100 million compounds, 100K protein structures, 2 million reactions, 1 million journal articles, 20 million patents and 15 billion substructures

Is 20TB really Big Data?

Noel O'Boyle, Daniel Lowe, John May and Roger Sayle

NextMove Software

BIG DATA IS...

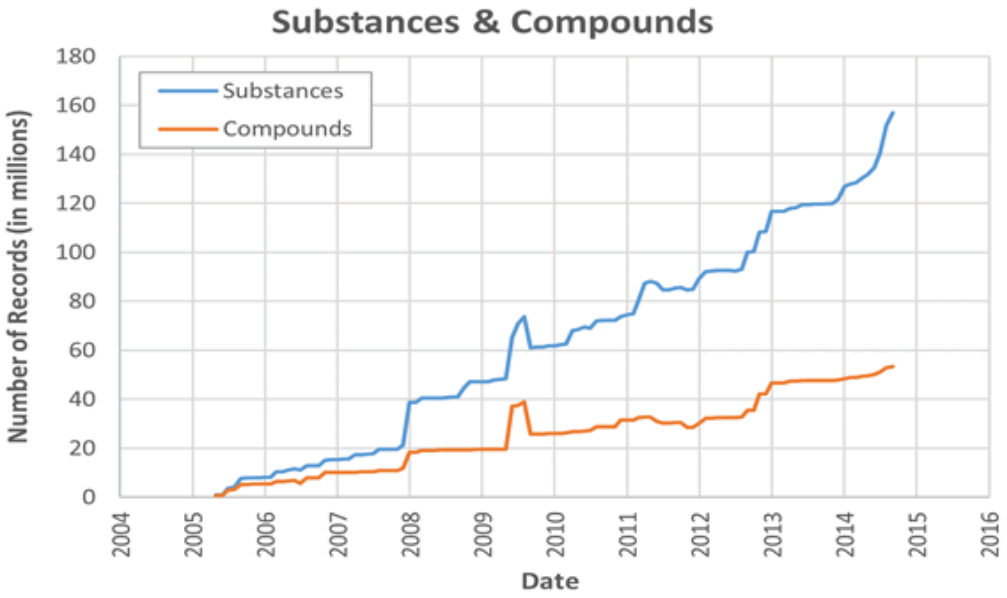
"...a broad term for data sets so large or complex that traditional data processing applications are inadequate." [Wikipedia]

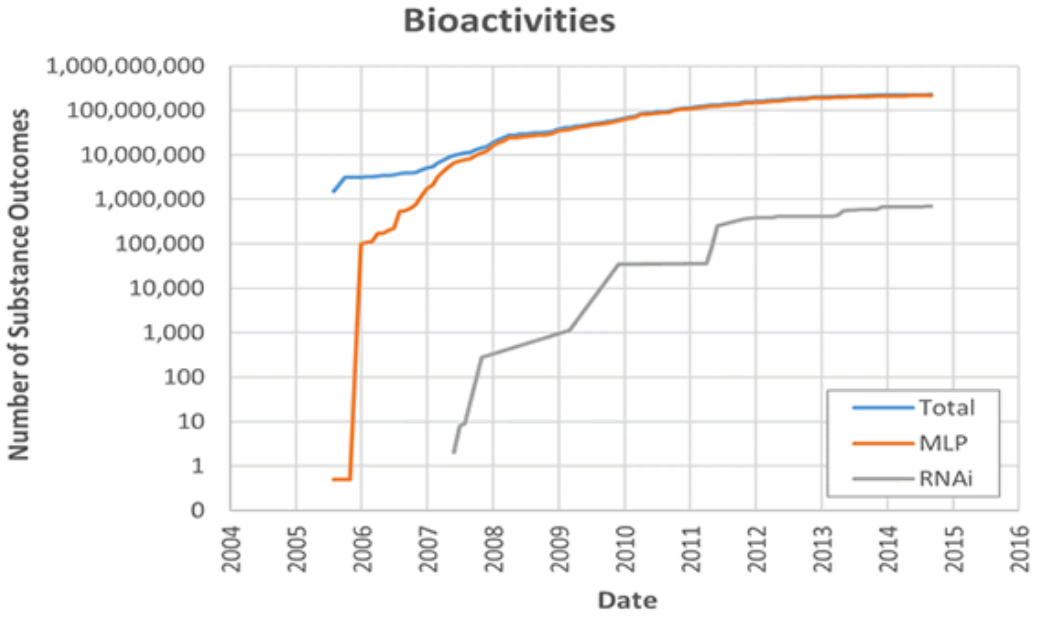
Any dataset could be considered *Big Data* without sufficiently efficient algorithms and tools

100K	protein structures	Swiss-Prot
2 million	reactions	Patents
1 million	journal articles	PubMed Central OA
20 million	patents	US, EU, JP, Kr
100 million	compounds	PubChem, UniChem
15 billion	substructures	PubChem, ChEMBL

20 Tb

700K	CSD entries
36 million	Wikipedia articles
47 billion	webpages indexed by Google
200 billion	tweets per year
22 Pb	EMBL-EBI's data



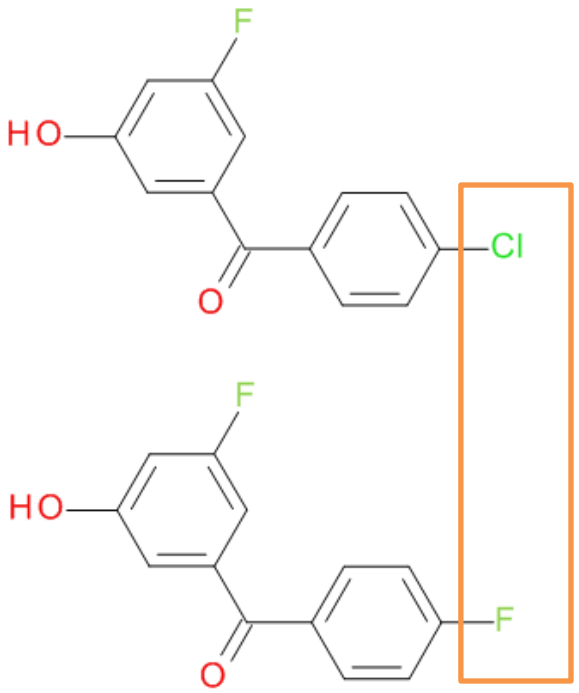


http://pubchemblog.ncbi.nlm.nih.gov/2014/09/16/ten-years-of-service/

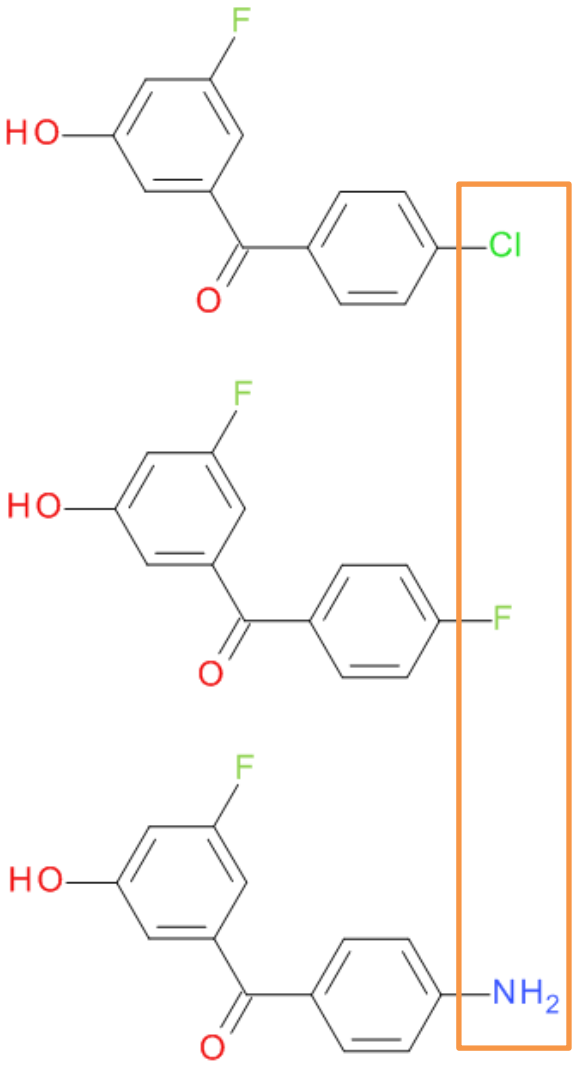
OVERVIEW

- Finding matched pairs/series
- Substructure searching
- Maximum common subgraph
- Chemical text-mining
- Naming reactions
- Canonicalisation

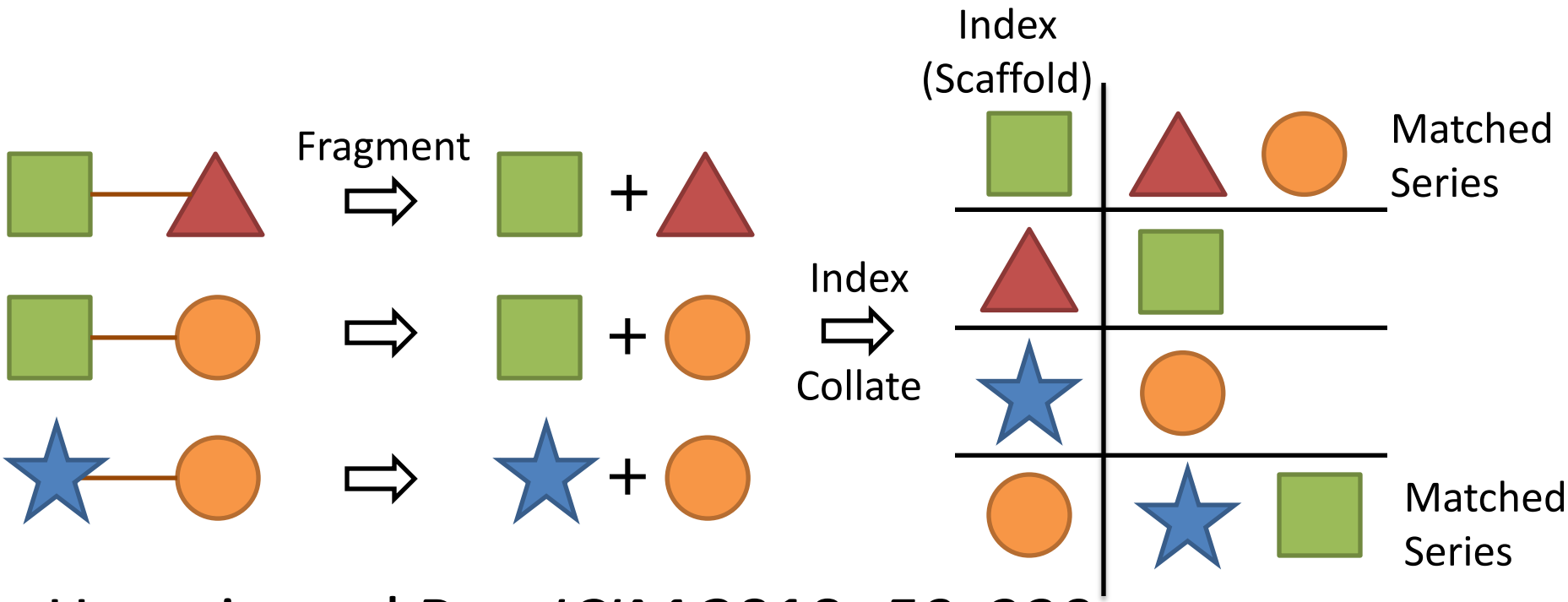
MATCHED PAIR



MATCHED SERIES OF LENGTH 3



FIND MATCHED PAIRS/SERIES



- Hussain and Rea *JCIM* **2010**, *50*, 339
- ChEMBL 20 IC_{50} data
  - 752 K datapoints from 64 K assays
- Processed in 12 minutes
  - Giving 391 K matched series

"Google searches are screamingly fast, so fast that the type-ahead feature is doing the search as you key characters in. Why are all chemical searches so sloooow? … Ideally, as you sketch your mol in, the searches should be happening at the same pace, like the typeahead feature."

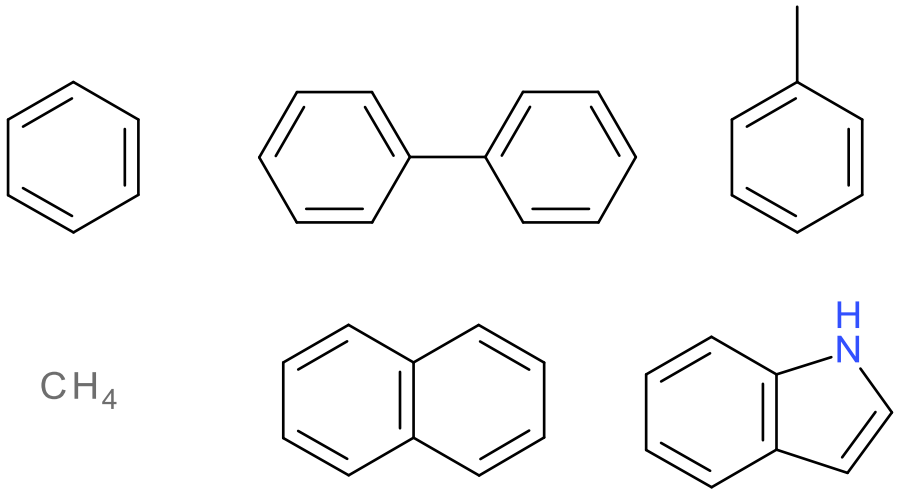
John Van Drie, Nov 2011

Via Rajarshi Guha's blog

http://blog.rguha.net/?p=993

SUBSTRUCTURE SEARCHING

- Approach: fingerprint screen (fast but false positives) then match (slow but exact)
- Pathological cases – many pass the screen
  - "denial of service queries" (Trung Nguyen)



  - Substructures which happen to map to the same bit as benzene, etc.

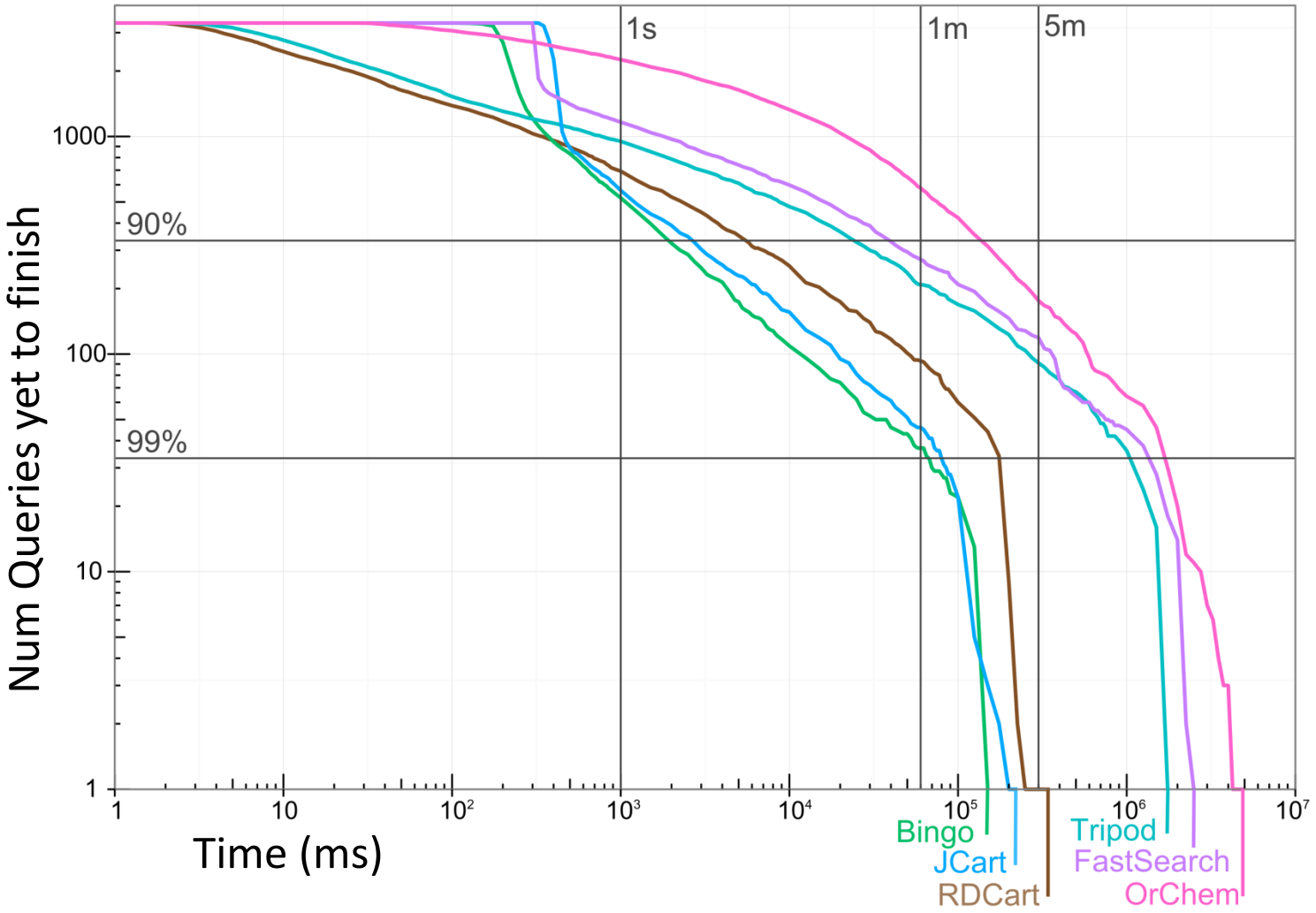
FASTER SUBSTRUCTURE SEARCHING

- Worst-case behaviour dominated by slow matching
  - This implies focus should be on faster matching
- Typical substructures can be expressed as SMARTS patterns
  - Arthor: fast SMARTS matching against a database
- Test set: Structure Query Collection (Andrew Dalke, BindingDB)
  - Time to count hits for 3323 queries against eMolecules (6.9 million)

OPTIMISED SMARTS MATCHING

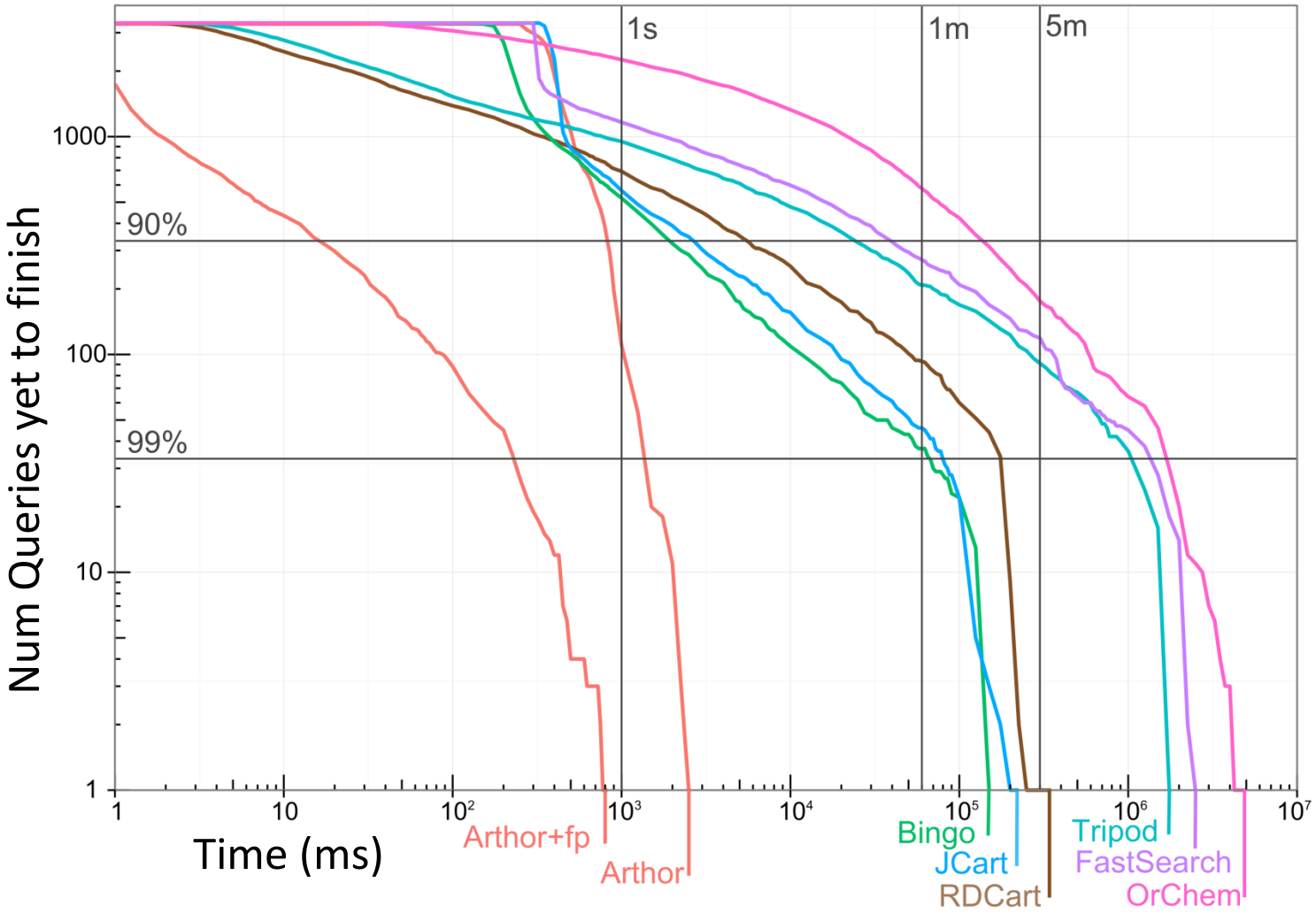
- Preprocess database so that matching is fast
  - Efficient binary representation (minimal I/O)
  - Matching done directly on binary representation (minimal malloc)
- Match rarer atom expressions first (*)
  - CCCCCBr → BrCCCCC
- Match rarer bond expressions first
  - CC#C → C#CC

* c.f. slide 31 of ICCS presentation
http://www.slideshare.net/NextMoveSoftware/efficient-matching-of-multiple-chemical-subgraphs



Total time for all 3323 queries

Bingo	2h 7m	Tripod	1d 6h
JCart	2h 42m	FastSearch	1d 15h
RDCart	5h 9m	OrChem	3d 1h

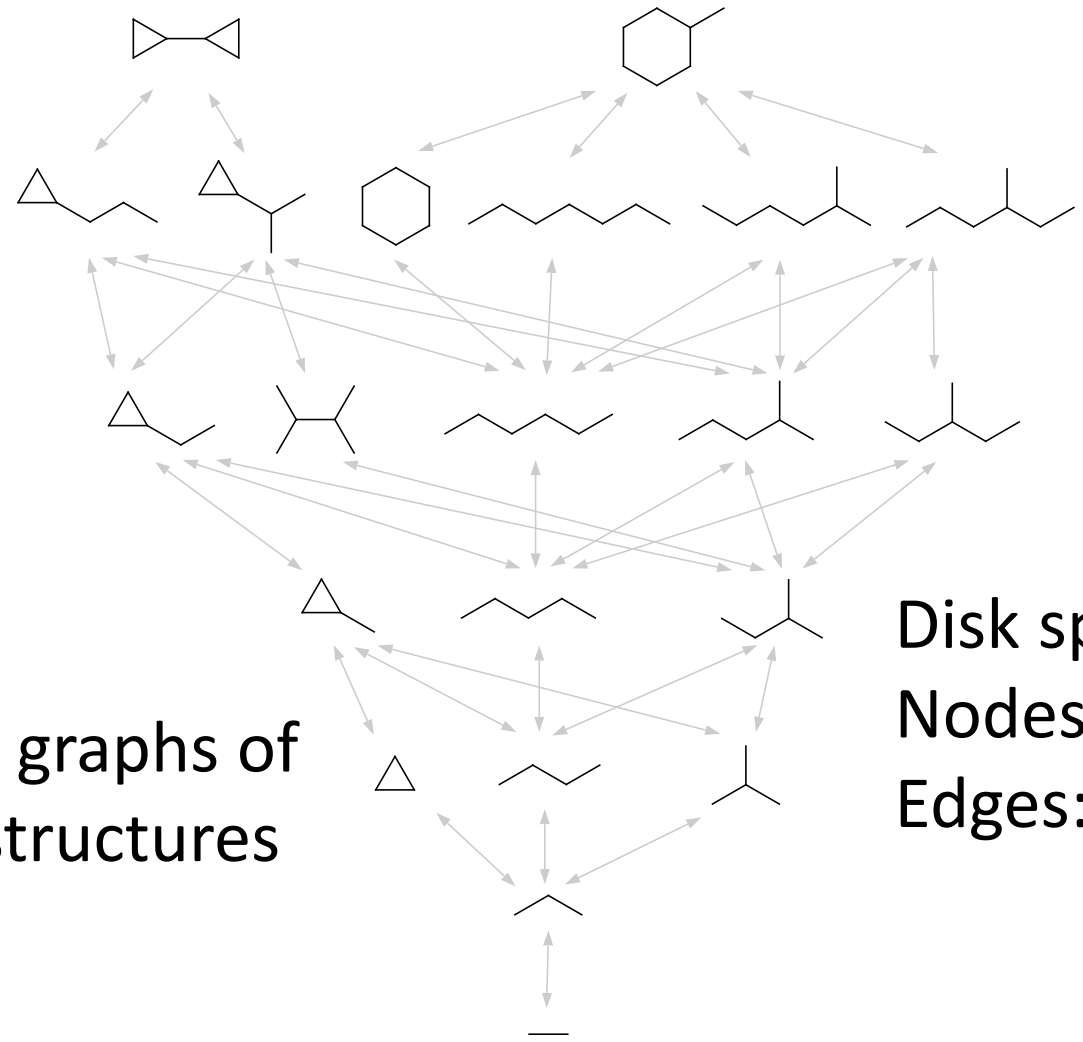


Arthor+fp	36s	Bingo	2h 7m	Tripod	1d 6h
Arthor	27m 31s	JCart	2h 42m	FastSearch	1d 15h
Total time for all 3323 queries		RDCart	5h 9m	OrChem	3d 1h

WHAT ABOUT EVEN LARGER DATABASES?

- Imagine a chemical database 100 times larger than PubChem
  - Search time scales linearly, so even Arthor will take 100 times longer to search it
- A completely different approach is needed
- SmallWorld: sublinear searching by precalculating all possible substructures and their relationships
  - The catch: disk space, and takes months of CPU-time to generate (here's one we made earlier)

SMALLWORLD – A GRAPH DATABASE



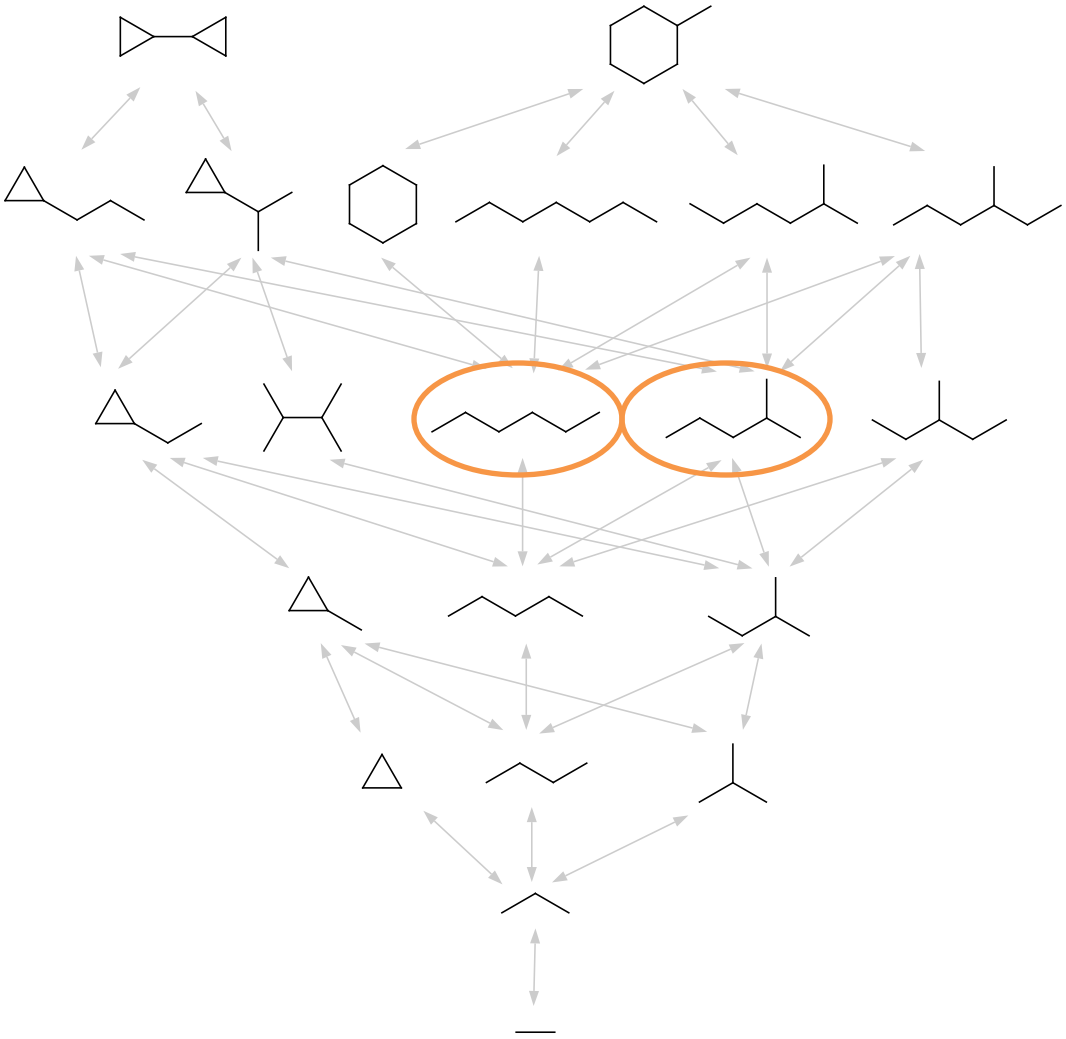
Nodes are anonymous graphs of known substructures

Disk space: 12TB
Nodes: 19.7 billion
Edges: 64.8 billion

MAXIMUM COMMON SUBGRAPH (MCS)

- Computationally expensive
- Previous approaches:
  - Backtracking algorithms
  - Clique detection
  - Dynamic programming
- However, with SmallWorld the computation has already been done in advance
  - MCS can be found in linear time relative to the number of atoms in the smaller molecule

SMALLWORLD – A GRAPH DATABASE



CHEMICAL TEXT-MINING BIG DATA

- LeadMine for chemical entity extraction from text
  - Performance: Finds ~90% of chemical structures (compared to inter-annotator agreement of 91%*)
- Method:
  - Dictionaries (e.g. list of common English words, trivial chemical names) and grammars (e.g. all possible IUPAC names)
  - Speed just depends on the number of dictionaries
  - Any dictionary can be used with spelling correction

* Krallinger et al, *J. Cheminf.* **2015**, *7*, S2

HOW LONG TO PROCESS?

- All Open Access papers available from PubMed Central
  - 1.0 million articles processed in 3h 38min (*)
  - 131K distinct compounds
- 2001-2015 USPTO applications (5 times larger)
  - 4.1 million patents processed in 22h 50min (*)
  - 4.3 million distinct compounds
  - 84.7 million compound mentions (with at least 10 heavy atoms)

* using 4 cores

AUTOMATIC NAMING OF REACTIONS

- Traditional approaches:
  - Atom-mapping (can be slow, can give wrong mapping)
  - Differences in fingerprints for reactants and products (fast but has limitations)
- Our approach uses compiled SMARTS matching:
  - Apply a particular reaction to the reactants and check whether the products appears on the right
  - Components are atom-mapped implicitly

Note: typically reactions in ELNs and patents are not balanced

PERFORMANCE

- Scales with the number of SMARTS patterns used
  - currently 802 patterns for 504 reactions
- Test set: 1.1 million reactions extracted from the USPTO applications 2001-2012
  - Processed in 11.2 h
  - 437 K reactions named

```
US20010000038A1 : 3.10.2 Friedel-Crafts alkylation [Friedel-Crafts reaction]
US20010000038A1 : 10.1.5 Wohl-Ziegler bromination [Halogenation]
```

A DIFFERENT TYPE OF "BIG" DATA

- Large macromolecules can be efficiently handled by cheminformatics tools

- Titin (35213 amino acids, 313 K atoms):

CC[C@H](C)[C@@H](C(=O)NCC(=O)N[C@@H](CCCCN)C(=O)N1CCC[C@H]1C(=O)N[C@@H](CO)C(=O)N[C@@H](Cc2c[nH]cn2)C(=O)N3CCC[C@H]3C(=O)N[C@@H](CO)C(=O)N[C@@H](CCC(=O)O)C(=O)N4CCC[C@H]4C(=O)N[C@@H](C(C)C)C(=O)N[C@@H](CC(C)C)C(=O)N[C@@H](C)C(=O)N[C@@H]([C@@H](C)CC)C(=O)N[C@@H](CC(=O)O)C(=O)N[C@@H](C)C(=O)N[C@@H](CS)C(=O)N[C@@H](CCC(=O)O)C(=O)N5CCC[C@H]5C(=O)N6CCC[C@H]6C(=O)N[C@@H](CCCNC(=N)N)C(=O)N[C@@H](CC(=O)N)C(=O)N[C@@H](C(C)C)C(=O)N[C@@H](CCCNC(=N)N)C(=O)N[C@@H]([C@@H](C)CC)C(=O)N[C@@H]([C@@H](C)O)C(=O)N[C@@H](CC(=O)O)C(=O)N[C@@H]([C@@H](C)CC)C(=O)N[C@@H](CO)C(=O)N[C@@H](CCCCN)C(=O)N[C@@H](CC(=O)N)C(=O)N[C@@H](CO)C(=O)N[C@@H](C(C)C)C(=O)N[C@@H](CC(=O)N)C(=O)N[C@@H](CC(C)C)C(=O)N[C@@H](CO)C(=O)N[C@@H](Cc7c[nH]c8c7cccc8)C(=O)N[C@@H](CCC(=O)N)C(=O)N[C@@H](CCC(=O)N)C(=O)N9CCC[C@H]9C(=O)N[C@@H](C)C(=O)N[C@@H](Cc1ccccc1)C(=O)N[C@@H](CC(=O)O)C(=O)NCC(=O)NCC(=O)N[C@@H](CO)C(=O)N[C@@H](CCCCN)C(=O)N[C@@H]([C@@H](C)CC)C(=O)N[C@@H]([C@@H](C)O)C(=O)NCC(=O)N[C@@H](Cc1ccc(cc1)O)C(=O)N[C@@H]([C@@H](C)CC)C(=O)N[C@@H](C(C)C)C(=O)N[C@@H](CCC(=O)O)C(=O)N[C@@H](CCCNC(=N)N)C(=O)N[C@@H](CCCNC(=N)N)C(=O)N[C@@H](CC(=O)O)C(=O)N[C@@H](CC(C)C)C(=O)N1CCC[C@H]1C(=O)N[C@@H](CC(=O)O)C(=O)NCC(=O)N[C@@H](CCCNC(=N)N)C(=O)N[C@@H](Cc1c[nH]c2c1cccc2)C(=O)N[C@@H]([C@@H](C)O)C(=O)N[C@@H](CCCCN)C(=O)N[C@@H](C)C(=O)N[C@@H](CO)C(=O)N[C@@H](Cc1ccccc1)C(=O)N[C@@H]([C@@H](C)O)C(=O)N[C@@H](CC(=O)N)C(=O)N[C@@H](C(C)C)C(=O)N[C@@H]([C@@H](C)CC)C(=O)N[C@@H](CCC(=O)O)C(=O)N[C@@H]([C@@H](C)O)C(=O)N[C@@H](CCC(=O)N)C(=O)N[C@@H](Cc1ccccc1)C(=O)N[C@@H]([C@@H](C)O)C(=O)N[C@@H](C(C)C)C(=O)N[C@@H](CO)C(=O)NCC(=O)N[C@@H](CC(C)C)C(=O)N[C@@H]([C@@H](C)O)C(=O)N[C@@H](CCC(=O)N)C(=O)N[C@@H](CC(=O)N)C(=O)N[C@@H](CO)C(=O)N[C@@H](CCC(=O)N)C(=O)N[C@@H](Cc1ccc(cc1)O)C(=O)N[C@@H](CCC(=O)O)C(=O)N[C@@H](Cc1ccccc1)C(=O)N[C@@H](CCCNC(=N)N)C(=O)N[C@@H](C(C)C)C(=O)N[C@@H](Cc1ccccc1)C(=O)N[C@@H](C)C(=O).....

A DIFFERENT TYPE OF "BIG" DATA

- Large macromolecules can be efficiently handled by cheminformatics tools

- Titin (35213 amino acids, 313 K atoms):
  - Canonical Smiles: 238s (state-of-the-art)
  - Canonical Smiles: 1.7s (our preliminary results)

- All Swiss-Prot entries (541K structures)
  - Remove those with ambiguous structures (e.g. Asx)
  - 453K protein structures canonicalised in 2m 57s (4 cores)

100 million compounds, 100K protein structures, 2 million reactions, 1 million journal articles, 20 million patents and 15 billion substructures

Is 20TB really Big Data?

Noel O'Boyle, Daniel Lowe, John May and Roger Sayle

NextMove Software

noel@nextmovesoftware.com

100 million compounds, 100K protein structures, 2 million reactions, 1 million journal articles, 20 million patents and 15 billion substructures

Is 20TB really Big Data?

With modern hardware and efficient algorithms, many classic cheminformatics problems can be handled with today's datasets.

Noel O'Boyle, Daniel Lowe, John May and Roger Sayle

NextMove Software

noel@nextmovesoftware.com