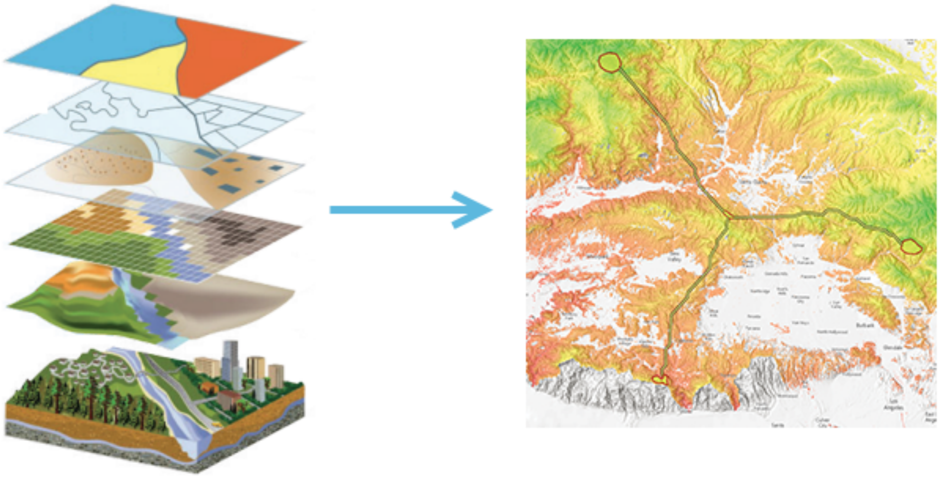
# TEST-DRIVEN DATA ANALYSIS

*Do you believe your analytical results?*



OLLIVIER & CO

Thanks to:
Nicholas Radcliffe
http://tdda.info
njr@StochasticSolutions.com
Dept of Mathematics, University of Edinburgh

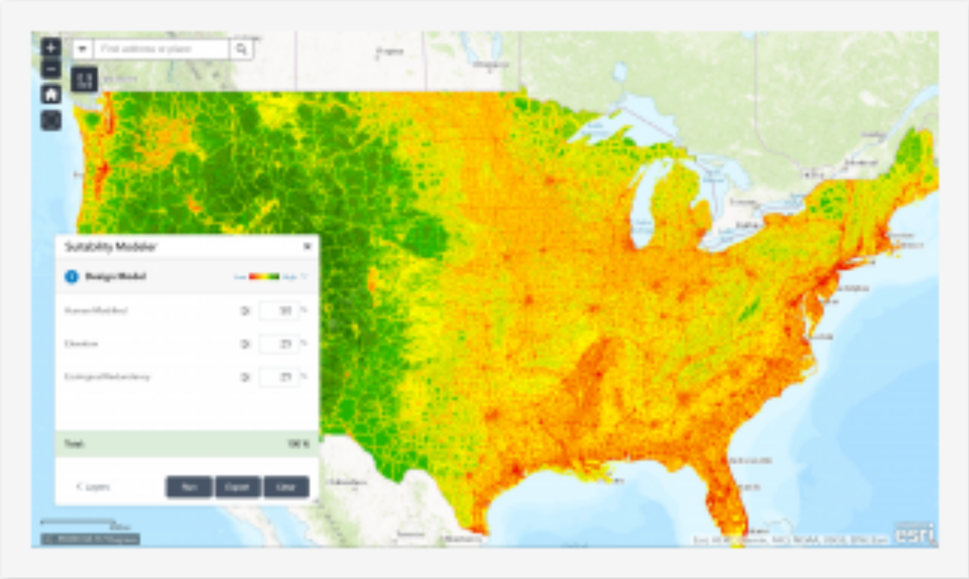
# GeoPlanner's Suitability Modeler is now part of Web AppBuilder

by Rob Stauder on June 29, 2017

Share 7 Tweet 88 Share

Imagine if, in a few clicks, you could answer multiple-factor spatial questions like *Where are the areas on low angled slopes, in shrubby vegetation and are far from roads*? What if you could do that and emphasize the importance of one of those factors over another?

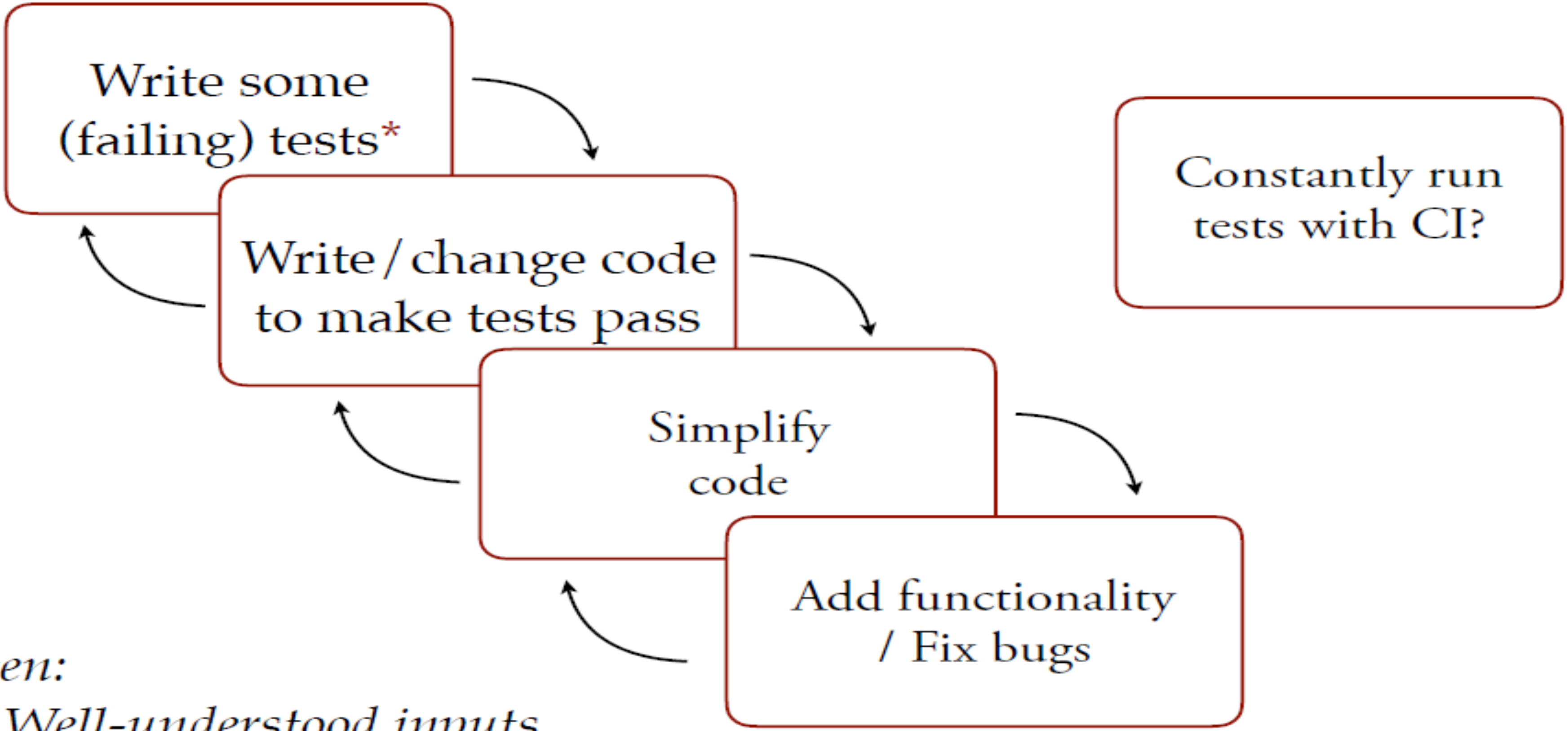
You would be the hero of your workplace!



# The Big Idea

*Transfer the ideas of*
*test-driven development*
*from software*
*development*
*to data analysis*

# SOFTWARE DEVELOPMENT (WITH TDD)



Write some (failing) tests*

Write / change code to make tests pass

Simplify code

Add functionality / Fix bugs

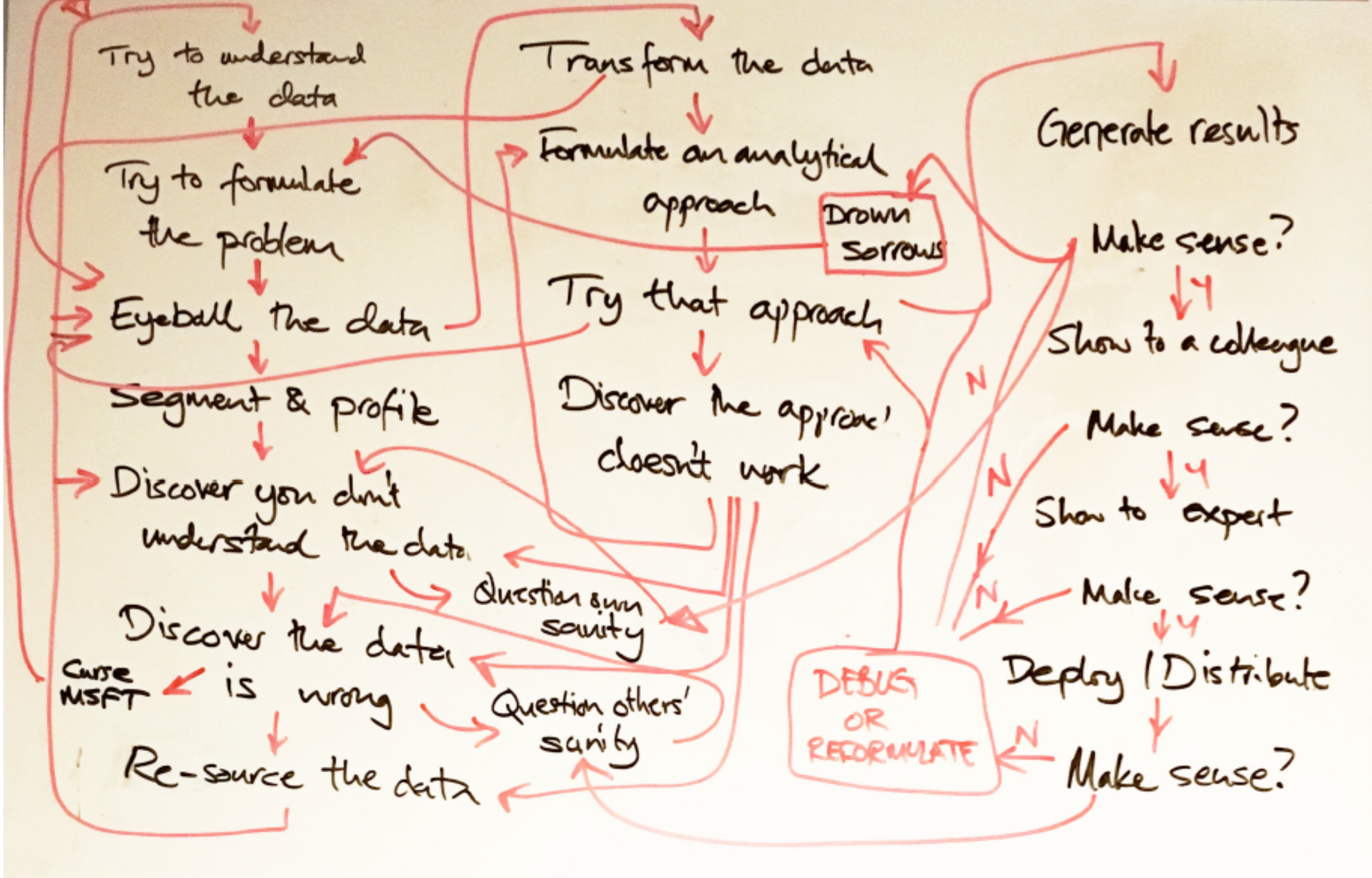
Constantly run tests with CI?

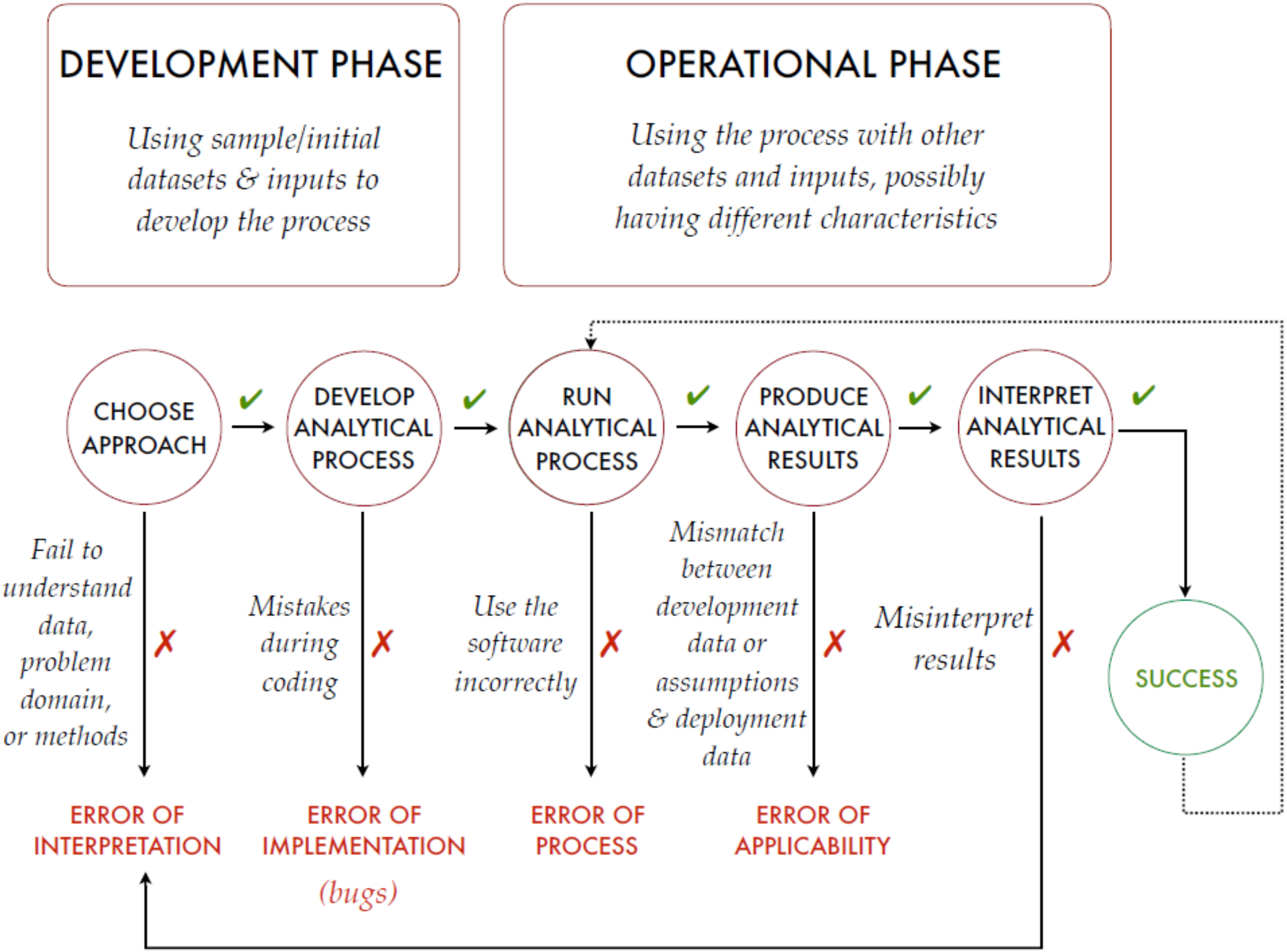
*Often:*

- *Well-understood inputs*
- *Well-understood goal*
- *Many kinds of errors/failures are unmistakable*

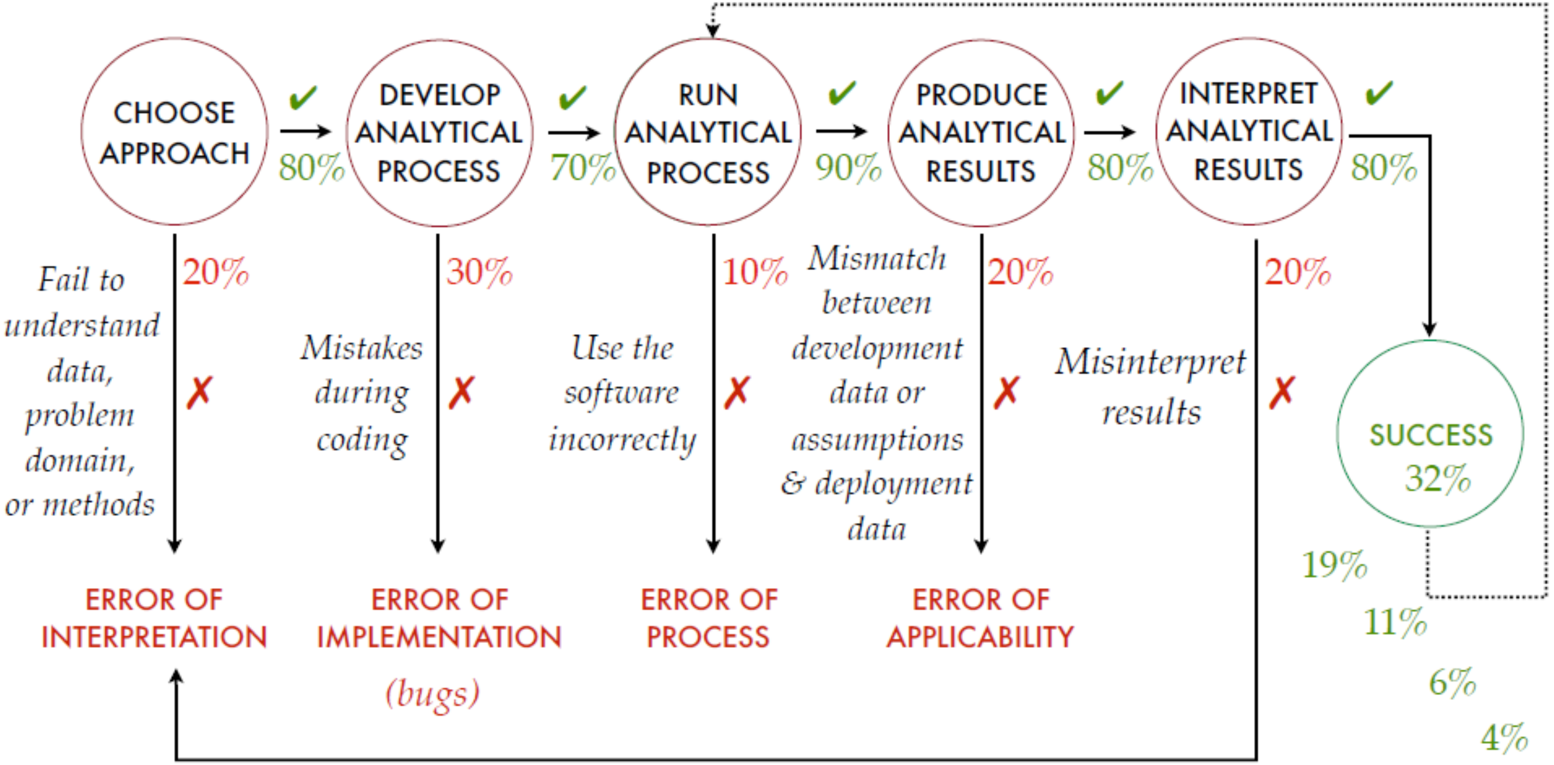
* While mocking almost everything

# How models are really developed

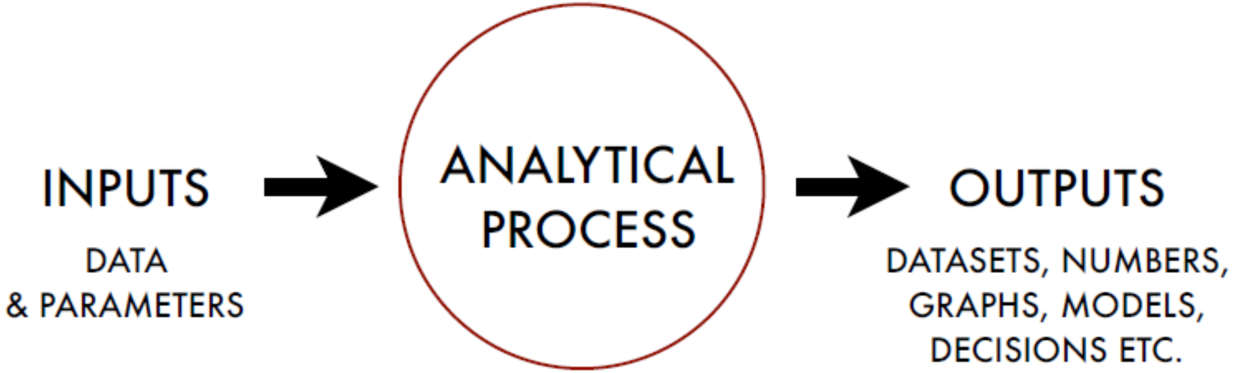



DEVELOPMENT PHASE
Using sample/initial datasets & inputs to develop the process
OPERATIONAL PHASE
Using the process with other datasets and inputs, possibly having different characteristics
CHOOSE APPROACH
✓
DEVELOP ANALYTICAL PROCESS
✓
RUN ANALYTICAL PROCESS
✓
PRODUCE ANALYTICAL RESULTS
✓
INTERPRET ANALYTICAL RESULTS
✓
Fail to understand data, problem domain, or methods
✗
Mistakes during coding
✗
Use the software incorrectly
✗
Mismatch between development data or assumptions & deployment data
✗
Misinterpret results
✗
SUCCESS
ERROR OF INTERPRETATION
ERROR OF IMPLEMENTATION
(bugs)
ERROR OF PROCESS
ERROR OF APPLICABILITY

*If you buy into this model, it's sobering to attach probability estimates to each transition and calculate the probability of success after a few runs . . .*



# TDDA: LEVEL ZERO



INPUTS

DATA
& PARAMETERS

ANALYTICAL PROCESS

OUTPUTS

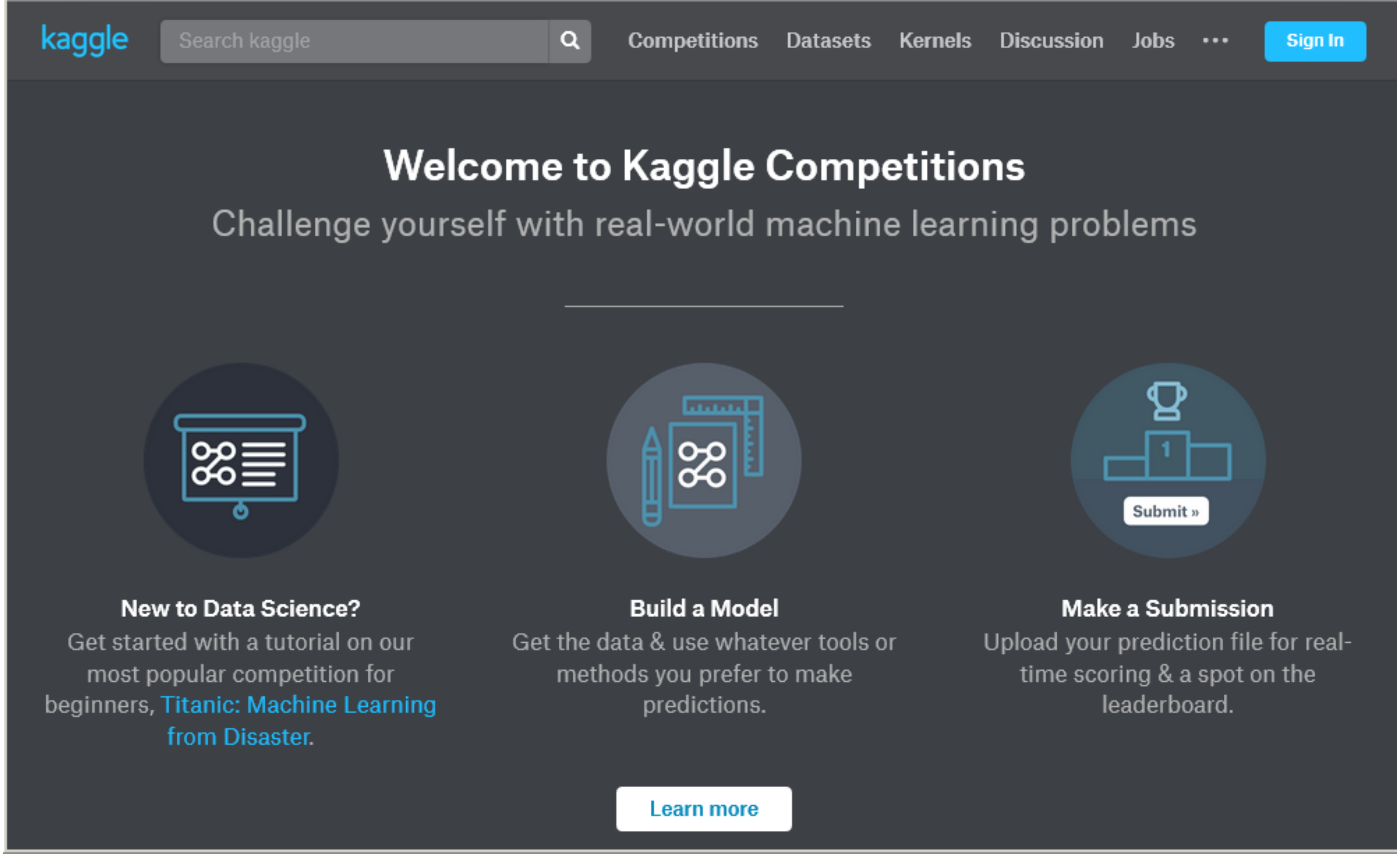
DATASETS, NUMBERS,
GRAPHS, MODELS,
DECISIONS ETC.

*Record inputs*

*Capture as scripted, parameterised executable procedure ("reproducible research")*

*Record ("reference") outputs*

*Develop a verification procedure (*`diff`*) and periodically rerun:*
*do the same inputs (still) produce the same outputs?*



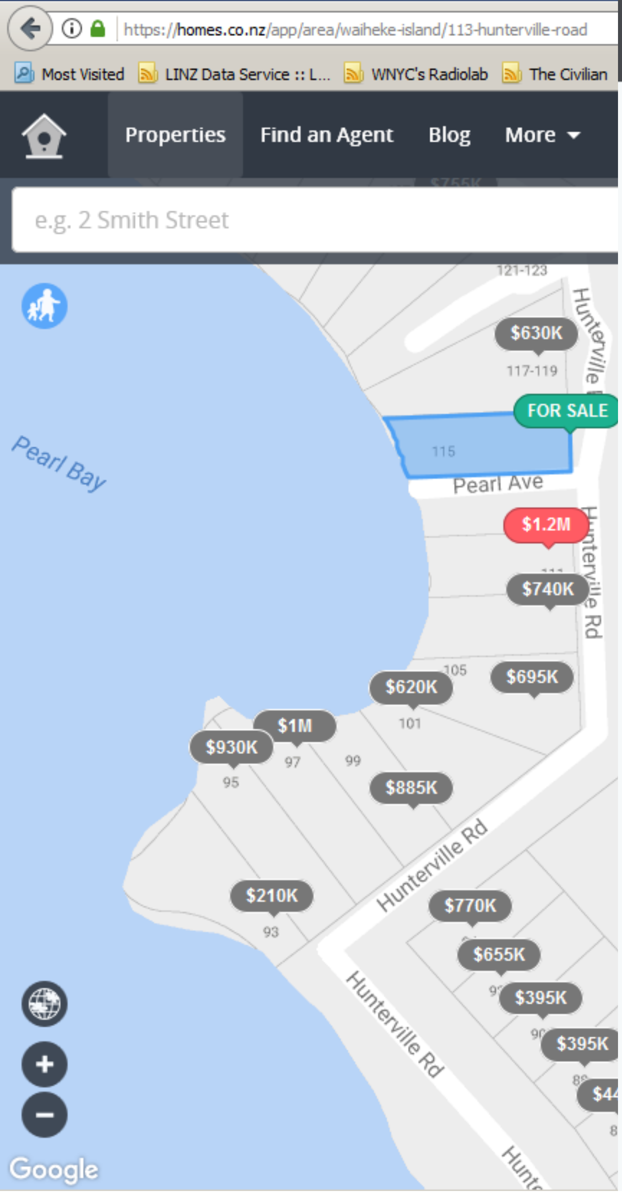
# Welcome to Kaggle Competitions

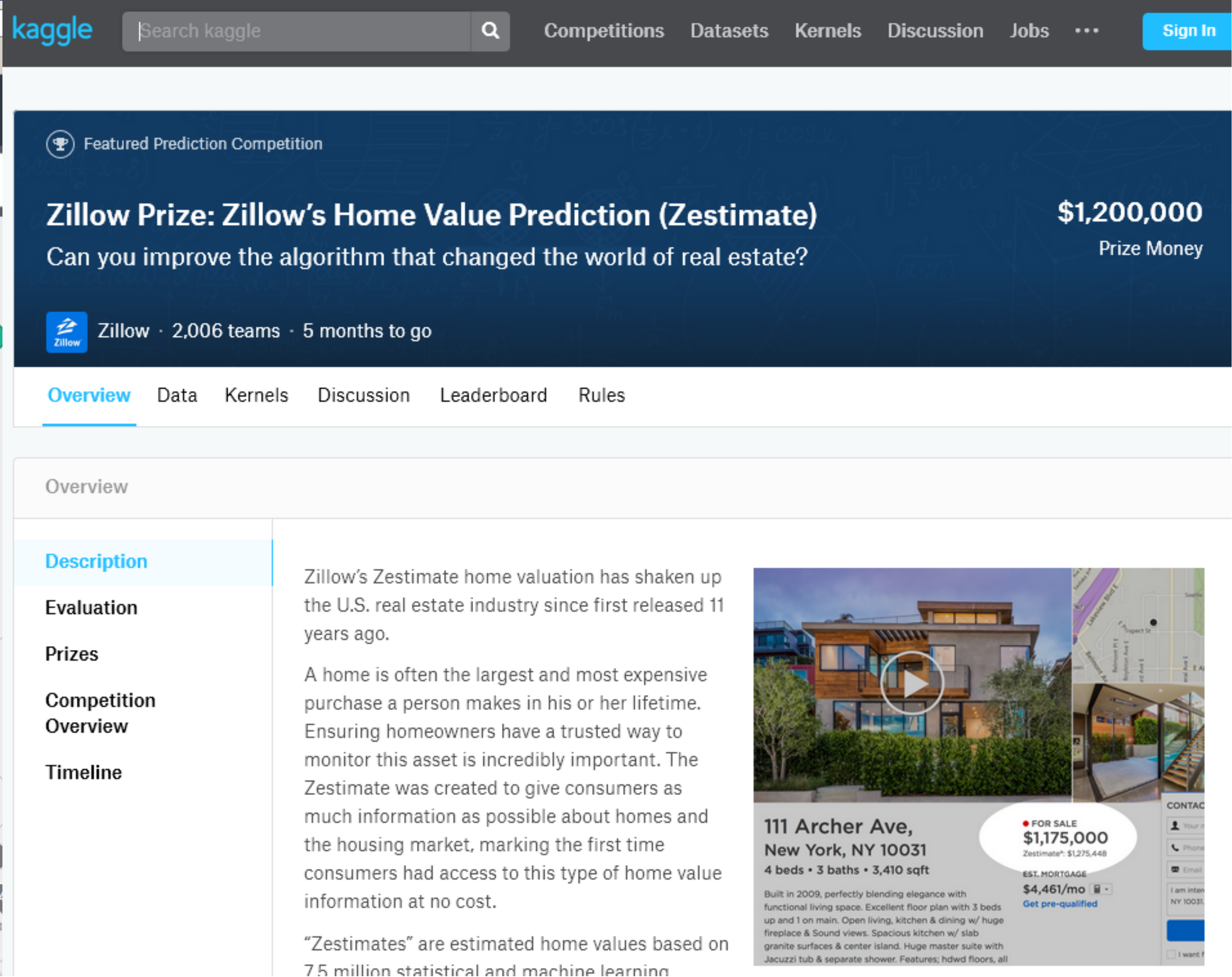
Challenge yourself with real-world machine learning problems

**New to Data Science?**
Get started with a tutorial on our most popular competition for beginners, Titanic: Machine Learning from Disaster.

**Build a Model**
Get the data & use whatever tools or methods you prefer to make predictions.

**Make a Submission**
Upload your prediction file for real-time scoring & a spot on the leaderboard.





## Zillow Prize: Zillow's Home Value Prediction (Zestimate)

Can you improve the algorithm that changed the world of real estate?

$1,200,000
Prize Money

Zillow · 2,006 teams · 5 months to go

Zillow's Zestimate home valuation has shaken up the U.S. real estate industry since first released 11 years ago.

A home is often the largest and most expensive purchase a person makes in his or her lifetime. Ensuring homeowners have a trusted way to monitor this asset is incredibly important. The Zestimate was created to give consumers as much information as possible about homes and the housing market, marking the first time consumers had access to this type of home value information at no cost.

"Zestimates" are estimated home values based on 7.5 million statistical and machine learning

111 Archer Ave,
New York, NY 10031
4 beds • 3 baths • 3,410 sqft

Built in 2009, perfectly blending elegance with functional living space. Excellent floor plan with 3 beds up and 1 on main. Open living, kitchen & dining w/ huge fireplace & Sound views. Spacious kitchen w/ slab granite surfaces & center island. Huge master suite with Jacuzzi tub & separate shower. Features; hdwd floors, all

● FOR SALE
$1,175,000
Zestimate®: $1,275,448
EST. MORTGAGE
$4,461/mo
Get pre-qualified

# tdda level1: CONSTRAINTS

*Look before you leap!*

Checking the data conforms to your assumptions before you start.

- Not just the obvious input, but also intermediate and output sets

This is tedious to generalise so there are tools to help...

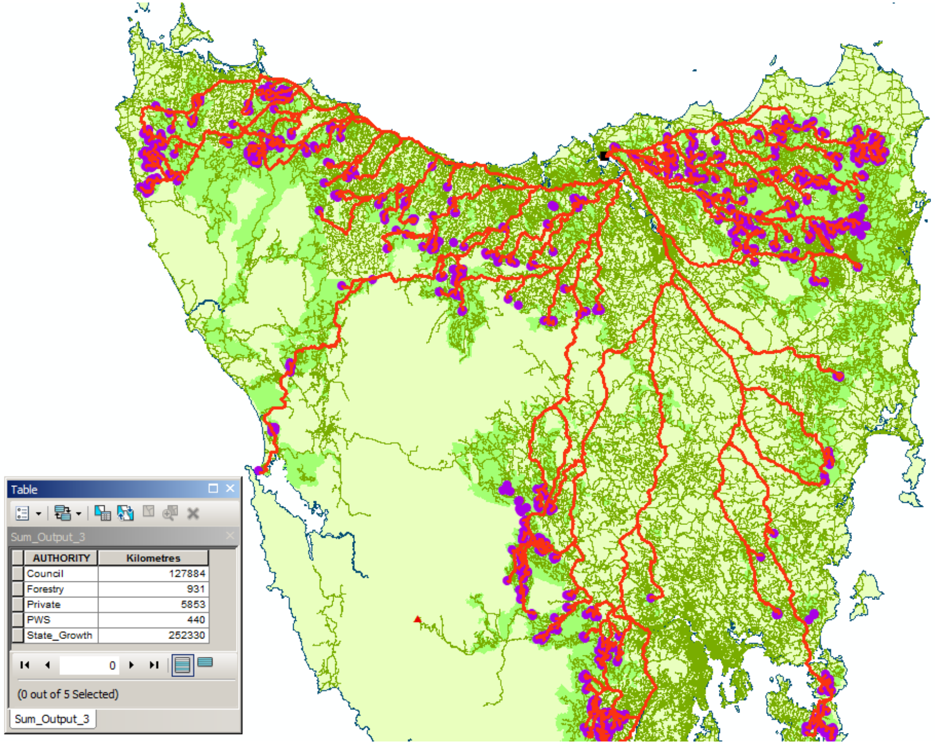
# Tools to automate L1 constraint tests

| Tool | Read input | Rules | Action | Flexibility, Complexity |
|---|---|---|---|---|
| ArcGIS | Input Dialog | Existence, schema | Block from starting | Low |
| RDBMS | Table schema | Field constraints, triggers | Rejection | Med |
| TDDA python module | Pandas framework | Regular expression generator | Report | Med |
| FME | Attribute Validator transformer | Choose from built-in rules, custom tests | Report, repair or filter | High |

# EXAMPLE CONSTRAINTS

| SINGLE FIELD CONSTRAINTS | DATASET CONSTRAINTS |
|---|---|
| $\text{Age} \leq 150$ | The dataset must contain field CID |
| type(Age) = int | Number of records must be 118 |
| $\text{CID} \neq \text{NULL}$ | One field should be tagged O |
| CID unique | Date should be sorted ascending |
| len(CardNumber) = 16 | **MULTI-FIELD CONSTRAINTS** |
| Base in {“C”, “G”, “A”, T”} | $\text{StartDate} \leq \text{EndDate}$ |
| Vote ≠ “Trump” | AlmostEqual(F, m * a, 6) |
| StartDate < tomorrow() | sum(Favourite*) = 1 |
| $v < 2.97\text{e}10$ | $\text{minVal} \leq \text{medianVal} \leq \text{maxVal}$ |
| $\text{Height} \sim N(1.8, 0.2)$ | $V \leq H * w * d$ |

# TEST-DRIVEN DATA ANALYSIS



Table

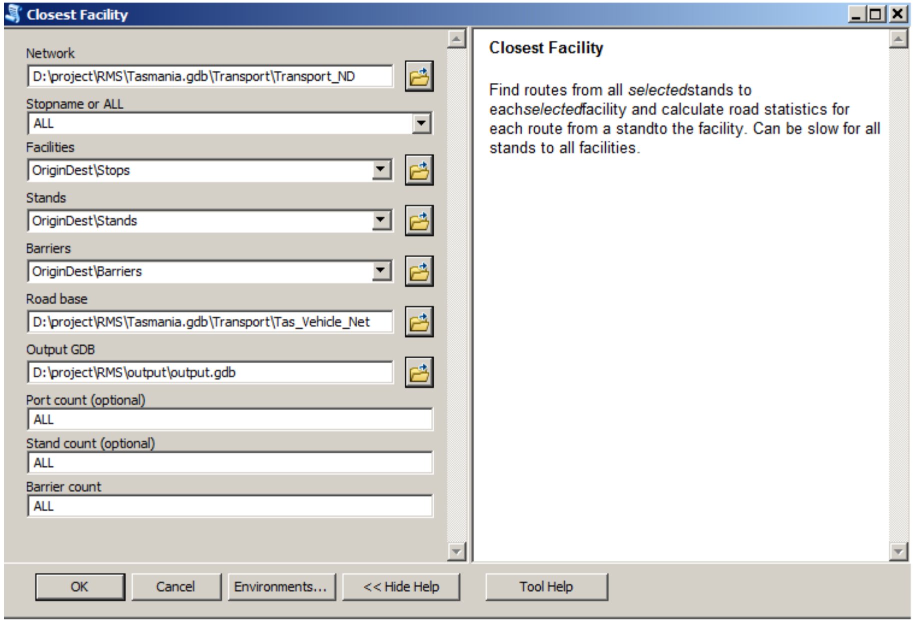
Sum_Output_3

| AUTHORITY | Kilometres |
|---|---|
| Council | 127884 |
| Forestry | 931 |
| Private | 5853 |
| PWS | 440 |
| State_Growth | 252330 |

0

(0 out of 5 Selected)

Sum_Output_3



Closest Facility

Network
D:\project\RMS\Tasmania.gdb\Transport\Transport_ND

Stopname or ALL
ALL

Facilities
OriginDest\Stops

Stands
OriginDest\Stands

Barriers
OriginDest\Barriers

Road base
D:\project\RMS\Tasmania.gdb\Transport\Tas_Vehicle_Net

Output GDB
D:\project\RMS\output\output.gdb

Port count (optional)
ALL

Stand count (optional)
ALL

Barrier count
ALL

**Closest Facility**

Find routes from all *selected*stands to each*selected*facility and calculate road statistics for each route from a standto the facility. Can be slow for all stands to all facilities.

OLLIVIER & CO