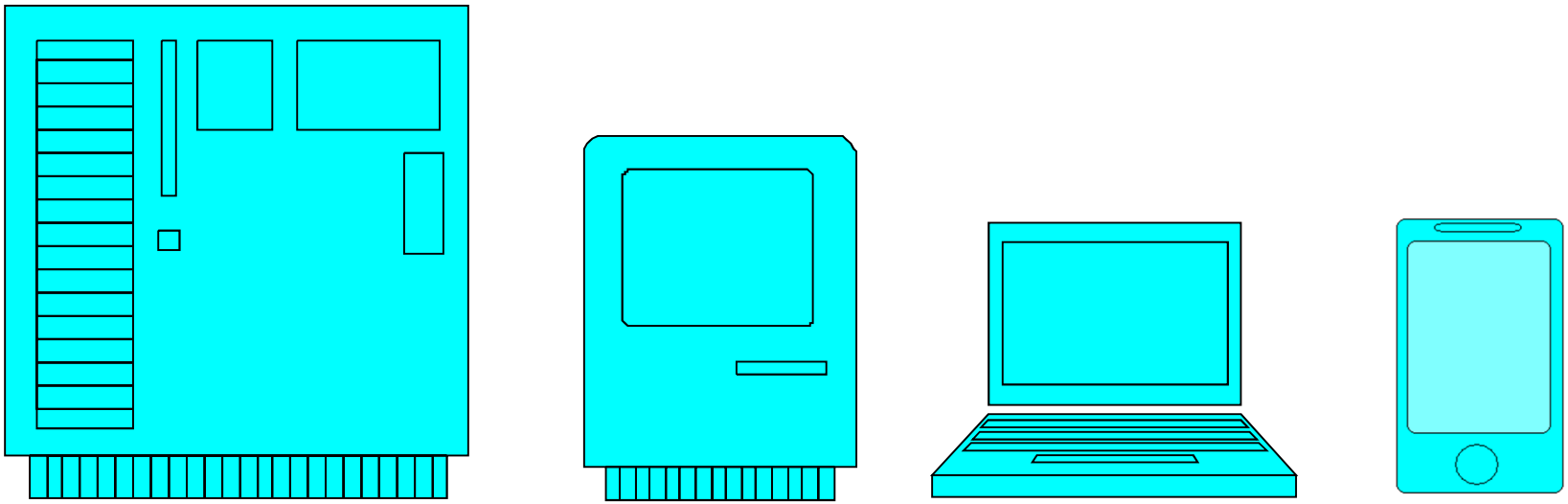
# WeWorkWithData

- Formed in Summer of 2009
- Steve Taylor, Mike Trotman & Antony Saccomani
- Straightforward work : Data Collection and Data Analysis
- Areas of speciality : Complex, bespoke reporting, PowerPoint automation, complex analysis & data consultancy (e.g. migration projects) & advertising research
- Clients from marketing consultancies, large agencies, “boutique” or speciality agencies and music industry
- SSS integral to all parts of our workflow

# Evolution of Data in Computing : 1



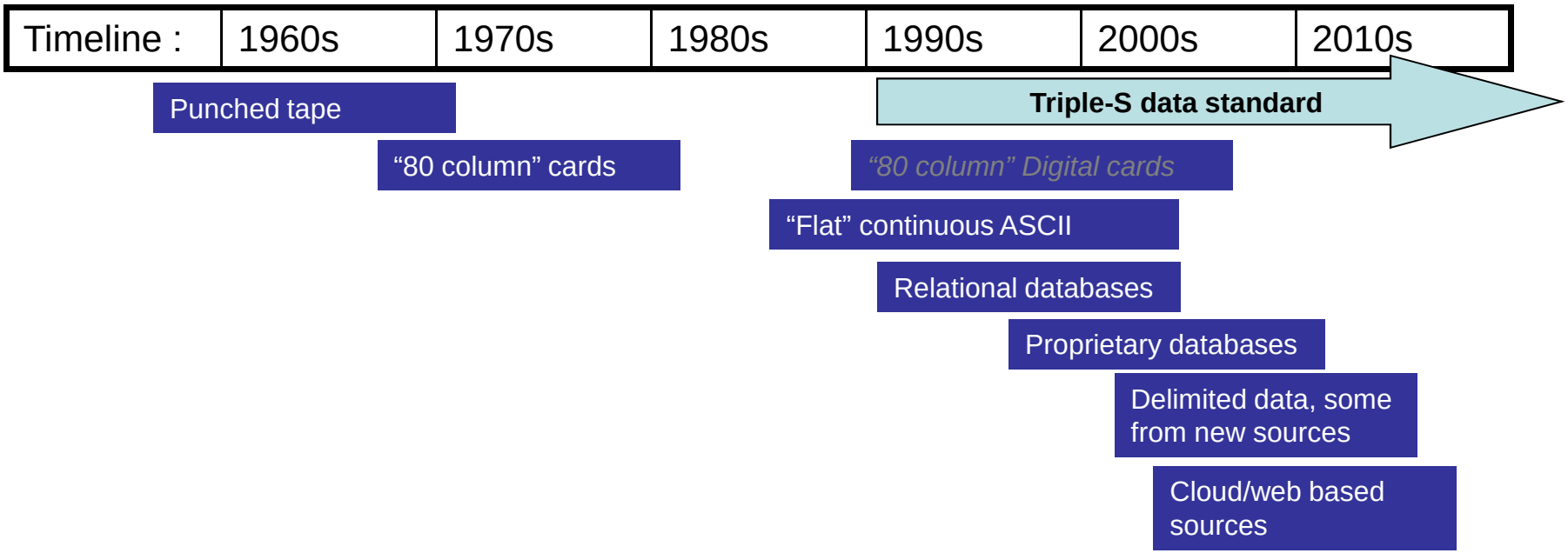
Over 30 years, the physical size of computers has reduced dramatically

Inversely, the raw computing power has increased

Disk storage has increased, along with speed of access

As cloud computing takes off, devices will become even smaller

# Evolution of Data in Computing : 2




Timelines are approximate. This is more the emergence of these types *in our work*

Main message is that data keeps evolving

Data has more utility now; changes in method enable more flexibility

More tools to view, manipulate and analyse data

# KISS : Keeping it simple (stupid!)

- There are No Prizes for complexity!
- Often found “macho” approach to spec-writing/data analysis:
    - Behold my enormous complex code!
    - Making oneself “indispensible”/protecting positions

- w3dL prefer to break things up into “process flows”
- Easier to check, easier to audit or fault-find
- Clear, readable code
- Agreed internal standards for work

Utilising readable and legible data forms part of this…

…therefore CSV is ideal

# Negative Issues with CSV data

- Once again, MR software can be slow to “catch up”
- As new modules in software start to read it, often encounter problems :
  - Embedded LF/CR characters within fields
  - When is a quote a quote? “ ` ‘ “
- Different types of CSV data from different platforms
- Can be some dangerous effects :
  - Column slippage – sometimes from quotes
  - Excel & other Microsoft software knowing your data better than you
    - Age band “12-16” suddenly becomes “16th December 2011”

THESE ARE SURMOUNTABLE

# Positive Issues with CSV data

- Compact – no data in some fields? Then NO FIELDS!
- Platform neutral – they are text files
- Since lifting of limits in Excel 2007 onwards, now have a powerful tool at our disposal that can filter, count and examine the raw data
- Everything, to some degree is “text”
- Fits into a Perl programming environment well, where there is greater “type” freedom or “weak typing”
- Faster
- Safer
- Less risk
- 1 step away from almost any tool

# Where's my bloody data?

## *Who cares?*

- "Where it is?" is not initially important but with some **important** caveats
- Death of data maps to some degree – in terms of layout
- Easier to get data from third party companies – most databases have this export as standard
- Relatively trivial to do an initial, basic read of data where the name of the variable, coded into the 1st row of the file becomes **important**
- Communicating conventions on variable naming to third parties becomes **massively important**
- Simplify naming conventions : no "funny" characters like underscores or punctuation – *not* lowest common denominator.

It is "keeping things simple"

# What's in a name?

- The correct name is very important
- On complex International projects, we take the initial stream and convert it to CSV
- From SPSS SAVs, column binary, ASCII… one read or conversion then preserve as CSV
- Makes examination very simple indeed and problems easier to explain to others

# “Losing your head”

- A header row is good practise, useful, but not necessarily conclusive
- SSSXML plugs the gap
- Not only “where” it is, but “what it is”
- Ease of finding and examining non-compliant data
- “Stop fretting about data maps” – whole layer of unnecessary detail in 2011
- Where have your problems with external data stemmed from? Names? Or Position of data? Solution is with CSV

# Ease of merging/blending new data

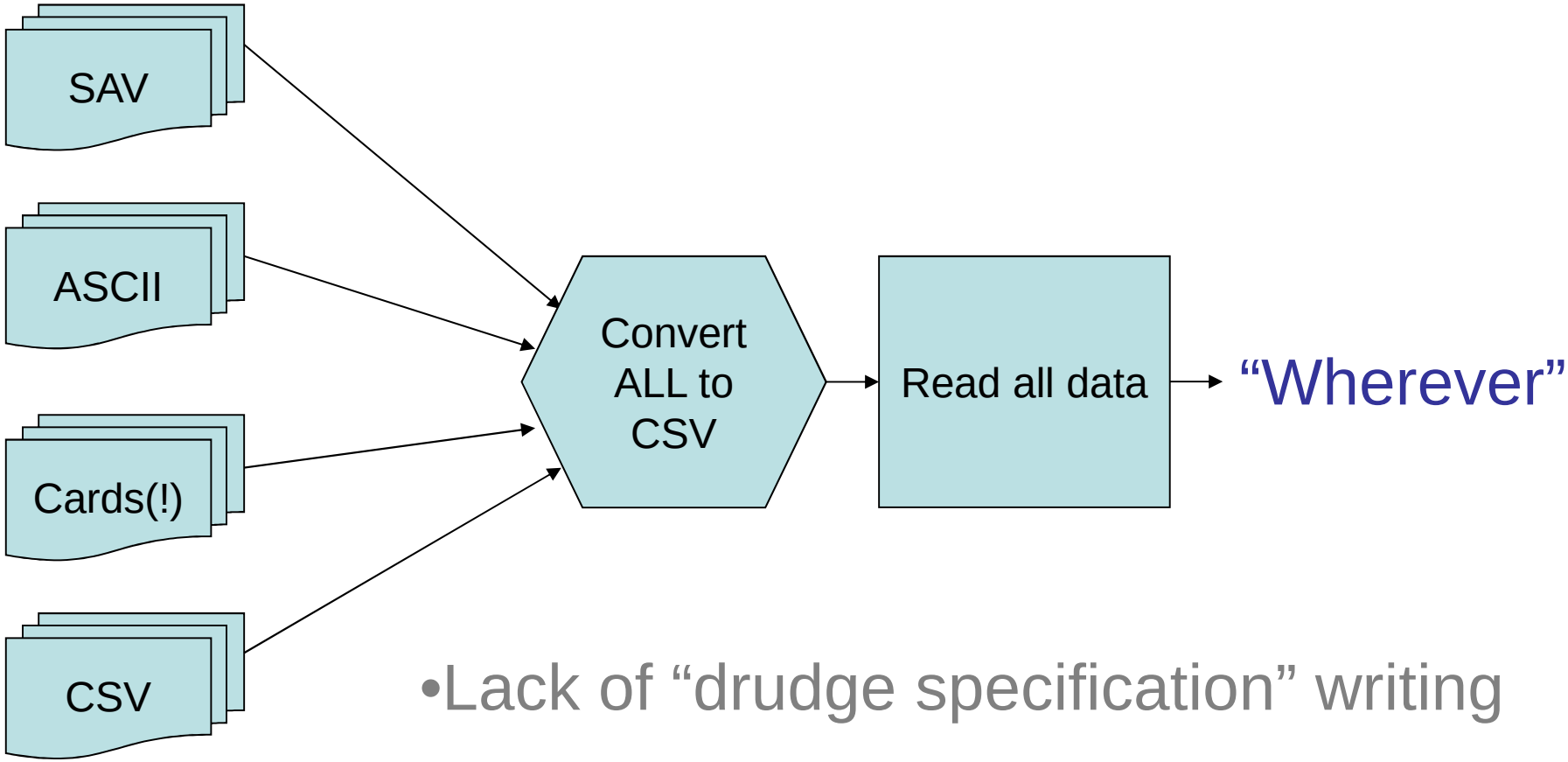
- CSV format makes it a lot easier to add new data to existing records
- Can use applications such as Excel (with serial number provisos!)
- Size of data & width of fields are not significant issues

It’s all about the CONTENT of the data

# SSSXML assisting Data Verify & Blend

- Simple to build new SSSXML to describe external or additional files
- We concentrate on names of variables and the codelists
    - i.e. Look at “q9”, not the “fifth variable”
- Using XML tools on incoming multi-country data we can :
    - Generate and compare lists of variables
    - Perform wholesale updates of variable names
    - Verify codelists for correct number/order of codes
    - Produce colourful HTML exception reports – then address issues
    - Re-order codelists if necessary
- All scripting can be re-used on newer revisions of datafiles

# Better process flow




- Lack of “drudge specification” writing
- “Post speccing”

# Conclusions

- We're hoping that there is far wider implementation in MR Software in the future
    - Both as input streams *and output streams*
- We work with lots of external data on projects, this is now invariably arriving as CSV data
- Stricter implementation of naming conventions for multiple field data
    - For once this is already in SPSS!
- Wider agreement between MR Software companies on this – informal at first?