系統設計的...

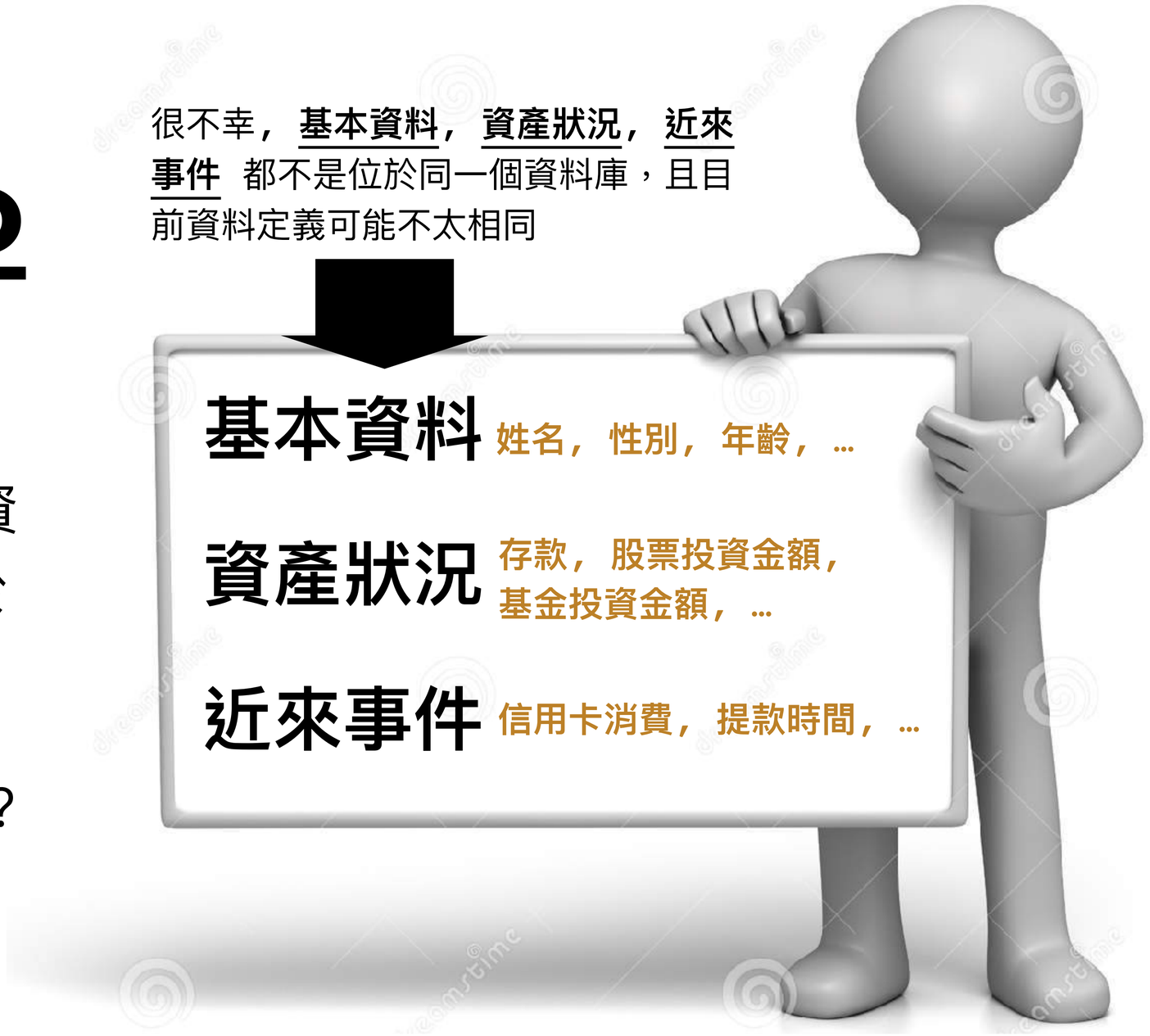
RC

Scenario

如果想要建立一個服務

給予**客戶ID**服務回傳相關資訊，如右邊所列，然後需於秒級內回覆。

那請問，我們要如何開始呢？



第一步會是

第二步會是

第三步會是

我的第一步

跟使用者再次確認需求

我的第二步

特性決定使用元件

- 秒級內回覆
- 水平擴充容易
- 各查詢服務依賴性低
- 動態決定每個欲查詢ID查詢路徑

我的第三步

系統架構圖

系統監控機制

資料流程圖

那我們今天就來實做以下情境

1 給予一個ID

2.1 查詢基本資料

2.2 查詢資產資料

2.3.a 查詢04/17 02:00–04:00的事件

2.3.b 查詢04/17 04:00–06:00的事件

3.1 彙整資料並儲存至 DB

我使用的工具

http://kafka-python.readthedocs.io/en/master/

```
>>> from kafka import KafkaProducer
>>> producer = KafkaProducer(bootstrap_servers='localhost:1234')
>>> for _ in range(100):
...     producer.send('foobar', b'some_message_bytes')
```

生產者

```
>>> from kafka import KafkaConsumer
>>> consumer = KafkaConsumer('my_favorite_topic')
>>> for msg in consumer:
...     print (msg)
```

消費者

notebook time

pymongo
kaka-python

Data

人

身分證字號

```
{“姓名”: RC,
 “性別”: 男性,
 “年齡”: 36,
 “存款”: 26,000,
 “基金”: 13,111,
 “股票”: 341,
 …
 …
}
```

資料

資料序號

人

身分證字號

```
{“姓名”: RC,
 “性別”: 男性,
 “年齡”: 36,
 “存款”: 26,000,
 “基金”: 13,111,
 “股票”: 341,
 …
 …
}
```

1. 照key值順序排列
2. 串接 value 值
3. 對其取 MD5 運算

資料

資料序號

人

DNA序列

```
{“姓名”: RC,
 “性別”: 男性,
 “年齡”: 36,
 “存款”: 26,000,
 “基金”: 13,111,
 “股票”: 341,
 …
 …
}
```

?

資料

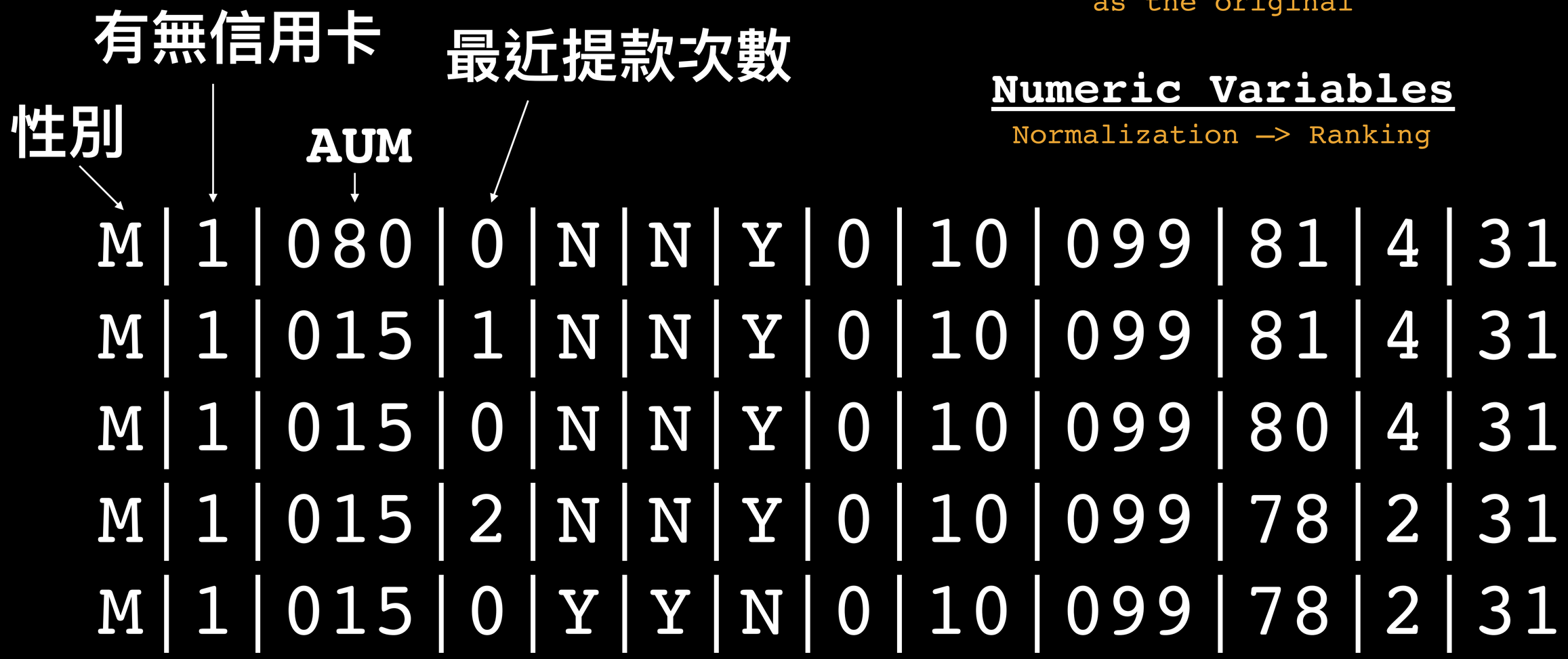
資料DNA

Categorical Variables

as the original

Numeric Variables

Normalization –> Ranking



M	1	080	0	N	N	Y	0	10	099	81	4	31
M	1	015	1	N	N	Y	0	10	099	81	4	31
M	1	015	0	N	N	Y	0	10	099	80	4	31
M	1	015	2	N	N	Y	0	10	099	78	2	31
M	1	015	0	Y	Y	N	0	10	099	78	2	31

Rolling Hash Algorithm

vs.

BLAST

https://zh.wikipedia.org/wiki/BLAST_(生物資訊學)

Rolling Hash Algorithm

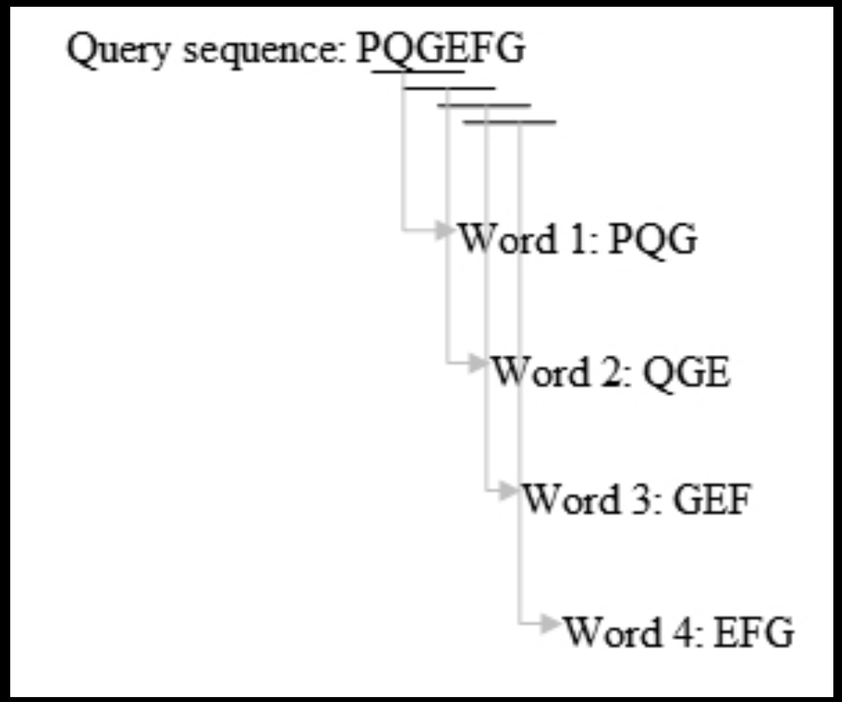
local sensitive hash algorithm

	MD5 Hash	Fuzzy Hash
abcdef	e80b5017098950fc58aad83c8c14978e	394,398,402
bcdefg	02b04a9c28417ea6b6657b638497e63e	398,402,406

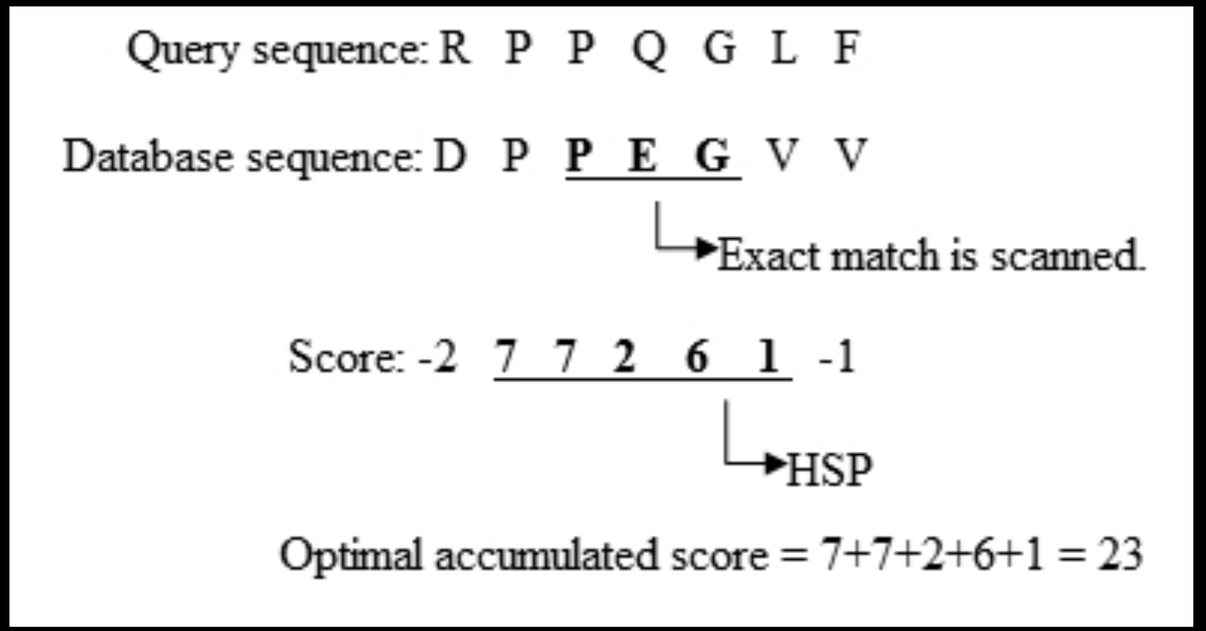
BLAST

Sequence - Sensitive

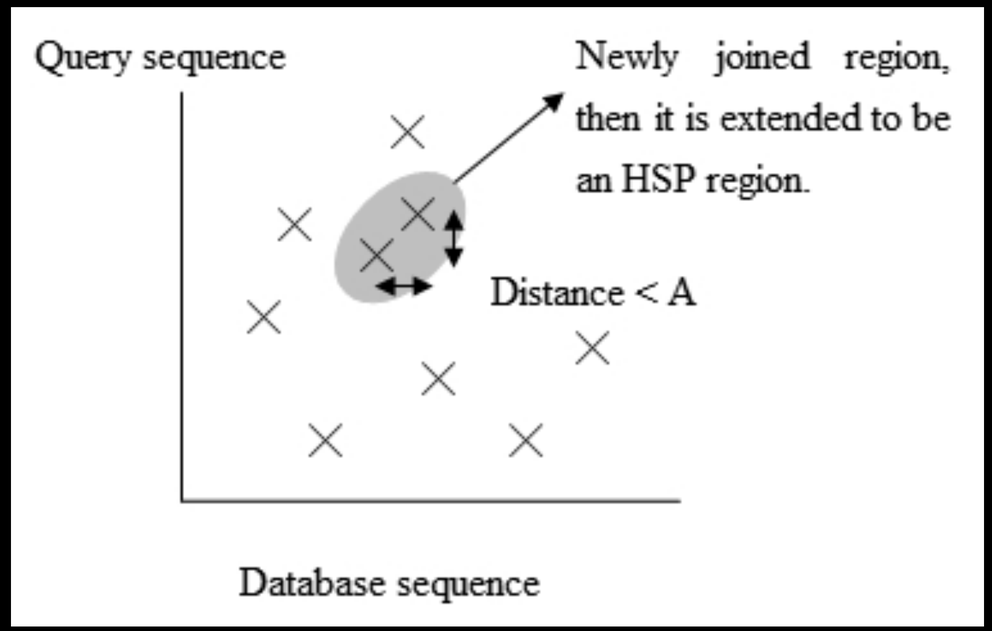
1



2



3



可能應用

- 快速找尋相似的人
  - 省卻複雜的 SQL 撰寫？
  - Ex.,
    - 我知道「Roger」是潛力VIP, 那麼我想要拿他去黏「我們資料庫」跟他相像的人
      - 過去用將屬性向量化，做距離計算
      - 以後給予 DNA 直接搜尋
- 建立模型效益的 **baseline**
  - 用目標 DNA（Rule Based） 找到的那群人，理論上應該有比亂猜好
  - 那麼模型應該要比這方法好

累

累·忙

累·忙·瞎