

Technology Science Information Networks Computing



Lecturer: Ting Wang (王挺)

利物浦大学计算机博士

清华大学计算机博士后

电子信息技术高级工程师

上海外国语大学网络与新媒体副教授

浙江清华长三角研究院海纳认知与智能研究中心主任

Intelligent Information Processing and Data Annotation

Ting WANG



School of Journalism and Communication
Shanghai International Studies University



Haina Cognition and Intelligence Research Center
Yangtze Delta Region Institute of Tsinghua University, Zhejiang

Contents

1. An overview.
2. What is data annotation?
3. Why data annotation?
4. How to annotate data?





An overview

brief introduction to AI in China

An Overview

Artificial Intelligence was born in Dartmouth College, USA, 1956

1956 Dartmouth Conference: The Founding Fathers of AI



John MacCarthy



Marvin Minsky



Claude Shannon



Ray Solomonoff



Alan Newell



Herbert Simon



Arthur Samuel



Oliver Selfridge



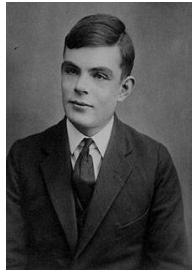
Nathaniel Rochester



Trenchard More

An Overview

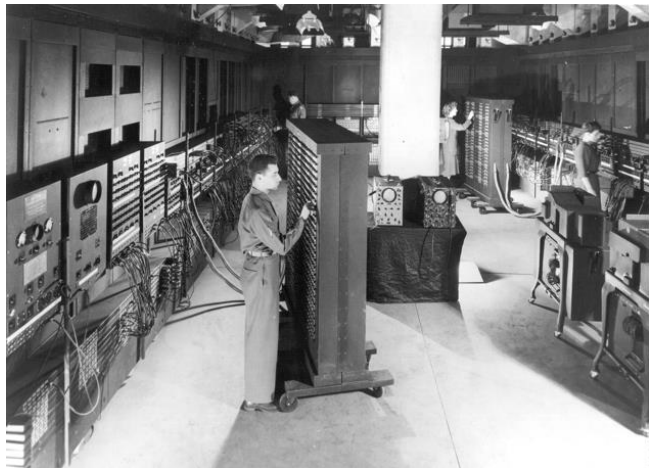
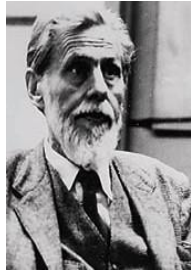
Three stages of AI



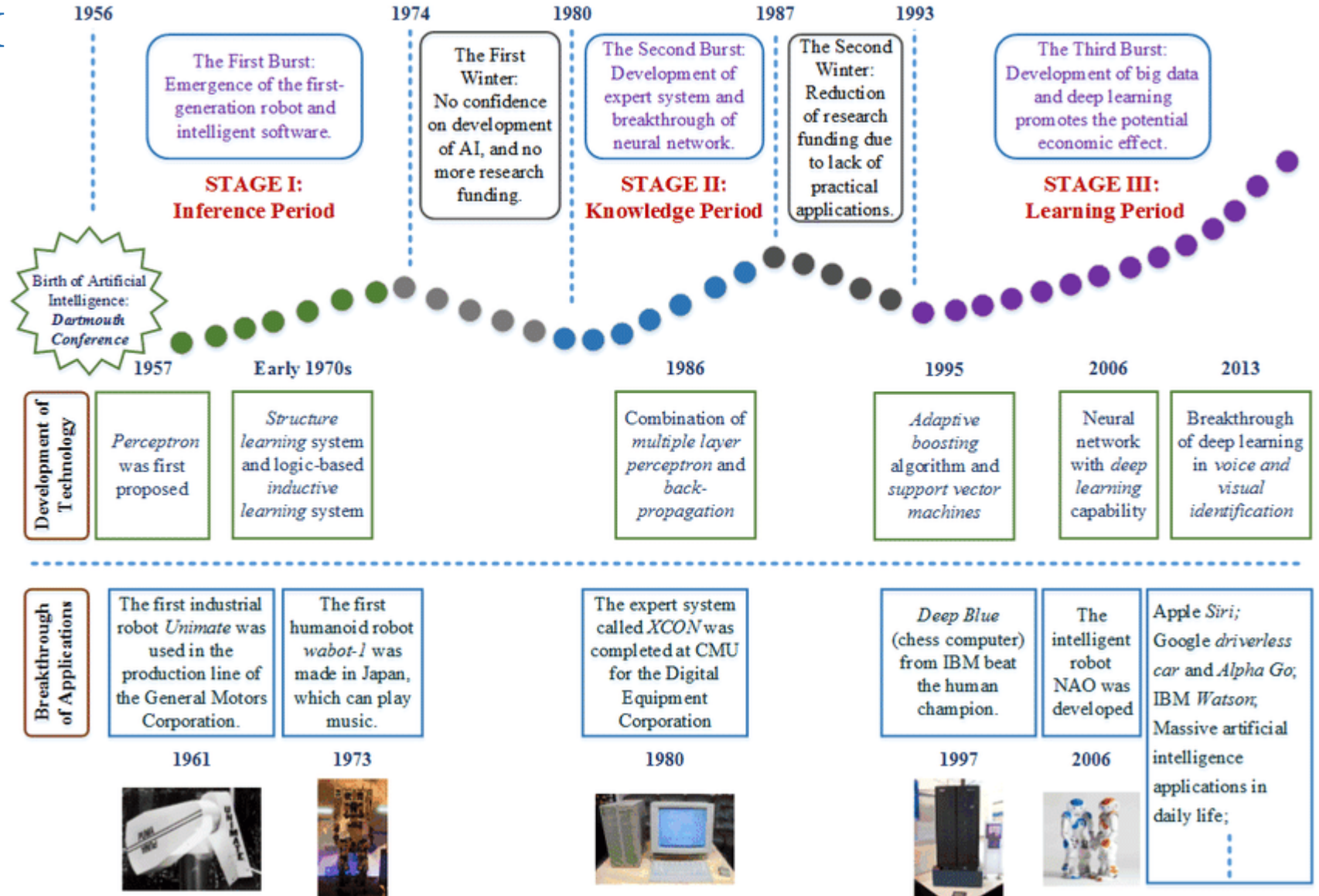
Alan Turing
Turing Machine
1936



Warren McCulloch, Walter Pitts
Artificial Neuron
1943

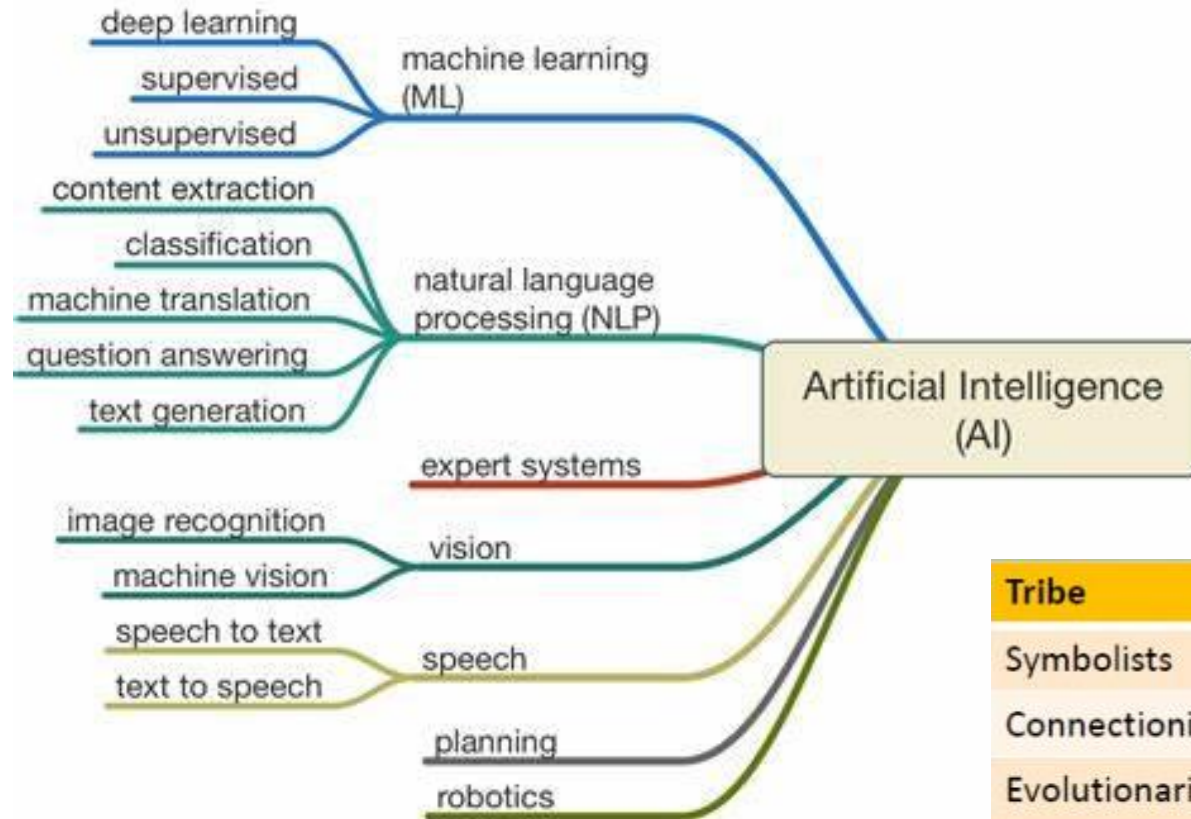


ENIAC
the first computer in the world, 1946



An Overview

Categories of AI

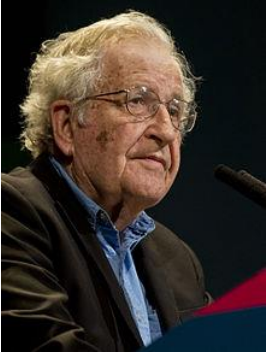


Prof. Pedro Domingos
University of Washington

2015, ACM

Tribe	Origins	Master Algorithm
Symbolists	Logic, philosophy	Inverse deduction
Connectionists	Neuroscience	Backpropagation
Evolutionaries	Evolutionary biology	Genetic programming
Bayesians	Statistics	Probabilistic inference
Analogizers	Psychology	Kernel machines

Why data annotation?



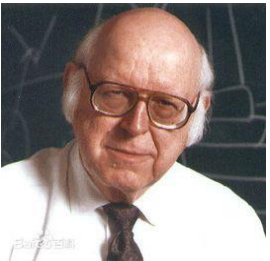
Avram Noam Chomsky

符号主义 1957
Symbolism



Herbert Simon

1. Plato is a man.
2. Man will die.
3. Plato will die.



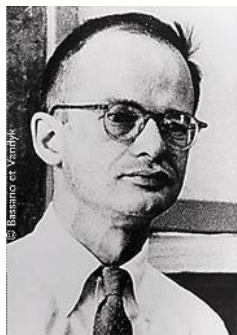
Allen Newell

Expert System
Universal Grammar and Chomsky Hierarchy

Why data annotation?



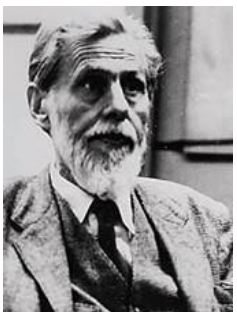
Donald Olding Hebb



Warren McCulloch



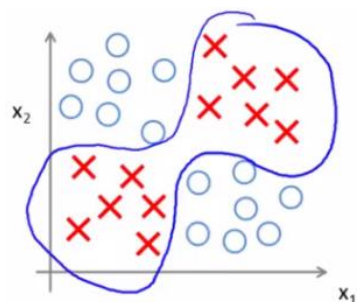
Frank Rosenblatt



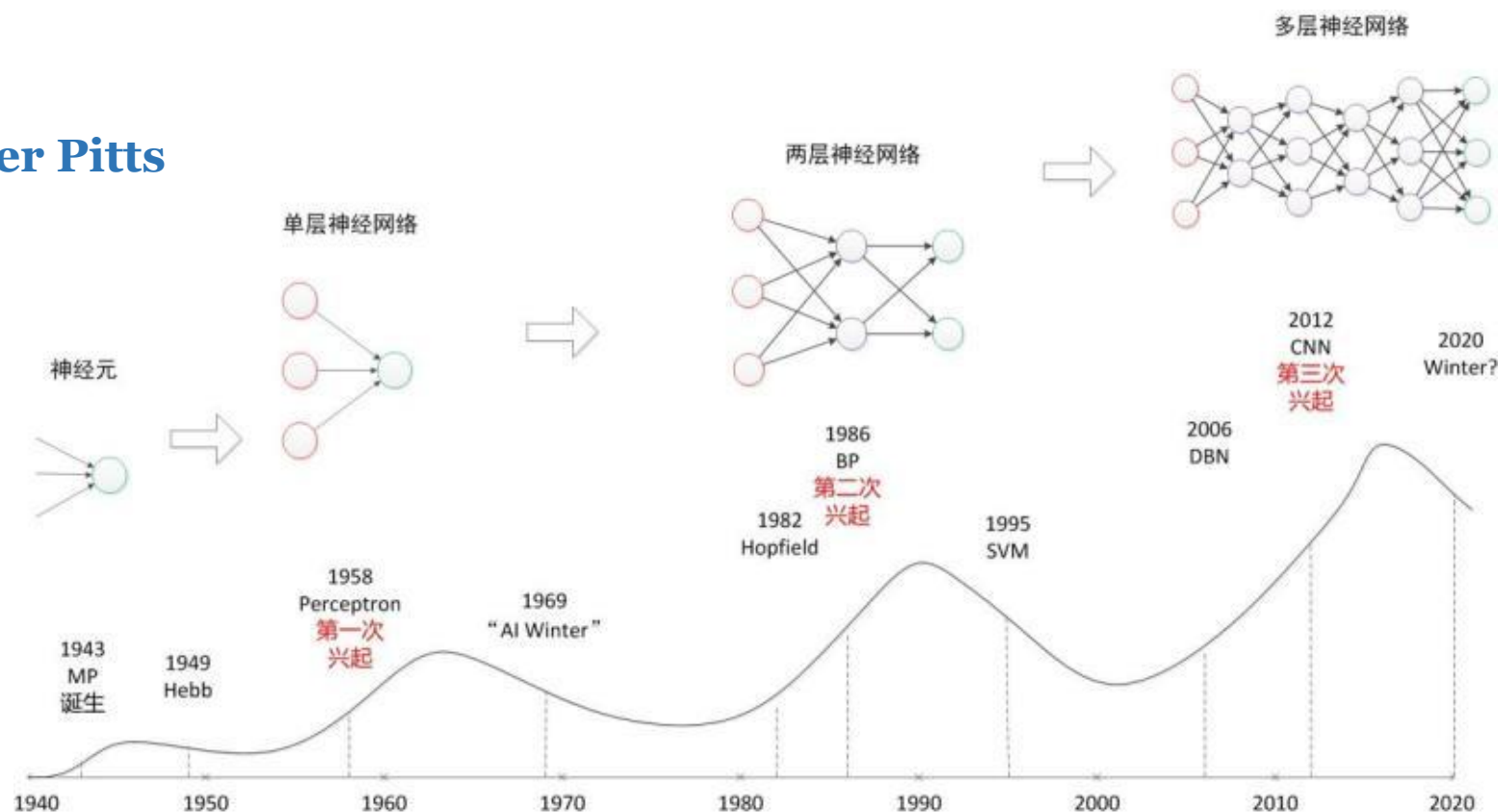
Walter Pitts



Marvin Lee Minsky



联结主义 1943 connectionism



Why data annotation?

进化主义 1970's *Evolutionism*



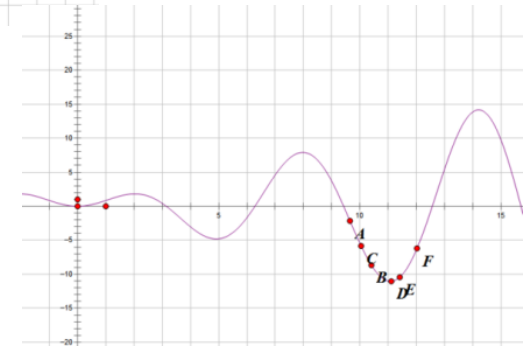
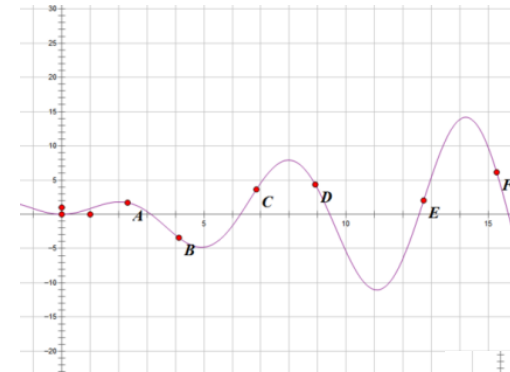
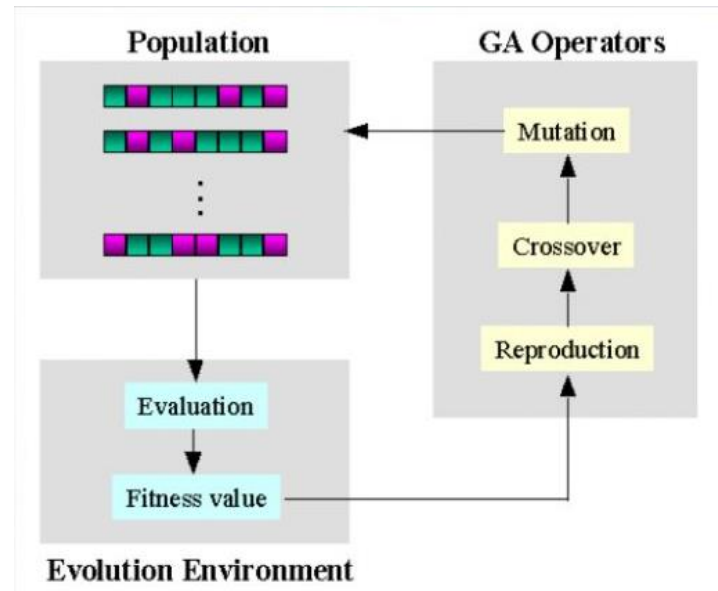
John Henry Holland

Genetic Algorithm
Optimization

Particle Swarm



Yuhui Shi



Why data annotation?

贝叶斯主义 1763 *Bayesianism*



Judea Pearl

Likelihood How probable is the evidence given that our hypothesis is true?	Prior How probable was our hypothesis before observing the evidence?
$P(H e) = \frac{P(e H) P(H)}{P(e)}$	
Posterior How probable is our hypothesis given the observed evidence? (Not directly computable)	Marginal How probable is the new evidence under all possible hypotheses? $P(e) = \sum P(e H_i) P(H_i)$

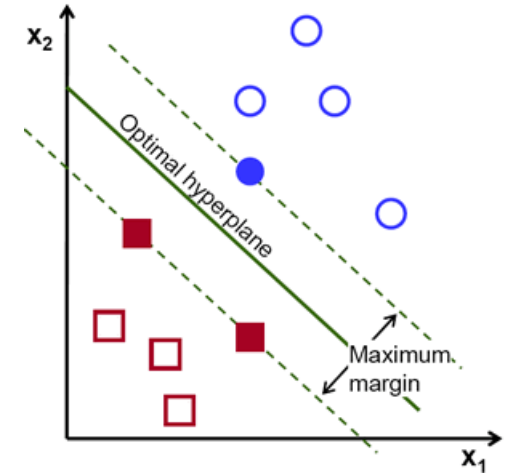
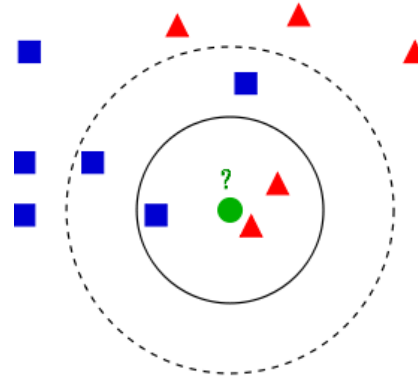
Why data annotation?

类推主义 1951 *Analogism*



Vladimir Vapnik

K-Nearest Neighbour
Support Vector Machine



An Overview

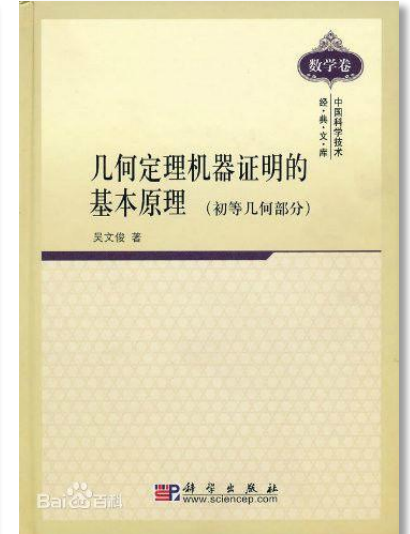
Artificial Intelligence started in China in 1978

The Highest Award on AI in China:

Wu Wen Jun AI Science & Technology Award



Ref: <http://www.wuwenjunkejijiang.cn/wj/index.aspx>



Wu Wenjun (Chinese: 吴文俊; 12 May 1919 – 7 May 2017), also commonly known as **Wu Wen-tsün**, was a Chinese mathematician and academician at the Chinese Academy of Sciences (CAS), best known for the Wu's method of characteristic set.

An Overview

AI History in China:

1978, Wu Wenjun was awarded by China Government.

1980, Some students were sent to Japan, Europe and US to learn AI.

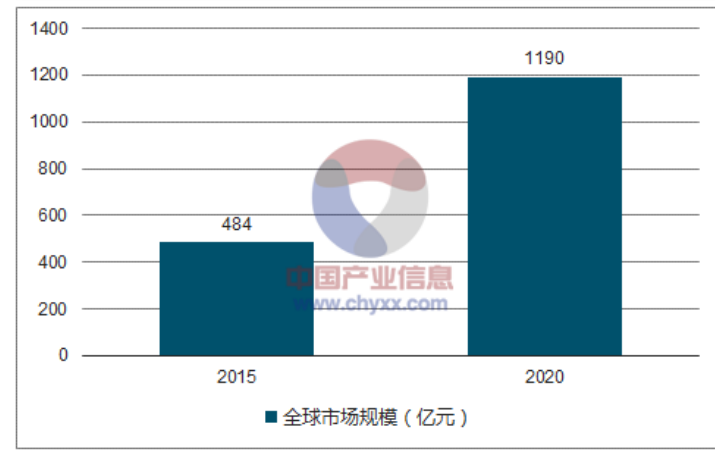
1981, Chinese Association for Artificial Intelligence, CAAI, established.

1986, Chinese National 863 Program started.

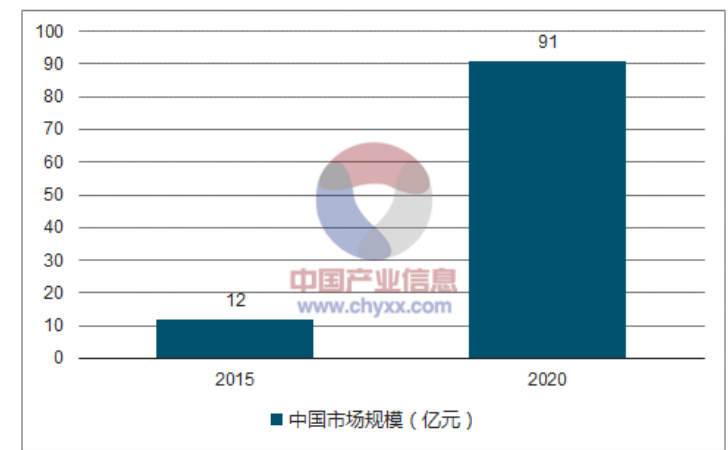
1987, First text book on AI was published by Tsinghua University.

2017, AI was upgraded to the Chinese National Strategy.

Market Size: (Year 2020)



World Market Size (100 million)

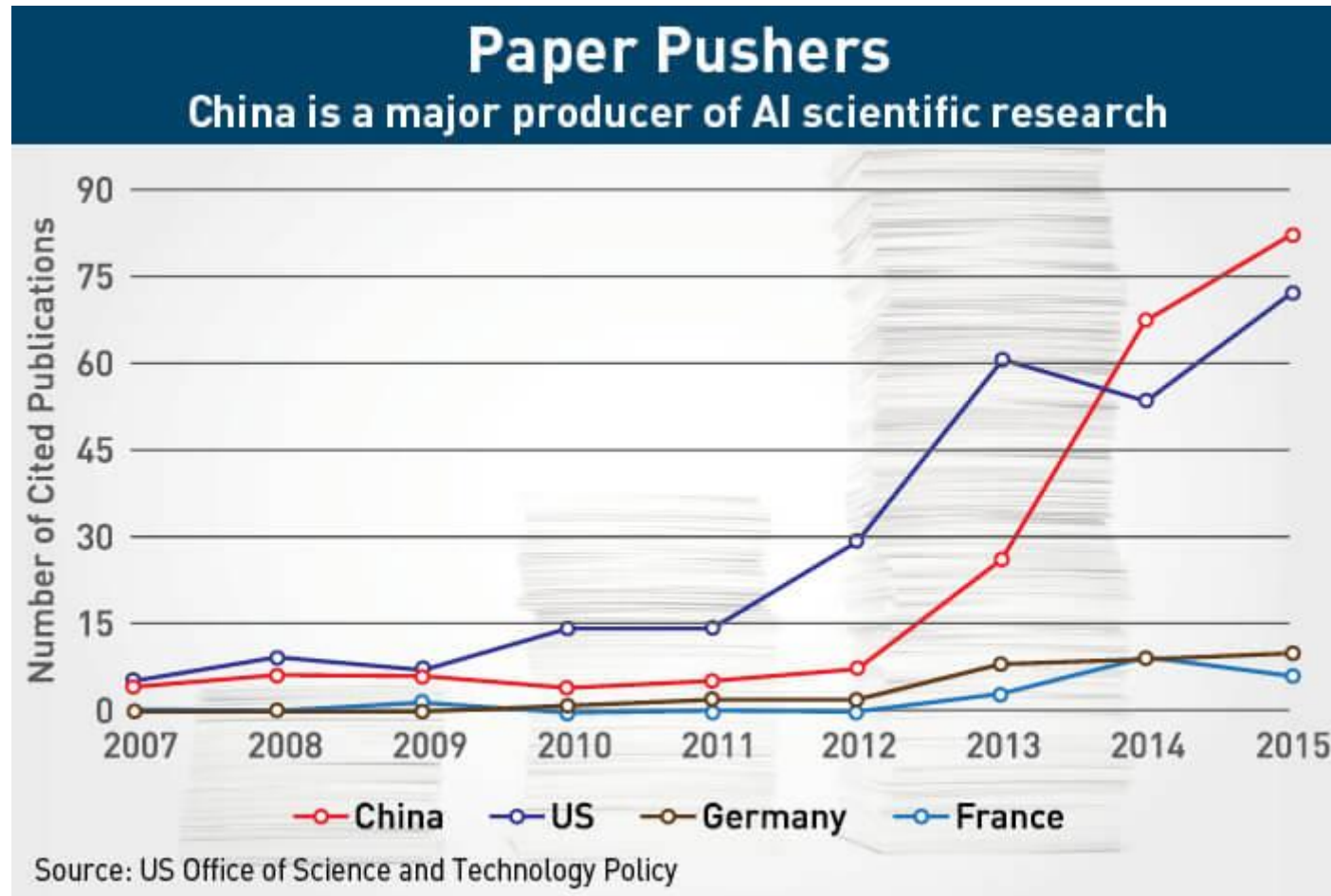


China Market Size (100 million)

Ref: <http://www.chyxx.com/industry/201803/619321.html>

An Overview

Is China now a leading country in AI?



Ref: <http://knowledge.ckgsb.edu.cn/2017/07/17/technology/ai-in-china-bringing-ai-real-world/>

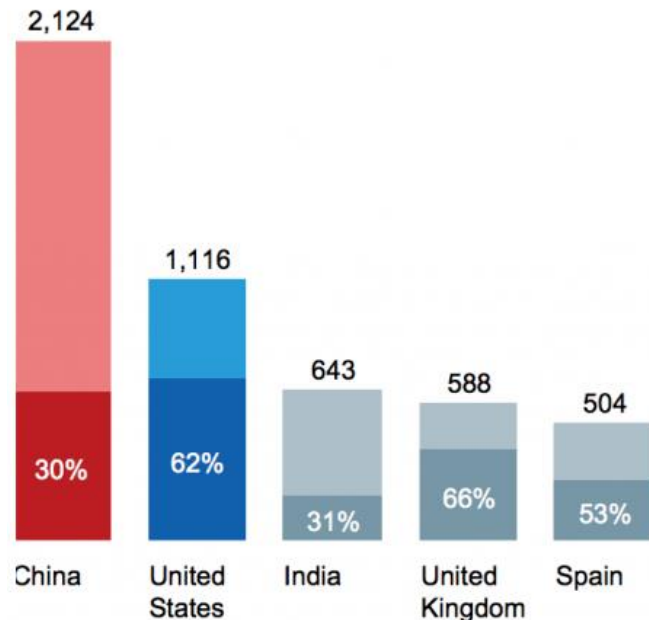
Artificial Intelligence in China: An Overview

China has less influence in AI research.
(data: 2017)

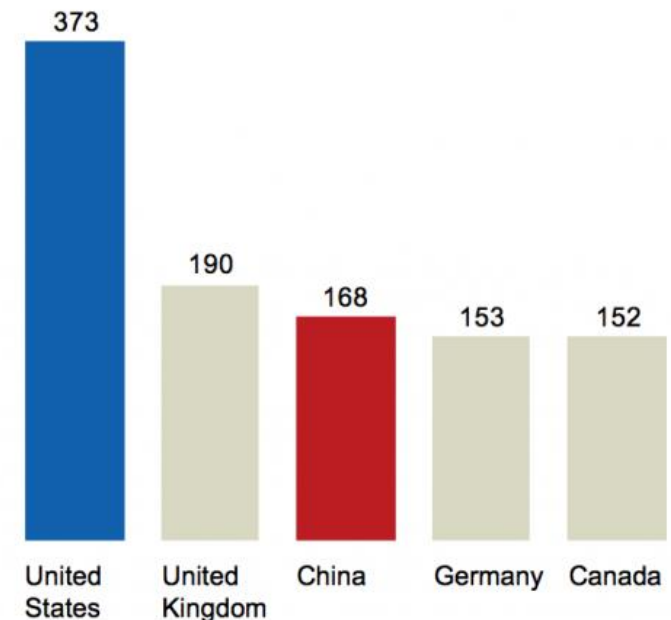
Although China produces a large number of widely cited AI-related papers, US and UK research remains more influential

While China ranks first for absolute AI citations, the United States holds an edge when self-citations are taken out
Number of AI publications cited

Self-citations¹
Other citations



Publication influence
H-index²



An Overview

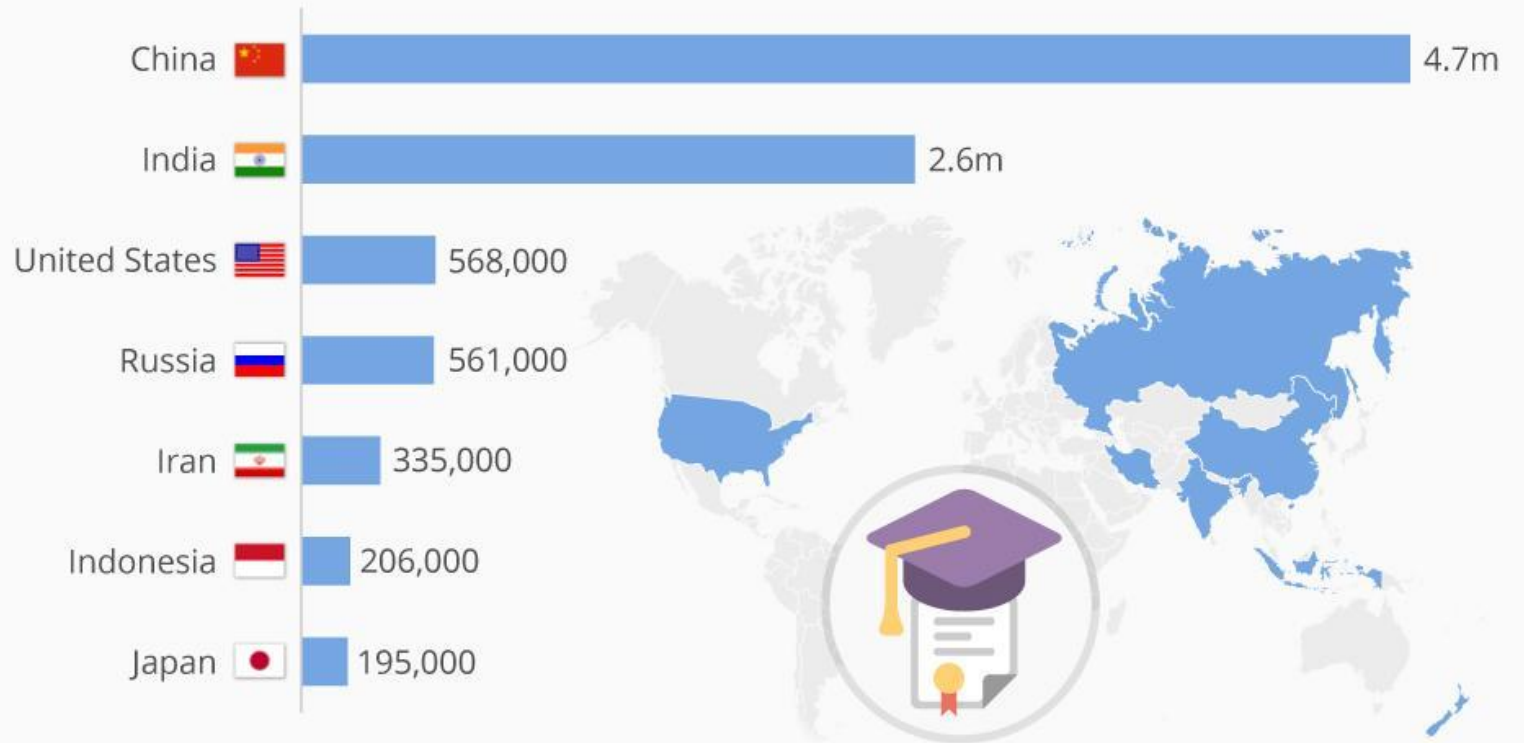
STEM Graduates

The US currently has 850,000 AI technical people while China has about 50,000. There are 70,000 overseas Chinese AI technical talents working in the US and China is lobbying to win them back.

By the end of 2018, more than 33 top universities has opened AI faculties.

The Countries With The Most STEM Graduates

Recent graduates in Science, Technology, Engineering & Mathematics (2016)

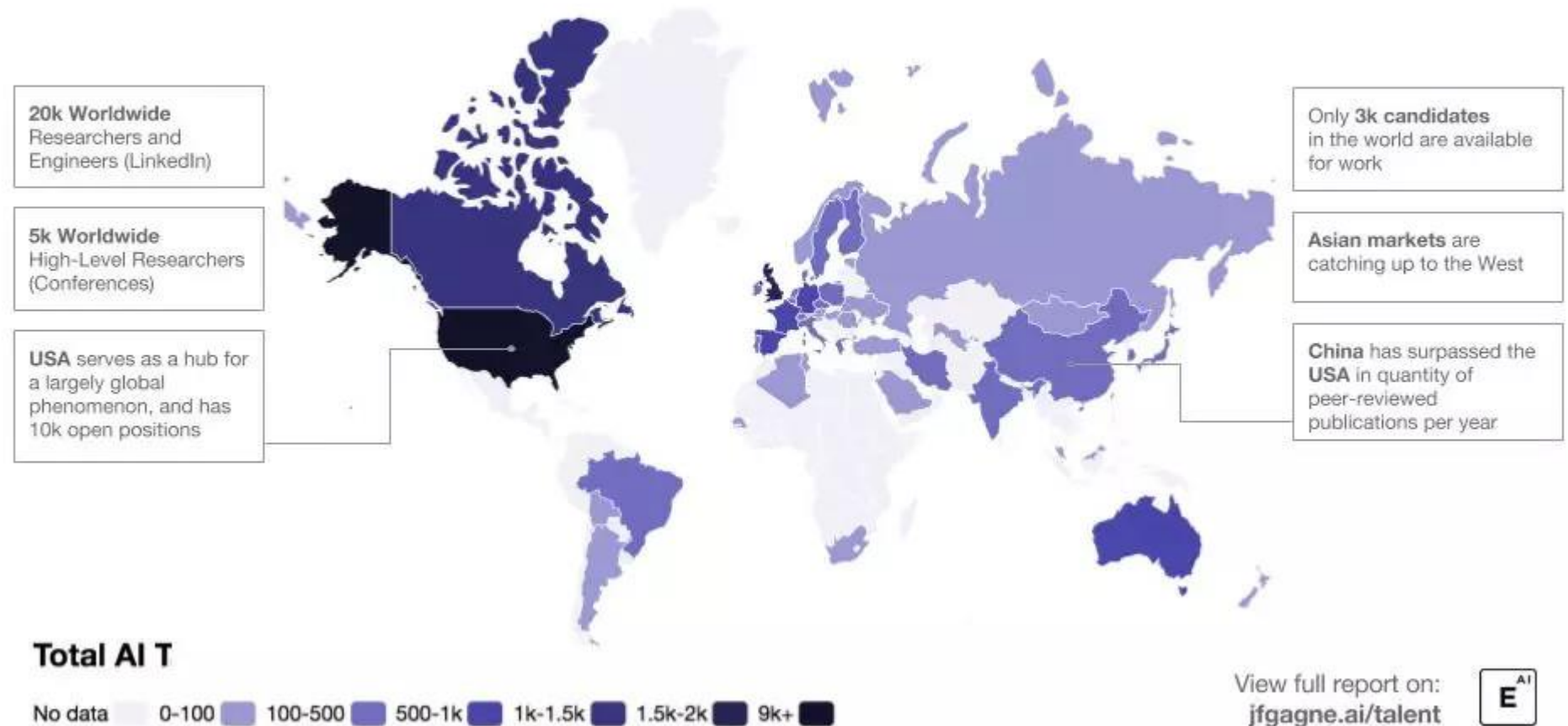


@StatistaCharts Source: World Economic Forum

Forbes statista

An Overview

Global AI Talent Pool Heat Map

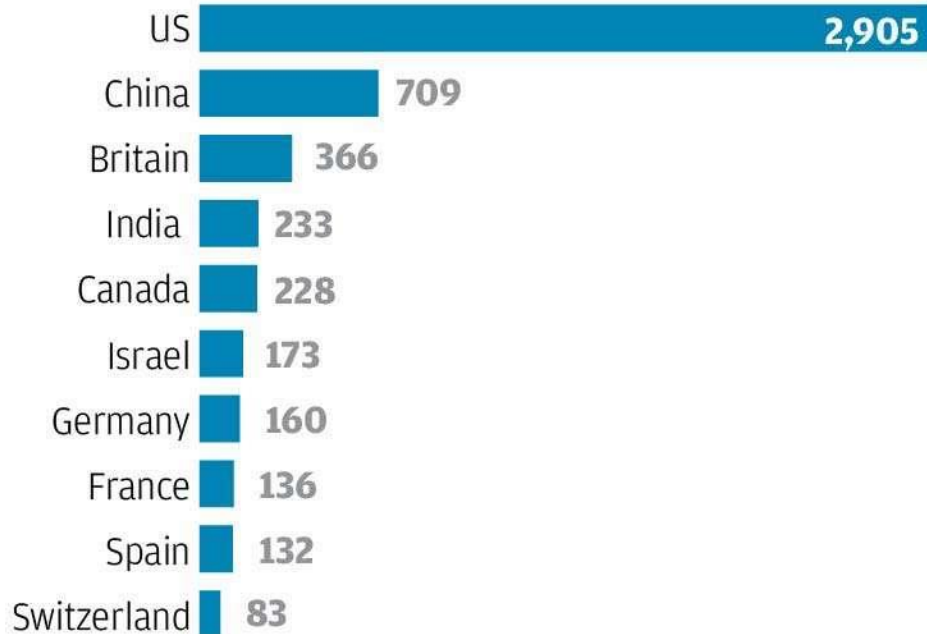


Ref: <https://www.elementai.com/news/2018/the-global-ai-talent-pool-going-into-2018>

An Overview

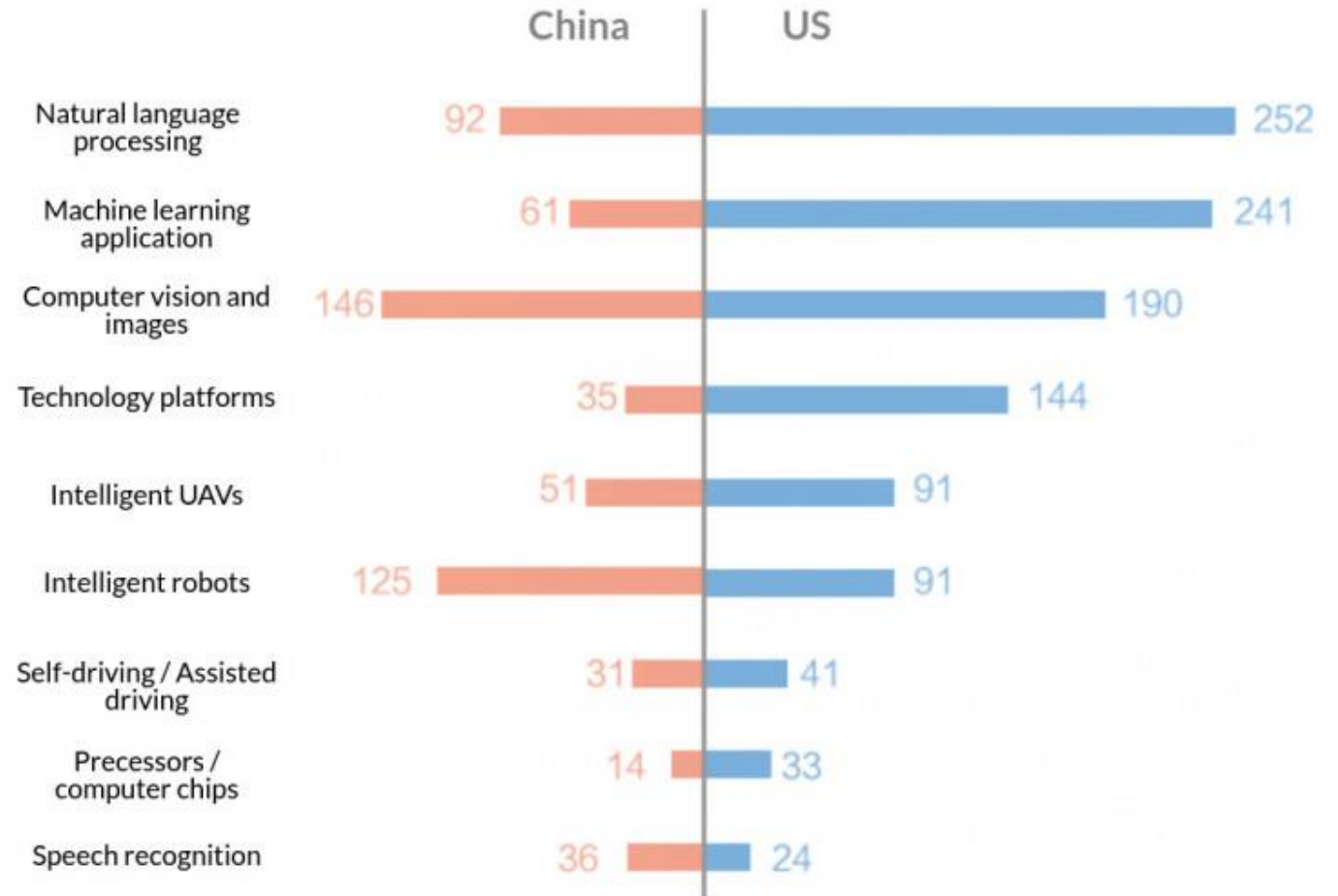
AI Companies

Total number of artificial intelligence companies



Source: Wuzhen Institute

SCMP



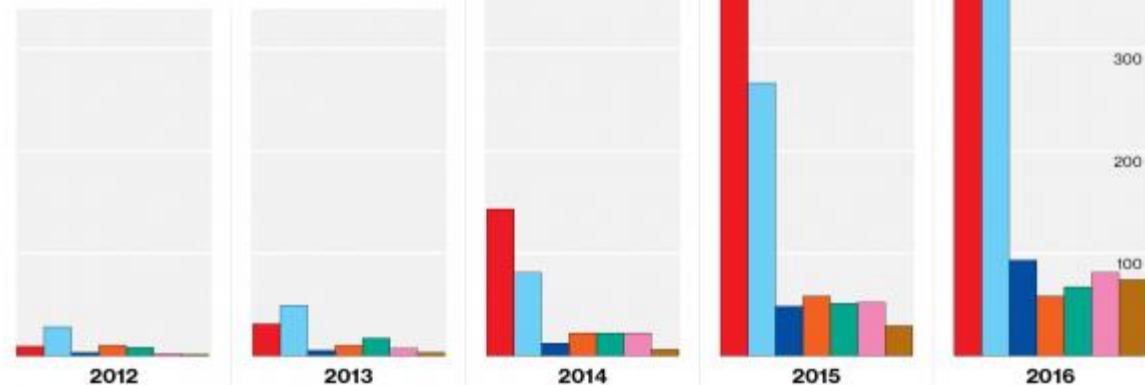
(Numbers indicate the number of enterprises)

Who Is Winning the AI Race?

China and the United States dominate the world of artificial-intelligence research. Microsoft, IBM, and Google are the leading companies.

China Learns Quickly

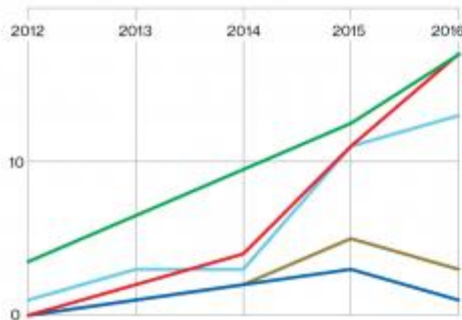
Since 2014 China has published the most research papers per year on deep learning, an advanced form of artificial intelligence.



The Big Three

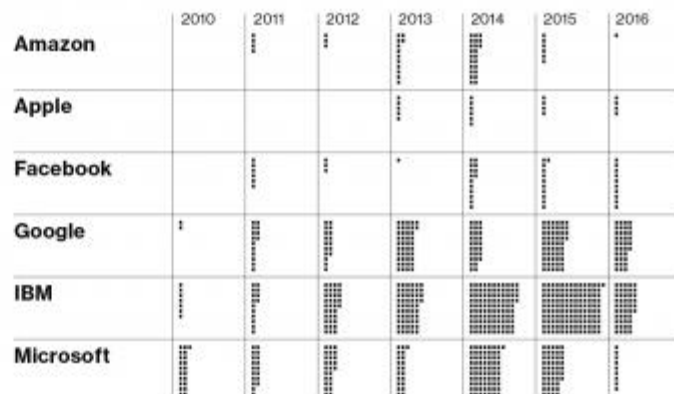
For years, Microsoft published the most deep-learning research papers, but Google and IBM have gained ground.

Microsoft | Google | IBM | Facebook | Baldu



The Fight for IP

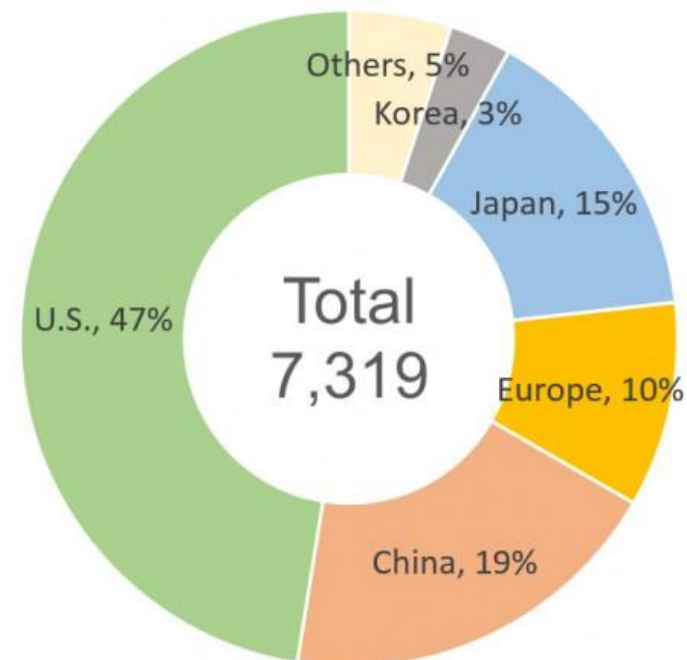
IBM has dominated U.S. patent activity in AI, but Google and Facebook may be closing the gap. Since the process involves a time lag before applications are published, records from 2014 onward are probably not complete.



An Overview

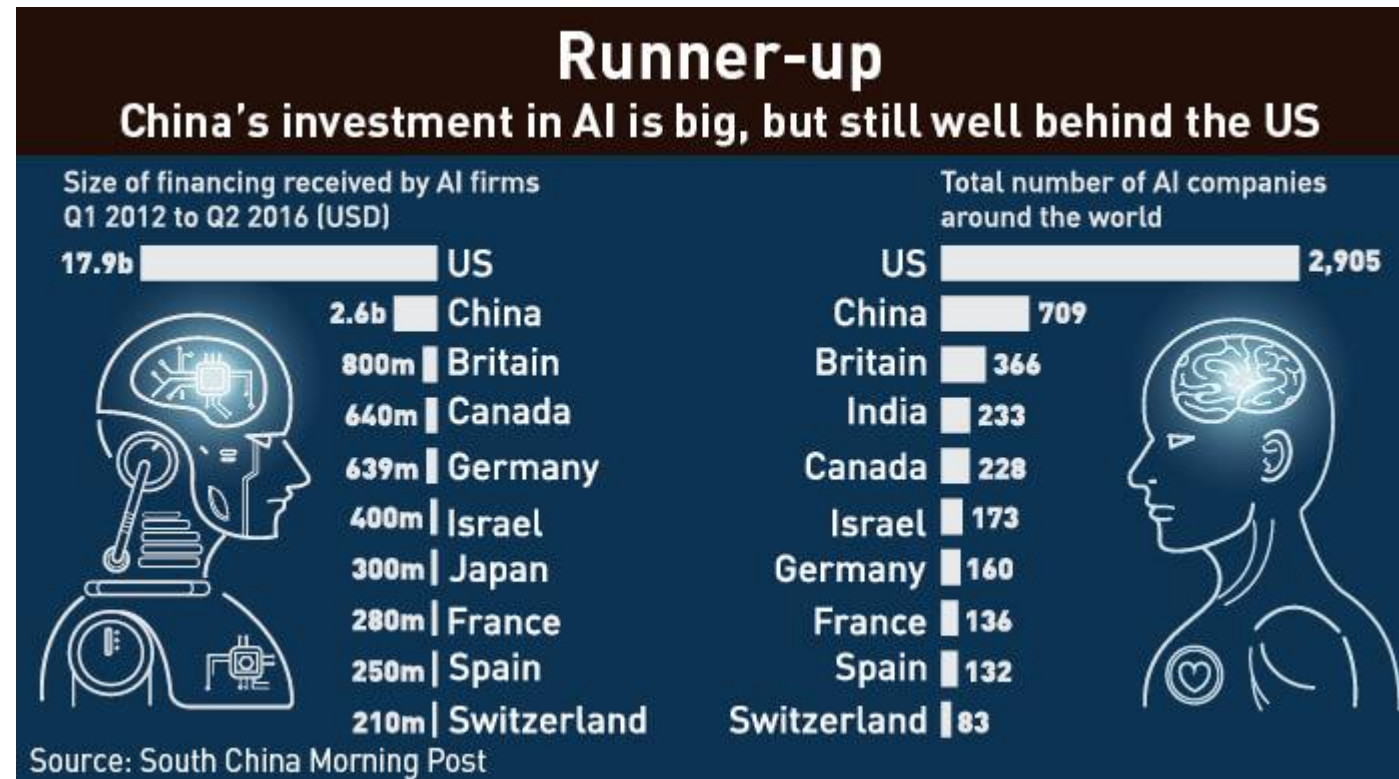
Patent

Japan Lags behind the U.S. and China in AI
– Patent application filings related to AI technologies –



An Overview

China's investment in AI



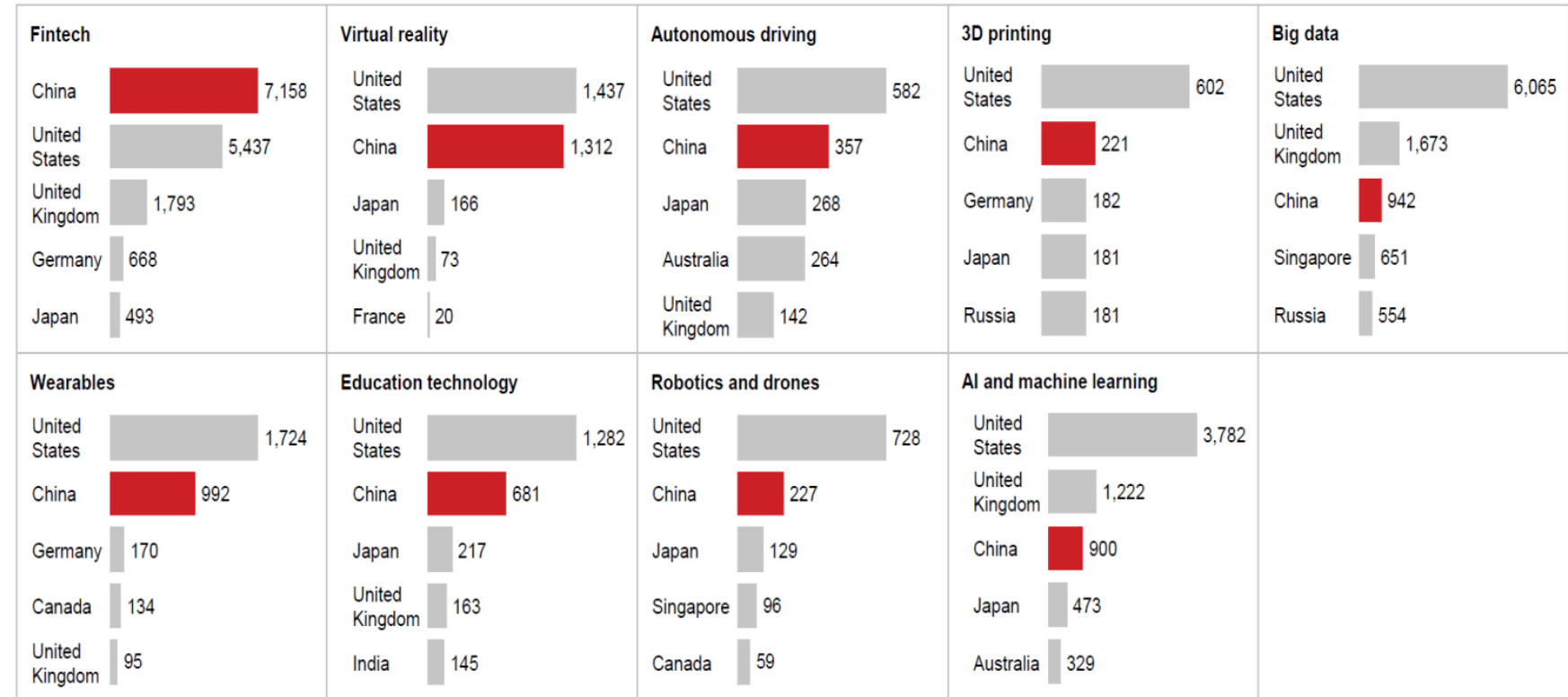
Ref: <http://knowledge.ckgsb.edu.cn/2017/07/17/technology/ai-in-china-bringing-ai-real-world/>

An Overview

Investment

China in Global Top Three for Venture Capital Investment in Key Technologies

Venture capital investment in leading technologies, 2016
US\$ million



Problems and Challenges of AI in China

1. Lacks of Core Technologies
2. Quick Result Investment
3. Conflicts brought by Unbalanced Development
4. Great Challenges in Legislation and Ethic
5. High Housing Price and Hukou Prevent STEM Graduates to 1-Tier Cities

New Hopes in China: the Future

1. Students prefer to select programs correlated to AI.
2. Many children start their Lego training when they are 3 years old, and start to learn computer programming when they are 6 years old.
3. Hundreds of education companies provide training programs on coding, robotics and algorithms.
4. Scratch is now a compulsory course in many primary schools, especially in the east coast of China.

New Job in AI

AI Trainer
Parameter Adjusting
Data Annotation

2022年人社部新职业“人工智能训练师”相关从业人员有望达500万

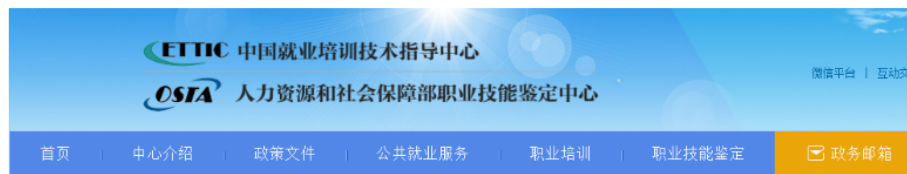


作者：人工智能培训

2020-01-03

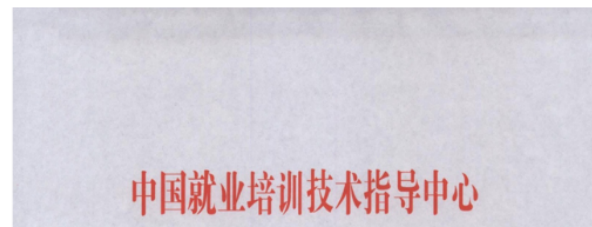
导读 人工智能训练师这一新职业和相关的技能标准，将有助于规范和引导人工智能应用的就业岗位，推动传统行业更好的拥抱人工智能，帮助实现社会生产力的整体跃升。

近日，中国就业培训技术指导中心发布《关于拟发布新职业信息公示的通告》，公布了16个新职位，人工智能训练师入围。



关于拟发布新职业信息公示的通告（中就培函〔2019〕67号）

稿件来源：本网 发布日期：2019-12-30



人工智能培训

专注于智能工程技术领域的多维教育，依托国际领先、具有自主知识产权的机器人核心算法和技术，将前沿技术及时转化为系统的培养方案和课程体系，旨在推动人工智能机器人的科普和专业教育。

+ 加关注

TA的热文

- 人工智能核心算法C/C++和Python哪家强？
- 教育部新增高职（专科）人工智能专业，2020
- 波士顿动力机器人视频闹剧深思：“伪人工智能
- 2020年起，发布AI造假视频要担责！B站和ZA
- 四川长宁地震预警系统立大功！但AI地震预测
- 大学报考人工智能专业应具备哪些条件？

http://www.qianjia.com/zhike/html/2020-01/3_18670.html

Requirements form Big Data

2016-2025年全球数据量的爆发式增长



- 人工智能模型以处理非结构化数据见长，但数据经过清洗与标注才能被唤醒价值
- 我国每年需要进行标注的语音数据超过**200万**小时，图片则达**数亿**张

来源：柱状图数据来自IDC，文字来自艾瑞自主研究。

An Overview

Data annotation industry in China

有多少智能，就有多少人工。

市场规模：300亿元

百度众测 京东众智

龙猫数据 Testin云测

倍赛BasicFinder

数据堂

小作坊

小镇青年是新职业主力军

50%新职业岗位来自三四五线城市



2018 年，有约 34% 的业务量流向专业做数据采标的第三方公司。

在距离贵阳市中心50公里的百鸟河数字小镇，有一个规模500人的“数据工场”，500名标注员中，近一半是附近一家扶贫高职“盛华职业学院”的学生。



位于贵阳的“数据工场”

他们很珍惜这个接近“白领”的兼职机会，1月能挣到1500元，经济上足以自立，省点还可以补贴家用，相比其他兼职选择：去餐厅辛苦端盘子或顶着风雨送外卖，数据标注相对轻松且体面。

<https://www.huxiu.com/article/233240.html>

<https://finance.sina.cn/stock/relnews/hk/2020-03-03/detail-iimxyqvz7570296.d.html>

Conclusions:

1. China is a leading country in AI;
2. Many jobs will be changed to new jobs by AI ;
3. Data annotation is a necessity to AI, also will be a large-scale industry in rural area.



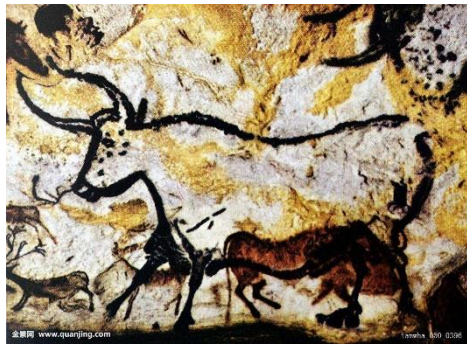
What is data annotation?

a definition to data annotation

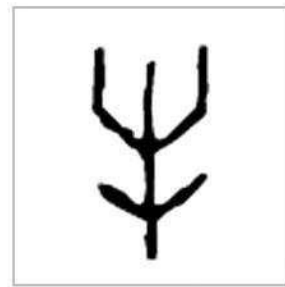
What is data annotation?

History of Data Annotation

Data Annotation is not new, which has a history as long as that of human beings.

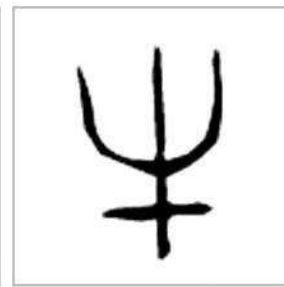


36,000 years ago



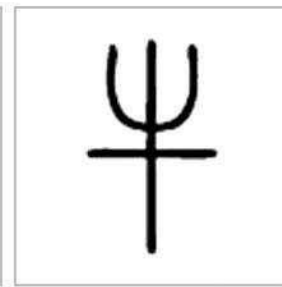
甲骨文

1500B.C.



金文

1000B.C



小篆

200B.C.



楷体

200A.D.-Today

MEANING		OUTLINE CHARACTER, B. C. 3500	ARCHAIC CUNEIFORM, B. C. 2500	ASSYRIAN, B. C. 700	LATE BABYLONIAN, B. C. 500
1.	The sun				
2.	God, heaven				
3.	Mountain				
4.	Man				
5.	Ox				
6.	Fish				

Persian Cuneiform

A	B	C	D	E	F	G	H	I	J	K	L	M
N	O	P	Q	R	S	T	U	V	W	X	Y	Z
a	b	c	d	e	f	g	h	i	j	k	l	m
n	o	p	q	r	s	t	u	v	w	x	y	z

What is data annotation?

Traditional Data Annotation

Dictionary is a traditional data annotation product.

The Essential of Data Annotation

- Abstraction
- Mapping

Conventional Data Annotation

- Conclusion
- Tagging
- Comments

piáu-lō 表_{ㄅㄧㄠˋ}露_{ㄌㄨˋ} make plain, to express,
expose

piáu-mōe (piáu-sio-mōe) 表_{ㄅㄧㄠˋ}妹_{ㄇㄨㄟˊ} daughter of
father's sister, of mother's brother or
sister, who is younger than oneself

piáu-pék 表_{ㄅㄧㄠˋ}白_{ㄅㄞˊ} express or state clearly,
explain, clear up, defend, clarify

piáu-phōe 表_{ㄅㄧㄠˋ}皮_{ㄆㄧˊ} epidermis, the cuticle (of
plants)

piáu-sī 表_{ㄅㄧㄠˋ}示_{ㄕㄨˋ} express, show, indicate, super-
scription, signify, expression

piáu-sī bóan-ì 表_{ㄅㄧㄠˋ}示_{ㄕㄨˋ}滿_{ㄇㄢˋ}意_{ㄧˋ} express or indi-
cate satisfaction

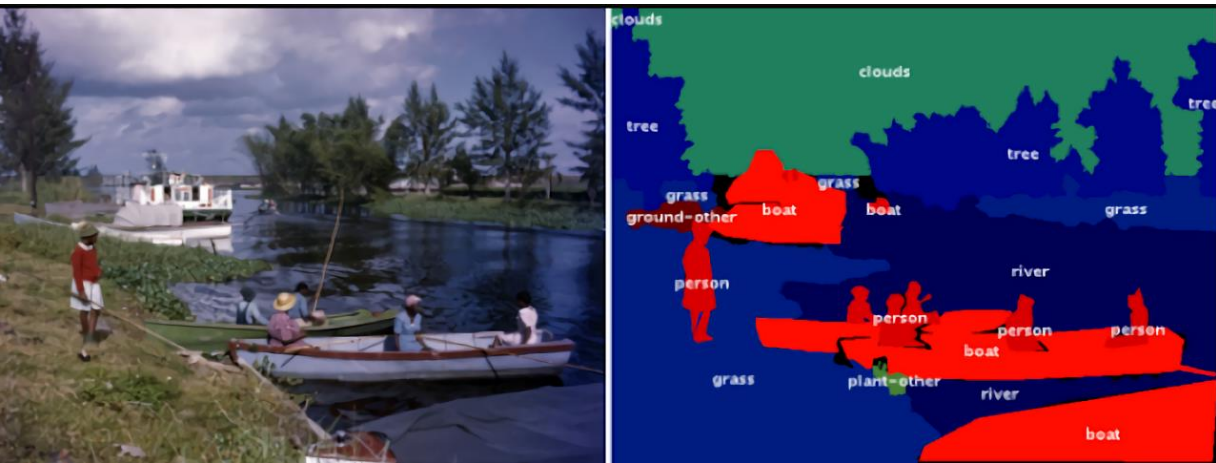
What is data annotation?

Data annotation is the task of labelling any type of data : images, audio, text, video, Generally, it is done by selecting a “zone” of the data, and adding a label to this specific zone.

From: <http://www.quora.com/What-is-data-annotation>

What is data annotation?

Different Types of Data Annotation



```
<acknowledgements>
  <acknowledgement PMCID=5567>
    <content>We thank G. Meissner for the generous gift of anti-RyR2 antibody C3-33 and Y. Chen for providing the photomicrograph used in Figure 2.
    </content>
    <annotations>
      <annotation label="person,donor">G. Meissner</annotation>
      <annotation label="antibody name">anti-RyR2</annotation>
      <annotation label="antibody name">C3-33</annotation>
      <annotation label="antibody">antibody</annotation>
      <annotation label="person">Y. Chen</annotation>
    </annotations>
  </acknowledgement>
</acknowledgements>
```

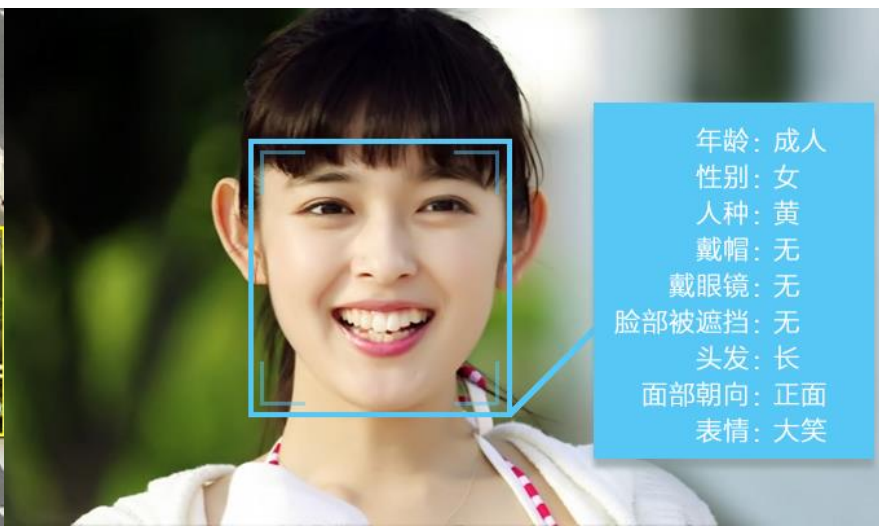
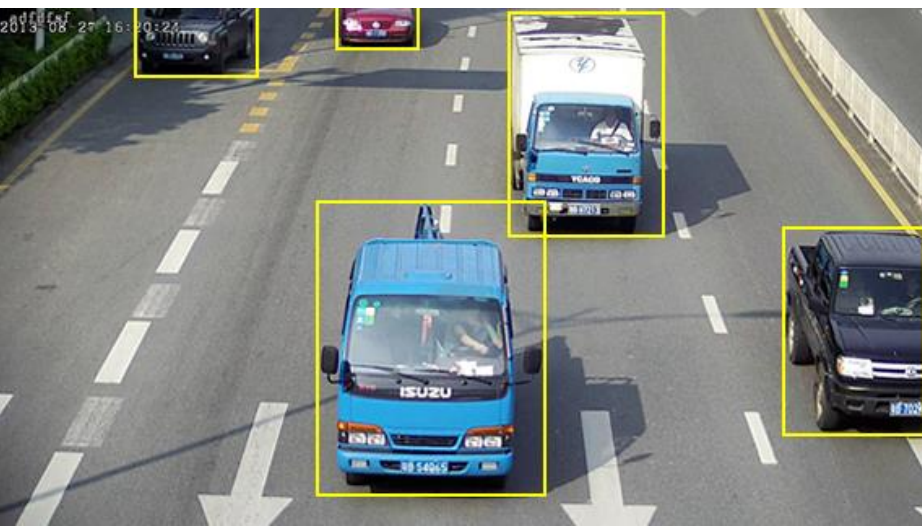
Area

Box

Classification(Comments)

Dot

Textual Entity





Why data annotation?

the objectives of data annotation

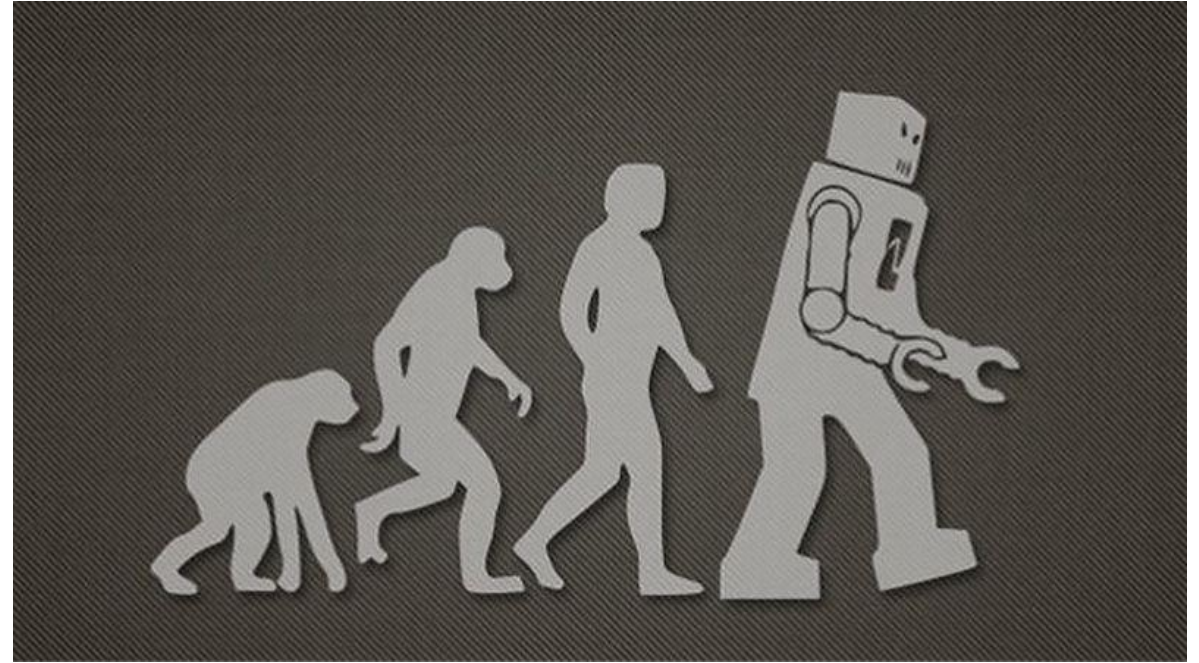
Why data annotation?

Traditional Objectives in Ancient Times

- To give comments to objects
- To add relations to semantic entities

Current Objectives

- To make machines more intelligent



Question 1: How to make machines more intelligent?

Question 2: What is intelligence?

Why data annotation?

Answer to Question 2:

There is no such a generally acceptable definition about “Intelligence”,

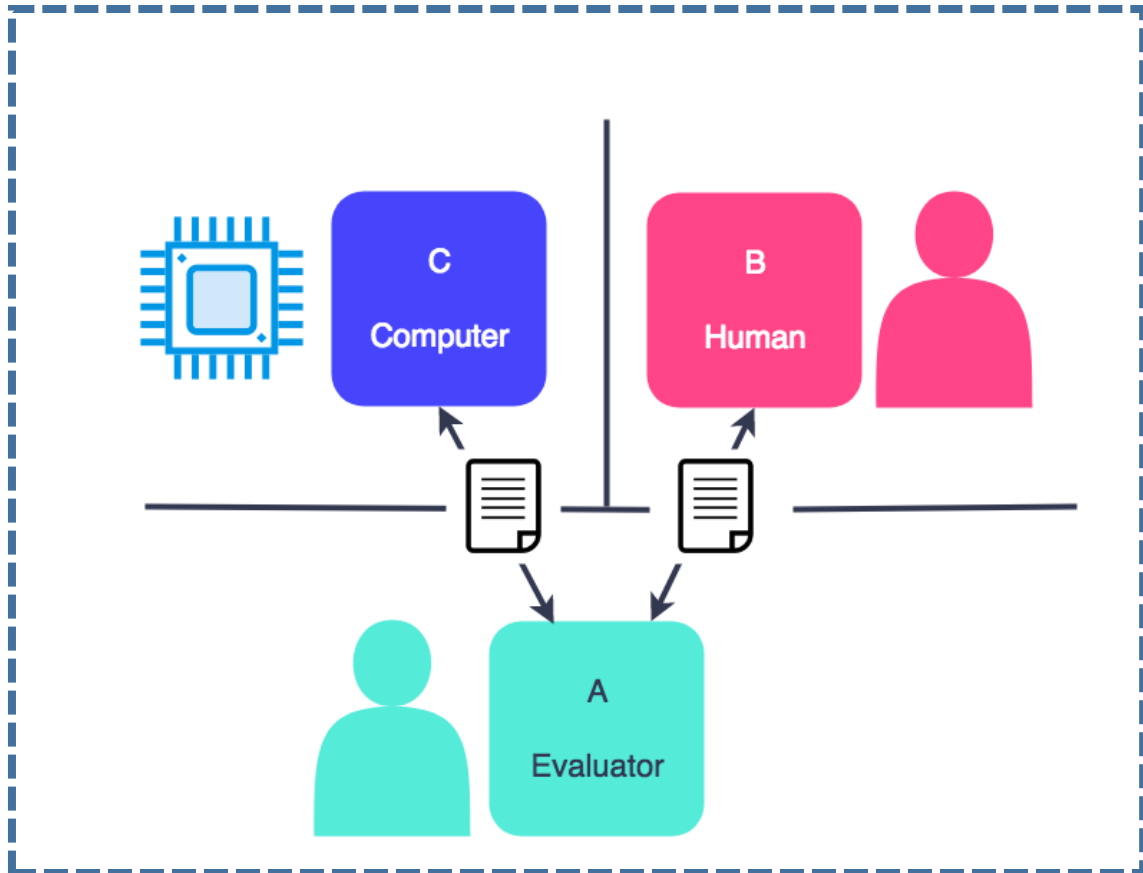
Artificial Intelligence: a modern approach to make machines more intelligent

- Artificial intelligence (AI) is the intelligence exhibited by machines or software.
[https://en.wikipedia.org/wiki/Artificial_intelligence_\(disambiguation\)](https://en.wikipedia.org/wiki/Artificial_intelligence_(disambiguation))
- Artificial intelligence (AI) makes it possible for machines to learn from experience, adjust to new inputs and perform human-like tasks.
https://www.sas.com/en_us/insights/analytics/what-is-artificial-intelligence.html
- Artificial intelligence (AI) is an area of computer science that emphasizes the creation of intelligent machines that work and react like humans.
<https://www.techopedia.com/definition/190/artificial-intelligence-ai>

Why data annotation?

The Turing Test

However, we have a generally acceptable description about “Intelligence”.



Alan Turing (1912-1954)

Alan Turing was an English mathematician, computer scientist, logician, cryptanalyst, philosopher and theoretical biologist. Turing was highly influential in the development of theoretical computer science, providing a formalisation of the concepts of algorithm and computation with the Turing machine, which can be considered a model of a general-purpose computer. Turing is widely considered to be the father of theoretical computer science and artificial intelligence.

Why data annotation?

Review: How to make a baby more intelligent?

Machine Learning



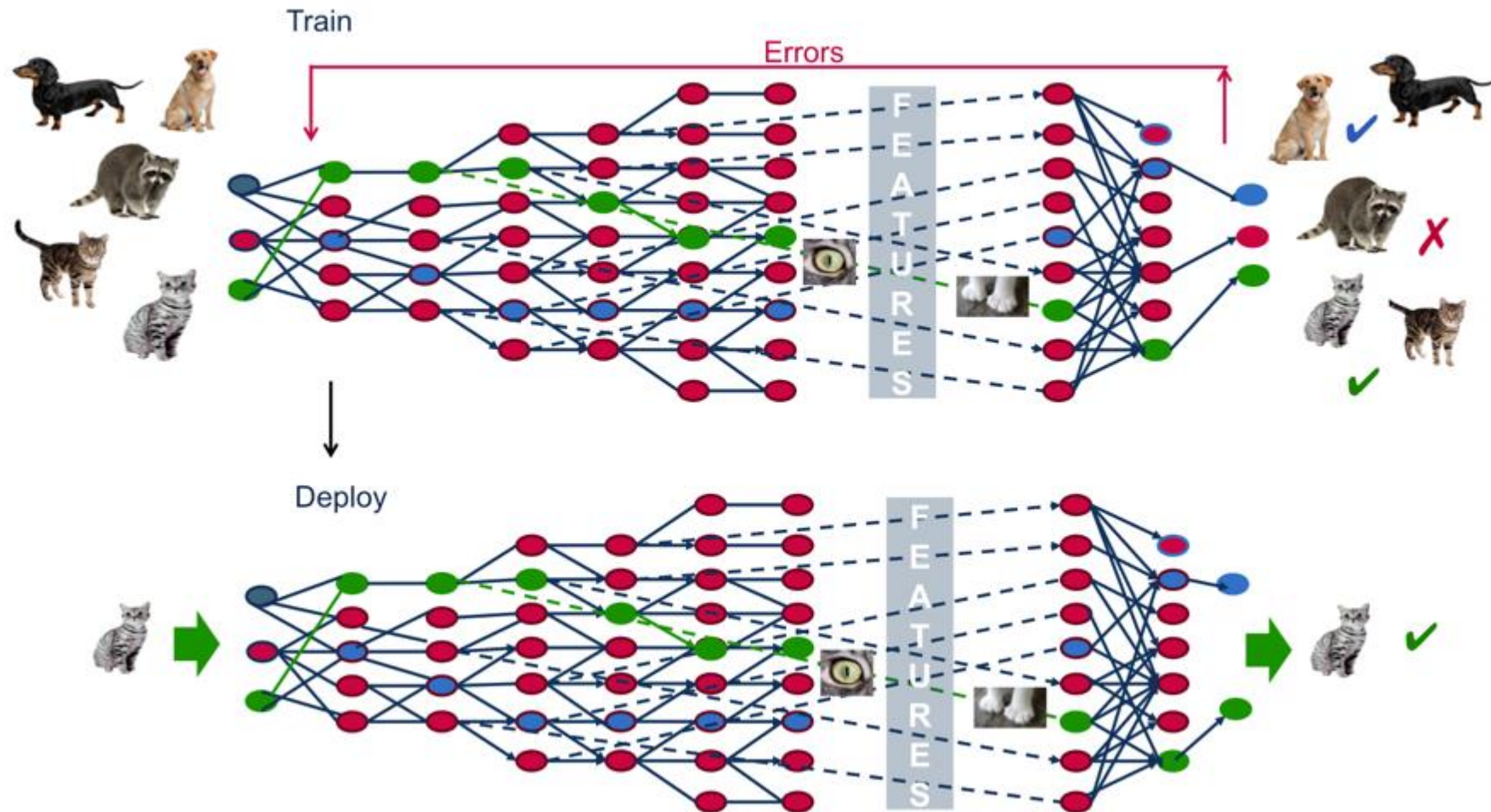
Learning by taught (Supervised Learning)

Automatic Learning (Unsupervised Learning)

Teaching the youth.

Why data annotation?

Data annotation in Deep Learning





How to annotate data?

an approach to data annotation

How to annotate data?

The Standard

1. National Standards
2. Industrial Standards
3. Enterprise Standards

ICS 35.240
L70

团 体 标 准

T/CESA 1040—2019

全国团体标准信息平台

信息技术 人工智能 面向机器学习的数据
标注规程

Information technology- Artificial intelligence- Code of practice for data annotation
of machine learning

How to annotate data?

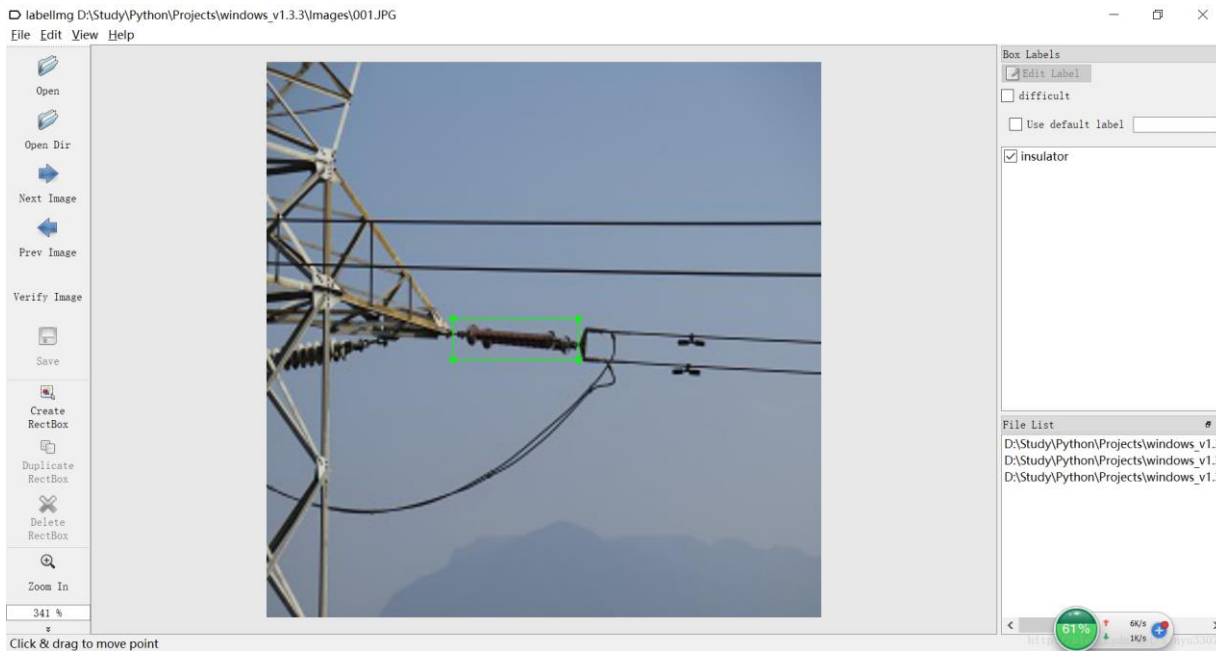
The tools for data annotation

1. image
2. audio
3. video
4. text/number



How to annotate data?

The tools for image



```
<?xml version="1.0"?>
- <annotation verified="no">
  <folder>Images</folder>
  <filename>001</filename>
  <path>D:\Study\Python\Projects\windows_v1.3.3\Images\001.JPG</path>
  - <source>
    <database>Unknown</database>
  </source>
  - <size>
    <width>256</width>
    <height>256</height>
    <depth>3</depth>
  </size>
  <segmented>0</segmented>
  - <object>
    <name>insulator</name>
    <pose>Unspecified</pose>
    <truncated>0</truncated>
    <difficult>0</difficult>
    - <bndbox>
      <xmin>86</xmin>
      <ymin>118</ymin>
      <xmax>144</xmax>
      <ymax>137</ymax>
    </bndbox>
  </object>
</annotation>
```

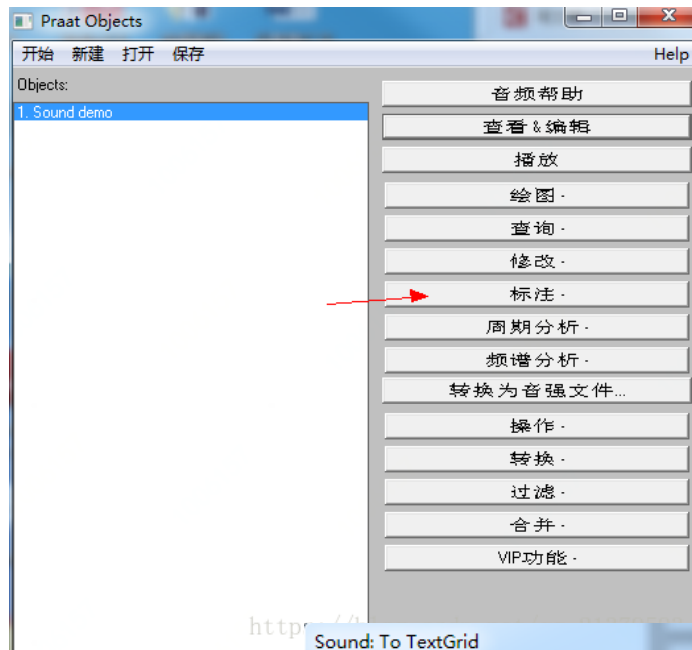
How to annotate data?

The tools for audio (1)

Praat: doing phonetics by computer

<http://www.fon.hum.uva.nl/praat/>

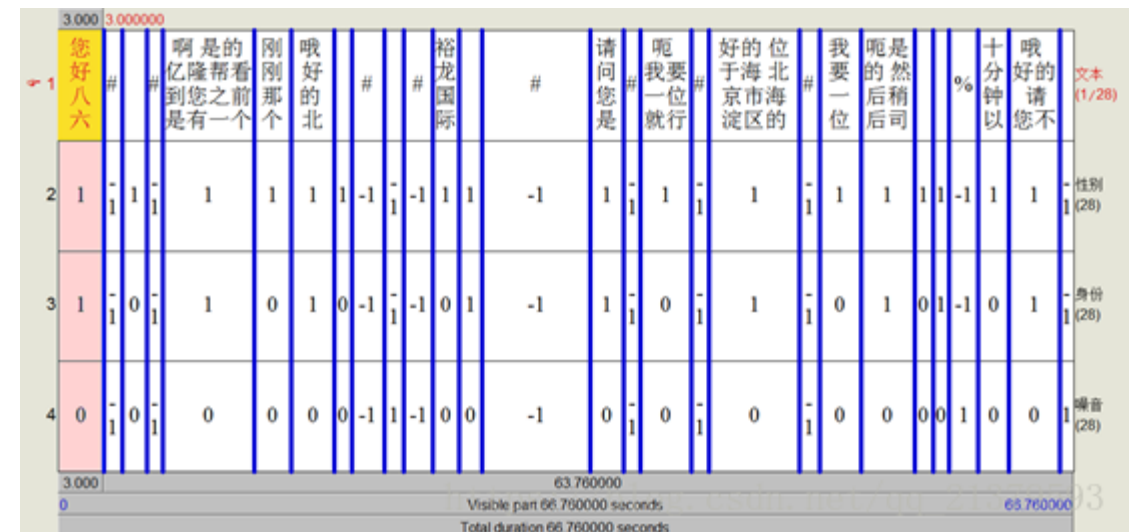
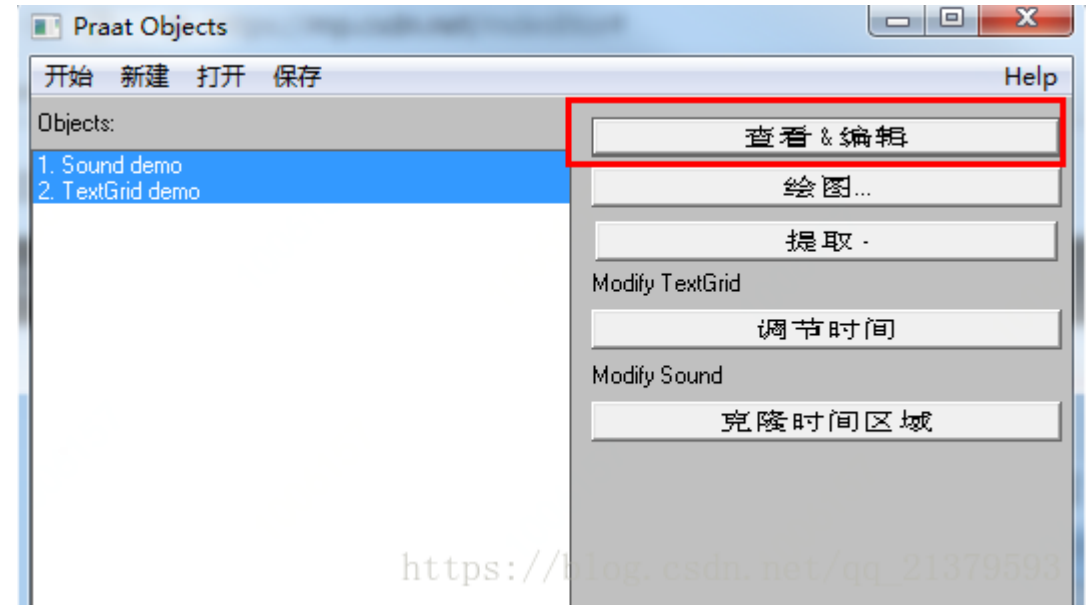
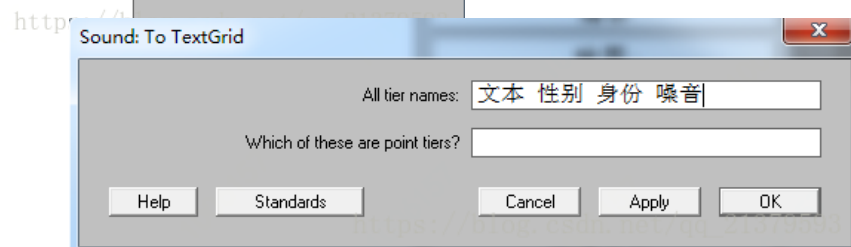
http://www.hejingzong.cn/blog/ViewBlog_54.aspx#vidio



PRAAT
doing phonetics by computer

version 5.3.72

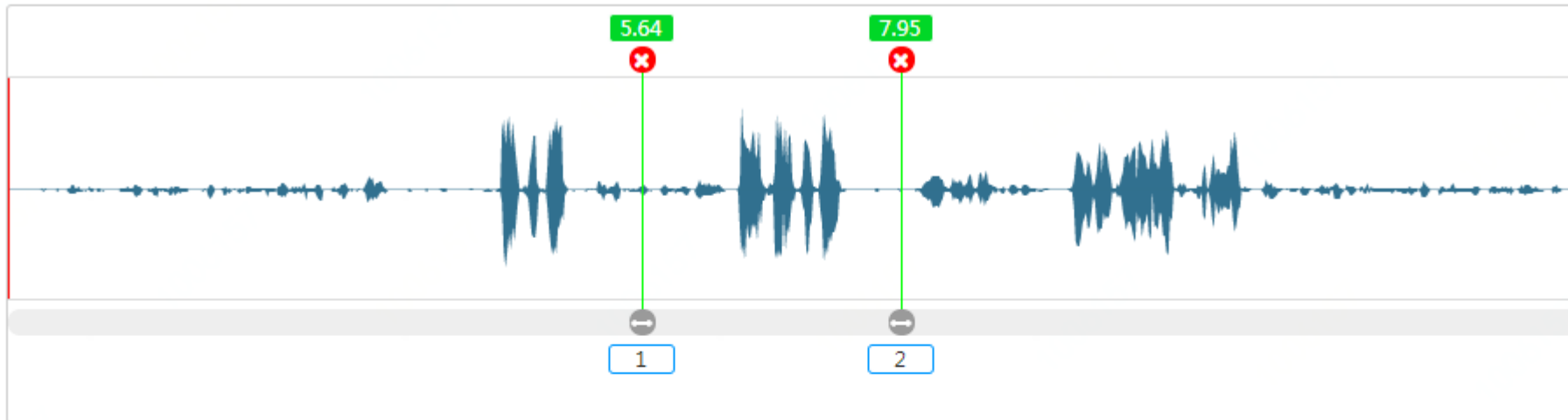
www.praat.org



How to annotate data?

The tools for audio (2)

京东众智 (<https://biao.jd.com/>)



播放进度(秒) : 0/14.00

【2】	角色	用户1 ▾	男	内容	<input type="text"/>	噪音 <input type="checkbox"/>	发音重叠 <input type="checkbox"/>	▶
【1】	角色	用户1 ▾	男	内容	<input type="text"/>	噪音 <input type="checkbox"/>	发音重叠 <input type="checkbox"/>	▶

用户1 : 男 ▾

How to annotate data?

The tools for video

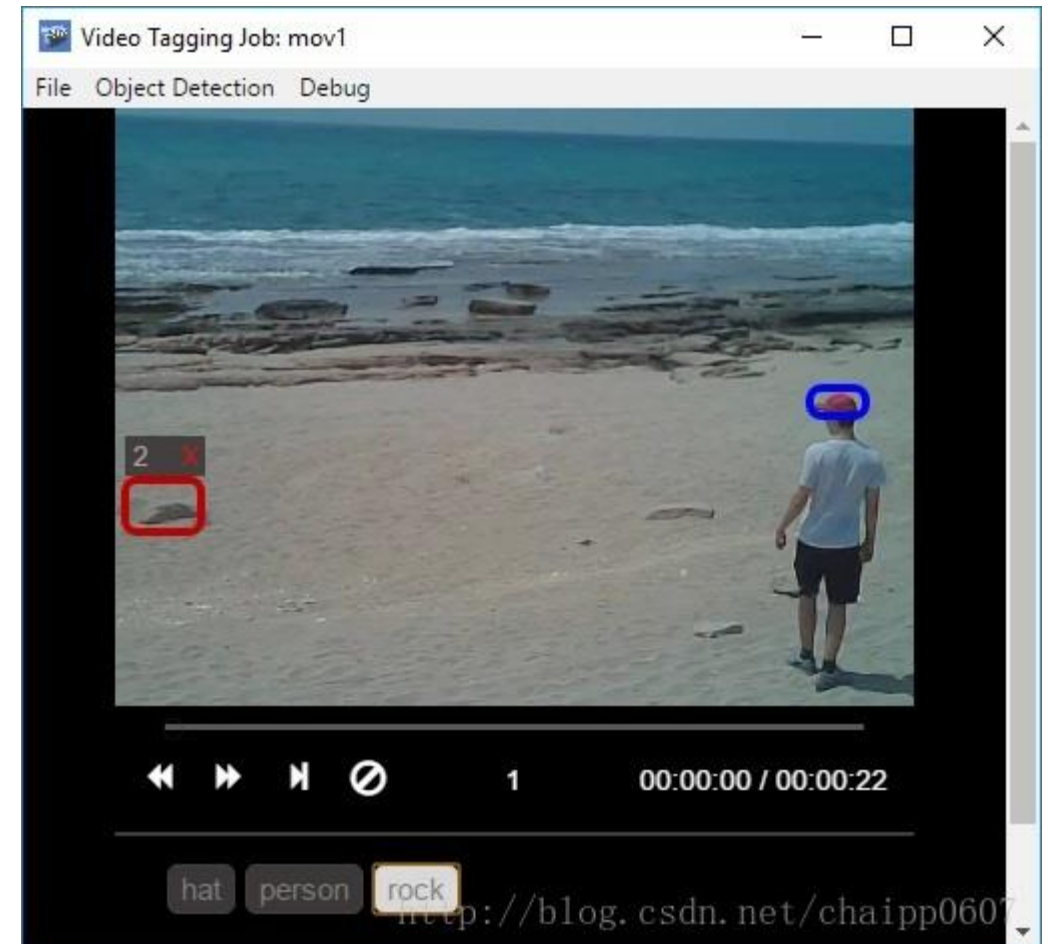
Vatic

<http://carlvondrick.com/vatic/>



VoTT

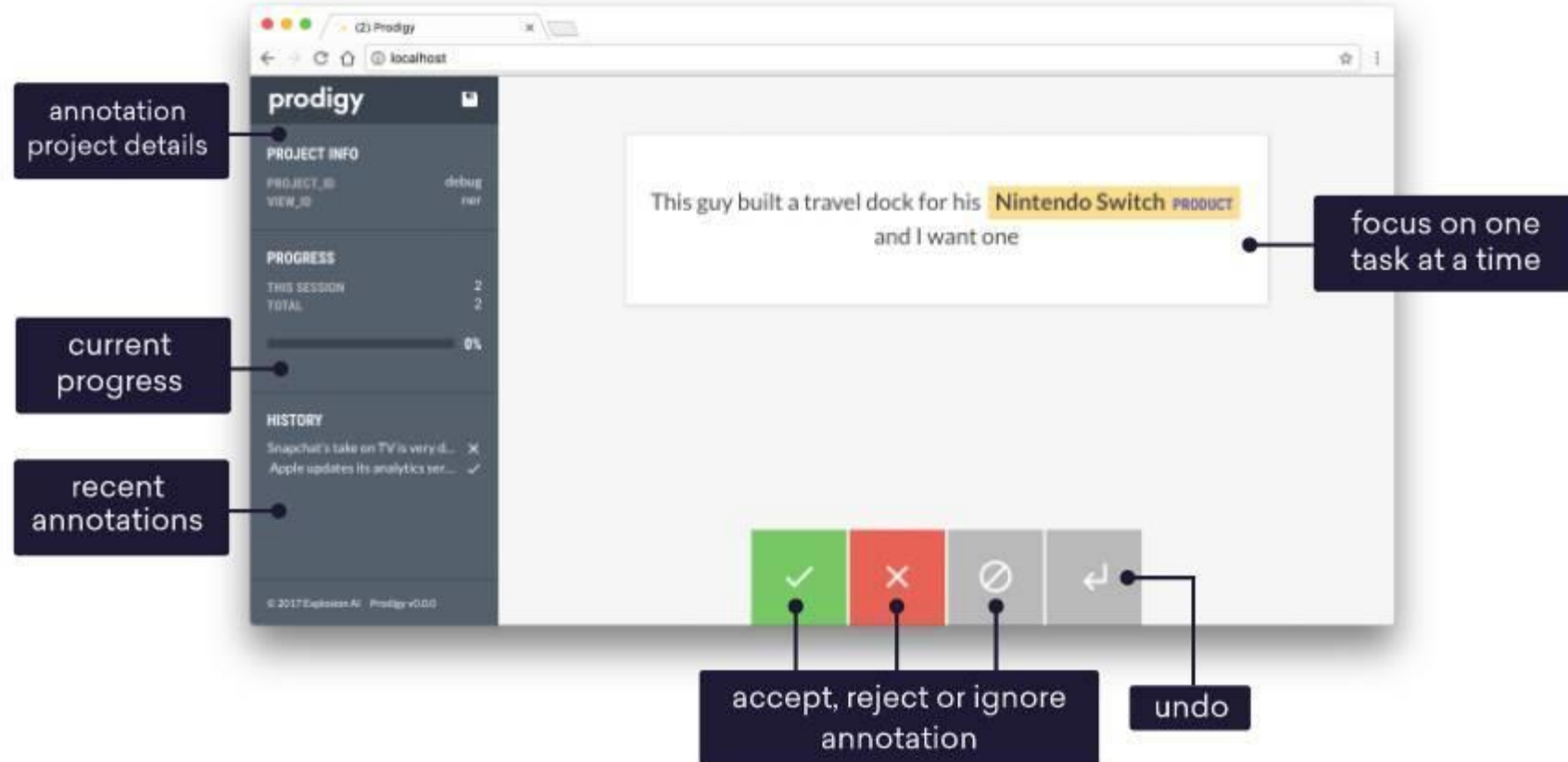
<https://github.com/Microsoft/VoTT/>



How to annotate data?

The tools for text/number (1)

Prodigy <https://prodi.gy/>
https://prodi.gy/demo?view_id=ner

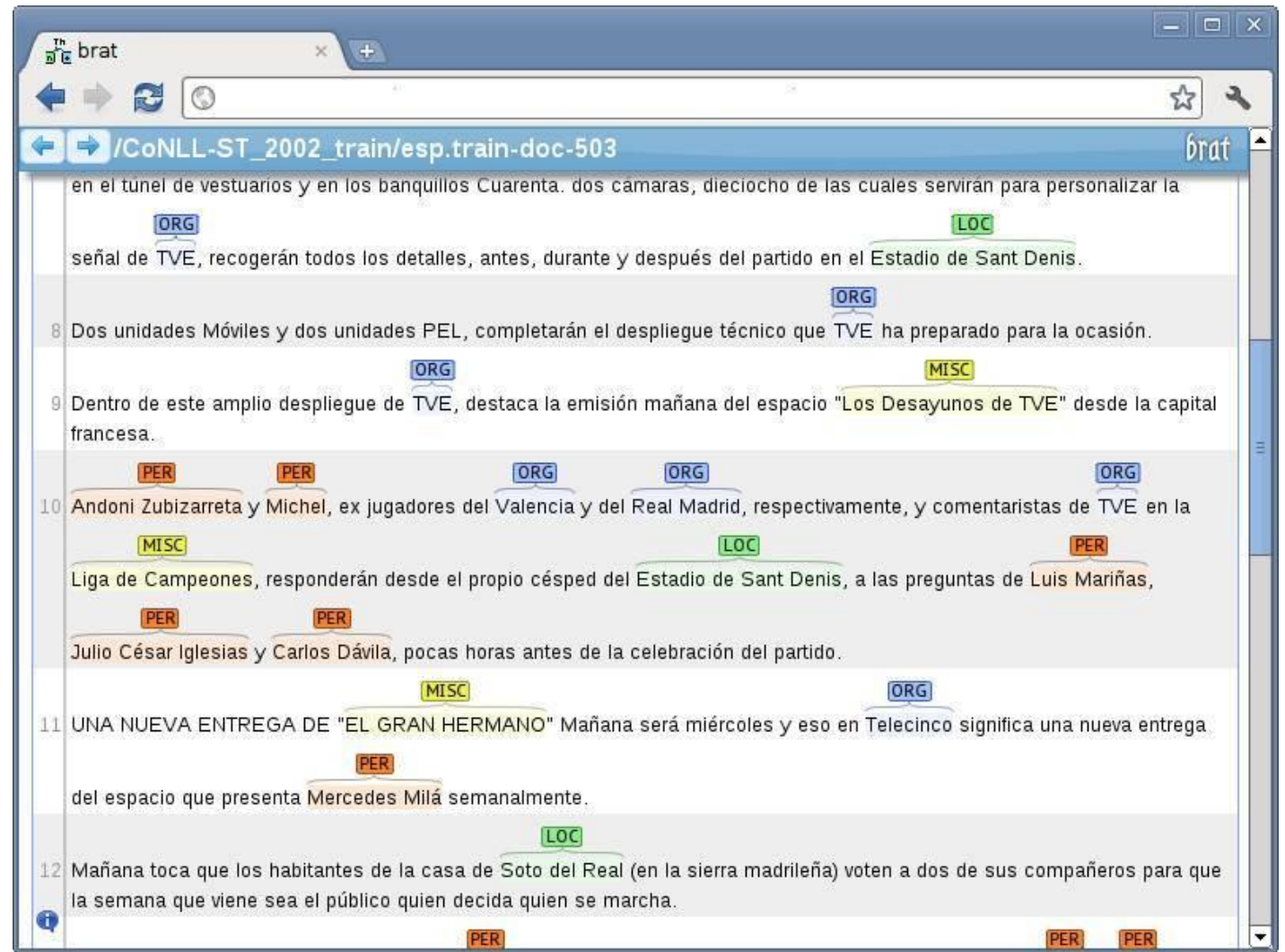


How to annotate data?

The tools for text/number (2)

BRAT

<http://brat.nlplab.org/index.html>



How to annotate data?

Automatic annotation

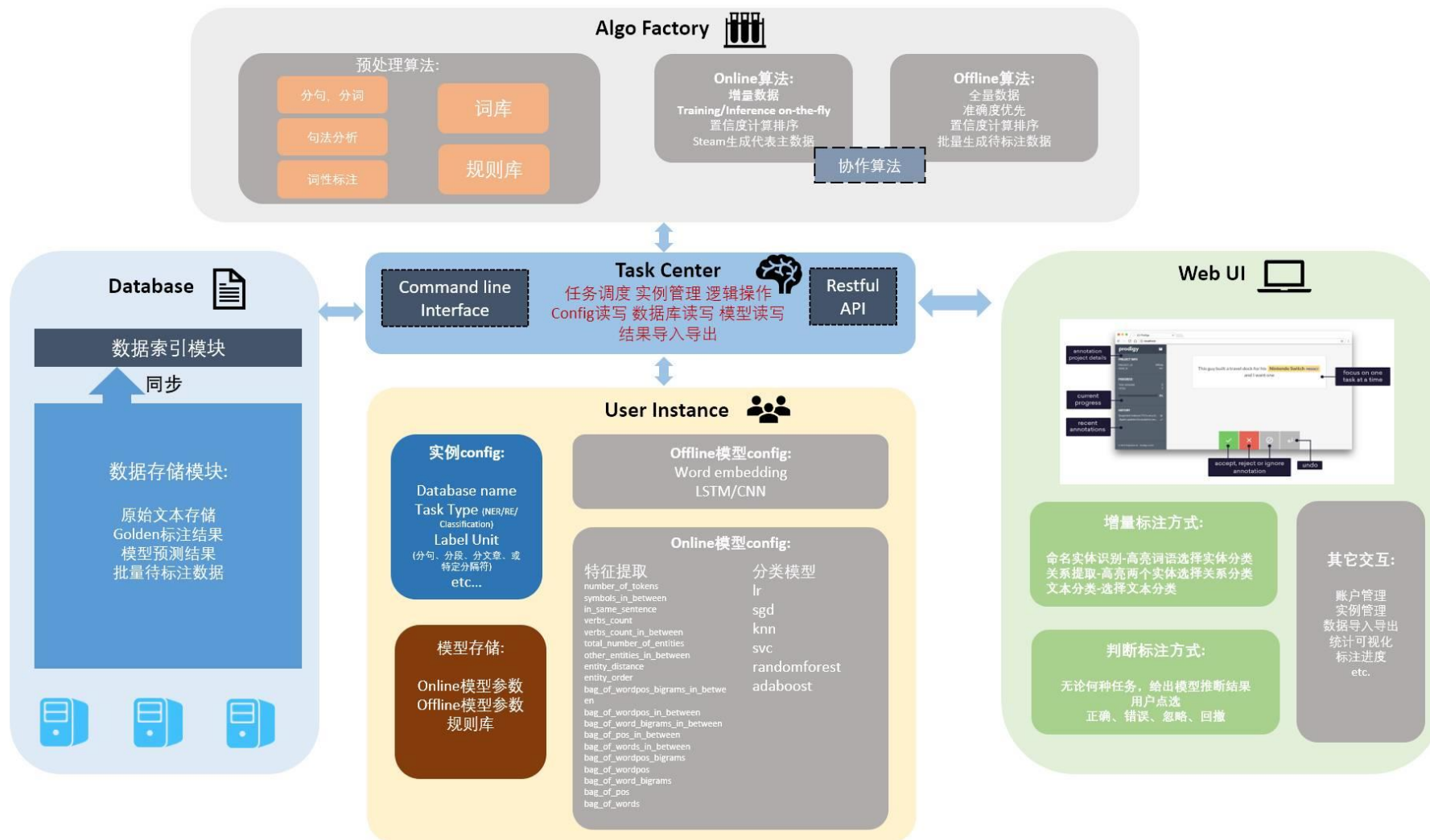
Based on rules

Based on machine learning



How to annotate data?

System structure of a data annotation system



How to annotate data?

How to design a tool for data annotation?

Important functions:

1. **Annotation interface:** a platform for annotation
2. **Statistical report:** show the work load of each annotator
3. **Process bar:** show the work progress of each annotator
4. **Save button:** temporary saving for uncertain data
5. **Submit button:** submit all the finished data
6. **Data import and export:** get data and release data
7. **QA service:** quality control



The End

Thank You

<http://www.wangting.ac.cn>