

# A Data Driven Research World – and ANDS Perspective

Ross Wilkinson

## The outline...

- Research is changing - data matters more
- Its used differently – no single lifecycle
- There are lots of services that might be relevant (some from ARCS and ANDS)
- ***How do we match services to needs and data lifecycles?***

## Role of data

- With more data online, more can be done
- Possible now to answer questions unrelated to reasons why data was collected originally
- Increasing focus on problems across disciplinary boundaries

...and researchers should care..

# Australian Code for the Responsible Conduct of Research

- It describes the responsibilities of institutions and researchers in the management of research data and primary materials
- Institutions are to retain research data, provide secure data storage, identify ownership, and ensure security and confidentiality of research data
- Researchers are to retain research data and primary materials, manage storage of research data and primary materials, maintain confidentiality of research data and primary materials

*[http://www.nhmrc.gov.au/publications/synopses/\\_files/r39.pdf](http://www.nhmrc.gov.au/publications/synopses/_files/r39.pdf)*

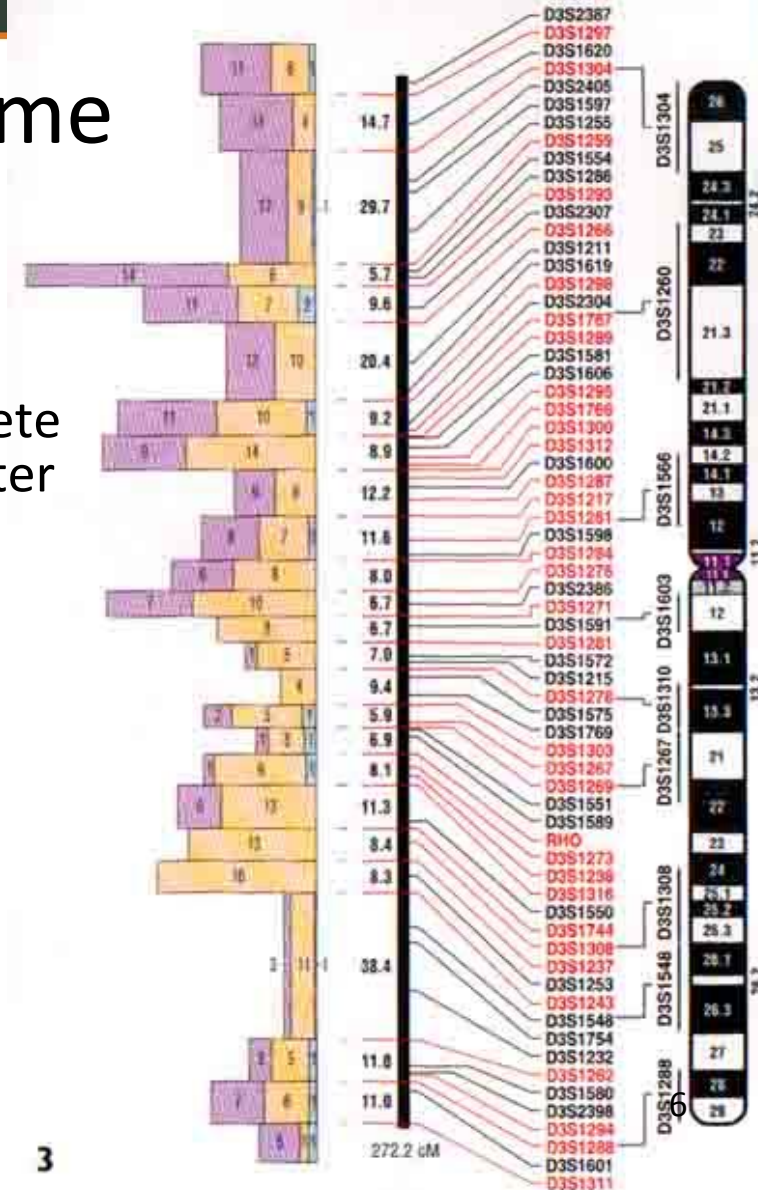
## Role of data citation – fame!

### Sharing Detailed Research Data Is Associated with Increased Citation Rate

- 48% of 85 cancer microarray clinical trial publications with publicly available microarray data received 85% of the aggregate citations
- *Piwowar HA, Day RS, Fridsma DB (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2(3): e308. doi:10.1371/journal.pone.0000308*

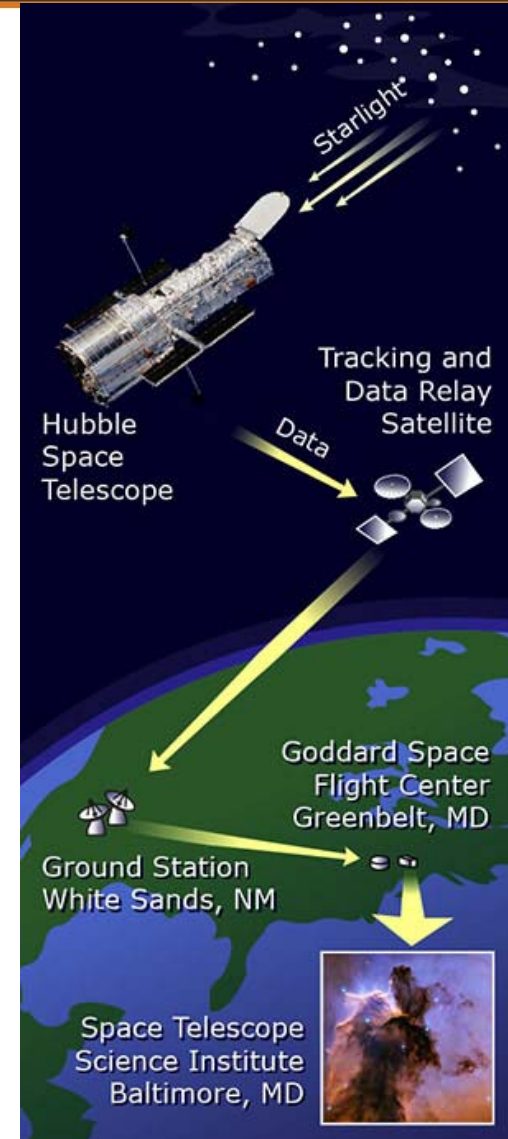
# Mapping the Human Genome

- Took a large team of scientists 10 years to map the 30,000 genes that describe the human body
- In 2007, Craig Venter, published his complete DNA sequence, unveiling the six-billion-letter genome of a single individual for the first time
- The work required a large team using new instruments to produce a large dataset – indeed 2 competing large teams!
- No single lab could have completed this project with available technology in a reasonable time



# The Hubble Telescope

- The Hubble telescope launched in 1990
- Increasing focus on cross-disciplinary science
- Observations are proposed, and if accepted, data is collected and made available to the proposers – who then write a research paper
- Each year around 1,000 proposals are reviewed and approximately 200 are selected, for a total of 20,000 individual observations
- The data is stored at the Space Telescope Science Institute
- There are more research papers written by “second use” of the research data, than by the use initially proposed



# Changing Data, Changing Research

## New scientific instruments

- Large Hadron Collider at CERN generates 1.5 gigabytes of data per second
- the Square Kilometre Array (1 EB/day!)

## New scientific Models

- The mapping of the Human Genome: A billion DNA letters in a human sequence
- Global climate models

## New teams

- 195 scientists mapped the genome of the fruit-fly

## New knowledge from unlocked data

- Most research from Hubble telescope data was not “first use”
- Common data sets unlocked the power of search technology – TREC

## Nature of Research

- Question -> Hypothesis -> Data -> Analysis -> Paper
- Problem -> Data -> Explore -> Publish -> Explore again
- Research Outputs: Papers, Data, Models, Artefacts

## Research Lifecycle for data intensive research

- *Conceive question*
- *Get grant*
- Conceive data
- Capture data
- Store data
- Manage data
- *Analyse data*
- *Publish results*
- Share data