MANGO SOLUTIONS

The Curse of the Unspecified Data Definition

Chris Campbell
Senior Consultant
ccampbell@mango-solutions.com
@CSJCampbell

How to Play

- R & RStudio
- Clone

Who are Mango?

- Data Science
- Tool/Framework Development
- Consulting
- Training R & Python

Software Development

- Identify challenge
- Specify behaviours of solution
- Describe ways of working
- Design tool
- Build, test, deploy

ParanormalDatabase.com

DATA

Calendar

Recent Entries

Misc Reports

SOCIAL

Facebook

Twitter

OTHER

Gallery

Further

The Paranormal Databas

...is a serious ongoing project to quantitatively document as many locations with paranormal / cryptozoological interest as possible, region by

Errors

- Protect
- Communicate

Errors

```
# provide informative message
stop(
  paste(
    "numeric input expected",
    "but x was",
    typeof(x)))
```

Data Definition

- Defined Inputs
- Handling Rules

Data Definition

Location	Type	Comments	Date	lon	lat

Data Definition

- Location: Description of Site
- Type: Sighting Category
- Comments: Description of Sighting
- Date: Date of Sighting
- lon: Site position East or West of the Greenwich Meridian
- lat: Site position North or South of the Equator

Types

- Storage Mode
  - Raw
  - Logical
  - Integer
  - Double
  - Complex
  - Character
  - List
- Class

Types

Location	Type	Comments	Date	lon	lat
Character	Factor	Character	Date	Numeric	Numeric

Missing Values

- Remove
- Impute
- Raise Error

Missing Values

Location	Type	Comments	Date	lon	lat
Character	Factor	Character	Date	Numeric	Numeric
Ignore NA +	Ignore NA +	Set blank ++	Ignore NA +	Remove NA *	Remove NA *

+ Ignore if missing

++ Set blank if missing

* Remove record if missing

$ Error if missing

Data Out of Range

- Business Rules
  - Treat as Missing
  - Correct
  - Update Specification

Data Out of Range

Location	Type	Comments	Date	lon	lat
Character	Factor	Character	Date	Numeric	Numeric
Ignore NA +	Ignore NA +	Set blank ++	Ignore NA +	Remove NA *	Remove NA *
0 ch	1 level	0 ch	1700-01-01	-180	-90
128 ch	1 level	2048 ch	2018-10-31	180	90
Truncate ff	Set NA £	Truncate ff	Set NA £	Remove #	Remove #

£ Missing if out of spec

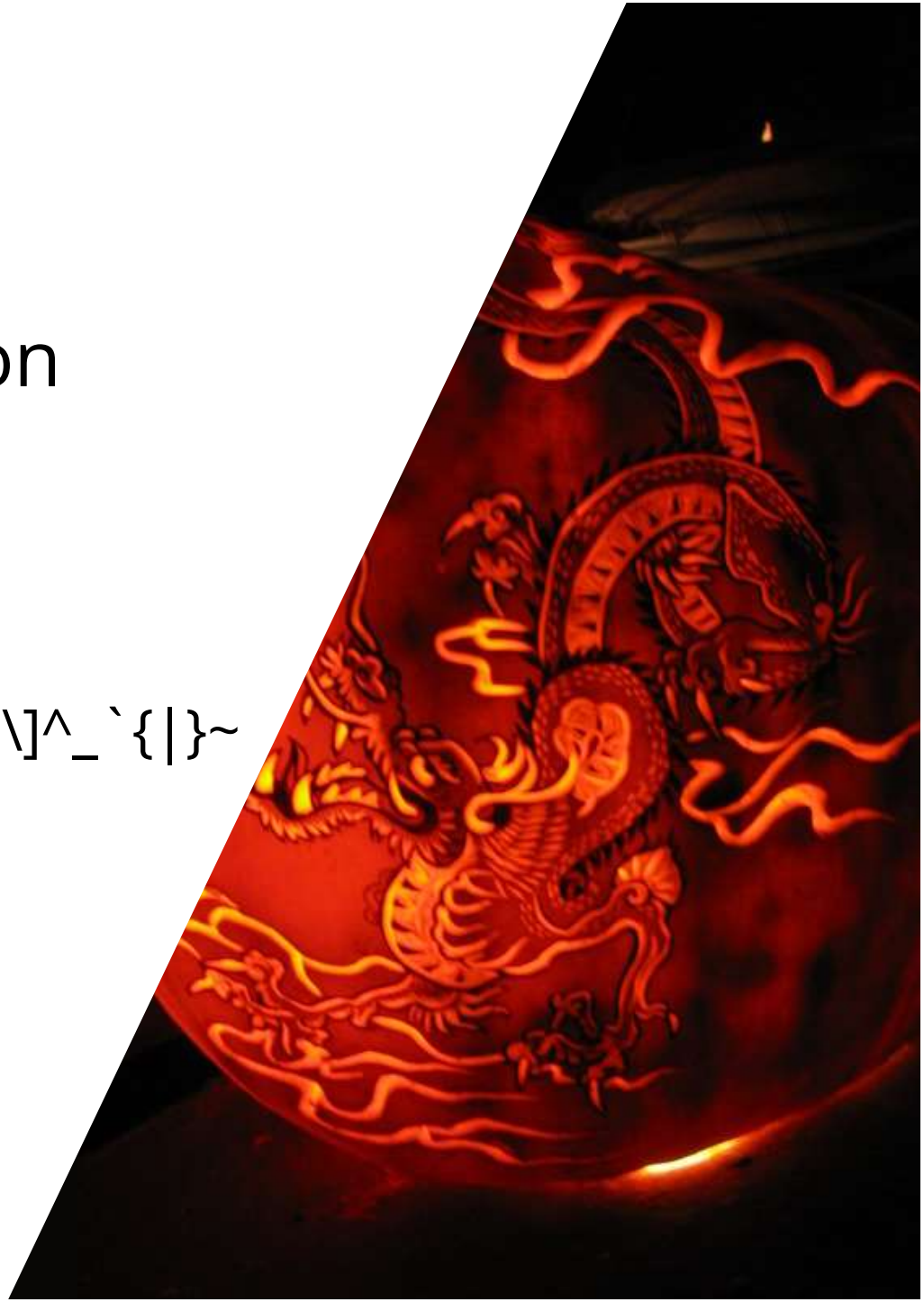
ff Truncate if out of spec

Remove record if out of spec

@ Error if out of spec

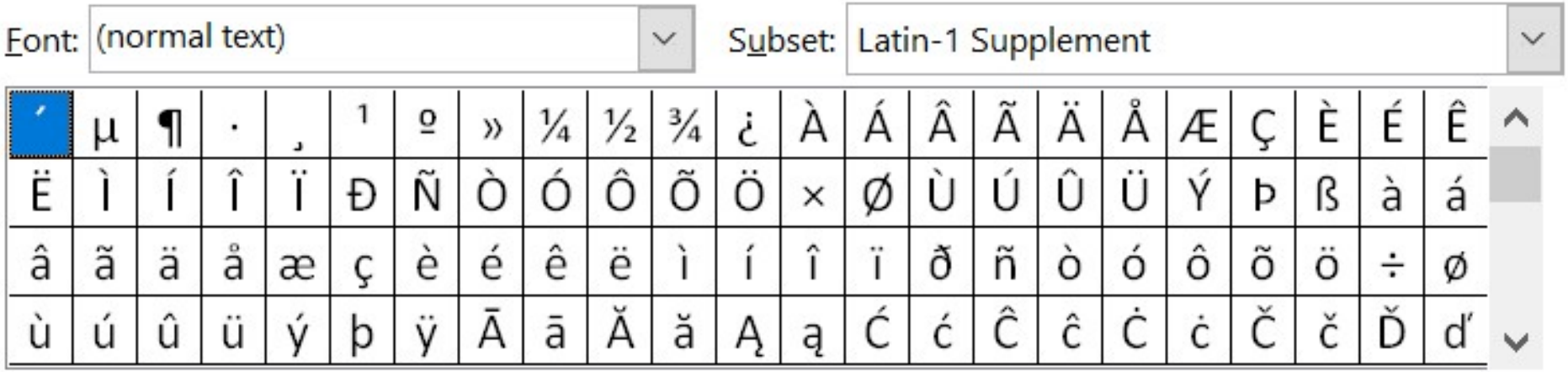
Encoding

- Internationalization
- ASCII
  - A-Za-z
  - 0-9
  - !"$%&'()*+,-./:;<=>?@[\]^_`{|}~



Encoding

- Latin-1 Supplement Unicode



Encoding

```
# non-ASCII encoding
x <- "Cardross - Laura's Café"
iconv(x,
   from = "latin1",
   to = "ASCII",
   sub = "byte")
# [1] "Cardross - Laura's Caf<e9>"
```

Encoding

Location	Type	Comments	Date	lon	lat
Character	Factor	Character	Date	Numeric	Numeric
Ignore NA +	Ignore NA +	Set blank ++	Ignore NA +	Remove NA *	Remove NA *
0 ch	1 level	0 ch	1700-01-01	-180	-90
128 ch	1 level	2048 ch	2018-10-31	180	90
Truncate ££	Set NA £	Truncate ££	Set NA £	Remove #	Remove #
ASCII	ASCII	ASCII			

Dates & Timestamps

- Diverse
- Ambiguous
- Timezones

Dates & Timestamps

```
# Date class
as.Date("2018-10-31")
[1] "2018-10-31"
# Days since 1970-01-01
as.Date(17835,
  origin = "1970-01-01")
[1] "2018-10-31"
```

Type:

14/03/2012
14/03/12
14.03.12
14-03-12
2012-03-14
mercredi 14 mars 2012
14 mars 12

Locale (location):

French (Luxembourg)

French (Haiti)
French (Luxembourg)
French (Mali)
French (Monaco)
French (Morocco)
French (Reunion)

time serial numbers as date values. Date formats that begin with an

Dates & Timestamps

Location	Type	Comments	Date	lon	lat
Character	Factor	Character	Date	Numeric	Numeric
Ignore NA +	Ignore NA +	Set blank ++	Ignore NA +	Remove NA *	Remove NA *
0 ch	1 level	0 ch	1700-01-01	-180	-90
128 ch	1 level	2048 ch	2018-10-31	180	90
Truncate ff	Set NA f	Truncate ff	Set NA f	Remove #	Remove #
ASCII	ASCII	ASCII	(%d) (%b) %Y %d %b $$		

$$ Record reoccurring sightings in 2016

Data Volume

- Time to respond
- Interpretability

Data Volume

Location	Type	Comments	Date	lon	lat
Character	Factor	Character	Date	Numeric	Numeric
Ignore NA +	Ignore NA +	Set blank ++	Ignore NA +	Remove NA *	Remove NA *
0 ch	1 level	0 ch	1700-01-01	-180	-90
128 ch	1 level	2048 ch	2018-10-31	180	90
Truncate ££	Set NA £	Truncate ££	Set NA £	Remove #	Remove #
ASCII	ASCII	ASCII	(%d) (%b) %Y %d %b		

Segment records so that no more than 200 are displayed simultaneously

Lift the Curse

- Define all inputs
- Define what happens when data is out of spec
- Implement