Small Statistics Big Data Curriculum

C.Foster
05 Oct 2018

Historic

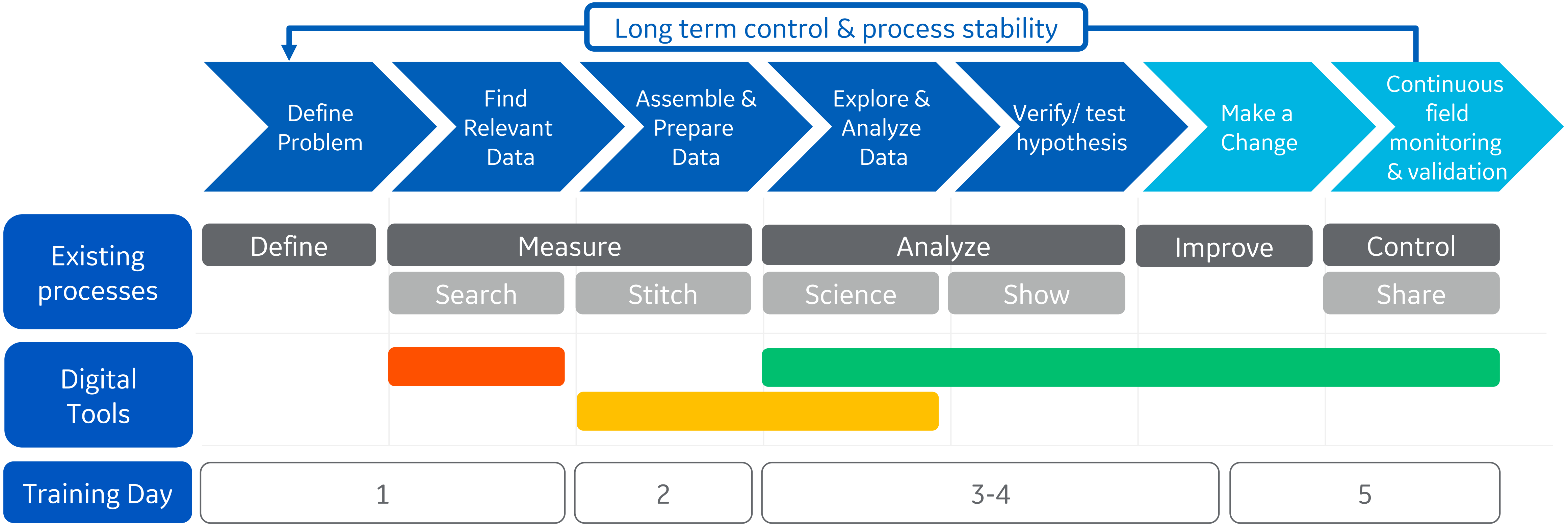
- Database access controlled and complex (Instance / Schema / Table)
  Software Engineer to organize and query data.
- Analysis in proprietary code (FEA/CFD Solvers)
  Analysis Engineer to setup and run solvers
- Statistical Analysis in special programs (JMP/Minitab)
  Statistician run analysis
- Publish results (HTML/Java)
  Programmer to create report or dashboard

No common process or link

Process & Tools

- Get Data
- Analyze and model
- Publish results

Target Audience: Engineers / Technicians / Scientists who are not primarily digital.



Long term control & process stability

Define Problem

Find Relevant Data

Assemble & Prepare Data

Explore & Analyze Data

Verify/ test hypothesis

Make a Change

Continuous field monitoring & validation

Existing processes

Define

Measure

Search

Stitch

Analyze

Science

Show

Improve

Control

Share

Digital Tools

Training Day

1

2

3-4

5

Query →	Norm / Imputing →	Aggregate →
	Data Engineering Process	
First N - Sampling	Z-Score 0 Replacement	Mean
	Applied Statistics Process	
Stratified Sampling	Min/Max Robust Scaling Imputation	Median IQR

→ Charts →	Tests →	Models →
	Data Visualization Process	
Scatter / Bar / Line / Combined	Color by... / Line by... / Shape by...	Trendlines
	Applied Statistics Process	
Quantile / Box Plot	Z-t Test / χ2 / FDR	Linear and Logistic Regression Classification and Regression Trees

Decision → Implementation → Control / Capability

Programming Process

Threshold

Recode and Test initial

Run Reports

Applied Statistics Process

Bootstrapped threshold

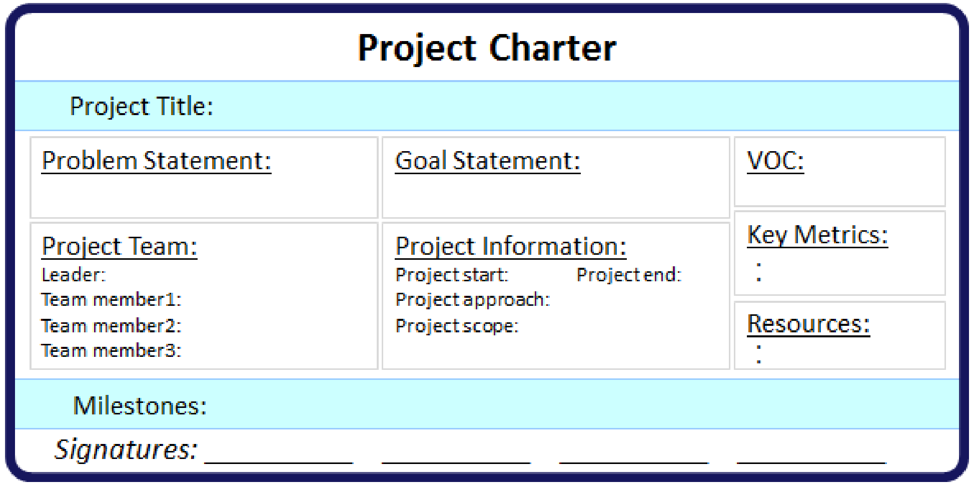
DOE Based Validation

Anomaly Detection / Residual Analysis

Integrate the Statistical / Data Science methods and the Tools and Process Activities

Four Examples

Define the goal in terms of a measurable outcome



Project Charter

Project Title:

Problem Statement:	Goal Statement:	VOC:
Project Team: Leader: Team member1: Team member2: Team member3:	Project Information: Project start: Project end: Project approach: Project scope:	Key Metrics: : Resources: :

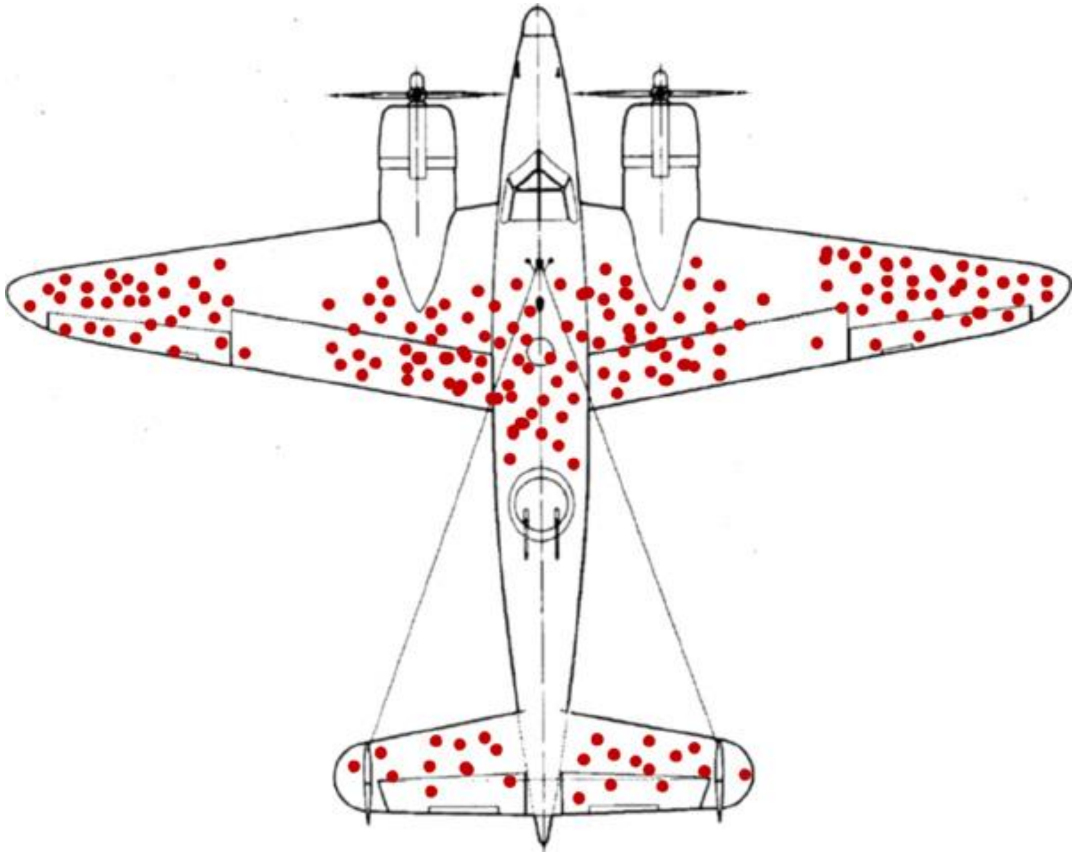
Milestones:

Signatures: __________ __________ __________ __________

1. N≠All	Something is missing
2. N is big	Statistical assumptions breakdown.
3. Act	What is wrong with the measurements? Are we solving the right problem?
4. Feedback	How will the system react and how will it effect the analytic?

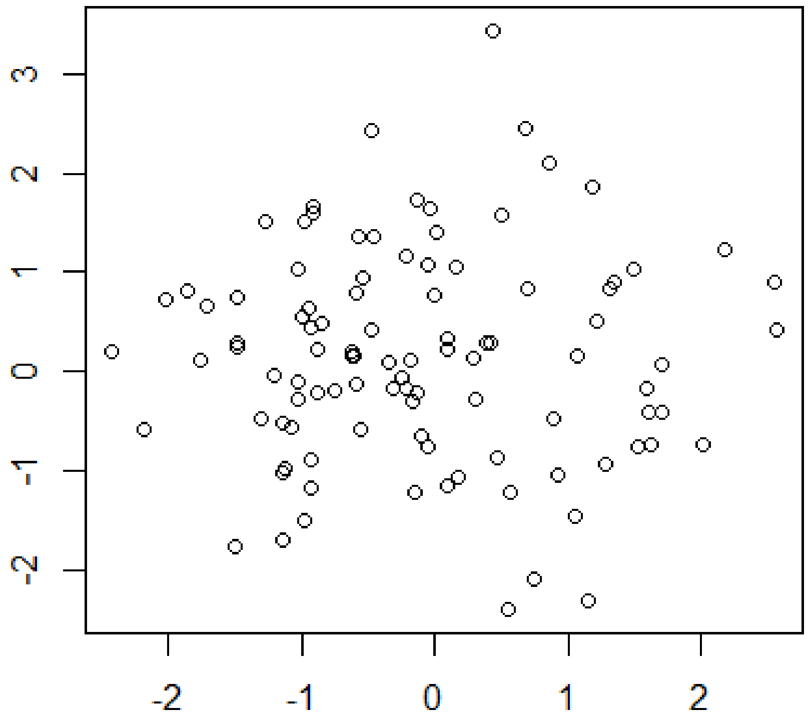
Models / Ownership

N≠All



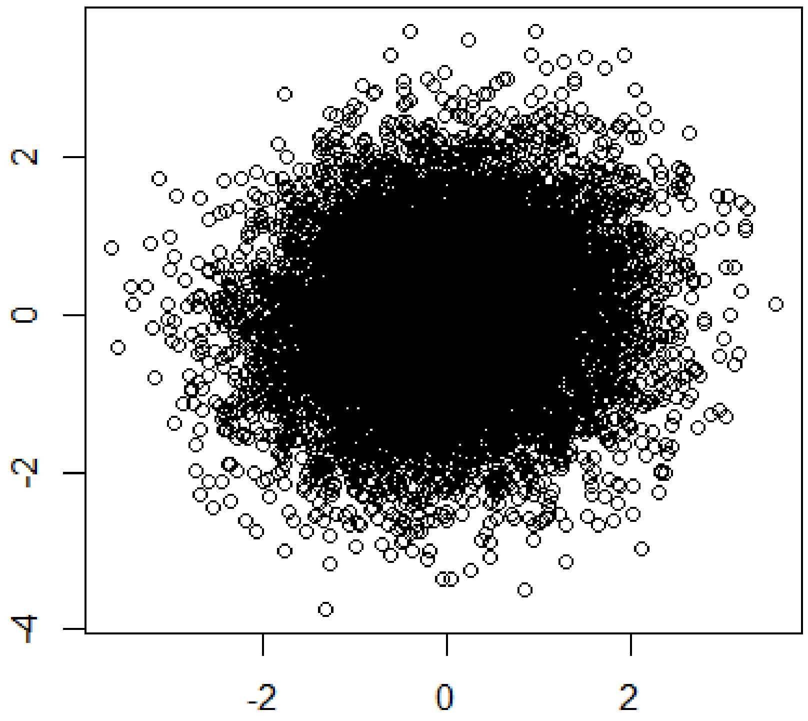
N is big

R = .1



100 points p=0.936

Not Significant



10000pts p<2e-16

VERY Significant

Act

Your task is to turn over as few cards as possible to verify whether the following statement is true.

Every card with a vowel on one side has an even number on the other.

A B 2 3

Which order would you turn over the cards?

1. AB23
2. AB32
3. A23B
4. A2B3
5. A3B2
6. A32B

Your task is to turn over as few cards as possible to verify whether the following statement is true.

Every part with a crack has corrosion as a cause

Part: Cracked Gear

Part: Worn Gear

Cause: Corrosion

Cause: Erosion

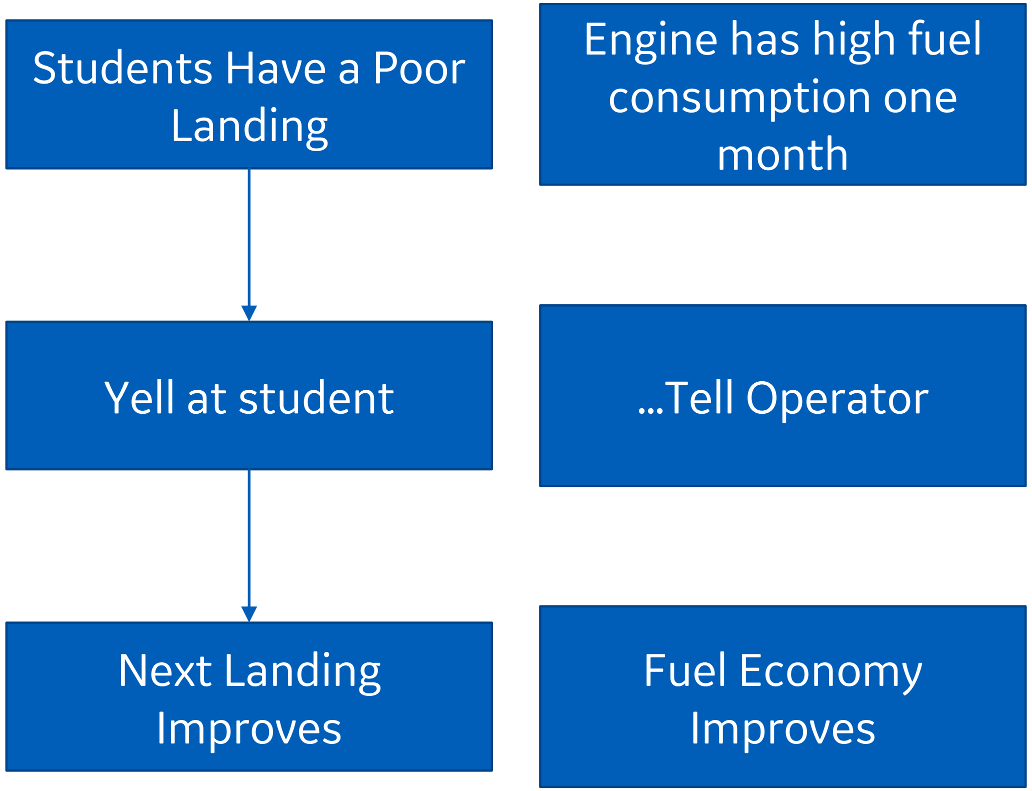
Which order would you turn over the cards?

Could a cause be crack? i.e. progressive damage.

Act

Chasing Noise...

All data are noisy



Students Have a Poor Landing

Yell at student

Next Landing Improves

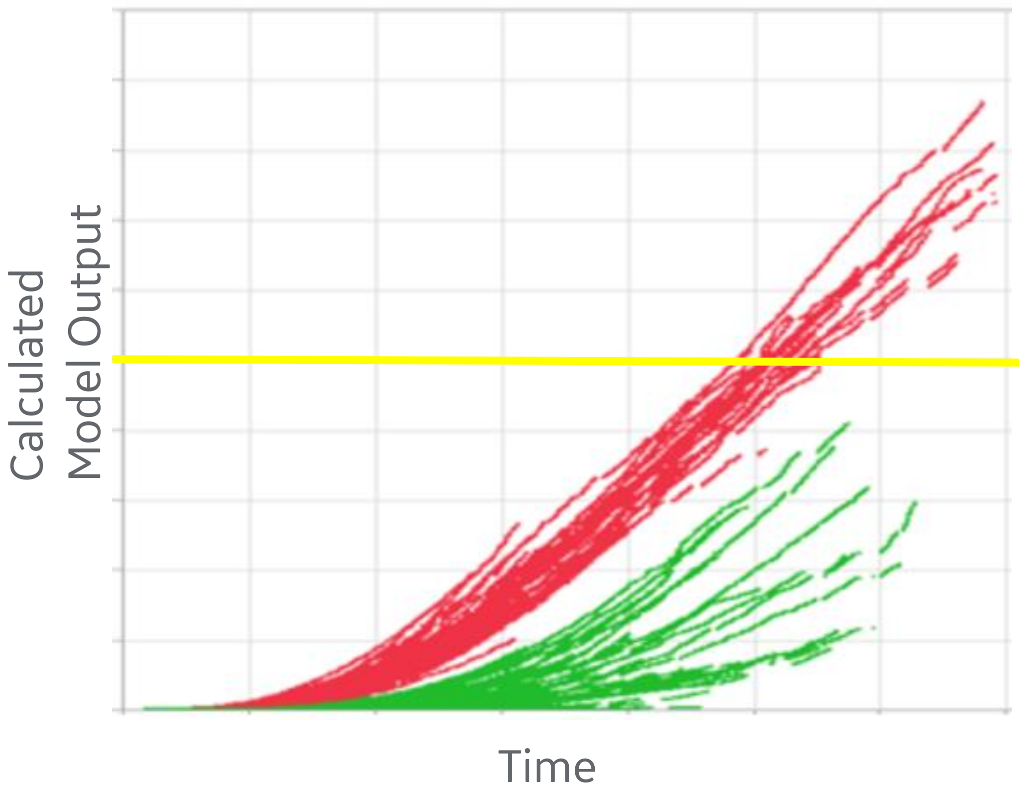
Engine has high fuel consumption one month

...Tell Operator

Fuel Economy Improves

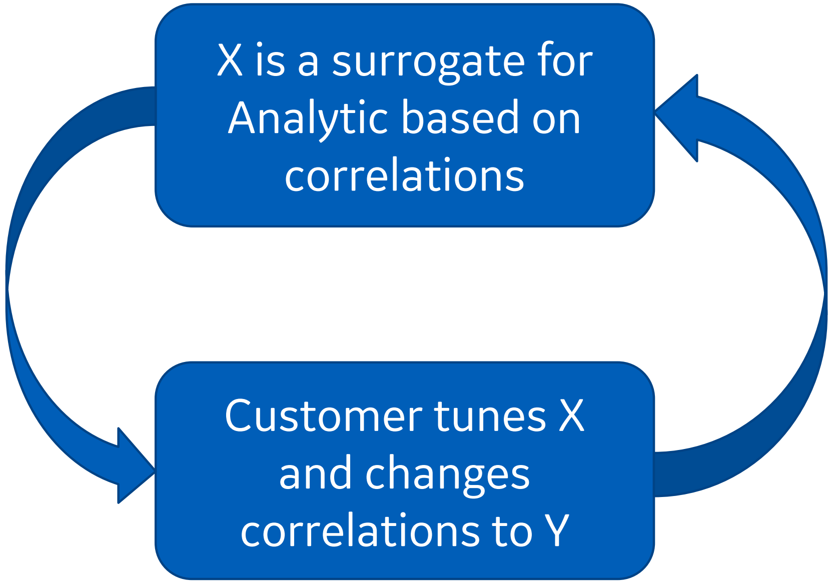
Feedback

Interactions with the System



If customer is charged more when they are here

They try to run here
Even if it does not change risk / cost

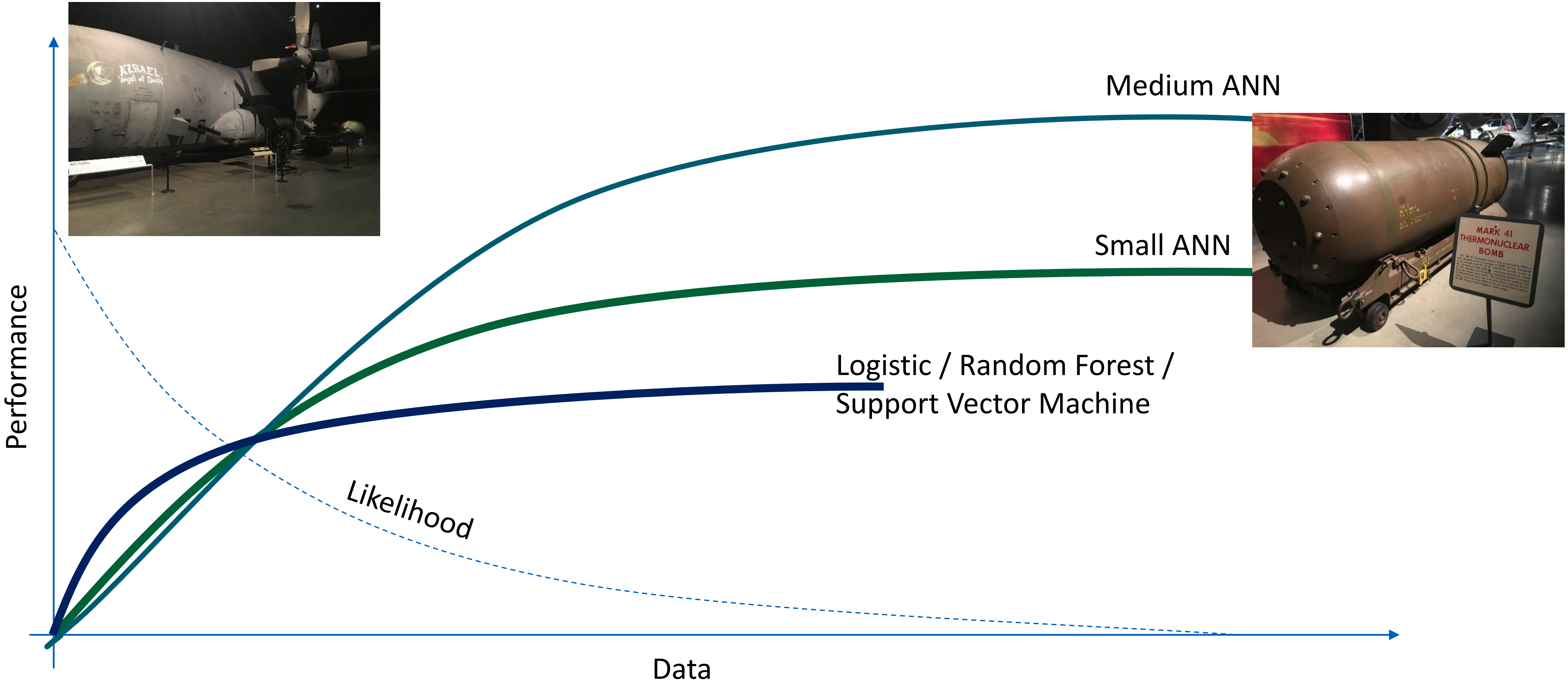

X is a surrogate for Analytic based on correlations

Customer tunes X and changes correlations to Y

Control is critical

Not predicting weather

Methods and Big Data



Gaussian Process
Random Forests

General use model

Pretty well most of the time

Fails obviously

Includes Uncertainty

Bayesian Calibration of Computer Models
https://rss.onlinelibrary.wiley.com/doi/10.1111/1467-9868.00294

Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife
http://jmlr.org/papers/volume15/wager14a/wager14a.pdf

The Analyst:

Owns the proposal

Owns knowing the data - and quality

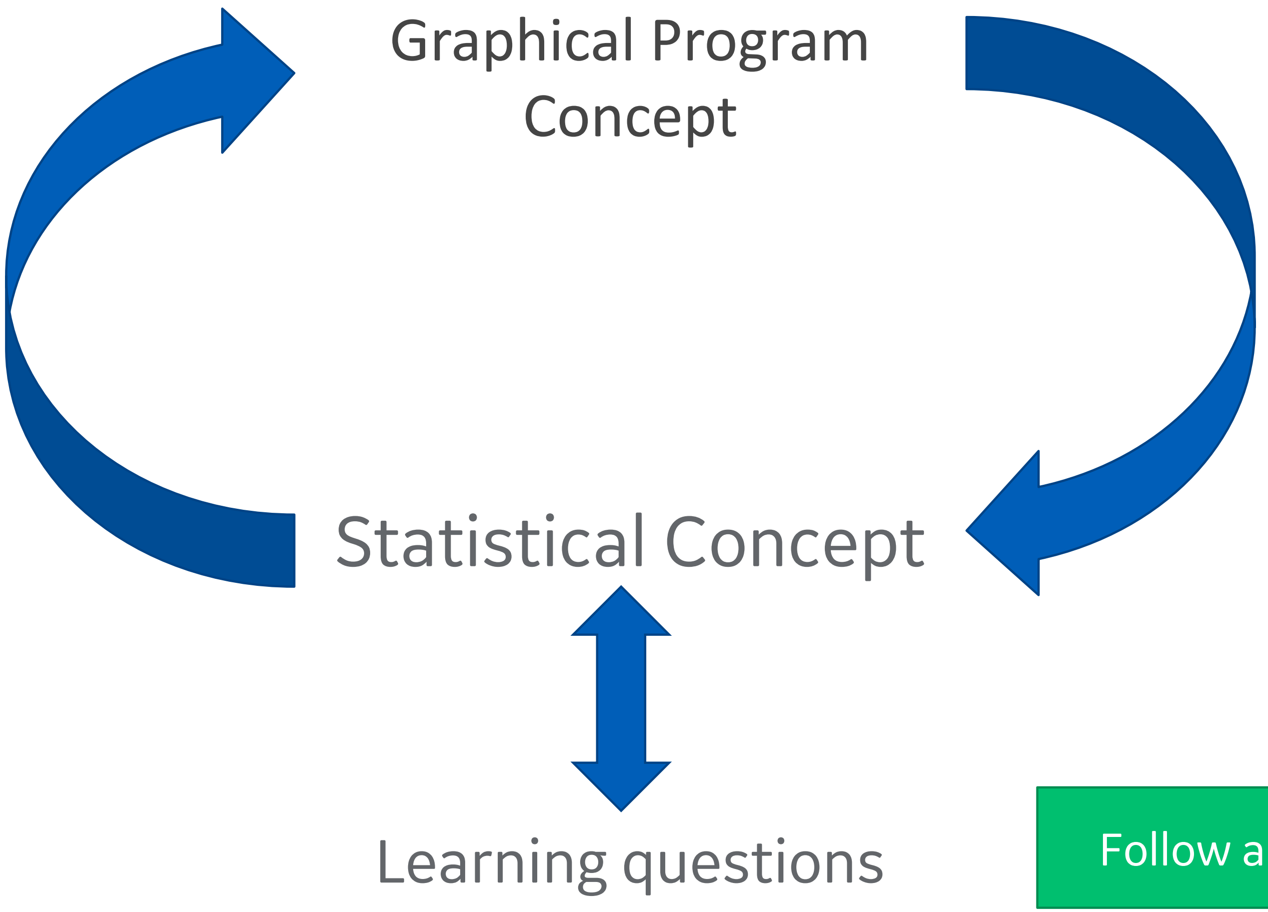
Gets the data from the lake

Stitches the sources

Creates the visualization

Integrate objectives

Training Process



Graphical Program Concept

Statistical Concept

Learning questions

Follow an available text

https://www.openintro.org/stat/textbook.php

Integrated Example

P-Value Sample

A poll by the National Sleep Foundation found that students on average sleep 7 hours per night.

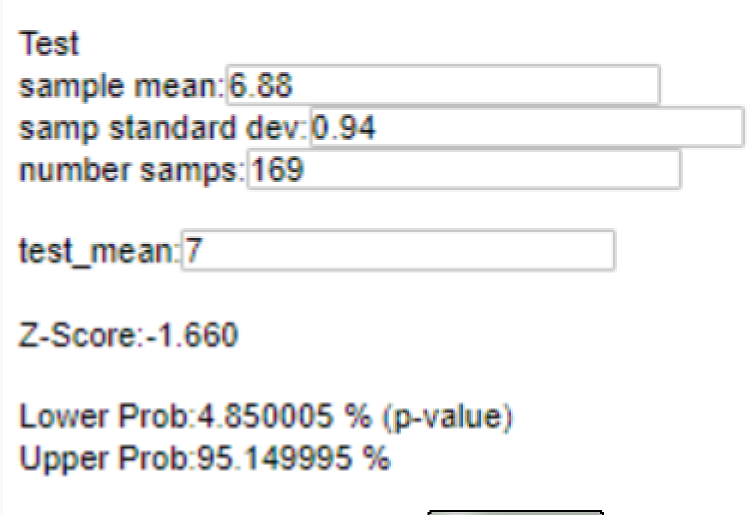
A sample of 169 students sleep for one night had an average of 6.88 hours and a standard deviation of 0.94 hours.

Assuming that this is a representative random sample, is there sufficient evidence to reject the null hypothesis that students on average sleep 7 hours per night?

Standard Error = sd/sqrt(samp)

What is the p-value for this hypothesis test?

What caveats would you add?



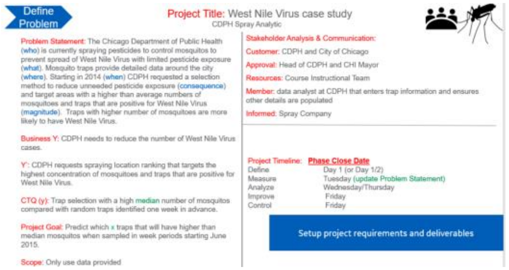
```
Test
sample mean:6.88
samp standard dev:0.94
number samps:169

test_mean:7

Z-Score:-1.660

Lower Prob:4.850005 % (p-value)
Upper Prob:95.149995 %
```

https://www.openintro.org/stat/textbook.php

Follow a case study

Student Case Study – West Nile Virus (with parallels to fleet management)

Setup project requirements and deliverables



Define Problem

Project Title: West Nile Virus case study
CDPH Spray Analytic

Problem Statement: The Chicago Department of Public Health (who) is currently spraying pesticides to control mosquitos to prevent spread of West Nile Virus with limited pesticide exposure (what). Mosquito traps provide detailed data around the city (where). Starting in 2014 (when) CDPH requested a selection method to reduce unneeded pesticide exposure (consequence) and target areas with a higher than average numbers of mosquitoes and traps that are positive for West Nile Virus (magnitude). Traps with higher number of mosquitoes are more likely to have West Nile Virus.

Business Y: CDPH needs to reduce the number of West Nile Virus cases.

Y': CDPH requests spraying location ranking that targets the highest concentration of mosquitoes and traps that are positive for West Nile Virus.

CTQ (y): Trap selection with a high median number of mosquitos compared with random traps identified one week in advance.

Project Goal: Predict which x traps that will have higher than median mosquitos when sampled in week periods starting June 2015.

Scope: Only use data provided

Stakeholder Analysis & Communication:

Customer: CDPH and City of Chicago

Approval: Head of CDPH and CHI Mayor

Resources: Course Instructional Team

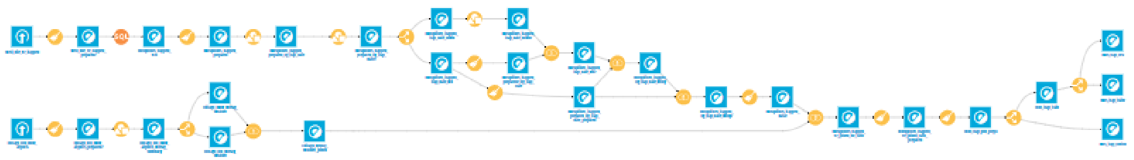
Member: data analyst at CDPH that enters trap information and ensures other details are populated

Informed: Spray Company

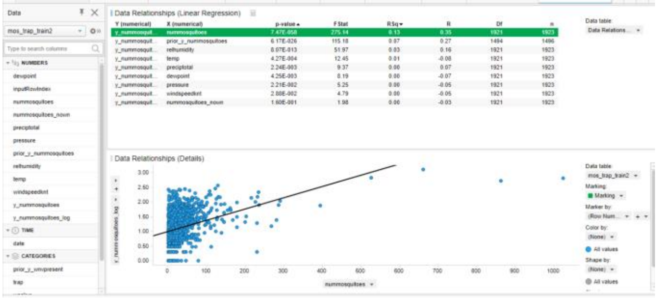
Project Timeline:	Phase Close Date
Define	Day 1 (or Day 1/2)
Measure	Tuesday (update Problem Statement)
Analyze	Wednesday/Thursday
Improve	Friday
Control	Friday

Setup project requirements and deliverables

Find and manipulate data



Visualize and build models

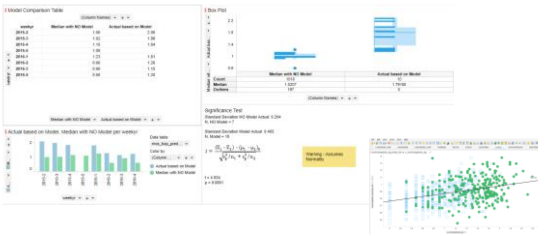


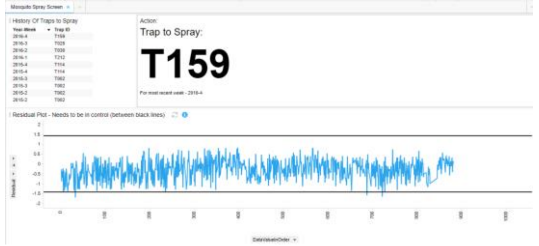
Data Relationships (Linear Regression)

Y (numerical)	X (numerical)	p-value	F Stat	RSq	R	Df	n
y_nummosquit...	nummosquitoes	7.47E-058	275.14	0.13	0.35	1921	1923
y_nummosquit...	prior_y_nummosquitoes	6.17E-026	115.18	0.07	0.27	1494	1496
y_nummosquit...	relhumidity	8.07E-013	51.97	0.03	0.16	1921	1923
y_nummosquit...	temp	4.27E-004	12.45	0.01	-0.08	1921	1923
y_nummosquit...	preciptotal	2.24E-003	9.37	0.00	0.07	1921	1923
y_nummosquit...	dewpoint	4.25E-003	8.19	0.00	-0.07	1921	1923
y_nummosquit...	pressure	2.21E-002	5.25	0.00	-0.05	1921	1923
y_nummosquit...	windspeedknt	2.88E-002	4.79	0.00	-0.05	1921	1923
y_nummosquit...	nummosquitoes_nown	1.60E-001	1.98	0.00	-0.03	1921	1923

Data Relationships (Details)

Validate and Publish Dashboard





Mosquito Spray Screen

History Of Traps to Spray

Year-Week	Trap ID
2016-4	T159
2016-3	T028
2016-2	T008
2016-1	T212
2015-4	T114
2015-4	T114
2015-3	T062
2015-3	T062
2015-2	T062
2015-2	T062

Action:

Trap to Spray:

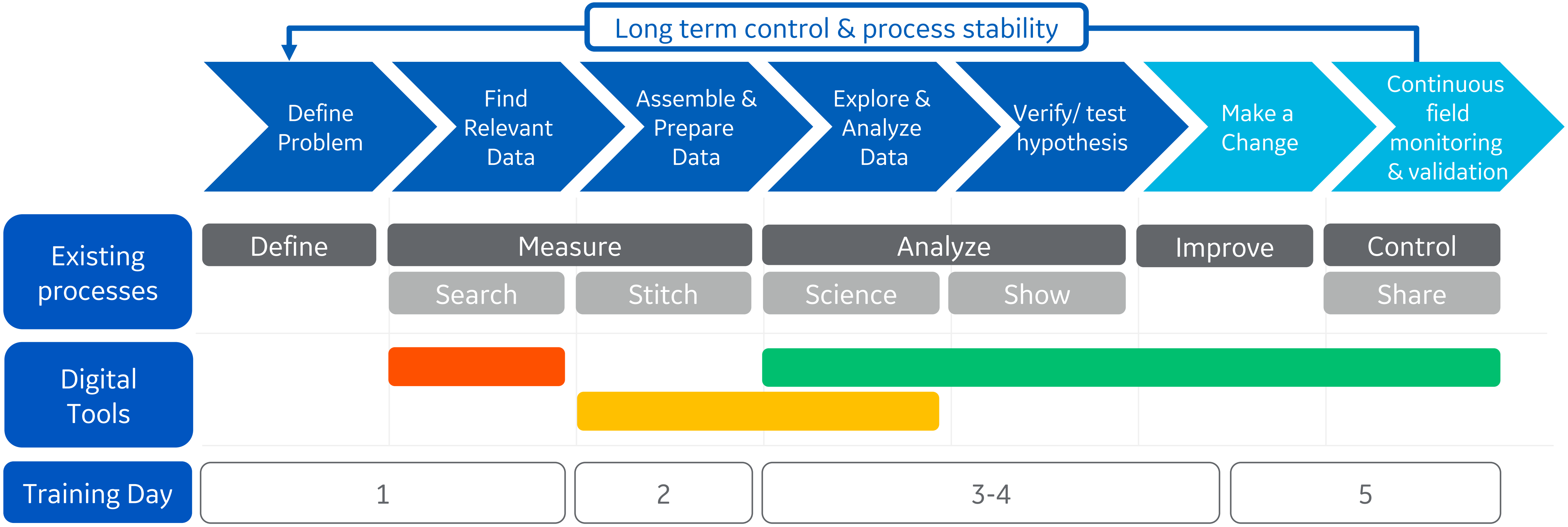
T159

For most recent week - 2016-4

Residual Plot - Needs to be in control (between black lines)

Complete process with tools

- Engineers and users are capable of entire analytic process
- Basic statistics knowledge and capability
- Integrated with current improvement process



Long term control & process stability

Define Problem

Find Relevant Data

Assemble & Prepare Data

Explore & Analyze Data

Verify/ test hypothesis

Make a Change

Continuous field monitoring & validation

Existing processes

Define

Measure

Search

Stitch

Analyze

Science

Show

Improve

Control

Share

Digital Tools

Training Day

1

2

3-4

5