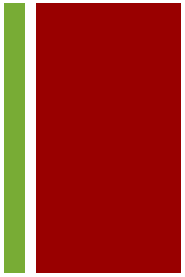
Tutorial 7: Data Quality Management

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

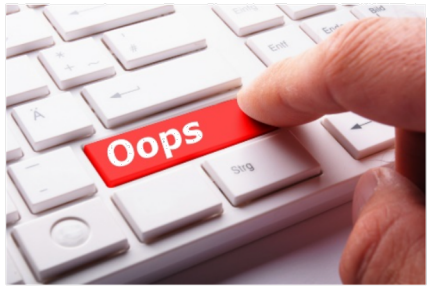
+ Q1. Quality Dimension



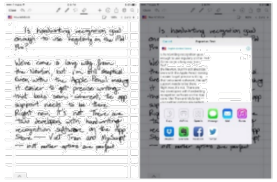
- Explain the following data quality dimensions with examples.
  1. Accuracy
  2. Representational Consistency
  3. Completeness
  4. Timeliness
  5. Accessibility

Q1-1. Accuracy

- The closeness between a value v and a value v'
  - v: the correct representation of the real-life phenomenon
  - v' aims to represent v
    - Input from Human
      - Data entry errors
      - Poor input validation
      - Lack of user training



    - Input from Photos & Apps
      - Character recognition in document management systems
      - OCR: Optical Character Recognition
      - Still not 100% Accurate



Q1-1. Accuracy

- Syntactic Accuracy
  - Check whether v' is any one of the values in domain D
    - Example: Employee (**#ID**, Country, …)
      - $v = (\#157,\ Japan,\ \ldots)$
      - $v' = (\#157,\ Switzerland,\ \ldots)$

        Syntactically Correct
      - $v = (\#157,\ Japan,\ \ldots)$
      - $v' = (\#157,\ Switzland,\ \ldots)$

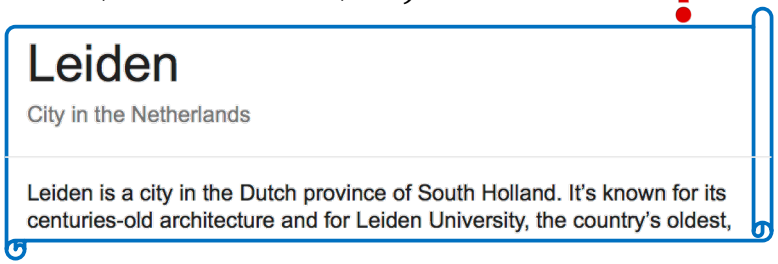
        Syntactically Incorrect

        Because $Switzland$ is not a valid country name

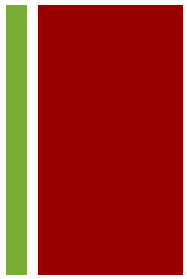
+ Q1-1. Accuracy

- Semantic Accuracy
  - More complex
  - True value has to be known
    (with additional knowledge)
    - Example: City (CityName, Country, …)
      - $v_1 = (\text{Tokyo}, \ Japan, \ \ldots)$
      - $v_2 = (\text{Osaka}, \ Japan, \ \ldots)$
      - $v_3 = (\text{Zürich}, \ Switzerland, \ \ldots)$
      - $v' = (\text{Tokyo}, \ Switzerland, \ \ldots)$

      Semantically Incorrect, Tokyo is not a city in Switzerland
      - $v' = (\text{Leiden}, \ Switzerland, \ \ldots)$?



+ Q1-2. Representational Consistency



- Different values that refers to the same entity
  - USA vs. United States
  - UQ vs. University of Queensland
  - 09/10/2018 vs. 9/10/2018

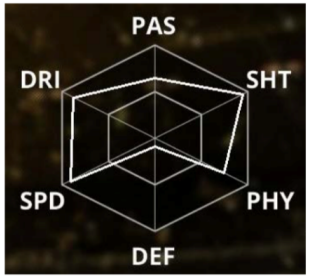
    09/10/2018 vs. 10/09/2018

    09/10/2018 vs. 09.10.2018

+ Q1-3. Completeness

- Sufficiency of data for the task at hand
  - Schema completeness
    - In football game
      - Player name, club, nationality, age, height, position, overall rating, passing, shooting, physical, defense, speed, dribbling…

Name	Club	Nat	Age	Height	Pos	Ovr	Pas	Sht	Phy
L. Messi			30	170	RWF	94	88	95	66
C. Ronaldo			32	185	LWF	94	83	95	87
L. Suárez			30	182	CF	92	82	95	87
Neymar			25	175	LWF	91	82	86	69
M. Neuer			31	193	GK	91	65	43	88
Z. Ibrahimović			36	195	CF	90	85	90	95



THE
OF
AUS

+ Q1-3. Completeness

- Sufficiency of data for the task at hand
  - Column completeness
    - Missing values

ID	Name	Surname	Birth Date	Email
1	John	Smith	03/17/1974	smith@abc.au
2	Edward	Monroe	02/03/1967	NULL
3	Anthony	White	01/01/1936	NULL
4	Marianne	Collins	11/20/1955	NULL

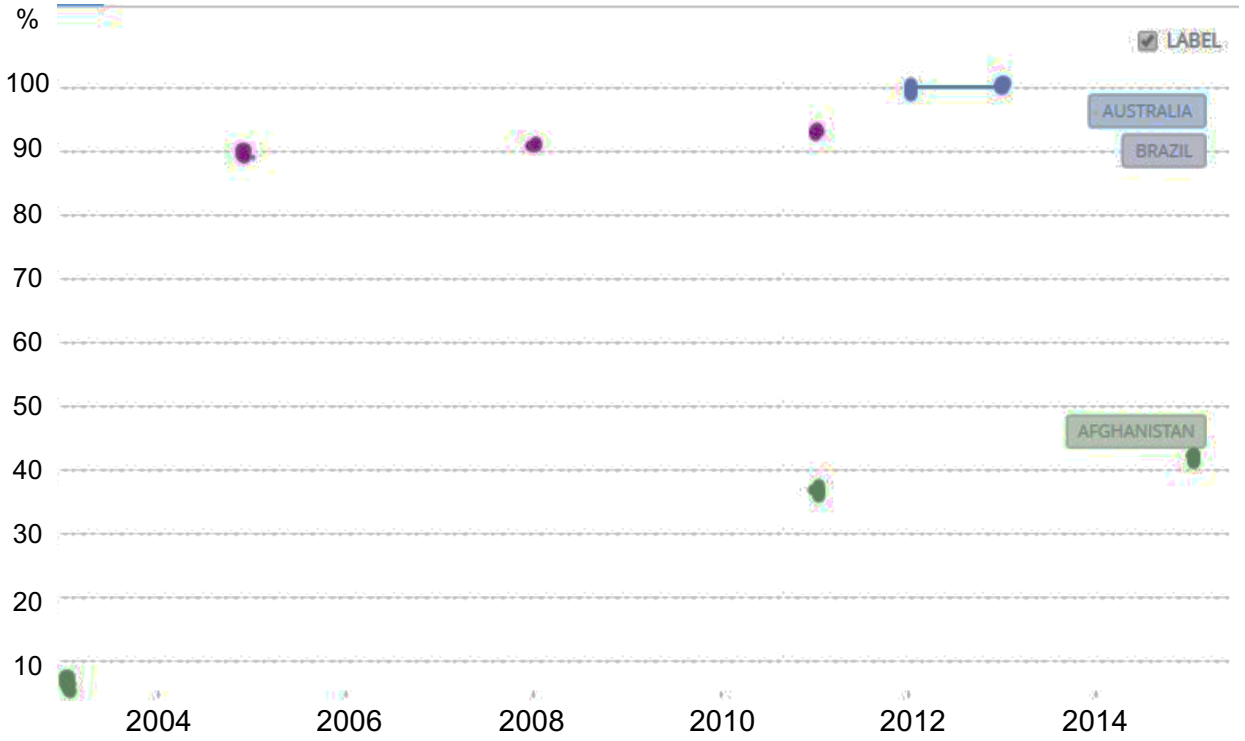
Not existing

Existing but unknown

Not known if existing

+ Q1-3. Completeness

- Sufficiency of data for the task at hand
  - Population completeness
    - Missing values with respect to a reference population
    - Completeness of birth registration from World Bank



THE
OF
AUS

+ Q1-4. Timeliness

- How current the data is for the task at hand
  - if late for specific usage → useless
  - measured by the time between when data is expected and when it is available/ready for use
  - decision making
- Example
  - 1 hour before football match begins, the starting lineups are submitted. C. Ronaldo is on substitution bench as rumored.
  - 15 minutes before match starts, some player got injured, and Ronaldo goes back to the starting lineup
  - The referee has the current data
  - The opponent does not
    - Even he does, it is too late

+ Q1-5. Accessibility

- The ability of the user to access the data from his own culture, physical status/functions, and technologies available
  - Mongolian user accesses Arabic data
  - Blind user accesses photo data
  - Server is down

+ Q1-a. Student Table

Surname	First Name	Age	Major	Degree	Gender
Barrat	John	22	Math	BSc	Male
Burns	Robert	24	CS	BSc	Male
Carter	Laura	20	Physics	MSc	Female
Davies	Michael	12	CS	BSc	Male

+ Q1-b. Driving Test Table

Surname	First Name	DoB	Driving Test Passed
Smith	J.	17/12/85	12/12/05
Smith	Jack	17/12/85	12/12/2005
Smith	Jock	17/12/95	12/12/2005

+ Q2. Edit Distance

Question 2: In a company, the *reference table* consists of the clean data on the surname and given name of its employees. Given a set of input records, the duplicates can be identified by checking against the reference table. In other words, two input records are regarded as duplicates if they are linked to the same tuple in the reference table. Suppose that we use Edit distance as the similarity measure, and use dynamic programming matrix to calculate Edit distance between two strings.

ID	Surname	Given name
1	Duboice	Nicholas
2	Hansson	William
3	Wang	Sean
4	Dubosh	Nick
5	Garnette	Kevin
6	Pitt	William
7	Hanschen	Williams
8	Aniston	Jennifer
9	Zhou	Xuan

reference table

ID	Surname	Given name	Expense
001	Duboise	Nicholas	130
002	Duboch	Nicolas	30
003	Hanson	Williams	47

input records

THE
OF C
AUST

Q2. Edit Distance

- **D**elete, **I**nsert, **R**eplace, Do **N**othing
  - 4 choice
  - Complexity $O(4^n)$ $\quad n = \max\{r,\ s\}$

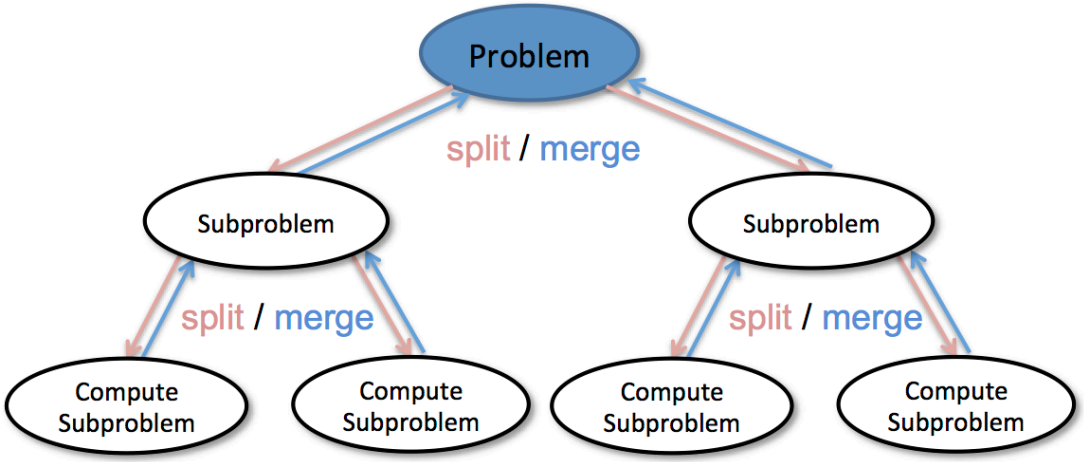
DDIR
DDIN
DDII
DDID
DDRR
DDRN
…

with memorization, $O(r * s)$

		j	o	h	n
	0	1	2	3	4
j	1	0	1	2	3
h	2	1	1	1	2
n	3	2	2	2	1

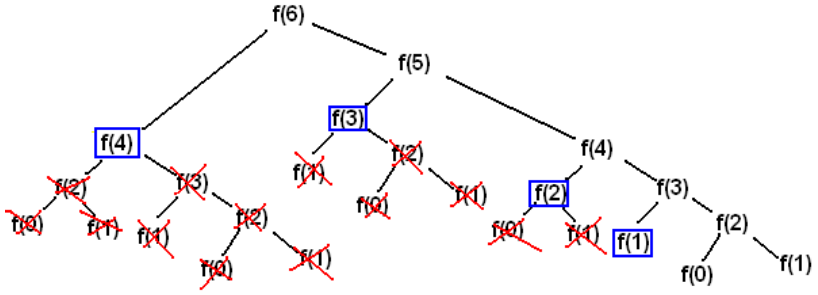
Q2. Edit Distance

- Divide-and-Conquer
  - Partition the problem into disjoint sub-problems
  - Solve the sub-problems **recursively**
  - Combine their solutions to solve the original problem



+ Q2. Dynamic Programming

- Dynamic Programming
  - The sub-problems share common sub-sub-problems
  - Solve each sub-sub-problems once
  - Save in a table, avoid re-computing
  - Use memory to trade time
  - Find the optimal solution among the solutions



		j	o	h	n
	0	1	2	3	4
j	1	0	1	2	3
h	2	1	1	1	2
n	3	2	2	2	1

+ Q2 Steps

1. Characterize the structure of an optimal solution.
2. Recursively define the value of an optimal solution.
3. Compute the value of an optimal solution, typically in a bottom-up fashion.
4. Construct an optimal solution from computed information

$$d_{ij} = \begin{cases} d_{i-1,\,j-1} & (a_j = b_i) \\ \min \begin{pmatrix} d_{i-1,\,j} + 1 \\ d_{i,\,j-1} + 1 \\ d_{i-1,\,j-1} + 1 \end{pmatrix} & (a_j \neq b_i) \end{cases}$$

		j	o	h	n
	0	1	2	3	4
j	1	0	1	2	3
h	2	1	1	1	2
n	3	2	2	2	1

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

+ Q2 Edit Distance

- (a) Considering surname only, are there any duplicates in the following input records?
- (b) When both surname and given name are considered (with the same importance), are there any duplicates in the following input records?

ID	Surname	Given name
1	Duboice	Nicholas
2	Hansson	William
3	Wang	Sean
4	Dubosh	Nick
5	Garnette	Kevin
6	Pitt	William
7	Hanschen	Williams
8	Aniston	Jennifer
9	Zhou	Xuan

reference table

ID	Surname	Given name	Expense
001	Duboise	Nicholas	130
002	Duboch	Nicolas	30
003	Hanson	Williams	47

input records

Edit distance

+ Q2–(a)

Edit distance

- $d_{1,\ 001} = 1/7 = 0.143$ $\quad d_{1,\ 002} = 2/7 = 0.286$
- $d_{4,\ 001} = 2/7 = 0.286$ $\quad d_{4,\ 002} = 1/6 = 0.167$
- $d_{2,\ 003} = 1/7 = 0.143$ $\quad d_{7,\ 003} = 3/8 = 0.375$

ID	Surname	Given name
1	Duboice	Nicholas
2	Hansson	William
3	Wang	Sean
4	Dubosh	Nick
5	Garnette	Kevin
6	Pitt	William
7	Hanschen	Williams
8	Aniston	Jennifer
9	Zhou	Xuan

reference table

ID	Surname	Given name	Expense
001	Duboise	Nicholas	130
002	Duboch	Nicolas	30
003	Hanson	Williams	47

ut records

$assume\ threshold = 0.2$

THE
OF C
AUST

Q2–(a)

- $d_{1,\ 001} = 1/7 = 0.143$
- $d_{4,\ 001} = 2/7 = 0.286$
- $d_{2,\ 003} = 1/7 = 0.143$

$d_{1,\ 002} = 2/7 = 0.286$

$d_{4,\ 002} = 1/6 = 0.167$

$d_{7,\ 003} = 3/8 = 0.375$

ID	Surname	Given name
1	Duboice	Nicholas
2	Hansson	William
3	Wang	Sean
4	Dubosh	Nick
5	Garnette	Kevin
6	Pitt	William
7	Hanschen	Williams
8	Aniston	Jennifer
9	Zhou	Xuan

reference table

ID	Surname	Given name	Expense
001	Duboise	Nicholas	130
002	Duboch	Nicolas	30
003	Hanson	Williams	47

input records

$assume\ threshold = 0.2$

+ Q2–(b)

- $d_{1,\ 001} = 1/16 = 0.063$

1. **Duboice#Nicholas**

001. **Duboise#Nicholas**

ID	Surname	Given name
1	Duboice	Nicholas
2	Hansson	William
3	Wang	Sean
4	Dubosh	Nick
5	Garnette	Kevin
6	Pitt	William
7	Hanschen	Williams
8	Aniston	Jennifer
9	Zhou	Xuan

reference table

ID	Surname	Given name	Expense
001	Duboise	Nicholas	130
002	Duboch	Nicolas	30
003	Hanson	Williams	47

input records

THE
OF C
AUST

+ Q2–(b)

- $d_{1,\ 001} = 1/16 = 0.063$ $d_{1,\ 002} = 3/16 = 0.188$

1. **Duboice#Nicholas**

002. **Dubo ch#Nic olas**

ID	Surname	Given name
1	Duboice	Nicholas
2	Hansson	William
3	Wang	Sean
4	Dubosh	Nick
5	Garnette	Kevin
6	Pitt	William
7	Hanschen	Williams
8	Aniston	Jennifer
9	Zhou	Xuan

reference table

ID	Surname	Given name	Expense
001	Duboise	Nicholas	130
002	Duboch	Nicolas	30
003	Hanson	Williams	47

input records

THE
OF C
AUST

+ Q2–(b)

- $d_{1,\ 001} = 1/16 = 0.063$ $\quad d_{1,\ 002} = 3/16 = 0.188$
- $d_{4,\ 001} = 6/16 = 0.375$

ID	Surname	Given name
1	Duboice	Nicholas
2	Hansson	William
3	Wang	Sean
4	Dubosh	Nick
5	Garnette	Kevin
6	Pitt	William
7	Hanschen	Williams
8	Aniston	Jennifer
9	Zhou	Xuan

reference table

ID	Surname	Given name	Expense
001	Duboise	Nicholas	130
002	Duboch	Nicolas	30
003	Hanson	Williams	47

input records

4. **Dubo sh#Nick**

001. **Duboise#Nicholas**

THE
OF C
AUST

+ Q2–(b)

- $d_{1,\ 001} = 1/16 = 0.063$ $\quad d_{1,\ 002} = 3/16 = 0.188$
- $d_{4,\ 001} = 6/16 = 0.375$ $\quad d_{4,\ 002} = 5/14 = 0.357$

ID	Surname	Given name
1	Duboice	Nicholas
2	Hansson	William
3	Wang	Sean
4	Dubosh	Nick
5	Garnette	Kevin
6	Pitt	William
7	Hanschen	Williams
8	Aniston	Jennifer
9	Zhou	Xuan

reference table

ID	Surname	Given name	Expense
001	Duboise	Nicholas	130
002	Duboch	Nicolas	30
003	Hanson	Williams	47

input records

4. **Dubosh#Nick**

002. **Duboch#Nicolas**

+ Q2–(b)

- $d_{1,\ 001} = 1/16 = 0.063$ $\quad d_{1,\ 002} = 3/16 = 0.188$
- $d_{4,\ 001} = 6/16 = 0.375$ $\quad d_{4,\ 002} = 5/14 = 0.357$
- $d_{2,\ 003} = 2/15 = 0.133$

ID	Surname	Given name
1	Duboice	Nicholas
2	Hansson	William
3	Wang	Sean
4	Dubosh	Nick
5	Garnette	Kevin
6	Pitt	William
7	Hanschen	Williams
8	Aniston	Jennifer
9	Zhou	Xuan

reference table

ID	Surname	Given name	Expense
001	Duboise	Nicholas	130
002	Duboch	Nicolas	30
003	Hanson	Williams	47

input records

2. **Hansson#William**

003. **Hans on#Williams**

THE
OF C
AUST

+ Q2–(b)

- $d_{1,\ 001} = 1/16 = 0.063$ $\quad d_{1,\ 002} = 3/16 = 0.188$
- $d_{4,\ 001} = 6/16 = 0.375$ $\quad d_{4,\ 002} = 5/14 = 0.357$
- $d_{2,\ 003} = 2/15 = 0.133$ $\quad d_{7,\ 003} = 3/17 = 0.176$

ID	Surname	Given name
1	Duboice	Nicholas
2	Hansson	William
3	Wang	Sean
4	Dubosh	Nick
5	Garnette	Kevin
6	Pitt	William
7	Hanschen	Williams
8	Aniston	Jennifer
9	Zhou	Xuan

reference table

ID	Surname	Given name	Expense
001	Duboise	Nicholas	130
002	Duboch	Nicolas	30
003	Hanson	Williams	47

input records

7. **Hanschen#Williams**

003. **Hans on#Williams**

THE
OF C
AUST

+ Q2–(b)

- $d_{1,\ 001} = 1/16 = 0.063$ $\quad d_{1,\ 002} = 3/16 = 0.188$
- $d_{4,\ 001} = 6/16 = 0.375$ $\quad d_{4,\ 002} = 5/14 = 0.357$
- $d_{2,\ 003} = 2/15 = 0.133$ $\quad d_{7,\ 003} = 3/17 = 0.176$

ID	Surname	Given name
1	Duboice	Nicholas
2	Hansson	William
3	Wang	Sean
4	Dubosh	Nick
5	Garnette	Kevin
6	Pitt	William
7	Hanschen	Williams
8	Aniston	Jennifer
9	Zhou	Xuan

reference table

ID	Surname	Given name	Expense
001	Duboise	Nicholas	130
002	Duboch	Nicolas	30
003	Hanson	Williams	47

input records

6. Pitt #William

003. Hanson#Williams

$d_{6,\ 003} = 7/15 = 0.467$

THE
OF C
AUST

+ Q2–(b)

$assume\ threshold = 0.2$

- $d_{1,\ 001} = 1/16 = 0.063$
- $d_{4,\ 001} = 6/16 = 0.375$
- $d_{2,\ 003} = 2/15 = 0.133$

$d_{1,\ 002} = 3/16 = 0.188$

$d_{4,\ 002} = 5/14 = 0.357$

$d_{7,\ 003} = 3/17 = 0.176$

ID	Surname	Given name
1	Duboice	Nicholas
2	Hansson	William
3	Wang	Sean
4	Dubosh	Nick
5	Garnette	Kevin
6	Pitt	William
7	Hanschen	Williams
8	Aniston	Jennifer
9	Zhou	Xuan

reference table

ID	Surname	Given name	Expense
001	Duboise	Nicholas	130
002	Duboch	Nicolas	30
003	Hanson	Williams	47

input records

6. Pitt #William

003. Hanson#Williams

$d_{6,\ 003} = 7/15 = 0.467$

+ Q2–(b)

$assume\ threshold = 0.2$

- $d_{1,\ 001} = 1/16 = 0.063$ $d_{1,\ 002} = 3/16 = 0.188$
- ~~$d_{4,\ 001} = 6/16 = 0.375$~~ ~~$d_{4,\ 002} = 5/14 = 0.357$~~
- $d_{2,\ 003} = 2/15 = 0.133$ $d_{7,\ 003} = 3/17 = 0.176$

ID	Surname	Given name
1	Duboice	Nicholas
2	Hansson	William
3	Wang	Sean
4	Dubosh	Nick
5	Garnette	Kevin
6	Pitt	William
7	Hanschen	Williams
8	Aniston	Jennifer
9	Zhou	Xuan

reference table

ID	Surname	Given name	Expense
001	Duboise	Nicholas	130
002	Duboch	Nicolas	30
003	Hanson	Williams	47

input records

6. **Pitt #William**

003. **Hanson#Williams**

$d_{6,\ 003} = 7/15 = 0.467$

THE
OF C
AUST

+ Q2–(b)

$assume\ threshold = 0.2$

- Conclusion
  - There are duplications in the input records
  - Both $input.001$ and $input.002$ are linked to $reference.1$

ID	Surname	Given name
1	Duboice	Nicholas
2	Hansson	William
3	Wang	Sean
4	Dubosh	Nick
5	Garnette	Kevin
6	Pitt	William
7	Hanschen	Williams
8	Aniston	Jennifer
9	Zhou	Xuan

reference table

ID	Surname	Given name	Expense
001	Duboise	Nicholas	130
002	Duboch	Nicolas	30
003	Hanson	Williams	47

input records