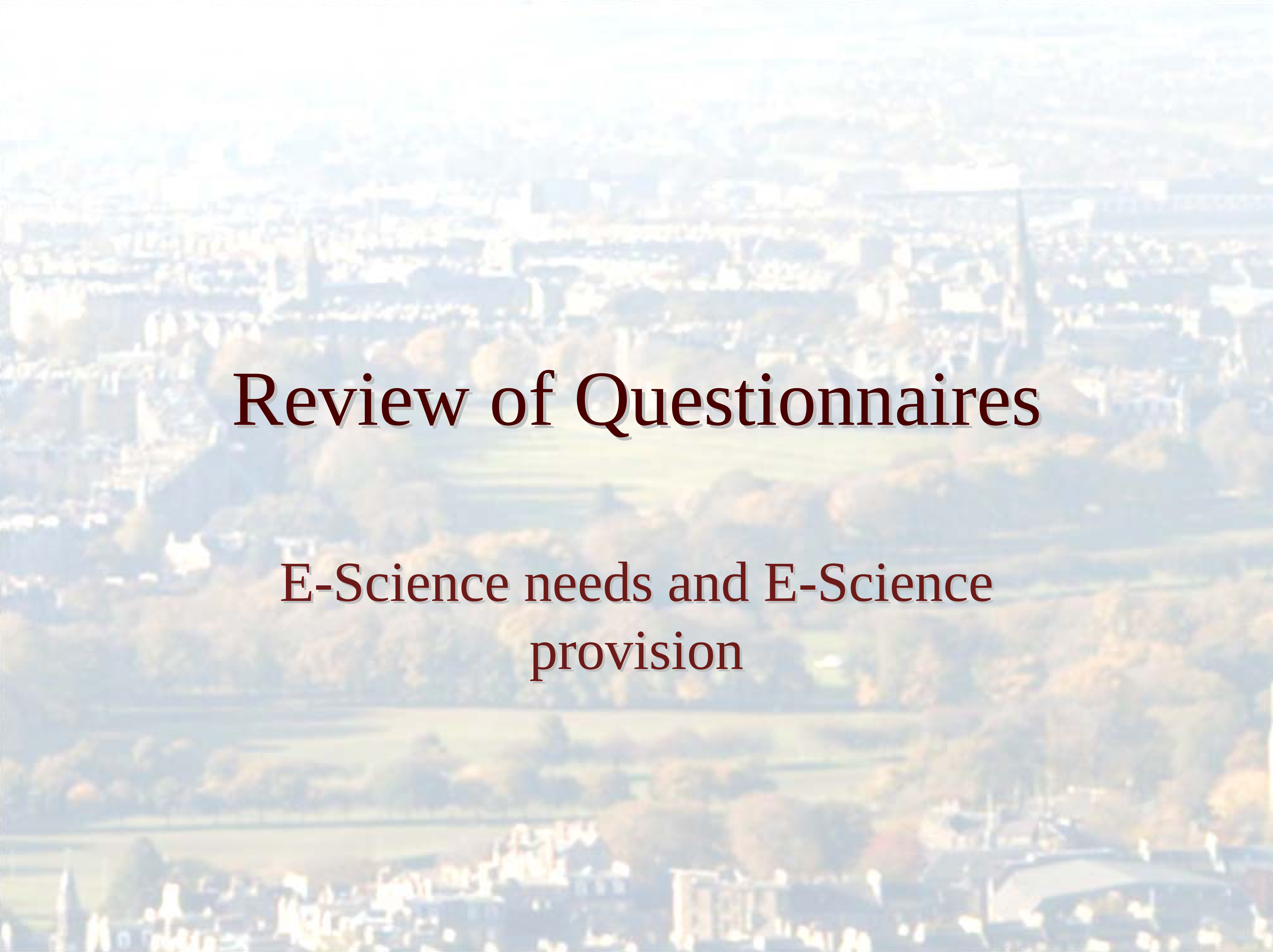
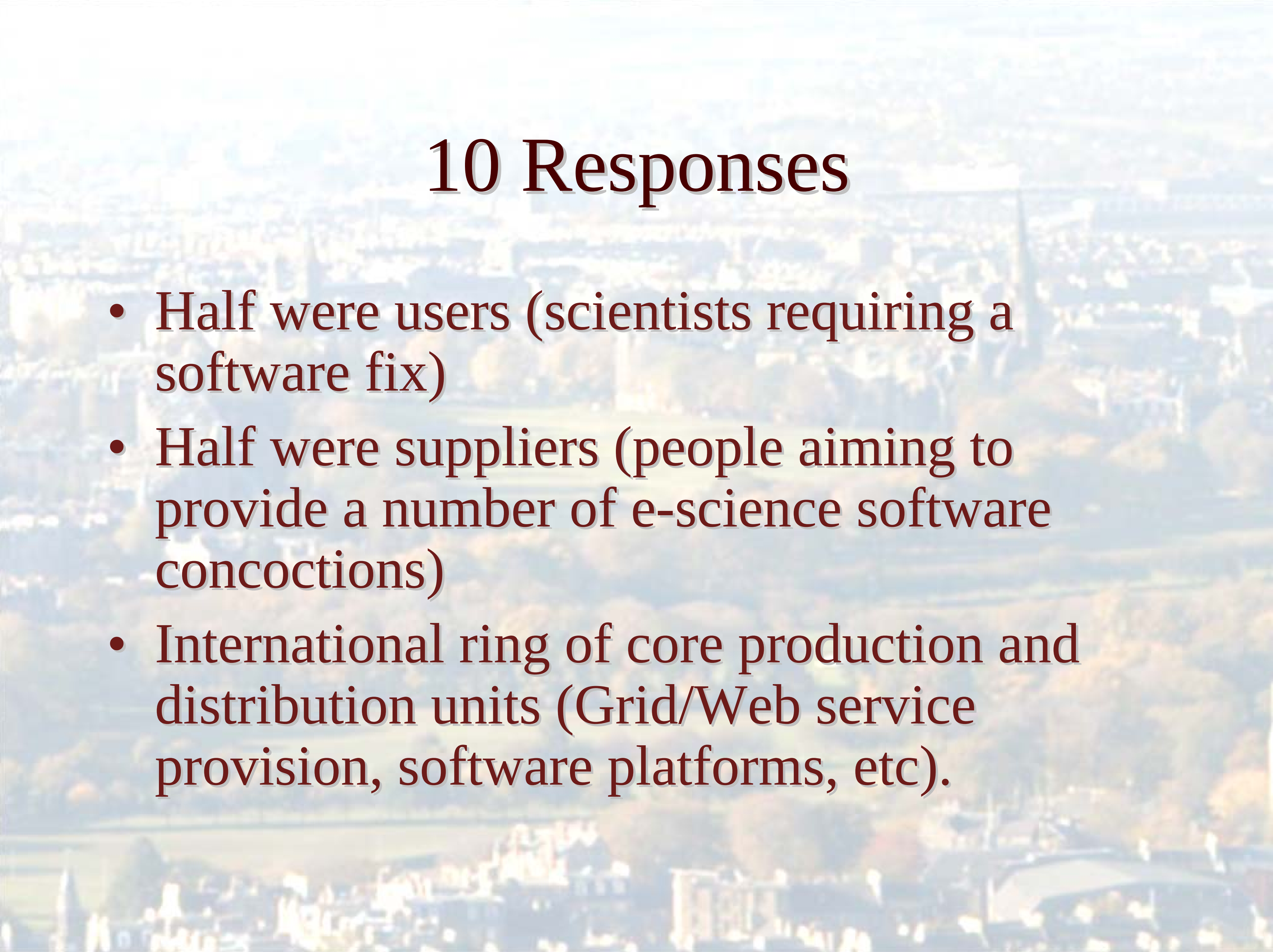


Review of Questionnaires

E-Science needs and E-Science provision

The Questionnaire

- Issues
  - Project drivers
  - Data mining drivers and importance
  - Data mining methods
  - Data types, sizes and format
  - Software and Platform

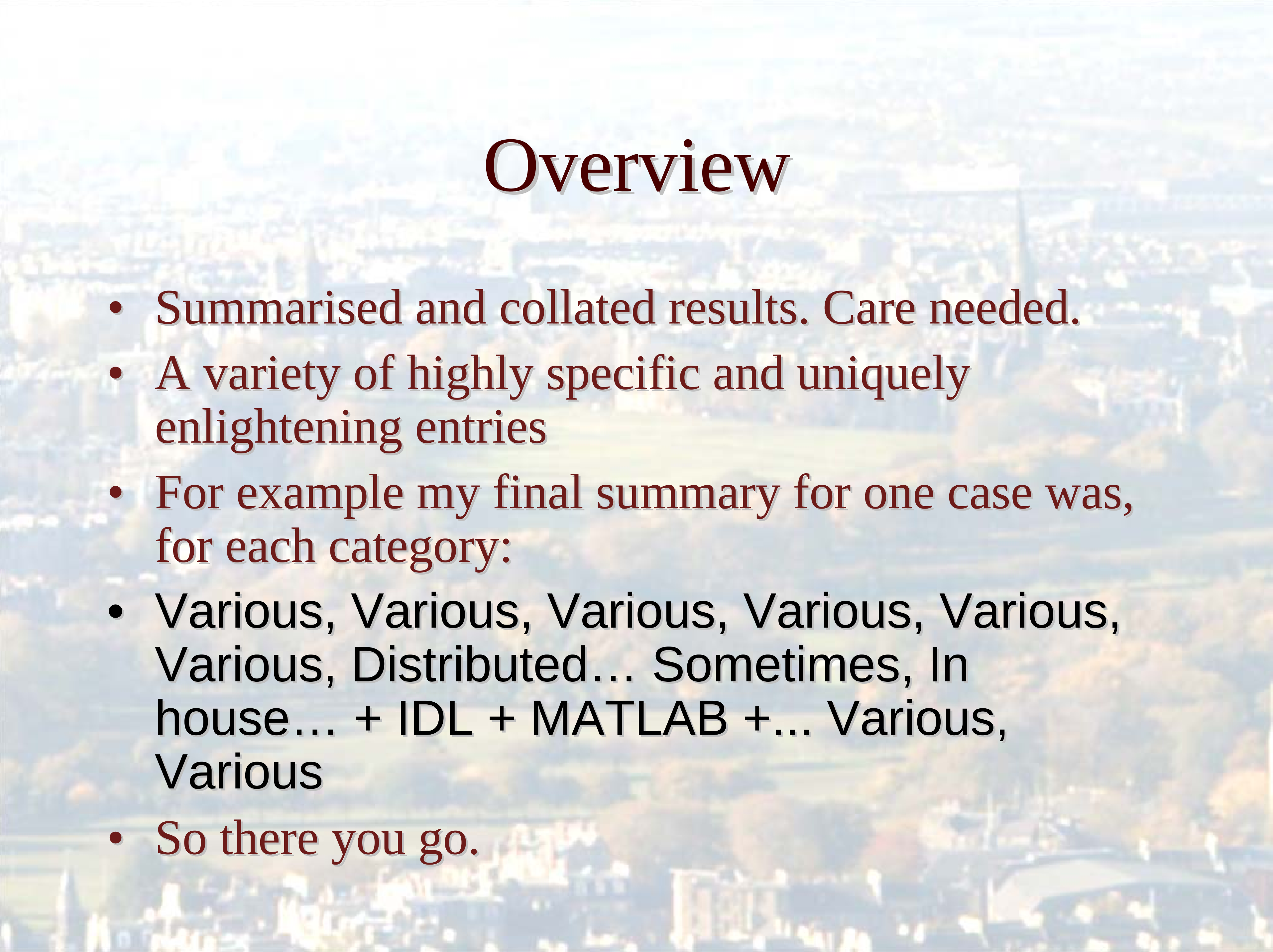


10 Responses

- Half were users (scientists requiring a software fix)
- Half were suppliers (people aiming to provide a number of e-science software concoctions)
- International ring of core production and distribution units (Grid/Web service provision, software platforms, etc).

Why?

- This is all about E-Science right?
- What E do scientists actually need.
- We need to make sure that the developments made are driven by real markets.
- Two possible serious risks:
  - Have user requirements which are not matched by technology.
  - Have a large technology infrastructure which does not match users simpler or specific requirements

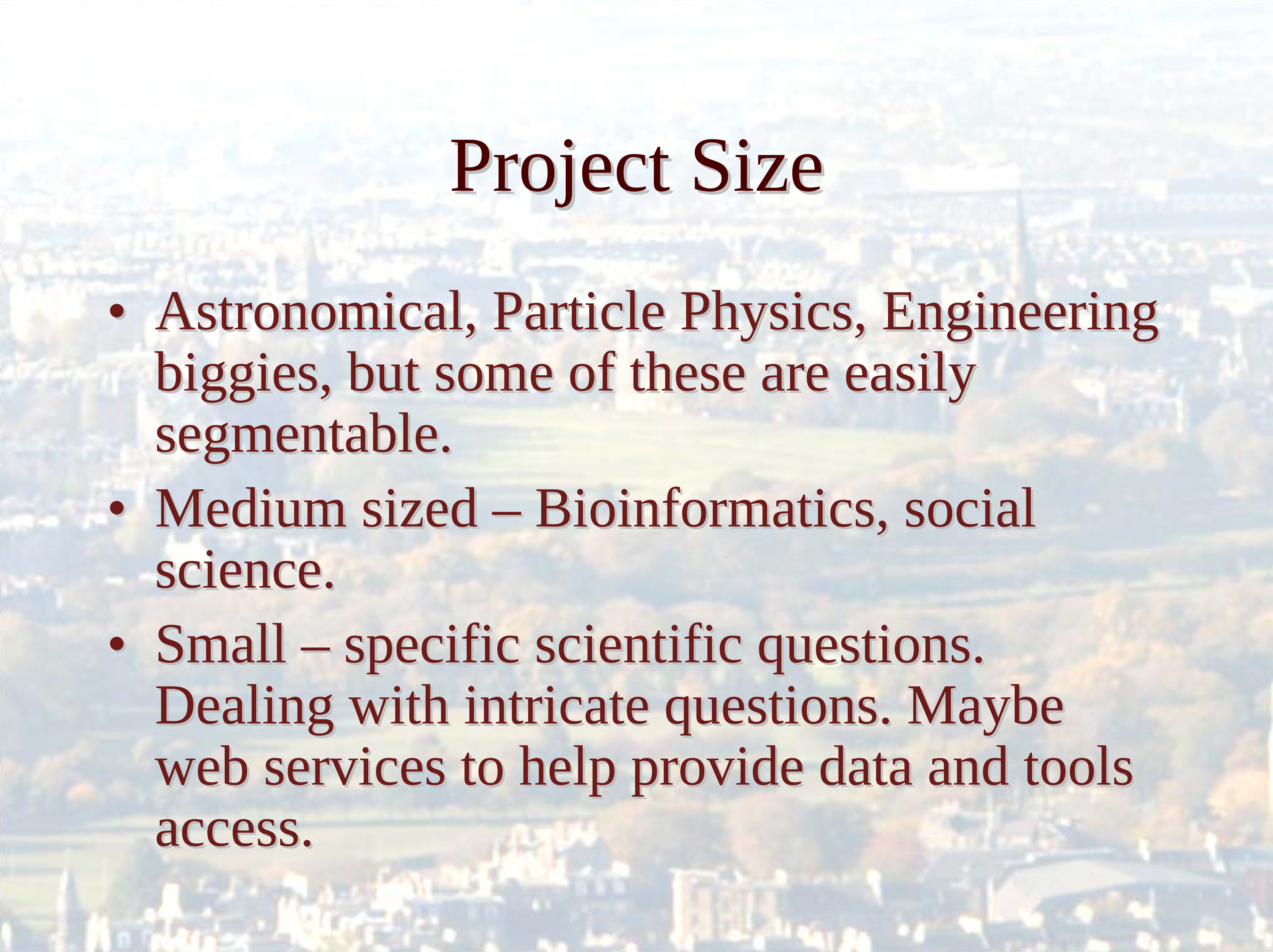


Overview

- Summarised and collated results. Care needed.
- A variety of highly specific and uniquely enlightening entries
- For example my final summary for one case was, for each category:
- Various, Various, Various, Various, Various, Various, Distributed… Sometimes, In house… + IDL + MATLAB +... Various, Various
- So there you go.

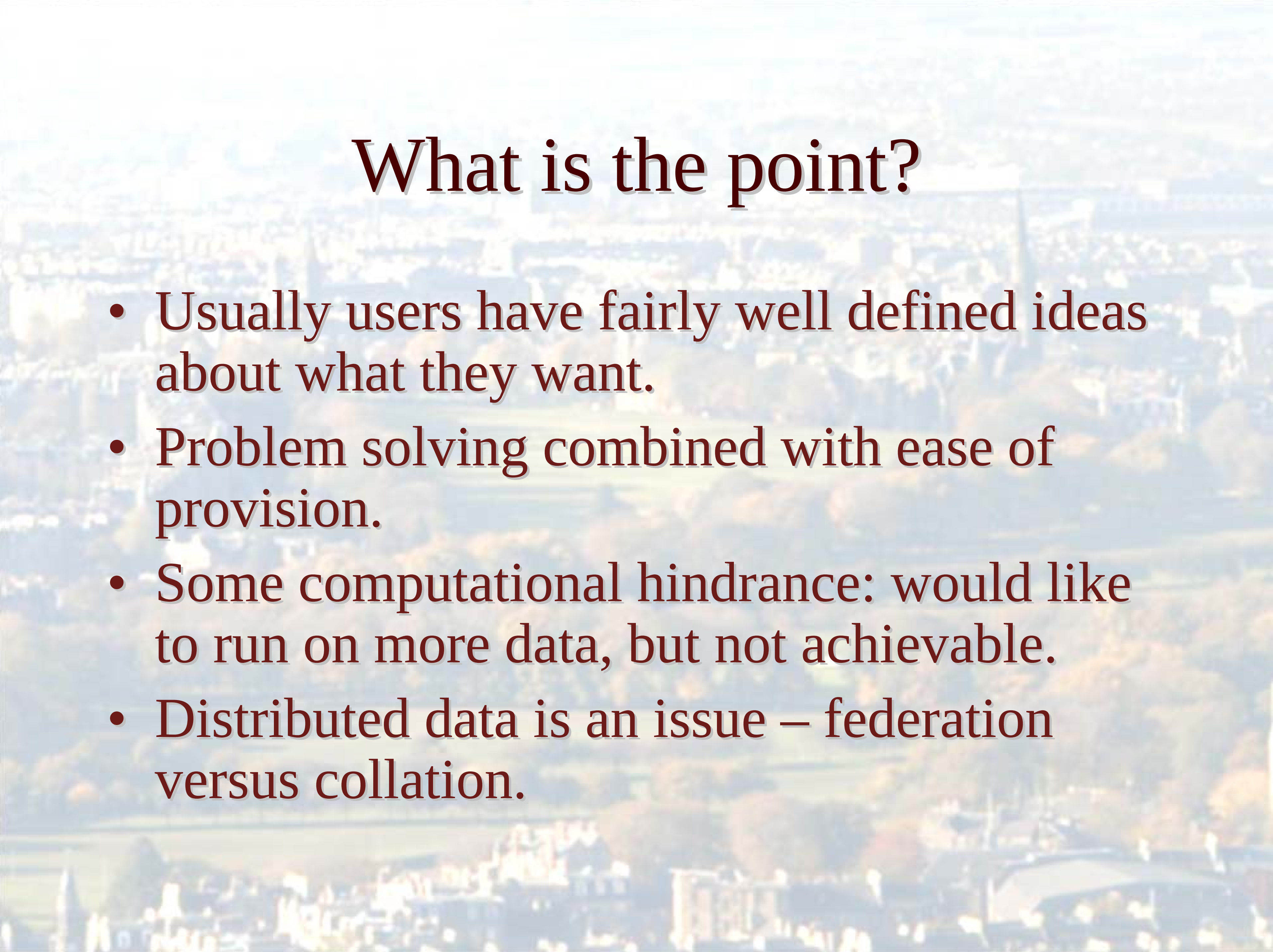
What can we really say

- Pinch of salt – small sample size BUT
- Difference between the providers and the users.
- Some big projects out there with large data sizes (>50GB), but many projects are small or medium in data size.
- Data integration a recurring theme
- In most cases we are talking about standard data mining tools.
- A lot of in-house software development.



Project Size

- Astronomical, Particle Physics, Engineering biggies, but some of these are easily segmentable.
- Medium sized – Bioinformatics, social science.
- Small – specific scientific questions. Dealing with intricate questions. Maybe web services to help provide data and tools access.



What is the point?

- Usually users have fairly well defined ideas about what they want.
- Problem solving combined with ease of provision.
- Some computational hindrance: would like to run on more data, but not achievable.
- Distributed data is an issue – federation versus collation.

Methods

- Standard Stuff:
  - Decision trees
  - Association rules
  - Neural Networks
  - K nearest neighbours
  - Clustering
  - Case based reasoning
- Selection Bias

Data Mining Drivers

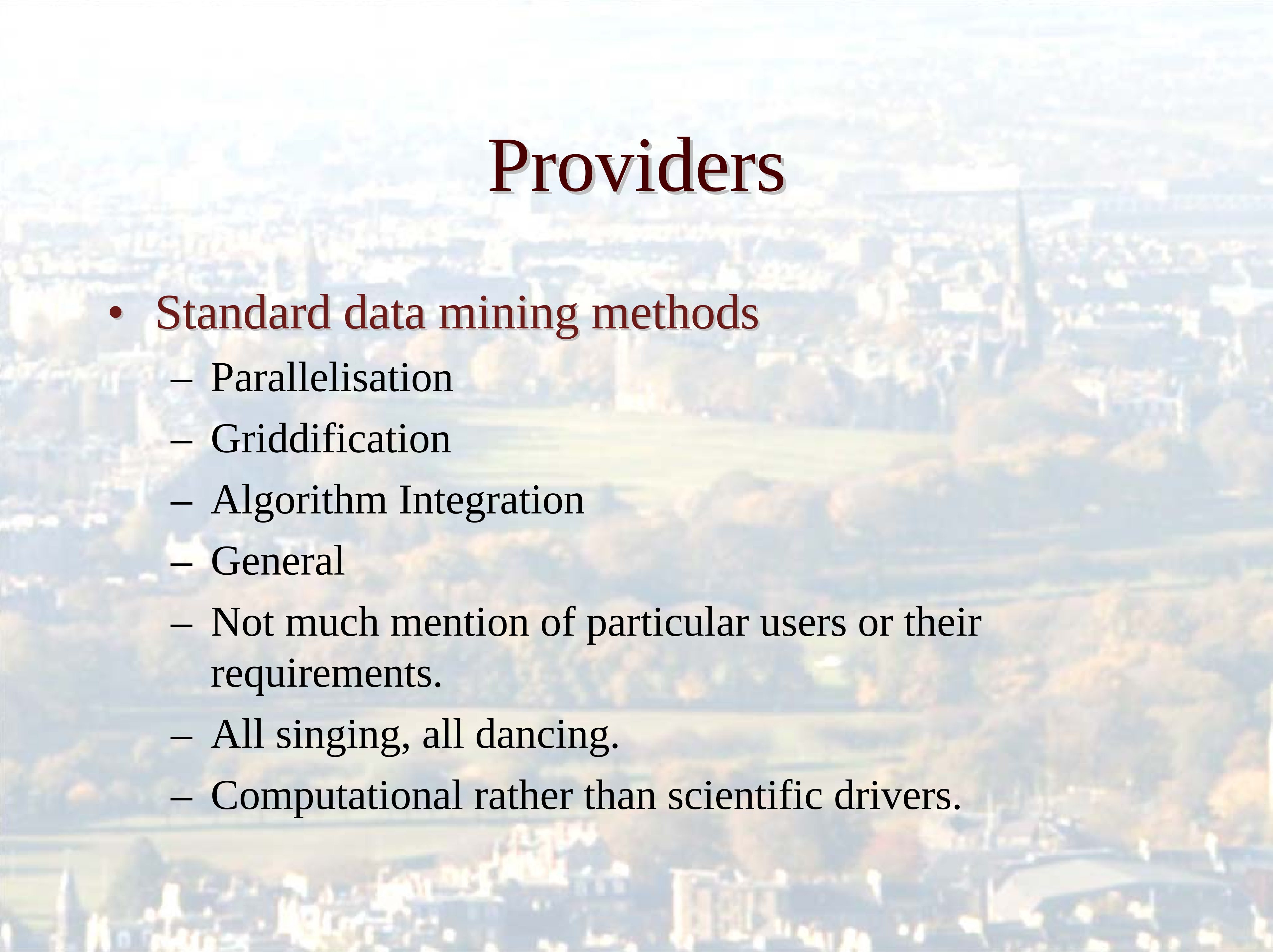
- Data Integration is a very common issue
  - Integration of different types of sources
  - Integrating sources from different locations
  - Data linkage – relationship discovery
- Helping with scientific inference
- Discovery of diagnostics

Type format location

- All sorts.
- Text, numeric, time series, hierarchical.
- XML, flat file, relational databases. Various.
- Often distributed.
- Some standalone. Grid services, Web services. Cluster approaches.

Software

- Lots of in-house
  - University bias – software development part of project deliverables.
  - Even so, suggests room for development.
- Other mentions: MATLAB, IDL, Weka, C5, Java, various open source.



Providers

- Standard data mining methods
  - Parallelisation
  - Griddification
  - Algorithm Integration
  - General
  - Not much mention of particular users or their requirements.
  - All singing, all dancing.
  - Computational rather than scientific drivers.

Summary

- Lots of people doing different things
- Users have very specific requirements
  - Can we provide generic tools to cover them, or do we need to engage with specialisms
- Size not the big issue.
- Data integration, accessibility of data and methods, dealing with distributed data are.
- Variety of work providing standard data mining tools in a more parallel or griddy way.
- Plenty of room for discussion.