

Pitfalls: Project Problems

Not Asking a Sharp Question (or Solving the Wrong Problem)

Make sure your answer is in the form of a number AND a range of uncertainty

Not Adapting the Question Once it's Already Failed

It's an iterative process. Go back, Jack and do it again.

Data is Owned, Not Governed

Sources for income information:

Social Security Data

IRS Data

Census Data

State Administrative Data

Current Population Survey

Data Does Not Contain the Information Needed

The Drunkard's search:

"I lost a \$2 bill down on Atlantic avenue," said the man.

"What's that?" asked the puzzled officer. "You lost a \$2 bill on Atlantic avenue? Then why are you hunting around here in Copley square?"

"Because," said the man as he turned away and continued his hunt on his hands and knees, "the light's better up here."

[-https://quoteinvestigator.com/2013/04/11/better-light/](https://quoteinvestigator.com/2013/04/11/better-light/)

https://en.wikipedia.org/wiki/Streetlight_effect

Not Relying on Inexpensive and Open Source Technologies



Timeline is Overly Optimistic

We never *really* know how models perform until the event occurs

Analysis is always the easiest part of any project.

80% of time in a data project is spent on cleaning data, while the other 20% is spent on complaining about cleaning data.

Inflated Expectations of Value

Sometimes insights aren't unexpected

Sometime the data just don't have anything to say

Sometimes the gains are real, but small

Scale, scale, scale

Expecting to Predict the Unpredictable



No amount of observations of white swans can allow the inference that all swans are white, but the observation of a single black swan is sufficient to refute that conclusion.

Overkill

$$y = \beta X + \varepsilon$$

Statistics

2009

$$y = \beta X + \varepsilon$$

**MACHINE
LEARNING**

2019