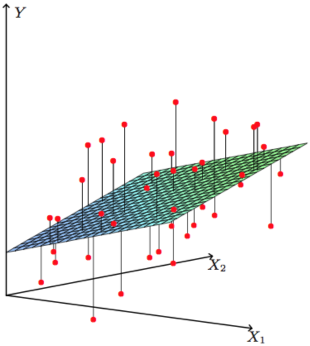
Data Mining and Machine Learning

Kuangnan Fang

Department of Statistics, Xiamen University
Email: xmufkn@xmu.edu.cn



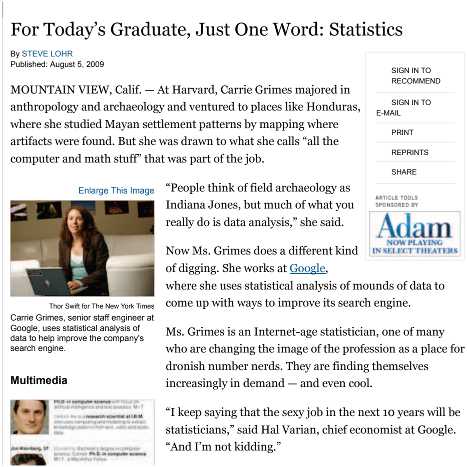
Statistics in the news

How IBM built Watson, its *Jeopardy*-playing supercomputer by Dawn Kawamoto DailyFinance 02/08/2011

Learning from its mistakes According to David Ferrucci (PI of Watson DeepQA technology for IBM Research), Watson's software is wired for more that handling natural language processing.

"It's machine learning allows the computer to become smarter as it tries to answer questions — and to learn as it gets them right or wrong."

不服来战
AlphaGo
樊麾
井山裕太
朴廷桓
柯洁
李世石
阿尔法围棋
（AlphaGo）
一款围棋人工智能程序，由谷歌旗下DeepMind公司开发



For Today's Graduate, Just One Word: Statistics

By STEVE LOHR
Published: August 5, 2009

SIGN IN TO RECOMMEND
SIGN IN TO E-MAIL
PRINT
REPRINTS
SHARE

MOUNTAIN VIEW, Calif. — At Harvard, Carrie Grimes majored in anthropology and archaeology and ventured to places like Honduras, where she studied Mayan settlement patterns by mapping where artifacts were found. But she was drawn to what she calls "all the computer and math stuff" that was part of the job.

Enlarge This Image

Thor Swift for The New York Times
Carrie Grimes, senior staff engineer at Google, uses statistical analysis of data to help improve the company's search engine.

"People think of field archaeology as Indiana Jones, but much of what you really do is data analysis," she said.

Now Ms. Grimes does a different kind of digging. She works at Google, where she uses statistical analysis of mounds of data to come up with ways to improve its search engine.

Ms. Grimes is an Internet-age statistician, one of many who are changing the image of the profession as a place for dronish number nerds. They are finding themselves increasingly in demand — and even cool.

"I keep saying that the sexy job in the next 10 years will be statisticians," said Hal Varian, chief economist at Google. "And I'm not kidding."

Multimedia

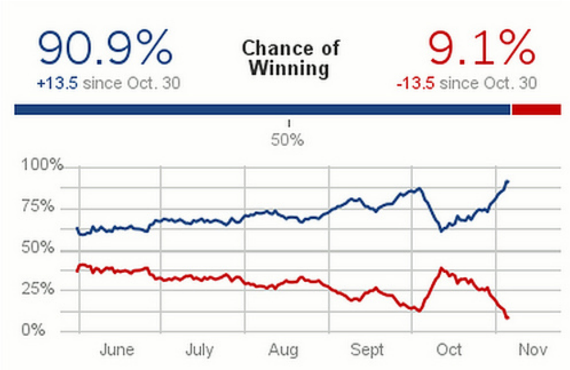
Ph.D. in computer science with focus on artificial intelligence and text analytics: M.I.T.
CAREER: He is a research scientist at I.B.M. who uses computing and modeling to extract knowledge patterns from text, video and audio data.
Jon Kleinberg, 37
EDUCATION: Bachelor's degree in computer science: Cornell. Ph.D. in computer science M.I.T.; a MacArthur Fellow

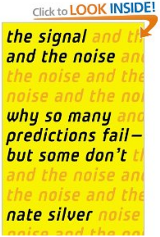
Quote of the Day, New York Times, August 5, 2009

"I keep saying that the sexy job in the next 10 years will be statisticians. And I'm not kidding." — HAL VARIAN, chief economist at Google.

FiveThirtyEight

Nate Silver's Political Calculus



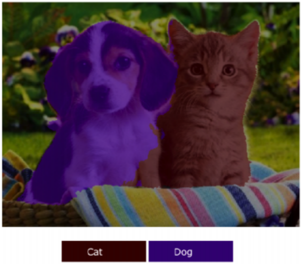

Click to LOOK INSIDE!

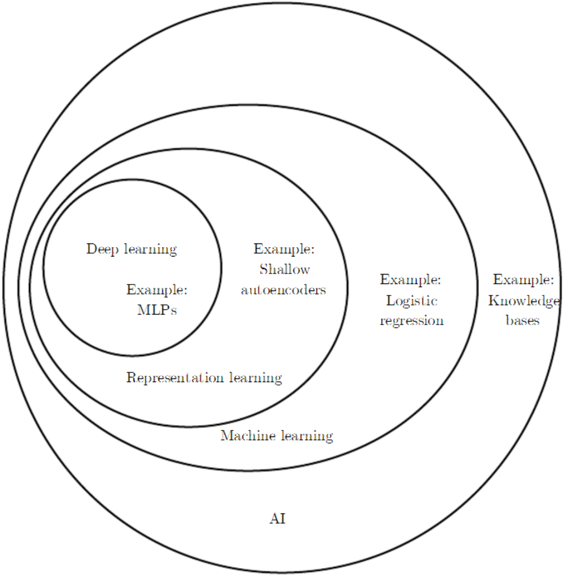
the signal and the noise

why so many predictions fail—but some don't

nate silver

What can I help you with?


Cat
Dog


Deep learning
Example: MLPs
Example: Shallow autoencoders
Representation learning
Example: Logistic regression
Machine learning
Example: Knowledge bases
AI

Statistical Learning versus Machine Learning

- Machine learning arose as a subfield of Artificial Intelligence.
- Statistical learning arose as a subfield of Statistics.
- *There is much overlap* — both fields focus on supervised and unsupervised problems:
  - Machine learning has a greater emphasis on *large scale* applications and *prediction accuracy*.
  - Statistical learning emphasizes *models* and their interpretability, and *precision* and *uncertainty*.
- But the distinction has become more and more blurred, and there is a great deal of "cross-fertilization".
- Machine learning has the upper hand in *Marketing!*

The Supervised Learning Problem

Starting point:

- Outcome measurement Y (also called dependent variable, response, target).
- Vector of p predictor measurements X (also called inputs, regressors, covariates, features, independent variables).
- In the *regression problem*, Y is quantitative (e.g price, blood pressure).
- In the *classification problem*, Y takes values in a finite, unordered set (survived/died, digit 0-9, cancer class of tissue sample).
- We have training data $(x_1, y_1), \ldots, (x_N, y_N)$. These are observations (examples, instances) of these measurements.

Objectives

On the basis of the training data we would like to:

- Accurately predict unseen test cases.
- Understand which inputs affect the outcome, and how.
- Assess the quality of our predictions and inferences.

Philosophy

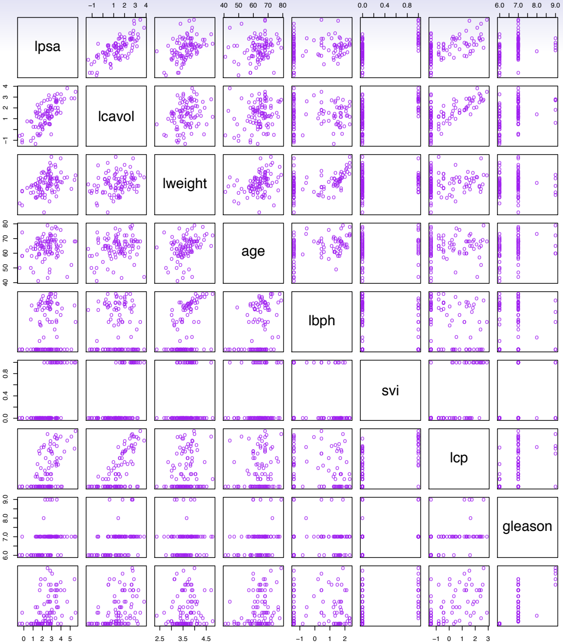
- It is important to understand the ideas behind the various techniques, in order to know how and when to use them.
- One has to understand the simpler methods first, in order to grasp the more sophisticated ones.
- It is important to accurately assess the performance of a method, to know how well or how badly it is working [simpler methods often perform as well as fancier ones!]
- This is an exciting research area, having important applications in science, industry and finance.
- Statistical learning is a fundamental ingredient in the training of a modern *data scientist.*

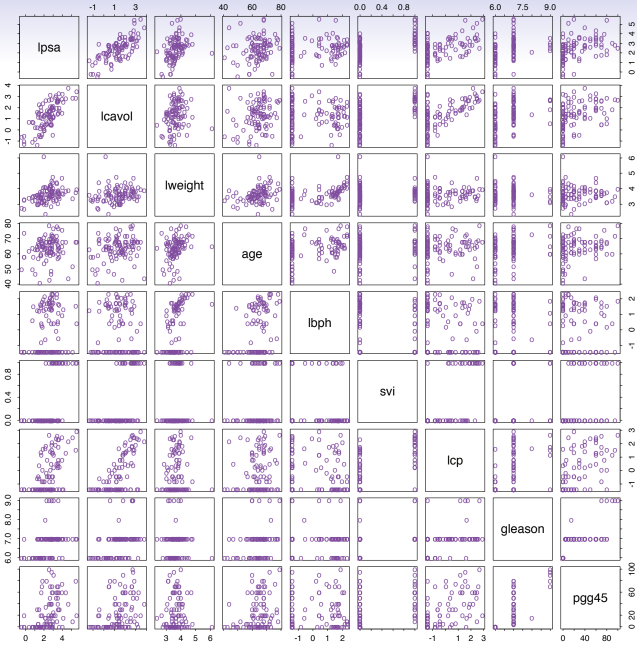
Unsupervised learning

- No outcome variable, just a set of predictors (features) measured on a set of samples.
- objective is more fuzzy — find groups of samples that behave similarly, find features that behave similarly, find linear combinations of features with the most variation.
- difficult to know how well your are doing.
- different from supervised learning, but can be useful as a pre-processing step for supervised learning.

Statistical Learning Problems

- Identify the risk factors for prostate cancer.
- Classify a recorded phoneme based on a log-periodogram.
- Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements.
- Customize an email spam detection system.
- Identify the numbers in a handwritten zip code.
- Classify a tissue sample into one of several cancer classes, based on a gene expression profile.
- Establish the relationship between salary and demographic variables in population survey data.
- Classify the pixels in a LANDSAT image, by usage.

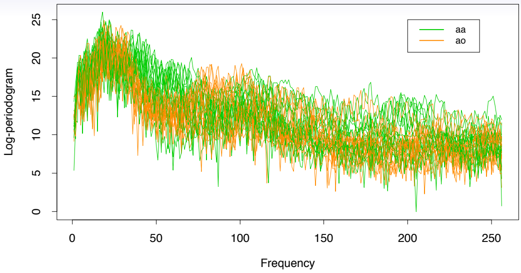

lpsa
lcavol
lweight
age
lbph
svi
lcp
gleason
−1
1
2
3
4
40
50
60
70
80
0.0
0.4
0.8
6.0
7.0
8.0
9.0
4
3
2
1
−1
80
70
60
50
40
0.8
0.4
0.0
9.0
8.0
7.0
6.0
0
1
2
3
4
5
2.5
3.5
4.5
−1
0
1
2
−1
0
1
2
3


lpsa
lcavol
lweight
age
lbph
svi
lcp
gleason
pgg45
-1
1
3
40
60
80
0.0
0.4
0.8
6.0
7.5
9.0
0
1
2
3
4
5
-1
0
1
2
3
4
6
5
4
3
40
50
60
70
80
2
1
0
-1
0.0
0.4
0.8
3
2
1
0
-1
6.0
7.0
8.0
9.0
100
60
0 20
0
2
4
3
4
5
6
-1
0
1
2
-1
1
2
3
0
40
80

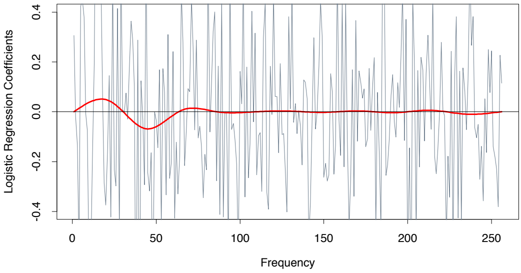
Statistical Learning Problems

- Identify the risk factors for prostate cancer.
- Classify a recorded phoneme based on a log-periodogram.
- Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements.
- Customize an email spam detection system.
- Identify the numbers in a handwritten zip code.
- Classify a tissue sample into one of several cancer classes, based on a gene expression profile.
- Establish the relationship between salary and demographic variables in population survey data.
- Classify the pixels in a LANDSAT image, by usage.

Phoneme Examples

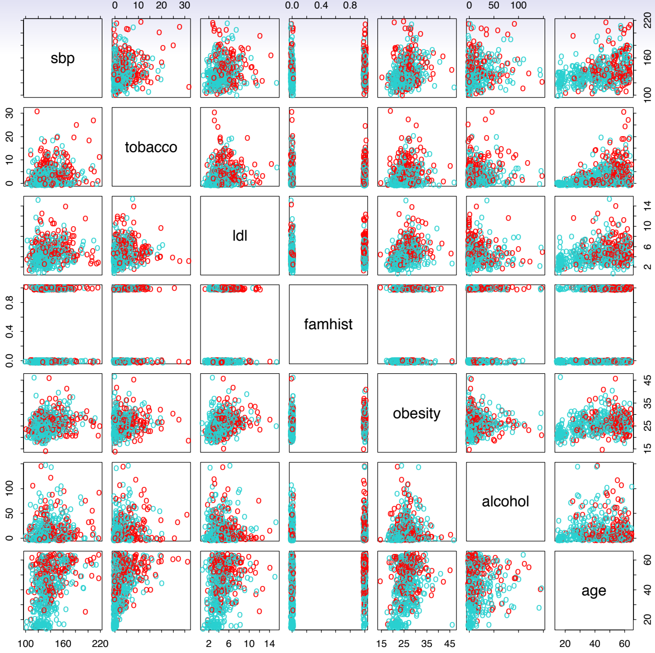


Phoneme Classification: Raw and Restricted Logistic Regression



Statistical Learning Problems

- Identify the risk factors for prostate cancer.
- Classify a recorded phoneme based on a log-periodogram.
- Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements.
- Customize an email spam detection system.
- Identify the numbers in a handwritten zip code.
- Classify a tissue sample into one of several cancer classes, based on a gene expression profile.
- Establish the relationship between salary and demographic variables in population survey data.
- Classify the pixels in a LANDSAT image, by usage.


sbp
tobacco
ldl
famhist
obesity
alcohol
age
0
10
20
30
0.0
0.4
0.8
0
50
100
220
160
100
30
20
10
0
14
10
6
2
0.8
0.4
0.0
45
35
25
15
100
50
0
60
40
20
100
160
220
2
6
10
14
15
25
35
45
20
40
60

Statistical Learning Problems

- Identify the risk factors for prostate cancer.

- Classify a recorded phoneme based on a log-periodogram.

- Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements.

- Customize an email spam detection system.

- Identify the numbers in a handwritten zip code.

- Classify a tissue sample into one of several cancer classes, based on a gene expression profile.

- Establish the relationship between salary and demographic variables in population survey data.

- Classify the pixels in a LANDSAT image, by usage.

Spam Detection

- data from 4601 emails sent to an individual (named George, at HP labs, before 2000). Each is labeled as *spam* or *email*.
- goal: build a customized spam filter.
- input features: relative frequencies of 57 of the most commonly occurring words and punctuation marks in these email messages.

	`george`	`you`	`hp`	`free`	`!`	`edu`	`remove`
`spam`	0.00	2.26	0.02	0.52	0.51	0.01	0.28
`email`	1.27	1.27	0.90	0.07	0.11	0.29	0.01

Average percentage of words or characters in an email message equal to the indicated word or character. We have chosen the words and characters showing the largest difference between `spam` *and* `email`*.*

Statistical Learning Problems

- Identify the risk factors for prostate cancer.

- Classify a recorded phoneme based on a log-periodogram.

- Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements.

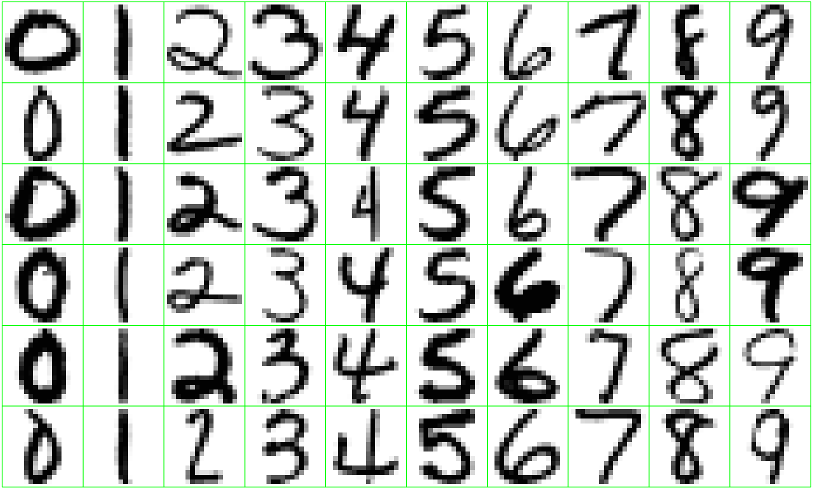
- Customize an email spam detection system.

- Identify the numbers in a handwritten zip code.

- Classify a tissue sample into one of several cancer classes, based on a gene expression profile.

- Establish the relationship between salary and demographic variables in population survey data.

- Classify the pixels in a LANDSAT image, by usage.



Statistical Learning Problems

- Identify the risk factors for prostate cancer.

- Classify a recorded phoneme based on a log-periodogram.

- Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements.

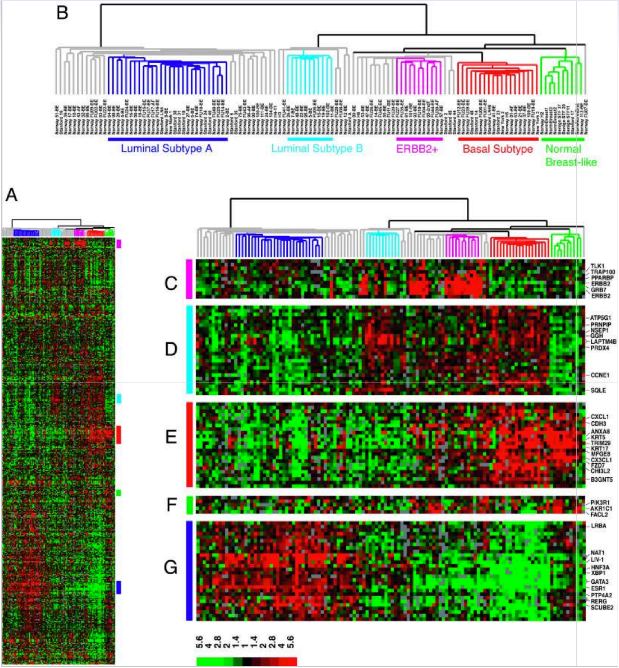
- Customize an email spam detection system.

- Identify the numbers in a handwritten zip code.

- Classify a tissue sample into one of several cancer classes, based on a gene expression profile.

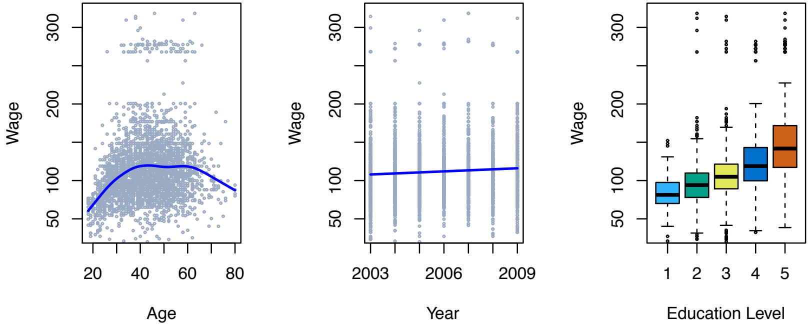
- Establish the relationship between salary and demographic variables in population survey data.

- Classify the pixels in a LANDSAT image, by usage.


B
Luminal Subtype A
Luminal Subtype B
ERBB2+
Basal Subtype
Normal Breast-like
A
C
D
E
F
G
TLK1
TRAP100
PPARBP
ERBB2
GRB7
ERBB2
ATP5G1
PRNPIP
NSEP1
GGH
LAPTM4B
PRDX4
CCNE1
SQLE
CXCL1
CDH3
ANXA8
KRT5
TRIM29
KRT17
MFGE8
CX3CL1
FZD7
CHI3L2
B3GNT5
PIK3R1
AKR1C1
FACL2
LRBA
NAT1
LIV-1
HNF3A
XBP1
GATA3
ESR1
PTP4A2
RERG
SCUBE2
5.6
4
2.8
2
1.4
1
1.4
2
2.8
4
5.6

Statistical Learning Problems

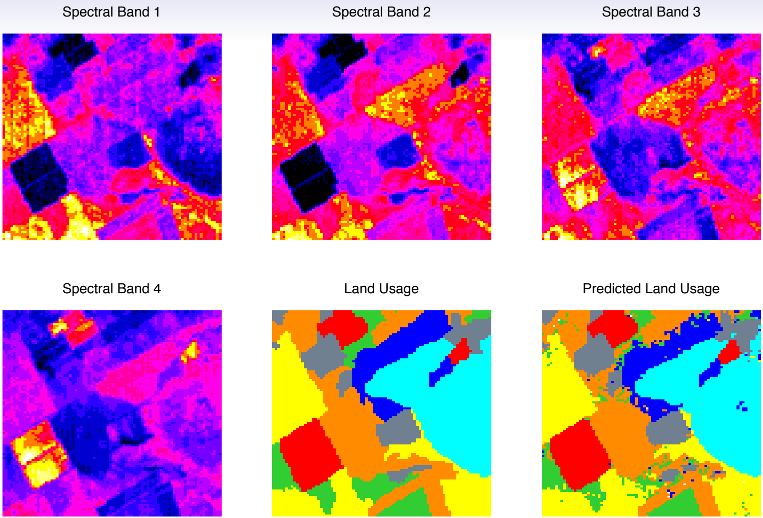
- Identify the risk factors for prostate cancer.
- Classify a recorded phoneme based on a log-periodogram.
- Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements.
- Customize an email spam detection system.
- Identify the numbers in a handwritten zip code.
- Classify a tissue sample into one of several cancer classes, based on a gene expression profile.
- Establish the relationship between salary and demographic variables in population survey data.
- Classify the pixels in a LANDSAT image, by usage.



Income survey data for males from the central Atlantic region of the USA in 2009.

Statistical Learning Problems

- Identify the risk factors for prostate cancer.
- Classify a recorded phoneme based on a log-periodogram.
- Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements.
- Customize an email spam detection system.
- Identify the numbers in a handwritten zip code.
- Classify a tissue sample into one of several cancer classes, based on a gene expression profile.
- Establish the relationship between salary and demographic variables in population survey data.
- Classify the pixels in a LANDSAT image, by usage.



Spectral Band 1

Spectral Band 2

Spectral Band 3

Spectral Band 4

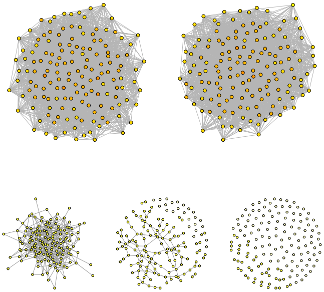
Land Usage

Predicted Land Usage

$Usage \in \{red\ soil,\ cotton,\ vegetation\ stubble,\ mixture,\ gray\ soil,\ damp\ gray\ soil\}$

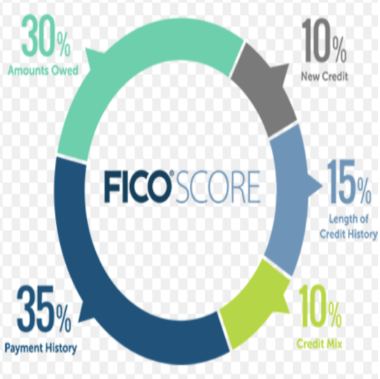
Graphical network

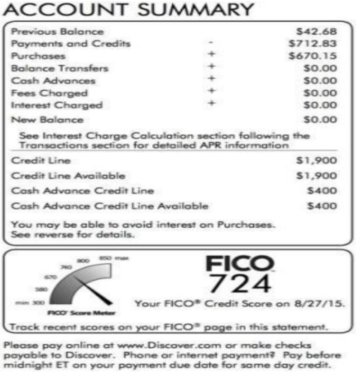
17,214 gene expressions and 22,247 CNVs are available from TCGA(X.Fan,K.Fang,S.Ma, Q. Zhang. Assisted Graphical Model for Gene Expression Data Analysis. Statistics in Medicine. 2019).



Top left panel: AGM; Top right panel: GM; Bottom left panel: difference between AGM and GM; Bottom middle: difference between moderate connections ; Bottom right: difference between strong connections.

The FICO score was first introduced in 1989 by Fair, Isaac, and Company, whic is used by the vast majority of banks and credit grantors, and is based on consumer credit files of the three national credit bureaus: Experian, Equifax, and TransUnion.





ACCOUNT SUMMARY

Previous Balance		$42.68
Payments and Credits	-	$712.83
Purchases	+	$670.15
Balance Transfers	+	$0.00
Cash Advances	+	$0.00
Fees Charged	+	$0.00
Interest Charged	+	$0.00
New Balance		$0.00

See Interest Charge Calculation section following the Transactions section for detailed APR information

Credit Line	$1,900
Credit Line Available	$1,900
Cash Advance Credit Line	$400
Cash Advance Credit Line Available	$400

You may be able to avoid interest on Purchases.
See reverse for details.

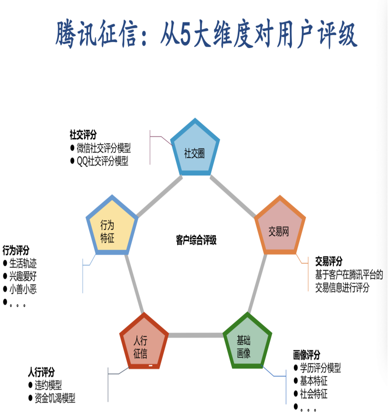
FICO
724

Your FICO® Credit Score on 8/27/15.

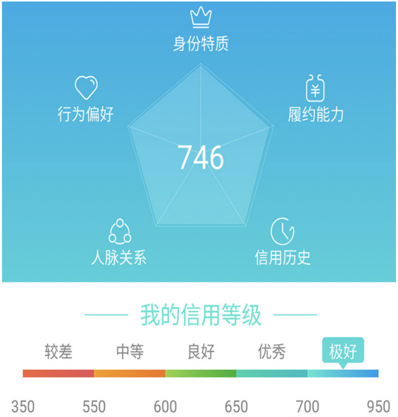
Track recent scores on your FICO® page in this statement.

Please pay online at www.Discover.com or make checks payable to Discover. Phone or internet payment? Pay before midnight ET on your payment due date for same day credit.

Credit score in China



腾讯征信：从5大维度对用户评级



Interesting problem for Credit scoring

- Variable selection
- reject inference
- reject option
- imbalance data
- label noise
- missing data
- fraud detection

Data integratoin

Two type of Data integration (Data fushion): Sample integration and Variable integration

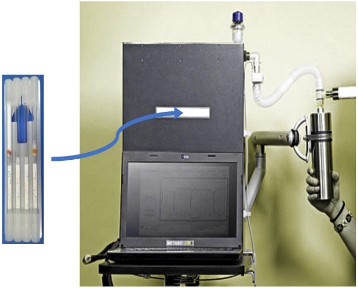
Table 1: 3 datasets on breast cancer from GEO

Breast	Gene	Sample	Case	Control
GSE9574	20995	29	14	15
GSE21947	20995	30	15	15
GSE5364	20995	196	183	13

Variable integration: Gene and CNA

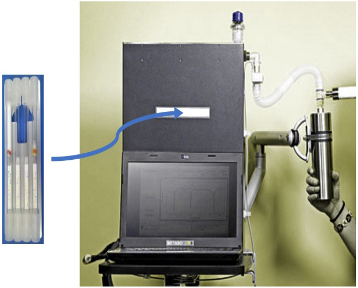
Integration of supervised and unsupervised learning

CSA Breath Data



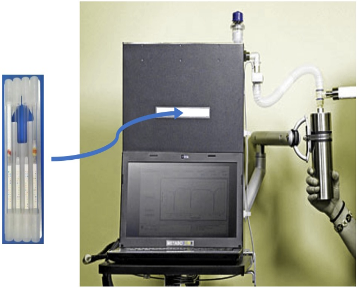
1. Tidal volume breathing for 5 minutes;

CSA Breath Data



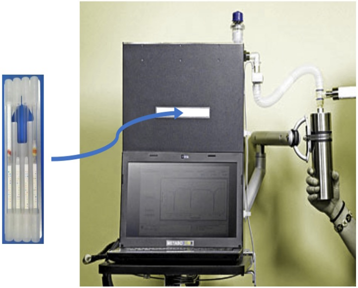
1. Tidal volume breathing for 5 minutes;
2. Exhaled breath drawn over the sensor array;

CSA Breath Data



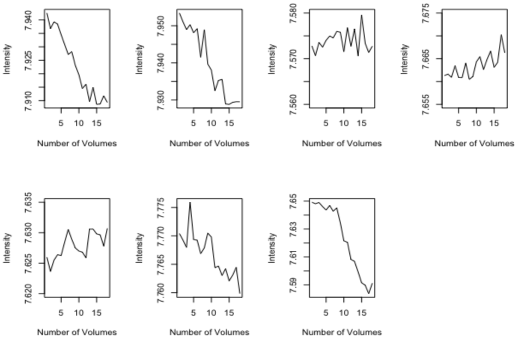
1. Tidal volume breathing for 5 minutes;
2. Exhaled breath drawn over the sensor array;
3. Images were converted to numerical values in the red, green, blue spectra, and 4 ultraviolet spectra.

CSA Breath Data

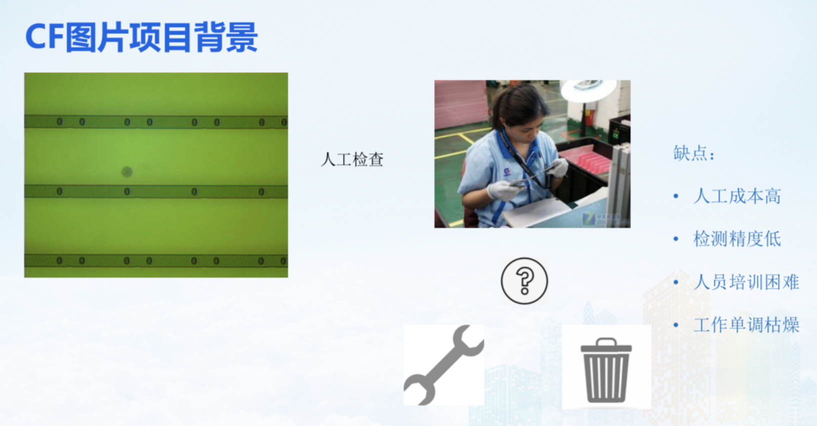


1. Tidal volume breathing for 5 minutes;
2. Exhaled breath drawn over the sensor array;
3. Images were converted to numerical values in the red, green, blue spectra, and 4 ultraviolet spectra.
4. Totally 128 (the number of colorants) × 7(changes in the red, green, blue, and 4 ultra-color spectrum of each colorant) $= 896$ groups.

CSA Breath Data



Object detection



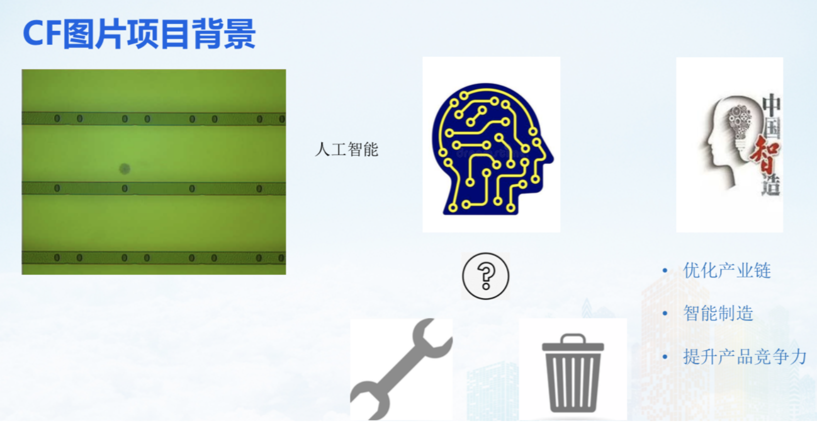
CF图片项目背景

人工检查

缺点：

- 人工成本高
- 检测精度低
- 人员培训困难
- 工作单调枯燥

Object detection



CF图片项目背景

人工智能

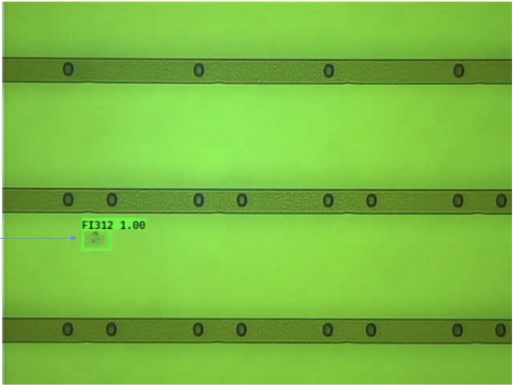
- 优化产业链
- 智能制造
- 提升产品竞争力

Object detection

标注示例——FI312

输出结果包含

- 分类标签FI312
- 分类置信度
- 目标所在的位置框



Network experience Index

II. Construction of the composite NEI/NPI index

Objective: Develop a new strategy for constructing the composite NEI/NPI index which can "inherit" the advantages of the existing approaches while overcoming their limitations.

Feature screening

- From Step I analysis: connectivity and Laplacian
- Experts' opinion

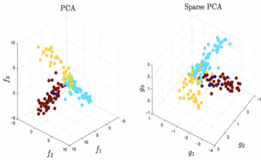
Feature selection and estimation

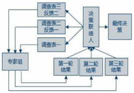
PCA (and other feature selection methods): identify relevant features and estimate their loadings.

Bayesian update

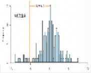
Delphi method

Collect prior information from experts through questionnaire.





Design of the loadings



VaR

Value at risk is a measure of the risk of loss for investments (with a given probability).

Extreme value theory

It assesses the probability of an event that is more extreme than previously observed. POT(Peaks Over Threshold) and BMM(Block Maxima Method are used to find the extreme values.

$$F_u(y) = P(X - u \le y \mid X > u) = \frac{P(u < X < u + y)}{P(X > u)} = \frac{F(u+y) - F(u)}{1 - F(u)} = \frac{F(x) - F(u)}{1 - F(u)}$$

$$\begin{aligned} F_1(x) &= P\left(X_{(1)} < x\right) = 1 - P\left(X_{(1)} \ge x\right) \\ &= 1 - P(X_1 > x, X_2 > x, \cdots, X_n > x) \\ &= 1 - [1 - F(x)]^n \end{aligned}$$

$$\begin{aligned} F_n(x) &= P\left(X_{(n)} \le x\right) \\ &= P(X_1 \le x, X_2 \le x, \cdots, X_n \le x) \\ &= F^n(x) \end{aligned}$$

[illegible]

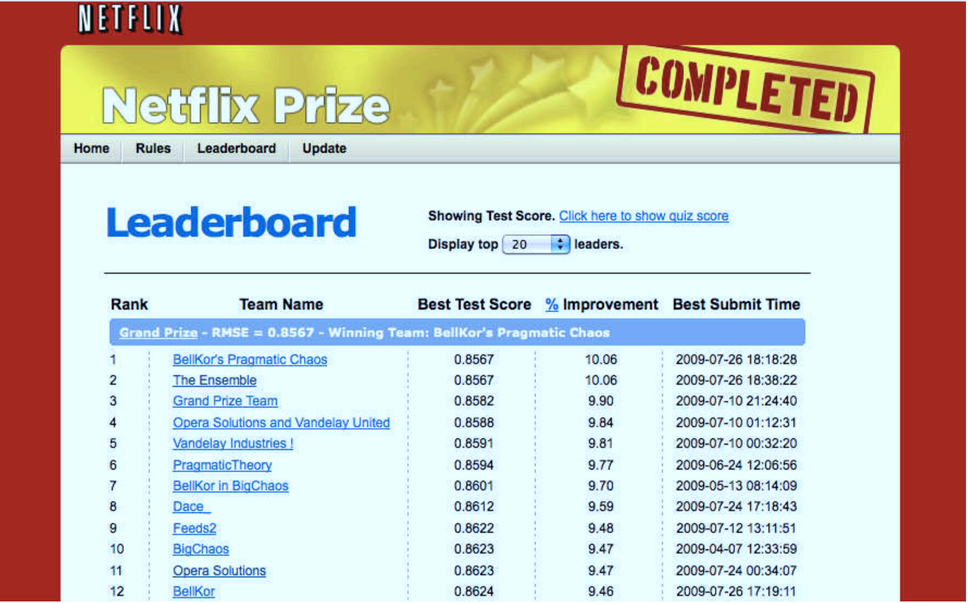
$$F(x) = (1 - F(u)) \times F_u(y) + F(u), \quad 0 \le y \le x_f - u,$$

Upper and lower bounds of indexes

Deliverable: a composite NEI/NPI index

The Netflix prize

- competition started in October 2006. Training data is ratings for $18,000$ movies by $400,000$ Netflix customers, each rating between 1 and 5.
- training data is very sparse— about 98% missing.
- objective is to predict the rating for a set of 1 million customer-movie pairs that are missing in the training data.
- Netflix's original algorithm achieved a root MSE of 0.953. The first team to achieve a 10% improvement wins one million dollars.
- is this a supervised or unsupervised problem?



Leaderboard

Showing Test Score. Click here to show quiz score

Display top 20 leaders.

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace_	0.8612	9.59	2009-07-24 17:18:43
9	Feeds2	0.8622	9.48	2009-07-12 13:11:51
10	BigChaos	0.8623	9.47	2009-04-07 12:33:59
11	Opera Solutions	0.8623	9.47	2009-07-24 00:34:07
12	BellKor	0.8624	9.46	2009-07-26 17:19:11

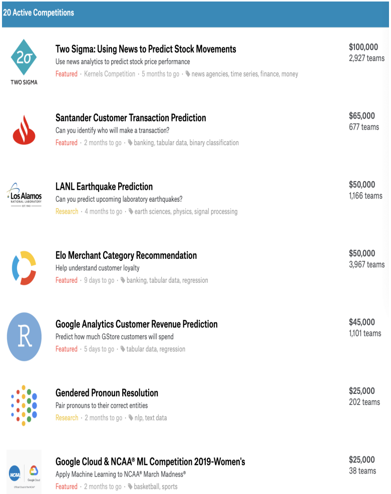
BellKor's Pragmatic Chaos wins, beating The Ensemble by a narrow margin.

Kaggle

kaggle Search Competitions Datasets

Kaggle is the place to do data science projects

See how it works



20 Active Competitions

Competition	Prize / Teams
Two Sigma: Using News to Predict Stock Movements Use news analytics to predict stock price performance Featured · Kernels Competition · 5 months to go · news agencies, time series, finance, money	$100,000 2,927 teams
Santander Customer Transaction Prediction Can you identify who will make a transaction? Featured · 2 months to go · banking, tabular data, binary classification	$65,000 677 teams
LANL Earthquake Prediction Can you predict upcoming laboratory earthquakes? Research · 4 months to go · earth sciences, physics, signal processing	$50,000 1,166 teams
Elo Merchant Category Recommendation Help understand customer loyalty Featured · 9 days to go · banking, tabular data, regression	$50,000 3,967 teams
Google Analytics Customer Revenue Prediction Predict how much GStore customers will spend Featured · 5 days to go · tabular data, regression	$45,000 1,101 teams
Gendered Pronoun Resolution Pair pronouns to their correct entities Research · 2 months to go · nlp, text data	$25,000 202 teams
Google Cloud & NCAA® ML Competition 2019-Women's Apply Machine Learning to NCAA® March Madness® Featured · 2 months to go · basketball, sports	$25,000 38 teams

Software