SPARK SUMMIT 2017

USING AI FOR PROVIDING INSIGHTS AND RECOMMENDATIONS ON ACTIVITY DATA

Alexis Roos, @alexisroos

Sammy Nammari

Agenda

- **Salesforce introduction**
- Inbox and email data
- Pricing request classifier pipeline
  - Labeling
  - Feature generation
  - Scoring

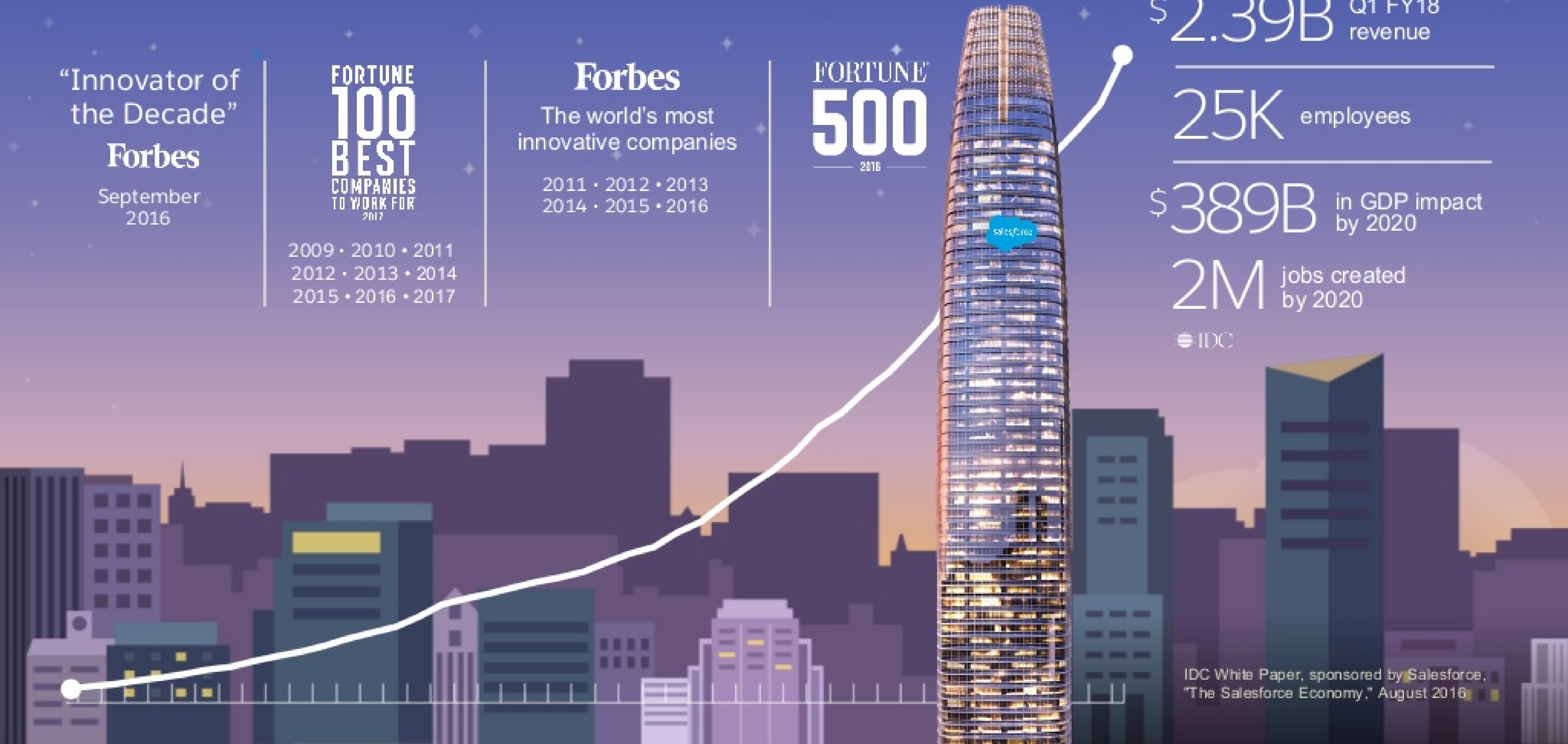
Together, We're Building a Path Forward

"Innovator of the Decade"
Forbes
September 2016

FORTUNE 100 BEST COMPANIES TO WORK FOR 2017
2009 • 2010 • 2011
2012 • 2013 • 2014
2015 • 2016 • 2017

Forbes
The world's most innovative companies
2011 • 2012 • 2013
2014 • 2015 • 2016

FORTUNE 500
2016



$2.39B Q1 FY18 revenue

25K employees

$389B in GDP impact by 2020

2M jobs created by 2020

IDC

IDC White Paper, sponsored by Salesforce, "The Salesforce Economy," August 2016

Agenda

- Salesforce introduction
- **Inbox and email data**
- Pricing request classifier pipeline
    - Labeling
    - Feature generation
    - Scoring

Salesforce Inbox Demo

What sorts of emails do salespeople receive?

- Emails from customers
  - Meeting requests, pricing requests, competitor mentioned, *etc*.
- Emails from coworkers
- Marketing emails
- Newsletters
- Telecom, Spotify, iTunes, Amazon purchases
- *Etc*

Pricing requests

We want to identify **pricing requests** from customers

Hey Ascander,

How much would it cost to add ten seats to the plan?

Thanks,
Gabe

Hello Eddie,

Can you send me that really important document?

Thanks,
Alexis

Welcome to Spotify!

Your new subscription is active.

Enjoy the music.

Agenda

- Salesforce introduction
- Inbox and email data
- Pricing request classifier pipeline
    - Labeling
    - Feature generation
    - Scoring

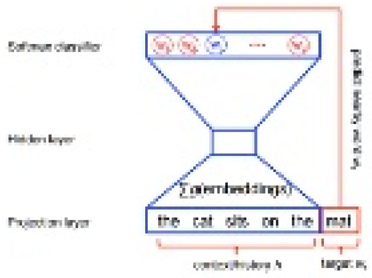
Data labeling pipeline

databricks

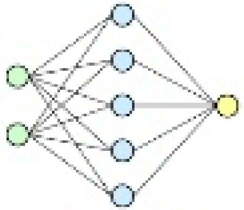
Spark

Emails

Filtering / Sampling

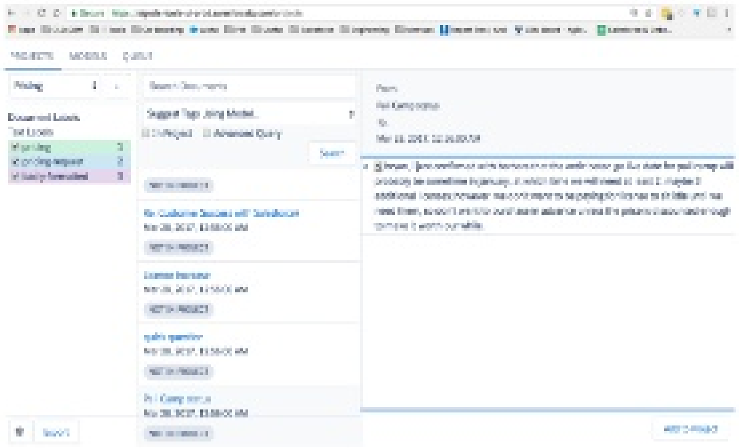


Word2Vec



GraphX

Labeling tool



Labeled Training Data

Data used

Billions of emails that we process over time

~2.5 million internal emails that we have anonymized and have explicit permission to label

Structure of an email

Intro	Hey Alexis,
Body	Let's meet with Ascander on Friday to discuss the $10,000/year rate. Ascander's phone number is (123) 456-7890.
Signature	Thanks, Noah Bergman Engineer at Salesforce (123) 456-7890
Confidentiality Notice	The contents of this email and any attachments are confidential and are intended solely for addressee…
Reply Chain	From: Alexis alexis@salesforce.com Date: April 1, 2017 Subject: Important Document Noah, how much does your product cost?

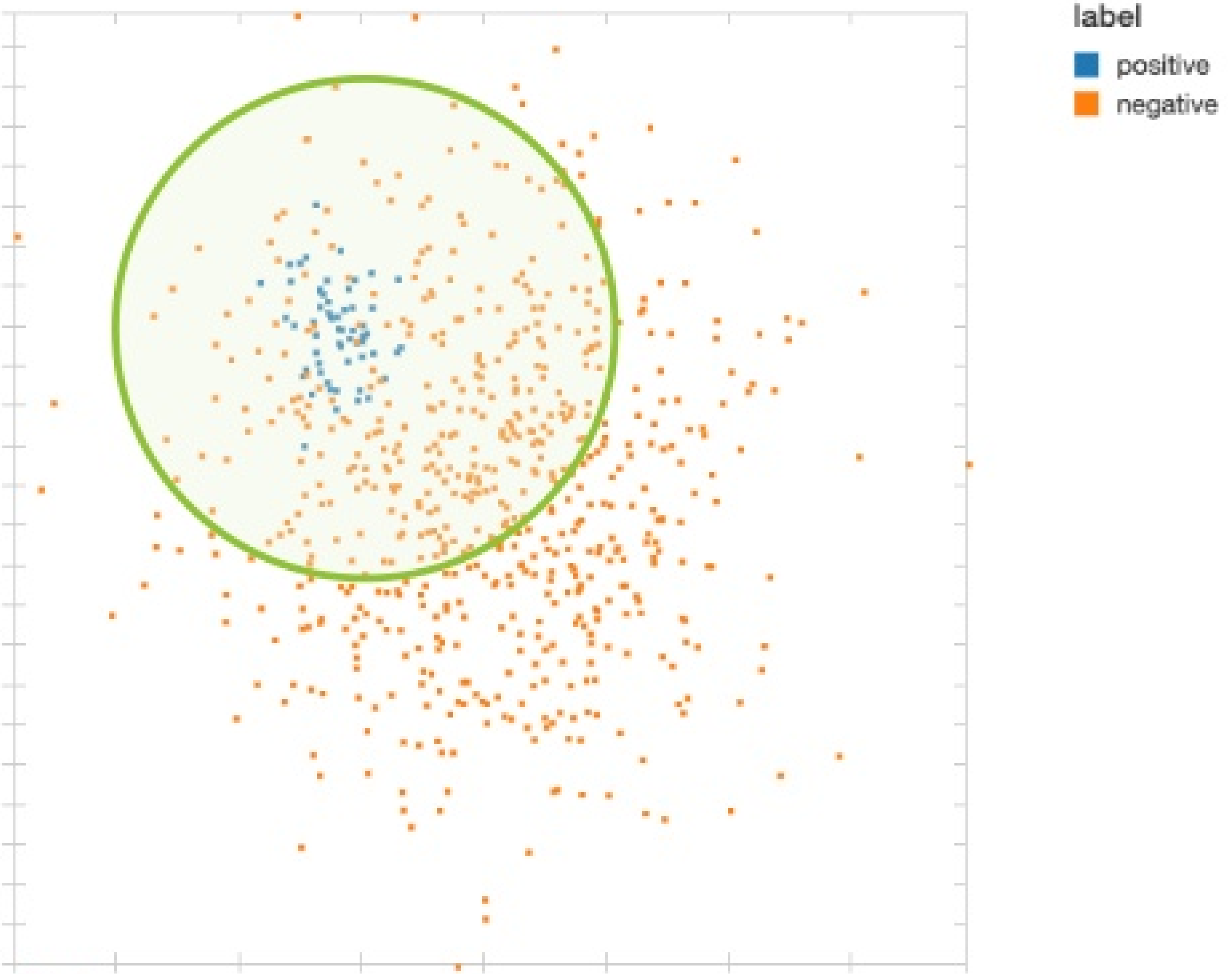
Labeling data

- No labels, and currently no mechanism to infer labels
- Pricing requests are very important, but relatively rare events
- Emails are sensitive — can’t mechanical turk

Hand-labeling impractical

Labeling data – high-recall filter

How can we get a higher yield of positive labels when labeling by hand?



Labeling data – high-recall filter

How do we build this green circle?

- Relationship graph (GraphX)
- Word2Vec

Labeling data – Word2Vec

What would be the total **cost** of a …

How much would it **cost** to add ten seats to the plan?

Does it **cost** a lot of money to …

Neural network that finds words similar to **cost** based on the context that it appears in

Labeling data – Word2Vec

- Train Word2Vec on unlabeled emails
- find words close in distance to “price”, “cost”, “license”, etc

Things we calculated after we got labels

Performance of this filter

- Our original dataset was **0.17%** positive labels

- Graph + Word2Vec reduced our dataset to **2%** of its original size, and increased the positive label rate to **11.2%**, with a recall of **0.93**

We've introduced some bias, but hand-labeling is now tractable!

Improving the output produced by Word2Vec

INTRO	Hey Alexis,
BODY	Let's meet with Ascander on Friday to discuss the $10,000/year rate. Ascander's phone number is (123) 456-7890.
SIGNATURE	Thanks, Noah Bergman Engineer at Salesforce (123) 456-7890
CONFIDENTIALITY NOTICE	The contents of this email and any attachments are confidential and are intended solely for addressee…
REPLY CHAIN	From: Alexis alexis@salesforce.com Date: April 1, 2017 Subject: Important Document Noah, how much does your product cost?

Improving the output produced by Word2Vec

"Let's meet with Ascander on Friday to discuss the $10,000/year rate. Ascander's phone number is (123) 456-7890."

Names, monetary values and phone numbers are noisy

Improving the output produced by Word2Vec

```
word2VecModel.findSynonyms(“cost”, 5)
```

```
$10
price
$85/month
$19.99
$15,000/year
```

Improving the output produced by Word2Vec

"Let's meet with Ascander on Friday to discuss the $10,000/year rate.
Ascander's phone number is (123) 456-7890."

Improving the output produced by Word2Vec

“Let’s meet with NAME on Friday to discuss the MONEY rate. NAME phone number is PHONE_NUMBER.”

Improving the output produced by Word2Vec

```
word2VecModel.findSynonyms("cost", 5)
```

```
MONEY
price
license
nominal
budget
```

Interleaving ngrams with unigrams

```
interleaveNGrams("hello my name is sammy", 2)
```

produces:

```
“hello hello-my my my-name name name-is is is-sammy sammy”
```

Improving the output produced by Word2Vec

```
word2VecModel.findSynonyms(“cost”, 5)
```

```
MONEY-per-month
price-of
license
month-to-month
price
```

Agenda

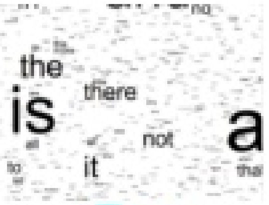
- Salesforce introduction
- Inbox and email data
- Pricing request classifier pipeline
  - Labeling
  - Features generation
  - Scoring

Generating feature vectors and model training

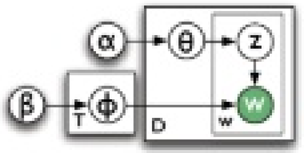
Emails

Labeled Training Data

Feature Engineering

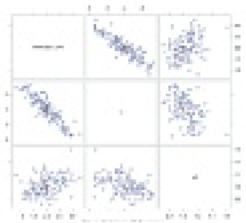


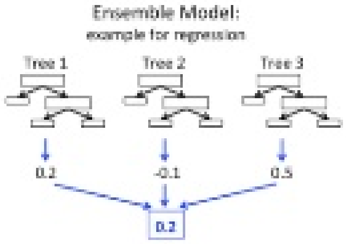
Text Processing / TF-IDF



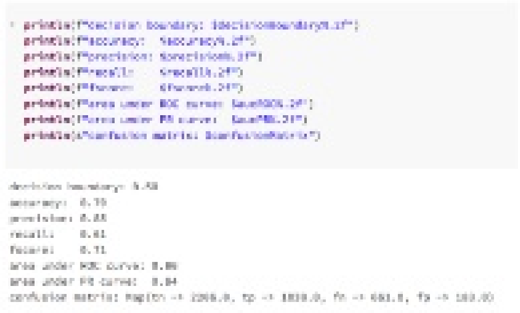
LDA

Model Training



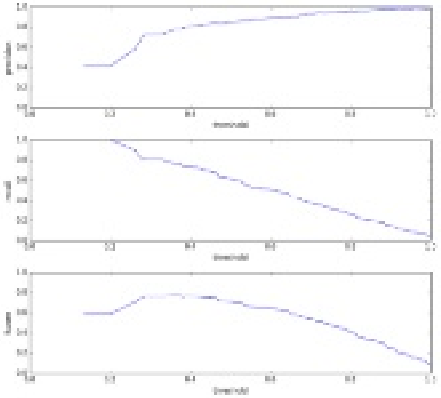


Model Evaluation



```
println(f"decision boundary: $decisionBoundary%.2f")
println(f"accuracy:  $accuracy%.2f")
println(f"precision: $precision%.2f")
println(f"recall:    $recall%.2f")
println(f"fscore:    $fscore%.2f")
println(f"area under ROC curve: $auROC%.2f")
println(f"area under PR curve:  $auPRC%.2f")
println(s"confusion matrix: $confusionMatrix")
```

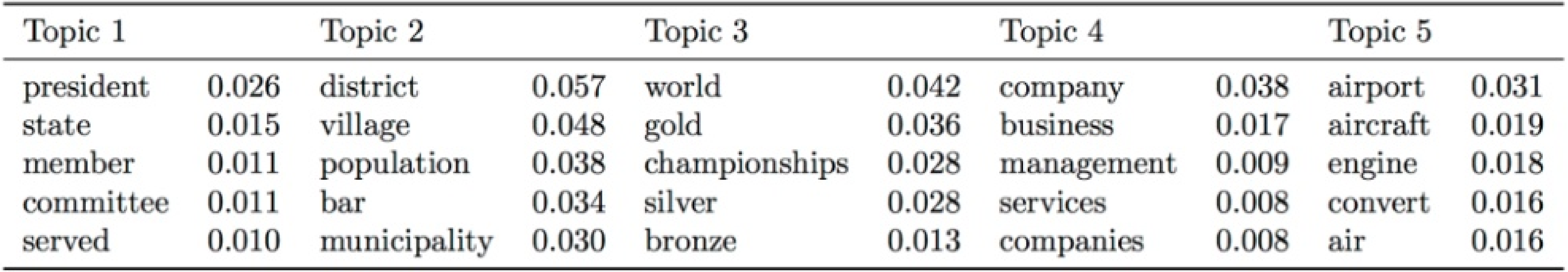
```
decision boundary: 0.50
accuracy:  0.79
precision: 0.88
recall:    0.61
fscore:    0.71
area under ROC curve: 0.86
area under PR curve:  0.84
confusion matrix: Map(tn -> 2266.0, tp -> 1839.0, fn -> 681.0, fp -> 183.0)
```



Latent Dirichlet Allocation (LDA)

takes a collection of text documents and seeks to group them by topic

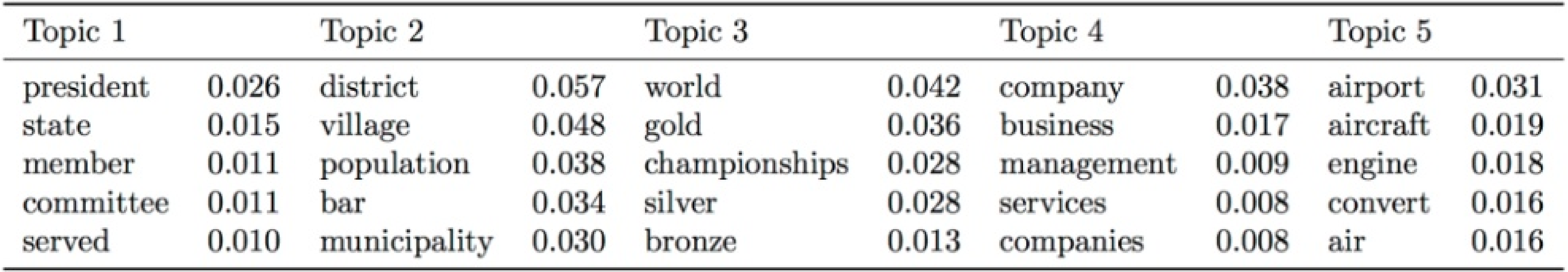
LDA on Wikipedia corpus yields:



Topic 1		Topic 2		Topic 3		Topic 4		Topic 5	
president	0.026	district	0.057	world	0.042	company	0.038	airport	0.031
state	0.015	village	0.048	gold	0.036	business	0.017	aircraft	0.019
member	0.011	population	0.038	championships	0.028	management	0.009	engine	0.018
committee	0.011	bar	0.034	silver	0.028	services	0.008	convert	0.016
served	0.010	municipality	0.030	bronze	0.013	companies	0.008	air	0.016

https://databricks.com/blog/2015/09/22/large-scale-topic-modeling-improvements-to-lda-on-apache-spark.html

Latent Dirichlet Allocation (LDA)



Topic 1		Topic 2		Topic 3		Topic 4		Topic 5	
president	0.026	district	0.057	world	0.042	company	0.038	airport	0.031
state	0.015	village	0.048	gold	0.036	business	0.017	aircraft	0.019
member	0.011	population	0.038	championships	0.028	management	0.009	engine	0.018
committee	0.011	bar	0.034	silver	0.028	services	0.008	convert	0.016
served	0.010	municipality	0.030	bronze	0.013	companies	0.008	air	0.016

https://databricks.com/blog/2015/09/22/large-scale-topic-modeling-improvements-to-lda-on-apache-spark.html

A document is a *probability distribution over topics*

Boeing: mixture of topics 4 and 5

Air Force One: mixture of topics 1 and 5

LDA

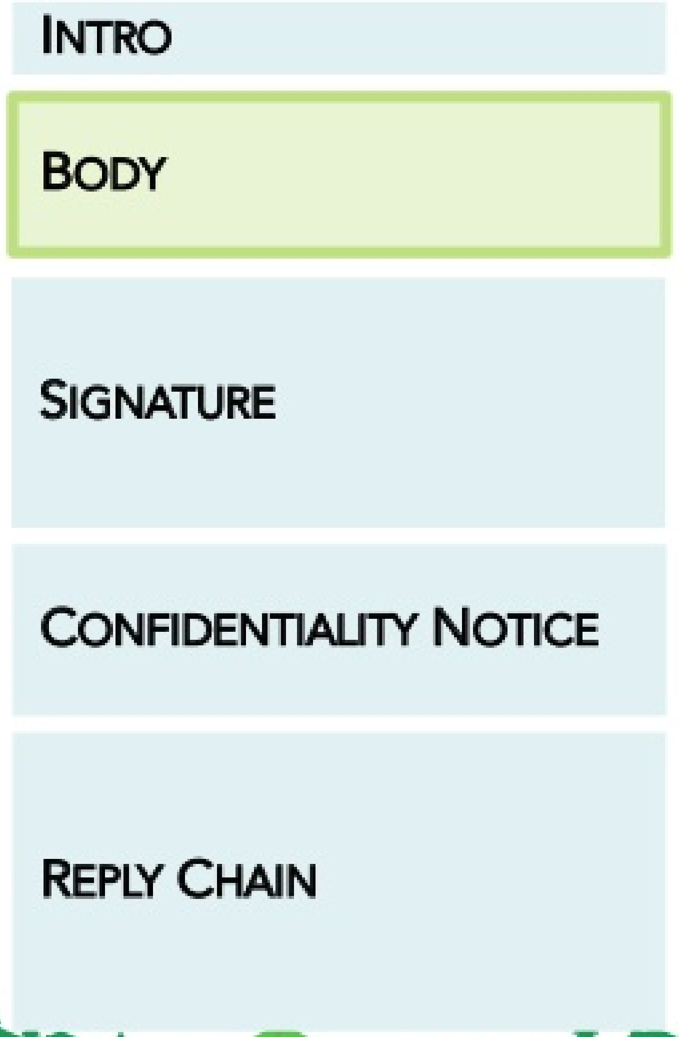
- Cannot (well, very hard to) select topics you want to identify in advance
- Can’t know what each topic is

Instead, include the **entire topic distribution** in the feature vector

Improving the topics identified by LDA

INTRO	Hey Alexis,
BODY	Let's meet with Ascander on Friday to discuss the $10,000/year rate. Ascander's phone number is (123) 456-7890.
SIGNATURE	Thanks, Noah Bergman Engineer at Salesforce (123) 456-7890
CONFIDENTIALITY NOTICE	The contents of this email and any attachments are confidential and are intended solely for addressee…
REPLY CHAIN	From: Alexis alexis@salesforce.com Date: April 1, 2017 Subject: Important Document Noah, how much does your product cost?

Improving the topics identified by LDA



INTRO

BODY

SIGNATURE

CONFIDENTIALITY NOTICE

REPLY CHAIN

- Common information blends topics together
- Reply chains add topics and oversample

In the past, we've identified "Sent from my iPhone" as a topic!

Upcoming improvements

- Investigate alternative methods of computing n-gram word vectors
- Use labeled data to generate high-recall filter
- Factor in user feedback

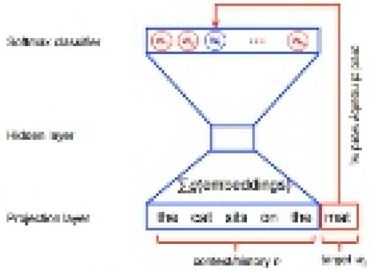
Agenda

- Salesforce introduction
- Inbox and email data
- Pricing request classifier pipeline
  - Labeling
  - Features generation
  - Scoring

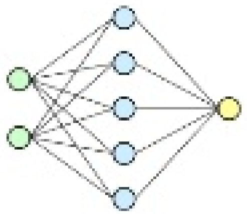
Scoring pipeline

Email Stream

Filtering

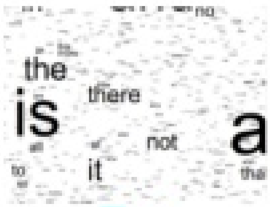


Word2Vec

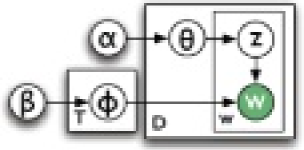


Graph

Feature Vector Generation

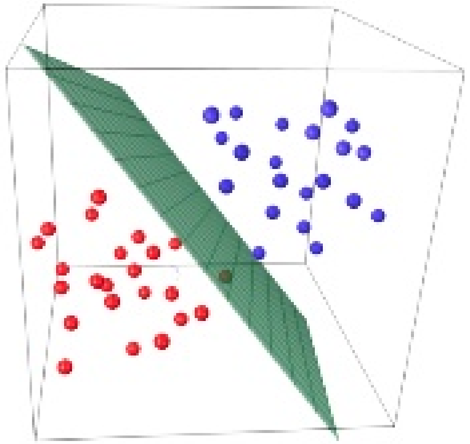


Text Processing / TF-IDF



LDA

Scoring

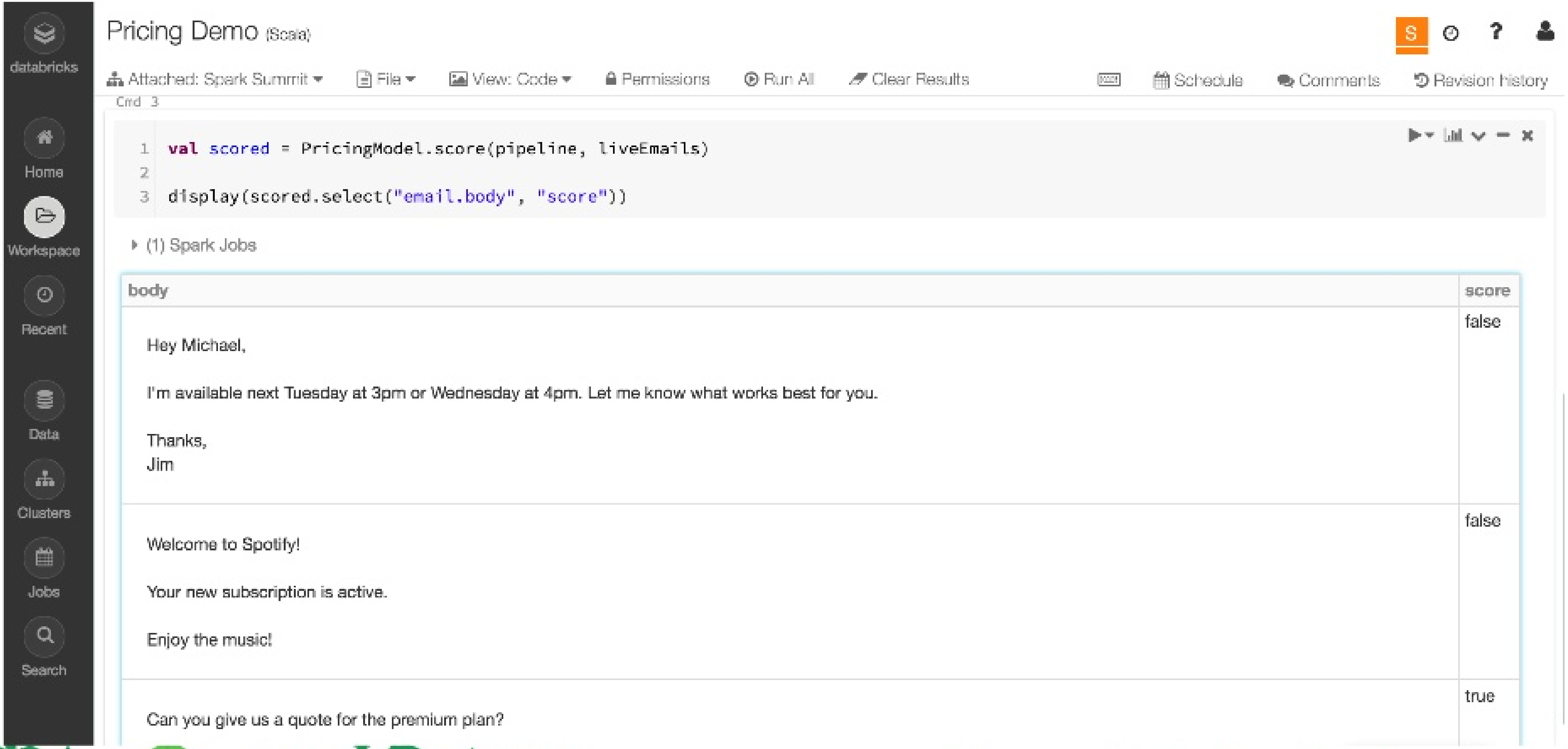


Scoring pipeline

```
val vectorizer: Dataset[Email] => DataFrame =
   ngramPipeline.transform _ andThen
   ldaPipeline.transform andThen
   assembler.transform

val featureVectors = vectorizer(emails)
val scored = model.transform(featureVectors)
```

Demo



```scala
val scored = PricingModel.score(pipeline, liveEmails)

display(scored.select("email.body", "score"))
```

body	score
Hey Michael, I'm available next Tuesday at 3pm or Wednesday at 4pm. Let me know what works best for you. Thanks, Jim	false
Welcome to Spotify! Your new subscription is active. Enjoy the music!	false
Can you give us a quote for the premium plan?	true

Some lessons learned

- High-recall filter
- Normalizing tokens
- Interleaving n-grams with unigrams
- Extracting bodies
- Filtering out reply chains
- ML pipeline

Thank You.

We’re hiring: salesforce.com/careers

data science and engineering