TEST-DRIVEN DATA ANALYSIS

Nicholas J. Radcliffe
Stochastic Solutions Limited
& Department of Mathematics, University of Edinburgh

TEST-DRIVEN DATA ANALYSIS

or

Why should anyone believe your analytical results?

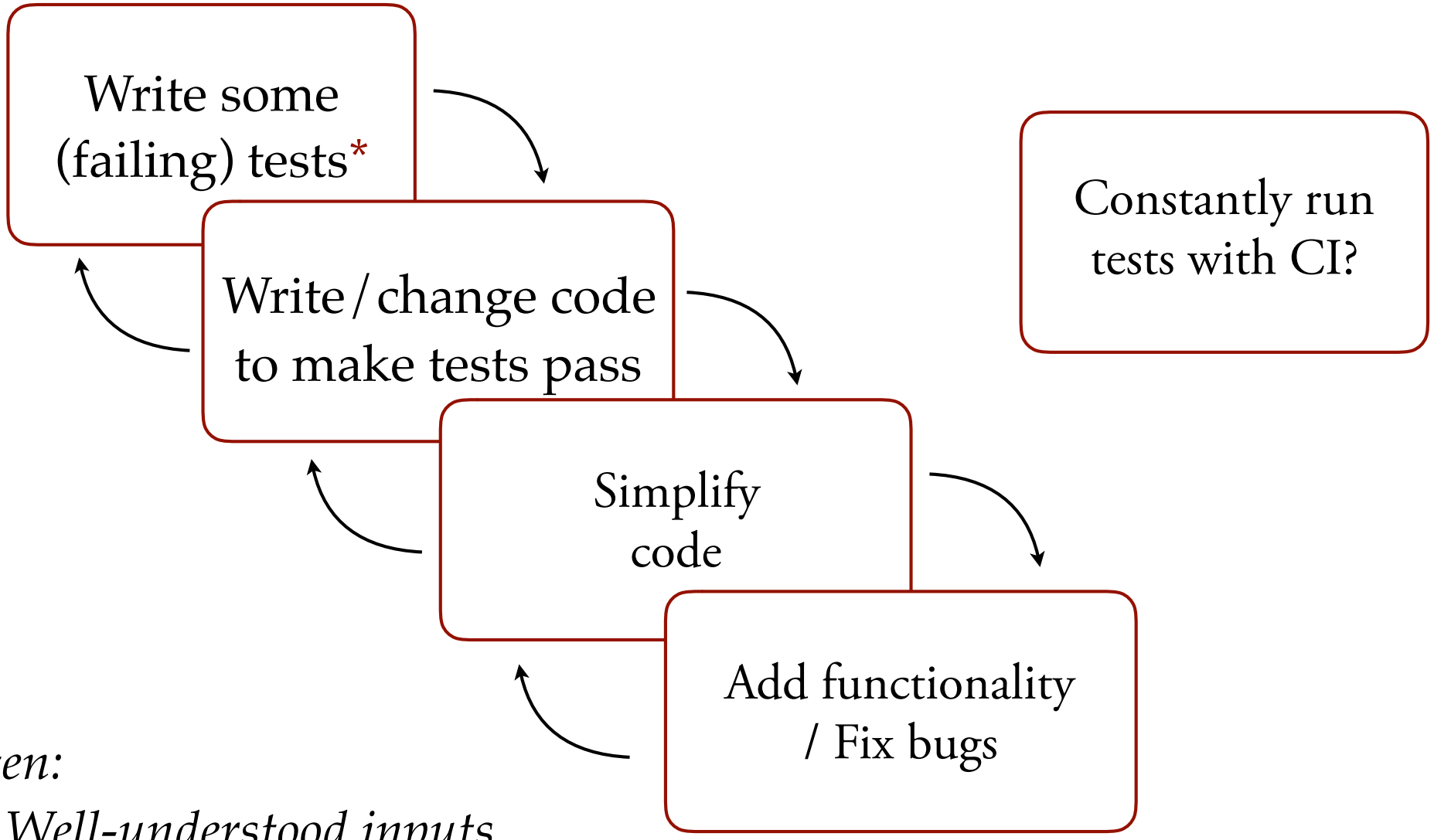
or

Data analysis as if the answers actually mattered

Image: Dave Briggs, Flickr, https://www.flickr.com/photos/theclosedcircle/3349126651/

Writing code
is not like
riding a bike

SOFTWARE DEVELOPMENT (WITH TDD)



Often:

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

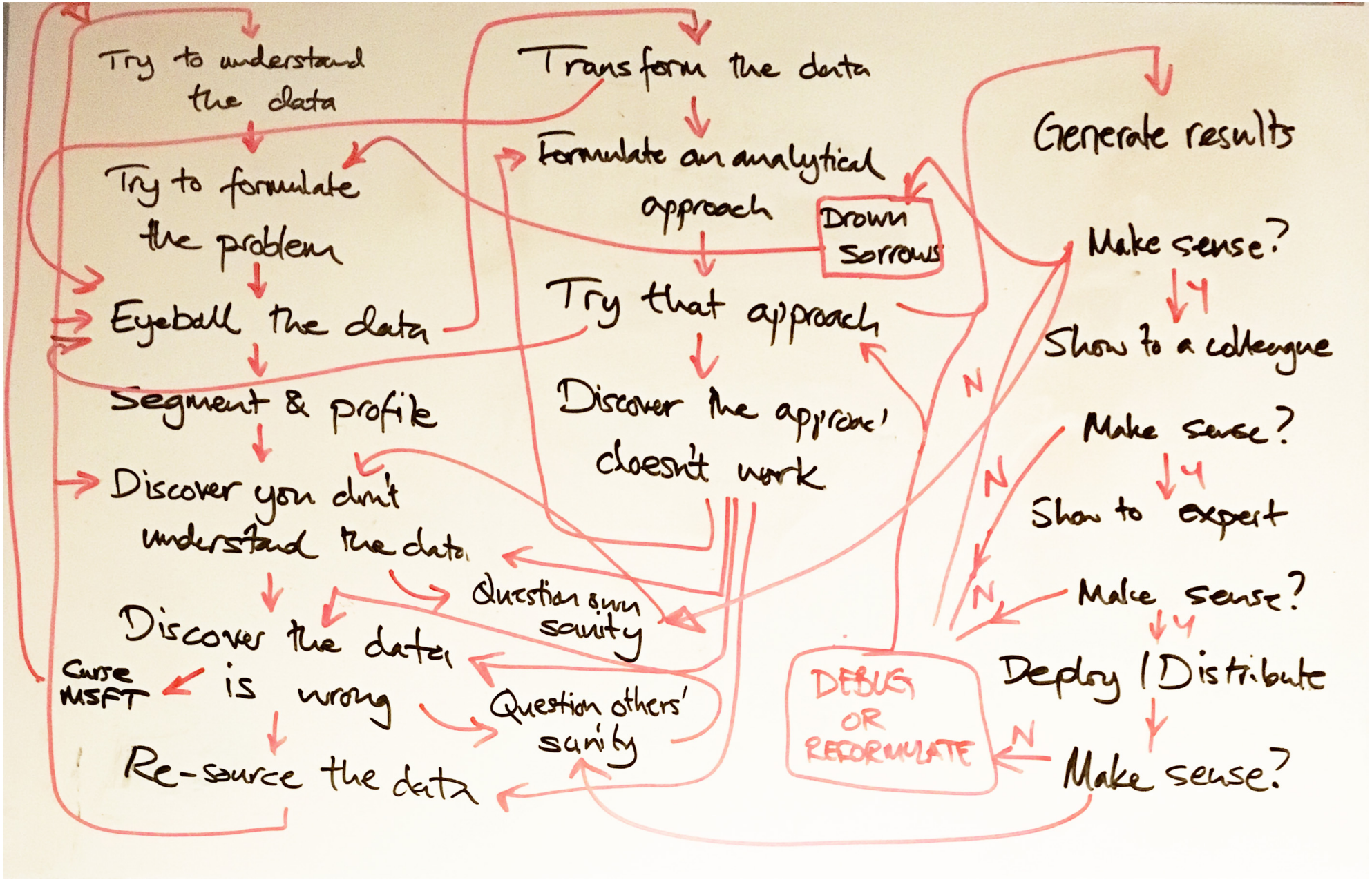
* While mocking almost everything

“Don’t mock. It’s not kind.”

— *Animal’s People,* Indra Sinha.

Why is this
lying bastard
lying to me?

—Jeremy Paxman


Try to understand the data
Try to formulate the problem
Eyeball the data
Segment & profile
Discover you don't understand the data
Discover the data is wrong
Curse MSFT
Re-source the data
Transform the data
Formulate an analytical approach
Drown sorrows
Try that approach
Discover the approach doesn't work
Question own sanity
Question others' sanity
DEBUG OR REFORMULATE
Generate results
Make sense?
Y
Show to a colleague
Make sense?
Y
Show to expert
Make sense?
Y
Deploy / Distribute
Make sense?
N
N
N
N

"Shouldn't we do test-driven data analysis?"

— Patrick Surry, c. 2010

THE BIG IDEA

Transfer the ideas of
test-driven development
from software development
to data analysis
(mutatis mutandis)*

* *those things which need to be changed having been changed*

CORRECTNESS

STURGEON'S REVELATION

"Ninety per cent of everything is crap"

— Theodore Sturgeon, Venture, 1958

TDD ↦ TDDA

We need to extend TDD's idea of testing for

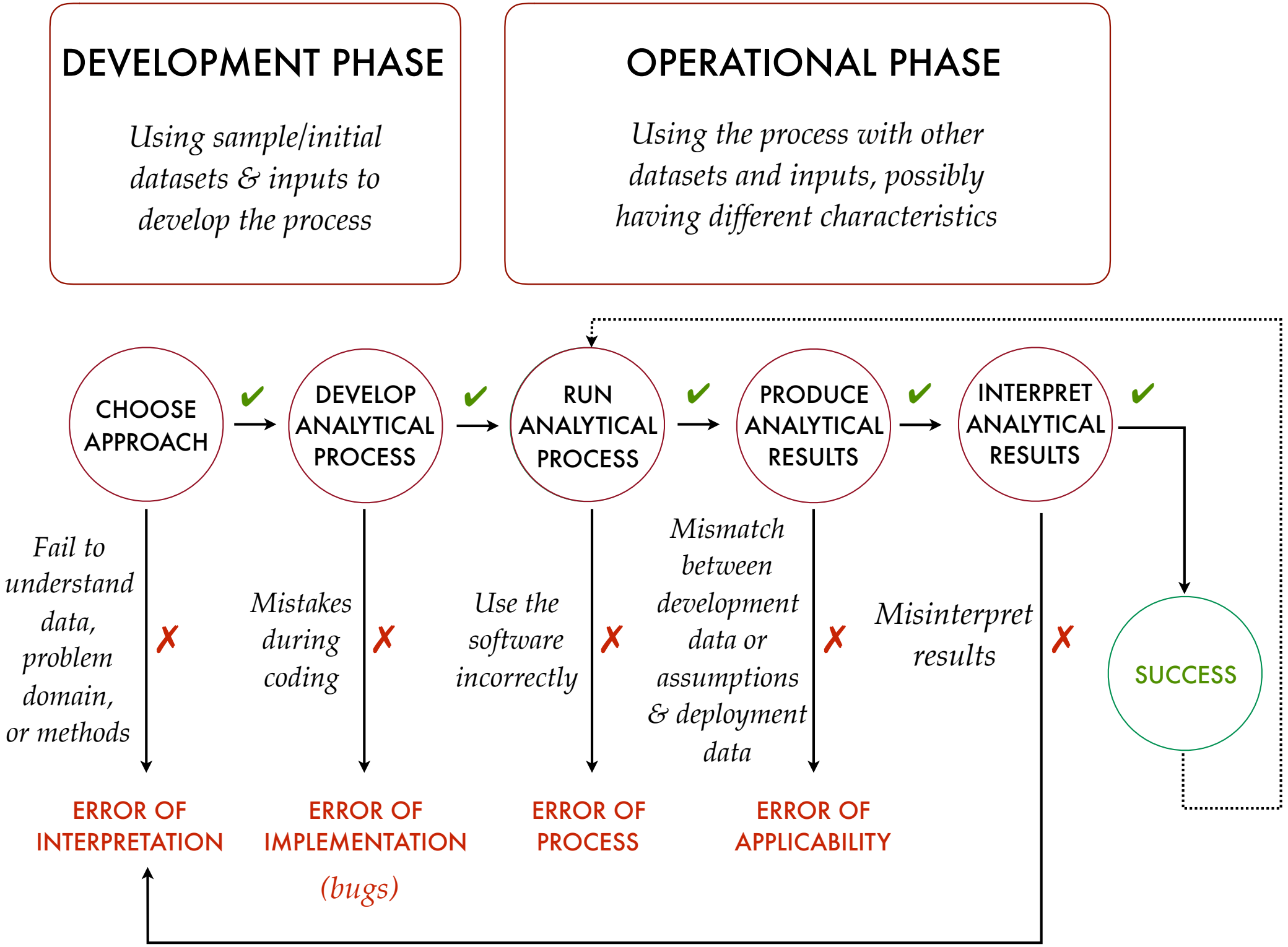
software correctness

with the idea of testing for

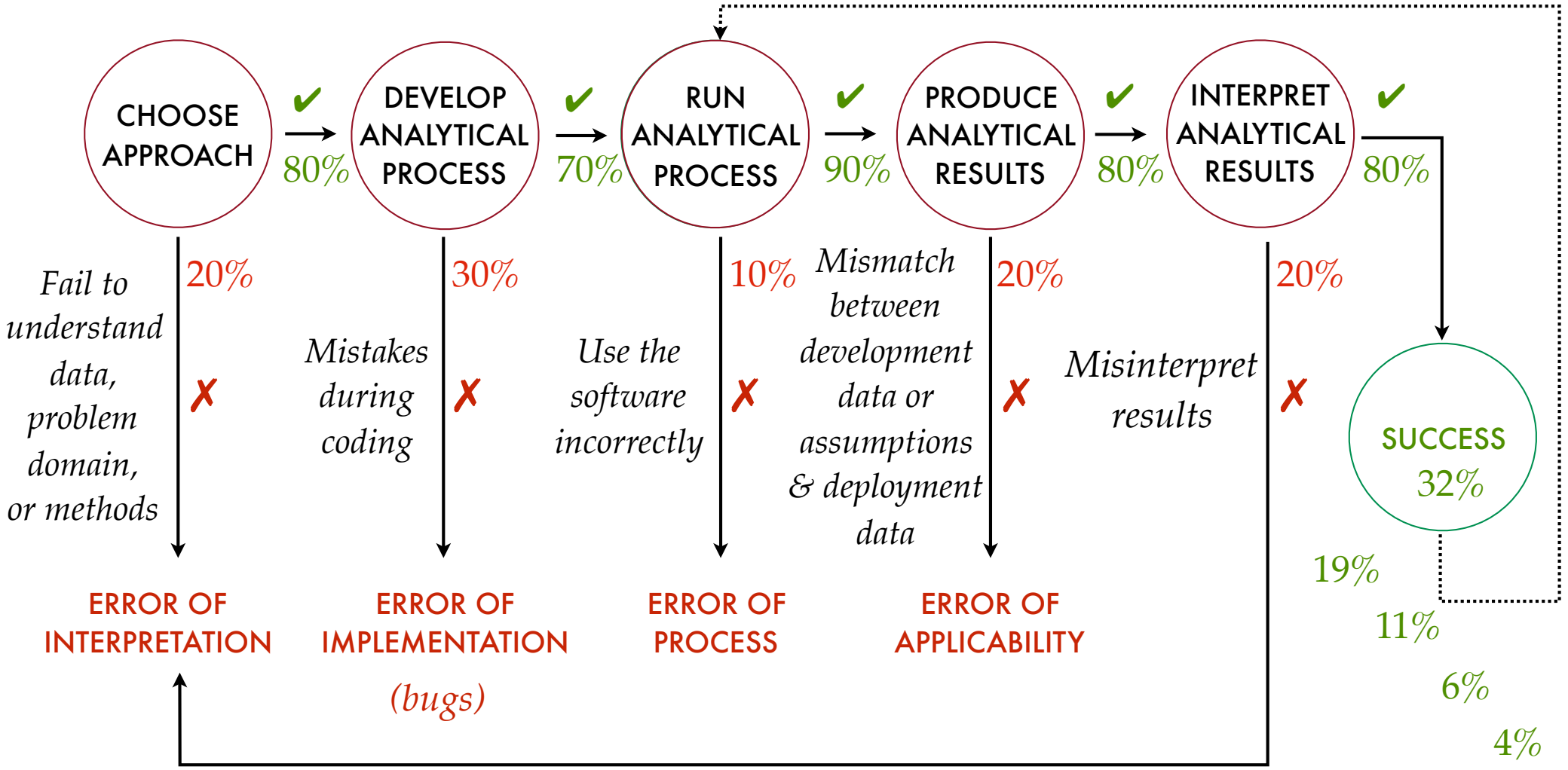
meaningfulness of analysis,

correctness and validity of input data,

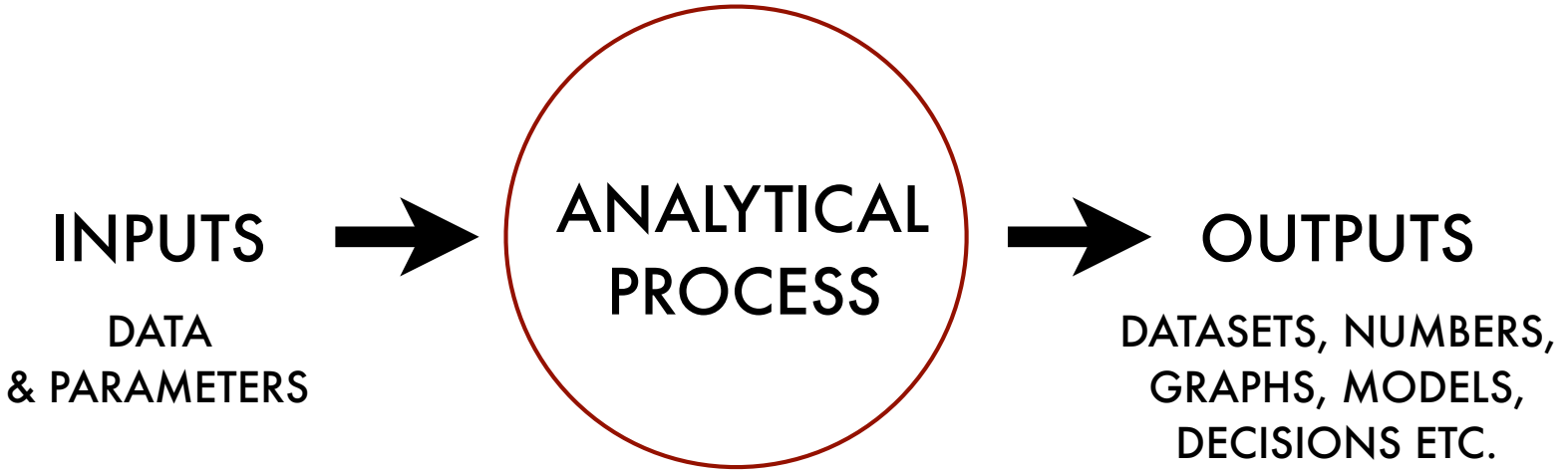
& *correctness of interpretation*.


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
SUCCESS
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
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



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

```
0 :$ python3 test_graphs.py
.....
File check failed.
Compare with "diff /Users/njr/pycon/testoutput/tsr-288-DS-5.svg /Users/njr/pycon/
reference/tsr-288-DS-5.svg".
Note exclusions: Copyright
Compare preprocessed with "diff /tmp/actual-tsr-288-DS-5.svg /tmp/expected-tsr-288-
DS-5.svg".
F
======================================================================
FAIL: testTSRGraphs (__main__.TestGraphs)
----------------------------------------------------------------------
Traceback (most recent call last):
  File "test_graphs.py", line 111, in testTSRGraphs
    maxPermutationCases=2)
  File "/Users/njr/python/tdda/writabletestcase.py", line 170, in check_file
    self.assertEqual(nFailures, 0)
AssertionError: 1 != 0

----------------------------------------------------------------------
Ran 9 tests in 11.320s

FAILED (failures=1)
```

```
1 :$ python3 test_graphs.py -W
Written /Users/njr/pycon/reference/pndate-graph.svg.
.Written /Users/njr/pycon/reference/test-pnhbar-graph.svg.
.Written /Users/njr/pycon/reference/pnd-295-SGLD-28.svg.
Written /Users/njr/pycon/reference/pnd-295-SI-1136.svg.
Written /Users/njr/pycon/reference/pnd-295-DS-235.svg.
Written /Users/njr/pycon/reference/pnd-428-BH-44.svg.
Written /Users/njr/pycon/reference/pnd-295-SI-1139.svg.
Written /Users/njr/pycon/reference/pnd-295-SGLD-8.svg.
Written /Users/njr/pycon/reference/pnd-295-BD-97.svg.
Written /Users/njr/pycon/reference/pnd-295-BD-94.svg.
Written /Users/njr/pycon/reference/pnd-428-BH-4.svg.
...
Written /Users/njr/pycon/reference/tsr-273-DS-4.svg.
Written /Users/njr/pycon/reference/tsr-278-DS-4.svg.
Written /Users/njr/pycon/reference/tsr-282-DS-3.svg.
Written /Users/njr/pycon/reference/tsr-287-DS-3.svg.
Written /Users/njr/pycon/reference/tsr-287-PW-3.svg.
Written /Users/njr/pycon/reference/tsr-288-DS-5.svg.
.
----------------------------------------------------------------------
Ran 9 tests in 11.039s

OK
```

```
0 :$ python3 test_graphs.py
.........
----------------------------------------------------------------------
Ran 9 tests in 11.658s

OK
0 :$
```

TDDA: LEVEL ONE

CONSTRAINTS

EXAMPLE CONSTRAINTS

SINGLE FIELD CONSTRAINTS	DATASET CONSTRAINTS
Age ≤ 150	The dataset must contain field **CID**
type(Age) = int	Number of records must be 118
CID ≠ 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"}	StartDate ≤ EndDate
Vote ≠ "Trump"	AlmostEqual(F, m * a, 6)
StartDate < tomorrow()	sum(Favourite*) = 1
v < 2.97e10	minVal ≤ medianVal ≤ maxVal
Height ~ N(1.8, 0.2)	V ≤ H * w * d

Z	Name	Symbol	Period	Group	Chemical Series	Atomic Weight	Etymology	Relative Atomic Mass	Melting Point C	Melting Point K	Boiling Point C	Boiling Point F	Density	Description	Colour
1	Hydrogen	H	1	1	Nonmetal	1.01	Greek hydrogenes	1.01	-258.98	14.20	-252.87	-423.17	0.00	gas	colorless
2	Helium	He	1	18	Noble gas	4.00	Greek helios	4.00	⌀	⌀	-268.93	-452.07	0.00	noble gas	⌀
3	Lithium	Li	2	1	Alkali metal	6.94	Greek lithos	6.94	180.70	453.90	1,342.00	2,448.00	0.53	0.6	silvery white/gray
4	Beryllium	Be	2	2	Alkaline earth metal	9.01	beryl	9.01	1,287.00	1,560.00	2,469.00	4,476.00	1.85	5.5	⌀
5	Boron	B	2	13	Metalloid	10.81	borax	10.81	2,300.00	2,570.00	3,927.00	7,101.00	2.34	9.3	black/brown/ amorphous boron is a brown powder, metallic boron is black
6	Carbon	C	2	14	Nonmetal	12.01	Latin carbo	12.01	3,675.00	3,948.00	4,027.00	7,280.00	2.27	1-2 graphite	black
7	Nitrogen	N	2	15	Nonmetal	14.01	Greek nitron	14.01	-209.86	63.30	-195.79	-320.42	0.00	gas	⌀
8	Oxygen	O	2	16	Nonmetal	16.00	Greek oxys	16.00	-222.65	50.50	-182.95	-297.31	0.00	gas	⌀
9	Fluorine	F	2	17	Halogen	19.00	Latin fluo	19.00	-219.52	53.60	-188.12	-306.62	0.00	halogen gas	yellow-green or yellowish brown
10	Neon	Ne	2	18	Noble gas	20.18	Greek neos	20.18	-248.45	24.70	-246.08	-410.94	0.00	noble gas	⌀
11	Sodium	Na	3	1	Alkali metal	22.99	Latin natrium	22.99	98.00	371.00	883.00	1,621.00	0.97	0.5	waxy, silvery white
12	Magnesium	Mg	3	2	Alkaline earth metal	24.31	Magnesia, Greece	24.31	650.00	923.00	1,090.00	1,994.00	1.74	2.5	silvery metallic
13	Aluminium	Al	3	13	Poor metal	26.98	Latin alumen	26.98	660.25	933.40	2,519.00	4,566.00	2.70	2.75	silvery
14	Silicon	Si	3	14	Metalloid	28.09	Latin silex	28.09	1,410.00	1,680.00	3,265.00	5,909.00	2.33	6.5 metalloid	dark gray, bluish tinge
15	Phosphorus	P	3	15	Nonmetal	30.97	Greek phosphoros	30.97	44.10	317.30	280.00	536.00	1.82	nonmetal	waxy white/ red/ black/ colorless
16	Sulfur	S	3	16	Nonmetal	32.07	Latin sulfur	32.07	115.36	388.51	444.60	832.30	2.07	2	lemon yellow
17	Chlorine	Cl	3	17	Halogen	35.45	Greek chloros	35.45	-100.84	172.00	-34.04	-29.27	0.00	halogen gas	yellowish green or greenish yellow
18	Argon	Ar	3	18	Noble gas	39.95	Greek argon	39.95	-189.19	84.00	-185.85	-302.53	0.00	noble gas	⌀
19	Potassium	K	4	1	Alkali metal	39.10	Latin kalium	39.10	63.35	336.50	759.00	1,398.00	0.86	0.4	silvery white
20	Calcium	Ca	4	2	Alkaline earth metal	40.08	Latin calx	40.08	839.00	1,112.00	1,484.00	2,703.00	1.54	1.75	gray

```
[2]> discover -l

declare (field-exists "Z")
declare (field-exists "Name")
declare (field-exists "Symbol")
declare (field-exists "Period")
declare (field-exists "Group")
declare (field-exists "ChemicalSeries")
declare (field-exists "AtomicWeight")
declare (field-exists "Etymology")
declare (field-exists "RelativeAtomicMass")
declare (field-exists "MeltingPointC")
declare (field-exists "MeltingPointKelvin")
declare (field-exists "BoilingPointC")
declare (field-exists "BoilingPointF")
declare (field-exists "Density")
declare (field-exists "Description")
declare (field-exists "Colour")
declare (= (length (field-names)) 16)
```

```
if (field-exists "Z")
   declare (= (type Z) "int")
   declare (>= (min Z) 1)
   declare (<= (max Z) 118)
   declare (= (countnull Z) 0)
   declare (non-nulls-unique Z)
fi

if (field-exists "Name")
   declare (= (type Name) "string")
   declare (>= (min (length Name)) 3)
   declare (<= (min (length Name)) 13)
   declare (= (countnull Name) 0)
   declare (non-nulls-unique Name)
fi

if (field-exists "Symbol")
   declare (= (type Symbol) "string")
   declare (>= (min (length Symbol)) 1)
   declare (<= (min (length Symbol)) 3)
   declare (= (countnull Symbol) 0)
   declare (non-nulls-unique Symbol)
fi
```

```
if (field-exists "Period")
   declare (= (type Period) "int")
   declare (>= (min Period) 1)
   declare (<= (max Period) 7)
   declare (= (countnull Period) 0)
fi

if (field-exists "Group")
   declare (= (type Group) "int")
   declare (>= (min Group) 1)
   declare (<= (max Group) 18)
fi

if (field-exists "ChemicalSeries")
   declare (= (type ChemicalSeries) "string")
   declare (>= (min (length ChemicalSeries)) 7)
   declare (<= (min (length ChemicalSeries)) 20)
   declare (= (countnull ChemicalSeries) 0)
   declare (= (countzero (or (isnull ChemicalSeries) (in ChemicalSeries (list
"Actinoid" "Alkali metal" "Alkaline earth metal" "Halogen" "Lanthanoid"
"Metalloid" "Noble gas" "Nonmetal" "Poor metal" "Transition metal")))) 0)
fi
```

Miro Log 001 (2016/09/14) dali

001 +1 003 004

Individual Field Constraints

Name	Type	Min	Max	Nulls	Sign	Unique	Values
Z	**int**	1	118	no nulls	pos	yes: 118 / 118 unique (100.00%)	
Name	**string**	length 3	length 13	no nulls		yes: 118 / 118 unique (100.00%)	
Symbol	**string**	length 1	length 3	no nulls		yes: 118 / 118 unique (100.00%)	
Period	**int**	1	7	no nulls	pos	no: 7 / 118 unique (5.93%)	
Group	**int**	1	18	28 nulls (23.73%)	pos	no: 18 / 90 unique (20.00%)	
ChemicalSeries	**string**	length 7	length 20	no nulls		no: 10 / 118 unique (8.47%)	"Actinoid" "Alkali metal" "Alkaline earth metal" "Halogen" "Lanthanoid" "Metalloid" "Noble gas" "Nonmetal" "Poor metal" "Transition metal"
AtomicWeight	**real**	1.007946	294.0	1 null (0.85%)	pos		
Etymology	**string**	length 4	length 53	1 null (0.85%)		no: 114 / 117 unique (97.44%)	
RelativeAtomicMass	**real**	1.007946	294.0	1 null (0.85%)	pos		
MeltingPointC	**real**	-258.975000	3675.0	20 nulls (16.95%)			
MeltingPointKelvin	**real**	14.200000	3948.0	20 nulls (16.95%)	pos		
BoilingPointC	**real**	-268.930000	5596.0	20 nulls (16.95%)			
BoilingPointF	**real**	-452.070000	10105.0	20 nulls (16.95%)			
Density	**real**	0.000089	41.0	5 nulls (4.24%)	pos		
Description	**string**	length 1	length 83	66 nulls (55.93%)		no: 36 / 52 unique (69.23%)	
Colour	**string**	length 4	length 80	85 nulls (72.03%)		no: 30 / 33 unique (90.91%)	

ABSENT CONSTRAINTS

Gregory (Scotland Yard detective): *"Is there any other point to which you would wish to draw my attention?"*
Holmes: *"To the curious incident of the dog in the night-time."*
Gregory: *"The dog did nothing in the night-time."*
Holmes: *"That was the curious incident."*

— *Silver Blaze*, in *Memoirs of Sherlock Holmes*
Arthur Conan Doyle, 1892.

njr@StochasticSolutions.com

http://tdda.info

http://github.com/tdda

@tdda0 @njr0

Correct interpretation: Zero

Error of interpretation: Letter "Oh"