**一、乱数据与脏数据的概念及区别**

**乱数据（Messy Data）  
指数据结构或组织形式不规范的原始数据，例如：**

* **格式混乱（如日期格式不统一：2023-01-01 vs 01/01/2023）**
* **多列合并存储（如将姓名和电话合并为“张三\_13800138000”）**
* **非结构化文本（如用户评论未分词或清洗）。**

**脏数据（Dirty Data）  
指数据内容存在错误或噪声，例如：**

* **重复记录（同一用户多次注册）**
* **异常值（年龄为-1或1000岁）**
* **逻辑矛盾（订单时间早于注册时间）。**

**区别：  
乱数据强调数据组织形式的混乱，需通过结构化处理解决；脏数据关注数据内容的错误，需通过清洗或修正解决。**

**二、机器学习中参数与超参数的概念及区别**

**参数（Parameters）**

* **模型内部通过训练数据自动学习的变量，例如：**
  + **线性回归中的权重（w）和偏置（b）**
  + **神经网络中的连接权重。**

**超参数（Hyperparameters）**

* **人工设定的控制模型训练过程的配置项，例如：**
  + **学习率（Learning Rate）**
  + **随机森林的树数量（n\_estimators）**
  + **K近邻算法的k值。**

**区别：  
参数是模型从数据中自动优化的，超参数需人工调整（如网格搜索、贝叶斯优化）。**

**三、数据缺失的三种类型及示例**

1. **完全随机缺失（MCAR, Missing Completely at Random）**
   * **缺失与任何变量无关，例如：设备故障导致某次实验数据随机丢失。**
2. **随机缺失（MAR, Missing at Random）**
   * **缺失与其他观测变量相关，例如：收入数据缺失可能与年龄相关（年轻人更可能隐瞒收入）。**
3. **不可忽略缺失（MNAR, Missing Not at Random）**
   * **缺失与缺失值本身相关，例如：高收入人群不愿填写收入字段，导致收入缺失与其实际值相关。**

**四、CAP 理论的全面解释**

**CAP 理论指出，在分布式系统中，以下三者无法同时满足：**

1. **一致性（Consistency）：所有节点访问同一最新数据。**
2. **可用性（Availability）：每次请求均能获得非错误响应。**
3. **分区容忍性（Partition Tolerance）：网络分区时系统仍能运行。**

**权衡场景：**

* **CP 系统（牺牲可用性）：如银行系统（强一致性优先，网络分区时拒绝服务）。**
* **AP 系统（牺牲一致性）：如社交媒体（允许临时数据不一致，保证高可用性）。**
* **CA 系统（牺牲分区容忍性）：仅适用于单节点或局域网场景（如传统数据库）。**

**核心结论：在分布式系统中，网络分区不可避免，实际需在 C 和 A 之间权衡。**