Opportunities to Collaborate in Data Science + Data Engineering

Kathy Copic
kathy@insightdatascience.com

The Insight Fellows Program
insightdatascience.com | insightdataengineering.com

insightdata.ai | insightthealtdatag.com
Insight Fellows across the US

750
DATA SCIENTISTS AND DATA ENGINEERS
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

... and Open Source!
What is Data Science?

Drew Conway
What is Data Engineering?

INSIGHT

**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

**High-Level MR**
1. Pig
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
4. Mesos
5. Hue

**Scheduling/Monitoring**
1. Luigi
2. Airflow
3. Nagios
4. Grafite
5. Azkaban

**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

**Ingestion**
1. Kafka
2. Logstash
3. RabbitMQ
4. Fluentd
5. AWS Kinesis

**File Format**
1. Avro
2. ProtoBuf
3. Thrift
4. Parquet
5. ORC Files

**Stream Processing**
1. Storm
2. Spark Streaming
3. AWS Lambda
4. Samza
5. Flink
Health Data: Explosion!

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

Falling fast

In the first few years after the end of the Human Genome Project, the cost of genome sequencing roughly followed Moore’s law, which predicts exponential declines in computing costs. After 2007, sequencing costs dropped precipitously.

Hayden, Technology: The $1,000 Genome, Nature (2014)
Segmentation in the Field
Segmentation in the Field

Data Analytics / Decision Science
Analysis to drive choices
Segmentation in the Field

Data Analytics / Decision Science
Analysis to drive choices

Data Products
Building Products used by people
Segmentation in the Field

Data Analytics / Decision Science
Analysis to drive choices

Data Products
Building Products used by people

AI / ML / Deep Learning Engineer
Next-level, “Reinforcement learning”
Segmentation in the Field

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
How to connect:
How to connect:

Go to Meet-ups or watch online.
How to connect:

Go to Meet-ups or watch online.

Subscribe to Data Science Weekly.
How to connect:

Go to Meet-ups or watch online.

Subscribe to Data Science Weekly.

Join LinkedIn! Meet up with people.
How to connect:

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

Subscribe to **Data Science Weekly**.

Join **LinkedIn**! Meet up with people.

**Podcasts**: Software Engineering Daily
How to connect:

Go to Meet-ups or watch online.

Subscribe to Data Science Weekly.

Join LinkedIn! Meet up with people.

Podcasts: Software Engineering Daily

Read Twitter!
How to connect:

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?
How to connect:

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!
How to connect:

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! :)
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

7 Weeks
Intro Week Data Project Company Visits

3-5 Weeks
Job Interviews

Bay Area, NYC, Boston and more!
Health Data Challenges

2.7 million patients in the US

Leading cause of stroke
Health Data Challenges

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
Health Data Challenges

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
Health Data Challenges
Health Data Challenges

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

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

Applied LDA topic modeling: a technique from Natural Language Processing
Many exciting questions:
Many exciting questions:

Memorial Sloan Kettering Cancer Center:
Who is a good candidate for clinical trials for treatments?
Many exciting questions:

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?
Many exciting questions:

**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?
Many exciting questions:

**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?
Many exciting questions:

**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?
Many exciting questions:

**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?
Many exciting questions:

**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?
Many exciting questions:

**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?
Many exciting questions:

**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?
Many exciting questions:

**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!