AI - Where it Can (and Can't) Give You a Competitive Edge in R&D

Thursday, September 05, 2024

Presented by

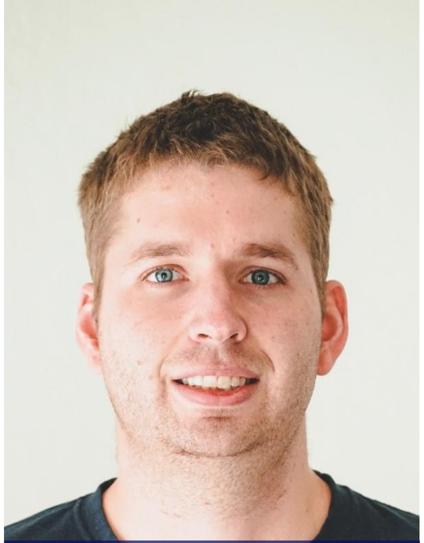
Noel Hollingsworth | CEO & Co-Founder | **Uncountable**

Agenda

Featured Speaker: Noel Hollingsworth

Noel Hollingsworth is Co-Founder and CEO at Uncountable. In his role, he works closely with Uncountable's customers to implement next-generation data management systems.

Prior to his work at Uncountable, Noel led data teams at startups and was awarded Forbes 30 under 30 for his work with machine learning and artificial intelligence.



CEO & Co-Founder

Noel Hollingsworth

uncountable®

About Uncountable

Founded in 2016
with offices in San Francisco, New York City, and Munich

100+ Customers Across Industries
R&D organizations including: paints & coatings, cosmetics & personal care, advanced materials, food & beverage, biotechnology & life sciences

One-of-a-Kind Platform
that centralizes R&D data and helps reduce new product development timelines

Proven Domain Expertise
Began as a data science company helping Fortune 500 materials companies accelerate development of new projects.

uncountable

Uncountable Proudly Supports Clients

That Span Across a Variety of Industries

CLARIANT | INDORAMA VENTURES | CABOT | PENN COLOR A WORLD OF COLOR | Sika | BEHR

remmers | INX A SAKATA INX COMPANY | codelpa COLORES DEL PACIFICO | SunChemical a member of the DIC group DIC Color & Comfort | Beiersdorf | AGC

SCG | coim | TRINSEO | syngenta | ZEON | TotalEnergies

CooperStandard | 6K | ROGERS CORPORATION | NAGASE | AlEn | Lohmann koester The Hygiene Brand

AirBoss Rubber Solutions | NAZDAR | NATURAL FIBER WELDING | 3D SYSTEMS | LESAFFRE | NOHBO

Greene Tweed | SOPREMA | REPSOL | tesa | pixelligent | ST. JUDE CHILDREN'S RESEARCH HOSPITAL

AI Today: What Is Causing All The Hype?

- **Deep Learning**
  - Large neural networks, taking advantage of large datasets to make predictions
  - Demonstrated revolutionary success in areas like image recognition and fraud detection in 2010s
- **Generative AI**
  - Popularized by ChatGPT in 2022
  - Capable of responding to text queries, summarizing data, creating images in "intelligent" manner
- **Tech Companies Rushing to Take Advantage of This**
  - 1 Trillion in CapEx over coming years

But Also Doubts?

Goldman Sachs | Global Macro Research

ISSUE 129 | June 25, 2024 | 5:10 PM EDT

TOP *of* MIND

GEN AI: TOO MUCH SPEND, TOO LITTLE BENEFIT?

What Characterizes These New Models?

Characterization	Fit for R&D?
Huge Data Sources Needed	If we can simulate experimental process - yes, otherwise no
Large Amount of Compute Required to Train	Yes
Black Box Explanations	It depends…
Results are not “Perfect”	Yes
Results Reflect Dataset	Yes, but it’s a limitation

What Does This Imply?

- **There will be places where these methods are revolutionary!**
  - Large amounts of pre-existing data
  - Places where we can run simulations or do extremely high throughput testing
  - Adjacent areas to experimentation, where we can collect data and learn insights
- **They are not a fit today for many scientific workflows**
  - Small data available
  - Simulation not possible
  - Want reasonable explanations for model behavior

Where Do We Go From Here?

3 Quotes from Peter Norvig (Google Research Director)

"We don't have better algorithms. We have more data."

"More data beats clever algorithms, but better data beats more data."

"Simple models and a lot of data trump more elaborate models based on fewer data."

Where do we go from here?

- **Get more data**
- **Make sure data is "better"**
- **Apply the right models to that data**
- **Understand that coatings R&D has inherent challenges:**
    - Could prevent an "AI revolution" taking place in an area with more data (e.g., drug discovery)

The Core Issue in R&D Orgs: Decentralized, Unstructured, & Fragmented Data

Data is collected & sits independently across the different teams and systems used throughout the entire R&D value chain…

Raw Materials

- Physical Attributes
- Chemical Attributes
- Batch History
- Prices
- Regulatory
- Compliance

Formulations

- Compositions
- Calculations
- Multi-part systems
- Order of Addition
- Dispersions & Intermediates

Processing

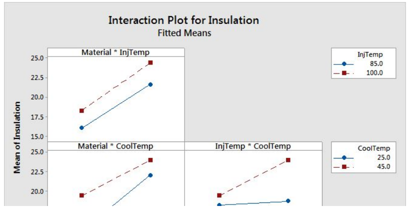
- Mixing Conditions
- Curing Parameters
- Metadata

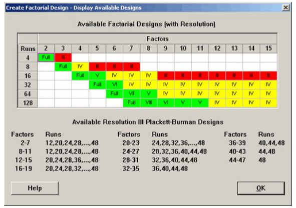
Formula Properties

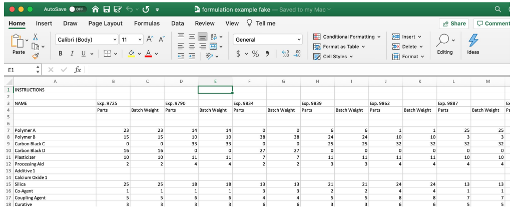
- Physical Properties
- Analytical Measurements
- Observations
- Images
- Curves & Reports

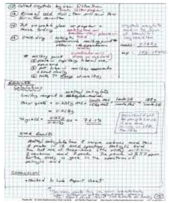
Product Properties

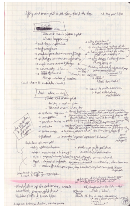
- Application Testing
- Customer Feedback
- Panel Testing

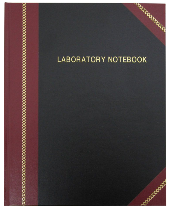












Coatings Trends & Technologies SUMMIT

uncountable.

Types of Data Systems: R&D Organizations

Spreadsheets

ELN/Lab Journals

LIMS

Inventory

Visualizations / Analysis

Statistical Tools

Predictive tools

Other Internal Databases

Data Systems: Structured vs. Unstructured

Unstructured

Examples

- Spreadsheets
- Word Documents
- PDFs
- Lab Journals/ELNs
- SharePoint/Shared Drive

Advantages

- Free
- Unrestricted entry of information
- Known/second nature "habitual"

Disadvantages

- Limited scope & scalability for application of info
- Ctrl+F keyword searching
- Limited collaboration
- Inability to innovate efficiently & at market-rate

Structured

Examples

- Databases
- LIMS
- Inventory Systems
- Uncountable

Advantages

- Instant access to specific information/data
- Shareable & scalable information
- Intelligent insights & reporting

Disadvantages

- Requires intentional/deliberate entry of information
- Change management
- Migration of historical data into new system
- Disciplined use

Top 3 Problems: Deploying AI Without Structured Data

- **Why Excel & Unstructured Data System Are Insufficient**

1. **Volume of Data**
    - A small data set with the best AI model in the world is worse than both expert scientists and simpler AI models applied to "big data"
        - The most important aspect of any AI model is its underlying data - both size and cleanliness

2. **Relevancy to Problems**
    - Will create desire to squeeze square peg in round hole - When we do have some data, we must apply AI, even if it's not a fit
        - AI is not a fit for all use cases!

3. **Scientist Trust**
    - Desire to be AI-first company without gathering appropriate data = scientist trust being lost
        - AI ends up being applied to projects that aren't good fits, or only to high priority projects that carry substantial failure risk - when there are issues, team loses faith in the process
        - Sufficient Data is important, but not the only prerequisite

Importance of Structuring Lab Data for AI: Example of Brookfield Viscosity

- **Standard Way Data Gets Recorded In Spreadsheets & Notebooks:**
    - Viscosity, 7D = 3000
    - Brookfield Visc. Sp #4 = 5500
    - BV, ON = 1800

- **Best Practices for Structuring Lab Data for AI:**
    - Brookfield Viscosity = 5000
        - Liquid Aging Time + Temperature: 7D at 23°C
        - Spindle #4
        - RPM: 150
        - Test Temperature: °23
            - Exact temperature and time
            - Machine SN, Operator

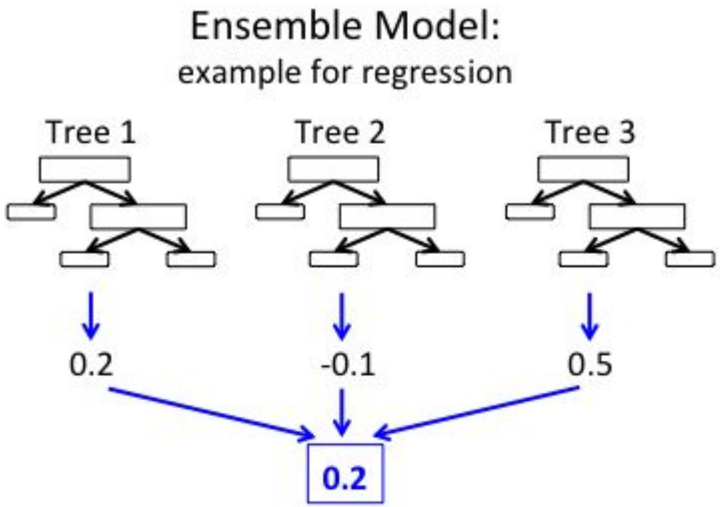
Coatings Trends & Technologies SUMMIT uncountable.

How Do We Select The Right Models To Use?

- **Can we just pick the one with the best performance on our test set?**
  - It’s not that easy!

Example: Gradient Boosted Trees

- **Often has best performance predicting results**
- **Decisions made by answering:**
  - "Yes" "No" Questions & Summing Results
- **Imagine this applied to a viscosity prediction**
  - If water < 50%: Viscosity is 100 cps
  - If 52% > water > 50% > Viscosity is 110 cps
- **Even if results are good…**
  - Does not reflect reality
  - Produces downstream issues



Example: Exploitation vs. Exploration

- **Once you have a model, you have to decide which points to test**
- **We could test the point our model thinks will perform best**
  - What if this is just a copy of our previous best result
  - Incidentally, what does “perform best” mean?
- **We could test the point that gives us the most new information**
  - What if this is something completely impractical to formulate?
- **Balancing between the two is a key goal in the field of Bayesian Optimization**
  - In our case, often done with many different physical, cost and regulatory constraints
- **Selecting a data point is significantly more difficult than predicting its performance**

So, What Do We Do?

- Don't just assume that good model performance = Good suggestions of experiments to run
- Performance on a toy example does not reflect performance on your data
- If you implement AI with a partner, emphasize:
  - **Is AI something they deeply understand or a trend they are chasing?**
    - Anyone can hire someone to be their "AI expert" - Was the company built with AI in mind, or was it added because of where we are in the hype cycle?
  - **Are they trying to sell you on AI for all use cases, or just where it's a fit?**
  - **Are they demanding you spend large amounts of money on nebulous AI projects before seeing results, or do they work with you to take it step by step?**

Considerations: Setting The Right Expectations

- **Too big of a search space**
  - 100s of ingredients, but limited data points
    - Either from collection, cleanliness, or standardization

- **Moonshot objectives**
  - What are you trying to achieve in this project vs long term goals
  - What are more reasonable targets that would allow you to claim “progress”

- **Perception of perfection**
  - Why would model suggest such a thing?
  - Why isn’t model more accurate?
  - Can it model pictures of exposure ratings?

How to Define & Create an AI Roadmap

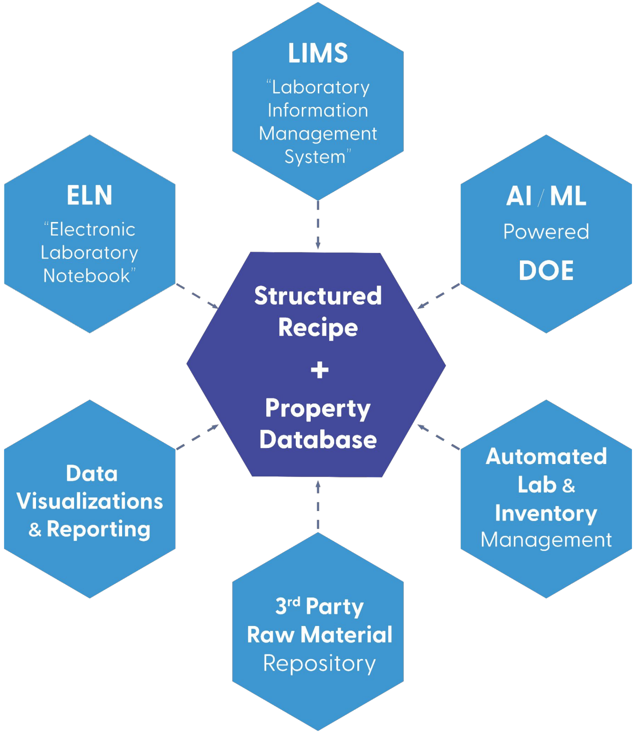
1. **Before (Preparation)**
   - Ensure structured data system in place
   - Verify all scientist work is being captured in a way fit for AI
     - All data points and all aspects of data
       - Example: Viscosity centipoise, temperature, spindle, rpm…
   - Utilize in-house expertise to understand/validate vendor and partner “claims”

2. **During (Deployment)**
   - Identify appropriate targets for AI - Example Criteria:
     - Large Amounts of Data
     - Known Success Criteria
     - Consistent Output Results
   - Ensure AI is embedded into daily workflows
     - Not judged off success in a project where majority of results are out of scientists control

3. **After (Maintenance)**
   - Identify areas where data capture is insufficient
   - Deploy systems and/or recurring procedures to collect data

Uncountable's All-In-One Platform?

A Platform to Centralize, Connect, And Structure All Types Of R&D Data.



Coatings Trends & Technologies SUMMIT

uncountable

New Modern Digitalization Tools: R&D Labs in Paints & Coatings

- **Drive learnings from the raw material level to improve performance**
  - Better understanding of the correlation between the data specific to raw materials and the desired results for formulas

- **Integration with raw materials/costs systems**
  - Enable world where scientists don't have to go to multiple systems to formulate with context

- **Unified Laboratory Informatics Platforms**
  - End-to-end web-based systems connecting recipes to results
  - Data is collected in standardized, streamlined, consistent way

- **AI/ML-powered DOE approach to efficiently explore the defined formulation space**
  - Enabling acceleration towards sustainability goals

Example 1

Platform-Wide Data

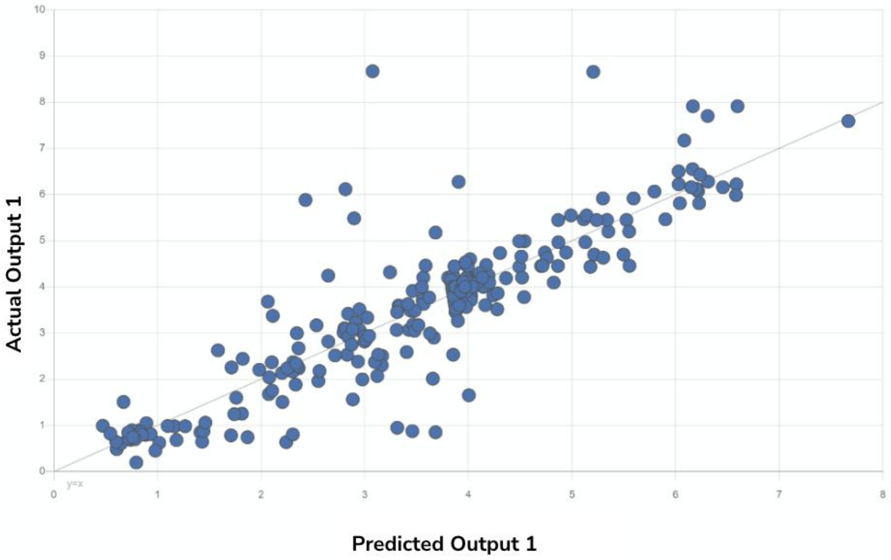
Example 1: Output Fits

Training Accuracy

Numeric Outputs

Plot	Name	Summary Statistics (Training Data)			Model Predictions		
		# Samples	Mean μ	Std. Dev. σ	RMSE	r² Score	Explained Error %
	Output 1	1360	172	145	54.4	0.86	**62.6%**
	Output 2	319	8.17	10.3	3.39	0.891	**67.1%**
	Output 3	254	16.4	7.38	5.03	0.533	**31.8%**
	Output 4	541	65.2	71.4	17.7	0.939	**75.2%**
	Output 5	359	8.02	5.4	2.96	0.698	**45.1%**
	Output 6	397	41.6	31.9	8.23	0.933	**74.2%**
	Output 7	1511	7520	12200	6070	0.752	**50.2%**
	Output 8	251	3.51	1.7	0.818	0.767	**51.8%**
	Output 9	895	7.25	4.03	2.12	0.725	**47.6%**
	Output 10	1475	8.26	6.71	2.9	0.813	**56.8%**
	Output 11	1497	0.00279	0.00112	0.000767	0.533	**31.7%**

High Predictive Accuracy Low Predictive Accuracy Add Comparison Model



Coatings Trends & Technologies SUMMIT uncountable.

Example 1: Linear Coefficients

Effect Sizes ⓘ

☑ Show linear coefficient approximations

Type	Name	Output 1	Output 2	Output 3	Output 4	Output 5	Output 6	Output 7	Output 8	Output 9	Output 10	Output 11	Output 12
Ingredient 1									-0.753			-1.51	0.000365
Ingredient 2									0.231			-0.0274	0.000324
Ingredient 3			10.8						0.0461			0.153	0.000214
Ingredient 4			27.1		1.32	-34.7	[illegible]	70	-1.59	[illegible]	-6.86	-0.872	-0.000199
Ingredient 5			-1.76		-0.0571		-0.223	0.233	0.229	-0.615	12.3	-0.781	-0.000187
Ingredient 6			4.98						3.21		-11.7	0.199	-0.000177
Ingredient 7			3.48						0.0483			0.18	0.00017
Ingredient 8					-0.157				-0.445	0.487	-3.6	0.233	0.000106
Ingredient 9			-25.7		-1.13	4.96	15.5	-3.46	2.95	20.8	76.7	-5.88	-0.0000996
Ingredient 10											-1.35	-0.603	-0.0000943
Ingredient 11			8.02		7.32	11.2		5.34	-1.12	5.63	2550	-10.7	0.0000875
Ingredient 12		-26.3	0.0219								-0.762	0.0724	0.0000671
Ingredient 13			5.72						-0.0281		-5.8	0.206	-0.0000477
Ingredient 14			3.5						-0.0947		-4.96	0.387	0.0000437
Ingredient 15			1.38		0.0708	0.793	-0.437	0.422	-0.0411	-0.893	4.04	-0.259	-0.000042
Ingredient 16			1.92					0.0139	0.192		-7.79	0.0894	0.0000417
Ingredient 17	2.2-1-2.4) +		-1.83			-0.064				0.0867	3.92	-0.158	0.0000402
Ingredient 18									39.5			0.171	0.0000366
Ingredient 19	rsion		-25.1		0.00226	-4.12	-1.95	0.0962	-1.25	-1.54	3.47	7.37	0.0000344
Ingredient 20		-25.3	0.224						0.0311		-0.561	0.0222	0.0000326
Ingredient 21			2.99						-0.276		-4.03	-0.0774	-0.0000317
Ingredient 22									0.193	0.183	6.13	0.373	-0.0000294
Ingredient 23			2.38								-2.82	0.335	0.0000285
Ingredient 24			0.168								-3.07	-0.0553	-0.0000255
Ingredient 25			1.73						-0.0281		-1.21	0.263	0.0000246
Ingredient 26			-0.915		0.0203	-0.41	-0.114	0.12	-0.0618	-0.687	-0.673	-0.38	-0.0000236
Ingredient 27			1.49						0.438		-3.2	0.526	0.0000232

Example 2

Targeted Experiments

Example 2: Suggested Experiments

Suggested Formulations

		Recipe 1	Recipe 2	Recipe 3	Recipe 4	Recipe 5	Recipe 6
Recipe Name							
Import Recipe?							
Ingredient 1	= 2.5	2.5	2.5	2.5	2.5	2.5	2.5
Ingredient 2	= 0.202	0.2021	0.2021	0.2021	0.2021	0.2021	0.2021
Ingredient 3	= 0.0121	0.01209	0.01209	0.01209	0.01209	0.01209	0.01209
Ingredient 4	[6.52, 10.2]		9.709	10.11	8.072	8.411	9.5
Ingredient 5	[10.3, 25]	22.71	22.71	17.27	18	18.62	18.69
Ingredient 6							
Ingredient 7	[4.78, 10]	9.658	8.146		6.247	6.307	
Ingredient 8	[5.84, 10.1]	6.138	6.347	6.328	6.235	5.906	5.877
Ingredient 9	[5.2, 50]	27.86	39.03	32.46	30.69	46.62	49.81
Ingredient 10	[5.1, 14.3]	10.26	11.35	9.108	10.64	11.42	13.41
Ingredient 11	[14.8, 32]	20.67		22	17.4		
		100	100	100	100	100	100
Calculation 1	[2, 4]	2.25	3.83	2.09	2.39	2.98	2.78
Calculation 2	[0.8, 1.2]	0.804	1.11	1.02	1.08	1.13	1.08
Calculation 3		3.16	3.87	2.89	3.21	3.01	2.81
Calculation 4		1.91	9.82	0.595	3.04	6.36	5.41
	Goal						
Predicted Output 1	≥ 300	221 ± 111	236 ± 111	224 ± 109	216 ± 107	210 ± 108	215 ± 109
Predicted Output 2	≥ 5	7.5 ± 2.75	5.82 ± 3.11	7.71 ± 2.88	7.39 ± 2.84	5.71 ± 3.08	5.94 ± 3.11
Predicted Output 3	≤ 4000	2100 ± 1390	1860 ± 1630	2140 ± 1580	2050 ± 1460	1810 ± 1760	1900 ± 1840
Predicted Output 4	≥ 2	2.5 ± 1.07	2.68 ± 1.15	1.94 ± 1.12	2.56 ± 1.07	2.46 ± 1.1	3.85 ± 1.18

Q&A

Just a few of many

Uncountable Long-Term Customers

Thank You!

Questions?

Email: info@uncountable.com
Inquiries: www.uncountable.com/contact-us

www.uncountable.com

Want a Demo?

Scan The QR Code