# Big Data Basics

Justin Post

# Where are we?

- Foundation in Python through `JupyterLab`

Now:

- What is Big Data?
- What are common things we'll run into with Big Data?
- What issues are we going to try and tackle?
- What software can be useful?

# What is Big Data?

Useful definition:

- Big data = data that you can't handle 'normally'

# What is Big Data?

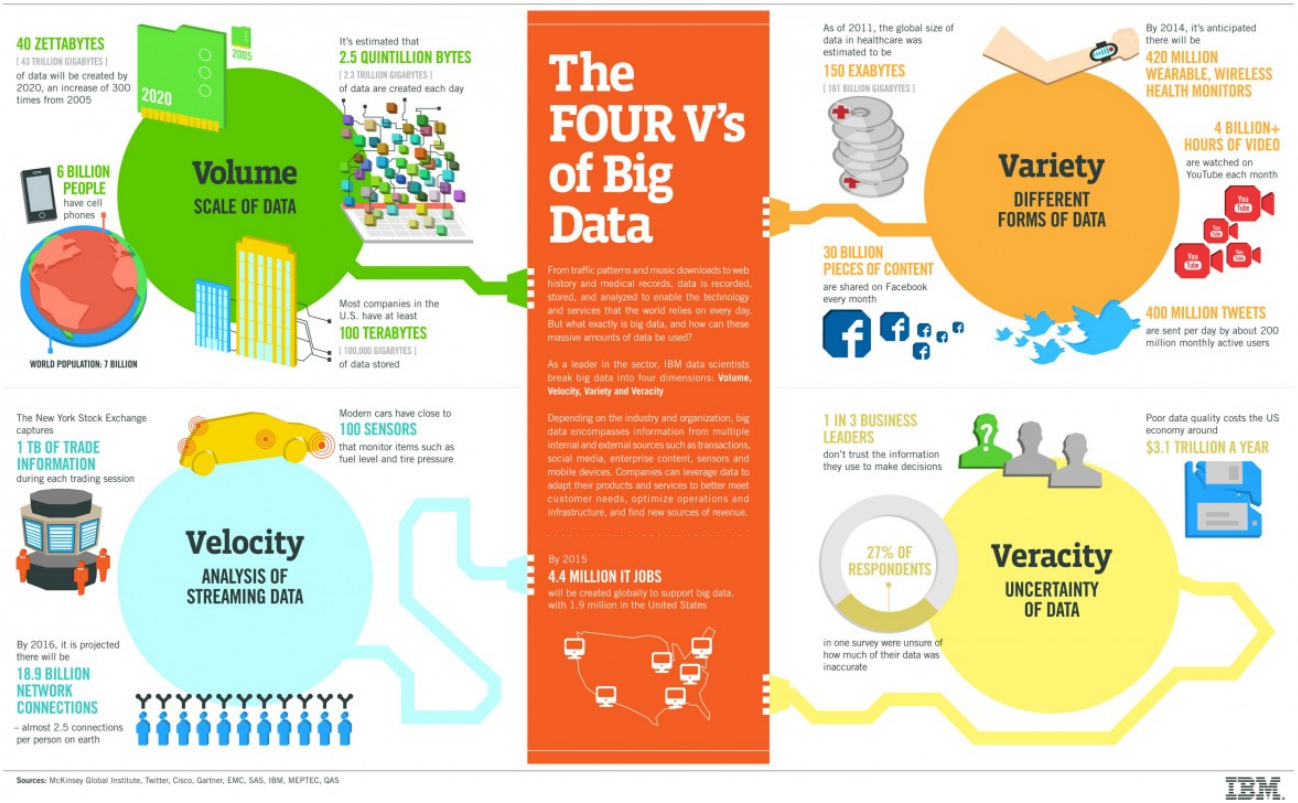
Useful definition:

- Big data = data that you can't handle 'normally'

- Big Data usually requires learning of new tools
- Want to feel overwhelmed?

# What are Common Attributes of Big Data?

- Fifth **V** = Value



# What are We Going to Consider?

- Basics of Big Data storage
- Basics of Big Data querying
- Basics of handling streaming data
- Summarizing and modeling Big Data

# Data and Model Pipelines

In the end, we should have a reasonable idea of how data comes in, gets transformed/combined/etc., can be used to build models or predict an outcome!

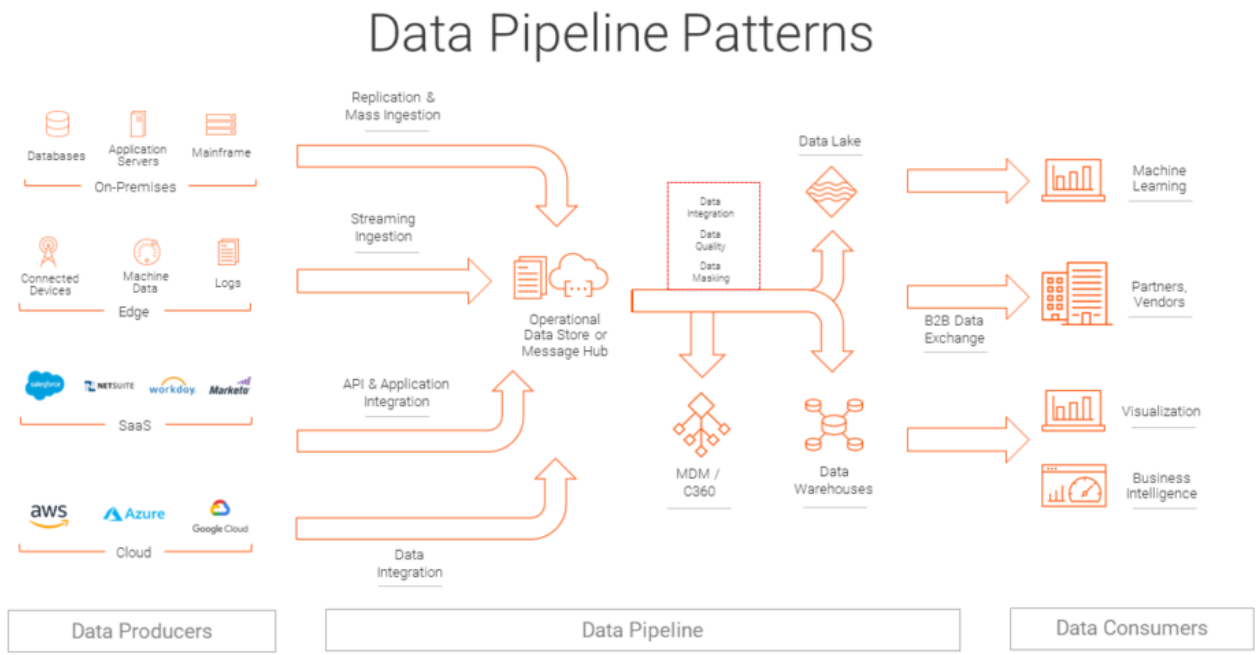



Image from informatica.com

# Recap

- Big Data requires learning new tools
- Big Data doesn't mean we have **everything**
- Statistics and machine learning can still be applied to Big Data but will require some modifications