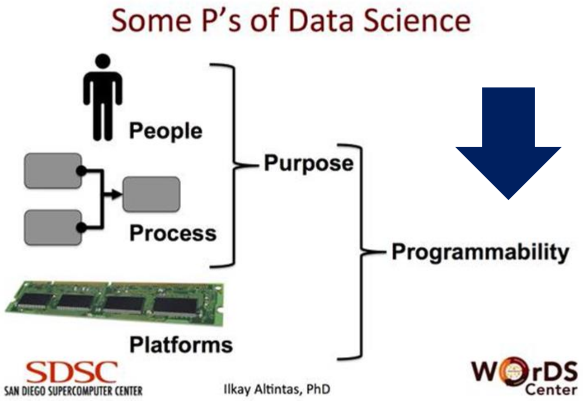
Data models (belongs to Programability)

Introduction to Big Data

Universidad
Francisco de Vitoria
UFV Madrid

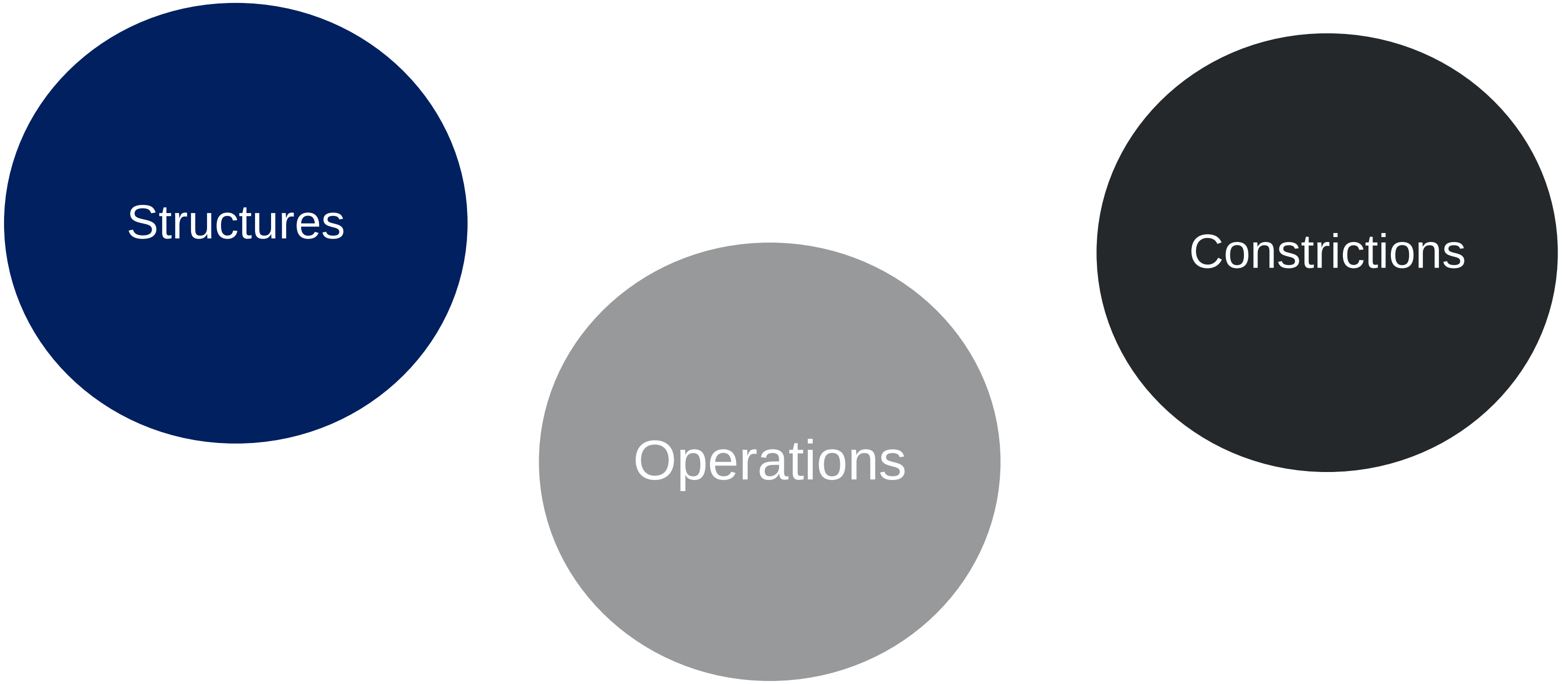
Index

- **01. Modeling**
- 01.01 Structures
- **02. Types of data models: structures**
- 02.01 Structured model: Relational model
  - 02.01 Operations
  - 02.02 Constrictions
- 03.01 Semi-structured model
- 02.03 Unstructured model: Text as a vector, intro to Vector Space model
- 03.02 Other models: Networks
- **03. Data model vs data format**



01.-Modeling

Data models describe data characteristics:



Structures

Operations

Constrictions

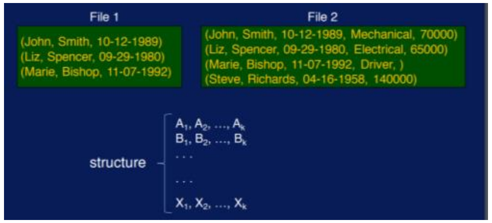
01.-Modeling

Structures

Structured

Semi structured

Unstructured



File 1

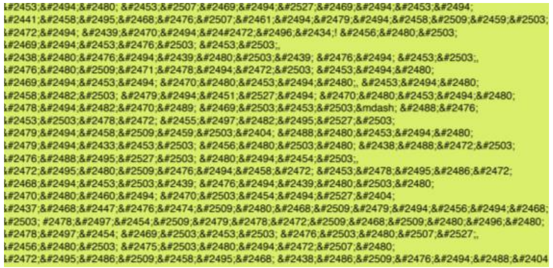
(John, Smith, 10-12-1989)
(Liz, Spencer, 09-29-1980)
(Marie, Bishop, 11-07-1992)

File 2

(John, Smith, 10-12-1989, Mechanical, 70000)
(Liz, Spencer, 09-29-1980, Electrical, 65000)
(Marie, Bishop, 11-07-1992, Driver,)
(Steve, Richards, 04-16-1958, 140000)

structure

$A_1, A_2, \ldots, A_k$
$B_1, B_2, \ldots, B_k$
. . .
. . .
$X_1, X_2, \ldots, X_k$

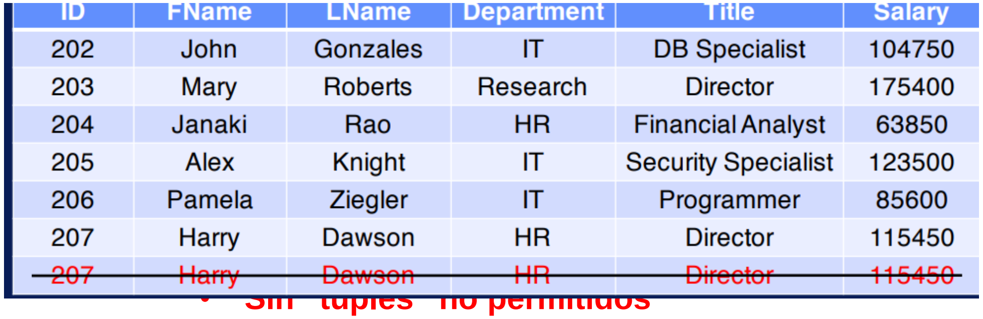


```
&#2453;&#2494;&#2480; &#2453;&#2507;&#2469;&#2494;&#2527;&#2469;&#2494;&#2453;&#2494;
&#2441;&#2458;&#2495;&#2468;&#2476;&#2507;&#2461;&#2494;&#2479;&#2494;&#2458;&#2509;&#2459;&#2503;
&#2472;&#2494; &#2439;&#2470;&#2494;&#24#2472;&#2496;&#2434;! &#2456;&#2480;&#2503;
&#2469;&#2494;&#2453;&#2476;&#2503; &#2453;&#2503;,
&#2438;&#2480;&#2476;&#2494;&#2439;&#2480;&#2503;&#2439; &#2476;&#2494; &#2453;&#2503;,
&#2476;&#2480;&#2509;&#2471;&#2478;&#2494;&#2472;&#2503; &#2453;&#2494;&#2480;
&#2469;&#2494;&#2453;&#2494; &#2470;&#2480;&#2453;&#2494;&#2480;, &#2453;&#2494;&#2480;
&#2458;&#2482;&#2503; &#2479;&#2494;&#2451;&#2527;&#2494; &#2470;&#2480;&#2453;&#2494;&#2480;
&#2478;&#2494;&#2482;&#2470;&#2489; &#2469;&#2503;&#2453;&#2503;&mdash; &#2488;&#2476;
&#2453;&#2503;&#2478;&#2472; &#2455;&#2497;&#2482;&#2495;&#2527;&#2503;
&#2479;&#2494;&#2458;&#2509;&#2459;&#2503;&#2404; &#2488;&#2480;&#2453;&#2494;&#2480;
&#2479;&#2494;&#2433;&#2453;&#2503; &#2456;&#2480;&#2503;&#2480; &#2438;&#2488;&#2472;&#2503;
&#2476;&#2488;&#2495;&#2527;&#2503; &#2480;&#2494;&#2454;&#2503;,
&#2472;&#2495;&#2480;&#2509;&#2476;&#2494;&#2458;&#2472; &#2453;&#2478;&#2495;&#2486;&#2472;
&#2468;&#2494;&#2453;&#2503;&#2439; &#2476;&#2494;&#2439;&#2480;&#2503;&#2480;
&#2470;&#2480;&#2460;&#2494; &#2470;&#2503;&#2454;&#2494;&#2527;&#2404;
&#2437;&#2468;&#2447;&#2476;&#2474;&#2509;&#2480;&#2468;&#2509;&#2479;&#2494;&#2456;&#2494;&#2468;
&#2503; #2478;&#2497;&#2454;&#2509;&#2479;&#2478;&#2472;&#2509;&#2468;&#2509;&#2480;&#2496;&#2480;
&#2478;&#2497;&#2454; &#2469;&#2503;&#2453;&#2503; &#2476;&#2503;&#2480;&#2507;&#2527;,
&#2456;&#2480;&#2503; &#2475;&#2503;&#2480;&#2494;&#2472;&#2507;&#2480;
&#2472;&#2495;&#2486;&#2509;&#2458;&#2495;&#2468; &#2438;&#2486;&#2509;&#2476;&#2494;&#2488;&#2404
```

02.-Data models: structures

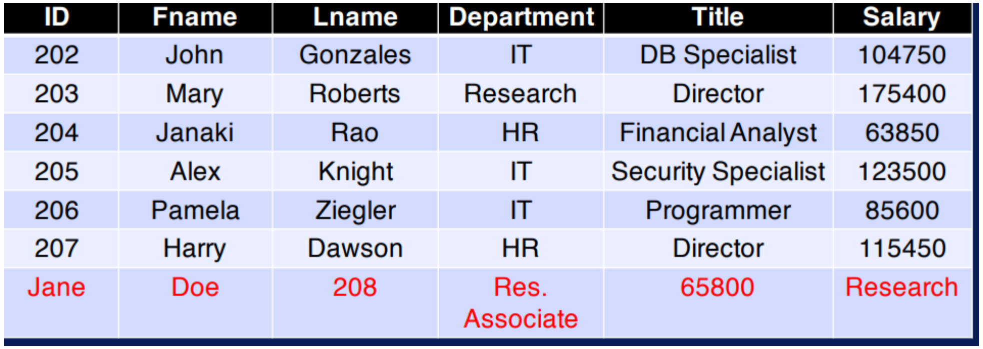
Structured

- **Relational models:**
  - **Table colection**
  - **Without duplicates**



ID	FName	LName	Department	Title	Salary
202	John	Gonzales	IT	DB Specialist	104750
203	Mary	Roberts	Research	Director	175400
204	Janaki	Rao	HR	Financial Analyst	63850
205	Alex	Knight	IT	Security Specialist	123500
206	Pamela	Ziegler	IT	Programmer	85600
207	Harry	Dawson	HR	Director	115450
~~207~~	~~Harry~~	~~Dawson~~	~~HR~~	~~Director~~	~~115450~~

- Sin "tuples" no permitidos

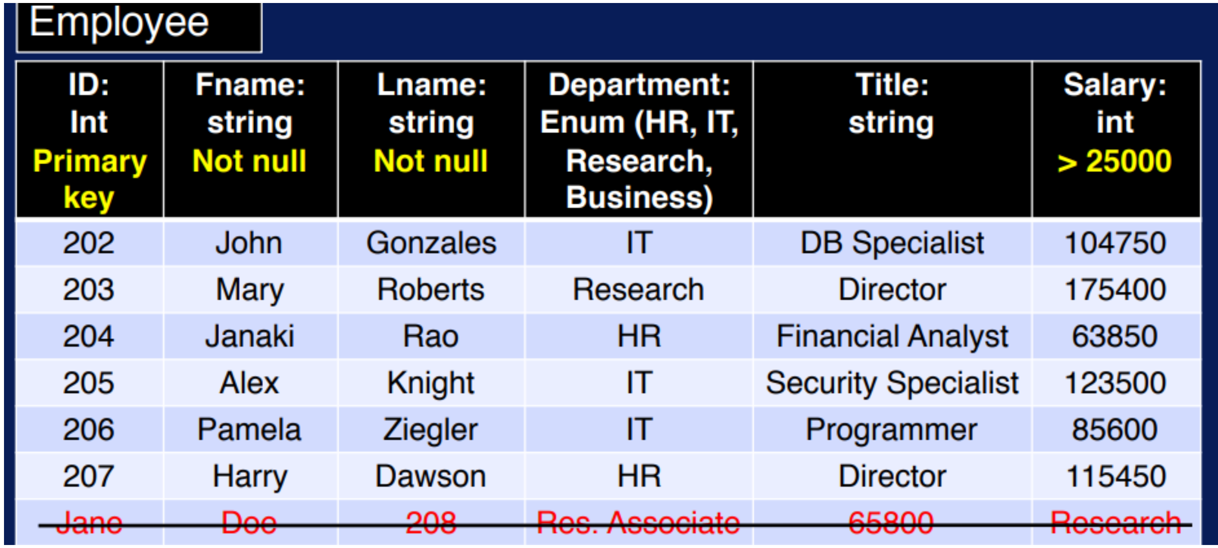


ID	Fname	Lname	Department	Title	Salary
202	John	Gonzales	IT	DB Specialist	104750
203	Mary	Roberts	Research	Director	175400
204	Janaki	Rao	HR	Financial Analyst	63850
205	Alex	Knight	IT	Security Specialist	123500
206	Pamela	Ziegler	IT	Programmer	85600
207	Harry	Dawson	HR	Director	115450
Jane	Doe	208	Res. Associate	65800	Research

02.-Data models: structures

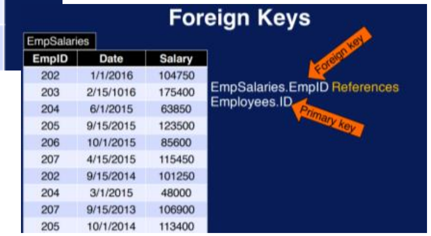
Structured

- **Relational model:**
  - **Primary Key**
  - **Constrictions**



Employee

ID: Int Primary key	Fname: string Not null	Lname: string Not null	Department: Enum (HR, IT, Research, Business)	Title: string	Salary: int > 25000
202	John	Gonzales	IT	DB Specialist	104750
203	Mary	Roberts	Research	Director	175400
204	Janaki	Rao	HR	Financial Analyst	63850
205	Alex	Knight	IT	Security Specialist	123500
206	Pamela	Ziegler	IT	Programmer	85600
207	Harry	Dawson	HR	Director	115450
~~Jane~~	~~Doe~~	~~208~~	~~Res. Associate~~	~~65800~~	~~Research~~



Foreign Keys

EmpSalaries

EmpID	Date	Salary
202	1/1/2016	104750
203	2/15/1016	175400
204	6/1/2015	63850
205	9/15/2015	123500
206	10/1/2015	85600
207	4/15/2015	115450
202	9/15/2014	101250
204	3/1/2015	48000
207	9/15/2013	106900
205	10/1/2014	113400

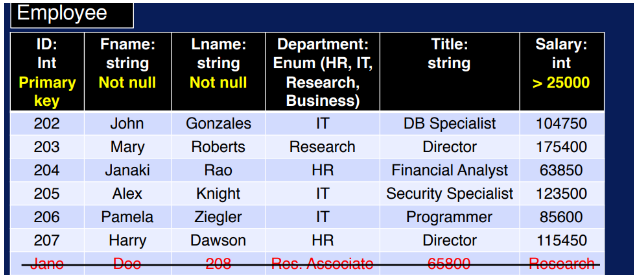
EmpSalaries.EmpID References Employees.ID

Foreign key
Primary key

02.-Data models: structures

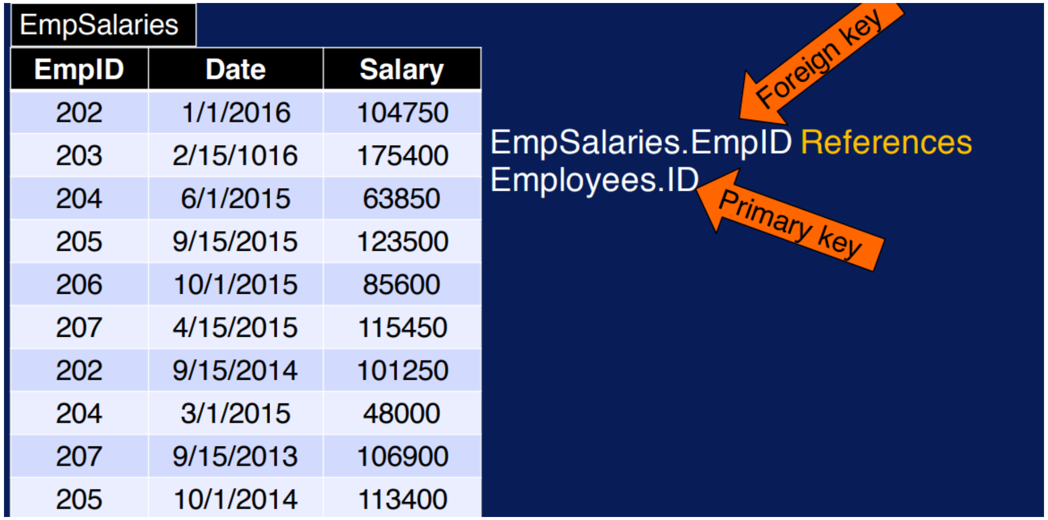
Structured

- Relational models:
  - Foreign key
  - Table relationships through the Foreing keys



Employee

ID: Int Primary key	Fname: string Not null	Lname: string Not null	Department: Enum (HR, IT, Research, Business)	Title: string	Salary: int > 25000
202	John	Gonzales	IT	DB Specialist	104750
203	Mary	Roberts	Research	Director	175400
204	Janaki	Rao	HR	Financial Analyst	63850
205	Alex	Knight	IT	Security Specialist	123500
206	Pamela	Ziegler	IT	Programmer	85600
207	Harry	Dawson	HR	Director	115450
~~Jane~~	~~Doe~~	~~208~~	~~Res. Associate~~	~~65800~~	~~Research~~



EmpSalaries

EmpID	Date	Salary
202	1/1/2016	104750
203	2/15/1016	175400
204	6/1/2015	63850
205	9/15/2015	123500
206	10/1/2015	85600
207	4/15/2015	115450
202	9/15/2014	101250
204	3/1/2015	48000
207	9/15/2013	106900
205	10/1/2014	113400

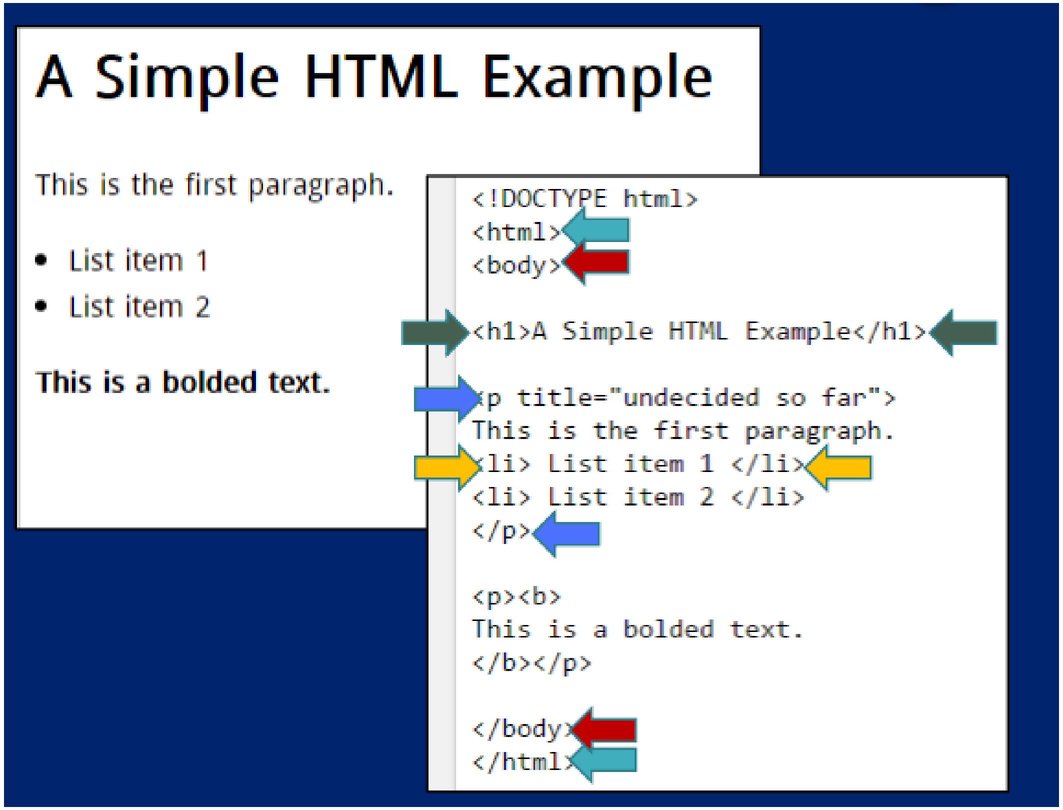
EmpSalaries.EmpID References Employees.ID

Foreign key
Primary key

02.-Data models: structures

Semi structured

- Semi structured
  - HTML, XML, JSON,…



A Simple HTML Example

This is the first paragraph.

- List item 1
- List item 2

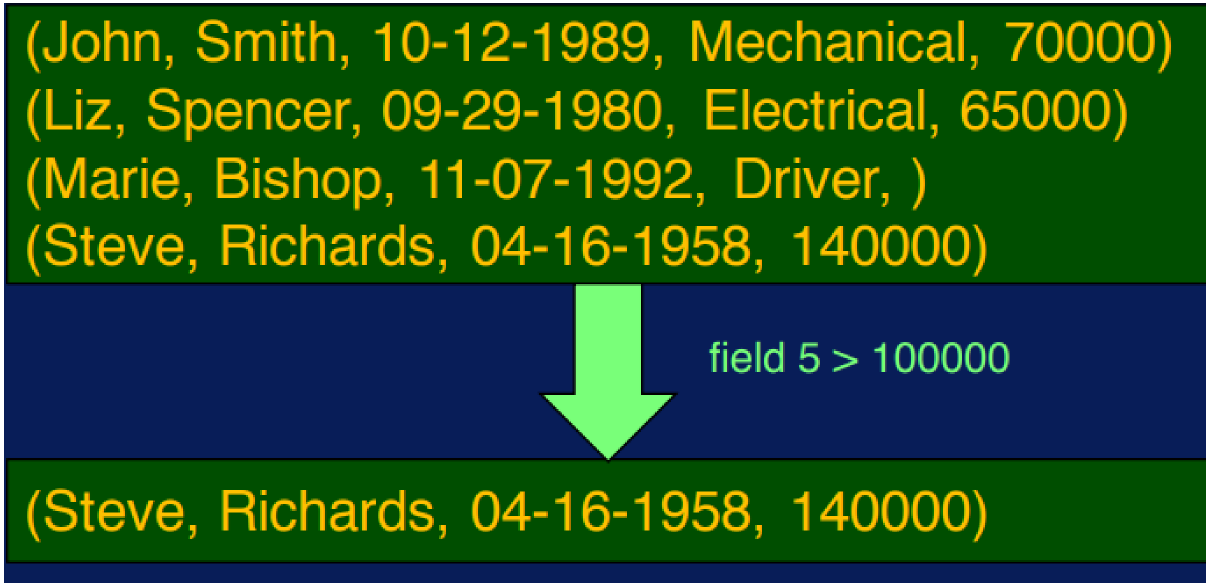
This is a bolded text.

```
<!DOCTYPE html>
<html>
<body>

<h1>A Simple HTML Example</h1>

<p title="undecided so far">
This is the first paragraph.
<li> List item 1 </li>
<li> List item 2 </li>
</p>

<p><b>
This is a bolded text.
</b></p>

</body>
</html>
```

01.-Modeling

Operations

- **Sub setting:** given a data set and a condition
  - Find a subset that fulfils the condition



(John, Smith, 10-12-1989, Mechanical, 70000)
(Liz, Spencer, 09-29-1980, Electrical, 65000)
(Marie, Bishop, 11-07-1992, Driver,)
(Steve, Richards, 04-16-1958, 140000)

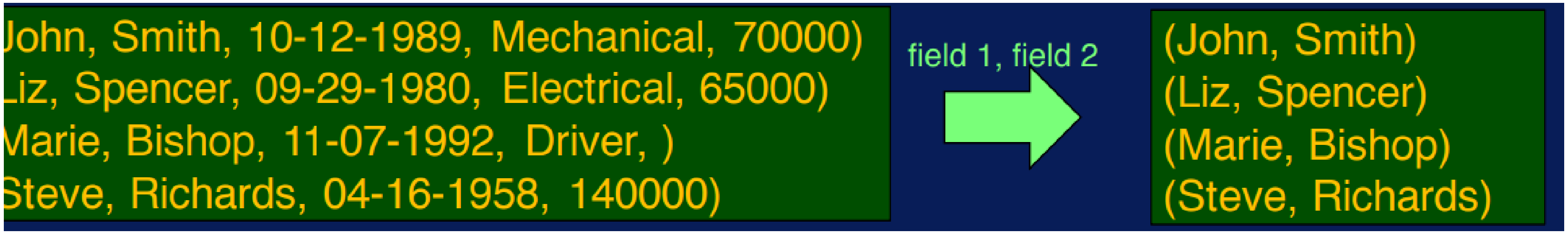
field 5 > 100000

(Steve, Richards, 04-16-1958, 140000)

01.-Modeling

Operations

- **Substructure extraction:** given a set of data with an
  - Extract a part of that structure with its elements



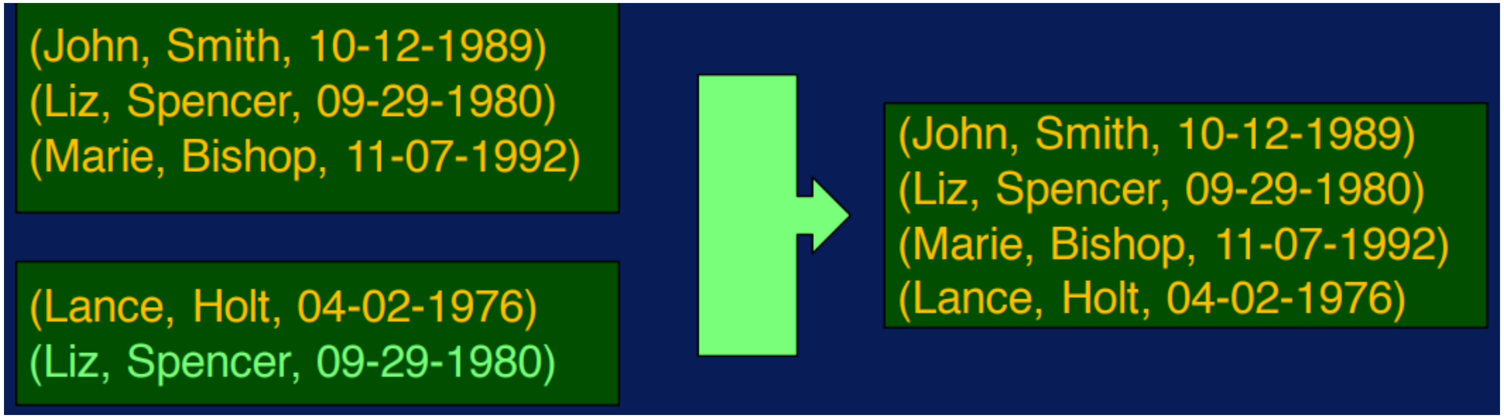
John, Smith, 10-12-1989, Mechanical, 70000)
Liz, Spencer, 09-29-1980, Electrical, 65000)
Marie, Bishop, 11-07-1992, Driver,)
Steve, Richards, 04-16-1958, 140000)

(John, Smith)
(Liz, Spencer)
(Marie, Bishop)
(Steve, Richards)

01.-Modeling

Operations

- **Union:** given two data sets
  - Create a new data set with elements from the two data sets
  - Erasing duplicates



(John, Smith, 10-12-1989)
(Liz, Spencer, 09-29-1980)
(Marie, Bishop, 11-07-1992)

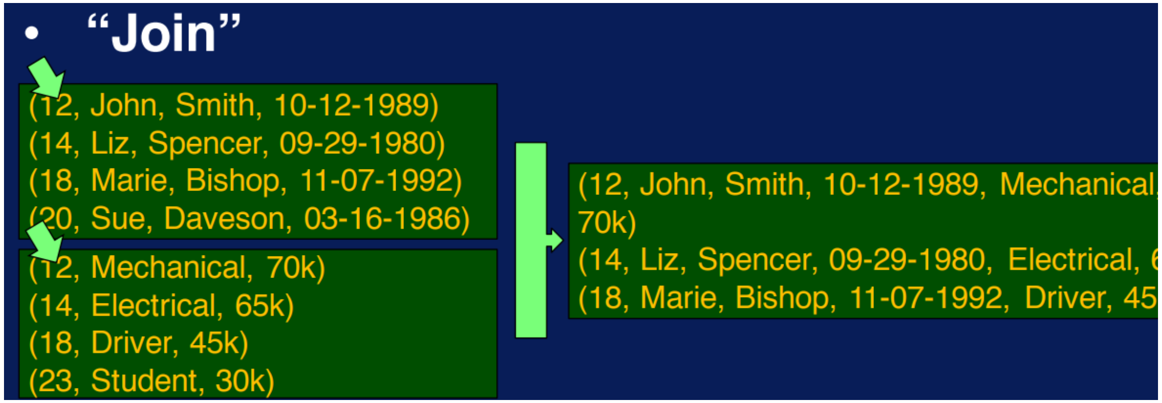
(Lance, Holt, 04-02-1976)
(Liz, Spencer, 09-29-1980)

(John, Smith, 10-12-1989)
(Liz, Spencer, 09-29-1980)
(Marie, Bishop, 11-07-1992)
(Lance, Holt, 04-02-1976)

01.-Modeling

Operations

- **Join:** given two data sets with complementary structure
  - Create a new group with elements of both data sets
  - Erasing the duplicates



- “Join”

(12, John, Smith, 10-12-1989)
(14, Liz, Spencer, 09-29-1980)
(18, Marie, Bishop, 11-07-1992)
(20, Sue, Daveson, 03-16-1986)

(12, Mechanical, 70k)
(14, Electrical, 65k)
(18, Driver, 45k)
(23, Student, 30k)

(12, John, Smith, 10-12-1989, Mechanical,
70k)
(14, Liz, Spencer, 09-29-1980, Electrical, 6
(18, Marie, Bishop, 11-07-1992, Driver, 45

01.-Modeling

Constrictions

- **Constrictions: there are logical propositions thta data must complain. example: each person has only one name**

- **Different models have different ways to express constrictions**

01.-Modeling

Constrictions

- **Constriction types**

Value constrictions:

Age can not be negative

Unicity:

each person has only one name

Cardinality:

each person has between 1 and 5 pone numbers

Type:

Surname is Alpha numeric

Domain:

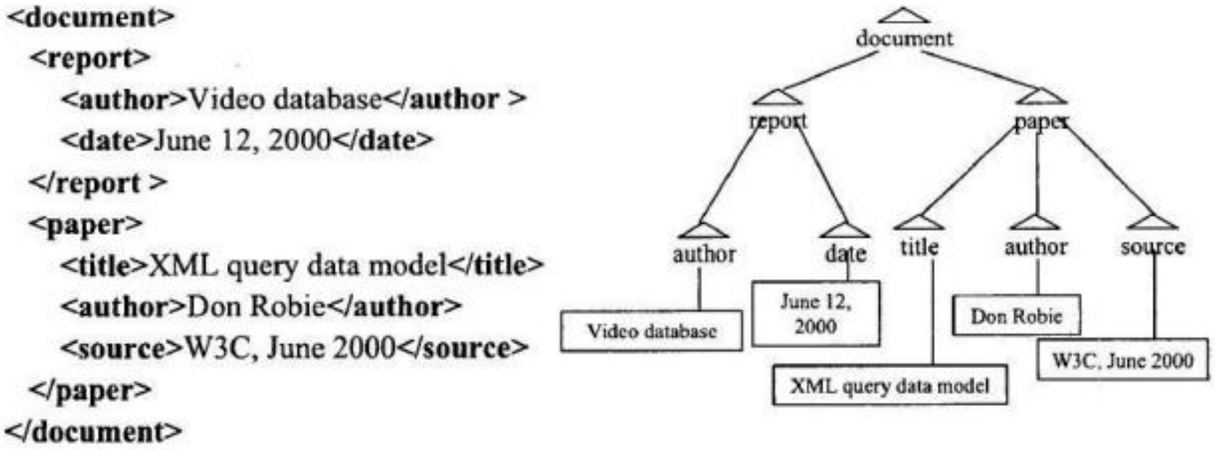
Days are from 1 to 31

Estructural: it sets restrictions to the structure instead to the data. Example: it must be a2x2 matrix

02.-Data models: structures

Semi structured

- **Semi structured data:**
  - **It is really a tree**
  - The queries actually involve navigating the tree until in common ancestor
  - **Example, query: June 12 2000 and author___document**

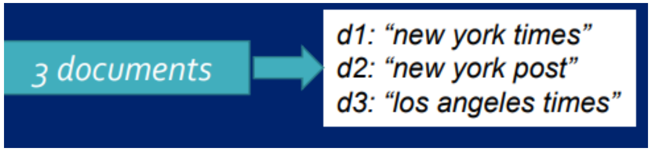


```
<document>
  <report>
    <author>Video database</author >
    <date>June 12, 2000</date>
  </report >
  <paper>
    <title>XML query data model</title>
    <author>Don Robie</author>
    <source>W3C, June 2000</source>
  </paper>
</document>
```

02.-Data models: structures

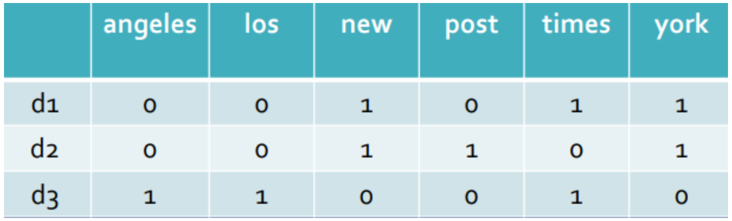
Unstructured

- **Vector Space Model: (we will review deeply in data analysis part next presentations)**
  - **Documents as vectors**

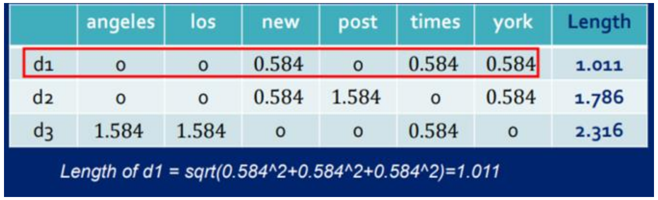


3 documents

d1: "new york times"
d2: "new york post"
d3: "los angeles times"



	angeles	los	new	post	times	york
d1	0	0	1	0	1	1
d2	0	0	1	1	0	1
d3	1	1	0	0	1	0



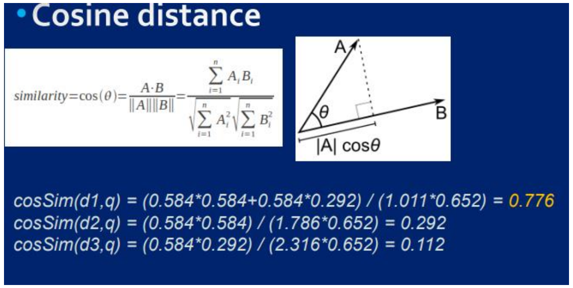
	angeles	los	new	post	times	york	Length
d1	0	0	0.584	0	0.584	0.584	1.011
d2	0	0	0.584	1.584	0	0.584	1.786
d3	1.584	1.584	0	0	0.584	0	2.316

Length of d1 = sqrt(0.584^2+0.584^2+0.584^2)=1.011

- **Queries as vectors for information retrieval**

query

q: new new york



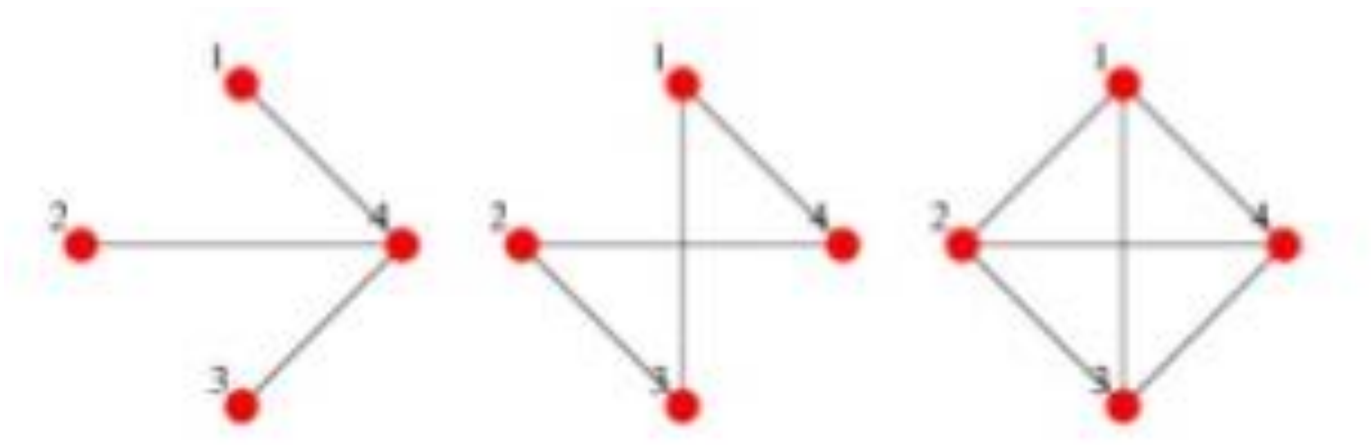
- Cosine distance

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

*cosSim(d1,q) = (0.584*0.584+0.584*0.292) / (1.011*0.652) = 0.776*
*cosSim(d2,q) = (0.584*0.584) / (1.786*0.652) = 0.292*
*cosSim(d3,q) = (0.584*0.292) / (2.316*0.652) = 0.112*

02.-Data models: structures

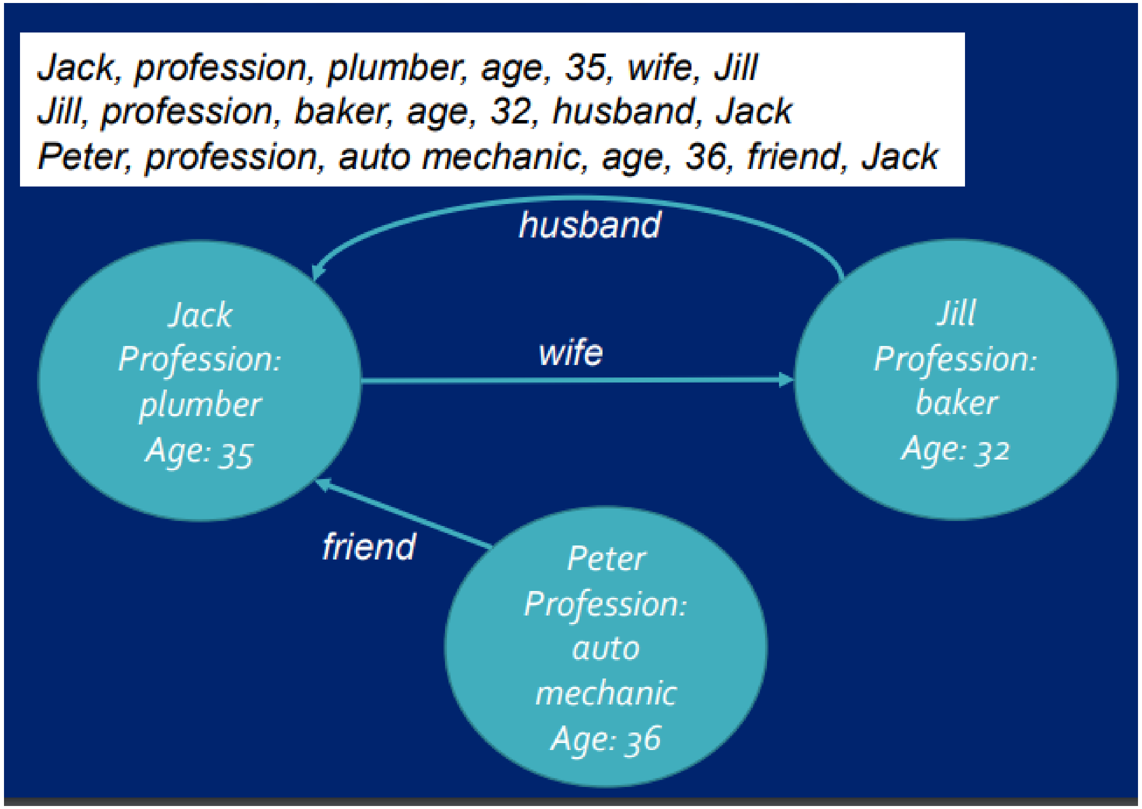
- **Other models: matrices**



$$\begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \quad \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix} \quad \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

03.-Data models vs data format

- **csv formal does not mean relational model : the format is not the model**



Jack, profession, plumber, age, 35, wife, Jill
Jill, profession, baker, age, 32, husband, Jack
Peter, profession, auto mechanic, age, 36, friend, Jack

BOOKs READINGS APPLIED TO THIS PRESENTATION AND INCLUDED FOR THE EXAMS

Book1: Big Data for Dummies

- Chapter 2:
  - Pg. 25-3
- Chapter 7: Operational data bases pg. 85-100 (some parts but the relevant content will be review in next presentations)