CONAMIC2015

EL FORO DE LAS FINANZAS POPULARES

Innovative Analytics for Traditional, Social, and Text Data

Dr. Gerald Fahner, Senior Director Analytic Science, FICO

Hot Trends in Predictive Analytics

Big Data – the Fuel

"is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."

Gartner

Machine Learning - the Engine

"is a scientific discipline that explores the construction and study of ***algorithms that can learn from data****. Such algorithms operate* ***by building a model from example inputs*** *and using that* ***to make predictions or decisions****, rather than following strictly static program instructions. Machine learning is closely related to and often overlaps with computational statistics; a discipline that also specializes in prediction-making."*

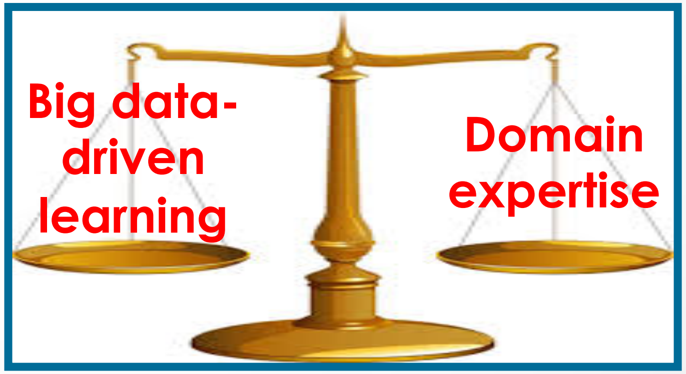
Wikipedia

Domain Expert – the Driver

Balance Number Crunching With Expertise

Big Data and Machine Learning can help you to ***Predict consumer behavior more accurately***

To unlock this value requires domain expertise to build ***Comprehensible models for justifiable decisions***



Deeper Insights - Stronger Predictions – Better Decisions

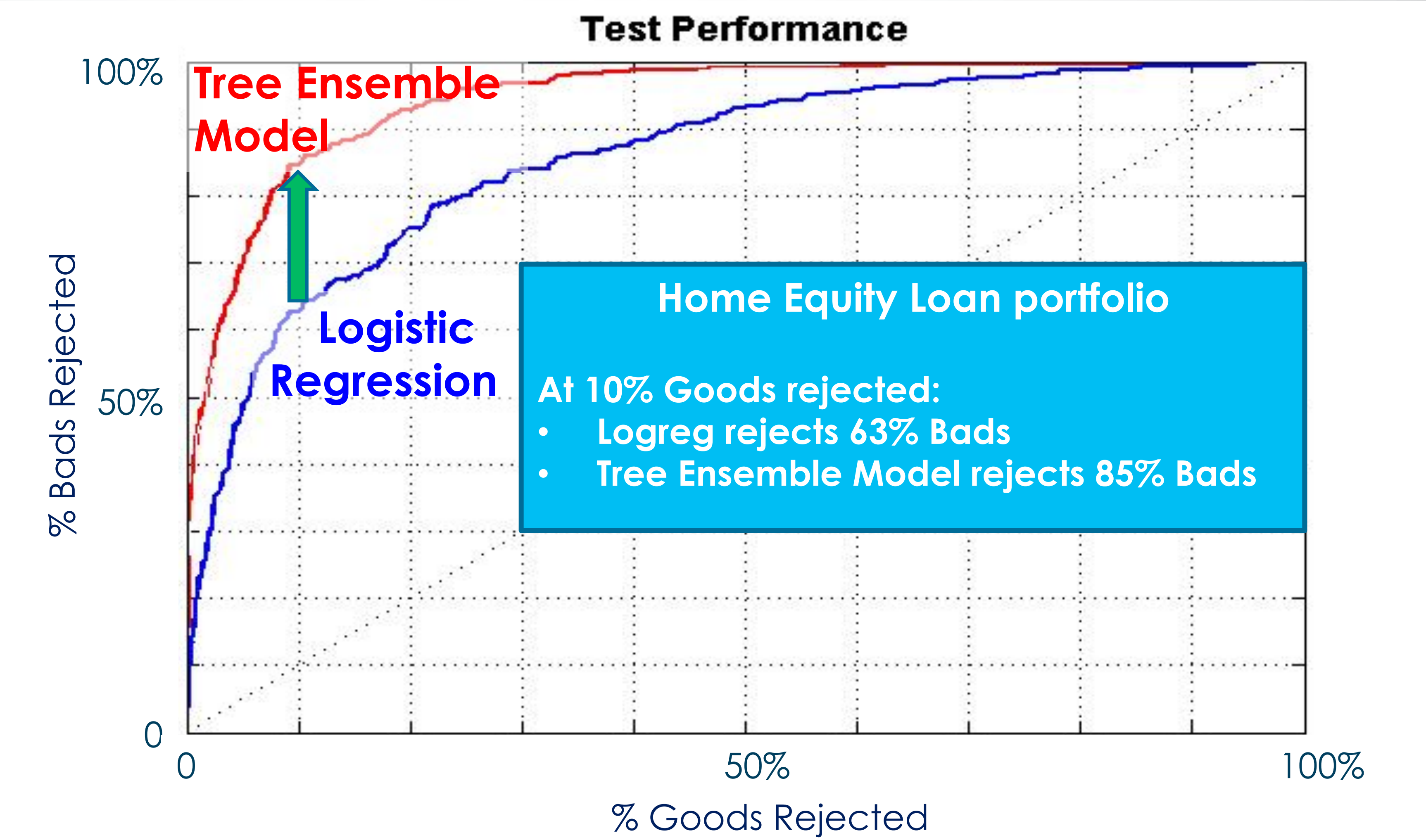
Case Studies

Informing Origination Risk Score Development by Machine Learning

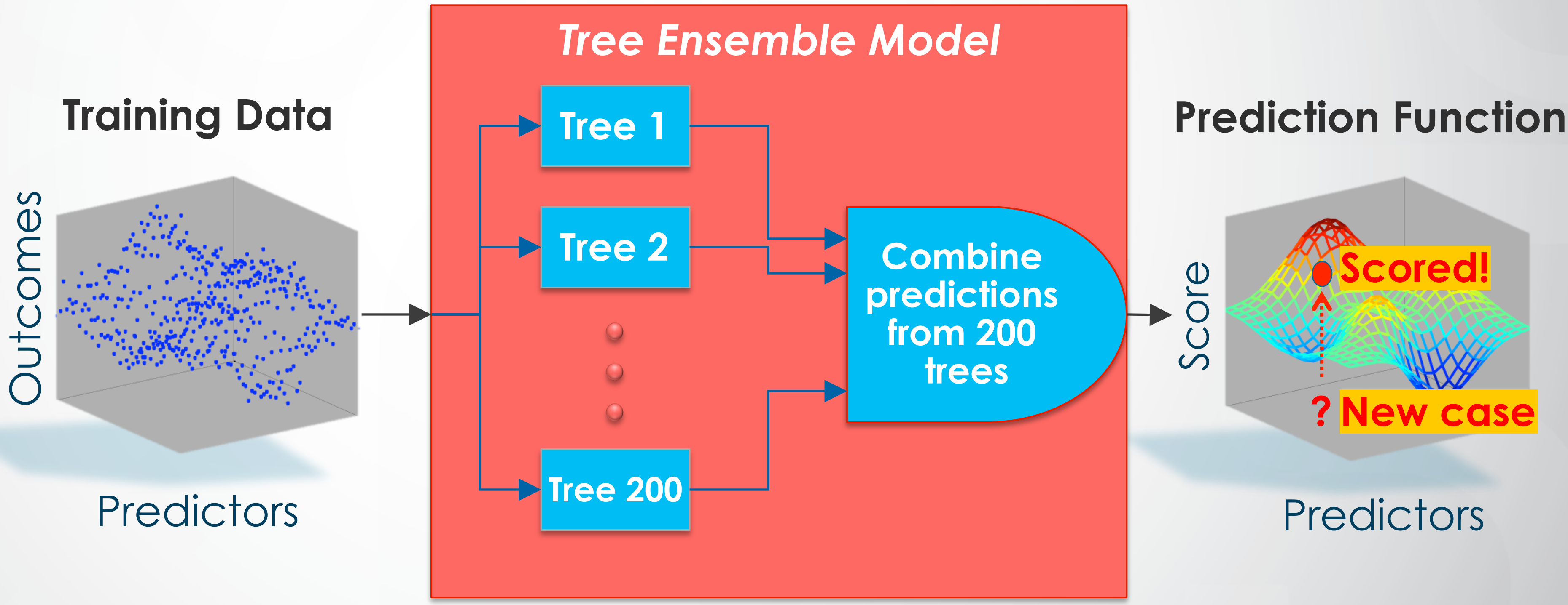
Improving Insurance Fraud Model With Social Network Variables

Evaluating Predictive Power of Text Data for a Peer Lending Network

Machine Learning Models Can Beat Simple Models by Substantial Margins



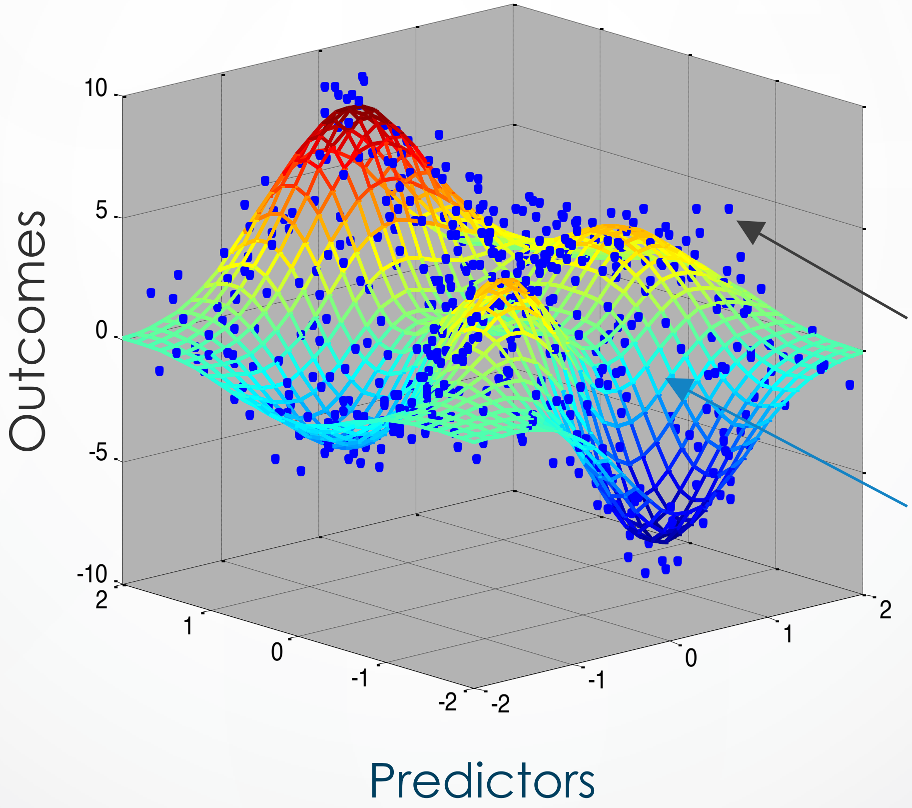
Anatomy of a Tree Ensemble Model



Random Forest

Gradient Boosting

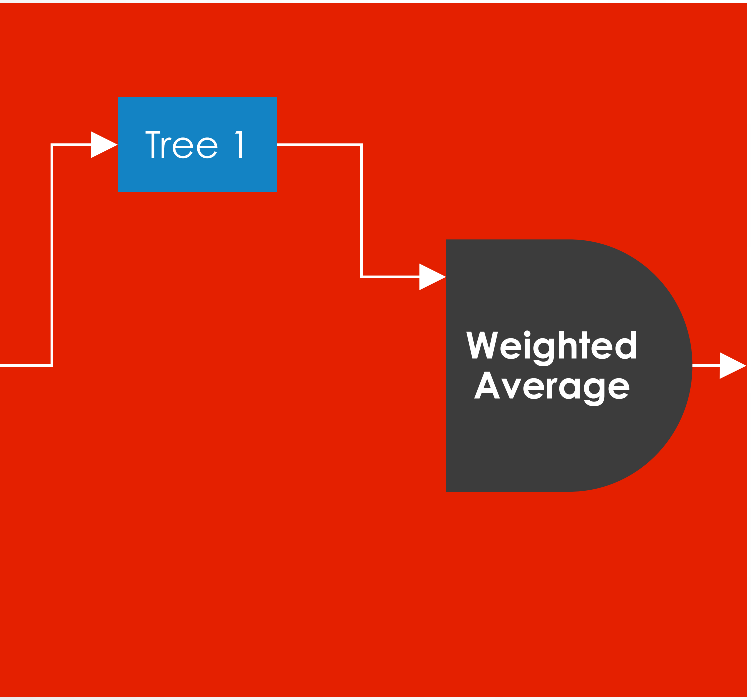
Demonstration Problem

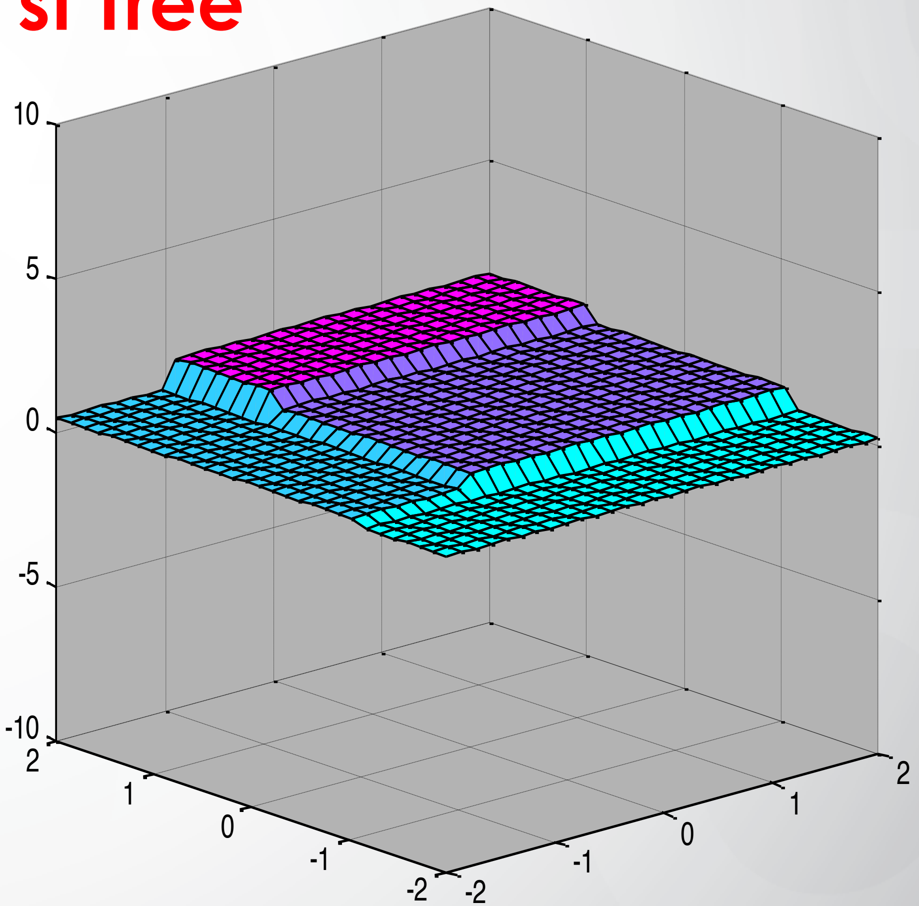


Noisy training samples

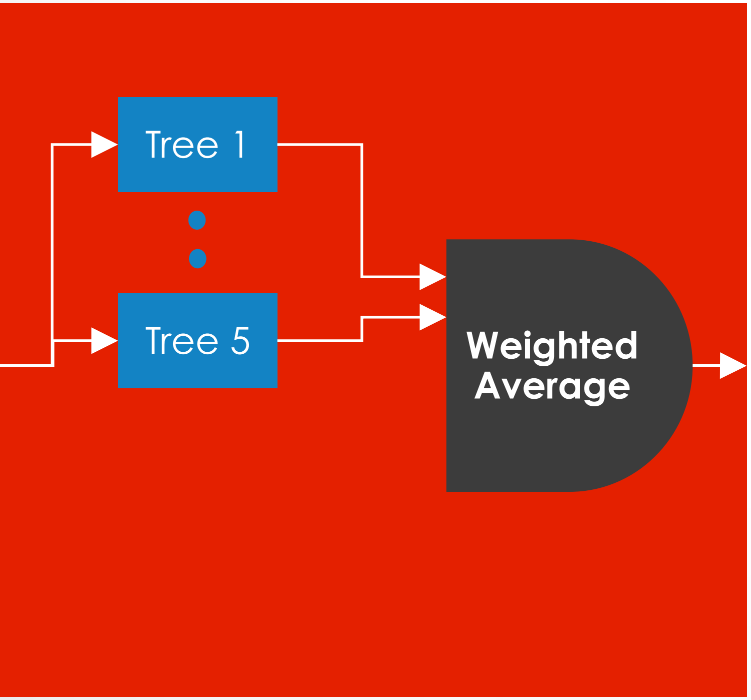
Predictive relationship from which data were generated ("ground truth")

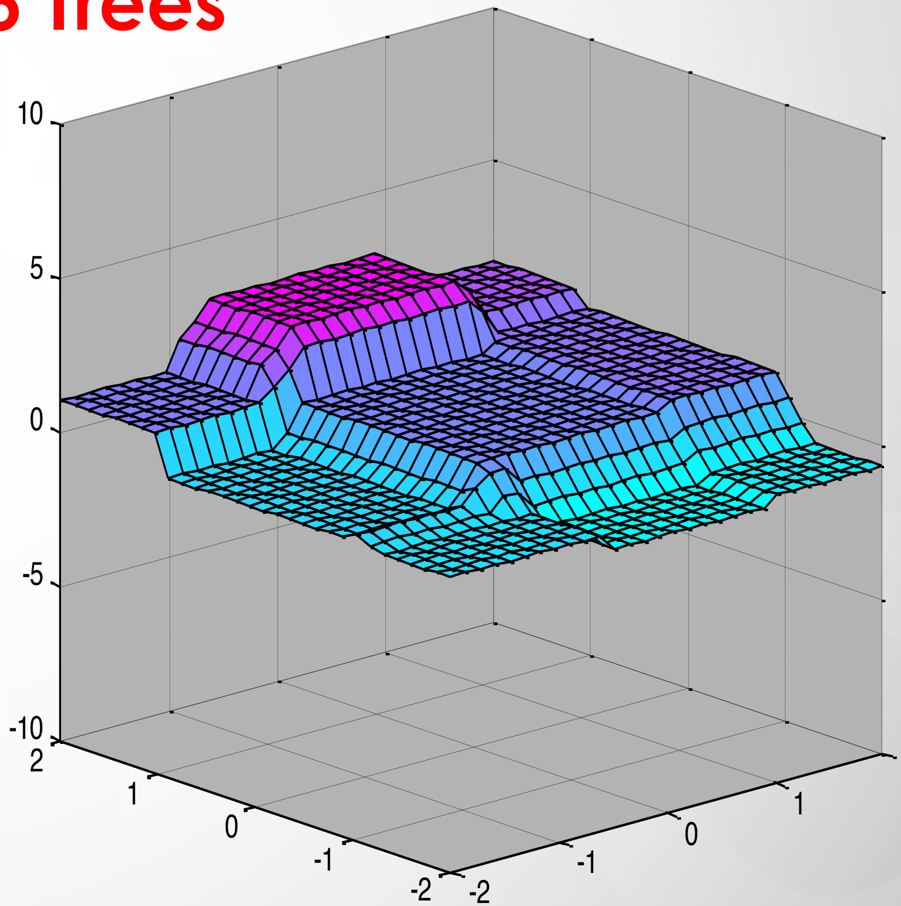
Stochastic Gradient Boosting
1'st Tree



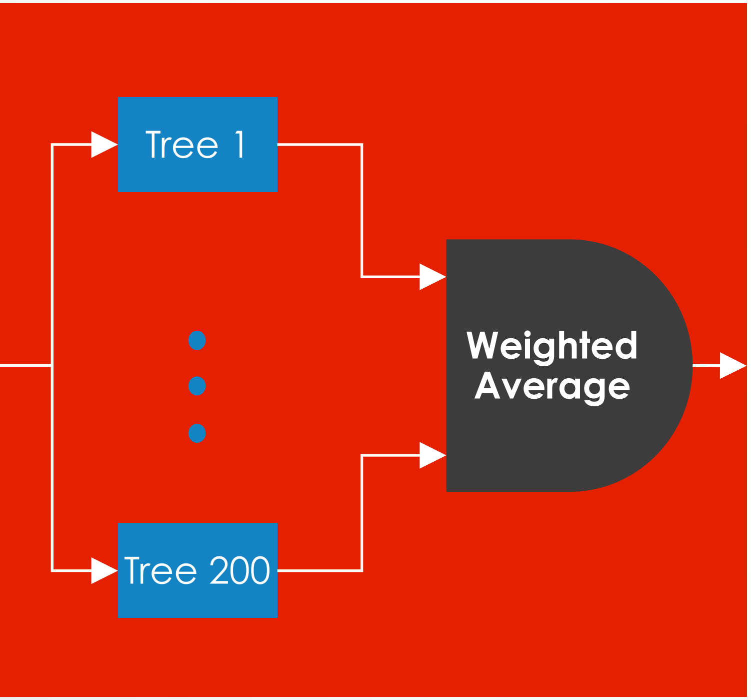


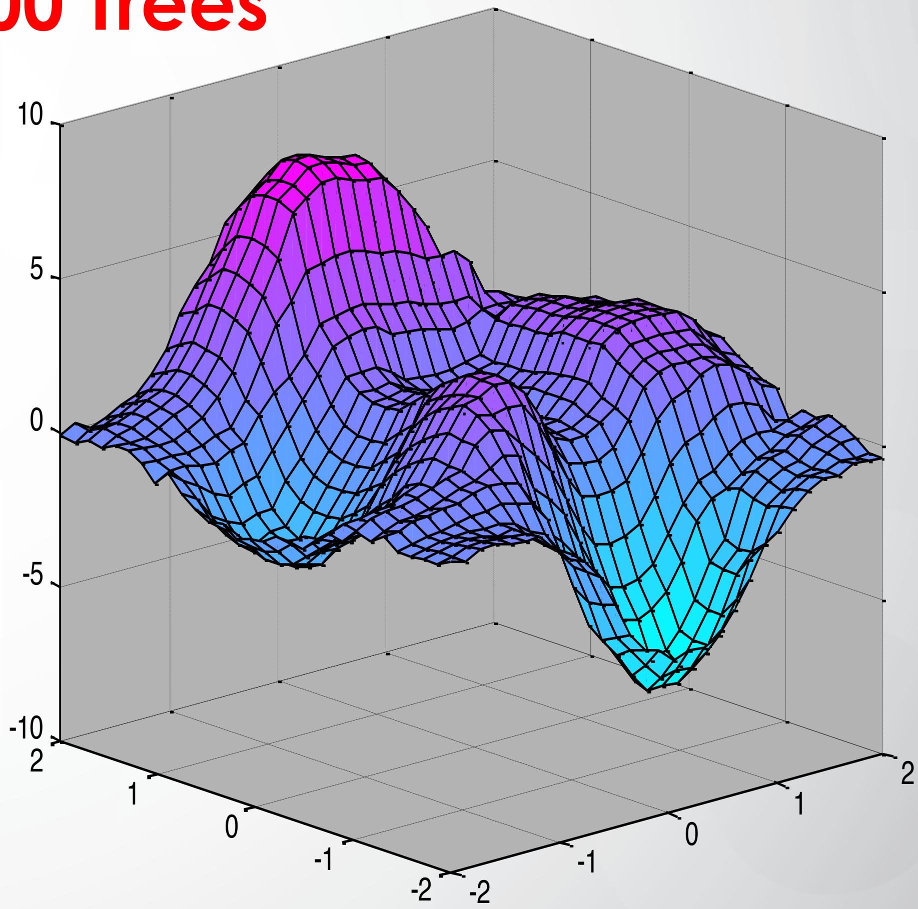
Stochastic Gradient Boosting
5 Trees





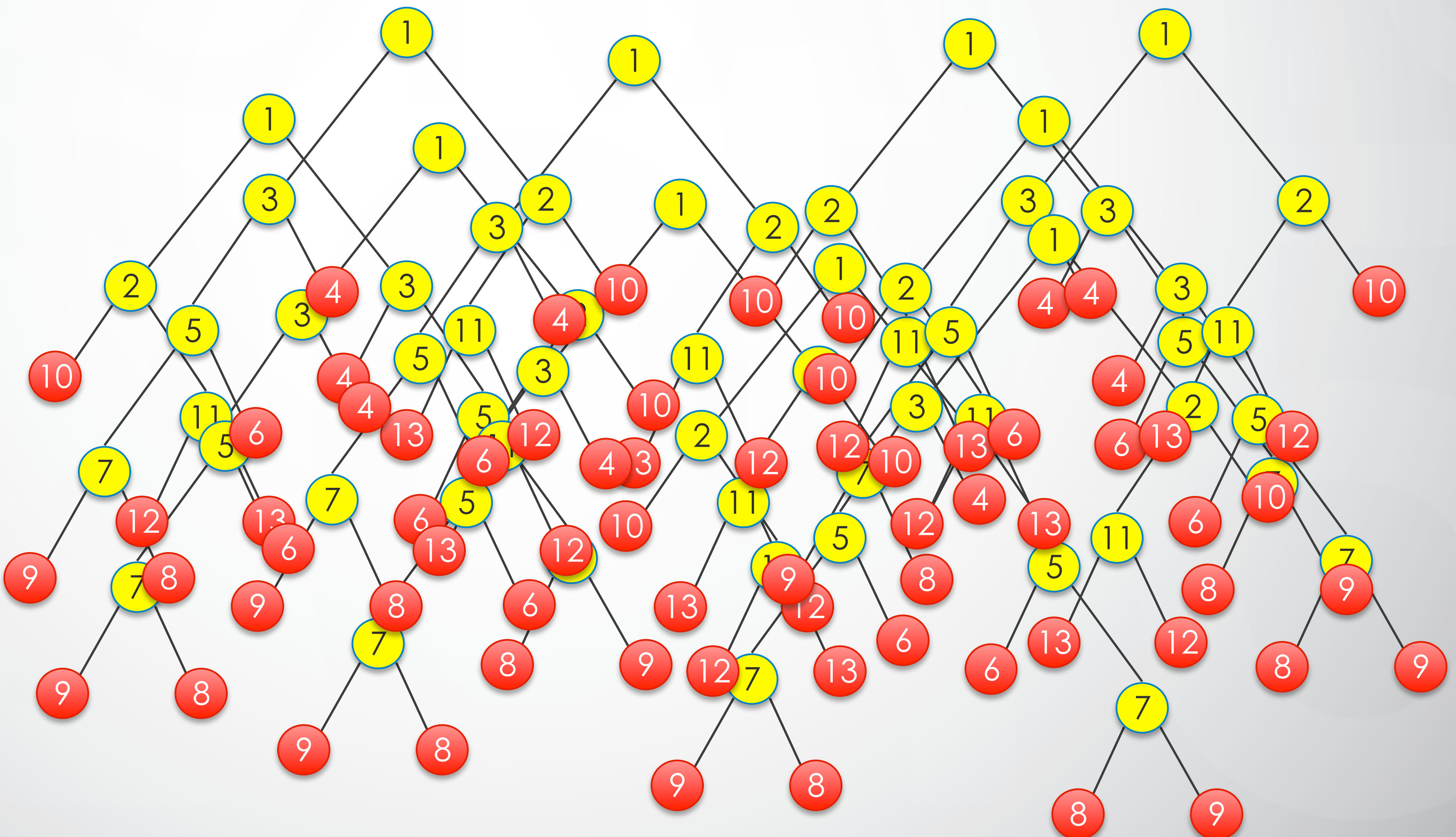
Stochastic Gradient Boosting
200 Trees





CONAMIC2015
EL FORO DE LAS FINANZAS POPULARES

Direct Inspection Yields a Black Box



Useful Diagnostic Information

Variable Importance

Relative to most important variable

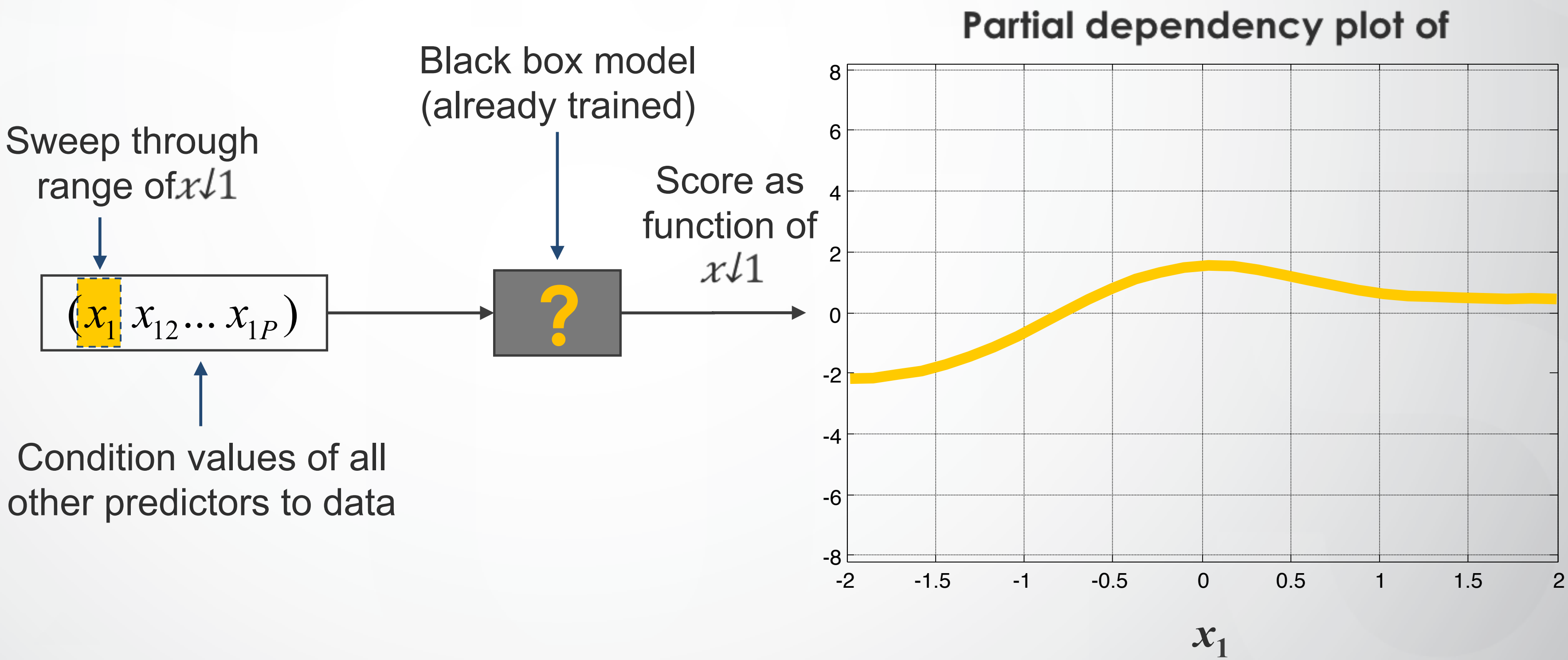
DEBTINC	1.00
DELINQ	0.49
VALUE	0.45
CLAGE	0.40
DEROG	0.34
LOAN	0.25
CLNO	0.24
MORTDUE	0.23
NINQ	0.23
JOB	0.20
YOJ	0.20
REASON	0.07

Interaction Test Statistics

Whether a variable interacts with other variables

DEBTINC	0.029
VALUE	0.022
CLNO	0.014
CLAGE	0.013
DELINQ	0.010
YOJ	0.008
MORTDUE	0.007
LOAN	0.007
DEROG	0.006
JOB	0.006

Input/Output Simulation Create Deeps Insight Into Black Box



Practical Pitfalls Hindering Deployment of Machine Learning Models

Non-intuitive Associations Can Lead to Unjustifiable Credit Decisions

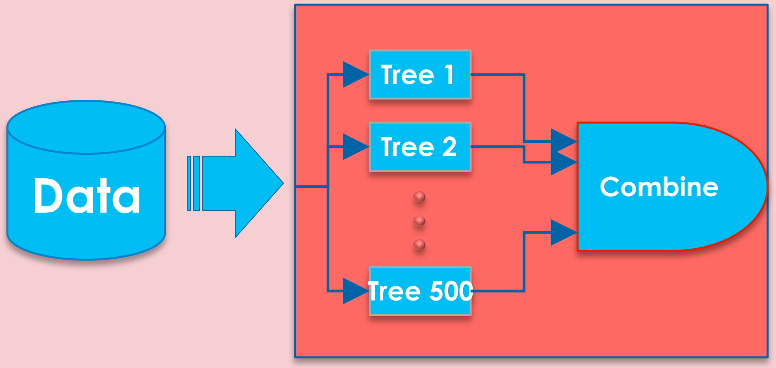
Association … (after controlling for all else)	… Could lead to decision
Loan Application: Consumers with 10% debt ratio have lower risk score than consumers with 30% debt ratio	Applicant is rejected because her debt ratio is ***not high enough(!?)***
Mortgage Lending: Consumers without previous mortgage have lower risk score than consumers with previous mortgage	Applicant ***can't get a mortgage because he doesn't have a mortgage(!?)***

Pros and Cons of Machine Learning Models for Credit Scoring

PROS	CONS
Highly accurate fit to data	Vulnerable to data limitations
Discovery of unexpected associations	May capture non-intuitive associations
Automated, productive analysis	Hard to impose domain expertise

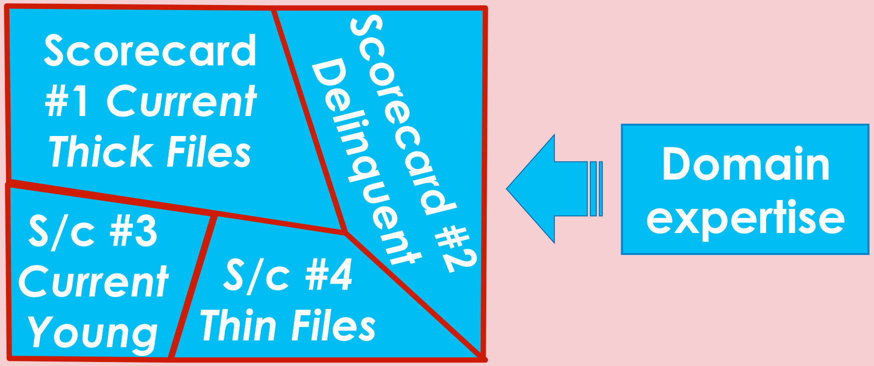
"Scorecardizer" Approach: Converts Machine Learnings into Powerful Comprehensible Scorecards

First train Machine Learning model



Completely data-driven discovery

Then convert learnings into (segmented) Scorecard(s)



Domain expertise can be injected to warrant model palatability

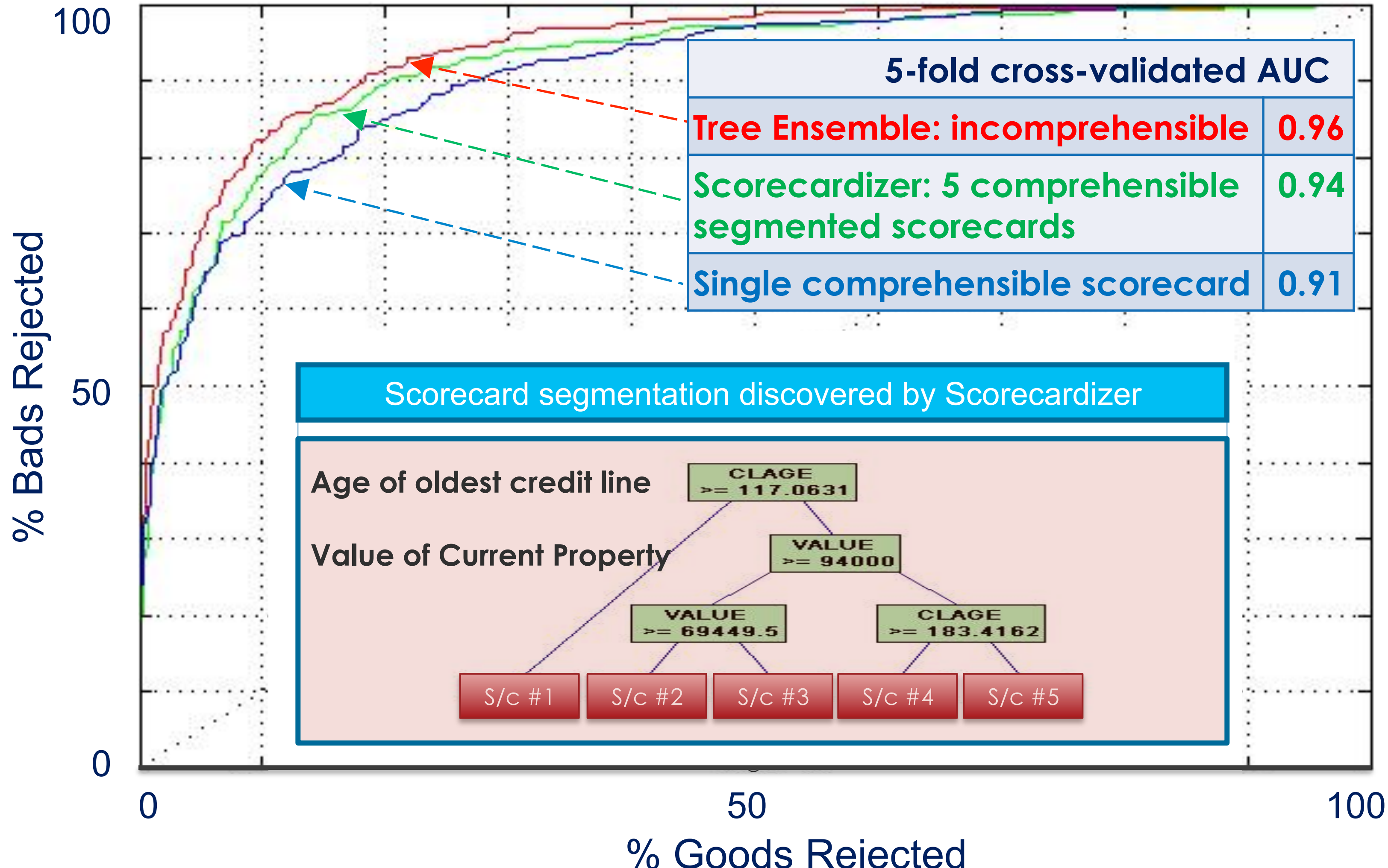
Can We Automate Expertise?

- In well-defined domains (e.g. credit scoring), can codify expertise
- Scorecardizer hooks into a database of expert rules, which enables **fully automatic construction of palatable models**
  - Manual refinement of the "end product" by domain experts is still possible before deployment

Some examples of codified expertise

- Everything else equal:
  - Score must not increase with higher Debt Ratio
  - Score must not decrease with Applicant Age above 60 years
  - For current accounts, a history of delinquency is bad
  - For 2-cycle delinquents, history of mild delinquency can be good (these have shown their capacity to recover)

Results for Home Equity Loan Portfolio



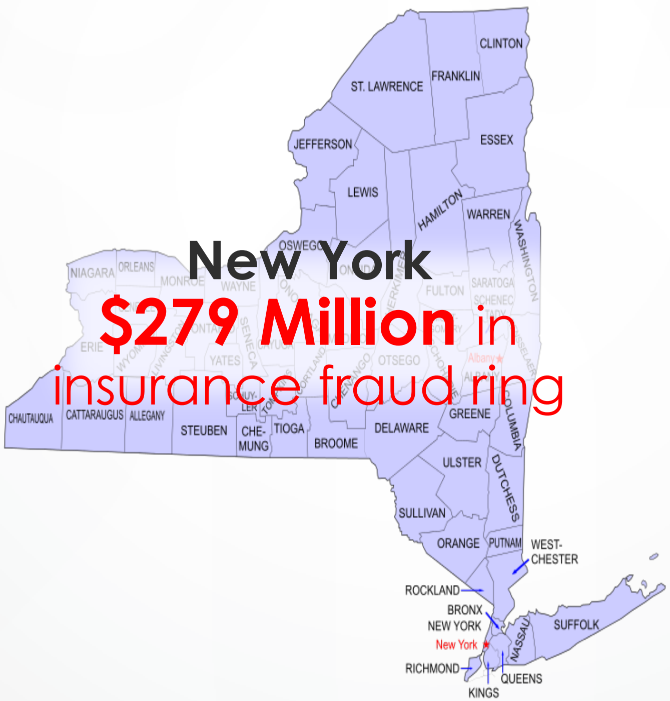
Case Studies

Informing Origination Risk Score Development by Machine Learning

Improving Insurance Fraud Model With Social Network Variables

Evaluating Predictive Power of Text Data for a Peer Lending Network

Insurance Fraud and Networks



New York

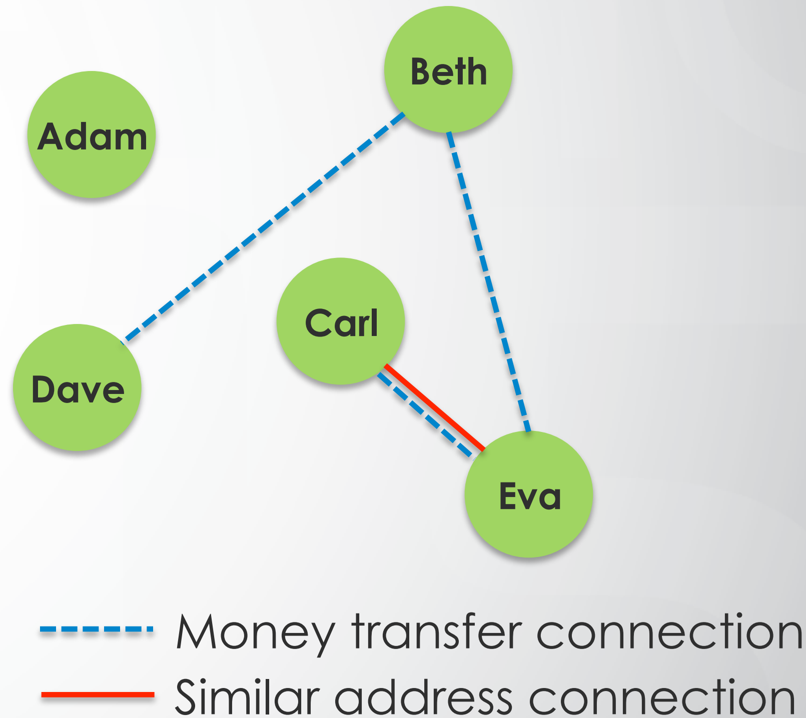
$279 Million in insurance fraud ring

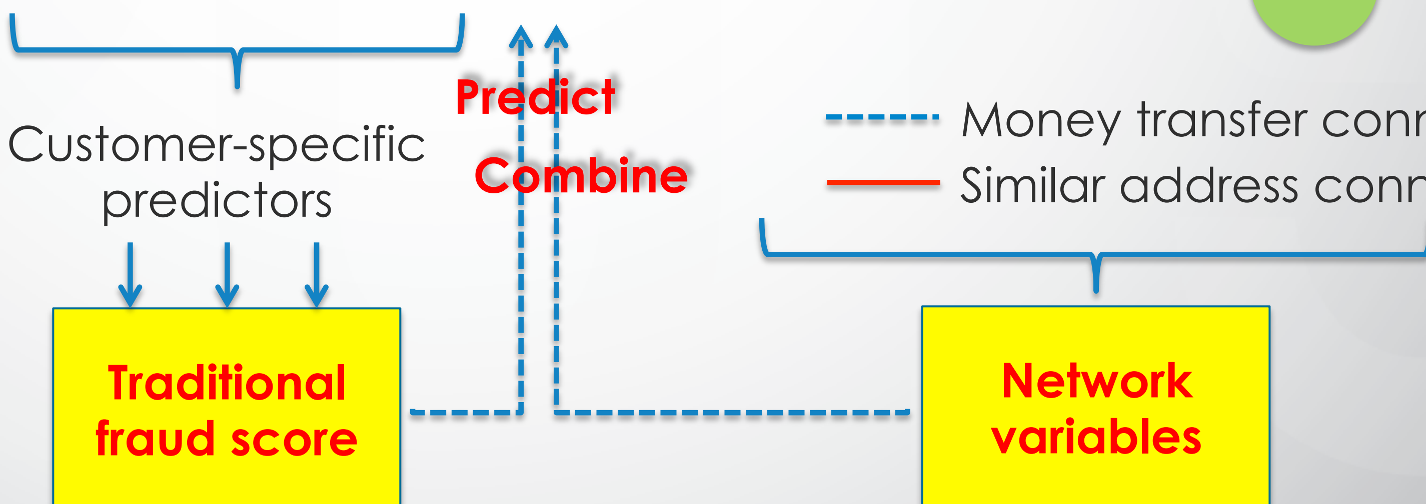
Florida

$1 Million in worker's compensation fraud ring

Predicting Fraud with Traditional and Network Data

	Age	Income	Credit Score	Fraud
Adam	21	7,000	680	0
Beth	36	3,200	744	0
Carl	62	9,000	803	1
Dave	32	4,250	713	0
Eva	49	800	720	0





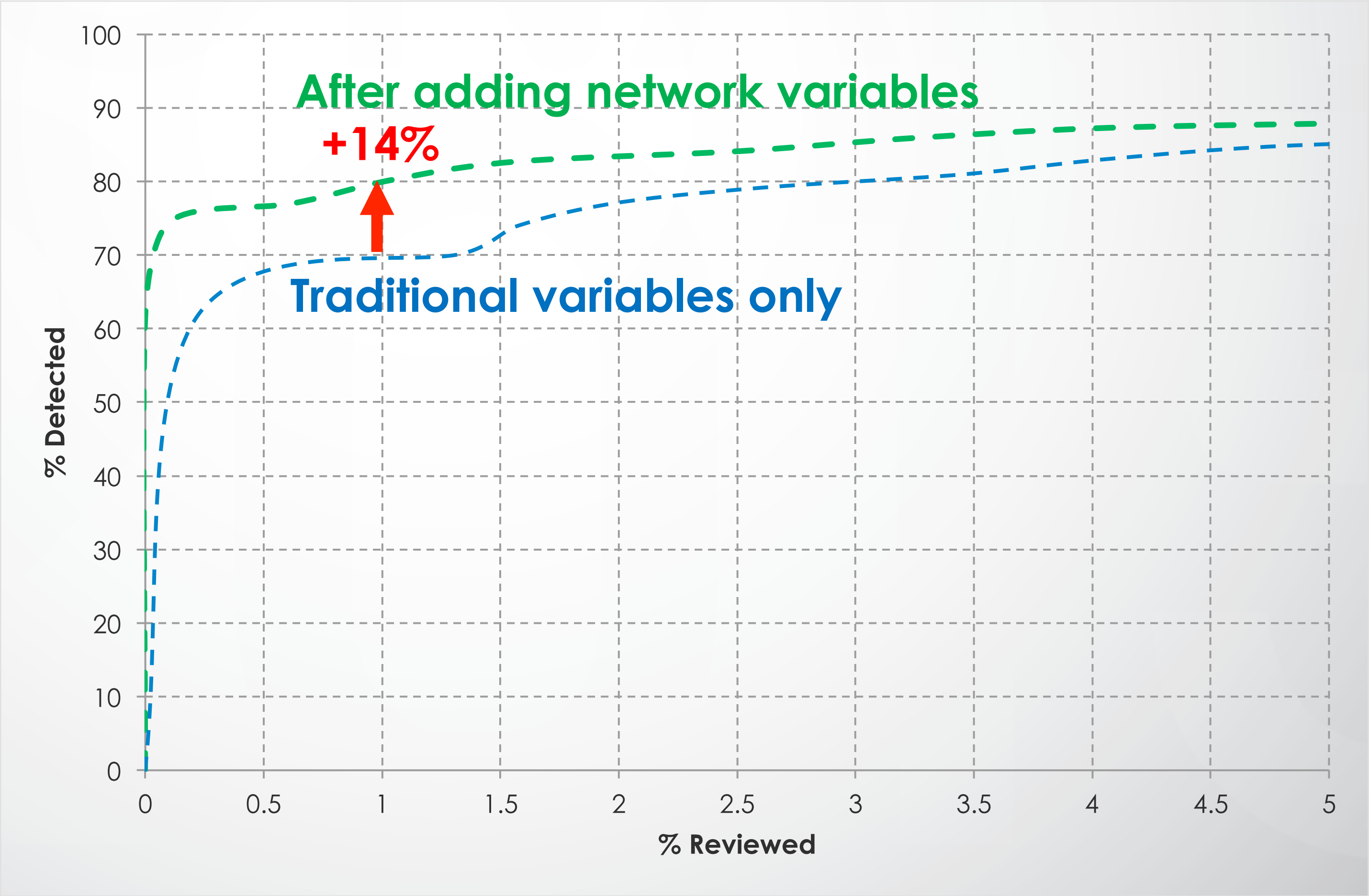
Enhancing Data with Network Variables

	Age	Income	Credit Score	Avg. Age of Connections	# Money Transfers in Network per Month	Fraud
Adam	21	7,000	680	23	2	0
Beth	36	3,200	744	45	3	0
Carl	62	9,000	803	21	7	1
Dave	32	4,250	713	37	3	0
Eva	49	800	720	55	4	0

Variables derived from network

- Graph algorithms
- Feature generation
- Feature evaluation
- Feature selection

Boosting Auto Insurance Fraud Detection with Network Information



Variable Importance
Top 10 Variables

Rank	Variable Name
1	Car Model
2	Total Paid Amount in Network
3	Policy Holder Occupation
4	Pre-accident Vehicle Value
5	Total # Payments in Network
6	# Phantom Vehicles in Network
7	ZIP Code of Policy Holder
8	Size of Network
9	Repairable Flag
10	Type of Accident

Network Variables

Case Studies

Informing Origination Risk Score Development by Machine Learning

Improving Insurance Fraud Model With Social Network Variables

Evaluating Predictive Power of Text Data for a Peer Lending Network

Ubiquitous Text Data Can Be Predictive and Yield New Insights

Call center records, claims, public records, collector notes, emails, blogs, social data, freeform comments, reviews, webpages, product descriptions, transcribed phone calls, news articles…

↓

Is there predictive value?
How can we leverage it for comprehensible models and justifiable decisions?

↓

Semantic Scorecards & Topic Analysis

Origination Risk for Peer-to-Peer Lending Network

Structured origination information

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Accounts_	CREDIT_G	Debt_To_Incom	Delinqu	Delin	Earliest_CREDIT_Li	Employme	FICO_Range	Home_Owners	Inquiries_	Loan_Purp	Monthly_	Months_S	Months_S	Open_CRE	Public_Re	Revolving	Revolving	Total_CRE
2	0	E2	14.28999996	1	0	10/27/2003 0:00	< 1 year	660-678	OWN	0	debt_cons	1833.33	11	0	7	0	4175	51.5	8
3	0	A2	3.720000029	0	0	11/19/1988 0:00	< 1 year	780+	MORTGAGE	0	other	16666.67	0	0	17	0	85607	0.7	26
4	0	A4	2.299999952	0	0	10/28/1998 0:00	< 1 year	714-749	MORTGAGE	0	debt_cons	8333.33	0	0	11	0	9698	19.4	20
5	0	C1	6.400000095	1	0	12/30/1986 0:00	3 years	679-713	RENT	1	credit_car	1500	5	0	6	0	8847	26.9	9
6	0	A4	11.32999992	0	0	11/11/1990 0:00	4 years	750-779	MORTGAGE	0	home_im	9166.67	0	0	13	0	7274	13.1	40
7	0	B1	15.55000019	0	0	5/24/1994 0:00	5 years	750-779	MORTGAGE	0	credit_car	6250	0	0	10	0	66033	23	29
8	0	A2	0.310000002	0	0	10/5/1997 0:00	1 year	780+	OWN	0	credit_car	7083.33	0	0	7	0	216	0.6	19
9	0	C4	1.210000038	0	0	7/12/1996 0:00	< 1 year	660-678	OWN	3	credit_car	6666.67	0	44	15	1	27185	16.1	29
10	0	B5	8.029999733	0	0	8/17/1995 0:00	4 years	679-713	MORTGAGE	1	debt_cons	4000	0	0	6	0	28329	48.6	6
11	0	B3	11.93000031	0	0	2/19/1995 0:00	2 years	714-749	MORTGAGE	1	home_im	15000	0	0	16	0	60568	39.2	38
12	0	A4	5.550000191	0	0	6/13/1996 0:00	< 1 year	714-749	MORTGAGE	0	home_im	15000	0	0	12	0	40934	26.3	39
13	0	A2	2.289999962	0	0	10/6/1997 0:00	8 years	750-779	MORTGAGE	0	debt_cons	10000	0	0	8	0	8379	16.9	16
14	0	B1	14.53999996	0	0	9/21/2000 0:00	< 1 year	750-779	MORTGAGE	1	small_bus	2083.33	0	0	10	0	3660	7.8	13
15	0	C1	0	1	0	2/11/1997 0:00	2 years	679-713	MORTGAGE	0	home_im	16666.67	19	0	5	0	0	0	8
16	0	C4	18.63999939	0	0	4/5/1993 0:00	1 year	679-713	OWN	0	credit_car	2500	0	0	10	0	15840	47.1	12
17	0	A5	14.36999989	0	0	2/9/1992 0:00	7 years	780+	MORTGAGE	0	credit_car	6166.67	0	0	15	0	6844	14.4	29
18	0	A3	3	0	0	3/29/1989 0:00	< 1 year	750-779	RENT	0	education	666.67	0	0	4	0	1321	16.5	4
19	0	A5	14.77999973	0	0	6/27/1995 0:00	4 years	750-779	RENT	0	vacation	2666.67	0	0	11	0	4737	23.9	22
20	0	B2	9.960000038	0	0	1/7/1999 0:00	1 year	714-749	RENT	0	credit_car	6083.33	0	0	21	0	23489	37.6	28
21	0	C1	10.69999981	0	0	9/17/2003 0:00	< 1 year	679-713	RENT	0	small_bus	2281.33	0	0	4	0	3534	54.4	4
22	0	C2	4.050000191	0	0	1/7/2000 0:00	2 years	750-779	RENT	0	small_bus	4000	0	0	5	0	2422	23.3	[illegible]
23	0	A2	3.829999924	0	0	7/21/2000 0:00	7 years	750-779	MORTGAGE	0	vacation	7916.67	0	0	8	0	3660	[illegible]	16
24	0	A2	0	0	0	12/25/1987 0:00	8 years	750-779	MORTGAGE	0	home_im	12500	0	0	2	0	[illegible]	19.5	19
25	0	B5	16.44000053	0	0	12/26/2002 0:00	< 1 year	679-713	RENT	1	education	1125	0	0	10	[illegible]	2864	41.1	16
26	0	B2	10	2	0	4/1/2003 0:00	1 year	714-749	MORTGAGE	0	other	20833.33	5	-9.9E+07	[illegible]	0	14354	36.6	7
27	0	F4	12.56999969	0	0	8/24/2003 0:00	1 year		RENT	1	debt_cons	4350	0	[illegible]	12	0	3075	92.3	13
28	0	C2	17.12000084	1	0	2/25/1969 0:00	3 years	714-749	RENT	2	small_bus	5000	[illegible]	0	10	0	17214	8.1	24
29	0	C5	2.039999962	0	0	5/25/2004 0:00	< 1 year	660-678	RENT	2	credit_car	1666.67	24	0	3	0	1153	75.8	4
30	0	E1	10.14999962	0	0	1/1/1996 0:00	6 years	679-713	MORTGAGE	0	credit_car	12083.33	-9.9E+07	-9.9E+07	17	0	41674	74.1	26
31	0	D3	15	0	0	4/7/1995 0:00	9 years	679-713	OWN	3	credit_car	6250	78	0	9	0	43039	93.4	24

lendingClubScored4.5-indx

Associated Loan Descriptions

Hi, thanks for considering my request. I'm a student in Southern California. I have a great credit score. I will use this loan to pay my rent, books and tuition expenses. I've secured a part

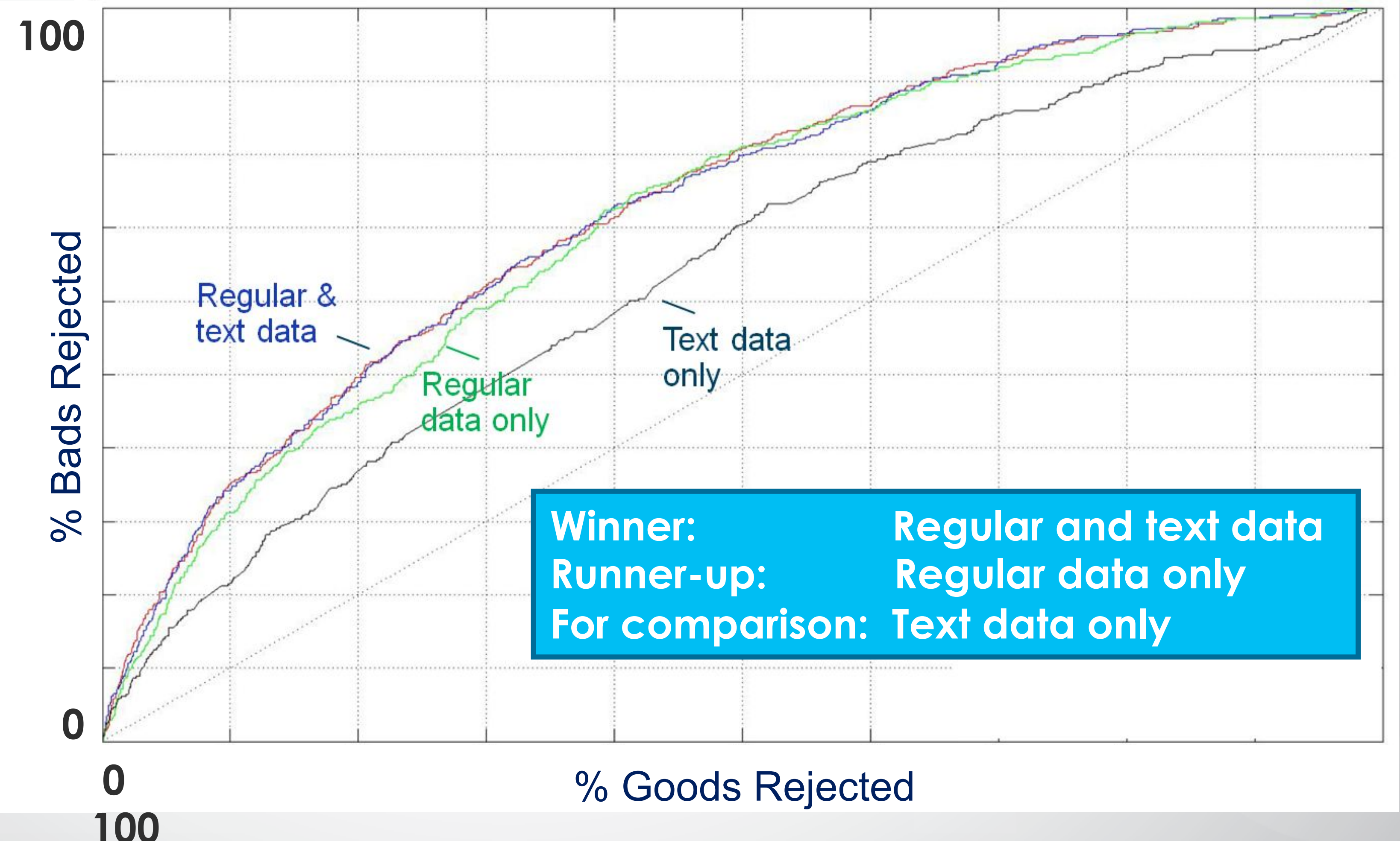
Prospect ID #5340164:
"I need this loan to pay off higher rate credit card debt - fixed rate at 15%, that's the only card I use"

Generate traditional variables

Extract keywords and topics

"Semantic Scorecard": Combine traditional variables and text-based features in a comprehensible model

Predictive Value of Text Data



CONAMIC2015
EL FORO DE LAS FINANZAS POPULARES

Top Predictors From Automatic Variable Selection

Lead to Unjustifiable Decisions

Traditional variables (black)
Keywords indicating elevated risk (red)
Keywords indicating reduced risk (green)

Rank	Variable Name
1	Credit Bureau Grade
2	Inquiries During Last 6 Months
3	Monthly Income
4	Months Since Last Record
5	need
6	Revolving Credit Balance
7	baby
8	Loan Purpose
9	business
10	would
11	Revolving Line Utilization
12	qualify
13	Length of Employment
14	sincerely
15	credit
16	provided
17	sales
18	job
19	open
20	stable
21	payday
22	card
23	www
24	Open Credit Lines
25	Total Credit Lines
26	shop

Gained New Insights Into Risk Topics

Results Using "Latent Dirichlet Allocation"

Reduced risk topic: "Credit card consolidation"

debt, free, consolidating, consolidated, card, credit, revolving, paying, payoff, sooner, quicker, clear, accumulated, accrued, completely, …

Elevated risk topic: "Business-related items"

business, equipment, sales, capital, store, marketing, experience, location, expand, owner, retail, advertising, partner, inventory, products, profit, shop, restaurant, ...

Discussion

- Machine learning, text- and social network analytics can yield deeper insights and stronger predictions, by combining traditional and novel data sources.

- Yet to make more profitable and justifiable decisions requires careful results interpretation and robust, explainable models.

- Domain expertise, combined with special methods and tools supporting interpretation, are of utmost importance for harnessing the potential of big data and machine learning.