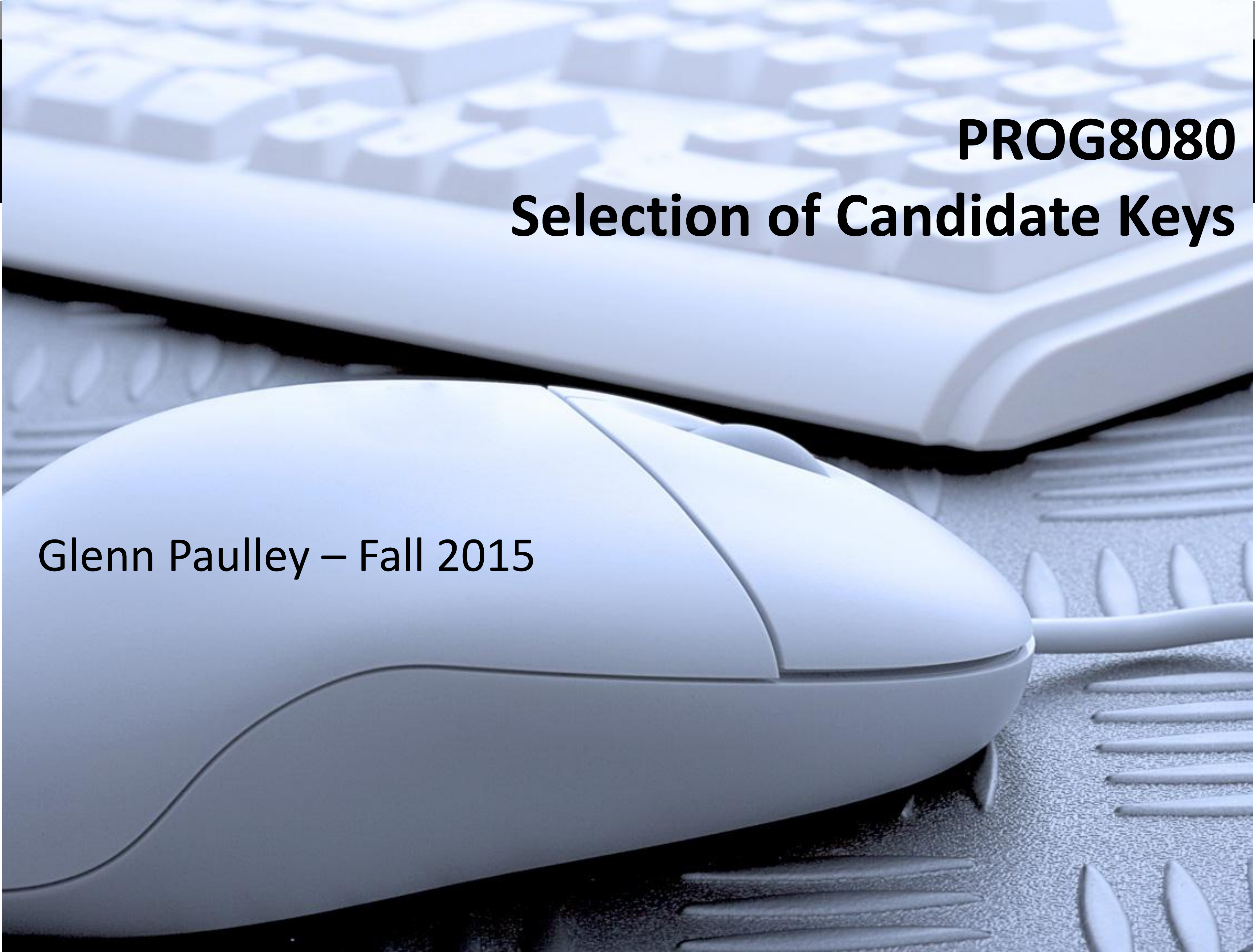

PROG8080
Selection of Candidate Keys

Glenn Paulley – Fall 2015

Characteristics of candidate keys

- Uniqueness
  - An obvious requirement for a key
  - Names, phone numbers are usually inappropriate, though still serve as useful, if not essential, search terms
    - Bell Canada: maintains both a client ID, and an account number, along with phone numbers
- Control
  - Is the key value under the application's complete control? Or is the value under the control of an independent organization, for example the Province of Ontario or the Government of Canada?
- Immutability
  - Updating primary keys is rarely a good idea
    - System maintenance can be complex – need to update all related tuples in other tables
    - Potential loss of historical information

Characteristics of candidate keys

- Data type
  - Alphabetic keys have implications for internationalization
  - Never use floating point values for keys
- Entropy and range
  - Running out of “room” is problematic because it necessitates both schema and application changes
  - How many bits does it take to store the key?
  - How many bits are required to effect a change in the key value?
  - How many bits are different between the letters ‘a’ and ‘b’ in 7-bit ASCII?
- Generation technique
  - System-generated or application-generated?
  - There are significant performance implications given the choices because an application-generated key will likely require additional SQL requests to insert a new row

Characteristics of candidate keys

- Composite or simple
    - Simple keys of a single column are significantly easier to deal with in an application
    - Simplifies the logical and physical schema
        - uses less storage

Characteristics of candidate keys

- Scope
  - Column, table, database, global?
  - IDENTITY column in Microsoft SQL Server: column
    - A fixed NUMERIC type, such as UNSIGNED BIGINT or UNSIGNED INTEGER, incremented by 1 with each insertion
  - SEQUENCE in ORACLE or SQL Server: database
    - Fixed numeric type, incremented by a pre-specified amount with each NEXT() function call
  - GUID: global
    - Many systems support a method call (such as NEWID()) that generates a 20-digit alphanumeric value that is (almost) guaranteed to be unique

Characteristics of candidate keys

- Format
  - Different key formats can be useful to differentiate between business entities
    - GHK009 – client identifier (note: alphanumeric; Latin characters; no vowels)
    - 897765 – group insurance policy number
  - Fast, easy differentiation is useful
- Self-checking
  - Self-checking keys can aid in administering client data
  - Examples: American Express numbers, Canadian SIN numbers
  - SINs in Canada: 8 digits plus a “check” digit
    - Computed using Luhn’s algorithm: see http://en.wikipedia.org/wiki/Luhn_Algorithm

Characteristics of candidate keys

- Embedded information
  - Often embedded information is held within a identifier; SIN is one example
    - First digit of a Canadian SIN number indicates the region in which the number was issued
- Input difficulty
  - Can help to reduce data entry errors; example: Canadian postal codes
    - Very difficult to touch-type a postal code even for superb typists
    - Significant reduction in misdirected mail