MSc in Official Statistics
Statistical Computing: Statistical Production Systems

Andrew Westlake
Survey & Statistical Computing
63 Ridge Road, London N8 9NP, UK
+44 (0) 20 8374 4723
AJW@SaSC.co.uk (E-Mail)
www.SaSC.co.uk

Statistical Production

- Statistical Production Systems
  - » What is needed
  - » How are they designed and built
  - » Role of the Statistician in the design and development process
- Different requirements in different contexts
- Data collection and processing technologies

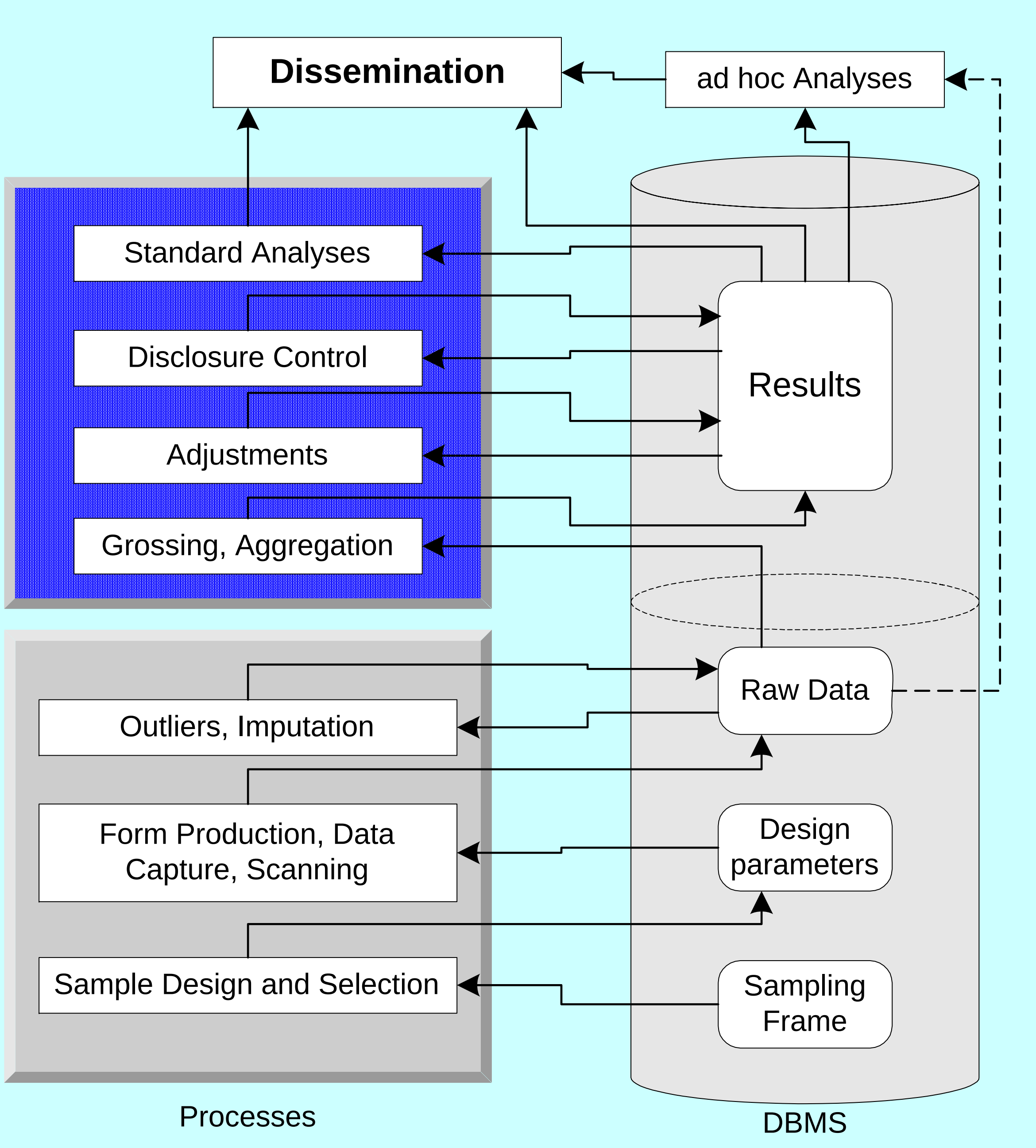
- Development methodologies
- UML as a design and specification tool

Statistical Production

- Different views of the overall process are possible
  - » Appropriate view depends on available skills and resources, and on the overall context
  - » Census, Sequences of surveys, Continuous surveys, Panels, Statutory inquiries, Statistical use of register systems
- Statistical and Production issues both important
- Correct Interpretation is vital
  - » Precision, Bias
  - » Imputation, Adjustment
  - » Derived Measures

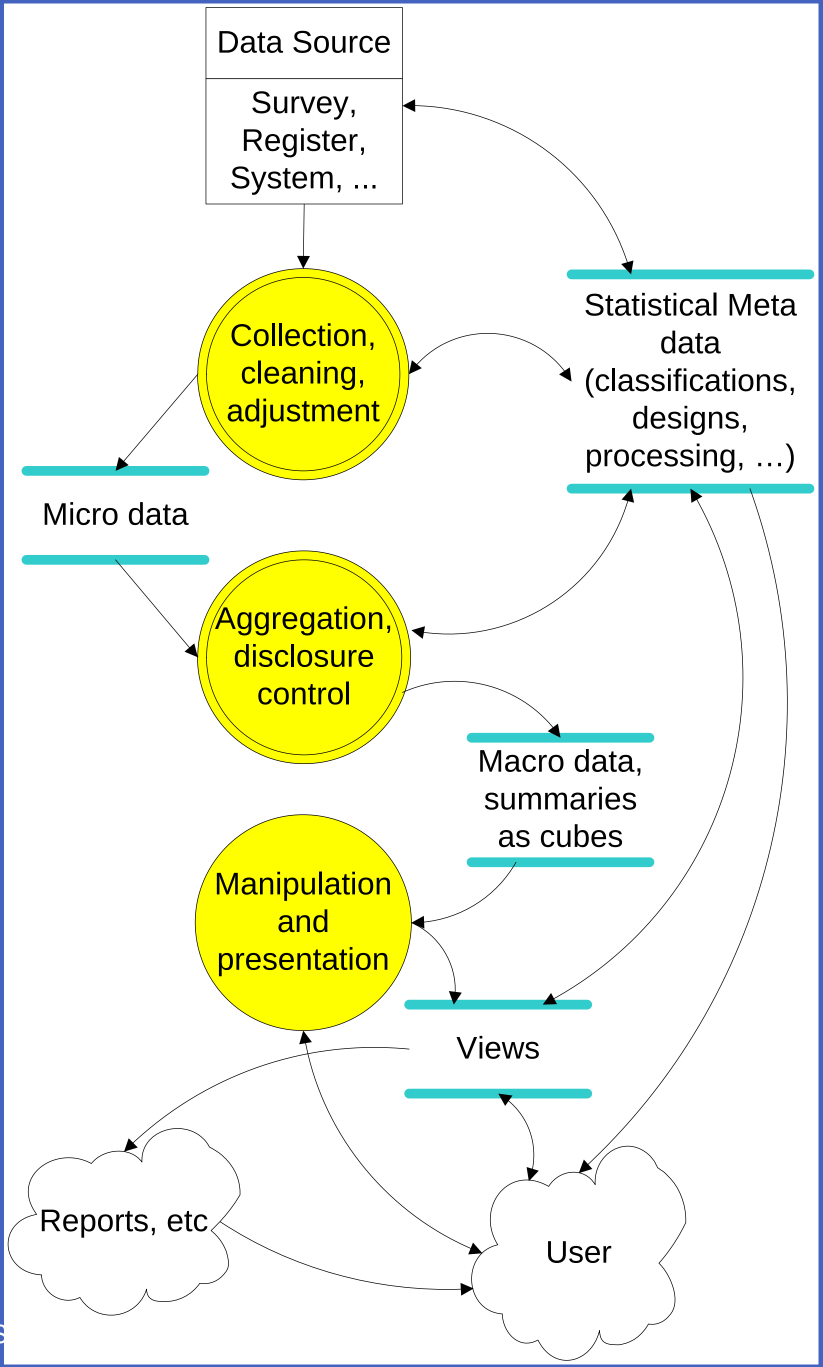
Processing Statistical Data

- Focus on production processes
  - » Centred on database / repository



Processing Statistical Data

- Focus on Structure and Functionality
  - » Production and ad-hoc use based on the same resources and functionality



Data Source

Survey, Register, System, ...

Collection, cleaning, adjustment

Statistical Meta data (classifications, designs, processing, …)

Micro data

Aggregation, disclosure control

Macro data, summaries as cubes

Manipulation and presentation

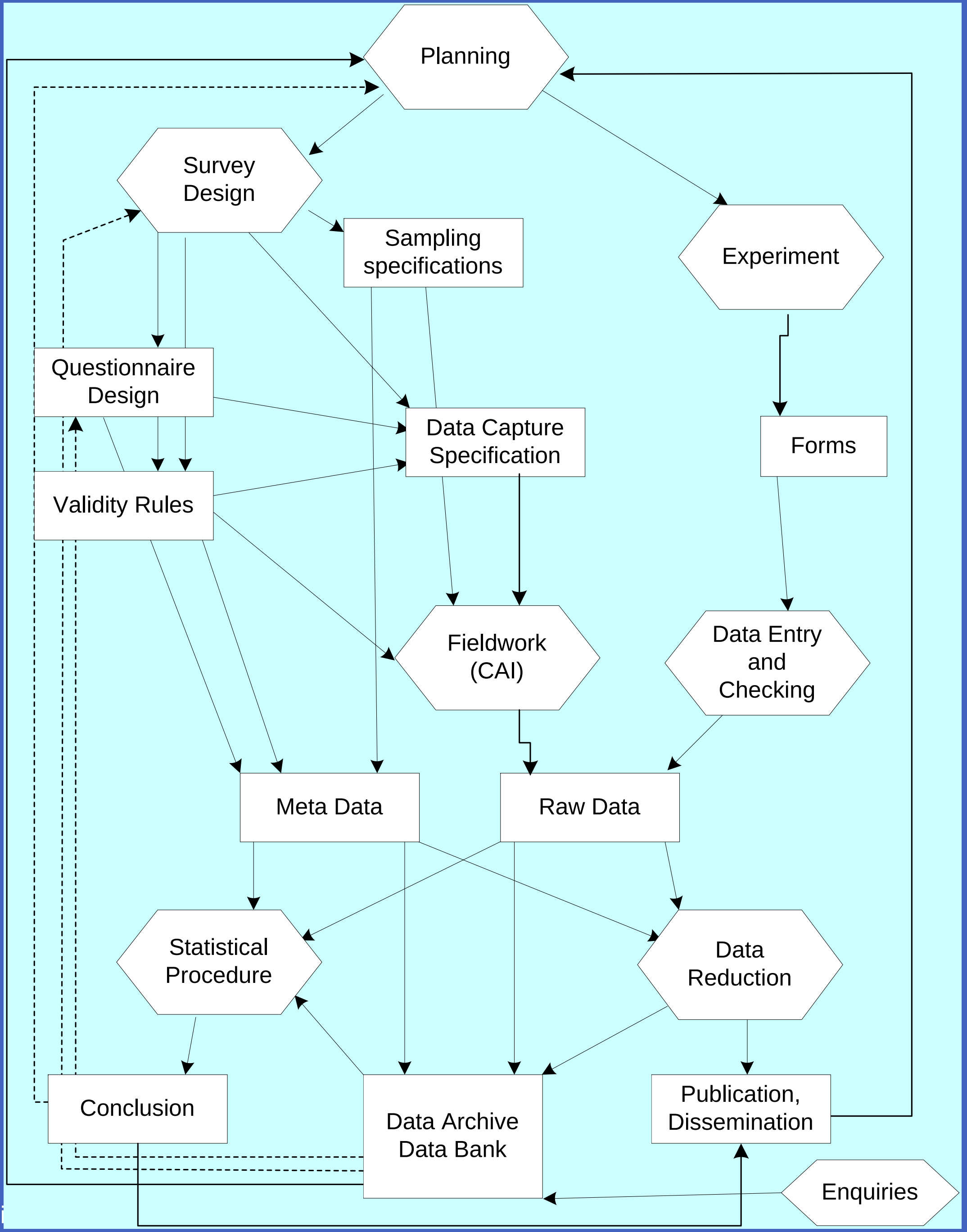
Views

Reports, etc

User

Processing Statistical Data

- Focus for Sequences of Survey Activities
  - » Each survey needs careful planning of all stages, every time



Processing statistical data

- Populations and sample design
- Data capture technologies
- Checking and imputation, missing values
- Post-processing
  - » Grossing for populations, adjustment for bias or different populations, indexing, standardisation and calibration
- Tabulation and report production
  - » Disclosure control
  - » Sampling errors – complex sample design
- Analysis and interpretation, dissemination
- Integration of steps into complete processing systems
  - » Scope for automation
  - » Need for human judgement

Technology choices for data capture

- Paper, Phone, Interviewer visit, e-mail form or program, web page
  - » Manual entry or scanning from paper
  - » Interviewer or Self-completion
- Cost of system set-up
- Cost of interview and data entry
- Implications for data validation
  - » How much intelligence to catch inconsistencies during entry, while the respondent is available for resolution

Data cleaning

- Outlier detection
  - What is an outlier?
    - External rules
    - Use of previous, prior or population information, to judge likelihood of observed value
- Outlier correction
  - Adjustment of observed value to ‘improve’ it
  - Windsorisation – fixed adjustment process
  - Empirical Bayes adjustment
    - Uses previous, prior or sample information, in a weighted combination of the observation and other evidence
- Trade-off between
  - Letting the data speak
  - Maintaining known (prior) distributions

Missing value methods

- Filling in data that should be present
    - It can be correct that data is missing
  - » Fixed values (e.g. mean)
  - » Hot-deck (similar records)
  - » Single and multiple imputation (Rubin)
    - Uses model of relationship of missing to other variables, fitted from present data
- Omit incomplete data from analysis
  - » By record or by analysis
- All methods are bad!
  - » All require some model of the relationship of the missing to the present data
    - Some models are good
  - » So imputed values merely reflect the model
    - **No additional information**
  - » Better to fit and use a good model directly to the available data
    - But not easy
  - » Analysis of expanded data should give same results (including precision) as correct analysis of available data
    - What is the correct sample size?

Adjustment from Sample to Target

- Weighting for selection probability
  - » Most statistical packages only support precision weights, not probability weights, though linear results are the same
- Weighting to known population size (Calibration)
  - » Post-stratification methods (Kalton)
  - » E.G. Calmar macro (SAS), g-Calib (SPSS)
- Adjustment for bias or different populations
  - » Sampled population may differ from Target Population
  - » e.g. VAT returns omit small businesses
- Indexing and standardisation
- All can be implemented as manipulations of appropriately aggregated data, do not need the micro data

Analysis functionality

- Use statistical packages for real statistical analysis
  - » Require links between the database and the package, e.g. ODBC
    - Metadata issues
    - Could be export for stand-alone data
- Much statistical reporting is just tabulation with commentary
  - » Decision makers want conclusions, not data
  - » Data users may have different questions, want more detail, more flexibility
- Suggests use of aggregate data as primary dissemination format
  - » Needs supporting manipulation and presentation functionality
- Role for Office products in presentation
  - » E.g. Manual editing in Excel, Word, PageMaker
  - » But not for production (could be used by a production system)
- Web dissemination of results and resources
  - » Basic CMS good for news and conclusions
  - » Need more functionality for exploration and discovery

Summary

- Different approaches and requirements in different contexts
  - » Need to be flexible in approach
- Lots of technical details in statistical processes
  - » These will be unfamiliar to IT specialists
- Detail or Generality? – do both, different levels of abstraction
  - » Top-down – broad scope, generalised functionality, inclusive
  - » Bottom-up – getting the details right for functions and users
- Implementation of real systems
  - » Usually select a subset of features for detailed implementation
  - » Important to retain the 'big picture' to facilitate further development
  - » Incremental implementation – current IT mantra, but not what accountants want
- Example
  - » Features for handling aggregate data in multi-way tables

Aggregated Results, as Multi-way Table

Disease Incidence Reports

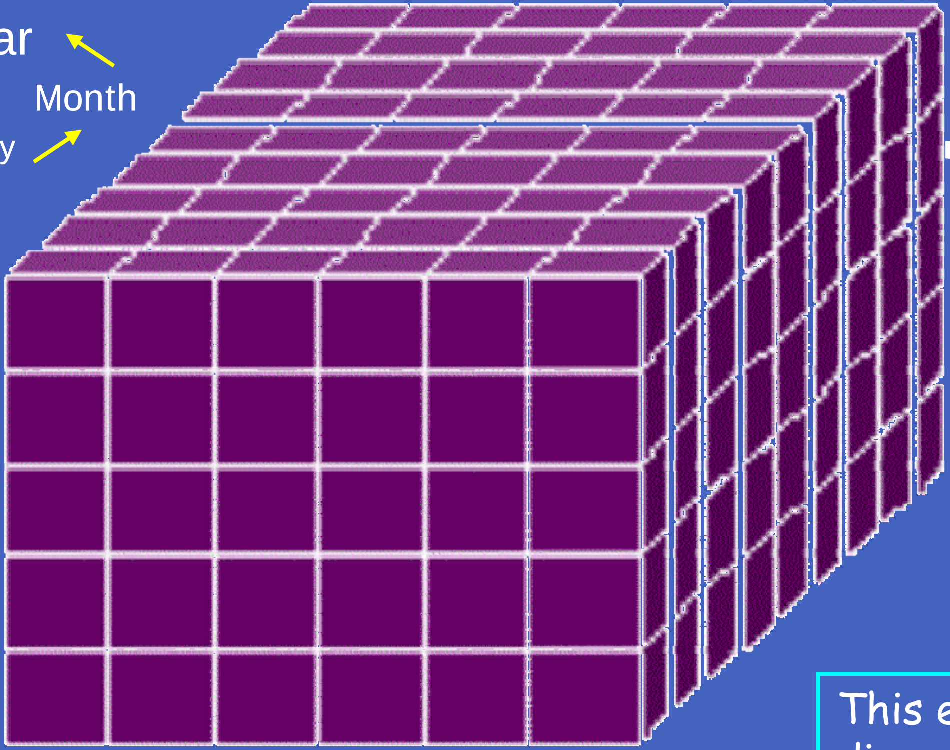
Period

Year

Week

Month

Day



Measures

- Reports received
- Population at risk
- Estimated Incidence rate
- SD of Incidence rate

Location

Country

Region

District

Detail

Minor Group

Major Group

Disease Classification (ICD)

This example has three dimensions (so that it can be visualised). In reality, for this application, we would need at least two more, Age and Gender.

Manipulation Functionality

- Store information with minimal aggregation
  - » Maximum detail in classifications
  - » Further aggregation (to less detail) on demand
    (may pre-compute for efficiency, may retain original records)
- Algebra for aggregation of classifications and measures is basically straight forward
- Aggregation of Measures (less detail)
  - » Everything based on summation can be regrouped
    (cf. updating algorithms, sufficient statistics)
  - » Some others, e.g Range
  - » Special issues for time: aggregate or cross sectional measures
- Derivations, across measures, cells, classifications, tables
- All aggregated tables are **proper** tables

Proper vs. Publication Tables

- Storage structure vs. formatting for presentation
- Proper tables – abstraction over which functionality is defined
  - Multidimensional structure (hyper-cube), all cells have same content
  - Based on aggregation or summarisation over a well-defined subset of data
  - Each case contributes once (through its weight)
  - Each dimension is an exhaustive classification (can have residual class)
  - Can have multiple measures in each cell, including derived measures
  - Don't (need to) store margins (totals), as these can be computed by further aggregation
  - Can create new proper tables by summarising or combining (compatible) existing ones
- Presentation tables – practical layout for using the information
  - Mapping and combination of proper tables to 2-dimensional page layout
  - Include margins, can abut classifications on dimensions
  - Further aggregation not usually possible without understanding the underlying proper tables

Publication Table from SPSS

				Source of Record					
				Main File			YP File		
				Type of sample			Type of sample		
				Main sample	Non-white sample	Total	Main sample	Youth sample	Total
Weight from YP File	Not Matched	YP Age	16-24	1	538	539	0	0	0
			Other Ages	37559	2961	40520	0	0	0
			Missing	48	12	60	0	0	0
			Total	37608	3511	41119	0	0	0
	Has matched Weight	YP Age	16-24	3346	0	3346	3349	2597	5946
			Other Ages	3	0	3	0	0	0
			Missing	0	0	0	0	1	1
			Total	3349	0	3349	3349	2598	5947
	Total	YP Age	16-24	3347	538	3885	3349	2597	5946
			Other Ages	37562	2961	40523	0	0	0
			Missing	48	12	60	0	1	1
			Total	40957	3511	44468	3349	2598	5947

Manipulation Functionality - for Processing

- Manipulation of Measures
  - Introduce measures from other tables with similar structure
  - Derive measures within cells
  - Not all combinations are meaningful
  - Complex Sampling Errors
- Choose appropriate levels in classification dimensions
  - E.g. regional or local rates, age groups, occupation of industry groups
- Combination of two tables
  - Do not always need to go back to original data to relate measures
    - Find common dimensions and classifications (may require some aggregation or mapping)
    - Choose one table as the detail table
    - Aggregate all non-common dimensions out of the 2nd table
    - Transfer measures from 2nd table, repeating values over missing classifications
- Meta-data to control validity of operations

Presentation Functionality

- Mapping from logical structure to presentation layout
  - » Rows, columns, pages (slices), margins
- Improper table combinations
  - » Combination of dissimilar dimensions e.g. Age groups by (SEG + Housing)
  - » Distinction between Classification and Measure is less important for presentation
- Medium
  - » Paper, Web, often with analysis (commentary)
  - » Machine readable (take away, not linked)
  - » Dynamic, for local or remote manipulation
- Associated material
  - » Generation of descriptions, footnotes, indexes, content lists
  - » Dynamic links to further or related results and metadata

Manipulation Functionality - for Exploration

- Dynamic viewing, linked to source aggregations
- Selection
  - » Subset of classification cells, and of measures
- Dynamic regrouping
  - » Roll up to combine existing groups to next level
  - » Drill down to get more detail in groups at lower level
  - » Operate independently, i.e. not all parts of a classification at the same level
  - » User-defined groupings
- All derivation and presentation facilities
- Specialist browsers, available for local data or over the Internet

Software

- Lots of software for survey processing
  - » Some good
  - » Good support for questionnaire design
  - » OK for basic statistical analysis, report production
  - » Weak on automation of processes
- Production systems
  - » Have been seen as database problems
  - » Inadequate analysis of statistical structures and processes
  - » Failure to see manipulations as generic processes
  - » Inadequate flexibility in making results and aggregate data available for further use
- Dissemination
  - » Various developments, Beyond 20/20, Nesstar, Super-*
  - » Need CMS, but with specialised functionality
- Metadata
  - » Nothing suitable for production use, but improving with tools for DDI and SDMX

Summary

- Much scope for recognising the generic nature of various statistical processes
  - » Particularly manipulations of aggregate data
  - » Judgement essential to interpretation of results
- Still lots of details to get right
- Need to see dissemination of quality data resources as the main function of most statistical data production systems
- Importance of metadata
  - » Integrate capture into the design and production process
  - » Use to help automate processes, particularly analysis
  - » Resource for dissemination and discovery
- Disclosure control – hard problem
- Need skills to communicate requirements to system developers