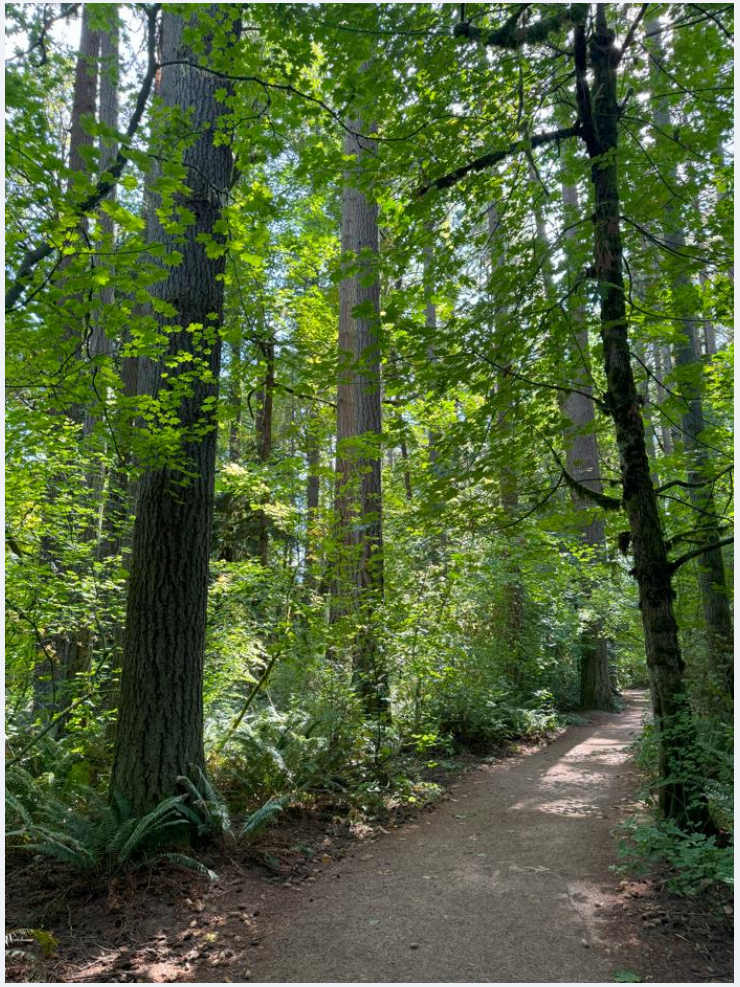
Silicon Crystal Balls: The Science, Math and Methods of Factory Future Prediction, Scheduling and Control

Dr. Holland M. Smith III
Director of Technical Marketing, INFICON IMS

Talk Outline

I. **Stochastic Environments**
   a) Motivating the problem
   b) Psychology of uncertainty
   c) Goals for talk
II. **Factory Future Forecasting**
   a) Predictive models
      - Static
      - Dynamic
   b) Prescriptive models
      - General optimization principles
      - Factory scheduling
III. **AI and Factory Future Forecasting**
   a) Machine learning
   b) Generative AI
IV. **Integrating Best Practices**
V. **Conclusions**



Consider the Following…

1. There is a perfect set of actions to take today that will make the most money one year from now.
2. Nobody knows what they are.
3. They will be perfectly clear one year from now.

What is a good strategy?

Consider the Following…

1. Learn about what has been tried previously, and choose from the options
2. Try to predict what will happen and build a strategy based on this
3. Ask ChatGPT
4. Run away!
5. Something else?

Before we go any futher…

Remember this and pack our humility for the remaining slides…

"Anyone offering automatic detailed foreknowledge of a genuinely complex system is not on the level."
- Jaron Lanier, *Who Owns the Future*

- *It is possible to do everything "right" and still get it wrong*

Acting in Uncertainty: Hard to Give Advice, Hard to Do It

Fun

Fear

Necessity

Decision Quality > Outcome Bias

Goals For This Talk

1) Suitable Questions
2) Models / techniques
3) Making models useful
4) What skills are needed to build them?
5) What is the impact of AI (ML, generative AI, etc.)?
6) Deployment best practices
7) ROI
8) Have fun!

Goals For This Talk

Goals For This Talk

Tools Required to Develop Solution

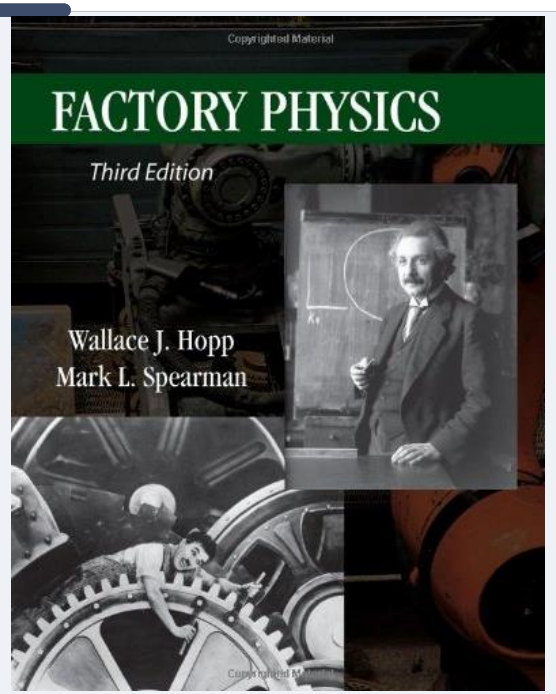
Excel

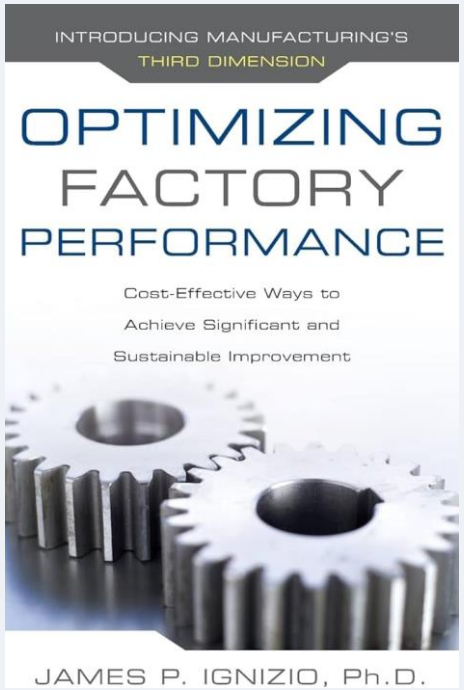
Python or other scripting language (R, JSL, Matlab, etc)

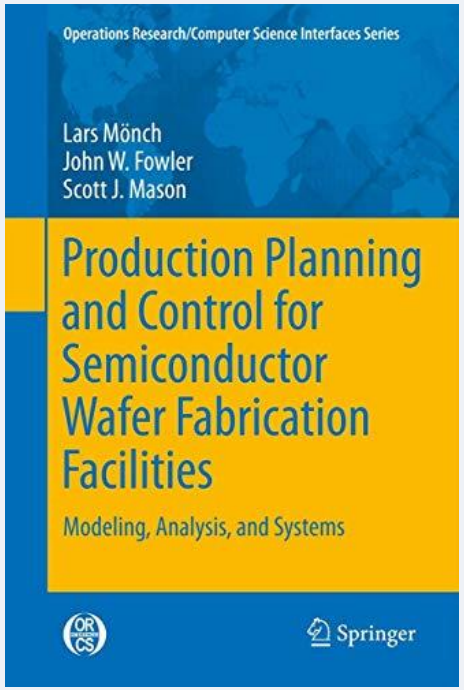
Python or other plus a solver (Gurobi, CPlex, NAG, etc.)

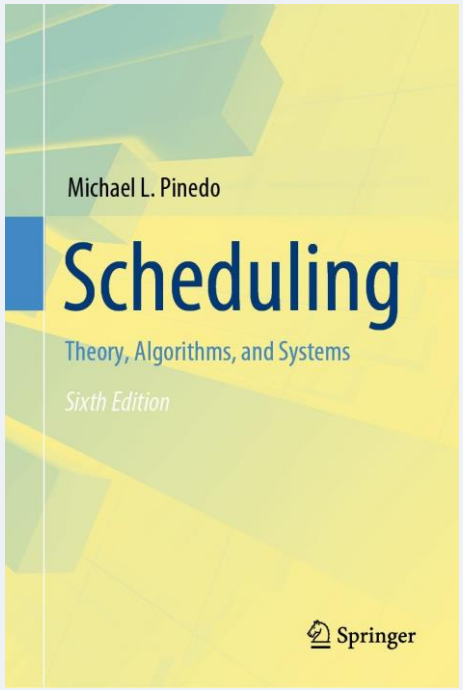
Full software engineering project

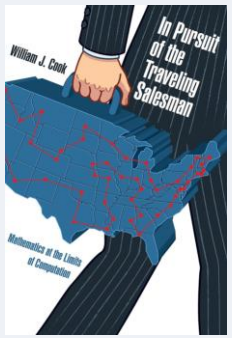
Books You Should Know

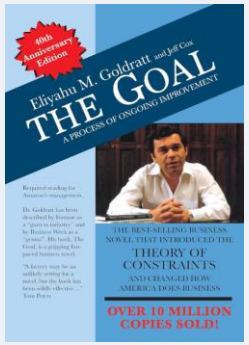


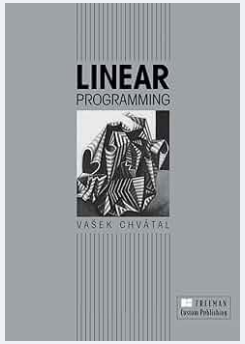


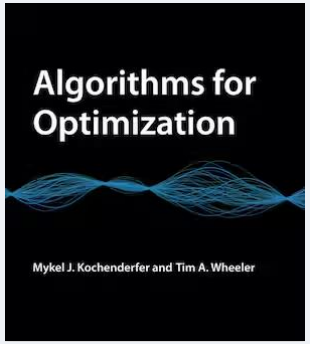


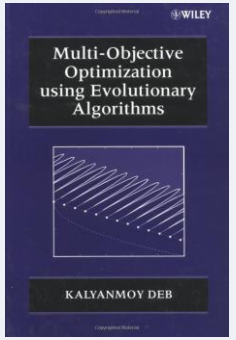






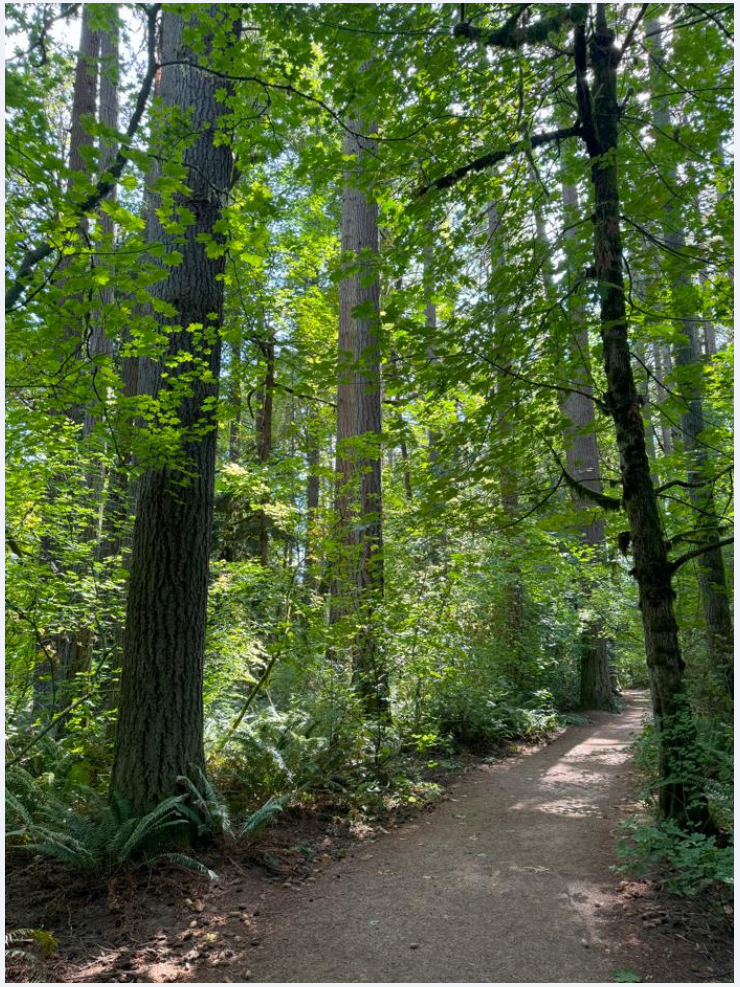




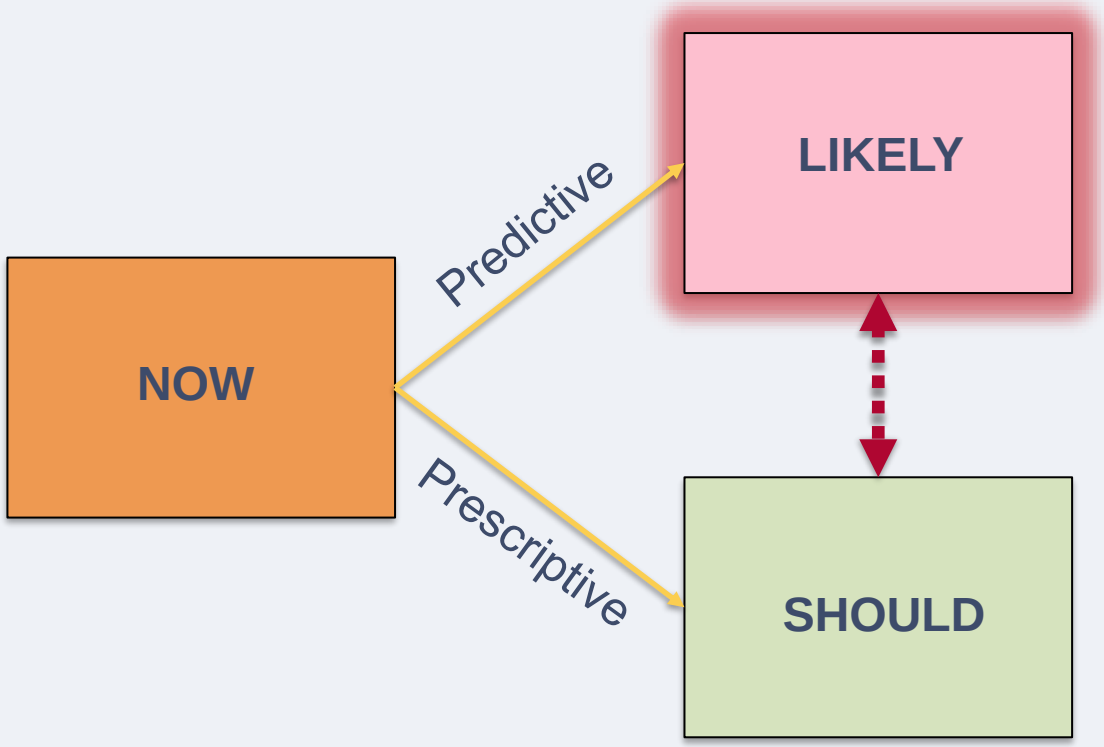


Talk Outline

I. **Stochastic Environments**
   a) Motivating the problem
   b) Psychology of uncertainty
   c) Goals for talk
II. **Factory Future Forecasting**
   a) Predictive models
      - Static
      - Dynamic
   b) Prescriptive models
      - General optimization principles
      - Factory scheduling
III. **AI and Factory Future Forecasting**
   a) Machine learning
   b) Generative AI
IV. **Integrating Best Practices**
V. **Conclusions**



Two Main Branches of Future Forecasting

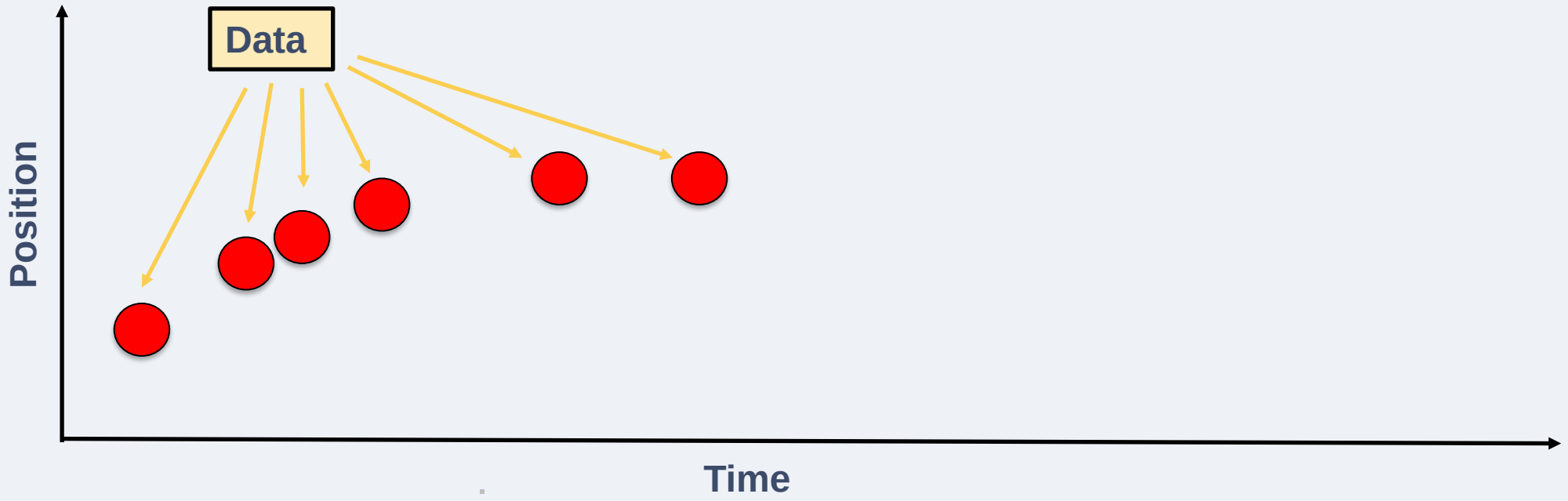


Examples:

- WIP forecasting
- Outs forecasting
- Impact of starts
- …

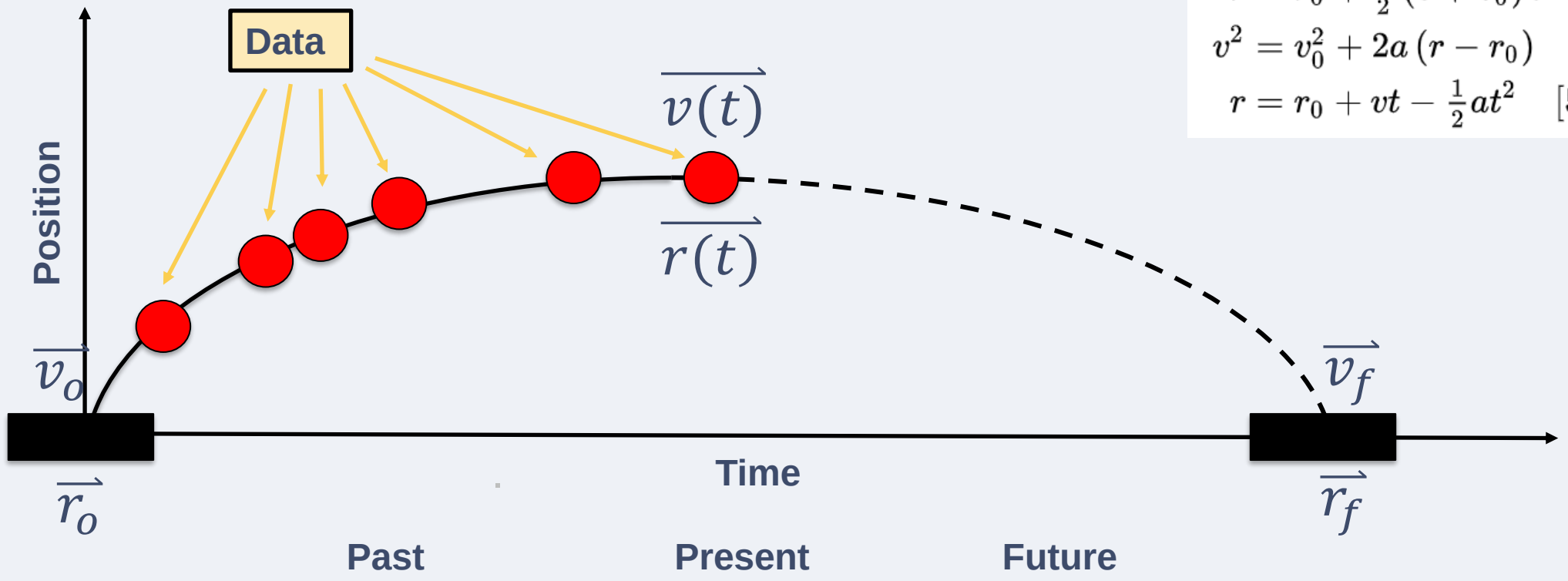
- Starts plan optimization
- Operations optimization
- What is the best thing for me to do ***right now***?
- …

Building a Predictive Model



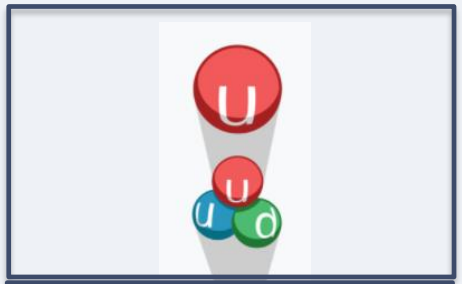
Building a Predictive Model

$$v = at + v_0 \quad [1]$$
$$r = r_0 + v_0 t + \tfrac{1}{2} a t^2 \quad [2]$$
$$r = r_0 + \tfrac{1}{2} (v + v_0)\, t \quad [3]$$
$$v^2 = v_0^2 + 2a\,(r - r_0) \quad [4]$$
$$r = r_0 + vt - \tfrac{1}{2} a t^2 \quad [5]$$



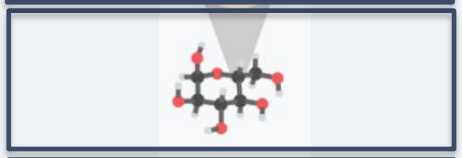
Our Focus: Factory as a Whole

Particle Physics:
Subatomic particles

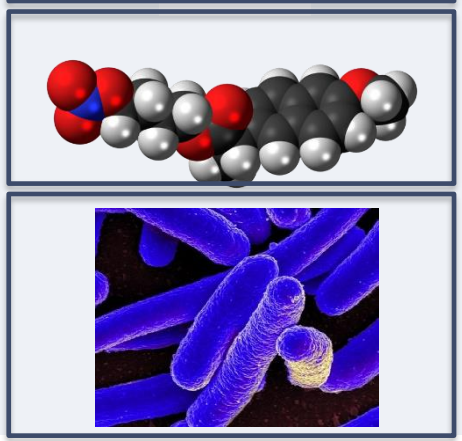


Nuclear Physics:
Atoms

Solid State Physics:
Crystals

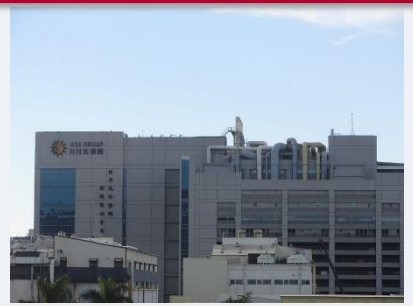


Physical Chemistry:
Molecules



Biology:
Organisms

Process and Equipment Engineering:
Tools



Industrial Engineering:
Factories

Enterprise Engineering:
Supply Chain

INFICON

https://commons.wikimedia.org/wiki/File:Massenskalen.png

https://commons.wikimedia.org/wiki/File:Naproxcinod_molecule_spacefill.png

https://commons.wikimedia.org/wiki/File:E._coli_Bacteria_(1657874457).jpg

https://commons.wikimedia.org/wiki/File:Advanced_Semiconductor_Engineering_Factory.JPG

https://commons.wikimedia.org/wiki/File:Supply_chain_%26_logistics.png

Factories: High-Level Functions

Sales

1. Order intake

Planning

2. Determine feasibility
3. Determine order sequence
4. Make commitments for out dates
5. Turn orders into lots, determine when to start in factory

Operations

6. Develop operational practices and technologies that achieve plan

Adapted from Fig 1.1 - Chapter 3: The MRP Crusade, Hopp and Spearman

Model #1: “Simple” Example – Rough Cut Capacity Planning

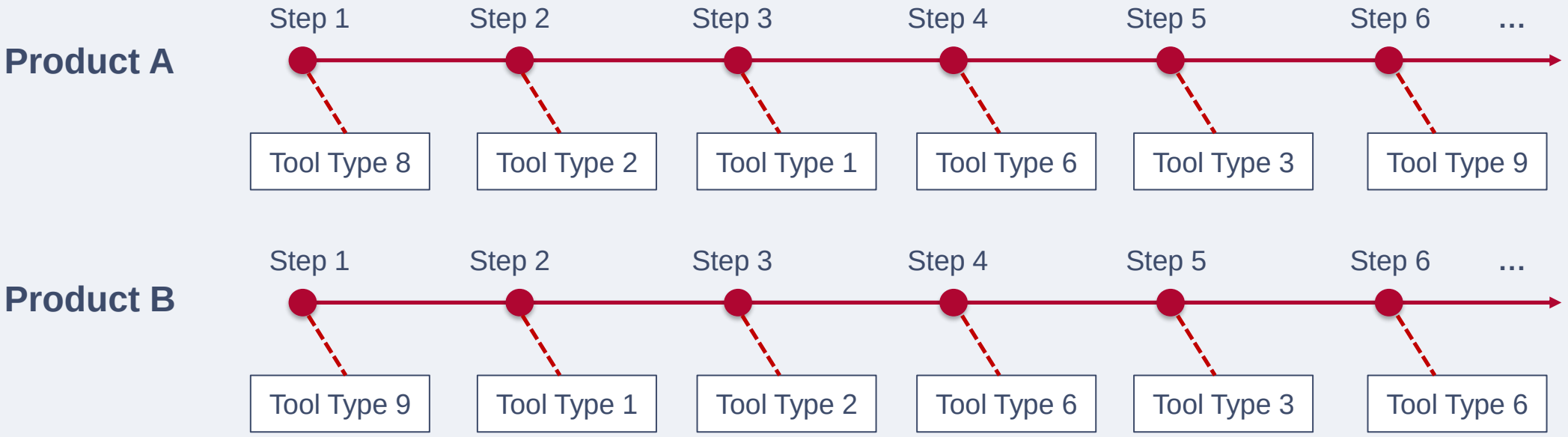
Determine Feasibility of Orders:

Factory has orders for 500 units of Product A and 200 units of Product B per week. Is it possible?

Model #1: “Simple” Example – Rough Cut Capacity Planning

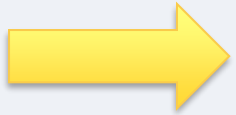
Determine Feasibility of Orders:

Factory has orders for 500 units of Product A and 200 units of Product B per week. Is it possible?



Model #1: “Simple” Example – Rough Cut Capacity Planning

Product	Qty	Step	Tool Type	Minutes per Unit	Total Process Minutes
A	500	1	8	15	7500
A	500	2	2	6	3000
A	500	3	1	30	15000
A	500	4	6	11	5500
A	500	5	3	12	6000
A	500	6	9	15	7500
B	200	1	9	20	4000
B	200	2	1	10	2000
B	200	3	2	11	2200
B	200	4	6	3	600
B	200	5	3	19	3800
B	200	6	6	4	800

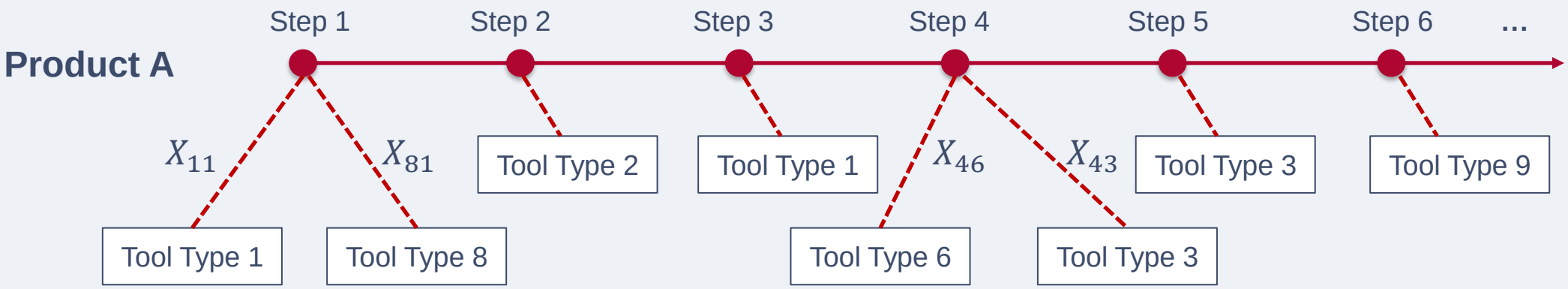


Tool Type	Total Minutes	Minutes in Week	Req. Utilization %
1	17000	10080	168.7
2	5200	10080	51.6
3	9800	10080	97.2
4	0	10080	0
5	0	10080	0
6	6100	10080	60.5
7	0	10080	0
8	7500	10080	74.4
9	11500	10080	114.1

Ok…

But what if a given step can run on multiple tool types? (very common)

Model #1: “Simple” Example – Rough Cut Capacity Planning



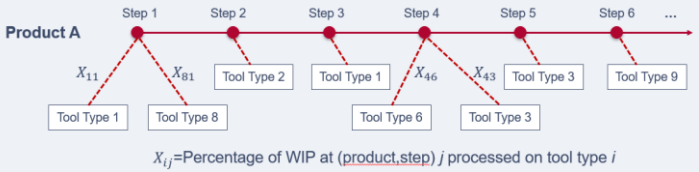
X_{ij}=Percentage of WIP at (product,step) *j* processed on tool type *i*

Our job now:
How can we determine the best x_{ij}?

Options:
1) 50-50?
2) History?
3) Some type of optimization?

Model #1: “Simple” Example – Rough Cut Capacity Planning

Optimization: Solve for the best x_{ij} (allowing WIP to be continuous rather than discrete)



$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leq b_i, \quad i = 1, \ldots, m. \end{aligned}$$

Option 1: Minimize Total Process Time

$$T_{total} = T_{(total,1)} + T_{(total,2)} + \cdots + T_{(total,M)}$$

Linear Programming

Option 2: Minimize Variance

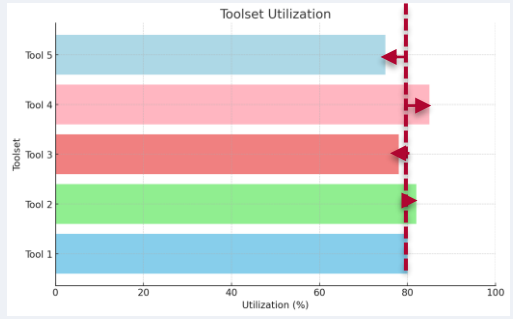
$$\theta = variance = \sum_{i=1}^{M} (T_i - \bar{T})^2 \quad \text{where} \quad \bar{T} = \frac{T_A + T_B + \cdots + T_M}{M_{(tool\ types)}}$$

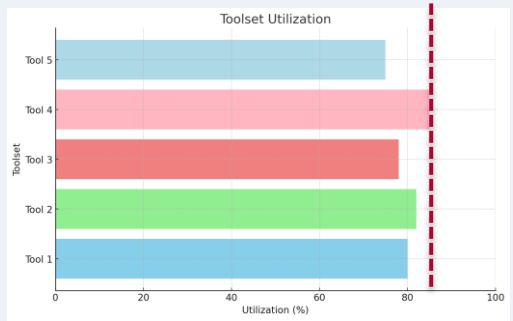
Quadratic Programming

Option 3: Minimize Maximum Utilization

$$\theta = \min[\max(T_i)] \; where \; i \; \epsilon \; \{A \ldots M\}$$

Mixed Integer Linear Programming / Dynamic Programming / Heuristics



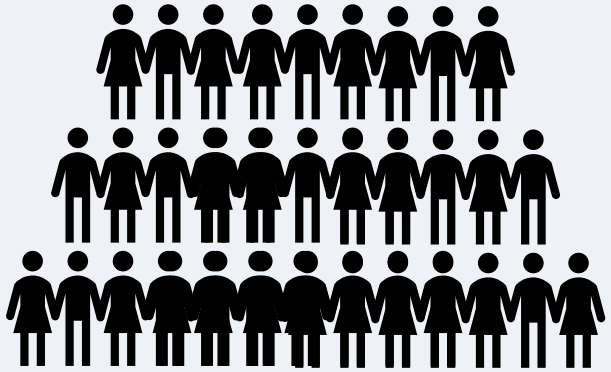


Model #1: “Simple” Example – Rough Cut Capacity Planning

“Simple” Question: Is a Starts Plan feasible?

No cross-qualified tools → **With cross-qualified tools** → **One more complication: Product substitutions**

Almost anyone



IE / Programmer / Convex Optimization Engineer

python + Solver

- Computer Scientist / Optimization Engineer

Well that got deep fast!

Seemingly “simple” model details can be very impactful in terms of solution time, cost and complexity.

Static and Dynamic Models

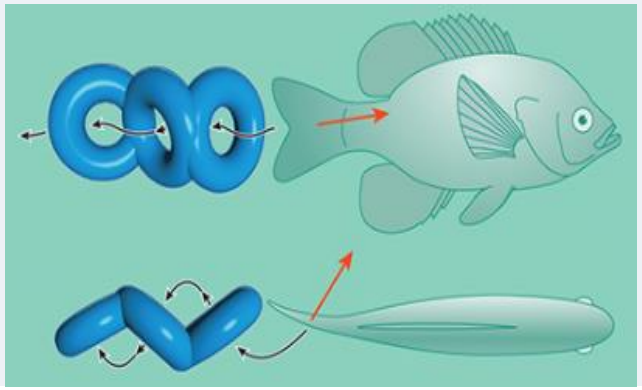
Rough Cut Capacity Modeling is an example of a **static model**

Static Models:
- Ignore factory dynamics (WIP transportation, flow, etc.)
- Focus on an abstracted static state

Dynamic Models:
- Attempt to represent the behavior of the factory over time
- Can investigate both transient and steady-state behavior

Static

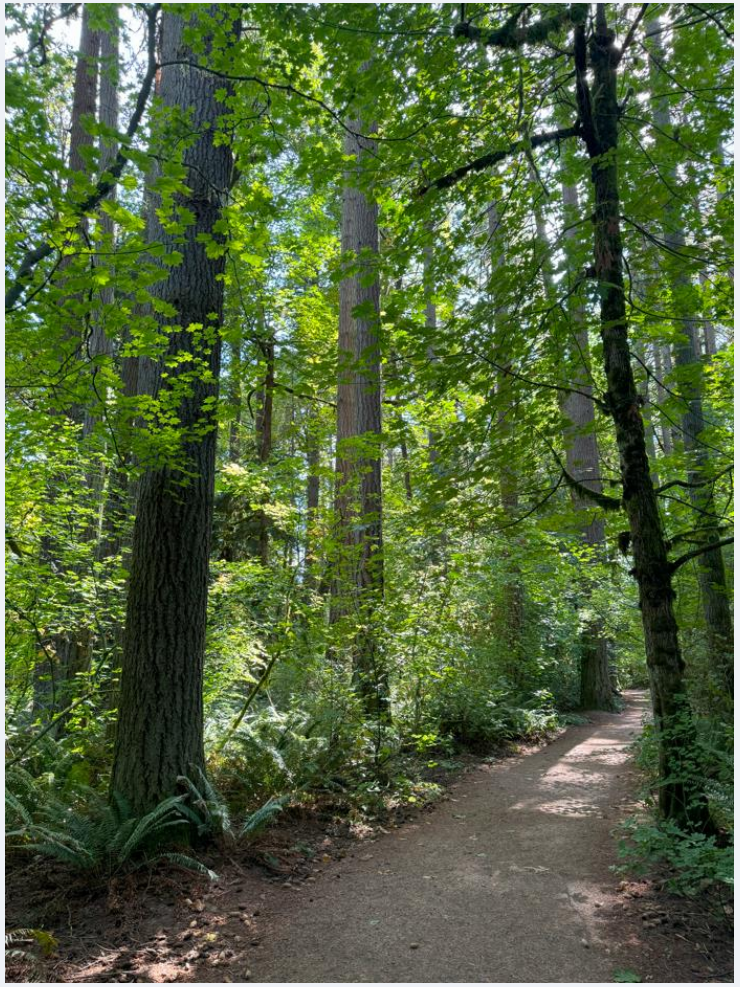
Dynamic - Transient



Dynamic - Steady State

Talk Outline

I. **Stochastic Environments**
   a) Motivating the problem
   b) Psychology of uncertainty
   c) Goals for talk
II. Factory Future Forecasting
   a) Predictive models
      - Static
      - **Dynamic**
   b) Prescriptive models
      - General optimization principles
      - Factory scheduling
III. AI and Factory Future Forecasting
   a) Machine learning
   b) Generative AI
IV. Integrating Best Practices
V. Conclusions



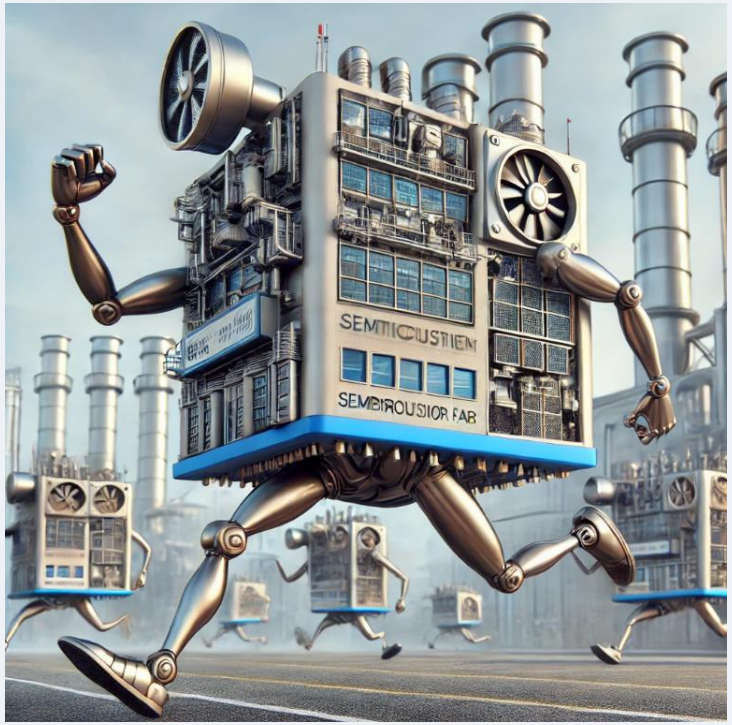
Dynamic Capacity Model Question

Determine Impact of Starts Plan Change:

I’ve decided I will add 500 units of Product A and 200 units of Product B per week to my starts plan. How will this impact my factory cycle time by product, by week?

Dynamic Model Options:

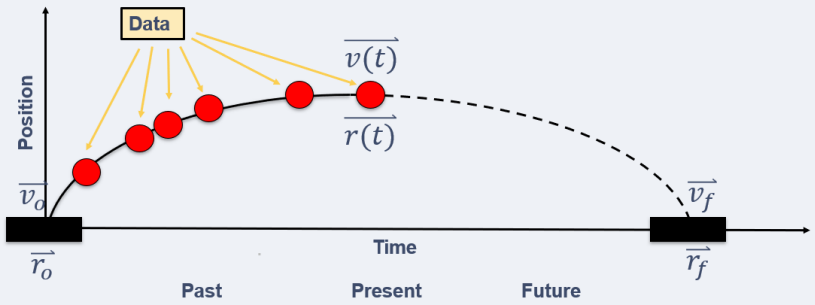
1. **Queueing Models**
2. **Discrete Event Simulation**
3. **Fluid Network Models**
4. **Historical Statistics**
5. **Other**



Queueing Models

The Newtonian Dream

$$v = at + v_0 \quad [1]$$
$$r = r_0 + v_0 t + \tfrac{1}{2} a t^2 \quad [2]$$
$$r = r_0 + \tfrac{1}{2}(v + v_0)\, t \quad [3]$$
$$v^2 = v_0^2 + 2a\,(r - r_0) \quad [4]$$
$$r = r_0 + vt - \tfrac{1}{2} a t^2 \quad [5]$$



THE SINGLE SERVER QUEUE IN HEAVY TRAFFIC

By J. F. C. KINGMAN

Communicated by M. F. Atiyah

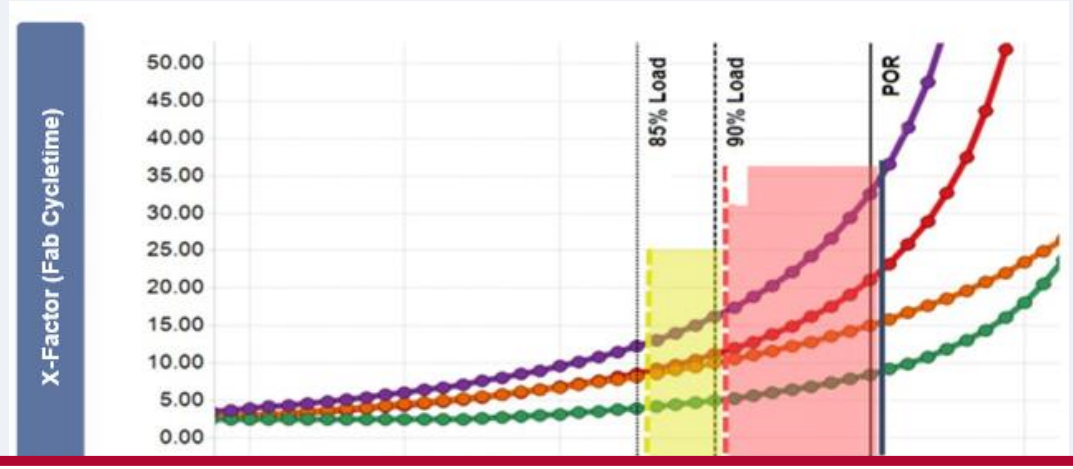
Received 23 November 1960

$$L = \lambda W$$

$$CT = V \cdot U \cdot T$$

$$\mathbb{E}(W_q) \approx \left(\frac{\rho}{1-\rho}\right)\left(\frac{c_a^2 + c_s^2}{2}\right)\tau$$

$$L = \rho + \frac{\rho^2 + \lambda^2 \operatorname{Var}(S)}{2(1-\rho)}$$



Goal: Find equations that describe cycle time as a function of measurable and computable parameters.

Queueing Models

Tuesday 2 pm

Saturday 12 pm

How much longer will my grocery trip take on Saturday versus Tuesday?

Can We Formalize Our Intuition?

$$f(num\ cars) = \int_{?}^{?} \frac{?}{(?-?)^{?}} *?\,?\,d?$$

Shopping Time!

What we would like to find

Incidentally, I would love it if a certain warehouse commercialized this idea…

Queueing Models

Cycle Time of Lot at Operation = Waiting (Queue) + Loading + Processing + Unloading + Hold + Transportation

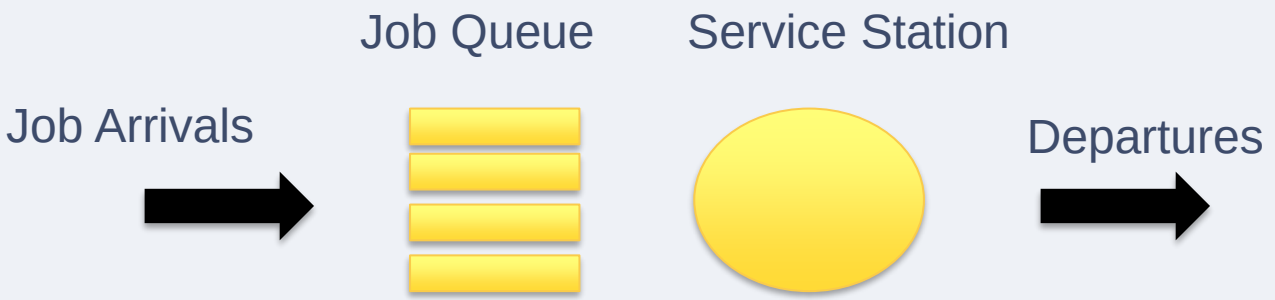
Some assumptions:

- Loading and unloading times are unavoidable and mostly fixed
- Ignore hold time
- Ignore transportation time
- Process time is mostly fixed and well known

Conclusion:

The waiting (queue) time is the single most impactful and variable contributor to overall lot cycle time.

Queueing Models



Given:

$r_a = rate\ of\ arrivals\ in\ jobs\ per\ unit\ time$
$c_a = arrival\ CV$
$m = number\ of\ parallel\ machines\ at\ station$
$b = buffer\ size\ (\max jobs\ in\ system)$
$t_e = mean\ effective\ process\ time$
$c_e = CV\ of\ effective\ process\ time$

We can study:

$\boldsymbol{CT_q = Expected\ wait\ time\ spent\ in\ queue}$
$WIP = average\ WIP\ level\ at\ station$

Some complexity:
Number of tools, buffer size, arrival characteristics (individual or batch), job times, etc.

Queueing Models

Under certain assumptions :

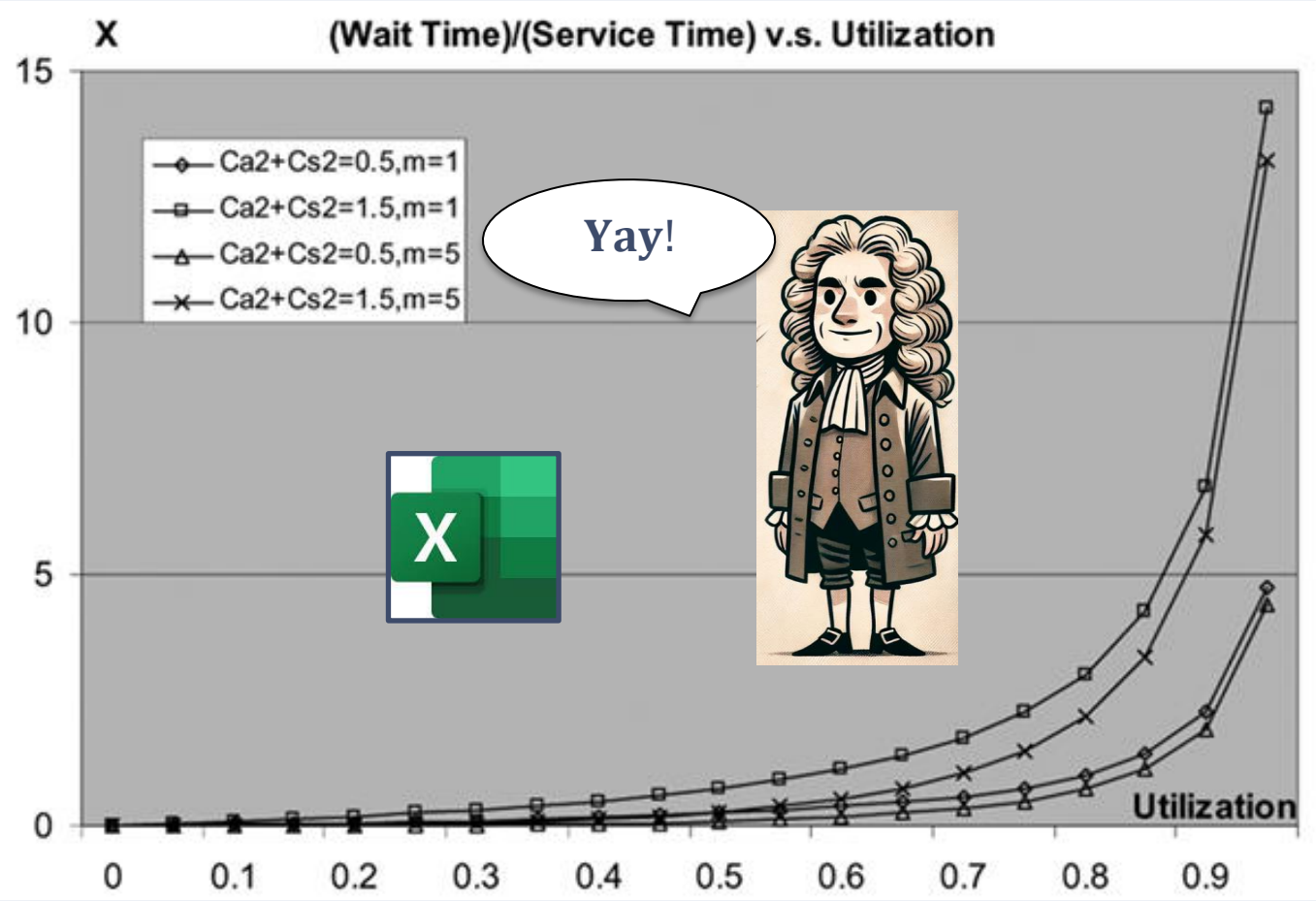
Can be computed **Can be measured**

$$\mathbb{E}(W_q) \approx \left(\frac{\rho}{1-\rho}\right)\left(\frac{c_a^2 + c_s^2}{2}\right)\tau$$

Analytic form

U V T

$\rho = Utilization$
$c_s = Service\ Time\ CV$
$c_a = Arrival\ CV$
$\tau = mean\ process\ time$



Shanthikumar, J. G.; Ding, S.; Zhang, M. T. (2007). "Queueing Theory for Semiconductor Manufacturing Systems: A Survey and Open Problems". IEEE Transactions on Automation Science and Engineering. 4 (4): 513

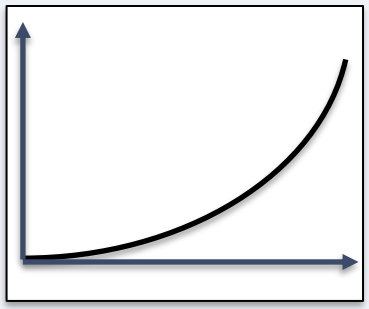
Victory?

Some Difficulties of Queueing Theory:

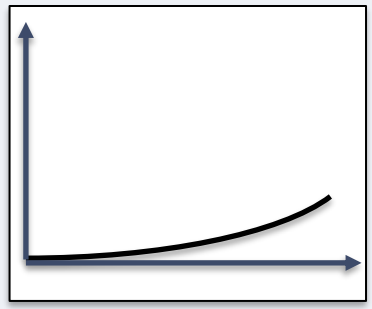
- Exact closed analytic form solutions only exist for a very small subset of queue types.
- Often the assumptions required to achieve an exact solution are not valid for a semiconductor fab work area
- Examples of real complications: tool dedication, waiting for metrology, variable batch sizes, sequence dependent setup changes, holds, rework, split and merge, reentrant processes, variable process times per recipe, etc…
- Approximations can be used in some cases…

Victory?

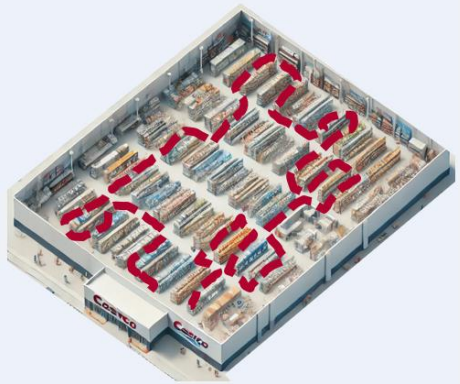
Park my car

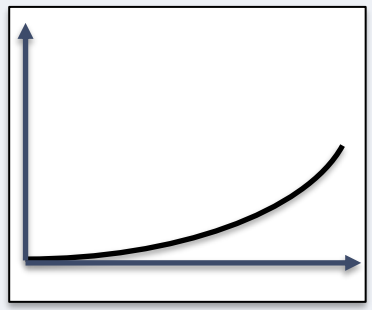


Show my card

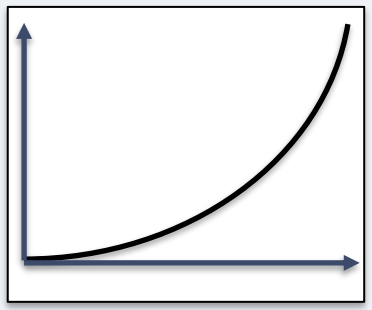


Shop

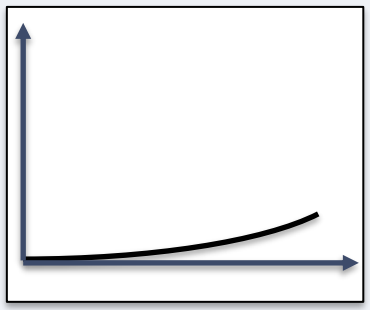




Check out



Receipt check



Networked Queueing Models

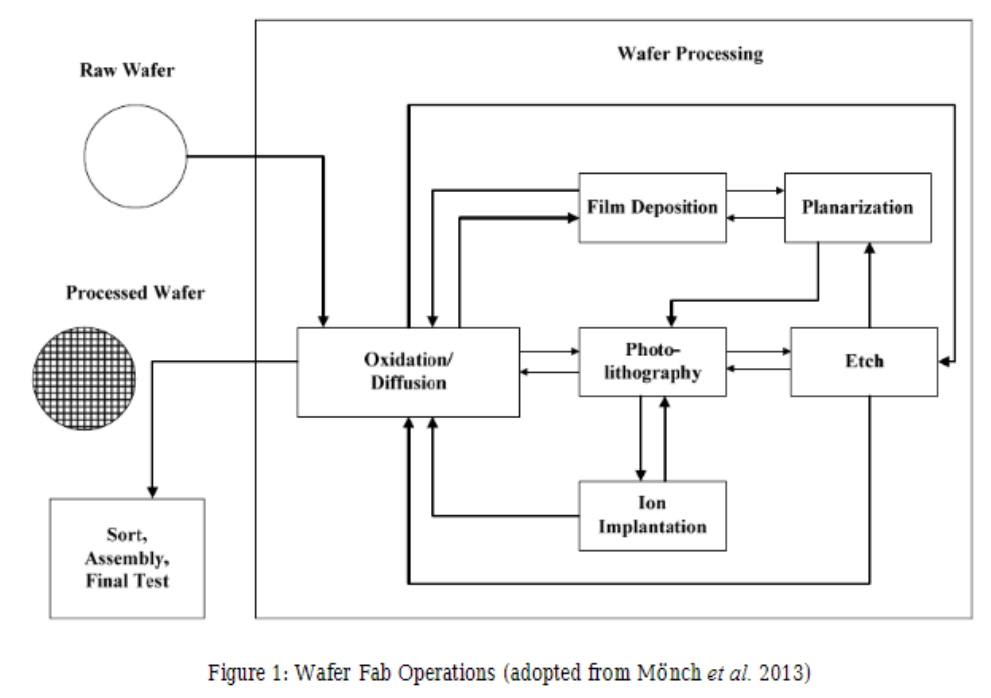
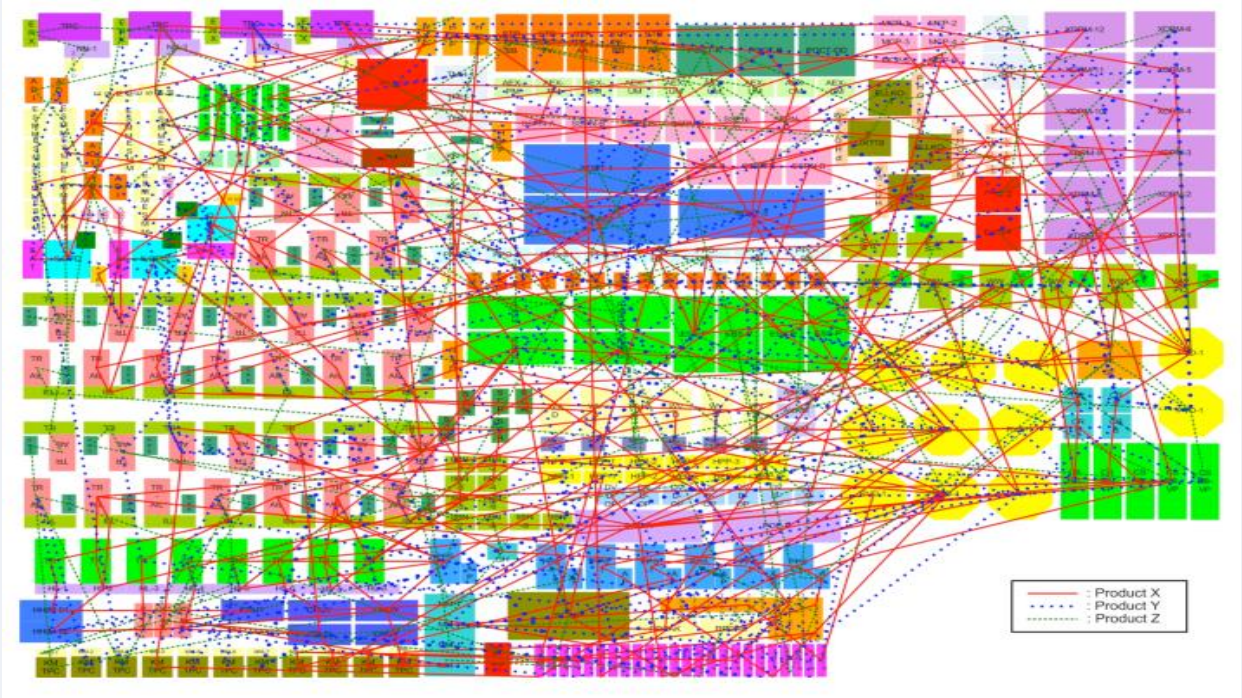
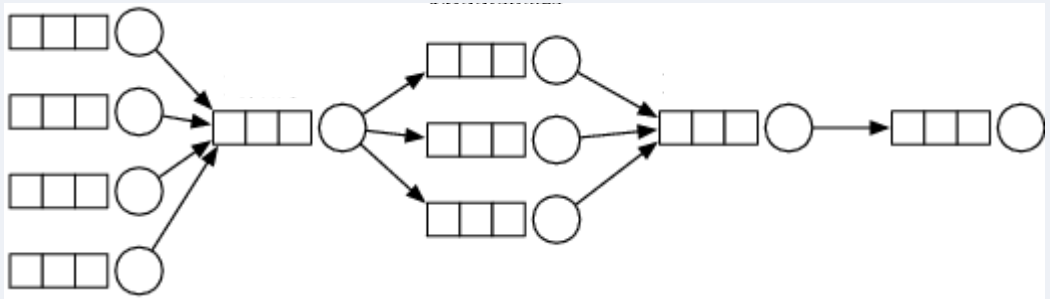


Figure 1: Wafer Fab Operations (adopted from Mönch *et al.* 2013)





$$Total\ Fab\ CT = CT_{q1} + (\lambda_{21} CT_{q21} + \lambda_{22} CT_{q22}) + \cdots$$

J.W. Fowler, L. Monsch, T. Ponsignon. "Discrete-Event Simulation for Semiconductor Wafer Fabrication Facilities: A Tutorial“ October 2015 The International Journal of Industrial Engineering: Theory, Applications and Practice 22(5):661-682

Queueing Models: Final Word

Pros

- Unquestionable value for **qualitative** understanding of factory cycle time behavior
- Can generate managerial insight (See Dr. Jennifer Robinson's CT Class!)
- Fast to work with
- Can be managed in Excel
- Remember that outputs are probabilistic

Cons

- Networked queueing model required to model whole factory behavior, introducing many more complications
- Exact analytic solutions only apply to small subset of workstations, all others must be approximated (or solved numerically)
- Input complexity, modeling limitations
- Mostly applicable to steady-state problems only (not transient)

Discrete Event Simulation

“Discrete-event simulation is a well-established and rather successful method in some semiconductor companies, while other companies do not use simulation at all.”

General Idea (Gross Simplification)

- Forget the project of finding closed analytic form equations of the future
- Instead, build a simulation model by defining discrete events that happen sequentially and can interact/interrupt
- Stochasticity can be comprehended
- Run the model and see what happens!
- Run it again and see what happens! (confidence intervals)

J.W. Fowler, L. Monsch, T. Ponsignon. "Discrete-Event Simulation for Semiconductor Wafer Fabrication Facilities: A Tutorial“ October 2015 The International Journal of Industrial Engineering: Theory, Applications and Practice 22(5):661-682

Discrete Event Simulation

Peeking at The Internals

Lot

- Waits to process at step 1
- Processes at step 1
- Waits to process at step 2
- Processes at step 2
- …

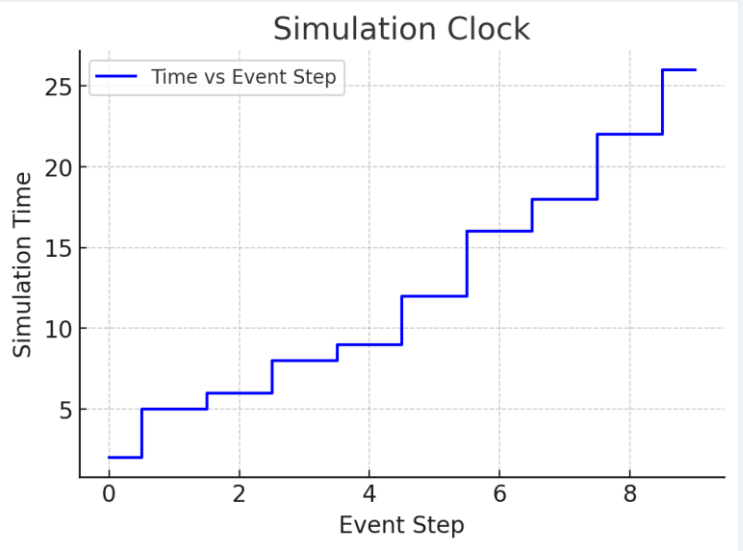
Tool

- Choose and load lot 1
- Choose and load lot 2
- Lot 1 finishes, choose/load lot 3
- Lot 2 finishes, choose/load lot 4
- …

Current Event List

Future Event List

Delay List



Event lists dynamically updated as simulation proceeds

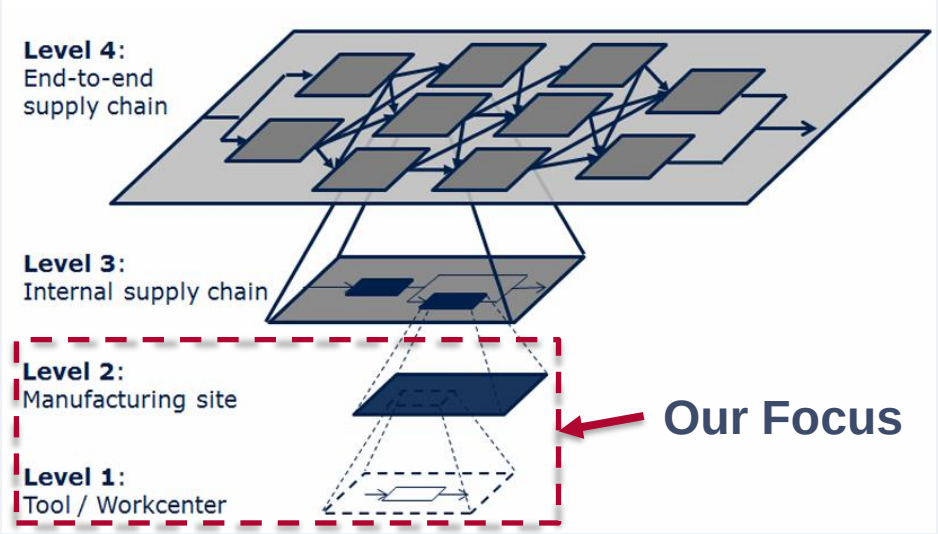
Discrete Event Simulation

Advantages

- Excels in capturing non-linear behavior, interdependent processes (complexity emerges from simply defined interactions)
- Flexibility and level of detail - can be extremely fine (e.g. wafer-level modeling on cluster tools) to high level (e.g. factory output over years)
- Able to study all types of dynamic questions – transient and steady state, what-if scenarios and trends
- Conceptually simpler than the math involved in more analytic approaches (but heavy on modeling & statistics)
- Several domain-specific off-the-shelf options for the core engine
- Can be paired with an optimization routine

Disadvantages

- Slow
- Time and resource/skill requirements
- Maintaining quality input data
- Confounding interactions of high number of dynamic elements
- Complexity of interpretation
- Earning trust of stakeholders

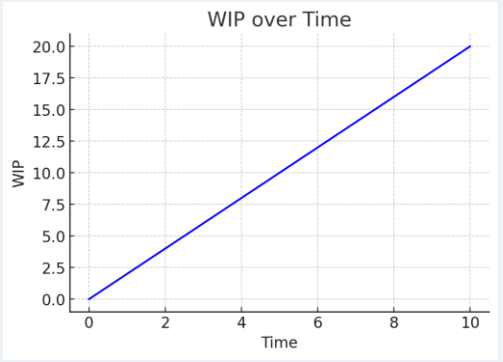


J.W. Fowler, L. Monsch, T. Ponsignon. "Discrete-Event Simulation for Semiconductor Wafer Fabrication Facilities: A Tutorial“ October 2015 The International Journal of Industrial Engineering: Theory, Applications and Practice 22(5):661-682

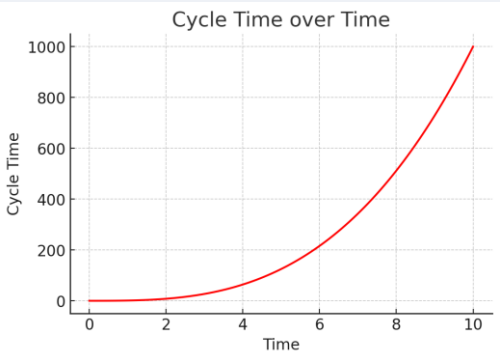
Discrete Event Simulation

Verification and Validation

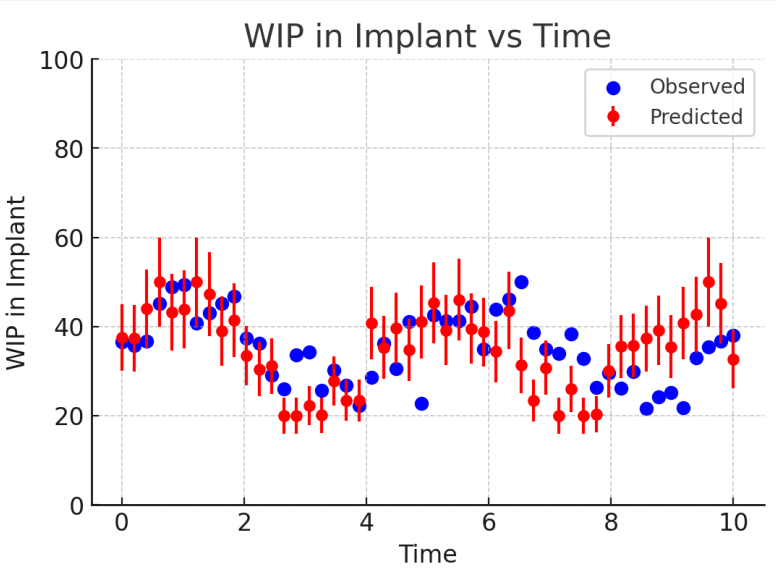
Verification:
Check that a model generates a correct output given a known input



DES Model



Validation:
Compare the output of the model to reality.



Discrete Event Simulation

Elaborating on challenges

"Why Is It So Hard to Build and Validate Discrete Event Simulation Models of Manufacturing Facilities"

- Nothing in a factory is truly static (even tool numbers)
- bagging equipment
- labor models
- labor qualifications
- specialist support
- Ever-changing product mix
- product maturity dynamics
- quality control plans
- Engineering activities
- Rework
- required level of detail
- run time
- software engineering required to modularize/deploy
- Validation
- gaining trust
- time to develop
- input data requirements
- model specialist education
- shift behavior differentials
- poor data
- incomplete data
- no two tools the same
- customer and stakeholder education
- counter-intuitive and unpopular findings
- confounding factors
- …

Fischbein, S.A., Yellig, E. (2011). Why Is It So Hard to Build and Validate Discrete Event Simulation Models of Manufacturing Facilities?. In: Kempf, K., Keskinocak, P., Uzsoy, R. (eds) Planning Production and Inventories in the Extended Enterprise. International Series in Operations Research & Management Science, vol 152. Springer, New York, NY.

Discrete Event Simulation: Final Words (for Operations Contexts)

Final Words

- Most powerful and detailed tool for exploring dynamic transient models
- Must acknowledge relatively low industry adoption (higher in bigger companies)
- Scope of time/labor/effort/skills – huge range!
  - Usage plan: one-time study or regular/deployed
  - Model length: short or long term
  - Model detail: higher or lower
  - Validation/verification: Able to be automated or mostly manual
  - Input data: automatically ingested or manually inputted
- Will revisit later in AI-section

A Word On Other Methods

1. Statistical Averaging of History

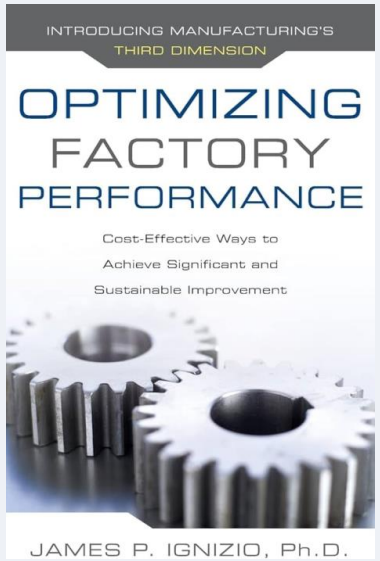
- Always possible to leverage past data to estimate future (processing times, cycle times, etc.)
- Built-in feature of a Digital Twin
- **Often sufficient**

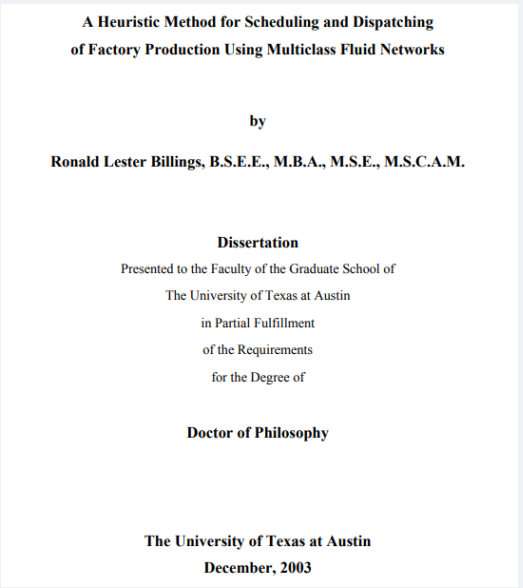
2. Fluid Network Models

- Dynamic steady-state
- Models WIP as continuous flow (no lot-level detail)
- Simpler inputs than queueing models, similar outputs
- Not widely used in industry (to my knowledge)
- I would love to learn more!!

3. Machine Learning

Later!





A Heuristic Method for Scheduling and Dispatching
of Factory Production Using Multiclass Fluid Networks

by

Ronald Lester Billings, B.S.E.E., M.B.A., M.S.E., M.S.C.A.M.

Dissertation

Presented to the Faculty of the Graduate School of
The University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

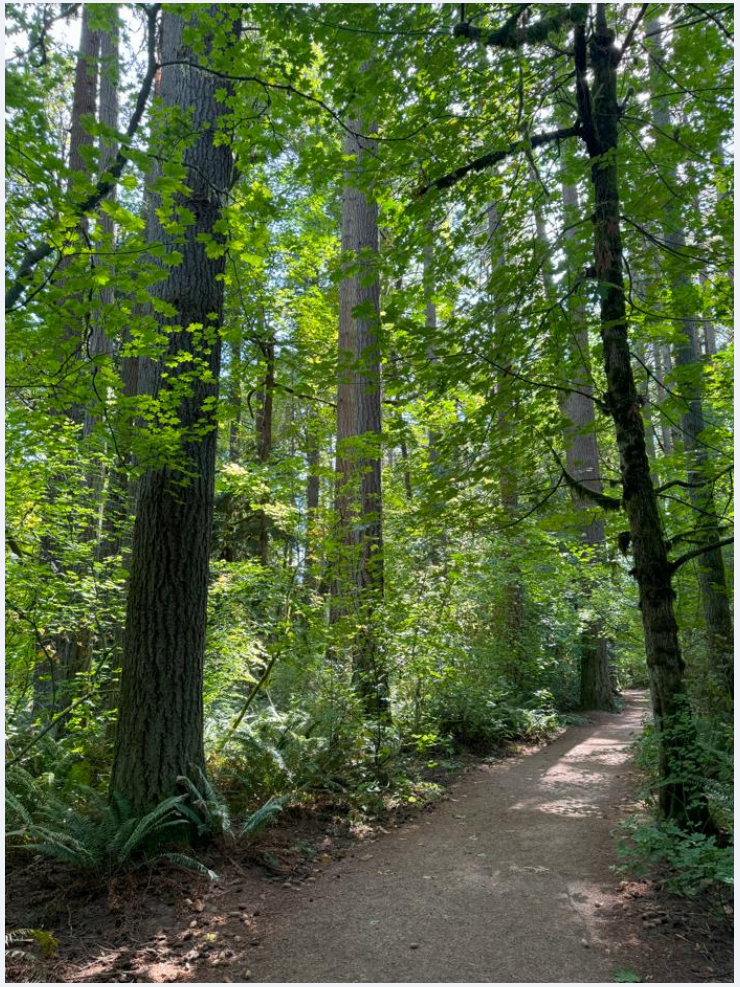
The University of Texas at Austin
December, 2003

*"…No, I don't use Muddle's simulation software… I'm using a simulation approach based on **fluid network modeling**, a type of continuous simulation. It's enormously faster than Muddle's discrete event-based simulation, and it's better suited to my work…I don't have the level of detail in my models that Muddle's simulation group does, but I can get what I want in a fraction of the time – and cost."*

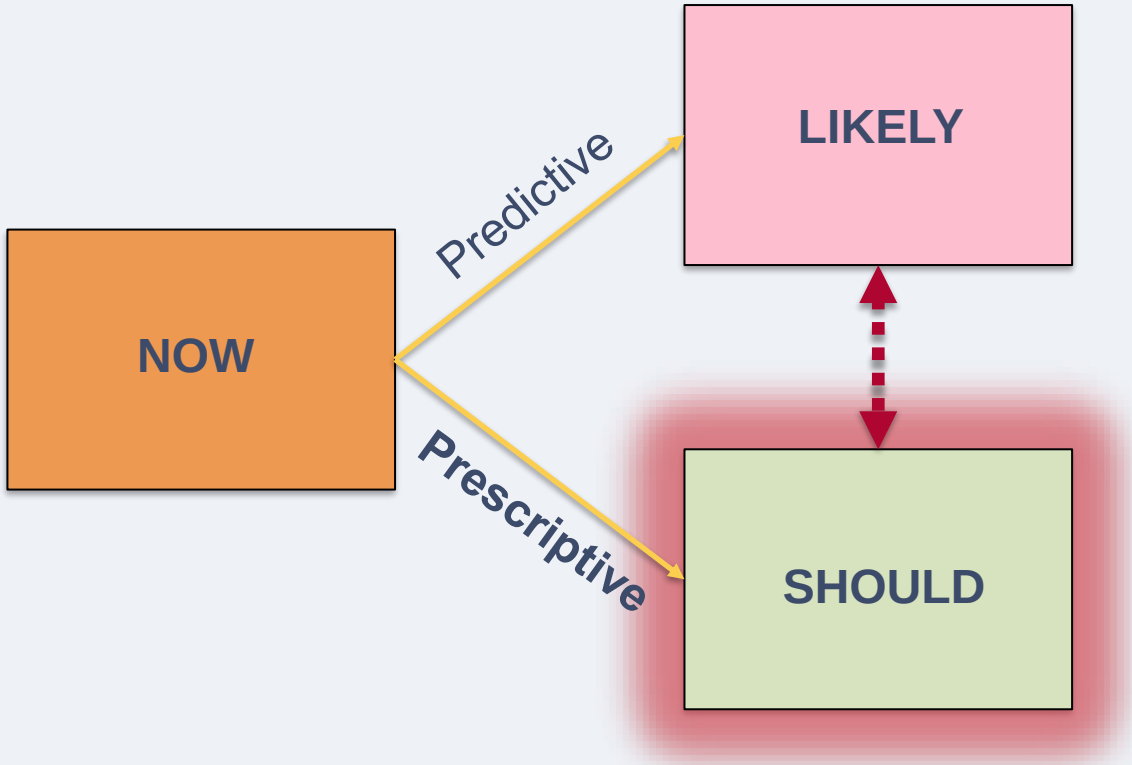
– Dr. Winston Smith

Talk Outline

I. **Stochastic Environments**
   a) Motivating the problem
   b) Psychology of uncertainty
   c) Goals for talk
II. Factory Future Forecasting
   a) Predictive models
      - Static
      - Dynamic
   b) Prescriptive models
      - General optimization principles
      - Factory scheduling
III. AI and Factory Future Forecasting
   a) Machine learning
   b) Generative AI
IV. Integrating Best Practices
V. Conclusions



Two Main Branches of Future Forecasting



Factory Problems of Finding “The Best”

Planning | Equipment | Operations | Maintenance | Quality/Yield

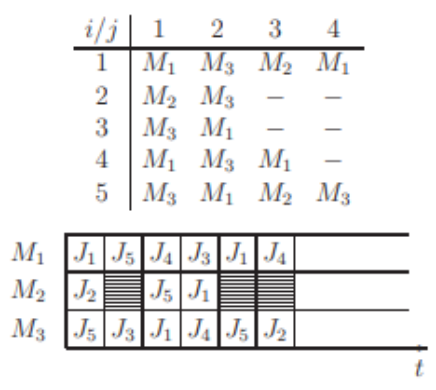
Scheduling

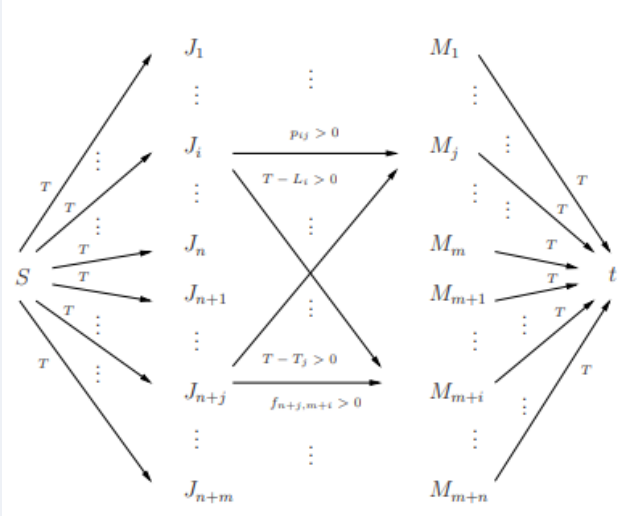
Scheduling is a *Worthy* Problem

The ultimate performance of a factory is nothing more than the sum of the individual decisions made minute-to-minute.

Our software makes those decisions.

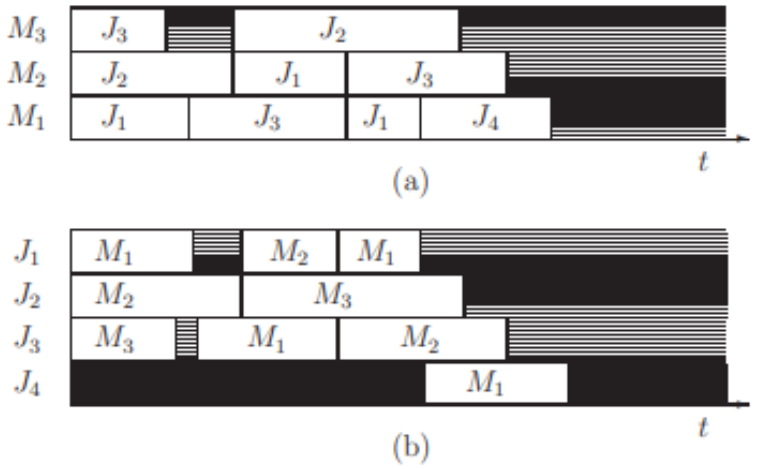
This is an enormous responsibility.



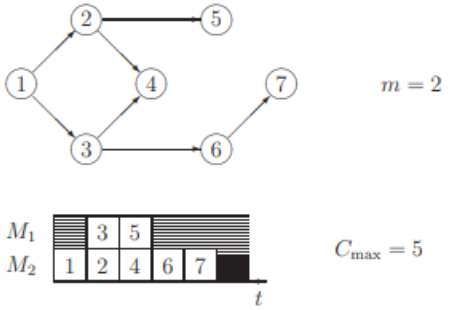


Job Shop Scheduling

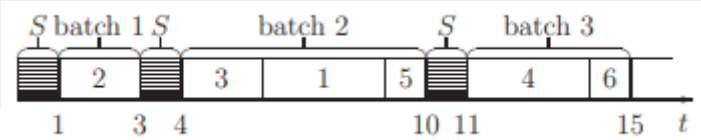
- **Complexity: Scheduling is NP-Hard**
  - Practically, this means that calculating a schedule for a typical situation from a real semiconductor factory CANNOT be guaranteed to finish in a reasonable amount of time.



- **Operations Research literature is full of studies on how to solve the job-shop scheduling problem in a variety of manifestations.**



- **Simple algorithm: Generate every possible schedule, choose the best one.**
- **Simple algorithm impractical because of the number of possible permutations.**



Combinatorics and Planning Scale

How many different ways to choose 10 lots from 100 possible?

Unique unordered combinations of 10 lots: $C(100,10) = \frac{100!}{10!\,(100-10)!} = \mathbf{1.73 * 10^{13}}$

How many different ways to arrange 10 lots from 100 possible on 1 tool?

Unique *ordered* permutations of 10 lots: $\mathrm{P}(100,10) = \frac{100!}{(100-10)!} = \mathbf{6.28 * 10^{19}}$

Assuming a single CPU cycle could generate a permutation (not realistic):

$$\sim 10^{19}\ permutations \times \frac{\sim 10^{-9} seconds}{permutation} = 10^{10}\ seconds = \mathbf{317.1\ years}$$

Bottom line: Combinatorial Explosion.
You cannot calculate (let alone evaluate) every possible schedule.

(Also: Quantum computing *won't* save us.)

At the Same Time… Don’t Throw Out Intuition

Our intuitions about combinatorial optimization problems are often pretty good…
(unlike with stochastic non-linear systems)

Chimpanzee Spatial Memory Organization

Abstract. *Juvenile chimpanzees, carried around an outdoor field and shown up to 18 randomly placed hidden foods, remembered most of these hiding places and the type of food that was in each. Their search pattern approximated an optimum routing, and they rarely rechecked a place they had already emptied of food.*

- Study: chimpanzee driven around an enclosure
- Watched researchers place food in 18 random locations
- Returned to enclosure entrance
- Cage opened, chimpanzee free to retrieve food

- **Each time, chimpanzee finds close to optimal solution**
- **Result is independent of researchers’ placement path!!**

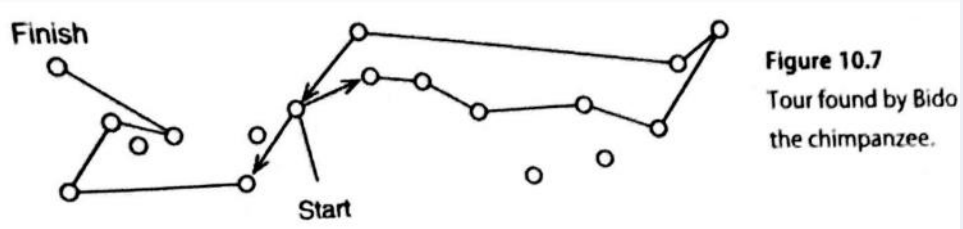


Figure 10.7
Tour found by Bido the chimpanzee.

INFICON

E.W. Menzel, “Chimpanzee Spatial Memory Organization”, 30 Nov 1973, Vol 182, Issue 4115, pp. 943-945

W. J. Cook, “In Pursuit of the Traveling Salesman: Mathematics at the Limits of Computation“, Princeton University Press, 2021, Chapter 10.

At the Same Time… Don’t Throw Out Intuition

Solution Techniques

- Widely established, commercial and in-house solutions
- Legacy technology
- Fast Execution

- State-of-the-Art
- Established commercially and growing quickly
- Some in-house
- Higher computational requirements

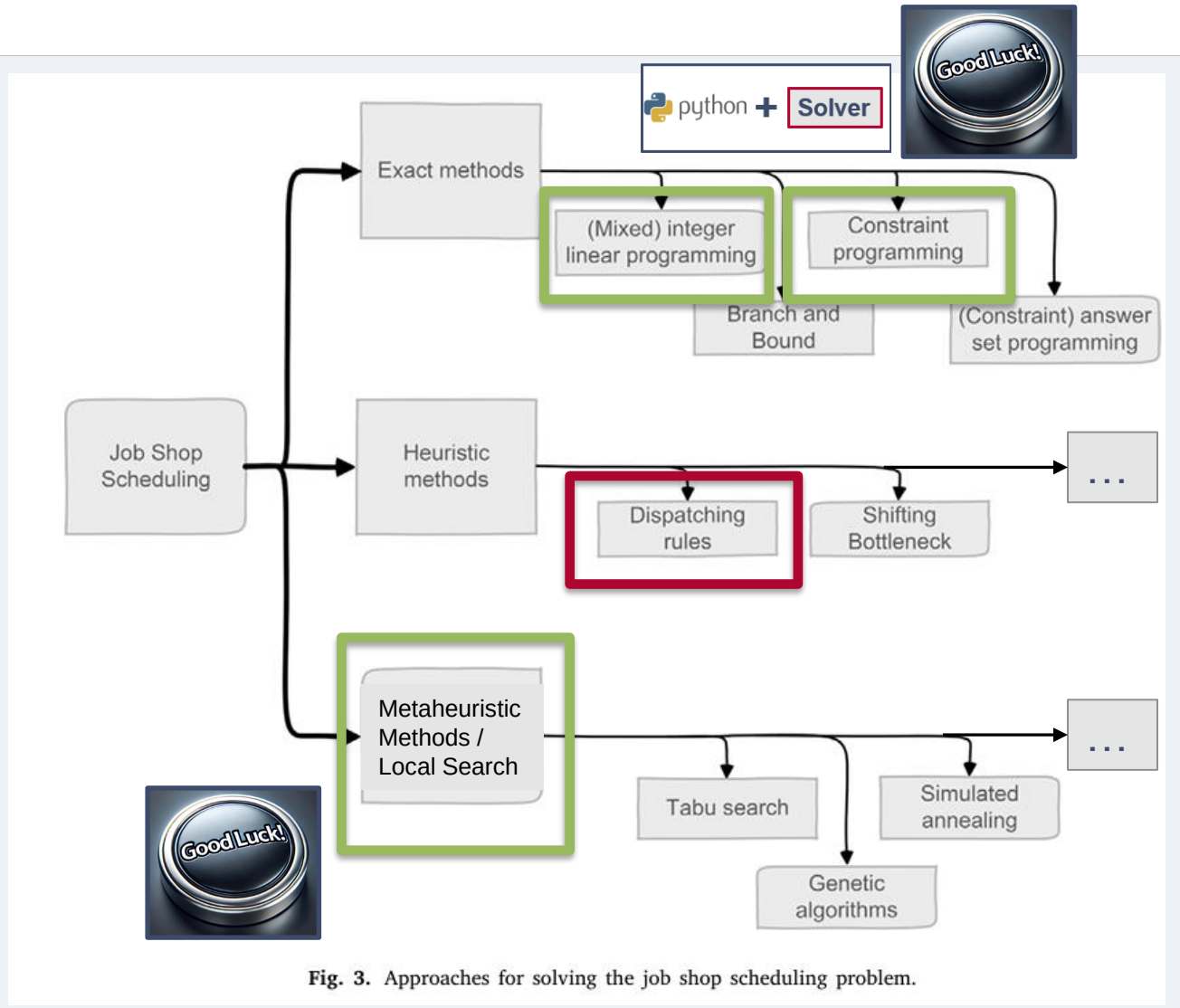


Fig. 3. Approaches for solving the job shop scheduling problem.

Giacomo Da Col, Erich C. Teppan, "Industrial-size job shop scheduling with constraint programming", Operations Research Perspectives, Volume 9, 2022

Important Complication: Time Scale Translation

In a typical Front End context:

Scheduling Horizon **~12-24 Hours**

Planning Horizon **~Months - Years**

Why is this a problem?

Planning statement:
I need to deliver 5000 units of Product A to Customer XYZ by Q4

Operations Perspective:
Ok, what do I need to do in the next shift to make sure we're on track?

Translating Objectives Between Time Scales

Suppose you want to lose 10 pounds in 2 months.

Meal Planner:

- Creates a detailed and delicious 2-day Meal Plan based on **daily calorie target**, items in fridge, weekly shopping list, etc., comprehending all dietary restrictions, preferences, need for variety, costs, etc.
- Calculation is too slow for anything longer than 2 days

~2 Days

Calorie Planner:

- Comes up with a **daily calorie target** based on your weight loss goal
- Calculation is fast enough to plan years of targets

~Months - Years

Translating Objectives Between Time Scales

Suppose you want to lose 10 pounds in 2 months.

Calorie Planner (Assume Linear Model):

- (10 pounds x 2500 calories / pound) / 60 Days = ~416 calories/ day
- Base Expenditure: 2200 calories
- 2200 – 416 = 1783 calories / day

Breakfast

Lunch

Dinner

Day 1 …

Will I lose the weight I want?

Translating Objectives Between Time Scales

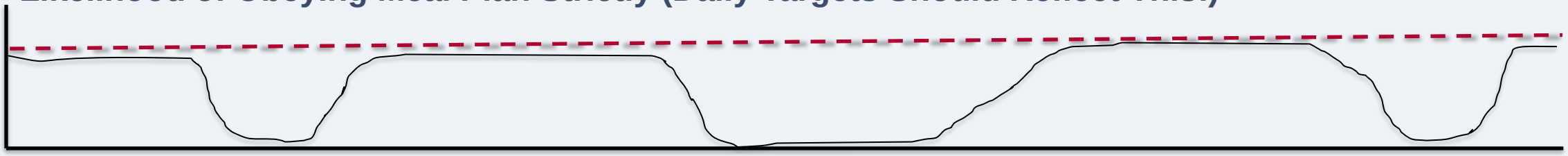
Maybe!... But probably not.

Work trip:
Conference

Family Vacation:
Alaskan Cruise!

Work Trip:
All Hands

Likelihood of Obeying Meal Plan Strictly (Daily Targets Should Reflect This!)



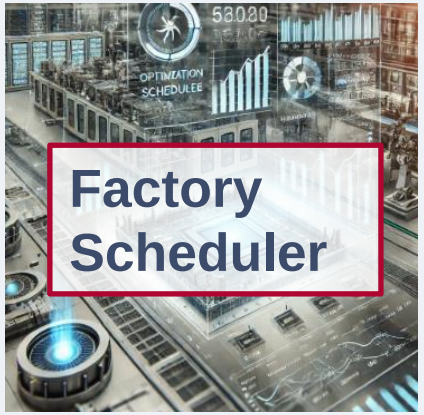
Time [t]

Translating Objectives Between Time Scales

Factories face the same challenge:

Long Term Wisdom

Short Term Action

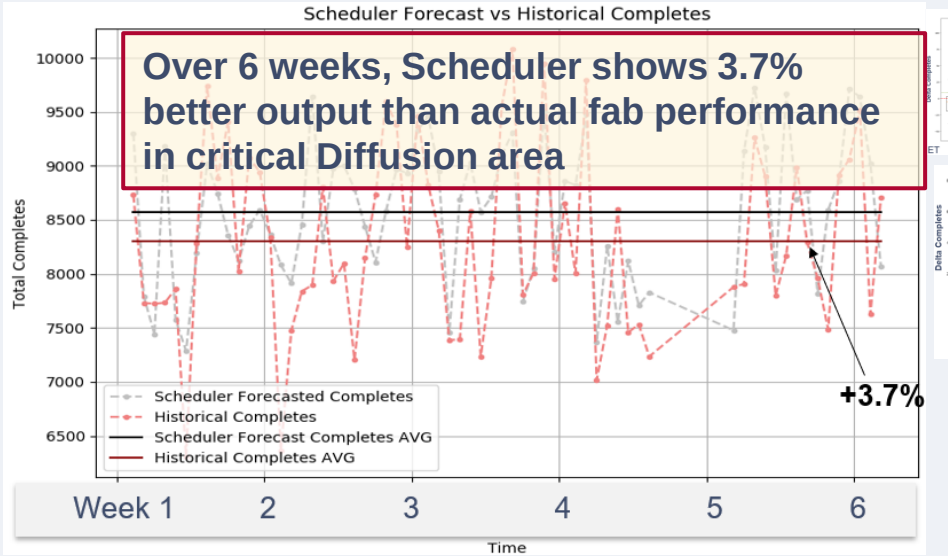


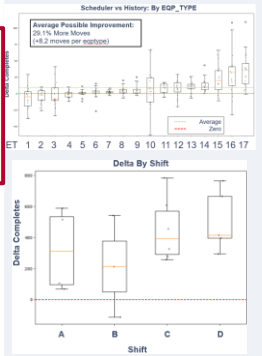
Overall Factory Performance

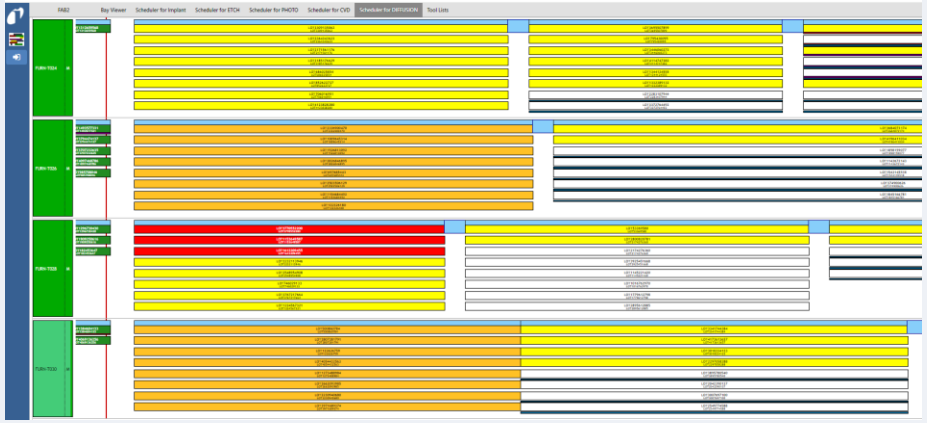
Quality of Scheduler “optimization” secondary to quality of guiding intelligence.

Deployment Best Practice: Pre-Adoption Verification and Validation

1. **Predicted output improvement**
   - By area
   - By shift
   - By equipment type
   - …
2. **Predicted OTD improvement**
3. **Predicted setup change improvements**
4. **Predicted queue timer improvements**
5. **Fundamental batching quality (Diffusion)**
6. **Etc…**

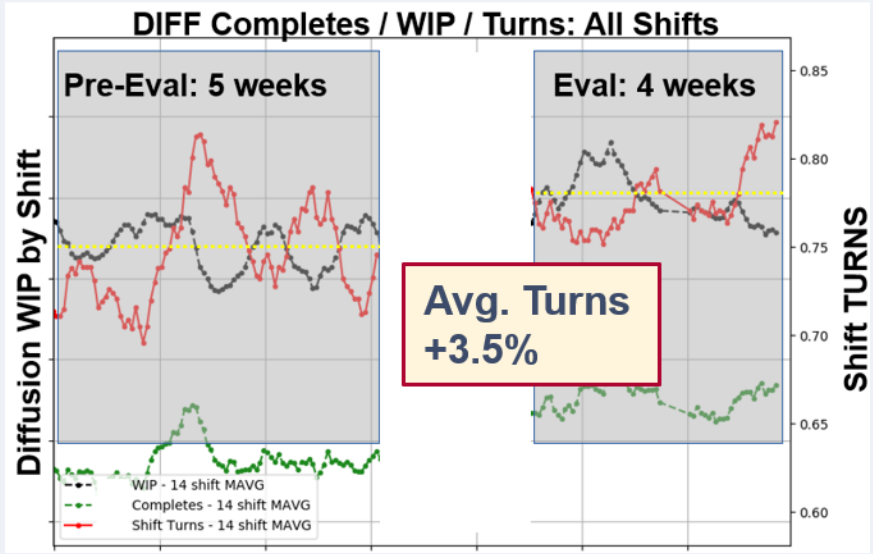


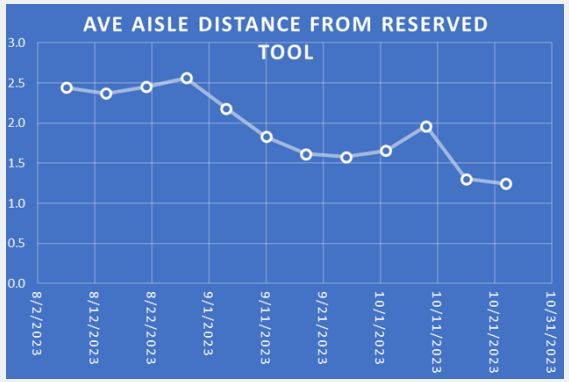


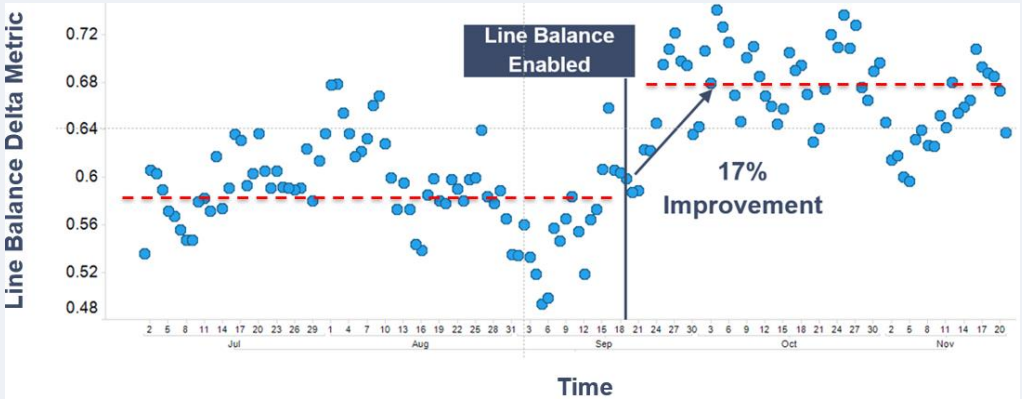


Deployment Best Practice: Post-Adoption Results

1. Output improvement – pre- and post-
2. OTD improvement – pre- and post-
3. Line balance – pre- and post-
4. Target completion – pre- and post-
5. Reduction of loss states (e.g. Idle W/WIP, Idle No WIP)
6. AMHS metrics
7. Etc…





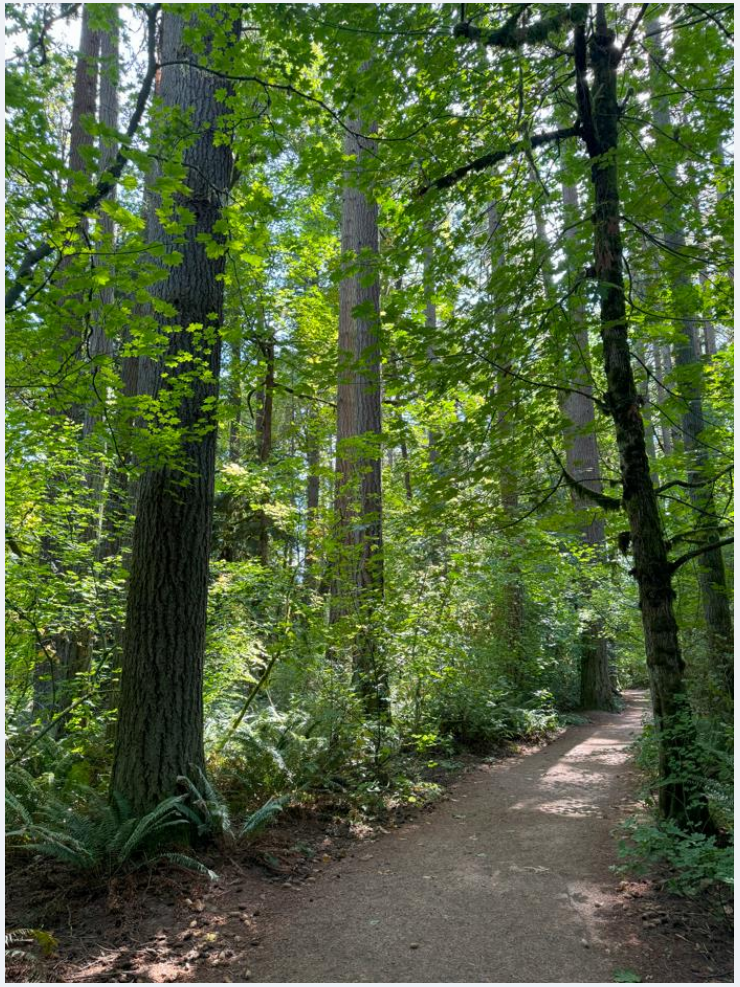


Scheduling: Highly Impactful and High ROI

1. A scheduler is the last-mile consolidator of all upstream optimization
2. While difficult to engineer, “easy” to productize and deploy (not a science experiment)
3. Must be guided by long-range intelligence (provided by predictive models), robustness over brittle/contingent “optimality”
4. Once in place, continuous possibilities for upstream improvement
5. Can do both validation and verification prior to rollout (shift-level forecasting means quick learning turns, as opposed to a one-year forecast!)
6. Easy to *quantify* ROI, even before rollout
7. Required for modernization, certainty of tool and time

Talk Outline

I. **Stochastic Environments**
   a) Motivating the problem
   b) Psychology of uncertainty
   c) Goals for talk
II. **Factory Future Forecasting**
   a) Predictive models
      - Static
      - Dynamic
   b) Prescriptive models
      - General optimization principles
      - Factory scheduling
III. **AI and Factory Future Forecasting**
   a) Machine learning
   b) Generative AI
IV. **Integrating Best Practices**
V. **Conclusions**



Recap: Factory Future Forecasting

Digital Twin

Predictive vs. **Prescriptive**
Qualitative vs. Quantitative
Deterministic vs. Non-Deterministic
Wafer-, Lot-, or Product-level
Static vs. Dynamic
Transient vs. Steady State
One Time vs. Continuous Use

Long Term – Months to Years

Optimization (starts, mix, etc.) | Statistics/ History | Queueing models | Discrete Event Simulation | Fluid Networks | Other

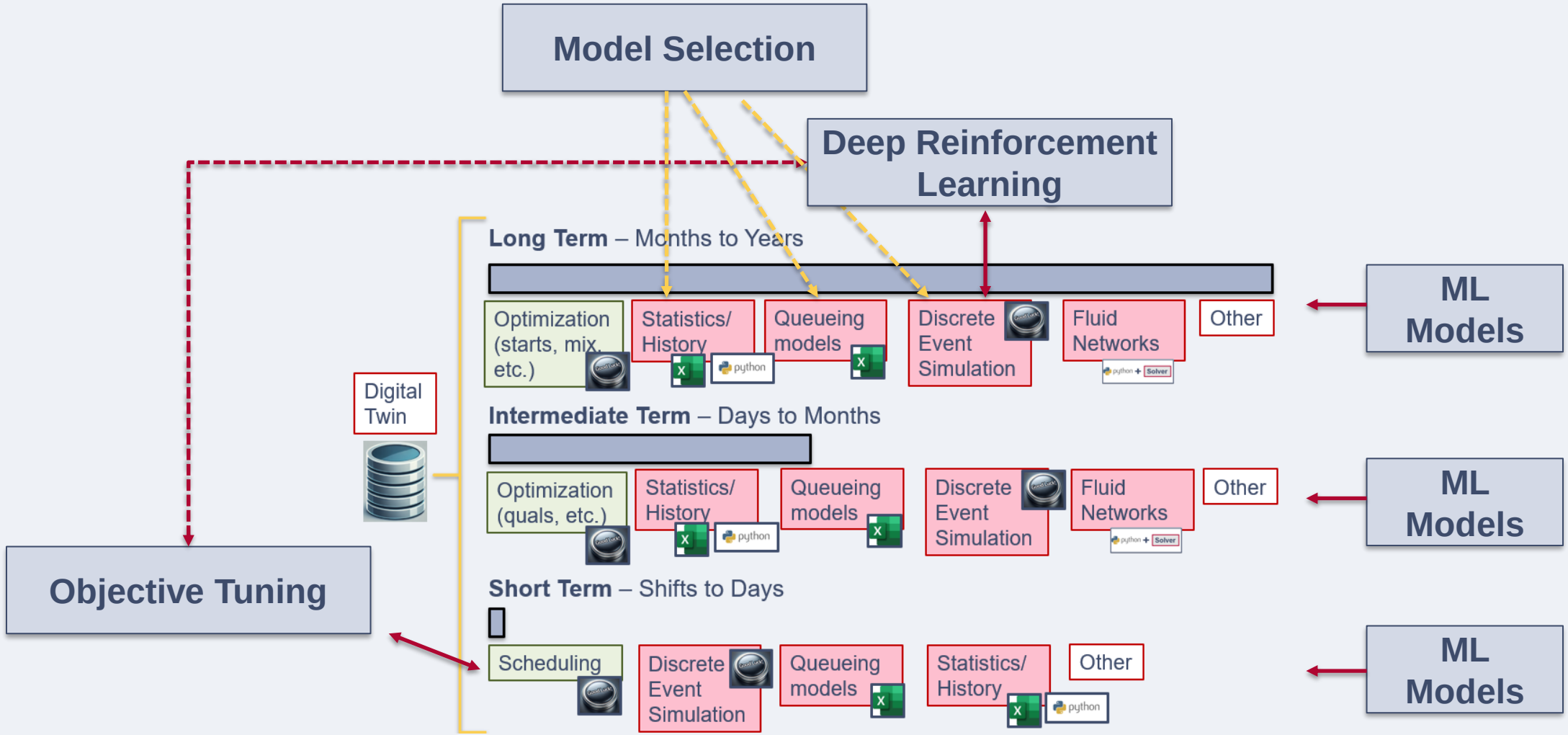
Intermediate Term – Days to Months

Optimization (quals, etc.) | Statistics/ History | Queueing models | Discrete Event Simulation | Fluid Networks | Other

Short Term – Shifts to Days

Scheduling | Discrete Event Simulation | Queueing models | Statistics/ History | Other

Where Does ML Fit?



AI-Assisted Reasoning

Our intuitions about non-linear stochastic systems are generally bad.

I'm considering bagging two of our diffusion furnaces given our factory loading. Will this have any surprising effects I'm not thinking about?

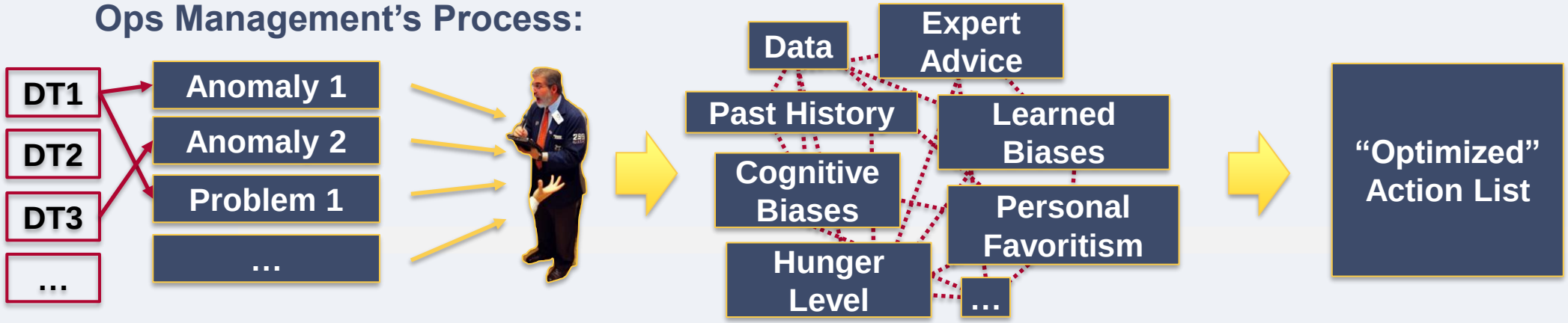
Two of my Photo operators called in sick today. What should I think about doing?

A line carrying etchant from the BCD to my etch tools has burst, taking all the etchers down for ~24 hours. What impacts on the line should I expect, and what can I do to mitigate them?

Anomaly Detection & Impact Management

Impact Management: Intelligence Layer

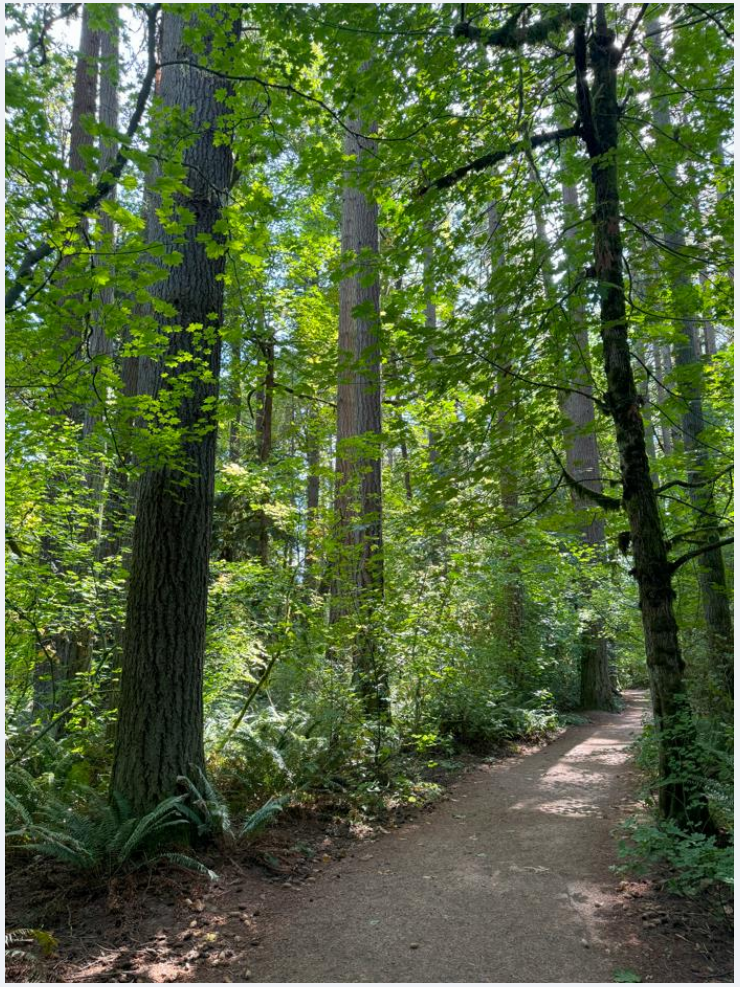
Ops Management's Process:



AI Impact Management Process (aspirational):

Talk Outline

I. **Stochastic Environments**
   a) Motivating the problem
   b) Psychology of uncertainty
   c) Goals for talk
II. **Factory Future Forecasting**
   a) Predictive models
      - Static
      - Dynamic
   b) Prescriptive models
      - General optimization principles
      - Factory scheduling
III. **AI and Factory Future Forecasting**
   a) Machine learning
   b) Generative AI
IV. **Integrating Best Practices**
V. **Conclusions**



Many Methods, Understand Nuances and Purposes

How long does it take to drive from Portland to Denver in general?

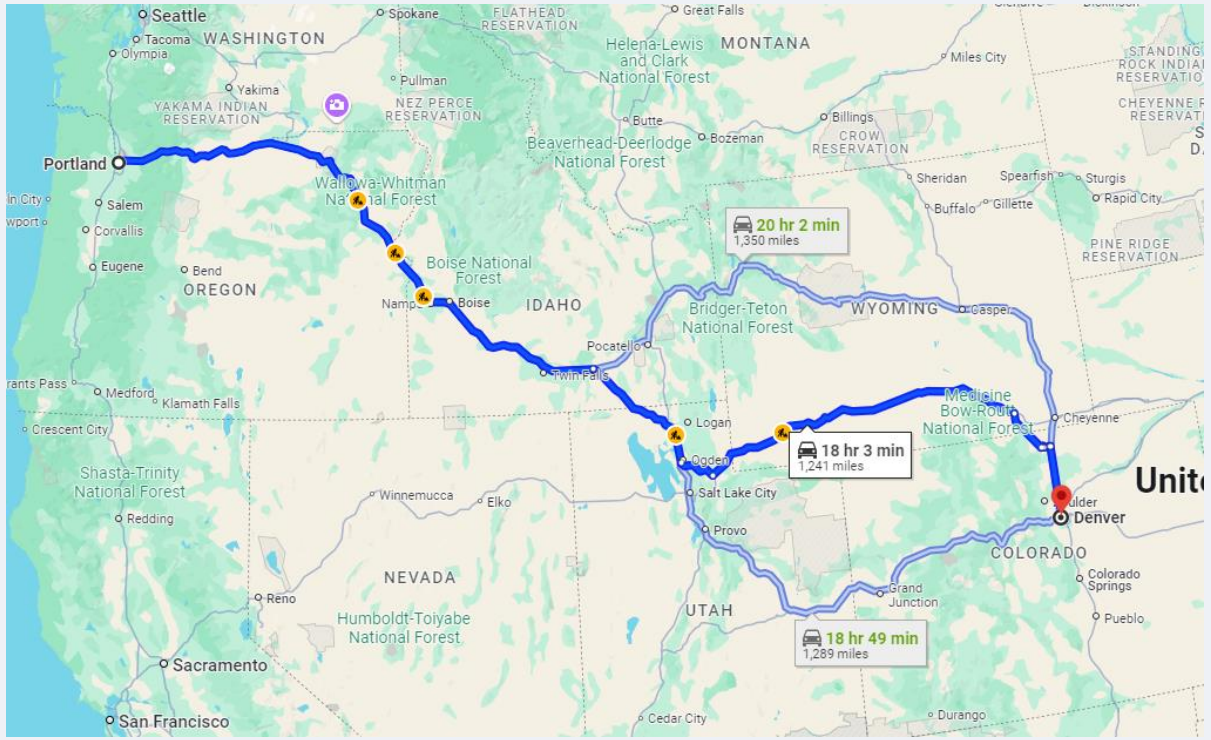
History

How long does it take to drive from Portland to Denver today?

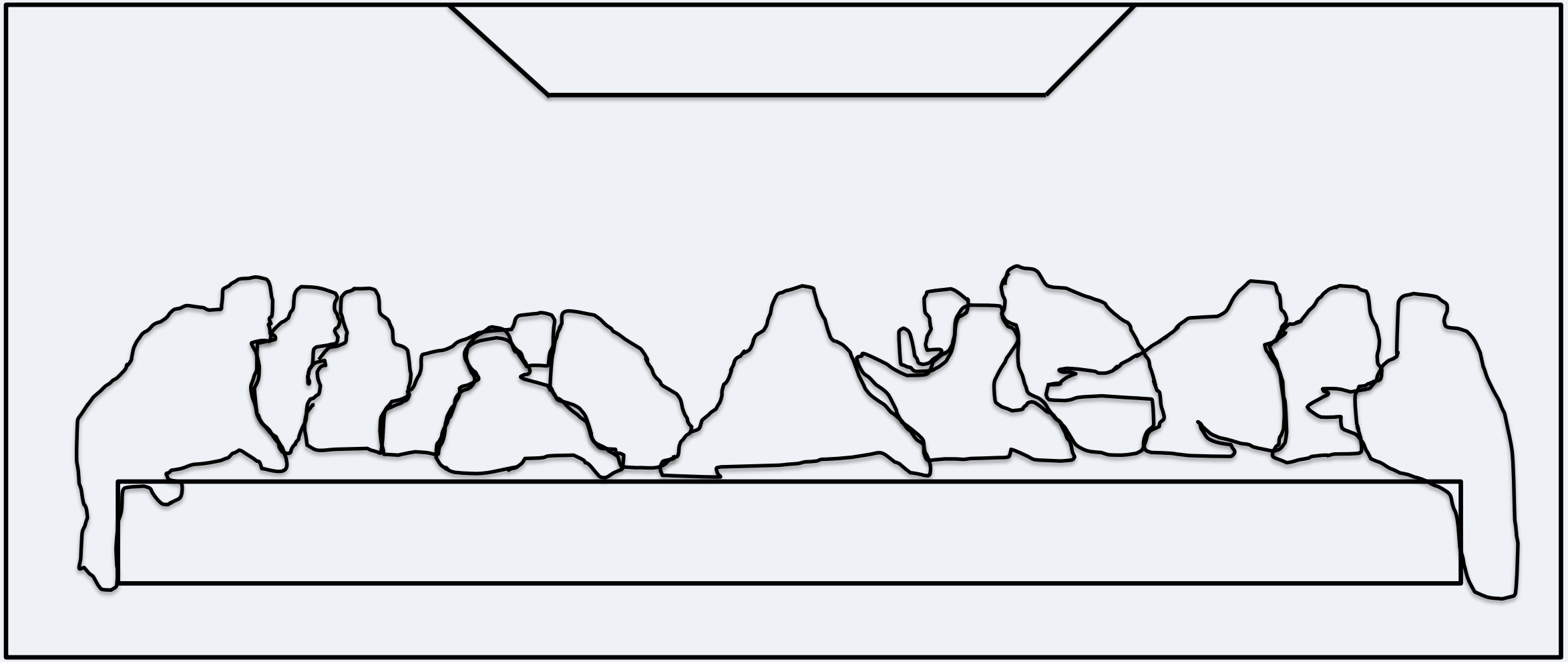
History **Queueing Models** **ML**

How long does it take to drive from Portland to Denver today given that I'm leaving at 8 am and will be going through Boise at 3 pm when there is a Taylor Swift concert starting?

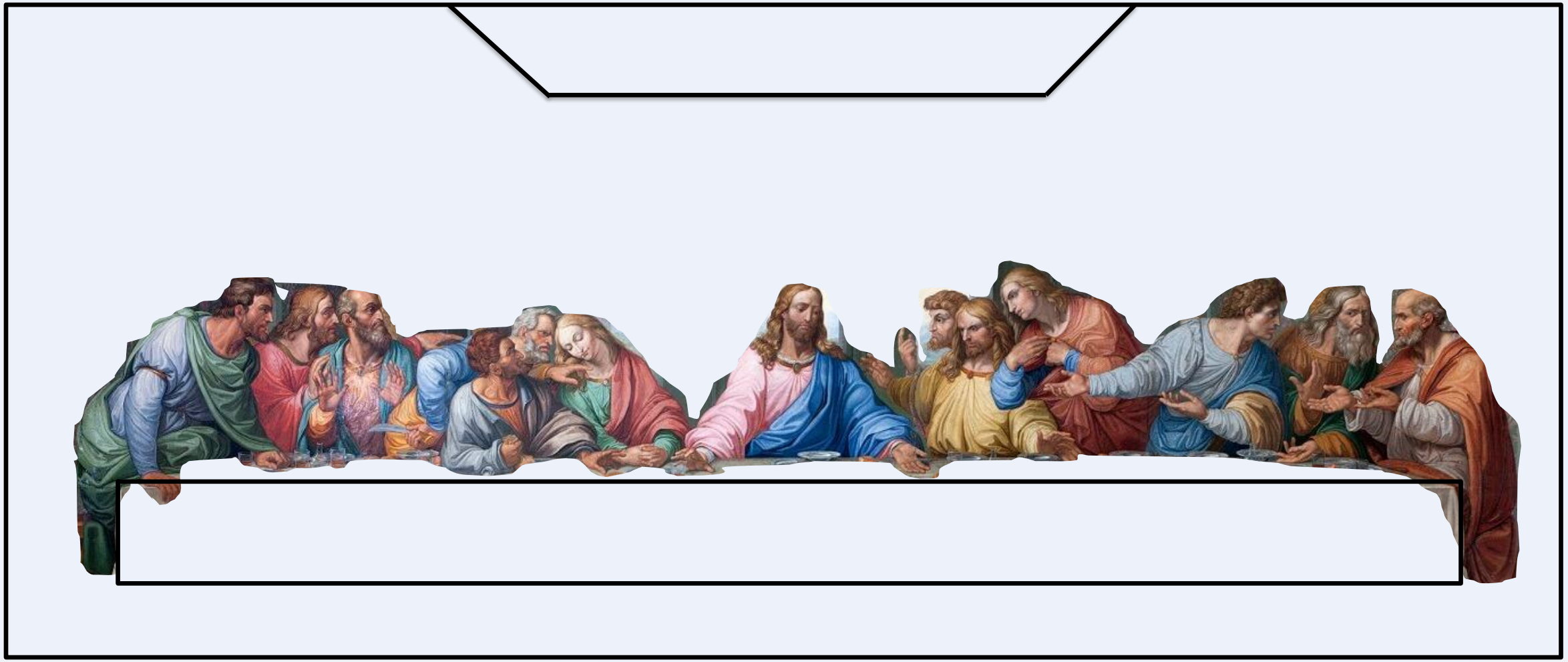
History **Queueing Models** **DES** **GEN AI** **ML** **etc.**



Forecasting + Optimization in Steps



Forecasting + Optimization in Steps



Forecasting + Optimization in Steps

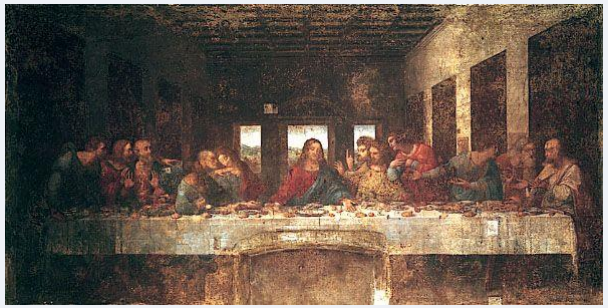
Always Validate!

The GREAT thing about Future Forecasting is that it can be compared to what actually happens!

Predictive Models:
Are they accurate? Are they useful?

Prescriptive Models:
Are they accurate? Are they being *used*?

Unvalidated

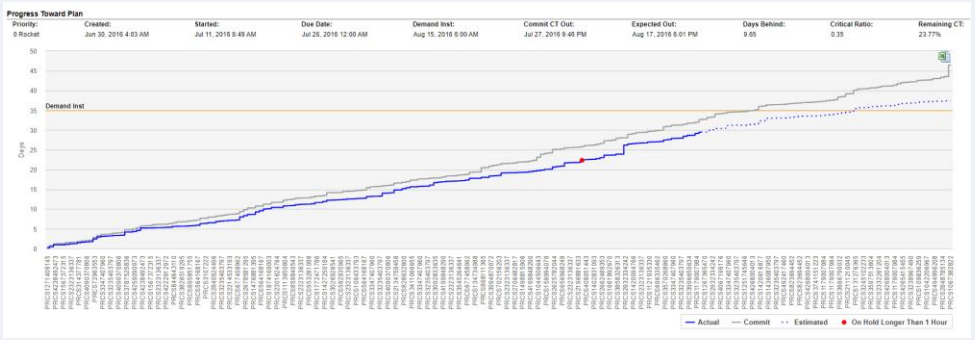


Unused

Speed and ROI: Leverage the Commercial Space

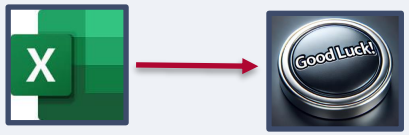
INFICON Digital Twin:

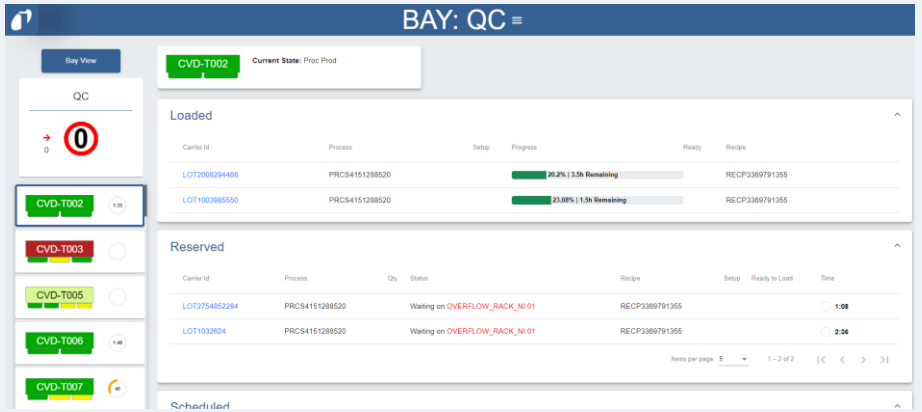
- Static and dynamic models
- Historical summarization
- Queueing models
- Scheduling
- Goal planning across scales

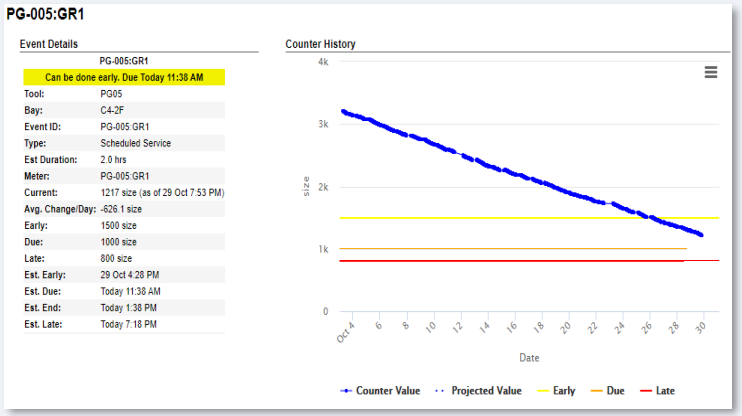


In Development:

- AI/ML-models
- Integrated Discrete Event Simulation

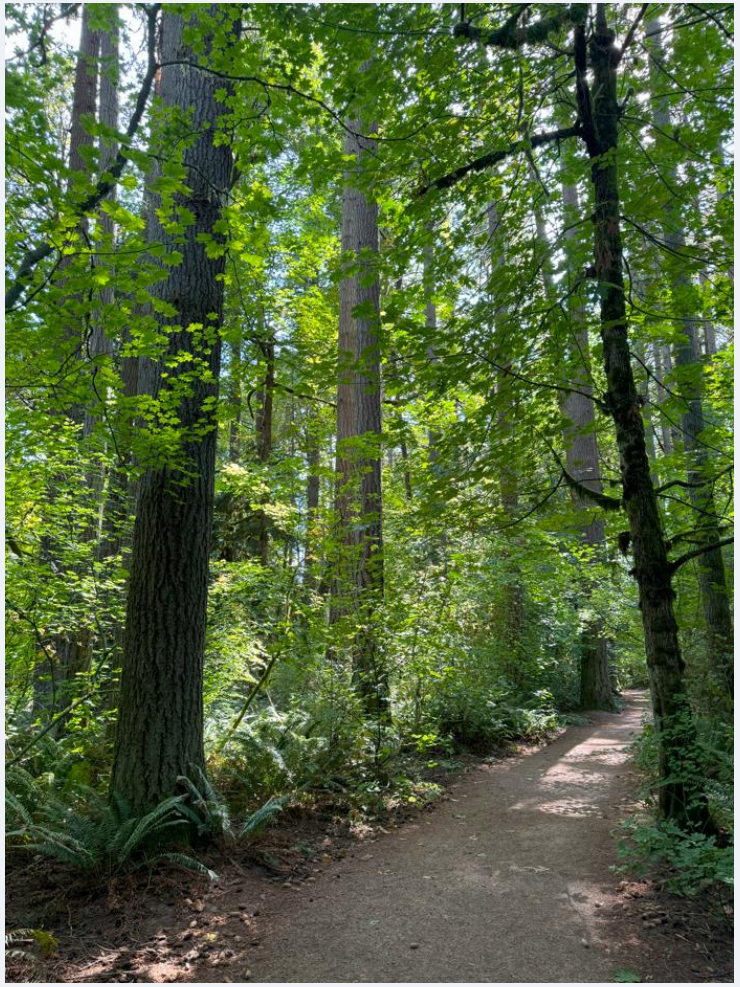






Talk Outline

I. **Stochastic Environments**
   a) Motivating the problem
   b) Psychology of uncertainty
   c) Goals for talk
II. **Factory Future Forecasting**
   a) Predictive models
      - Static
      - Dynamic
   b) Prescriptive models
      - General optimization principles
      - Factory scheduling
III. **AI and Factory Future Forecasting**
   a) Machine learning
   b) Generative AI
IV. **Integrating Best Practices**
V. **Conclusions**



How Should We Behave Given Stochastic Circumstances

1. Start investing early
2. Create a plan with long term goals
3. Diversify approach
4. As much as feasible, test goals against models
5. Trust the plan, automate actions and ignore day-to-day fluctuations
6. Only re-evaluate during critical decision points
   - Buying a house
   - Retiring / job change
   - Life changes
7. Focus on **Decision Quality** vs. Outcome Bias
8. Act strategically - not impulsively

How Should We Behave Given Stochastic Circumstances

1. We cannot know the future (but we do have to act)
2. There are a diversity of techniques to help forecast with probabilistic projections at different scales
3. Understand:
    - what you have (strengths and weaknesses)
    - what you need (and costs per improvement)
4. Identify a path of highest ROI for forecasting capabilities
5. For each forecasting application, automate as much as possible:
    - Verification / validation
    - Data ingestion
    - **Action!** (control)
6. Focus on **Decision Quality** vs. Outcome Bias
7. Understand when to intervene! (Deming's Funnel)
8. Be kind to yourself and others

Thank you!

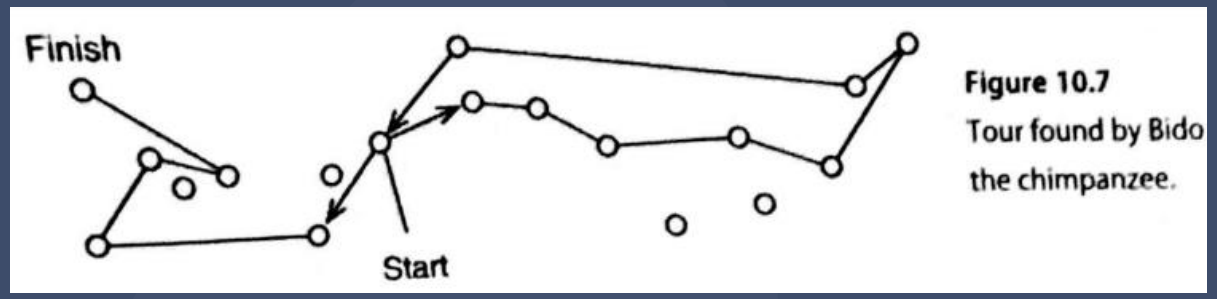


Figure 10.7
Tour found by Bido the chimpanzee.

Questions?

INFICON

Inspired by visions. Proven by success.

Connect with us!

www.inficon.com