Data Science in Action

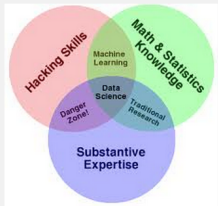
Ji Li

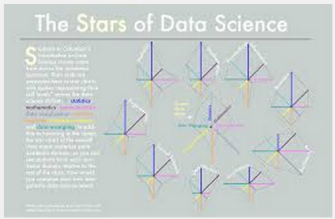
Data Scientist

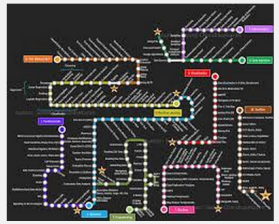
March 25, 2015

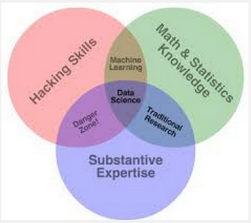
Table of Contents

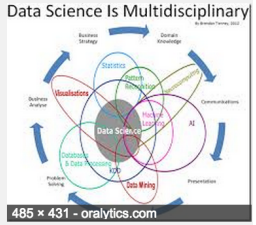
Data science on Google search

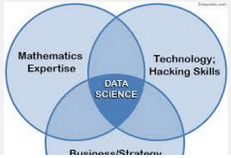












Data science from my point of view

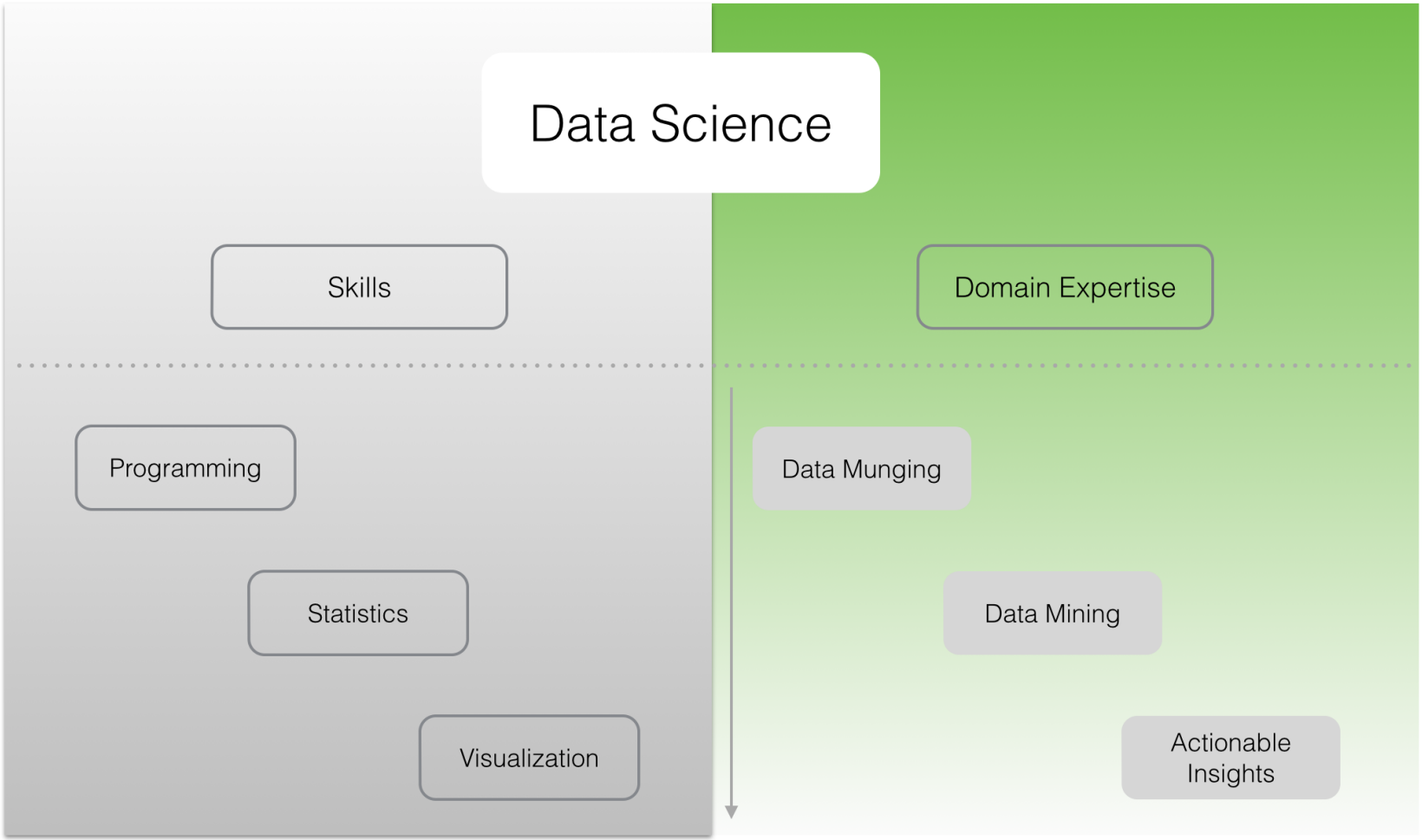


Table of Contents

Data Munging

Data Munging

..also called Data Wrangling, Data Preparation, usually means some or all of the following tasks that mark the beginning of a data science project:

- ETL

ETL

ETL is the process of extract, transform, and load data. - For one thing, we might want to acquire data from external sources to assist with data mining. On the other hand, there are often multiple data sources internally.

Data Munging

Data Munging

..also called Data Wrangling, Data Preparation, usually means some or all of the following tasks that mark the beginning of a data science project:

- ETL
- Data Integration

Data Integration

After ETL, we need to combine data from disparate sources into meaningful and valuable information.

Data Munging

Data Munging

..also called Data Wrangling, Data Preparation, usually means some or all of the following tasks that mark the beginning of a data science project:

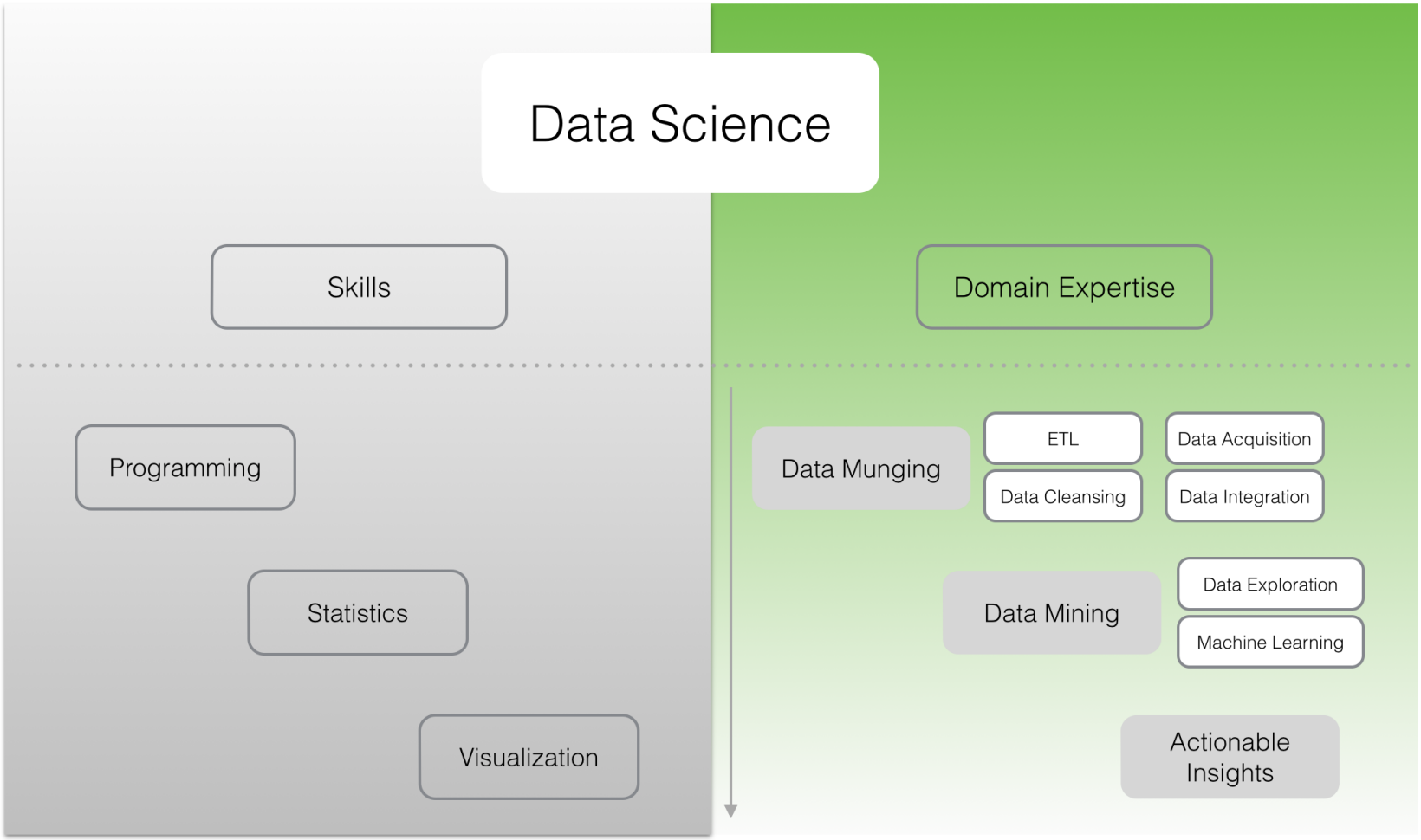
- ETL
- Data Integration
- Data Cleansing

Data Cleansing

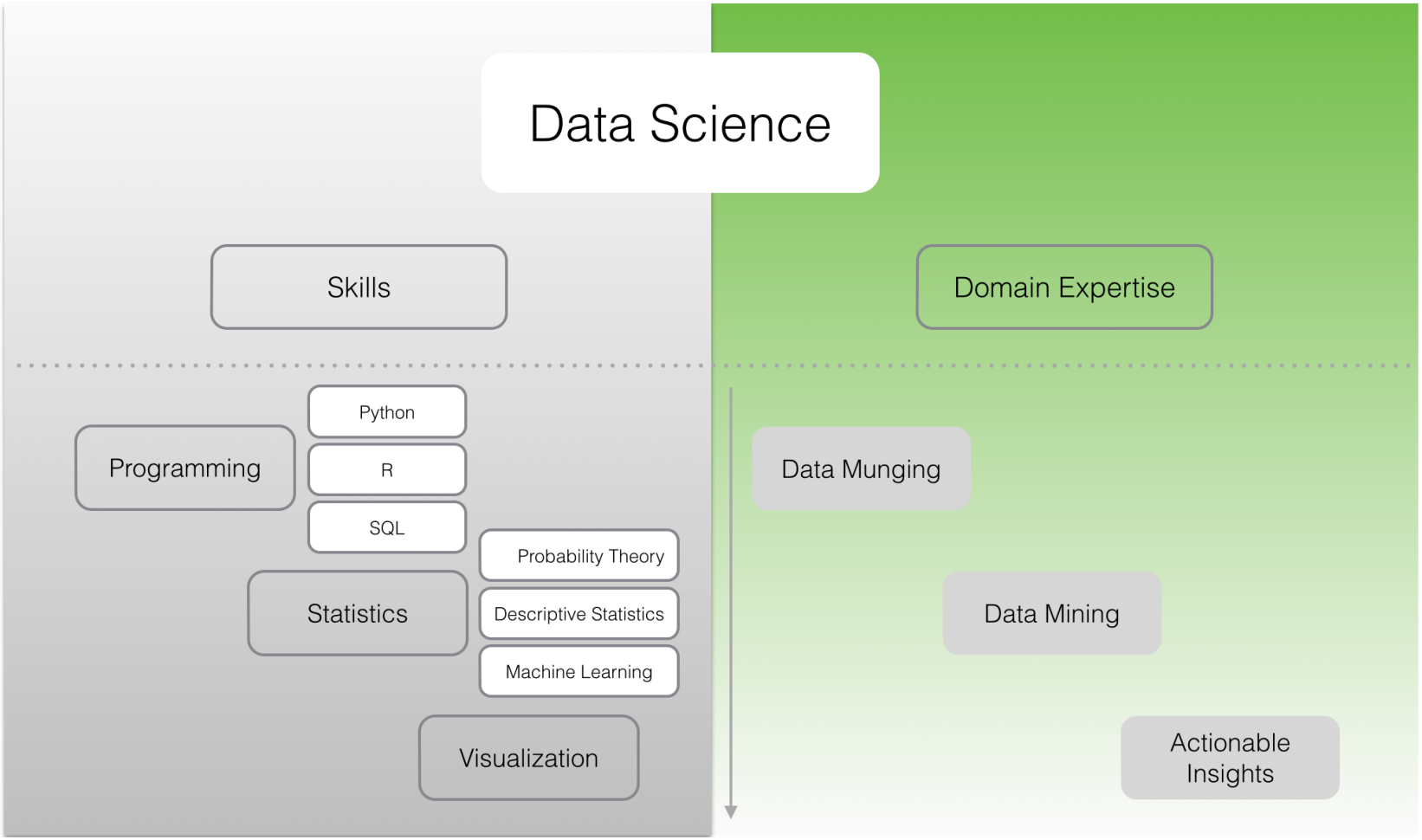
Data cleansing, also called data scrubbing, is the process of amending or removing data in a database that is incorrect, incomplete, improperly formatted, or duplicated.

Table of Contents

Domain expertise of a data scientist



Skills of a data scientist



Doing data science

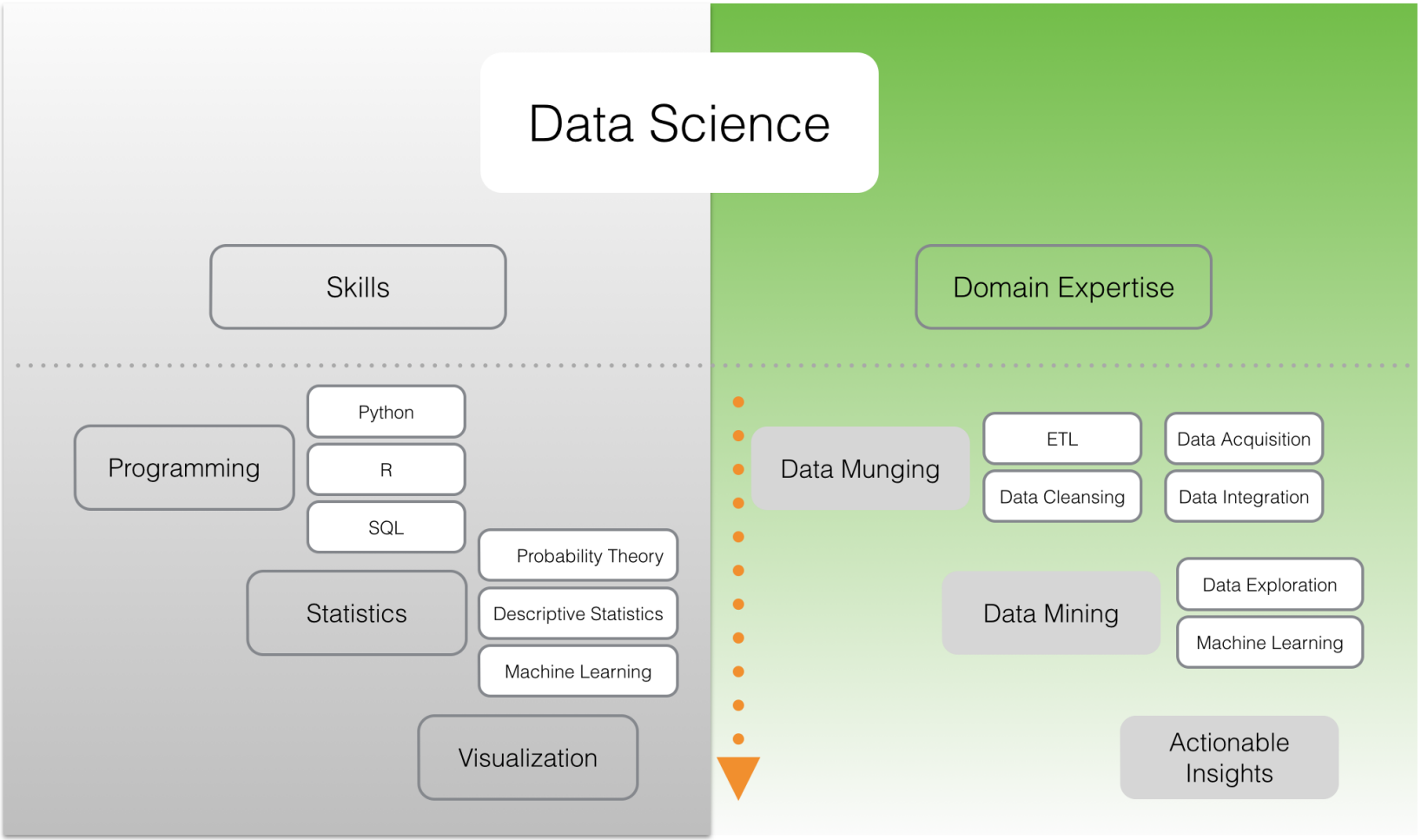


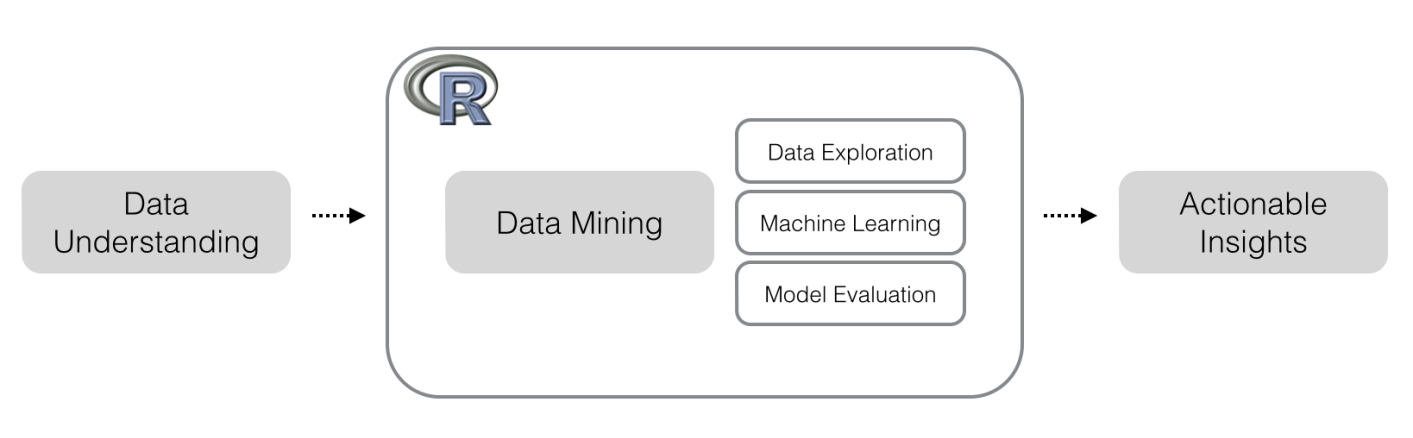
Table of Contents

The Goal

Goal

- To predict which customers will churn.

Workflow

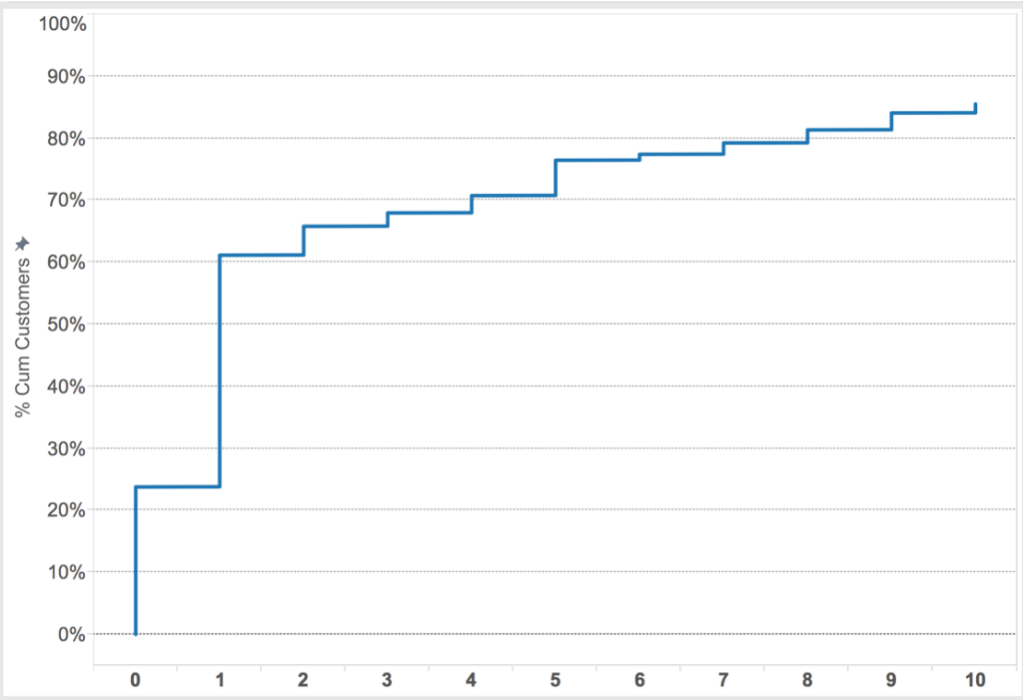


Customer renewal data

customer_id	is_churn	days_renew
213	FALSE	5
102	TRUE	NULL
31	TRUE	NULL
921	FALSE	5
…	…	…

- **Days_Renew** is the number of days after subscription expiration before the customer renewed.
- If the customer has churned, then this value will be NULL.

Customer Renewal Schedule

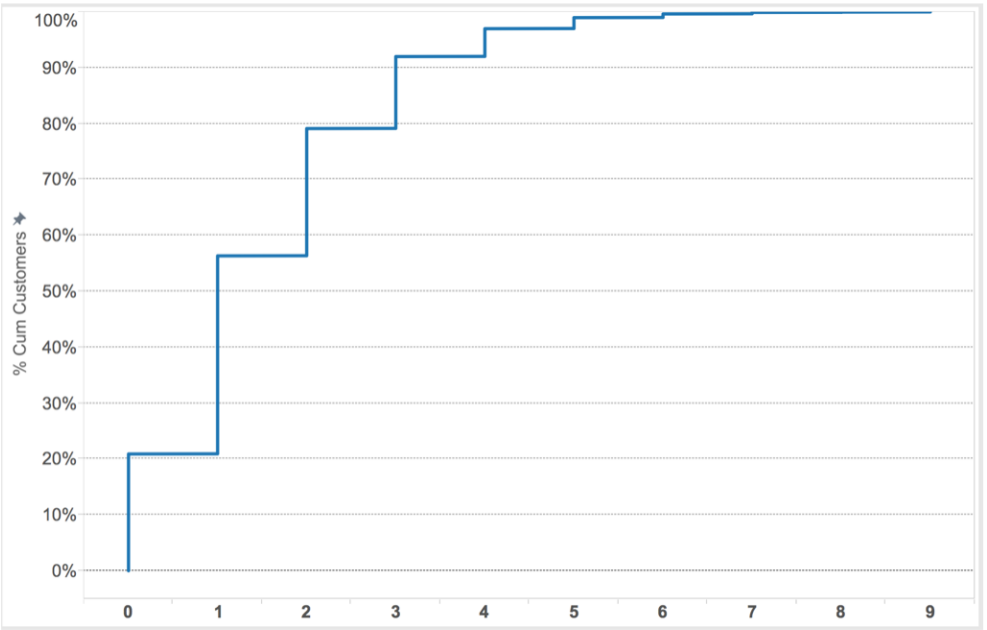


Customer service call data

customer_id	cust_surv_calls
213	0
102	3
31	1
921	2
…	…

- 80% of customers made at most 2 phone calls to the customer service.
- On average, customers made 1.6 customer service calls.
- Some customers made as many as 9 calls to customer service.

Customer Call
Cumulative Distribution



Customer plan type and demographic

customer_id	state	account_en	area_code	is_intl_plan	is_vmail_plan
1	KS	128	415	no	yes
2	OH	107	415	no	yes
3	NJ	137	415	no	no
4	OH	84	408	yes	no
5	OK	75	415	yes	no
6	AL	118	510	yes	no
7	MA	121	510	no	yes
8	MO	147	415	yes	no
9	LA	117	408	no	no
10	WV	141	415	yes	yes
…	…	…	…	…	…

Customer usage data

customer_id	vmail_messages	day_mins	day_calls	day_charge	eve_mins	eve_calls	eve_charge	night_mins	night_calls	night_charge	intl_mins	intl_calls	intl_charge
1	25	265.1	110	45.07	197.4	99	16.78	244.7	91	11.01	10	3	2.7
2	26	161.6	123	27.47	195.5	103	16.62	254.4	103	11.45	13.7	3	3.7
3	0	243.4	114	41.38	121.2	110	10.3	162.6	104	7.32	12.2	5	3.29
4	0	299.4	71	50.9	61.9	88	5.26	196.9	89	8.86	6.6	7	1.78
5	0	166.7	113	28.34	148.3	122	12.61	186.9	121	8.41	10.1	3	2.73
6	0	223.4	98	37.98	220.6	101	18.75	203.9	118	9.18	6.3	6	1.7
7	24	218.2	88	37.09	348.5	108	29.62	212.6	118	9.57	7.5	7	2.03
8	0	157	79	26.69	103.1	94	8.76	211.8	96	9.53	7.1	6	1.92
9	0	184.5	97	31.37	351.6	80	29.89	215.8	90	9.71	8.7	4	2.35
10	37	258.6	84	43.96	222	111	18.87	326.4	97	14.69	11.2	5	3.02
…	…	…	…	…	…	…	…	…	…	…	…	…	…

Churn data structure

```
$ customer_id    : int  1 2 3 4 5 6 7 8 9 10 ...
$ state          : Factor w/ 51 levels "AK","AL","AR",..: 17 36 32 36 37 2 20 25 19 50 ...
$ account_len    : int  128 107 137 84 75 118 121 147 117 141 ...
$ area_code      : int  415 415 415 408 415 510 510 415 408 415 ...
$ phone          : Factor w/ 3333 levels "327-1058","327-1319",..: 1927 1576 1118 1708 111 2254 1048 81
$ is_intl_plan   : Factor w/ 2 levels "no","yes": 1 1 1 2 2 2 1 2 1 2 ...
$ is_vmail_plan  : Factor w/ 2 levels "no","yes": 2 2 1 1 1 1 2 1 1 2 ...
$ vmail_messages : int  25 26 0 0 0 0 24 0 0 37 ...
$ day_mins       : num  265 162 243 299 167 ...
$ day_calls      : int  110 123 114 71 113 98 88 79 97 84 ...
$ day_charge     : num  45.1 27.5 41.4 50.9 28.3 ...
$ eve_mins       : num  197.4 195.5 121.2 61.9 148.3 ...
$ eve_calls      : int  99 103 110 88 122 101 108 94 80 111 ...
$ eve_charge     : num  16.78 16.62 10.3 5.26 12.61 ...
$ night_mins     : num  245 254 163 197 187 ...
$ night_calls    : int  91 103 104 89 121 118 118 96 90 97 ...
$ night_charge   : num  11.01 11.45 7.32 8.86 8.41 ...
$ intl_mins      : num  10 13.7 12.2 6.6 10.1 6.3 7.5 7.1 8.7 11.2 ...
$ intl_calls     : int  3 3 5 7 3 6 7 6 4 5 ...
$ intl_charge    : num  2.7 3.7 3.29 1.78 2.73 1.7 2.03 1.92 2.35 3.02 ...
$ cust_surv_calls: int  1 1 0 2 3 0 3 0 1 0 ...
$ is_churn       : Factor w/ 2 levels "False.","True.": 1 1 1 1 1 1 1 1 1 1 ...
$ days_renew     : int  0 0 0 0 0 0 0 0 0 0 ...
```

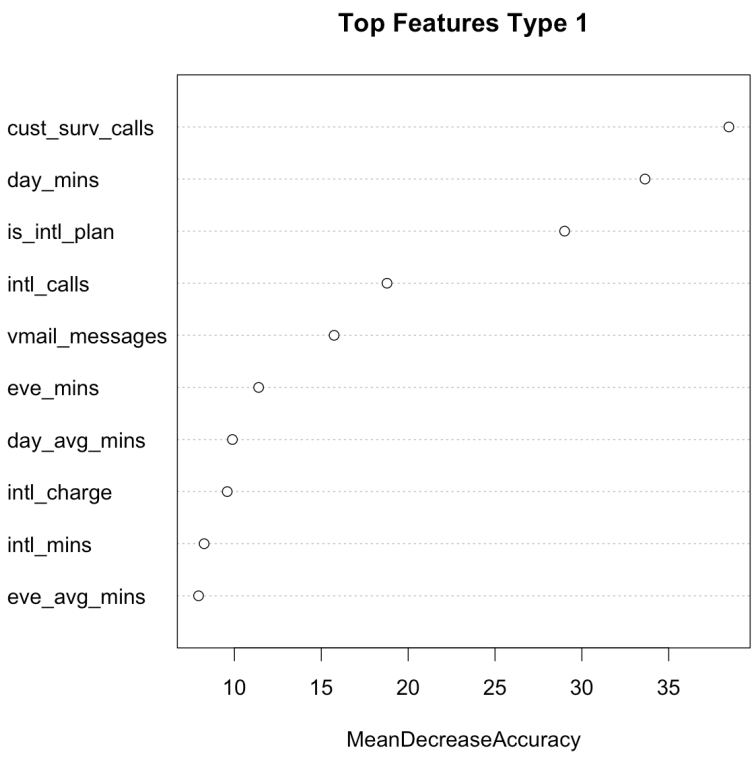
Table of Contents

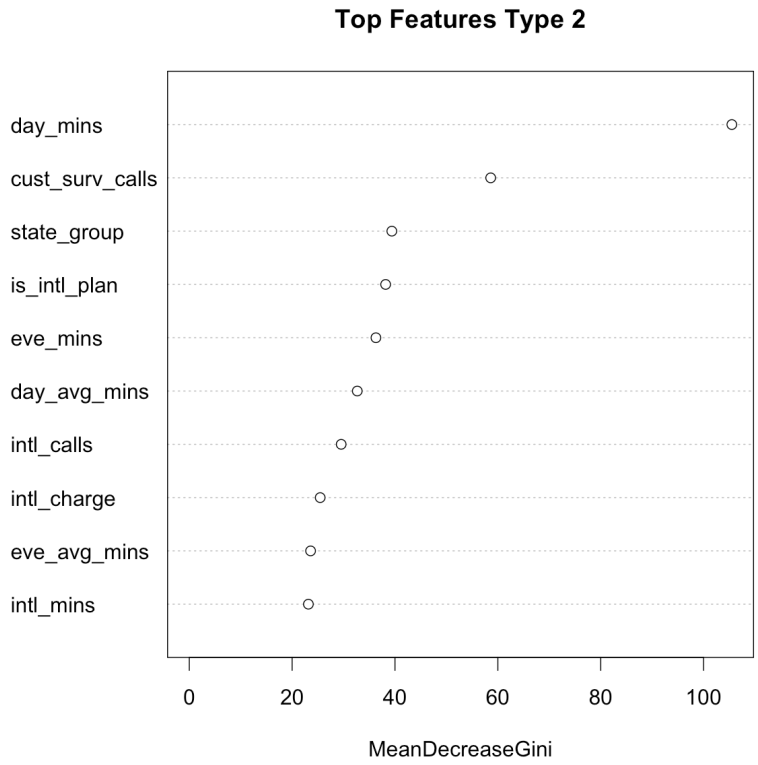
R Script

https://github.com/vieplivee/Data-Science-in-Action/blob/master/src/churn.R

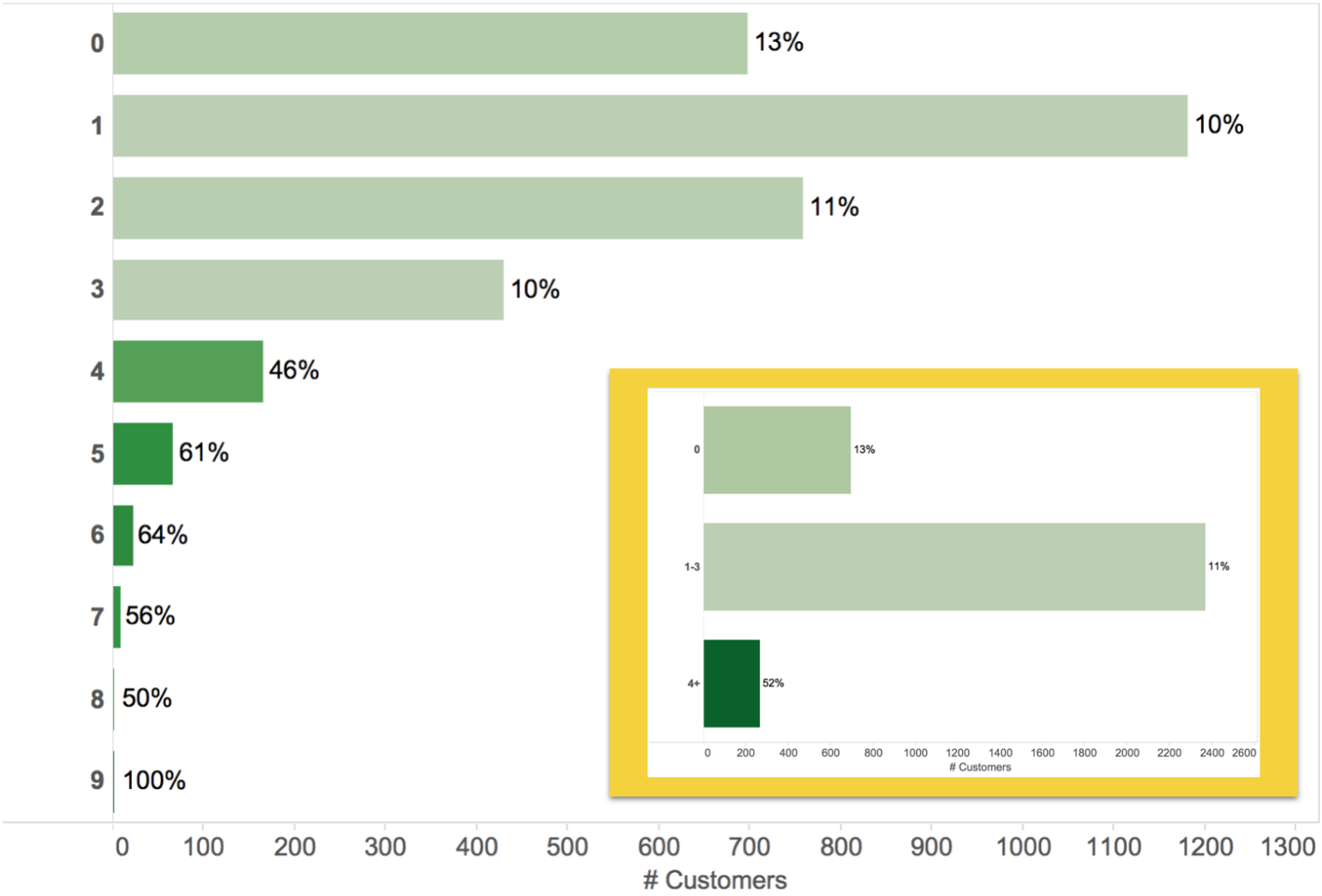
Table of Contents

Variable importance from random forest

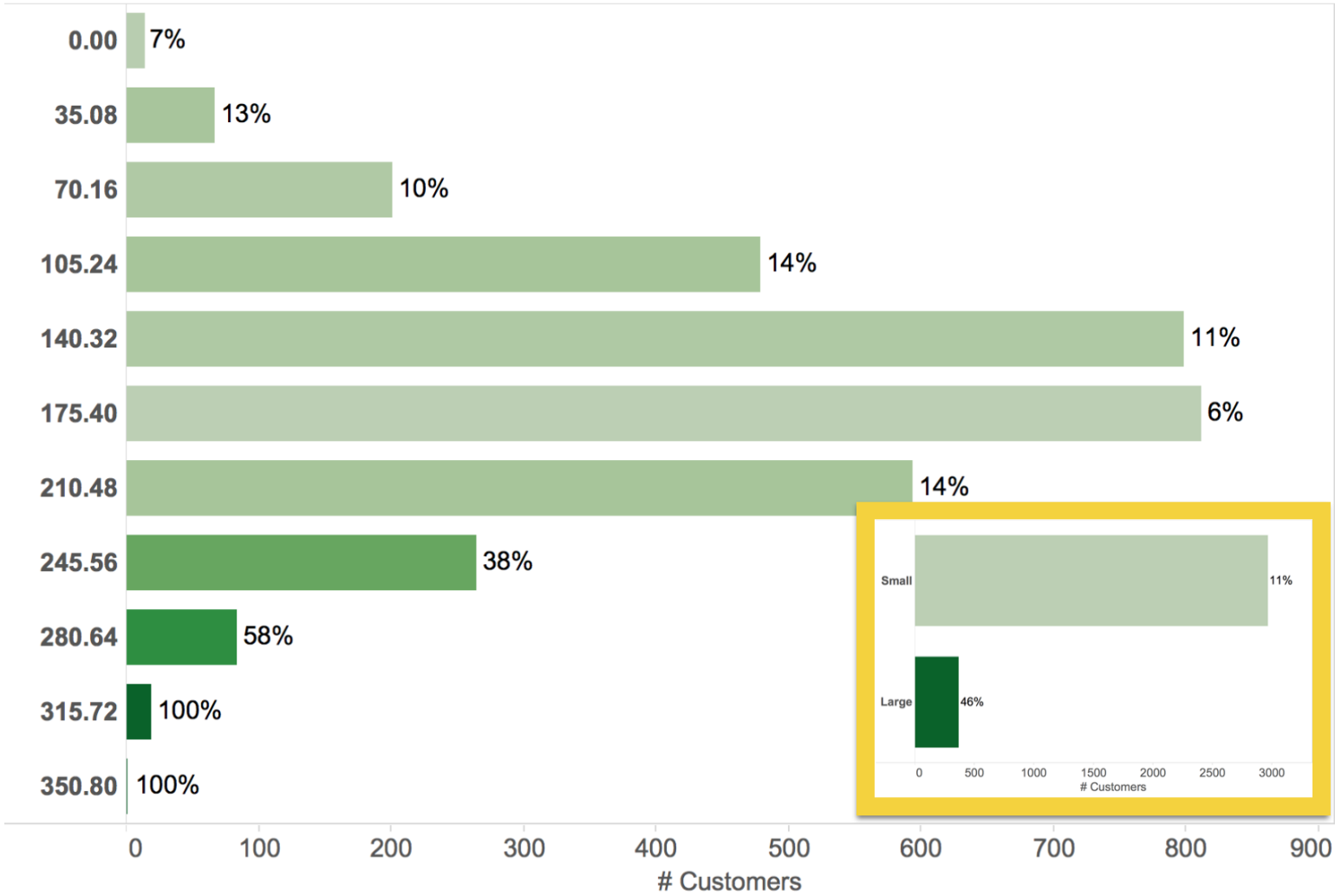




Top feature - customer service calls



Top feature - customer day time minutes



Top feature - international plan and international calls

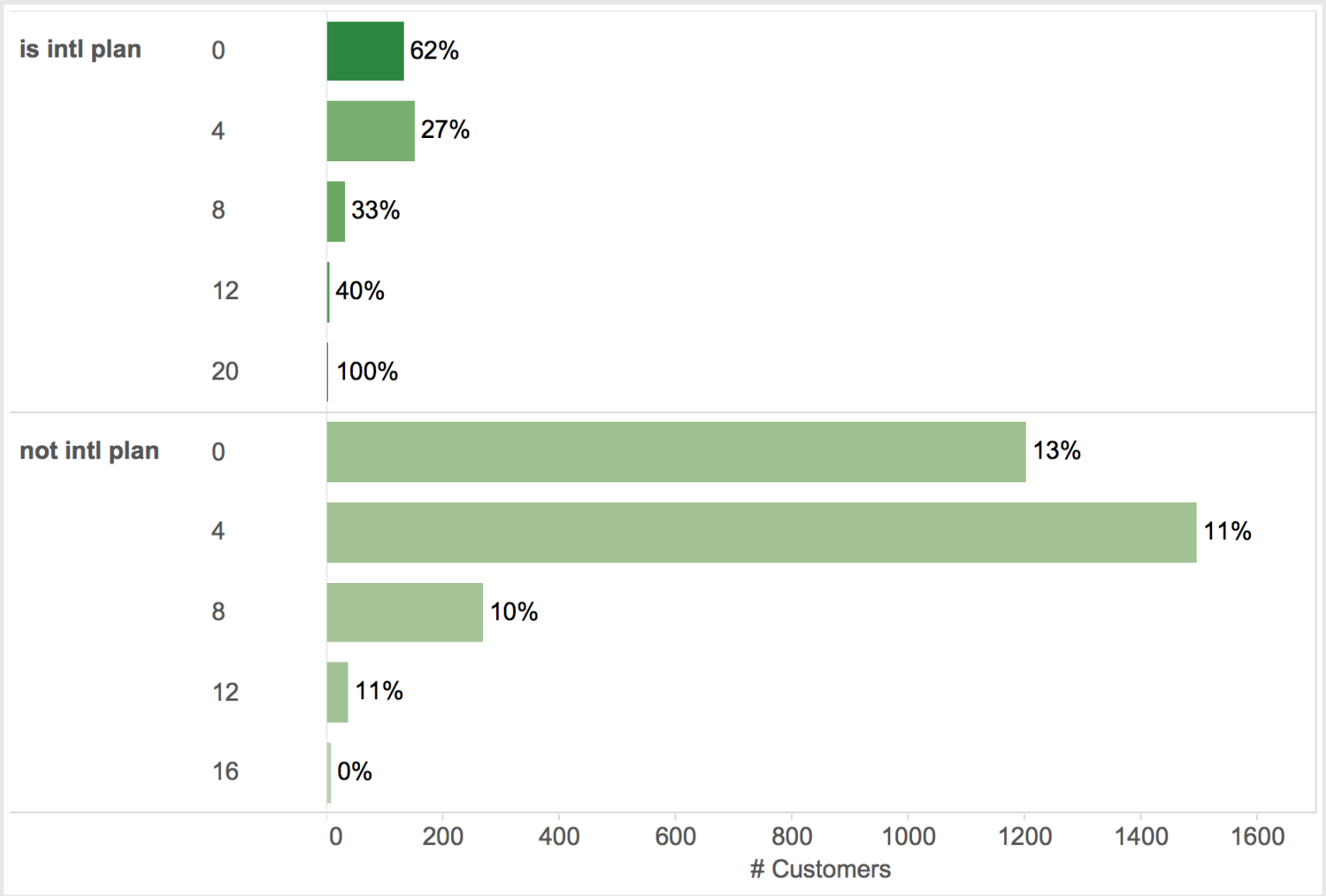


Table of Contents

Churn analysis

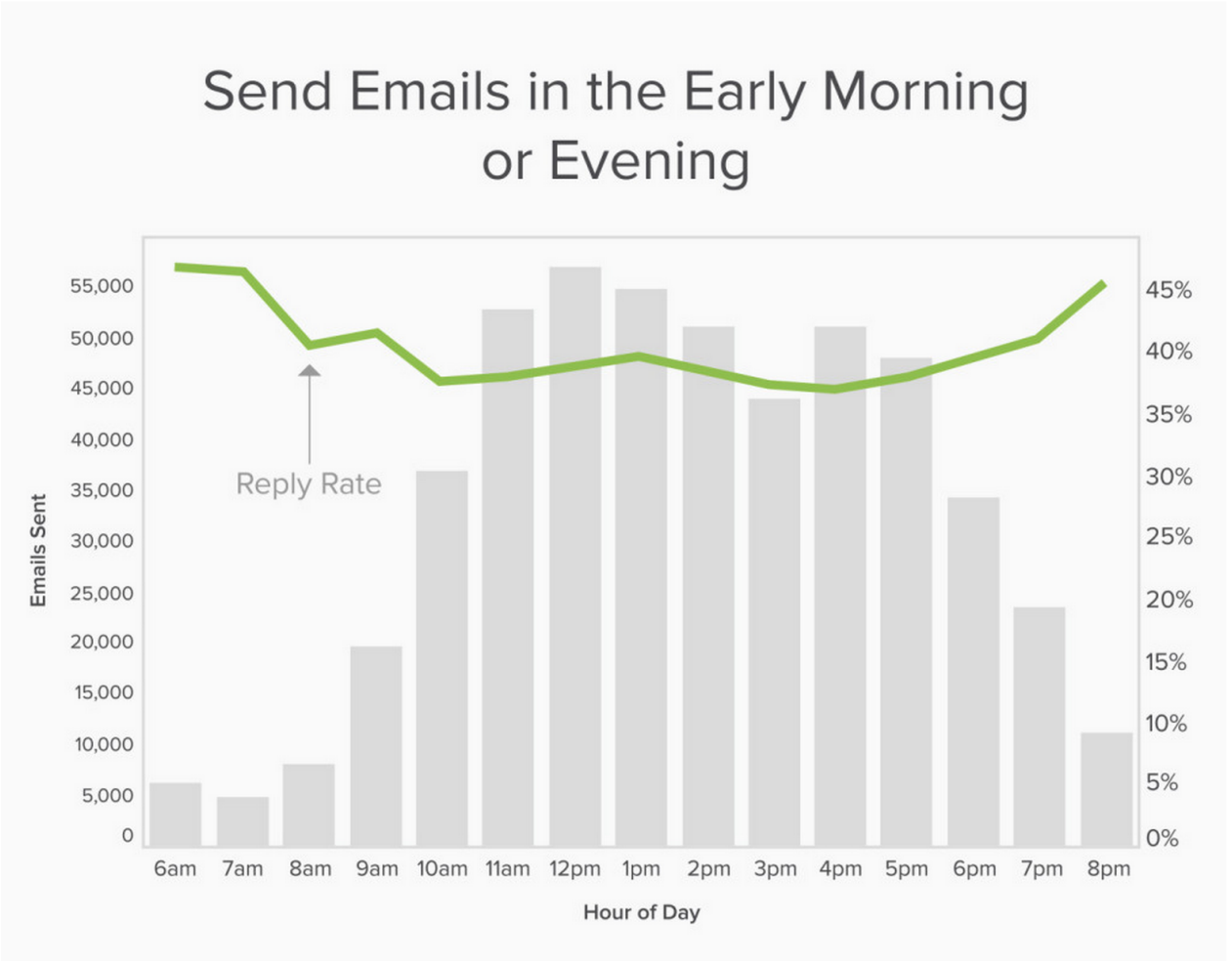
Is Intl Plan	Cust Surv Calls (group)	Day Mins Group	% Churn	Number of Records	Summary
is intl plan	0	Large	60%	15	Small Group
		Small	44%	68	Small Group
	1-3	Large	56%	36	Small Group
		Small	34%	176	
	4+	Large	100%	1	Outlier
		Small	67%	27	Small Group
not intl plan	0	Large	48%	60	Small Group
		Small	4%	554	Low Churn
	1-3	Large	43%	229	
		Small	4%	1,928	Low Churn
	4+	Large	44%	27	Small Group
		Small	50%	212	
Grand Total			**14%**	**3,333**	

Churn analysis

Is Intl Plan	Cust Surv Calls (group)	Day Mins Group	% Churn	# Customers	Summary	Avg Charge	Revenue if we got 10% of the churned customers back
is intl plan	0	Large	60%	15	Small Group	$ 72	$ 65
		Small	44%	68	Small Group	$ 56	$ 168
	1-3	Large	56%	36	Small Group	$ 72	$ 144
		Small	34%	176		$ 55	$ 325
	4+	Large	100%	1	Outlier	$ 78	$ 8
		Small	67%	27	Small Group	$ 56	$ 101
not intl plan	0	Large	48%	60	Small Group	$ 73	$ 212
		Small	4%	554	Low Churn	$ 55	$ 132
	1-3	Large	43%	229		$ 72	$ 706
		Small	4%	1,928	Low Churn	$ 55	$ 418
	4+	Large	44%	27	Small Group	$ 74	$ 89
		Small	50%	212		$ 54	$ 578
Grand Total			**14%**	**3,333**		$ 57	$ 2,660

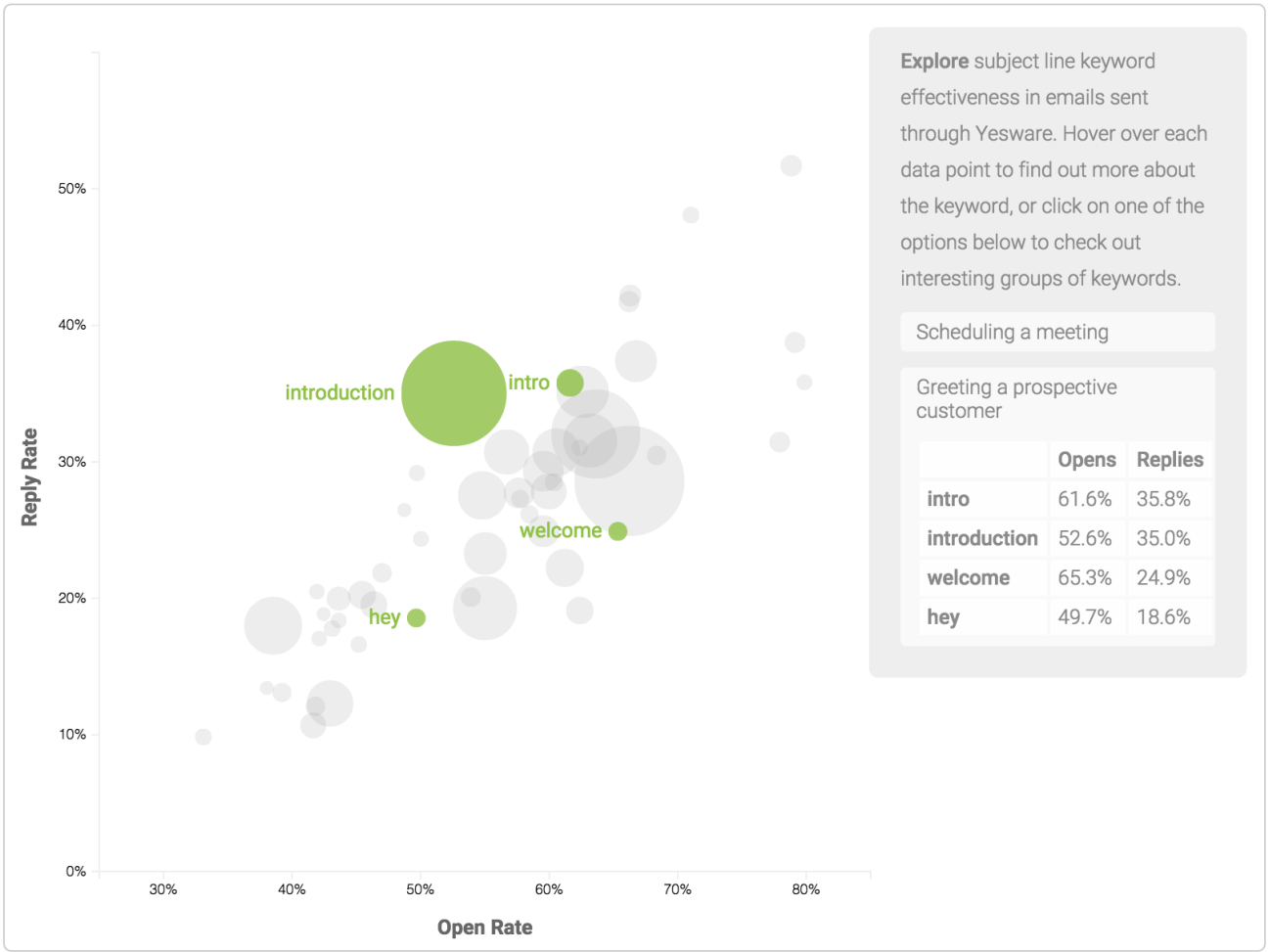
Table of Contents

When to send an email



Best time to send email: `http://goo.gl/lVdD31`

D3 Visualization: Subject Line Keywords



Subject line key words: `http://goo.gl/PK9xh0`

Table of Contents

From academia to industry

	Academia	Industry
Goal	Improve human knowledge	Make money
Success Criteria	Publish papers	Create and deliver business value
Approach	Finding a better way to do a hard thing	Finding the fastest way to do easy things
Importance of Speed	**Not as important as other things**	**Very important**

Tools that help you do data science fast

	Python	R	Unix	SQL	Scala
Powerful Packages / Library	★★★★★	★★★★★	★★	★	★★★★
Community Support	★★★★★	★★★★★	★★★★	★★★	★★★★
Data Munging	★★★★	★★★	★★★★	★★★★	★★★★
Data Exploration	★★★★	★★★★★	★★	★★★	★★★
Machine Learning	★★★★	★★★★★	★	★★	★★★★

Data visualization tools

	Excel	R	Tableau	D3
Ease of Learning	★★★★★	★★★	★★★★★	★
Is Free	No	Yes	No	Yes
Good for Data Exploration	★★	★★★★	★★★★★	★★★
Flexibility in Data Representation	★★	★★★★	★★★★	★★★★★
Good for Reporting and Sharing	★★★★	★★★★	★★★	★★★★

D3 `http://d3js.org/`

Table of Contents

Data Democracy

> *Data provides us a window into the world, and we should not deny people access to that world just because they don't know how to use certain tools – this is not democratization.*

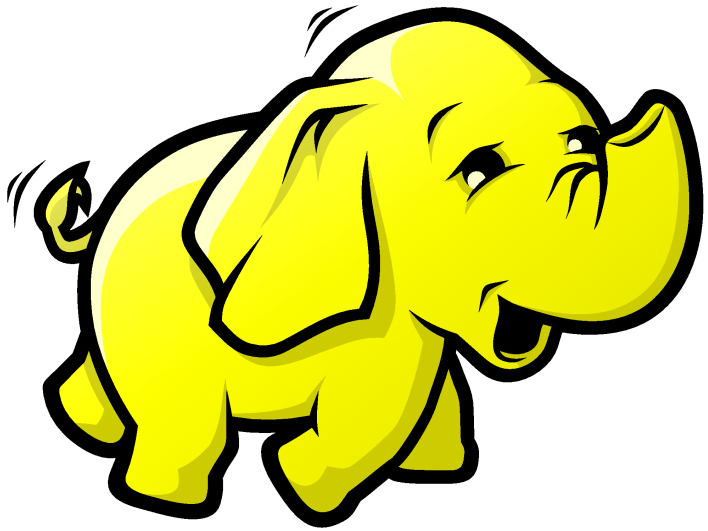
Democratization requires that we provide people with an easy way to understand the data. It requires sharing information in a form that everyone can read and understand. It requires timely communication about what is happening in a relevant and personal way. It means giving people the stories that are trapped in the data so they can do something with the information.

http://www.narrativescience.com/blog/democratization-data/

Big Data

Big Data Ecosystem

- Hadoop - file system



Big Data

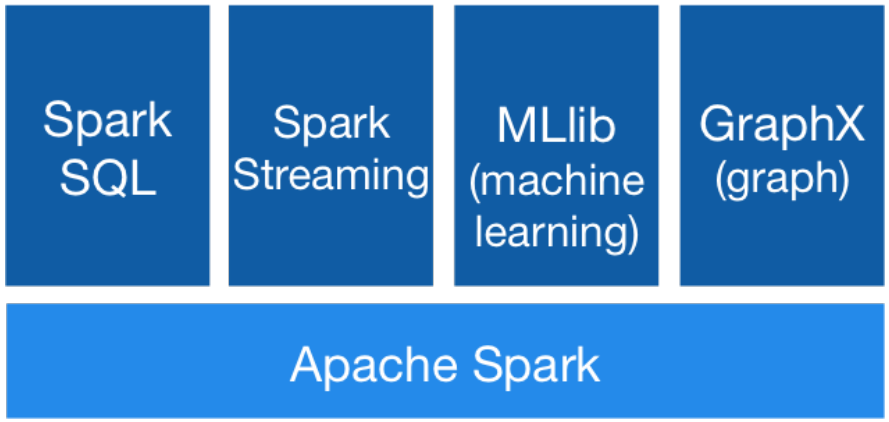
Big Data Ecosystem

- Hadoop - file system
- Spark - computing system

Big Data

Big Data Ecosystem

- Hadoop - file system
- Spark - computing system
- Spark Stack

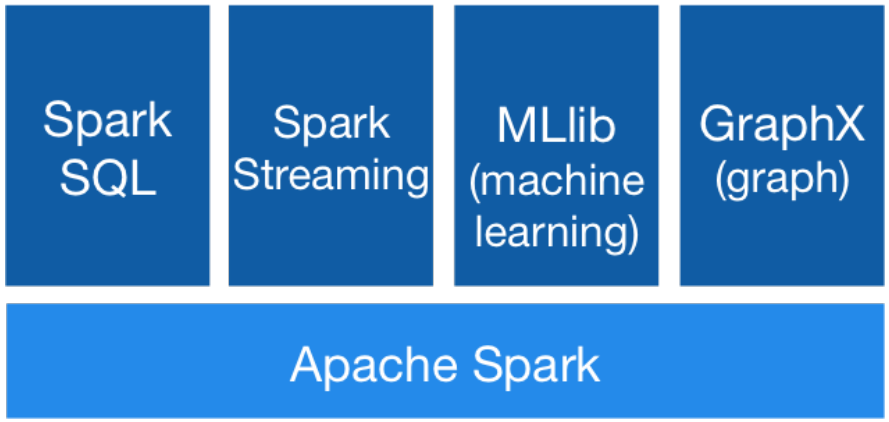


`https://spark.apache.org/`

Big Data

Big Data Ecosystem

- Hadoop - file system
- Spark - computing system
- Spark Stack
  - Spark SQL - Data Munging
  - Spark Streaming - Real Time Processing
  - MLlib - Machine Learning
  - GraphX - Visualization



`https://spark.apache.org/`

Thank You

Please send your questions and feedback to me!