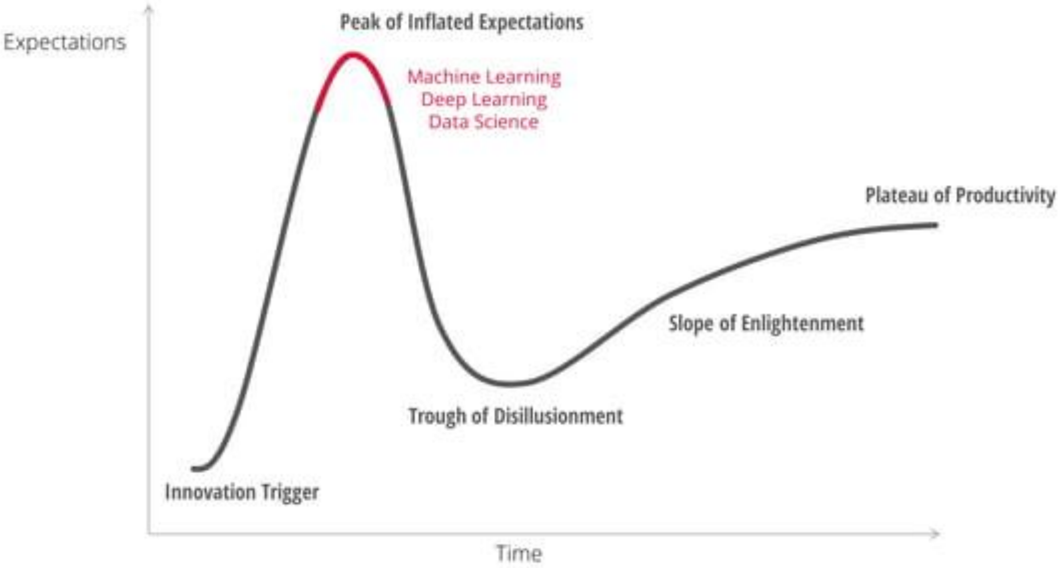
Exploring The Data Science Process


Data
Science

Gartner's Hype Cycle



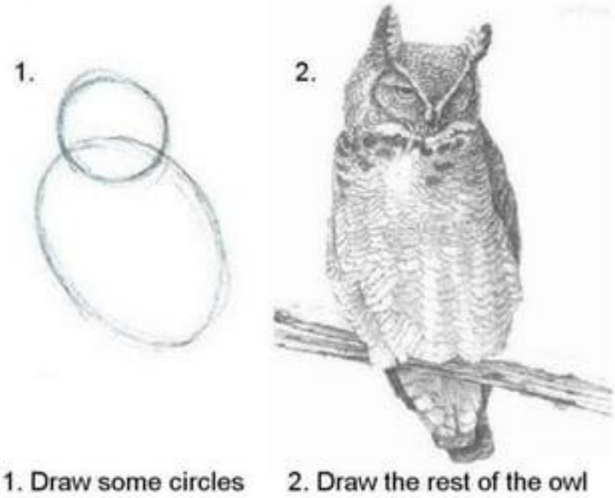
1

BUILD
A MACHINE LEARNING MODEL
IN JUST THREE
QUICK AND EASY STEPS
USING [...]!!!

– Most tutorials

How to Become a Data Scientist?

How to Draw An Owl



2

50% of analytic projects fail.

– Gartner, 2015

Data + Machine Learning = Profit

https://xkcd.com/1838/

On September 21, 2009, the grand prize of US$1,000,000 was given to the BellKor's Pragmatic Chaos team which bested Netflix's own algorithm for predicting ratings by 10.06%.

"[T]he additional accuracy gains that we measured did not seem to justify the engineering effort needed to bring them into a production environment."

Netflix Technology Blog

Learn more about how Netflix designs, builds, and operates our systems and engineering organizations

Apr 5, 2012

Analytic projects fail because...

...they aren't completed within budget or on schedule,

or because they fail to deliver the features and benefits

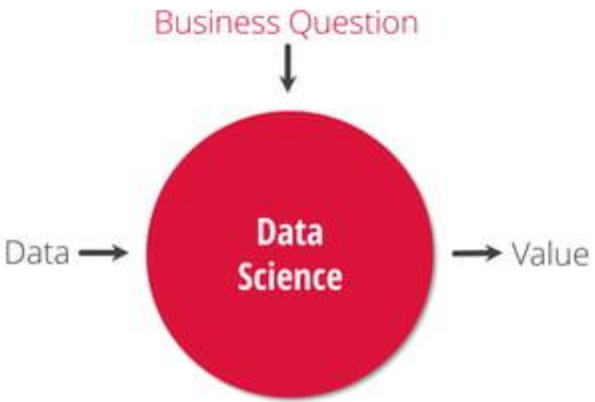
that are optimistically agreed on at their outset.

How to Avoid Failure?

1. Build with Organizational Buy-in
2. Build with End In Mind
3. Build with a Structured Approach

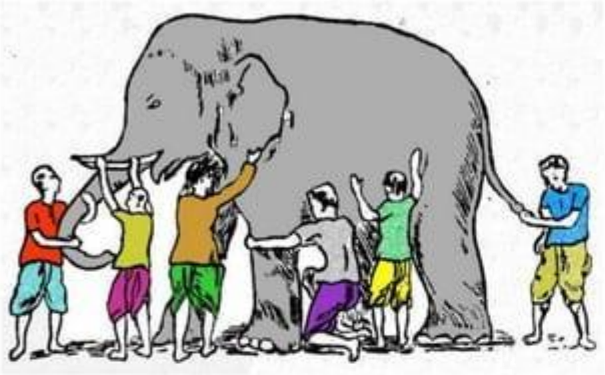
How to Avoid Failure?

1. Build with Organizational Buy-in
2. Build with End In Mind
3. Build with a Structured Approach


Business Question
Data
Data
Science
Value


Data
Science

The Blind Men and the Elephant

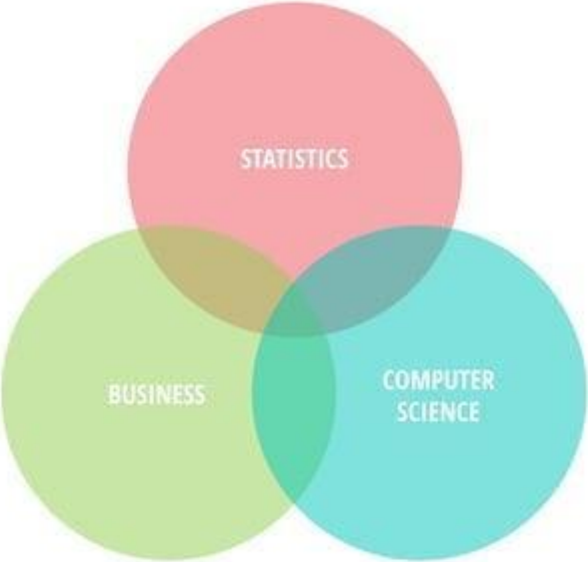


It was six men of Indostan
To learning much inclined,
Who went to see the Elephant
(Though all of them were blind),
That each by observation
Might satisfy his mind.

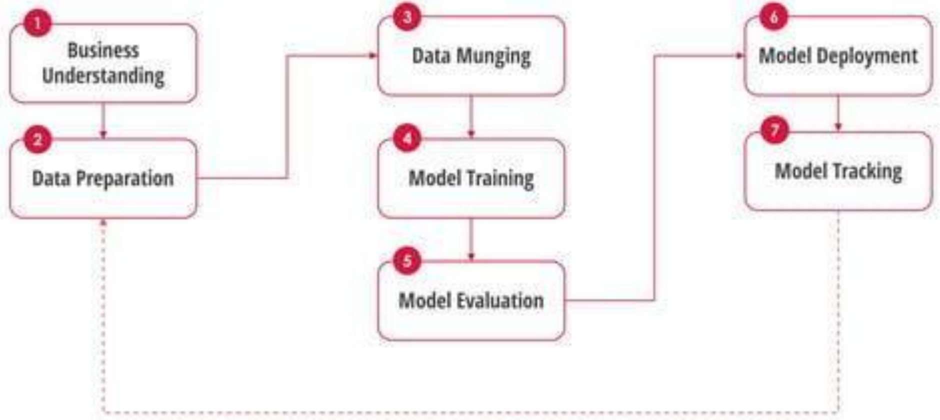
And so these men of Indostan
Disputed loud and long,
Each in his own opinion
Exceeding stiff and strong,
Though each was partly in the right
And all were in the wrong!

– John Godfrey Saxe

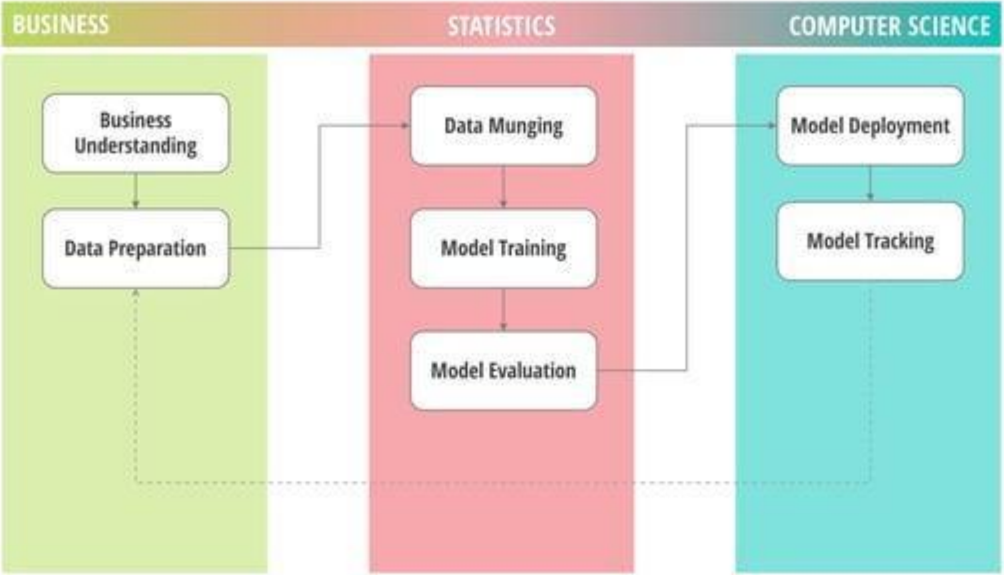
Data Science



Data Science Process



The Data Science Process



Business Understanding	Data Preparation	Data Munging	Model Training	Model Evaluation	Model Deployment	Model Tracking

Far better

an **approximate** answer to the **right** question

than

an **exact** answer to the **wrong** question.

– John Tukey

1 DETERMINE

2 UNDERSTAND

3 MAP

What does the client want to achieve?

1 DETERMINE

2 UNDERSTAND

3 MAP

Primary Objective

- Reduce attrition
- Customized targeting
- Plan future media spend
- Prevent fraud
- Recommend Products

1 DETERMINE

2 **UNDERSTAND**

- Understand **success criteria**.
  - Specific, measurable, time-bound
- List **assumptions, constraints,** and **important factors**.
- Identify **secondary or competing objectives**.
- Study **existing solutions** (if any).

3 MAP

1 DETERMINE

2 UNDERSTAND

3 MAP

Business Objective → Technical Objective

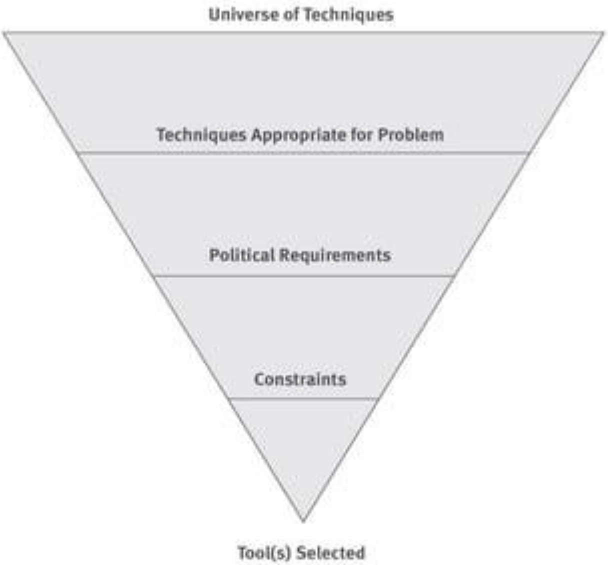
- State the **project objective(s) in technical terms**.
- Describe how the data science project will **help solve the business problem**.
- Explore **successful scenarios**.

1 DETERMINE

2 UNDERSTAND

3 MAP

OBJECTIVE →	TECHNIQUE →	EXAMPLES
Predict Values	Regression	Linear regression, Bayesian regression, Decision Trees
Predict Categories	Classification	Logistic regression, SVM, Decision Trees
Predict Preference	Recommender System	Collaborative / Content-based filtering
Discover groups	Clustering	*k*-means, Hierarchical clustering
Identify unusual data points	Anomaly Detection	*k*-NN, One-class SVM
...		


Universe of Techniques
Techniques Appropriate for Problem
Political Requirements
Constraints
Tool(s) Selected

If all you have is a hammer
then everything looks like a nail.

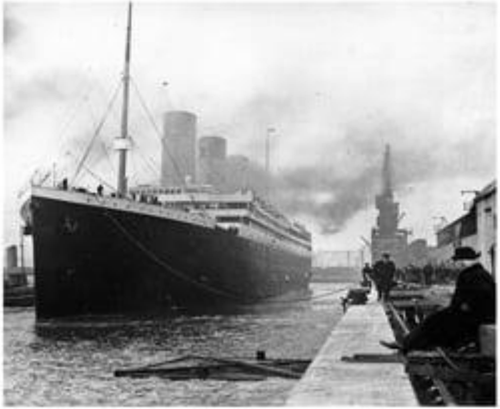
- **Primary Objective:** Prevent attrition → Increase subscription renewals
- **Competing Objective:** High value customers are also targeted for up-sell
- **Constraints:** Avoid targeting customers too close to their contract expiration
- **Success Criteria:** Current renewal rate = 65% → Improve by 8% within the next quarter
- **Existing Solution:** Business-rule-based targeting
- **Data Science Objective:** Build a binary classification model to identify customers who are not likely to renew their subscriptions at least three months in advance of their contract expiration.
- **Success Scenario:** The model correctly identifies 80% of the future attritors; the promotional campaign targets all likely attritors, and successfully converts 20% of them into non-attritors.

Project Plan

- Duration
- Inventory of resources
- Tools and techniques
- Risks and contingencies
- Costs and benefits
- Milestones

The thought that disaster is impossible often leads to an unthinkable disaster.

– Gerald Weinberg



Titanic at Southampton docks, prior to departure

1. IDENTIFY
2. COLLECT
3. ASSESS
4. VECTORIZE

1 IDENTIFY

2 COLLECT

3 ASSESS

4 VECTORIZE

- **Data sources, formats**
  - Database, Streaming API's, Logs, Excel files, Websites, etc.
- **Entity Relationship Diagram** (ERD)
- Identify **additional data sources**.
  - Demographics data appends,
  - Geographical data,
  - Census data, etc.
- Identify **relevant data**.
- Record **unavailable data.**
- How long a history is available and one should use?

1 IDENTIFY

2 COLLECT

- Access or acquire all relevant data in **a central location**
- **Quality control checks and tests**
  - File formats, delimiters
  - Number of records, columns
  - Primary keys

3 ASSESS

4 VECTORIZE

1 IDENTIFY

2 COLLECT

3 ASSESS

4 VECTORIZE

First look at the data

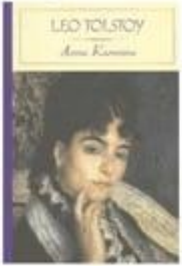
- **Get familiar** with the data.
- Study **seasonality**.
  - Monthly/weekly/daily patterns
  - Unexplained gaps or spikes
- Detect **mistakes**.
  - Extreme or outlier values
  - Unusual values
  - Special missing values
- Check **assumptions**.
- Review **distributions**.

Trust, but verify.

Tidy datasets
~~Happy families~~ are all alike;

Every ~~unhappy family~~ is ~~unhappy~~ in its own way.
messy dataset messy

- Hadley Wickham



GOAL: Create the Analysis Dataset

1 IDENTIFY

2 COLLECT

3 ASSESS

4 **VECTORIZE**

$$y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix}$$

Outcome
Target / Labels
Independent Variable

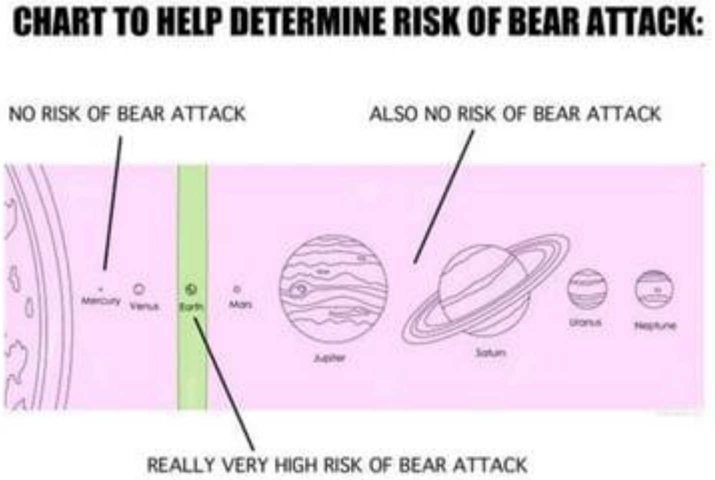
$$X = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1j} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2j} \\ x_{31} & x_{32} & x_{33} & \dots & x_{3j} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{nj} \end{pmatrix}$$

Inputs
Features / Attributes
Dependent Variables

Target Definition

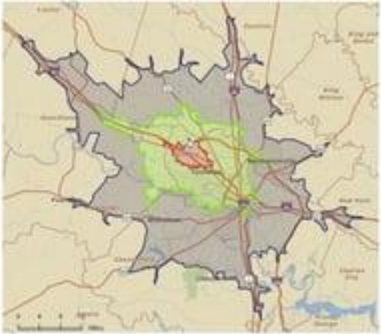
- **Churn = 90 days of consecutive inactivity** (for a pre-paid telecom customer)
- What's **inactivity**?
  - Incoming and outgoing calls
  - Data usage
  - Incoming text
  - Promotional texts
  - Voicemail usage
  - Call forwarding
  - Etc.
- Customers may **change their device** or phone number.
  - Churn at the individual (person) level, or at the device (phone) level?
- Customers may return (become active again) after 90 days of inactivity?
- Prediction window
  - Predict 90 days of consecutive inactivity?
  - Would 10 days of consecutive inactivity suffice?
  - How many customers return after *x* days of inactivity?
- Fraud, Involuntary churn
- Etc.

Accurate but not Precise

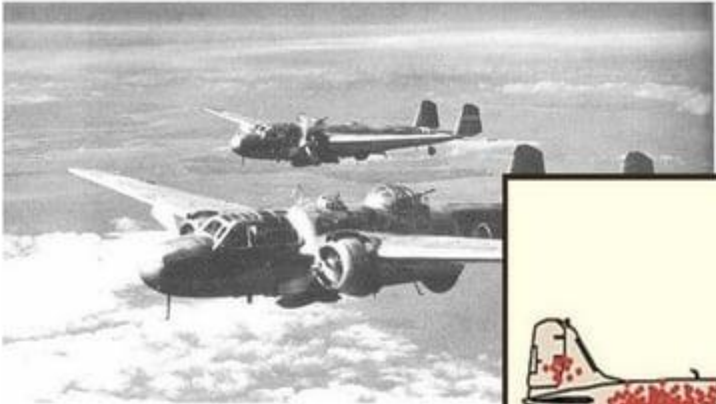


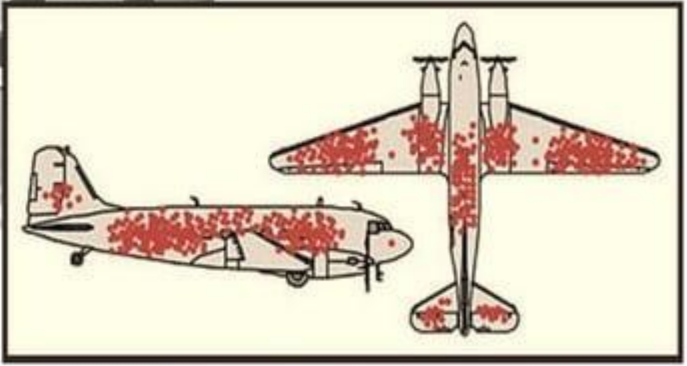
Modeling Sample

- **Historical trends and seasonality**
  - Are there certain timeframes that should be discarded?
  - The model should be generalizable.
- **Eligible, relevant population**
  - Must align with the business goals
- **Eligible, relevant markets**
  - Must align with the business goals
  - E.g., within a certain drive-time distance
- **Outdated products or events**



Selection Bias





Abraham Wald's Work on Aircraft Survivability
Journal of the American Statistical Association Vol. 79, No. 386 (Jun., 1984)

Information Leakage

- The leading indicators must be calculated from the timeframe *leading up to* the event – it must not overlap with the prediction window.
- Beware of proxy events, e.g., future bookings.

Data Aggregation

- **Attribute creation**
  - Derived attributes: Household income / Number of adults = Income per adult
- **Brainstorm with team members** (both technical and non-technical)

$$X = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1j} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2j} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3j} \\ . & . & . & & . \\ . & . & . & & . \\ . & . & . & & . \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nj} \end{pmatrix}$$

Data Aggregation

CUSTOMER ID	PURCHASE DATE
1001	02-12-2015:05:20:39
1001	05-13-2015:12:18:09
1001	12-20-2016:00:15:59
1002	01-19-2014:04:28:54
1003	01-12-2015:09:20:36
1003	05-31-2015:10:10:02
...	...

CUSTOMER ID	x_1	x_2	...	x_j
1001	...	...		...
1002	...	...		...
1003	...	...		...
...	...	...	...	...

1. Number of transactions (Frequency)
2. Days since the last transaction (Recency)
3. Days since the earliest transaction (Tenure)
4. Avg. days between transaction
5. # of transactions during weekends
6. % of transactions during weekends
7. # of transactions by day-part (breakfast, lunch, etc.)
8. % of transactions by day-part
9. Days since last transaction / Avg. days between transactions
10. ....

1 IDENTIFY

2 COLLECT

3 ASSESS

4 VECTORIZE

OUTPUT: The Analysis Dataset

$$y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix} \qquad X = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1j} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2j} \\ x_{31} & x_{32} & x_{33} & \dots & x_{3j} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{nj} \end{pmatrix}$$

kaggle

coursera

Business Understanding | Data Preparation | Data Munging | Model Training | Model Evaluation | Model Deployment | Model Tracking

Give me six hours to chop down a tree

and I will spend the first four sharpening the axe.

– Anonymous

- **Descriptive statistics**
  - Review with the client
- **Correlation analysis**
  - Review with the client
  - Watch out for data leakage
- **Impute missing values**
- **Trim extreme values**
- **Process categorical attributes**
- **Transformations** (square, log, etc.)
  - Binning / variable smoothing
- **Multicollinearity**
  - Reduce redundancy
- **Create additional feature**
- **Interactions**
- **Normalization** (scaling)

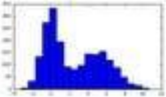
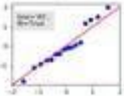
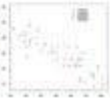
via @vboykis

Machine learning experts display cleaned data samples in preparation for modeling.

Annibale Caracci, c.1600

	Univariate	Multivariate

Non-Graphical	○ Categorical: Tabulated frequencies
	○ Quantitative:
	    ○ Central tendency: mean, median, mode
	    ○ Spread: Standard deviation, inter-quartile range
	    ○ Skewness and kurtosis
		○ Cross-tabulation
		○ Univariate statistics by category
		○ Correlation matrices

Graphical	○ Histograms ○ Box plots, stem-and-leaf plots ○ Quantile-normal plots   	○ Univariate graphs by category (e.g., side-by-side box-plots) ○ Scatterplots ○ Correlation matrix plots   

- **Feature Reduction:** The process of selecting a subset of features for use in model construction
  - Useful for both supervised and unsupervised learning problems

Art is the elimination of the unnecessary.

– Pablo Picasso

Feature Reduction: Why

- **True dimensionality <<< Observed dimensionality**
  - The abundance of redundant and irrelevant features
- **Curse of dimensionality**
  - With a fixed number of training samples, the predictive power reduces as the dimensionality increases. [Hughes phenomenon]
  - With d binary variables, the number of possible combinations is $O(2^d)$.
- **Goal of the Analysis**
  - Descriptive → Diagnostic → Predictive → Prescriptive

    Hindsight Insight Foresight
- **Law of Parsimony** [Occam's Razor]
  - Other things being equal, simpler explanations are generally better than complex ones.
- **Overfitting**
- **Execution time (Algorithm and data)**

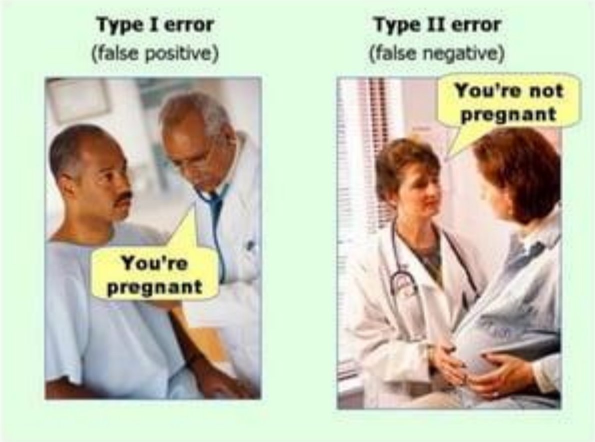
Feature Reduction Techniques

A practical guide to dimensionality reduction techniques – Vishal Patel

1. Percent missing values
2. Amount of variation
3. Pairwise correlation
4. Multicolinearity
5. Principal Component Analysis (PCA)
6. Cluster analysis
7. Correlation (with the target)
8. Forward selection
9. Backward elimination
10. Stepwise selection
11. LASSO
12. Tree-based selection

- Try **more than one machine learning technique**.
- Fine-tune **parameters**.
- Assess **model performance**.
- Avoid **Over-fitting**.

Assess Model Performance



- **New Age:** Area Under the ROC Curve (AUC), Confusion Matrix, Precision, Recall, Log-loss, etc.
- **Old School:** Model Lift, Model Gains, Kolmogorov-Smirnov (KS), etc.

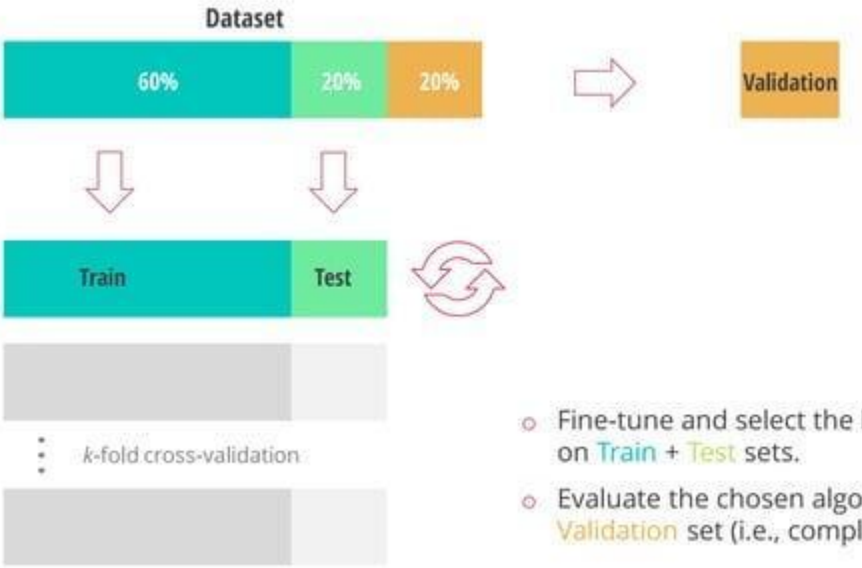
When a measure becomes a target,

it ceases to be a good measure.

Goodhart's law

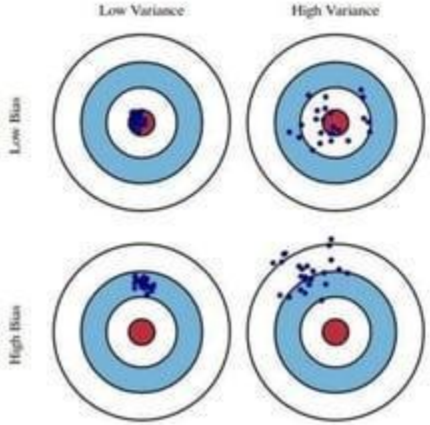
Pic Courtesy: Bizaresis

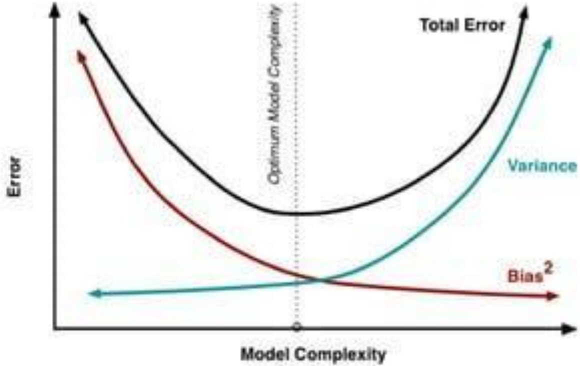
Tri-fold Partition



- Fine-tune and select the best model based on Train + Test sets.
- Evaluate the chosen algorithm on the Validation set (i.e., completely unseen data).

Bias-Variance Tradeoff





With four parameters I can fit an elephant,

and with five I can make him wiggle his trunk.

- John von Neumann

1. MODEL SELECTION

2. ASSESSMENT

3. PRESENTATION

1 **MODEL SELECTION**

- **Law of Parsimony** (Occam's Razor)
- **Model execution time**
- **Deployment complexity**

2 ASSESSMENT

Build the simplest solution that can adequately answer the question.

3 PRESENTATION

1 MODEL SELECTION

2 **ASSESSMENT**

3 PRESENTATION

1 MODEL SELECTION

2 ASSESSMENT

3 PRESENTATION

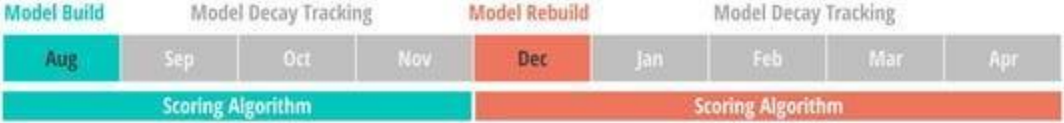
- **AUC, etc.**
- **Cumulative Gains Chart / Lift Chart**
- **Compare against existing business rules/model**
- **Predictor Importance**
- **Each predictor's relationship with the target**
- **Reason-coding**
- **Model usage recommendations**
  - Decile reports
- **Personify**
- **Model peer-review** (Quality Control)

Interpret results as they relate to the business application.

- **Model production cycle**
- **Scoring code, or publish model as a web service**
  - Hand-off
- **Model Documentation** (Technical Specifications)
  - Data preparation, transformations, imputations, parameter settings, etc.
- **Reproducibility**
  - Docker containers
- **Model Persistence vs. Model Transience**

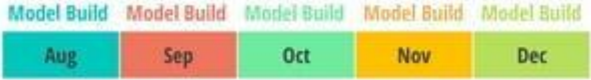
Model Persistence vs. Model Transience

Model Persistence



- Traditional approach
- Provides stability
- Less resource intensive

Model Transience



- Modern approach
- Able to capture recent trends
- Resource intensive

1. MONITOR

2. MAINTAIN

3. TEST

1 MONITOR

- **Model decay tracking (monitoring) plan**
  - Model performance over time
  - Predictor distribution

2 MAINTAIN

3 TEST

1 MONITOR

2 MAINTAIN

- Model maintenance plan
- Adding new data sources
- Version control

3 TEST

1 MONITOR

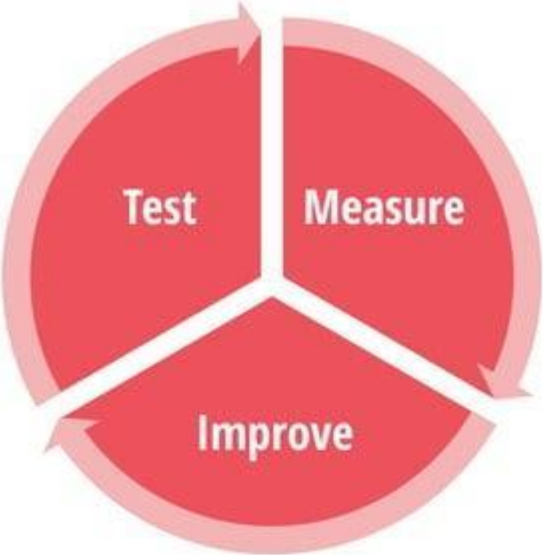
2 MAINTAIN

3 TEST

- **Campaign Set-up and Execution**
  - Experimental Design (A/B tests, Fractional Factorial)

Experimental Design

	Marketing Treatment	No Treatment
Selection Based on Model	A Test	B Selection Hold-out
No Selection (Random)	C Control	D Random Hold-out


Test
Measure
Improve

Data Science Process: Recap

Business Understanding	Data Preparation	Data Munging	Model Training	Model Evaluation	Model Deployment	Model Tracking
Determine	Identify	Impute	Train	Evaluate	Deploy	Monitor
Understand	Collect	Transform	Assess	Peer Review	Document	Maintain
Map	Assess	Reduce	Select	Present		Test
	Vectorize					

DISCUSS COLLATE WRANGLE PERFORM COMMUNICATE EXECUTE TRACK

Process as Proxy

"Good process serves you so you can serve customers.

But if you're not watchful, the process can become the proxy for the result you want.

You stop looking at outcomes and just make sure you're doing the process right.

Gulp.

It's always worth asking, do we own the process or does the process own us?"

– Jeff Bezos

THANK YOU!

vishal@derive.io

www.linkedIn.com/in/VishalJP

DΣRIVE