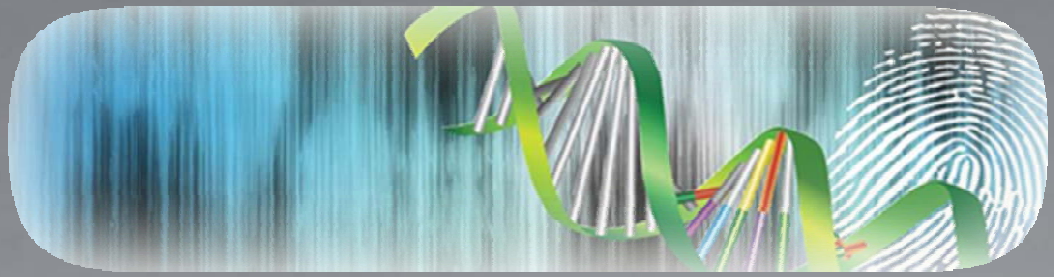
Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Two Strategies for Signature Discovery:
Small and Large Problem Spaces



ALEJANDRO HEREDIA-LANGNER

Kristin H. Jarman, Robert G. Ewing, Marvin G. Warner, Brett G. Amidan, Shari Matzner, Nathan A. Baker

May 2014

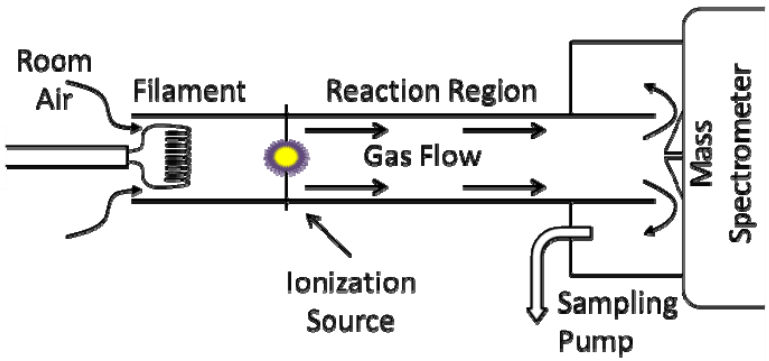
Motivation

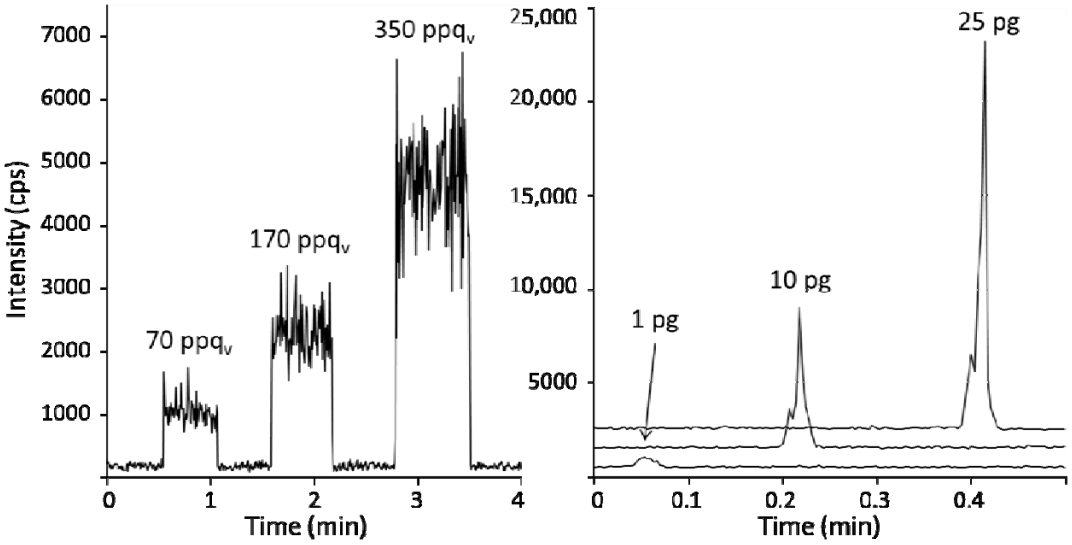
The Signature Discovery Initiative seeks to develop a systematic process for the rapid discovery of new signatures in any domain

- We present work related to two aspects of the signature discovery process
- Statistically designed experiments to explore small(er) spaces
- Fishing for features in large spaces using Genetic Algorithms

Small Problem Spaces
Statistically Designed Experiments

- Optimization of a signal from an atmospheric flow tube Mass Spectrometer (AFT-MS) sensor for RDX detection
- Detection of vapors from explosives is an important goal in screening
- It would allow for non-contact screening and sampling of large areas quickly
- Concentrations in air are considerably lower than saturated levels

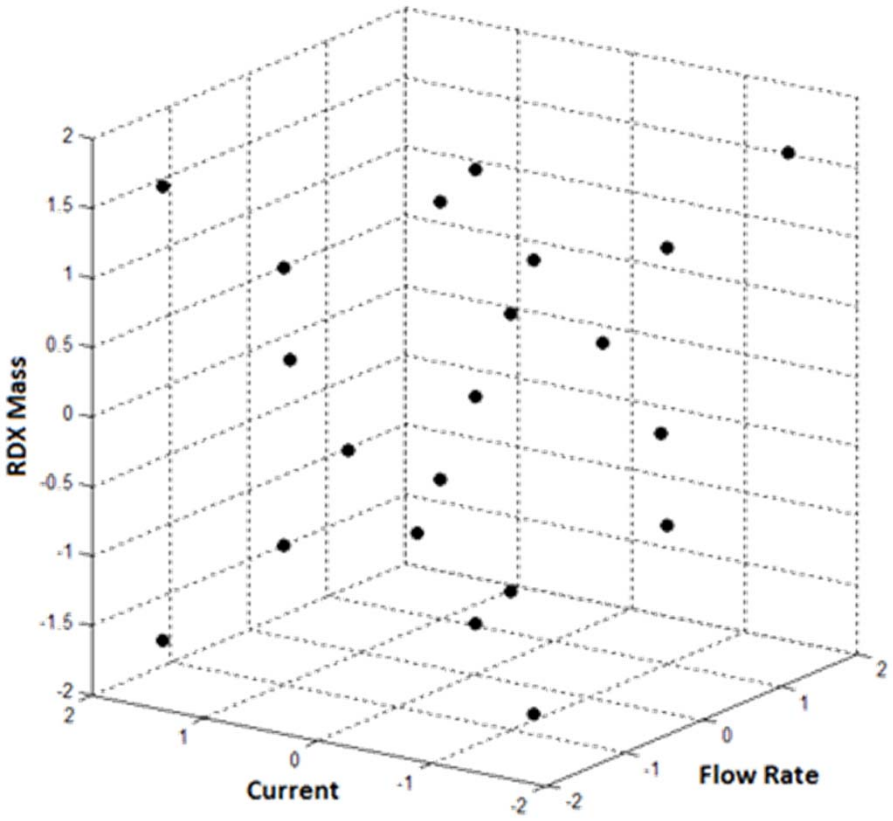




We need to understand the effect of experimental factors

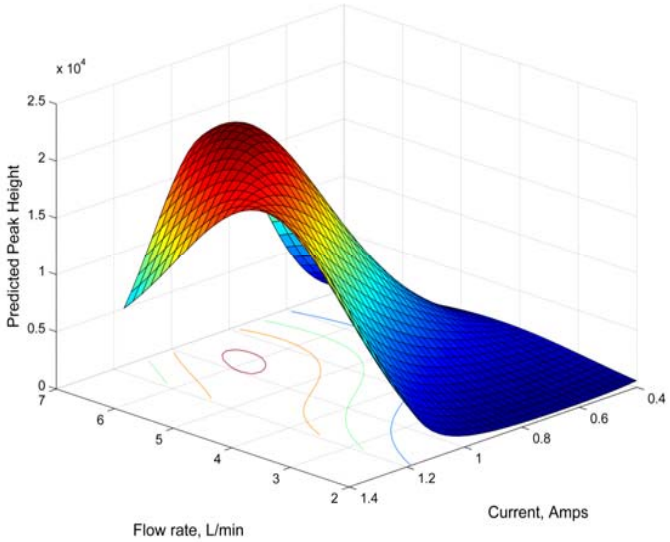
One of the experimental designs

- Three relevant factors
- Seek conditions that maximize peak height



Flow Rate (L/min)	Current (A)	RDX mass (pg)	Peak Height
4.5	0.8	25	9080
3	0.56	10	1060
3	1.04	40	2100
6	1.04	10	5900
4.5	0.8	25	9200
6	0.56	40	5040
3	0.56	40	2140
3	1.04	10	3220
6	1.04	40	20880
4.5	0.8	25	10100
6	0.56	10	1180
4.5	0.8	25	5740
2	0.8	25	720
4.5	0.8	25	7480
4.5	0.8	50	24000
4.5	0.4	25	1380
4.5	0.8	0.2	0
7	0.8	25	15420
4.5	1.2	25	18200
4.5	0.8	25	10220
7	0.4	50	2980
2	0.4	50	1500
2	1.2	50	5800
7	1.2	0.2	960
2	0.4	0.2	0
2	1.2	0.2	0

Prediction model and results



- Region of maximum signal is relatively broad and remains stable for varying RDX masses
- RDX amounts below 0.2 pg were difficult to distinguish from noise
- %RSD of signal remained 20-40%

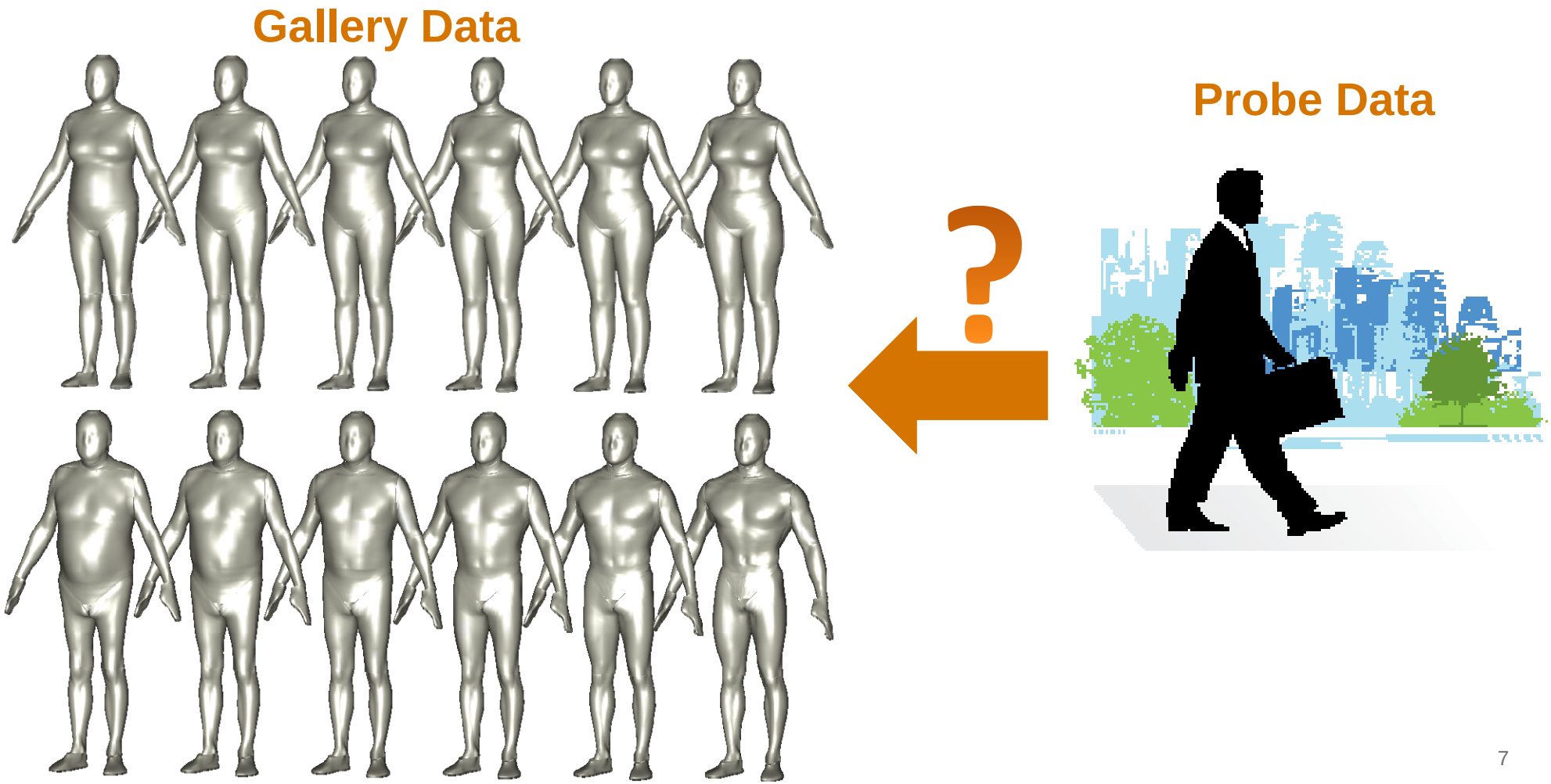
- Air samples collected from a shipping container showing RDX concentrations in the range of 1 to 50 ppqv
- Less than 5 min sampling time

What if we cannot use designed experiments?

- In many cases, we may not even know which variables to examine
- The problem space may be too large to consider optimization using highly fractionated factorial (or other types) of designs

Large Problem Spaces
A different approach needed

- Consider a re-identification problem: Recognizing a person (or object) previously observed, and for whom some information is available

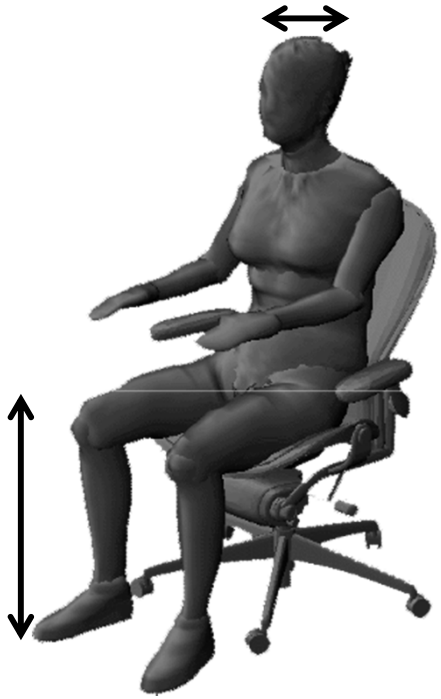


Re-identification when Gallery and Probe Sets Differ

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

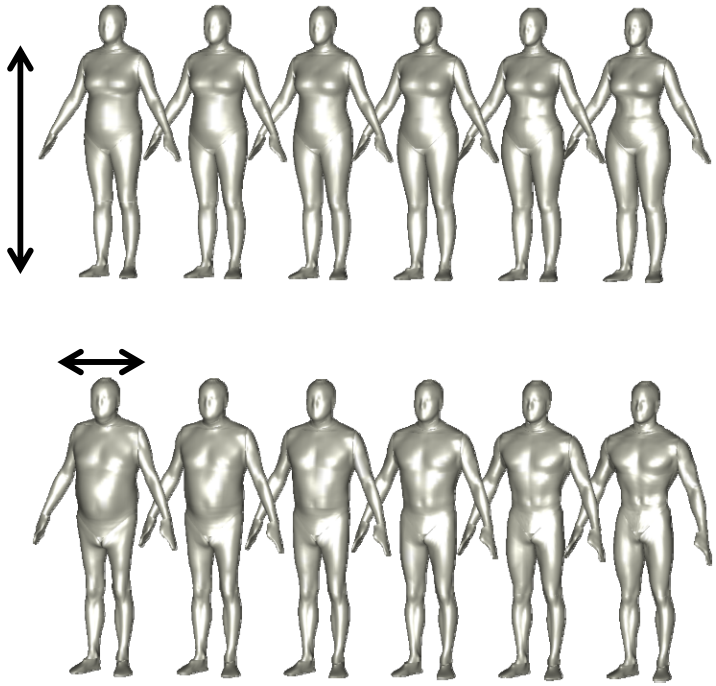
CAESAR database: 2378 individuals, data for seated & standing positions

Probe Data



16 Features Measured

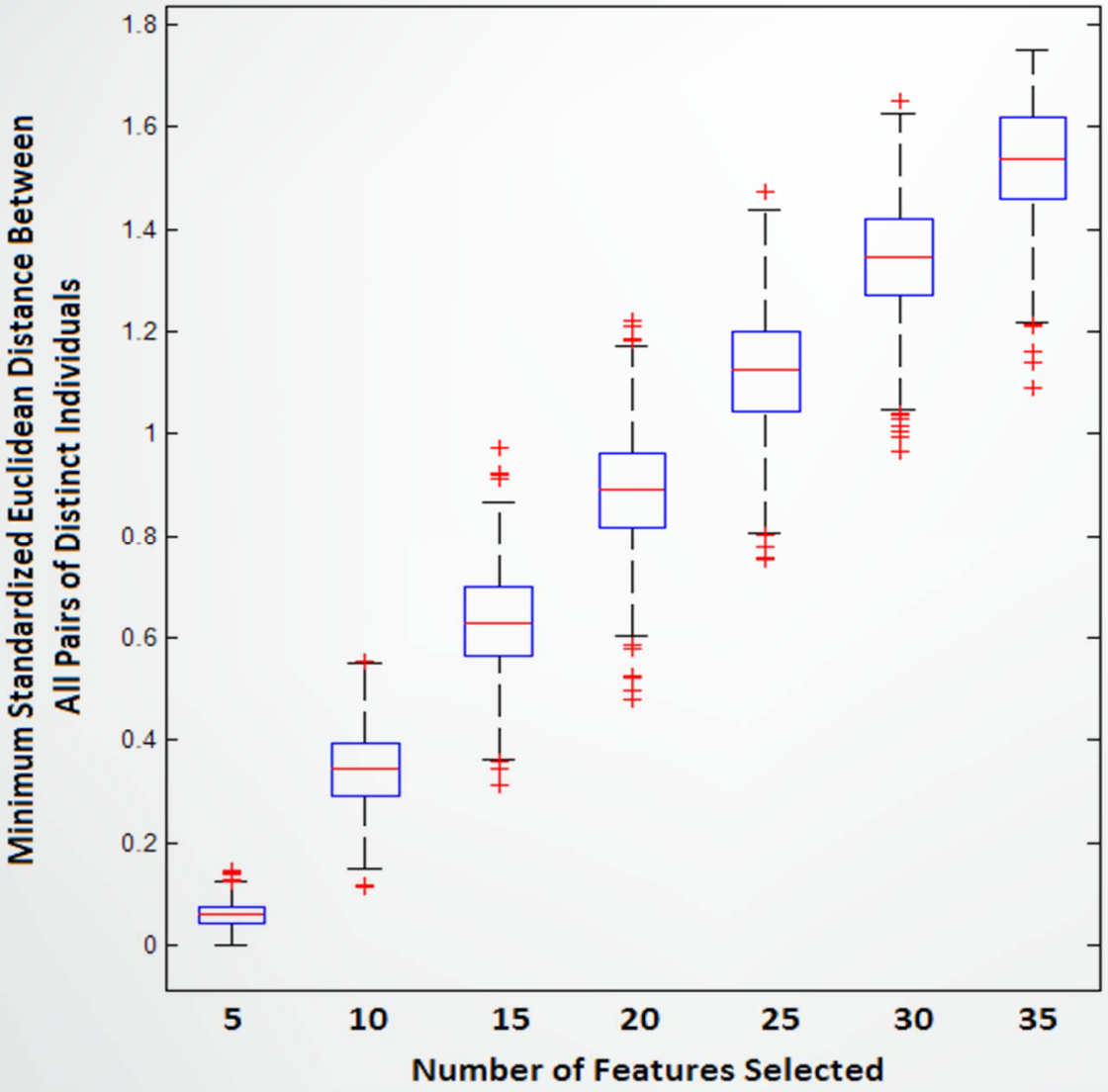
Gallery Data



43 Features Measured

- How do we re-identify a person in our gallery if all we have is probe set data?
- Solution usually expressed as *Rank*

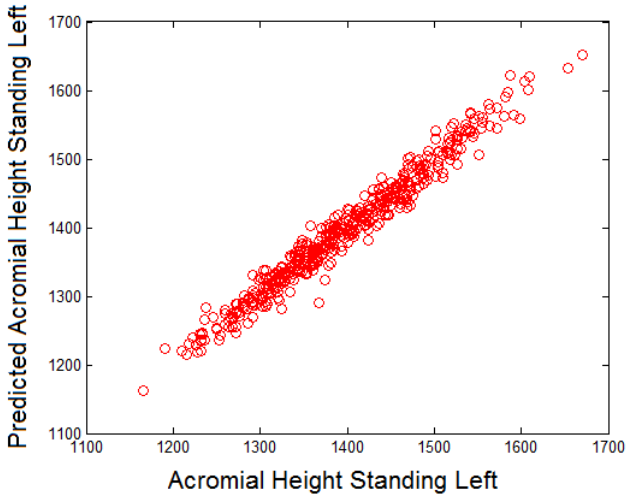
Evidence indicates very few features are needed to unambiguously re-identify an individual if gallery set data is available and remains unchanged



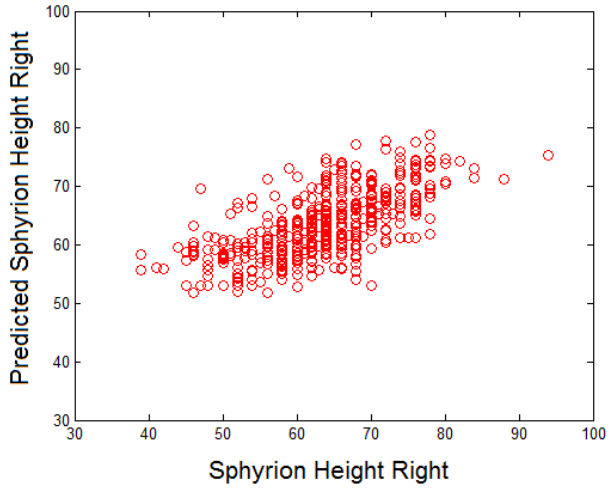
Modeling

Gallery Set

Feature Name	Feature Name	Feature Name
Acromial Ht Stand Lt	Bitrochant.Brth Stand	Malleolus Med Rt
Acromial Ht Stand Rt	Bustpoint Brth	Neck Ht
Acromion-Radiale Len Lt	Cervicale Ht	Radiale-Stylion Lt
Acromion-Radiale Len Rt	Chest Ht Stand	Radiale-Stylion Rt
Ankle Ht Lt Malleolus,Lat Lateral)	Elbow Ht Stand Lt	Sellion Suprasternion
Ankle Ht Rt Malleolus.Lat Lateral)	Elbow Ht Stand Rt	Sleeve Outseam Lt
Arm Inseam Lt	Foot Brth Lt	Sleeve Outseam Rt
Arm Inseam Rt	Foot Brth Rt	Sphyrion Ht Lt
Axilla Ht Lt	Infraorbitale Ht Lt Stand	Sphyrion Ht Rt
Axilla Ht Rt	Infraorbitale Ht Rt Stand	Suprasternale Ht
Biacromial Brth	Inter-pupillary Dst	Trochanterion Ht Lt
Biceps Brth	Interocye Dst Stand	Trochanterion Ht Rt
Bigonial Brth	Knee Ht Stand Lt	Waist Back
Bispinous Brth	Knee Ht Stand Rt	
Bitragion Brth	Malleolus Med Lt	



$$f(\vec{x})$$



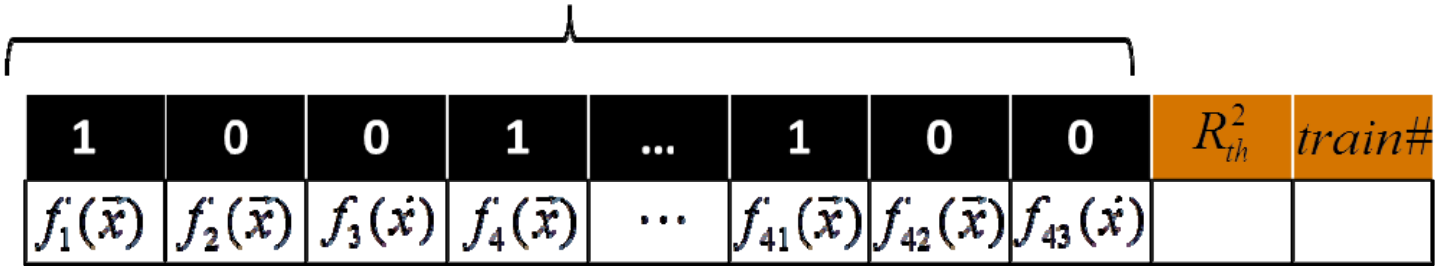
Probe Set

Feature Name
Acromial Ht Sit Lt
Acromial Ht Sit Rt
Bi-lateral Femoral Epicondyle Brth Sit
Bi-lateral Humeral Epicondyle Brth Sit
Bitrochanteric Brth Sit
Buttock to Trochanter Lth
Femoral Epicondyle Lat to Malleolus Lat Lt
Femoral Epicondyle Lat to Malleolus Lat Rt
Infraorbitale Ht Sit Lt
Infraorbitale Ht Sit Rt
Trochanter to Femoral Epicondyle Lat Lt
Trochanter to Femoral Epicondyle Lat Rt
Trochanter to Seated Surface Lt
Trochanter to Seated Surface Rt
Elbow Ht Sit Lt
Elbow Ht Sit Rt

Genetic Algorithms (GA)

- Optimization approach loosely based on Darwin's theory of evolution
- Useful to explore large problem spaces
- No guarantee of optimality, computationally expensive

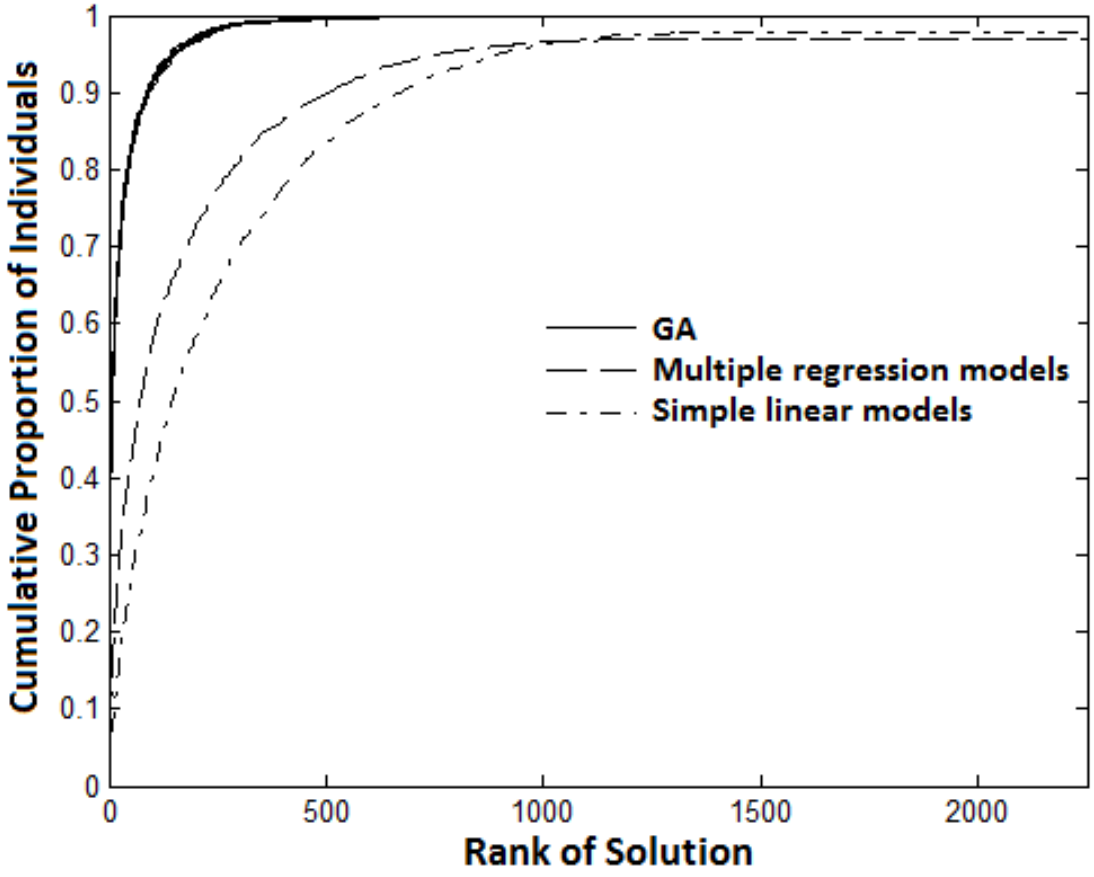
Models for Gallery features, use/not use(1/0)



- Recombination, selection, mutation – apply until convergence
- Running algorithm multiple times is not a bad idea

Results

- Multiple regression: All features, all individuals
- Simple linear regression: All features, all individuals
- GA
  - R2 threshold ~0.87
  - ~300-400 individuals training set
  - ~12-20 Gallery features selected



Feature Name	Feature Name
Acromial Ht Stand Lt	Infraorbitale Ht Lt Stand
Acromial Ht Stand Rt	Infraorbitale Ht Rt Stand
Acromion-Radiale Length Lt	Knee Ht Stand Rt
Acromion-Radiale Length Rt	Sleeve Outseam Len Lt
Axilla Ht Lt	Trochanterion Ht Lt
Bitrochanteric Brth Stand	Trochanterion Ht Rt

Problem Space searched by GA has ~9×10^{12} solutions

Final remarks

- Designed experiments and GA both change factor levels simultaneously
- Both estimate factor effects – explicitly or implicitly
- Both guide the search to an optimal region
- Statistically Designed Experiments – Search is more systematic and thorough & allows for testing of specific hypotheses
- GA are very flexible and a good choice when deterministic methods cannot be used

Thank you!

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Learn more about the Signature Discovery Initiative at http://signatures.pnnl.gov or contact Nathan Baker nathan.baker@pnnl.gov

- Comments for Alejandro (Alejandro.Heredia-Langner@pnnl.gov)

The research described in this presentation is part of the Signature Discovery Initiative at Pacific Northwest National Laboratory (PNNL). It was conducted under the Laboratory Directed Research and Development Program at PNNL, a multi-program national laboratory operated by Battelle for the U.S. Department of Energy.