数据分析（科学）

数据科学的定义强调的是：

- 统计推断
- 数据可视化
- 实验设计
- 领域知识
- 沟通

数据科学家可能会使用一些简单的工具：他们计算报告百分比，并根据SQL查询制作线性图。 他们也可以使用非常复杂的方法：使用分布式数据存储来分析数以万亿计的数据记录，开发尖端的统计技术，并构建交互式可视化模型。 无论他们使用什么，目标都是更好地解读他们的数据。

机器学习

机器学习是属于预测领域的：“给定具有特定特征的样本X，预测它的Y值”。这些预测可能是关于未来（“预测这个病人是否会得败血症”），但也可能是一些对计算机而言的弱势领域（“例如预测这个图像中是否有鸟”）。几乎Kaggle的所有比赛项目都可以被认为是机器学习问题：他们提供一些训练数据，然后看看选手们能否使用自己的模型对新的示例做出准确的预测

人工智能

诸多“人工智能”定义中的一个共同点是：一个模拟人类智能的智能体代理，它能自主执行任务，并能根据行为作出反馈。（Poole，Mackworth和Goebel 1998，Russell and Norvig 2003）。一些我认为应该描述为人工智能的系统包括：

- 游戏演算法（Deep Blue，AlphaGo）
- 机器人和控制理论（运动规划，双足走路机器人）
- 优化（Google地图选择驾驶路线）
- 自然语言处理
- 强化学习


人工智能
机器学习
深度学习
1950's
1960's
1970's
1980's
1990's
2000's
2010's

NLP

自然语言处理(Natural Language Processing，NLP)属于人工智能的一个子领域，是指用计算机对自然语言的形、音、义等信息进行处理，即对字、词、句、篇章的输入、输出、识别、分析、理解、生成等的操作和加工。它对计算机和人类的交互方式有许多重要的影响。

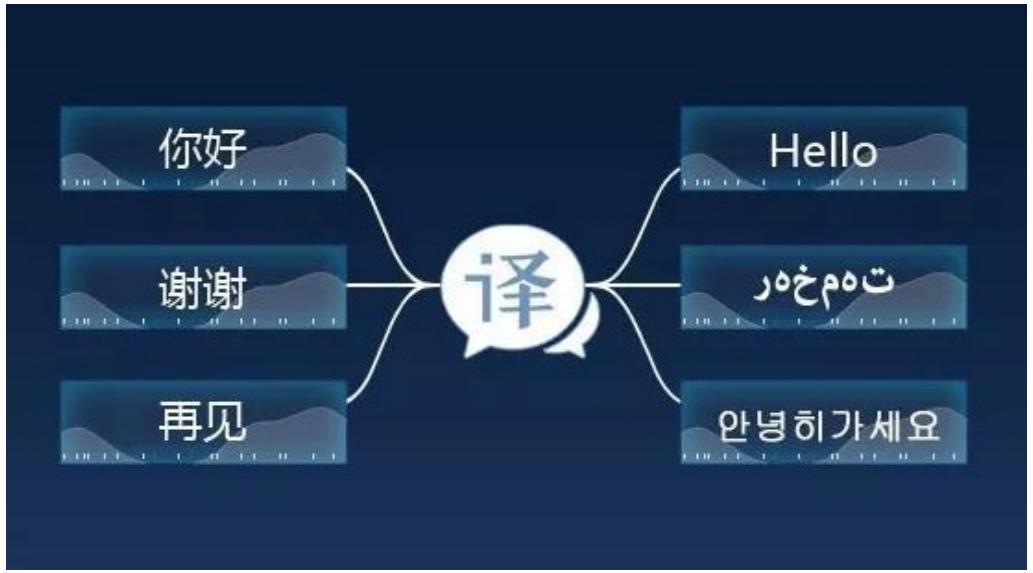
自然语言处理（Natural Language Processing, NLP）
自然语言理解（Natural Language Understanding, NLU）
自然语言生成（Natural Language Generation, NLG）

NLU: 理解文本中的意思
NLG: 根据意思生成文本

NLP = NLU + NLG

从NLU到NLG

应用场景



机器翻译

机器翻译 随着通信技术与互联网技术的飞速发展、信息的急剧增加以及国际联系愈加紧密，让世界上所有人都能跨越语言障碍获取信息的挑战已经超出了人类翻译的能力范围。

机器翻译因其效率高、成本低满足了全球各国多语言信息快速翻译的需求。目前，谷歌翻译、百度翻译、搜狗翻译等人工智能行业巨头推出的翻译平台逐渐凭借其翻译过程的高效性和准确性占据了翻译行业的主导地位。

应用场景

垃圾邮件过滤

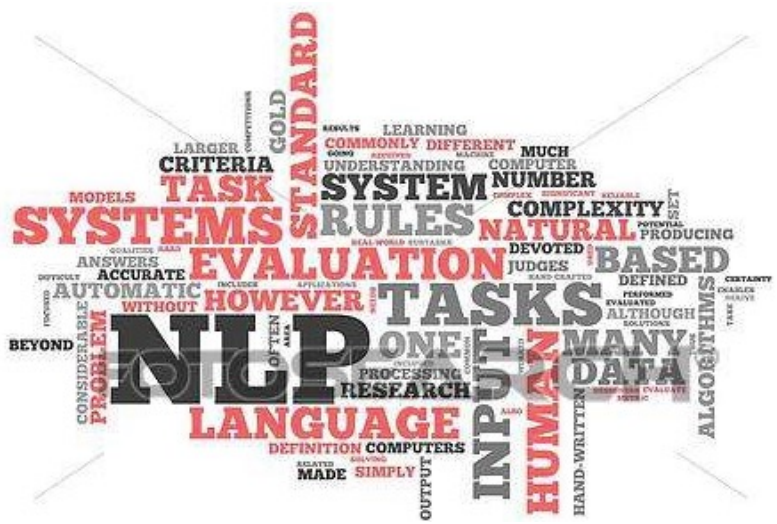
垃圾邮件 随着通信技术与互联网技术的飞速发展、信息的急剧垃圾邮件过滤器已成为抵御垃圾邮件问题的第一道防线。不过，有许多人在使用电子邮件时遇到过这些问题：不需要的电子邮件仍然被接收，或者重要的电子邮件被过滤掉。事实上，判断一封邮件是否是垃圾邮件，首先用到的方法是"关键词过滤"，如果邮件存在常见的垃圾邮件关键词，就判定为垃圾邮件。但这种方法效果很不理想，一是正常邮件中也可能有这些关键词，非常容易误判，二是将关键词进行变形，就很容易规避关键词过滤。

应用场景

情感分析

情感分析 基于上下文感知的情感分析、跨领域跨语言情感分析、基于深度学习的端到端情感分析、情感解释、反讽分析、立场分析等；比如，企业分析消费者对产品的反馈信息，或者检测在线评论中的差评信息等。

应用场景



自动摘要

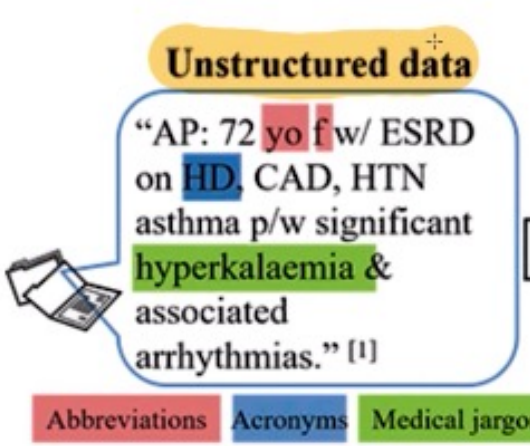
自动文摘 新闻的摘要要求编辑能够从新闻事件中提取出最关键的信息点，重新组织语言来写摘要；paper的摘要需要作者从全文中提取出最核心的工作，然后用更加精炼的语言写成摘要；综述性的paper需要作者通读N篇相关topic的paper之后，用最概括的语言将每篇文章的贡献、创新点写出来，并且对比每篇文章的方法各有什么优缺点。

应用场景

自动问答机器人

自动问答 随着互联网的快速发展，网络信息量不断增加，人们需要获取更加精确的信息。传统的搜索引擎技术已经不能满足人们越来越高的需求，而自动问答技术成为了解决这一问题的有效手段。自动问答是指利用计算机自动回答用户所提出的问题以满足用户知识需求的任务，在回答用户问题时，首先要正确理解用户所提出的问题，抽取其中关键的信息，在已有的语料库或者知识库中进行检索、匹配，将获取的答案反馈给用户。

应用场景



Meaningful information

Attributes	Medical text
Age	72 years old (yo)
Gender	female (f)
Condition	hypertensive disease (HD)
Symptom	hyperkalemia

信息抽取

诗歌生成

突然想起唐伯虎的那句诗：人不轻狂枉少年。我想作为成年人的我们，怀念的或许不止是少年悠闲的时光，更是那时初生牛犊不怕虎，愿意勇敢尝试不计得失的精神。因而用了“人再少年”这四个字作为诗篇每句的开始作为对少年时代的缅怀。

人生何所贵，再议良媒后。少壮乃不甘，年华复谁惜。

计算机视觉（CV）

但是相机看到的则是这样的：

194	210	201	212	199	213	215	195	178	158	182	209
180	189	190	221	209	205	191	167	147	115	129	163
114	126	140	188	176	165	152	140	170	106	78	88
87	103	115	154	143	142	149	153	173	101	57	57
102	112	106	131	122	138	152	147	128	84	58	66
94	95	79	104	105	124	129	113	107	87	69	67
68	71	69	98	89	92	98	95	89	88	76	67
41	56	68	99	63	45	60	82	58	76	74	65
20	41	69	75	56	41	51	73	55	70	63	44
50	50	57	69	75	75	73	74	53	68	59	37
72	59	53	66	84	92	84	74	57	72	63	42
67	61	58	65	75	78	76	73	59	75	69	50

CV应用

人脸识别（Face Recognition）是基于人的面部特征信息进行身份识别的一种生物识别技术。采集含有人脸的图片或视频流，并自动在图片中检测和跟踪人脸，进而对检测到的人脸进行面部识别的一系列相关技术。人脸识别可提供图像或视频中人脸检测定位、人脸属性识别、人脸比对、活体检测等。

CV应用

视频监控分析是利用机器视觉技术对视频中的特定内容信息进行快速检索、查询、分析的技术。由于摄像头得到了广泛的应用，它们产生的视频数据是一个天文数字，这些数据蕴藏的价值巨大，靠人工根本无法统计，机器视觉技术的逐步成熟，让视频分析成为可能，这项技术使得公安部门在海量的监控视频中搜寻到罪犯有了可能，在大量人群流动的交通枢纽，该技术也被广泛用于人群分析、防控预警等。

CV应用

拍立淘。拍立淘是手机淘宝的一个应用，主要解决的问题是通过图片来代替文字进行搜索，来帮助用户搜索无法用简单文字描述的需求。比如你看到了一个裙子很好看， 但又很难用简单的语言文字来描述这个裙子的样子，那么这个时候可以使用拍立淘，可以很轻松地在淘宝上搜出同款裙子，或是和它非常接近的款式。

CV应用

自动驾驶是一种通过计算机实现无人驾驶的智能汽车，自动驾驶汽车依靠人工智能、机器视觉、雷达、监控装置和全球定位系统协同合作，让计算机可以在没有任何人类主动的操作下，自动安全地操作机动车辆。机器视觉的快速发展推动了自动驾驶技术的成熟，使无人驾驶在未来成为可能。

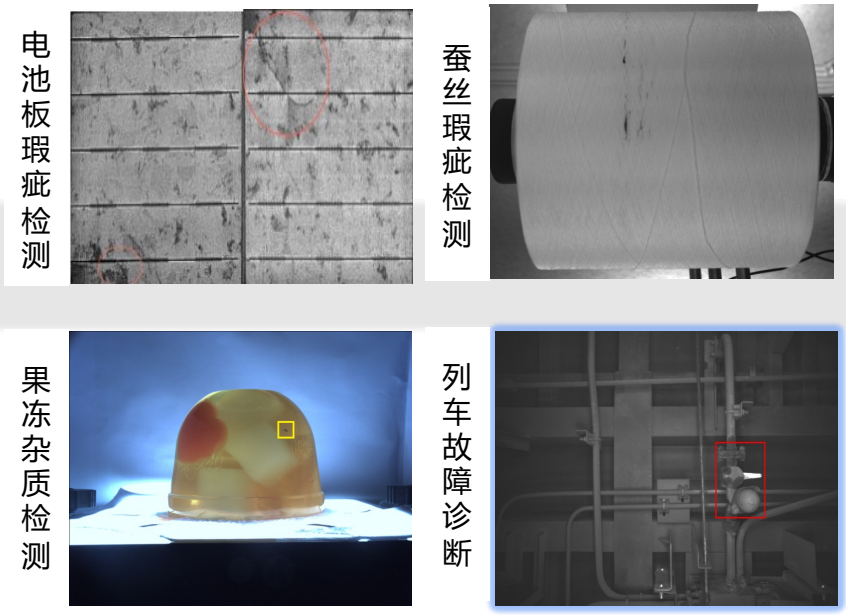
CV应用

计算机文字识别，俗称光学字符识别(Optical Character Recognition)，它是利用光学扫描技术把票据、报刊、书籍、文稿及其它印刷品的文字转化为图像信息，再利用文字识别技术将图像信息转化为可以使用的计算机输入技术。

编号	识别结果
1	明珠塔路
2	MingZhu ta
3	滨江大道
4	binjiang Ave
5	陆家嘴环路
6	Lujiazui Ring Rd
7	陆家嘴西路

瑕疵检测

瑕疵检测是科技新型研发瑕疵检测系统。主要是通过工业摄像头，定位系统，以及目前正在实践的基于深度学习的瑕疵智能检测方案。



深度学习和图像处理算法

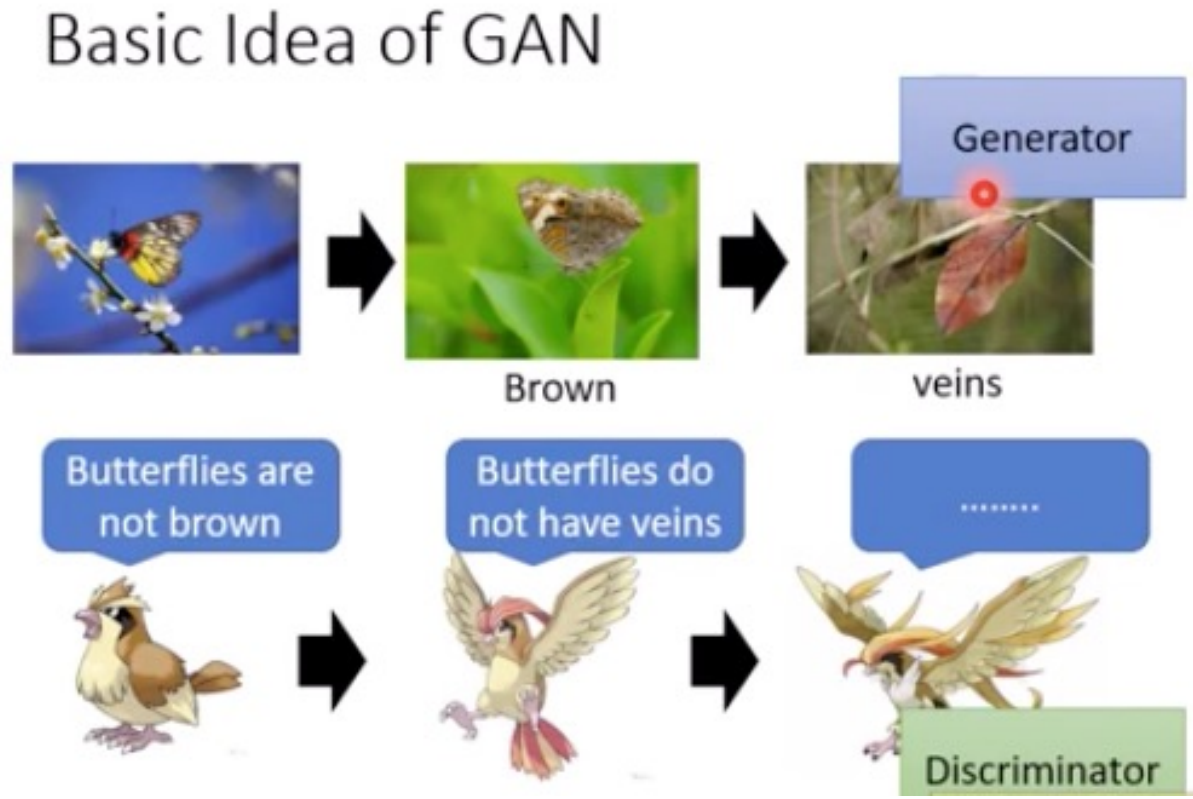
自动识别瑕疵或故障

全自动生产流程

节省人力，提高精度和稳定性

对抗网络

生成对抗网络（Generative Adversarial Net，GAN) 是近年来深度学习中一个十分热门的方向，卷积网络之父、深度学习元老级人物LeCun Yan说过“GAN is the most interesting idea in the last 10 years in machine learning”。



适合做什么

Why distribution?

(The same input has different outputs.)

- Especially for the tasks needs "*creativity*"

Drawing

Character with red eyes → Network →

Chatbot

你知道輝夜是誰嗎? → Network →

她是秀知院學生會 ...

她開創了忍者時代 ...

图计算（知识图谱）

图计算中的图英文是Graph，用英文完整的表达就是Graph Computing。
图计算是研究客观世界当中的任何事物和事物之间的关系，对其进行完整的刻划、计算和分析的一门技术。
图计算技术主要是由点和边来组成的。举例来说，点可以是坐在演播室里的三个人，这三个人就是三个点，所谓的边就是这三个人之间的关系。比如说，同事关系、亲戚关系、夫妻关系等。

图计算（知识图谱）

图计算的应用场景

①社交网络分析

社交网络是十分常见的一类图数据，代表着各种个人或组织之间的社会关系，而图数据能够呈现复杂的社交网络关系，进而易于用户进行进一步的分析。例如，在一个典型的社交网络中，常常会存在“谁认识谁，谁上过什么学校，谁常住什么地方”。Facebook、Twitter、Linkedin用它来管理社交关系，实现好友推荐。

②电子购物应用

电子购物是互联网中的一类核心业务，在这类场景中，节点分为两类：用户和商品，存在的关系有浏览、收藏、购买等。用户与商品之间可以存在多重关系，如既存在收藏关系也存在购买关系。这类复杂的数据场景可以用属性图轻松描述。电子购物催生了一项大家熟知的技术应用—推荐系统。用户与商品之间的交互关系，反映了用户的购物偏好。例如，经典的啤酒与尿布的故事：爱买啤酒的人通常也更爱买尿布。

③交通网络应用

交通网络具有多种形式，比如地铁网络中将各个站点作为节点，站点之间的连通性作为边。通常在交通网络中我们比较关注的是路径规划相关的问题：比如最短路径问题，再如我们将车流量作为网络中节点的属性，去预测未来交通流量的变化情况。

联邦学习

谷歌最开始提出联邦学习时是为了解决C端用户终端设备上模型训练的问题。C端用户手机上的智能软件提供服务时背后都得依靠模型，而模型的训练学习全部要基于用户的数据。比如手机上的输入法，基于不同人的打字拼音习惯，输入法会不停更新会慢慢和每个人的打字习惯进行匹配，用户会觉得输入法越来越智能；
那么过去这些手机输入法是如何进行模型训练的了？

过去的做法：将用户每天产生的行为数据全部上传至云端服务器，部署在服务器上的模型基于上传的数据进行训练，然后更新模型，最终实际应用时本地需要请求云端服务。

上述这种模型训练的方式，我们也叫做“集中式模型训练”，这种方式有两个弊端：
•**无法保证用户的数据隐私**：服务商将用户的数据全部采集到了服务器上进行统一管理。这种方式在监管对个人数据隐私管控越来越严的情况下，会越来越受限；
•**实时性难以保证**：模型在应用时需要通过网络请求云端的模型，在网络延迟或者没有网络的情况下，模型没办法发挥它的作用；
为了解决上述的弊端，谷歌提出了一种新的解决方案，并将它命名为“Federated Learning”。总的来说就是：**用户数据不出本地，所有模型的训练都是在设备本地进行。本地模型训练完毕后将得到的模型参数or下降梯度，经过加密上传至云端，云端模型接收到所有上传的加密参数or梯度后，结合所有的参数值进行统一的聚合，比如通过加权平均得到新的模型参数or下降梯度，然后将新的结果再重新下发到本地，本地更新得到一个全新的模型。**

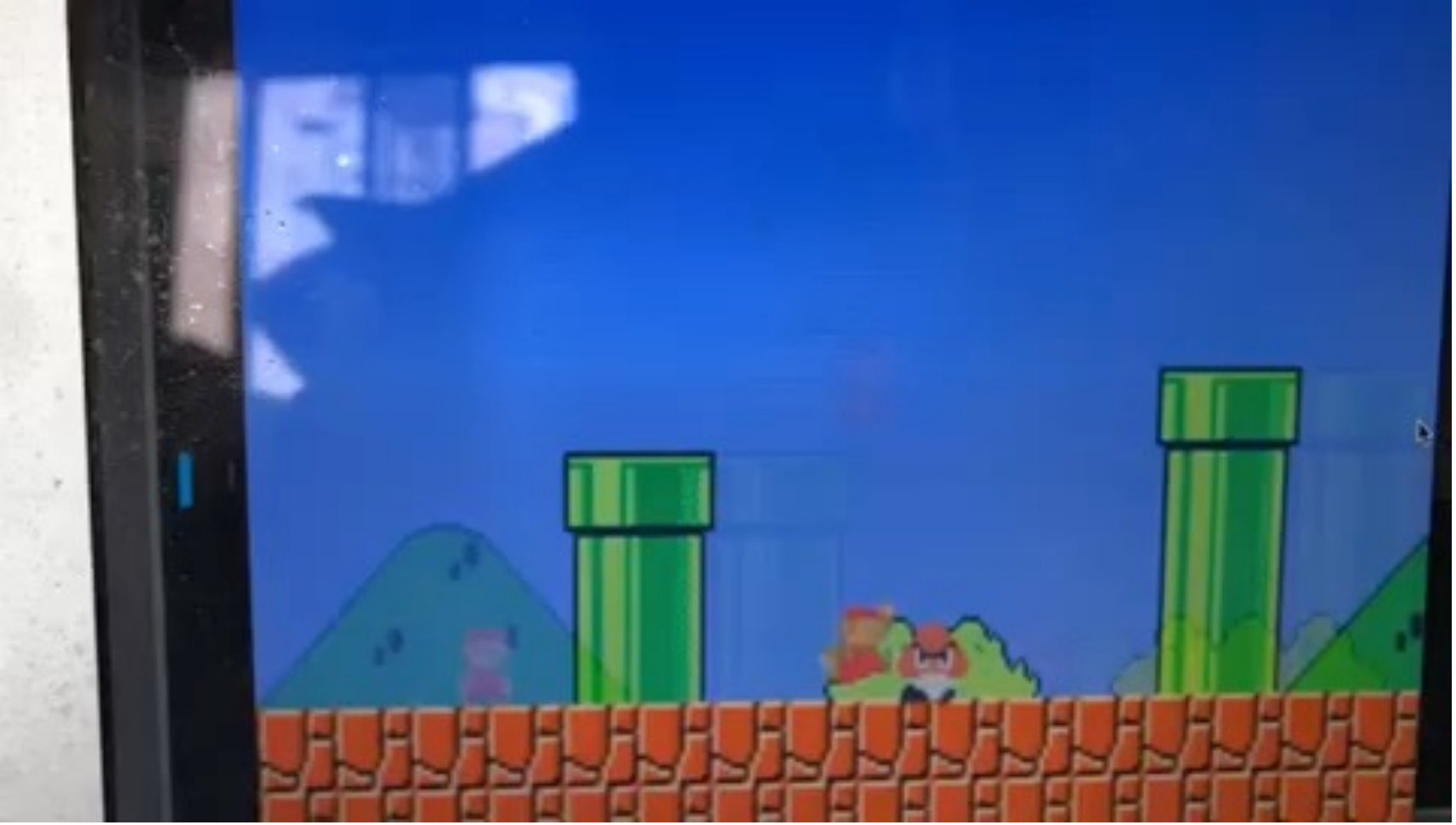
知识蒸馏

各种模型算法，最终目的都是要为某个应用服务。在买卖中，我们需要控制收入和支出。类似地，在工业应用中，除了要求模型要有好的预测(收入)以外，往往还希望它的「支出」要足够小。具体来说，我们一般希望部署到应用中的模型使用较少的计算资源(存储空间、计算单元等)，产生较低的时延。
在深度学习的背景下，为了达到更好的预测，常常会有两种方案：1. 使用参数化的深度神经网络，这类网络学习能力非常强，因此往往加上一定的正则化策略(如dropout)；2. 集成模型(ensemble)，将许多弱的模型集成起来，往往可以实现较好的预测。这两种方案无疑都有较大的「支出」，需要的计算量和计算资源很大，对部署非常不利。这也就是模型压缩的动机：我们希望有一个规模较小的模型，能达到和大模型一样或相当的结果。当然，从头训练一个小模型，从经验上看是很难达到上述效果的，也许我们能先训练一个大而强的模型，然后将其包含的知识转移给小的模型呢—知识蒸馏

强化学习

强化学习主要由智能体（Agent）、环境（Environment）、状态（State）、动作（Action）、奖励（Reward）组成。智能体执行了某个动作后，环境将会转换到一个新的状态，对于该新的状态环境会给出奖励信号（正奖励或者负奖励）。随后，智能体根据新的状态和环境反馈的奖励，按照一定的策略执行新的动作。上述过程为智能体和环境通过状态、动作、奖励进行交互的方式。

强化学习



非监督学习

什么是无监督学习?

现实生活中常常会有这样的问题：缺乏足够的先验知识，因此难以人工标注类别，或是进行人工类别标注的成本太高，例如从庞大的样本集合中选出具有代表性的样本加以标注用于分类器的训练、在无类别信息情况下，寻找明显特征或将所有样本自动分为不同的类别等。

因此就会希望能由计算机来全部或部分地完成这些工作，无监督学习也就应运而生，其本质上是一个统计手段，无监督学习主要是通过根据类别未知(没有被标记)的训练样本来解决模式识别中的各种问题。

非监督学习

常见的无监督学习适用场景可以包括：

1.发现异常

如果对于大量数据进行分析，并通过人力去找出其中的异常是一件成本很高很复杂的事情。通过无监督学习，我们可以快速把行为特征进行分类，虽然可能不知道这些分类意味着什么，但是通过这种分类，可以快速排出正常的用户，更有针对性的对异常行为进行深入分析。

2.用户细分

对于很多广告平台，不仅可以把用户按性别、年龄、地理位置等维度进行用户细分，还可以通过用户行为对用户进行分类，通过无监督学习对用户从多维度进行细分，广告投放可以更有针对性，效果更好。

3.推荐系统

相信大家都听过”啤酒+尿不湿”的故事，这个故事就是根据用户的购买行为来推荐相关的商品的一个典型例子。

比如大家在购物软件上网购的时候，总会根据你的浏览行为推荐一些相关的商品，有些商品就是无监督学习通过聚类来推荐出来的。系统会发现一些购买行为相似的用户，推荐这类用户最”喜欢”的商品。

Theano

- Theano是在BSD许可证下发布的一个开源项目，是由LISA集团（现MILA）在加拿大魁北克的蒙特利尔大学开发。它是用一个希腊数学家的名字命名的。
- 在过去的很长一段时间内，Theano 是深度学习开发与研究的行业标准。而且，由于出身学界，它最初是为学术研究而设计，这导致深度学习领域的许多学者至今仍在使用 Theano。但随着 Tensorflow 在谷歌的支持下强势崛起，Theano 日渐式微，使用的人越来越少。这过程中的标志性事件是：创始者之一的Ian Goodfellow 放弃 Theano 转去谷歌开发 Tensorflow了。

MXNet

- MXNet 是亚马逊（Amazon）的李沐带队开发的深度学习框架。它拥有类似于 Theano 和 Tensorflow 的数据流图，为多 GPU 架构提供了良好的配置，有着类似于 Lasagne 和 Blocks 的更高级别的模型构建块，并且可以在你想象的任何硬件上运行（包括手机）。对 Python 的支持只是其冰山一角，MXNet 同样提供了对 R、Julia、C++、Scala、Matlab、Golang和 Java 的接口。
- MXNet以其超强的分布式支持，明显的内存、显存优化为人所称道。同样的模型，MXNet往往占用更小的内存和显存，并且在分布式环境下，MXNet展现出了明显优于其他框架的扩展性能。
- 总结来看：文档比较混乱导致不太适合新手入门，但其分布性能强大，语言支持比较多，比较适合在云平台使用。
- 项目主页：https://mxnet.incubator.apache.org/。

Keras

- Keras并不能称为一个深度学习框架，它更像一个深度学习接口，它构建于第三方框架之上。Keras的缺点很明显：过度封装导致丧失灵活性。Keras最初作为Theano的高级API而诞生，后来增加了TensorFlow和CNTK作为后端。为了屏蔽后端的差异性，提供一致的用户接口，Keras做了层层封装，导致用户在新增操作或是获取底层的数据信息时过于困难。同时，过度封装也使得Keras的程序过于缓慢，许多BUG都隐藏于封装之中，在绝大多数场景下，Keras是本节介绍的所有框架中最慢的一个。
- 项目主页：https://keras.io。

Caffe2

- Caffe2出自Facebook 人工智能实验室与应用机器学习团队，贾杨清仍是主要贡献者之一。Caffe2 在工程上做了很多优化，比如运行速度、跨平台、可扩展性等等，它可以看作是 Caffe 更细粒度的重构，但在设计上其实Caffe2和TensorFlow更像。目前代码已开源。
- 总结来说：Caffe至今工业界和学界仍有很多人在使用，Caffe2的出现给我们提供了更多的选择。
- 项目地址：Caffe： http://caffe.berkeleyvision.org/
- Caffe2： https://caffe2.ai/

Pytorch

- Pytorch 是一个 Python 优先的深度学习框架，能够在强大的 GPU 加速基础上实现张量和动态神经网络。
- Pytorch 是一个 Python 软件包，其提供了两种高层面的功能：
- 使用强大的 GPU 加速的 Tensor 计算（类似 Numpy）。
- 构建于基于 tape 的 autograd 系统的深度神经网络。
- 活跃的社区：PyTorch提供了完整的文档，循序渐进的指南，作者亲自维护的论坛 供用户交流和求教问题。Facebook 人工智能研究院对PyTorch提供了强力支持，作为当今排名前三的深度学习研究机构，FAIR的支持足以确保PyTorch获得持续的开发更新，不至于像许多由个人开发的框架那样昙花一现。

- **总结来看：如果说 TensorFlow的设计是“Make It Complicated”，Keras的设计是“Make It Complicated And Hide It”，那么Pytorch的设计真正做到了“Keep it Simple，Stupid”。**
- 项目地址：http://pytorch.org/。

TensorFlow

- 2015年11月10日，Google宣布推出全新的机器学习开源工具Tensorflow。 Tensorflow 最初是由 Google 机器智能研究部门的 Google Brain 团队开发，基于Google 2011年开发的深度学习基础架构DistBelief构建起来的。Google几乎在所有应用程序中都使用Tensorflow来实现机器学习。 例如，如果您使用到了Google照片或Google语音搜索，那么您就间接使用了Tensorflow模型。它们在大型Google硬件集群上工作，在感知任务方面功能强大。

- **总结来说，凭借Google着强大的推广能力，Tensorflow已经成为当今最炙手可热的深度学习框架，不完美但是最流行，目前来看各公司使用的框架也不统一，有必要多学习几个流行框架作为知识储备，无疑Tensorflow就是一个不错的选择。**
- 项目地址：https://github.com/tensorflow/tensorflow。

学什么

从本次课程里能学到什么，***灵魂拷问***的应该是自己。

有的人也列出了一些非常简明扼要的判断标准，比如：
- 什么方向工资高，就学什么
- 现在媒体上宣传什么，就学什么
- 大师说学什么，就学什么
- 招聘网站上什么岗位多，就学什么

假如时光倒退20年，你学什么？

籍曰："书足以记名姓而已，剑一人敌，不足学，学万人敌。"

——《史记·项羽本纪》

每个领域，都会有一些沉淀下来的基础知识和技能，这是要学的。此外，更重要的是“学习能力”，学习内容是增长学习能力的载体，学习能力就是“万人敌”。
因此，**本课程从始至终，我希望贯彻一个宗旨：提升学习能力**，不是唠唠叨叨地诉说知识点。

课程寄语

最后送上励志名言：

努力面前，忘记背后，向着标杆直跑。

当遇到貌似无法克服的困难时，当怀疑自己是否能够学会时，当受到其他诱惑想放弃时，请来读一读这句话。