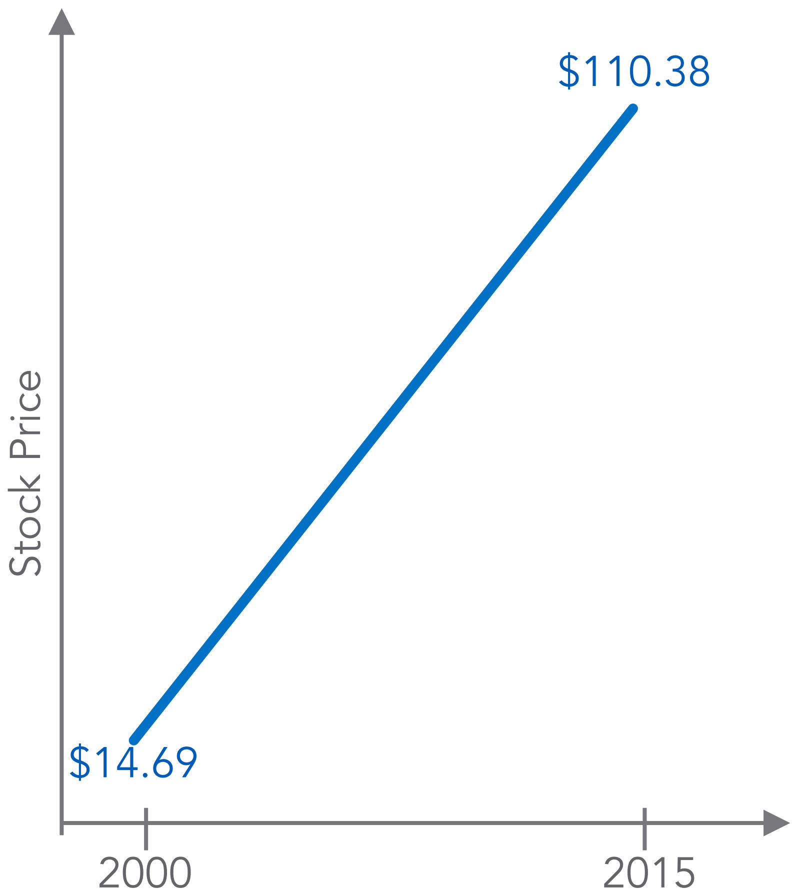
Big Math For Big Problems

Predicting Predictions: The Use of Bayesian Model Averaging To Select Models

Diversification In Retirement Planning

Concentrated Portfolio: AAPL



Diversified Portfolio

TOYOTA

Microsoft

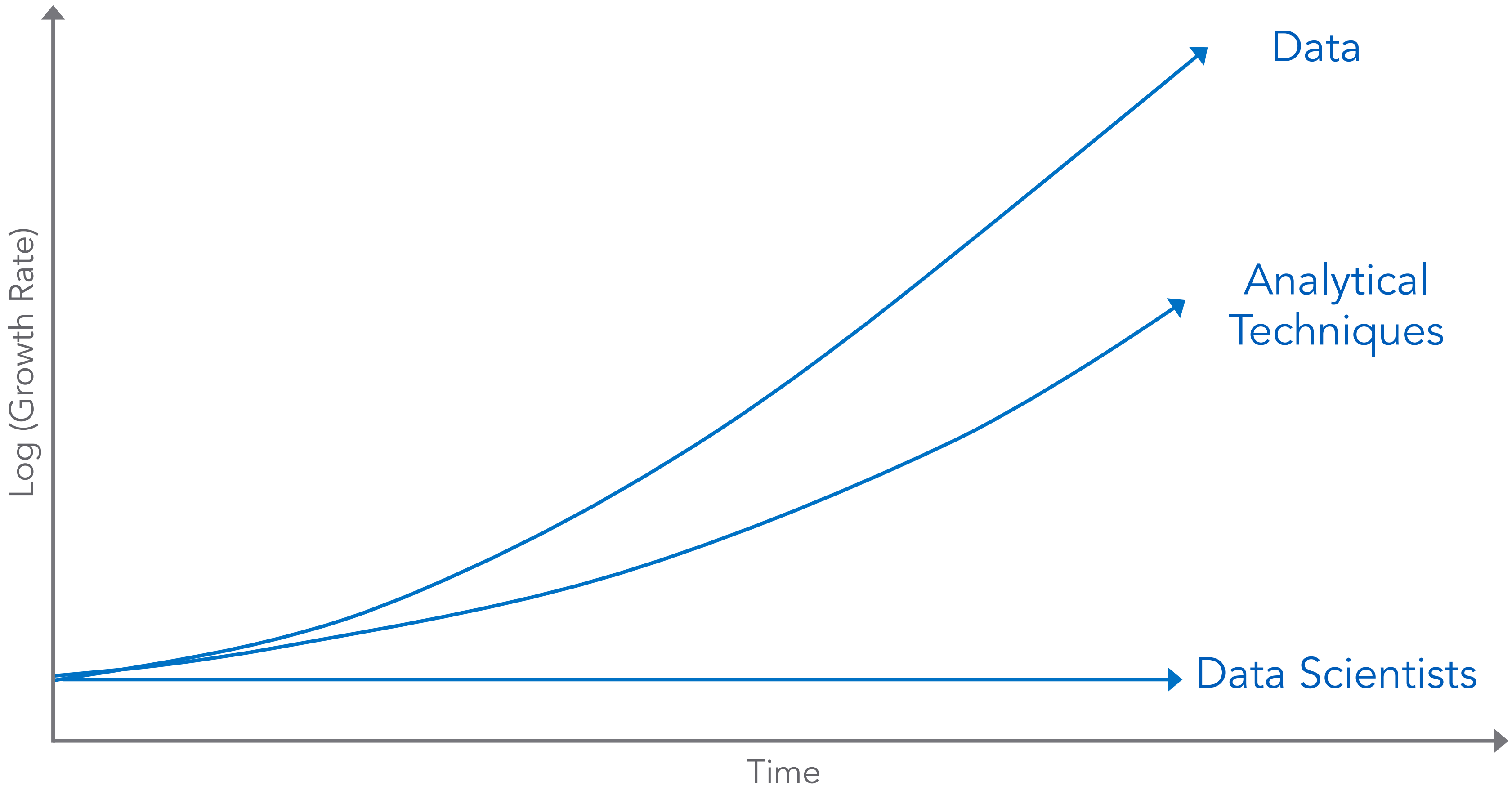
biogen idec

Bonds

Commodities

Concentration Is Common In Modeling

Explosive growth in digital data is driving a search for better software models.



Data Grows In Two Ways

Long Data

Rows

	A	B	C	D
1	Individual Sales made in Store B	Number of Items in Sale	Specific Items Sold	Date
2	$235.90	6	4, 11, 25, 75, 13, 2	1/15/13
3	$25.00	1	18	1/15/13
4	$260.00	7	13, 2, 18, 111, 5, 1, 215	1/15/13
5	$210.00	5	71, 9, 41, 69, 23	1/15/13
6	$190.00	5	72, 35, 11, 12, 34	1/15/13
7	$27.00	1	37	1/15/13
8	$240.00	6	7, 13, 25, 75, 13, 2	1/15/13
9	$22.00	1	56	1/15/13
10	$21.00	1	72	1/15/13
11	$250.00	6	4, 11, 25, 75, 13, 2	1/15/13
12	$25.50	1	51	1/15/13
13	$250.55	6	1, 23, 22, 75, 13, 2	1/15/13
14	$52.00	1	17	1/15/13
15	$52.40	1	72	1/15/13
16	$261.00	7	75, 2, 18, 111, 5, 1, 215	1/15/13
17	$271.20	7	111, 5, 1, 215, 75, 13, 2	1/15/13
18	$25.00	1	72	1/15/13
19	$70.77	2	67, 89	1/15/13
20	$250.00	6	4, 11, 25, 75, 13, 2	1/15/13
21	$78.00	2	31, 95	1/15/13
22	$210.00	5	71, 9, 41, 69, 23	1/15/13
23	$190.00	5	72, 35, 11, 12, 34	1/15/13
24	$54.00	1	72	1/15/13
25	$72.00	2	45, 23	1/15/13
26	$220.00	6	10, 3, 29, 51, 13, 2	1/15/13
27	$210.00	5	71, 9, 41, 69, 23	1/15/13
[illegible]	[illegible]		[illegible]	[illegible]

Wide Data

	A	B	C	D	E	F	G	H	I	J	K
1	Store	Total Sales	Location	Inventory Total (units)	Size of Store (sq. ft.)	Operation Hours	Days Store is Open	Peak Hours	Reps in Store at Peek Hours	POS System	Location Specs
2	A	$2,359,000	City 1	786333	43685	9 - 9	M, T, W, Th, F, S, S	12 - 7	6	2012	Strip M
3	B	$250,000	City 2	83333	4630	10 - 6	M, W, Th, F	2 - 4	1	2000	Stand Alon
4	C	$2,600,000	City 3	866667	48148	9 - 9	M, T, W, Th, F, S, S	12 - 7	4	2012	Mall
5	D	$2,100,000	City 4	700000	38889	9 - 9	M, T, W, Th, F, S, S	2 - 4	2	2000	Mall
6	E	$1,900,000	City 5	633333	35185	9 - 9	M, T, W, Th, F, S, S	12 - 7	8	2012	Mall
7	F	$270,000	City 6	90000	5000	10 - 6	M, W, Th, F	2 - 4	1	2000	Strip Mall
8	G	$2,400,000	City 7	800000	44444	9 - 9	M, T, W, Th, F, S, S	12 - 7	4	2012	Stand A
9	H	$220,000	City 8	73333	4074	10 - 6	M, W, Th, F	1 - 3	2	2000	Stand Alone
10	I	$210,000	City 9	70000	3889	10 - 6	M, W, Th, F	1 - 3	3	2000	Stand Al
11	J	$2,500,000	City 10	833333	46296	9 - 9	M, T, W, Th, F, S, S	12 - 7	7	2012	Stand A
12	K	$255,000	City 11	85000	4722	10 - 6	M, W, Th, F	1 - 3	5	2000	Strip Mall
13	L	$2,505,500	City 12	835167	46398	9 - 9	M, T, W, Th, F, S, S	1 - 5	6	2000	Strip Ma
14	M	$520,000	City 13	173333	9630	10 - 6	M, W, Th, F	2 - 4	6	2000	Mall
15	N	$524,000	City 14	174667	9704	10 - 6	M, W, Th, F	1 - 3	6	2000	Mall
16	O	$2,610,000	City 15	870000	48333	9 - 9	M, T, W, Th, F, S, S	12 - 7	1	2000	Strip M
17	P	$2,712,000	City 16	904000	50222	9 - 9	M, T, W, Th, F, S, S	1 - 5	7	2000	Stand Alo
18	Q	$250,000	City 17	83333	4630	10 - 6	M, W, Th, F	2 - 4	8	2000	Stand Alone
19	R	$707,700	City 18	235900	13106	9 - 9	M, T, W, Th, F, S, S	1 - 5	1	2000	Strip M
20	S	$2,500,000	City 19	833333	46296	9 - 9	M, T, W, Th, F, S, S	12 - 7	2	2000	Strip Mall

Variables

Implications Of Wide Data

As the width of a data set grows, the number of possible relationships among the variables grows super-exponentially.

10 variables = 1,024 possible models

~large jar of jelly beans

Common problems with a small number of variables like retail sales predictions

100 variables >1 x 10^{29} possible models

> grains of sand on Earth

Rare events, like a power plant failure or a security breach

1,000 variables = nearly infinite models

> all the atoms in the universe

Crop genetics problems routinely have 10,000+ variables

Diversification Mitigates Risk

Best Possible Model Ever?

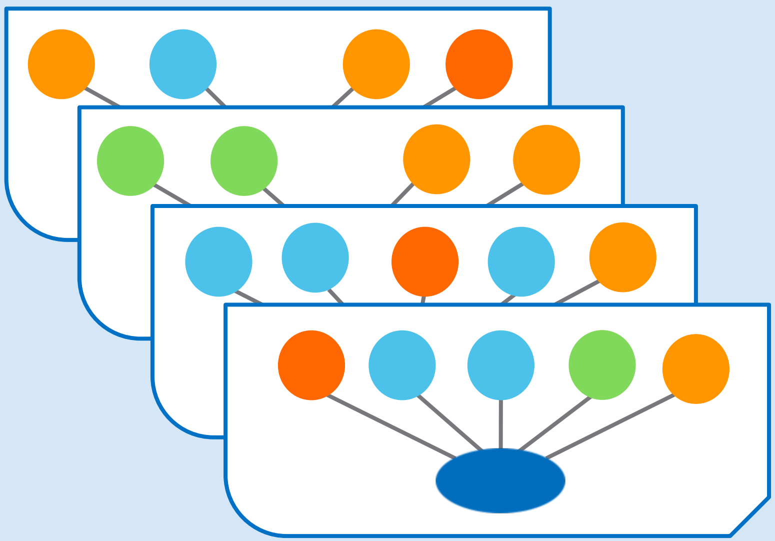
Best Model for Hot Weather

Best Model 3 Years from Now

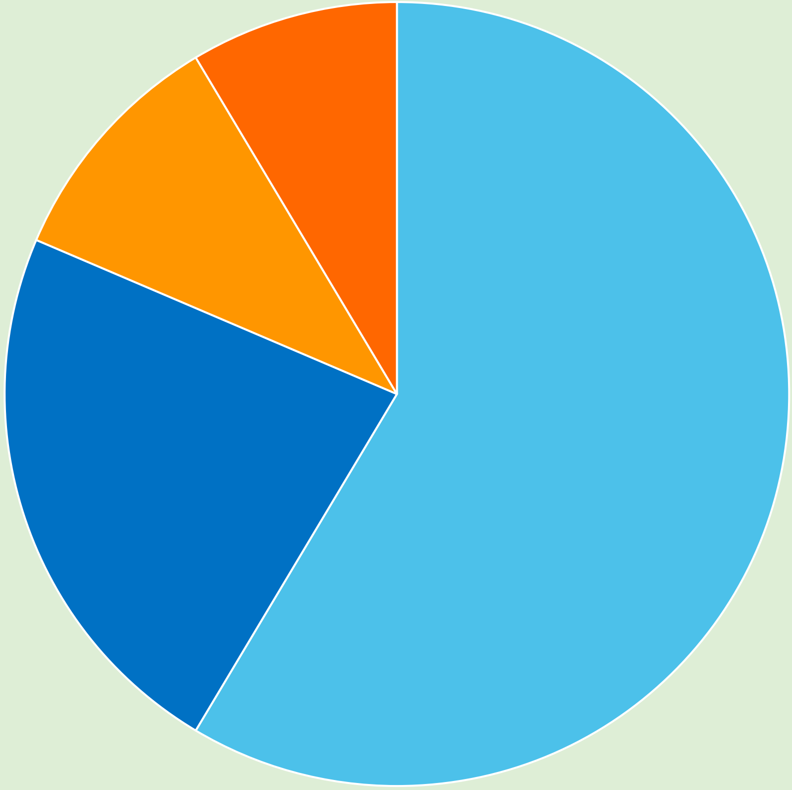
Best Model When Energy Prices are Stable

Ensemble vs. Portfolio Diversification

Ensemble: Same data sources with same analytical approach



Portfolio: Diverse data sources and diverse analytical approaches

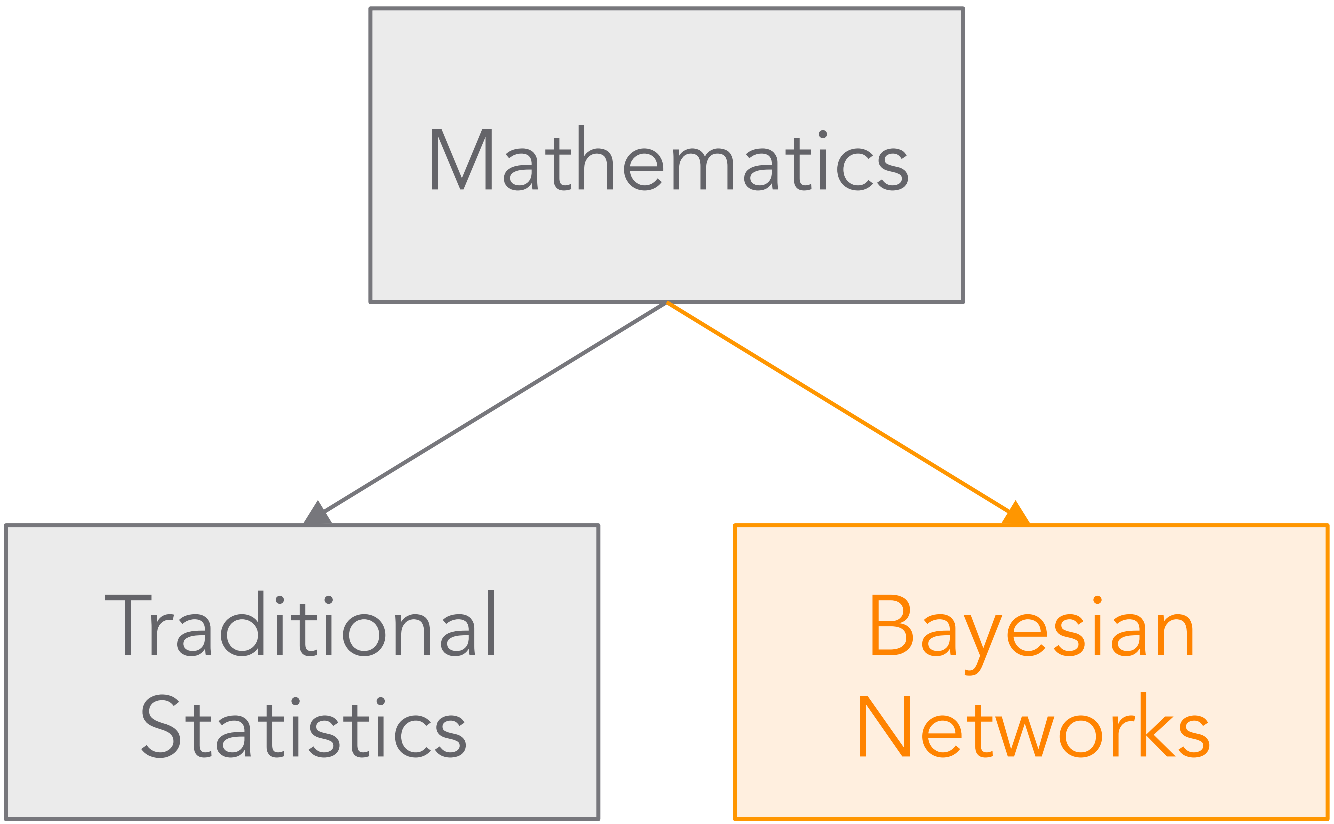


The Catch

Common Issues In Portfolio Management		
Correlation / Co-Linearity	Negative Weights	Hypothesis Bias

Proposed Solution

The combination of advanced machine learning and Bayesian networks can address common portfolio issues.



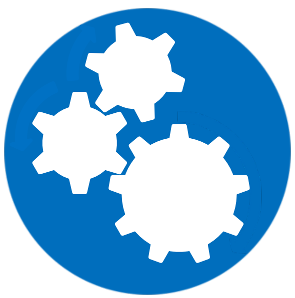
Example Of Portfolio Of Models

A portfolio would track performance against important scenarios.

	A	B	C	D	E
1	Model 1	Model 2	Model n	Weather	Energy Price
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
21					
22					
23					
24					

Beyond Algorithms

Ideally, our learning regarding algorithm performance goes beyond predictive performance but also measures outcomes from interventions.



$$x = Ay^2 + Bz + \varepsilon$$

The Benefits For Big Problems

A portfolio approach to software modeling is worth the effort when done well.

- Risk management
- Rigor and scale with which models can be compared
- Ability to integrate latest algorithmic advances
- Closed loop learning

VIA SCIENCE

Colin Gounden

- cgounden@viascience.com

Jeremy Taylor

- jtaylor@viascience.com