Data Science Expertise

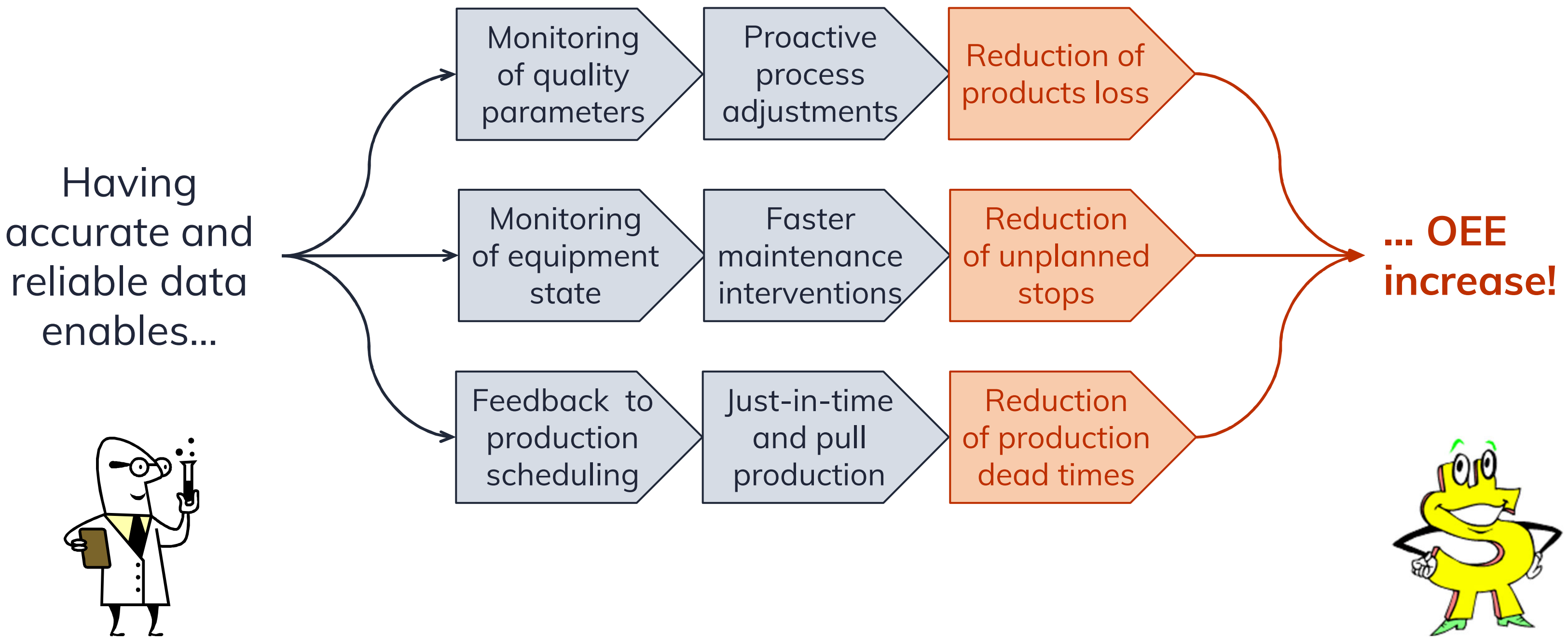
What happened, what will happen?

Last update: 2021-05-11

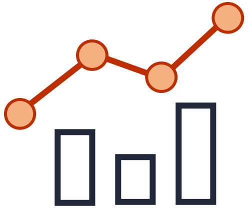
“The” question: why using data?

- Methodic de-bottlenecking and economic ramp-up
- Asset condition monitoring and increase of uptime
- Develop smart solutions for mines, plants, buildings, cities, …
- Diagnose and optimize process controls
- Locate, track and quantify wait times & unproductive delays
- Improve scheduling & supply chain efficiency
- Forecast demand and anticipate productivity
- Democratic information sharing between people & systems

Data and OEE are interrelated

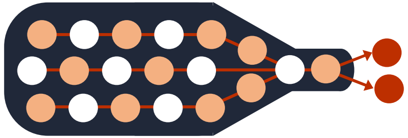


The science of using data serves....



PROCESS
INTELLIGENCE

PREDICTIVE
MAINTENANCE



BUSINESS
IMPROVEMENT

DECISION SUPPORT

= Process control
= Troubleshooting, debottlenecking
= Production optimization
= Maintenance planning
= Engineering and expansion

Examples of data science tasks

- Reducing unplanned maintenance by developing predictive models of track conditions combined with rail health monitoring dashboards
- Increasing productivity through optimization of start-up sequencing for flow paths in complex high throughput bulk-material conveyors system
- Improving process control in difficult to measure situations of various process plants through the development and deployment of soft sensors
- Optimizing process throughput by determining the ideal robust control loop setpoints based on historical data
- Forecasting ground level pollutant concentrations under complex weather conditions and recommending process changes to alleviate impacts

Demystifying Data Science

Too many buzzwords, you think?

Common terms explained

Artificial intelligence

Any technique enabling computers to "decide by themselves" and to mimic human intelligence

Machine Learning

Subset of data science tools aiming at automatically discovering features and predicting from data

Deep Learning

Specialized machine learning technique (neural networks with "lots" of layers = deep); natural language processing, image recognition

Data Science

Use of modern computer capabilities and algorithms to apply manage, clean, analyze, model and predict using data; extension of traditional statistic

Big Data

High volume of (un)structured data collected at a high rate (real time) and stored without being cleansed

Volume-Variety-Velocity-Veracity

Data Engineering

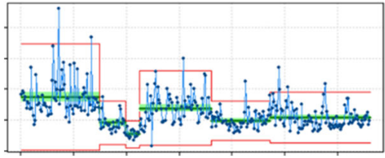
IT techniques making possible to capture, retrieve, store in warehouses and extract data

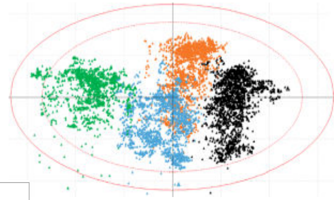
Internet-of-Things

Interconnections between data generators, warehouses, analyzers and users; all the above would be boring without it

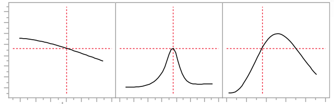
Common data science tasks

- Regroup/segment/decompose
  - Group observations in clusters, combine variables in components
  - Find hidden stable subsets or subperiods from apparent chaos

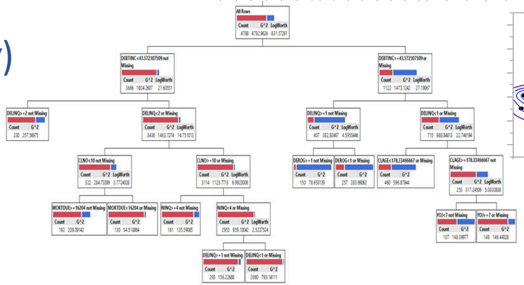


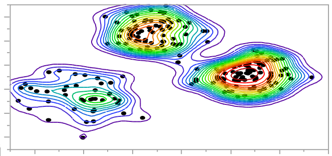


- Classify/sort
  - Tell which observations belong to which category/class
  - Match patterns and images against templates



- Quantify
  - Predict future values (levels and variability)
  - Explain what contributes to these values

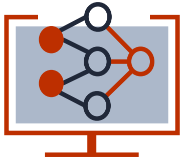




Datamining = learning from data

- Understand the past (descriptive)
  - What happened? Why? When?
  - Which factors explain variability?

- Anticipate the future (predictive)
  - What could happen? When? How certain?
  - How well can we estimate future results?



- Find “the” recipe (prescriptive)
  - How to make things happen as wished?
  - What is the best move onward?

First-things-first roadmap to AI

Descriptive Analytics

Current state assessment

- How does it work today?
- What is the performance?
- What data is available?
- What data is missing?
- What is the future state?

Process mapping
Exploratory data analysis
Gage R&R and process capability
Performance metrics deployment
Design functional specifications

Predictive Analytics

Computerized scenario evaluation

- Forecast what, where and when
- Dig performance data from DB
- Extrapolate from orders & status
- Predict future system status
- Assess predictions uncertainty

Statistical programming
Datamining and algorithms
Machine learning
Forecasting techniques
Monte Carlo simulation

Prescriptive Analytics

Integrated advisory systems

- Repeated scenario evaluations
- Automated “what if” processors
- Reduce resources idle time
- Reduce delivery times
- Recommend lowest cost solution

Discrete event simulation
Knowledge-based heuristics
Production planning and scheduling
Optimization under constraints

People-driven decision process → Data-driven decision tools → Model-driven intelligent systems

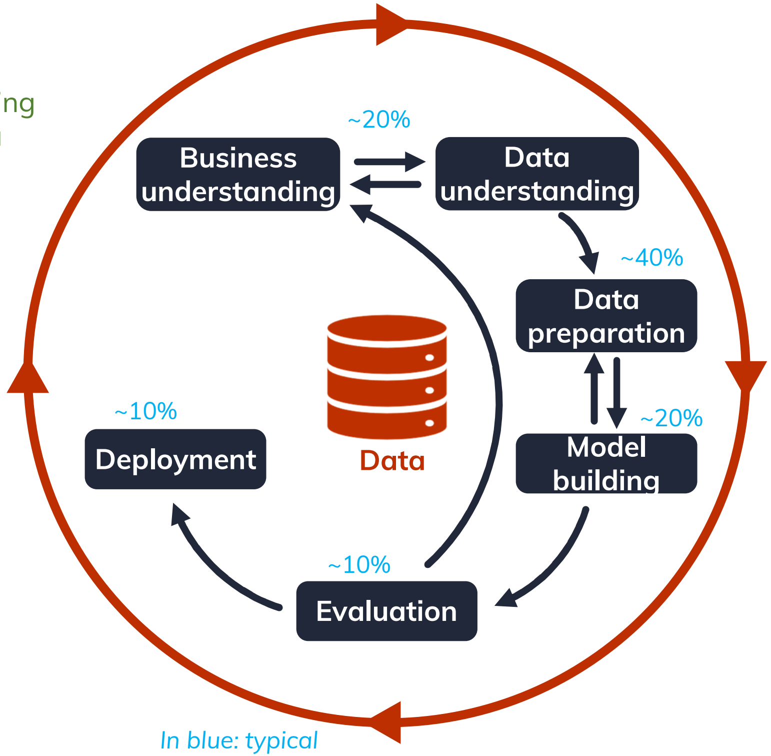
Down-to-earth Approach

Because we are the ones doing it...

Industry standard datamining process

Interviewing people, defining the goals, collecting the data, verifying data quality and meaning

Structuring, recoding, handling of missing values, detecting outliers, aggregating, organizing tables



Developing whatever has to be developed

Transferring ownership to the client, automating the analyzes, making it easy to consume

Checking if the model is meeting the goals, rectifying and refining is needed

In blue: typical effort breakdown

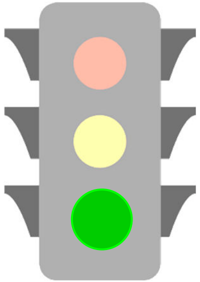
Reality check before starting!

- Before deploying complex things:
  - Do we have the right data vs expected information?
  - Do we have the accurate data?
  - What is the appropriate data volume to consider?
  - How will data “flow” from generators to final display?
  - Is the pipeline there or will humans be the pipeline?

  - Do we have subject matter experts to ensure the success?
  - Just ask: why do we want AI, ML, DS, etc.? What is the need for it?

To keep in mind while going!

- Data science is not magic, it’s just science
    - It won’t make data better…
    - But it will use the most of what’s in it



- Without a context and goals, data is just numbers…
    - Deploying complex autonomous models is great only is people use it
    - People will use data science that was tailored to their reality and needs

- Data science supports decision-making. It should be augmented intelligence: evaluate, predict, advise and let people decide!

Key points of our approach

≡ Sequential (iterative) transformation steps

- Avoid jumping from nothing to full-scale AI, this always fail
- Prevent from disruptive interventions that will break the system
- Define progressive changes, let people adapt to new systems

≡ Ensure productivity and efficiency gains at all stages

- Promote auto-financing as much as possible
- Split project in targeted bite-size high ROI phases
- Put it place the proper KPIs to quantity benefits

≡ Fully customized approach

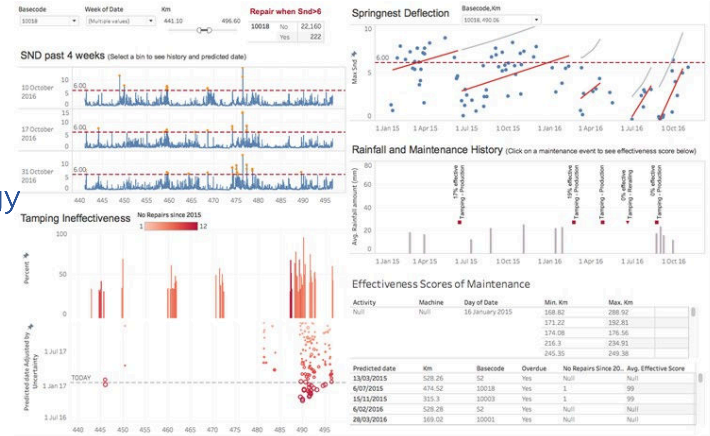
- Generic solutions exist on the market, but generic clients do not exist!
- We adapt existing solutions in collaboration with the client

Selected Case Studies

Not convinced yet? Have a look at this!

Predictive tracks maintenance

≡ Actual inefficient system relying on visual inspections
  - 1,500 km of tracks in 10 m sections
  - Many disparate data sources
  - No way to quantify the effectiveness of maintenance strategy
  - No way to identify where and when to repair predictively

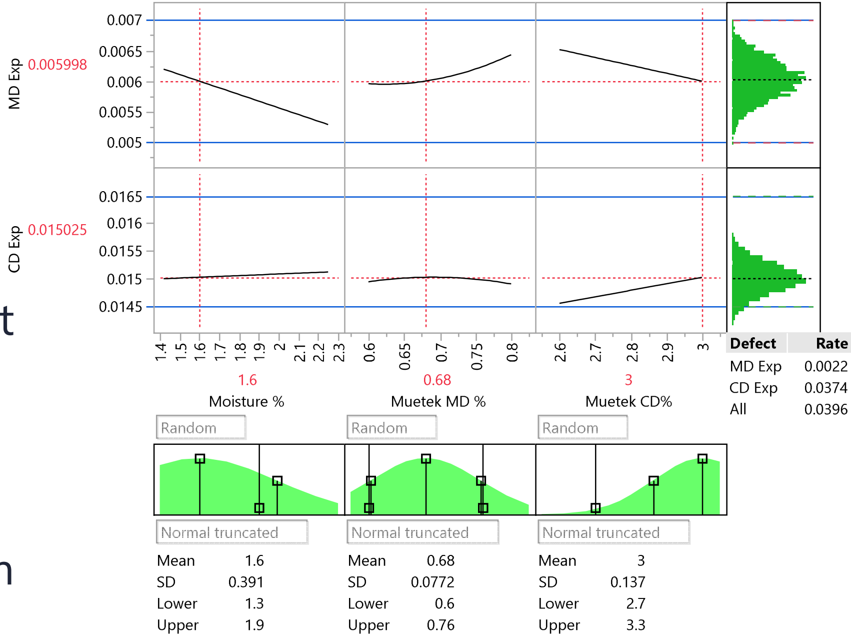


≡ New interactive dashboard
  - Combine and integrate all useful data sources
  - Forecast track condition using machine learning
  - Timely advise required interventions

≡ Live health monitoring and risk assessment
  - Operational and engineering risk factors (including track geometry) are combined to give a total risk score.
  - Dashboards get auto-updated when new data comes in.

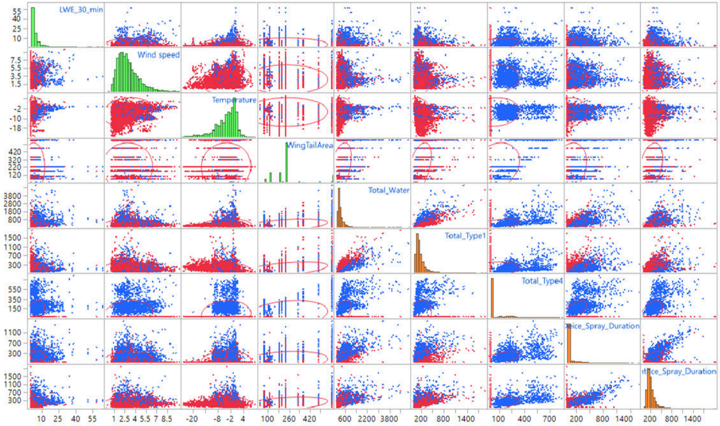
Reducing defect in paper making

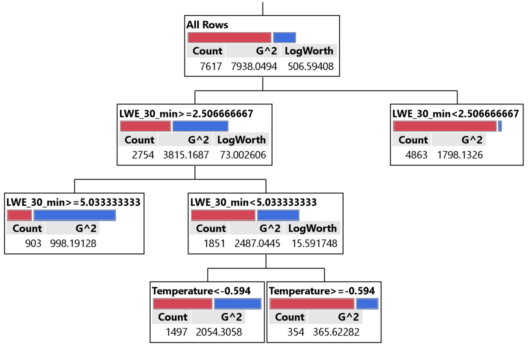
- Client reports many sheet breakages when using specialty paper
  - Skimmed historical data to select the variables related to the client-side sheet quality
- Designed an experiment to test several process adjustments
- Built a predictive model to determine most desirable operating conditions
  - Monte-Carlo simulation is used to estimate the proportion of out-of-specs rolls
  - Robust settings and interactions are identified
- Sheet breakage occurrences reduced from 16% of rolls to 4%



Predicting aircraft deicing duration

- Deicing operators need to anticipate duration and required fluid quantities
  - To coordinate the flow incoming aircrafts
  - To ensure optimal truck refilling strategies – reduce lost time at station and airport outbound delays
  - To dispatch crews according to live fluid levels and deicing tasks
- Investigated available data sources
  - Operations data from deicing crews
  - Weather data from meteorological monitoring stations
- Developed accurate prediction models
  - Less than 3% error when guessing pilots decisions
  - Advisory service developed to challenge possibly improper decision-making by pilots



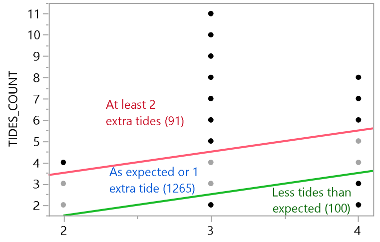


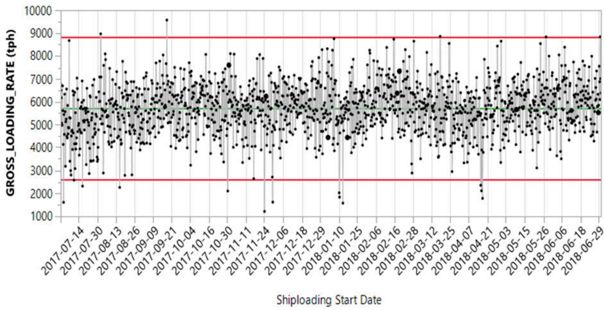
Bulk material port capacity uplift

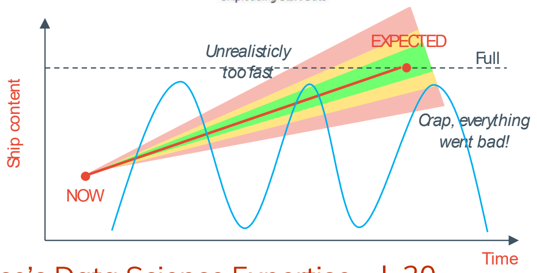
≡ Merged data from multiple sources and inconsistent formats:
- Datasets with 170,000 – 450,000 rows and 16 – 88 cols
- Delays, movements, setpoints, vessels and rakes cycles

≡ Built a portrait of actual shiploading performance
- Evaluated capability to meet tide highs
- Estimated the disruptive impact of dual shiploading
- Assessed the uplifting potential without CAPEX

≡ Developed improvement roadmap:
- Optimizing conveying route start-up + smart controller notifications
- Dashboard and risk assessment: Monte-Carlo simulation to estimate time remaining with chances to meet upcoming tide times

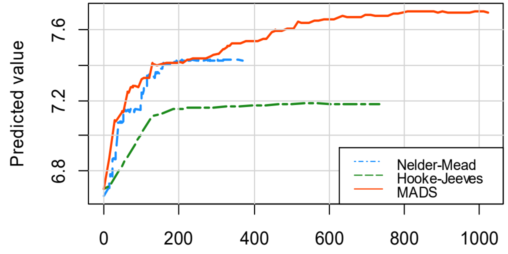


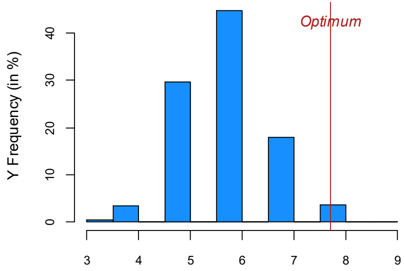


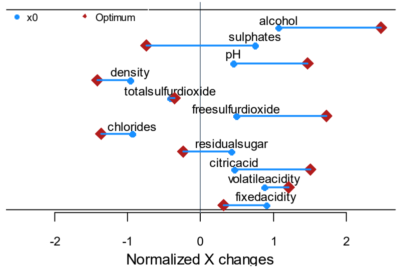


Robust data-driven optimization

- Developed a user-friendly automated framework to optimize performance directly from (big) data
  - Using k-means clustering to discover how data is spreading
  - Using random forests to predict and interpolate new points
  - Using Monte-Carlo simulation to assess new points robustness
  - Using black-box derivative-free optimizers
- Published the methodology and R code
  - Implemented MADS algorithm into the “dfoptim” package (MADS = mesh adaptive direct searches)
  - Issued a scientific paper to describe and justify the approach (to appear in proceedings of the 2019 WinterSim Conference)
- Applied the framework in several projects

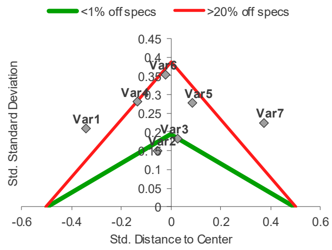


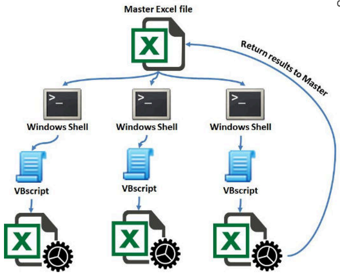


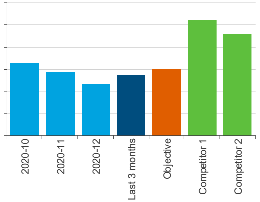


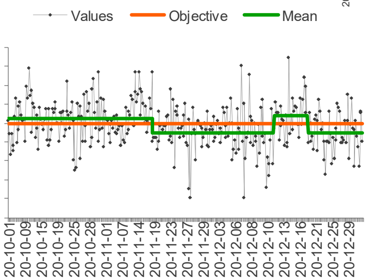
Automated datamaing quality report

- Automation of a complex quality data analysis and report building using Excel VBA
  - Data dumped from SAP in CSV format
  - Time-based data from multiple quality tests performed on grades, plants and brands is re-structured
  - Capability analysis, goal plots and detection of shifts in the process mean is conducted on each grade-site combination
  - A summary workbook with charts and tables is generated
- Advances in data computing with VBA
  - A recursive bootstrapping algorithm to detect changes in the process mean was implemented
  - A multi-threaded job manager for Excel was created to exploit the full potential of multi-core computers for the intensive work
- Benefits of this work
  - Before: a person could terminate 5 reports in 3 days of work
  - After: at least 50 reports are generated in less than 10 minutes
  - Lots of customizability and standardized output









Optimal job scheduling tool

≡ **Development of a job scheduling tool**

- Task durations and setup times modelled by statistical distributions based on historical data
- Possibility to reduce the setup time depending on the previous product family (based on compatibilities)
- Calculation of the total delivery time
- Estimation of the per-order slack times

≡ **Advanced analytical tools!**

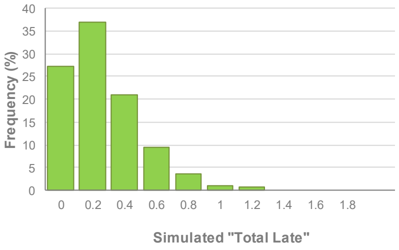
- Implementation of a Genetic Algorithm to minimize the total delays of late orders
- Implementation of a Monte-Carlo simulator to estimate the probability of finishing a job late given the predecessors

≡ **Benefits of this work**

- The genetic algorithm outperformed the expert planners
- Visual aids provide help to assess the risks of finishing late
- An optimal job sequencing is generated in several minutes instead of taking hours by a person to obtain a sub-optimal one

Job sequence		Reset	Refresh		TOTAL LATE:	0.37	
Order ID	SKU	Setup (h)	Production (h)	Total time (h)	Estimated completion (d)	Days late	Late prob. (%)
16	AP02	4.5	6.5	11	0.46	-5.54	0
2	AP02	0	6.5	6.5	0.73	-3.27	0
3	AP02	0	7.8	7.8	1.05	-2.95	0
1	BX08	2.5	9	11.5	1.53	-4.47	0
7	BX08	0	21.6	21.6	2.43	-1.57	0
11	BX05	1.6	9	10.6	2.88	-1.13	0
14	BX05	0	6	6	3.13	-6.88	0
5	BX22	3.6	12	15.6	3.78	-0.22	1.6
13	RT00	3.3	24	27.3	4.91	-1.09	0
6	GD04	3.1	12.6	15.7	5.57	-0.43	0.1
8	NU57	2.4	14.7	17.1	6.28	0.28	96.7
4	NU57	0	12.6	12.6	6.80	-3.20	0
12	AH01	2.9	5.5	8.4	7.15	-2.85	0
9	AH01	0	4.4	4.4	7.34	-2.66	0
15	BX22	3.6	18	21.6	8.24	-1.76	0
17	BX05	1.6	4.5	6.1	8.49	-1.51	0
18	AP02	4.5	11.7	16.2	9.17	-0.83	0
10	RT00	3.3	16.8	20.1	10.00	0.00	52.8

Calculation controls	
Mode	Simulation
Iterations	1000
Conf.Level	0.95



différence
consulting in statistics

Différence is a society offering coaching, consulting and training services in statistic, data science, simulation and continuous improvement.

We promote the use of quantitative tools that can be applied at the different steps of an improvement and variability reduction project.

Powerful methods

Adapted approach

Combining hard work with fun

For more information, you can contact:

Martin Carignan, M.Sc., MBA
Principal Associate
514.795.8000
mcarignan@difference-gcs.com
linkedin.com/in/martincarignan

Vincent Béchard, MASc.
Associate, Analytical Decision Specialist
438.521.5829
vbechard@difference-gcs.com
linkedin.com/in/vincentbechard