Final Project – Group #48

Topic E - Predicting Project Success

Visualizing Project Success using Iterative Bayesian Estimation

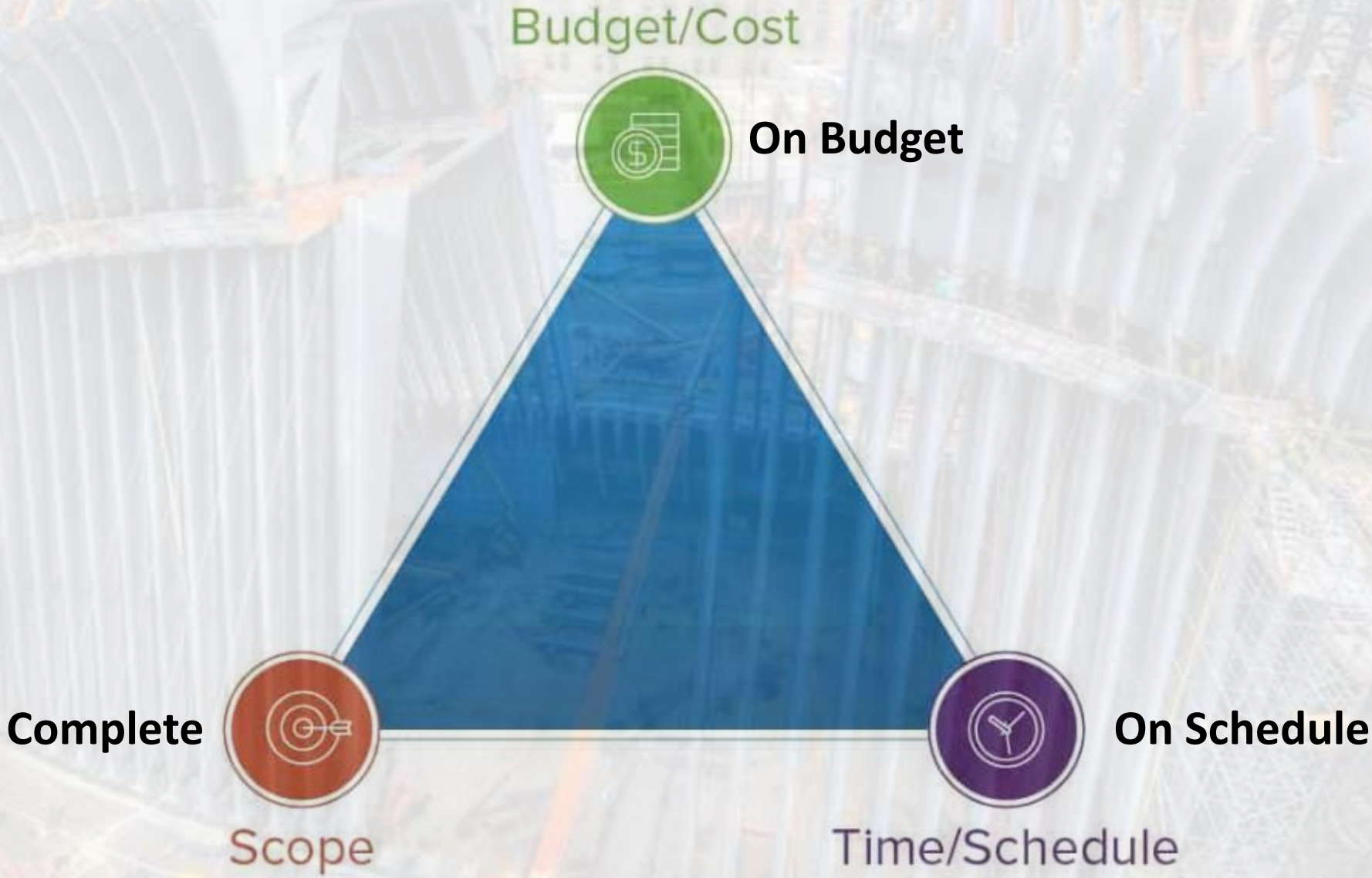
Michael Lee **Micah Nickerson** **Daniel Olal**

HARVARD
School of Engineering
and Applied Sciences

CSCI-E-109B - Advanced Topics in Data Science - Spring 2020

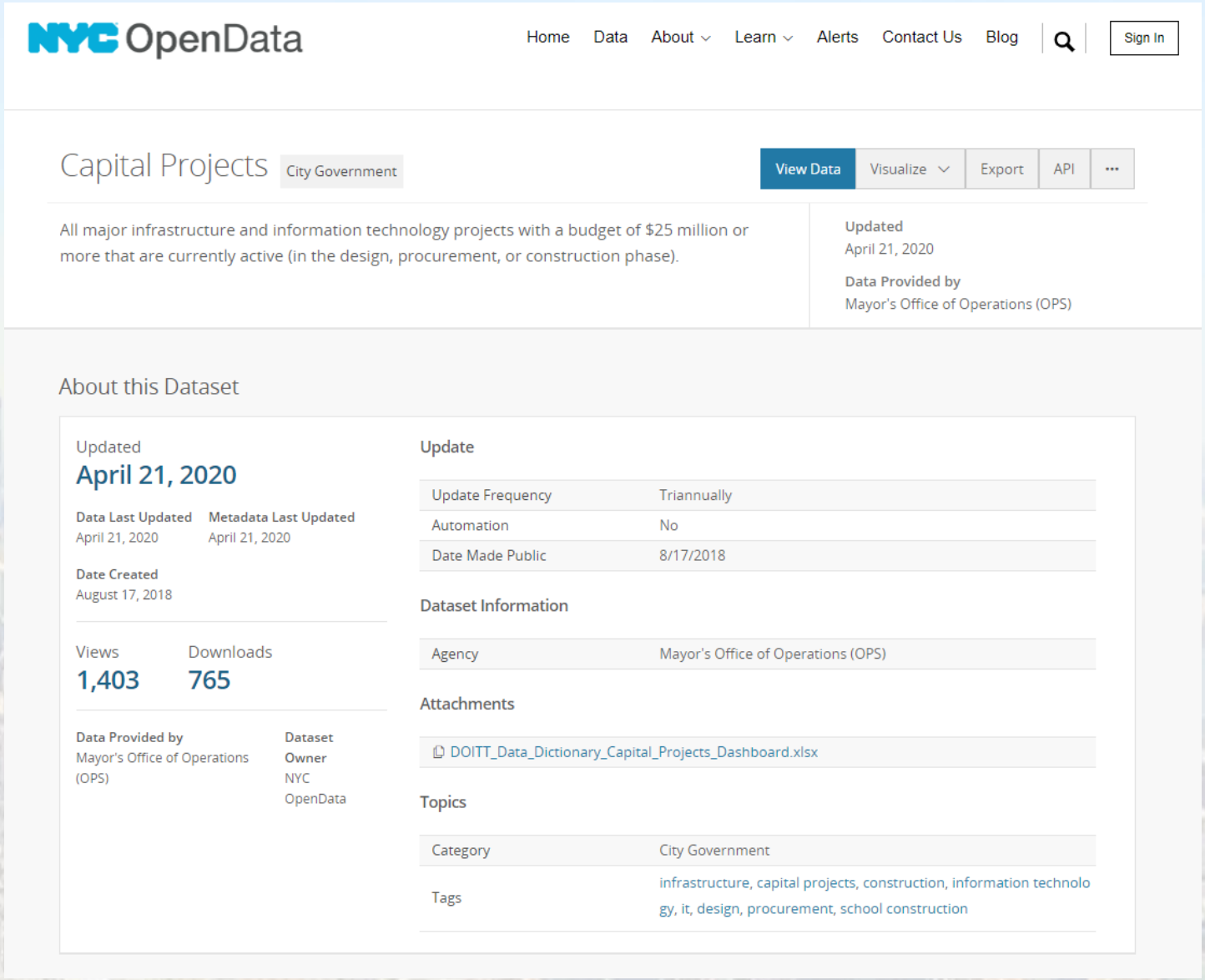
Prof. Pavlos Protopapas, Mark Glickman, Chris Tanner

Project Success



Dataset – NYC Capital Projects

- **Capital Projects: 378**
  - Roadways
  - Bridges
  - Schools
  - Parks
- **Managed by NYC agencies**
- **Budget: $25M or More**
- **Schedule: 1993 - 2020**



Method of Procedure

1

Enhancing the dataset

- *How can we understand what makes projects unique?*
- **Data** augmentation with **text analysis** and adding descriptors

2

Predicting Project Success

- *Will we on-time and on-budget?*
- **Recurrent neural networks** using project descriptions

3

Modeling Project Error

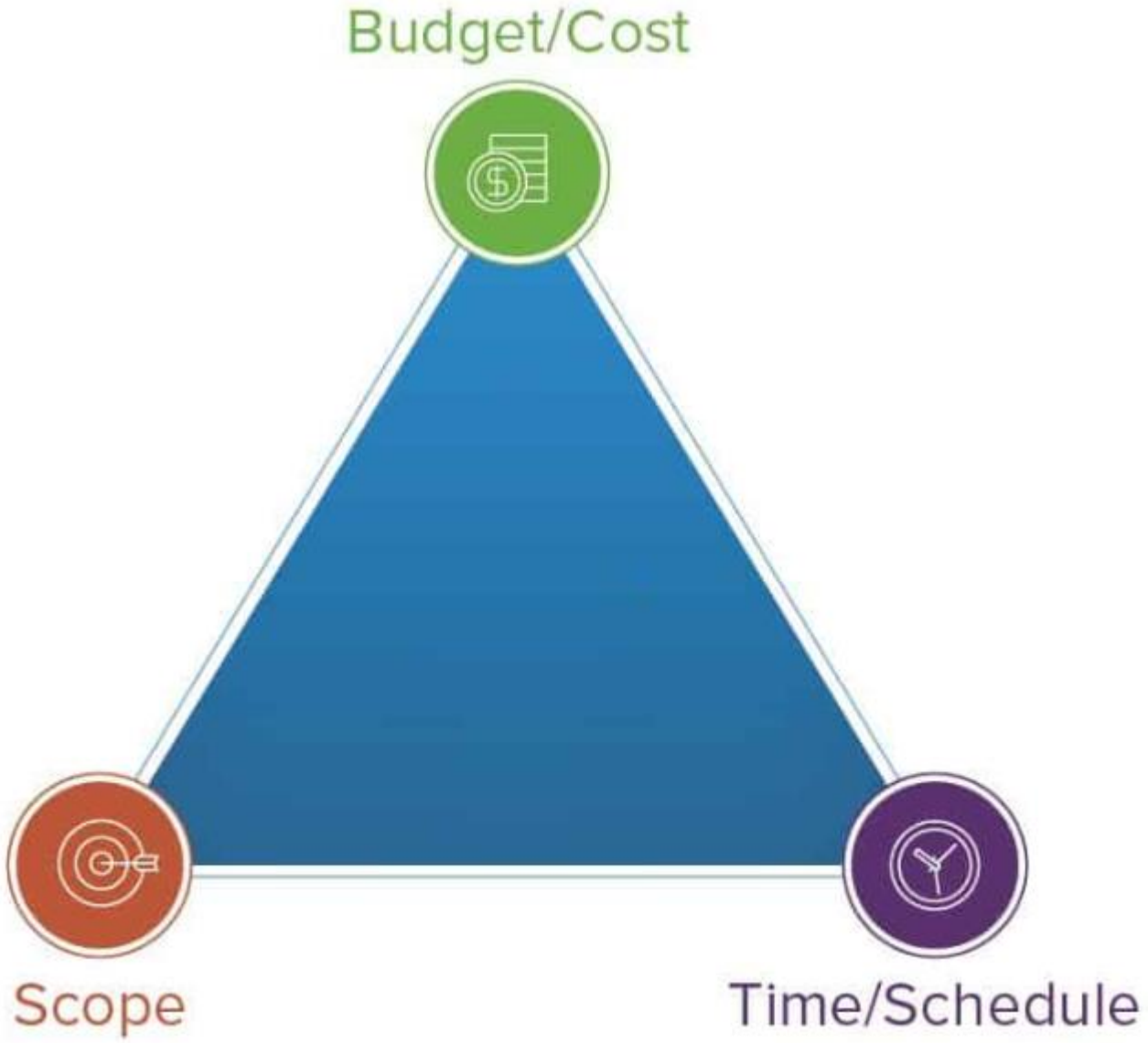
- *How correct are our schedule and budget forecasts?*
- **Linear Regression** Statistical Modeling on Percentage Error

4

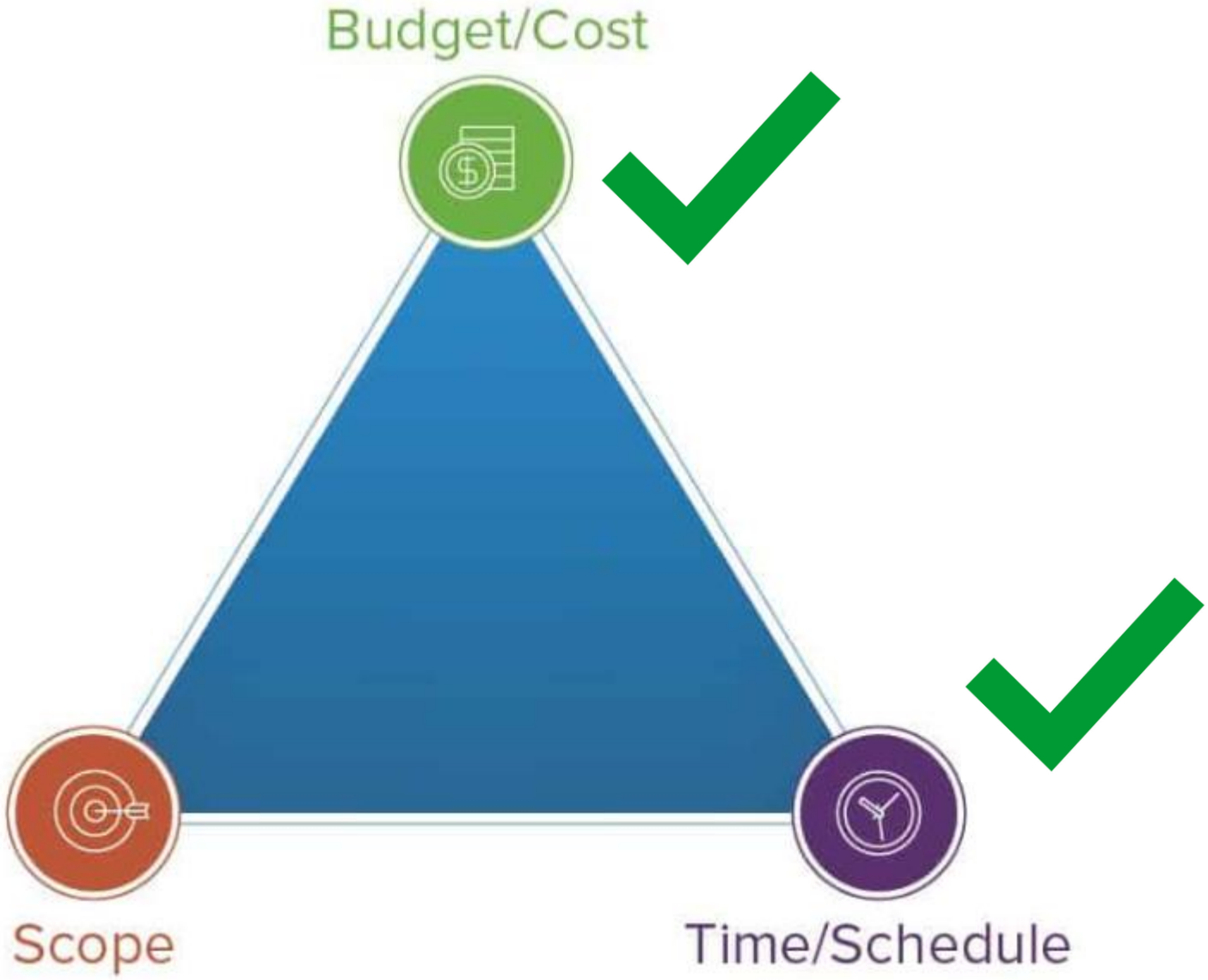
Visualizing Project Success

- *How can we make better project decisions?*
- **Iterative** statistical modeling using **Bayesian** estimation

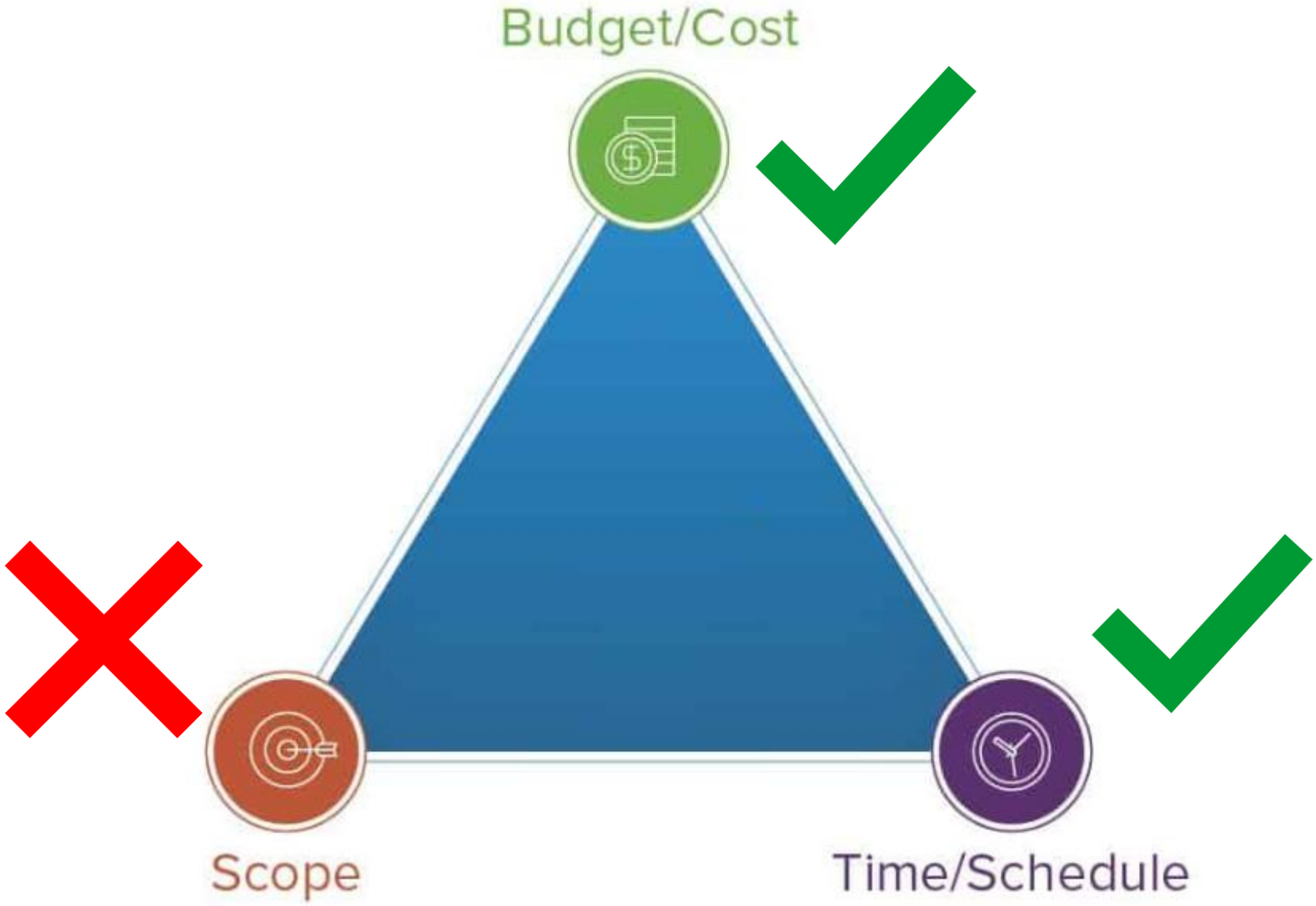
Project Success



Project Success



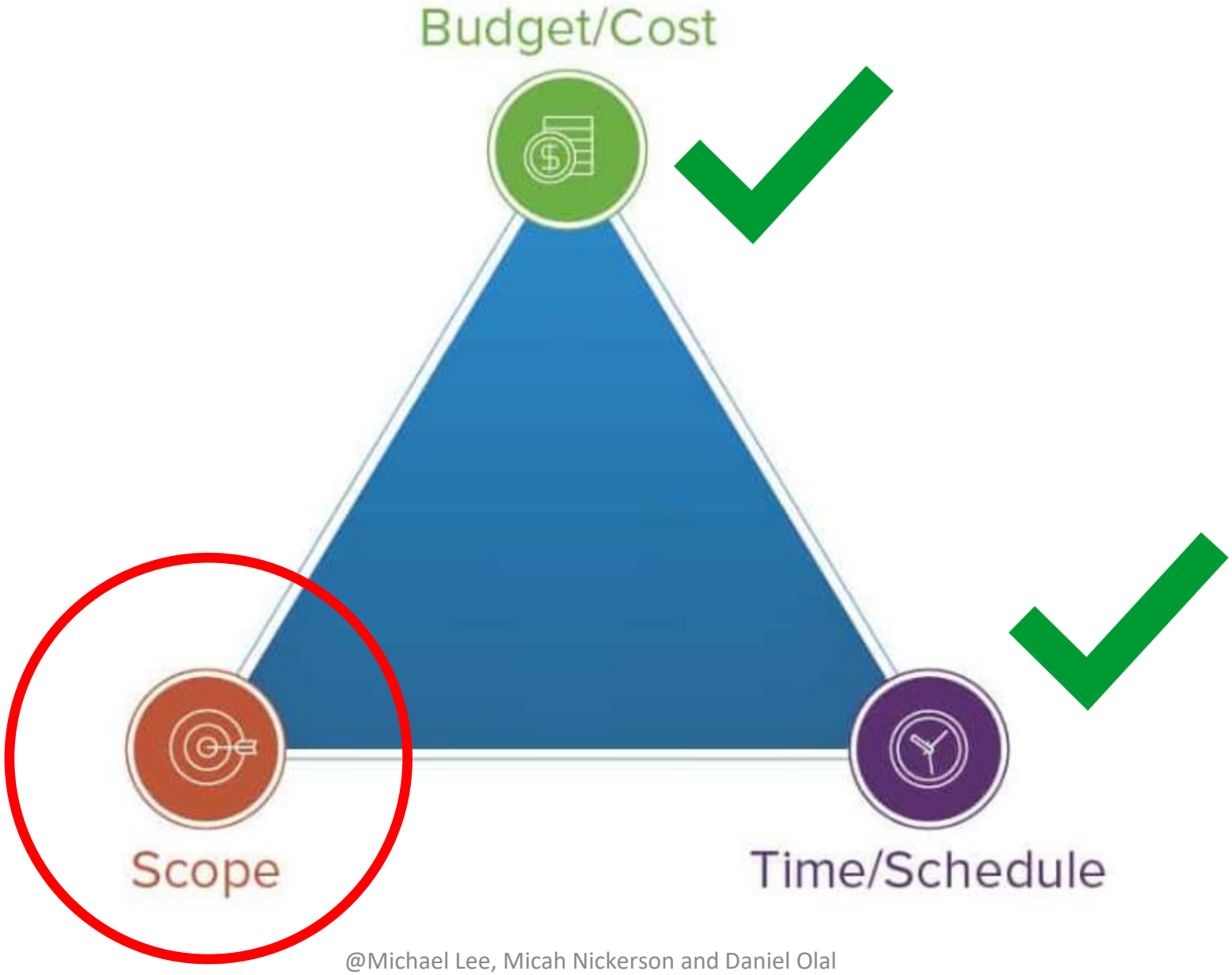
Project Success



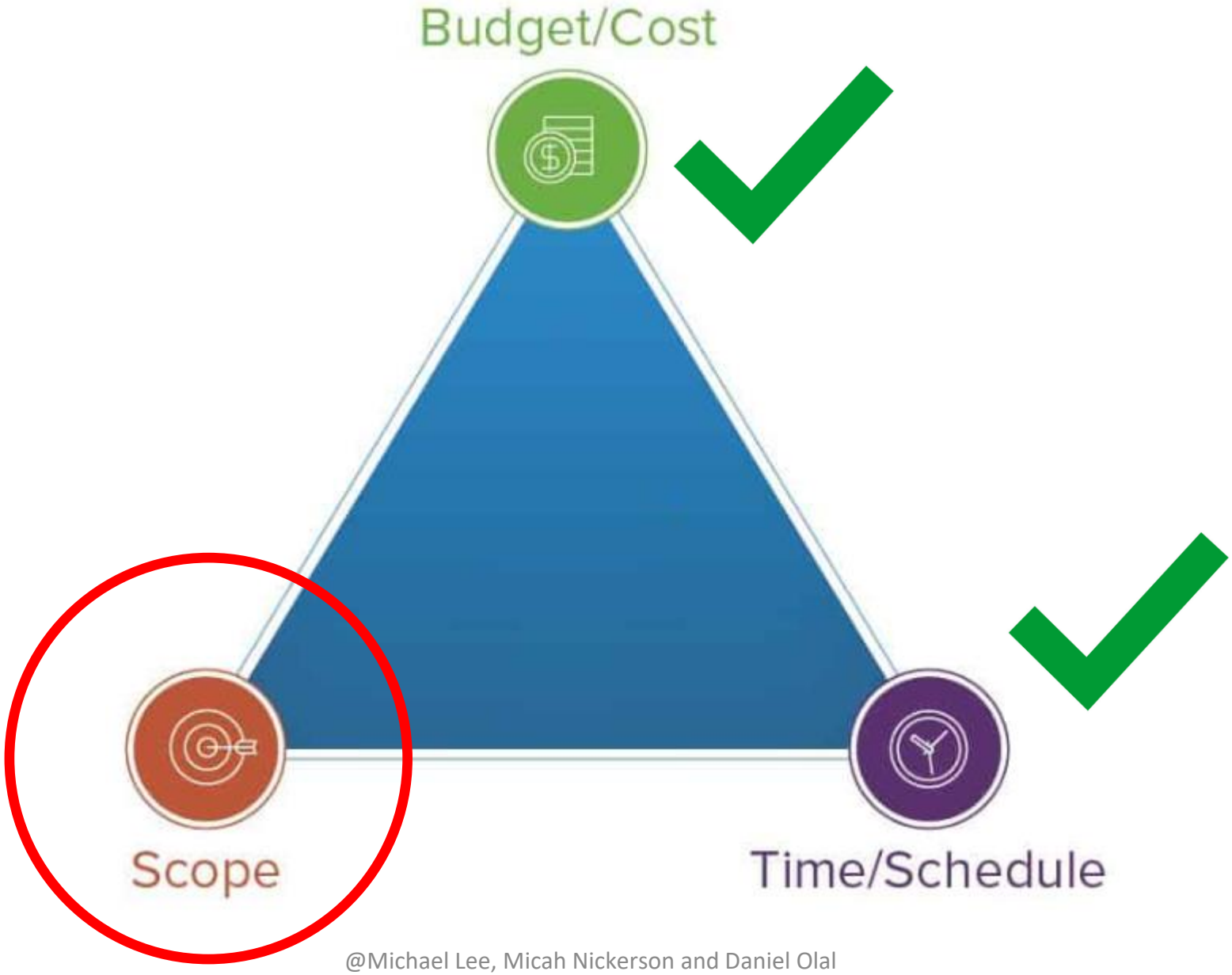
Project Success

<u>Scope:</u>

Deliverable ✓



Project Success



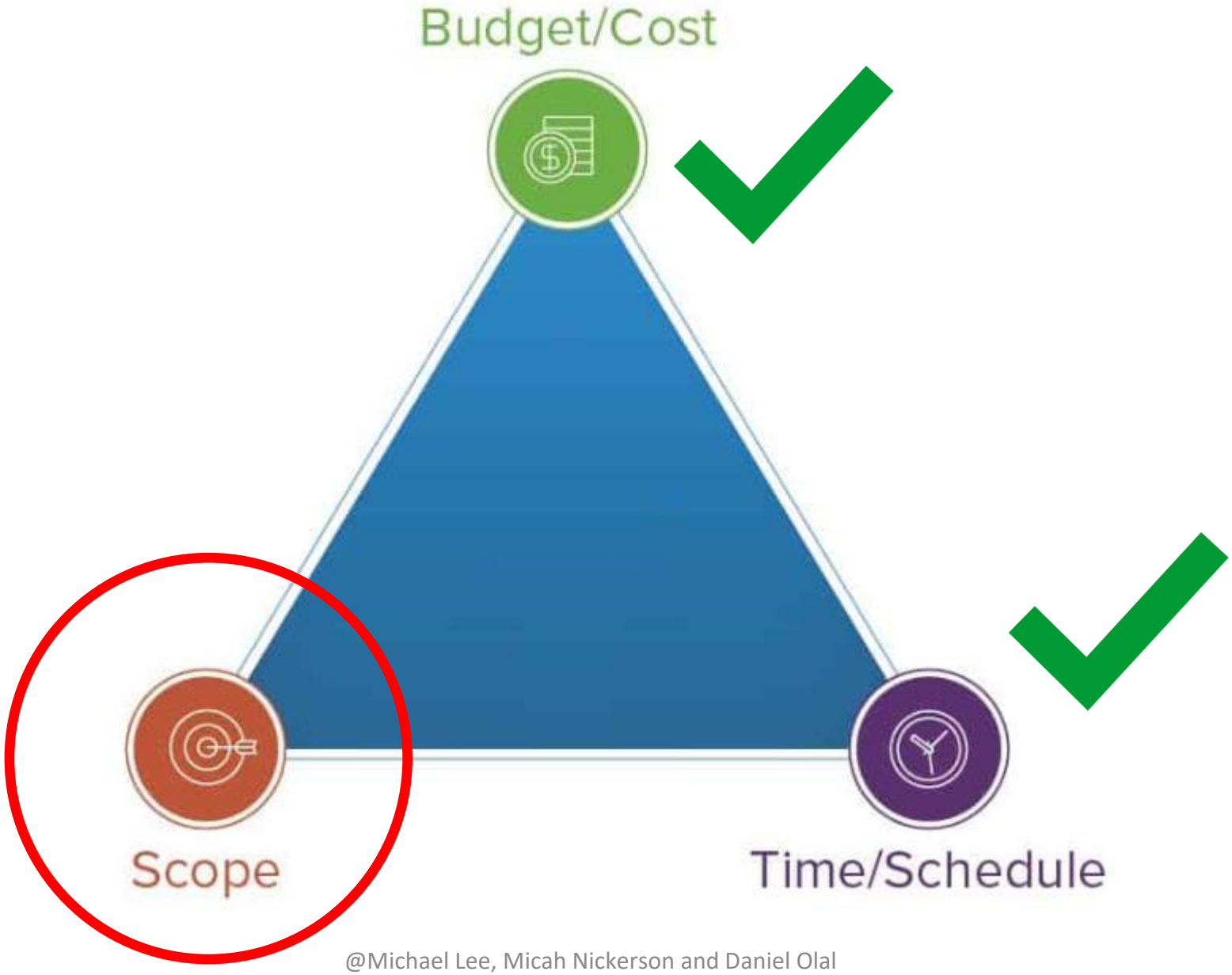
Scope:

Deliverable ✓

What Tasks? ✗

No. of Steps? ✗

Project Success



Scope:

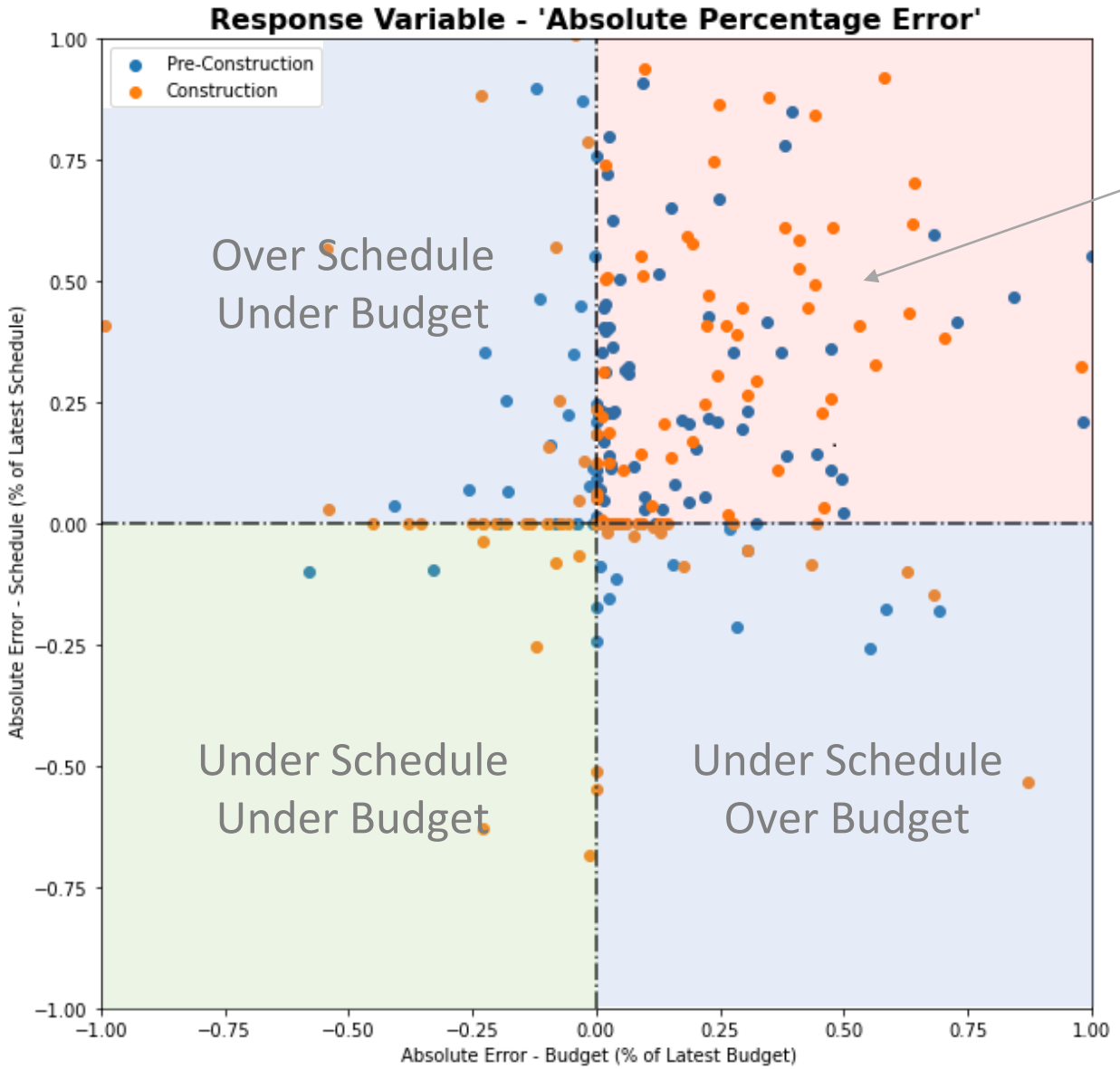
Deliverable ✓

What Tasks? ✗

No. of Steps? ✗

Tasks must be Estimated

Capital Projects - Exploratory Data Analysis

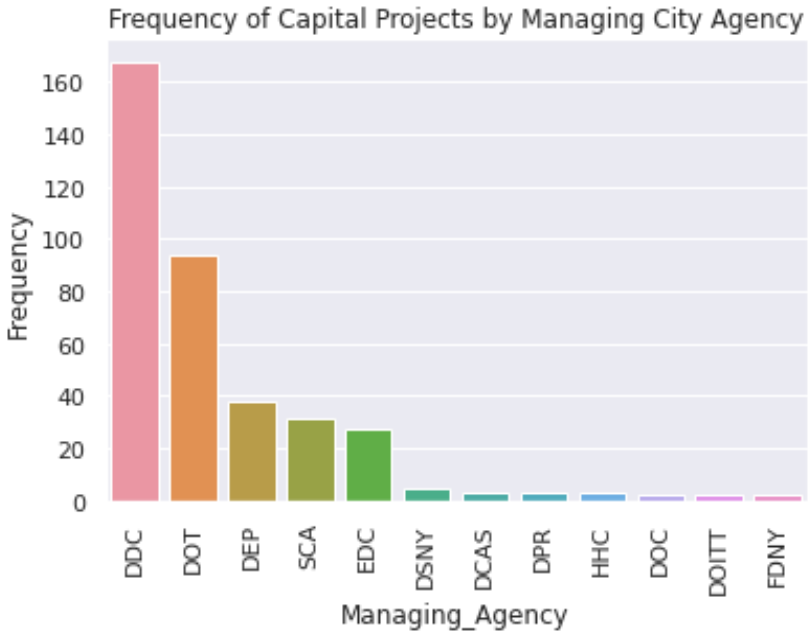


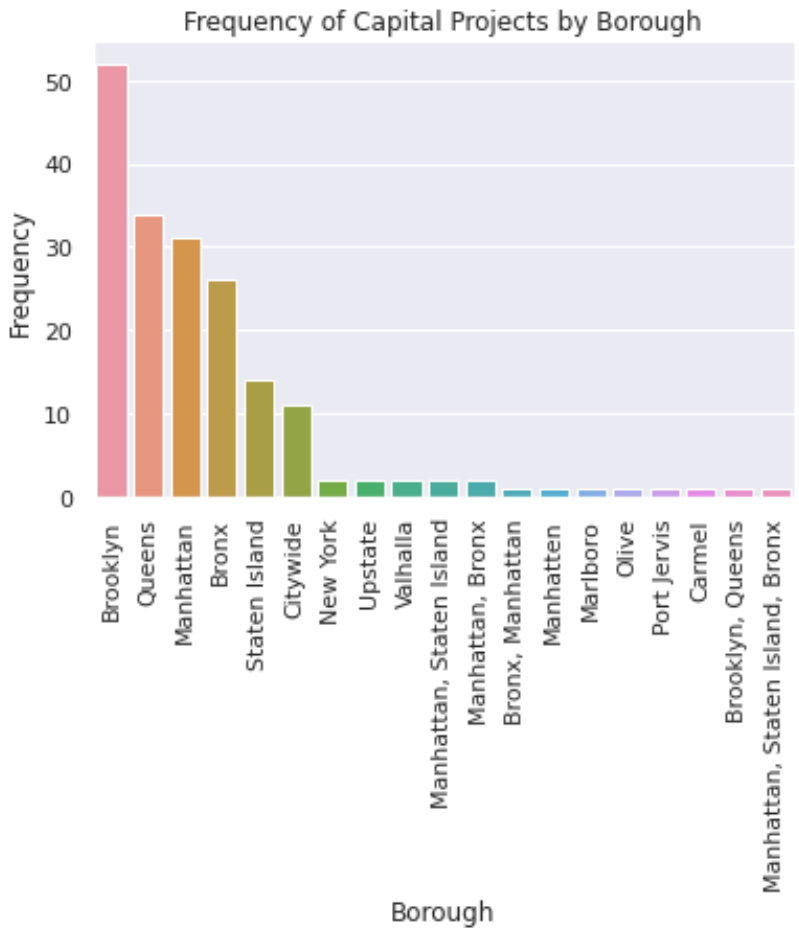
Over Schedule
Over Budget

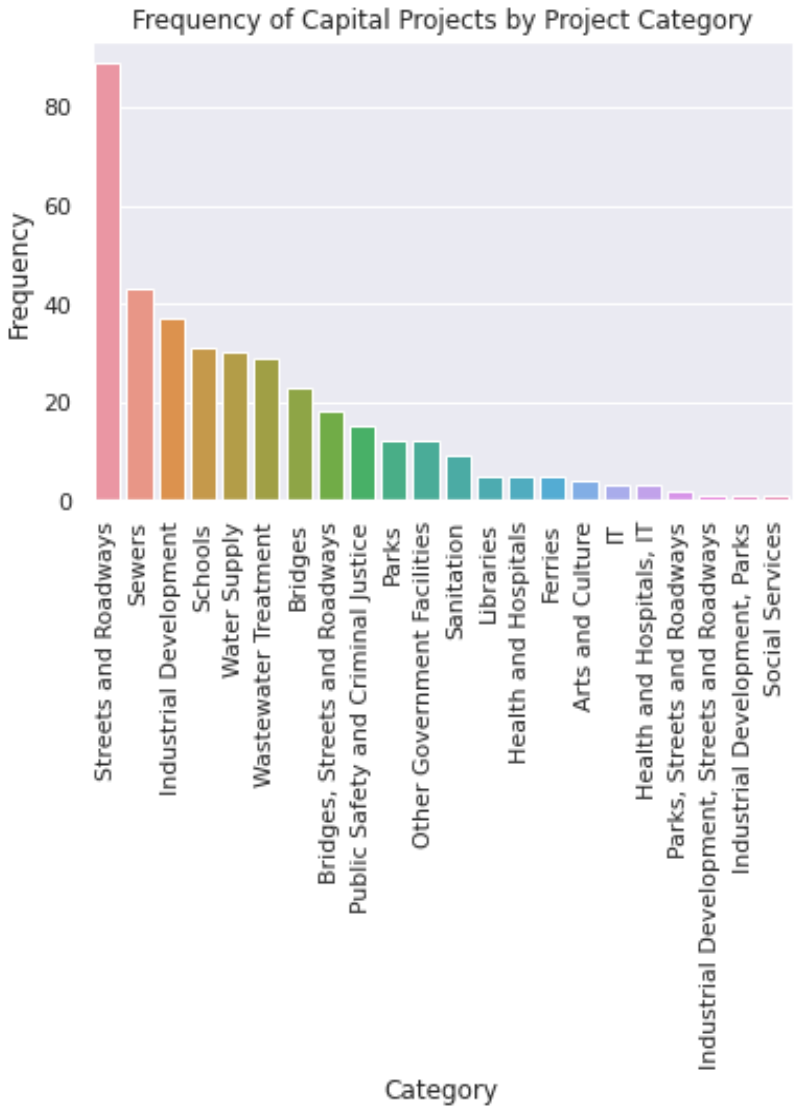
Table 1 - Data Distribution (by Project Success Measurement)	
Subset of Data	**Total (Percentage of Dataset)**
Total Projects	**378 (100.00%)**
Projects *Over Budget*	264 *(69.84%)*
Project *Over Schedule*	267 *(70.63%)*
Projects ***Both*** *Over Budget and Over Schedule*	233 *(61.64%)*

Imbalanced Classes

Capital Projects - Exploratory Data Analysis



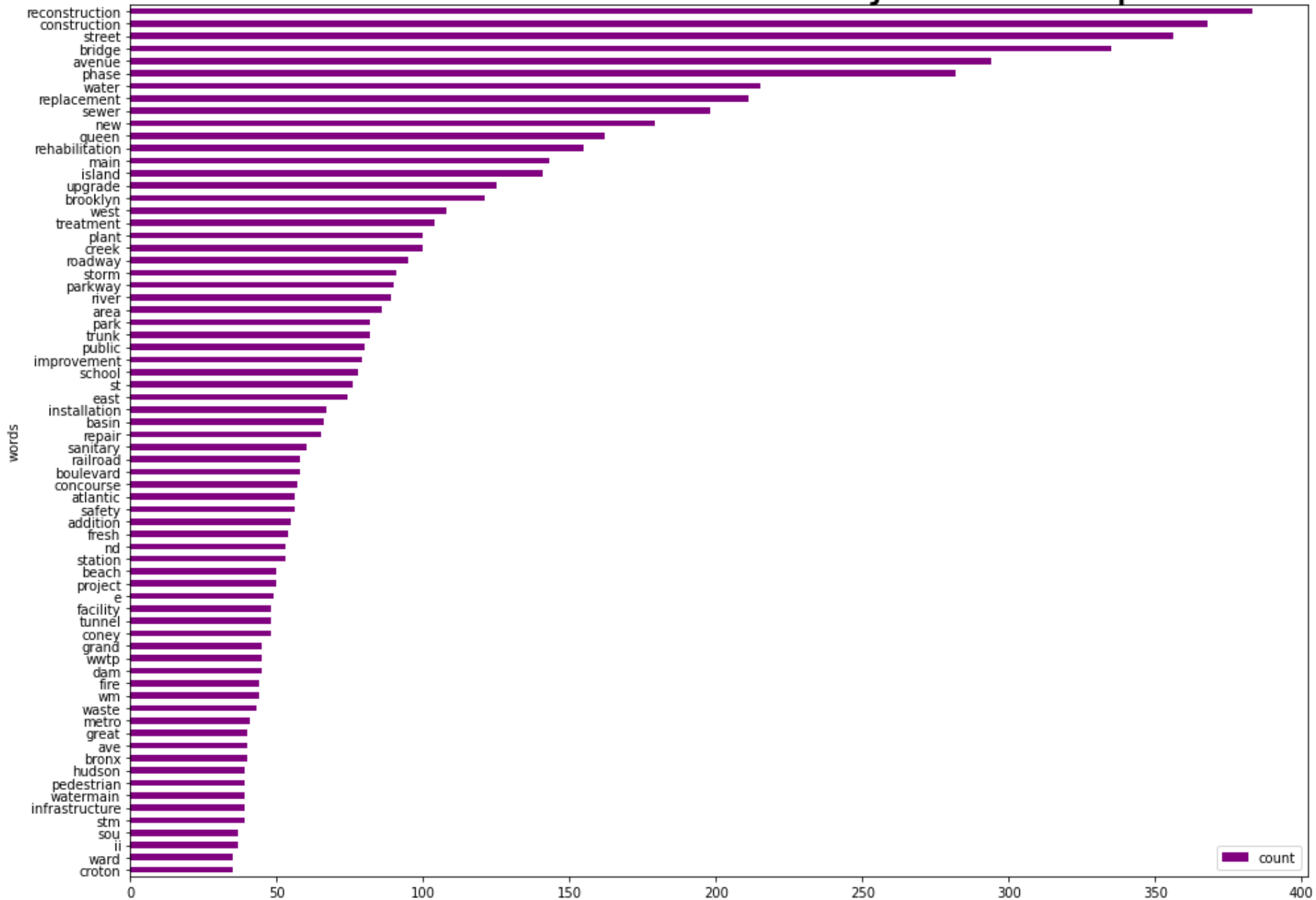




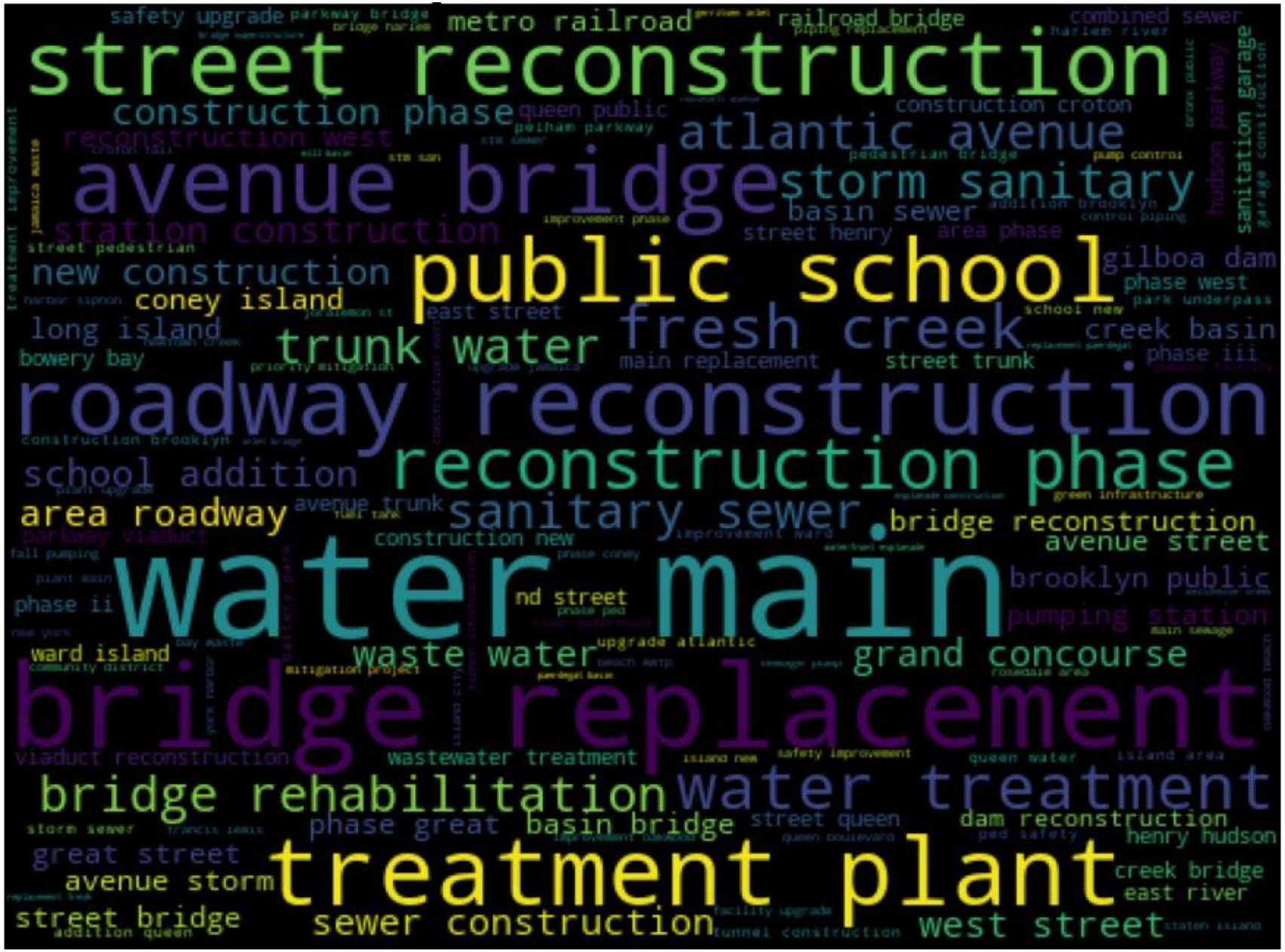
How can we understand what makes projects unique?

1 **Text Analysis + Data Augmentation**

Common Words Found in Project Description



Project Word Cloud



Representing Scope: Added Descriptors

Scope

- New
- Renovation
- Rehabilitation
- Restoration
- Repair
- Reconstruction
- Upgrade
- Building

CSI Divisions

- Electrical
- Landscaping
- Painting
- Electronic
- Communications
- Security
- Safety
- Water
- Fire
- Power
- Waste
- Transport

Logistical Complexity

- **No** Site Storage Available
- **Limited** Site Storage Available
- **Large** Site Storage Available

Project Type

- Building
- Civil
- Digital
- Park
- Street
- Building

Will we be on time and on budget?

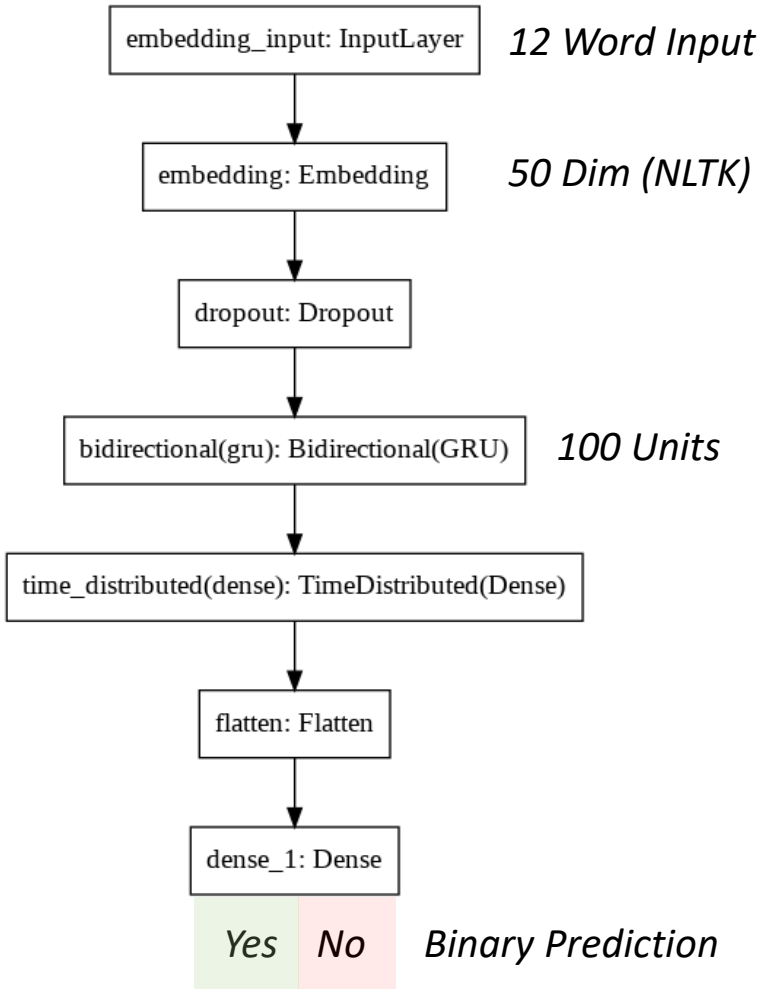
2

Recurrent Neural Networks

to **Predict Project Success**

BiDirectional Gated Recurrent Unit Neural Network

Model Structure
(Identical for Schedule and Budget)



Example Project Description:

"... Repair/replacement of bridge superstructure and substructure and new concrete-filled steel deck ... "

Data Preparation

- **Class Balancing**
  - Equal number of successful and unsuccessful observations.
  - Both Budget and Schedule

- **Project Description** – Sentence Processing
  - Removed
    - *Alphanumerics*
    - *Stop Words*
  - **Tokenized** Project Descriptions

- **Word Embeddings**
  - 50 Dimensional Word Embedding

Will we be on budget?

BiGRU Model – **Budget** Success

Confusion Matrix

Actual	Predicted: On Budget	Predicted: Over Budget
On Budget	123	1
Over Budget	24	260

Model Performance - Accuracy

	precision	recall	f1-score	support
0.0	0.84	0.99	0.91	124
1.0	1.00	0.92	0.95	284
accuracy			0.94	408
macro avg	0.92	0.95	0.93	408
weighted avg	0.95	0.94	0.94	408

Model Predictions

```
This model has: 383 correct predictions
This model has: 25 incorrect predictions
```

Will we be on time?

BiGRU Model – **Schedule** Success

Confusion Matrix

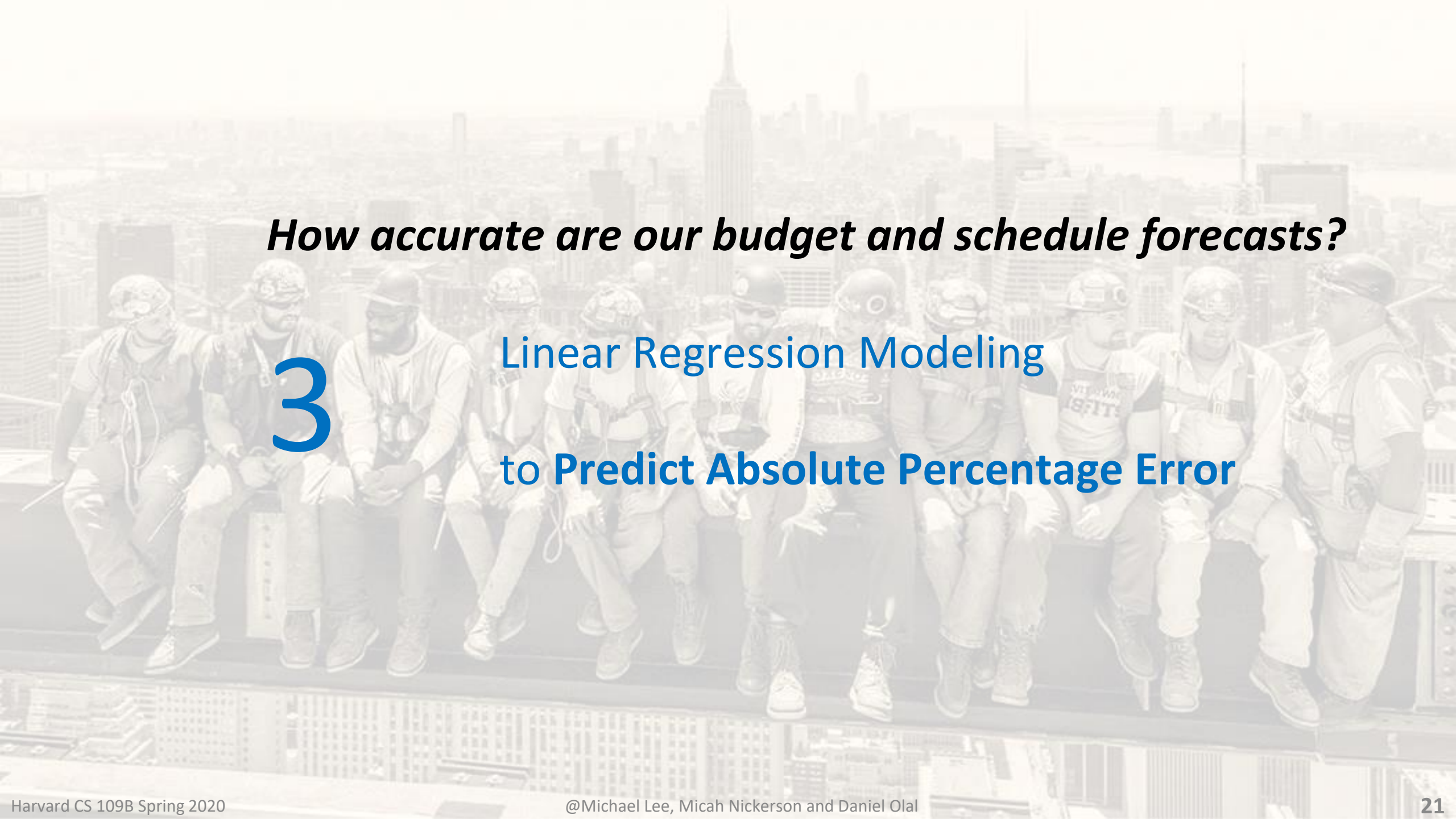
Actual	Predicted: On Schedule	Predicted: Over Schedule
On Schedule	79	3
Over Schedule	19	307

Model Performance - Accuracy

	precision	recall	f1-score	support
0.0	0.81	0.96	0.88	82
1.0	0.99	0.94	0.97	326
accuracy			0.95	408
macro avg	0.90	0.95	0.92	408
weighted avg	0.95	0.95	0.95	408

Model Predictions

```
This model has: 386 correct predictions
This model has: 22 incorrect predictions
```



How accurate are our budget and schedule forecasts?

3

Linear Regression Modeling

to **Predict Absolute Percentage Error**

How accurate are our budget and schedule forecasts?

Absolute Percentage Error (APE)

$$\text{Absolute Percentage Error (\%)} = \frac{|\text{Estimated Value} - \text{Actual Value}|}{\text{Actual Value}} \times 100$$

Schedule

Budget

How accurate are our budget forecasts?

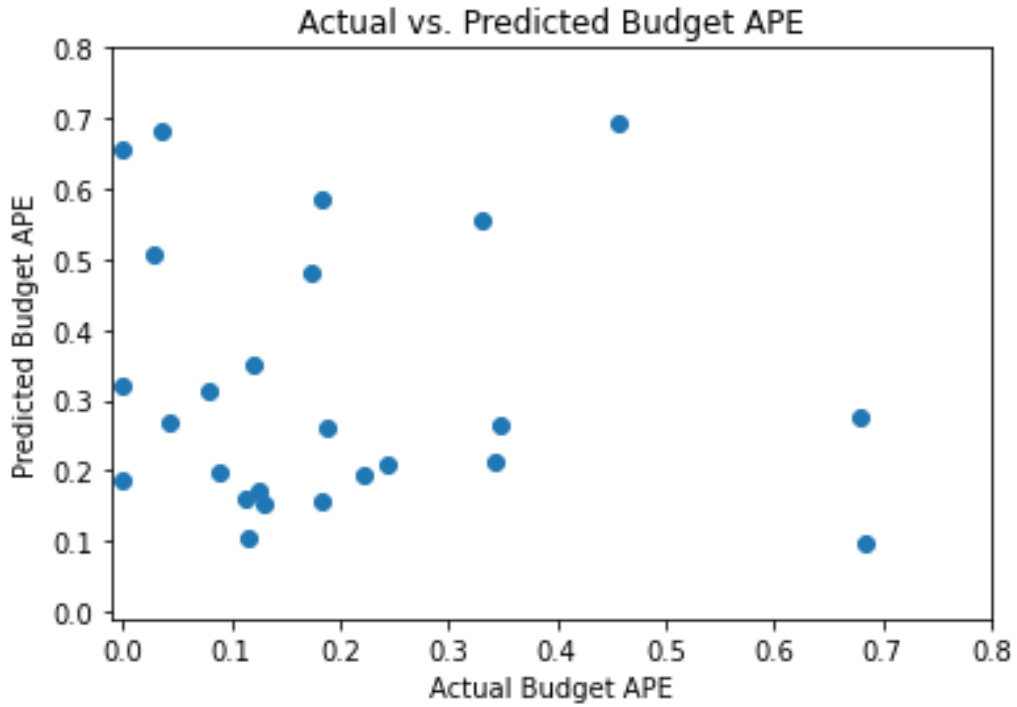
Budget APE Linear Model

OLS Regression Results

Dep. Variable:	Budget_Absoulte_Error	**R-squared:**	0.335
Model:	OLS	**Adj. R-squared:**	0.265
Method:	Least Squares	**F-statistic:**	4.764
Date:	Sat, 09 May 2020	**Prob (F-statistic):**	3.84e-05
Time:	03:44:21	**Log-Likelihood:**	7.2443
No. Observations:	95	**AIC:**	5.511
Df Residuals:	85	**BIC:**	31.05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2659	0.024	10.936	0.000	0.218	0.314
Latest_Schedule_Changes	-0.0638	0.029	-2.201	0.030	-0.121	-0.006
Total_Schedule_Changes	0.0839	0.027	3.120	0.002	0.030	0.137
Category_Public Safety and Criminal Justice	-0.0604	0.029	-2.099	0.039	-0.118	-0.003
Category_Streets and Roadways	0.1039	0.045	2.330	0.022	0.015	0.193
Managing_Agency_DDC	0.0649	0.027	2.413	0.018	0.011	0.118
Client_Agency_DCAS	0.0623	0.026	2.397	0.019	0.011	0.114
Client_Agency_DOT	-0.0865	0.046	-1.877	0.064	-0.178	0.005
Client_Agency_Mayor's Office	0.0627	0.025	2.556	0.012	0.014	0.111
Site Storage_1	-0.0576	0.025	-2.321	0.023	-0.107	-0.008

Omnibus:	40.519	**Durbin-Watson:**	1.890
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	109.407
Skew:	1.502	**Prob(JB):**	1.75e-24
Kurtosis:	7.315	**Cond. No.**	3.77




```
Budget Model Train RMSE: 0.05
Budget Model Test RMSE: 0.12
```

How accurate are our schedule forecasts?

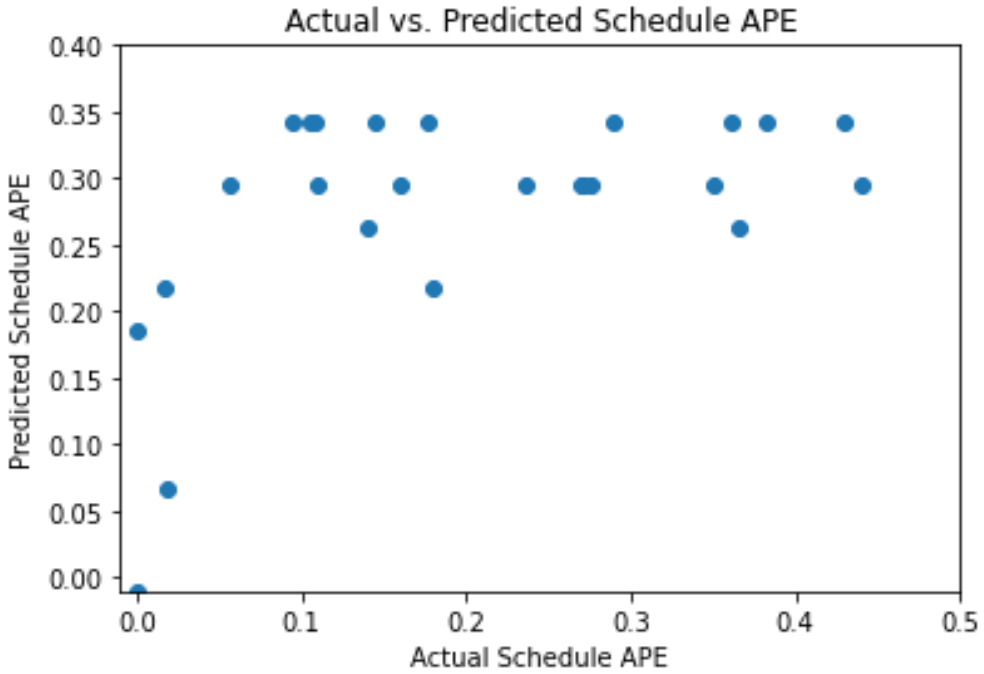
Schedule APE Linear Model

OLS Regression Results

Dep. Variable:	Schedule_Absoulte_Error	**R-squared:**	0.306
Model:	OLS	**Adj. R-squared:**	0.258
Method:	Least Squares	**F-statistic:**	6.456
Date:	Sat, 09 May 2020	**Prob (F-statistic):**	1.15e-05
Time:	03:58:20	**Log-Likelihood:**	48.084
No. Observations:	95	**AIC:**	-82.17
Df Residuals:	88	**BIC:**	-64.29
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2474	0.016	15.913	0.000	0.217	0.278
Current_Phase_Design	0.0374	0.016	2.355	0.021	0.006	0.069
Category_Sanitation	-0.0314	0.016	-1.931	0.057	-0.064	0.001
Category_Schools	-0.0281	0.006	-4.857	0.000	-0.040	-0.017
Managing_Agency_SCA	-0.0281	0.006	-4.857	0.000	-0.040	-0.017
Client_Agency_DEP	-0.0228	0.018	-1.284	0.203	-0.058	0.013
Client_Agency_DOE	-0.0281	0.006	-4.857	0.000	-0.040	-0.017
Client_Agency_ORR	-0.0348	0.016	-2.208	0.030	-0.066	-0.003
Divison_8	-0.0167	0.016	-1.055	0.294	-0.048	0.015

Omnibus:	1.524	**Durbin-Watson:**	2.181
Prob(Omnibus):	0.467	**Jarque-Bera (JB):**	1.365
Skew:	0.292	**Prob(JB):**	0.505
Kurtosis:	2.938	**Cond. No.**	6.44e+18




```
Schedule Model Train RMSE: 0.02
Schedule Model Test RMSE: 0.03
```

How can we make better project decisions?

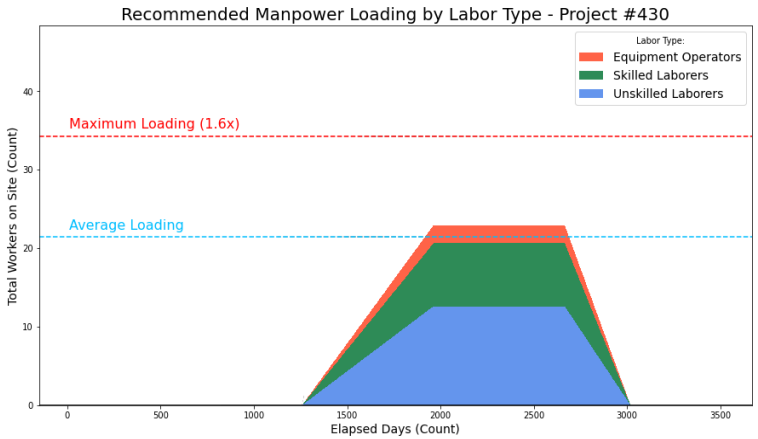
4

Bayesian Iterative Statistical Modeling

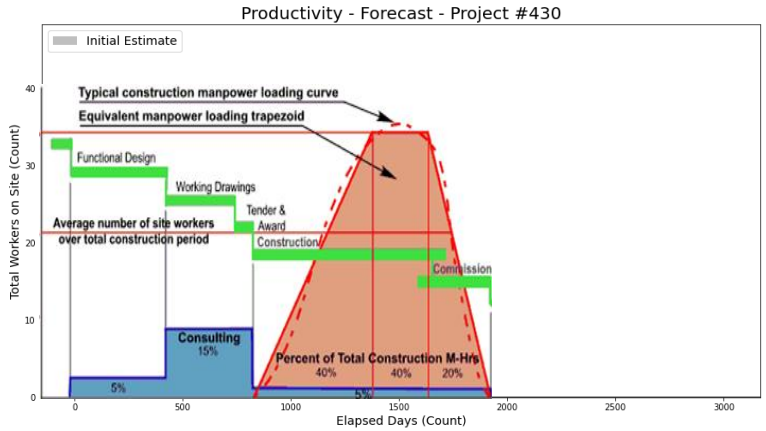
to **Visualize Project Success**

Bayesian Statistical Approach

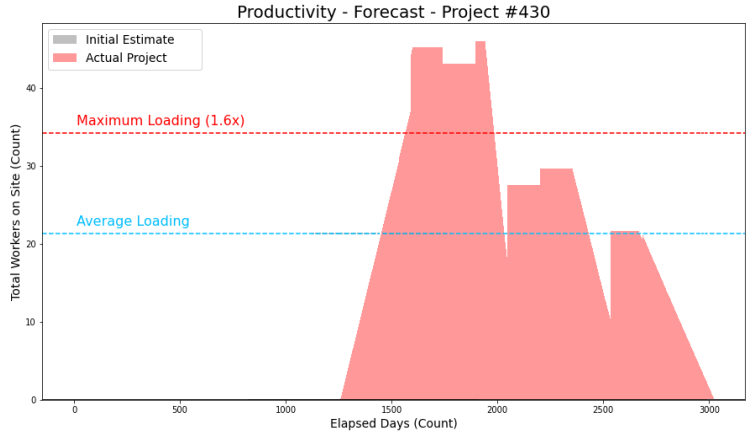
Posterior $\propto$ **Prior** **x** **Likelihood**



$\propto$

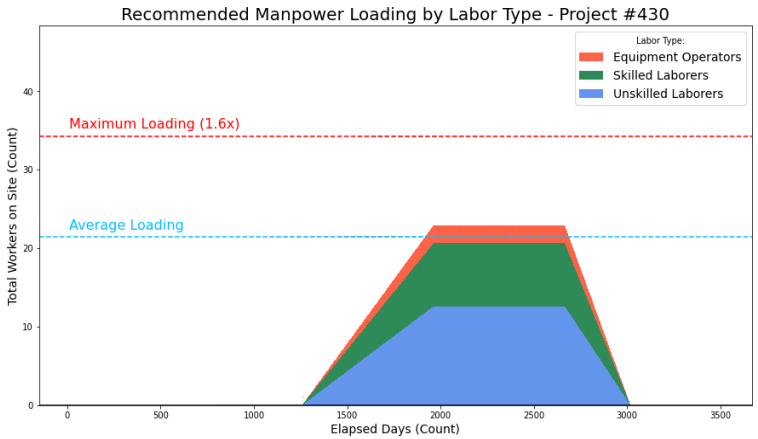


x

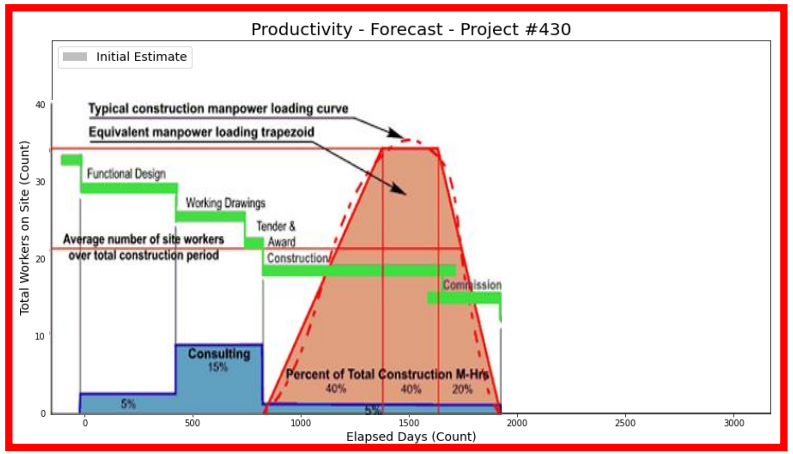


Bayesian Statistical Approach

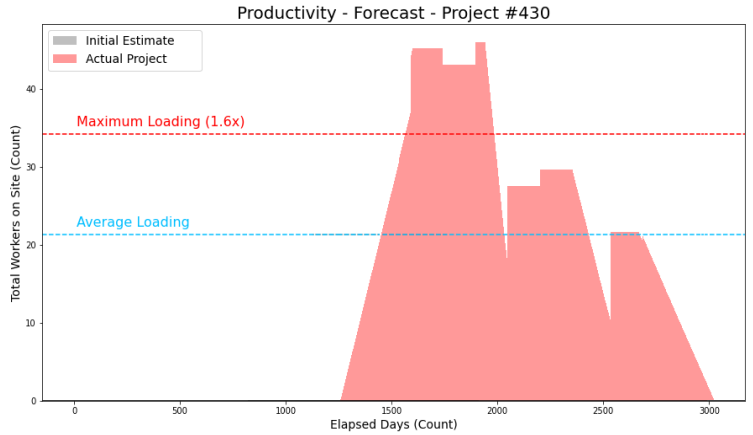
Posterior $\propto$ **Prior** x Likelihood



$\propto$

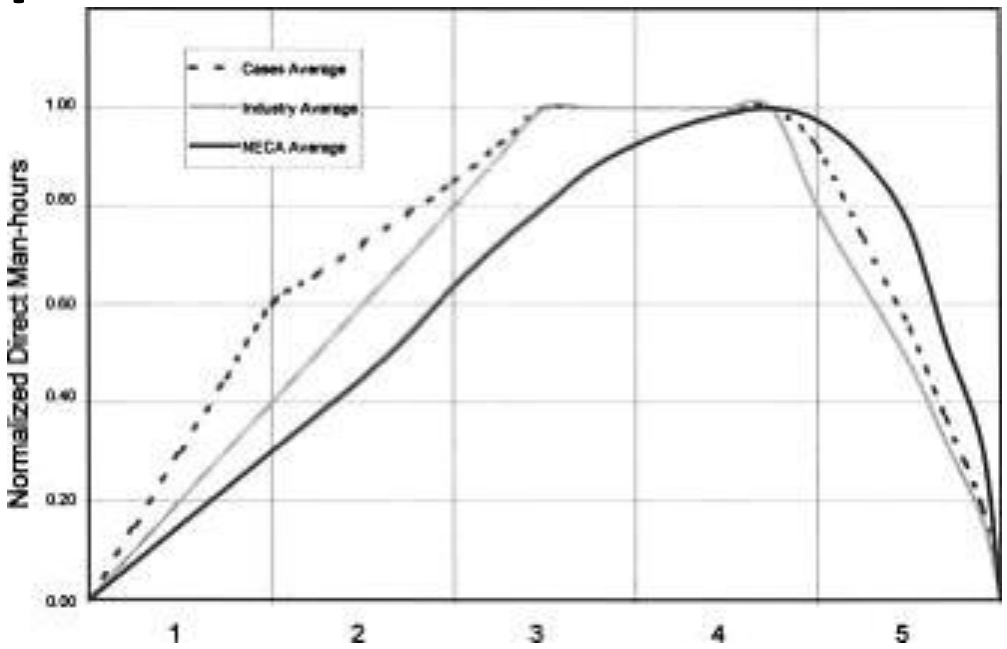


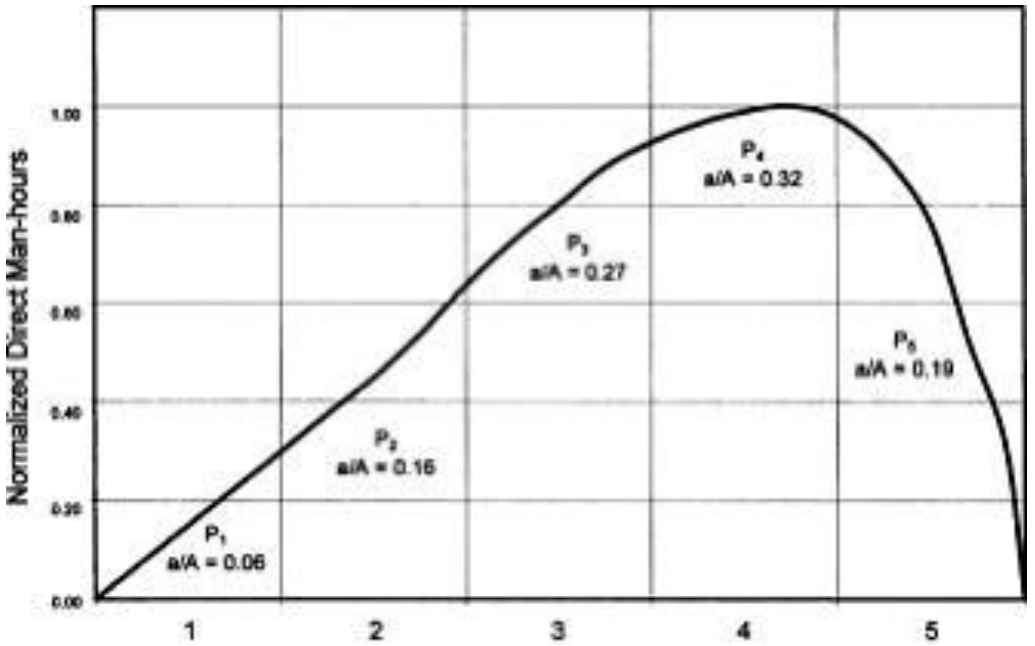
x

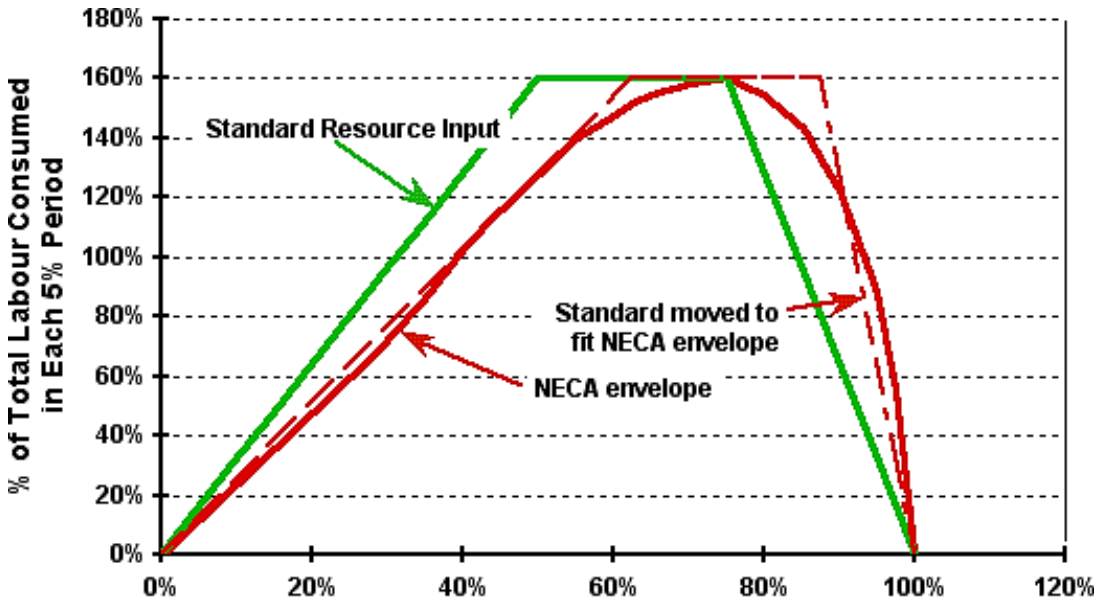


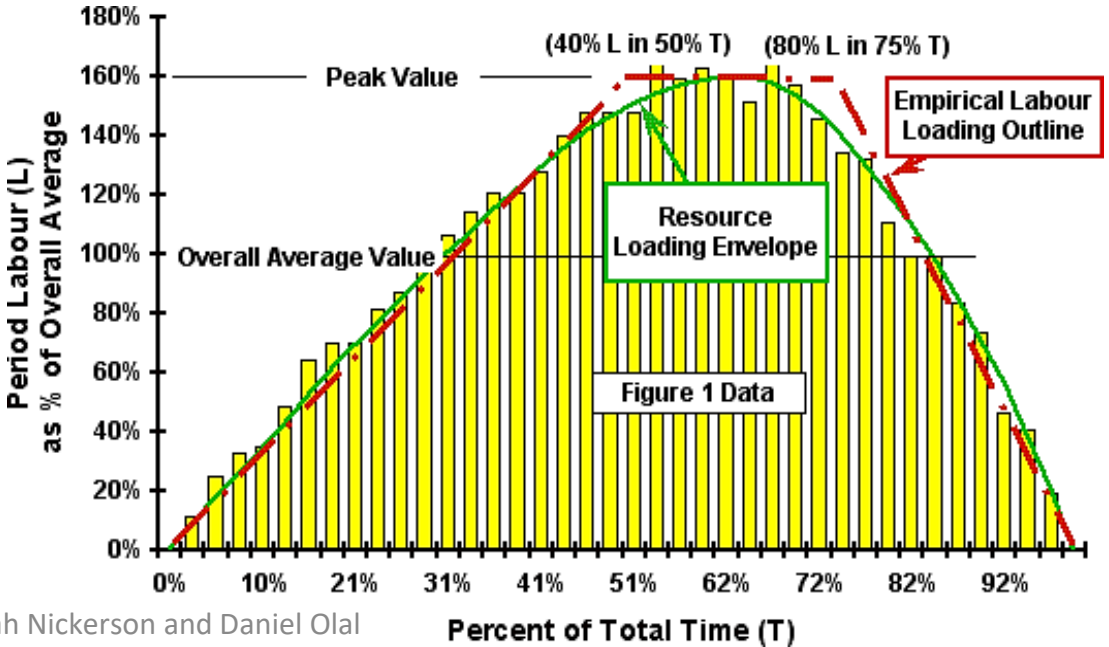
Prior

SCOPE - “Labor Loading Curves” -- How much activity per day?



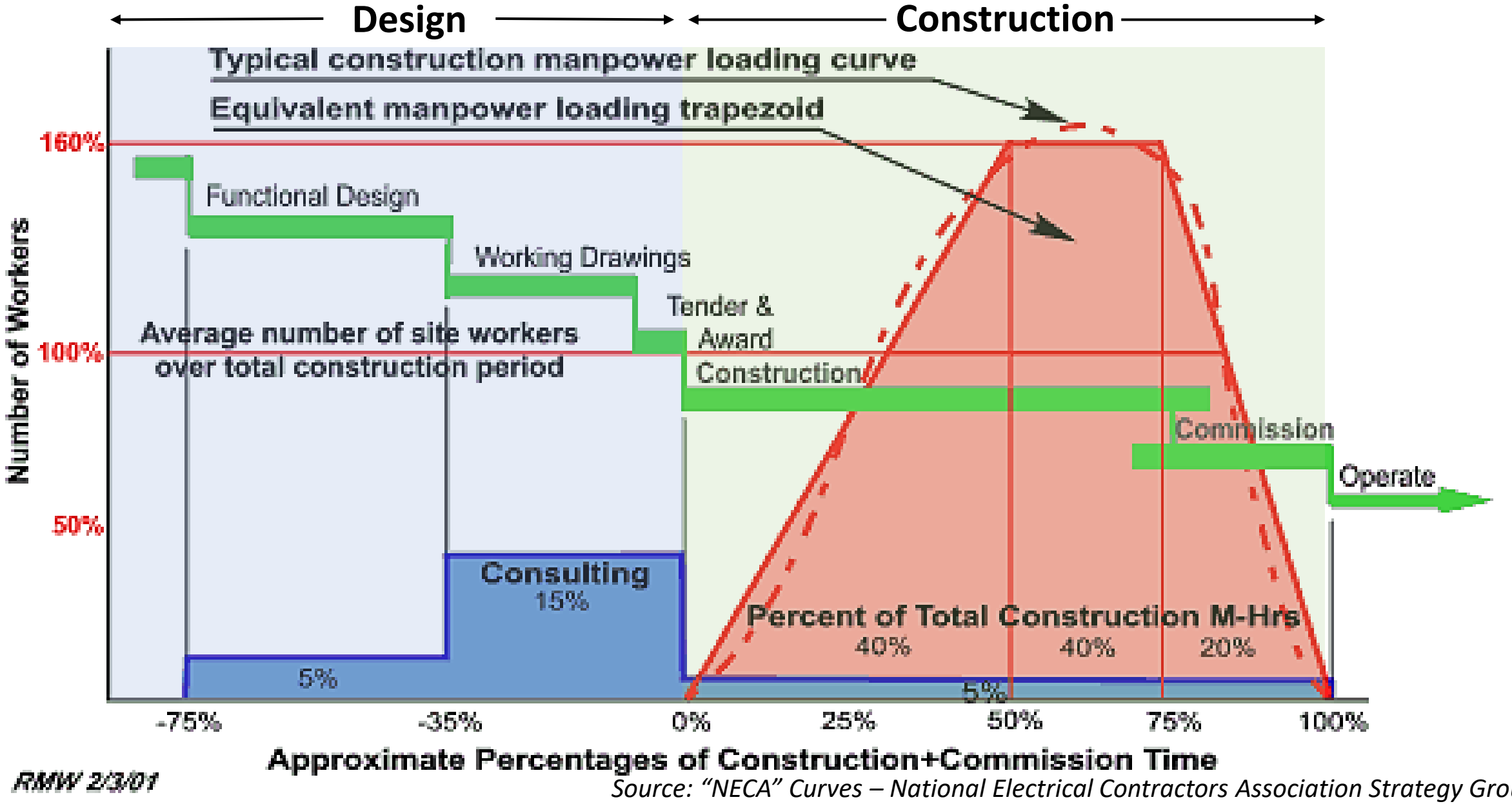






Prior

"A Typical NYC Project"



Source: "NECA" Curves – National Electrical Contractors Association Strategy Group

Prior

Project Types

Civil

Streetwork (Roadways and Sewers)

Buildings

Parks

Prior

Budgets per Project Type

Civil

total cost = 15% soft costs + 37.5% labor + 35% material + 12.5% heavy machinery

Streetwork (Roadways and Sewers)

total cost = 20% soft costs + 52% labor + 20% material + 8% heavy machinery

Buildings

total cost = 20% soft costs + 52% labor + 25% material+tools + 3% heavy machinery

Parks

total cost = 20% soft costs + 45% labor + 30% material+tools + 5% heavy machinery

Prior

Budgets per Project Type

Civil

total cost = 15% soft costs + **37.5% labor** + 35% material + 12.5% heavy machinery

Streetwork (Roadways and Sewers)

total cost = 20% soft costs + **52% labor** + 20% material + 8% heavy machinery

Buildings

total cost = 20% soft costs + **52% labor** + 25% material+tools + 3% heavy machinery

Parks

total cost = 20% soft costs + **45% labor** + 30% material+tools + 5% heavy machinery

Prior

Labor Breakdowns per Project Type

Civil

total labor = 32.5% Unskilled Labor + 55% Skilled Labor + 12.5% equipment operators

Streetwork (Roadways and Sewers)

total labor = 55% Unskilled Labor + 35% Skilled Labor + 10% equipment operators

Buildings

total labor = 30% Unskilled Labor + 65% Skilled Labor + 5% equipment operators

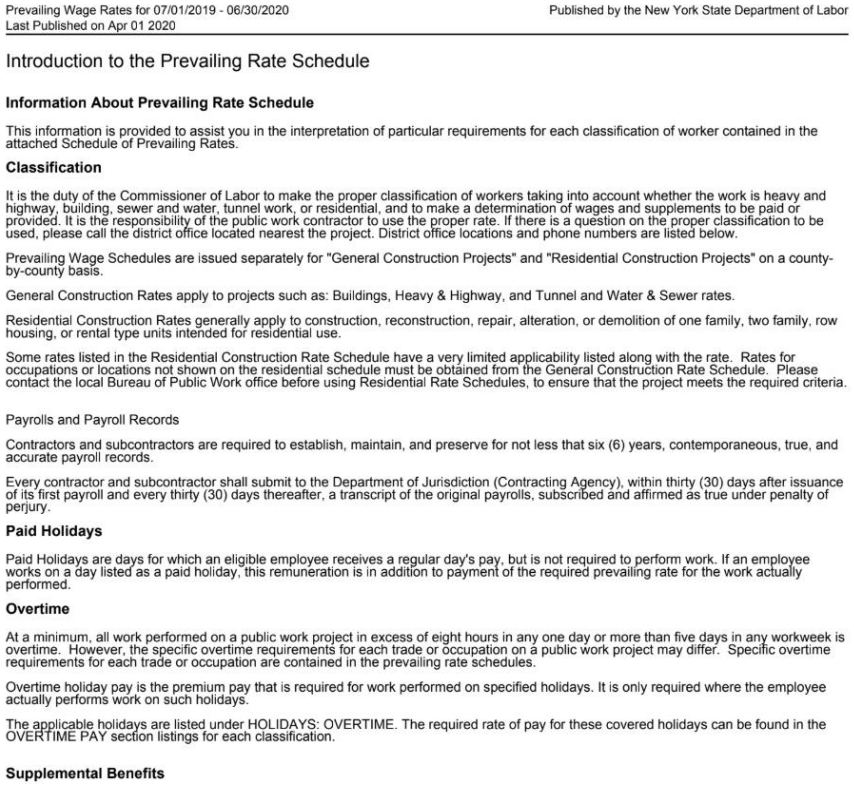
Parks

total labor = 40% Unskilled Labor + 50% Skilled Labor + 10% equipment operators

Prior

Labor Costs per Project Type

	civil			building			street			park		
	unskilled_labor	skilled_labor	equipment_operators	unskilled_labor	skilled_labor	equipment_operators	unskilled_labor	skilled_labor	equipment_operators	unskilled_labor	skilled_labor	equipment_operators
base	47.15	62.66	92.43	44.65	61.82	90.36	42.65	55.658	41.185	42.65	56	41.185
benefits	44.48	42.45	28.5	44.48	27.8	28.5	44.48	36.0825	46.7225	44.48	56.54	46.7225
total	91.63	105.11	120.93	89.13	89.62	118.86	87.13	91.7405	87.9075	87.13	112.54	87.9075



Prevailing Wage Rates for 07/01/2019 - 06/30/2020 Published by the New York State Department of Labor
Last Published on Apr 01 2020

Introduction to the Prevailing Rate Schedule

Information About Prevailing Rate Schedule

This information is provided to assist you in the interpretation of particular requirements for each classification of worker contained in the attached Schedule of Prevailing Rates.

Classification

It is the duty of the Commissioner of Labor to make the proper classification of workers taking into account whether the work is heavy and highway, building, sewer and water, tunnel work, or residential, and to make a determination of wages and supplements to be paid or provided. It is the responsibility of the public work contractor to use the proper rate. If there is a question on the proper classification to be used, please call the district office located nearest the project. District office locations and phone numbers are listed below.

Prevailing Wage Schedules are issued separately for "General Construction Projects" and "Residential Construction Projects" on a county-by-county basis.

General Construction Rates apply to projects such as: Buildings, Heavy & Highway, and Tunnel and Water & Sewer rates.

Residential Construction Rates generally apply to construction, reconstruction, repair, alteration, or demolition of one family, two family, row housing, or rental type units intended for residential use.

Some rates listed in the Residential Construction Rate Schedule have a very limited applicability listed along with the rate. Rates for occupations or locations not shown on the residential schedule must be obtained from the General Construction Rate Schedule. Please contact the local Bureau of Public Work office before using Residential Rate Schedules, to ensure that the project meets the required criteria.

Payrolls and Payroll Records

Contractors and subcontractors are required to establish, maintain, and preserve for not less that six (6) years, contemporaneous, true, and accurate payroll records.

Every contractor and subcontractor shall submit to the Department of Jurisdiction (Contracting Agency), within thirty (30) days after issuance of its first payroll and every thirty (30) days thereafter, a transcript of the original payrolls, subscribed and affirmed as true under penalty of perjury.

Paid Holidays

Paid Holidays are days for which an eligible employee receives a regular day's pay, but is not required to perform work. If an employee works on a day listed as a paid holiday, this remuneration is in addition to payment of the required prevailing rate for the work actually performed.

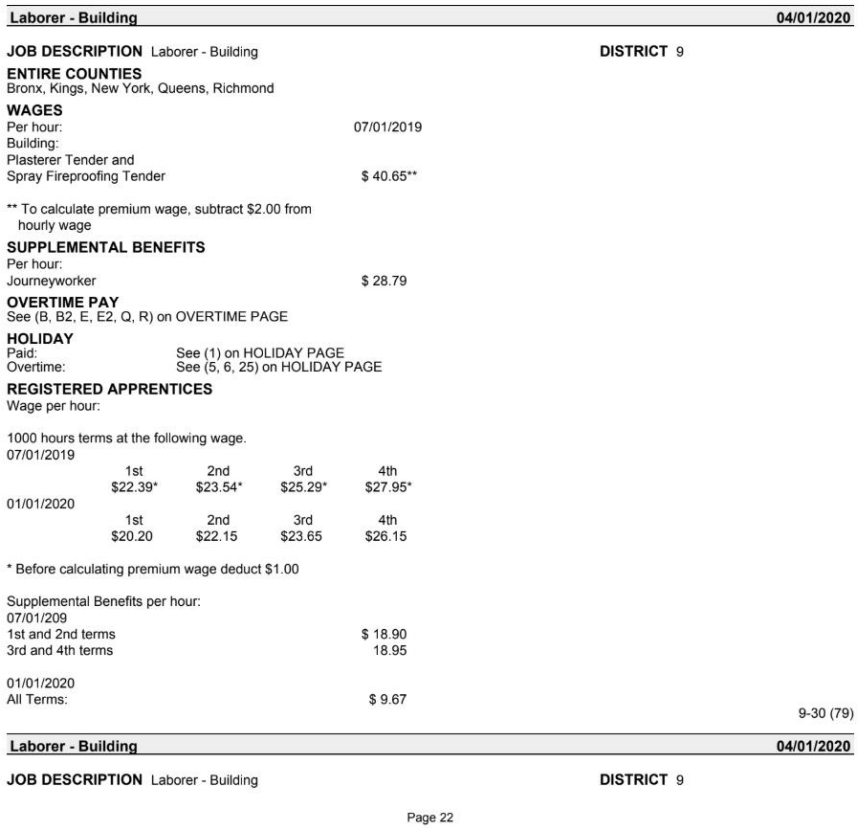
Overtime

At a minimum, all work performed on a public work project in excess of eight hours in any one day or more than five days in any workweek is overtime. However, the specific overtime requirements for each trade or occupation on a public work project may differ. Specific overtime requirements for each trade or occupation are contained in the prevailing rate schedules.

Overtime holiday pay is the premium pay that is required for work performed on specified holidays. It is only required where the employee actually performs work on such holidays.

The applicable holidays are listed under HOLIDAYS: OVERTIME. The required rate of pay for these covered holidays can be found in the OVERTIME PAY section listings for each classification.

Supplemental Benefits



Laborer - Building **04/01/2020**

JOB DESCRIPTION Laborer - Building **DISTRICT** 9

ENTIRE COUNTIES
Bronx, Kings, New York, Queens, Richmond

WAGES
Per hour: 07/01/2019
Building:
Plasterer Tender and
Spray Fireproofing Tender $ 40.65**

** To calculate premium wage, subtract $2.00 from hourly wage

SUPPLEMENTAL BENEFITS
Per hour:
Journeyworker $ 28.79

OVERTIME PAY
See (B, B2, E, E2, Q, R) on OVERTIME PAGE

HOLIDAY
Paid: See (1) on HOLIDAY PAGE
Overtime: See (5, 6, 25) on HOLIDAY PAGE

REGISTERED APPRENTICES
Wage per hour:

1000 hours terms at the following wage.

	1st	2nd	3rd	4th
07/01/2019	$22.39*	$23.54*	$25.29*	$27.95*
01/01/2020	$20.20	$22.15	$23.65	$26.15

* Before calculating premium wage deduct $1.00

Supplemental Benefits per hour:
07/01/209
1st and 2nd terms $ 18.90
3rd and 4th terms 18.95

01/01/2020
All Terms: $ 9.67

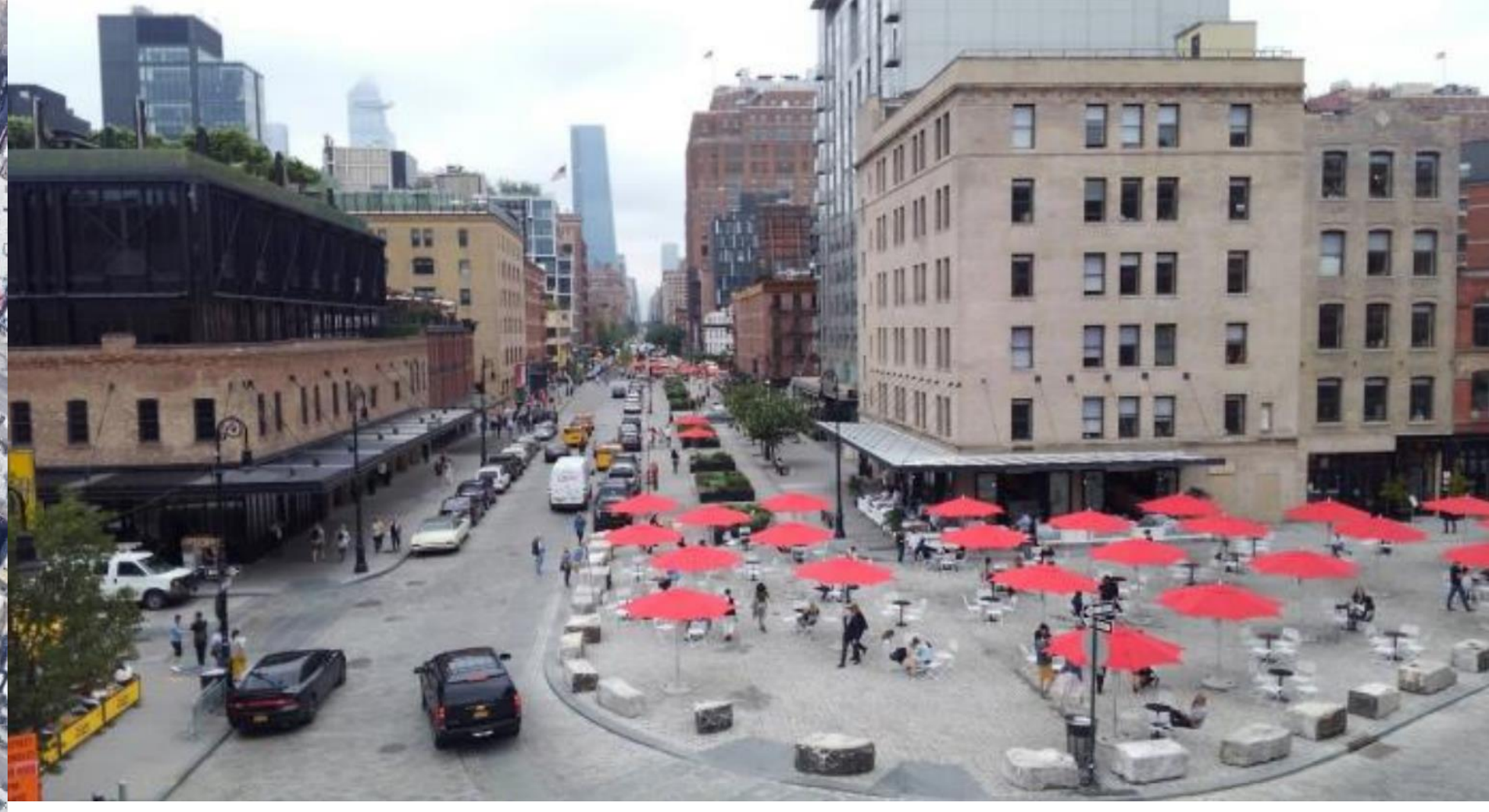
9-30 (79)

Laborer - Building **04/01/2020**

JOB DESCRIPTION Laborer - Building **DISTRICT** 9

Page 22

Source: New York State 2020 Prevailing Wage Tables

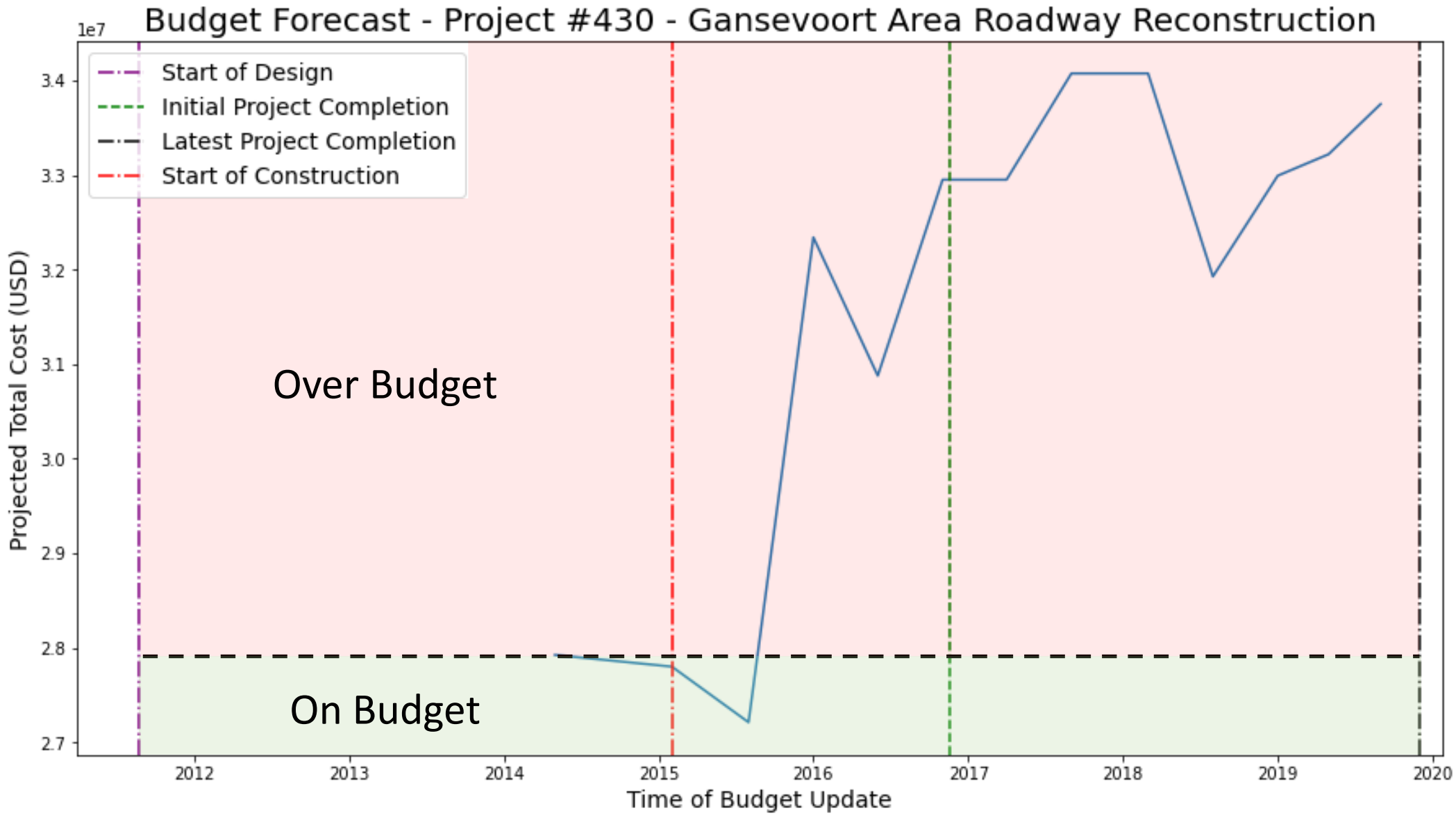


430

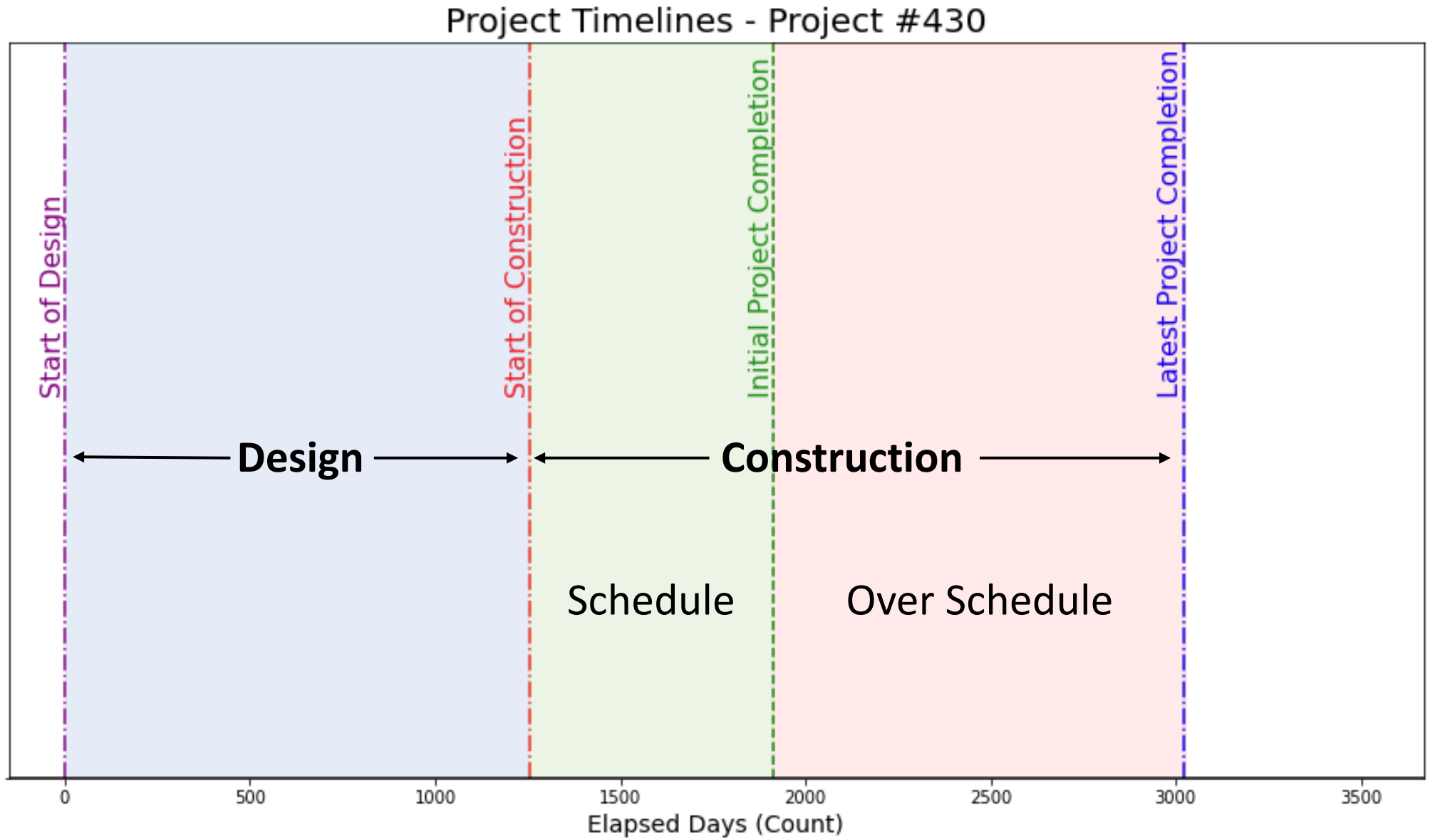
“Gansevoort Area Roadway Reconstruction”

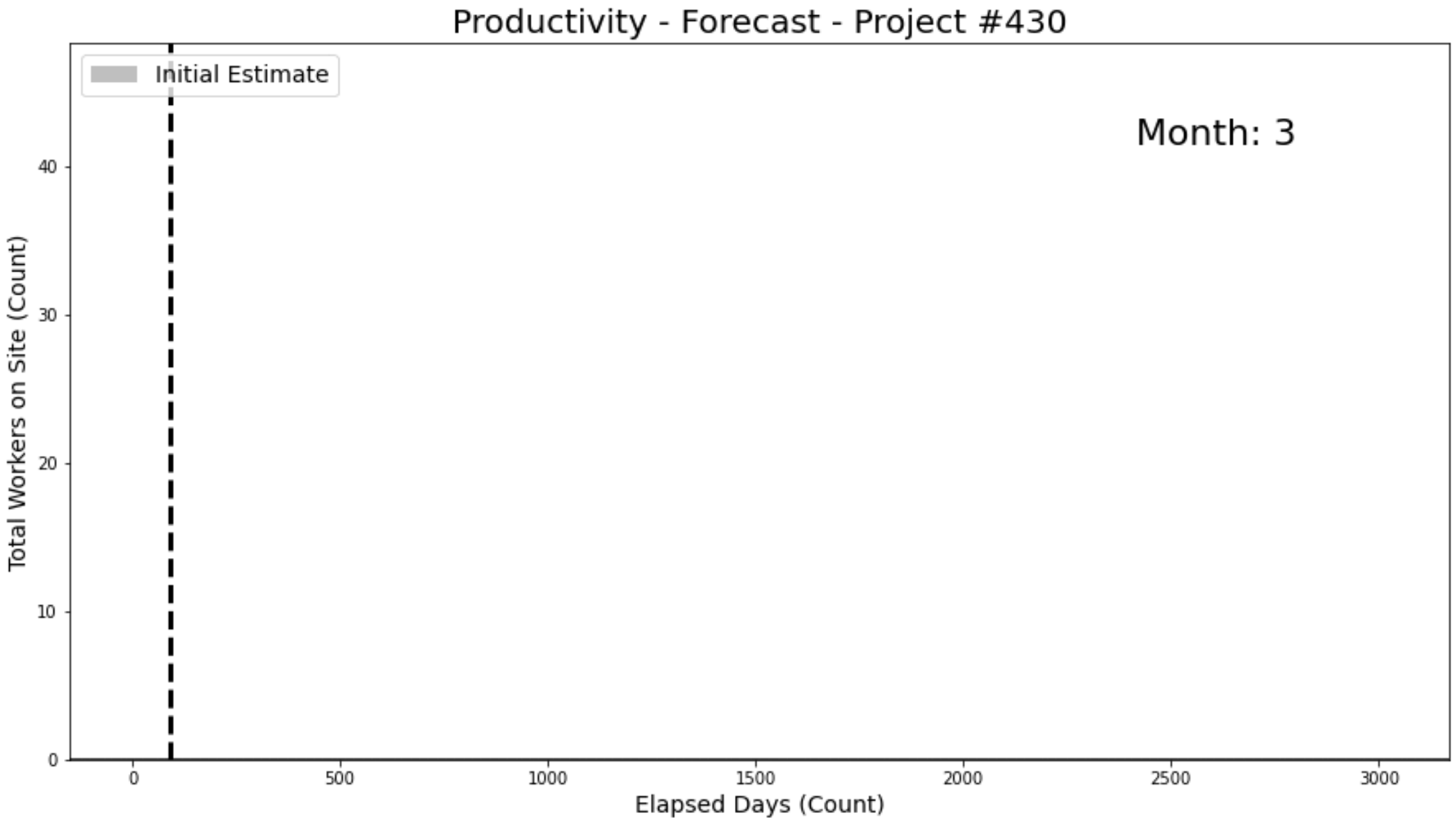
Streets and Roadways

Gansevoort Area Roadway Reconstruction

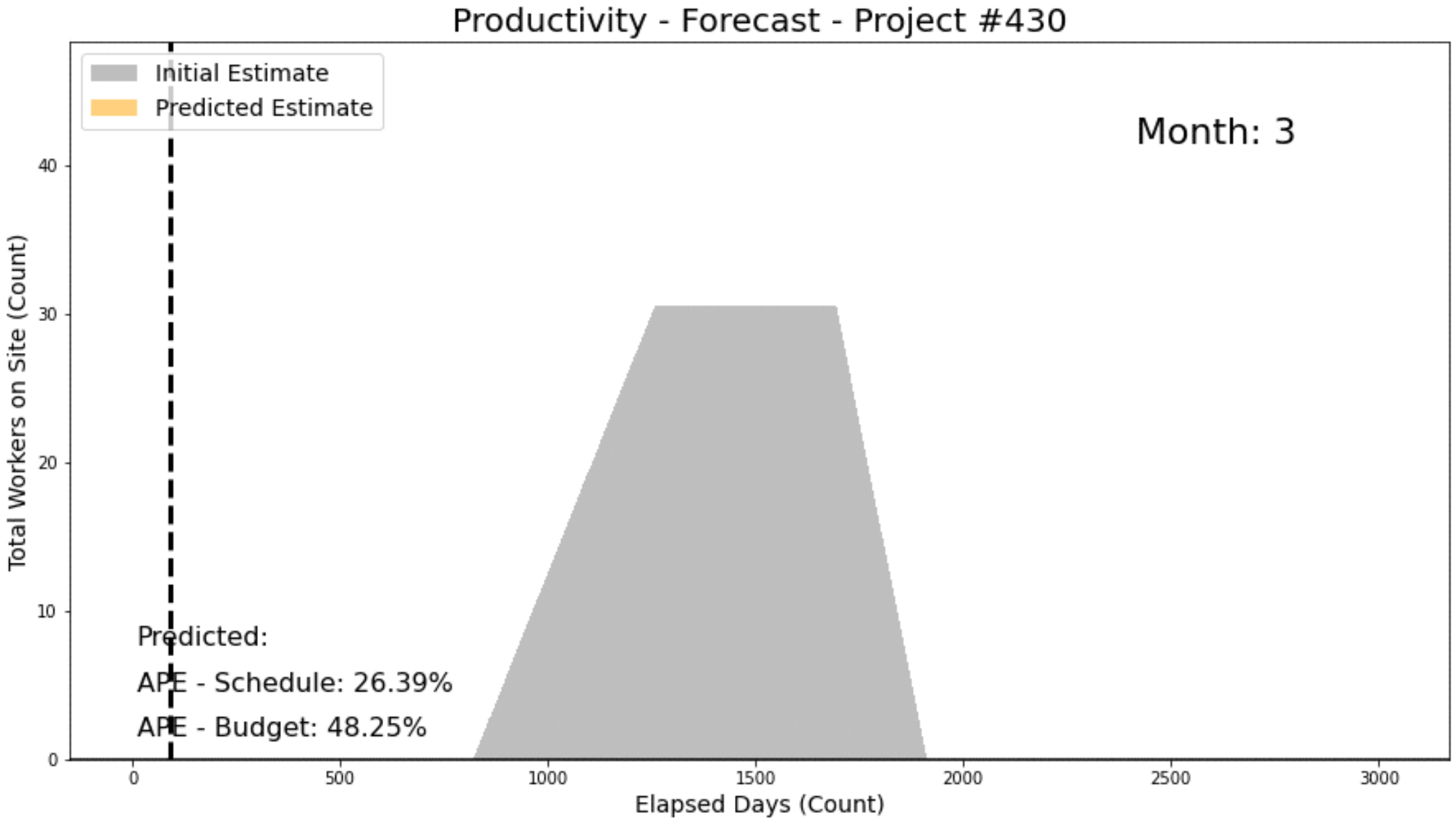


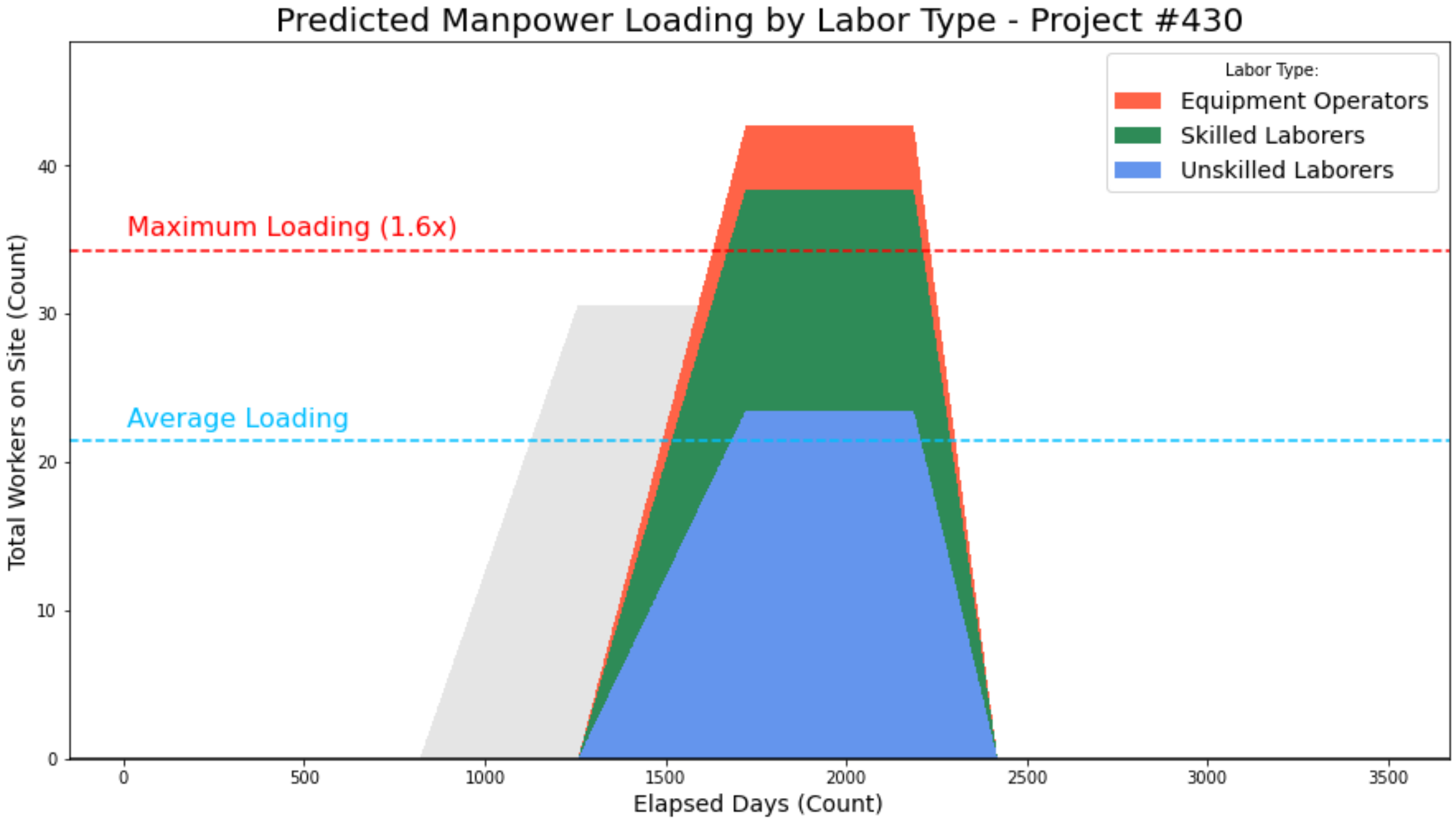
Gansevoort Area Roadway Reconstruction

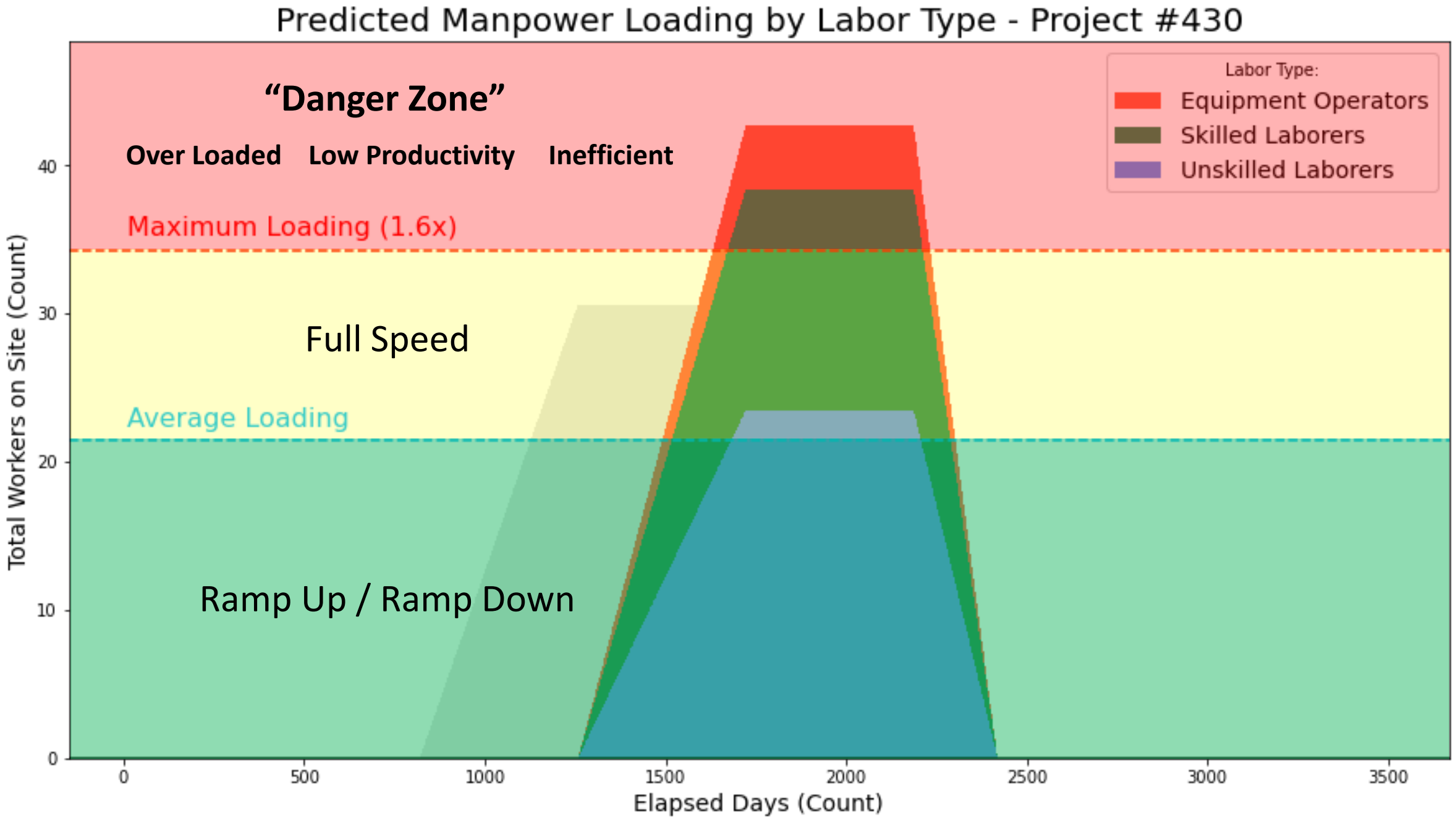


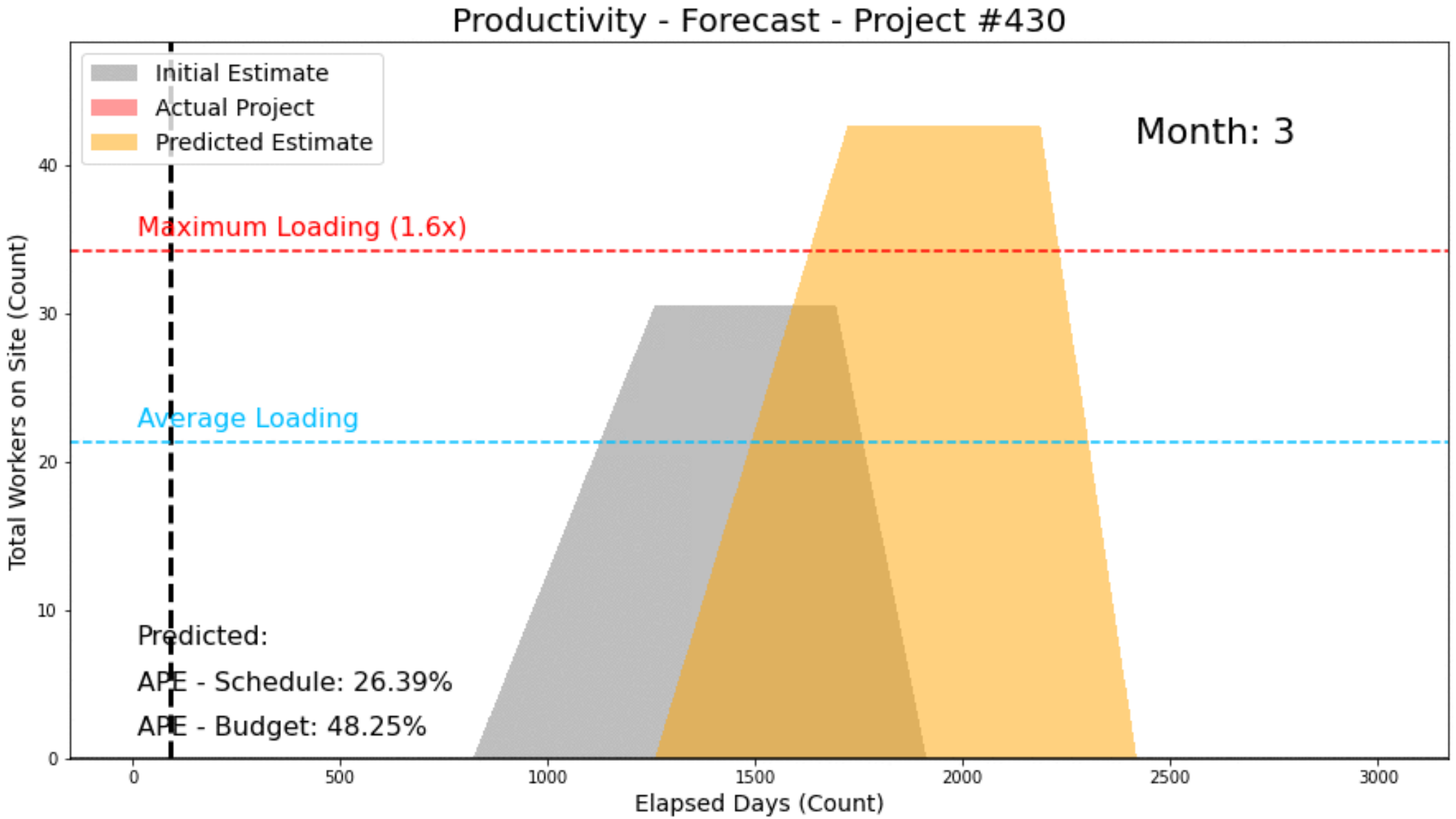

Productivity - Forecast - Project #430
Initial Estimate
Month: 3
Total Workers on Site (Count)
0
10
20
30
40
0
500
1000
1500
2000
2500
3000
Elapsed Days (Count)

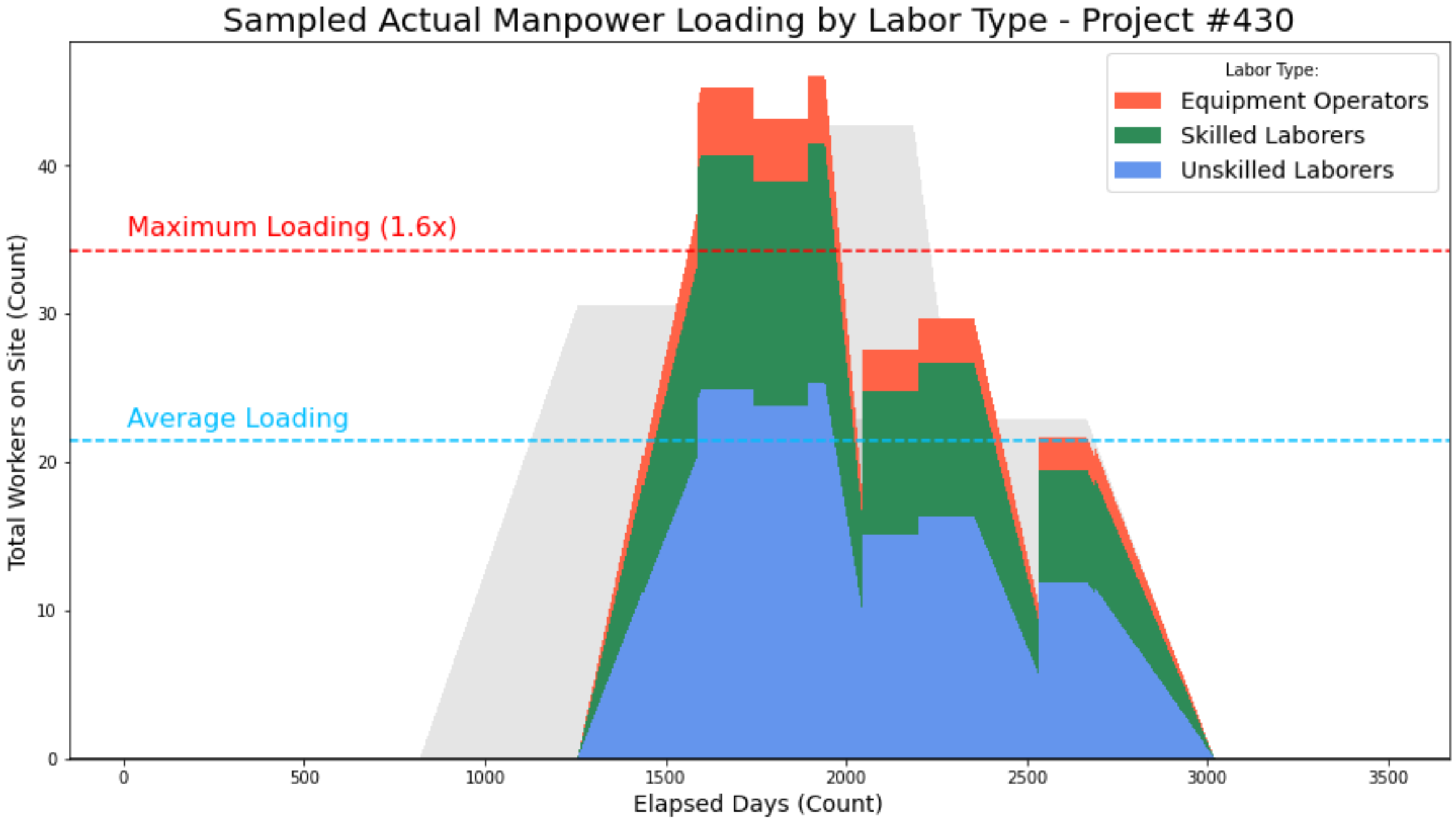
Gansevoort Area Roadway Reconstruction

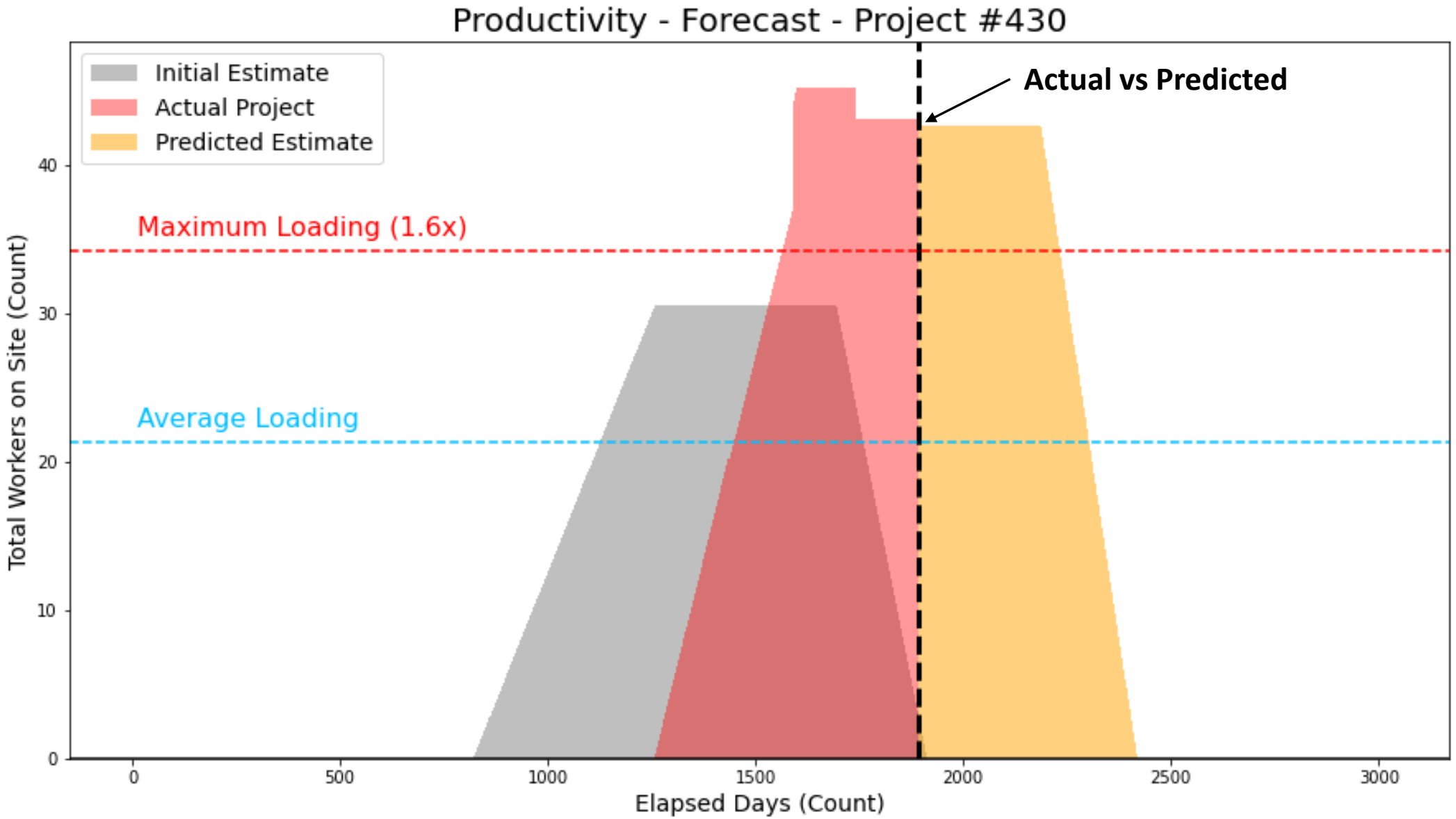


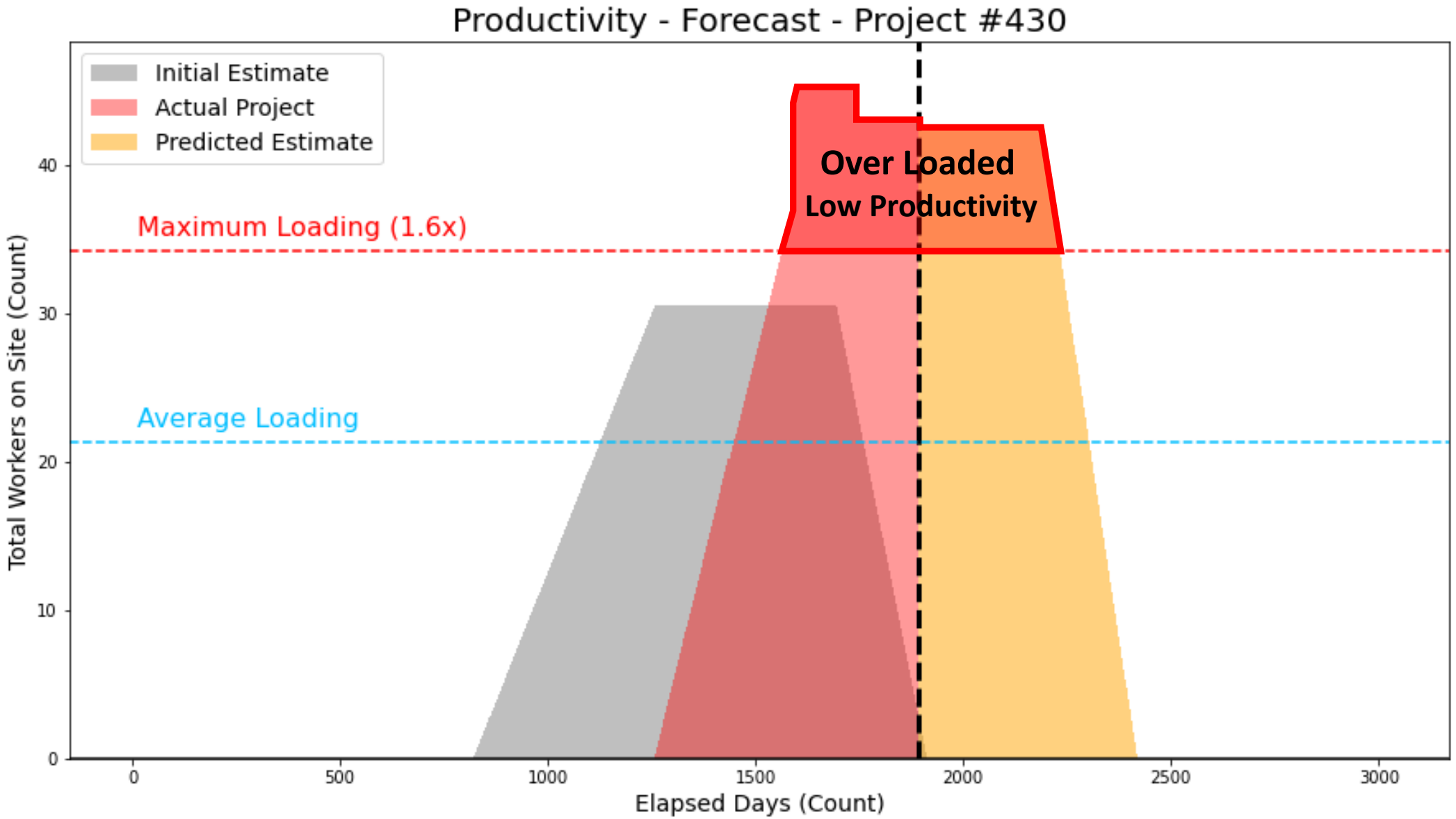

Predicted Manpower Loading by Labor Type - Project #430
Labor Type:
Equipment Operators
Skilled Laborers
Unskilled Laborers
Maximum Loading (1.6x)
Average Loading
Total Workers on Site (Count)
0
10
20
30
40
0
500
1000
1500
2000
2500
3000
3500
Elapsed Days (Count)

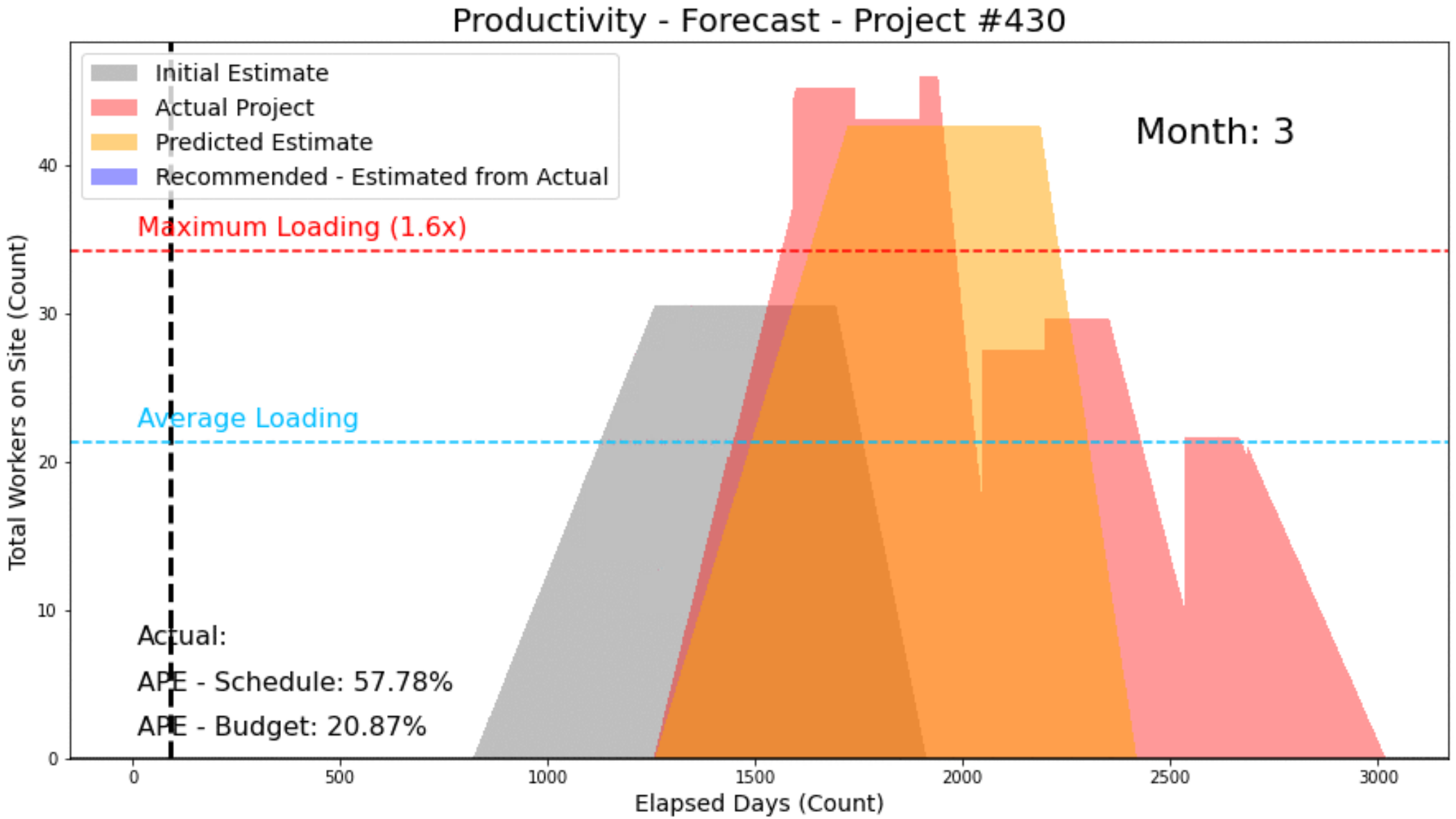

Predicted Manpower Loading by Labor Type - Project #430
"Danger Zone"
Over Loaded
Low Productivity
Inefficient
Maximum Loading (1.6x)
Full Speed
Average Loading
Ramp Up / Ramp Down
Labor Type:
Equipment Operators
Skilled Laborers
Unskilled Laborers
Total Workers on Site (Count)
Elapsed Days (Count)
0
10
20
30
40
0
500
1000
1500
2000
2500
3000
3500


Productivity - Forecast - Project #430
Initial Estimate
Actual Project
Predicted Estimate
Month: 3
Maximum Loading (1.6x)
Average Loading
Predicted:
APE - Schedule: 26.39%
APE - Budget: 48.25%
Total Workers on Site (Count)
Elapsed Days (Count)
0
10
20
30
40
0
500
1000
1500
2000
2500
3000

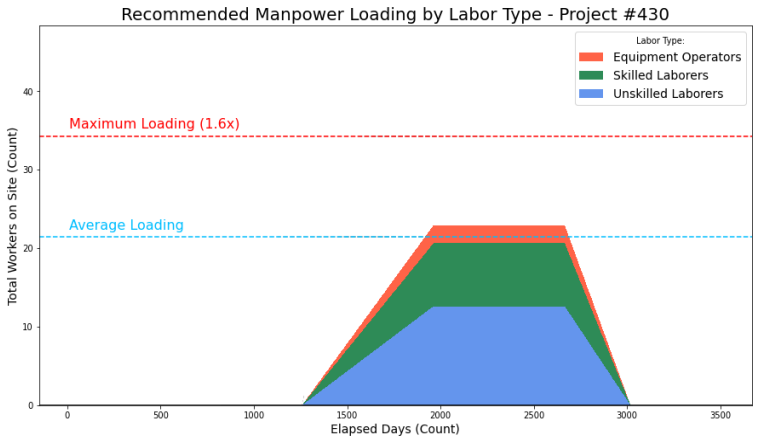

Sampled Actual Manpower Loading by Labor Type - Project #430
Labor Type:
Equipment Operators
Skilled Laborers
Unskilled Laborers
Maximum Loading (1.6x)
Average Loading
Total Workers on Site (Count)
0
10
20
30
40
0
500
1000
1500
2000
2500
3000
3500
Elapsed Days (Count)


Productivity - Forecast - Project #430
Initial Estimate
Actual Project
Predicted Estimate
Actual vs Predicted
Maximum Loading (1.6x)
Average Loading
Total Workers on Site (Count)
Elapsed Days (Count)
0
10
20
30
40
0
500
1000
1500
2000
2500
3000

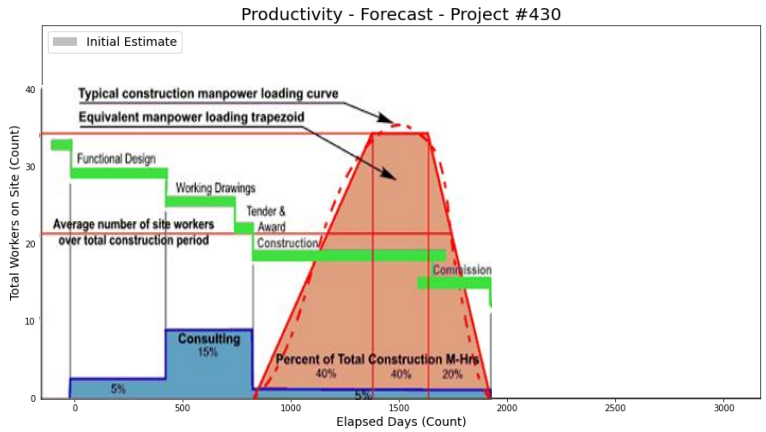

Productivity - Forecast - Project #430
Initial Estimate
Actual Project
Predicted Estimate
Over Loaded
Low Productivity
Maximum Loading (1.6x)
Average Loading
Total Workers on Site (Count)
Elapsed Days (Count)
0
10
20
30
40
0
500
1000
1500
2000
2500
3000


Productivity - Forecast - Project #430
Initial Estimate
Actual Project
Predicted Estimate
Recommended - Estimated from Actual
Month: 3
Maximum Loading (1.6x)
Average Loading
Actual:
APE - Schedule: 57.78%
APE - Budget: 20.87%
Total Workers on Site (Count)
Elapsed Days (Count)
0
10
20
30
40
0
500
1000
1500
2000
2500
3000

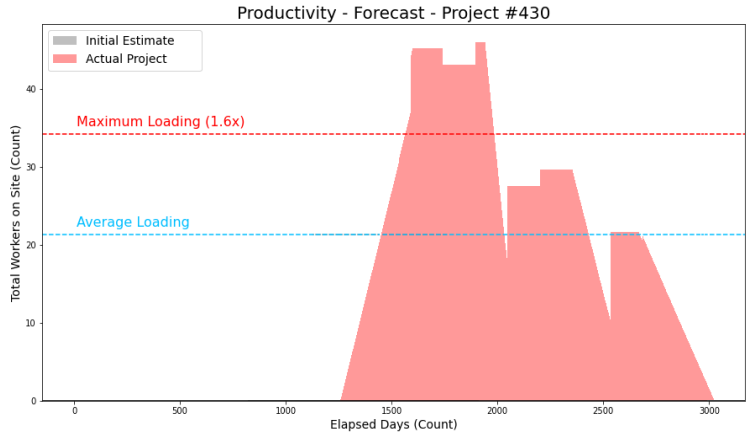
Posterior ∝ Prior x Likelihood

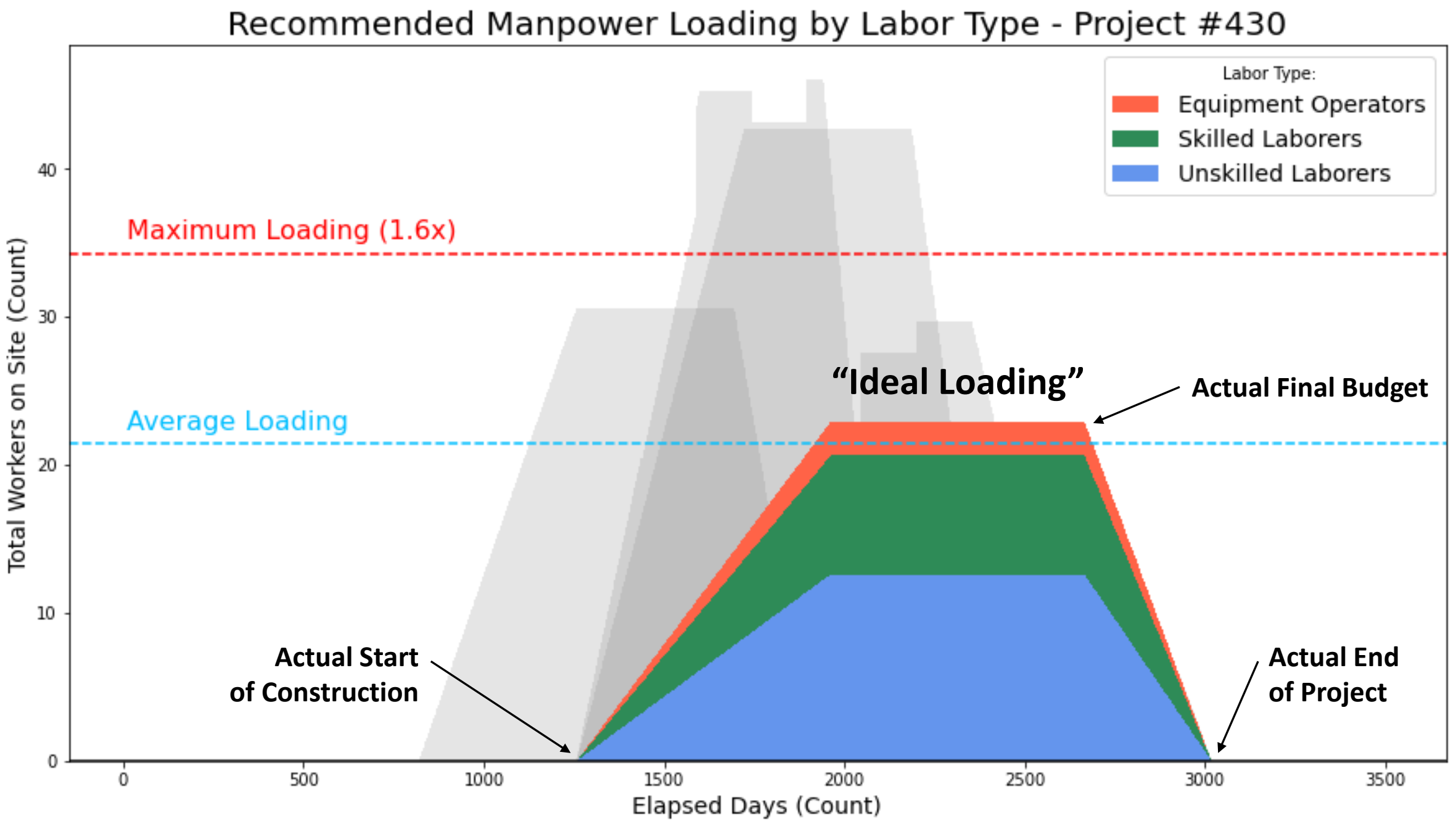


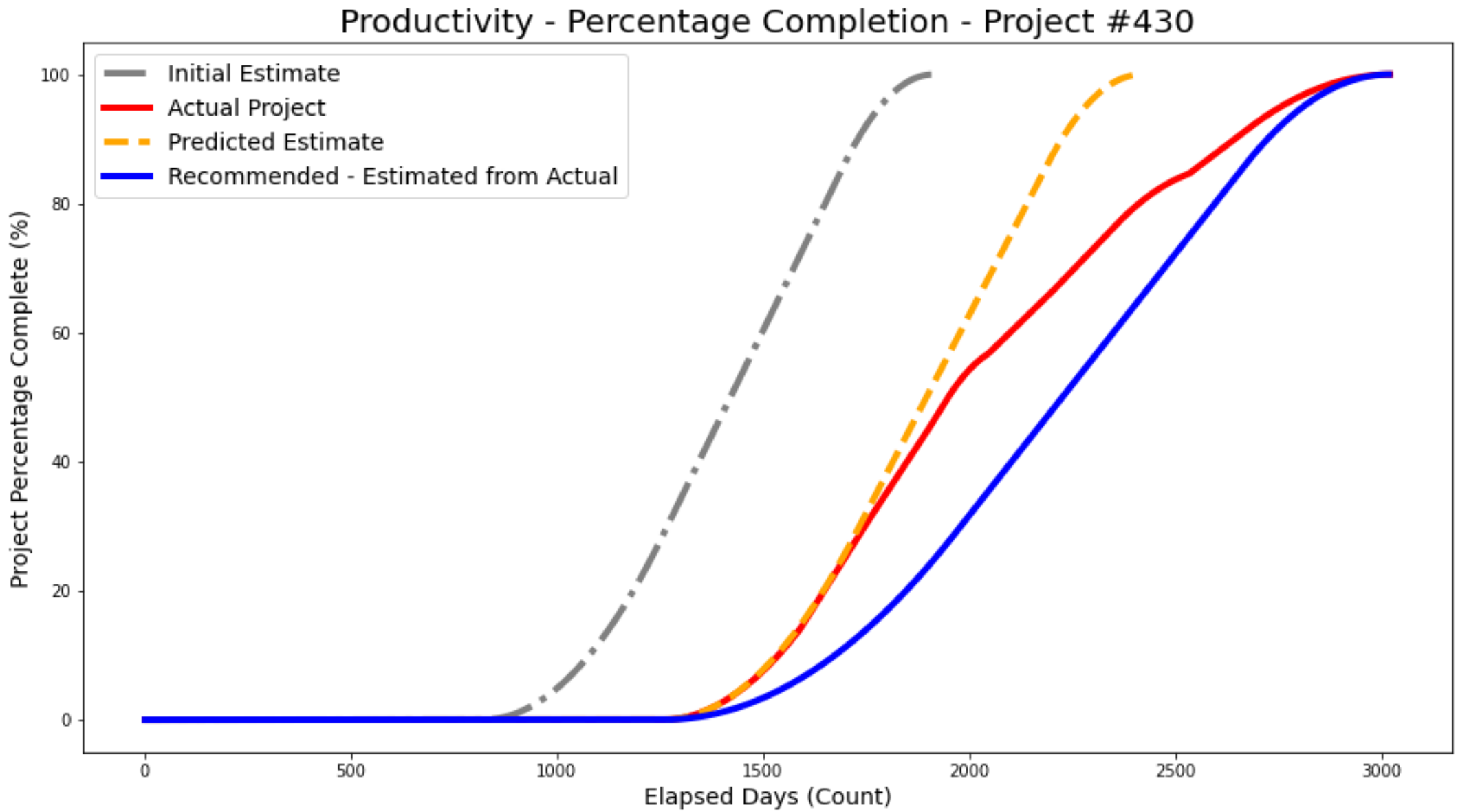
∝

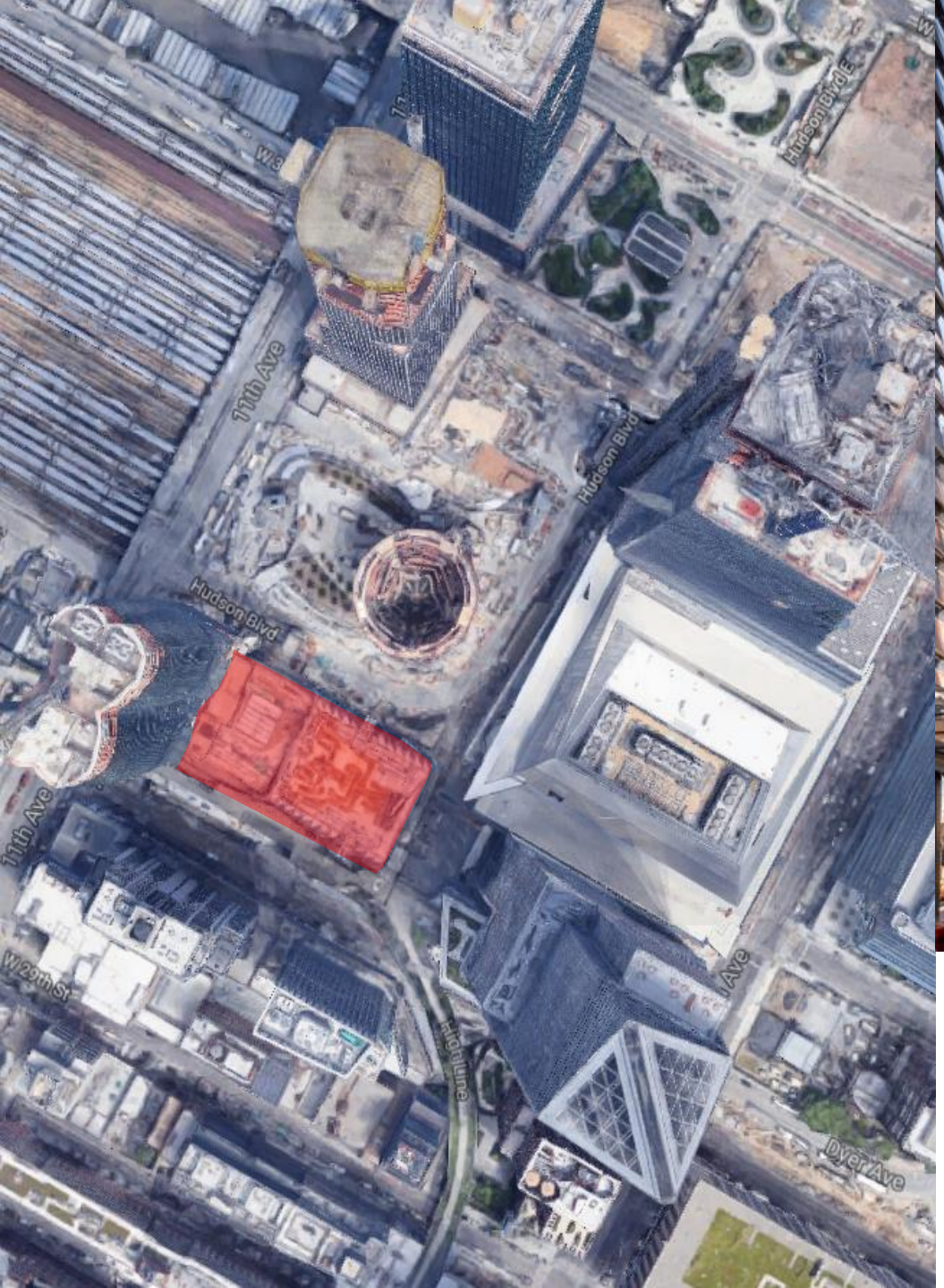


x




Recommended Manpower Loading by Labor Type - Project #430
Labor Type:
Equipment Operators
Skilled Laborers
Unskilled Laborers
Maximum Loading (1.6x)
Average Loading
"Ideal Loading"
Actual Final Budget
Actual Start of Construction
Actual End of Project
Total Workers on Site (Count)
Elapsed Days (Count)
0
10
20
30
40
0
500
1000
1500
2000
2500
3000
3500

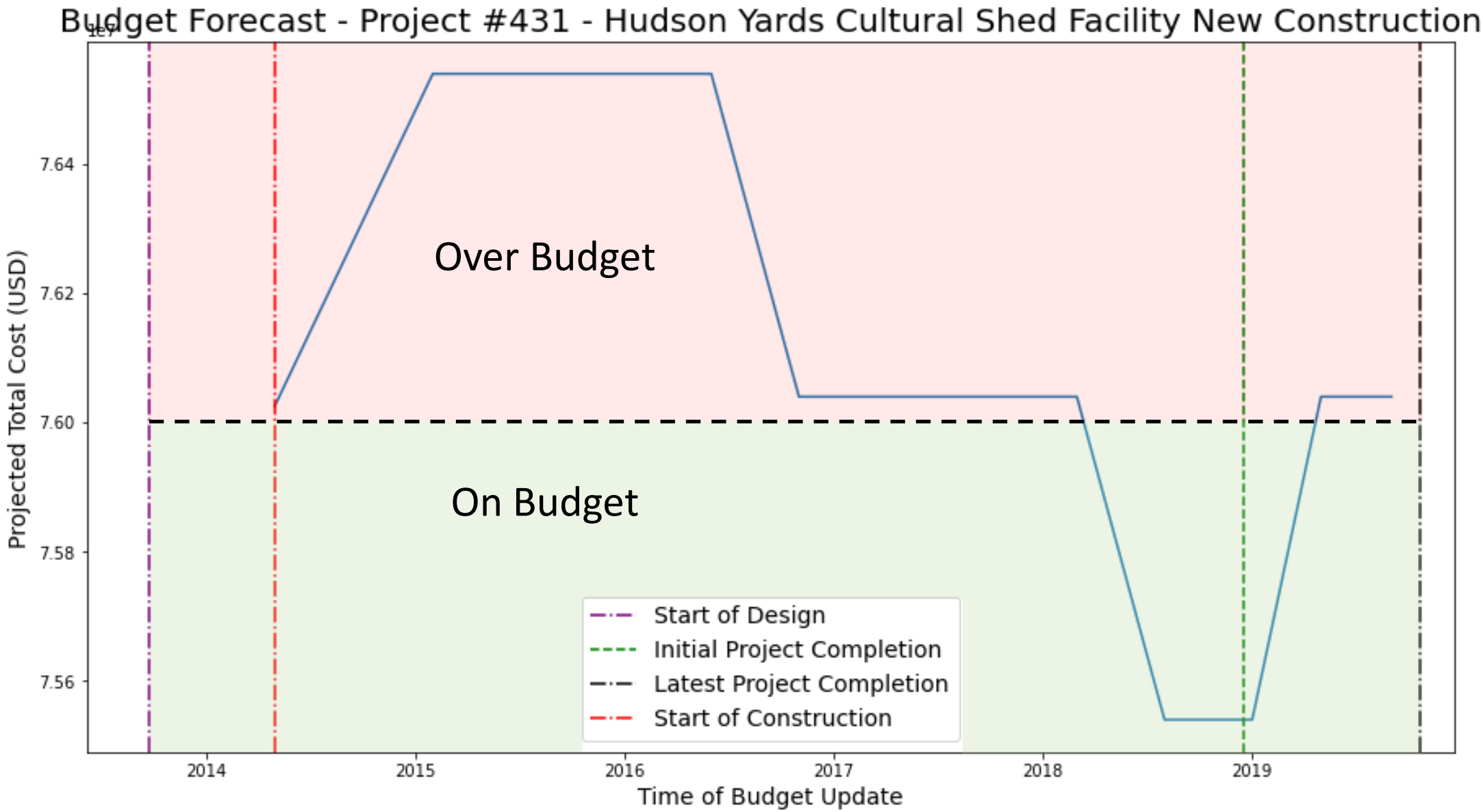

Productivity - Percentage Completion - Project #430
Initial Estimate
Actual Project
Predicted Estimate
Recommended - Estimated from Actual
Project Percentage Complete (%)
0
20
40
60
80
100
0
500
1000
1500
2000
2500
3000
Elapsed Days (Count)

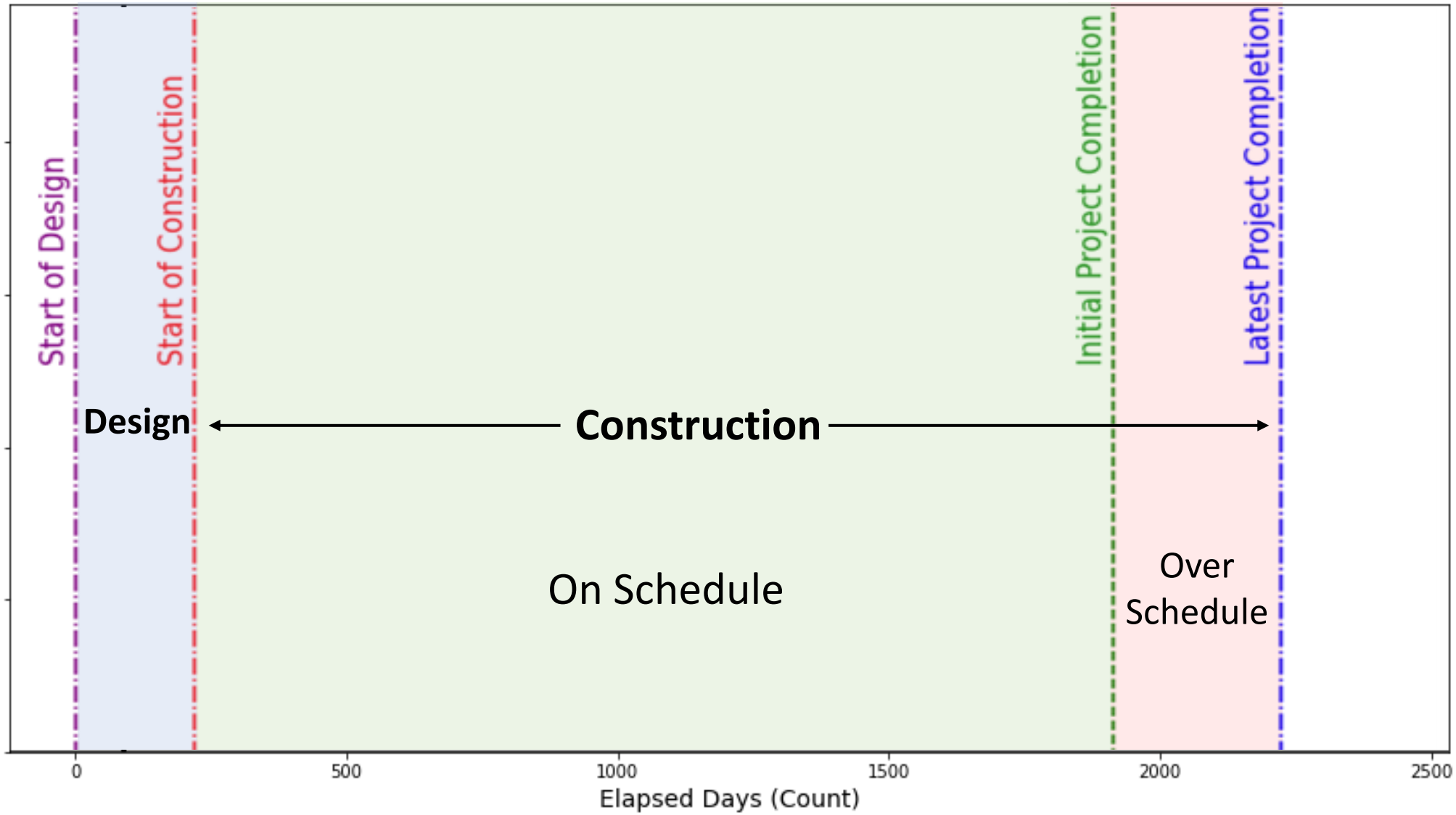


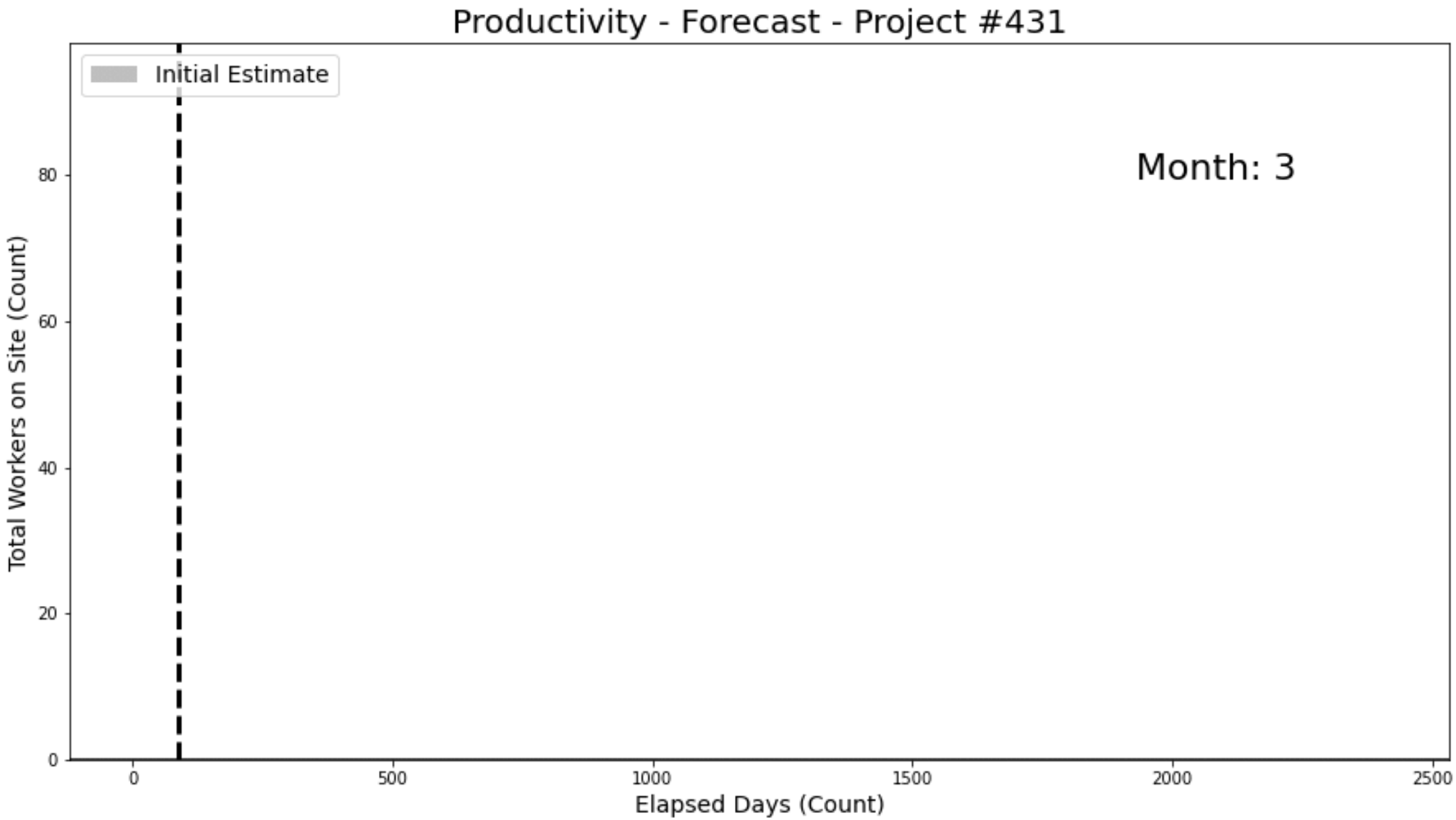
431

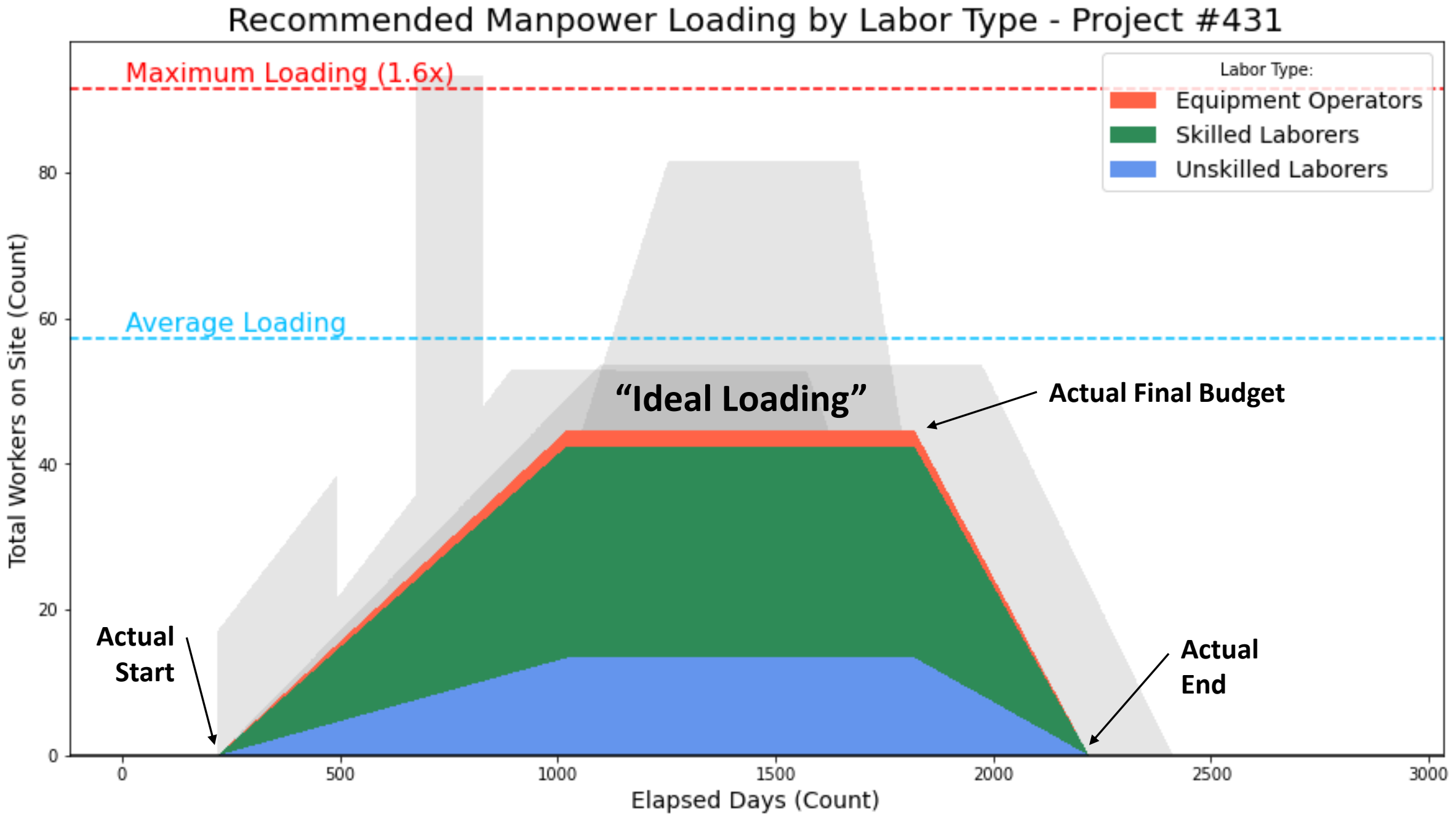
“Hudson Yards Cultural Shed”

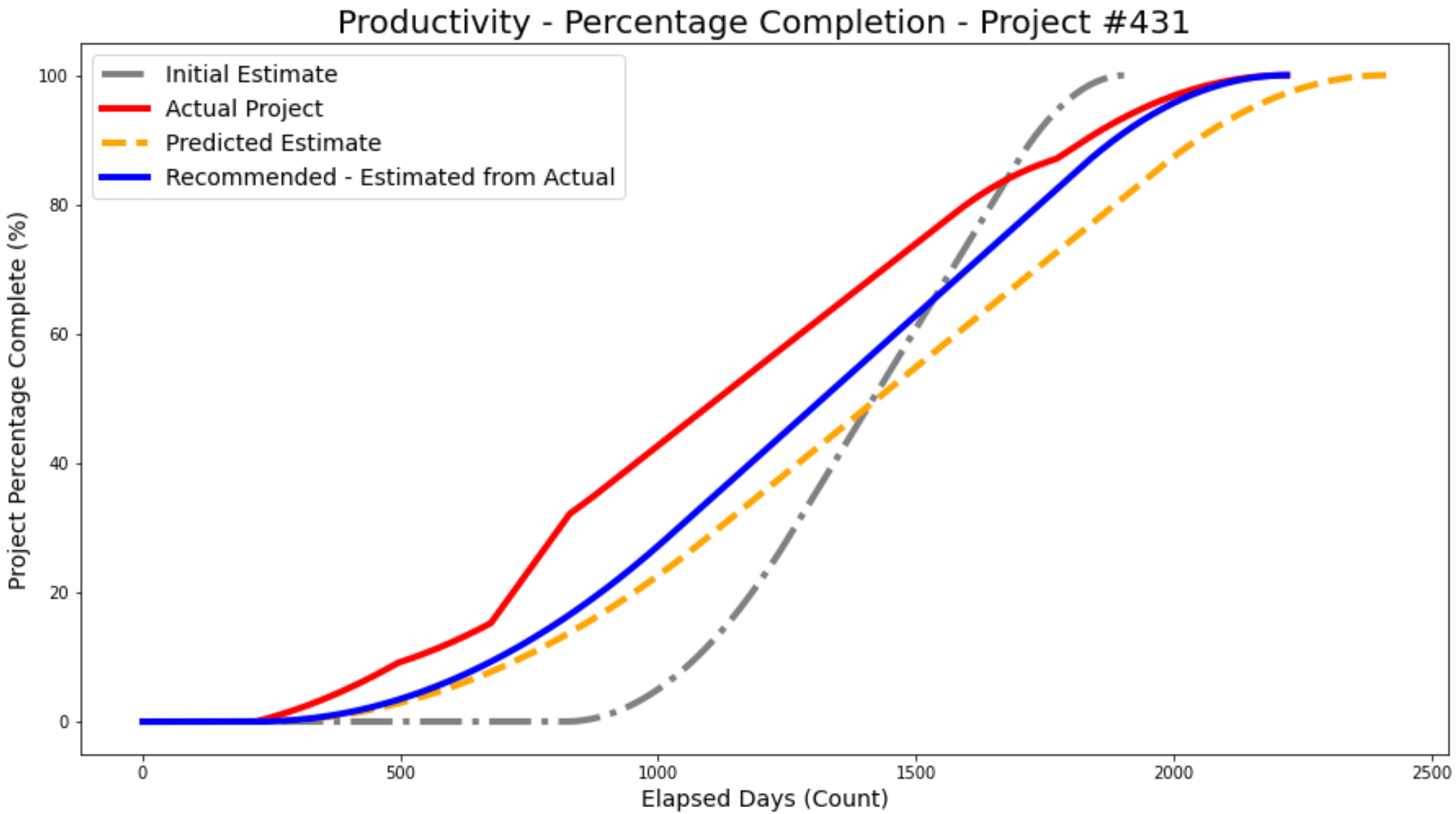
Building – New Construction


Budget Forecast - Project #431 - Hudson Yards Cultural Shed Facility New Construction
1e9
7.64
7.62
7.60
7.58
7.56
Projected Total Cost (USD)
Over Budget
On Budget
Start of Design
Initial Project Completion
Latest Project Completion
Start of Construction
2014
2015
2016
2017
2018
2019
Time of Budget Update


Start of Design
Start of Construction
Initial Project Completion
Latest Project Completion
Design
Construction
On Schedule
Over Schedule
0
500
1000
1500
2000
2500
Elapsed Days (Count)


Productivity - Forecast - Project #431
Initial Estimate
Month: 3
Total Workers on Site (Count)
0
20
40
60
80
0
500
1000
1500
2000
2500
Elapsed Days (Count)


Recommended Manpower Loading by Labor Type - Project #431
Maximum Loading (1.6x)
Average Loading
Labor Type:
Equipment Operators
Skilled Laborers
Unskilled Laborers
"Ideal Loading"
Actual Final Budget
Actual Start
Actual End
Total Workers on Site (Count)
Elapsed Days (Count)
0
20
40
60
80
0
500
1000
1500
2000
2500
3000

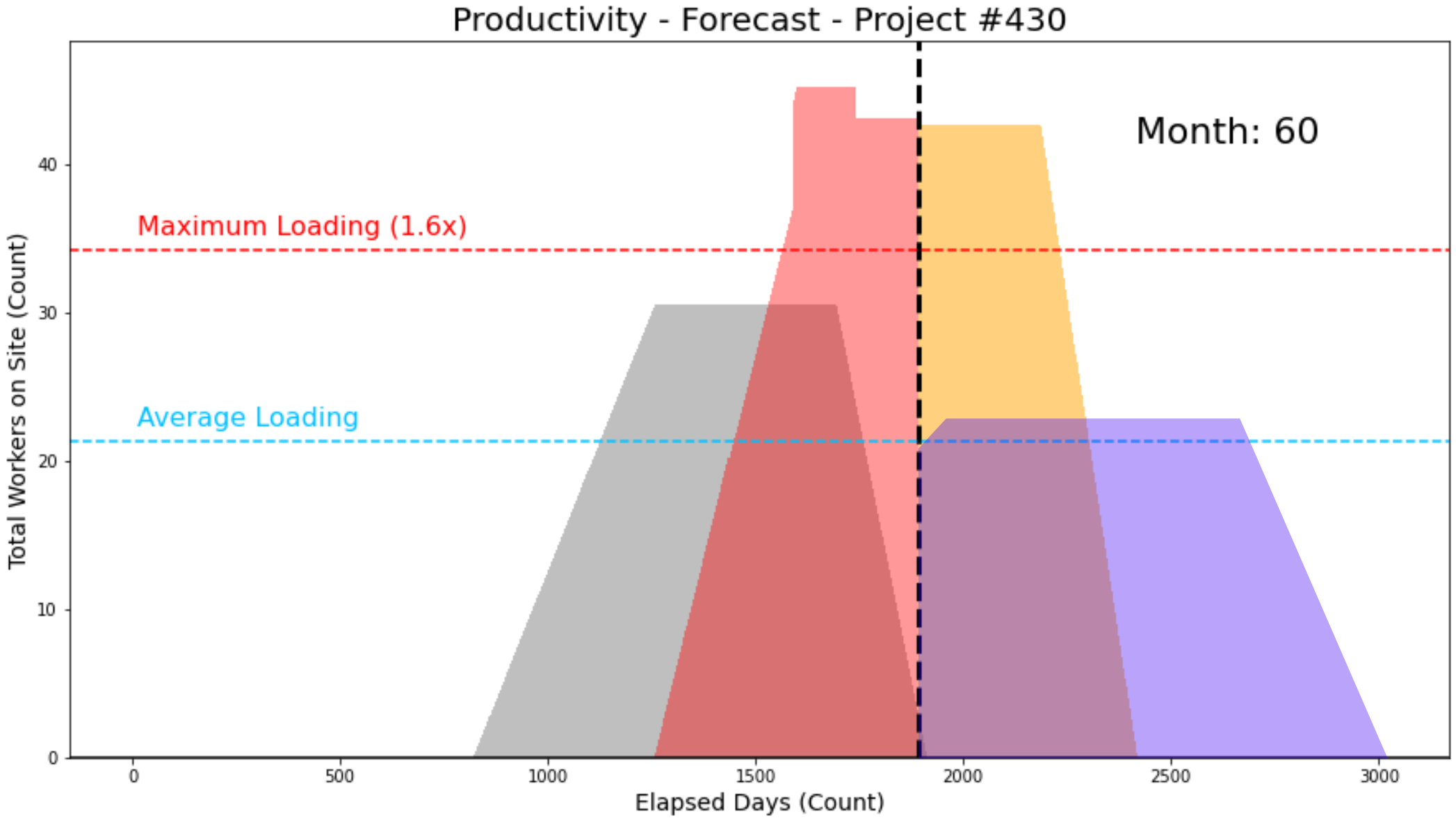

Productivity - Percentage Completion - Project #431
Initial Estimate
Actual Project
Predicted Estimate
Recommended - Estimated from Actual
Project Percentage Complete (%)
0
20
40
60
80
100
0
500
1000
1500
2000
2500
Elapsed Days (Count)

Conclusion

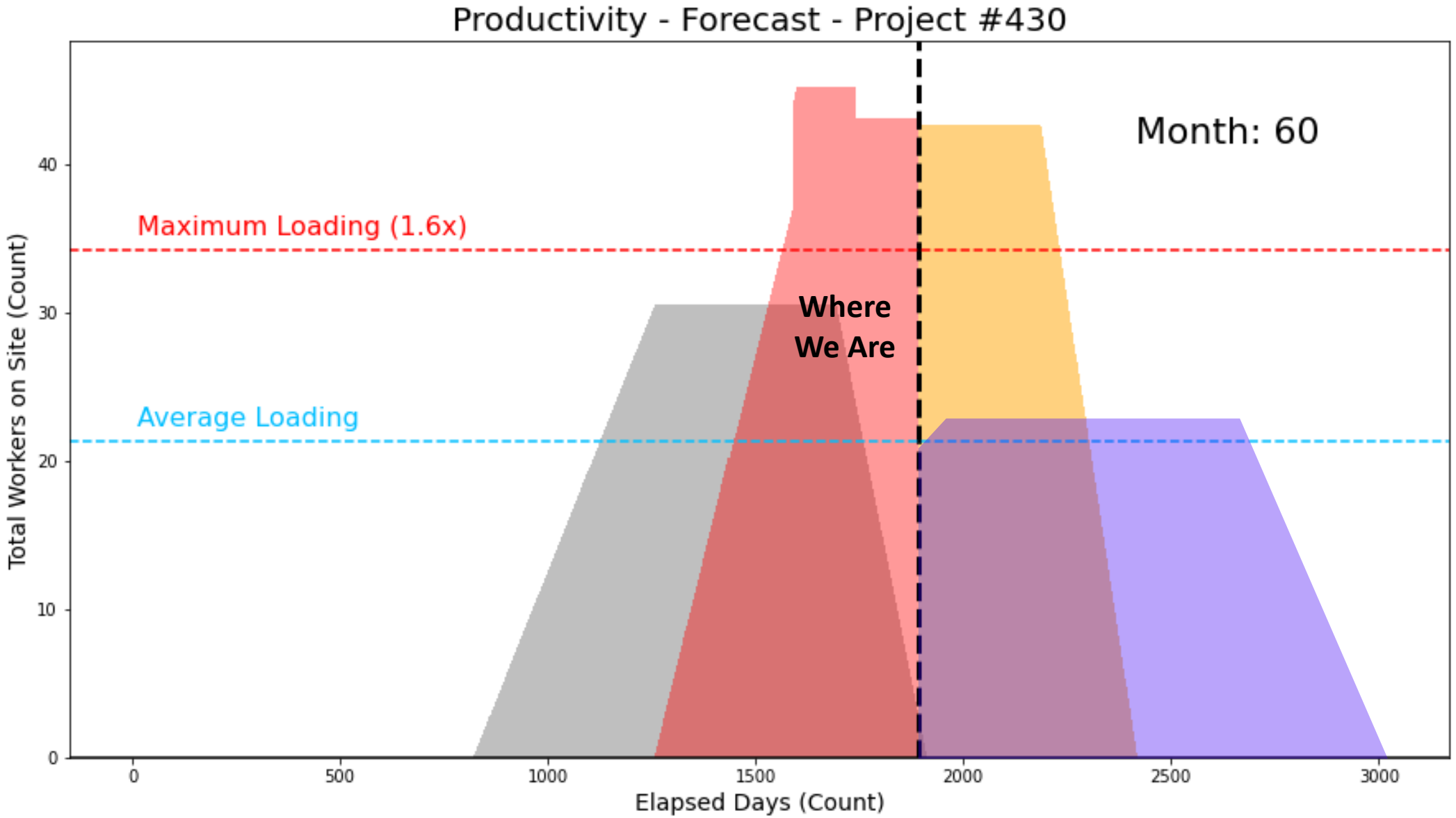
5

Understanding Project Success

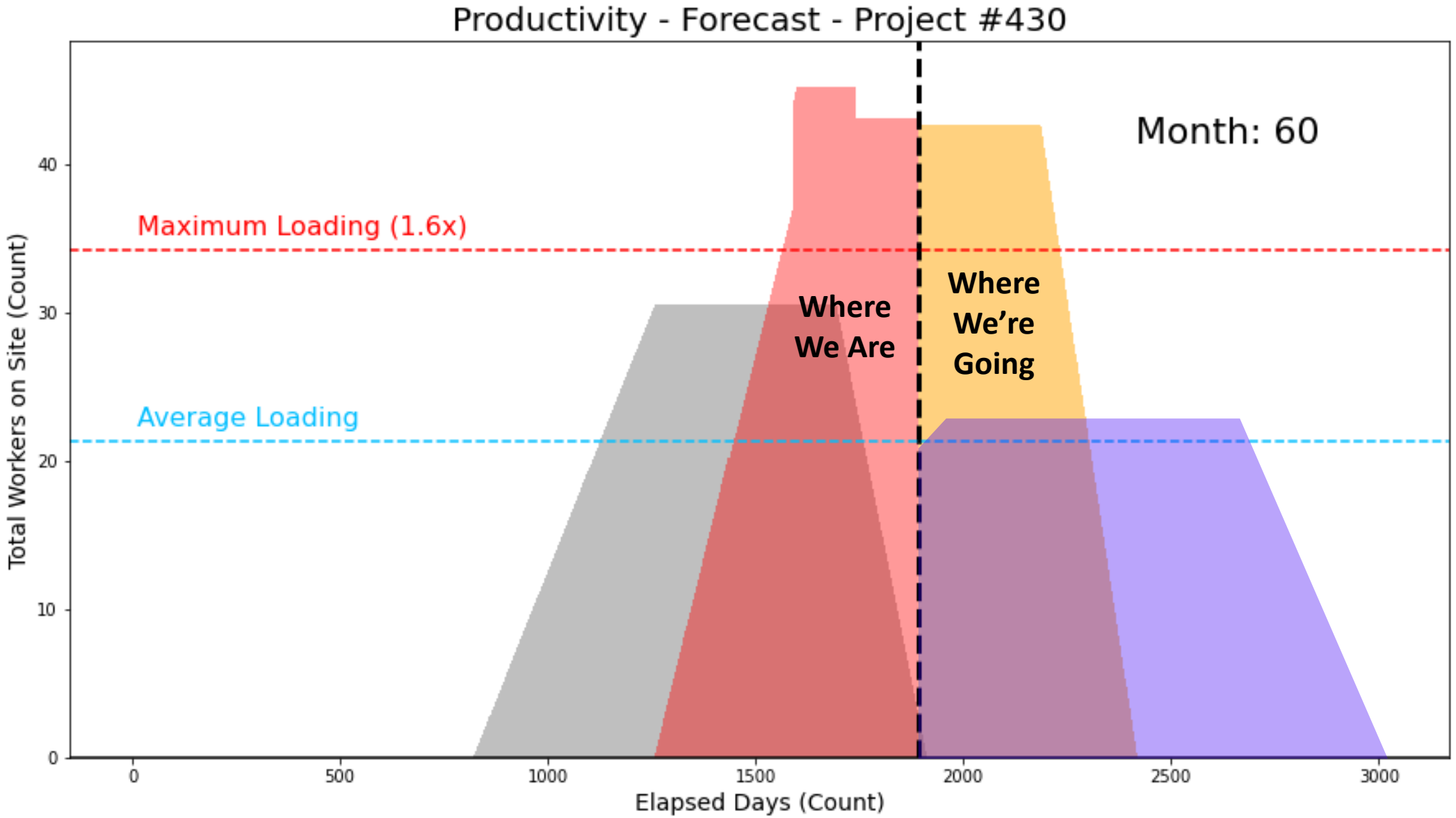
How can we make our project successful?



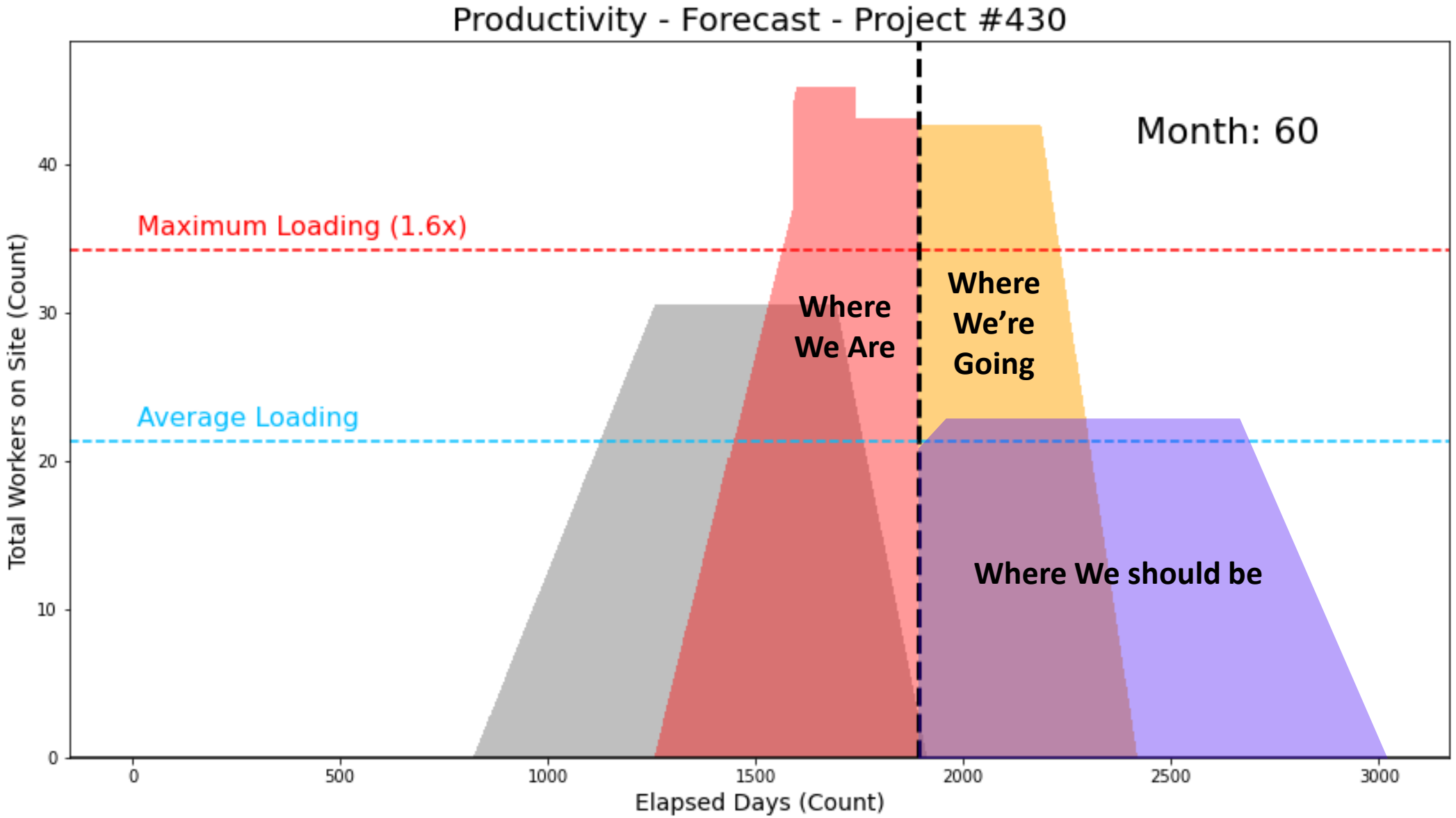
How can we make our project successful?



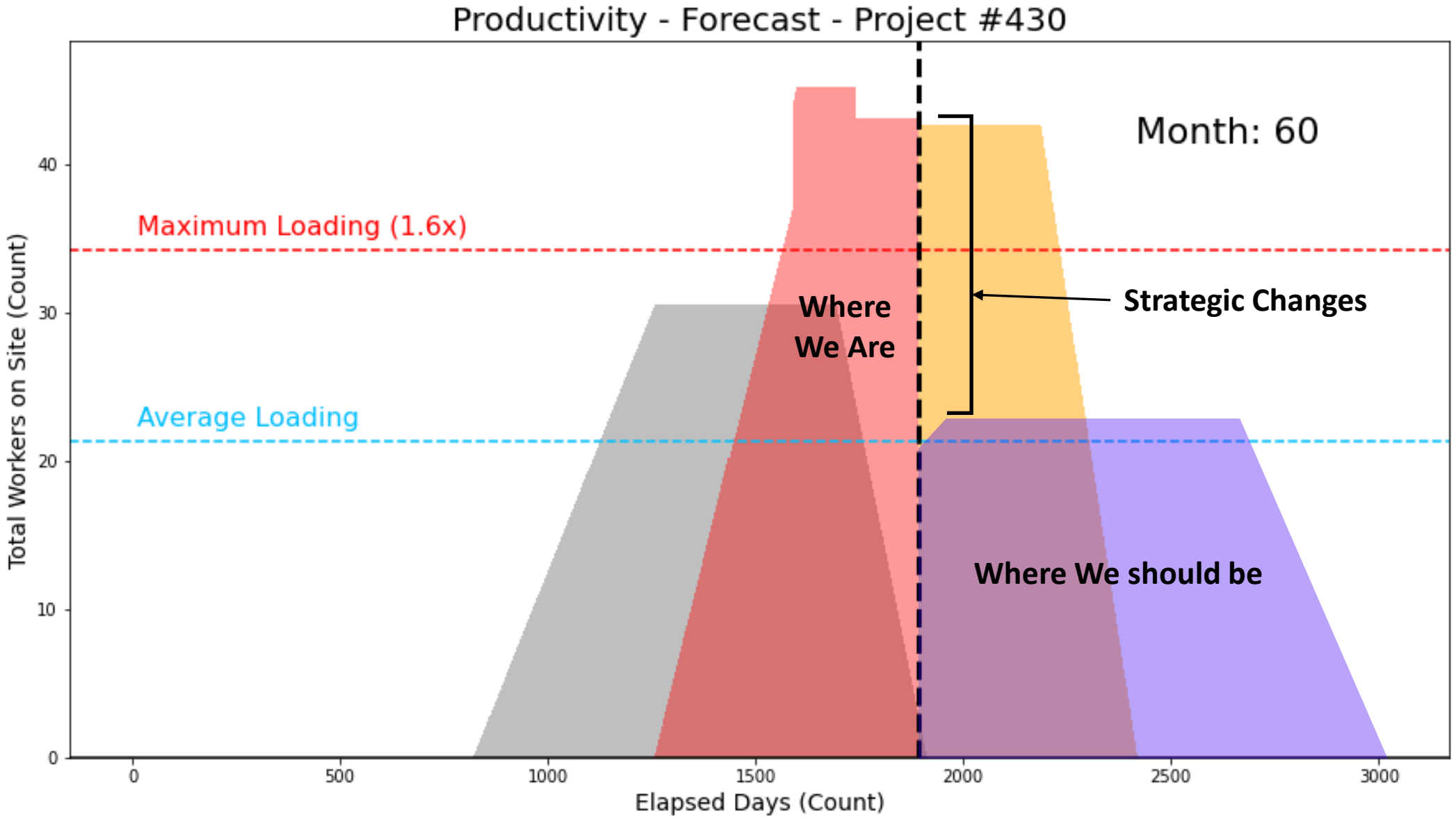
How can we make our project successful?



How can we make our project successful?

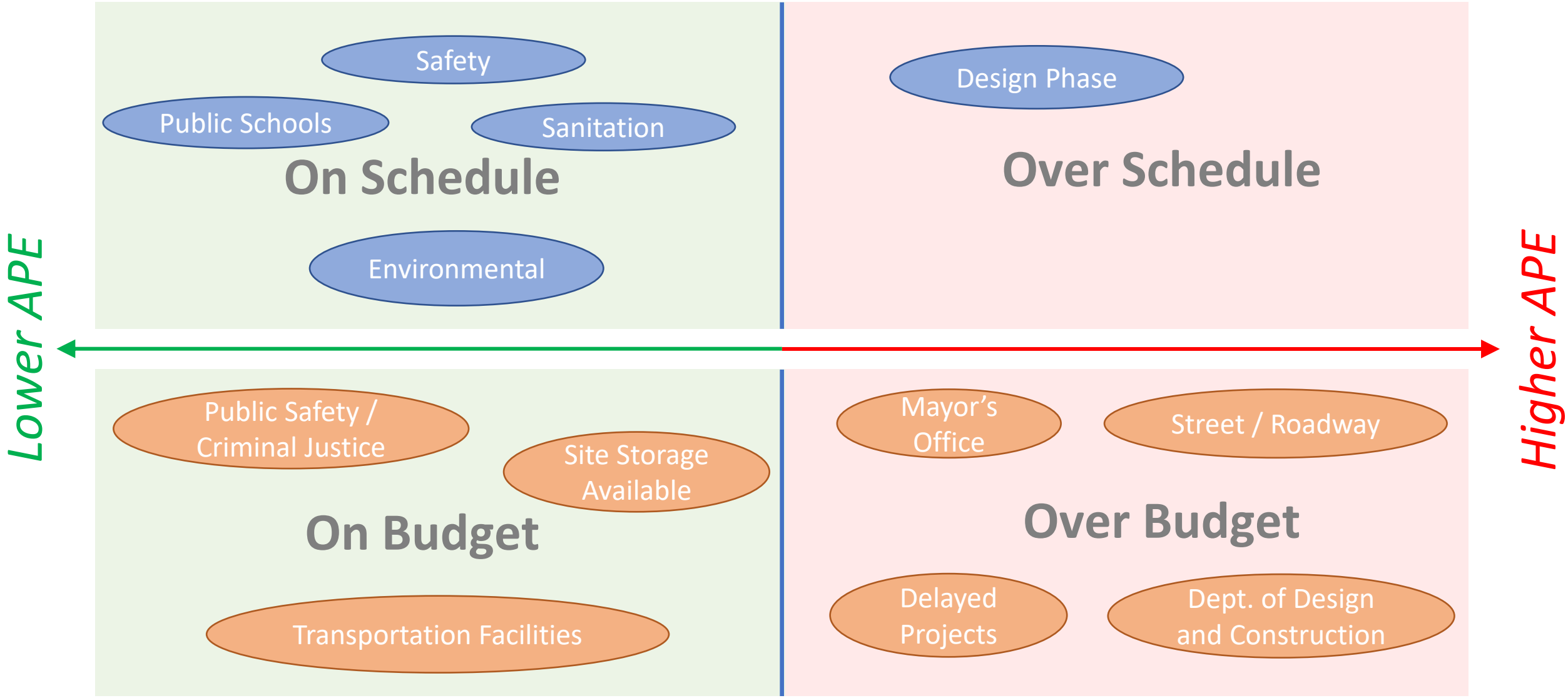


How can we make our project successful?

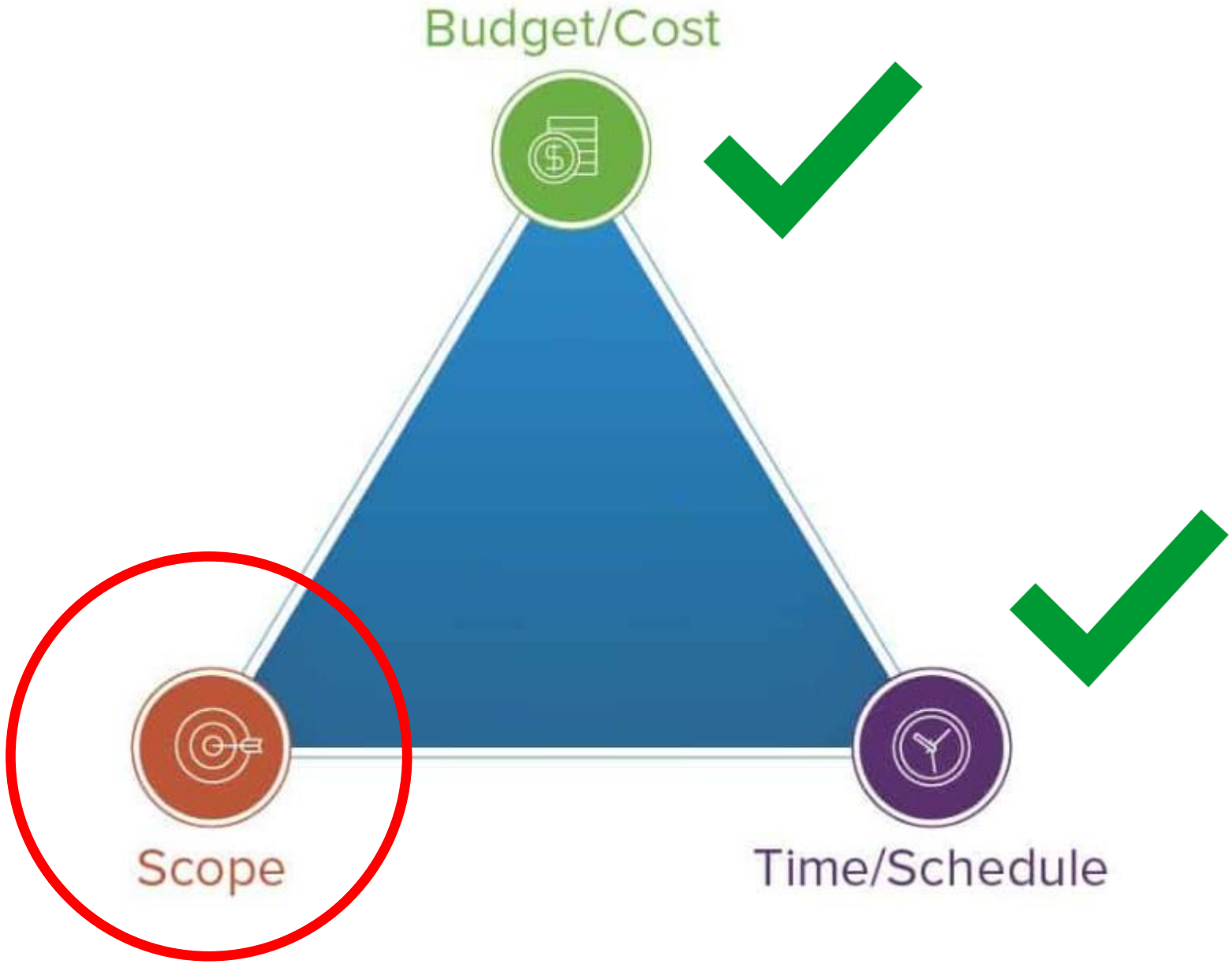


Project Success Indicators

Per APE Linear Models



Next Steps



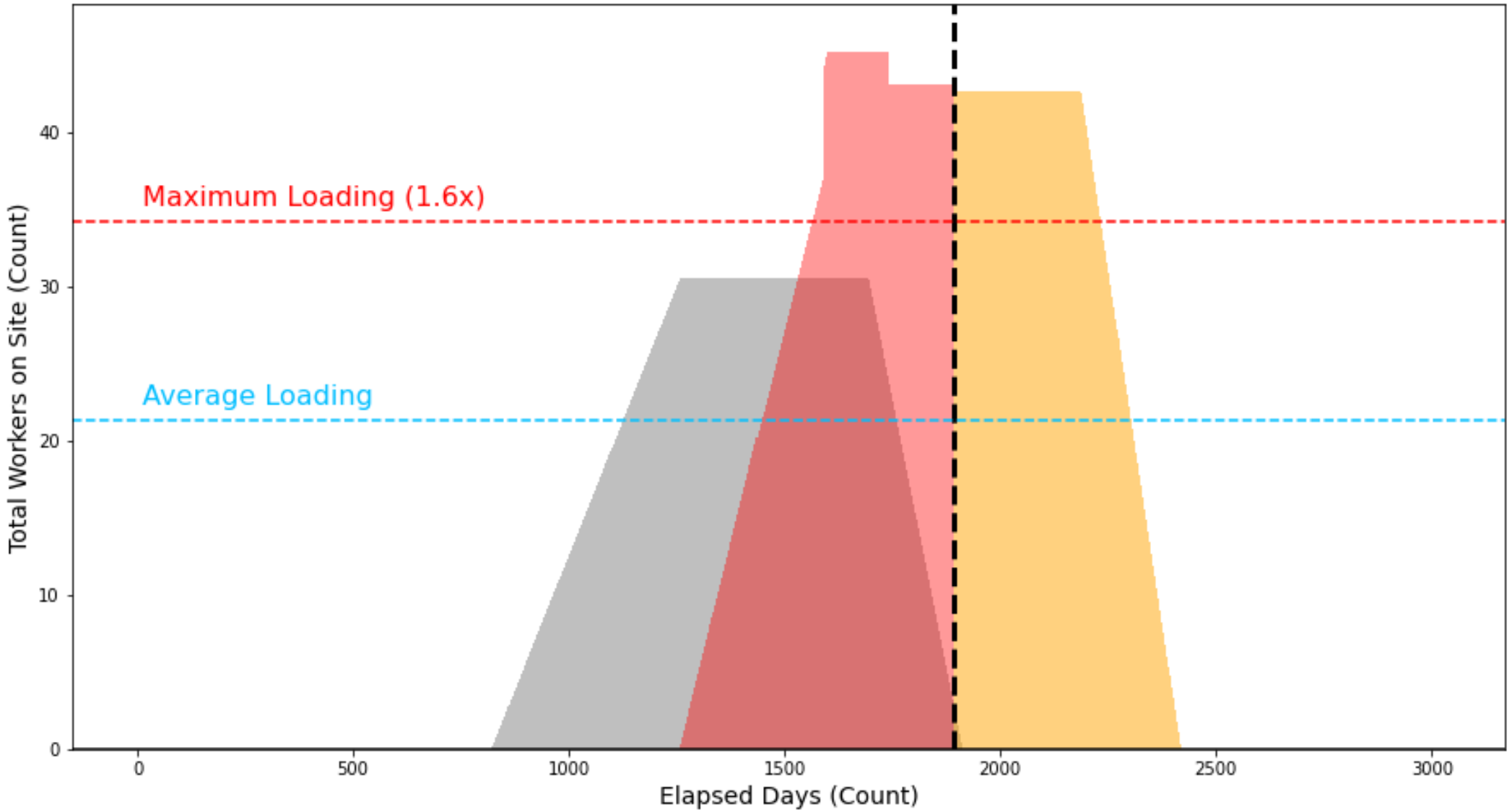
Scope Changes are Unknown

Next Step:

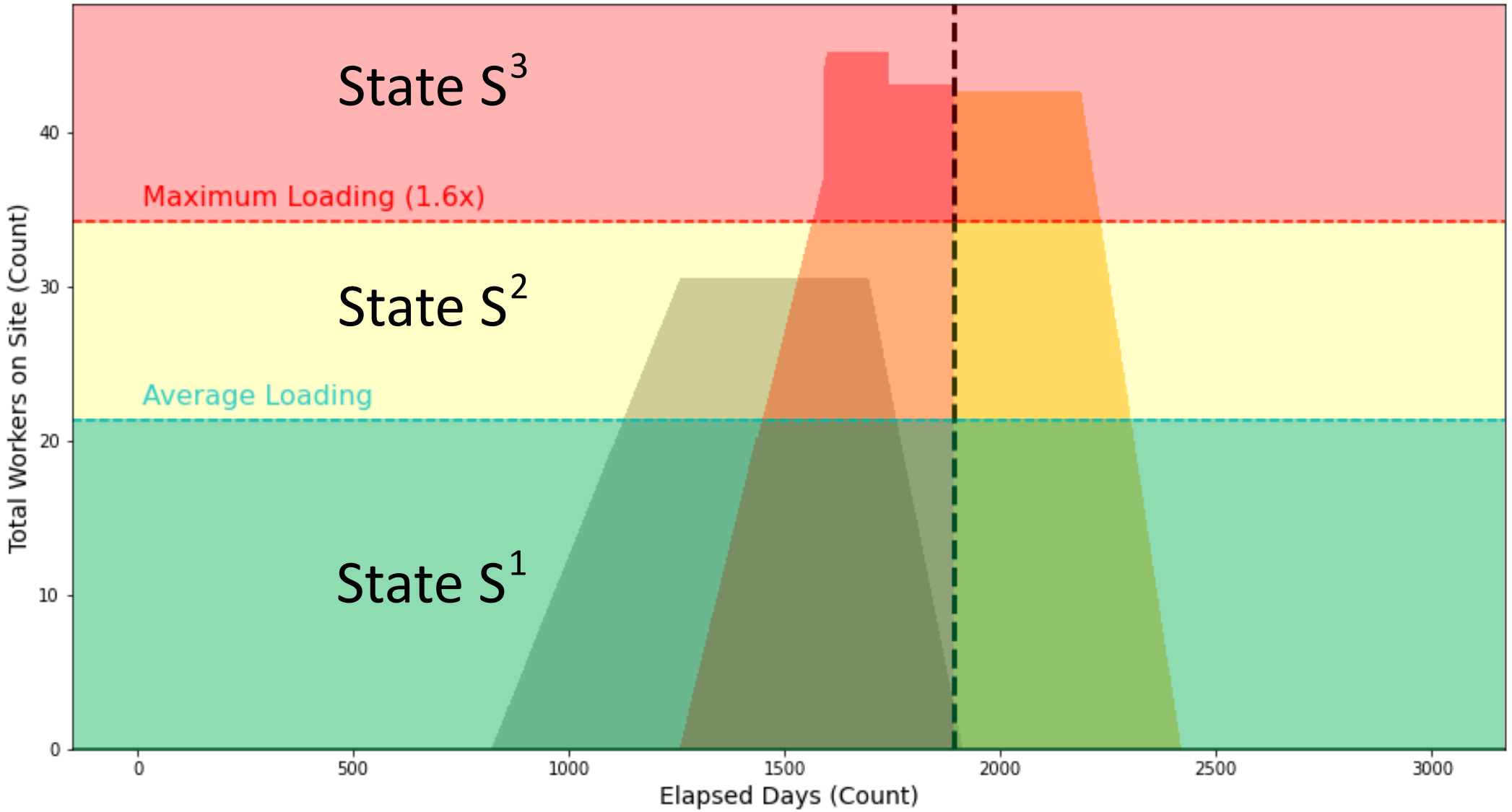
Find scope specific info

- *Actual Critical Path*
- *Specific tasks list*
- *Actual Gantt chart*

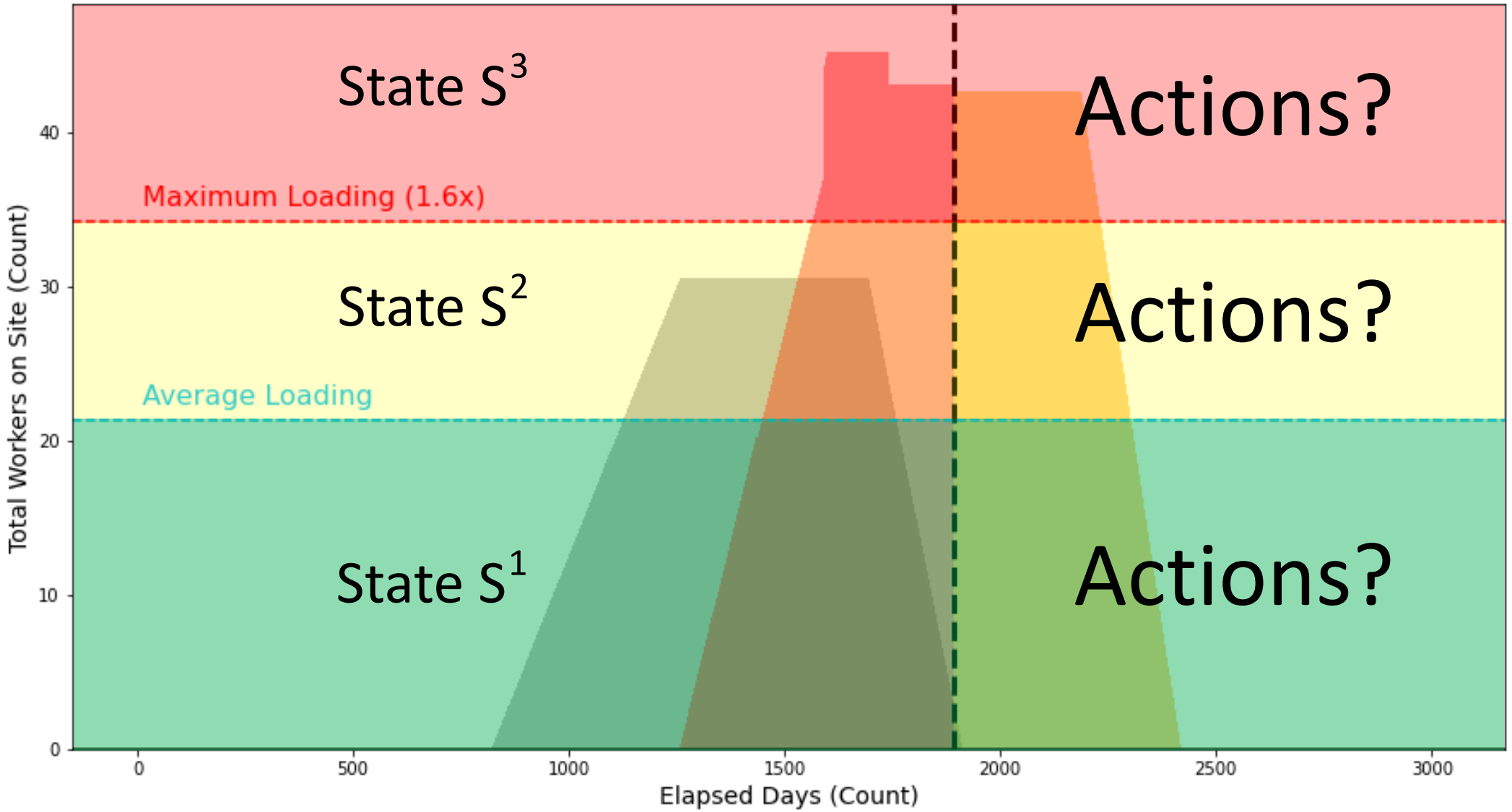
Next Step - Reinforcement Learning



Next Step - Reinforcement Learning



Next Step - Reinforcement Learning



Project Index

- **Predicting Project Success – Group #48**
  1. Michael Lee
  2. Micah Nickerson
  3. Daniel Olal

- **Reference Links:**
  - Dataset : "NYC Open Data – Capital Projects"
    - https://data.cityofnewyork.us/City-Government/Capital-Projects/n7gv-k5yt
  - Project Code (Github):
    - https://github.com/mjnickerson/csci-109b-final_project

CSCI-E-109B - Advanced Topics in Data Science - Spring 2020

Prof. Pavlos Protopapas, Mark Glickman, Chris Tanner