Real-World Data Modeling

Mark Kotanchek

Evolved Analytics LLC

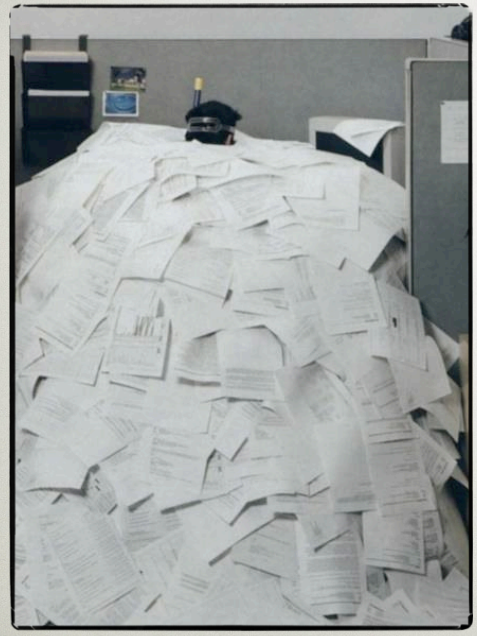
www.evolved-analytics.com

GECCO'12 Companion, July 7–11, 2012, Philadelphia, PA, USA.
ACM 978-1-4503-1178-6/12/07.

Why are we here?

Money

Specifically, we want to convert data into money



Dealing with the Data Deluge

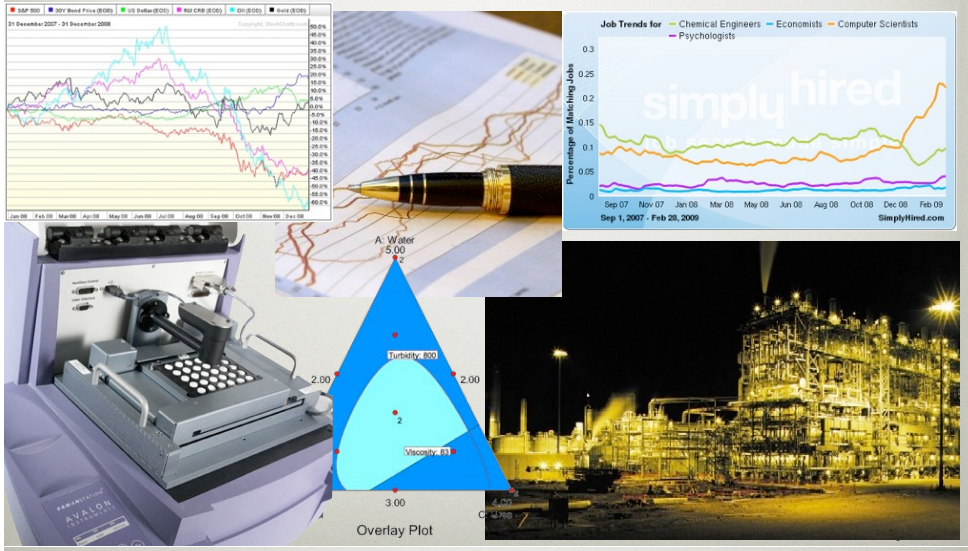
Conventional technologies struggle ...

- too much data
- too many variables
- ugly data
- too little time
- the simplifying assumptions are not valid
- models are big piles of "trust-me"

Multi-objective symbolic regression ...

- can help to make sense ... and money ... out of the data deluge.
- trustable models are possible
- cheaper-better-faster really IS possible!

Data, Data Everywhere



The Essence is Numbers (and some Context)

symbol	name	BookValu	FloatShar	ForwardP	MarketCa	PEGRatio	PERatio	PriceTarg	PriceToBo	PriceToSa	ShortRatii	YearPERa	EarningsF	EBITDA	ForwardE	QuarterFo	YearEarni	Dividend\	DividendF	Dividend
A	Agilent Te	7.311	3.4E+08	9.23	5.7E+09	0.85	8.74	26.86	2.23	0.99	1.2	11.11	1.868	1.0E+09	1.77	0.33	1.47		0	2.057
AA	Alcoa Inc	18.755	7.9E+08	16.69	7.8E+09	0.35	4.9	17.27	0.55	0.28	1.9	6.68	2.113	4.0E+09	0.62	-0.15	1.55	0.0657	0.68	0.17
AAPL	Apple Inc	23.674	8.8E+08	13.82	7.9E+10	0.87	16.64	142.48	3.77	2.44	0.4	17.08	5.358	6.7E+09	6.45	1.17	5.22		0	
ABC	Amerisou	17.348	1.5E+08	9.76	5.3E+09	0.94	22.09	39.04	1.96	0.08	1.3	10.85	1.542	9.2E+08	3.49	0.89	3.14	0.0095	0.325	0.1
ABT	Abbott La	12.468	1.5E+09	14.19	8.1E+10	1.31	17.87	61	4.18	2.8	1.2	15.68	2.914	7.6E+09	3.67	0.76	3.32	0.027	1.405	0.36
ACAS	American	24.426	2.0E+08	1.29	6.4E+08	0.19		11.65	0.12	0.53	5.7	1.07	-8.386		2.37	0.62	2.85	1.0131	3.09	1.05
ACS	Affiliated (	24.304	8.8E+07	9.81	4.2E+09	0.84	11.77	51.44	1.71	0.65	1.6	11.37	3.526	1.1E+09	4.23	0.93	3.65		0	
ADBE	Adobe Sy	8.225	5.3E+08	10.74	1.1E+10	0.75	14.36	26.42	2.64	3.22	0.9	11.98	1.51	1.3E+09	2.02	0.45	1.81		0	0.0065
ADI	Analog De	8.312	2.8E+08	15.14	5.3E+09	1.02	7.44	20.82	2.37	2.22	0.9	18.74	2.646	7.8E+08	1.3	0.24	1.05	0.0396	0.78	0.2
ADM	Archer Da	21.097	6.2E+08	9.84	1.8E+10	0.68	7.59	29.36	1.34	0.23		8.24	3.735	4.1E+09	2.88	0.64	3.44	0.0183	0.52	0.13
ADP	Automatic	9.551	5.0E+08	14.27	1.9E+10	1.17	15.99	42.68	3.9	2.11	2.2	15.65	2.33	2.1E+09	2.61	0.81	2.38	0.0322	1.2	0.33
ADSK	Autodesk	6.187	2.2E+08	11.93	4.2E+09	0.86	11.64	18.36	3.1	1.79	2	9.85	1.651	6.0E+08	1.61	0.37	1.95		0	0.015
AEE	Ameren C	33.395	2.1E+08	9.77	6.9E+09	2.81	10.4	32.06	0.98	0.91	1.3	11.17	3.136	2.1E+09	3.34	0.63	2.92	0.0779	2.54	0.635
AEP	American	27.089	4.0E+08	9.26	1.3E+10	1.82	8.34	38.11	1.11	0.84		9.5	3.623	4.2E+09	3.26	0.94	3.18	0.0543	1.64	0.41
AES	The Aes (	6.387	6.6E+08	6.49	5.3E+09	0.47	3.98	14.6	1.18	0.31	1.1	6.91	1.891	4.4E+09	1.16	0.28	1.09		0	
AET	Aetna Inc	20.162	4.6E+08	5.99	1.2E+10	0.44	7.46	40.09	1.21	0.37	0.7	6.21	3.274	3.1E+09	4.08	0.95	3.93	0.0016	0.04	0.04
AFL	Aflac Inco	13.64	4.5E+08	10.14	2.1E+10	0.78	15.66	55.89	3.41	1.33	3.2	11.6	2.971	2.3E+09	4.59	1.11	4.01	0.0206	0.96	0.24
AGN	Allergan I	13.147	3.0E+08	13.51	1.1E+10	1	19.3	48.25	2.81	2.53	1.6	14.47	1.991	1.2E+09	2.73	0.61	2.55	0.0041	0.15	0.05
AIG	American	26.463	2.4E+09	1.35	4.5E+09			5.96	0.07	0.09			-16.523	-4E+10	1.3	0.03	-3.62	0.3543	0.62	0.22
AIV	Apartmen	9.843	8.2E+07	5.11	1.1E+09	1.03	3.7	13.87	1.43	0.68	3	4.59	3.803	9.0E+08	2.76	0.7	3.07	0.5614	7.91	1.8
AIZ	Assurant	31.797	1.2E+08	3.86	3.1E+09	0.39	7.75	47.19	0.79	0.35	0.3	4.43	3.238	6.5E+08	6.51	1.65	5.67	0.0215	0.54	0.14
AKAM	Akamai Te	8.904	1.6E+08	9.28	2.6E+09	0.59	21.1	20.64	1.8	3.57	2.1	9.98	0.761	3.3E+08	1.73	0.41	1.61		0	
AKS	Ak Steel H	12.787	1.1E+08	4.52	1.1E+09	0.27	2.3	23.79	0.87	0.16	1	2.7	4.815	1.1E+09	2.45	0.4	4.1	0.0181	0.2	0.05
ALL	Allstate C	31.601	5.3E+08	5.57	1.6E+10	1.19	78.86	39.73	0.94	0.5	1.1	8.3	0.378		5.35	1.45	3.59	0.055	1.64	0.41
ALTR	Altera Co	2.902	2.9E+08	18.15	4.8E+09	0.92	15.31	19.15	5.75	3.61	2.7	14.27	1.991	4.3E+08	0.92	0.22	1.17	0.0114	0.19	0.05
AMAT	Applied M	5.673	1.3E+09	14.42	1.3E+10	3.29	15.26	13.09	1.88	1.74	1.1	39.52	0.699	1.7E+09	0.74	0.04	0.27	0.0225	0.24	0.06
AMD	Advanced	2.202	4.7E+08		1.4E+09			3.86	1.04	0.21	4.5		-5.743	8.5E+08	-1.76	-0.58	-1.55		0	
AMGN	Amgen In	18.727	1.1E+09	12.48	6.1E+10	1.28	15.54	71.39	3.12	4.13	1.9	12.84	3.76	6.7E+09	4.68	1.16	4.55		0	
AMP	Ameriprisi	31.005	2.2E+08	6.61	4.6E+09	0.43	8.56	35.25	0.71	0.58	0.8	6.9	2.563	1.1E+09	3.32	0.73	3.18	0.0292	0.64	0.17
AMT	American	7.245	3.9E+08	41.57	1.1E+10	4.89	47.33	42.07	3.96	7.28	1.6	56.24	0.606	1.0E+09	0.69	0.16	0.51		0	
AMZN	Amazon (	5.89	3.2E+08	35.93	2.2E+10	1.62	36.47	55.45	9.03	1.26	2.5	38.54	1.458	1.1E+09	1.48	0.31	1.38		0	
AN	Autonatio	12.028	1.6E+08	9.89	1.7E+09	0.95		7.21	0.81	0.11	8.1	9.06	-7.021	6.4E+08	0.98	0.25	1.07		0	0.68
ANF	Abercrom	20.686	8.0E+07	10.67	2.0E+09	0.58	4.85	24	1.1	0.52	1.5	6.84	4.684	8.6E+08	2.13	0.22	3.32	0.0308	0.7	0.175
AOC	Aon Corp	22.044	2.5E+08	12.11	1.1E+10	1.36	7.84	51.9	1.92	1.46	0.8	14.87	5.404	1.3E+09	3.5	0.99	2.85	0.0142	0.6	0.15
APA	Apache C	55.77	3.3E+08	10.57	2.3E+10	0.63	5.28	99.35	1.33	1.84	1	5.99	14.023	1.0E+10	7	1.46	12.36	0.0095	0.7	0.15
APC	Anadarko	39.672	4.6E+08	21.55	1.7E+10	1.21	6.74	58.2	0.99	1.24	1.3	7.52	5.822	8.5E+09	1.82	0.34	5.22	0.0092	0.36	0.09

www.evolved-analytics.com

We want to convert numbers into ...

Identify Outliers

Variable Selection

Prediction

Guidance

Money

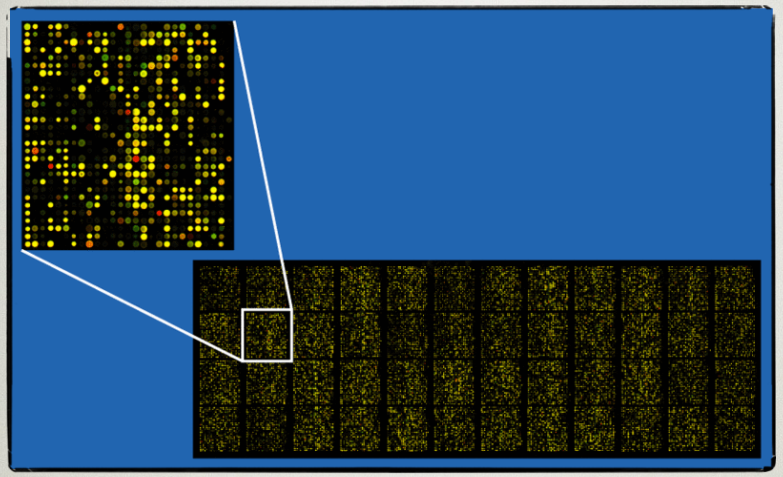
Variable Relationships

Insight & Understanding

Emulators & Optimization

Early Warning & Risk Management

www.evolved-analytics.com



Variable Selection

Simply knowing which of the inputs REALLY matter can be important and valuable

www.evolved-analytics.com

Variable Relationships

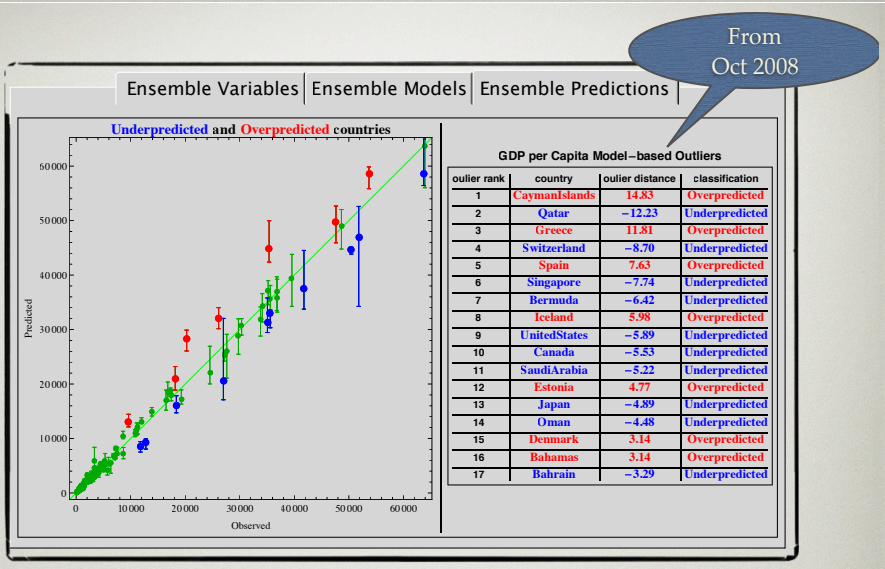
Identifying key variable combinations (metavariables) can lead to insights and better models

www.evolved-analytics.com

Prediction

Inferring from observables the current or future quality, price, temperature, demand, etc. can be valuable.

www.evolved-analytics.com



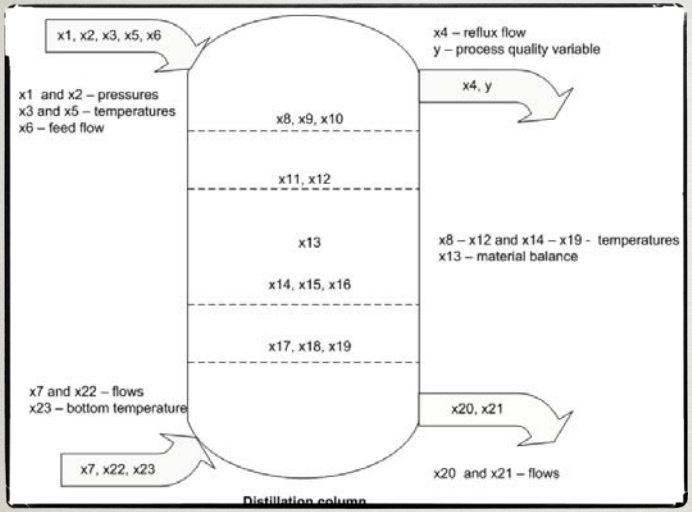
GDP per Capita Model–based Outliers

oulier rank	country	oulier distance	classification
1	CaymanIslands	14.83	Overpredicted
2	Qatar	−12.23	Underpredicted
3	Greece	11.81	Overpredicted
4	Switzerland	−8.70	Underpredicted
5	Spain	7.63	Overpredicted
6	Singapore	−7.74	Underpredicted
7	Bermuda	−6.42	Underpredicted
8	Iceland	5.98	Overpredicted
9	UnitedStates	−5.89	Underpredicted
10	Canada	−5.53	Underpredicted
11	SaudiArabia	−5.22	Underpredicted
12	Estonia	4.77	Overpredicted
13	Japan	−4.89	Underpredicted
14	Oman	−4.48	Underpredicted
15	Denmark	3.14	Overpredicted
16	Bahamas	3.14	Overpredicted
17	Bahrain	−3.29	Underpredicted

Outlier Detection

An outlier is either a nuisance or a critical nugget of information. The first step in that assessment is to identify them

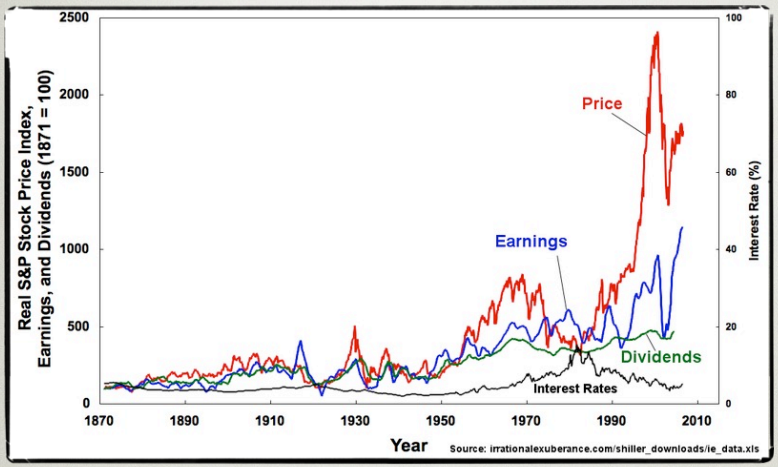
www.evolved-analytics.com



Emulators & Optimization

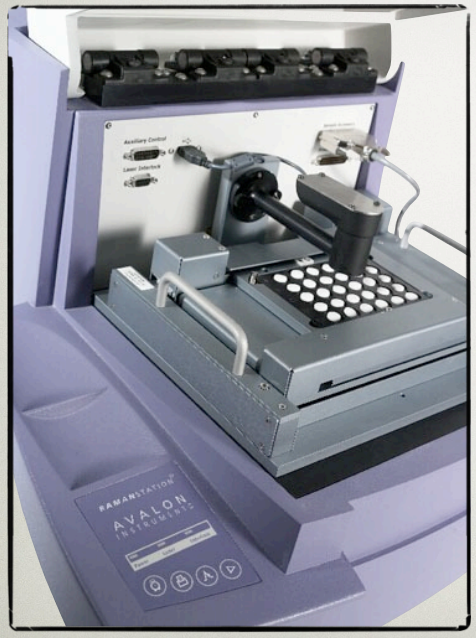
A surrogate for a real system can be useful for coarse optimization or training or exploring what-if scenarios

www.evolved-analytics.com



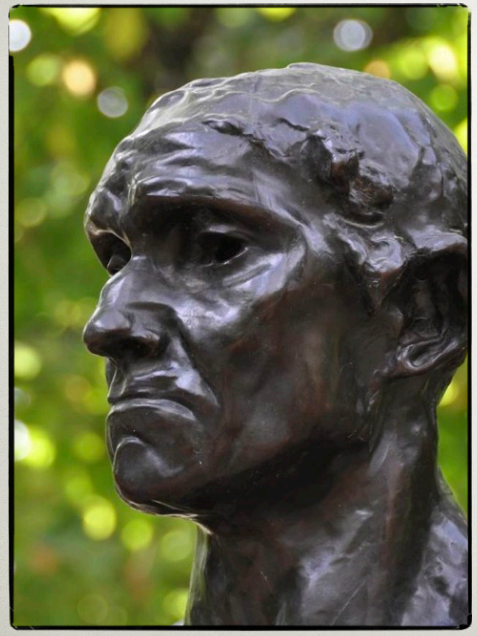
Risk Management

Knowing when the the system has changed or the model SHOULDN'T be trusted can be very valuable.



Guidance

Where should we look for better solutions?
How can we drive uncertainty out of the models?
What are the key factors?



Insight & Understanding

The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka!' (I found it!) but 'That's funny ...'

— Isaac Asimov (1920 - 1992)

The Human Factor

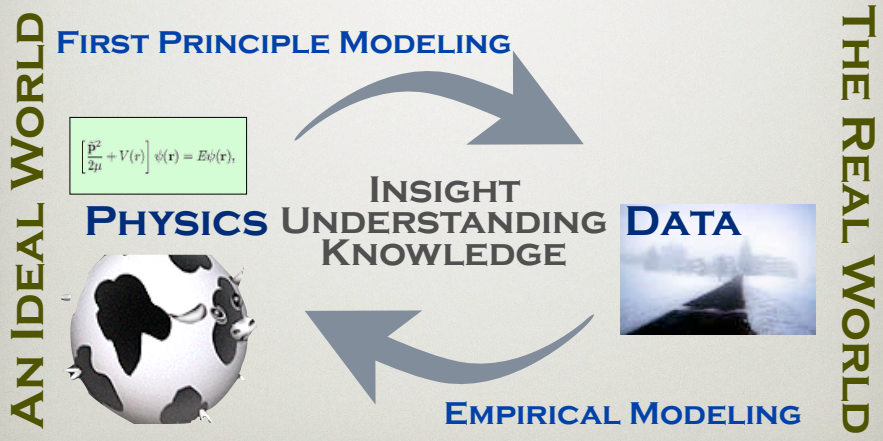
If you don't know where you are going any path will get you there

Data is never to be trusted

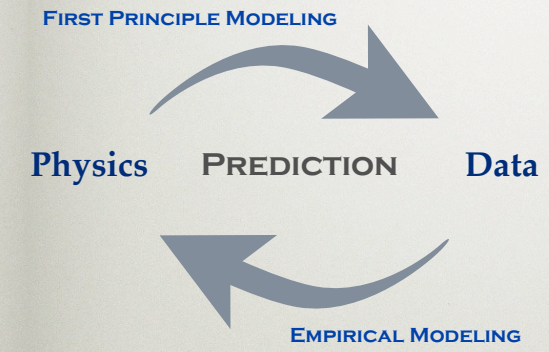
- Modeling Objective
- Data Characteristics
- Context & *a priori* Knowledge

Context-free analysis leads to confidently wrong conclusions

Why Modeling?

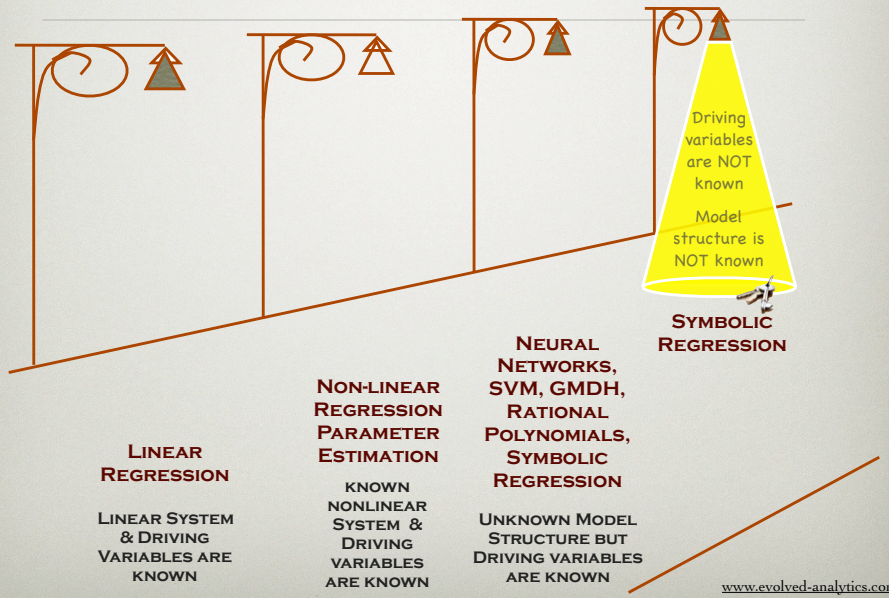


Why Modeling?



- Insight
- Actionable Data
- Direction
- Optimization
- Design

Modeling Options



What is Symbolic Regression?

- Must discover BOTH the model structure AND the embedded coefficients
- This is a MUCH more difficult problem than conventional regression or neural network modeling — the search space is infinite & an infinite number of models will fit the data
- The Basic Problem:
  - How do we efficiently search for models
  - How do we know when we are done?

Agenda

- Modeling & Motivations
- Conventional GP Overview & Limitations
- The ParetoGP (Multi-Objective) Perspective
- Ensembles & Trustable Models
- Trustability & Active DOE
- Outlier Identification
- Case Studies

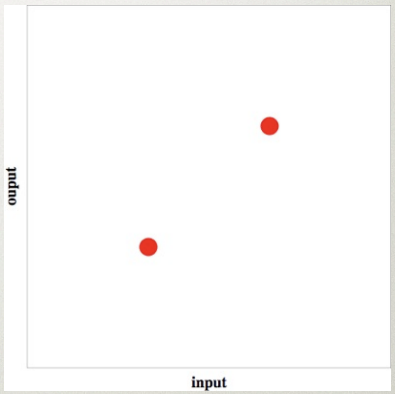
Tutorial Goals

- Understanding of the **modeling context** & motivation
- Understanding of GP-based **symbolic regression**
  - vanilla GP
  - Pareto GP
  - other algorithmic variants
  - common issues, problems, good practices and rules-of-thumb
- Foundations and implications of **ensembles & trustable models**
- Make it tangible
  - **case studies** using industrial data
- **Awareness** that GP-based symbolic regression (ParetoGP) is something truly special

What is Truth?

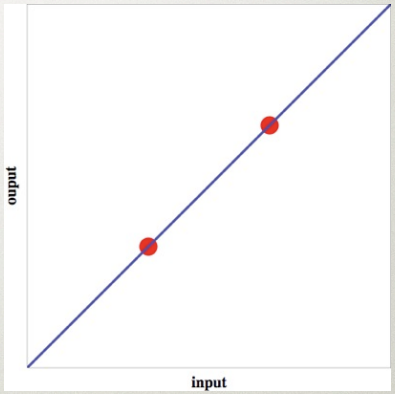
What is truth?

We only know the data



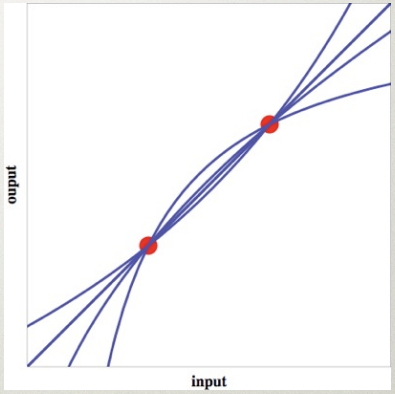
What is truth?

Maybe the linear model is correct?



What is truth?

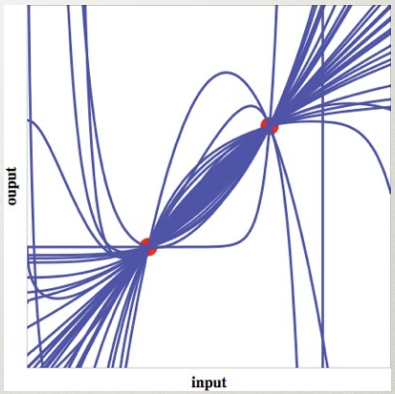
Perhaps we need to be more open-minded?



What is truth?

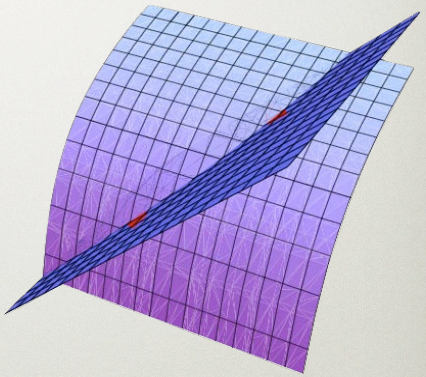
Each of these models fits the data EXACTLY

We need more information to make an intelligent model selection



Confounding factors

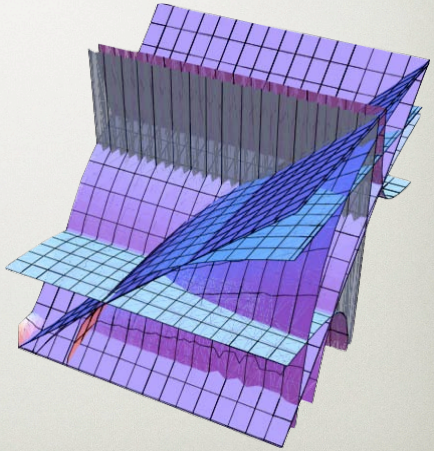
With many variables and limited data, determining which inputs are truly drivers and which correlations are accidental is difficult



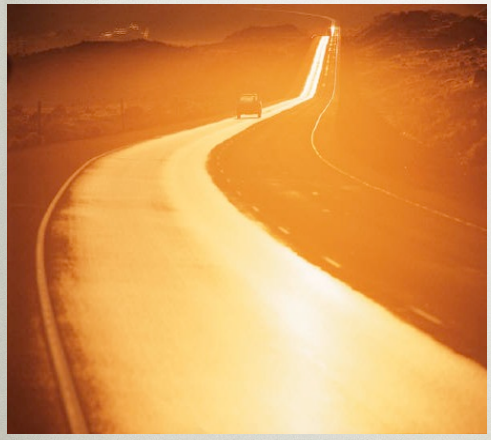
Confounding factors

Again, all models fit the data EXACTLY

How do we determine which variables are real?



Trust & Trepidation



- Using data-driven models is like driving fast with the windshield painted over and using only the rear-view mirrors
- This might work as long as the model is perfectly accurate and there aren't any curves
- Are you feeling lucky?

Key Point

All of the previous response surfaces were generated via symbolic regression.

The only constraint is the supplied building blocks.

Symbolic Regression ⇒ **Hypothesis Generator**

Human limits of imagination & possibility are not imposed!

We can exploit this creativity to produce trustable data models.

What Are Good Models?

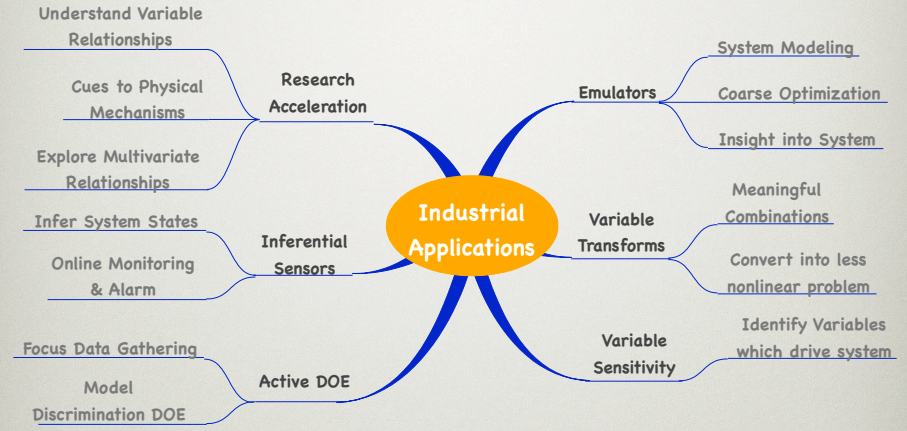
Evolved Analytics
Evolving Daily. Learning Always.

What are good Models?

- Credible
- Interpretable
- Cost-effective
- Extrapolate Well
- Robust
- Self-Assessing

Symbolic Regression can deliver against all of these objectives — which makes it very unique in its capabilities!

Reasons to Model



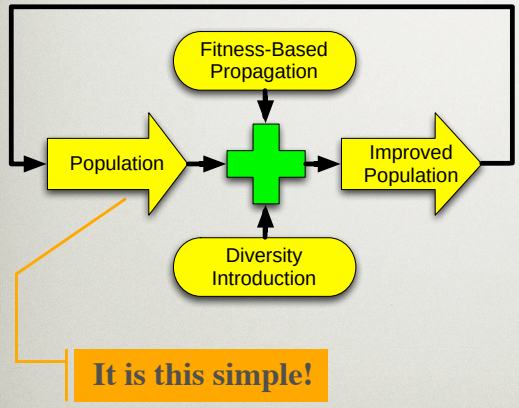
There are lots of good reasons to develop empirical models; however, not knowing WHICH reason(s) are applicable is a recipe for disaster

Vanilla GP

In the beginning, there was the void and Darkness.

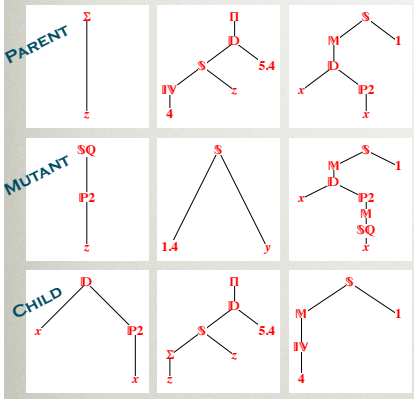
And Koza spoke ...

Evolutionary Computing



- EC Variants:
  - Genetic Algorithms (GA)
  - Evolutionary Strategies (ES)
  - Evolutionary Programming (EP)
  - Genetic Programming (GP)
  - Particle Swarm Optimization (PSO)
  - etc.
- Genetic Programming
  - Genome (genetic code) evolves
  - Phenotype (realization) judged for fitness
  - Goal is to evolve programs which solve problems
  - The search space is infinite!
  - Symbolic regression is one GP application
  - Many Variants of GP Now

Symbolic Regression via GP



Parent	z	$0.185107\left(\frac{1}{4}-z\right)$	$-1-\frac{1}{x}$
Mutant	$\sqrt{z^2}$	$1.37669-y$	-2
Child	$\frac{1}{x}$	0	$-\frac{5}{4}$

Summary

There are many possible variants of symbolic regression. These can either make the model search more efficient or they can make the search ineffective and slow.

Conventional Genetic Programming Problems

- Relatively Slow Discovery
  - Computational demands are intense
- Selection of "Quality" Solutions
  - Trade-off of Complexity vs. Performance
- Good-but-not-Great Solutions
  - Other nonlinear techniques outperform in raw performance
- Bloat
  - The "best" model explodes in complexity
- Stagnation
  - The search cannot innovate out of local optima.

Representations

- There are many GP variants which use different underlying genome structures
  - Tree-Based (the Koza original)
  - GE — Grammatical Evolution
  - Cartesian GP — graph based
  - GEP — Gene Expression Programming (closer to a GA representation)
  - LinearGP (evolve machine code directly)
  - Stack-based (e.g., PUSH)
- Functionally, they are pretty similar despite the passion of some of their proponents
- HOWEVER, implementation details can produce orders-of-magnitude differences in performance & model quality

The way we USED to do Data Modeling

1. Use "stacked analytic networks" to identify the driving variables
2. Use support vector regression (SVR) to identify the key data records
3. Apply GP (slowly) to the reduced data set
4. Painfully search through the developed models for simple and high-quality expressions or to extract insight.
5. Think that there HAD to be a better way

Pareto (Multi-Objective) GP

3+ orders-of-magnitude speed improvement opens lots of possibilities!

What do we REALLY want?

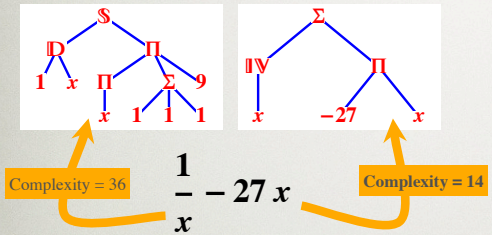
- Accurate
- Simple
- No spurious variables
- Robust
- Limited nonlinearity
- Dimensional consistency
- Smoothness
- etc.

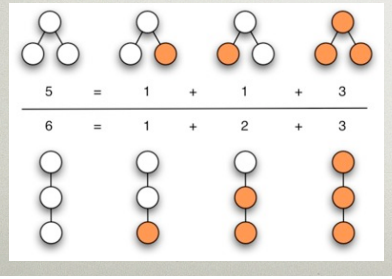
Accuracy, of course, is dominant

If we focus on model simplicity, then most of the other goals will be achieved as a side-effect

However, we have and do use more than just simplicity as the additional goal

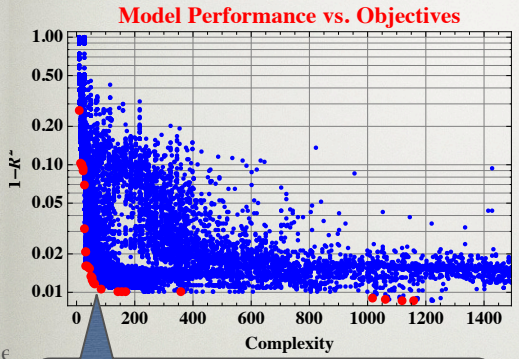
Model Complexity





- What is complexity?
  - # of nodes?
  - Tree depth?
  - Included functions?
  - Number of variables?
  - Combinations?
- Chosen function is sum of sum of node counts of genome
  - Provides more resolution at low end of complexity than simply using node count
  - Rewards flatter structures with fewer layers
  - Maarten Keijzer has renamed this "visitation length"

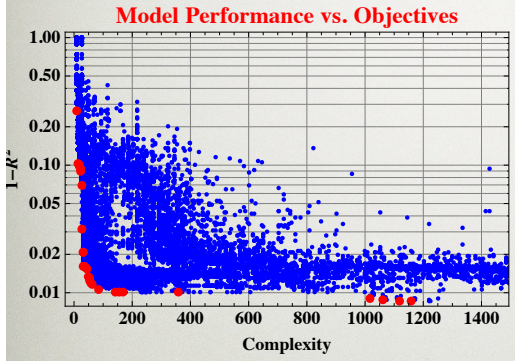
The Pareto Front



Models at the "knee" of the Pareto front are generally preferred since they have the best balance; however, we don't know the complexity vs. accuracy trade-off until AFTER we have built the models.

- Identifies trade-off surface between competing objectives
  - e.g., accuracy vs. complexity
- Pareto front solutions are the best "bang-for-the-buck"
- Unwarranted complexity is punished automatically
  - spurious variables
  - introns
  - unnecessary terms

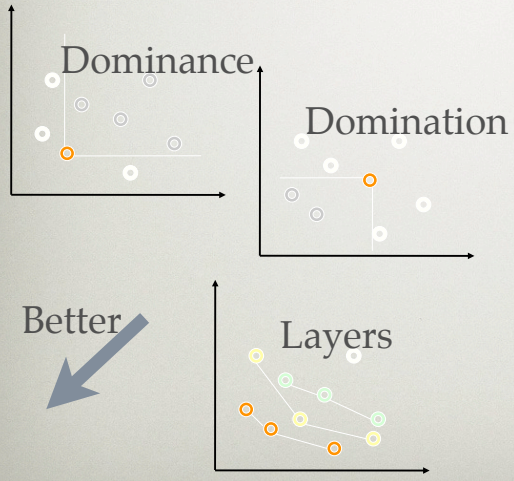
The Pareto Front



How do we exploit the Pareto front?

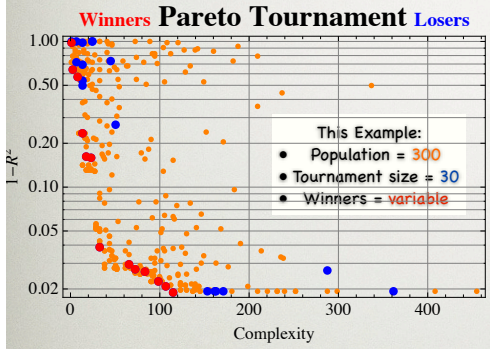
- Identifies trade-off surface between competing objectives
  - e.g., accuracy vs. complexity
- Pareto front solutions are the best "bang-for-the-buck"
- Unwarranted complexity is punished automatically
  - spurious variables
  - introns
  - unnecessary terms

Pareto Performance



- Multiple ways to characterize Pareto Performance
- Computational Issues
  - Brute force is MN^2
  - Can do $M\ N\ \log_{M-1}(N)$ or $M\ N\ \log_{M-2}(N)$ if clever
    - M = # of objectives
    - N = population size
- Global comparison is expensive & doesn't scale well
- Pareto tournaments are an alternative

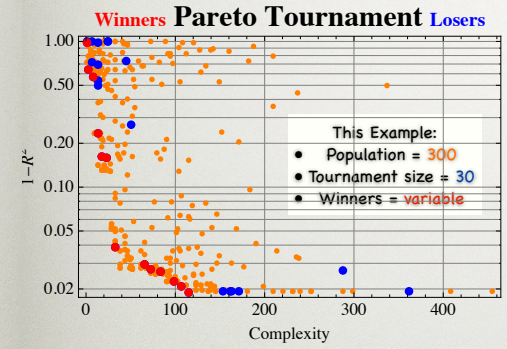
Pareto Tournaments



Pareto Tournaments
1. select a random pool of N models to compete
2. models on the multi-objective Pareto front getting breeding rights
3. repeat until sufficient parents are selected

- Tournaments are the standard in EC
  - simple
  - robust
  - tunable selectivity
  - scalable — NO global awareness
- Pareto tournaments have the same properties

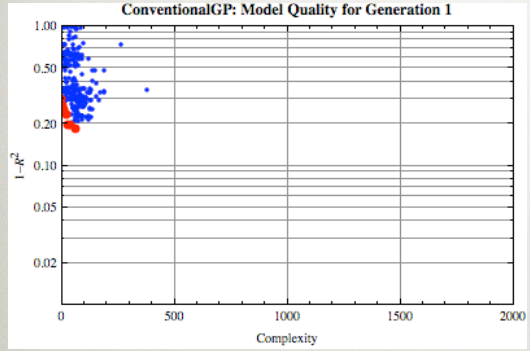
Pareto Tournaments



Pareto Tournaments
1. select a random pool of N models to compete
2. models on the multi-objective Pareto front getting breeding rights
3. repeat until sufficient parents are selected

- Identifying Pareto front of a local tourney is easier & faster than establishing the global pecking order
- The selection focus is automatically the knee of the Pareto front
- Note that this is NOT the commonly used approach of pairwise comparison — which is NOT tunable

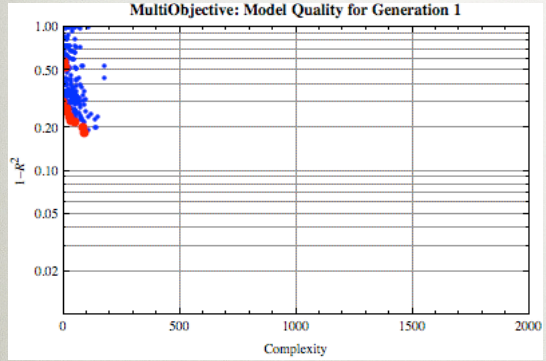
Conventional GP Evolution



- Inferential sensor data set
- Single objective tournaments based upon accuracy
- Upper bound on complexity
- Best model survives across generations
- "Best" models are WAY too complex

Lots of Effort & Minimal Results!

Pareto Tournament Impact



- Same data set
- Minor changes:
  - Pareto Tourney (30)
  - Pareto front survives across generations
- 50% more model search in same time
- Continual improvement
- MUCH better models

Less Effort & Better Results

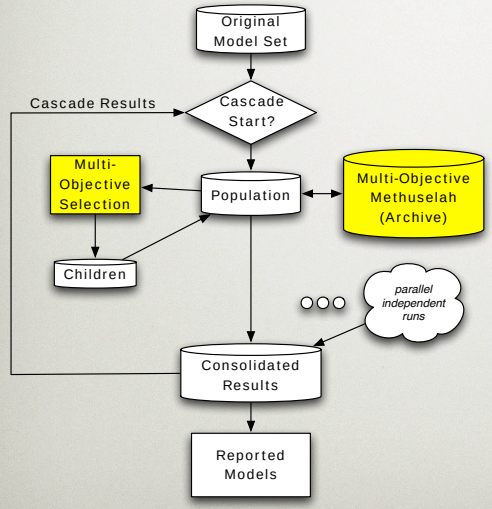
But that's not all...

Behind door #2, we have ...

Evolved Analytics

Evolving Daily: Learning Always.

The ParetoGP Essence



- Multi-Objective Selection
- Multi-Objective Archive
- Niching
- Continual New Genetics Influx

Good Ideas

Archives (Methuselahs)
- preserve good models

Multiple Objectives
- focus on the real goal
- primary/secondary objectives
- alternating objectives

Niching
- model protection during development

Scale & Translation Invariance
- e.g., absolute correlation
- calculating scale & translation is straightforward

Eliminate redundant models
- clean out duplicates

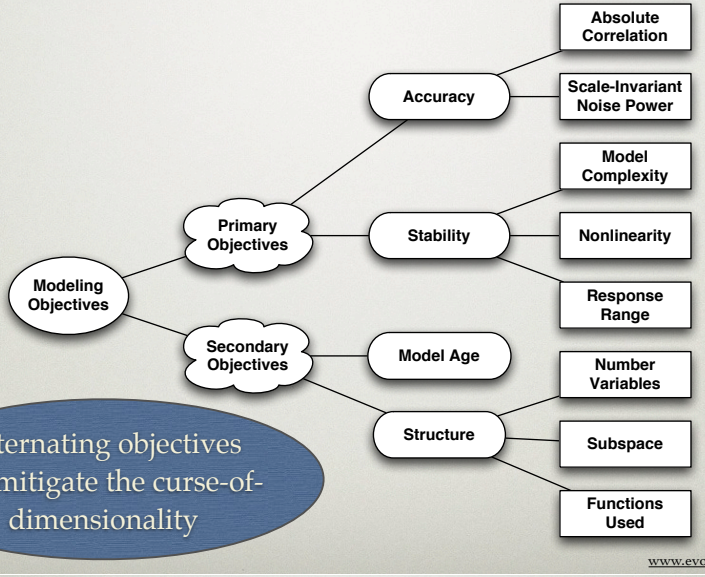
Rescale
- mapping variables to a common range can make them more interchangeable
- this will create bigger expressions when returning to the natural scaling

Lamarckian Evolution
- simplify & optimize models during evolution

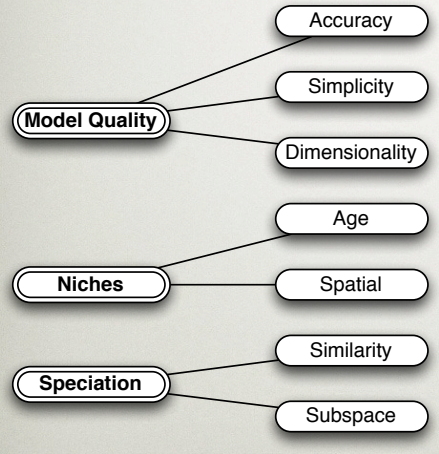
Interval Arithmetic
- focus on models which can be deployed safely

OrdinalGP & Essence
- dynamic environment
- data balancing

Multiple Objectives



Niching



- Niching is a foundation of diversity and continuous innovation
- There are many ways to implement niching
  - Multiple Objectives (ParetoGP)
  - HFC — Hierarchical Fair Competition
  - ALPS —Age-Layered Population Structure
  - Simple Geographies
  - Islands
  - etc.
- The key is to let models develop in a protected way

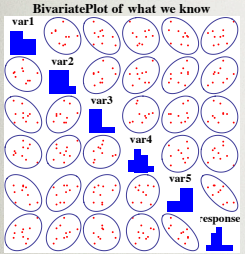
Interval Arithmetic & Robustness

```
rangeMap = {
   x1 → Interval[{1, 4}],
   x2 → Interval[{2, 10}]
  };
x1
──  /. rangeMap
x2

Interval[{1/10, 2}]

x1^2 - x1 /. rangeMap

Interval[{-3, 15}]
```

- Rapid test for model stability
  - two-orders of magnitude faster than nonlinear search
  - assumes hypercube of parameter ranges
- Test is conservative — rejects well behaved models
  - restricts model structure
  - can be implemented during model evolution for rapid pathology rejection

Implications

Rapid Modeling

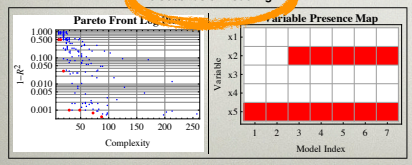


The efficiency can be remarkable!

What we Know is NOT that $y = \left(\frac{x2}{x5+1} + 2\right)^2$

record #	var1	var2	var3	var4	var5	response
1	0.329258	8.11465	7.12105	1.03291	3.40657	14.757
2	4.84938	4.66287	0.0189522	5.90275	2.36361	11.4668
3	7.93757	0.973329	0.982196	3.88811	7.52169	4.46992
4	0.324429	9.32384	3.88377	7.67598	0.162064	100.471
5	7.82642	6.71559	2.22499	1.36682	3.29299	12.7043
6	5.79551	5.14337	5.42646	8.24253	4.30972	8.813
7	8.373	5.68092	3.48945	5.25569	6.61319	7.54158
8	1.30168	7.21797	2.95179	7.11198	5.94553	9.23689
9	1.1921	1.87904	5.53493	7.25667	1.84713	7.07547
10	2.04239	3.23745	4.26614	6.1096	6.89175	5.80922

15 Seconds of Modeling



Model Selection Table

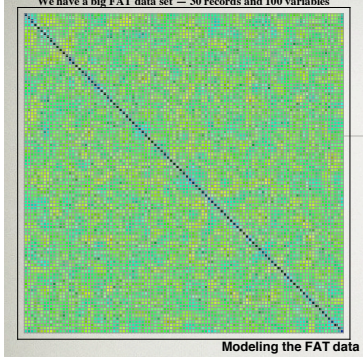
	Complexity	$1-R^2$	Vars	Function
1	11.000	0.513	x_5	$\frac{11.791}{x_5} + 5.977$
2	15.000	0.512	x_5	$15.104 + -2.808\sqrt{x_5}$
3	19.000	0.031	x_2 x_5	$\frac{3.846\,x_2}{x_5} + 4.198$
4	29.000	0.001	x_2 x_5	$\frac{4.741\,(x_2 \cdot 1)}{x_5} + 4.233$
5	48.000	9.646×10^{-4}	x_2 x_5	$\frac{4.994\left(x_2 - \sqrt[8]{x_2}\right)}{x_5} + 4.246$
6	72.000	7.671×10^{-4}	x_2 x_5	$-4.926\left(\frac{\sqrt[8]{\frac{x_2}{x_5}}}{x_5} - \frac{x_2}{x_5}\right) + 4.096$
7	88.000	5.427×10^{-4}	x_2 x_5	$5.091\left(\frac{x_2}{x_5} - \frac{\sqrt[8]{\frac{x_2 \cdot x_5}{x_2}}}{x_5}\right) + 4.105$

Evolved Analytics
Evolving Daily. Learning Always.

www.evolved-analytics.com

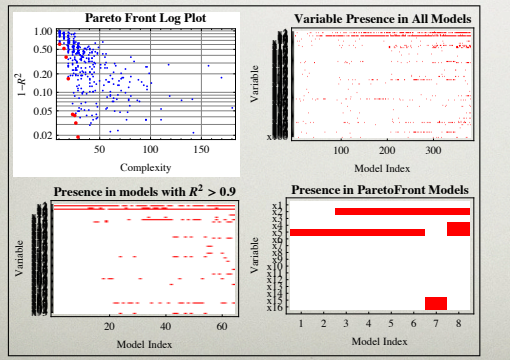
Fat Data Sets

We have a big FAT data set — 30 records and 100 variables



One minute of CPU Time

Modeling the FAT data array



Summary from the initial fat array SymbolicRegressio

variables having >10% presence in models with R^2>0.8

	# Models	% of Models	Variable	Meaning
1	113	91.1	x5	var5
2	95	76.6	x2	var2
3	27	21.8	x14	var14
4	24	19.4	x86	var86
5	24	19.4	x46	var46
6	23	18.5	x16	var16
7	22	17.7	x95	var95
8	16	12.9	x4	var4
9	16	12.9	x30	var30
10	15	12.1	x15	var15
11	13	10.5	x29	var29

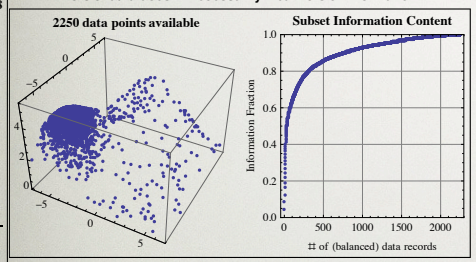
Model Selection Report

	Complexity	$1-R^2$	Function
1	11	0.605	$8.498 + \frac{8.448}{\text{var5}}$
2	15	0.516	$56.185 + -19.686\sqrt{\text{var5}}$
3	17	0.368	$0.981 + 0.100\,\text{var15}\,\text{var16}\,\text{var2}$
4	19	0.164	$7.318 + \frac{1.825\,\text{var2}}{\text{var5}}$
5	23	0.044	$-0.509 + \frac{4.430\,\text{var2}}{\sqrt{\text{var5}}}$
6	26	0.043	$1.901 + \frac{0.602\,\text{var2}\,\text{var4}}{\sqrt{\text{var5}}}$
7	27	0.032	$2.166 + \frac{0.482\,\text{var2}^2}{\sqrt{\text{var5}}}$
8	29	0.019	$6.837 + \frac{0.003\,\text{var2}^4}{\text{var5}}$

www.evolved-analytics.com

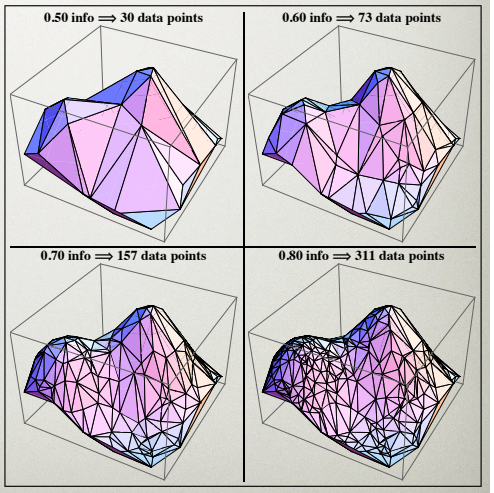
Redundant Data

Lots of data doesn't necessarily mean lots of information



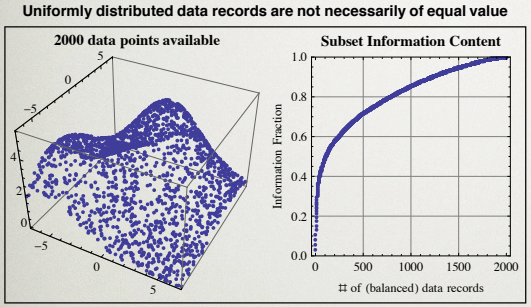
Redundant data is common in closed loop systems like industrial processes

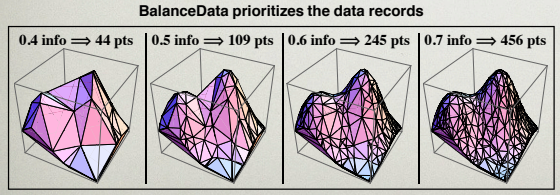
Balanced Data ⟹ Faster & Better Models



www.evolved-analytics.com

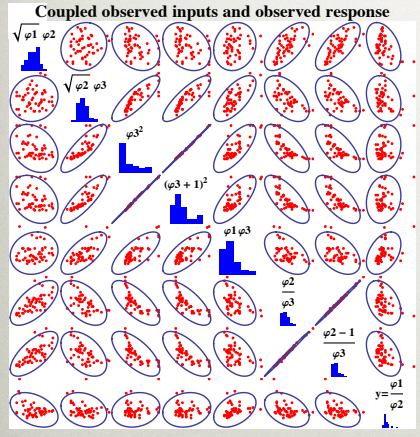
Large Data Set Strategies





- More CPU Time
- Balanced Data Subset
- OrdinalGP (randomly changing subsets)
- ESSENCE (incrementally adding balanced data)

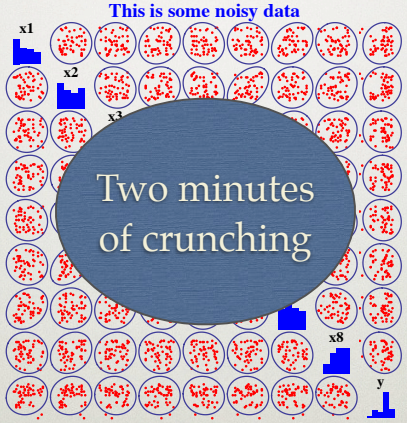
Correlated Data



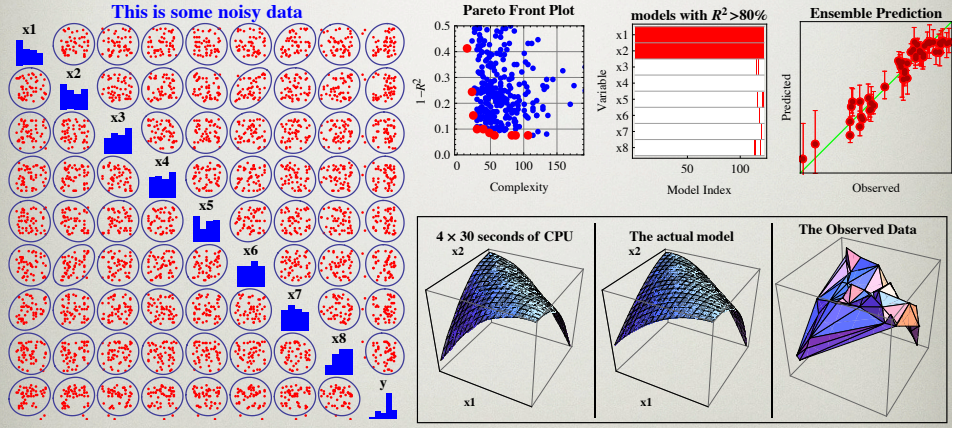
15 Seconds of Modeling

	Complexity	$1-R^2$	Vars	Function
1	11.000	0.564	x_7	$\frac{0.334}{x_7} + 0.633$
2	19.000	0.167	x_2 x_5	$-0.482 + \frac{0.620\,x_5}{x_2}$
3	27.000	0.089	x_2 x_5	$\frac{0.106\,x_5^2}{x_2^2} + 0.294$
4	29.000	0.000	x_3 x_5 x_6	$\frac{1.000\,x_5}{x_3\,x_6}$

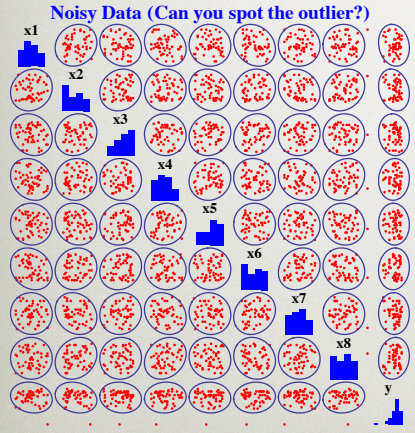
Noisy Data

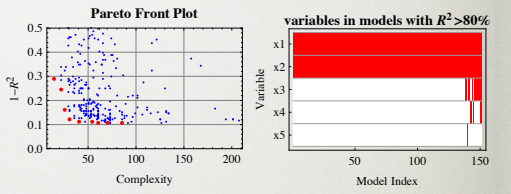


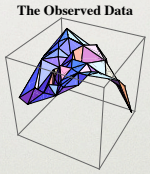
Noisy Data

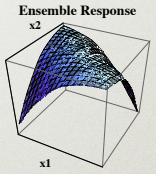


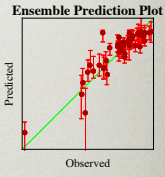
Outlier Detection







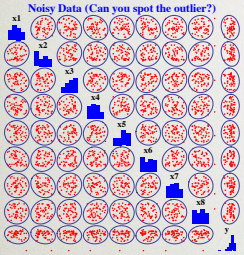




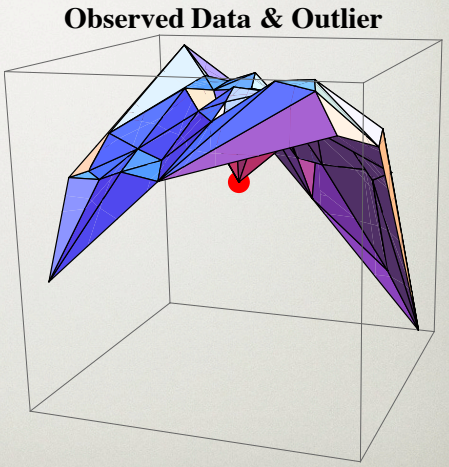
Data Outlier Analysis Table

oulier rank	index	strangeness	outlier distance	data record	observed response
1	51	1.96	2.97	{0, 0, 0, 0, 0, 0, 0, 0}	−2.02

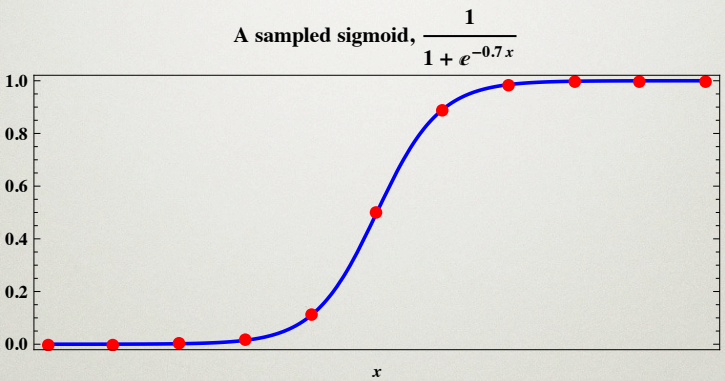
Outlier Detection



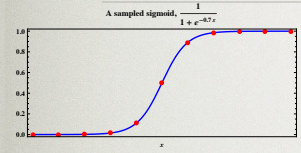
Outlier detection HAS to be model-based!

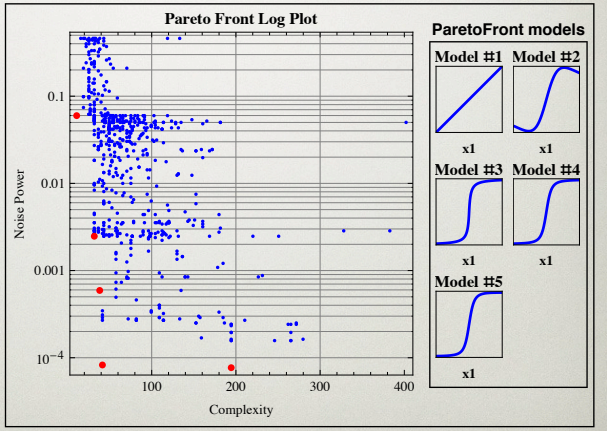


Trustable Models

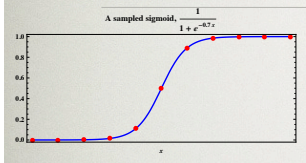


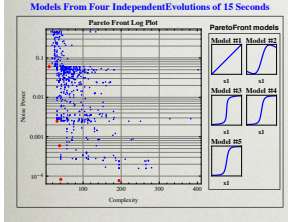
Trustable Models

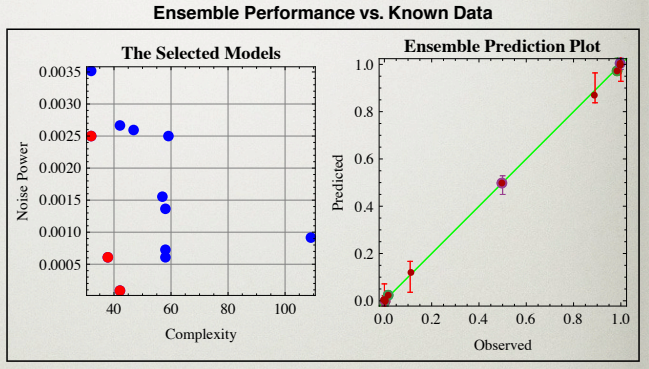




Trustable Models

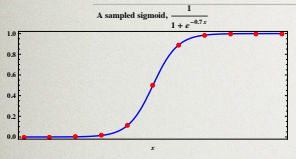


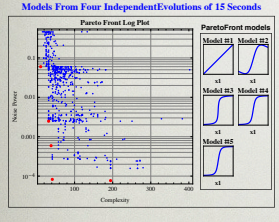


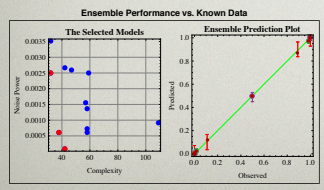


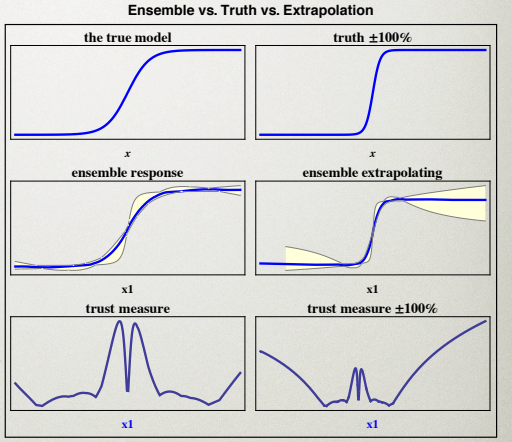
www.evolved-analytics.com

Trustable Models



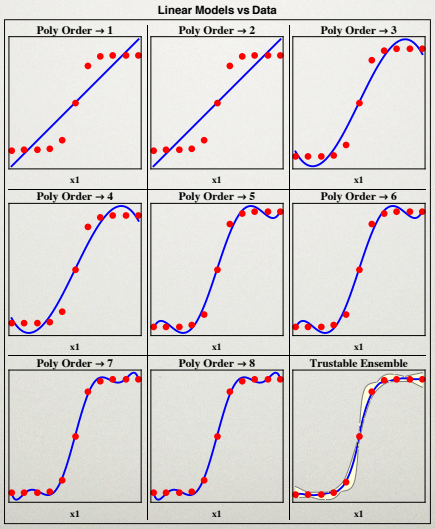






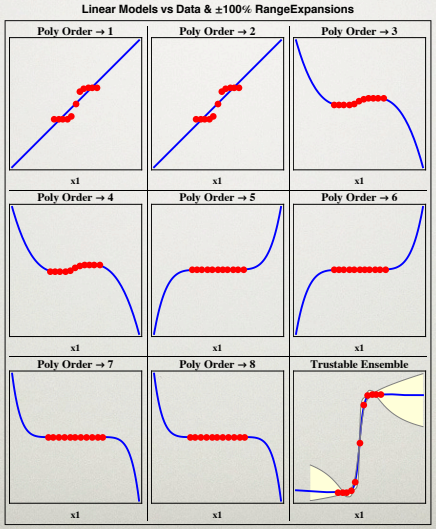
www.evolved-analytics.com

Extrapolation



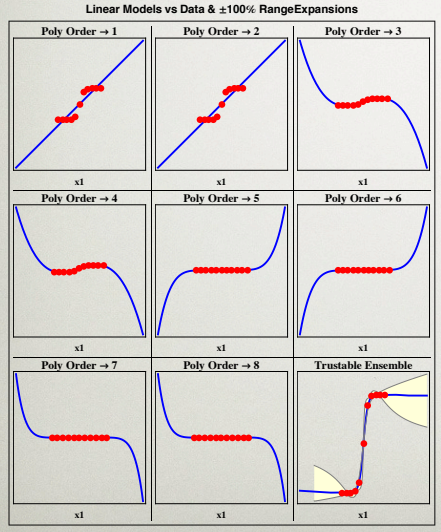
www.evolved-analytics.com

Extrapolation



www.evolved-analytics.com

Extrapolation



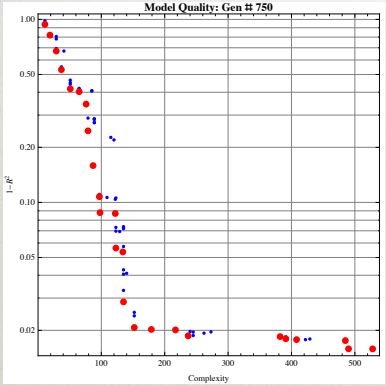
Nature does not know it is supposed to be a low-order polynomial!

Ensembles

Many working together can achieve great things

Evolved Analytics
Evolving Daily. Learning Always.

How do we choose the model?



- Many models are developed during the evolutionary search
- These models have different structures despite having comparable performance
- Different evolutions will discover different model forms

How do we choose the model?

We don't!

How do we choose THE model?

- Instead, we exploit the DIVERSITY of structural forms via model ensembles

Recall that the models are constrained by the data and NOT by physics or preconceived assumptions!

Other Things to Consider

- model forms for INSIGHT ...
- variable presence for DRIVERS
- variable combinations (metavariables)
- Pareto front shape (modeling potential & difficulty)

What is an ensemble?

- A collection of models which are ...
  - **diverse** and
  - "**good enough**" complexity and accuracy

Why is an ensemble special?

- The constituent models ...
  - **AGREE** where constrained by the observed data (otherwise, they would not be good models)
  - **DIVERGE** where they are not constrained by experience (otherwise, they would not be diverse)
- Thus, we have a high-quality prediction **AND** a trust metric to accompany that prediction!

TRUSTABLE MODELS!!

- No more "Trust me"
  - This is a classic issue for use of data-derived models
- This is VERY unique and made possible by the diverse structures coming out of ParetoGP
- Knowing when NOT to trust the prediction can be very valuable
- We can also exploit ensembles for adaptive data collection
- Identified systems also tend to extrapolate well

ENSEMBLES ARE/ARE NOT ...

ARE NOT ...

- collections of weak-learners
- diverse due to using different training data
- collections of locally-good models

ARE ....

- extrapolate reasonably
- white-box
- robust predictors (use median average for prediction)
- diverse due to uncorrelated prediction residuals

DIVERSE MODELS

Diversity Aspects

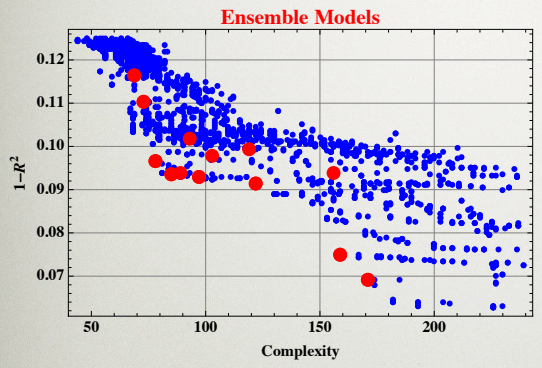
- model structure
- constituent variables
- (lack of) error correlation
  - observed data
- prediction disagreement
  - synthetic data

A Definition Algorithm

- build covariance matrix
- select most uncorrelated pair
- delete all models not uncorrelated (within threshold) to pair
- repeat until no models left
- Divide-and-conquer for large numbers of models to avoid scaling problems

ALTERNATE: choose a reference and work from there

ASSEMBLING THE ENSEMBLE

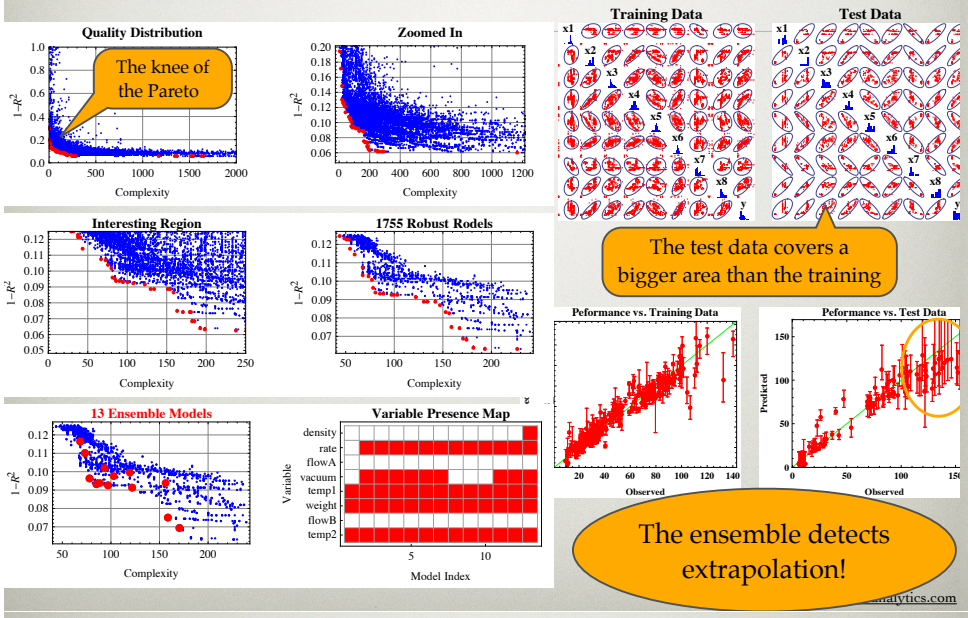


The definition of "uncorrelated" needs to be relaxed from the conventional statistics criteria.

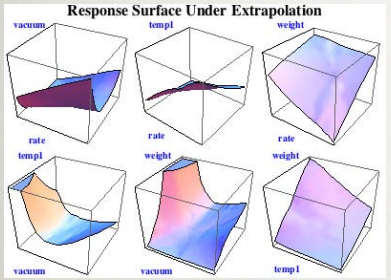
- A strategy that seems to work is:
  - uncorrelated models from model population
  - uncorrelated models from Pareto front
  - "most typical" model from Pareto front

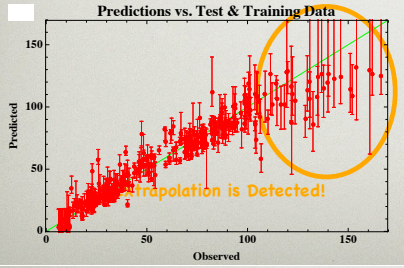
Making it Tangible

(An Inferential Sensor Example)

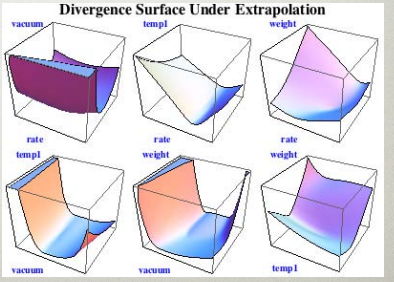


Evaluating the Ensemble





- Response:
  - median, median average or mean
- Consensus:
  - extrema range, inter-quantile range or standard deviation



Summary: Trustable Models!

- Ensembles of diverse models address a fundamental problem with empirical models
- We have outlined a straight-forward approach to assembling those models
- We can now use models with an awareness of the prediction risk
- The windshield has been cleaned!

Foundations Summary

ParetoGP

- Appropriate complexity models
- Automatic variable selection
- Able to handle "fat" data sets
- continual innovation

Ensembles

- More aware models (no need for data partitioning during model development?)
- Trustable predictions
- Interesting models for expert insight — "hypothesis selection"
- Ability to implement Active DOE

Summary

- A smart ParetoGP implementation is >1,000x faster than vanilla GP and produces better models
- That speed improvement CHANGES what is possible and redefines best practices!

Active Design-of-Experiments

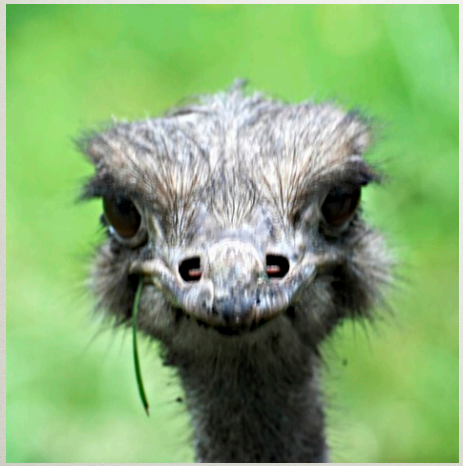
Accelerating Learning via FOcused Data Collection

Evolved Analytics
Evolving Daily. Learning Always.

The Curse of Dimensionality

- Adding more variables makes the problem difficulty explode
- It is often difficult or impossible to initially collect enough of the right data

What to Do???



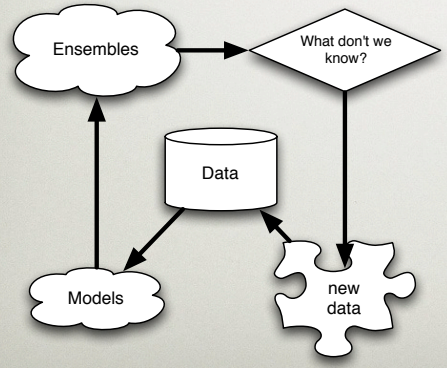
Target data collection to:

- identify the true driving variables
- reject spurious models
- increase model fidelity (global accuracy)

Guiding Principle: We want a diverse collection of accurate but simple models

HOW to get there?

Adaptive Data Collection



- ParetoGP lets us:
  - explore diverse model structures and
  - different variable combinations
- Diverse model ensembles let us:
  - identify regions of lower model fidelity (the models are presumably accurate in the sense that they fit the observed data)

Motivations & Scenarios: Operations Optimization

- Operating plants "walk away" from their optimal set-points
- Many available variables with unknown interactions
- Most historical data is essentially identical
- New data is precious & risky

Data collection is severely constrained

Motivations & Scenarios: High-Throughput Research

- LOTS of options being explored
- Want to quickly identify key factors & relationships
- Want to search for global rather than local optimum

Many variables of unknown significance

Motivations & Scenarios: Design & Emulation

- The REAL system takes a long time to generate data
- Want an emulator for coarse optimization & human insight
- Want better products and faster time-to-market

Limited ability to collect data

Motivations & Scenarios: Bioreactor Startup

- Nonlinear dynamics
- Lots of control settings
- Full-scale reactors VERY different from lab-scale used for development
- New yeast strain likely behaves differently than previous

Time is Money

Motivations & Scenarios: Maximizing Model Validity

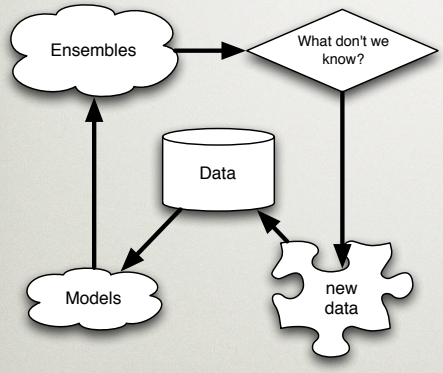
- Knowing that the model is not a good predictor is not enough
- We want the model to be accurate over the entire operating range
- Example: modeling price-volume elasticity

Need to maximize the model fidelity & range

Motivations & Scenarios: Common Themes

- Data is precious
- Many variables of unknown significance
- Data collection is difficult and/or expensive
- BOTH variable selection AND a good model is desired
- Delivery time pressure (Aren't you done YET?)

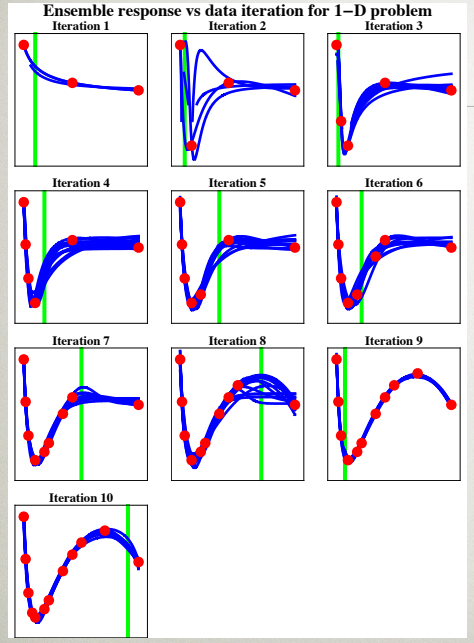
Adaptive Data Collection Sequence



1. build models from the available data
2. assemble ensembles from those models
3. identify the locations of model uncertainty and potential optima
4. collect new data
5. repeat

What we HAVEN'T done

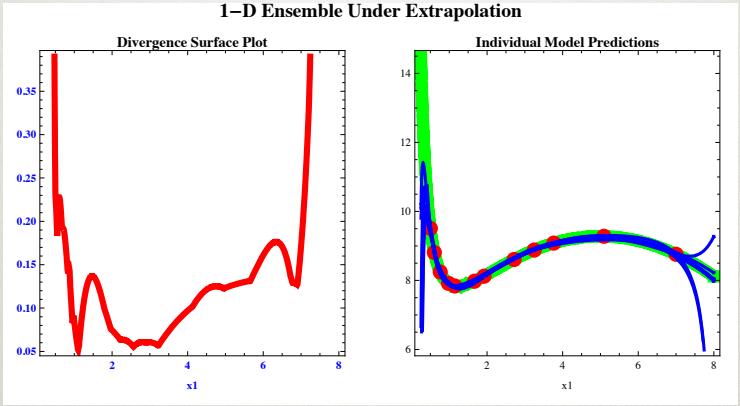
- assumed which variables are important
- harshly constrained the number of variables
- assumed variables are independent & uncorrelated
- assumed a model form
- collected data blindly



A 1-D Adaptive DOE

- Using only the divergence in the constituent models for guidance, the adaptive DOE process was able to collect data to produce a high-quality predictor

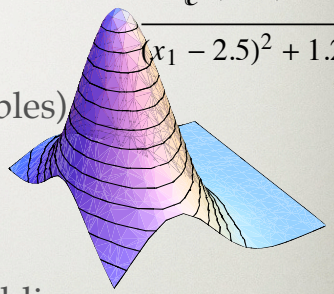
Ensemble under extrapolation



Note that the ensemble models agree where there is data and diverge where not constrained by data

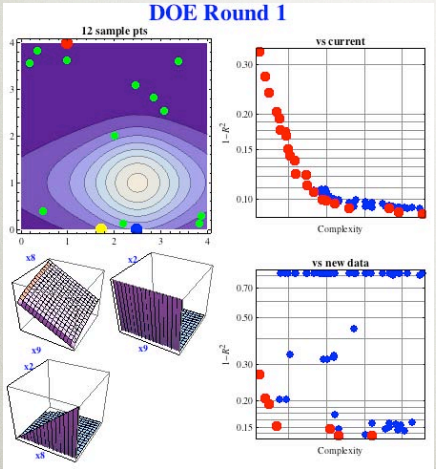
Making the problem harder

$$\frac{e^{-(x_2-1)^2}}{(x_1-2.5)^2+1.2}$$



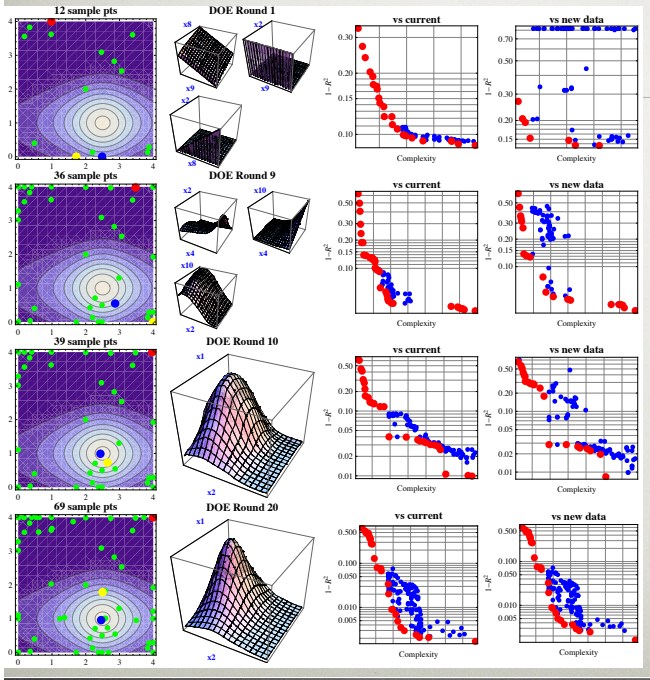
- We have 10 dimensions (variables)
- Only two variables are real
- We start with 12 data points
- After each round of model building, we collect data at the predicted maximum, predicted minimum and the point of maximum ensemble uncertainty

20 Rounds of Active DOE



- Only the data projection into the true variables is shown
- All data inputs are 10-D
- Note that the early stages feature high accuracy but low fidelity
- Response surface only shown for variables which are in >50% of models in the "interesting region"

Fidelity is not the same as accuracy!



Selected Rounds

- In the early stages the new data tends to shatter many candidate models
- In later stages the incremental is less destructive to candidate models
- Spurious variables may appear dominant in the early going but additional data leads to their rejection

What have we seen?

- Exploiting the DIVERSITY of candidate models generated by ParetoGP,
- we can use a model ensemble prediction divergence to identify regions of UNCERTAINTY
- and target data collection at reducing that uncertainty.
- This results in HIGH-FIDELITY and TRUSTABLE models and the REJECTION of spurious inputs.

Case Studies

The proof is in the pudding

The selected data consists of 109 attributes from 132 countries

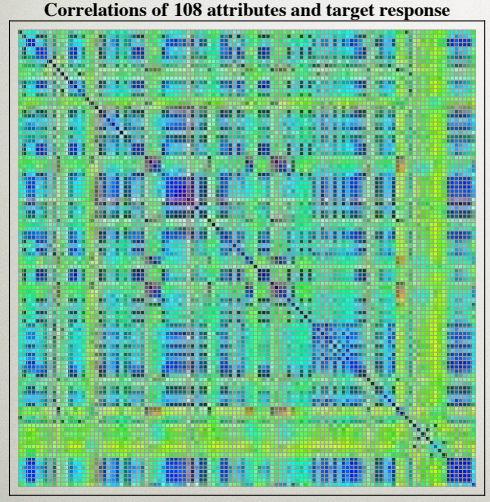
CountryIndex	AdultPopulation	Airports	AMRadioStations
AnnualBirths	AnnualDeaths	ArableLandArea	ArableLandFraction
Area	BirthRateFraction	BoundaryLength	CallingCode
CellularPhones	ChildPopulation	CoastlineLength	CropsLandArea
CropsLandFraction	DeathRateFraction	EconomicAid	ElderlyPopulation
ElectricityConsumption	ElectricityExports	ElectricityImports	ElectricityProduction
ExchangeRate	ExportValue	ExternalDebt	FemaleAdultPopulation
FemaleChildPopulation	FemaleElderlyPopulation	FemaleInfantMortalityFraction	FemaleLifeExpectancy
FemaleLiteracyFraction	FemaleMedianAge	FemalePopulation	FixedInvestment
FMRadioStations	GovernmentConsumption	GovernmentExpenditures	GovernmentReceipts
GovernmentSurplus	GrossInvestment	HighestElevation	HouseholdConsumption
ImportValue	InfantMortalityFraction	InflationRate	InternetHosts
InternetUsers	InventoryChange	IrrigatedLandArea	IrrigatedLandFraction
LaborForce	LandArea	LifeExpectancy	LiteracyFraction
LowestElevation	MaleAdultPopulation	MaleChildPopulation	MaleElderlyPopulation
MaleInfantMortalityFraction	MaleLifeExpectancy	MaleLiteracyFraction	MaleMedianAge
MalePopulation	MedianAge	MigrationRateFraction	MilitaryAgeMales
MilitaryExpenditureFraction	MilitaryFitMales	NaturalGasConsumption	NaturalGasExports
NaturalGasImports	NaturalGasProduction	NaturalGasReserves	OilConsumption
OilExports	OilImports	OilProduction	PavedAirports
PavedRoadLength	PhoneLines	Population	PopulationGrowth
PriceIndex	RadioStations	RoadLength	ShortWaveRadioStations
TelevisionStations	TotalConsumption	TotalFertilityRate	UnemploymentFraction
UNNumber	WaterArea	ExpenditureFractions- ExportValue	ExpenditureFractions- FixedInvestment
ExpenditureFractions- GovernmentConsumption	ExpenditureFractions- GrossInvestment	ExpenditureFractions- HouseholdConsumption	ExpenditureFractions- ImportValue
ExpenditureFractions- InventoryChange	ExpenditureFractions- TotalConsumption	PavedAirportLengths- 3000To5000Feet	PavedAirportLengths- 5000To8000Feet
PavedAirportLengths- 8000To10000Feet	PavedAirportLengths- Over10000Feet	PavedAirportLengths- Total	PavedAirportLengths- Under3000Feet
GDPPerCapita			

GDP Per Capita

- Source: *Mathematica* CountryData
- Looking at GDP Per Capita
- Using numeric input and output

GDP Per Capita

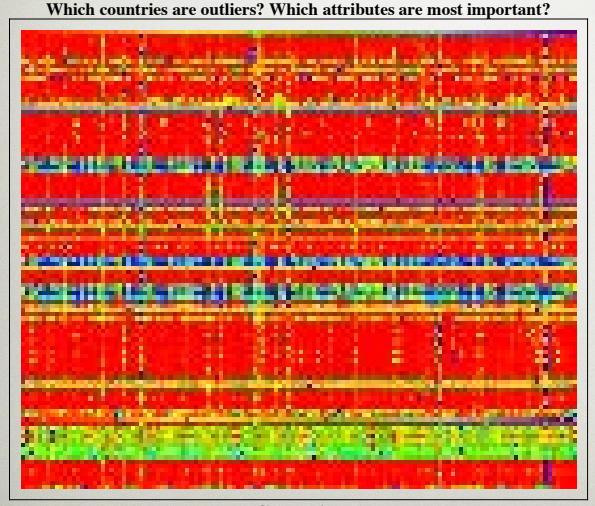
Correlations of 108 attributes and target response



- Correlated inputs!

GDP Per Capita

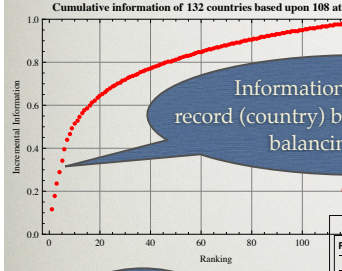
Which countries are outliers? Which attributes are most important?



- FAT data —nearly as many inputs as data records
- Highly correlated inputs
- Which variables are important?
- Which variables are sufficient?
- Which countries are most typical?
- Which countries are most atypical?
- A linear modeling approach is not viable

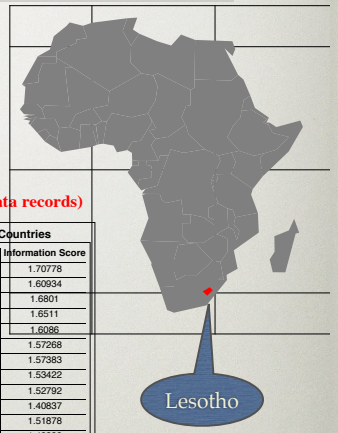
GDP Per Capita
Information COntent

Cumulative information of 132 countries based upon 108 attributes



Information content of each record (country) based upon SMITS data balancing algorithm

Each added country provides the maximum information content to the previously selected data records



The top & bottom ranked countries (data records)

Top Countries

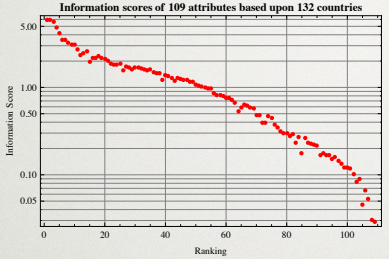
Rank	Attribute	Information Score
1	UnitedStates	116.833
2	China	65.0173
3	Lesotho	57.4641
4	India	56.4142
5	Russia	54.6371
6	Canada	54.0747
7	Japan	42.9039
8	UnitedKingdom	29.7679
9	Brazil	24.2928
10	Turkmenistan	23.0736
11	Italy	16.2112
12	Iran	17.263
13	Indonesia	15.077
14	SaudiArabia	13.9068
15	Australia	12.4714
16	France	11.9526
17	Peru	11.5849
18	Qatar	11.6882
19	Germany	10.9553
20	Nigeria	9.8855

Bottom Countries

Rank	Attribute	Information Score
1	Honduras	1.70778
2	CostaRica	1.60934
3	Nicaragua	1.6801
4	Namibia	1.6511
5	Panama	1.6086
6	Lithuania	1.57268
7	Barbados	1.57383
8	Bahamas	1.53422
9	Aruba	1.52792
10	Tunisia	1.40837
11	ElSalvador	1.51878
12	Dominica	1.49238
13	Slovenia	1.36972
14	Slovakia	1.31135
15	Jamaica	1.2543
16	Cameroon	1.24504
17	SaintLucia	1.24261
18	Belize	1.19521
19	Malta	1.04347
20	Fiji	0.916755

GDP Per Capita
Attribute Information Content

Information scores of 109 attributes based upon 132 countries



"The top & bottom ranked attributes"

Top Attributes for Country Discrimination

Rank	Attribute	symbol	Information Score
1	ExchangeRate	x25	5.917010432358875
2	GovernmentSurplus	x41	5.861501158096805
3	FemaleLiteracyFraction	x33	5.676381766048056
4	GDPPerCapita	y	4.851577865999762
5	CallingCode	x12	4.216983042161471
6	CountryIndex	x1	3.510115825552460 5
7	LifeExpectancy	x55	3.507746554253089
8	ArableLandFraction	x8	3.27015331765554
9	HighestElevation	x43	3.120580355648746
10	UNNumber	x93	3.067479622996432
11	CropsLandFraction	x17	2.709289493488639
12	FemaleMedianAge	x34	2.340805328238699
13	ExpenditureFractions- TotalConsumption	x102	2.474298354325172 4
14	BirthRateFraction	x10	2.591075587720015
15	MedianAge	x66	1.984850046084618 6
16	ExpenditureFractions- InventoryChange	x101	2.157947985238084 5
17	MilitaryExpenditureFraction	x69	2.182781412172800 6
18	ElectricityExports	x22	2.255899471910555 4
19	ExpenditureFractions- FixedInvestment	x96	2.155455789828333
20	UnemploymentFraction	x92	2.109238932955205 5

Bottom Attributes for Country Discrimination

Rank	Attribute	symbol	Information Score
90	ElectricityConsumption	x21	0.214968502532835 17
91	FixedInvestment	x36	0.166095682266068 27
92	OilConsumption	x76	0.176171153059166 06
93	GrossInvestment	x42	0.167092445853205 1
94	TotalConsumption	x90	0.169230206949292 22
95	MaleElderlyPopulation	x60	0.150175165464828
96	MilitaryFitMales	x70	0.159875871835683 86
97	PavedAirportLengths- 5000To8000Feet	x104	0.143324905664671 38
98	Population	x83	0.135284168376346 63
99	HouseholdConsumption	x44	0.122601363613522 9
100	MaleChildPopulation	x59	0.119944625509610 3
101	ElderlyPopulation	x20	0.117387255885551 35
102	PavedAirports	x80	0.102405819617770 87
103	MalePopulation	x65	0.083044929661966 36
104	FemalePopulation	x35	0.090137154512178 03
105	FemaleAdultPopulation	x28	0.045243978337267 8
106	AdultPopulation	x2	0.066764423260164 43
107	PavedAirportLengths- Total	x107	0.053245390606786 71
108	MilitaryAgeMales	x68	0.030618966600265 297
109	MaleAdultPopulation	x58	0.028652520738616 023

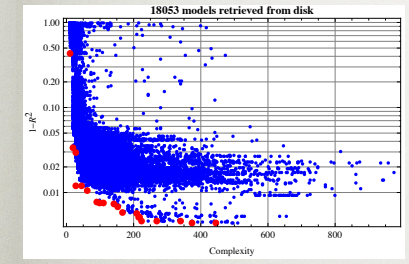
- Scores using SMITS algorithm
- Ranking of attributes for discriminating between countries

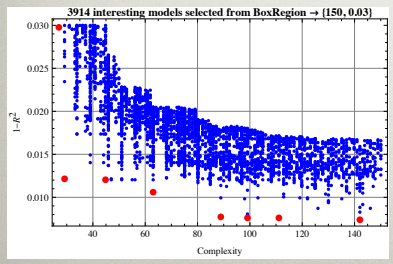
Summary: Information Content & Data Balancing

- We have a means to create a **balanced data** set from unbalanced data
- Balanced data avoids inappropriate weighting of data regions ⇒ better models AND faster modeling
- We can also detect (some) outliers during the data balancing
- The SMITS algorithm does not assume a data model (just an information metric)
- Identifying the dominant records can be insightful — these are the prototypes

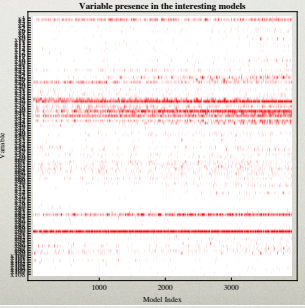
www.evolved-analytics.com

GDP Per Capita Modeling Results





- Lots of 10 minute runs
- Note the variable selection
- 1207 different variable combinations in developed models

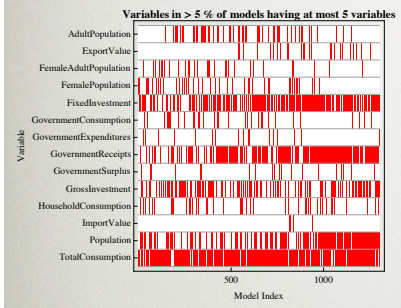


Variables present in >5% of models

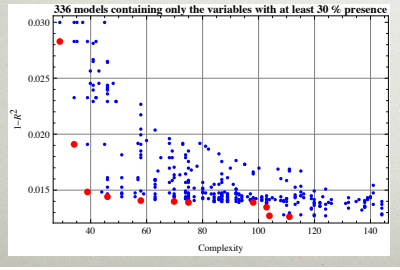
	# Models	% of Models	Variable	Meaning
1	3384	86.5	x90	TotalConsumption
2	2686	68.6	x36	FixedInvestment
3	2308	59.0	x40	GovernmentReceipts
4	1788	45.7	x83	Population
5	1670	42.7	x42	GrossInvestment
6	1095	28.0	x2	AdultPopulation
7	839	21.4	x35	FemalePopulation
8	828	21.2	x44	HouseholdConsumption
9	803	20.5	x28	FemaleAdultPopulation
10	625	16.0	x41	GovernmentSurplus
11	503	12.9	x38	GovernmentConsumption
12	499	12.7	x26	ExportValue
13	334	8.5	x45	ImportValue
14	243	6.2	x34	GovernmentExpenditures

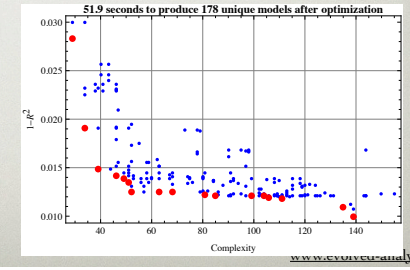
www.evolved-analytics.com

GDP Per Capita Variable Selection & Niching



- Despite the fact that we have a **fat array and correlated inputs**, we managed to identify and isolate driving variables
- **Diverse model structures** were uncovered that fit the data
- The discovered models feature **diversity** in numbers of models (dimensionality) and different variable combinations (subspaces)
- We can select models for use which have limited dimensionality or targeted variable combinations





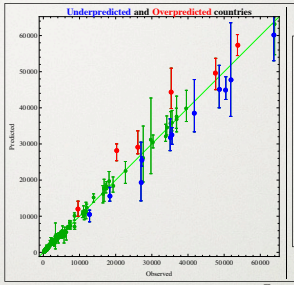
www.evolved-analytics.com

GDP Per Capita Outlier Detection

Outlier Detection HAS to be model-based.

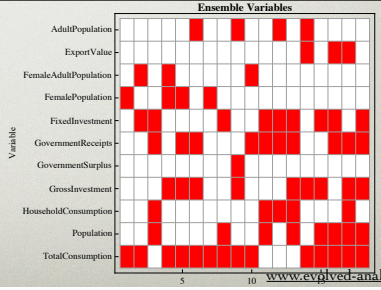
Models in the Selected Ensemble

	Complexity	1-R²	Function
1	27	0.030	$-190.723 + \frac{0.652\,\text{TotalConsumption}}{4668.020\,\text{FemalePopulation}}$
2	29	0.023	$-155.975 + \frac{0.315\,(\text{FixedInvestment}\cdot\text{TotalConsumption})}{\text{FemaleAdultPopulation}}$
3	34	0.017	$110.337 + \frac{0.015\,(\text{FixedInvestment}\cdot\text{GovernmentReceipts}\cdot\text{HouseholdConsumption})}{\text{Population}}$
4	34	0.028	$18.236 + \frac{0.493\,(\text{FemaleAdultPopulation}\cdot\text{GrossInvestment}\cdot\text{TotalConsumption})}{\text{FemalePopulation}}$
5	34	0.029	$370.370 + \frac{0.335\,(\text{GovernmentReceipts}\cdot\text{GrossInvestment}\cdot\text{TotalConsumption})}{\text{FemalePopulation}}$
6	34	0.030	$130.922 + \frac{0.472\,(\text{GovernmentReceipts}\cdot\text{GrossInvestment}\cdot\text{TotalConsumption})}{\text{AdultPopulation}}$
7	37	0.030	$-193.790 + \frac{1.304\,\text{TotalConsumption}}{10042.900\,\text{-}2\,\text{FemalePopulation}}$
8	37	0.030	$-93.049 + \frac{-0.954\,(\text{FixedInvestment}\cdot\text{TotalConsumption})}{0.271\,\text{Population}}$
9	39	0.023	$-93.434 + \frac{0.642\,(2\,\text{-GovernmentSurplus}\cdot\text{GrossInvestment}\cdot\text{TotalConsumption})}{\text{AdultPopulation}}$
10	41	0.025	$54.369 + \frac{0.166\,(\text{GovernmentReceipts}\cdot2\,\text{TotalConsumption})}{\text{FemaleAdultPopulation}}$
11	46	0.014	$36.660 + \frac{0.663\,(2.136\,\text{FixedInvestment}\cdot\text{GovernmentReceipts}\cdot\text{HouseholdConsumption})}{\text{Population}}$
12	46	0.026	$-79.935 + \frac{0.307\,(3\,\text{FixedInvestment}\cdot\text{GovernmentReceipts}\cdot\text{HouseholdConsumption})}{\text{AdultPopulation}}$
13	62	0.014	$33.122 + \frac{0.662\,(3\,\text{FixedInvestment}\cdot\text{GovernmentReceipts}\cdot\text{GrossInvestment}\cdot\text{HouseholdConsumption}\cdot\text{Population})}{\text{Population}}$
14	63	0.018	$-371.255 + \frac{0.164\,(\text{AdultPopulation}\cdot\text{ExportValue}\cdot5\,\text{GrossInvestment}\cdot3\,\text{TotalConsumption})}{\text{AdultPopulation}}$
15	64	0.017	$-416.607 + \frac{-0.811\,(\text{FixedInvestment}\cdot\text{GrossInvestment}\cdot\text{TotalConsumption})}{\text{FixedInvestment}\cdot\text{Population}}$
16	80	0.012	$-60.304 + \frac{0.087\,(\text{ExportValue}\cdot14\,\text{FixedInvestment}\cdot5\,\text{GovernmentReceipts}\cdot\text{Population}\cdot7\,\text{TotalConsumption})}{\text{Population}}$
17	80	0.014	$70.673 + \frac{0.132\,(\text{ExportValue}\cdot5\,\text{GovernmentReceipts}\cdot3\,\text{GrossInvestment}\cdot\text{HouseholdConsumption}\cdot3\,\text{TotalConsumption})}{\text{Population}}$
18	87	0.014	$-13.581 + \frac{0.152\,(7\,\text{FixedInvestment}\cdot3\,\text{GovernmentReceipts}\cdot2\,\text{GrossInvestment}\cdot4\,\text{TotalConsumption})}{\text{Population}}$



GDP per Capita Model–based Outliers

outlier rank	country	outlier distance	classification
1	CaymanIslands	16.99	Overpredicted
2	Greece	14.77	Overpredicted
3	Singapore	−14.60	Underpredicted
4	Switzerland	−10.59	Underpredicted
5	Qatar	−7.53	Underpredicted
6	Ireland	−6.67	Underpredicted
7	Iceland	6.34	Overpredicted
8	Bermuda	−6.66	Underpredicted
9	Canada	−6.38	Underpredicted
10	UnitedStates	−5.95	Underpredicted
11	Spain	5.41	Overpredicted
12	Japan	−5.72	Underpredicted
13	Bahrain	−5.11	Underpredicted
14	Estonia	4.10	Overpredicted
15	SaudiArabia	−4.35	Underpredicted
16	Denmark	3.09	Overpredicted
17	NewZealand	−3.10	Underpredicted



www.evolved-analytics.com

Symbolic Regression: Summary Benefits

- Compact Nonlinear Models
  - Compact empirical models can be suitable for online implementation
  - Model(s) can be used as an emulator for coarse system optimization
- Driving Variable Selection & Identification
  - Appropriate models may be developed from poorly structured data sets (too many variables & not enough measurements)
  - Identified driving variables may be used as inputs into other modeling tools
- Metasensor (Variable Transform) Identification
  - Identifying variable couplings can give insight into underlying physical mechanisms
  - Identified metavariables can enable linearizing transforms to meld symbolic regression and more traditional statistical analysis
- Diverse Model Ensembles
  - The independent evolutions will produce independent models. Independent (but comparable) models may be stacked into ensembles whose divergence in prediction may be an indicator of extrapolation & model trustworthiness. This is an issue in high dimensional parameter spaces.
- Human Insight
  - The transparency of the evolved models as well as the explicit identification of the model complexity-accuracy trade-off is very compelling
  - Examining an expression can be viewed as a visualization technique for high-dimensional data
- Rapid Modeling
  - Exploitation of the Pareto front has resulted in orders-of-magnitude in the symbolic regression performance relative to more traditional GP. This greatly increases the range of possible applications.
- Metavariables can also be used as inputs into other modeling tools

www.evolved-analytics.com

EC in Dow Chemical

Application Domains	Examples
Material Design	• Color Matching • Appearance Engineering • Polymer Design • Synthetic Leather
Materials Research	• Diverse Chemical Library Selection • Fundamental Model Building • Reaction Kinetics Modeling • Combi-Chem Catalyst Exploration • Combi-Chem Data Analysis
Production Design	• Acicular Mullite Emulator • EDC/VCM Nonlinear DOE • Bioreactor Optimization
Production Monitoring & Analysis	• Epoxy Holdup Monitoring • Isocyanate Level Estimation • FTIR Calibration Variable Selection • Poly-3 Volatile Emission Monitoring • Epoxy Intelligent Alarm Processing • PerTet Emulator for Online Optimization • Emissions Monitoring
Business Modeling	• Diffusion of Innovation • Hydrocarbon Trading & Energy Systems Optimization • Scheduling Heuristics • Plant Capacity Drivers

More Information

www.evolved-analytics.com

Acknowledgement

Guido Smits
Katya Vladislavleva
Arthur Kordon