# Dealing with the Wealth of Open Source Data

John Holliday
University of Sheffield

# Contents

- Sheffield Group Research
- Explosion in Data Resources
- Curation
- Formats
- Benchmark Datasets
- Conclusions

# Sheffield Research Areas

- Molecular Similarity/Diversity, Clustering, Data Fusion
- Graph Matching, Hyperstructures
- Machine Learning, HTS Profiling
- 3D – Alignment, Protein Surface Comparisons, Pharmacophore Modelling, 3D Similarity, De Novo Design, Hypermolecules, Docking
- ADME – SAR in Toxicology

# Pre Open Source – at Sheffield

Circa 1999/2000

| | |
|---|---|
| MDDR 102K | (WDI) |
| IDAlert 11.5K | (DNP) |
| NCI AIDS99/Cancer99 42K/32K | |
| Starlist 9.5K | SciFinder et al |
| Bioster | |

# ‘Open’ Resources

SuperNatural BindingDB PubChem ChemDB

SuperDrug NIST KiBank MUV ZINC Brenda

Kegg

Ligand Expo ChemBank TTD

BiaDB NPACT BioCYC DUD ChEMBL

HIT

HMDB SMPDB PharmaGKB

DrugBank

NCI

PDB-Bind ChemSpider

PDBeChem

HPRD

ChemIDPlus CHMIS-C

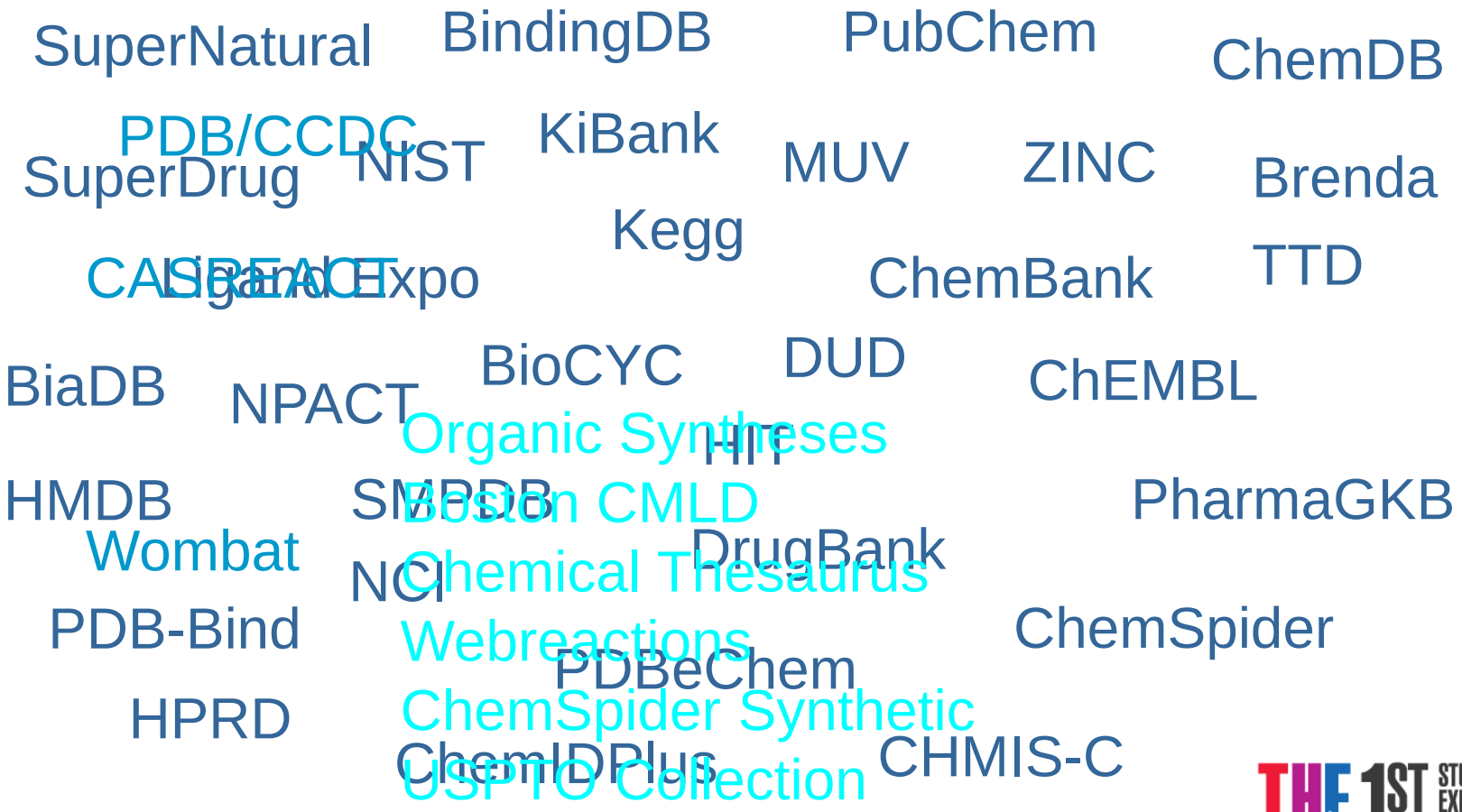
 

# Current Sheffield Research

| | |
|---|---|
| Hyperstructures/Clique Detection | MDDR, Wombat, MUV |
| GAs and GPs for VS | MDDR, Wombat, ChEMBL |
| Dimensionality on SS | Wombat/ChEMBL /MUV ? |
| Order Theory in Data Fusion | Wombat/ChEMBL /MUV ? |
| Orphan Drug Status | (MDDR) |
| Seizures in Zebrafish | ChEMBL, ZINC, SciFinder, ZFIN |
| Bio-isosteric Similarity of Fragments | Chembl, PDB with CCDC, corporate DB |
| Transformation-based De Novo Design | Chembl, ChemSpider, PubChem, ZINC |
| Reaction Networks | CASREACT |

# ‘Open’ Resources



# Cross-database Integration

- Multiple formats
  - Consistency
  - Correlation
  - Integrity
- Data completeness
- Data confidence

# External Data Sources

- Unstructured data
  - new formats
  - new data types
  - nature of the web
  - software formats
- Standardisation
- Will some rise in popularity?

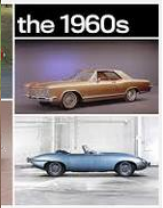
# Hardware and Software

- Managed repository
- Backup, security
- Duplication
- Consistency/standards
- Search/metadata
- Full management system

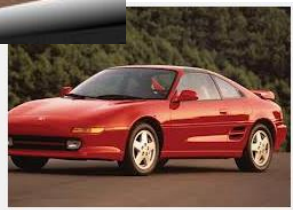

 

# Management System

## Classic D

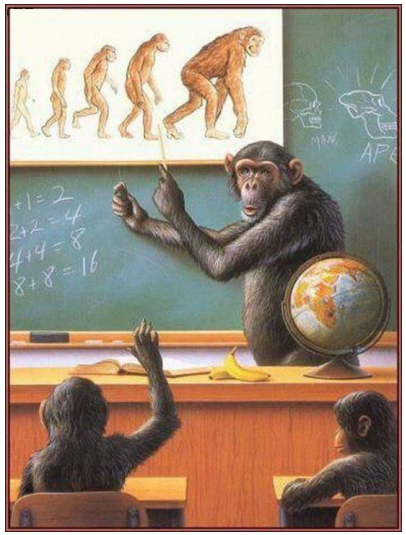




## Compute



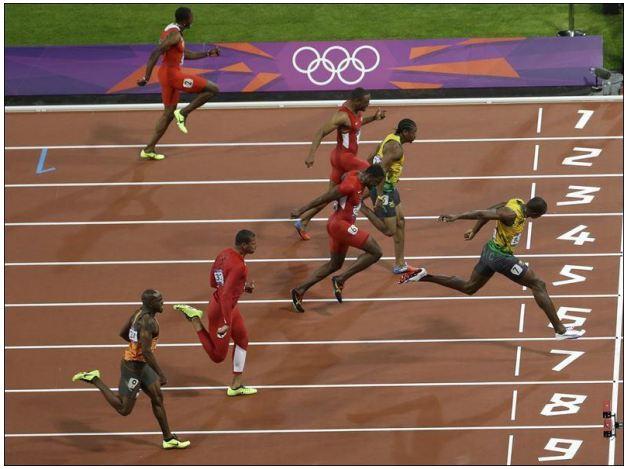
 

# Data Analysis, Visualisation, Organisation



- Analysis, aggregation, classification, visualisation
- Tailored presentation
  - We're all individuals
  - You must speak the right language
- Data must be contextually correct

# Benchmarking



- Evolving databases
  - moving goalposts
- Several benchmark sets exist
  - Carefully chosen for a variety of challenges
  - Well referenced and becoming standard for comparison
  - Temporal change still an issue

# Conclusions

- Increase in ‘data awareness’ in the discipline
- We are becoming data scientists
  - But we can’t forget our roots
- Decrease in programming skills and increase in the right ‘tools’