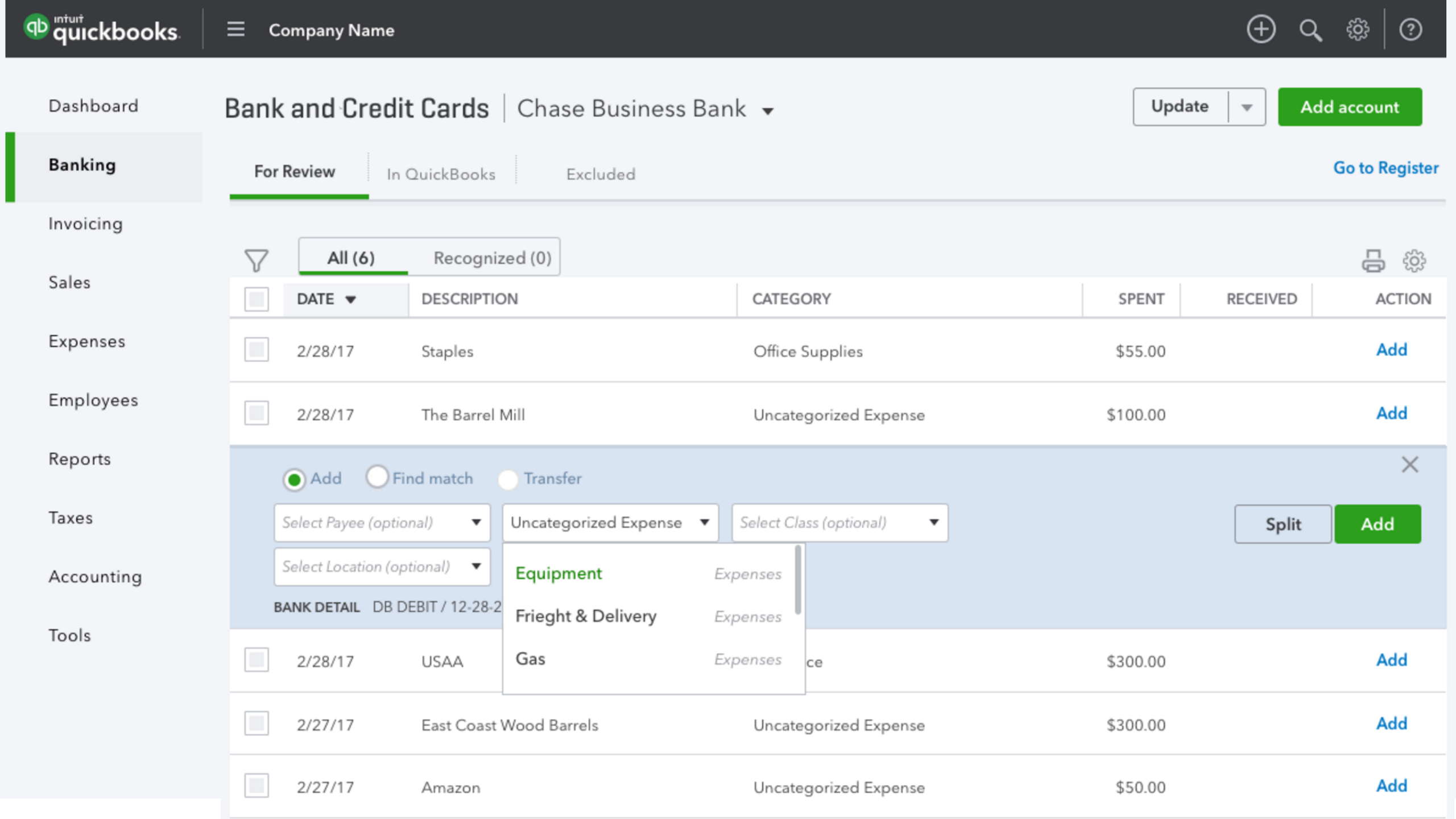
INTUIT

turbotax quickbooks proconnect mint

Data Science Automating Accounting

Oct. 25th 2017

Chris Lesner, Marko Rukonic, Wei Wang and Alex Ran


quickbooks
Company Name
Dashboard
Banking
Invoicing
Sales
Expenses
Employees
Reports
Taxes
Accounting
Tools
Bank and Credit Cards
Chase Business Bank
Update
Add account
For Review
In QuickBooks
Excluded
Go to Register
All (6)
Recognized (0)
DATE
DESCRIPTION
CATEGORY
SPENT
RECEIVED
ACTION
2/28/17 Staples Office Supplies $55.00 Add
2/28/17 The Barrel Mill Uncategorized Expense $100.00 Add
Add
Find match
Transfer
Select Payee (optional)
Uncategorized Expense
Select Class (optional)
Split
Add
Select Location (optional)
BANK DETAIL DB DEBIT / 12-28-2
Equipment Expenses
Frieght & Delivery Expenses
Gas Expenses
2/28/17 USAA ce $300.00 Add
2/27/17 East Coast Wood Barrels Uncategorized Expense $300.00 Add
2/27/17 Amazon Uncategorized Expense $50.00 Add

Find a match

TFR 92861732040100

BP CITY GATES 4228 MACKAY

interac purchase-7996 Memory Express

NEFT OW -Mohan M Upadhy-P17030607803370

Oeiras I

Takealot 485442*4328 26 SEP

UMBRELLA BEACH BAR

C2NMVbX8

TIM HORTONS 2312 QTH MISSION

Water Heaters

Rent or Lease

Repair & Maintenance

Stationery & Printing

Supplies

Taxes & Licenses

Travel

Travel Meals

Utilities

Shipping, Delivery Income

Job Materials

Other General and Admin Expenses

Prepaid Expenses

Transactions

Variable Format

City, State, Address

Transaction # / Reference #

Payee Name and/or ID#

Payer Name and/or ID#

Payment Method / Card or Account #

Date/Time, Amounts/Unit costs

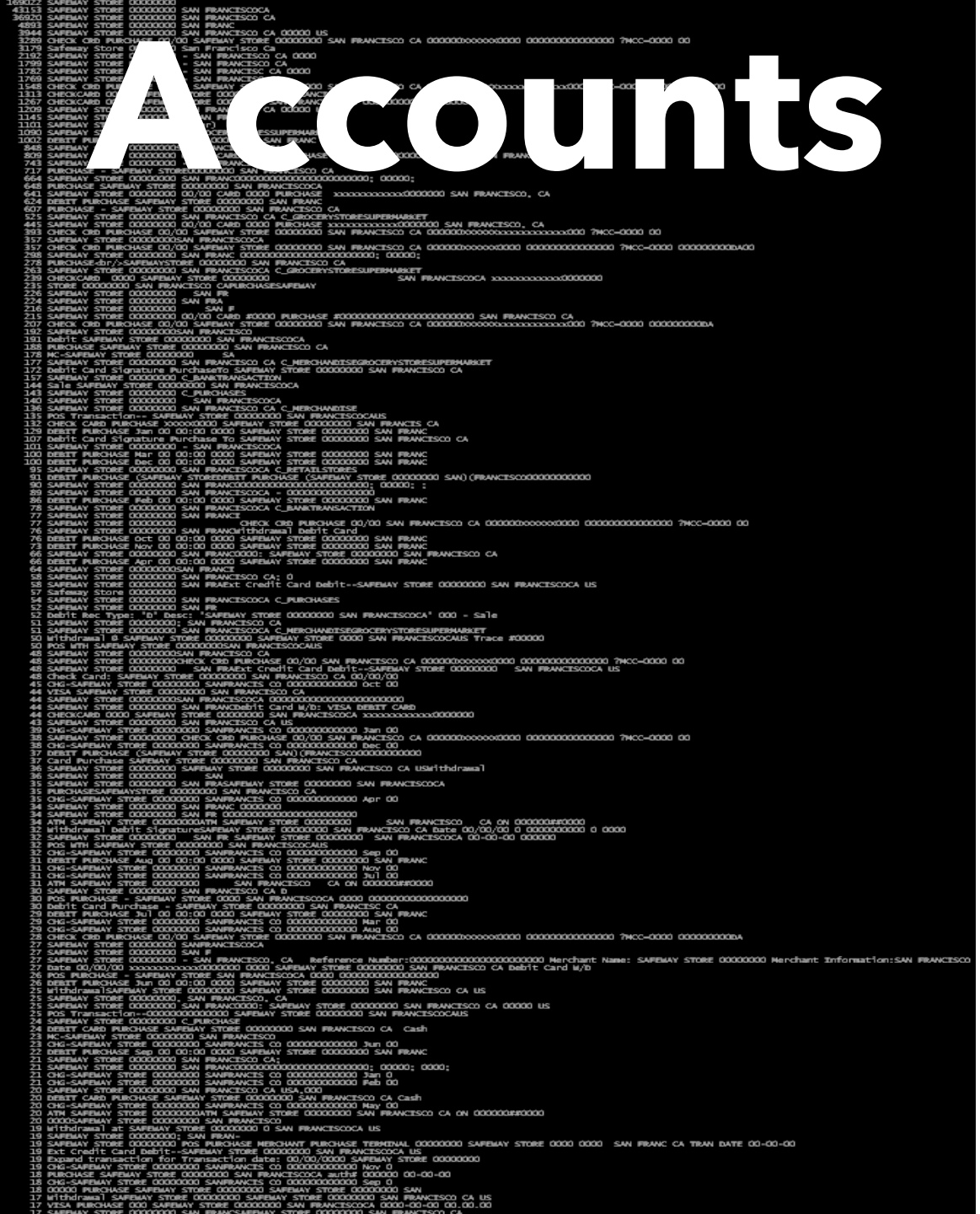
Errors: Token Merges,
Duplication, Truncation

300M transactions

9M unique descriptions (cleaned)

5M descriptions occur only once

Accounts



Same names but **different** purposes
(with different transaction types)

Different names but **same** purpose
(with same transaction types)

Overly generic / specific
(G&A, Misc, Ask accountant)

Stats:
97M accounts / 24M unique names
21M names occur only once
1.5M “vehicle” accounts

Same Safeway represented 300+ ways!

```
SAFEWAY STORE 00000000

SAFEWAY STORE 00000000 SAN FRANCISCOCA

SAFEWAY STORE 00000000 SAN FRANCISCO CA

SAFEWAY STORE 00000000 SAN FRANC

SAFEWAY STORE 00000000 SAN FRANCISCO CA 00000 US

CHECK CRD PURCHASE 00/00 SAFEWAY STORE 00000000 SAN FRANCISCO CA 000000XXXXXX0000
000000000000000 ?MCC=0000 00

CHECK CRD PURCHASE 00/00 SAFEWAY STORE 00000000 SAN FRANCISCO CA 000000XXXXXXxxxxxxxxxxxx000
?MCC=0000 000000000DA00

SAFEWAY STORE 00000000 SAN FRANC SAFEWAY STORE 00000000 SAN FRANCISCO 00000000000 000 Oct 00 @
0:00pm

SAFEWAY STORE 00000000 SAN FRANCISCO, CA 00.00 USD @ 0.000000

PURCHASE<br/>SAFEWAYSTORE 00000000 SAN FRANCISCO CA

SAFEWAY STORE 00000000 - SAN FRANCISCO, CA   Reference Number:000000000000000000000000 Merchant
Name: SAFEWAY STORE 00000000 Merchant Information: SAN FRANCISCO CA Category:
Retail/Department Stores
```

Same Safeway represented 300+ ways!

SAFEWAY STORE 00000000

SAFEWAY STORE 00000000 SAN FRANCISCOCA

SAFEWAY STORE 00000000 SAN FRANCISCO CA

SAFEWAY STORE 00000000 SAN FRANC

SAFEWAY STORE 00000000 SAN FRANCISCO CA 00000 US

CHECK CRD PURCHASE 00/00 SAFEWAY STORE 00000000 SAN FRANCISCO CA 000000XXXXXX0000 0000000000000000 ?MCC=0000 00

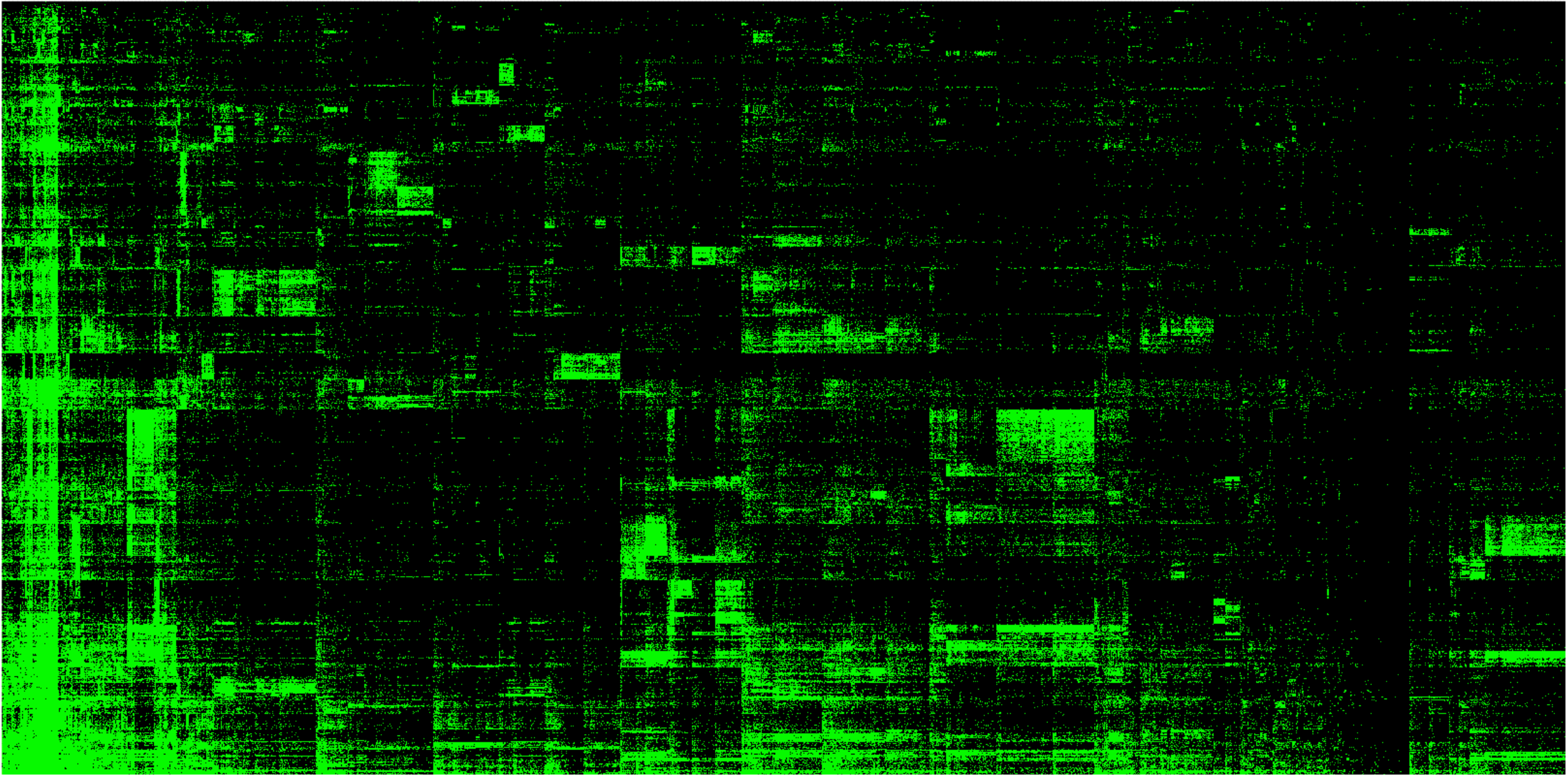
CHECK CRD PURCHASE 00/00 SAFEWAY STORE 00000000 SAN FRANCISCO CA 000000XXXXXXxxxxxxxxxxxxx000 ?MCC=0000 000000000DA00

SAFEWAY STORE 00000000 SAN FRANC SAFEWAY STORE 00000000 SAN FRANCISCO 00000000000 000 Oct 00 @ 0:00pm

SAFEWAY STORE 00000000 SAN FRANCISCO, CA 00.00 USD @ 0.000000

PURCHASE
SAFEWAYSTORE 00000000 SAN FRANCISCO CA

SAFEWAY STORE 00000000 - SAN FRANCISCO, CA Reference Number:000000000000000000000000 Merchant Name: SAFEWAY STORE 00000000 Merchant Information: SAN FRANCISCO CA Category: Retail/Department Stores



Merchants (rows) to Accounts (columns)

Transaction Categorization Today

Source (Bank)		
Date	Payee	Amount
6/14	BOOMERAMOUNTAIN VIEW CA	$74.87
6/15	GOOGLE *ADWS6379576361	$44.87

Transaction Categorization Today

Source (Bank)			*Intuit - FDS*	
Date	Payee	Amount	Schedule C cat.	Clean payee
6/14	BOOMERAMOUNTAIN VIEW CA	$74.87	10016 Other Bus. Expense	Boomeramountain View Ca
6/15	GOOGLE *ADWS6379576361	$44.87	10020 Uncategorized	Adws

Transaction Categorization Today

Source (Bank)			*Intuit - FDS*		*QBO*	
Date	Payee	Amount	Schedule C cat.	Clean payee	Company id	Chart of Accts
6/14	BOOMERAMOUNTAIN VIEW CA	$74.87	10016 Other Bus. Expense	Bomeramountain View Ca	58361	Business Expenses
6/15	GOOGLE *ADWS6379576361	$44.87	10020 Uncategorized	Adws	58361	Uncategorized Expense

intuit cto dev

intuit QuickBooks Online

Community Categorization

Intuit - FDS		*QBO*	
Schedule C cat.	Clean payee	Company id	Chart of Accts
10016 Other Bus. Expense	Boomeramount ain View Ca	58361	Business Expenses
10020 Uncategorized	Adws	58361	Uncategorized Expense

Community Categorization

Intuit - FDS		*QBO*	
Schedule C cat.	Clean payee	Company id	Chart of Accts
10016 Other Bus. Expense	Boomeramo untain View Ca	58361	Business Expenses
10020 Uncategorized	Adws	58361	Uncategorized Expense

Community Categorization

Intuit - FDS		*QBO*	
Schedule C cat.	Clean payee	Company id	Chart of Accts
10016 Other Bus. Expense	Boomeramo untain View Ca	58361	Business Expenses

Community Categorization

Intuit - FDS		*QBO*	
Schedule C cat.	Clean payee	Company id	Chart of Accts
10016 Other Bus. Expense	Boomeramo untain View Ca	58361	Business Expenses
		768767	Software

Community Categorization

Intuit - FDS		*QBO*	
Schedule C cat.	Clean payee	Company id	Chart of Accts
10016 Other Bus. Expense	Boomeramo untain View Ca	58361	Business Expenses
		768767	Software
		4873264	Services

Community Categorization

Intuit - FDS		*QBO*	
Schedule C cat.	Clean payee	Company id	Chart of Accts
10016 Other Bus. Expense	Boomeramo untain View Ca	58361	Business Expenses
		768767	Software
		4873264	Services
		2354736	Software

Community Categorization

Intuit - FDS		*QBO*	
Schedule C cat.	Clean payee	Company id	Chart of Accts
10016 Other Bus. Expense	Boomeramo untain View Ca	58361	Business Expenses
		768767	Software
		4873264	Services
		2354736	Software
		1723647	Software

Community Categorization

Intuit - FDS		*QBO*	
Schedule C cat.	Clean payee	Company id	Chart of Accts
10016 Other Bus. Expense	Boomeramo untain View Ca	58361	Business Expenses
		768767	**Software**
		4873264	Services
		2354736	**Software**
		1723647	**Software**

Community Categorization

Intuit - FDS		*QBO*	
Schedule C cat.	Clean payee	Company id	Chart of Accts
10016 Other Bus. Expense	Boomeramountain View Ca	58361	Business Expenses

IRIS
Chart of Accts from community
Software

Community Categorization

Intuit - FDS		*QBO*		*IRIS*
Schedule C cat.	Clean payee	Company id	Chart of Accts	Chart of Accts from community
10016 Other Bus. Expense	Boomeramo untain View Ca	58361	Business Expenses	Software
10020 Uncategorized	Adws	58361	Uncategorize d Expense	Advertising

Another look at the data

- Over the last year, QBO customers categorized *500M transactions with 17M distinct payees* into *22M different accounts that have 4M unique names*
- On average, a QBO company categorized 400 transactions with 90 distinct payees into 28 accounts out of 62 in their chart of accounts
- Each company has a unique Chart of Accounts which defines a unique set of categories for transactions.

  Categories are defined (implicitly) by the company and not an external agency like the government in the case of tax categories

How can we use very few examples (~90) to handle millions (~17M) of future possibilities?

Categorization Prior to Using Machine Learning

In the past Intuit has developed 2 systems:

- A system that memorizes company-specific manual categorization and repeats the same choice when applicable. This approach is very reliable but it only applies to repeated transactions - about half all of all company transactions and yields ***accuracy of about 50%***
- An expert system that determines merchant category and assigns transactions to the account with the name linguistically most similar to merchant category. This approach has lower reliability and only applies to about two thirds of all transactions and yields ***accuracy of about 30%***

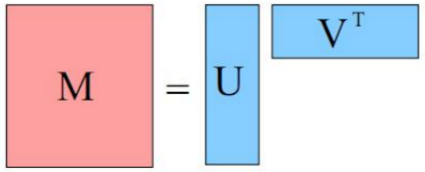
Combined Approach Delivered Accuracy of ~60%

Can we use machine learning to improve on the past?

Formalizing the intuition: Estimate probability of account given transaction

$$P(Ai \mid Tj) = \frac{P(Ai)P(Tj \mid Ai)}{P(Tj)}$$

	A1	A2	A3	…	An
T1	P(T1|A1)	P(T1|A2)	P(T1|A3)		P(T1|An)
T2	P(T2|A1)	P(T2|A2)	P(T2|A3)	…	P(T2|An)
…	…	…	…	…	…
Tm	P(Tm|A1)	P(Tm|A2)	…	…	P(Tm|An)



$$\begin{pmatrix} p(t_1 \mid a_1) & \cdots & \cdots & \cdots & p(t_1 \mid a_n) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ p(t_m \mid a_1) & \cdots & \cdots & \cdots & p(t_m \mid a_n) \end{pmatrix} = \begin{pmatrix} u_{11} & \cdots & u_{1d} \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \\ u_{1m} & \cdots & u_{md} \end{pmatrix} * \begin{pmatrix} v_{11} & \cdots & \cdots & \cdots & v_{1n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ v_{d1} & \cdots & \cdots & \cdots & v_{dn} \end{pmatrix}$$

$$p(t_i \mid a_j) = \sum_k u_{ik} \cdot v_{kj}$$

$$n * m \gg (n + m) * d$$

Similar companies assign similar transactions to similar accounts

businessday

Industries | Small Business | Money | Property | Farming | World | Innovation | Opinion & Analy

Xero finds machine-learning from the cloud will be trickier than expected

TOM PULLAR-STRECKER
Last updated 13:44, March 30 2017

Xero

Xero chief executive Rod Drury. The company announced on Wednesday that it had notched up its millionth

the **differences in the ways businesses apply account codes are far greater than Xero had expected.** One company's travel expense may be another firm's entertainment budget item.

…

"We're currently **only providing models that learn from the practices of individual businesses**," Gumbley explains.

…

"…there is huge variation in practice and encoding between different businesses – far greater than we expected."

…

If you could trust data had a common definition, it would potentially make it easy to apply other forms of artificial intelligence to provide actionable answers to a myriad of interesting questions that rely on business' accounts being like-for-like.

How does this work?

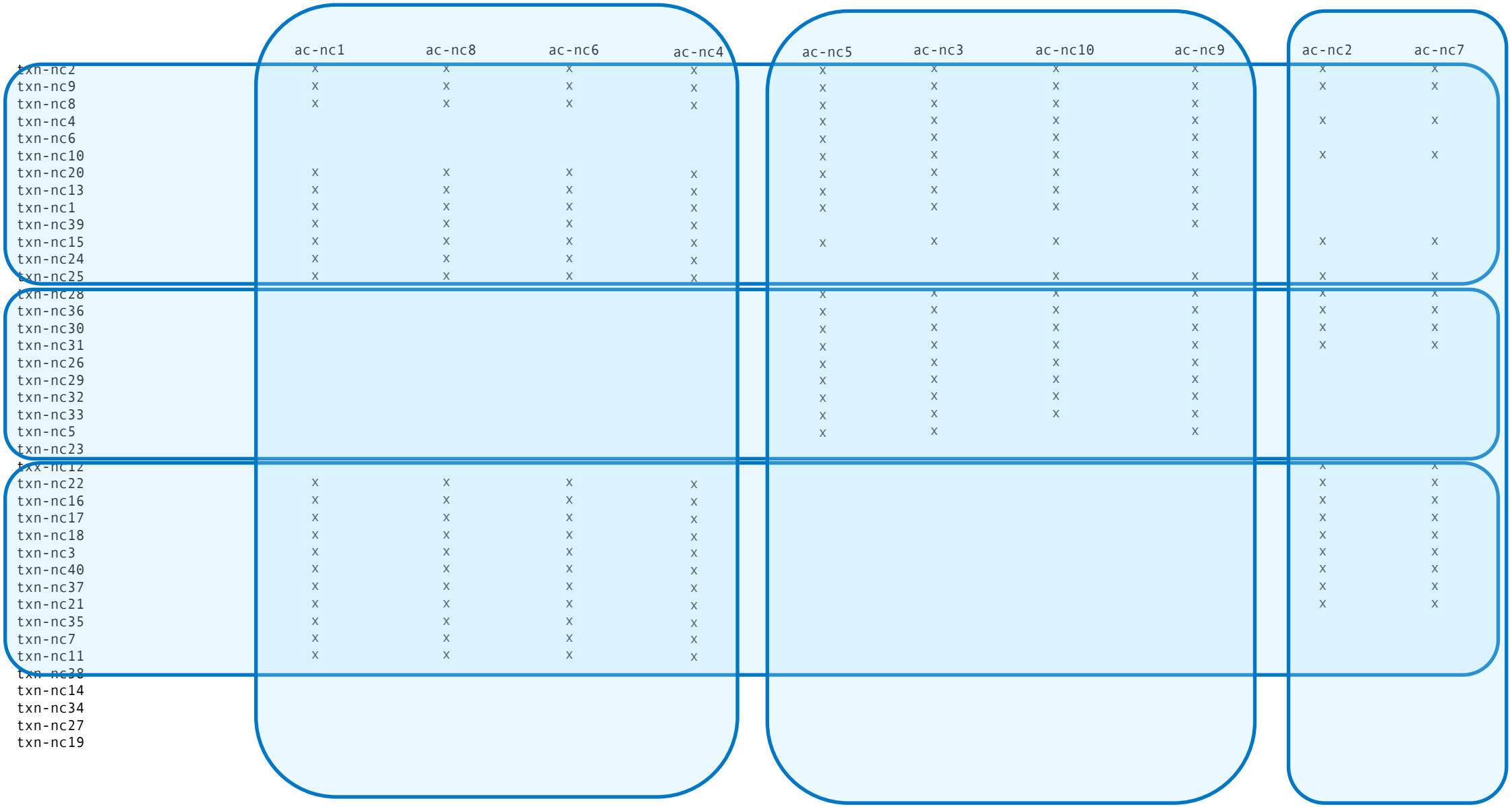
Content Clustering: Vectorization

	A1	A2	A3	…	An
T1	P(T1|A1)	P(T1|A2)	P(T1|A3)		P(T1|An)
T2	P(T2|A1)	P(T2|A2)	P(T2|A3)	…	P(T2|An)
…	…	…	…	…	…
Tm	P(Tm|A1)	P(Tm|A2)	…	…	P(Tm|An)

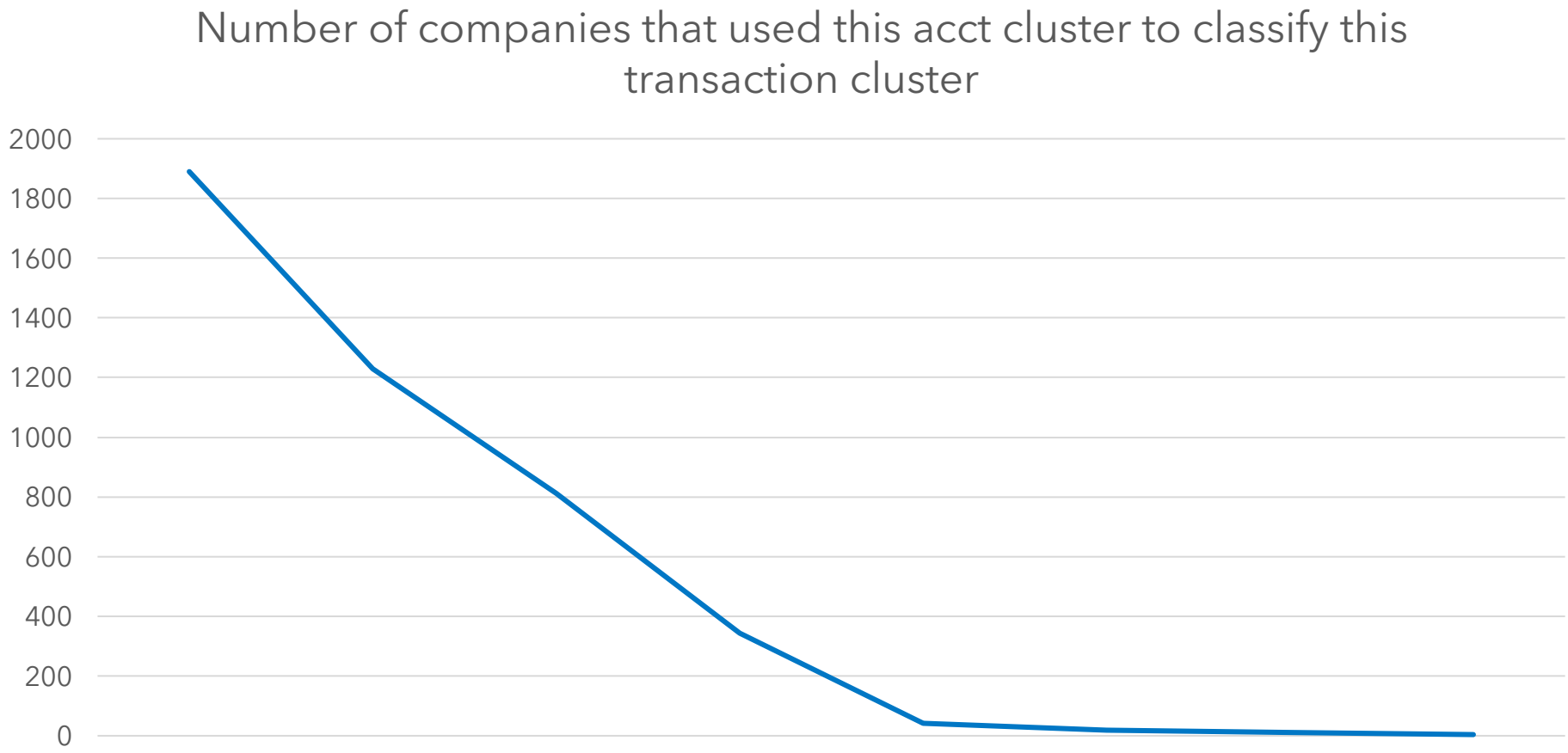
T1:	P(T1|A1)	P(T1|A2)	P(T1|A3)	---	…	P(T1|An)
A3:	P(T1|A3)	P(T2|A3)	P(T3|A3)	---	…	P(Tm|A3)

TF-IDF normalization: $P(Ti \mid Aj) \rightarrow P(Ti \mid Aj) * \ln(n / \sum_{k} (P(Ti \mid Ak) > \varepsilon)$

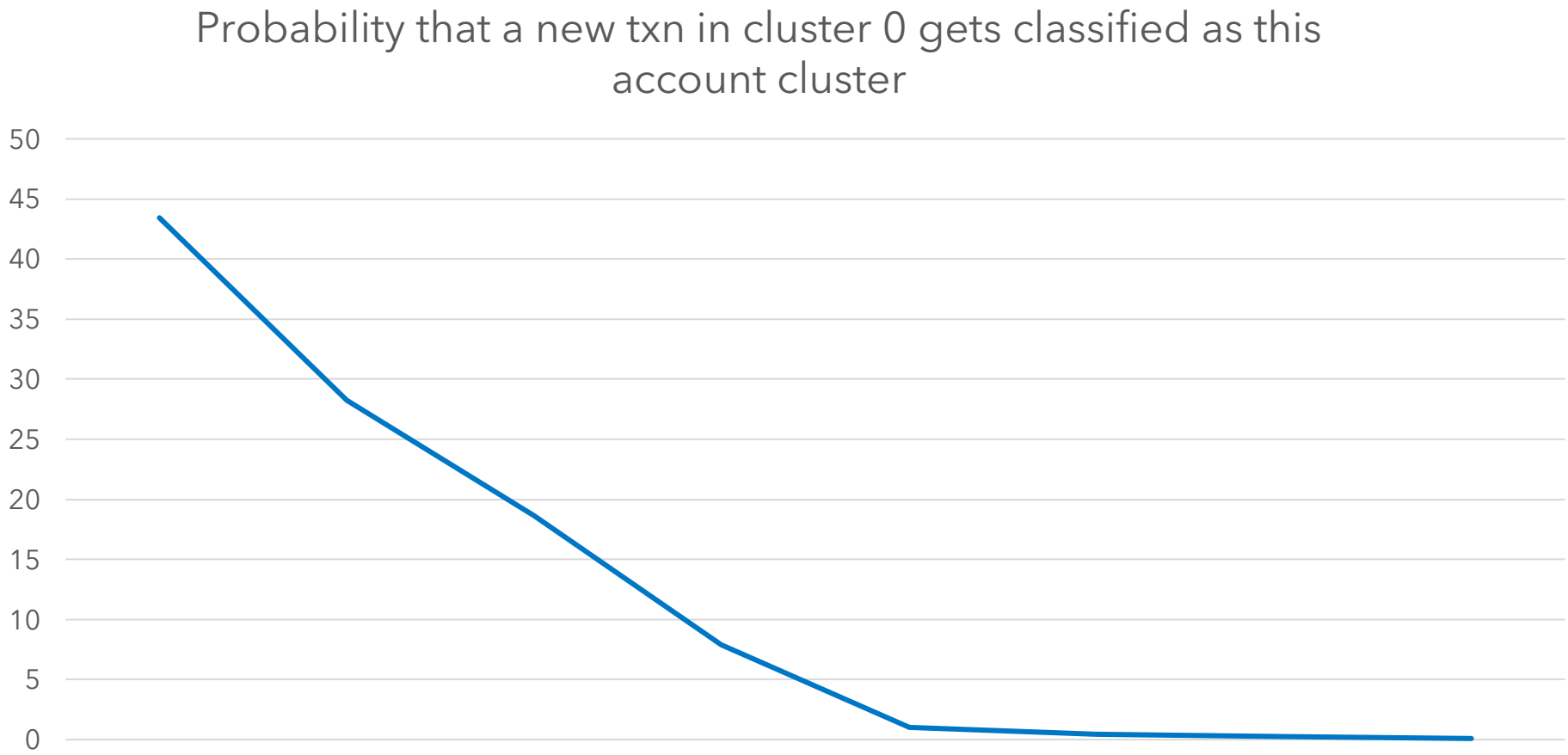
Similar transactions – similar accounts



	ac-cc0	ac-cc1	ac-cc2	ac-cc3	ac-cc4	ac-cc5	ac-cc6	ac-cc7
txn-cc0	1890	1229	812	344	43	18	12	4



	ac-cc0	ac-cc1	ac-cc2	ac-cc3	ac-cc4	ac-cc5	ac-cc6	ac-cc7
txn-cc0	1890	1229	812	344	43	18	12	4



Iris classifier – A case study

Company ID 21979721
“Amazon Services23897213 12-22-2015 OXXXX-XXXX-XXX”

FDS downloads this transaction from the bank On behalf of company 21979721

Iris classifier – A case study

Company ID 21979721
“Amazon Services23897213-Kindle 12-22-2015
OXXXX-XXXX-XXX”

Amazon.com, business category “Other Business Expenses”

Iris classifier – A case study

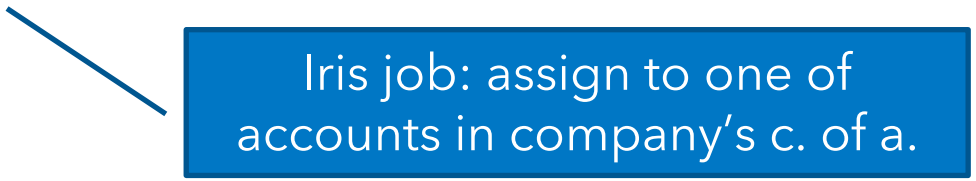
Company ID 21979721
“Amazon Services23897213-Kindle 12-22-2015
OXXXX-XXXX-XXX”

Amazon.com, business category “Other Business Expenses”

Iris classifier – A case study

Company ID 21979721
“Amazon Services23897213-Kindle 12-22-2015
OXXXX-XXXX-XXX”

Amazon.com, business category “Other Business Expenses”



Iris classifier – A case study

Amazon.com, business category “Other Business Expenses”

↓

Name cluster: c850aaea4b06

Iris assigns name cluster
(N->1 map)

Iris classifier – A case study

Amazon.com, business category “Other Business Expenses”

↓

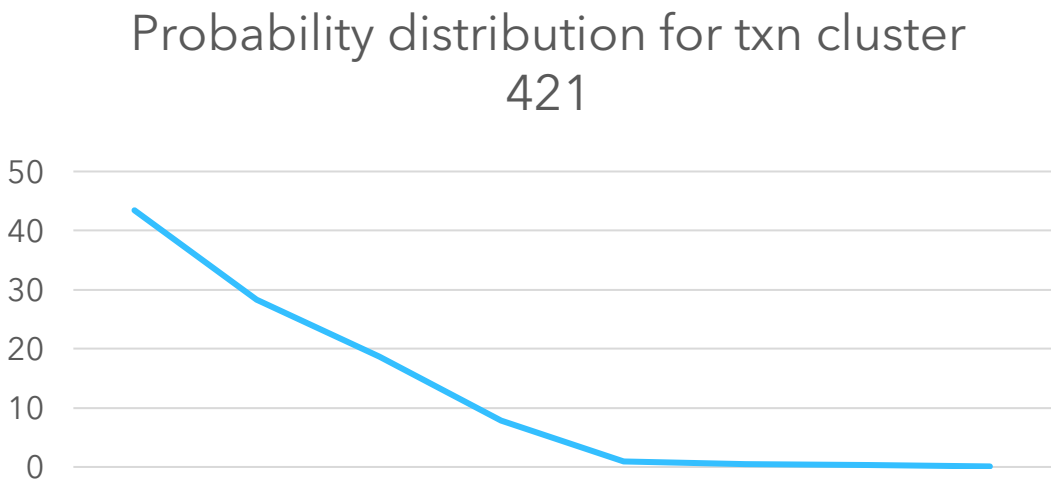
Name cluster: c850aaea4b06 → Content cluster: 421

Iris assigns content cluster (calculated previously) N->1

Iris classifier – A case study

Amazon.com, business category “Other Business Expenses”

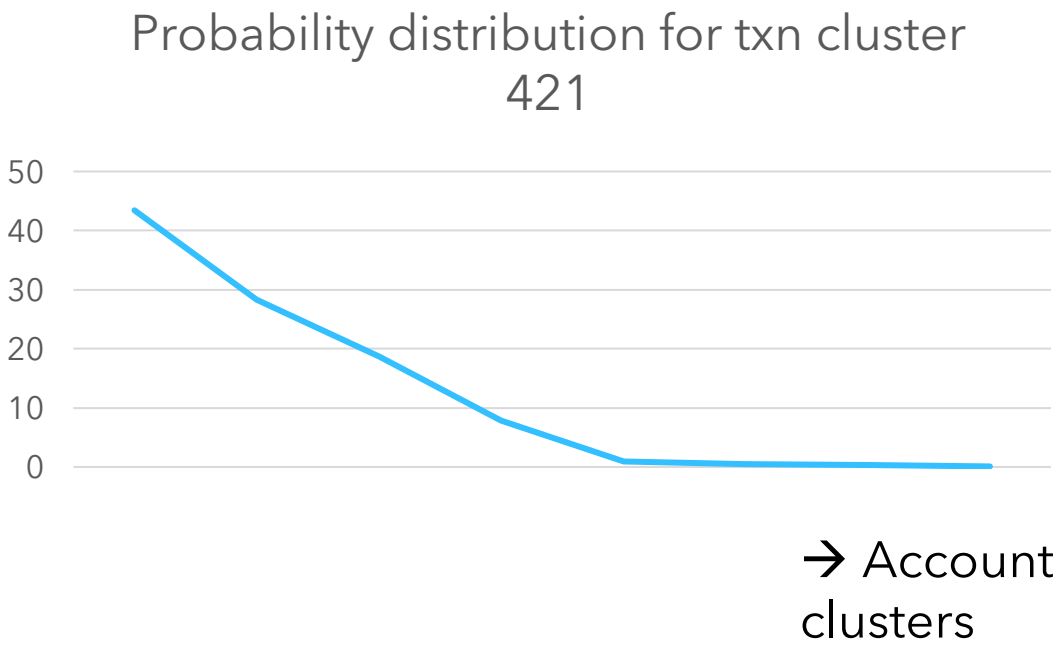
Name cluster: c850aaea4b06 → Content cluster: 421



Iris classifier – A case study

Amazon.com, business category “Other Business Expenses”

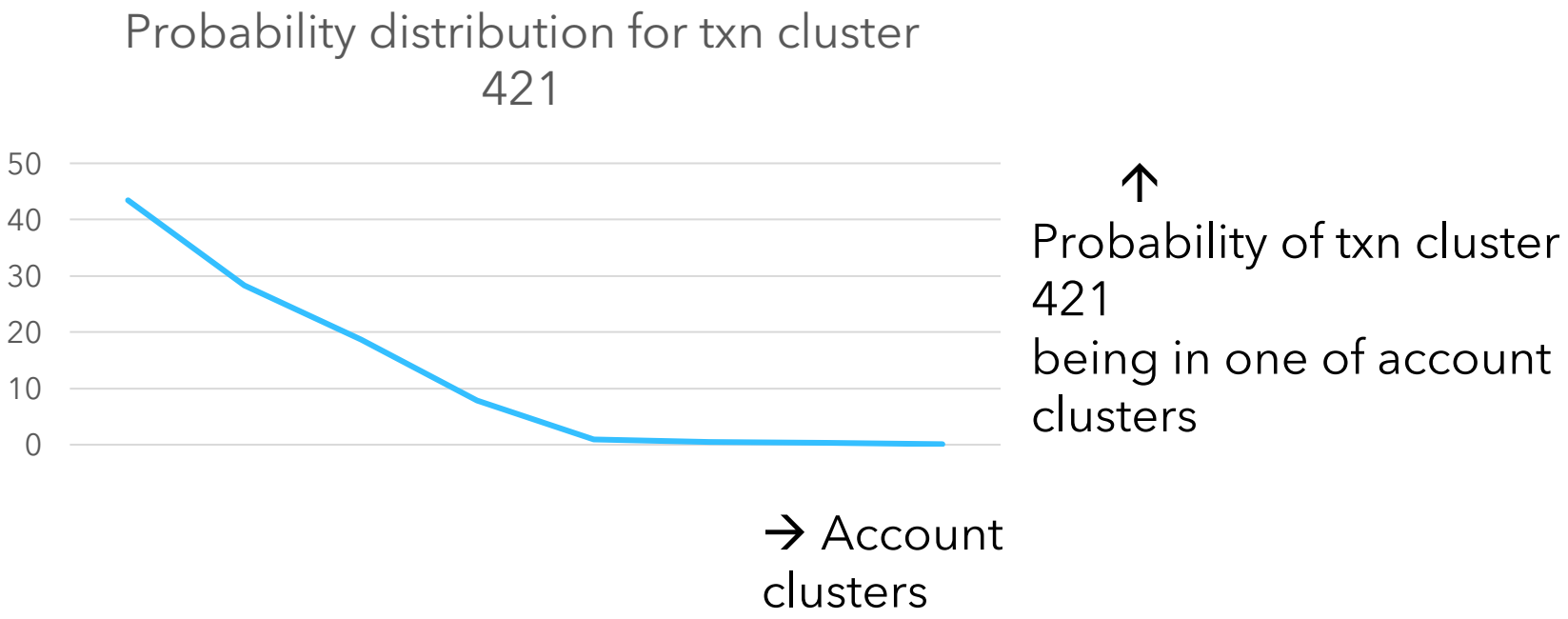
Name cluster: c850aaea4b06 → Content cluster: 421



Iris classifier – A case study

Amazon.com, business category “Other Business Expenses”

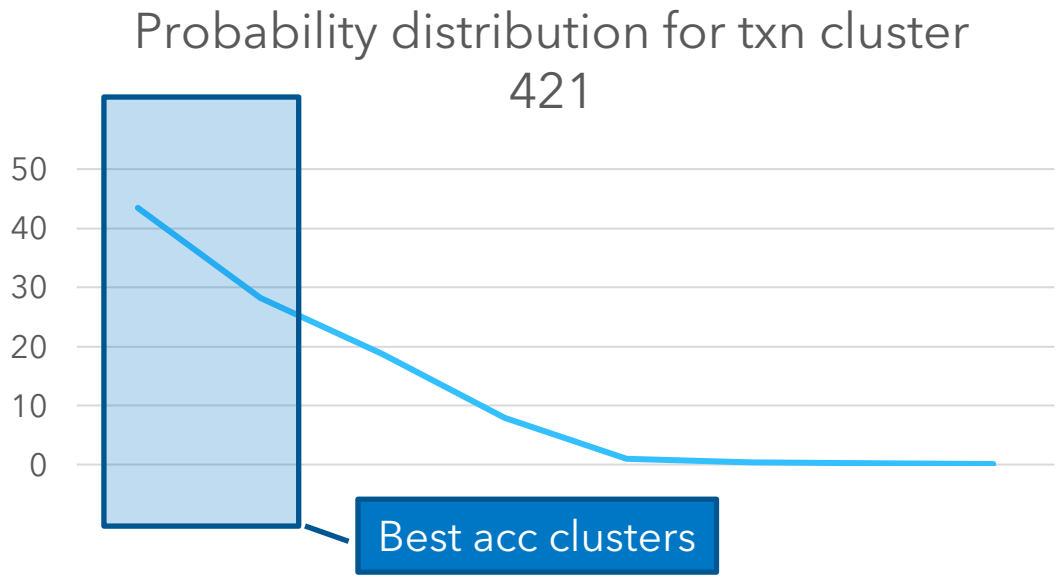
Name cluster: c850aaea4b06 → Content cluster: 421



Iris classifier – A case study

Amazon.com, business category “Other Business Expenses”

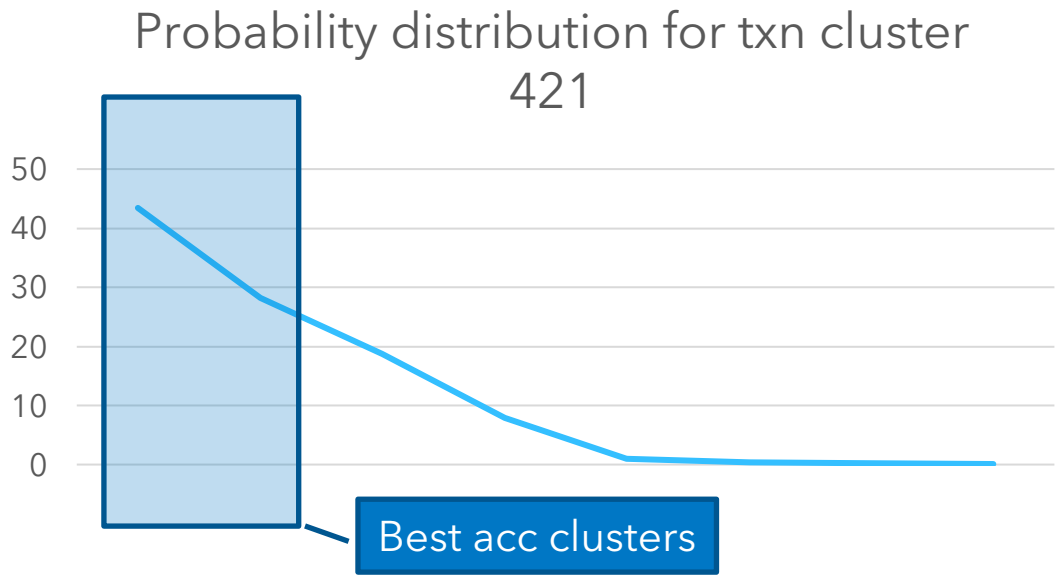
Name cluster: c850aaea4b06 → Content cluster: 421



Iris classifier – A case study

Amazon.com, business category “Other Business Expenses”

Name cluster: c850aaea4b06 → Content cluster: 421

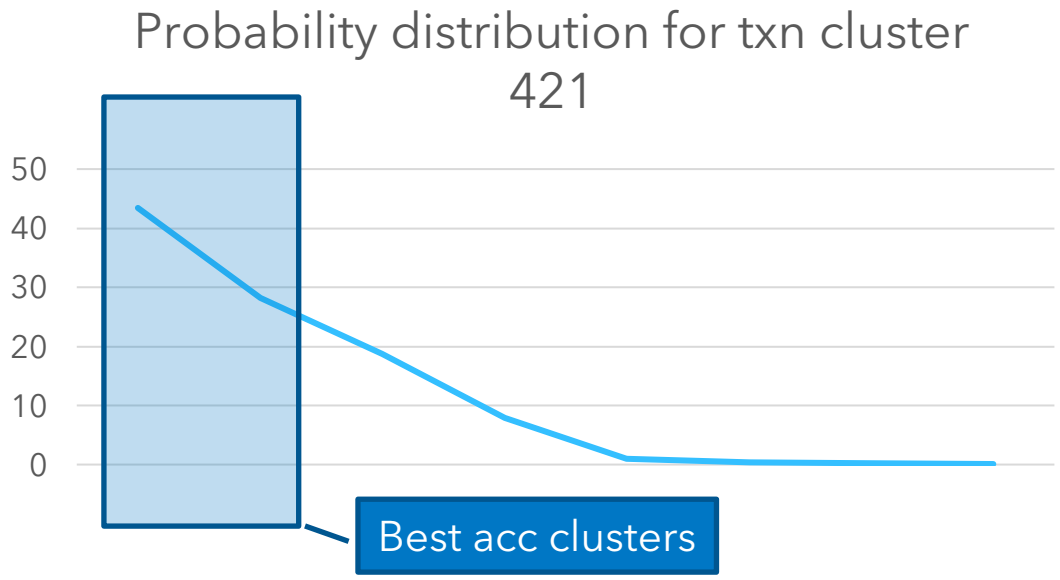


Acct cluster
12
433
94
864
765
31
24

Iris classifier – A case study

Amazon.com, business category “Other Business Expenses”

Name cluster: c850aaea4b06 → Content cluster: 421

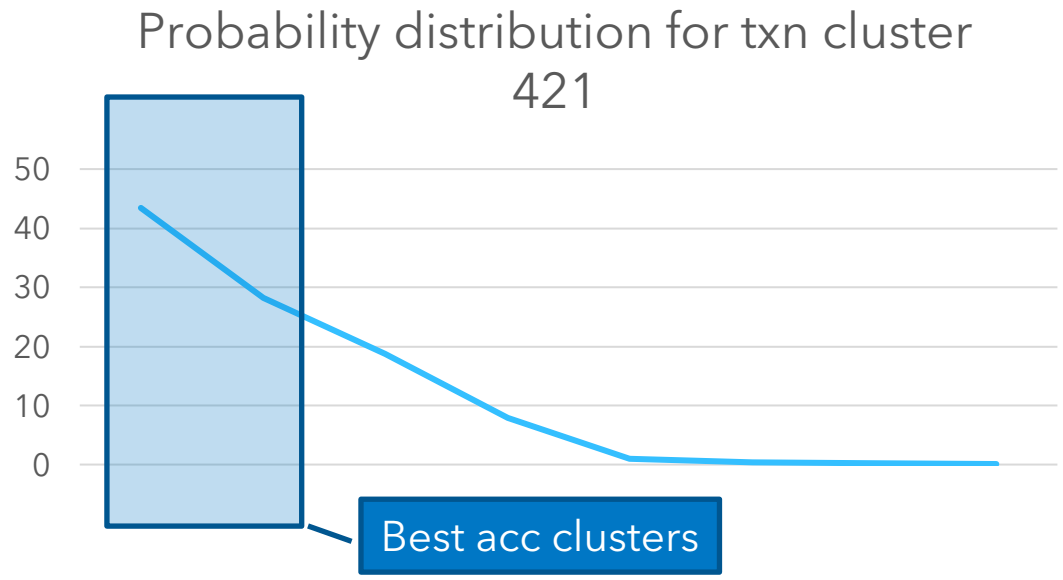


Acct cluster	In company c. of a.?
12	Yes
433	No
94	No
864	Yes
765	Yes
31	No
24	Yes

Iris classifier – A case study

Amazon.com, business category “Other Business Expenses”

Name cluster: c850aaea4b06 → Content cluster: 421

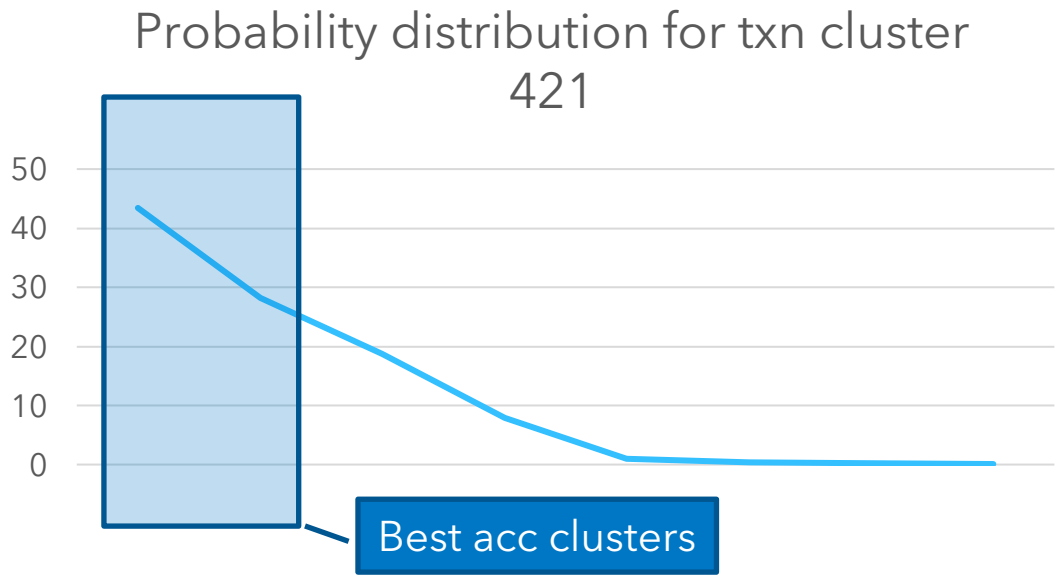


Acct cluster	In company c. of a.?
12	Yes
433	No
94	No
864	Yes
765	Yes
31	No
24	Yes

Iris classifier – A case study

Amazon.com, business category “Other Business Expenses”

Name cluster: c850aaea4b06 → Content cluster: 421



Acct name	In company c. of a.?
Lotions/oils	Yes
Hosting	Yes
Services	Yes
Materials	Yes

Iris classifier – A case study

Amazon.com, business category “Other Business Expenses”

Account recommendation for 100% of transactions (coverage)

Acct name	In company c. of a.?
Lotions/oils	Yes
Hosting	Yes
Services	Yes
Materials	Yes

Iris classifier – A case study

Amazon.com, business category “Other Business Expenses”

Our 1st acct recommendation is 80% accurate

Acct name	In company CoA?
Lotions/oils	Yes
Hosting	Yes
Services	Yes
Materials	Yes

Iris classifier – A case study

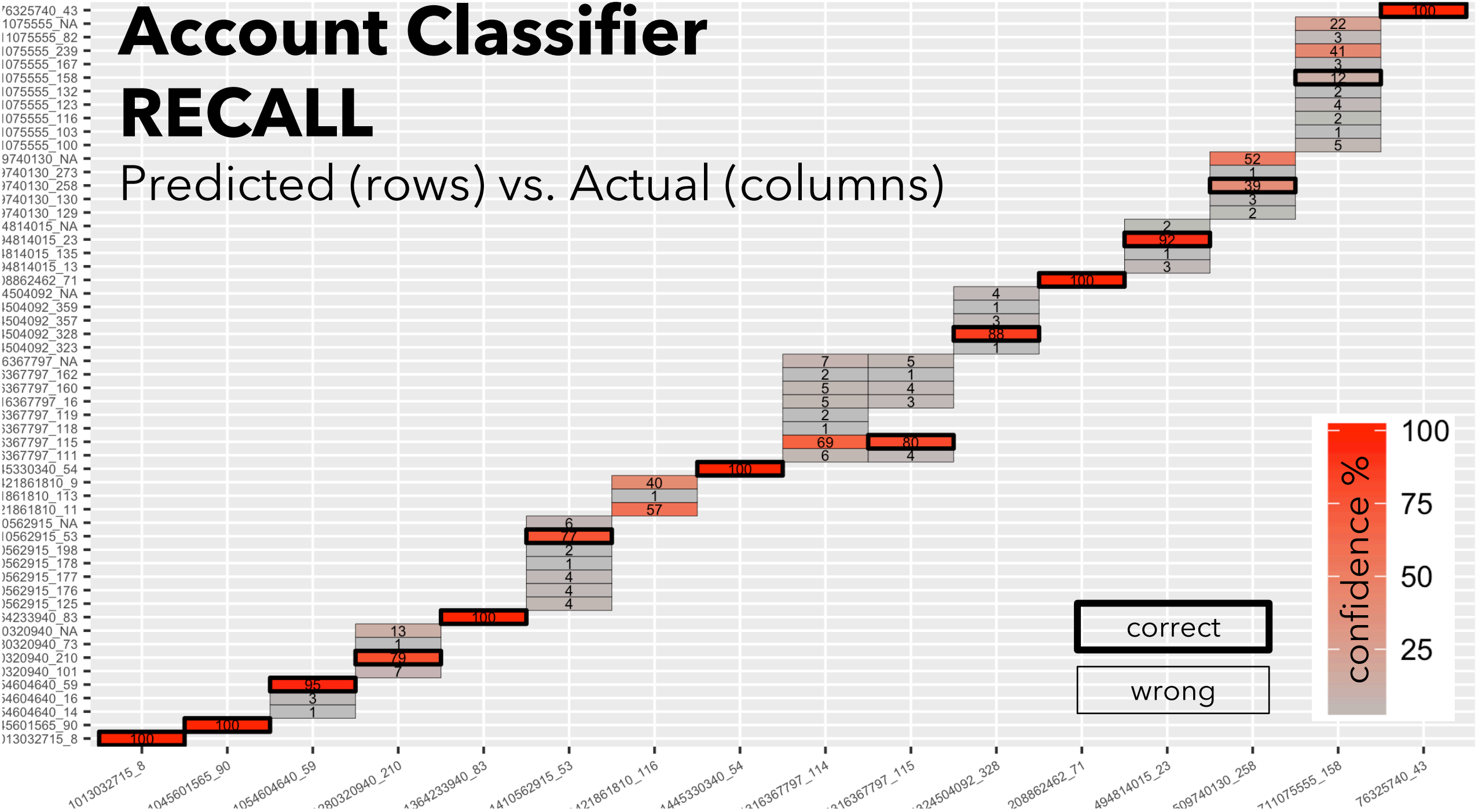
Amazon.com, business category “Other Business Expenses”

Our top-5 recommendations are 90% accurate

Acct name	In company CoA?
Lotions/oils	Yes
Hosting	Yes
Services	Yes
Materials	Yes

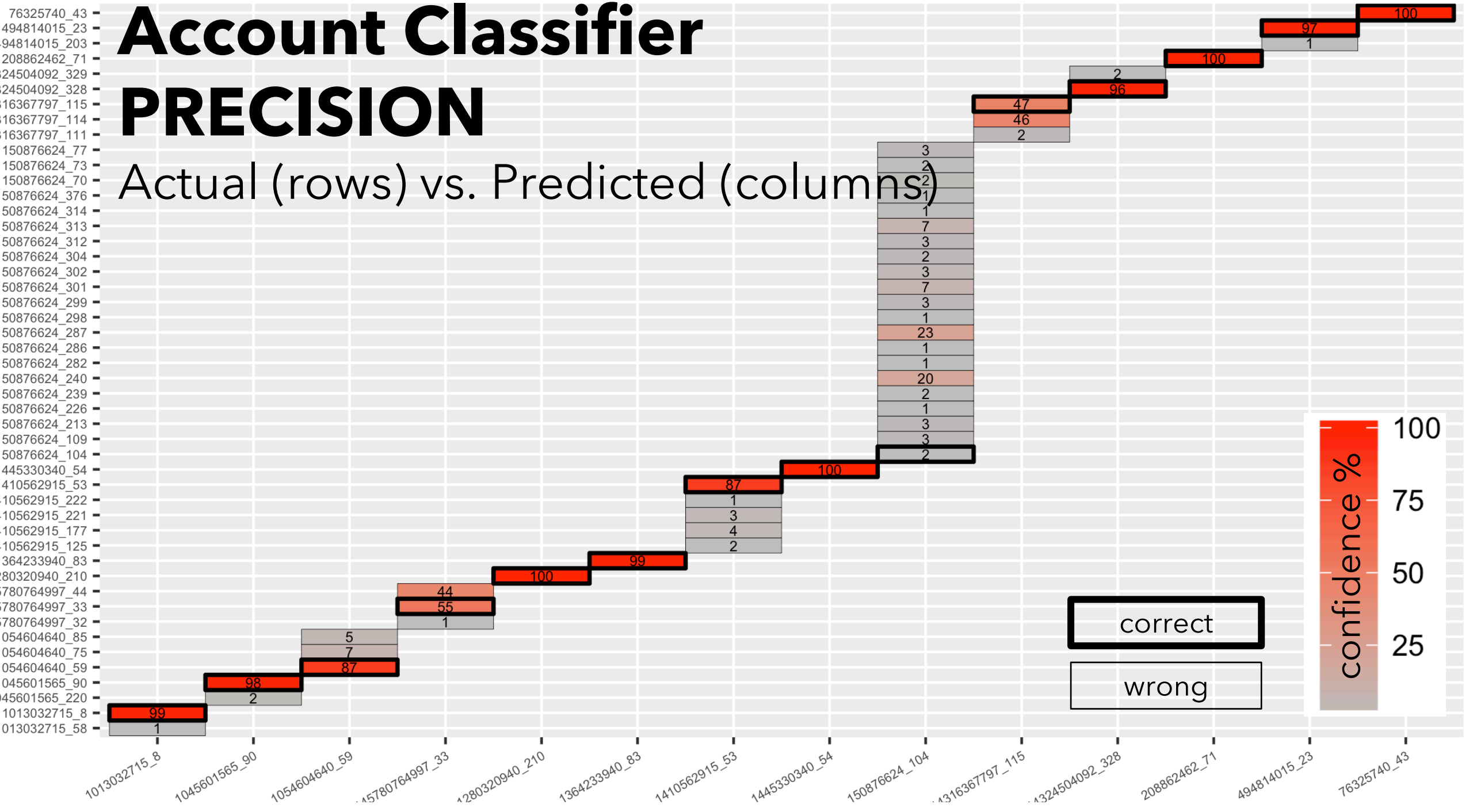
Account Classifier RECALL

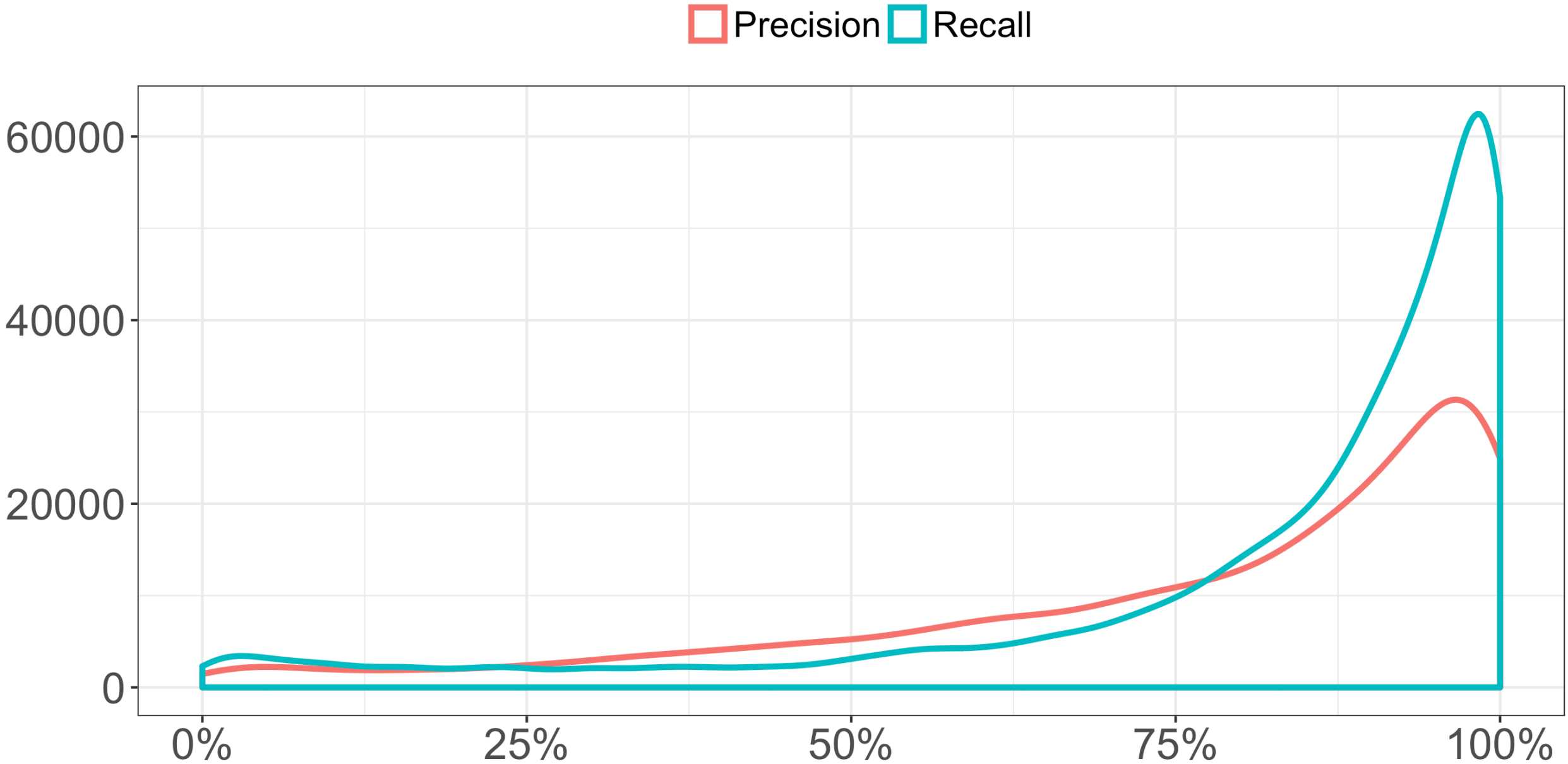
Predicted (rows) vs. Actual (columns)

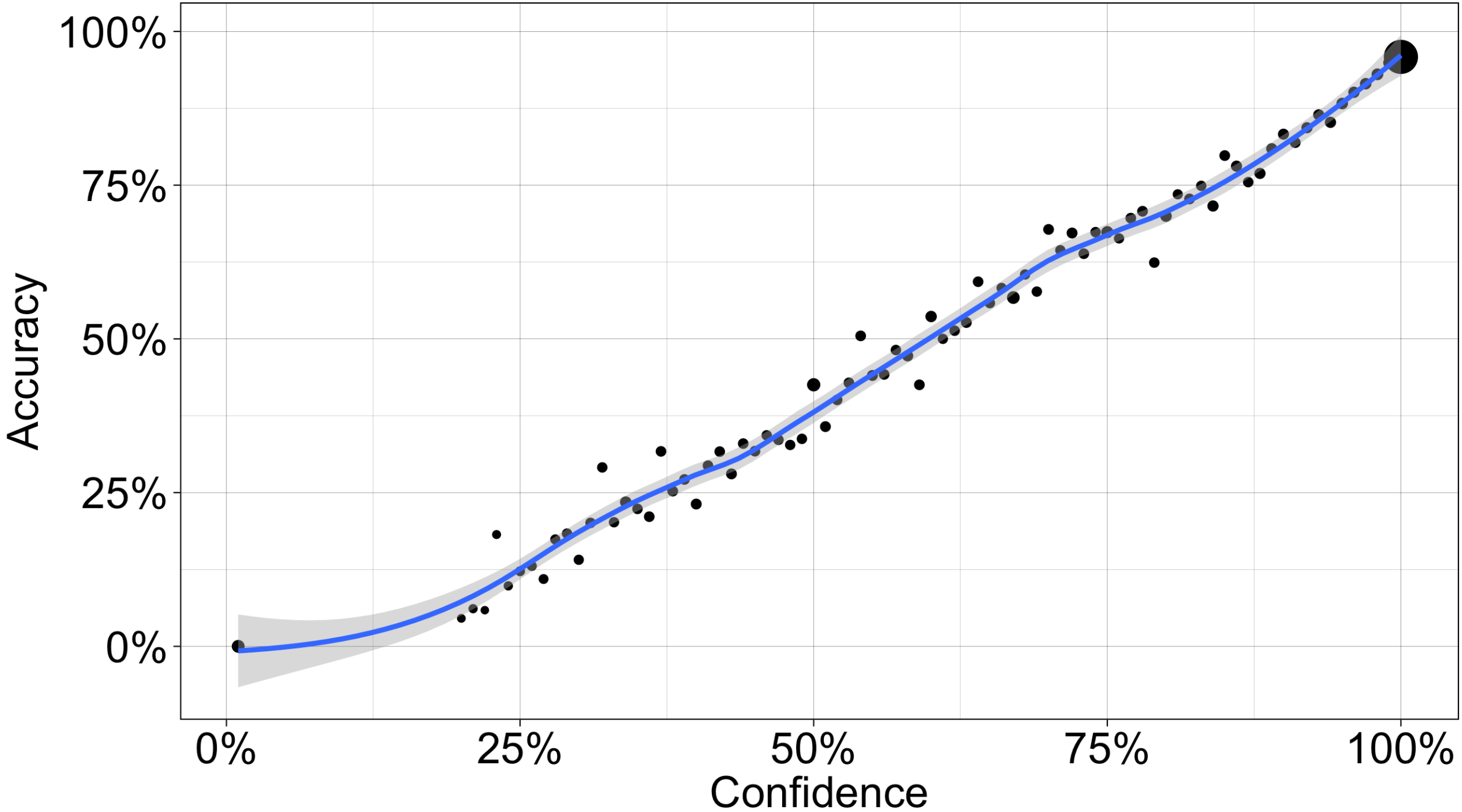


Account Classifier
PRECISION

Actual (rows) vs. Predicted (columns)




Precision
Recall
60000
40000
20000
0
0%
25%
50%
75%
100%


100%
75%
50%
25%
0%
Accuracy
0%
25%
50%
75%
100%
Confidence

Machine learning in service of accounting

Accounts with unlikely collections of transactions

Companies with poorly organized accounts

Chart of Accounts Repair

- Missing Account Recommendations
- Account Hierarchy Recovery
- Account Splitting

Thank you!