Modeling, Data Analytics, and Machine Learning for Process Development and Verification

Richard D. Braatz

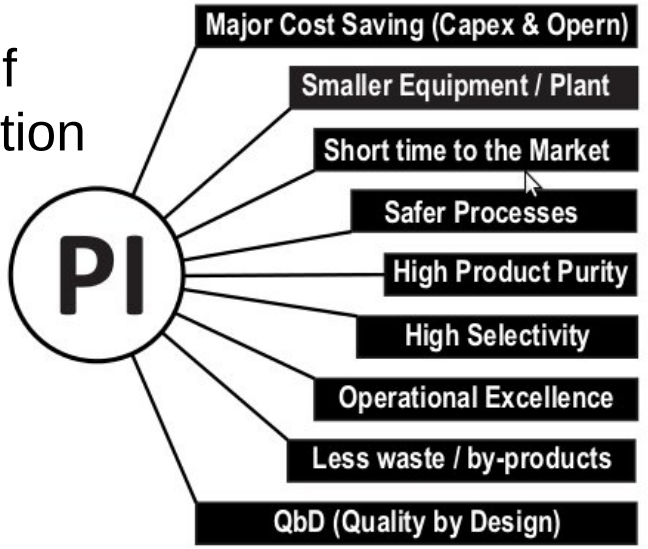
Massachusetts Institute of Technology

Outline

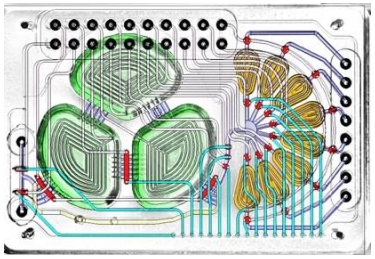
- Current state of the art in using models, data analytics, and machine learning
- Looking to the future in process models and data analytics

Current State of the Art in Process Development

- Greatly increased understanding & optimization of each unit operation, exploiting process intensification

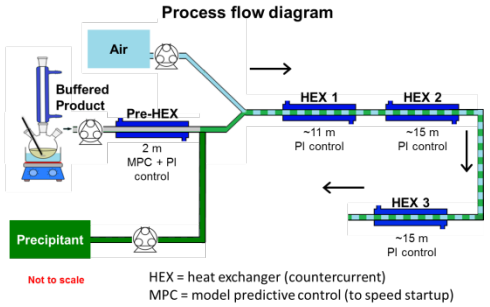


- Automated high-throughput microscale technology for fast process R&D

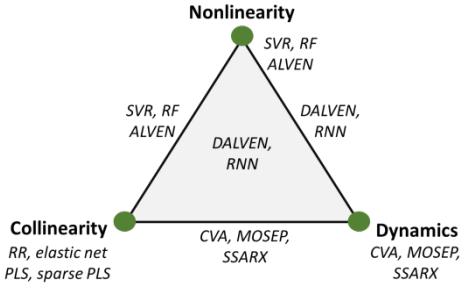


- Plug-and-play modules w/integrated control & monitoring to facilitate deployment

- Dynamic models (w/uncertainty) for unit operations for automated plant-wide simulation & control design



- Autonomous/smart data analytics/machine learning



All strategies based heavily on models

PI image from pi-inc.co; puzzle from web; other images from Braatz & R Ram pubs

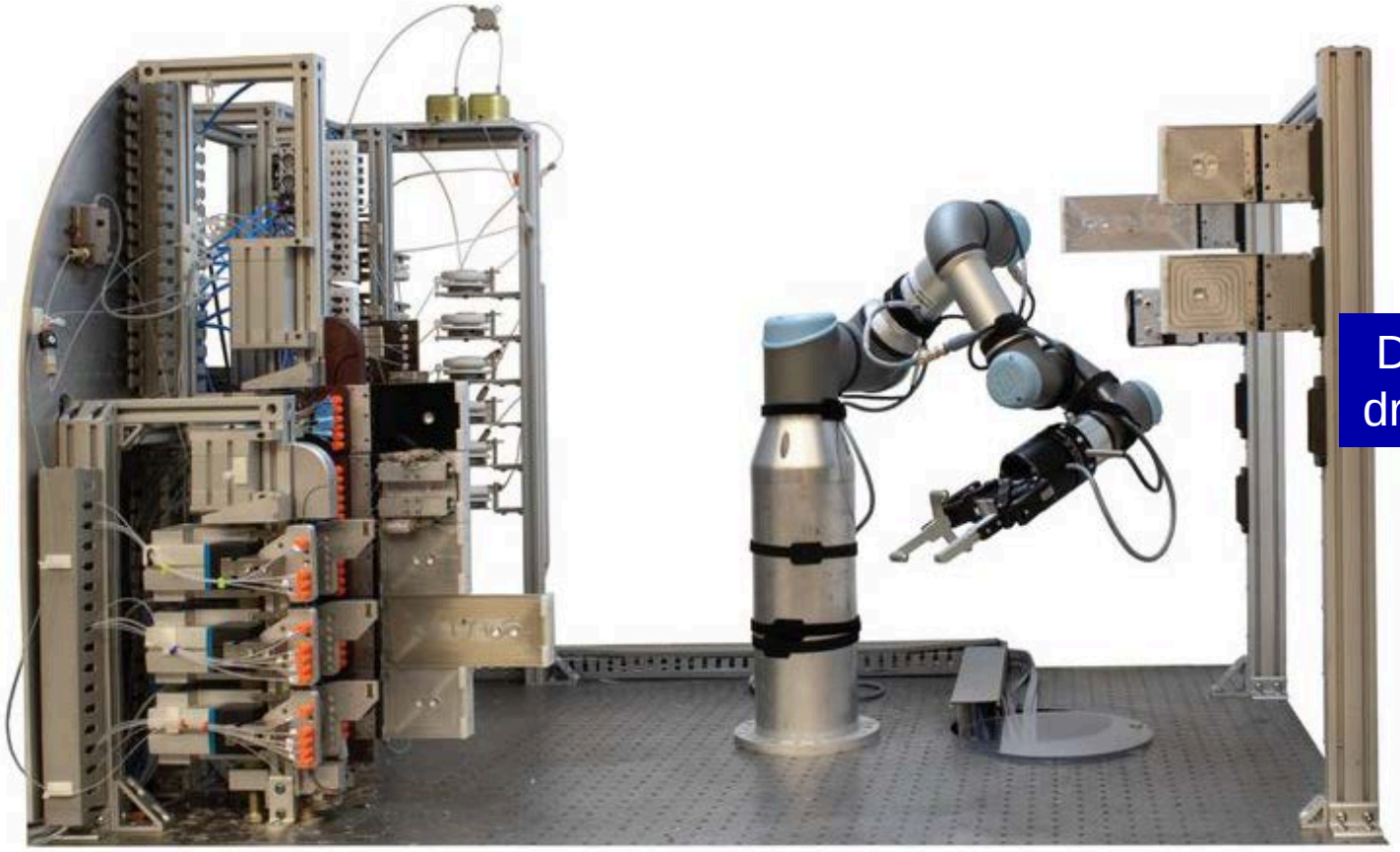
Many published applications of the strategies to pharmaceutical manufacturing

- Costs reduced by ~50%
- Methods are being applied in the pharma industry*

S Mascia et al., *Angewandte Chemie*, 52, 12359-12363, 2013; Highlight in *Nature*, 502, 274, 2013
* E Içten et al., *AIChE Annual Meeting*, 2019

Many published applications of the strategies to pharmaceutical manufacturing

An automated molecular synthesizer produces, purifies, and characterizes using flowsheet models, process intensification, optimized plug-and-play fluidic modules, and feedback control



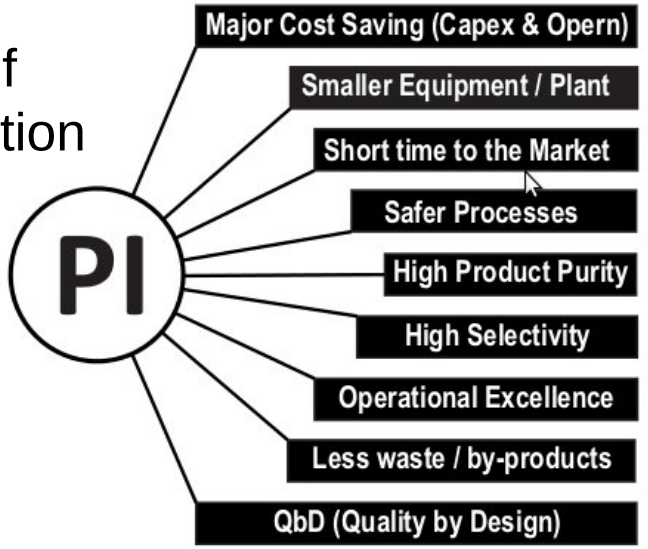
Demonstrated for 15 drug products in 2019

Software tools being used by academia and companies, mlpds.mit.edu/

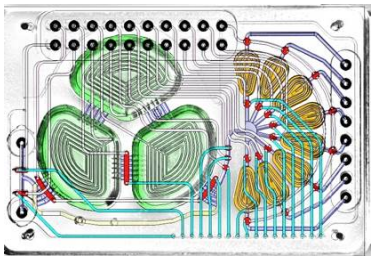
C Coley, ..., T Jamison, K Jensen, *Science*, 365(6453), 2019

Current State of the Art in Process Development

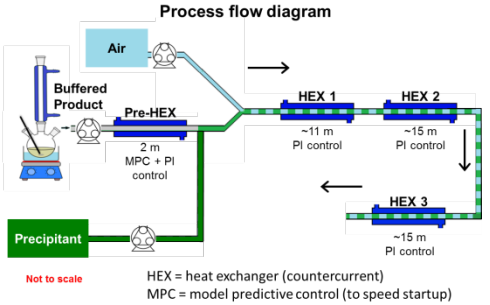
- Greatly increased understanding & optimization of each unit operation, exploiting process intensification



- Automated high-throughput microscale technology for fast process R&D

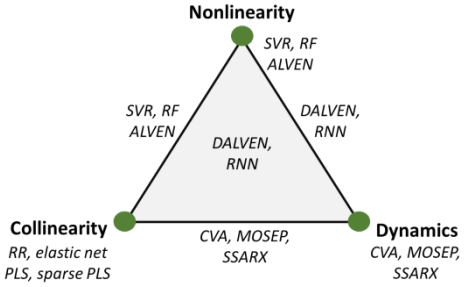


- Plug-and-play modules w/integrated control & monitoring to facilitate deployment



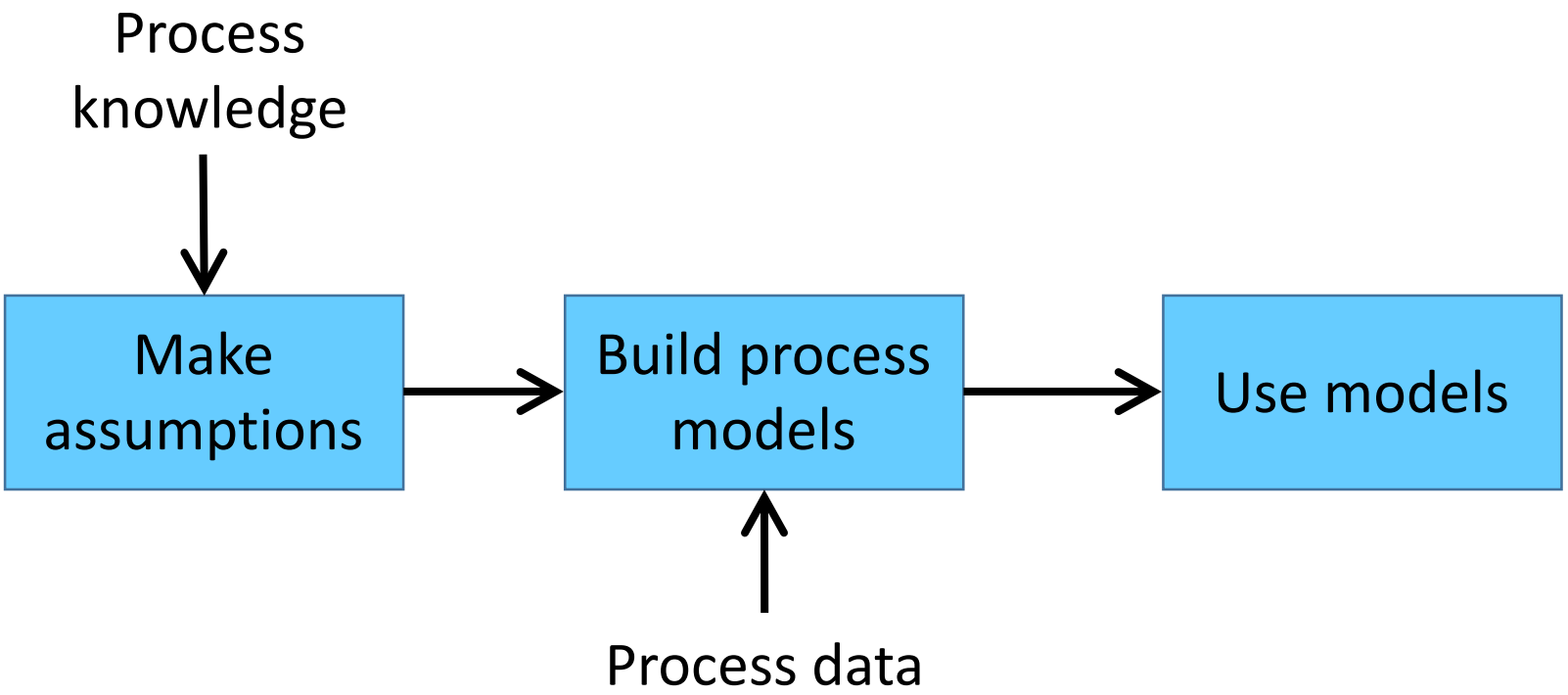
- Dynamic models (w/uncertainty) for unit operations for automated plant-wide simulation & control design

- Autonomous/smart data analytics/machine learning

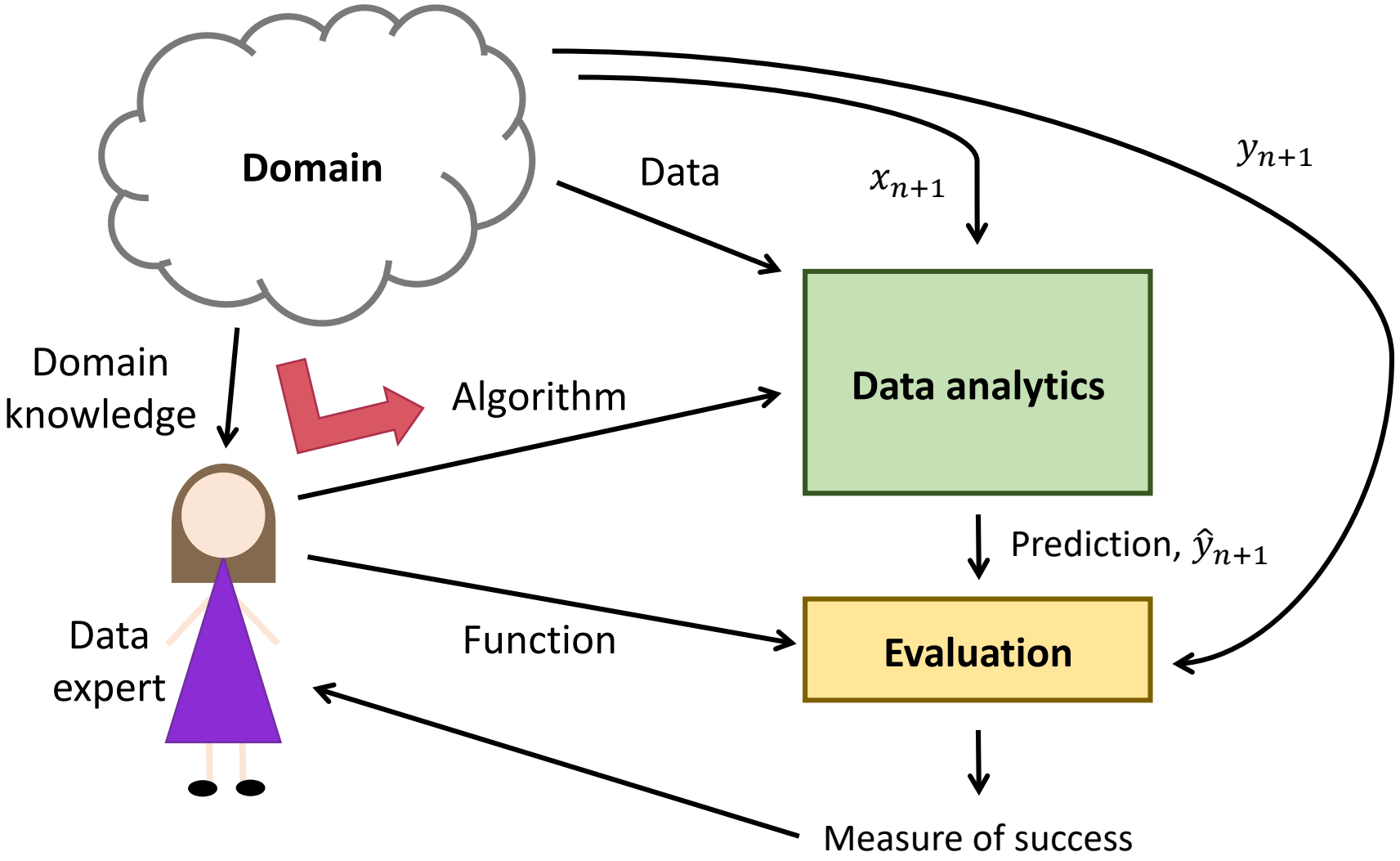


PI image from pi-inc.co; puzzle from web; other images from Braatz & R Ram pubs

Current status of process data analytics in the pharmaceutical industry



Current status of process data analytics in the pharmaceutical industry



Based on L. Kaelbling's lecture on September 2014 and K. Severson's lecture on May 2018

Selecting the best data analytics tool

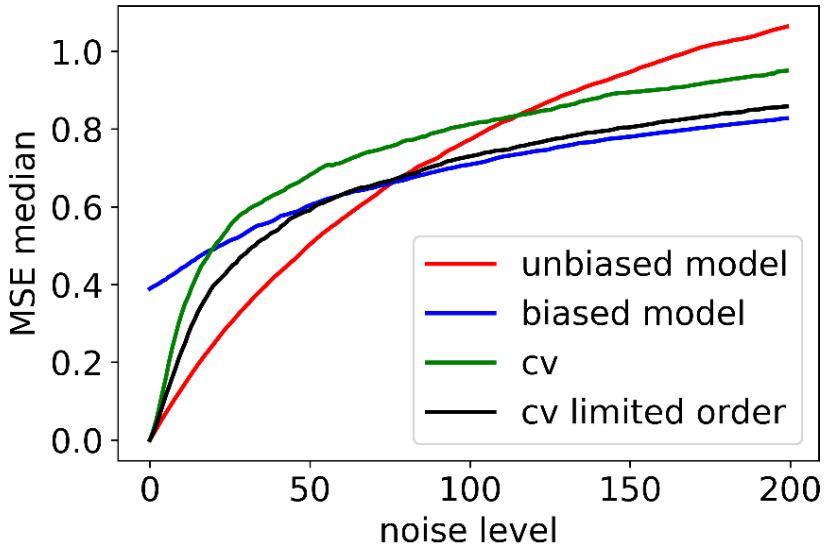
- A substantial level of expertise is required to select the best data analytics tool for a particular application
- Tools come from chemometrics, applied statistics, pattern classification, computer science, etc.
- In practice, users apply the tool(s) that they know, which can produce suboptimal results
- The alternative is a systematic approach for process data analytics tool selection that allows the user to focus on objectives rather than methods

Alternative 1: Selecting the best of many models by minimizing observed cross-validation error

Consider a process $y = 0 + x + 0.5x^2 + 0.1x^3 + 0.05x^4 + \epsilon$ with 30 samples

The data were fit to four types of models:

1) 2nd order ($y = ax + bx^2 + k$, biased)
2) 4th order ($y = ax + bx^2 + cx^3 + dx^4 + k$, unbiased)
3) polynomial order of 1 to 10 selected by MCCV
4) polynomial order of 2 and 4 selected by MCCV

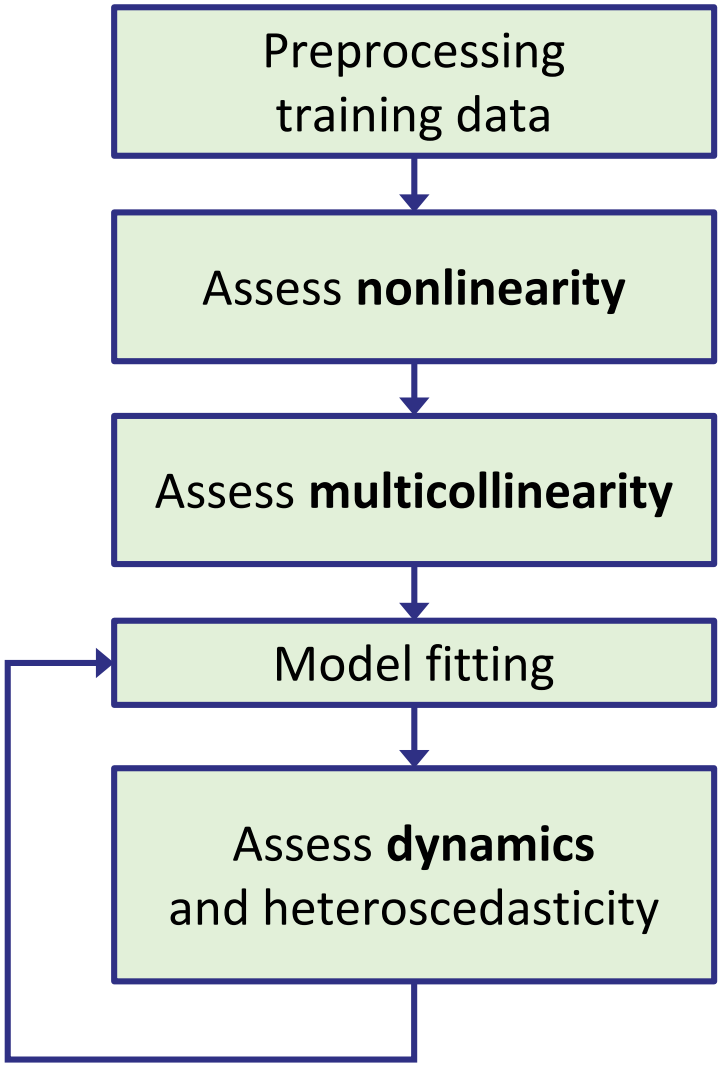


Key takeaways:

- Selecting over many methods results in **larger true prediction errors**
- The best method for a specific problem depends on the specific data characteristics
- Pre-select a **small** number of methods known to be suitable for the data and type of application

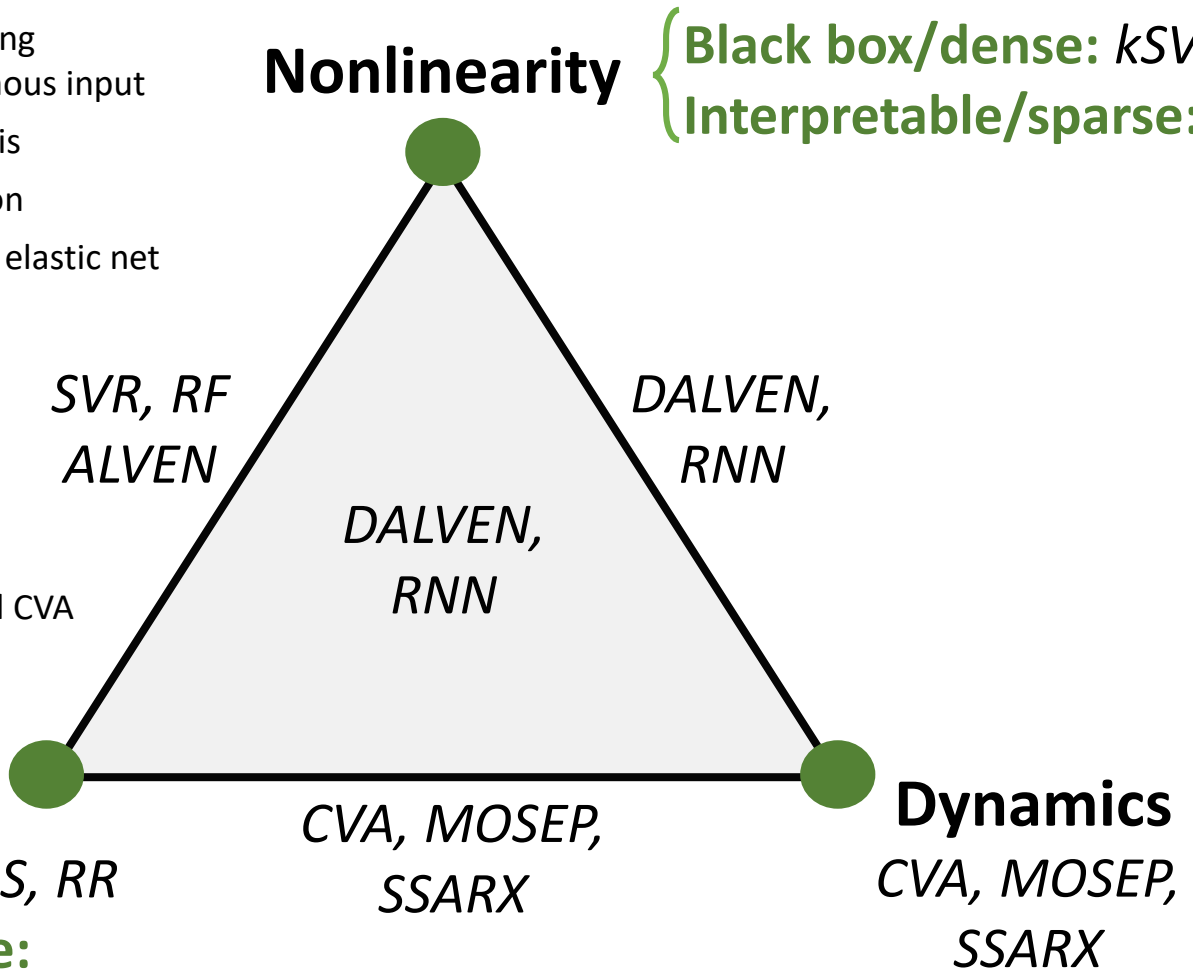
Alternative 2: Select best-in-class method based on the problem type and data characteristics

- First interrogate the dataset to ascertain its characteristics
- Second, apply the best-in-class DA/ML methods for data with those characteristics



Alternative 2: Select best-in-class method based on the problem type and data characteristics

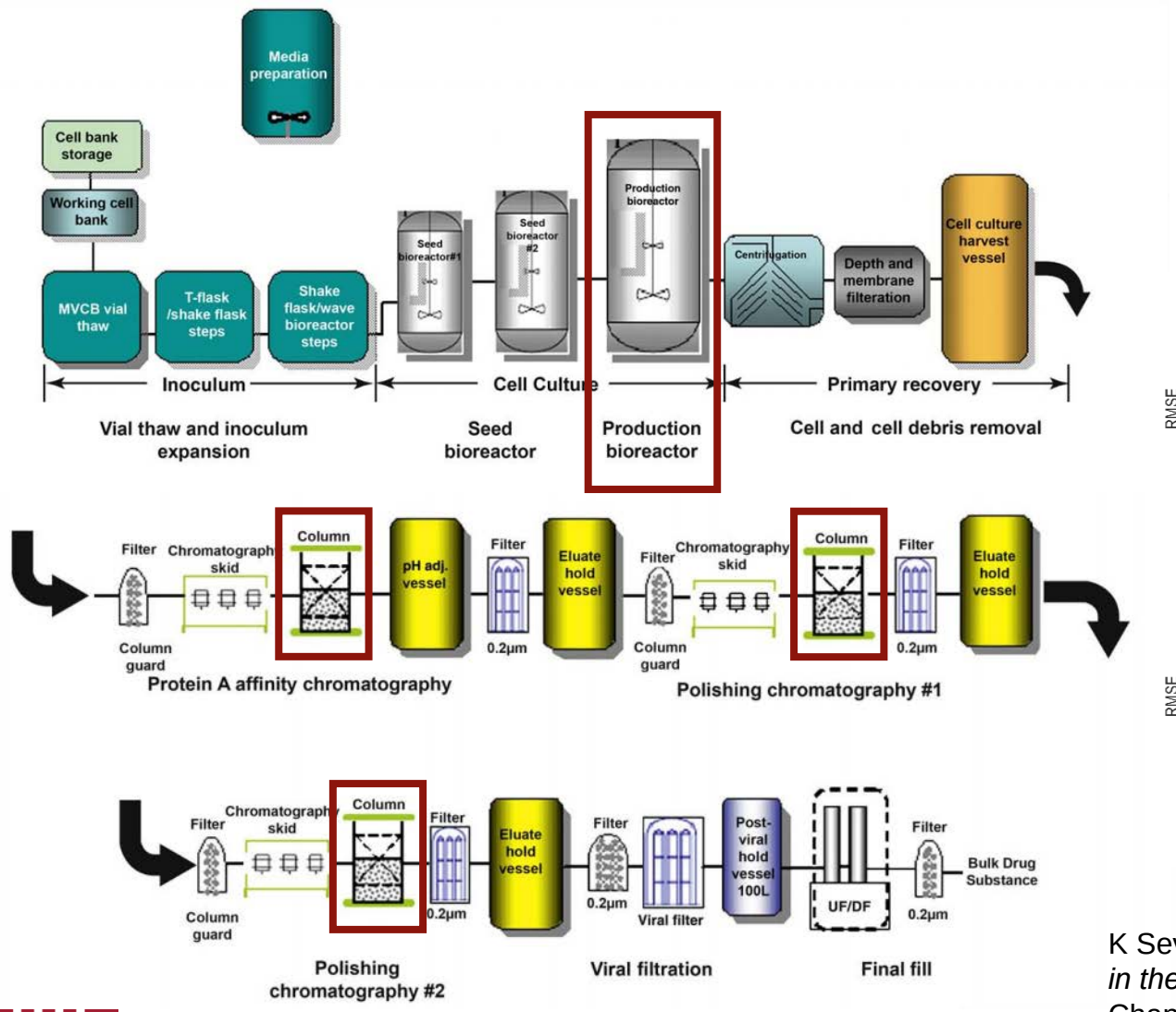
- ARMAX = autoregressive moving average with exogenous input
- CVA = canonical variate analysis
- SVR = support vector regression
- ALVEN = algebraic learning via elastic net
- RF = random forest
- PLS = partial least squares
- RR = ridge regression
- NCVA = nonlinear CVA
- NARMAX = nonlinear ARMAX
- NRCVA = nonlinear regularized CVA



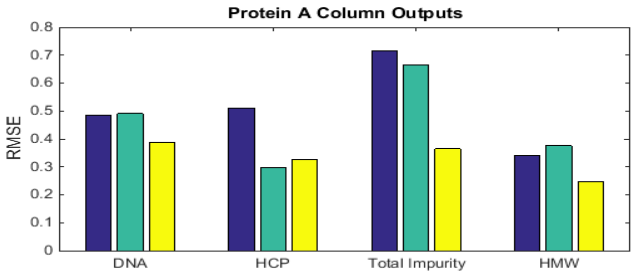
Sun, Braatz, FOPAM, 2019

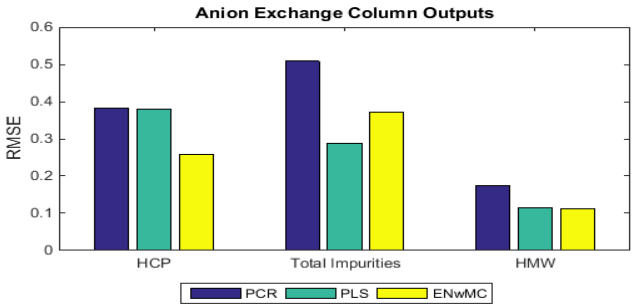
PLS = partial least squares; RR = ridge regression; SVR = support vector regression; RF = random forest;
kSVR = kernel support vector regression; CVA = canonical variate analysis; SSARX = subspace ARX
ALVEN = algebraic learning via elastic net for nonlinear modelling, RNN = recurrent NN

Case study: Biopharmaceutical monoclonal antibody manufacturing CQA prediction at Biogen



Outperformed PLS, PCR, and RSM for a mAb at Biogen

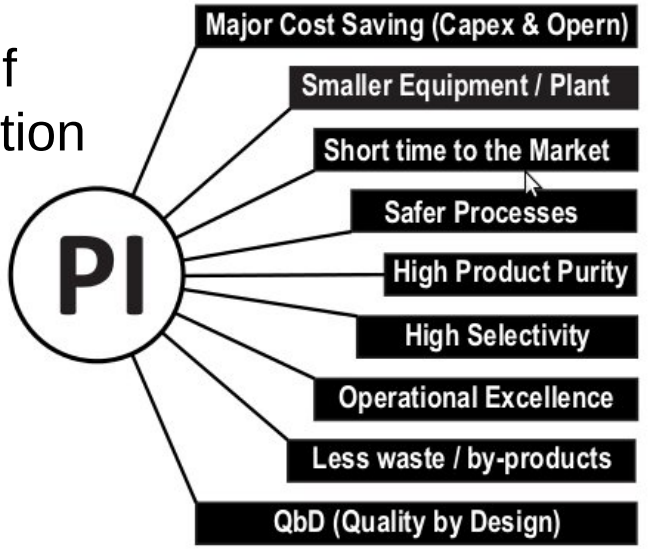




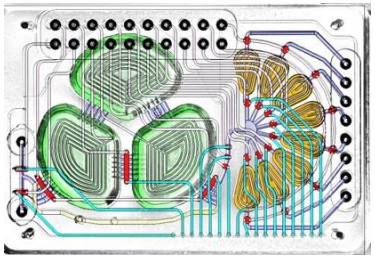
K Severson et al., in *Multivariate Analysis in the Pharmaceutical Industry*, Elsevier, Chapter 12, 2018

Current State of the Art in Process Development

- Greatly increased understanding & optimization of each unit operation, exploiting process intensification

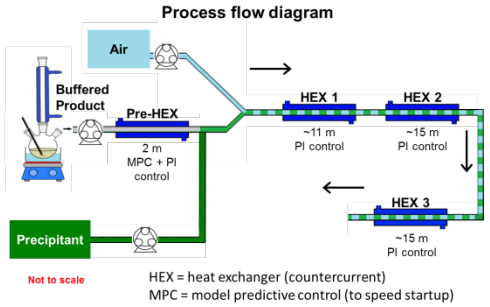


- Automated high-throughput microscale technology for fast process R&D

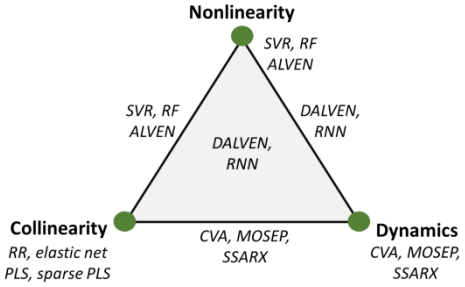


- Plug-and-play modules w/integrated control & monitoring to facilitate deployment

- Dynamic models (w/uncertainty) for unit operations for automated plant-wide simulation & control design



- Autonomous/smart data analytics/machine learning



All strategies based heavily on models

PI image from pi-inc.co; puzzle from web; other images from Braatz & R Ram pubs

State-of-the-art process development strategies support all stages of process validation

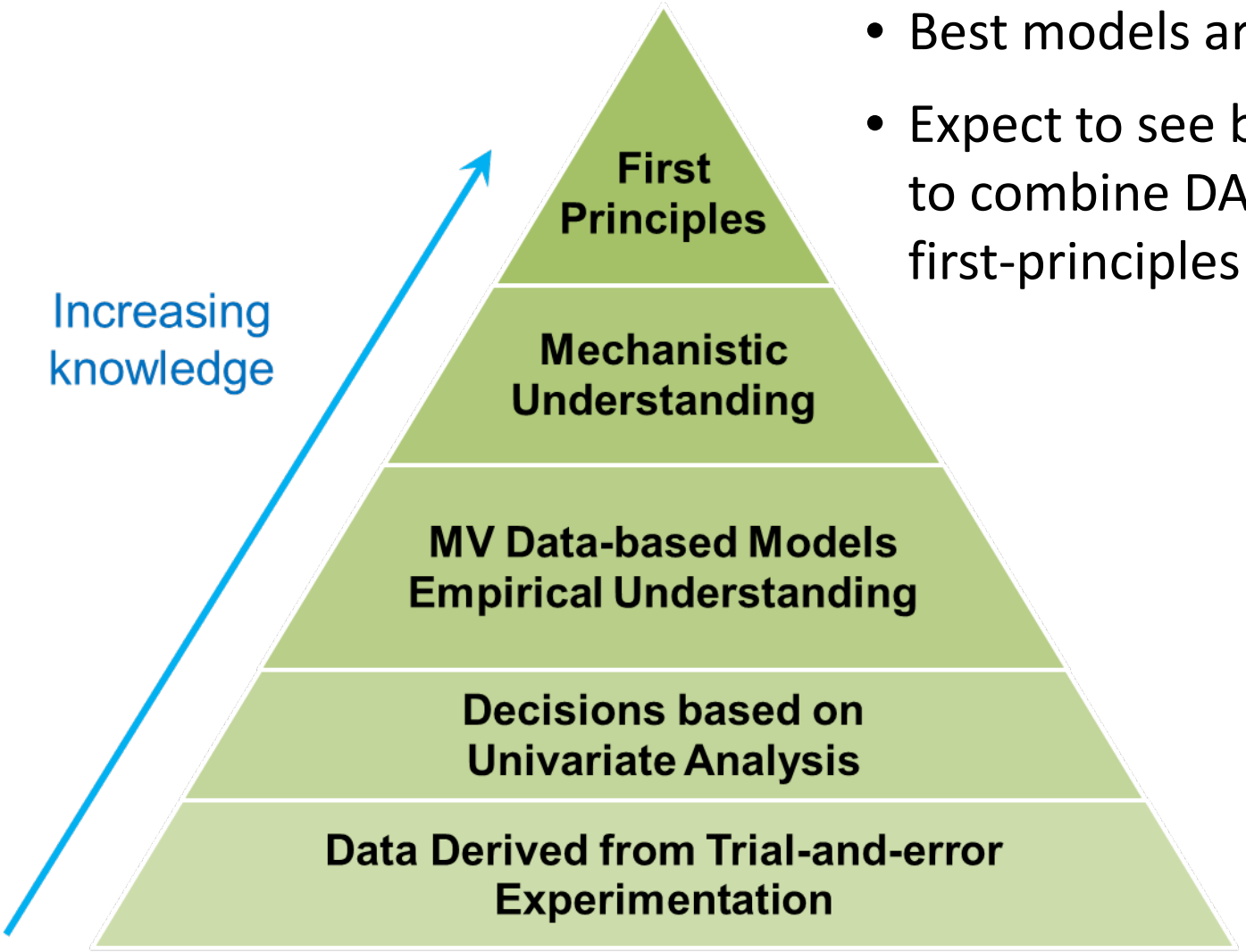
"Process validation is the analysis of data gathered throughout the design and manufacturing of a product in order to confirm that the process can reliably output products of a determined standard"

1. Process design gathers and analyzes data during process development to define the commercial manufacturing process
2. Process qualification confirms quality and output capabilities for all production processes and manufacturing equipment
3. Continued process verification (CPV) is the ongoing collection and analysis of end-to-end production components and processes data to ensure product outputs satisfy quality limits

Outline

- Current state of the art in using models, data analytics, and machine learning
- Looking to the future in process models and data analytics

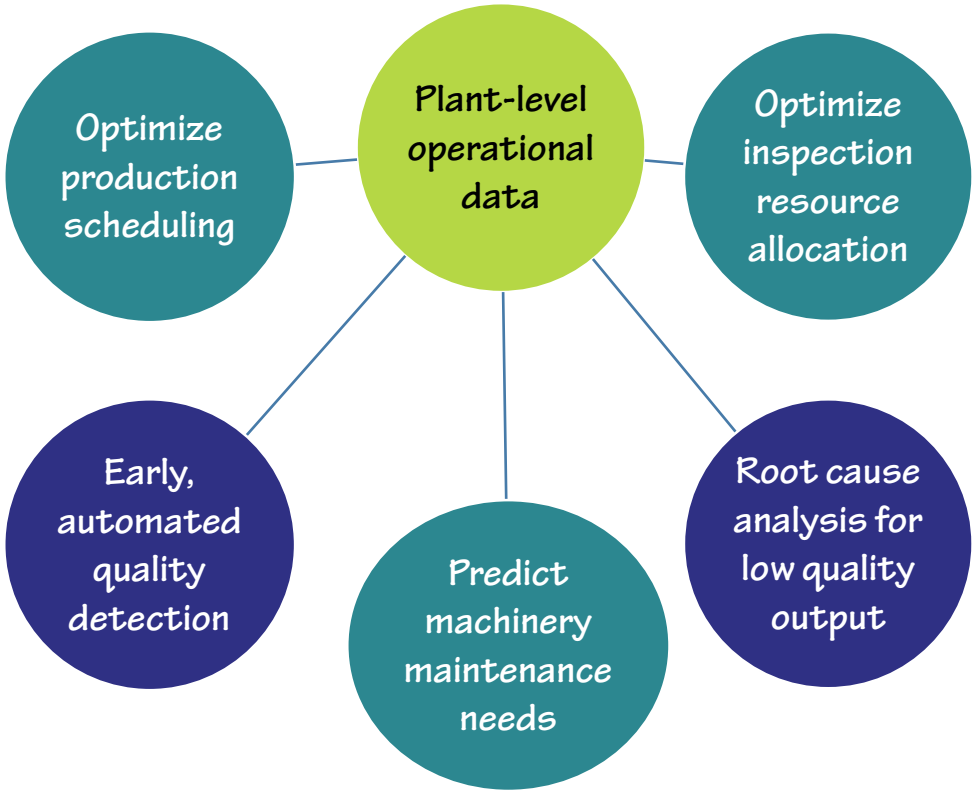
Knowledge pyramid favors 1st principles



- Best models are usually hybrid
- Expect to see better ways to combine DA/ML with first-principles models

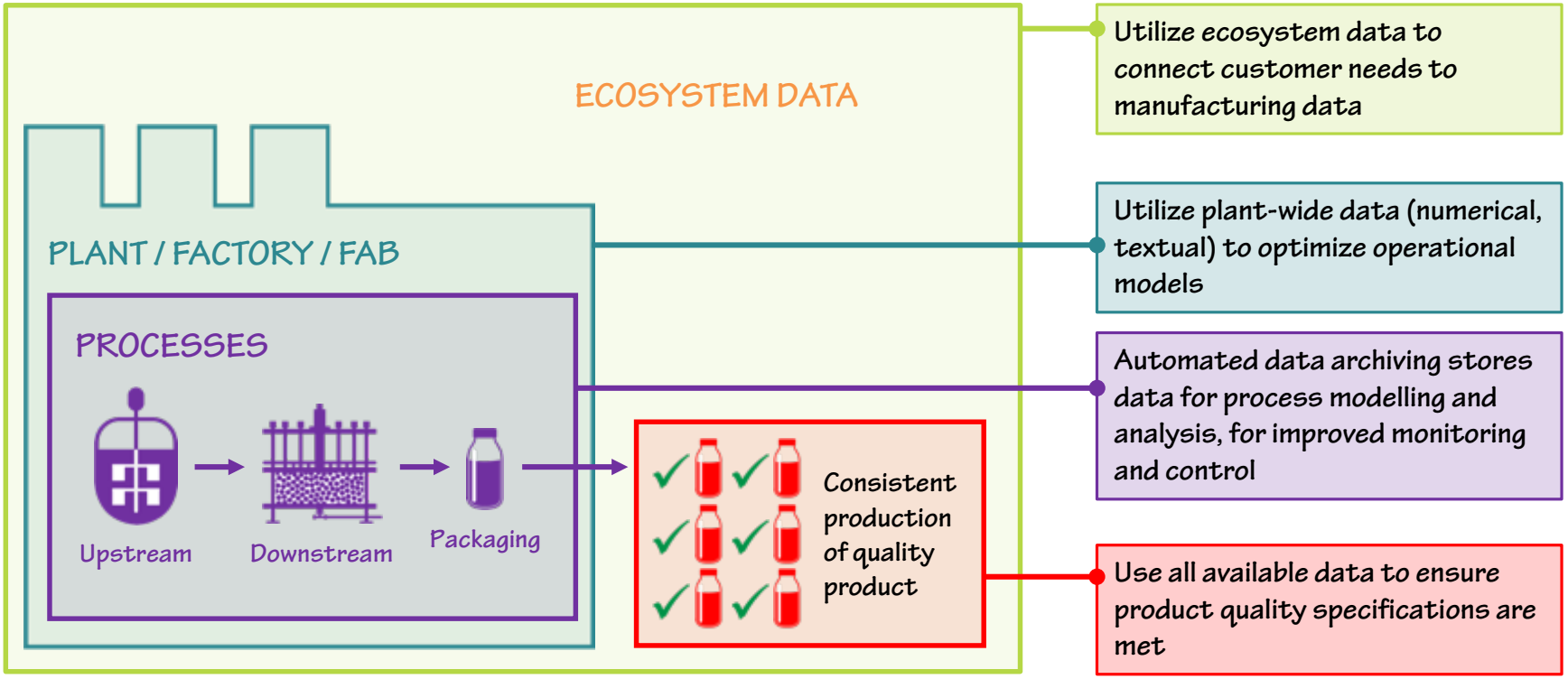
Adapted from Basic Principles of GMP, World Health Organization, May 2008

Broadening the scope of data analytics and machine learning in pharma applications



- All data brought into one database for easy access
- Correlate plant data to off-line product quality specs
- Connect product quality data to the supply chain
- Quickly find causes of off-spec product (e.g. raw materials, operator error)
- Propose design or control changes to reduce operational problems
- Optimize raw material selection
- Design predictive maintenance schedules
- Facilitate continuous improvement

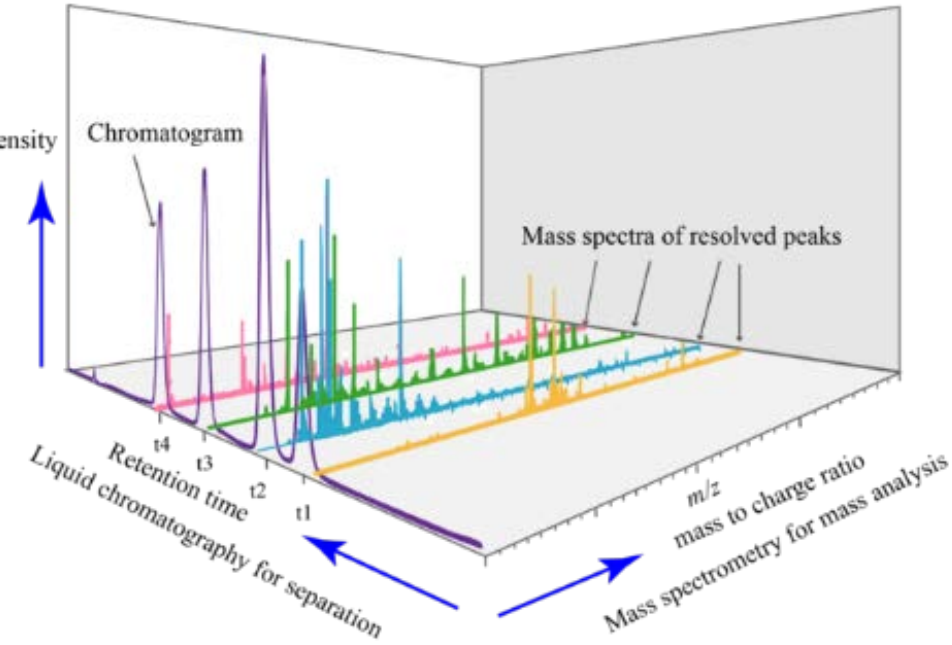
Broadening the scope of data analytics and machine learning in pharma applications



Objectives: Diagnostics/Prognostics, Continuous Improvement, and Optimal Decision Making

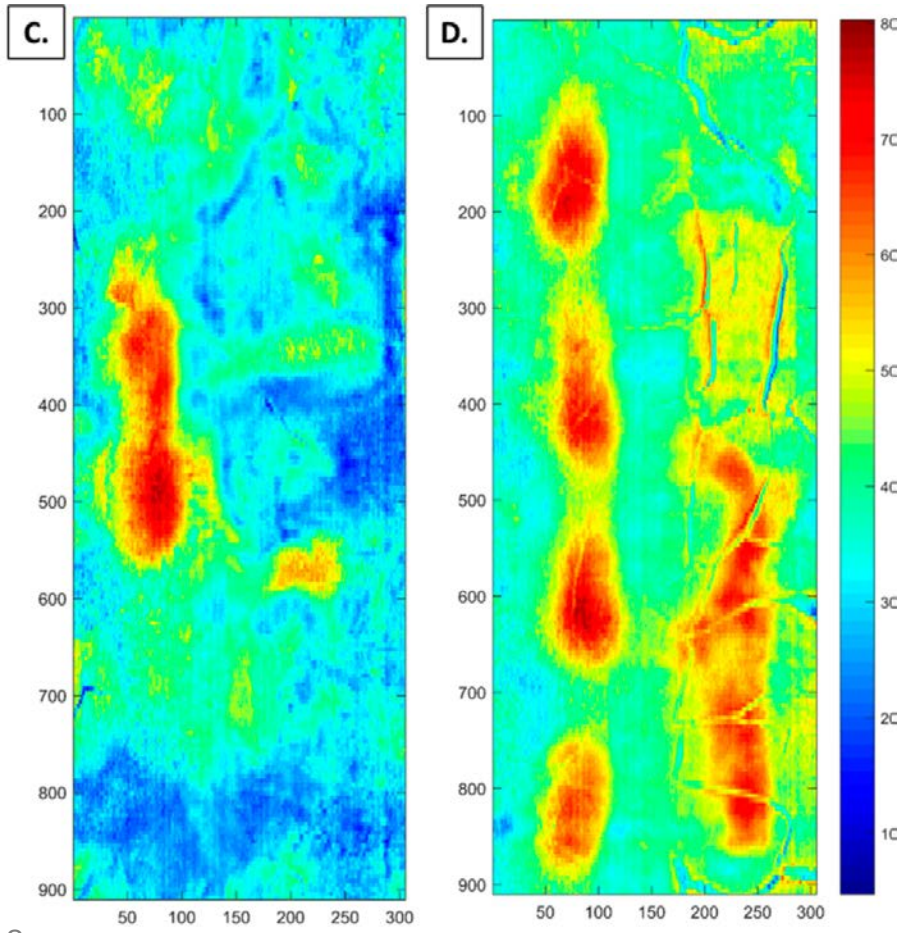
Increased use of tensorial datastreams

LC-MS is becoming more powerful, less expensive, and easier to implement



With autosamplers, allows dynamic measurement from small molecules to protens

b (blue) and g (red) mannitol concentrations in lyophilized samples by NIR chemical imaging



Brouckaert et al., Analytical Chemistry, 90, 4354-4362, 2018

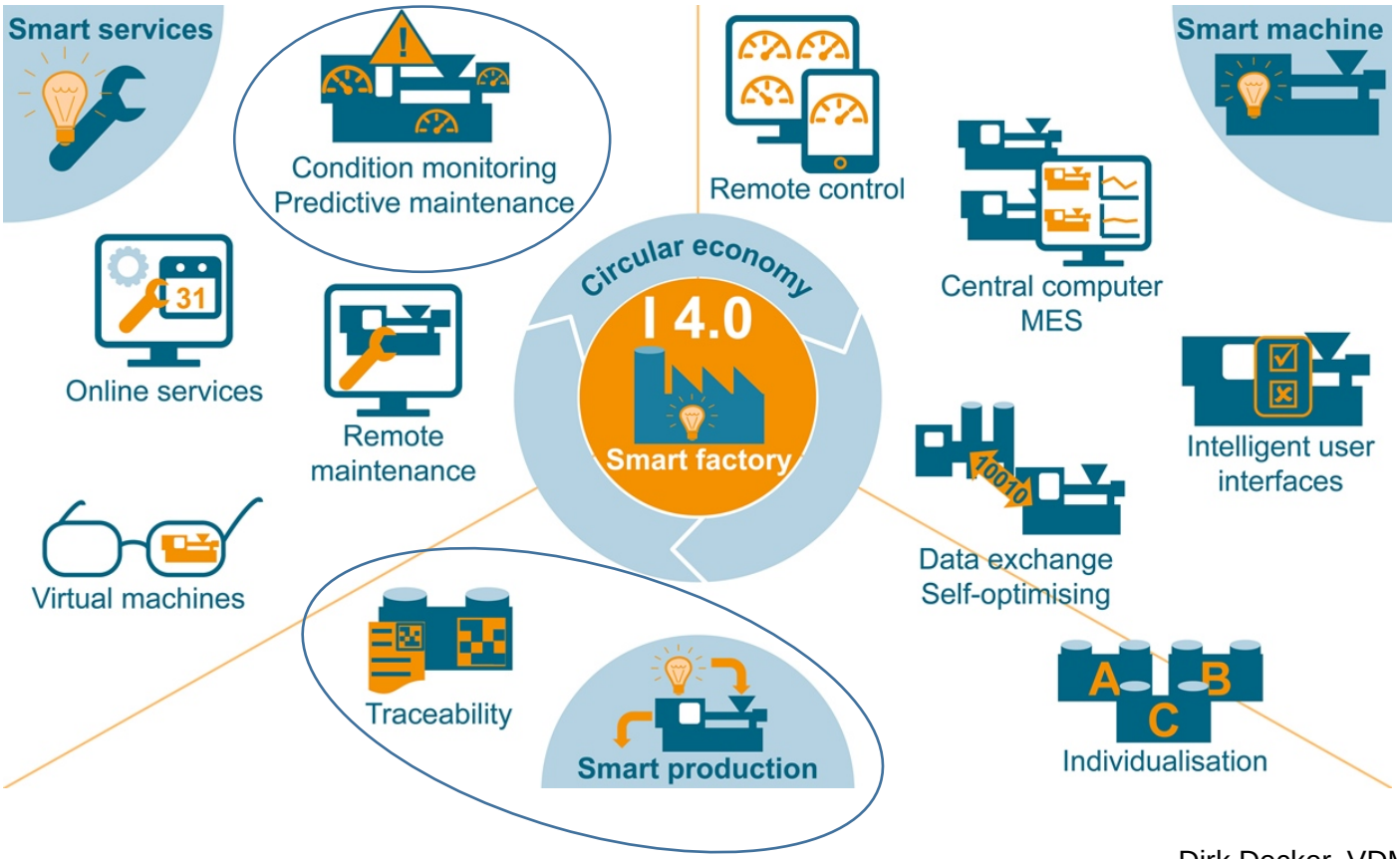
Summary

- Current state of the art in using models, data analytics, and machine learning
  - Key to strategies for accelerating process development
  - Systematic approaches to DA/ML method selection are available
  - Supports all stages of process validation
- Looking to the future in process models and data analytics
  - Expect to see better ways to combine DA/ML with first-principles models
  - Broadening the scope of data analytics and machine learning in pharma applications
  - Increased use of tensorial datastreams

Extra Slides

Industry 4.0, aka Smart Factory

"the current trend of automation and data exchange that includes cyber-physical systems, Internet of Things, cloud computing, and cognitive computing"



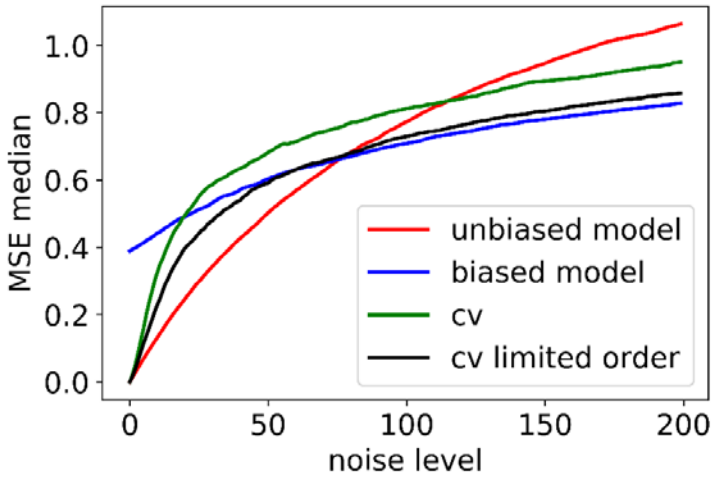
Dirk Decker, VDMA

Selecting the best of many models by using cross-validation overfits data

Consider a process $y = 0 + x + 0.5x^2 + 0.1x^3 + 0.05x^4 + \epsilon$ with 30 samples

The data were fit to four types of models:

1) 2nd order ($y = ax + bx^2 + k$, biased)
2) 4th order ($y = ax + bx^2 + cx^3 + dx^4 + k$, unbiased)
3) polynomial order of 1 to 10 selected by MCCV
4) polynomial order of 2 and 4 selected by MCCV



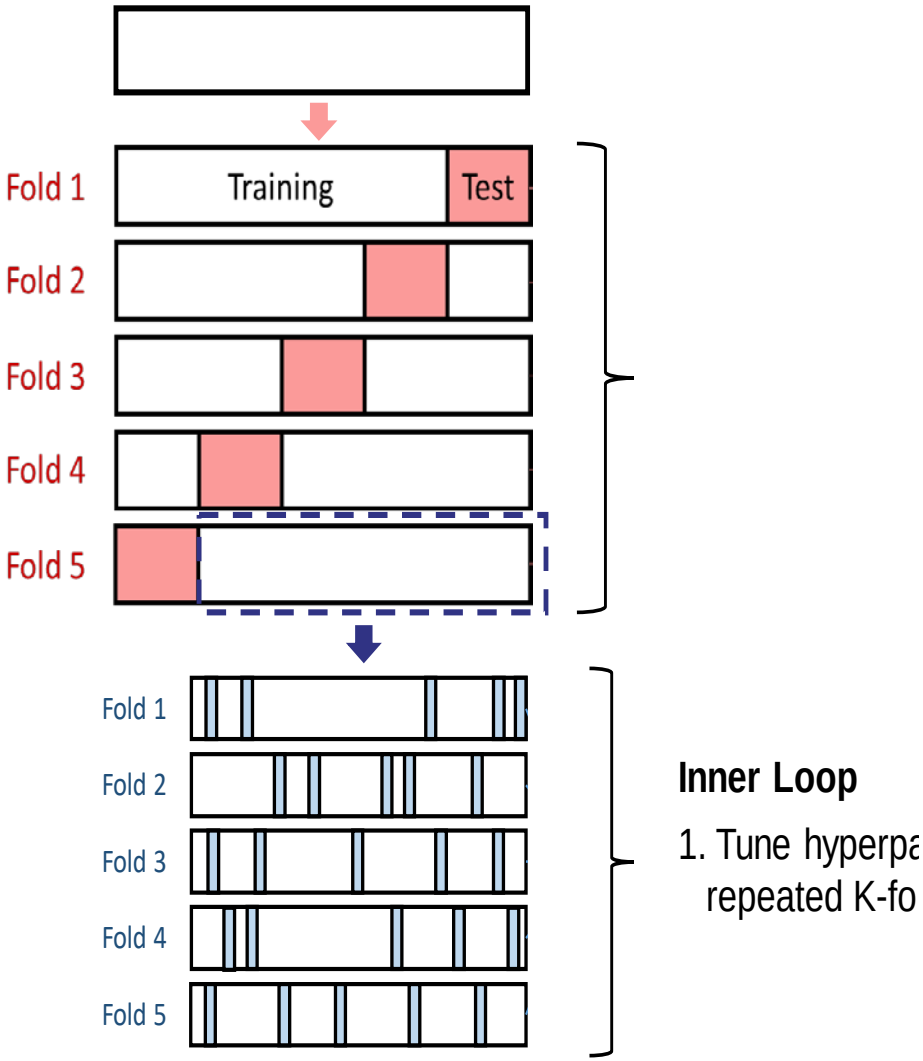
Key takeaways:

- The best model for a specific problem depends on the specific data characteristics
- Pre-select a small number of methods known to be suitable for the type of application

Smart Process Analytics combines best practices with interrogation of data properties

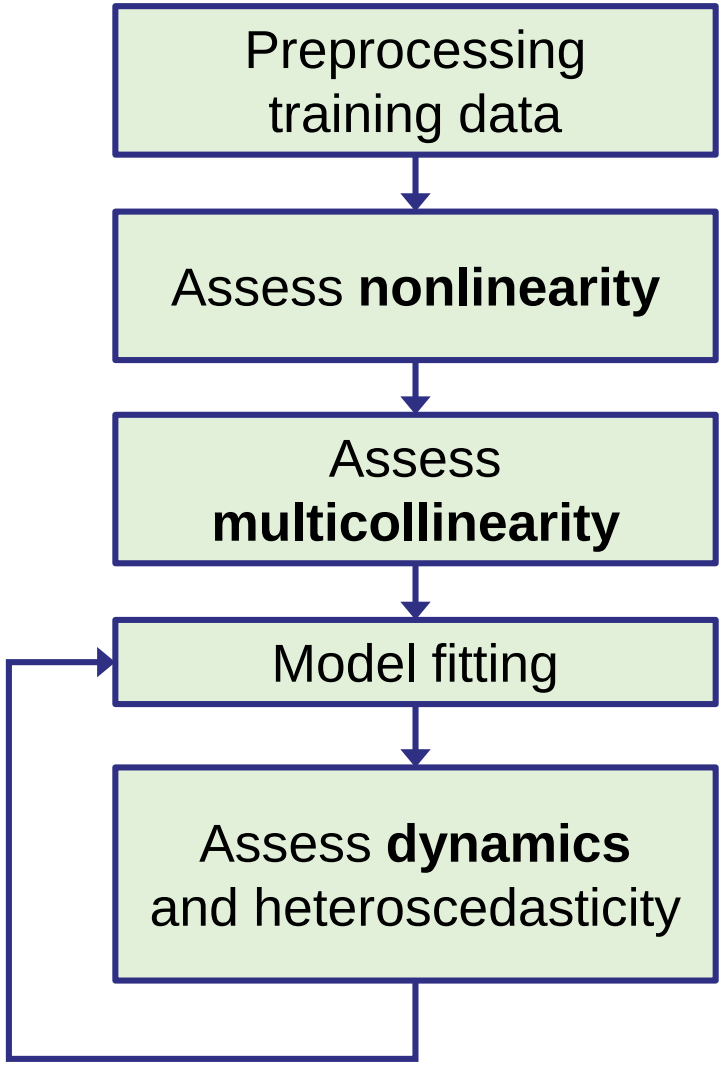
MCCV: Monte Carlo cross-validation

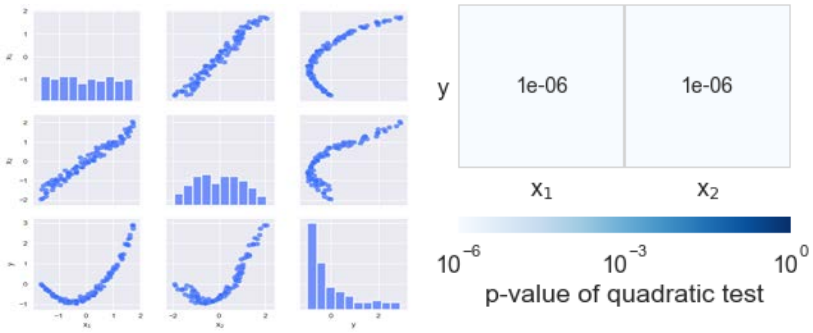
Nested cross-validation for robust method selection and error estimation

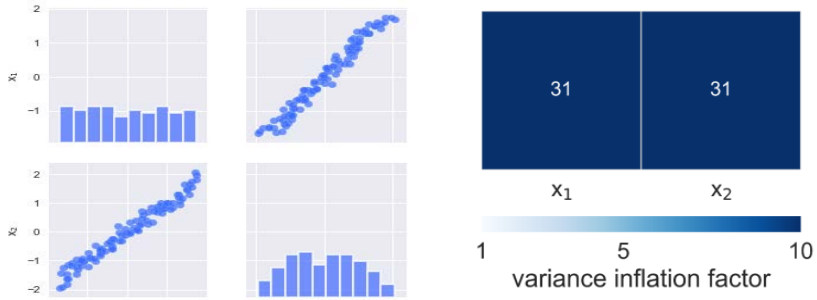


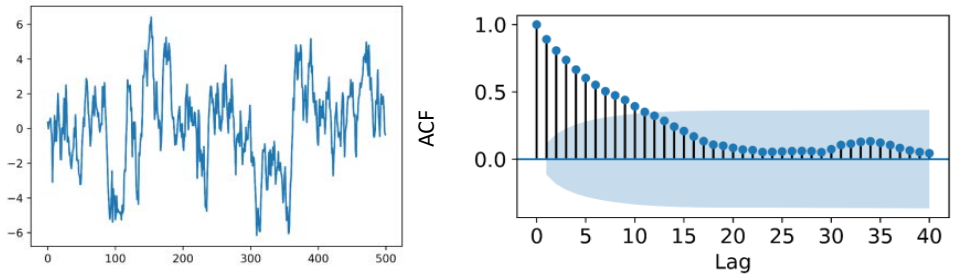
- Monte Carlo cross-validation to minimize overfitting
- k-fold cross-validation (never use leave-one-out)
- Nested cross-validation to estimate model prediction error
- Block Monte Carlo reduces effects of biased data (and can find such data)
- Data sufficiency should be assessed

Alternative 2, Step 1: Interrogate the dataset to ascertain its characteristics

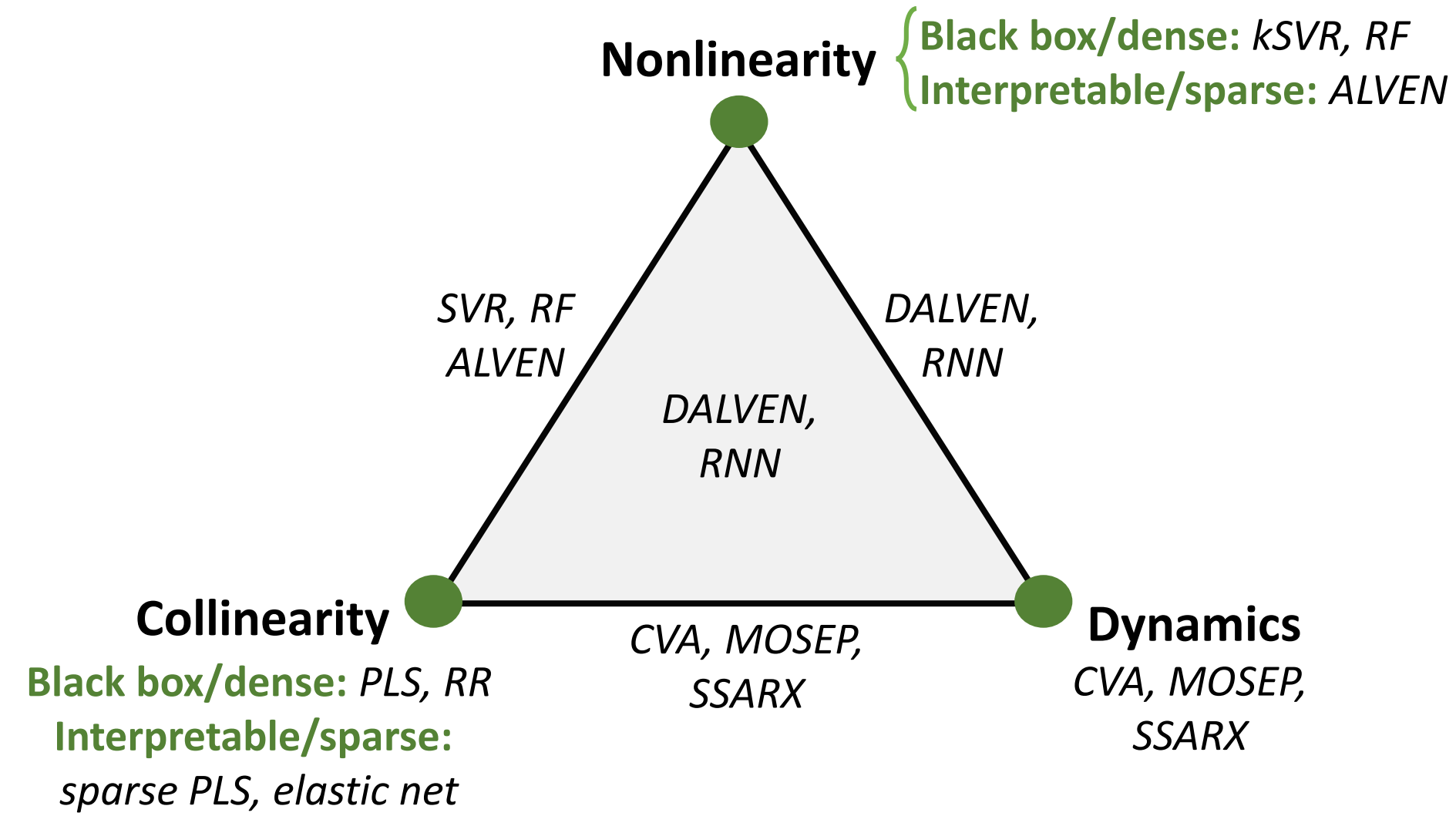






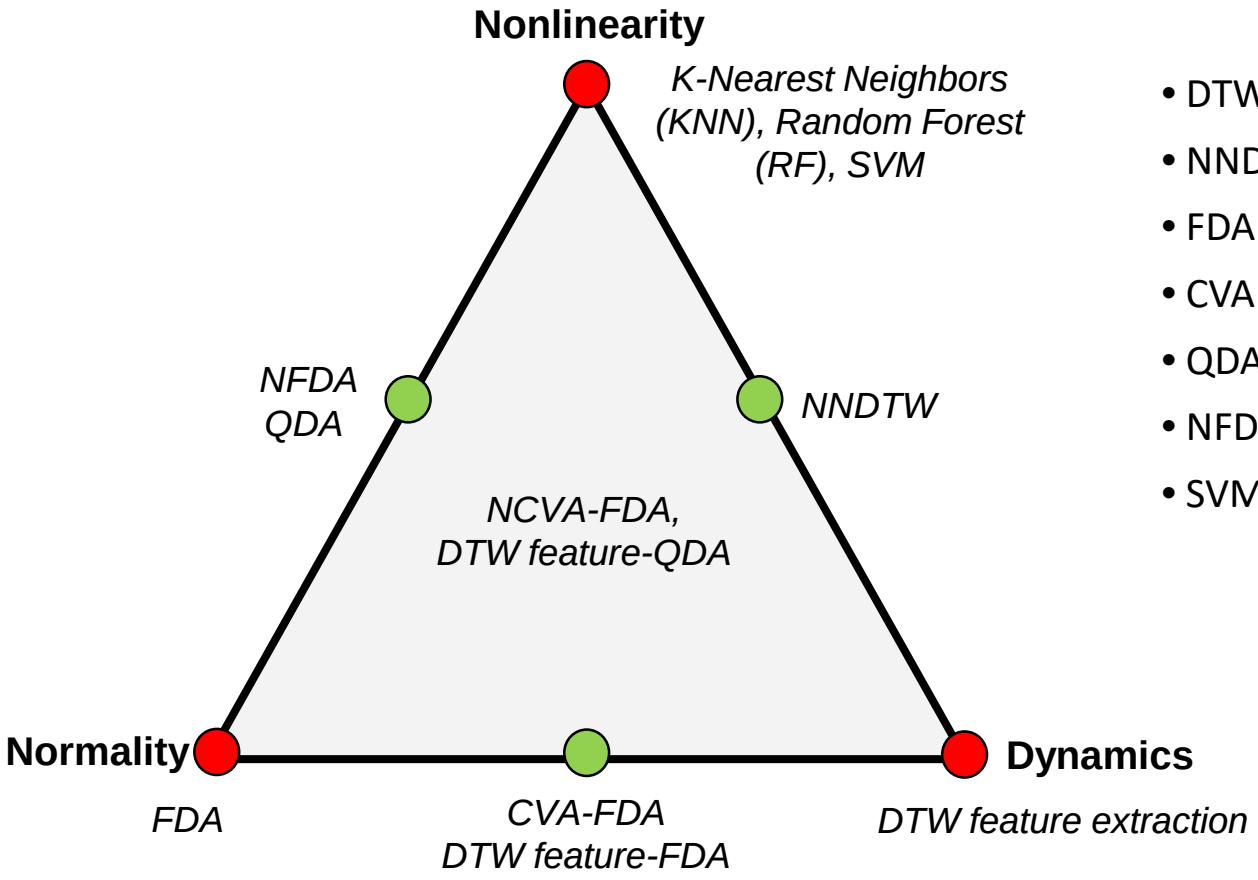


Smart Process Data Analytics (for prediction)



PLS = partial least squares; RR = ridge regression; SVR = support vector regression
SVR = kernel support vector regression; NRCVA = nonlinear regularized canonical variate analysis
ALVEN = algebraic learning via elastic net for nonlinear modelling

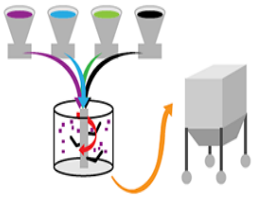
Optimal selection of data analytics methods for supervised classification (v1.0)



- DTW = dynamic time warping
- NNDTW = neural network DTW
- FDA = Fisher discriminant analysis
- CVA = canonical variate analysis
- QDA = quadratic discriminant analysis
- NFDA = nonlinear FDA
- SVM = support vector machine

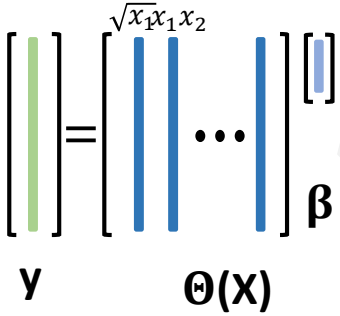
ALVEN: Algebraic Learning via Elastic Net for Nonlinear Modelling

I. Data Collection

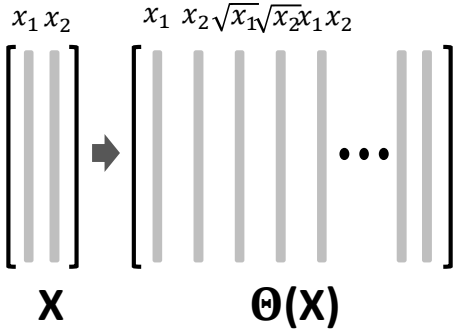


X CPPs **y** CQA

II. Model fitting



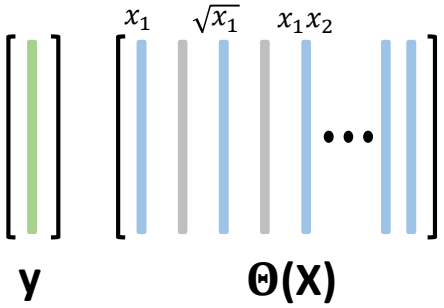
(1) Nonlinear Basis Expansion



Nonlinear transformation family (up to 3rd order)

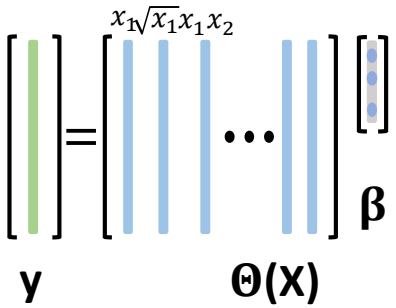
- Basic: $x, \sqrt{x}, \log x, \frac{1}{x}$
- Single feature interaction: $x^{1.5}, \frac{\log x}{x}, \ldots$
- Feature interaction: x_1x_2

(2) Feature Pre-screening



Univariate test $f = \frac{r^2}{1-r^2}(n-2)$

(3) Sparse Regression (EN)



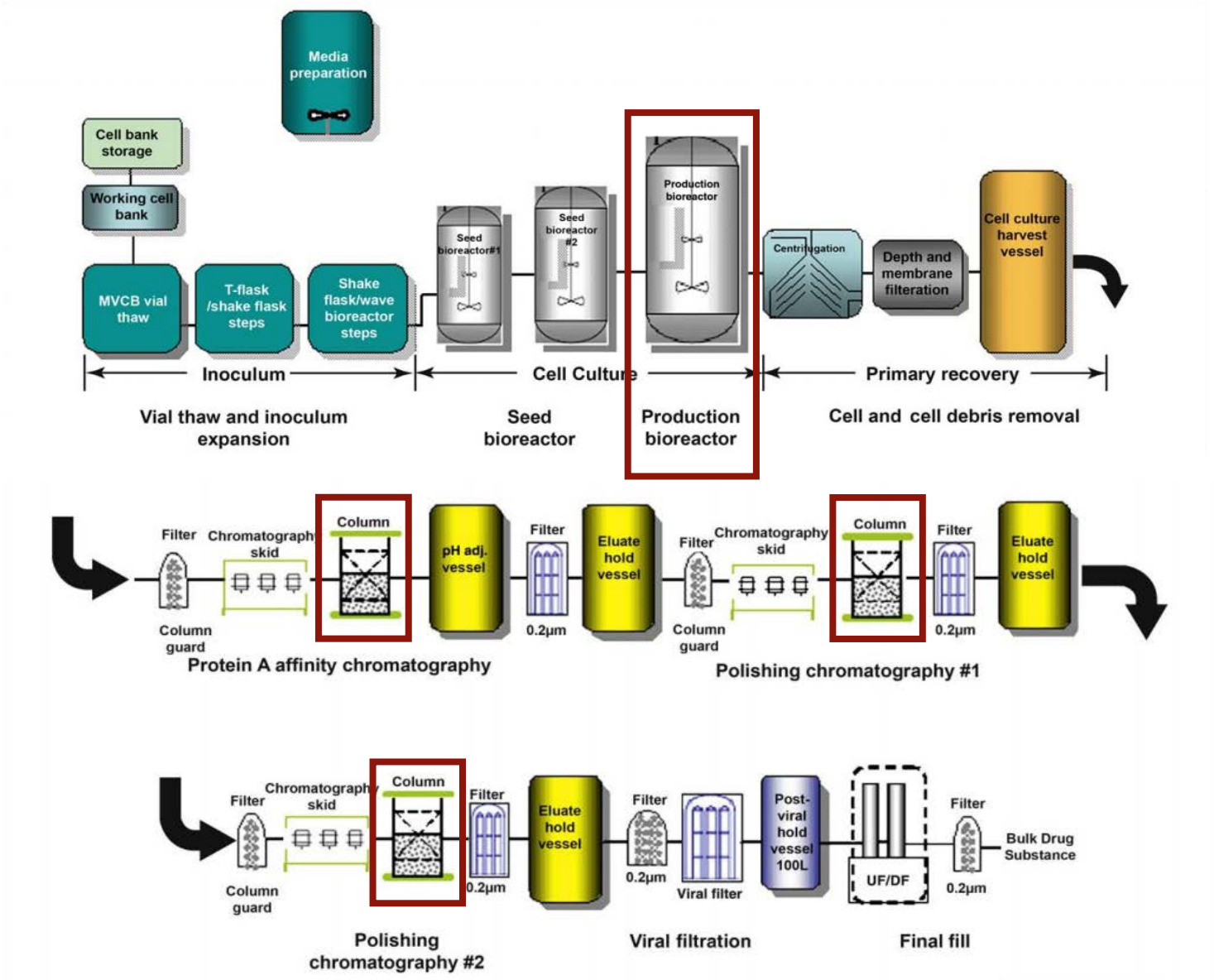
Feature selection with cross validation

ALVEN provides model prediction for a nonlinear process with interpretability. Especially useful when limited data are available.

Case study: Biopharmaceutical mAb manufacturing modeling at Biogen

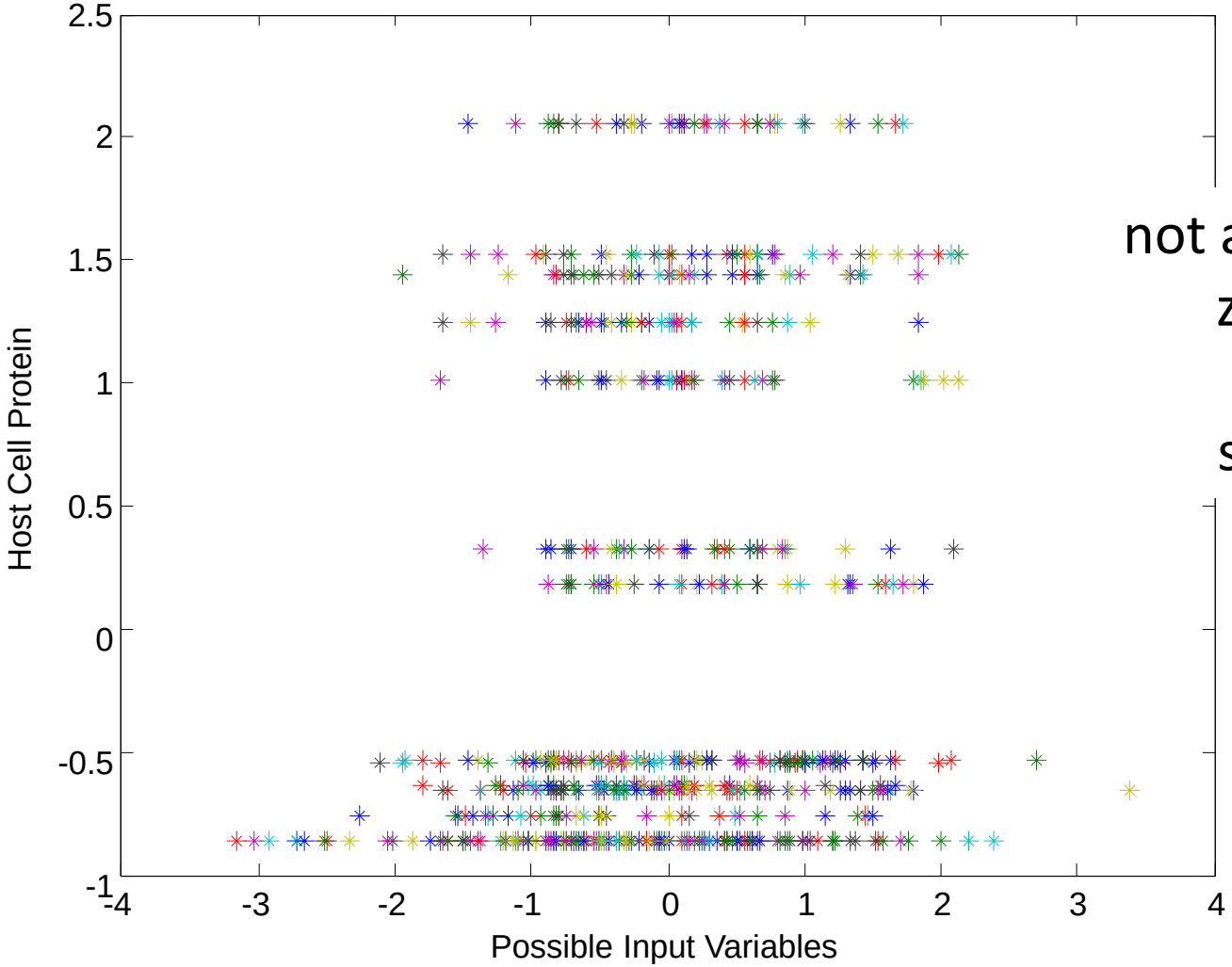
- Application: model critical quality attributes in a monoclonal antibody production process
- Modeling goal: understand the parameters that affect product quality, for use in control
- The approach selects elastic net as the data analytics tool, which outperforms the response surface methodology and PLS methods commonly applied in the biopharma industry

Production-Scale Data for a mAb



Shukla , Thommes. Recent advances in large-scale production of monoclonal antibodies and related proteins, *Trends in Biotechnology*, 2010

Production-Scale Data

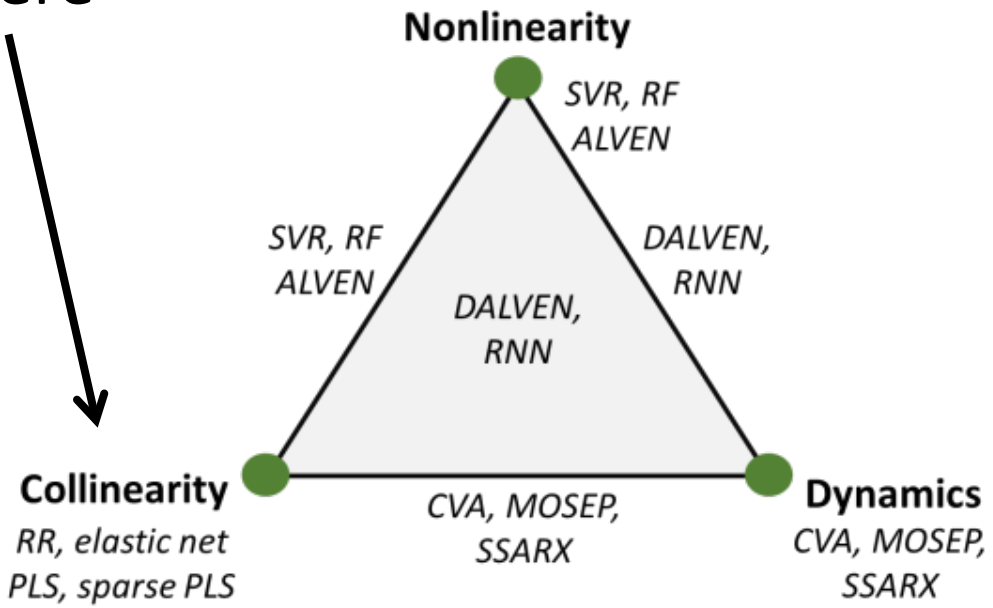


not a factorial design
z-scored data
static data
small dataset

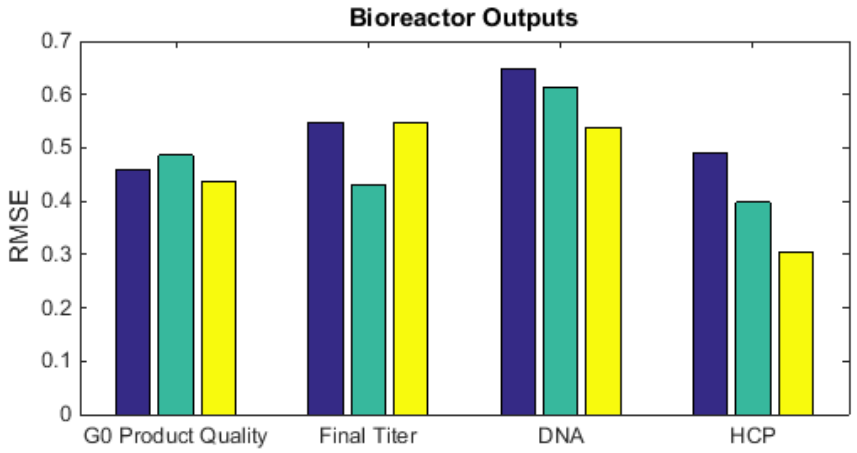
K. Severson, ..., R.D. Braatz, CACE 2015

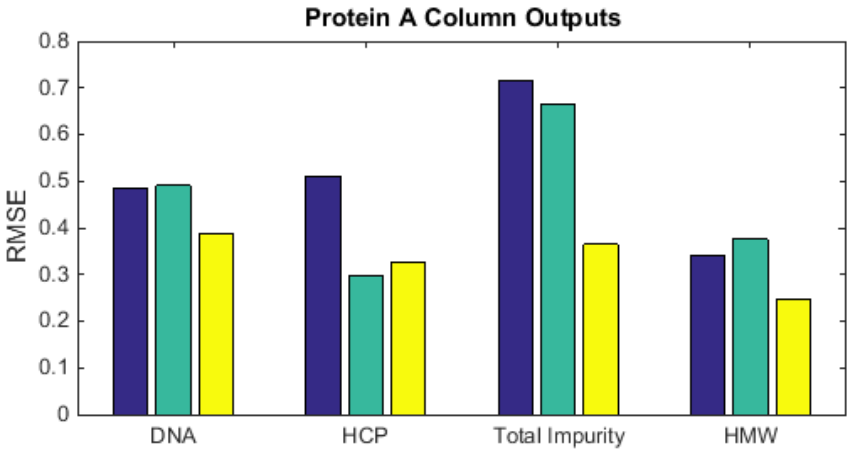
Method Selection

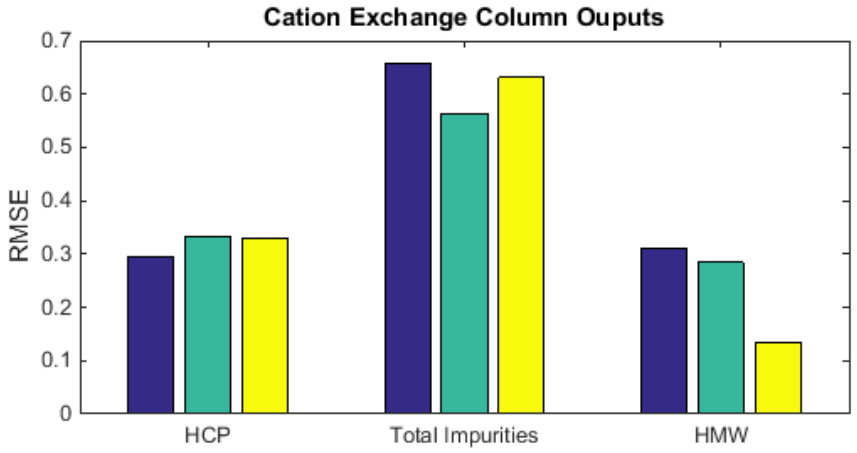
- Goal: Find most accurate model for predicting CQAs
- Data interrogation indicates collinearity so apply a method from here
- Use elastic net (EN) to throw out bad inputs & enable interpretability
- Cross-validation via Monte Carlo sampling

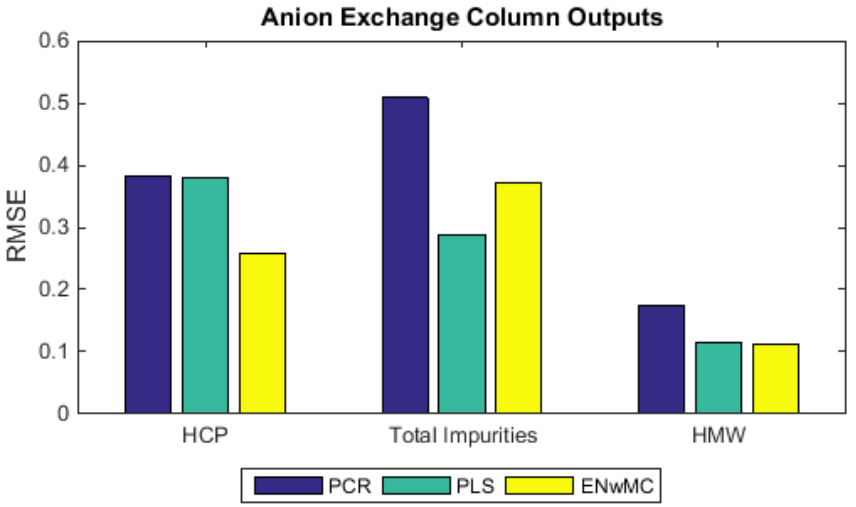


Prediction error using all upstream inputs









K. Severson, ..., R.D. Braatz, CACE 2015

Full Process vs. Modular

- Modular process model restricts the input variables to only the inputs to the unit operation and the output of the previous unit in which data are available

	Full Process Model				Modular Process Model		
CQAs	Model Coeff.	Inputs	RMSE		Model Coeff.	Inputs	RMSE
HCP	3	G0 product quality, antibody conc. entering Protein A, VCD F4	0.26		3	Total impurity exiting CEX, HMW exiting CEX, antibody conc. entering AEX	0.58
Total impurity	4	Total impurity exiting Protein A, N-1 run duration, final % viability, HMW exiting Protein A	0.37		1	Total impurity exiting CEX	0.65
High molecular weight impurities (HMW)	2	Final % viability, HMW exiting protein A	0.11		2	HMW exiting CEX, AEX column loading	0.23

- Most accurate CQA predictions obtained by using bioreactor data as inputs

Summary of Case Study

- Sparse modeling tools such as elastic net can more accurately predict CQAs than response surface methodology and PLS
- Thorough cross-validation should be carried out, to construct models with the highest prediction accuracy
- Maintain traceability in the dataset to capture correlations between upstream operations to downstream product attributes