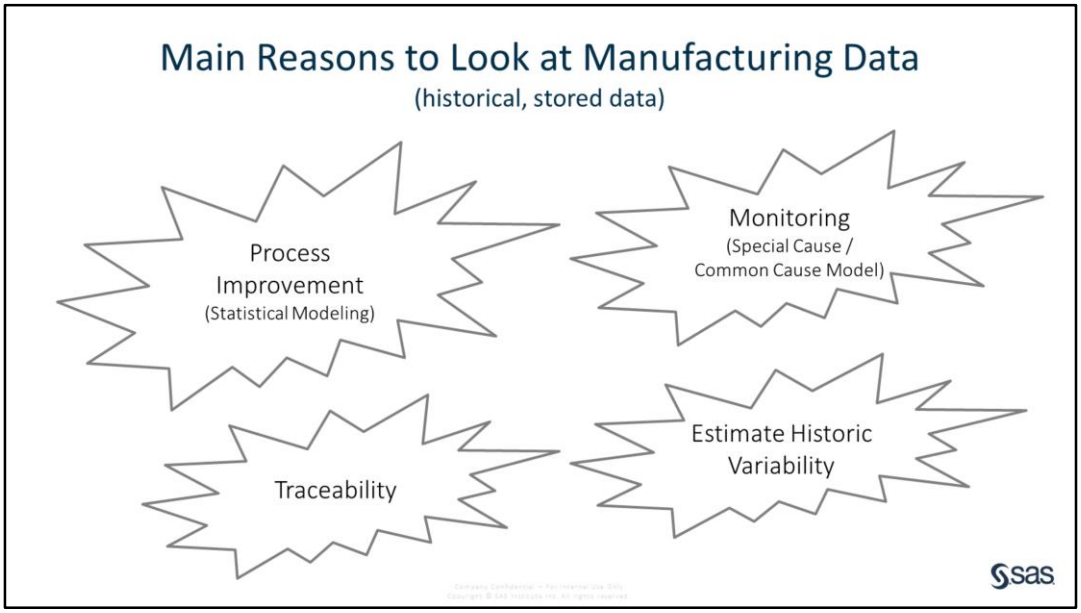
The Opportunities and Challenges of Analyzing Manufacturing (Big) Data

Sam Edgemon
sam.edgemon@sas.com

Tony Cooper
tony.cooper@sas.com

Summary Details

- This is an ongoing project
- We are going to be talking about historical, stored, “big” data
- The response variable will be (what the clients call) “scrap”
- There are 2 kinds of models
  1. Prediction
  2. Explanatory
- We use models to establish what to monitor (in real time)
- Exploratory Data Analysis (EDA) is very important!
  1. Learning about data (myself AND data owners)
  2. Communication channels are established and confidence instilled



Big Data in manufacturing was virtually unheard of twenty years ago, and it is still cutting edge.

Monitoring

- On going evaluation of the stability of process operation.

Process Improvement

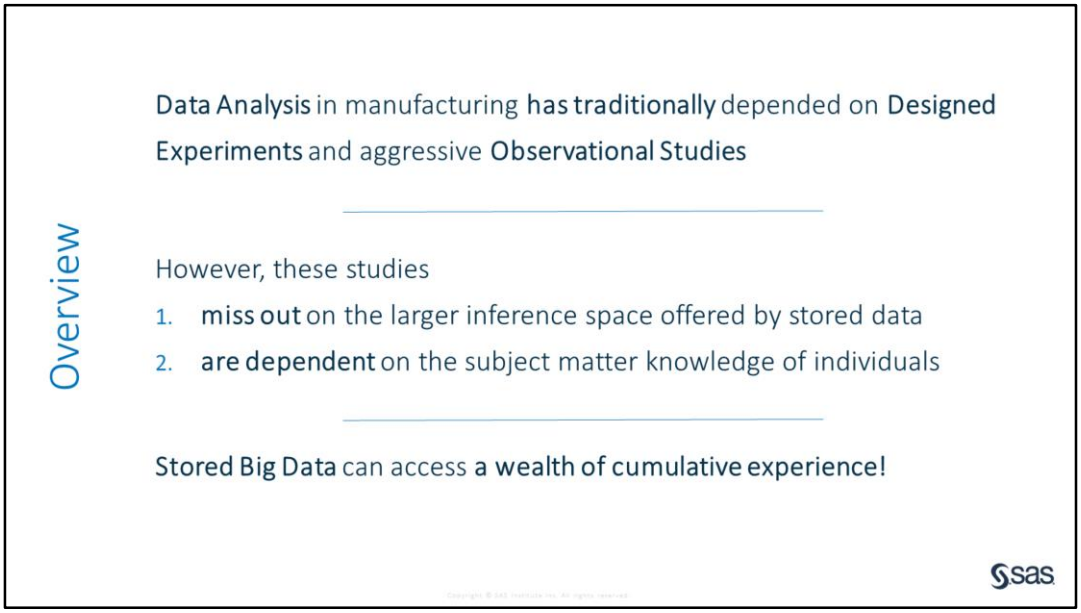
- Generate theories and improve understanding of Causal Structure
- Understand inference
  - Representative data
  - Levels and ratios experienced
- Inform experimentation

Traceability

- What was happening last year when field complaints seemed to rise?
- Machines used? Setting? Raw material source?

Estimate Historic Variability

- See how much inputs have moved over time
- If you were to come with a new robust design – what do you need robust to?

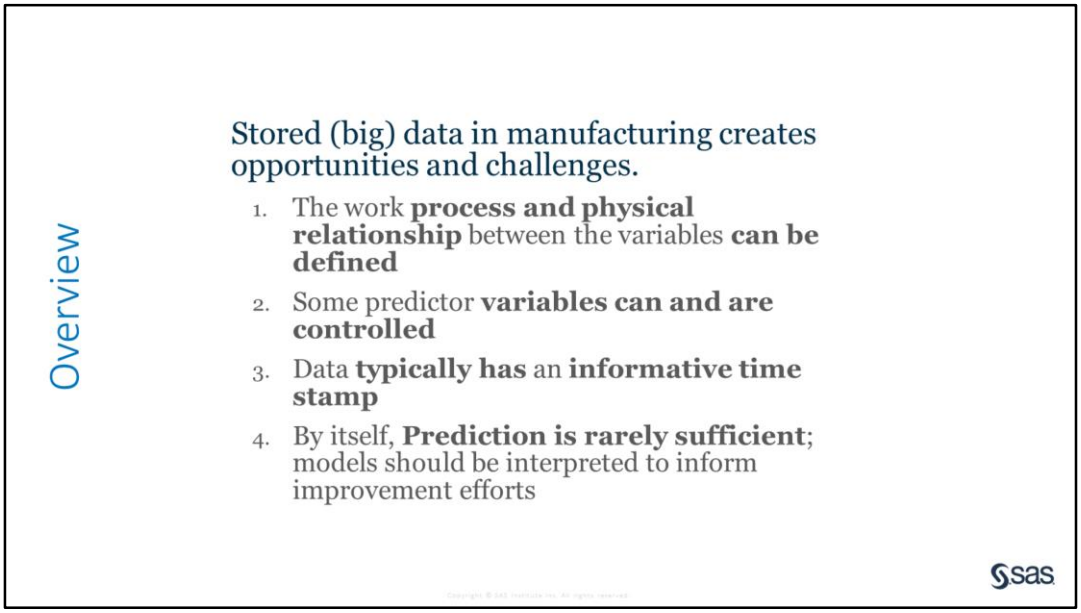


Manufacturing processes can and should be understood.
Contrast modeling First Pass Yield (FPY) at the end of a wafer line to modeling churn at a telecom. The wafer manufacturing process may be complex, but it is much more tangible that getting in the heads of telecom customers.

It is recommended that the analyst develop a process map, listing variables, before even looking at the database.

It is well understood that missing information, variables that are not recorded, can lead to misleading analysis.
Data preparation (getting data out of historians, matching with lab analytics and time stamps, summarizing and transposing and retaining) may all be required and are dictated by an understanding of the process.

The objective of statistical analysis and thus statistical tools to be used may be particular. For instance, the goal of data mining is often explanatory modeling vs. predictive modeling



Manufacturing data is typically collected over time

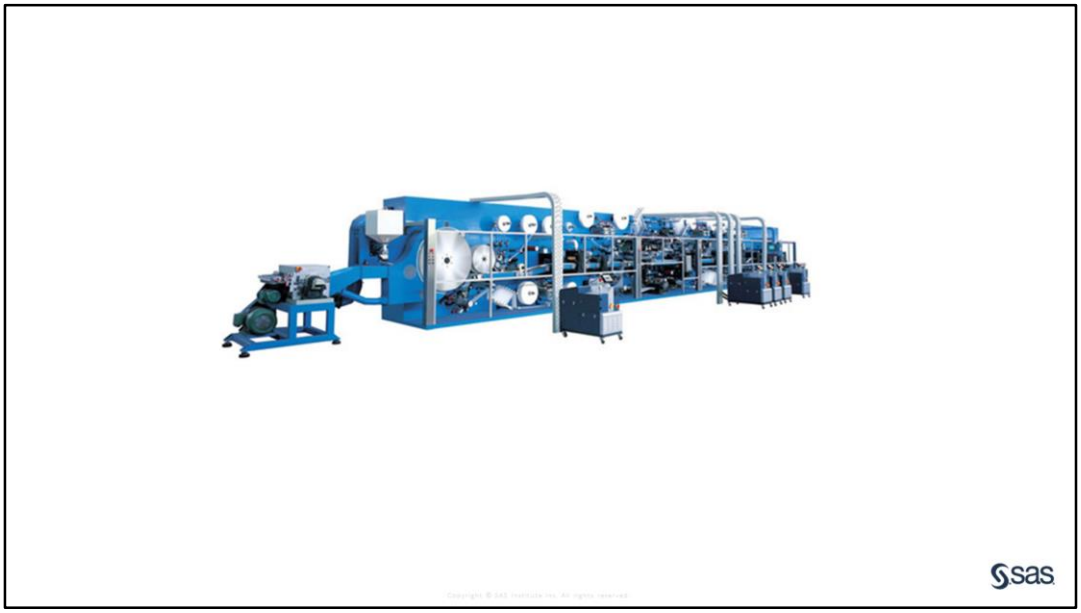
- Time is a powerful surrogate for changing x's and noise
- Exploratory Data Analysis should always plot the data over time
- Modeling is typically a snapshot combining all time (or at least large chunks)

Manufacturing data differs from typical data mining data, such as customer buying habits, in that the organization controls the variation in many of the input variables, e.g., an organization will learn about the effect of a temperature, develop specs., and hold it there.

In most big data, multicollinearity is an issue;

- This could be by underlying principles; older employees may have longer tenure, the reactor temperature will effect the reactor pressure.
- This could occur in practice; common practice is to allocate 70% of a budget increase to task A and the remainder to Task B, operators are advised to increase temperature and mix rate when a certain quality characteristic becomes a problem

Manufacturing typically has continuous response variables. Typical data on customers have binomial responses (did they buy or not).



This is an “Internet of Things” story.

The **Internet of Things** (**IoT**) is the network of physical devices, vehicles, and other items which may be embedded with electronics, software, sensors, actuators, and network connectivity which enable these objects to collect and exchange data

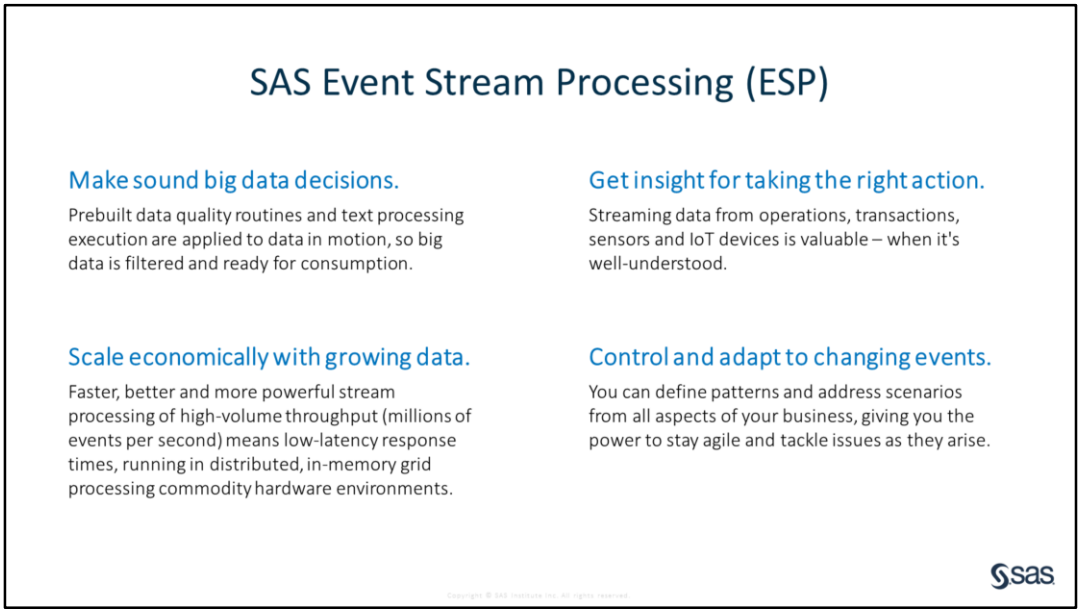
The story goes like this:

- A series of these machines in a single (large) room produce 2,000 “things” every second!
    - 15 machines producing product
    - Product moving through each machine at over 83 MPH
- Imagine the data these machines might produce!
- Now, consider that these machines have 3,000 sensors strategically placed to provide that data!
- That amounts to gigabytes of information per minute…

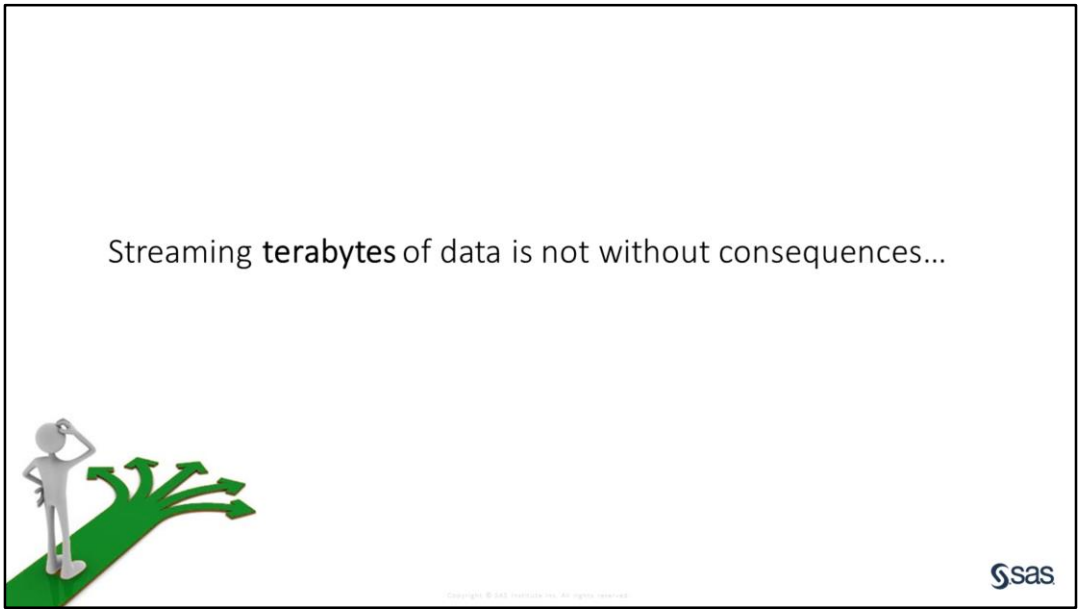
This process creates challenges:

1. Storing the data is a major issue!
2. Using it is the second challenge – “how can we learn about thousands of fields of data?”
    a. 3,000 sensor producing data means there were at least 3,000 fields.

Doesn’t this sound like a great opportunity!

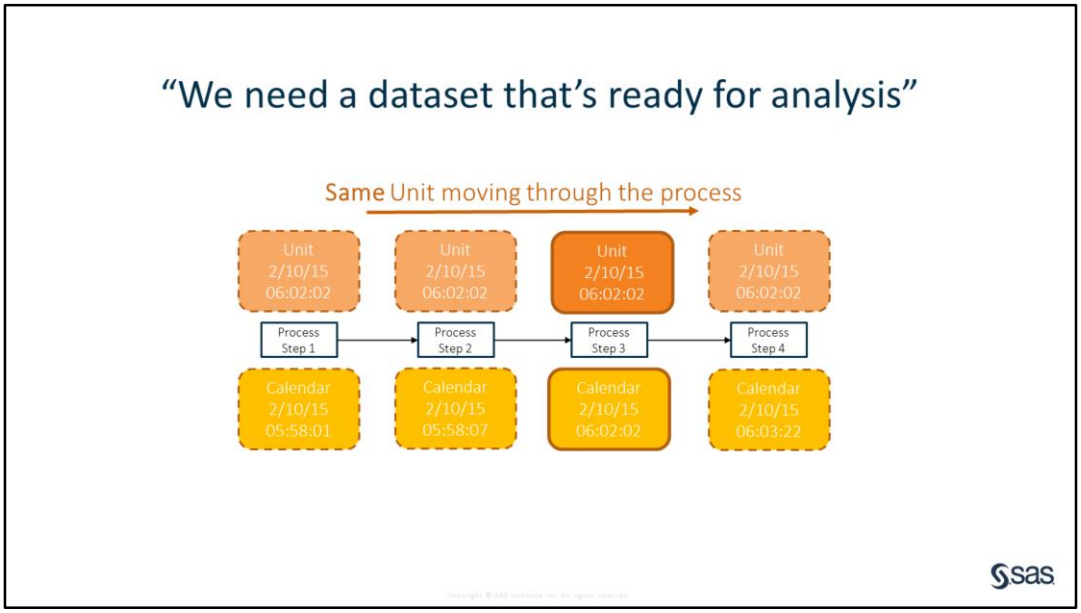


Learn more about SAS ESP at:
https://www.sas.com/en_us/software/event-stream-processing.html



Streaming Terabytes of Data produces great opportunities for analysis, but there are also problems that must be addressed. The most apparent are:

1. How and where do you store it, and
2. Given that we've got thousands of fields … how do we use it?



Data sets will have one row per ID (timestamp)

The inputs and the outputs must be at the same frequency

Data Integrity MUST be Addressed
- ✓ Mistakes in the data
- ✓ Data quality issues
- ✓ New features

The **question** was "**how?**" (do we get this dataset ready for analysis)
The **answer** was "**JMP**"

Live Demo