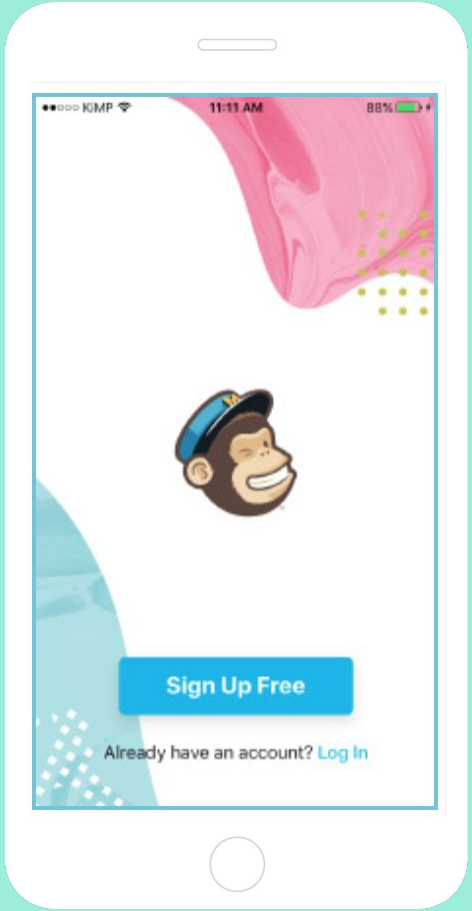
The Evolution of a Data Science App

Matt Mastin, PhD | Data Science

MailChimp

The world’s largest marketing automation platform.

- Email
- Digital Ads
- Automation



Data Science at MailChimp

We research, design, and build systems that provide data science services to MailChimp.

- User facing features:
  - Send Time Optimization
  - Predicted Demographics
  - Audience Expansion
  - Product Recommendation

Data Science at MailChimp

We research, design, and build systems that provide data science services to MailChimp.

- User facing features:
  - Send Time Optimization
  - Predicted Demographics
  - Audience Expansion
  - Product Recommendation

Data Science at MailChimp

We research, design, and build systems that provide data science services to MailChimp.

- Compliance:
  - Bot detection
  - List quality prediction

This is the story...

of an early data science win and a few lessons we have learned as MailChimp has grown.

This is the story...

of an early data science win and a few lessons we have learned as MailChimp has grown.

Spoiler Alert: The moral is that data engineering is important.

New Business

Forever

Free

Create beautiful, professional campaigns and marketing automations for free—no design or coding experience necessary. It's so easy, you can start sending today.

Learn More →

On Sept. 1, 2009 MailChimp began offering free accounts.

- Great for small businesses
  - Democratization of marketing tools
- Great for MailChimp
  - 5x increase in users in 1 year
- Potentially great for spammers
  - How do we stop bad actors?

New Business

Forever

Free

Create beautiful, professional campaigns and marketing automations for free—no design or coding experience necessary. It's so easy, you can start sending today.

Learn More →

On Sept. 1, 2009 MailChimp began offering free accounts.

- Great for small businesses
  - Democratization of marketing tools
- Great for MailChimp
  - 5x increase in users in 1 year
- Potentially great for spammers
  - How do we stop bad actors?

New Business

Forever

Free

Create beautiful, professional campaigns and marketing automations for free—no design or coding experience necessary. It's so easy, you can start sending today.

Learn More →

On Sept. 1, 2009 MailChimp began offering free accounts.

- Great for small businesses
  - Democratization of marketing tools
- Great for MailChimp
  - 5x increase in users in 1 year
- Potentially great for spammers
  - How do we stop bad actors?

New Business

Forever

Free

Create beautiful, professional campaigns and marketing automations for free—no design or coding experience necessary. It's so easy, you can start sending today.

Learn More →

On Sept. 1, 2009 MailChimp began offering free accounts.

- Great for small businesses
  - Democratization of marketing tools
- Great for MailChimp
  - 5x increase in users in 1 year
- Potentially great for spammers
  - How do we stop bad actors?

An Example of Bad Behavior

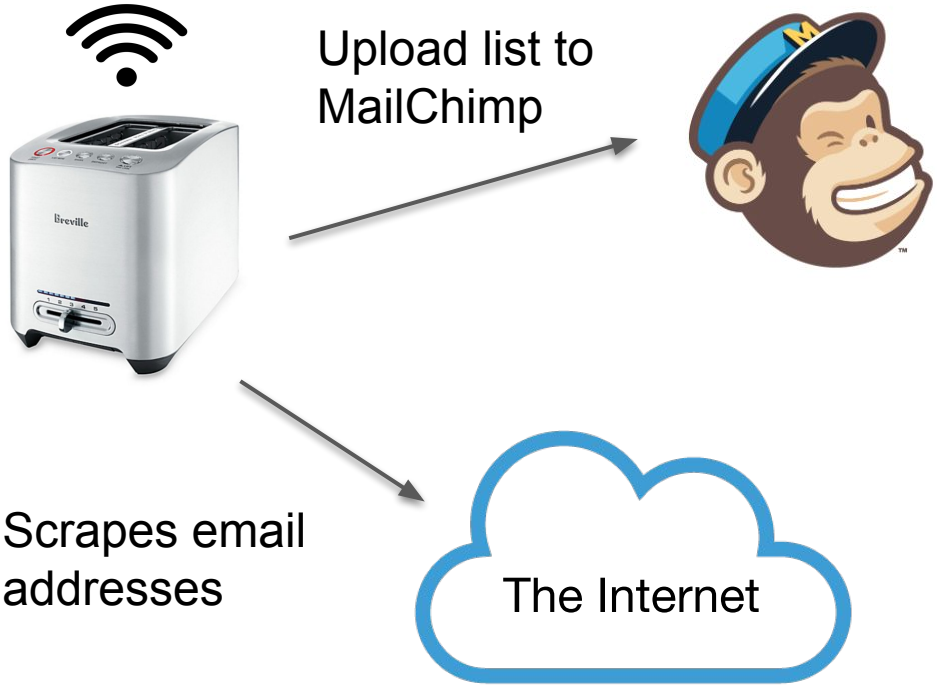
An Example of Bad Behavior

An Example of Bad Behavior

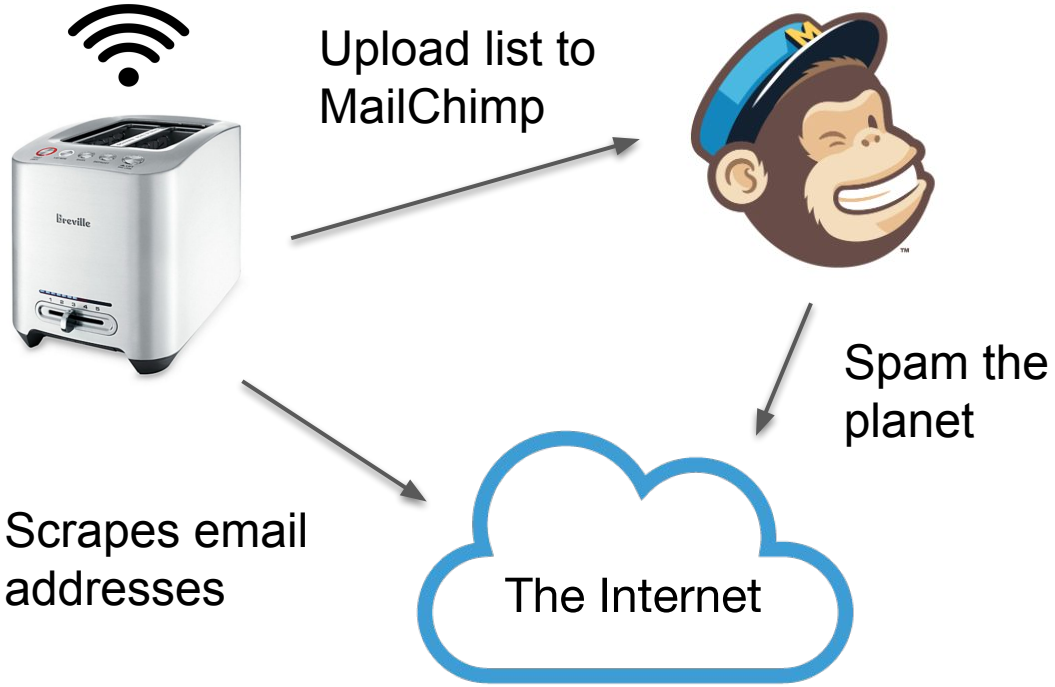
Scrapes email addresses

The Internet

An Example of Bad Behavior



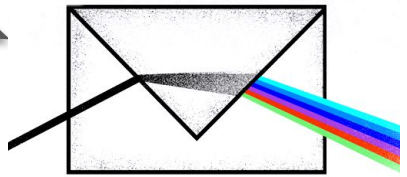
An Example of Bad Behavior



The Email Genome Project (EGP)

Upload list to MailChimp

Scan the list



Scrapes email addresses

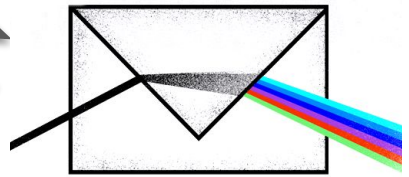
The Internet

We like logos, this is EGP's

The Email Genome Project (EGP)

Upload list to MailChimp

Scan the list



Don’t spam the planet

Scrapes email addresses

The Internet

We like logos, this is EGP’s

What does EGP do?

Predict the quality of a list

- How “stale” is the list?
  - Distribution of email age
  - Frequency of delivery attempts
- How engaged is the list in terms of opens and clicks?
- Similarity to known bad lists
- Bounce Rate
  - What percentage of the emails on the list will fail to deliver?

What does EGP do?

Predict the quality of a list

- How “stale” is the list?
  - Distribution of email age
  - Frequency of delivery attempts
- How engaged is the list in terms of opens and clicks?
- Similarity to known bad lists
- Bounce Rate
  - What percentage of the emails on the list will fail to deliver?

What does EGP do?

Predict the quality of a list

- How “stale” is the list?
  - Distribution of email age
  - Frequency of delivery attempts
- How engaged is the list in terms of opens and clicks?
- Similarity to known bad lists
- Bounce Rate
  - What percentage of the emails on the list will fail to deliver?

What does EGP do?

Predict the quality of a list

- How “stale” is the list?
  - Distribution of email age
  - Frequency of delivery attempts
- How engaged is the list in terms of opens and clicks?
- Similarity to known bad lists
- Bounce Rate
  - What percentage of the emails on the list will fail to deliver?

What does EGP do?

Predict the quality of a list

- How “stale” is the list?
  - Distribution of email age
  - Frequency of delivery attempts
- How engaged is the list in terms of opens and clicks?
- Similarity to known bad lists
- Bounce Rate
  - What percentage of the emails on the list will fail to deliver?

What does EGP do?

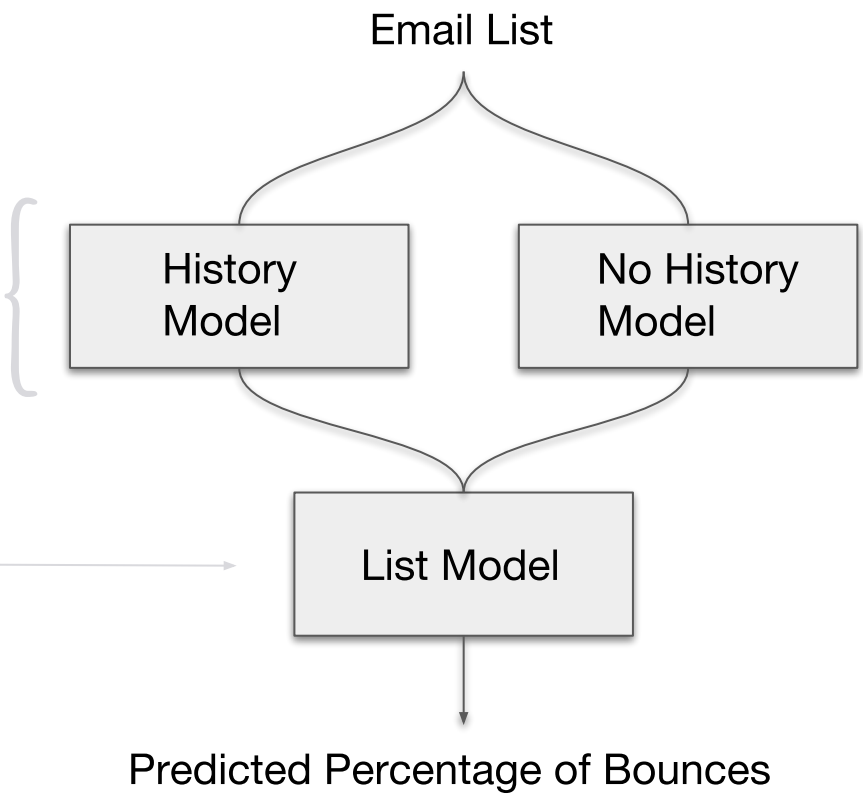
Predict the quality of a list

- How “stale” is the list?
  - Distribution of email age
  - Frequency of delivery attempts
- How engaged is the list in terms of opens and clicks?
- Similarity to known bad lists
- Bounce Rate
  - What percentage of the emails on the list will fail to deliver?

Bounce Prediction

Three random forests

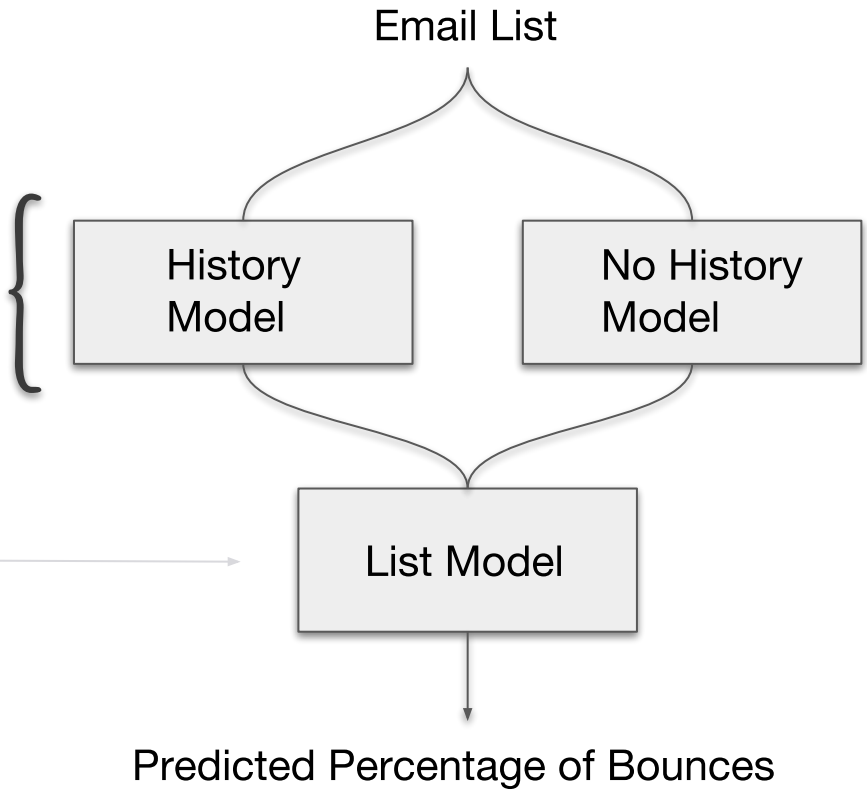
- Predict the probability of a bounce or delivery for each address
- Predict the percentage of addresses on a list that will bounce



Bounce Prediction

Three random forests

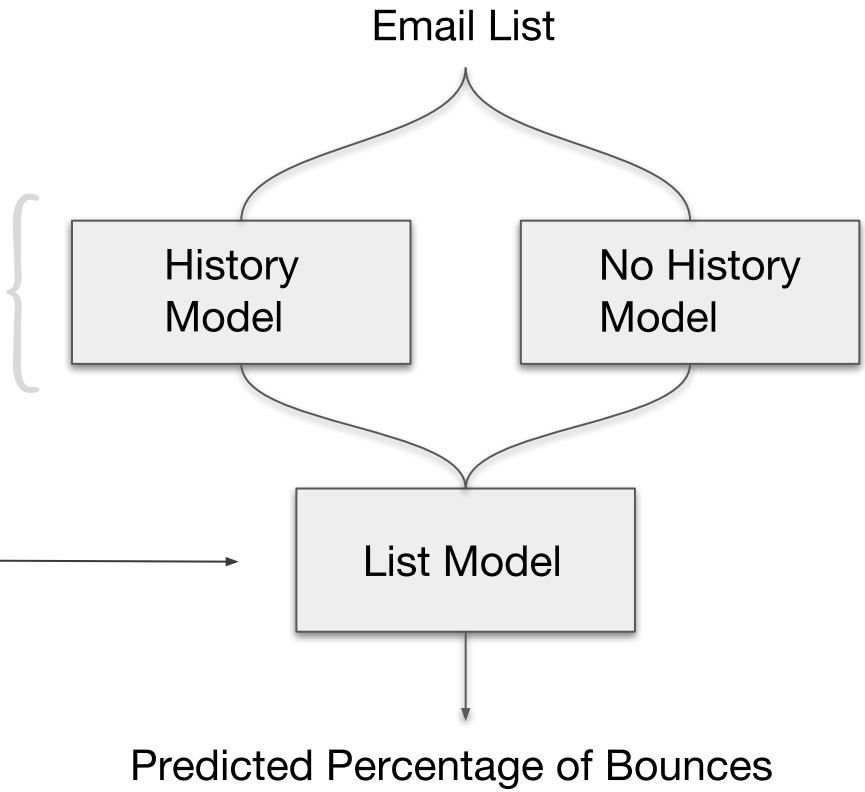
- Predict the probability of a bounce or delivery for each address
- Predict the percentage of addresses on a list that will bounce



Bounce Prediction

Three random forests

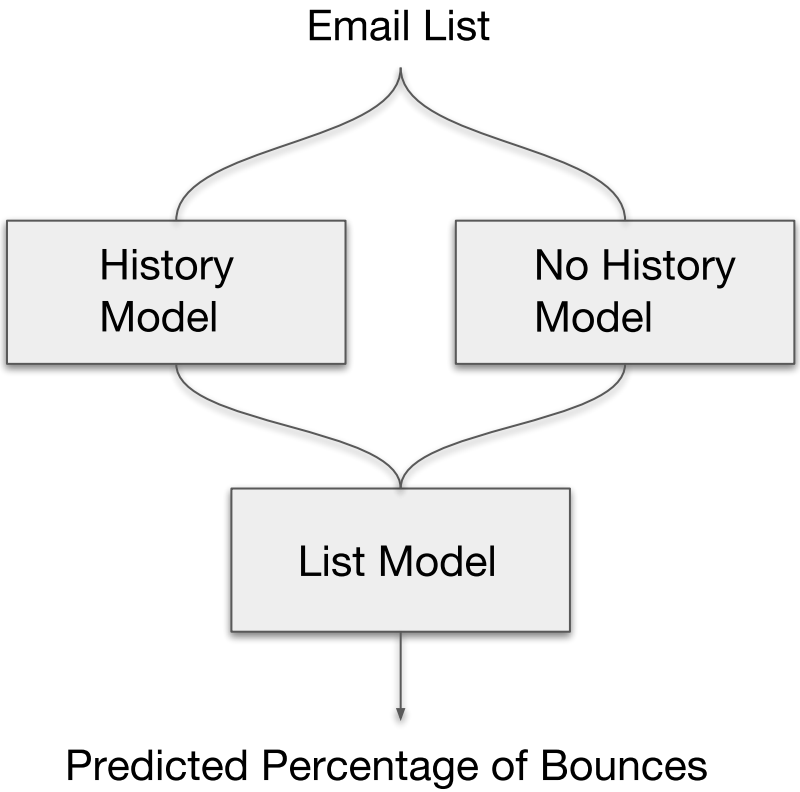
- Predict the probability of a bounce or delivery for each address
- Predict the percentage of addresses on a list that will bounce



Bounce Prediction

Challenges

- Class imbalance in individual models
  - Layered design lessens the impact
- Overfitting in the list model
  - Properly sample lists of different sizes
  - We care most about accuracy for new users

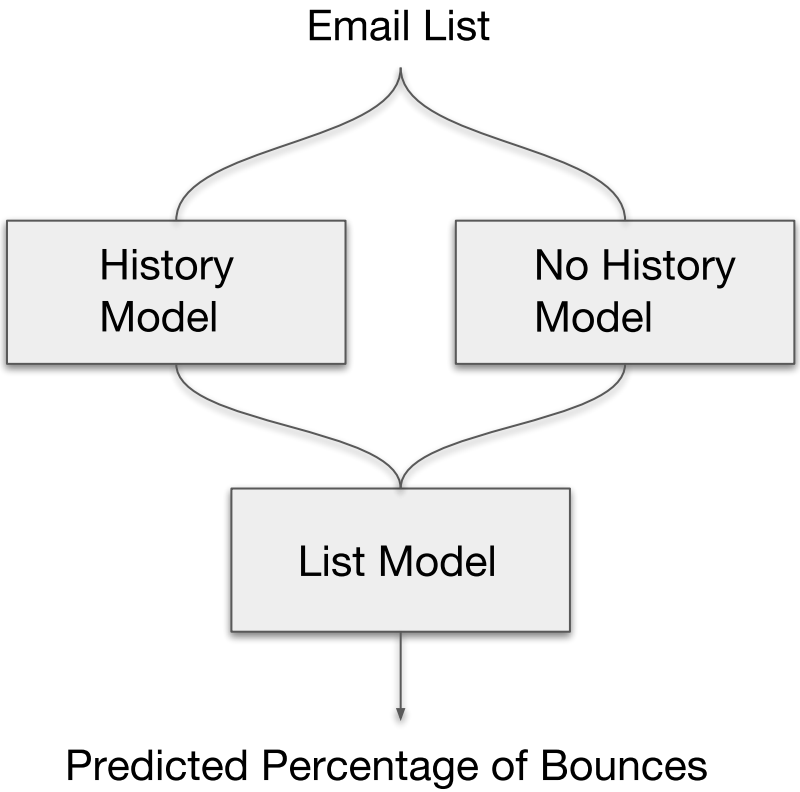


Bounce Prediction

Challenges

- Class imbalance in individual models
  - Layered design lessens the impact

- Overfitting in the list model
  - Properly sample lists of different sizes
  - We care most about accuracy for new users

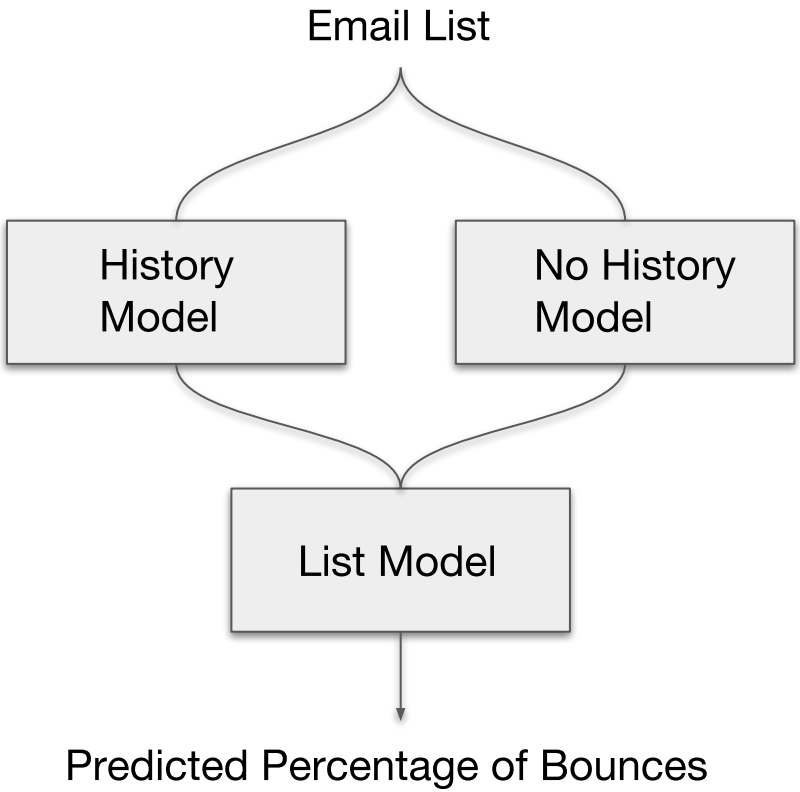


Bounce Prediction

Challenges

- Class imbalance in individual models
  - Layered design lessens the impact

- Overfitting in the list model
  - Properly sample lists of different sizes
  - We care most about accuracy for new users



We have a model, now what?

We need to aggregate data by email address. This is challenging for several reasons:

- MailChimp organizes data by user
- There are actually many MailChimps
- We send a lot of email

We have a model, now what?

We need to aggregate data by email address. This is challenging for several reasons:

- MailChimp organizes data by user
- There are actually many MailChimps
- We send a lot of email

We have a model, now what?

We need to aggregate data by email address. This is challenging for several reasons:

- MailChimp organizes data by user
- There are actually many MailChimps
- We send a lot of email

We have a model, now what?

We need to aggregate data by email address. This is challenging for several reasons:

- MailChimp organizes data by user
- There are actually many MailChimps
- We send a lot of email

29 million

29 million

Number of emails sent per *day* when EGP was built (2012)

1 billion

1 billion

Number of emails sent per *day* now

3

Percentage of EGP code that implements predictive models

What is the rest?

- Moving data
- Aggregating data
- Archiving raw events for training
- API
- Deployment

What is the rest?

- Moving data
- Aggregating data
- Archiving raw events for training
- API
- Deployment

What is the rest?

- Moving data
- Aggregating data
- Archiving raw events for training
- API
- Deployment

What is the rest?

- Moving data
- Aggregating data
- Archiving raw events for training
- API
- Deployment

What is the rest?

- Moving data
- Aggregating data
- Archiving raw events for training
- API
- Deployment

What is the rest?

- Moving data
- Aggregating data
- Archiving raw events for training
- API
- Deployment

Growing Pains

EGP is a successful system, but...

- It has been fairly isolated
  - Didn’t take advantage of improving support structures
  - Knowledge of the system was isolated as well

- Maintenance and scaling take a lot of effort
  - Largely due to the data architecture
  - Less time to work on new projects

Growing Pains

EGP is a successful system, but...

- It has been fairly isolated
  - Didn’t take advantage of improving support structures
  - Knowledge of the system was isolated as well

- Maintenance and scaling take a lot of effort
  - Largely due to the data architecture
  - Less time to work on new projects

Growing Pains

EGP is a successful system, but...

- It has been fairly isolated
  - Didn’t take advantage of improving support structures
  - Knowledge of the system was isolated as well

- Maintenance and scaling take a lot of effort
  - Largely due to the data architecture
  - Less time to work on new projects

Growing Pains

EGP is a successful system, but...

- It has been fairly isolated
  - Didn't take advantage of improving support structures
  - Knowledge of the system was isolated as well

- Maintenance and scaling take a lot of effort
  - Largely due to the data architecture
  - Less time to work on new projects

Growing Pains

EGP is a successful system, but...

- It has been fairly isolated
  - Didn't take advantage of improving support structures
  - Knowledge of the system was isolated as well

- Maintenance and scaling take a lot of effort
  - Largely due to the data architecture
  - Less time to work on new projects

Growing Pains

EGP is a successful system, but...

- It has been fairly isolated
  - Didn't take advantage of improving support structures
  - Knowledge of the system was isolated as well

- Maintenance and scaling take a lot of effort
  - Largely due to the data architecture
  - Less time to work on new projects

Important Point

I am not saying that bad decisions were made.

"organizations which design systems ... are constrained to produce designs which are copies of the communication structures of these organizations."

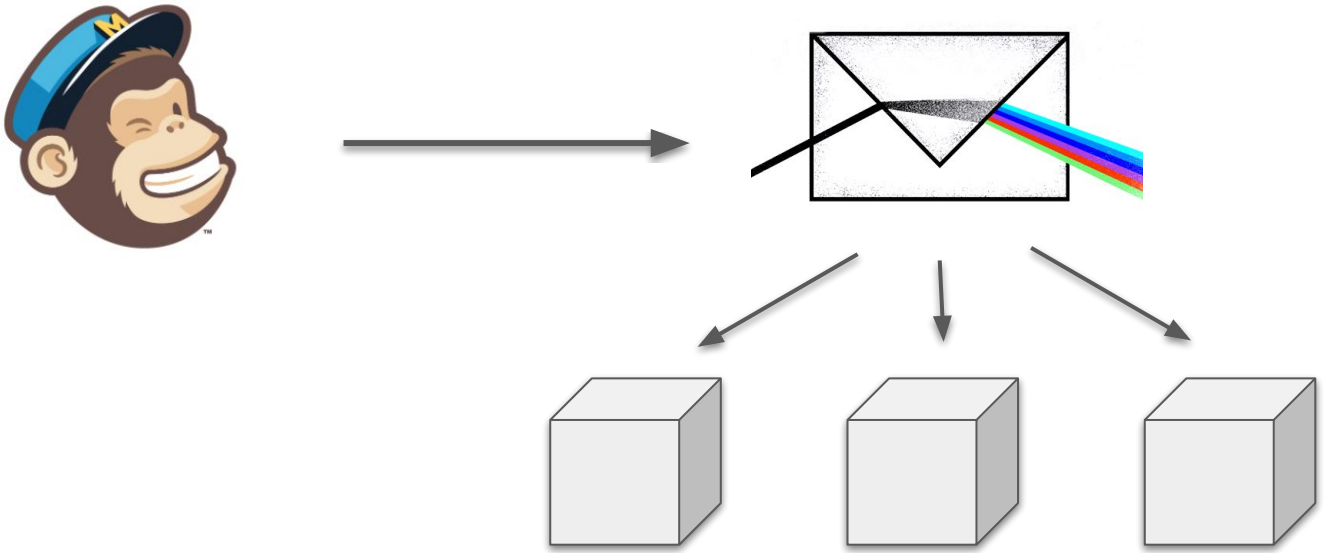
— M. Conway

Wikipedia

“We want to do a thing. EGP has that data, let’s do it there!”

- Us… many times

Data is like gravity



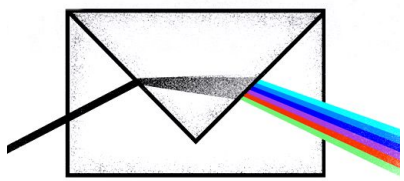
This is an anti-pattern

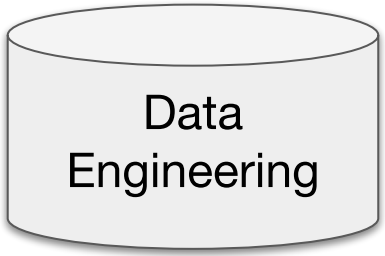
- If an application is serving data downstream, then this either...
    - imposes constraints on any additional projects that rely on that data
    - introduces unnecessary complexity into the “gatekeeper” application

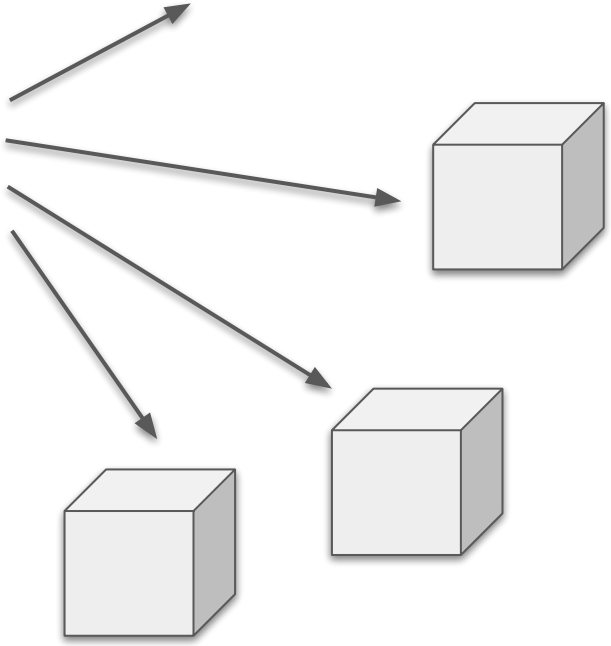
This is an anti-pattern

- If an application is serving data downstream, then this either...
  - imposes constraints on any additional projects that rely on that data
  - introduces unnecessary complexity into the “gatekeeper” application

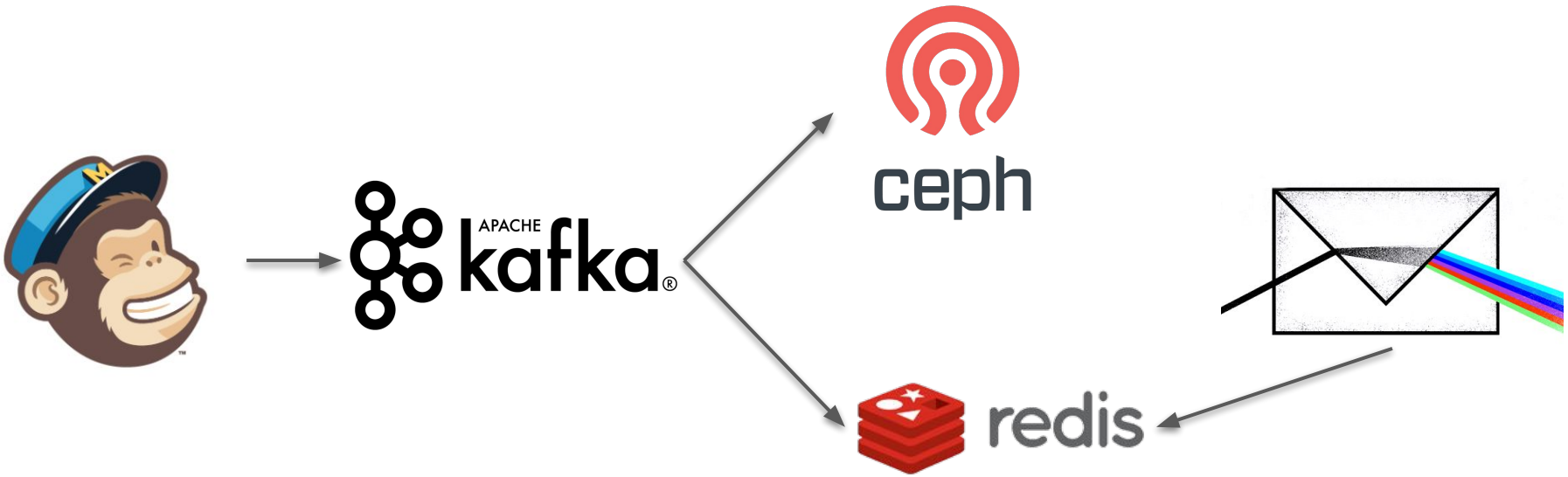
A better pattern



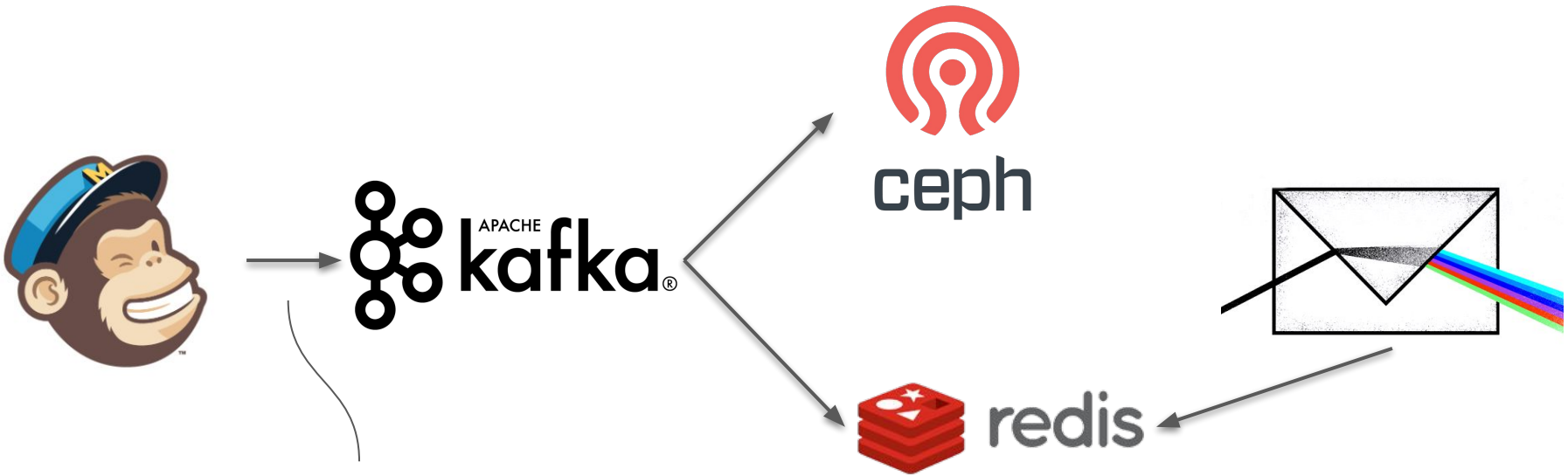




Work in Progress

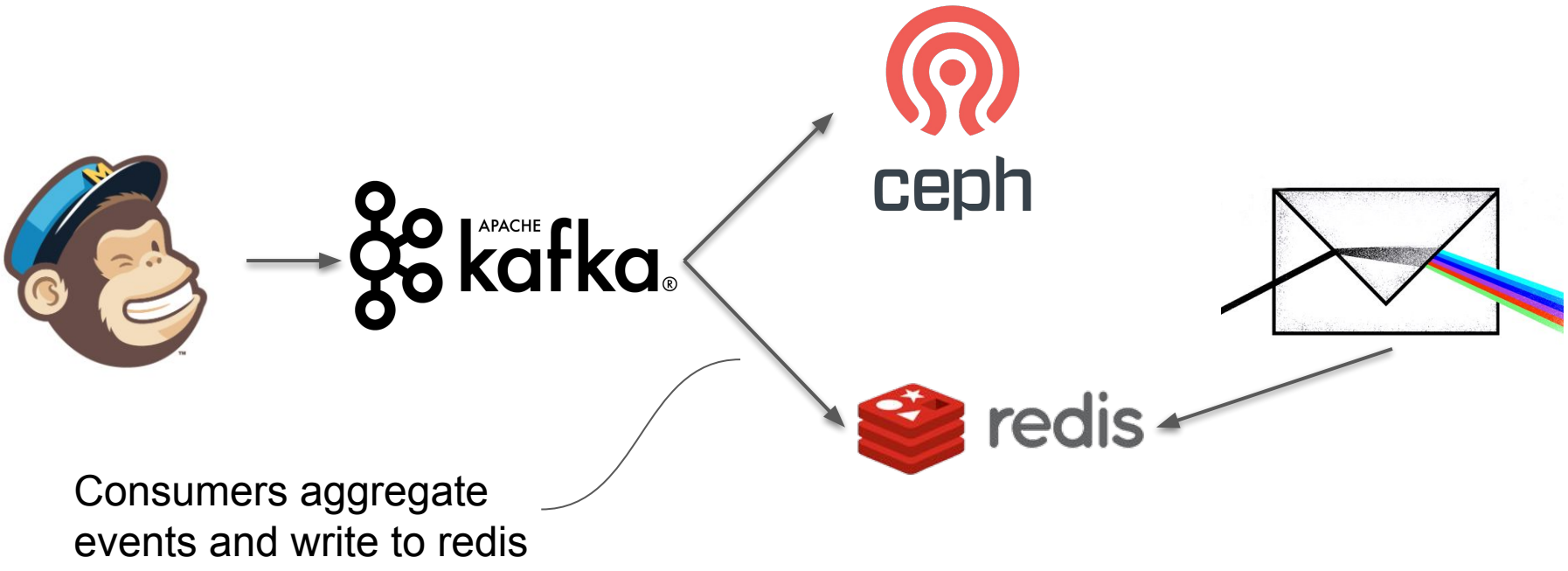


Work in Progress

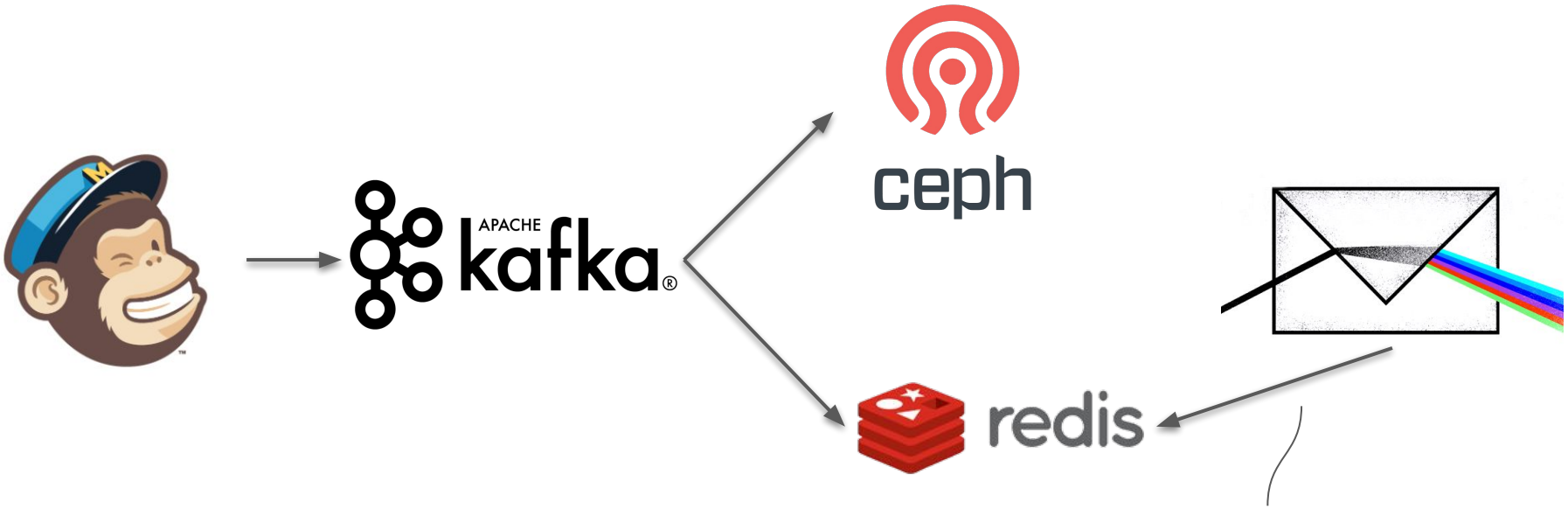


Event data is produced
from MailChimp into Kafka

Work in Progress

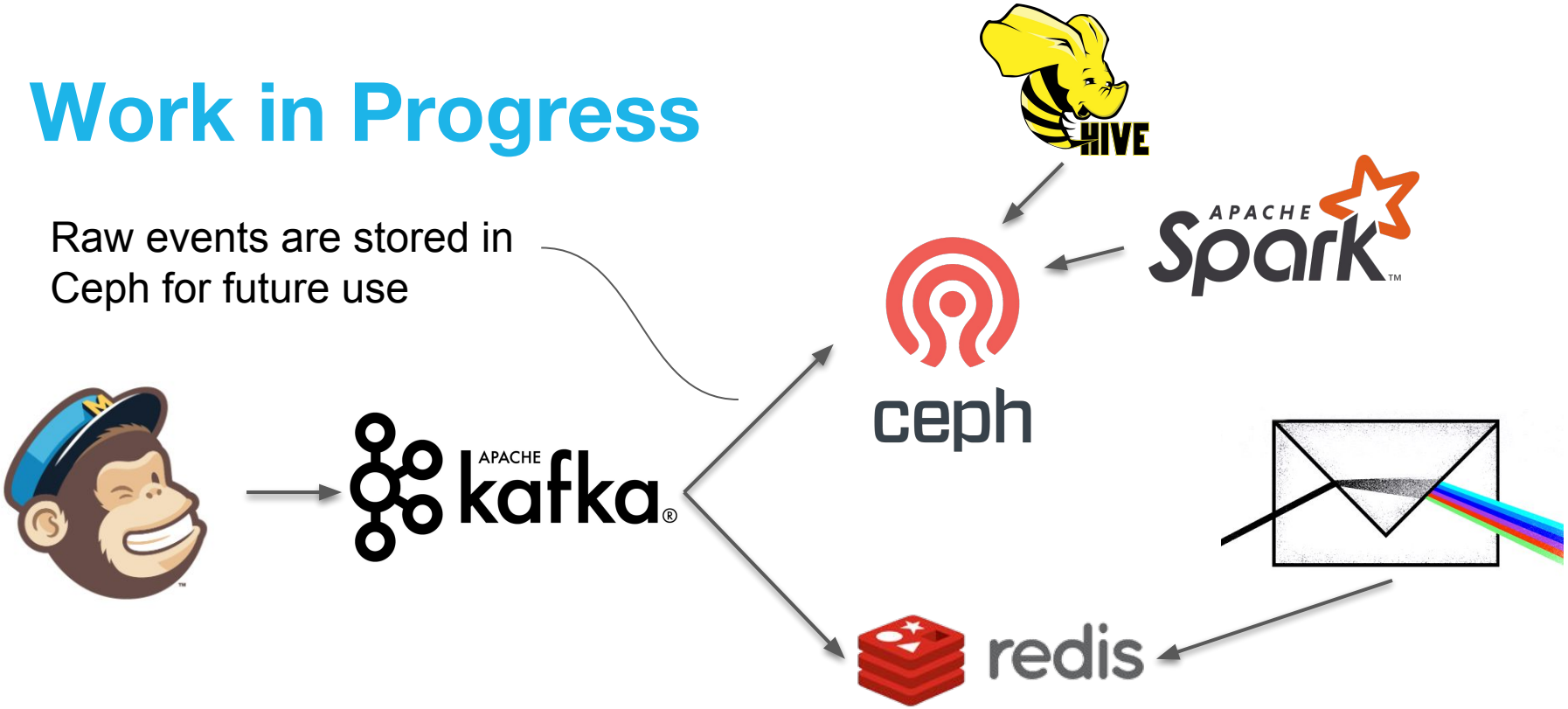


Work in Progress



Work in Progress

Raw events are stored in Ceph for future use



Wrapping up

- Data Science at MailChimp helps to
  - power user facing features
  - prevent abuse

- We have learned valuable lessons and are putting that knowledge to work
  - More attention to curating data
  - More collaboration between teams

- Conway's Law is not a joke
  - We should be mindful of the interfaces between both applications and teams

Wrapping up

- Data Science at MailChimp helps to
  - power user facing features
  - prevent abuse

- We have learned valuable lessons and are putting that knowledge to work
  - More attention to curating data
  - More collaboration between teams

- Conway's Law is not a joke
  - We should be mindful of the interfaces between both applications and teams

Wrapping up

- Data Science at MailChimp helps to
  - power user facing features
  - prevent abuse

- We have learned valuable lessons and are putting that knowledge to work
  - More attention to curating data
  - More collaboration between teams

- Conway’s Law is not a joke
  - We should be mindful of the interfaces between both applications and teams

Speaking of teams...

Apr 14, 2018

How to Manage Your Data Science and Software Engineering Teams to be More Productive?

(Panel) Trey Grainger, Frank Hinek , Coty Rosenblath, Rumman Chowdhury

4:30pm - 5:30pm

Coty Rosenblath is the Director of Data Systems and Data Science at MailChimp

Email: matt.mastin@mailchimp.com
Twitter: @mathmastin