ATA German Language Division
11 March 2023, Mainz

MT, AI and the Language Professional

How to fit in to the emerging language services landscape

Jay Marciano

Director, Machine Translation Outreach & Strategy

Lengoo GmbH

jay.marciano@lengoo.com

Introductions

Jay Marciano

- Director of Machine Translation, Lengoo
- 25 years' experience in the development & application of MT
- President of AMTA (Association of Machine Translation in the Americas)
- Avowed language geek on a mission to increase understanding, cooperation, and collaboration among translators and interpreters, MT researchers and developers, language service companies, and customers

lengoo

- Founded 2014
- Headquartered in Berlin
- Language Technology and Services Provider
- ~100 employees

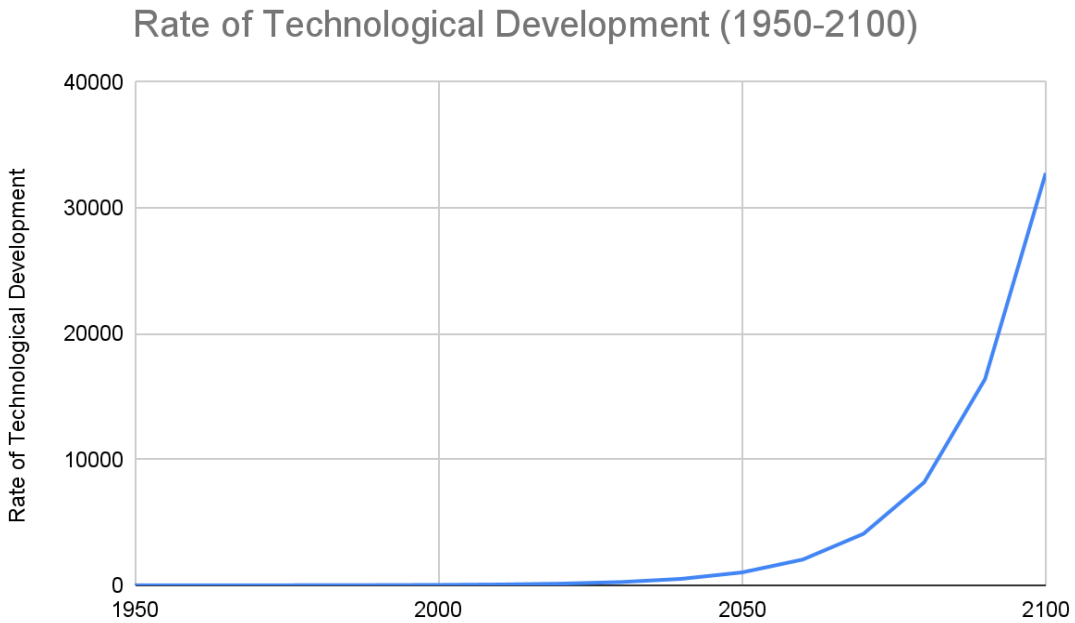
One thing has become very, very clear to us in the last few years …

Technology is advancing faster and faster

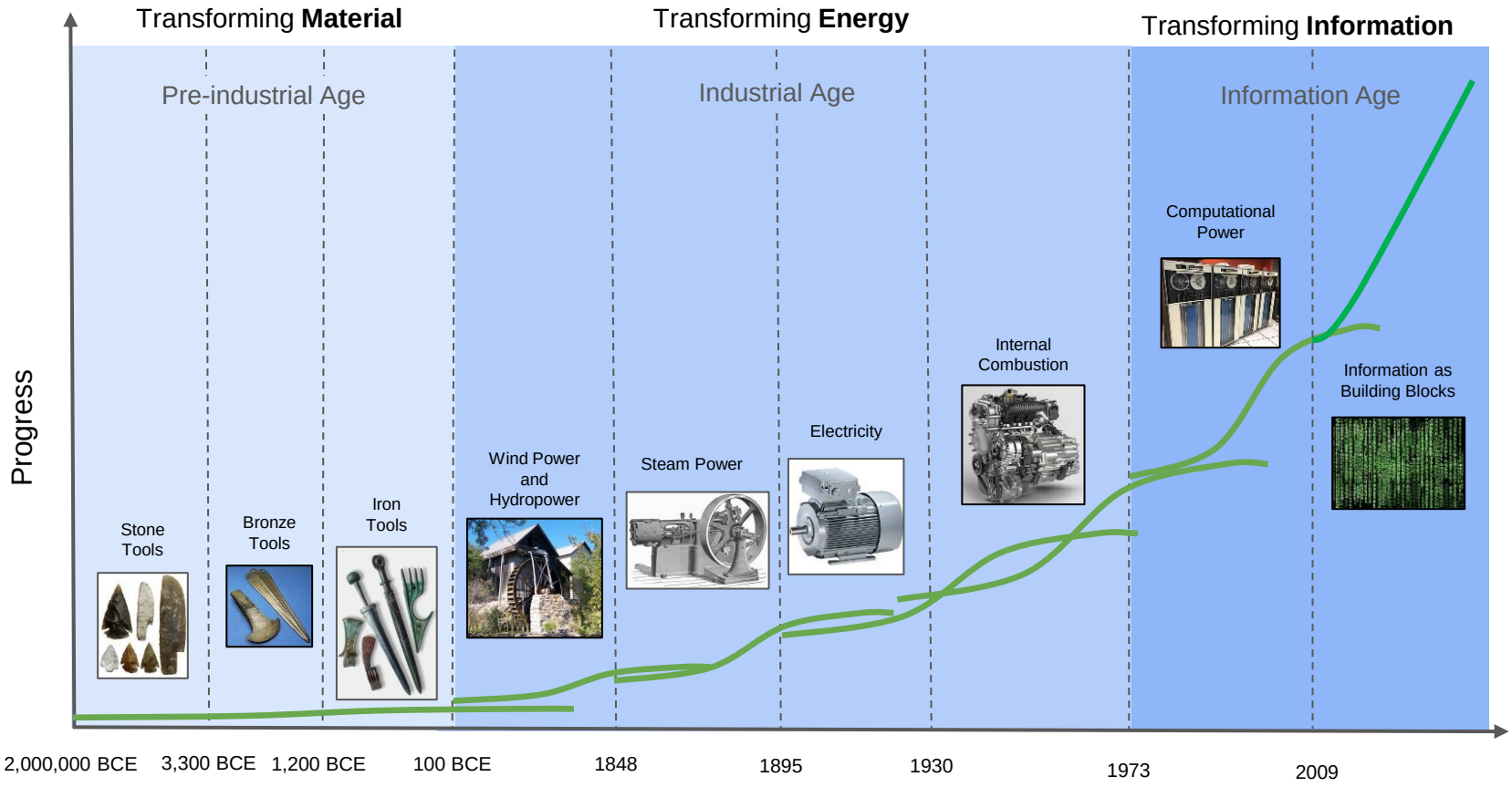
Rate of tech development has since doubled

Kurzweil's Rate of Technological Change

- In 2003, Ray Kurzweil estimated that the rate of technological advance doubles every decade.
- Assuming that's true, it is exponential growth, with the technological advance of the past decade helping to increase the rate of development in the next decade.
- In 2023, we're developing technology twice as fast as we were in 2013, and 128 times faster than in the 1950s, when MT was first demonstrated.
- And from now until 2033, the rate of change will double again.



Drivers of Progress

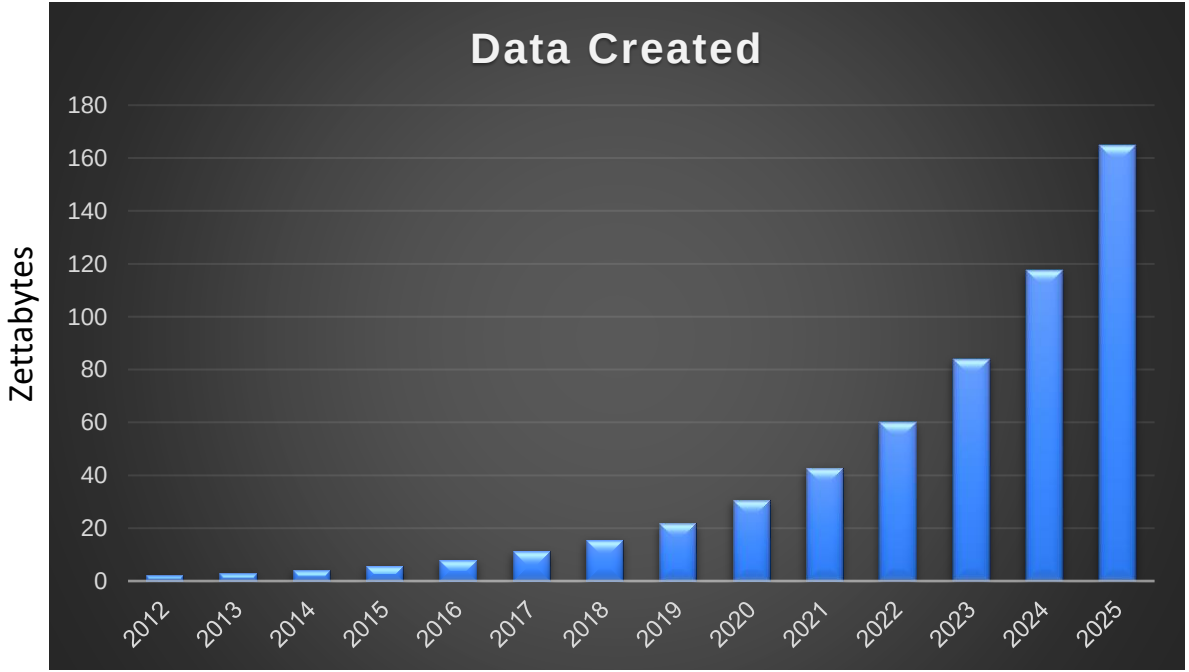


In the Information Age,

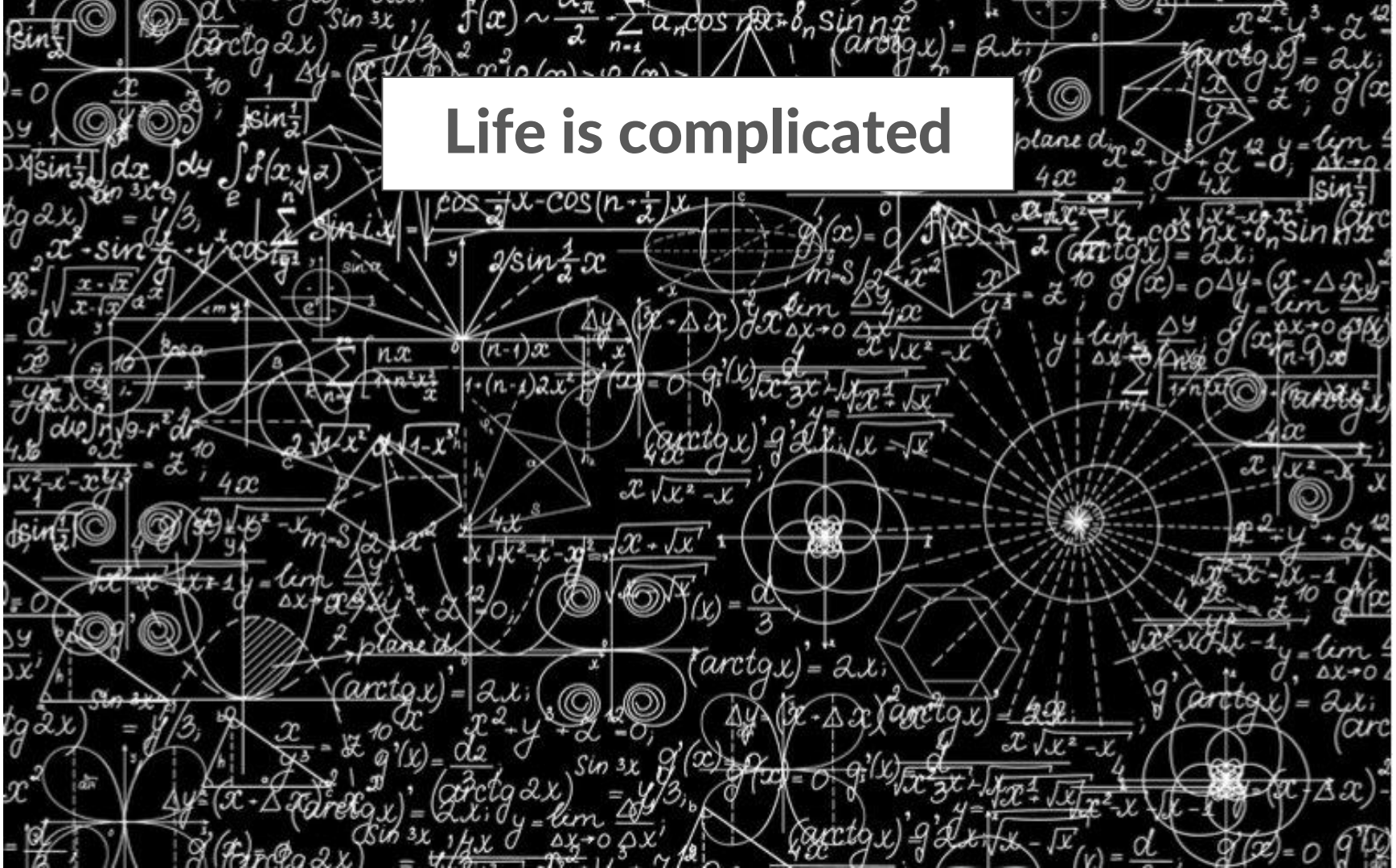
data

is the primary driver of progress.

And it’s an abundant resource!



Let’s think about Complexity


Life is complicated

The intersection of simplicity and endless variability

How many ways can you shuffle a deck of cards?

How many ways can you shuffle a deck of cards?

$$x = 52 \times 51 \times 50 \times 49 \times 48 \ldots \times 3 \times 2 \times 1$$

$$x = \mathbf{52!}$$

How many ways can you shuffle a deck of cards?

Would it surprise you to hear that there are

824 TRILLION

ways to shuffle a deck of cards?

How many ways can you shuffle a deck of cards?

And what if that 824 trillion were only this much of the true number?

80,658,175,170,943,878,571,660,636,856,403,766,975,289,505,440,883,277,**824,000,000,000,000**

You think that's a big number?

Language is unimaginably more complicated!

Characteristic	A Deck of Cards	A Human Language
Number of "words"	52	Let's say 100,000
Possible lengths	1	Let's say 100
Repetition of "words"?	Impossible	Possible
Possible combinations	8.0658175×10^{67}	Incalculably higher

And yet you're able to understand multiple languages.

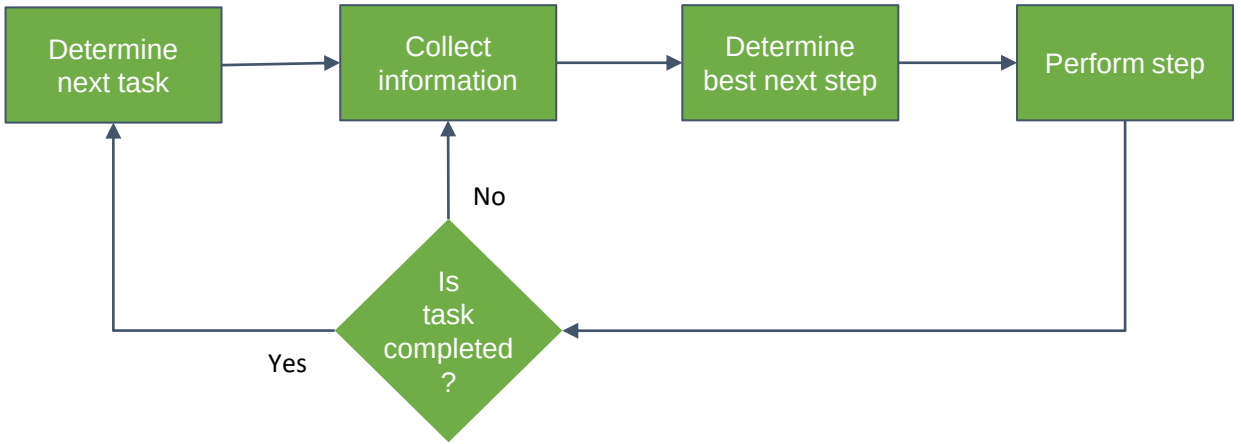
Life is complicated

- You have a million things spinning around in your head.
- But you need to GET STUFF DONE.
- So you calm down and decide to attack a problem.
- You focus on the task at hand and say to yourself, “Take it one step at a time … just figure out what to do next and you’ll get there eventually.”

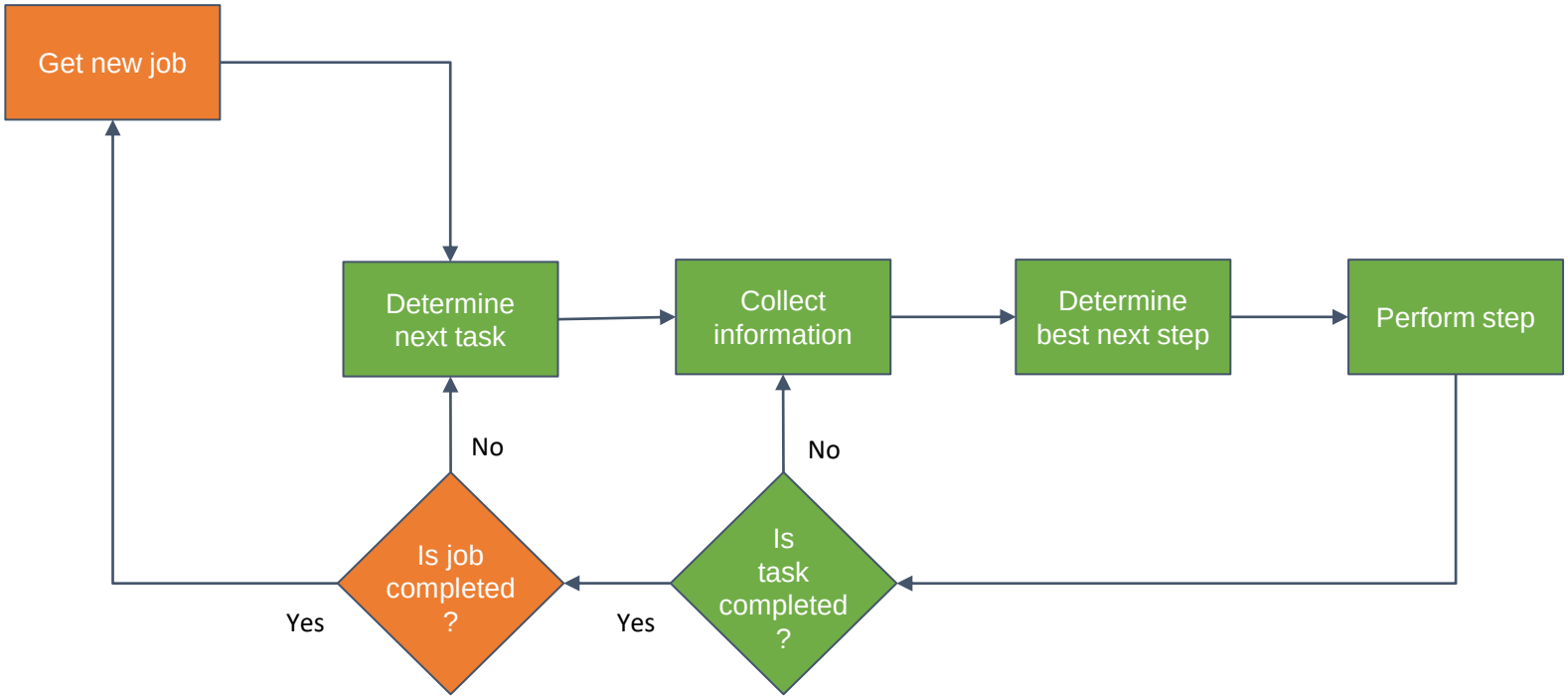
You're already thinking algorithmically

algorithm *n.* a procedure or formula for solving a problem, based on conducting a sequence of specified actions

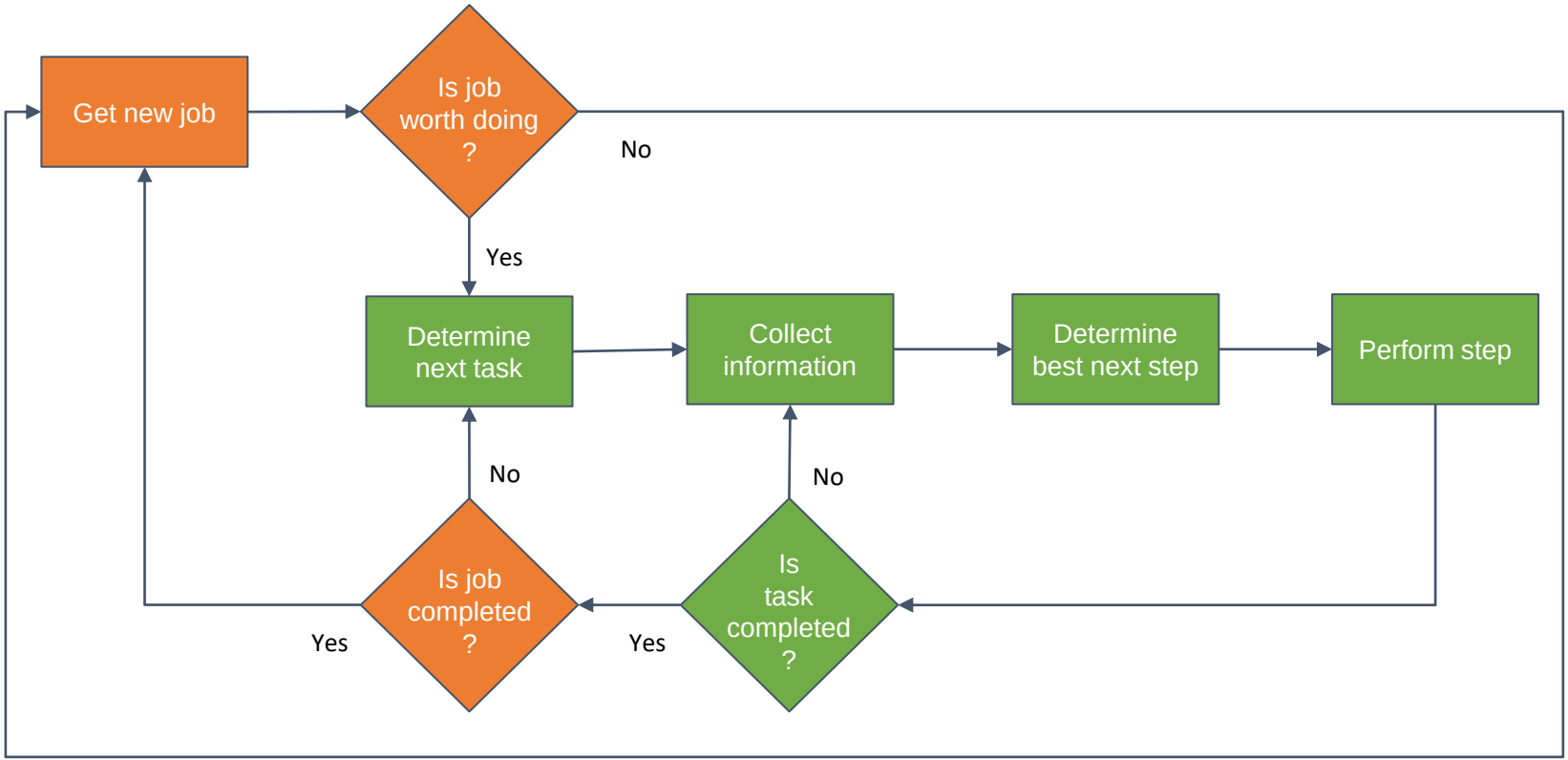
Algorithmic task completion



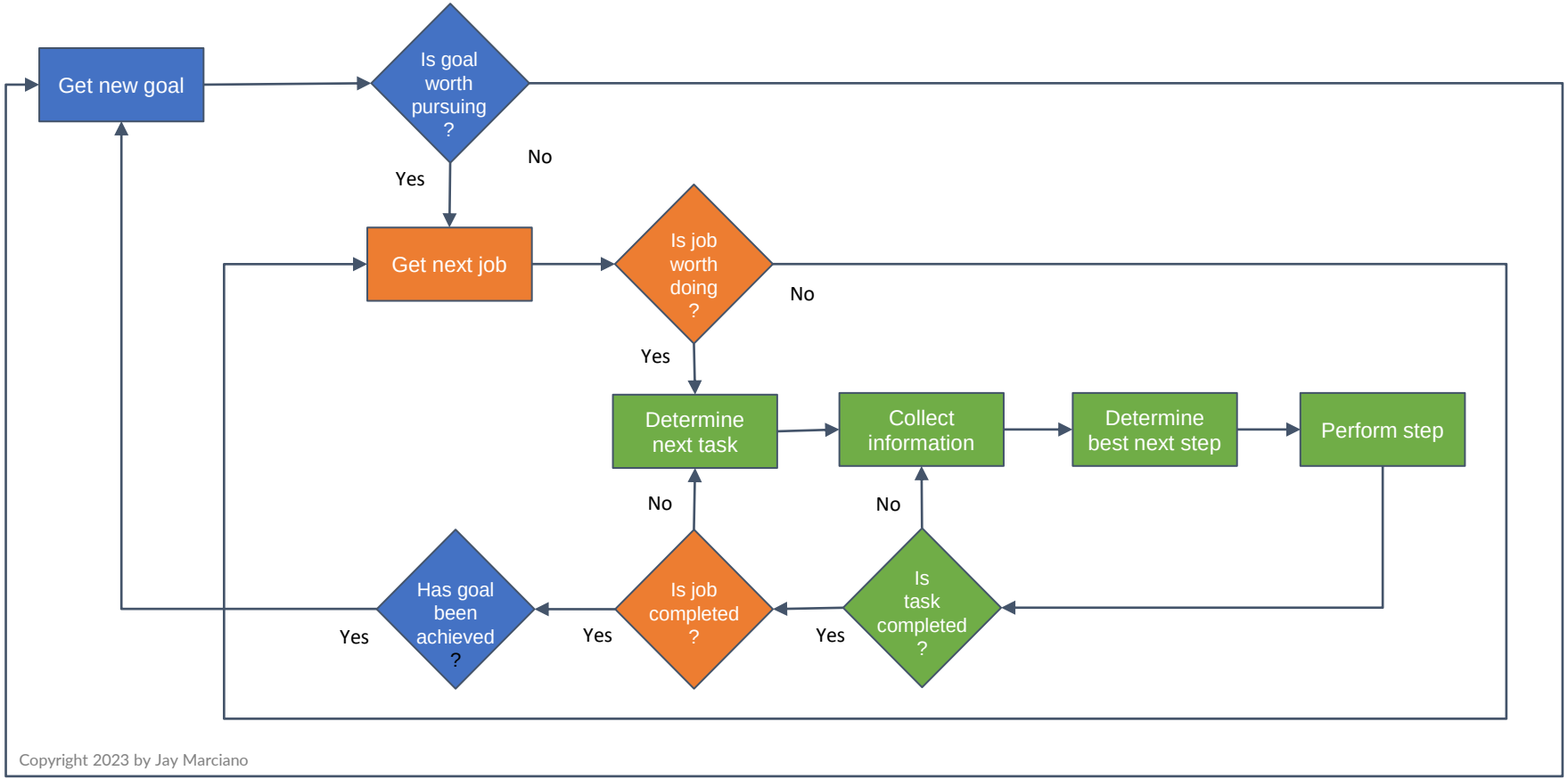
Algorithmic job completion



Algorithmic job selection and completion



Algorithmic goal selection and completion



And that’s what AI and Neural Networks do …

- Learn about a task
- Develop a method for:
  - Determining what the best next step is
  - Executing that step
  - And repeating that process until
    - the task is done,
    - or the job is done,
    - or, maybe someday, an overarching goal is achieved.

Let’s think about Machine Translation

Quickening development of Machine Translation

Rules-based MT

Build a translation with language-specific algorithms and bilingual dictionaries

Was state-of-the-art for ~50 years

Statistical MT

Find the most probable translation using statistical modeling

Was state-of-the-art for ~17 years

Neural MT

Predict a translation using a deep neural network

State-of-the-art for ~6 years

Let’s build an NMT model

Monolingual training material for NMT

More specifically, the presidency currently believes that, should this type of intelligence capability be established within the EU, we would have to investigate the procedures that would need to be implemented in order to ensure democratic monitoring of these activities.

These procedures would naturally have to take into account the particular nature of the intelligence field which, in order to function effectively, requires that information gathered remains confidential.

Mr President, to sum up and in conclusion, I would like to congratulate Mr Watson once again for his excellent work and I hope that you have an interesting debate.

I shall be staying in the House to listen to what you have to say.

Mr President, ladies and gentlemen, recently, when I had to explain the function of our Echelon Committee, I used the following example: I said, 'Imagine you are a detective sergeant.

In short, it is a rather difficult case.

This is tantamount to what happened when journalists began to tell us that there was a global interception system called Echelon which was operated by the United States, the United Kingdom, Canada, Australia and New Zealand and which could intercept and analyse phone, fax and e-mail communications worldwide.

Individuals, businesses and institutions, they said, were subjected to systematic

福建厦门即日起到今年年底将采取临时性限购措施，规定每个购房家庭只能在本市新购买一套商品住房。当地主管部门表示，之后将视市场情况对限购政策进行调整。

10月1号，广州市国土房管局对当地在售楼盘进行了检查，要求各开发商在购房须知上对新一轮房地产调控政策有所体现。目前他们正密切关注楼市动态，一旦发现房价出现大幅反弹，不排除采取限购等措施，保持市场稳定。

深圳日前出台了房产限购令，规定拥有深圳市户籍的居民家庭，限购两套房，持有一年以上纳税证明或者社会保险缴纳证明的非户籍居民家庭，限购一套房。深圳银监局表示，一旦发现有银行违反房贷政策及有关监管规定，将严肃查处。

各地楼市观望气氛浓厚

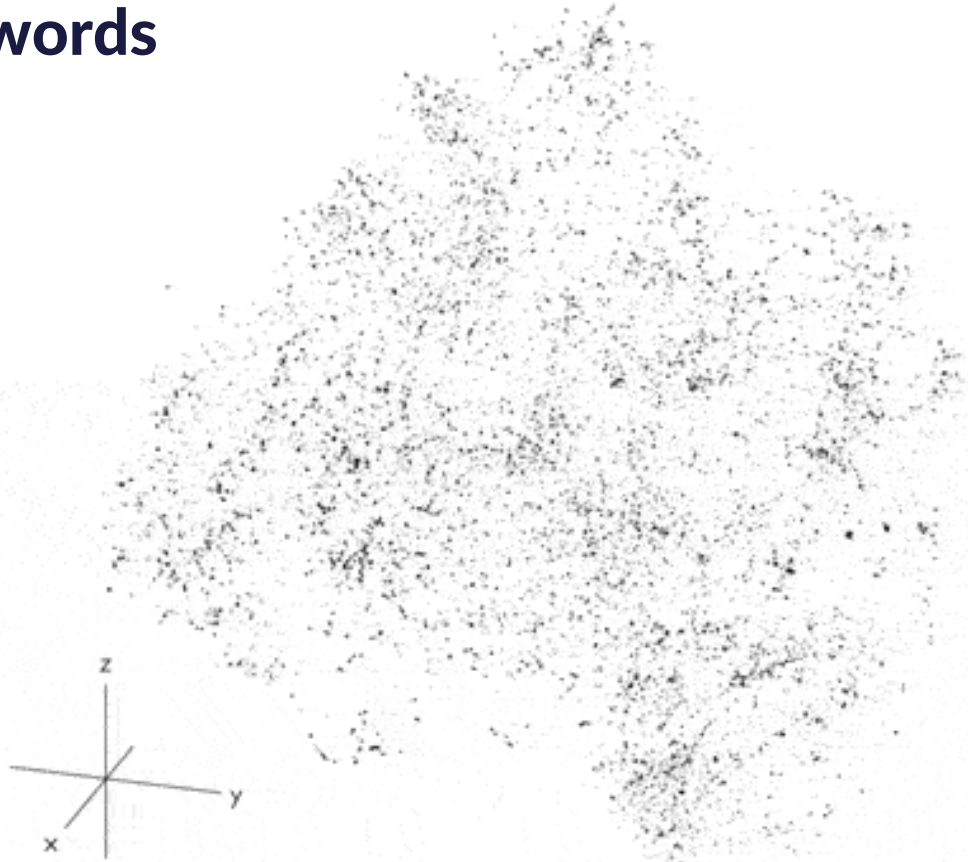
热热闹闹的北京楼市在“黄金周”突然遇冷。根据北京市房地产交易管理网的数据显示，9月30日，因为“末班车”效应，北京二手房签约量出现井喷，达到1067套，也是新政后首次突破千套大关。但到了10月1日和10月2日，交易量就迅速跌至谷底，两天共成交了18套。

在上海，被称为楼市“晴雨表”的“假日楼市?秋季房展”3日上午开幕。尽管这被认为是近3年来最热闹的一次房展会，尽管部分楼盘优惠幅度巨大，“直降10万”，但几天前出台的楼市二次调控政策

What is “learned” about words

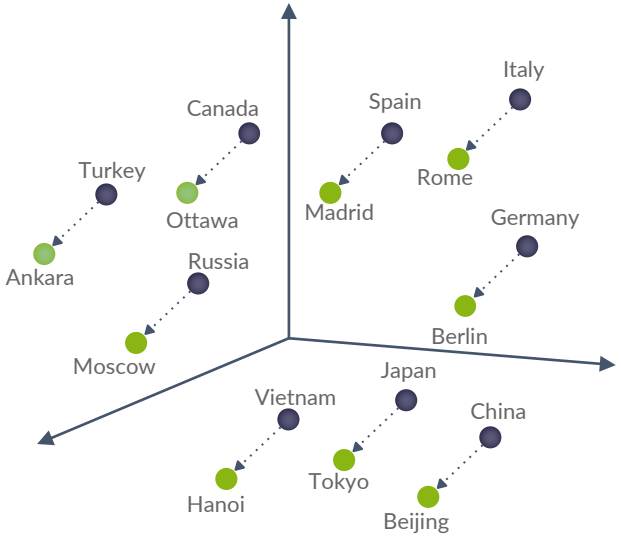
For every word in the training material, the Deep Learning system calculates a **word embedding**, a vector that contains semantic and grammatical information and indicates relationships among the words.

This information provides a multidimensional map of each supported language, showing the relationships among all of the words in those languages.



Word embeddings

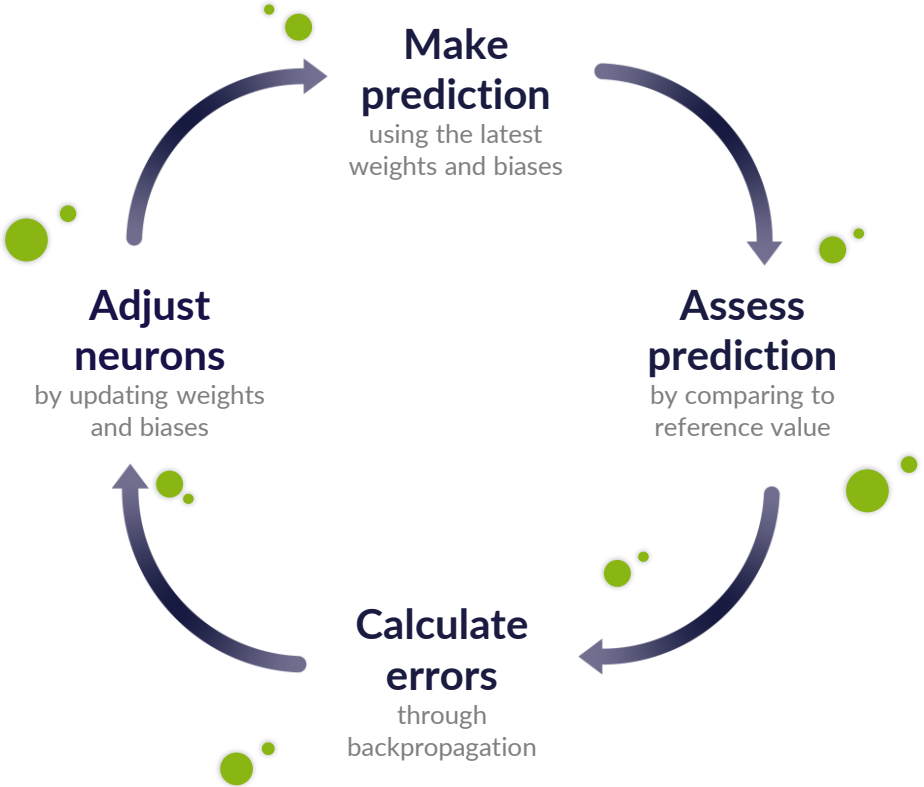
- Word embeddings can be mapped in multidimensional space
- Similar words have similar values (or locations)
- The mathematical relationship between words that have a related meaning resembles the relationship between two other words that share that semantic relationship
    - Example: Countries and their capital cities



Bilingual training material for NMT

怡景宾馆位于香港最繁华核心的商业购物区-**旺角，交通非常便利，集吃喝玩**乐于一体。	Yee King Hotel is located at the heart of Mong Kok, Hong Kong's busiest commercial shopping district. Very convenient public transportation nearby for visiting, dining, and entertainment.
附近金银珠宝店林立，名牌服饰，化妆品商铺琳琅满目。	Close to a wide range of jewelry stores, designer clothing shops, and cosmetics shops.
毗邻朗豪坊、旺角电器街、女人街、波鞋街等，周边还有各种人气超火的小吃店、餐厅，**莎莎、卓悦、屈臣氏、万宁、周生生、周大福、豊**泽、百老匯。	Adjacent to Langham Place, Mongkok electronics district, Ladies Market, Sneaker Street, and a variety of popular eateries, restaurants, as well as Sasha, Bonjour, the Watsons, Mannings, Chow Sang Sang, Chow Tai Fook, Feng Chak, and Broadway.
为您赴港游提供周到的服务，营造美好愉悦的度假心情。	Comprehensive traveling services for your most wonderful vacation in Hong Kong.
怡景宾馆位于香港最繁华核心的商业购物区-**旺角，交通非常便利，附近就有旺角地**铁站**、火**车站、A21 **机**场巴士站、中港直通车站。	Yee King Hotel is located in Mong Kok at the heart of Hong Kong's busiest commercial shopping district. Very convenient public transportation, close to Mongkok MTR station, railway station, A21 airport bus station, and Sino-Hong Kong express bus station.
环中线五和地铁 **温馨** mini 阁楼	Cozy mini loft close to Wuhe Station at Metro line no.5
看图，实拍。	Please refer to the photos for the room interior.
本人是一名软装设计师,**兼**职微商,**做**护肤品因为本身是个非常讲究的人,对生活品质有自**己的要求,平**时最喜欢的就是布置房间，希望入住的朋友能跟着我一样喜欢"**她**们"爱护"**她**们"	I am an interior designer and sell skin care product through micronet as my part-time job. I'm very particular in the quality of living and I love to decorate my living space. I hope you'll appreciate and cherish my place like I do.
亲爱的房客们请务必认真看介绍:**以免入住期**间引起不必要的误会，房型为小阁楼,**在城中村里面，由民房改造,楼下**为客厅,**楼上卧室,屋内布置温馨,厨房和洗手**间为后建的，不是实墙，免费 wifi,**免**费提供洗漱用品,24 **小**时热水,**有茶具提供,厨房有所有做**饭用的厨具。	Dear tenant, to avoid unnecessary misunderstanding, please read through the introduction carefully. The room is a small loft located the Urban Village, renovated from a family house. The living room is downstairs and the bedroom is upstairs. Kitchen and bathroom are built later without concrete wall. Free WiFi, complimentary

Training an artificial neural network



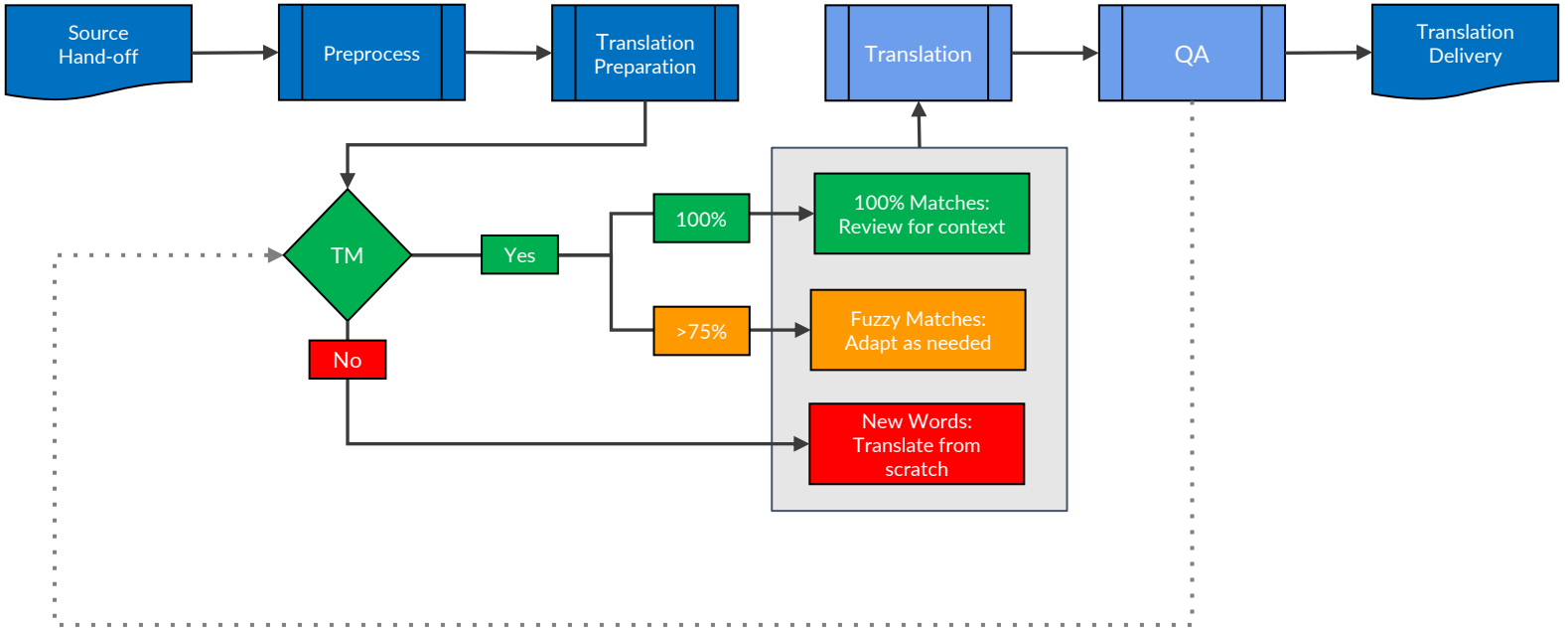
Mastermind

Try it: https://www.webgamesonline.com/mastermind/

Let’s focus on using MT

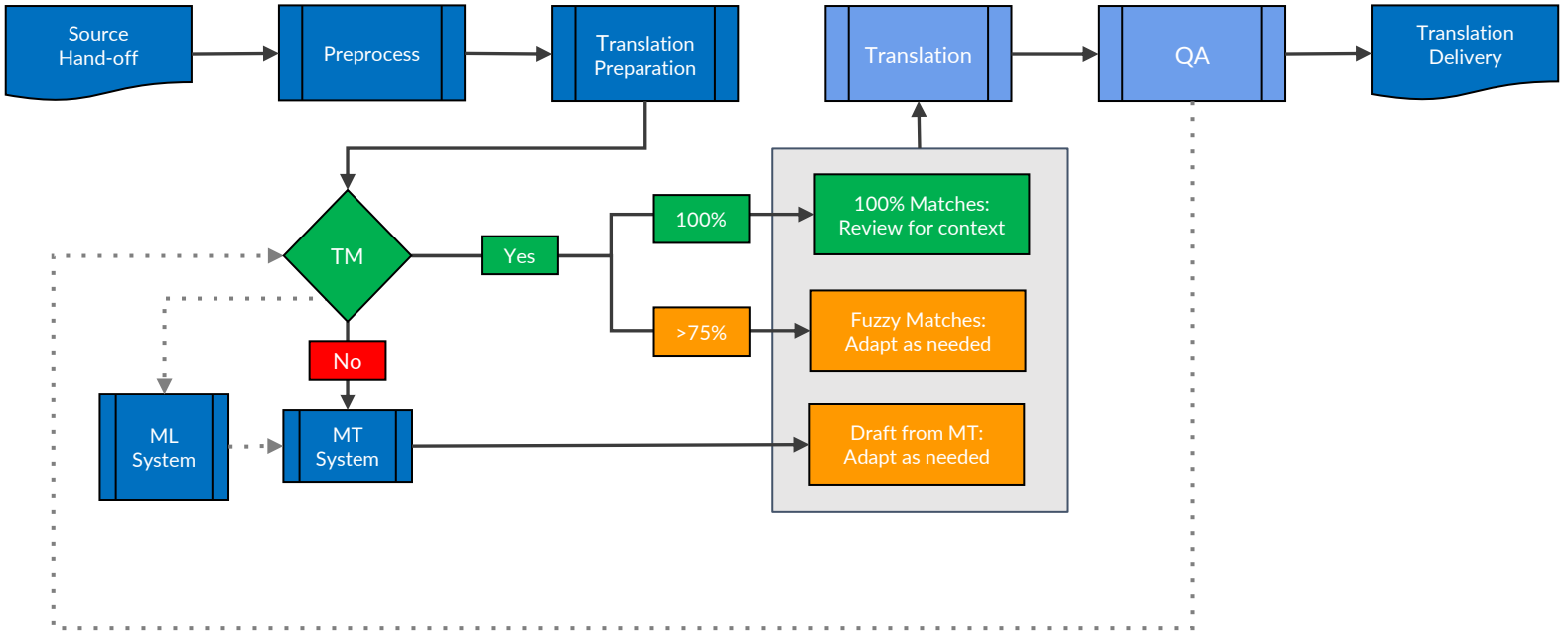
How can Machine Translation be applied to a project?

Part 1: Simplified view of a workflow with TM but ***without*** MT

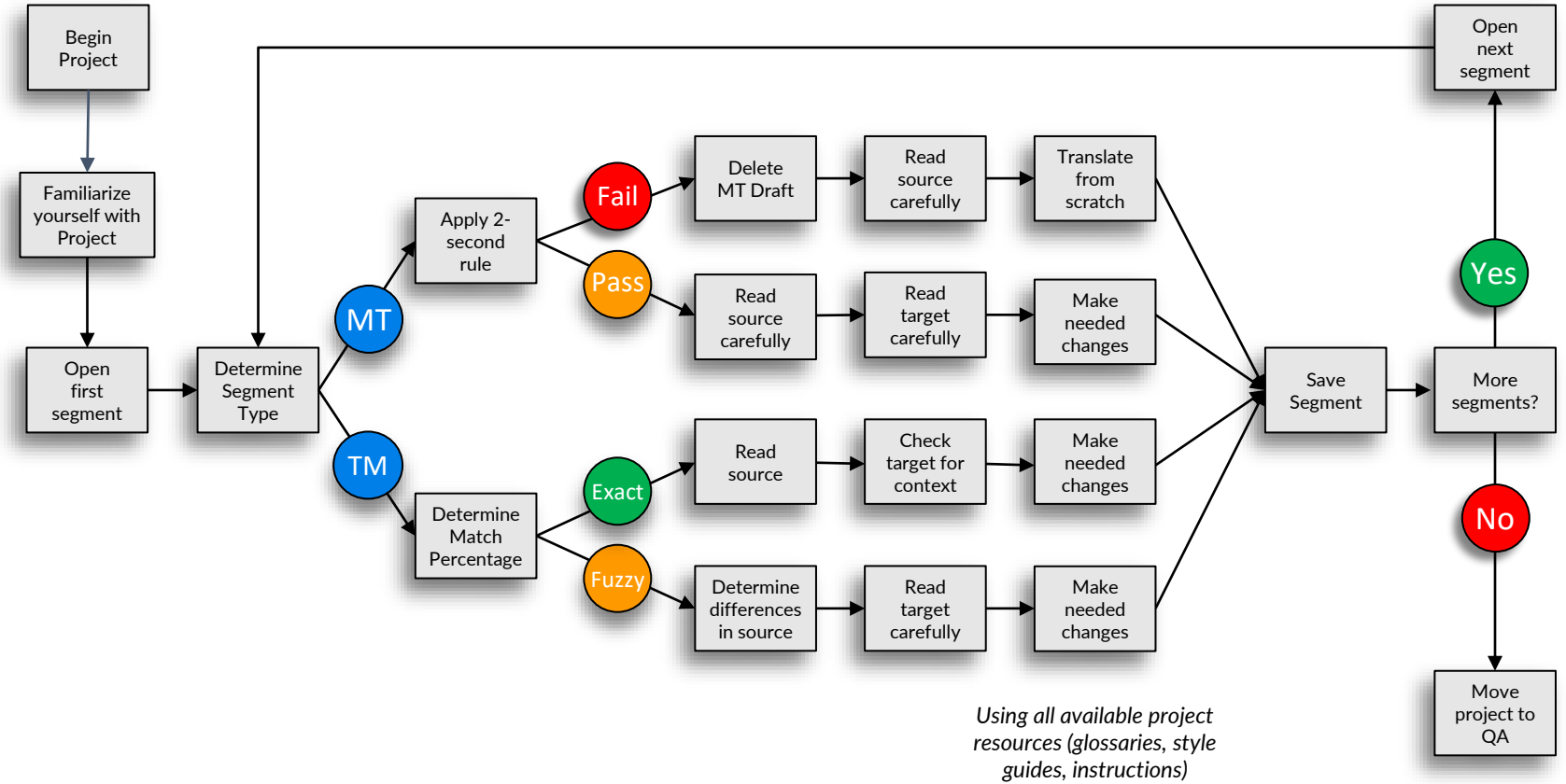


How can Machine Translation be applied to a project?

Part 2: Simplified view of a workflow ***with*** MT and post-editing



Translation Tasks with TM and MT (old-school)

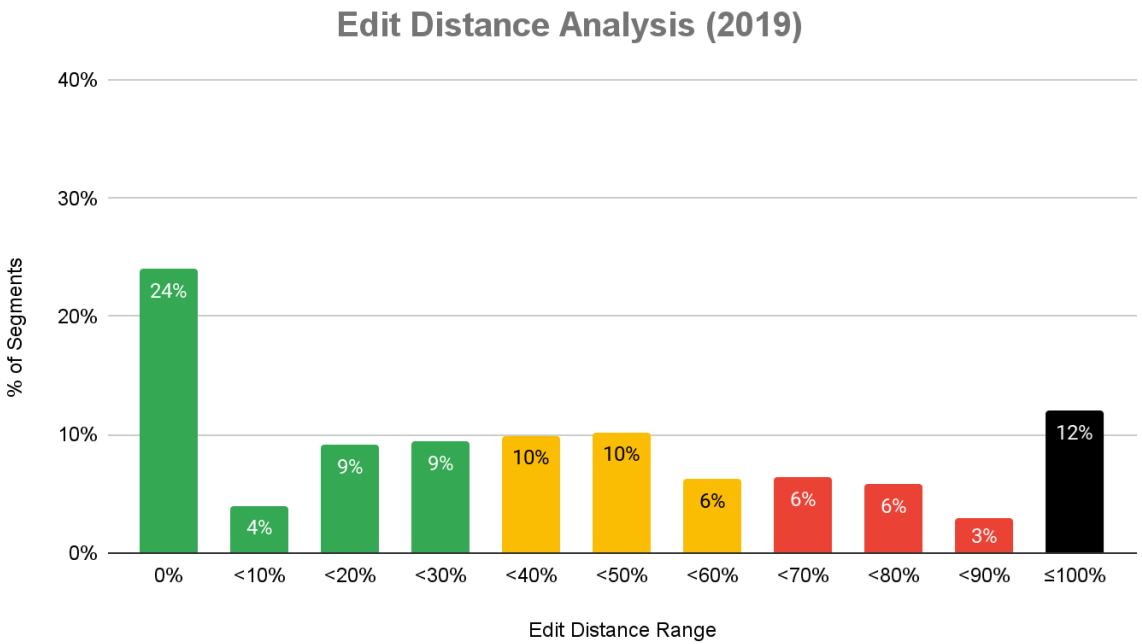


Using all available project resources (glossaries, style guides, instructions)

MT Quality: Where were we four years ago?

Edit-distance analysis at large LSP (2019)

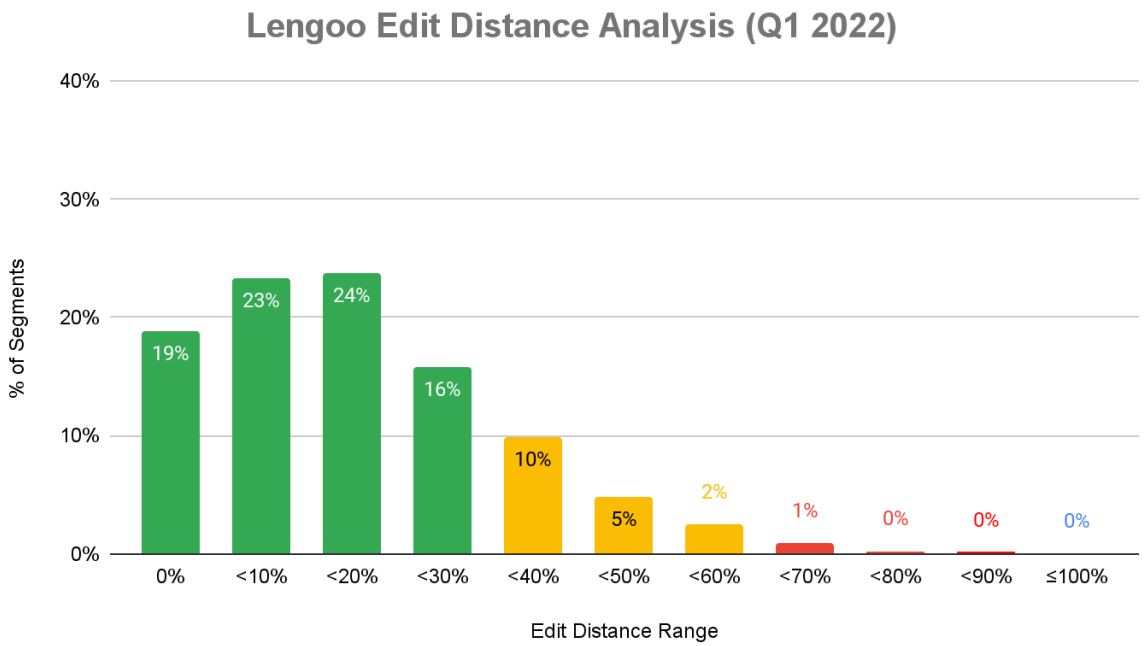
- Google (AutoML), Microsoft Custom Translator, DeepL
- All customers
- All languages
- All content



MT Quality: Where are we now?

Edit-distance analysis at Lengoo (Q1 2022)

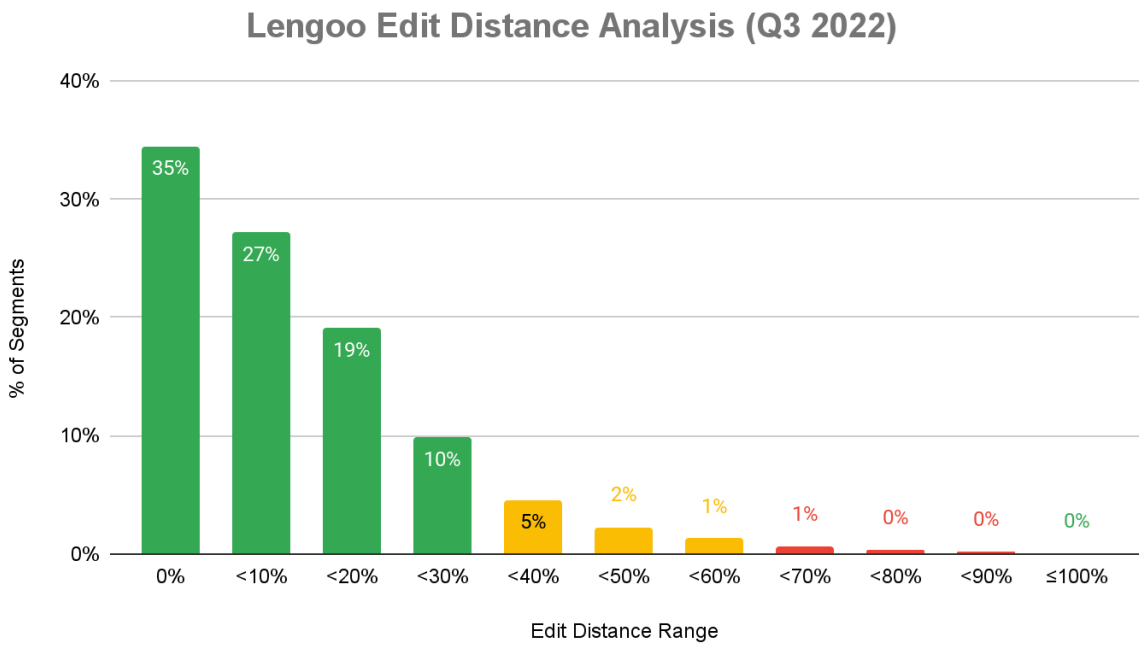
- Lengoo NMT
- All customers
- All languages
- All content



MT Quality: Where are we now?

Edit-distance analysis at Lengoo (Q3 2022)

- Lengoo NMT
- All customers
- All languages
- All content



Conditions that you need to adjust to now

1. MT is still improving
2. Reliable automatic quality estimation will revolutionize translation services

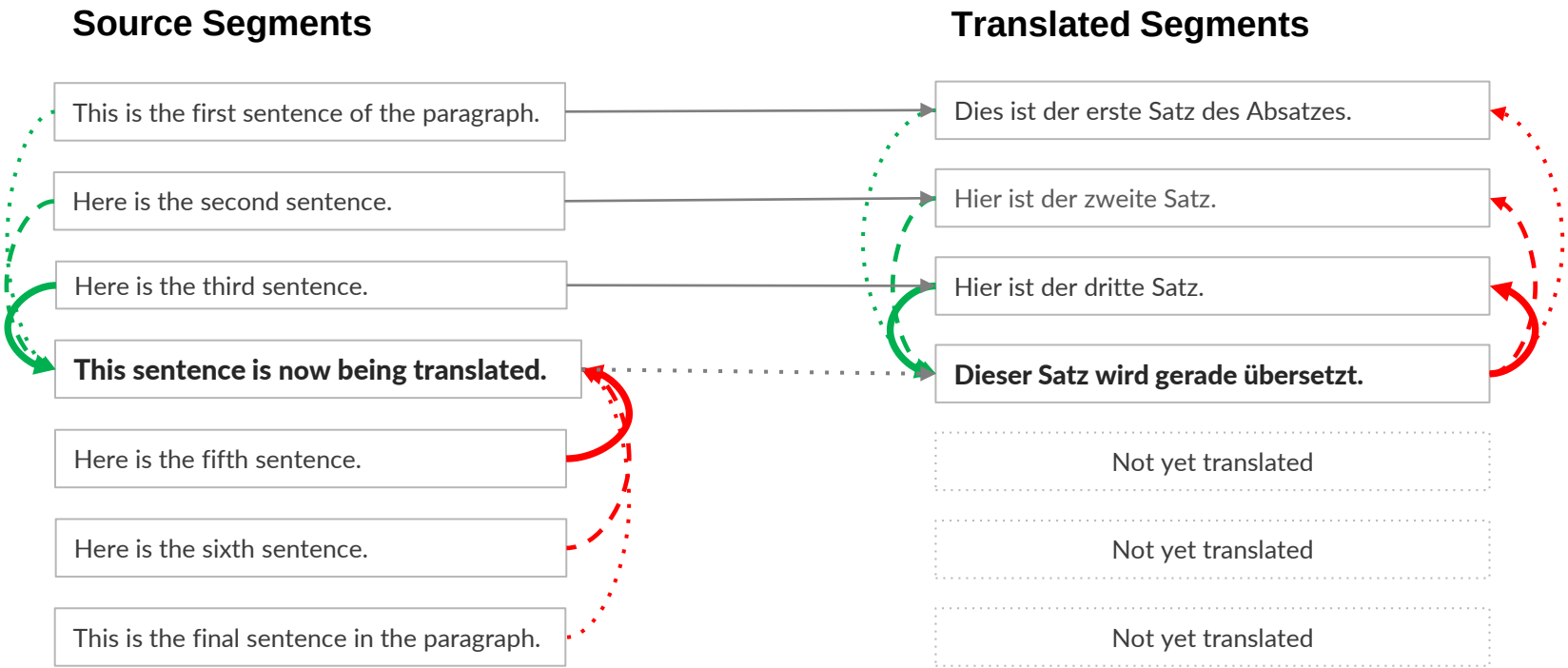
1. Improving MT: Context sensitivity

Information throughout a document informs the translation of every sentence

- MT has traditionally worked on a sentence-by-sentence basis
- Potentially valuable information from preceding or following sentences is not currently leveraged
- This is changing and will help in
    - Resolving pronoun antecedents from previous sentences
    - Reducing inconsistencies
    - Assist in domain resolution for vocabulary choice

1. Improving MT: Context sensitivity

Information throughout a document informs the translation of every sentence



2. Reliable automatic quality estimation

Reducing the “cognitive load” of post-editing on a high percentage of segments

A neural network trained on LQA data, including:

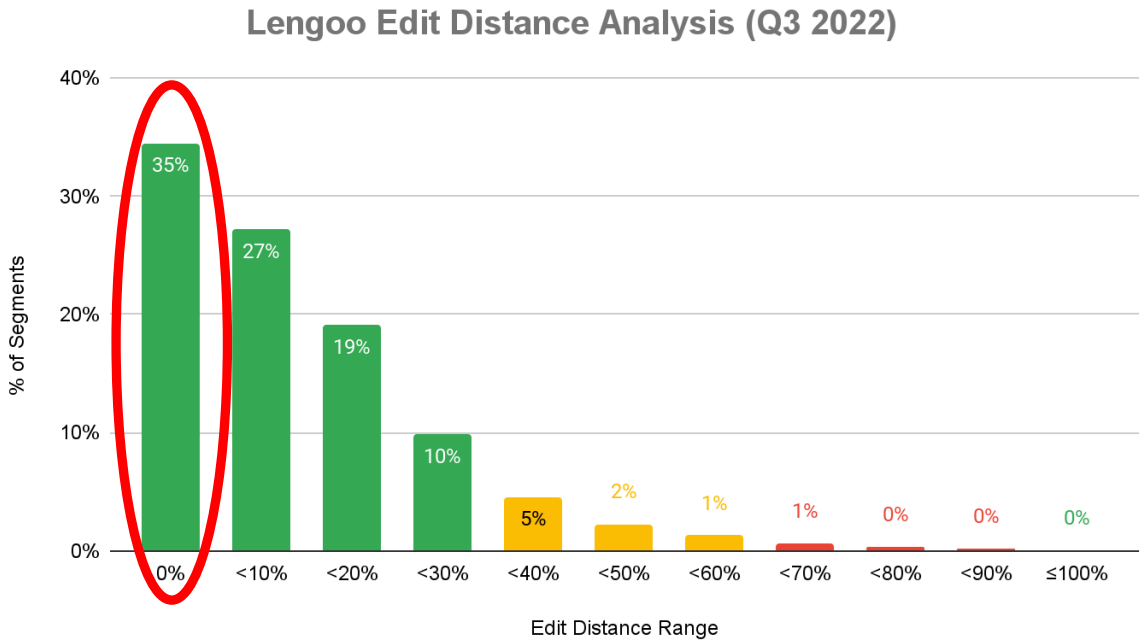
- Source segments
- Unedited MT output
- Post-edited MT output
- Edit distance and other automated metrics
- Language Quality Assessments where available

That predicts:

- whether the machine translation of a new sentence will be in the group of segments that require no change by the translator

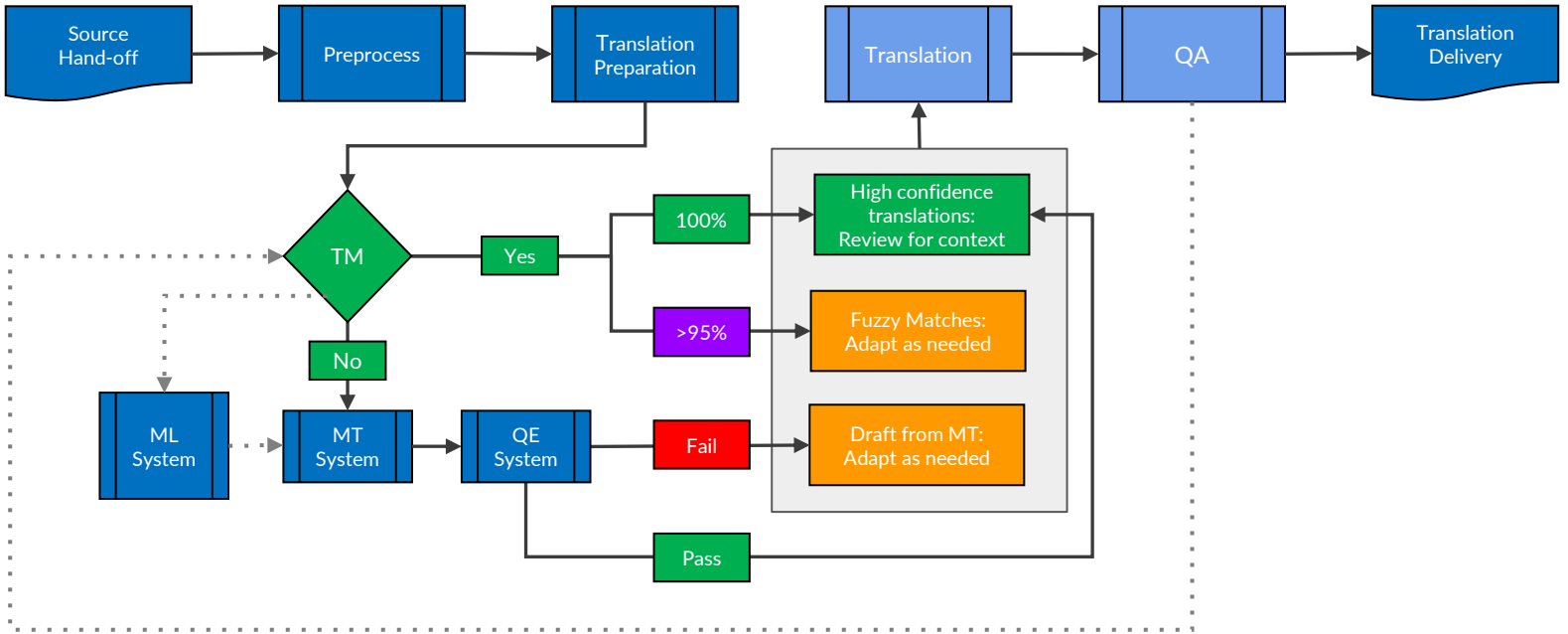
2. Reliable automatic quality estimation

When systems can confidently predict which sentences will fall into the 0% Edit Distance Range, the translation market will change radically.

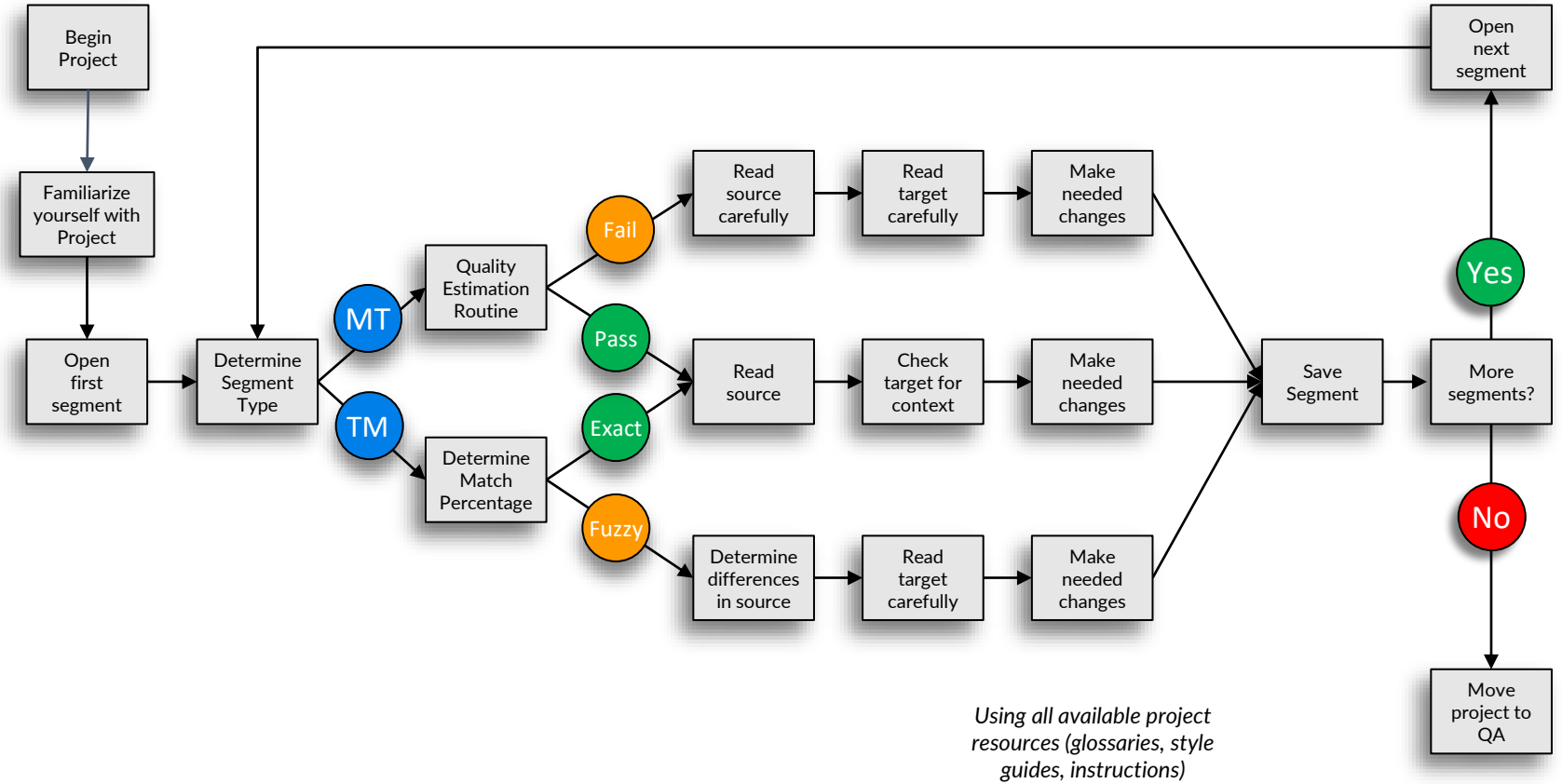


2. Reliable automatic quality estimation

Part 2: Simplified view of a workflow with MT, QE, and post-editing



Translation Tasks with TM, MT and QE



Using all available project resources (glossaries, style guides, instructions)

Now let’s move beyond MT

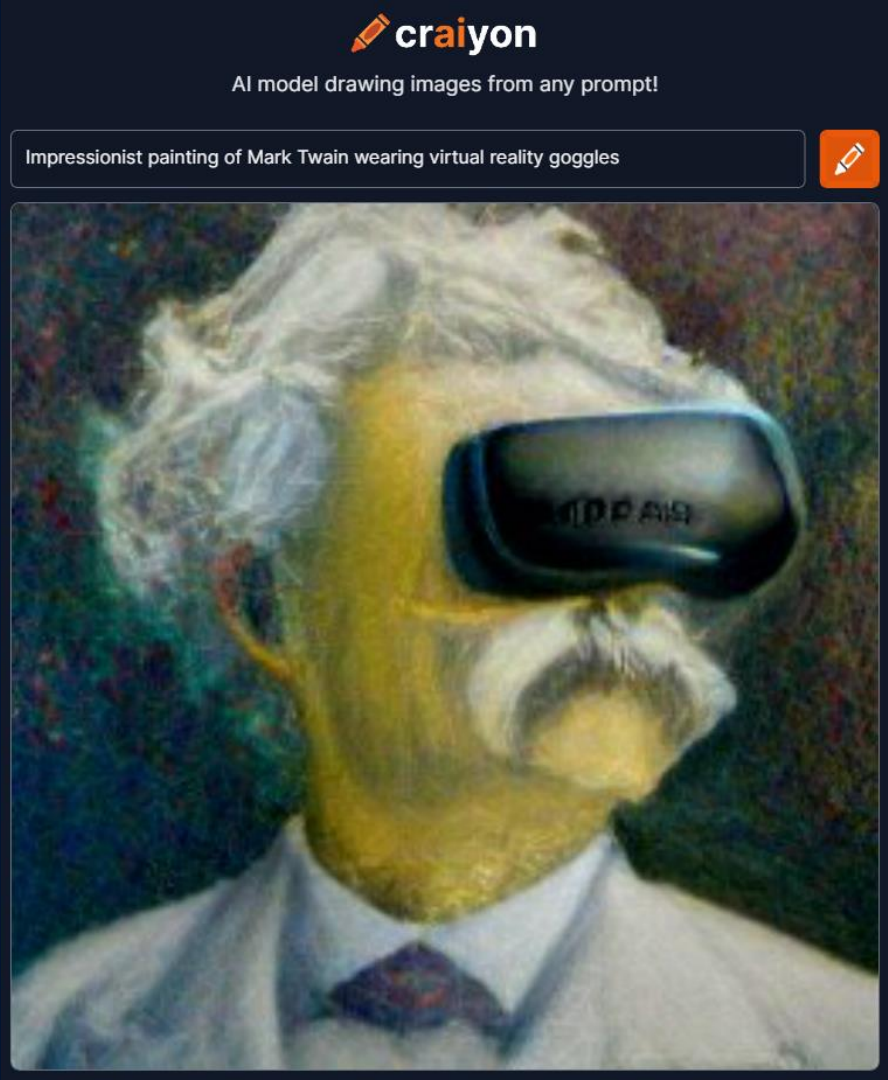
Conditions that you need to adjust to

1. MT is still improving
2. Reliable automatic quality estimation will revolutionize translation services
3. Multimodal “Foundation Models” will provide the next MT platform
4. MT will become just one part of a much larger AI offering
5. Translation data and metadata will be considered a critical asset, giving today’s tech-savvy language professionals a very bright future

3. Multimodal Foundation Models

The next platform for MT … and so much more

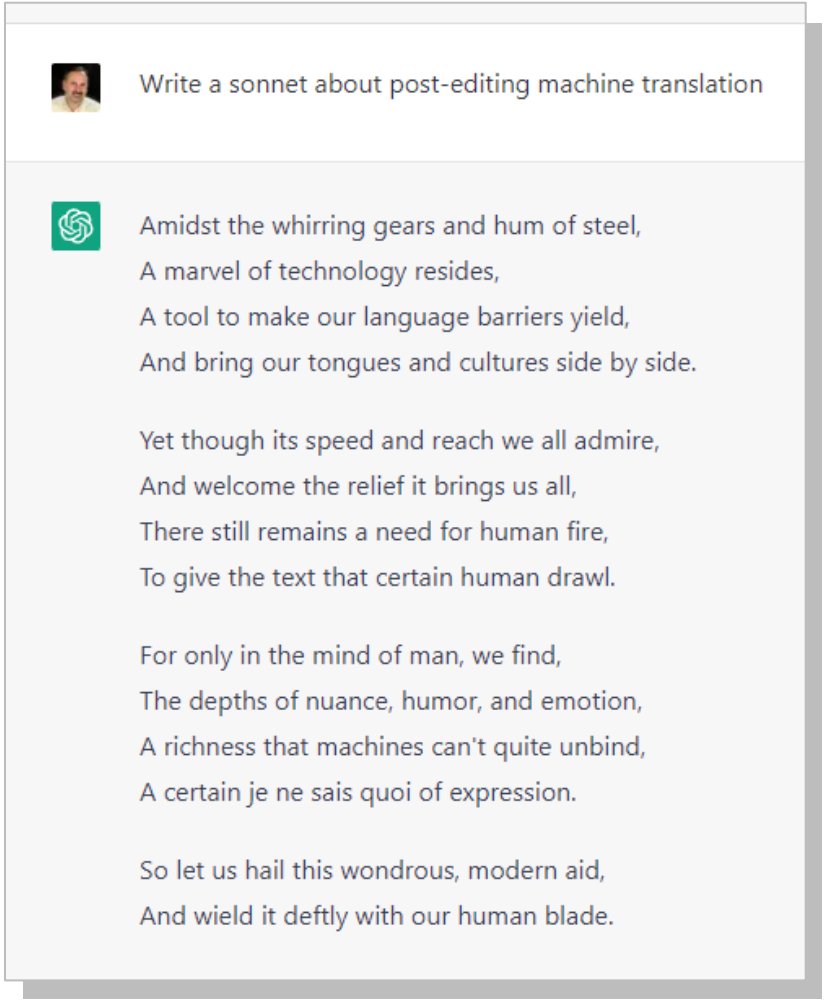
- A paradigm shift in AI
- Massive pre-trained models provide the foundation for task-specific models
- Initial training is unsupervised and can include all kinds of data but does not include a specific task
- Can be “trained” for a downstream task by simply providing it with a natural language description of the task
- Spooky: this ability to do tasks based on verbal prompts was neither specifically trained for nor anticipated to arise.



ChatGPT

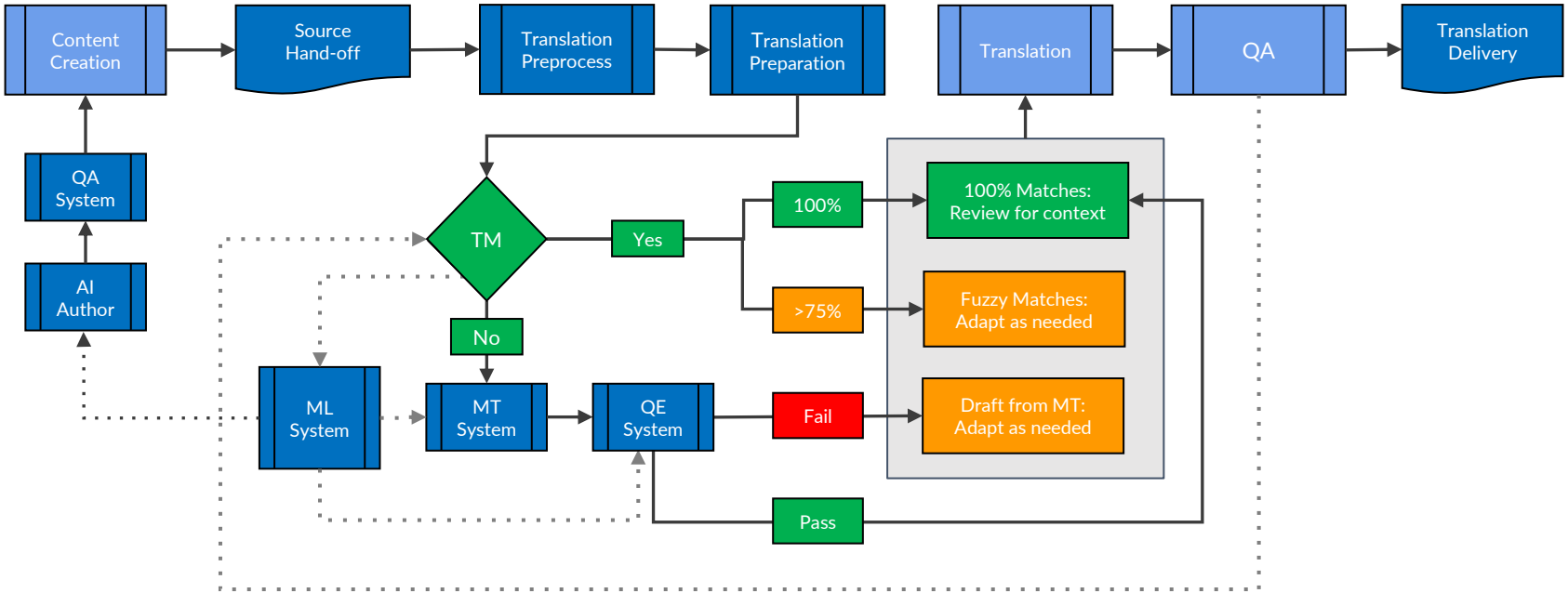
- Developed by OpenAI
- Released in November 2022
- Based on GPT-3
  - Transformer Neural Network
  - “**G**enerative **P**re-trained **T**ransformer”
  - Trained with massive
  - LLM with 175 billion “parameters”
- Generates text in response to prompts

PS: GPT-4 is being released next week.



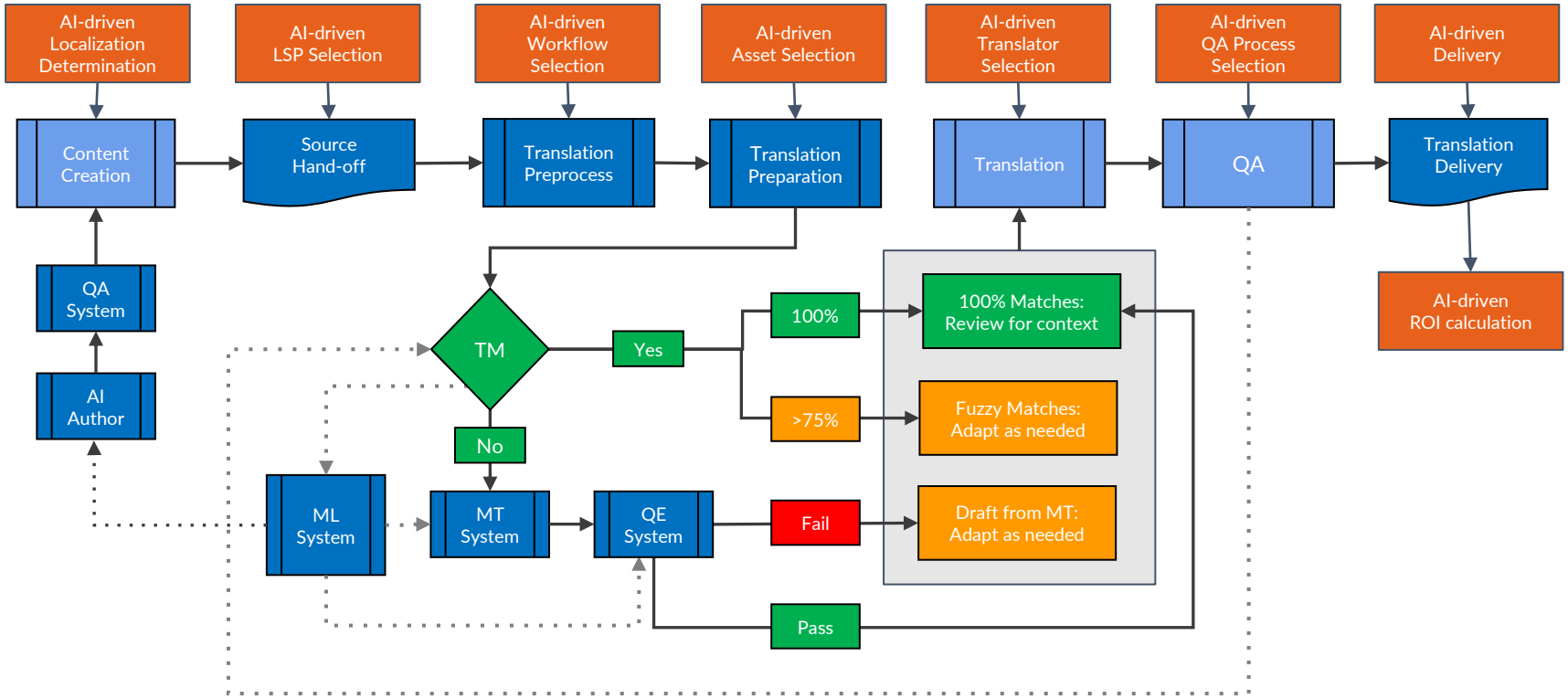
How will foundation models be applied to projects?

Part 4: Simplified view of a workflow ***with*** AI-assisted content creation, MT, QE, and post-editing



How will foundation models be applied to projects?

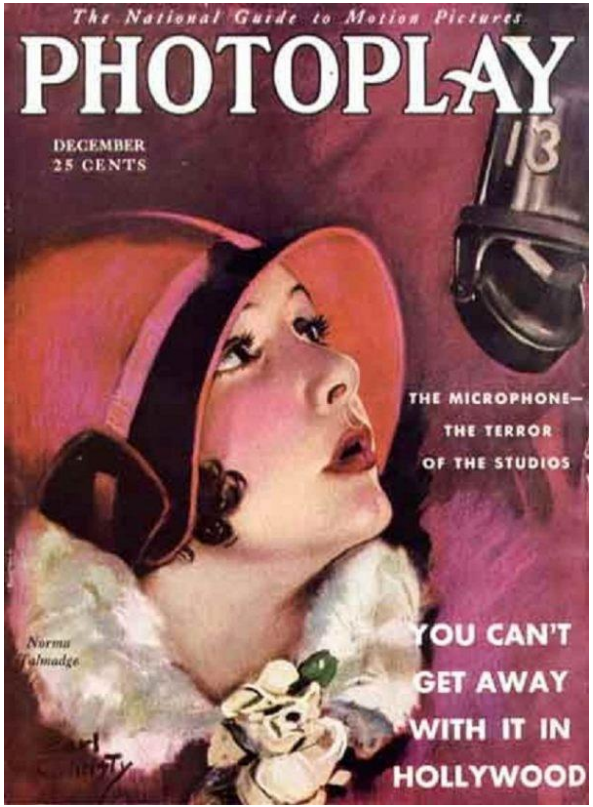
Part 5: Simplified view of a workflow ***with all kinds of AI-driven functionality***



A 100-year-old lesson about technology-driven change

- At the first Academy Awards (1929), all of the nominees for Best Picture were silent movies

A 100-year-old lesson about technology-driven change



- At the first Academy Awards (1929), all of the nominees for Best Picture were silent movies
- At the second Academy Awards (1930) none of them were
- Actors unable to transition to the world of talkies lost work, including Charlie Chapin, Rudolf Valentino, Mary Pickford, Lillian Gish, Clara Bow
- But many, many new jobs were created.

What kind of jobs will be created by AI in our industry?

- Data Curator
- Data Scientist
- Terminologist
- Corpus Linguist
- Computational Linguist
- Language Engineer
- Communication Analyst
- Semantic Analyst
- Prompt Designer
- Translation Technology Expert
- Language Technology Analyst
- Language Process Analyst
- Machine Learning Evaluator
- AI Ethicist
- Translation Quality Assessor
- “Gatekeeper” Translator
- “Gatekeeper” Interpreter

Data Curator

Tasks:

- Ensure data quality and consistency through the implementation of data governance practices
- Define and enforce data styling and terminology standards
- Develop and maintain a data catalog to ensure efficient data discovery and accessibility
- Work with stakeholders to understand data requirements and ensure data availability and accuracy
- Collaborate with data scientists and machine learning teams to support the development of new models and features
- Manage data retention and archival policies to ensure compliance with data privacy regulations

Requirements:

- Experience in data management and data governance
- Familiarity with machine learning and data science methodologies
- Strong understanding of data storage and retrieval systems
- Experience with SQL and NoSQL databases
- Excellent written and verbal communication skills
- Ability to work with cross-functional teams and manage stakeholder relationships
- Strong problem-solving and analytical skills
- Bachelor's or Master's degree in Computer Science, Data Science, or a related field

Machine Learning Evaluator

Tasks:

- A Machine Learning Evaluator evaluates the performance and effectiveness of machine learning models, particularly Neural Machine Translation (NMT) and Large Language Models.
- Develop and implement testing methodologies to evaluate the performance and accuracy of NMT and Large Language Models.
- Analyze and interpret the output of machine learning models to identify strengths, weaknesses, and areas for improvement.
- Collaborate with engineering and research teams to provide recommendations and feedback for improving machine learning models.
- Conduct experiments and A/B testing to evaluate the impact of changes to machine learning models.

Requirements:

- Bachelor's or Master's degree in Computer Science, Mathematics, or a related field.
- 2+ years of experience in machine learning evaluation or related field.
- Strong understanding of machine learning concepts and techniques.
- Experience with NMT and Large Language Models.
- Proficiency in programming languages such as Python or R.
- Familiarity with machine learning tools and frameworks such as TensorFlow, PyTorch, and Keras.
- Strong analytical and problem-solving skills.
- Excellent communication and collaboration skills.

Gatekeeper Translator

Tasks:

- Ensure that automatically translated texts meet quality standards for their respective use cases
- Sample translations using traditional language quality assessment techniques
- Ensure compliance with corporate style and terminological guidelines
- Evaluate the suitability of translations for the intended audience
- Provide feedback to the translation team to improve the quality of future translations
- Stay up-to-date with the latest advances in translation technology and quality assessment techniques

Requirements:

- Bachelor's or Master's degree (or equivalent experience) in Translation, Modern Languages, Linguistics, or a related field
- Strong background in translation and language quality assessment
- Excellent written and verbal communication skills in source and target languages
- Ability to work with cross-functional teams and manage stakeholder relationships
- Strong problem-solving and analytical skills
- Familiarity with machine learning and NLP techniques

Language Engineer

Tasks:

- Design and implement AI models to process and understand multimodal language data, including translation memories, monolingual texts, voice recordings, and video transcriptions.
- Develop and implement algorithms for natural language understanding, natural language processing, and natural language generation.
- Work with cross-functional teams to understand customer requirements and design solutions that meet those needs.
- Collaborate with data scientists and engineers to ensure high quality and accuracy of language models.
- Keep up to date with the latest developments in AI and natural language processing.

Requirements:

- Strong programming skills in Python and/or other programming languages.
- Experience with NLP libraries such as NLTK, spaCy, or PyTorch.
- Strong understanding of natural language processing, including syntax, semantics, and pragmatics.
- Experience working with machine learning algorithms, including deep learning, decision trees, and random forests.
- Excellent written and verbal communication skills.
- Bachelor's or Master's degree in Computer Science, Linguistics, or a related field.

We’ve already come so far

An IBM Electronic Calculator speeds through thousands of intricate computations so quickly that on many complex problems it's like have 150 EXTRA Engineers.
No longer must valuable engineering personnel ... now in critical shortage ... spend priceless creative time at routine repetitive figuring.
Thousands of IBM Electronic Business Machines ... vital to our nation's defense ... are at work for science, industry, and the armed forces, in laboratories, factories, and offices, helping to meet urgent demands for greater production.

1953 advertisement for the IBM 701 mainframe

Is there an enemy here?

You are not in competition with AI.

You are in competition with people who use AI better than you do.

What strengths will be required to thrive in these jobs?

- All the great skills you already have
  - Subtle and sophisticated knowledge of language
  - Excellent proficiency in two or more languages
  - No fear of other languages
- Deep curiosity
- An appreciation for “algorithmic thinking”
- Comfort level (or better) with data and databases
- The audacity to work on skills that will make the old you redundant

Sure, but what else?

Imagination

Thank you for your attention!
Let’s keep the conversation going!

https://www.linkedin.com/in/jaymarciano/