optibrium

Data Visualisation: Saying it all in a bite-sized chunk

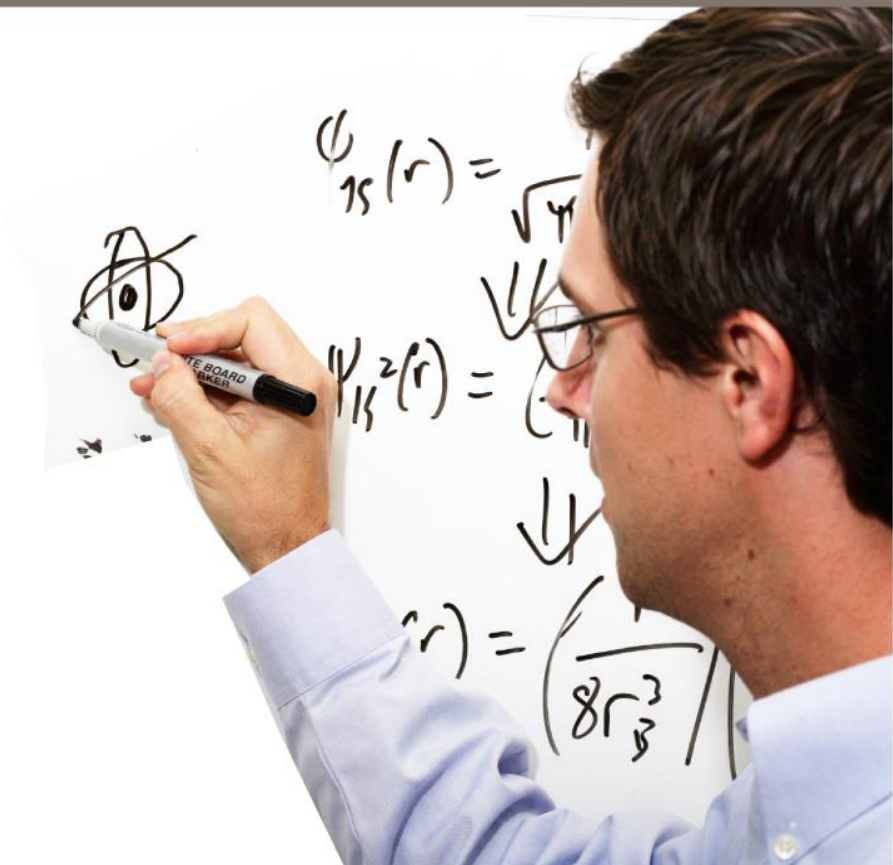
Ed Champness, Matt Segall & Peter Hunt

Overview

- Data visualisation
- Less is more?
- The many dimensions of drug discovery data
- Multi-Parameter Optimisation
- Conclusions

Data Visualisation

optibrium



Why use data visualisation?

- Visualisation systems provide visual representations of data sets designed to help people carry out tasks more efficiently
- Visualisation allows people to analyse data when they don't know exactly what they are looking for
- …when there is a need to augment human capabilities rather than replace people with computational decision-making methods

Visualization Analysis and Design – Tamara Munzner (UBC)

Use of data visualisation

When the statistics deceive us… Anscombe's quartet

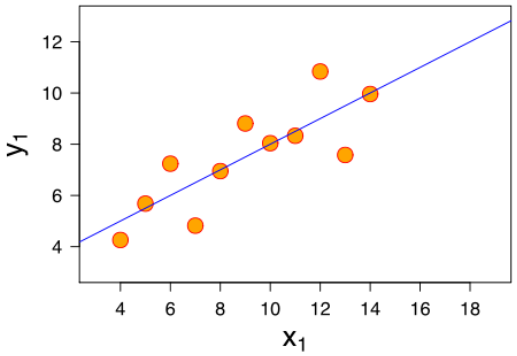
1		2		3		4	
X1	Y1	X2	Y2	X3	Y3	X4	Y4
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11		
11.0	8.33	11.0	9.26	11.0	7.81		
14.0	9.96	14.0	8.10	14.0	8.84		
6.0	7.24	6.0	6.13	6.0	6.08		
4.0	4.26	4.0	3.10	4.0	5.39		
12.0	10.84	12.0	9.13	12.0	8.15		
7.0	4.82	7.0	7.26	7.0	6.42		
5.0	5.68	5.0	4.74	5.0	5.73		

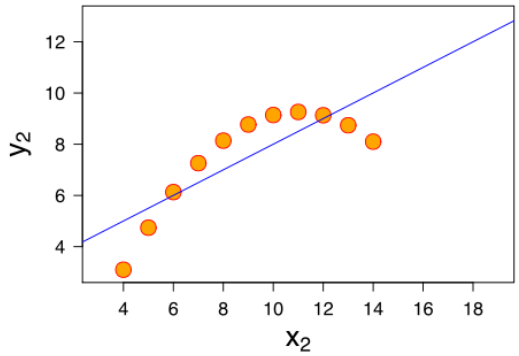
Property	Value
Mean(x)	9
Variance(x)	11
Mean(y)	7.5 (2dp)
Variance(y)	4.122 or 4.127 (3dp)
Pearson r	0.816 (3dp)
Linear regression	y = 3 + 0.5x (2dp)

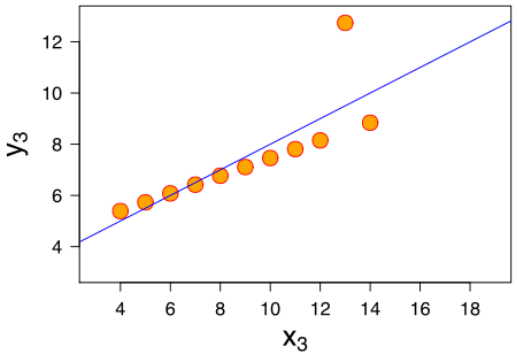
en.wikipedia.org/wiki/Anscombe%27s_quartet

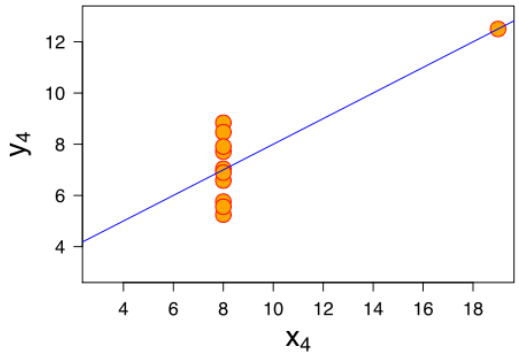
Use of data visualisation

When the statistics deceive us… Anscombe's quartet







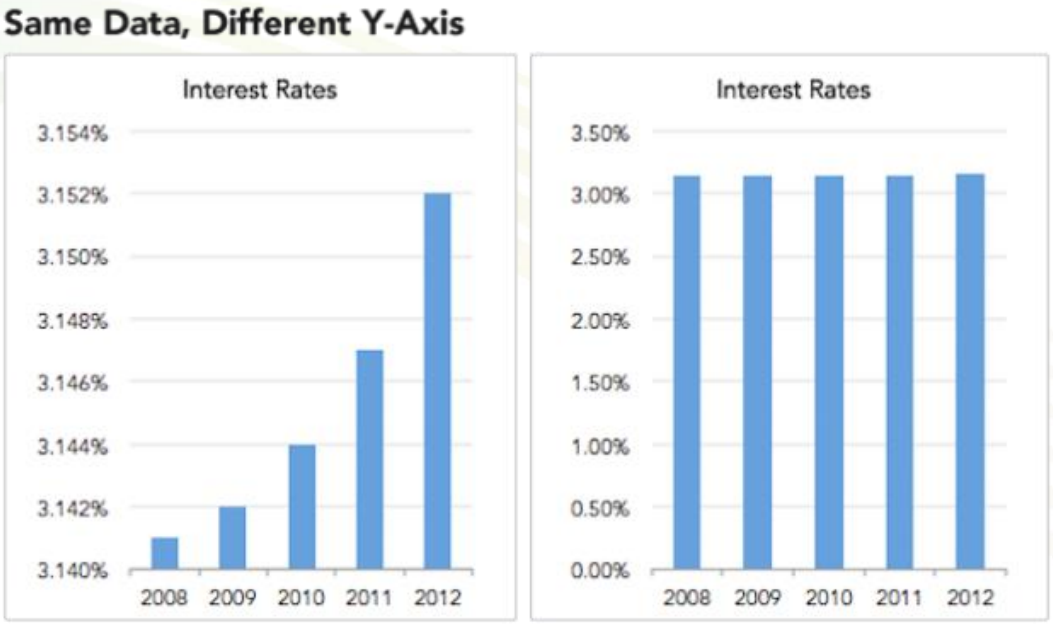


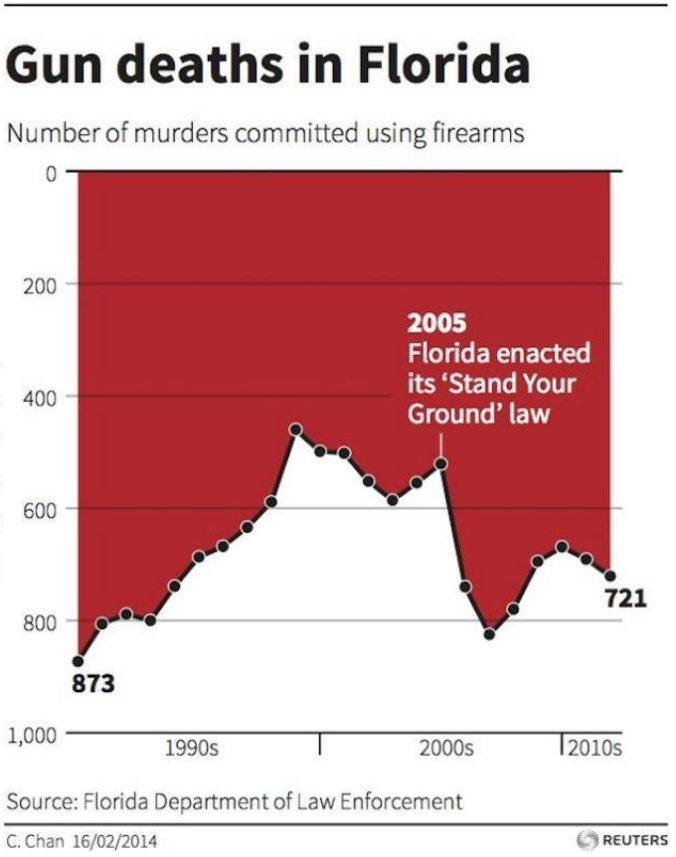
Property	Value
Mean(x)	9
Variance(x)	11
Mean(y)	7.5 (2dp)
Variance(y)	4.122 or 4.127 (3dp)
Pearson r	0.816 (3dp)
Linear regression	y = 3 + 0.5x (2dp)

en.wikipedia.org/wiki/Anscombe%27s_quartet

A few (obvious?) do's and don'ts

- Don't distort the data

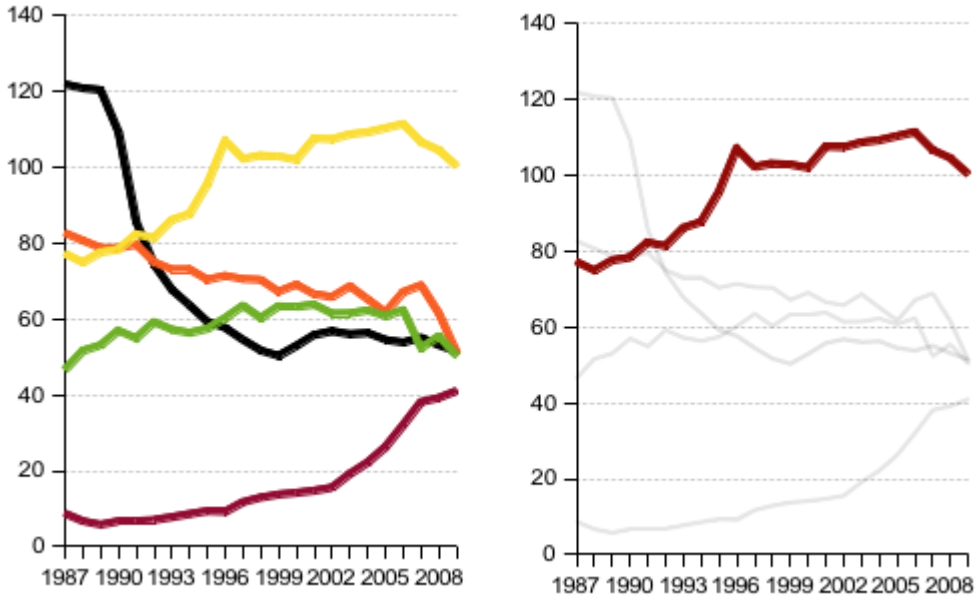




Ravi Parikh - Heap Analytics: gizmodo.com/how-to-lie-with-data-visualization-1563576606

A few (obvious?) do's and don'ts

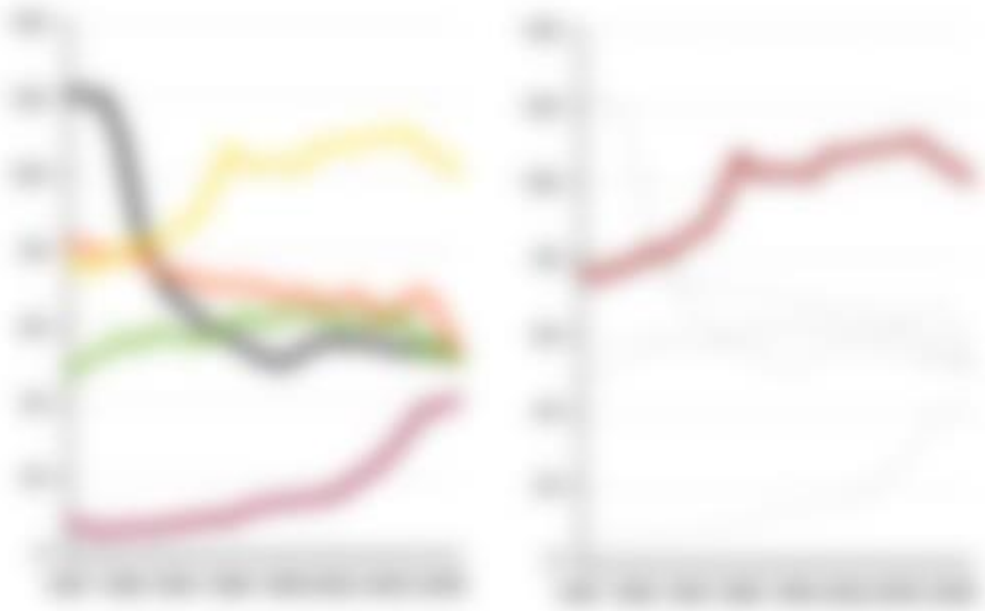
- Don't distort the data
  - But do simplify the less important information



https://www.vis4.net/blog/posts/doing-the-line-charts-right/

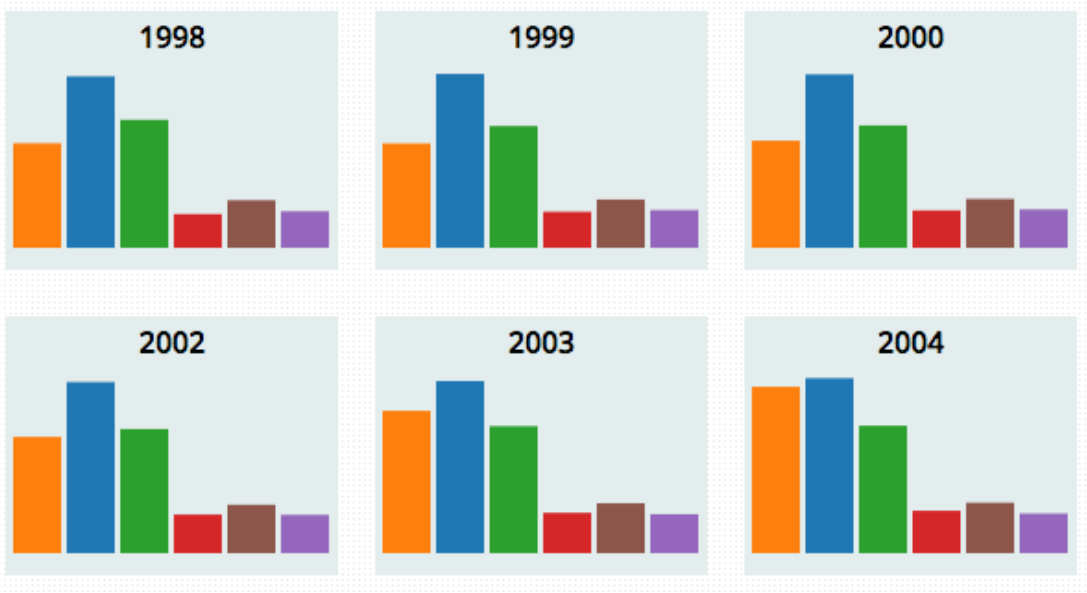
A few (obvious?) do's and don'ts

- Don't distort the data
  - But do simplify the less important information
- Consider the squint test



A few (obvious?) do's and don'ts

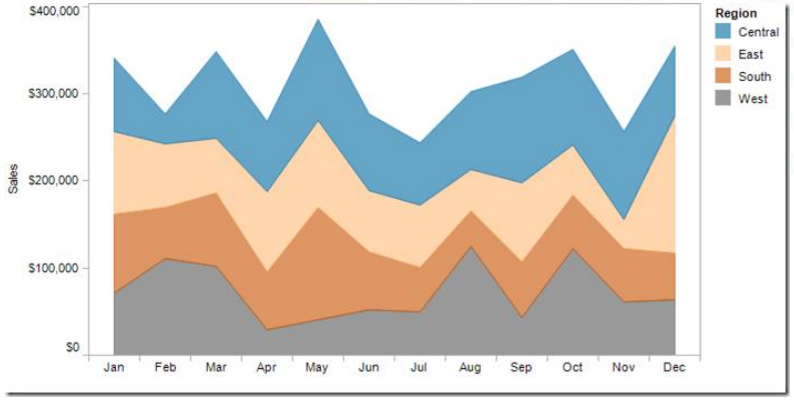
- Don't distort the data
  - But do simplify the less important information
- Consider the squint test
- Use consistent layout



http://vallandingham.me/small_multiples_with_details.html

A few (obvious?) do's and don'ts

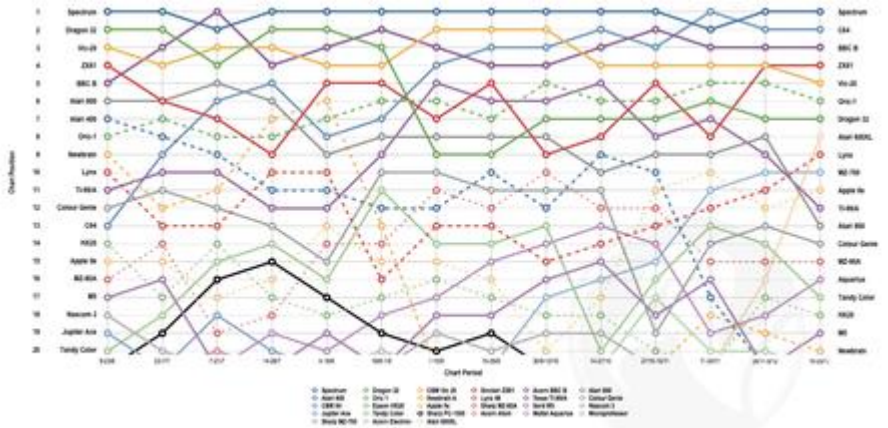
- Don't distort the data
  - But do simplify the less important information
- Consider the squint test
- Use consistent layout
- Don't force the reader to calculate differences



http://www.vizwiz.com/2012/10/stacked-area-chart-vs-line-chart-great.html

A few (obvious?) do's and don'ts

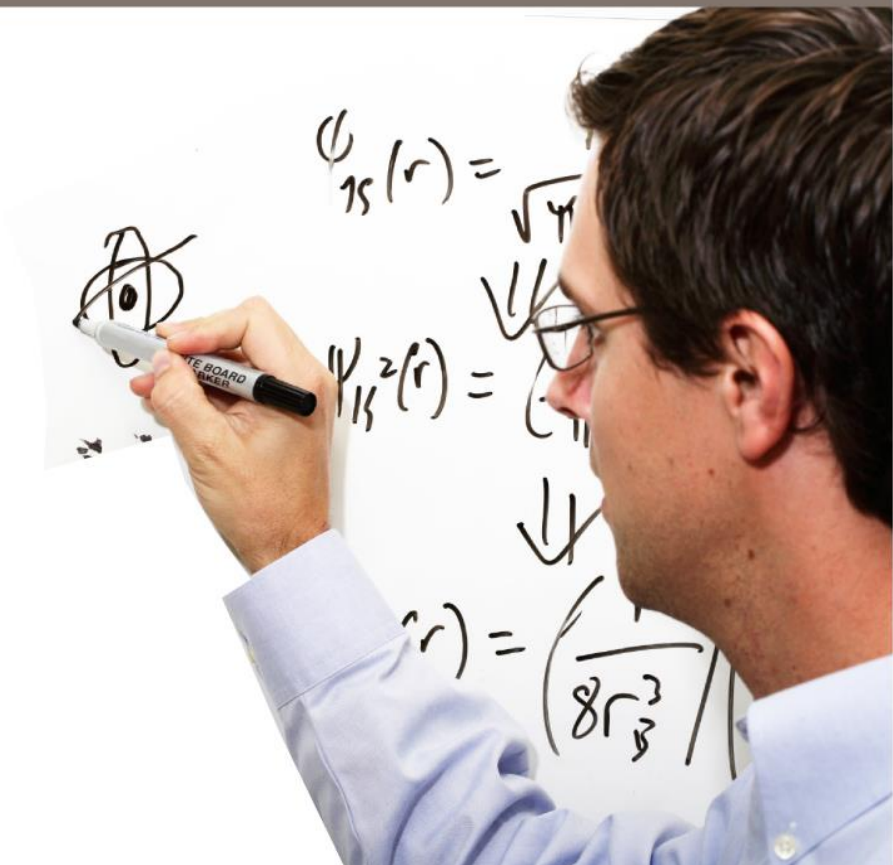
- Don't distort the data
  - But do simplify the less important information
- Consider the squint test
- Use consistent layout
- Don't force the reader to calculate differences
- Don't overload the chart



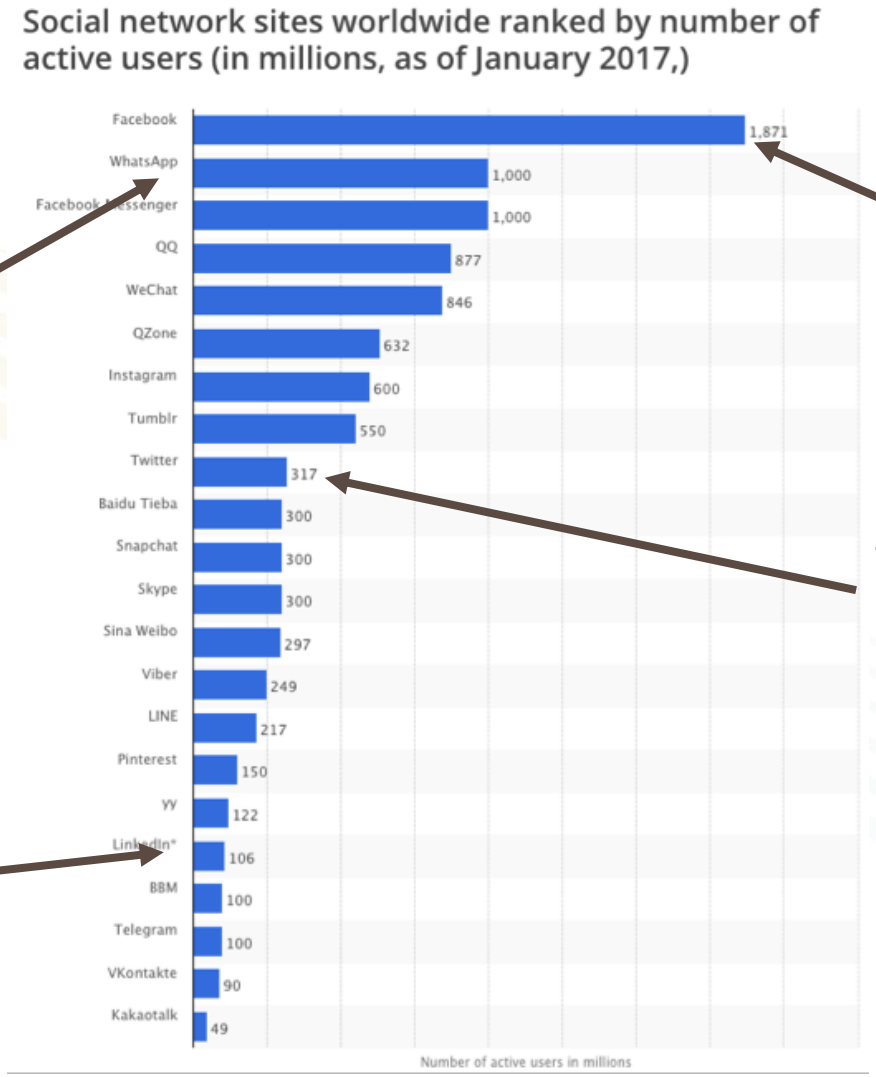
http://junkcharts.typepad.com – Kaiser Fung

Less is more?

optibrium



The predominance of social media



WhatsApp
1,000

Facebook
1,871

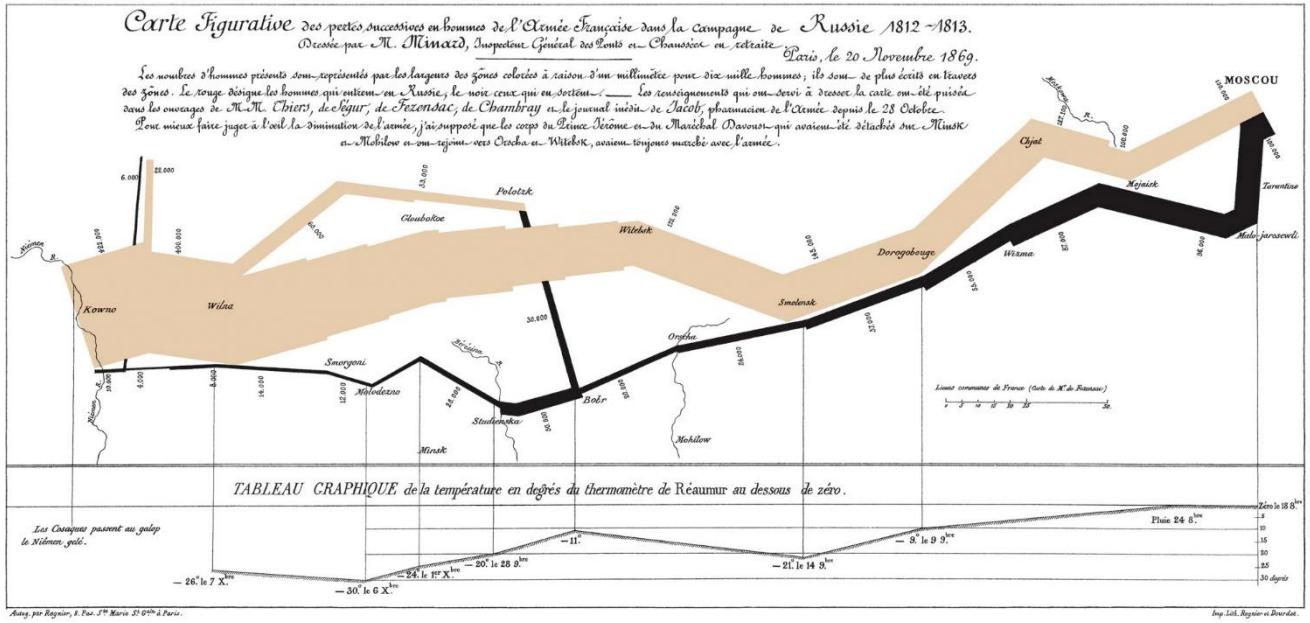
Twitter
317

LinkedIn
106

www.smartinsights.com

Designing a Visualisation

- Tufte: What makes for such graphical elegance? … Good design has two key elements: Graphical elegance is often found in simplicity of design and complexity of data.

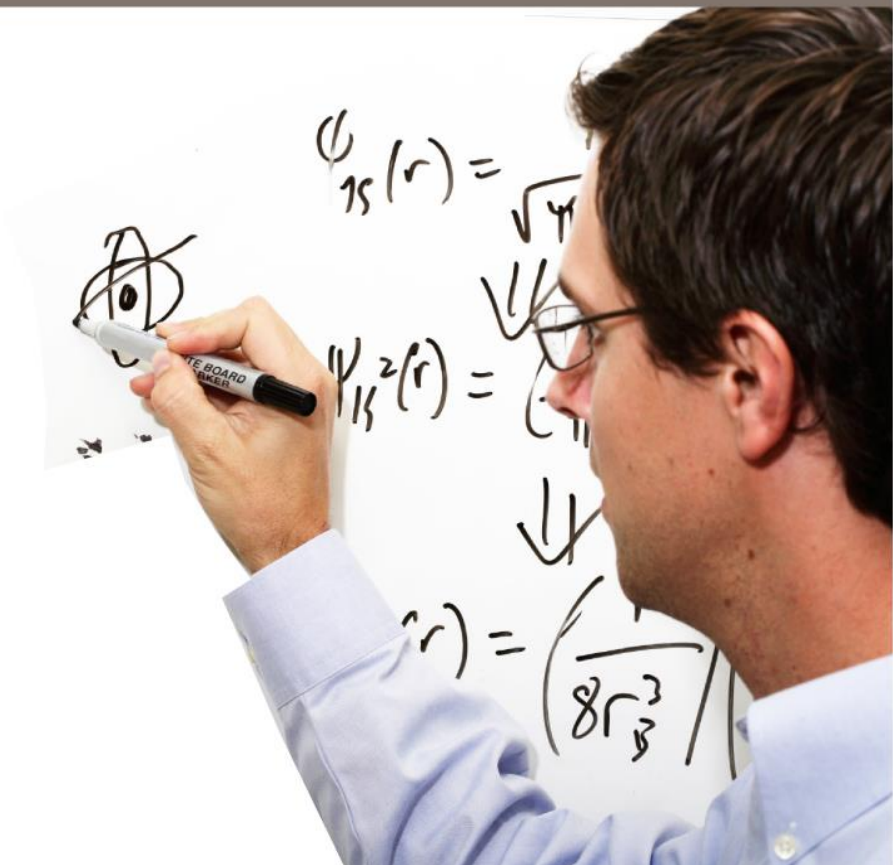


- Minard's graphic of Napoleon's Russian march and retreat

The Visual Display of Quantitative Information – Edward Tufte

The many dimensions of drug discovery data

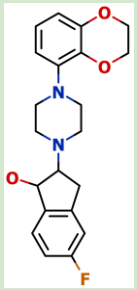
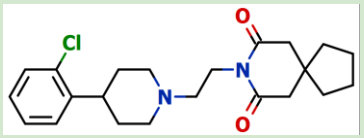
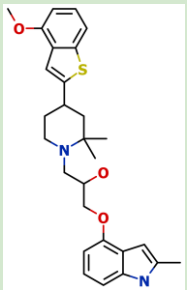
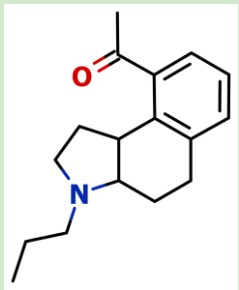
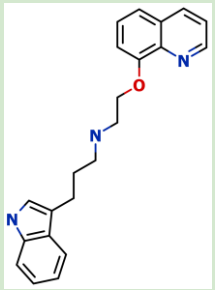
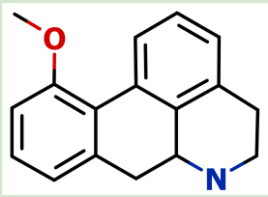
optibrium



Drug discovery example

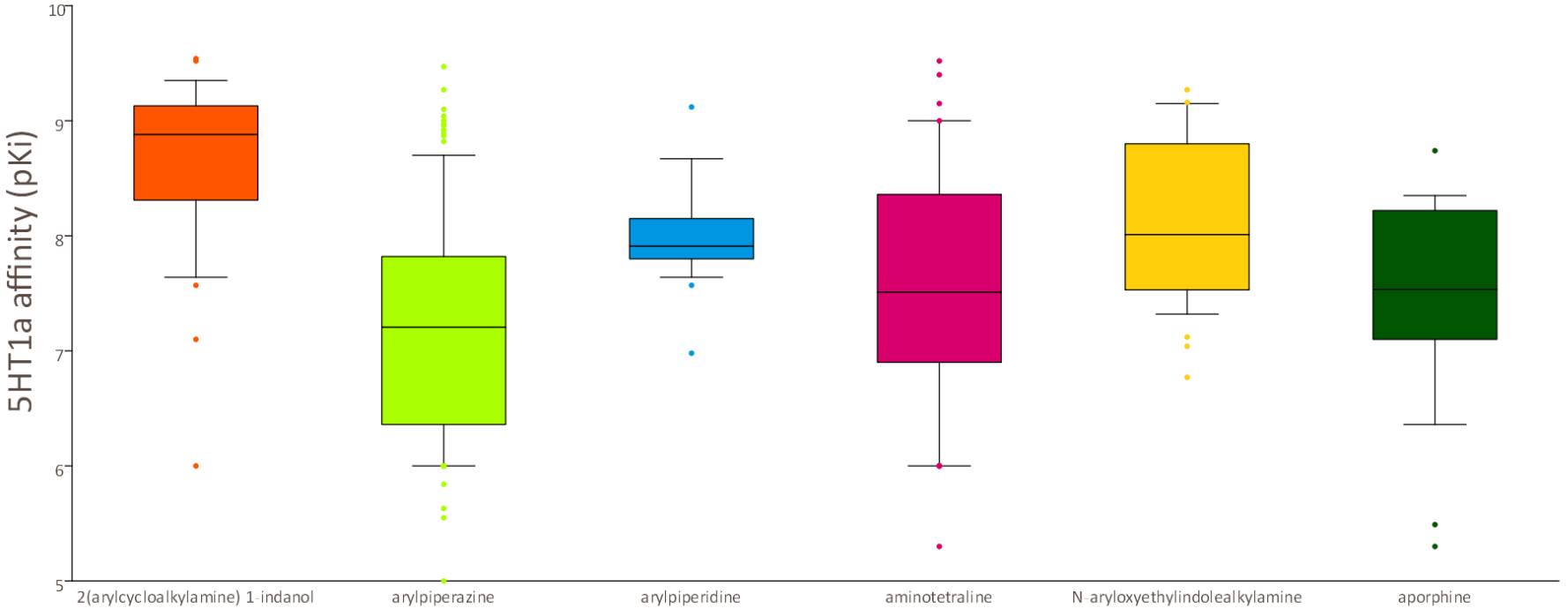
Let's looks at some drug discovery data:

- Library of 264 5HT1a compounds
- Measured potencies and other ADME/physicochemical properties
- Six different chemotypes:

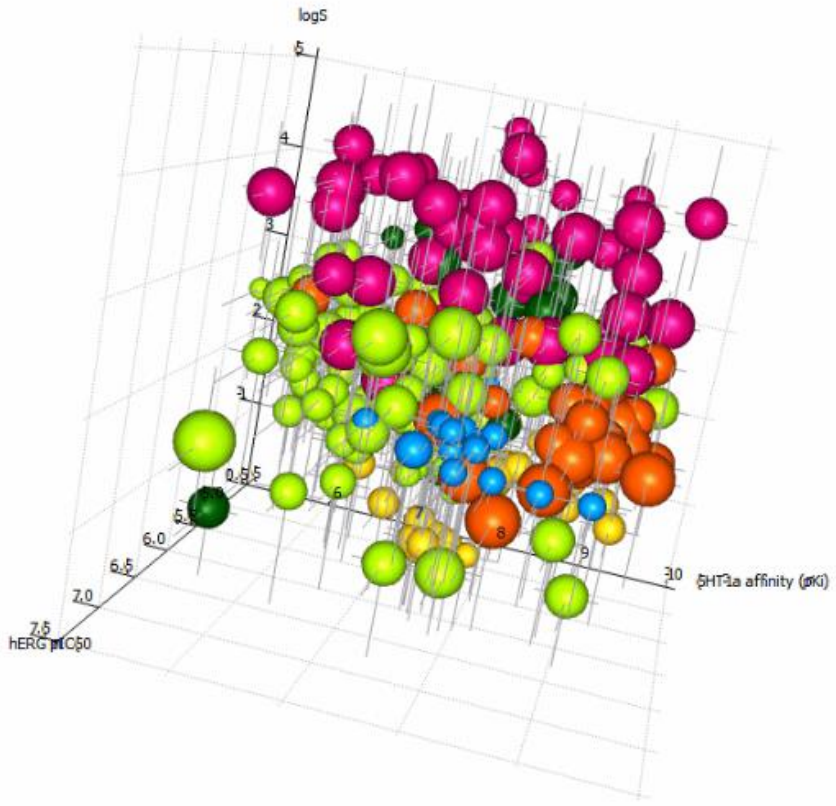
2(arylcycloalkylamine) 1-indanols (27)	Arylpiperazines (120)	Arylpiperidines (17)	Aminotetralines (51)	N-aryloxyethylindolealkylamines (29)	Aporphines (20)
					

Let's think about how we might prioritise these…

Potencies of chemotypes



…but we have many properties



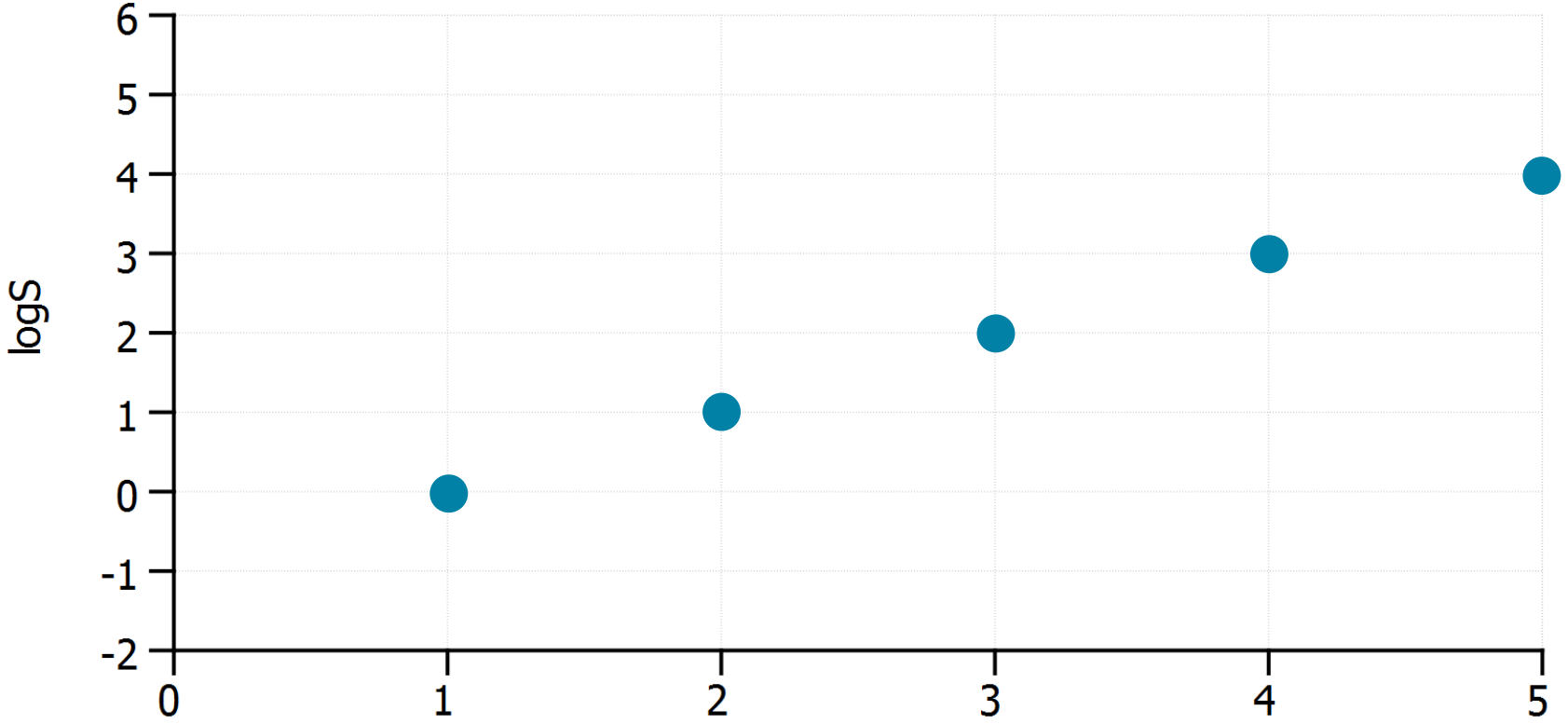
Data in drug discovery

- What's certain?
  - We know some simple properties of our compounds
- What's not so certain?
  - *In vitro*/*In vivo* measurements
    - experimental variability
  - *In silico* predictions
    - statistical error
  - Inference/translation

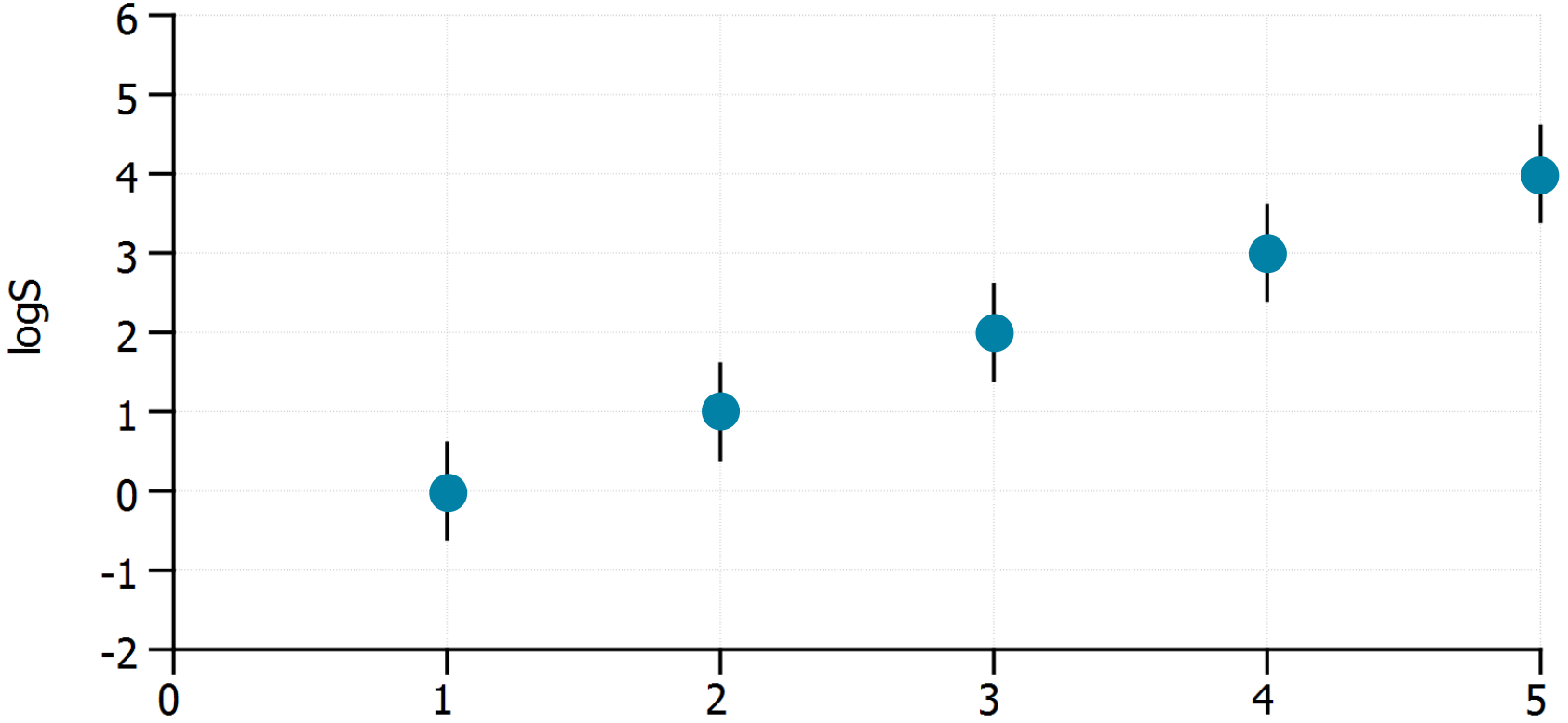
Uncertain data

- A good RMSE for logS (solubility) is 0.6
- Assuming normal distribution a logS value of 2 (that's 100µM) means:
  - 68% of the time this represents an actual value between 1.4 and 2.6 (25µM to 400µM)
  - 95% of the time this represents an actual value between 0.8 and 3.2 (6µM to 1.6mM)
  - 99% of the time this represents an actual value between 0.2 and 3.8 (1.6µM to 6.3mM)

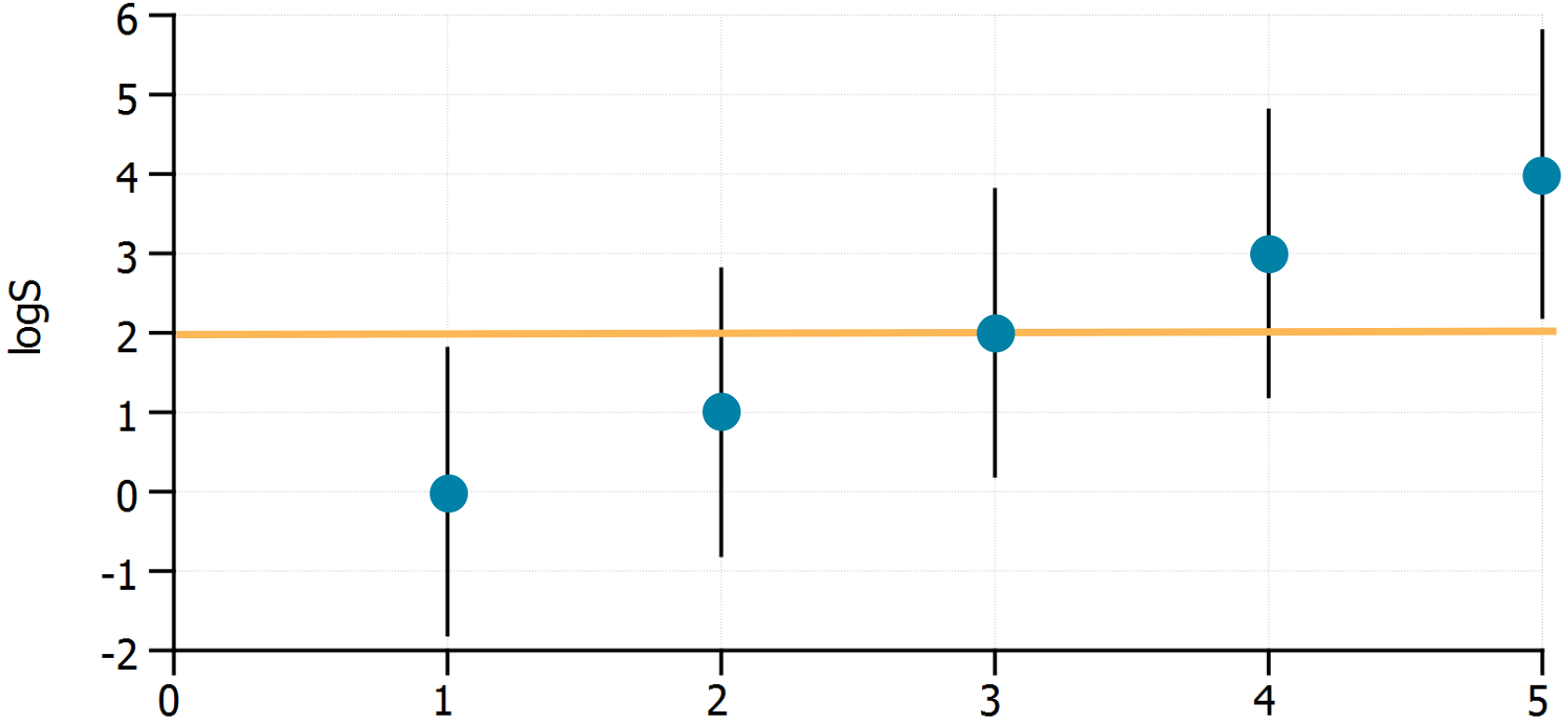
LogS – Without Error Bars



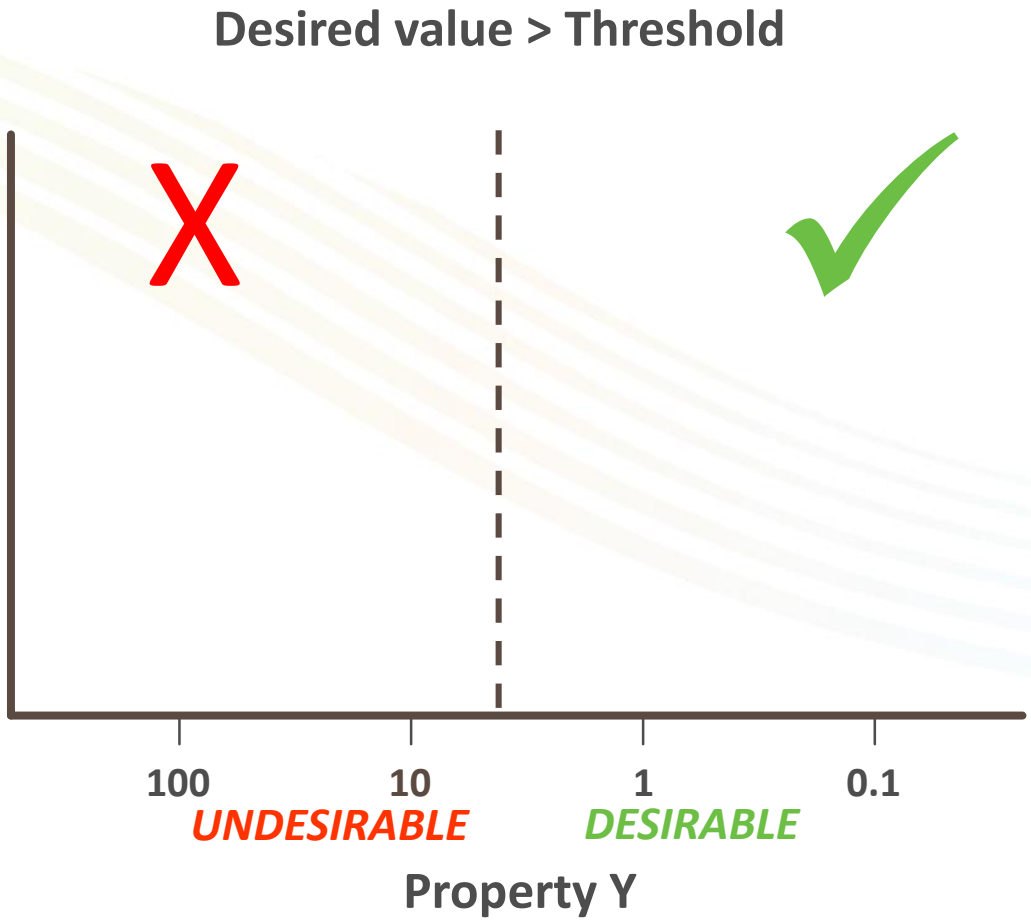
LogS – 1 Standard Deviation



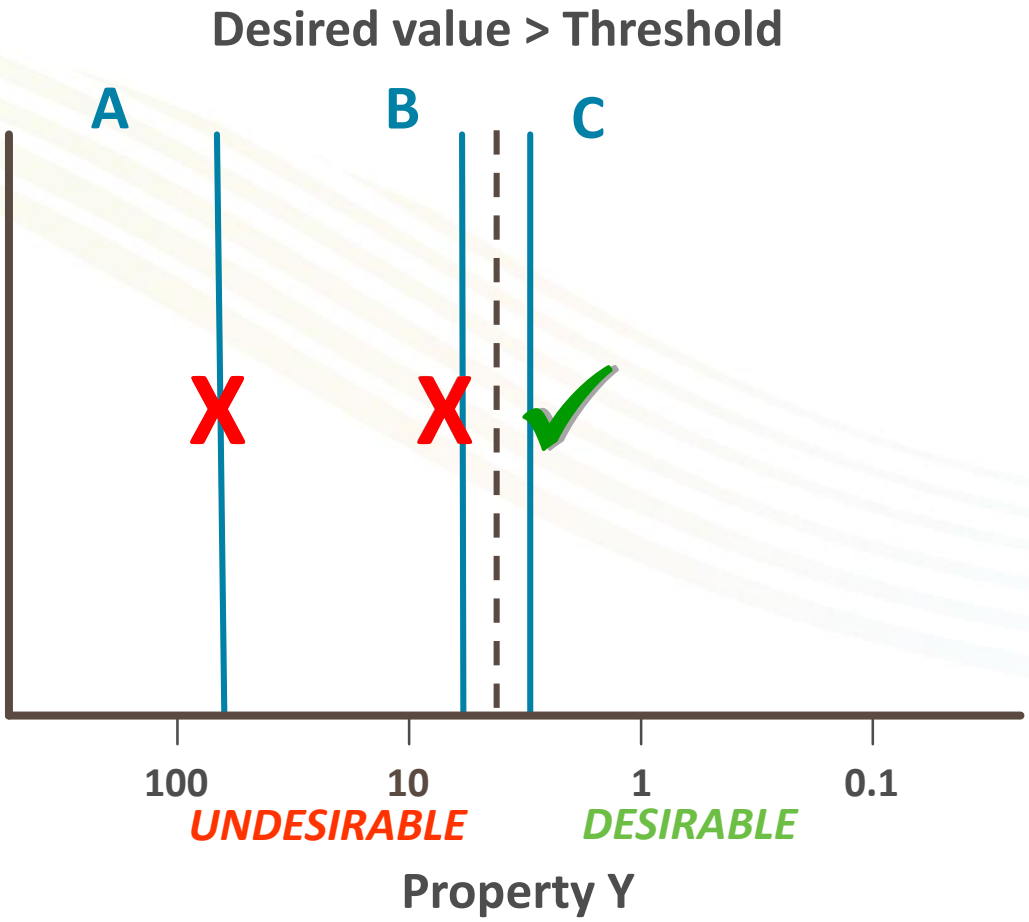
LogS – 3 Standard Deviations



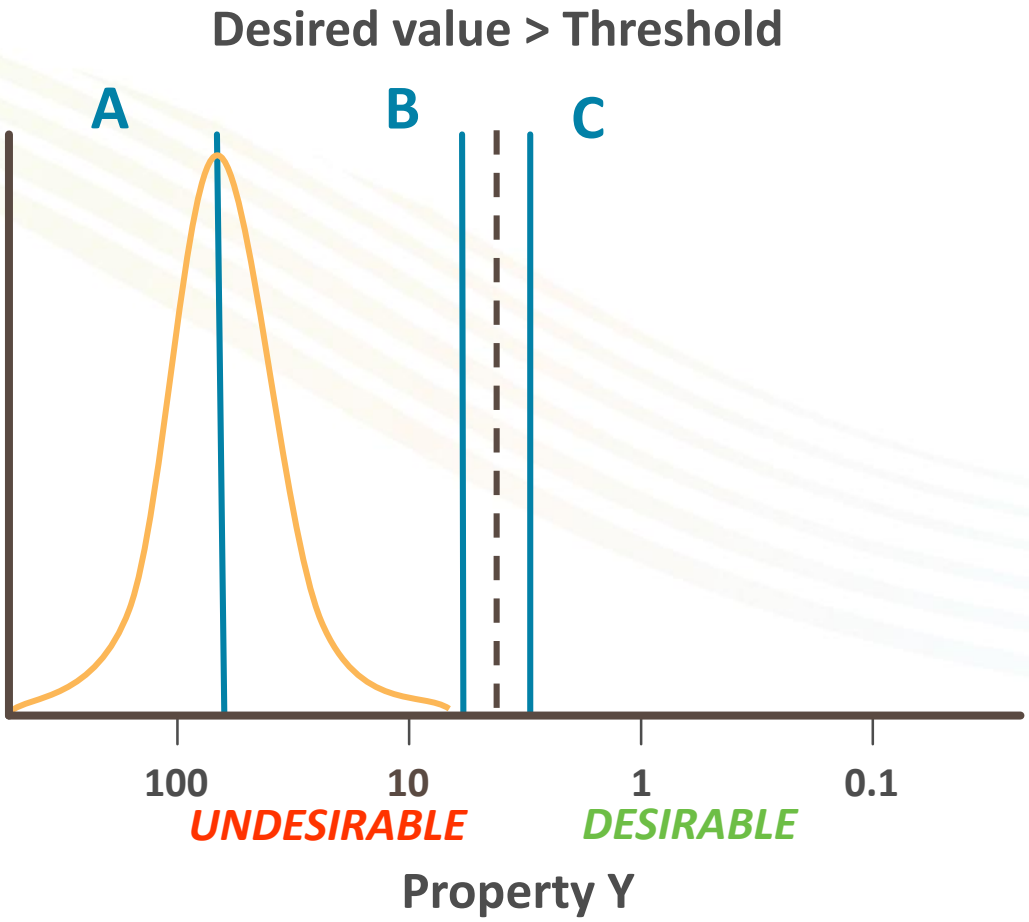
Importance of Uncertainty



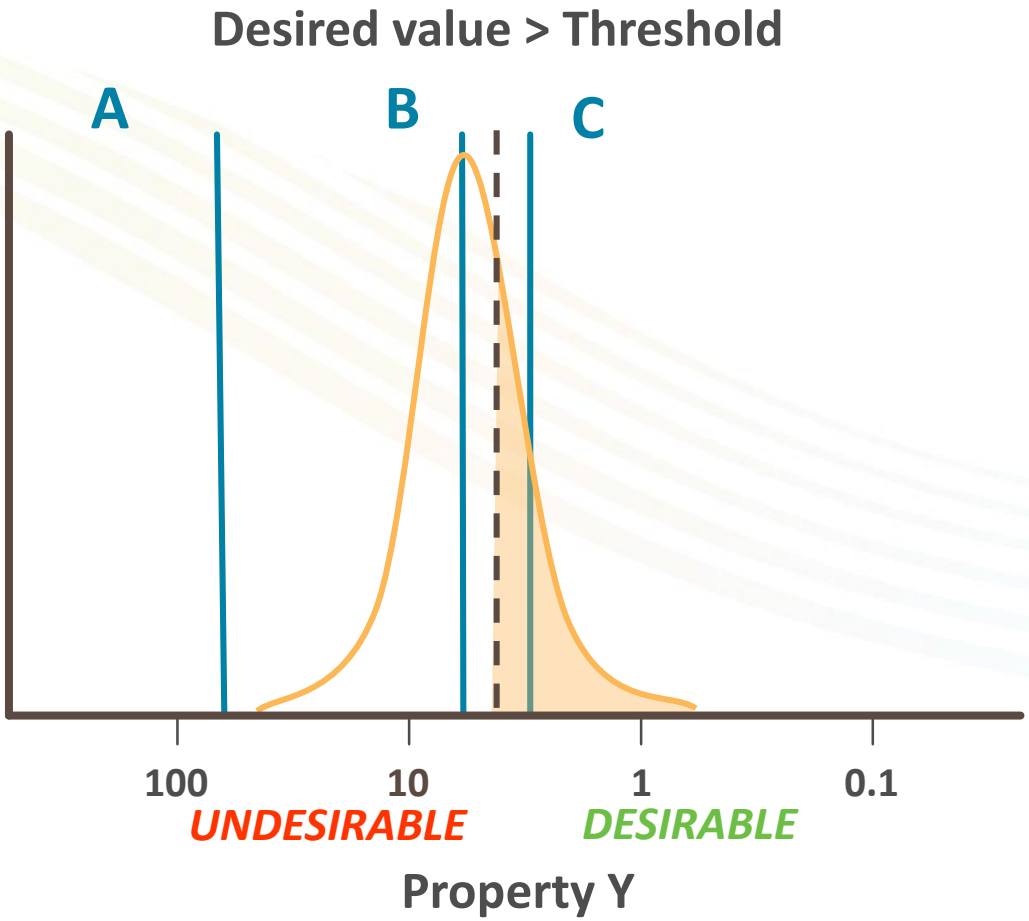
Importance of Uncertainty



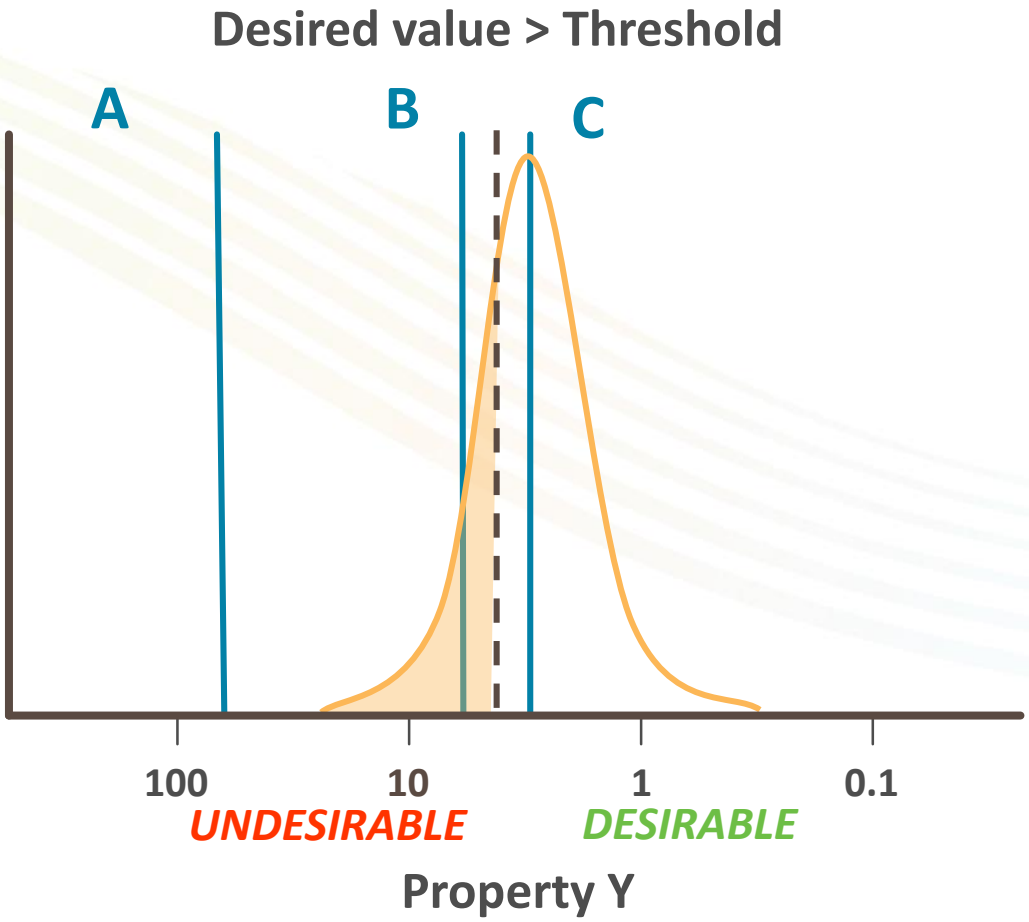
Importance of Uncertainty



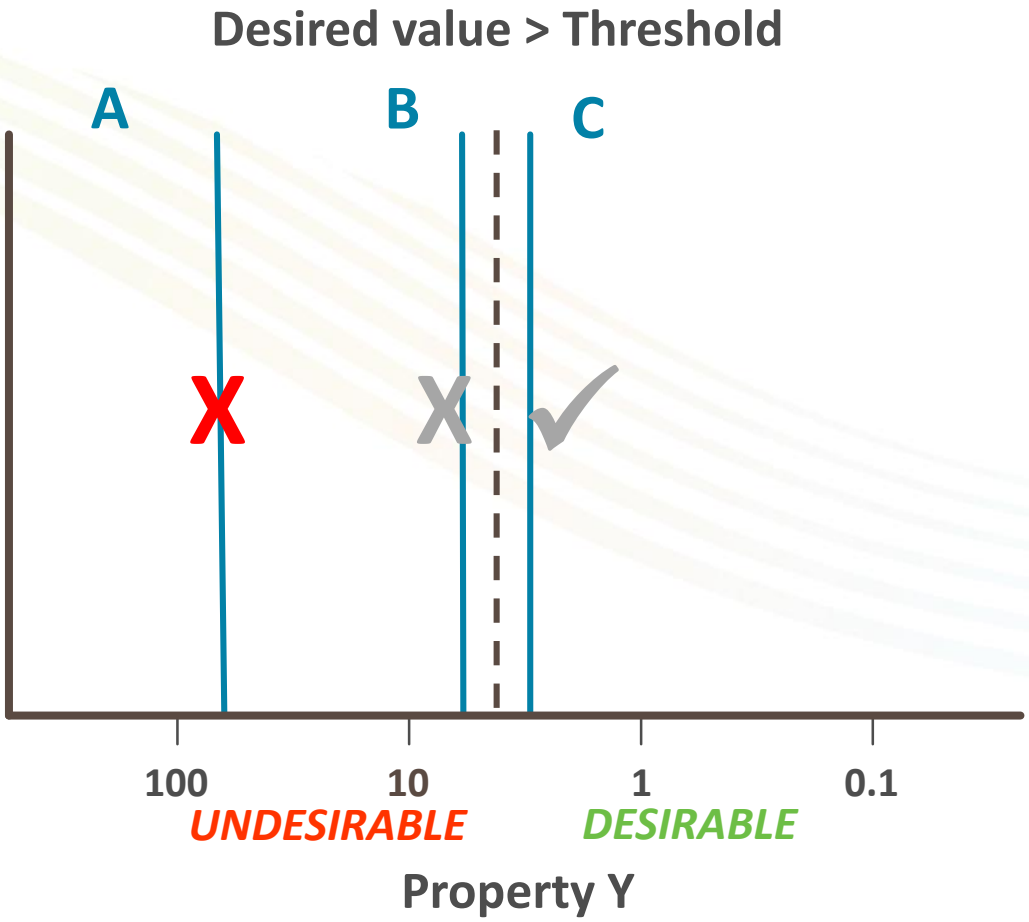
Importance of Uncertainty



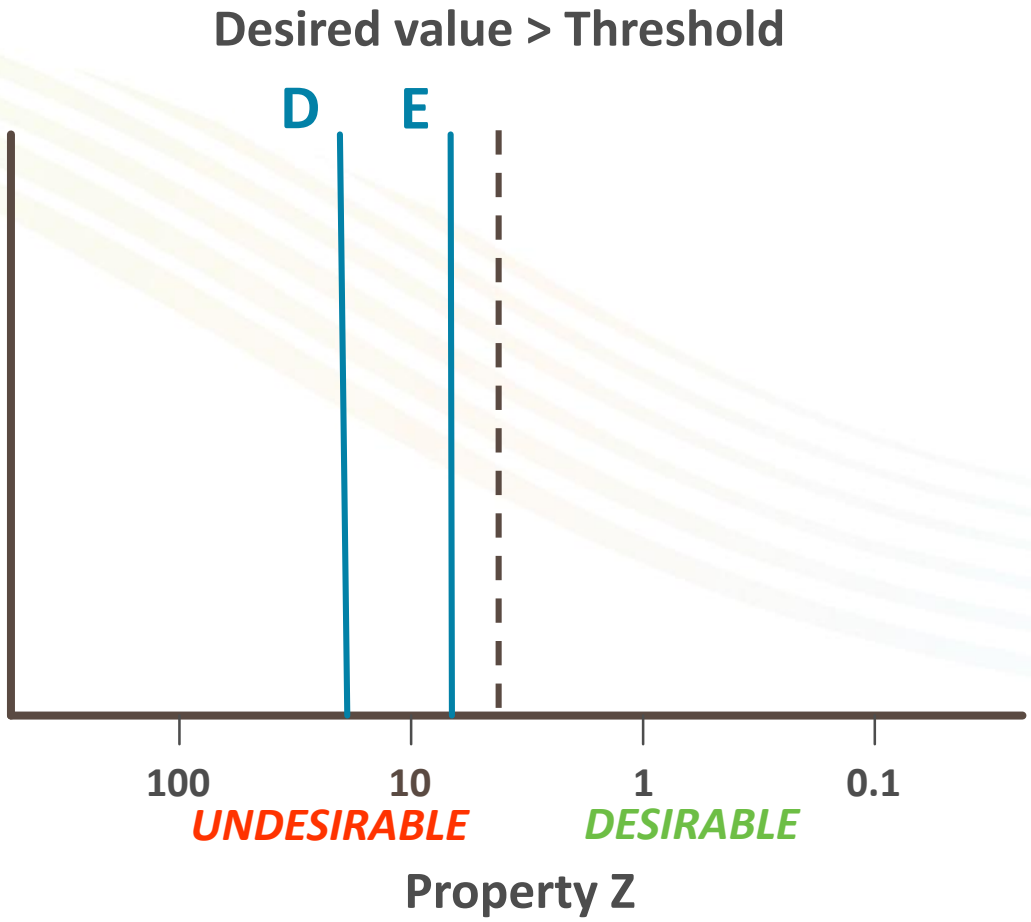
Importance of Uncertainty



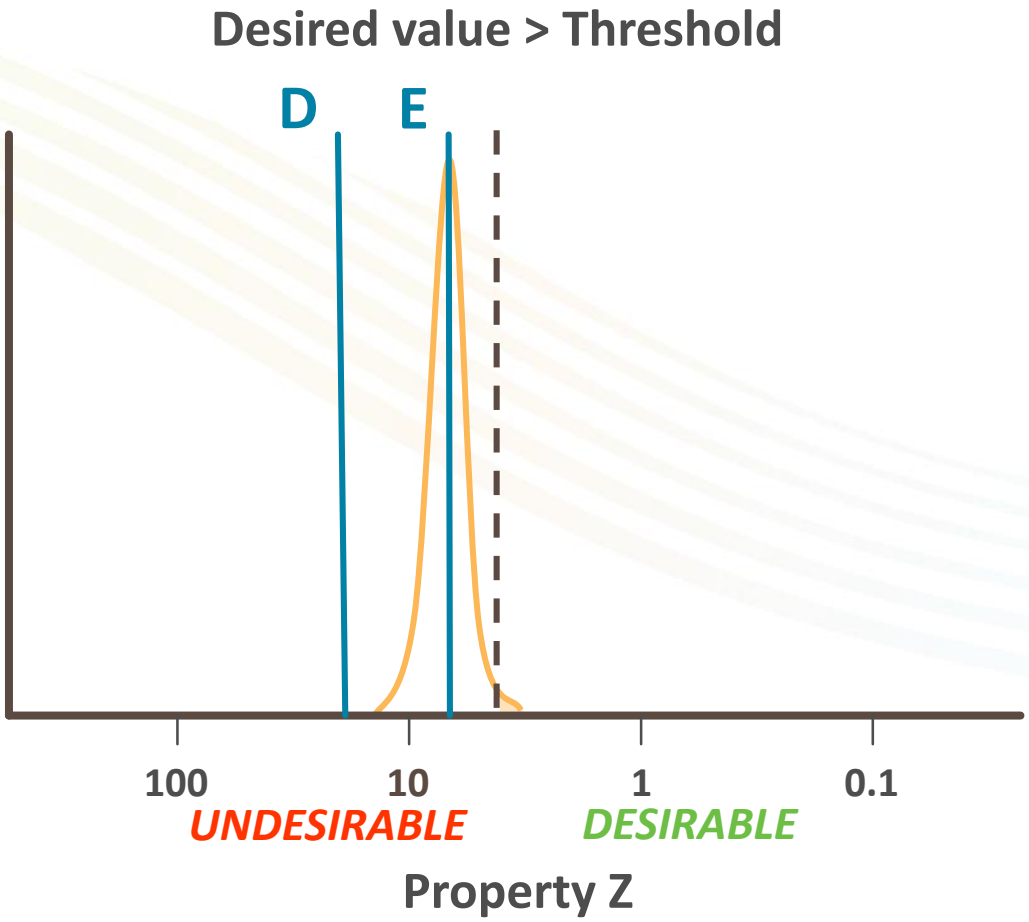
Importance of Uncertainty



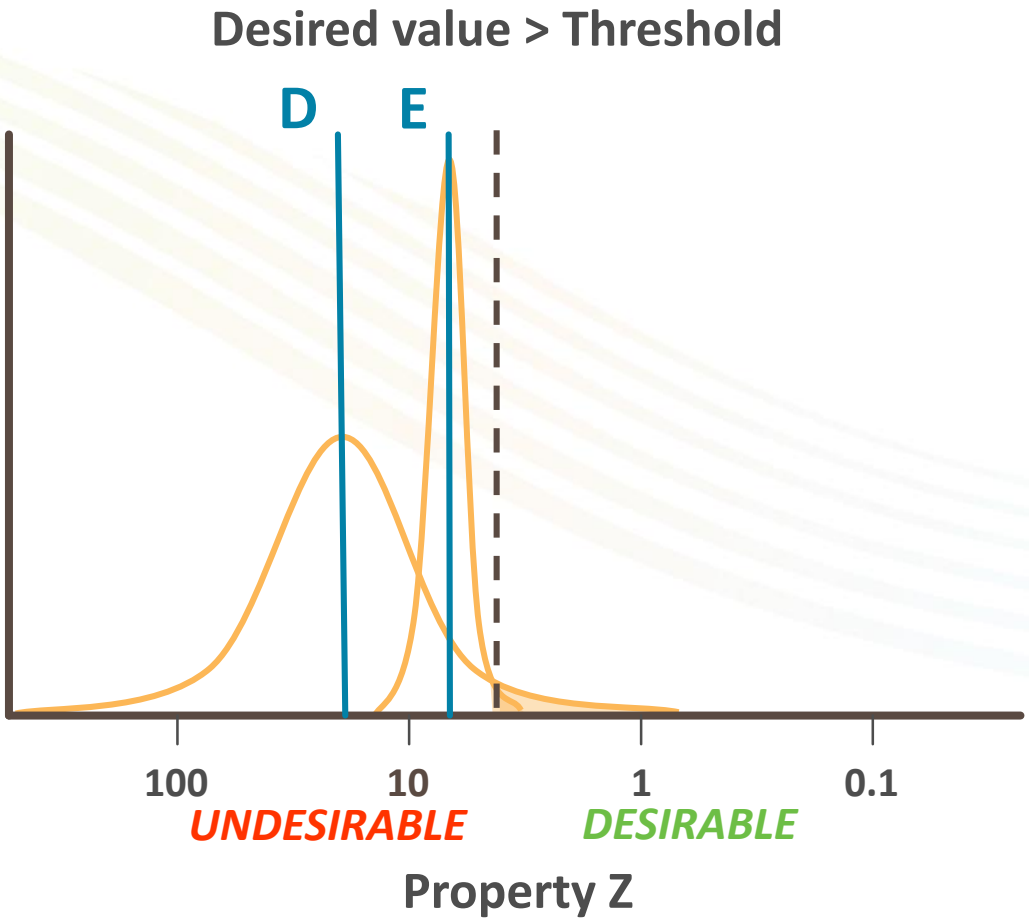
Importance of Uncertainty



Importance of Uncertainty



Importance of Uncertainty

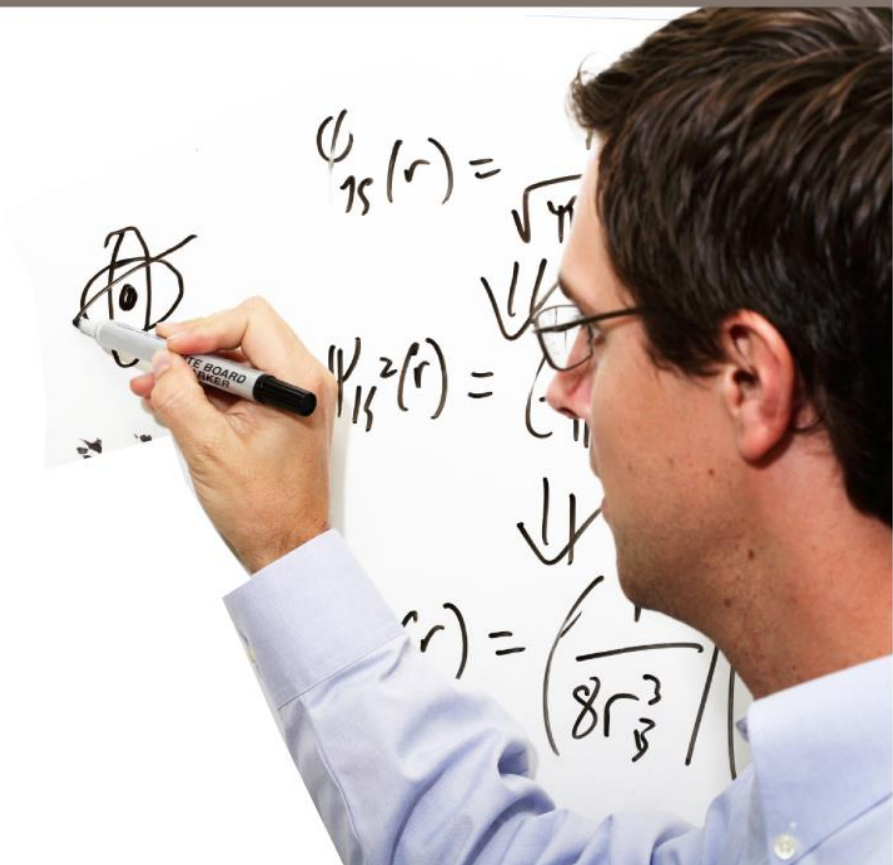


…and don't forget

- We probably have quite a few properties we need to optimise!
  - Each will have their own uncertainty
  - Each will have its own criteria we'd like to achieve
  - Each will have its own level of importance relative to the other properties

Multi-Parameter Optimisation

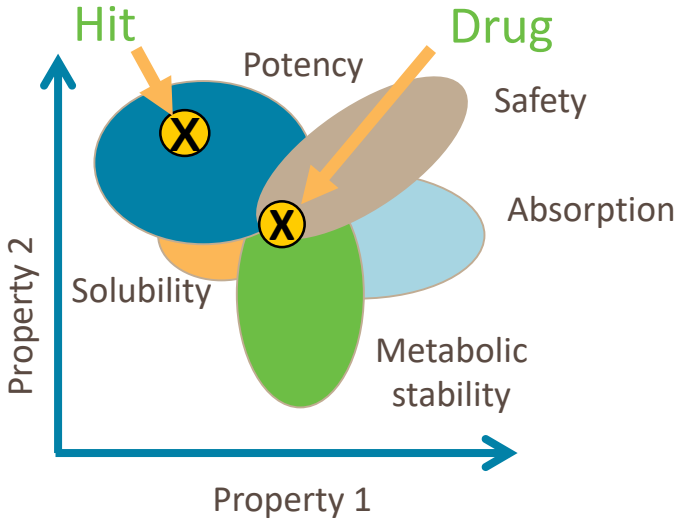
optibrium



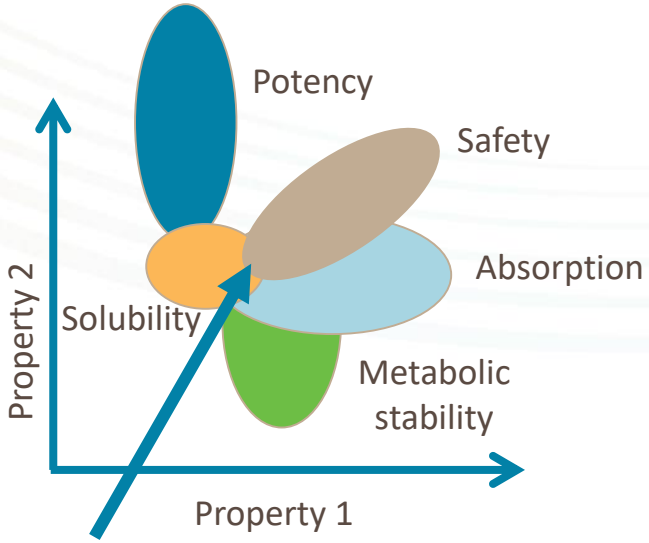
Guiding Decisions in Compound Optimisation

Multi-parameter optimisation

- Identify chemistries with an optimal **balance** of properties



- Quickly identify situations when such a balance is not possible
  - –Fail fast, fail cheap
  - –Only when **confident**
  - –Avoid **missed opportunities**



No good drug

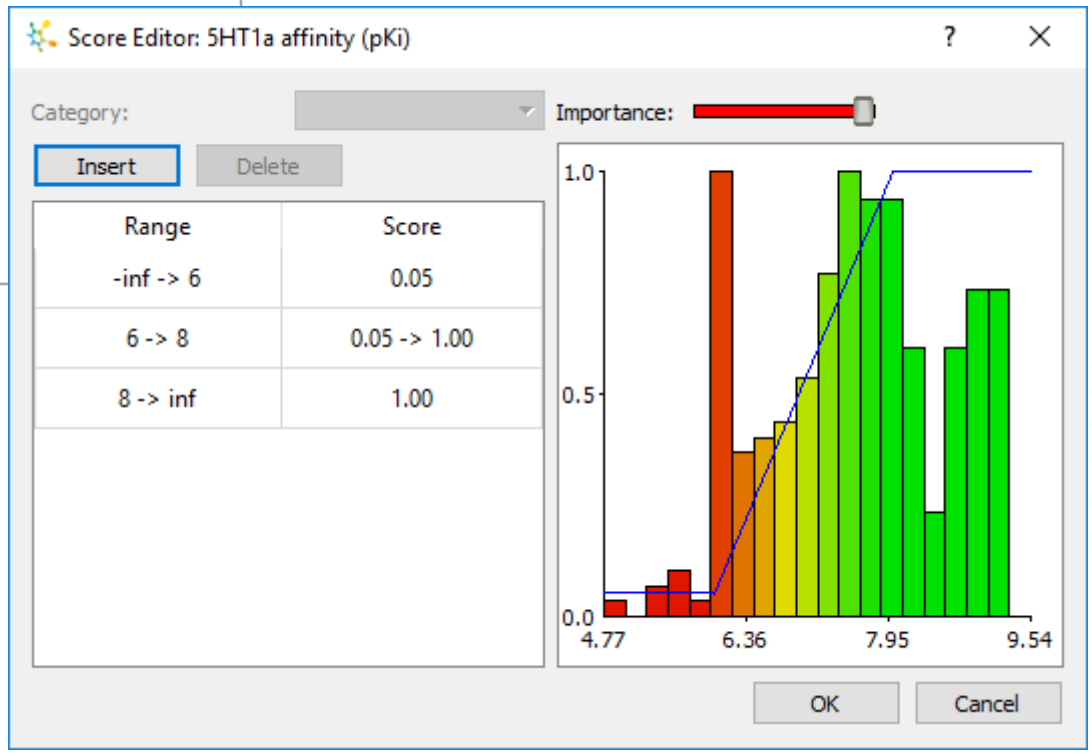
Back to our 5HT1a library

- Example criteria we might like to achieve for an ideal compound

Property	Desired value	Importance
Potency (pK_i)	> 7	High
logS (log µM)	> 1	High
Human Intestinal Absorption (category)	+	High
BBB log([brain]:[blood])	-0.2 -> 1	High
logP	0 -> 3.5	Medium
P-gp (category)	No	Medium
hERG pIC50	≤ 5	Medium
2C9 pK_i	≤ 6	Low
2D6 affinity (category)	Low/Medium	Low
Plasma protein binding (category)	Low	Low

Putting it all together (MPO):
Probabilistic Scoring* – Scoring Profile

Property	Desired Value	Importance
5HT1a affinity (pKi)	> 7	
logS	> 1	
HIA category	+	
logP	0 -> 3.5	
BBB log([brain]:[blood])	-0.2 -> 1	
BBB category	+	
P-gp category	no	
hERG pIC50	⩽ 5	
2C9 pKi	⩽ 6	
2D6 affinity category	low medium	
PPB90 category	low	



* Segall *et al.* (2009) Chem. & Biodiv. **6** p. 2144

Multi-parameter Optimisation
Probabilistic Scoring*

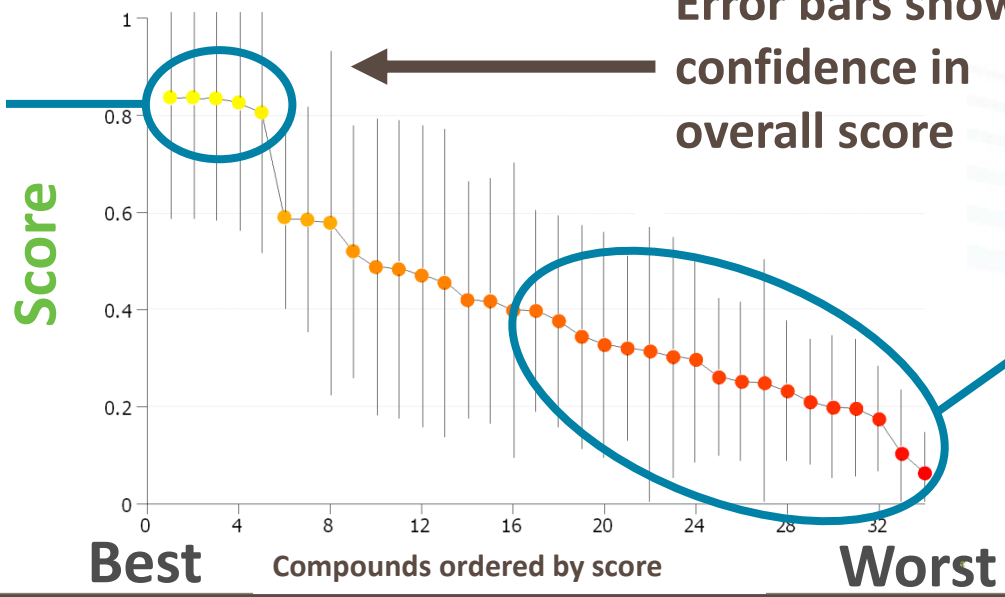
- **Property data**
  - Experimental or predicted
- **Criteria for success**
  - Relative importance
- **Uncertainties in data**
  - Experimental or statistical

- **Score (Likelihood of Success)**
- **Confidence in score**

Data do not separate these as error bars overlap

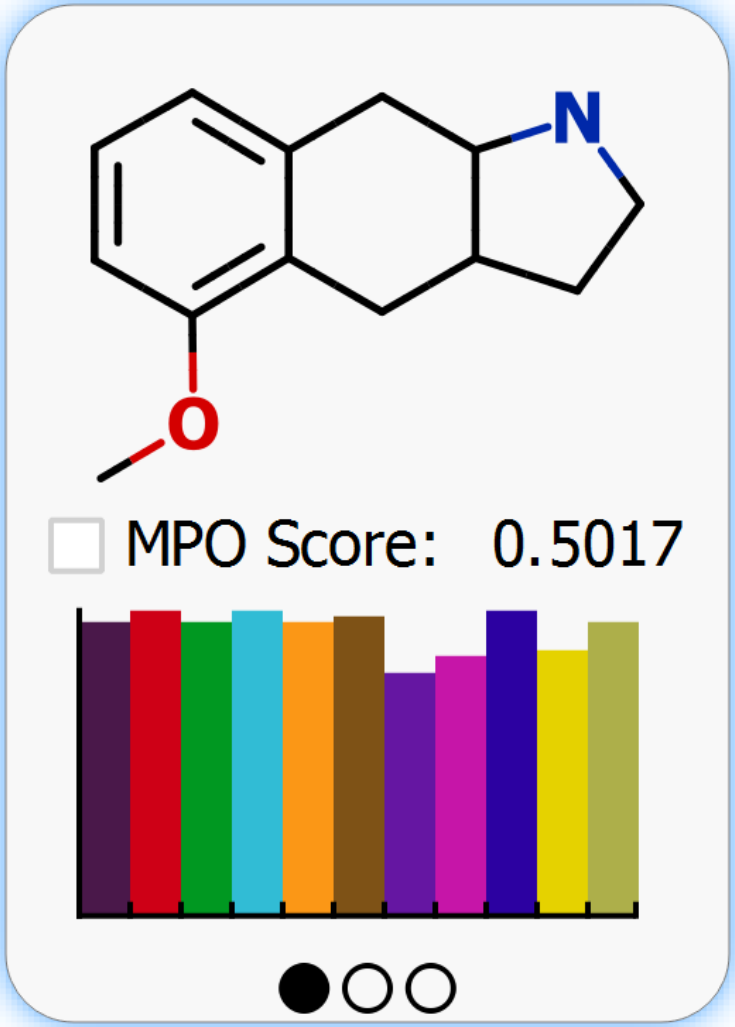
Error bars show confidence in overall score

Bottom 50% may be rejected with confidence



*M.D. Segall (2012) Curr. Pharm. Des. **18**(9) pp. 1292-1310

The best compound



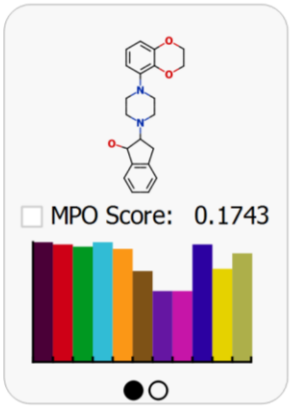
The context:

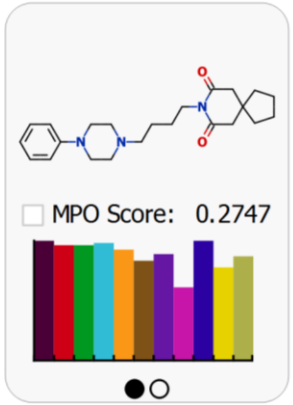
Property	Desired Value	Importance
5HT1a affinity (pKi)	> 7	
logS	> 1	
HIA category	+	
logP	0 -> 3.5	
BBB log([brain]:[blood])	-0.2 -> 1	
BBB category	+	
P-gp category	no	
hERG pIC50	⩽ 5	
2C9 pKi	⩽ 6	
2D6 affinity category	low medium	
PPB90 category	low	

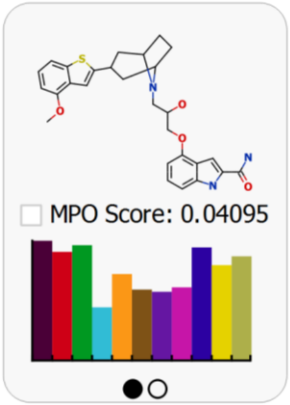
The best of each chemotype

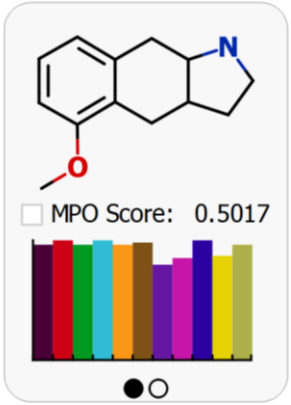
The context:

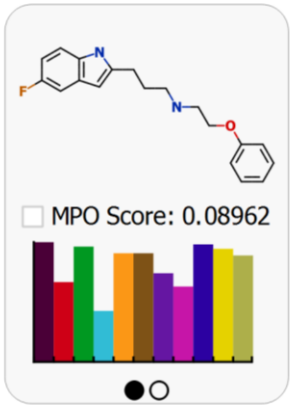
Property	Desired Value	Importance
5HT1a affinity (pKi)	> 7	
logS	> 1	
HIA category	+	
logP	0 -> 3.5	
BBB log([brain]:[blood])	-0.2 -> 1	
BBB category	+	
P-gp category	no	
hERG pIC50	⩽ 5	
2C9 pKi	⩽ 6	
2D6 affinity category	low medium	
PPB90 category	low	

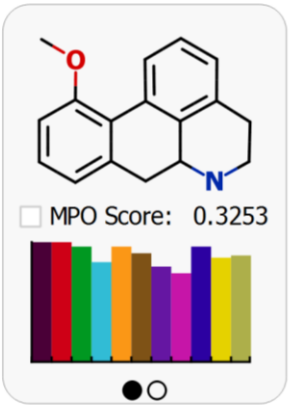




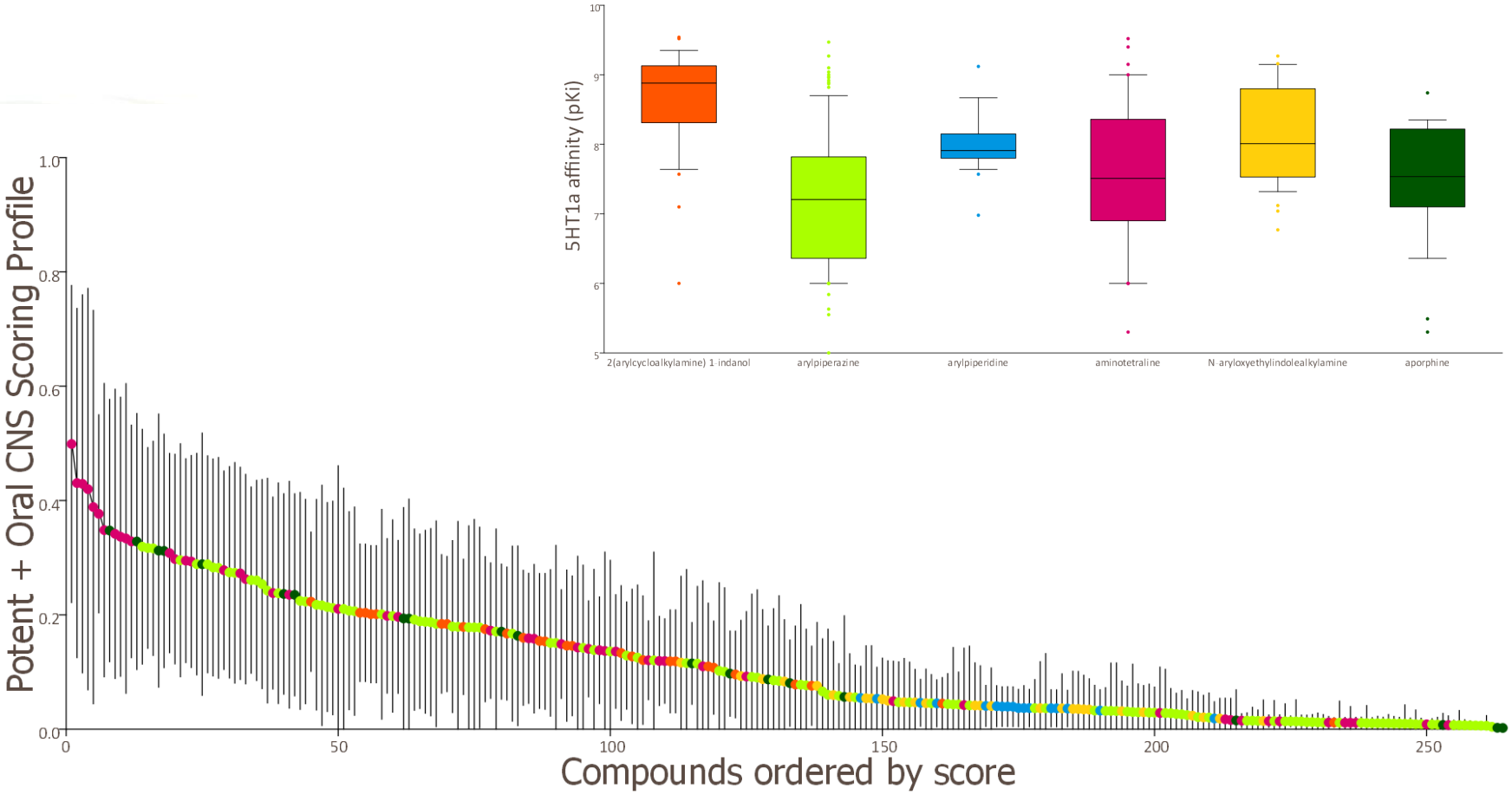




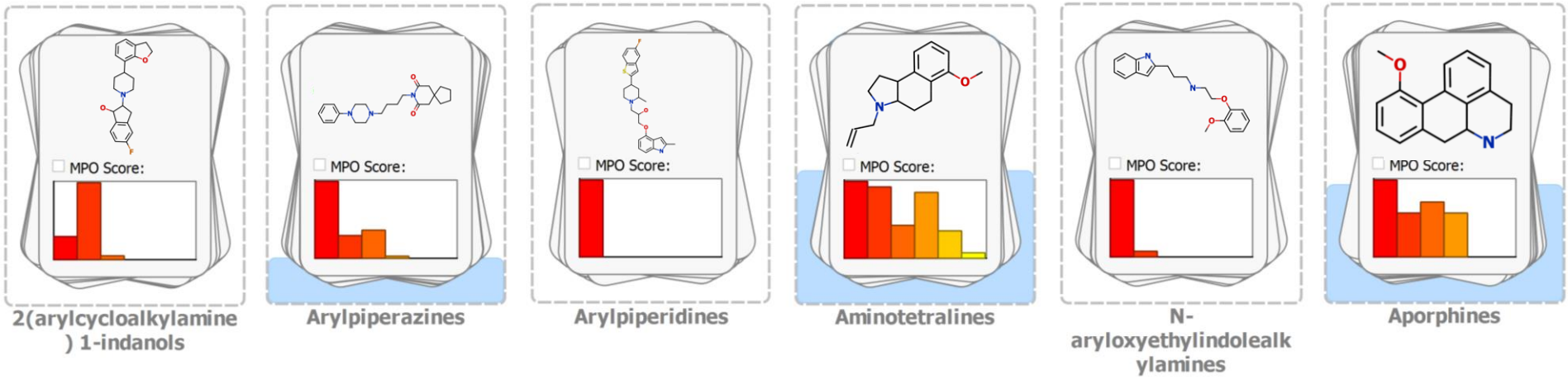




Snake plot for complete library



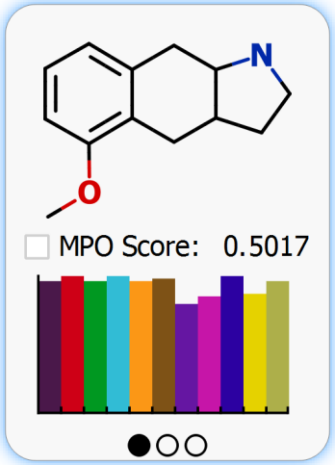
An appropriate selection?



(N.B. There are 2.77×10^{54} possible ways to select 50 compounds)

Conclusions

- We are always dealing with many dimensions of data while making decisions in drug discovery and so our visualisations should reflect this appropriately
- Only a small proportion of the possible visualisations we can create will be relevant/informative to our decision-making
- We tend to be draw towards information in small, easily-digestible chunks so…
- …could we tweet this?



Acknowledgements

- Matt Segall
- Peter Hunt
- Chris Leeding
- James Chisholm
- Alex Elliott
- Fayzan Ahmed
- Nick Foster
- Tamsin Mansley