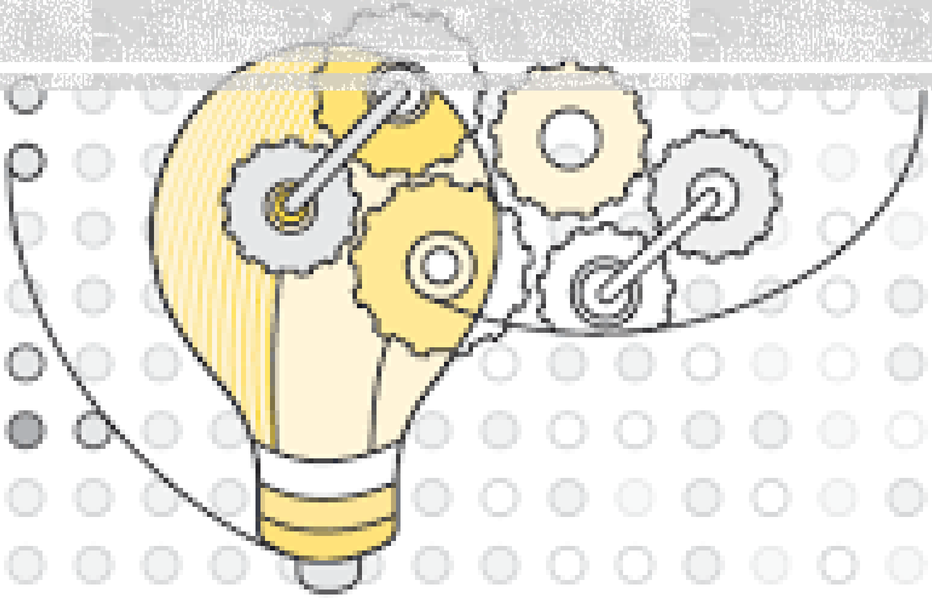
MAJOR THEMES

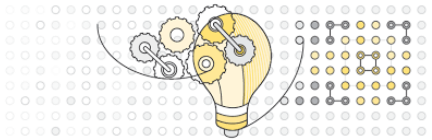


COURSE GOALS

1. Curious to discover more
2. Confident of doing it yourself
3. Contemplative of the theory
4. Cautious of the dangers

KEY LESSONS

1. Machine learning is
   the design of algorithms
   that improve their performance
   at some task with experience.
2. The goal of machine learning
   is generalization.



RELATED FIELDS

Probability. The theory of randomness.

Statistics. Collection, analysis, interpretation, presentation, organization of data.

Data Science. The study of everything related to data, including data sources, statistics, computation, infrastructure, human-factors.

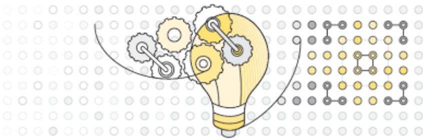
Science of Intelligence. Data science, machine learning, machine intelligence, computer science, neuroscience, cognitive science, ethics.

Artificial Intelligence. Computers making decisions autonomously.

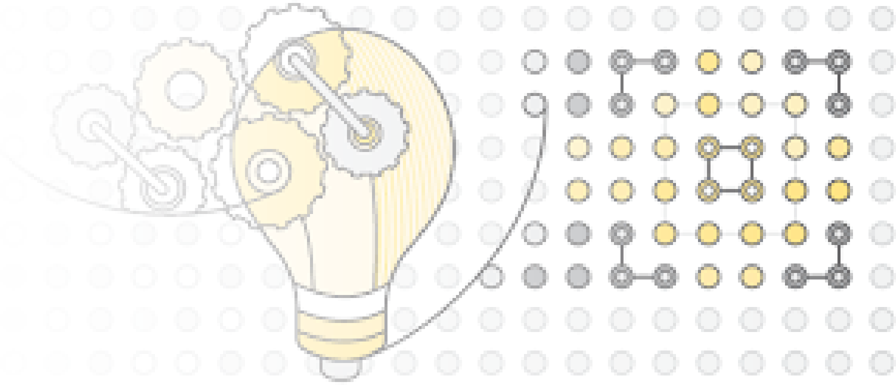
Machine Learning. Computers making sense of data with human help.

Data Mining. Humans making sense of data with machine help.

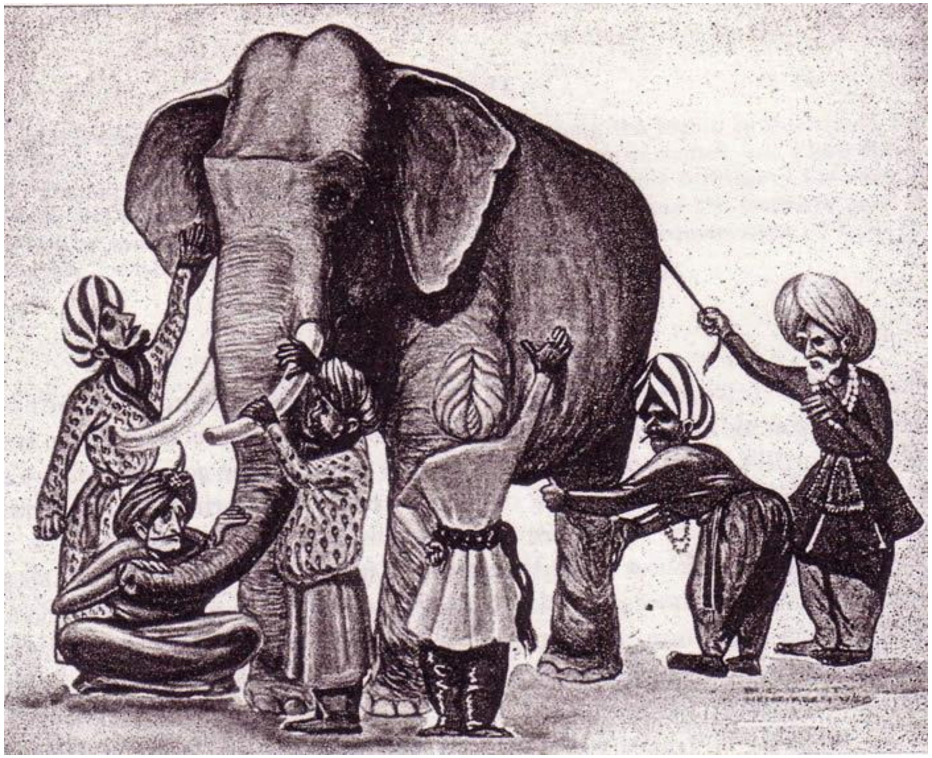
Data Analytics. Humans making decisions, with or without machines.



DATA

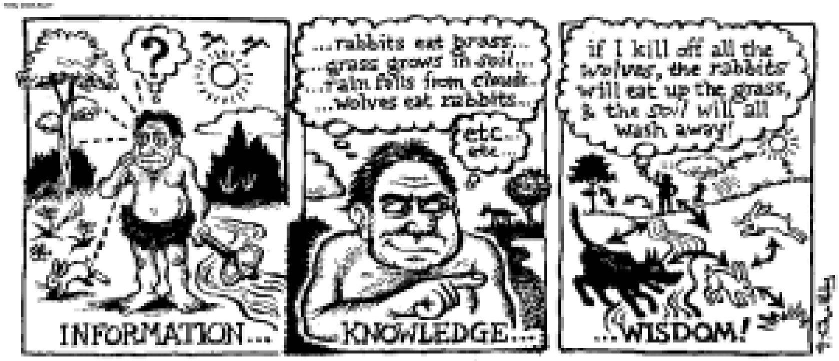


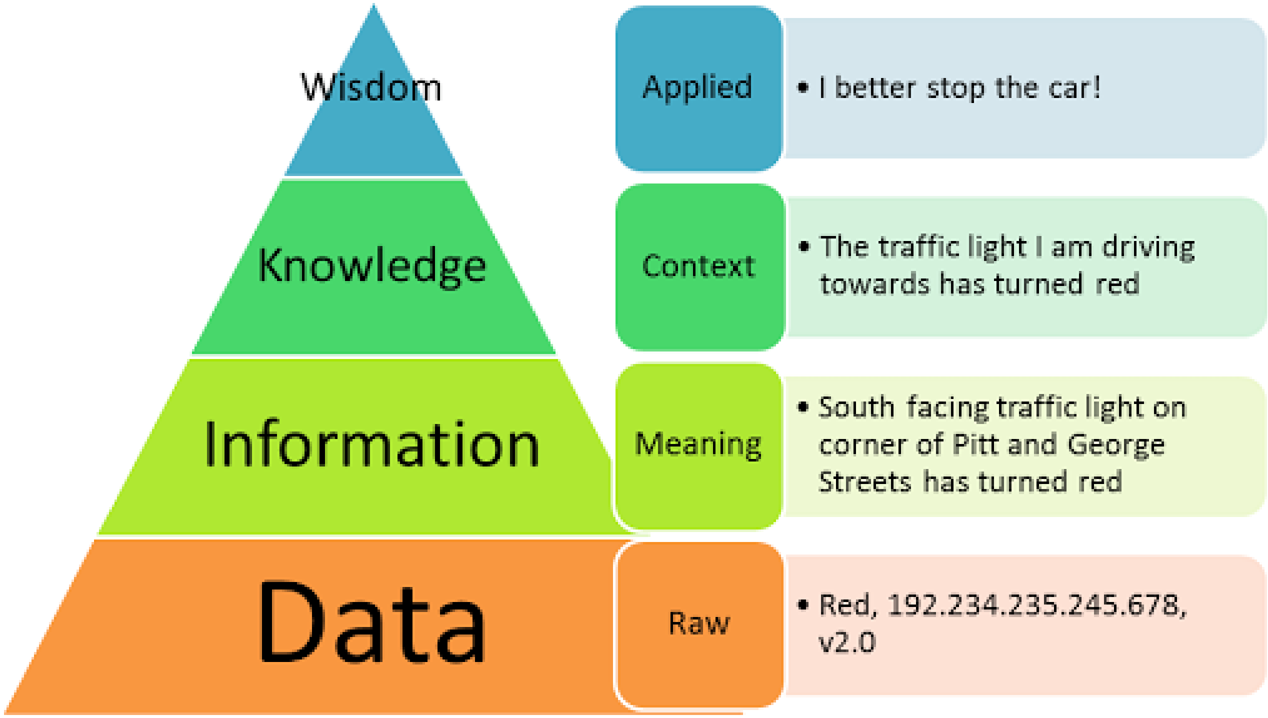
DATA SOURCES



- Sensor data
- Questionaires
- Domain experts
- Mechanical Turks
- News
- Wikipedia
- Linked Open Data
- World Wide Web (e.g. unsupervised learning)
- Social media (e.g. Twitter, Facebook, …)
- Historical data (beyond current timeframe)
- Geospatial data (beyond current region)

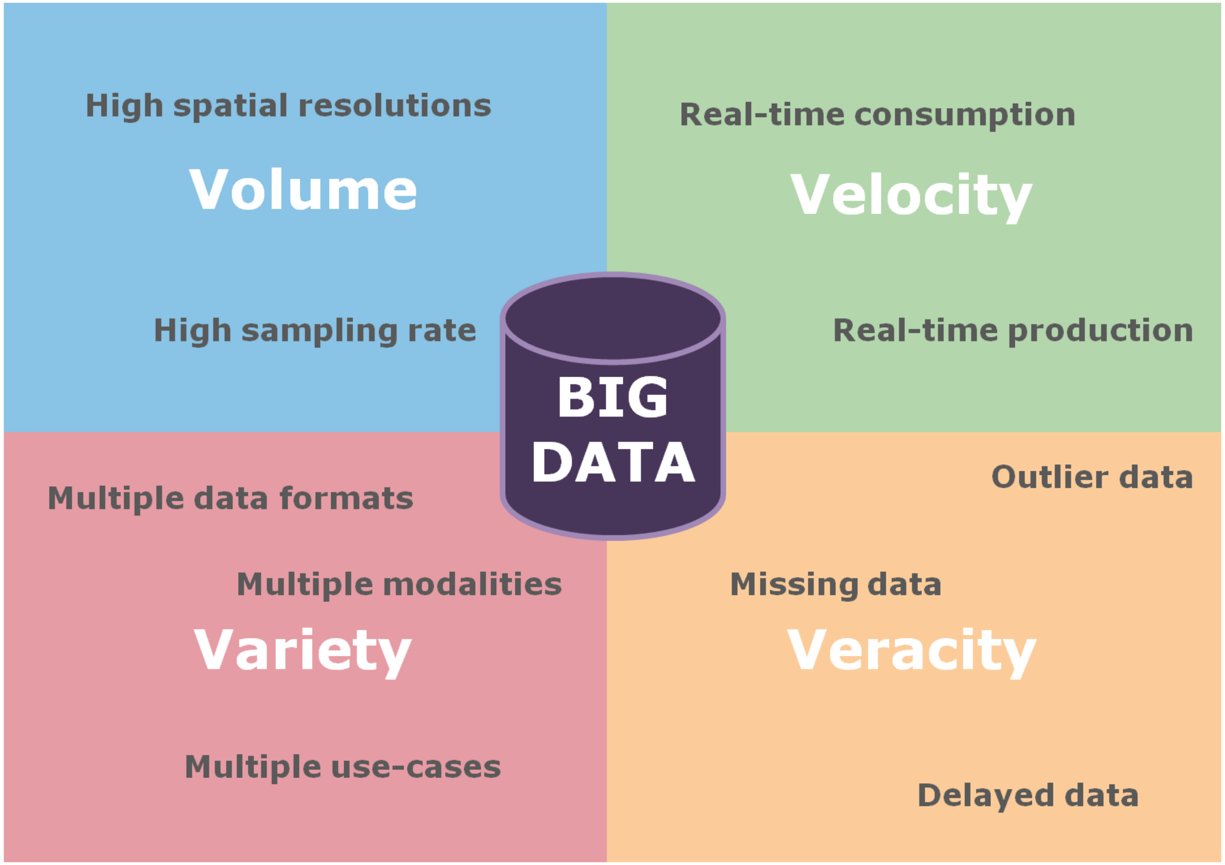
DIKW HIERARCHY


?
INFORMATION...
...rabbits eat grass...
...grass grows in soil...
...rain falls from clouds...
...wolves eat rabbits...
etc... etc...
KNOWLEDGE...
if I kill off all the wolves, the rabbits will eat up the grass, & the soil will all wash away!
...WISDOM!



Wisdom	Applied	• I better stop the car!
Knowledge	Context	• The traffic light I am driving towards has turned red
Information	Meaning	• South facing traffic light on corner of Pitt and George Streets has turned red
Data	Raw	• Red, 192.234.235.245.678, v2.0

BIG DATA



BIG DATA

Volume

High spatial resolutions

High sampling rate

Velocity

Real-time consumption

Real-time production

Variety

Multiple data formats

Multiple modalities

Multiple use-cases

Veracity

Outlier data

Missing data

Delayed data

KEY STEPS IN DATA PROCESS

Prepare

- ask questions
- collect data
- organize data
- cleanse data

Analyze

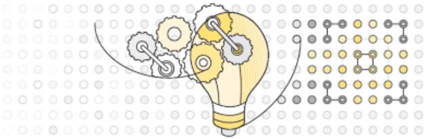
- find patterns
- find relationships
- filter data
- summarize data
- calculate formulas
- create charts

Apply

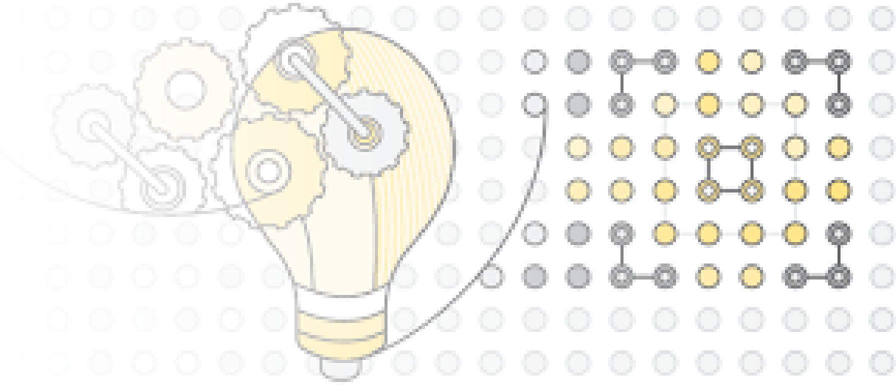
- make decisions
- share results
- visualize data

80%

20%

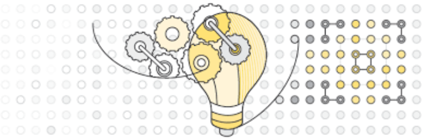


THEORY



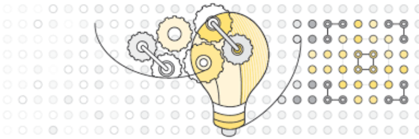
PREPARATION

1. Outliers
2. Missing Values (Collaborative Filtering)
   - Mean, Median
   - K-Nearest Neighbors
   - Matrix Factorization
3. Feature Engineering
4. Dimensionality Reduction (Unsupervised Learning)
   - K-Means Clustering
   - EM Clustering
   - Matrix Factorization
   - Deep Learning



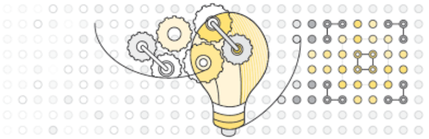
LEARNING

1. Data
   - Test Set, Validation Set, Training Set
   - Features, Responses, Labels
2. Model
   - Parameters, Hypotheses, Predictors
3. Training
   - Loss Functions, Metrics (Distance), Kernels (Similarity)
   - Point Loss, Training Loss
   - Point Gradient, Training Gradient
4. Prediction
   - Test Error, Validation Error, Training Error
5. Generalization



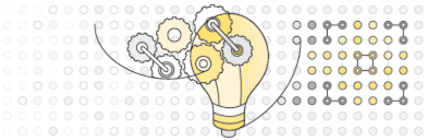
STATISTICS

1. Generative Models
2. Training
   - Maximum Likelihood
   - Expectation-Maximization (Hidden Variables)
3. Prediction
   - Log Likelihood Ratio (Classification)
   - Conditional Expectation (Regression)
4. Smoothing
   - Priors, Posteriors



METHODS

1. Supervised Learning
   - Regression – Linear Regression, Ridge Regression
   - Classification – Perceptron, Hinge Loss
2. Unsupervised Learning
   - Clustering – K-Means Clustering
   - Collaborative Filtering – K-Nearest Neighbors, Matrix Factorization
3. Reinforcement Learning
4. Transfer Learning
5. Breakthroughs
   - Support Vector Machines – Max Margin, Duality, Kernels
   - Deep Learning – Multilayer Networks, Backpropagation, Autoencoders

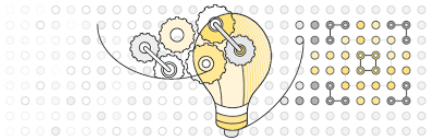


OPTIMIZATION

1. Local, Global Minima
2. Exact Solution
3. Gradient Descent
   - Learning Rate
   - Sub-gradients
   - Second Order Methods (e.g. BFGS)
4. Stochastic Gradient Descent
   - Momentum
5. Coordinate Descent
6. Convex Optimization
7. Constrained Optimization
8. Duality

GENERALIZATION

1. Model Selection
   - Overfitting, Underfitting
2. Regularization
   - Weight Decay, Sparsity
3. Hyperparameters
4. Simple Validation
5. Cross Validation
6. Bayesian Information Criterion



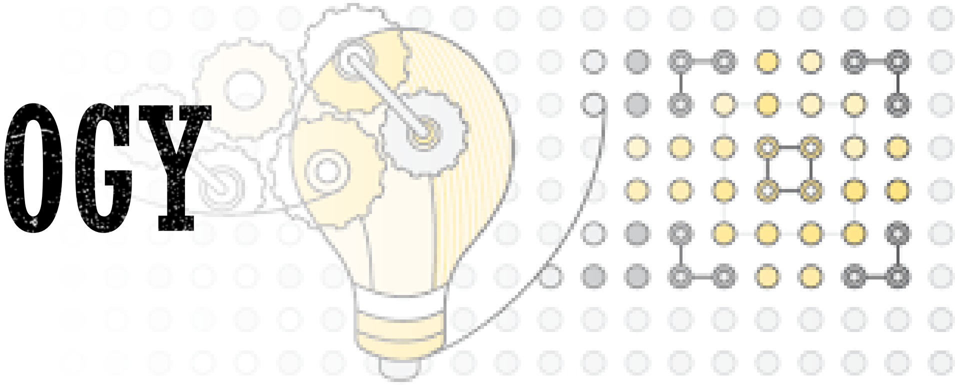
COMPUTATION

1. Automatic Differentiation
2. Parallelization
3. Kernel Trick

PRIORITIES

1. Structure
2. Data
3. Computation
4. Algorithm

METHODOLOGY



BUILDING A PORTFOLIO

1. Unified sensing platform (Jurong Lake District)

   – sharing of sensor network infrastruture among agencies
2. Environmental monitoring (Noise mapping)

   – to enforce noise laws and measure extent of noise pollution
3. Urban planning (Punggol microclimate)

   – combining modelling with sensing for better design of towns
4. Structural health monitoring (Port cranes)

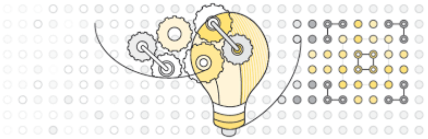
   – sensors and analytics to detect faults early
5. Transport infrastructure (Sensors on wheels)

   – cars with sensors to detect potholes, broken lights, etc.
6. Public cleanliness (Smart bins)

   – detect fullness of bins to minimize manpower for cleaning
7. High-tech agriculture (Vertical farming)

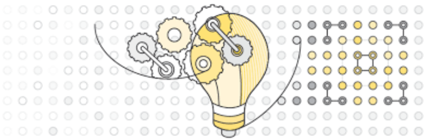
   – sensors in green houses to improve crop yields per unit area
8. Healthcare and aging (Behavioural sensing)

   – home sensors to monitor elderly for falls, depression, etc.



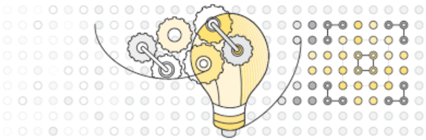
ADVICE

1. There is no ‘right’ method – only what works or doesn’t work. Use common sense, not textbook answers or black boxes. The machine is only as smart as the human who designed it.
2. Ask the right questions, before looking for answers. After finding some answers, sharpen your questions.
3. Visualize, visualize, visualize! Learning new visualization techniques should be a priority.
4. First explore, then focus. When you first brainstorm, do not throw out anything. After seeing everything, prioritize.



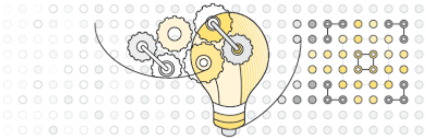
ADVICE

5. Look for the unexpected, not what you already know. Outliers are outliers for a reason. Data is missing for a reason.
6. Look beyond correlations. Learn ‘explanations’ for the correlations, e.g. user attributes in Netflix challenge.
7. Look for indirect ways of measuring a feature of interest, e.g. instead of asking if a user is interested in a product, meaure how long and how often he visits the product page.
8. Apply what you learn from the data. Ask ‘what can I do differently in my operations now that I know this?’

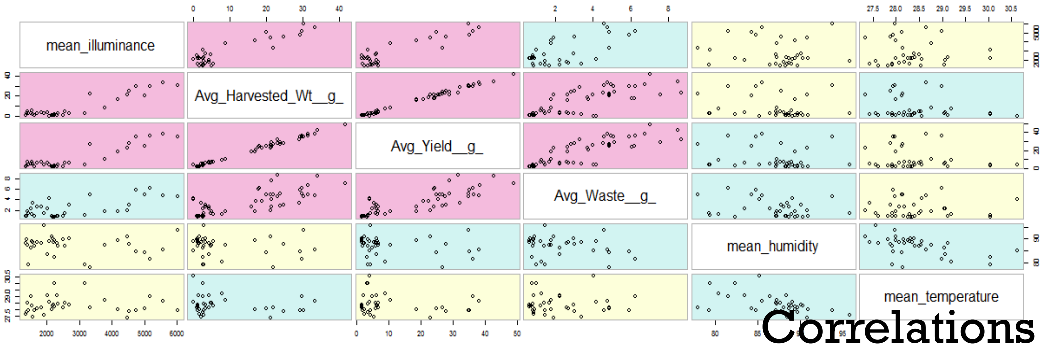


ADVICE

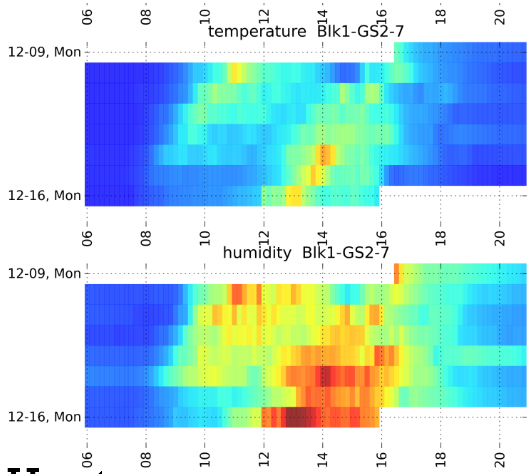
9. Document the journey. Not only will it help in seeing the big picture, it will also convince others of your conclusions.
10. Publish your findings. Use Jupyter Notebook or R Studio or create a web app that allow others to explore the data.
11. Leverage on power tools. Learn to use new software like Apache Spark, or Storm, or Kafka, or Tableau, or TensorFlow.
12. Be objective. Measure. Apply statistical methods for decision-making. Do not jump to conclusions from pictures. Conduct new experiments to test your hypothesis if necessary.



VISUALIZATION TOOLS

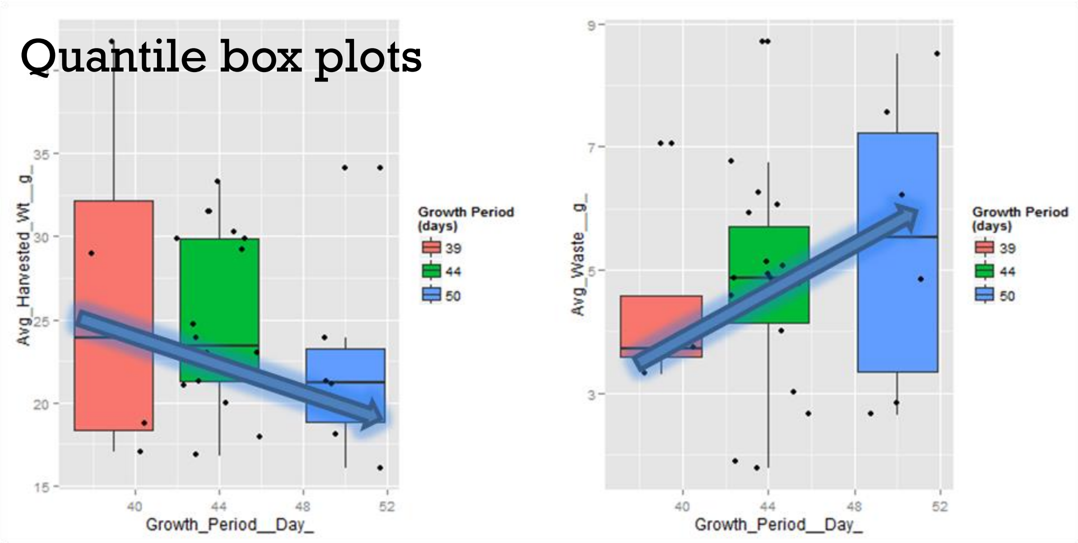


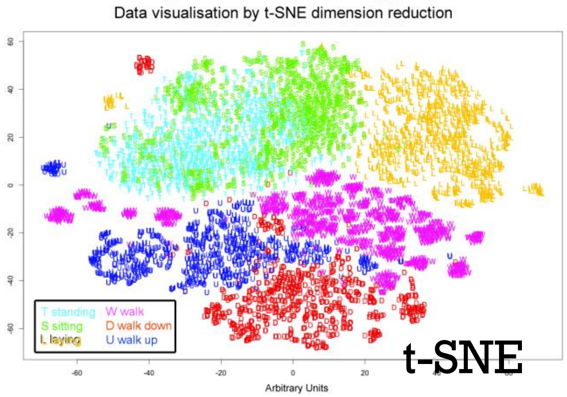
Correlations



Heat maps

Quantile box plots

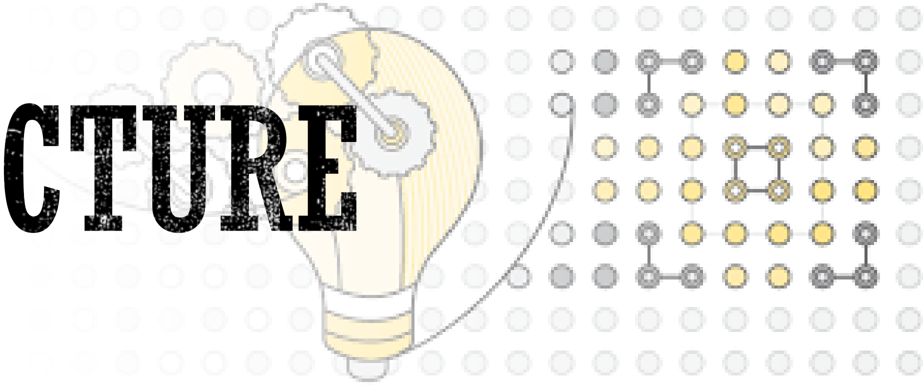




t-SNE

http://scikit-learn.org/stable/auto_examples/manifold/plot_lle_digits.html#sphx-glr-auto-examples-manifold-plot-lle-digits-py
https://distill.pub/2016/misread-tsne/

INFRASTRUCTURE



WEB API

- Set of programming instructions and standards for accessing a Web-based software application or Web tool
- REST (Representational State Transfer) APIs
  use URLs for resources and
  HTTP verbs for actions
- Responses are typically in JSON or XML format
- Extreme form: microservices architecture

Action	HTTP verb
Create	POST
Read	GET
Update	PUT
Delete	DELETE

NOSQL DATABASES

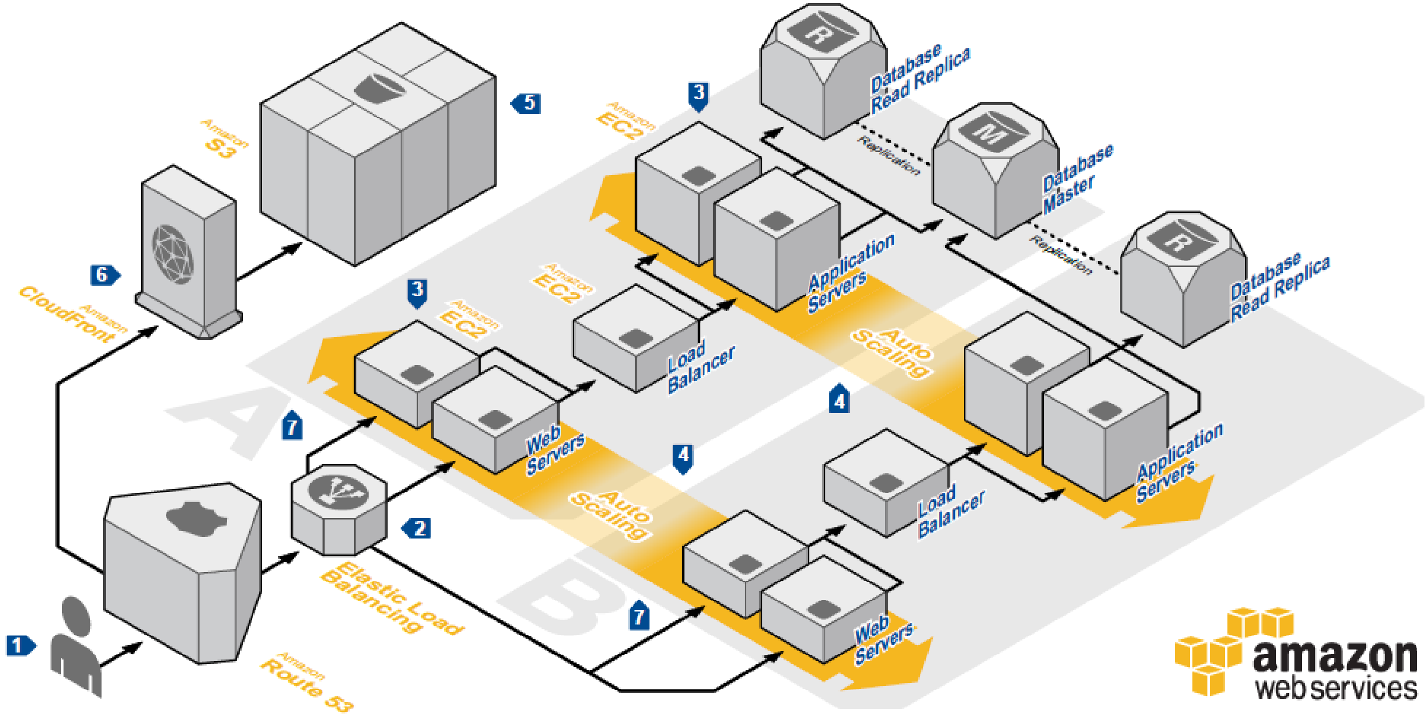
Multiple Data Formats

- Document stores
- Key-value stores
- Wide column stores
- Graph databases

Apache CouchDB
mongoDB
hadoop
redis
ArangoDB
Couchbase
amazon DynamoDB
Cassandra
riak

SCALABLE CLOUD

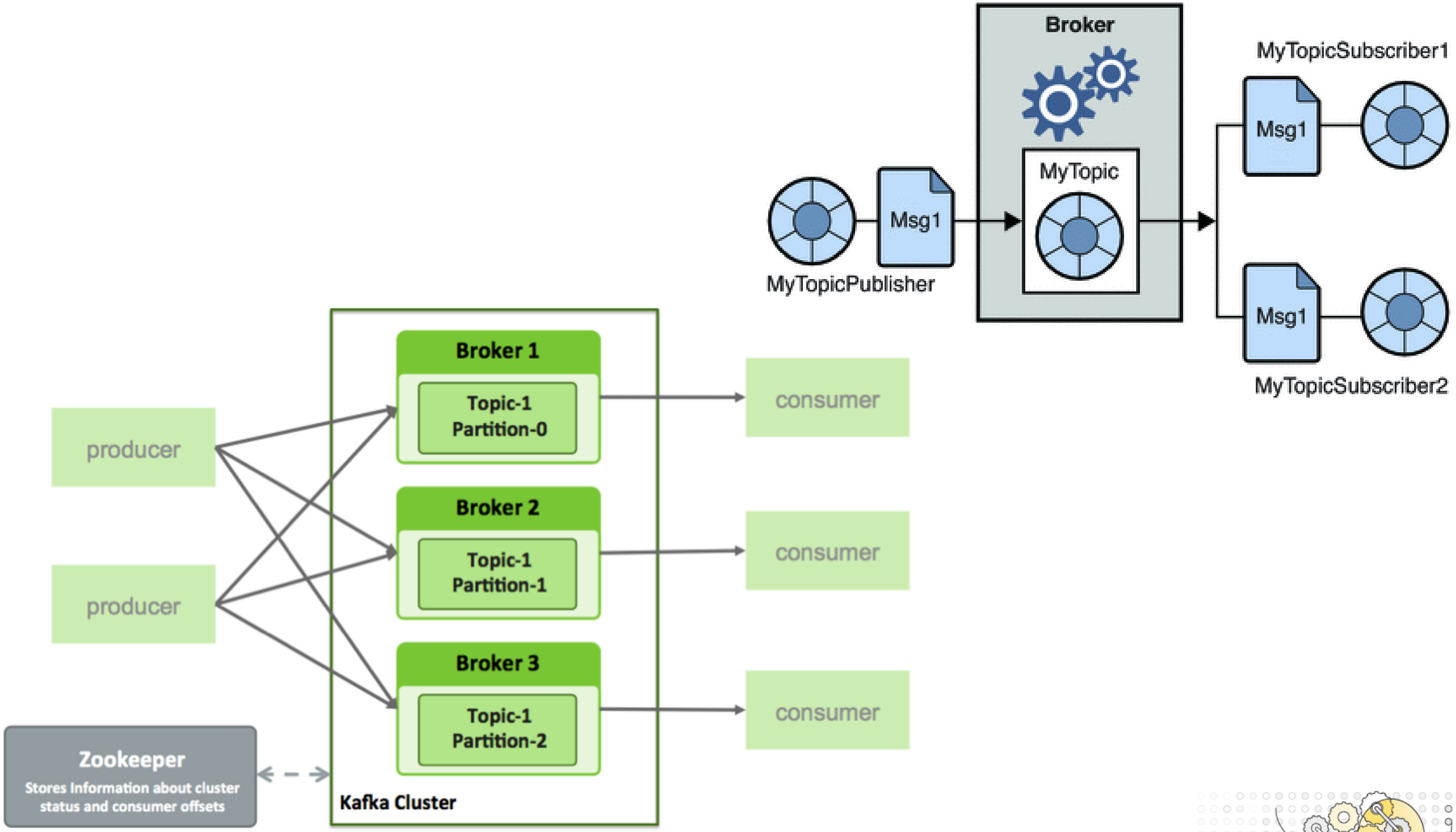
Balancing
Data Access



amazon web services

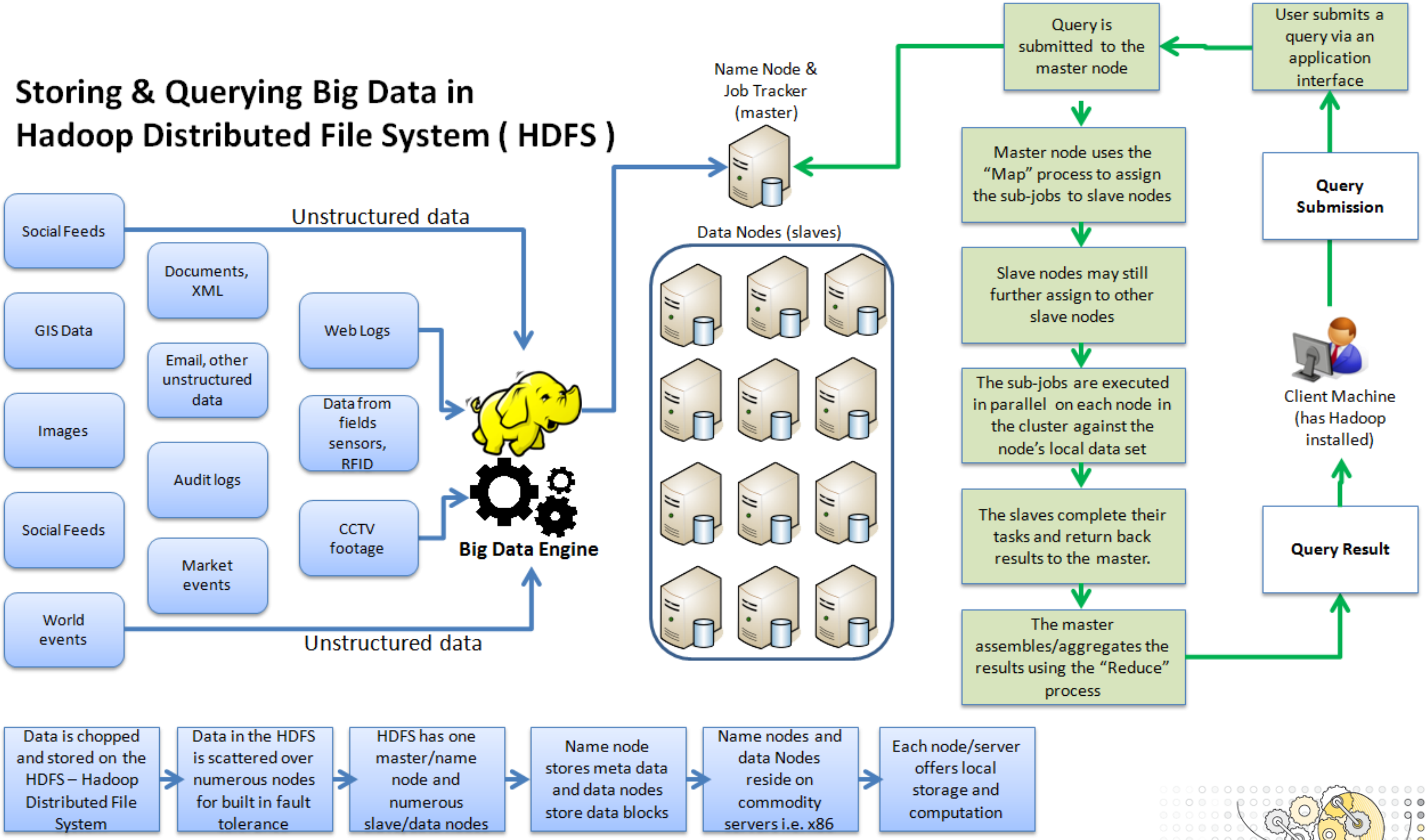
PUBLISH-SUBSCRIBE

Stream Processing



MAP-REDUCE

Balancing Computation



Storing & Querying Big Data in Hadoop Distributed File System (HDFS)

Social Feeds

GIS Data

Images

Social Feeds

World events

Documents, XML

Email, other unstructured data

Audit logs

Market events

Web Logs

Data from fields sensors, RFID

CCTV footage

Unstructured data

Unstructured data

Big Data Engine

Name Node & Job Tracker (master)

Data Nodes (slaves)

User submits a query via an application interface

Query is submitted to the master node

Master node uses the "Map" process to assign the sub-jobs to slave nodes

Slave nodes may still further assign to other slave nodes

The sub-jobs are executed in parallel on each node in the cluster against the node's local data set

The slaves complete their tasks and return back results to the master.

The master assembles/aggregates the results using the "Reduce" process

Query Submission

Client Machine (has Hadoop installed)

Query Result

Data is chopped and stored on the HDFS – Hadoop Distributed File System

Data in the HDFS is scattered over numerous nodes for built in fault tolerance

HDFS has one master/name node and numerous slave/data nodes

Name node stores meta data and data nodes store data blocks

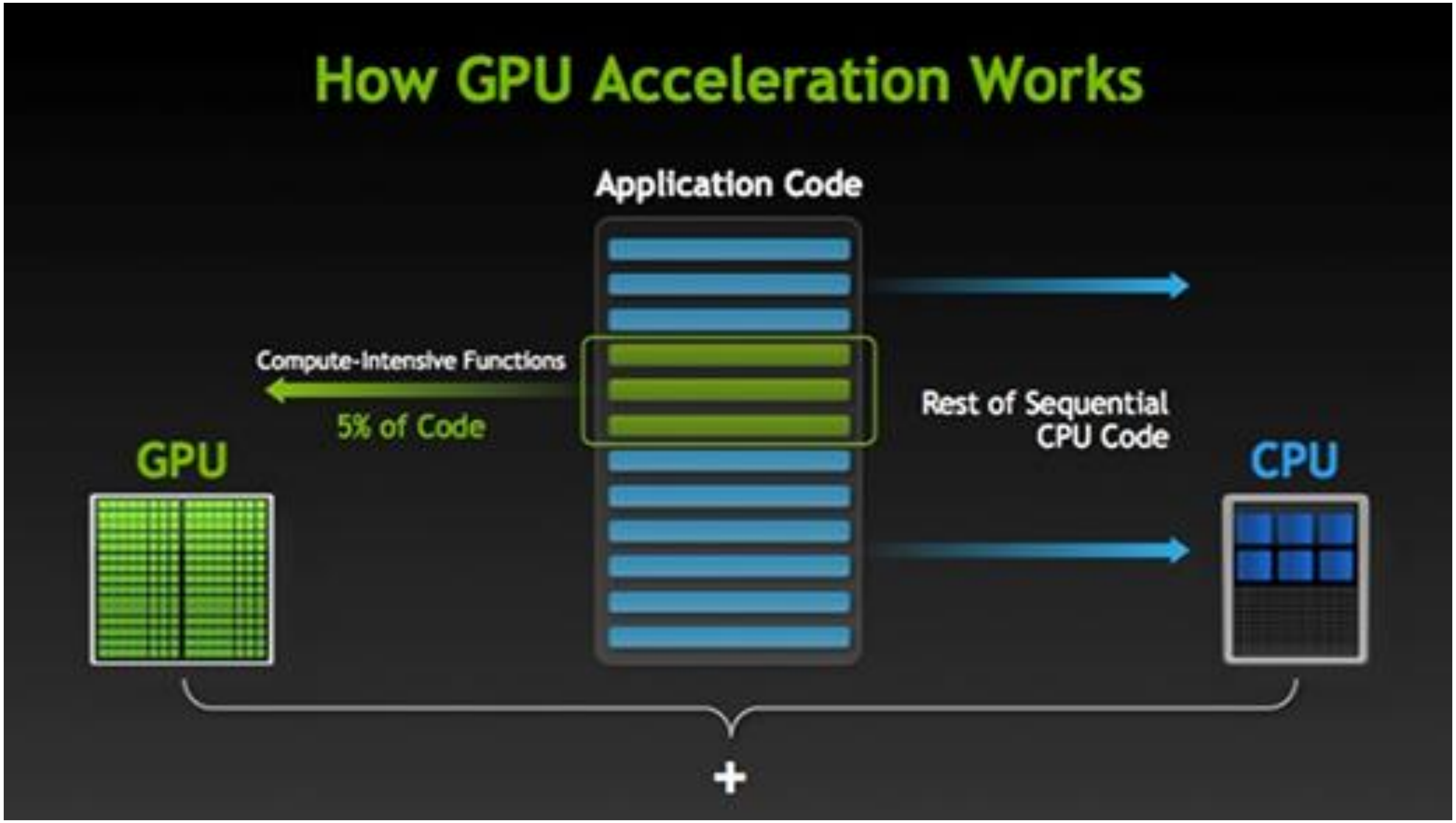
Name nodes and data Nodes reside on commodity servers i.e. x86

Each node/server offers local storage and computation

Designed by Sri Prakash, November 2012

CPU+GPU COMPUTING

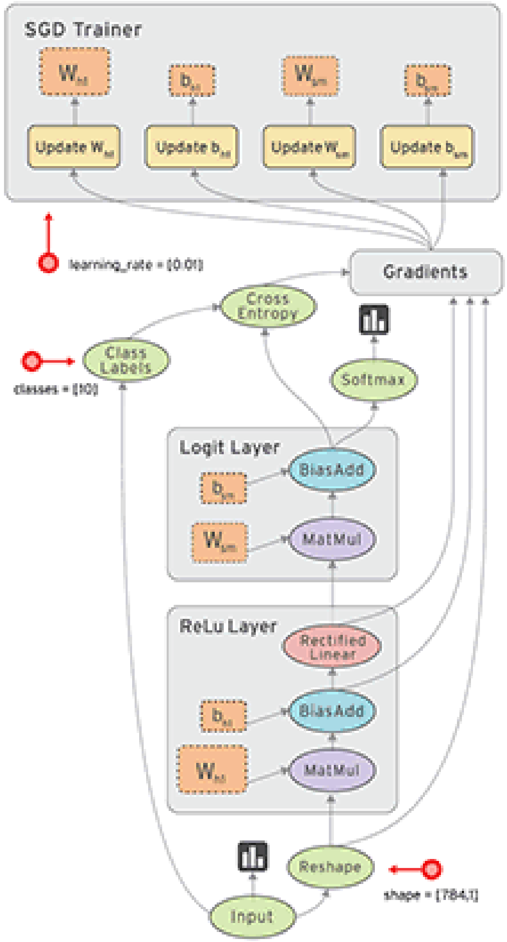
Balancing Computation



How GPU Acceleration Works
Application Code
Compute-Intensive Functions
5% of Code
GPU
Rest of Sequential CPU Code
CPU
+

COMPUTATION GRAPH

Balancing Computation

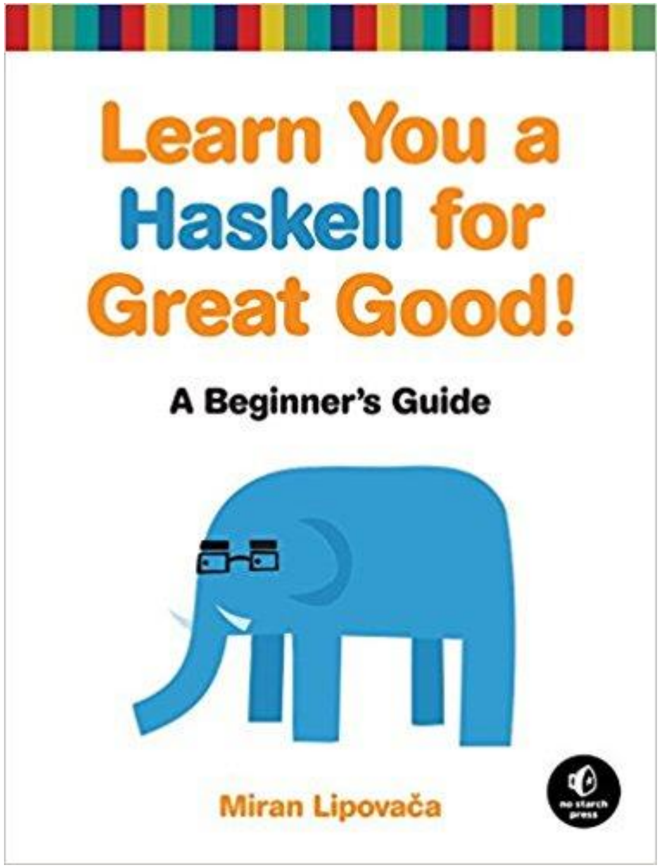


FUNCTIONAL PROGRAMMING

- Python, R
  (not really…)
- Julia
- Clojure
- Haskell, …

Higher-order languages
(Dependent types,
inductive types)

- Coq, Agda, …



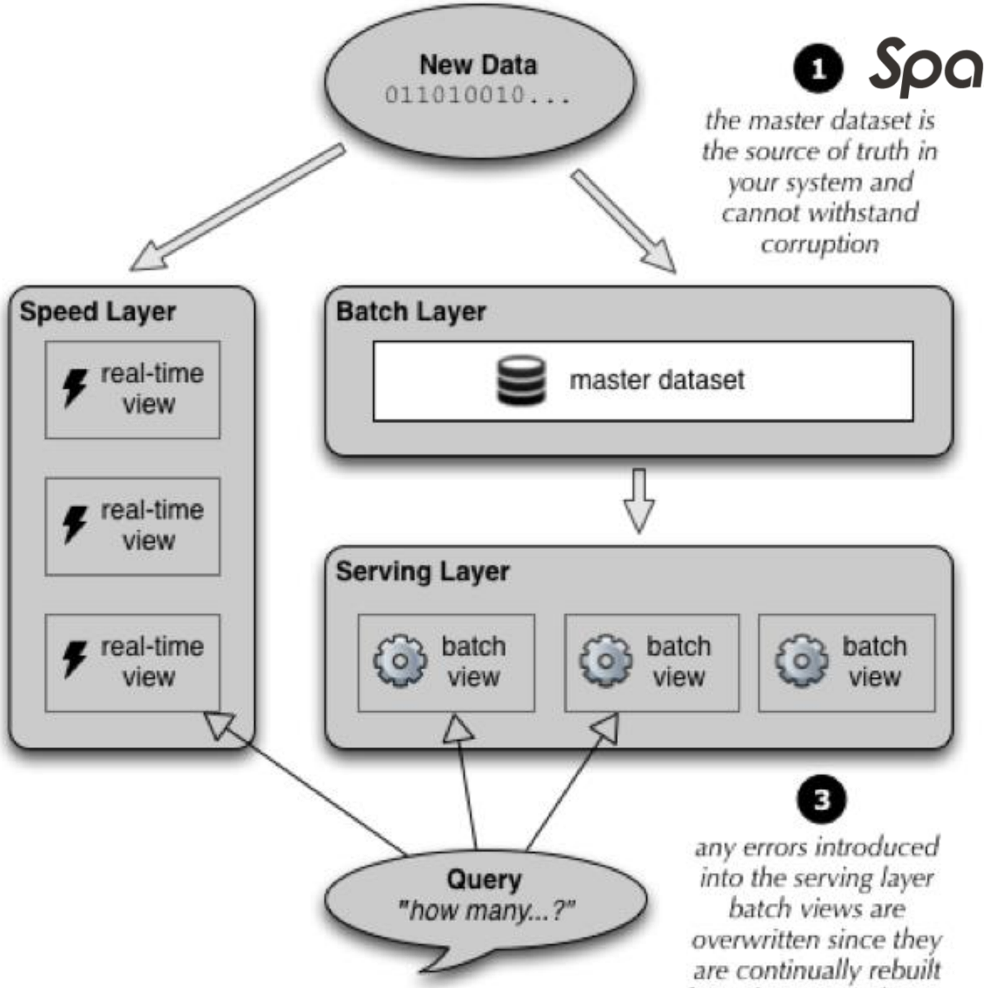
LAMBDA ARCHITECTURE

Putting it all together

STORM

2

the data in the speed layer real-time views has a high turnover rate, so any errors are quickly expelled

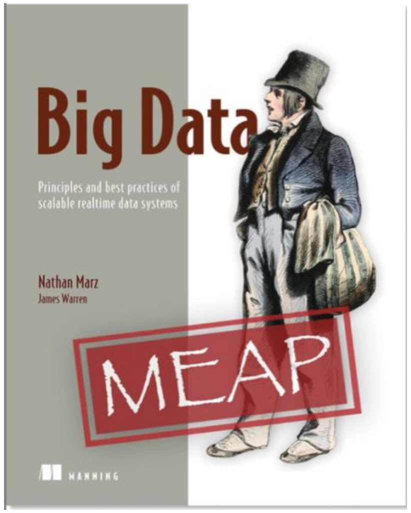


1 Spark

the master dataset is the source of truth in your system and cannot withstand corruption

3

any errors introduced into the serving layer batch views are overwritten since they are continually rebuilt from the master dataset



Big Data

Principles and best practices of scalable realtime data systems

Nathan Marz
James Warren

MEAP

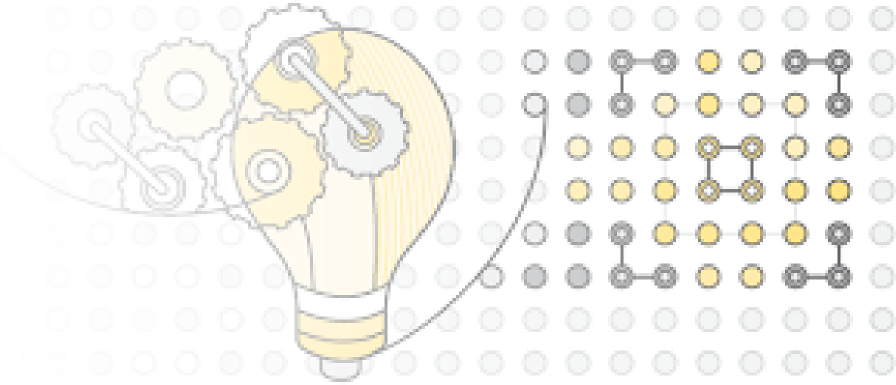
MANNING

LINKED DATA

Interoperability

Manu Sporny, "What is Linked Data?" June 2012. <https://www.youtube.com/watch?v=4x_xzT5eF5Q>

META



NOT IN THIS COURSE, UNFORTUNATELY

1. Time Series, Stochastic Processes
2. Principal Component Analysis
3. Quantile Regression
4. Random Forest
5. Decision Theory
6. Sparsity, Regularization, Dimensionality Reduction
7. Cloud, Fog Computing
8. Knowledge Graphs
9. Machine Reasoning

ETHICS

- With great power comes great responsibility
- Democratization of intelligence?
- Disruption of human history, culture, values