Opportunities to Collaborate in Data Science + Data Engineering

Kathy Copic

kathy@insightdatascience.com

The Insight Fellows Program

insightdatascience.com | insightdataengineering.com

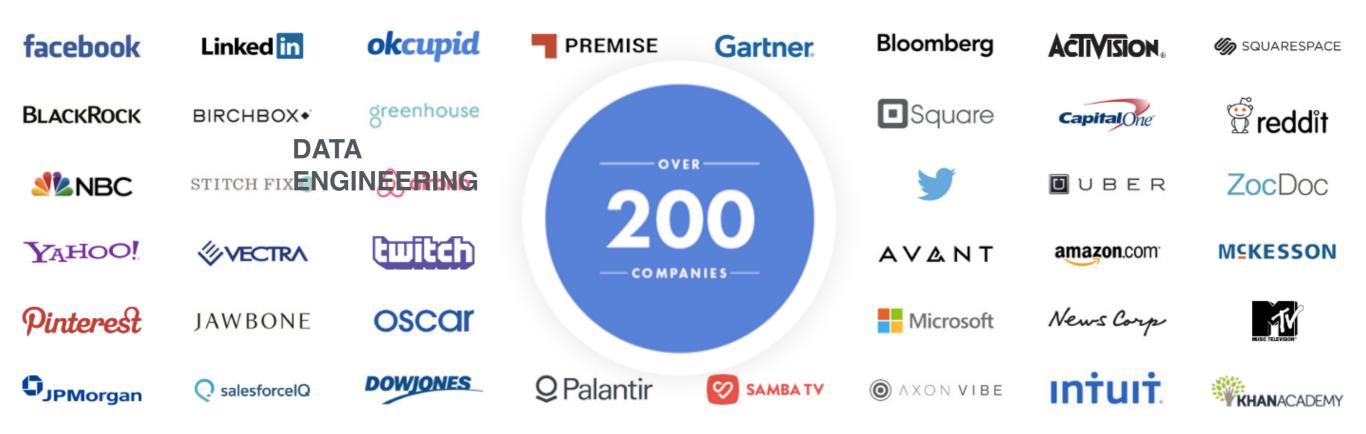
insightdata.ai insighthealthdata.com



Insight Fellows across the US



Insight Fellows are Data Scientists and Data Engineers at:



Silicon Valley • New York • Boston • Los Angeles • Seattle + many others...

Tremendous innovation and production of tools

INSIGHT

Find out more about the Insight Data Engineering Fellows Program and Data Labs

File Format

2. ProtoBuf

4. Parquet

5. ORC Files

1. Avro

3. Thrift

Ingestion

- Kafka
- 2. Logstash
- 3. RabbitMQ
- 4. Fluentd
- 5. AWS Kinesis

Mouse over each box for more details.

Click on each technology for resources to get started.

Last updated Oct 26th, 2015 You can also find the previous version here.

High-Level MR

- 1. Pia 2. Cascading
- 3. Hadoop Streaming
- 4. Cascalog

Batch ML

- 1. H2O 2. Mahout
- 3. Spark MLlib
- 4. FlinkML

Batch Graph

- 1. GraphLab 2. Giraph
- 3. Spark GraphX

4. Hama

Batch SQL

- 1. Hive 2. Presto
- 3. Drill
- 4. Impala

General management tools for data pipelines

Cluster Management

- 1. Docker
- 2. Zookeeper 3. YARN
- 5. Hue
- 4. Mesos

Scheduling/Monitoring

- 1. Luiai
- 2. Airflow 3. Nagios
- 4. Graphite
- 5. Azkaban

File System

- 1. HDFS 2. AWS S3
- 3. Azure
- 4. Tachyon
- 5. Ceph

Batch Processing

- 1. Spark
 - 2. Hadoop MapReduce
 - 3. AWS EMR
 - 4. Flink
- 5. Tez

Stream Processing

2. Spark Streaming 3. AWS Lambda

1. Storm

4. Samza

5. Flink

Data Store

Transactions Analytics

Uptime Critical

Search Graph

Geospatial

Time Series Cache

Web Framework

- 1. Ruby on Rails
- 2. Node.is
- 3. Diango
- 4. AngularJS
- 5. Flask

Data Visualization

- 1. D3
- 2. Tableau
- 3. Leaflet
- 4. Highcharts
- 5. Kibana

Transactions

- 1. MvSQL
- 2. Oracle
- 3. PostgreSQL

Analytics 1. AWS Redshift

- 2. Vertica
 - 3. HBase

- 1. Cassandra
- 2. Riak
- 3. AWS DynamoDB

Uptime Critical

Search

- 1. Elasticsearch
- 2. Solr
- 3. MongoDB

Graph

- 1. Neo4i
- 2. OrientDB 3. ArangoDB

Geospatial

- 1. CouchDB 2. PostGIS
- 3. Elasticsearch

Time Series

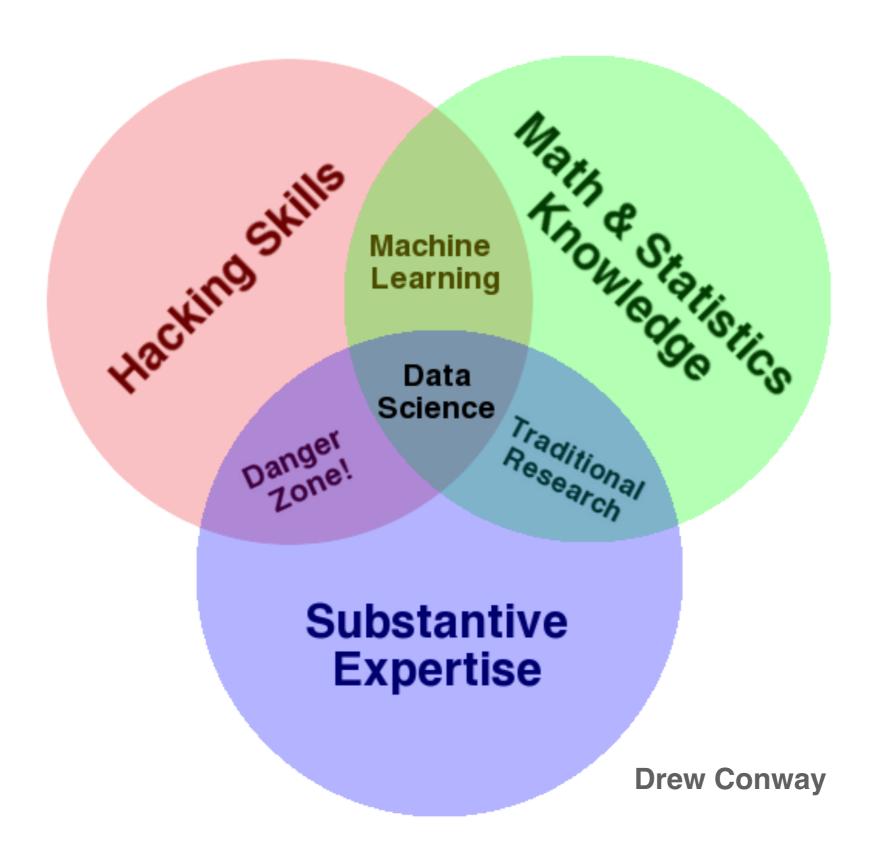
- 1. InfluxDB 2. Cassandra
- 3. Druid

Cache

- 1. Redis 2. Memcached
- 3. Hazelcast

... and Open Source!

What is Data Science?



What is Data Engineering?

INSIGHT

High-Level MR

- 1. Pia
- 2. Cascading
- 3. Hadoop Streaming
- 4. Cascalog

Batch ML

- 1. H2O
- 2. Mahout
- 3. Spark MLlib 4. FlinkML

Batch Graph 1. GraphLab

3. Spark GraphX

2. Giraph

- 1. Hive
- 2. Presto

Batch SQL

- 3. Drill
- 4. Impala

General management tools for data pipelines

Cluster Management

- 1. Docker
- 2. Zookeeper
- 3. YARN
- 4. Mesos
- 5. Hue

Scheduling/Monitoring

- 1. Luigi
- 2. Airflow
- Nagios
- 4. Graphite
- 5. Azkaban

Ingestion

- Kafka
- 2. Logstash
- 3. RabbitMQ
- 4. Fluentd
- 5. AWS Kinesis

File Format

- 1. Avro
- 2. ProtoBuf
- 3. Thrift
- 4. Parquet
- 5. ORC Files

File System Batch Processing

- 1. HDFS
- 2. AWS S3
- 3. Azure
- 4. Tachyon
- 5. Ceph

- 1. Spark
- 2. Hadoop MapReduce
- 3. AWS EMR
- 4. Flink
- 5. Tez

Stream Processing

- 1. Storm
- 2. Spark Streaming
- 3. AWS Lambda
- 4. Samza
- 5. Flink

Data Store

Transactions Analytics

Uptime Critical

Search Graph

Geospatial Time Series

Cache

Web Framework

- 1. Ruby on Rails
- 2. Node.js
- 3. Django
- 4. AngularJS
- 5. Flask

Data Visualization

- 1. D3
- 2. Tableau
- 3. Leaflet
- 4. Highcharts
- 5. Kibana

Transactions

- 1. MySQL
- 2. Oracle
- 3. PostareSQL
- Analytics
- 1. AWS Redshift
- 2. Vertica
- 3. HBase

Uptime Critical

- 1. Cassandra 2. Riak
- 3. AWS DynamoDB

Search

- 1. Elasticsearch 2. Solr
- 3. MongoDB

Graph

- 1. Neo4i
- 2. OrientDB
- 3. ArangoDB

Geospatial

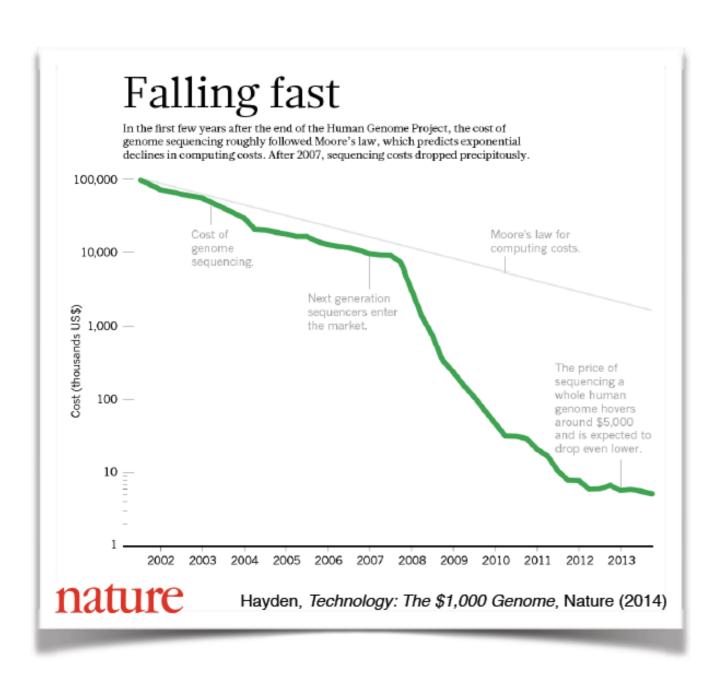
- 1. CouchDB 2. PostGIS
- 3. Elasticsearch

Time Series

- 1. InfluxDB 2. Cassandra
- 3. Druid

- Cache 1. Redis
- 2. Memcached
- Hazelcast

Health Data: Explosion!



- Genome sequences
- Electronic medical records
- Claims data
- Quantified-self data
- Longitudinal studies
- Drug screens
- Clinical trial data

Data Analytics / Decision Science Analysis to drive choices

Data Analytics / Decision Science Analysis to drive choices

Data Products

Building Products used by people

Data Analytics / Decision Science Analysis to drive choices

Data Products

Building Products used by people

AI / ML / Deep Learning Engineer Next-level, "Reinforcement learning"

Data Analytics / Decision Science Analysis to drive choices

Data Products

Building Products used by people

AI / ML / Deep Learning Engineer Next-level, "Reinforcement learning"

Data Engineers

Deal with size, speed, quality

Big difference in Big vs Small Companies

Go to **Meet-ups** or watch online.

Go to **Meet-ups** or watch online.

Subscribe to **Data Science Weekly**.

Go to **Meet-ups** or watch online.

Subscribe to **Data Science Weekly**.

Join LinkedIn! Meet up with people.

Go to **Meet-ups** or watch online.

Subscribe to **Data Science Weekly**.

Join LinkedIn! Meet up with people.

Podcasts: Software Engineering Daily

Go to **Meet-ups** or watch online.

Subscribe to Data Science Weekly.

Join LinkedIn! Meet up with people.

Podcasts: Software Engineering Daily

Read Twitter!

Go to **Meet-ups** or watch online.

Subscribe to **Data Science Weekly**.

Join LinkedIn! Meet up with people.

Podcasts: Software Engineering Daily

Read Twitter!

Do nearby universities have a center for data science?

Go to **Meet-ups** or watch online.

Subscribe to **Data Science Weekly**.

Join LinkedIn! Meet up with people.

Podcasts: Software Engineering Daily

Read Twitter!

Do nearby universities have a center for data science?

Make open source contributions, get involved!

Go to **Meet-ups** or watch online.

Subscribe to Data Science Weekly.

Join LinkedIn! Meet up with people.

Podcasts: Software Engineering Daily

Read **Twitter!**

Do nearby universities have a center for data science?

Make open source contributions, get involved!

Read our Insight Blog!:)



PROGRAMS



DATA SCIENCE

FOR PHDS



DATA ENGINEERING

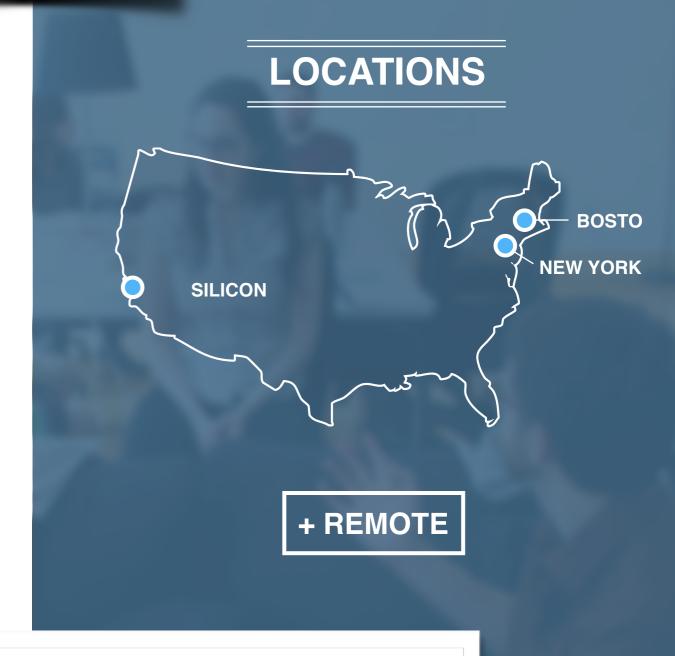


HEALTH ANALYTICS

FOR PHDS + MDS



ARTIFICIAL INTELLIGENCE

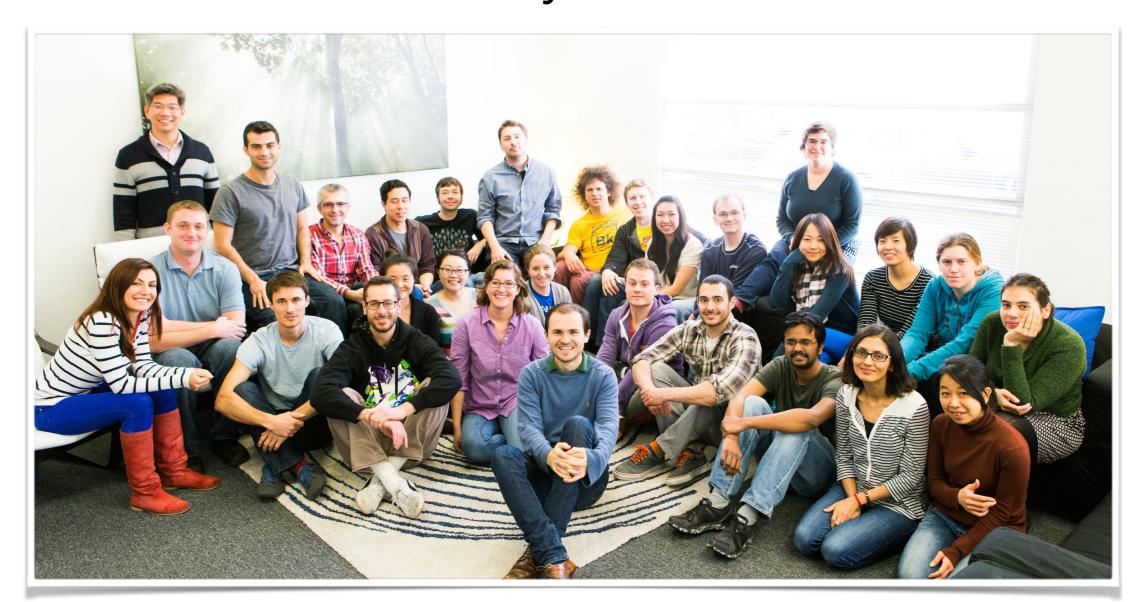


Upcoming Session: January 17, 2017

Application Deadline: October 24, 2016
Visit our website to apply or sign up for notifications

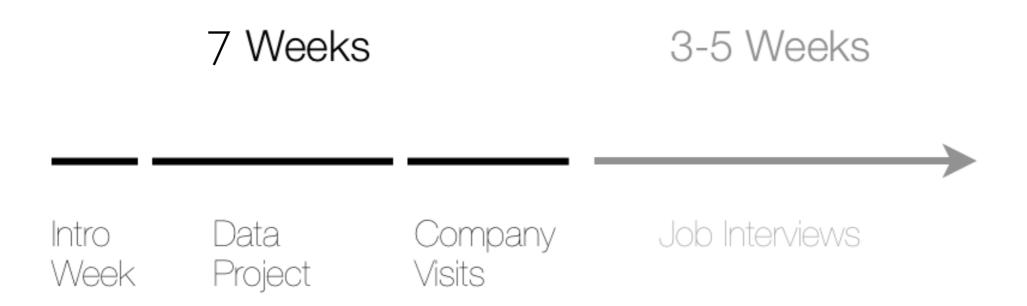
Contact me!

kathy@insightdatascience.com
Kathy Copic: LinkedIn!
@KathyScientist

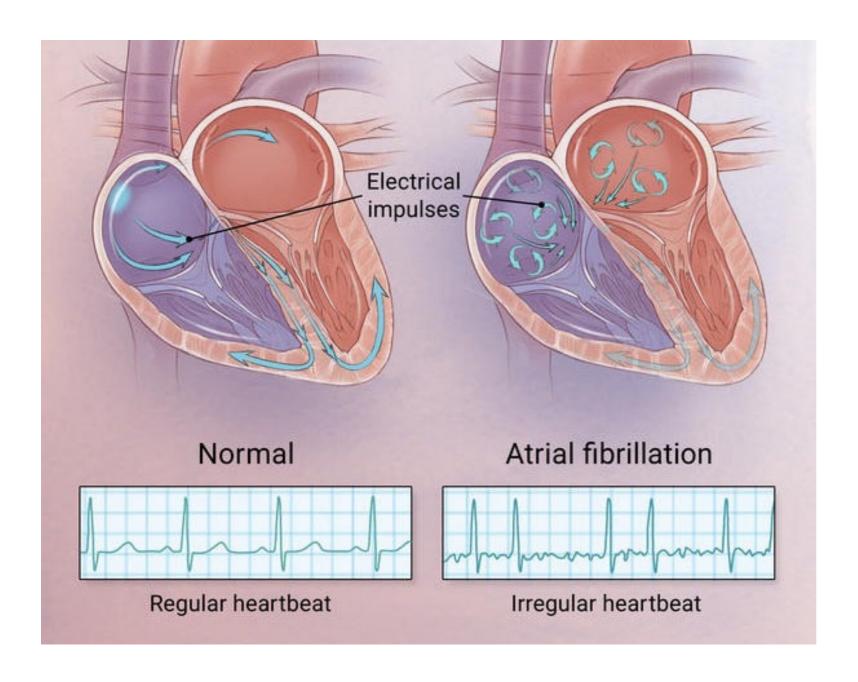




Data Science & Data Engineering programs

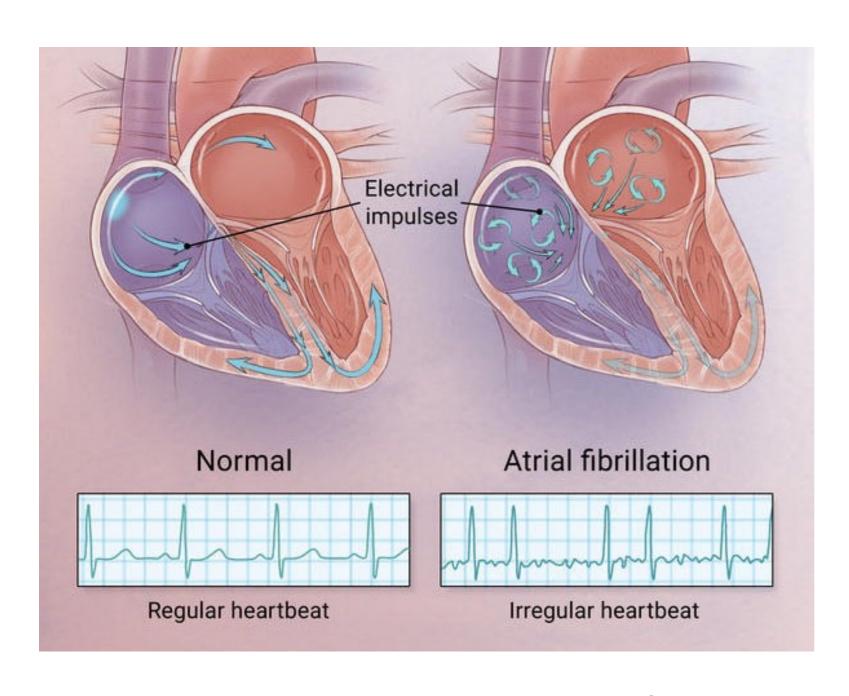


Bay Area, NYC, Boston and more!



2.7 million patients in the US

Leading cause of stroke

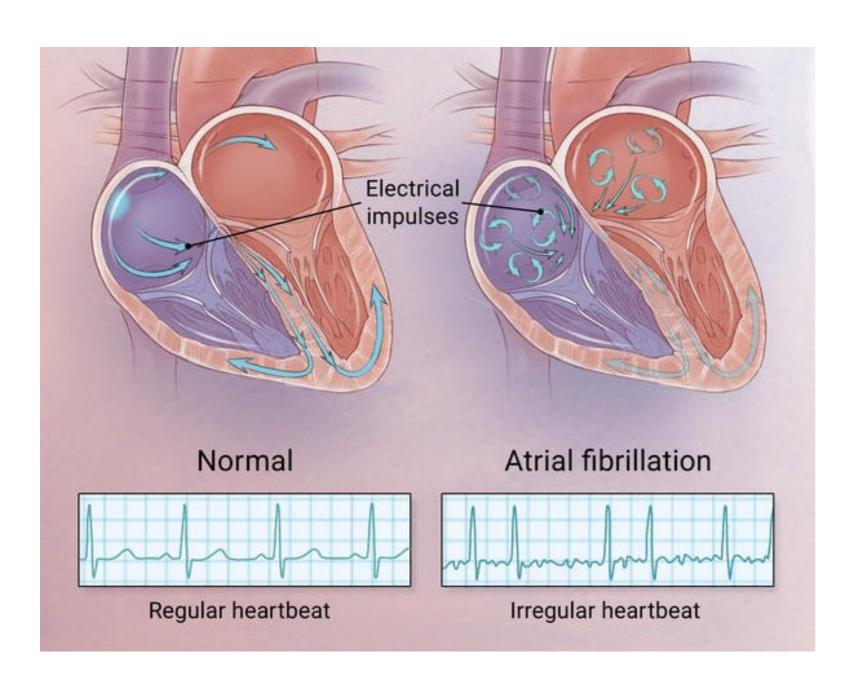


UCSF's Health eHeart Study:

13.5 million heart rate measurements from ~500 users

2.7 million patients in the US

Leading cause of stroke



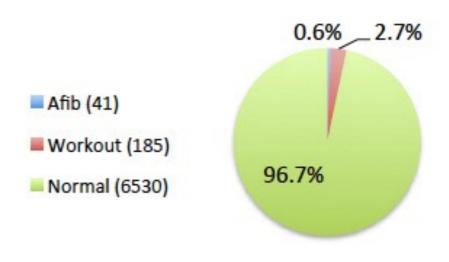
2.7 million patients in the US

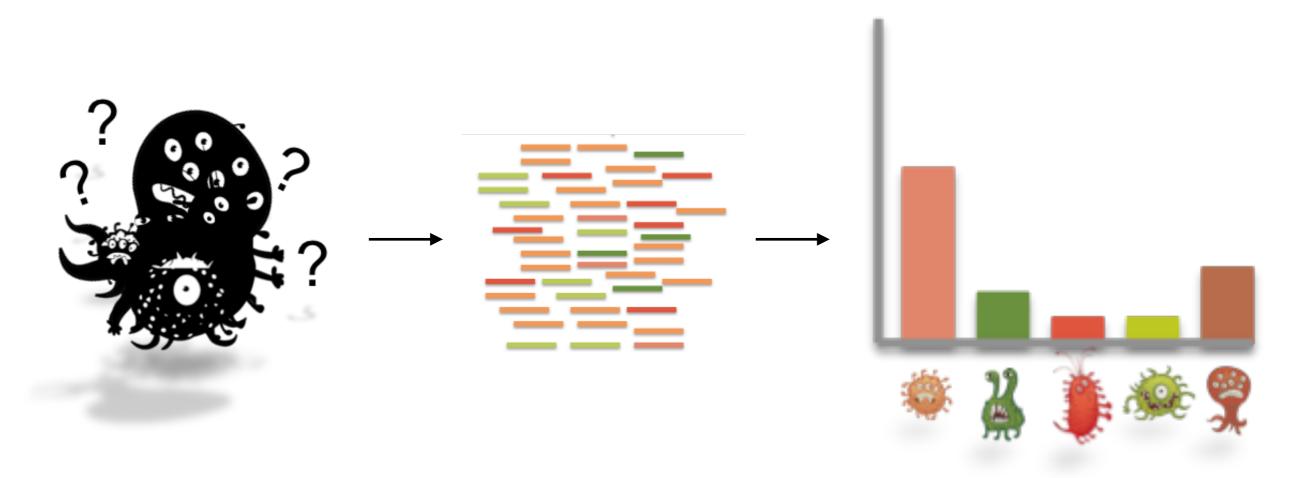
Leading cause of stroke

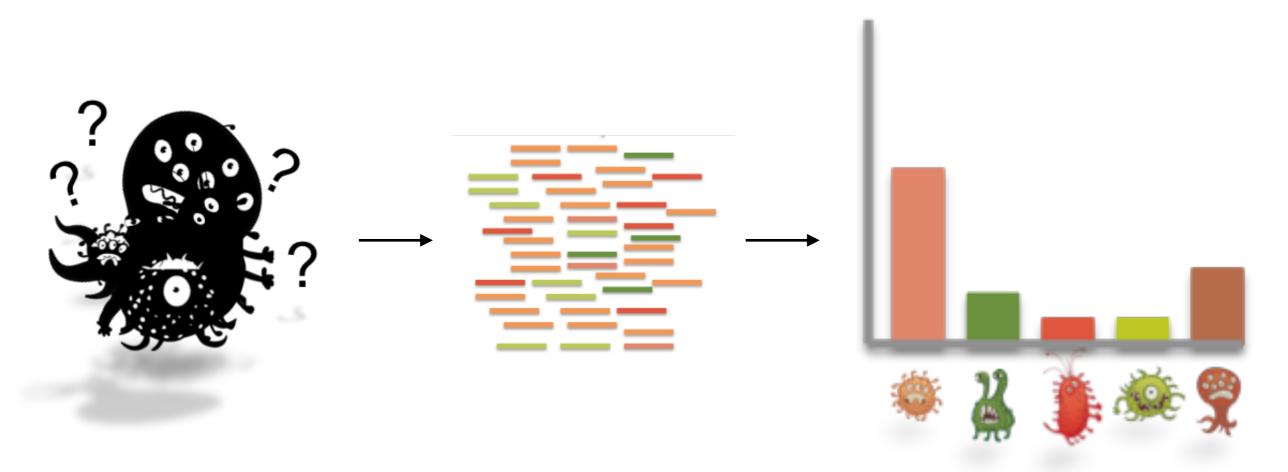
UCSF's Health eHeart Study:

13.5 million heart rate measurements from ~500 users

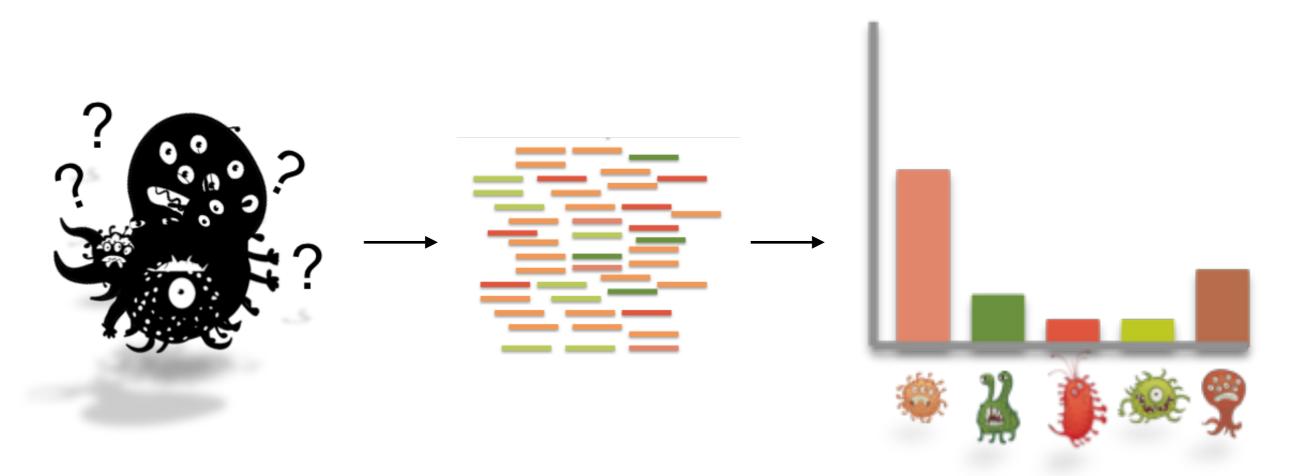
Only 41 Afib cases







YC startup One Codex — 500 FASTQ sequencing files, ~1M+ reads



YC startup One Codex — 500 FASTQ sequencing files, ~1M+ reads

Applied LDA topic modeling: a technique from Natural Language Processing

Memorial Sloan Kettering Cancer Center:

Who is a good candidate for clinical trials for treatments?

Memorial Sloan Kettering Cancer Center:

Who is a good candidate for clinical trials for treatments?

Biogen: If you combine data from genomics, clinical research, environmental studies, etc, are there new approaches to treat disease?

Memorial Sloan Kettering Cancer Center:

Who is a good candidate for clinical trials for treatments?

Biogen: If you combine data from genomics, clinical research, environmental studies, etc, are there new approaches to treat disease?

Constant Therapy: Can an iPad app deliver highly personalized, continuous therapy to patients with traumatic brain injury, stroke, etc?

Memorial Sloan Kettering Cancer Center:

Who is a good candidate for clinical trials for treatments?

Biogen: If you combine data from genomics, clinical research, environmental studies, etc, are there new approaches to treat disease?

Constant Therapy: Can an iPad app deliver highly personalized, continuous therapy to patients with traumatic brain injury, stroke, etc?

Memorial Sloan Kettering Cancer Center:

Who is a good candidate for clinical trials for treatments?

Biogen: If you combine data from genomics, clinical research, environmental studies, etc, are there new approaches to treat disease?

Constant Therapy: Can an iPad app deliver highly personalized, continuous therapy to patients with traumatic brain injury, stroke, etc?

Trace Genomics: Can we identify microbes and genetic mutations that cause or protect crops from disease, improving yield and shelf life?

Memorial Sloan Kettering Cancer Center:

Who is a good candidate for clinical trials for treatments?

Biogen: If you combine data from genomics, clinical research, environmental studies, etc, are there new approaches to treat disease?

Constant Therapy: Can an iPad app deliver highly personalized, continuous therapy to patients with traumatic brain injury, stroke, etc?

Trace Genomics: Can we identify microbes and genetic mutations that cause or protect crops from disease, improving yield and shelf life?

Memorial Sloan Kettering Cancer Center:

Who is a good candidate for clinical trials for treatments?

Biogen: If you combine data from genomics, clinical research, environmental studies, etc, are there new approaches to treat disease?

Constant Therapy: Can an iPad app deliver highly personalized, continuous therapy to patients with traumatic brain injury, stroke, etc?

Trace Genomics: Can we identify microbes and genetic mutations that cause or protect crops from disease, improving yield and shelf life?

Everywhere: Data architecture — how can the data be stored, streamed, accessed?

Memorial Sloan Kettering Cancer Center:

Who is a good candidate for clinical trials for treatments?

Biogen: If you combine data from genomics, clinical research, environmental studies, etc, are there new approaches to treat disease?

Constant Therapy: Can an iPad app deliver highly personalized, continuous therapy to patients with traumatic brain injury, stroke, etc?

Trace Genomics: Can we identify microbes and genetic mutations that cause or protect crops from disease, improving yield and shelf life?

Everywhere: Data architecture — how can the data be stored, streamed, accessed?

Memorial Sloan Kettering Cancer Center:

Who is a good candidate for clinical trials for treatments?

Biogen: If you combine data from genomics, clinical research, environmental studies, etc, are there new approaches to treat disease?

Constant Therapy: Can an iPad app deliver highly personalized, continuous therapy to patients with traumatic brain injury, stroke, etc?

Trace Genomics: Can we identify microbes and genetic mutations that cause or protect crops from disease, improving yield and shelf life?

Everywhere: Data architecture — how can the data be stored, streamed, accessed?

Memorial Sloan Kettering Cancer Center:

Who is a good candidate for clinical trials for treatments?

Biogen: If you combine data from genomics, clinical research, environmental studies, etc, are there new approaches to treat disease?

Constant Therapy: Can an iPad app deliver highly personalized, continuous therapy to patients with traumatic brain injury, stroke, etc?

Trace Genomics: Can we identify microbes and genetic mutations that cause or protect crops from disease, improving yield and shelf life?

Everywhere: Data architecture — how can the data be stored, streamed, accessed?

Tremendous potential for impact!