Experiences with Big Data

Murat Kulahci

Technical University of Denmark
and
Luleå University of Technology

Disclaimer

- This talk is based on experiences we have had over the years working with industrial problems
- The applications are purely in data analytics and more importantly in production statistics
- Being well aware of the danger in generalizing based on limited data, small variation in experiences is comforting in doing so
- Yet some of the conclusions may be perceived as controversial and that is a good thing to start up the discussion

DTU

LULEÅ UNIVERSITY OF TECHNOLOGY

Production Statistics

- The use of statistics in production has been widespread particularly in
  - Process Understanding
  - Product Development
  - Process Improvement
  - Process Surveillance
  - Quality Control
  - Reliability Engineering
  - Maintenance Scheduling and Planning

DTU

LULEÅ UNIVERSITY OF TECHNOLOGY

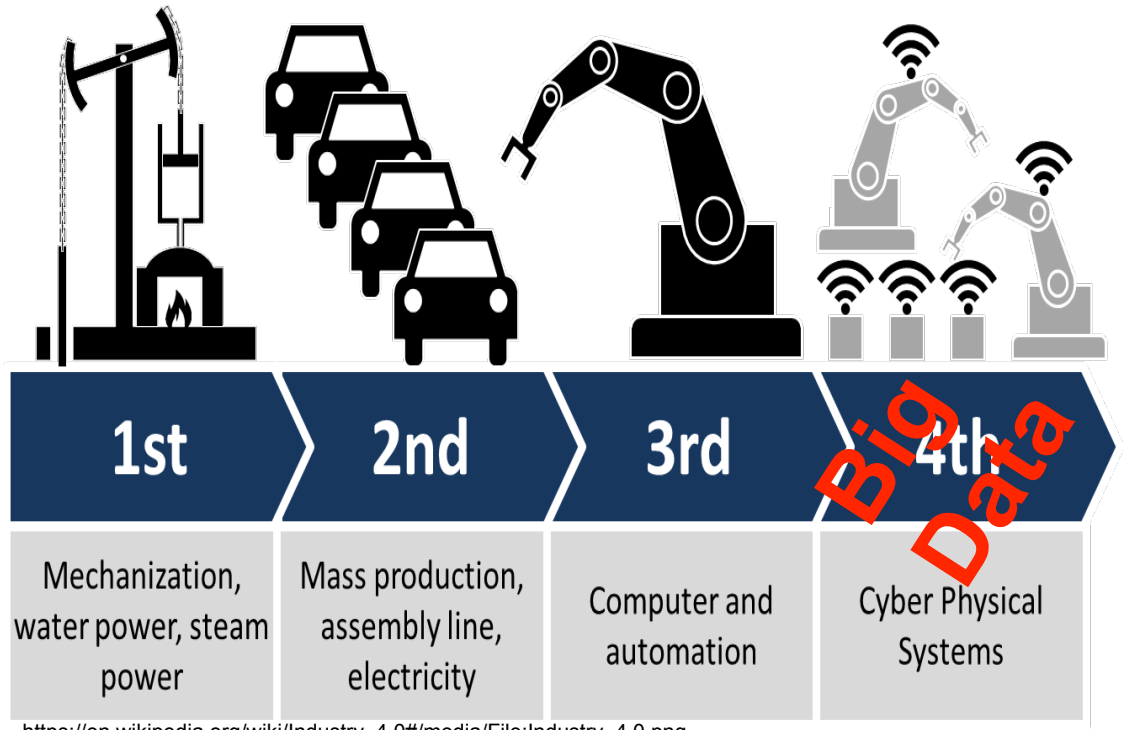
Statistical Tools

- An array of tools have been used in these endeavors
  - Simple descriptive statistics with exploratory plots
  - Design of experiments
  - Statistical modeling for predictive purposes
  - Statistical Process Surveillance
  - Acceptance sampling
- Many of these tools are in need of updating to be effectively used in modern production

DTU

LULEÅ UNIVERSITY OF TECHNOLOGY

4th Industrial Revolution (Industry 4.0)



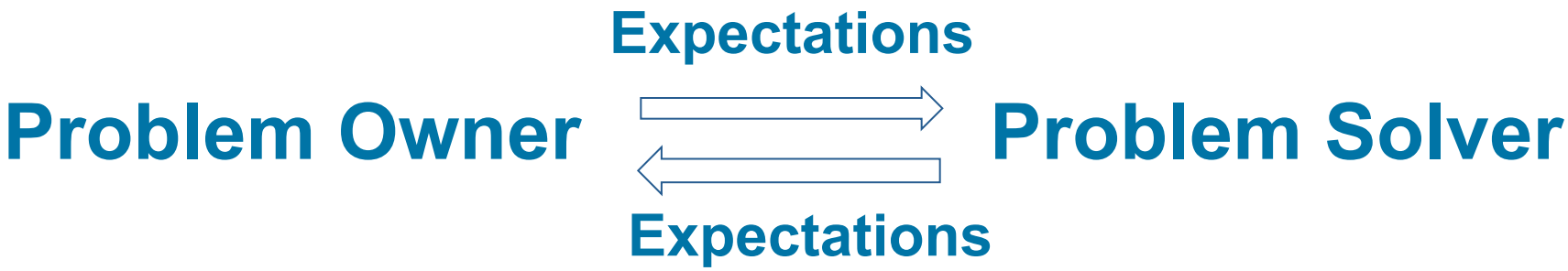
https://en.wikipedia.org/wiki/Industry_4.0#/media/File:Industry_4.0.png

DTU

LULEÅ UNIVERSITY OF TECHNOLOGY

Industry and Academia

- Relationship involves two parties

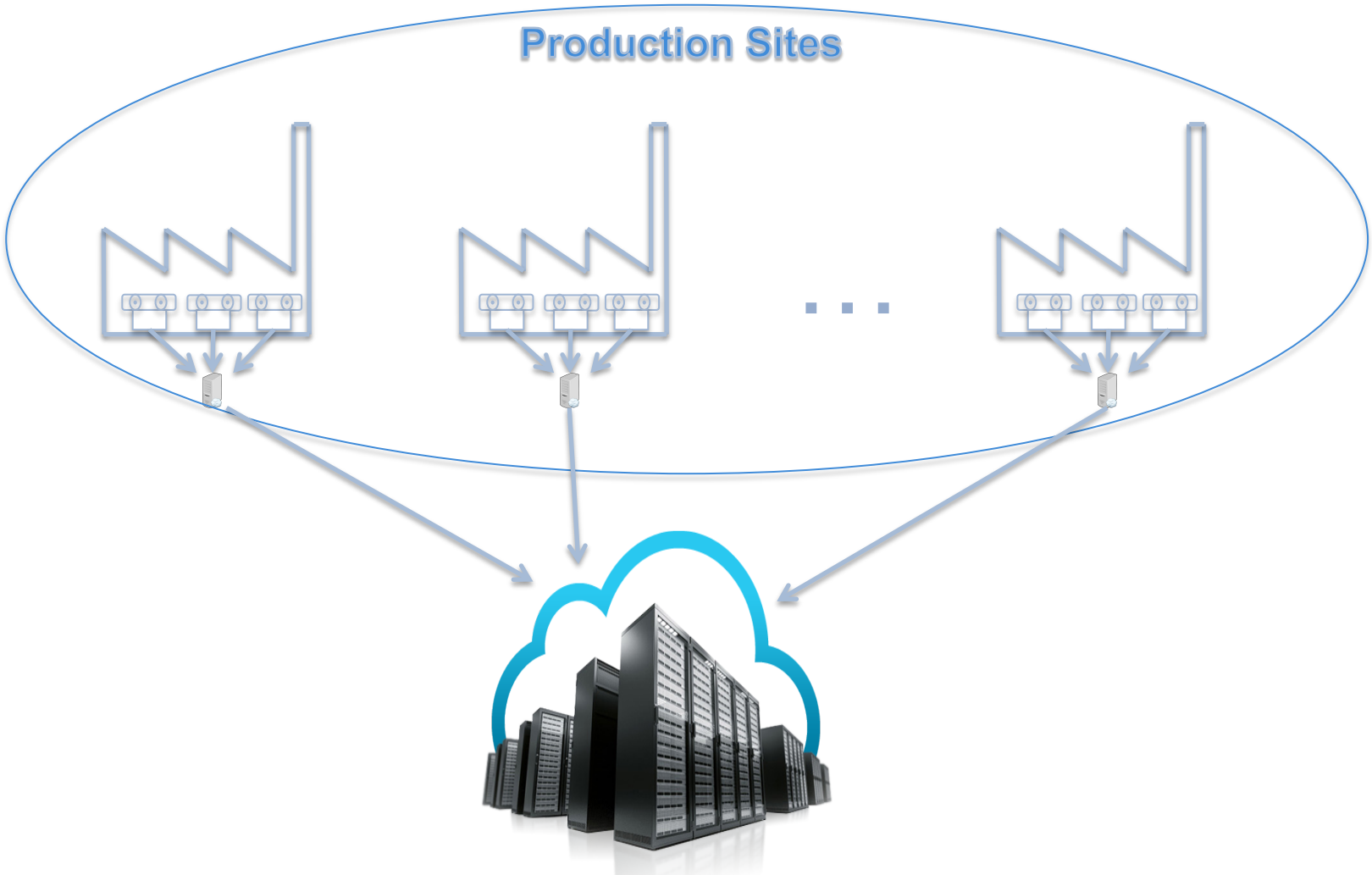


- Key to success is also simple

MATCHING EXPECTATIONS!

MISMATCH IN EXPECTATIONS

Initially: Building up Databases (Historians)



Computer clip arts are obtained from http://clipart-library.com/cloud-server-cliparts.html

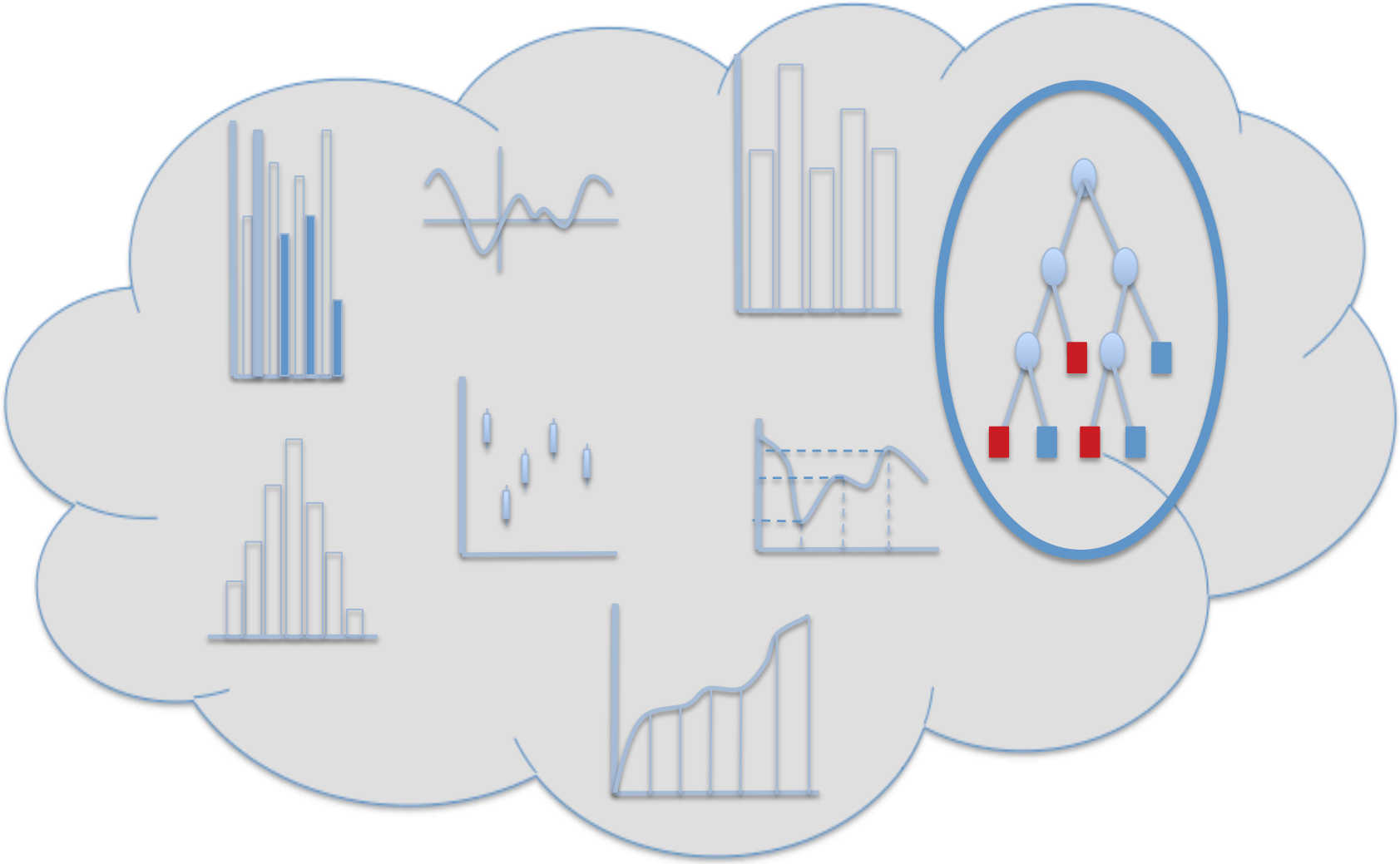
The edict has been

There must be some answers to some questions

DTU

LULEÅ UNIVERSITY OF TECHNOLOGY

Searching for Patterns



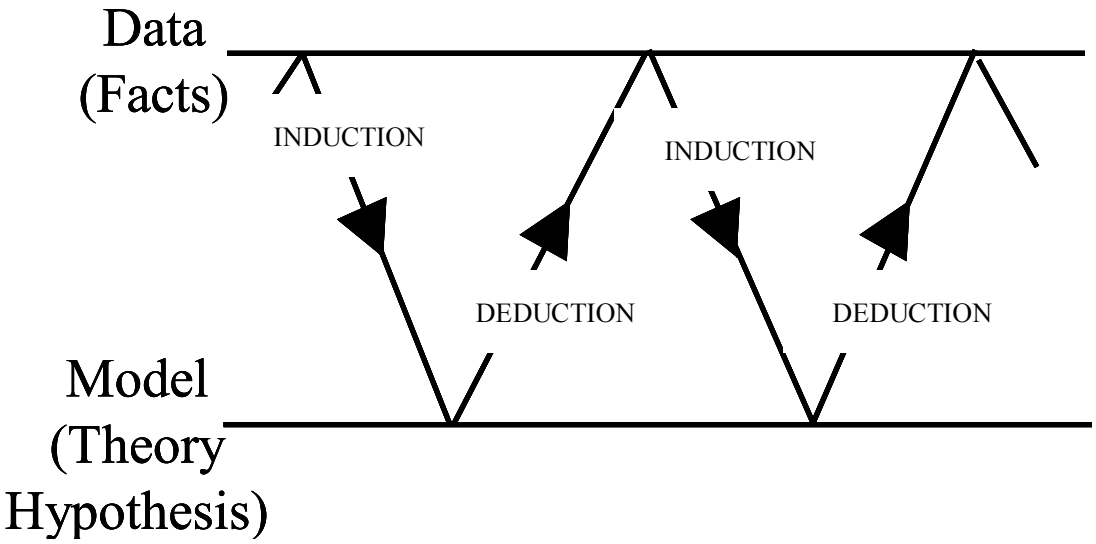
DTU

LULEÅ UNIVERSITY OF TECHNOLOGY

What we do

- Many statisticians (and engineers) are classically trained in scientific investigation

  (Scientific Method)
- A sequential method of induction and deduction heading from observation to hypothesis generation



Data Scientists in general

- We are often not the problem owners nor are we the subject matter experts
- We assume the supporting role and come into play when the hypothesis has already been defined
- We usually recommend that as the first step "The problem should be defined and then …"
- Similar approach is adopted by popular quality management methodologies

DTU

LULEÅ UNIVERSITY OF TECHNOLOGY

Quality Management

- One of the key aspects of Total Quality Management is PDSA cycle: Plan, Do, Study and Act
- The whole thing starts with PLAN and then comes DO
- Six Sigma quality management approach primarily revolves around DMAIC; Define, Measure, Analyze, Improve and Control
- Again the whole process starts with DEFINE followed by MEASURE

DTU

LULEÅ UNIVERSITY OF TECHNOLOGY

First mismatch

- We are conditioned to deduce and not necessarily induce at least at the beginning of the learning process
- Hence “Do something!” did not make initial sense
- Considerable amount of time was wasted in aligning the expectations
- **Then came the data related issues**

DTU

LULEÅ UNIVERSITY OF TECHNOLOGY

Database Issues

- Retrieving data from existing databases has proven to be quite challenging
- Connection to various databases with varying protocols requires different expertise than most data analyst may have been trained for
- Security concerns were well-abounding when handling sensitive production data
- Physical location of the data analyst often required remote access

DTU

LULEÅ UNIVERSITY OF TECHNOLOGY

Merging Databases

- Data from different sites become available
- But often those sites are subjected to different operating conditions
- These sometimes severely impaired the ability to combine data from different sites
- Focus was given on more stable production sites, which is the the right approach at least initially
- But site to site differences are also valuable to extract in order to eventually minimize them

DTU

LULEÅ UNIVERSITY OF TECHNOLOGY

Historical Data

- The claim that large amounts of historical data being available can be misleading
- Usually operating conditions can drastically change to make parts of the historical data incompatible
- Data collection schemes and measurement systems do also get changed and modernized over the years
- This can once again render different parts of the data being incompatible for further analysis

DTU

LULEÅ UNIVERSITY OF TECHNOLOGY

Multi-stage Processes

- Many industrial processes consist of multiple stages
- Historically focus has been on unit operations sometimes taken care of different groups in production
- Data collection is seldom standardized hence connecting all these data becomes extremely challenging
- We do also have different types of data; continuous, categorical, text, etc.
- These represent huge hurdles for many analysis methods

DTU

LULEÅ UNIVERSITY OF TECHNOLOGY

Traceability

- True cradle to grave traceability is rare particularly for chemical processes
- That is, for each product knowing all process variables the raw material was exposed to is usually not available
- Often different production streams are combined and/or split to make the traceability almost impossible
- Connection between process variables and product characteristics require a sensible level of traceability

DTU

LULEÅ UNIVERSITY OF TECHNOLOGY

Data Manipulation

- Data sometimes is not recorded in its raw form
- Manipulations are made to ease storage issues
- Data compression is a common practice particularly in chemical industry
- This results in process variables of various lengths and irregular sampling frequencies
- Many analysis approaches can simply not handle such data

DTU

LULEÅ UNIVERSITY OF TECHNOLOGY

Lack of Specialty Data

- The aim of many data analytics studies is to classify the process or the product to be good or bad as in process surveillance
- “Bad process data” is surprising hard to come by
- Then the methods relying almost solely on good process data are reduced to declaring that “there is something out of the ordinary”
- This becomes particularly important when classification of defect types in a product for example
- That is, going beyond detection towards diagnosis

DTU

LULEÅ UNIVERSITY OF TECHNOLOGY

Process and Product Data

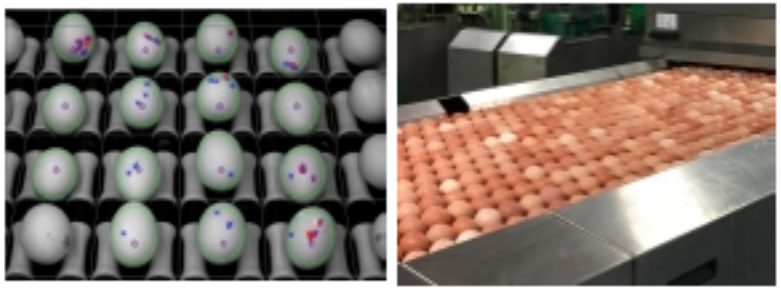
- How do we connect process information with product quality?
- Production rate seems to be an issue
- Fast production rate makes it difficult to obtain product quality for each product
- Moreover type of inspection certainly affects feasibility for more frequent inspection

DTU

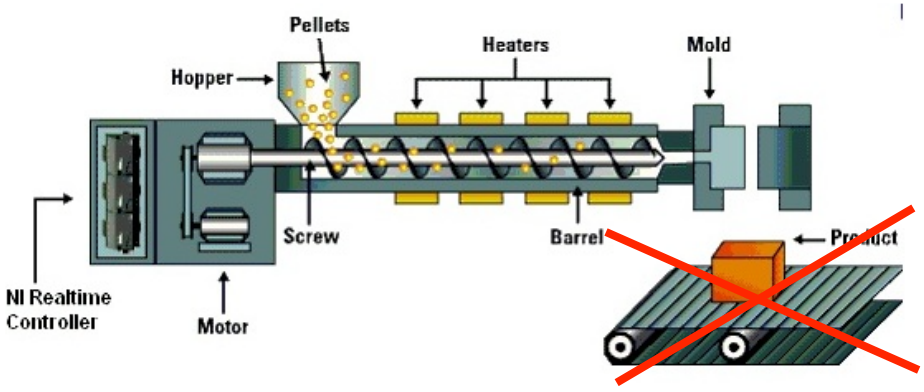
LULEÅ UNIVERSITY OF TECHNOLOGY

Type of Inspection

- 100% inspection
  - Leaky and dirty eggs being sorted out through Image Analysis*



- What if inspection requires more detailed measurements?



IH Foods http://www.ihfood.dk/eggs
The picture for the injection molding machine is taken from http://www.screw-barrel.com/html/Screw-Barrel-For-Injection-Molding%20Machine.html

Sampling


Process Variables
Quality Measurements
X
X1
y1
X2
new data
?
Supervised learning
+
Unsupervised learning
Semi-supervised learning

Semi-supervised Learning

- Combining supervised and unsupervised data is proven to be difficult
- Eventually we can hope for all supervised data, i.e., direct connection between the product characteristics and corresponding process variables data
- More sensorics applications are needed to accomplish that

DTU

LULEÅ UNIVERSITY OF TECHNOLOGY

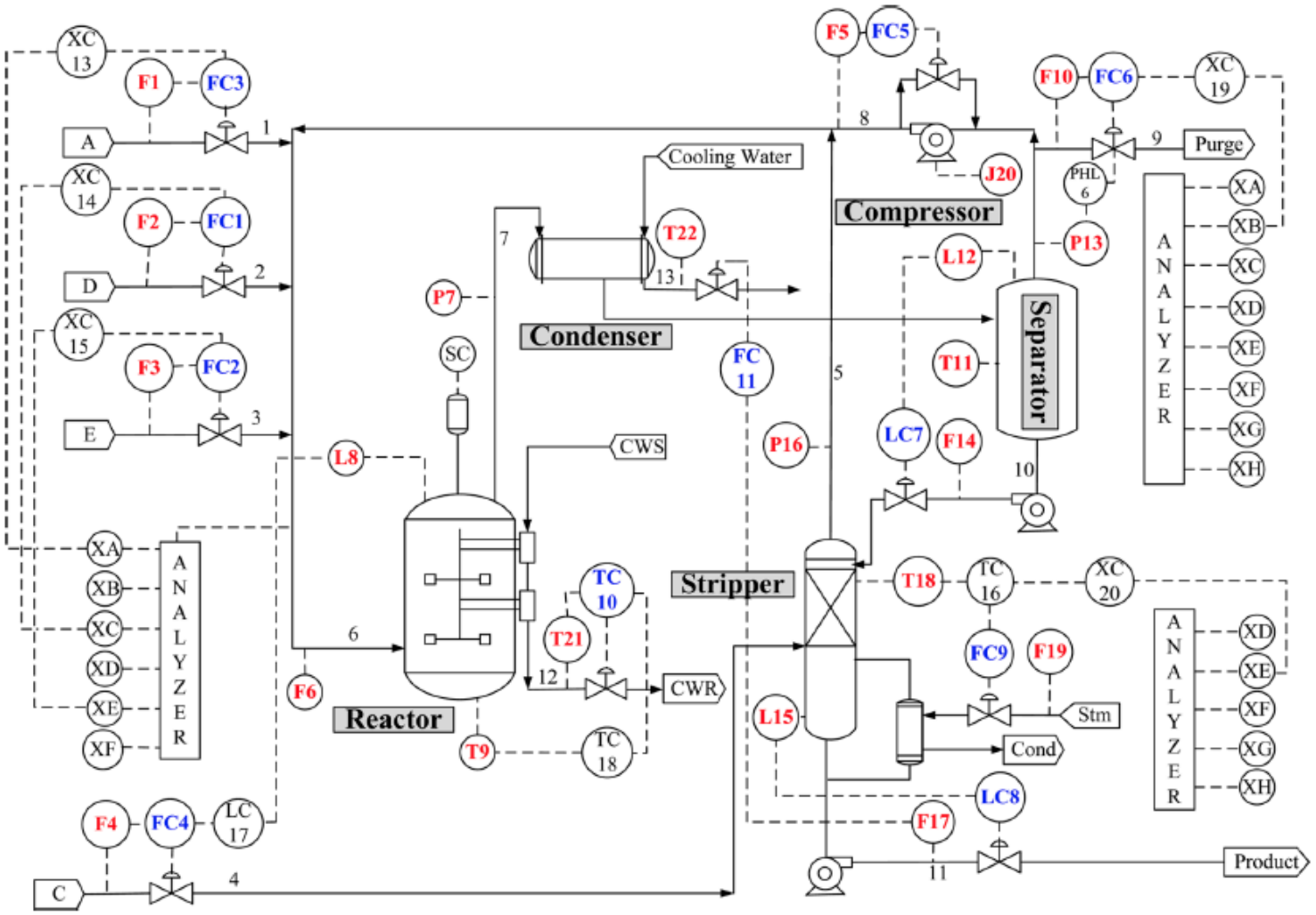
Process Complexity

- Perceived complexity of the processes favors correspondingly complicated approaches in data analysis
- Solutions then tend to be case specific or at least fine-tuned to solve a particular situation rendering generalization difficult
- Process expertise can help alleviate this

DTU

LULEÅ UNIVERSITY OF TECHNOLOGY

An almost realistic process



DTU

LULEÅ UNIVERSITY OF TECHNOLOGY

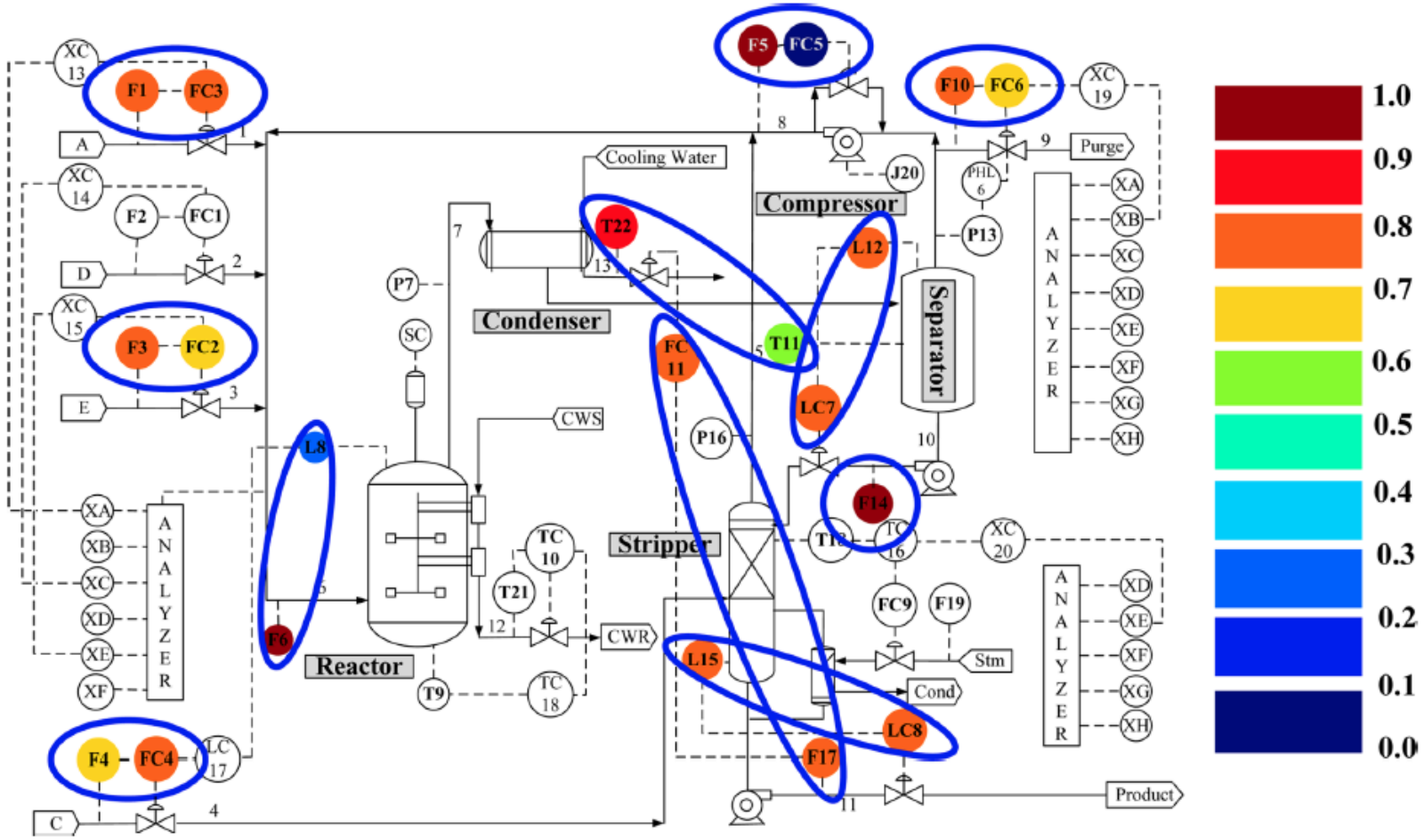
Tennessee Eastman Process

- Accepted to be highly realistic due to its complexity and used in many academic studies for methodology development
- Up to 50 variables can be considered and many simulated faults can be introduced
- All suggested methods attempt to tackle the problem at once
- A slightly skeptical approach reveals a different picture about the interdependencies in the system

DTU

LULEÅ UNIVERSITY OF TECHNOLOGY

Behind the scene



What can we conclude?

- Many of the relationships are in fact known and physically introduced in terms of controlled/manipulated variables pairs
- Isolation of these relationships revealed essential process related correlations
- This in turn provided a clearer focus for example in process surveillance efforts

DTU

LULEÅ UNIVERSITY OF TECHNOLOGY

Correlation and Causality

- Observational data allows for predictive models
- This is achieved through unearthed correlations between the inputs and the outputs
- These models can be quite valuable in making predictions about the future state of the process and hence for risk management
- Expecting more than that can be foolish
- Controlled experiments is usually the way to go

DTU

LULEÅ UNIVERSITY OF TECHNOLOGY

Final Thoughts

- Some of these issues are things of the past and some do linger
- We are working on all of these
- We are however no longer excited with the news of "We have Big Data"
- What is exciting is to have the ability to collect as much relevant data as needed

BIG DATA ⇒ Intelligently collected BIG DATA

DTU

LULEÅ UNIVERSITY OF TECHNOLOGY

Final Thoughts

- Big Data applications should be forward looking
- Furthermore, it is essential to understand that this work is interdisciplinary involving IT, sensorics and data analytics
- So far the concern has been in gathering the data, extracting information and making inference
- How to covert that into actionable decision involves more disciplines such as operations management

DTU

LULEÅ UNIVERSITY OF TECHNOLOGY

Thank You!