Data Science : A Perspective

Casey Stella

February 25, 2014

Table of Contents

Hi, I'm Casey

- Education
  - B.S. in Computer Science from University of Louisiana at Monroe
  - M.S. in Mathematics from Texas A&M University with an emphasis in Theoretical Computer Science

Hi, I'm Casey

- Education
  - B.S. in Computer Science from University of Louisiana at Monroe
  - M.S. in Mathematics from Texas A&M University with an emphasis in Theoretical Computer Science
- I tend to work with "Big Data"
  - I am a consulting data scientist and big data architect at Hortonworks, an open source software company
  - I was an architect/software engineer on the high performance indexing team at Explorys, a medical informatics startup based in the Cleveland Clinic
  - I was a Research Geophysicist in the Oil Industry doing signal processing
  - I built a VOIP network oriented toward WoW gamers

Data Science : The Skin

"A data scientist is a peculiar blend of developer and statistician that is capable of turning data into awesome."

— Josh Wills, Director of Data Science at Cloudera

Data Science : Pithy Definitions

"A data scientist is a statistician who lives in San Francisco."

— Snarky Internet Troll

Data Science : Pithy Definitions

"Data Science is an apple."

— Me

Data Science : The Meat

- Gathering and repurposing data that was otherwise lost or forgotten.

Data Science : The Meat

- Gathering and repurposing data that was otherwise lost or forgotten.
- Processing said data using
  - Traditional statistics

Data Science : The Meat

- Gathering and repurposing data that was otherwise lost or forgotten.
- Processing said data using
  - Traditional statistics
  - Machine learning

Data Science : The Meat

- Gathering and repurposing data that was otherwise lost or forgotten.
- Processing said data using
  - Traditional statistics
  - Machine learning
  - Natural Language Processing

Data Science : The Meat

- Gathering and repurposing data that was otherwise lost or forgotten.
- Processing said data using
  - Traditional statistics
  - Machine learning
  - Natural Language Processing
- Visualizing said data in a way which allows for an insight to be derived.

The Core : Data Gathering

- Traditionally data analytics has been focused on limited, mostly numeric, data.

The Core : Data Gathering

- Traditionally data analytics has been focused on limited, mostly numeric, data.
- Traditional statistical analysis focused on predicting some numerical quantity from curated, well planned data.

The Core : Data Gathering

- Traditionally data analytics has been focused on limited, mostly numeric, data.
- Traditional statistical analysis focused on predicting some numerical quantity from curated, well planned data.
- Increased computing power and decreased cost left many data that could be gathered without a home in existing statistical models and therefore dropped.

The Core : Data Gathering

- Traditionally data analytics has been focused on limited, mostly numeric, data.
- Traditional statistical analysis focused on predicting some numerical quantity from curated, well planned data.
- Increased computing power and decreased cost left many data that could be gathered without a home in existing statistical models and therefore dropped.
- Data science is the expansion of traditional analytics to include data that is being dropped or is unstructured.

The Core : Unstructured Data

- Some of the most interesting data gathered is not simple numerical or categorical data.
  - For example, doctor's notes, radiologist reports, Facebook postings.

The Core : Unstructured Data

- Some of the most interesting data gathered is not simple numerical or categorical data.
  - For example, doctor's notes, radiologist reports, Facebook postings.
- Natural language processing is a technique to make a computer begin to understand the written word.

The Core : Unstructured Data

- Some of the most interesting data gathered is not simple numerical or categorical data.
  - For example, doctor's notes, radiologist reports, Facebook postings.
- Natural language processing is a technique to make a computer begin to understand the written word.
- Sometimes the outputs are structured data and sometimes the outputs are insights themselves.

The Core : Machine Learning

- A technique in computer science and statistics whereby a computer algorithm is presented data and learns patterns about the data.

The Core : Machine Learning

- A technique in computer science and statistics whereby a computer algorithm is presented data and learns patterns about the data.
  - e.g. One could use machine learning to predict whether a given tweet originated from a person based on their twitter history.

The Core : Machine Learning

- A technique in computer science and statistics whereby a computer algorithm is presented data and learns patterns about the data.
  - e.g. One could use machine learning to predict whether a given tweet originated from a person based on their twitter history.
- Traditional statistical models are generally well defined by a human and run over the data.

The Core : Machine Learning

- A technique in computer science and statistics whereby a computer algorithm is presented data and learns patterns about the data.
  - e.g. One could use machine learning to predict whether a given tweet originated from a person based on their twitter history.
- Traditional statistical models are generally well defined by a human and run over the data.
- Machine learning models have a human defining the input, but the machine develops an internal model based on examples of the data.

The Core : Machine Learning

- A technique in computer science and statistics whereby a computer algorithm is presented data and learns patterns about the data.
  - e.g. One could use machine learning to predict whether a given tweet originated from a person based on their twitter history.
- Traditional statistical models are generally well defined by a human and run over the data.
- Machine learning models have a human defining the input, but the machine develops an internal model based on examples of the data.
- Sometimes these approaches are at odds, but both techniques have merit and are used in the field.

The Seeds: The Skillset

- Statistics/Mathematics
- Computer Science
- Domain expertise
  - Data science is applied to a domain, so domain expertise is a necessity.

Who is eating the apple?

- Computational power and storage has made keeping and analyzing massive amounts of data feasible.

Who is eating the apple?

- Computational power and storage has made keeping and analyzing massive amounts of data feasible.
- More and more industries are interested in leveraging this data to make decisions.
  - Retail
  - Healthcare
  - Finance
  - Oil and Gas

An apple looking for a tree

- Data science skillsets are necessarily cross-disciplinary.

An apple looking for a tree

- Data science skillsets are necessarily cross-disciplinary.
- Universities are just starting to widen their programs to create cross-domain training for data science.

An apple looking for a tree

- Data science skillsets are necessarily cross-disciplinary.
- Universities are just starting to widen their programs to create cross-domain training for data science.
- Furthermore, data science as a practice is hard to commoditize/productize.

An apple looking for a tree

- Data science skillsets are necessarily cross-disciplinary.
- Universities are just starting to widen their programs to create cross-domain training for data science.
- Furthermore, data science as a practice is hard to commoditize/productize.
- This adds up to high demand and low supply.

Challenges

- Not all analytics are possible
  - This can be jarring to those consuming data science.

Challenges

- Not all analytics are possible
  - This can be jarring to those consuming data science.
- Extremely hyped discipline.

Challenges

- Not all analytics are possible
  - This can be jarring to those consuming data science.
- Extremely hyped discipline.
- Realistic methodologies to predict expected time to completion for data science tasks do not exist or are flawed deeply.

Challenges

- Not all analytics are possible
  - This can be jarring to those consuming data science.
- Extremely hyped discipline.
- Realistic methodologies to predict expected time to completion for data science tasks do not exist or are flawed deeply.
- Tooling is either extremely expensive (restricting range of analytics) or free and harder to use.

Challenges

- Not all analytics are possible
  - This can be jarring to those consuming data science.
- Extremely hyped discipline.
- Realistic methodologies to predict expected time to completion for data science tasks do not exist or are flawed deeply.
- Tooling is either extremely expensive (restricting range of analytics) or free and harder to use.
- Explaining insights gained can be extremely challenging.

Conclusion

- Thanks for your attention
- Questions?