European Language Resource Coordination
Connecting Europe Facility

THE POTENTIAL OF LANGUAGE TECHNOLOGY AND AI

European Commission

where we are, where we are heading / should TRY TO head

Khalid CHOUKRI
ELRA/ELDA
choukri@elda.org

For the European Language Resource Consortium

http://www.elra.info/ www.lr-coordination.eu

European Language Resource Coordination
Connecting Europe Facility

OUTLINE OF THE PRESENTATION

European Commission

I. Human Languages and Technologies … Big successes and achievements

- Language Technologies
- Data driven approach
- **Artificial Intelligence approaches**
- Some illustrations and examples for MT over the ages

II. Trends and Challenges

- Market analysis and European position
- Trends and Roadmaps

III. When Will AI Exceed Human Performance ?

Dimensions of the Human Languages

- Speech
- Text inc. documents management (structure)
- Signs
- Handwriting and OCR
- Gestures … pointing
- Images
- Biometrics
- …. Multimodal & Multimedia
- ……

Multilinguality

EXAMPLES OF LTs 1/2

- **Speech Technologies**
  - **Speech Recognition (Speech-to-text)**
  - **Speech Synthesis (text-to-speech)**
  - **Speech to Speech/Text Translation**
  - **Speaker Identification /Verification**
- **Translation Technologies**
  - **Machine Translation**
  - **Computer Aided Translation (CAT) tools**
  - **Translation Memories**
  - **Alignment Tools**
  - **Translation Workflow management**
  - **Authoring Tools**
- Terminology Technologies
  - Terminology Management Systems
  - Terminology Extraction

Examples of LTs 2/2

- Localisation technologies
  - Localisation tools applied to Websites
  - Localisation tools applied to Software
  - Localisation tools applied to Forms
  - Localisation tools applied to Subtitling/Dubbing production
- Natural Language Understanding (NLU) Technologies
  - Chatbot / Virtual Assistant
  - **Automatic Summarisation tools**
- **Text Analytics Technologies**
  - Text Mining tools
  - **Sentiment Analysis tools**
  - Text Prediction tools
  - Authorship Attribution tools
- **Multilingual and Semantic Search Technologies**
  - **Question Answering System**
  - **Search Engine**
- **Optical Character Recognition (OCR)**

Example Interpretation /Minutes …

Multilingual Meetings … Lectures…

- **We can: Listen and Transcribe spoken/audio signals (minutes)**
- **We can identify the speakers**
- **We can translate the transcription and/or interpret (speech to speech or Speech to Text translations)**

What is common to all these technologies

All are based on

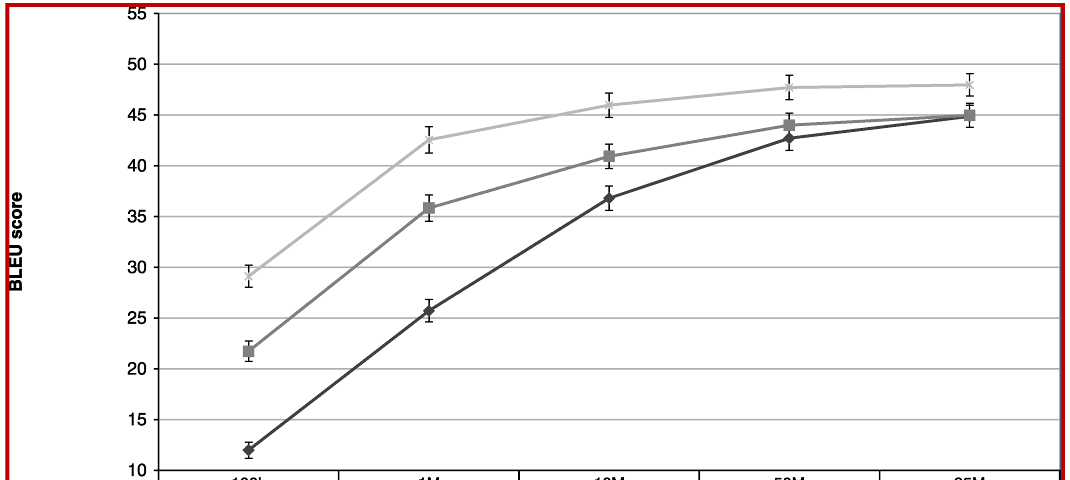
MACHINE (Deep) LEARNING FROM DATA

(AI and the DATA driven Paradigm)

IMPORTANCE OF DATA AND RE-USABILITY

✓ **Almost all technologies are data driven and based on statistical paradigms … (modeling based on huge amounts of date)**

Let us look at MT performance when "simply" adding data



	100k	1M	10M	50M	95M
Baseline	11.99	25.72	36.81	42.72	44.87
PlusExtracted	21.73	35.83	40.93	44	44.97
UpperBound	29.12	42.56	45.97	47.71	47.97

Initial Corpus Size

MT performance improvements for Arabic-English
(Courtesy Dragos Stefan Munteanu and Daniel Marcu)

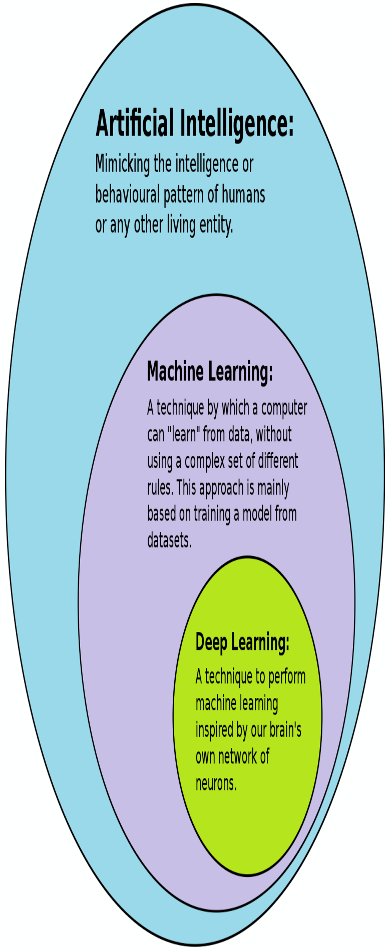
ARTIFICIAL INTELLIGENCE

WHAT IS AI? LANGUAGE AND INTELLIGENCE

"If a conversation with a device cannot be differentiated from a similar conversation with a human being, then the device can be called intelligent" *(Alan Turing, roughly)*

- How to apply this to (Human) Language ?
- Let us see for Machine Translation

Positioning AI, Machine Learning, Deep Learning



ARTIFICIAL INTELLIGENCE
Any technique that enables computers to mimic **human-like** intelligence.

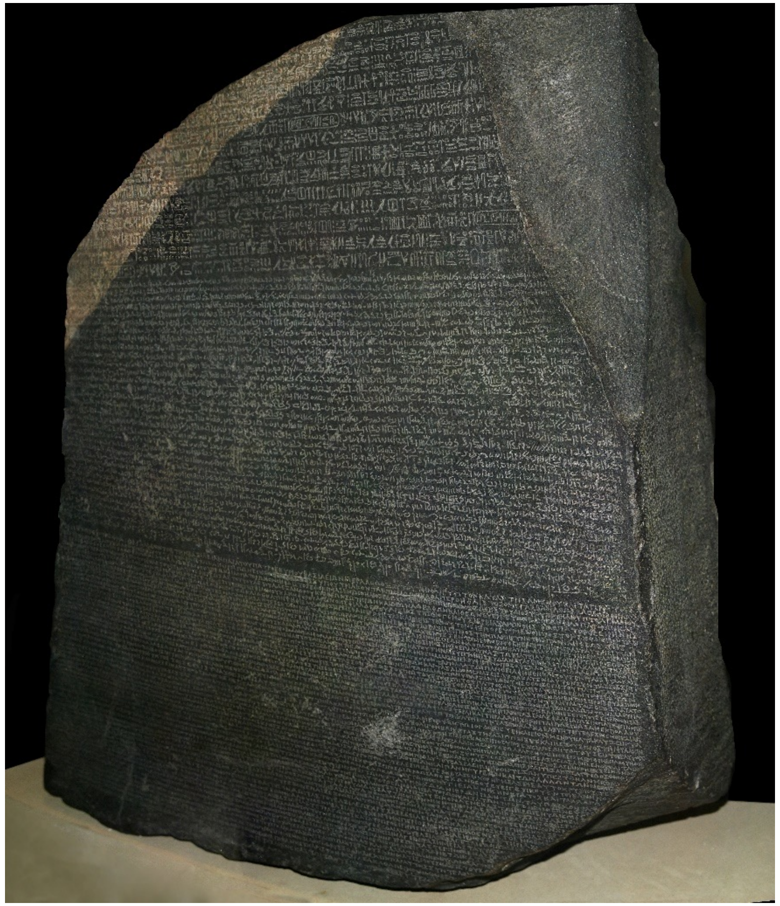
MACHINE LEARNING
A subset of Artificial Intelligence that includes complex statistical techniques that enable machines to learn – improve at tasks with experience – but without being explicitly programmed to do so. There are various types of machine learning, including supervised learning, unsupervised learning and renforcement learning.

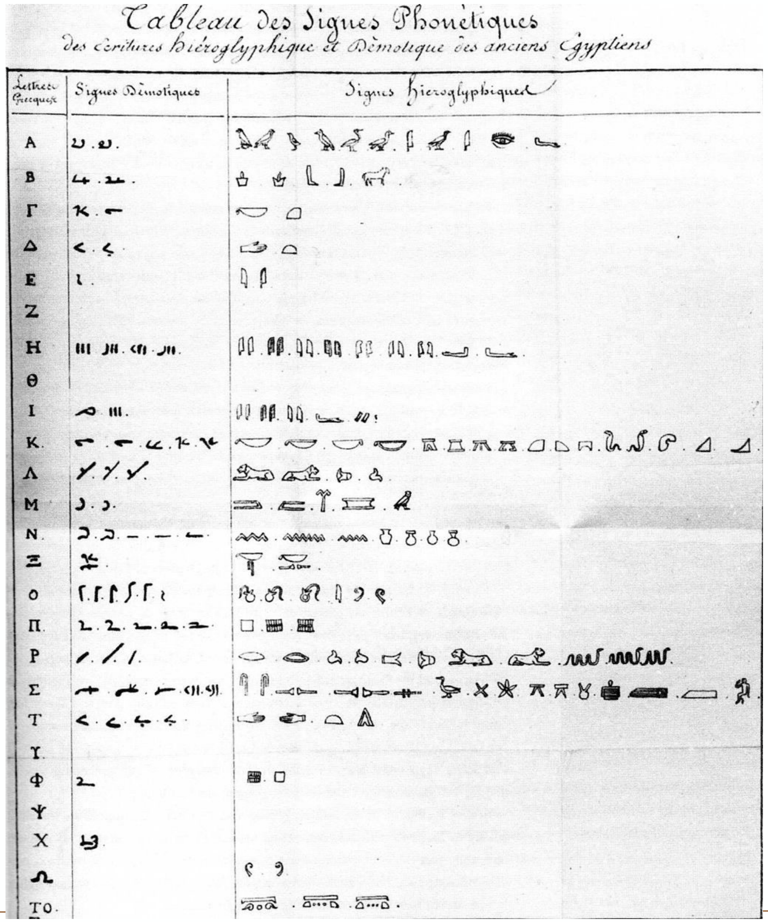
DEEP LEARNING
A subset of Machine Learning composed of algorithms that permit software to train itself to perform tasks (like speech and image recognition). Inspired by the human brain, deep learning works by exposing multi-layered neural networks to vast amounts of data.

1950s 1960s 1970s 1980s 1990s 2000s 2010s

MT and the different ages

Champollion & The Rosetta Stone





How MT can Learn from Data?

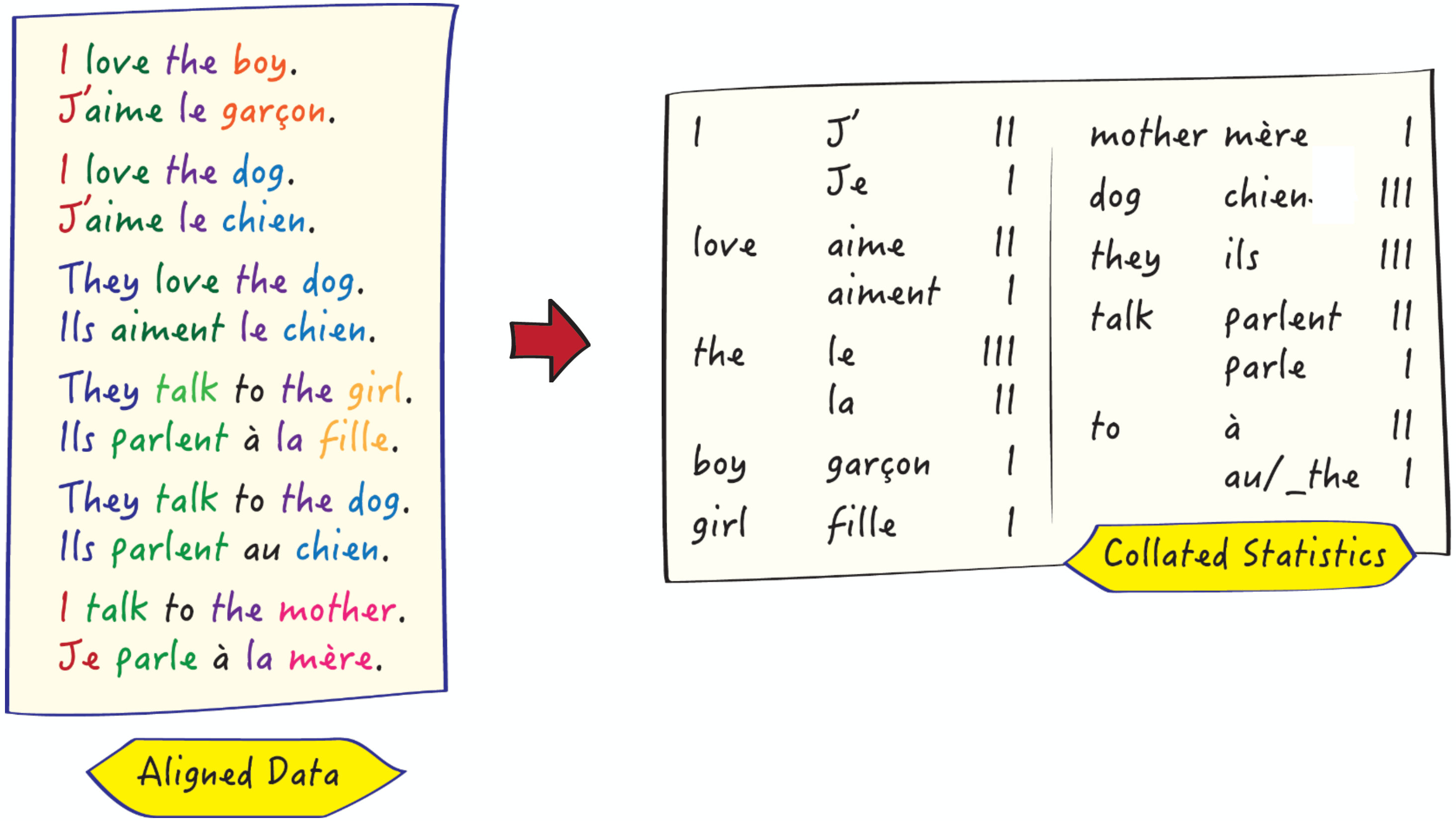
Statistical MT learns from data:
- Source documents and their human translations
- Target language collections

The more data the better!
Also: the right kind of data!

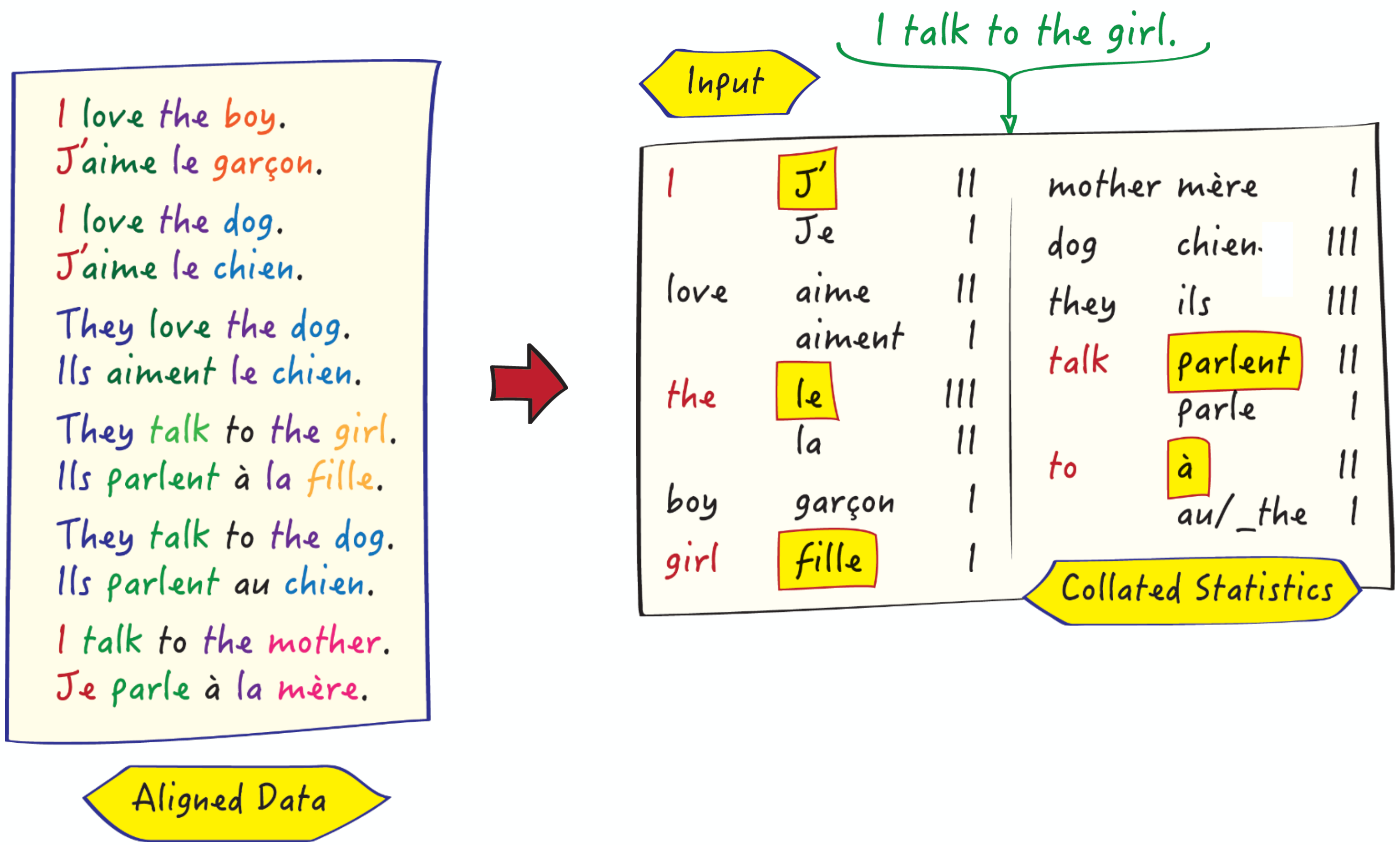
- Which sentences translate as which: sentence alignment
- Which words translate as which: word alignment + translation probabilities => translation model
- What do good target sentences look like: language model

GERMAN	ENGLISH	FRENCH
Einleitung	Introduction	Introduction
I. Von dem Unterschiede der reinen und empirischen Erkenntnis	*I. Of the difference between Pure and Empirical Knowledge*	*I. De la différence de la connaissance pure et de la connaissance empirique.*
Daß alle unsere Erkenntnis mit der Erfahrung anfange, daran ist gar kein Zweifel; denn wodurch sollte das Erkenntnisvermögen sonst zur Ausübung erweckt werden, geschähe es nicht durch Gegenstände, die unsere Sinne rühren und teils von selbst Vorstellungen bewirken, teils unsere Verstandestätigkeit in Bewegung bringen, diese zu vergleichen, sie zu verknüpfen oder zu trennen, und so den rohen Stoff sinnlicher Eindrücke zu einer Erkenntnis der Gegenstände zu verarbeiten, die Erfahrung heißt? Der Zeit nach geht also keine Erkenntnis in uns vor der Erfahrung vorher, und mit dieser fängt alle an.	That all our knowledge begins with experience there can be no doubt. For how is it possible that the faculty of cognition should be awakened into exercise otherwise than by means of objects which affect our senses, and partly of themselves produce representations, partly rouse our powers of understanding into activity, to compare to connect, or to separate these, and so to convert the raw material of our sensuous impressions into a knowledge of objects, which is called experience? In respect of time, therefore, no knowledge of ours is antecedent to experience, but begins with it.	Que toute notre connaissance commence avec l'expérience, cela ne soulève aucun doute. En effet, par quoi notre pouvoir de connaître pourrait-il être éveillé et mis en action, si ce n'est par des objets qui frappent nos sens et qui, d'une part, produisent par eux-mêmes des représentations et, d'autre part, mettent en mouvement notre faculté intellectuelle, afin qu'elle compare, lie ou sépare ces représentations, et travaille ainsi la matière brute des impressions sensibles pour en tirer une connaissance des objets, celle qu'on nomme l'expérience? Ainsi, chronologiquement, aucune connaissance ne précède en nous l'expérience et c'est avec elle que toutes commencent.

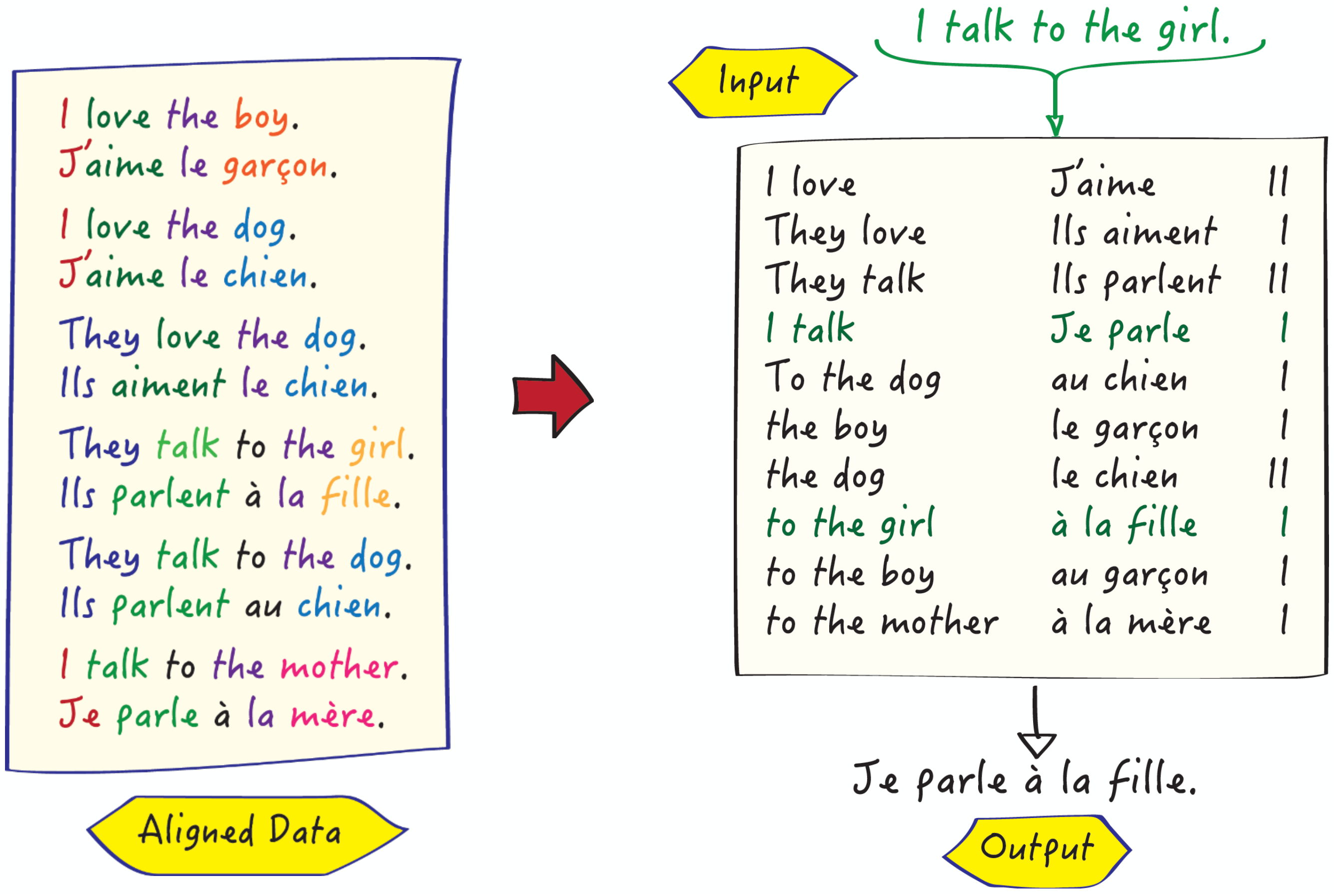
Statistical Machine Translation



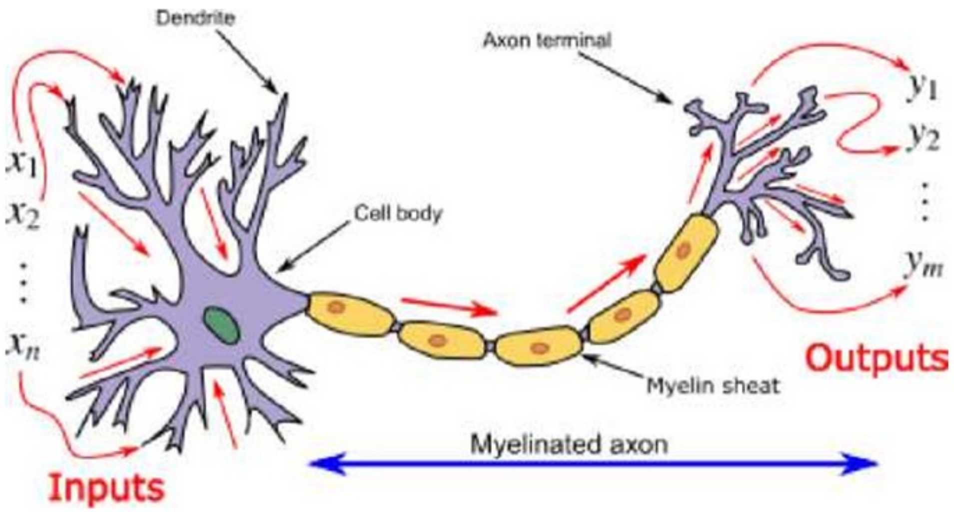
STATISTICAL MACHINE TRANSLATION

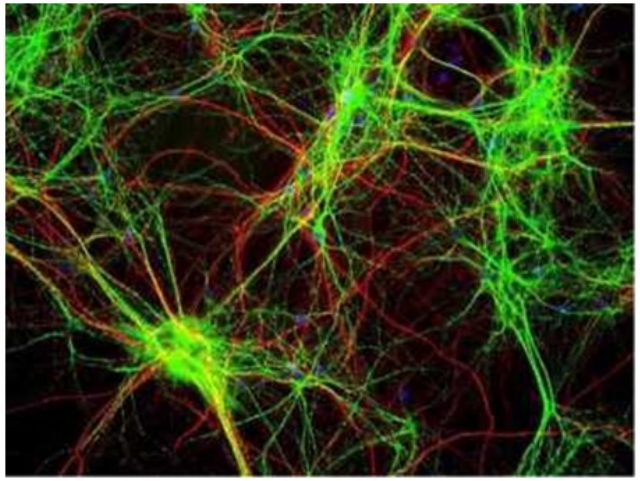


PHRASE-BASED SMT

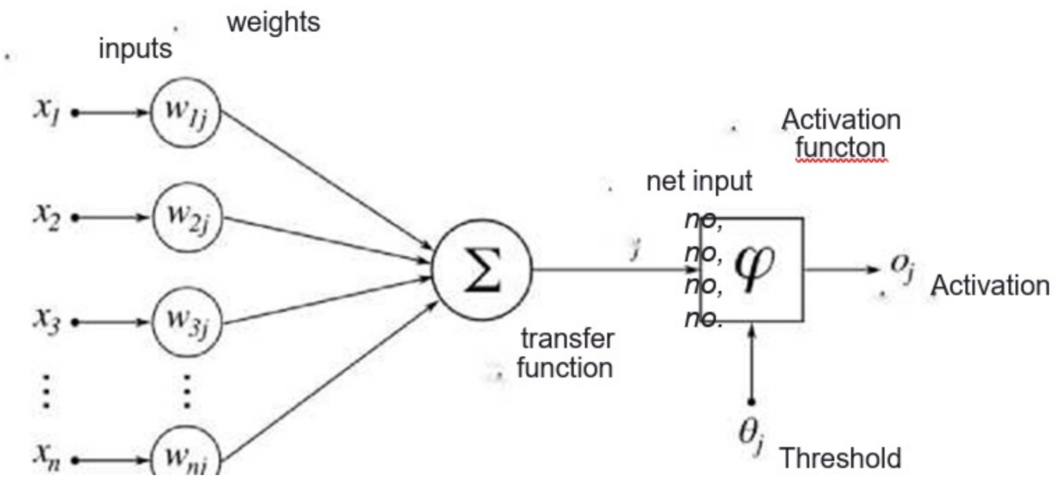


HOW MACHINES ARE TRANSLATING (NEURAL NETWORK AGE)





Source: Wikimedia



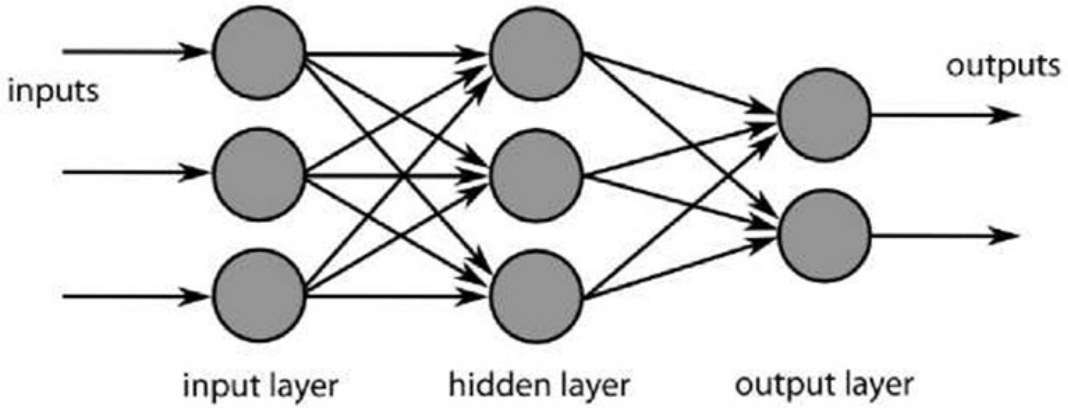
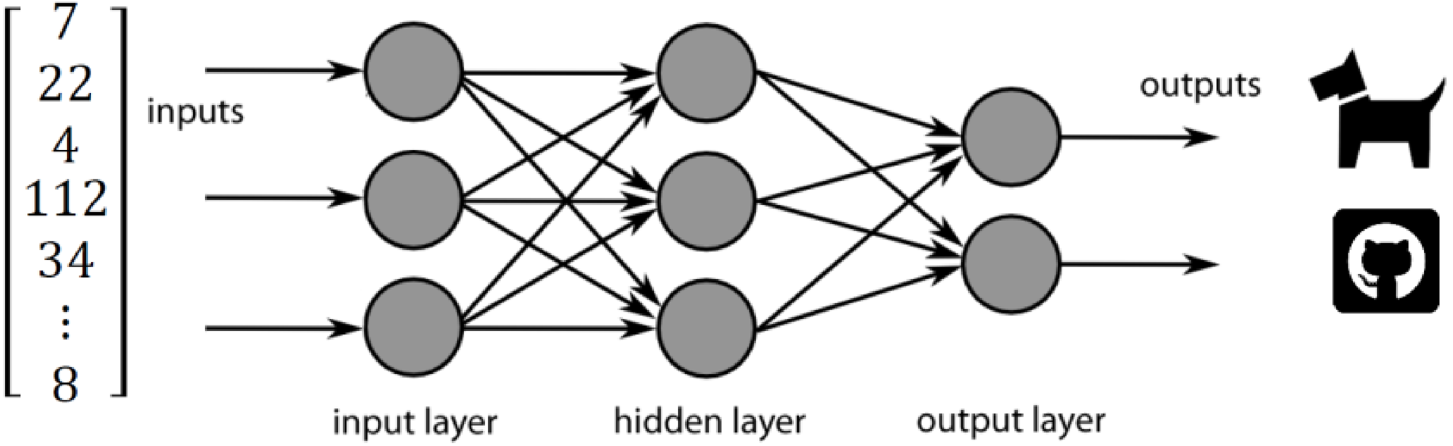
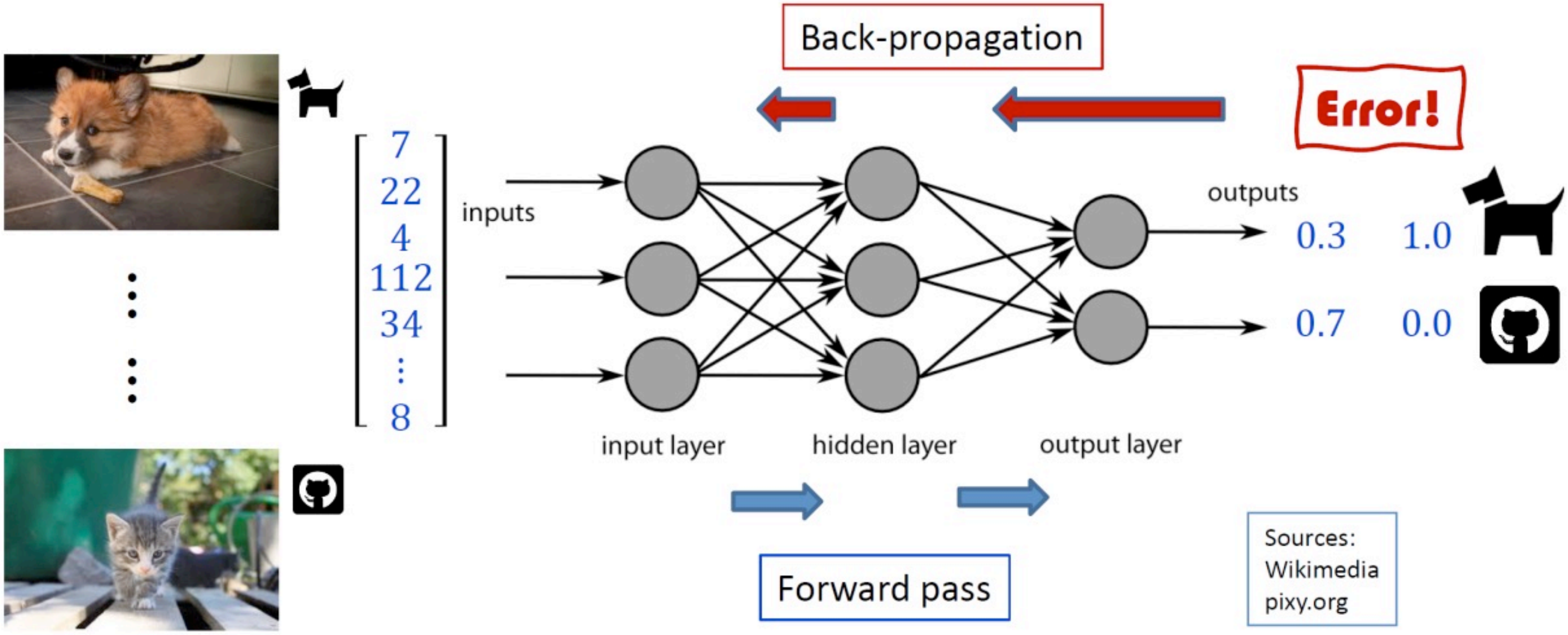
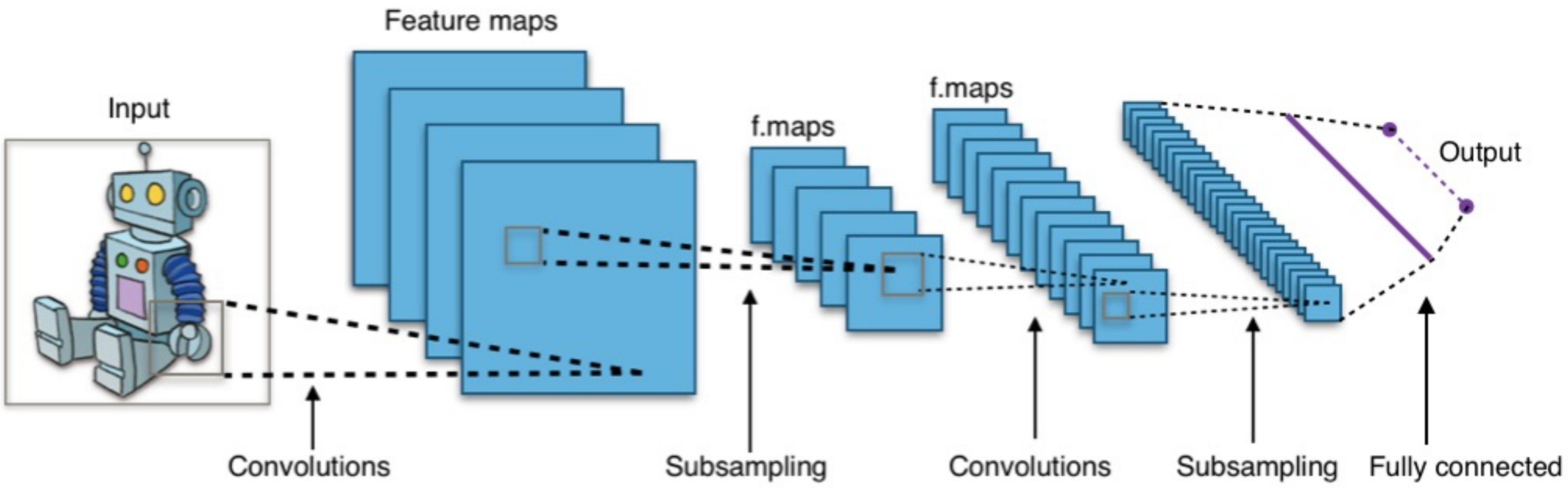


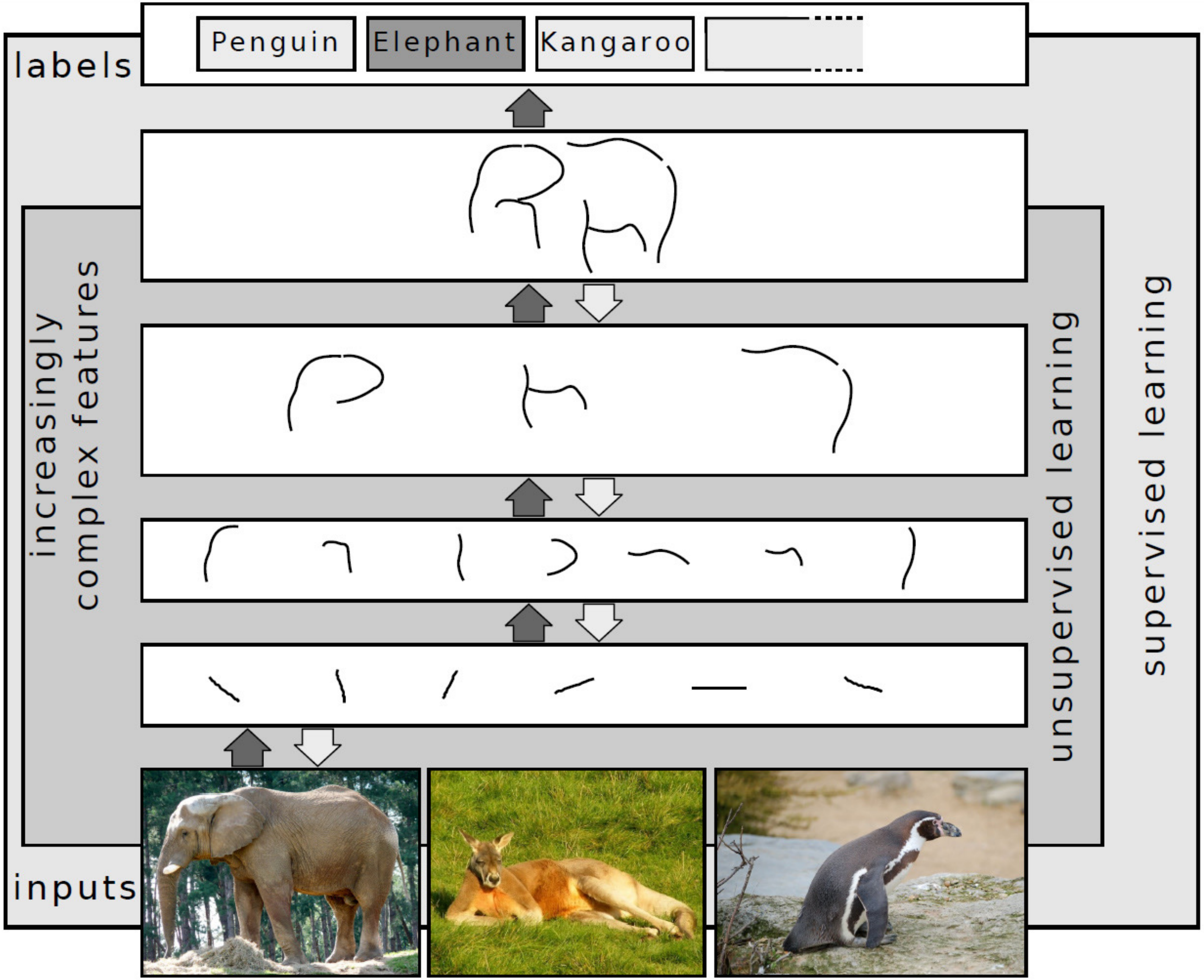
Illustration of Neural Network Image Recognition ...



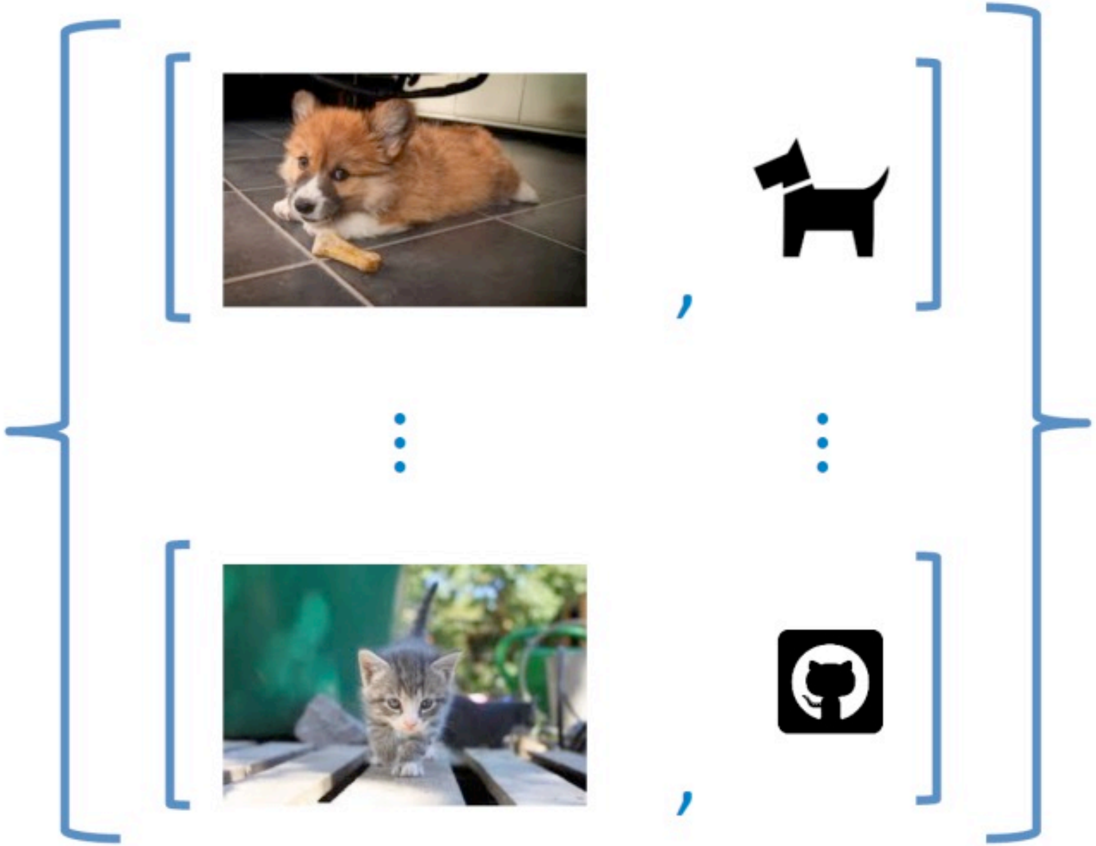
Sources:
Wikimedia
pixy.org


Back-propagation
Error!
inputs
7
22
4
112
34
⋮
8
outputs
0.3
1.0
0.7
0.0
input layer
hidden layer
output layer
Forward pass
Sources:
Wikimedia
pixy.org


Feature maps
Input
f.maps
f.maps
Output
Convolutions
Subsampling
Convolutions
Subsampling
Fully connected


labels
Penguin
Elephant
Kangaroo
increasingly complex features
unsupervised learning
supervised learning
inputs

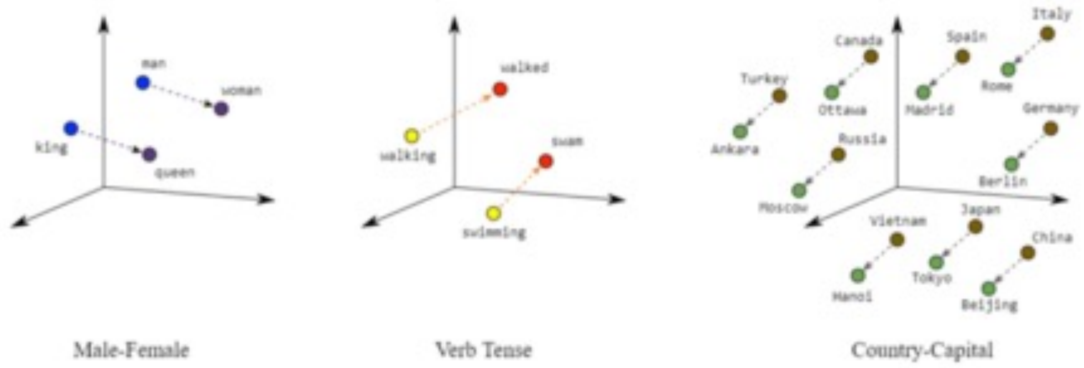
DATA IS THE BASIS …. FUEL



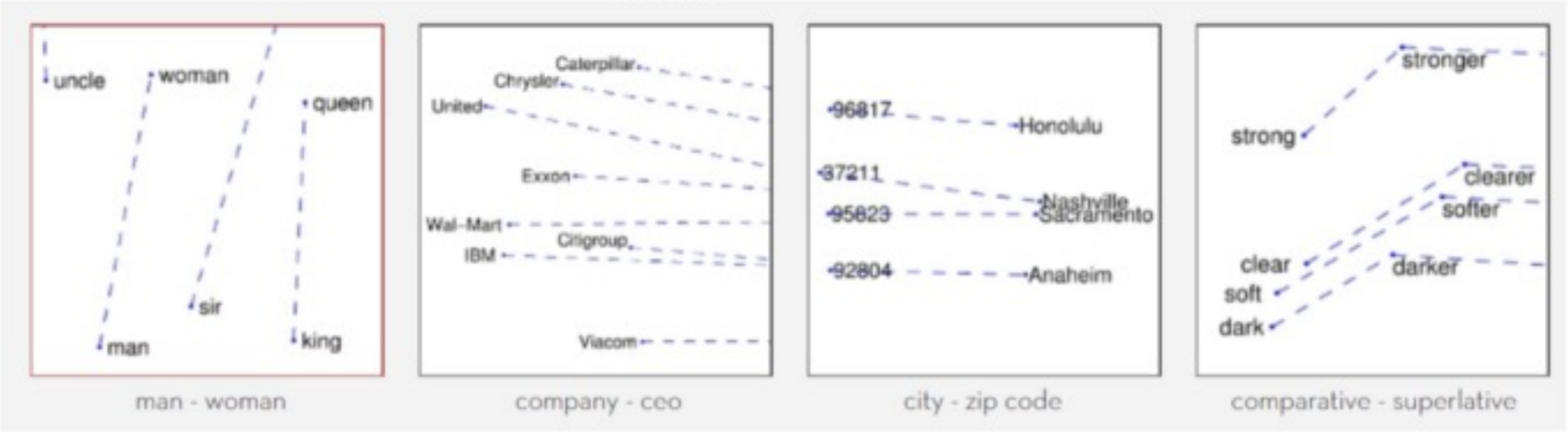
Supervised ML
Labelled training data
[data , label]
Label = supervision signal

WORD-EMBEDDINGS-FOR-NLP

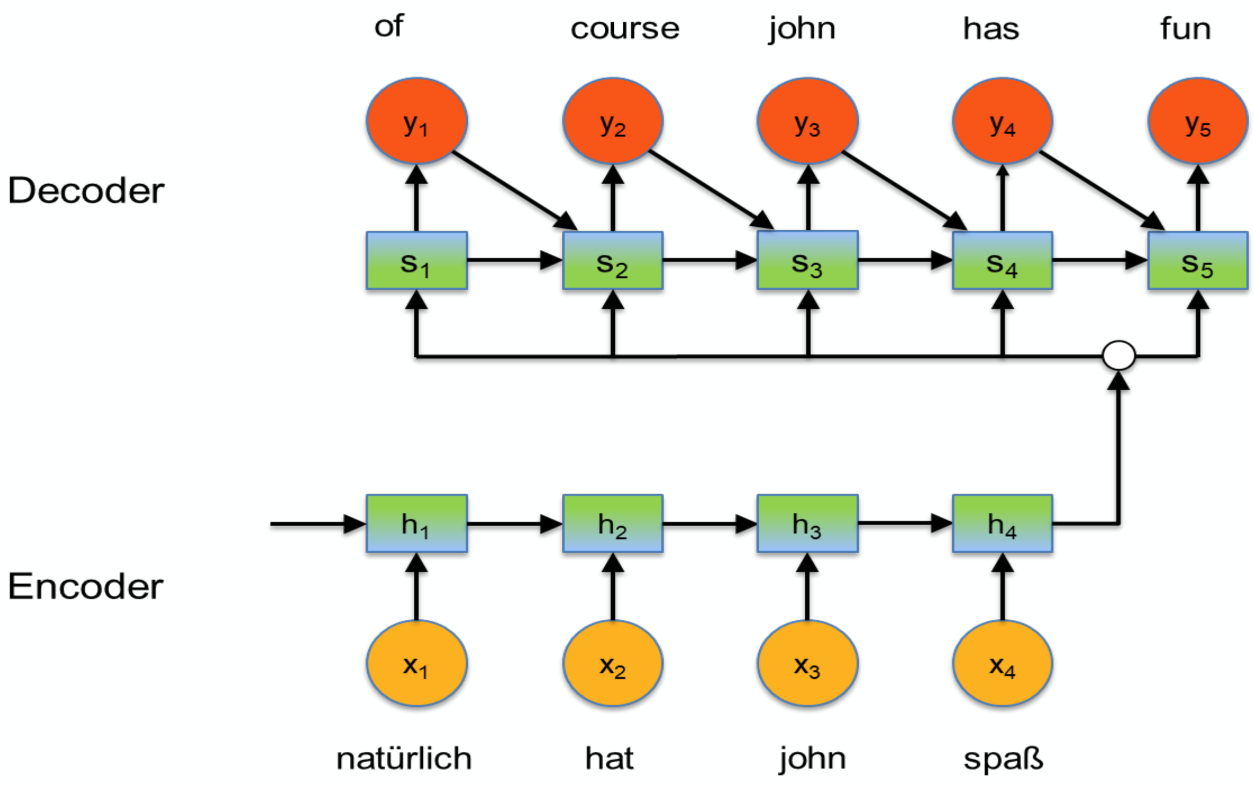
Word2Vec



GloVe



How machines translate today (state of the art):



MT and Human Parity?

http://www.statmt.org/wmt19/pdf/53/WMT01.pdf

WMT 2019, Florence, Italy
Example for News Translation Task

English→German

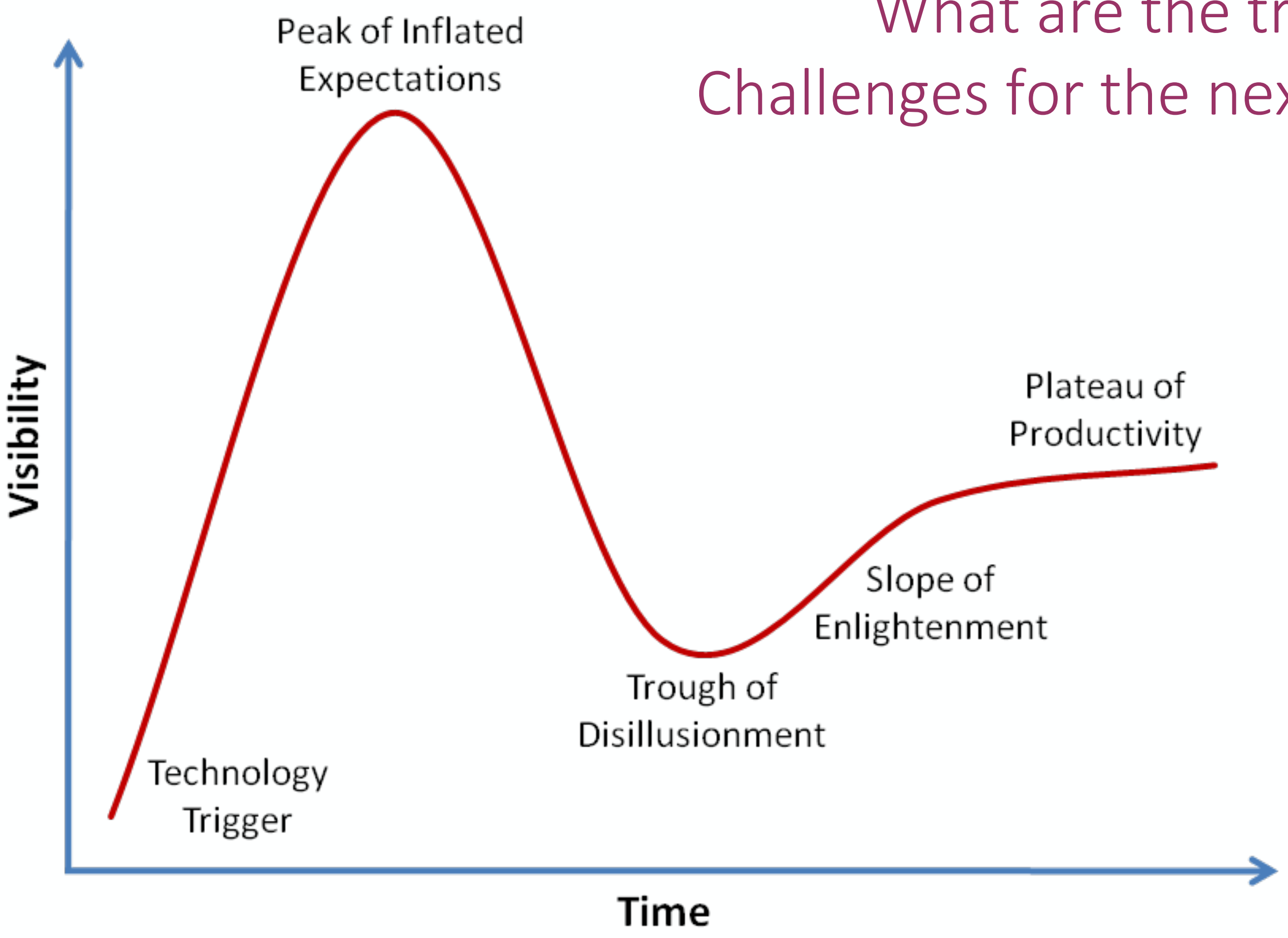
Ave.	Ave. z	System
90.3	0.347	Facebook-FAIR
93.0	0.311	Microsoft-WMT19-sent-doc
92.6	0.296	Microsoft-WMT19-doc-level
90.3	0.240	HUMAN
87.6	0.214	MSRA-MADL
88.7	0.213	UCAM
89.6	0.208	NEU
87.5	0.189	MLLP-UPV
87.5	0.130	eTranslation
86.8	0.119	dfki-nmt
84.2	0.094	online-B
86.6	0.094	Microsoft-WMT19-sent-level
87.3	0.081	JHU
84.4	0.077	Helsinki-NLP
84.2	0.038	online-Y
83.7	0.010	lmu-ctx-tf-single
84.1	0.001	PROMT-NMT
82.8	−0.072	online-A
82.7	−0.119	online-G
80.3	−0.129	UdS-DFKI
82.4	−0.132	TartuNLP-c
76.3	−0.400	online-X
43.3	−1.769	en-de-task

English→Lithuanian

Ave.	Ave. z	System
90.5	1.017	HUMAN
72.8	0.388	tilde-nc-nmt
69.1	0.387	MSRA-MASS-uc
68.0	0.262	tilde-c-nmt
68.2	0.259	MSRA-MASS-c
67.7	0.155	GTCOM-Primary
62.7	0.036	eTranslation
59.6	−0.054	NEU
57.4	−0.061	online-B
47.8	−0.383	TartuNLP-c
38.4	−0.620	online-A
39.2	−0.666	online-X
32.6	−0.805	online-G

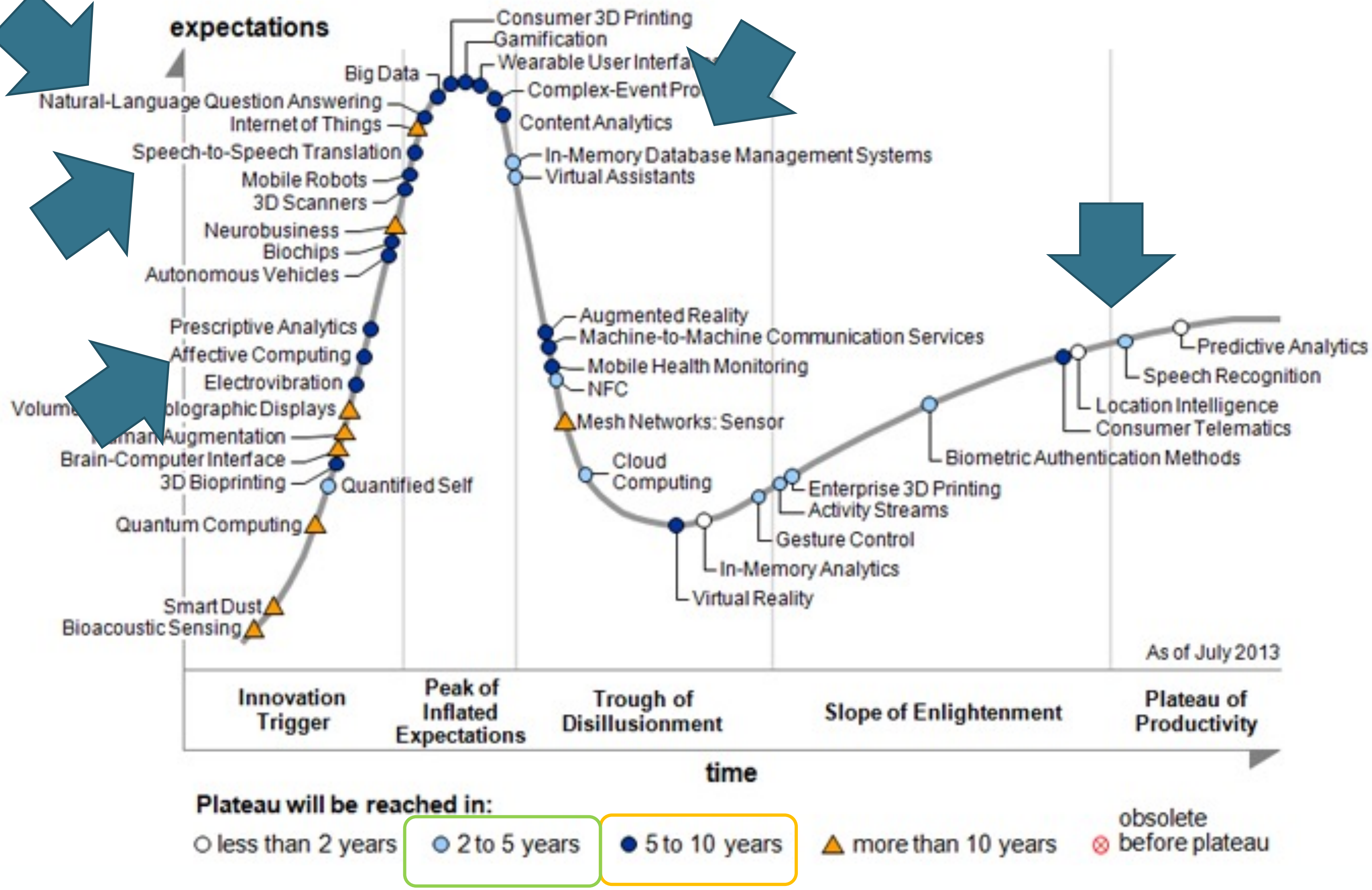
WHERE DO WE STAND TODAY …
TECHNO TRENDS

What are the trends …
Challenges for the next « decade! »



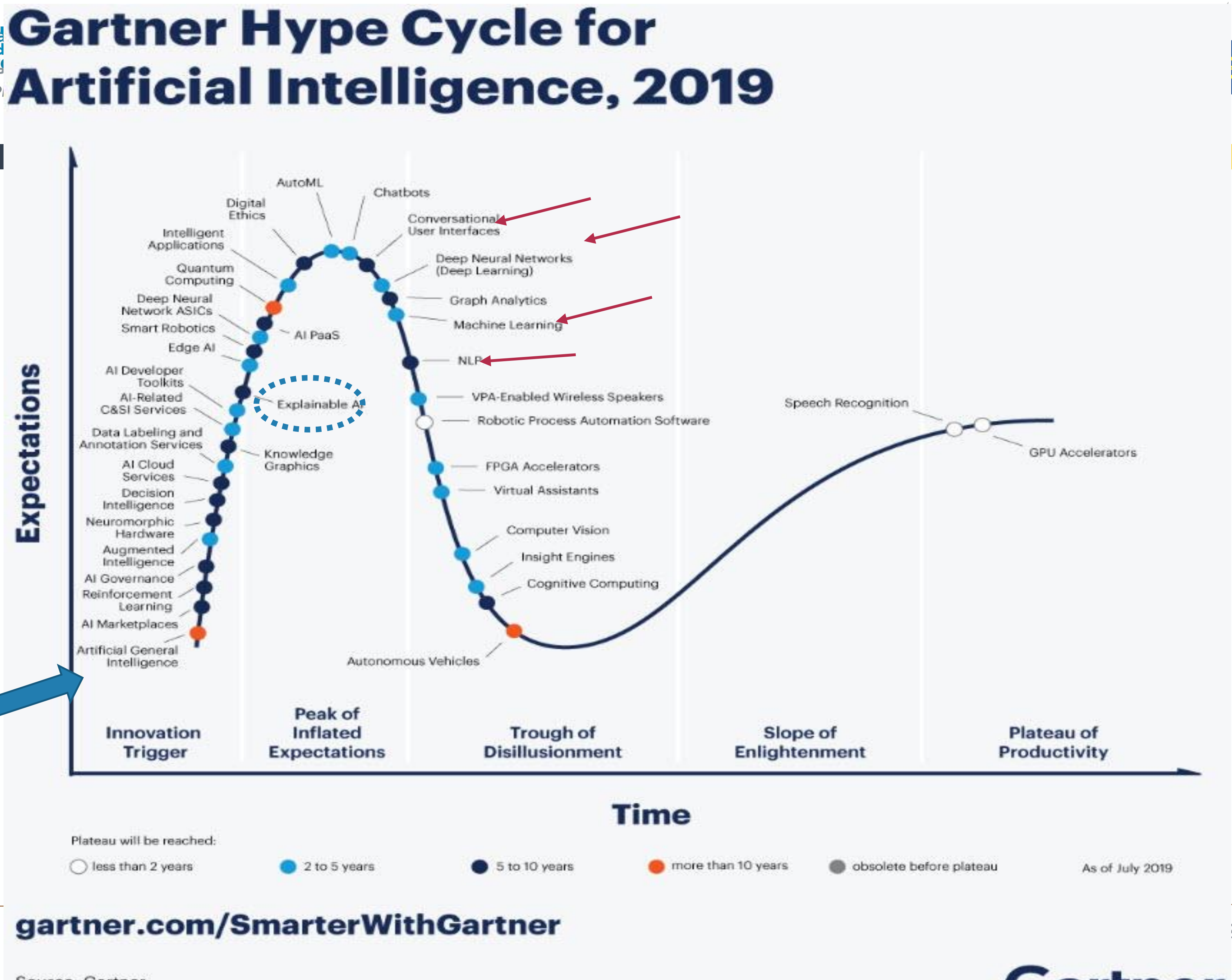
The Gartner Hype Cycle
https://www.gartner.com/smarterwithgartner/5-trends-drive-the-gartner-hype-cycle-for-emerging-technologies-2020/

European Commission

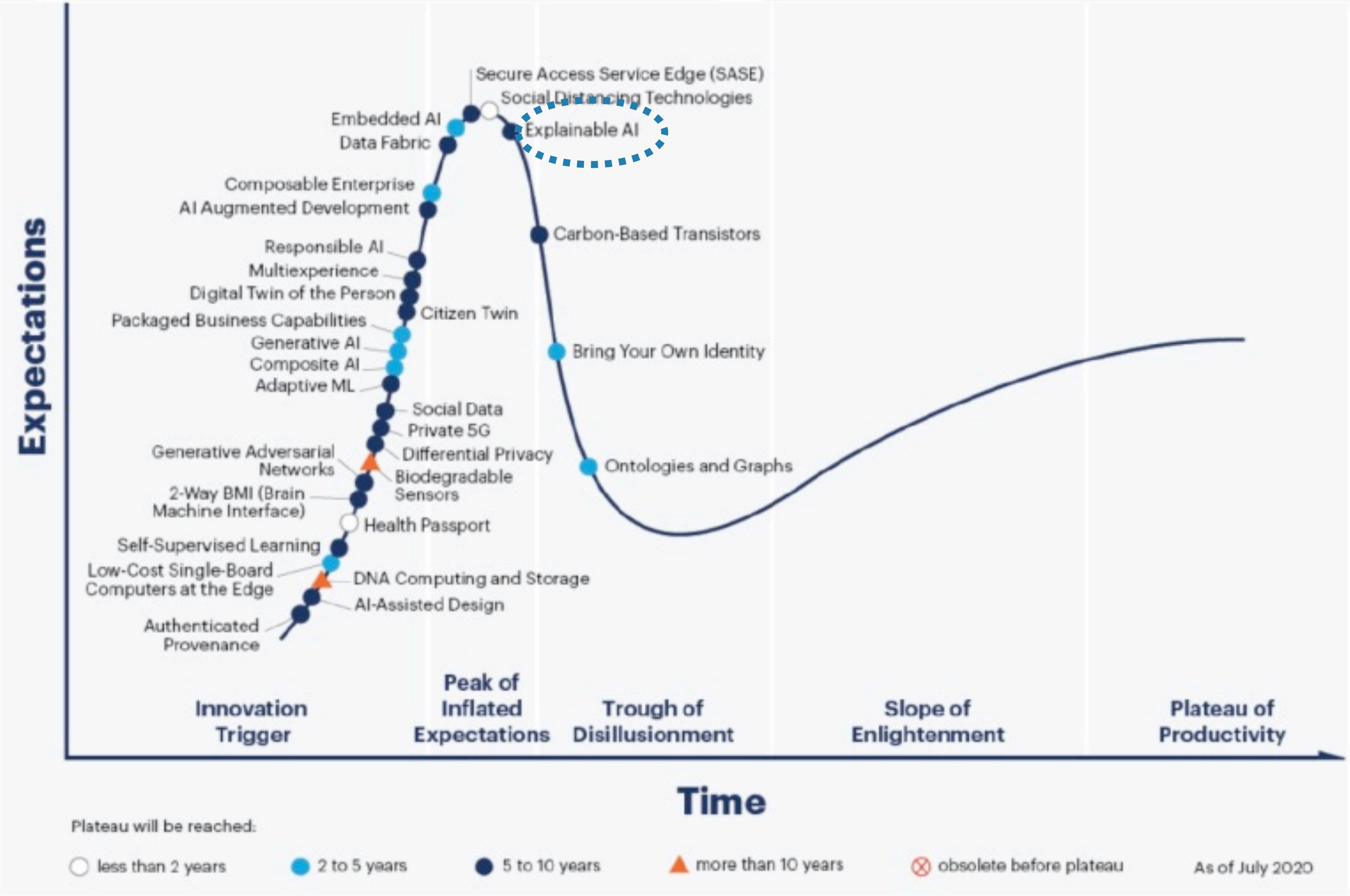


Source: Gartner August 2013

The 2013 Emerging Technologies Hype Cycle highlights technologies


Gartner Hype Cycle for Artificial Intelligence, 2019
Expectations
AutoML
Digital Ethics
Chatbots
Conversational User Interfaces
Intelligent Applications
Quantum Computing
Deep Neural Networks (Deep Learning)
Graph Analytics
Deep Neural Network ASICs
Machine Learning
Smart Robotics
AI PaaS
Edge AI
NLP
AI Developer Toolkits
Explainable AI
VPA-Enabled Wireless Speakers
AI-Related C&SI Services
Robotic Process Automation Software
Speech Recognition
Data Labeling and Annotation Services
Knowledge Graphics
GPU Accelerators
AI Cloud Services
FPGA Accelerators
Decision Intelligence
Virtual Assistants
Neuromorphic Hardware
Computer Vision
Augmented Intelligence
Insight Engines
AI Governance
Cognitive Computing
Reinforcement Learning
AI Marketplaces
Artificial General Intelligence
Autonomous Vehicles
Innovation Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
Time
Plateau will be reached:
less than 2 years
2 to 5 years
5 to 10 years
more than 10 years
obsolete before plateau
As of July 2019
gartner.com/SmarterWithGartner

AND 2020


Expectations
Secure Access Service Edge (SASE)
Social Distancing Technologies
Explainable AI
Embedded AI
Data Fabric
Composable Enterprise
AI Augmented Development
Responsible AI
Multiexperience
Digital Twin of the Person
Citizen Twin
Packaged Business Capabilities
Generative AI
Composite AI
Adaptive ML
Social Data
Private 5G
Differential Privacy
Generative Adversarial Networks
Biodegradable Sensors
2-Way BMI (Brain Machine Interface)
Health Passport
Self-Supervised Learning
Low-Cost Single-Board Computers at the Edge
DNA Computing and Storage
AI-Assisted Design
Authenticated Provenance
Carbon-Based Transistors
Bring Your Own Identity
Ontologies and Graphs
Innovation Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
Time
Plateau will be reached:
less than 2 years
2 to 5 years
5 to 10 years
more than 10 years
obsolete before plateau
As of July 2020

When Will AI Exceed Human Performance?

"High-level machine intelligence" (HLMI) is achieved

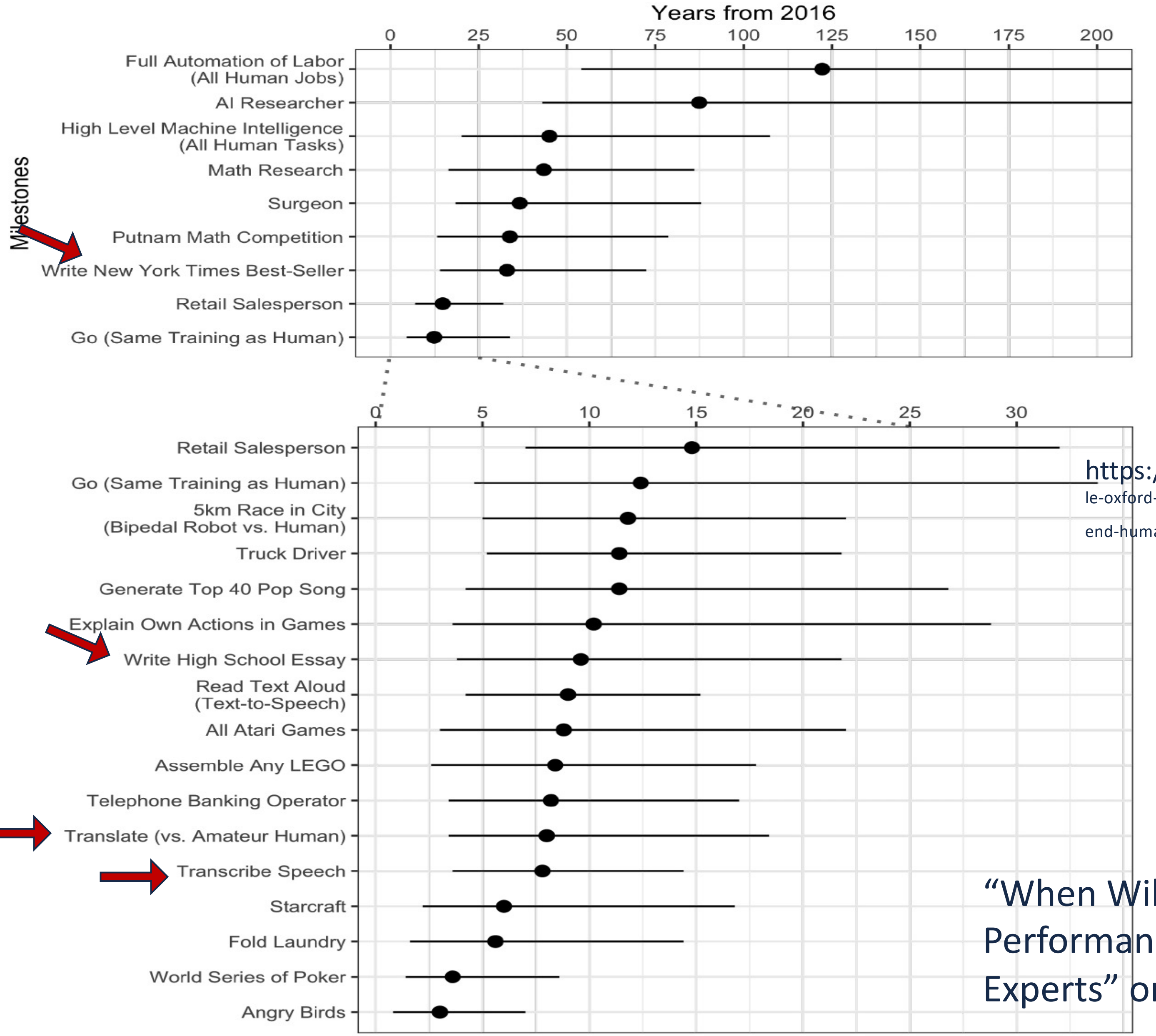
when unaided machines can accomplish every task

better and more cheaply than human workers.

Results from Surveys and experts' opinions

European Language Resource Coordination
Connecting Europe Fa

European Commission



https://slator.com/academia/yale-oxford-enter-business-predicting-end-human-translator/

"When Will AI Exceed Human Performance? Evidence from AI Experts" on May 24, 2017.

European Language Resource Coordination
Connecting Europe Fa

“High-level machine intelligence” (HLMI) is achieved when unaided machines can accomplish every task better and more cheaply than human workers.

European Commission

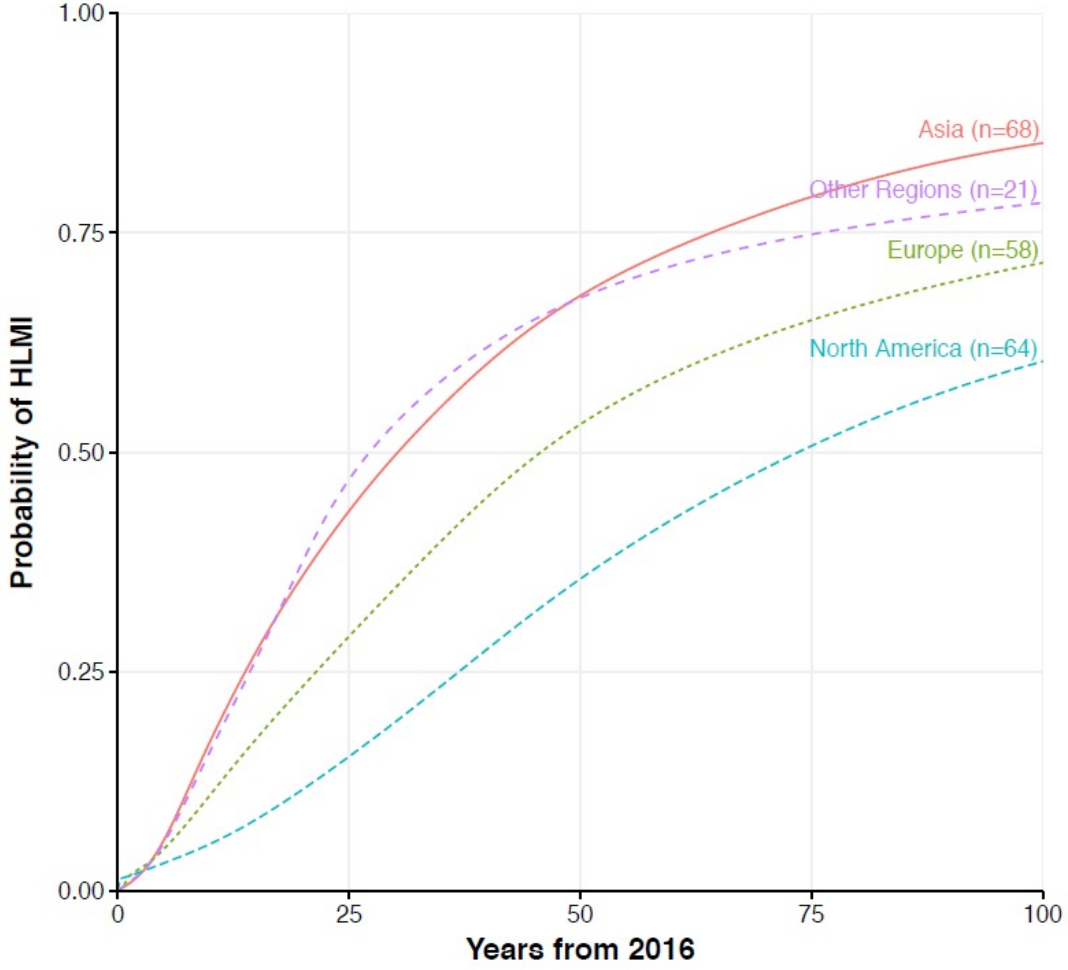


Figure 3: Aggregate Forecast (computed as in Figure 1) for HLMI, grouped by region in which respondent was an undergraduate. Additional regions (Middle East, S. America, Africa, Oceania) had much smaller numbers and are grouped as “Other Regions.”

TRENDS AND LONG TERM PERSPECTIVES

- New hot topics and trends
  - More languages (not only ~300 out of the 7000) , under-resourced,
    - see UNESCO Decade of activities on Indigeneous Languages (LT4ALL initiative

      https://en.unesco.org/LT4All & Proceedings of 2019: https://lt4all.org/en/)
    - European Language Equality project (https://european-language-equality.eu/)
  - Focus on social networks and other media ….
    - Hate speech detection and media monitoring

AI & LT FOR EU

- Identify strategic sectors with EU strength e.g. Multilingualism
- Develop an EU-centric LT and data policies with
  - international partnerships
  - Not only Market-driven
  - Particular attention to non-official languages
- Easy to understand AI regulations (AI transparency)
- Real funding for EU players (e.g. Public Procurements)

The language of Europe is translation.

Assises de la traduction littéraire à Arles (France) le 14 novembre 1993,

European Language
Resource Coordination
Connecting Europe Facility

European Commission

Website: www.lr-coordination.eu

Twitter: @LR_Coordination

Email: info@lr-coordination.eu

THE ROLE OF MACHINE TRANSLATION

MT is the <u>only viable solution</u> for:

- quick and cheap access to information in foreign languages.
- understanding information received in a foreign language that otherwise could not be used or would require substantial time and costs to translate.
- making multilingual use of websites possible
- facilitating cross-lingual information search and analytics.

That is why machine translation (MT) is a
critically important technology for multilingual Europe

MACHINE TRANSLATION USERS

- Do not understand the source language
  - Quick gisting (requiring validation)
  - Decide on relevance and routing
- Understand both source and target language
  - Professional translators
    - Use as input for HQ translation
  - Other expert users
    - **Post-edit for consultation** (e.g. in international working groups)

Decide

Publish

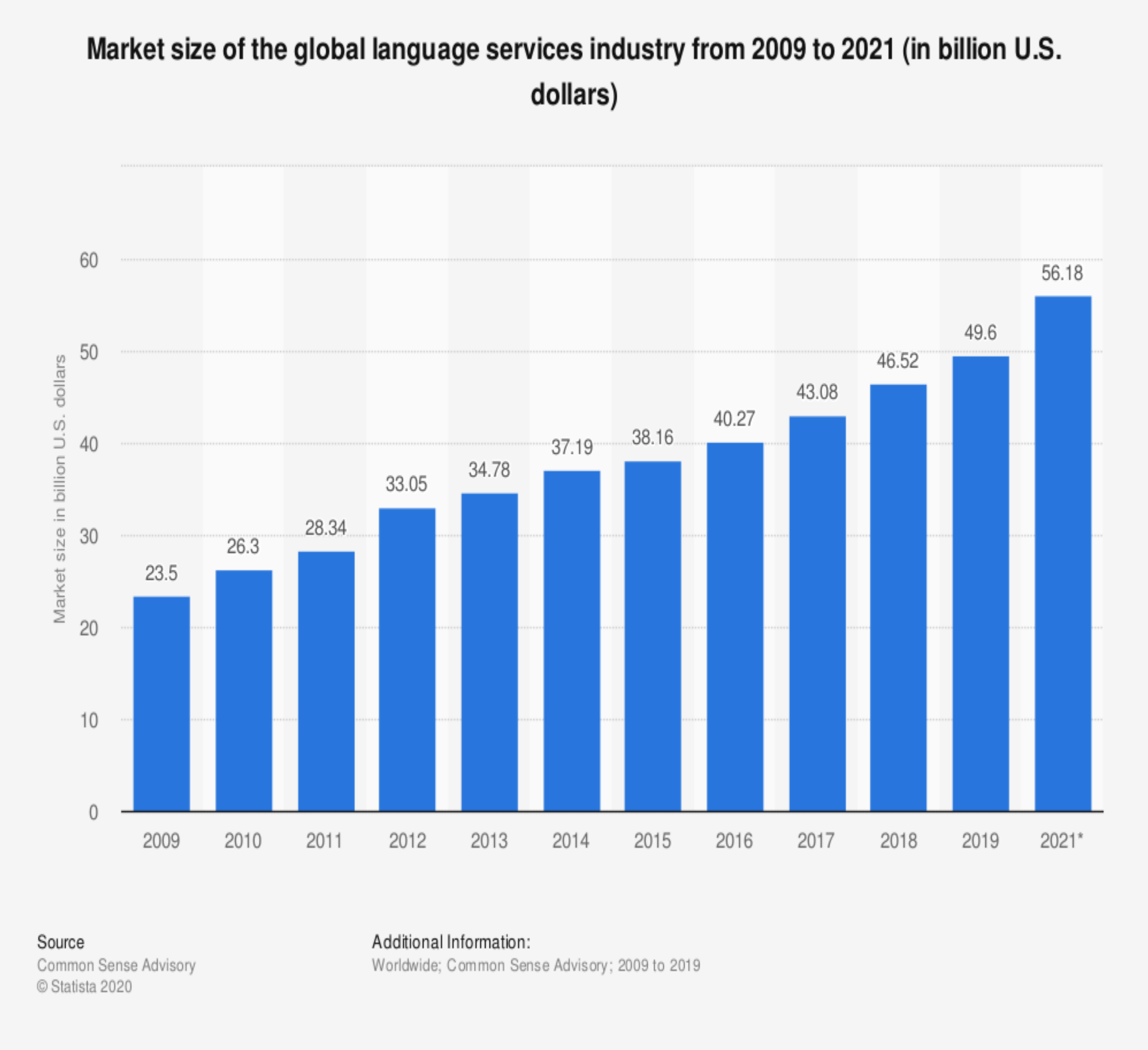
SKYPE TRANSLATOR

Language Technologies' Market

- Difficulty to define the market perimeter
- Often market research institutions compile and consolidate data from different segments inc. non-technological ones (human translations, localization, etc.)
- Different timelines
- Different geographical areas
- The most lucrative ones:
  - Machine Translation technology
  - Speech technologies
  - Multilingual and semantic search technology
  - Text and Speech Analytics

European Language Resource Coordination
Connecting Europe Facility

European Commission




Global Market for Natural Language Processing (NLP)
Market forecast to grow at CAGR of 10.3%
USD 13 Billion
USD 25.7 Billion
2020
2027
https://www.researchandmarkets.com/reports/3502818
RESEARCH AND MARKETS
THE WORLD'S LARGEST MARKET RESEARCH STORE

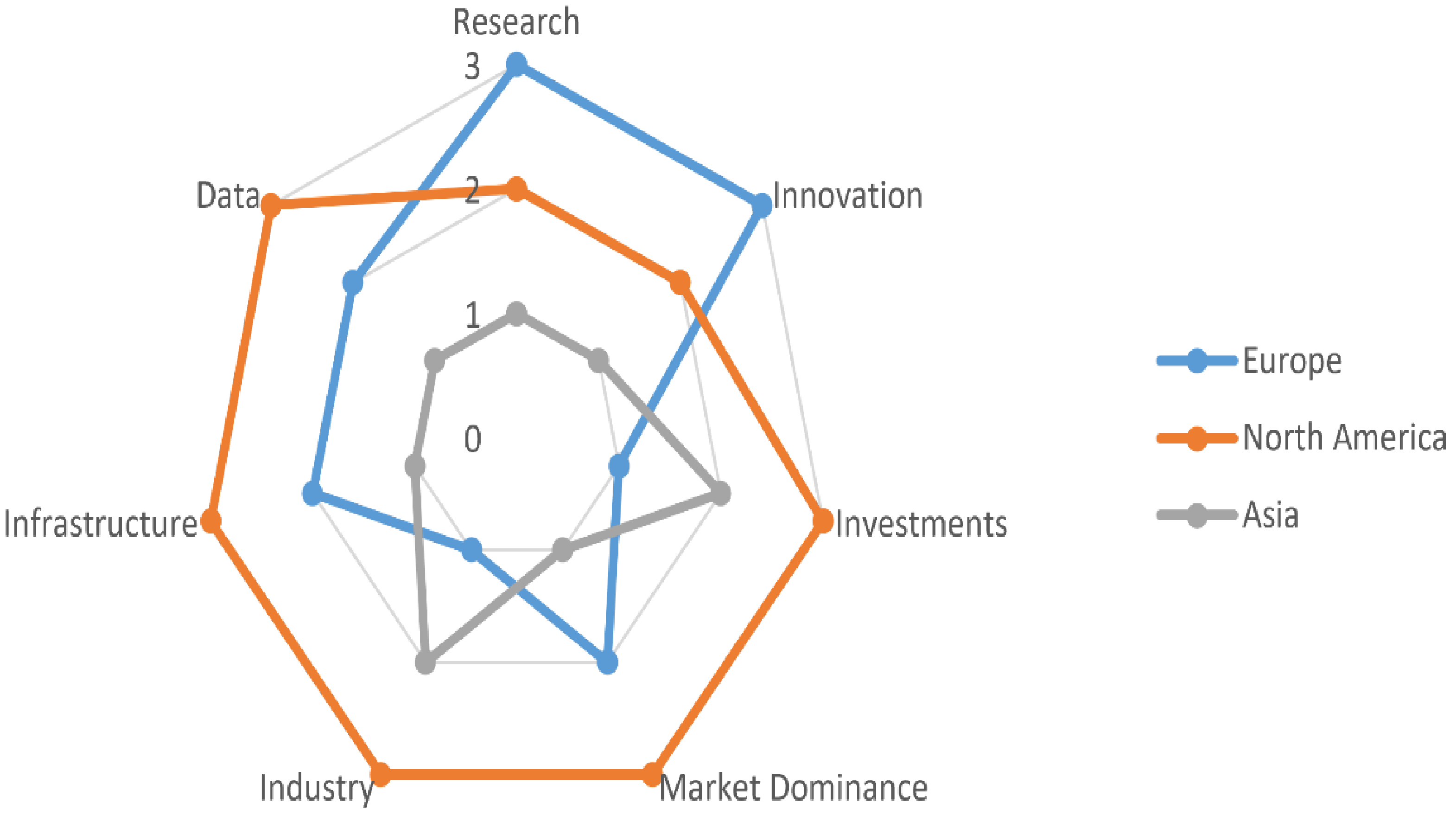
Market dimensions and Analysis

We have identified the following 7 dimensions to decompose the LT markets:

- Research
- Innovations
- Investments
- Market dominance
- Industry
- Infrastructure
- Open data

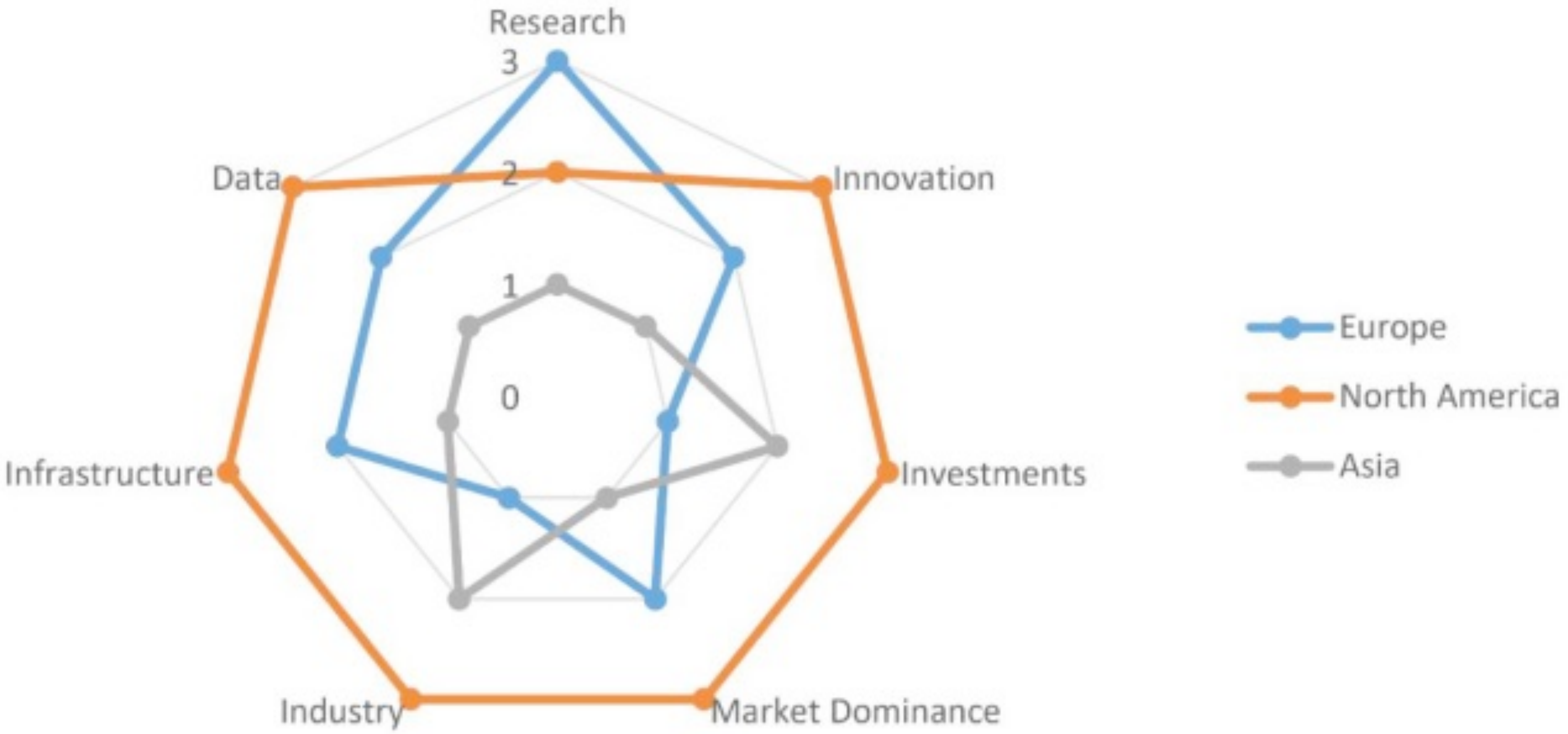
Market analyzed in the context of global competitiveness, highlighting particularly the most important achievements and gaps of the LT ecosystem

MACHINE TRANSLATION AREA

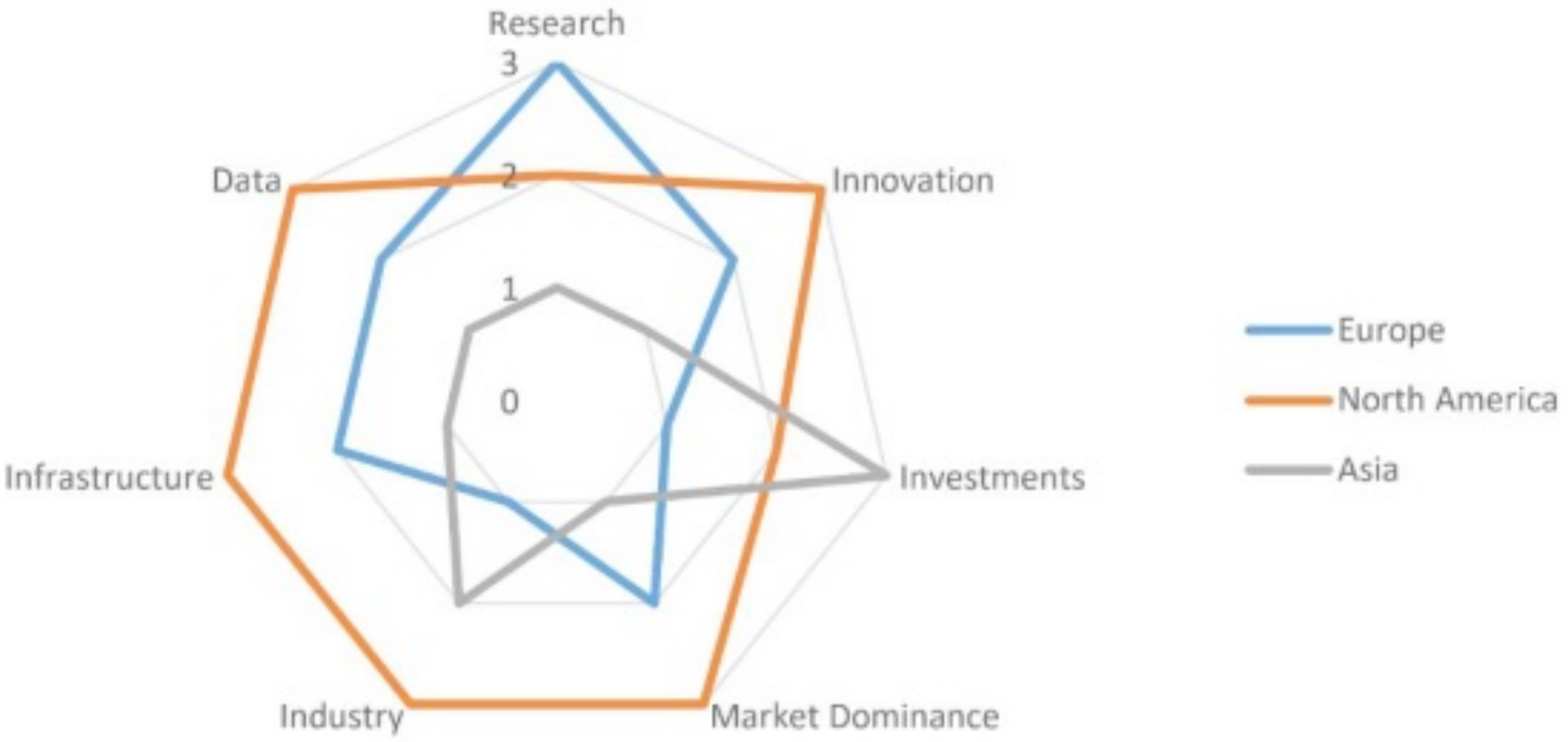


- Research
- Innovations
- Investments
- Market dominance
- Industry
- Infrastructure
- Open data

SPEECH TECHNOLOGIES



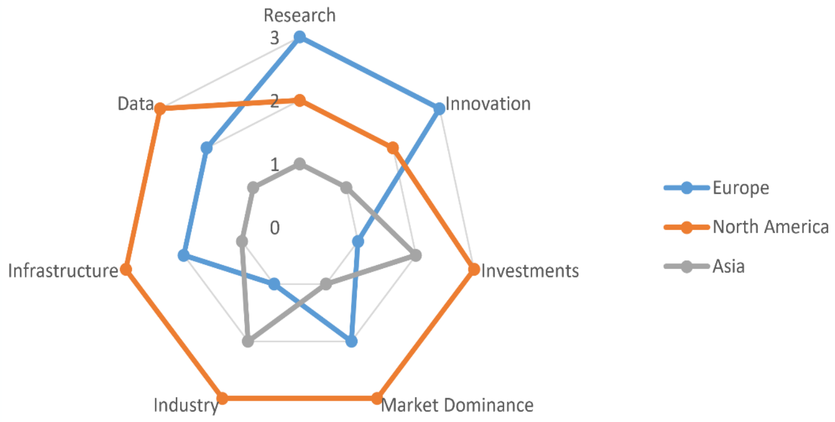
Search Technologies



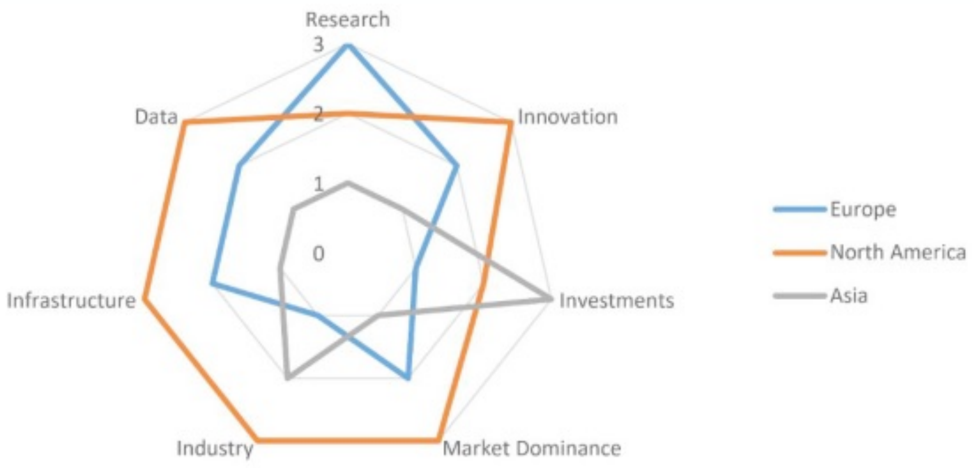
Speech, Search and Translation Technologies

Speech Technologies

Translation Technologies



Search Technologies



- ➢ (Secretary General of) European Language Resources Association [ELRA]

- an infrastructure for **Language Resources (LRs) sharing & Technology evaluation**

- Created in February 1995

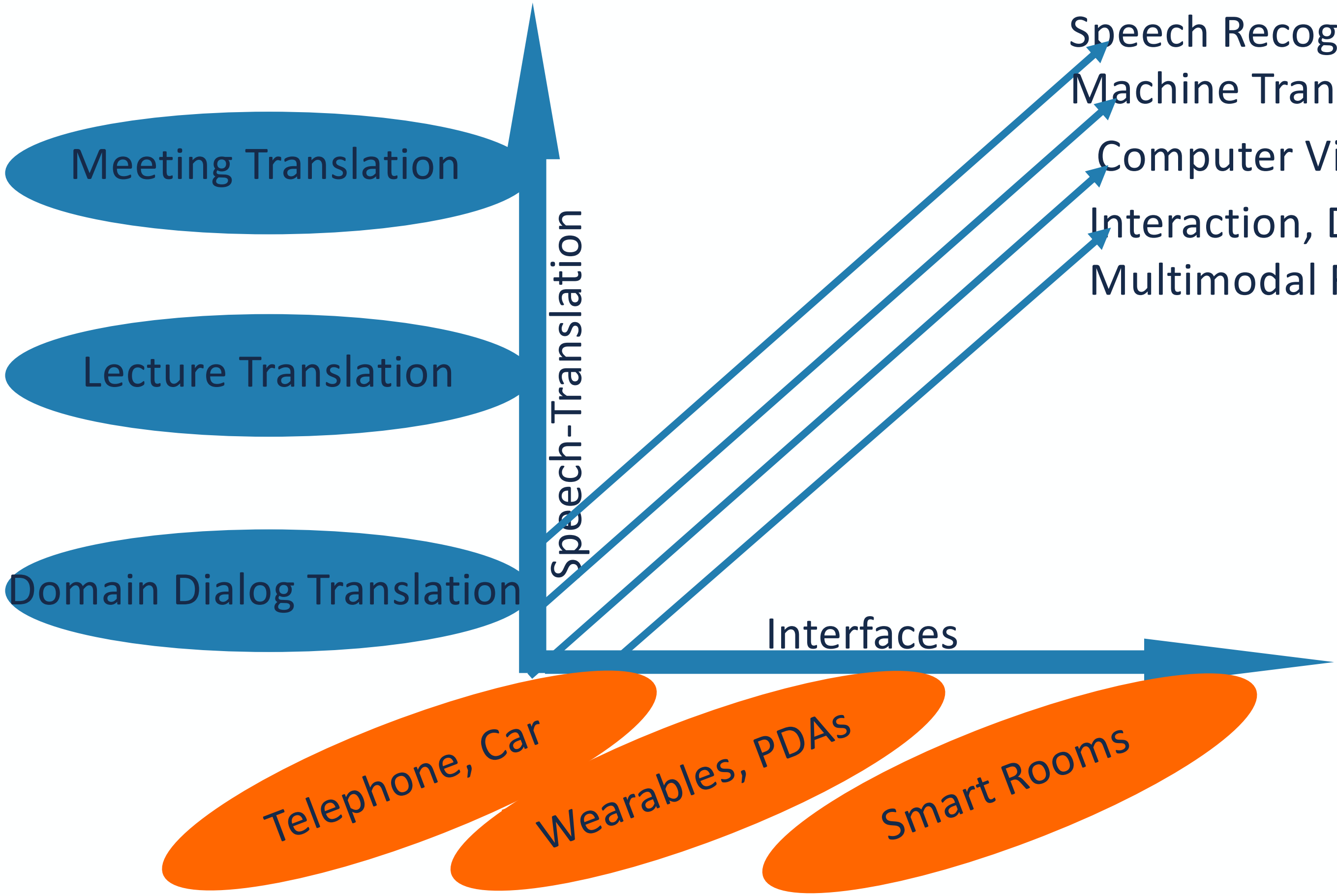
- Main rationale: bring into focus the need for a mutual exchange and use of LRs

- A (not for profit) Association of Users of Language Resources for Research/ Technology Development

- A Repository for Language Resources needed by Language Technologists (Research & Industry)

- Infrastructure for the evaluation of Human Language Technologies

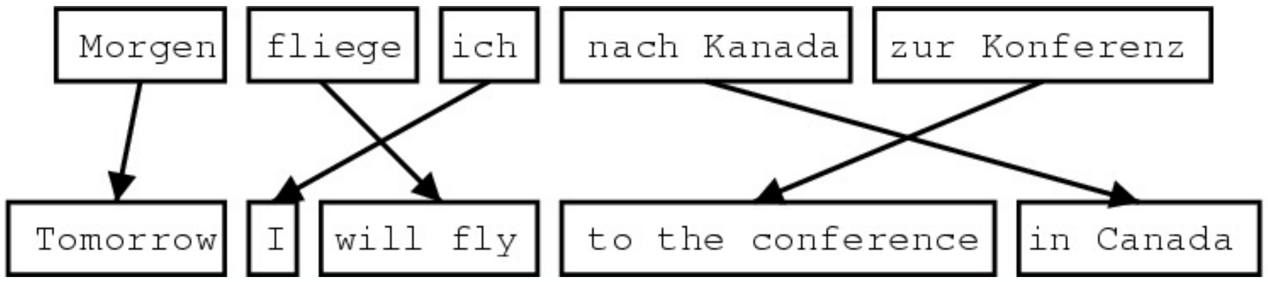
GRAND CHALLENGES

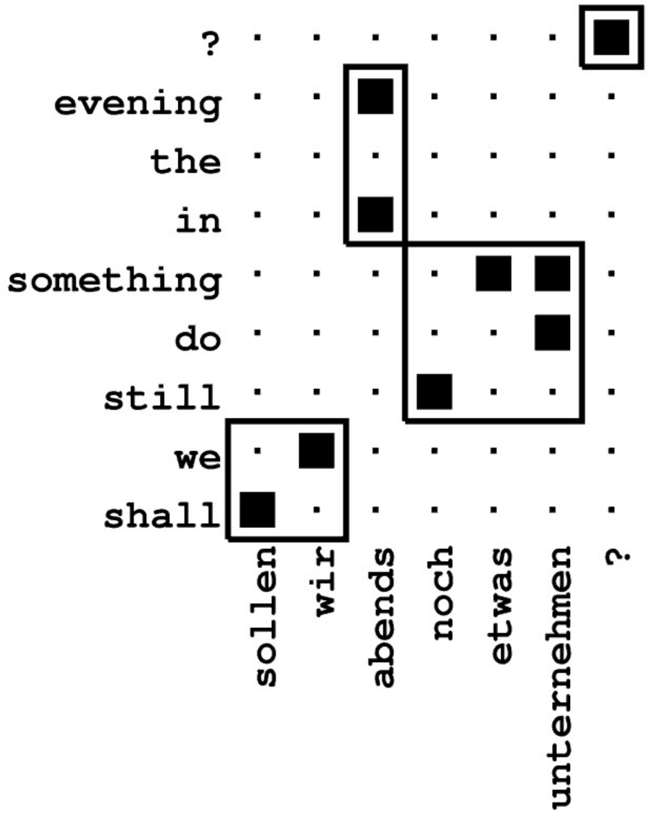


Babel Fish ….

and probably the oddest thing in the Universe

How machines used to translate (Statistics' age) :





MT and Human Parity? 2019 /2020

http://www.statmt.org/wmt19/pdf/53/WMT01.pdf
http://www.statmt.org/wmt20/pdf/2020.wmt-1.1.pdf

WMT 2019 & 2020
Example for News Translation Task

English→German

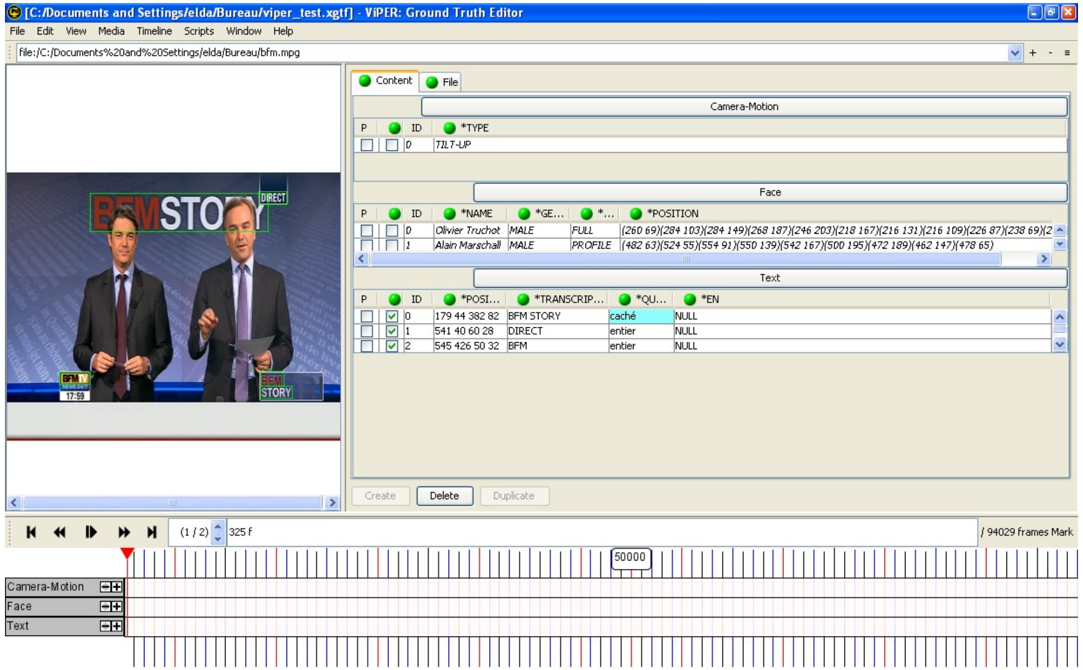
Ave.	Ave. z	System
90.3	0.347	Facebook-FAIR
93.0	0.311	Microsoft-WMT19-sent-doc
92.6	0.296	Microsoft-WMT19-doc-level
90.3	0.240	HUMAN
87.6	0.214	MSRA-MADL
88.7	0.213	UCAM
89.6	0.208	NEU
87.5	0.189	MLLP-UPV
87.5	0.130	eTranslation
86.8	0.119	dfki-nmt
84.2	0.094	online-B
86.6	0.094	Microsoft-WMT19-sent-level
87.3	0.081	JHU
84.4	0.077	Helsinki-NLP
84.2	0.038	online-Y
83.7	0.010	lmu-ctx-tf-single
84.1	0.001	PROMT-NMT
82.8	−0.072	online-A
82.7	−0.119	online-G
80.3	−0.129	UdS-DFKI
82.4	−0.132	TartuNLP-c
76.3	−0.400	online-X
43.3	−1.769	en-de-task

English→German

Ave.	Ave. z	System
90.5	0.569	HUMAN-B
87.4	0.495	OPPO
88.6	0.468	Tohoku-AIP-NTT
85.7	0.446	HUMAN-A
84.5	0.416	Online-B
84.3	0.385	Tencent-Translation
84.6	0.326	VolcTrans
85.3	0.322	Online-A
82.5	0.312	eTranslation
84.2	0.299	HUMAN-paraphrase
82.2	0.260	AFRL
81.0	0.251	UEDIN
79.3	0.247	PROMT-NMT
77.7	0.126	Online-Z
73.9	−0.120	Online-G
68.1	−0.278	zlabs-nlp
65.5	−0.338	WMTBiomedBaseline

Multimodal technologies

- **TV Broadcast**
  - Head localization & identification
  - Embeded text localization & transcription
  - Speech transcription & annotation
  - Machine Translation (Speech2Text/Speech)



European Language Resource Coordination
Connecting Europe Facility

EXAMPLE MT OVER THE YEARS
RULE-BASED AGE

European Commission

➢ Rule Based Machine Translation

- **Direct Systems** (Dictionary Based Machine Translation) map input to output with basic rules.
- **Transfer RBMT Systems** (Transfer Based Machine Translation) employ morphological and syntactical analysis.

- Basically: … Analysis …….. dictionary …. Generation

Source Language

Bilingual Dictionaries

Target Language