MANCHESTER
1824
The University of Manchester

COMP60411

Modelling Data On The Web

Uli Sattler & Bijan Parsia

Week 1 Introduction, Data Models, Tables, and SQL

Topic Overview

- What is a fundamental **data model**?
- Some key data models
  - **Flat**: flat files
  - **Table** based: relational
  - **Tree** based: XML and a bit of JSON
  - **Graph** based: RDF
- **Trade offs** (esp. representational) between them
  - Looking for the *pain points* and *sweet spots*

Course Goals:

Knowledge & Understanding

- This course unit aims to give you a
  - good understanding of core concepts of data modelling
  - some familiarity with formalisms, APIs, and languages
    - for modelling data on the web
    - design/representation issues that arise

Course Goals: Skills

- This course unit aims to give you the ability/skill to
  - compare different data modelling formalisms,
  - design or analyse a data management system,
    - does it make good use of the formalism's features?
    - does it fit its purpose?

Course Structure

- Lectures
    - Active learning
- Lab
    - Make sure you understand the coursework!
- Readings
    - All readings available online
    - Core: the "Learning" eBook series
        - Learning SQL (or here)
        - Learning XML (or here)
        - Learning SPARQL

Assessment

- Coursework (50%, ≈200 marks)
  - Each week, a mixture
    1. MCQ quizzes (≈10 marks)
    2. Short essays (≈5 marks)
    3. A modelling assignment (≈10 marks)
    4. A programming assignment (≈15 marks)
  - Precise mark breakdown varies
- Exam (50%)
  - Taken online
  - Very like 1 & 2

Materials & Blackboard

- All course materials are available online on the materials page
- We use **Blackboard** for
  - Coursework
  - Online forums
    - Use these!
  - Exam

Variant Circumstances

- Disability (Equality Act):
  - any condition which has a significant, adverse and long-term effect on a person’s ability to carry out normal day-to-day activities.
  - Disability Advisory and Support Service
    - Exam & Study support & more
    - Great, helpful people
- Counselling service
- SSO and Mitigating Circumstances process

...feel free to ask us: we're *happy* to advise!

Assistance & Help

- Early intervention is more effective
  - If you are having challenges of any sort
    - the sooner they are identified *and*
    - communicated to us
    - the more likely we can find a good resolution
- This is very true for mitigating circumstances
  - If something is interfering, document it!
  - Fill out the form *when* things are happening
  - There is a "too late" here!

...when in doubt, ask us and SSO for MitCircs

Expected Conduct

- We expect of you (and ourselves) to
  - be fair minded
  - treat each other well & with respect
  - avoid **academic malpractice**
  - take responsibility for course duties
  - be engaged, curious, and active
- If you have a problem or issue
  - please raise it with us
  - if that doesn't help, contact your programme director

Preliminaries

We all have to start somewhere

Data Management (1)

- Almost every program must do some data management
  - If only config files!
  - Many are *information heavy*
    - And must deal with that information over time
- Database Management Systems (DBMSs)
  - Separate (or separable) component
  - Specialised for variables purposed
    - Secondary storage, scaling, complexity, etc.

Data Management: Lifetime

- Some data is (typically) **transient** or **ephemeral**
  - Position of the cursor on the screen
- Some data is (typically) **persistent**
  - Bank records, addresses, health data, library entries
  - Cursor position can be!
    - (If you are recording the screen...)

*We're focused on data that leans toward **persistent***

Data Management: Structure

- Some data is (more or less) **informationally opaque**
    - E.G., images, video, text, audio
    - The information content isn't (immediately) available
        - You typically must do some *extraction*
    - Such is called **unstructured data**
- Some data is **informationally transparent**
    - The information content is programmtically explicit
    - Such is called **structured data**
        - We will later distinguish
            - **Structured**
            - **Semi-structured**

Out Of Scope

- There is lots of DM that's outside our scope
  1. Performance & Scaling: see COMP62421
  2. Concurrency
     - Thus *transactions*
       - (You should read up on ACIDity)
  3. Tuning, indeed most physical level stuff
  4. Cleansing
  5. Integration
     - Except for a tiny bit, around *merging*

These considerations *do* affect modelling!

Data And The Web

- The Web is a collaborative information structure
    - Largely decentralised
    - Immense
    - Growing rapidly
    - Changing rapidly
- The Web produces new data challenges
    - Scale of data
    - Kind of data
    - Shape of data
    - Use of data

Data On, From, Behind The Web

- **On** the Web
  - data.gov, data.gov.uk, …
- **From** the Web
  - Log files
- **Behind** the Web
  - Data(base) backed Websites
    - The filesystem is a kind of database
  - Content Management Systems
    - Wordpress
  - Sites as Database Front Ends
    - See Amazon

What Is A Data Model?

- Three Key Aspects
  1. Underlying Data Structure, "**Core DM**"
  2. Data Integrity
  3. Data Manipulation
  4. (Plus a fourth!) Data Sharing
     - More important on the Web *

”Data Model” Is Ambiguous

- Data model is used to refer to...

  1. a complete data representation and manipulation approach (we do this!)
  2. just the **core data model**
  3. a particular data representation for a domain or application, also called the **domain model**

     - "Does your calendar data model include leap years?"

  Generally, you can tell from context, (2) is rare.

Kinds Of Data

- Data can lend itself to different **shapes**
  - Array-like
  - Tree-like
  - Graph-like
  - Document-like
- Data can have different **volumes**
  - Small to "big" data
- Data can have different **velocities**
  - Static/offline to streaming
- Data can have different **use patterns**
  - Many readers/few writers or the reverse or other!

Polyglot Persistence

> *...we are gearing up for a shift to polyglot persistence — **where any decent sized enterprise will have a variety of different data storage technologies for different kinds of data.** There will still be large amounts of it managed in relational stores, but increasingly we'll be **first asking how we want to manipulate the data** and only then figuring out what technology is the best bet for it.*
>
> — *Martin Fowler*

Polyglot Persistence (2)

> *This polyglot [e]ffect will be apparent even within a single application. A complex enterprise application uses different kinds of data, and already usually integrates information from different sources.* ***Increasingly we'll see such applications manage their own data using different technologies depending on how the data is used.***
>
> — *Martin Fowler*

Poly -Glot/-System Persistence

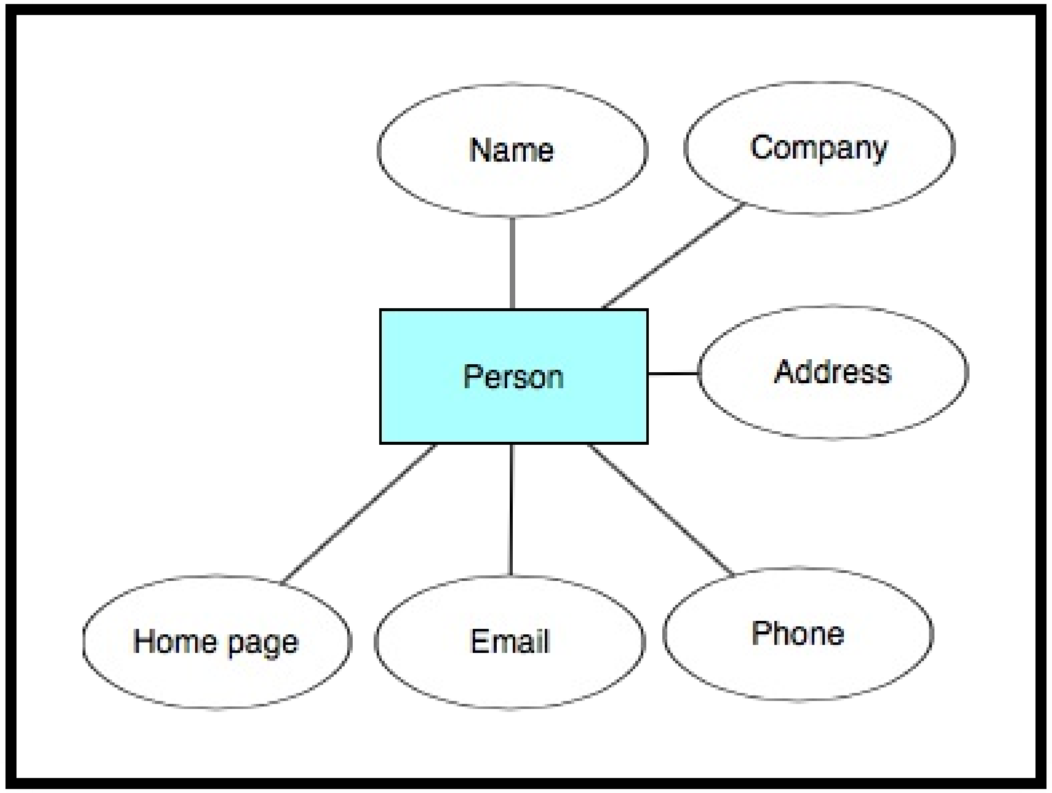
- Even with a single core data model
  - Multiple systems with different characteristics
  - Multiple, overlapping, domain models
  - Multiple, overlapping owners, versions, variants

This is particularly true in on the Web!

"Flat Files" – A Simple Model

A Sample Domain

- We start with a classic example: The Address Book
  - People and information about them
  - Names and contact information
- We can do a first cut as a diagram

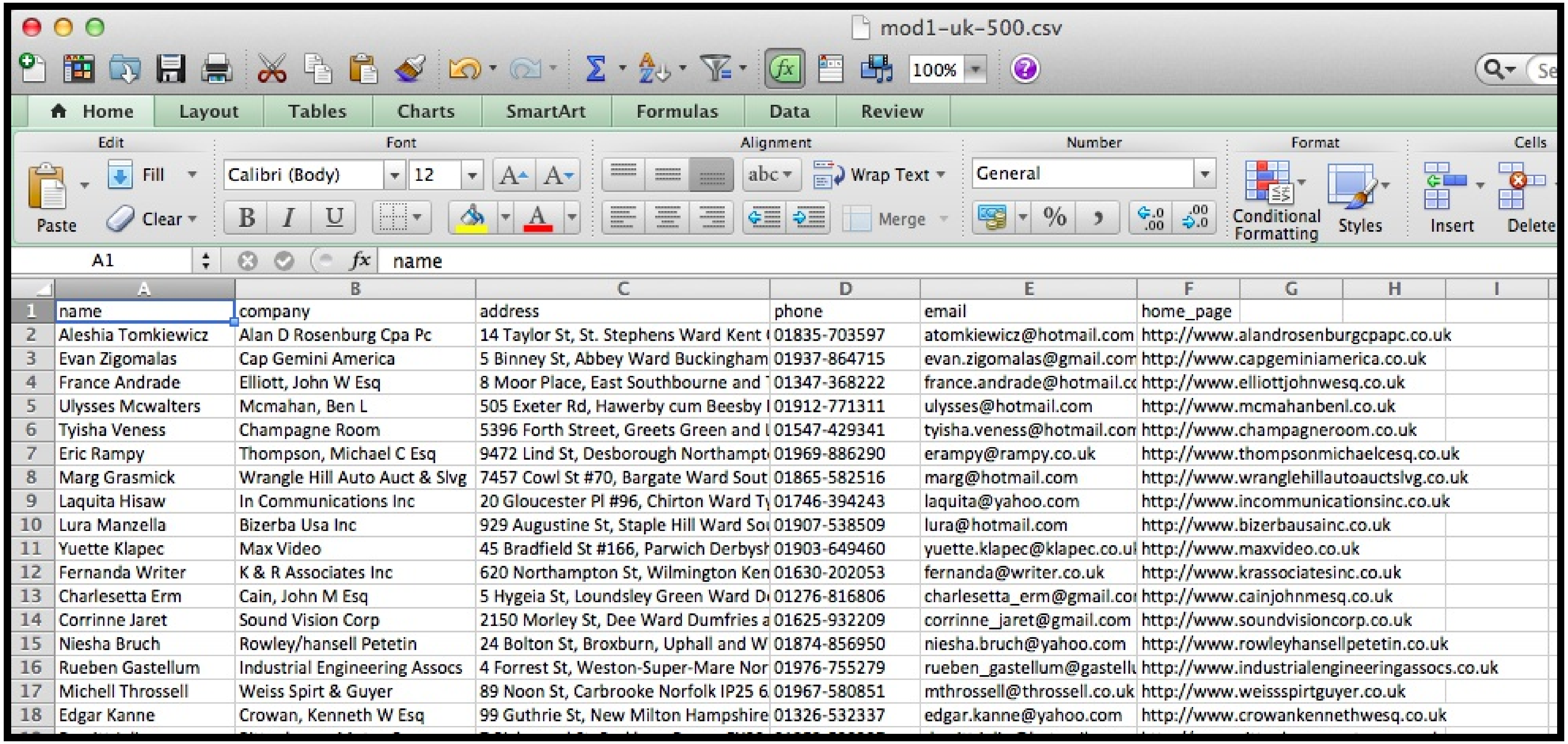


For Example

- Bijan!
  - Name: Bijan Parsia
  - Company: University of Manchester
  - Email: bijan.parsia@manchester.ac.uk
  - ...
- Uli!
  - Name: Uli Sattler
  - Company: University of Manchester
  - Email: uli.sattler@manchester.ac.uk

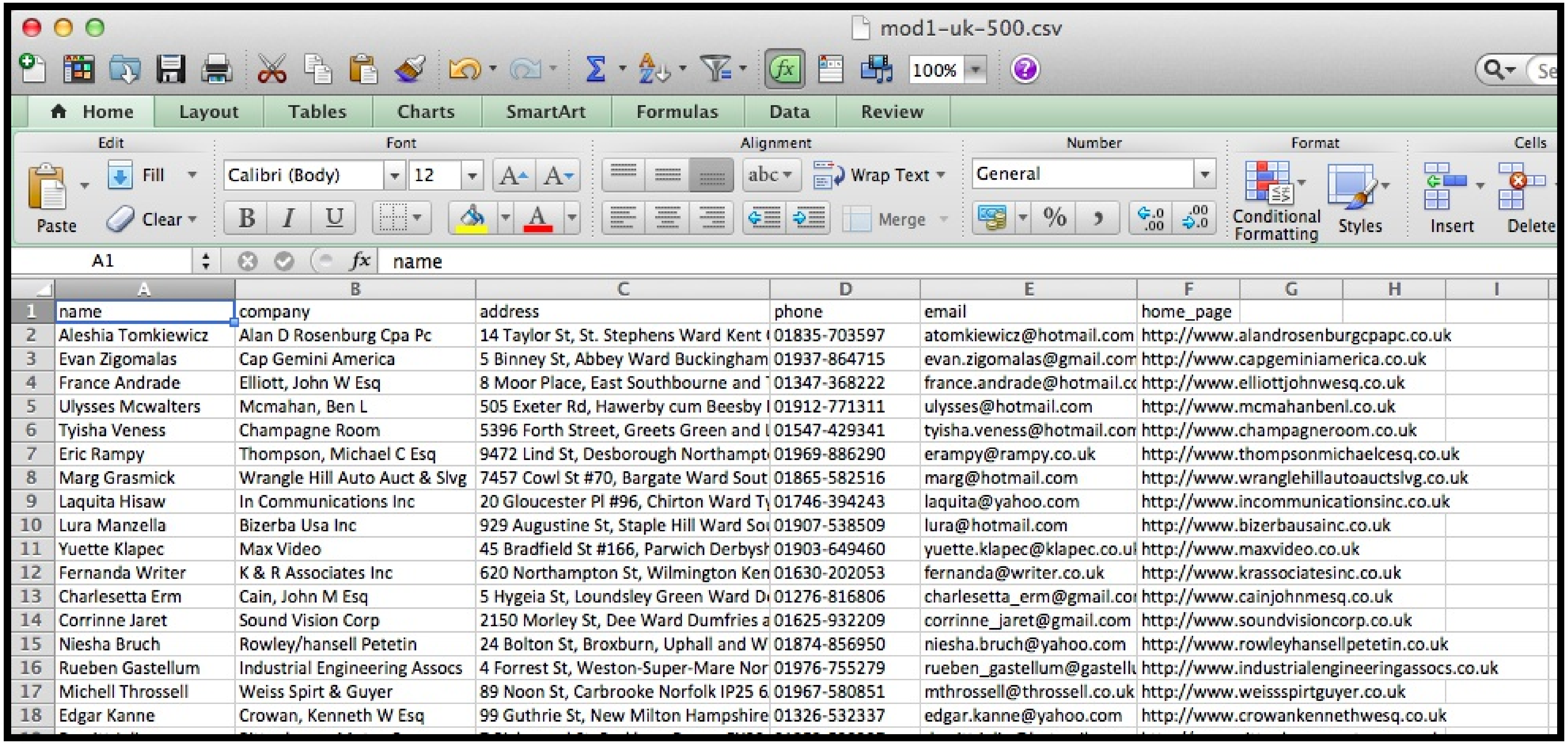
Storing!

- Slides are not a good storage place for data
- We have an array like structure so…
    - How about a spreadsheet!
        - 1 entity/record/person per **row**
        - Each field/attribute is a **column**
- We have software that works well with this!



	A	B	C	D	E	F	G	H	I
1	name	company	address	phone	email	home_page			
2	Aleshia Tomkiewicz	Alan D Rosenburg Cpa Pc	14 Taylor St, St. Stephens Ward Kent (	01835-703597	atomkiewicz@hotmail.com	http://www.alandrosenburgcpapc.co.uk			
3	Evan Zigomalas	Cap Gemini America	5 Binney St, Abbey Ward Buckingham	01937-864715	evan.zigomalas@gmail.com	http://www.capgeminiamerica.co.uk			
4	France Andrade	Elliott, John W Esq	8 Moor Place, East Southbourne and T	01347-368222	france.andrade@hotmail.co	http://www.elliottjohnwesq.co.uk			
5	Ulysses Mcwalters	Mcmahan, Ben L	505 Exeter Rd, Hawerby cum Beesby I	01912-771311	ulysses@hotmail.com	http://www.mcmahanbenl.co.uk			
6	Tyisha Veness	Champagne Room	5396 Forth Street, Greets Green and l	01547-429341	tyisha.veness@hotmail.com	http://www.champagneroom.co.uk			
7	Eric Rampy	Thompson, Michael C Esq	9472 Lind St, Desborough Northampt	01969-886290	erampy@rampy.co.uk	http://www.thompsonmichaelcesq.co.uk			
8	Marg Grasmick	Wrangle Hill Auto Auct & Slvg	7457 Cowl St #70, Bargate Ward Sout	01865-582516	marg@hotmail.com	http://www.wranglehillautoauctslvg.co.uk			
9	Laquita Hisaw	In Communications Inc	20 Gloucester Pl #96, Chirton Ward Ty	01746-394243	laquita@yahoo.com	http://www.incommunicationsinc.co.uk			
10	Lura Manzella	Bizerba Usa Inc	929 Augustine St, Staple Hill Ward Sou	01907-538509	lura@hotmail.com	http://www.bizerbausainc.co.uk			
11	Yuette Klapec	Max Video	45 Bradfield St #166, Parwich Derbysh	01903-649460	yuette.klapec@klapec.co.uk	http://www.maxvideo.co.uk			
12	Fernanda Writer	K & R Associates Inc	620 Northampton St, Wilmington Ken	01630-202053	fernanda@writer.co.uk	http://www.krassociatesinc.co.uk			
13	Charlesetta Erm	Cain, John M Esq	5 Hygeia St, Loundsley Green Ward D	01276-816806	charlesetta_erm@gmail.co	http://www.cainjohnmesq.co.uk			
14	Corrinne Jaret	Sound Vision Corp	2150 Morley St, Dee Ward Dumfries a	01625-932209	corrinne_jaret@gmail.com	http://www.soundvisioncorp.co.uk			
15	Niesha Bruch	Rowley/hansell Petetin	24 Bolton St, Broxburn, Uphall and W	01874-856950	niesha.bruch@yahoo.com	http://www.rowleyhansellpetetin.co.uk			
16	Rueben Gastellum	Industrial Engineering Assocs	4 Forrest St, Weston-Super-Mare Nor	01976-755279	rueben_gastellum@gastellu	http://www.industrialengineeringassocs.co.uk			
17	Michell Throssell	Weiss Spirt & Guyer	89 Noon St, Carbrooke Norfolk IP25 6	01967-580851	mthrossell@throssell.co.uk	http://www.weissspirtguyer.co.uk			
18	Edgar Kanne	Crowan, Kenneth W Esq	99 Guthrie St, New Milton Hampshire	01326-532337	edgar.kanne@yahoo.com	http://www.crowankennethwesq.co.uk			

Interacting With The Data



	A	B	C	D	E	F	G	H	I
1	name	company	address	phone	email	home_page			
2	Aleshia Tomkiewicz	Alan D Rosenburg Cpa Pc	14 Taylor St, St. Stephens Ward Kent (	01835-703597	atomkiewicz@hotmail.com	http://www.alandrosenburgcpapc.co.uk			
3	Evan Zigomalas	Cap Gemini America	5 Binney St, Abbey Ward Buckingham	01937-864715	evan.zigomalas@gmail.com	http://www.capgeminiamerica.co.uk			
4	France Andrade	Elliott, John W Esq	8 Moor Place, East Southbourne and T	01347-368222	france.andrade@hotmail.cc	http://www.elliottjohnwesq.co.uk			
5	Ulysses Mcwalters	Mcmahan, Ben L	505 Exeter Rd, Hawerby cum Beesby I	01912-771311	ulysses@hotmail.com	http://www.mcmahanbenl.co.uk			
6	Tyisha Veness	Champagne Room	5396 Forth Street, Greets Green and L	01547-429341	tyisha.veness@hotmail.com	http://www.champagneroom.co.uk			
7	Eric Rampy	Thompson, Michael C Esq	9472 Lind St, Desborough Northampt	01969-886290	erampy@rampy.co.uk	http://www.thompsonmichaelcesq.co.uk			
8	Marg Grasmick	Wrangle Hill Auto Auct & Slvg	7457 Cowl St #70, Bargate Ward Sout	01865-582516	marg@hotmail.com	http://www.wranglehillautoauctslvg.co.uk			
9	Laquita Hisaw	In Communications Inc	20 Gloucester Pl #96, Chirton Ward Ty	01746-394243	laquita@yahoo.com	http://www.incommunicationsinc.co.uk			
10	Lura Manzella	Bizerba Usa Inc	929 Augustine St, Staple Hill Ward Sou	01907-538509	lura@hotmail.com	http://www.bizerbausainc.co.uk			
11	Yuette Klapec	Max Video	45 Bradfield St #166, Parwich Derbysh	01903-649460	yuette.klapec@klapec.co.ul	http://www.maxvideo.co.uk			
12	Fernanda Writer	K & R Associates Inc	620 Northampton St, Wilmington Ken	01630-202053	fernanda@writer.co.uk	http://www.krassociatesinc.co.uk			
13	Charlesetta Erm	Cain, John M Esq	5 Hygeia St, Loundsley Green Ward De	01276-816806	charlesetta_erm@gmail.co	http://www.cainjohnmesq.co.uk			
14	Corrinne Jaret	Sound Vision Corp	2150 Morley St, Dee Ward Dumfries a	01625-932209	corrinne_jaret@gmail.com	http://www.soundvisioncorp.co.uk			
15	Niesha Bruch	Rowley/hansell Petetin	24 Bolton St, Broxburn, Uphall and W	01874-856950	niesha.bruch@yahoo.com	http://www.rowleyhansellpetetin.co.uk			
16	Rueben Gastellum	Industrial Engineering Assocs	4 Forrest St, Weston-Super-Mare Nor	01976-755279	rueben_gastellum@gastellu	http://www.industrialengineeringassocs.co.uk			
17	Michell Throssell	Weiss Spirt & Guyer	89 Noon St, Carbrooke Norfolk IP25 6	01967-580851	mthrossell@throssell.co.uk	http://www.weissspirtguyer.co.uk			
18	Edgar Kanne	Crowan, Kenneth W Esq	99 Guthrie St, New Milton Hampshire	01326-532337	edgar.kanne@yahoo.com	http://www.crowankennethwesq.co.uk			

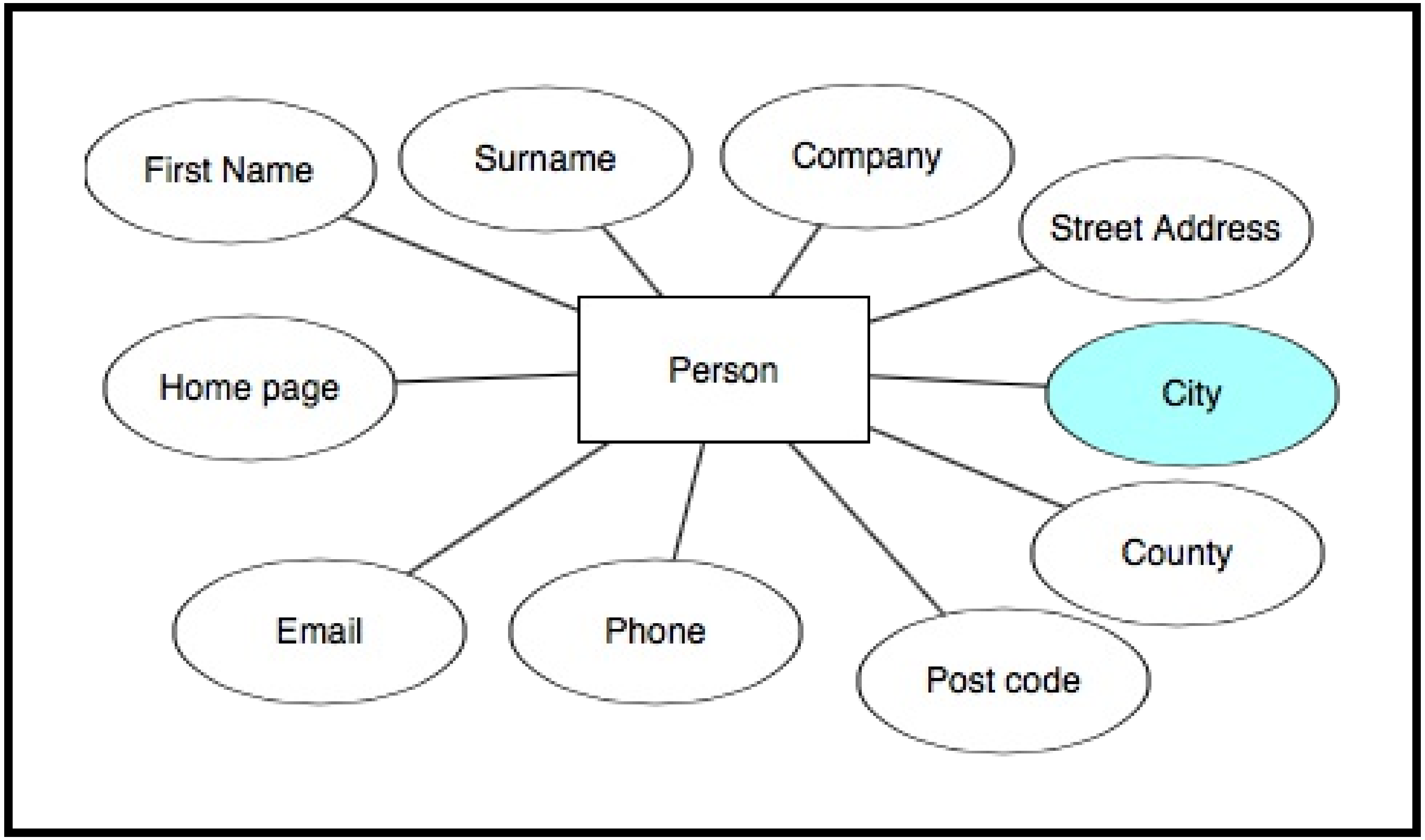
To the demo!

Pain Points

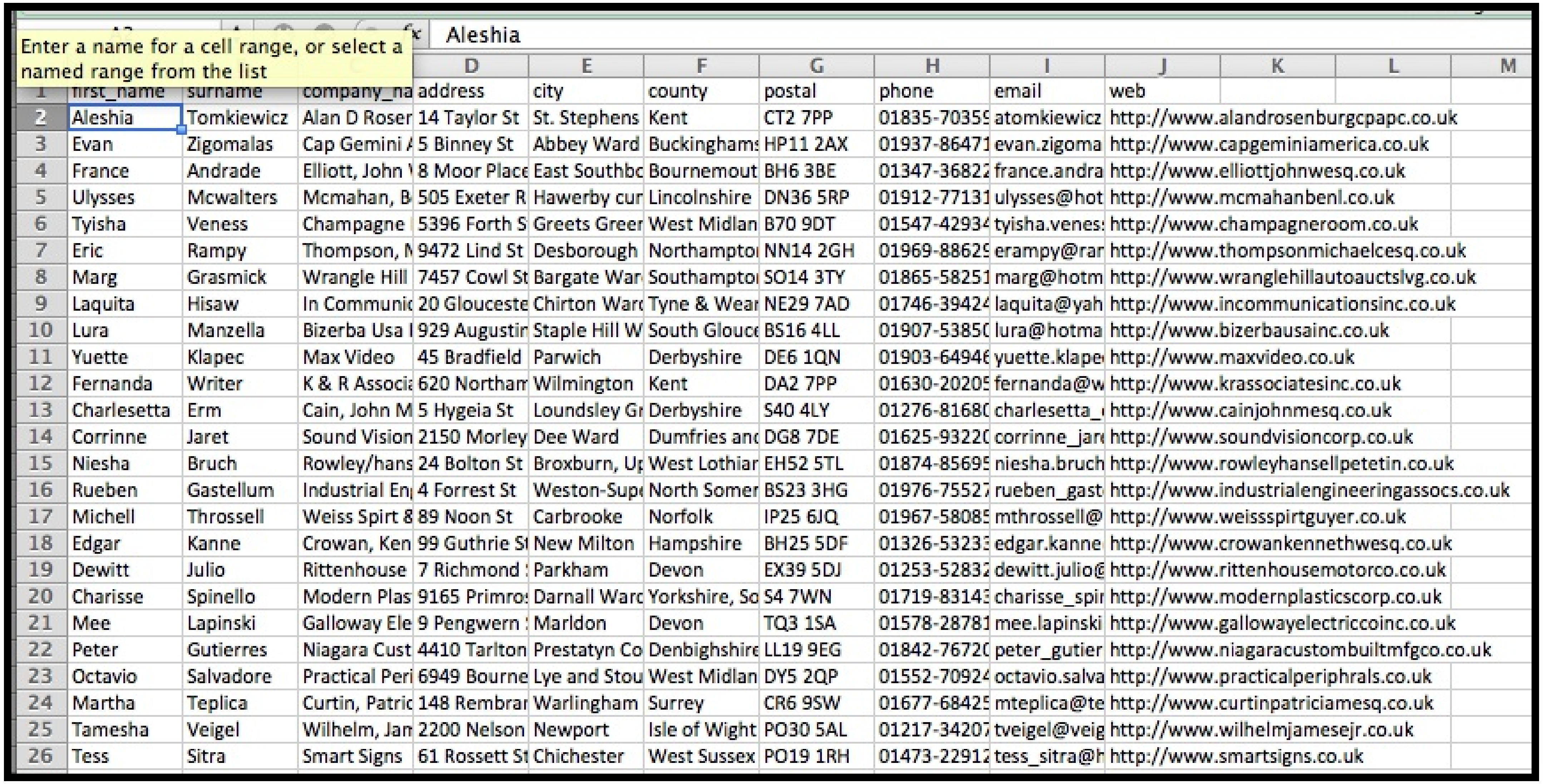
- Around "name"
  - Sorting
    - Sorting is on **columns**
      - Can't sort by last name
  - Filtering
    - *Can* filter by **names** beginning with Z
    - Cannot by **surname's** beginning with Z
- Around "address"
  - Can't sort or filter by **postcode**
  - Can't sort or filter by **city**
  - Can't sort or filter by **county**

These are *problems with our model*

Fixing The Domain Model



Interacting!



Enter a name for a cell range, or select a named range from the list

Aleshia

				D	E	F	G	H	I	J	K	L	M
1	first_name	surname	company_na	address	city	county	postal	phone	email	web			
2	Aleshia	Tomkiewicz	Alan D Roser	14 Taylor St	St. Stephens	Kent	CT2 7PP	01835-70359	atomkiewicz	http://www.alandrosenburgcpapc.co.uk			
3	Evan	Zigomalas	Cap Gemini A	5 Binney St	Abbey Ward	Buckinghams	HP11 2AX	01937-86471	evan.zigoma	http://www.capgeminiamerica.co.uk			
4	France	Andrade	Elliott, John W	8 Moor Place	East Southbo	Bournemout	BH6 3BE	01347-36822	france.andra	http://www.elliottjohnwesq.co.uk			
5	Ulysses	Mcwalters	Mcmahan, B	505 Exeter R	Hawerby cur	Lincolnshire	DN36 5RP	01912-77131	ulysses@hot	http://www.mcmahanbenl.co.uk			
6	Tyisha	Veness	Champagne	5396 Forth S	Greets Greer	West Midlan	B70 9DT	01547-42934	tyisha.venes	http://www.champagneroom.co.uk			
7	Eric	Rampy	Thompson, M	9472 Lind St	Desborough	Northamptor	NN14 2GH	01969-88629	erampy@rar	http://www.thompsonmichaelcesq.co.uk			
8	Marg	Grasmick	Wrangle Hill	7457 Cowl St	Bargate War	Southamptor	SO14 3TY	01865-58251	marg@hotm	http://www.wranglehillautoauctslvg.co.uk			
9	Laquita	Hisaw	In Communic	20 Glouceste	Chirton Ward	Tyne & Wear	NE29 7AD	01746-39424	laquita@yah	http://www.incommunicationsinc.co.uk			
10	Lura	Manzella	Bizerba Usa I	929 Augustin	Staple Hill W	South Glouce	BS16 4LL	01907-53850	lura@hotma	http://www.bizerbausainc.co.uk			
11	Yuette	Klapec	Max Video	45 Bradfield	Parwich	Derbyshire	DE6 1QN	01903-64946	yuette.klape	http://www.maxvideo.co.uk			
12	Fernanda	Writer	K & R Associa	620 Northam	Wilmington	Kent	DA2 7PP	01630-20205	fernanda@w	http://www.krassociatesinc.co.uk			
13	Charlesetta	Erm	Cain, John M	5 Hygeia St	Loundsley Gr	Derbyshire	S40 4LY	01276-81680	charlesetta_	http://www.cainjohnmesq.co.uk			
14	Corrinne	Jaret	Sound Vision	2150 Morley	Dee Ward	Dumfries and	DG8 7DE	01625-93220	corrinne_jar	http://www.soundvisioncorp.co.uk			
15	Niesha	Bruch	Rowley/hans	24 Bolton St	Broxburn, Up	West Lothiar	EH52 5TL	01874-85695	niesha.bruch	http://www.rowleyhansellpetetin.co.uk			
16	Rueben	Gastellum	Industrial En	4 Forrest St	Weston-Supe	North Somer	BS23 3HG	01976-75527	rueben_gast	http://www.industrialengineeringassocs.co.uk			
17	Michell	Throssell	Weiss Spirt 8	89 Noon St	Carbrooke	Norfolk	IP25 6JQ	01967-58085	mthrossell@	http://www.weissspirtguyer.co.uk			
18	Edgar	Kanne	Crowan, Ken	99 Guthrie St	New Milton	Hampshire	BH25 5DF	01326-53233	edgar.kanne	http://www.crowankennethwesq.co.uk			
19	Dewitt	Julio	Rittenhouse	7 Richmond S	Parkham	Devon	EX39 5DJ	01253-52832	dewitt.julio@	http://www.rittenhousemotorco.co.uk			
20	Charisse	Spinello	Modern Plas	9165 Primros	Darnall Ward	Yorkshire, So	S4 7WN	01719-83143	charisse_spir	http://www.modernplasticscorp.co.uk			
21	Mee	Lapinski	Galloway Ele	9 Pengwern S	Marldon	Devon	TQ3 1SA	01578-28781	mee.lapinski	http://www.gallowayelectriccoinc.co.uk			
22	Peter	Gutierres	Niagara Cust	4410 Tarlton	Prestatyn Co	Denbighshire	LL19 9EG	01842-76720	peter_gutier	http://www.niagaracustombuiltmfgco.co.uk			
23	Octavio	Salvadore	Practical Peri	6949 Bourne	Lye and Stou	West Midlan	DY5 2QP	01552-70924	octavio.salva	http://www.practicalperiphrals.co.uk			
24	Martha	Teplica	Curtin, Patric	148 Rembrar	Warlingham	Surrey	CR6 9SW	01677-68425	mteplica@te	http://www.curtinpatriciamesq.co.uk			
25	Tamesha	Veigel	Wilhelm, Jam	2200 Nelson	Newport	Isle of Wight	PO30 5AL	01217-34207	tveigel@veig	http://www.wilhelmjamesejr.co.uk			
26	Tess	Sitra	Smart Signs	61 Rossett St	Chichester	West Sussex	PO19 1RH	01473-22912	tess_sitra@h	http://www.smartsigns.co.uk			

Demo encore!

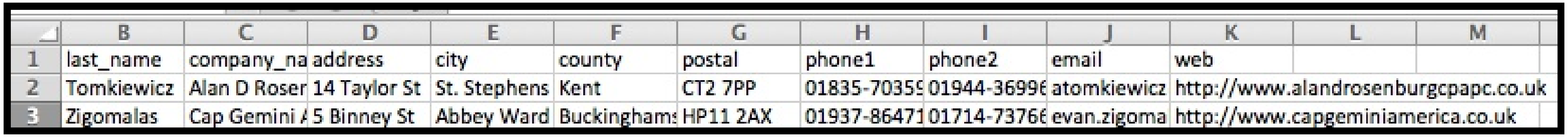
New Pain Points

- Variable numbers of the "same" attribute
    - Phone number
    - Email address
    - Web page
    - Inserting columns is painful
        - Lots of partial columns
        - Sheer number sucks
- Companies have addresses!
    - More than one!
    - And phone numbers, etc.

More *problems with our model*

Bad Model

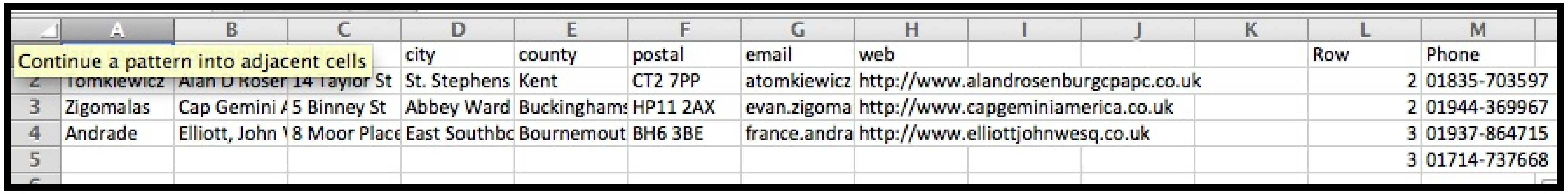
- Bad



	B	C	D	E	F	G	H	I	J	K	L	M
1	last_name	company_na	address	city	county	postal	phone1	phone2	email	web		
2	Tomkiewicz	Alan D Roser	14 Taylor St	St. Stephens	Kent	CT2 7PP	01835-70359	01944-36996	atomkiewicz	http://www.alandrosenburgcpapc.co.uk		
3	Zigomalas	Cap Gemini A	5 Binney St	Abbey Ward	Buckinghams	HP11 2AX	01937-86471	01714-73766	evan.zigoma	http://www.capgeminiamerica.co.uk		

Fixing The Model 2

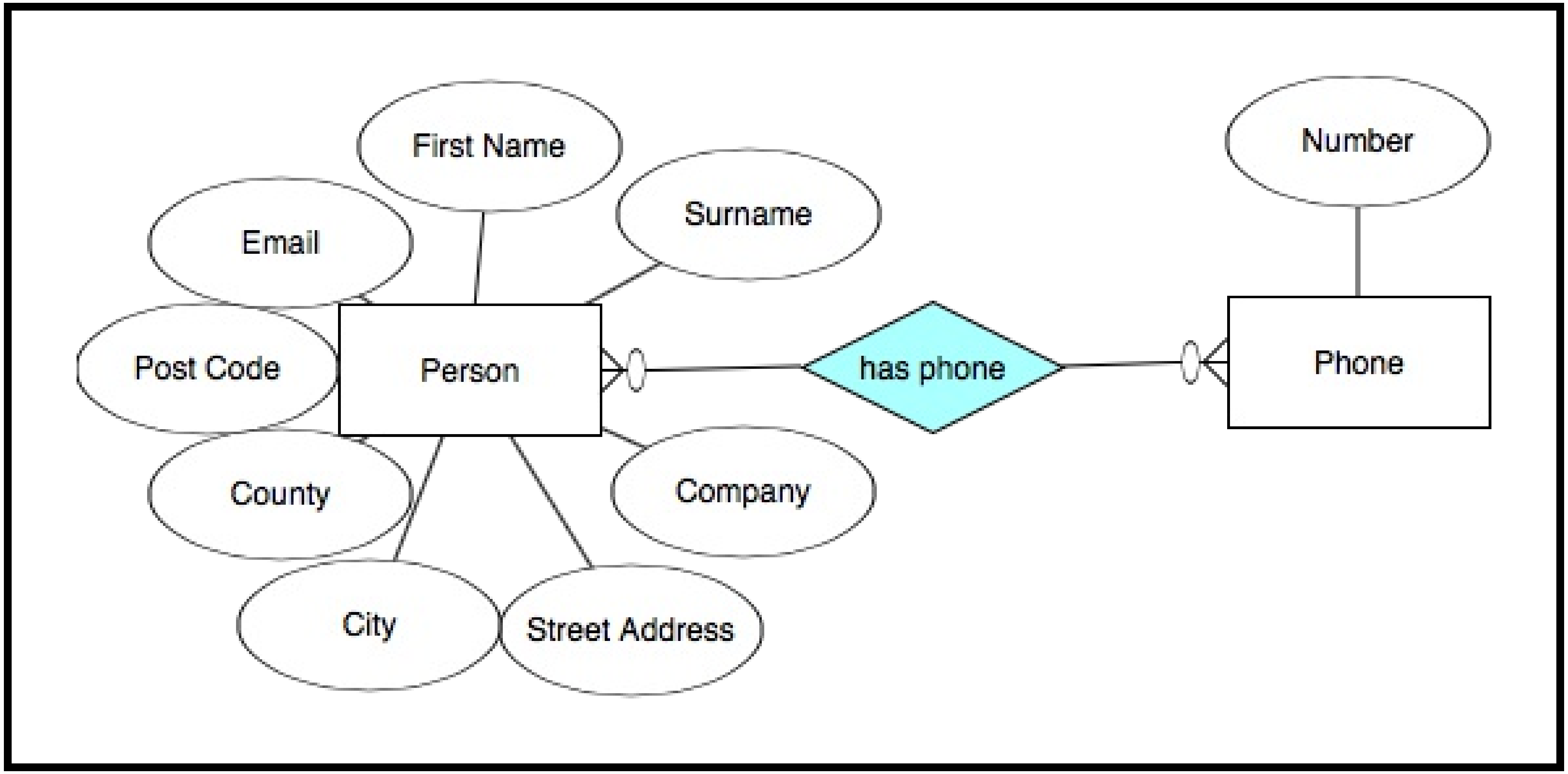
- We want adding a (similar) column to be easy!
  - Easy as adding a row!
  - Make a *new table* just for phone numbers
  - Index numbers with person rows



Continue a pattern into adjacent cells

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	last_name	company_na	address	city	county	postal	email	web				Row	Phone
2	Tomkiewicz	Alan D Roser	14 Taylor St	St. Stephens	Kent	CT2 7PP	atomkiewicz	http://www.alandrosenburgcpapc.co.uk				2	01835-703597
3	Zigomalas	Cap Gemini A	5 Binney St	Abbey Ward	Buckinghams	HP11 2AX	evan.zigoma	http://www.capgeminiamerica.co.uk				2	01944-369967
4	Andrade	Elliott, John \	8 Moor Place	East Southbo	Bournemout	BH6 3BE	france.andra	http://www.elliottjohnwesq.co.uk				3	01937-864715
5												3	01714-737668

Fixing The Model Again



Pain Points

- Sorting **destroys** the relationship
  - We used row numbers to connect
  - Sorting changes the row number!
- Hard to see the record
- No longer a simple flat file
  - CSV format makes assumptions

These are (mostly) **implementation** problems!

When A Domain Model Fails

- Failure must be analysed!
  - Did we
    - get the domain wrong?
    - fit it wrong into our core DM?
    - pick the wrong core CM to model it in?
  - Is it
    - unworkable?
    - workable but requires a lot of application code?
    - reasonable with some workarounds?

How much **technical debt** are we piling up?

What's the **cost of switching**?

Broken Core Data Model

- If you are
  - always "fighting" the system
  - use lots of application code to hack things
  - live in an error rich environment
  - have increasing amounts of workaround support in your data

Your data model might not be a good fit for your domain and application!

The Rest Of The DBMS

- Even if your core data model isn't a good fit
    - You might
        - be stuck with the system
            - You paid good money for that Oracle database!
        - need features of the implementation
            - is there an XML database with transactions?
            - what's the support contract?
        - be stuck with the model
            - critical legacy apps

Just because the **model** is broken doesn't mean that the **system** is

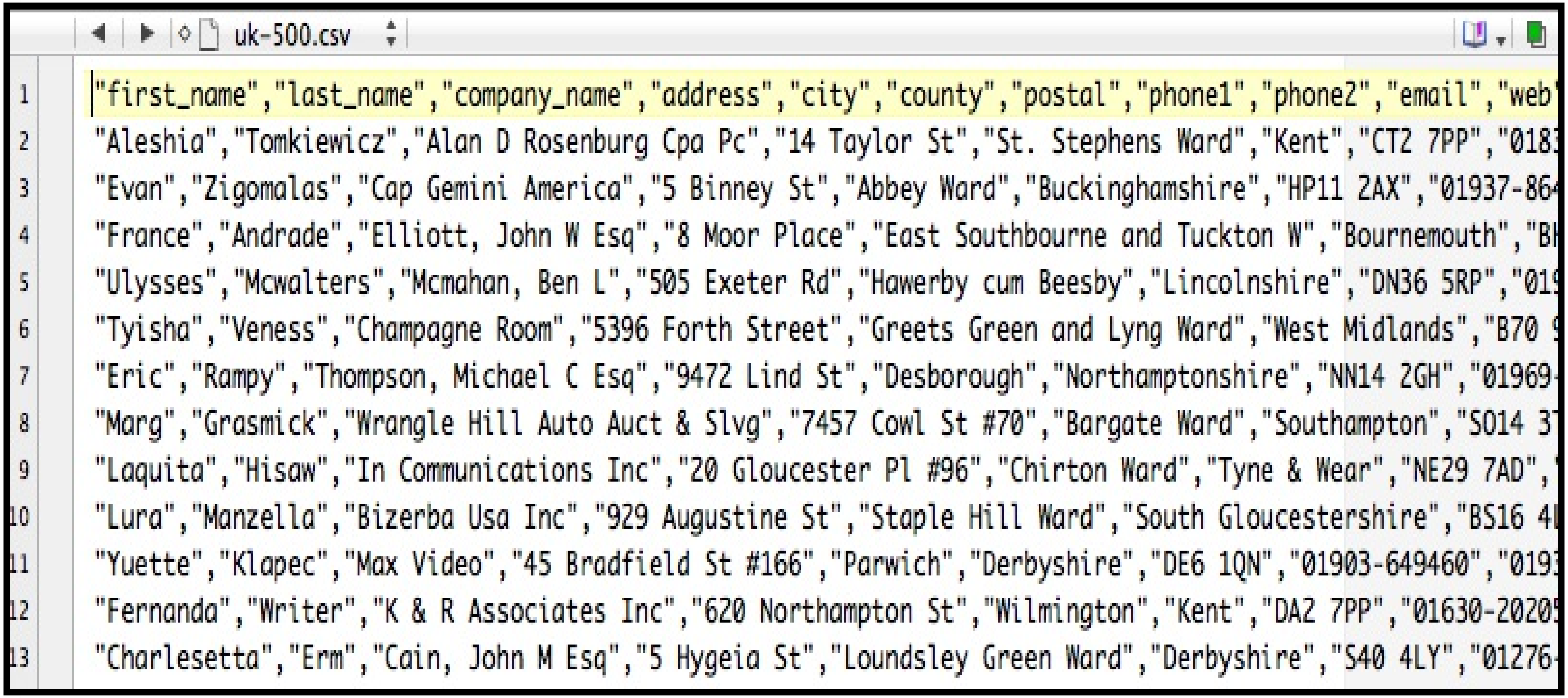
Or is **broken enough** to justify a switch

Flat File Programming

Sharing Our Databases

- Spreadsheets?
  - **Propriatoryish** (Excel, Google Doc, OpenOffice)
- Lingua franca: **CSV**
  - Comma (or Tab) Delimited Values
  - *Exactly* the (pure) flat file model
  - Format:
    - Text file
    - 1 record per line
    - First line can be special (column names)
    - Each column separated by a ","
      - We may need to quote cells (with commas)

CSV Example



```
"first_name","last_name","company_name","address","city","county","postal","phone1","phone2","email","web
"Aleshia","Tomkiewicz","Alan D Rosenburg Cpa Pc","14 Taylor St","St. Stephens Ward","Kent","CT2 7PP","018
"Evan","Zigomalas","Cap Gemini America","5 Binney St","Abbey Ward","Buckinghamshire","HP11 2AX","01937-86
"France","Andrade","Elliott, John W Esq","8 Moor Place","East Southbourne and Tuckton W","Bournemouth","B
"Ulysses","Mcwalters","Mcmahan, Ben L","505 Exeter Rd","Hawerby cum Beesby","Lincolnshire","DN36 5RP","01
"Tyisha","Veness","Champagne Room","5396 Forth Street","Greets Green and Lyng Ward","West Midlands","B70
"Eric","Rampy","Thompson, Michael C Esq","9472 Lind St","Desborough","Northamptonshire","NN14 2GH","01969
"Marg","Grasmick","Wrangle Hill Auto Auct & Slvg","7457 Cowl St #70","Bargate Ward","Southampton","SO14 3
"Laquita","Hisaw","In Communications Inc","20 Gloucester Pl #96","Chirton Ward","Tyne & Wear","NE29 7AD",
"Lura","Manzella","Bizerba Usa Inc","929 Augustine St","Staple Hill Ward","South Gloucestershire","BS16 4
"Yuette","Klapec","Max Video","45 Bradfield St #166","Parwich","Derbyshire","DE6 1QN","01903-649460","019
"Fernanda","Writer","K & R Associates Inc","620 Northampton St","Wilmington","Kent","DA2 7PP","01630-2020
"Charlesetta","Erm","Cain, John M Esq","5 Hygeia St","Loundsley Green Ward","Derbyshire","S40 4LY","01276-
```

Programmatic Manipulation

- If we store our databases as CSV
    - We can load and parse them into structures
    - Manipulate our data from programs
        - *Our* programs, instead of Excel
- E.g., using the Apache Commons CSV

```
Reader in = new FileReader("path/to/file.csv");
Iterable<CSVRecord> records = CSVFormat.EXCEL.parse(in);
for (CSVRecord record : records) {
    String surname = record.get("surname");
    String firstName = record.get("first_name");
    ...
}
```

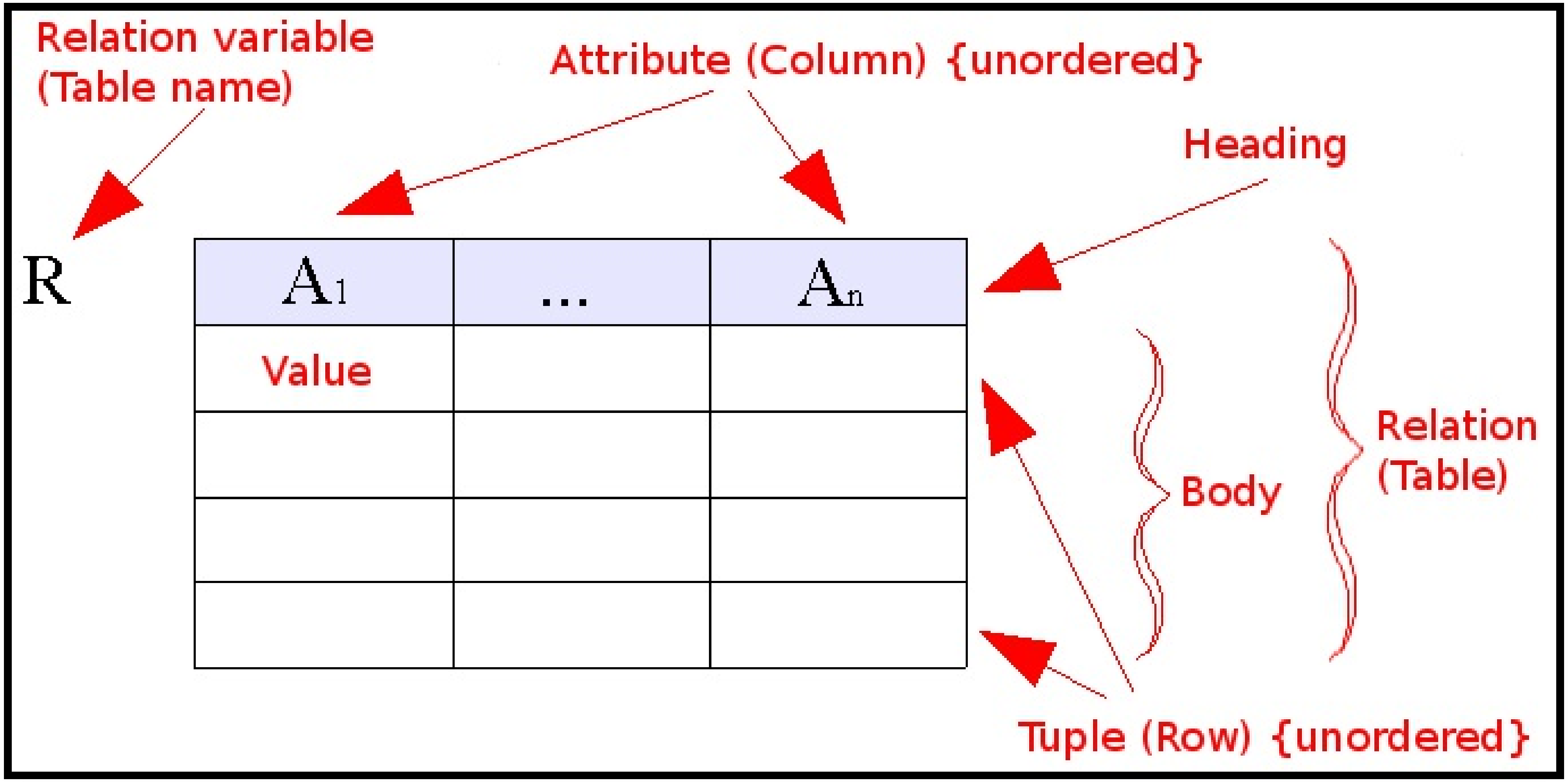
Solving Problems

- This solves some problems!
    - Inserting/removing columns a "small matter of programming"
        - Or we could use multiple arrays with pointers
    - We can split/combine fields at will
        - Well, with a bit of programming
    - We can control sorting well enough
        - Use pointers to connect
- Lots of work!

Against Bespoke Programming

- This is all at the wrong level
  - Flat files and flat file++ are ubiquitous
  - We shouldn't be coding complex functions
    - Over and over again!
- Even if we can program our way around problems
  - Doesn't eliminate the problems
  - Some solutions (pointers) effectively change the model!

A Relational Model



Relation variable (Table name)

Attribute (Column) {unordered}

Heading

R

A_1	...	A_n
Value		

Body

Relation (Table)

Tuple (Row) {unordered}

Tables

- Table (or **relation**) is the core data structure
  - A table is a **set** of **tuples**
  - A tuple is
    - an n-ary **sequence**
    - a **set** of key-value **pairs**
- Flat file had **one** table
  - We allow many!
  - **Named** tables
  - Aka **relations**

Relations!

- (We use **table** and **relation** interchangeably)
- Relations are like First Order Logic (FOL) **predicates**
  - Relation name == Predicate name
  - Number of columns == Arity of predicate
    - Person(bijan, u_o_manchester, ...)
  - Predicate is **true** (or false!) of its arguments
    - Relation is "true" of tuples which occur in it
  - Predicates can have
    - **definitions** (intensional!)
    - **facts** (extensional!)

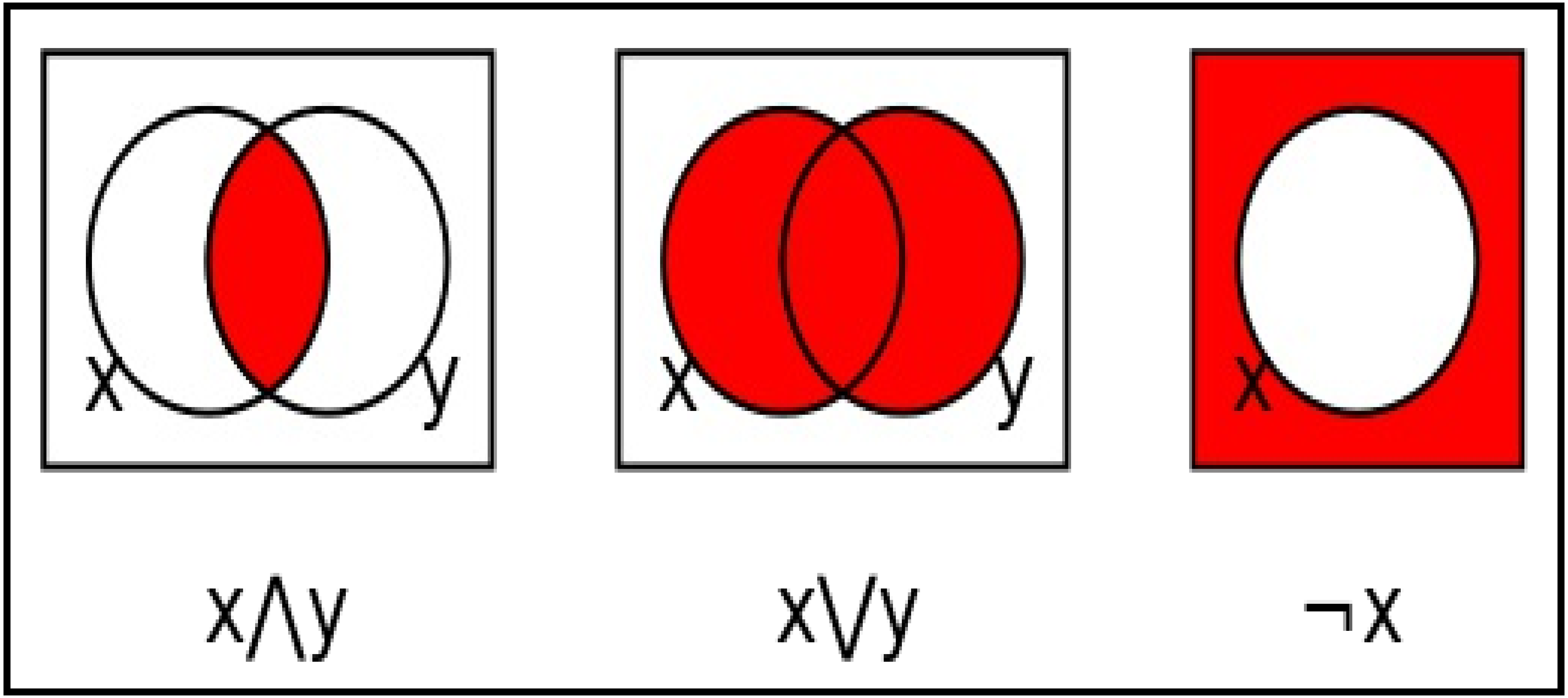
Order And Identity

- Records/Rows/Entities need **identity**
    - In Excel, we had the **row label**
        - The order or position of a record was significant
    - In our model, we need **distinguishing attributes**
        - We push identity *into* the data: a **key**
        - Either a naturally unique set of attributes
            - i.e., a definite description
        - or a made up one: an **ID**
- **Order** is always a property of the
    - data values
    - implementation

Multiple Tables

- Actions on multiple tables:
  - **Splitting** at
    - design time: try to normalize your DB
    - run time: dropping bits
  - **Combining**
    - Take two tables and produce a new table
- The key to relational domain modelling
  - **Decompose** your problem into "base" tables
  - **Derive** new tables for specific needs

A Relational Formalism



What Is A Formalism?

- A formal system (or *formalism*):
  - **syntax**: what can we write?
  - **semantics**: what does our writing mean?
  - with precise (mathematical) definitions
  - designed to capture a coherent set of operations
  - ("syntax" is loose, e.g., we might just have a collection of operators)

Key Goals Of A Formalism

1. to be **clear** about **what we mean**
   - In our spreadsheet is "1" a number, a string, either, both, something else?
2. to allow the determination of **key properties**
   - e.g., complexity of query answering
3. to **abstract** away from particular implementions
   - e.g., allow us to determine when wildly different implementations are *correct* thus can *interoperate*

Formalism Vs. Language

- Formalisms are often **abstract**
  - This can be an advantage!
  - Can be hard to use if **only** abstract
  - Concrete instances typically involve **compromise**
- We focus on concrete languages
  - Formalisms are the **theory**
  - Languages are the **practice**
  - Other Quotes On Theory vs Practice
    - Well, it may be all right in practice, but it will never work in theory.
    - In theory, there is no difference between theory and practice. But, in practice, there is.

SQL: A Language For Tables

- Schema
  - `CREATE TABLE` *table_name*
- Update
  - `INSERT INTO` *table_name*
  - `DELETE FROM` *table_name*
  - `UPDATE` *table_name*
  - …
- Query
  - `SELECT ... FROM` *table_name*

SQL operations (largely) are closed over tables

An Infelicity

There is a lot of lingo with slight different meanings. Concepts get divided up in slightly different ways.

Our talk	Common	Learning SQL p.10
Core Data Model		
Data Integrity	Data Definition	SQL schema statements "CREATE"
Data Manipulation	Query/Update Language	SQL Data statements

A Sample SQL Program

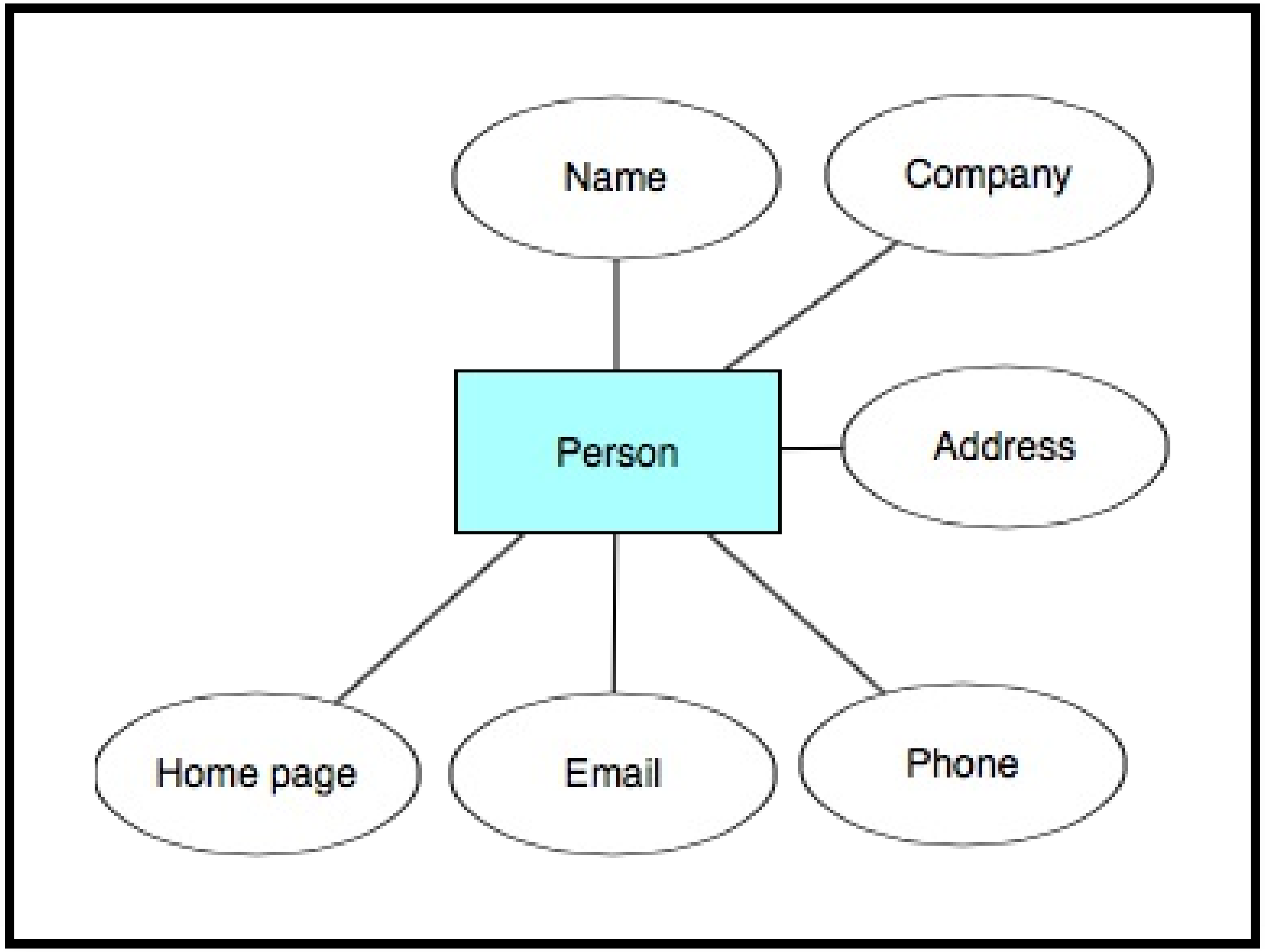
```
CREATE TABLE People (
    name varchar(255),
    company varchar(255),
    address varchar(255),
    phone varchar(255),
    email varchar(255),
    home_page varchar(255));

INSERT INTO People
    VALUES ('Aleshia Tomkiewicz', 'Alan D Rosenburg Cpa Pc',
            '14 Taylor St, St. Stephens Ward, Kent CT2 7PP',
            '01835-703597','atomkiewicz@hotmail.com',
            'http://www.alandrosenburgcpapc.co.uk');
SELECT name FROM People
```

- You must Define before Update before Query
    - I.e., CREATE before INSERT before SELECT

Modelling With SQL

- SQL lets us express models at the **logical** to (some of the) **physical** level
  - Specifying indices is a bit physical
  - Knowledge about implementation may inform modelling choices
- SQL has no mechanisms for **conceptual** level

Domain Model 1 In SQL



Domain Model 1 In SQL

```sql
 CREATE TABLE People (
    name varchar(255),
    company varchar(255),
    address varchar(255),
    phone varchar(255),
    email varchar(255),
    home_page varchar(255));

INSERT INTO People
    VALUES ('Aleshia Tomkiewicz', 'Alan D Rosenburg Cpa Pc',
            '14 Taylor St, St. Stephens Ward, Kent CT2 7PP',
            '01835-703597','atomkiewicz@hotmail.com',
            'http://www.alandrosenburgcpapc.co.uk');
...
```

Can we do all that we did in the spreadsheet?

SQL Manipulation Of DM 1

- Count records in your People table:

```
SELECT COUNT(*) FROM People
```

- Search for items:

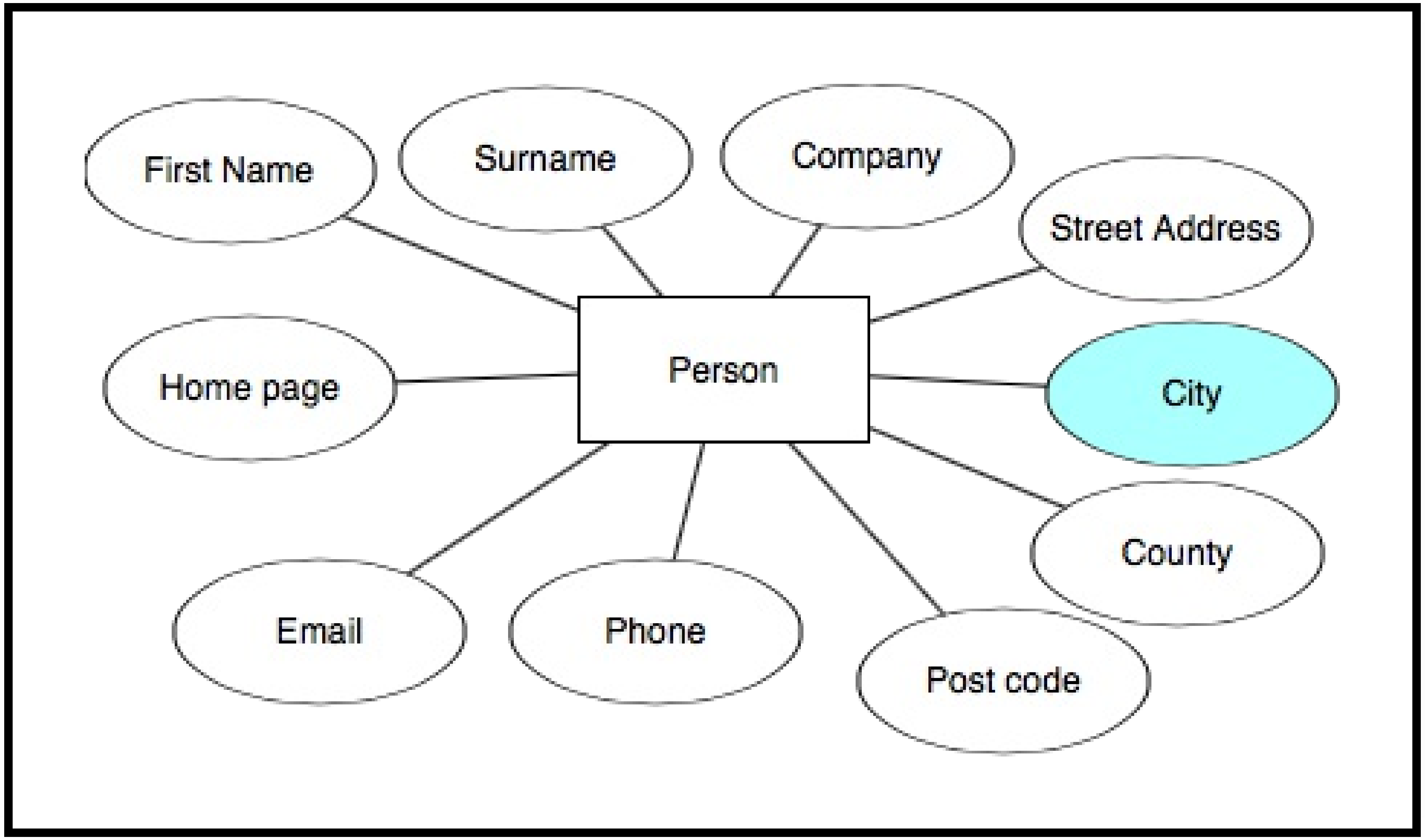
```
SELECT *  FROM People
WHERE name like 'Aleshia%'
```

```
SELECT *  FROM People
WHERE name like '%Tomkiewicz'
```

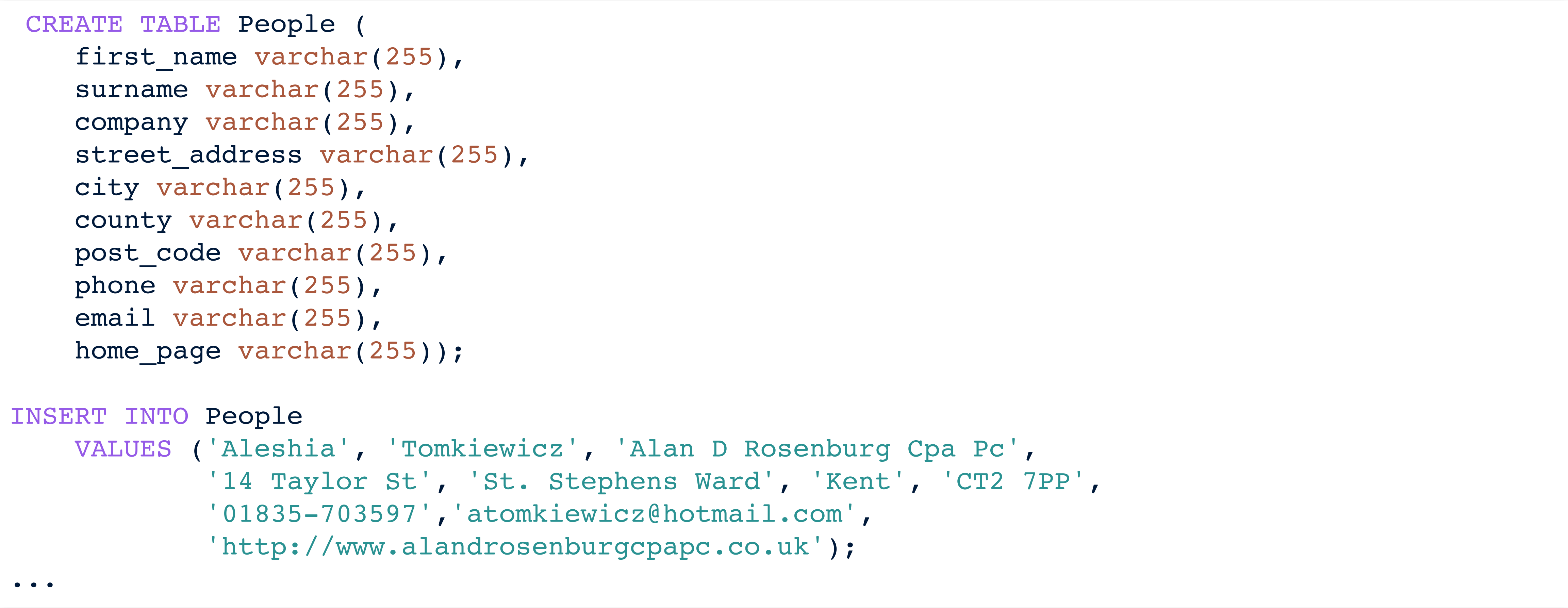
- Sort the table!

```
SELECT *  FROM People
ORDER BY name asc
```

Domain Model 2 In SQL



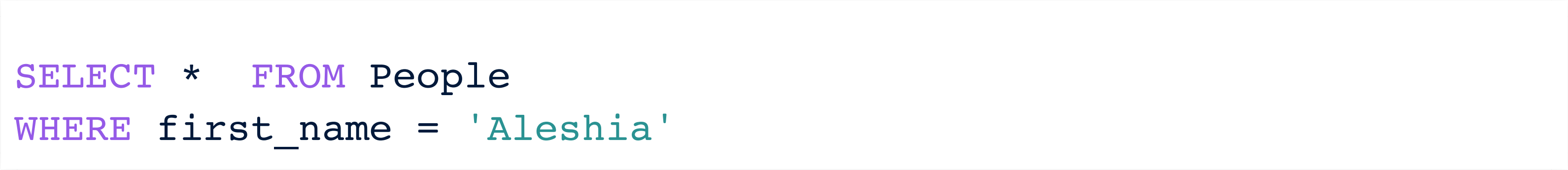
Domain Model 2 In SQL



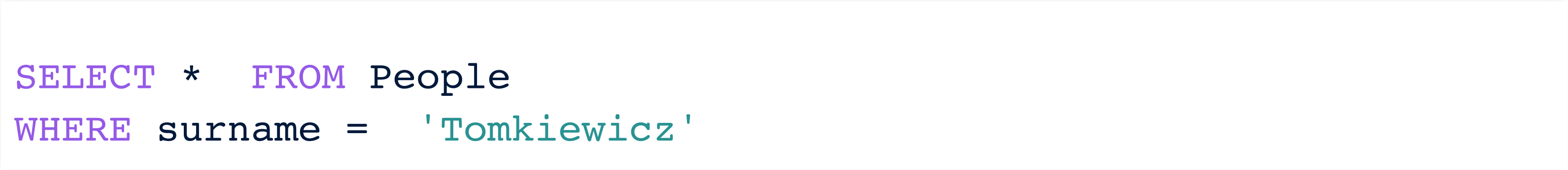
```
 CREATE TABLE People (
    first_name varchar(255),
    surname varchar(255),
    company varchar(255),
    street_address varchar(255),
    city varchar(255),
    county varchar(255),
    post_code varchar(255),
    phone varchar(255),
    email varchar(255),
    home_page varchar(255));

INSERT INTO People
    VALUES ('Aleshia', 'Tomkiewicz', 'Alan D Rosenburg Cpa Pc',
            '14 Taylor St', 'St. Stephens Ward', 'Kent', 'CT2 7PP',
            '01835-703597','atomkiewicz@hotmail.com',
            'http://www.alandrosenburgcpapc.co.uk');
...
```

SQL Manipulation Of DM 2

- The old queries work, but we can improve them
  - Search for items:

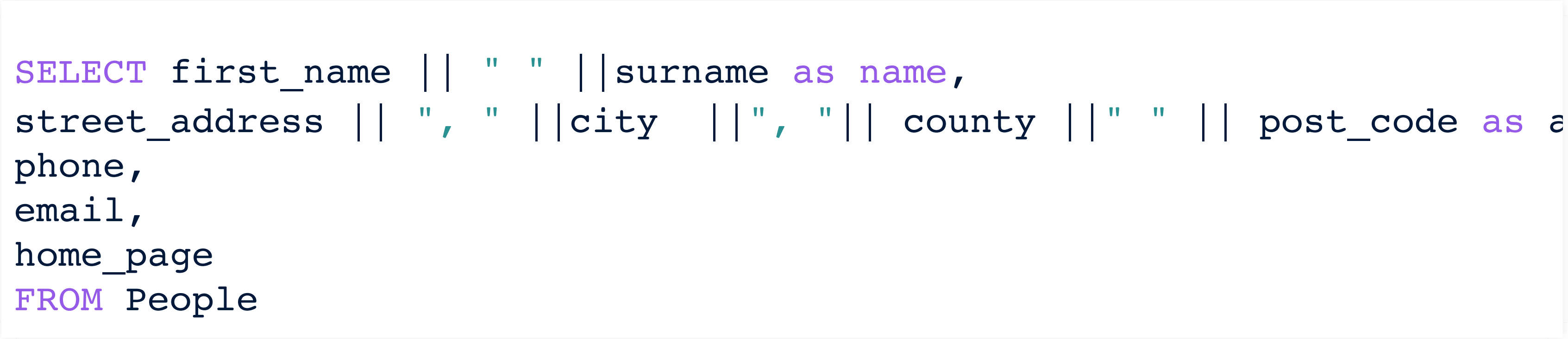


```
SELECT *  FROM People
WHERE first_name = 'Aleshia'
```



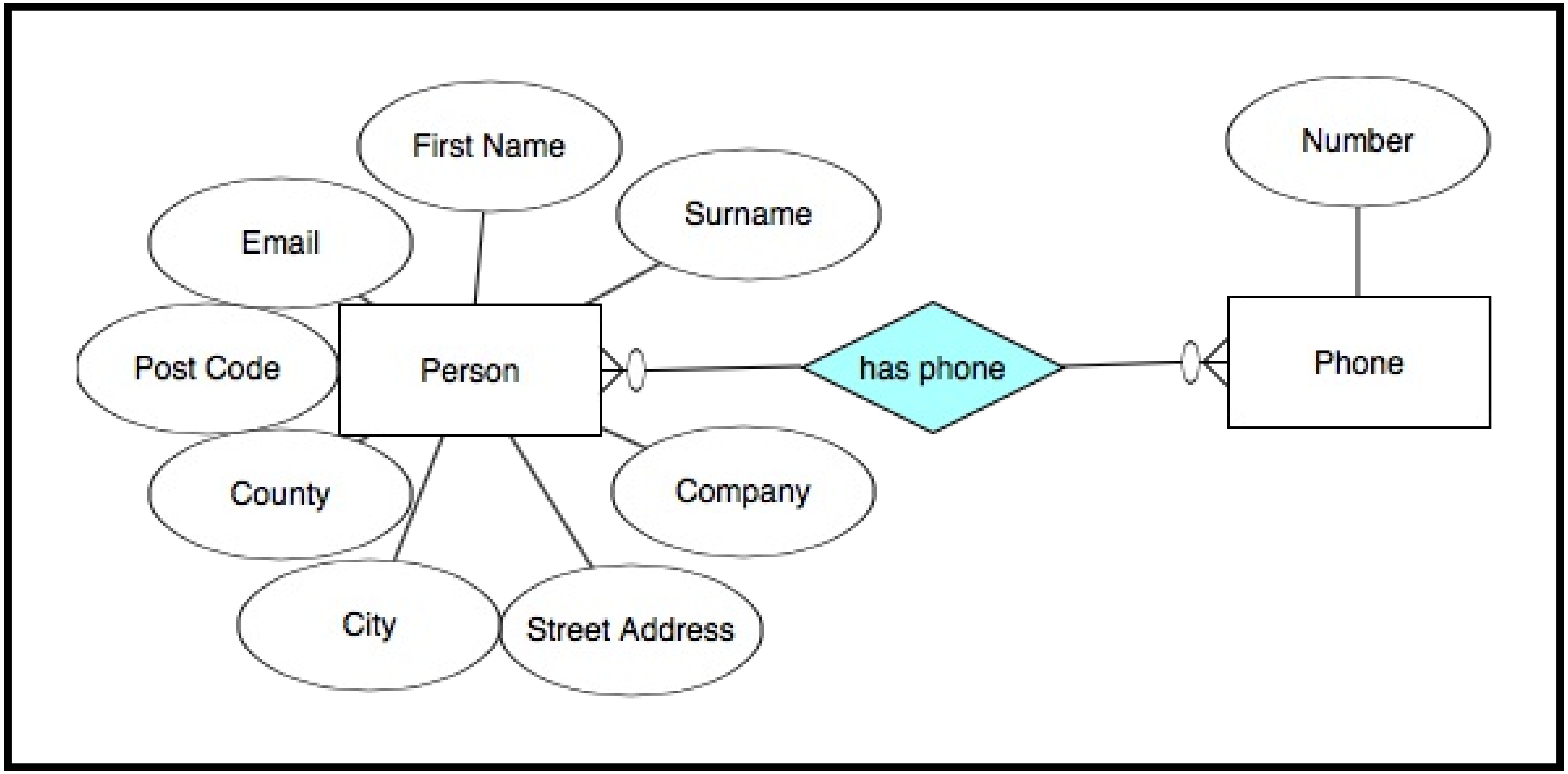
```
SELECT *  FROM People
WHERE surname =  'Tomkiewicz'
```

- We can recreate DM 1!

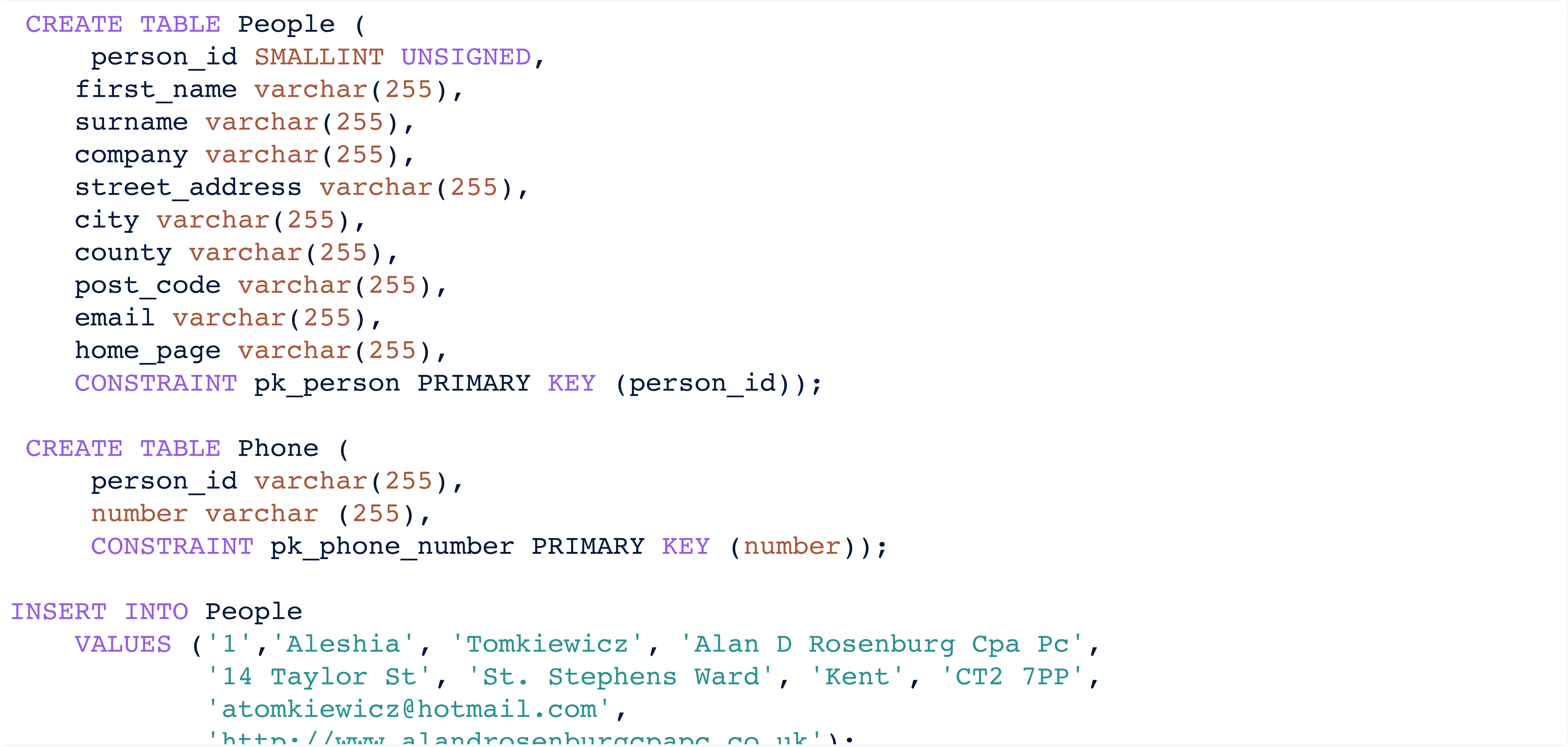


```
SELECT first_name || " " ||surname as name,
street_address || ", " ||city  ||", "|| county ||" " || post_code as a
phone,
email,
home_page
FROM People
```

Domain Model 3 In SQL



Domain Model 3 In SQL



```
 CREATE TABLE People (
     person_id SMALLINT UNSIGNED,
    first_name varchar(255),
    surname varchar(255),
    company varchar(255),
    street_address varchar(255),
    city varchar(255),
    county varchar(255),
    post_code varchar(255),
    email varchar(255),
    home_page varchar(255),
    CONSTRAINT pk_person PRIMARY KEY (person_id));

 CREATE TABLE Phone (
     person_id varchar(255),
     number varchar (255),
     CONSTRAINT pk_phone_number PRIMARY KEY (number));

INSERT INTO People
    VALUES ('1','Aleshia', 'Tomkiewicz', 'Alan D Rosenburg Cpa Pc',
            '14 Taylor St', 'St. Stephens Ward', 'Kent', 'CT2 7PP',
            'atomkiewicz@hotmail.com',
            'http://www.alandrosenburgcpapc.co.uk');
```

SQL Manipulation Of DM 3

- Recreate DM 1 and DM 2: easy
- Find everyone with same phone number
- Can we have unassigned numbers?

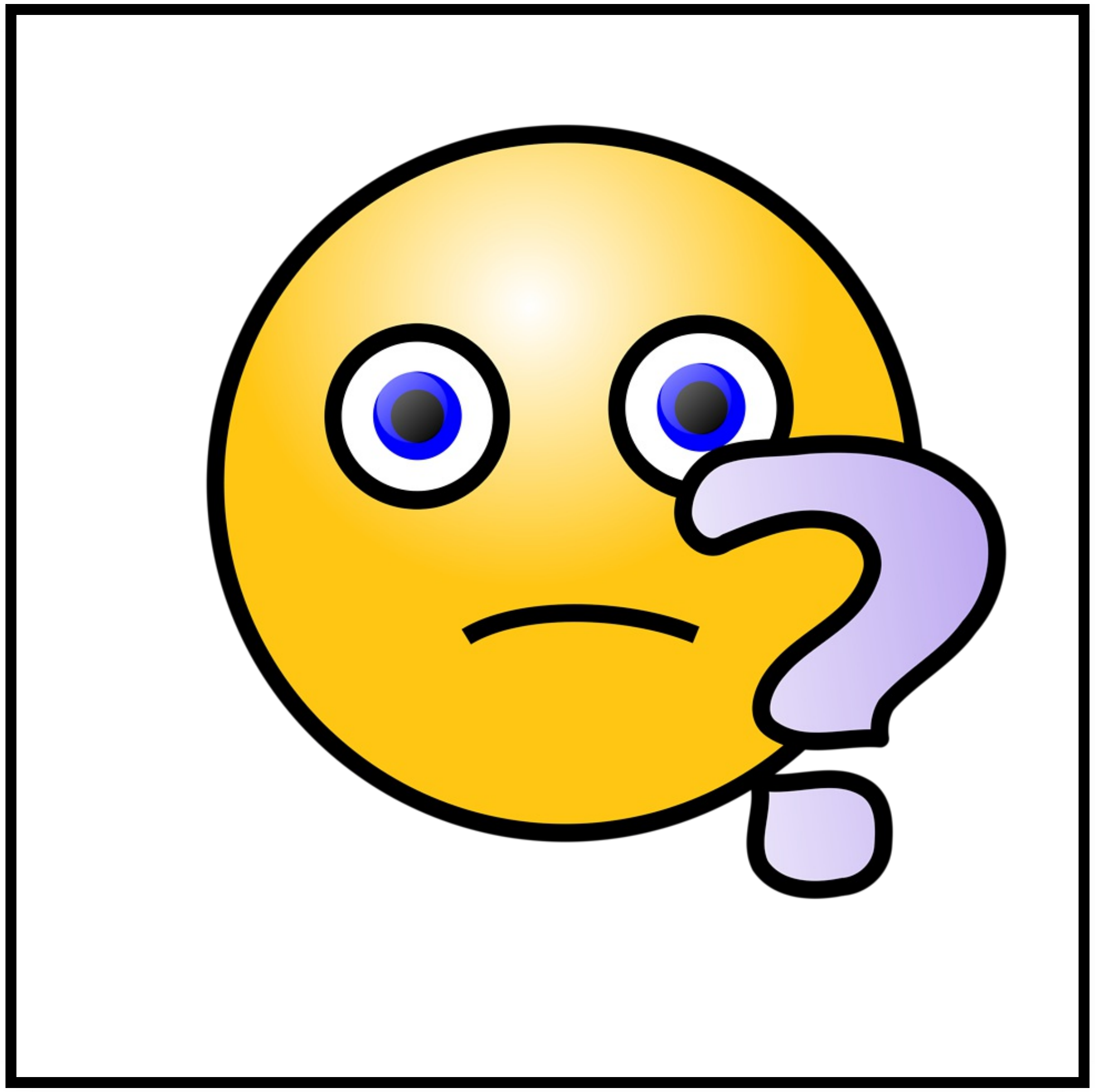
How’d DM Do?

- Core DM/Data structure
  - Tables seem to work
- SQL and Relational Model
  - We can do everything!
    - All queries in all models
    - Model 3 has 2 tables/requires joins
- Domain Model 3
  - Neater inserting and deleting
    - Can have as many phones as you want!
  - Every other domain model can be derived
    - Just write the query:
    - define as a view!

Expressive Power

- SQL is **expressive**
  - The core data model is rich
    - Composing and filtering tables does a lot!
    - Operators and functions helpful
      - Without concat, there'd be trouble!
  - The language is **powerful**
    - Reasonably **composable**
    - Lots of features
    - Extended and extensible in many implementations
      - Interop problems!

Querying With SQL



Schemas Vs. Queries

- CREATE statements
  - "create" *empty* tables
  - out of nothing at all
  - with certain constraints
  - with some expectation of permanence
- SELECT statements
  - "generate" *new* tables (possibly with data)
  - out of existing tables
  - according to some constraints
  - with no expectation of permanence

Closed Over Tables

- SQL is (mostly) **closed** over tables
  - Most SQL constructs take tables and produce tables
  - Clear exception: Functions!
- Manipulation is manipulation of tables
  - Not rows, columns, or cells directly
  - Rows, columns, and cells are "degenerate tables"...

Filtering V

- Key operation `SELECT`: ignoring some parts
  - Basically "find"
  - Can filter rows or columns or both
  - Requires "testing" functions on values

Filtering Columns

- "Projection"
- Specified in the SELECT clause
  - Keep all columns:

    ```
      SELECT * FROM People
    ```

  - Just a single column:

    ```
    SELECT county FROM People
    ```

  - Multiple columns:

    ```
    SELECT name, county FROM People
    ```

  - Rename columns:

    ```
    SELECT street_address AS address
    FROM People
    ```

Filtering Rows

- Just called "filter" or "selecting"
- Specified in the WHERE clause of your query:
  - Equality:

    ```
      SELECT * FROM People
      WHERE surname = "Smith"
    ```

  - Range:

    ```
    SELECT * FROM People
    WHERE heartrate > 95
    ```

  - Compound criteria:

    ```
    SELECT * FROM People
    WHERE heartrate > 95 AND county="Kent"
    ```

Building Tables With Cross Join

- The fundamental operation is Cartesian product
  People x Phone
- This makes a new row out of **every** pair of rows between the two table
  - What's the size of the result?
- Not really a user-oriented feature
  - "Incidentally" cross joins are dangerous!

Building Tables With Inner Join

- An **inner join** is a join *filtered* on common columns
  - Useful for our phone records!

```
SELECT * FROM People, Phone
INNER JOIN ON People.person_id = Phone.person_id
```

- (Special case called a "natural" join.)

Building Tables With Outer Join

- An **outer join** is like an inner join but it returns also rows that don't have a match in the other table
  - *left outer* different from *right outer*

```
SELECT * FROM People, Phone
RIGHT OUTER JOIN ON People.person_id = Phone.person_id
```

- will return also people who have no phone!

Building And Filtering

- Once we've built a table we can filter things we need:

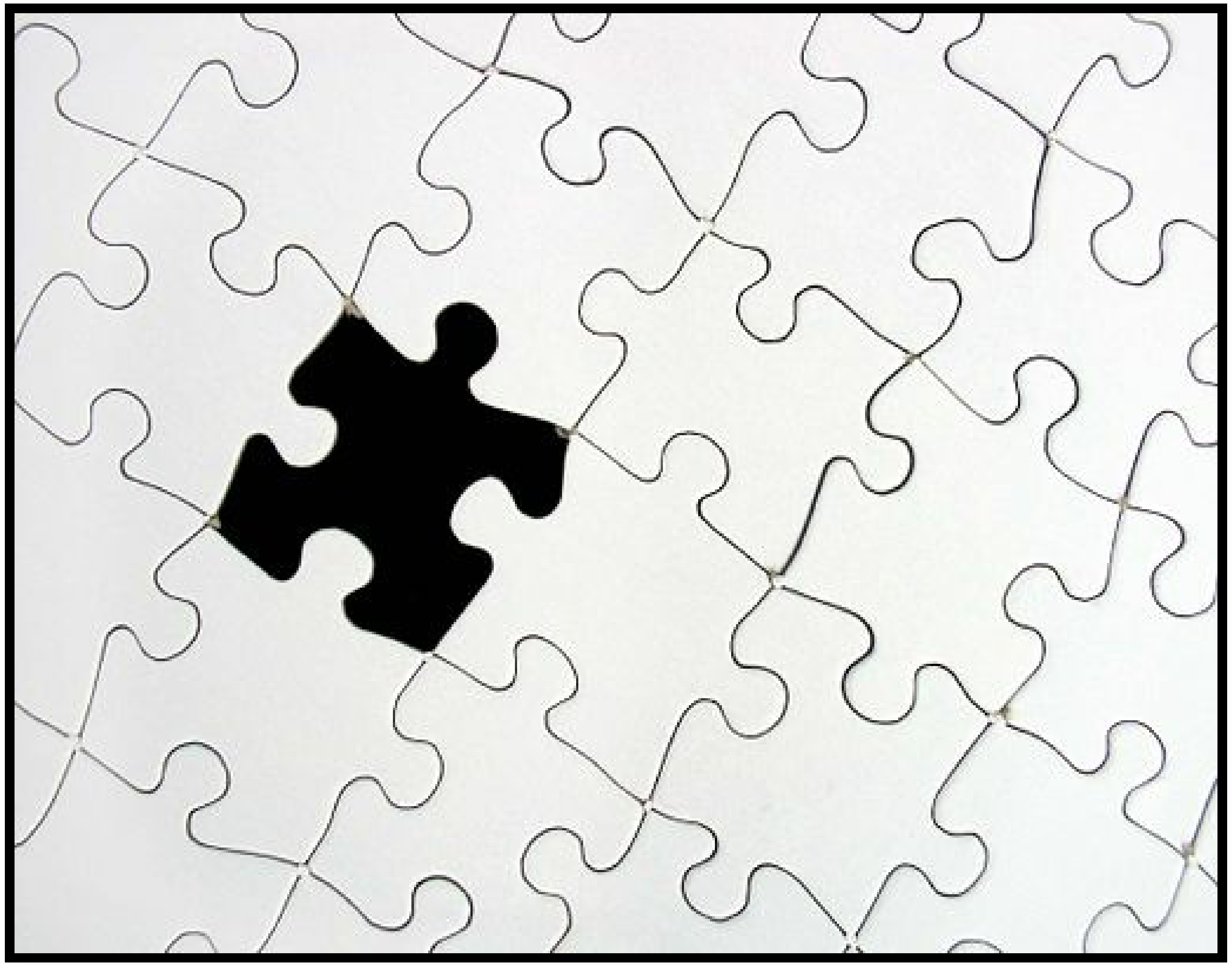
```
SELECT * FROM People, Phone
RIGHT OUTER JOIN ON People.person_id = Phone.person_id
WHERE People.surname = "Smith"
```

- ...you knew that already!?

The Cost

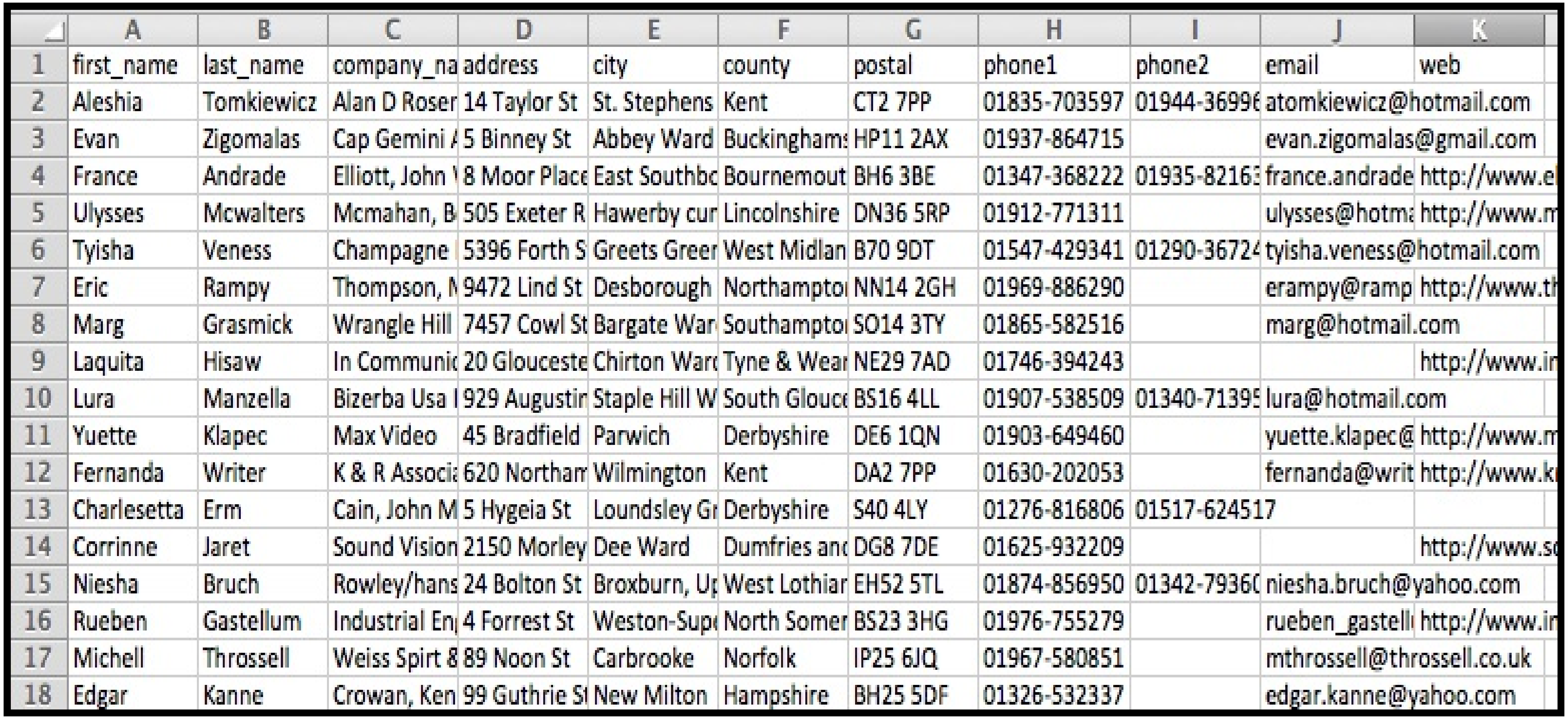
- A **key issue** with joins
  - Worse case is a CROSS
  - Even if you don't **generate** the CROSS
    - You might have to **consider** all the pairs
    - (If you aren't careful)
- **Good optimisers** avoid both
  - Considering lots of matches (think indexes)
  - Generating large intermediate tables

Incomplete Data



Multiple Phone Columns

- Some people have **none or one**
- Or no **email** or **web page**



	A	B	C	D	E	F	G	H	I	J	K
1	first_name	last_name	company_na	address	city	county	postal	phone1	phone2	email	web
2	Aleshia	Tomkiewicz	Alan D Roser	14 Taylor St	St. Stephens	Kent	CT2 7PP	01835-703597	01944-36996	atomkiewicz@hotmail.com	
3	Evan	Zigomalas	Cap Gemini /	5 Binney St	Abbey Ward	Buckinghams	HP11 2AX	01937-864715		evan.zigomalas@gmail.com	
4	France	Andrade	Elliott, John \	8 Moor Place	East Southbc	Bournemout	BH6 3BE	01347-368222	01935-82163	france.andrade	http://www.e
5	Ulysses	Mcwalters	Mcmahan, B	505 Exeter R	Hawerby cur	Lincolnshire	DN36 5RP	01912-771311		ulysses@hotma	http://www.m
6	Tyisha	Veness	Champagne	5396 Forth S	Greets Greer	West Midlan	B70 9DT	01547-429341	01290-36724	tyisha.veness@hotmail.com	
7	Eric	Rampy	Thompson, N	9472 Lind St	Desborough	Northamptoi	NN14 2GH	01969-886290		erampy@ramp	http://www.th
8	Marg	Grasmick	Wrangle Hill	7457 Cowl St	Bargate War	Southamptoi	SO14 3TY	01865-582516		marg@hotmail.com	
9	Laquita	Hisaw	In Communic	20 Glouceste	Chirton Warc	Tyne & Wear	NE29 7AD	01746-394243			http://www.in
10	Lura	Manzella	Bizerba Usa I	929 Augustir	Staple Hill W	South Glouc	BS16 4LL	01907-538509	01340-71395	lura@hotmail.com	
11	Yuette	Klapec	Max Video	45 Bradfield	Parwich	Derbyshire	DE6 1QN	01903-649460		yuette.klapec@	http://www.m
12	Fernanda	Writer	K & R Associa	620 Northam	Wilmington	Kent	DA2 7PP	01630-202053		fernanda@writ	http://www.k
13	Charlesetta	Erm	Cain, John M	5 Hygeia St	Loundsley Gr	Derbyshire	S40 4LY	01276-816806	01517-624517		
14	Corrinne	Jaret	Sound Vision	2150 Morley	Dee Ward	Dumfries anc	DG8 7DE	01625-932209			http://www.sc
15	Niesha	Bruch	Rowley/hans	24 Bolton St	Broxburn, Up	West Lothiar	EH52 5TL	01874-856950	01342-79360	niesha.bruch@yahoo.com	
16	Rueben	Gastellum	Industrial En	4 Forrest St	Weston-Supe	North Somer	BS23 3HG	01976-755279		rueben_gastell	http://www.in
17	Michell	Throssell	Weiss Spirt 8	89 Noon St	Carbrooke	Norfolk	IP25 6JQ	01967-580851		mthrossell@throssell.co.uk	
18	Edgar	Kanne	Crowan, Ken	99 Guthrie St	New Milton	Hampshire	BH25 5DF	01326-532337		edgar.kanne@yahoo.com	

No Surname

- Even if we normalised that away
    - Some people don't have a surname!

Null

- `null` is a distinguished value which can mean:
  - "Value not yet known"
  - "Not applicable to this entity"
  - "Value undefined"
  - check out LSQL
- Key property: Unequal to everything
  - `null = null` is **never** true
  - Match on not `null`, rather than `null`

Strange value!

Outer Joins

- If you have no `nulls` in your base tables
  - you can't get them in tables derived by inner join
- However, the 2 phone column table **is** derivable
  - We use the **outer** join
  - Outer joins take a table T
    - for each row in T
      - extend it with the (projected) columns from another table
      - *If* there's a match, add the matched values
      - *else, add `null`s
- See Learning SQL Chapter 10 for some worked examples

Null Proliferation

- `null` never matches
    - So iterated outer joins proliferate `null`s
        - As you get wider, you get sparser
            - If you are matching on a sparse attribute
- `null`s pose challenge for relational theory
    - And somewhat for practice
    - Starts moving from the sweet spot

SQL And The Web

A brief tour

SQL Driven Websites

- Many websites are **backed by** a database
  - PHP makes it easy
  - Consider WordPress and other CMSs
- Lots of **unstructured** content
  - Stuff in blobs and text fields
- Key properties
  - Scaling
  - ACID: Atomicity, Consistency, Isolation, Durability
    - Transactions
  - Concurrent access

There is a key historical text that is still good reading, esp chps 11-12

CSV & SQL Programs On The Web

- UN Data repository
- Other government repositories:
  - data.gov
  - data.gov.uk
- Scientific sites
  - cinicaltrials.gov
  - uniprot.org
  - ...

Google Query Viz Language

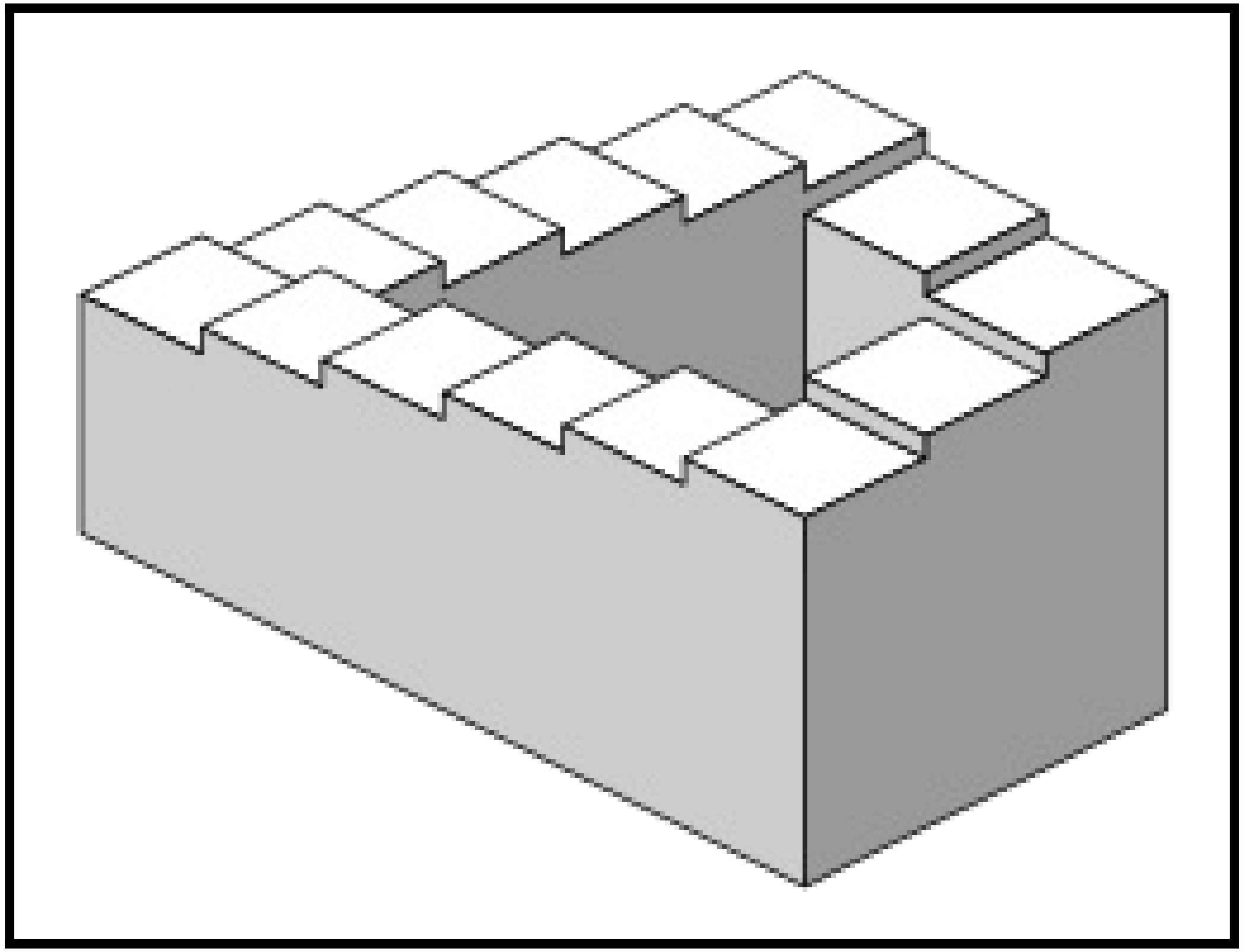
- A SQL like language
  - Used in Google Docs Spreadsheet
  - QUERY function takes queries as argument

WebSQL

The WhatWG and W3C tried to standardize WebSQL

> *This specification introduces a set of APIs to manipulate client-side databases using SQL.*
>
> ```
> function prepareDatabase(ready, error) {
>   return openDatabase('documents', '1.0', 'Offline document storage', 5*1024*1024
>     db.changeVersion('', '1.0', function (t) {
>       t.executeSql('CREATE TABLE docids (id, name)');
>     }, error);
>   });
> }
> ```
>
> *Local database backed web apps*
>
> - *For offline use*
> - *Just increased capabilities*

Next Steps



Reading

There is a key historical text that is still good reading, esp chps 11-12

Any Questions So Far?

Labs & Coursework

- Next, we go to the Labs
- You look in BB at Week 1 coursework:
  - Quiz Q1
  - Short Essay SE1
  - Small Modelling exercise M1
  - Some querying CW1
- Read, think, ask us!