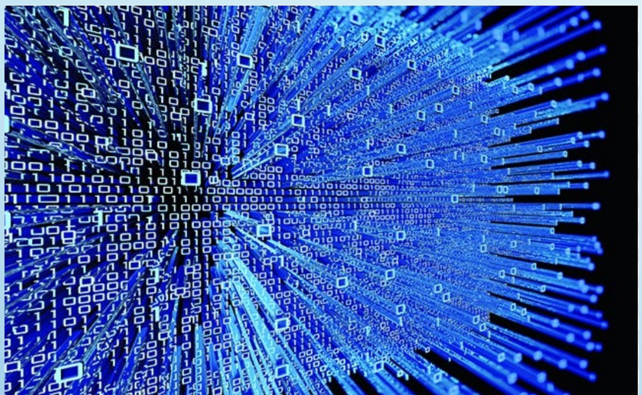
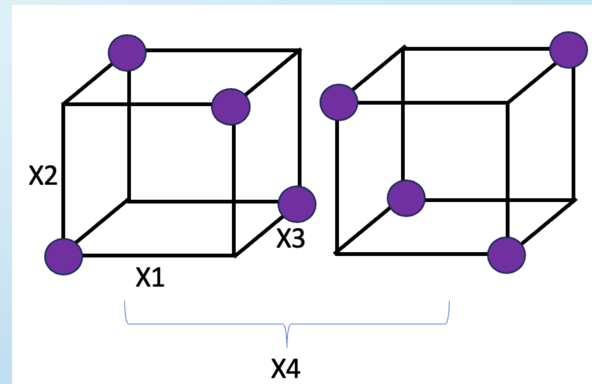
SYNERGIES OF DESIGNED DATA COLLECTION AND BIG DATA



Christine M. Anderson-Cook

candcook@gmail.com



JMP STATISTICALLY SPEAKING SERIES

JULY 2023

BASED ON THE DISCUSSION PAPER IN QREI

Quality and Reliability Engineering International (2023) 39(4) 1085-1119.

WILEY

DISCUSSION

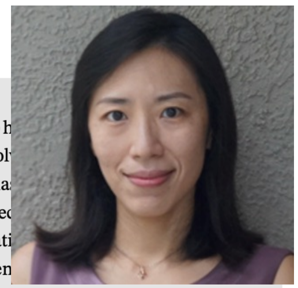
Is designed data collection still relevant in the big data era?

Christine M. Anderson-Cook[1] | **Lu Lu**[2]

[1] Los Alamos, New Mexico, USA

[2] University of South Florida, Tampa, Florida, USA

Correspondence
Christine M. Anderson-Cook, Los Alamos, New Mexico, USA.
Email: candcook@gmail.com



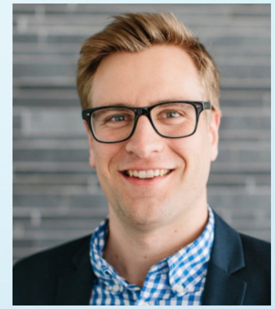
Abstract

The excitement and buzz around big data (BD) h
BD has become thought of as a panacea for sol
in fact limitations to what BD can do, which ha
potentially suboptimal problem-solving and dec
to the question posed in the title is an emphati
use of designed data collection (DDC) represen
studies involving BD are planned, executed, and interpreted to deliver more satisfying solutions. In this paper, we outline eleven opportunities for integrating DDC strategies, including statistical design of experiments (DoE) and sampling techniques, with BD at all stages of the study. There are opportunities for building background understanding of the problem and improving the quality of the data via strategic planning before acquiring BD. In addition, incorporating DDC into the collection and processing of the actual BD, and strengthening conclusions or augmenting with supplementary designed data after BD also represent key opportunities to increase the impact of BD in the era of Industry 4.0 and Quality 4.0.

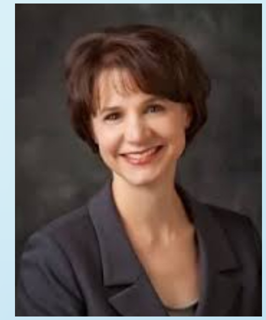
KEYWORDS
causality connections, data fidelity, design of experiments, Industry 4.0, Quality 4.0, sampling, sequential data collection

Thank you to the discussants for their insights …

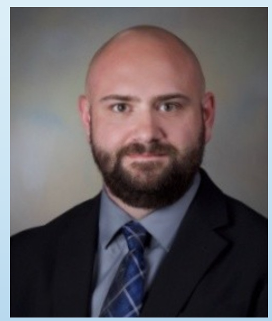
Laura Freeman



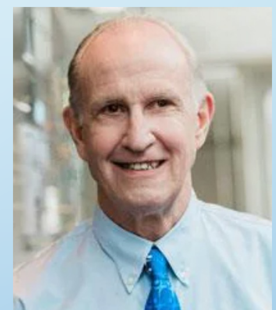
Nathaniel Stevens



Allison Jones-Farmer



Caleb King



Bradley Jones

OUTLINE

- Defining "Big Data" (BD) and "Designed Data Collection" (DDC)
- Situations with synergy between Big Data and Designed Data Collection

1. Before Big Data

a. Pre-planning to build understanding

b. Measurement assessment / data fidelity

2. Actual Big Data

a. Generating BD

b. Designed Simulation experiments

c. Sequential DoE

d. Sequentially ordered execution

e. Adaptive experiments

f. Sampling to downsize big data

g. Management and storage of BD

3. After Big Data

a. Validation of observational BD

b. Supplementing BD for new problems

- Illustrative Examples
- Conclusions

WHAT IS "BIG DATA"?

	A	B	C	D	E	F	G	H	I	J	K
1	Order #	First Name	Last Name	Email	Country	IP address	Total	Item #	Payment	Shipping	Status
2	1	Dalton	Kramer	dalton@email.com	France	211.91.226.108	99	868	Card	Regular	In progress
3	2	Gita	Tetterton	gita@email.com	USA	222.153.179.100	99	537	Card	Regular	Delivered
4	3	West						16	Paypal	Regular	Delivered
5	4	Brad						73	Card	Fast	Delivered
6	5	Mary						29	Bank	Regular	In progress
7	6	Allys						40	Card	Regular	In progress
8	7	Lucile						48	Paypal	Fast	In progress
9	8	Mick						53	Paypal	Fast	Delivered
10	9	Clarin						17	Bank	Regular	Delivered
11	10	Kimberly	Penny	kimberly@email.com	France	[illegible]	99	998	Bank	Regular	In progress
12	11	Colleen	Kellough	colleen@email.com	USA	73.51.152.185	49	14	Paypal	Regular	In progress
13	12	Nettie	Edmonds	nettie@email.com	Spain	94.133.138.234	99	670	Card	Fast	Delivered
14	13	Duncan	Rickenbacker	duncan@email.com	France	211.91.226.108	199		Card	Regular	Delivered
15	14	Marchelle	Diedrich	marchelle@email.com	Italy	222.153.179.100	29		pal	Regular	Delivered
16	15	Mariano	Murrell	mariano@email.com	Italy	203.123.236.1	99	47		Fast	Delivered

Types of data:
- Numeric (discrete, categorical, continuous)
- Text
- Functional

Large number of measures per item:

Examples:
- Characterization of materials
- Genomics

Very large number of observations:

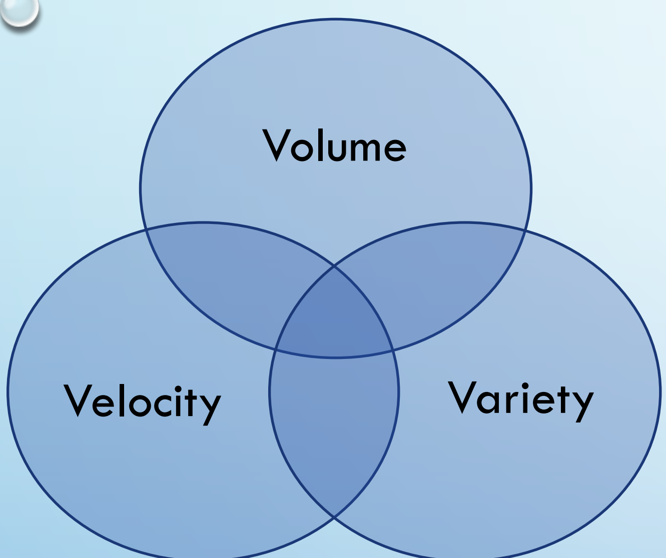
Examples:
- Credit card transactions
- Assembly line measurements

Many observations, many measurements:

Examples:
- Loyalty program grocery summaries
- Government census

"Big" is in the eyes of the beholder (and changing as computers increase in power)

CHALLENGES OF BIG DATA



Volume: So much data! Storing, organizing, managing

Velocity: Speed of accumulation (batch, near time, real time, streaming)

Variety: Different types combined (structured, unstructured, semi-structured, different frequency of measurement)

Other V's:

- **Validity (data quality)**
- **Veracity (accuracy)**
- **Variability (precision)**
- **Value (relevance)**
- **Venue (access to dispersed sources)**

THE COSTS OF BIG DATA

Costs declining steadily as computer technology evolves
***Primary focus when touting emergence and benefits of big data**

1. Collecting, storing and managing the data

2. Analysis of data (primarily labor and time, computer costs)

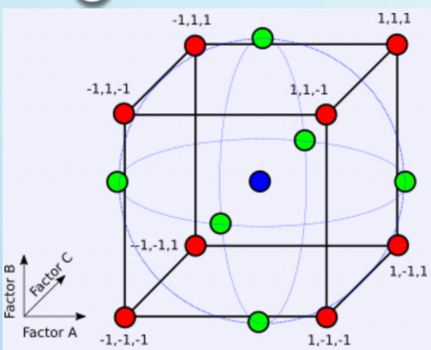
3. Opportunity cost of collecting the wrong data (waste and delay)

4. Incorrect conclusions which lead to sub-optimal actions, decisions

Opportunities for designed data collection to have a powerful impact

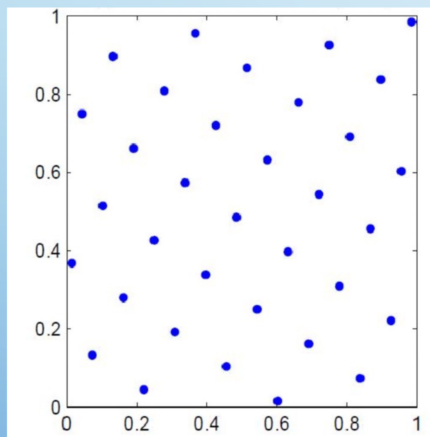
WHAT IS "DESIGNED DATA COLLECTION"?

DESIGNED EXPERIMENTS:



Response surface

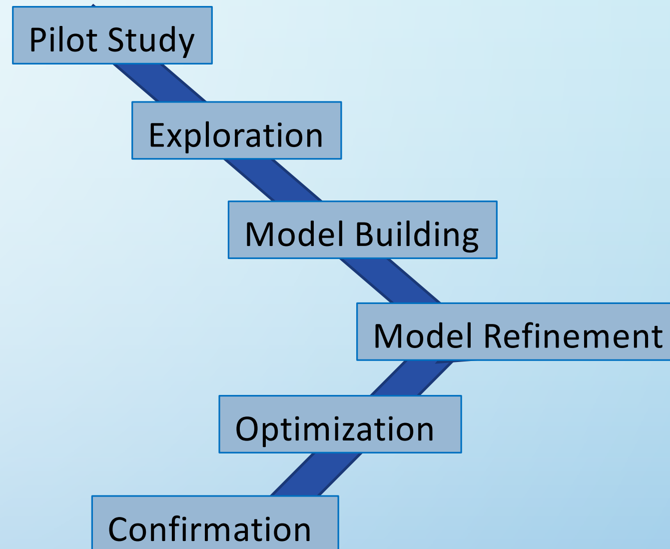
Space filling



SAMPLING:

First Name	Last Name	Email
Dalton	Kramer	dalton@email.com
Gita	Tetterton	gita@email.com
Weston	Jurgens	weston@email.com
Brad	Chupp	brad@email.com
Marybeth	Baumann	marybeth@email.com
Allyson	Feder	allyson@email.com
Lucile	Folks	lucile@email.com
Mickey	Rusk	mickey@email.com
Clarine	Esslinger	clarine@email.com
Kimberly	Penny	kimberly@email.com
Colleen	Kellough	colleen@email.com
Nettie	Edmonds	nettie@email.com
Duncan	Rickenbacker	duncan@email.com
Marchelle	Diedrich	marchelle@email.com
Weston	Jurgens	weston@email.com
Brad	Chupp	brad@email.com
Marybeth	Baumann	marybeth@email.com
Allyson	Feder	allyson@email.com
Lucile	Folks	lucile@email.com
Mickey	Rusk	mickey@email.com
Clarine	Esslinger	clarine@email.com
Kimberly	Penny	kimberly@email.com
Colleen	Kellough	colleen@email.com

SEQUENTIAL EXPERIMENTS:



WHAT IS NEEDED FOR SUCCESS WITH BIG DATA

Data Management
(cleaning, storage, manipulation of data)

Statistical Expertise
(modeling, handling variation & uncertainty)

Engineering Problem-solving
(mindset, using the right tools in combination to solve complex problems)

Analytics Translator
(integrate data/information with business values)

OUTLINE

- Defining "Big Data" (BD) and "Designed Data Collection" (DDC)
- Situations with synergy between Big Data and Designed Data Collection

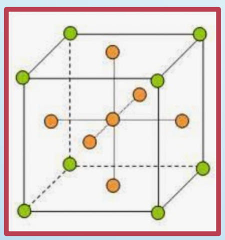
1. **Before Big Data**
   a. Pre-planning to build understanding
   b. Measurement assessment / data fidelity
2. **Actual Big Data**
   a. Generating BD
   b. Designed Simulation experiments
   c. Sequential DoE
   d. Sequentially ordered execution
   e. Adaptive experiments
   f. Sampling to downsize big data
   g. Management and storage of BD
3. **After Big Data**
   a. Validation of observational BD
   b. Supplementing BD for new problems

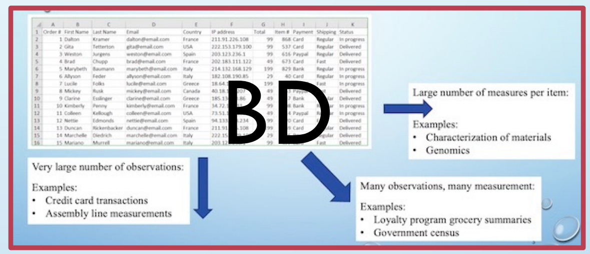
- Illustrative Examples
- Conclusions

BEFORE BIG DATA

1.a. Pre-Planning to Build Understanding:

What preliminary set of data should be collected to help refine the focus, before vast resources are spent collecting the Big Data?





Key points:

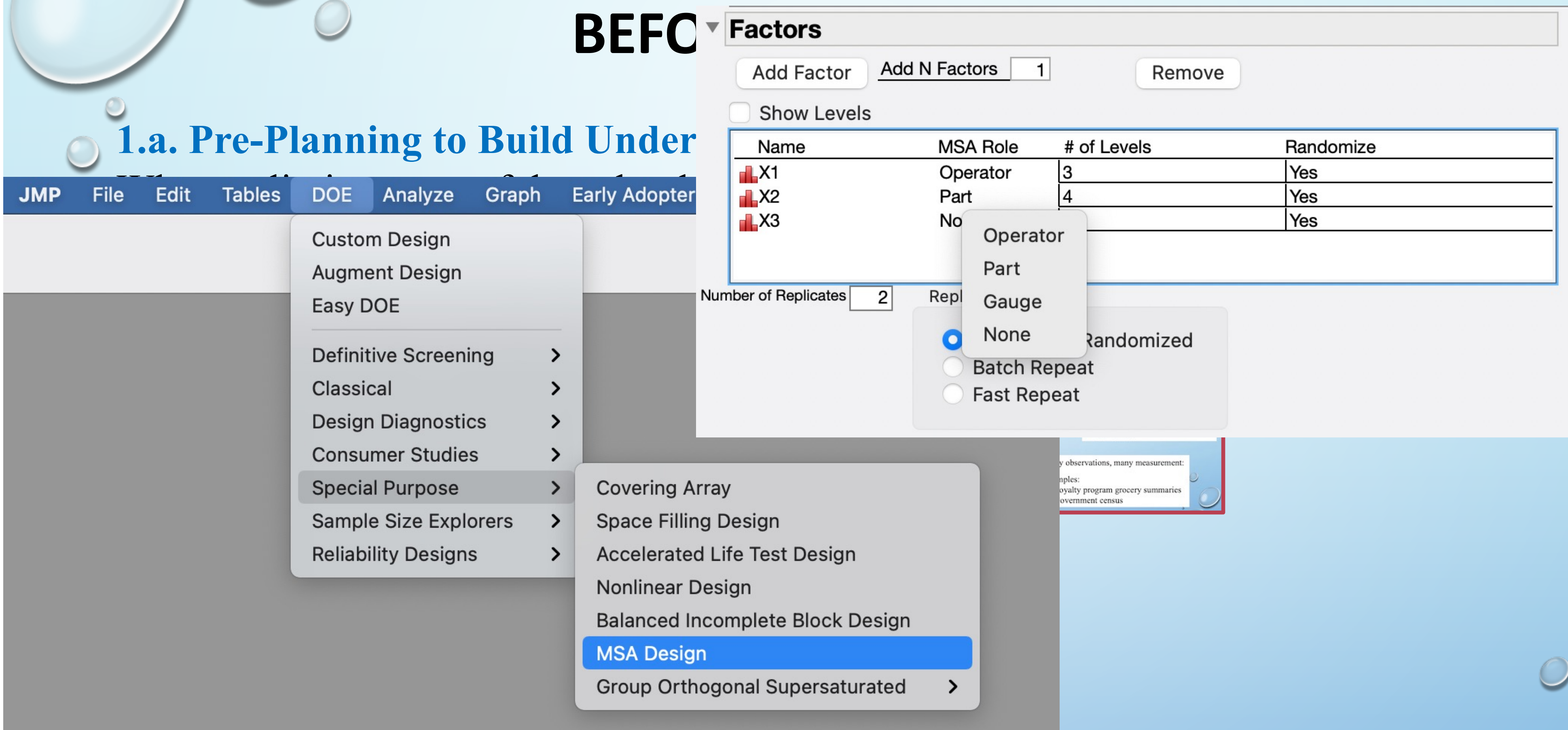
- Reduce data collection costs
- Inform effective collection of big data
- Early opportunity to stop, question, decide the purpose of data
- Resolve issues around fidelity and precision of data

1.b. Measurement Assessment:

Will the Big Data be of sufficient quality (accuracy & precision)?

BEFO

1.a. Pre-Planning to Build Under



1.b. Measurement Assessment:

Will the Big Data be of sufficient quality (accuracy & precision)?

CARBON CAPTURE IN INDUSTRY

<u>Goal</u>: Develop chemical and physics-based technologies (solvents and materials) to remove CO_2 from industrial smokestacks and sequester it

<u>Keys</u>: - leverage what we know from earlier experiments
- use precious runs at Full Scale strategically
- build design sequentially* to learn within each scale

<u>Data</u>: steady-state and dynamic, many attributes, long duration, functional data

*Lu & Anderson-Cook, 2021

Bench-scale → Lab-scale → Pilot Scale → Full Scale

OUTLINE

- Defining "Big Data" (BD) and "Designed Data Collection" (DDC)
- Situations with synergy between Big Data and Designed Data Collection

1. **Before Big Data**
   a. Pre-planning to build understanding
   b. Measurement assessment / data fidelity

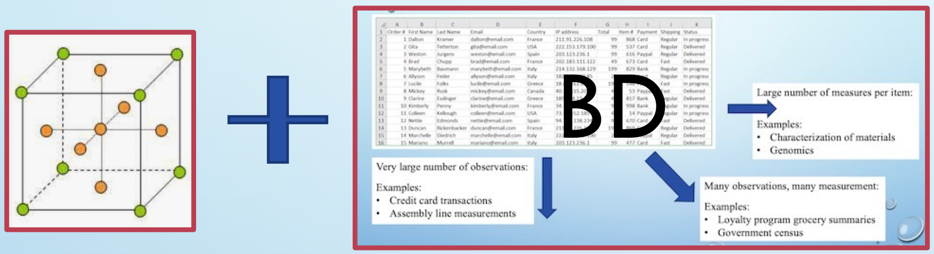
2. **Actual Big Data**
   a. Generating BD
   b. Designed Simulation experiments
   c. Sequential DoE
   d. Sequentially ordered execution
   e. Adaptive experiments
   f. Sampling to downsize big data
   g. Management and storage of BD

3. **After Big Data**
   a. Validation of observational BD
   b. Supplementing BD for new problems

- Illustrative Examples
- Conclusions

DESIGNED DATA COLLECTION + BIG DATA

2.a. Generating Big Data:
Intentionally select what data to collect for Big Data



Key points:

- Need to have the ability to manipulate inputs to generate data
- Allows causal connection between inputs and responses to be drawn
- Strategic selection of data to collect that
  - (a) reflects the input space of interest,
  - (b) targets the questions of interest and
  - (c) is manageable for storage/sharing/planned analysis

URBAN RADIATION DETECTION

Problem: At major events, radiation detectors are deployed to search for dirty bomb in a cluttered urban environment. Data from detectors is noisy because of variable background environment. Goals were to

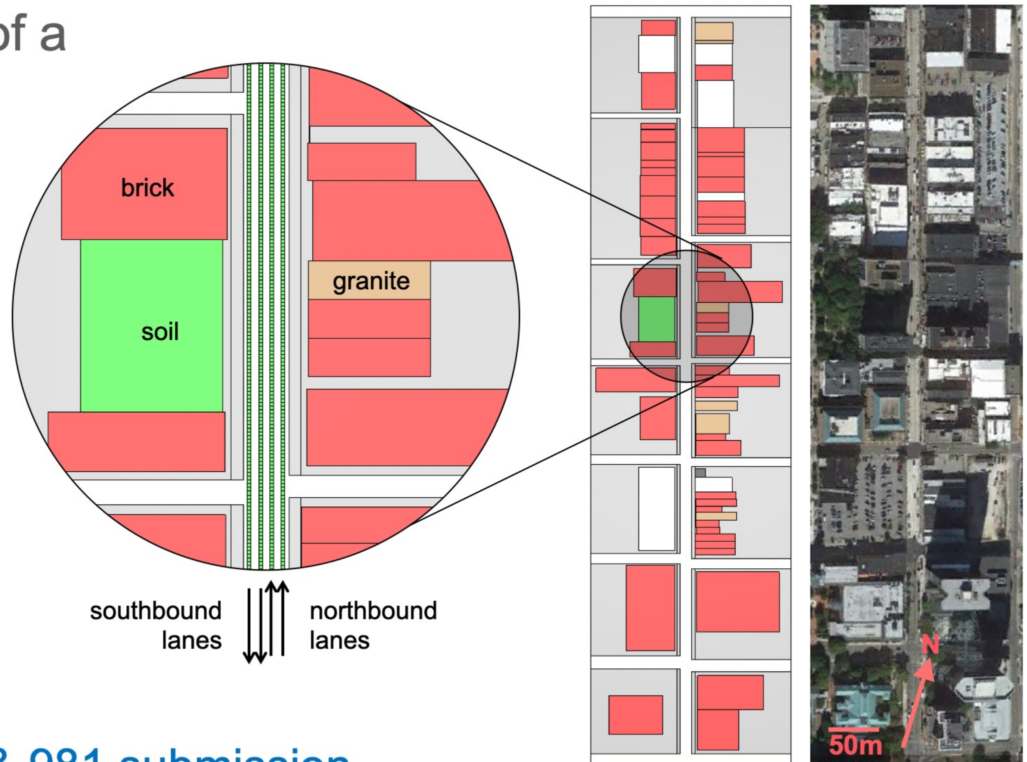
(1) improve algorithms to **detect**, **identify** and **locate** radioactive sources, while maintaining an acceptable **false positive rate**, and
(2) compare algorithms

Solution: Fielded a data competition topcoder a wipro company with prize money to drive improvements, and use data as standardized data set for comparison.

Technical Challenges:

- Generating realistic data with known answers
- Want algorithms to perform well in very diverse environments
- Design of experiments on massive scale (~26 000 runs)

- ORNL* designed multiple versions of a 0.5 mile street model with characteristics similar to a street in Knoxville.
- 56 buildings
  - –48 brick, 7 granite, 1 concrete
- Side streets, sidewalks, 6 parking areas.
- Ability to vary levels of K, U, Th.



2 competitions:

Government (2018) – 16 participants & 981 submission

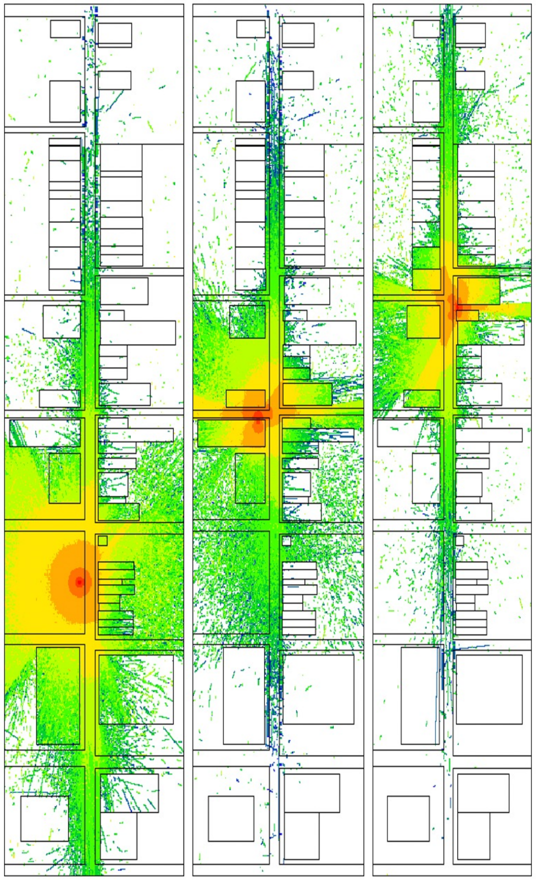
TopCoder (2019) – 71 participants & 1614 submissions

* Oak Ridge National Laboratory

A. **Background:** 8 background models, each with 82 parameters

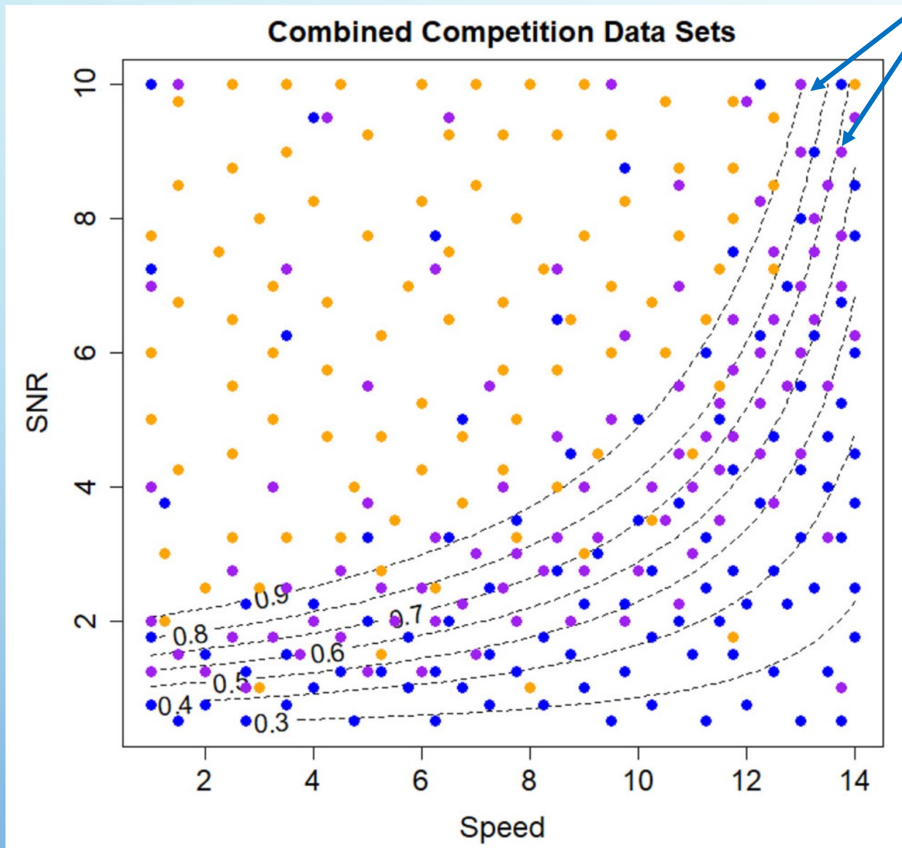
B. Source
- **6 types** (2 weapon, medical and industrial isotopes)
- **2 shielding settings** (On/Off)
- **strength**
- **15 source locations**

C. Other
- **speed** of detector
- **proximity to source** (lane for detector)
- length of path, starting location

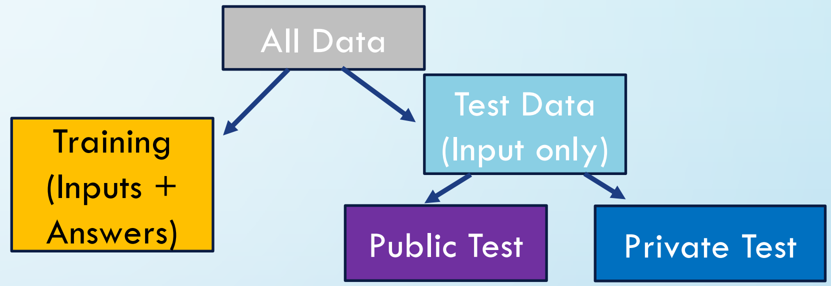


Design of Experiments challenge:
Create complementary data sets with ~10000 runs for training, ~6700 + ~9300 runs for public & private test

Sample Solution using **Non-Uniform Space Filling*** designs: unshielded Pu



Contours show P(correct detection) for current state of the art algorithm performance



Yellow: training data – generally easiest

Purple: public test – moderately difficult

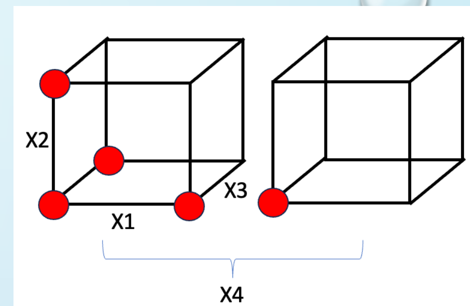
Blue: private test – most difficult, with ability to evaluate adapting to new cases

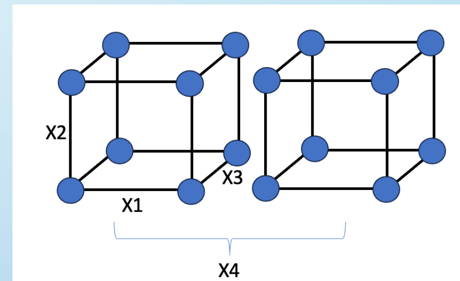
*Lu et al, 2020

DESIGNED DATA COLLECTION + BIG DATA

2.b. Designing Simulation Studies:

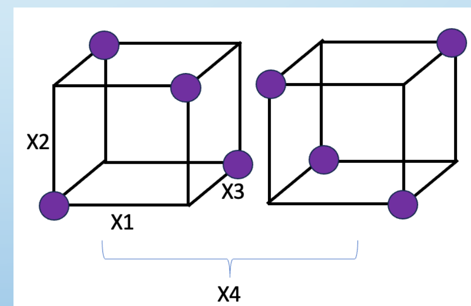
Goal: Demonstrate capability of new method and/or compare to alternatives





Key points:

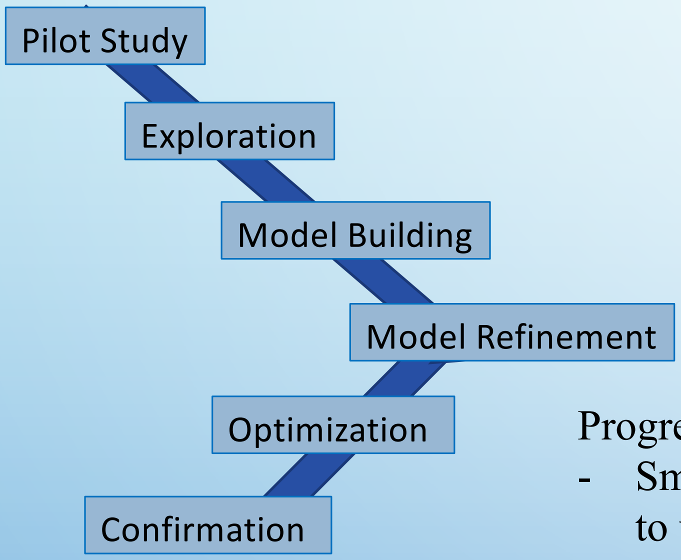
- Intentionally select what aspects to manipulate, and what levels to show – want to characterize a broad region of interest
- Avoid haphazard choices
- Avoid cherry-picking scenarios where new method performs well
- Factorial structure can grow very quickly – take advantage of DOE structure (eg. fractional factorial)
- Model results instead of using large tables



DESIGNED DATA COLLECTION + BIG DATA

2.c. Sequential Design of Experiments:

Learn as you go with small experiments, each with a specific objective



Carbon Capture Example

Progression:

- Small exploratory experiment (space-filling) to verify science from smaller scale

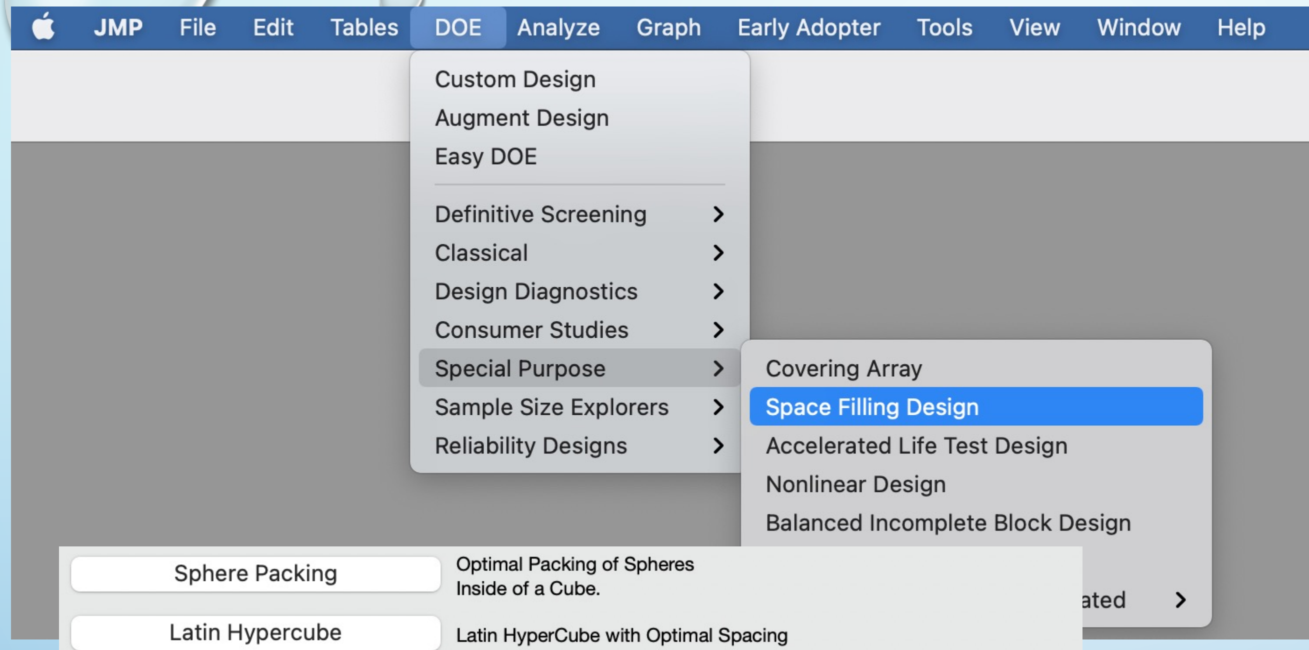
BIG DATA

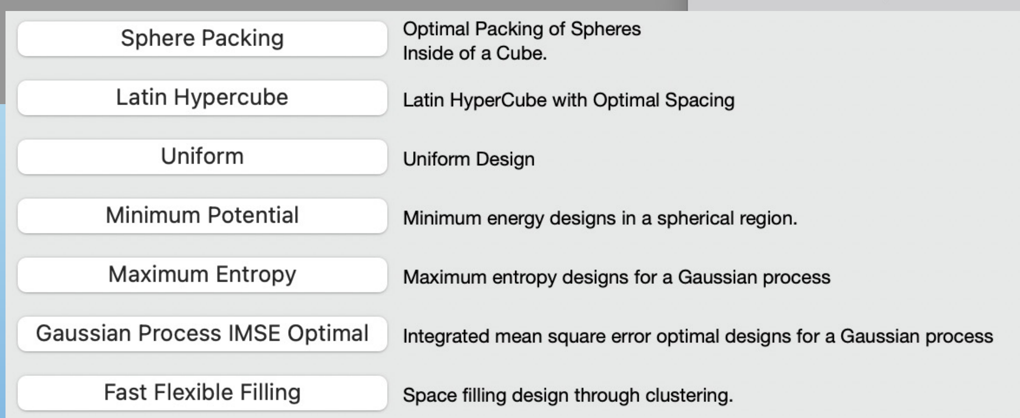
ecific objective

Carbon Capture Example

eriment (space-filling)

smaller scale

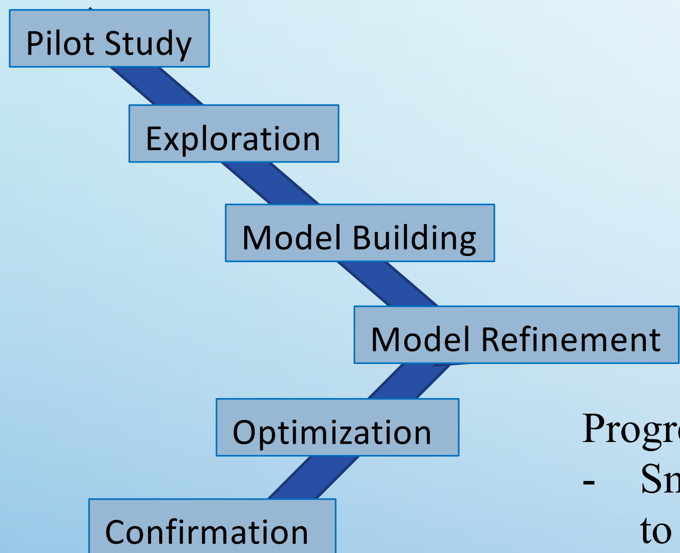


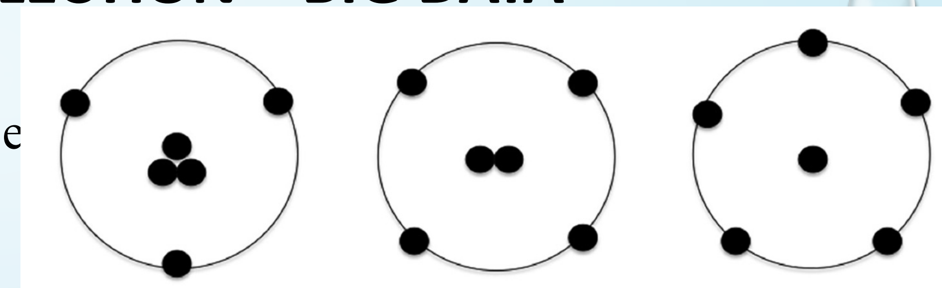


DESIGNED DATA COLLECTION + BIG DATA

2.c. Sequential Design of Experiments:

Learn as you go with small experiments, e





(Stevens & Anderson-Cook, 2019)

Carbon Capture Example

Progression:

- Small exploratory experiment (space-filling) to verify science from smaller scale
- Refine model with more data targeting where scale changes relationships
- Optimize to find best settings for this scale
- Confirm optimal settings for production

DESIGNED DATA COLLECTION + BIG DATA

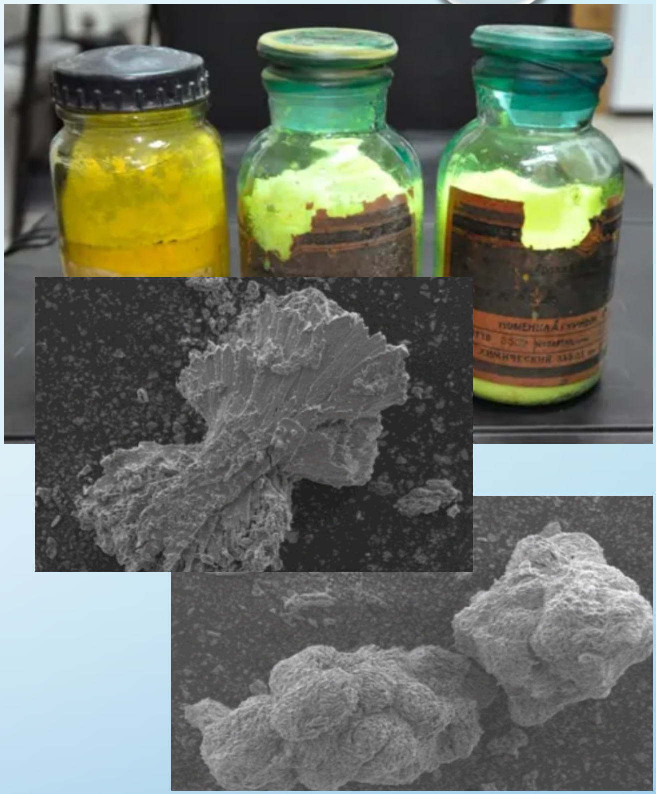
2.d. Sequentially Ordered Execution

Intentionally select the order that the data will be collected to allow for preliminary analysis of results, verifying strategy matches expectations

Key points:

- Good when overall data set needs to be defined BEFORE start of experiment
- Data often collected over an extended period of time
  Why not be strategic about what data to collect first?
- Original design is optimal for the global purpose.
  Then broken into blocks that address subgoals early.
- Often no net increase in cost,
  but allows for improved understanding from earlier analyses
- Need to adapt analysis to reflect data collection structure

NUCLEAR FORENSICS



Problem: When contraband nuclear materials are seized internationally, it is important to be able to establish its pedigree to understand who produced it with what process

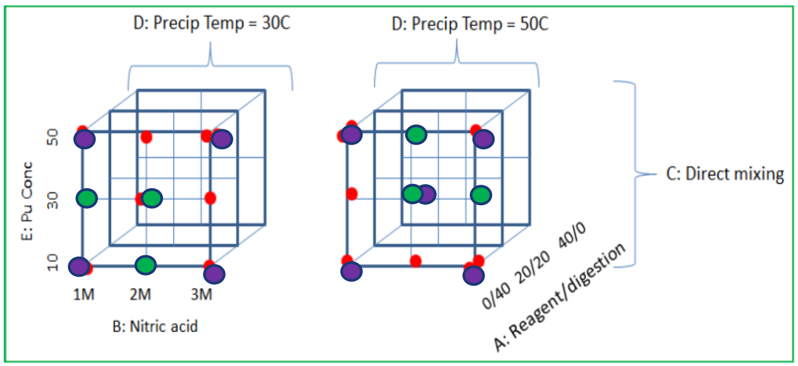
Solution: A designed experiment to produce materials across known range of production methods. Measure the *physical, chemical and morphological characteristics* of each material to compare to new samples and establish pedigree.

Technical Challenges:

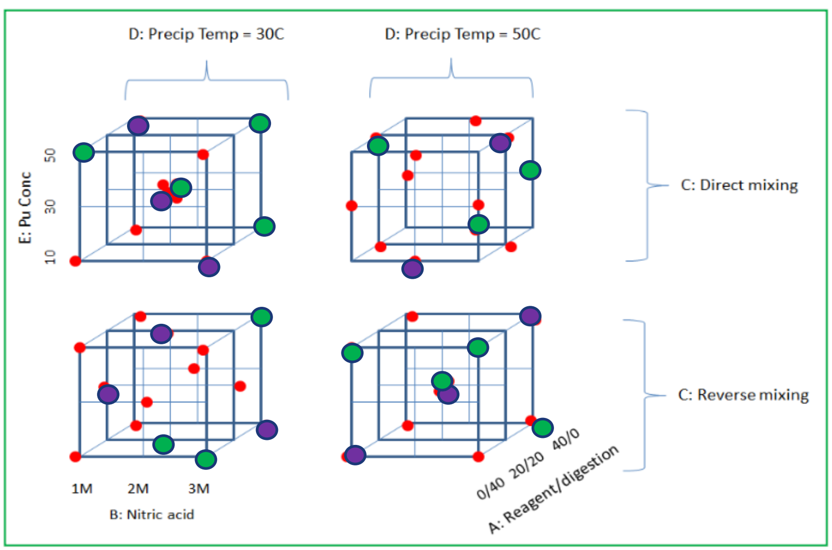
- Producing / measuring each run is very expensive and time-consuming
- Managing the amount of nuclear material produced constrained experiment size
- Limits of possible production ranges not precisely defined

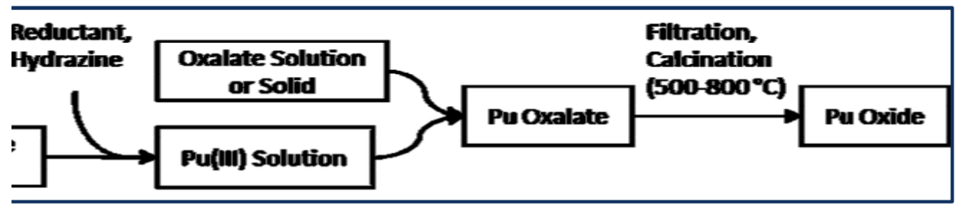
Example: Pu (III) Oxalate Bench-Scale Study

Oxalate Feed Solid



Oxalate feed in solution





Factors

Oxalate feed:
Solution (0.9M) & Solid

Reagent add./dig. timing:
0/40, 20/20, 40/0

Nitric acid concentration:
1M, 2M, 3M

Mixing sequence:
Direct & Reverse

Precipitation Temperature:
30°C & 50°C

Feed concentration of Pu:
10g/L, 30g/L, 50g/L

Stage 1: Test ranges of factors, see initial trends

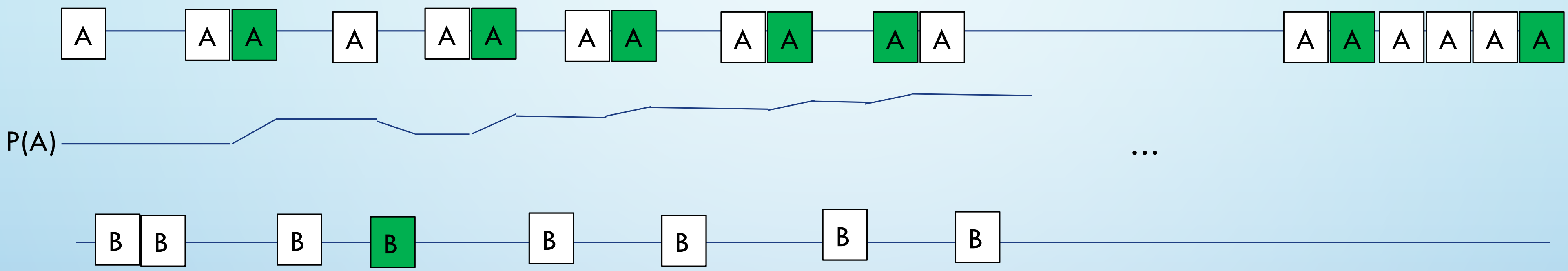
Stage 2: Estimate first order model

Stage 3: Estimate full model

DESIGNED DATA COLLECTION + BIG DATA

2.e. Adaptive Experiments:

Also know as Online Controlled Experimentation (OCE) or A/B or A/B/n testing



Key points:

- Exploit large volume of data possible, but often with tiny treatment effect
- Originally proposed to balance exploration vs exploitation
- Now much broader class of applications (estimate long-term or heterogeneous treatment effects, or average treatment effects in the presence of network interference (Larsen et al., 2022)

- Huge opportunities to help this new area to evolve

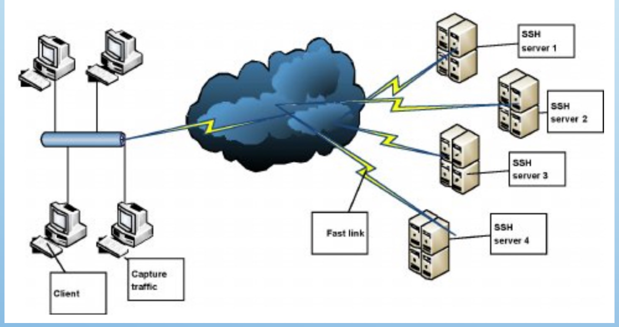
DESIGNED DATA COLLECTION + BIG DATA

2.f. Sampling to Downsize Big Data:

Selecting subsets of the big data for manageable analysis, checking for consistency of results

Key points:

- Many data sets have large "uninteresting" proportions, with a small "interesting" fraction
- What size is manageable / sufficient for valid results – dependent on the analysis selected?
- Ensuring representativeness of sample(s) for effective conclusions



NETWORK TRAFFIC

- High volume and velocity but high degree of redundancy
- 2 goals:
  - **Characterizing network traffic** (simple random sampling, reservoir sampling, sliding stratified sampling, chain sampling)
  - **Prioritize anomaly detection** (importance sampling)
- A single sampling strategy is unlikely to be successful at both goals

DESIGNED DATA COLLECTION + BIG DATA

2.g. Management and Storage of Big Data:

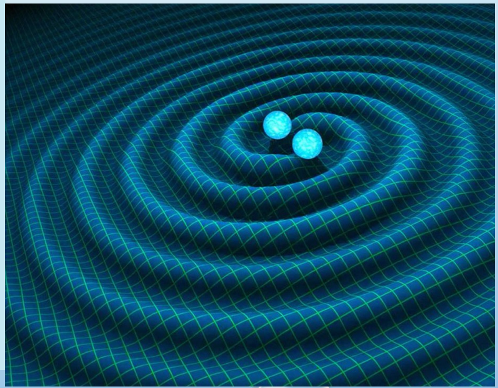
Effective storage of results when size or velocity of incoming data is unmanageable

Key points:

- For truly BIG data with high volume and high velocity, the constraint for managing the data may be storage, not analysis capability
- Focus here is on **triage of data** to divide into “keep” or “don’t keep”

STUDYING COSMIC CHIRPS

- Data from Very Large Array, near Socorro, NM
- Listening for a “Cosmic chirp” (anomalous sounds (eg. 2 black holes colliding)
- Trying to detect a very rare “Fast Radio Burst” (on the order of milliseconds)
- High volume streaming data arriving from 27 radio antennae
- Goal: triage data into 3 categories:
  - Definitely store (3%) for later analysis
  - Definitely trash (90%)
  - Temporary keep (7%) for later exploration with potential to store
- Requirements:
  - Able to store < 12%
  - Make decision in real time



OUTLINE

- Defining "Big Data" (BD) and "Designed Data Collection" (DDC)
- Situations with synergy between Big Data and Designed Data Collection

1. Before BD

a. Pre-planning to build understanding

b. Measurement assessment / data fidelity

2. Actual BD

a. Generating BD
b. Designed Simulation experiments
c. Sequential DoE
d. Sequentially ordered execution
e. Adaptive experiments
f. Sampling to downsize big data
g. Management and storage of BD

3. After BD

a. Validation of observational BD

b. Supplementing BD for new problems

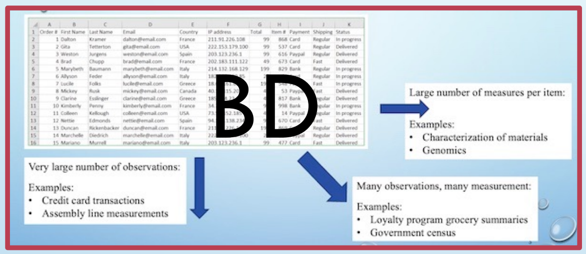
- Illustrative Examples
- Conclusions

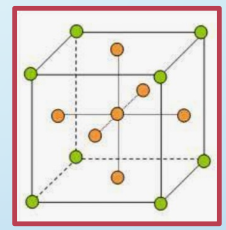
DESIGNED DATA COLLECTION + BIG DATA

3.a. Big Data Validated by Design of Experiments:

Analysis of Big Data suggests a solution or set of optimal settings.
DoE provides an opportunity to *validate* the result and *establish causality* of relationship

Observational data





Key points:

- Observational big data often reveal relationships suggestive of the desired outcome
- Running a designed experiment to verify the causal connection BEFORE implementing the solution can increase confidence and avoid waste.

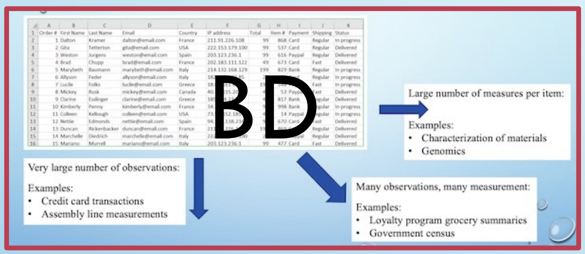
Eg. Factory production line

DESIGNED DATA COLLECTION + BIG DATA

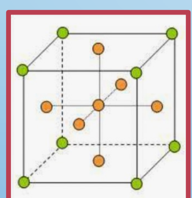
3.b. Big Data combined with Small Data:

Available Big Data summarize some of problem space of interest, but augmenting with new strategic data provides complementary sets of data to match intended goals

Observational legacy data



+

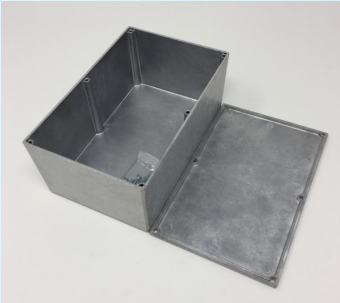


Combined data cover desired input space of interest

Key points:

- Original big data often not an exact match for question of interest
- Complementary data helps broaden understanding of relationship to better match study goals
- Cheaper than only DoE
- Better answer than big data alone

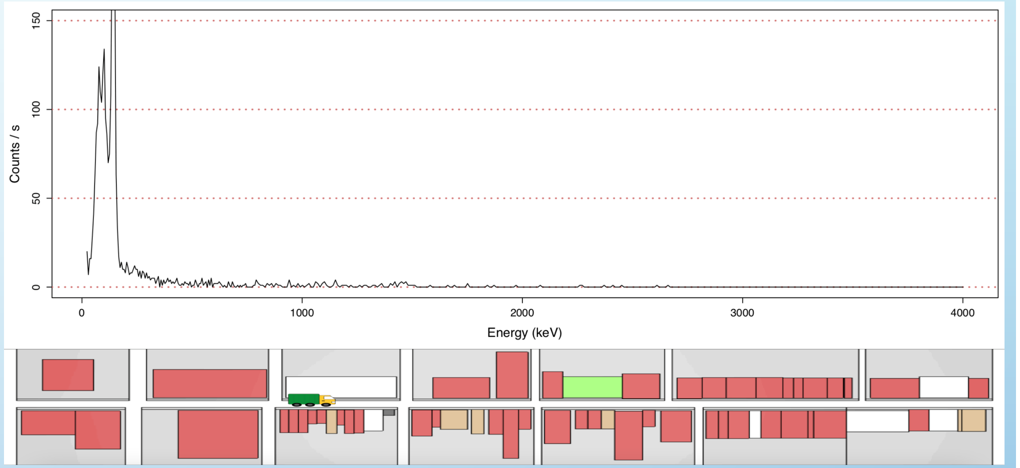
URBAN RADIATION DETECTION



After the data competition ended, there was interest in understanding how effective a new type of shielding would be for avoiding detection

Solution:

- Create an additional designed experiment that focuses on the new type of shielding across the other factors (different materials, backgrounds, detector speed, etc.)
- Run existing algorithms on new data to characterize performance



Allowed broader conclusions after competition closed

*May need caveats about how to interpret data from different sources

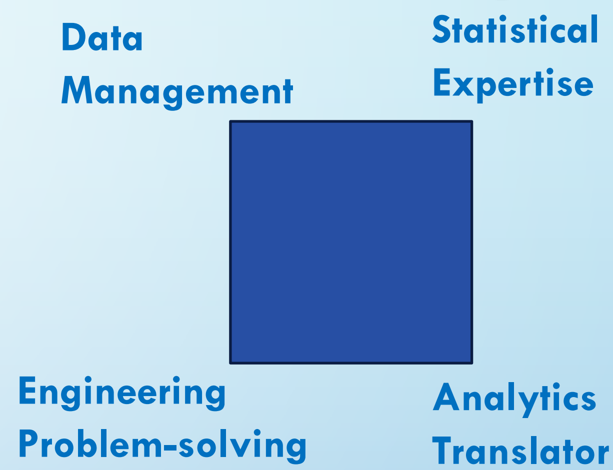
THERE WILL ALWAYS BE A ROLE FOR INTENTIONALLY COLLECTED SMALL DATA

Situations where small data makes sense:

- Stronger conclusions with causality and interpretability are needed
- Available resource restrictions
  - High cost ($) per observation
  - Time and/or labor-intensive
  - Safety constraints

CONCLUSIONS

- Big data are here to stay! So is designed data collection!
- Look for opportunities to be strategic about collecting or supplementing big data through designed data collection (before, during or after)
- Keys to success:
  - Clear definition of the goal of the study / analysis – will we know if we have been successful?
  - Collaboration with subject matter experts to incorporate important knowledge
  - The right statistical tools to answer the key questions
  - Problem-solving expertise to blend the right tools to find the right solution
- You don't have to do it all alone - build teams that create this collective expertise!



REFERENCES

1. Anderson-Cook, CM, Lu, L (2023) "Is Designed Data Collection Still Relevant in the Big Data Era? **Quality and Reliability Engineering International 39(4)** 1085-1119.
2. Lu, L, Anderson-Cook, CM, Robinson, TJ (2011) "Optimization of Designed Experiments Based on Multiple Criteria Utilizing a Pareto Frontier" ***Technometrics*** 53 353-365.
3. Lu, L, Anderson-Cook, CM (2021) "Strategies for Sequential Design of Experiments and Augmentation" **Quality and Reliability Engineering International** 37(5) 1740-1757.
4. Lu, L, Anderson-Cook, CM, Ahmed, T (2021) "Non-Uniform Space-Filling Designs" ***Journal of Quality Technology*** 53(3) 309-330.
5. Lu, L, Anderson-Cook, CM, Martin, M, Ahmed, T (2021) "Practical Choices for Space Filling Designs" **Quality and Reliability Engineering International** 38(3) 1165-1188.
6. Lu, L, Anderson-Cook, CM (2021) "Incorporating Uncertainty for Enhanced Leaderboard Scoring and Ranking in Data Competitions" **Quality Engineering** 33(2) 189-207.
7. Larsen, N, Stallrich, JW, Sengupta, S, Deng, A, Kohavi R, Stevens, NT, (2022) "Statistical Challenges in Online Controlled Experiments: A Review of A/B Testing Methodology" **arXiv:2212.11366**
8. Morgan, JC, Omell, B, Matuszewski, M, Miller, DC, Shah, MI, Benquet, C, Knarvik, ABN, de Cazenove, T, Anderson-Cook, CM, Ahmed, T, Tong, C, Ng, B, Bhattacharyya, D (2021) "Application of Sequential Design of Experiments (SDoE) to Large Pilot-Scaled Solvent-Based CO2 Capture Process at Technology Centre Mongstad (TCM)" **Proceedings of 15th International Conference on Greenhouse Gas Control Technology, GHGT-15**
9. Myers, RH, Montgomery, DC, Anderson-Cook, CM (2016). ***Response Surface Methodology: Process and Product Optimization Using Designed Experiments***. 4th Ed. New York: Wiley.
10. Stevens, NT, Anderson-Cook, CM (2019) "Design and Analysis of Confirmation Experiments" **Journal of Quality Technology** 51(2) 109-124.