1. **Data privacy**

**Definition**: Freedom from unauthorized intrusion

**Aspects**:

1. Information privacy(收集和处理个人信息，如医疗记录)；
2. Bodily privacy(保护物理信息被侵入)；
3. Privacy of communication(通信内容中的隐私)；
4. Territorial privacy(个人居所等领地限制进入)；

**Current situation**: 以Google为代表的互联网公司收集了大量的用户信息，然而却未声明如何使用和处置。



**Privacy protection measures (high layer categories)**

1. Access control: Restricted access to the data to authorized users;

Check the identification of user during an authentication process to prevent unauthorized users from accessing the data.

1. **Inference control**: Restrict inference from accessible data to additional data;

**针对disclosing data from inference attack，一类主要的威胁衍生的主要手段。**

**研究工作主要围绕：**

1. **Data anonymization** (通过包括generalization在内的多种方式对数据进行匿名);
2. **Statistical database** (设计统计数据库的访问和输出方式);
3. **Privacy preserving data mining** (在数据挖掘过程中保护隐私).
4. Flow control: Prevent information from authorized use to unauthorized use;
5. Encryption: Use cryptography to protect information from unauthorized disclosure while in transmit and in storage.
6. **Inference Control**

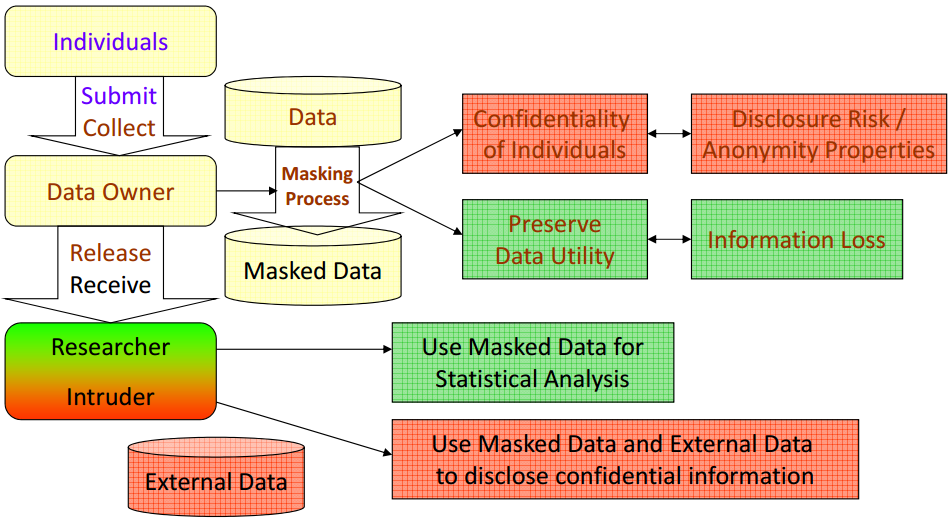
**Need**: publish useful information;

**Concern**: Private information disclosure through inference on publish data and external data.

**What is disclosure**?

The identity of an entity or something new about the entity is disclosed.

**2.1 data anonymization**

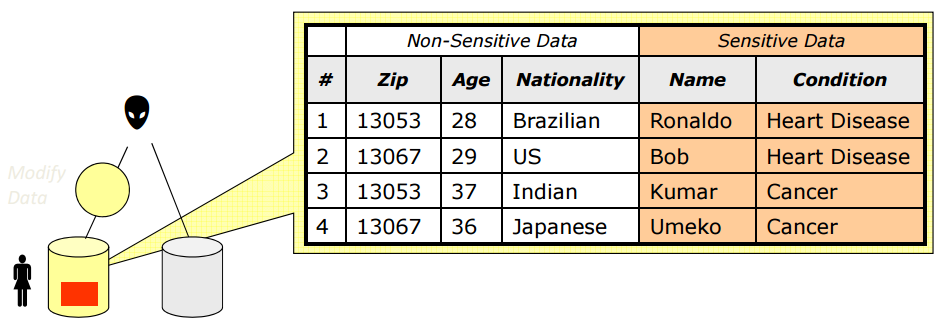


收集的用户数据在发布公开使用之前，需要进行masking，这个过程中要注意的两个方面是：

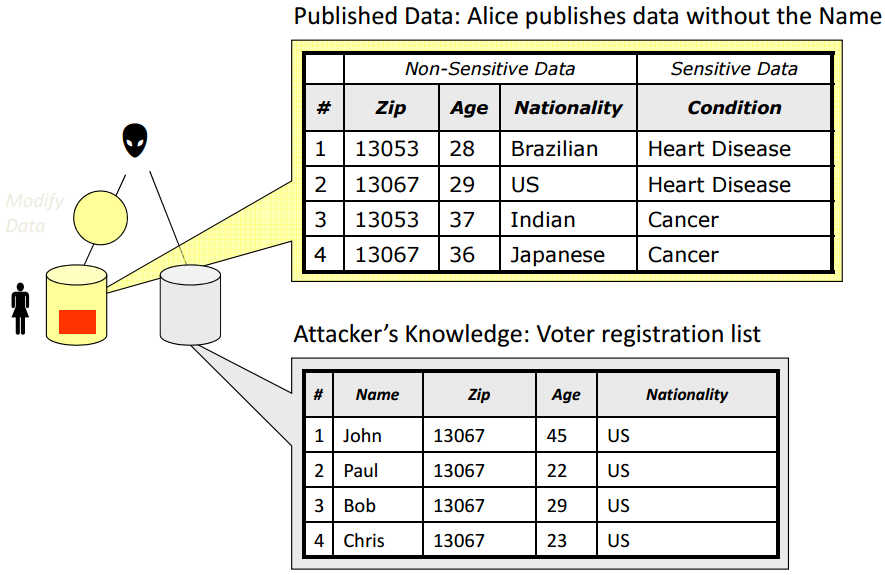
1. 进行部分属性的移除或修改，以防止隐私的泄露；
2. 在修改原始信息时保持数据的可用性，引入尽量少的信息丢失。

从上图可见，一个intruder通过将发布的masked data与external data（通过其它途径获取的microdata），分析出隐私数据。

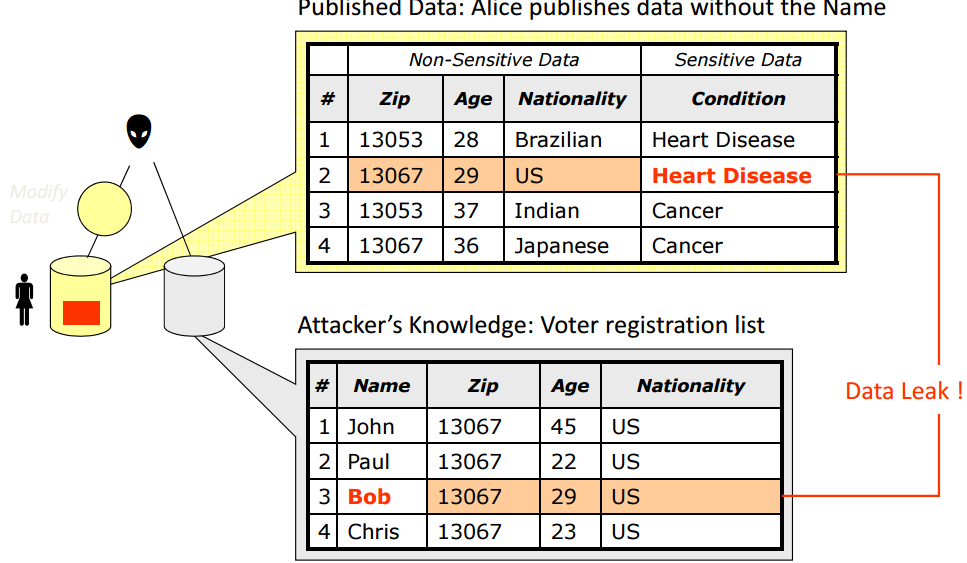
**以下是一个此类威胁的例子**：



黄色图形对应的数据库在发布数据之前进行过了masking过程，将敏感数据中的Name这一项去除；



Intruder可以同时获取masked data和灰色数据库对应的原始信息

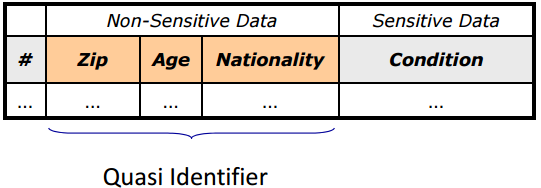


通过分析，找到重合的属性域，可以disclose entity’s identification(如图中的Bob)

**Source of the inference problem**:

即使在数据发布去除了用户身份信息，可是发布的数据中的某些字段可以唯一的指认出用户的身份，一个潜在的intruder可以通过将这些字段与其它数据资源比较，推断用户身份信息。

对于上面的例子，邮编、年龄和国籍信息即构成了这样一个**近似标识符**(quasi identifier)。

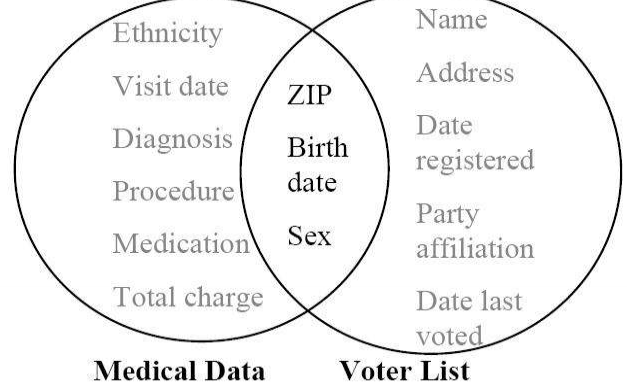


近似标识符：非直接身份信息（identifier，标识符），是一系列key属性的集合。

在LBS服务中，虽然用户的身份经过了匿名，可是location或者轨迹信息却构成了与身份相对应的identifier

**真实的隐私泄露例子**：

1. Massachusetts GIC Incident；



利用投票信息推测出发布的医疗数据中部分实体的身份。

1. AOL Query Log Release

**K-匿名**

A dataset satisfies k-anonymity for k > 1 if at least k records exist for each combination of quasi-identifier values.

对每个公开的属性的组合的查询都会返回至少K个记录，使得对应到真实身份上的可能低于1/K。

1998年提出，重要文献：K-Anonymity: A Model for Protecting Privacy

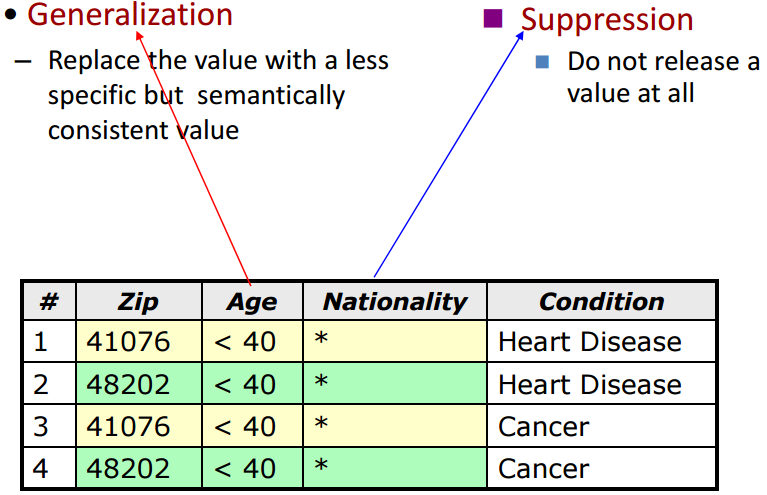
**L-diversity**

经过K匿名之后的数据集仍然可能受到Homogeneity attack，即K个record中的敏感信息是相同的。

针对该问题，L-多样性保证映射到一个K个记录构成的组中的敏感信息的值具有多样性。

**Inference control技术**

1. Remove identifier，移除身份标识符；
2. Sampling，通过采样的方式选出数据集的子集；
3. Microaggregation，微聚合，对某个属性进行聚合，该属性同时使用聚合后的结果，例如，将所有记录的工资求平均之后，用均值替代原本属性值；
4. Data swapping，将部分record的某个属性的值进行交换；
5. **Generalization and suppression**，是实现K-匿名的两种方式，因为将至少K个record的可能成为quasi identifier的属性变为相同的区域或者直接变成一个通配符，所以不再构成一个唯一的标识符。



* 1. **Cluster/microaggregation for anonymization**

**critique of generalization and suppression**:

1. 使用generalization/suppression实现K-匿名是NP难的；
2. 选择哪些字段进行generalization是难以确定的；
3. 对数据进行泛化和压缩的过程会降低数字的可用性，即使K个记录中仅仅两个tuple有相同的值。
4. 。。。。。

**Microaggregation/clustering步骤**

1. *Partition original dataset into* ***clusters*** *of similar records containing at least K records*;
2. *For each cluster,* ***compute an aggregation operation*** *and use it to replace the original records*.

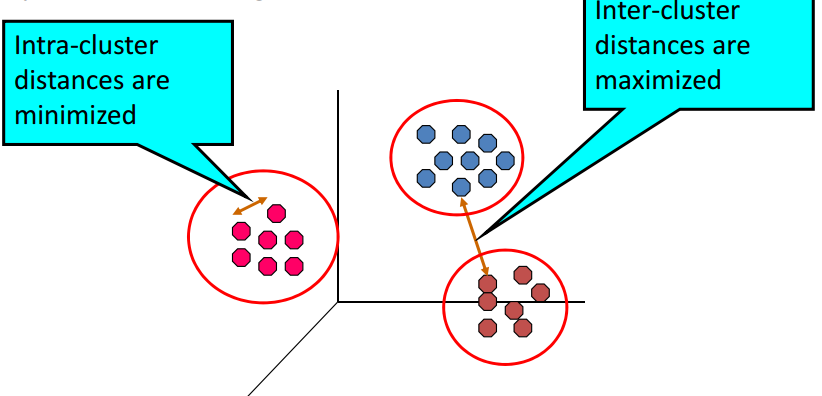
优势：

存在近乎于最优的启发式；

相比于泛化和压缩法，降低了data distortion。

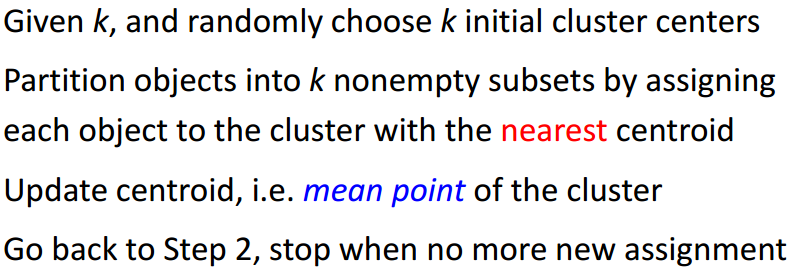
**Something about clustering**：

1. 评价：输出的多个簇中，簇内部的相似度很高；簇之间的相似度很低。

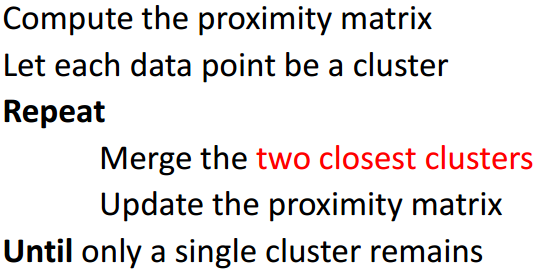


1. 聚类方法：
2. Partitioning approach

**k-means**



1. Hierarchical approach



下图给出有6个数据点时，如何通过5次（红色标记）层次化的迭代将其聚为1类。实际迭代中，可以选择一个距离阈值，超过阈值的两个簇之间不合并。

