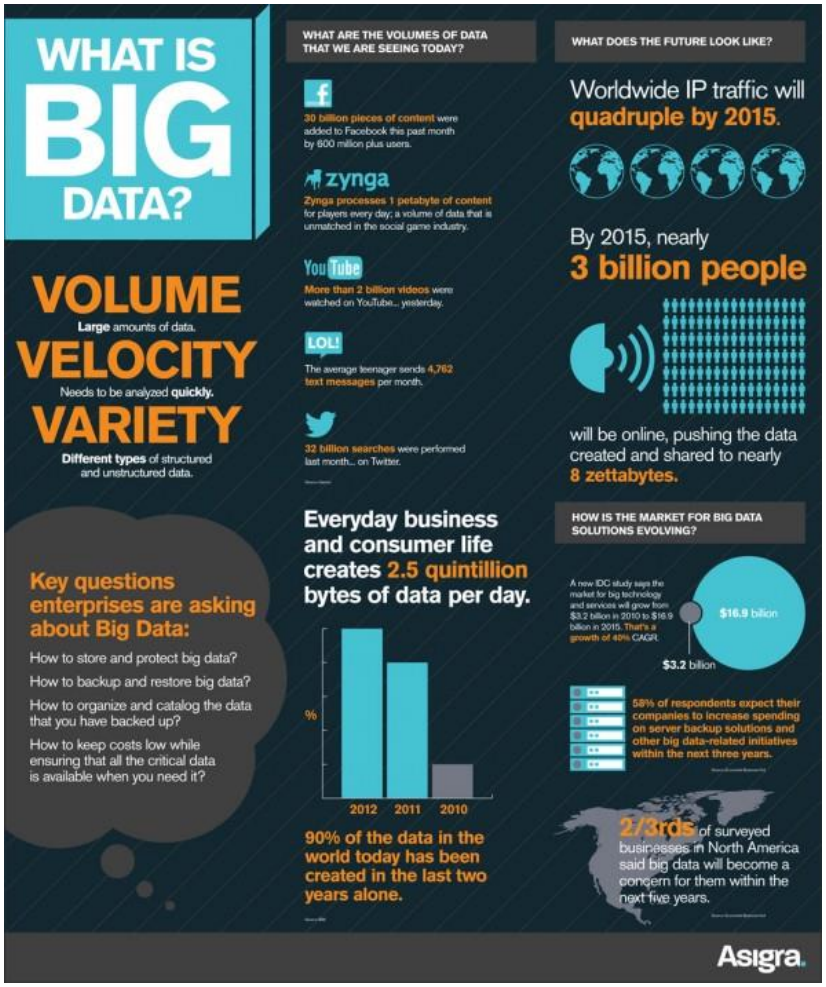
第一讲
机器学习简介

夏睿
计算机科学与工程学院
南京理工大学
rxia@njust.edu.cn
http://www.nustm.cn/member/rxia

大数据时代



- 每天大约有50亿tweets被传送，也就是说每秒钟有超过5700条tweets。
- Facebook上有超过11.5亿的活跃用户在互动。
- 超过50亿人群正在打电话、发短信、发微博、在手机上浏览网站。
- VISA每天大约有172,800,000的信用卡业务需要处理。
- United Parcel Service（美国联合包裹服务）每天能收到大约395,000,00包裹。
- RFID (射频识别系统产生高达常规条形码系统数据的1,000倍。

什么是机器学习?

数据

↓

机器学习

↓

认知

机器学习的定义

- Arthur Samuel (1959)定义机器学习是一个

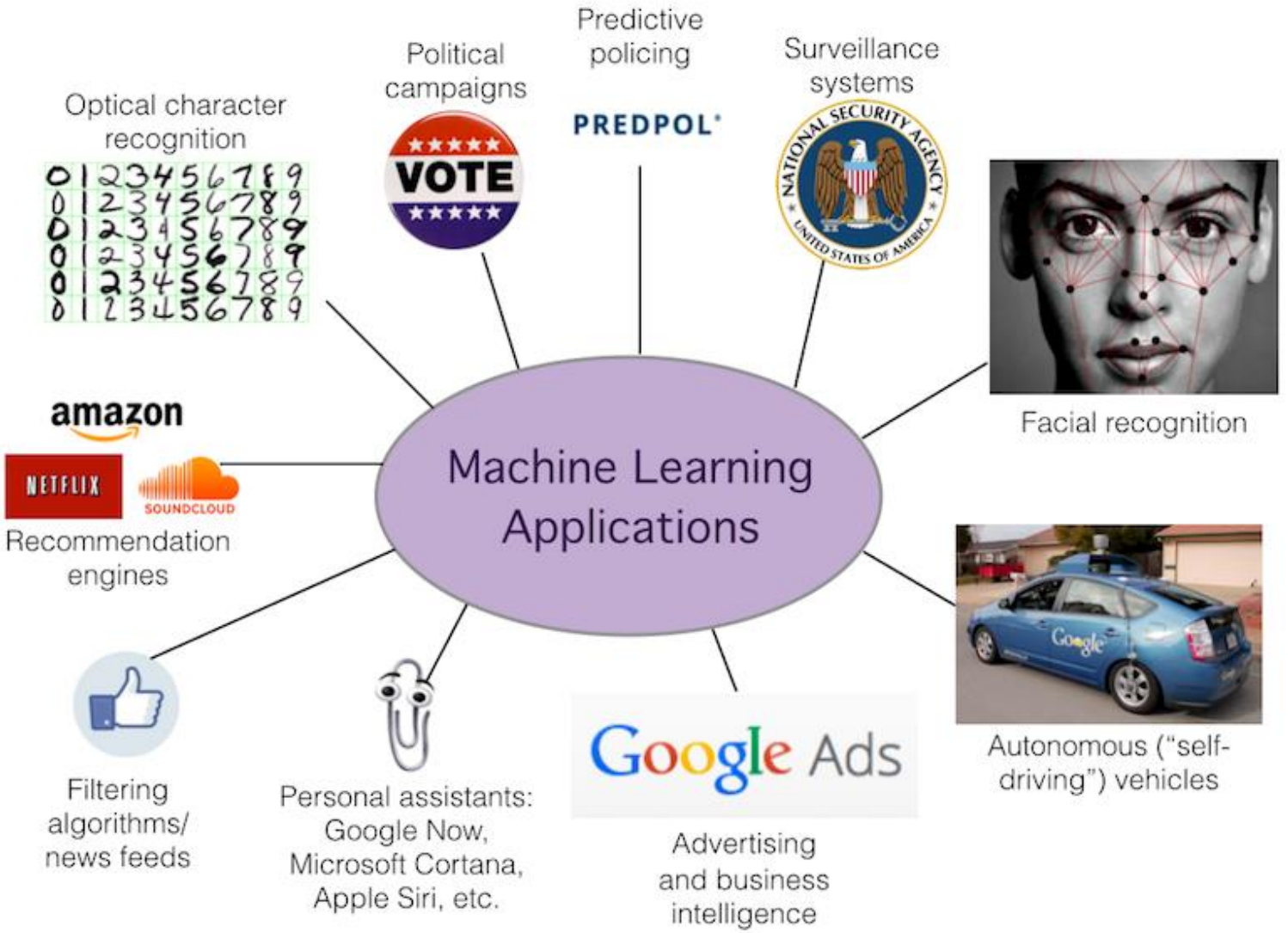
  “无需明确编程而给予计算机学习能力的研究领域。”

- Tom M. Mitchell (1997)提供了一个更广泛的引证，更规范的定义是：

  “如果它在任务T中的性能被P测试，随着经验E的增加而提高，则被认为是计算机程序是从经验E中学习某些类别的任务T和性能度量P。”

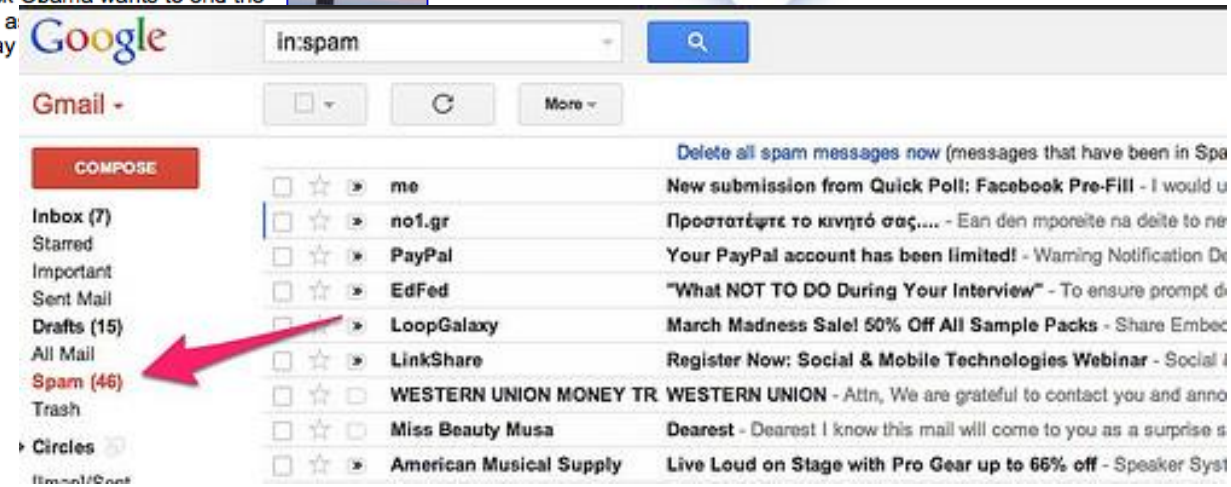
机器（电脑）了解我们（人类）可以学习什么（从数据）？

机器学习的应用


Optical character recognition
Political campaigns
VOTE
Predictive policing
PREDPOL
Surveillance systems
NATIONAL SECURITY AGENCY
UNITED STATES OF AMERICA
Facial recognition
amazon
NETFLIX
SOUNDCLOUD
Recommendation engines
Machine Learning Applications
Filtering algorithms/ news feeds
Personal assistants: Google Now, Microsoft Cortana, Apple Siri, etc.
Google Ads
Advertising and business intelligence
Google
Autonomous ("self-driving") vehicles

文本分类


Google news
Search News
Search the Web
U.S.
Top Stories
Starred
World
U.S.
Business
Sci/Tech
Entertainment
Sports
Health
Spotlight
Most Popular
› All news
Headlines
Top Stories
Obama's budget proposal draws rapid fire
Richard Wolf, Steve Hebert - 2 hours ago
WASHINGTON - President Obama's proposed $3.8 trillion budget ran into immediate trouble in Congress on Monday among lawmakers who said it tries to do too much while cutting the deficit too little.
The Hindu
Video: Obama Budget Would Create Highest-ever Deficit
The Associated Press
Wealthy Face Tax Increase
Wall Street Journal
Milwaukee Journal Sentinel - Bradenton Herald - ABC News - Daily Caller
all 4,573 news articles »
Email this story
Md. stands to gain despite Obama budget cuts
Baltimore Sun - Paul West - 1 hour ago
WASHINGTON - - President Barack Obama wants to end the nation's troubled program to return a
but NASA officials indicated Monday


Google
in:spam
Gmail -
More
COMPOSE
Inbox (7)
Starred
Important
Sent Mail
Drafts (15)
All Mail
Spam (46)
Trash
Circles
Delete all spam messages now (messages that have been in Spa
me
New submission from Quick Poll: Facebook Pre-Fill - I would u
no1.gr
Προστατέψτε το κινητό σας.... - Ean den mporeite na delte to ne
PayPal
Your PayPal account has been limited! - Warning Notification De
EdFed
"What NOT TO DO During Your Interview" - To ensure prompt d
LoopGalaxy
March Madness Sale! 50% Off All Sample Packs - Share Embed
LinkShare
Register Now: Social & Mobile Technologies Webinar - Social
WESTERN UNION MONEY TR
WESTERN UNION - Attn, We are grateful to contact you and anno
Miss Beauty Musa
Dearest - Dearest I know this mail will come to you as a surprise s
American Musical Supply
Live Loud on Stage with Pro Gear up to 66% off - Speaker Syst

情感分析和观点挖掘


twitrratr
TRACKING OPINIONS ON TWITTER
SEARCH
SEARCHED TERM
socialmouths
POSITIVE TWEETS
17
NEUTRAL TWEETS
165
NEGATIVE TWEETS
0
TOTAL TWEETS
182
9.34% POSITIVE
90.66% NEUTRAL
0.00% NEGATIVE
None Found
Ads by Google
View ads about:

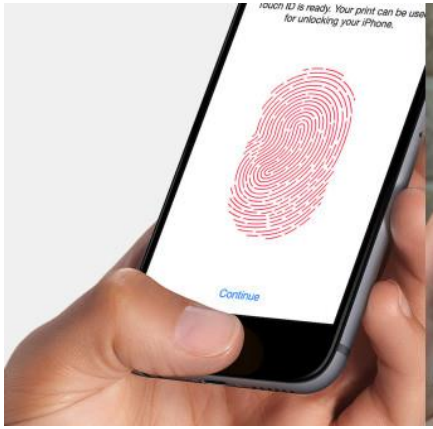

100 %
50%
0 %
SENTIMENT


商品详情
包装和参数
累计评价1624
月成交记录1007件
给我推荐
与描述相符
4.8
大家都写道
行货正品(419)
性价比很高 (28)
包装不错哦(21)
很轻便(13)
系统流畅(12)
外观靓丽(12)
屏幕不错(9)
通话质里好(3)
外设不错(2)
性能好(1)
电池耐用(1)
通话质里一般(8)
外观一般(8)
电池一般(6)
屏幕一般(5)
不清楚是否行货(5)
系统不流畅(2)
包装一般(2)
配件一般(1)
性价比一般 (1)
查看追加(65)
有内容评价(10字及以上)
默认
按时间
按信用

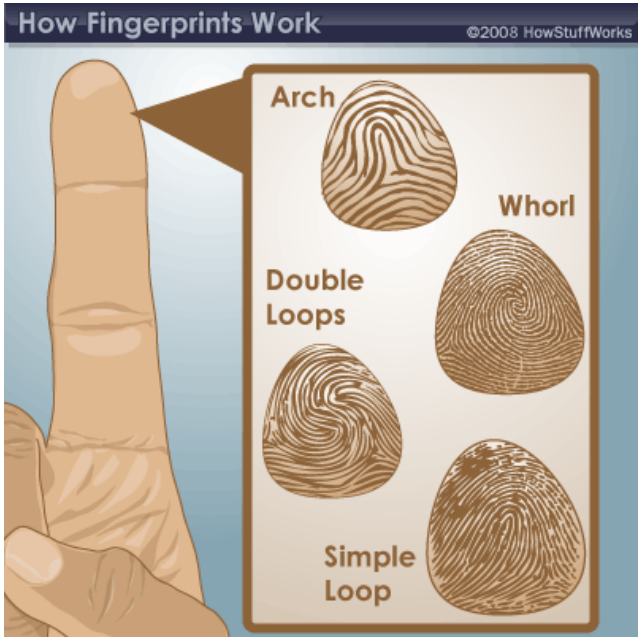
社会媒体分析

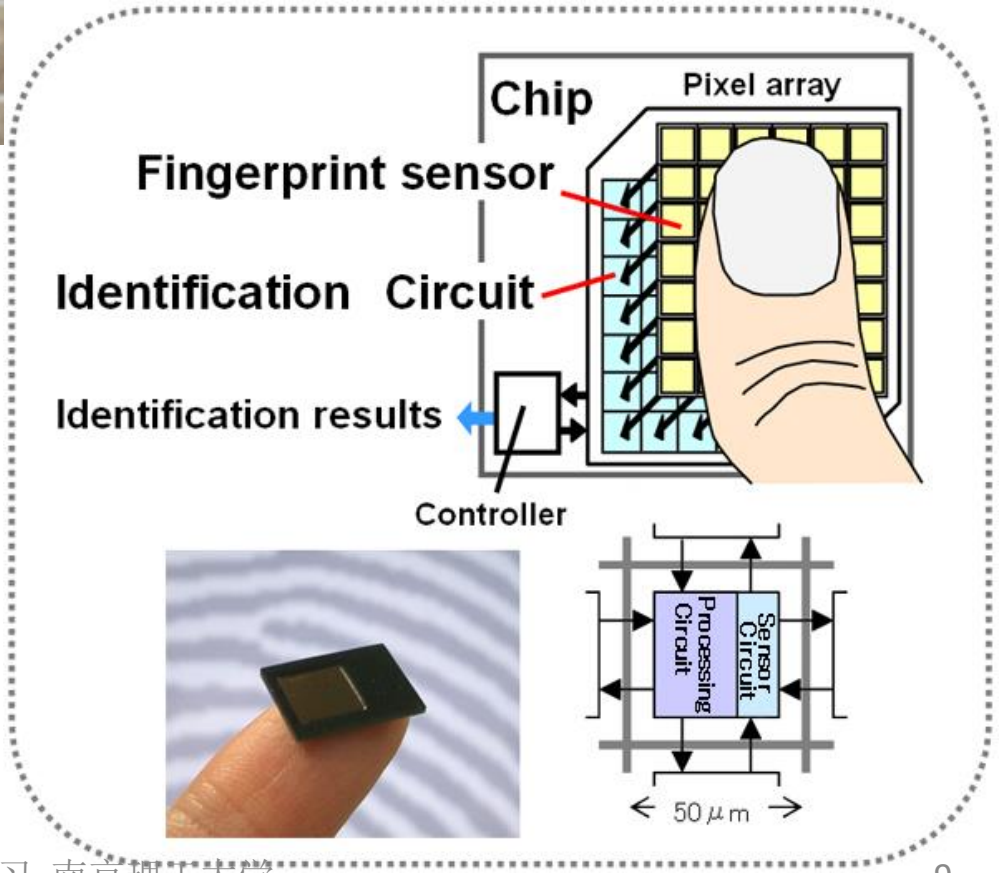

可口可乐 (Coca-Cola)
May 25 2011
Welcome Jennifer
Logout
English
Captive Audience
Mentions
Competition
Tencent Weibo
Influencers
Person
Han Han
Xiaoming Xu
Chengpeng Li
Xianping Lang
Jinglei Xu
Gender Profanity
Sentiment & Tension
Gender Geotagging
Frequency Analysis
Network Analysis
Census
@theethicalchef
@rcmamarkets
@c3sc
@getgood
@swpeast
@grangecardiff
@cardiffcouncil
@ashleygovier
@doughty4csp
@swpolice
@vaughangething
@kevinbrennanmp
@sarulislam
@pontcannaranger
@grangetownplaid
@luke4csp
@lukenicholas87
@fishat85
Network Type
Retweets
Mentions
Show tension
Export Network
Network Metric
Investigate
degree centrality
The degree centrality of a node indicates the importance of the node based on the number of edges connected to it.
Nodes
Size nodes by degree centrality
Filter nodes by degree centrality
Dim
filtered-out nodes and edges
Fewer nodes
More nodes
Node Labels
Screen Name D.C. B.C. C.C.
@cardiffcouncil 1.00 1.00 0.42
@swpolice 0.72 0.02 1.00
@grangetownplaid 0.69 0.11 0.36
@c3sc 0.68 0.02 0.37
@fishat85 0.67 0.02 0.81
@vaughangething 0.67 0.28 0.38
@swpeast 0.66 0.13 0.30
@kevinbrennanmp 0.64 0.21 0.34
@luke4csp 0.58 0.05 0.27
@doughty4csp 0.58 0.10 0.36
@theethicalchef 0.57 0.02 1.00
@rcmamarkets 0.54 0.10 0.24
@pontcannaranger 0.53 0.02 0.32
@lukenicholas87 0.51 0.01 0.22
@ashleygovier 0.51 0.20 0.46

指纹识别


Continue

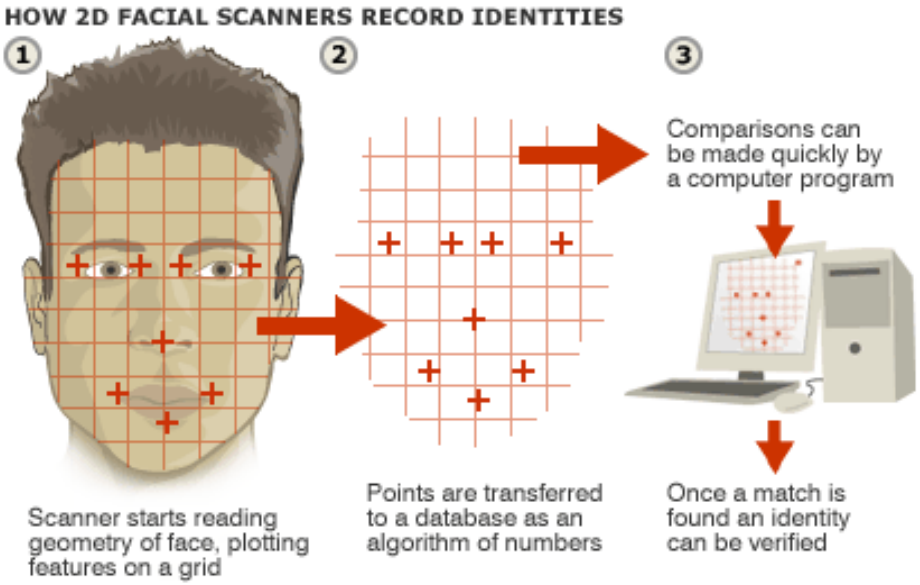

Touch ID or Enter Passcode
Emergency
Cancel


How Fingerprints Work
©2008 HowStuffWorks
Arch
Whorl
Double Loops
Simple Loop


Chip
Pixel array
Fingerprint sensor
Identification Circuit
Identification results
Controller
Processing Circuit
Sensor Circuit
50 μm

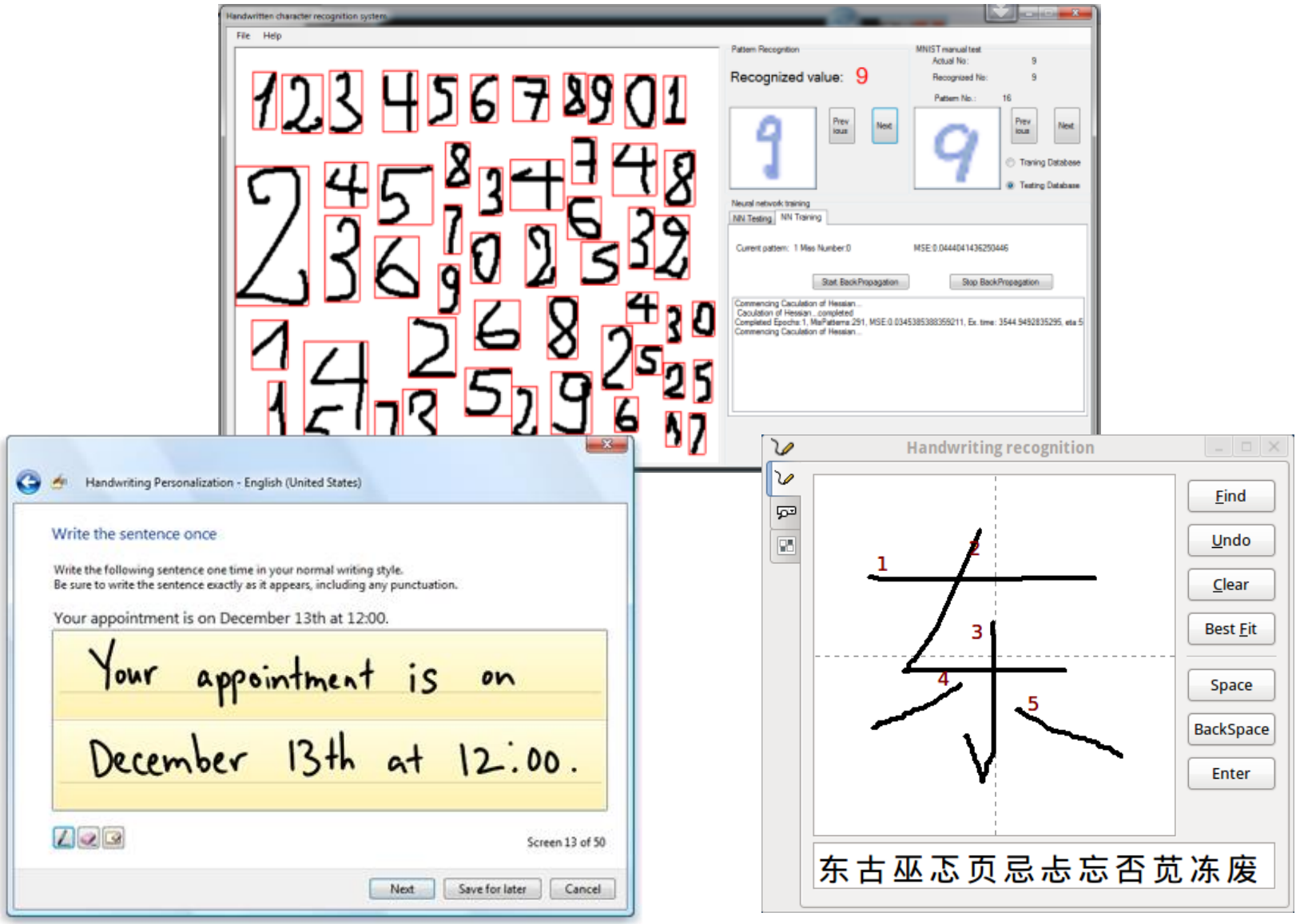
人脸识别


03
52


HOW 2D FACIAL SCANNERS RECORD IDENTITIES
1
2
3
Comparisons can be made quickly by a computer program
Scanner starts reading geometry of face, plotting features on a grid
Points are transferred to a database as an algorithm of numbers
Once a match is found an identity can be verified

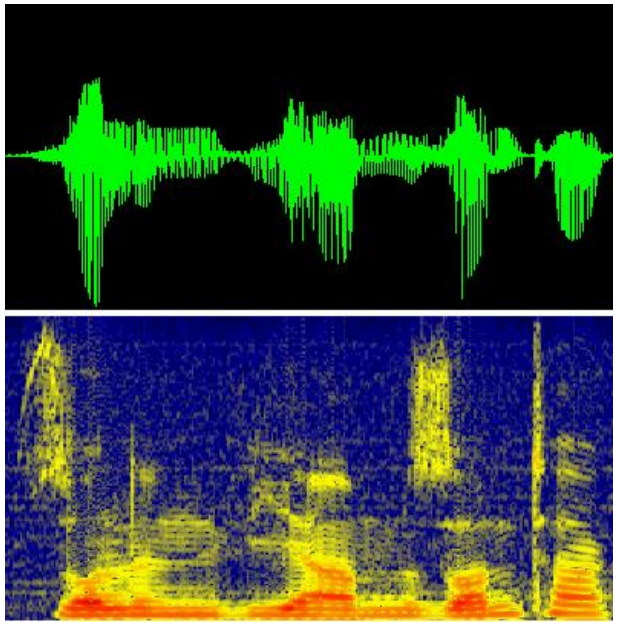

AuthenMetric

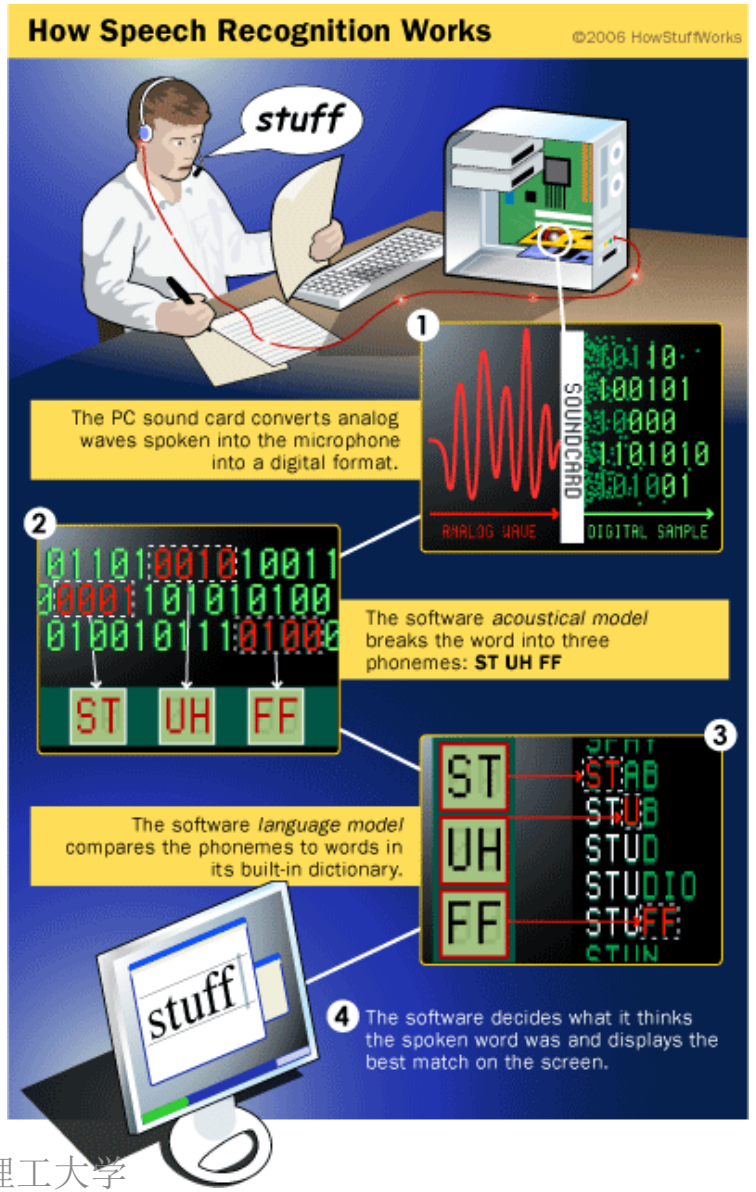
手写字符识别


Handwritten character recognition system
File Help
Pattern Recognition
Recognized value: 9
MNIST manual test
Actual No: 9
Recognized No: 9
Pattern No.: 16
Prev Icon
Next
Training Database
Testing Database
Neural network training
NN Testing
NN Training
Start BackPropagation
Stop BackPropagation
Handwriting Personalization - English (United States)
Write the sentence once
Write the following sentence one time in your normal writing style.
Be sure to write the sentence exactly as it appears, including any punctuation.
Your appointment is on December 13th at 12:00.
Your appointment is on
December 13th at 12:00.
Screen 13 of 50
Next
Save for later
Cancel
Handwriting recognition
Find
Undo
Clear
Best Fit
Space
BackSpace
Enter
1
2
3
4
5
东古巫丕页忌忐忘否苋冻废

语音识别


VOICE RECOGNITION
Speak now
Cancel
Working
Cancel



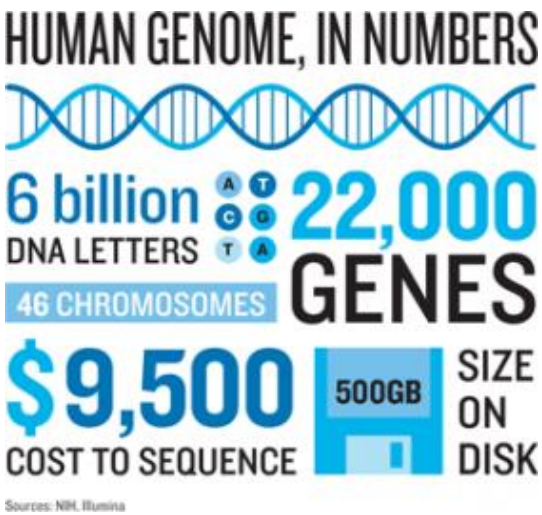

How Speech Recognition Works
©2006 HowStuffWorks
stuff
1
SOUNDCARD
ANALOG WAVE
DIGITAL SAMPLE
The PC sound card converts analog waves spoken into the microphone into a digital format.
2
ST
UH
FF
The software acoustical model breaks the word into three phonemes: ST UH FF
3
ST
UH
FF
STAB
STUB
STUD
STUDIO
STUFF
The software language model compares the phonemes to words in its built-in dictionary.
stuff
4
The software decides what it thinks the spoken word was and displays the best match on the screen.

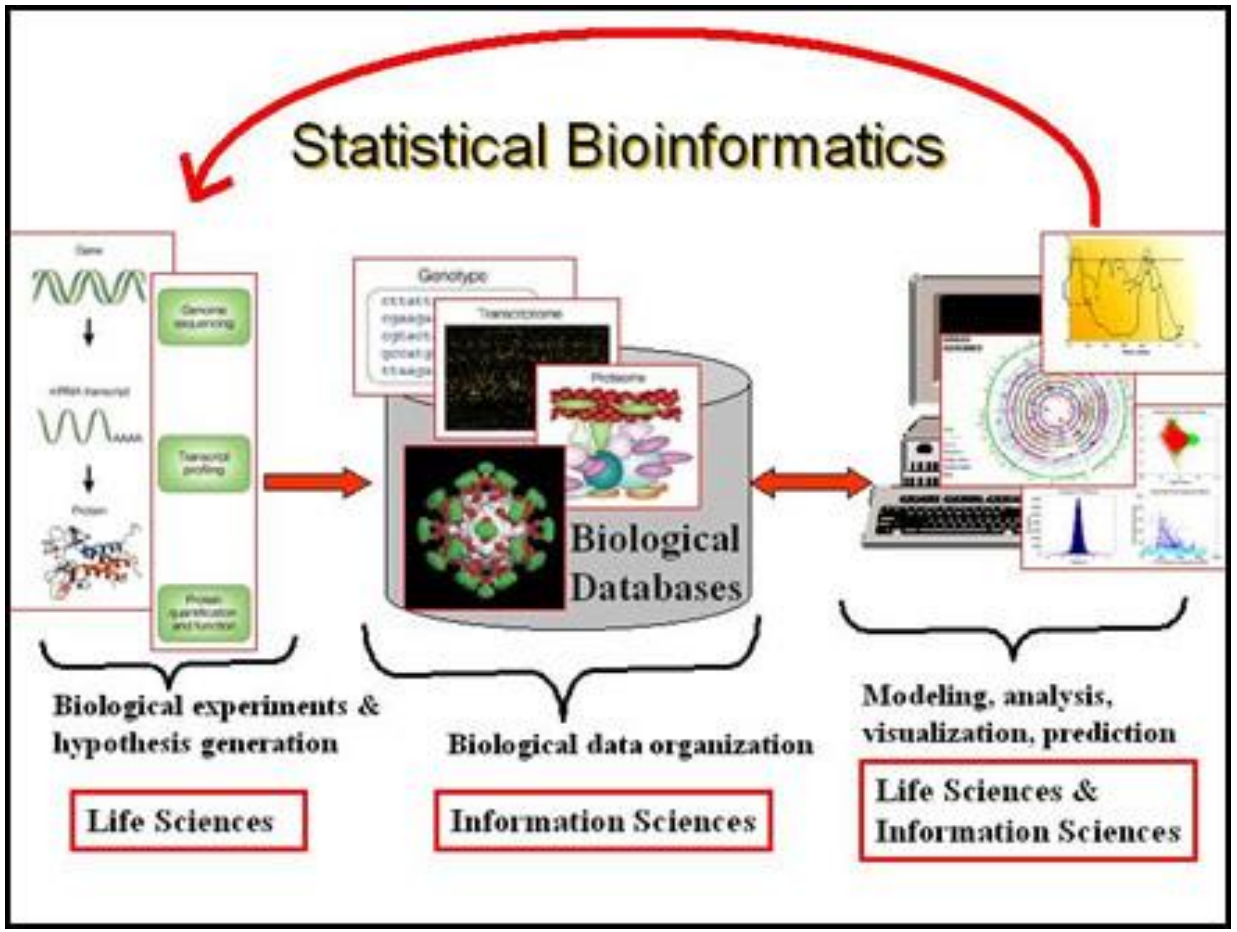
图片识别

https://www.zhihu.com/question/51020471

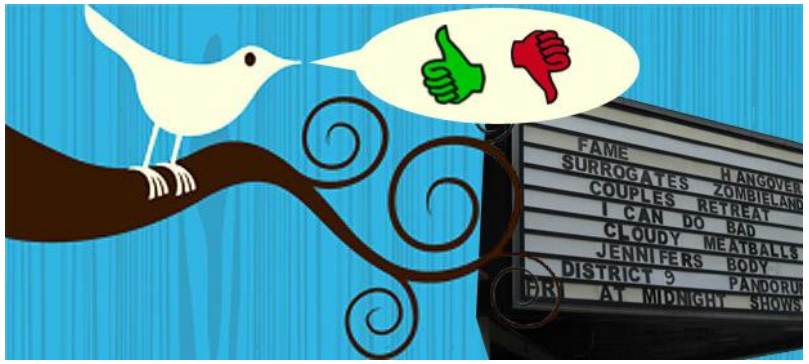
生物信息学

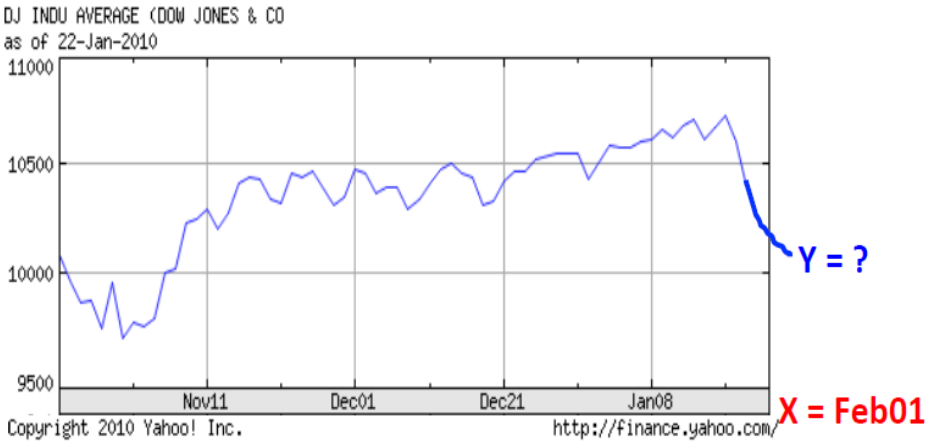

bio informatics

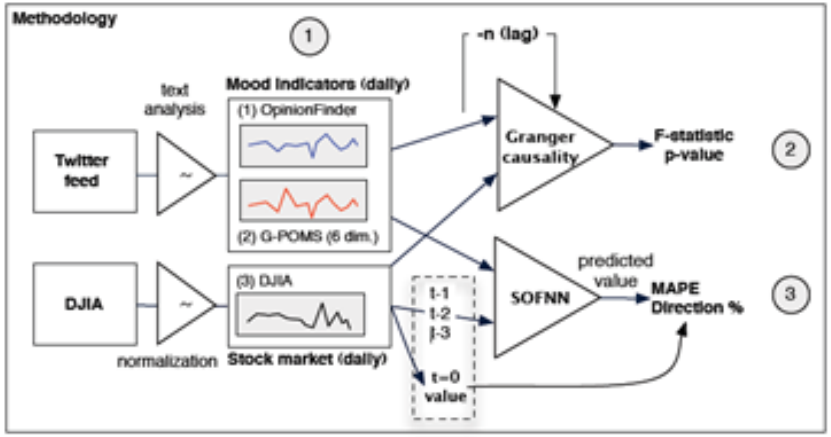

HUMAN GENOME, IN NUMBERS
6 billion
DNA LETTERS
22,000
GENES
46 CHROMOSOMES
$9,500
COST TO SEQUENCE
500GB
SIZE ON DISK
Sources: NIH, Illumina

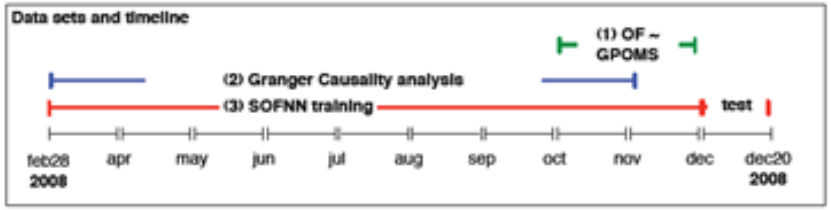

Statistical Bioinformatics
Genotype
Biological Databases
Biological experiments & hypothesis generation
Biological data organization
Modeling, analysis, visualization, prediction
Life Sciences
Information Sciences
Life Sciences & Information Sciences

股市预测


FAME
SURROGATES
COUPLES
I CAN DO BAD
CLOUDY MEATBALLS
JENNIFERS BODY
DISTRICT 9
AT MIDNIGHT


DJ INDU AVERAGE (DOW JONES & CO
as of 22-Jan-2010
11000
10500
10000
9500
Nov11
Dec01
Dec21
Jan08
Y = ?
X = Feb01
Copyright 2010 Yahoo! Inc.
http://finance.yahoo.com/


Methodology
1
-n (lag)
text analysis
Mood indicators (daily)
(1) OpinionFinder
(2) G-POMS (6 dim.)
Twitter feed
Granger causality
F-statistic p-value
2
DJIA
(3) DJIA
normalization
Stock market (daily)
t-1
t-2
t-3
t-0 value
SOFNN
predicted value
MAPE
Direction %
3


Data sets and timeline
(1) OF ~ GPOMS
(2) Granger Causality analysis
(3) SOFNN training
test
feb28 2008
apr
may
jun
jul
aug
sep
oct
nov
dec
dec20 2008

人类-机器 比赛

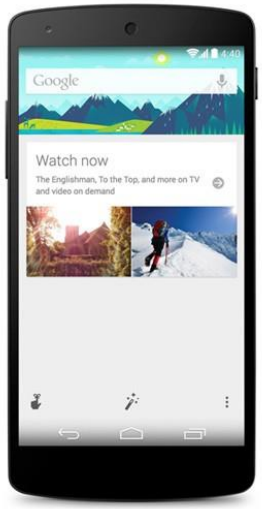

►韓國棋王李世石三
戰連敗
美聯社


PIENSE
THINK
DENKE
PENSER
$300,000
$1,000,000
$200,000
KEN
WATSON
BRAD

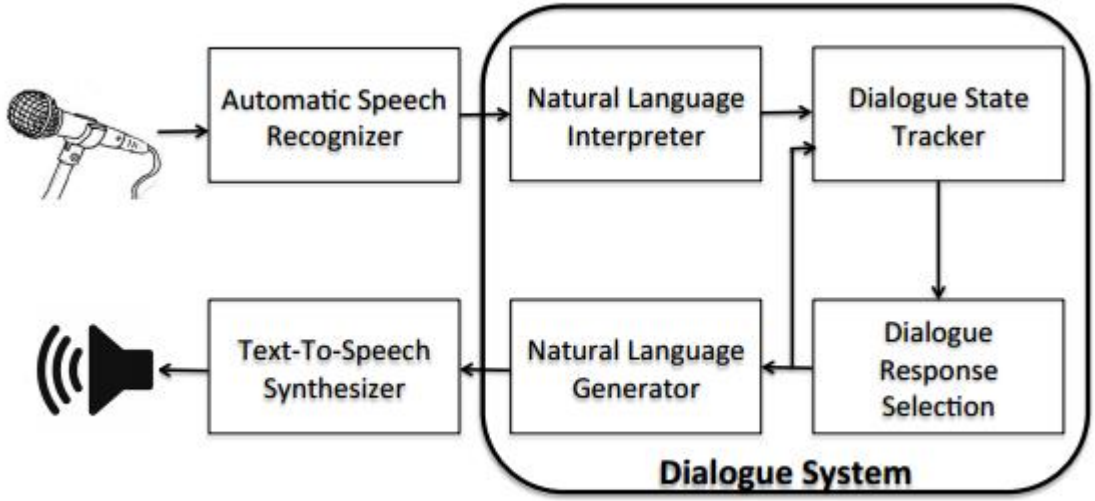
对话系统

Apple Siri

Google Now



Windows Cortana



机器翻译

Google

Translate

Turn off instant translation

Chinese | English | Spanish | Detect language

English | Chinese (Simplified) | Spanish | Translate

【谷歌NMT，见证奇迹的时刻】
微信最近疯传人工智能新进展：谷歌翻译实现重大突破！值得关注和庆贺。mt 几乎无限量的自然带标数据在新技术下，似乎开始发力。报道说：

十年前，我们发布了 Google Translate（谷歌翻译），这项服务背后的核心算法是基于短语的机器翻译（PBMT:Phrase-Based Machine Translation）。

自那时起，机器智能的快速发展已经给我们的语音识别和图像识别能力带来了巨大的提升，但改进机器翻译仍然是一个高难度的目标。

今天，我们宣布发布谷歌神经机器翻译（GNM
Neural Machine Translation）系统，该系统使
先进的训练技术，能够实现到目前为止机器翻
大提升。我们的全部研究结果详情请参阅我们
《Google`s Neural Machine Translation Syst
Bridging the Gap between Human and Machi

[Google NMT, witness the miracle of the moment]
Recent advances in microblogging crazy biography of artificial intelligence: Google translation to achieve a major breakthrough! Worthy of attention and celebration. Mt almost unlimited number of natural standard data in the new technology, it seems to start force. The report says:

Ten years ago, we released Google Translate, the core algorithm behind this service is PBMT: Phrase-Based Machine Translation.

Since then, the rapid development of machine intelligence has given us a great boost in speech recognition and image recognition, but improving machine translation is still a difficult task.

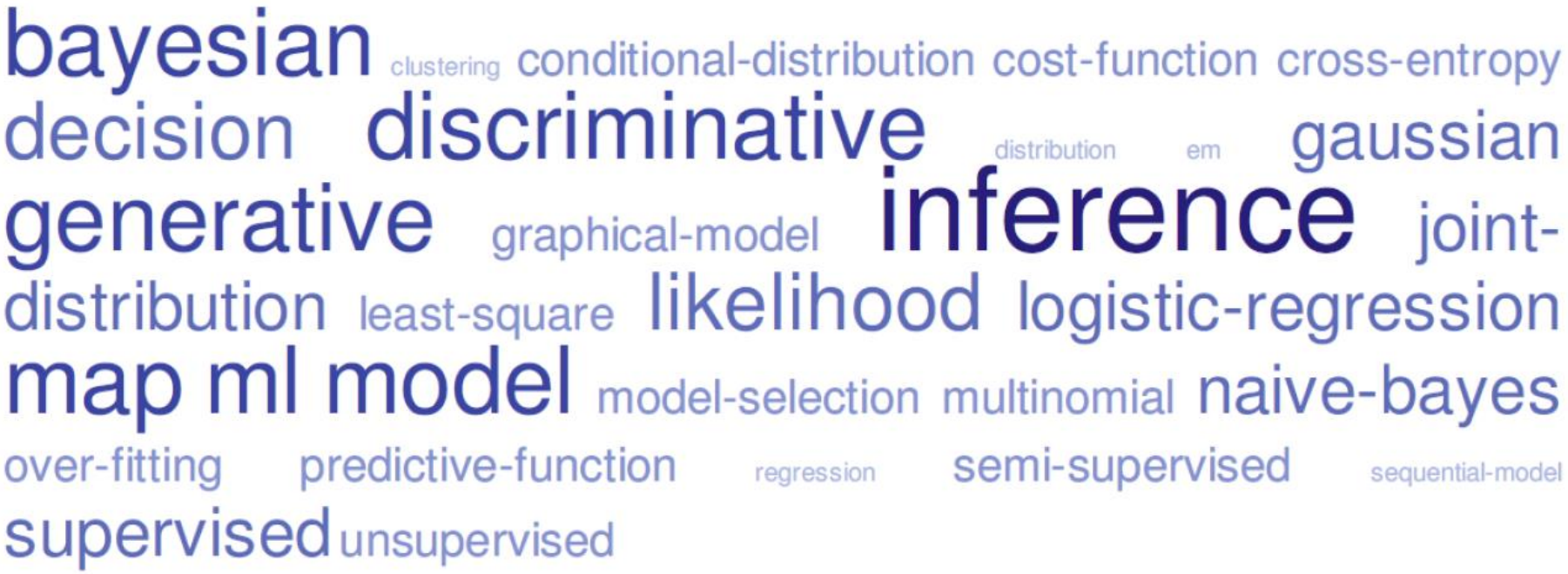
Today, we announced the release of the Google Neural Machine Translation (GNMT) system, which utilizes state-of-the-art training techniques to maximize the quality of machine translation so far. For a full review of our findings, please see our paper "Google`s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation."

无人车

更多应用

- 网络搜索和信息检索
- 自然语言理解
- 机器翻译
- 在图像和视频中定位/追踪/识别目标
- 金融预测和商业智能
- 医疗诊断，媒体图像分析
- 推荐系统
- …

机器学习的关键词/概念



Tag-crowd of Bi-shop's PRML Book

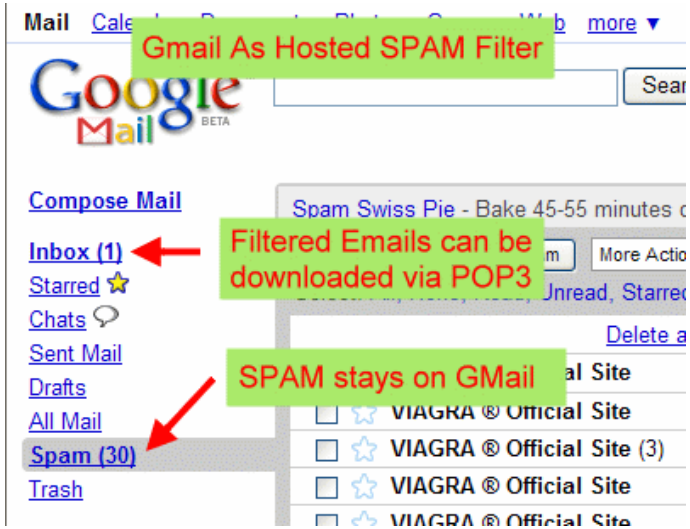
(模式识别与机器学习)

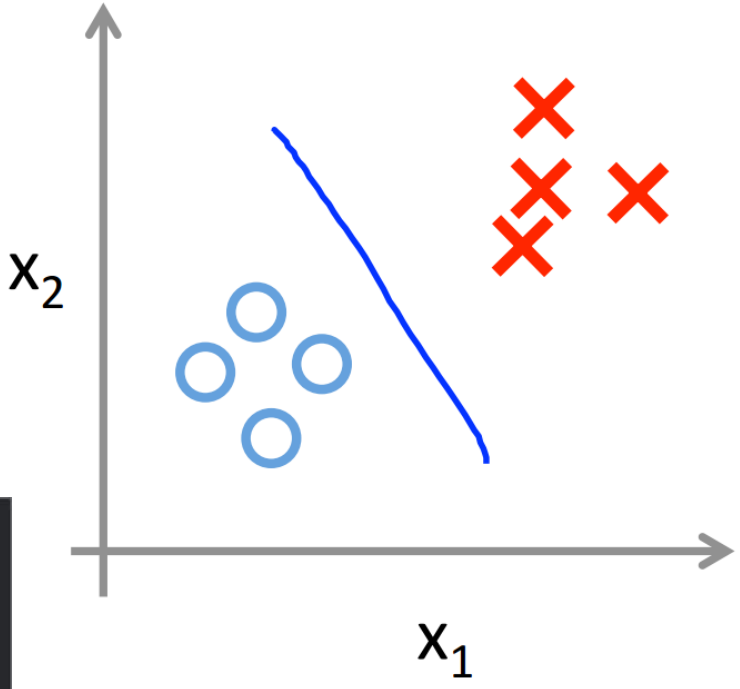
http://research.microsoft.com/en-us/um/people/cmbishop/PRML

机器学习分类

- 监督学习: 给予一个输入对应一个输出结果的例子，然后从新的输入预测输出结果。
  - 分类，回归等
- 非监督学习: 给予一个输入对应一个输出结果的例子，然后从新的输入预测输出结果
  - 聚类，密度估计等.
- 半监督学习
- 集成学习
- 主动学习
- 迁移学习
- 增强学习
- 深度学习
- …

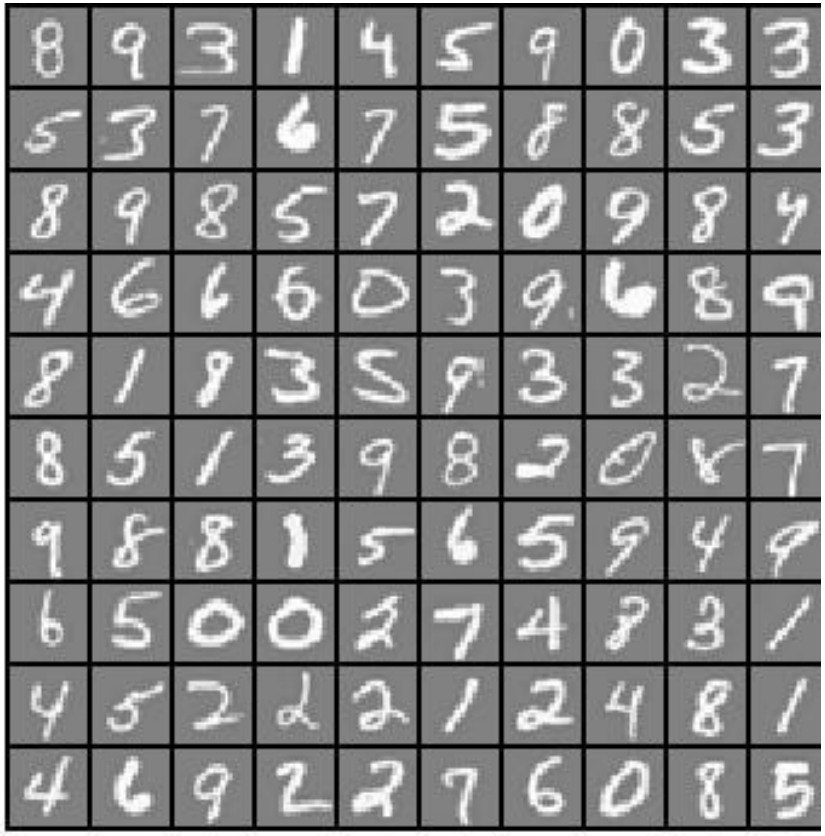
二分类

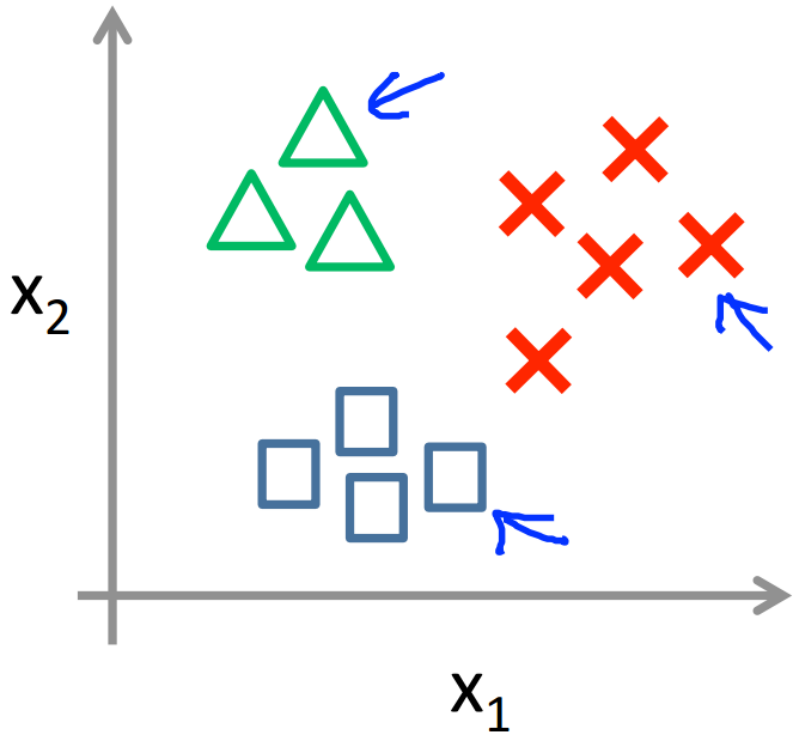

Gmail As Hosted SPAM Filter
Filtered Emails can be downloaded via POP3
SPAM stays on GMail
Compose Mail
Inbox (1)
Starred
Chats
Sent Mail
Drafts
All Mail
Spam (30)
Trash
VIAGRA ® Official Site (3)
VIAGRA ® Official Site


x_2
x_1

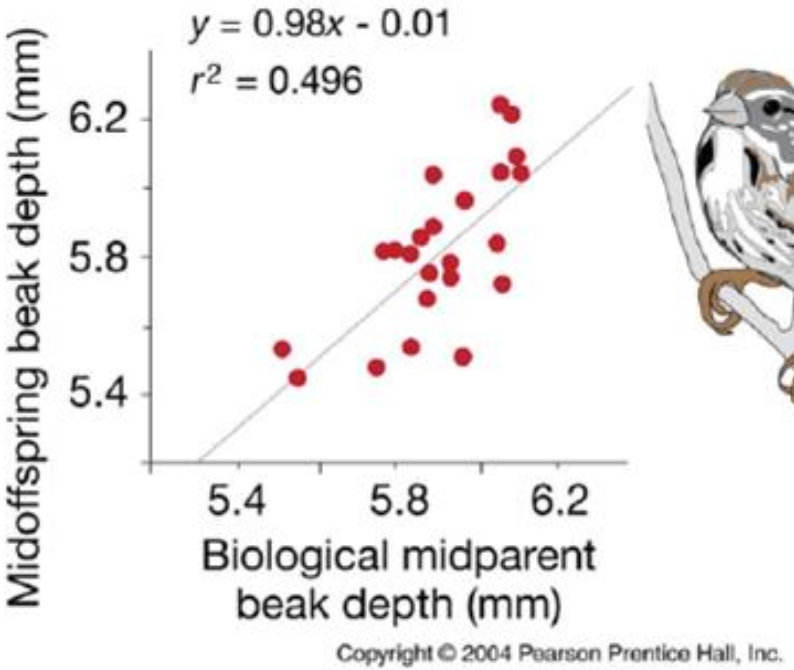

tweetfeel
obama
Search
Try some Twitter trends: Justin Bieber Easter Caddyghini's Desmond
41
33
55%
RT @jeremyscahill: read leaked CIA report on Europe/Afghanistan. CIA sez apathy key 2 US war, obama is great war salesman, sez use Afghan women more 2 sell war
I really love obama!!!!!!! I am just bitter that he beat out Hilary.
#ijustwannaknowhy tavis smiley hate obama so much
RT @jeremyscahill: read leaked CIA report on Europe/Afghanistan. CIA sez apathy key 2 US war, obama is great war salesman, sez use Afghan women more 2 sell war

多分类



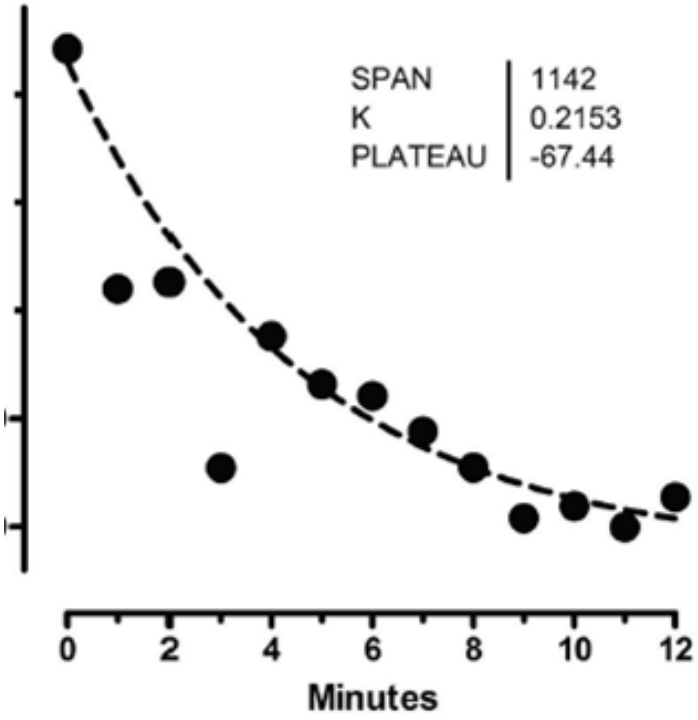


回归



线性回归

非线性回归

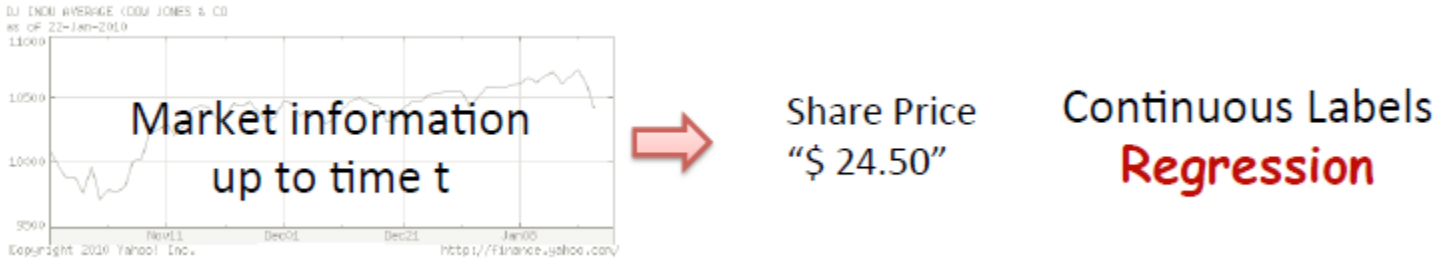


分类 vs. 回归

- 分类


Feature Space $\mathcal{X}$
Words in a document
Label Space $\mathcal{Y}$
"Sports"
"News"
"Science"
...
Discrete Labels
Classification

- 回归

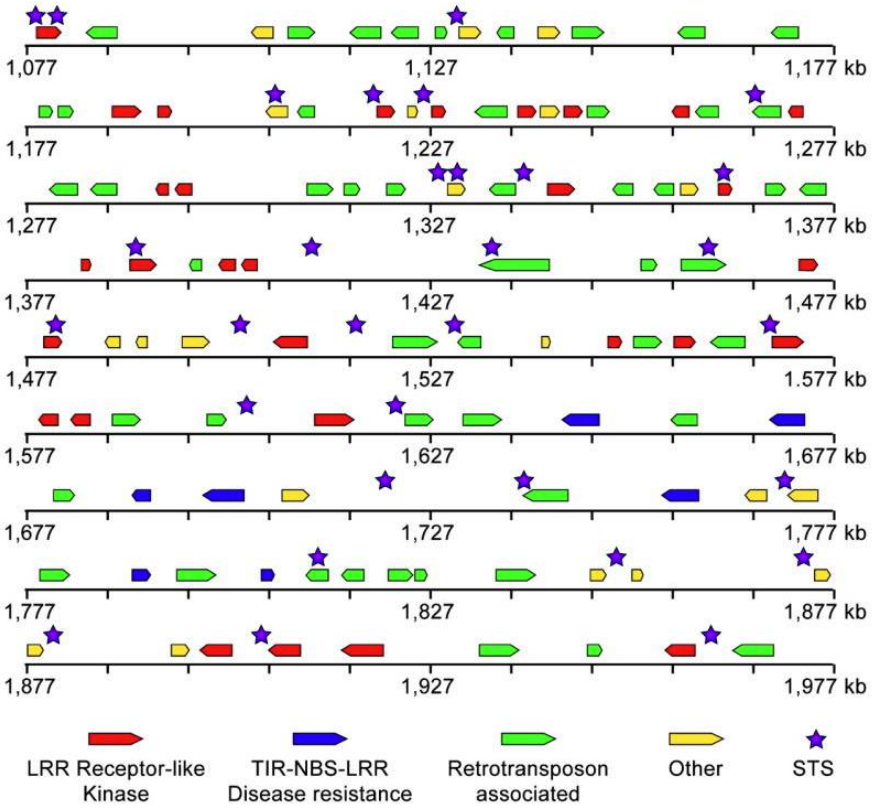

Market information
up to time t
Share Price
"$ 24.50"
Continuous Labels
Regression

聚类

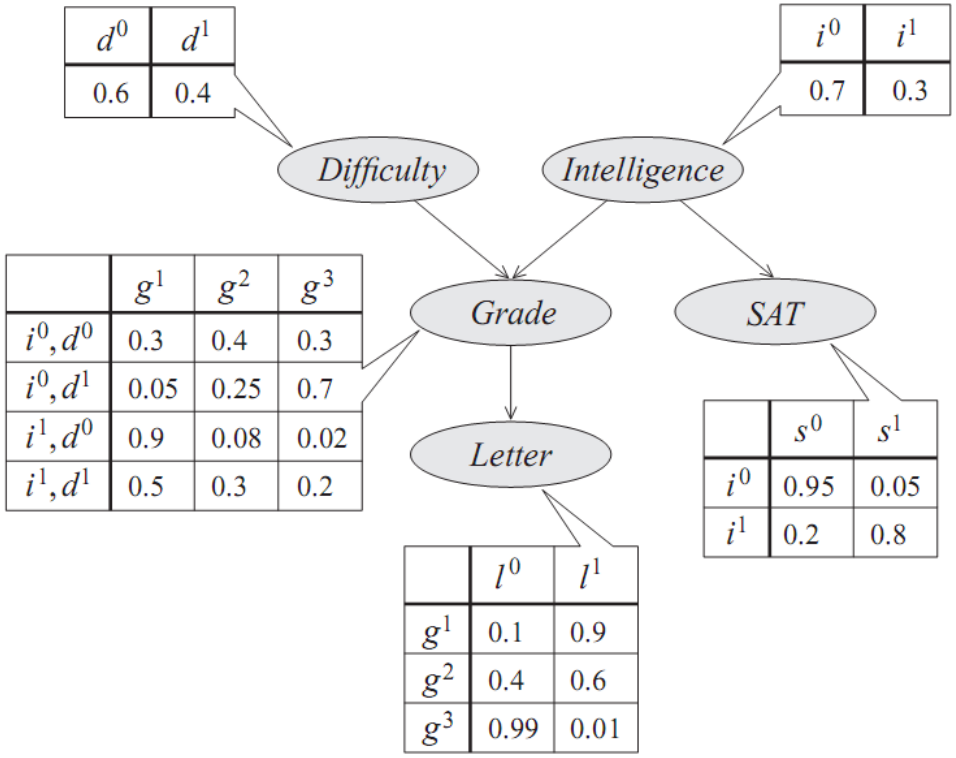
将相似的模型分组（单词，图像...）

序列、图的学习

- Sequential 模型

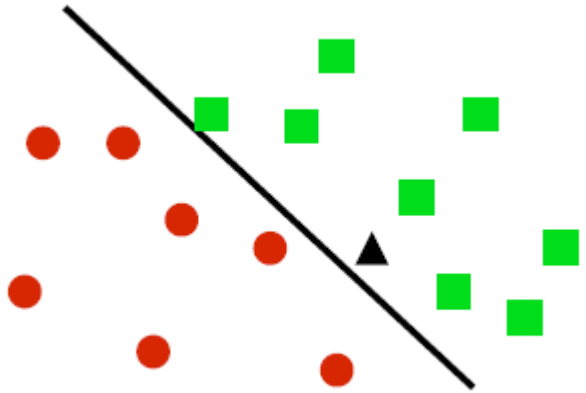


- Graphical 模型



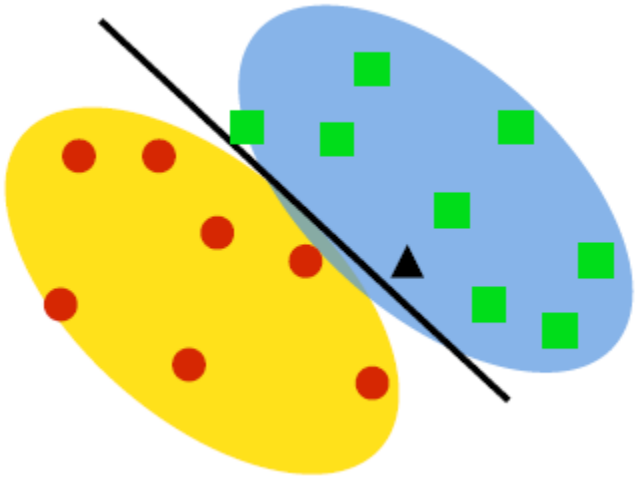
生成模型 vs. 判别模型

- 判别模型



用给定观察p(y|x)标签的后验概率进行建模

- 生成模型



标签和观察 $p(x, y)$ 的联合概率模型，然后用贝叶斯法则$p(y|x) = p(x, y)/p(x)$来预测

机器学习阶段

- 训练阶段(用训练数据)

  你可以从“黄金标准”中提供数据，并通过将输入与预期输出匹配来训练你的模型

- 测试阶段(用测试数据)

  为了估计你的模型被训练的如何，并估计模型属性（如回归的平均误差，分类的准确性）

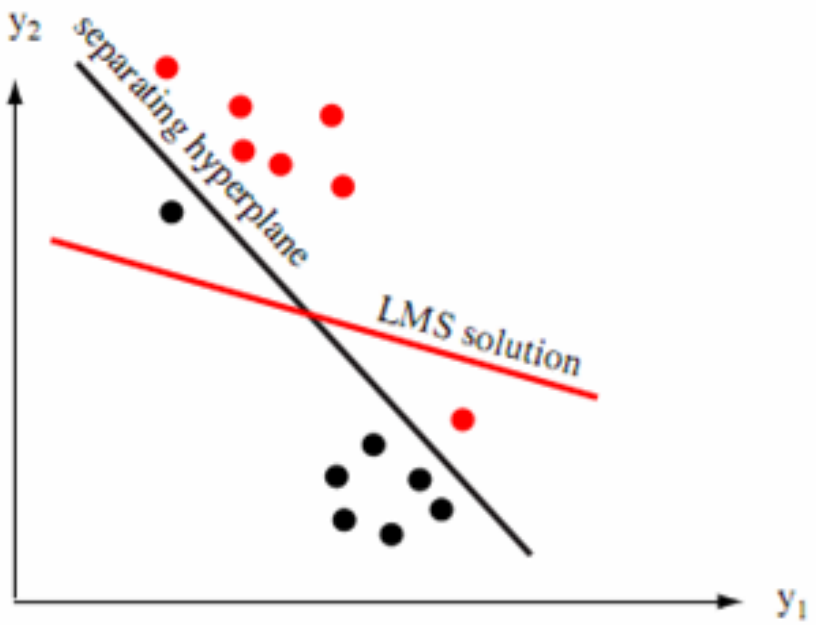
- 应用阶段(为将来的数据)

  现在，你可以将新开发的模型应用于现实世界的数据并获得结果

假设-学习-决策

- 假设
  - 具有（未知）参数（或结构）的数学模型
- 学习 (用来评估参数)
  - 最大似然估计（MLE）, MAP, 贝叶斯估计
  - 代价函数优化
- 决策
  - 贝叶斯决策规则
  - 直接预测函数

学习准则



- 感知机准则
- 最小均方(LMS)
- 最小交叉熵(CE)
- 最大间距准则
- 最大似然
- …

哪一个线性超平面更好?

选择哪个学习准则?

优化方法

- 分析解决方案
- 梯度下降
- 随机梯度下降
- 牛顿法
- 拟牛顿法(BFGS)
- 有限内存BFGS (L-BFGS)
- 共轭梯度
- GIS
- IIS
- …

基本的数学知识

- 微积分（微分，积分）
- 线性代数
- 概率论
- 优化方法

欢迎提问！