

Unleashing Modern AI in Pharmaceutical and Chemical Industries

Mikhail Golovnya
Senior Advisory Data Scientist

AI Suitcase



- **Reactive Machines:**

- These are basic rule-based systems that operate based on predefined rules.

- **Expert Systems:**

- These are computer systems that mimic the decision-making ability of a human expert in a specific domain.

- **Machine Learning (ML) Systems:**

- ML is a subset of AI that focuses on developing algorithms and models that enable computers to learn from data.
- Types of ML systems include **supervised learning (PA)**, unsupervised learning, and reinforcement learning.

- **Neural Networks:**

- Inspired by the human brain, neural networks are a key component of many AI systems.

- **Narrow/Generative LLM AI (Weak AI):**

- These AI systems are designed and trained for a specific task or a narrow set of tasks.
- Examples include virtual personal assistants, image recognition software, and language translation services.

- **Limited Memory:**

- These AI systems can learn from historical data to make better decisions.
- Self-driving cars often use limited memory AI to navigate based on past experiences.
- Can be integrated into robots to enable them to learn and interact with the environment.

- **Self-aware AI:**

- This refers to hypothetical AI systems with self-awareness and consciousness.

- **Theory of Mind:**

- This is a more advanced form of AI that can understand human emotions, beliefs, intentions, and thoughts.

- **General AI (Strong AI):**

- General AI systems can understand, learn, and apply knowledge across diverse domains.
- They can perform any intellectual task that a human being can do.

- **Superintelligent AI:**

- This is a theoretical AI that surpasses human intelligence in every aspect.

AI Suitcase



- **Reactive Machines:**

- These are basic rule-based systems that operate based on predefined rules.

- **Expert Systems:**

- These are computer systems that mimic the decision-making ability of a human expert in a specific domain.

- **Machine Learning (ML) Systems:**

- ML is a subset of AI that focuses on developing algorithms and models that enable computers to learn from data.
- Types of ML systems include **supervised learning (PA)**, unsupervised learning, and reinforcement learning.

Numbers In
Numbers Out

- **Neural Networks:**

- Inspired by the human brain, neural networks are a key component of many AI systems.

- **Narrow/Generative LLM AI (Weak AI):**

- These AI systems are designed and trained for a specific task or a narrow set of tasks.
- Examples include virtual personal assistants, image recognition software, and language translation services.

Text In
Text Out

- **Limited Memory:**

- These AI systems can learn from historical data to make better decisions.
- Self-driving cars often use limited memory AI to navigate based on past experiences.
- Can be integrated into robots to enable them to learn and interact with the environment.

- **Self-aware AI:**

- This refers to hypothetical AI systems with self-awareness and consciousness.

- **Theory of Mind:**

- This is a more advanced form of AI that can understand human emotions, beliefs, intentions, and thoughts.

- **General AI (Strong AI):**

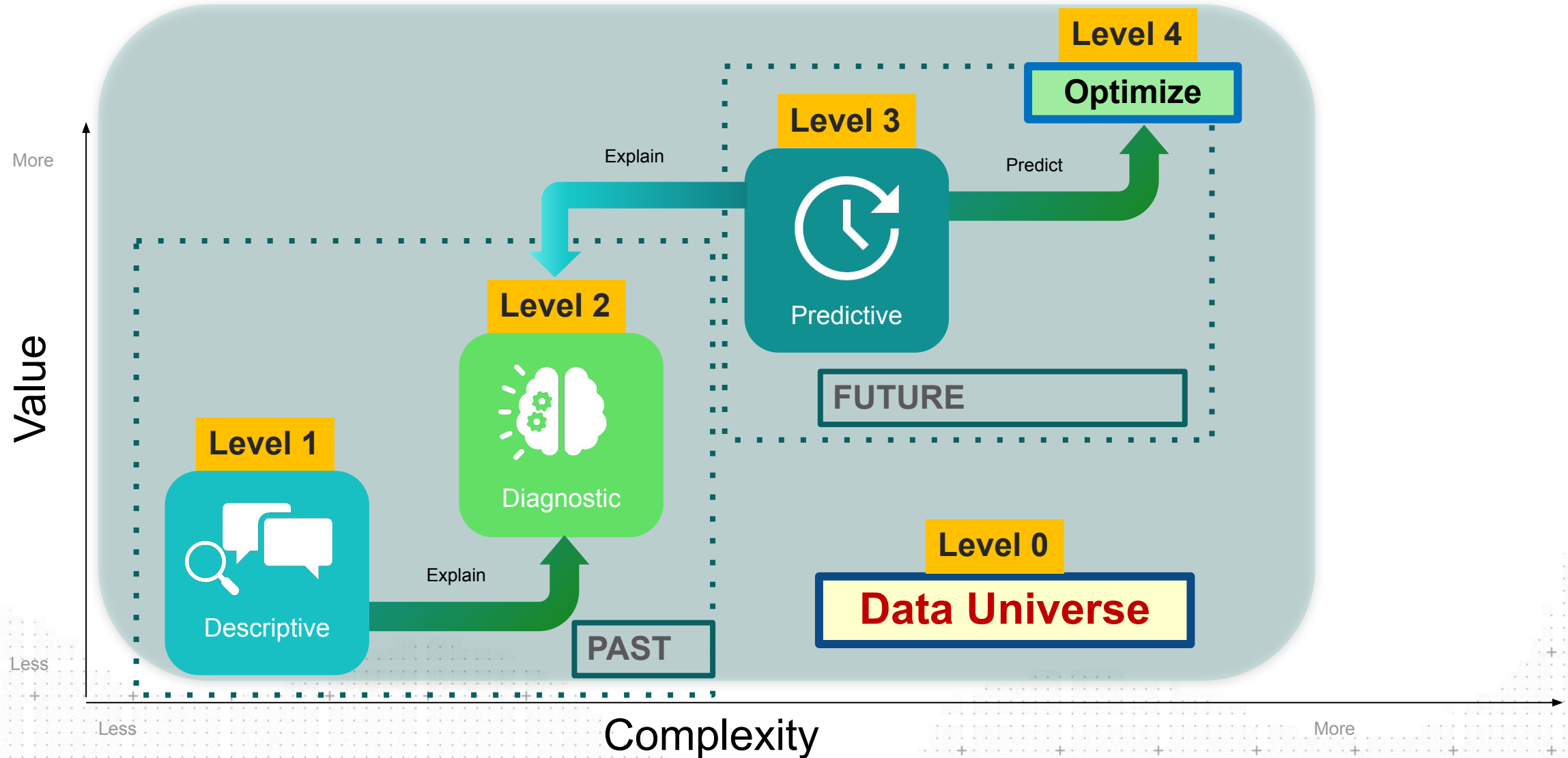
- General AI systems can understand, learn, and apply knowledge across diverse domains.
- They can perform any intellectual task that a human being can do.

- **Superintelligent AI:**

- This is a theoretical AI that surpasses human intelligence in every aspect.

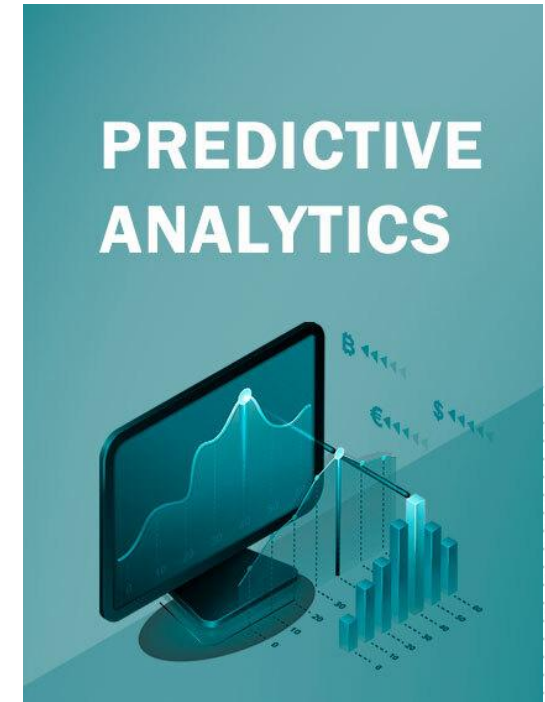
AGI / Science Fiction / Religion

Levels of Data Analytics



Predictive AI: The Essence and Utility

- ▶ **Predict** [**something**] based on [**something else**]
- ▶ **Direct Application (Level 3)**: use a PA model **to predict** the outcome of interest
 - Emphasis on the **accuracy** of predictions
- ▶ Link to **Diagnostic Analytics (Level 2)**: use a PA model **to explain** why such and such predictions are made
 - Emphasis on the **explanation** of predictions
- ▶ Link to **Response Optimization (Level 4)**: use a PA model **to discover optimal inputs** to achieve a desired outcome
 - Emphasis on the **discovery** of the optimal inputs
- ▶ **Key tradeoff**: simple and easy to explain models tend to be less accurate!

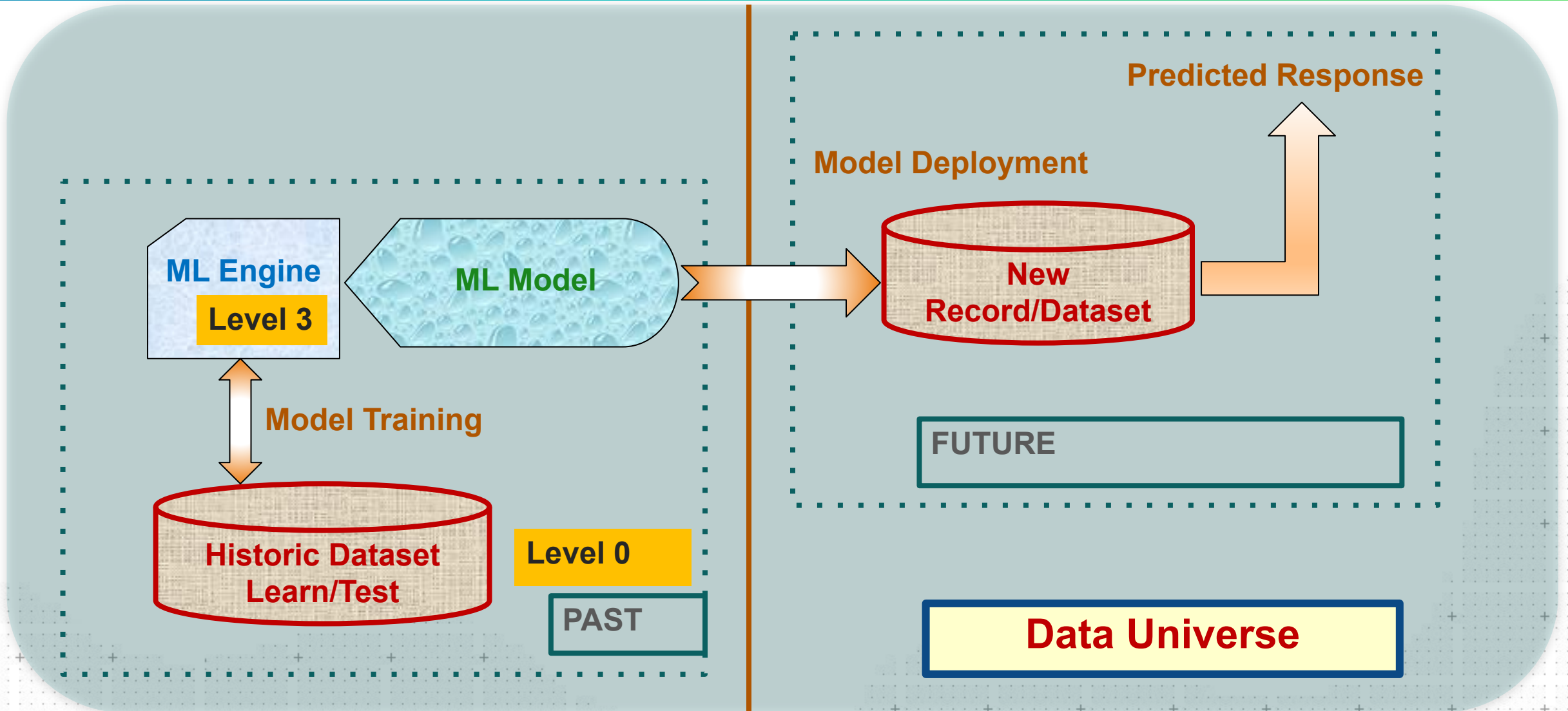


Traditional Predictive AI (a.k.a. Machine Learning)

**Numbers In
Numbers Out**

- ▶ To engage in conventional predictive AI, we **must have**:
- ▶ **Clear Problem Definition** – predict **something** based on **something else**
- ▶ **Historical Dataset** – a fixed grid of **observations** by **variables**
- ▶ **Response/Target variable** – what we want to predict (**all sides** must be captured by the historical data)
 - **Regression Problem** – the target variable is **continuous** (Energy, Score, etc.)
 - **Classification Problem** – the target variable is **categorical** (Region, Color, etc.)
- ▶ **Predictor variables** – variables used to predict the response (any combination of continuous and categorical variables)

Machine Learning Process



Classical vs Modern Approaches in ML



▶ **Classical approach:** I will dictate what the model structure is, the data will be used to fit in the details

- Multiple linear regression
- Logistic regression
- PLS, Lasso, etc.

**Dictate to
the data**

**Easy to explain
Easy to fail
Compute Light**

▶ **Modern approach:** I will let the data tell me what the model is

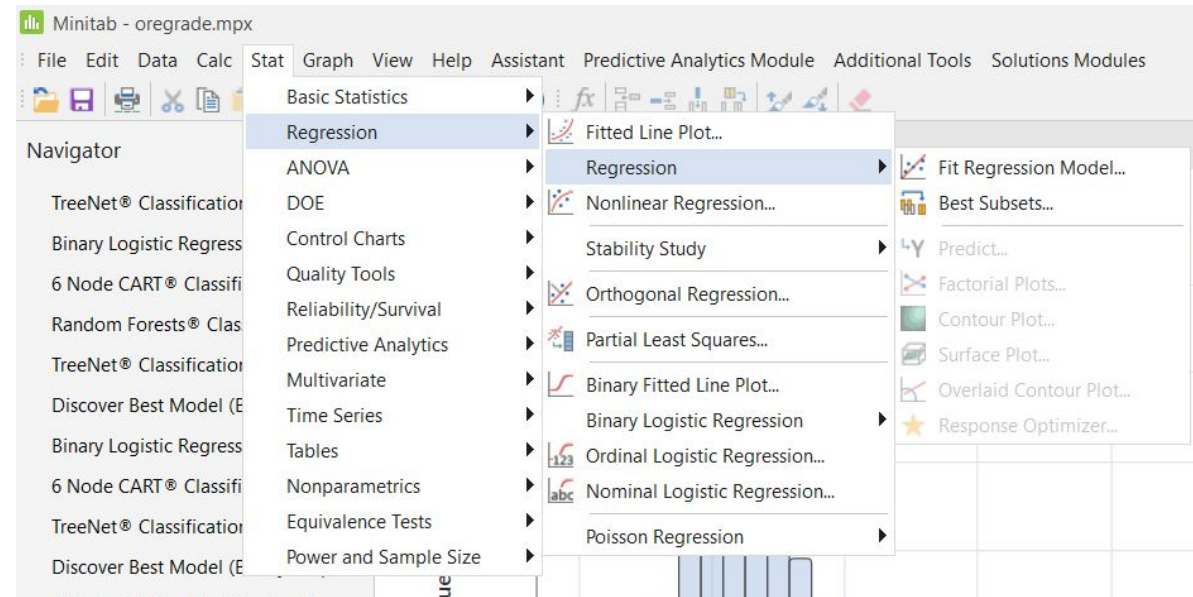
- **CART** – **C**lassification **A**nd **R**egression **T**rees
- **MARS** – **M**ultivariate **A**daptive **R**egression **S**plines
- **RF** – **R**andom **F**orest
- **TN** – **T**ree **N**et (a.k.a. Stochastic Gradient Boosting)

**Learn from
the data**

**Hard to explain
Hard to fail
Compute Intense**

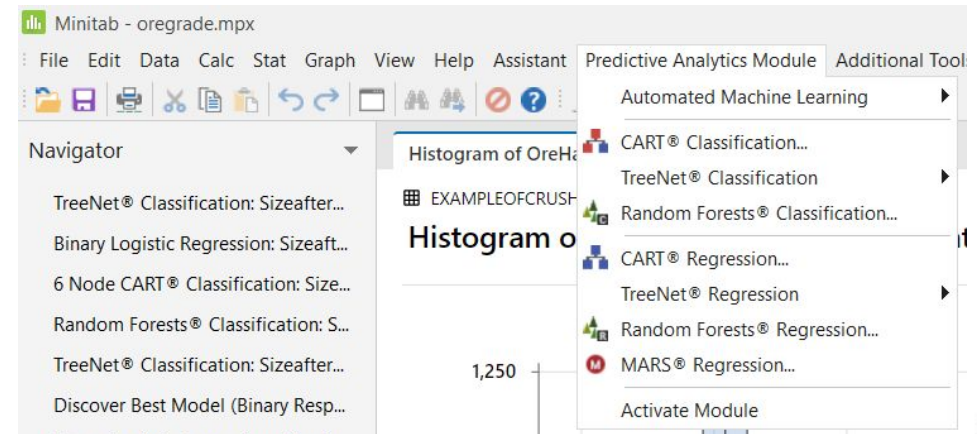
Classical Algorithms: Common Features

- ▶ **Parametric** in nature (well-defined equations)
- ▶ Based on certain statistical/distributional **assumptions**
- ▶ **Assume expert knowledge of the problem**
- ▶ The model shape (response surface) is **dictated** by the analyst
- ▶ All observations contribute **globally**
- ▶ Focus solely on reducing the **variance** of estimates
- ▶ Give **baseline accuracy**
- ▶ Work great on **small** datasets



Modern Algorithms: Common Features

- ▶ **Non-parametric** in nature (no equations)
- ▶ **Do not rely** on statistical/distributional assumptions
- ▶ **Learn from the data**
- ▶ The model shape (response surface) is **developed dynamically** by scrutinizing the data
- ▶ Observations contribute **locally**
- ▶ Focus on the **bias/variance tradeoff** by giving a **range** of models
- ▶ Give **superior accuracy**
- ▶ Work great on **medium to large** datasets




Uses of Modern ML in Pharma and Chemical Industries

- ▶ Drug Discovery and Development
 - **QSAR** (Quantitative Structure-Activity Relationship) Modeling based on molecular descriptors and fingerprints
 - **Virtual Screening** (Hit Discovery) – to filter large compound libraries to identify potential candidates
 - **ADMET Predictions** (Absorption, Distribution, Metabolism, Excretion, Toxicity)
- ▶ Chemical Manufacturing and Process Optimization
 - Process Control and Yield Prediction
 - Fault Detection and Root Cause Analysis
- ▶ Clinical Research and Personalized Medicine
 - Ranking genomic, proteomic, or metabolomic features associated with treatment response
 - Clinical trial optimization – predicting patient dropout, adverse effects, treatment response
- ▶ Toxicology and Environmental Risk Assessment
- ▶ Pharmaceutical Formulation, Polymer and Material Design



Published Uses of ML in Recent Research

Industry/Application	Method	Highlights & Outcomes	
Pharmaceutical shipping optimization	LightGBM (GBM)	96% accuracy in mode selection SpringerLink +1	
Chemical functional use prediction (QSUR)	Random Forest	Screened ~6,400 chemicals; strong performance actiac.org	
Sulphonation product quality modeling	Random Forest	MAE 0.089, corr. 0.978 arXiv	
Soft-sensing (detergent quality)	Random Forest	Top 3 features achieved strong prediction arXiv	
Iron oxide synthesis (phase & size control)	Random Forest	96% phase, 81% size accuracy; parameter recommendations arXiv	
Zinc filtration moisture modeling	Random Forest	Outperformed SVR in predictive accuracy arXiv	

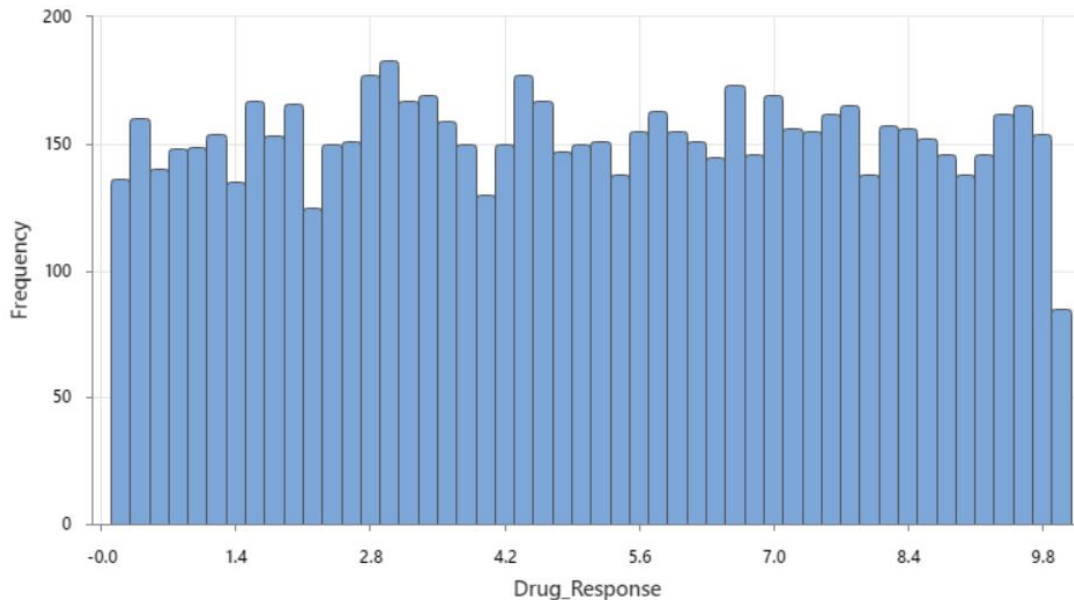


RESEARCH

Case Study 1: Personalized Medicine

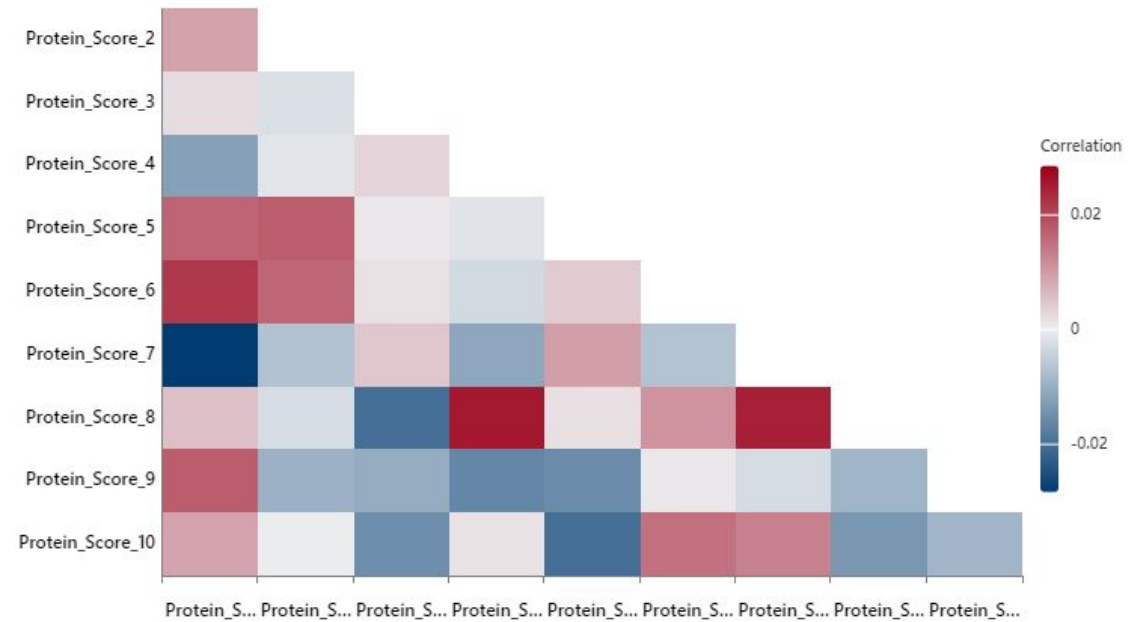
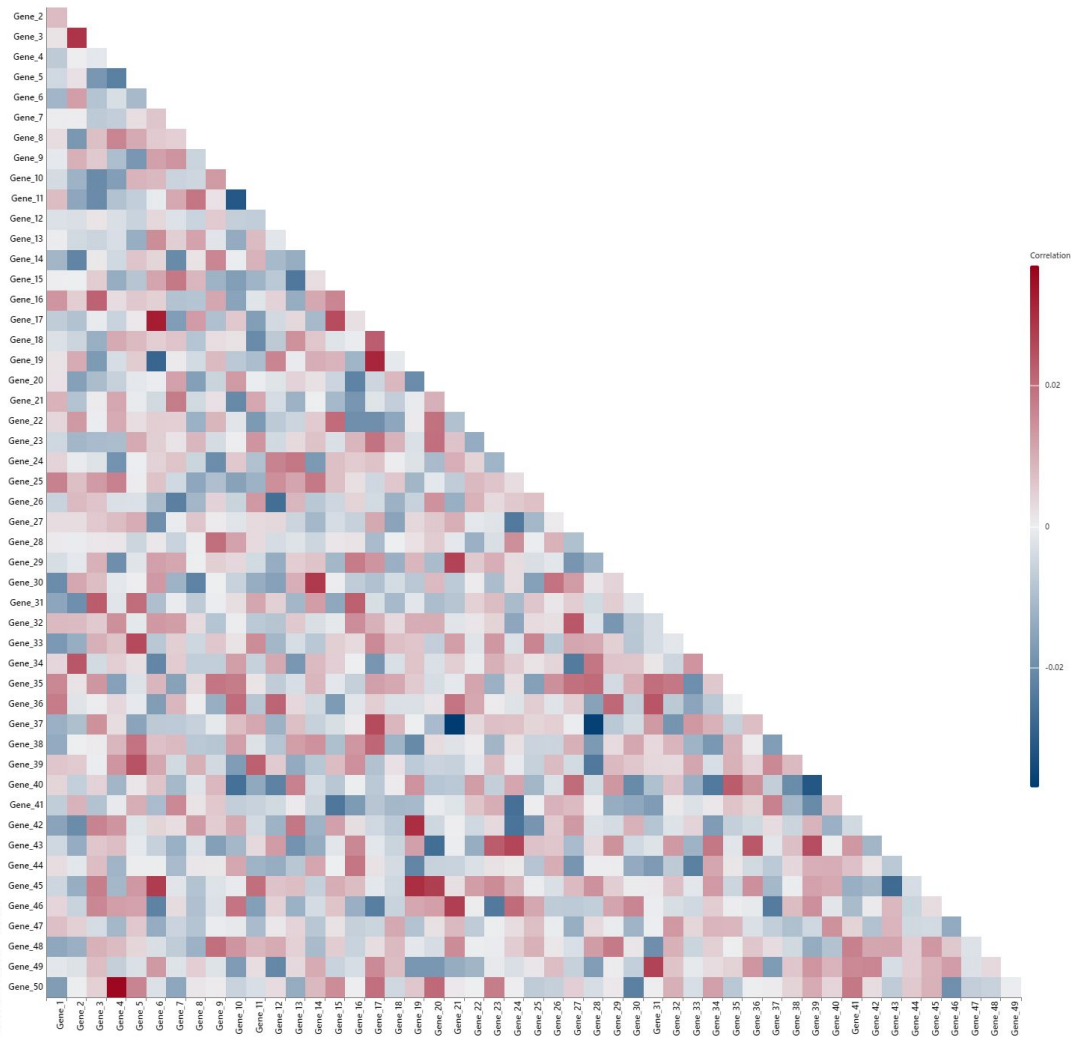
Personalized Medicine Research

C70	C71	C72-T	C73-T	C74-T	C75-T	C76-T	C77
Protein_Score_9	Protein_Score_10	Pathway_1	Pathway_2	Pathway_3	Pathway_4	Pathway_5	Drug_Response
0.194099	0.589134	Low	High	Moderate	Moderate	Low	9.44071
0.391995	0.223219	Low	Low	High	High	Low	8.91781
0.952073	0.440213	High	Low	High	Low	Low	9.06261
0.867519	0.323174	Moderate	Moderate	High	Low	High	9.03005
0.669718	0.769588	Low	Low	High	Moderate	High	8.75017
0.180836	0.519773	High	Low	Low	High	High	7.13843
0.779635	0.725349	Moderate	High	Low	Low	Moderate	0.93278



- ▶ A dataset containing 7643 patient profiles
 - 50 key gene expressions
 - 10 mutation indicators
 - 10 protein scores
 - 5 pathways (low, moderate, high)
- ▶ Want to predict continuous drug response
- ▶ This is a regression problem with 60 continuous and 15 categorical predictors

Level 1: Descriptive



Conventional Regression Model

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)	Test S	Test R-sq
2.84268	1.31%	0.97%	0.61%	2.82412	0.00%

- ▶ Unfortunately, nothing useful/significant is detected!
- ▶ **Our main challenge:** how do we know whether we should continue looking for signal or abandon the project?
- ▶ Let's unleash the power of modern data-driven PA algorithms

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	6.081	0.267	22.81	0.000	
Gene_1	0.0272	0.0134	2.03	0.042	1.00
Gene_9	-0.0231	0.0135	-1.71	0.087	1.00
Gene_11	-0.0215	0.0135	-1.60	0.110	1.00
Gene_15	-0.0318	0.0135	-2.35	0.019	1.00
Gene_17	-0.0199	0.0136	-1.46	0.144	1.00
Gene_23	-0.0302	0.0135	-2.23	0.026	1.00
Gene_26	-0.0233	0.0134	-1.74	0.082	1.00
Gene_29	-0.0248	0.0134	-1.85	0.065	1.00
Gene_31	-0.0259	0.0135	-1.92	0.054	1.00
Gene_35	0.0217	0.0135	1.61	0.107	1.00
Gene_39	-0.0328	0.0133	-2.46	0.014	1.00
Gene_41	-0.0293	0.0134	-2.18	0.029	1.01
Protein_Score_1	0.328	0.136	2.41	0.016	1.00
Protein_Score_10	-0.324	0.136	-2.38	0.017	1.00
Mutation_3					
1	0.1347	0.0779	1.73	0.084	1.00
Mutation_6					
1	-0.1779	0.0778	-2.29	0.022	1.00
Pathway_2					
Low	0.2110	0.0953	2.21	0.027	1.34
Moderate	0.0498	0.0955	0.52	0.602	1.34

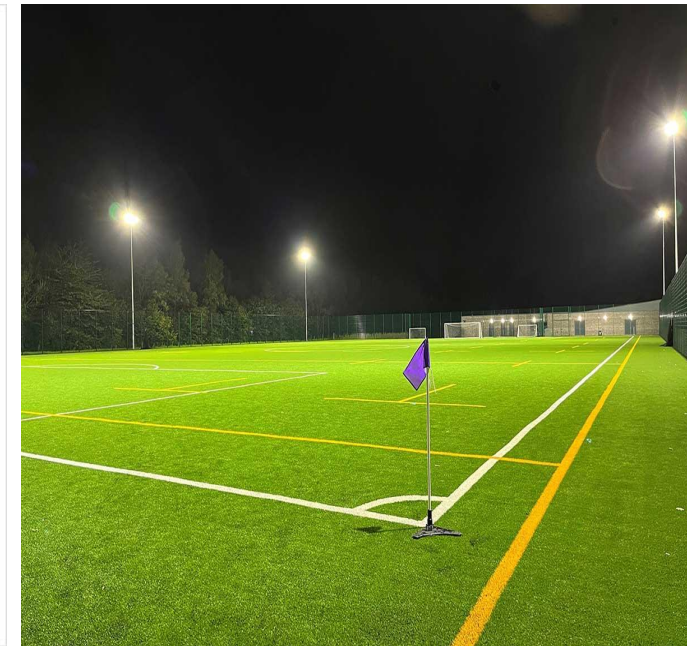
Modern PA Model (TreeNet)

Model Summary

Total predictors 75
Important predictors 5
Number of trees grown 300
Optimal number of trees 1

Statistics

	Training	Test
R-squared	0.13%	0.00%
Root mean squared error (RMSE)	2.8347	2.8520
Mean squared error (MSE)	8.0354	8.1340
Mean absolute deviation (MAD)	2.4571	2.4669
Mean absolute percent error (MAPE)	1.6172	1.5236

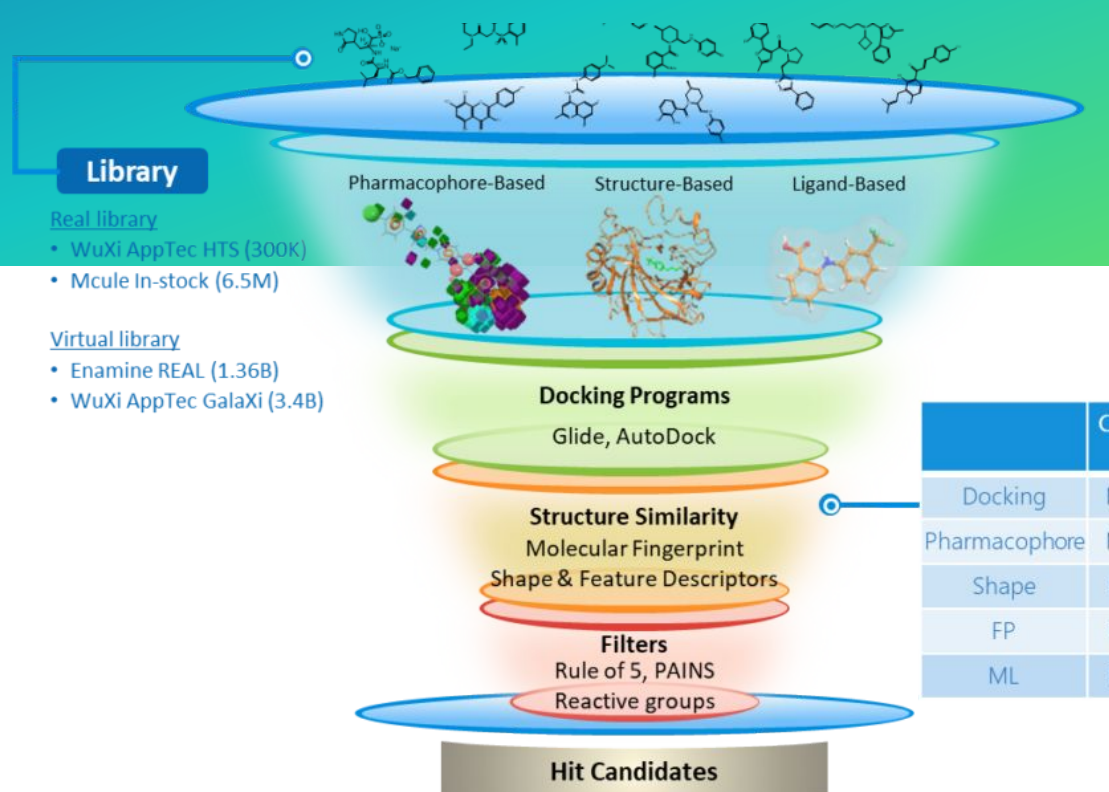
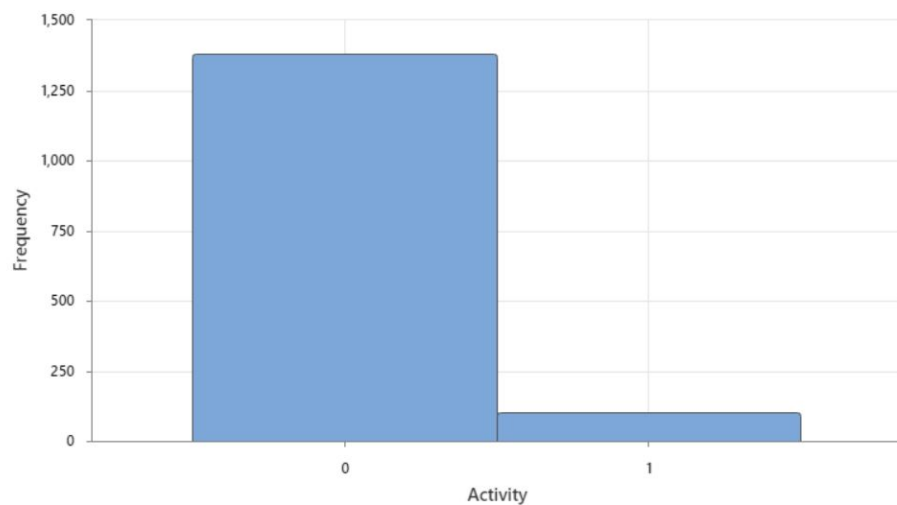


- ▶ We have a **final confirmation**: there is no empirical evidence to support our initial hypothesis that the drug response can be different depending on the genomic and proteomic profile of the patient
- ▶ Further research might focus on **finding different genes and proteins** by collecting more data

Case Study 2: Virtual Screening Using Molecular Fingerprints

Virtual Screening Dataset

C1020	C1021	C1022	C1023	C1024	C1025
1019	1020	1021	1022	1023	Activity
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	1
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0



- ▶ A dataset containing 1484 binary fingerprints of different compounds
 - 1024 indicators for each fingerprint
- ▶ Want to predict binary activity response for further screening
- ▶ This is a binary classification problem with 1024 binary predictors

Conventional Logistic Regression Model

Binary Logistic Regression: Activity versus 0, 1, 2, 3, 4, 5, 6, 7, 8, 9,

** ERROR * The training data set must contain all levels of the categorical predictors.*

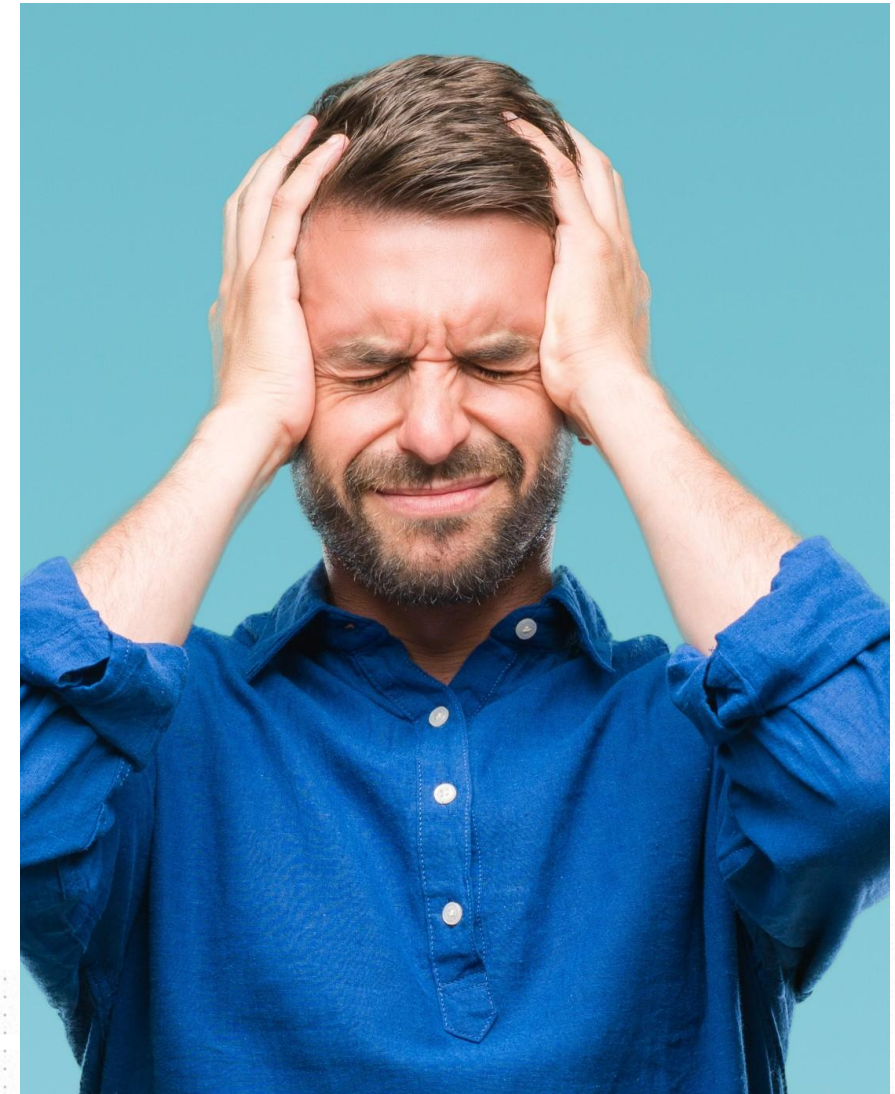
Binary Logistic Regression: Activity versus 0, 1, 2, 3, 4, 5, 6,

** ERROR * Continuous predictors must have more than one distinct value.*

Binary Logistic Regression: Activity versus 0, 1, 2, 3, 4, 5, 6,

** ERROR * Categorical predictors must have more than one distinct value.*

- ▶ Unfortunately, we struggle with the **data requirements!**
- ▶ **Solution:** Unleash modern PA on it!



Modern PA Model: TreeNet

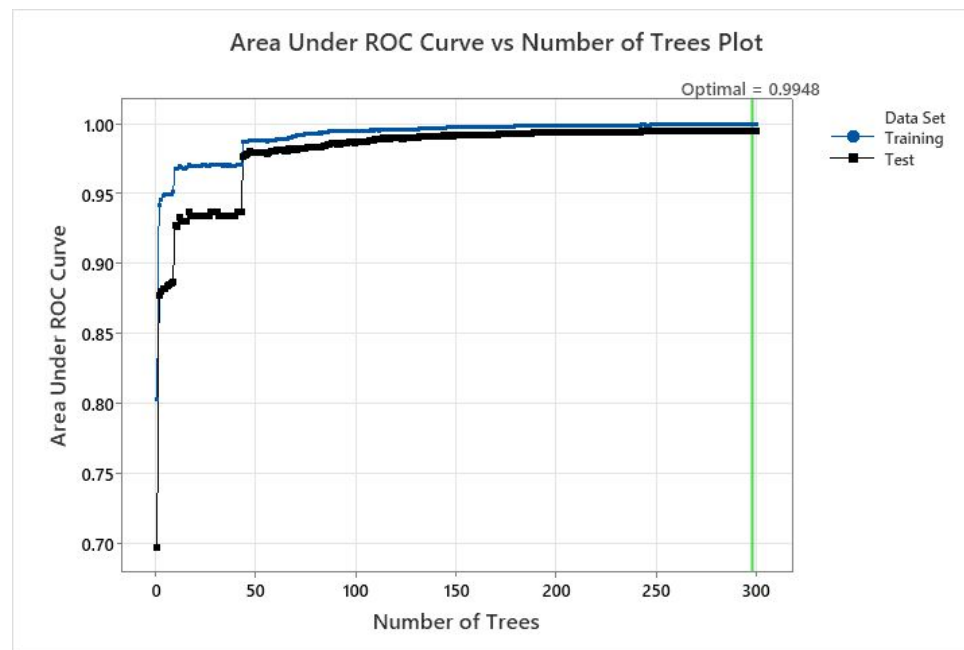
Binary Response Information

Variable Class	Training		Test	
	Count	%	Count	%
Activity 1 (Event)	74	7.07	30	6.86
0	973	92.93	407	93.14
All	1047	100.00	437	100.00

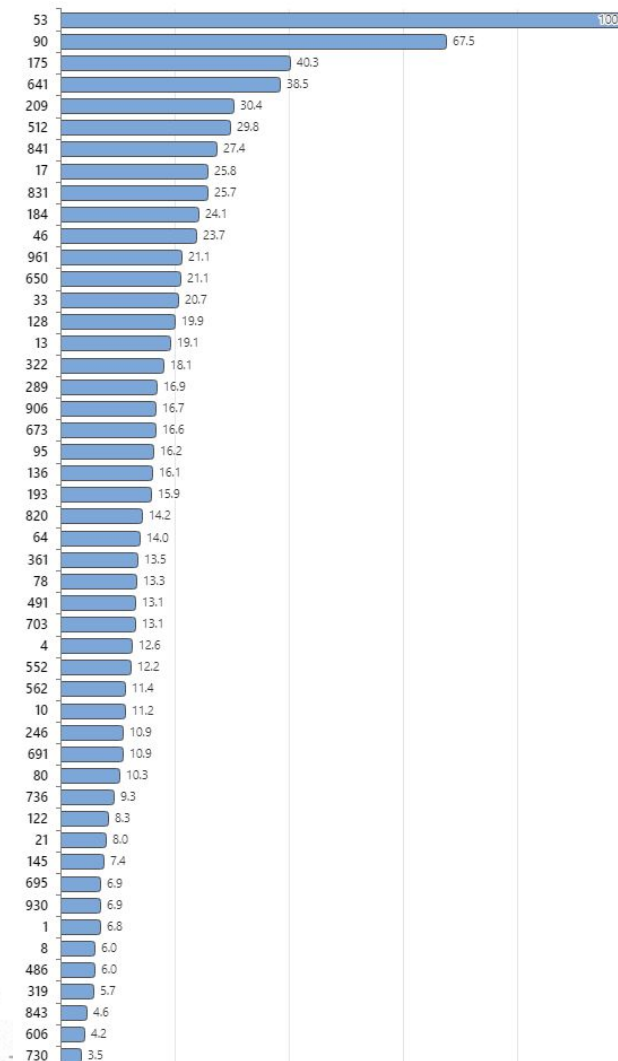
Model Summary

Total predictors	1024
Important predictors	58
Number of trees grown	300
Optimal number of trees	298

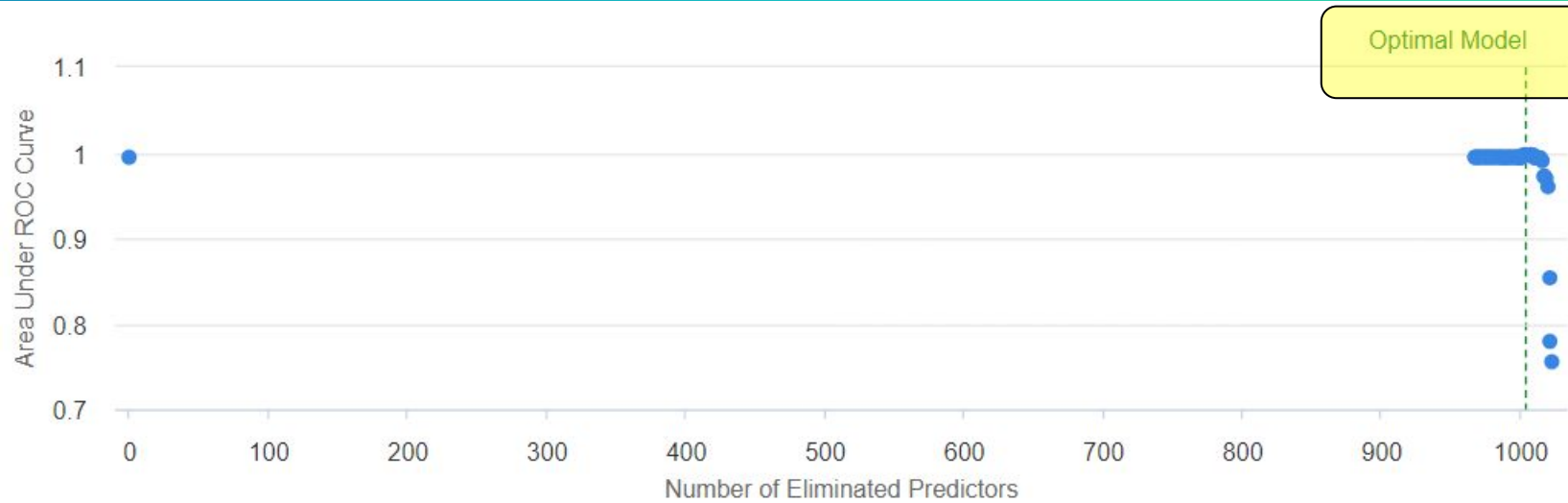
Statistics	Training	Test
Average -loglikelihood	0.0393	0.0553
Area under ROC curve	0.9995	0.9948
95% CI	(0.9988, 1)	(0.9884, 1)
Lift	10.0000	9.3333
Misclassification rate	0.0086	0.0137



Relative Variable Importance



Predictor Discovery AI



Model	Optimal Number of Trees	Area Under ROC Curve	Number of Predictors	Eliminated Predictors
51	150	0.999999	8	55
52	275	0.972072	7	841
53	291	0.972891	6	13
54	299	0.970147	5	209
55	144	0.960278	4	175
56	2	0.853726	3	512
57	1	0.779402	2	90
58	1	0.75561	1	641

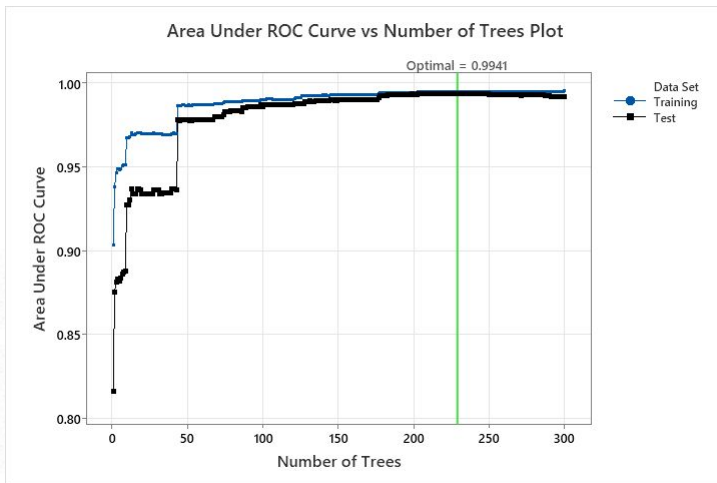
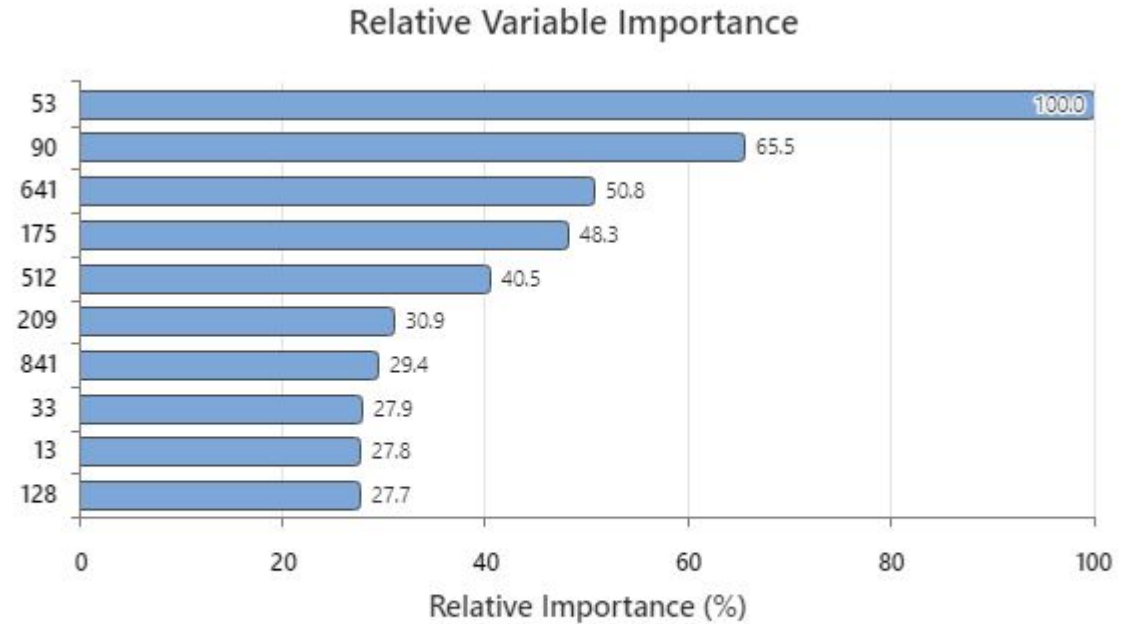


Discovering the 10-Marker Model

Model Summary

Total predictors 10
 Important predictors 10
 Number of trees grown 300
 Optimal number of trees 229

Statistics	Training	Test
Average -loglikelihood	0.0527	0.0628
Area under ROC curve	0.9949	0.9941
95% CI	(0.9913, 0.9985)	(0.9880, 1)
Lift	9.4206	9.6125
Misclassification rate	0.0162	0.0252



▶ The entire discovery process is finished in minutes!

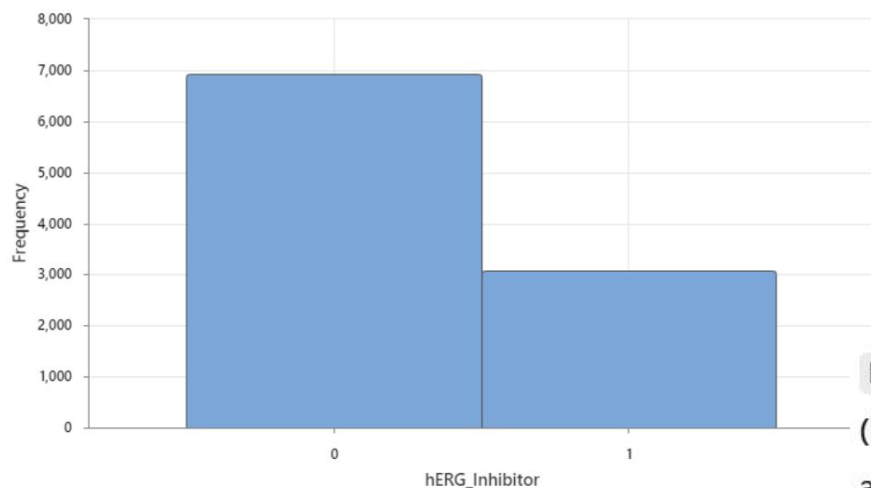
Case Study 3: Predicting Drug Toxicity

Drug Toxicity Dataset

C1-T	C2-T	C3	C4	C5	C6	C7	C8	C9	C10
compound_id	Chemotype	MW	LogP	LogD_7_4	TPSA	HBD	HBA	RotatableBonds	AromaticRings
CMPD00001	C	282.199	2.01064	-2.53610	77.851	0	4	6	2
CMPD00002	D	355.483	1.98273	-0.82796	42.140	2	7	6	0
CMPD00003	C	249.926	1.94915	1.94915	74.323	0	9	4	1
CMPD00004	D	223.309	2.54028	-0.56605	83.361	3	4	8	1
CMPD00005	A	400.597	1.29458	0.71939	42.529	0	3	8	2
CMPD00006	A	312.426	3.40643	3.40643	20.457	1	7	7	2



- ▶ A dataset containing 10000 drug trial compounds
 - 6 binary target responses related to toxicity and efficacy
 - 30 categorical and continuous predictors describing the molecular structure and chemical characteristics of each compound
- ▶ Want to understand what factors correlate with different types of toxicity



`hERG_Inhibitor` is a binary flag indicating whether a compound is predicted to inhibit the cardiac hERG (KCNH2) potassium channel. Inhibition of hERG reduces the I_{Kr} current, can prolong the QT interval, and is a well-known cardiotoxicity liability (linked to torsades de pointes).

Descriptive AI

Graph Builder

Gallery

See all the different ways to visualize your data.

Variables

Chemotype x MW x LogP x

LogD_7_4 x TPSA x HBD x HBA x

RotatableBonds x AromaticRings x

Gallery Histogram Probability Plot Boxplot Interval Plot Individual Value Plot Line Plot Pareto

Scatterplot

MW vs LogP by Chemotype

Binned Scatterplot

MW vs LogP

Bubble Plot

MW vs LogP by Chemotype

Matrix Plot

MW, LogP, LogD_7_4, TPSA, HBD, ...

Correlogram

MW, LogP, LogD_7_4, TPSA, HBD, ...

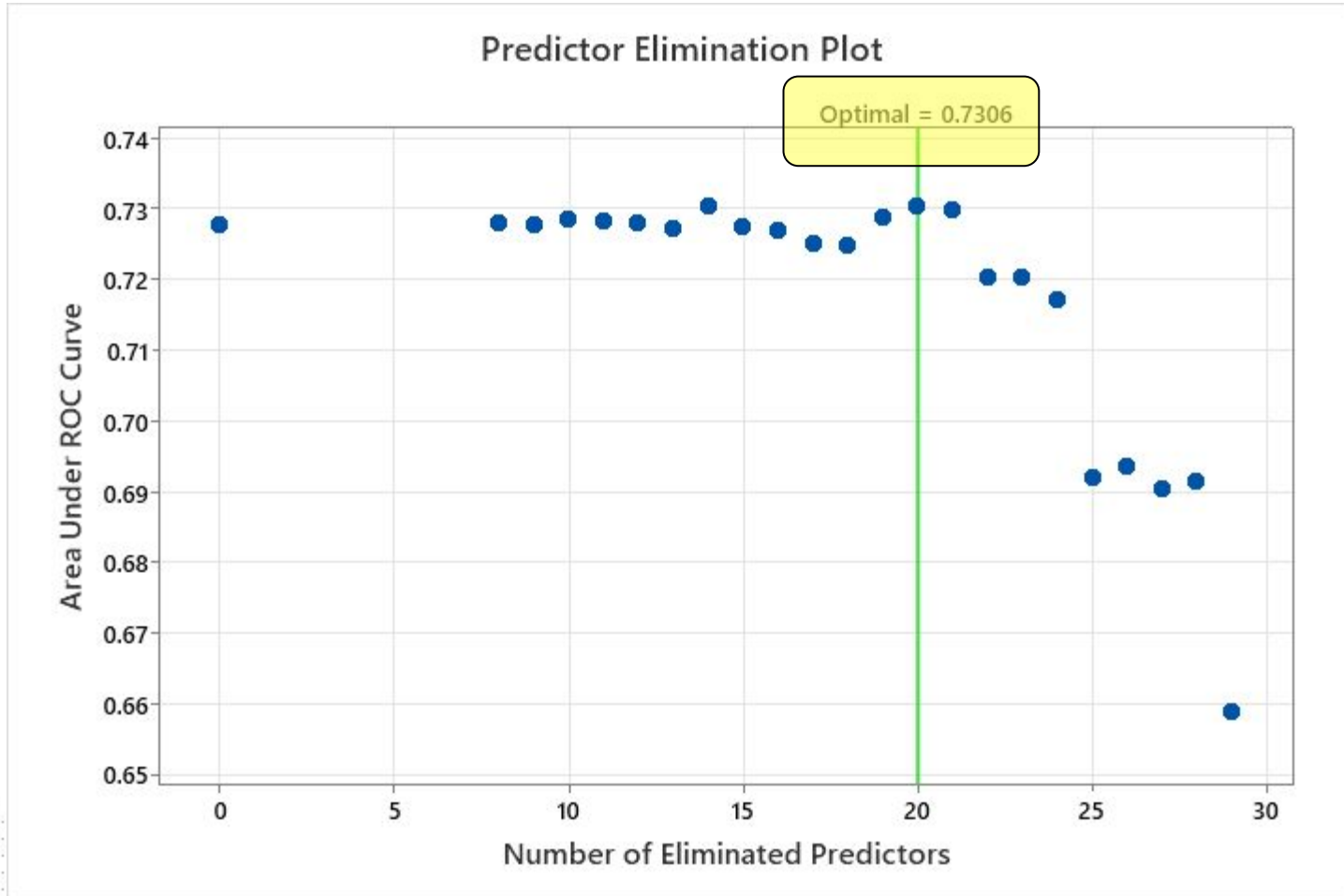
Heatmap

MW vs Chemotype

Chemotype A B C D

Help Reset Create Cancel

Predictor Discovery AI

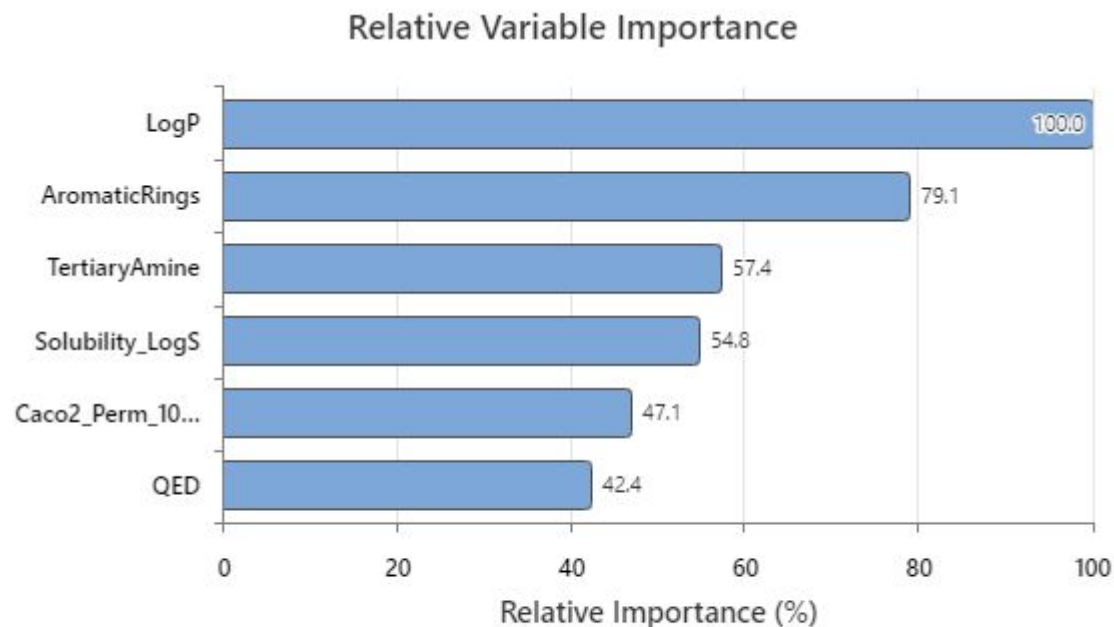
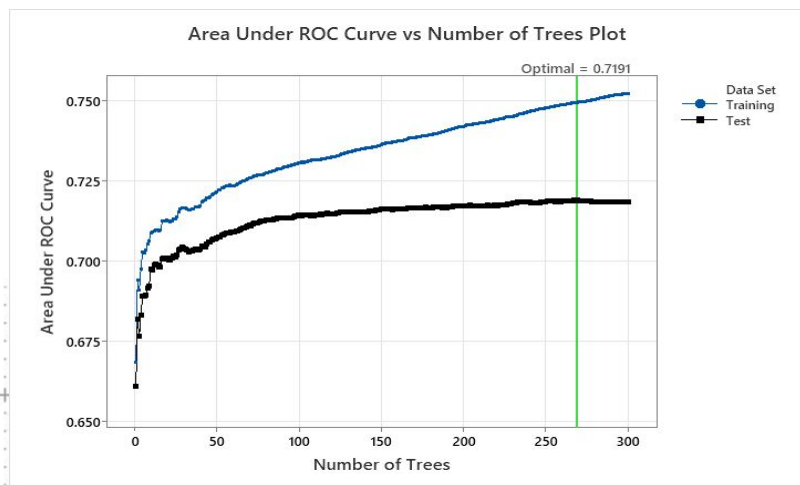


The Final Model Candidate

Model Summary

Total predictors 6
Important predictors 6
Number of trees grown 300
Optimal number of trees 269

Statistics	Training	Test
Average -loglikelihood	0.5324	0.5462
Area under ROC curve	0.7496	0.7191
95% CI	(0.7373, 0.7618)	(0.6993, 0.7388)
Lift	2.3734	2.1359
Misclassification rate	0.2631	0.2712

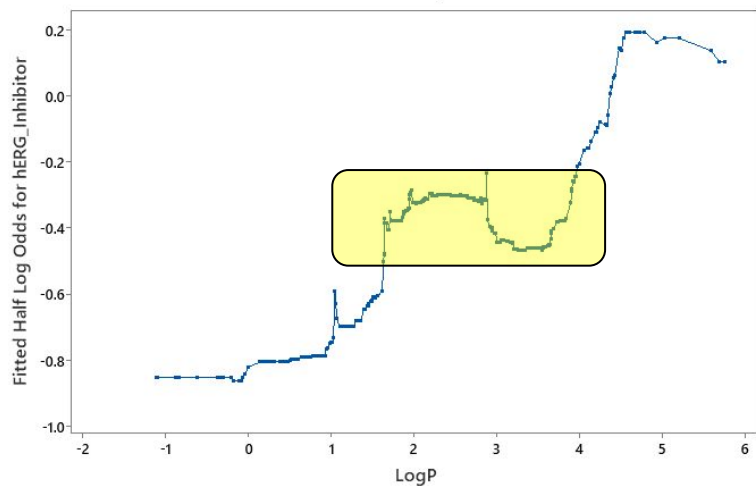


Predictors of Interest

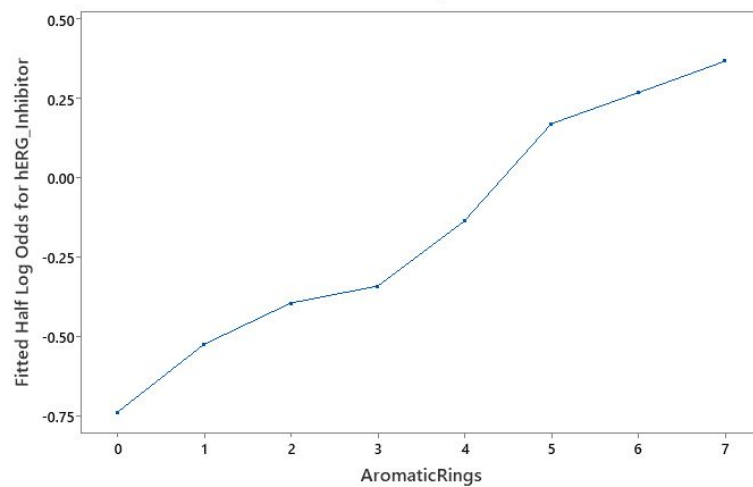
- ▶ **LogP** is a measure of a molecule's **lipophilicity** (how “fat-loving” or hydrophobic it is).
- ▶ **AromaticRings** (in the dataset) is an **integer count** of how many **aromatic rings** a compound has.
- ▶ **TertiaryAmine** (in the dataset) is a **binary flag (0/1)** that marks whether the compound contains a **tertiary amine functional group**.
- ▶ **Solubility_LogS** is the **base-10 logarithm of aqueous solubility expressed in mol/L** (molar solubility):
- ▶ **Caco2_Perm_10e6cms** is the **Caco-2 apparent permeability (P_{app})** measured across a Caco-2 cell monolayer, reported in **units of 10^{-6} to 10^{-6} cm/s**.
- ▶ **QED** stands for **Quantitative Estimate of Drug-likeness**. It's a single number in **[0, 1]** that summarizes how “drug-like” a **small-molecule** is based on a handful of simple physicochemical properties. Higher is more drug-like *by this heuristic*.

TN Model Interpretation

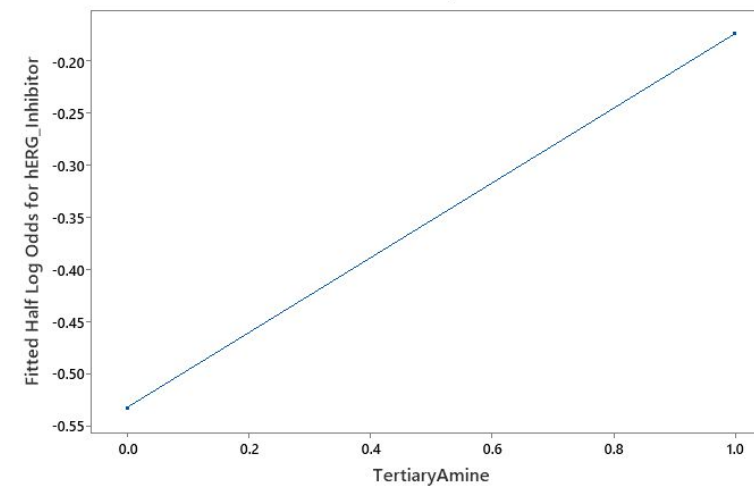
One Predictor Partial Dependence Plot



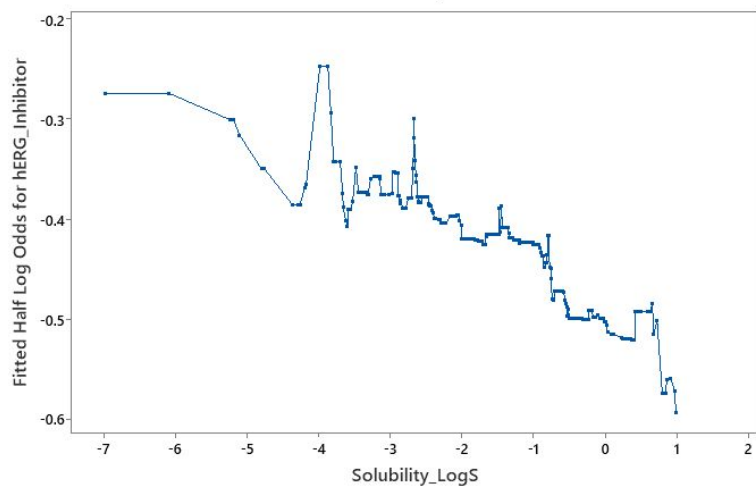
One Predictor Partial Dependence Plot



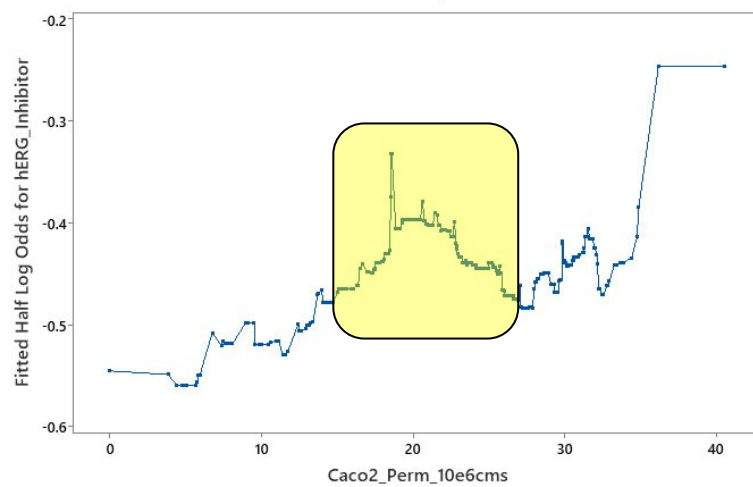
One Predictor Partial Dependence Plot



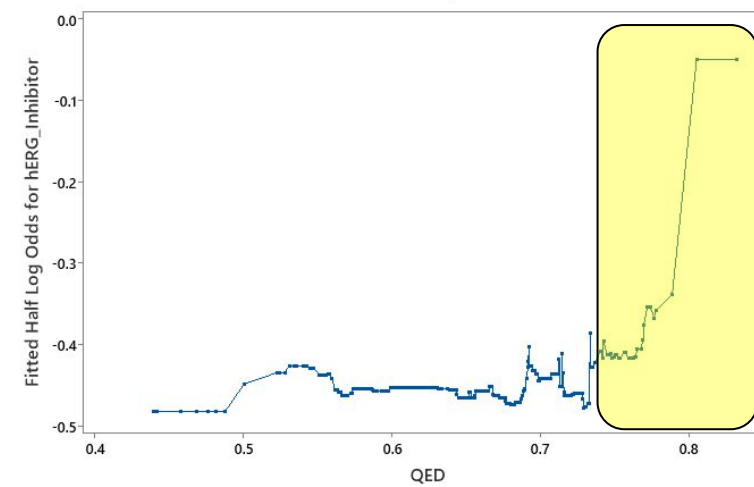
One Predictor Partial Dependence Plot



One Predictor Partial Dependence Plot



One Predictor Partial Dependence Plot



Logistic Regression Model

Model Summary

Deviance		Test			Test Area	
R-Sq	R-Sq(adj)	AIC	AICc	BIC	Area Under ROC Curve	Deviance Under ROC Curve
9.96%	9.89%	7798.17	7798.19	7846.15	0.7070	8.58%
						0.6935

Regression Equation

$$P(1) = \frac{\exp(Y')}{1 + \exp(Y')}$$

$$Y' = -4.372 + 0.3451 \text{ LogP} + 0.3506 \text{ AromaticRings} + 0.8924 \text{ TertiaryAmine} + 2.414 \text{ QED} - 0.0578 \text{ Solubility_LogS} + 0.00947 \text{ Caco2_Perm_10e6cms}$$

Coefficients

Term	Coef	SE Coef	Z-Value	P-Value	VIF
Constant	-4.372	0.344	-12.70	0.000	
LogP	0.3451	0.0599	5.76	0.000	5.95
AromaticRings	0.3506	0.0214	16.37	0.000	1.05
TertiaryAmine	0.8924	0.0622	14.35	0.000	1.01
QED	2.414	0.495	4.88	0.000	1.05
Solubility_LogS	-0.0578	0.0553	-1.04	0.296	7.00
Caco2_Perm_10e6cms	0.00947	0.00463	2.04	0.041	1.47

- Comments:
- Reduced model performance (0.69 vs 0.72)
 - Global monotone structure
 - AI already did all the predictor discovery heavy lifting

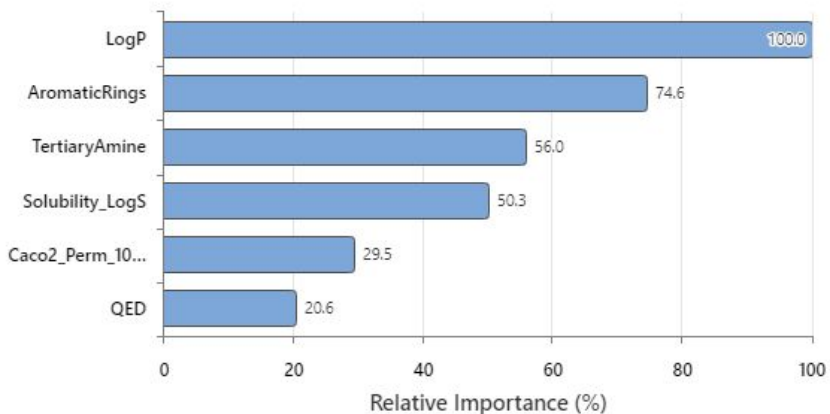
Assessing the Impact of Interactions

Model Summary

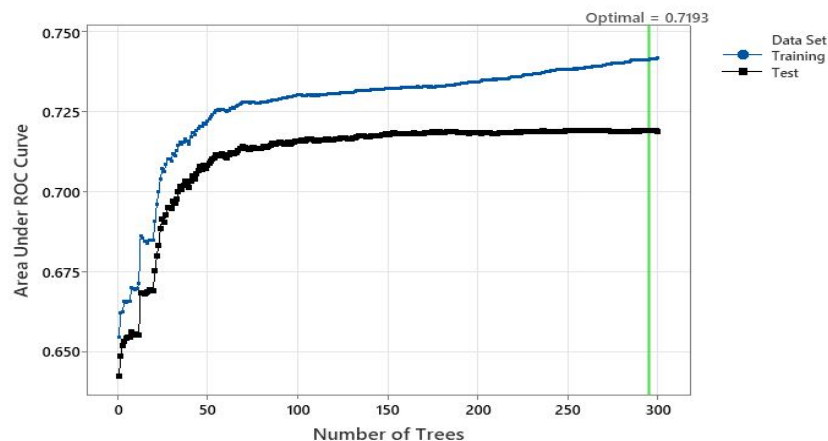
Total predictors 6
 Important predictors 6
 Number of trees grown 300
 Optimal number of trees 295

Statistics	Training	Test
Average -loglikelihood	0.5399	0.5463
Area under ROC curve	0.7413	0.7193
95% CI	(0.7289, 0.7537)	(0.6996, 0.7390)
Lift	2.3243	2.1105
Misclassification rate	0.2668	0.2735

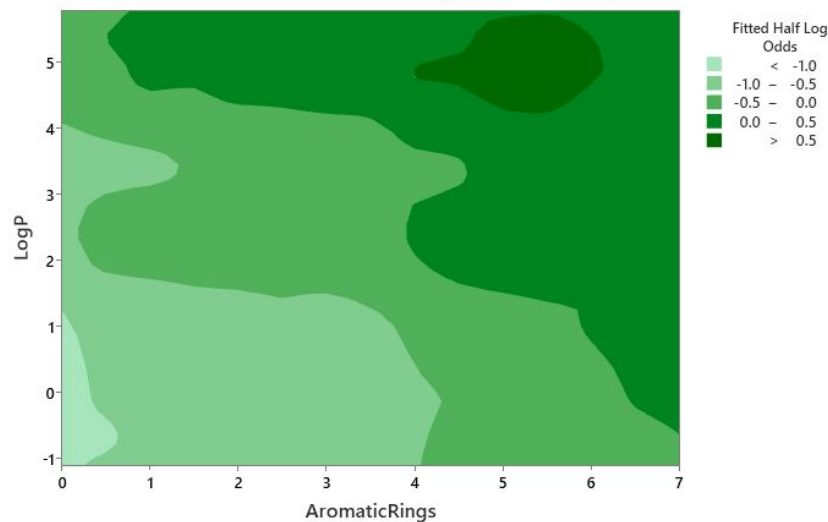
Relative Variable Importance



Area Under ROC Curve vs Number of Trees Plot



Contour Plot of Fitted Half Log Odds for hERG_Inhibitor



Generative AI



What About Generative AI?

- **Reactive Machines:**

- These are basic rule-based systems that operate based on predefined rules.

- **Expert Systems:**

- These are computer systems that mimic the decision-making ability of a human expert in a specific domain.

- **Machine Learning (ML) Systems:**

- ML is a subset of AI that focuses on developing algorithms and models that enable computers to learn from data.
- Types of ML systems include **supervised learning (PA)**, unsupervised learning, and reinforcement learning.

- **Neural Networks:**

- Inspired by the human brain, neural networks are a key component of many AI systems.

- **Narrow/Generative AI (Weak AI):**

- These AI systems are designed and trained for a specific task or a narrow set of tasks.
- Examples include virtual personal assistants, image recognition software, and language translation services.

Text In
Text Out

- **Limited Memory:**

- These AI systems can learn from historical data to make better decisions.
- Self-driving cars often use limited memory AI to navigate based on past experiences.
- Can be integrated into robots to enable them to learn and interact with the environment.

- **Self-aware AI:**

- This refers to hypothetical AI systems with self-awareness and consciousness.

- **Theory of Mind:**

- This is a more advanced form of AI that can understand human emotions, beliefs, intentions, and thoughts.

- **General AI (Strong AI):**

- General AI systems can understand, learn, and apply knowledge across diverse domains.
- They can perform any intellectual task that a human being can do.

- **Superintelligent AI:**

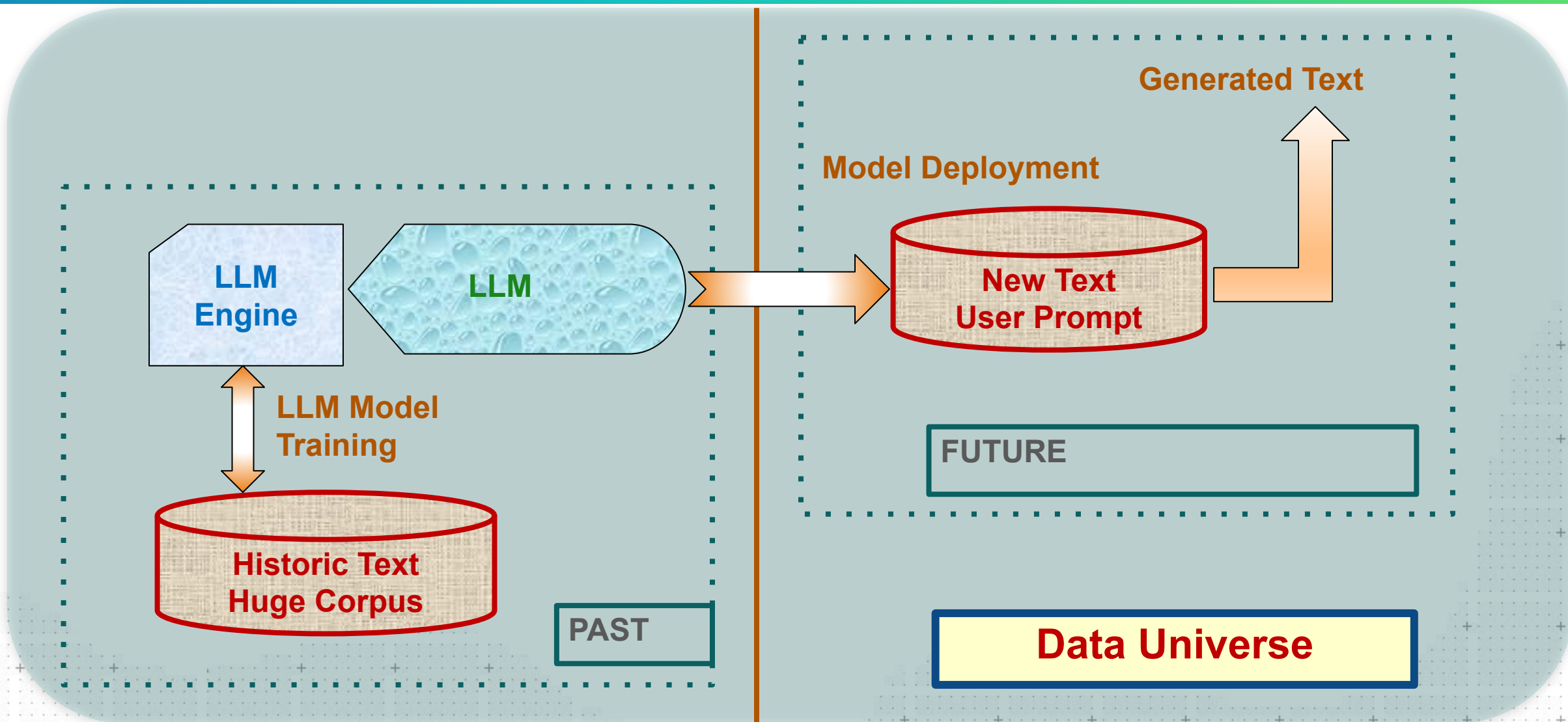
- This is a theoretical AI that surpasses human intelligence in every aspect.

A More General Case of PA

Text In
Text Out

- ▶ To engage in **Generative AI**, you must have:
- ▶ **Clear Problem Definition** – predict **Output Text** based on **Input Text**
- ▶ **Historical Dataset** – a fixed corpus of documents
- ▶ Due to the special nature of the historical dataset (unstructured text), the target predictors requirements have been generalized:
 - **Response/Target variable** – **Output Text**
 - **Predictor variables** – **Input Text**

Large Language Model Process (Gen AI)



Generations of Information Retrieval (IR)

- ▶ Libraries
 - As old as ancient Greece and Egypt
 - Culminated in the Library of Congress Classification
- ▶ Pre-digital Mechanization (19th-20th centuries: index cards, punch cards, etc.)
- ▶ Early Computerized IR (1940s-1960s: Boolean searches, etc.)
- ▶ Search Databases (1970s-1980s: electronic abstracts, indices, etc.)
- ▶ Web-search era (1990s-2000s: Web Crawler, Alta Vista, Yahoo, Google, etc.)
- ▶ Big Data and ML (2010s: collaborative filtering, recommender systems, personal assistants, etc.)
- ▶ Generative AI and Conversational Retrieval (2020s: ChatGPT and its likes)



I don't know,
Google it.



How Should We Handle Generative AI?

Divination



Interrogation



Examples of ChatGPT

1. Fabricated Legal Precedents & Decisions

- In May 2023, a lawyer submitted a brief to a court citing *precedents* that were entirely generated by ChatGPT, leading to a court ruling on the conduct, noting the fraudulent nature of the

3. Fabricated Biographical, Historical, and Scientific Information

- ChatGPT has produced wildly inaccurate information, including:
 - Claiming Leonardo da Vinci painted the Mona Lisa
 - Stating that George Washington invented the potato
 - Creating fake academic-sounding text that even fooled experts until the truth was discovered

TRUST BUT VERIFY



YOU MUST



its

substitute for table salt, and ChatGPT advised to use table salt for toxicity. The AI gave this advice which was not accurate enough to be reported in a medical journal.

AI-generated Summaries

Researchers have found that AI tools—including ChatGPT—can generate summaries, and produce misleading summaries. One study encouraged users to rely on reliable sources, making fact-checking more difficult.

GenAI Fundamental Flaws

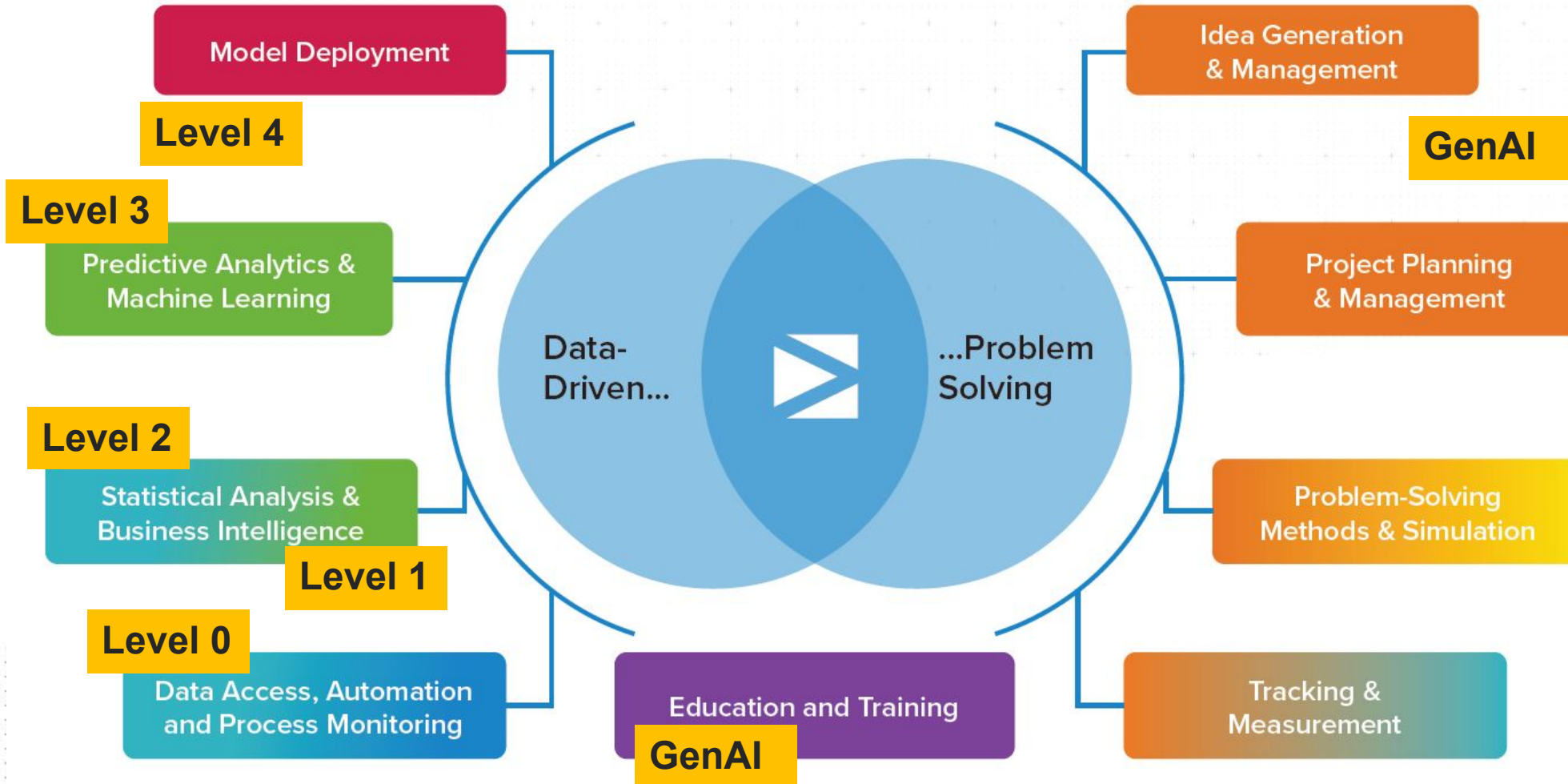
- ▶ The end of scalability with the release of GPT 5.0 (Moore's law yielding to the 90/10 rule)
- ▶ Creation vs Regurgitation
- ▶ Unacceptable Liability (Code Generation vs Medical Advice)
- ▶ Hacking and code/prompt injection attacks
- ▶ Emotional attachments
- ▶ Mega Hitler phenomenon and other ethical issues
 - Ghosting
 - Red-lining
- ▶ "Don't Hallucinate" oxymoron
- ▶ Flooding the Data Universe with the superfluous (garbage) content
 - Junk books on Amazon (book mills)
 - Junk science papers (paper mills)
 - Junk reviews and bots (content mills)



MINITAB

Data Driven Problem Solving

A Suite of Solutions



Why Minitab?

- ▶ Trusted “down to Earth” digital transformation tools and techniques
- ▶ Realistic innovation in data science and machine learning
- ▶ Focus on understanding your business needs and offering practical data-driven solutions
- ▶ Versatile project-management tools, both individual and corporate
- ▶ Variety of data access, data gathering, and data storage tools
- ▶ Real-time sensor-based process monitoring
- ▶ Intuitive descriptive statistics, graphs, control charts, and dashboards
- ▶ State of the art powerful machine learning algorithms
- ▶ Scalable model deployment and model monitoring facilities
- ▶ Discrete event simulation tools
- ▶ Python, R, and ChatGPT connectors
- ▶ Information and education aids to bring you up to speed





- ▶ In this presentation, Mikhail will break down the complexities of modern AI into digestible concepts, making the subject accessible to all. He will then illustrate the key points with an insightful case study from medical research. The following topics will be discussed:
 - The fundamental principles of AI and machine learning.
 - Analyst-driven vs data-driven paradigms in machine learning.
 - A case study to highlight the usefulness of predictive AI in medical research.

You have data. We have solutions. Imagine the possibilities.

At Minitab, we help customers around the world leverage the power of data analysis to gain insights and make a significant impact on their organizations. By unlocking the value of data, Minitab enables organizations to improve performance, develop life changing innovations and meet their commitments of delivering high quality products and services and outstanding customer satisfaction.



You have data.

[analytics]

[dashboards]

[machine learning]

We have **Solutions Analytics™**.

[training]

[visualizations]

[innovation]

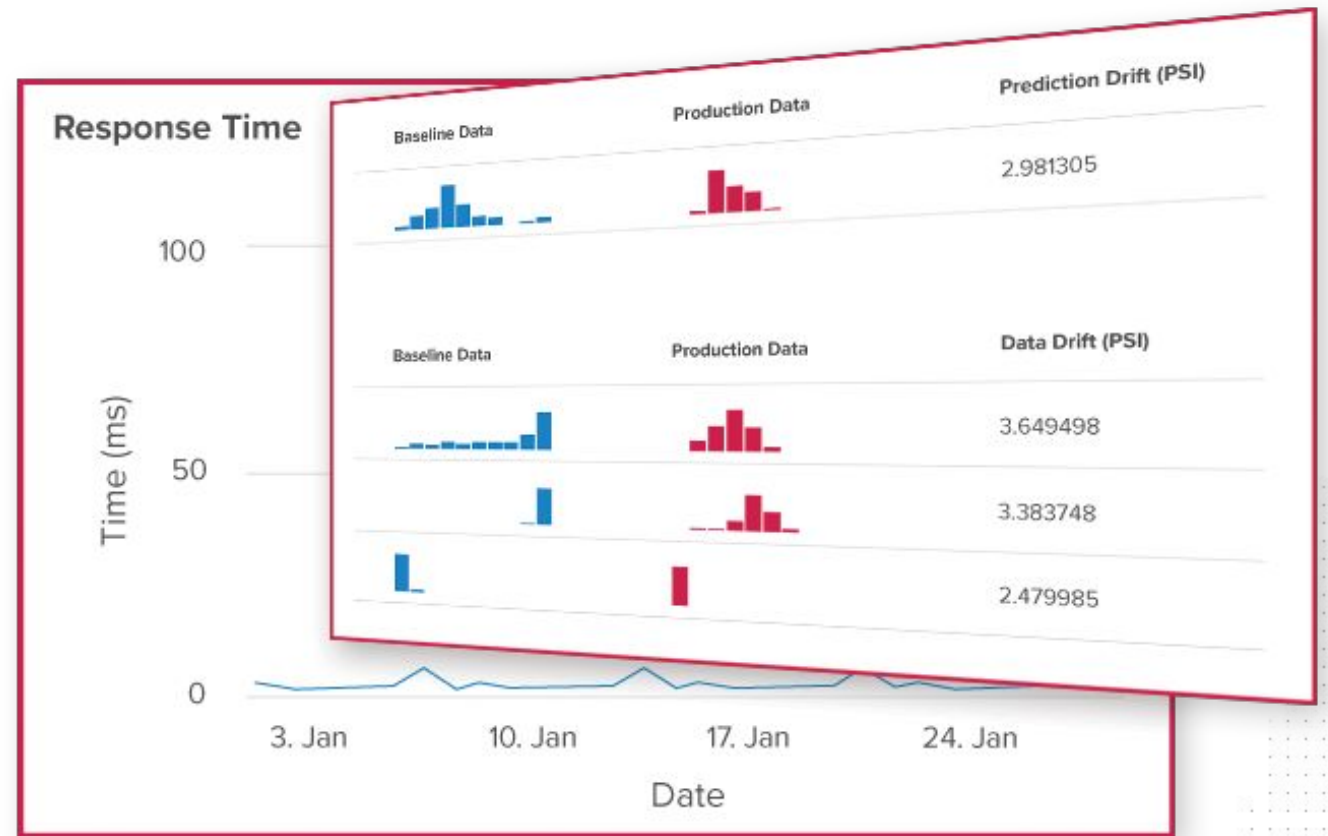
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut semper urna dictum, efficitur augue at, vestibulum nunc. Nunc volutpat finibus erat, nec aliquet elit congue nec.

- ▶ Sed in felis auctor enim ornare molestie
- ▶ Duis ipsum velit, posuere sed nunc
- ▶ Vivamus eu lorem vel turpis facilisis vehicular

Curabitur vel dui augue. Quisque vel gravida diam. Vivamus tristique enim ut lorem sodales luctus. Duis ultrices lectus eu felis ornare, non pharetra mauris scelerisque. Integer accumsan, nisl ut ultrices aliquet, lacus leo mollis dui, a ultricies tortor sapien sed sem. Etiam id interdum sem. Fusce et risus ut augue placerat faucibus.

Donec tincidunt vel diam sed suscipit. Donec eget neque rhoncus, pretium sem a, mollis est. Aenean justo lorem, convallis et magna nec, lacinia ultricies est.

- Fusce at nibh eget purus aliquam congue
- Quisque in accumsan massa
- Fusce at nibh eget purus aliquam congue



thank you

Gracias

ευχαριστώ

Danke

Grazie

благодаря

Hvala

Obrigado

Kiitos

شكراً

Tak

Ahsante

Teşekkürler

متشكراً

Salamat Po

감사합니다

Cám ơn

شكريه

Terima Kasih

Dank u Wel

Děkuji

நன்றி

Köszönöm

ありがとう
ございます

ขอขอบคุณครับ

Dziękuję

谢谢

Tack

Mulțumesc

спасибо

Merci

תודה

多謝晒

дядкую

Ďakujem