The Digicene: the Age of Big Data in the Geosciences

Lee Allison
Arizona Geological Survey
USGIN Foundation, Inc.
A system that works for one geologist, in the field and at their desk...

...and for users of streaming, High Performance Computing
~20% of earth scientists use/need HPC capabilities; 80% rely dominantly on desktop software and data sets they collect themselves.
Big Data: Data Integration ("interoperability")

Capabilities for moving data between data warehousing, business analytics, master data management, enterprise applications, and custom applications.

Integrating structured data in relational databases with social media data, weblogs, and various unstructured data.
“Big Data” is not just large data sets

Digitizing, organizing, analyzing, modeling, visualizing [small] data from disparate sources
National Geothermal Data System

Free online access to:
Maps, data, & documents from 65+ providers nationwide

• >10 million data records
• >3 million oil & gas wells
• >47,000 maps & reports

Distributed network

Powered by USGIN
For mainstream science

Not *centralized* “big iron,” but *decentralized data wrangling*

Not “one ring to rule them all” but “small pieces loosely joined”

**Mass democratisation** of the means of access, storage & processing of data

A *distributed* ecosystem of information, an ecosystem of *small data*

Distributed models not centralized ones

Collaboration not control, and *small data not big data*
Barriers to discovery, access, and integration of data have shaped the scientific practice since its beginnings.

Most of us work in small data because there is no viable alternative.

Tackling interdisciplinary/multidisciplinary problems was not realistic.

Many large geoscience databases are not really ‘big data’.

Open data access and interoperability are changing the paradigm.

Digitization, integration, and modeling are creating unprecedented opportunities and challenges.

The new generation of geoscientists increasingly will need data science skills.
Turning legacy data....
....into this
Challenges to Big Data in the Geosciences
- technical and social

- Creating common standards and protocols
- Engaging the vast number of distributed data resources
- Establishing practices for recognition of and respect for intellectual property
- Developing simple data and resource discovery and access systems
- Building mechanisms to encourage development of web service tools and workflows for data analysis
- Brokering the diverse disciplinary service buses
- Creating sustainable business models for maintenance and evolution of information resources
- Integrating the data management life-cycle into the practice of science
There will be a shortage of talent necessary for organizations to take advantage of big data. By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions.

McKinsey Global Institute, 2011
Global Cyberinfrastructure and Data Management Initiatives

- Environment
- Earth Observation
- Geoscience
- Marine
- Petroleum/Energy
- All

FOCUS
- National/Continental
- Cross Continental
- Global

Geoscience Information Council
- Belmont e-Infrastructure
- GEO/GEOSS
- OneGeology

EarthCube (US only)
- Earth Science Information Partners (ESIP) (US)

GeoSeas (EU)

Renewable Energy Agency US

Inspire (EU only)

IGSN (EU, US, AUS)

ODIP (US, EU, AUS)

ICSU World data System

Energistics

CODATA

National Data Repositories

Research Data Alliance

ICSU World data System

ONE Geology

OneGeology agenda

Ones GA follows

Ones GA is actively involved in

Ones mentioned in OneGeology agenda

Ones not on OneGeology Agenda

USGIN involvement

Geoscience Australia
Geoscience Cyberinfrastructure
Opportunities

Everything digital, online, & interoperable

Productivity increase

Enabling inter- and multidisciplinary research

Enabling unheard-of analytical, computational, visualization, and modeling
Challenges

Digitization
Open access
Integration/Interoperability
Modeling, visualization, analytical capacity
Capacity building
Social construct
GLOBAL CONVERGENCE IN GEOSCIENCE CYBERINFRASTRUCTURE

DAWN OF THE DIGICENE