



Transportation Data Science at NREL

Adam Duran, Kenneth Kelly, Caleb Phillips

5/22/2018

Why do you need a Big Data factory?

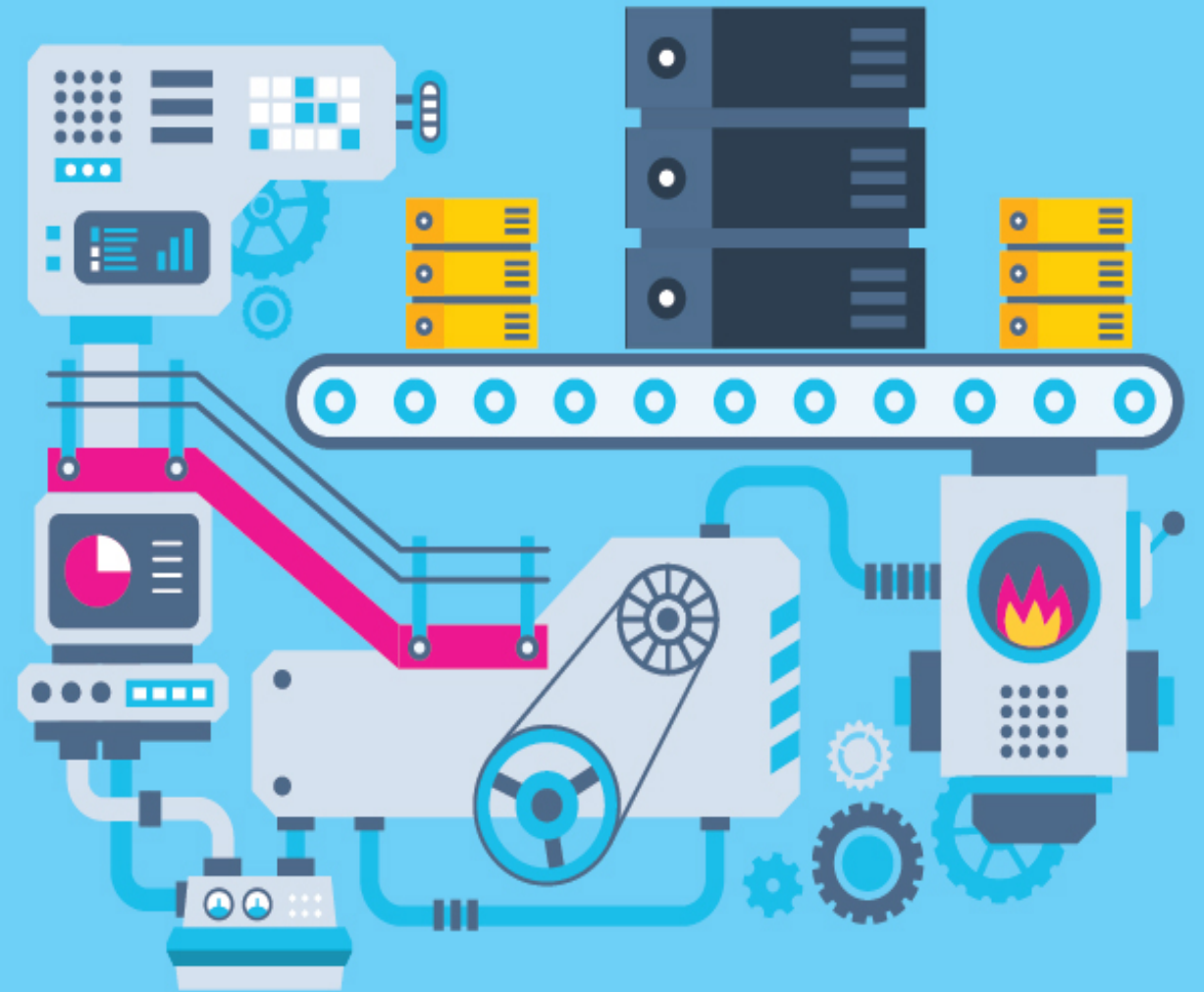
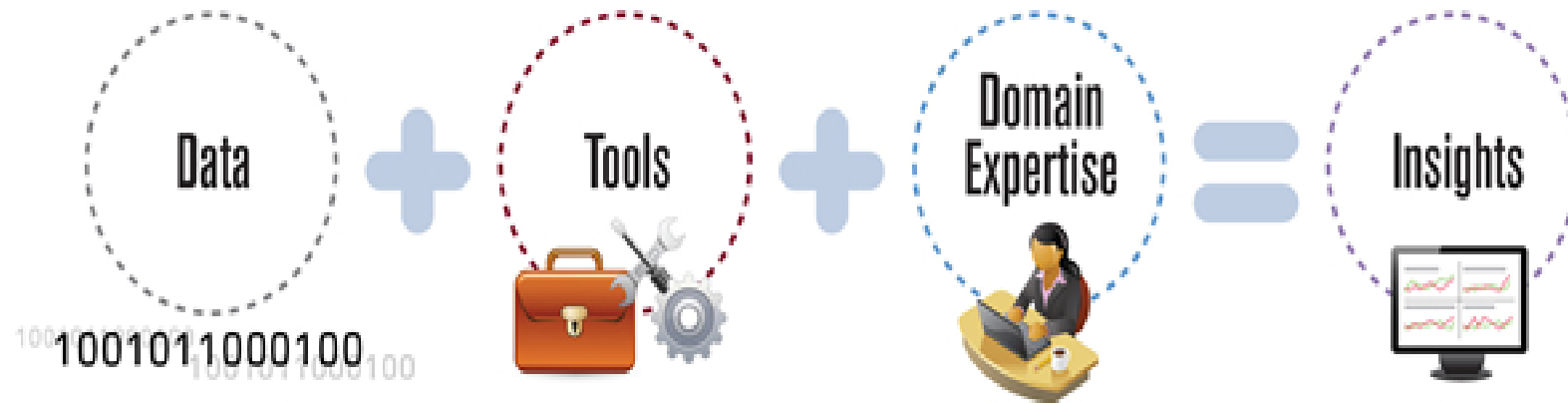


Image - <http://www.cmsitservices.com/blog/why-do-you-need-a-big-data-factory/>

BLANKET ANY BECOMES COLLECTION
 TOOLS TERM TRANSFER ANALYSIS INCLUDE
 DATA SHARING CAPTURE
STORAGE MANAGEMENT PROCESSING
BIG DATA AND DEPENDENCE
 TEAM
 TIME ENABLE TRIG FWHAT CAPABIL
 ENTERPRISE BUS
 DECISI PACKAGES GEN
 OPTIMIZA PETA
 3V
 COMPLEX
 FOR ASSETS
 RELATIONAL METEOROLOGY RANGE
VOLUME THAT ENVIRONMENTAL
 NEW SOURCES DOUG LANAY
 SYSTEMS SIZE GOVERNMENT



How Big is Big? The 5 Vs of Big Data

Do you have big data?

Volume – how big?

Variety – what type and nature?

Velocity – how fast does it arrive?

Variability – are there inconsistencies?

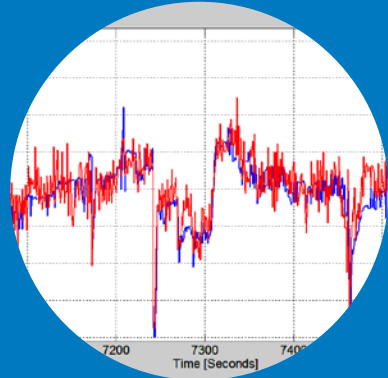
Veracity – challenging to assure quality?

Big data and machine learning challenges exist **across all industries**

Machine Learning ❤️ Big Data



Do you need a truck?



Timeseries

- 1 Hz CAN/OBD and Instrument Data
- Fuel Rates
- Vehicle Speed
- Engine and Emissions Parameters



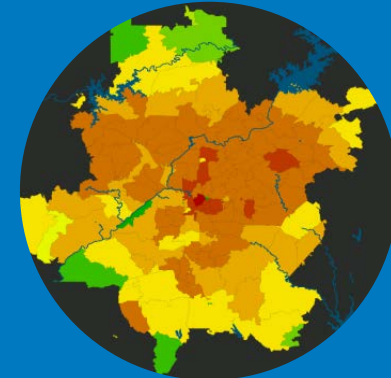
Geospatial

- 1Hz GPS Data
- Latitude
- Longitude
- Elevation
- Heading



Categorical

- Vehicle Classifiers for Sorting Results
- Weight Class
- Transmission
- Fuel
- Body



Supplemental

- Road Networks
- Infrastructure
- Solar Exposure
- Climate and Temperature



Structured vs. Unstructured Data

Structured:

- Traditional Databases (SQL)

Semi-Structured:

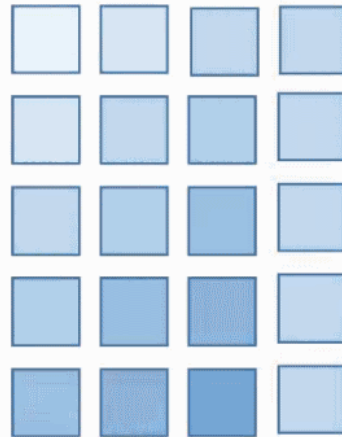
- XML
- JSON

Unstructured:

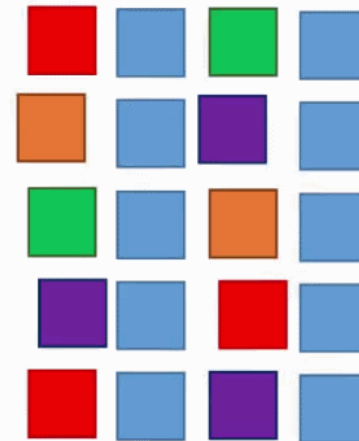
- Text
- Images
- Audio

Structured, Unstructured and Semi-Structured

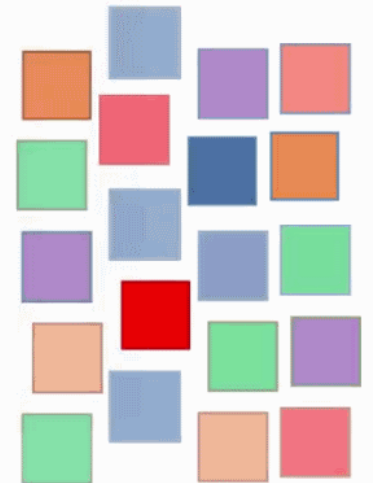
Structured Data



Semi-Structured Data



Unstructured Data



What do we do with all of this data? Enter the Data Scientist!

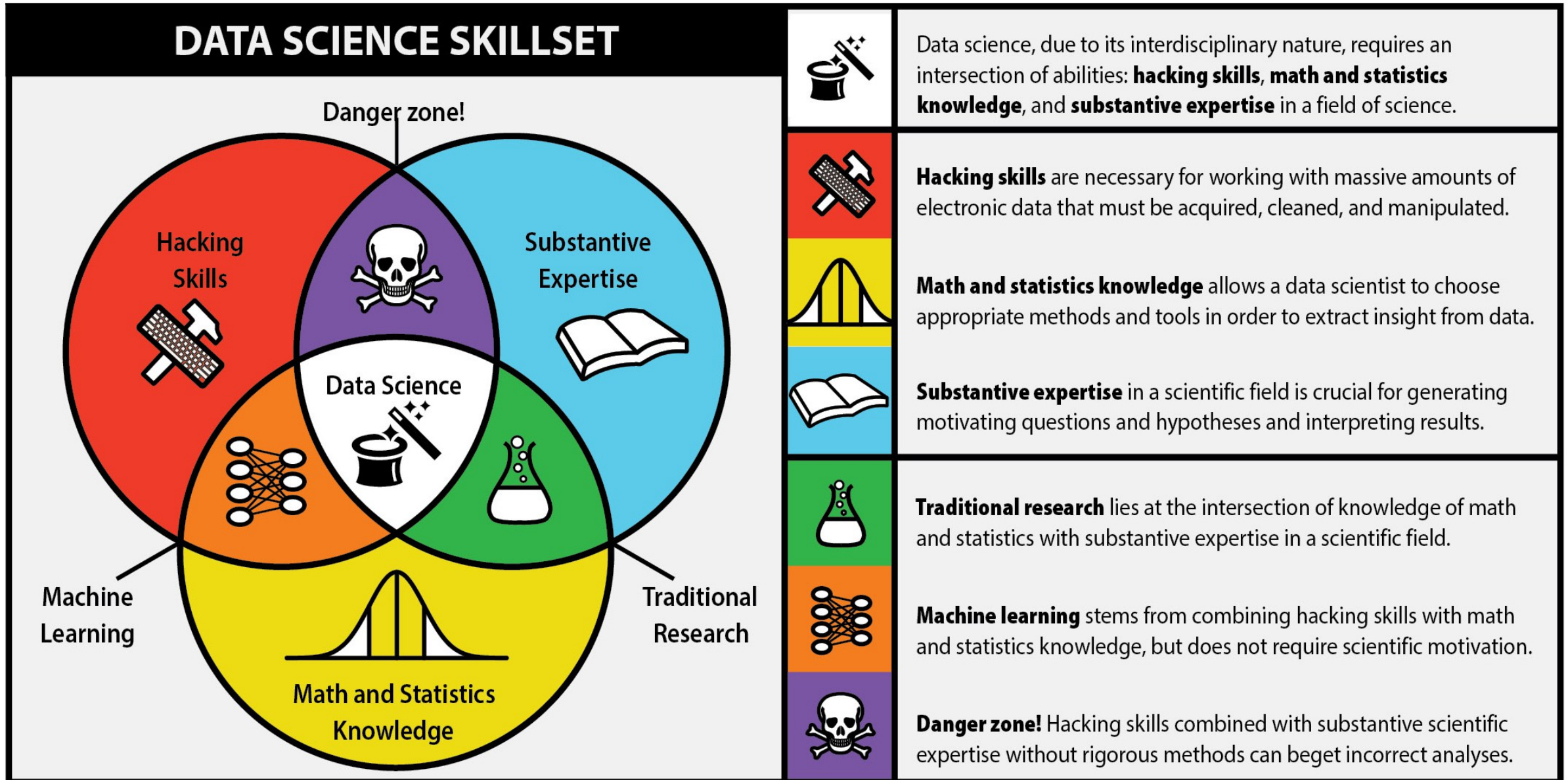


Image - <http://berkeleysciencereview.com/how-to-become-a-data-scientist-before-you-graduate/>

Computational Sciences Center

- HPC Systems and Operations
- **Data Analysis and Visualization**
- Simulation and Optimization
- Algorithms and Fluid Dynamics

ESIF High Performance Computing (HPC) Center



NREL Data Resource Landscape

Established

- Peregrine
 - Parallel File system
 - Mass Storage
 - Visualization



l.u.s.t.r.e.
File System

- Hitachi Storage
- Relational Database Servers



PostgreSQL

- Timeseries Cluster
- ESIF Data Repository
- Data Relays



- APIs & Web services



- Invites external collaborators

NREL Data Resource Landscape

Established

- Peregrine
 - Parallel File system
 - Mass Storage
 - Visualization
- Hitachi Storage
- Relational Database Servers
- Timeseries Cluster
- ESIF Data Repository
- Data Relays
- APIs & Web services
- Invites external collaborators



l.u.s.t.r.e.
File System



PostgreSQL



Emerging

- Sparkplug
 - Openstack
 - Spark
 - Hadoop
 - Kafka
- Scalable Attached (Object) Storage
- Peregrine 2 (August!)
 - HPC -> Big Data
- Scalable Relational Databases



ceph

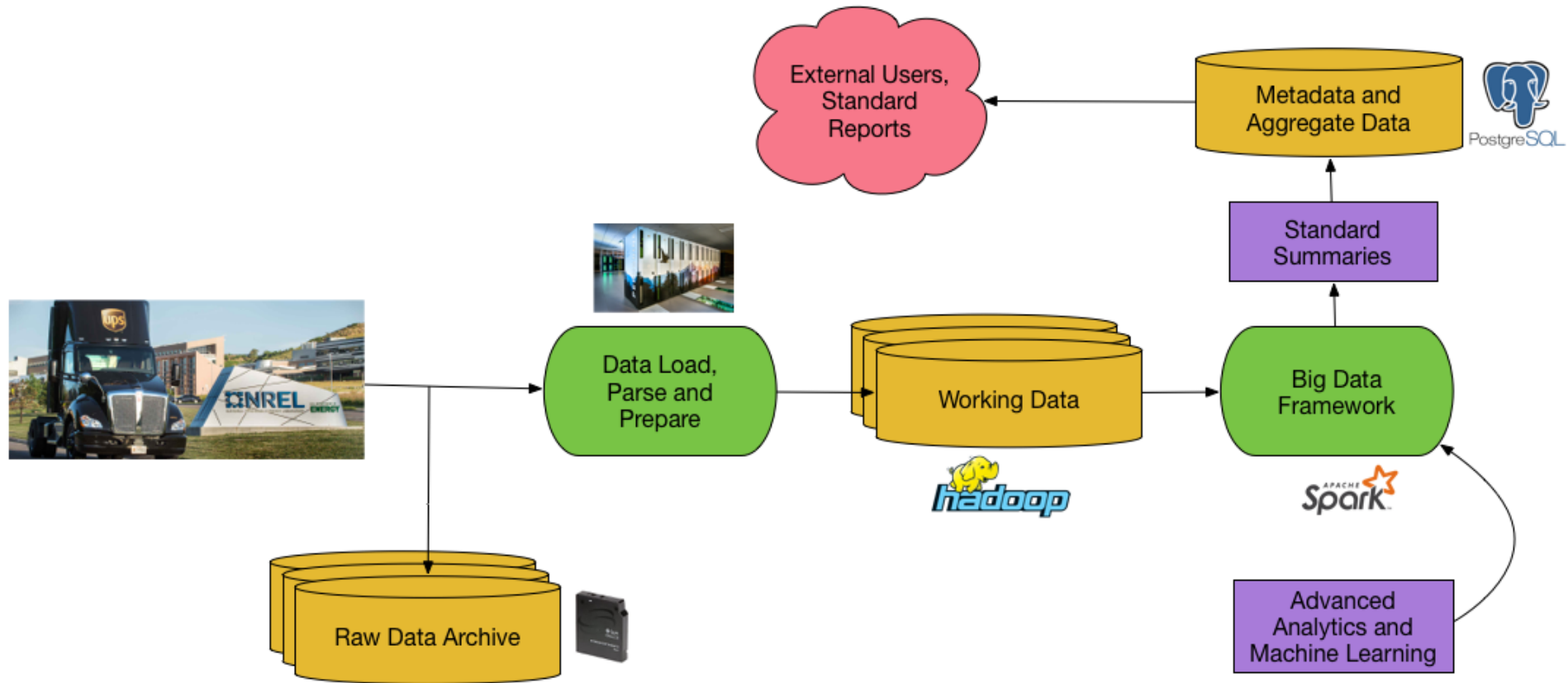


Cloud Compatibility Allows Arbitrary Scalability

- Amazon and Competitors offer Hundreds of Services
- Increasing adoption by large companies
- NREL approach: cloud-replicable infrastructure
- Key services:
 - S3 – Scalable Object-based Storage
 - EC2 – Scalable Compute
 - Lambda – Pay per 'function' execution
 - Marketplace Gateway -- Monetize data access

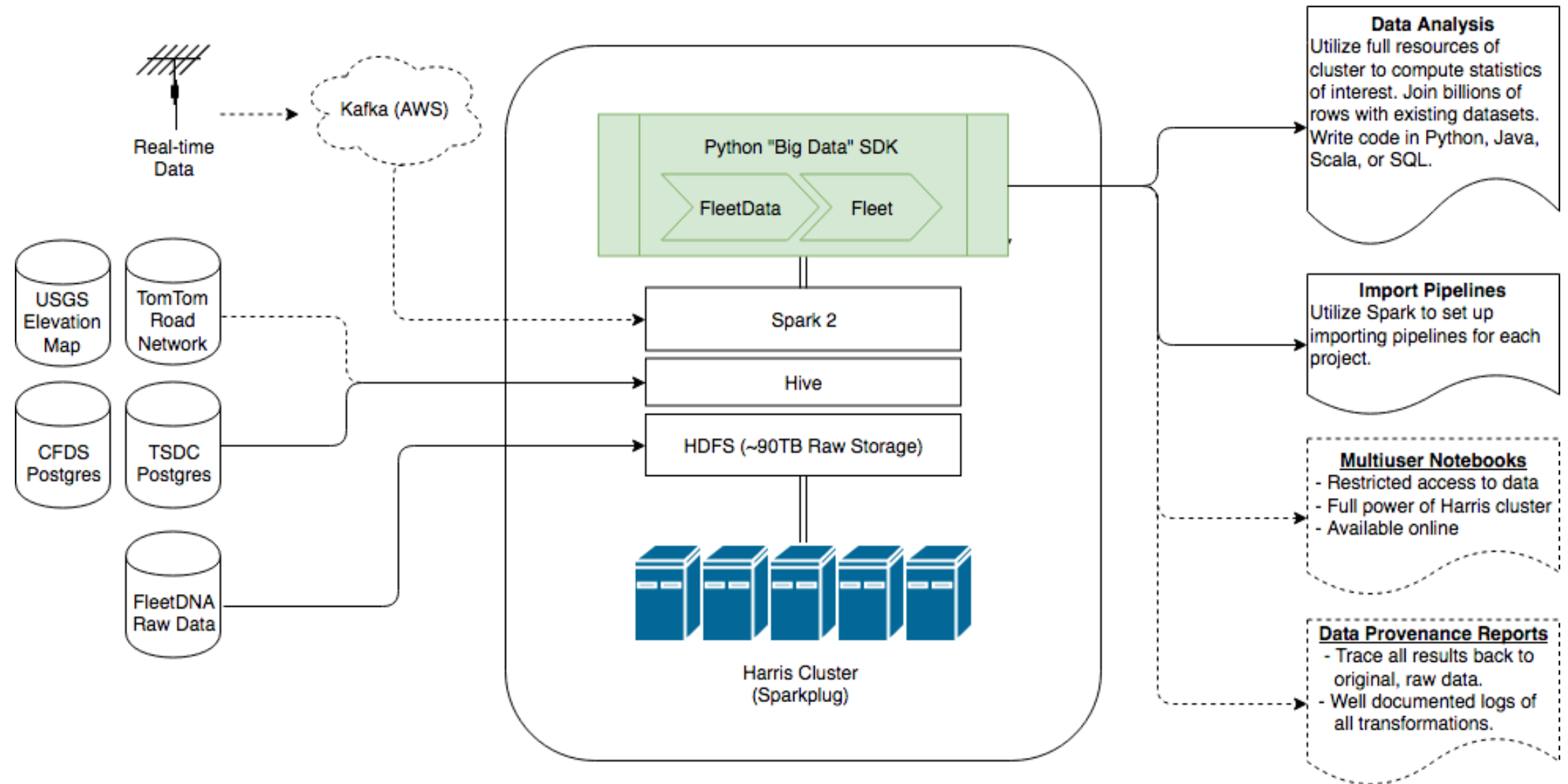


Big IIOT Data – For transportation systems



- Fully compatible with cloud services – industry standard technology
- Can process streaming (high velocity) or offline (high volume) data
- Designed for petabyte-scale (or bigger) datasets]
- Can support traditional HPC or Big-Data use cases
- Promotes collaboration with external users

Current Status: 'Big Data SDK' for Transportation Data



'Big Data SDK' for Transportation Data

Query - master on jperrsau@localhost:5432 *

SQL Editor

```
select count(gpspeed), avg(gpspeed) from combined_gps_points.combined_gps_points;
```

Output pane

	count bigint	avg double precision
1	304052425	47.4307721246939

```
hive> select count(gpspeed), avg(gpspeed) from tsrc_combined_gps_points_v5;
Query ID = jperrsau_20180202101248_34ce3754-cc4c-4297-84d6-99f276fbc5f5
Total jobs = 1
Launching Job 1 out of 1

Status: Running (Executing on YARN cluster with App id application_1516207399240_0128)
```

	VERTICES	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	SUCCEEDED	40	40	0	0	0	0
Reducer 2	SUCCEEDED	1	1	0	0	0	0

=====>] 100% ELAPSED TIME: 10.94 s

```
In [2]: %%time
from lib.projects.paccar import Paccar
df = Paccar().dataframe.cache()
df.printSchema()

root
 |-- lat: string (nullable = true)
 |-- lon: string (nullable = true)
 |-- speed: string (nullable = true)
 |-- ts: long (nullable = true)
 |-- import_projects: string (nullable = false)
 |-- lag_ts: long (nullable = true)
 |-- lag_speed: string (nullable = true)
 |-- lag_vdir: string (nullable = true)
 |-- vdir: string (nullable = true)
 |-- trip: integer (nullable = true)
 |-- start: long (nullable = true)
 |-- end: long (nullable = true)

CPU times: user 157 ms, sys: 14.6 ms, total: 172 ms
Wall time: 26.2 s

Compute the number of rows in Paccar as well as the number of unique vehicles.

In [3]: %%time
from pyspark.sql.functions import lit, sum
rows = df.count()
vehicles = df.groupBy("vdir").agg(sum(lit(1))).count()
print "There are {0} rows and {1} vehicles.".format(rows, vehicles)

There are 5221641 rows and 12 vehicles.
CPU times: user 2.13 ms, sys: 4.8 ms, total: 6.94 ms
Wall time: 17.5 s

Speed Calcs using Fleet class

In [4]: from lib.fleet import Fleet

paccar_fleet = Fleet(sqlContext, df=df, metadata=False)
display( paccar_fleet.trip_speed_calcs().toPandas() )
```

	vdir	avg_driving_speed	max_speed	percent_zero	driving_speed_standard_deviation	ca_standard	as_standard	percent_distance_below_55	stops_per_mile
0	51	87.152345	99.992	50.466242	27.987210	0.223395	141.012115	6.399828	0.070547
1	54	75.054231	99.964	28.405015	28.694433	0.226294	126.475453	14.093510	0.063974
2	34	5.799703	8.0938	97.990223	2.281908	0.303282	10.129556	100.000000	51.726789
3	35	64.944577	99.992	18.003997	24.163895	0.267748	147.802765	3.900609	0.056571

Primary Interface:



Or, if you prefer:

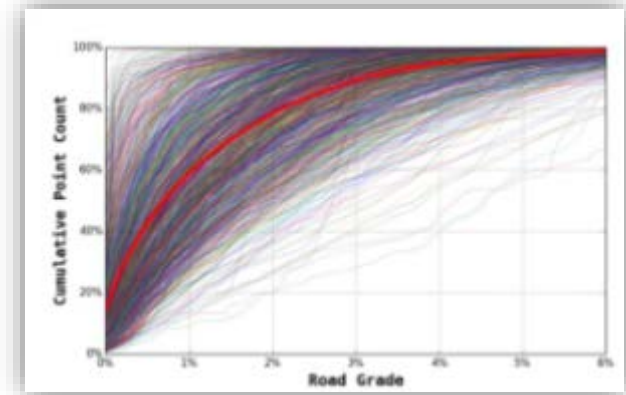
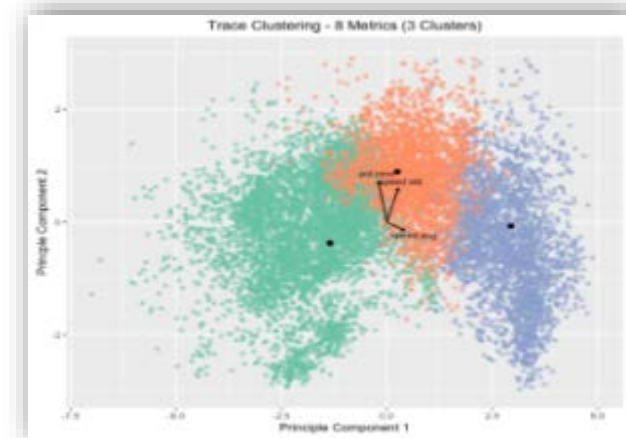


Scientific Approach & Accomplishment

- NREL Fleet DNA data and analytic expertise provided information crucial to EPA's development of Phase II GHG and fuel efficiency standards for medium- and heavy-duty vehicles.
- NREL segmented vocational vehicle drive-cycle characteristics into multi-dimensional operating groups—including urban, mixed urban, and highway driving conditions—to develop a series of transient drive cycles with weighting factors representative of the acceleration rates, speed distributions, and idle times seen in real-world commercial vehicle driving.
- NREL applied map-matching techniques with USGS elevation data and then weighted the profiles using freight activity data.
- Statistically representative highway segments were identified for on-road testing, and road grade profiles were incorporated into EPA certification cycles.

Significance & Impact

- Analysis of Fleet DNA vocational vehicle data helped EPA ensure Phase II GHG regulations are more representative of real-world driving conditions.



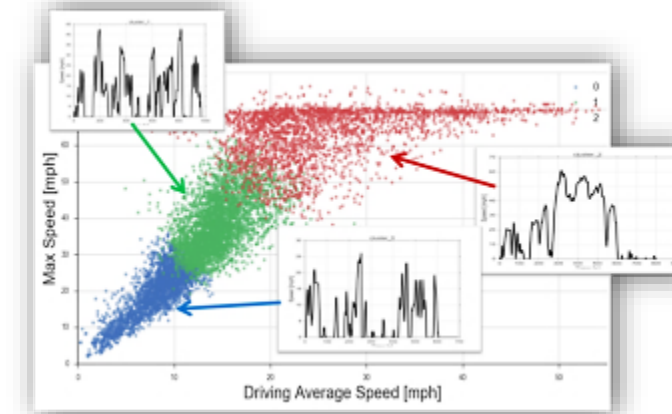
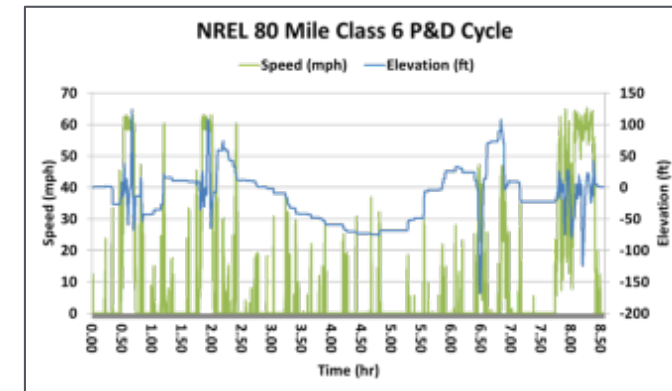
This work tapped into Fleet DNA data, fused with national road network and freight activity data using NREL's Peregrine high-performance computing system.

Scientific Approach & Accomplishment

- Leveraging Fleet DNA data to characterize real-world duty cycles from urban delivery vehicles, NREL applied the k-mediod clustering algorithm to segment in-use driving profiles into operational modes and developed representative drive cycles for various modes using the DRIVE tool.
- NREL developed analytical methods to incorporate other parameters, such as road grade, idle time, and key status into the drive cycles.
- NREL's drive cycles are being used to size drivetrain components and optimize energy storage control strategies to meet performance requirements and validate performance relative to program objectives.

Significance & Impact

- This work was conducted as part of two industry partnerships under DOE FOAs led by **Cummins and Robert Bosch** to develop commercially viable, range-extended EVs for urban delivery applications targeting a 50% efficiency improvement.



NREL-developed representative drive cycles are used by **Cummins and Bosch** in powertrain optimization and performance evaluations.

Big Data Applications – Forecasting – National Scale Platooning Potential

Scientific Approach & Accomplishment

- NREL analyzed fuel-savings data from six independent platooning studies conducted between 2013 and 2016 with Class 8 tractor trailers, including four independent track test studies, wind tunnel results from LLNL, and CFD simulations from Denso.
- NREL followed up track testing efforts with large scale (50k+ vehicles) evaluating real world potential for platooning on US roadways.

Significance & Impact

- NREL evaluation and analysis have characterized platooning performance under a range of speeds, loads, and following distances, including reduced benefits at very close following distances.
- NREL platooning data and analysis are being used in an **ARPAe NEXTCAR project with Purdue, Cummins, and Peloton** to develop next-generation adaptive platooning technologies and in other research efforts at **LLNL, LBNL, and FHWA**.



Platooning reduces aerodynamic drag by decreasing the driving distance between vehicles.

