Spark Capabilities
(i.e. Hadoop shortcomings)

- Performance
  - First, use RAM
  - Also, be smarter

- Ease of Use
  - Python, Scala, Java first class citizens

- New Paradigms
  - SparkSQL
  - Streaming
  - MLib
  - GraphX
  - ...more

But using Hadoop as the backing store is a common and sensible option.
Same Idea (improved)

Driver
Python
Scala
Java

CPU
RAM

RDD
Resilient Distributed Dataset
Spark Formula

1. Create/Load RDD
   Webpage visitor IP address log

2. Transform RDD
   "Filter out all non-U.S. IPs"

3. But don’t do anything yet!
   Wait until data is actually needed
   Maybe apply more transforms ("Distinct IPs")

4. Perform Actions that return data
   Count "How many unique U.S. visitors?"
Simple Example

>>> lines_rdd = sc.textFile("nasa_19950801.tsv")

Spark Context

The first thing a Spark program requires is a context, which interfaces with some kind of cluster to use. Our pyspark shell provides us with a convenient \texttt{sc}, using the local filesystem, to start. Your standalone programs will have to specify one:

```python
from pyspark import SparkConf, SparkContext
conf = SparkConf().setMaster("local").setAppName("Test_App")
sc = SparkContext(conf = conf)
```

You would typically run these scripts like so:

```
spark-submit Test_App.py
```
Simple Example

```python
>>> lines_rdd = sc.textFile("nasa_19950801.tsv")
>>> stanfordLines_rdd = lines_rdd.filter(lambda line: "stanford" in line)
>>> stanfordLines_rdd.count()
47
>>> stanfordLines_rdd.first()
'glim.stanford.edu\t\t807258357\tGET\t/shuttle/missions/61-c/61-c-patch-small.gif'
```

**Lambdas**

We’ll see a lot of these. A lambda is simply a function that is too simple to deserve its own subroutine. Anywhere we have a lambda we could also just name a real subroutine that could go off and do anything.

When all you want to do is see if “given an input variable line, is “stanford” in there?”, it isn’t worth the digression.

Most modern languages have adopted this nicety.
## Common Transformations

<table>
<thead>
<tr>
<th>Transformation</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>map(func)</code></td>
<td>Return a new RDD by passing each element through <code>func</code>.</td>
</tr>
<tr>
<td><code>filter(func)</code></td>
<td>Return a new RDD by selecting the elements for which <code>func</code> returns true.</td>
</tr>
<tr>
<td><code>flatMap(func)</code></td>
<td><code>func</code> can return multiple items, and generate a sequence, allowing us to “flatten” nested entries (JSON) into a list.</td>
</tr>
<tr>
<td><code>distinct()</code></td>
<td>Return an RDD with only distinct entries.</td>
</tr>
<tr>
<td><code>sample(...)</code></td>
<td>Various options to create a subset of the RDD.</td>
</tr>
<tr>
<td><code>union(RDD)</code></td>
<td>Return a union of the RDDs.</td>
</tr>
<tr>
<td><code>intersection(RDD)</code></td>
<td>Return an intersection of the RDDs.</td>
</tr>
<tr>
<td><code>subtract(RDD)</code></td>
<td>Remove argument RDD from other.</td>
</tr>
<tr>
<td><code>cartesian(RDD)</code></td>
<td>Cartesian product of the RDDs.</td>
</tr>
<tr>
<td><code>parallelize(list)</code></td>
<td>Create an RDD from this (Python) list (using a spark context).</td>
</tr>
</tbody>
</table>

Full list at [http://spark.apache.org/docs/latest/api/python/pyspark.html#pyspark.RDD](http://spark.apache.org/docs/latest/api/python/pyspark.html#pyspark.RDD)
<table>
<thead>
<tr>
<th>Action</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>collect()</td>
<td>Return all the elements from the RDD.</td>
</tr>
<tr>
<td>count()</td>
<td>Number of elements in RDD.</td>
</tr>
<tr>
<td>countByValue()</td>
<td>List of times each value occurs in the RDD.</td>
</tr>
<tr>
<td>reduce(func)</td>
<td>Aggregate the elements of the RDD by providing a function which combines any two into one (sum, min, max, ...).</td>
</tr>
<tr>
<td>first(), take(n)</td>
<td>Return the first, or first n elements.</td>
</tr>
<tr>
<td>top(n)</td>
<td>Return the n highest valued elements of the RDDs.</td>
</tr>
<tr>
<td>takeSample(...)</td>
<td>Various options to return a subset of the RDD.</td>
</tr>
<tr>
<td>saveAsTextFile(path)</td>
<td>Write the elements as a text file.</td>
</tr>
<tr>
<td>foreach(func)</td>
<td>Run the $func$ on each element. Used for side-effects (updating accumulator variables) or interacting with external systems.</td>
</tr>
</tbody>
</table>

Full list at [http://spark.apache.org/docs/latest/api/python/pyspark.html#pyspark.RDD](http://spark.apache.org/docs/latest/api/python/pyspark.html#pyspark.RDD)
Pair RDDs

- Key/Value organization is a simple, but often very efficient schema, as we mentioned in our NoSQL discussion.

- Spark provides special operations on RDDs that contain key/value pairs. They are similar to the general ones that we have seen.

- On the language (Python, Scala, Java) side key/values are simply tuples. If you have an RDD *all* of whose elements happen to be tuples of two items, it is a Pair RDD and you can use the key/value operations that follow.
### Pair RDD Transformations

<table>
<thead>
<tr>
<th>Transformation</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>reduceByKey(func)</td>
<td>Reduce values using <code>func</code>, but on a key by key basis. That is, combine values with the same key.</td>
</tr>
<tr>
<td>groupByKey()</td>
<td>Combine values with same key. Each key ends up with a list.</td>
</tr>
<tr>
<td>sortByKey()</td>
<td>Return an RDD sorted by key.</td>
</tr>
<tr>
<td>mapValues(func)</td>
<td>Use <code>func</code> to change values, but not key.</td>
</tr>
<tr>
<td>keys()</td>
<td>Return an RDD of only keys.</td>
</tr>
<tr>
<td>values()</td>
<td>Return an RDD of only values.</td>
</tr>
</tbody>
</table>

Note that all of the regular transformations are available as well.
Pair RDD Actions

As with transformations, all of the regular actions are available to Pair RDDs, and there are some additional ones that can take advantage of key/value structure.

<table>
<thead>
<tr>
<th>Action</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>countByKey()</td>
<td>Count the number of elements for each key.</td>
</tr>
<tr>
<td>lookup(key)</td>
<td>Return all the values for this key.</td>
</tr>
</tbody>
</table>

Full list at http://spark.apache.org/docs/latest/api/python/pyspark.html#pyspark.RDD
<table>
<thead>
<tr>
<th>Transformation</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>subtractByKey(otherRDD)</td>
<td>Remove elements with a key present in other RDD.</td>
</tr>
<tr>
<td>join(otherRDD)</td>
<td>Inner join: Return an RDD containing all pairs of elements with matching keys in self and other. Each pair of elements will be returned as a ((k, (v_1, v_2))) tuple, where ((k, v_1)) is in self and ((k, v_2)) is in other.</td>
</tr>
<tr>
<td>leftOuterJoin(otherRDD)</td>
<td>For each element ((k, v)) in self, the resulting RDD will either contain all pairs ((k, (v, w))) for (w) in other, or the pair ((k, (v, None))) if no elements in other have key (k).</td>
</tr>
<tr>
<td>rightOuterJoin(otherRDD)</td>
<td>For each element ((k, w)) in other, the resulting RDD will either contain all pairs ((k, (v, w))) for (v) in this, or the pair ((k, (None, w))) if no elements in self have key (k).</td>
</tr>
<tr>
<td>cogroup(otherRDD)</td>
<td>Group data from both RDDs by key.</td>
</tr>
</tbody>
</table>
Joins Are Quite Useful

Any database designer can tell you how common joins are. Let's look at a simple example. We have (here we create it) an RDD of our top purchasing customers.

And an RDD with all of our customers' addresses.

To create a mailing list of special coupons for those favored customers we can use a join on the two datasets.

```python
>>> best_customers_rdd = sc.parallelize(["Joe", "$103"], ["Alice", "$2000"], ["Bob", "$1200"])

>>> customer_addresses_rdd = sc.parallelize(["Joe", "23 State St."], ["Frank", "555 Timer Lane"], ["Sally", "44 Forest Rd."], ["Alice", "3 Elm Road"], ["Bob", "88 West Oak"])

>>> promotion_mail_rdd = best_customers_rdd.join(customer_addresses_rdd)

>>> promotion_mail_rdd.collect()
[('Bob', ('$1200', '88 West Oak')), ('Joe', ('$103', '23 State St.')), ('Alice', ('$2000', '3 Elm Road'))]
```
Shakespeare, a Data Analytics Favorite

Applying data analytics to the works of Shakespeare has become all the rage. Whether determining the legitimacy of his authorship (it wasn’t Marlowe) or if Othello is actually a comedy (perhaps), it is amazing how much publishable research has sprung from the recent analysis of 400 year old text.

We’re going to do some exercises here using a text file containing all of his works.
Who needs this Spark stuff?

As we do our first Spark exercises, you might think of several ways to accomplish these tasks that you already know. For example, Python Pandas is a fine way to do our following problem, and it will probably work on your laptop reasonably well.

However we are learning how to leverage scalable techniques that work on very big data. Shortly, we will encounter problems that are considerable in size, and you will leave this workshop knowing how to harness very large resources.

Searching the Complete Works of William Shakespeare for patterns is a lot different from searching the entire Web (perhaps as the 800TB Common Crawl dataset).

So everywhere you see an RDD, realize that it is a actually a parallel databank that could scale to PBs.
Some Simple Problems

We have an input file, Complete _Shakespeare.txt, that you can also find at http://www.gutenberg.org/ebooks/100. You might find http://spark.apache.org/docs/latest/api/python/pyspark.html#pyspark.RDD useful to have in a browser window.

If you are starting from scratch on the login node:
1) interact 2) cd BigData/Shakespeare 3) module load spark 4) pyspark
...

```python
>>> rdd = sc.textFile("Complete_Shakespeare.txt")
```

Let’s try a few simple exercises.

1) Count the number of lines

2) Count the number of words (hint: Python "split" is a workhorse)

3) Count unique words

4) Count the occurrence of each word

5) Show the top 5 most frequent words

These last two are a bit more challenging. One approach is to think “key/value”. If you go that way, think about which data should be the key and don’t be afraid to swap it about with value. This is a very common manipulation when dealing with key/value organized data.
Some Simple Answers

>>> lines_rdd = sc.textFile("Complete_Shakespeare.txt")
>>> lines_rdd.count()
124787
>>> words_rdd = lines_rdd.flatMap(lambda x: x.split())
>>> words_rdd.count()
904061
>>> words_rdd.distinct().count()
67779

Next, I know I'd like to end up with a pair RDD of sorted word/count pairs:

(23407, 'the'), (19540, 'I'), (15682, 'to'), (15649, 'of') ...

Why not just words_rdd.countByValue()? It is an action that gives us a massive Python unsorted dictionary of results:

... 1, u'precious-princely': 1, u'christenings?': 1, 'empire': 11, u'vaunts': 2, u'Lubber's': 1, u'poet.': 2, u'Toad!': 1, u'leaden': 15, u'captains': 1, u'leaf': 9, u'Barnes,' : 1, u'lead': 101, u'"Hell": 1, u'wheat,' : 3, u'lean': 28, u'Toad,' : 1, u'trencher!': 2, u'1.F.2.: 1, u'leas': 2, u'leap': 17, ...

Where to go next? Sort this in Python or try to get back into an RDD? If this is truly BIG data, we want to remain as an RDD until we reach our final results. So, no.
Things data scientists do.

- Turn these into k/v pairs
- Reduce to get words counts
- Flip keys and values so we can sort on wordcount instead of words.
Some Homework Problems

To do research-level text analysis, we generally want to clean up our input. Here are some of the kinds of things you could do to get a more meaningful distinct word count.

1) Remove punctuation. Often punctuation is just noise, and it is here. Do a Map and/or Filter (some punctuation is attached to words, and some is not) to eliminate all punctuation from our Shakespeare data. Note that if you are familiar with regular expressions, Python has a ready method to use those.

2) Remove stop words. Stop words are common words that are also often uninteresting ("I", "the", "a"). You can remove many obvious stop words with a list of your own, and the MLlib that we are about to investigate has a convenient StopWordsRemover() method with default lists for various languages.

3) Stemming. Recognizing that various different words share the same root ("run", "running") is important, but not so easy to do simply. Once again, Spark brings powerful libraries into the mix to help. A popular one is the Natural Language Tool Kit. You should look at the docs, but you can give it a quick test quite easily:

```python
import nltk
from nltk.stem.porter import *
stemmer = PorterStemmer()
stems_rdd = words_rdd.map(lambda x: stemmer.stem(x))
```
Optimizations

We said one of the advantages of Spark is that we can control things for better performance. There are a multitude of optimization, performance, tuning and programmatic features to enable better control. We quickly look at a few of the most important.

- Persistence
- Partitioning
- Parallel Programming Capabilities
- Performance and Debugging Tools
Persistence

- Lazy evaluation implies by default that all the RDD dependencies will be computed when we call an action on that RDD.

- If we intend to use that data multiple times (say we are filtering some log, then dumping the results, but we will analyze it further) we can tell Spark to persist the data.

- We can specify different levels of persistence: `MEMORY_ONLY, MEMORY_ONLY_SER, MEMORY_AND_DISK, MEMORY_AND_DISK_SER, DISK_ONLY`

```python
>>> lines_rdd = sc.textFile("nasa_19950801.tsv")
>>> stanfordLines_rdd = lines.filter(lambda line: "stanford" in line)
>>> stanfordLines_rdd.persist(StorageLevel.MEMORY_AND_DISK)
>>> stanfordLines_rdd.count()
47
>>> stanfordLines_rdd.first(1)
[u'glim.stanford.edu\t\t807258394\tGET\t/shuttle/.../orbiters-logo.gif\t200\t1932\t\t'
  
  .
  
  .
>>> stanfordLines.unpersist()
```
Partitions

• Spark distributes the data of your RDDs across its resources. It tries to do some obvious things.

• With key/value pairs we can help keep that data grouped efficiently.

• We can create custom partitioners that beat the default (which is probably a hash or maybe range).

• Use persist() if you have partitioned your data in some smart way. Otherwise it will keep getting re-partitioned.
Spark has several parallel programming features that make it easier and more efficient to do operations in parallel in a more explicit way.

Accumulators are variables that allow many copies of a variable to exist on the separate worker nodes.

It is also possible to have replicated data that we would like all the workers to have access to. Perhaps a lookup table of IP addresses to country codes so that each worker can transform or filter on such information. Maybe we want to exclude all non-US IP entries in our logs. You might think of ways you could do this just by passing variables, but they would likely be expensive in actual operation (usually requiring multiple sends). The solution in Spark is to send an (immutable, read only) broadcast variable.

```python
log = sc.textFile("logs")
blanks = sc.accumulator(0)
def tokenizeLog(line):
    global blanks
    if (line == "")
        blanks += 1
    return line.split(" ")
entries = log.flatMap(tokenizeLog)
entries.saveAsTextFile("parsedlogs.txt")
print "Blank entries: %d" blanks.value
```

```python
log = sc.textFile("log.txt")
IPtable = sc.broadcast(loadIPTable())
def countryFilter(IPentry, IPtable):
    return (IPentry.prefix() in IPTable)
USentries = log.filter(countryFilter)
```
Performance & Debugging

We will give unfortunately short shrift to performance and debugging, which are both important. Mostly, this is because they are very configuration and application dependent.

Here are a few things to at least be aware of:

- **SparkConf() class.** A lot of options can be tweaked here.
- **Spark Web UI.** A very friendly way to explore all of these issues.
Spark has an impressive, and growing, list of input/output formats it supports. Some important ones:

- Text
- CSV
- SQL type Query/Load
  - JSON (can infer schema)
  - Parquet
  - Hive
  - XML
  - Sequence (Hadoop key/value)
  - Databases: JDBC, Cassandra, HBase, MongoDB, etc.
- Compression (gzip...)

And it can interface directly with a variety of filesystems: local, HDFS, Lustre, Amazon S3,...
Spark Streaming

Spark addresses the need for streaming processing of data with an API that divides the data into batches, which are then processed as RDDs.

There are features to enable:

- Fast recovery from failures or timeouts
- Load balancing
- Integration with static data and interactive queries
- Integration with other components (SQL, Machine Learning)

15% of the "global datasphere" (quantification of the amount of data created, captured, and replicated across the world) is currently real-time. That number is growing quickly both in absolute terms and as a percentage.
MLib rolls in a lot of classic machine learning algorithms. We barely have time to touch upon this interesting topic today, but they include:

- Useful data types
- Basic Statistics
- Classification (including SVMs, Random Forests)
- Regression
- Algorithms (SGD,...)
- Clustering...
One of the reasons we use spark is for easy access to powerful data analysis tools. The MLlib library gives us a machine learning library that is easy to use and utilizes the scalability of the Spark system.

It has supported APIs for Python (with NumPy), R, Java and Scala.

We will use the Python version in a generic manner that looks very similar to any of the above implementations.

There are good example documents for the clustering routine we are using here:

http://spark.apache.org/docs/latest/mllib-clustering.html

And an excellent API reference document here:

http://spark.apache.org/docs/latest/api/python/pyspark.mllib.html#pyspark.mllib.clustering.KMeans

I suggest you use these pages for all your Spark work.
Clustering is a very common operation for finding grouping in data and has countless applications. This is a very simple example, but you will find yourself reaching for a clustering algorithm frequently in pursuing many diverse machine learning objectives, sometimes as one part of a pipeline.
Clustering

As intuitive as clustering is, it presents challenges to implement in an efficient and robust manner.

You might think this is trivial to implement in lower dimensional spaces.

But it can get tricky even there.

Sometimes you know how many clusters you have to start with. Often you don’t. How hard can it be to count clusters? How many are here?

We will start with 5000 2D points. We want to figure out how many clusters there are, and their centers. Let’s fire up pyspark and get to it...
Finding Clusters

Using Python version 2.7.5 (default, Nov 20 2015 02:00:19)
SparkContext available as sc, HiveContext available as sqlContext.

```python
>>> rdd1 = sc.textFile('5000_points.txt')
>>> rdd2 = rdd1.map(lambda x: x.split())
>>> rdd3 = rdd2.map(lambda x: [int(x[0]), int(x[1])])
```

```
br06% interact
...
```

```
r288% module load spark
r288% pyspark
```
```python
>>> rdd1 = sc.textFile("5000_points.txt")
>>> rdd1.count()
5000
>>> rdd1.take(4)
[u' 664159  550946', u' 665845  557965', u' 597173  575538', u' 618600  551446']
>>> rdd2 = rdd1.map(lambda x:x.split())
>>> rdd2.take(4)
[[u'664159', u'550946'], [u'665845', u'557965'], [u'597173', u'575538'], [u'618600', u'551446']]
>>> rdd3 = rdd2.map(lambda x: [int(x[0]),int(x[1])])
>>> rdd3.take(4)
[[664159, 550946], [665845, 557965], [597173, 575538], [618600, 551446]]
```
Finding Clusters

Using Python version 2.7.5 (default, Nov 20 2015 02:00:19)
SparkContext available as sc, HiveContext available as sqlContext.

```python
>>> rdd1 = sc.textFile("5000_points.txt")
>>> rdd2 = rdd1.map(lambda x:x.split())
>>> rdd3 = rdd2.map(lambda x: [int(x[0]),int(x[1])])
>>> from pyspark.mllib.clustering import KMeans
```

```
class KMeans

     Train a k-means clustering model.

     Parameters:
     * k -- Number of clusters to create.
     * maxIterations -- Maximum number of iterations allowed. (default: 100)
     * runs -- This param has no effect since Spark 2.0.0.
     * initializationMode -- The initialization algorithm. This can be either "random" or "k-means||". (default: "k-means||")
     * seed -- Random seed value for cluster initialization. Set as None to generate seed based on system time. (default: None)
     * initializationSteps -- Number of steps for the k-means|| initialization mode. This is an advanced setting -- the default of 5 is almost always enough. (default: 5)
     * epsilon -- Distance threshold within which a center will be considered to have converged. If all centers move less than this Euclidean distance, iterations are stopped. (default: 1e-4)
     * initModel -- Initial cluster centers can be provided as a KMeansModel object rather than using the random or k-means|| initializationModel. (default: None)
```
Finding Clusters

![Graph showing data points representing finding clusters.](image-url)
Finding Clusters

Let’s see results for 1-30 cluster tries

```
version 1.6.0

Using Python version 2.7.5 (default, Nov 20 2015 02:00:19)
SparkContext available as sc, HiveContext available as sqlContext.

>>> rdd1 = sc.textFile("5000_points.txt")
>>> rdd2 = rdd1.map(lambda x:x.split())
>>> rdd3 = rdd2.map(lambda x: [int(x[0]),int(x[1])])

>>> from pyspark.mllib.clustering import KMeans
>>> for clusters in range(1,30):
...    model = KMeans.train(rdd3, clusters)
...    print (clusters, model.computeCost(rdd3))
```
```python
>>> for trials in range(10):
...     print
...     for clusters in range(12,18):
...         model = KMeans.train(rdd3,clusters)
...         print (clusters, model.computeCost(rdd3))
```

<table>
<thead>
<tr>
<th>Clusters</th>
<th>Compute Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>2.45472346524e+13</td>
</tr>
<tr>
<td>13</td>
<td>2.00175423869e+13</td>
</tr>
<tr>
<td>14</td>
<td>1.90313863726e+13</td>
</tr>
<tr>
<td>15</td>
<td>1.52746006962e+13</td>
</tr>
<tr>
<td>16</td>
<td>8.67526114029e+12</td>
</tr>
<tr>
<td>17</td>
<td>8.49571894386e+12</td>
</tr>
<tr>
<td>12</td>
<td>2.62619056924e+13</td>
</tr>
<tr>
<td>13</td>
<td>2.9031673822e+13</td>
</tr>
<tr>
<td>14</td>
<td>1.52308079405e+13</td>
</tr>
<tr>
<td>15</td>
<td>8.91765957989e+12</td>
</tr>
<tr>
<td>16</td>
<td>8.70736515113e+12</td>
</tr>
<tr>
<td>17</td>
<td>8.49616440477e+12</td>
</tr>
<tr>
<td>12</td>
<td>2.5524719797e+13</td>
</tr>
<tr>
<td>13</td>
<td>2.14332949698e+13</td>
</tr>
<tr>
<td>14</td>
<td>1.11070395905e+13</td>
</tr>
<tr>
<td>15</td>
<td>1.47792736325e+13</td>
</tr>
<tr>
<td>16</td>
<td>1.85736955725e+13</td>
</tr>
<tr>
<td>17</td>
<td>8.42795740134e+12</td>
</tr>
<tr>
<td>12</td>
<td>2.31466520037e+13</td>
</tr>
<tr>
<td>13</td>
<td>1.91856542103e+13</td>
</tr>
<tr>
<td>14</td>
<td>1.49332023312e+13</td>
</tr>
<tr>
<td>15</td>
<td>1.3506302755e+13</td>
</tr>
<tr>
<td>16</td>
<td>8.7757678836e+12</td>
</tr>
<tr>
<td>17</td>
<td>1.60075548613e+13</td>
</tr>
<tr>
<td>12</td>
<td>2.5187054064e+13</td>
</tr>
<tr>
<td>13</td>
<td>1.83498739266e+13</td>
</tr>
<tr>
<td>14</td>
<td>1.96076943156e+13</td>
</tr>
<tr>
<td>15</td>
<td>1.4172566214e+13</td>
</tr>
<tr>
<td>16</td>
<td>1.41986217172e+13</td>
</tr>
<tr>
<td>17</td>
<td>8.46755159547e+12</td>
</tr>
<tr>
<td>12</td>
<td>2.38234539188e+13</td>
</tr>
<tr>
<td>13</td>
<td>1.85101922046e+13</td>
</tr>
<tr>
<td>14</td>
<td>1.91732620477e+13</td>
</tr>
<tr>
<td>15</td>
<td>8.91769396968e+12</td>
</tr>
<tr>
<td>16</td>
<td>8.64876051004e+12</td>
</tr>
<tr>
<td>17</td>
<td>8.54677681587e+12</td>
</tr>
<tr>
<td>12</td>
<td>2.5187054064e+13</td>
</tr>
<tr>
<td>13</td>
<td>2.04031903147e+13</td>
</tr>
<tr>
<td>14</td>
<td>1.95213876047e+13</td>
</tr>
<tr>
<td>15</td>
<td>1.93000628589e+13</td>
</tr>
<tr>
<td>16</td>
<td>2.07670831868e+13</td>
</tr>
<tr>
<td>17</td>
<td>8.47797102908e+12</td>
</tr>
<tr>
<td>12</td>
<td>2.39830397362e+13</td>
</tr>
<tr>
<td>13</td>
<td>2.00248378195e+13</td>
</tr>
<tr>
<td>14</td>
<td>1.34867313762e+13</td>
</tr>
<tr>
<td>15</td>
<td>2.09299321238e+13</td>
</tr>
<tr>
<td>16</td>
<td>1.32266735736e+13</td>
</tr>
<tr>
<td>17</td>
<td>8.50857884943e+12</td>
</tr>
</tbody>
</table>
for trials in range(10):  # Try ten times to find best result
...    for clusters in range(12, 16):  # Only look in interesting range
...        model = KMeans.train(rdd3, clusters)
...        cost = model.computeCost(rdd3)
...        centers = model.clusterCenters
...        if cost<1e+13:  # Let's grab cluster centers
...            print (clusters, cost)  # If result is good, print it out
...            for coords in centers:
...                print (int(coords[0]), int(coords[1]))
...            break
...
We are going to find a recurring theme throughout machine learning:

- Our data naturally resides in higher dimensions
- Reducing the dimensionality makes the problem more tractable
- And simultaneously provides us with insight

This last two bullets highlight the principle that "learning" is often finding an effective compressed representation.

As we return to this theme, we will highlight these slides with our Dimensionality Reduction badge so that you can follow this thread and appreciate how fundamental it is.
Why all these dimensions?

The problems we are going to address, as well as the ones you are likely to encounter, are naturally highly dimensional. If you are new to this concept, let's look at an intuitive example to make it less abstract.

<table>
<thead>
<tr>
<th>Category</th>
<th>Purchase Total ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children's Clothing</td>
<td>$800</td>
</tr>
<tr>
<td>Pet Supplies</td>
<td>$0</td>
</tr>
<tr>
<td>Cameras (Dash, Security, Baby)</td>
<td>$450</td>
</tr>
<tr>
<td>Containers (Storage)</td>
<td>$350</td>
</tr>
<tr>
<td>Romance Book</td>
<td>$0</td>
</tr>
<tr>
<td>Remodeling Books</td>
<td>$80</td>
</tr>
<tr>
<td>Sporting Goods</td>
<td>$25</td>
</tr>
<tr>
<td>Children's Toys</td>
<td>$378</td>
</tr>
<tr>
<td>Power Tools</td>
<td>$0</td>
</tr>
<tr>
<td>Computers</td>
<td>$0</td>
</tr>
<tr>
<td>Garden</td>
<td>$0</td>
</tr>
<tr>
<td>Children's Books</td>
<td>$180</td>
</tr>
</tbody>
</table>

This is a 2900 dimensional vector.
Why all these dimensions?

If we apply our newfound clustering expertise, we might find we have 80 clusters (with an acceptable error).

People spending on “child’s toys” and “children’s clothing” might cluster with “child’s books” and, less obvious, "cameras (Dashcams, baby monitors and security cams)", because they buy new cars and are safety conscious. We might label this cluster "Young Parents". We also might not feel obligated to label the clusters at all. We can now represent any customer by their distance from these 80 clusters.

<table>
<thead>
<tr>
<th>Customer Representation</th>
<th>80 dimensional vector.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster</td>
<td>Young Parents</td>
</tr>
<tr>
<td>Distance</td>
<td>0.02</td>
</tr>
</tbody>
</table>

We have now accomplished two things:

• we have compressed our data
• learned something about our customers (who to send a dashcam promo to).
This is a good time to point out how our intuition can lead us astray as we increase the dimensionality of our problems - which we will certainly be doing - and to a great degree. There are several related aspects to this phenomenon, often referred to as the *Curse of Dimensionality*. One root cause of confusion is that our notion of Euclidian distance starts to fail in higher dimensions.

These plots show the distributions of pairwise distances between randomly distributed points within differently dimensioned unit hypercubes. Notice how all the points start to be about the same distance apart.

Once can imagine this makes life harder on a clustering algorithm!

There are other surprising effects: random vectors are almost all orthogonal; the unit sphere takes almost no volume in the unit square. These cause all kinds of problems when generalizing algorithms from our lowly 3D world.
Even the definition of distance (the *metric*) can vary based upon application. If you are solving chess problems, you might find the Manhattan distance (or taxicab metric) to be most useful.

For comparing text strings, we might choose one of dozens of different metrics. For spell checking you might want one that is good for phonetic distance, or maybe edit distance. For natural language processing (NLP), you probably care more about tokens.

For genomics, you might care more about string sequences.

Some useful measures don’t even qualify as metrics (usually because they fail the triangle inequality: \( a + b \geq c \)).
Alternative DR: Principal Component Analysis

3D Data Set

Maybe mostly 1D!
Alternative DR: Principal Component Analysis

Flatter 2D-ish Data Set

View down the 1st Princ. Comp.
Let's look at one more example today. Suppose we are trying to do a Zillow type of analysis and predict home values based upon available factors. We may have an entry (vector) for each home that captures this kind of data:

<table>
<thead>
<tr>
<th>Home Data</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Latitude</td>
<td>4833438 north</td>
</tr>
<tr>
<td>Longitude</td>
<td>630084 east</td>
</tr>
<tr>
<td>Last Sale Price</td>
<td>$ 480,000</td>
</tr>
<tr>
<td>Last Sale Year</td>
<td>1998</td>
</tr>
<tr>
<td>Width</td>
<td>62</td>
</tr>
<tr>
<td>Depth</td>
<td>40</td>
</tr>
<tr>
<td>Floors</td>
<td>3</td>
</tr>
<tr>
<td>Bedrooms</td>
<td>3</td>
</tr>
<tr>
<td>Bathrooms</td>
<td>2</td>
</tr>
<tr>
<td>Garage</td>
<td>2</td>
</tr>
<tr>
<td>Yard Width</td>
<td>84</td>
</tr>
<tr>
<td>Yard Depth</td>
<td>60</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

There may be some opportunities to reduce the dimension of the vector here. Perhaps clustering on the geographical coordinates...
Principal Component Analysis Fail

1st Component Off
Data Not Very Linear

H x W Is Not Linear
But (HxW) Fits Well

Non-Linear PCA?
A Better Approach Tomorrow!
As mentioned earlier, an appreciation for having some defined structure to your data has come back into vogue. For one, because it simply makes sense and naturally emerges in many applications. Often even more important, it can greatly aid optimization, especially with the Java VM that Spark uses.

For both of these reasons, you will see that the newest set of APIs to Spark are DataFrame based. Sound leading-edge? This is simply SQL type columns. Very similar to Python pandas DataFrames (but based on RDDs, so not exactly).

We haven't prioritized them here because they aren't necessary, and some of the pieces aren't mature. But some of the latest features use them.
Creating DataFrames

It is very pretty intuitive to utilize DataFrames. Your elements just have labeled columns.

A **row RDD** is the basic way to go from RDD to DataFrame, and back, if necessary. A "row" is just a tuple.

```python
>>> row_rdd = sc.parallelize([ ("Joe","Pine St.","PA",12543), ("Sally","Fir Dr.","WA",78456), ("Jose","Elm Pl.","ND",45698) ])
```

```python
>>> aDataFrameFromRDD = spark.createDataFrame( row_rdd, ["name", "street", "state", "zip"] )
```

```python
>>> aDataFrameFromRDD.show()
+---------+-------+-----+----+
| name    | street| state|  zip|
|---------+-------+-----+----+
|  Joe    | Pine St. | PA  |12543|
| Sally   | Fir Dr.  | WA  |78456|
| Jose    | Elm Pl.  | ND  |45698|
+---------+-------+-----+----+
```
Creating DataFrames

You will come across DataFrames created without a schema. They get default column names.

```python
>>> noSchemaDataFrame = spark.createDataFrame(row_rdd)
>>> noSchemaDataFrame.show()
+-----+--------+-----+-----+
| _1  | _2     | _3  | _4  |
+-----+--------+-----+-----+
| Joe | Pine St.| PA  | 12543|
| Sally| Fir Dr.| WA  | 78456|
| Jose| Elm Pl.| ND  | 45698|
+-----+--------+-----+-----+
```

And you can create them inline as well.

```python
>>> directDataFrame = spark.createDataFrame([("Joe","Pine St.","PA",12543),
                                           ("Sally","Fir Dr.","WA",78456),
                                           ("Jose","Elm Pl.","ND",45698)],
                                           ["name", "street", "state", "zip"])
```

### Datasets

Spark has added a variation (technically a superset) of DataFrames called Datasets. For compiled languages with strong typing (Java and Scala) these provide static typing and can detect some errors at compile time.

This is not relevant to Python or R.
Speaking of pandas, or SciPy, or...

Some of you may have experience with the many Python libraries that accomplish some of these tasks. Immediately relevant to today, *pandas* allows us to sort and query data, and *SciPy* provides some nice clustering algorithms. So why not just use them?

The answer is that Spark does these things in the context of having potentially huge, parallel resources at hand. We don't notice it as Spark is also convenient, but behind every Spark call:

- every RDD could be many TB in size
- every transform could use many thousands of cores and TB of memory
- every algorithm could also use those thousands of cores

So don't think of Spark as just a data analytics library because our exercises are modest. You are learning how to cope with *Big Data*. 
Run My Programs Or Yours

exec()

[urbanic@r001 ~]$ pyspark
Python 3.7.4 (default, Aug 13 2019, 20:35:49)
Type 'copyright', 'credits' or 'license' for more information
IPython 7.8.0 -- An enhanced Interactive Python. Type '?' for help.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Welcome to

Welcome to

version 3.0.0-preview2

Using Python version 3.7.4 (default, Aug 13 2019 20:35:49)
SparkSession available as 'spark'
In [1]: exec(open("./clustering.py").read())
1 5.76807041184e+14
2 3.73234816206e+14
3 2.13508993715e+14
4 1.38250712993e+14
5 1.2632806251e+14
6 7.97690150116e+13
7 7.14156965883e+13
8 5.7815194802e+13
...
...
How does all this fit together?

Big Data

Character Recognition
Capchas
Chess
Go

Deep Learning

Artificial Intelligence