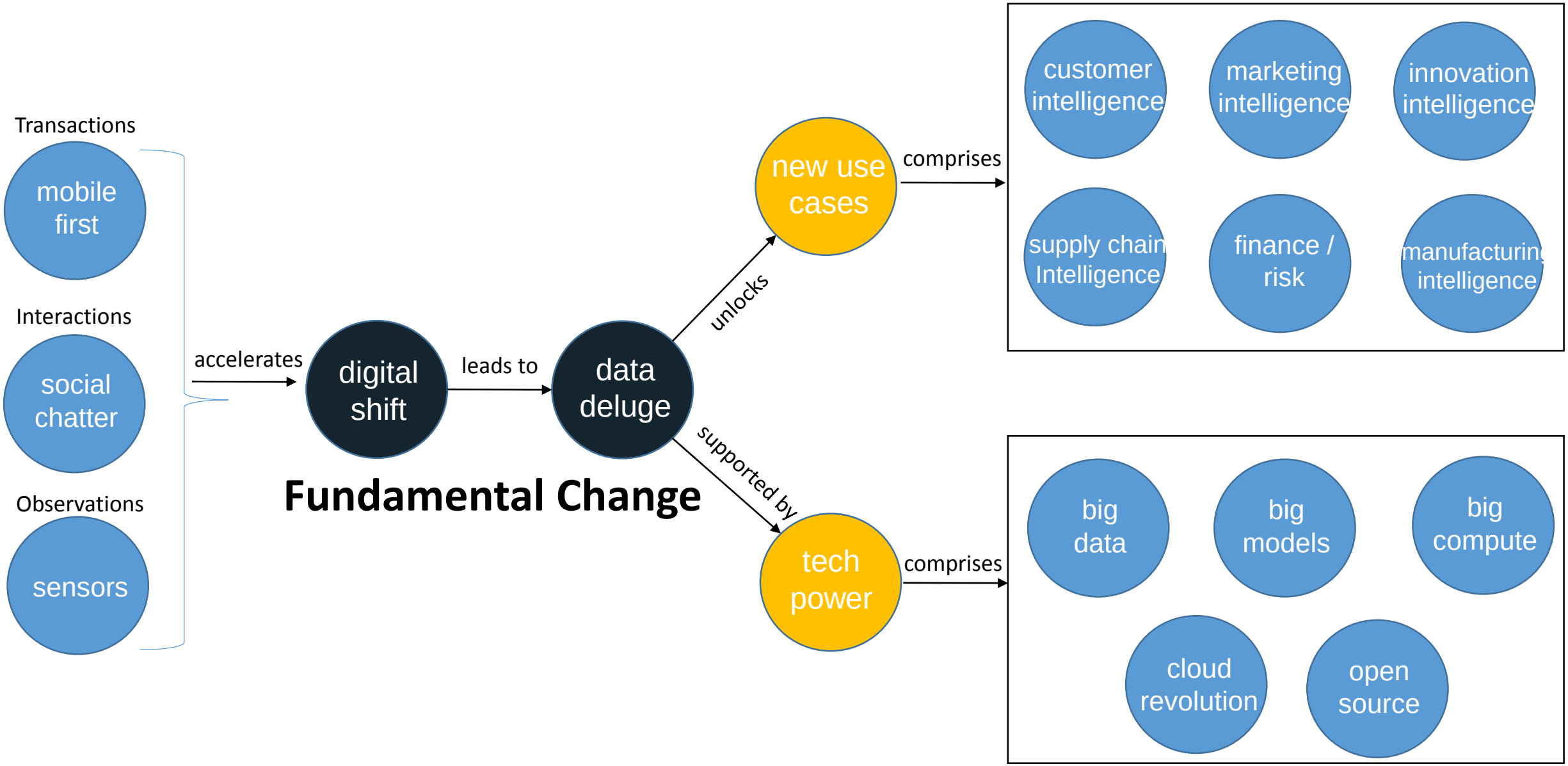
Natural Language Processing – Motivation & Mechanics

By Karthikeyan Sankaran

Why is Analytics fundamental & fascinating?



Data Science & ML can have great impact on industries

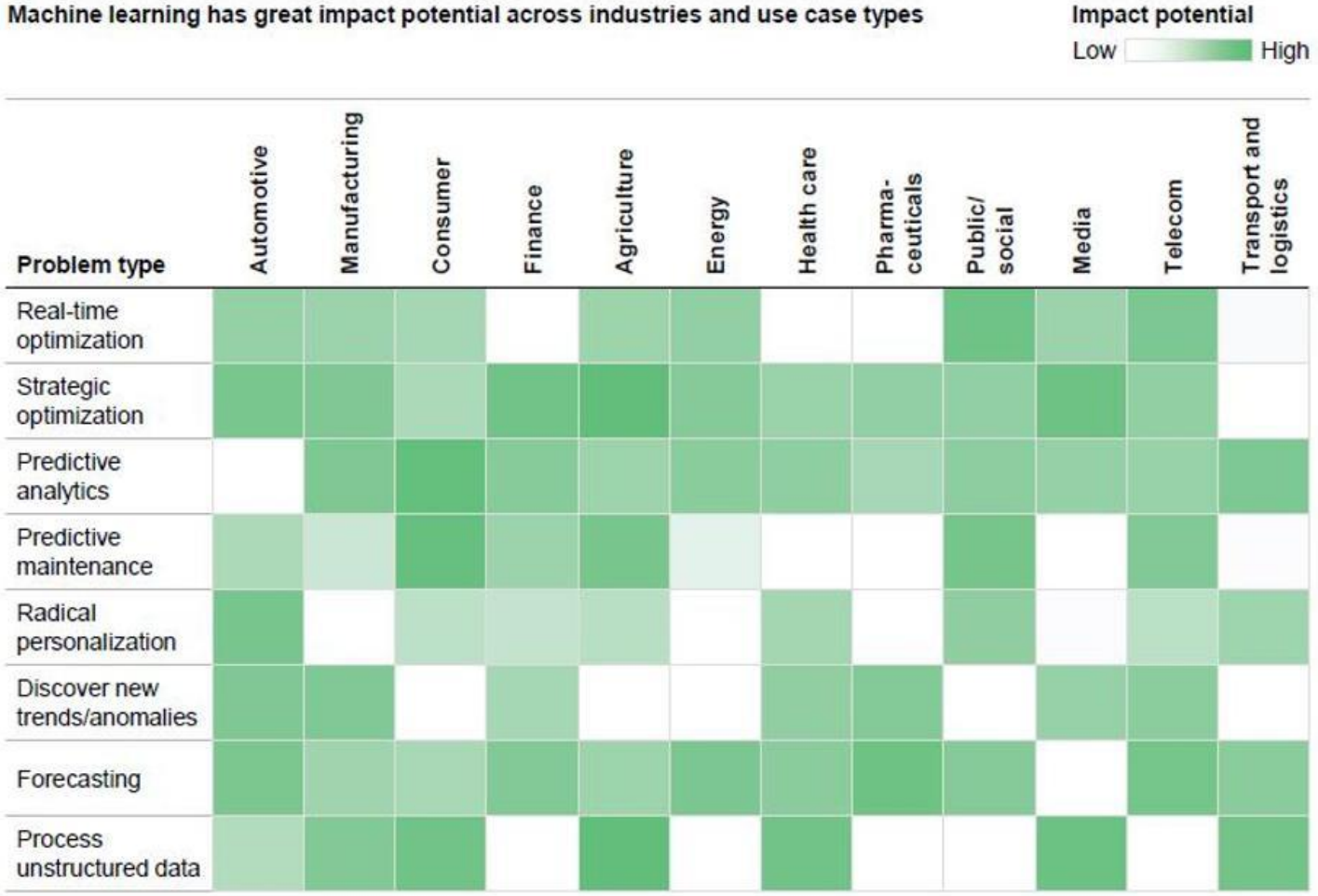
McKinsey&Company

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



SOURCE: McKinsey Global Institute analysis

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

Map to make sense of the data science space

Navigating the Data Science World
http://bit.ly/datascienceworld

Focus in on Data

- 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: Reinforcement Learning / AI
- Category 7: Scalable ML
- Category 8: Optimization
- Category 9: ML in Production

→ **Natural Language Processing (NLP)**

Focus in on System (generating the data)

- Category 10: Bayesian Inference / Probabilistic Programming
- Category 11: Business Process Simulations

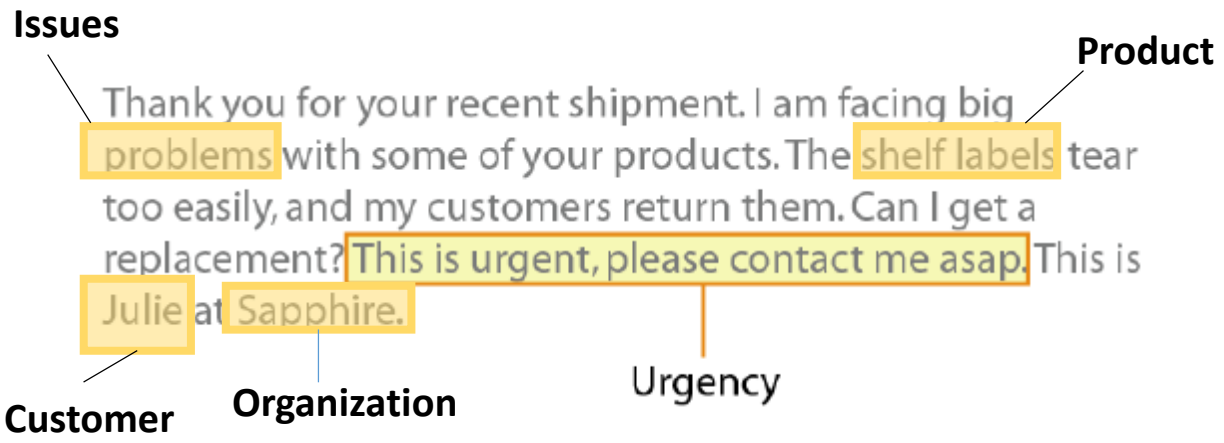
Focus in on Process

- Category 12: Process Mining

Why is Text Analytics Important

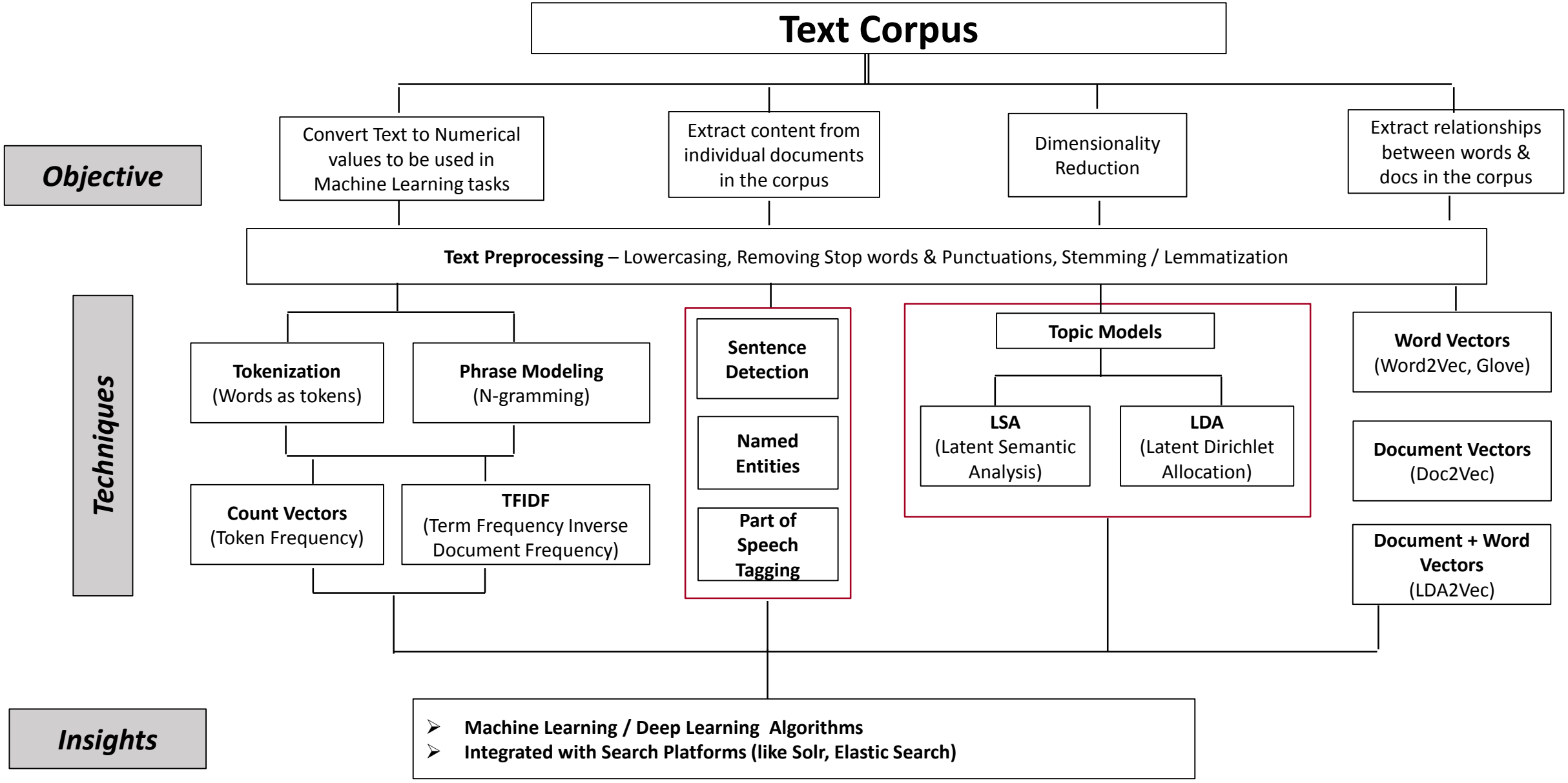
1 **Lots of unstructured text** – 80% of data in organization is unstructured and text data is rapidly growing

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



Let’s start with this sample text data

Record ID	Attribute 1 (Cust ID)	…	Attribute N	Text Data (Ex: Review)	Label (Ex:Rating)
1	1345	…	Place 1	The staff is friendly and the joint is always clean! Kudos to the place…	4
1	1345	…	Place 1	Horrible - don't waste your time and money.	1
1	1345	…	Place 1	Great place to have a couple drinks and watch any and all sporting events as the walls are covered with TV's..	5

Reviews Record Set

*Label might or might not be present

Structured Data

Unstructured Data

Document = Text Data corresponding to each row in the table

Corpus = Collection of all documents across all rows in the table

Words = Each word in the Text Data Row

Action 1: Extract Text, Pre-process, Create DTM & TFIDF

Reviews Record Set

Extract Text

Document ID	Text Data (Ex: Review)
Doc1	The staff is friendly and the joint is always clean! Kudos to the place…
Doc2	Horrible - don't waste your time and money.
Doc3	Great place to have a couple drinks and watch any and all sporting events as the walls are covered with

1) Lowercase
2) Remove Punctuation & Stopwords
3) Stemming
4) Lemmatization

Document ID	Text Data (Ex: Review)
Doc1	staff friend joint always clean kudo place
Doc2	horrible dont waste your time money.
Doc3	great place have couple drink watch all sport events walls covered tv

Create the DTM

Document ID	staff	horrible	sport	place	…	Word N
Doc1	1	0	0	1	…	1
Doc2	0	1	0	0	…	0
Doc3	0	0	1	1	…	0

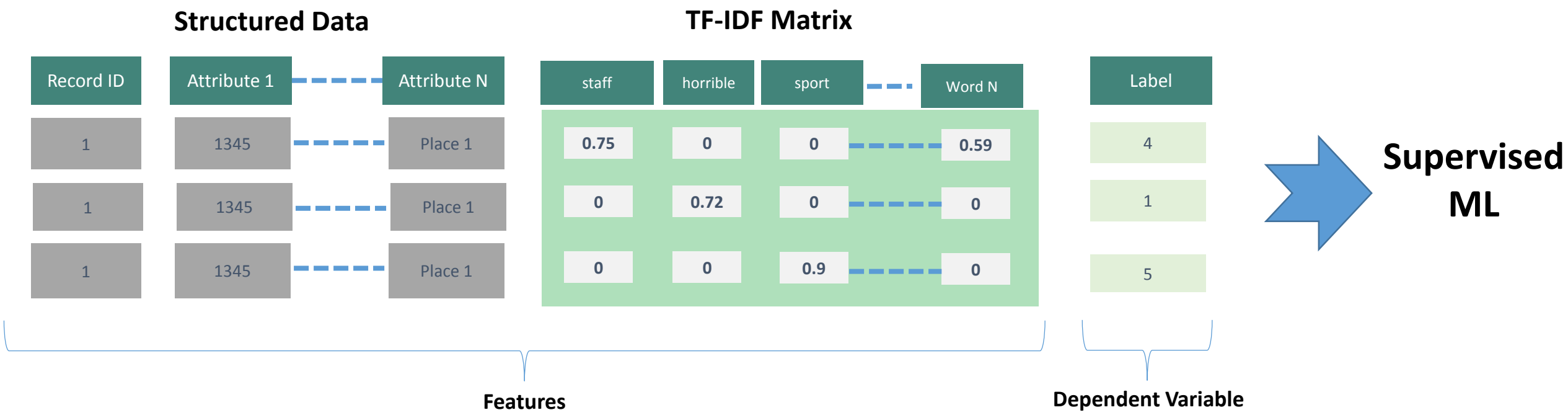
Document Term Matrix (DTM)

TFIDF

Document ID	staff	horrible	sport	…	Word N
Doc1	0.75	0	0	…	0.59
Doc2	0	0.72	0	…	0
Doc3	0	0	0.9	…	0

Document Term Matrix (DTM) weighted by TF-IDF

Action 2: Utilize the Bag of Words Model for Supervised ML



Two Problems with the Bag of Words modelling:

1) Number of words in the corpus can be very high dimensional
2) Till now we are treating words as being independent of one another, which is not a good assumption in the case of text / language

Action 3: Dimensionality Reduction

Document Term Matrix (DTM)

Document ID	staff	horrible	sport	place	...	Word N
Doc1	1	0	0	1	...	1
Doc2	0	1	0	0	...	0
Doc3	0	0	1	1	...	0

Latent Semantic Analysis (LSA)

(Basically SVD on the DTM)

Document ID	Dim 1	Dim 2	Dim 3	...	Dim K
Doc1	0.75	0.39	0.02	...	0.1
Doc2	-0.27	-0.93	0.34	...	0.44
Doc3	0.45	0.78	-0.01	...	0.23

- **K Dimensions where K < N**
- **Dimensions may not be interpretable**

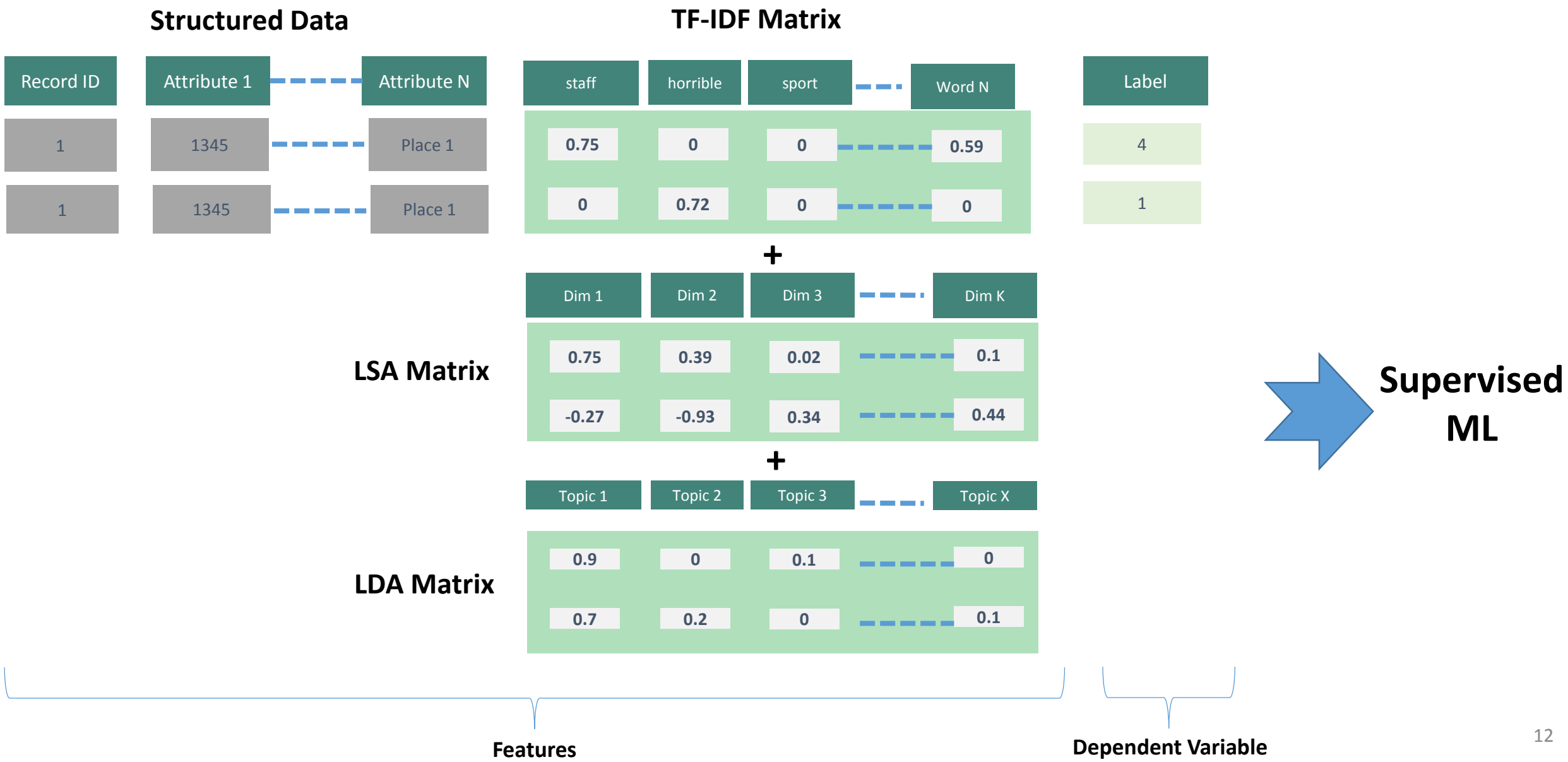
Latent Dirichlet Allocation (LDA)

(Uses Probability Distributions to predict words in topics. Softmax to assign probabilities)

Document ID	Topic 1	Topic 2	Topic 3	...	Topic X
Doc1	0.9	0	0.1	...	0
Doc2	0.7	0.2	0	...	0.1
Doc3	0.5	0.2	0.3	...	0

- **X Topics where X < N**
- **Documents → Topics → Words**
- **Topics can be interpreted by visualizing the collection of words**

Action 4: Utilize All / Some Features for Supervised ML



Action 4: Derive Similarities (among words, documents & topics)

Document ID	Text Data (Ex: Review)
Doc1	staff friend joint always clean kudo place
Doc2	horrible dont waste your time money.
Doc3	great place have couple drink watch all sport events walls covered tv

Continuous sequence of words

staff	friend	joint	always	clean	kudo	place
horrible	dont	waste	your	time	money	
word1	word 2	word 3	word 4			

Word Embeddings
(Word2Vec)

Words	Dim 1	Dim 2	Dim 3	Dim (N)
staff	0.75	0.39	0.02	0.1
friend	-0.27	-0.93	0.34	0.44
joint	0.45	0.78	-0.01	0.23

Salient Points:

1. Each word has a N dimensional vector associated with it. That is each word lives within a N-dimensional space
2. Intuition is that words that are similar / replacement of one another (not synonyms) are close to each other
3. The N-dimensions are not interpretable by a human
4. Word Algebra is possible as it is just addition or subtraction of vectors (Famous example: King – Man + Woman = Queen)
5. We can find similar words both semantically (Land:Run::Water:Swim) and syntactically (walk:walking::run:running)

- Continuous Bag of Words (CBOW) – Predicting the 'word' given the 'context'
- Skip-gram – Predicting the 'context' given the 'word'

Action 5: Derive Similarities (among words, documents & topics)

Doc2Vec – Finding similarities among documents

Document ID	Text Data (Ex: Review)
Doc1	staff friend joint always clean kudo place
Doc2	horrible dont waste your time money.
Doc3	great place have couple drink watch all sport events walls covered tv

Document Vectors
(Doc2Vec)

Docs	D1	D2	D3	----	D(N)
Doc 1	0.75	0.39	0.02	----	0.1
Doc2	-0.27	-0.93	0.34	----	0.44
Doc3	0.45	0.78	-0.01	----	0.23

Helps to find similar documents as each document is represented by a vector

Topic + Word Vectors (lda2Vec) – Illustration with an example

Let's take the word "Germany". Word2Vec model will come up with France, Italy etc. as output based on the 'country' similarity

But if the word 'Germany' occurs in a document which has the topic "Airline" as the predominant topic, then we can add the word vector for "Airline" to word vector of "Germany"

So instead of searching for just 'Germany' in the word2vec space, we can search for 'Germany + Airline' thus getting back "Lufthansa"

Combines LDA with Word2Vec to produce more interpretable output

Some Useful Resources

1. **Patrick's Harrison's talk at PyData 2016** https://www.youtube.com/watch?v=6zm9NC9uRkk

2. **Chris Moody's (Stichfix) talk** https://www.youtube.com/watch?v=vkfXBGnDplQ

3. **DiceTech GitHub** https://github.com/DiceTechJobs/ConceptualSearch

4. **John Savage talk on Kaggle Home Depot competition** https://www.youtube.com/watch?v=_td_w2TpQsw

Leading up to the Social Media Analytics Workshop…

Unstructured Social Media Conversations

Lexical Analysis

Keywords Generation
(through secondary research and Radian6 Sample Extraction)

Data Extraction
(through Radian6)

Pre-processing
(removal of prepositions, stop words, spam words etc.)

N-Grams Extraction
Disintegrating the posts into Unigrams, Bi-Grams and Tri-Grams
(e.g. "energized", "fresh morning" etc.)

Dimension Tagging
("Balance breakfast", "Good health" = Enhance Health)

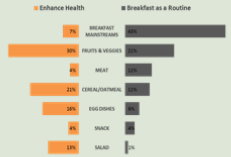
Theme Mapping
(Domain-specific Concepts and Relationships)

Reconciliation
(using keyword mappings)

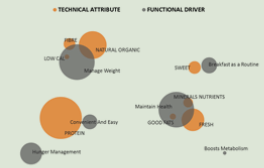
Validation
(manual checks on random samples)

Insights

- Top Drivers
- Top Activities
- Top Consumptions
- Cross Theme Analysis



- Theme Association



- Key Takeaways

Thank You

- Karthikeyan Sankaran, Director, LatentView Analytics
- Email ID – Karthikeyan.Sankaran@latentview.com
- LinkedIn – http://in.linkedin.com/in/karthikeyansankaran
- Twitter – @karthikonbi

What are the dimensions of Analytics?

Business	Use Case Formulation	Interpret Analytics Output	Domain Expertise
Data	Understand datasets	Data Engineering & Architecture	Data Quality & Governance
Math / Quant	Understand the algorithms	Select the right techniques & code	Evaluating the output of algos
Tech / Software	Understand the IT Ecosystem	Build the tech infrastructure	Software Engineering / SDLC