# COMP 499 Introduction to Data Analytics

## Lecture 2 — Data Wrangling

Greg Butler

Data Science Research Centre

and

Centre for Structural and Functional Genomics

and

Computer Science and Software Engineering
Concordia University, Montreal, Canada

gregb@cs.concordia.ca

# Overview of Lecture

1. Data Wrangling Overview
   - Discover
   - Structure
   - Cleanse
   - Enrich
   - Validate
   - Publish
2. Context
   - Measurement Scales
   - Normalization
   - Accuracy & Precision
   - Significant Digits
3. Data Cleaning — Professor Skiena lecture

# Data Wrangling — Discovery

Discover what data is available

Extract step of ETL

# Data Wrangling — Structure

Organize data into suitable format

Transform step of ETL

# Data Wrangling — Cleanse

Clean the data

Iterative step with basic data analysis

# Data Wrangling — Enrich

Discover and include related data

Integrate new data sets and data types

add more data fields

# Data Wrangling — Validate

## Check data is consistent and complete

### Consistency

Does your data fit into expected values for it?
Do field values match the data type for the column?
Are values within acceptable ranges?
Are rows unique? Duplicated?

### Completeness

Are all expected values included in your data?
Are some fields missing values?
Are there expected values that are not present in the dataset?

### Test routines for your data wrangling process

# Data Wrangling — Publish

Make available for analysis

Load step of ETL

into data warehouse in traditional business intelligence setting

# Context — Measurement Scales

## Nominal aka Categorical

Values have *names* as in enum or scalar type
equality testing allowed
mode is measure of central tendency

## Ordinal

Ranked values, such as *good, better, best*
equality and comparison allowed
median is measure of central tendency
mean and deviation do not make sense

## Interval

Difference between values can be determined, eg integers
equality, comparison, $+$, $-$ allowed
mean is measure of central tendency; deviation makes sense

## Ratio

Value is a ratio of continuous values, eg real number
also $\times$, $/$ allowed
geometric mean is measure of central tendency

# Context — Normalization

## A normal form ...

is a unique representation for an entity

## Examples

a string *" the Happiest day of My Life "*
to all lower case
and without leading or trailing blanks
and only one blank between words
*"the happiest day of my life"*

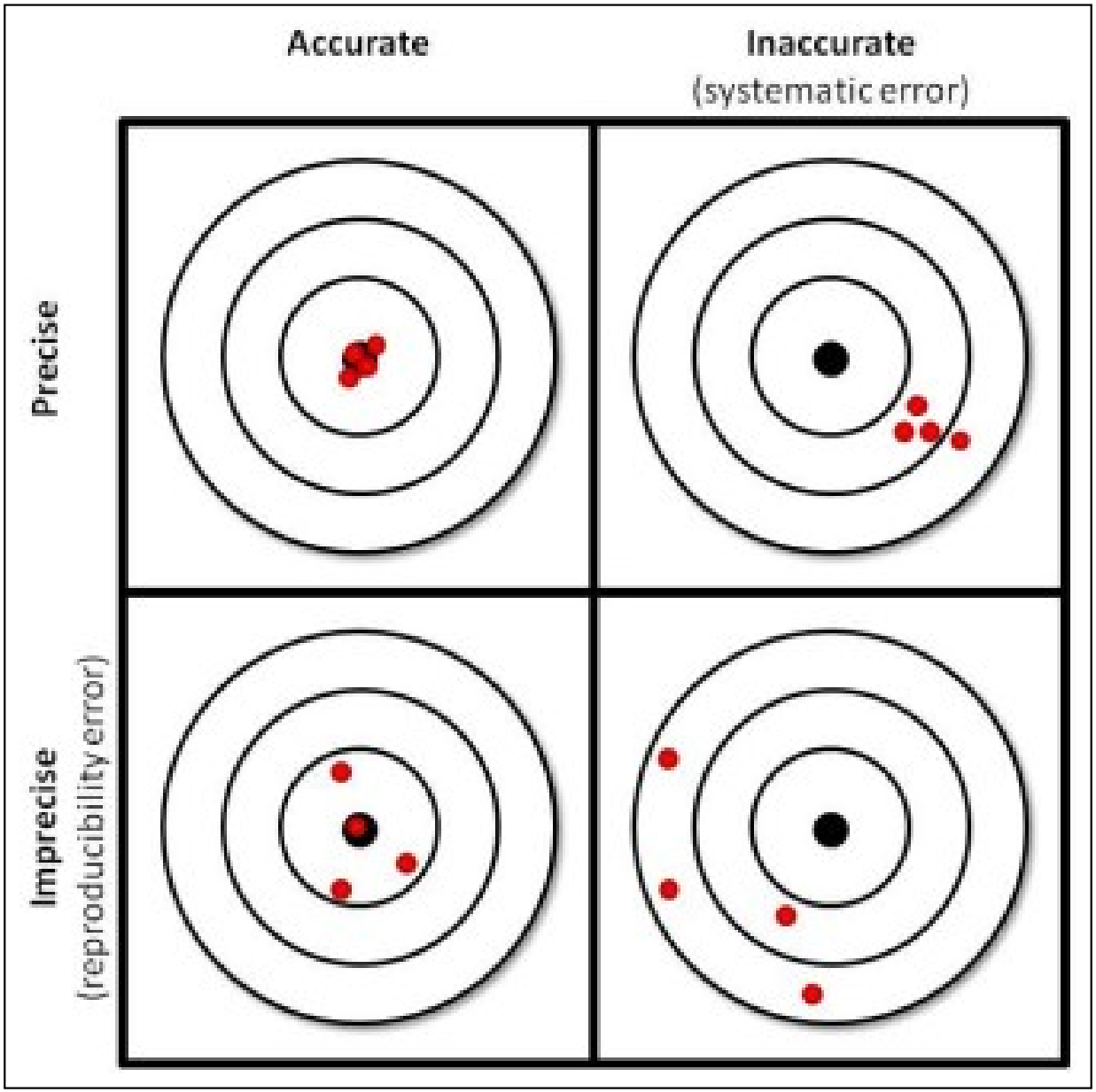
## Normalization creates a normal form

allows simple test for equality

## More Examples

Names
Dates
Currency
Metric vs Imperial measurements

# Context — Accuracy and Precision




http://www.geographer-miller.com/accuracy-vs-precision/

# Context — Significant Digits

## Problem

Showing more digits in a number than are meaningful

Especially in decimal component

## Examples

0.046 has two significant digits

4009 kg has four significant digits

7.90 has three significant digits

8200 has 2, 3, or 4 significant digits (**unclear**)

$8.200 \times 10^3$ has four significant digits

$8.20 \times 10^3$ has three significant digits

$8.2 \times 10^3$ has two significant digits

## Problem

Need to know significant digits for input data

Need to keep track of sig. digits in arithmetic

Be careful formatting output

## Reference

`https://www.physics.uoguelph.ca/tutorials/sig_fig/`