



# Analyzing unstructured data

Data Engineering on Google Cloud Platform

©Google Inc. or its affiliates. All rights reserved. Do not distribute.  
May only be taught by Google Cloud Platform Authorized Trainers.

## **Notes:**

25 slides + 1 lab: 1 hour

## Remember?

### Human

Real-time insight into supply chain operations.  
Which partner is causing issues?

Drive product decisions.  
How do people really use feature X?

### Easy counting problems

Did error rates decrease after the bug fix was applied?

Which stores are experiencing long delays in payment processing?

### Harder counting problems

Are programmers checking in low-quality code?

Which stores are experiencing lacking of parking space?

### Notes:

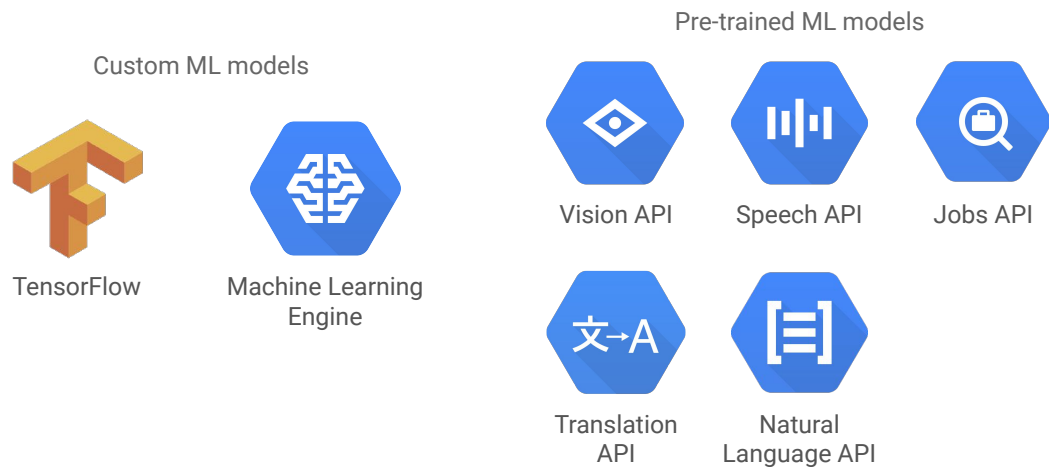
Compare these with the ones on the previous slide. Structured vs. unstructured.

Low-quality could be determined by bad sentiment in code reviews and often through programmer's own negative comments in the code. They are checking in the code because they need to move on to their next project .... But there are also tools out there that will look for code-smells. Those tools can be run at scale on Dataproc.

# Agenda

Infuse your business with Machine Learning + Lab

# Pretrained models are a fast way to magical experiences



## Notes:

We'll look at TF & MLE next. But right now, let's talk about pre-trained ML models.

# ML APIs are accessed through REST APIs -- no machine learning knowledge required

## Invoke Vision API

The Vision API can work off an image in Cloud Storage or embedded directly into a POST message. I'll use C



. That photograph is from <http://www.publicdomainpictures.net/view-image.php?image=15842>

Image  
(GCS/embedded)

JSON request

```
# Running Vision API
import base64
IMAGE="gs://cloud-training-demos/vision/sign2.jpg"
vservice = build('vision', 'v1', developerKey=APIKEY)
request = vservice.images().annotate(body={
    'requests': [
        {
            'image': {
                'source': {
                    'gcs_image_uri': IMAGE
                },
            },
            'features': [
                {
                    'type': 'TEXT_DETECTION',
                    'maxResults': 3,
                }
            ]
        }
    ]
})
responses = request.execute(num_retries=3)
print responses
```

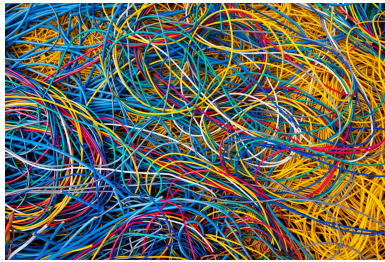
```
{
  "responses": [
    {
      "textAnnotations": [
        {
          "locale": "zh",
          "description": "请您爱护和保护卫生创建优美水环境",
          "boundingPoly": {
            "vertices": [
              {
                "x": 183,
                "y": 654
              },
              {
                "x": 150,
                "y": 654
              },
              {
                "x": 178,
                "y": 241
              },
              {
                "x": 322,
                "y": 41
              }
            ]
          },
          "confidence": 0.98
        }
      ]
    }
  ]
}
```

JSON response

## Notes:

This was a lab in the fundamentals course. They've done this already.

# Build on top of Google



ML API



Images  
Audio  
Video  
Free-form text



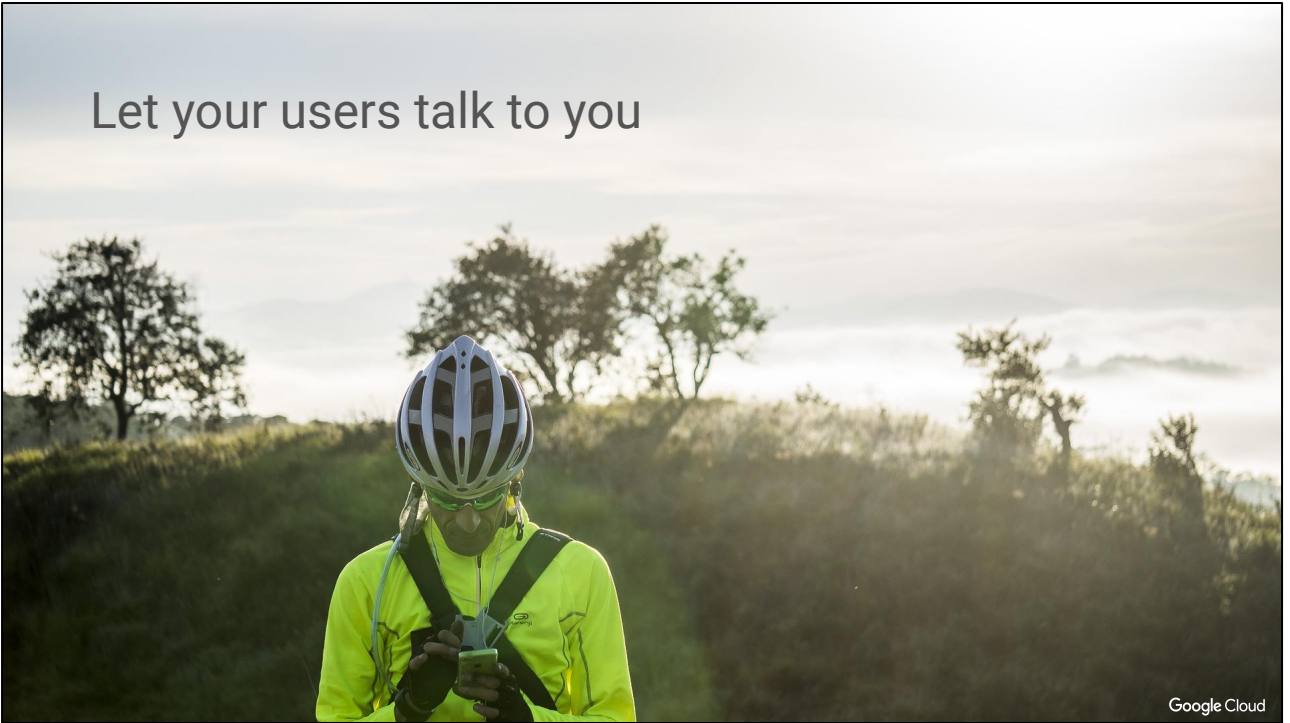
Places  
Labels  
People  
Events  
...

**COUNT THESE A LOT  
EASIER!**

## Notes:

<https://pixabay.com/en/list-zettelbox-note-leaves-stack-1925395/> (cc0)

# Let your users talk to you



## Notes:

<https://pixabay.com/en/cycling-bike-trail-sport-sol-1533268/> (cc0)

idea : voice-enable your applications

Use Speech API

Just a REST call, so easy to incorporate

Voice-navigation?

Know your user/from their app/anticipate their need/carry on conversation = ASSISTANT story

## Go from speech to action with bots



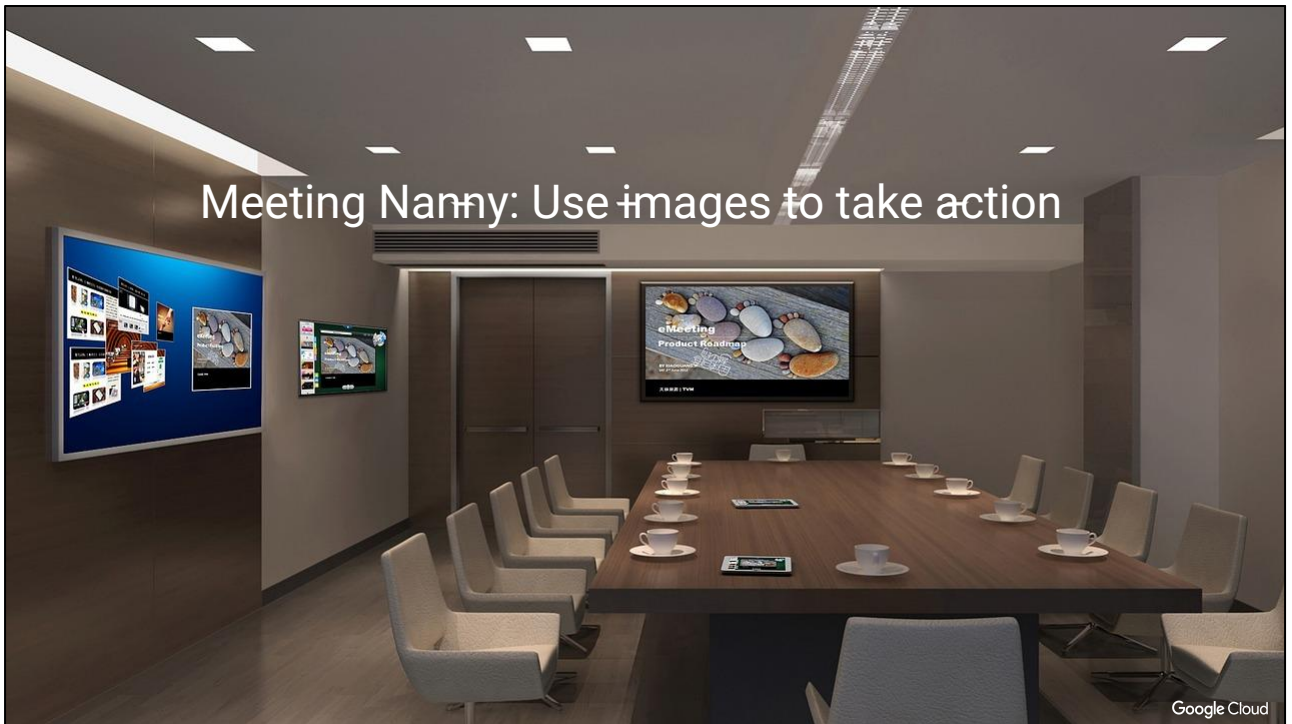
<https://dialogflow.com/>

<https://cloud.google.com/dialogflow-enterprise/>

### Notes:

<https://cloud.google.com/dialogflow-enterprise/>





## Notes:

Image from <https://pixabay.com/en/interior-design-tv-multi-screen-828545/> (cc0)

From:  
<https://g3doc.corp.google.com/java/com/google/corp/bizapps/reus/spaces/aver/g3doc/gvc.md?cl=head>

We do occupancy detection via motion detection (by the VC camera) and by call ID matching. Every 30 seconds, the VC unit sends a Pubsub notification whether motion was detected or not. It also sends a Pubsub notification when a call started or ended together with whether the call ID matched the meeting ID.

If motion is detected between 6 and 8 minutes after the meeting start time, the room counts as occupied. Otherwise, it's empty.

## Ocado routes emails based on NLP



Improves natural language processing of customer service claims

"Hi Ocado,  
I love your website. I have children so it's easier for me to do the shopping online. Many thanks for saving my time!  
Regards"

Feedback

Customer is happy

### Notes:

- World's largest online only grocery supermarket
- Goal for best customer service
- Customers call or email their contact center (Social media, landline, email, SMS)
- Types: General feedback, refunds, redeliver, payment issues
- No forms or self categorization...all emails in a central mailbox
- Traditionally, each email gets addressed and routed. Can't scale, longer delays, poor experience
- Sifting through email is a repetitive task
- Ocado Technology w/ 1000+ developers, engineers, data scientists
- Used Natural Language processing: combines computer science, artificial intelligence, and computational linguistics
- Parse through the body of emails, tags and routes to help contact center reps determine the priority and context

# Wootric collects both numeric and qualitative feedback



How likely are you to recommend API Editor to colleague?

Not at all likely 0 1 2 3 4 5 6 7 8 9 10 Extremely likely

powered by wootric

**EASY TO COMPUTE NET  
PROMOTER SCORE**

Great. What is most satisfying about it?

Editor handles simple API well but it loads really slow for complex API definition. I was able to get some tips from docs though.

Not at all likely 0 1 2 3 4 5 6 7 8 9 10 Extremely likely

powered by wootric

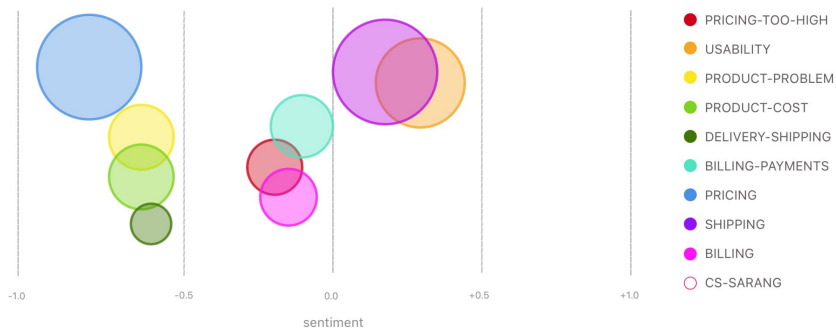
SEND

**FREE FORM TEXT -- NOT  
AS EASY TO HANDLE**

<https://cloud.google.com/blog/big-data/2017/03/analyzing-customer-feedback-using-machine-learning>

They use NLP API and custom ML models to classify sentiment, topic, and support personnel

Sentiment Volume



### Notes:

Support personnel is not shown, but the idea is that if the feedback mentions "jessica", they know who is being talked about.

## Lab 6: Add Machine Learning (ML)

- Simplify machine learning tasks using the NLP API
- Incorporate several machine learning services from the Natural Language API
- Use sentiment analysis and entity analysis, to produce meaningful results from unstructured data.



cloud.google.com