**Big data** usually includes data sets with sizes beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time.”

a = array(v, dim = c(3,4,3), dimnames = list(rn,cn,mn)):

unValues <- data[!complete.cases(data),] 输出带empty的

3.**Big data’s 3V: Volume, Variety, Velocity**

relational database: a set of relations

Relation(2 parts): Schema, Instance

Integrity constraints(ICS): condition that must be true for any instance of the database

a legal instance of a relation is one that satisfies all specified ICs

Primary key:

Super key: a combination of columns that uniquely defines any row within a table

Candidate key:

Foreign key: must correspond to a primary key in an another table

Normalization:

1NF: all elements atomic

2NF: every non-prime attribute depends on a candidate key or another non-prime attribute

3NF: some redundancy but dependency preserving

BCNF: no redundancy but not dependency preserving

SQL

DDL (Data Description Language): create, drop, alter   
 DML (Data Manipulation Language): insert, update, delete, select  
 DCL (Data Control Language): grant, revoke  
 TCL (Transaction Control Language): commit, rollback

ACID transaction: atomicity, consistency, isolation, durability

NoSQL

Not using the relational model; schemaless; running well on clusters, tend to be open-source; aggregates - nested data stored together

application development productivity; large-scale data

An aggregate is a collection of data that we interact with as a unit.

1) Key-value database  
2) Document database

3) Column-family stores

4) Graph database

NoSQL differs to RDBMS in the way entities get distributed and that no consistency is enforced across those entities

When to use RDBMS: table based; relations between distinct table entities and rows; referential integrity; ACID transactions; arbitrary queries and joins

When to use NoSQL:

If you just want to store your application entities in a persistent and consistent way   
 If you have hierarchical application objects and need some query capability into them   
 If you ever tried to store large trees or networks you will know that an RDBMS is not the best solution here   
 If you are running in the Cloud and need to run a distributed database for durability and availability.   
 You might already use a data warehouse for your analytics. If your data grows to large to be processed on a single machine, you might look into hadoop or any other solution that supports distributed Map/Reduce.

4.

Definitions of Data warehouse:

Subject-oriented, integrated, time-variant, non-volatile;

Copy of transaction data, specifically structured for query and analysis;

Single, complete and consistent store of data.

• The goal of data warehousing is to integrate enterprise wide corporate data into a single repository from which users can easily run queries.

• Data warehouse is an organization’s (enterprise’s) memory.

**Expectations:**

* Data should be integrated across the enterprise
* Summary data has a real value to the organization
* Historical data holds the key to understanding data over time
* What-if capabilities are required

**Warehousing issues:**

* **Semantic Integration**: When getting data from multiple sources, must eliminate mismatches, e.g., different units (temperature, weight, currency).
* Heterogeneous Sources: Must access data from a variety of source formats and repositories
* Load, Refresh, Purge: Must load data, periodically refresh it, and purge too-old data
* Metadata Management: Must keep track of sources, loading time, and other information for all data in the warehouse

Data warehouse architecture

* Data warehouse server  
  – almost always a relational DBMS, rarely flat files
* OLAP servers

-Multidimensional data model

-Collection of numeric measures which depend on a set of dimensions

– to support and operate on multi-dimensional data

* structures • Clients
* – Query and reporting tools – Analysis tools  
  – Data mining tools

Data warehouse schema:

* “Star” schema
* **“Fact constellation” schema**
* “Snowflake” schema

**Operations in multidimensional data model：**

* **Aggregation**
* **Selection**
* **Navigation**
* **Visualization**

**OLAP queries**

* **Drill-down**
* **Pivoting**

5.

**Measures of central tendency:**

• mode

• median   
• mean μ  
• x% trimmed mean

**Measures of spread:**

• ranges: crude range: highest, lowest

Extended/corrected range: adds one unit to the range (to account for a possible error in measurement)

trimmed ranges: drop x% of extreme points on both sides

* variance σ2 =Σi(xi-μ)2/n
* standard deviation σ =sqrt(σ2)
* average deviation Σi (xi-μ)/n

**Other measures of probability distribution:**

• kurtosis: a descriptor of the shape of a probability distribution

• skewness: a descriptor of asymmetry of the probability distribution

**Normal/Gaussain distribution ( σ>0)**

1. μ = median = mode

2. ***3-sigma rule***: 68% within 1σ, 95% within 2σ, 99.7% within 3σ

**Linear regression**

1. Correlation does not mean causation

**2. Least-squares regression:** minimize the sum of squares of the deviations of the data points from the line in the vertical direction

**3. Asymmetry of regression**

**4. Outliers,** typically removed manually.

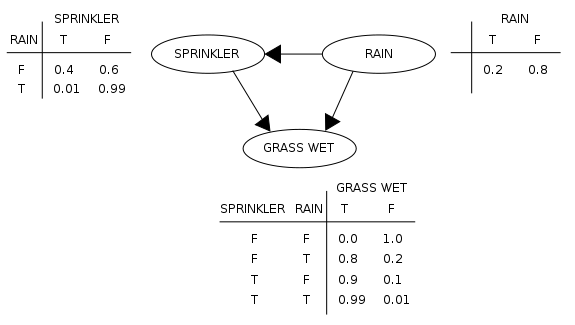
**Bayesian networks**

1. P(A,B,C,D)=P(A|B,C,D) P(B|C,D) P(C|D) P(D)

Arc from the conditioning variables to the variables in the factorization

2. Prior probability distribution tables for nodes without predecessors

Conditional probability distributions tables for nodes with predecessors



3. when there are independences in the domain, Bayesian networks are much more efficient than probability tree.

4. Causal Markov condition

For a Bayesian network states that every node in a Bayesian network is conditionally independent\* of its nondescendents, given its parents.

\*In the standard notation of probability theory, ***R* and *B* are conditionally independent** given *Y* if and only if

P(R∩B|Y) = P(R|Y) P(B|Y)

R, B and Y are represented by the areas shaded red, blue and yellow respectively.

5. Directed graphs

Advantages: reflect the causal structure, accommodate representation of uncertainty, can be reconfigured as needed

7.

• The classical statistics: Significance testing.

**Null hypothesis (H0) (no effect) and its complement (H1).**

**Significance level (α, p value).**

Statistical power (1- β), probability of rejecting H0 given that it should be rejected.

Sample size n / Effect size.

**If P(data|H0)<α, reject H0 (the data are unlikely/surprising if H0 is true).**

**Type I errors (reject H0 when true).  
Type II errors (accept H0 when false).**

Decision making: we need to consider uncertainty and preferences.

Probability is a measure of uncertainty

Utility is a measure of preference that combines with probability as mathematical expectation.

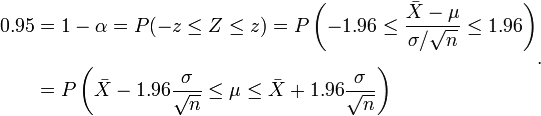
determine the **prior probability of the hypothesis**

**Decision-theoretic (Bayesian) view allows to explore the exact relation between the significance level and the decision.**

**Confidence Intervals**

from the point of view of the **classical hypothesis testing**

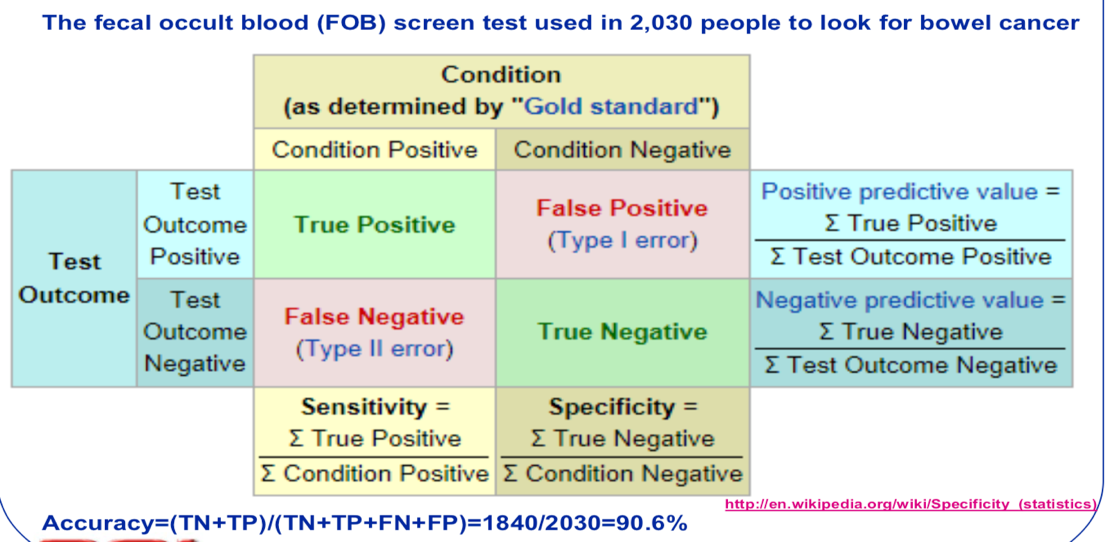
***I'm 95% sure that the true mean is inside this interval***

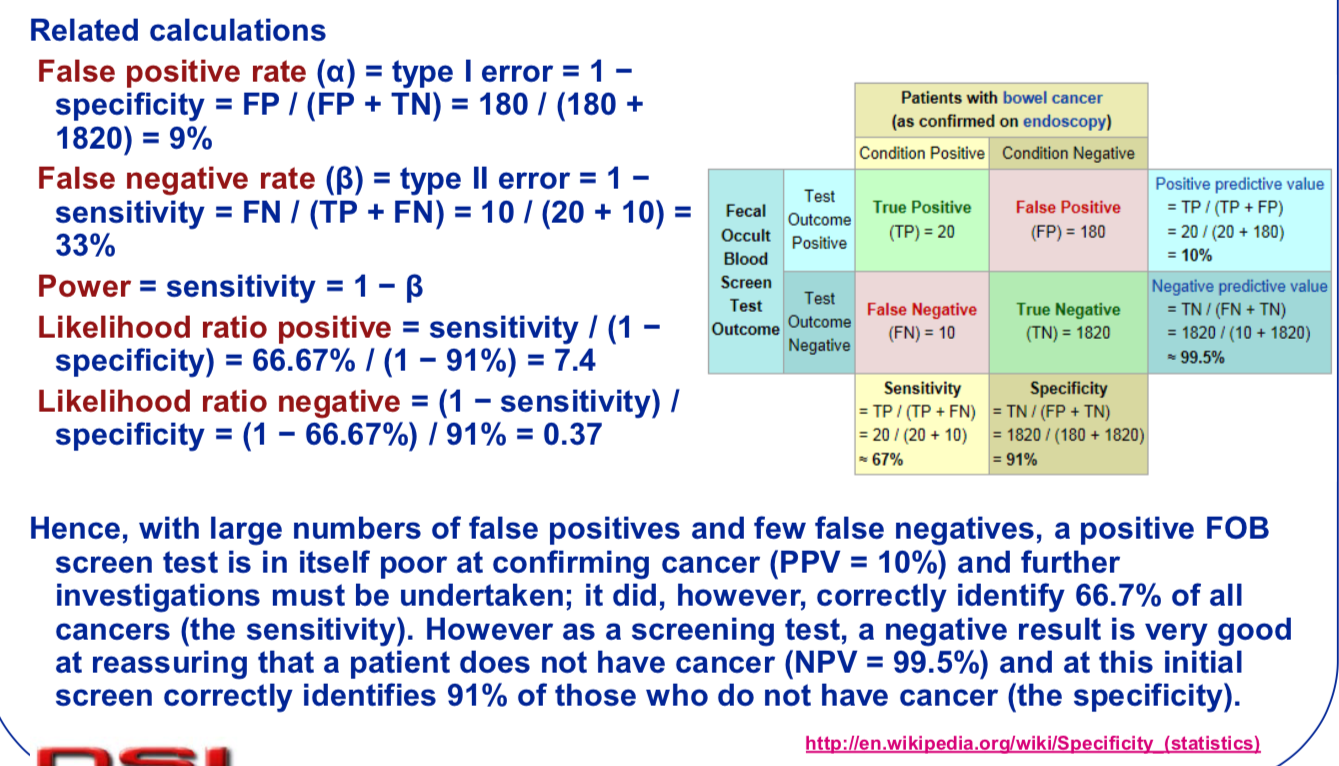


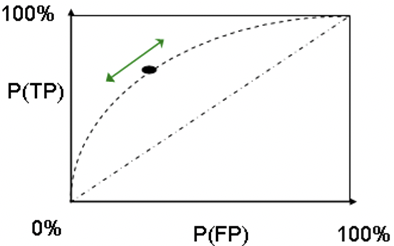
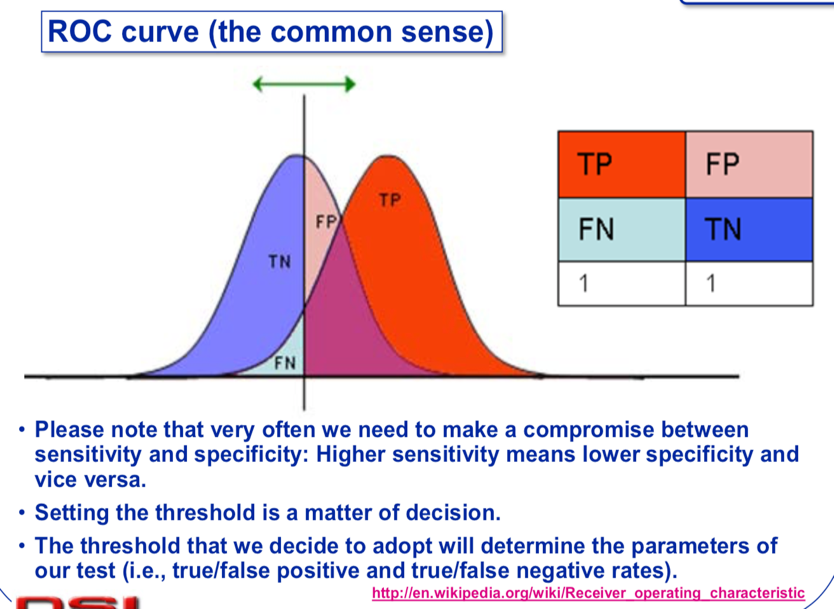
**Sensitivity**: recall rate, measures the proportion of actual positives which are correctly identified

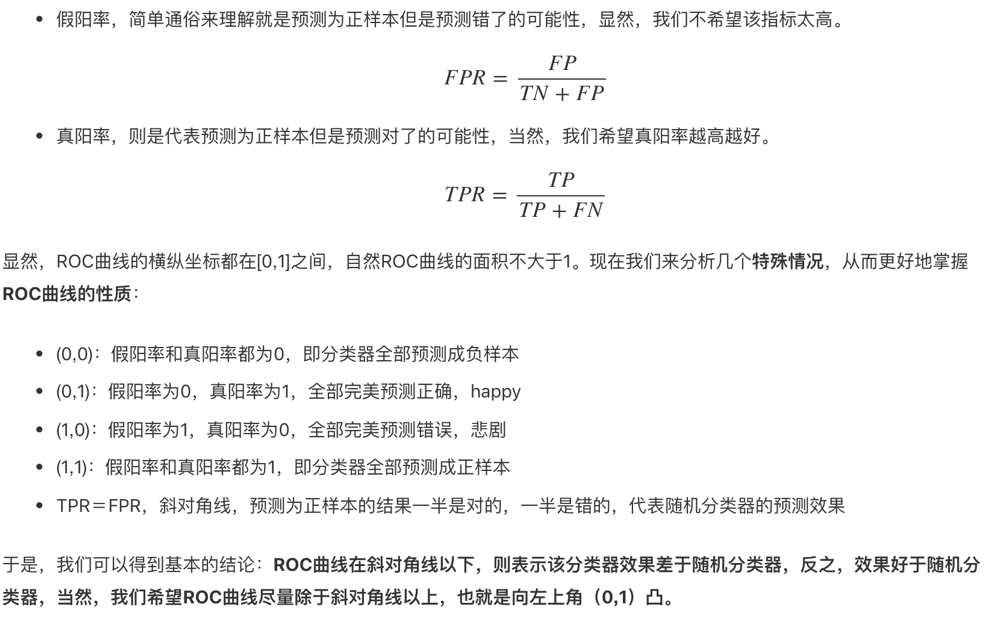
**Specificity:** measures the proportion of negatives which are correctly identified

A perfect predictor would be described as 100% senitivity and 100% specificity.

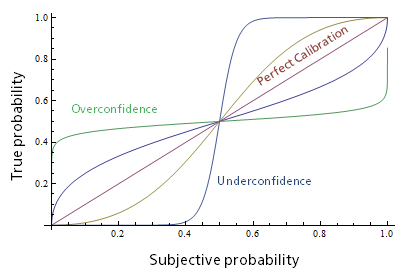








**AUC: Area Under the (ROC) Curve**

* AUC ＝ 1，代表完美分类器
* 0.5 < AUC < 1，优于随机分类器
* 0 < AUC < 0.5，差于随机分类器
* **Calibration**
* We plot the frequencies observed in the data (y axis) against the probabilities calculated by the system (x axis).
* 
* **Cross-validation:**
* Testing a model on the same data that we used for training it does not seem fair.
* It will favor most complex models that fit the data best.
* Simpler model may actually fit future instances of data better than complex models.
* **Cross-validation prevents over-fitting**
* **k-Fold cross-validation**
* **“Leave-One-Out” cross-validation**
* Uses effectively n-1 instances for training and tests the model on all n instances, one at a time (an extreme case of k-fold, k=n).
* **Bootstrap cross-validation**
* Repetitive test set method, each repetition involves selecting a new test set from among all records.