Continuous Data Quality Improvement with R

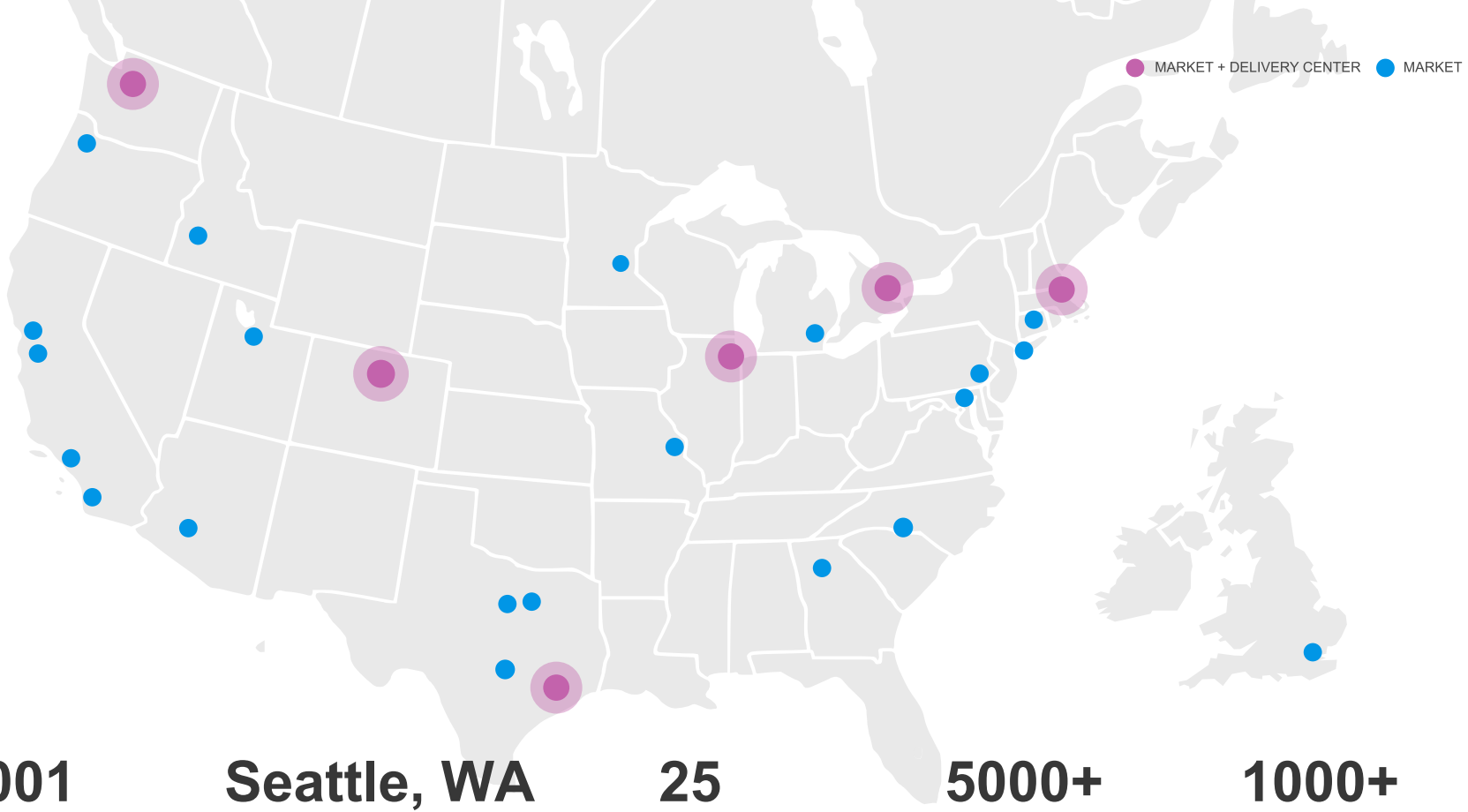
Frank Farach, PhD

@frankfarach

Cascadia R Conf | June 2018

Photo by Edu Lauton on Unsplash

slalom



2001
YEAR FOUNDED

Seattle, WA
HEADQUARTERS

25
LOCAL MARKETS

5000+
CONSULTANTS

1000+
CLIENTS

Photo by rawpixel on Unsplash

@frankfarach

“Quality is everyone’s responsibility.”

W. Edwards Deming

Better Data
Better Decisions
Better Outcomes
(probabilistically)

Data Quality = Fitness for use

Do the pipes work?

What's in the pipes?

Sources of quality expectations

Logical entailment
Domain experts
Business/legal requirements
User testing
Prior experience

“Testing should be addictive, so you do it all the time.”

README.md - *testthat*

testthat

build passing | build unknown | codecov 83% | CRAN 2.0.0

covr

build passing | build unknown | codecov 80% | CRAN 3.1.0

Track test coverage for your R package and view reports locally or (optionally) upload the results to codecov or coveralls.

Do the pipes work?

What's in the pipes?

Assertions to check expectations

Structural

- Shape
- Missing values
- Data types

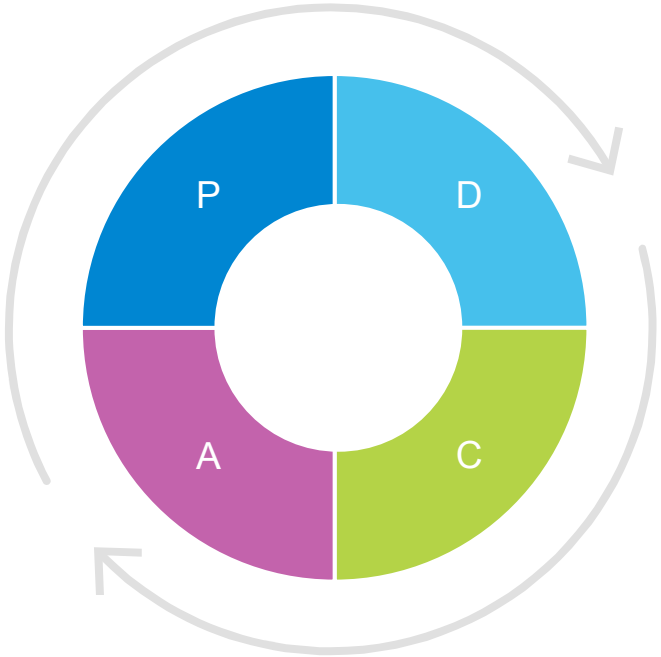
Statistical

- Aggregates
- Distributions

Logical

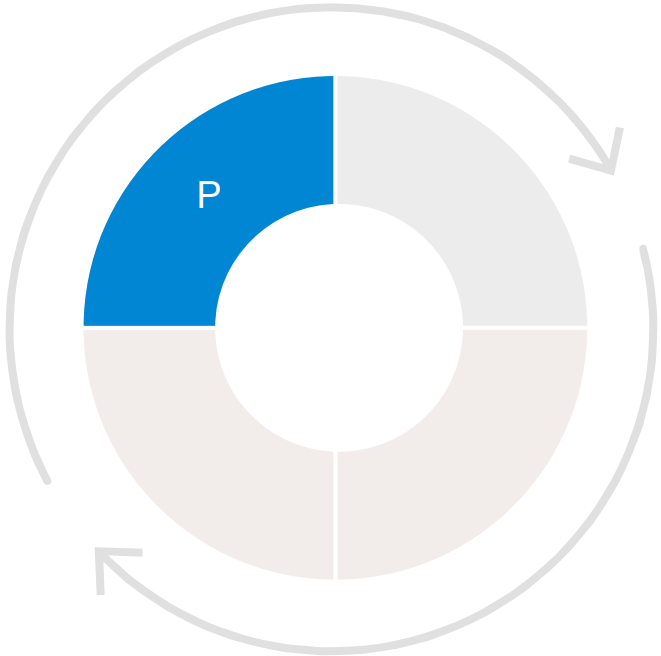
- Value domains
- Ref. integrity

Build Quality in with PDCA



1. **Plan**
2. **Do**
3. **Check**
4. **Act**

1. PLAN a specific improvement



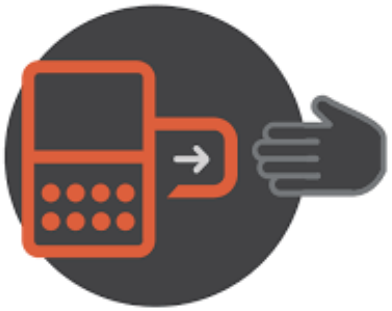
Select an objective

Formulate assertions

Specify error behavior

1. PLAN a specific improvement

BikeTown: How do customers access bike rentals?



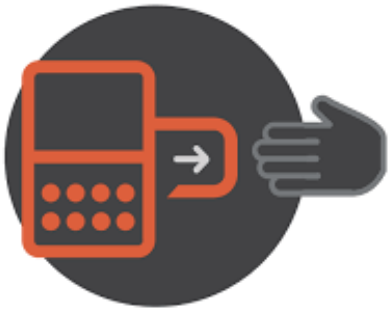
keypad

mobile

web

1. PLAN a specific improvement

Assertion	Error Behavior
More than 1,000 rows	Fail on error
Valid values for rental access path	Fail on error



keypad

mobile

web

2. DO: Run the validation

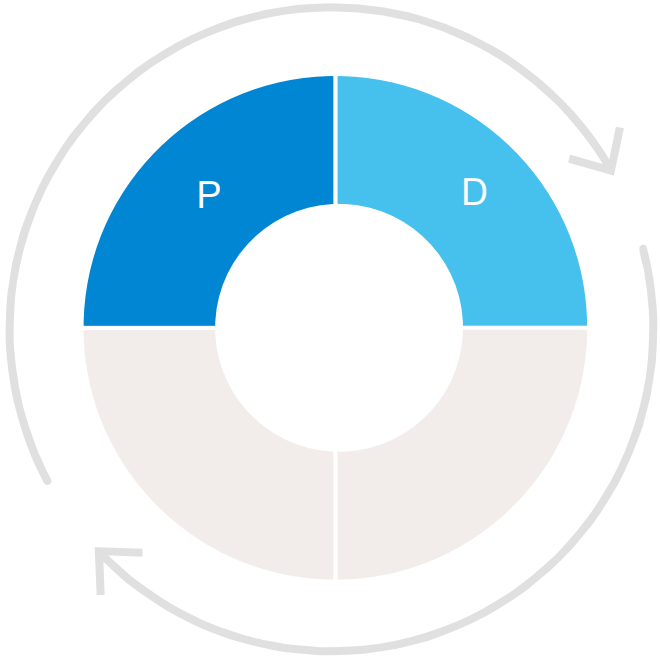
assertr

```
install.packages("assertr")
library(assertr)
```

magrittr


```
install.packages("magrittr")
library(magrittr)
```

2. DO: Run the data validation



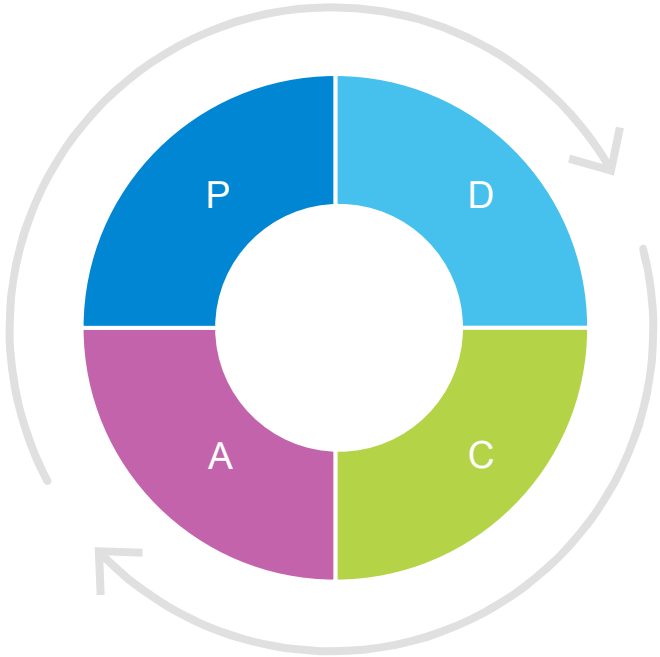

```
valid_vals <-
c(“keypad”, “mobile”, “web”)

data %>%
  verify(nrow(.) > 1000) %>%
  assert(in_set(valid_vals),
    RentalAccessPath)
```

3. CHECK: What happened?

Assertion	Result
More than 1,000 rows	OK
Valid values for rental access path	Bad value “admin” in row 9641

```
Column 'RentalAccessPath' violates assertion 'in_set(valid_vals)' 1 time
    verb redux_fn         predicate           column index value
1 assert       NA in_set(valid_vals) RentalAccessPath  9641 admin
Error: assertr stopped execution
```

4. ACT: Make an informed change



Root cause analysis

Check with larger sample

Exploratory data analysis

Productionize assertions

Continuous: Never Done

Photo by chrissie kremer on Unsplash


KEEP
CALM
AND
PLAN-DO-
CHECK-ACT

slalom

slalom.com

@frankfarach

frankfarach.com