

# Making Sense of Text

Data Science Symposium

Braga, March, 28th 2019

**amplemarket**

*“ pretty much anything you could do with a second of thought, we can probably now or soon automate using supervised learning, using this input-output mapping. ”*

**Andrew Ng**

Email:

mica@amplemarket.com

Education:

Physics

Occupation:

Co-founder @ Amplemarket and Fermat's Library

Amplemarket: Virtual Assistant for Sales  
Professionals

Fermat's Library: Platform to Illuminate  
Academic Papers

Backed by:



Combinator

# amplemarket

[amplemarket.com](https://amplemarket.com)

## FERMAT'S LIBRARY

[fermatlibrary.com](https://fermatlibrary.com)



San Francisco



Lisbon

# Email Automation? Why?

# Data!

Email is the preferred business communication method.

3 Billion + active email addresses.

**235 Billion Emails are exchanged daily!**



## Daily Email Volume

2015: 205.6

2016: 215.3

2017: 225.3

2018: 235.6

2019: 246.5

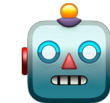
\*emails/day (in billion)

# ctrl-c, ctrl-v

Most of business email is **support and sales related**.

These teams spend most of their time on email and **deal with 100s of emails on daily basis**.

They are **copying/pasting snippets of text** and filling in the blanks.



Support and Sales Professionals already behave like robots.

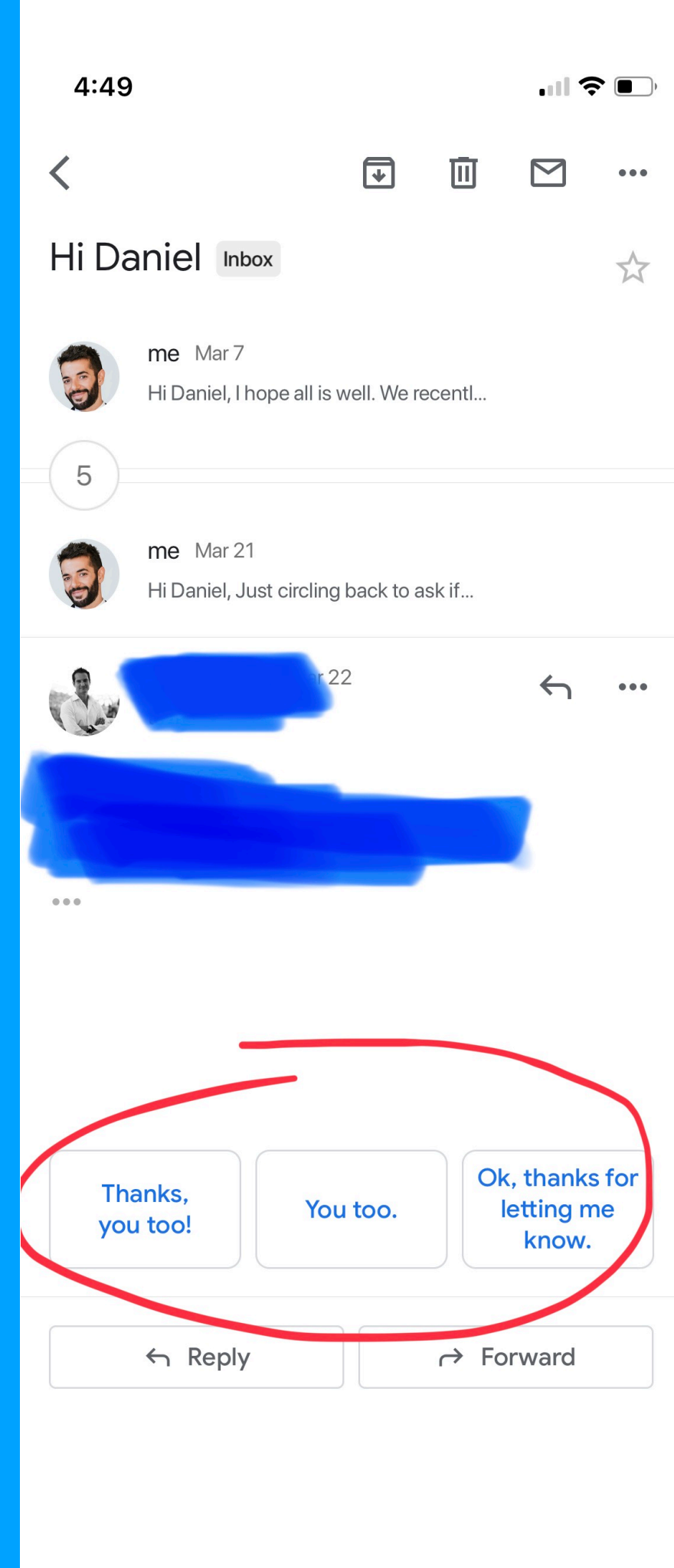
Most of their tasks add no value. Why not automate them?

# It's already here

Smart replies constitute **10% of all messages** sent over Gmail. (Set 2018)

Google and LinkedIn users are already using smart replies.

Leverage recent ML developments and open source projects.



# The Dataset.

### Out Of Office

I am currently out of the office. I will check email as I'm able to during this time. Please call my cell or John Doe at (123) 456-7890 for anything time sensitive.

Hi John! I'm on vacation right now. I'm back on June 25th.

Thank you for your email! I am traveling for business and have limited access to my email. If you have anything urgent, please send me an sms (+123 456 7890).

👉 Different emails with different wording have the same meaning!



### Out Of Office

I am currently out of the office. I will check email as I'm able to during this time. Please call my cell or John Doe at (123) 456-7890 for anything time sensitive.

Hi John! I'm on vacation right now. I'm back on June 25th.

Thank you for your email! I am traveling for business and have limited access to my email. If you have anything urgent, please send me an sms (+123 456 7890).

👉 Different emails with different wording have the same meaning!

👉 Can we build a finite number of categories where we can group emails?

### Out Of Office

I am currently out of the office. I will check email as I'm able to during this time. Please call my cell or John Doe at (123) 456-7890 for anything time sensitive.

Hi John! I'm on vacation right now. I'm back on June 25th.

Thank you for your email! I am traveling for business and have limited access to my email. If you have anything urgent, please send me an sms (+123 456 7890).

👉 Different emails with different wording have the same meaning!

👉 Can we build a finite number of categories where we can group emails?

👉 Unique dataset of mappings to finite number of categories:

Out Of Office

Interested

Not Interested

Introduction

Product

Pricing

Meeting

Circle Back

**The Model.**

**Supervised Machine Learning**  
on unique dataset of text to  
class mappings.

**Data pre-processing:**

- ✓ Tokenize
- ✓ Stemming
- ✓ Vectorize
- ✓ TF-IDF

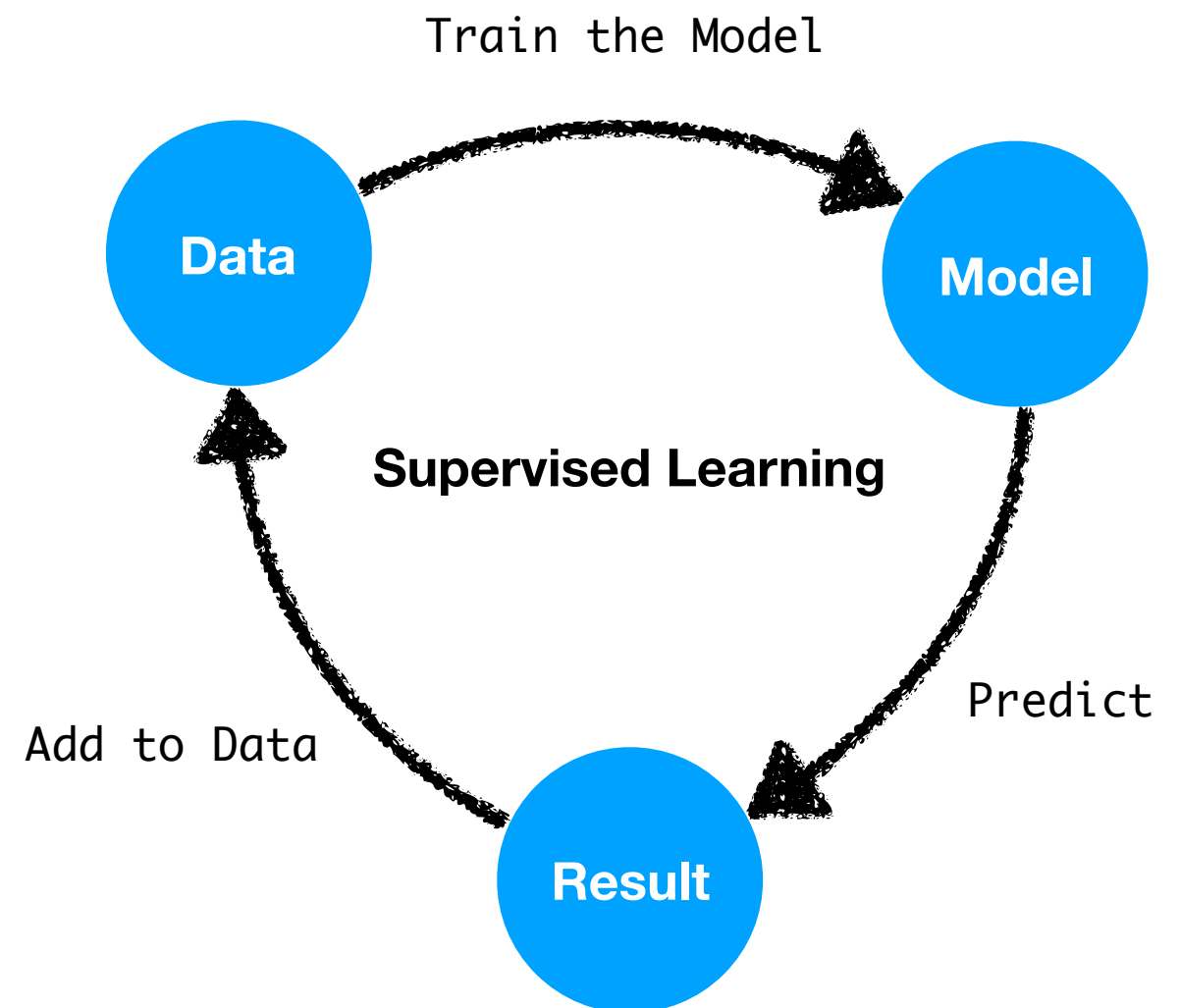
**Supervised Machine Learning**  
on unique dataset of text to  
class mappings.

**Data pre-processing:**

- ✓ Tokenize
- ✓ Stemming
- ✓ Vectorize
- ✓ TF-IDF

**Create and Train** a model with  
Supervised Learning.

Apply model to predict the class  
of an unseen email.



**An example.**

from: john@company.com

to: mike@companyB.com

email:

I'd love to learn more about your company.  
Can we schedule a quick call next week  
sometime? My schedule is pretty flexible  
every day except Monday afternoon and Tuesday  
morning.

from: john@company.com  
to: mike@companyB.com  
email:

I'd love to learn more about your company.  
Can we schedule a quick call next week  
sometime? My schedule is pretty flexible  
every day except Monday afternoon and Tuesday  
morning.



[love learn more about your compani can we schedul  
quick call next week sometim my schedul pretti  
flexibl everi day except monday afternoon and  
tuesday morn]



from: john@company.com  
to: mike@companyB.com  
email:

I'd love to learn more about your company.  
Can we schedule a quick call next week  
sometime? My schedule is pretty flexible  
every day except Monday afternoon and Tuesday  
morning.



[love learn more about your compani can we schedul  
quick call next week sometim my schedul pretti  
flexibl everi day except monday afternoon and  
tuesday morn]



{Vectorize, TF-IDF}

from: john@company.com  
to: mike@companyB.com  
email:

Interested

I'd love to learn more about your company.  
Can we schedule a quick call next week  
sometime? My schedule is pretty flexible  
every day except Monday afternoon and Tuesday  
morning.

[love learn more about your compani can we schedul  
quick call next week sometim my schedul pretti  
flexibl everi day except monday afternoon and  
tuesday morn]

{Vectorize, TF-IDF}

Classify

Result: Interested

Probabilities for each class:

```
{  
  "circle_back_later":0.0002683432087,  
  "forwarded_email":0.0001808586523,  
  "hard_no":0.000901196182,  
  "interested":0.997172092,  
  "introduction":4.60500148e-05,  
  "not_interested":0.0001858046449,  
  "not_the_right_person":0.000904631910,  
  "ooo":0.0003410228265,  
}
```

**Looking at strings.**

“Let’s grab coffee tomorrow afternoon at  
10 AM at Sightglass on 7th Street.”

Possible Features:

- PERSONS
- ORGANIZATIONS
- LOCATIONS
- DATES
- TIMES
- QUANTITIES
- MONETARY VALUES
- PERCENTAGES

## Meeting

“Let’s grab coffee tomorrow afternoon at  
10 AM at Sightglass on 7th Street”

DATE  
TIME  
ORG  
LOC





Possible Features:

- PERSONS
- ORGANIZATIONS
- LOCATIONS
- DATES
- TIMES
- QUANTITIES
- MONETARY VALUES
- PERCENTAGES

## Meeting

“Let’s grab coffee <sup>DATE</sup> tomorrow afternoon at  
<sup>TIME</sup> 10 AM at <sup>ORG</sup> Sightglass on <sup>LOC</sup> 7th Street ”



-  Coffee John <> Mica  
Saturday, Mar 30 · 10–11:10 AM
-  Sightglass Coffee  
270 7th St, San Francisco, CA 94103, USA
-  10 minutes before
-  mica@amplemarket.com

8:02



≡ March ▾



SAT  
30

9 AM

10 AM

Coffee John <> Mica  
Sightglass Coffee

11 AM

12 PM

**A Challenge.**

# Learn by doing. 🧐

**Challenge:** create a classifier for support tickets (clean data set in PT).

1. Read “Multi Class Text Classification”: <https://towardsdatascience.com/multi-class-text-classification-with-scikit-learn-12f1e60e0a9f>
2. Source Code: [https://github.com/susanli2016/Machine-Learning-with-Python/blob/master/Consumer\\_complaints.ipynb](https://github.com/susanli2016/Machine-Learning-with-Python/blob/master/Consumer_complaints.ipynb)
3. Try your model on this data set: <https://www.kaggle.com/samuelhei/dataset-for-text-tagging-phone-company-ptbr/version/4>

Feel free to email me at [mica@amplemarket.com](mailto:mica@amplemarket.com), happy to help!

1	é consertar o meu telefone	reparar-linha
2	como faço para ter a linha controle oitenta	comprar-linha
3	não estou conseguindo fazer uma ligação	reparar-linha
4	é adquirir uma linha cadeada	comprar-linha
5	não o problema é o chiado no meu telefone	reparar-linha
6	quero saber um problema com uma linha	reparar-linha
7	é fazer o reparo da minha linha telefônica	reparar-linha
8	como faço para consertar o telefone	reparar-linha



# Thank you!

[mica@amplemarket.com](mailto:mica@amplemarket.com)

**amplemarket**

# Other challenges...

**amplemarket**

Look-a-like audiences

Parsing text data

Template conversions

FERMAT'S LIBRARY

Reference Extraction

Paper recommendations