The Data Science ABCs

AN INTRODUCTION TO ESSENTIAL DATA SCIENCE CONCEPTS, FROM A TO Z.

Introduction

The promise of data science is as elusive as it is extraordinary. Not only are data scientists unicorns who possess a perfect balance of data savvy and business acumen, they have the best, sexiest job in the United States. The field they represent is posited as a business-saving competency, yet executives and data scientists alike struggle to implement it successfully in an enterprise setting. Even the most well-known players in the space have run into challenges.

It’s strange to consider, then, that data science is just a blend of statistics and computer science applied to business problems. It’s not magic. Data scientists spend the majority of their time hunting down and cleaning data. From there, it’s a matter of selecting and fine tuning the framework that will allow a data scientist to find the important information — or patterns — in the data. The complex, messy processes that come after are what keep data-driven insights from reaching an executive’s desk (or better yet, the systems that power a business).

In this ebook, we will explore everything from the buzzwords that are inextricably linked to data science — such as artificial intelligence — to the tools and concepts that support it, like Jupyter notebooks and zero downtime, all in alphabet order.

A is for...

ARTIFICIAL INTELLIGENCE

Artificial intelligence (AI) is a popular catch-all term for systems that perform complex tasks that once required the input of a human, such as chatting online with customers or playing chess. It's often used interchangeably with machine learning and deep learning, two of its subfields. To a data scientist, AI is an important tool for tackling projects that can save time and improve business operations on a massive scale.

Is Your Company Ready To Invest in AI?

ALGORITHM

An algorithm is a mathematical operation or procedure that solves a problem in a series of steps. Essential to data science, algorithms like random forest, logistic regression, and k-means clustering apply different kinds of decision-making operations to a dataset in order to arrive at an output. For example, the k-means clustering algorithm can be used to find distinct groups of customers in a large dataset by identifying inherent similarities between those groups, such as interests or purchase history.

An Introduction to Machine Learning Algorithms

B is for...

BUSINESS INTELLIGENCE

Business intelligence (BI) is a logical first step for companies that want to dabble in big data, but are not yet equipped to manage the technology and personnel requirements of data science. BI analysts are largely concerned with the exploration of past trends, while data scientists are tasked with finding the predictors and significance of those trends. The two types of data professionals pair well together at companies that are looking to forecast business outcomes, refine products, and scale operations.

Data Scientists vs. BI Analysts: What's the Difference?

BOTS

One increasingly popular application of data science in a business setting is the bot. Bots help automate customer service by leveraging a machine learning technique called natural language processing (NLP). Trained by a data scientist using text data like online chat or phone logs from customer support interactions, bots can learn the meaning of human language and even attribute sentiment to words or phrases. They can then “chat” with customers by assessing new text inputs, scoring possible outputs, and selecting the best output to respond to a customer’s question.

Building a Chatbot for Business

C is for...

CHIEF DATA OFFICER

Chief data officers (CDOs) are a recent addition to the C-suite, but Gartner reports that 90% of large global organizations will have hired one by 2019. The CDO's role, though still largely undefined, is centered on managing the complexities of data science, such as technology requirements, data access, and expectations.

5 Challenges Your Chief Data Officer is Likely to Face

D is for...

DATA SCIENTIST

Data scientists have the sexiest job of the 21st century, according to Harvard Business Review. They have topped Glassdoor's 50 Best Jobs in America list three years in a row. But, oddly, there is no universally accepted definition of a data scientist. These analytical unicorns tend to have graduate degrees in computer science, mathematics, or economics; R or Python coding skills; and work primarily in the tech or industrial sectors. Companies typically seek to hire one of two types: Data scientists that focus on answering the question of why something is happening and how to improve it, and data scientists that build machine learning applications that can be deployed into production.

Study: What Skills Does the Typical Data Scientist Have?

DIGITAL TRANSFORMATION

The majority of companies today are striving for digital transformation, an overhaul of processes, programs, and organizational structures that seeks to better leverage digital technologies to improve performance. Digital transformation is the key to staying competitive, and it's driven, in part, by data science: The successful implementation of algorithmic decision making is an indisputable sign of a successful digital transformation.

Fireside Chat: How to Algorithmically Drive Your Digital Transformation

D is for...

DATAOPS

DataOps, or data operations, focus on cultivating data management practices that improve the speed and accuracy of data science and analytics, including data access, quality control, automation, integration, and, ultimately, model deployment and management. At its core, DataOps is about aligning the way data is managed with the goals a business has for that data, resulting in fewer bottlenecks and faster results.

What is DataOps? Everything You Need to Know

DEEP LEARNING

A machine learning technique loosely based on the structure of a biological nervous system, deep learning passes data through interconnected layers of "nodes" that make up a neural network. Each node performs an operation before passing the result on to the next node. Deep learning models require large data sets with which to learn and are typically used in applications that undertake complex deductions, including facial recognition.

What is Deep Learning?

E is for...

ETHICS

Businesses are in the midst of a revolution where the collection and analysis of data can drive more personalized customer experiences, optimize operations, and more. But the proliferation of data also has the potential to create privacy and fairness issues. Any company that delves into data science today has to consider the ethics of how its data is being used — and create practical guidelines that are morally sound, compliant with a growing number of data security laws, and beneficial to the business.

Why a Data Science Code of Ethics is Good for (Your) Business

F is for...

FORECASTING

Forecasting financial results, product performance, and other business outcomes is a popular corporate pastime. When paired with data science, forecasting reaches new levels of accuracy. Data scientists leverage not only historical information, but any number of relevant data sources, such as customers' social media posts and online reviews. Machine learning provides a way for forecasters to glean patterns or trends that would be impossible to pinpoint otherwise due to the size and breadth of a dataset.

A Data Science Framework for Forecasting Opening Box Office Revenue

G is for...

GIT

A version control system originally intended for tracking the progress of software development projects, Git is becoming essential to enterprise data science teams. Using a Git solution like GitHub, Bitbucket, or GitLab, data scientists can work together in parallel on model building and analysis, all while maintaining transparency, code quality, and the reproducibility of their work.

Version Control for Enterprise Data Science Teams

GPUs

Any data scientist who has tried to deploy a machine learning model at scale has experienced some degree of processing lag. In fact, machine learning advancements hit a plateau in the early 2000s due to computing power limitations. It wasn't until the field incorporated graphical processing units (GPUs) that innovation reached today's levels. GPUs can perform computations in parallel, meaning they can process large datasets much faster than traditional central processing units (CPUs).

CPU vs. GPU in Machine Learning

H is for...

HIRING (A DATA SCIENTIST)

One of the most difficult — and essential — parts of bringing data science into a business setting is hiring the right people. That's easier said than done, however. Data scientists have one of the most in-demand roles today, with the number of job posts rising 75 percent from 2015 to 2018, according to Indeed.com. In addition, there is no universally agreed upon job description for data scientists; companies are looking for part statisticians, part software developers, and everything in between.

How to Hire the Right Data Scientist

I is for...

INTERNET OF THINGS (IOT)

Physical devices that connect to a network — like smartphones, smart home appliances, connected cars, and Smart City sensors — are known collectively as the Internet of Things (IoT). The number of IoT devices in the world is projected to grow to over 75 billion by 2025, ultimately providing data scientists with a plethora of new and interesting datasets related to usage and functionality. For example, IoT devices can be used to track assets or monitor system health, providing the data needed to uncover inefficiencies or create preventative maintenance schedules, respectively.

How IoT Reinvents Supply Chain Management

IMAGE RECOGNITION

Thanks to smartphone cameras and social media sites, a huge amount of today's data is visual. It's no surprise, then, that the need for large-scale image recognition systems — such as Facebook's "tag photos" feature — has grown exponentially. Data scientists support this endeavor by building deep learning models and training them to recognize objects in photos or videos. These models require huge amounts of data in order to learn; in some cases, millions of images.

Image Recognition with Deep Learning

J is for...

JUPYTER

Jupyter is just one of the open source notebook options available to data scientists. In a notebook, data scientists can create shareable documents that contain live code and visualizations, as well as plain text explanations of the work they have done. When you consider that collaboration is one of the biggest challenges to implementing data science across an organization, it’s no wonder that tools like Jupyter are vital to the success of such initiatives.

3 Reasons Jupyter Adoption is on the Rise

K is for...

KEY PERFORMANCE INDICATORS (KPIS)

Key performance indicators (KPIs) are just as important to data science teams as they are to other business units. Teams simply perform better when everyone understands the primary objective of a project — and due to the complexity of machine learning, this is especially true for data science projects. Step One of bridging the gap between data scientists and executives is defining KPIs before work even starts.

4 Things to Remember When Defining KPIs

L is for...

LIFETIME VALUE

Lifetime value has been an important metric, especially for marketers, for quite some time. But with the help of data science, predicting the value of future profits generated by a customer is no longer based on past averages. Instead, data scientists can account for variation in customer behavior and different types of business models — such as subscription businesses — to arrive at more accurate conclusions about customer lifetime value. Those conclusions can help businesses focus their efforts on prospects that resemble their most valuable patrons.

An Introduction to Predictive Customer Lifetime Value Modeling

M is for...

MACHINE LEARNING

Machine learning is a branch of artificial intelligence focused on building systems that “learn” — or improve performance — based on the data they consume. Because machine learning models can adapt without being explicitly programmed, they are ideal for quickly transforming large datasets into highly accurate predictions about future outcomes. For example, if a product generates thousands of online reviews and social media posts, a machine learning model could parse that text in order to identify key phrases — and even learn what sentiment is attached to them.

Machine Learning: An Overview

MODELING

Modeling starts with a dataset and ends with a framework that can be used to understand the processes of a business mathematically — and to make predictions about how it will operate in the future. The modeling process is a technical one that requires a data scientist or statistician with knowledge of algorithms. However, creating a model that is truly representative of a business also calls for input from executives or decision makers who are familiar with how the business operates and the problems the model is intended to solve.

An Executive’s Guide to Predictive Modeling

N is for...

NEURAL NETWORKS

In the context of data science, neural networks are frameworks that mimic the structure of biological nervous systems and can be paired with machine learning algorithms to find meaning in large datasets. A neural network passes data through interconnected layers of nodes; each node performs an operation before passing the output onto the next node, eventually arriving at a final result. When a neural network is made of up of a large number of layers, it is considered to be performing deep learning.

Introduction to Neural Nets

NATURAL LANGUAGE PROCESSING

Natural language processing (NLP) is a set of techniques that allows computers to understand human language, whether it is in a text or audio format. NLP is particularly useful for business applications related to customer satisfaction, such as building customer service chatbots and parsing online reviews for information that can inform product or service changes. Systems that use NLP continuously learn from their interactions with new text or audio data, allowing them to improve over time.

3 Business Applications for Natural Language Processing

O is for...

OPEN SOURCE

At one time, proprietary solutions were the mainstay of data analysis at enterprise companies. But with the rise of data science, open source projects that are available to anyone for free — like R, Python, Spark, Jupyter, and TensorFlow — are being embraced by teams that build predictive models and applications. While there are many reasons for this shift, one major one is innovation. Because the open source community is highly collaborative, tools and projects are quickly improved.

Open Source Tools for Enterprise Data Science

P is for...

PYTHON

Python and R are undoubtedly data scientists' most loved programming languages. Why? Both are open source, meaning they are free to use and are constantly being expanded and improved. Created in 1989, Python was built to emphasize code readability and efficiency. Unlike R, it's an object-oriented programming language, which means it groups data and code into objects that can interact with — and even modify — one another. Java, C++, and Scala are other examples of object-oriented languages. This sophisticated approach allows data scientists to execute tasks with better stability, modularity, and code readability.

R vs. Python: What Language is Best for Building Data Models?

Q is for...

QUERY

In order to assemble data sources that can help fuel experimentation, data scientists query, or request information from, a database. Queries must have parameters that tell the database what information is relevant; these are either selected in a database's menu or a querying tool, or written in a query language like SQL. Queries can retrieve data or perform operations on that data.

To SQL Or Not To SQL, That Is The Question

R is for...

RECOMMENDATION ENGINE

Recommendation engines can be found on e-commerce and content sites all over the internet. These systems predict how likely a user is to purchase an item or engage with certain videos or images, and then subsequently serve up the most relevant items to that user. They may sound simple, but recommender systems can be powerful: Netflix reports that the recommendation engine behind its TV and movie suggestions keeps customers from leaving its service to the tune of $1 billion per year.

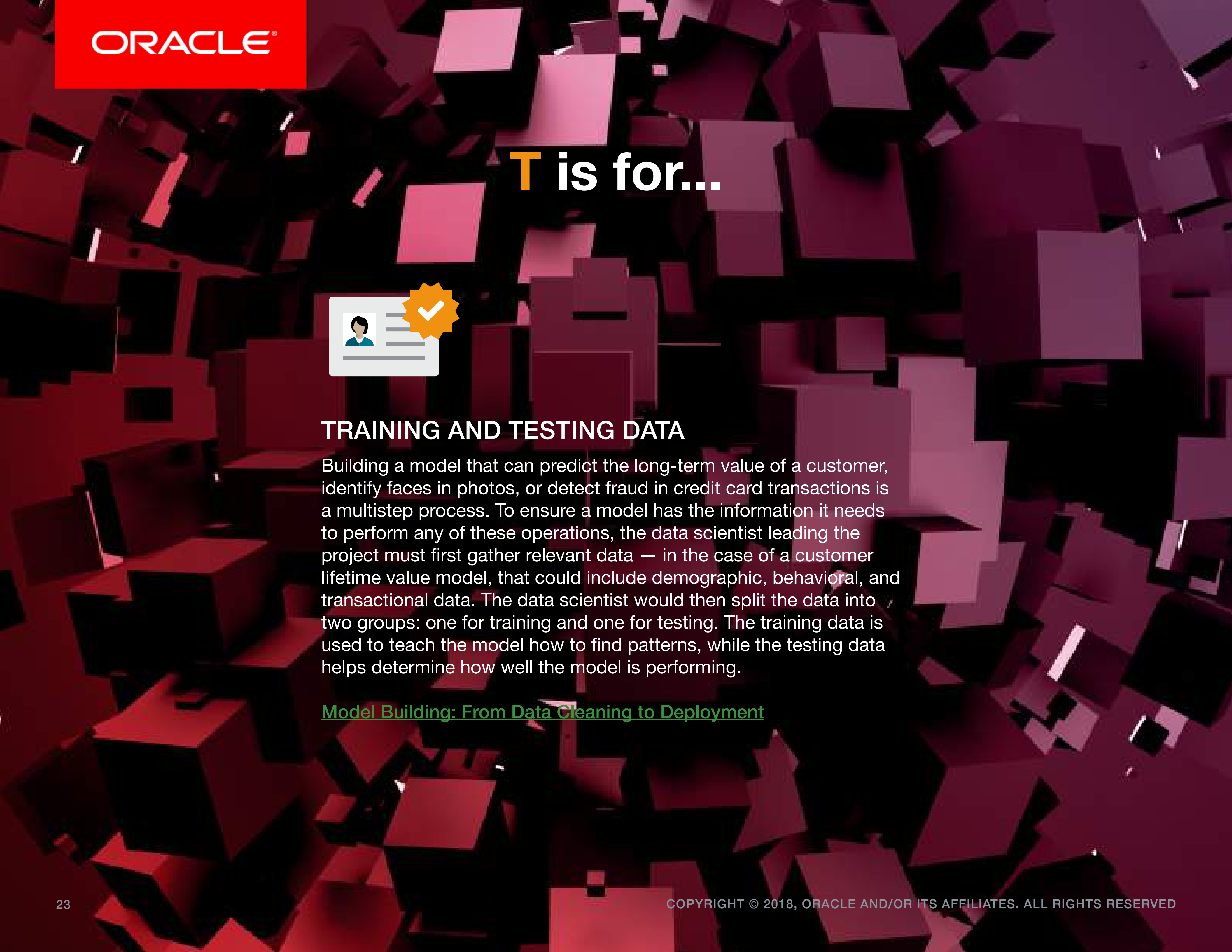
Introduction to Recommendation Engines

S is for...

SCALE

Performing data science at scale means making it the engine that powers every decision across a business. Often, data scientists are performing one-off analyses that get lost in the shuffle of day to day operations. When a data science team successfully scales its efforts, predictive models are integrated into the systems used by employees and decision makers in every department. For example, the outputs of a customer churn model can be pushed to call center software, giving customer service agents more insight into whether a customer is at risk of canceling his or her service.

Scaling Data Science at Airbnb



T is for...

TRAINING AND TESTING DATA

Building a model that can predict the long-term value of a customer, identify faces in photos, or detect fraud in credit card transactions is a multistep process. To ensure a model has the information it needs to perform any of these operations, the data scientist leading the project must first gather relevant data — in the case of a customer lifetime value model, that could include demographic, behavioral, and transactional data. The data scientist would then split the data into two groups: one for training and one for testing. The training data is used to teach the model how to find patterns, while the testing data helps determine how well the model is performing.

Model Building: From Data Cleaning to Deployment

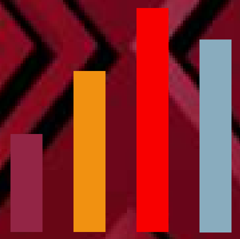
U is for...

UNSUPERVISED MACHINE LEARNING

Unsupervised machine learning is the idea that a computer can learn to identify complex processes and patterns without a human to provide guidance along the way, which is essential to the idea of artificial intelligence. For example, while a supervised classification algorithm learns to ascribe labels to images based on labels provided by a data scientist, its unsupervised counterpart will look at inherent similarities between images and separate them into groups accordingly, assigning its own new label to each group.

Supervised vs. Unsupervised Machine Learning

V is for...



VISUALIZATION

Data visualizations are a great way for data scientists to present their findings to non-technical decision makers so that complex information is more understandable. Data visualizations are pictorial or graphical representations of data that can reveal patterns, trends, and correlations that might otherwise have gone unnoticed. They are an essential part of analyzing data in a business setting.

5 Best Practices for Data Visualization

W is for...

WRANGLING

Data scientists spend the majority of their time not on model building or delivering insights, but on data preparation tasks like cleaning and wrangling. Data wrangling is the practice of transforming raw data into a useable format for the task at hand, whether that be data visualization, analysis, or model training and testing.

Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says

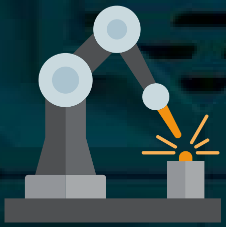
X is for...

XGBOOST

XGboost is an open source machine learning tool that increases model performance and computational speed. Since becoming the tool of choice for several winning teams across a number of machine learning contests, XGBoost has gained popularity among data scientists and enthusiasts alike.

XGBoost, a Top Machine Learning Method on Kaggle, Explained

Y is for...



YIELD

Data science has made its way into nearly every industry, from finance and hospitality to gaming and manufacturing. It’s even helping agricultural businesses predict and mitigate issues that impact an essential metric: crop yield, the agricultural output of a unit of land. Data scientists are helping agriculturalists improve yield through a number of innovations, including machine learning models that are trained to identify common corn diseases. With such a capability, a farmer could take a photo of a crop, receive a diagnosis, and take action before yield is negatively impacted.

Inside Monsanto's Digital Transformation

Z is for...

ZERO DOWNTIME

Data science shouldn’t live and die on a data scientist’s laptop. Ideally, the machine learning models built by data scientists should be deployed in a production environment where they can integrate with other systems and inform decision making. Consequently, data science is inextricably intertwined with IT tasks like cluster management, the act of tracking resources like memory, CPU, or storage. System downtime — an IT challenge that costs companies an estimated $700 billion annually — can get in the way of data science value by taking important applications offline. Achieving zero downtime, in which all systems are up and running continuously, is vital.

Taking an Agile Approach to Data Science

Conclusion

Congratulations! Now you know your data science ABCs.

For more data science content, please visit our resources library.

To receive more ebooks like this one and other content in your inbox, join our mailing list.