

# Collecting clean and usable data for better reliability modelling

*Data Scientist & founder*  
*carles@reliabledynamics.com*



*carlescg*



*@carles\_*



# Agenda

- Who we are?
- Modeling 101 - Quick introduction
- So, what can we do?
- Six steps to improve any modeling
- Demo/Case
- Recap

# Who we are?

We believe that **reliability engineering** practitioners should focus on delivering value and developing critical thinking in O&M environments. **Software should not be a barrier** to apply this techniques in any organization. We take care to eliminate that barrier.

We believe that the competitive advantage is to have a **wide range of apps**, that increase the reliability inside any organization, regardless of size and budget.

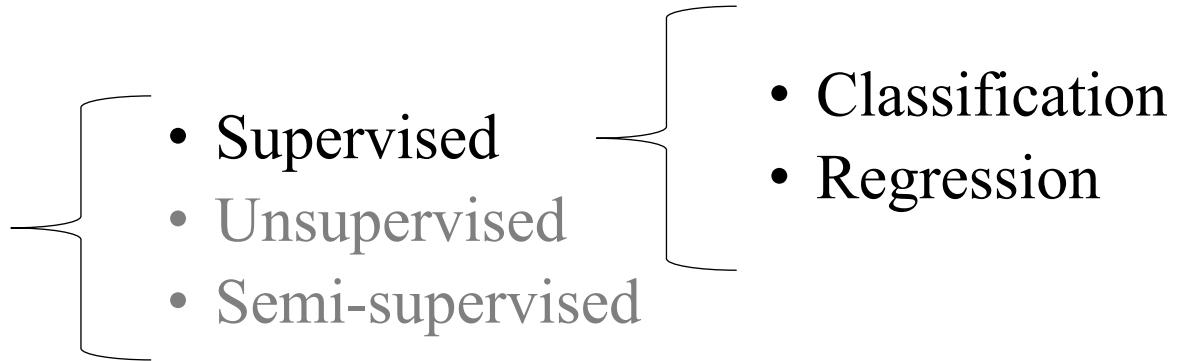
We develop **basic applications** with **modern technologies**. Through standard applications or custom solutions. Our products include apps from statistical calculations, reporting, to warranty predictions and predictive maintenance.



**RELIABLE DYNAMICS**

- Reliability modeling
  - Life data analysis (statistical)
  - Vibration (deterministic)
  - Oil sampling
- Algorithms
  - Machine learning
  - Deep learning
  - Reinforced learning

**Output = function ( Input )**



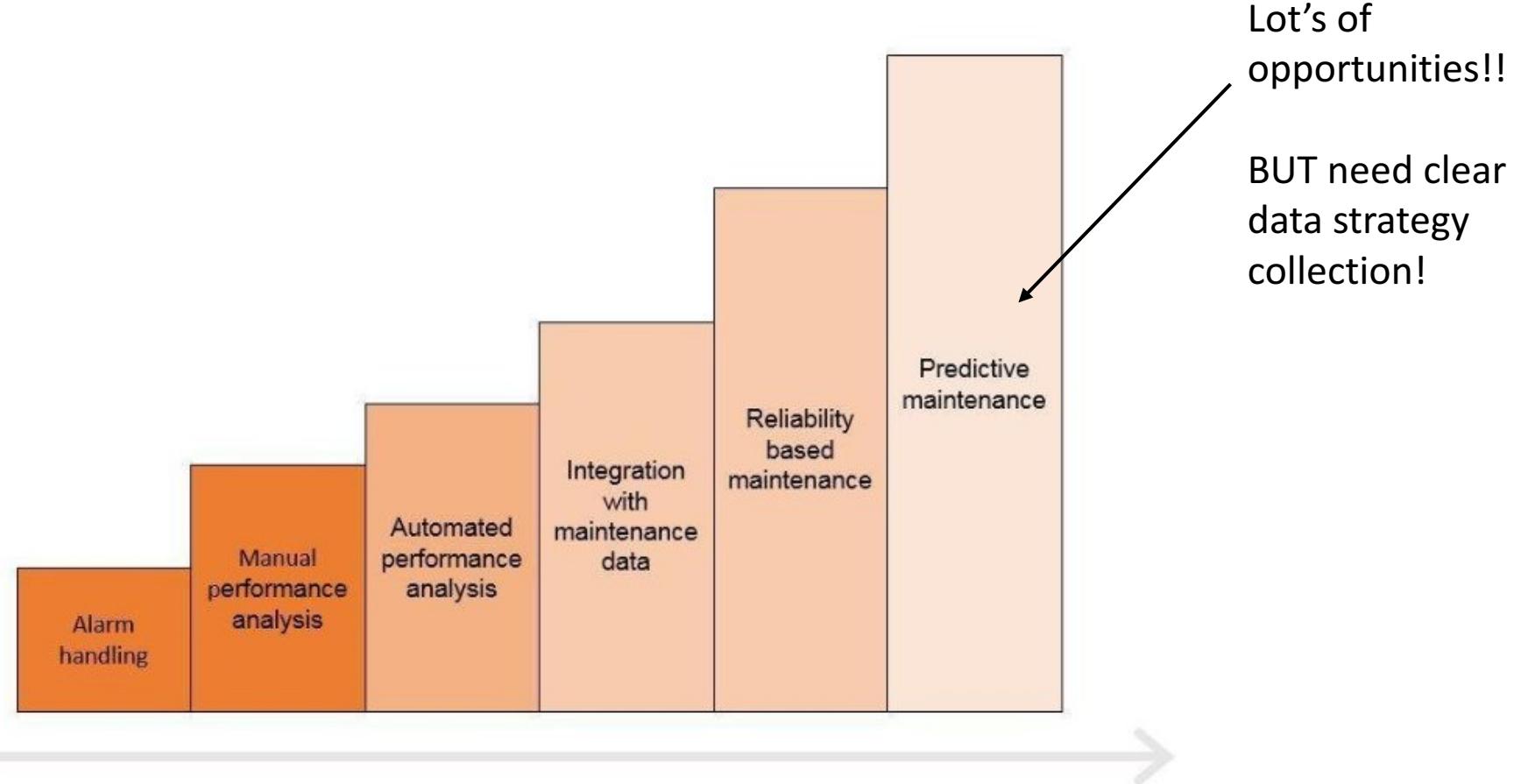
- Data driven services require integrated workflows
- Garbage in / garbage out
- No free lunch theorem
- Confusion Matrix

		Actual	
		Positive	Negative
Predicted	Positive	100	20
	Negative	10	50

A hand-drawn curly arrow points from the bottom right corner of the matrix (labeled 50) up and to the left, ending near the top-left cell (labeled 100).

Keep track of miss-detection!

# Modeling in Operation & Maintenance



# So what can we do??



Crunch time in France  
Ten years on: banking after the crisis  
South Korea's unfinished revolution  
Biology, but without the cells

MAY 6TH-12TH 2017

## The world's most valuable resource

Facebook

### The node pole: inside Facebook's Swedish hub near the Arctic Circle

Remote datacentre in Luleå cools itself using freezing outside air, has a fence to keep moose out and processes your selfies

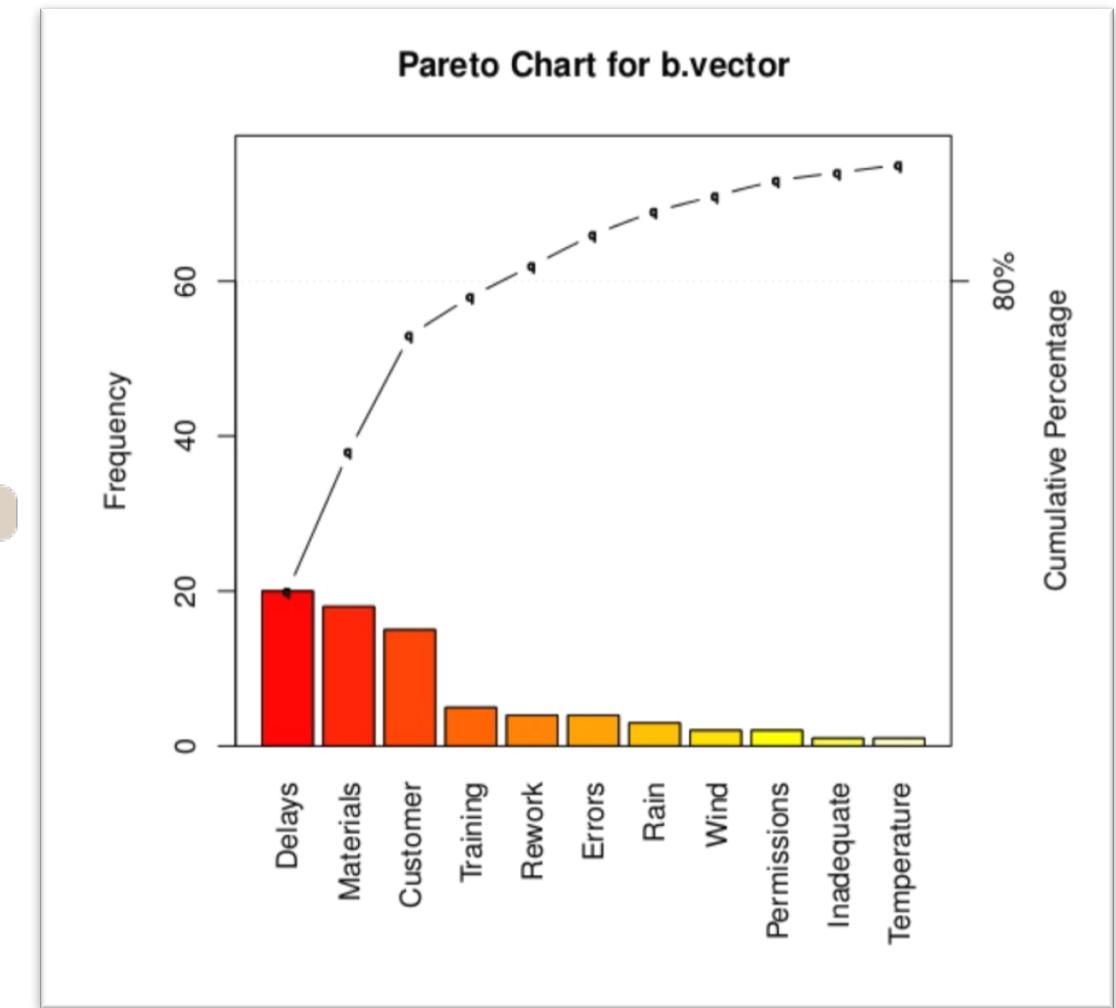
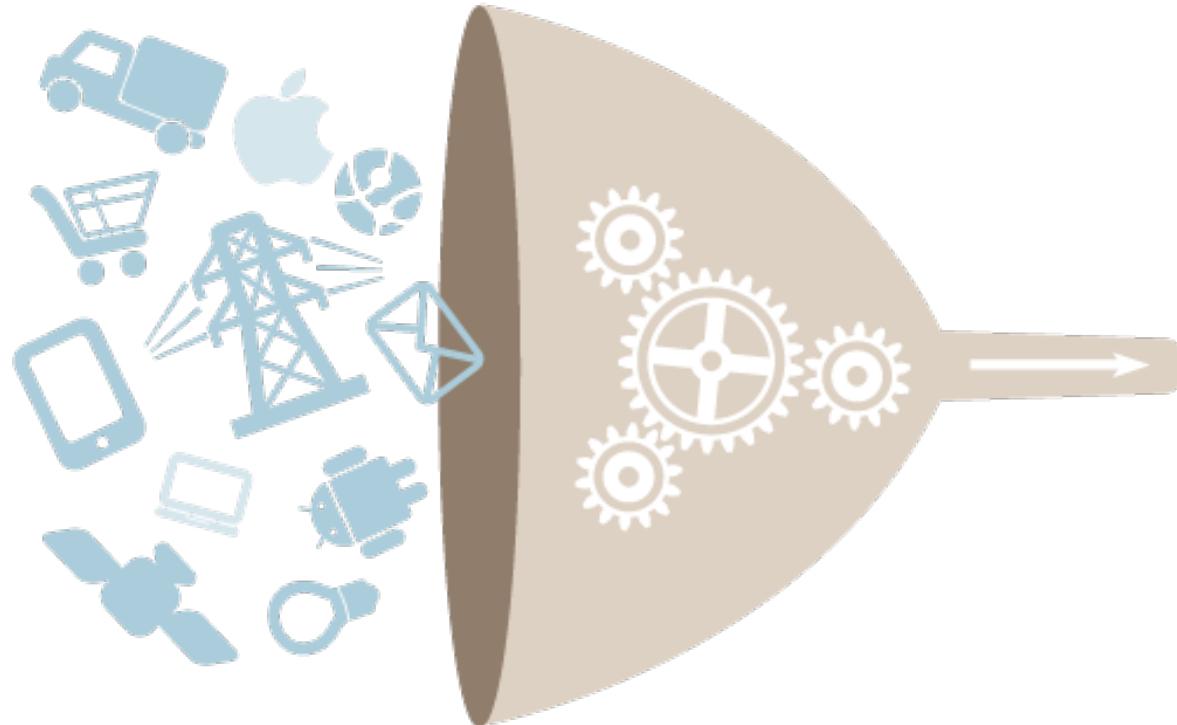
"By sharing their algorithms, Facebook and Google are merely sharing the recipe. Someone has to provide the eggs and flour and provide the baking facilities (which in Google and [Facebook's case](#) are vast data-computation facilities, often located near hydroelectric power stations for cheaper electricity)."

The Guardian.

"This is probably why Facebook and Google have so freely shared their methodologies: they know that the real value in their companies is the vast quantities of data they retain about each one of us."

The Guardian.

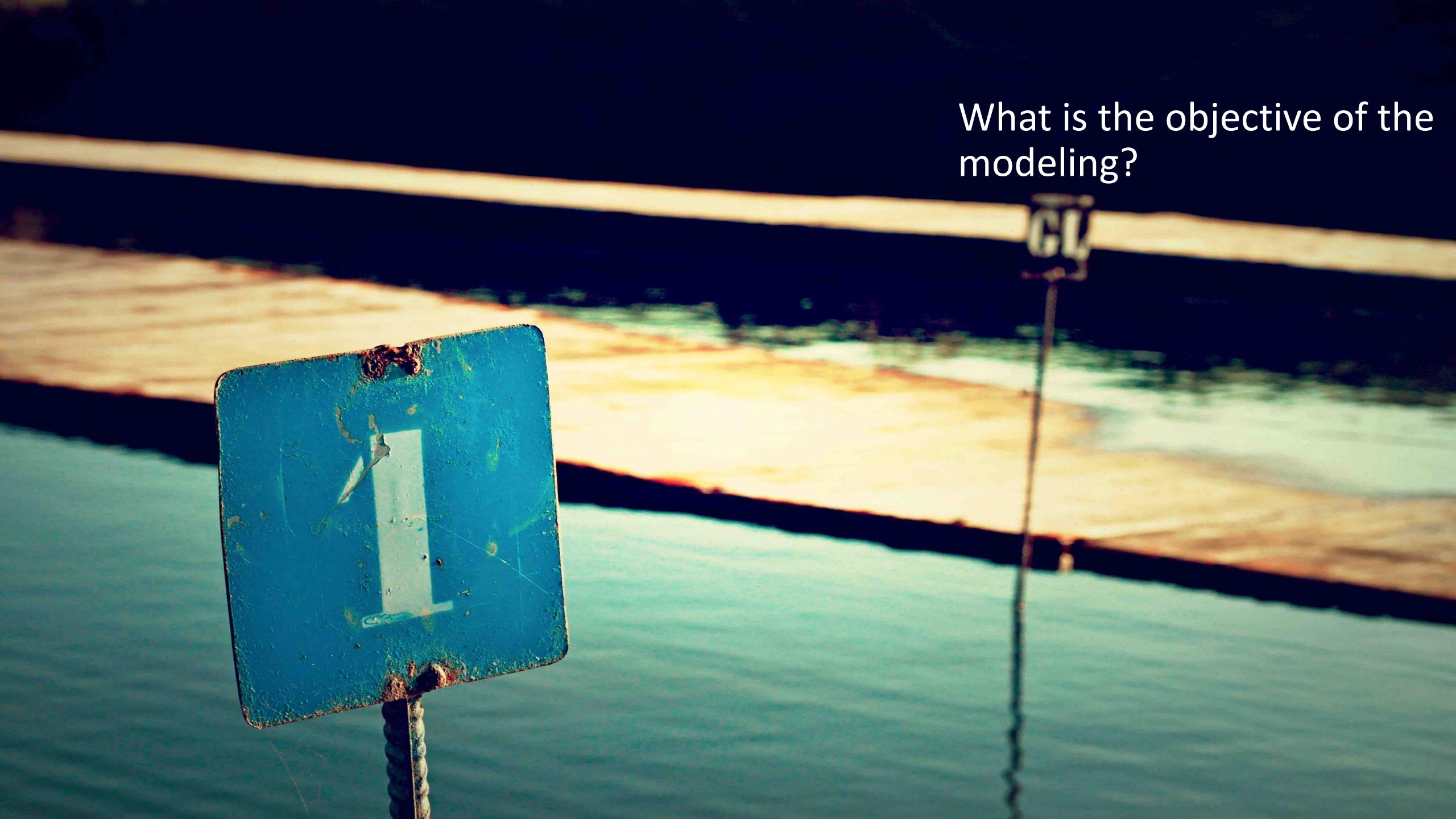
# Data Centric



# Six steps to improve your modeling



*Photo by Clem Onojeghuo*

A photograph of a weathered blue metal signpost on a wooden pier. The signpost has two signs: one facing forward with a white 'CL' logo on a dark background, and another sign further back with a similar logo. The pier is made of light-colored wooden planks and railings, extending into a body of water under a clear sky.

What is the objective of the modeling?

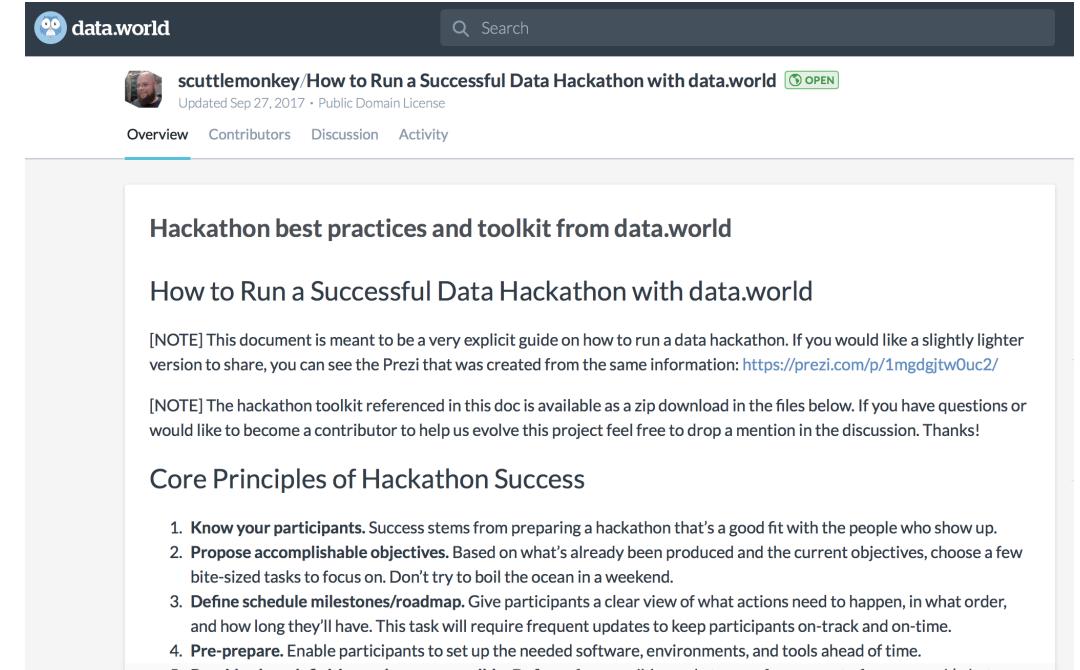
A close-up photograph of a person's hand making a thumbs-up gesture. The hand is positioned in the lower-left foreground, with the thumb pointing upwards and the fingers curled. The background is bright and out of focus, creating a lens flare effect that highlights the hand.

What is the ideal data set?

# Step 2 – Ideal data set

Cluster	Parameter
Time	Age of component Time
Stress	Full load hours Shear modulus Deviations
Environment	AMB temperature Wind speed Wave height Wake effect
Maintenance	Crane/non-crane components Rate/degree/effort of maintenance Human factor

Guide [Data sharing](#) by Jeff Leek



The screenshot shows a data.world document page. At the top, there's a header with the data.world logo, a search bar, and a navigation bar with links for Overview, Contributors, Discussion, and Activity. The main content area has a title 'Hackathon best practices and toolkit from data.world' and a sub-section 'How to Run a Successful Data Hackathon with data.world'. Below this, there are two notes: one about a Prezi presentation and another about a hackathon toolkit available as a zip download. The final section is titled 'Core Principles of Hackathon Success' with a numbered list of four items.

**Hackathon best practices and toolkit from data.world**

**How to Run a Successful Data Hackathon with data.world**

[NOTE] This document is meant to be a very explicit guide on how to run a data hackathon. If you would like a slightly lighter version to share, you can see the Prezi that was created from the same information: <https://prezi.com/p/1mgdgjtw0uc2/>

[NOTE] The hackathon toolkit referenced in this doc is available as a zip download in the files below. If you have questions or would like to become a contributor to help us evolve this project feel free to drop a mention in the discussion. Thanks!

**Core Principles of Hackathon Success**

1. Know your participants. Success stems from preparing a hackathon that's a good fit with the people who show up.
2. Propose accomplishable objectives. Based on what's already been produced and the current objectives, choose a few bite-sized tasks to focus on. Don't try to boil the ocean in a weekend.
3. Define schedule milestones/roadmap. Give participants a clear view of what actions need to happen, in what order, and how long they'll have. This task will require frequent updates to keep participants on-track and on-time.
4. Pre-prepare. Enable participants to set up the needed software, environments, and tools ahead of time.



Is there domain  
expertise available?

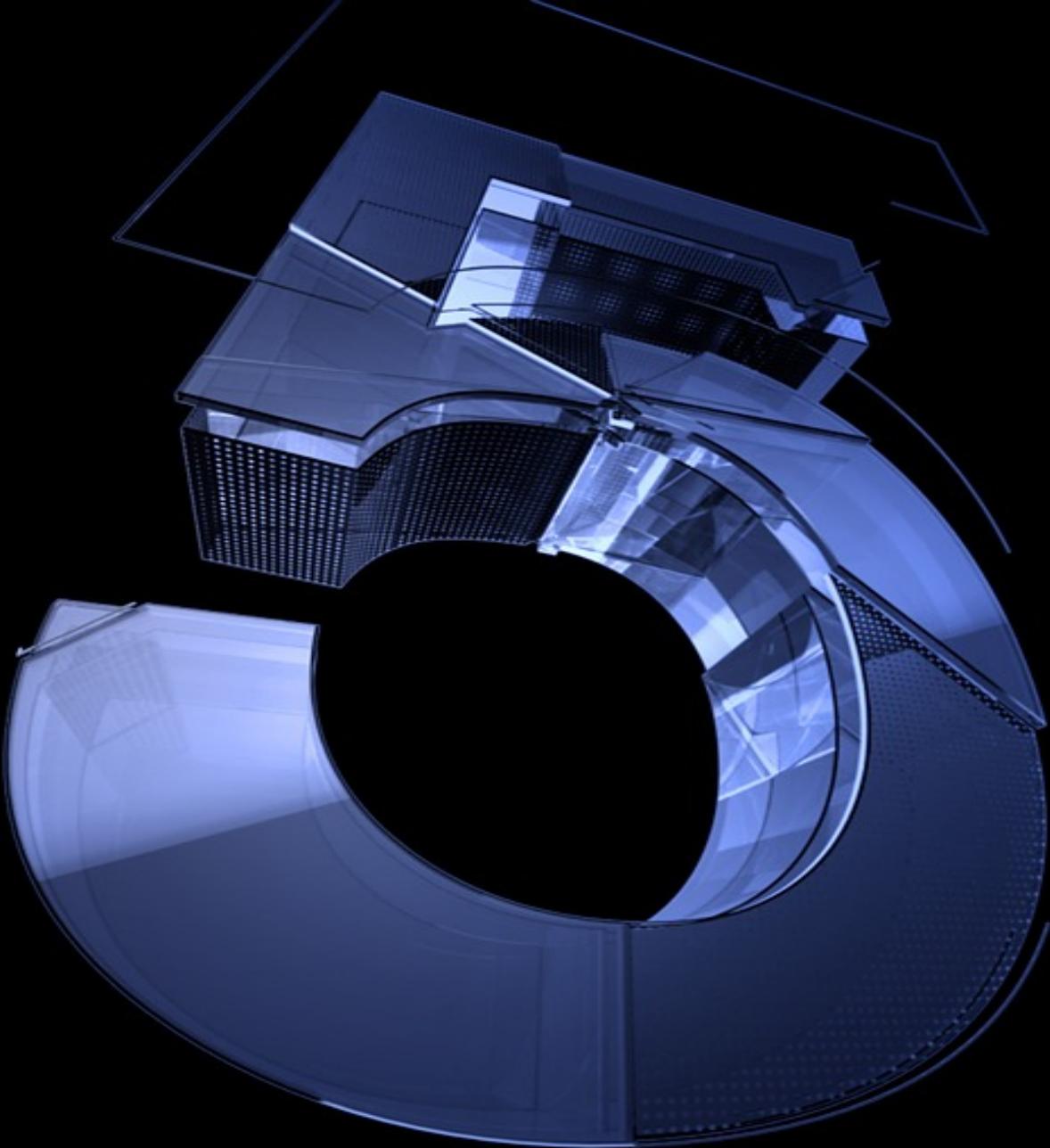
Elevator





What data is  
available?





Is the data clean  
& usable?

# Step 5 – Data cleaning

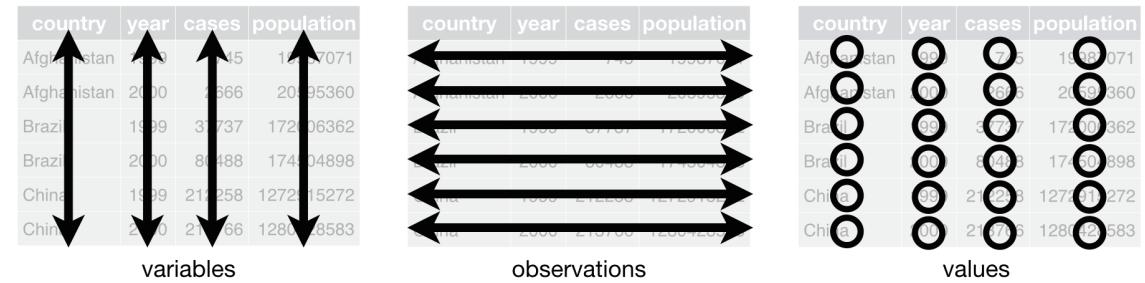
## The life of a data scientist

Data scientists, according to interviews and expert estimates, spend from 50 percent to 80 percent of their time mired in this more mundane labor of collecting and preparing unruly digital data, before it can be explored for useful nuggets.

-- "For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insight" - The New York Times, 2014

# Step 5 – Data cleaning

- Guides
  - [Data sharing by Jeff Leek](#)
  - [The Quartz guide to bad data](#)
  - Reliability Centered Maintenance
    - Asset data register
- Data steward
- Is it tidy data?
  - Transactional data
  - SCADA data
  - Failure data (time, energy)



1. Each variable in the data set is placed in its own column
2. Each observation is placed in its own row
3. Each value is placed in its own cell

Discovery Loop

1

2

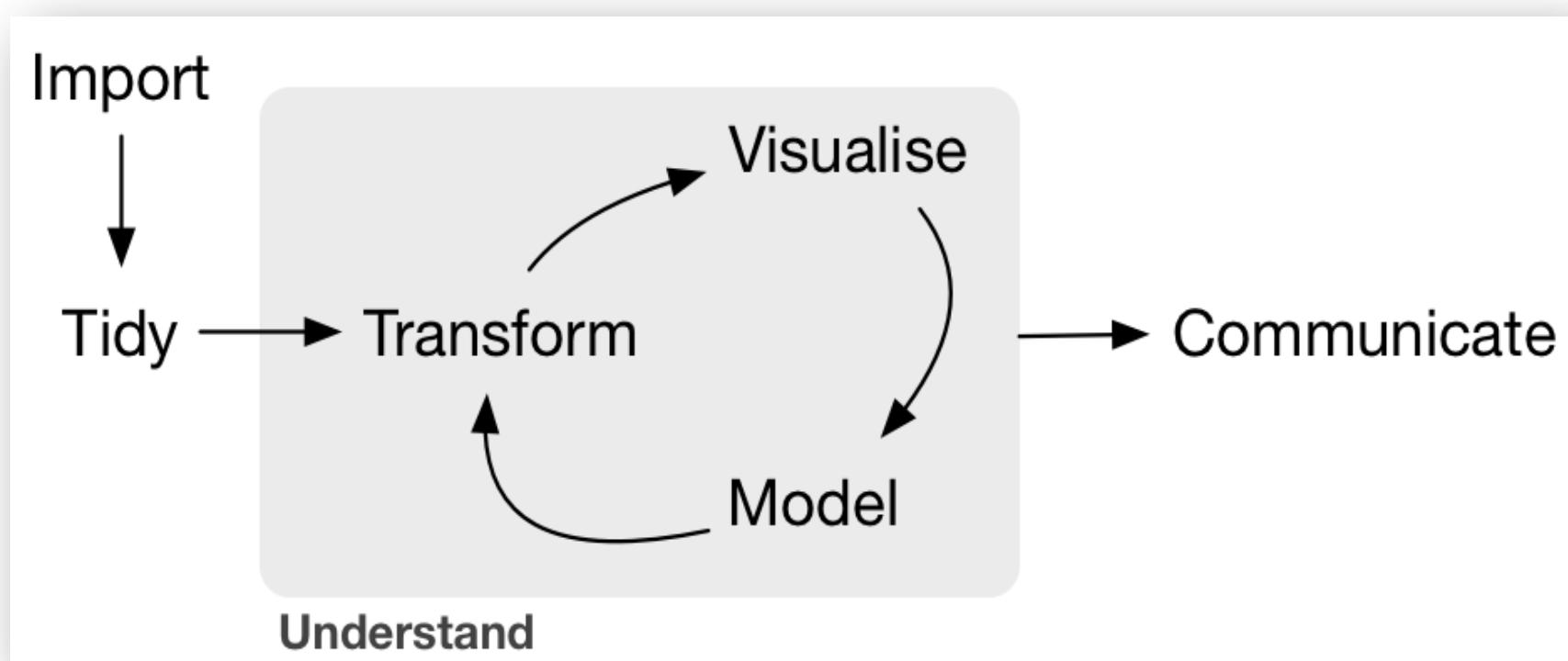
3

4

5

6

# Step 6 – Discovery loop



by Hadley Wickham

A photograph showing several people's hands and legs around a table, looking at a laptop screen. The laptop screen displays a virtual keyboard. The background is a dark, textured surface.

www.makeaweibull.com

Demo

#makeaweibull

# Main Takeaways

- There is a process!
  - Step 1 – Define objective of the modeling
  - Step 2 – Define ideal data set
  - Step 3 – Add domain expert (iterate step 1 & 2)
  - Step 4 – Document available data
  - Step 5 – Clean data & make usable
  - Step 6 – Discovery loop
- Base models go a long way!

[bit.ly/BarcelonaR](http://bit.ly/BarcelonaR)

# Thank you!

## Q&A



**RELIABLE DYNAMICS**

*Carles CG*

*Data Scientist & founder*  
*carles@reliabledynamics.com*



*carlescg*

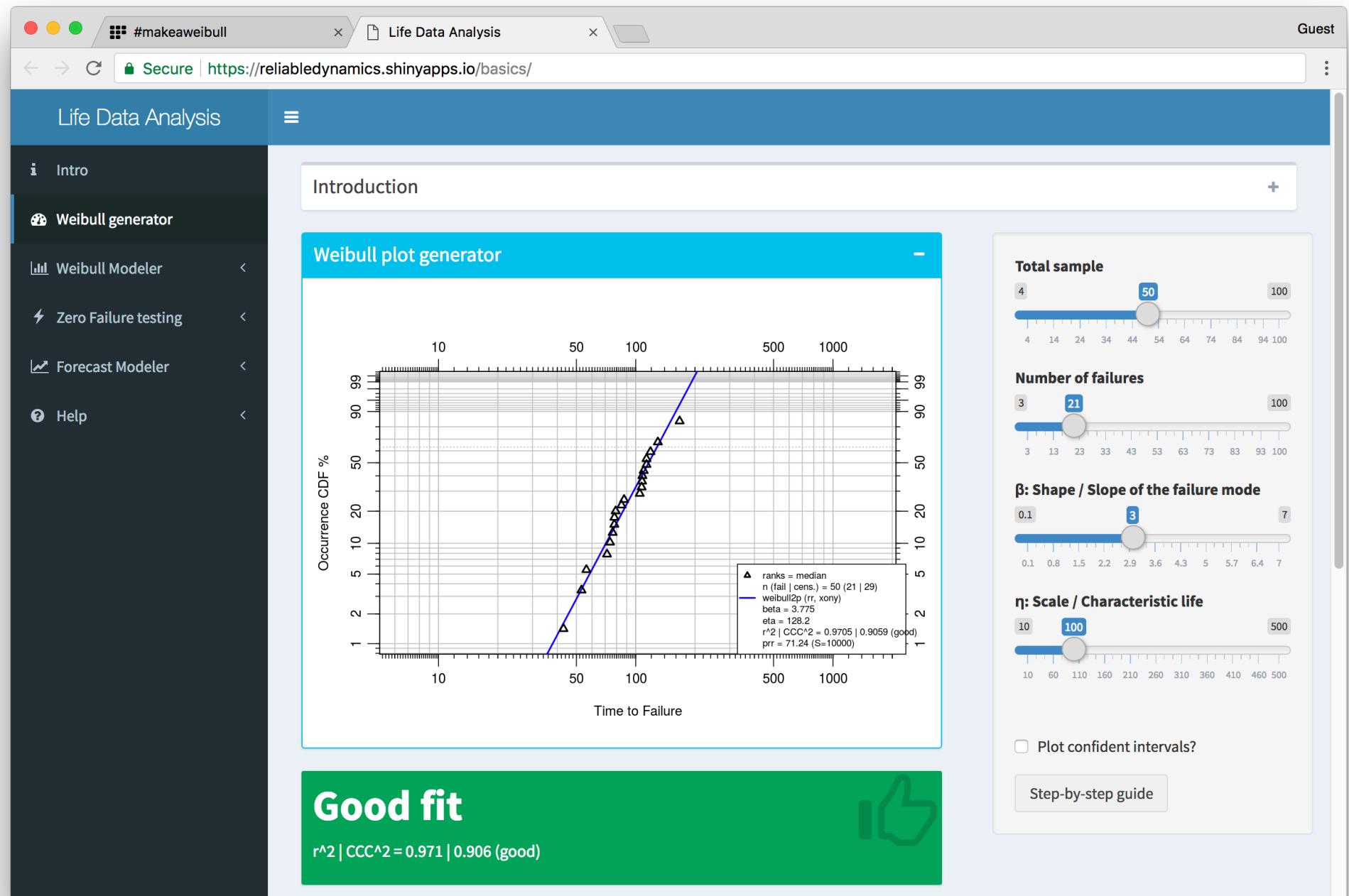


*@carles\_*

[makeaweibull.com](http://makeaweibull.com)

Guest

#makeaweibull https://www.makeaweibull.com



Good fit

$r^2 | CCC^2 = 0.971 | 0.906$  (good)



#makeaweibull

Secure | https://reliabilitydynamics.shinyapps.io/basics/

Guest

Introduction

Weibull plot generator

The modules are composed into two parts: **inputs** and **outputs**.

Inputs will be generally in this side of the module. They can be composed of sliders, numerical inputs, checkboxes, etc. They are meant to be interactive and they will perform calculations on-the-fly on the server.

**Total sample**

Number of failures

**$\beta$ : Shape / Slope of the failure mode**

**$\eta$ : Scale / Characteristic life**

Plot confident intervals?

**Good fit**

$r^2 | CCC^2 = 0.971 | 0.906$  (good)

**Contour plot**

Step-by-step guide

#makeaweibull

Secure | https://reliabilitydynamics.shinyapps.io/basics/

## Life Data Analysis

- Intro
- Weibull generator
- Weibull Modeler
- Upload data**
- Calculate
- Zero Failure testing
- Forecast Modeler
- Help

### Data

row_id	part_id	time	event	failure_mode
1	A10	412	1	Mode A
2	A09	551	1	Mode A
3	A08	858	1	Mode A
4	A08	600	1	Mode A
5	A08	700	1	Mode A
6	A08	100	1	Mode A
7	A01	913	0	None
8	A02	913	0	None
9	A03	913	0	None
10	A04	913	0	None
11	A05	913	0	None
12	A06	913	0	None
13	A07	913	0	None
14	A08	55	0	None
15	A09	269	0	None

Show 25 entries Search:

**Choose CSV File**

Browse... Weibull\_template (2).csv Upload complete

Header

**Separator**

Comma  Semicolon  Tab

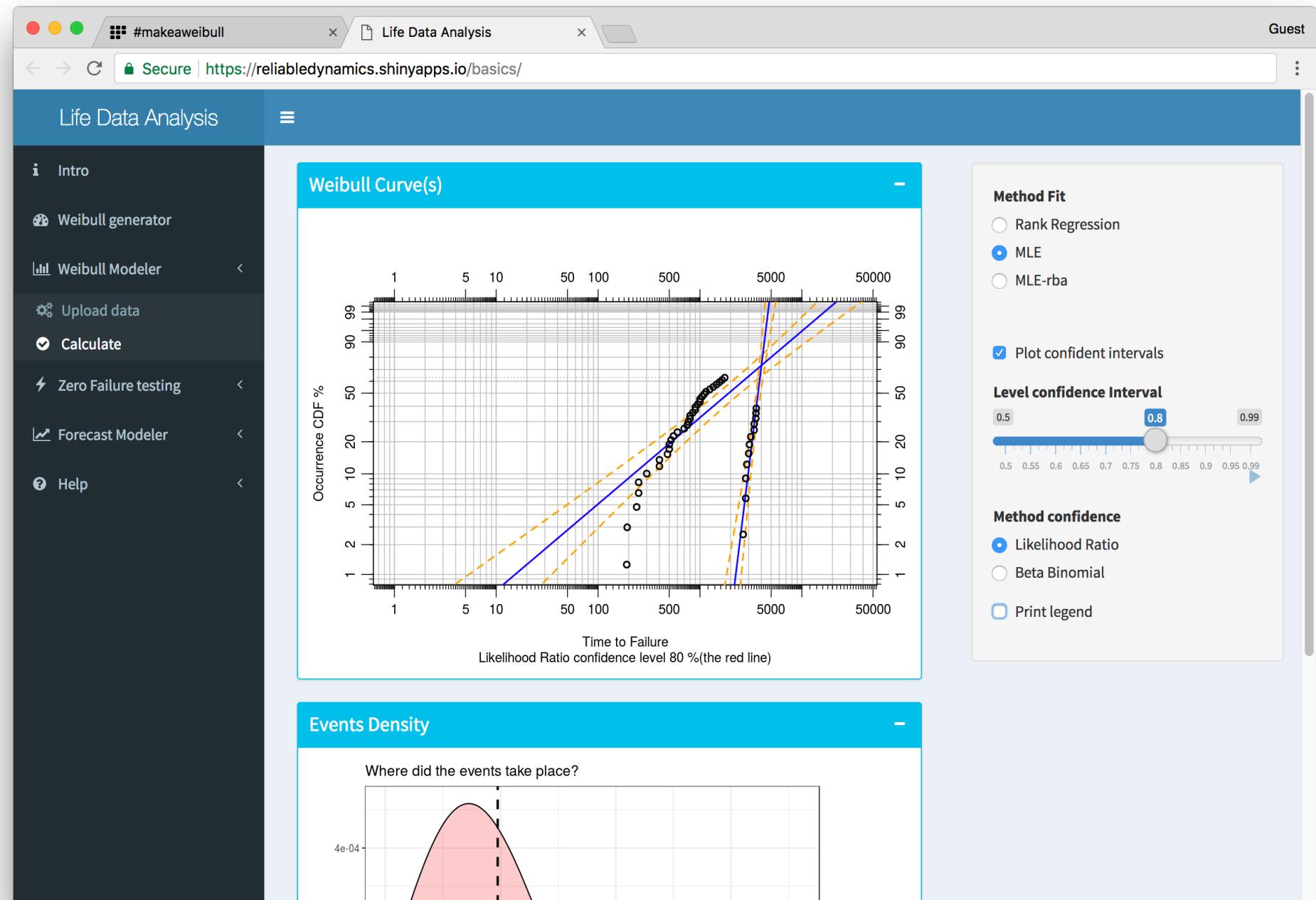
**Quote**

None  Double Quote  Single Quote

Need a template?

[Download](#)

[Guide](#)



#makeaweibull

Secure | https://reliabilitydynamics.shinyapps.io/basics/

## Life Data Analysis

- Intro
- Weibull generator
- Weibull Modeler
- Zero Failure testing
- Time testing complete
- Time testing partial
- Forecast Modeler
- Help

### Introduction

#### Partial Testing Curves

Time	Approximate Components Required for 0.85 Confidence
120	~100
150	~80
180	~70
210	~65
240	~60
270	~55
300	~50

Number components to test against confidence level, depending on the time available. In general, the less time to test the more components are necessary

**Time:**

- 120
- 150
- 180
- 210
- 240
- 270
- 300

**Confidence:** 0.7 0.85 0.99

**Buttons:**

- Zero failure test
- Download report

Demo End