

# Data Curation: Skill-sets and Workforce Needs

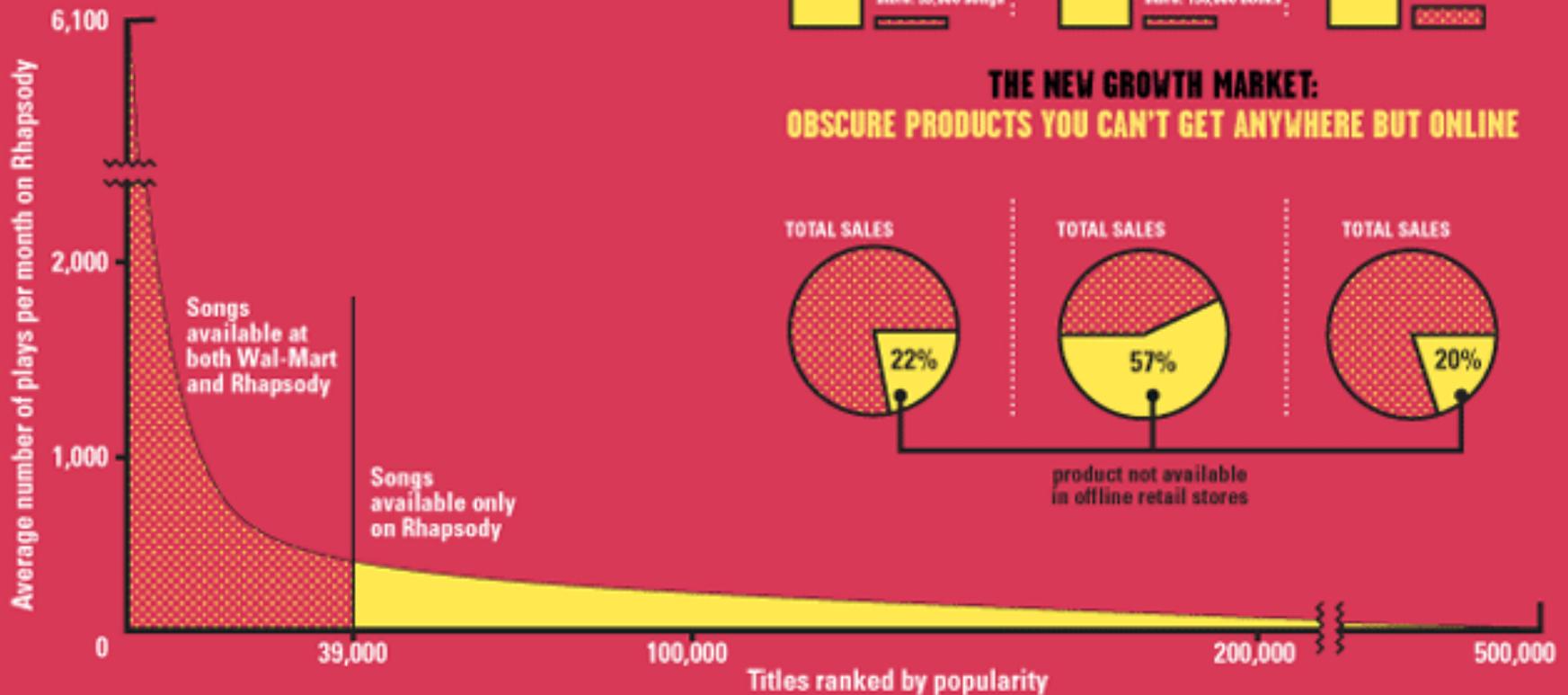
**David E. Schindel, Executive Secretary**

National Museum of Natural History  
Smithsonian Institution

[SchindelD@si.edu](mailto:SchindelD@si.edu);  
<http://www.barcoding.si.edu>  
202/633-0812; fax 202/633-2938

# ANATOMY OF THE LONG TAIL

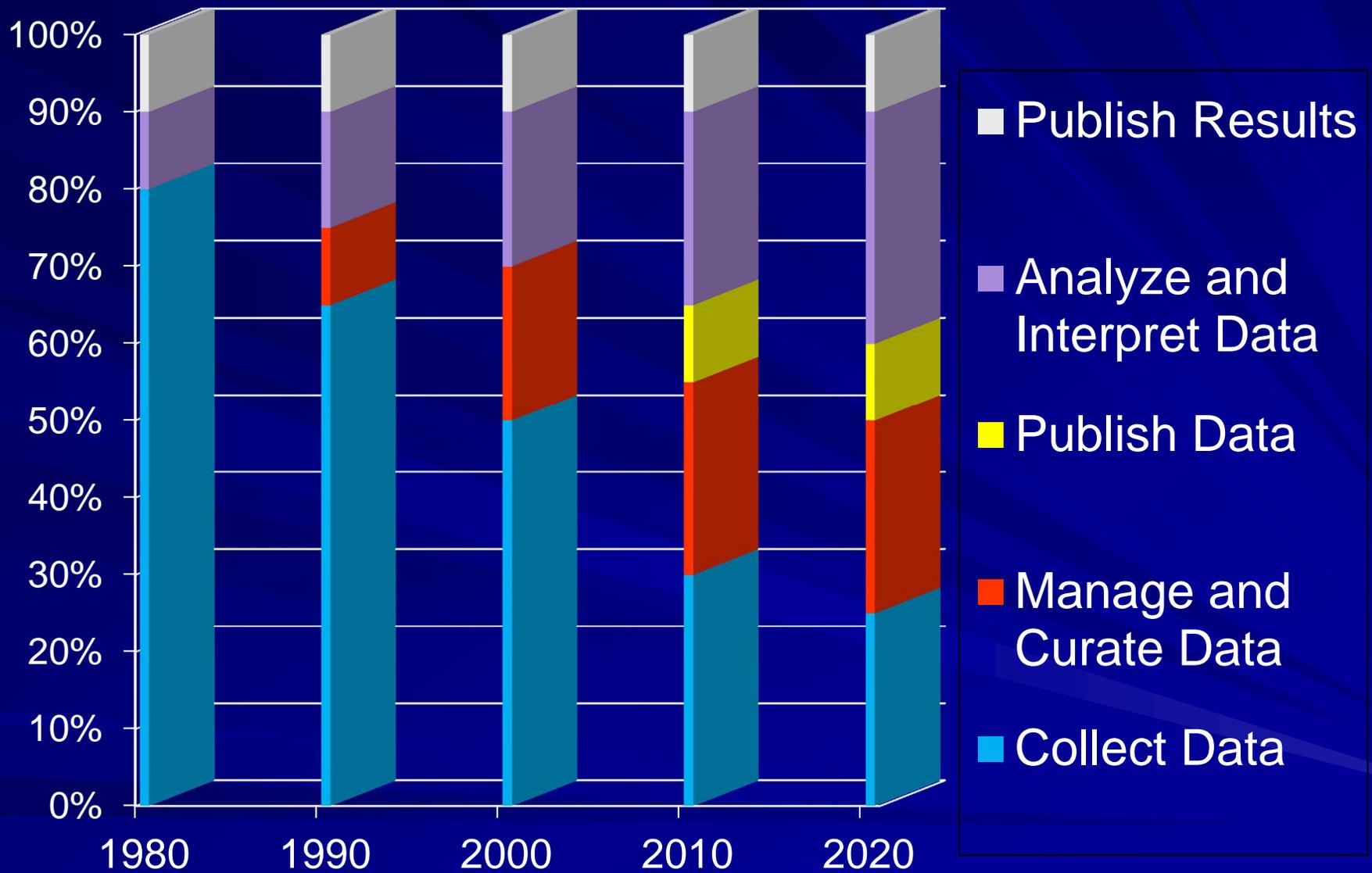
Online services carry far more inventory than traditional retailers. Rhapsody, for example, offers 19 times as many songs as Wal-Mart's stock of 39,000 tunes. The appetite for Rhapsody's more obscure tunes (charted below in yellow) makes up the so-called Long Tail. Meanwhile, even as consumers flock to mainstream books, music, and films (right), there is real demand for niche fare found only online.



Chris Anderson, 2004, "The Long Tail:", Wired Magazine

# Main Points

- Not all Big Data are Born Big
- Data Curation is a real, pervasive, emerging need
  - Not yet a designated occupation title
  - Some full-time practitioners
  - Near-universally required skill for researchers
  - Increasingly a necessary life-skill for all
- Examples from biodiversity research community



# Data Curation as a Career

- McKinsey report characterization:
  - Deep Analytical Talent
  - Big Data Savviness
  - Supporting Technology
- Not yet a recognized ‘occupation’
  - Digital librarian
  - Database manager
  - Information specialist
  - Data analyst
  - Computer programmer

# Data Curation in Research

- Sometimes by people, sometimes automated
- Data preservation, longevity, migration to new media
- Error-trapping for consistency, accuracy
- Metadata and documentation
- Compliance with standards
- Interoperability with related databases
- Web exposure, web services

# EU Survey of Data Curators

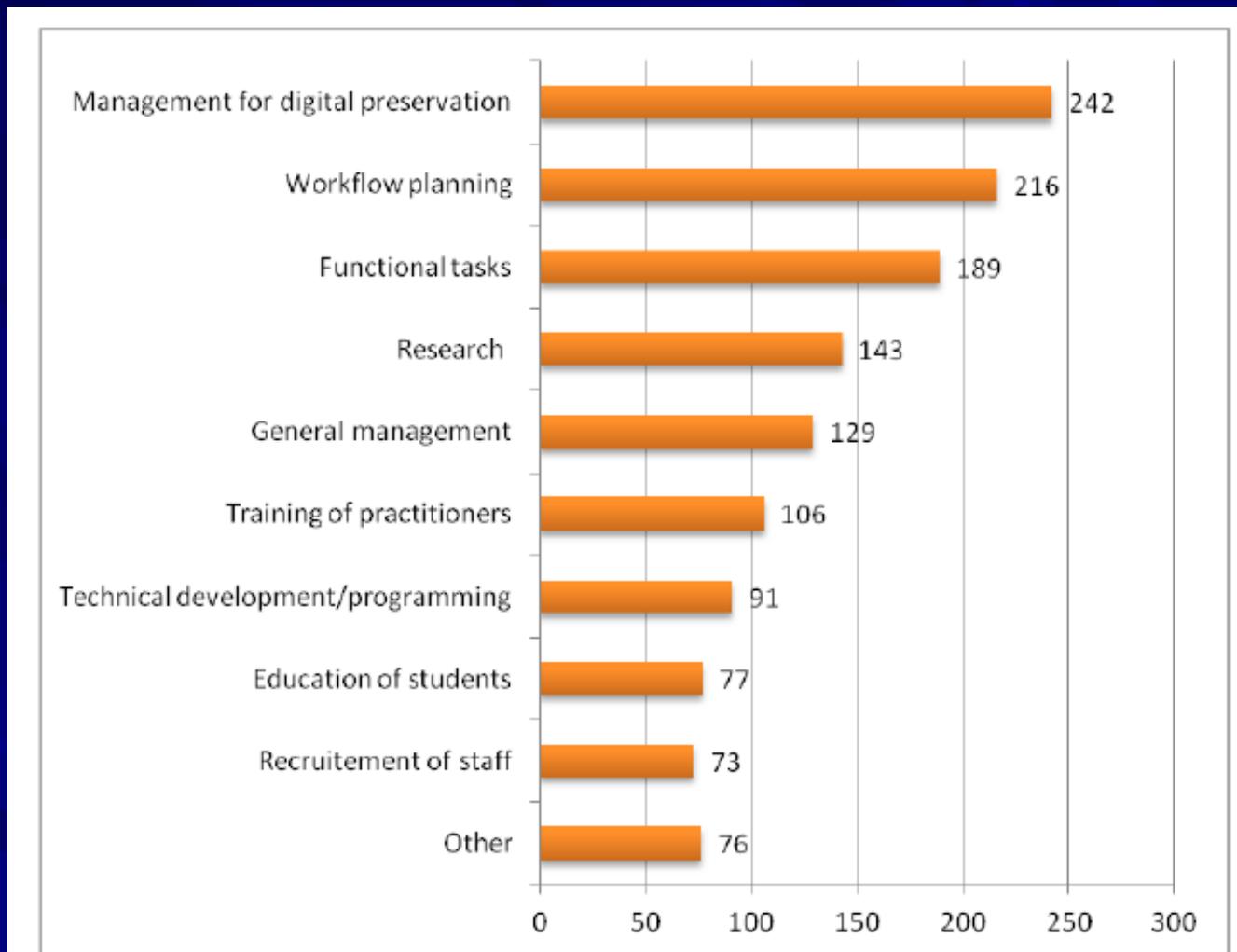


Fig. 4: Tasks the respondents are responsible for

# Distribution of Needs

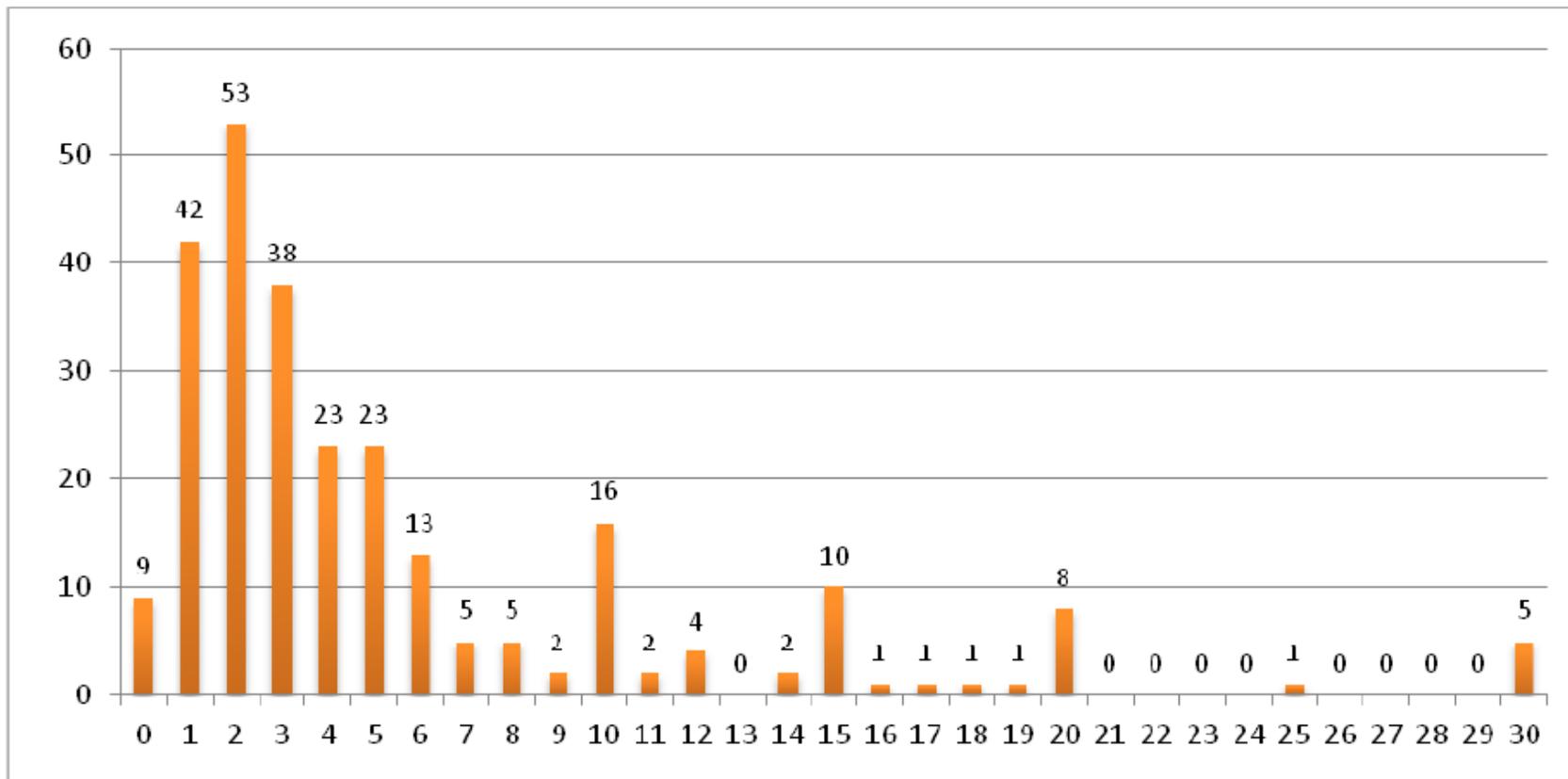


Fig. 7: Number of digital preservation staff

# The Rising Data Tide

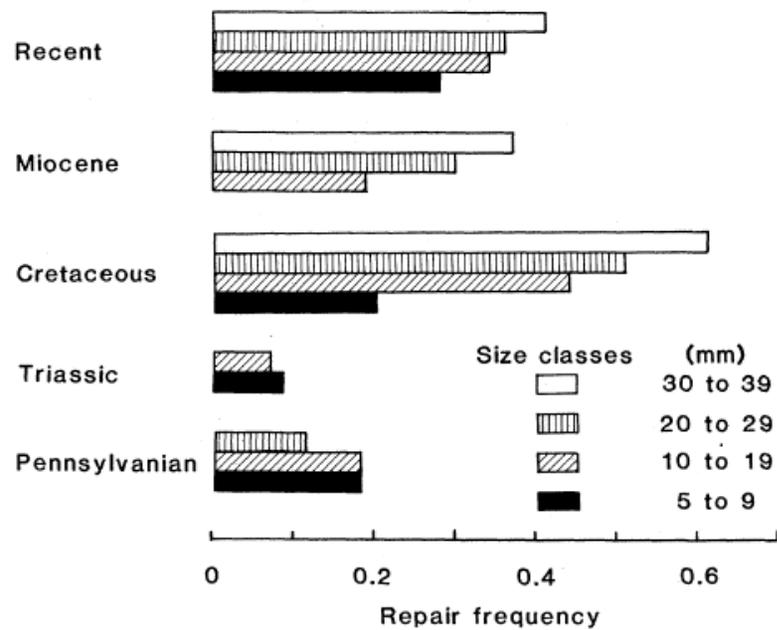
## The Era

- 1980: Me and my microscope: Access-limited
- 1990: Email, microcomputer: Network-limited
- 2000: Web-based data, sensors: Provider-limited
- 2010: Aggregator portals: Dialect-limited

## The Needed Skills

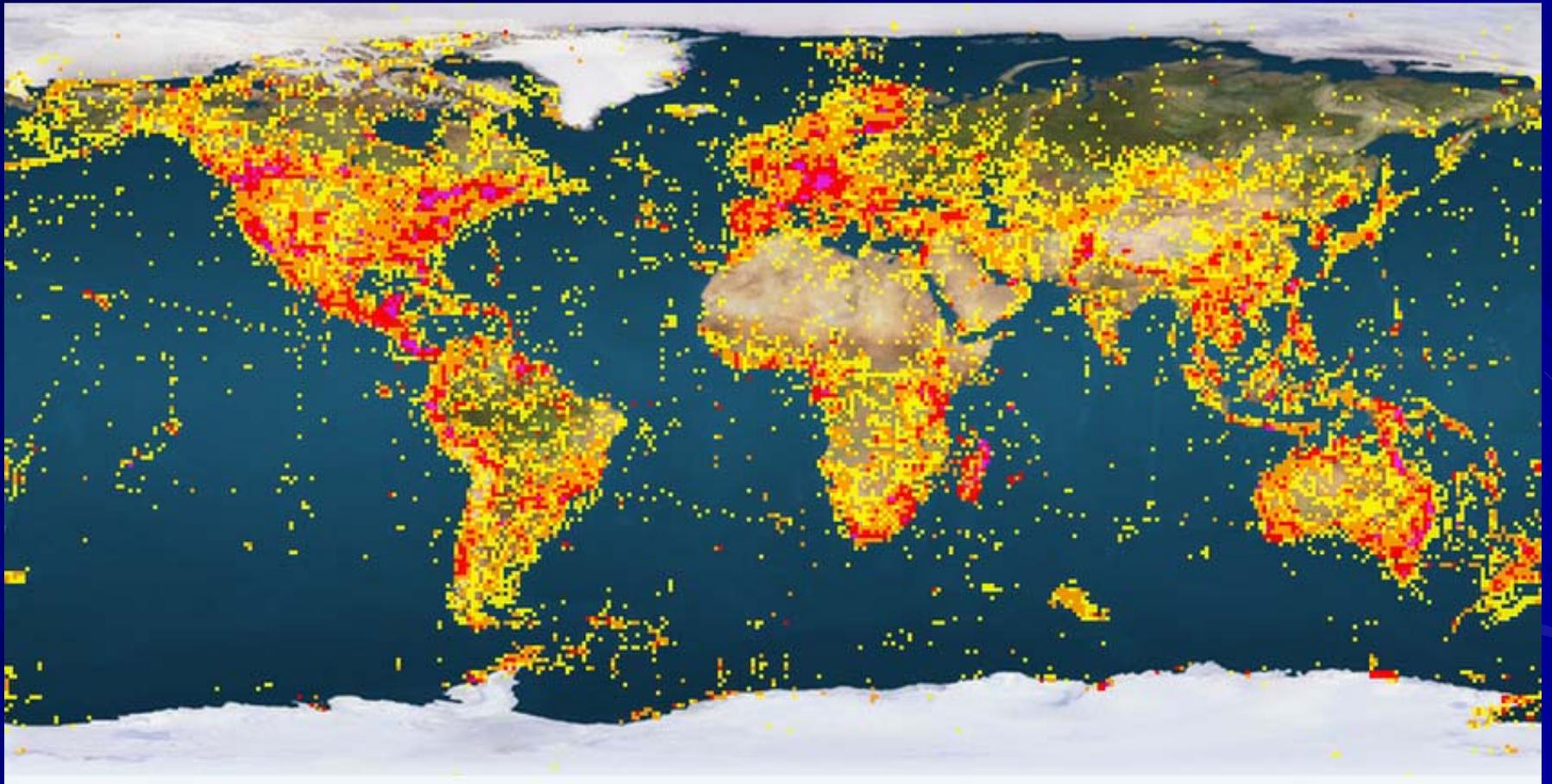
- Domain knowledge, setting priorities, time management
- Basic computer skills, social networking
- Hardware system design, data capture, web interfaces, negotiating data release
- Data integration and re-use, negotiating partnerships

# 1980



Assemblage	Shell repair in							
	5- to 9-mm class		10- to 19-mm class		20- to 29-mm class		Complete sample	
	<i>N</i>	<i>F</i>	<i>N</i>	<i>F</i>	<i>N</i>	<i>F</i>	<i>N</i>	<i>F</i>
Recent, all species	3	.40	26	.27	15	.33	53	.28
Venado Beach, Panama			8	.37	7	.44	18	.37
Wom, Papua New Guinea			6	.12			13	.08
Pujada Bay, Mindanao			6	.31			12	.25
Dodinga Bay, Halmahera			5	.38			7	.44
Tumon Bay, Guam					4	.27	9	.29
Miocene, all species			16	.20	14	.34	19	.40
Gatun 1			5	.14	10	.33	17	.33
Gatun 2			7	.18	10	.33	12	.36
Gatun 3							9	.19
Cretaceous, Ripley (14 sites)	7	.20	7	.41	3	.50	24	.35
Triassic, St. Cassian Group	10	.08	7	.06			11	.08
Pennsylvanian, all species	8	.15	15	.11	4	.08	16	.13
Grindstone Creek	5	.21	7	.05			12	.17
Wolf Mountain			5	.14			5	.11
Colony Creek 1							5	.09
Colony Creek 2	5	.11					8	.09
Finis 1			6	.07			7	.06
Finis 2			5	0			7	.04
Wayland			6	.32			7	.31

# 2012: Aggregated Records

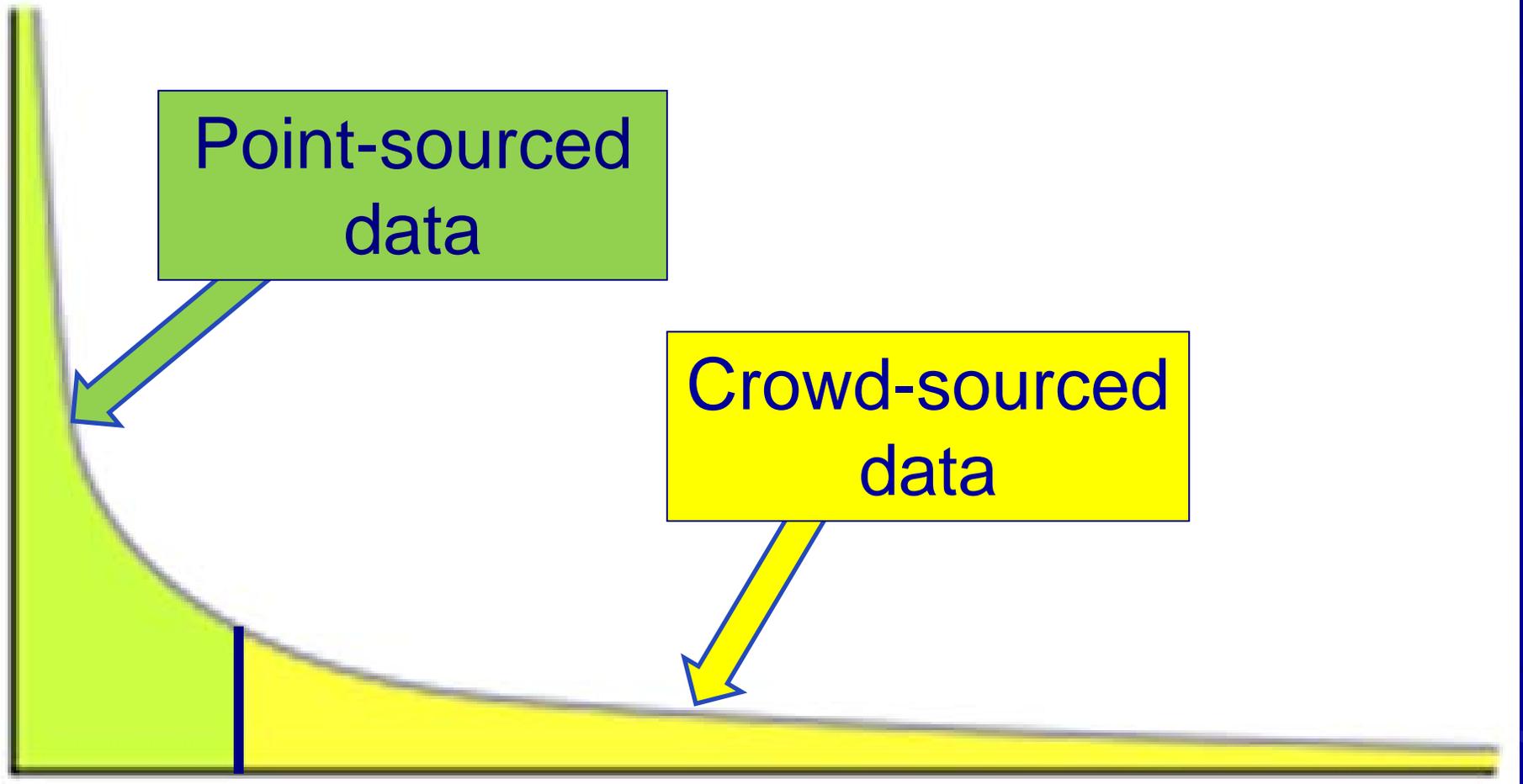


# Sources of Big Data in Biology

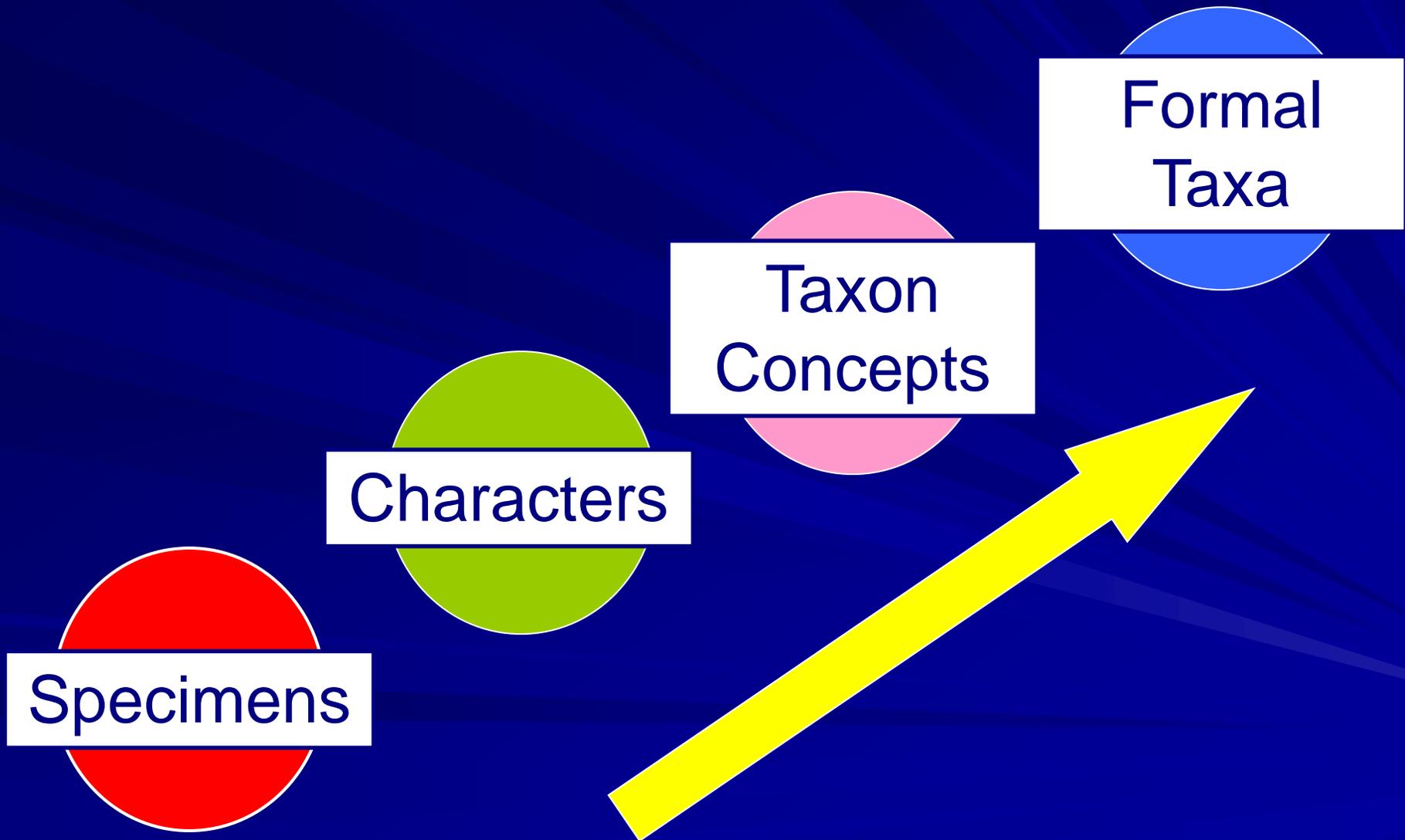
- New high-volume sensors
  - Next-Generation Sequencing
  - Sensors for environmental monitoring
  - Remote sensing
- Data mobilization
  - Digitization of analog records (NSF ADBC)
  - Networking of databases (e.g., VertNet)
  - Rise of Aggregators (GBIF, EOL)
- Data archiving (DataOne, data.gov)

Point-sourced  
data

Crowd-sourced  
data

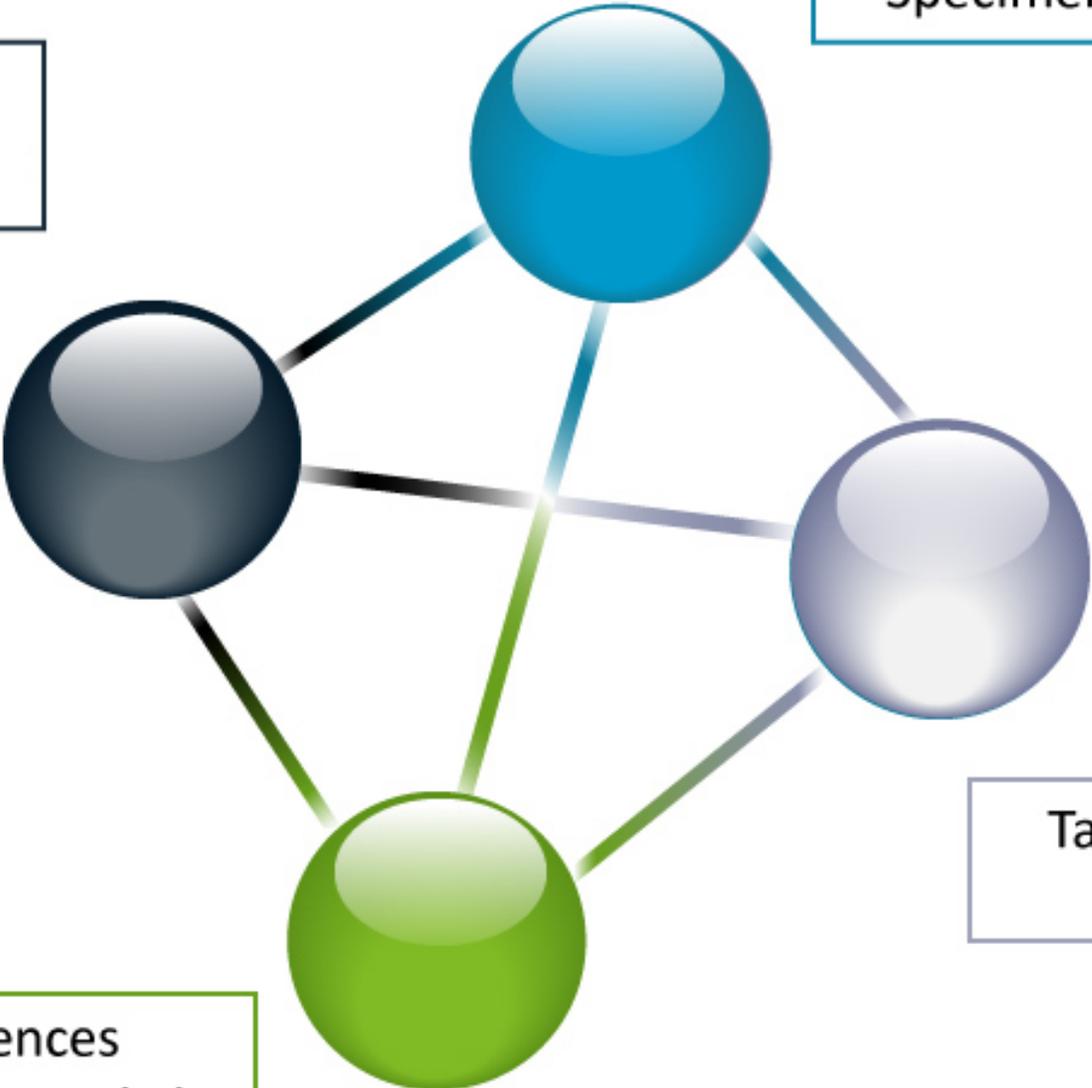


# Taxonomic Processes



**Publications**

**Specimens**



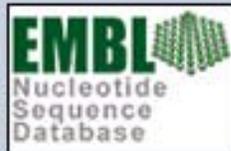
**Taxonomic names**

**DNA sequences and other characteristics**

Next-Generation  
Sequence data

DNA BARCODE  
data

**A DNA barcode is a  
short gene sequence  
taken from  
standardized portions  
of the genome,  
used to identify species**



## International Nucleotide Sequence Database Collaboration

- The International Nucleotide Sequence Databases (INSD) have been developed and maintained collaboratively between [DDBJ](#) , [EMBL](#) , and [GenBank](#) for over 18 years.
- The INSDC advisory board, the [International Advisory Committee](#) , is made up of members of each of the databases' advisory bodies. At their most recent meeting, members of this committee unanimously endorsed and reaffirmed the existing data-sharing policy of the three databases that make up the INSDC , which is stated below.
- Individuals submitting data to the international sequence databases should be aware of [INSDC policy](#) .

### How to submit data

- For full details of how to submit data to the databases, please select a collaborating partner.
- [DDBJ](#) , [EMBL](#) , [GenBank](#)
- The INSDC Feature Table Definition Document is available [here](#) .

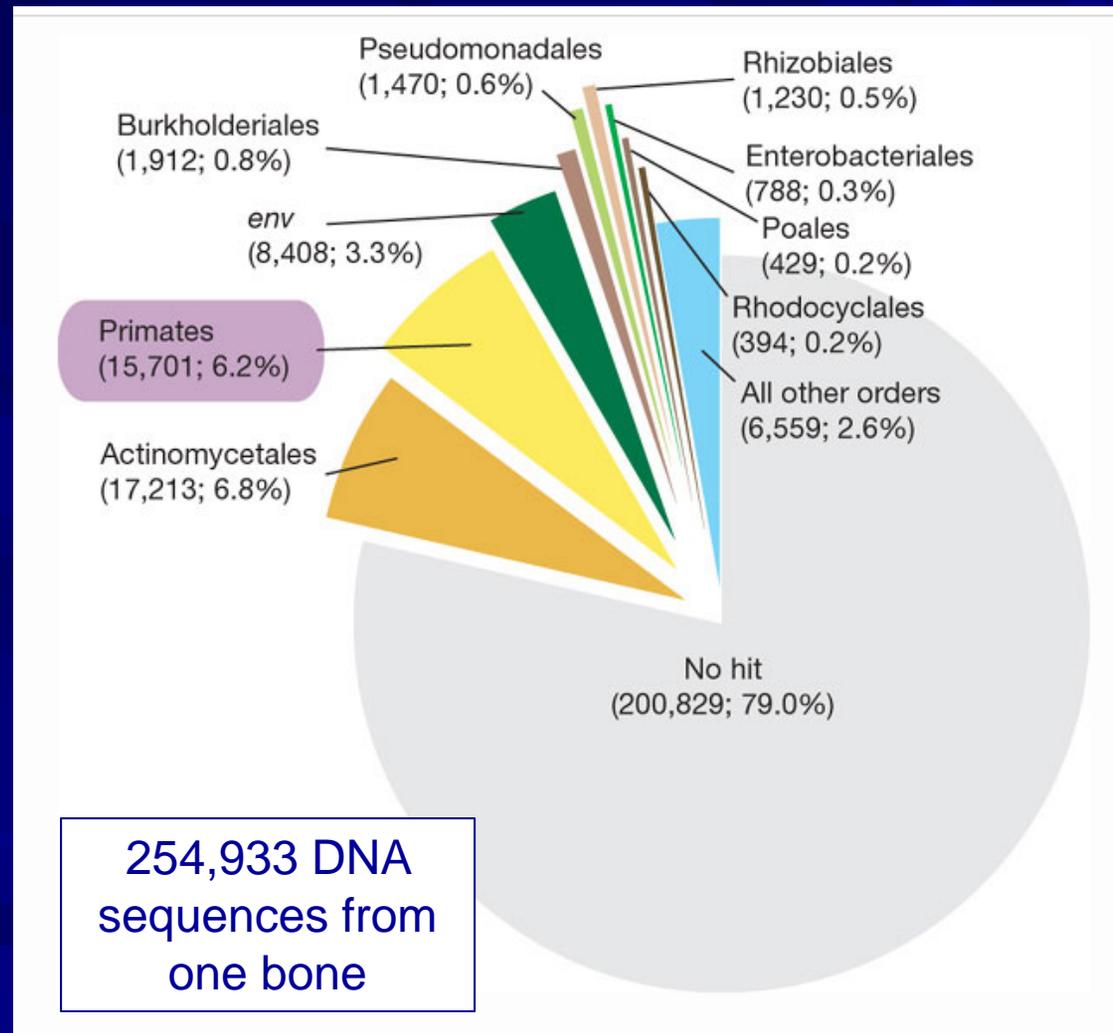
# How much information in GenBank?

- Whole Genome Sequencing:
  - 191 billion bases
  - 62 million records
- “Traditional” GenBank – small projects:
  - 127 billion bases
  - 135 million sequence records

# What's the Distinction?

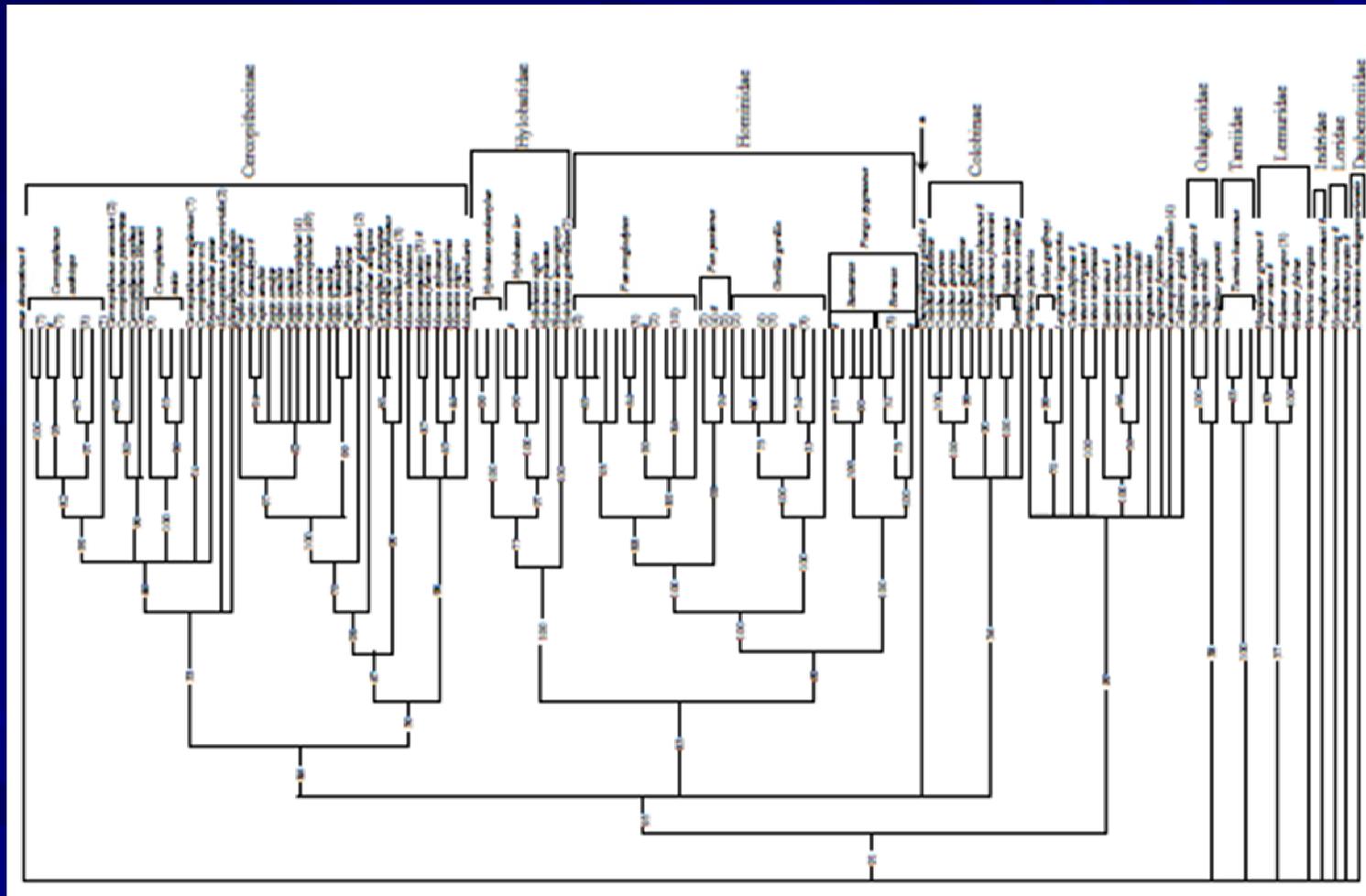
- Technology of sensors
- Popular demand
- Supply rate
- Cost of production
- Ease of access
- Inter-connectivity among data records
- Viral buzz

# Neanderthal DNA Study



Green, et al., *Nature* 444:330-336 (16 November 2006)

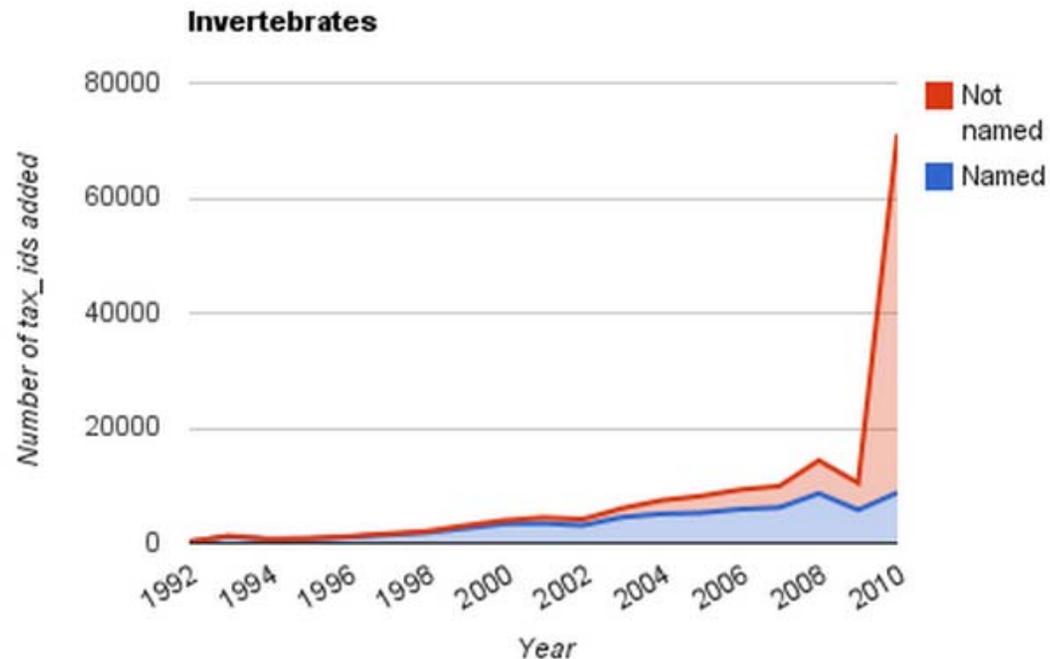
# 225 DNA barcodes, 20 species



Lorenz et al., Phil. Trans. R. Soc. B (2005) 360:1869–1877

# Rod Page's 'Dark Taxa'

For "invertebrates" 2010 saw an explosive growth in the number of new taxa sequenced, with nearly 71,000 new taxa added to GenBank.



This coincides with a spectacular drop in the number of properly-named taxa, but even before 2010 the proportion of named invertebrate species in GenBank was in decline: in 2009 just over a half of the species added had binomials.

# A 30-year Trend

- Increasing access to publications
- Release of data
- Development of data standards, ontologies
- Growth of data aggregators, networks, centers
- Mandates for open access publication
- Mandates for data management plans

# Fort Lauderdale Principles, 2003

## Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility

Report of a meeting organized by the Wellcome Trust  
and held on 14–15 January 2003 at Fort Lauderdale, USA.

- ‘Win-Win’ roles of funders, producers and users
- ‘Community Resource Projects’ to generate data for re-use
- Global public goods versus self-interest of researchers
- Pre-publication release of publicly-funded genomic research through GenBank, EMBL, DDBJ
- Protect researcher rights through publication of ‘Project Descriptions’ to describe plans, timeline for publication

# Data Aggregators and Networks

	Major Data Repositories	Data Standards
DNA and Genes	<a href="#">International Nucleotide Sequence Database Collaboration</a> (GenBank, EMBL, DDBJ)	<a href="#">Gene Ontology</a> <a href="#">Genomic Standards Consortium</a>
Species	<a href="#">Global Biodiversity Information Facility</a>	<a href="#">Darwin Core Standards</a>
Populations	<a href="#">Global Population Dynamics Database</a>	<a href="#">Ecological Markup Language</a>
Ecological communities	<a href="#">DataOne</a>	<a href="#">Ecological Markup Language</a>
Coastal and Marine Environments	<a href="#">Digital Coast</a>	<a href="#">Coastal and Marine Ecological Classification Standard</a>
Habitats	<a href="#">European habitats</a>	<a href="#">EUNIS Habitat Classification</a>
Geospatial landscape data	<a href="#">US National Geospatial Program</a> <a href="#">European Environment Agency data/maps</a>	<a href="#">Open Geospatial Consortium</a>

# Global Biodiversity Information Facility



The screenshot shows the website's header with a green logo on the left and navigation links (SPECIES, COUNTRIES, DATASETS, OCCURRENCES, SETTINGS, ABOUT) on the right. Below the navigation is a blue area with a circuit-like pattern, containing XML code and four small images of a gorilla, a butterfly, a dolphin, and a flower. At the bottom of this section is the text "... free and open access to biodiversity data".

GLOBAL BIODIVERSITY INFORMATION FACILITY

SPECIES COUNTRIES DATASETS OCCURRENCES SETTINGS ABOUT

```
<?xml version="1.0" encoding="UTF-8"
<response xmlns="http://rs.tdwg.org/t
<header>
<source accesspoint="http://145.18.162/
<software name="TapirLink" version="0.2(re
```

... free and open access to biodiversity data

389,467,366 data records  
340,426,764 with coordinates

[Return to Search](#)

[Email Query](#)

[Bookmark Query](#)

[RSS Feed for Query](#)

[Help](#)

Query: text : rodent AND ( datasource :( urn\:node\:LTER ) )

[Hide Filters](#)

Filter by author	Filter by project	Filter by keywords	Filter by Originator
<a href="#">David Lightfoot (54)</a> <a href="#">David Tilman (21)</a> <a href="#">Brandon Bestelmeyer (7)</a> <a href="#">Paul Stapp (7)</a> <a href="#">Esteban Muldavin (6)</a> <a href="#">Don W. Duszynski (5)</a> <a href="#">Ana Davidson (4)</a>	<a href="#">Sevilleta LTER (89)</a> <a href="#">Jornada LTER (7)</a> <a href="#">Coweeta Long Term (2)</a> <a href="#">Long-Term Ecological Research (2)</a> <a href="#">Small mammal population (1)</a>	<a href="#">LTER (27)</a> <a href="#">Sevilleta National Wildlife Refuge (26)</a> <a href="#">New Mexico (25)</a> <a href="#">SEV (25)</a> <a href="#">Ecology (21)</a> <a href="#">Biodiversity (20)</a>	<a href="#">Department of Biological Science (7)</a> <a href="#">Andrews Forest LTER Site (1)</a> <a href="#">Dept. of Biology (1)</a>

Sort By: [Relevance](#) [Date](#) [Member Node](#)

Viewing Documents 1 - 10 out of 132  
[Prev](#) [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#) [Next](#)

**PINO GATE PRAIRIE DOG STUDY: LANDSCAPE PLOT GROUND-DWELLING ARTHROPOD DATA** 06/01/2000 - 10/01/2001

**Datasource:** LTER NETWORK MEMBER NODE

Keystone species have large impacts on community and ecosystem properties, and create important ecological interactions with other species. Prairie dogs (*Cynomys* spp.) and banner-tailed kangaroo rats (*Dipodomys spectabilis*) are considered keystone species of grassland ecosystems, and create a mosaic of unique habitats on the landscape. These habitats are known to attract a number of animal species, but little is known about how they affect arthropod communities. Our research evaluated the keystone roles of prairie dogs and kangaroo rats on arthropods at the Sevilleta National Wildlife Refuge...



[View full metadata](#)

[Data Files \(0\)](#)

**SMES CRYPTOGAM CRUST DATA** N/A - N/A

# Training Needs

- Some specialist advanced degrees
  - Staffing for data centers, aggregators
- Coursework for all doctoral candidates
- Undergrad concentration/majors
  - Library/information schools
  - Joint with Computer Science
- Undergrad courses for science credit
- K-12 life skills in math/science, personal finances
- YouTube instruction, Citizen Science