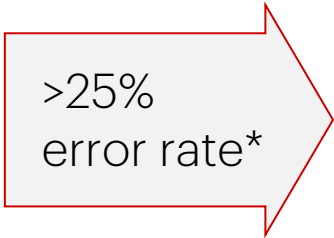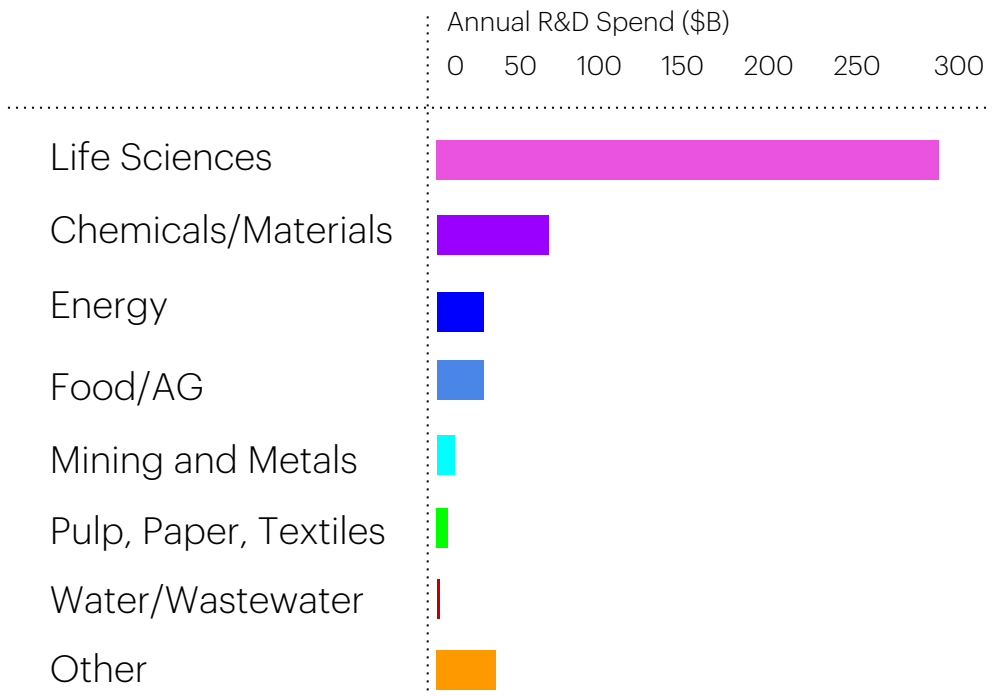*"Data that goes into a database should obey what I call the CAP principle. It should be complete, it should be accurate, and it should be permanent, so you never have to do it again.*

Otherwise, there's no progress."

— Sydney Brenner, Nobel Laureate, 2004

The problem in science today: >$100B of lost R&D productivity each year

Annual R&D Spend ($B)

Life Sciences	
Chemicals/Materials	
Energy	
Food/AG	
Mining and Metals	
Pulp, Paper, Textiles	
Water/Wastewater	
Other	

$420B
R&D Spend

>25%
error rate*

$105B
R&D losses

* Error rate = R&D results that fail to reproduce, leading to wasted resources, failed tech transfers and missed discoveries.

Sources:
2013 R&D Magazine Global Funding Forecast
NSF Science and Engineering Indicators 2012
CIA World Fact Book
Prinz, et al., Believe it or not: how much can we rely on published data on potential drug targets?, Nat. Rev. Drug Disc., 2011
Begley & Ellis, Raise Standards for Preclinical Cancer Research, Nature, 483, 2012
Ioannidis & Khoury, Improving Validation Practices in "Omics" Research, Science, 334, 2011
Halford, The Second Annual State of Translational Research, Sigma-Aldrich / AAAS, 2014
Freedman, et al., The Economics of Reproducibility in Preclinical Research, PLOS Biology, 2015
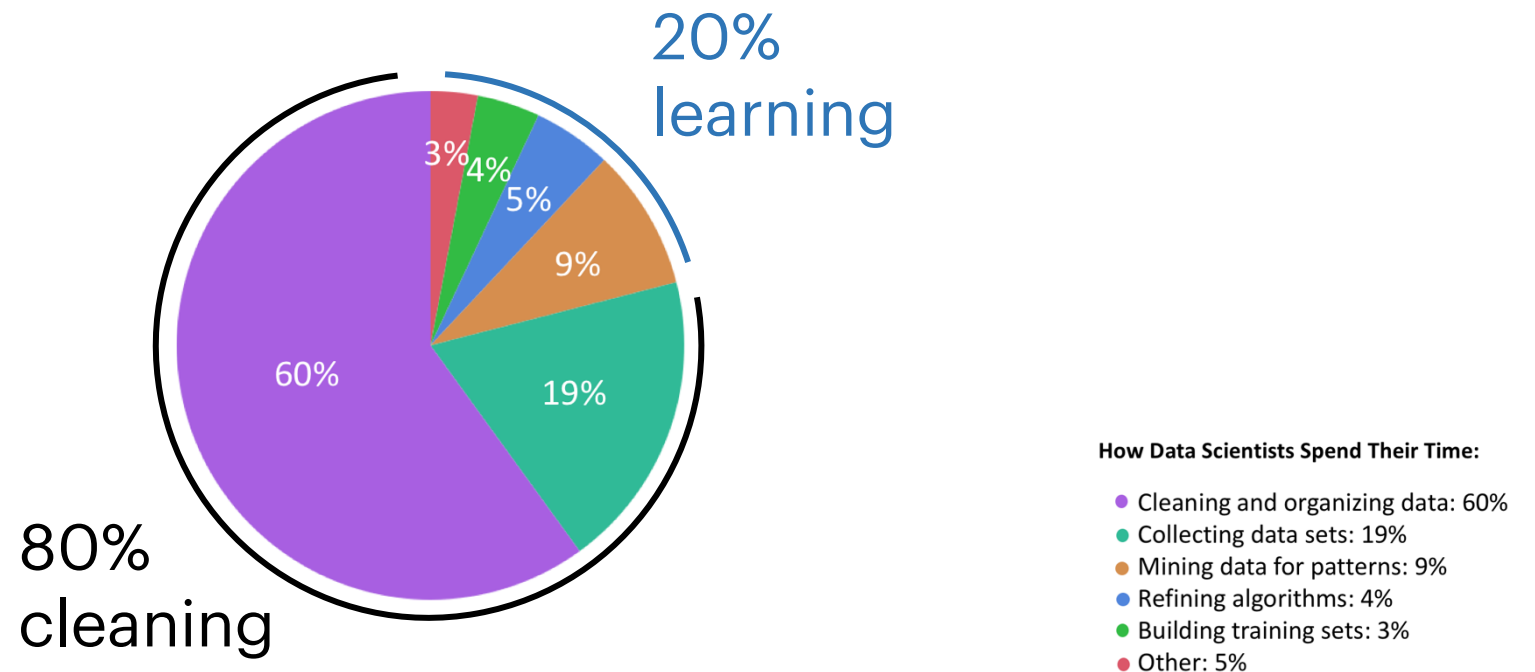
"It often takes us 2-3 months to assemble the data from a single development batch"

- senior scientist global pharmaceutical company

Researchers spend 80% of their time cleaning and organizing data, instead of learning from it

20%
learning

80%
cleaning

60%

19%

9%

5%

4%

3%

How Data Scientists Spend Their Time:

- Cleaning and organizing data: 60%
- Collecting data sets: 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Building training sets: 3%
- Other: 5%

Data from survey of ~80 data scientists conducted by CrowdFlower

Wrangler

Data Wrangler

Statements of pain in R&D

Biologics Bioprocessing

"Assessing the impact of batch variations on product quality is quite challenging."

Animal DMPK studies

"We want integrated data for rapid correlative analysis across workflow & sample genealogy."

Screening / Fermentation

"We are trying to escape Excel hell."

Antibody Bio-panning

"Our existing systems are too inflexible to support our changing processes."

R&D data issues
- quality
- access
- integration
- interpretation
- system flexibility

Clean data starts with quality experiments

But we, as a society, are failing to

to teach the principles,
develop the tools, &
build the culture

of quality in R&D

Examples from the front lines of science

The bad, the ugly and the hopeful

Examples selected from stages of the typical biotech development lifecycle

cell line
screening

bioprocess
process
development

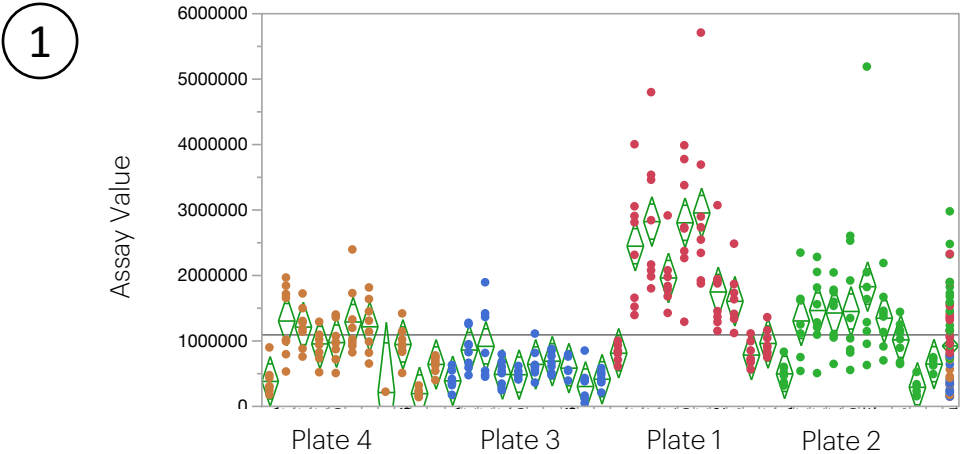downstream
process
development

analytical
chemistry

animal DMPK
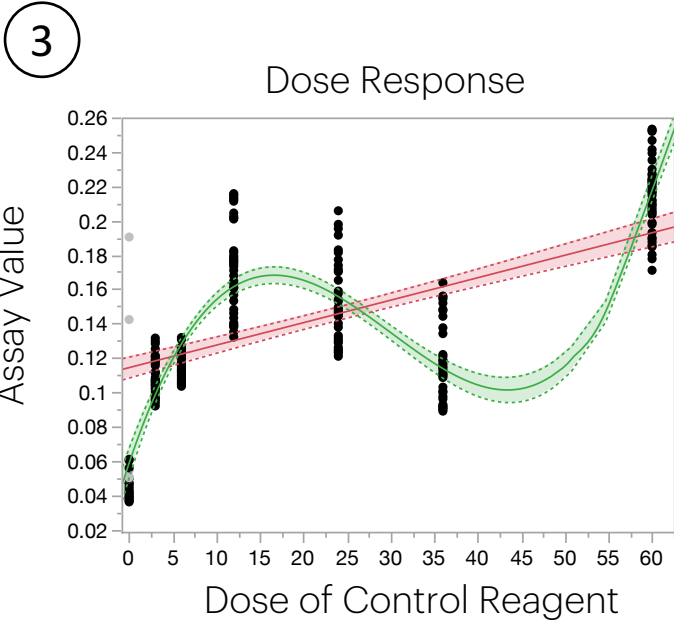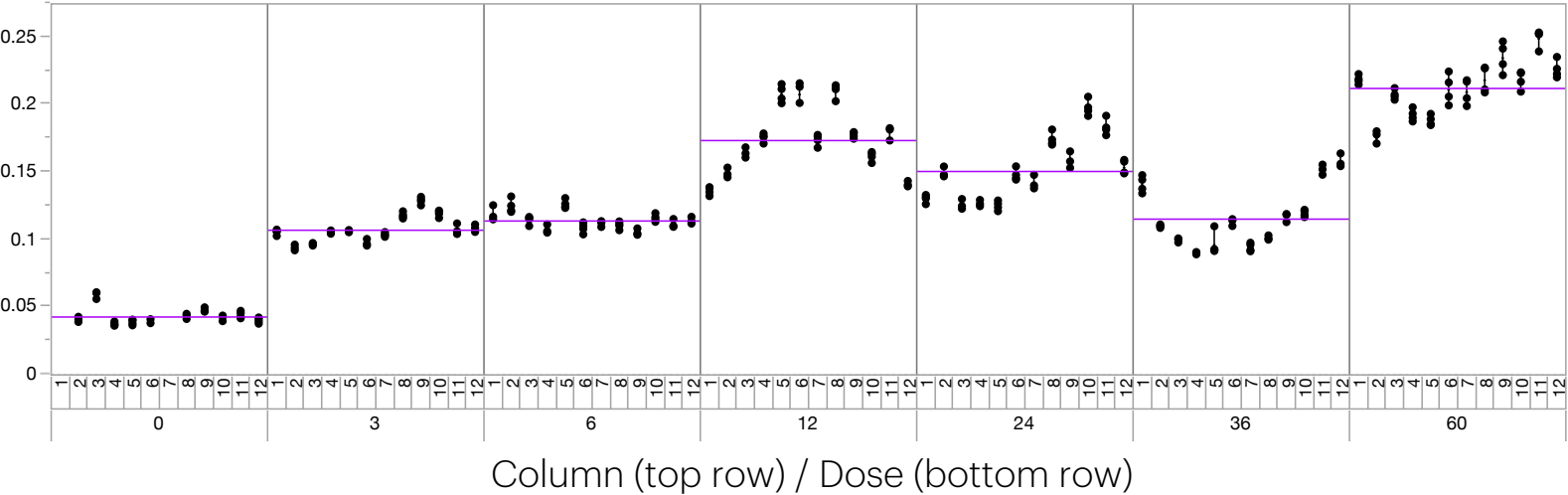studies

Example 1:
Unreliable microtiter plate assay
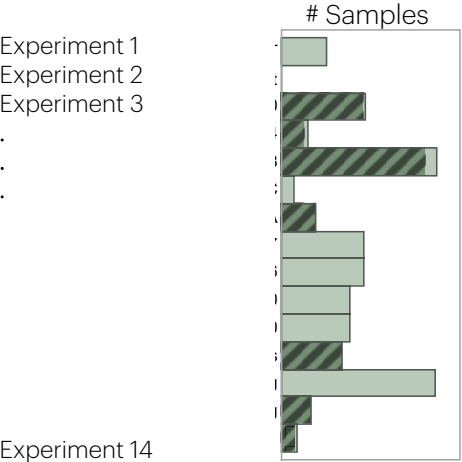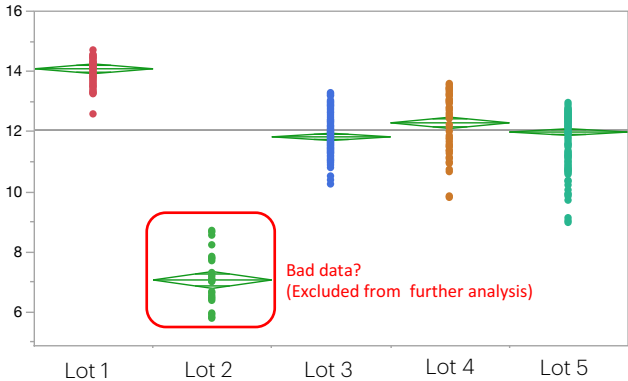

cell line screening

Example 2:
"Regression to the mean" – screening hits don't hold up over time

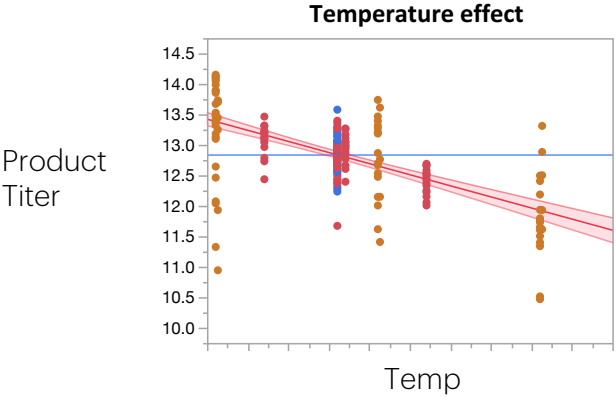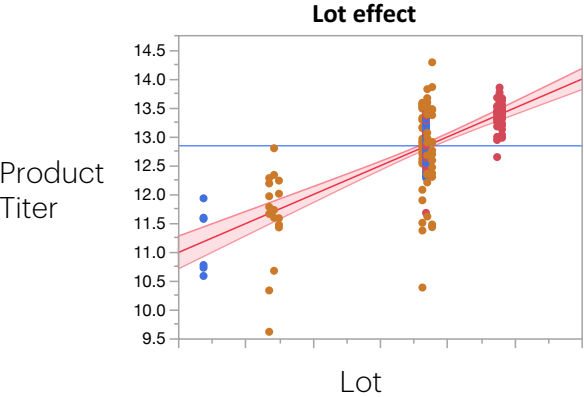**1** Collected aggregate experiment data set from Riffyn. Identified common controls

2 Assessed media lot-to-lot effects on control strains

Bad data?
(Excluded from further analysis)

<0A-> Prepare corn mash | i | corn mash | resource name

3 Model fit to control for lot-effects and temperature effects. Most of variation in product titer attributable to these 2 factors, not the candidate strains


4 CONCLUSION: Candidate strains thought to be hits were not hits (indistinguishable from parent).

Strain							Least Sq Mean
Strain 1	A						13.629426
Strain 2	A	B	C				13.579121
Strain 3	A	B	C				13.571188
.	A	B	C	D			13.454254
.	A	B		D			13.284823
.	A	B	C	D	E		12.946421
	A	B	C	D	E		12.901854
		B	C	D	E		12.897631
Reference Strain 1		B		D	E		12.896804
	A	B	C	D	E		12.822106
	A	B	C	D	E		12.802635
.	A	B	C	D	E	F	12.793868
	A	B	C	D	E	F	12.764988
	A	B	C	D	E		12.764021
.	A	B	C	D	E	F	12.759558
.	A	B	C	D	E	F	12.747076
.	A	B	C	D	E	F	12.737596
	A	B	C	D	E		12.722621
	A	B	C	D	E	F	12.715914
	A	B	C	D	E	F	12.712887
	A	B	C	D	E		12.705254
	A	B	C	D	E	F	12.666850
	A	B	C	D	E	F	12.659794
	A	B	C	D	E		12.636188
	A	B	C	D	E	F	12.512988
	A	B	C	D	E	F	12.494354
	A	B	C	D	E	F	12.473626
	A	B	C	D	E	F	12.423954
	A	B	C	D	E	F	12.340154
	A	B	C	D	E	F	12.319121
	A	B	C	D	E	F	12.307854
	A	B	C	D	E	F	12.299754
	A	B	C	D	E	F	12.268421
Reference Strain 2			C			F	12.231741
.	A	B	C	D	E	F	12.196254
.	A	B	C	D	E	F	12.195421
	A	B	C	D	E	F	12.190721
.Strain N	A	B	C	D	E	F	12.176354

Example 3:
Pre-clinical testing results don't match across independent sites

animal DMPK studies

Same compound and protocol across 2 R&D sites

Testing result

MEAN

MEAN

○ Testing Day 1
○ Testing Day 2
○ Testing Day 3

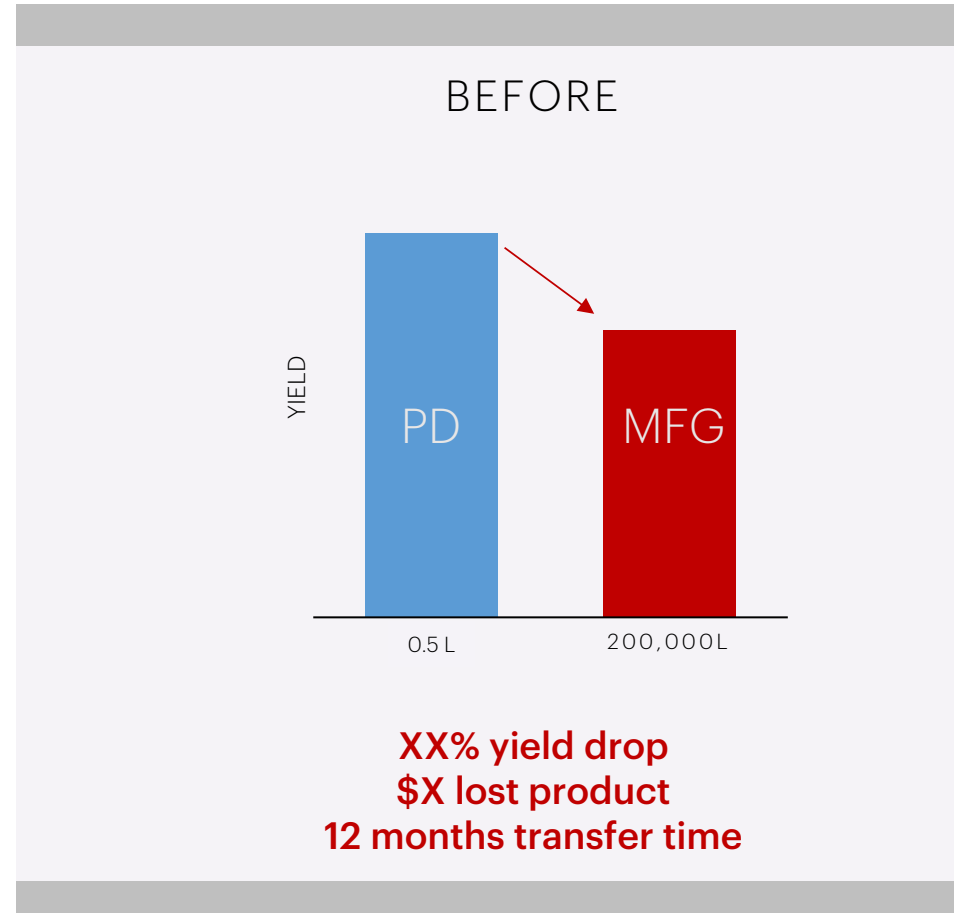Site 1   Site 2
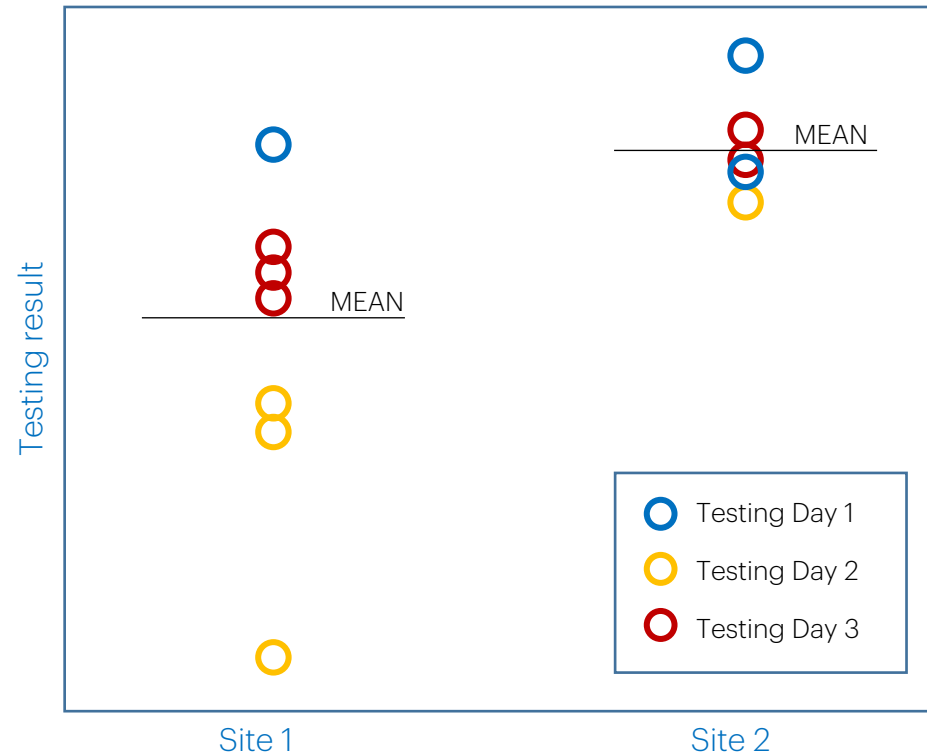
Example 4:
Fermentation scale-up—troublesome tech transfers to full-scale manufacturing

bioprocess
process
development
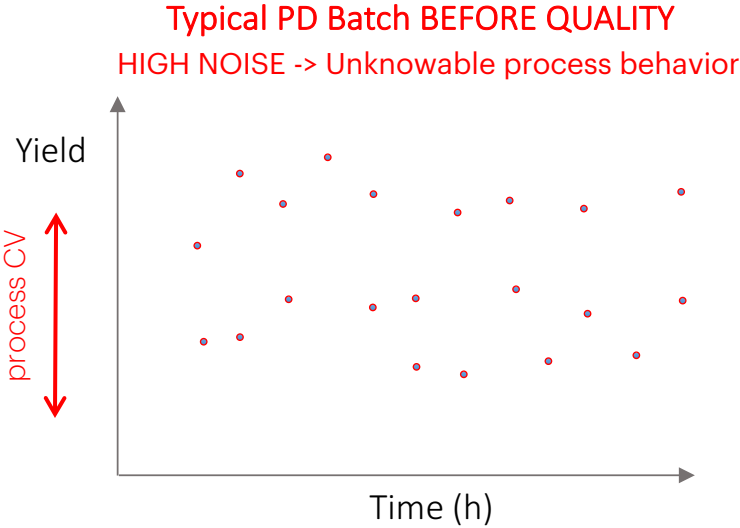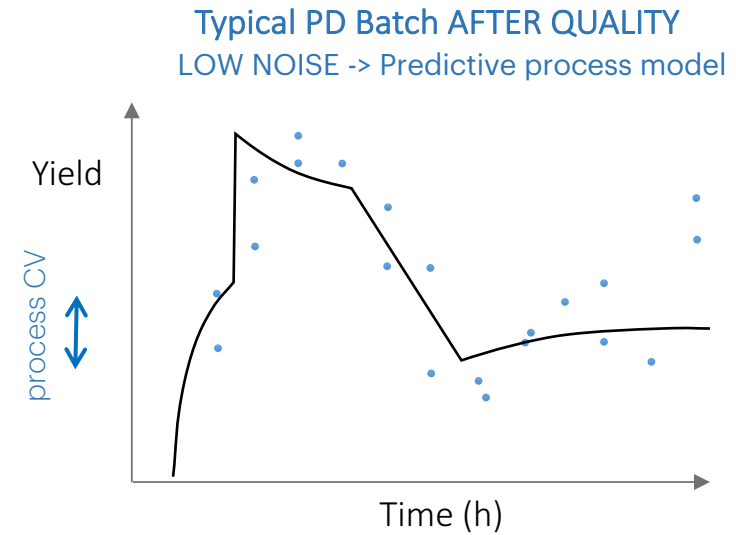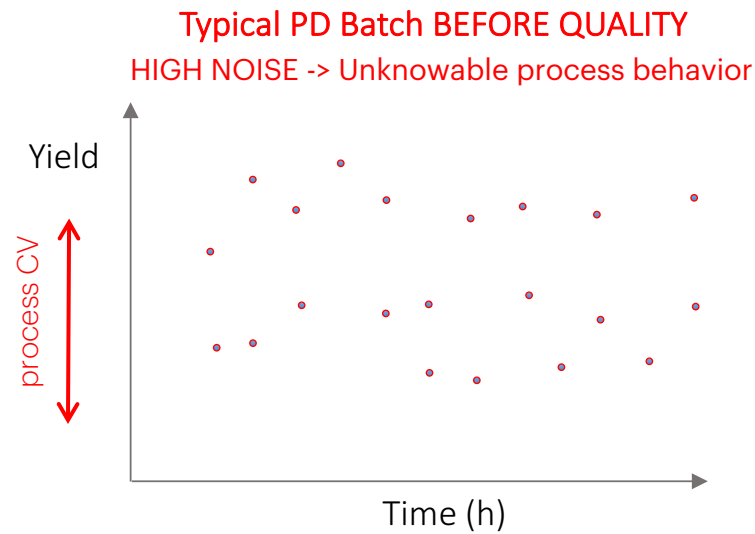
analytical
chemistry

BEFORE

YIELD

PD

MFG

0.5 L

200,000L

XX% yield drop
$X lost product
12 months transfer time

Example 4:
Fermentation process development—initial state

Quality improvement campaign

Typical PD Batch BEFORE QUALITY

HIGH NOISE -> Unknowable process behavior

Yield

process CV

Time (h)

Original product assay

% Recovery of Standard

2.5% CV

98% accuracy

Redeveloped product assay

% Recovery of Standard

0.3% CV

99.9% accuracy

16

Example 4:
Fermentation process development—after quality campaign

Typical PD Batch BEFORE QUALITY

HIGH NOISE -> Unknowable process behavior

Yield

process CV

Time (h)

Typical PD Batch AFTER QUALITY

LOW NOISE -> Predictive process model

Yield

process CV

Time (h)

Example 4:
Fermentation scale-up: before & after quality improvement and data integration

AFTER

YIELD

PD

MFG

0.5 L

200,000L

**Exact performance match
across scales
3-4 months transfer time**
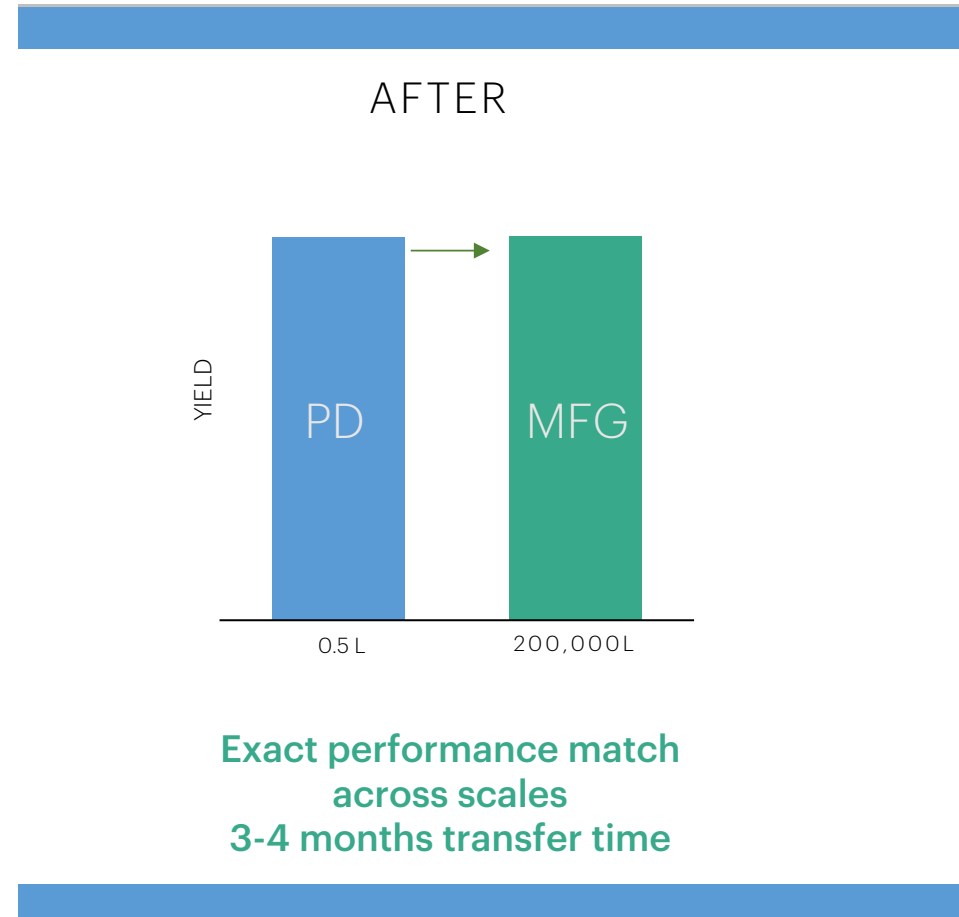
Example 5:
High-throughput strain screening artifacts and unreproducible hits

96 well plate data

Root cause

Assay value
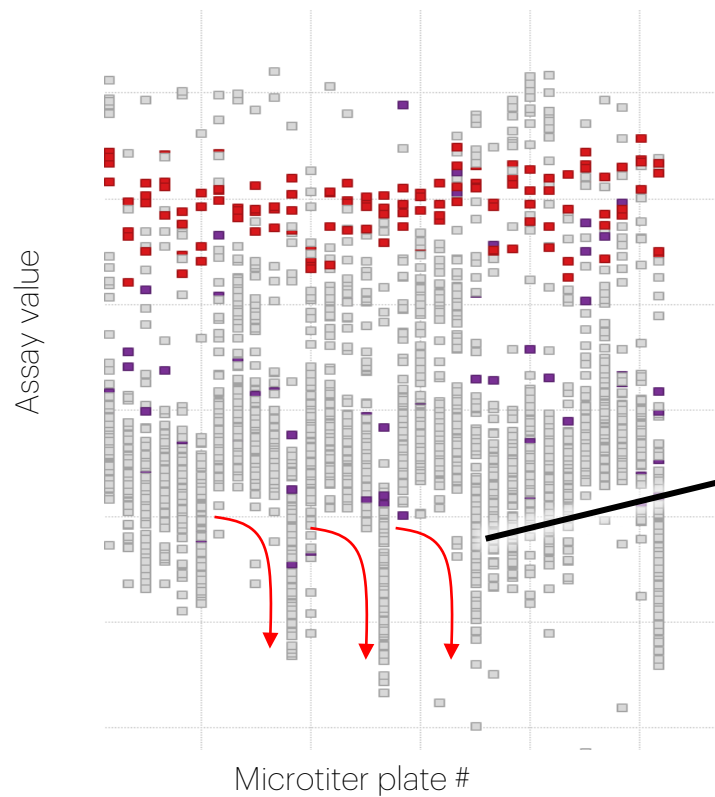
Microtiter plate #
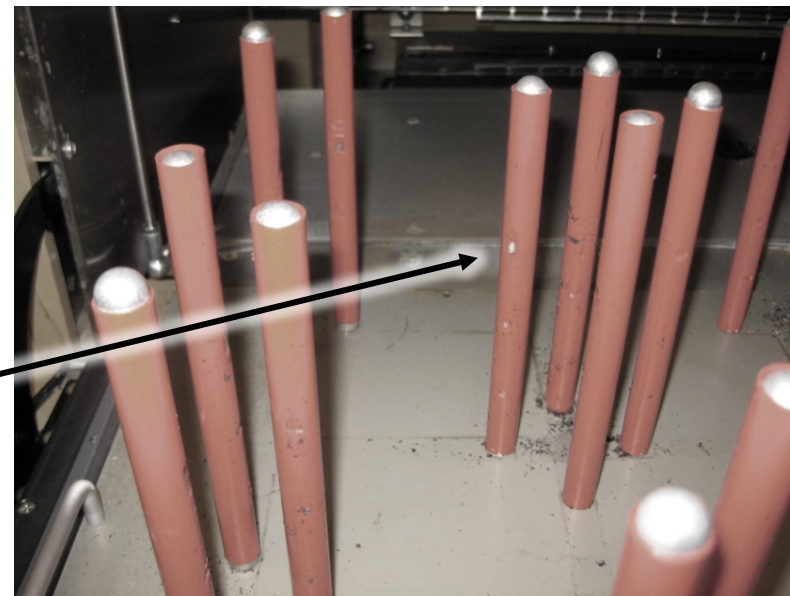
6X reduction in screening error doubles rate of strain improvement

BEFORE VARIANCE REDUCTION
1X R&D pace

AFTER VARIANCE REDUCTION
2X R&D pace

30% relative error

5% relative error

6X error reduction

2X

Product yield of a strain

1

0.5

0

2007   2008   2009   2010

Year
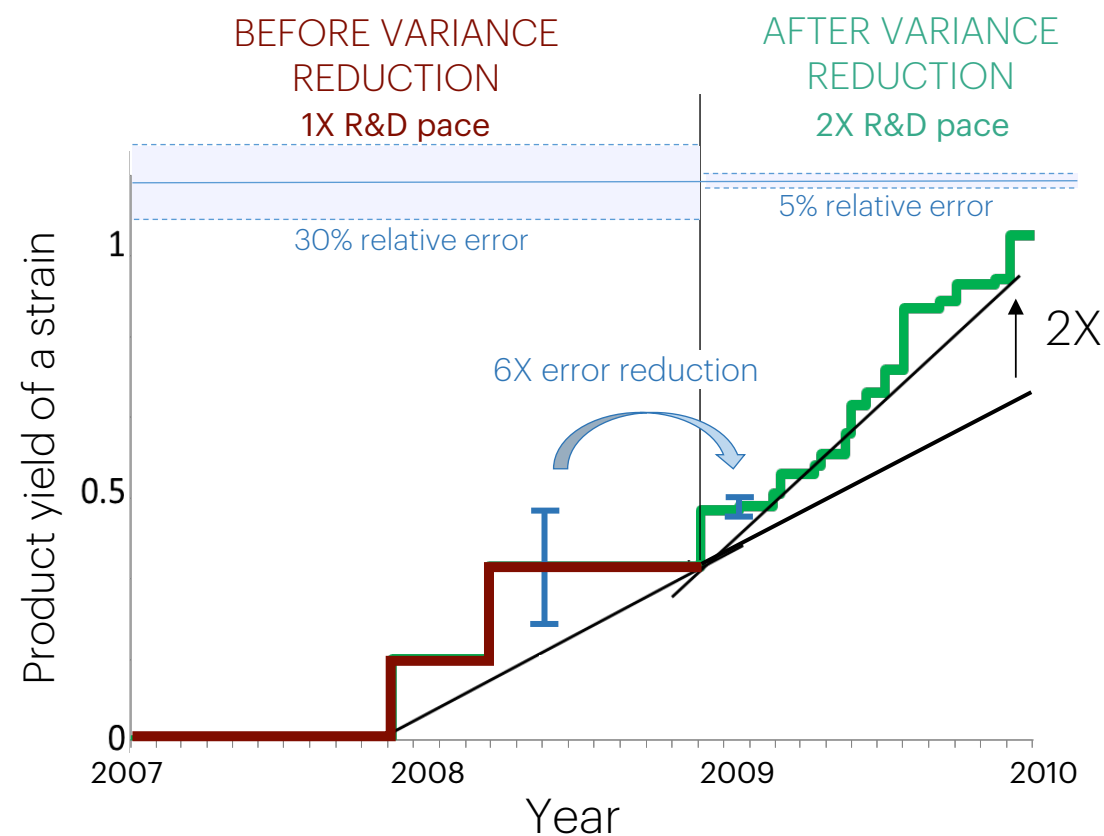
Source: Gardner, TS (2013) Trends in Biotech. 31:3, 123-125.

Improved data quality and integration revolutionizes the auto industry

Automobile manufacturing speed
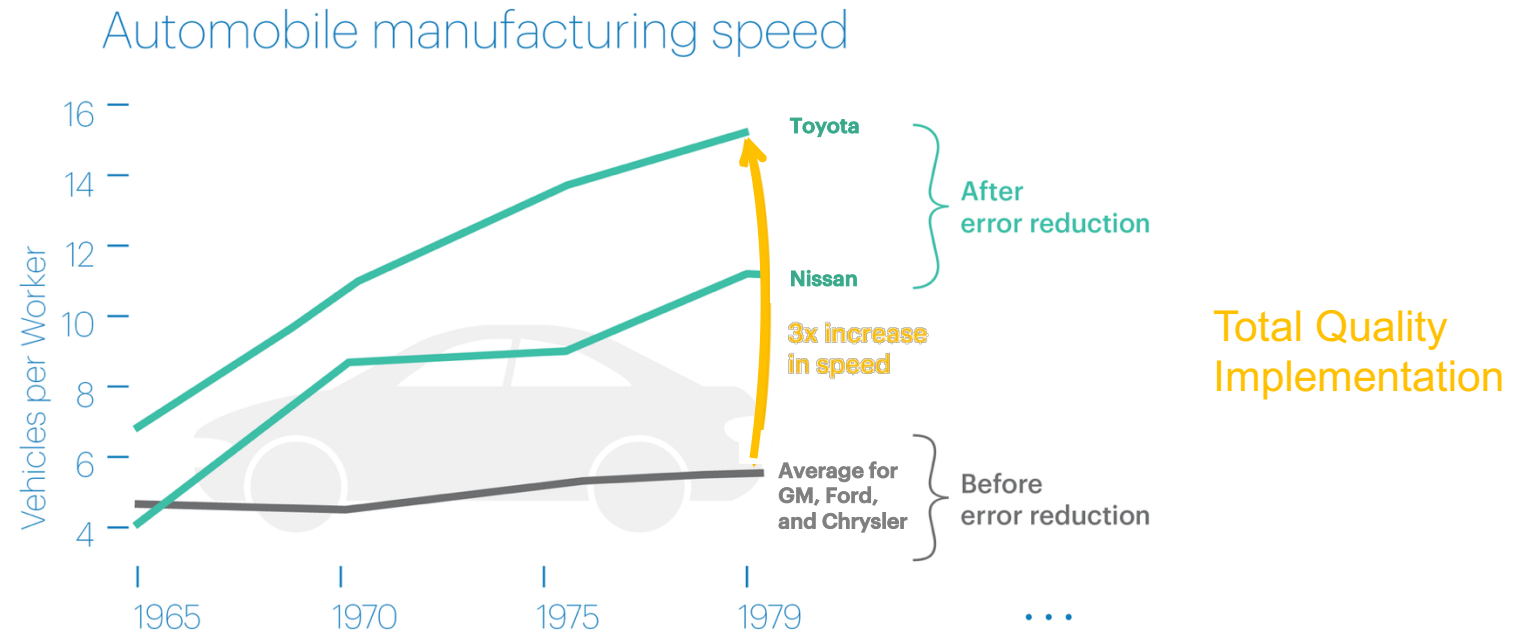
Derived from: M.A. Cusuman (1985) The Japanese Automobile Industry: Technology and Management at Nissan and Toyota, Harvard University Press

Lessons of manufacturing quality

DMAIC (6-sigma) (Mfg)	TRANSLATION TO R&D (R&D)
~~Define~~ →	Design
Measure	Measure
Analyze / Improve	Analyze / Improve
~~Control~~ →	Share
	Iterate

Clean data starts with quality experiments

But we, as a society, are failing to

to teach the principles,
develop the tools, &
build the culture

of quality in R&D

We need to recognize our experiments as **measurement systems**

and teach our future scientists accordingly

Riffyn's mission is to help scientists deliver research we can trust and build upon efficiently.

Riffyn.

Never Miss a Discovery™