

词典

散列：冲突

11-A3

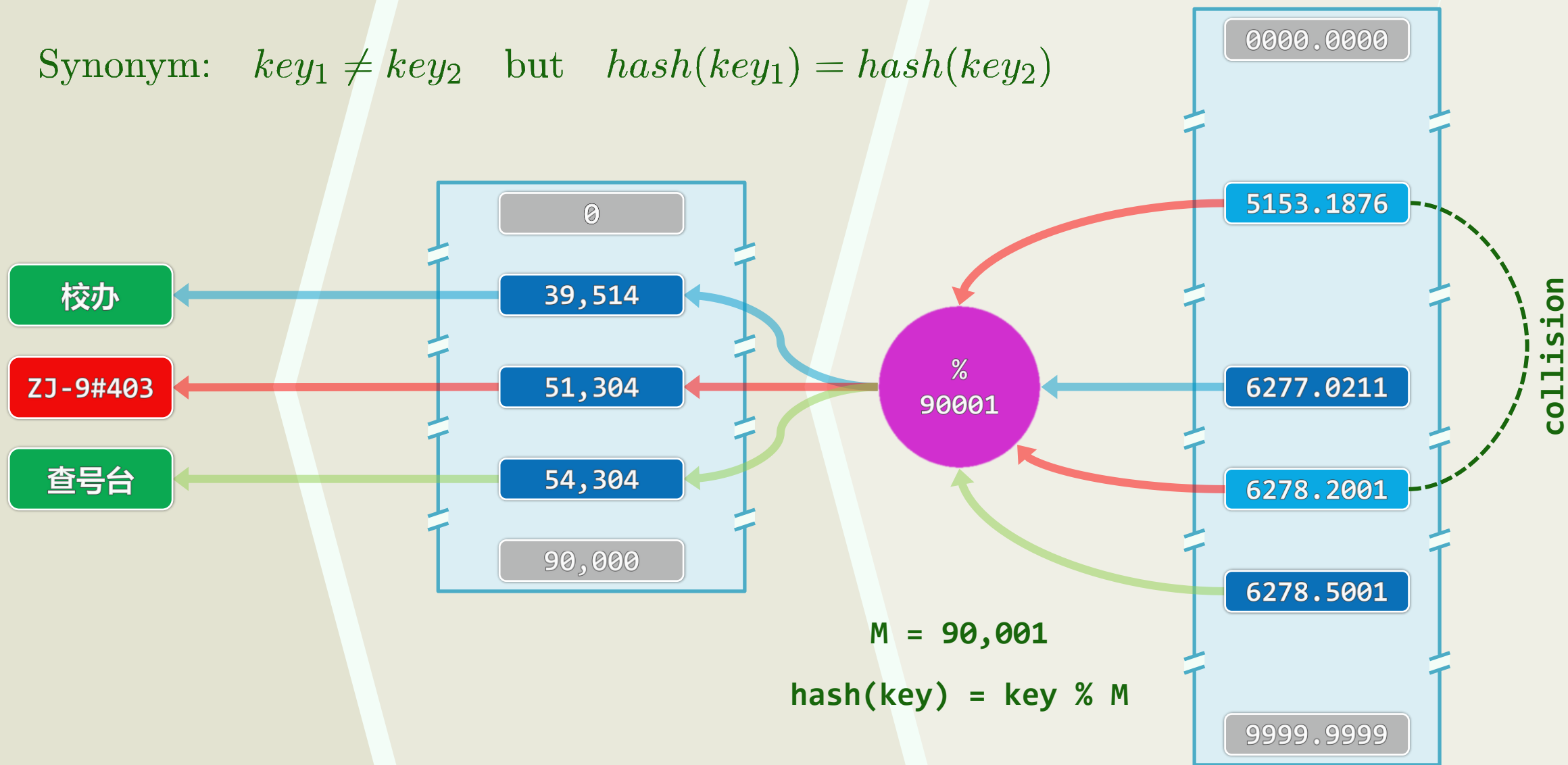
宝玉道：“已经完了，怎么又作揖？”袭人笑道：“这是他来给你拜寿。今儿也是他的生日，你也该给他拜寿。”宝玉听了，喜的忙作下揖去，说：“原来今儿也是姐姐的芳诞。”

邓俊辉

deng@tsinghua.edu.cn

同义词

Synonym: $key_1 \neq key_2$ but $hash(key_1) = hash(key_2)$



装填因子 vs. 冲突

❖ load factor : $\lambda = \mathcal{N}/\mathcal{M}$

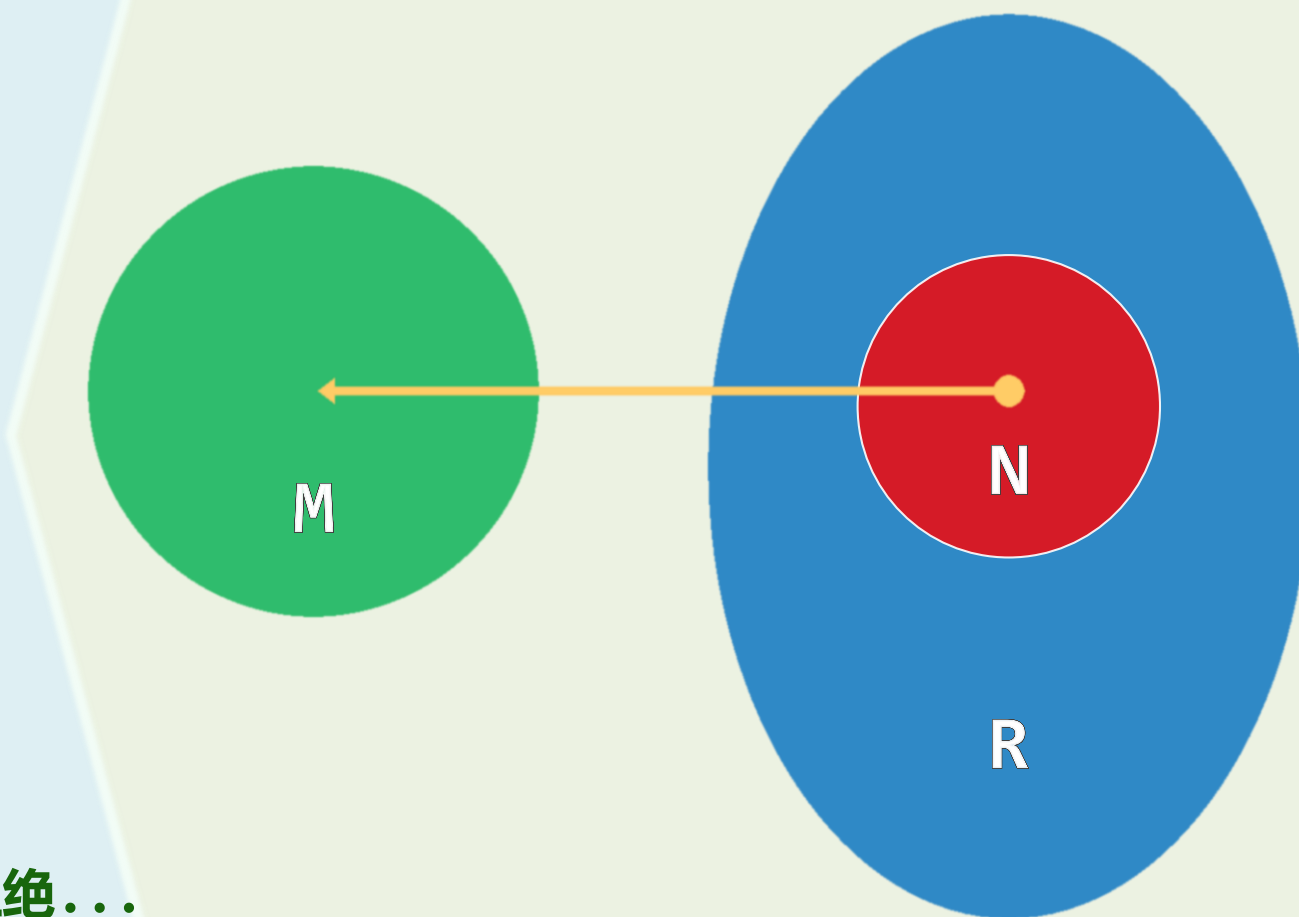
❖ λ 选多大才合适 ?

❖ λ 越大/小

- 空间利用率越高/低
- 冲突的情况沿越严重/轻微

❖ 通过降低 λ , 冲突状况将会大大改善

但只要数据集在动态变化 , 就无法彻底杜绝...



完美散列

❖ 在某些条件下，的确可以实现

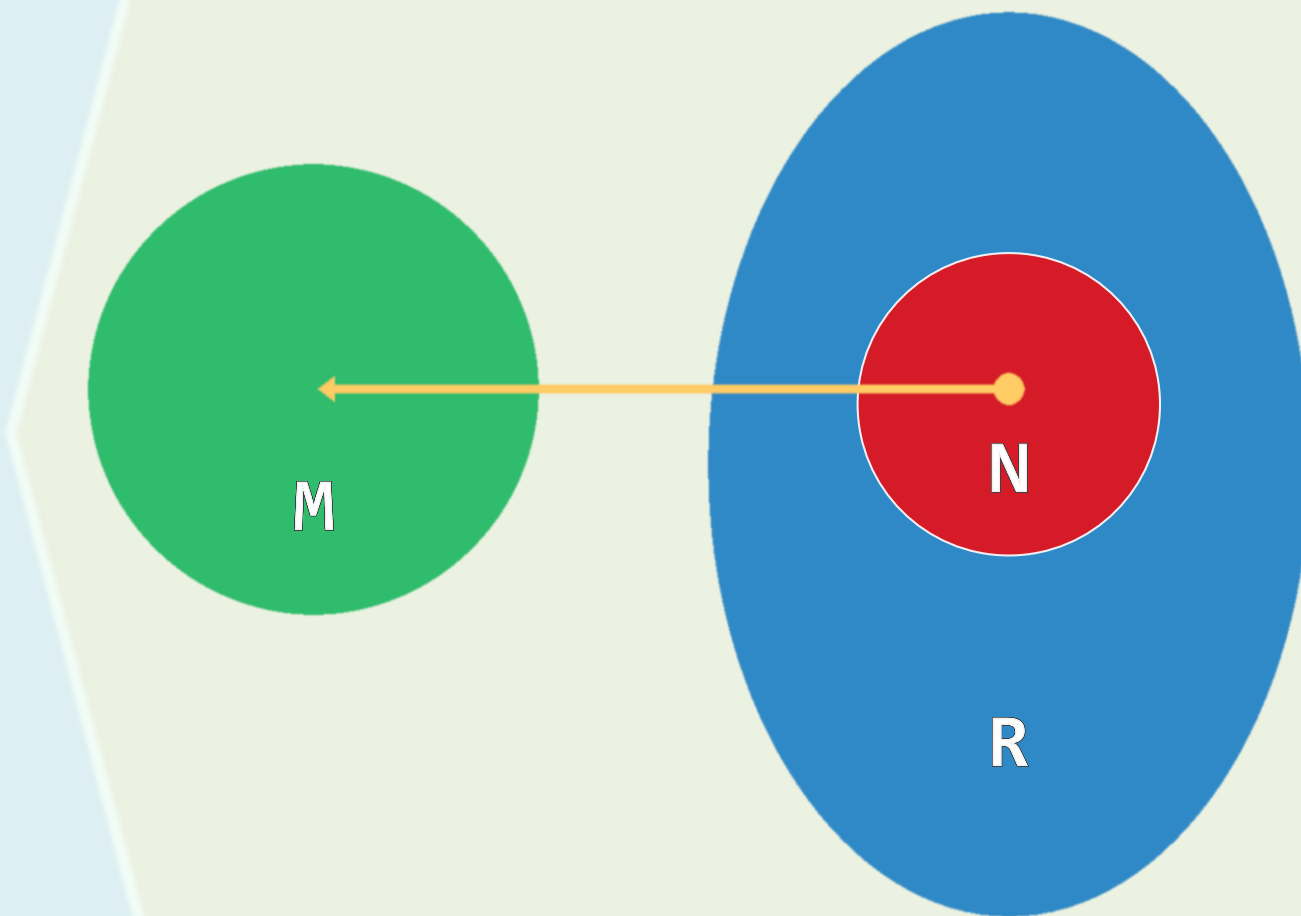
单射 (injection) 式散列，比如...

❖ 数据集已知且固定时，可实现

完美散列 (perfect hashing)

- 采用两级散列模式
- 仅需 $O(n)$ 空间
- 关键码之间互不冲突
- 最坏情况下的查找时间也不过 $O(1)$

❖ 不过在一般情况下，完美散列可期不可求...



生日悖论

❖ 将在座同学（对应的词条）按生日（月/日）做散列存储

散列表长 $M = 365$ ，装填因子 = 在场人数 $N / 365$

❖ 冲突（至少有两位同学生日相同）的可能性 $P_{365}(n) = ?$

// 概率论与数理统计讲义第一章，清华大学数学系王晓峰

$P_{365}(21) = 44.4\%$ ， $P_{365}(22) = 47.6\%$ ，...， $P_{365}(23) = 50.7\%$ // $23/365 = 6.3\%$

❖ 100人的集会： $1 - p_{365}(100) = 0.000,031\%$

- 自7岁起，不吃不喝、无休无息，每小时参加四次
- 到100岁，才有可能期望遇到一次无冲突的集会

❖ 因此，在装填因子确定之后，散列策略的选取将至关重要，散列函数的设计也很有讲究...

两项基本任务

- ❖ 首先（下一节）：精心**设计**散列表及散列函数，尽可能**降低**冲突的概率
- ❖ 同时（再下节）：制定可行的**预案**，以便在发生冲突时，能够尽快予以**排解**

