Moore's Law and Data Science

In the last few years, there has been a burst in the amount of data that's available. Whether its web server logs, tweet streams, online transaction records, data from sensors, government data, or some other source, the problem isn't getting data, it's figuring out what to do with it.

The question facing every company today even the startups and non-profit, is how to use data effectively—not just their own data, but all the data that's available and relevant. Using data effectively requires something different from traditional statistics.

What differentiates data science from statistics is that data science is a holistic approach. We're increasingly finding data in the wild, and data scientists are involved with gathering data, massaging it into a manageable form, finding a story/trend from it, and presenting the story to others.

Much of the data we currently work with is the direct consequence of Web 2.0, and of Moore's Law as applied to data. The web has people spending more time online, and leaving a trail of data wherever they go. Mobile applications leave an even richer data trail, since many of them are annotated with geo-location, or involve video or audio, all of which can be mined. Point-of-sale devices and frequent-shopper's cards make it possible to capture all of your retail transactions, not just the ones you make online. All of this data would be useless if we couldn't store it, and that's where Moore's Law comes in.

The importance of Moore's law as applied to data isn't a theory anymore. Data expands to fill the space you have to store it. The more storage is available, the more data you will find to put into it. The data exhaust you leave behind whenever you surf the web, friend someone on Facebook, or make a purchase in your local supermarket, is all carefully collected and analyzed. Increased storage capacity demands increased sophistication in the analysis and use of that data. That's the foundation of data science.

Data scientists combine entrepreneurship with patience, the willingness to build data products incrementally, the ability to explore, and the ability to iterate over a solution. They are inherently inter-disciplinary. They can tackle all aspects of a problem, from initial data collection and data conditioning to drawing conclusions.

The future belongs to the companies who figure out how to collect and use data successfully. Google, Amazon, Facebook, and LinkedIn have all tapped into this and all other companies are trying to catch up with this at a war footing.