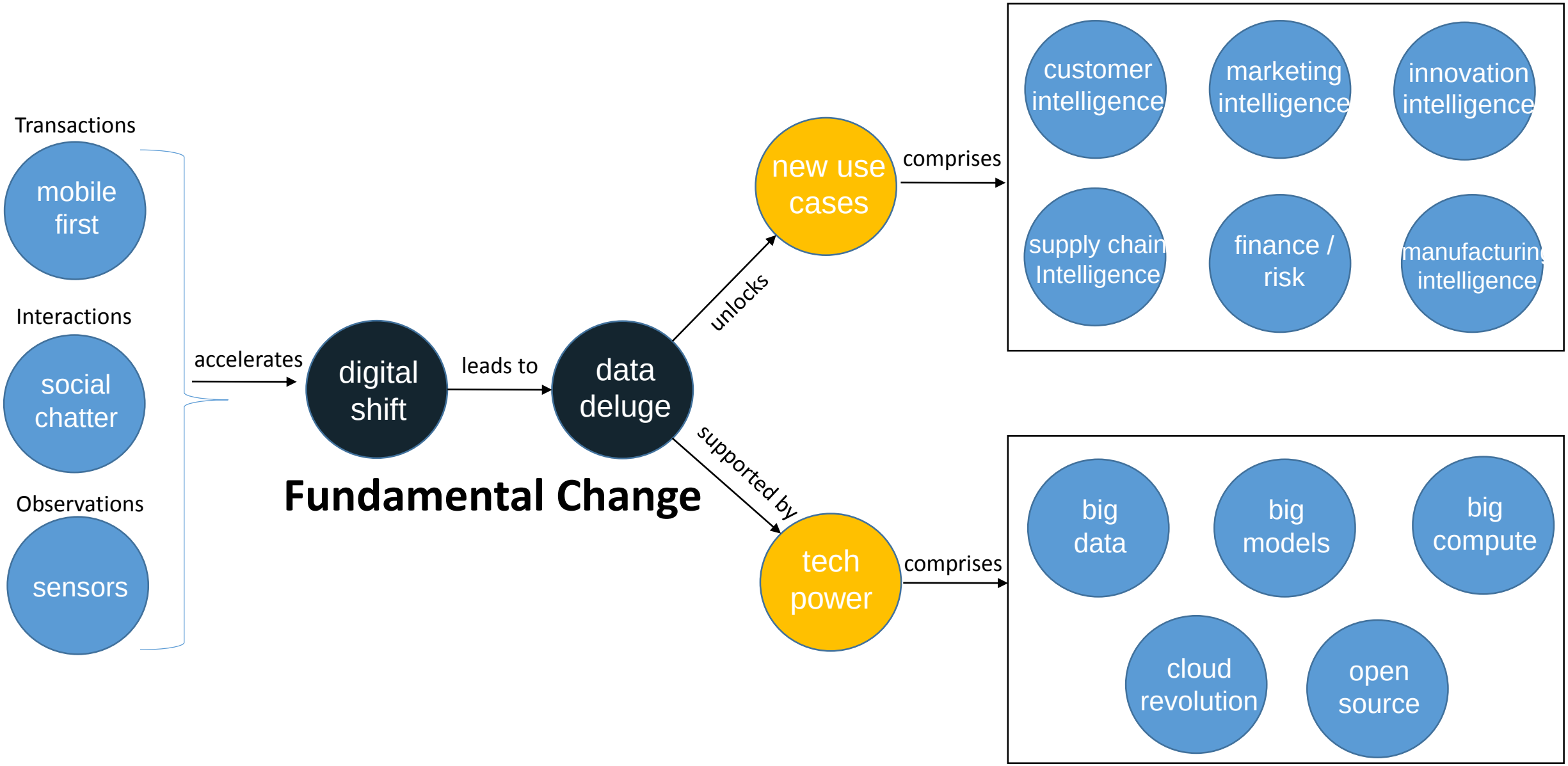
Navigating the Data Science World

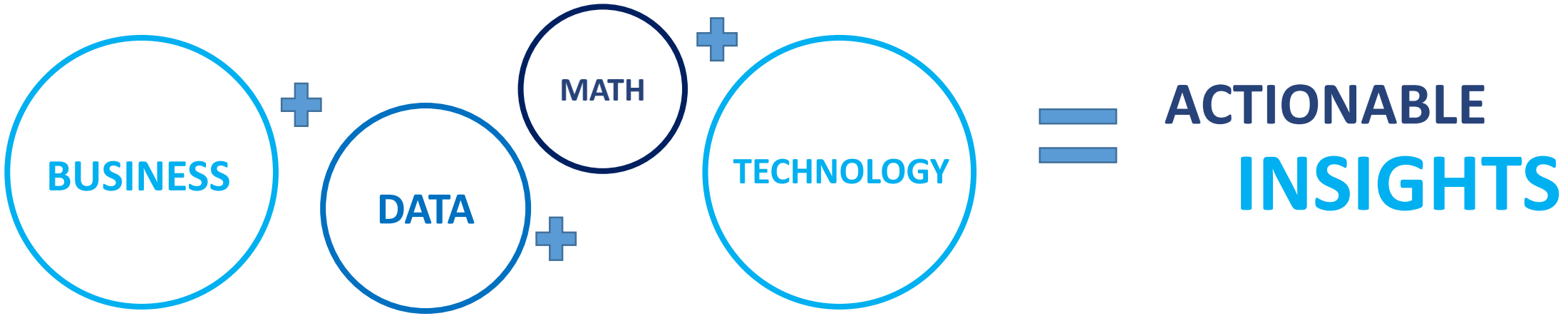
November 2017

By Karthikeyan Sankaran

Data Science & Analytics – 'Clear & Present' Opportunity



What does it take to produce ‘Actionable Insights’



What are the dimensions of Analytics?

Business	Use Case Formulation	Interpret Analytics Output	Domain Expertise
Data	Data Acquisition	Data Exploration & Visualization	Data Pre-processing
Math / Quant	Understand the algorithms	Select the right techniques & code	Evaluating the output of algos
Tech / Software	Understand the IT Ecosystem	Data Engineering & Architecture	Software Engineering / SDLC

My Analytics Mindmap

- Global Trends in Society
- Macro-economy
- Business Fundamentals
- Specific Industry Domain
- Analytical use cases

- Data Management
- Reporting & Self-service
- Quantitative Techniques
- Performance Mgmt
- Insight Delivery

Analytics for Business Value
http://bit.ly/31KArT8

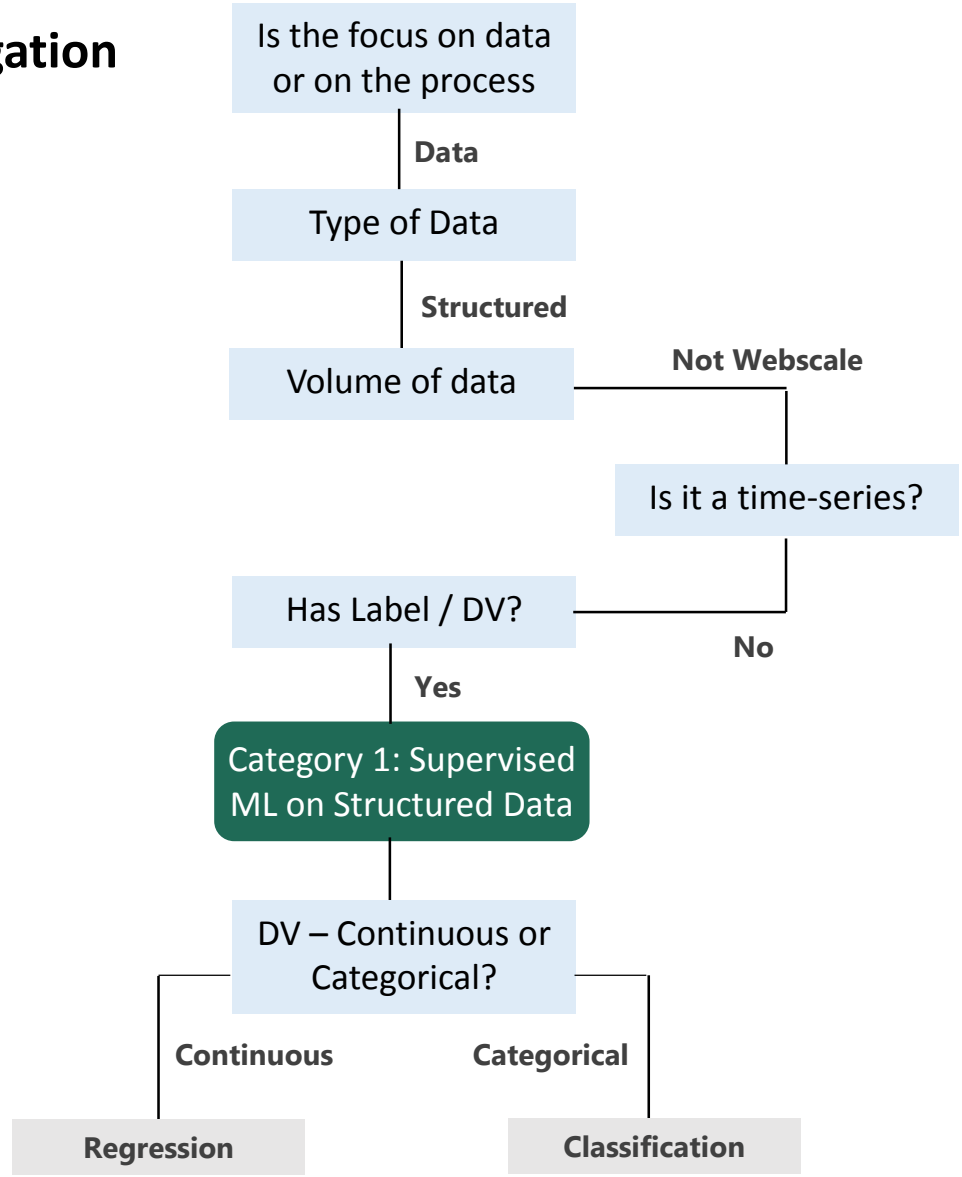
- Scan for New Products
- Evaluate Maturity

- Monitor Ecosystem
- Leverage Resources

Data Science Techniques - What's the real problem?

Category 1

Navigation

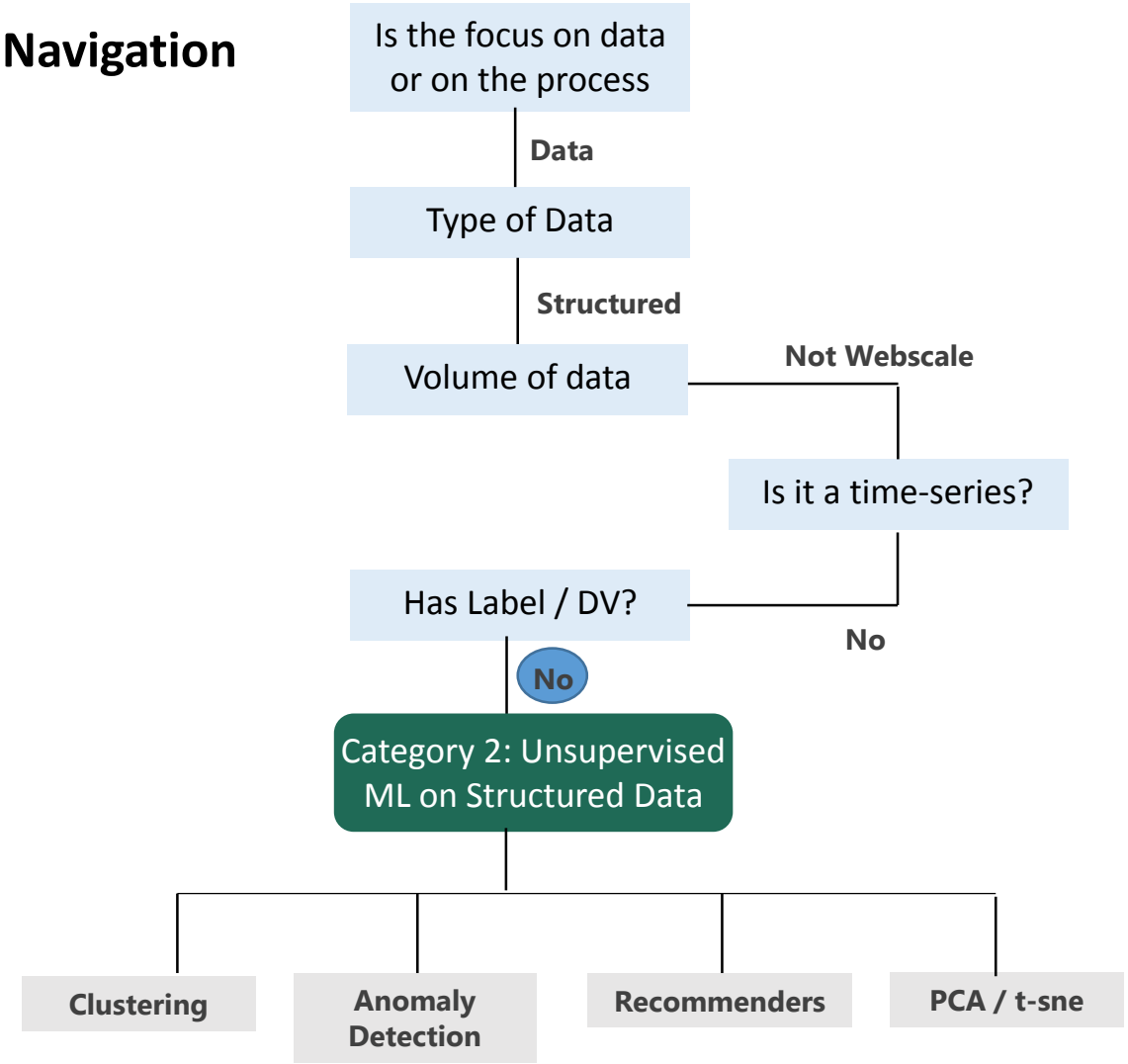


Details

- Exploratory Data Analysis (EDA)
- Data Pre-processing – Outliers, Missing data, Variable Transformations
- Feature Selection & Dimensionality Reduction
- Feature Engineering
- Algorithms – Standalone vs Ensembles
- Algorithms – Parametric vs Non-Parametric
- Algorithms – Linear vs Non-linear
- Cross validation
- Hyper-parameter Tuning
- Predict on Test set

Category 2

Navigation

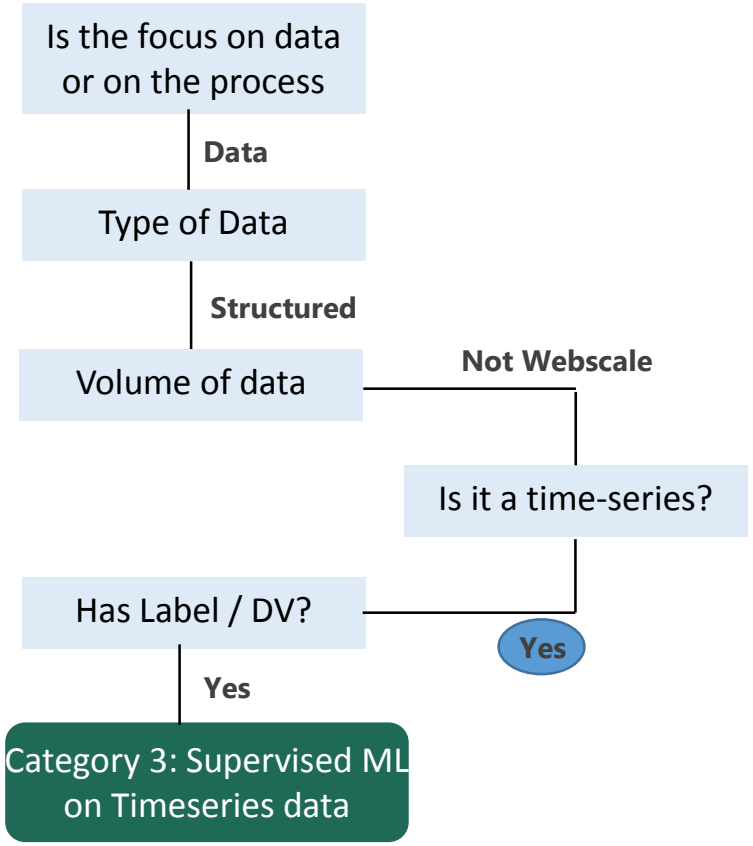


Details

- Exploratory Data Analysis (EDA)
- Clustering – K-Means, Hierarchical, many others…
- Anomaly Detection – Isolation Forest, LoF, many others…
- Recommenders – Content based, Collaborative Filtering, Hybrids
- Self-Organizing Maps (SOMs) – Use Deep Learning for structure discovery

Category 3

Navigation

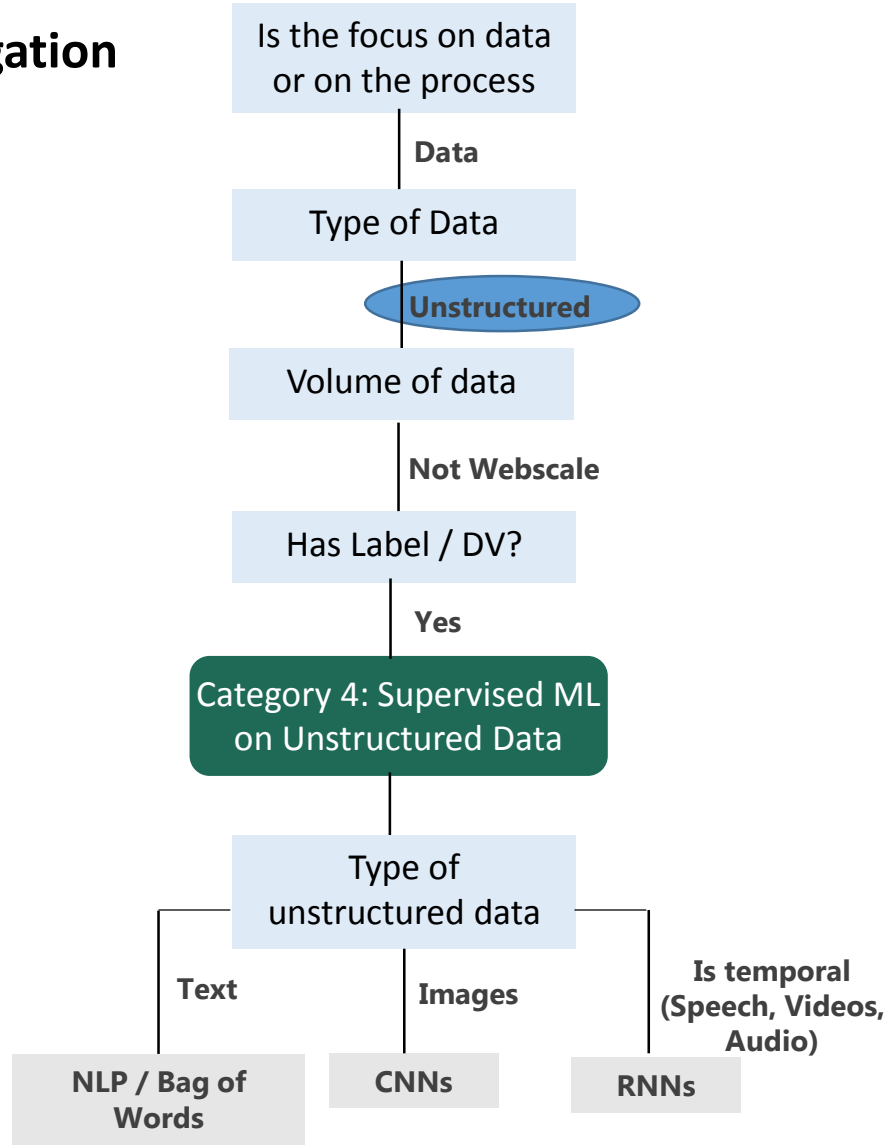


Details

- Univariate vs Bivariate timeseries
- Visualizing and Decomposing Time series
- Stationary & Non-stationary Time series
- Plot Auto-correlation plots to find optimal differencing parameters
- Feature engineering – Creating features like lag, moving average etc.
- Build Timeseries forecasting models like ARIMA, Holt-Winters etc.
- Deep Learning Techniques like Recurrent Neural Networks (RNN)

Category 4

Navigation



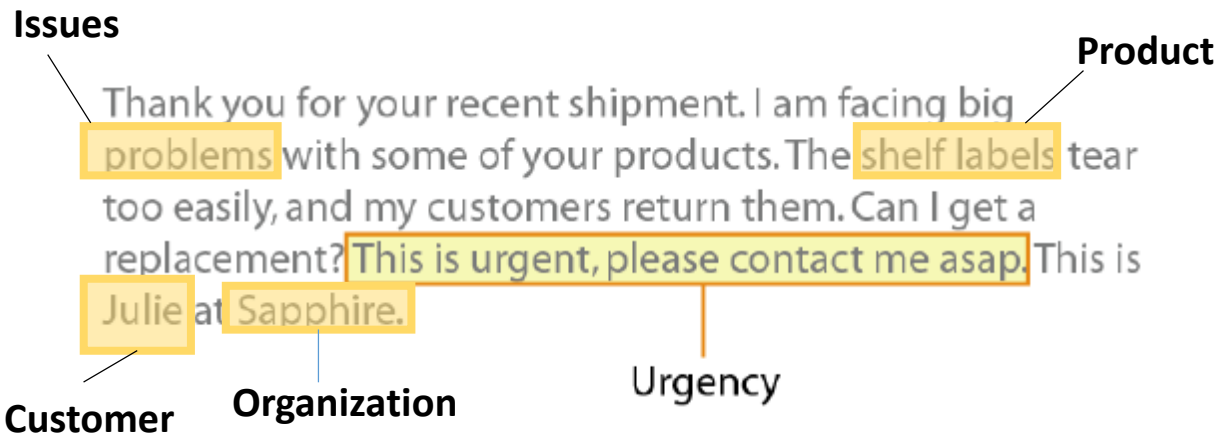
Details

- NLP – Natural Language Processing. Used in areas like email spam detection, sentiment analysis, sarcasm detection etc.
- CNNs – Convolutional Neural Networks - Specialized Neural Network Architecture with Convolutional layer, Pooling layer, Flattening and Fully Connected layers to detect image features
- RNNs – Recurrent Neural Networks – Specialized ANN architecture with feedback loops that helps in short-term memory. (Ex: LSTMs)

Brief Detour - Why is Text Analytics Important

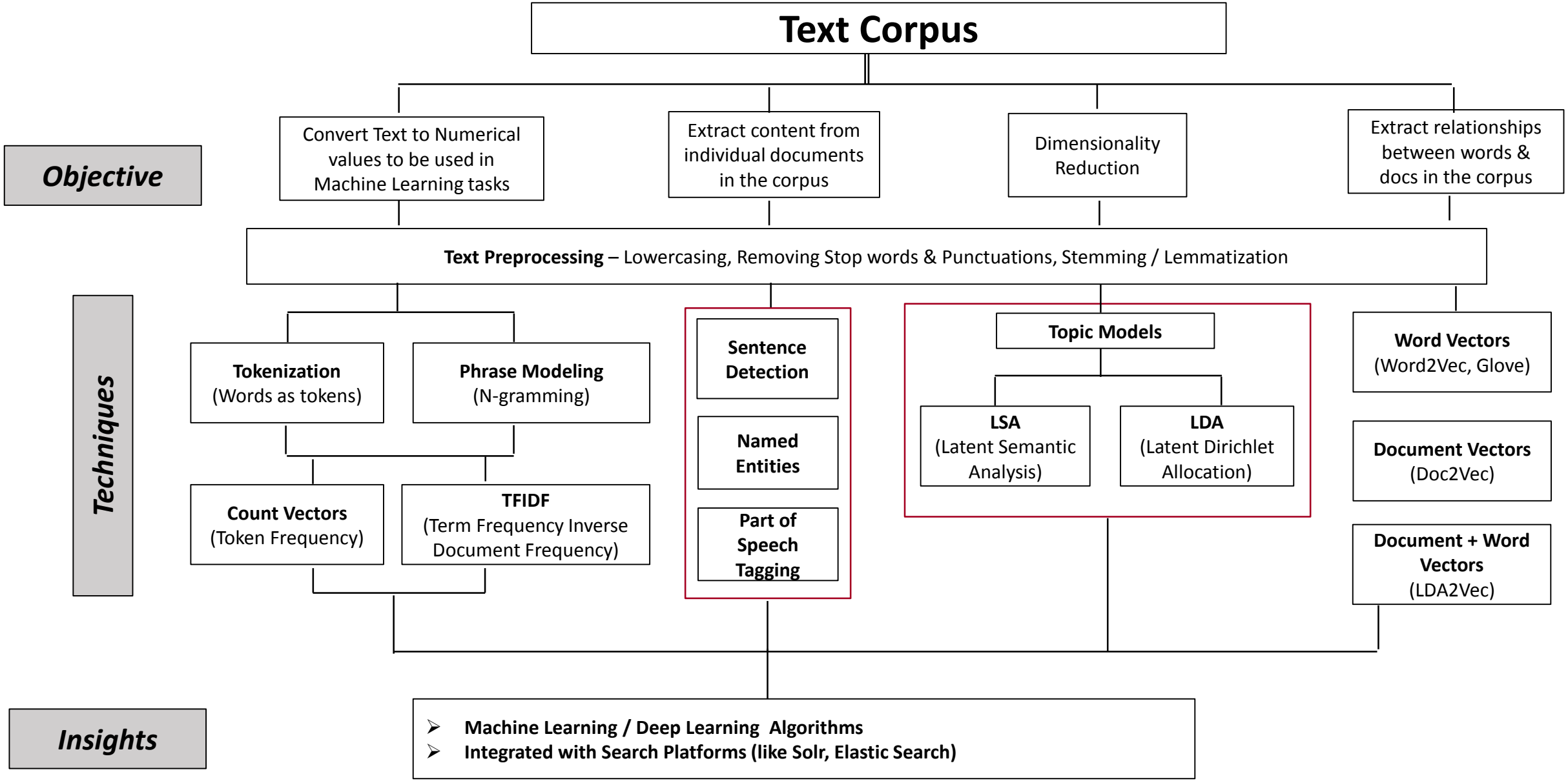
1 **Lots of unstructured text** – 80% of data in organization is unstructured and text data is rapidly growing

2 **Text packs a lot of information**



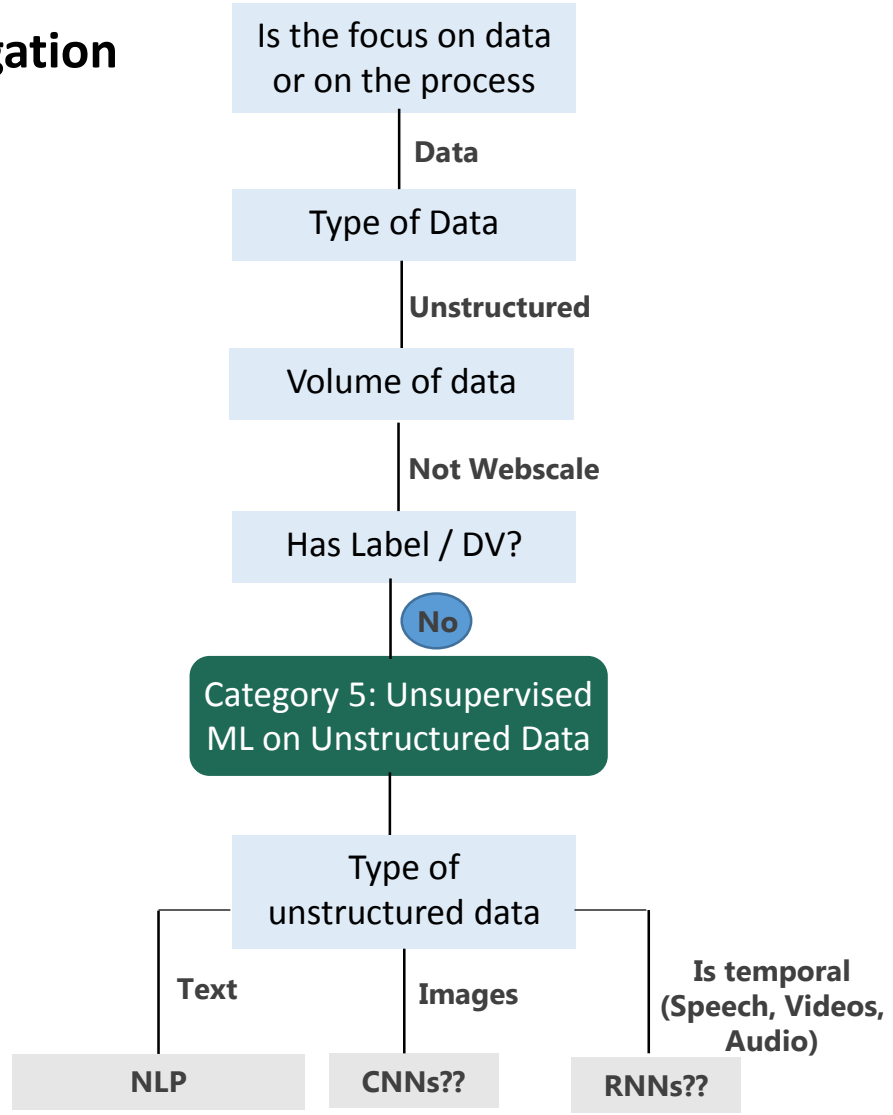
3 **Text Analytics is the foundation to higher levels of cognitive technologies & to artificial intelligence**

NLP in 1 Slide



Category 5

Navigation

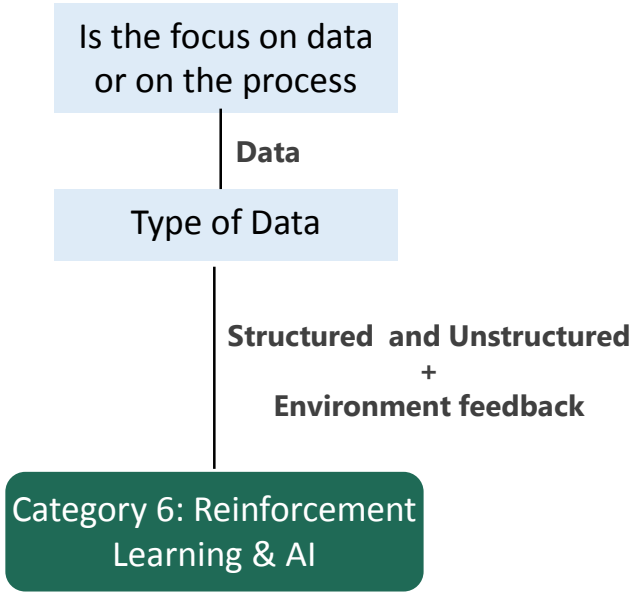


Details

- Full range of NLP can be utilized – Phrase Modeling, Entity Extraction, Part of Speech tagging, Topic Models, Word embeddings, etc.
- GANS – Generative Adversarial Networks. Artificial Neural Network architectures that generate text, art etc.
- Restricted Boltzmann Machines, Auto Encoders etc.

Category 6

Navigation

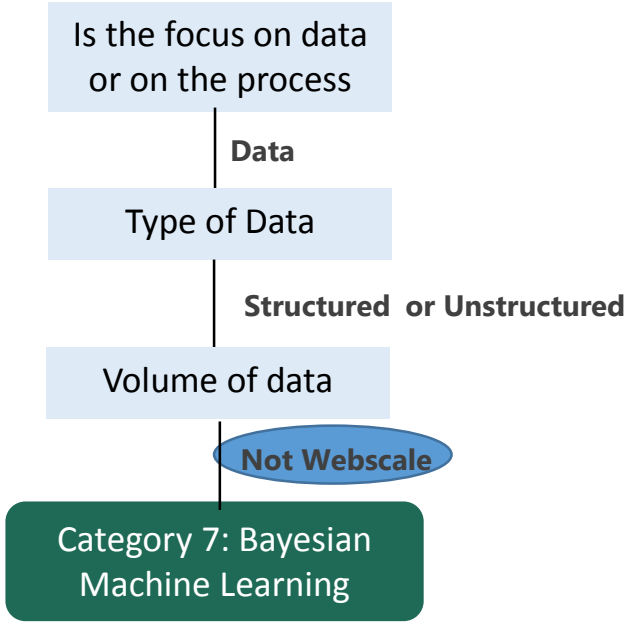


Details

- Delayed Feedback from environment
- States, Policies, Actions
- Key Terms / Techniques – Multi-armed bandits, Q Learning, Bellman's equation, Markov Decision Process, etc.
- In my view, AI is also in this category as all AI systems needs to work with structured & unstructured data and take actions based on their perception of the environment
- AlphaGo, Autonomous cars etc.
- Platforms – OpenAI, Gym, Universe etc.

Category 7

Navigation



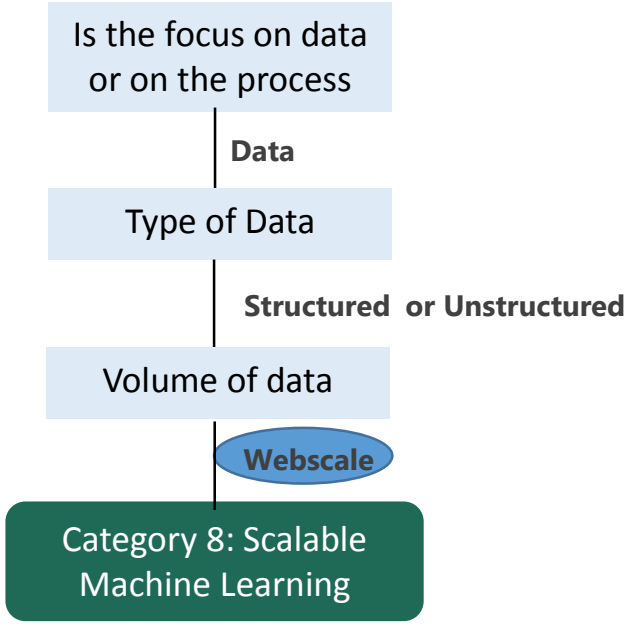
Details

- Fundamental idea is to use Bayes Theorem to estimate:

  Probability (Parameters | Data)
- All ML algorithms have a Bayesian variation – Ex: Bayesian Linear Regression, Bayesian Logistic Regression, Bayesian Decision Trees etc.
- Estimating uncertainty is critical for business decision making
- Packages for computation – Stan, PyMC3 etc.

Category 8

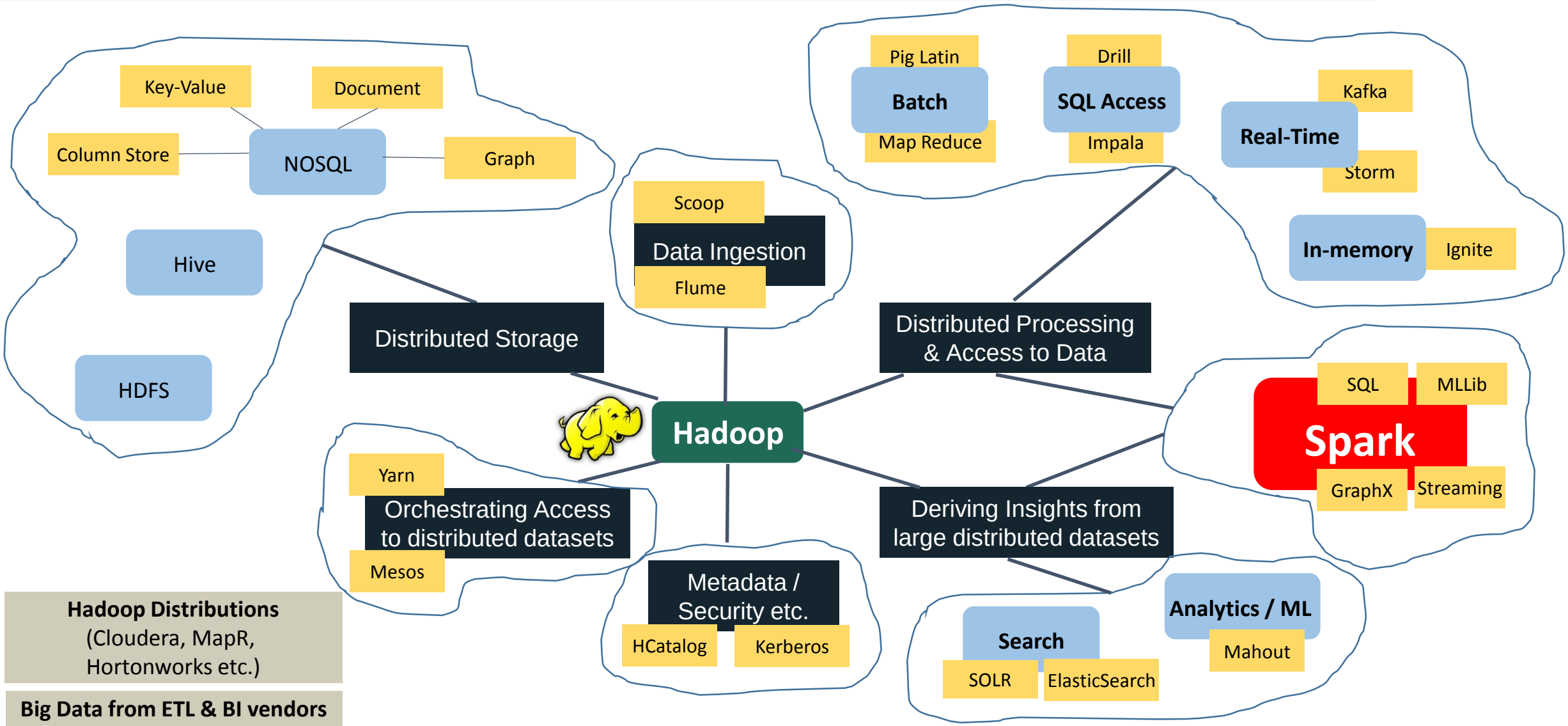
Navigation



Details

- Scaling both Supervised and Unsupervised Machine Learning Algorithms
- Spark, Big Data, Cloud – All become relevant in addition to specific algorithmic techniques
- Check out H2o.ai – Open Source platform build ground up for Machine Learning
- I would also put ‘Enterprise Search’ (ELK) into this category as elasticsearch, solr etc. are about providing structure to massive amounts of data so that it can be easily retrieved

Big Data – Key Technology Enabler

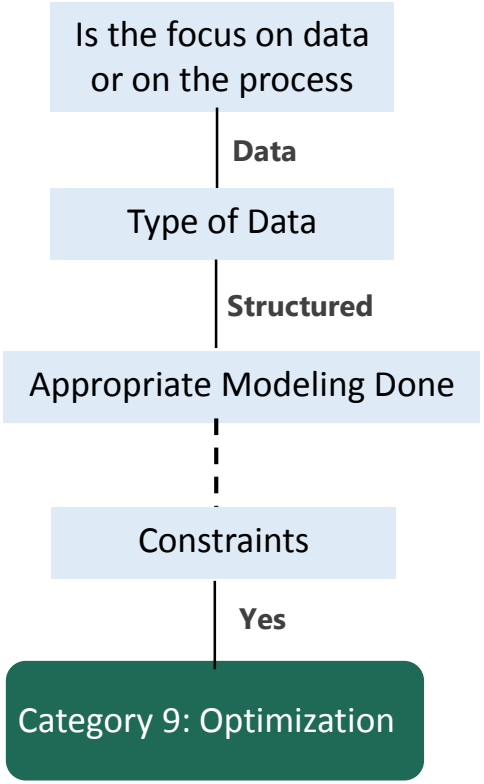


Hadoop Distributions
(Cloudera, MapR, Hortonworks etc.)

Big Data from ETL & BI vendors
(Infa, Teradata, MSTR, Tableau etc.)

Category 9

Navigation



Details

- Very relevant in business context as they always have constraints
- Many optimization methods are available – most common one is Linear Optimization (Simplex) with its variants like Integer, Quadratic optimization etc.
- Ex: PuLP package in Python helps to execute Linear Programming

Category 10

Navigation

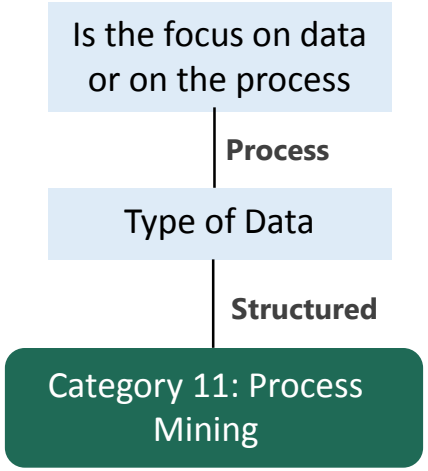
Category 10: ML in Production

Details

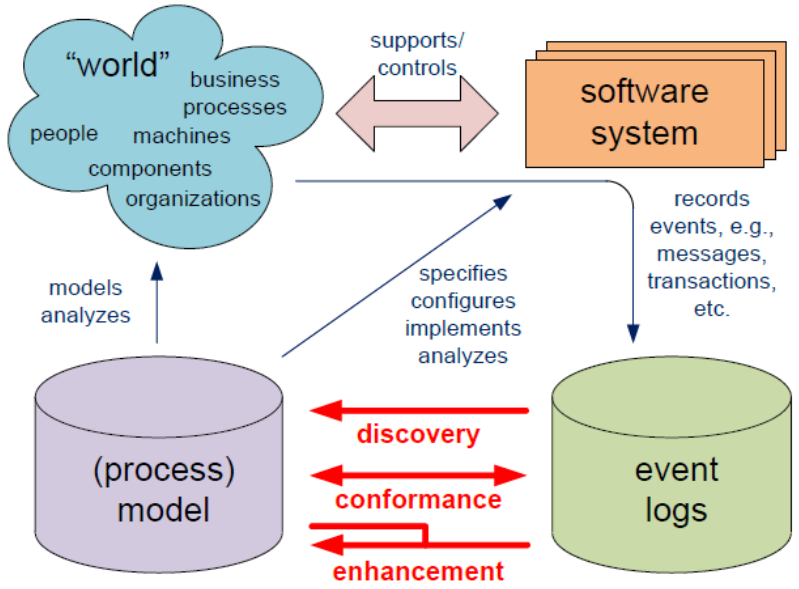
- Architectural considerations:
  1. Data Products vs Applications
  2. User access (Concurrent etc.)
  3. Frequency of predictions
  4. Frequency of model calibration
  5. How many models are deployed (Is is one per customer segment?, one per product? etc.)
  6. Logging? Monitoring? Error Handling? Fault Tolerance?
  7. Software Engineering Principles
- Data Pipelines
- Webservices / APIs
- Containers & its orchestration (Ex: Dockers, Kubernetes etc.)
- Hardware – GPUs
- Notebooks (Ex: Zeppelin)
- Streaming data (Ex: Storm, Kafka etc.)

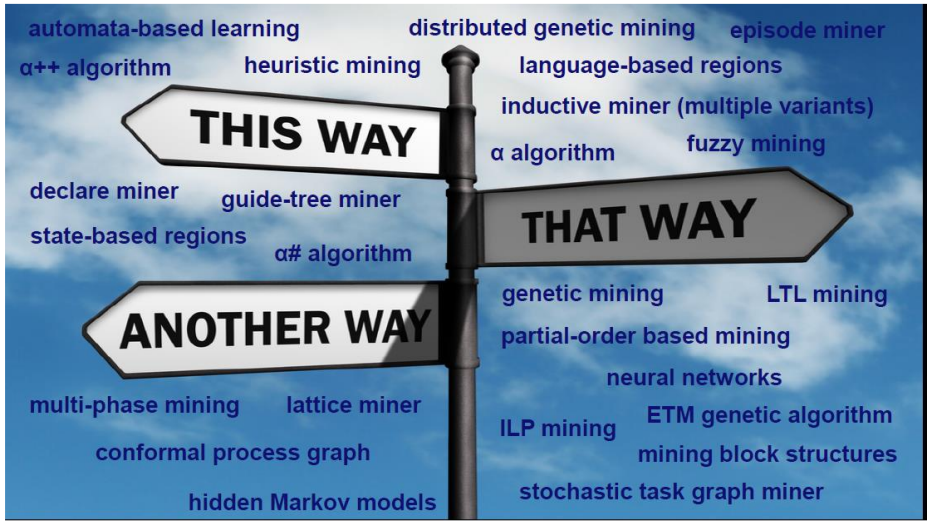
Category 11

Navigation



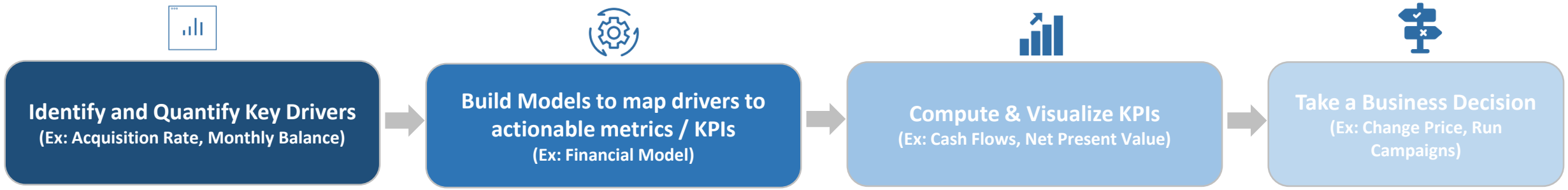
Details



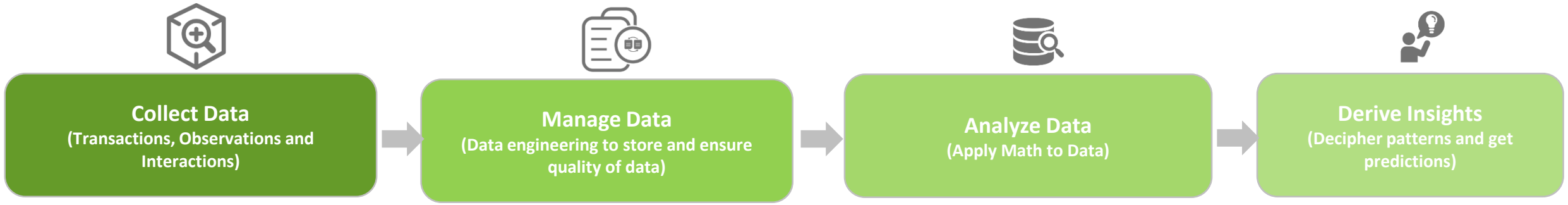


Key Skill #1: Business Connect

Business Decision Making Pipeline



Gap that needs to be bridged to deliver maximum business value



Data Science Pipeline

Key Skill #2 - Hands-on Knowledge

Categorization of Data Science Topics

Categories

- Category 1: Supervised ML on Structured Data
- Category 2: Unsupervised ML on Structured Data
- Category 3: Supervised ML on Time-series data
- Category 4: Supervised ML on Unstructured Data
- Category 5: Unsupervised ML on Unstructured Data
- Category 6: Scalable Machine Learning
- Category 7: Reinforcement Learning / AI
- Category 8: Business Optimization
- Category 9: Machine Learning in Production
- Category 10: Process Mining

Programmers

R, Python, etc.

Non-Programmers

Azure ML, BigML etc.

Key Skill #3 – Think Technology Landscape

- Cloud
- Mobility
- Web Technologies
- Embedded Analytics in Applications
- Legacy Systems

Q & A

- Karthikeyan Sankaran, Director, LatentView Analytics
- Email ID – Karthikeyan.Sankaran@latentview.com
- LinkedIn – http://in.linkedin.com/in/karthikeyansankaran
- Github – github.com/skkeyan-mlai

Data Science & ML can have great impact on industries

McKinsey&Company

MCKINSEY GLOBAL INSTITUTE

THE AGE OF ANALYTICS: COMPETING IN A DATA-DRIVEN WORLD

DECEMBER 2016

Machine learning has great impact potential across industries and use case types

Impact potential: Low — High

SOURCE: McKinsey Global Institute analysis