CRUISER

巡洋舰科技

主讲人：许铁
巡洋舰科技Founder & CEO

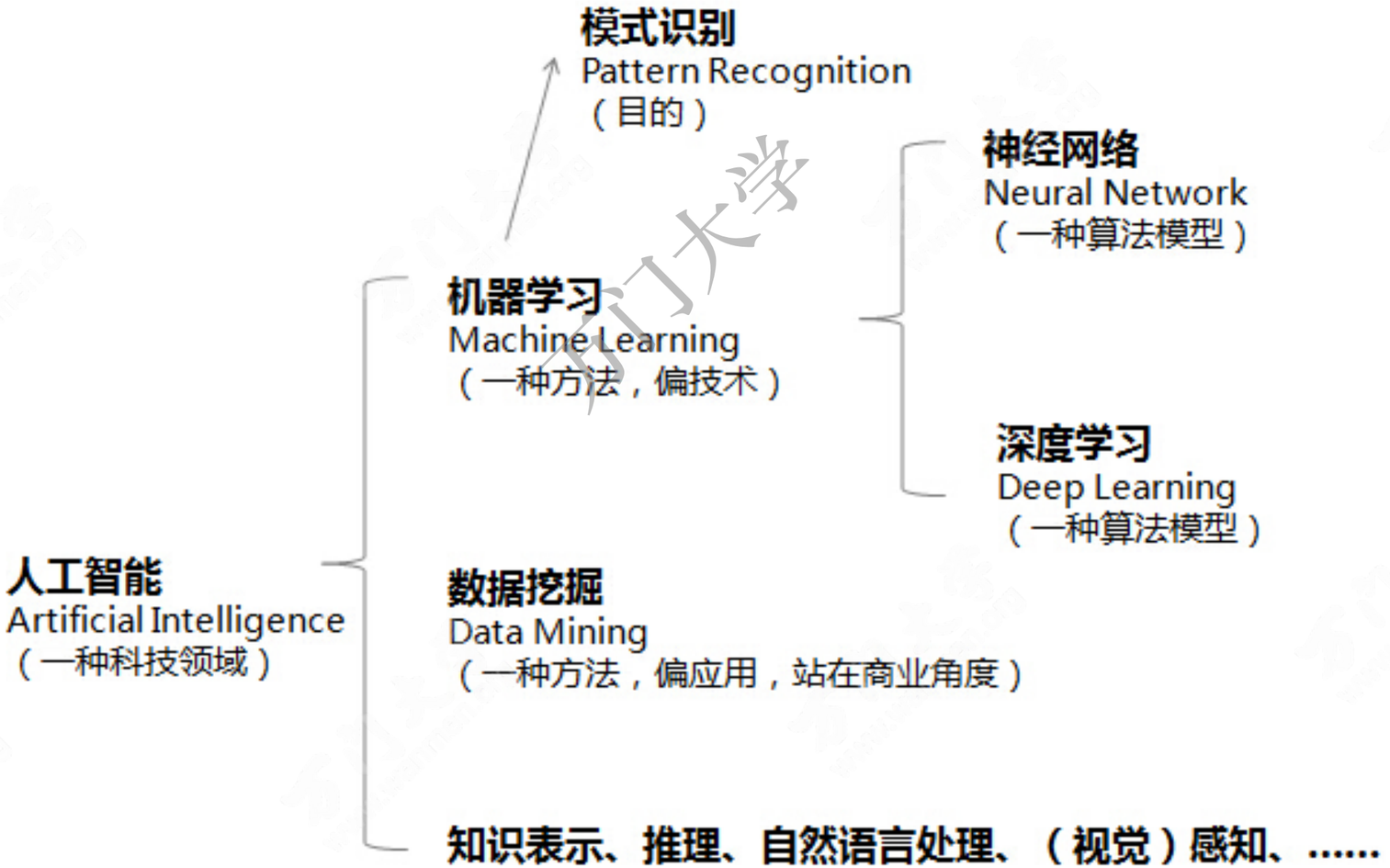
Syllabus

理解机器学习含义

理解机器学习训练

理解线性回归

书写你的第一个机器学习程序


模式识别
Pattern Recognition
（目的）
神经网络
Neural Network
（一种算法模型）
机器学习
Machine Learning
（一种方法，偏技术）
深度学习
Deep Learning
（一种算法模型）
人工智能
Artificial Intelligence
（一种科技领域）
数据挖掘
Data Mining
（一种方法，偏应用，站在商业角度）
知识表示、推理、自然语言处理、（视觉）感知、……

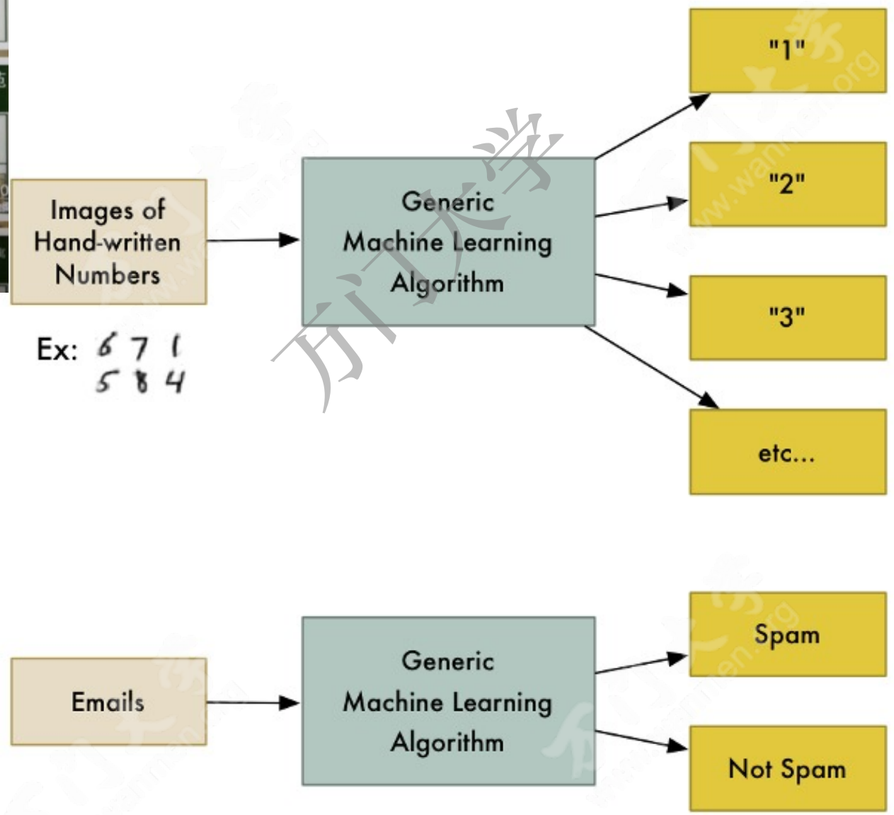
CRUISER

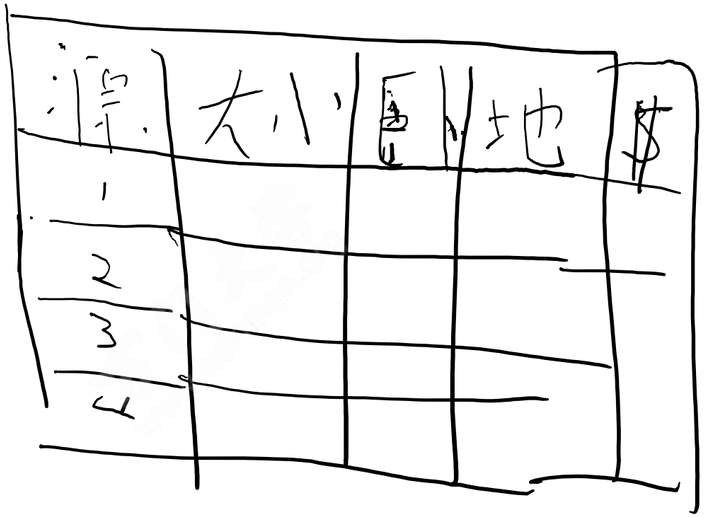
婴儿是怎么学习的

错误是成功之母

CRUISER

什么是·机器学习？





机器学习算法是个黑盒，可以重用来解决很多不同的分类问题。
“机器学习”是一个涵盖性术语，覆盖了大量类似的泛型算法

机器学习为什么work？
一个故事讲给你听

CRUISER

房价的预测

监督学习

Bedrooms	Sq. feet	Neighborhood	Sale price
3	2000	Normaltown	$250,000
2	800	Hipsterton	$300,000
2	850	Normaltown	$150,000
1	550	Normaltown	$78,000
4	2000	Skid Row	$150,000

Bedrooms	Sq. feet	Neighborhood	Sale price
3	2000	Hipsterton	???

存在不存在一个计算机程序， 从前三者算出最后？

CRUISER

基于规则的代码

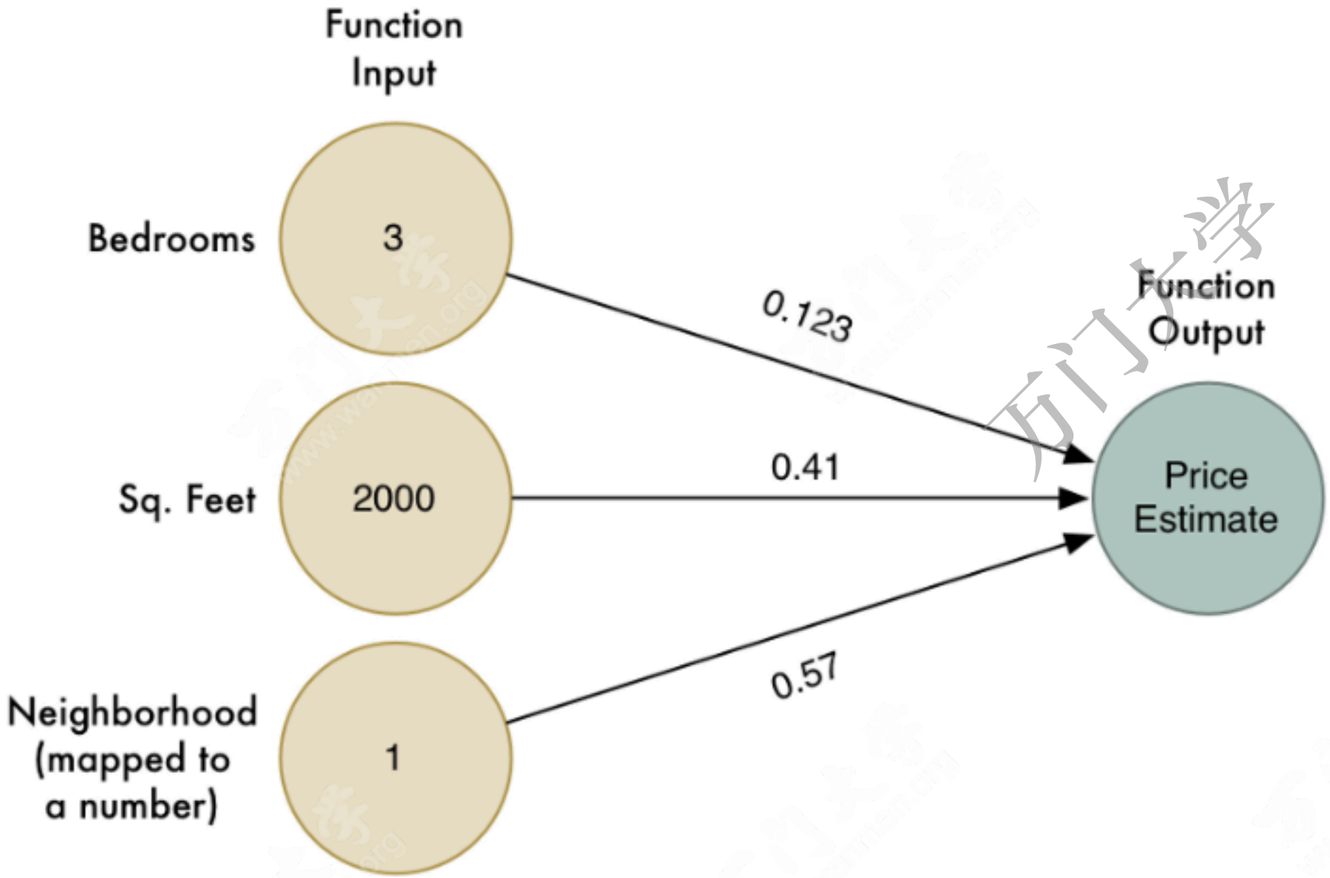
如果你对机器学习一无所知....你的代码将会是...

```python
def estimate_house_sales_price(num_of_bedrooms, sqft, neighborhood):
  price = 0

  # In my area, the average house costs $200 per sqft
  price_per_sqft = 200

  if neighborhood == "hipsterton":
    # but some areas cost a bit more
    price_per_sqft = 400

  elif neighborhood == "skid row":
    # and some areas cost less
    price_per_sqft = 100

  # start with a base price estimate based on how big the place is
  price = price_per_sqft * sqft

  # now adjust our estimate based on the number of bedrooms
  if num_of_bedrooms == 0:
    # Studio apartments are cheap
    price = price-20000
  else:
    # places with more bedrooms are usually
    # more valuable
    price = price + (num_of_bedrooms * 1000)

return price
```

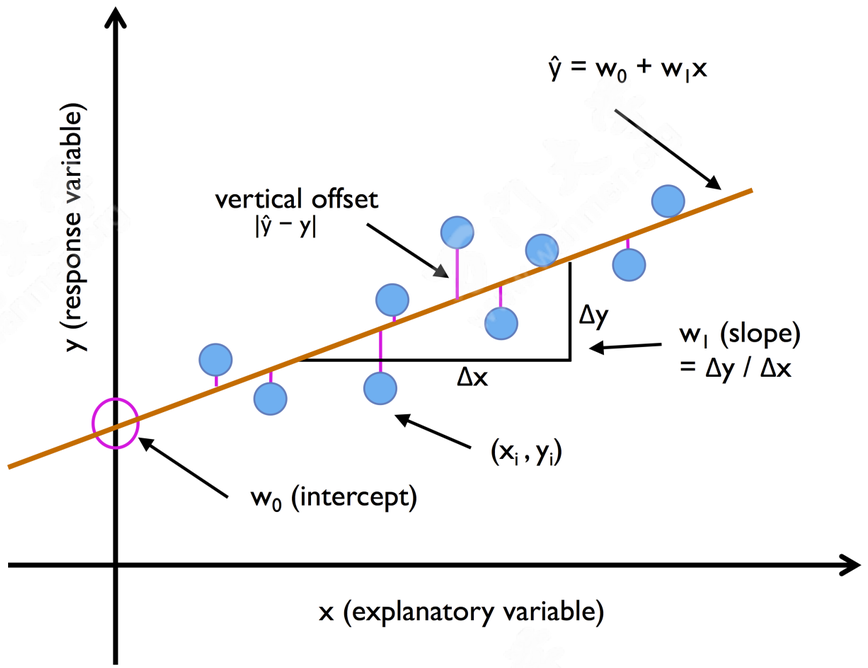
缺点：房屋价格变动时很难维护

CRUISER

换个思路，建立一个模型



$$P = w1x1 + w2x2 + w3x3 + b$$



CRUISER

用模型来预测

房屋价格是一个整体

房间数量 房屋大小等等 都只是一个个特征(Feature)

```python
def estimate_house_sales_price(num_of_bedrooms, sqft, neighborhood):
 price = 0

 # a little pinch of this
 price += num_of_bedrooms * .841231951398213

 # and a big pinch of that
 price += sqft * 1231.1231231

 # maybe a handful of this
 price += neighborhood * 2.3242341421

 # and finally, just a little extra salt for good measure
 price += 201.23432095

 return price
```

在线性模型里， 这些数字就是**权重**

找出对于每间房子都适用的权重，我们就能预测所有的房价

所以问题来了

对于整个数据集，我们该怎么找到权重？

CRUISER

聪明办法： 解线性方程组！

m 个cases，N 个数据

Math Quiz #1 - Teacher's Answer Key

	A	B	C			A	B	C		
1)	$2W_1$	$4W_2$	$5W_3$	=	3	5)	6	2	2	=	10
2)	$5W_1$	$2W_2$	$8W_3$	=	2	6)	3	1	1	=	2
3)	$2W_1$	$2W_2$	$1W_3$	=	3	7)	5	3	4	=	11
4)	$4W_1$	$2W_2$	$2W_3$	=	6	8)	1	8	1	=	7

如果数据变到成千上万.... $\begin{cases} 2x+3y=1 \\ 3x+y=0 \end{cases}$

CRUISER

笨办法：监督学习

如果能让计算机找出实现上述函数功能的办法，这样岂不更好？

只要返回的房价数字正确，谁会在乎函数具体干了些什么呢？

```python
def estimate_house_sales_price(num_of_bedrooms, sqft, neighborhood):
  price = <computer, plz do some math for me>

  return price
```

监督学习是一种归纳法！

CRUISER

第一步，我很蠢

把每个权重都设置成1

```python
def estimate_house_sales_price(num_of_bedrooms, sqft, neighborhood):
  price = 0

  # a little pinch of this
  price += num_of_bedrooms * 1.0

  # and a big pinch of that
  price += sqft * 1.0

  # maybe a handful of this
  price += neighborhood * 1.0

  # and finally, just a little extra salt for good measure
  price += 1.0

  return price
```

$w_1=1 \quad w_2=1 \quad w_3=1 \quad b=0$

$y=w_1x_1+w_2x_2+w_3x_3+b$

CRUISER

第二步，我有多蠢

将每栋房产带入你的函数运算，检验估算值与正确价格的偏离程度：

Bedrooms	Sq. feet	Neighborhood	Sale price	My Guess
3	2000	Normaltown	$250,000	$178,000
2	800	Hipsterton	$300,000	$371,000
2	850	Normaltown	$150,000	$148,000
1	550	Normaltown	$78,000	$101,000
4	2000	Skid Row	$150,000	$121,000

例如：上表中第一套房产实际成交价为25万美元，你的函数估价为17.8万，这一套房产你就差了7.2万。
将你的数据集中的每套房产估价偏离值平方后求和。假设数据集中有500套房产交易，估价偏离值平方求和总计为86,123,373美元。这就反映了你的函数现在的“正确”程度。
现在，将总计值除以500，得到每套房产的估价偏离平均值。将这个平均误差值称为你函数的代价。
如果你能调整权重使得这个代价变为0，你的函数就完美了。它意味着，根据输入的数据，你的程序对每一笔房产交易的估价都是分毫不差。而这就是我们的目标——尝试不同的权重值以使代价尽可能的低。

所以...怎么变聪明？

刚刚的代价函数

$$\text{Cost} = \frac{\sum_{i=1}^{500} \left(\text{MyGuess}(i) - \text{RealAnswer}(i)\right)^2}{500 \cdot 2}$$

对上面的式子进行改写，
J(θ) 表示的是当前权重值对应的代价

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

CRUISER

第三步，暴力穷举

不断重复第二部，**尝试所有可能的权重值组合**。
哪一个组合使得代价最接近于0，它就是你要使用的，
你只要找到了这样的组合，问题就得到了解决!

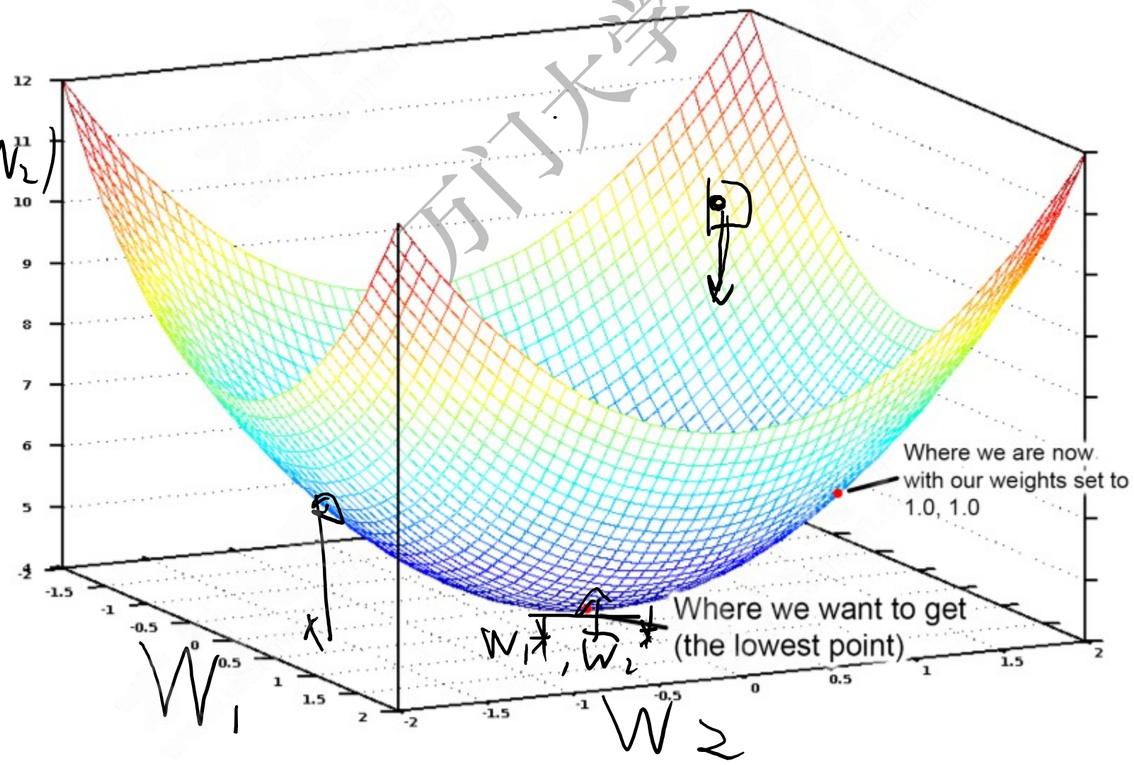
Bedrooms	Sq. feet	Neighborhood	Sale price	My Guess
3	2000	Normaltown	$250,000	$178,000
2	800	Hipsterton	$300,000	$371,000
2	850	Normaltown	$150,000	$148,000
1	550	Normaltown	$78,000	$101,000
4	2000	Skid Row	$150,000	$121,000

太麻烦了！ 我需要一个个的试验吗？

CRUISER

梯度下降：更聪明的变聪明方法

如果将所有赋给卧室数和面积的可能权重值以图形形式显示，
我们会得到类似下图的图表



图中蓝色的最低点就是代价最低的地方——即我们的程序偏离最小。最高点意味着偏离最大。
所以，如果我们能找到一组权重值带领我们到达图中的最低点，我们就找到了答案！

使用梯度，一种不那么蠢的办法

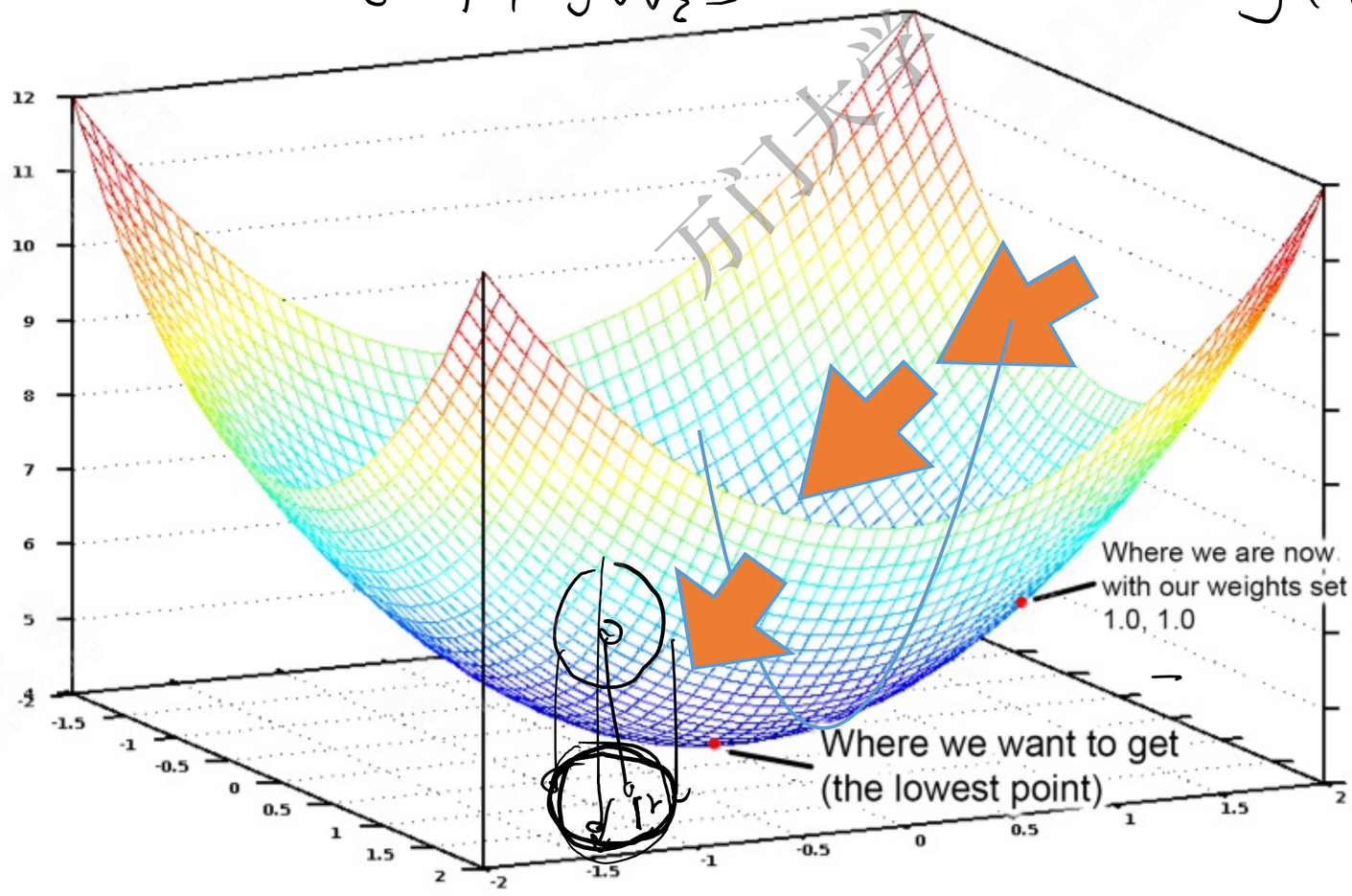
$$J(w_1, w_2) = -\left(\frac{\partial J}{\partial w_1}, \frac{\partial J}{\partial w_2}\right)$$

$J(\theta) \rightarrow J(\theta^+)$

$J(\theta) \rightarrow J(\theta^-)$

$J(\theta')$

$J(\theta^+) = \min J(\theta')$



CRUISER

什么是机器学习？

假如我们刚刚的程序里没有类似“面积”和“卧室数”这样的参数，而是接受了一组数字。

假设每个数字代表的是你车顶安装的摄像头捕捉的画面中的一个像素，再将预测的输出不称为“价格”而是叫做“方向盘转动度数”，

这样你就得到了一个程序可以自动操纵你的汽车了！

还有错误，怎么回事？需不需要追求完美？

The unknown under beneath….

机器学习总会成功吗?不!

CRUISER

数据是机器学习的能源

- 大数据时代1.0:数据的积累和呈现
- 大数据时代2.0:机器学习：用经验数据预测未来

- DT时代数据即财富
- 机器学习给数据赋予价值

CRUISER

数据标注是第一生产力