Understanding the Power of Data

Webinar Reference Document

Published September 23, 2020

Prepared by Mary Willcock

The Volume of Data

44 zettabytes or 44 trillion GB.

- This is not just your email address that marketers always try to get from you, but pictures, videos, voice commands to alexa, etc
- 1.7MB of data is created every second by every person during 2020.
- In the last two years alone, the astonishing 90% of the world's data has been created.
- 2.5 quintillion bytes of data are produced by humans every day.
- 463 exabytes of data will be generated each day by humans as of 2025.
- 95 million photos and videos are shared every day on Instagram.
- By the end of 2020, 44 zettabytes will make up the entire digital universe.
- Every day, 306.4 billion emails are sent, and 5 million Tweets are made.

Retailers who choose to leverage the full potential of big data analytics can optimize their operating margins by approximately 60%.

As of this moment, only 0.5% of all accessible data is analyzed and used. Imagine the potential here.

The Value of Data

Rising Value of Data

Value – more than oil – Back in 2017, *The Economist* published a story titled, "The world's most valuable resource is no longer oil, but data."
In 2020, up to 90% of the world's largest enterprises are expected to generate income from data-as-a-service (DaaS).
Drastic changes to all industries - Google buying Walmart and amazon buying whole foods for the data

Wisdom Pyramid

Information architecture data to wisdom pyramid

1. Data – raw – 'red'
2. Information – meaning – 'The south facing traffic light at the corner of North and Main st has turned red'
3. Knowledge – context – 'The traffic light I am driving toward has turned red'
4. Wisdom – applied – 'I'd better stop the car'

Stitch fix example

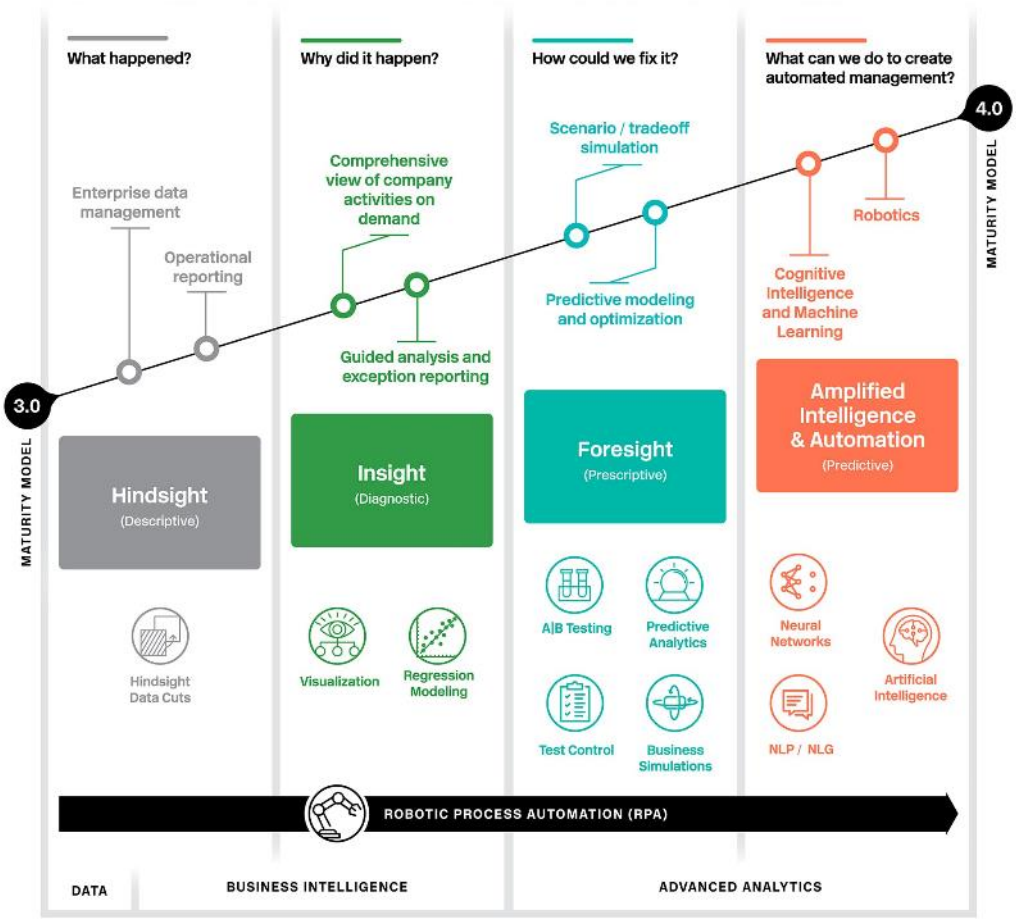
HBR Article

Katrina Lake at Harvard for MBA told the model would fail, don't get into retail.

Stitch Fix sold $730 million worth of clothing in 2016 and $977 million worth in 2017. One hundred percent of our revenue results directly from our recommendations, which are the core of our business. We have more than 2 million active clients in the United States, and we carry more than 700 brands. We're not upselling you belts that match that blouse you just added to your cart, or touting a certain brand because you've bought it before, or using browsing patterns to intuit that you might be shopping for a little black dress—all activities that have low conversion rates. Instead we make unique and personal selections by combining data and machine learning with expert human judgment.

Analytics Maturity Chart

https://www.bakertilly.com/specialties/data-and-analytics



Ethics in Advanced Analytics and AI

Newspaper Trained Model

The researches built a model to predict how an analogy should end using gender specific words, such as “man is to king as woman is to _____” where the model would predict the female equivalent word “queen”. After training the model with Google News articles the researchers then used non-gendered professions to see the female equivalent. For “man is to computer programmer as woman is to _____” the model predicted “homemaker”. Other extreme “female” professions in the model included: “nurse”, “receptionist”, “librarian” and “hairdresser” while extreme male professions included: “maestro”, “skipper”, “philosopher” and “captain”. These models reflect the societal biases found in commonly used data sources such as Google News that are used in many business, government and research contexts without awareness or correction for gender biases.

Facial Recognition Error Rates

Joy Bualomwini's facial recognition study

Error Rates in Commercial Gender Classification Products

	Microsoft	Face++	IBM
Dark Skinned Female	20.8%	34.5%	34.7%
Light Skinned Female	1.7%	6.0%	7.1%
Dark Skinned Male	6.0%	0.7%	12.0%
Light Skinned Male	0.0%	0.8%	0.3%

In June IBM announced they will no longer offer, develop, or research facial recognition technology and will reevaluate selling the technology to law enforcement.

Criminal Sentencing

Correctional Offender Management Profiling for Alternative Sanctions, or COMPAS. Pro-Republica article first investigating the bias.

Northpointe's risk of recidivism score had an accuracy rate of 68 percent in a sample of 2,328 people. **COMPAS** has been used by the U.S. **states** of New York, Wisconsin, California, Florida's Broward County.

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Cathy O'Neil's book Weapons of Math Destruction.

Sweden's Snow Removal

Caroline Criado Perez's book Invisible Women.

Karlskoga, Sweden assessed their snow removal policies, and found bias. The major arteries get the first snow removal while the walkers–more female than male–slip, slide and get injured. One Swedish city reported that women experienced 69% of the injuries that occurred during winter with two-thirds of those individuals slipping on ice or snow.

Core Values

Simplicity

The most simple solution should always be used. When we choose a flashier solution that does not provide a better solution we create unnecessarily complex systems.

Transparency

Many AA & AI solutions aren't simple. However, they can and should be readily explained and made accessible. Real solutions require communication and collaboration, which requires transparency on the part of the technical team.

Hypothesis Driven

Keep the science in data science – what do you think will happen, why do you think this is happening? Has it been proven or not?

Use Cases

Predictive Maintenance

Unplanned Downtime

Unplanned downtime costs businesses an average of $2 million over the last three years. In 2014 the average downtime cost per hour was $164,000. By 2016, that statistic had exploded by 59% to $260,000 per hour.

35% of orgs reported in 2017 they are unaware of the impact downtime has for them.

According to the US National Center for Manufacturing Sciences (NCMS), 39% of all cyber attacks in 2016 were against the manufacturing industry. Since January of 2017 attacks increased by 24% catapulting the manufacturing industry ahead of healthcare as the most sought-after victim by today's sophisticated digital criminals.

Identifying causes and timing of future outages provides the opportunity to reduce unplanned downtime.

Pricing Optimization

Looking to get the most sales for the highest price per item. Finding the sweet spot here is tough, but can be improved with a data driven focus.

Retail example

- Old way: We put shorts on sale every memorial day weekend
- New way: We advertise shorts memorial day weekend and put them on sale the week after

What is the impact of the market?

- Dell predicted the dot com bubble burst of 2000-2001

- They adjusted by dropping their prices as low as they could and maintained on a 3% quarter loss compared with competitors at 35% and higher

Natural Language Processing

Conversation Evaluation utilizing Natural Language Processing

- Before: We knew customers talked to expert specialists for support.
- After: We know how often different customers talk to experts, we know what topics the customer and experts talk about, we know when customers engage in key topics, we know what experts are most responsive and resolve problems efficiently.