

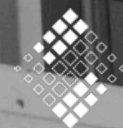
Building the World's Largest Database of Car Features from PDFs

John Akred

Chief Technology Officer
Silicon Valley Data Science

Robert Munro

Chief Executive Officer
Idibon



SILICON VALLEY
DATA SCIENCE

iDIBON



JOHN AKRED

@BigDataAnalysis

Founder & CTO

Silicon Valley Data Science

Consulting firm of elite data science and engineering teams who specialize in data-driven product development and business transformation.



Robert Munro

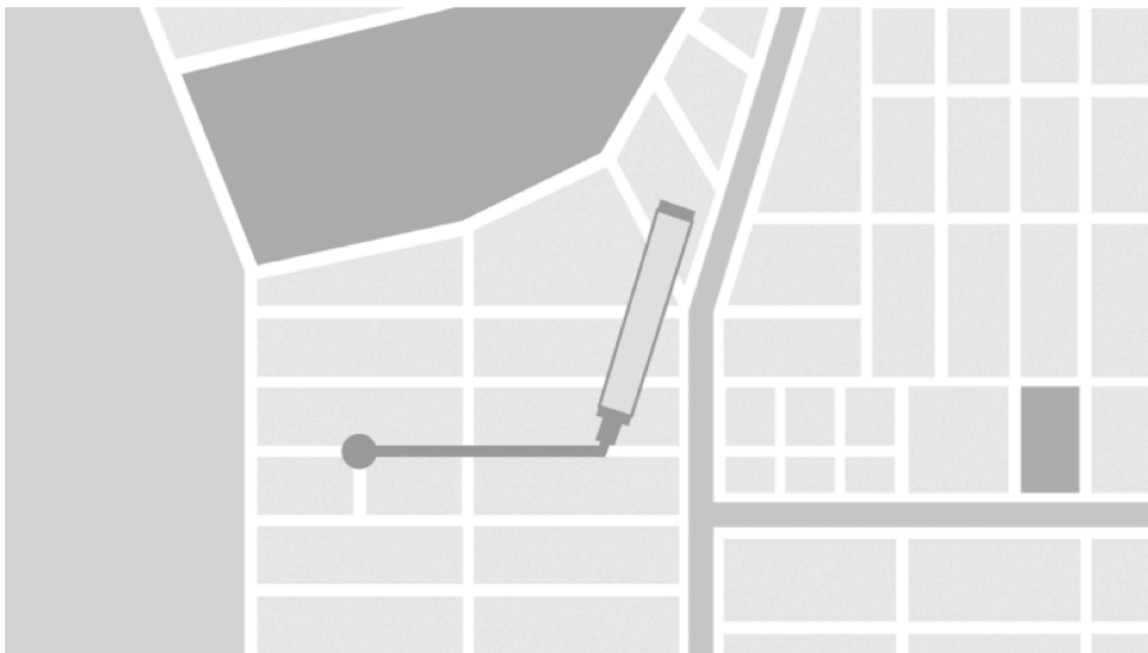
@WWRob

Founder & CEO
Idibon

*Developers of cloud-based
natural language
processing services that
adapt to business-specific
problems, in any language.*

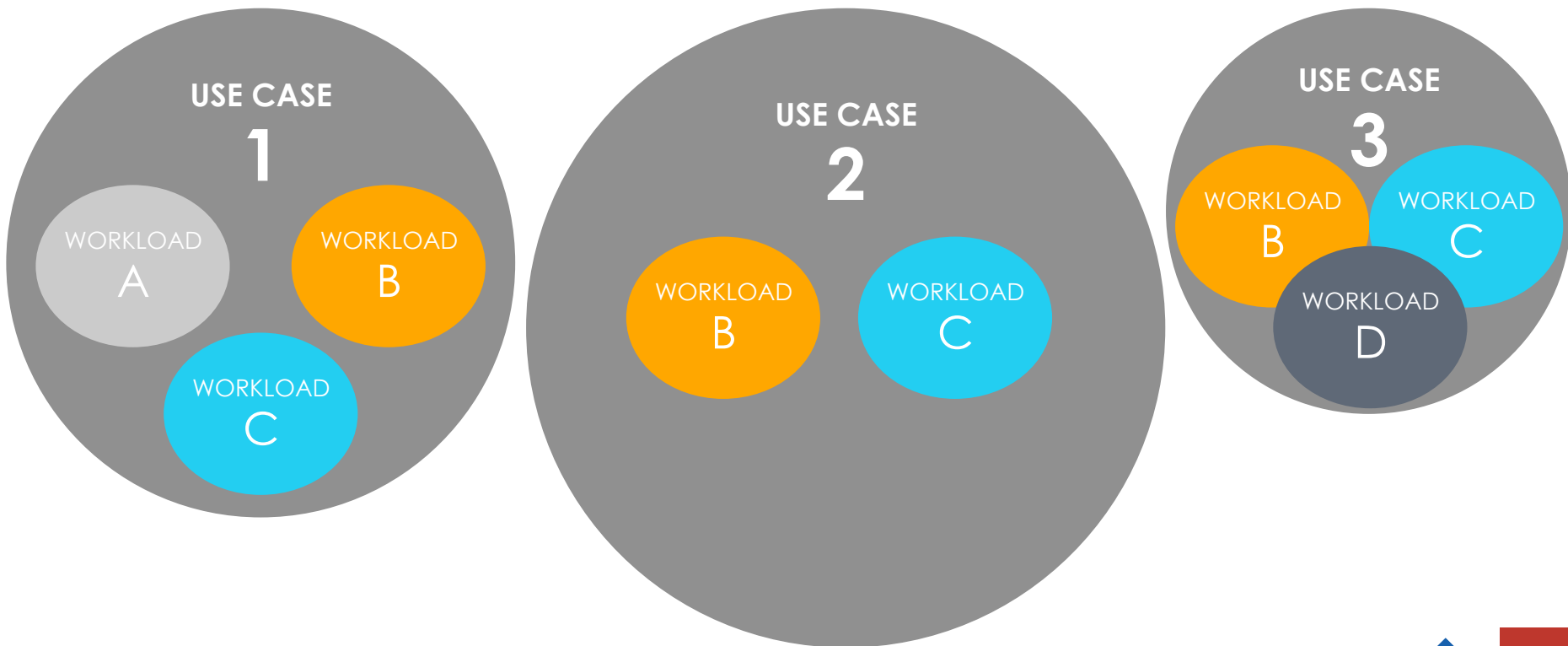
DATA STRATEGY

*and choosing projects for
maximum impact*



DEFINE YOUR ROADMAP

IDENTIFY STRATEGIC WORKLOADS





PRIORITIES



DIMENSIONS



ASSUMPTIONS
(overcome them)



LATHER, RINSE, REPEAT

Putting this into action for
EDMUNDS.com



edmunds.com Price Promise™ Get upfront pricing Make ▼ Model ▼ Year ▼ Car Type ▼ Used Cars Car Research ▼ Live Help Search

Pricing **Reviews** [View Incentives](#)

3 Most popular configurations in your area
2015 Mercedes-Benz C-Class C250 2dr Coupe (1.8L 4cyl Turbo 7A)

C250 2dr Coupe (1.8L 4cyl Turbo 7A)	C250 2dr Coupe (1.8L 4cyl Turbo 7A)	C250 2dr Coupe (1.8L 4cyl Turbo 7A)
18 Cars Nearby	18 Cars Nearby	14 Cars Nearby
Options and Packages <ul style="list-style-type: none"> Rear Trunk Lid Spoiler Multimedia Package info Wheel Locks Lighting Package info Blind Spot Assist info Appearance Package info See Standard Features Exterior colors available:	Options and Packages <ul style="list-style-type: none"> Becker MAP PILOT Pre-Wiring See Standard Features Exterior colors available:	Options and Packages <ul style="list-style-type: none"> Rear Trunk Lid Spoiler Multimedia Package info Keyless-go info See Standard Features Exterior colors available:
Estimated MSRP \$47,605 Compare Dealer Prices	Estimated MSRP \$40,850 Compare Dealer Prices	Estimated MSRP \$44,200 Compare Dealer Prices

ADVERTISEMENT
 The all-new 2015 C-Class
 Starting at \$40,400*
[VIEW INVENTORY](#) [EXPLORE](#) Mercedes-Benz

Other C-Class Years
[2014 Mercedes-Benz C-Class](#)
[2013 Mercedes-Benz C-Class](#)
[2012 Mercedes-Benz C-Class](#)

ADVERTISEMENT

Existing revenue streams:

- Ads
- Price quotes (leads)

Shopping is the focus:

- Need real-time inventory
- Accurately described VIN's

THE PDFs

[illegible]

THE CHALLENGE

- Couldn't keep pace with original equipment manufacturers (OEMs)
- Approach was largely manual, and backlogs would develop
- ~6.5% of VIN's being held back
- Content operations team was a silo



THE QUESTION

What is the
minimum viable data
required to get a VIN live?

(Then come back to add
features and specs.)



A FRESH LOOK

A black and white photograph of a man in a suit and hat standing on a rocky outcrop, looking out over a forested valley. The man is positioned on the left side of the frame, facing right. The background shows a dense forest of trees under a cloudy sky.

- Worked with Product Owners to define shell products (aka *minimally viable data*)
- Leveraged NLP to automate OEM data translations
- Fully integrated NLP workflow into existing tooling

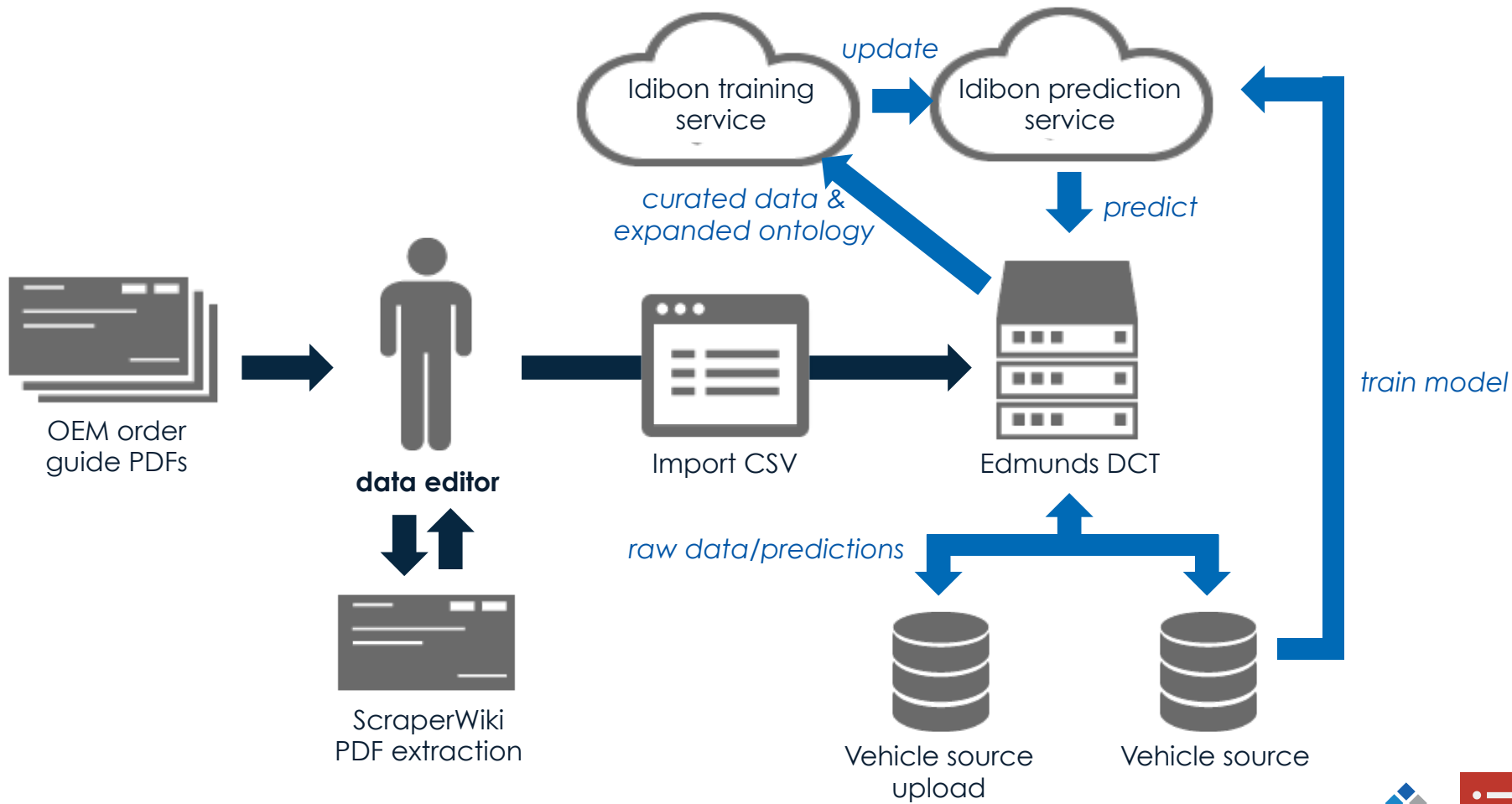


Capabilities **have** to be integrated into the business process.

Ultimately, you may be able to make predictions, but putting those into action is the real thing.

TRAINING HUMANS IS TRICKY





Data Entry Tool (ver. 2.3.96)

Product Tree

Product Type Editor

Issues

All Predictions

Partner translation

Product Tree

Attributes



Save Expand All Collapse All

- 2011 #200000082
- 2012 #200005741
- 2013 #200374079
- 2014 #200389127
- 2015 #200481161
 - 2 Series #200676054
 - 3 Series #200704009
 - 3 Series Gran Turismo #200706926
 - 4 Series #200704560
 - 4 Series Gran Coupe #200676095
 - 5 Series #200705185
 - 528i 4dr Sedan (2.0L 4cyl Turbo 8A
 - 528i xDrive 4dr Sedan AWD (2.0L 4
 - 535d 4dr Sedan (3.0L 6cyl Turbodie
 - 535d xDrive 4dr Sedan AWD (3.0L 6
 - 535i 4dr Sedan (3.0L 6cyl Turbo 8A
 - 535i xDrive 4dr Sedan AWD (3.0L 6
 - 550i 4dr Sedan (4.4L 8cyl Turbo 8A
 - 550i xDrive 4dr Sedan AWD (4.4L 8
 - Engines
 - Transmissions

- New Product
- Copy Product
- Cut Product
- Paste Product
- Remove Product
- Import From CSV
- Resurrect Product
- Product Editor
- Color Bindings
- Edit Rules
- Changes History

Attribute



CSV Layout config

Save Choose Vehicle Trim Choose Descriptions Column Choose Availability Symbol Products: 528i 4dr Sedan (2.0L 4cyl Turbo 8A)

Page 0

0	1	2	3	4
Name: [8220]Table 1 of 1 on page	Name: [8220]Table 1 of 1 on page	Name: [8220]Table 1 of 1 on page	Name: [8220]Table 1 of 1 on page	Name: [8220]Table 1 of 1 on page
Table: 5	Table: 5	Table: 5	Table: 5	Table: 5
5 Series [8211] 528i/xDrive, 550i/xDrive	Standard equipment Optional equipment	Standard equipment Optional equipment	Standard equipment Optional equipment	Standard equipment Optional equipment
Comfort and convenience	528i	x5D2ri8vie	550i	x5D5ri0vie
Engine Start/Stop button with Keyless	•	•	•	•
2-way power moonroof with remote	•	•	•	•
interior sunshade				
Power windows with key-off and [8220]	•	•	•	•
opening from remote, and closing from				
Power trunk lid opening and closing	o	o	o	o
2-zone automatic climate control for	•	•	•	•
Micro-filter ventilation system with re	•	•	•	•
Automatic tilt-down of passenger[8220]	•	•	•	•
Park Distance Control (front and rear)	o	o	•	•
Parking Assistant (requires Park Di	o		o	
Automatic high beams	o	o	o	o
Heated steering wheel	o	o	o	o
Comfort Access keyless entry with	o	o	o	o



CSV Layout config



Save



Choose Vehicle Trim



Choose Descriptions Column



Choose Availability Symbol

Products: Karim's test feature



Page 0

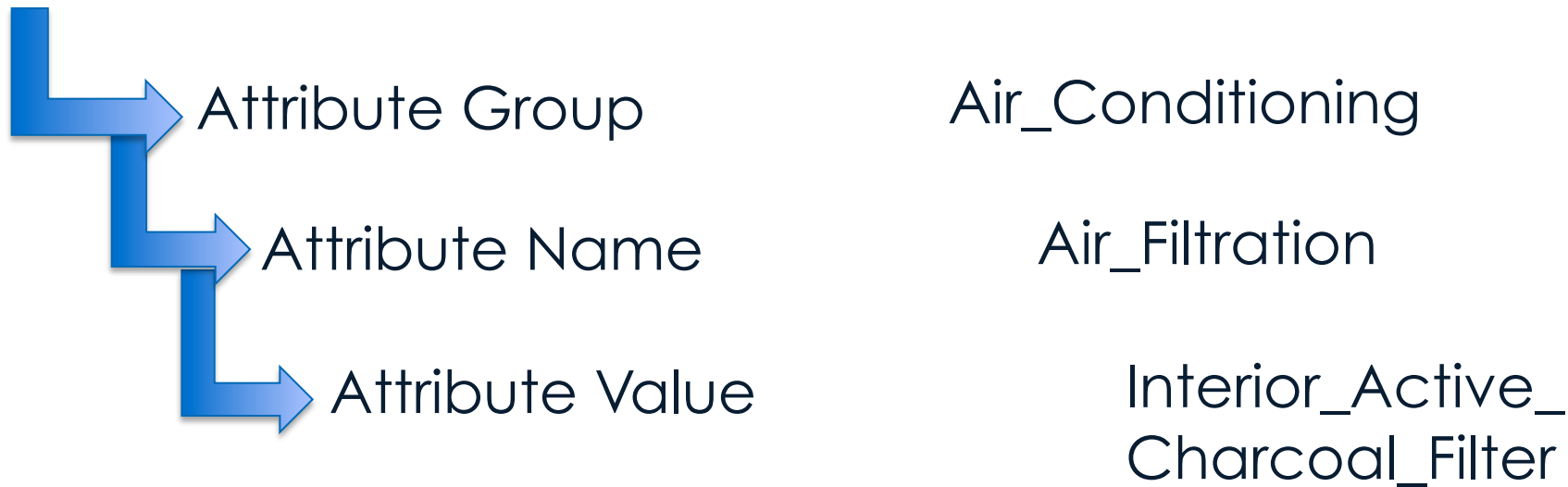
0	1	2	3	4
Name: [8220]Table 1 of 1 on page :	Name: [8220]Table 1 of 1 on page :	Name: [8220]Table 1 of 1 on page :	Name: [8220]Table 1 of 1 on page :	Name: [8220]Table 1 of 1 on page :
Table: 5	Table: 5	Table: 5	Table: 5	Table: 5
5 Series [8211] 528i/xDrive, 550i/xD	Standard equipment Optional equ	Standard equipment Optional equ	Standard equipment Optional equ	Standard equipment Optional equ
Comfort and convenience	528i	x5D2ri8vie	550i	x5D5ri0vie
Engine Start/Stop button with Keyle	•	•	•	•
2-way power moonroof with remote	•	•	•	•
interior sunshade				
Power windows with key-off and [8	•	•	•	•
opening from remote, and closing fr				
Power trunk lid opening and closing	○	○	○	○
2-zone automatic climate control fo	•	•	•	•
Micro-filter ventilation system with r	•	•	•	•
Automatic tilt-down of passenger[8	•	•	•	•
Park Distance Control (front and re	○	○	•	•
Parking Assistant (requires Park Di	○		○	
Automatic high beams	○	○	○	○



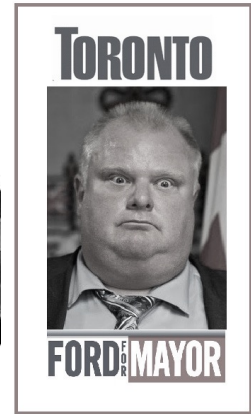
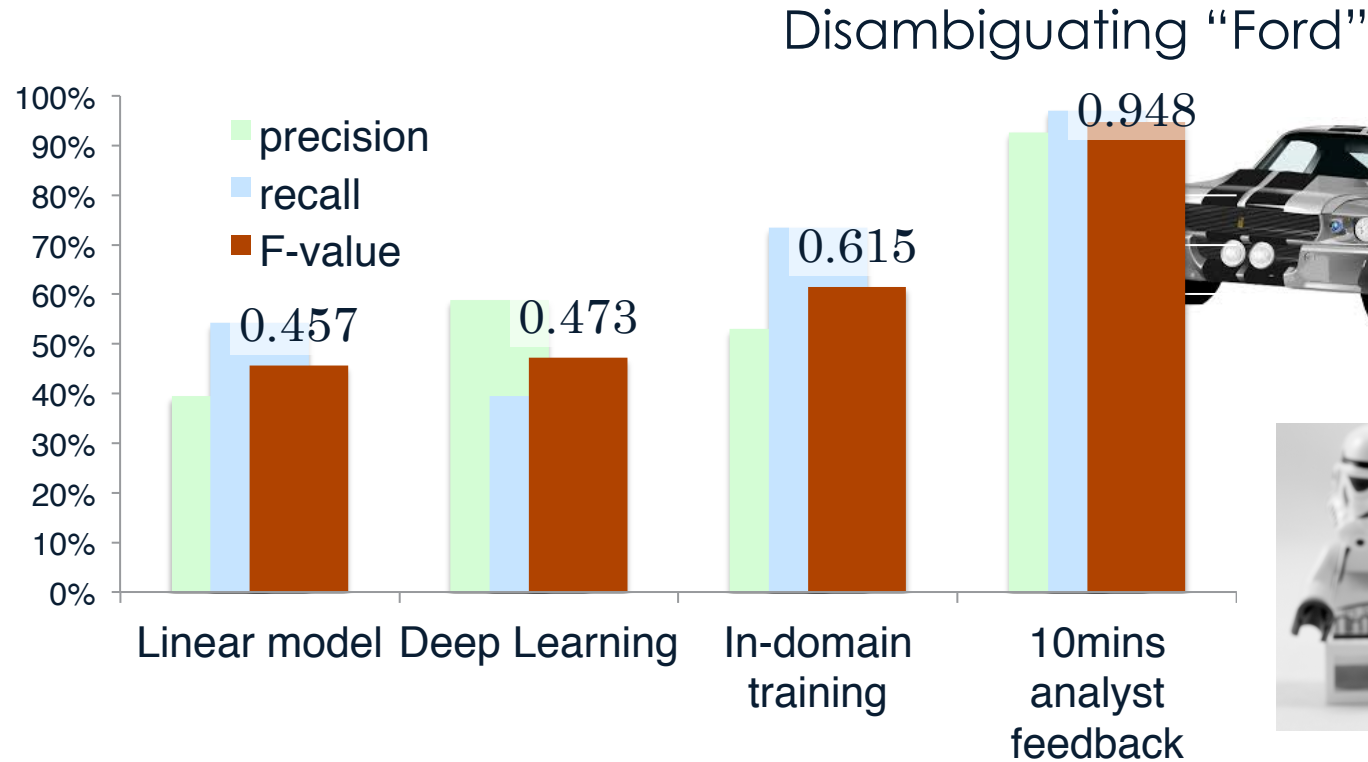
Product Tree Product Type Editor Issues All Predictions Partner translation Label Product Karim's test feature						
Save Submit						
#	Source Text	Attribute Group	Attribute Name	Predicted Value	Correct Prediction	Score
1	Engine Start/Stop button with Keyless-go feature				Add Prediction	
2	Power windows with key-off and [8220]one touch[8221] up and down operation both front and rear, anti-trapping feature, opening from remote, and closing from exterior lock				Add Prediction	
3	2-zone automatic climate control for driver and front passenger	Front Passenger Seat	Cooled Passenger Seat	passenger	<input type="checkbox"/>	0.82006632187...
					Add Prediction	
4	Micro-filter ventilation system with replaceable active-charcoal filters				Add Prediction	
5	Automatic-dimming interior rear-view mirror and exterior side-view mirrors	Misc. Exterior Features	Exterior Camera	rear and side view camera	<input type="checkbox"/>	0.40087529414...
					Add Prediction	
6	Ambiance lighting				Add Prediction	
7	Interior courtesy lights with automatic dimming function				Add Prediction	
8	Dual front sun visors with illuminated mirrors	Trailer Towing Equipment	Tow Hooks	front	<input type="checkbox"/>	0.17822636533...
					Add Prediction	
9	Dual cupholders in front and rear				Add Prediction	
10	Automatic headlight on/off control				Add Prediction	
11	Power outlet in front passenger[8217]s footwell, front center console storage compartment, rear center console, and in trunk	Airbags	Head Airbags	front and rear	<input type="checkbox"/>	0.57523750852...
		Storage	Cupholders Location	front and rear	<input type="checkbox"/>	0.12471087869...
					Add Prediction	
12	Additional 12-V power outlets	Driver Seat	Driver Seat Easy Entry	power driver seat	<input type="checkbox"/>	0.46132767690...
		Mirrors	Exterior Extending Side mir...	power	<input type="checkbox"/>	0.10462681315...
					Add Prediction	

Hierarchical Entity Resolution

'Micro-filter ventilation system with replaceable active-charcoal filters'



Data beats algorithms; feedback beats data



Edmunds ontology

'Micro-filter ventilation system with replaceable active-charcoal filters'

Air_Conditioning (depth 1)

Air_Filtration (depth 2)

Interior_Active_Charcoal_Filter (depth 3)

...

Climate_Control_Memory

...

Front_Air_Conditioning

Front Zone

Rear_Air_Conditioning

Rear Zones

...

4000 (!) car features

20,000 labeled items to date

- requires expertise; accuracy with few data points

Features:

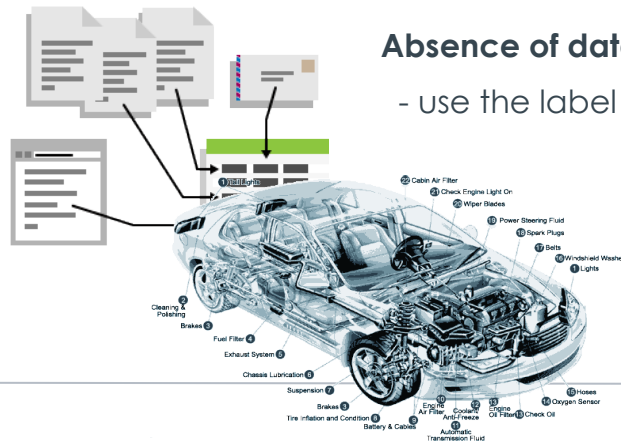
- words, n-grams, word shapes, taxonomies

Negative sampling:

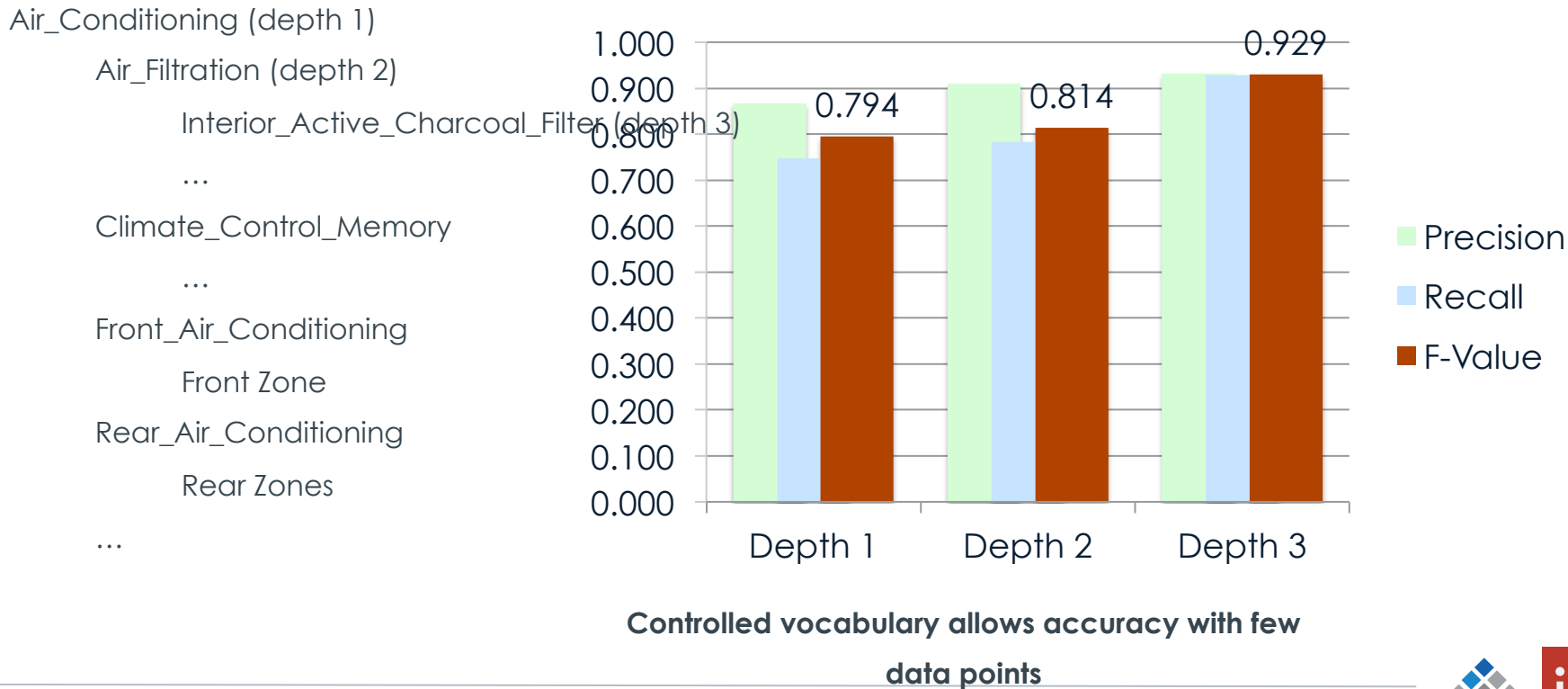
- other items within the same parent

Absence of data:

- use the label name as the first data item



Accuracy at depths in the ontology



- 1–2 days to get a new Car Model live online vs. 2 weeks (~85% reduction)
- 95% reduction in backlog
- API for making predictions on unstructured vehicle data using Edmunds' ontology assets

NEW CAPABILITIES

- Use data to highlight the real issues
- Change is difficult, agile methods help users embrace new approaches
- New processes must adapt to workflows

TAKEAWAYS

working with

UNSTRUCTURED DATA

Entity Resolution

Recognizing references to objects:
"BMW 435" = 2014 BMW 4 Series
435i xDrive 2dr Coupe AWD (3.0L
6cyl Turbo 8A)

Concept Discovery

AWD = 4WD = xDrive = quattro
Manual transmission = stick shift

What Others Are Saying

Consumer Reviews for the BMW 4 Series

★★★★★ [See all 7 reviews](#)

1 of 1 people found this review helpful

★★★★★

Try the 428i

by **professorx** on Sep 3, 2014

Vehicle: 2014 BMW 4 Series 428i SULEV 2dr Coupe (2.0L 4cyl Turbo 8A)

Picked up my 428i yesterday. Was worried that the 4 cyl wouldn't be enough and that I would not be

Sentiment Analysis

Positive comment
referred to
exterior design of
competitor's
model

8 of 22 people found this review helpful

★★★★★

I (almost) love everything about

by **jrbskisummit** on Apr 13, 2014

Vehicle: 2014 BMW 4 Series 435i xDrive 2dr Coupe AWD (3.0L 6cyl Turbo 8A)

As far as I could tell there were only two cars that met my desires for a coupe with AWD, Manual transmission and forced induction for Colorado altitudes. The BMW 435 ended up being my choice. The other choice was the A5/S5. The 428 would have been a competitor for the A5 if it was available with manual. Anyway I drove a S5 and compared it with the 435. I preferred the S5 looks but engine dynamics are noticeably better in the 435 (if the sport mode is active). My car has a manual, so sport mode doesn't fool with the transmission shift points as it does in the Automatic cars. As expensive as it is the 435 seems a better value than the S5. I wish it was all wonderful but it could be better.





THANK YOU

John Akred

john@svds.com

Robert Munro

rob@idibon.com

Want these slides? Go to:

TO ADD

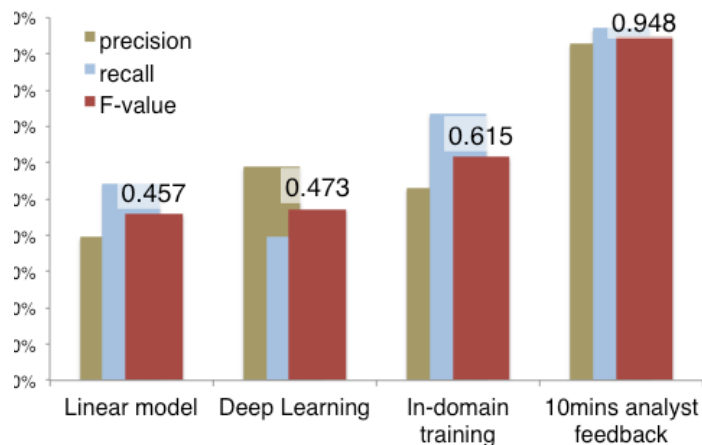




How do we get the analysts' feedback?

How many ways could we *ask* someone to distinguish the right “Ford”?

Get it right, and save 90% of the labor \approx 90% of total cost.



Distinguishing “Ford” the company from people called “Ford”



Super-user, guided by model suggestions

idibon MUC7_partial Signout

Read only

A new "mother ship" to carry the Woods Hole Oceanographic Institute's manned submersible, Alvin, and deep-diving robot explorers is scheduled to be launched Thursday in Moss Point, Miss., where the \$50 million state-of-the-art ship was built.

The 274-foot research vessel Atlantis will replace Atlantis II, a smaller ship being retired after 33 years of service for the institute and the US Navy.

When work on its systems is completed in 1997, the new Atlantis will be among the most sophisticated oceanographic research ships in the world. It has four times the on-board laboratory space of Atlantis II, can carry more scientists, and has twice the endurance at sea.

"This is a great event for us," said Richard Pittenger, associate director of marine operations for the institute, in a telephone interview Wednesday. "We just took a tour, and the ship is absolutely magnificent," he said.

Following the traditional christening with champagne, shipyard workers plan to send the 274-foot research vessel down greased skids in a spectacular sideways launch, splashing it into a bayou where it

Spans

NER

LOCATION

X Miss.

Goto document list

Annotation filters



Targeting one label at a time

idibon MUC7_v5 Signout

Work Unit #1

UPPER MARLBORO, Md. &MD; The Navy said here on Thursday that it had shipped equipment to the Dominican Republic to help find the flight data recorder of the Boeing 757 that crashed offshore on Tuesday night, killing all 189 people on board.

Navy officials said they had been asked to locate the so-called black box, but so far have not been asked to help retrieve it or any other parts of the wreckage. In the meantime, discussions continued about who would pay to salvage parts of the jet.

The Turkish-owned charter jet, which was leased to a Dominican airline, was carrying German tourists home from a Caribbean vacation when it plunged into the Atlantic shortly after takeoff from Puerto Plata in the Dominican Republic.

Late on Thursday afternoon in Puerto Plata, the Coast Guard said it had given up its search, and it remained uncertain exactly how many bodies had been accounted for.

Dominican officials said 128 bodies had been recovered. A Pentagon spokesman said local people in boats appeared to have stripped some bodies of money and identification.

Highlight every location name in the document.
(More)

Comments:

I'm done



Simple accept/reject annotation for fast annotation

idibon

Car_Pricing_Tweets

Signout

Work Unit #1

Dodge Ram 1500 Used for Sale Whittier Ca 90605: Dodge Ram 1500 Hemi Sport Used Price: \$13,500... <http://t.co/aNfgsGypIk>

Please judge if the text has been properly classified.

Style_Pricing: Make

Accept

Reject

