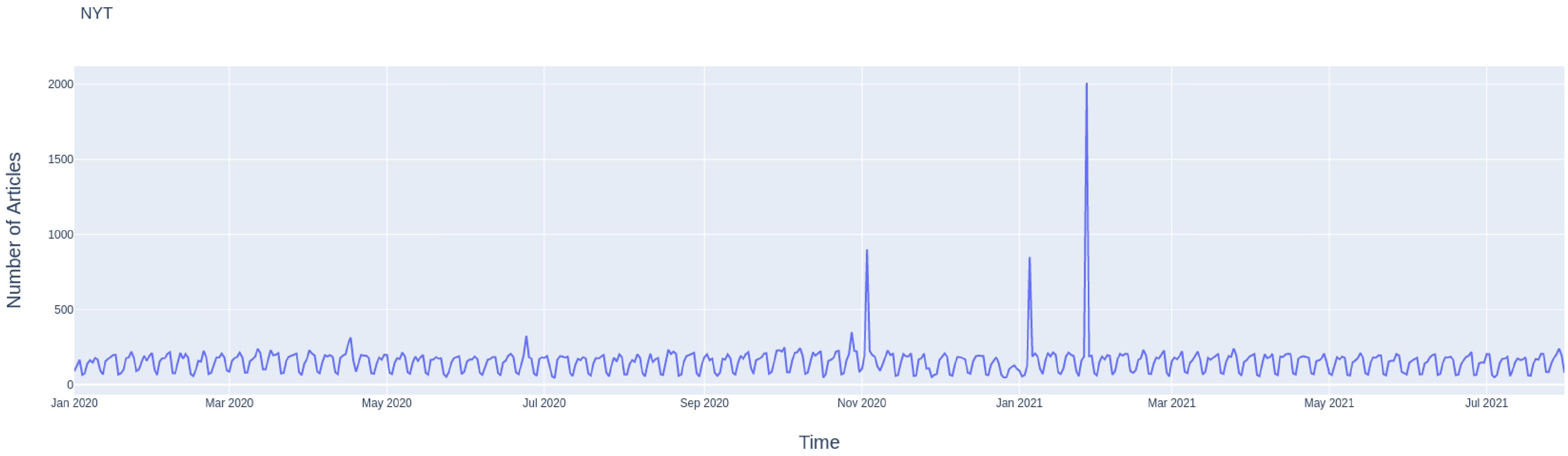
Introduction to Data Mining

Kaihua Ding, Ph.D.

Too much data too little time – data mining



- NYT publish around 155 articles per day
- Spend 10 minutes per article -> over 25 hours a day to analyze
- ...

Too much data too little time – data mining

"Computers have promised us a fountain of wisdom but delivered a flood of data."

"It has been estimated that the amount of information in the world doubles every 20 months."

--- Frawley, Piatetsky-Shapiro, Matheus, 1992

What is not data mining?

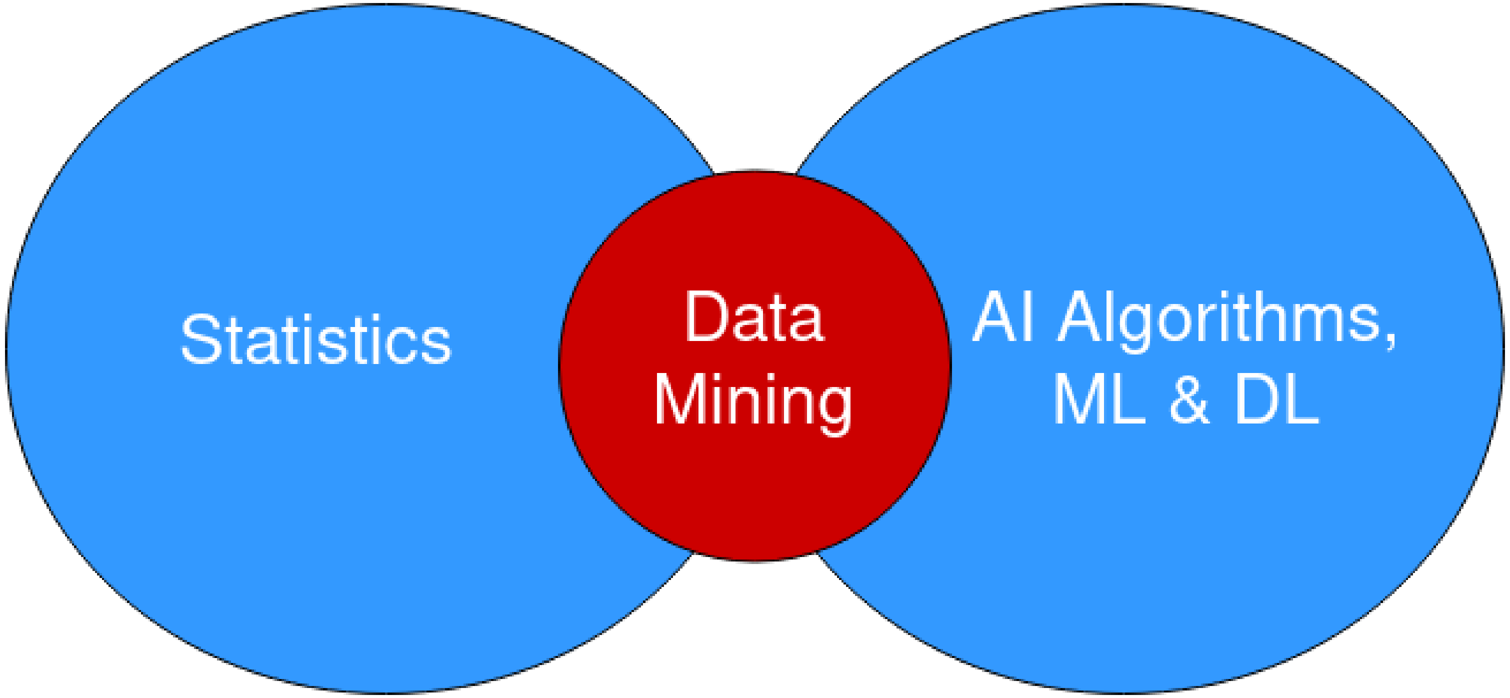
"An unethical econometric practice of massaging and manipulating the data to obtain the desired results."

-- William Brown, Introducing Econometrics

"Torturing data until it confesses ... and if you torture it enough, it will confess to anything."

--Jeff Jonas, IBM fellow

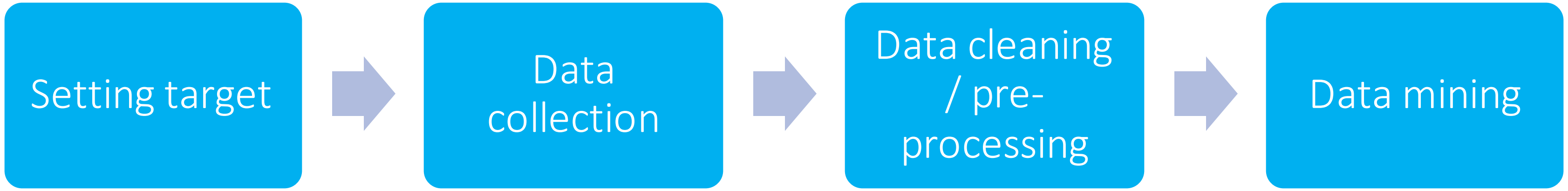
Origins of data mining



Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems

Many definitions of data mining

- Non-trivial extraction of implicit, previously unknown and potentially useful information from data
- Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns
- ...



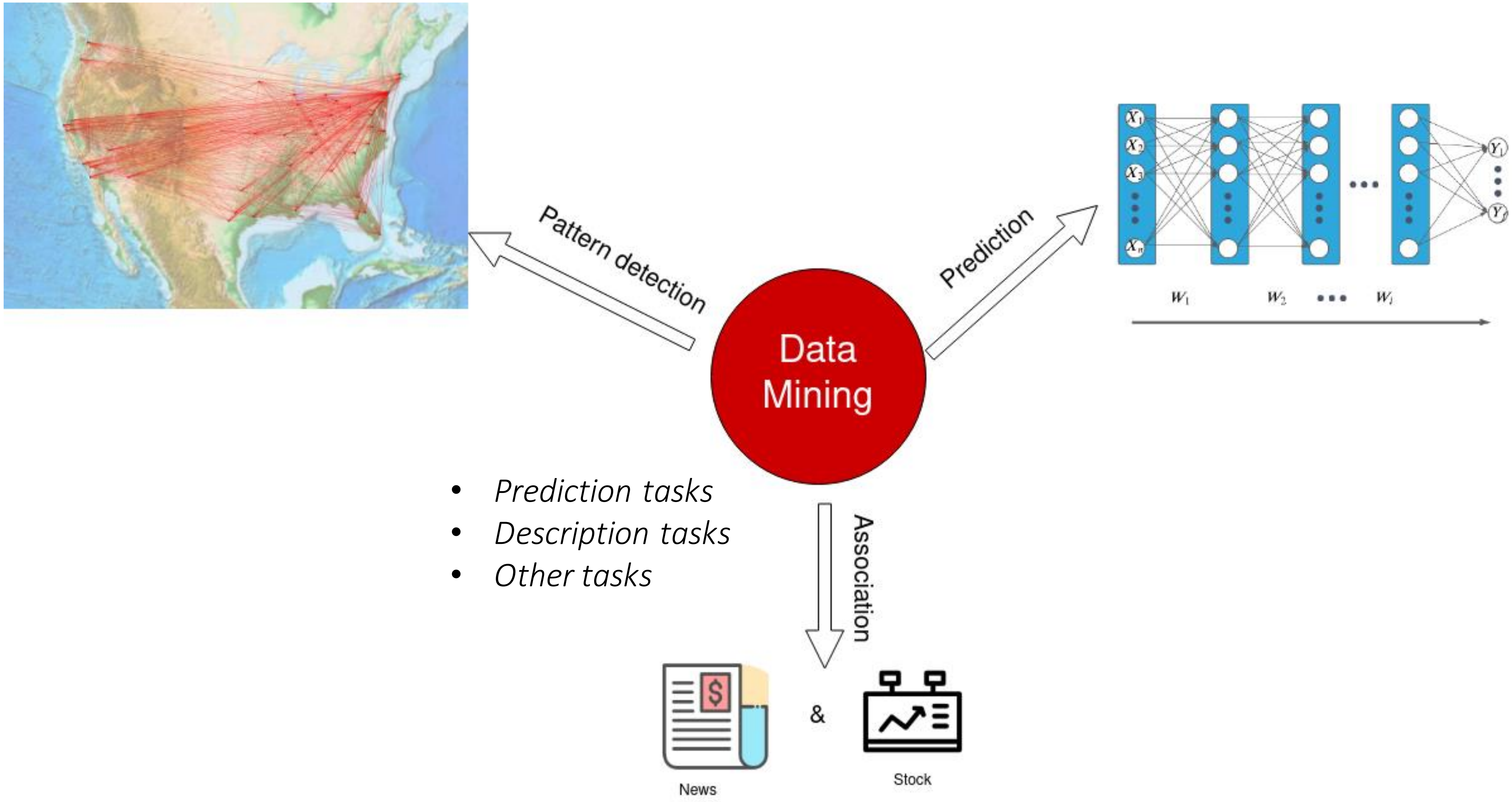
Setting target → Data collection → Data cleaning / pre-processing → Data mining

- 📚
- 🎥
- 🌐
- 🎼
- ...

- Web scrapping
- Data generator
- ...

- Graph
- Articles
- ...

Many tasks of data mining



- *Prediction tasks*
- *Description tasks*
- *Other tasks*

Workshop materials

Github repo,
https://github.com/rcc-uchicago/introduction_to_data_mining

Google Colab,
https://colab.research.google.com/drive/1WItmbi5QqntkcZbJqjeAg8VrUJVxknVn?usp=sharing

Tutorial 0, setting target

News!

Web scraping vs API Zoom Poll

Tutorial 1.0 data collection

API packages

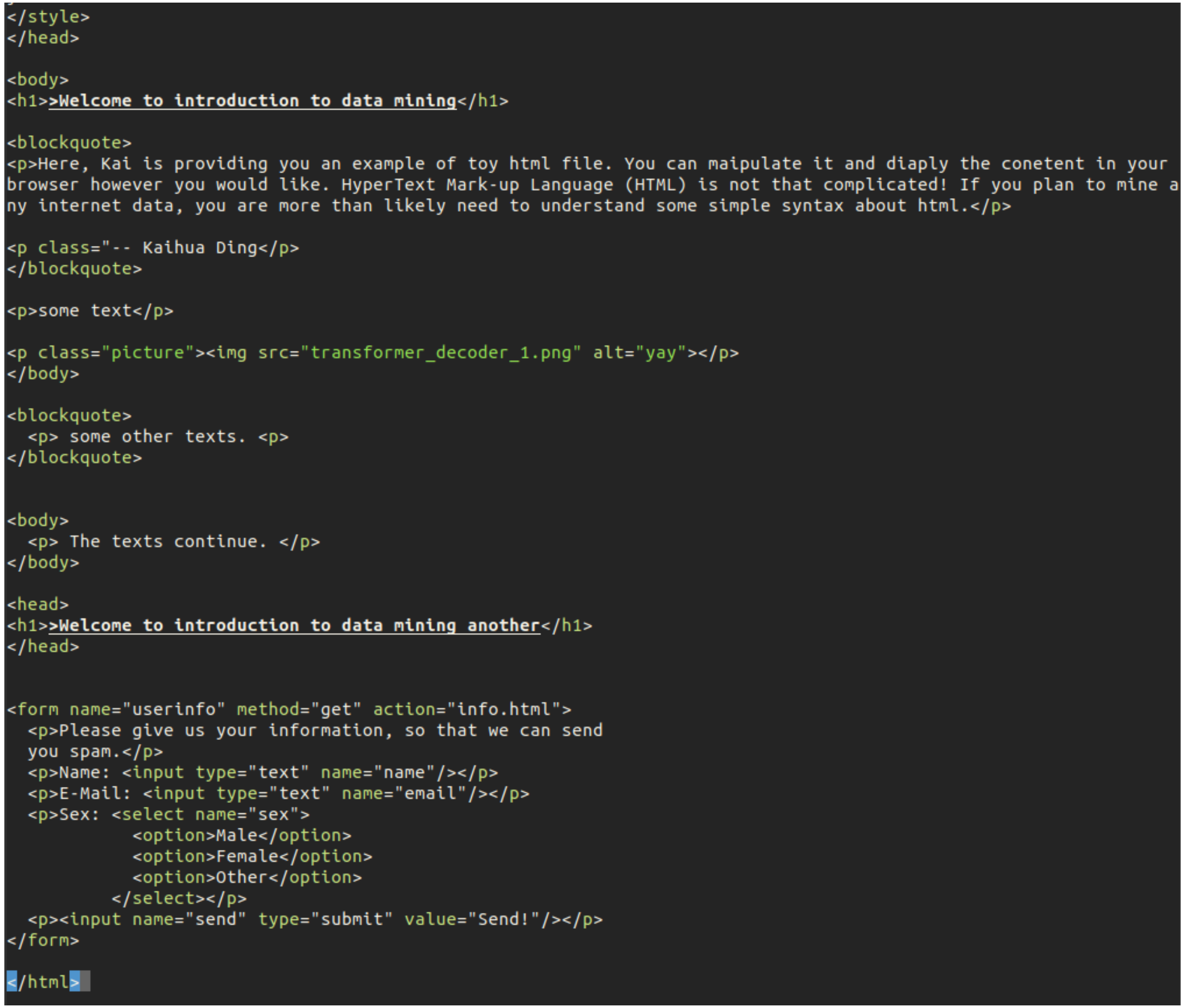
- Python requests package
- Beautiful soup
- …

Command line software

- Wget (gnu, standalone, more feature)
- cURL (faster, a library)
- …

In most cases, they are just sending http "GET" requests to a server.

Tutorial 1.1 data collection



```
</style>
</head>

<body>
<h1>>Welcome to introduction to data mining</h1>

<blockquote>
<p>Here, Kai is providing you an example of toy html file. You can maipulate it and diaply the conetent in your
browser however you would like. HyperText Mark-up Language (HTML) is not that complicated! If you plan to mine a
ny internet data, you are more than likely need to understand some simple syntax about html.</p>

<p class="-- Kaihua Ding</p>
</blockquote>

<p>some text</p>

<p class="picture"><img src="transformer_decoder_1.png" alt="yay"></p>
</body>

<blockquote>
  <p> some other texts. <p>
</blockquote>


<body>
  <p> The texts continue. </p>
</body>

<head>
<h1>>Welcome to introduction to data mining another</h1>
</head>


<form name="userinfo" method="get" action="info.html">
  <p>Please give us your information, so that we can send
  you spam.</p>
  <p>Name: <input type="text" name="name"/></p>
  <p>E-Mail: <input type="text" name="email"/></p>
  <p>Sex: <select name="sex">
            <option>Male</option>
            <option>Female</option>
            <option>Other</option>
          </select></p>
  <p><input name="send" type="submit" value="Send!"/></p>
</form>

</html>
```

API packages

- parsed

Command line software

- parse yourself

Tutorial 2.0 data cleaning / pre-processing

*A **regular expression** (regex or regexp, also referred to as rational expression) is a sequence of characters that specifies a search pattern. Usually such patterns are used by string-searching algorithms for "find" or "find and replace" operations on strings, or for input validation. It is a technique developed in theoretical computer science and formal language theory.*

https://en.wikipedia.org/wiki/Regular_expression

https://docs.python.org/3/library/re.html

In some cases, data pre-processing or pre-examination can make a huge difference. E.g., data centric ai competition.

Any questions?

(5-minute break)

Tutorial 3.0 – data mining

Description tasks

- *Summarization*
- …

Prediction tasks

- *Classification (e.g. sentiment)*
- …

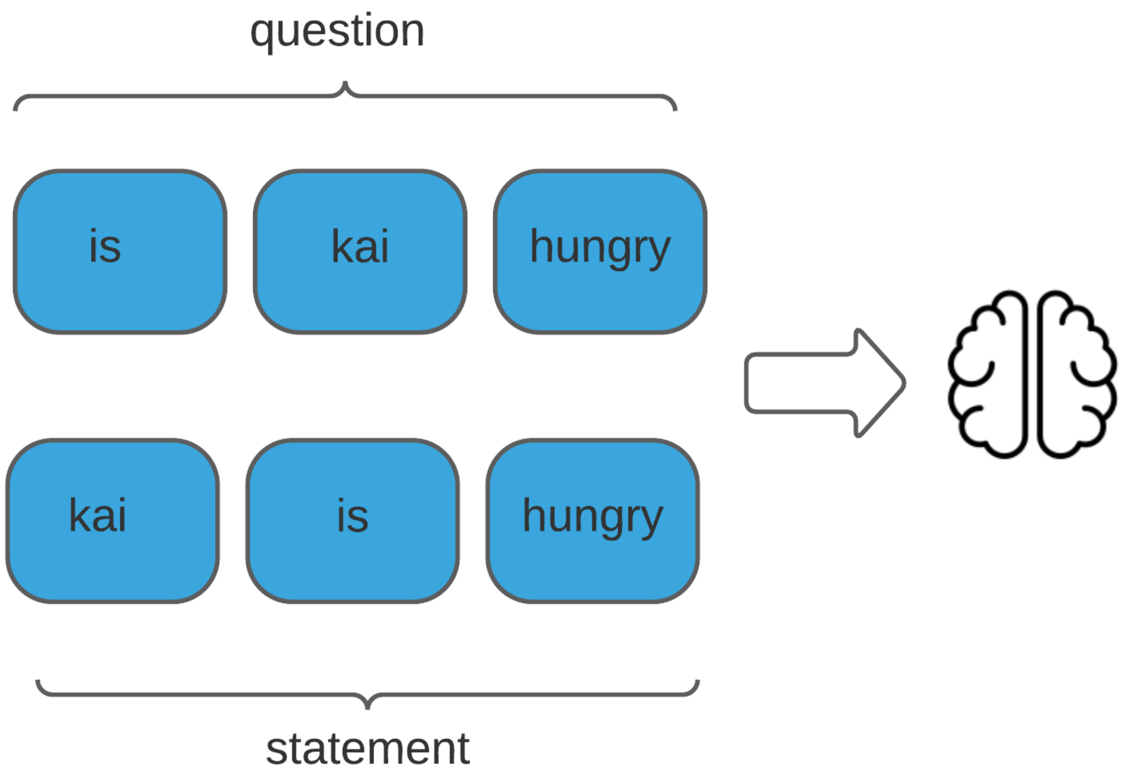
Other tasks

- …

→ Machine learning is a natural solution!

Spolier alert, it turns out some ML / DL methods are better than others...

Data mining, texts example – sequence with highly flexible and complex rules (AI)



question

is kai hungry

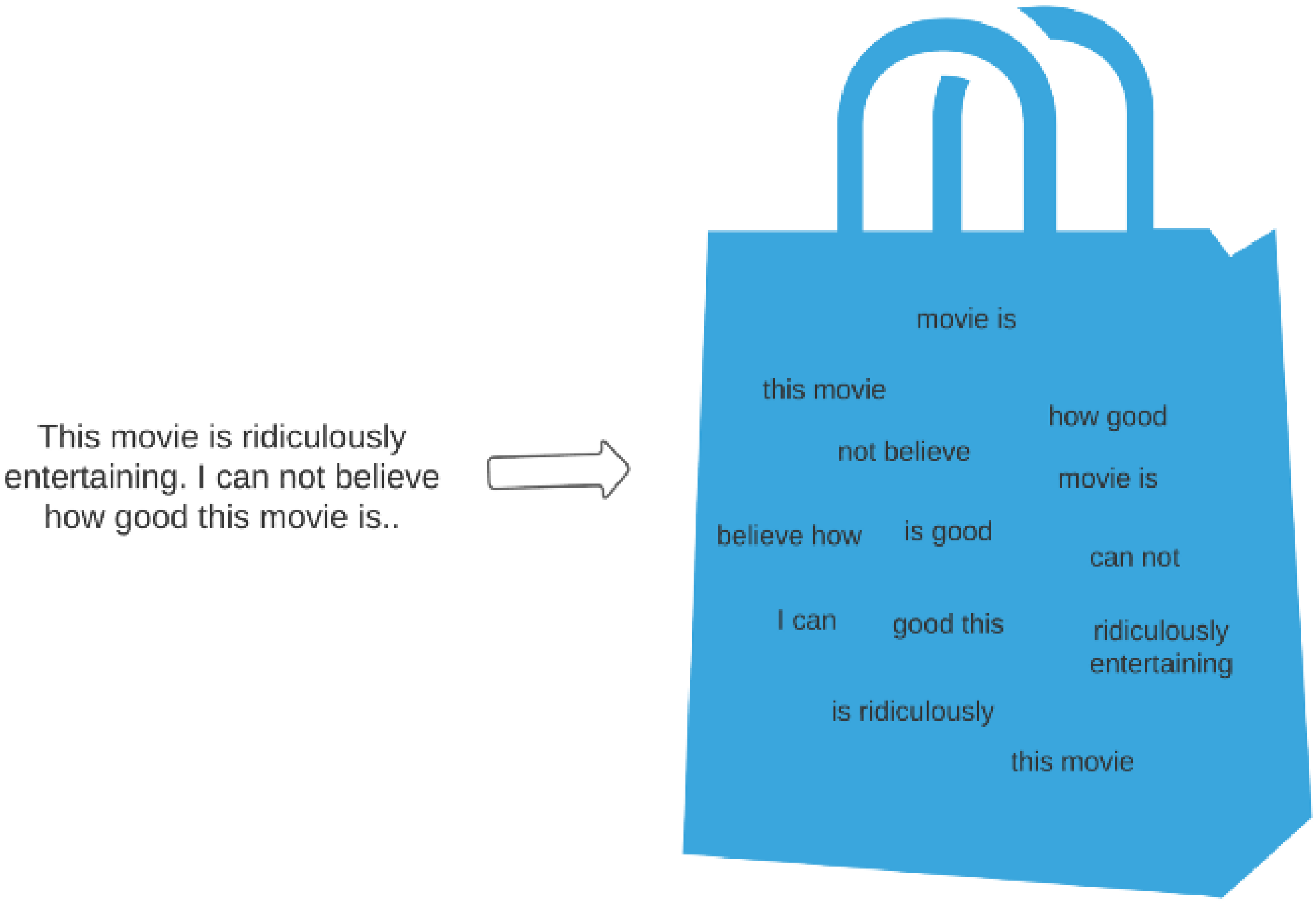
kai is hungry

statement

Same words with different sequence lead to different meaning!

Medium -- Kaihua Ding, 2021

Data mining, texts example – NLP applications were not pervasive 10 years ago



This movie is ridiculously entertaining. I can not believe how good this movie is..

movie is

this movie

how good

not believe

movie is

believe how

is good

can not

I can

good this

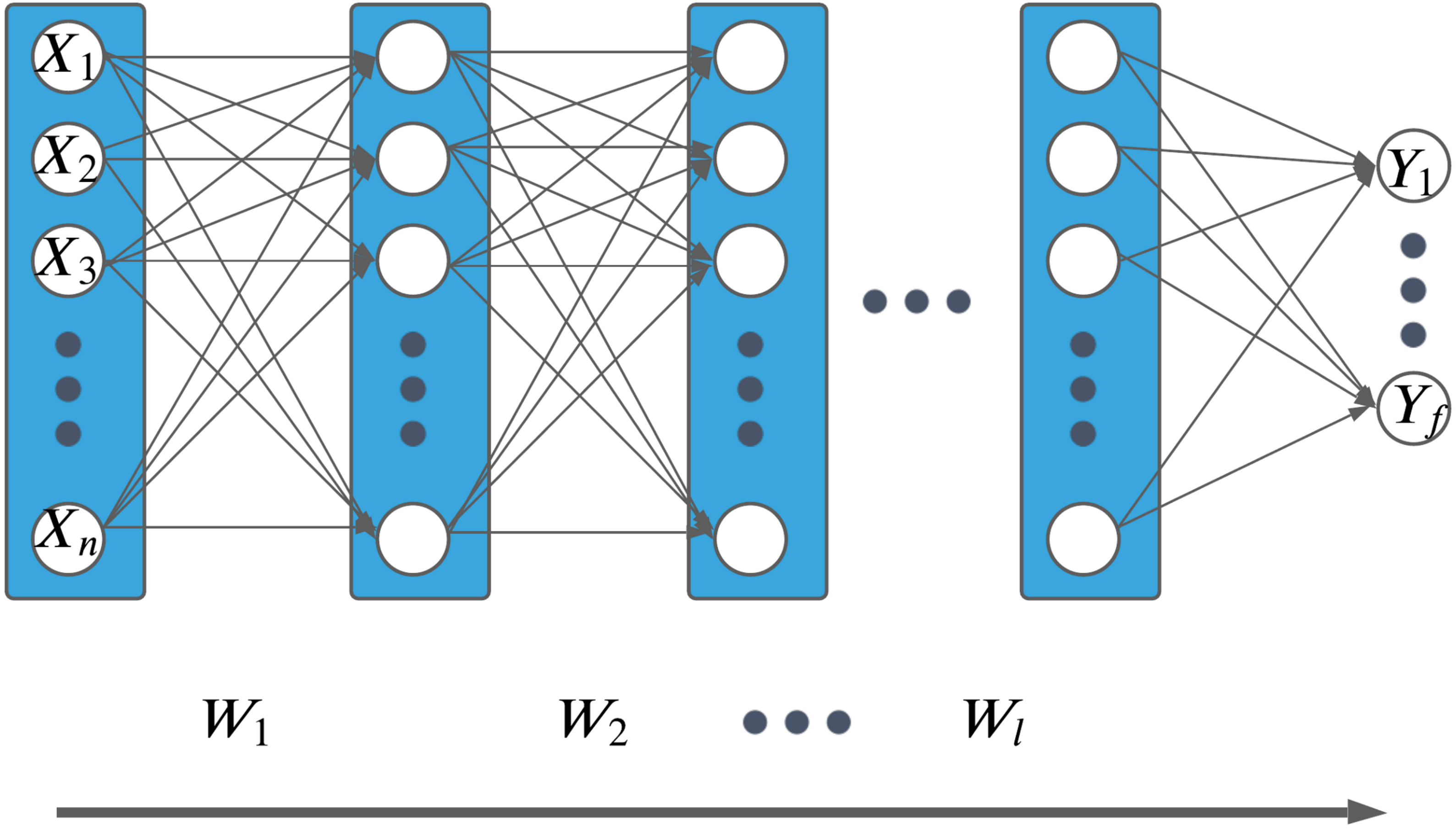
ridiculously entertaining

is ridiculously

this movie

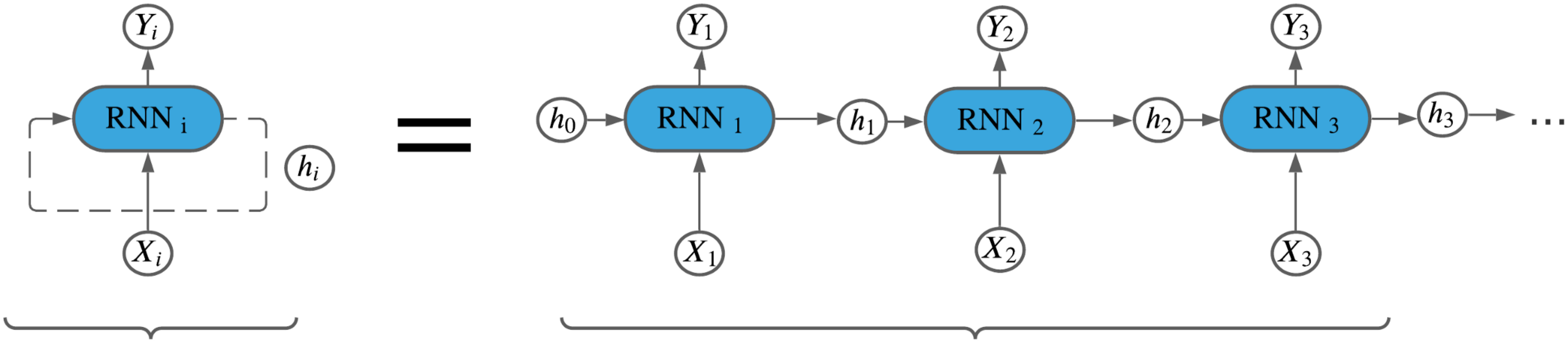
Medium -- Kaihua Ding, 2021

Data mining, texts example -- a generic neural net



[Medium -- Kaihua Ding, 2021]

Data mining, texts example -- sequence with highly flexible and complex rules (AI)

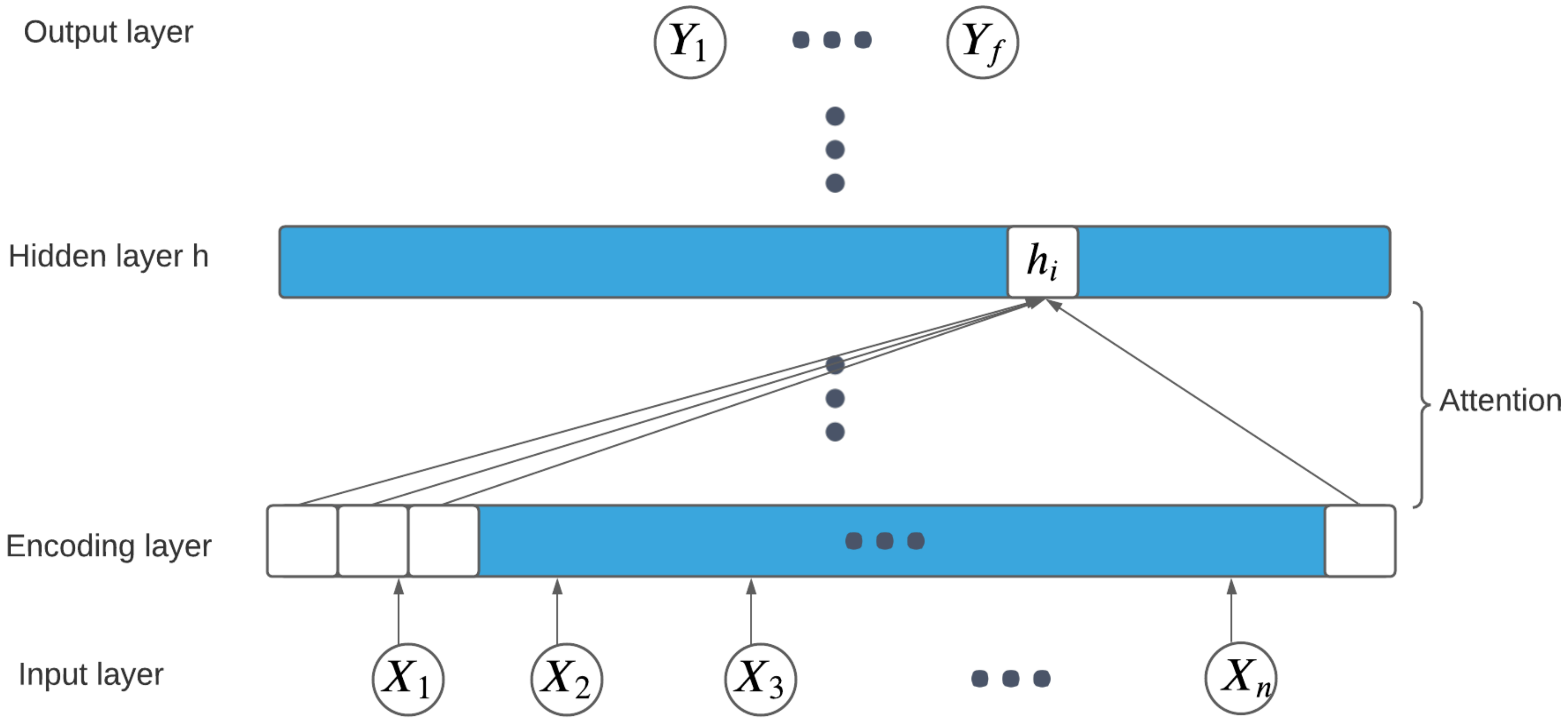


Recurrent unit, e.g. vanilla RNN, LSTM, GRU, etc

An unrolled recurrent neural network.

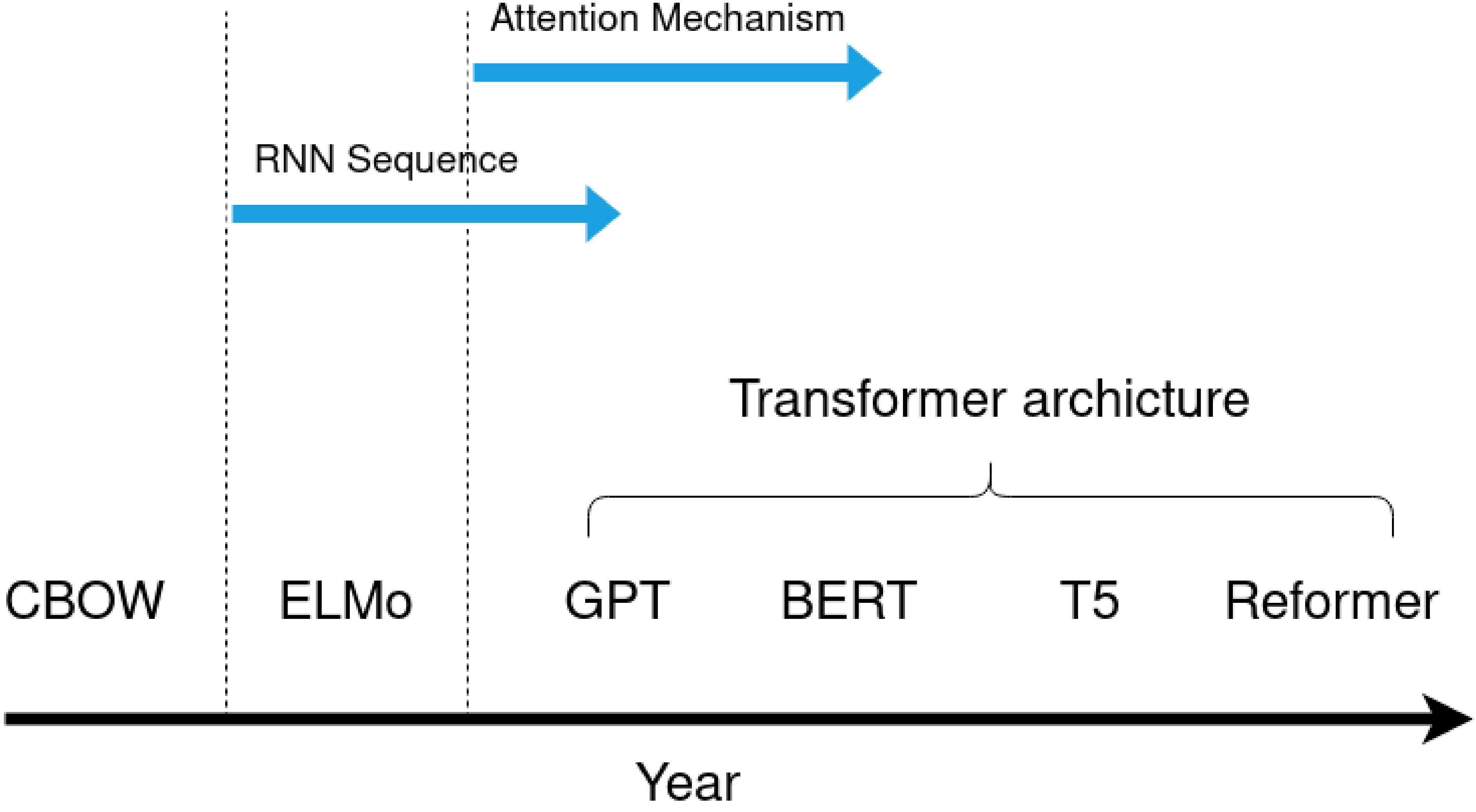
Medium -- Kaihua Ding, 2021

Data mining, texts example -- sequence with highly flexible and complex rules (AI)



Medium -- Kaihua Ding, 2021

Models are converging

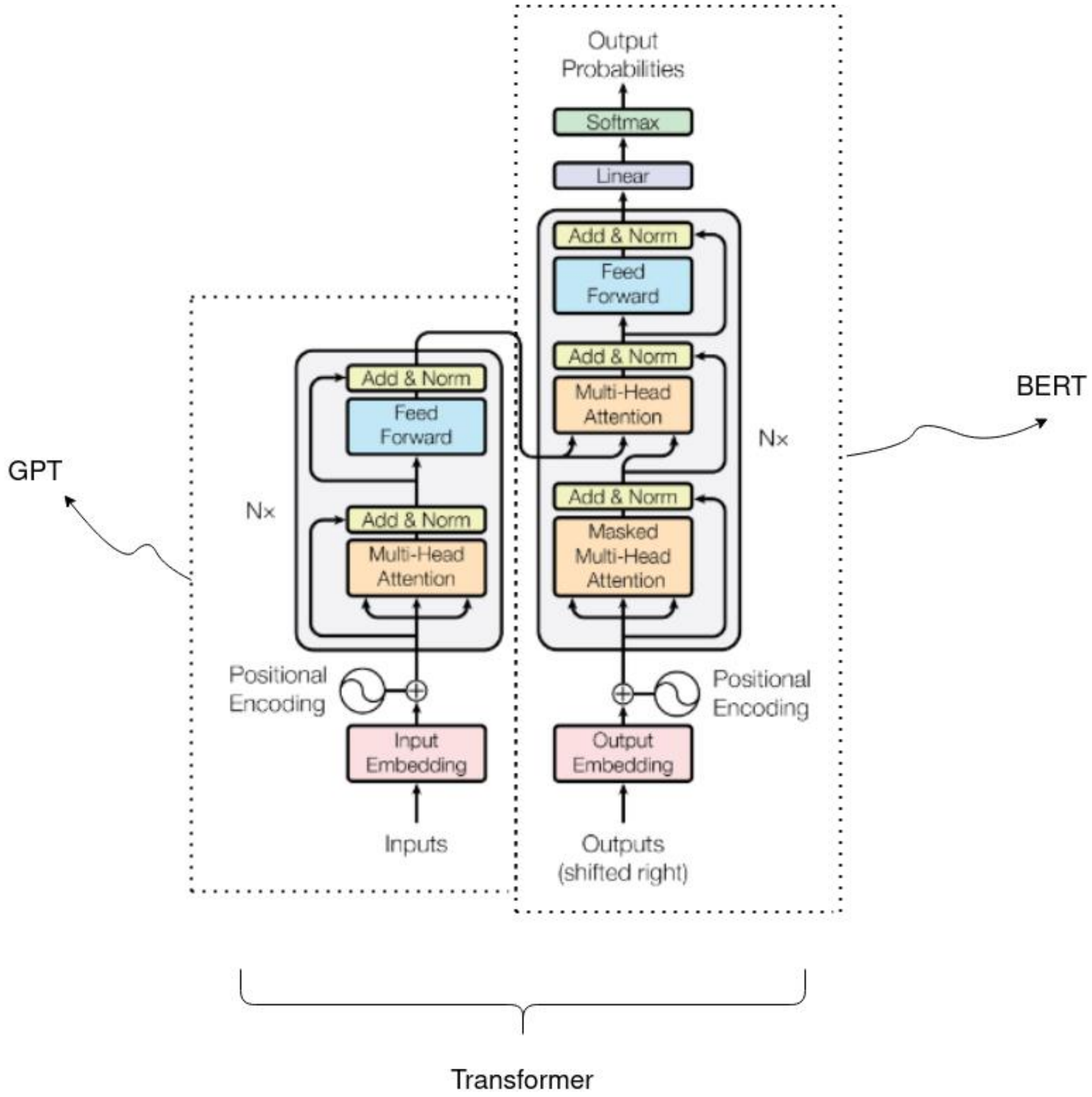


TextBlob vs Attention Zoom Poll

Tutorial 3.0 texts – NLP accuracy of older models are unflattering

Take a look at tutorial 3.0

Text algorithm -- transformer architecture

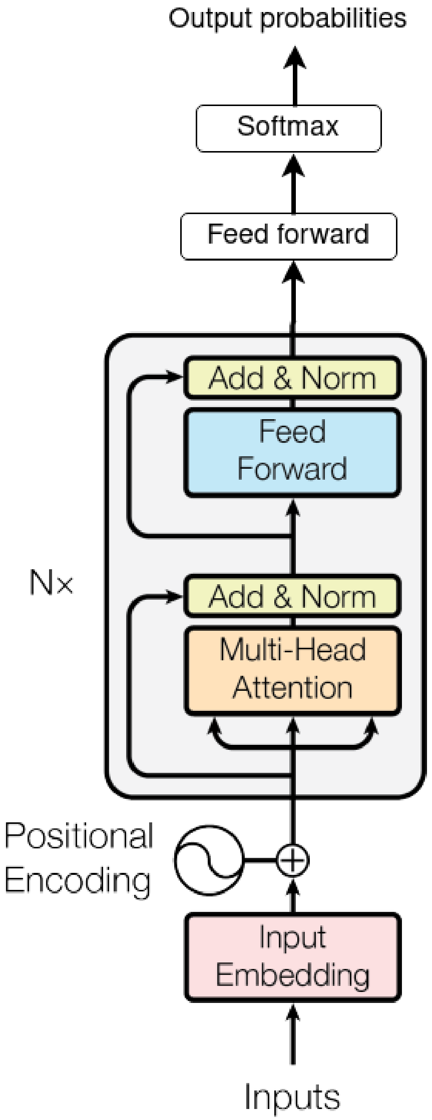


[Attention is all you need]

Any questions?

(5-minute break)

GPT-2 (transformer) – Tutorial 3.1



- Attention mechanism -- matrix multiplication enable each entries to pay "attention" to the receipient
- Multiple heads – allow for parallel computing (embarrassingly parallel)
- Positional encoding – sequence order

[GPT-2]()

[*Attention is all you need*]()

Generative Pre-trained Transformer 2 (GPT-2)

- The GPT-2 model has 1.5 billion parameters and was trained on a dataset of **8 million web pages**.
- Expensive to train and basically impossible to train on single CPU or GPU.
- If you insist, here are some tips to train large models, 1) parallel computing, 2) data generator, 3) store checkpoints, and 4) grid search

Tutorial 3.2.0 data mining

Description tasks

- Summarization ---> AI NLP with transformers

Prediction tasks

- Classification ---> sentiment classification with transformers

Tutorial 3.2 texts – data mining

Description tasks

- Summarization ---> AI NLP with transformers

Prediction tasks

- Classification ---> sentiment classification with transformers

Tutorial 3.3 data mining -- classification

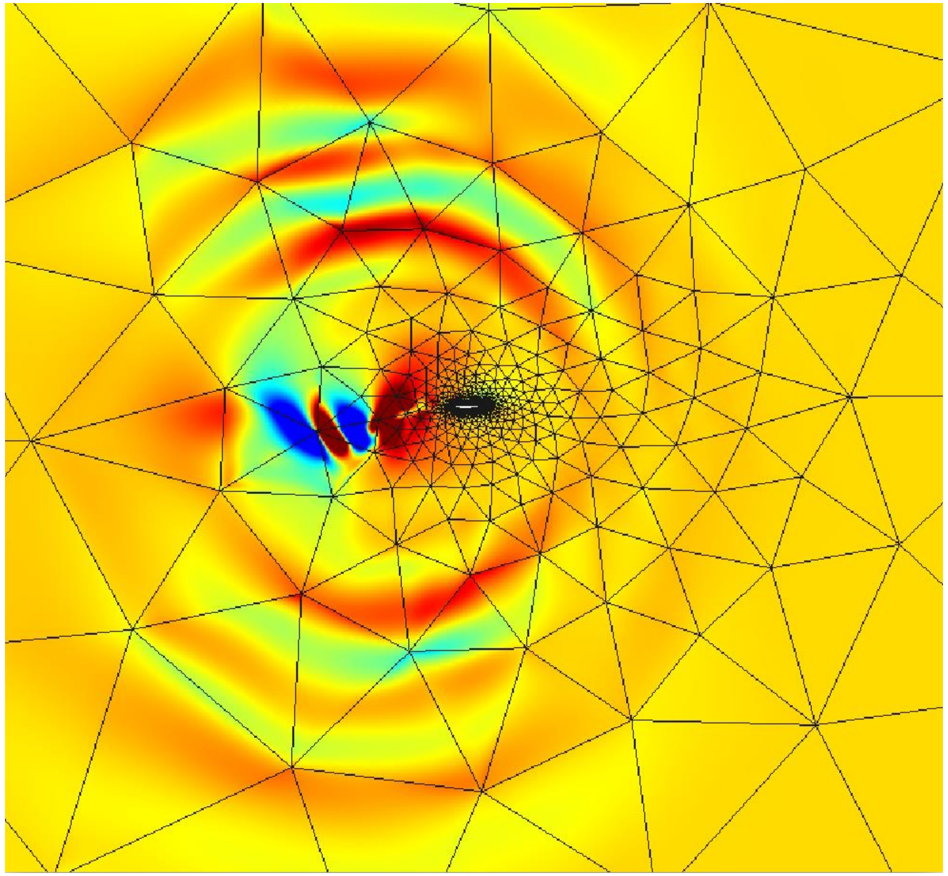
Description tasks

- Topic modeling ---> statics method
- Summarization ---> AI NLP with transformers

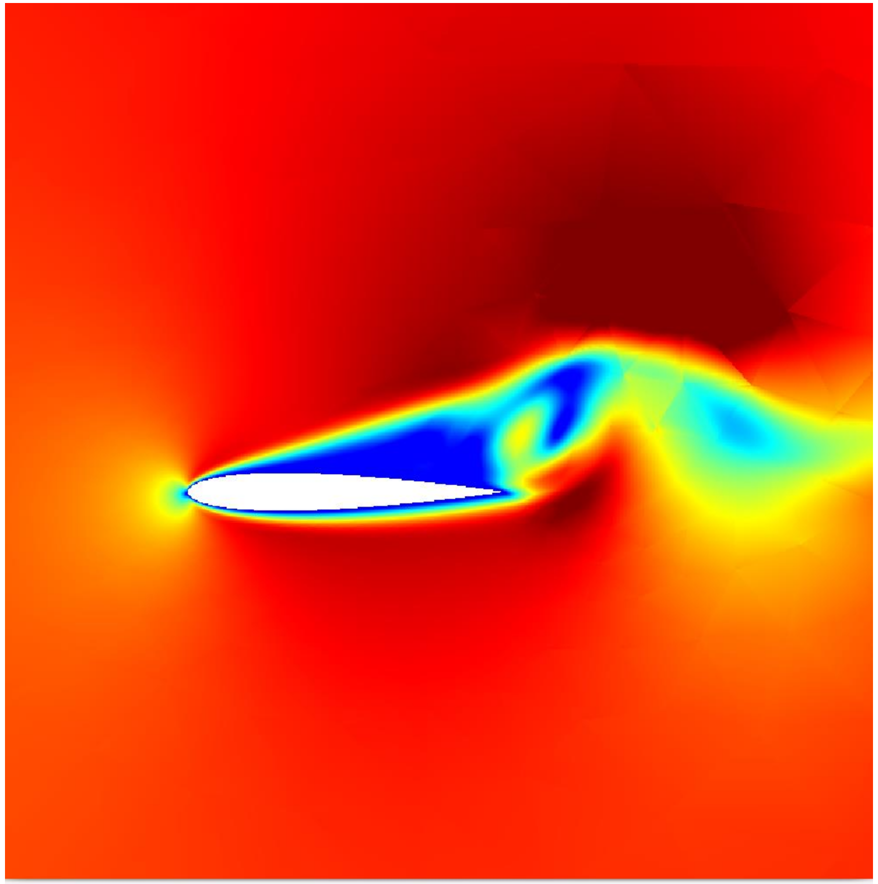
Prediction tasks

- Classification ---> sentiment classification with transformers

Data mining -- you might need to come up with your own methods …



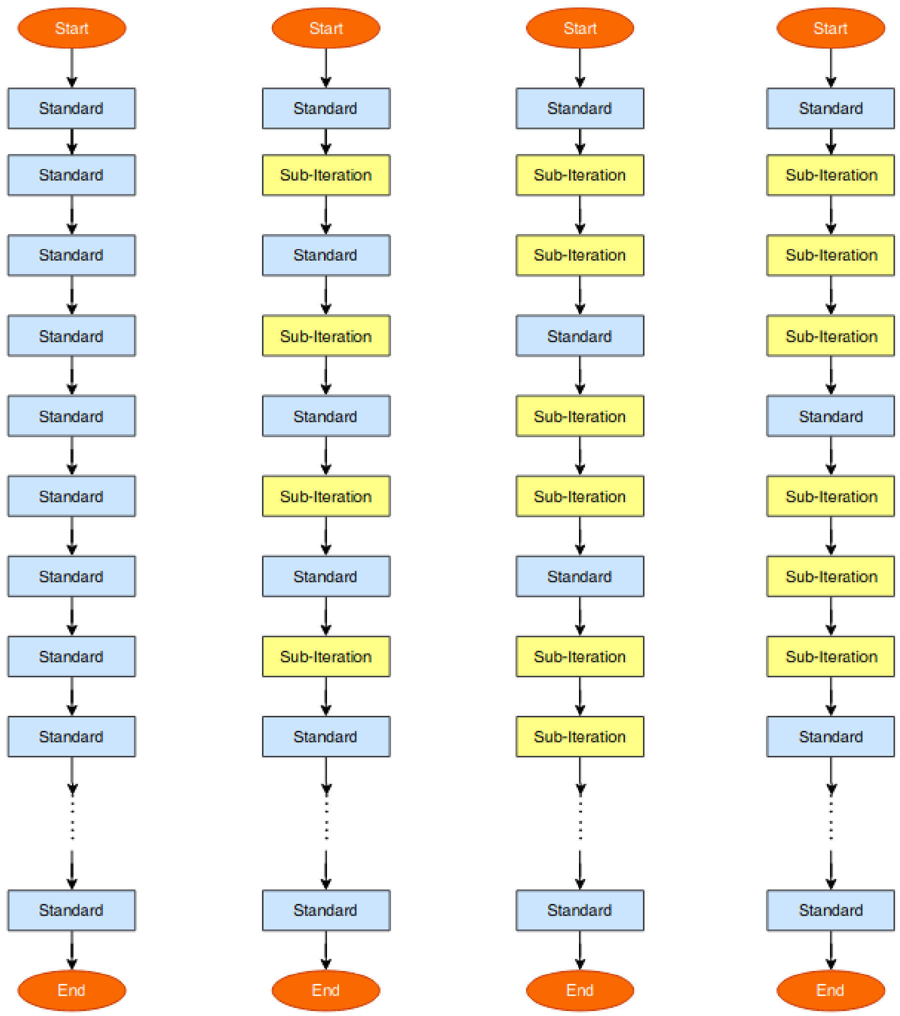
Airflow is highly flexible and unpredictable

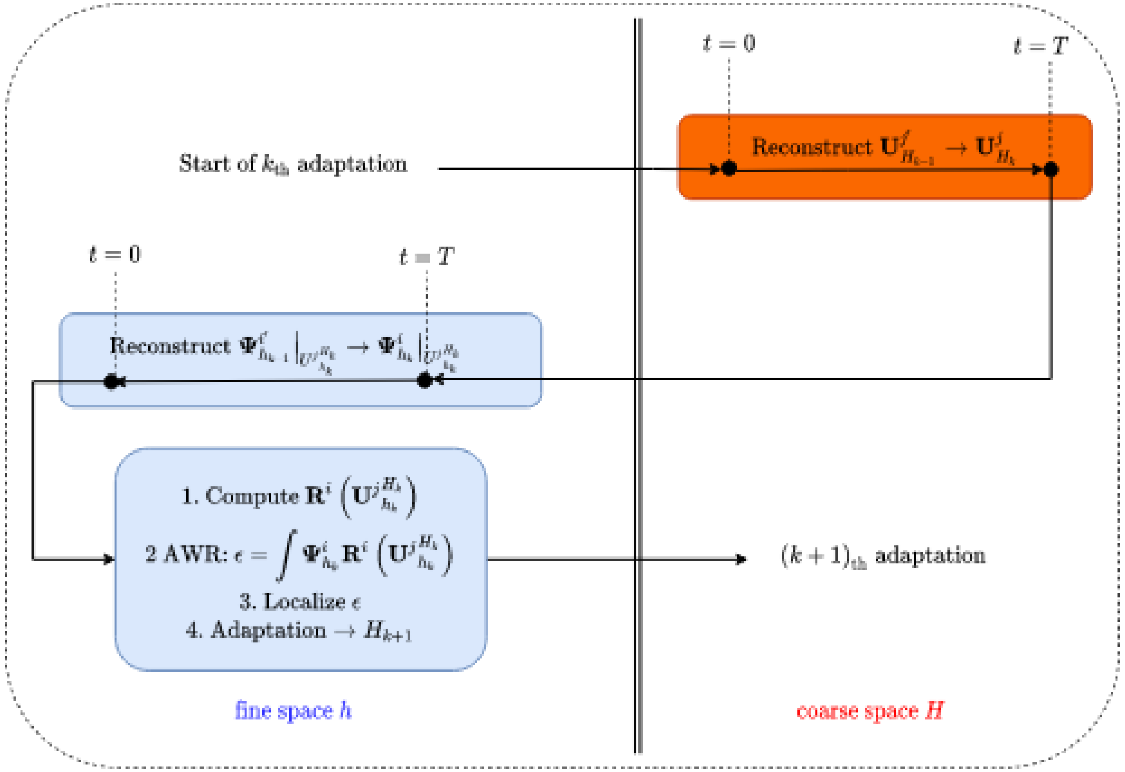


Vortex shedding and turbulence are intractable

Kaihua Ding, et al, 2021

Data mining -- your own methods …




$t = 0$
$t = T$
Reconstruct $\mathbf{U}^{f}_{H_{k-1}} \rightarrow \mathbf{U}^{i}_{H_k}$
Start of k_{th} adaptation
$t = 0$
$t = T$
Reconstruct $\mathbf{\Psi}^{f}_{h_{k-1}}\big|_{U^{j}{}^{H_k}_{h_k}} \rightarrow \mathbf{\Psi}^{i}_{h_k}\big|_{U^{j}{}^{H_k}_{h_k}}$
1. Compute $\mathbf{R}^{i}\left(\mathbf{U}^{j}{}^{H_k}_{h_k}\right)$
2 AWR: $\epsilon = \int \mathbf{\Psi}^{i}_{h_k}\mathbf{R}^{i}\left(\mathbf{U}^{j}{}^{H_k}_{h_k}\right)$
3. Localize ϵ
4. Adaptation $\rightarrow H_{k+1}$
$(k+1)_{\text{th}}$ adaptation
fine space h
coarse space H

Residual network that you just implemented!

Kaihua Ding, et al, 2021

How can you extract meanings (mine)?

- Come up with you own algorithms?
  - advantage: highly tailored and potentially powerful
  - disadvantage: time consuming, and error-prone for non-experts
- Use packages
  - advantage: fast prototyping
  - disadvantage: packages accuracy might surprise you (sanity check)

Thank you!

A brief self-assessment with 10 questions (< 5 minutes to finish)

dingkaihua@uchicago.edu