Building the Utopian Dream – Toward the Integrated, Paperless Drug Discovery Lab.

Allan Jordan
Head of Chemistry
CR-UK Manchester Institute, UK

MANCHESTER 1824
The University of Manchester

CANCER RESEARCH UK | MANCHESTER INSTITUTE

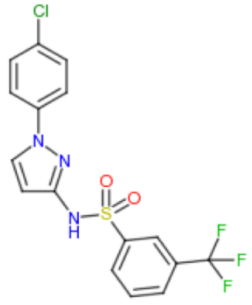
CRUK-MI Dotmatics History

- Workflow built from scratch around Dotmatics (2009-11)

- Appeal based on several factors:
  - Encompassed chemistry and biology
  - Implementable despite no dedicated chemo-informatics expertise
  - Scalable
  - Little IT overhead
  - Readily configurable
  - Single platform?
  - SAR interrogation/visualisation tools
  - Affordable

- Advantages:
  - No legacy data
  - De-novo setup
    - No software integration pre-requisites
    - No legacy hardware
- Diasdvantages
  - No internal Oracle expertise
    - University Oracle expertise “sporadic”
  - No chemo-informatics expertise
  - “Different” IT needs….

CRUK MI DDU LSD1 Project



Latest assay date

PDD00014643	project name: LSD1	Chemist: jhitchin

COMPOUND NAME: N-[1-(4-chlorophenyl)pyrazol-3-yl]-3-(trifluoromethyl)benzenesulfonamide

MW	FORMULA	HAC	HBA	HBD	LIPINSKI	ROTBONDS	TPSA (NOPS)	XLogP
401.79	C16H11ClF3N3O2S	26	5	1	Pass	4	72.4	4.76

Target	LE	LE modifier
LSD1	0.22	
LSD1	0	>
MAO-A	0	>

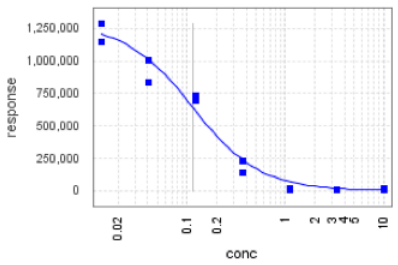
Sample ID	Conc µM	Assay result (avg % inh)	Assay format	Target
no results				

Sample ID	Analysis Name	Target	q	Assay format	Result (Geo Mean) µM	Result (n of m)	Hill Slope (geo mean)	Slope (n of m)
PDD00014643-001-001	IC50	LSD1		HTRF	56.91	2 of 2	0.87	2 of 2
PDD00014643-001-001	IC50	LSD1	>	LANCE Ultra	187.5	2 of 2	3.58	2 of 2
PDD00014643-001-001	IC50	MAO-A	>	FI	500	2 of 2	1.85	2 of 2

Dotmatics

not validated
props
comments
1 (Plate 1)

PDD00014494-003-001	1285603.0	0.01372uM	1285603.0
PDD00014494-003-001	1148267.0	0.01372uM	1148267.0
PDD00014494-003-001	1002218.0	0.04115uM	1002218.0
PDD00014494-003-001	833122.0	0.04115uM	833122.0
PDD00014494-003-001	731439.0	0.12346uM	731439.0
PDD00014494-003-001	696156.0	0.12346uM	696156.0
PDD00014494-003-001	231901.0	0.37037uM	231901.0
PDD00014494-003-001	143825.0	0.37037uM	143825.0
PDD00014494-003-001	13444.0	1.11111uM	13444.0
PDD00014494-003-001	14561.0	1.11111uM	14561.0
PDD00014494-003-001	10530.0	3.33333uM	10530.0
PDD00014494-003-001	12005.0	3.33333uM	12005.0
PDD00014494-003-001	12340.0	10.0uM	12340.0
PDD00014494-003-001	15011.0	10.0uM	15011.0

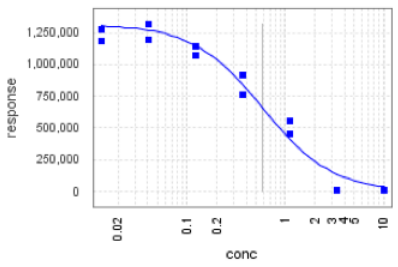


Cytotoxicity: 0.1147 pass

slope=1.259 min=0.000 max=1285603.000 err=>10000 R^2=0.9748

not validated
props
comments
1 (Plate 1)

PDD00016955-002-001	1278965.0	0.01372uM	1278965.0
PDD00016955-002-001	1180721.0	0.01372uM	1180721.0
PDD00016955-002-001	1315318.0	0.04115uM	1315318.0
PDD00016955-002-001	1192347.0	0.04115uM	1192347.0
PDD00016955-002-001	1144970.0	0.12346uM	1144970.0
PDD00016955-002-001	1072629.0	0.12346uM	1072629.0
PDD00016955-002-001	919233.0	0.37037uM	919233.0
PDD00016955-002-001	757352.0	0.37037uM	757352.0
PDD00016955-002-001	555248.0	1.11111uM	555248.0
PDD00016955-002-001	450135.0	1.11111uM	450135.0
PDD00016955-002-001	12798.0	3.33333uM	12798.0
PDD00016955-002-001	14219.0	3.33333uM	14219.0
PDD00016955-002-001	12140.0	10.0uM	12140.0
PDD00016955-002-001	15192.0	10.0uM	15192.0

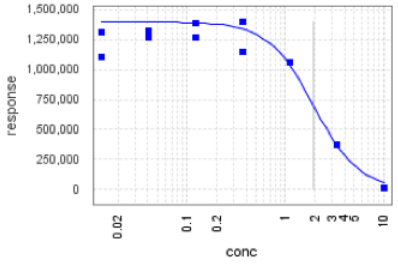


Cytotoxicity: 0.5908 pass

slope=1.229 min=0.000 max=1315318.000 err=>10000 R^2=0.9727

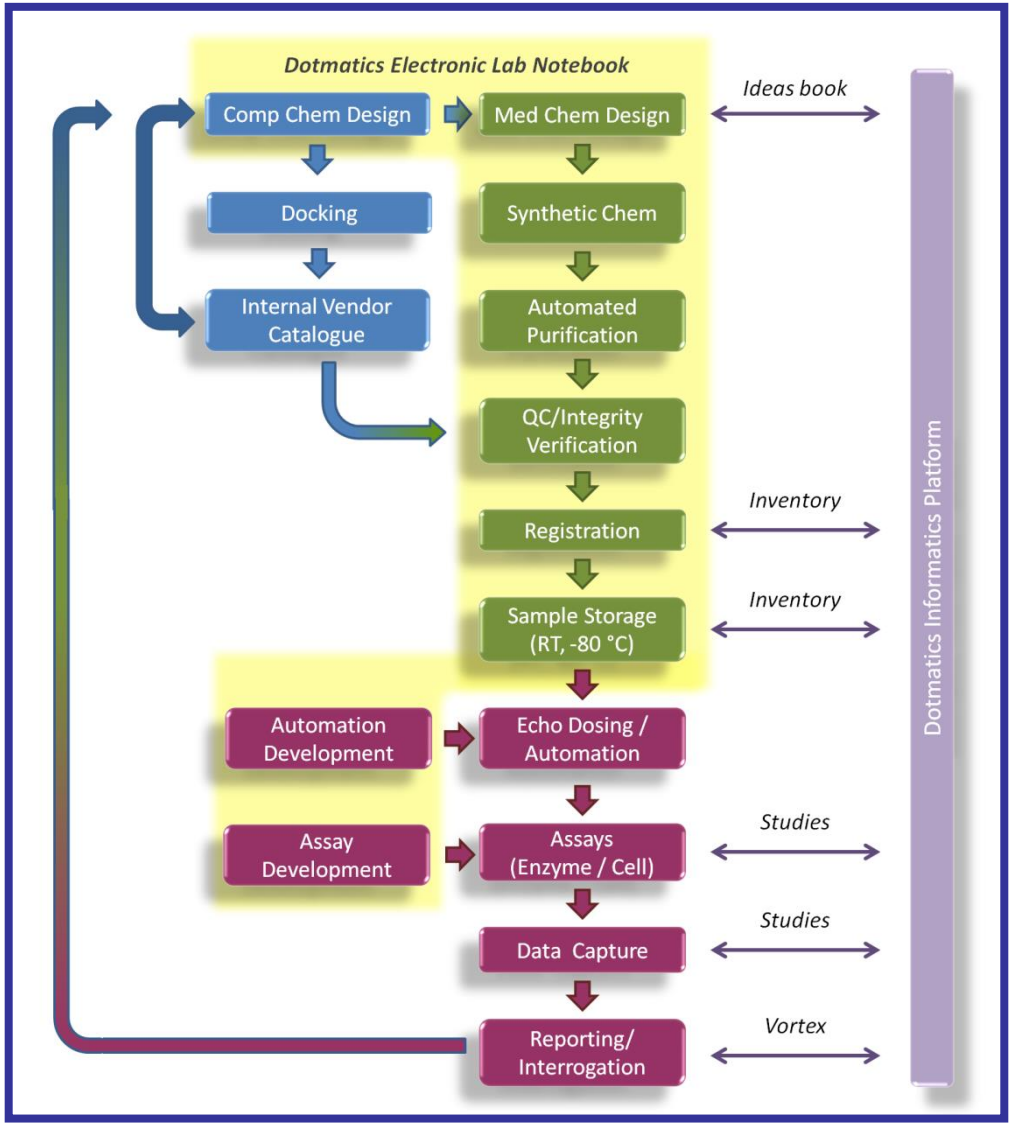
not validated
props
comments
1 (Plate 1)

PDD00017570-002-001	1315749.0	0.01372uM	1315749.0
PDD00017570-002-001	1107849.0	0.01372uM	1107849.0
PDD00017570-002-001	1323759.0	0.04115uM	1323759.0
PDD00017570-002-001	1263938.0	0.04115uM	1263938.0
PDD00017570-002-001	1393385.0	0.12346uM	1393385.0
PDD00017570-002-001	1266761.0	0.12346uM	1266761.0
PDD00017570-002-001	1397435.0	0.37037uM	1397435.0
PDD00017570-002-001	1150931.0	0.37037uM	1150931.0
PDD00017570-002-001	1061291.0	1.11111uM	1061291.0
PDD00017570-002-001	1062088.0	1.11111uM	1062088.0
PDD00017570-002-001	370674.0	3.33333uM	370674.0
PDD00017570-002-001	375628.0	3.33333uM	375628.0
PDD00017570-002-001	16241.0	10.0uM	16241.0
PDD00017570-002-001	19114.0	10.0uM	19114.0



Cytotoxicity: 1.9475 pass

slope=1.883 min=0.000 max=1397435.000 err=>10000 R^2=0.9488


Dotmatics Electronic Lab Notebook
Comp Chem Design
Med Chem Design
Docking
Synthetic Chem
Internal Vendor Catalogue
Automated Purification
QC/Integrity Verification
Registration
Sample Storage (RT, -80 °C)
Automation Development
Echo Dosing / Automation
Assay Development
Assays (Enzyme / Cell)
Data Capture
Reporting/ Interrogation
Ideas book
Inventory
Inventory
Studies
Studies
Vortex
Dotmatics Informatics Platform

Bugbears from a previous life…

- Screening data is easy to capture, process and disseminate

- But:
  - Access to data stored on hard drives or filing cabinets?
    - Desktops of "downsized" staff?
  - Uniform processing of "non-standardised" or "fuzzy" data?
  - Capture of "soft" knowledge?

- Reliance on fixed, out-of-date and incomplete data repositories for decision-making

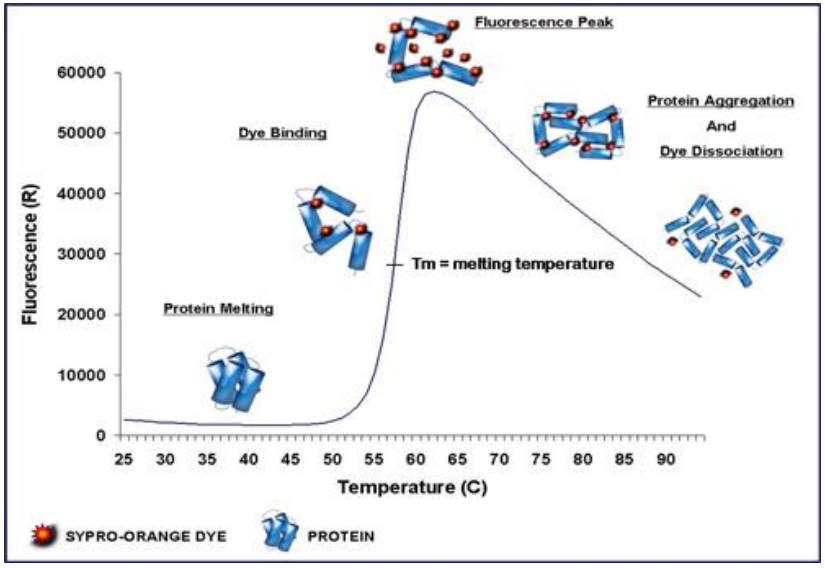
Drug Discovery Informatics Conclusions…

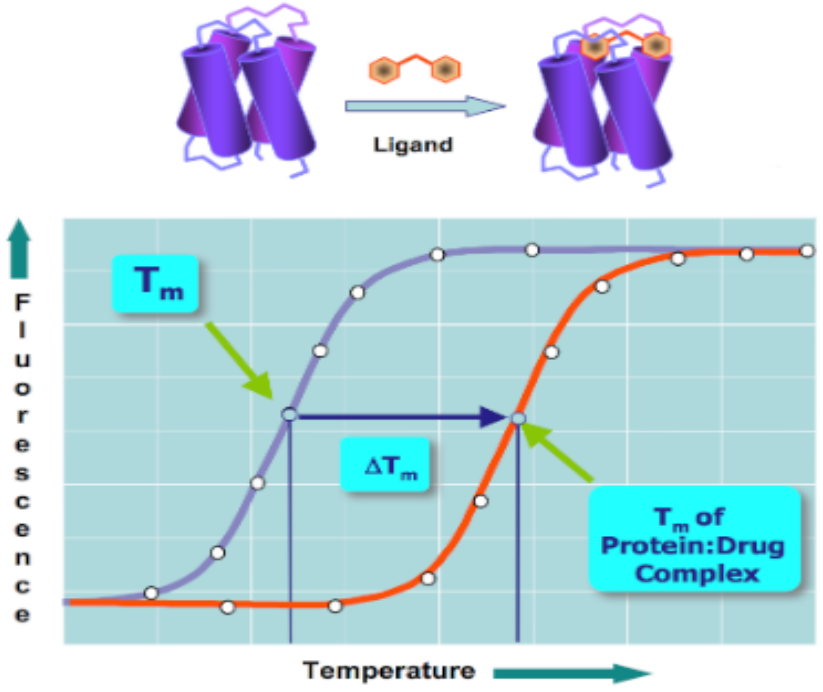
- Drug discovery databases are great except for:
- Getting data in.
- Getting data out.
- Particularly:
  - Importing "non-standard" data
    - "unusual" data curves, in vivo data
  - Complex data retrieval
    - Pivoting and/or aggregating across multiple datasources
- Our daily users are not informaticians.

GETTING DATA IN…

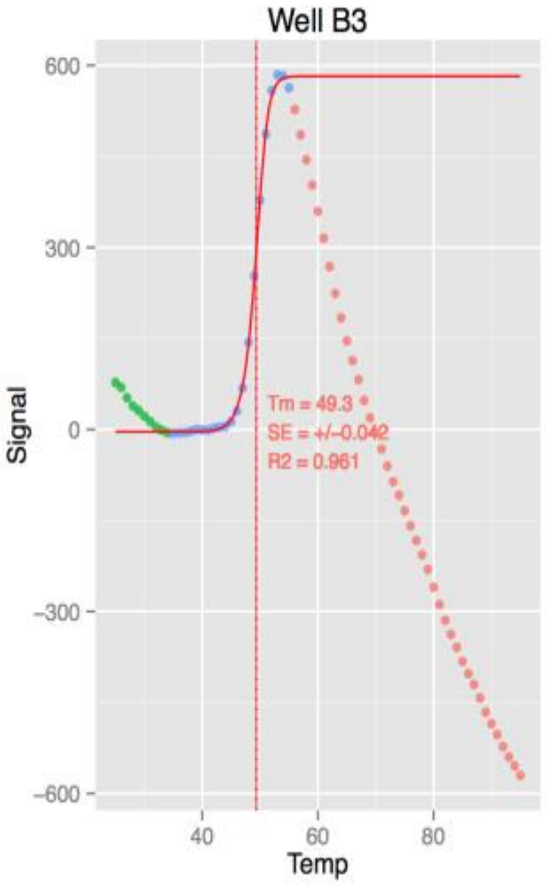
- Thermal Shift Data
- Incucyte Cell screening

Thermal Shift Assay





How it works




```
#get the PastMax values
max.cycle <- mydata[mydata[,2] == max(mydata[,2]),1]
mydata[mydata$Temp > max.cycle +2, 'Status'] <- 'PastMax'

#get the PreMin values
min.signal <- min(mydata[mydata$Status=='Use',2])
min.cycle <- mydata [ mydata$Status == 'Use' & mydata$Signal == min.signal, 1 ]
mydata[mydata$Temp < min.cycle , 'Status'] <- 'PreMin'

#if less than 5  'use' value, reset all data points to use so that user can see data
if (nrow(subset(mydata,mydata$Status=='Use')) < 5) {
  mydata$Status <- 'Use'
}

#now do the curve fitting
mydata.m1 <- drm(Signal ~ Temp, data = mydata[mydata$Status == 'Use',],
                 fct=LL.5(names=c('Hillslope','Bottom','Top','EC50','Symmetry')))
```

- Find the max value
- Find the min value
- Exclude data points either side
  Sanity check there are enough points
- Fit the curve – 5 parameter log logistic

```
for (i in 1:(ncol(df)-1)) {

  if (i==1) {
    output.list <- list()}

  tm <- thermal.melt(df[,c(1,i+1)],pm)

  if (!is.null(tm)) { #check that something has come back from the thermal.melt

    output.list[[tm$wellid]] <- tm

  }


}
```

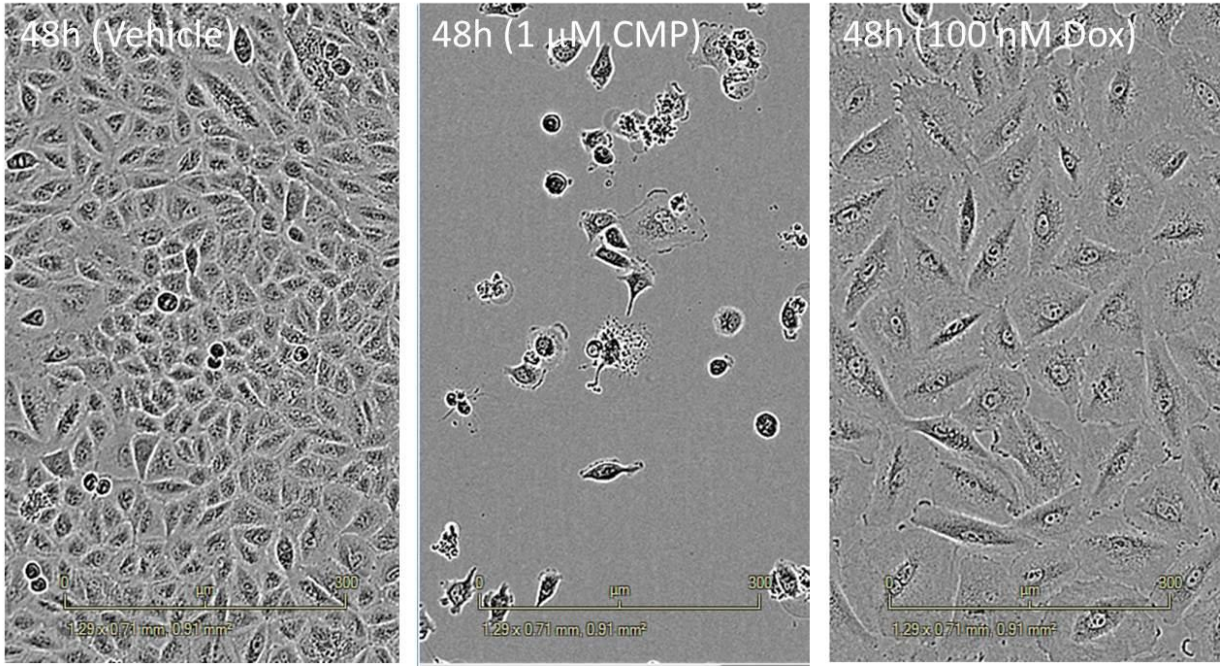
Loop through all wells

Incucyte Cell Screening

- "The best cancer cell is a dead one"
- How to measure cell death?
- Staining of living cells, then count at a specific time
- Better way – live cell imaging
  - Monitor cell growth rates directly in the incubator, in real time
- How to capture this type of data?

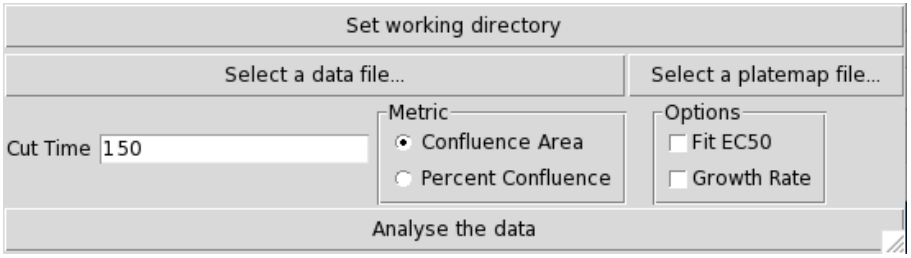
Incucyte Live Cell Imaging

- Effect of a compound monitored in living cells in real time
- Many, many data points: 300 images per well of a 384-well plate!

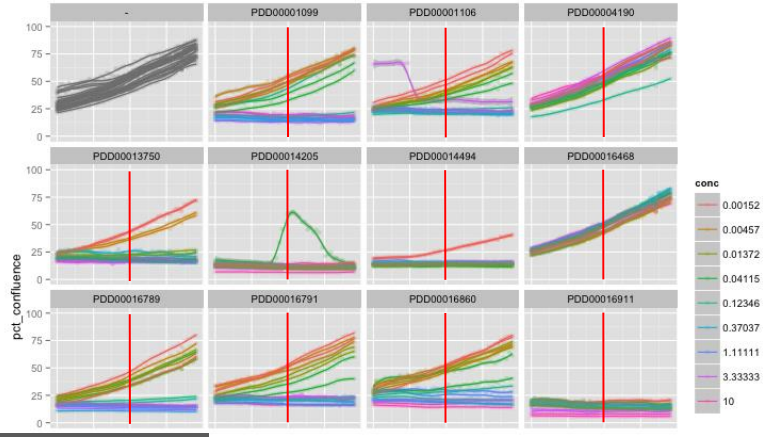


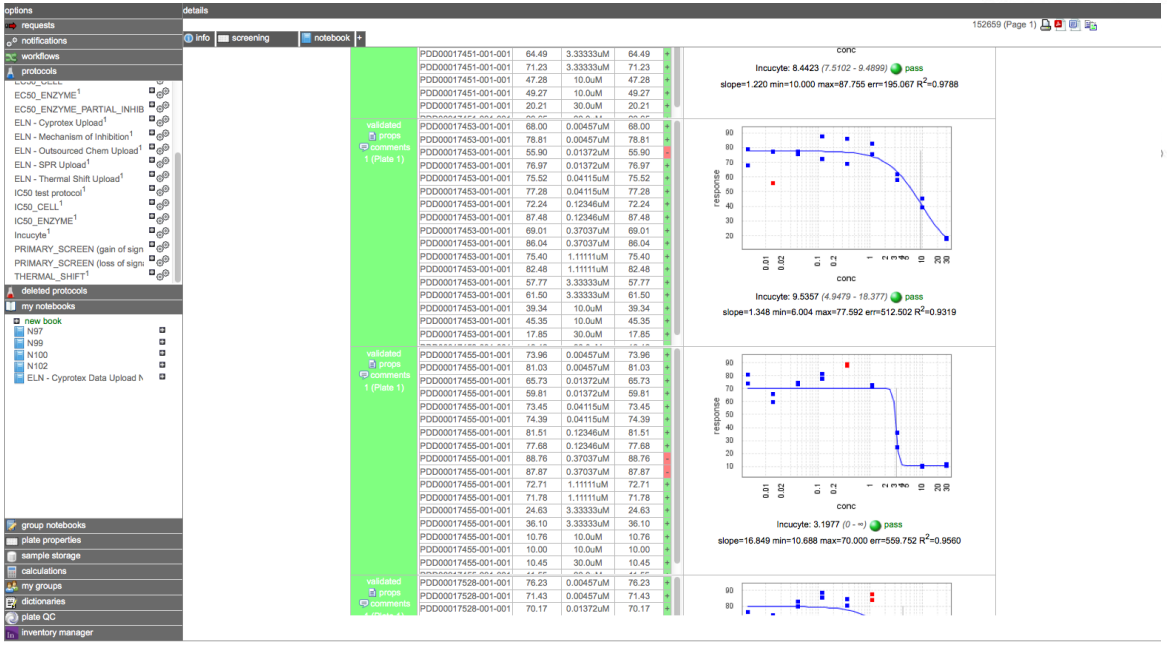
- How to parse this into "standard" data for database insertion / analysis?

Incucyte Live Cell Imaging



User selects raw data files and options





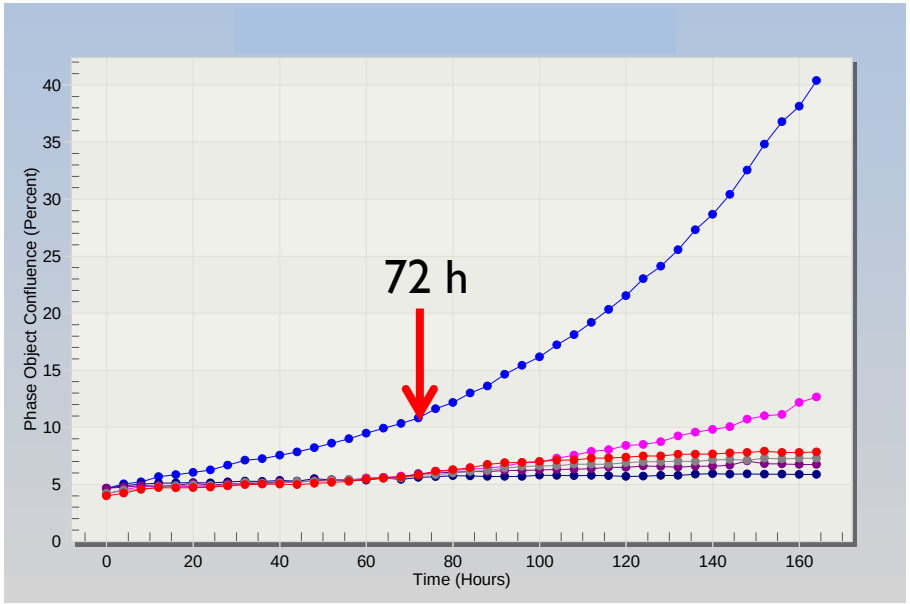
R script plots growth curves and data points extracted at specified time

Data imported into Dotmatics for dose response analysis
- **Hybrid solution**

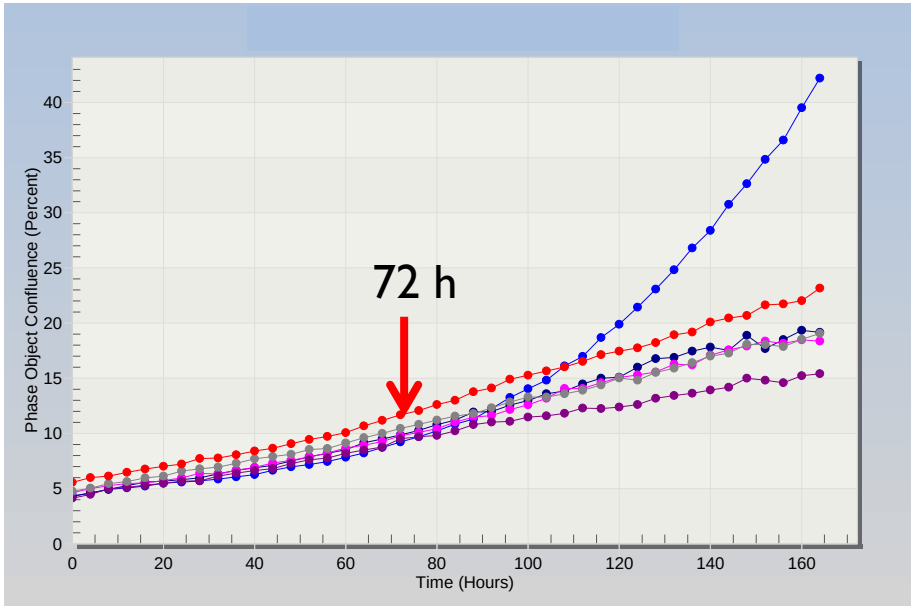
So far so good…

- But different tumour types grow at different rates…
- Moreover, different potential therapeutic agents take different numbers of cell replication cycles to manifest their effects…

Kinase inhibitor



DNA repair inhibitor

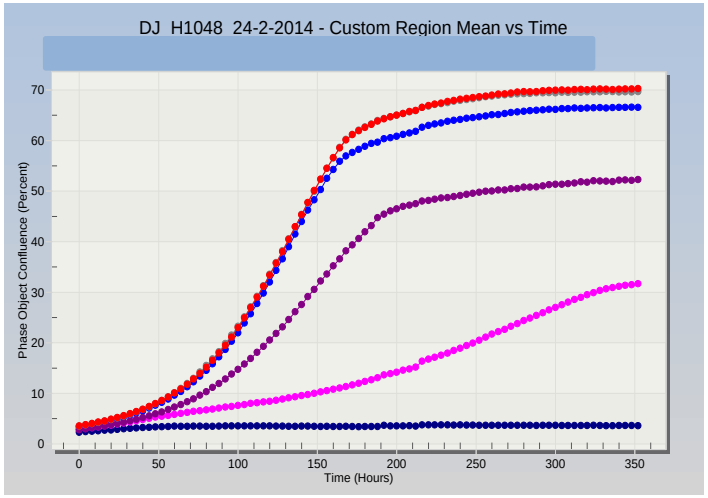


Example: DNA damage repair inhibitors often require longer timepoints

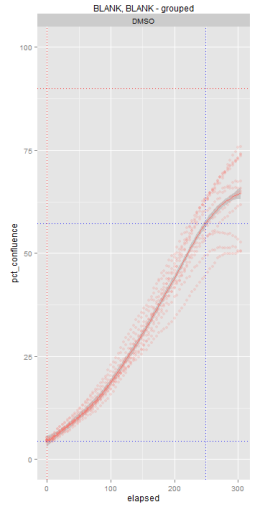
So far so good…

- But different tumour types grow at different rates…
- Different agents take different numbers of cell cycles to manifest their effects…
- Can we capture this?
  - Relate to population doubling times of untreated cells
  - Determine for each specific cell line
  - Assess drug action at this specific and unique timepoint

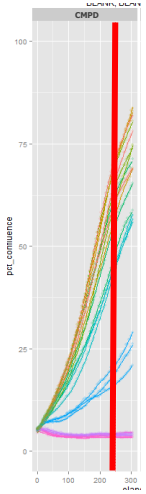
Incucyte Analysis – per-cell line growth rates



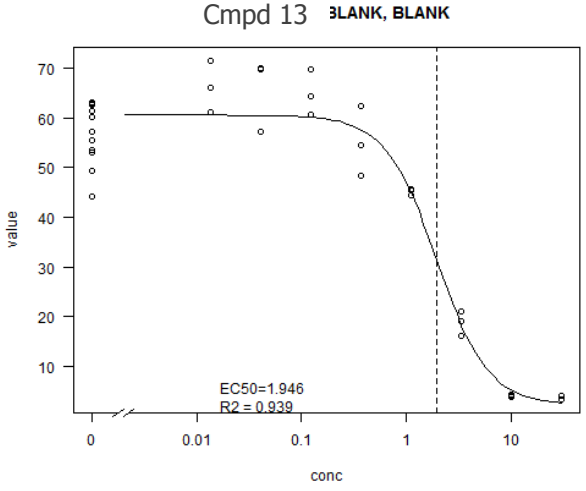
All data imported into R
(bespoke, in-house package)



Cut-off before growth inhibition
Number of population doublings calculated (in controls)



doublings chosen
Growth of dosed cells calculated



Data uploaded and analysed in Dotmatics

IC_{50}

Benefits of Scripting

- Automated, fast and easy to deploy

- Consistent processing
  - Today, tomorrow, next year
  - Absolute data consistency
  - Absolute user consistency
  - *Avoids Excel transposition errors*

- Less user bias, higher data integrity

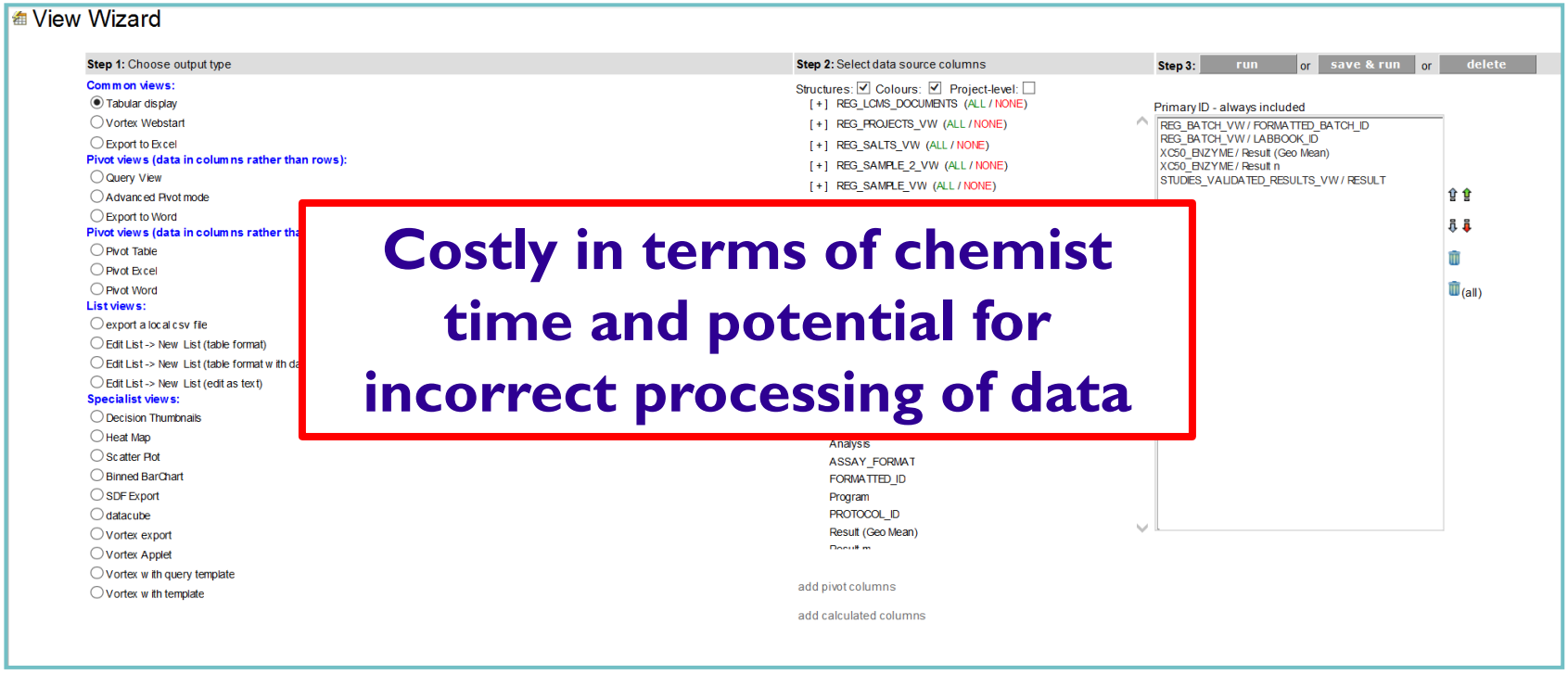
- **Less variance over time**
- **Quality decision-making data**

GETTING DATA OUT…

- Data Warehousing

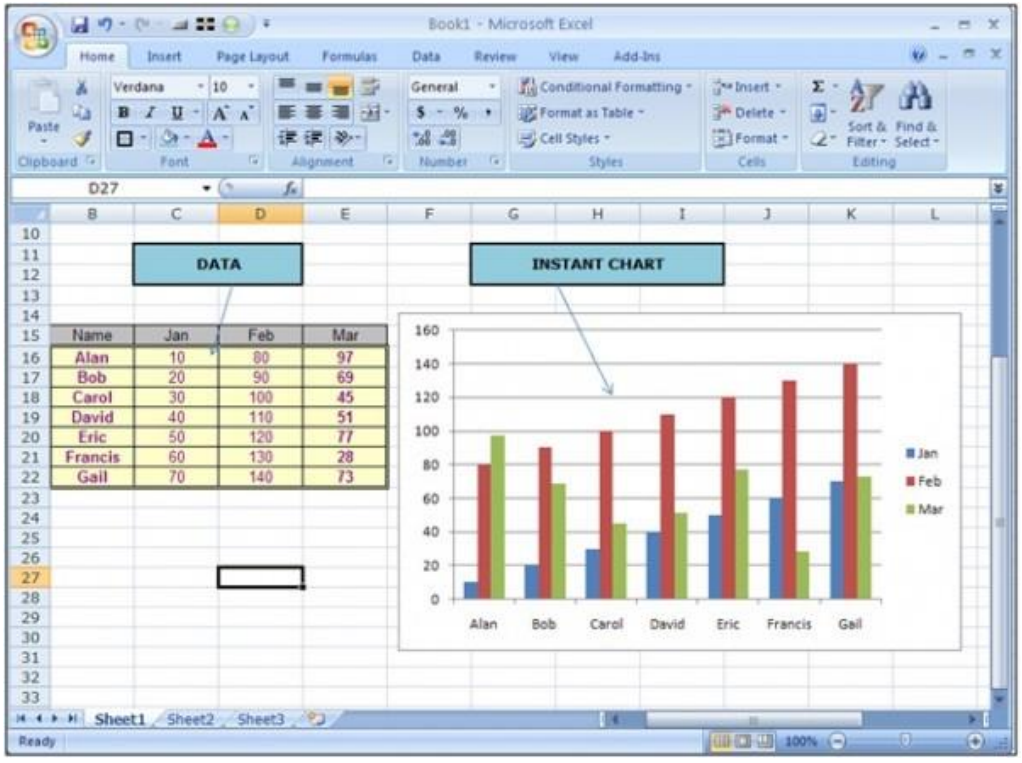
Data Export

- For single compounds, our database browser is great.
- For lots of compounds, view wizard – not so great:



- Need to understand the data structures
- Then what to do with it?
  - Pivoting, joining, aggregating…

Data Analysis



Issues

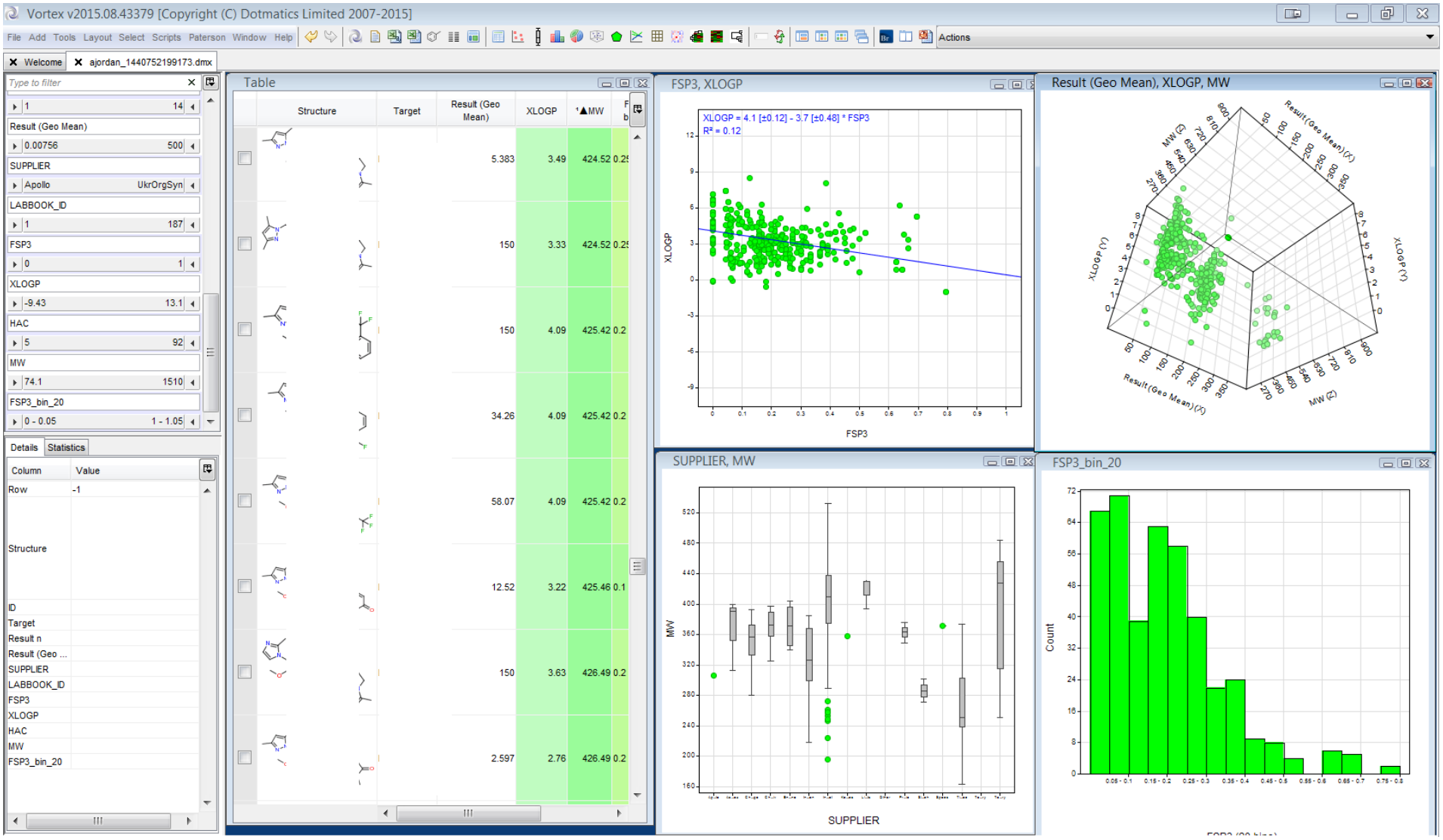
Breaking free from chemical spreadsheets

Matthew Segall, Ed Champness, Chris Leeding, James Chisholm, Peter Hunt, Alex Elliott, Hector Garcia-Martinez, Nick Foster and Samuel Dowling

Optibrium Ltd, 7221 Cambridge Research Park, Beach Drive, Cambridge CB25 9TL, UK

- Limited data interpretation tools
- Structural awareness requires additional plug-ins

Vortex


Vortex v2015.08.43379 [Copyright (C) Dotmatics Limited 2007-2015]
File Add Tools Layout Select Scripts Paterson Window Help
Actions
Welcome
ajordan_1440752199173.dmx
Type to filter
Result (Geo Mean)
SUPPLIER
Apollo UkrOrgSyn
LABBOOK_ID
FSP3
XLOGP
HAC
MW
FSP3_bin_20
Details
Statistics
Column
Value
Row
Structure
ID
Target
Result n
Result (Geo ...
SUPPLIER
LABBOOK_ID
FSP3
XLOGP
HAC
MW
FSP3_bin_20
Table
Structure
Target
Result (Geo Mean)
XLOGP
MW
5.383 3.49 424.52
150 3.33 424.52
150 4.09 425.42
34.26 4.09 425.42
58.07 4.09 425.42
12.52 3.22 425.46
150 3.63 426.49
2.597 2.76 426.49
FSP3, XLOGP
XLOGP = 4.1 [±0.12] - 3.7 [±0.48] * FSP3
R² = 0.12
XLOGP
FSP3
Result (Geo Mean), XLOGP, MW
Result (Geo Mean) (X)
XLOGP (Y)
MW (Z)
SUPPLIER, MW
MW
SUPPLIER
FSP3_bin_20
Count

Issues

Breaking free from chemical spreadsheets

Matthew Segall, Ed Champness, Chris Leeding, James Chisholm, Peter Hunt, Alex Elliott, Hector Garcia-Martinez, Nick Foster and Samuel Dowling

Optibrium Ltd, 7221 Cambridge Research Park, Beach Drive, Cambridge CB25 9TL, UK

- Limited data interpretation tools
- Structural awareness requires additional plug-ins
- Legacy / historical point of reference
- Who has the "right" data?
  - Batch/compound level aggregation
  - End-user averaging

Dotmatics R Integration

- Use ROracle package to make connection to Oracle database
- SQL directly in R then pulls data out
- Once data is in R, can then aggregate, filter, pivot and export to text file which can be visualised (eg Shiny, Vortex).
- Standard scripts run as an automated job, then copies output to accessible location – 'Data Warehouse'
- Applications:
  - **Nightly** output of all project compounds and key data
  - Generate standard plots and visualisations

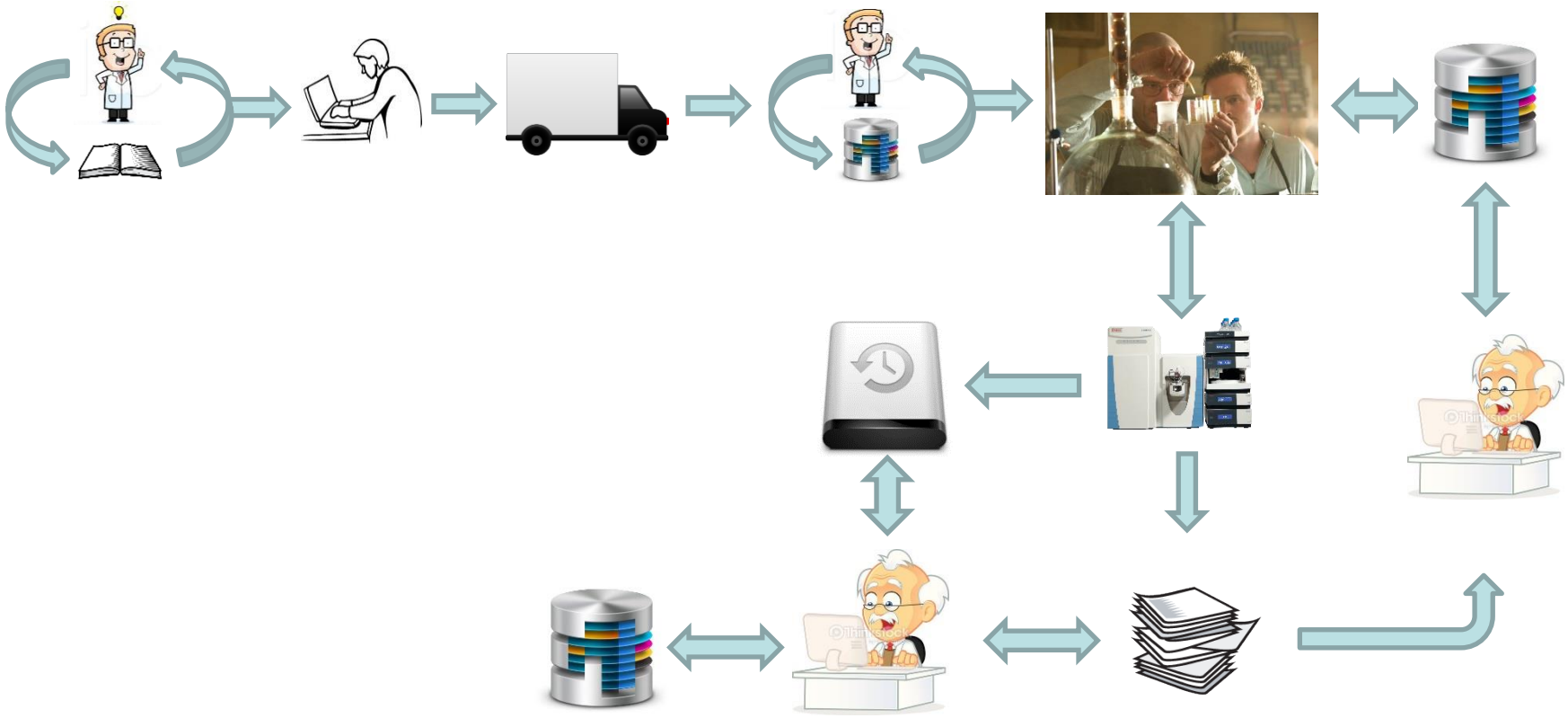
Dotmatics R Integration

- Benefits of approach:
  - All of the data analysis capability of R can be applied to data in Dotmatics
  - Standardised: no need for scientists to have to do their own piecemeal pivoting, aggregating, visualisation etc.
  - Consistent: scientists interrogate the same data
  - Easy to customise as we add new data types

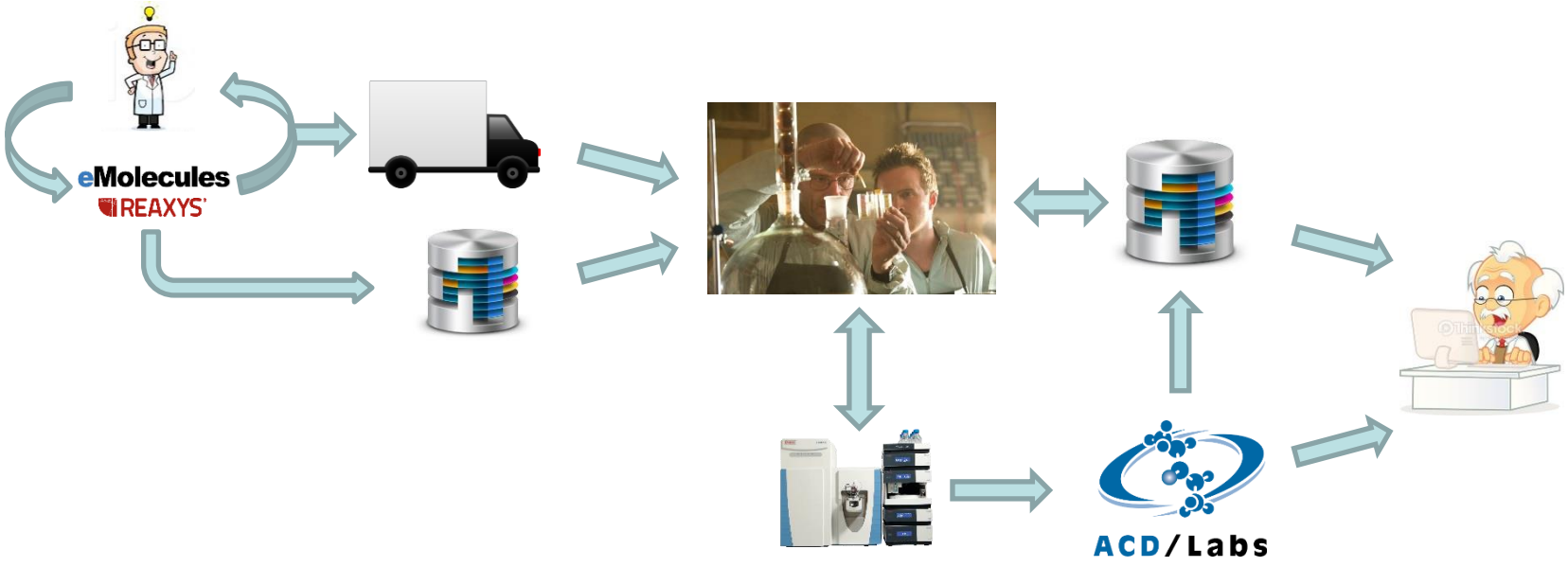
- Vortex **always** points to the very latest data

FUTURE DIRECTIONS…

Present chemistry workflow



What are we working toward?



Stumbling blocks…

- Data exchange formats
  - The Universal Format for Interchangable Data??
  - Mp3/JPG/CSV for the lab?
  - Platereaders
    - Instrument header; row; column; value; timestamp
  - Analytical Instrumentation
    - Interpretable, standard raw data output per technique
  - SMILES/SMARTS
    - Differing interpretations…
- How do we drive standardisation with the vendors?

Summary

Drug Discovery is not about making and testing compounds

- It is about making decisions…
- Right question, of the right compound, at the right time
  - Requires easy access to all relevant data
  - Requires easy access to the *latest* data
- Issues getting data in and out are endemic across laboratories
- Demands innovative solutions to overcome
  - *But these need not be overly complex or cumbersome*

Better data capabilities ➔ Better Decisions ➔ Medicines to patients, faster

Acknowledgements

Drug Discovery Unit:

Donald Ogilvie

Ian Waddell

Allan Jordan

Jonathan Ahmet
Roger Butlin
Niall Hamilton
James Hitchin
Shaun Johns
Stuart Jones
Chris Kershaw
Alison McGonagle
Daniel Mould
Rebecca Newton
Ali Raoof
Kate Smith
Bohdan Waszkowycz

Phil Chapman
Adnana Tudose

Ben Acton
Mentxu Aiertza
Habiba Begum
Elizabeth Blaikely
Charlotte Burt
Mark Cockerill
Emma Fairweather
Samantha Fritzl
Louise Griffiths
Nicola Hamilton
Gemma Hopkins
Dominic James
Paul Kelly
Nikki March
Helen Small
Alex Stowell
Graeme Thomson
Mandy Watson

dotmatics knowledge solutions
Jon Steadman
Dan Ormsby

CANCER RESEARCH UK

ACD/Labs
Veronica Paget

eMolecules
Haydn Boehm