

Dr. Phil Barry
George Mason University
SYST530: Data-Driven Risk Estimation: The PERIL Database for Project Failures

Contents

Introduction	2
Lesson Learning Objectives	2
Data-Driven Decision Making	2
Project Failure Database: PERIL	2
Jupyter Notebooks and access to PERIL	2
Understanding the PERIL Database	3
Questions for Exploration and Analysis	4
Modify Your Risk Estimates	5
APPENDIX A How to access PERIL data set via Jupyter Notebook	6
Option 1: Static View via nbviewer	6
Option 2: Web Interactive with BINDER	7
How to Save Your Work in BINDER	8
APPENDIX B Coding Example for Filtering Data	10
APPENDIX C Helpful Links	11
APPENDIX D ID Template Used for this Lesson	12
APPENDIX E Instructor Guide for Teaching This Lesson	13
Version History	14

Introduction

In last week's assignment you identified a representative set of risk events that could affect the project and assess each risk in terms of probability (likelihood), impact, and controllability. In this lesson, you will use data-driven techniques to strengthen your risk estimates.

Lesson Learning Objectives

After participating in this lesson, you will be able to:

- Explore root causes of project failures using real-world data
- Use data visualizations as a risk assessment tool
- Modify risk estimates for autonomous project using data-driven techniques

Data-Driven Decision Making

Estimates rooted in real world data is preferable to those from models and guestimates. You can use data to test and modify already assumed probabilities of risk events and to reshape your risk mitigation plan. Rick Tison's "Data-Driven Estimating" [article](#) presents data-driven decision making in the construction industry.

Project Failure Database: PERIL

In this lesson you will be using the PERIL database¹ of project failures.

Per Tom Kendrick's [Failure-proof Projects](#) site, "The information in the PERIL database comes primarily from participants in classes and workshops on project risk management, representing a wide range of project types. Slightly more than half the projects are product development projects, with tangible deliverables. The rest are information technology, customer solution, or process improvement projects..."

You can read a Tom's article which summarize this data [here](#). Please focus on assumptions, category definitions, and limitations.

Jupyter Notebooks and access to PERIL

You will be able to explore project failure data through powerful visualizations within a Jupyter Notebook environment using the Python scripting language. You won't be asked to code but to perform simple data filtering to answer queries (see **Appendix B**). See **Appendix C** for an overview of Jupyter Notebooks and Python.

Please note: The PERIL database you will be using has been modified from the original for this lesson. Two variables have been added: **TRL**² and **COST**. These variables do not appear in the original database described in the article and on the Failure-proof Project site.

¹ Partially adapted from Kendrick, Tom (2003). PERIL Database. Unpublished dataset, <http://www.failureproofprojects.com>.

² See <http://acqnotes.com/acqnote/tasks/technology-readiness-level>

jupyter Phil_Barry_NB (unsaved changes) Logout

File Edit View Insert Cell Kernel Help Not Trusted Python 3

Learning how to understand a project management dataset and build models in Python for further analysis

Today we will be exploring a synthetically generated dataset based on the PERIL database.

To accomplish our task of understanding the data, we will need to use some Python libraries. Python libraries can be summarized as code other people have written and shared that we can use to save us time. We can import and use that code with a small import statement. Some of the packages we will use today include [Pandas](#), [Scipy](#), [Numpy](#), and [Matplotlib](#). If you want to learn more about how these different packages work with examples, click [here](#). If you would like to see the code that imports these packages - click the show code button below.

Out [1]: The raw code for this notebook is by default hidden for easier reading. To toggle on/off the raw code, click [show code](#).

Out [84]: (5000, 10)

Out [85]:

	Parameter	Category	Sub cat	Impact	TRL	Description	Region Numeric	Region	Project	Date
0	Resource	Money	Limitation	3	7	Did not have sufficient resources and the cont...	2	Eur/ME	IT/Solution	2001
1	Resource	Outsourcing	Delayed start	13	8	Contractor setup delayed by a week	0	Americas	Prod. Dev.	2006
2	Resource	Outsourcing	Late or poor output	16	5	Contractor did not spend time on the project b...	0	Americas	IT/Solution	2007
3	Resource	Outsourcing	Late or poor output	16	3	Contractors failed to show up as committed	1	Asia	IT/Solution	2014
4	Resource	Outsourcing	Late or poor output	26	4	Third-party vendor inadvertently introduced a ...	2	Eur/ME	IT/Solution	2009
5	Resource	Outsourcing	Late or poor output	19	3	Outsourced staff lacked the necessary skillset	3	Africa	IT/Solution	2014
6	Resource	People	Late start	4	1	Planning delayed due to staff being still tied...	1	Asia	Prod. Dev.	2015
7	Resource	People	Loss	9	3	Chef quit two days before the café was schedul...	3	Africa	IT/Solution	2017
8	Resource	People	Motivation	15	3	Work at customer site had to be done by union ...	3	Africa	IT/Solution	2003
9	Resource	People	Queuing	21	5	Critical task assigned to a heavily booked expert	1	Asia	IT/Solution	2010

What are our column names?

Parameter
Category

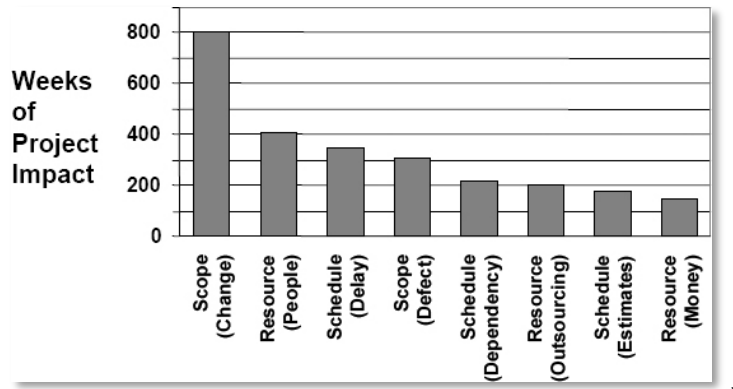
Understanding the PERIL Database

The key dependent variable in this model is **IMPACT**. The columns in the table are:

- Parameter (risk: Scope, Schedule, Resource)
- Category
- Sub cat
- TRL: The Technology Readiness Description
- Region Numeric
- Region
- Project
- Date
- Cost (correlated to IMPACT)

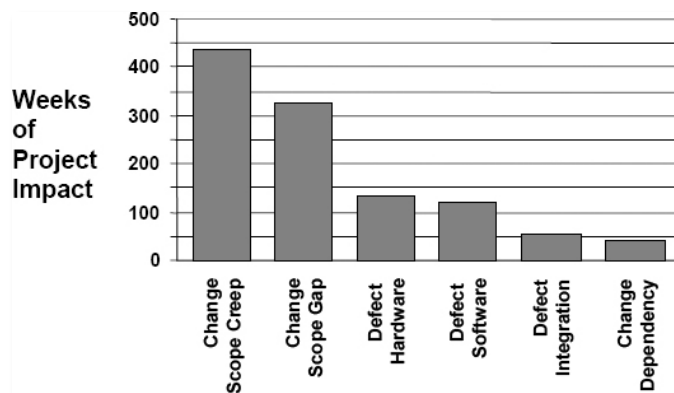
The PERIL database categorizes **risk** in terms of **three project parameters**: Scope, Schedule, and Resource. **Impact** is measured as weeks of project delay.

Categories identify the root cause of the risk. For example, in this chart categories of Change, People Delay, Defect, Dependency, Outsourcing, Estimates, and Money are root causes:



3

Each **Subcategory** (under each category) further explains the root cause. For example, considering Scope risk: SCOPE => Change => Creep



Questions for Exploration and Analysis

Explore the PERIL database to determine which categories under each of the three project parameters best describe your project.

- The box & whisker plot displays the distribution of **cost**. Which risk subcategories contribute to median impact? Which contribute to the outliers?
- Explore impact by **Region**. Would switching the project to another region change impact?
- Explore impact by **TRL**.⁴ How does TRL affect impact? What can you say about TRL as a predictor of impact?
- How does the cost vary from year to year? Does it change with Region? What subcategories contribute to the low/high costs in years 2012 and 2015 for the Americas?

³ Pareto charts from: Kendrick, C. (2003). Overcoming project risk: lessons from the PERIL database. Paper presented at PMI® Global Congress 2003—North America, Baltimore, MD. Newtown Square, PA: Project Management Institute. Retrieved from <https://www.pmi.org/learning/library/overcoming-project-risk-lessons-peril-database-7713> on 4/22/2019.

⁴ See <http://acqnotes.com/acqnote/tasks/technology-readiness-level>

Create a set of focused research questions to test various hypotheses related to your project. You want to extract probabilities of a risk event given a combination of root causes in PERIL. Email your instructor your research questions and/or post to the class blog.

Modify Your Risk Estimates

Risks can be accepted, controlled, or transferred. Your project plan should certainly include risk analysis and recommended controls, but the risks properly belong to the project stakeholders. As the project manager, you should focus on documenting the risks and providing reasonable project controls for those risks that aren't accepted or transferred.

Based on your exploration of the PERIL database,

- How might you modify your previous risk likelihood and impact assessment?
- Do your risk events change?
- Do the features you've chosen change your determination of dominance?

APPENDIX A How to access PERIL data set via Jupyter Notebook

You can access the Jupyter Notebook and PERIL in two ways:

1. Static – view Notebook in browser via **nbviewer** site
2. Web interactive – view and interact with Notebook in browser via **Binder** site – nothing to download

Technical Support: Please contact Ali Zaidi (szaidi@mitre.org) or Joe Garner (garnerj@mitre.org). Please write **SYST530** in the subject line.

Option 1: Static View via nbviewer

You will receive a link to the static version of the Notebook. You can view tables, graphs, and other visualizations but cannot change or interact with the Notebook.

Link to main site: <https://nbviewer.jupyter.org/>


Link for SYS530: (active for 3 weeks from April 26, 2019):

<https://nbviewer.jupyter.org/github/szaidimitre/SYST530-Peril-Notebook/blob/master/SYST530%20Notebook.ipynb>



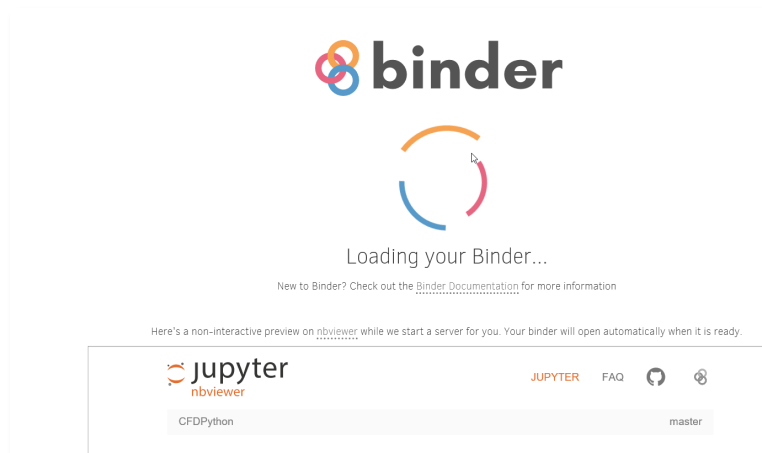
Option 2: Web Interactive with BINDER

You will receive a link. Each link will be unique to each project group. For better performance, please use

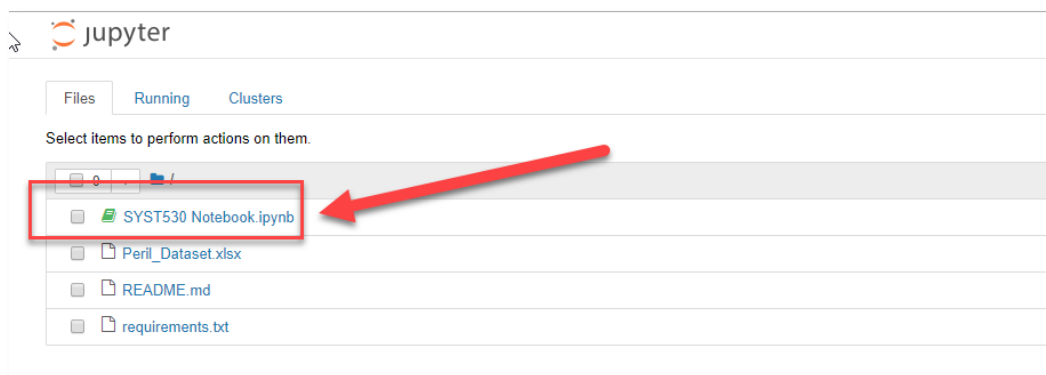
Chrome  . Expect up to 3 minutes to load!

Link for SYS530: (active for 3 weeks from April 26, 2019):

<https://mybinder.org/v2/gh/szaidimitre/SYST530-Peril-Notebook/master>



The file structure is as below. Click on the Notebook.



jupyter Phil_Barry_NB (unsaved changes)

File Edit View Insert Cell Kernel Help Not Trusted Python 3

Learning how to understand a project management dataset and build models in Python for further analysis

Today we will be exploring a synthetically generated dataset based on the PERIL database.

To accomplish our task of understanding the data, we will need to use some Python libraries. Python libraries can be summarized as code other people have written and shared that we can use to save us time. We can import and use that code with a small import statement. Some of the packages we will use today include [Pandas](#), [Scipy](#), [Numpy](#), and [Matplotlib](#). If you want to learn more about how these different packages work with examples, click [here](#). If you would like to see the code that imports these packages - click the show code button below.

Out [1]: The raw code for this notebook is by default hidden for easier reading. To toggle on/off the raw code, click [show code](#).

Out [84]: (5000, 10)

Out [85]:

	Parameter	Category	Sub cat	Impact	TRL	Description	Region Numeric	Region	Project	Date
0	Resource	Money	Limitation	3	7	Did not have sufficient resources and the cont...	2	EurME	IT/Solution	2001
1	Resource	Outsourcing	Delayed start	13	8	Contractor setup delayed by a week	0	Americas	Prod. Dev.	2006
2	Resource	Outsourcing	Late or poor output	16	5	Contractor did not spend time on the project b...	0	Americas	IT/Solution	2007
3	Resource	Outsourcing	Late or poor output	16	3	Contractors failed to show up as committed	1	Asia	IT/Solution	2014
4	Resource	Outsourcing	Late or poor output	26	4	Third-party vendor inadvertently introduced a ...	2	EurME	IT/Solution	2009
5	Resource	Outsourcing	Late or poor output	19	3	Outsourced staff lacked the necessary skillset	3	Africa	IT/Solution	2014
6	Resource	People	Late start	4	1	Planning delayed due to staff being still tied...	1	Asia	Prod. Dev.	2015
7	Resource	People	Loss	9	3	Chef quit two days before the café was schedul...	3	Africa	IT/Solution	2017
8	Resource	People	Motivation	15	3	Work at customer site had to be done by union ...	3	Africa	IT/Solution	2003
9	Resource	People	Queueing	21	5	Critical task assigned to a heavily booked expert	1	Asia	IT/Solution	2010

What are our column names?

Parameter
Category

IMPORTANT: When you first open the Notebook, run all the cells!

jupyter 00_Quick_Python_Intro (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help

Run Cells
Run Cells and Select Below
Run Cells and Insert Below
Run All
Run All Above
Run All Below
Cell Type
Current Outputs
All Output

Python C

Hello! This is a q

There are two wa

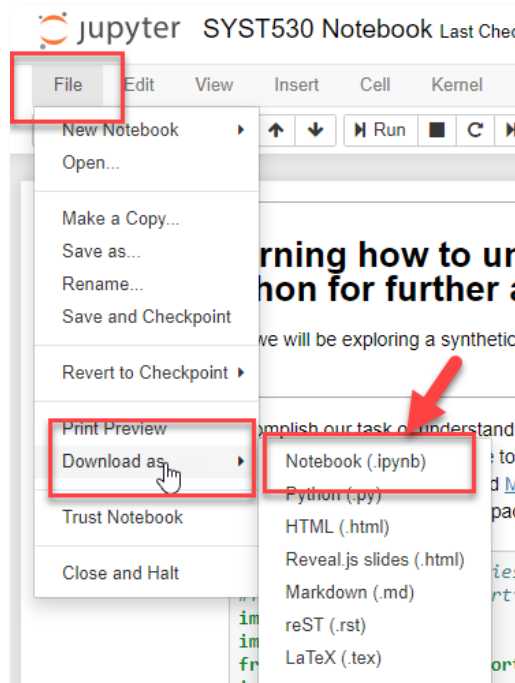
1. You can dow
2. You can run Python in the cloud using [vstack](#) web-based data analy

After Cell => Run All, you will now have run all the cells in the Notebook! You can proceed with the lesson.

How to Save Your Work in BINDER

- **WARNING:** Your Notebook will automatically shut down in BINDER after 10 minutes of inactivity so make sure you keep clicking in the cells or "RUN ALL"
- To save any changes you have made to your Notebook

- Download the Notebook file to your PC or MAC
- RENAME it
- Upload the renamed Notebook to BINDER



APPENDIX B Coding Example for Filtering Data

Here is an example of how to perform simple filtering using Python in the Notebook. You can use the Python Quick Reference Guide⁵ to read the code and/or use the links in **Appendix C**.

You would like to display data for the **Americas**. Thus, Region = 'Americas'. You may also wish to restrict your analysis to the first **50 projects** in the database. The first column displayed is just the row count in the results (e.g., 0 to 49 for the first 50 projects we selected).

```
In [20]: region_filter = dataset.loc[dataset['Region'] == 'Americas']
region_filter = region_filter.sort_values(['Cost'], ascending = False)
region_filter = region_filter.reset_index(drop = True)
region_filter.head(50) #Let's look at the first 50 projects in the Americas
```

Sort by descending COST

row count	Project Number	Parameter	Category	Sub cat	Impact	TRL	Description	Region Numeric	Region	Project	Year	Month	Cost
0	4813	Schedule	Delay	Hardware	26	7	No problems in test, but production problems r...	0	Americas	Prod. Dev.	2010	July	2556034
1	3722	Resource	People	Loss	26	4	Customer did not describe in detail what exact...	0	Americas	Prod. Dev.	2005	May	2494102
2	3895	Scope	Change	Gap	25	7	All students did not complete training as sche...	0	Americas	IT/Solution	2017	October	2493700
3	4798	Scope	Defect	Hardware	26	8	Problem detected in production due to transact...	0	Americas	Prod. Dev.	2015	May	2491710

Now you'd like to filter down to which projects in the **Americas** failed due to **Money** as the root cause.

```
In [21]: category_filter = dataset.loc[(dataset['Category'] == 'Money') & (dataset['Region'] == 'Americas')]
category_filter = category_filter.sort_values(['Cost'], ascending = False)
category_filter = category_filter.reset_index(drop = True)
category_filter.head(50) #Let's look at the first 50 projects where the region is the Americas and the category is Mo
```

Project Number	Parameter	Category	Sub cat	Impact	TRL	Description	Region Numeric	Region	Project	Year	Month	Cost	
0	4120	Resource	Money	Limitation	23	6	"Customer supplied" hardware does not work. R...	0	Americas	Prod. Dev.	2016	January	2187760
1	4387	Resource	Money	Limitation	24	9	fittings and assemblies were not up to project...	0	Americas	Prod. Dev.	2008	July	1895856
2	4361	Resource	Money	Limitation	18	4	Details of requirements were submitted incorre...	0	Americas	Prod. Dev.	2002	June	1762812

You can filter on many other variables (categories, subcategories)!

⁵ Quick Reference Guide: <https://www.python.org/ftp/python/doc/quick-ref.1.3.html>

APPENDIX C Helpful Links

Here are resources to learn about Jupyter Notebook, Python, and Machine Learning.

Quick introduction to Jupyter Notebook	https://www.youtube.com/watch?v=jZ952vChhul
Python: YouTube video series	https://www.youtube.com/watch?v=k9TUPpGqYTo&list=PL-osiE80TeTskrapNbzxhwoFUiLCjGgY7&index=2
Python: Free Online Book	https://automatetheboringstuff.com/
Python: Online EdX Course	https://www.edx.org/course/introduction-to-python-absolute-beginner-3
Intro to ML (first learn to code in Python then attempt this):	https://developers.google.com/machine-learning/crash-course/

APPENDIX D ID Template Used for this Lesson

Learning Objectives	Discussion(s)	Assignment(s)	Learning Resources + Media
<ul style="list-style-type: none"> Explore root causes of project failures using real-world data Use data visualizations as a risk assessment tool Modify risk estimates for autonomous project using data-driven techniques 	<p>There are many approaches to establishing probabilities of risk events without real world data. Some project teams take different approaches to defining risk events: delayed delivery, vehicle catches fire, System problems, Project-related: staff or PM availability or attrition, poor cost estimates</p> <ol style="list-style-type: none"> From last week's assignment, what was your approach to defining root causes of risk events & probabilities for your project? What's the difference between project risk vs. product or service risk? What is data-driven risk analysis? Before this assignment, what was your plan to justify your estimates? (e.g., search the web, cite a research paper) How can PERIL help inform your risk estimates and mitigation plan? 	<p>Estimates rooted in real world data is preferable to those from models and guestimates. You can use data to test and modify already assumed probabilities of risk events and to reshape your risk mitigation plan.</p> <p>To prepare for this assignment:</p> <ul style="list-style-type: none"> Read Rick Tison's article about data-driven decision making Present your risk events, estimates of likelihood of risk events, & risk mitigation plan from last week's assignment. Read Tom Kendrick's article describing PERIL database. Focus on assumptions, category definitions, and limitations Explore trends for root causes using PERIL database <p>Assignment: Create a set of focused research questions to test various hypotheses related to your project. You want to extract probabilities of a risk events given a combination of root causes in PERIL. Email your instructor research questions and/or post to the class blog</p> <ul style="list-style-type: none"> The box & whisker plot displays the distribution of cost. Which risk subcategories contribute to median impact? Which contribute to the outliers? Explore impact by Region. Would switching the project to another region change impact? Explore impact by TRL. How does TRL affect impact? What can you say about TRL as a predictor of impact? How does the cost vary from year to year? Does it change with Region? What subcategories contribute to the low/high costs in years 2012 and 2015 for the Americas? <p>Risks can be accepted, controlled, or transferred. Your project plan should certainly include risk analysis and recommended controls, but the risks properly belong to the project stakeholders. As the project manager, you should focus on documenting the risks and providing reasonable project controls for those risks that aren't accepted or transferred.</p> <p>Based on your exploration of the PERIL database,</p> <ul style="list-style-type: none"> How might you modify your previous risk likelihood and impact assessment? Do your risk events change? Do the features you've chosen change your determination of dominance? 	<p>Rick Tison's "Data-Driven Estimating" https://www.fminet.com/fmi-quarterly/article/2015/06/data-driven-estimating/</p> <p>Tom Kendrick's "Overcoming Project Risk" https://www.pmi.org/learning/library/overcoming-project-risk-lessons-peril-database-7713</p> <p>Jupyter Notebook: PERIL database with classifiers + canned filters</p> <p>Video & QRG: Filtering with Python; Compare to using other means to search/query data e.g. SQL COUNT, SUM, AVG</p> <p>PPT Lecture: Illustrate data science workflow: data cleaning and transformation, descriptive statistics, predictive analytics</p> <p>Live Demo: Extracting probabilities: How many product development projects fail due to staff attrition?</p> <ul style="list-style-type: none"> Filter on "Prod Dev" + "Americas" => 1279 records found Then filter on subcategory "Loss" => 78 records found Extract probability based on simple ratio: $78/1279 = 6\%$ for staff loss

APPENDIX E Instructor Guide for Teaching This Lesson

This lesson is taught over two weeks. Each week has a live, online session through WEBEX. Here is a suggested structuring of the **Week 1 session** (1.5 hrs) using **ROPES**: Review, Overview, Presentation, Exercise, Summary.

Week 1 Session (time in minutes)	Learning Objective	Instructional Method	Activity + Media + Learning Resources
10	Review	Q&A & Presentation	Students present risk estimates for their projects
10	(1) Explore root causes of project failures using real-world data (2) Use data visualizations as a risk assessment tool (3) Modify risk estimates	(R) Question	Prompt for students to explain project vs. product risk calculations
10		(O) Discussion	Whole class discussion about defining risk estimates
10		(P) Whole class demonstration	Whole class demonstration of dataset for project failures; relate to project assignment
35		(E) Dataset exploration	Explore dataset with Jupyter Notebook ; work with classifiers; create and test focused research questions; review job aids for Notebook
15		(S) Discussion	Clarify understanding of hypothesis testing; refine research questions

Version History

Version	Editor	Changes	Date
1.0	JGarner	Created official version 1.0 Added Appendix B & C Added Introduction, notes about TRL, Cost	May 1, 2019
2.0	JGarner	Added Instructor Guide	May 7, 2019
3.0	JGarner	Added ID Template	June 6, 2019