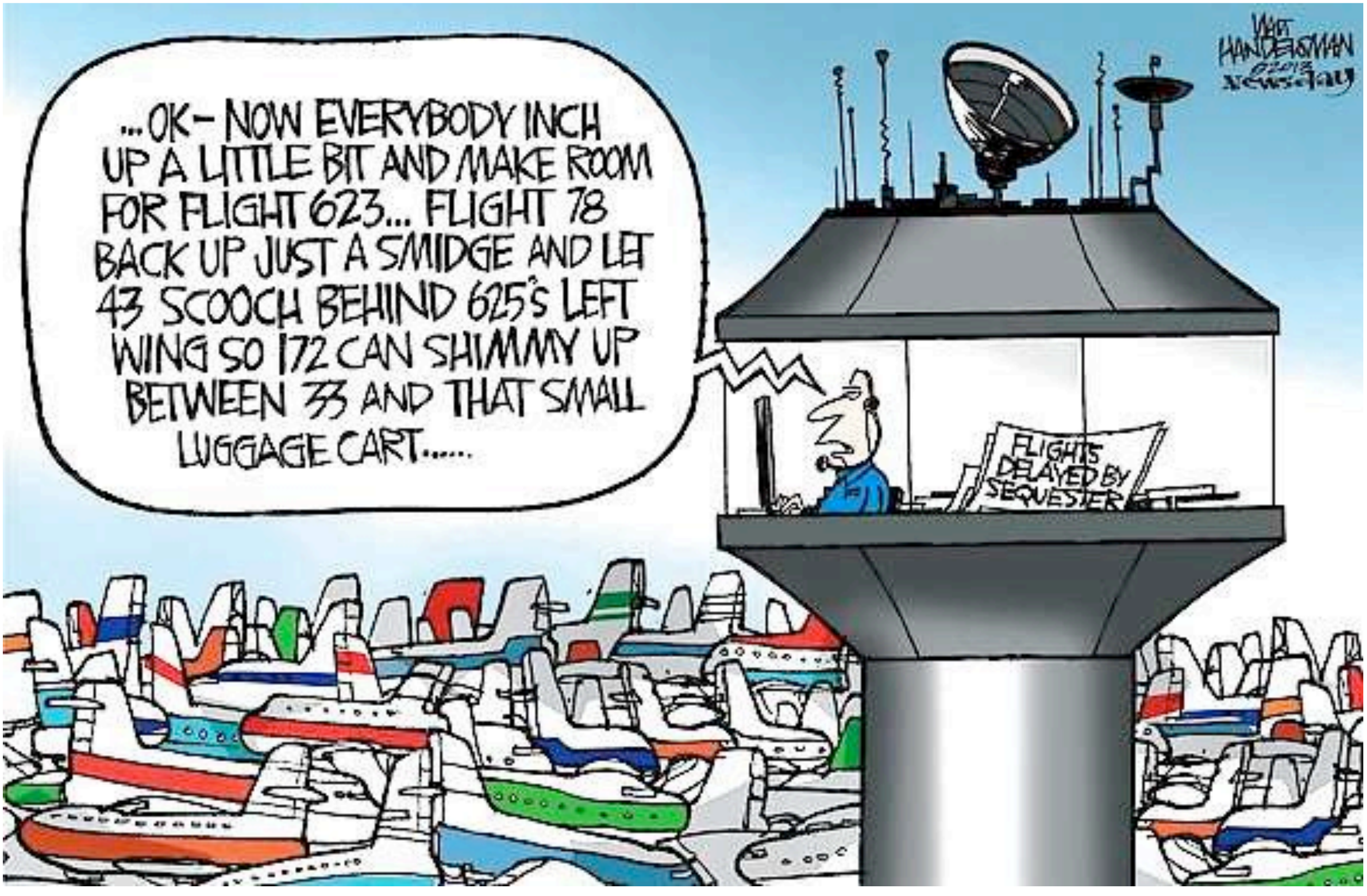
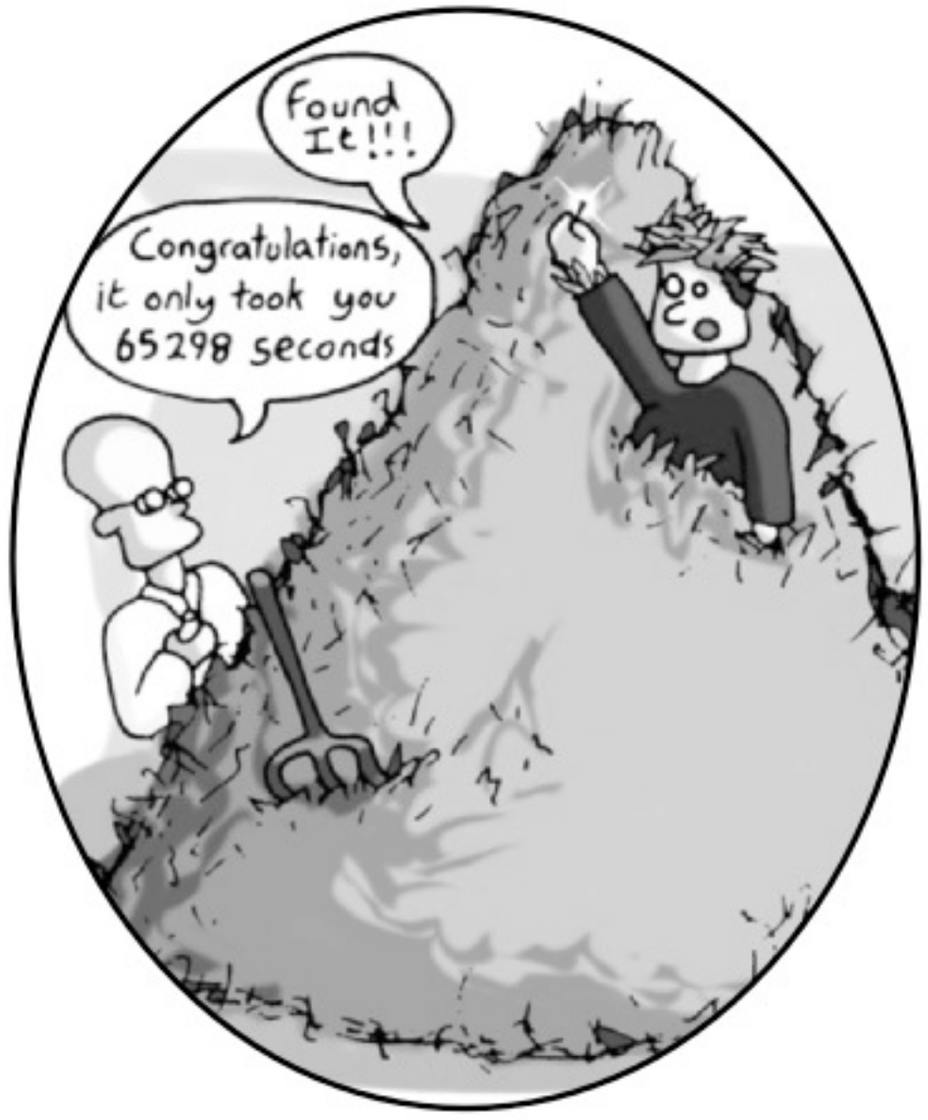
# Air Traffic Management with Big Data Analytics

Alessandro Ferreira Leite

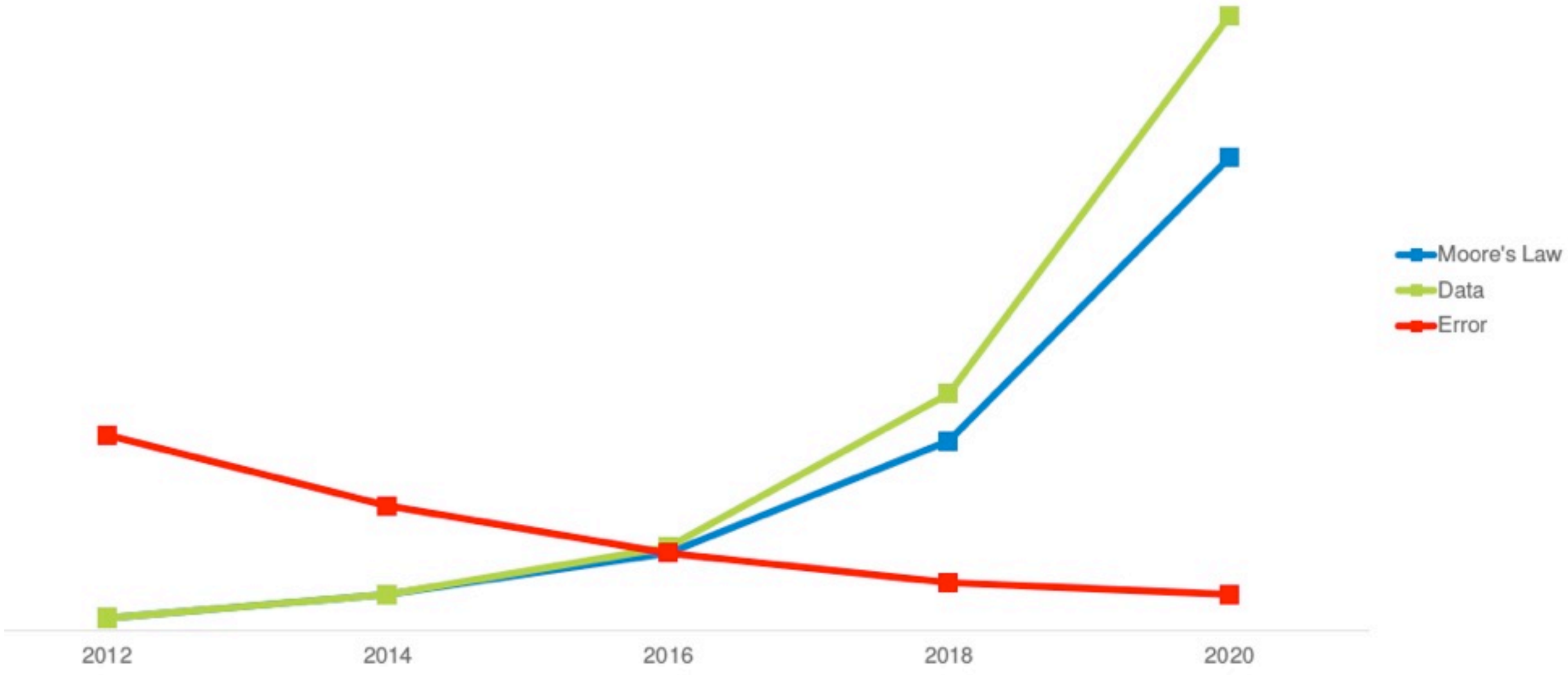

...OK- NOW EVERYBODY INCH UP A LITTLE BIT AND MAKE ROOM FOR FLIGHT 623... FLIGHT 78 BACK UP JUST A SMIDGE AND LET 43 SCOOCH BEHIND 625's LEFT WING SO 172 CAN SHIMMY UP BETWEEN 33 AND THAT SMALL LUGGAGE CART.....
FLIGHTS DELAYED BY SEQUESTER

# Understanding why a traffic is delayed is a difficult task

- Historical information
- Weather
- Availability of airplanes
- Concurrent flights
- ....

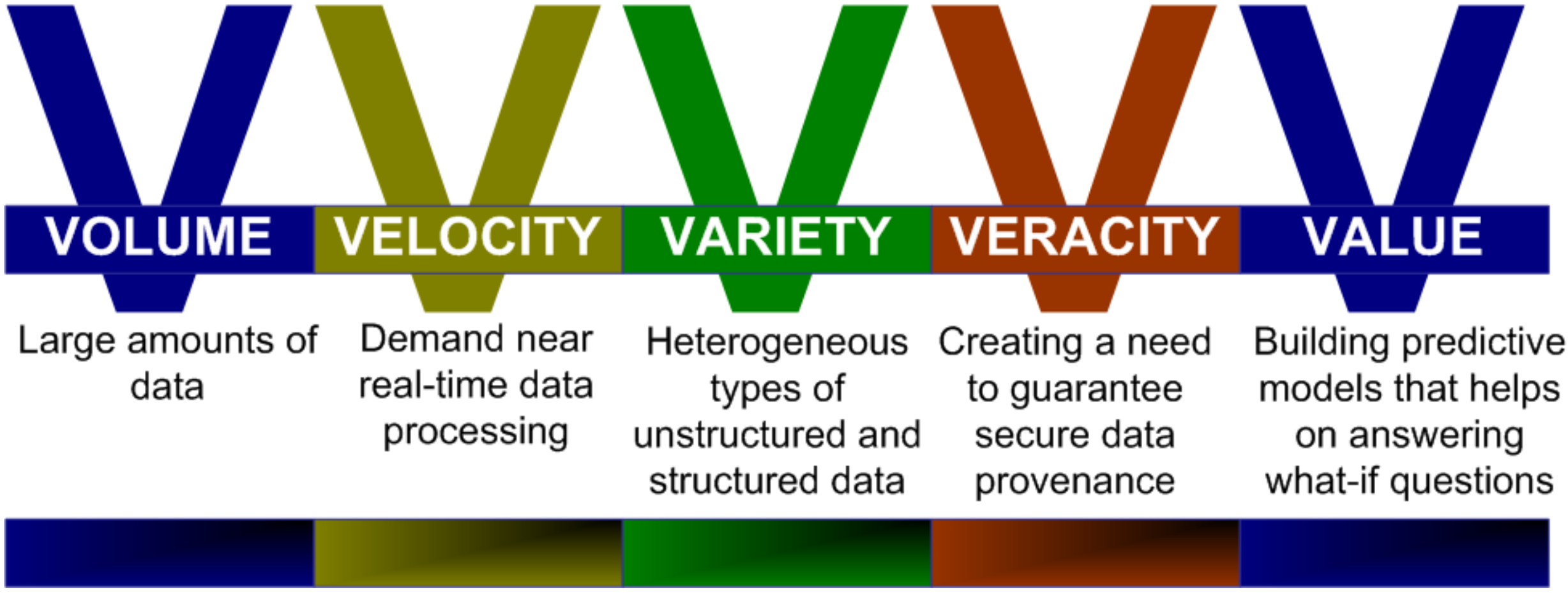


# Data is growing faster than Moore's law




**Source**: https://amplab.cs.berkeley.edu/for-big-data-moores-law-means-better-decisions/

# **Data** has always been **big**



# Big Data Examples

- Facebook's daily logs: 60 TB
- Google web index: 10+ PB
- Cost of 1 TB of disk: ~$35
- Time to read 1 TB from disk: 3 hours (100 MB/s)

Big data V's

volume | velocity | variety | veracity | value

Big data V's

volume | velocity | variety | veracity | value

not enough space to store all data

Big data V's

volume | velocity | variety | veracity | value

not enough space to store all data

not enough idle time to finish proper tuning

Big data V's

volume | velocity | variety | veracity | value

not enough space to store all data

not enough idle time to finish proper tuning

unpredictable workload change

Big data V's

volume | velocity | variety | veracity | value

not enough space to store all data

not enough idle time to finish proper tuning

unpredictable workload change

not enough resources to process all data

Big data V's

volume | velocity | variety | veracity | value

not enough space to store all data

not enough idle time to finish proper tuning

unpredictable workload change

not enough resources to process all data

One possible solution is to distribute data over multiple machines

Big data V's

volume | velocity | variety | veracity | value

not enough space to store all data

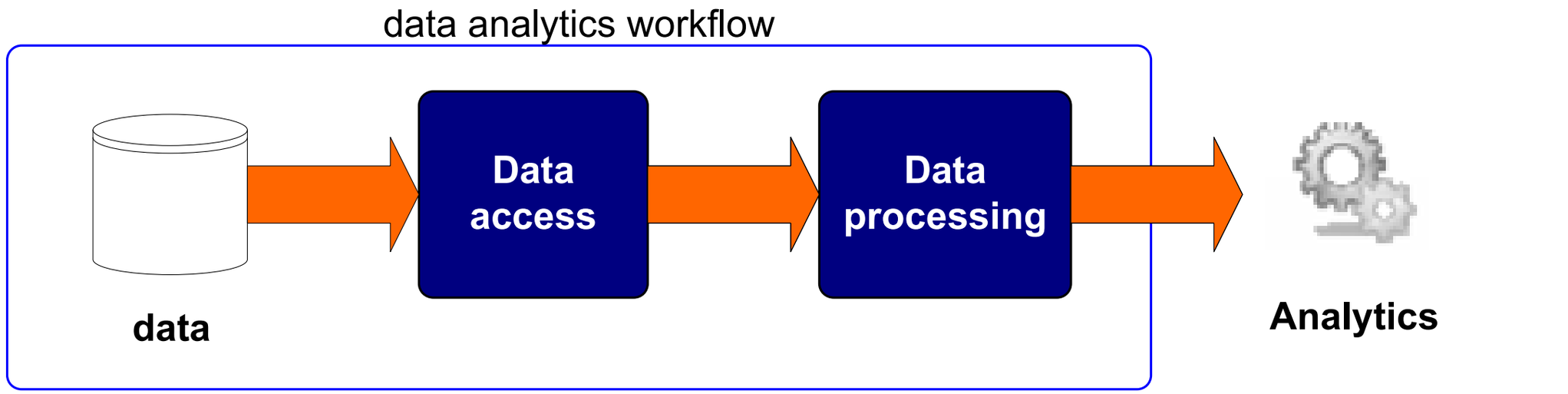
not enough idle time to finish proper tuning

unpredictable workload change

not enough resources to process all data

One possible solution is to distribute data over multiple machines

# How do we split work across machines?




- Descriptive statistics
- Machine Learning
- MapReduce

# How do we find the longest flight for each company?

# How do we find the longest flight for each company?

| | | | | | | |
|---|---|---|---|---|---|---|
| **1503** | **UA** | LAX | -5 | -10 | ... | **2536** |
| **540** | **PS** | BUR | 13 | 5 | | **186** |
| **1920** | **DL** | BOS | 10 | 32 | | **1876** |
| **1840** | **DL** | SFO | 0 | 13 | | **568** |
| **272** | **US** | BWI | 4 | -2 | | **359** |
| **784** | **PS** | SEA | 7 | 3 | | **176** |
| **796** | **PS** | LAX | -2 | 2 | | **237** |
| **1525** | **UA** | SFO | 3 | -5 | | **1867** |
| **632** | **US** | SJC | 2 | -4 | | **245** |
| **1610** | **UA** | MIA | 60 | 34 | | **1365** |
| **2032** | **DL** | EWR | 10 | 16 | | **789** |
| **2134** | **DL** | DFW | 6 | 6 | | **914** |

# How do we find the longest flight for each company?

| Flight ID | | | | | | |
|---|---|---|---|---|---|---|
| **1503** | **UA** | LAX | -5 | -10 | ... | **2536** |
| **540** | **PS** | BUR | 13 | 5 | | **186** |
| **1920** | **DL** | BOS | 10 | 32 | | **1876** |
| **1840** | **DL** | SFO | 0 | 13 | | **568** |
| **272** | **US** | BWI | 4 | -2 | | **359** |
| **784** | **PS** | SEA | 7 | 3 | | **176** |
| **796** | **PS** | LAX | -2 | 2 | | **237** |
| **1525** | **UA** | SFO | 3 | -5 | | **1867** |
| **632** | **US** | SJC | 2 | -4 | | **245** |
| **1610** | **UA** | MIA | 60 | 34 | | **1365** |
| **2032** | **DL** | EWR | 10 | 16 | | **789** |
| **2134** | **DL** | DFW | 6 | 6 | | **914** |

# How do we find the longest flight for each company?

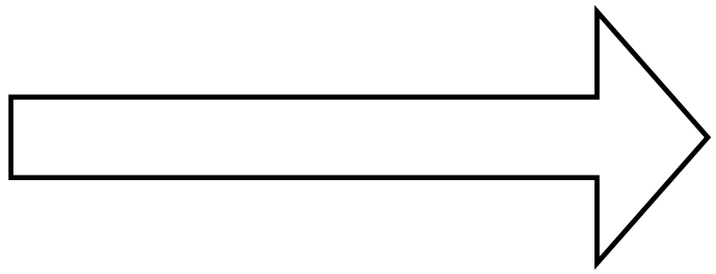
| Flight ID | Airline ID | | | | | |
|---|---|---|---|---|---|---|
| **1503** | **UA** | LAX | -5 | -10 | ... | **2536** |
| **540** | **PS** | BUR | 13 | 5 | | **186** |
| **1920** | **DL** | BOS | 10 | 32 | | **1876** |
| **1840** | **DL** | SFO | 0 | 13 | | **568** |
| **272** | **US** | BWI | 4 | -2 | | **359** |
| **784** | **PS** | SEA | 7 | 3 | | **176** |
| **796** | **PS** | LAX | -2 | 2 | | **237** |
| **1525** | **UA** | SFO | 3 | -5 | | **1867** |
| **632** | **US** | SJC | 2 | -4 | | **245** |
| **1610** | **UA** | MIA | 60 | 34 | | **1365** |
| **2032** | **DL** | EWR | 10 | 16 | | **789** |
| **2134** | **DL** | DFW | 6 | 6 | | **914** |

# How do we find the longest flight for each company?

| Flight ID | Airline ID | | | | | Distance |
|---|---|---|---|---|---|---|
| **1503** | **UA** | LAX | -5 | -10 | ... | **2536** |
| **540** | **PS** | BUR | 13 | 5 | | **186** |
| **1920** | **DL** | BOS | 10 | 32 | | **1876** |
| **1840** | **DL** | SFO | 0 | 13 | | **568** |
| **272** | **US** | BWI | 4 | -2 | | **359** |
| **784** | **PS** | SEA | 7 | 3 | | **176** |
| **796** | **PS** | LAX | -2 | 2 | | **237** |
| **1525** | **UA** | SFO | 3 | -5 | | **1867** |
| **632** | **US** | SJC | 2 | -4 | | **245** |
| **1610** | **UA** | MIA | 60 | 34 | | **1365** |
| **2032** | **DL** | EWR | 10 | 16 | | **789** |
| **2134** | **DL** | DFW | 6 | 6 | | **914** |

# How do we find the longest flight for each company?

| Flight ID | Airline ID | | | | | Distance |
|---|---|---|---|---|---|---|
| **1503** | **UA** | LAX | -5 | -10 | ... | **2536** |
| **540** | **PS** | BUR | 13 | 5 | | **186** |
| **1920** | **DL** | BOS | 10 | 32 | | **1876** |
| **1840** | **DL** | SFO | 0 | 13 | | **568** |
| **272** | **US** | BWI | 4 | -2 | | **359** |
| **784** | **PS** | SEA | 7 | 3 | | **176** |
| **796** | **PS** | LAX | -2 | 2 | | **237** |
| **1525** | **UA** | SFO | 3 | -5 | | **1867** |
| **632** | **US** | SJC | 2 | -4 | | **245** |
| **1610** | **UA** | MIA | 60 | 34 | | **1365** |
| **2032** | **DL** | EWR | 10 | 16 | | **789** |
| **2134** | **DL** | DFW | 6 | 6 | | **914** |



# How do we find the longest flight for each company?

| Flight ID | Airline ID | | | | | Distance |
|---|---|---|---|---|---|---|
| **1503** | **UA** | LAX | -5 | -10 | ... | **2536** |
| **540** | **PS** | BUR | 13 | 5 | | **186** |
| **1920** | **DL** | BOS | 10 | 32 | | **1876** |
| **1840** | **DL** | SFO | 0 | 13 | | **568** |
| **272** | **US** | BWI | 4 | -2 | | **359** |
| **784** | **PS** | SEA | 7 | 3 | | **176** |
| **796** | **PS** | LAX | -2 | 2 | | **237** |
| **1525** | **UA** | SFO | 3 | -5 | | **1867** |
| **632** | **US** | SJC | 2 | -4 | | **245** |
| **1610** | **UA** | MIA | 60 | 34 | | **1365** |
| **2032** | **DL** | EWR | 10 | 16 | | **789** |
| **2134** | **DL** | DFW | 6 | 6 | | **914** |

{
UA: 2356,
PS: 237,
...
}

# And, what if the datasets are really big?

# And, what if the datasets are really big?

| | | | | | | |
|---|---|---|---|---|---|---|
| **1503** | **UA** | LAX | -5 | -10 | ... | **2536** |
| **540** | **PS** | BUR | 13 | 5 | | **186** |
| **1920** | **DL** | BOS | 10 | 32 | | **1876** |
| **1840** | **DL** | SFO | 0 | 13 | | **568** |
| **272** | **US** | BWI | 4 | -2 | | **359** |
| **784** | **PS** | SEA | 7 | 3 | | **176** |
| **796** | **PS** | LAX | -2 | 2 | | **237** |
| **1525** | **UA** | SFO | 3 | -5 | | **1867** |
| **632** | **US** | SJC | 2 | -4 | | **245** |
| **1610** | **UA** | MIA | 60 | 34 | | **1365** |
| **2032** | **DL** | EWR | 10 | 16 | | **789** |
| **2134** | **DL** | DFW | 6 | 6 | | **914** |

# And, what if the datasets are really big?

| Flight ID | | | | | | |
|---|---|---|---|---|---|---|
| **1503** | **UA** | LAX | -5 | -10 | ... | **2536** |
| **540** | **PS** | BUR | 13 | 5 | | **186** |
| **1920** | **DL** | BOS | 10 | 32 | | **1876** |
| **1840** | **DL** | SFO | 0 | 13 | | **568** |
| **272** | **US** | BWI | 4 | -2 | | **359** |
| **784** | **PS** | SEA | 7 | 3 | | **176** |
| **796** | **PS** | LAX | -2 | 2 | | **237** |
| **1525** | **UA** | SFO | 3 | -5 | | **1867** |
| **632** | **US** | SJC | 2 | -4 | | **245** |
| **1610** | **UA** | MIA | 60 | 34 | | **1365** |
| **2032** | **DL** | EWR | 10 | 16 | | **789** |
| **2134** | **DL** | DFW | 6 | 6 | | **914** |

# And, what if the datasets are really big?

| Flight ID | Airline ID | | | | | |
|---|---|---|---|---|---|---|
| **1503** | **UA** | LAX | -5 | -10 | ... | **2536** |
| **540** | **PS** | BUR | 13 | 5 | | **186** |
| **1920** | **DL** | BOS | 10 | 32 | | **1876** |
| **1840** | **DL** | SFO | 0 | 13 | | **568** |
| **272** | **US** | BWI | 4 | -2 | | **359** |
| **784** | **PS** | SEA | 7 | 3 | | **176** |
| **796** | **PS** | LAX | -2 | 2 | | **237** |
| **1525** | **UA** | SFO | 3 | -5 | | **1867** |
| **632** | **US** | SJC | 2 | -4 | | **245** |
| **1610** | **UA** | MIA | 60 | 34 | | **1365** |
| **2032** | **DL** | EWR | 10 | 16 | | **789** |
| **2134** | **DL** | DFW | 6 | 6 | | **914** |

# And, what if the datasets are really big?

| Flight ID | Airline ID | | | | | Distance |
|---|---|---|---|---|---|---|
| **1503** | **UA** | LAX | -5 | -10 | ... | **2536** |
| **540** | **PS** | BUR | 13 | 5 | | **186** |
| **1920** | **DL** | BOS | 10 | 32 | | **1876** |
| **1840** | **DL** | SFO | 0 | 13 | | **568** |
| **272** | **US** | BWI | 4 | -2 | | **359** |
| **784** | **PS** | SEA | 7 | 3 | | **176** |
| **796** | **PS** | LAX | -2 | 2 | | **237** |
| **1525** | **UA** | SFO | 3 | -5 | | **1867** |
| **632** | **US** | SJC | 2 | -4 | | **245** |
| **1610** | **UA** | MIA | 60 | 34 | | **1365** |
| **2032** | **DL** | EWR | 10 | 16 | | **789** |
| **2134** | **DL** | DFW | 6 | 6 | | **914** |

# And, what if the datasets are really big?

| Flight ID | Airline ID | | | | | Distance |
|---|---|---|---|---|---|---|
| **1503** | **UA** | LAX | -5 | -10 | ... | **2536** |
| **540** | **PS** | BUR | 13 | 5 | | **186** |
| **1920** | **DL** | BOS | 10 | 32 | | **1876** |
| **1840** | **DL** | SFO | 0 | 13 | | **568** |
| **272** | **US** | BWI | 4 | -2 | | **359** |
| **784** | **PS** | SEA | 7 | 3 | | **176** |
| **796** | **PS** | LAX | -2 | 2 | | **237** |
| **1525** | **UA** | SFO | 3 | -5 | | **1867** |
| **632** | **US** | SJC | 2 | -4 | | **245** |
| **1610** | **UA** | MIA | 60 | 34 | | **1365** |
| **2032** | **DL** | EWR | 10 | 16 | | **789** |
| **2134** | **DL** | DFW | 6 | 6 | | **914** |

# And, what if the datasets are really big?

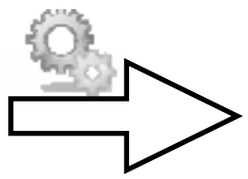
| Flight ID | Airline ID | | | | | Distance |
|---|---|---|---|---|---|---|
| **1503** | **UA** | LAX | -5 | -10 | ... | **2536** |
| **540** | **PS** | BUR | 13 | 5 | | **186** |
| **1920** | **DL** | BOS | 10 | 32 | | **1876** |
| **1840** | **DL** | SFO | 0 | 13 | | **568** |
| **272** | **US** | BWI | 4 | -2 | | **359** |
| **784** | **PS** | SEA | 7 | 3 | | **176** |
| **796** | **PS** | LAX | -2 | 2 | | **237** |
| **1525** | **UA** | SFO | 3 | -5 | | **1867** |
| **632** | **US** | SJC | 2 | -4 | | **245** |
| **1610** | **UA** | MIA | 60 | 34 | | **1365** |
| **2032** | **DL** | EWR | 10 | 16 | | **789** |
| **2134** | **DL** | DFW | 6 | 6 | | **914** |

# And, what if the datasets are really big?

| Flight ID | Airline ID | | | | | Distance |
|---|---|---|---|---|---|---|
| **1503** | **UA** | LAX | -5 | -10 | ... | **2536** |
| **540** | **PS** | BUR | 13 | 5 | | **186** |
| **1920** | **DL** | BOS | 10 | 32 | | **1876** |
| **1840** | **DL** | SFO | 0 | 13 | | **568** |
| **272** | **US** | BWI | 4 | -2 | | **359** |
| **784** | **PS** | SEA | 7 | 3 | | **176** |
| **796** | **PS** | LAX | -2 | 2 | | **237** |
| **1525** | **UA** | SFO | 3 | -5 | | **1867** |
| **632** | **US** | SJC | 2 | -4 | | **245** |
| **1610** | **UA** | MIA | 60 | 34 | | **1365** |
| **2032** | **DL** | EWR | 10 | 16 | | **789** |
| **2134** | **DL** | DFW | 6 | 6 | | **914** |

# And, what if the datasets are really big?

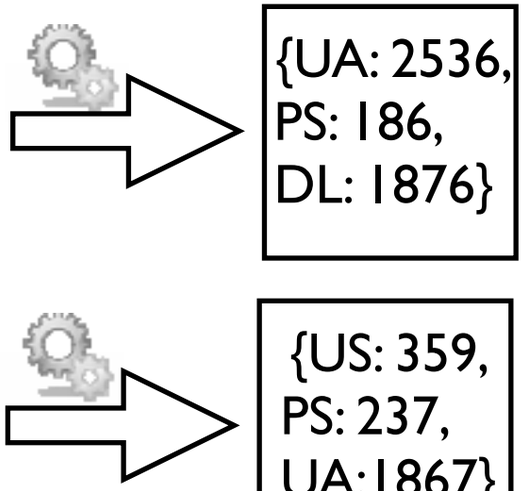
| Flight ID | Airline ID | | | | | Distance |
|---|---|---|---|---|---|---|
| **1503** | **UA** | LAX | -5 | -10 | ... | **2536** |
| **540** | **PS** | BUR | 13 | 5 | | **186** |
| **1920** | **DL** | BOS | 10 | 32 | | **1876** |
| **1840** | **DL** | SFO | 0 | 13 | | **568** |
| **272** | **US** | BWI | 4 | -2 | | **359** |
| **784** | **PS** | SEA | 7 | 3 | | **176** |
| **796** | **PS** | LAX | -2 | 2 | | **237** |
| **1525** | **UA** | SFO | 3 | -5 | | **1867** |
| **632** | **US** | SJC | 2 | -4 | | **245** |
| **1610** | **UA** | MIA | 60 | 34 | | **1365** |
| **2032** | **DL** | EWR | 10 | 16 | | **789** |
| **2134** | **DL** | DFW | 6 | 6 | | **914** |



{UA: 2536,
PS: 186,
DL: 1876}

# And, what if the datasets are really big?

| Flight ID | Airline ID | | | | | Distance |
|---|---|---|---|---|---|---|
| **1503** | **UA** | LAX | -5 | -10 | ... | **2536** |
| **540** | **PS** | BUR | 13 | 5 | | **186** |
| **1920** | **DL** | BOS | 10 | 32 | | **1876** |
| **1840** | **DL** | SFO | 0 | 13 | | **568** |
| **272** | **US** | BWI | 4 | -2 | | **359** |
| **784** | **PS** | SEA | 7 | 3 | | **176** |
| **796** | **PS** | LAX | -2 | 2 | | **237** |
| **1525** | **UA** | SFO | 3 | -5 | | **1867** |
| **632** | **US** | SJC | 2 | -4 | | **245** |
| **1610** | **UA** | MIA | 60 | 34 | | **1365** |
| **2032** | **DL** | EWR | 10 | 16 | | **789** |
| **2134** | **DL** | DFW | 6 | 6 | | **914** |

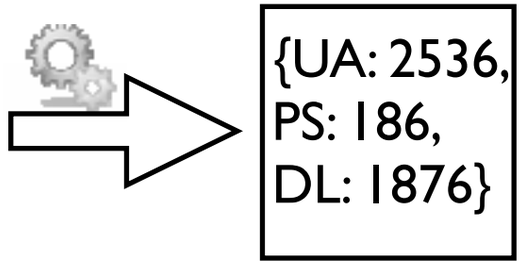


{UA: 2536,
PS: 186,
DL: 1876}

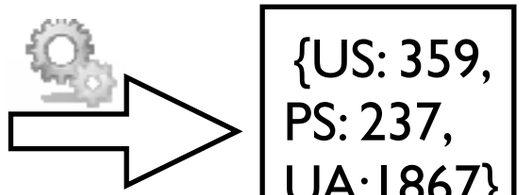
{US: 359,
PS: 237,
UA:1867}

# And, what if the datasets are really big?

| Flight ID | Airline ID | | | | | Distance |
|---|---|---|---|---|---|---|
| 1503 | UA | LAX | -5 | -10 | ... | 2536 |
| 540 | PS | BUR | 13 | 5 | | 186 |
| 1920 | DL | BOS | 10 | 32 | | 1876 |
| 1840 | DL | SFO | 0 | 13 | | 568 |
| 272 | US | BWI | 4 | -2 | | 359 |
| 784 | PS | SEA | 7 | 3 | | 176 |
| 796 | PS | LAX | -2 | 2 | | 237 |
| 1525 | UA | SFO | 3 | -5 | | 1867 |
| 632 | US | SJC | 2 | -4 | | 245 |
| 1610 | UA | MIA | 60 | 34 | | 1365 |
| 2032 | DL | EWR | 10 | 16 | | 789 |
| 2134 | DL | DFW | 6 | 6 | | 914 |



{UA: 2536,
PS: 186,
DL: 1876}



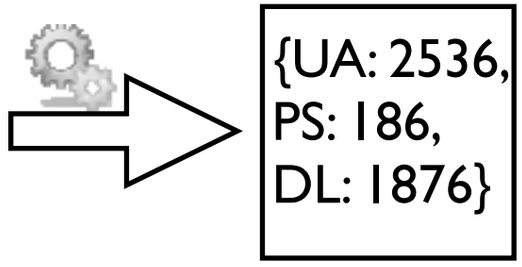
{US: 359,
PS: 237,
UA:1867}

{US: 245,
UA: 1365,
DL: 914}

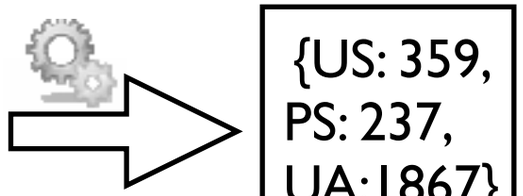
# And, what if the datasets are really big?

| Flight ID | Airline ID | | | | | Distance |
|---|---|---|---|---|---|---|
| **1503** | **UA** | LAX | -5 | -10 | ... | **2536** |
| **540** | **PS** | BUR | 13 | 5 | | **186** |
| **1920** | **DL** | BOS | 10 | 32 | | **1876** |
| **1840** | **DL** | SFO | 0 | 13 | | **568** |
| **272** | **US** | BWI | 4 | -2 | | **359** |
| **784** | **PS** | SEA | 7 | 3 | | **176** |
| **796** | **PS** | LAX | -2 | 2 | | **237** |
| **1525** | **UA** | SFO | 3 | -5 | | **1867** |
| **632** | **US** | SJC | 2 | -4 | | **245** |
| **1610** | **UA** | MIA | 60 | 34 | | **1365** |
| **2032** | **DL** | EWR | 10 | 16 | | **789** |
| **2134** | **DL** | DFW | 6 | 6 | | **914** |

Machines 1 to 3



{UA: 2536,
PS: 186,
DL: 1876}

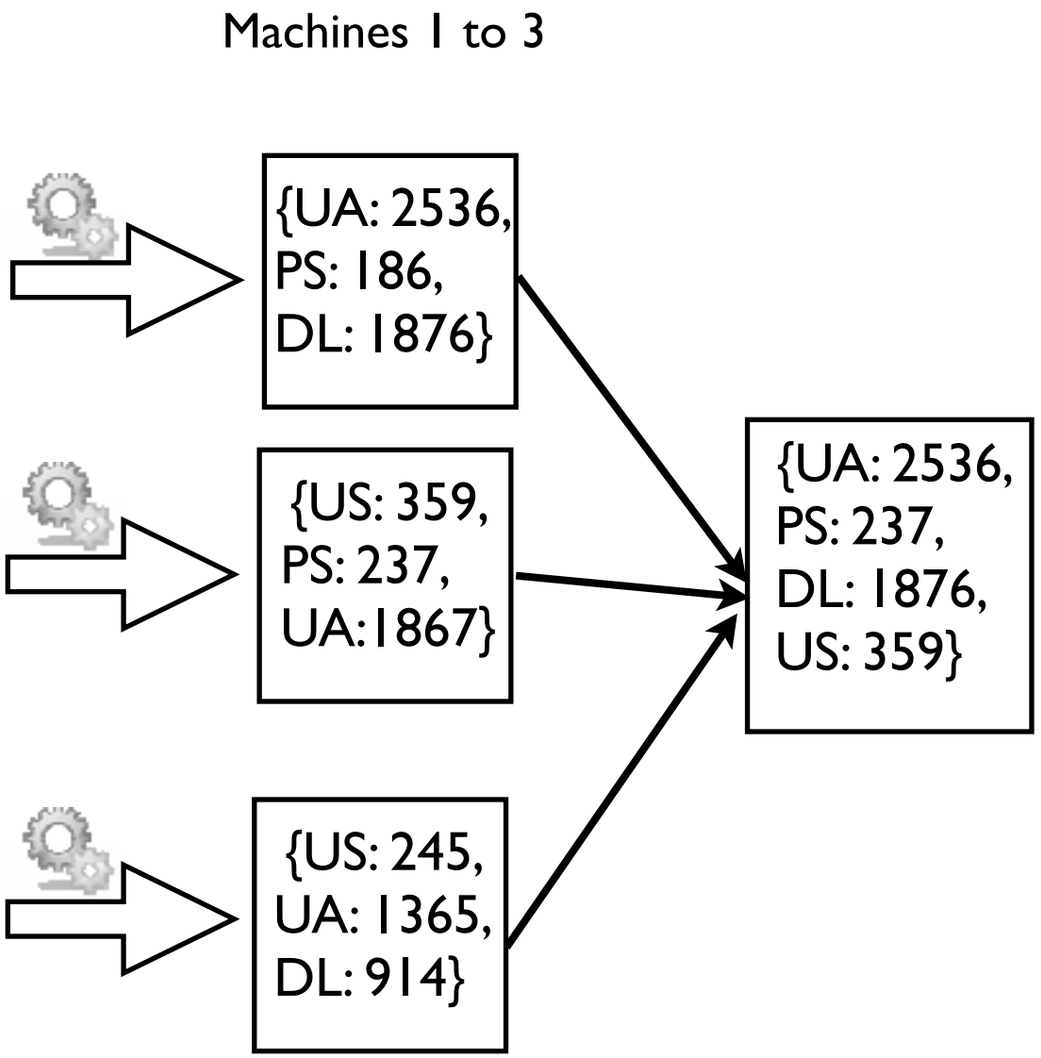


{US: 359,
PS: 237,
UA:1867}

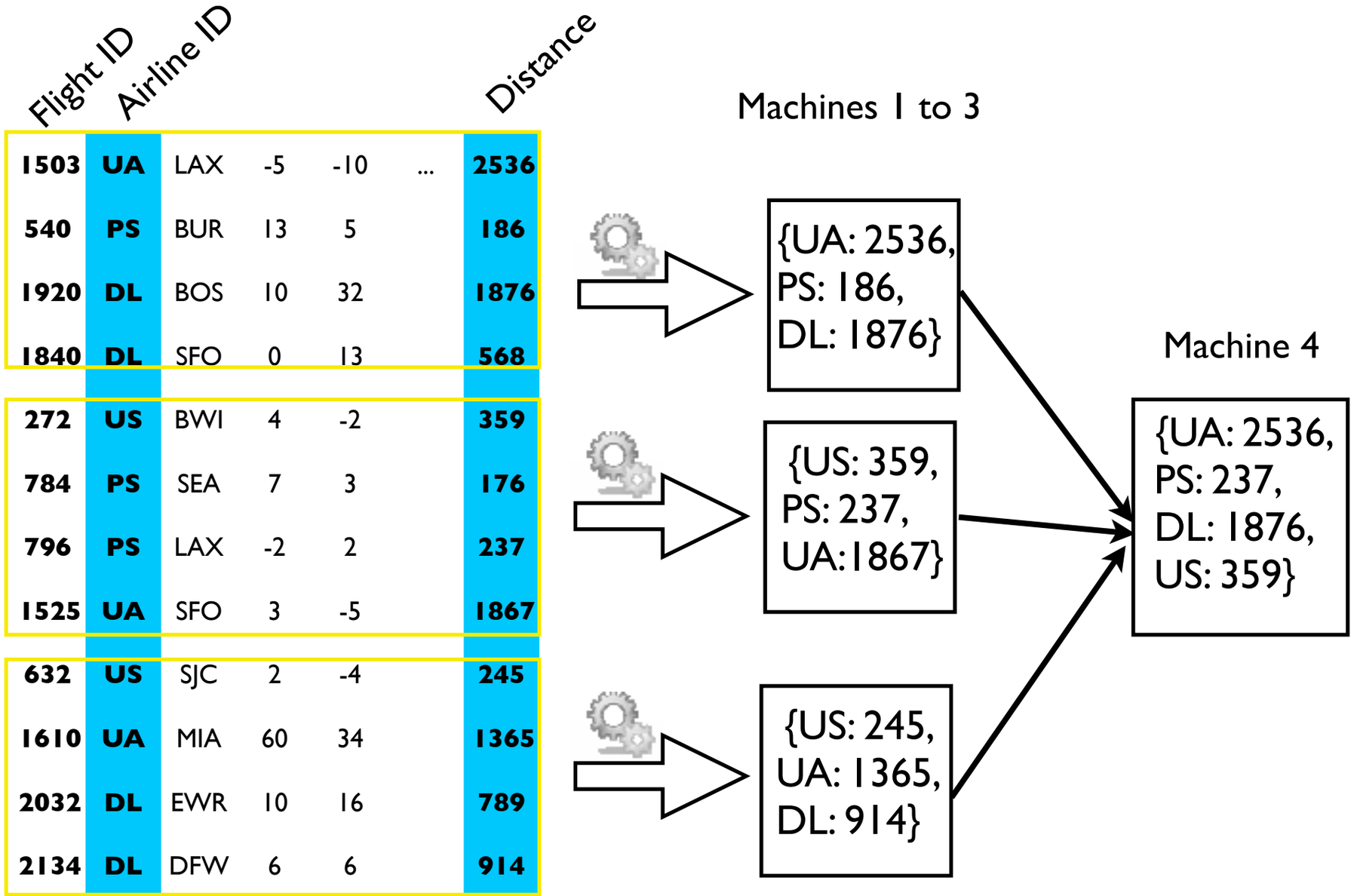
{US: 245,
UA: 1365,
DL: 914}

# And, what if the datasets are really big?

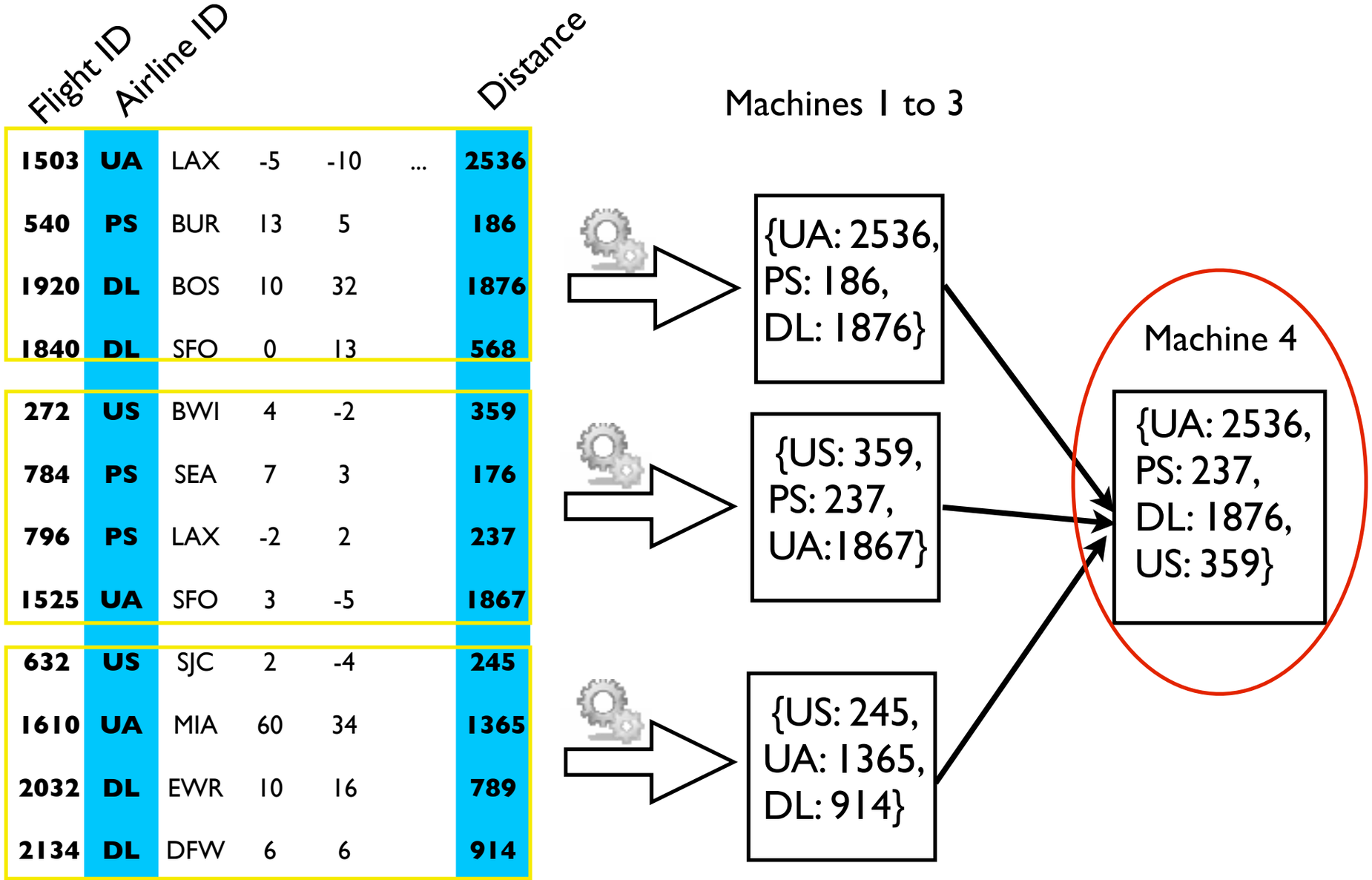
| Flight ID | Airline ID | | | | | Distance |
|---|---|---|---|---|---|---|
| 1503 | UA | LAX | -5 | -10 | ... | 2536 |
| 540 | PS | BUR | 13 | 5 | | 186 |
| 1920 | DL | BOS | 10 | 32 | | 1876 |
| 1840 | DL | SFO | 0 | 13 | | 568 |
| 272 | US | BWI | 4 | -2 | | 359 |
| 784 | PS | SEA | 7 | 3 | | 176 |
| 796 | PS | LAX | -2 | 2 | | 237 |
| 1525 | UA | SFO | 3 | -5 | | 1867 |
| 632 | US | SJC | 2 | -4 | | 245 |
| 1610 | UA | MIA | 60 | 34 | | 1365 |
| 2032 | DL | EWR | 10 | 16 | | 789 |
| 2134 | DL | DFW | 6 | 6 | | 914 |



# And, what if the datasets are really big?

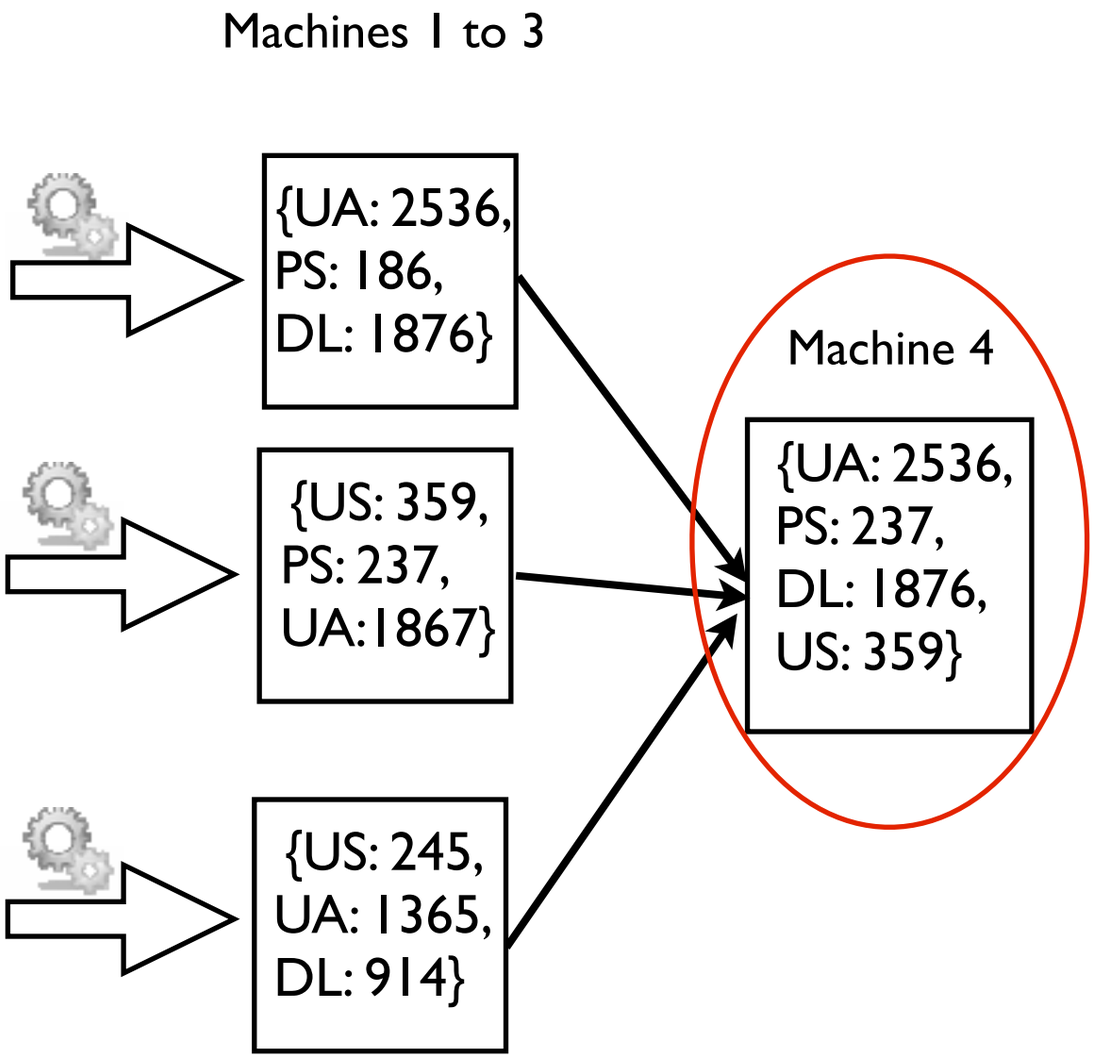


# And, what if the datasets are really big?



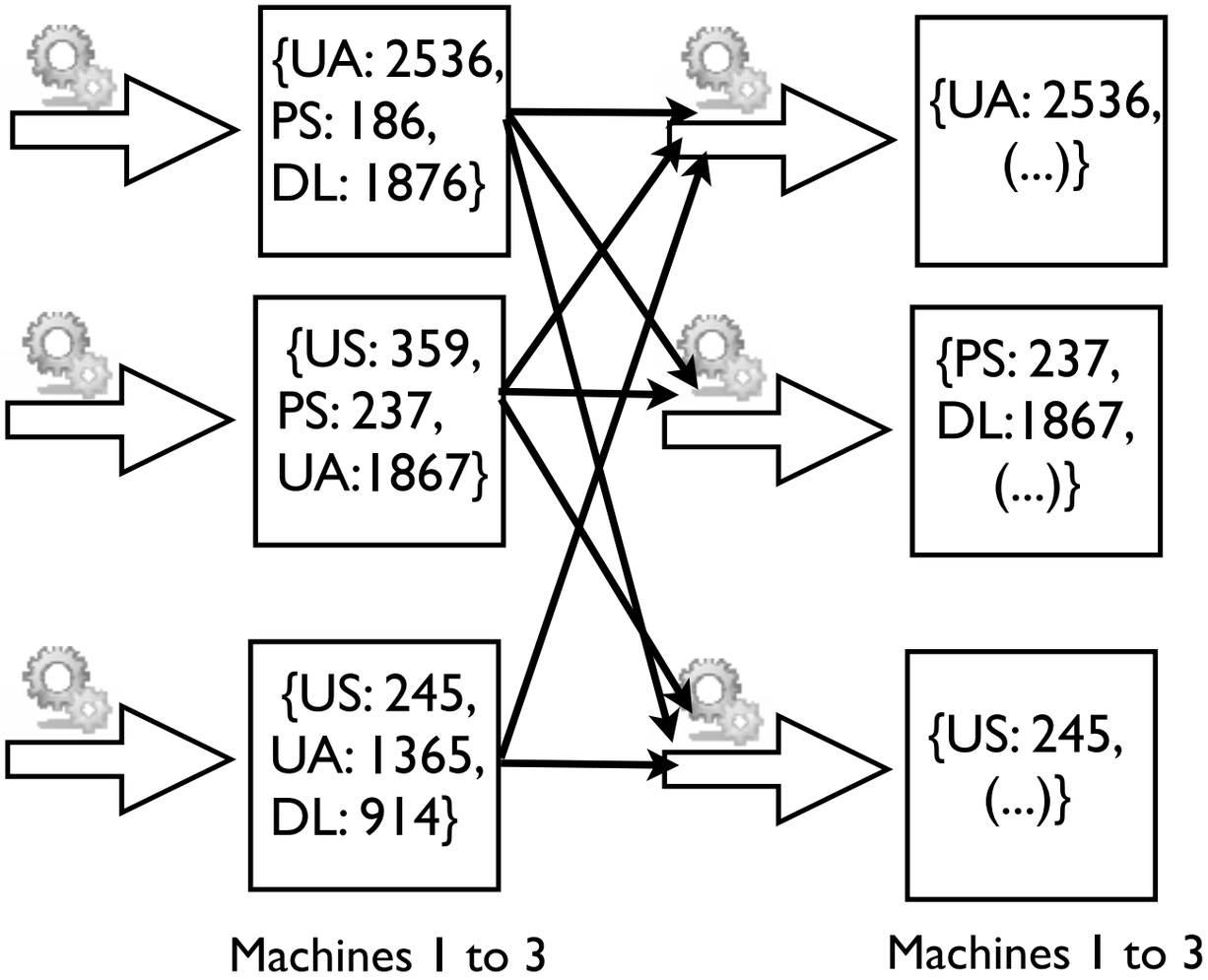
# And, what if the datasets are really big?

| Flight ID | Airline ID | | | | | Distance |
|---|---|---|---|---|---|---|
| 1503 | UA | LAX | -5 | -10 | ... | 2536 |
| 540 | PS | BUR | 13 | 5 | | 186 |
| 1920 | DL | BOS | 10 | 32 | | 1876 |
| 1840 | DL | SFO | 0 | 13 | | 568 |
| 272 | US | BWI | 4 | -2 | | 359 |
| 784 | PS | SEA | 7 | 3 | | 176 |
| 796 | PS | LAX | -2 | 2 | | 237 |
| 1525 | UA | SFO | 3 | -5 | | 1867 |
| 632 | US | SJC | 2 | -4 | | 245 |
| 1610 | UA | MIA | 60 | 34 | | 1365 |
| 2032 | DL | EWR | 10 | 16 | | 789 |
| 2134 | DL | DFW | 6 | 6 | | 914 |



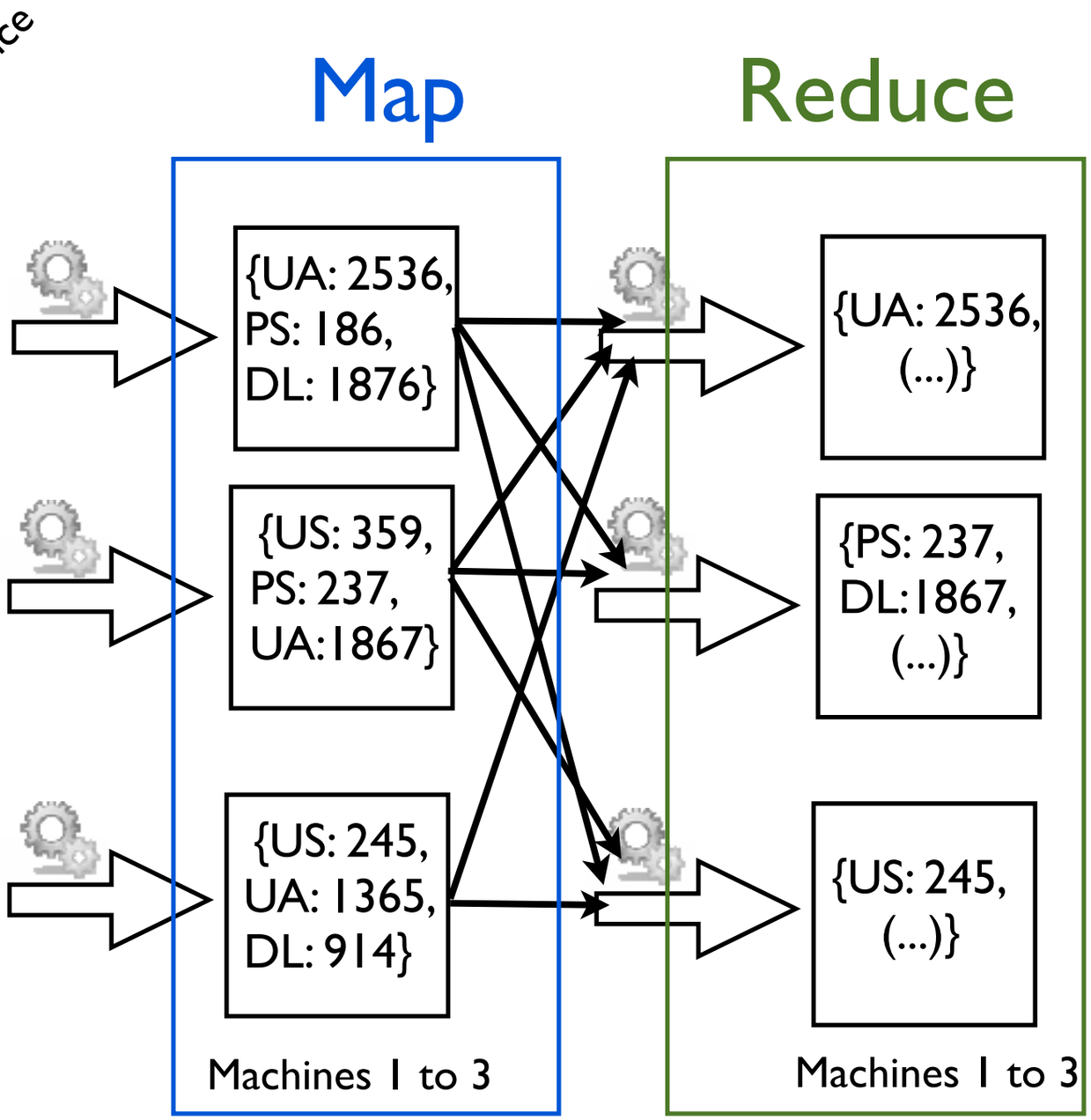
# We can employ the divide and conquer strategy to deal with memory limitation

| Flight ID | Airline ID | | | | | Distance |
|---|---|---|---|---|---|---|
| **1503** | **UA** | LAX | -5 | -10 | ... | **2536** |
| **540** | **PS** | BUR | 13 | 5 | | **186** |
| **1920** | **DL** | BOS | 10 | 32 | | **1876** |
| **1840** | **DL** | SFO | 0 | 13 | | **568** |
| **272** | **US** | BWI | 4 | -2 | | **359** |
| **784** | **PS** | SEA | 7 | 3 | | **176** |
| **796** | **PS** | LAX | -2 | 2 | | **237** |
| **1525** | **UA** | SFO | 3 | -5 | | **1867** |
| **632** | **US** | SJC | 2 | -4 | | **245** |
| **1610** | **UA** | MIA | 60 | 34 | | **1365** |
| **2032** | **DL** | EWR | 10 | 16 | | **789** |
| **2134** | **DL** | DFW | 6 | 6 | | **914** |

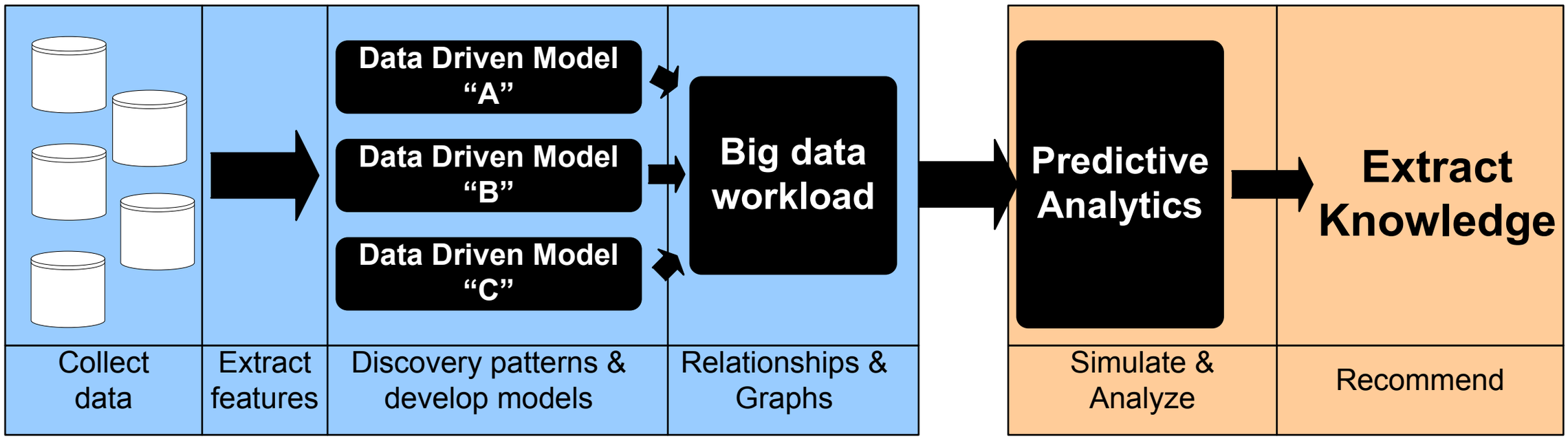


# We can employ the divide and conquer strategy to deal with memory limitation

| Flight ID | Airline ID | | | | | Distance |
|---|---|---|---|---|---|---|
| 1503 | UA | LAX | -5 | -10 | ... | 2536 |
| 540 | PS | BUR | 13 | 5 | | 186 |
| 1920 | DL | BOS | 10 | 32 | | 1876 |
| 1840 | DL | SFO | 0 | 13 | | 568 |
| 272 | US | BWI | 4 | -2 | | 359 |
| 784 | PS | SEA | 7 | 3 | | 176 |
| 796 | PS | LAX | -2 | 2 | | 237 |
| 1525 | UA | SFO | 3 | -5 | | 1867 |
| 632 | US | SJC | 2 | -4 | | 245 |
| 1610 | UA | MIA | 60 | 34 | | 1365 |
| 2032 | DL | EWR | 10 | 16 | | 789 |
| 2134 | DL | DFW | 6 | 6 | | 914 |



# High Performance Analytics Workflow



# Strengths & Limitations

- Strengths
  - analytics are made easy when they fit in the map reduce approach
  - enables easy design of data exploration systems
  - there are mature data and cluster processing systems (e.g., Apache Spark, Apache Flink)
    - the data processing frameworks automatically take into account data locations when distributing the task
- Limitations
  - considerable learning curve
  - we can still face some scalability problem

# That's all Folks!