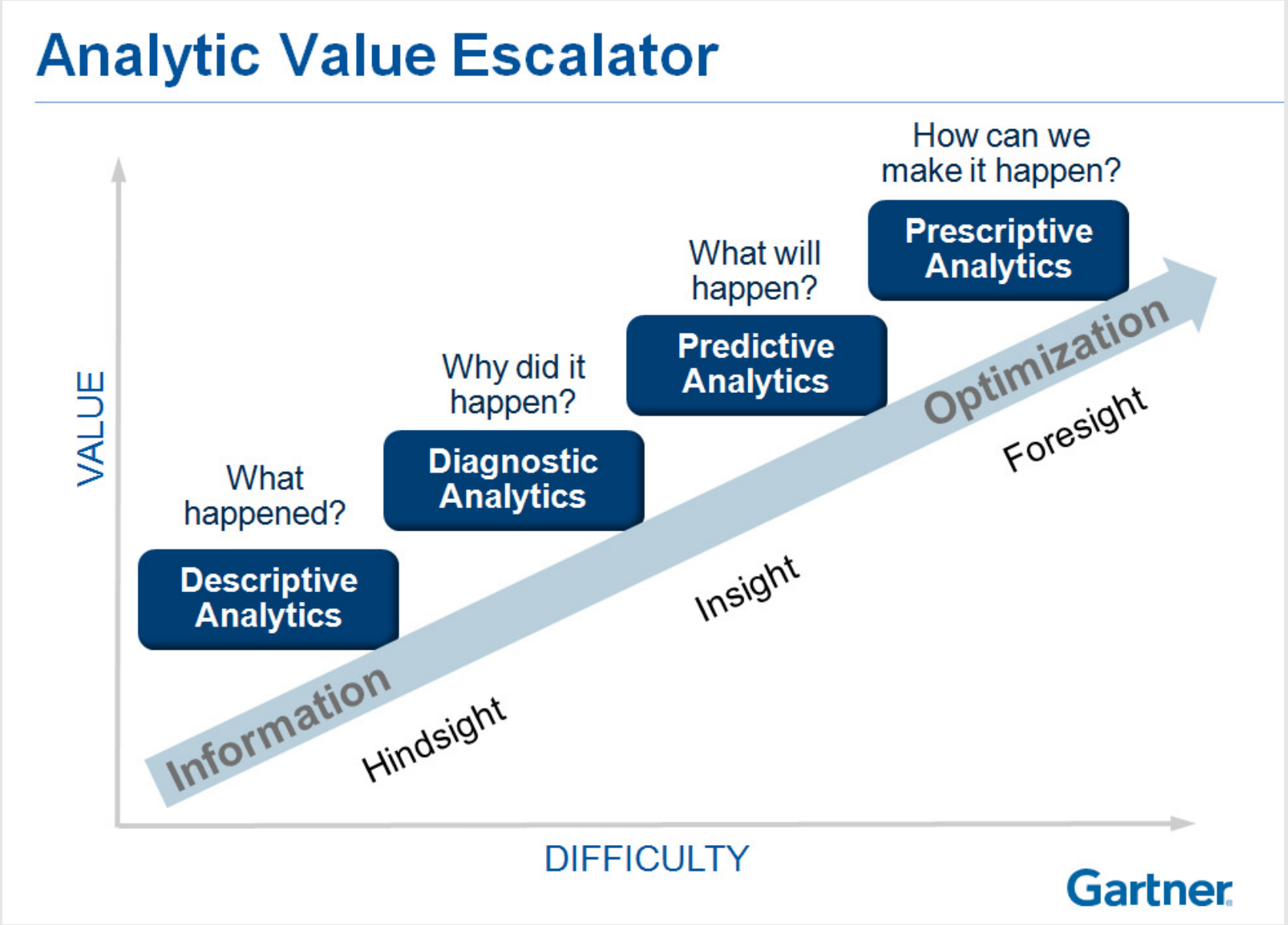
Session 1: Introduction to Machine Learning in Finance

Analytics

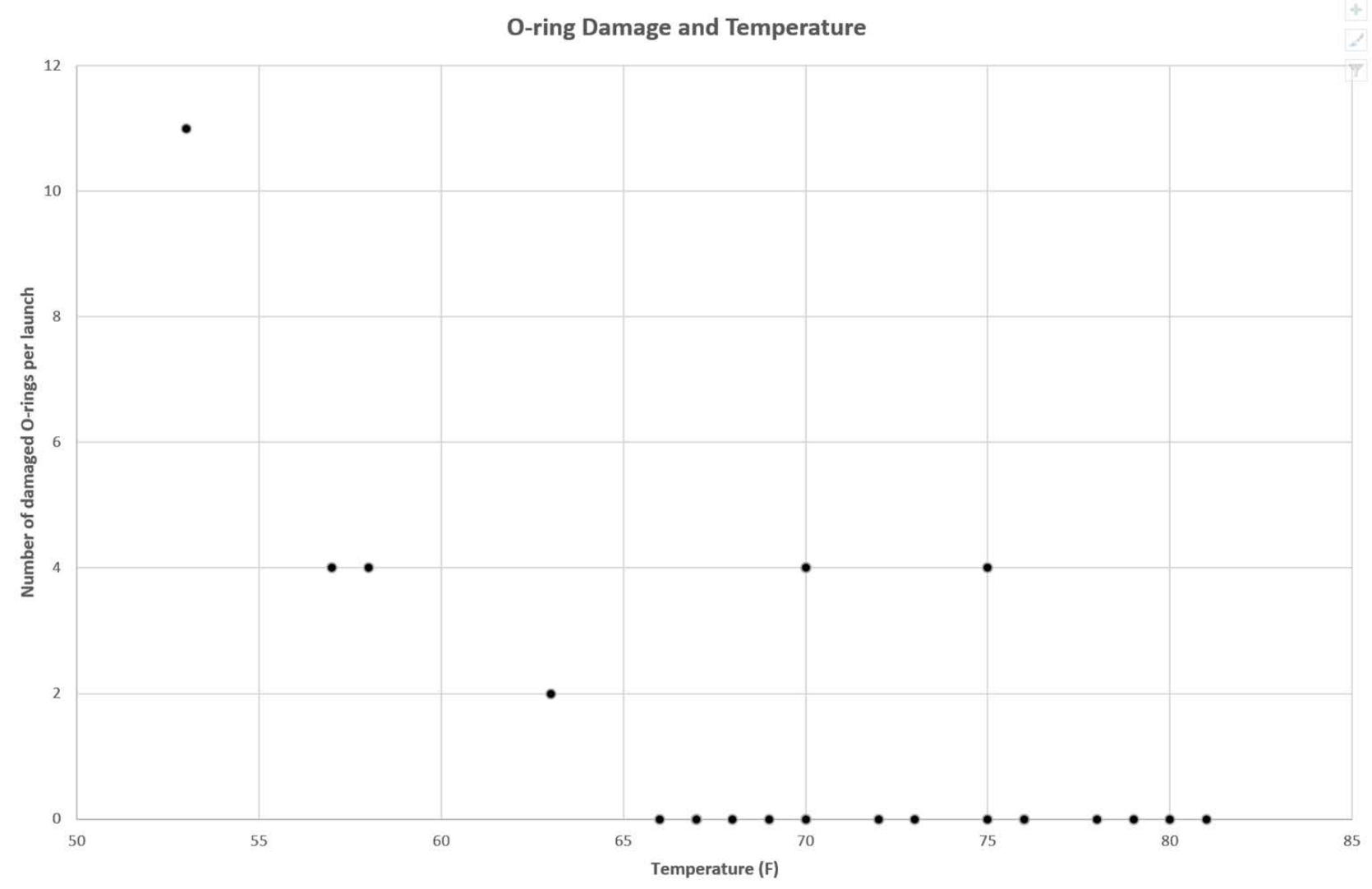
Types of Analytics



Types of Analytics

- Descriptive Analytics
  - Make use of data aggregation and data mining to provide insight into the past and answer: “What has happened?”
- Predictive Analytics
  - Make use of statistical models and forecasts techniques to understand the future and answer: “What could happen?”
- Prescriptive Analytics
  - Make use of optimisation and simulation algorithms to advice on possible outcomes and answer: “What should we do?”

Diagnostic Analysis



Data Analysis

What is Data Analysis?

- Sometimes called **Data Analytics**
- A process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making ~ Wikipedia
- Data mining is a particular data analysis technique that focuses on modelling and knowledge discovery for predictive rather than purely descriptive purposes
- Business intelligence covers data analysis that relies heavily on aggregation, focusing on business information.

What is Data Analysis?

- In statistical applications data analysis can be divided into
  - Descriptive statistics,
  - Exploratory data analysis (EDA),
  - Confirmatory data analysis (CDA).
- EDA focuses on discovering new features in the data
- CDA on confirming or falsifying existing hypotheses.
- Predictive analytics focuses on application of statistical models for predictive forecasting or classification
- Text analytics applies statistical, linguistic, and structural techniques to extract and classify information from textual sources, a species of unstructured data.

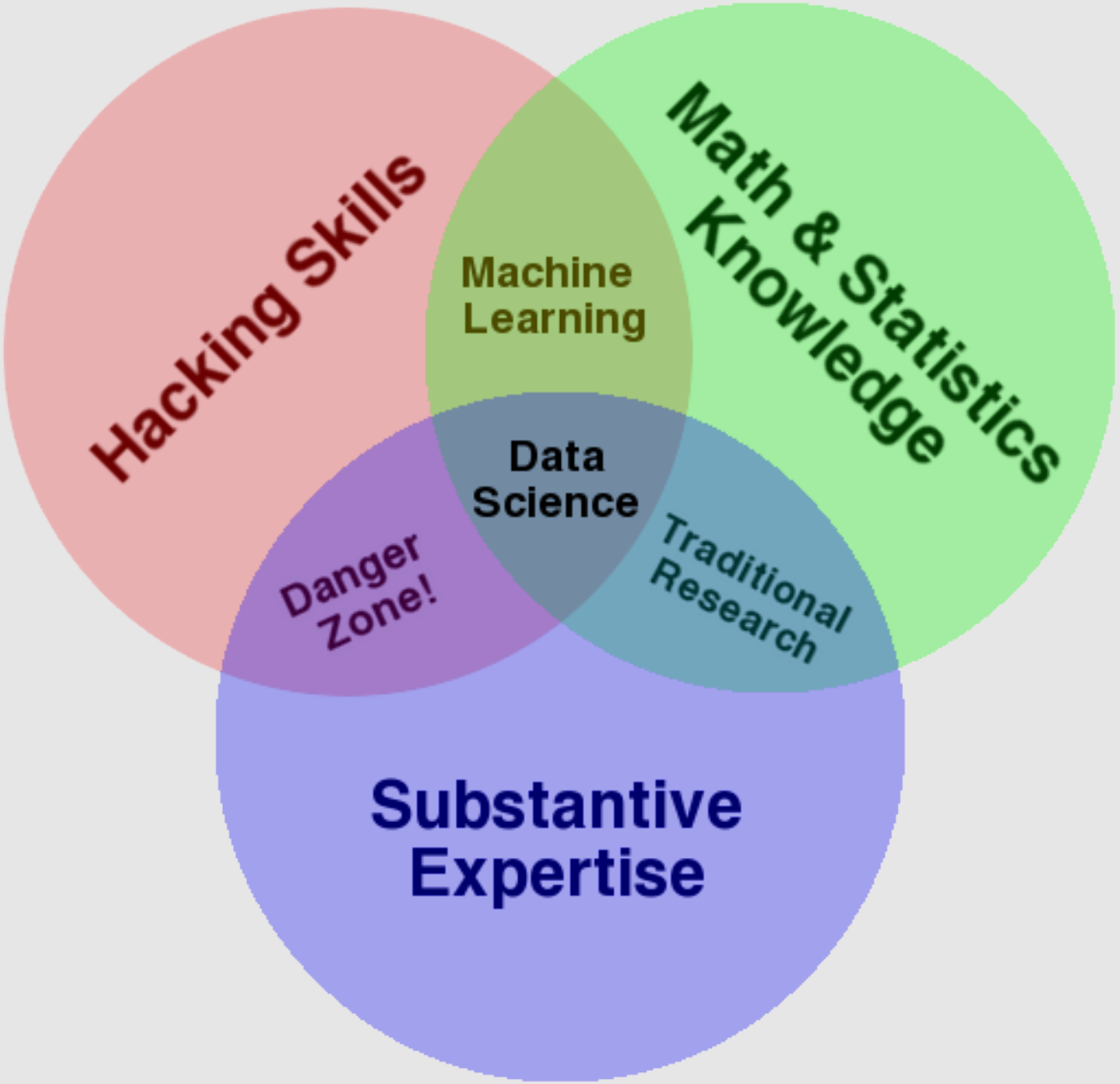
What is Data Analysis?

- A/B Testing
- Association Rule Learning
- Classification
- Cluster Analysis
- Data mining
- Ensemble learning
- Genetic algorithms
- Machine learning
- Natural language processing (NLP)
- Neural networks
- Network analysis
- Optimisation
- Pattern recognition
- Predictive modelling
- Regression
- Sentiment analysis
- Signal processing
- Spatial analysis
- Supervised learning
- Simulation
- Time series analysis
- Time series forecasting
- Unsupervised learning
- Visualisation

Data Science

What is Data Science?

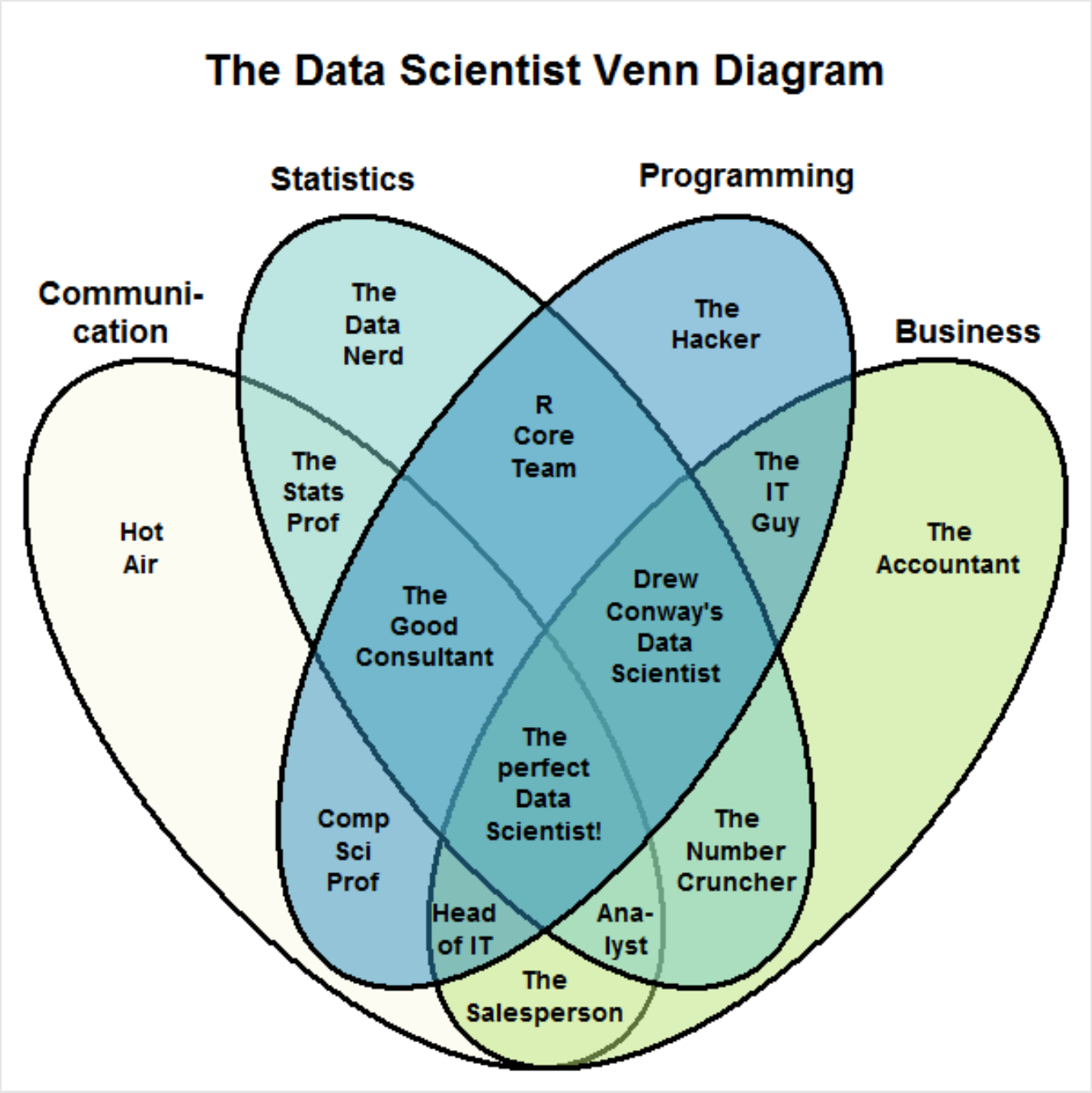
- Includes data analysis
- Goes much beyond
- Data Scientist end goal? Insight discovery
- Data Scientist
  - Interests is aligned with company goals
  - Proposed the hypothesis or questions
- Data Analyst goes about answering those questions



Source: [Drew Conway]

Some Terminologies

- Analytics
  - Analytics is the discovery, interpretation, and communication of meaningful patterns in data. Especially valuable in areas rich with recorded information, analytics relies on the simultaneous application of statistics, computer programming and operations research to quantify performance.
  - Multidisciplinary.
  - Extensive use of data analysis.
  - Insights used to recommend action or to guide business decision making.
  - Individual analyses forms part of it
  - Concerned with the entire methodology.
- Analysis
  - Extensive use of mathematics and statistics
  - Make use of descriptive techniques and predictive models



Source: [Stephan Kolassa]

What is Data Science?

- Data science, also known as data-driven science, is an interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining.
- Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyse actual phenomena" with data.

Source: Wikipedia

Machine Learning

[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed.

Arthur Samuel, 1959

Why Machine Learning?

- Set rules versus ML
  - E.g., Spam filtering
- ML is great for:
  - Problems the requires a lot of hand-tuning or long list of rules
  - Complex problems
  - Evolving environment: Adapts to new data
  - Pattern discovery / Provides insights

Good Resource: McKinsey's An executive's guide to AI

Types of Machine Learning Algorithms – When to use it?

- Supervised
  - You know how to classify the input data and the type of behaviour you want to predict, but you need the algorithm to calculate it for you on new data
- Unsupervised
  - You do not know how to classify the data, and you want the algorithm to find patterns and classify the data for you
- Reinforcement Learning
  - You don't have a lot of training data; you cannot clearly define the ideal end state; or the only way to learn about the environment is to interact with it

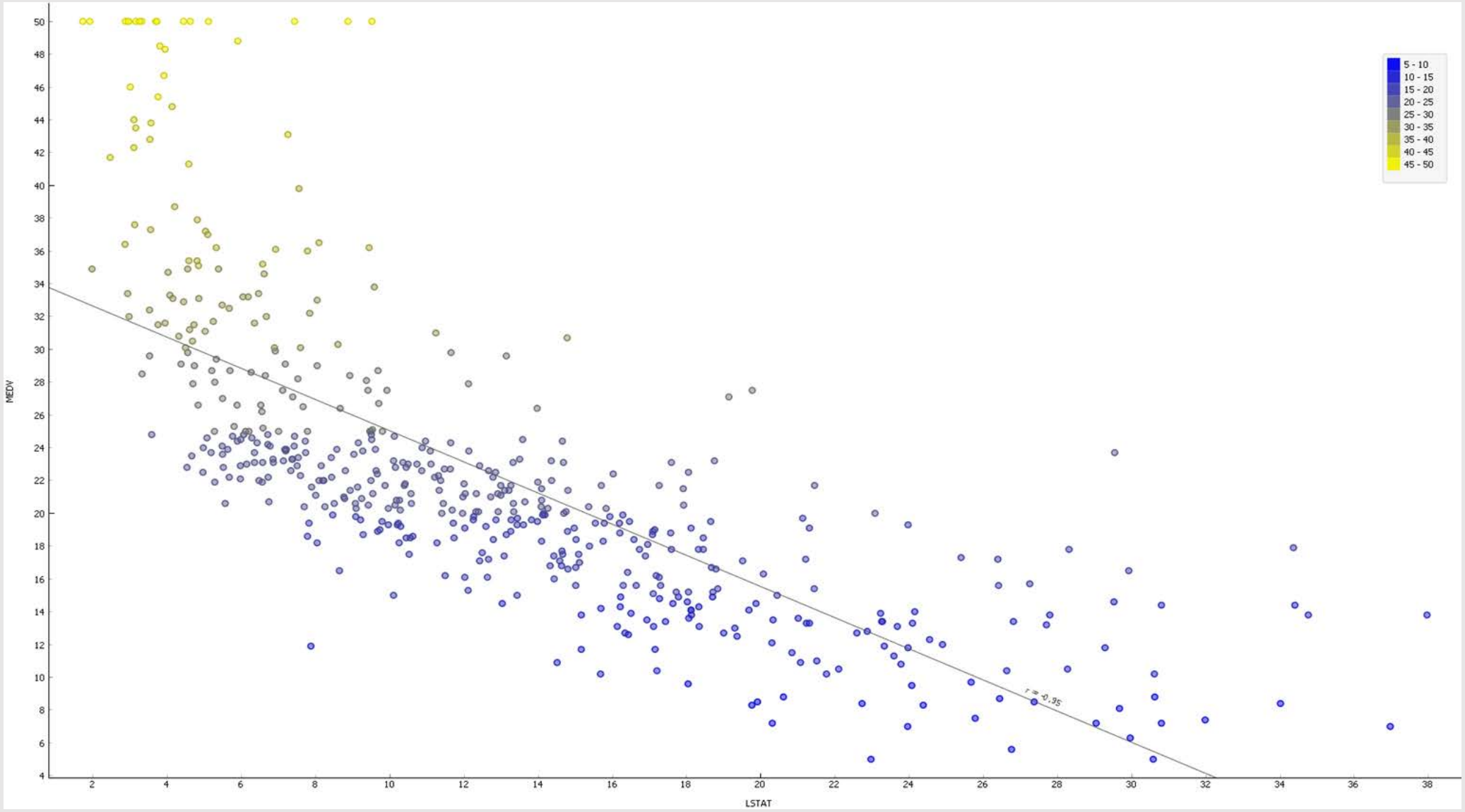
Source: McKinsey's An executive's guide to AI

Types of Machine Learning Algorithms

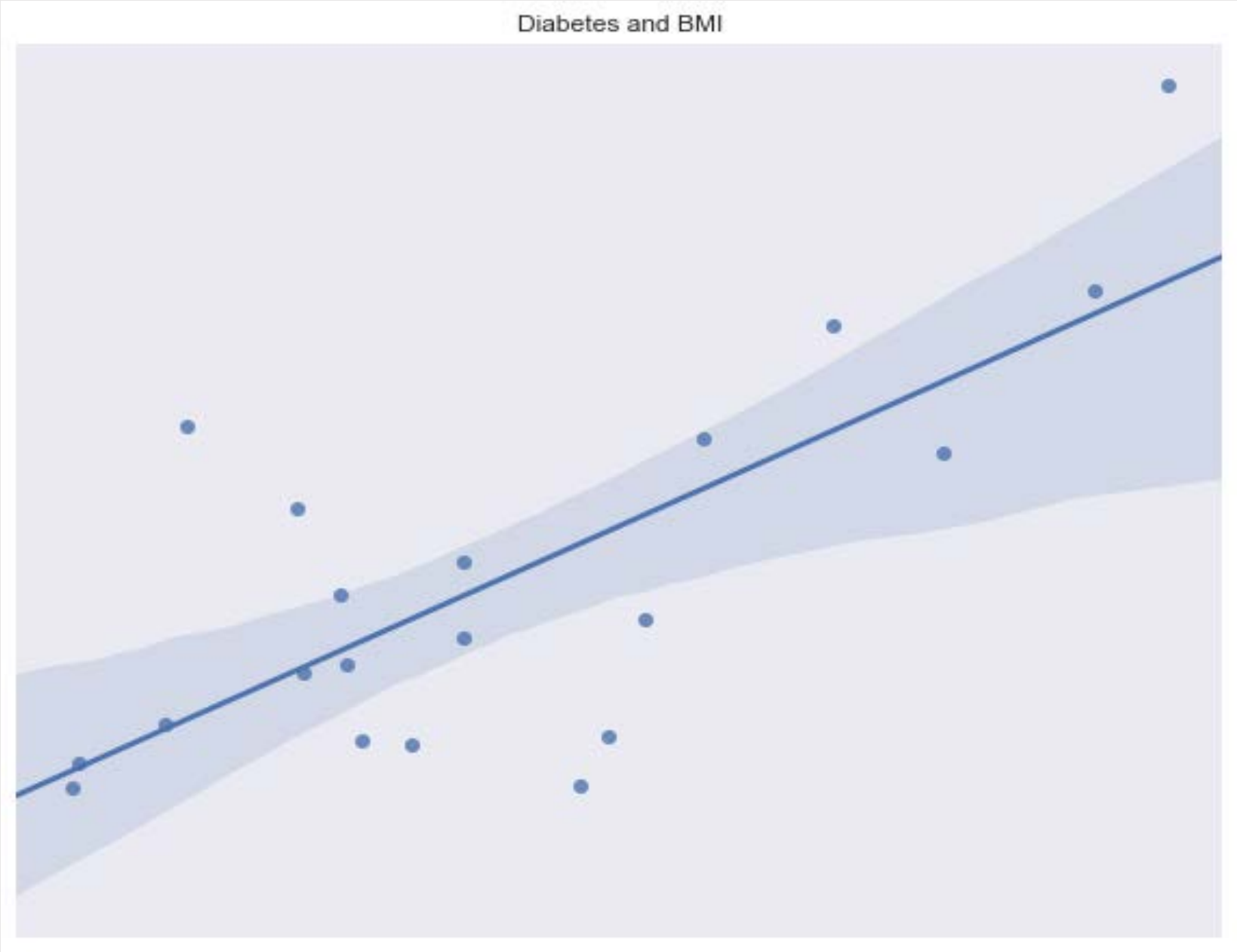
- Supervised
  - Regression
    - Linear Regression, Penalised Methods, Tree based
  - Classification
    - Logistic Regression, Tree based, Bagging, Boosting
- Unsupervised
  - Clustering
  - Visualisation / Dimensionality Reduction
  - Association rule learning
- Semi-supervised
- Reinforcement Learning

Good Resource: McKinsey's An executive's guide to AI

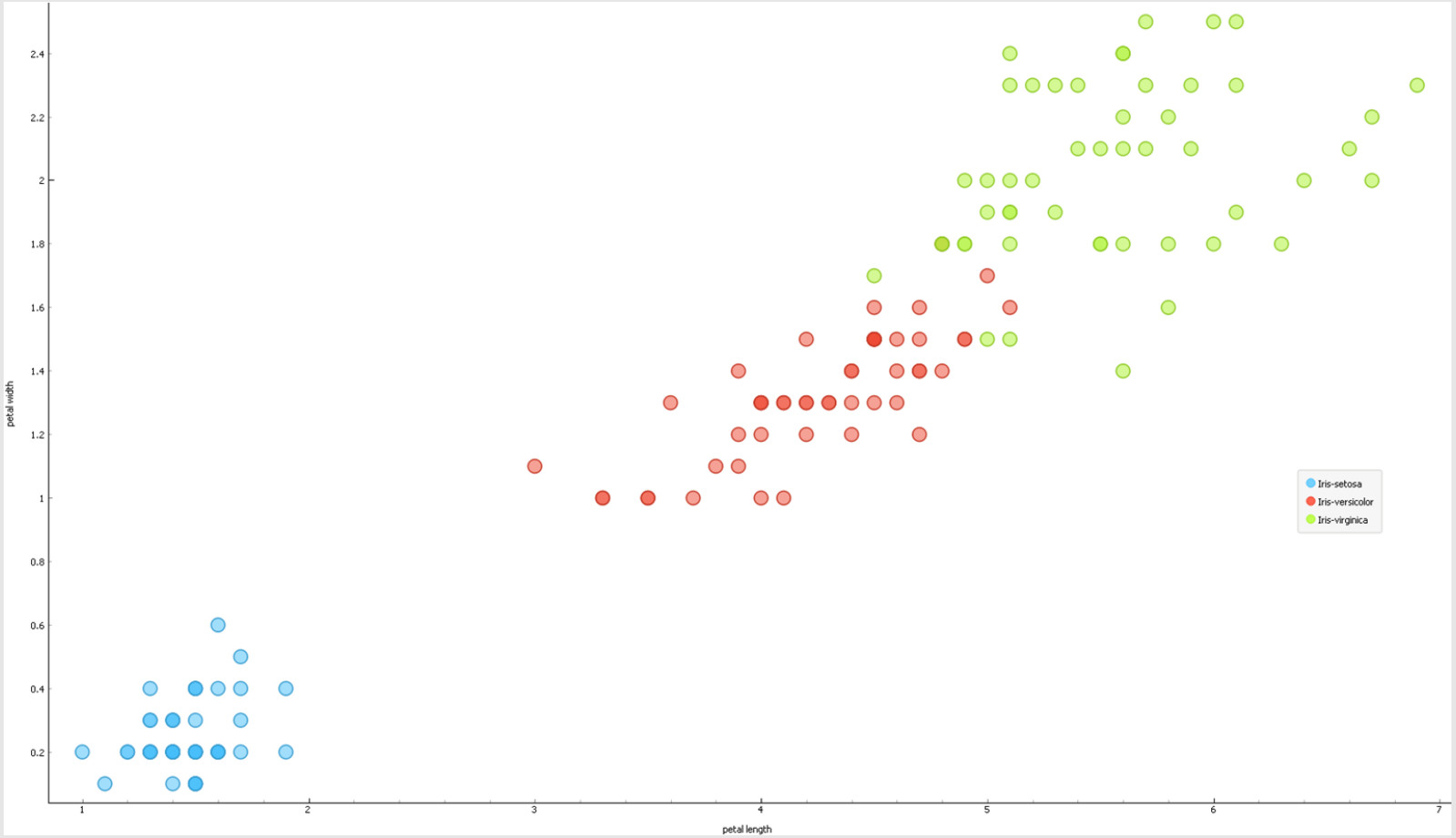
Regression



Regression



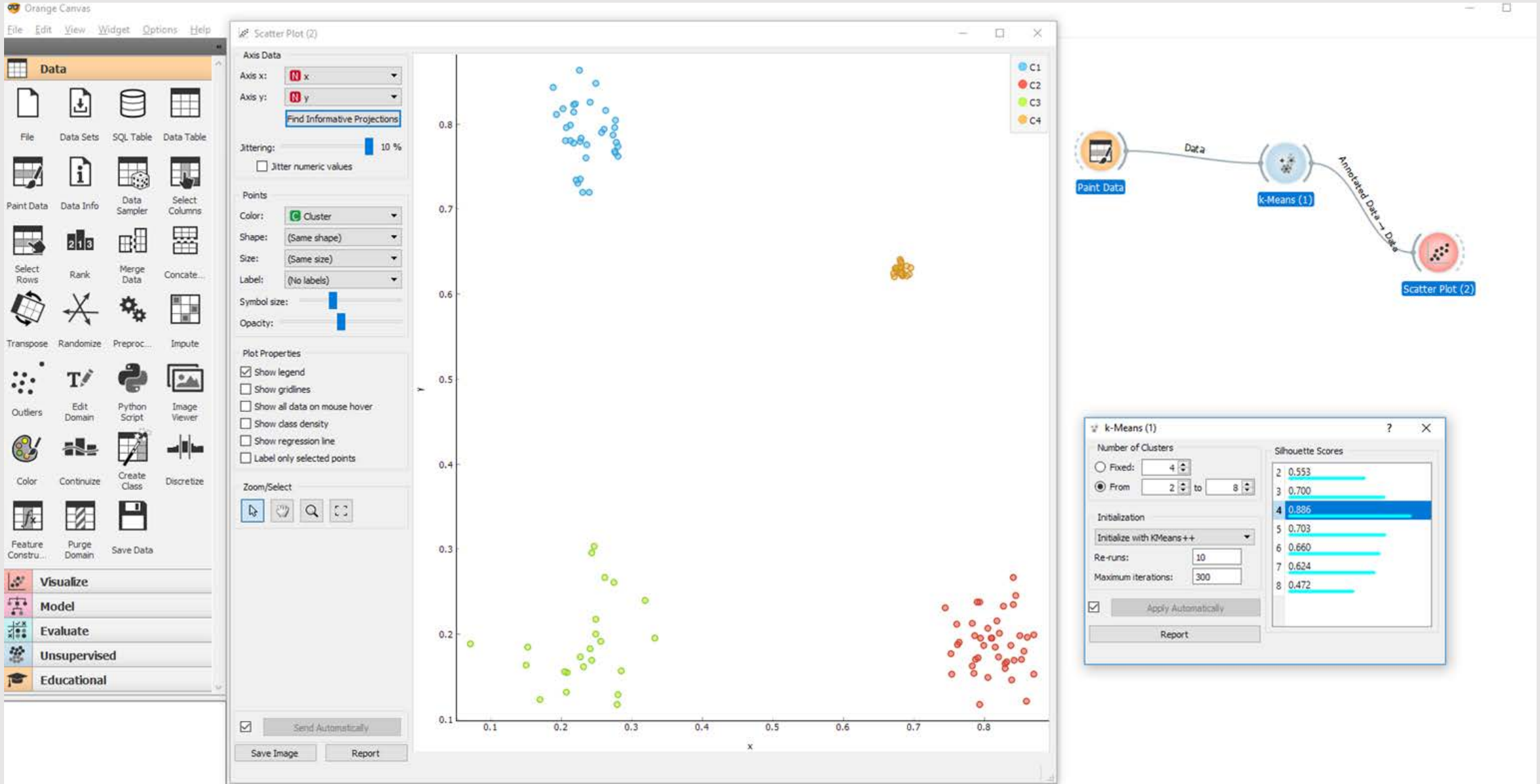
Classification



Deep Learning - Convolutional NN

Source: https://samirashabanian.wordpress.com/2015/02/15/convolutional-neural-network/

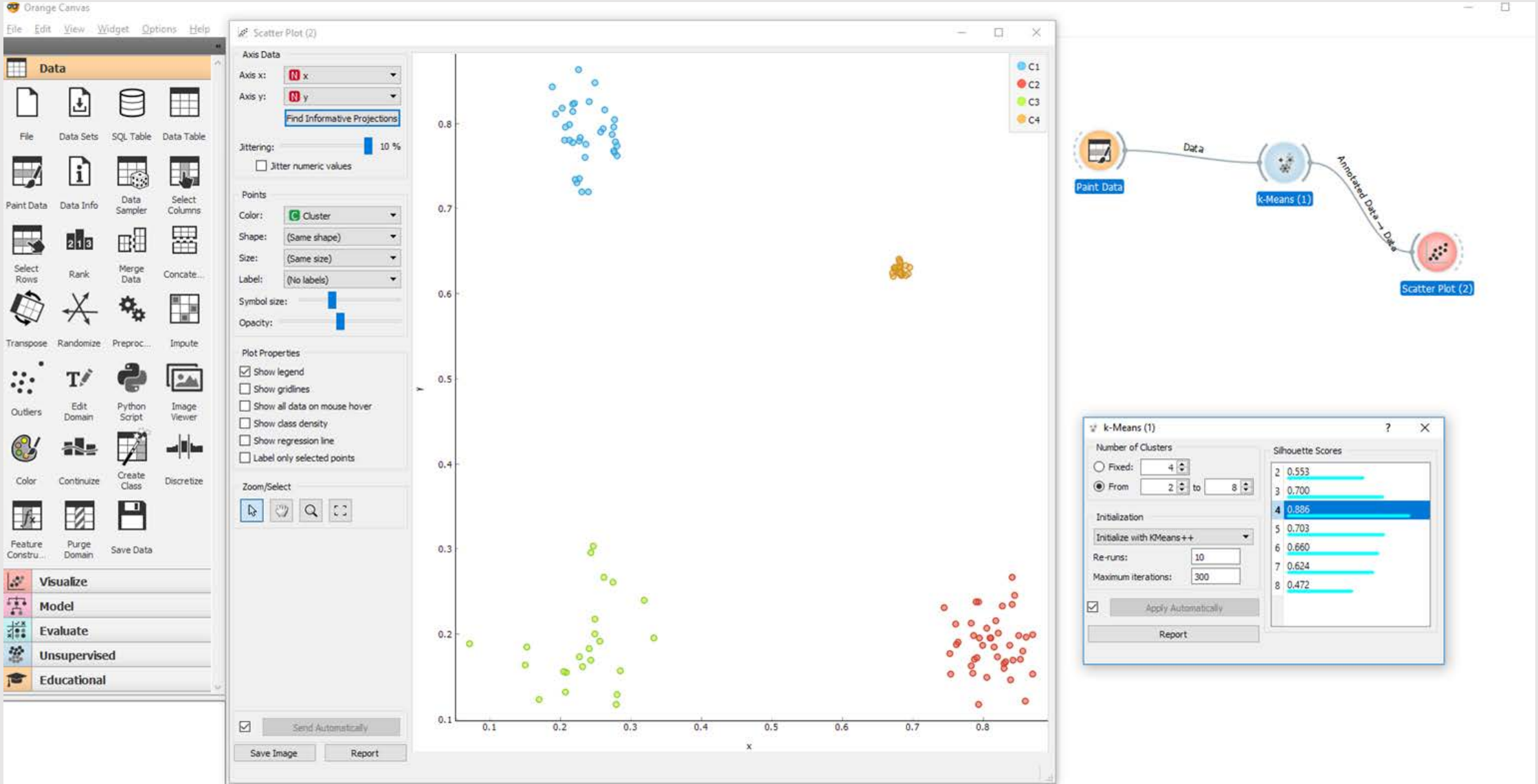
Unsupervised



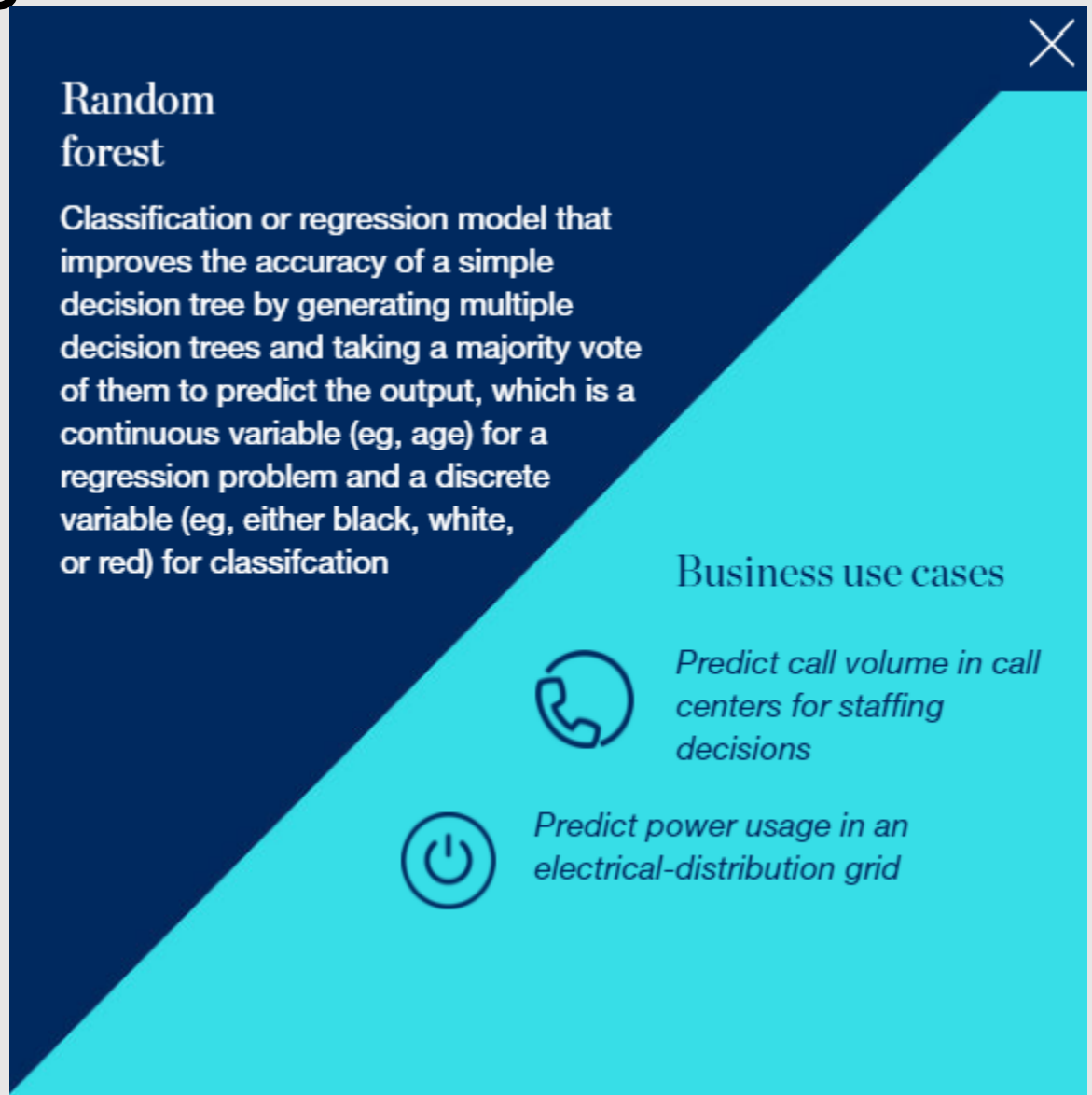
Reinforcement Learning - 2015

Reinforcement Learning - 2017

Unsupervised



Supervised Learning – Use Case



Source: McKinsey's An executive's guide to AI

Supervised Learning – Use Case

Source: McKinsey's An executive's guide to AI

Supervised Learning – Use Case

Source: McKinsey's [An executive's guide to AI]

Supervised Learning – Use Case

Source: McKinsey's An executive's guide to AI

Supervised Learning – Use Case

Source: McKinsey's An executive's guide to AI

Supervised Learning – Use Case

Source: McKinsey's An executive's guide to AI

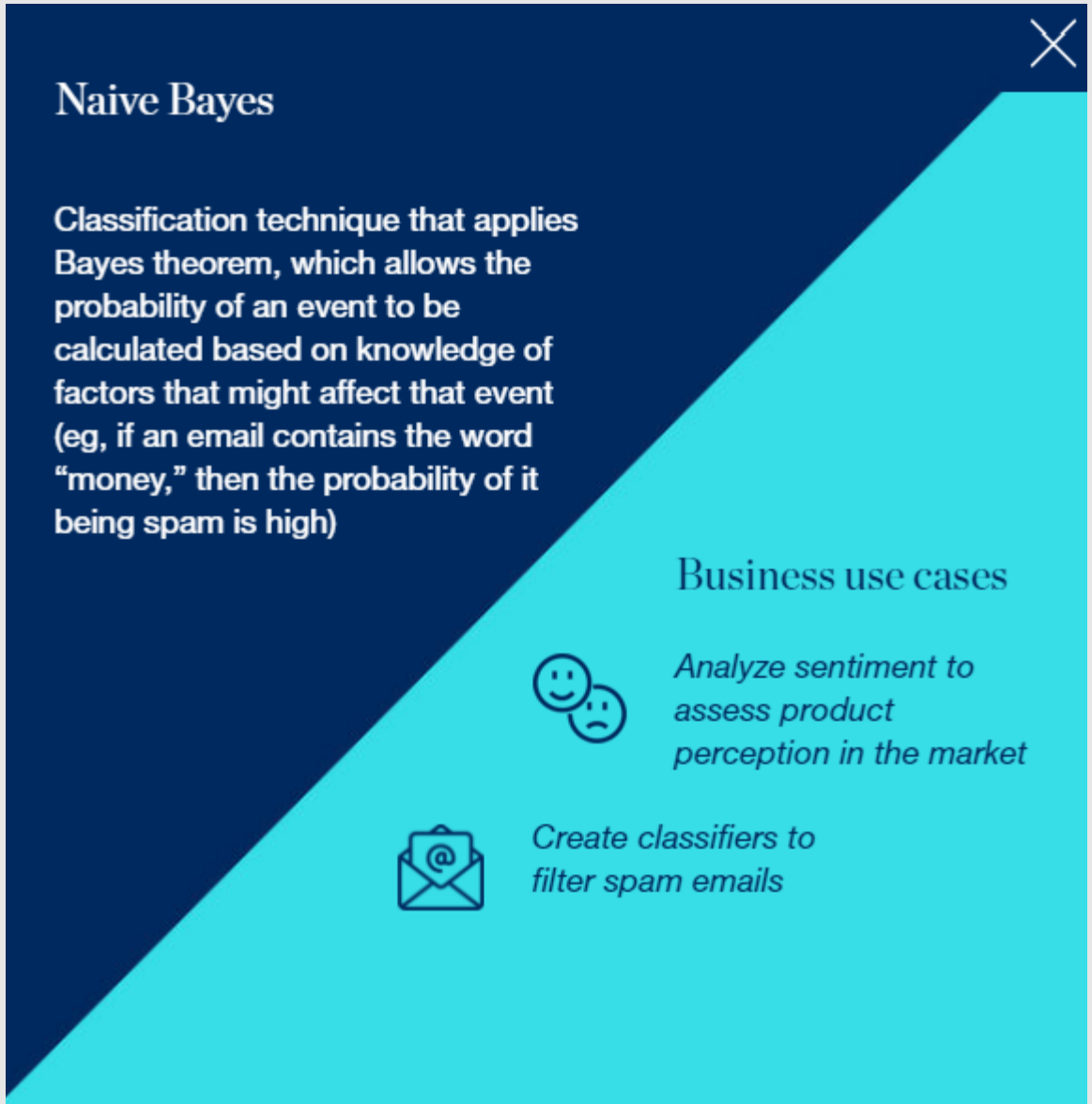
Supervised Learning – Use Case

Source: McKinsey's An executive's guide to AI

Supervised Learning – Use Case

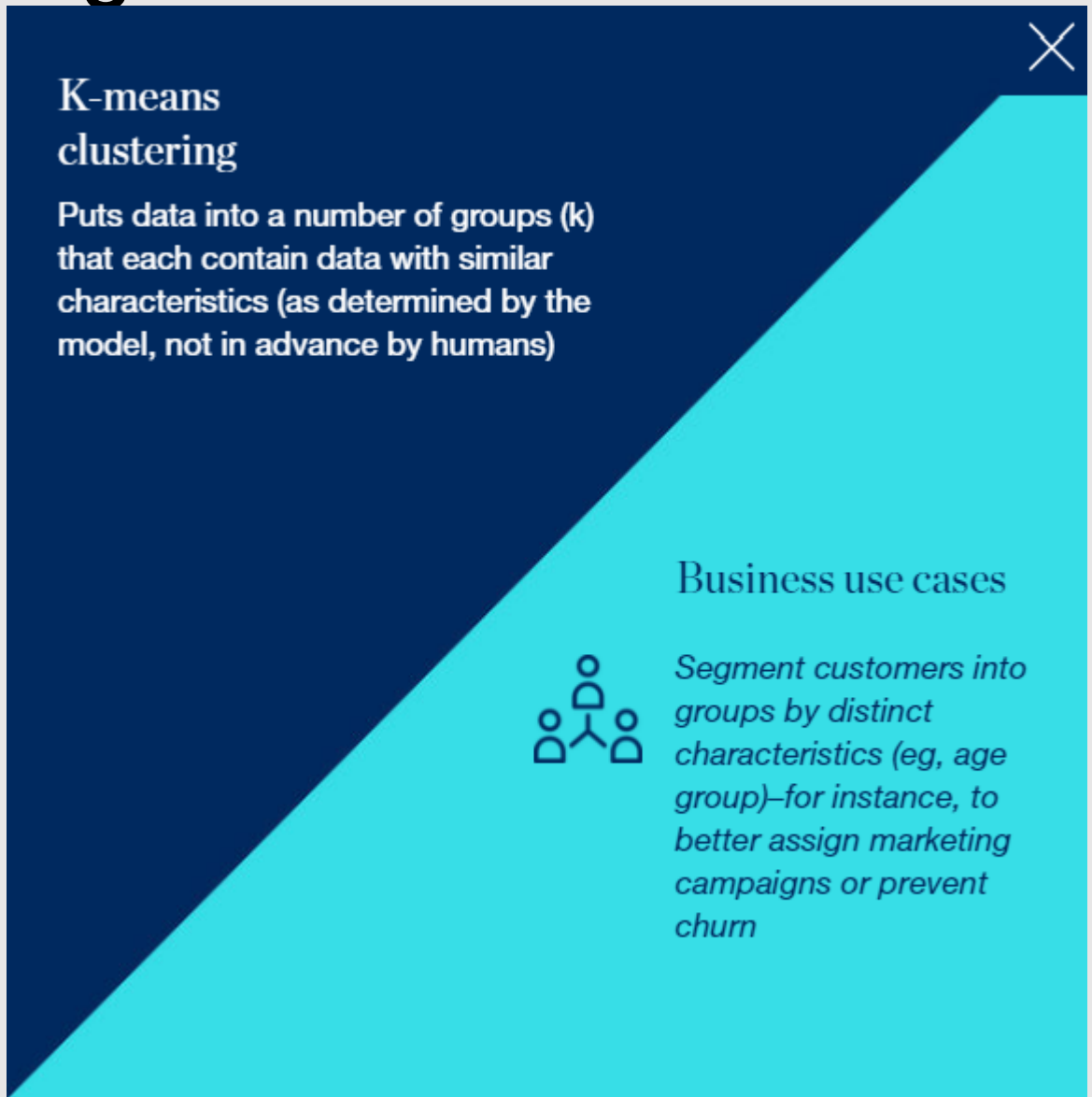
Source: McKinsey's An executive's guide to AI

Supervised Learning – Use Case



Source: McKinsey's An executive's guide to AI

Unsupervised Learning – Use Case

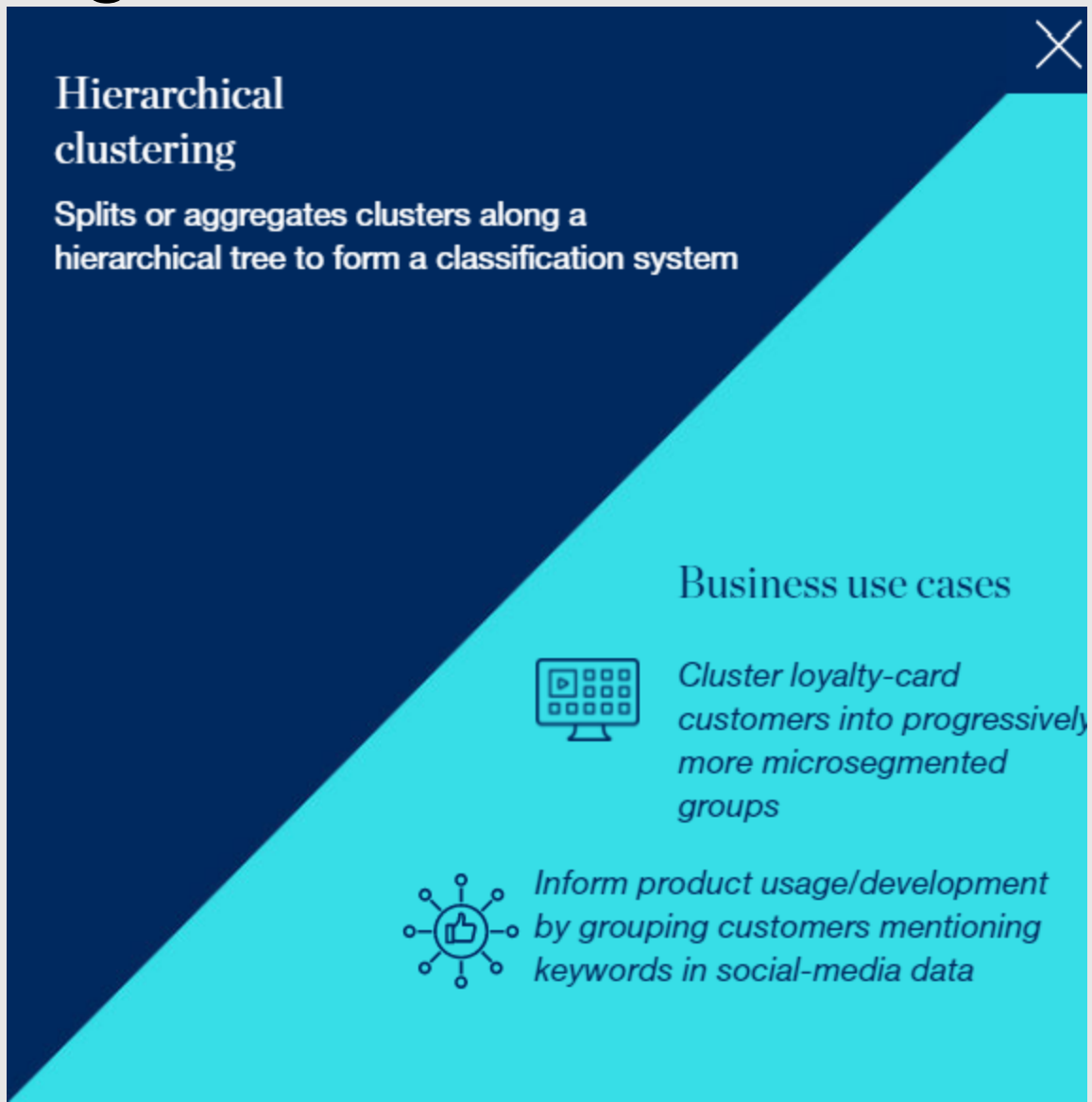


Source: McKinsey's An executive's guide to AI

Unsupervised Learning – Use Case

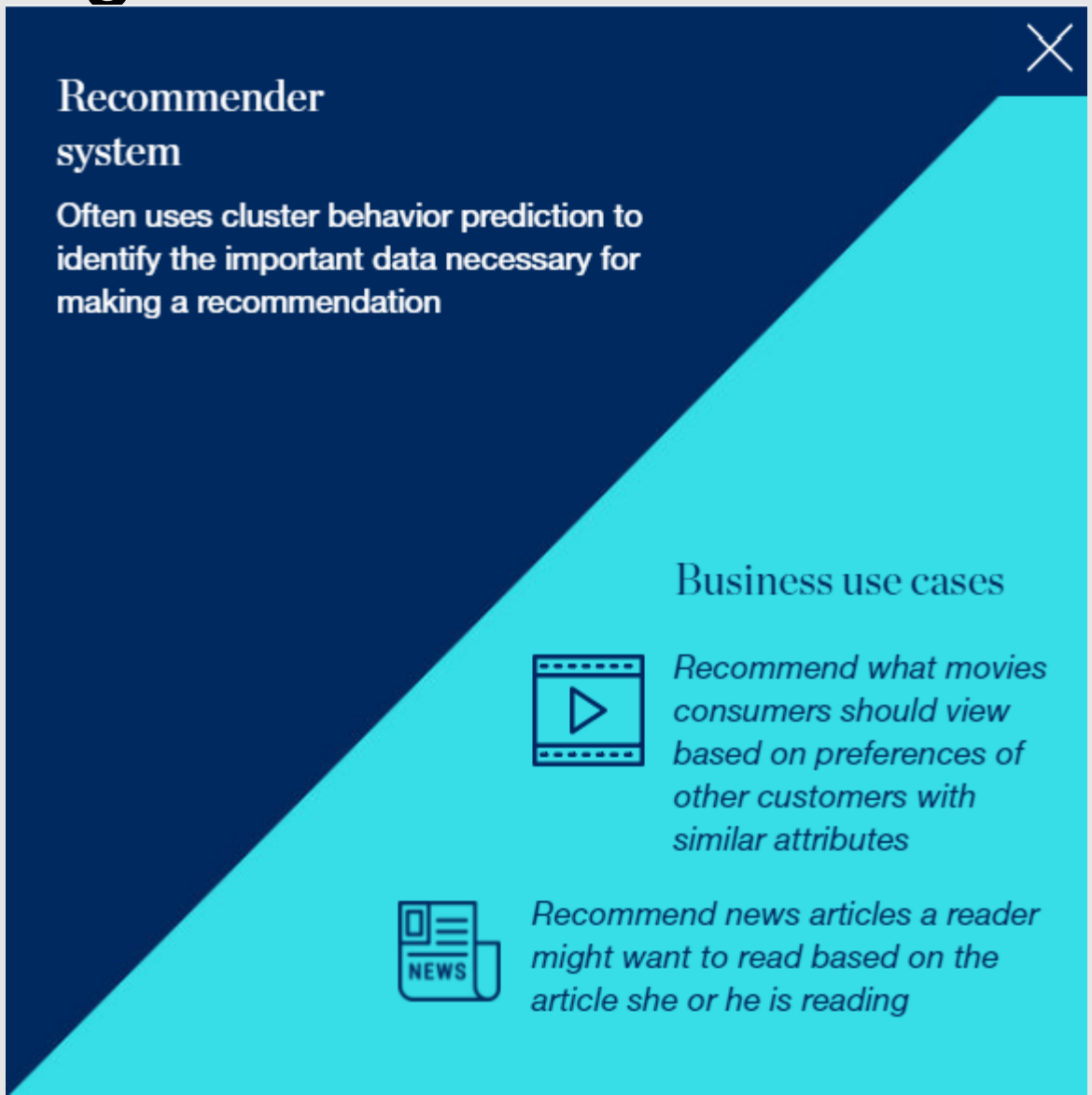
Source: McKinsey's An executive's guide to AI

Unsupervised Learning – Use Case



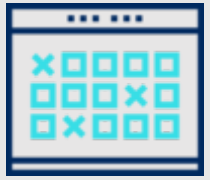
Source: McKinsey's An executive's guide to AI

Unsupervised Learning – Use Case



Source: McKinsey's An executive's guide to AI

Reinforcement Learning – Use Case



Optimise the trading strategy for an options-trading portfolio

Balance the load of electricity grids in varying demand cycles

Stock and pick inventory using robots

Optimise the driving behaviour of self-driving cars

Optimise pricing in real time for an online auction of a product with limited supply

Source: McKinsey's An executive's guide to AI

Deep Learning General Use Case

Sound	**Industry**
Voice recognition	UX/UI, Automotive, Security, IoT
Voice search	Handset maker, Telecoms
Sentiment analysis	CRM
Flaw detection (engine noise)	Automotive, Aviation
Fraud detection (latent audio artefacts)	Finance, Credit Cards
Time Series	**Industry**
Log analysis/Risk detection	Data centres, Security, Finance
Enterprise resource planning	Manufacturing, Auto., Supply chain
Predictive analysis using sensor data	IoT, Smart home, Hardware manufacturing
Business and Economic analytics	Finance, Accounting, Government
Recommendation engine	E-commerce, Media, Social Networks

Source: https://deeplearning4j.org/use_cases

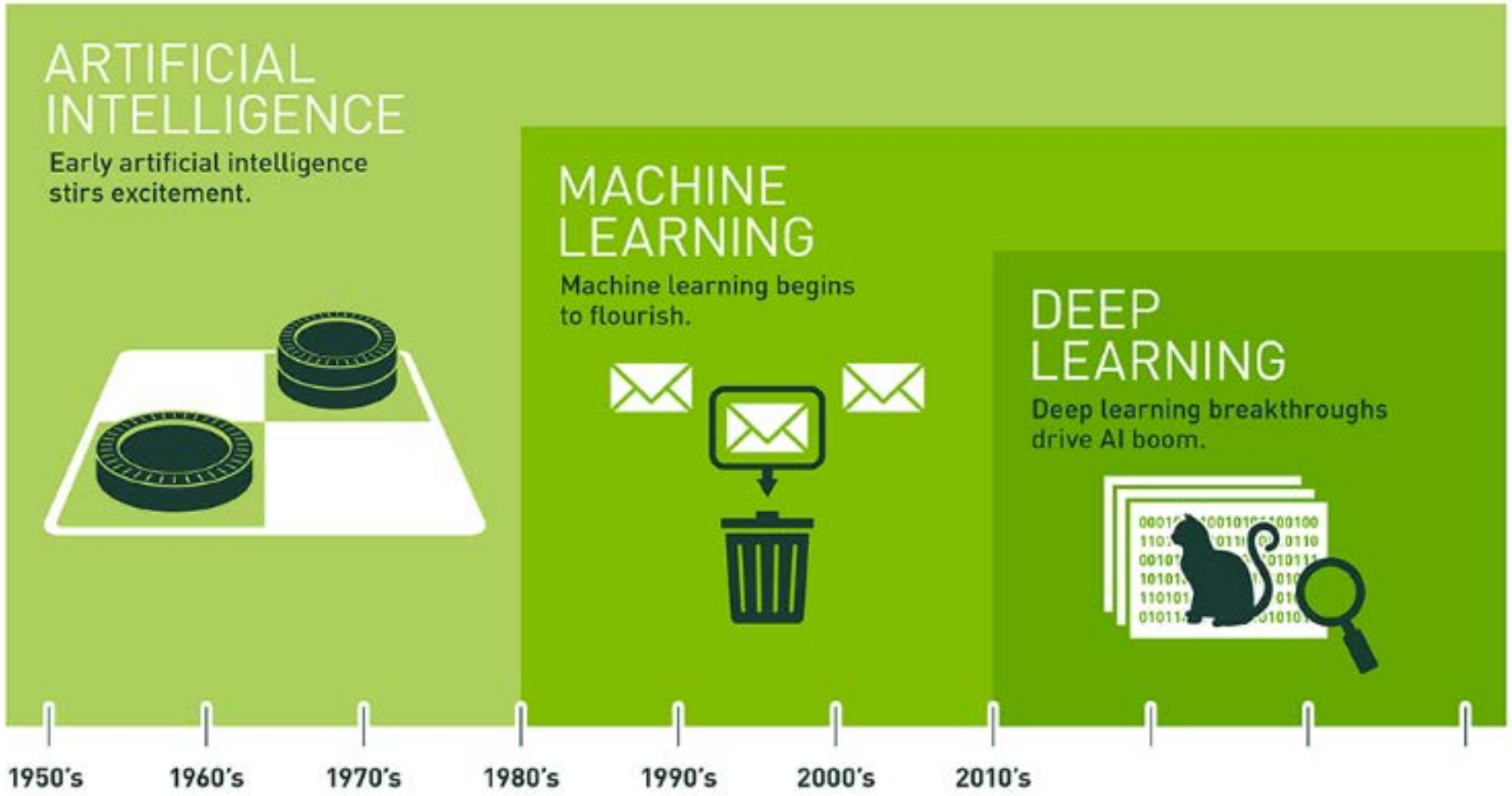
Deep Learning General Use Case

Text	**Industry**
Sentiment Analysis	CRM, Social media, Reputation mgt.
Augmented search, Theme detection	Finance
Threat detection	Social media, Govt.
Fraud detection	Insurance, Finance
Image	**Industry**
Facial recognition, Image search	Social media
Machine vision	Automotive, aviation
Photo clustering	Telecom, Handset makers
Video	**Industry**
Motion detection	Gaming, UX, UI
Real-time threat detection	Security, Airports

Source: https://deeplearning4j.org/use_cases

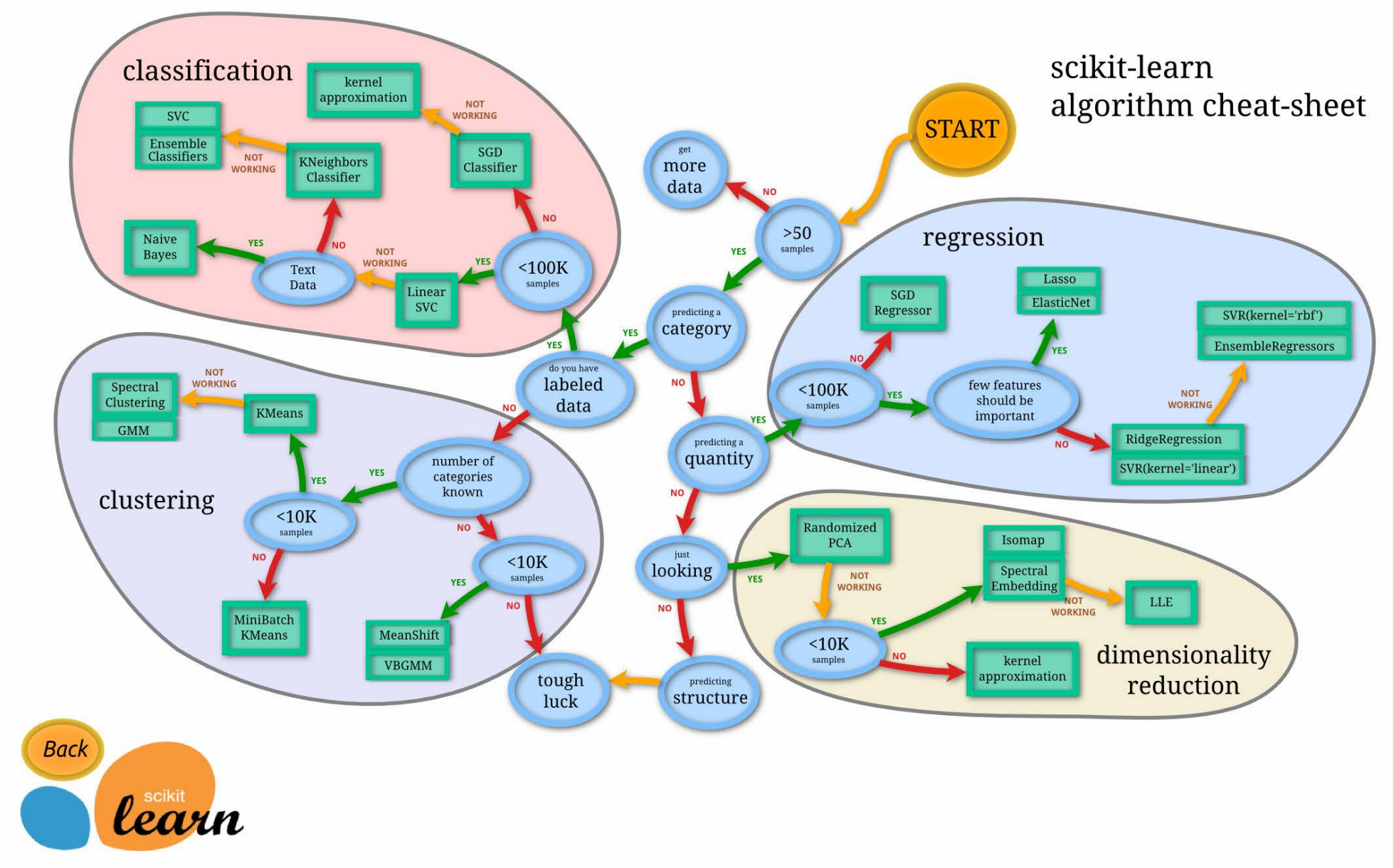
Machine Learning Algorithms

Anthony Ng @ 2018



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Source: Nvidia https://bit.ly/2aoQwIx



Source: http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

Machine Learning in Finance

Supervised Machine Learning

Credit Scoring

Credit Scoring

- Statistical forecast of individual credit risk
- To objectively evaluate creditworthiness of borrowers
  - Ability
  - Willingness
- Used by lenders in
  - Loan approval
  - Evaluating risk based charge (E.g., interest rate benchmark + margin)

Source: WPS Analytics

Credit Scoring - Prehistoric

- Performed by loan manager
- 5Cs
  - Character
  - Capital – deposit and loan quantum
  - Collateral – Security being offered
  - Capacity – Repaying ability
  - Condition – Macroeconomic consideration

Source: Credit Scoring Using Machine Learning, https://arrow.dit.ie/sciendoc/137/

Credit Scoring - Prehistoric

- Shortcomings:
  - Subject to human emotion
  - Decisions lacked consistency / difficult to replicate
  - No set process & difficult to formalise
  - Not scalable

Source: Credit Scoring Using Machine Learning, https://arrow.dit.ie/sciendoc/137/

Credit Scoring

- Regulations drove the need for a more objective and transparent process
- Equal Credit Opportunity Act (ECOA) 1974
  - Prohibits a creditor from discriminating against an applicant on the basis of race, colour, religion, national origin, and sex.
  - Statistically credit scoring systems were proposed
- The US Federal Reserve's Regulation B1 (Section 202.2) required that credit scoring systems must be, amongst other things:
  - based on data that are derived from an empirical comparison of sample groups or the population of creditworthy and noncreditworthy applicants who applied for credit within a reasonable preceding period of time
  - Developed and validated using accepted statistical principles and methodology

Source: Credit Scoring Using Machine Learning, https://arrow.dit.ie/sciendoc/137/

Credit Scoring

- Empirically derived and other credit scoring systems
- A credit scoring system is a system that evaluates an applicant's creditworthiness
  - Mechanically
  - Based on key attributes of the applicant
  - Aspects of the transaction
  - Information about the applicant
- To qualify as an empirically derived, demonstrably and statistically sound, credit scoring system, the system must be:
  - Based on data that are derived from an empirical comparison of sample groups or the population of creditworthy and non-creditworthy applicants who applied for credit within a reasonable preceding period of time
  - Developed for the purpose of evaluating the creditworthiness of applicants
  - Developed and validated using accepted statistical principles and methodology
  - Periodically revalidated to maintain predictive ability.

Source: Credit Scoring Using Machine Learning, https://arrow.dit.ie/sciendoc/137/

Credit Scoring

Table 3.1: Application scorecard with a credit score for applicant X

Feature	*Attribute*	*Points*	*Attribute value for applicant X*	*Points for applicant X*
Age	< 25	69		
	25 - 29	77		
	30 - 34	84	34	84
	35 - 41	93		
	42 - 50	104		
	50+	110		
Bank Customer	Yes	29	Yes	29
	No	20		
Credit limit on credit card	Blank	60		
	< 2,000	55		
	2,000 - 3,750	59	3,500	59
	3,751 - 6,000	64		
	6,001 - 10,000	71		
	> 10,000	74		
Years at current job	< 1	20		
	1 - 3	24		
	4 - 6	29	4	29
	7+	36		
Accommodation Status	Own	42		
	Rent	28	Rent	28
	Parents	32		
	Other	34		
Self-employed	Yes	25		
	No	41	No	41
Gross Monthly Income	< 2,500	71		
	2,500 - 3,150	79		
	3,151 - 3,850	85	3,750	85
	3,851 - 4,350	92		
	4,351 - 5,100	103		
	> 5100	111		
Score				**355**

Source: *Credit Scoring Using Machine Learning.* Kenneth Kennedy.
https://arrow.dit.ie/sciendoc/137/

Scorecard Model Development

- Data Quality
  - Accuracy – degree of precision of measurements of a feature to its true value
  - Completeness – the extent to which values are missing in the data
  - Consistency – multiple data sources are used and due to a lack of standardisation, two or more data items may conflict with each other
- Data discretisation / classing
- Feature selection
- Train / Test dataset
- Model development and validation

Source: Building a Scorecard in Practice, https://bit.ly/2iKyw4n

Common Models

- Logistic regression
  - Most commonly used algorithm within the consumer credit scoring industry
- Supervised ML
  - Linear Discriminant Analysis (LDA)
  - Support Vector Machine
  - Neural Network
  - Decision Tree
  - Random forests

Alternative Credit Scoring in the US

SharedLending

cignifi

ecredable
Make Every Payment Count

firstaccess

FACTOR TRUST

happy mango

Neener Analytics

zest finance

Juvo

trooly

trustingsocial

Alternative

Alternative Data

- Utilities (gas, water, electricity)
- Telecom (TV, mobile, broadband)
- Rent
- Property/asset record: including value of owned assets
- Public records: beyond the limited public records information already found in standard credit reports
- Alternative lending payments (e.g., payday, instalment loan, rent-to-own, buy-here-pay-here auto loans, auto title loans): including both on-time and derogatory payment data
- Demand deposit account (DDA) information: including recurring payroll deposits and payments, average balance, etc.

Source: Alternative Data and the Unbanked https://owy.mn/2Gxwwn7

Alternative Data

- Transaction Data
- Telecom / Rent / Utility Data
- Social Profile Data
- Social Network Data
- Clickstream Data
- Audio and Text Data
- Survey Data
- Mobile App Data

Source: http://www.fico.com/en/blogs/analytics-optimization/using-alternative-data-in-credit-risk-modeling/

Transaction Data

- How customers use their credit or debit cards.
- Not often mined to extract the maximum predictive value.
- Can be used to generate a wide range of predictive characteristics
  - Ratios of Cash to Total Spend in last X week(s)
  - Ratios of Spend in last X week(s) to last Y week(s)
  - Characteristics based on
    - Number
    - Frequency
    - Value of transactions at different retailer types

Source: http://www.fico.com/en/blogs/analytics-optimization/using-alternative-data-in-credit-risk-modeling/

Social Profile Data

- Facebook, LinkedIn, Twitter, Instagram, Snapchat or other social media sites
- Concerns:
  - Regulatory hurdles
  - Privacy issues
- Easy to manipulate

Source: http://www.fico.com/en/blogs/analytics-optimization/using-alternative-data-in-credit-risk-modeling/

Social Network Analysis

- Map a consumer's network
- Identify all the files and accounts for a single customer
- Identify the individual's connections with other people
- Especially useful when evaluating a new credit applicant with no or little credit history. The credit ratings of the applicant's network can provide useful information.
- Potential regulatory issue

Source: http://www.fico.com/en/blogs/analytics-optimization/using-alternative-data-in-credit-risk-modeling/

Other Data Sources

- **Clickstream Data.**
  - How a customer moves through your website, where they click and how long they take on a page can be predictive.
- **Audio and Text Data.**
  - Information found on credit applications
  - In recorded customer service or collections calls.
  - Complement “thin files” / lacked credit history

Source: http://www.fico.com/en/blogs/analytics-optimization/using-alternative-data-in-credit-risk-modeling/

“Alternative”

- Data showing **trends or patterns** in traditional loan repayment data.
- **Payment data** relating to non-loan products requiring regular (typically monthly) payments, such as telecommunications, rent, insurance, or utilities.
- Checking account transaction and **cashflow data** and information about a consumer’s assets, which could include the regularity of a consumer’s cash inflows and outflows, or information about prior income or expense shocks.

Source: https://files.consumerfinance.gov/f/documents/20170214_cfpb_Alt-Data-RFI.pdf

“Alternative”

- Data that some consider to be related to a consumer’s **stability**, which might include information about the frequency of changes in residences, employment, phone numbers or e-mail addresses.
- Data about a consumer’s **educational** or **occupational** attainment, including information about schools attended, degrees obtained, and job positions held.
- **Behavioural data** about consumers, such as how consumers interact with a web interface or answer specific questions, or data about how they shop, browse, use devices, or move about their daily lives.
- Data about consumers’ friends and associates, including data about **connections** on social media.

Source: https://files.consumerfinance.gov/f/documents/20170214_cfpb_Alt-Data-RFI.pdf

Characteristics of Good Alternative Data

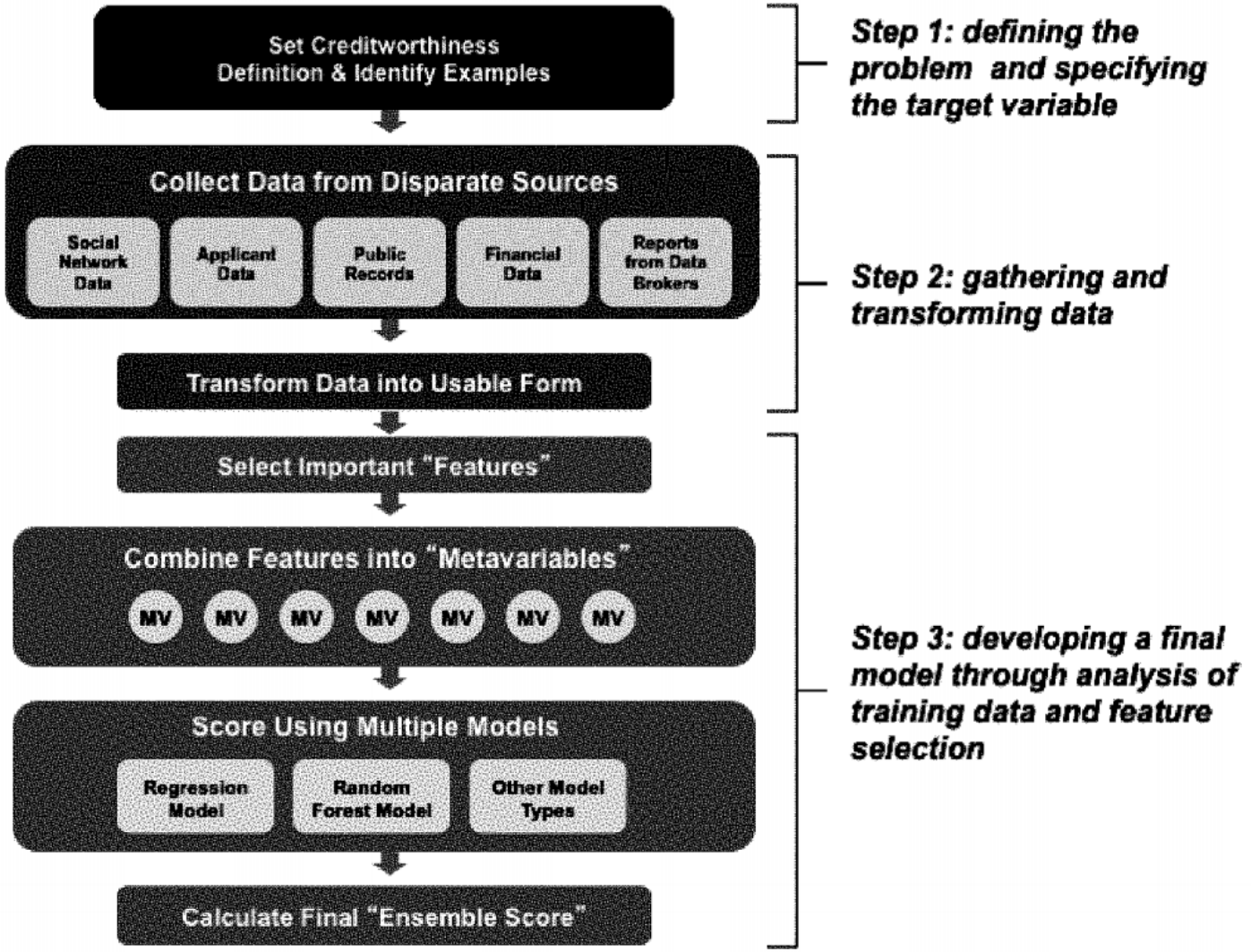
- **Coverage**: a new data source will ideally have broad and consistent coverage
- **Specificity**: a data source should ideally contain detailed data elements about an individual – data elements that provide part of a full picture of the borrower (e.g., ontime and late payments over a significant time series, or specific asset or income data)
- **Accuracy and timeliness**: data should be accurate and frequently updated
- **Predictive power ('signal')**: most important, data should contain information relevant to the behaviour that you're trying to predict
- **Orthogonality**: ideally, the data source should be additive to traditional bureau data; this means that using it will improve the predictive accuracy of any new score by improving the signal-to-noise ratio
- **Regulatory compliance**: data sources must comply with existing regulations for consumer credit (i.e., Fair Credit Reporting Act, Equal Credit Opportunity Act, Gramm-Leach-Bliley Act)

Source: Alternative Data and the Unbanked https://owy.mn/2Gxwwn7

Company & Product	Example Data Inputs
LexisNexis – RiskView	Residential stability, asset ownership, life-stage analysis, property deeds and mortgages, tax records, criminal history, employment and address history, liens and judgments, ID verification, and professional licensure.
FICO – Expansion Score	Purchase payment plans, checking accounts, property data, public records, demand deposit account records, cell and landline utility bill information, bankruptcy, liens, judgments, membership club records, debit data, and property asset information.
Experian – Income Insight	Rental payment data, public record data.
Equifax – Decision 360	Telco utility payments, verified employment, modeled income, verified income, spending capacity, property/asset information, scheduled monthly payments, current debt payments, debt-to-income ratio, bankruptcy scores.
TransUnion – CreditVision	Address history, balances on trade lines, credit limit, amounts past due, actual payment amount.
ZestFinance	Major bureau credit reports and thousands of other variables" including financial information, technology usage, and how quickly a user scrolls through terms of service.
LendUp	Major bureau credit reports, social network data, how quickly a user scrolls through its site.
Kreditech (Not available in U.S.)	Location data (e.g., GPS), social graphing (likes, friends, locations, posts), behavioral analytics (movement and duration on a webpage), e-commerce shopping behavior, device data (apps installed, operating systems).
Earnest	Current job, salary, education history, balances in savings or retirement accounts, online profile data (e.g., LinkedIn), and credit card information.
Demyst Data	Credit scores, occupation verification, fraud checks, employment stability, work history, and online social footprint.

Source: Credit Scoring in The Era of Big Data, https://bit.ly/2vdxhi9

Fig. 1: ZestFinance's modeling & scoring process[91]



Source: Credit Scoring in The Era of Big Data, https://bit.ly/2vdxhi9

Machine Learning in Finance

Supervised Machine Learning

Anthony Ng @ 2018

Machine Learning in Finance

Algorithmic Trading and Portfolio Allocation

Anthony Ng @ 2018

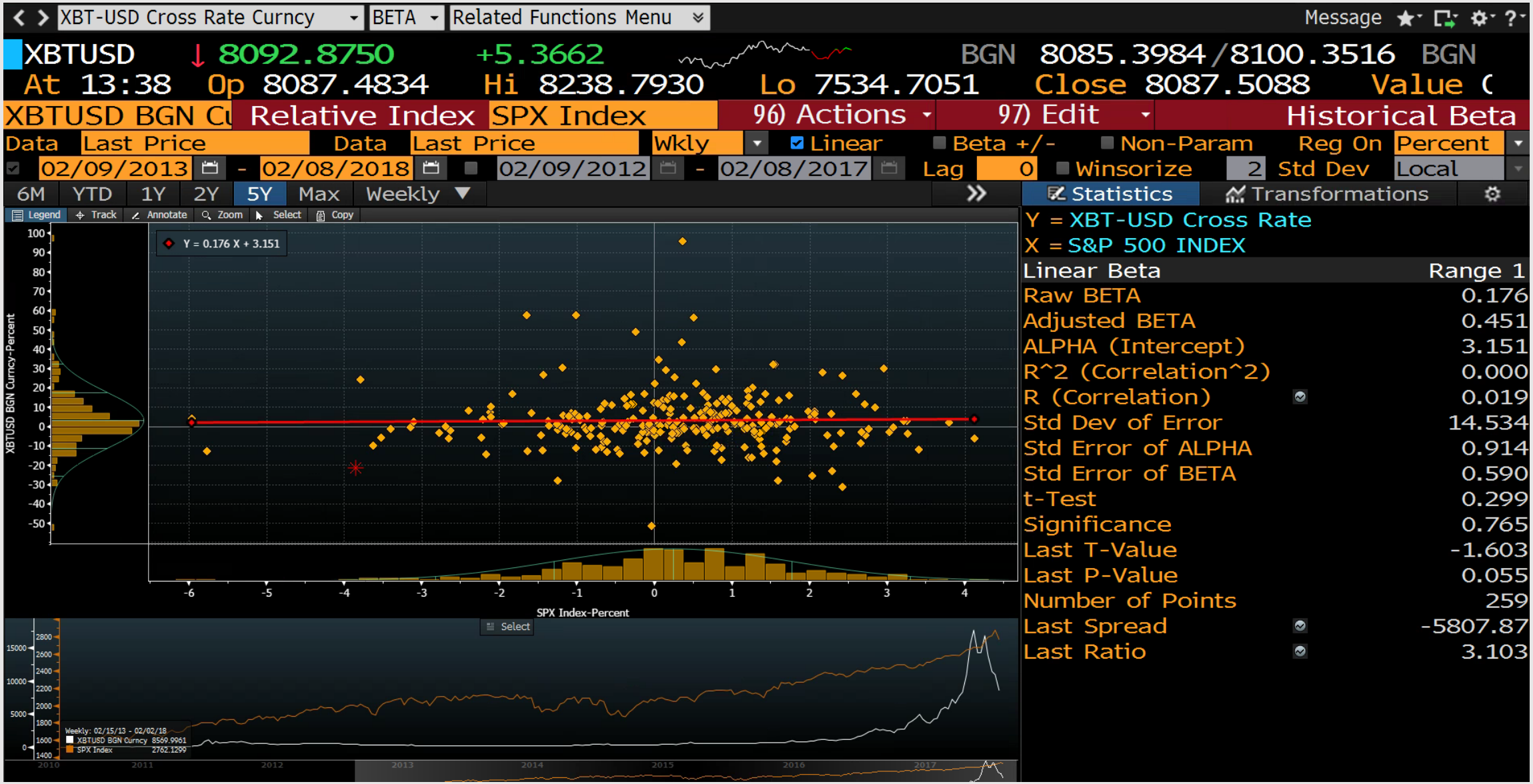
Algorithmic Trading

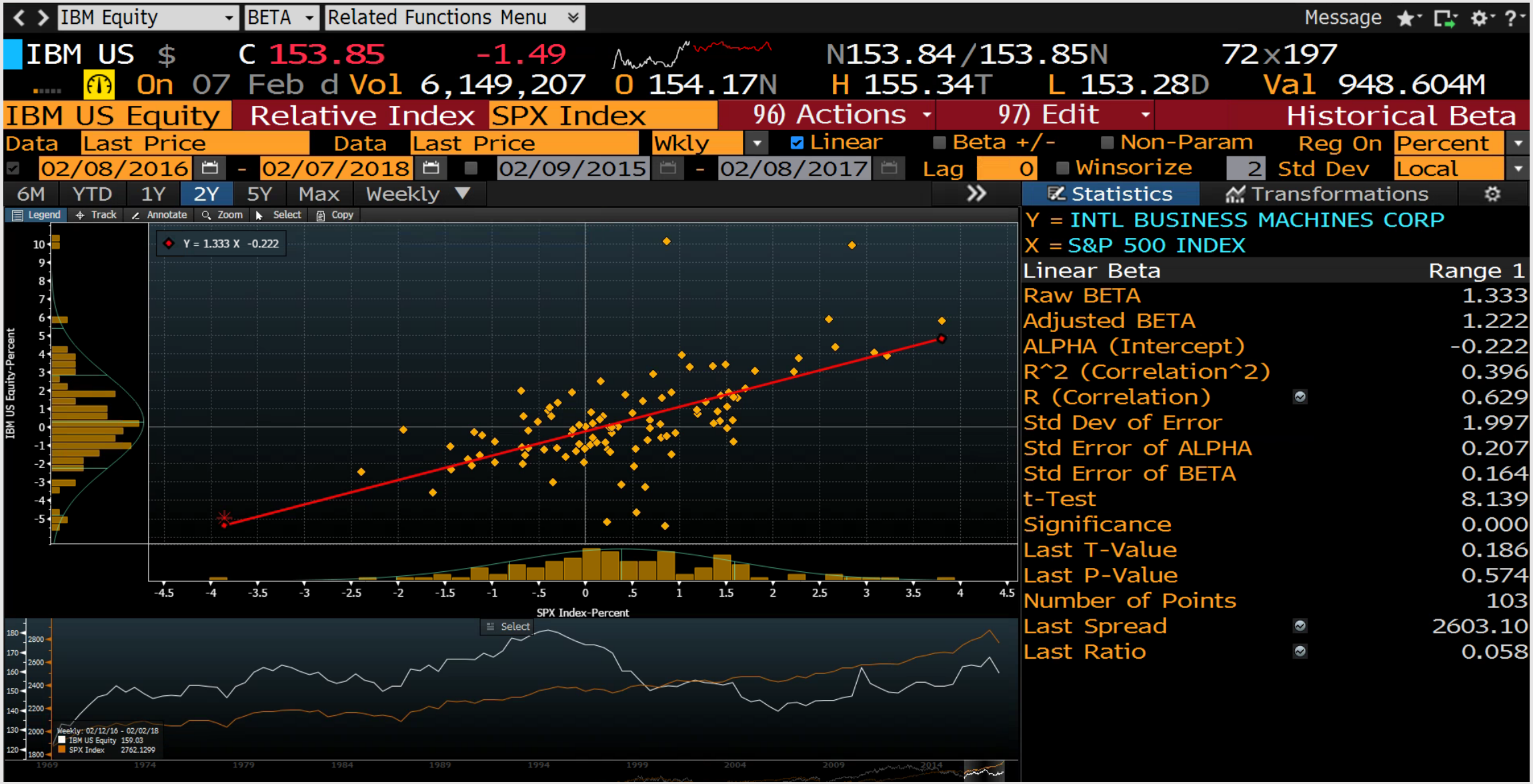
What is Algorithmic Trading?

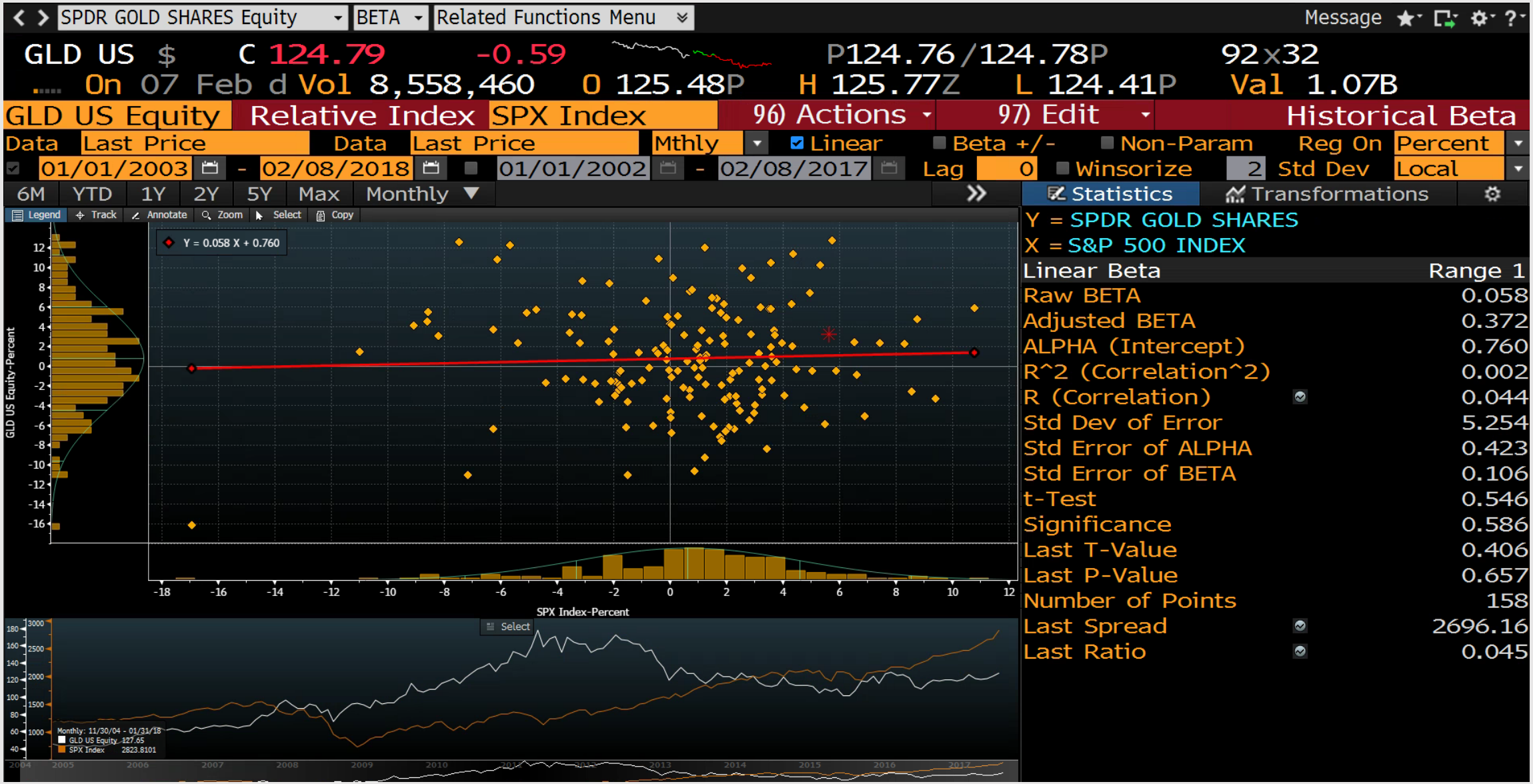
- Covered the method of execution large orders by breaking them to small trade size and filling them without significantly moving the market.
- Systematic trading with set rules in the form of algorithm(s) and little or no human intervention with a profit objective.

Pairs Trading as a Motivational Example

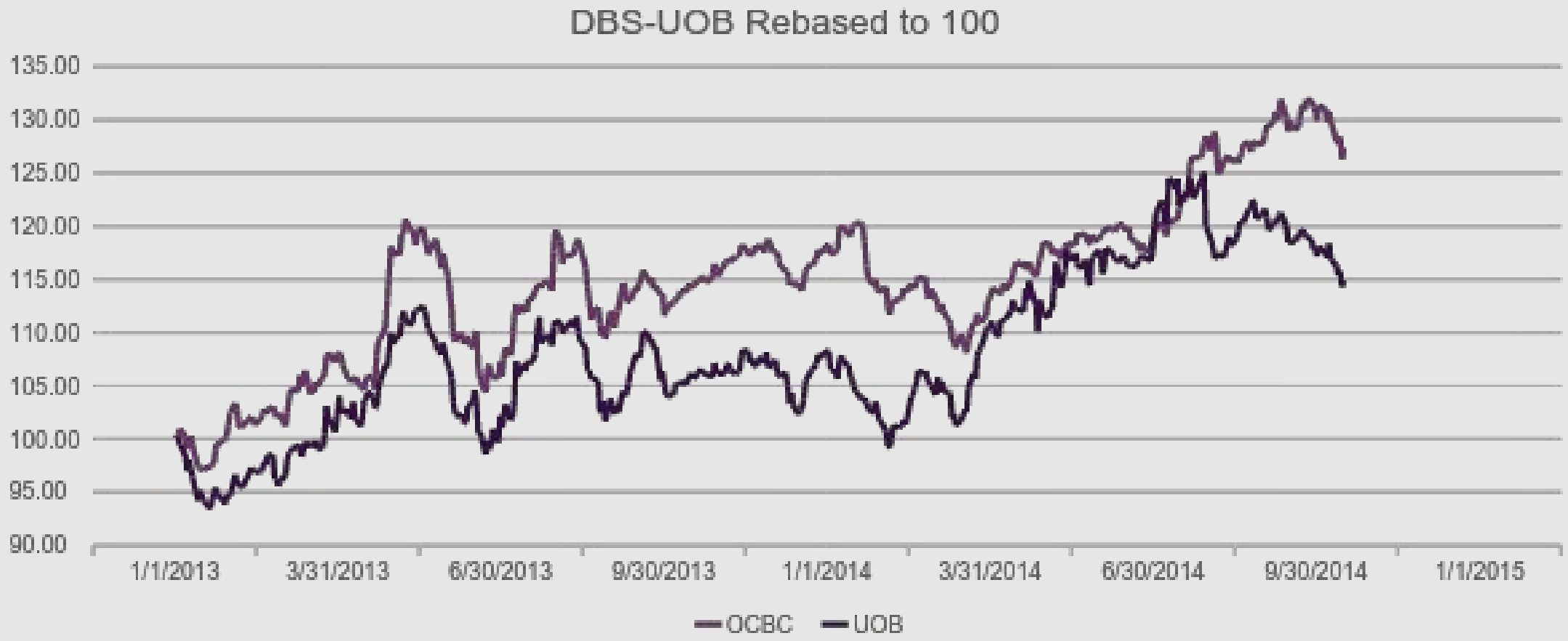
- Market neutral trading strategy
  - Uncorrelated with the market return. In CAPM terminology, beta is zero or near zero
- Profit in all market conditions
- Constructed with only two assets
- Concurrently long and short


XBT-USD Cross Rate Curncy
BETA
Related Functions Menu
Message
XBTUSD ↓ 8092.8750 +5.3662 BGN 8085.3984/8100.3516 BGN
At 13:38 Op 8087.4834 Hi 8238.7930 Lo 7534.7051 Close 8087.5088 Value (
XBTUSD BGN Cu Relative Index SPX Index 96) Actions 97) Edit Historical Beta
Data Last Price Data Last Price Wkly Linear Beta +/- Non-Param Reg On Percent
02/09/2013 - 02/08/2018 02/09/2012 - 02/08/2017 Lag 0 Winsorize 2 Std Dev Local
6M YTD 1Y 2Y 5Y Max Weekly
Statistics Transformations
Legend Track Annotate Zoom Select Copy
Y = 0.176 X + 3.151
XBTUSD BGN Curncy-Percent
SPX Index-Percent
Select
Weekly: 02/15/13 - 02/02/18
XBTUSD BGN Curncy 8569.9961
SPX Index 2762.1299
Y = XBT-USD Cross Rate
X = S&P 500 INDEX
Linear Beta Range 1
Raw BETA 0.176
Adjusted BETA 0.451
ALPHA (Intercept) 3.151
R^2 (Correlation^2) 0.000
R (Correlation) 0.019
Std Dev of Error 14.534
Std Error of ALPHA 0.914
Std Error of BETA 0.590
t-Test 0.299
Significance 0.765
Last T-Value -1.603
Last P-Value 0.055
Number of Points 259
Last Spread -5807.87
Last Ratio 3.103

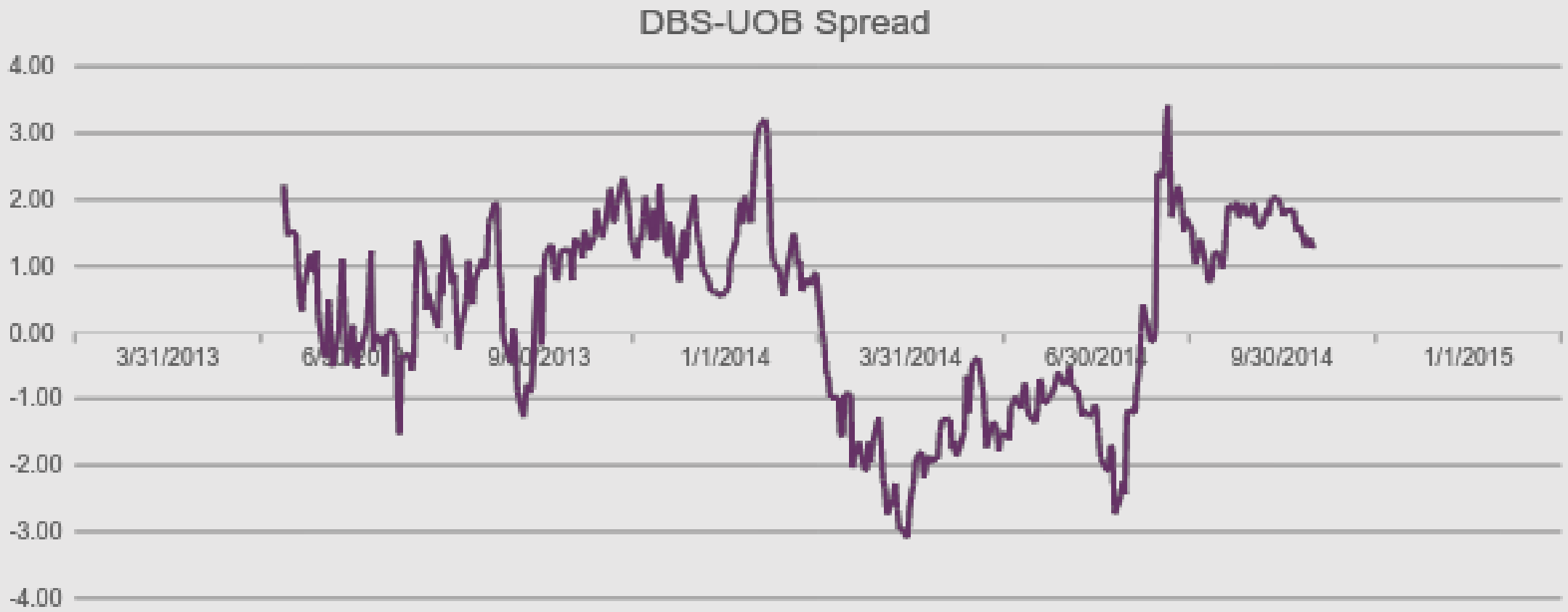

IBM Equity
BETA
Related Functions Menu
Message
IBM US $ C 153.85 -1.49
N153.84/153.85N 72x197
On 07 Feb d Vol 6,149,207 O 154.17N H 155.34T L 153.28D Val 948.604M
IBM US Equity
Relative Index SPX Index
96) Actions
97) Edit
Historical Beta
Data Last Price
Data Last Price
Wkly
Linear
Beta +/-
Non-Param
Reg On Percent
02/08/2016 - 02/07/2018
02/09/2015 - 02/08/2017
Lag 0
Winsorize 2
Std Dev Local
6M YTD 1Y 2Y 5Y Max Weekly
Statistics
Transformations
Legend Track Annotate Zoom Select Copy
Y = 1.333 X -0.222
IBM US Equity-Percent
SPX Index-Percent
Select
Weekly: 02/12/16 - 02/02/18
IBM US Equity 159.03
SPX Index 2762.1299
Y = INTL BUSINESS MACHINES CORP
X = S&P 500 INDEX
Linear Beta Range 1
Raw BETA 1.333
Adjusted BETA 1.222
ALPHA (Intercept) -0.222
R^2 (Correlation^2) 0.396
R (Correlation) 0.629
Std Dev of Error 1.997
Std Error of ALPHA 0.207
Std Error of BETA 0.164
t-Test 8.139
Significance 0.000
Last T-Value 0.186
Last P-Value 0.574
Number of Points 103
Last Spread 2603.10
Last Ratio 0.058


SPDR GOLD SHARES Equity
BETA
Related Functions Menu
Message
GLD US $ C 124.79 -0.59
P124.76/124.78P 92x32
On 07 Feb d Vol 8,558,460 O 125.48P H 125.77Z L 124.41P Val 1.07B
GLD US Equity
Relative Index SPX Index
96) Actions
97) Edit
Historical Beta
Data Last Price
Data Last Price
Mthly
Linear
Beta +/-
Non-Param
Reg On Percent
01/01/2003 - 02/08/2018
01/01/2002 - 02/08/2017
Lag 0
Winsorize 2
Std Dev Local
6M YTD 1Y 2Y 5Y Max Monthly
Statistics
Transformations
Legend Track Annotate Zoom Select Copy
Y = 0.058 X + 0.760
GLD US Equity-Percent
SPX Index-Percent
Select
Monthly: 11/30/04 - 01/31/18
GLD US Equity 127.65
SPX Index 2823.8101
Y = SPDR GOLD SHARES
X = S&P 500 INDEX
Linear Beta Range 1
Raw BETA 0.058
Adjusted BETA 0.372
ALPHA (Intercept) 0.760
R^2 (Correlation^2) 0.002
R (Correlation) 0.044
Std Dev of Error 5.254
Std Error of ALPHA 0.423
Std Error of BETA 0.106
t-Test 0.546
Significance 0.586
Last T-Value 0.406
Last P-Value 0.657
Number of Points 158
Last Spread 2696.16
Last Ratio 0.045

Simple Pairs Trading Example

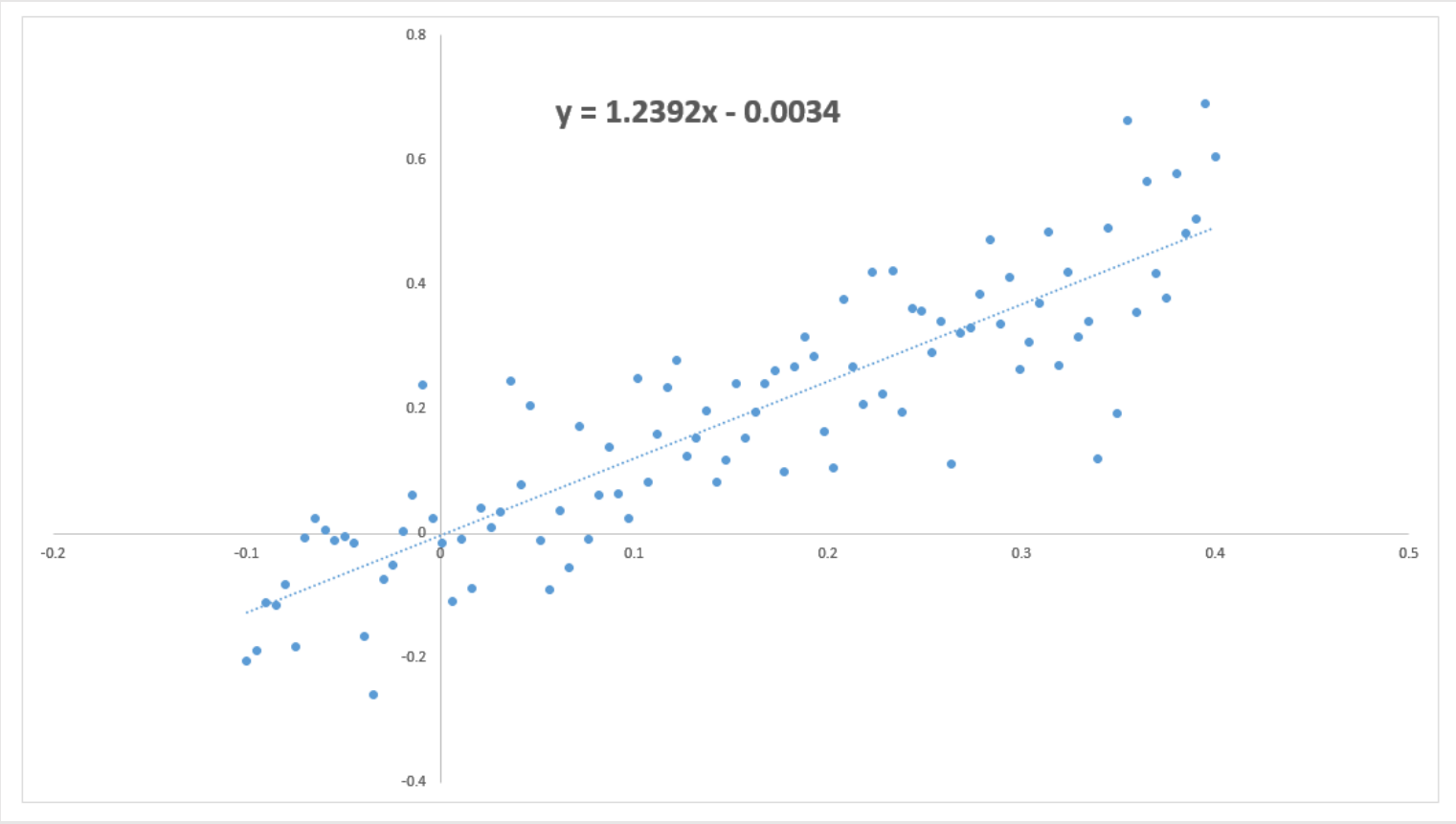


Simple Pairs Trading Example

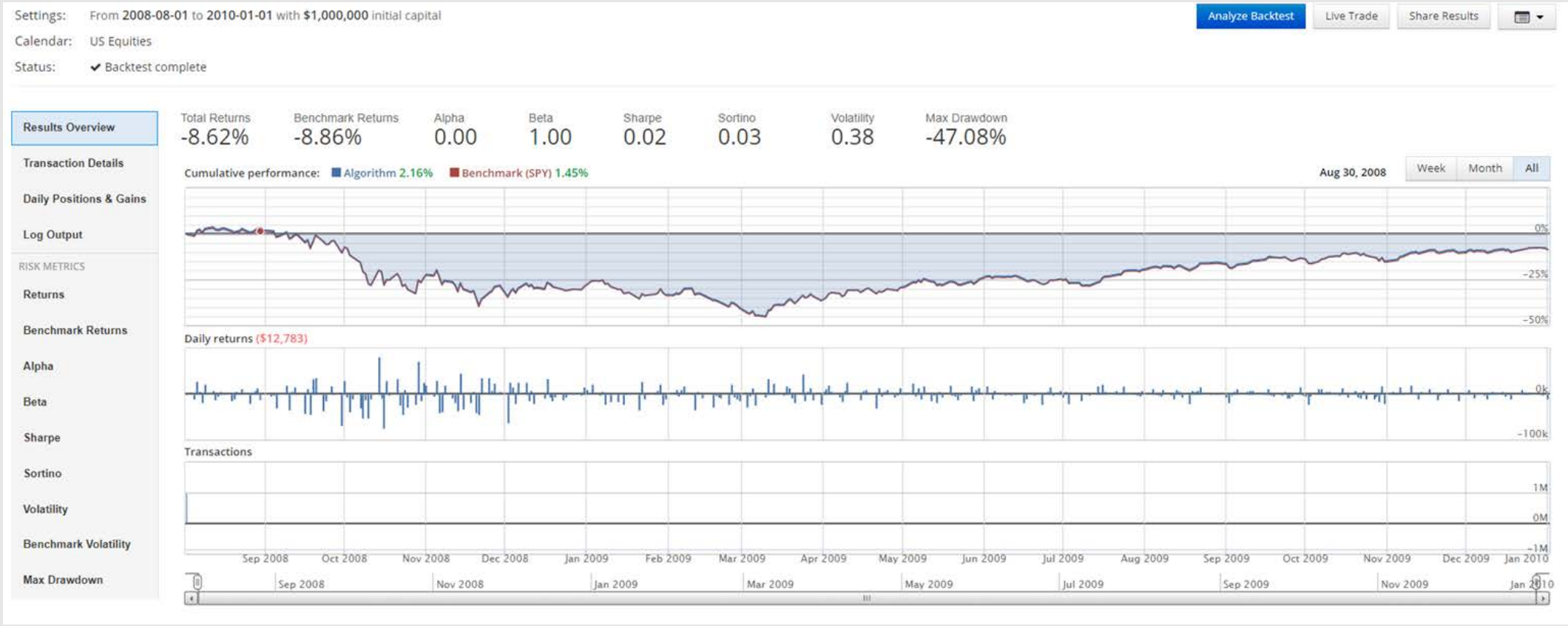


Performance Metrics

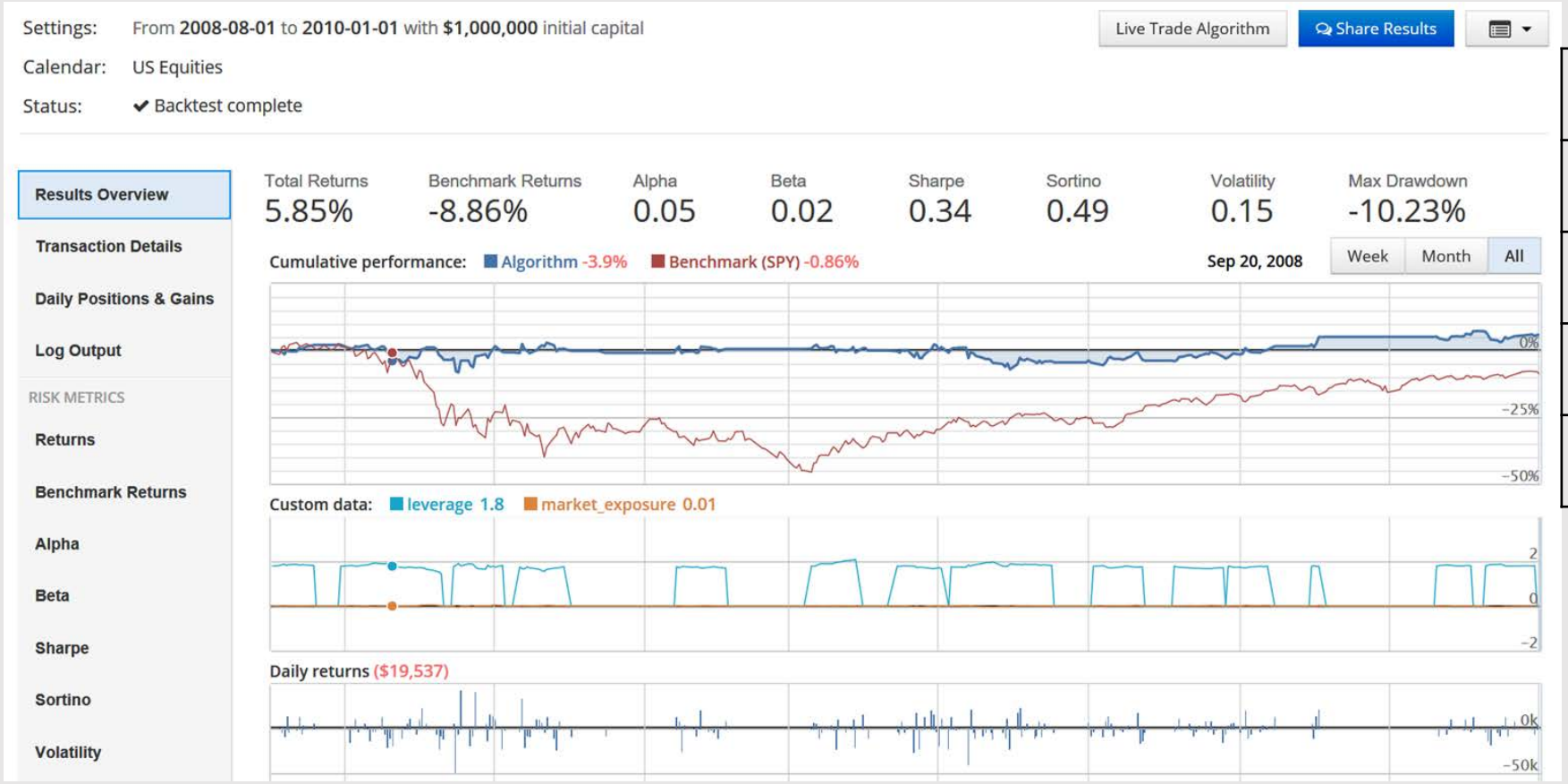
- Total Returns
- Benchmark Returns
- Volatility
- Sharpe Ratio
  - (Risk Premium) / SD



Benchmark – Long Only SPY



Pairs Trading Backtest Example



	SPY	Pairs
Total Returns	-8.62%	5.85%
Sharpe	0.02	0.34
Volatility	38%	15%
Max Drawdown	-47%	-10%

Pairs Trading Technology

- Distance Method
- Cointegration Method
  - Linear Regression
  - Penalised Regression – Ridge, Lasso, ElasticNet
  - Kalman Filters
  - Cointegrated Augmented Dickey Fuller
- Companies similarity
  - Clustering
  - Sentiment based
- Market Regime Detection
  - Hidden Markov Models
  - RNN / LSTM
- Reinforcement Learning

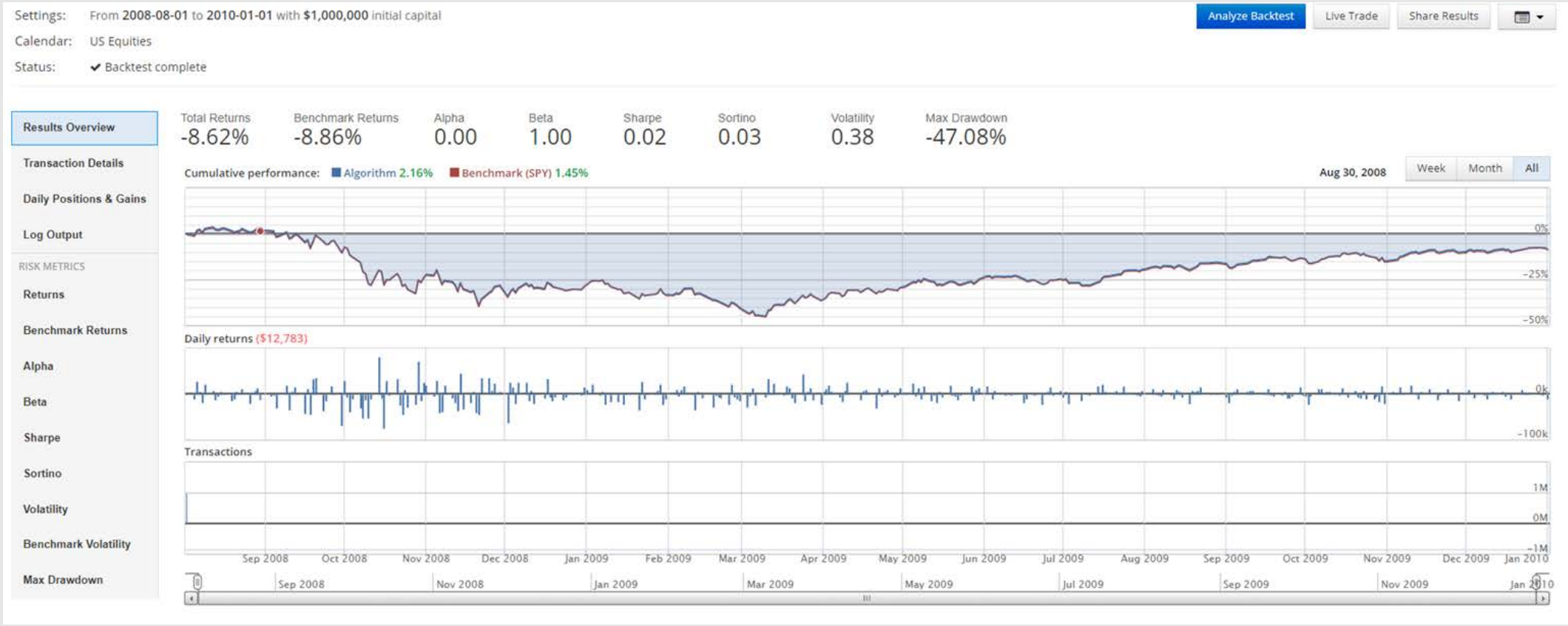
Quant Workflow

- Data
- Universe Definition
- Alpha Discovery
- Alpha Combination
- Portfolio Construction
- Trading

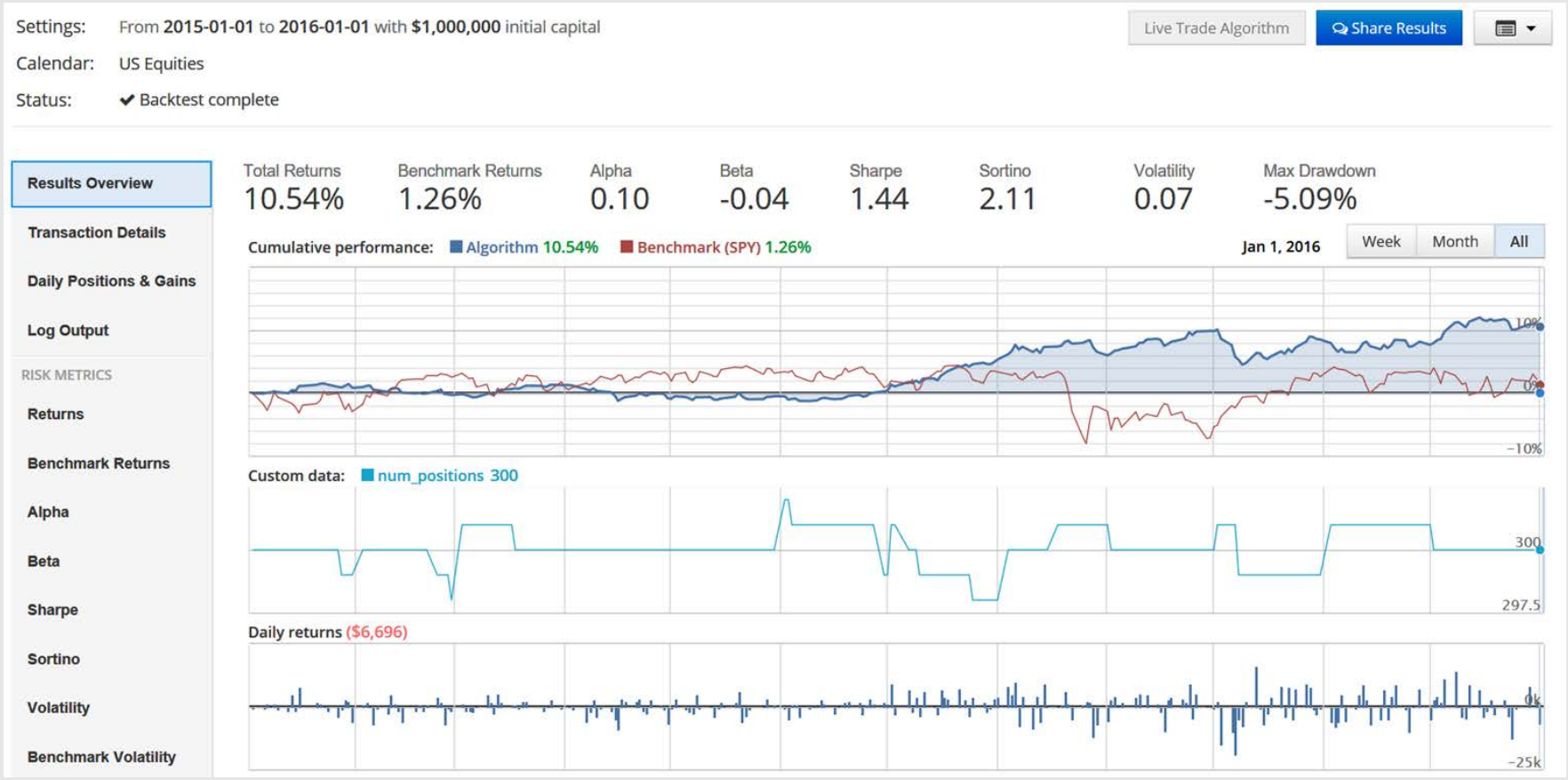
Factor Based Investing

- Exposure to factor risk earns a risk premium
- Types of Factors
  - Macroeconomic Factors: Growth, Inflation, Political Risk, etc.
  - Static Factors: Equities, Bonds, etc.
  - Dynamic Factors: Value-Growth, Momentum, Low Volatility, Size, Liquidity, Credit Risk, etc.
- CAPM
  - Single Factor: Market Factor
- Multifactor Models:
  - Fama-French:
    - Market
    - Size (**S**mall [Market Capitalisation] **M**inus **B**ig)
    - Book/Price (**H**igh pbook-to-market] **M**inus **L**ow). High book-to-market = value stocks.
  - $r_i = r_f + \beta_M\left(r_M - r_f\right) + \beta_{SMB}SMB + \beta_{HML}HML + \alpha$
  - Both SMB and HML are zero-cost portfolios. That's another term for dollar neutral.

Benchmark – Long Only SPY



Long-Short Backtest Example



	SPY	L/S
Total Returns	0.33%	10.54%
Sharpe	0.10	1.44
Volatility	16%	7%
Max Drawdown	-11.8%	-5%

Factor Based Technology

- Factor identifications
  - Deep learning
  - Alternative datasets
    - Slides (https://slides.com/anthonytyng/alternative_data/live)
    - Video (https://youtu.be/6GUKFoZKf5I)
    - Consumer transaction. E.g., POS data
    - Social Media / Sentiment
    - Online search
    - IoT
    - Satellite & Weather
    - etc.