CARPE DATA

Next Generation
Data for Insurance

1st Annual UCSB InsurTech Summit

May 3, 2019

Overview of ML Systems and Products at Carpe Data

Adam P Tashman

Contents

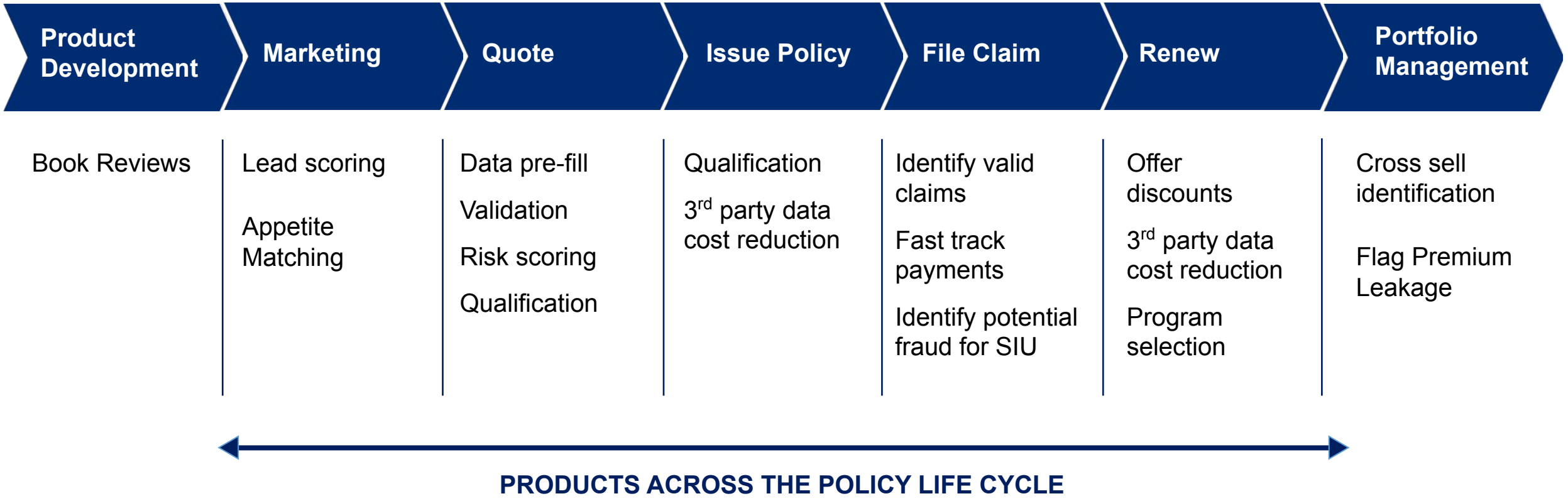
It Started with Background Checks

Mission: Map Emerging Data to Impactful Products

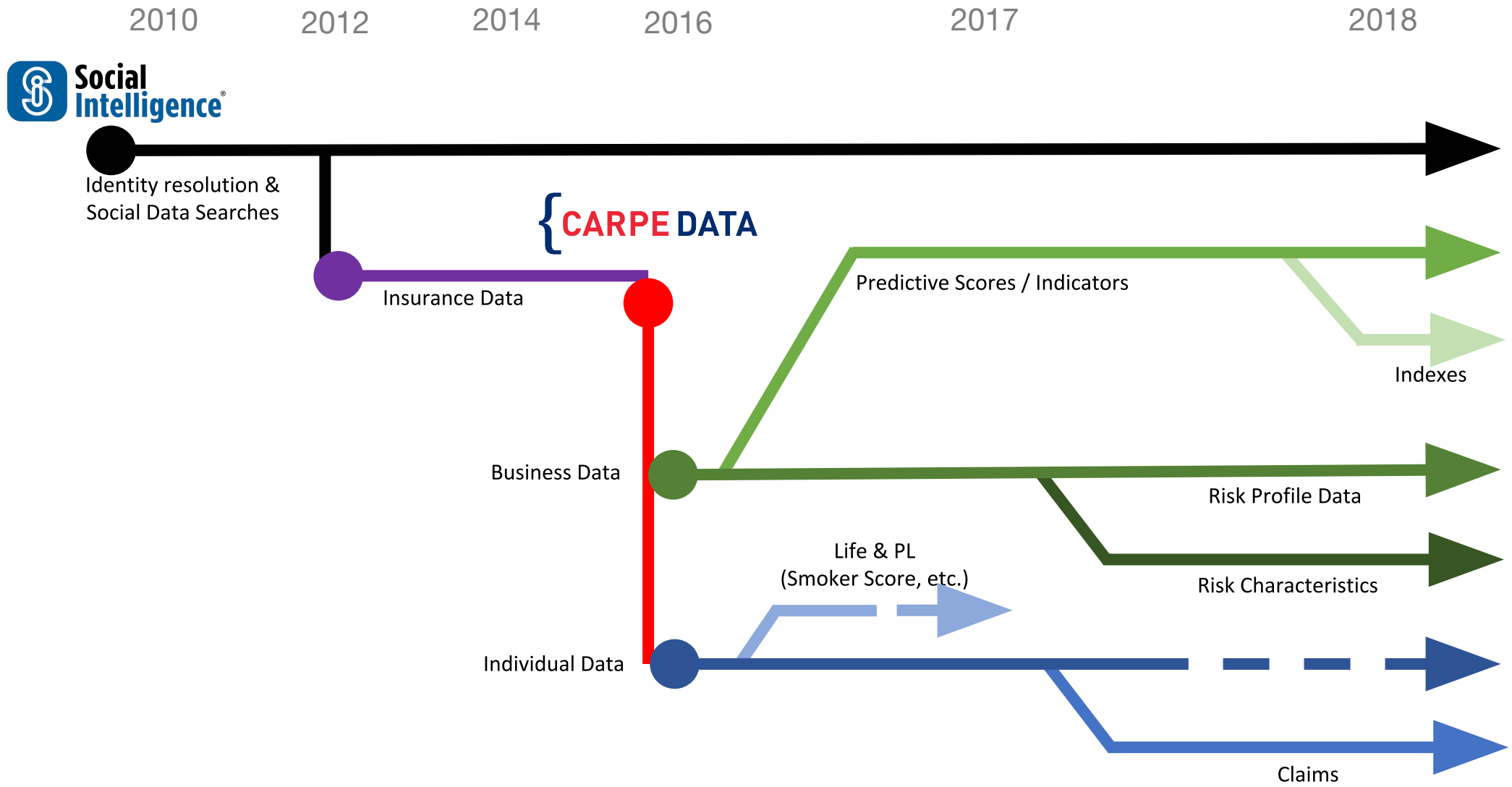
SCORES | INDEXES | RISK CHARACTERISTICS

INSURANCE CARRIER

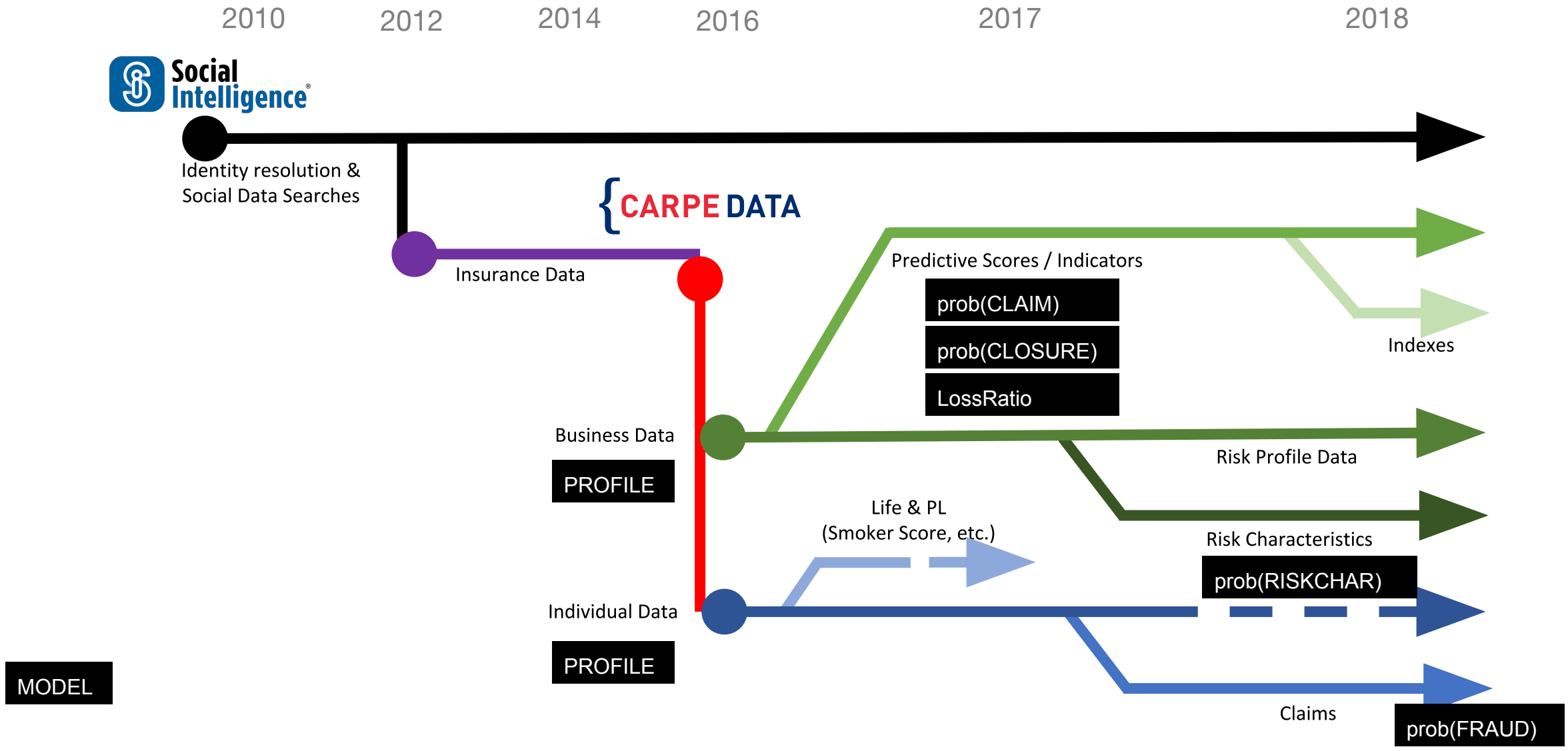
The Insurance Policy Continuum



Capabilities Timeline



Capabilities Timeline



Commercial Data Store: Process Overview

1	Entity Resolution at **business location** level
2	Profile data collection: location, contact information, hours of operation, social data, ...
3	Business type classification (NAICS code)
4a	Given NAICS, collect relevant **risk characteristics** full-service restaurant → {serves alcohol? open late? delivery?}
4b	Risk characteristics inference from various methods: 1) regular expressions 2) text classifiers 3) image classifiers
5	Compute indexes, which are composites of data elements (e.g., customer review index)
6	Compute risk scores (probability of claim, loss ratio, …)

{ Risk Characteristic Extraction

Multiple methods considered for retrieval. Sometimes methods can corroborate (images with captions).

	Regular Expressions	Text Classifier	Image Classifier
RECALL	generally **LOW** - regexes don't learn	**HIGH** given sufficient data	**HIGH** given sufficient data
PRECISION	generally **HIGH** if developer carefully tests	can be **LOW** (high false positives)	can be **LOW** (high false positives)
DATA VOLUME REQUIRED	**LOW** since developed from a priori knowledge	**HIGH**	**HIGH**
DEV TIME	**HIGH** since all patterns must be enumerated	**LOW** (collect data, train, save model)	**HIGH** due to complex models, large set of hyperparams, large set of training data
NOTE			can produce embarrassing errors

Risk Characteristic Extraction

Multiple methods considered for retrieval. Sometimes methods can corroborate (images with captions).

	Regular Expressions	Text Classifier	Image Classifier
RECALL	generally **LOW** - regexes don't learn	**HIGH** given sufficient data	**HIGH** given sufficient data
PRECISION	generally **HIGH** if developer carefully tests	can be **LOW** (high false positives)	can be **LOW** (high false positives)
DATA VOLUME REQUIRED	**LOW** since developed from a priori knowledge	**HIGH**	**HIGH**
DEV TIME	**HIGH** since all patterns must be enumerated	**LOW** (collect data, train, save model)	**HIGH** due to complex models, large set of hyperparams, large set of training data
NOTE			can produce embarrassing errors

Building Better Text Classifiers

1. **Carefully think through each step of preprocessing**
   Use case: Does the business offer 3D printing services?
   Better not strip numeric!

2. **Feature engineering**
   Several ways to quantitatively represent text including n-gram count, n-gram presence, TF-IDF

3. **Word embeddings can be powerful (Word2Vec, GloVe)**
   In essence, words are combined with their context to generate representations useful for semantic and syntactic similarity

   Particularly in web data, the dictionary can be much larger than a proper dictionary (slang, misspellings)
   Hard to control for these characteristics, but word embeddings can help

4. **Class Balancing**
   In particular for rare classes, balancing can offer drastic improvement

Building Better Image Classifiers

1. **Class Selection** When building datasets, include from three areas:

Class	Example
contains label	swimming pool
does not contain label	people, cars, logos, ...
edge cases	beach, lake

2. **Augmentation**
   Include perturbed images in training set (rotations, translations, cropping).
   Be sure these transformations preserve the label

3. **Image Selection**
   Train on “challenging” images. Many Google Images are simplistic with subject against a monochrome background.

4. **Transfer Learning**
   Load a model trained on a massive dataset, repurpose the intermediate features, and add customized layers

5. **Dropout** This is a powerful regularization method which randomly drops neurons during training.

Carpe Data Indexes

A suite of indexes targeting dimensions of risk to insurance carriers

- Targeted low correlation between indexes
- Tuned by segment & geography

Indexes Provide Scalable Data Integration

Our collected data elements are in the thousands, and **rapidly** growing

Counting business presence on websites, we collect **millions** of data elements

Indexes Provide Scalable Data Integration contd.

Given entity, pull from data store

Cafe ABC

Data aggregation*

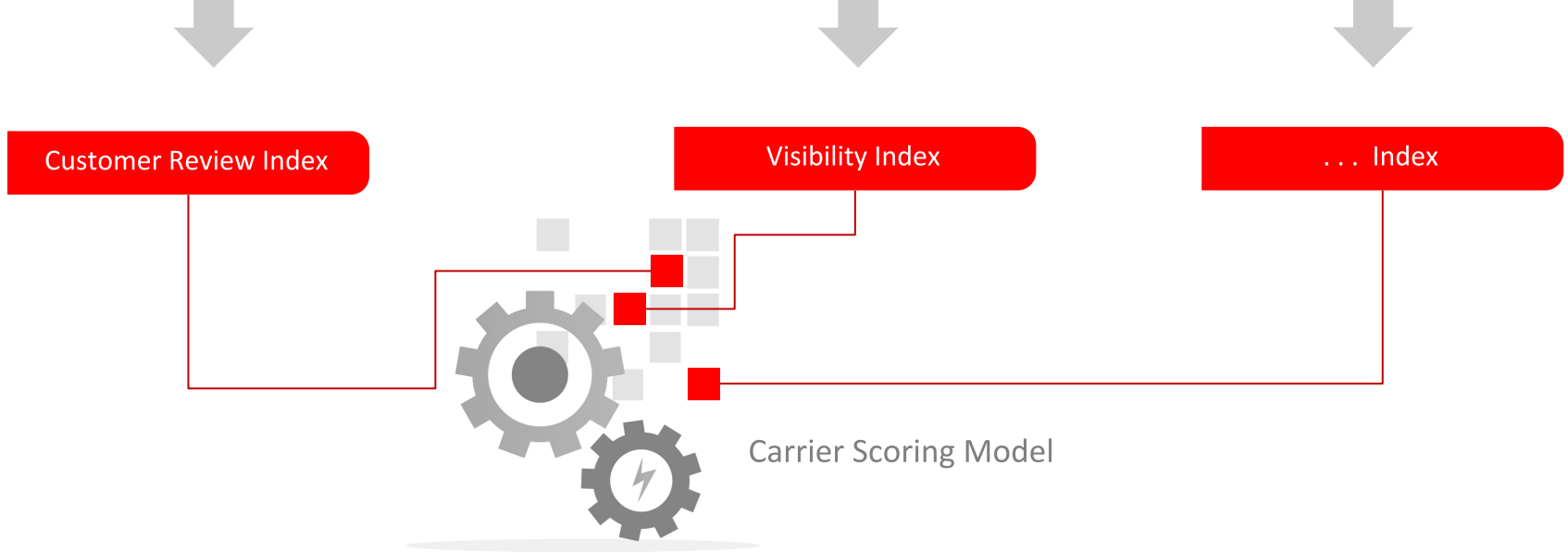


*Illustration purposes only

{ Q&A

{ Thank you!