Questions to be addressed:

* Is it all machine learning, AI, and coding?

The ML/AI and coding are means to an end which is solving a business/research/policy problem. The business component of the mix is more important.

* Does the data have to be big? and what is "big data" anyway?

Big data does provide a promise for having a larger, more connected and coherent picture of reality. However, just the size doesn't matter without meaning.

Meaningful data [that carries enough information about phenomena/process/reality] is more important.

* What background is most suitable for data science?

Individuals can enter into this field from a variety of different angles.

Data Scientist job requires a robust skill set touching Statistics, Mathematics [Core AI and ML], Computer Science [Applied].

Other background would need to invest time and effort to build up the basic knowledge base in above mentioned.

* What is the need for statisticians? mathematicians? operations research specialists?
* How essential are skills in SAS, SQL, matlab, python, and r ?

Tools to implement, productize and maintain solutions are not less important than theoritcal understanding of how machine learning algorithms

are derived, extended and specialized.

Without a

* What are the new “must have” skills?
  + Depending on the DS process phase:
    - Data Management
    - Modeling
    - Deployment
    - Domain Knowledge: ML works only in well defined, well understood problems
* Is a graduate degree a must?

Yes. But not sufficient. Build an online portfolio of projects

* Are data science bootcamps replacing graduate degrees to get people into the field quickly?

A graduate degree in Mathematics, Statistics, Computing is necessary prerequisite without which topics in data science would not be understandable.

Therefore, bootcamps and certificate programs can brush up and tune the knowledge-base in addition to providing an introductory experience into

the tools but they cannot replace.

* How is data science different than data mining? business intelligence? analytics?

It is at the intersection of multiple disciplines.

Data Analysis: A human process of finding hidden patterns and str

Analytics: Automating the analysis process

* How does/should data science fit into an organization?
  + Sectors:
    - Online Service: Netflix
    - Manufacturing: Anomaly Detection, Pre-emptive Crisis Management
    - Retail: Consumer Analytics, Supply Chain Management, Loyalty Management
    - Agriculture:
    - Medicine:
  + DS Phases:
    - Introduction
    - maintenance
    - scaling
  + IT-Centric DS initiatives:
    - DS is composed as a component of IT Department
    - MLaas (ML as a Service) Approach
      * Economical
      * Commodification of ML by Google, Amazon, Microsoft elevates the need to have private infra and expertise
      * Fast training and easy deployment and maintenance of models
  + Integrated DS Initiative
    - DS and IT are loosely-coupled, interdependent staff
    - DS team prepare data, train models
    - IT incharge of interfaces, and infra to deploy and service the models
    - Advance ML tools (TensorFlow, Theano)
    - Highly specific business problems that need custom-built solutions
  + Dedicated DS department
    - Highest cost: infrastructure & talent (acquisition, engagement, retention)
    - Specialized DS teams for different business units (Security, Customer Analytics, SCM)

Basic Philosophy

* Move compute to data

Hadoop 1, 2, 3

* Alpha, Beta, GA (Generally Available)

10 Years of Hadoop

Hadoop 3: <https://www.slideshare.net/Hadoop_Summit/apache-hadoop-yarn-present-and-future-77278025>

Typical Hadoop Cluster

<https://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.1.3/bk_cluster-planning-guide/content/typical-hadoop-cluster-hardware.html>

YARN Latest

YARN is often called the operating system of Hadoop because it is responsible for managing and monitoring workloads, maintaining a multi-tenant environment, implementing security controls, and managing high availability features of Hadoop.

<https://medium.com/@markobonaci/the-history-of-hadoop-68984a11704>

<https://www.slideshare.net/Hadoop_Summit/apache-hadoop-yarn-present-and-future-77278025>

Replication:

Replication-factor: number of times each block ok data is replicated

A trade-off: Reliability (Fault-Tolerance) Read/write bandwidth

Default: First replica on same node: Second on different rack, Third on different node of same rack

HDFS Cluster

Yahoo Data center

<http://blog.damac.com/true-strength-behind-server-rack/>

Rack-aware Replication

<http://blog.dbtechnosolutions.com/2015/07/rack-awareness-and-configuration-in.html>

<http://www.bigdataplanet.info/2013/10/Hadoop-Tutorial-Part-3-Replication-and-Read-Operations-in-HDFS.html>