

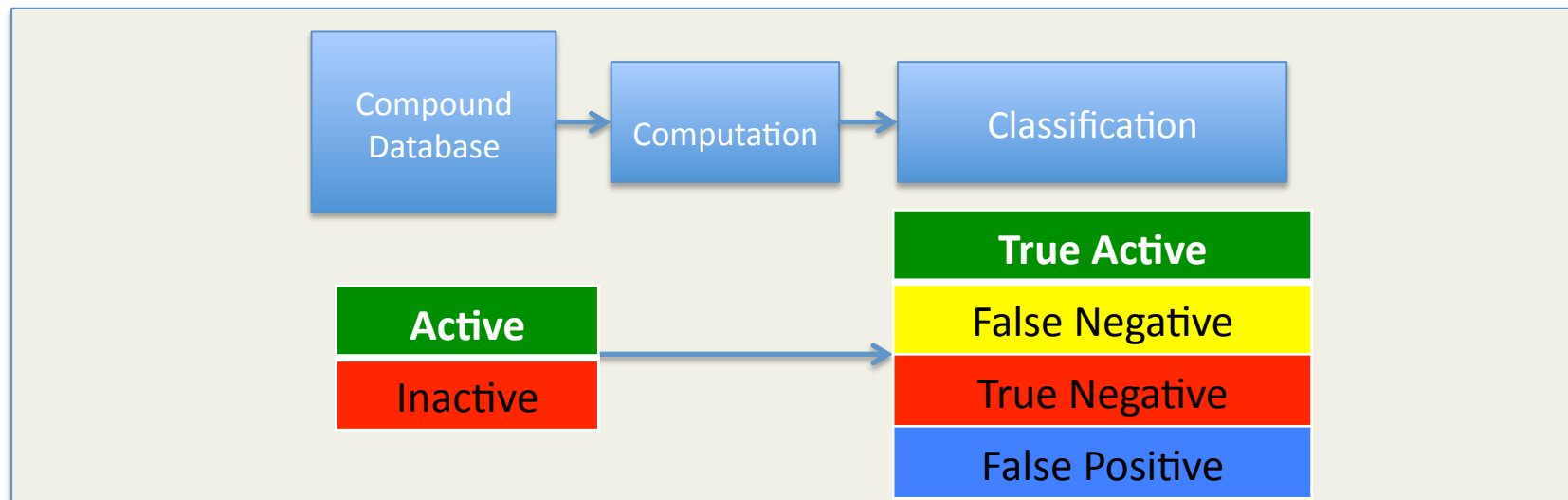
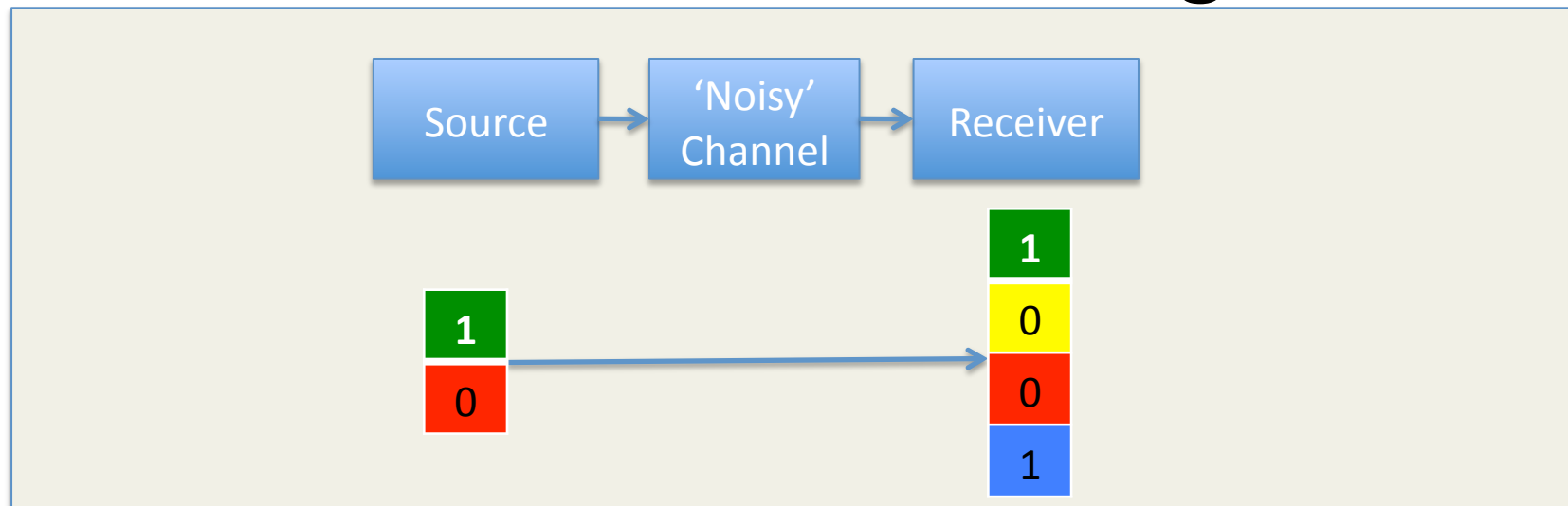
Statistical Reasoning with Ligands

Ligands and Information Content

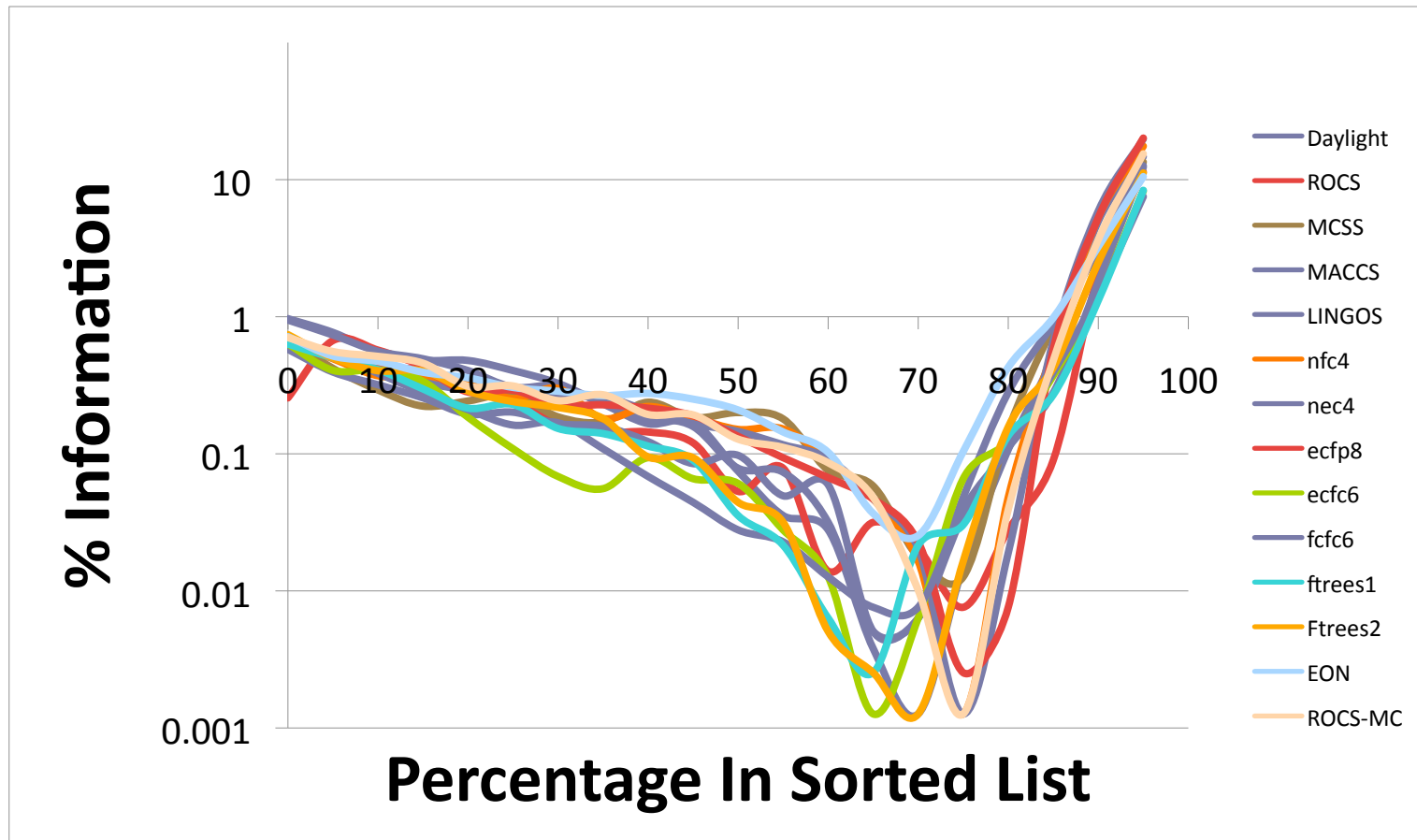
Statistics & Information

- Statistics deals with I. I. D. objects
 - Identical and Independently distributed objects
- Ligands *aren't*
- Information theory gives a framework for dealing with this
- Increasingly interesting

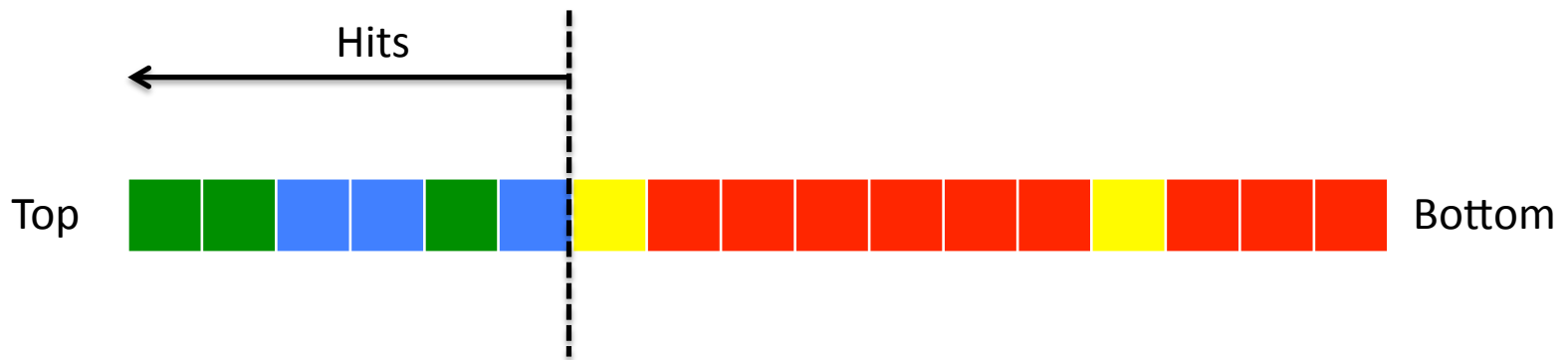
Telecommunication and Virtual Screening



Log Linear Information vs % Rank

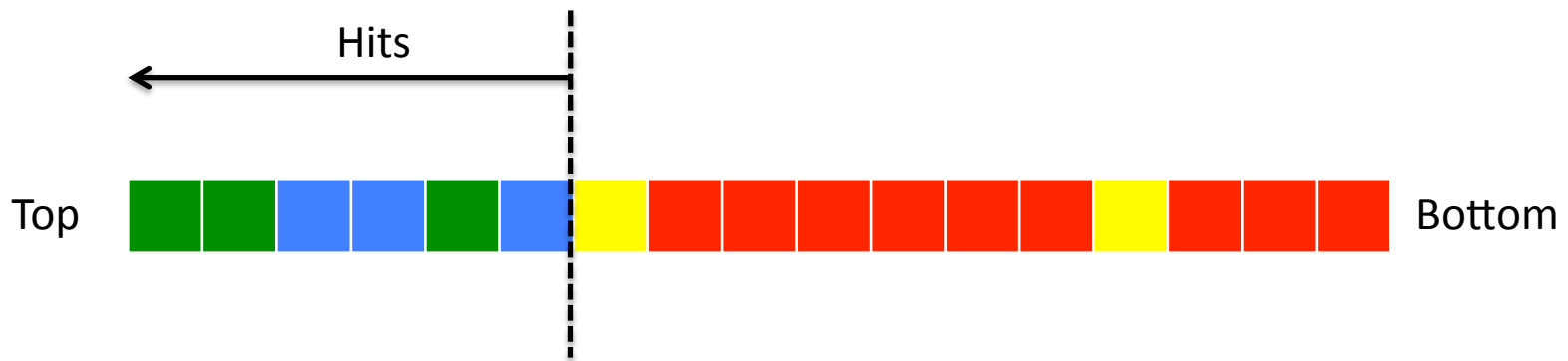


Information



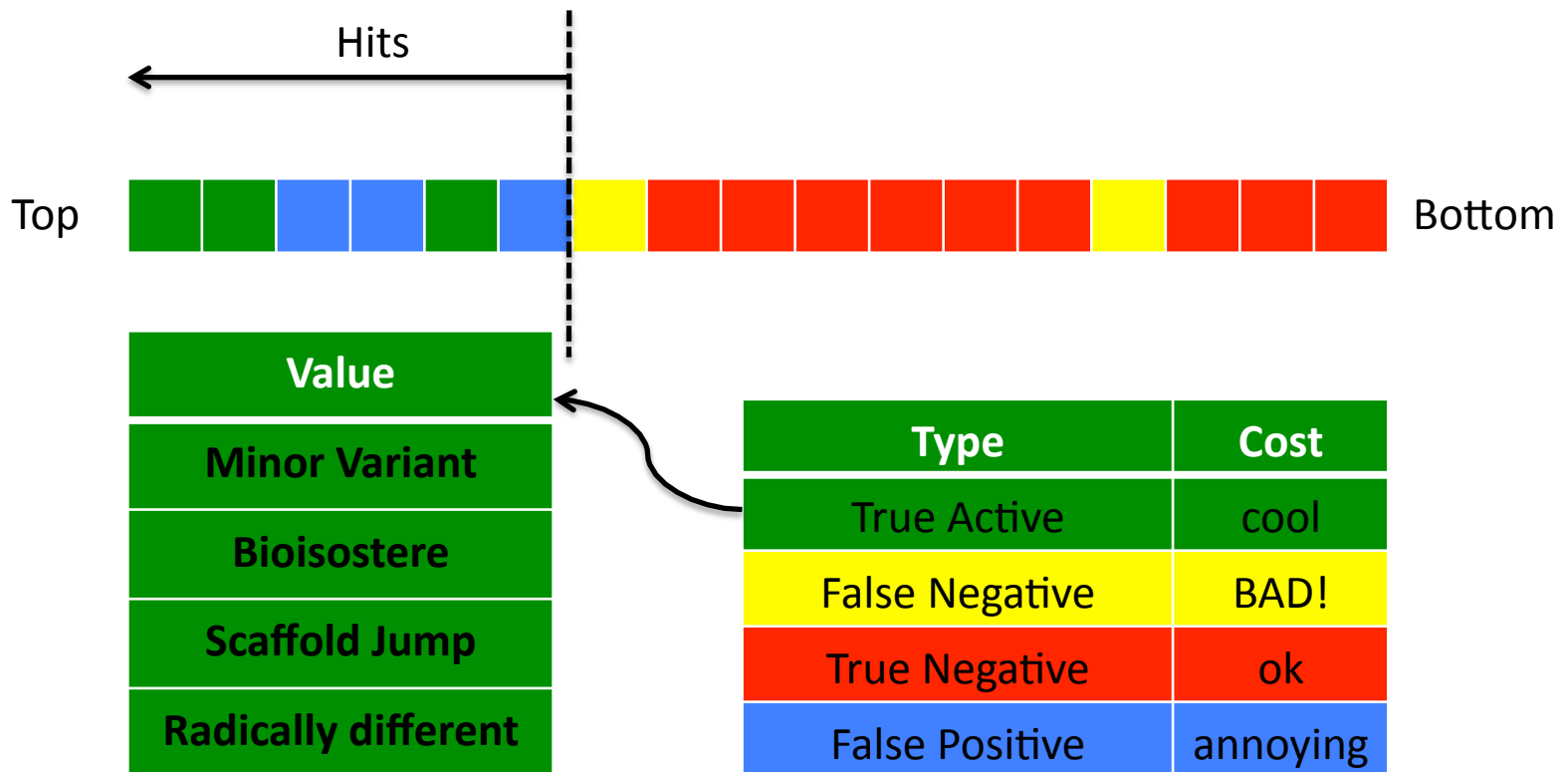
Type	Cost
True Active	1
False Negative	0
True Negative	1
False Positive	0

The Value of Information



Type	Cost
True Active	cool
False Negative	BAD!
True Negative	ok
False Positive	annoying

The Value of Information

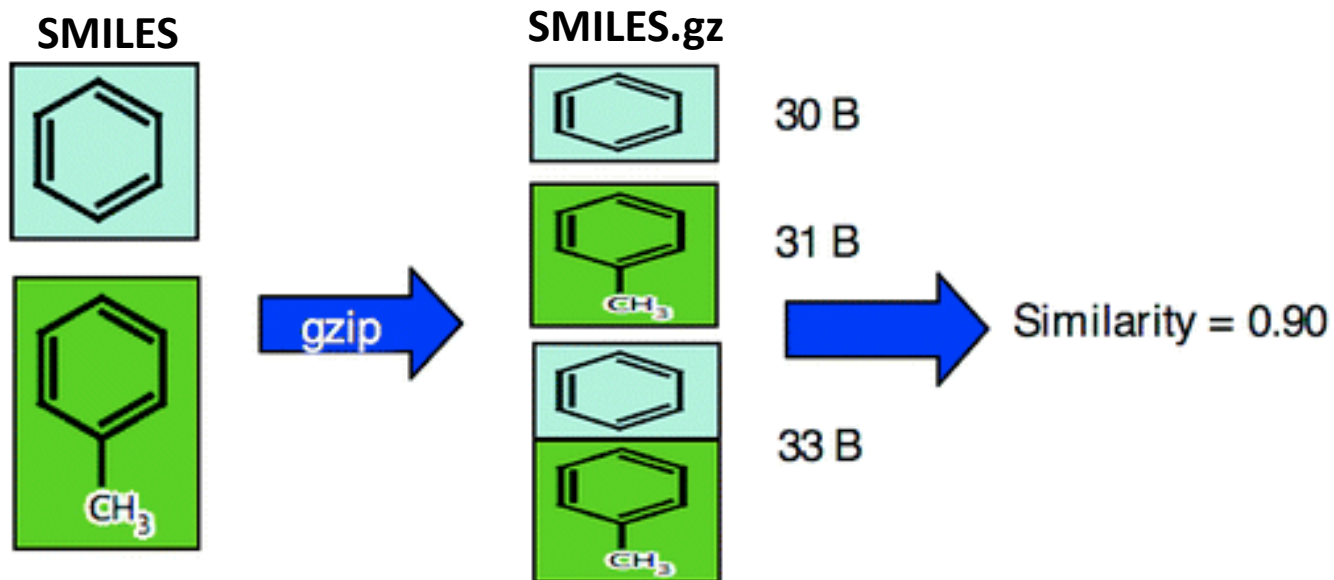


Valuing Hits

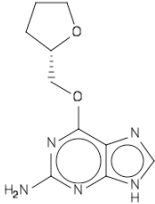
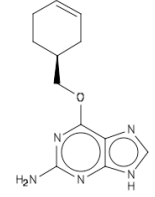
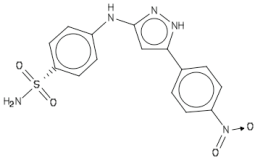
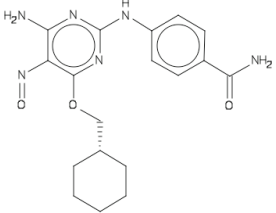
Value Judgments	Protagonist
First one in a series	Andy Good/ Tudor Oprea
First in a series=1, second= ½, etc..	Bob Clark
How quickly to X% of hits	Friesner
How quickly to M series	Christopher Bayly
Weight by information content	Me

Similarity and Compression

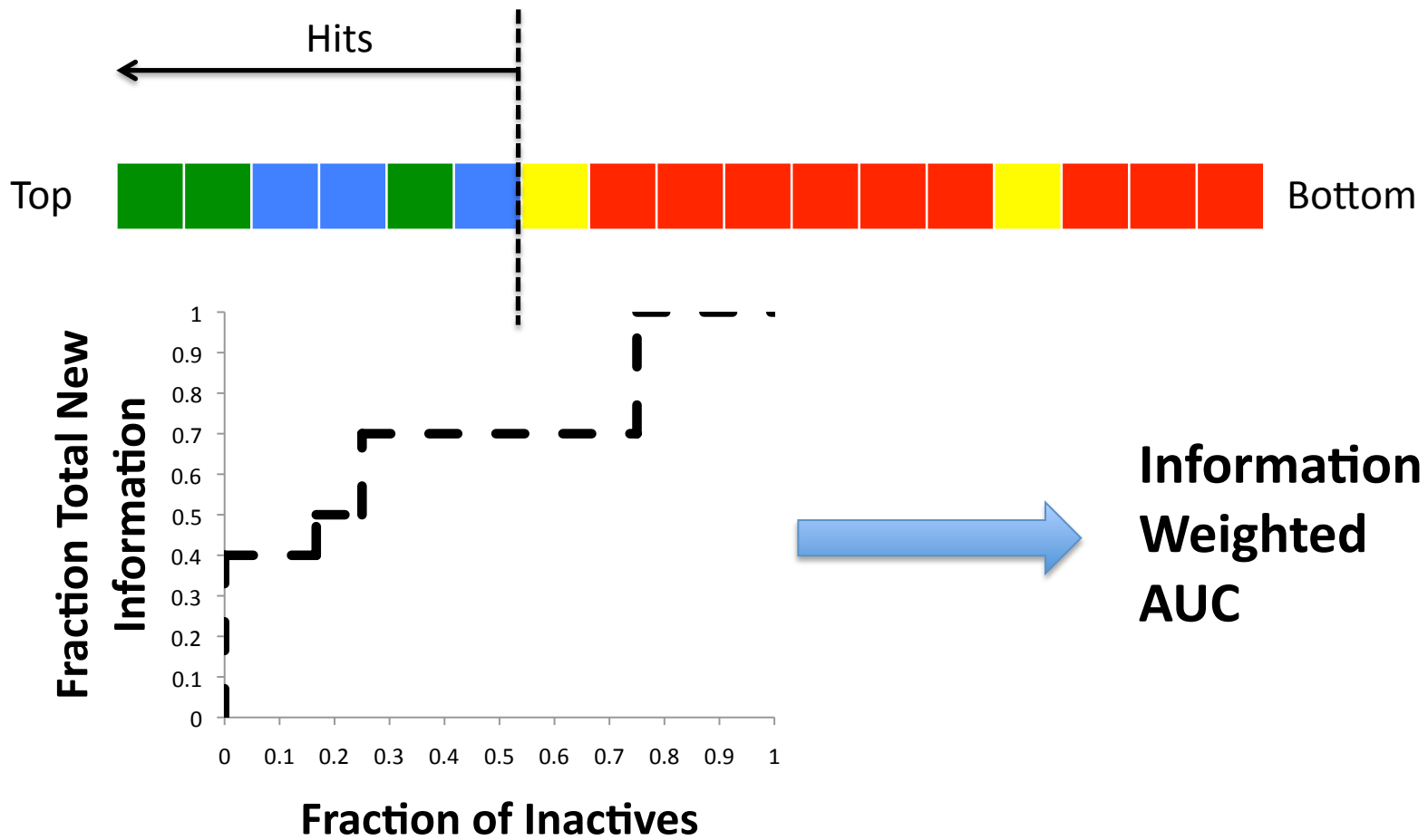
Melville, J.L., Riley, J.F. & Hirst, J.D., *JCIM*, 47, 25-33 (2007)



Novelty Estimation

Molecule		SMILES	CONCATENATED SMILES	GZIP SIZE	Added Bits
Known Active		<chem>c1[nH]c2c(n1)c(nc(n2)N)OC[C@@H]3CCCO3</chem>	<chem>c1[nH]c2c(n1)c(nc(n2)N)OC[C@@H]3CCCO3</chem>	29	
First Hit		<chem>c1[nH]c2c(n1)c(nc(n2)N)OC[C@H]3CCCC=C3</chem>	<chem>c1[nH]c2c(n1)c(nc(n2)N)OC[C@@H]3CCCO3.c1[nH]c2c(n1)c(nc(n2)N)OC[C@H]3CCCC=C3</chem>	36	7
Second Hit		<chem>c1cc(ccc1c2cc(n[nH]2)Nc3ccc(cc3)[S@@](=O)(=O)N)[N+](=O)[O-]</chem>	<chem>c1[nH]c2c(n1)c(nc(n2)N)OC[C@@H]3CCCO3.c1[nH]c2c(n1)c(nc(n2)N)OC[C@H]3CCCC=C3.c1cc(ccc1c2cc(n[nH]2)Nc3ccc(cc3)[S@@](=O)(=O)N)[N+](=O)[O-]</chem>	74	38
Third Hit		<chem>c1cc(ccc1C(=O)N)Nc2nc(c(c(n2)OC[C@@H]3CCCCC3)N=O)N</chem>	<chem>c1[nH]c2c(n1)c(nc(n2)N)OC[C@@H]3CCCO3.c1[nH]c2c(n1)c(nc(n2)N)OC[C@H]3CCCC=C3.c1cc(ccc1c2cc(n[nH]2)Nc3ccc(cc3)[S@@](=O)(=O)N)[N+](=O)[O-].c1cc(ccc1C(=O)N)Nc2nc(c(c(n2)OC[C@@H]3CCCCC3)N=O)N</chem>	90	16

Information Retrieval



AUC and AUC-I over DUD

	AUC	Δ AUC-I
LINGOS	0.725	0.028 ± 0.016
ROCS	0.735	0.029 ± 0.022
MACCS	0.733	0.034 ± 0.017
FRED	0.685	0.057 ± 0.021

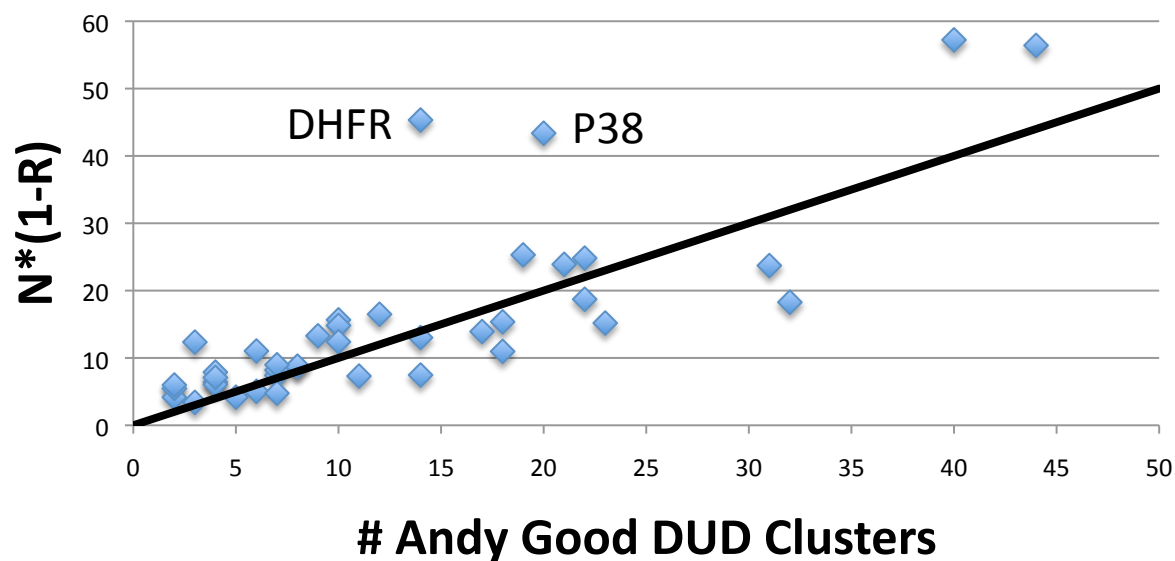
Abbott Pair Data: Information Asymmetry

A=1-10nm	AB - A	AB - B
B<10A	25.4	25.9
10A<B<100A	26.3	27.5
100A<B<10 ³ A	25.5	29.3
10 ³ A<B<10 ⁴ A	24.3	32.3
10 ⁴ A<B<10 ⁵ A	22.8	35.0
10 ⁵ A<B<10 ⁶ A	19.3	56.2

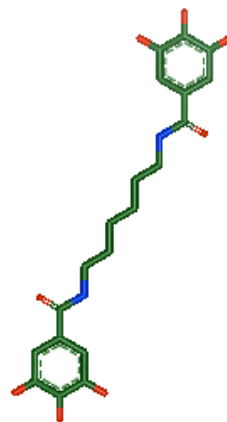
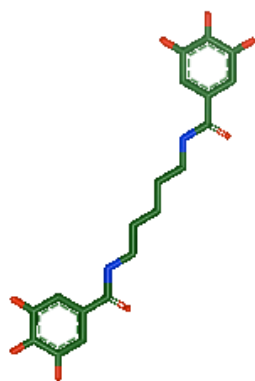
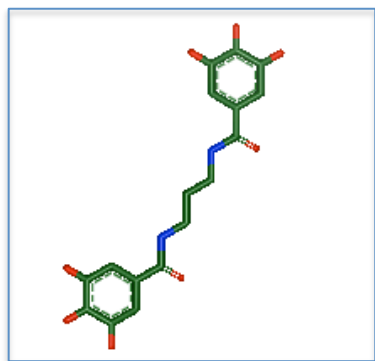
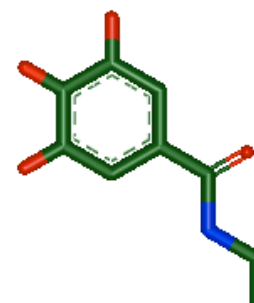
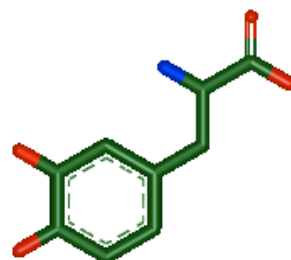
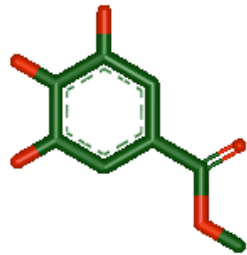
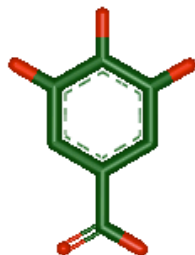
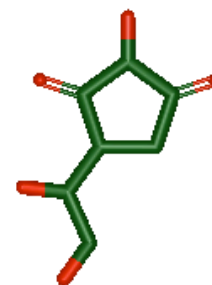
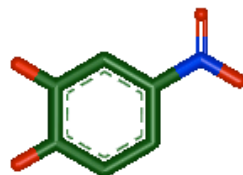
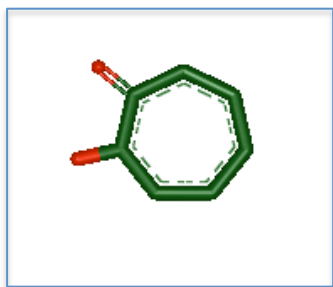
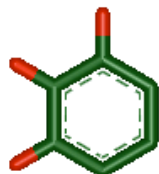
Information Redundancy

$$R = 1 - \frac{\text{Best Gzip of } N \text{ concatenated SMILES Length}}{\text{Sum of } N \text{ Gzipped SMILES Lengths}}$$

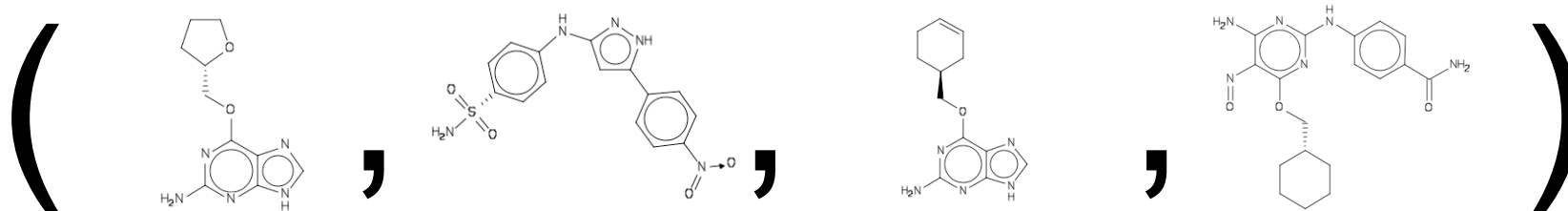
$N^*(1-R)$ = “Effective Number of Novel Molecules”



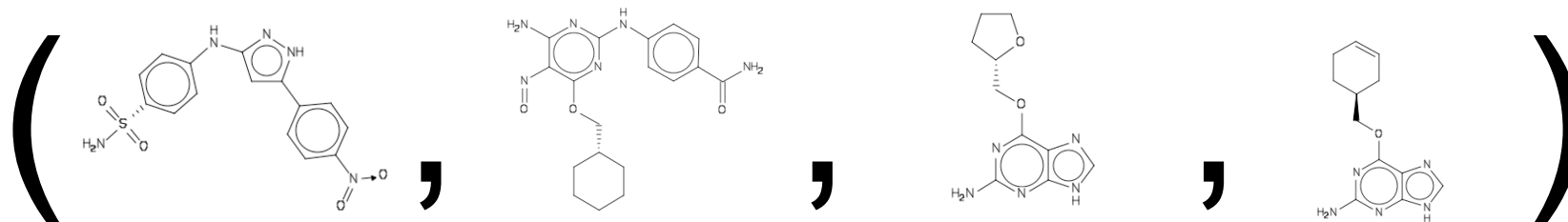
Andy Good DUD Clusters: COMT



Clustering



Reorder to Maximize compression:



Clustering by Compression by Cilibiasi, R. and Vitanyi, P. M. B,
IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 51, NO 4, APRIL 2005, 1523–1545

Final Thoughts

1) Active Ligands = A lot of information

2) Statistics on protocols treat them either as:

- Identical objects
- Objects with subjective or arbitrary value

3) Information theory gives:

- Objective similarity/ difference
- Way to measure novelty in Virtual Screening
- Interesting asymmetry in similarity
- An objective measure of cluster centers
- A novel way to cluster