Challenges and perspectives in machine learning

Alexis Zubiolo
alexis.zubiolo@gmail.com

Data Science Team Lead @ Adcash

February 16, 2017

Purpose of this talk

Illustrate recent **trends** and **challenges** in machine learning on **practical examples**.

Supervised learning paradigm

Supervised learning paradigm

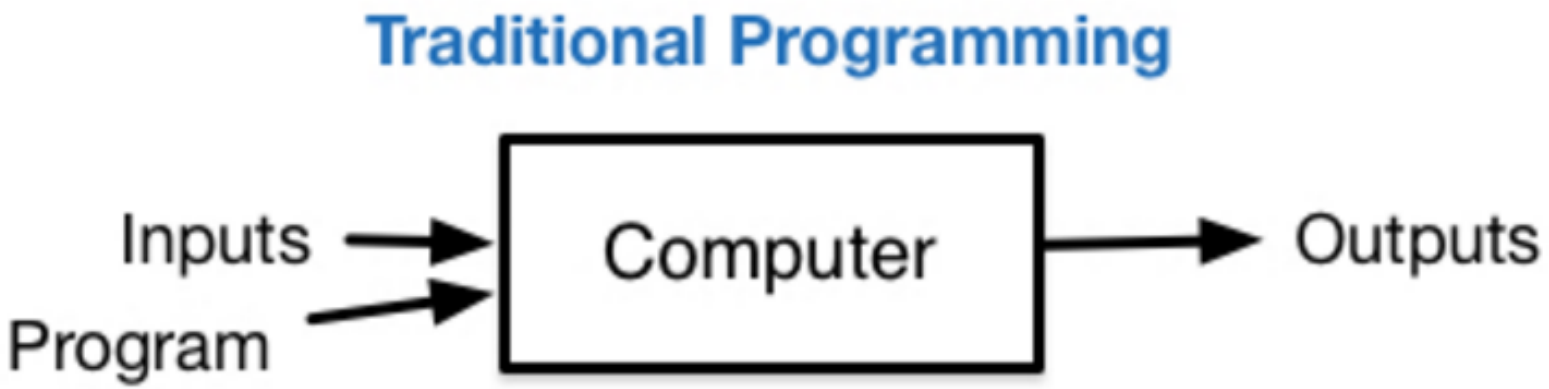
Goal: Build algorithms that can

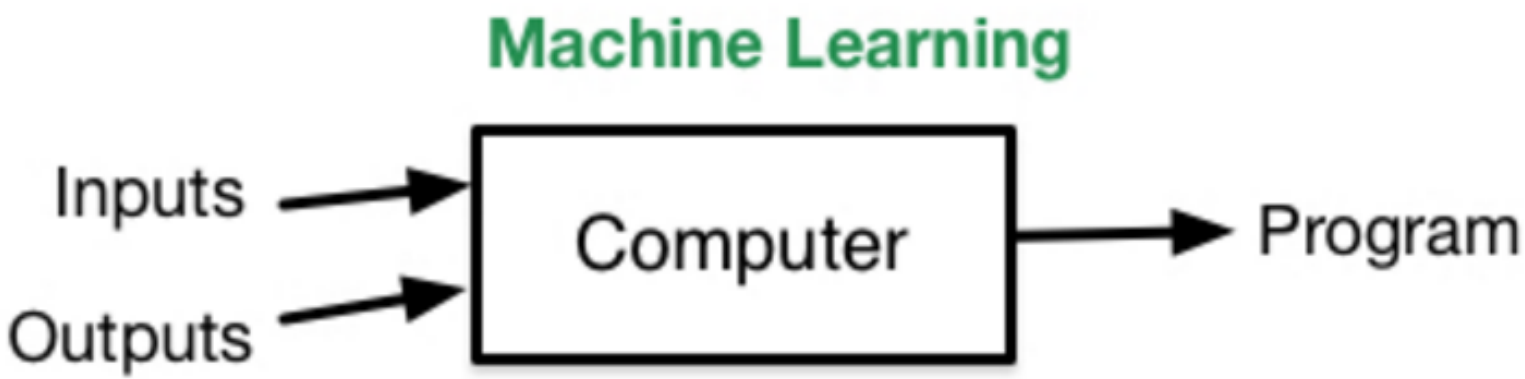
- **learn** from data
- **make predictions** on data

Supervised learning paradigm

Goal: Build algorithms that can

- **learn** from data
- **make predictions** on data





Example 1: Machine translation

Machine translation

Machine learning is a way to perform automated translation:

- Input = reference text
- Output = translated text

Machine translation

Machine learning is a way to perform automated translation:

- Input = reference text
- Output = translated text

There are differences between languages at different levels:

- Genders, e.g. neutral, masculine, feminine
- Conjugation
- Cases
- ...

Machine translation: A tricky example

Take a close look at these two sentences:

Machine translation: A tricky example

Take a close look at these two sentences:

The trophy could not fit in the bag because **it** is too small.

and

The trophy could not fit in the bag because **it** is too big.

Machine translation: A tricky example

Take a close look at these two sentences:

The trophy could not fit in the bag because **it** is too small.

and

The trophy could not fit in the bag because **it** is too big.

What does the word "**it**"refer to in both cases?

Machine translation: A tricky example

Take a close look at these two sentences:

The trophy could not fit in **the bag** because **it** is too small.

and

The trophy could not fit in the bag because **it** is too big.

What does the word "**it**"refer to in both cases?

Machine translation: A tricky example

Based on the context, we can guess what "**it**"refers to.

Machine translation: A tricky example

Based on the context, we can guess what "**it**"refers to.

This is key because, if we wanted to translate this sentence into Bulgarian, we would be faced to the following challenge:

- Trophy = трофей is masculine
- Bag = чанта is feminine

Machine translation: A tricky example

Based on the context, we can guess what "**it**"refers to.

This is key because, if we wanted to translate this sentence into Bulgarian, we would be faced to the following challenge:

- Trophy = трофей is masculine
- Bag = чанта is feminine

Consequence: "it"would refer to a masculine or a feminine noun, depending on the context.

Machine translation: A tricky example

Let's translate this sentence **from English to Bulgarian**... Here is my attempt:

Machine translation: A tricky example

Let's translate this sentence **from English to Bulgarian**... Here is my attempt:

The trophy could not fit in the **bag** because **it** is too small.

Machine translation: A tricky example

Let's translate this sentence **from English to Bulgarian**... Here is my attempt:

The trophy could not fit in the **bag** because **it** is too small.
= Трофеят не може да се побере в **чантата**, защото (**тя**) е твърде **малка**.

Machine translation: A tricky example

Let's translate this sentence **from English to Bulgarian**... Here is my attempt:

The trophy could not fit in the **bag** because **it** is too small.
= Трофеят не може да се побере в **чантата**, защото (**тя**) е твърде **малка**.

The **trophy** could not fit in the bag because **it** is too big.

Machine translation: A tricky example

Let's translate this sentence **from English to Bulgarian**... Here is my attempt:

The trophy could not fit in the **bag** because **it** is too small.
= Трофеят не може да се побере в **чантата**, защото (**тя**) е твърде **малка**.

The **trophy** could not fit in the bag because **it** is too big.
= **Трофеят** не може да се побере в чантата, защото (**той**) е твърде **голям**.

Machine translation: A tricky example

Let's translate this sentence **from English to Bulgarian**... Here is my attempt:

The trophy could not fit in the **bag** because **it** is too small.
= Трофеят не може да се побере в **чантата**, защото (**тя**) е твърде **малка**.

The **trophy** could not fit in the bag because **it** is too big.
= **Трофеят** не може да се побере в чантата, защото (**той**) е твърде **голям**.

This is often referred to as **long short-term memory** (LSTM).

Another example

Classic challenge in machine translation: **homonyms**.

Another example

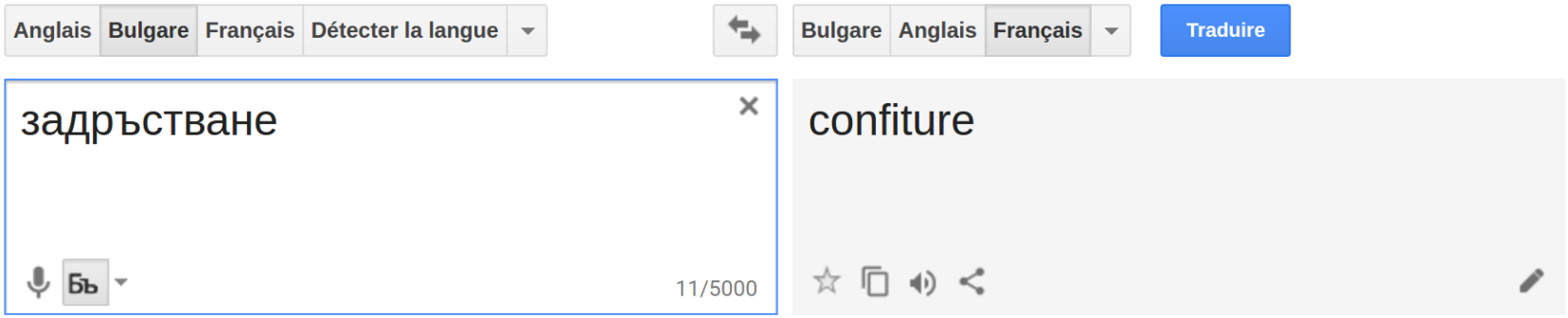
Classic challenge in machine translation: **homonyms**.

Example:

Another example

Classic challenge in machine translation: **homonyms**.

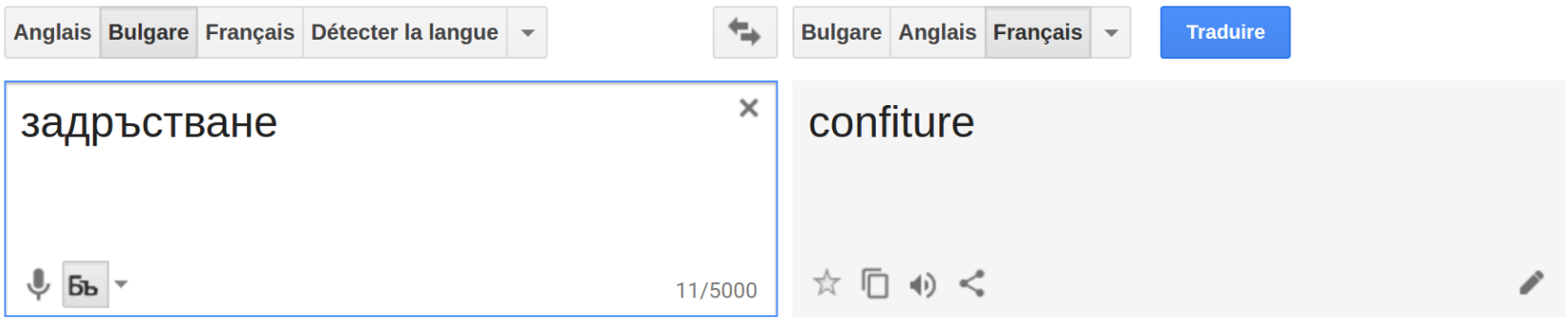
Example:



Another example

Classic challenge in machine translation: **homonyms**.

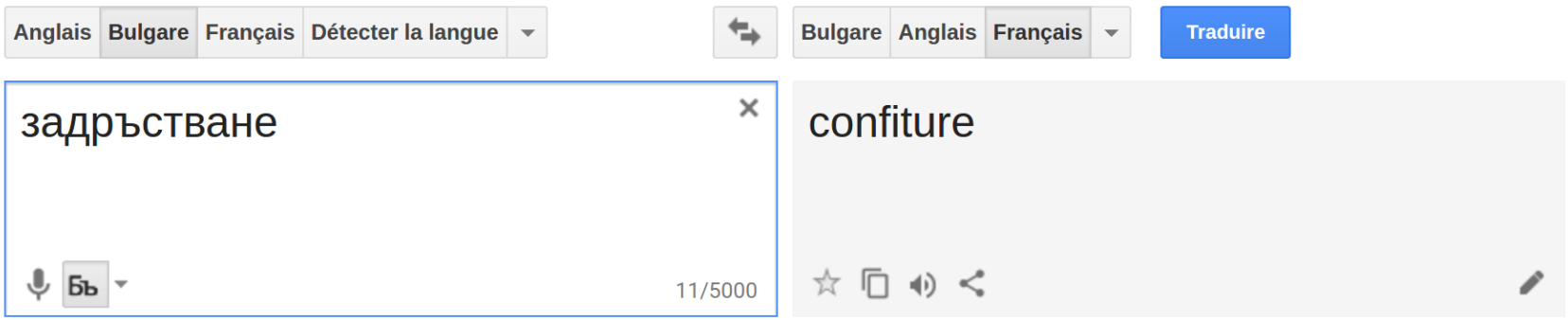
Example:



Another example

Classic challenge in machine translation: **homonyms**.

Example:

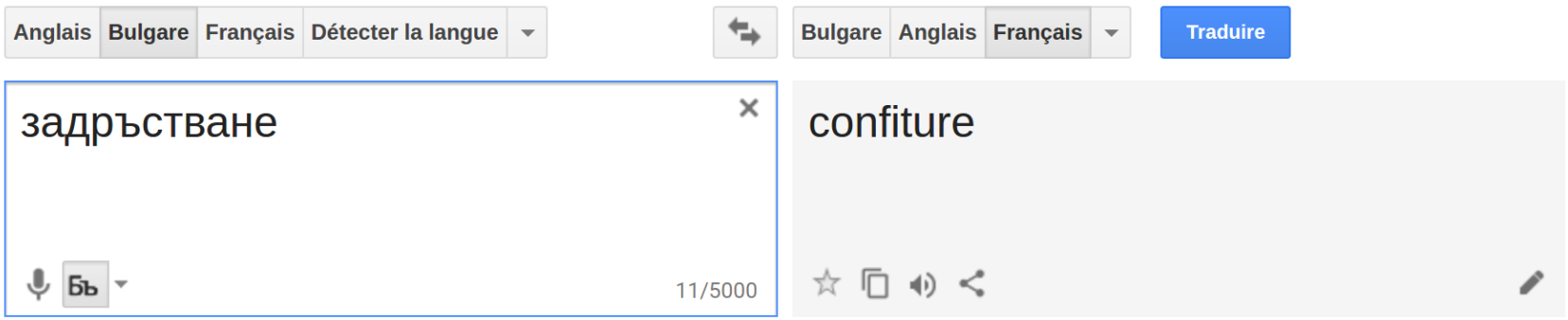


How could it be?

Another example

Classic challenge in machine translation: **homonyms**.

Example:



How could it be? This is **jam** in English.

Example 2: Acquiring data

Acquiring data

Let us switch to another example:

Acquiring data

Let us switch to another example: Reading house numbers in Google street.

Acquiring data

Let us switch to another example: Reading house numbers in Google street.

Acquiring data

Why would we do that?

Acquiring data

Why would we do that? In order to locate houses accurately.

Acquiring data

Why would we do that? In order to locate houses accurately.

Acquiring data

There exist really good machine learning algorithms to read numbers automatically but . . .

Acquiring data

There exist really good machine learning algorithms to read numbers automatically but . . . They need **training data**, that is:

- inputs (images with numbers on it): They **already have**

Acquiring data

There exist really good machine learning algorithms to read numbers automatically but . . . They need **training data**, that is:

- inputs (images with numbers on it): They **already have**
- outputs (value): They **don't have**

Acquiring data

There exist really good machine learning algorithms to read numbers automatically but . . . They need **training data**, that is:

- inputs (images with numbers on it): They **already have**
- outputs (value): They **don't have**

What did they do to obtain this labelled data?

Acquiring data

There exist really good machine learning algorithms to read numbers automatically but . . . They need **training data**, that is:

- inputs (images with numbers on it): They **already have**
- outputs (value): They **don't have**

What did they do to obtain this labelled data? Propose a free **CAPTCHA** system . . .

Acquiring data

There exist really good machine learning algorithms to read numbers automatically but . . . They need **training data**, that is:

- inputs (images with numbers on it): They **already have**
- outputs (value): They **don't have**

What did they do to obtain this labelled data? Propose a free **CAPTCHA** system . . .

Acquiring data

There exist really good machine learning algorithms to read numbers automatically but . . . They need **training data**, that is:

- inputs (images with numbers on it): They **already have**
- outputs (value): They **don't have**

What did they do to obtain this labelled data? Propose a free **CAPTCHA** system . . .

. . . so that real human can provide labels (outputs) to their images (inputs)!

Conclusion

Machine learning is not only about designing super sophisticated algorithms (even if it helps!).

Conclusion

Machine learning is not only about designing super sophisticated algorithms (even if it helps!).

Gathering **a significant amount of (clean) data** is **crucial**!

Conclusion

Machine learning is not only about designing super sophisticated algorithms (even if it helps!).

Gathering **a significant amount of (clean) data** is **crucial**!

Understanding **corner cases** is **key**!

Thank you! Questions?

alexis.zubiolo@gmail.com