

# Multilayered Analysis of HCS Data: An Integrated Approach

Amit Bahl, Wendy Bailey, Bonnie Howell, Ed Keough, Irene Pak,  
Ansu Bagchi

Merck & Co., Inc.

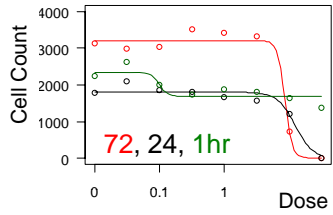
January 13, 2010

# Outline

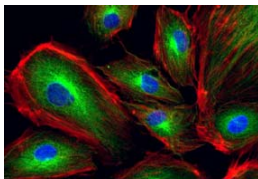
- **Data Analysis**
  - Data from multi-factorial experiments
  - Multivariable data – cannot ignore correlations
  - Univariate vs. Multivariate strategies
- **Assay Quality**
  - What goes in...
- **Data / Information Management**
  - Storage / Retrieval / Navigation
- **Putting it all together**

# Data Analysis

# HCS Workflow



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC	BD	BE	BF	BG	BH	BI	BJ	BK	BL	BM	BN	BO	BP	BQ	BR	BS	BT	BU	BV	BW	BX	BY	BZ	CA	CB	CC	CD	CE	CF	CG	CH	CI	CJ	CK	CL	CM	CN	CO	CP	CQ	CR	CS	CT	CU	CV	CW	CX	CY	CZ	DA	DB	DC	DD	DE	DF	DG	DH	DI	DJ	DK	DL	DM	DN	DO	DP	DQ	DR	DS	DT	DU	DV	DW	DX	DY	DZ	EA	EB	EC	ED	EE	EF	EG	EH	EI	EJ	EK	EL	EM	EN	EO	EP	EQ	ER	ES	ET	EU	EV	EW	EX	EY	EZ	FA	FB	FC	FD	FE	FF	FG	FH	FI	FJ	FK	FL	FM	FN	FO	FP	FQ	FR	FS	FT	FU	FV	FW	FX	FY	FZ	GA	GB	GC	GD	GE	GF	GG	GH	GI	GJ	GK	GL	GM	GN	GO	GP	GQ	GR	GS	GT	GU	GV	GW	GX	GY	GZ	HA	HB	HC	HD	HE	HF	HG	HH	HI	HJ	HK	HL	HM	HN	HO	HP	HQ	HR	HS	HT	HU	HV	HW	HX	HY	HZ	IA	IB	IC	ID	IE	IF	IG	IH	II	IJ	IK	IL	IM	IN	IO	IP	IQ	IR	IS	IT	IU	IV	IW	IX	IY	IZ	JA	JB	JC	JD	JE	JF	JG	JH	JI	IJ	JK	JL	JM	JN	JO	JP	JQ	JR	JS	JT	JU	JV	JW	JX	JY	JZ	KA	KB	KC	KD	KE	KF	KG	KH	KI	KJ	KK	KL	KM	KN	KO	KP	KQ	KR	KS	KT	KU	KV	KW	KX	KY	KZ	LA	LB	LC	LD	LE	LF	LG	LH	LI	LJ	LK	LL	LM	LN	LO	LP	LQ	LR	LS	LT	LU	LV	LW	LX	LY	LZ	MA	MB	MC	MD	ME	MF	MG	MH	MI	MJ	MK	ML	MM	MN	MO	MP	MQ	MR	MS	MT	MU	MV	MW	MX	MY	MZ	NA	NB	NC	ND	NE	NF	NG	NH	NI	NJ	NK	NL	NM	NN	NO	NP	NQ	NR	NS	NT	NU	NV	NW	NX	NY	NZ	OA	OB	OC	OD	OE	OF	OG	OH	OI	OJ	OK	OL	OM	ON	OO	OP	OQ	OR	OS	OT	OU	OV	OW	OX	OY	OZ	PA	PB	PC	PD	PE	PF	PG	PH	PI	PJ	PK	PL	PM	PN	PO	PP	PQ	PR	PS	PT	PU	PV	PW	PX	PY	PZ	QA	QB	QC	QD	QE	QF	QG	QH	QI	QJ	QK	QL	QM	QN	QO	QP	QQ	QR	QS	QT	QU	QV	QW	QX	QY	QZ	RA	RB	RC	RD	RE	RF	RG	RH	RI	RJ	RK	RL	RM	RN	RO	RP	RQ	RR	RS	RT	RU	RV	RW	RX	RY	RZ	SA	SB	SC	SD	SE	SF	SG	SH	SI	SJ	SK	SL	SM	SN	SO	SP	SQ	SR	SS	ST	SU	SV	SW	SX	SY	SZ	TA	TB	TC	TD	TE	TF	TG	TH	TI	TJ	TK	TL	TM	TN	TO	TP	TQ	TR	TS	TT	TU	TV	TW	TX	TY	TZ	UA	UB	UC	UD	UE	UF	UG	UH	UI	UJ	UK	UL	UM	UN	UO	UP	UQ	UR	US	UT	UU	UV	UW	UX	UY	UZ	VA	VB	VC	VD	VE	VF	VG	VH	VI	VJ	VK	VL	VM	VN	VO	VP	VQ	VR	VS	VT	VU	VV	VW	VX	VY	VZ	WA	WB	WC	WD	WE	WF	WG	WH	WI	WJ	WK	WL	WM	WN	WO	WP	WQ	WR	WS	WT	WU	WV	WW	WX	WY	WZ	XA	XB	XC	XD	XE	XF	YG	YH	YI	YJ	YK	YL	YM	YN	YO	YP	YQ	YR	YS	YT	YU	YV	YW	YX	YY	YZ	ZA	ZB	ZC	ZD	ZE	ZF	ZG	ZH	ZI	ZJ	ZK	ZL	ZM	ZN	ZO	ZP	ZQ	ZR	ZS	ZT	ZU	ZV	ZW	ZX	ZY	ZZ
--	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----



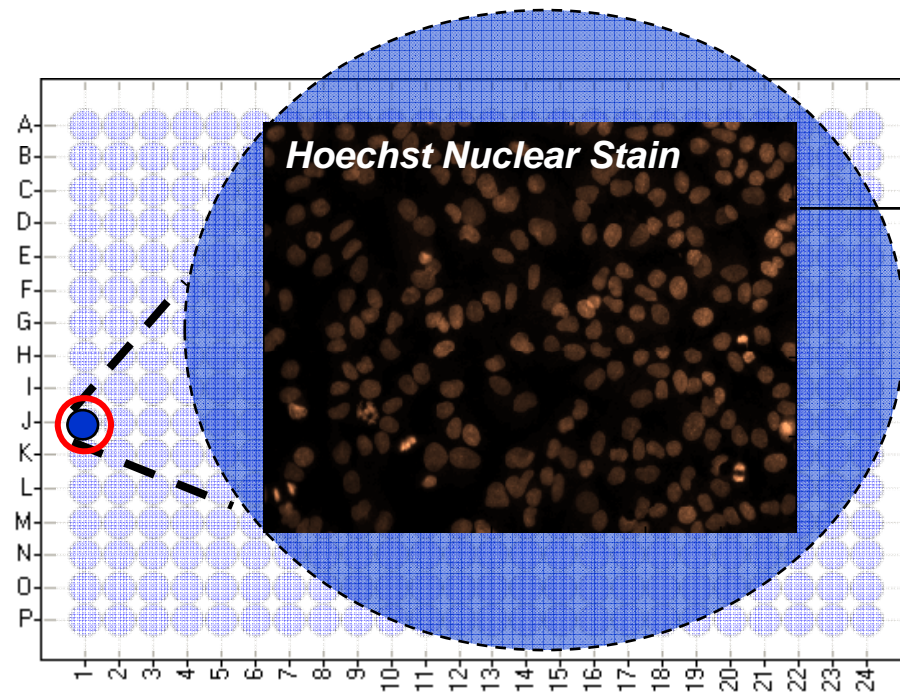
## Scientific Insight!

# Statistical Analysis / Data Mining

## Image Analysis

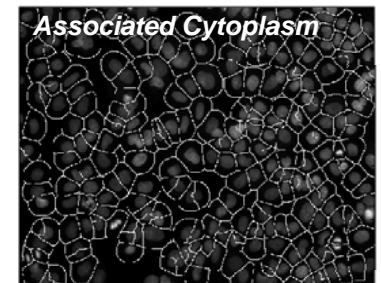
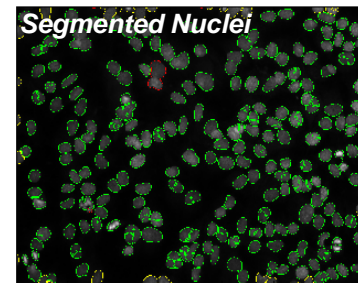
# HCScreen

# High Content Screens



**Assay and Image Acquisition**

## Automated Image Analysis



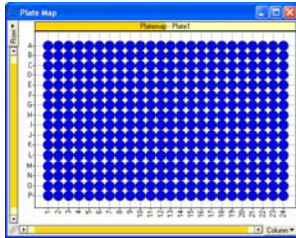
	cell #	1	2 ...
Cell area		40.1	42.5
Cell perimeter		35.3	38.9
Cell aspect ratio		1.56	2.01
Actin content		4510	4939
Actin texture		16.8	17.2
Cell solidity		0.99	1.03
Cell extent		0.68	0.35
Nuclear area		10.9	11.1
Nuclear perimeter		1.0	1.1
Nuclear aspect ratio		1.56	2.01

**Cell-level Data**

Systematic titration study of compounds, or other therapeutic agents, across multiple readouts, cell lines, or several time points

# Data Scope

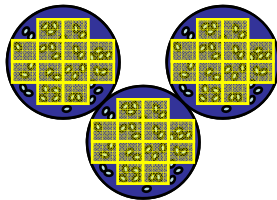
## Plate level



### Plate setup

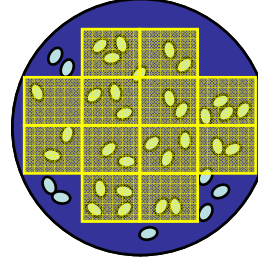
- Treatment type
- Treatment ID
- Concentration
- Reagents

## Replicate level



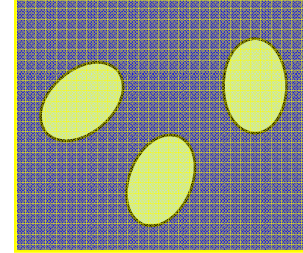
Replicate summary (basic statistics on all the cells pooled together from one or more experimentally-identical wells) for feature of interest generated by the user.

## Well level



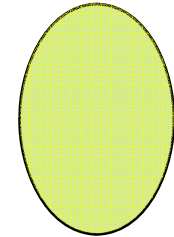
Well summary (basic statistics on all the cells imaged in this well) for each feature measured by the instrument or generated by the user.

## Field level



Field summary (basic statistics on the cells imaged in a single field) for each feature measured by the instrument or generated by the user.

## Cell level



Numerical data (a.k.a. "descriptors") for all features measured for a single cell by the instrument.

Extracted from  
Platemap File

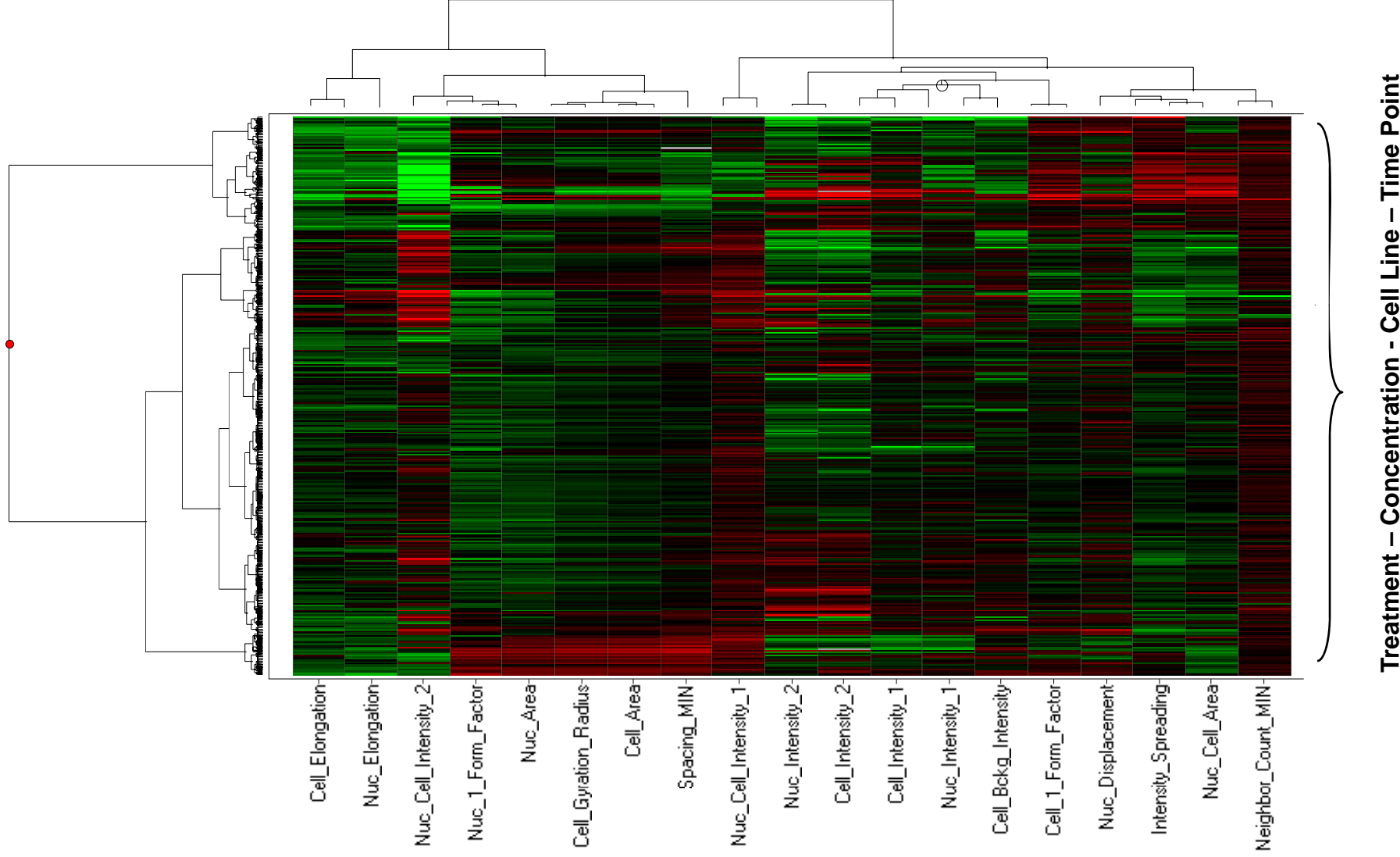
Extracted or derived from  
Instrument Data

# Scientific Goals

What questions are we interested in asking?

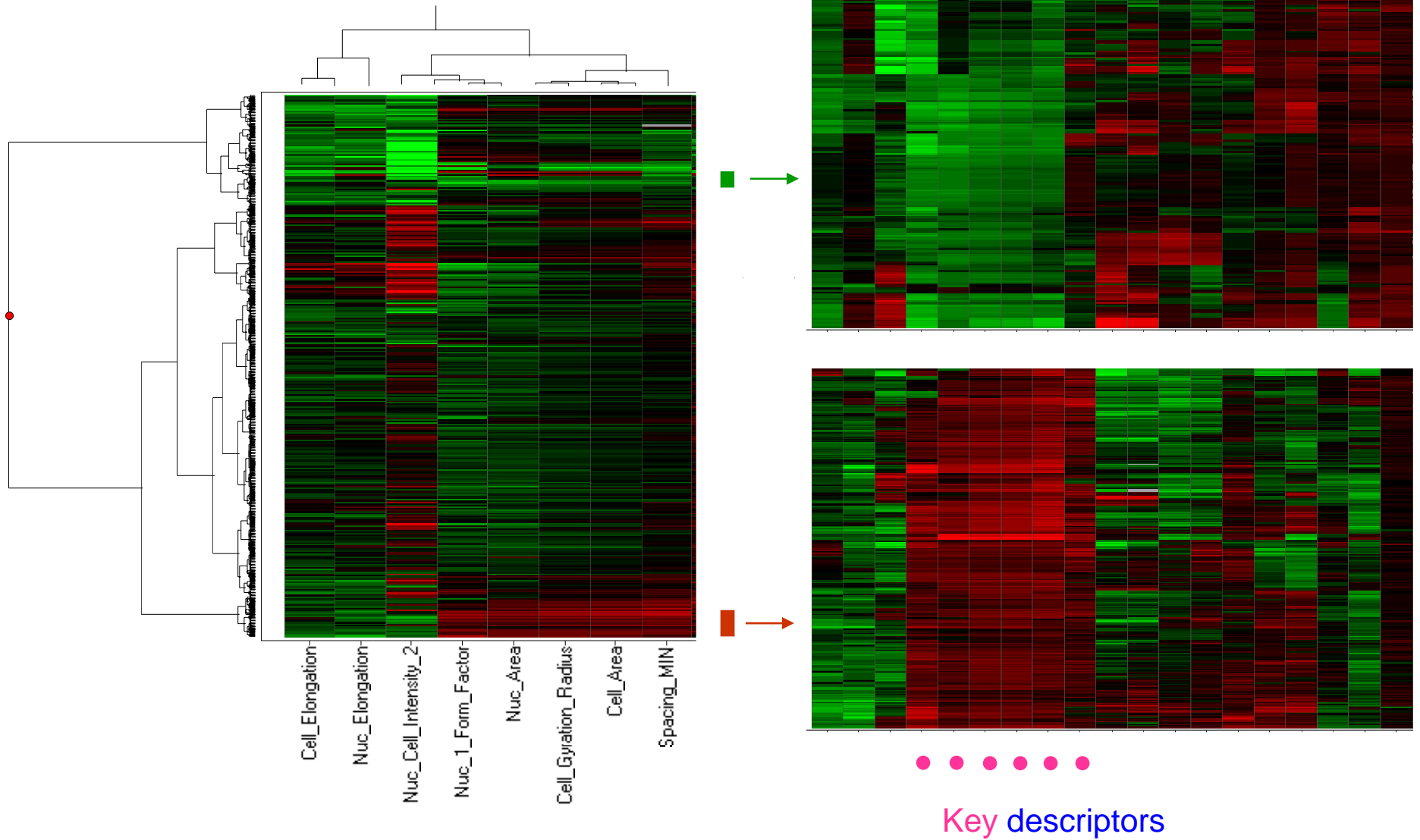
- Classification of therapeutic agents?
  - Into multiple pre-determined groups
- Differentiation between therapeutic agents?
  - Identify key phenotypic characteristics, and
  - The concentration of the test agent, at which the phenotype emerges
- To find items that are similar/dissimilar to positive/negative controls?
  - Controls with known bio-chemical characteristics (good/bad)
- Characterize MOA?

# Univariate: Clustering

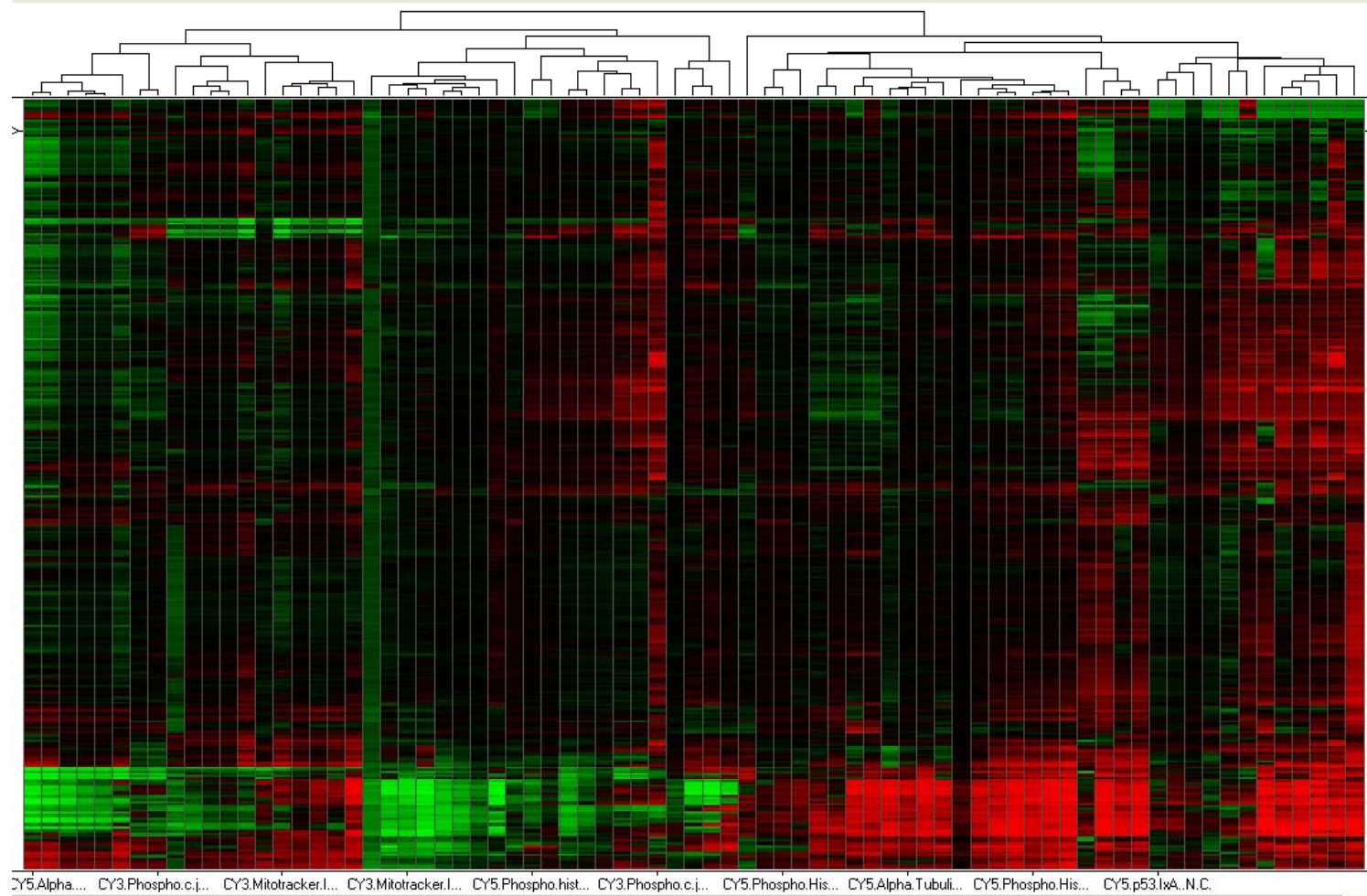




# Univariate: Differentiation

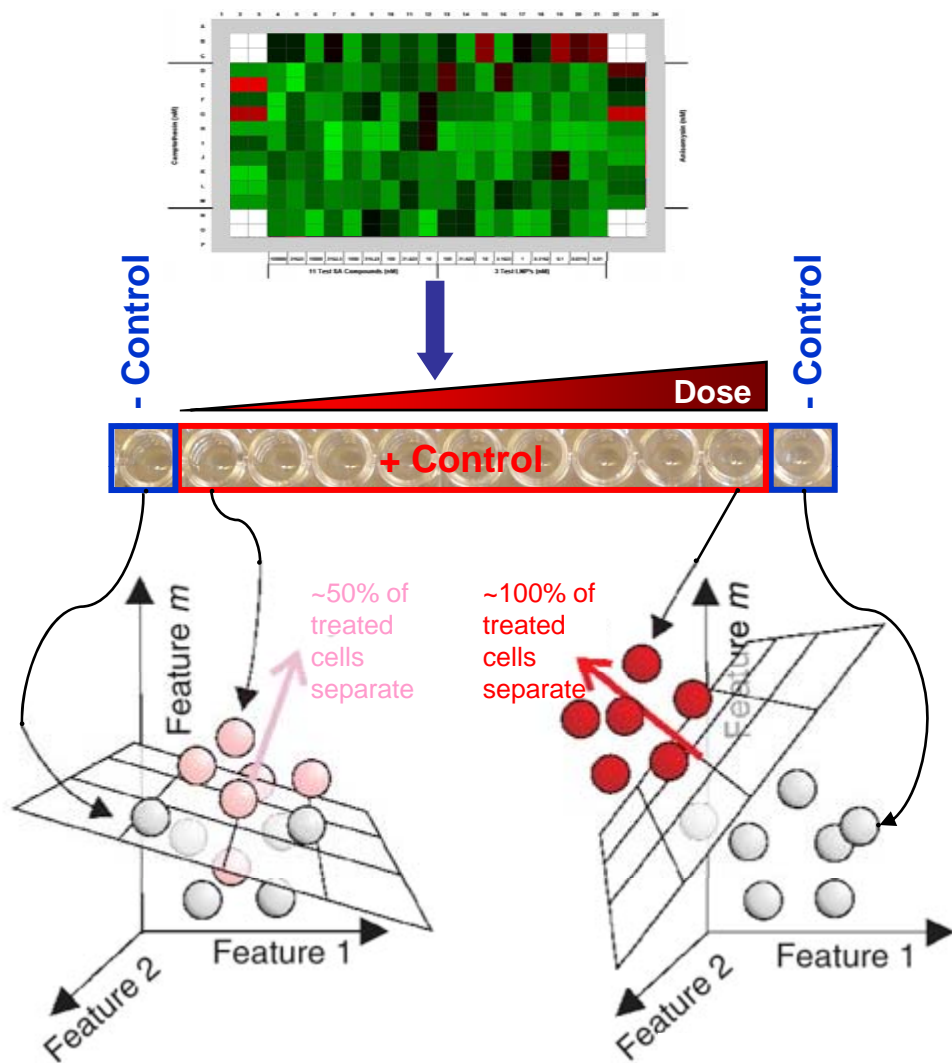


# But: Many features are correlated



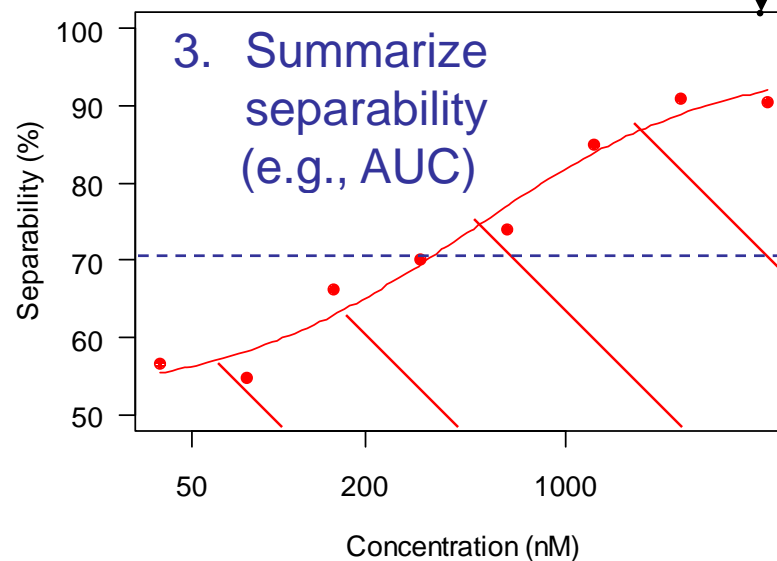
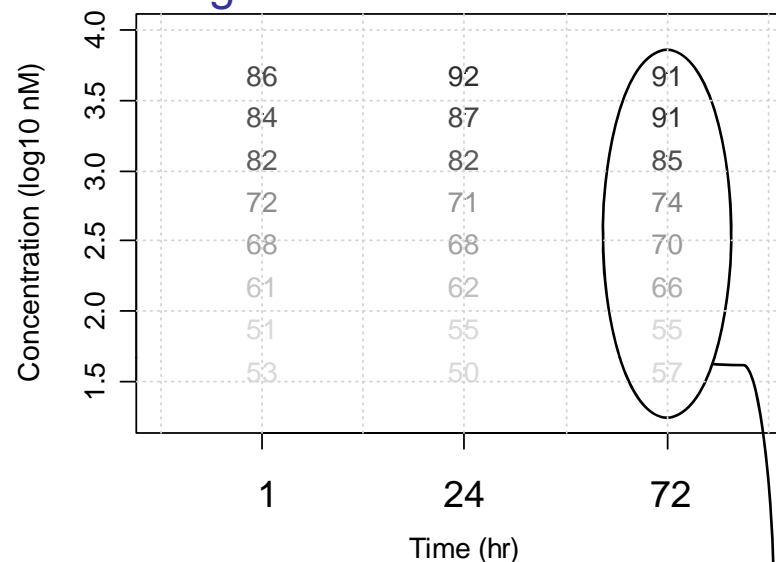
Consider multivariate methods: Support Vector Machine/Random Forest

# Multivariate: Evaluate Separability



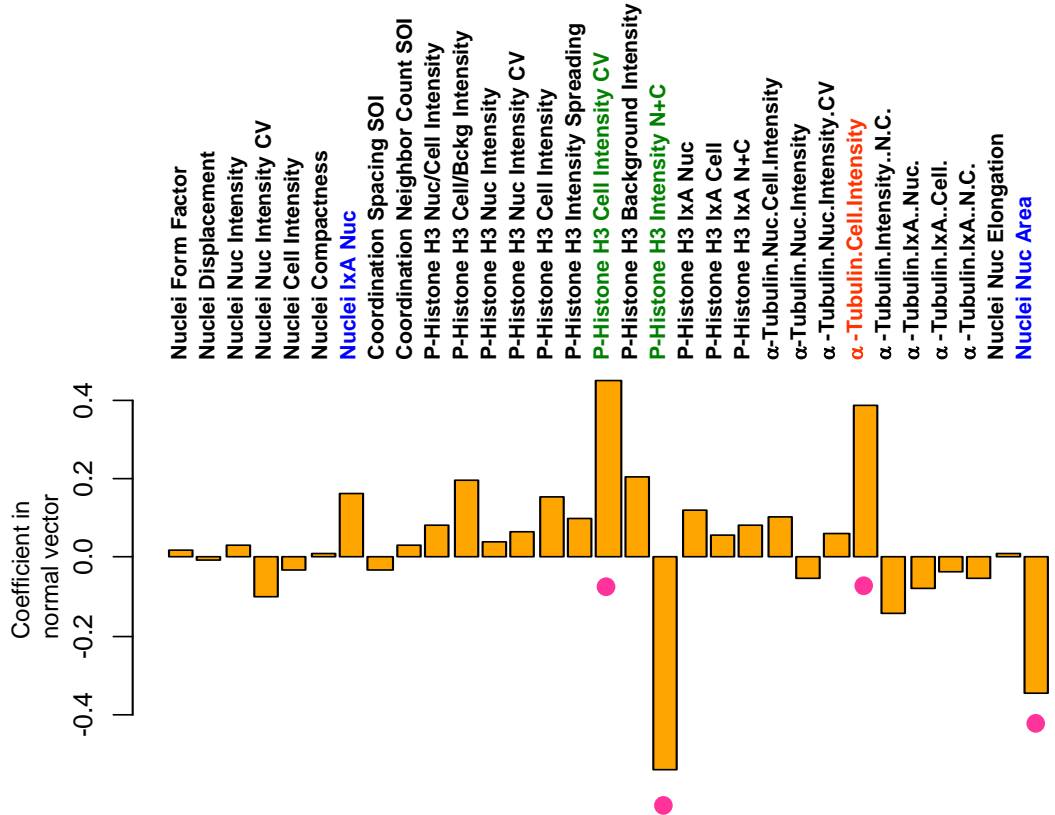
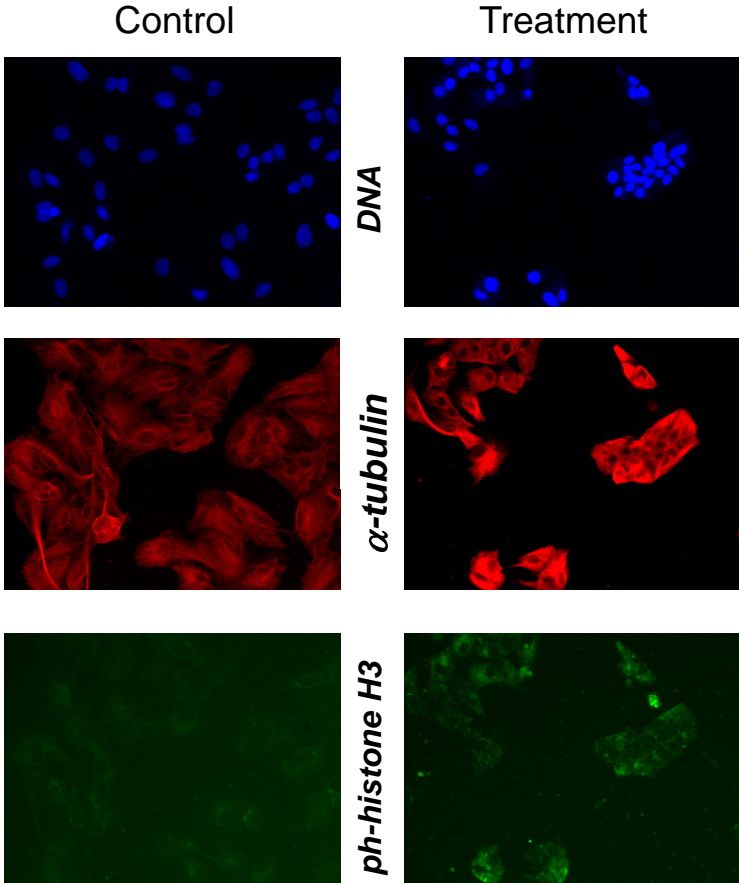
1. For each sample well, a SVM is constructed relative to control wells.

2. Resulting *separability* is a surrogate for differential behavior



3. Summarize separability (e.g., AUC)

## Extract meaningful distinguishing parameters



Images captured on INCell 1000 (GE Healthcare) - 20x objective

# Univariate vs. Multivariate

## Well Summary (KS)

### *Pros*

Very fast, space/memory efficient	No subpopulation analysis
Focuses on large-scale population shifts	Single summary for interleaving CDFs may be inadequate
Can combine plates/probe-sets	One variable at a time
Existing capabilities in SF	Multiple measures of separability

### *Cons*

## Cellular level (SVM/RF)

### *Pros*

Range of ML methods	Slower, memory intensive
Multivariate	Interpretability of results
Subpopulation analysis	Requires careful integration
Single continuous measure of separability	Relatively less familiar to biologists

### *Cons*

An optimal approach is often a combination of both

# Assay Quality

# Z'-factor

## **Integration of Multiple Readouts into the Z' Factor for Assay Quality Assessment**

**ANNE KÜMMEL, HANSPETER GUBLER, PATRICIA GEHIN,  
MARTIN BEIBEL, DANIELA GABRIEL, and CHRISTIAN N. PARKER**

Methods that monitor the quality of a biological assay (i.e., its ability to discriminate between positive and negative controls) are essential for the development of robust assays. In screening, the most commonly used parameter for monitoring assay quality is the Z' factor, which is based on 1 selected readout. However, biological assays are able to monitor multiple readouts. For example, novel multiparametric screening technologies such as high-content screening provide information-rich data sets with multiple readouts on a compound's effect. Still, assay quality is commonly assessed by the Z' factor based on a single selected readout. This report suggests an extension of the Z' factor, which integrates multiple readouts for assay

*Journal of Biomolecular Screening, November 25, 2009*

- Z'-factors are standard measures of assay quality in high throughput screens.
  - Reduce “amplitude and variability” of each assay measurement to a single parameter
- HCS provides multiple readouts
  - Not optimal to calculate Z'-factors separately
- Need for multivariate data analyses implies advanced methods to compute Z'.

# Z'-factors for HCS

## Probeset-specific Z'-factors

$$Z' = 1 - 3 ( (\sigma^+ + \sigma^-) / (|\mu^+ + \mu^-|) )$$

*where  $\sigma^+, \mu^+$  (or,  $\sigma^-, \mu^-$ ) represent the standard deviation and mean, respectively, of the calculated biological measures (e.g., AUCs) relative to the positive (or, negative) control*

Time (hr)	MMP	Cell Cycle Arrest	Cyto-skeletal Integrity	Stress Kinase Pathway	Oxidative Stress	Nuclear Integrity	DNA Damage
1	0.89	0.84	0.88	0.92	0.86	0.83	0.92
24	0.90	0.90	0.88	0.84	0.88	0.92	0.93
72	0.89	0.94	0.84	0.91	0.94	0.92	0.96

- Z' measures using multivariate techniques suggest very high assay quality

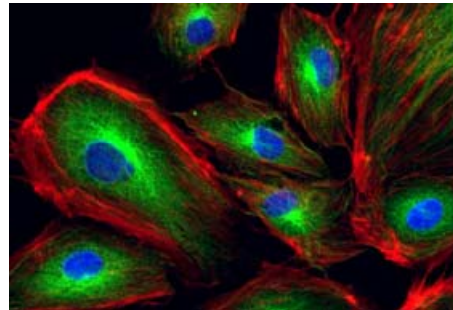
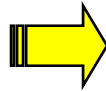


# Data Management *and* Navigation/Analysis Tool

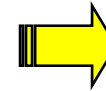
# Navigation through Multilayered Data



Set up and run experiment

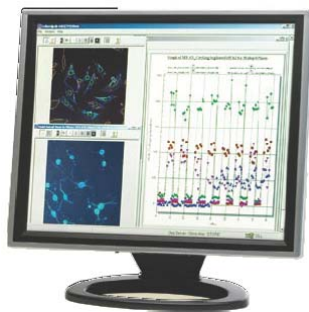
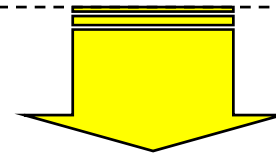


Capture images of cells

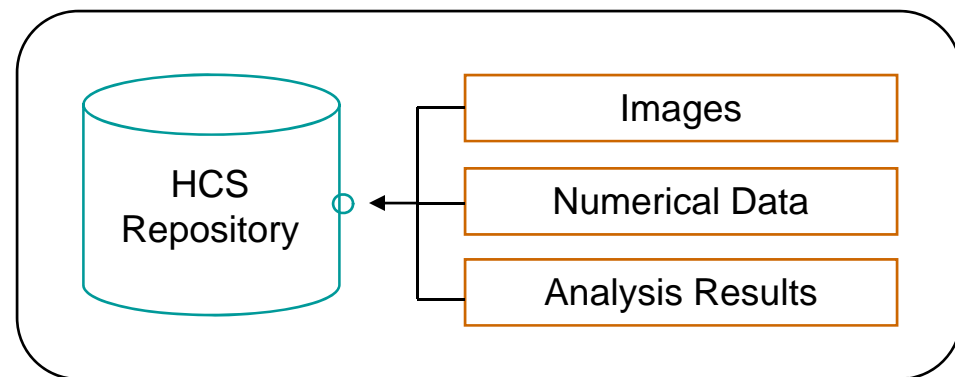
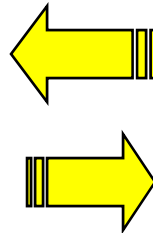


	W	C	N	N	C	N	I	J	C	I	M	N	
28	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate
29	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate
30	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate
31	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate
32	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate
33	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate
34	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate
35	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate
36	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate
37	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate
38	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate
39	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate
40	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate
41	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate
42	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate
43	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate
44	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate
45	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate
46	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate
47	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate
48	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate
49	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate
50	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate
51	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate
52	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate
53	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate
54	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate
55	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate
56	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate
57	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate
58	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate
59	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate
60	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate	Rate

Numeric descriptors



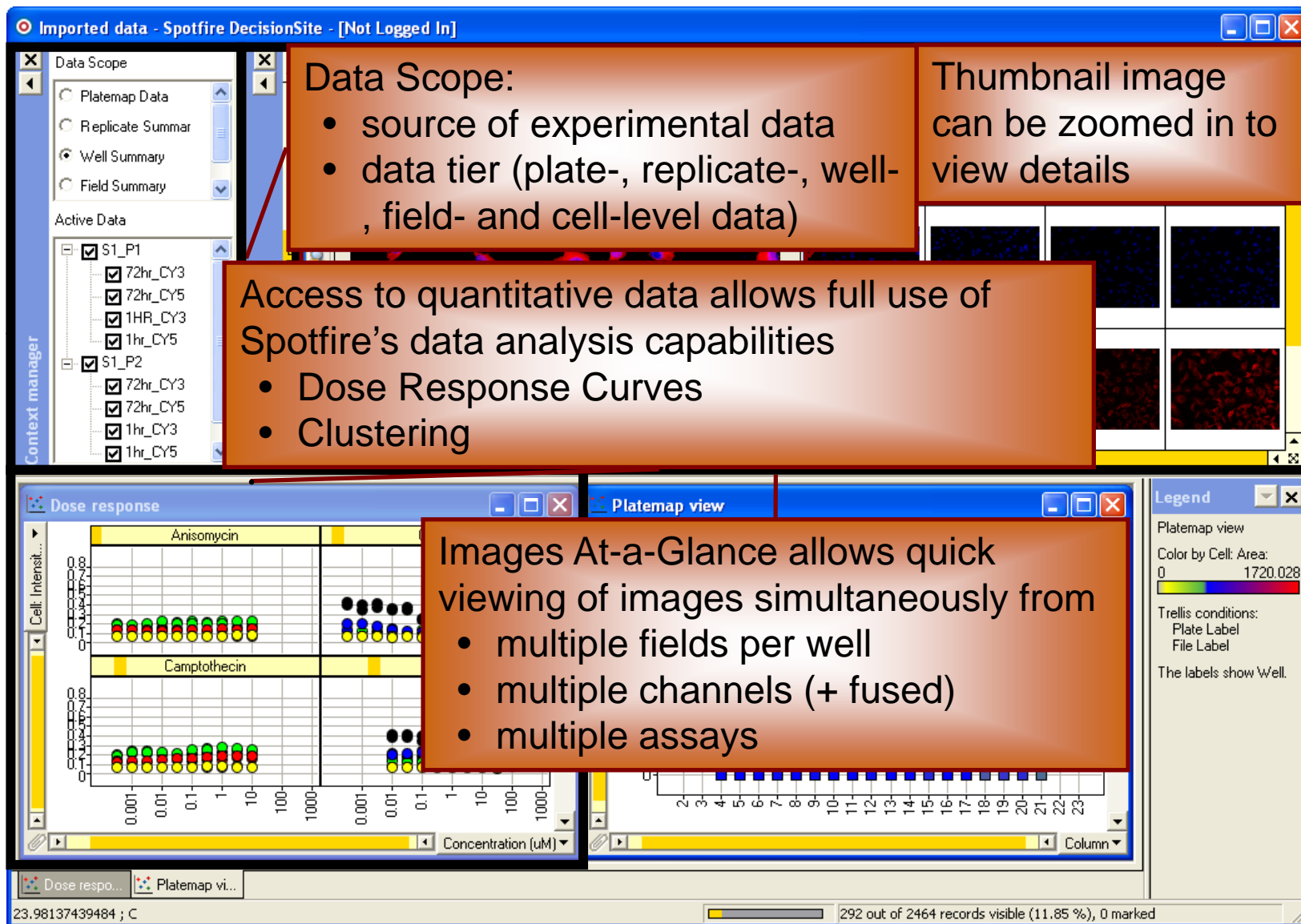
Visualization and Analysis Tool



# Tool Objectives

- To facilitate appropriate analyses of HCS experiments on a *single* platform:
  - navigation between raw data, processed data, and images
  - quality control of experimental data
  - visualization and analysis of data
  - Access to data mining tools
  - data export and sharing of analysis results

# HCS Tool At-a-Glance



# Desired Tool Characteristics

## What you do *not* want

- Multiple applications used: Excel, Prism, Spotfire, IN CELL Developer
- Image and numerical data can only be referenced by well label, with no linkage to experimental details– there is no associated “context”
- Well image can only be brought up one field at a time, one assay at a time
- Additional effort required to generate fused image
- Analytics performed only at the well-summary level [no capability to examine data beyond well summary]
- User can only examine data from one assay at a time
- Clunky, laborious, and error-prone

## HCS Tool that we would love to have

- Single platform (e.g., in Spotfire) to bring together numerical data and images
- Annotations are automatically applied to the data (via platemap) – images and numerical data can be examined “in context”
- Well images from multiple fields and assays can be viewed simultaneously
- Fused images are generated automatically
- Fast context switching between different data tier allowed: platemap, replicate, well-summary, field-summary, [cell data]
- User can examine multiple assays at a time (batch import allowed)
- Simple, easy, and error-proof

# Summary

- Seek balance between
  - Data formats: standardized vs. customized
  - Mining tools: device specific vs. robust
  - Analytical strategy: univariate vs. multivariate
  - Efficiency: high vs. low throughput (data volume)
  - Required level of rigor: high vs. low resolution
  - Scientific Goals: signature identification or characterization vs. classification
- Ease of navigation is highly desired
  - Connect the dots ...
  - Through complex, layered, multi-type data

# Acknowledgements

Amy Aslamkhan, **Amit Bahl**, **Wendy Bailey**, Randy Crawford, **Bonnie Howell**,  
**Ed Keough**, Adam Kozakov, Andrew Kraynak, Constantine Kreatsoulas, Steve  
Krotzer, Matt Kuhls, Kevin Leach, Liping Liu, Jim Monroe, **Irene Pak**, Francesca  
Santini, Joe Sina, Tom Skopek, Richard Storer, Dianne Umbenhauer, Jim Xu

**Laura Sepp-Lorenzino**

**Frank Sistare**

**Jeff Saltzman**