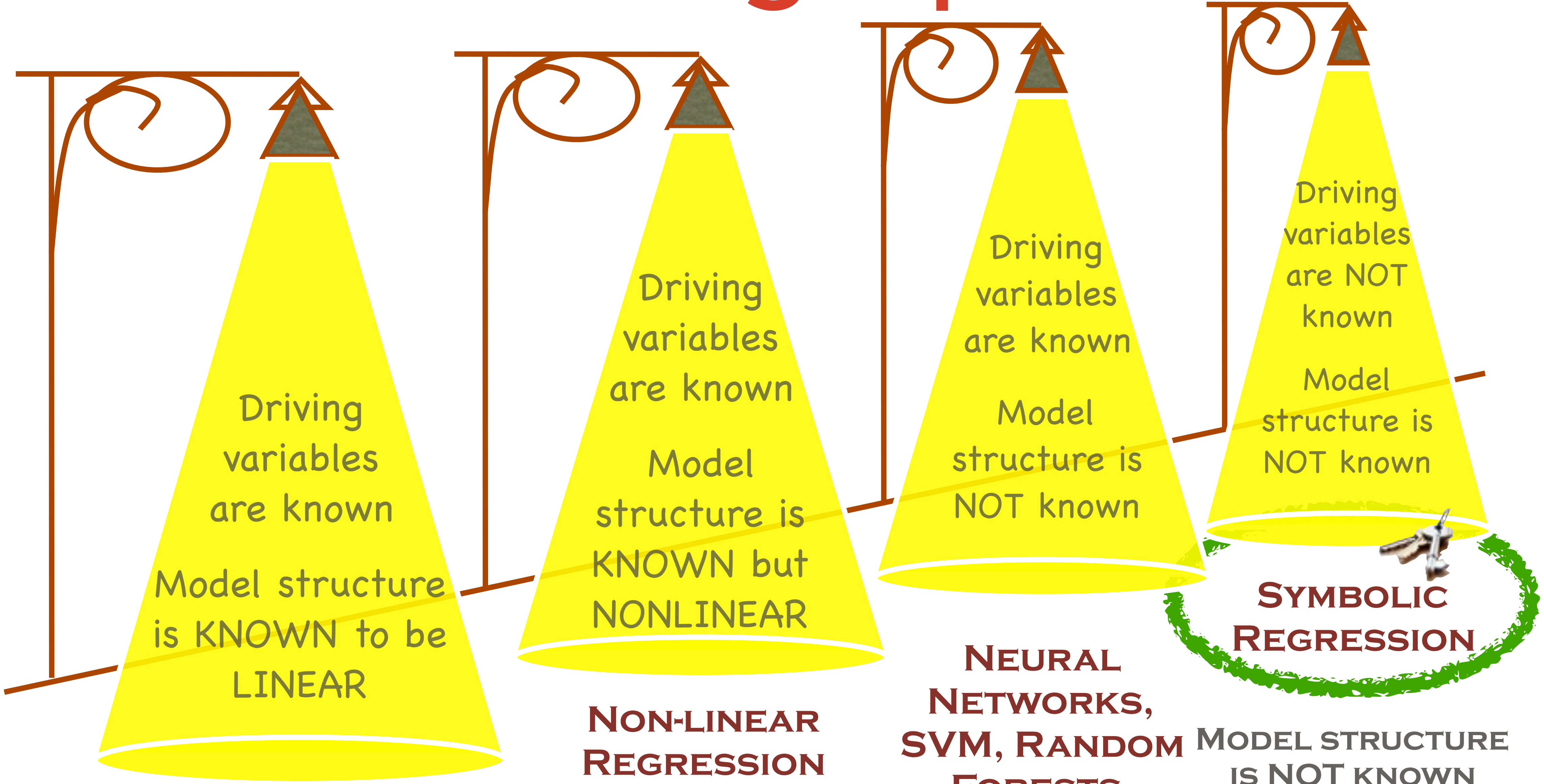
Evolved Analytics LLC

Big Insight vs. Big Data

Mark Kotanchek

Modeling Options



Driving variables are known

Model structure is KNOWN to be LINEAR

Linear Regression

Linear System & Driving Variables are known

Driving variables are known

Model structure is KNOWN but NONLINEAR

Non-linear Regression Parameter Estimation

known nonlinear System & Driving variables are known

Driving variables are known

Model structure is NOT known

Neural Networks, SVM, Random Forests, Symbolic Regression

Unknown Model Structure but Driving variables are known

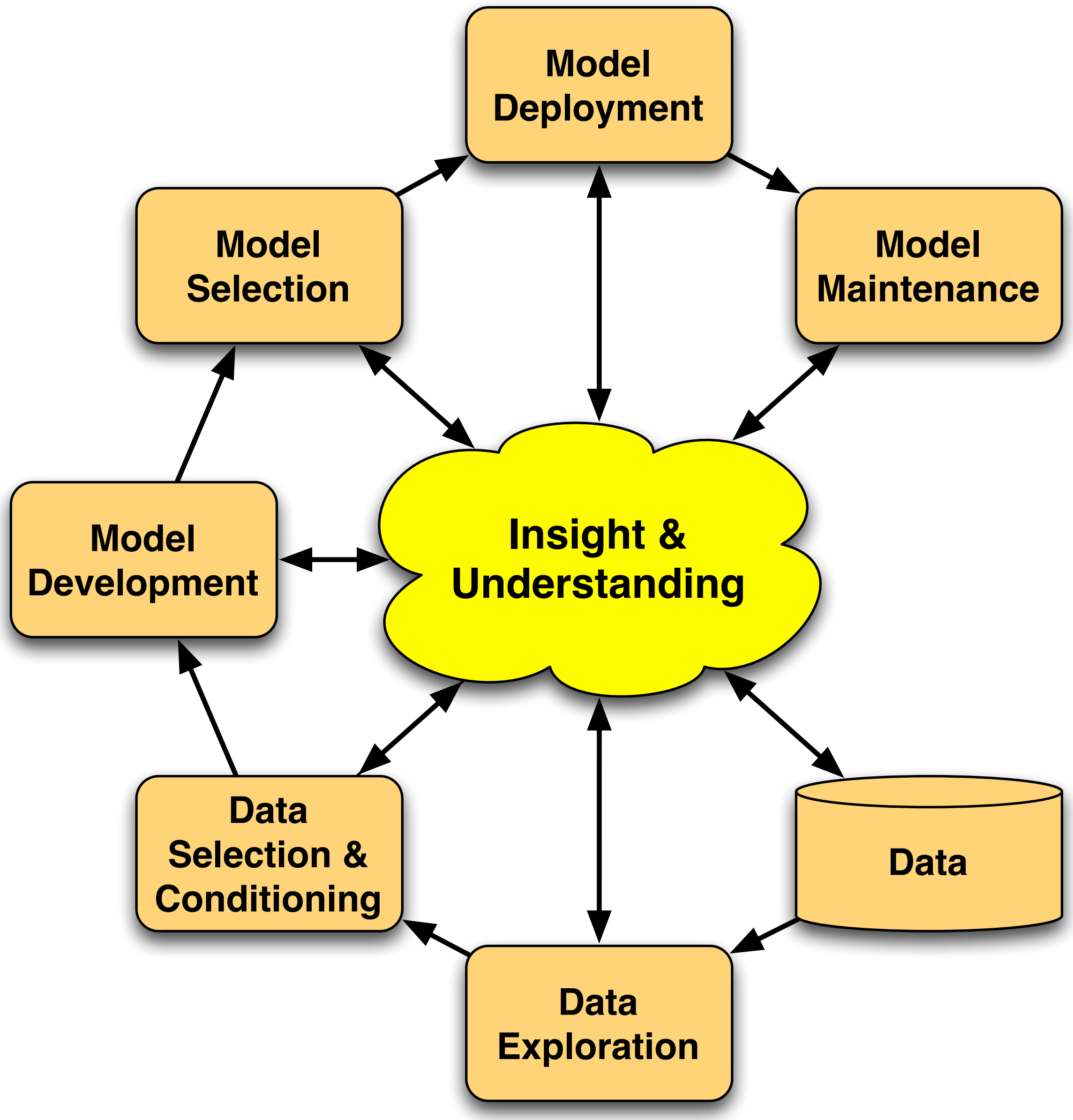
Driving variables are NOT known

Model structure is NOT known

Symbolic Regression

Model structure is NOT known & Driving variables are NOT known

Modeling Workflow is Everything!



- **Awareness & execution** across all aspects
- Analysis flow is **iterative**
- Utilize **visualization** to guide analysis
- An **audit trail** is fundamental

Data in the Real World

Missing Elements

Wide

Lots of Records

Too Little Data

Correlated Variables

Wrong Data

Noisy

Unreliable Sensors

Key Point

Symbolic Regression ⇒ Hypothesis Generator

The only constraint is the supplied building blocks.

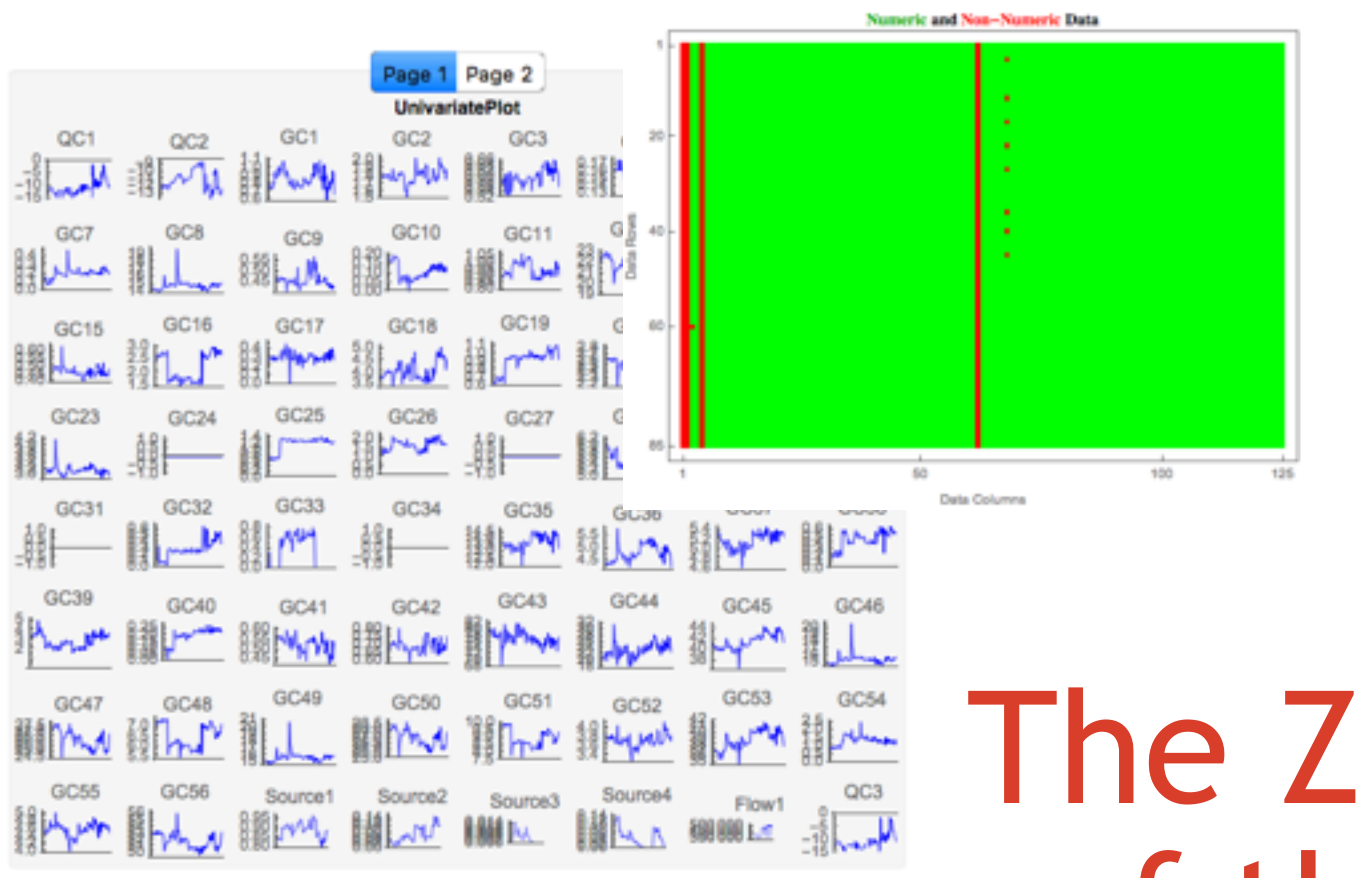
Human limits of imagination & possibility are not imposed!

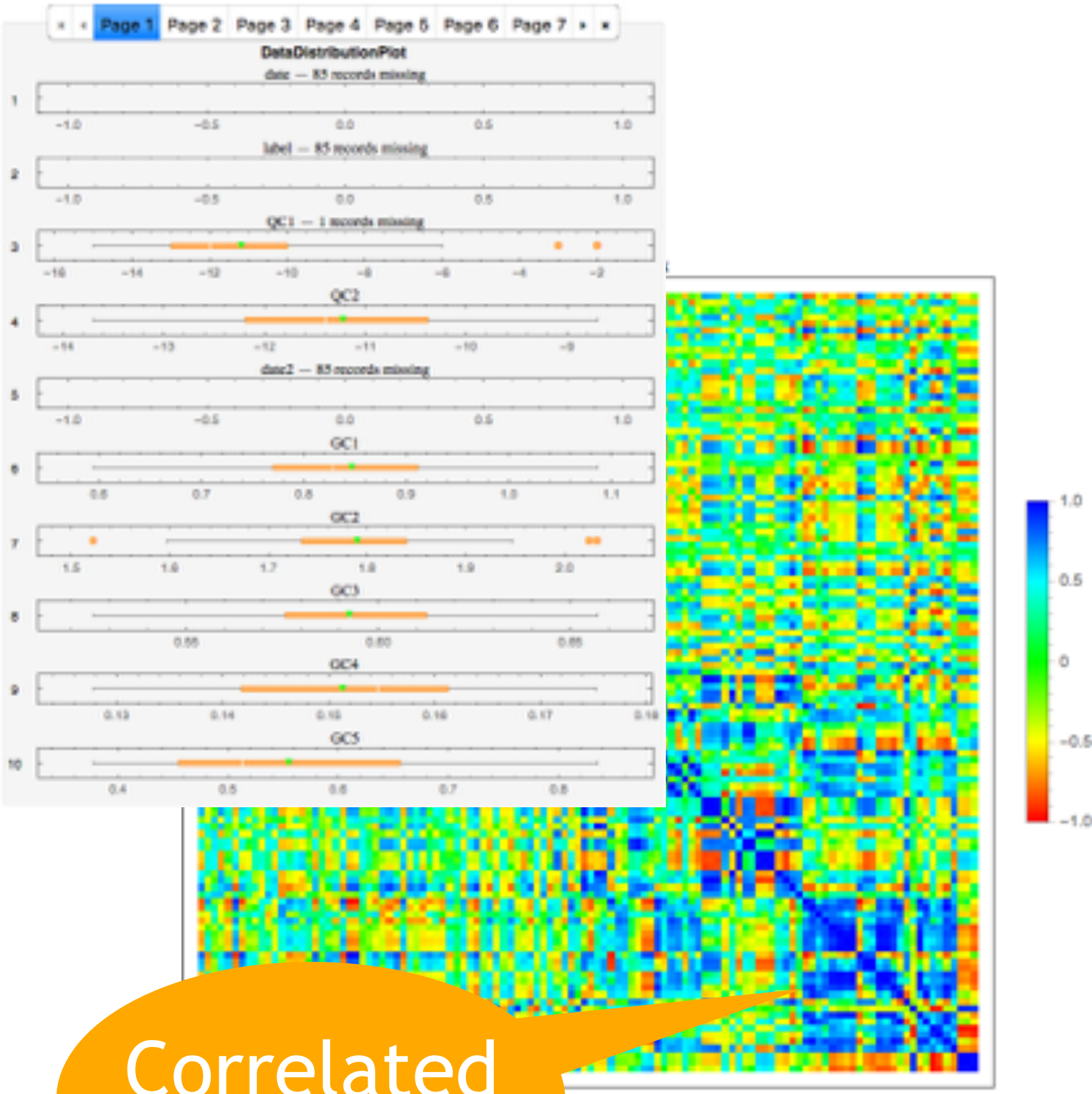
We can exploit this creativity to produce trustable data models.

The Illustration Data

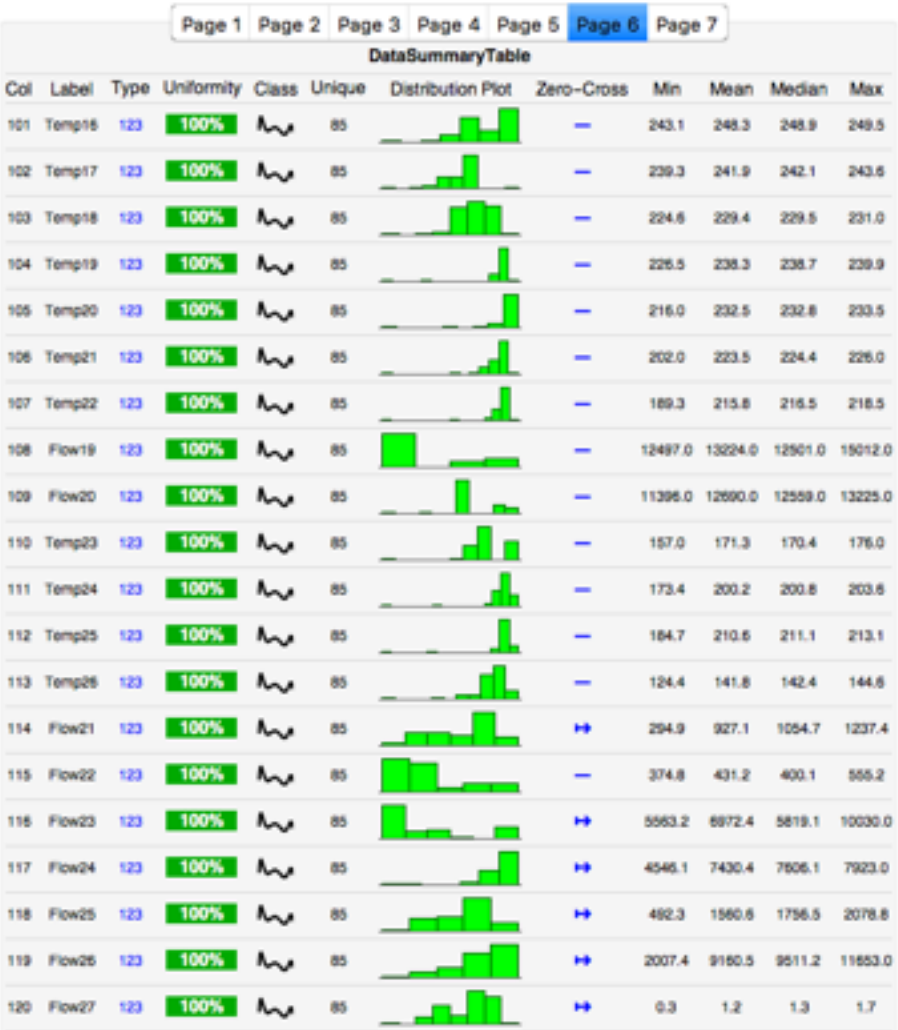
- Data from an industrial chemical reactor
- Having problems with product quality (QC)
- 125 variables sampled over three months
- Chemical composition from gas chromatography (GC)
- Process information from plant (flows, temps & source)
- Two users:
  - Analytical Chemists (Why??)
  - Production Engineers (What to do??)

The Zen of the data



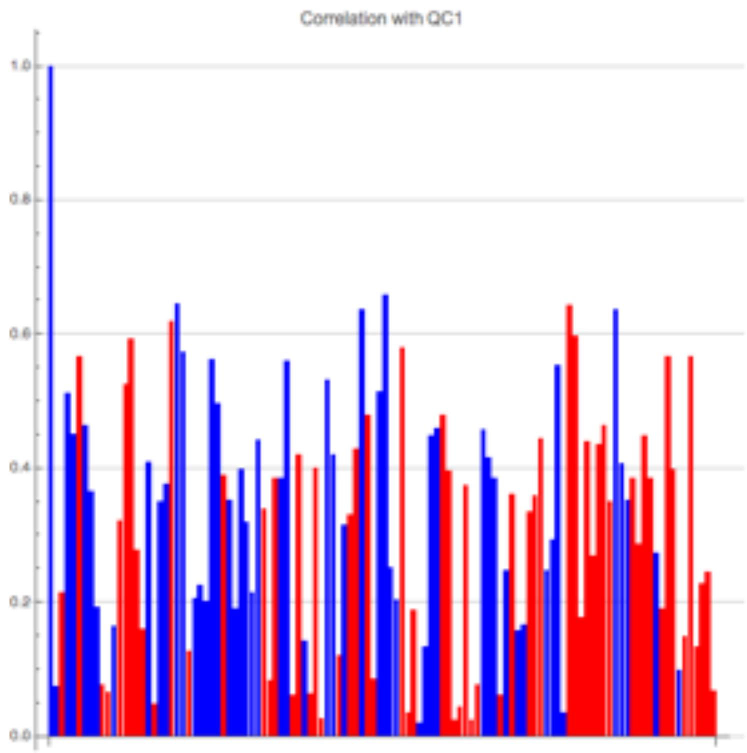


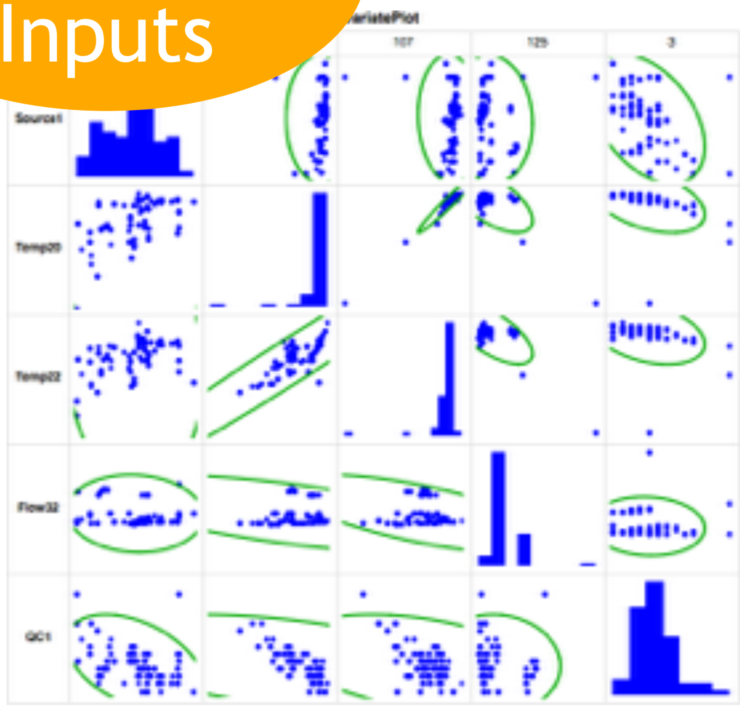
Correlated Inputs

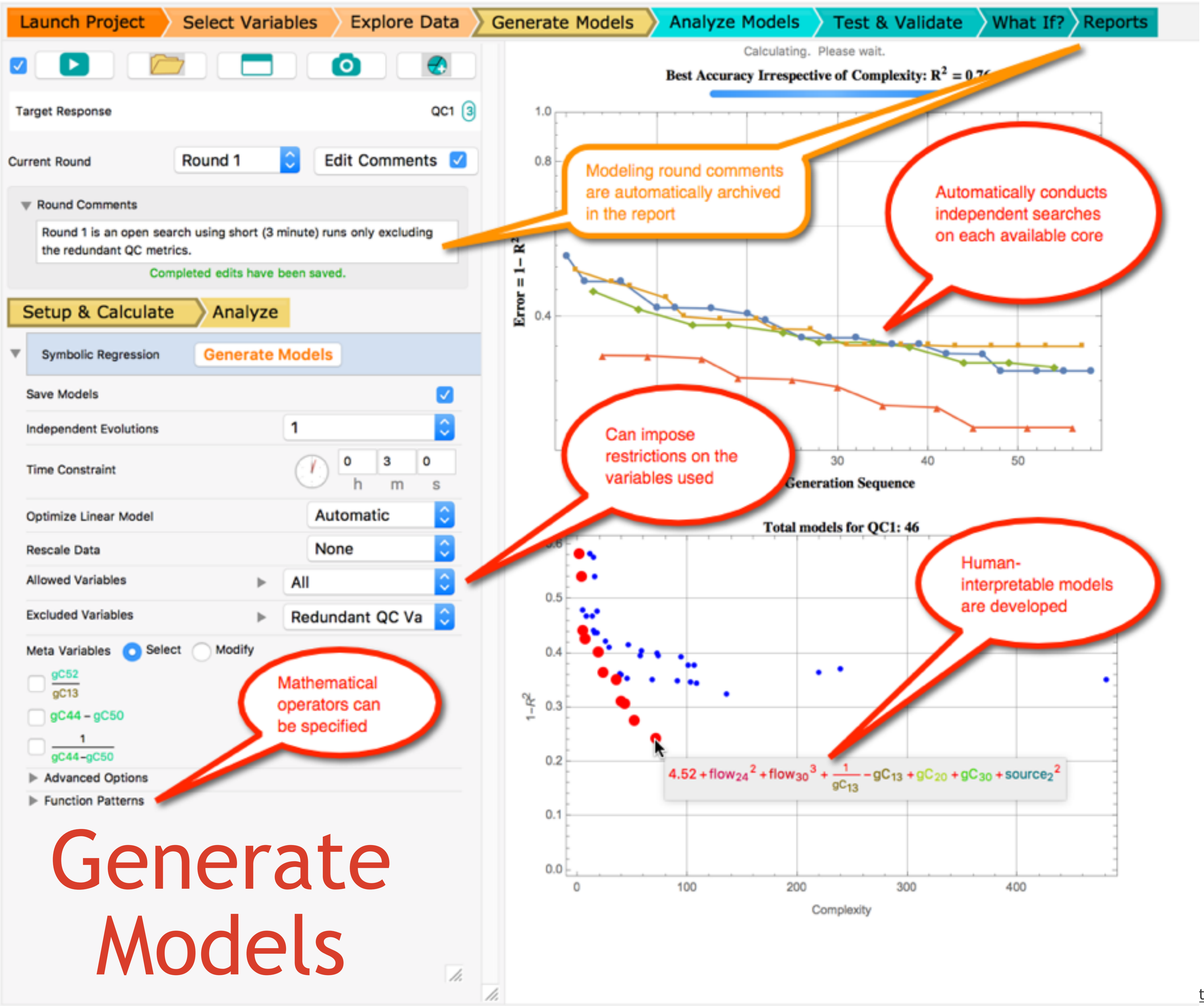


DataSummaryTable

Col	Label	Type	Uniformity	Class	Unique	Distribution Plot	Zero-Cross	Min	Mean	Median	Max
101	Temp16	123	100%		85		—	243.1	248.3	248.9	249.5
102	Temp17	123	100%		85		—	239.3	241.9	242.1	243.6
103	Temp18	123	100%		85		—	224.6	229.4	229.5	231.0
104	Temp19	123	100%		85		—	226.5	238.3	238.7	239.9
105	Temp20	123	100%		85		—	216.0	232.5	232.8	233.5
106	Temp21	123	100%		85		—	202.0	223.5	224.4	226.0
107	Temp22	123	100%		85		—	189.3	215.8	216.5	218.5
108	Flow19	123	100%		85		—	12497.0	13224.0	12501.0	15012.0
109	Flow20	123	100%		85		—	11396.0	12690.0	12559.0	13225.0
110	Temp23	123	100%		85		—	157.0	171.3	170.4	176.0
111	Temp24	123	100%		85		—	173.4	200.2	200.8	203.6
112	Temp25	123	100%		85		—	184.7	210.6	211.1	213.1
113	Temp26	123	100%		85		—	124.4	141.8	142.4	144.6
114	Flow21	123	100%		85		↦	294.9	927.1	1054.7	1237.4
115	Flow22	123	100%		85		—	374.8	431.2	400.1	555.2
116	Flow23	123	100%		85		↦	5563.2	6972.4	5819.1	10030.0
117	Flow24	123	100%		85		↦	4546.1	7430.4	7606.1	7923.0
118	Flow25	123	100%		85		↦	492.3	1560.6	1756.5	2078.8
119	Flow26	123	100%		85		↦	2007.4	9160.5	9511.2	11653.0
120	Flow27	123	100%		85		↦	0.3	1.2	1.3	1.7







☑

Target Response QC1 3

Current Round: Round 1

Edit Comments ☑

Round Comments

Round 1 is an open search using short (3 minute) runs only excluding the redundant QC metrics.

Completed edits have been saved.

Setup & Calculate | Analyze

Symbolic Regression | Generate Models

Save Models ☑

Independent Evolutions: 1

Time Constraint: 0 h 3 m 0 s

Optimize Linear Model: Automatic

Rescale Data: None

Allowed Variables: All

Excluded Variables: Redundant QC Va

Meta Variables: ● Select ○ Modify

□ $\frac{gC52}{gC13}$

□ gC44 − gC50

□ $\frac{1}{gC44-gC50}$

Advanced Options

Function Patterns

$4.52 + flow_{24}^2 + flow_{30}^3 + \frac{1}{gC_{13}} - gC_{13} + gC_{20} + gC_{30} + source_2^2$

Modeling round comments are automatically archived in the report

Automatically conducts independent searches on each available core

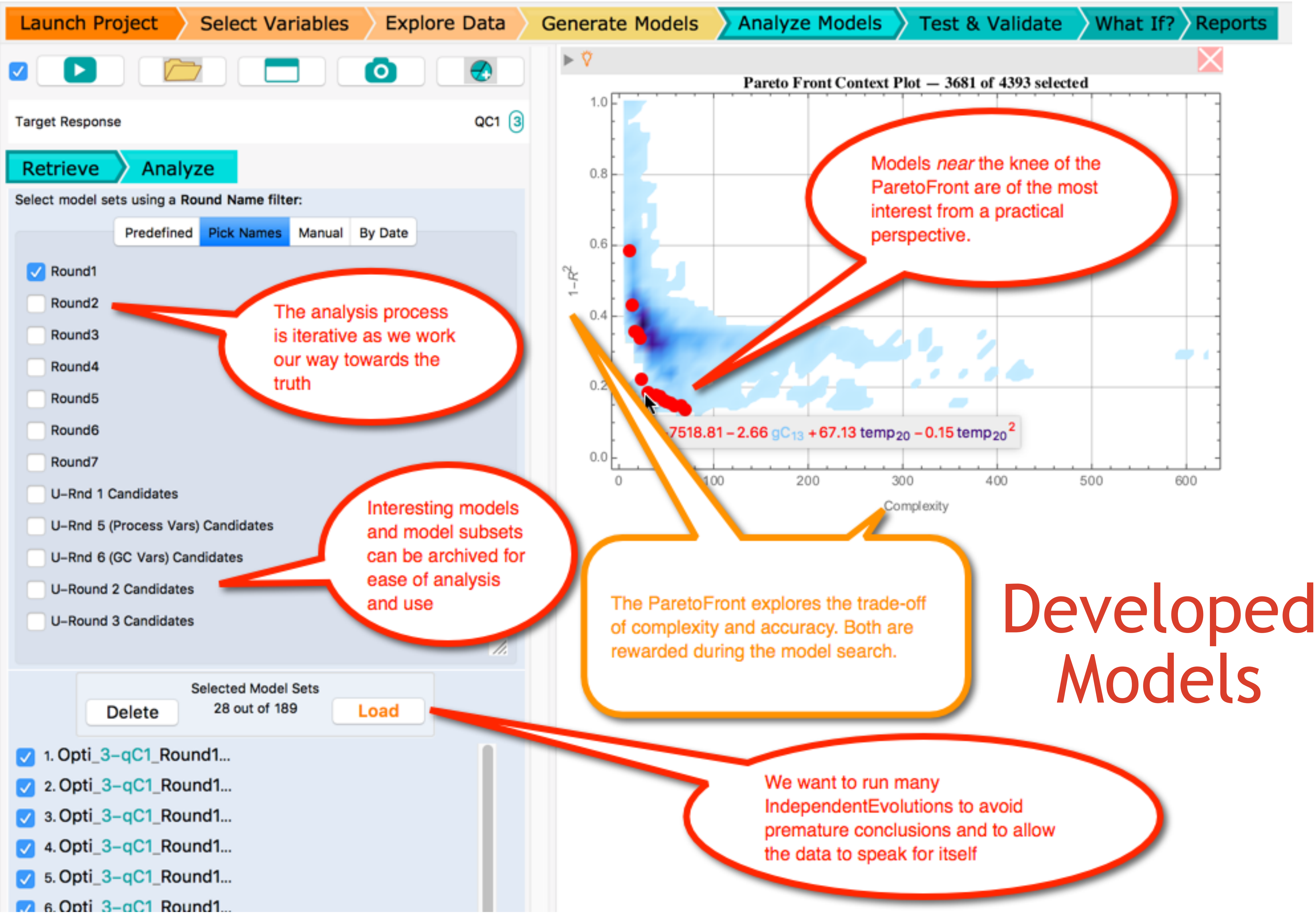
Can impose restrictions on the variables used

Human-interpretable models are developed

Mathematical operators can be specified

Generate Models

Developed Models



Target Response QC1 3

Select model sets using a **Round Name** filter:

☑ Round1
□ Round2
□ Round3
□ Round4
□ Round5
□ Round6
□ Round7
□ U-Rnd 1 Candidates
□ U-Rnd 5 (Process Vars) Candidates
□ U-Rnd 6 (GC Vars) Candidates
□ U-Round 2 Candidates
□ U-Round 3 Candidates

Selected Model Sets
28 out of 189

☑ 1. Opti_3-qC1_Round1...
☑ 2. Opti_3-qC1_Round1...
☑ 3. Opti_3-qC1_Round1...
☑ 4. Opti_3-qC1_Round1...
☑ 5. Opti_3-qC1_Round1...
☑ 6. Opti_3-qC1_Round1...

$7518.81 - 2.66\, gC_{13} + 67.13\, temp_{20} - 0.15\, temp_{20}^{2}$

The analysis process is iterative as we work our way towards the truth

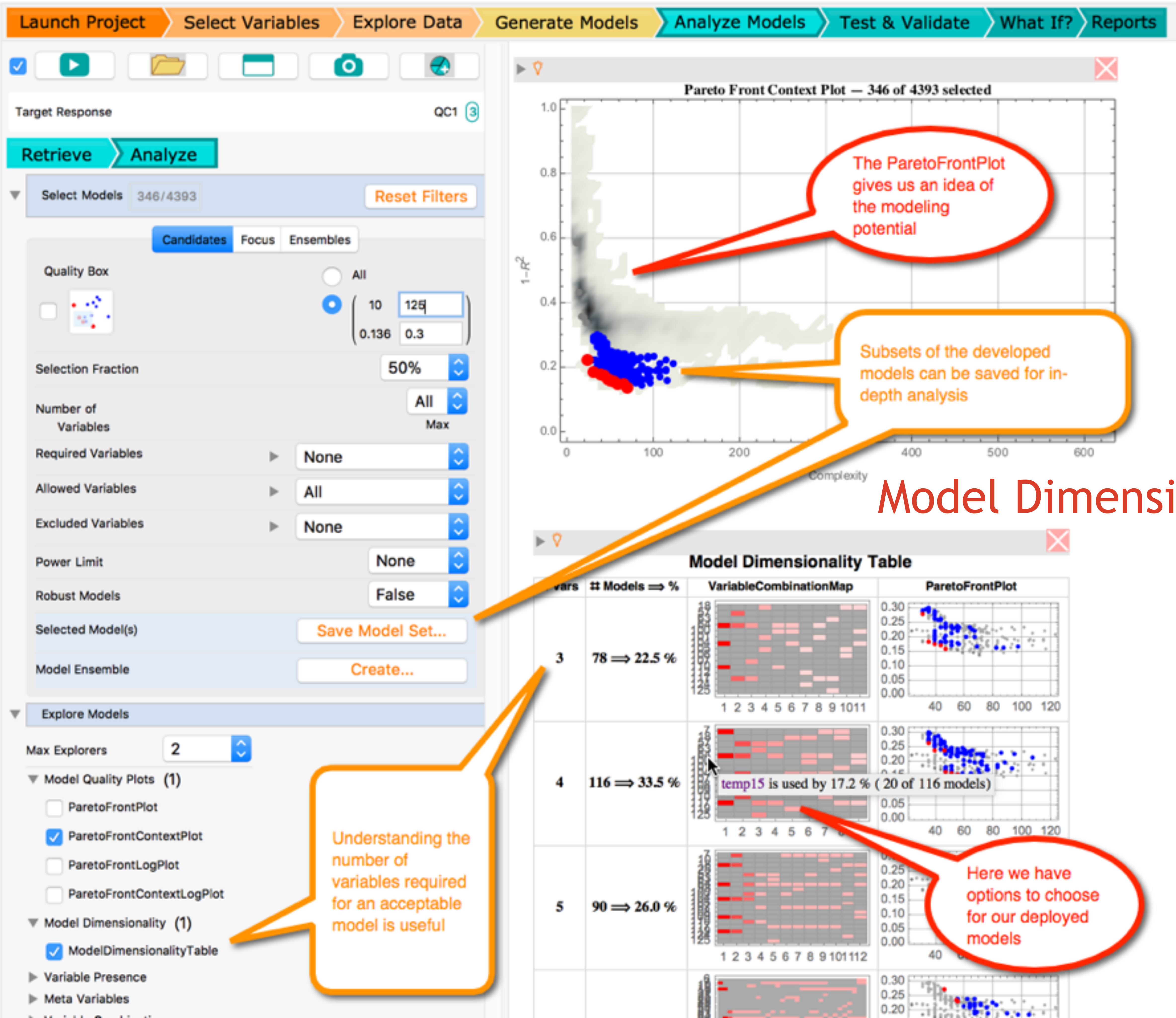
Interesting models and model subsets can be archived for ease of analysis and use

Models *near* the knee of the ParetoFront are of the most interest from a practical perspective.

The ParetoFront explores the trade-off of complexity and accuracy. Both are rewarded during the model search.

We want to run many IndependentEvolutions to avoid premature conclusions and to allow the data to speak for itself

Model Dimensionality



Target Response QC1 3

Select Models 346/4393 Reset Filters

Quality Box
○ All
● 10 125
0.136 0.3

Selection Fraction 50%

Number of Variables All Max

Required Variables None

Allowed Variables All

Excluded Variables None

Power Limit None

Robust Models False

Selected Model(s) Save Model Set...

Model Ensemble Create...

Explore Models

Max Explorers 2

Model Quality Plots (1)
☐ ParetoFrontPlot
☑ ParetoFrontContextPlot
☐ ParetoFrontLogPlot
☐ ParetoFrontContextLogPlot

Model Dimensionality (1)
☑ ModelDimensionalityTable

Variable Presence

Meta Variables

Pareto Front Context Plot — 346 of 4393 selected

The ParetoFrontPlot gives us an idea of the modeling potential

Subsets of the developed models can be saved for in-depth analysis

Model Dimensionality Table

Vars	# Models ⟹ %	VariableCombinationMap	ParetoFrontPlot
3	78 ⟹ 22.5 %		
4	116 ⟹ 33.5 %		
5	90 ⟹ 26.0 %		

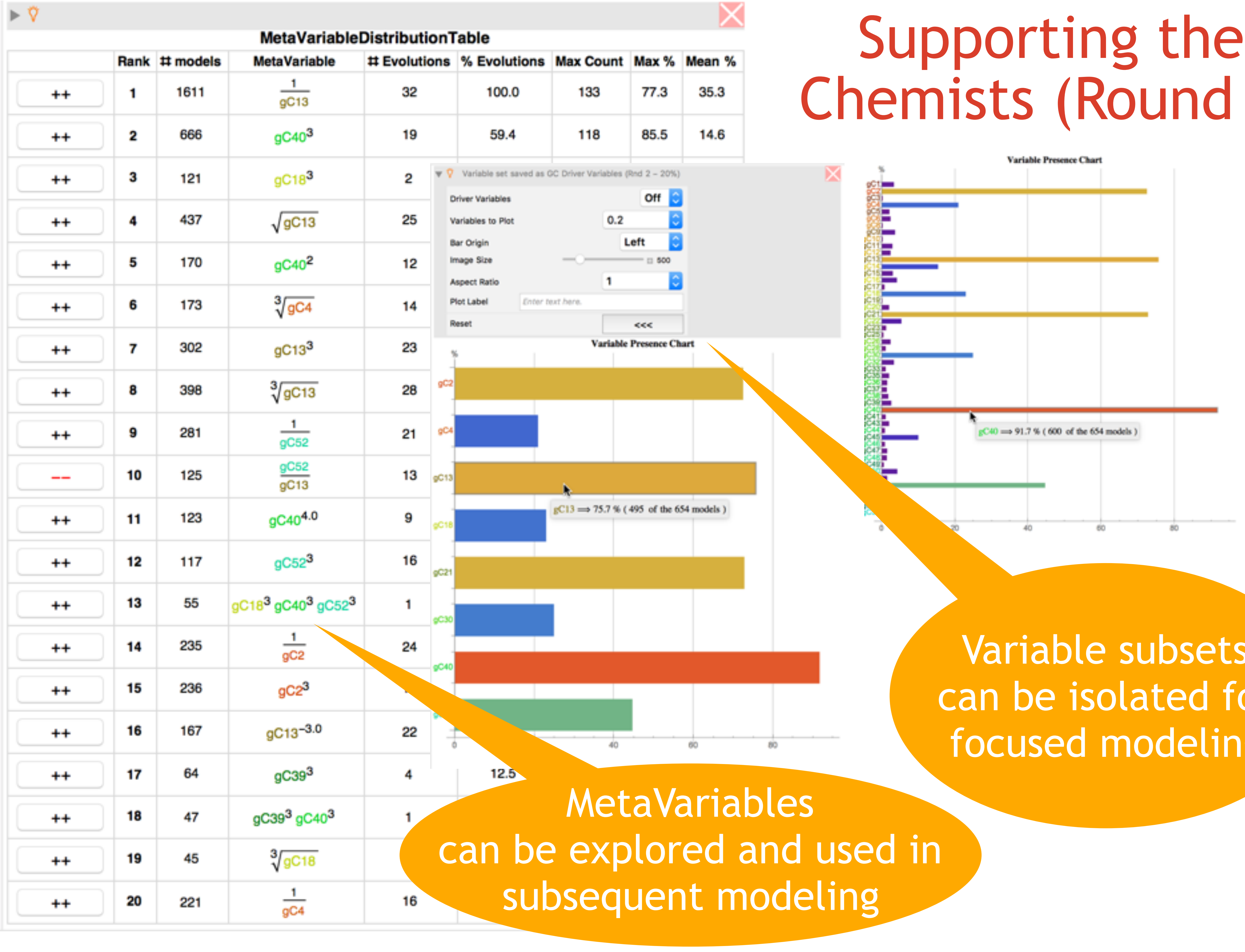
temp15 is used by 17.2 % (20 of 116 models)

Understanding the number of variables required for an acceptable model is useful

Here we have options to choose for our deployed models

volved-analytics.com

Supporting the Chemists (Round 2)



MetaVariableDistributionTable

	Rank	# models	MetaVariable	# Evolutions	% Evolutions	Max Count	Max %	Mean %
++	1	1611	$\frac{1}{gC13}$	32	100.0	133	77.3	35.3
++	2	666	$gC40^{3}$	19	59.4	118	85.5	14.6
++	3	121	$gC18^{3}$	2				
++	4	437	$\sqrt{gC13}$	25				
++	5	170	$gC40^{2}$	12				
++	6	173	$\sqrt[3]{gC4}$	14				
++	7	302	$gC13^{3}$	23				
++	8	398	$\sqrt[3]{gC13}$	28				
++	9	281	$\frac{1}{gC52}$	21				
--	10	125	$\frac{gC52}{gC13}$	13				
++	11	123	$gC40^{4.0}$	9				
++	12	117	$gC52^{3}$	16				
++	13	55	$gC18^{3}\ gC40^{3}\ gC52^{3}$	1				
++	14	235	$\frac{1}{gC2}$	24				
++	15	236	$gC2^{3}$					
++	16	167	$gC13^{-3.0}$	22				
++	17	64	$gC39^{3}$	4	12.5			
++	18	47	$gC39^{3}\ gC40^{3}$	1				
++	19	45	$\sqrt[3]{gC18}$					
++	20	221	$\frac{1}{gC4}$	16				

Variable set saved as GC Driver Variables (Rnd 2 – 20%)

Driver Variables: Off

Variables to Plot: 0.2

Bar Origin: Left

Image Size: 500

Aspect Ratio: 1

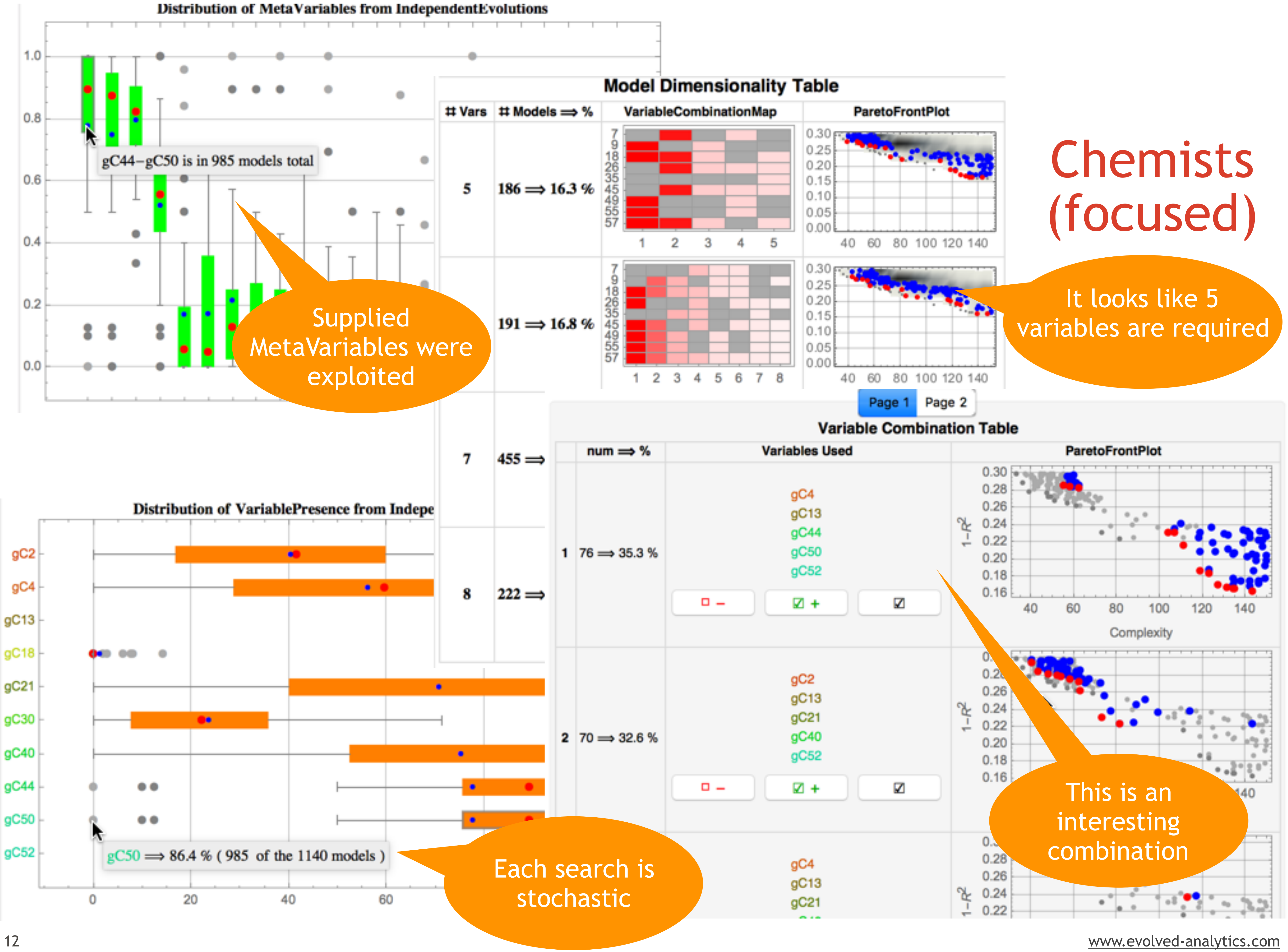
Plot Label: Enter text here.

Reset: <<<

Variable subsets can be isolated for focused modeling

MetaVariables can be explored and used in subsequent modeling

Chemists (focused)



gC44–gC50 is in 985 models total

gC50 ⟹ 86.4 % (985 of the 1140 models)

Model Dimensionality Table

# Vars	# Models ⟹ %	VariableCombinationMap	ParetoFrontPlot
5	186 ⟹ 16.3 %		
	191 ⟹ 16.8 %		
7	455 ⟹		
8	222 ⟹		

Page 1 Page 2

Variable Combination Table

	num ⟹ %	Variables Used	ParetoFrontPlot
1	76 ⟹ 35.3 %	gC4 gC13 gC44 gC50 gC52 □ – ☑ + ☑	
2	70 ⟹ 32.6 %	gC2 gC13 gC21 gC40 gC52 □ – ☑ + ☑	
		gC4 gC13 gC21	

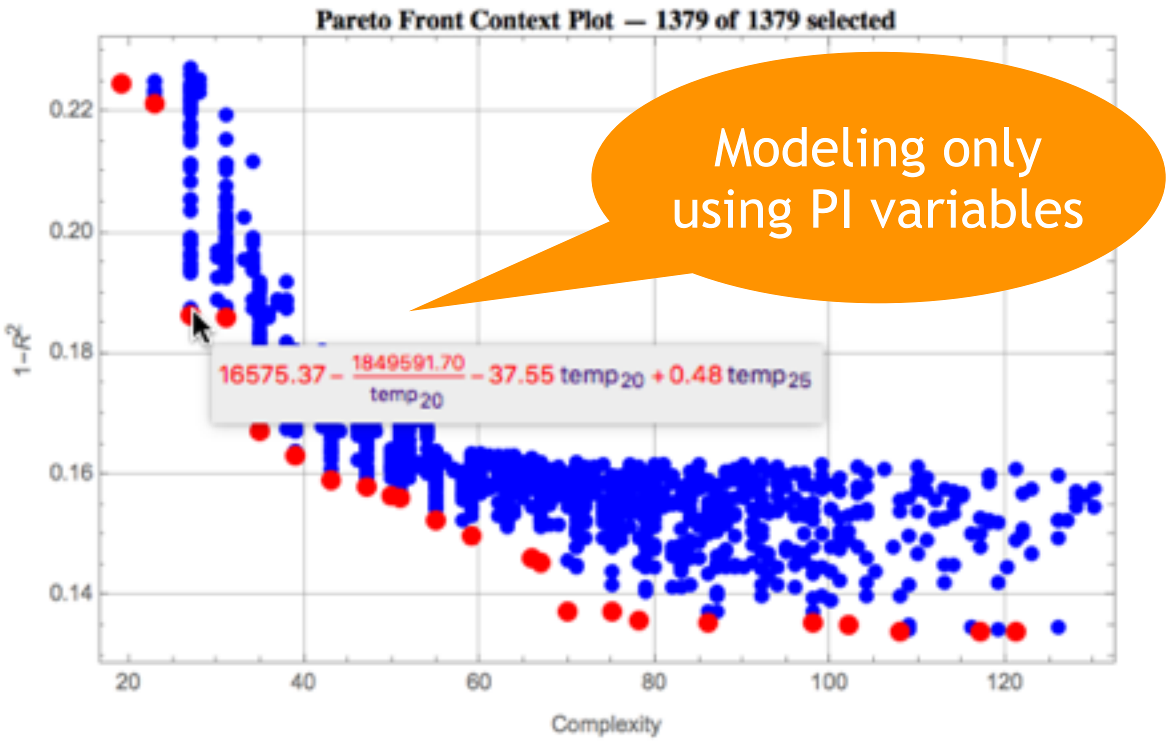
Supplied MetaVariables were exploited

It looks like 5 variables are required

This is an interesting combination

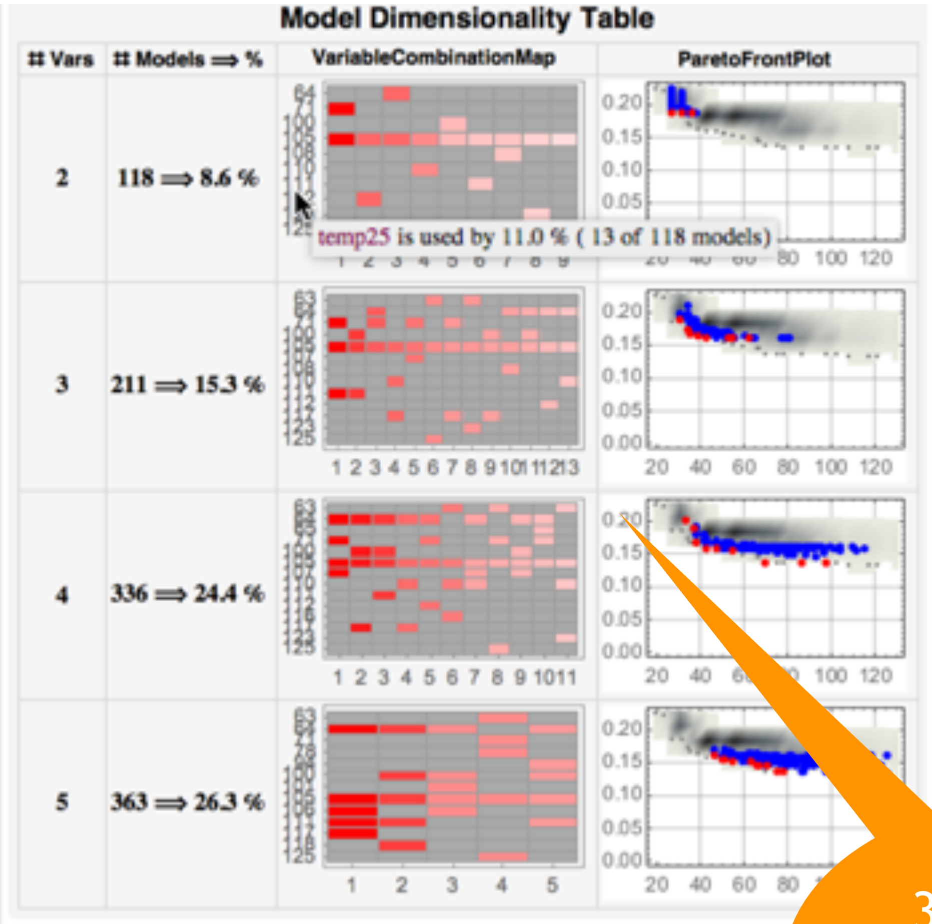
Each search is stochastic

Supporting Production



$16575.37 - \frac{1849591.70}{temp_{20}} - 37.55\, temp_{20} + 0.48\, temp_{25}$

Modeling only using PI variables



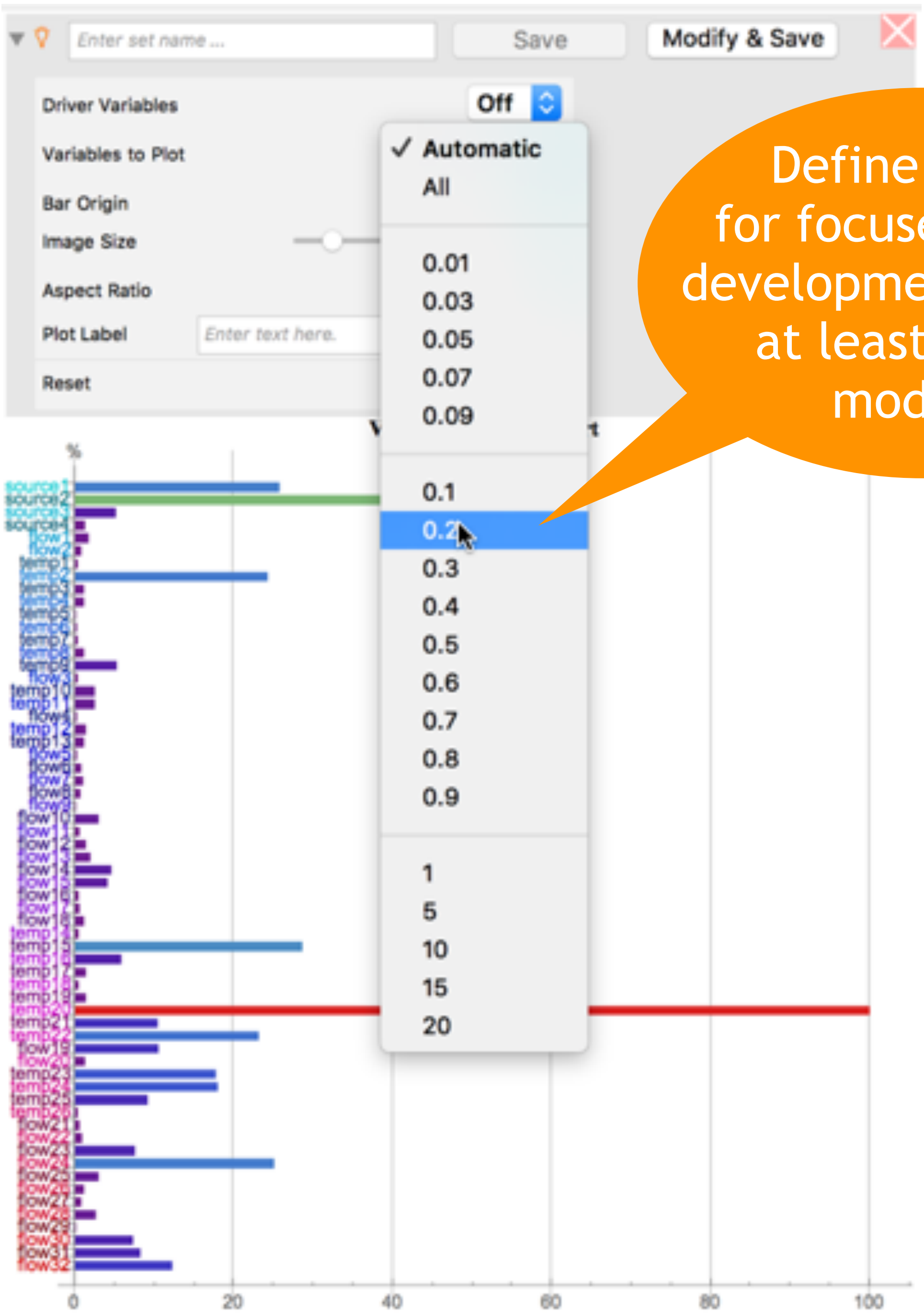
Model Dimensionality Table

# Vars	# Models ⟹ %	VariableCombinationMap	ParetoFrontPlot
2	118 ⟹ 8.6 %		
3	211 ⟹ 15.3 %		
4	336 ⟹ 24.4 %		
5	363 ⟹ 26.3 %		

temp25 is used by 11.0 % (13 of 118 models)

0.20
0.15
0.10
0.05
0.00
20 40 60 80 100 120

3 or 4 vars needed for good-enough models



Enter set name ...

Save

Modify & Save

Driver Variables: Off

Variables to Plot: ✓ Automatic

Bar Origin

Image Size

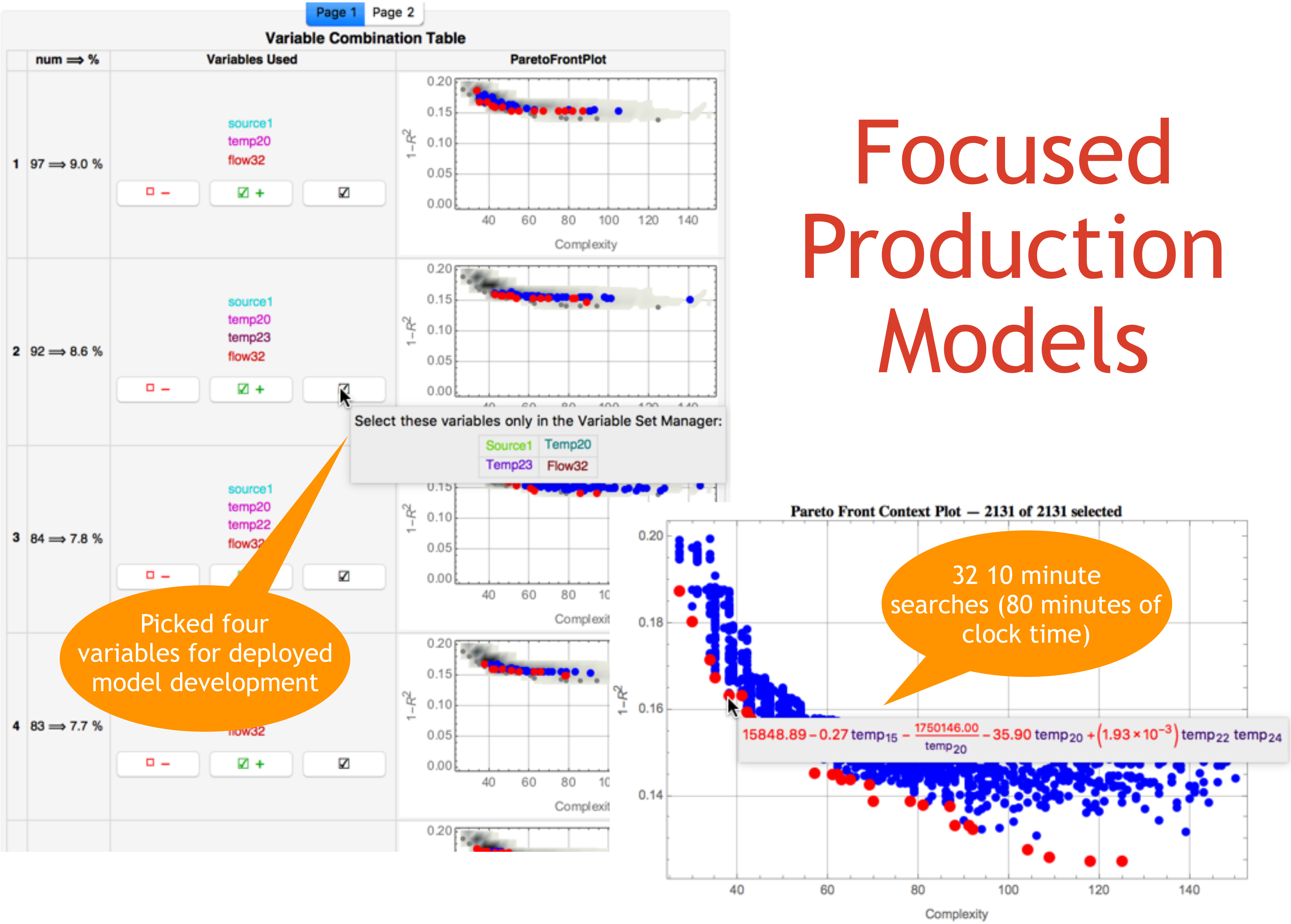
Aspect Ratio

Plot Label: Enter text here.

Reset

Define inputs for focused model development (vars in at least 20% of models)

Focused Production Models



Variable Combination Table

	num ⇒ %	Variables Used	ParetoFrontPlot
1	97 ⇒ 9.0 %	source1 temp20 flow32 □ – ☑ + ☑	
2	92 ⇒ 8.6 %	source1 temp20 temp23 flow32 □ – ☑ + ☑	
3	84 ⇒ 7.8 %	source1 temp20 temp22 flow32 □ – ☑	
4	83 ⇒ 7.7 %	flow32 □ – ☑ + ☑	

Select these variables only in the Variable Set Manager:

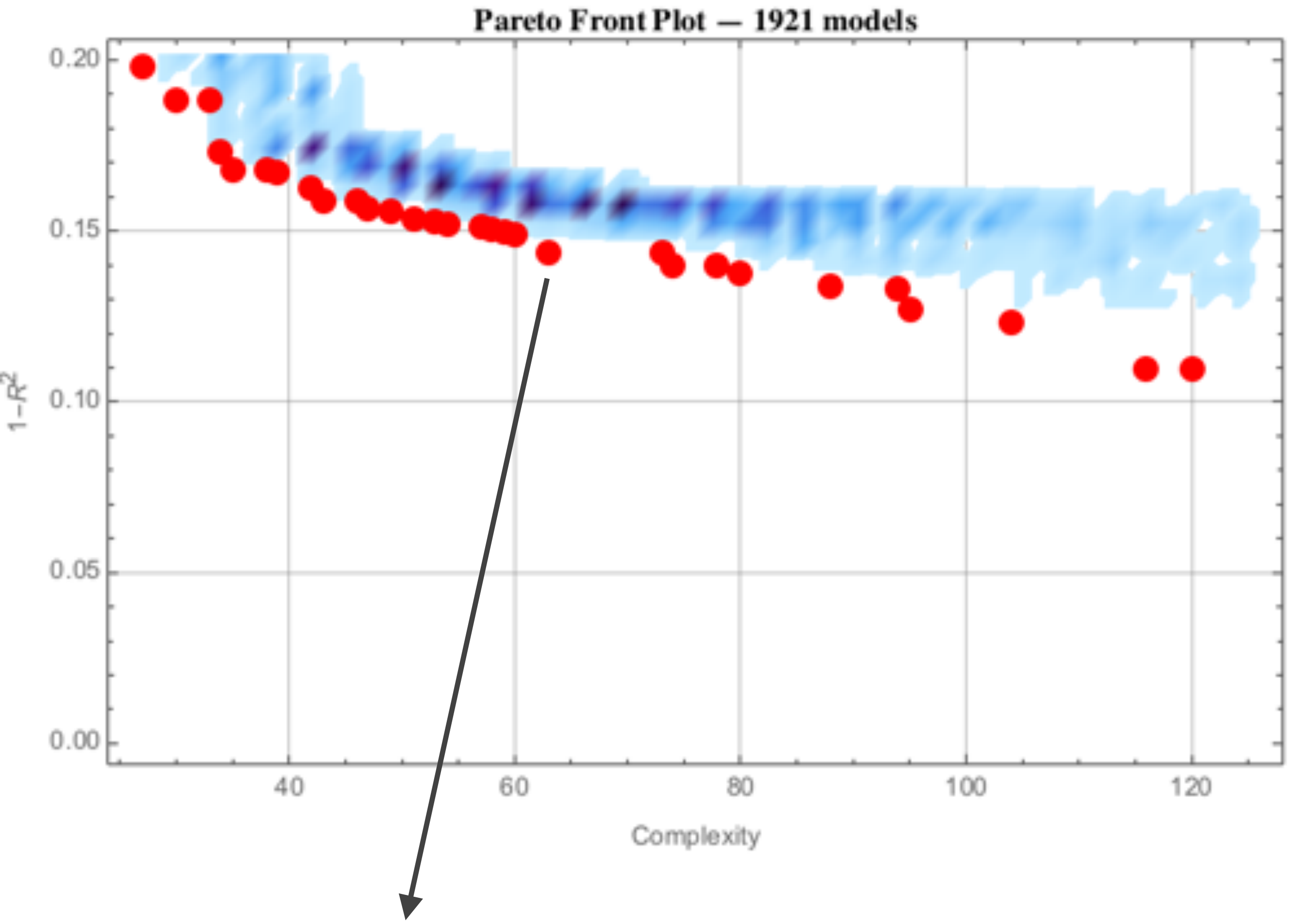
Source1	Temp20
Temp23	Flow32

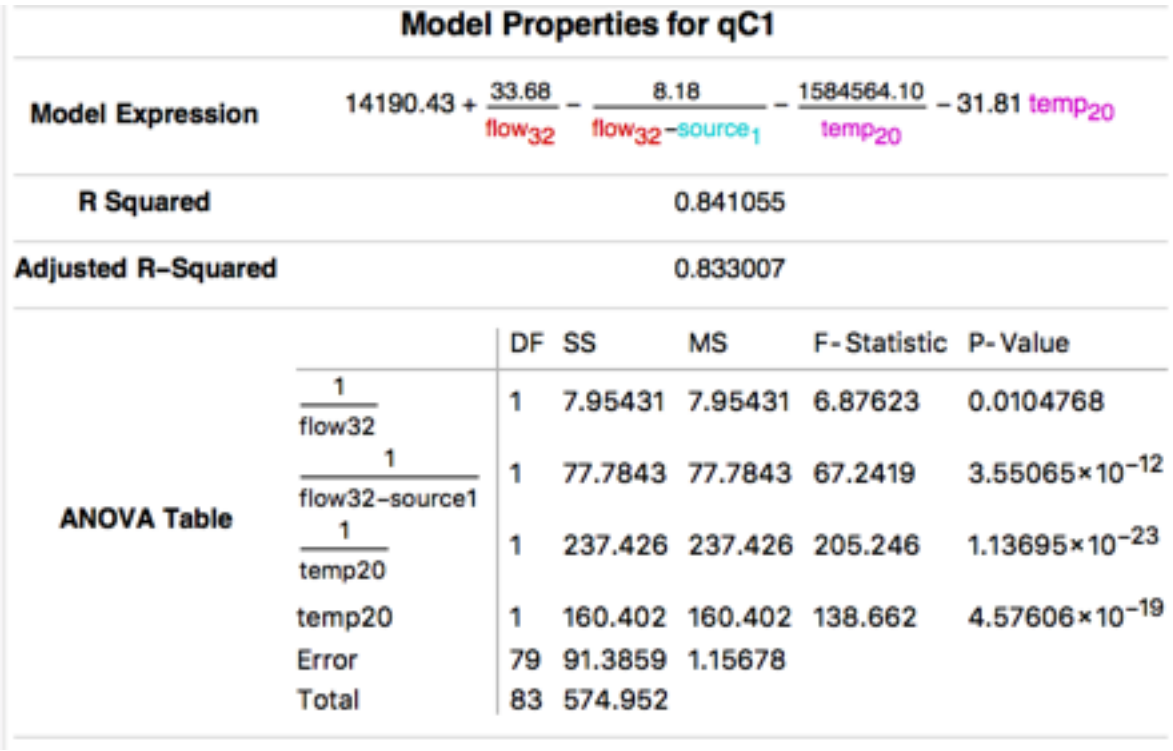
Picked four variables for deployed model development

Pareto Front Context Plot — 2131 of 2131 selected

$$15848.89 - 0.27\,\text{temp}_{15} - \frac{1750146.00}{\text{temp}_{20}} - 35.90\,\text{temp}_{20} + \left(1.93\times10^{-3}\right)\text{temp}_{22}\,\text{temp}_{24}$$

32 10 minute searches (80 minutes of clock time)





Model Properties for qC1

Model Expression	$14190.43 + \frac{33.68}{flow_{32}} - \frac{8.18}{flow_{32} - source_1} - \frac{1584564.10}{temp_{20}} - 31.81\, temp_{20}$
R Squared	0.841055
Adjusted R-Squared	0.833007

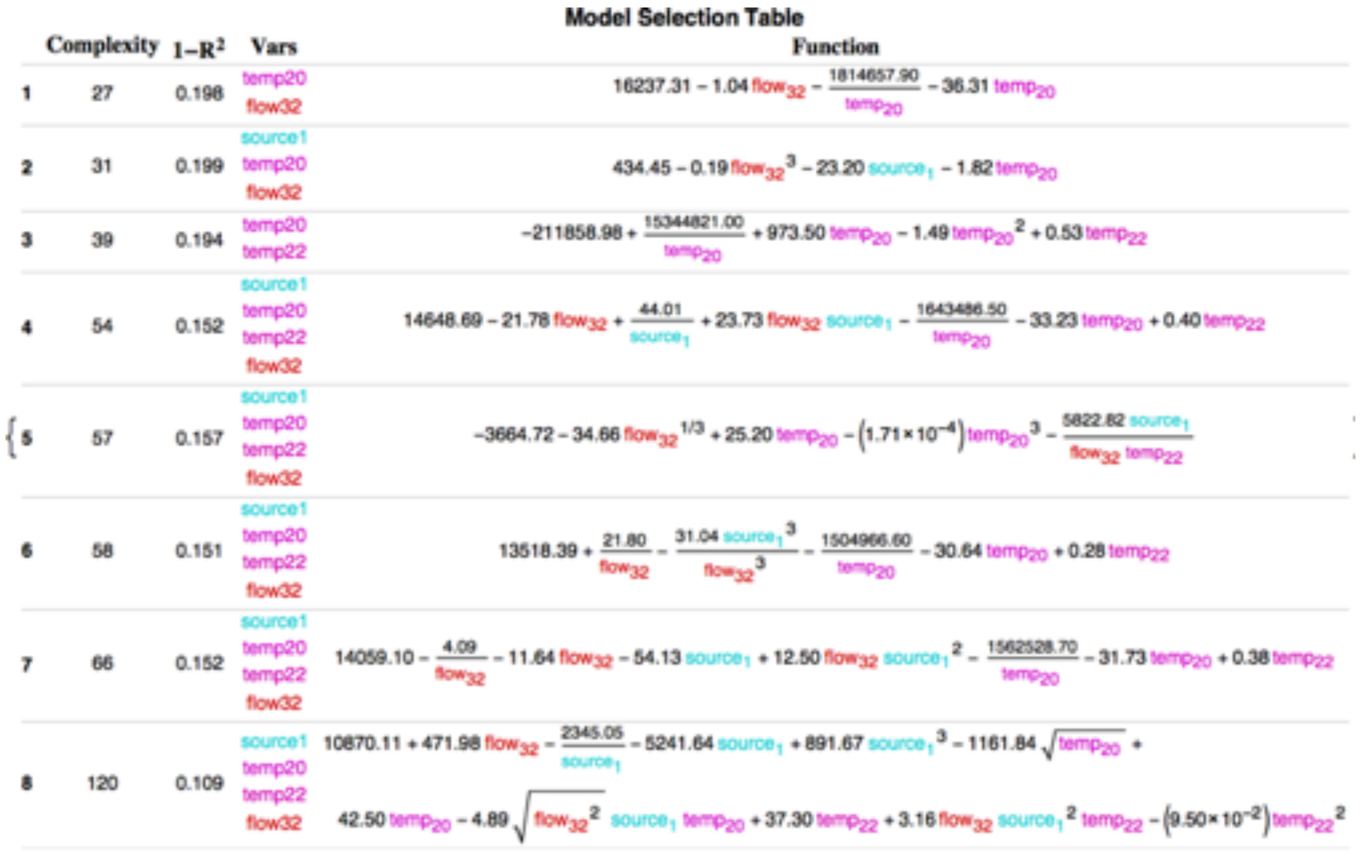
ANOVA Table

	DF	SS	MS	F-Statistic	P-Value
$\frac{1}{flow32}$	1	7.95431	7.95431	6.87623	0.0104768
$\frac{1}{flow32-source1}$	1	77.7843	77.7843	67.2419	3.55065×10^{-12}
$\frac{1}{temp20}$	1	237.426	237.426	205.246	1.13695×10^{-23}
temp20	1	160.402	160.402	138.662	4.57606×10^{-19}
Error	79	91.3859	1.15678		
Total	83	574.952			

Deployable Models

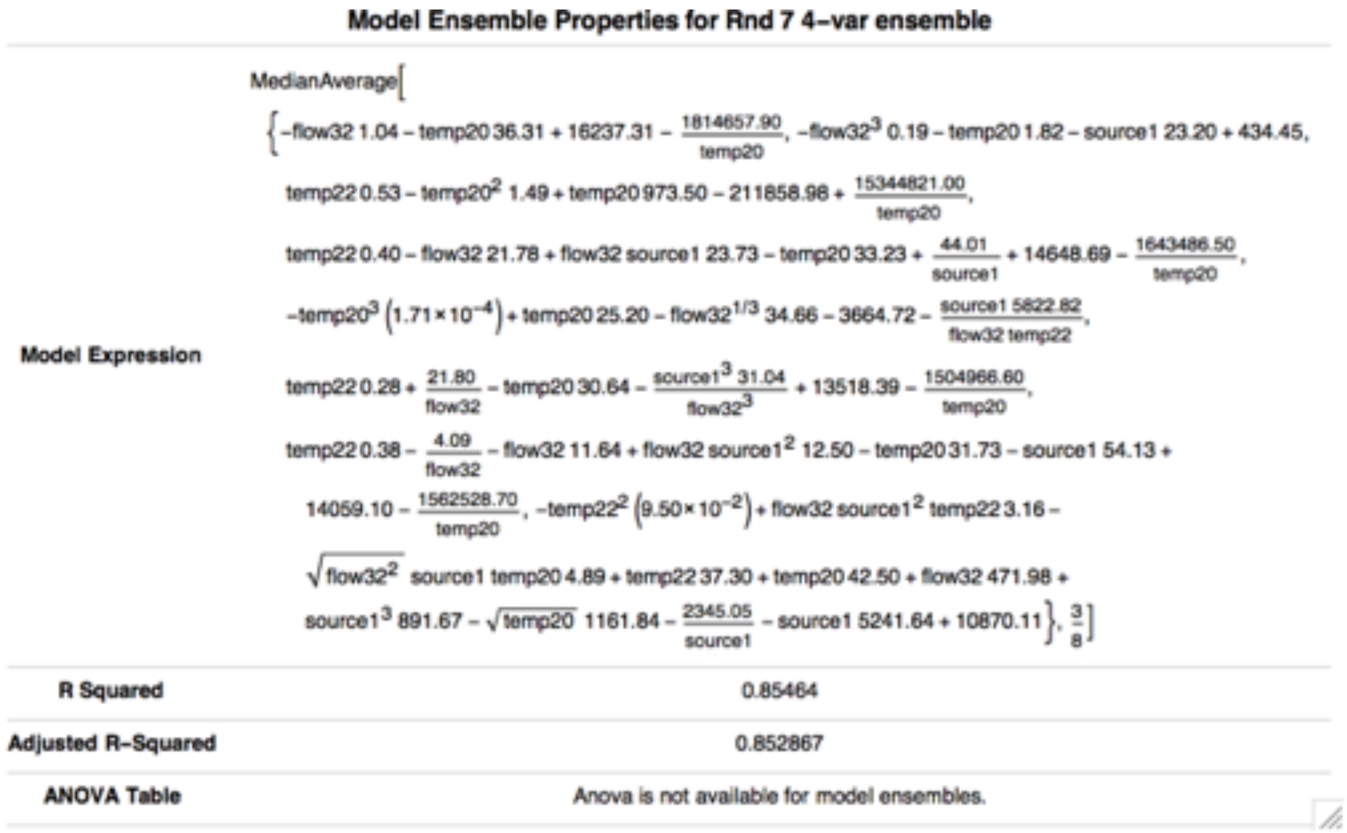
- 80 minutes of model development (32 independent searches of 10 minutes on a quad-core laptop)
- Models were rewarded for simplicity and accuracy
- The individual models are good; however, we want trustable models based upon ensembles

An Ensemble



Model Selection Table

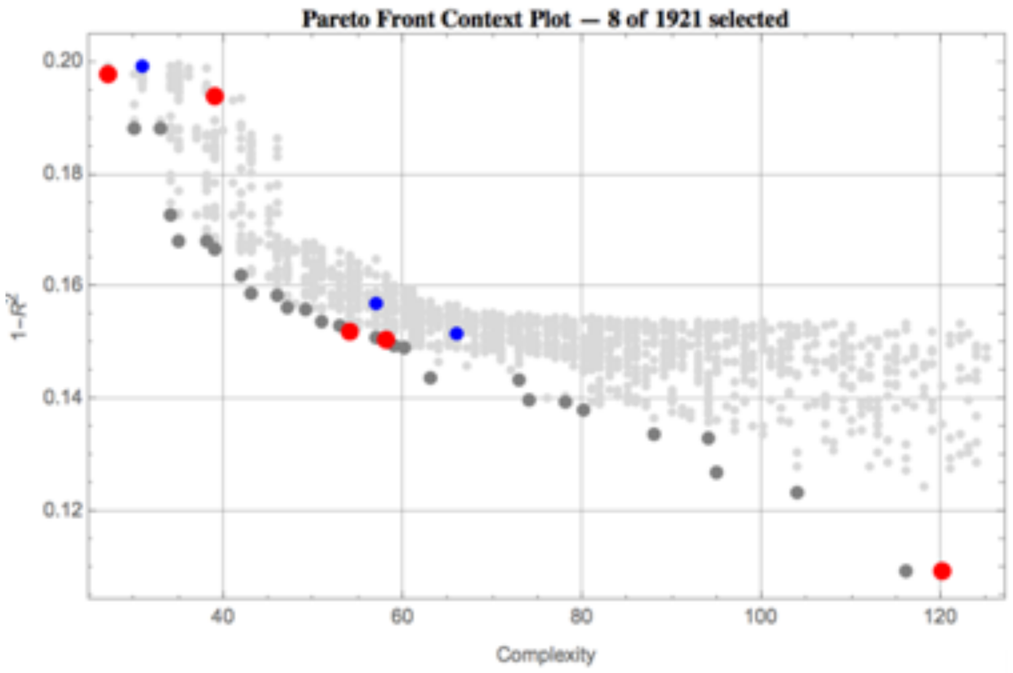
	Complexity	1−R²	Vars	Function
1	27	0.198	temp20 flow32	$16237.31 - 1.04\,\mathrm{flow}_{32} - \frac{1814657.90}{\mathrm{temp}_{20}} - 36.31\,\mathrm{temp}_{20}$
2	31	0.199	source1 temp20 flow32	$434.45 - 0.19\,\mathrm{flow}_{32}{}^{3} - 23.20\,\mathrm{source}_{1} - 1.82\,\mathrm{temp}_{20}$
3	39	0.194	temp20 temp22	$-211858.98 + \frac{15344821.00}{\mathrm{temp}_{20}} + 973.50\,\mathrm{temp}_{20} - 1.49\,\mathrm{temp}_{20}{}^{2} + 0.53\,\mathrm{temp}_{22}$
4	54	0.152	source1 temp20 temp22 flow32	$14648.69 - 21.78\,\mathrm{flow}_{32} + \frac{44.01}{\mathrm{source}_{1}} + 23.73\,\mathrm{flow}_{32}\,\mathrm{source}_{1} - \frac{1643486.50}{\mathrm{temp}_{20}} - 33.23\,\mathrm{temp}_{20} + 0.40\,\mathrm{temp}_{22}$
5	57	0.157	source1 temp20 temp22 flow32	$-3664.72 - 34.66\,\mathrm{flow}_{32}{}^{1/3} + 25.20\,\mathrm{temp}_{20} - \left(1.71\times10^{-4}\right)\mathrm{temp}_{20}{}^{3} - \frac{5822.82\,\mathrm{source}_{1}}{\mathrm{flow}_{32}\,\mathrm{temp}_{22}}$
6	58	0.151	source1 temp20 temp22 flow32	$13518.39 + \frac{21.80}{\mathrm{flow}_{32}} - \frac{31.04\,\mathrm{source}_{1}{}^{3}}{\mathrm{flow}_{32}{}^{3}} - \frac{1504966.60}{\mathrm{temp}_{20}} - 30.64\,\mathrm{temp}_{20} + 0.28\,\mathrm{temp}_{22}$
7	66	0.152	source1 temp20 temp22 flow32	$14059.10 - \frac{4.09}{\mathrm{flow}_{32}} - 11.64\,\mathrm{flow}_{32} - 54.13\,\mathrm{source}_{1} + 12.50\,\mathrm{flow}_{32}\,\mathrm{source}_{1}{}^{2} - \frac{1562528.70}{\mathrm{temp}_{20}} - 31.73\,\mathrm{temp}_{20} + 0.38\,\mathrm{temp}_{22}$
8	120	0.109	source1 temp20 temp22 flow32	$10870.11 + 471.98\,\mathrm{flow}_{32} - \frac{2345.05}{\mathrm{source}_{1}} - 5241.64\,\mathrm{source}_{1} + 891.67\,\mathrm{source}_{1}{}^{3} - 1161.84\sqrt{\mathrm{temp}_{20}} + 42.50\,\mathrm{temp}_{20} - 4.89\sqrt{\mathrm{flow}_{32}{}^{2}}\,\mathrm{source}_{1}\,\mathrm{temp}_{20} + 37.30\,\mathrm{temp}_{22} + 3.16\,\mathrm{flow}_{32}\,\mathrm{source}_{1}{}^{2}\,\mathrm{temp}_{22} - \left(9.50\times10^{-2}\right)\mathrm{temp}_{22}{}^{2}$



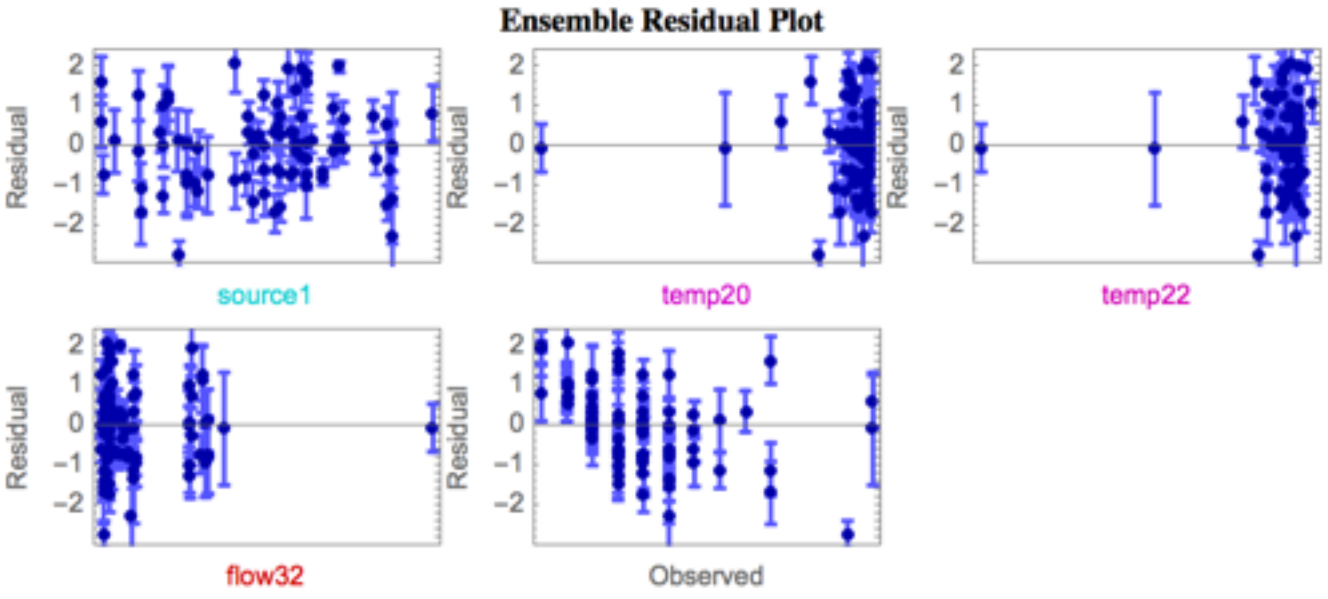
Model Ensemble Properties for Rnd 7 4-var ensemble

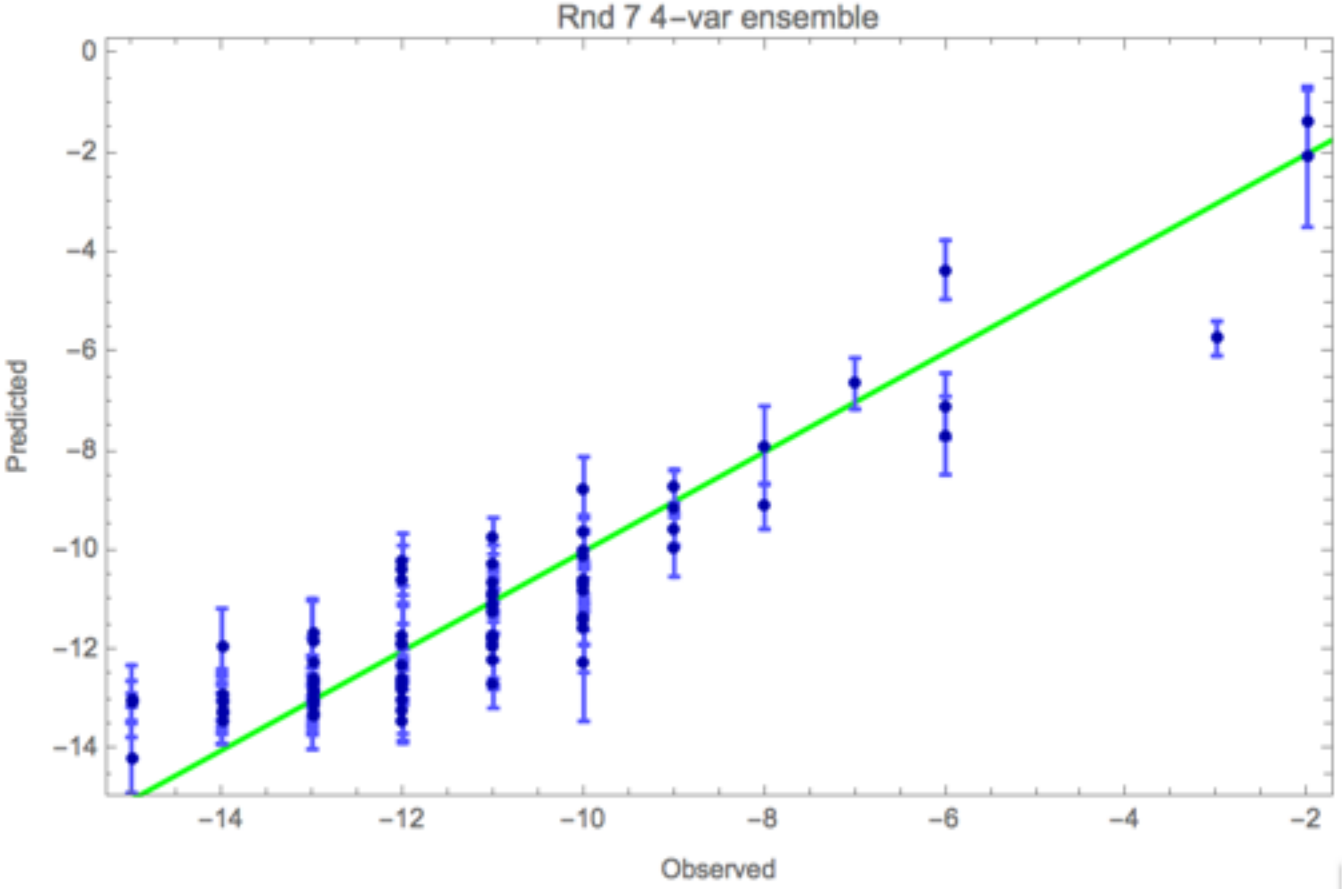
Model Expression	$\mathrm{MedianAverage}\Big[\Big\{-\mathrm{flow32}\,1.04 - \mathrm{temp20}\,36.31 + 16237.31 - \frac{1814657.90}{\mathrm{temp20}},\ -\mathrm{flow32}^3\,0.19 - \mathrm{temp20}\,1.82 - \mathrm{source1}\,23.20 + 434.45,\ \mathrm{temp22}\,0.53 - \mathrm{temp20}^2\,1.49 + \mathrm{temp20}\,973.50 - 211858.98 + \frac{15344821.00}{\mathrm{temp20}},\ \mathrm{temp22}\,0.40 - \mathrm{flow32}\,21.78 + \mathrm{flow32}\,\mathrm{source1}\,23.73 - \mathrm{temp20}\,33.23 + \frac{44.01}{\mathrm{source1}} + 14648.69 - \frac{1643486.50}{\mathrm{temp20}},\ -\mathrm{temp20}^3\left(1.71\times10^{-4}\right) + \mathrm{temp20}\,25.20 - \mathrm{flow32}^{1/3}\,34.66 - 3664.72 - \frac{\mathrm{source1}\,5822.82}{\mathrm{flow32}\,\mathrm{temp22}},\ \mathrm{temp22}\,0.28 + \frac{21.80}{\mathrm{flow32}} - \mathrm{temp20}\,30.64 - \frac{\mathrm{source1}^3\,31.04}{\mathrm{flow32}^3} + 13518.39 - \frac{1504966.60}{\mathrm{temp20}},\ \mathrm{temp22}\,0.38 - \frac{4.09}{\mathrm{flow32}} - \mathrm{flow32}\,11.64 + \mathrm{flow32}\,\mathrm{source1}^2\,12.50 - \mathrm{temp20}\,31.73 - \mathrm{source1}\,54.13 + 14059.10 - \frac{1562528.70}{\mathrm{temp20}},\ -\mathrm{temp22}^2\left(9.50\times10^{-2}\right) + \mathrm{flow32}\,\mathrm{source1}^2\,\mathrm{temp22}\,3.16 - \sqrt{\mathrm{flow32}^2}\,\mathrm{source1}\,\mathrm{temp20}\,4.89 + \mathrm{temp22}\,37.30 + \mathrm{temp20}\,42.50 + \mathrm{flow32}\,471.98 + \mathrm{source1}^3\,891.67 - \sqrt{\mathrm{temp20}}\,1161.84 - \frac{2345.05}{\mathrm{source1}} - \mathrm{source1}\,5241.64 + 10870.11\Big\},\ \frac{3}{8}\Big]$
R Squared	0.85464
Adjusted R-Squared	0.852867
ANOVA Table	Anova is not available for model ensembles.

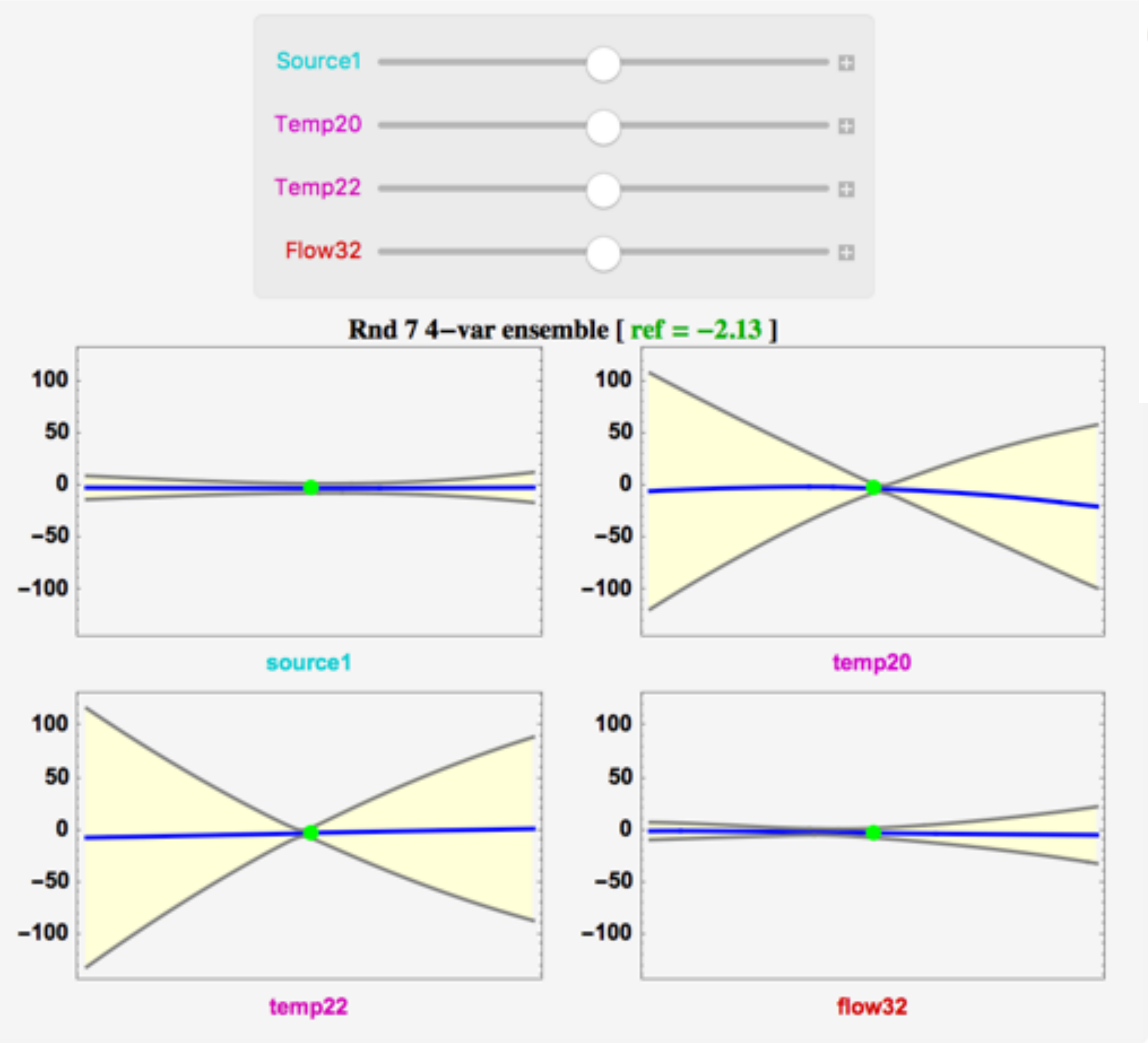
- From candidate (accurate and simple) models, models chosen for their diversity
- Ensemble has better prediction accuracy than individual models
- Divergence of models provides a trust metric!



Ensemble Performance







- Ensemble predictions have a trust metric based upon divergence.
- Temperatures are coupled so they cannot be varied independently ⇒ prediction spread greatly increases if we try to do that!

Conclusions

- One more thing …
  - Modeling and data results can be archived to analysis report at the click of a button
  - A function package is available for use in a notebook front end as well as to facilitate automated analysis flows
- For more information or trial licenses for DataModeler, contact
  - info@evolved-analytics.com
- We also do consulting and custom analysis system development
- We have offices in the US and Europe