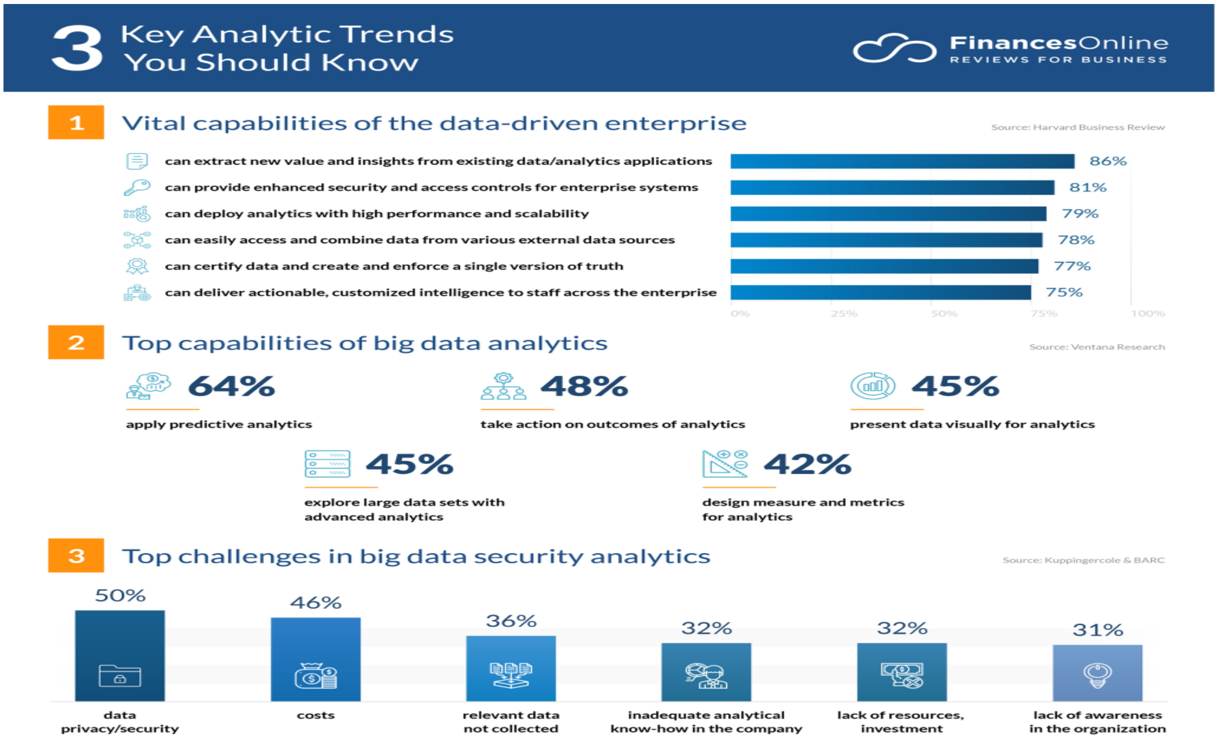
Week #1 – Introduction to Data Science

Introduction



3 Key Analytic Trends You Should Know

FinancesOnline REVIEWS FOR BUSINESS

1 Vital capabilities of the data-driven enterprise

Source: Harvard Business Review

can extract new value and insights from existing data/analytics applications	86%
can provide enhanced security and access controls for enterprise systems	81%
can deploy analytics with high performance and scalability	79%
can easily access and combine data from various external data sources	78%
can certify data and create and enforce a single version of truth	77%
can deliver actionable, customized intelligence to staff across the enterprise	75%

2 Top capabilities of big data analytics

Source: Ventana Research

64% apply predictive analytics

48% take action on outcomes of analytics

45% present data visually for analytics

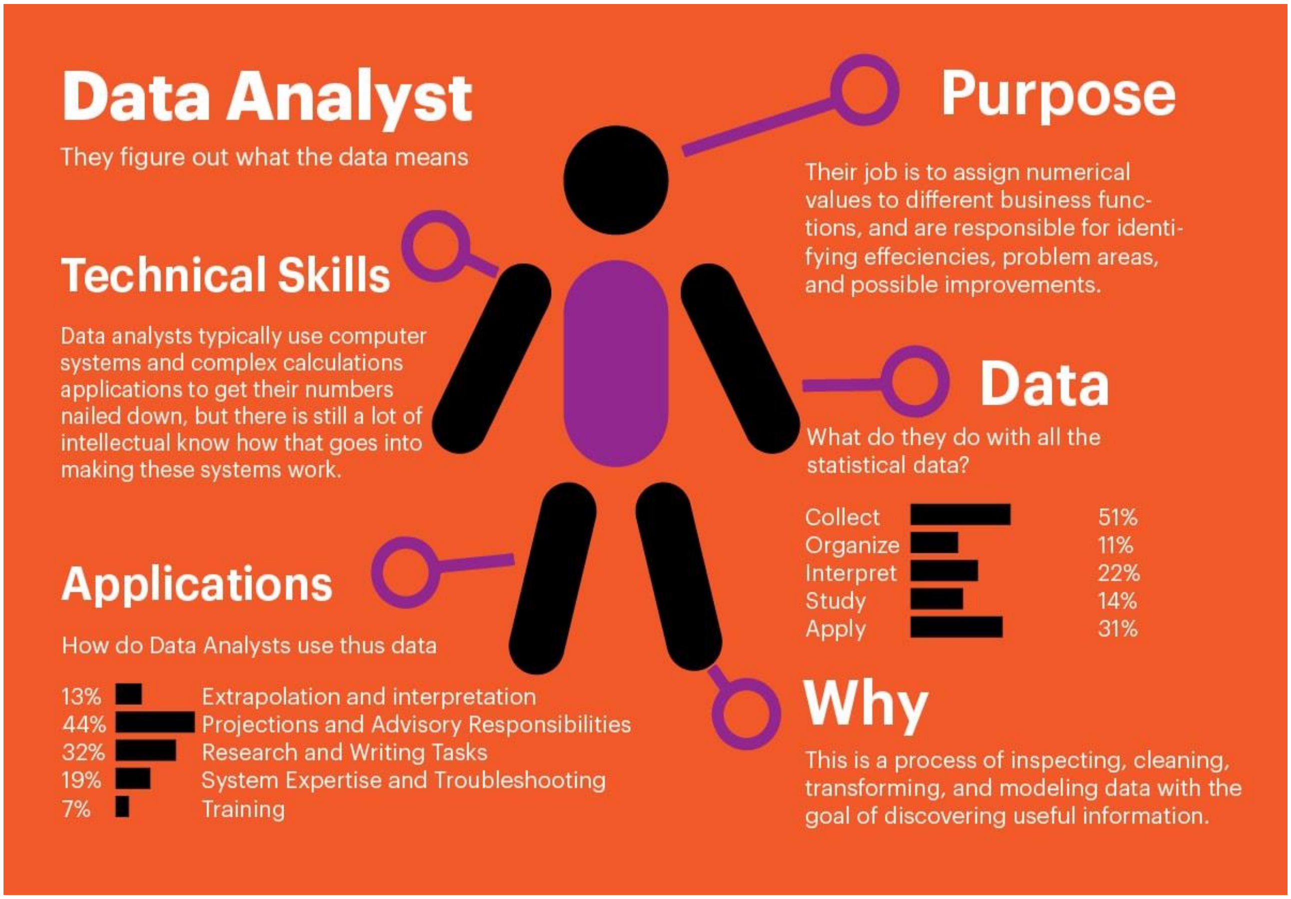
45% explore large data sets with advanced analytics

42% design measure and metrics for analytics

3 Top challenges in big data security analytics

Source: Kuppingercole & BARC

data privacy/security	50%
costs	46%
relevant data not collected	36%
inadequate analytical know-how in the company	32%
lack of resources, investment	32%
lack of awareness in the organization	31%



Data Analyst

They figure out what the data means

Purpose

Their job is to assign numerical values to different business functions, and are responsible for identifying effeciencies, problem areas, and possible improvements.

Technical Skills

Data analysts typically use computer systems and complex calculations applications to get their numbers nailed down, but there is still a lot of intellectual know how that goes into making these systems work.

Data

What do they do with all the statistical data?

Collect	51%
Organize	11%
Interpret	22%
Study	14%
Apply	31%

Applications

How do Data Analysts use thus data

13%	Extrapolation and interpretation
44%	Projections and Advisory Responsibilities
32%	Research and Writing Tasks
19%	System Expertise and Troubleshooting
7%	Training

Why

This is a process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information.

Data Science

Data Science is the analytical discipline which utilizes the principles, processes and techniques of understanding a phenomenon through automated analysis of data. The goal of data science is to enable data-driven decision making.

Data Science Concepts

Concept	Description	Example
Data-driven decision making	Practice of making decisions based on sound analysis and insights from the data	In 2004, Walmart prepared for the impact of Hurricane Frances by mining the trillions of bytes worth of data on shopper history to decide on stocking up on strawberry pop tarts and beer. Walmart used data-driven decision making to infer consumer purchase patterns
Data scientist skills	A wide variety of skills in addition to core math and analytical thinking.	Skills in data extraction, manipulation, statistical modeling, visualization, story-telling, communication and ethical thinking
Data Engineering vs. Data Science	A lot of data processing might support data science but actually falls under data engineering.	Although big data technologies such as Hadoop may be occasionally used for data science tasks, they are more often used for data processing and data engineering which are in support of data science activities.

Big Data

Big Data refers to data that exhibit specific characteristics of volume, velocity and variety. Volume refers to the size of the data i.e., data that is too large to be processed by traditional computers and cannot be analyzed by traditional statistical techniques is considered big data. Velocity refers to how quickly the data is refreshed or arrives for processing and how quickly it may be considered stale. Big data generally refers to data with high velocity. Variety refers to the type of data that is being processed and analyzed. Generally big data is characterized by both structured and unstructured data which includes numbers, pictures, text, logs etc.

Where Big Data is Useful

a. Big data is useful in situations where data is available from a variety of sources and using all the data to predict an outcome would improve the results. For instance, in the case of predicting whether a customer is likely to leave a telecommunication company, data from the customer profile, their call history, customer’s social network, their online social media activity, their billing history etc. might be used to achieve higher accuracy in the prediction. In this case, a big data approach is appropriate.

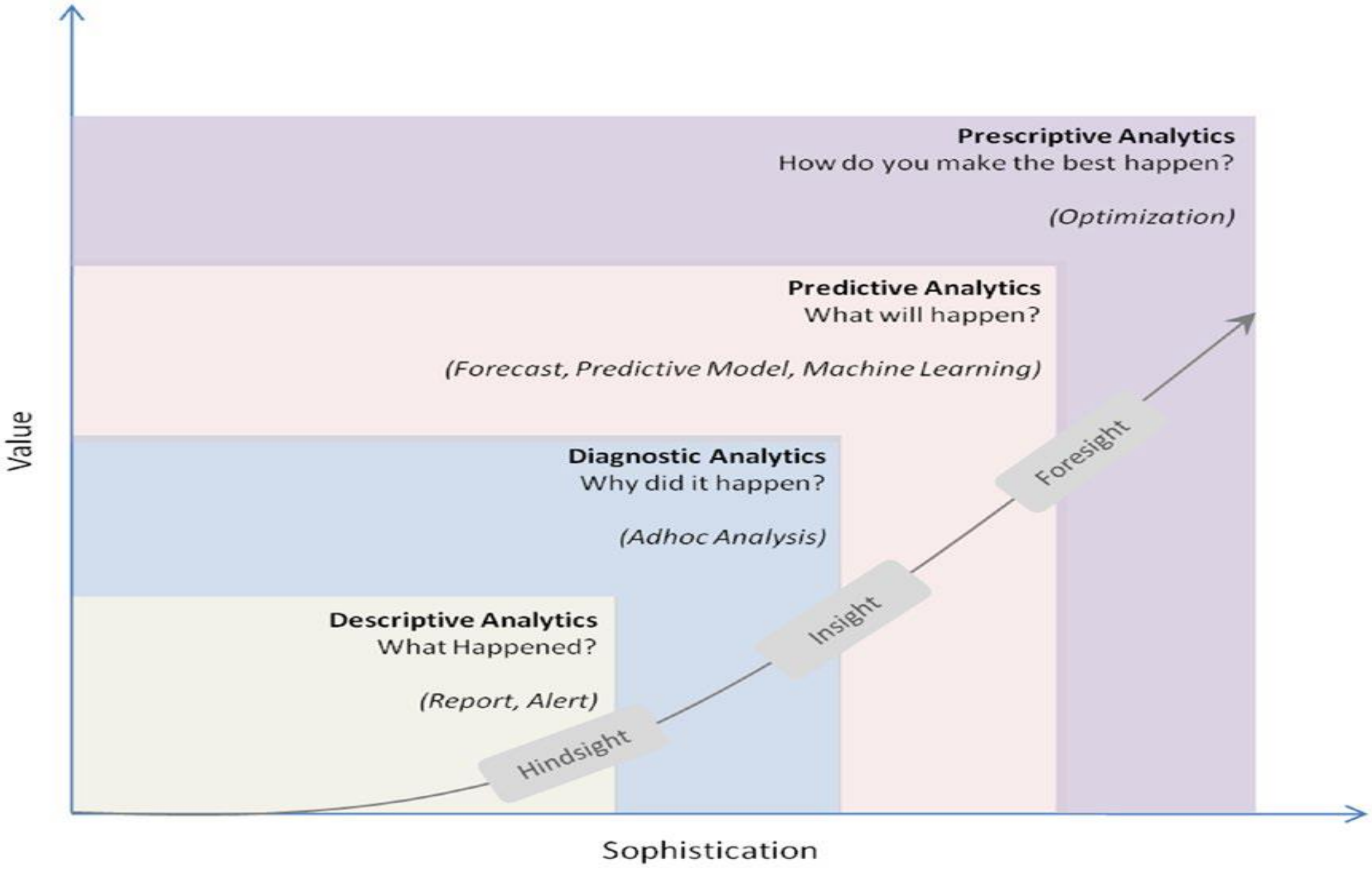
b. Although a big data approach will work well with averaging techniques such as regression or neural networks and some nonparametric techniques, its real advantage is in extremity-based or tail-based modeling techniques. Such tail-based techniques are useful when we are trying to predict rare events (e.g., fraud), or trying to optimize where the use of the whole dataset rather than a sample is fruitful.

c. Big data is also useful in situations where quick iterations are facilitated by the distributed technologies of big data, which improves data analyst productivity.

Data Analytics Types

Data-Analytic thinking refers to the process of thinking systematically about a business problem and using data to improve performance. Data-Analytic thinking is essential for considering the opportunities for business strategy and for making data-driven investment decisions. Some of the fundamental concepts of data analytic thinking include:

- A systematic process can be followed for extracting useful knowledge from data to inform business decision-making
- From a large dataset, information technology tools can be used to identify attributes and characteristics about entities of interest (e.g., entities could be customers, products, etc.)
- By working with a dataset for a considerable amount of time, some patterns or insights can be gleaned but that may not be applicable to new or unseen situations
- Thinking data-analytically involves thinking carefully about the context in which the data was produced and will be used.



Source: Mastering Machine Learning with Python in Six Steps

The Disruption of Data Analytics and Big Data

Data Analytics and big data are disrupting every industry and functional area imaginable. Some of the common use cases in a few industries are provided below:

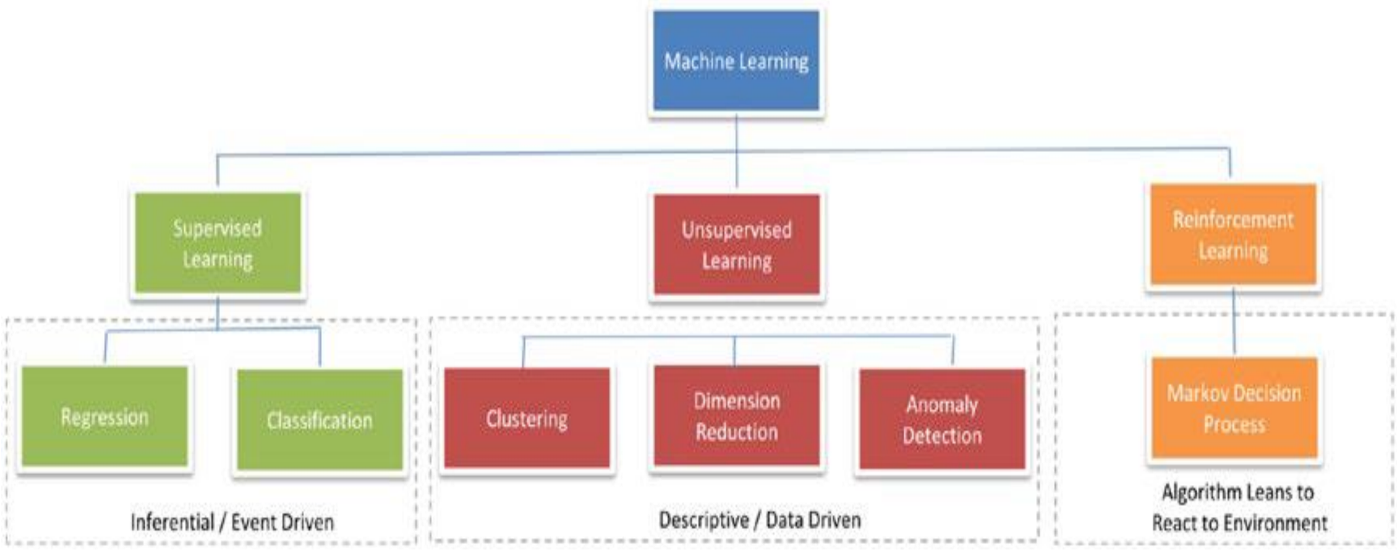
- **Pharmaceuticals and Healthcare** – Big data and data science plays an important part in drug discovery and clinical trials in the pharmaceutical industry. For example, drugs that will be effective on 2 % of the population are now being pursued today because it is possible to identify that segment of the population as well as what the impact would be. In healthcare, big data enables identifying trends in population health and demographics. For instance, data can be used to analyze long term trends such as the trends in aging populations so that policy makers and healthcare providers may cater to that segment of the population better. Additionally, personalized medicine and disease management is possible with data-driven decision making.

- **Credit cards and Finance** – In the credit card industry, predictive models make it possible to recognize fraud based on past fraudulent transactions even if the case of fraud is less prevalent. Predictive models and big data is also being used in stock picking and forecasting of stock market returns so that better trading decisions can be made.

- **Telecommunications** – In telecommunications, predictive models are used for predicting who is likely to leave (churn) and launch retention campaigns to try and save the customer. Similarly, upsell predictive models are used for predicting who is likely to take up an offer in marketing campaigns.

- **Retail** – Big data and data science is used in retail for store display optimization and merchandizing optimization. It is also used for managing and tracking the results of offering coupons and discounts.

Machine Learning



Source: Mastering Machine Learning with Python in Six Steps