles University
  - Hungarian-English corpus extended by Morphologic
  - more data from European Union / European Commission

→ **good translation quality possible** with this data

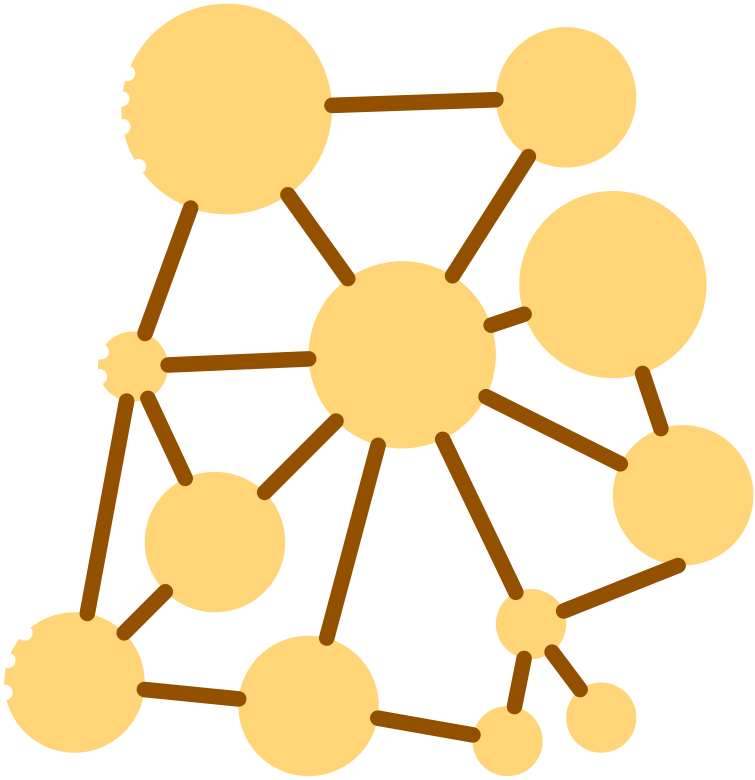
Data from Commercial Sources?

- All large corpora are from **governments**, international institutions
- **Commercial** sources are hard to come by
  - **ownership** between original author, translator
  - intellectual property rights and privacy concerns
  - data is seen as **competitive advantage**
- What could be done:
  - **randomizing** the order of sentences
  - **anonymizing** named entities
- **User generated** data?

Open $ource

Academic Research

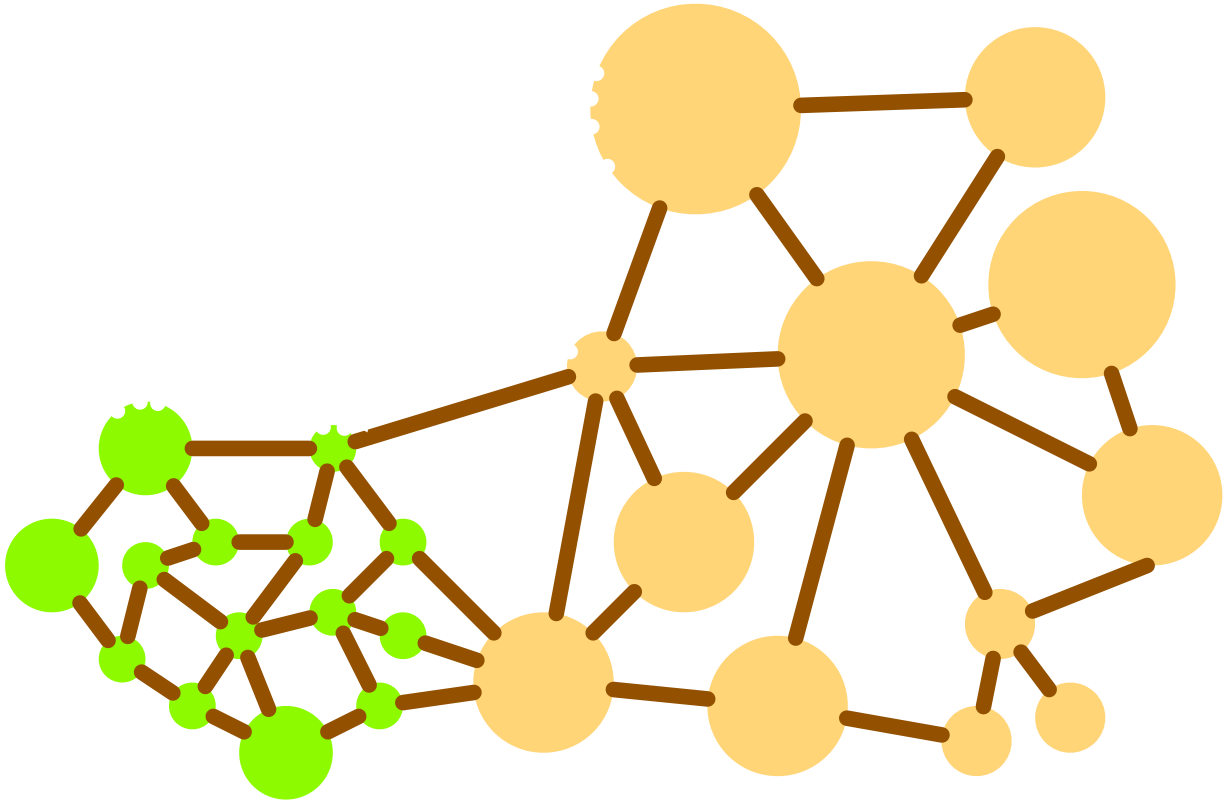
Open $ource



Academic Research

Open Source

Open $ource

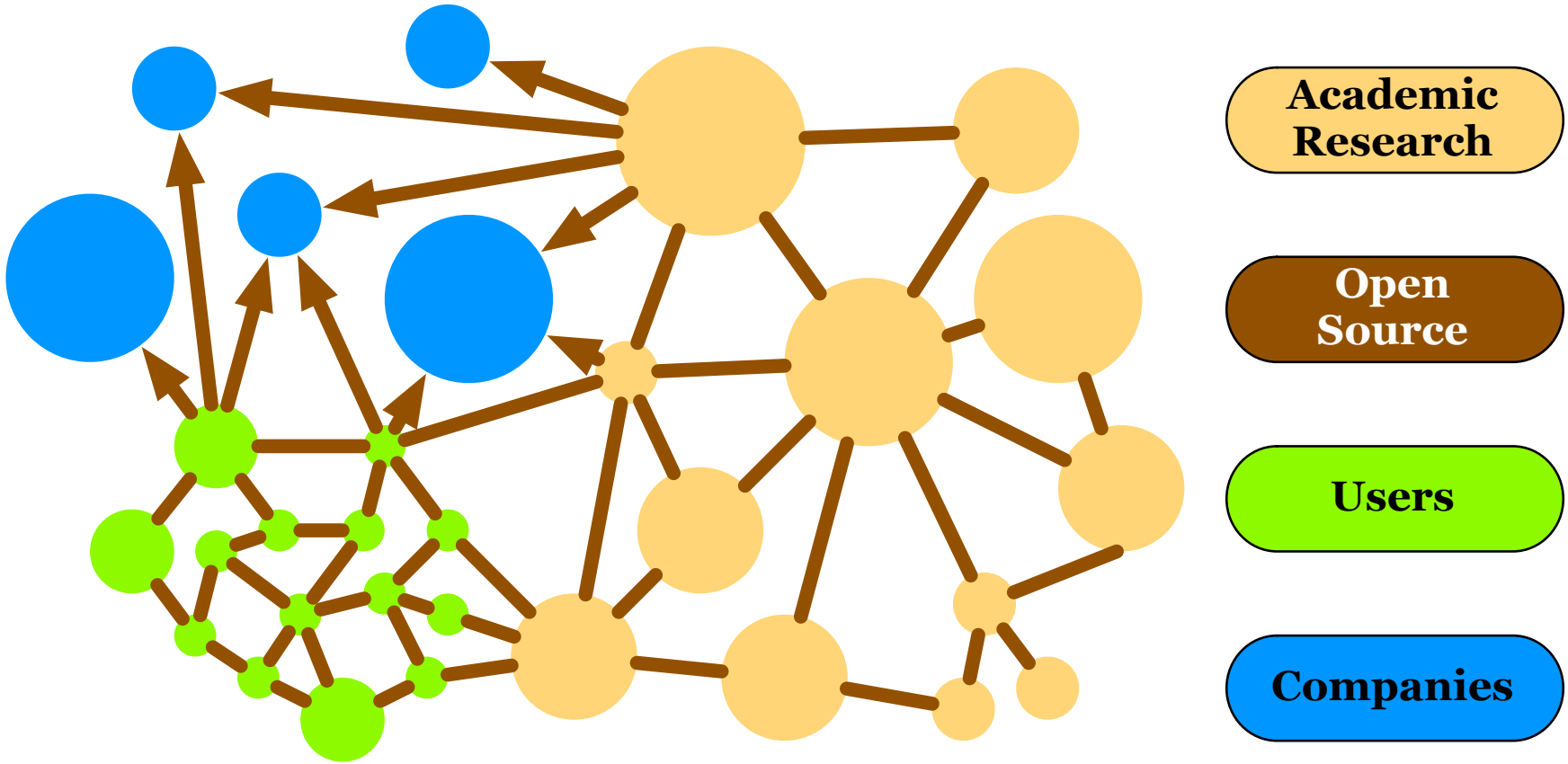


Academic Research

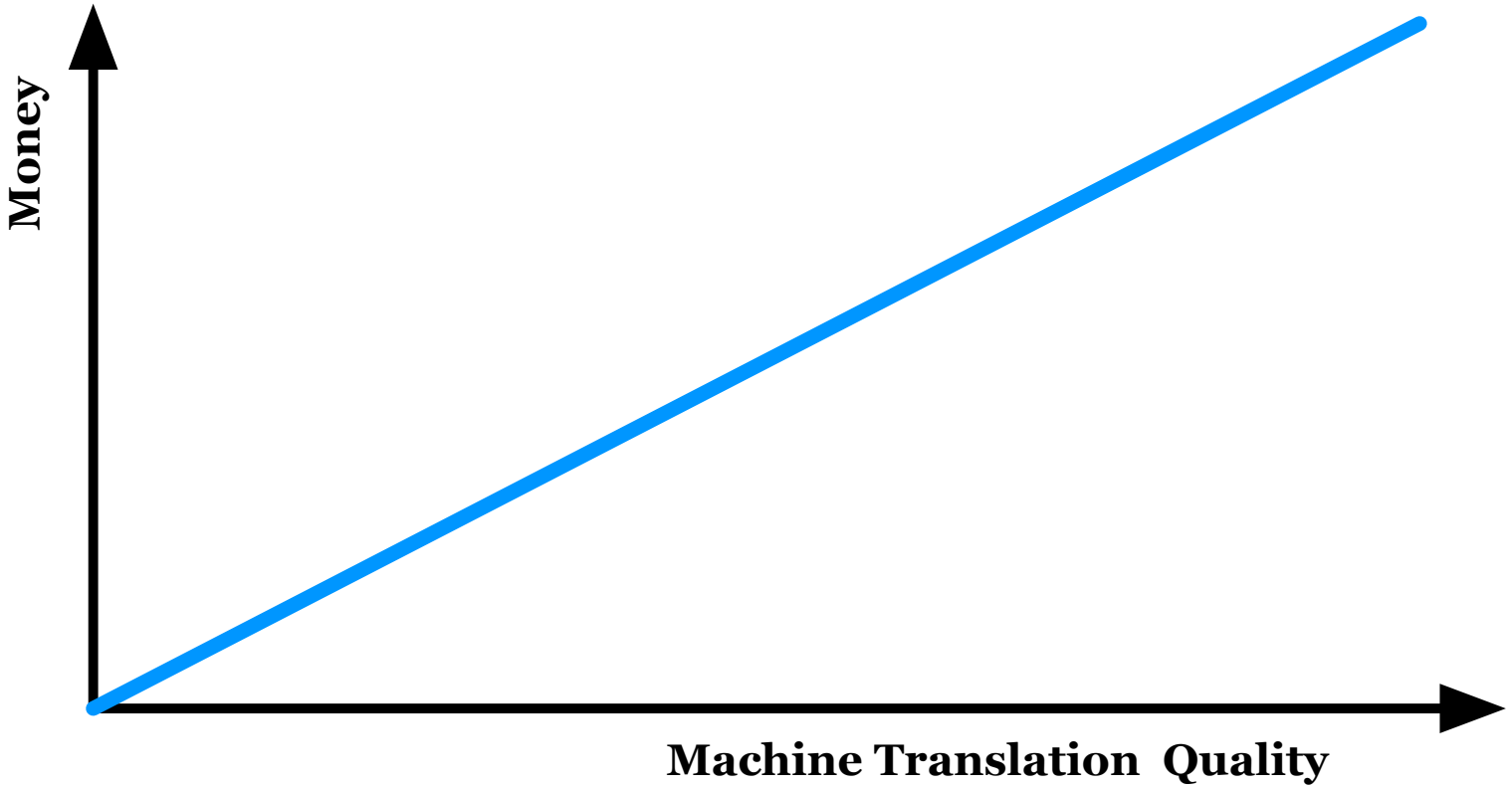
Open Source

Users

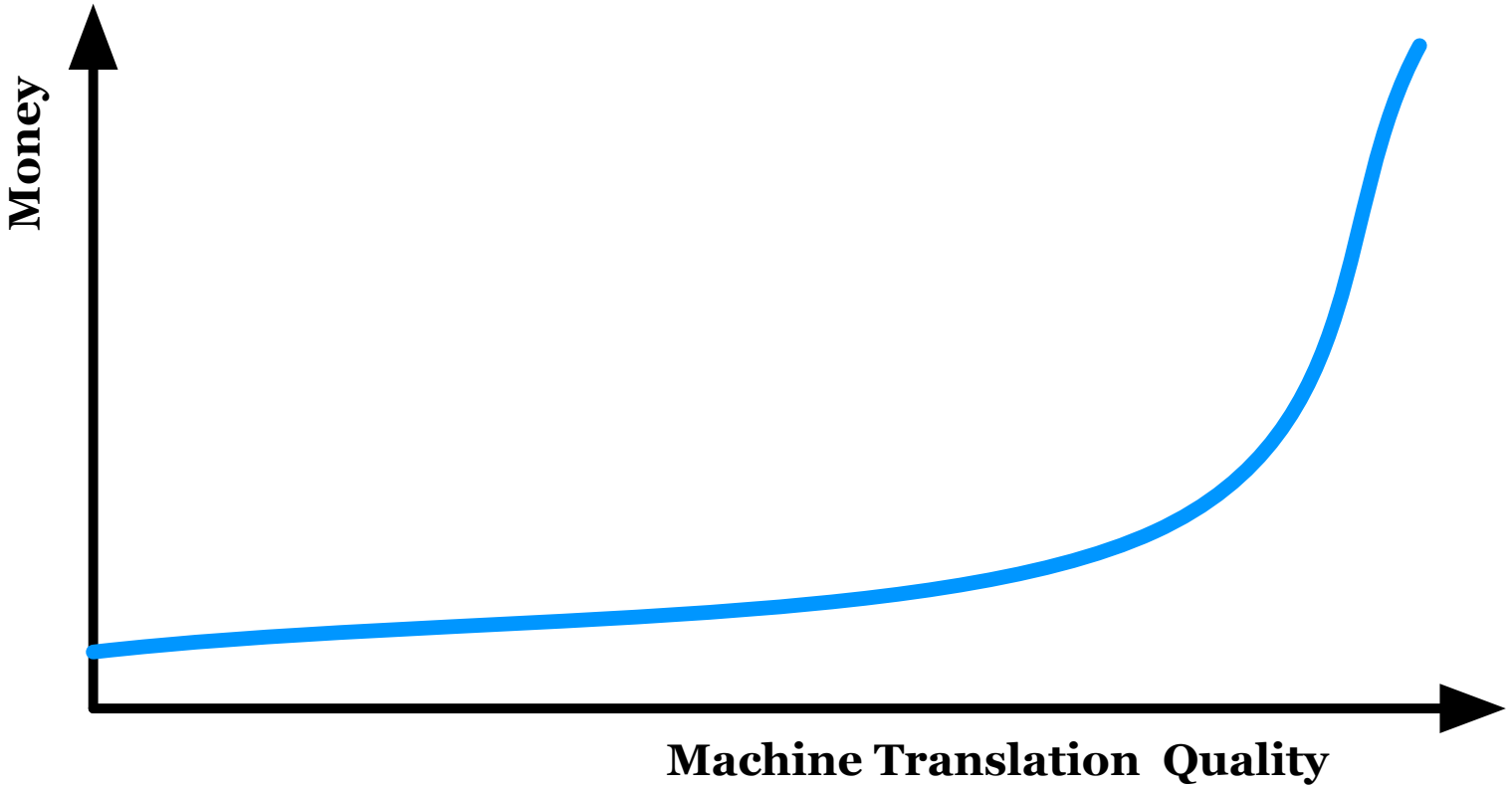
Open $ource



The Tipping Point for MT



The Tipping Point for MT



Thank you

Questions?