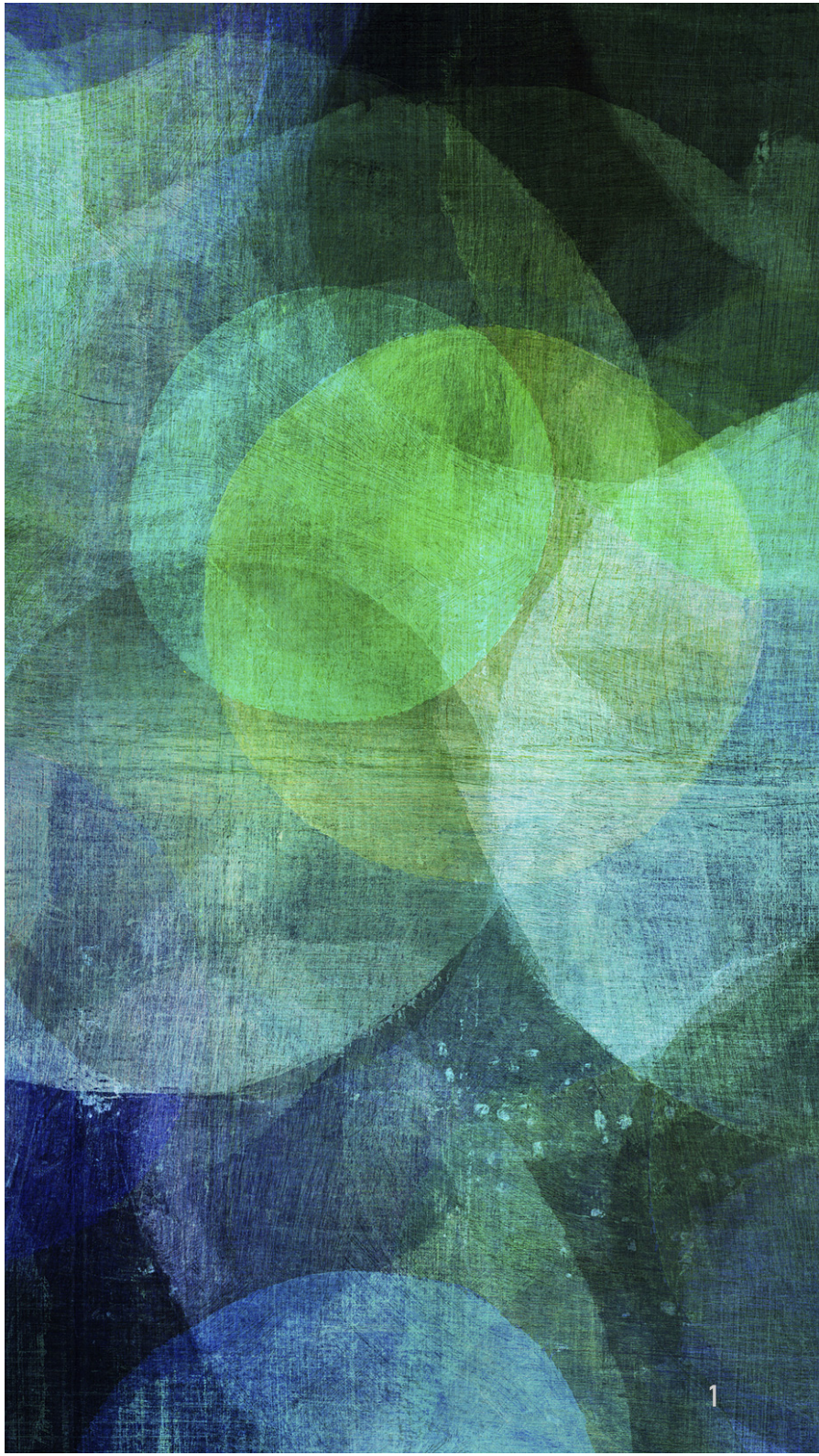
DATA AND "GOOD DATA"

avoid "good data" being misunderstood or creating false facts

G Benoît



YOUR QUESTIONS . . .

- What constitutes “good” data?
  - Requires looking at the data sources …
  - the “Quality” of the data versus the “goodness” of the data [values for computation]
  - Statistics - “data robustness”
    - Are these data suitable for this statistical test?

- Your data & visualizations are ethical issues …
  - But that’s also a social issue … ensuring truthfulness and trust in the data is a challenge
  - People rely on your data … your ethics … your visualizations!

Part of a data table …

Subject	height	weight	triclicerides
1	52cm	150k	3.25
2	55cm	172k	1.23
3	60cm	204k	5.23
4	52	150k	3.00

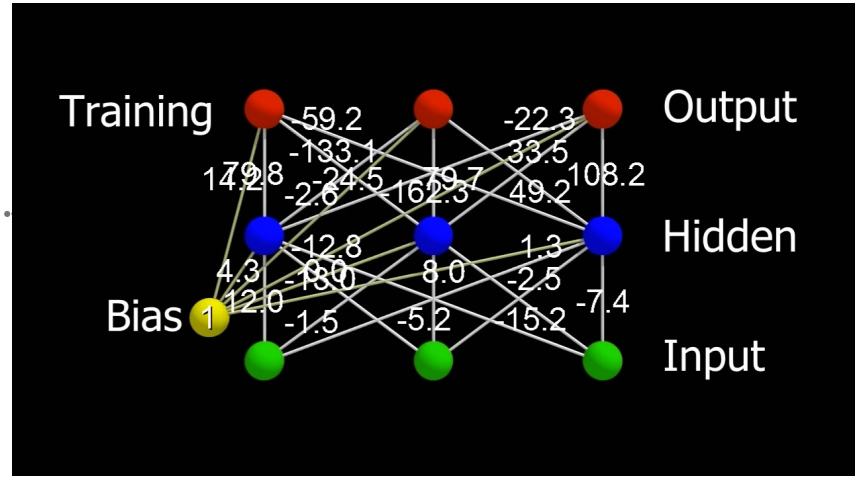
DATA

- In response to a student question (and a great one at that!) this week we're going to explore data in general …

- The content of this slide set ranges from general observations to statistical/math models to address the class's range of statistical and research backgrounds.

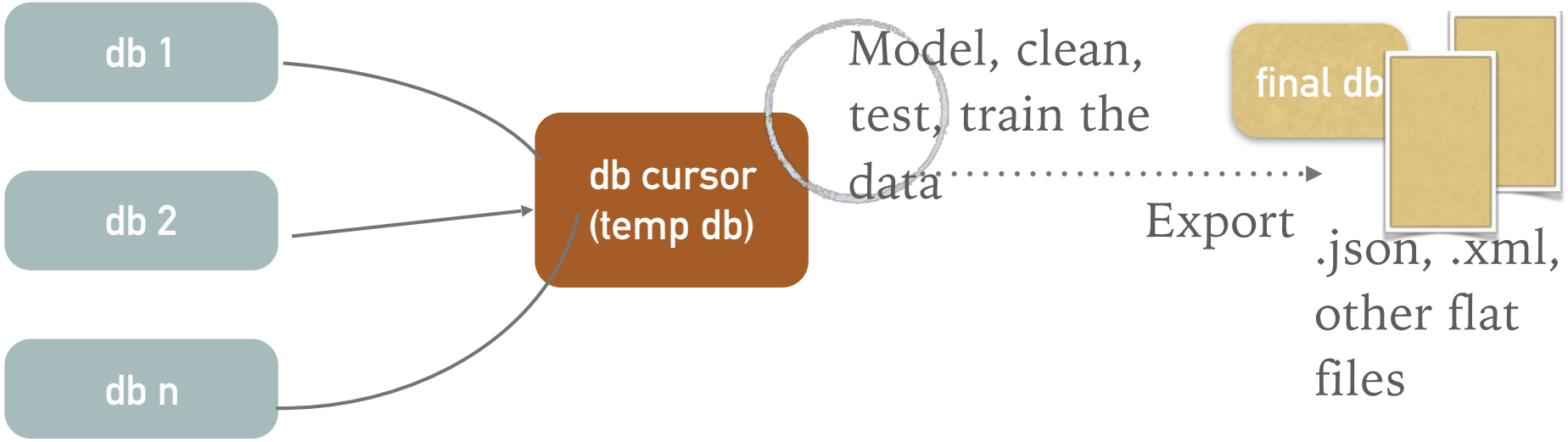
- Both art and math of data are in play …
  - and the people who use our visualizations

OUTLINE FOR THESE SLIDES



- Sources of data
- Issues: Statistics and data
- Data cleansing, warehousing, and mining
- Odds & ends about data…

1 SOURCES OF DATA . . . EXAMPLE: EXTRACTING FROM RDBMS



- Exported from single relational database table ("mytable") - E.g. of adding a bit of noise to the data to prevent multiplication by 0, "Not a Number" (NaN) error, or stopping code
  - `UPDATE TABLE mytable SET fieldname = 0.3 WHERE filename IS NULL`
- Exported and integrated from multiple databases

1 SOURCES OF DATA . . . EXAMPLE: USE OF THE EXPORTED DATA

- Integrated from flat-files
  - .json, .xml, .dat, .txt
    - Use extensible markup language & transformation (.xsl) files as a way of storing the manipulating the .xml file(s) created in response to a DB query
  - E.g., transaction log analysis
  - E.g., life & medical sciences
    - esp epidemiology
- What kinds of issues can we expect? - Let's view a few main points: nulls, cleaning techniques, threats to validity, and some advanced models [to give an overview of the many approaches]

2 DATA AND POTENTIAL ISSUES – EXAMPLE: DB TO SAS

- Data that cannot establish any kind of pattern is of primary concern …
- Computationally cleaning your data …
  - Example using SAS

```
PROF FREQ DATA=PATIENTS;
    TITLES “FREQUENCY COUNTS”;
    TABLES GENDER DX AE / NOCUM NOPERCENT;
RUN;
```

  - [No cumulative stats; no percentages] Identifies what’s missing

Cody, R., et al. Data Cleaning 101

2 DATA AND POTENTIAL ISSUES – DB DATA

- Preparation of data depends on the software and the model you select
- Common issues:
  - **Nominal** data: survey data collecting "M" or "F" and the end-users don't reply … options?
  - **Interval** Data: survey data collecting 0…5 (or any n) and the respondent tries to use "in-between" values (say 3.5)
    - Rounding up or down increases error.
  - **Interval** & **Ratio** Data out of range:
    - Input errors - check the domain and range of your data…

2 DATA AND POTENTIAL ISSUES – EXAMPLE: TABLEAU

- Tableau has tools for addressing null string literals and number type nulls
- http://kb.tableau.com/articles/knowledgebase/replacing-null-literalsclass
  - Check out the above site.

2 DATA AND POTENTIAL ISSUES

- Leads to “how to determine an actual outlier from a mistake?”
  - Test the “robustness” of the data and the parameters of the statistical test (e.g., critical values, chi-square tests)

- Some basic issues:
  - case: say the data are input in lower case but your code looks for upper case? [e.g., `if (temp == “S”) … ]`
    - How would you prep your data?
    - What if values are missing?
    - What if your visualization uses calculated data?
      - (hint: calculated data aren’t usually stored…)

3 PROBLEM EXAMPLE: MISSING VALUES

- loss of efficiency
- complications in handling & analyzing the data
- bias resulting in differences between missing & complete data
  - Examples: MCAR: Missing Completely at Random [non-uniform error]
  - MAR: Missing at Random [probability of a record having a missing value for an attribute could depend on the observed data, but not on the value of the missing data]
  - How might you address the many approaches to missing values?

London School of Hygiene & Tropic Medicine, 2013; Horton & Kleinman, 2007; Lakshminarayan et al. 1999; Horton & Kleinman, 2007); see especially Kantardzic, 2003

3 PROBLEM EXAMPLE: MISSING VALUES

- Reducing the data set [Kantardzic, 2003]
- Treating missing attribute values as special values [Grymala-Busse, 2001)] - for values that aren't expected to influence future analyses
- Replacing missing value with mean [most common approach]
  - Replacing missing value with (a) mean or (b) median for the given class
- Replacing missing values with most common attribute value
  - (a most common value is inserted)
- Closest Fit …

3 MISSING VALUES

- Imputation using k-nearest neighbor
  - (replacing the null with the closed data set points to y)
  - (Hand, Mannil, & Smyth, 2001)

- Closest Fit:

$$dist(x,y) = \left\{ \sum_{i=1}^{n} dist(x_i, y_i) \right.$$

$$\text{distance}(x_i, y_i) = \begin{cases} 0 \text{ if } x_i = y_i, \\ 1 \text{ if x and y are symbolic and } x_i \neq y_i \text{ or } x_i = ? \text{ or } y_i = ? \\ \dfrac{|x_i - y_i|}{r} \text{ if } x_i \text{ and } y_i \text{ are numbers and } x_i \neq y_i \end{cases}$$

Point: replace missing value in column with something.

(x,y) refer to x,y coordinates in an exported table

3 MISSING VALUES IMPUTATION WITH NEURAL NETWORKS

Multilayered networks & back propagation algorithm:

1: Collect all training cases without any missing value and call them the complete set.

2: Collect all training and test cases with at least one missing value and call them the incomplete set.

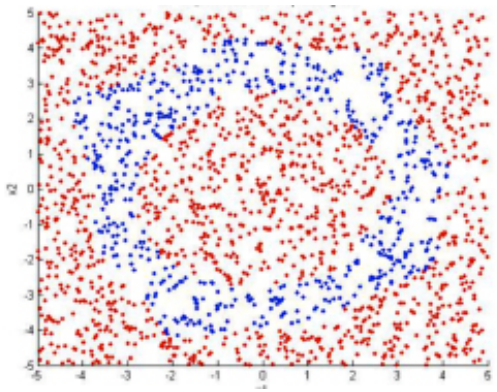
3: For each pattern of missing values, construct a multi-layered network with the number of input nodes in the input layer equal to the number of non-missing attributes, and the number of output nodes in the output layer equal to the number of missing attributes. Each input node is used to accept one non-missing attribute, and each output node to represent one missing attribute.

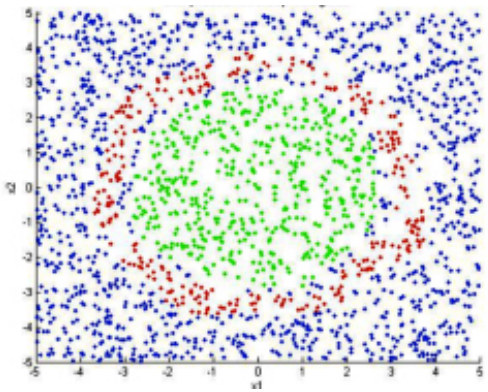
4: Use the complete set and the backpropagation algorithm to train each network constructed in step 3. Since the complete set does not have missing values, different patterns of input-output pairs can be obtained from the complete set to satisfy the input-output requirements for different networks from step 3. As the output of a network is between 0 and 1, data have to be converted to values between 0 and 1 for this reconstruction procedure.

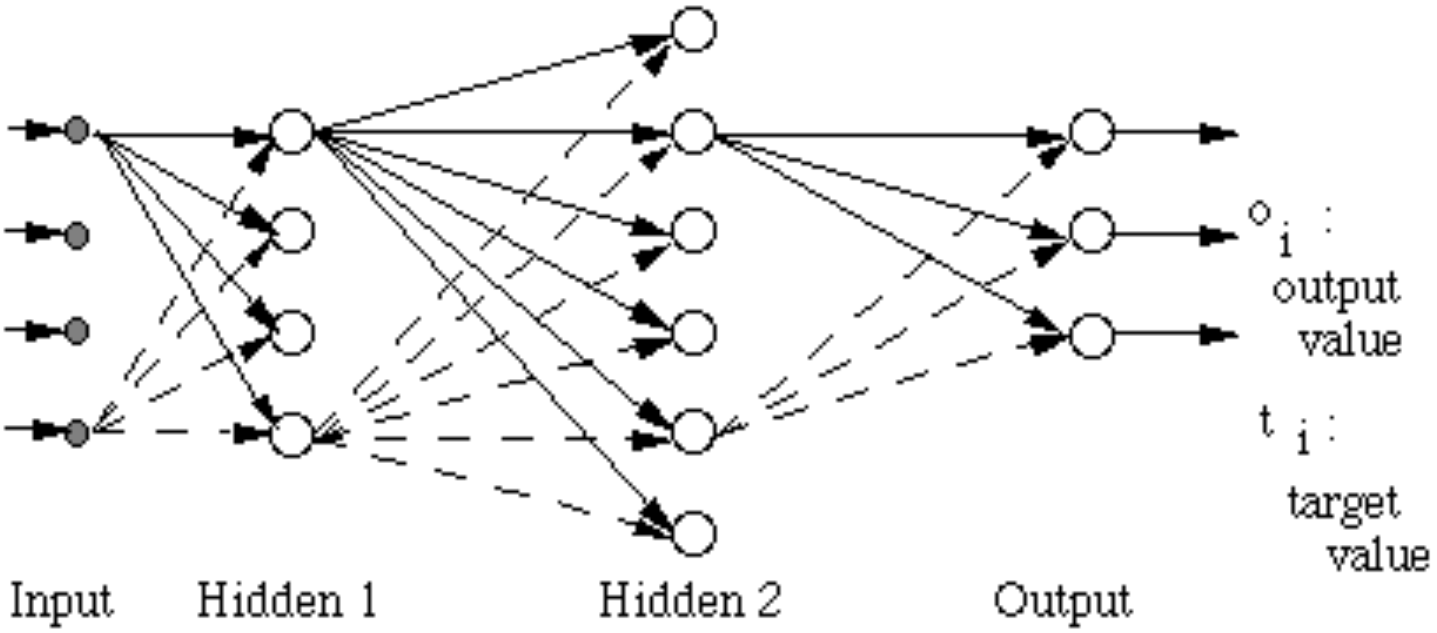
Chen, Z., 2001

5: Use the trained networks from step 4 to calculate the missing values in the incomplete set.

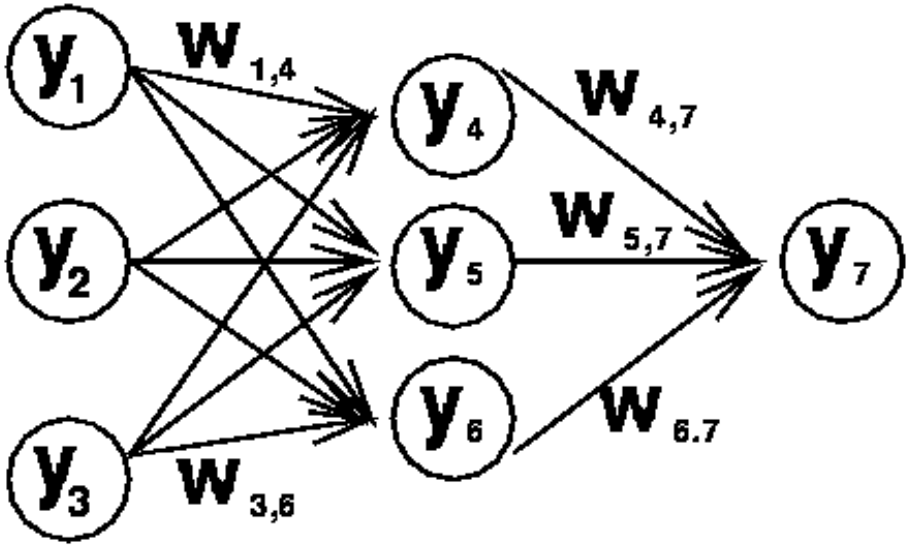
http://www.gierad.com/projects/neural-networks-in-action/







Calculating the feed-forward, back-propagation values



Big View: inputs (y); weights (w) and steps (arrows). The values calculated using a subset of the data (a "training set") and then used to fill missing values in the big data set.

http://www.cs.ru.nl/~ths/rt2/col/h10/10neurENG.html

$$\frac{\partial E_p}{\partial Y_{ji}} = \frac{\partial E_p}{\partial net_{(j+1)1}} \frac{\partial net_{(j+1)1}}{\partial Y_{ji}} + \frac{\partial E_p}{\partial net_{(j+1)2}} \frac{\partial net_{(j+1)2}}{\partial Y_{ji}} + \ldots \quad (2.10)$$

$$= \sum_{a=1}^{N_{j+1}} \left[\frac{\partial E_p}{\partial net_{(j+1)a}} \frac{\partial net_{(j+1)a}}{\partial Y_{ji}} \right] (2.11)$$

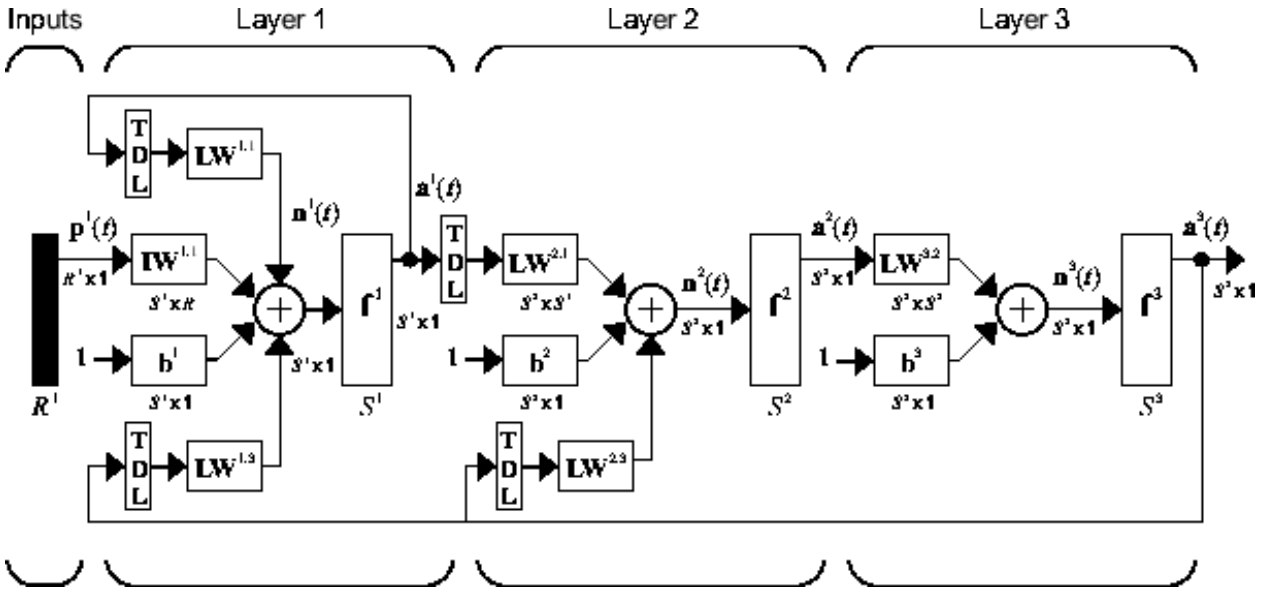
$$= \sum_{a=1}^{N_{j+1}} \left[-\delta_{(j+1)a} \frac{\partial}{\partial Y_{ji}} (W_{(j+1)a0} Y_{j0} + \ldots + W_{(j+1)ai} Y_{ji} + \ldots) \right] (2.12)$$

$$= \sum_{a=1}^{N_{j+1}} \left[-\delta_{(j+1)a} \frac{\partial}{\partial Y_{ji}} (W_{(j+1)ai} Y_{ji}) \right] (2.13)$$

$$= \sum_{a=1}^{N_{j+1}} \left[-\delta_{(j+1)a} W_{(j+1)ai} \right] . (2.14)$$

Just fyi: a model and formula … to suggest the complexity of it all…

Calculating the feed-forward, back-propagation values



3 MISSING VALUES IMPUTATION USING ASSOCIATION RULES

- An association rules is a probabilisitic statement about the co-occurrence of certain events - e.g., a binary variables rule (Hand et al., 2001) `if A = 1 AND B = 1 AND C = 1,`
- where A, B, C, are variables, then
  - `p = p(C=1 | C = 1 | B = 1)`
  - (a conditional probability) …

3 OTHER APPROACHES

- Hidden Markov Models
- Genetic Algorithms
- Replace missing values by the average of all available data of that variable in the training set [Gupta, 1996]
- Fuzzy sets [Pay-del-Castillo, 2009]

- SUMMARY : No null values
  - If "stuffed" with temp values, (e.g., 0.3), then we know the error in the data (which can be accounted for)
  - AND we're assured of no multiplication by 0 or by null {which really nixes our data!}

4 PROBLEM EXAMPLE: CHARS & STRINGS

- Case… use your preferred programming language's promotion to UC or to LC
  - e.g., var s = "fishlips";
  - e.g., `String s = s.toUpperCase();`
  - Make sure your program or system isn't case-sensitive (unless that's meaningful to your data)
  - One option is to update all the data to upper case when ingesting the data; another option is to use the toUpperCase() option in the program.
  - Mixed languages and encodings of integrated data sets may be a problem, too. [E.g., mixing Japanese Industrial Standard (JIS), KOI-8 (Russian), Big5, TwinBridge (Chinese), Latin-1 (an ISO standard), IIS (Indian Industrial Standard), old 7-bit ASCII, UTF-8, etc.]

5 PROBLEM EXAMPLE: NUMBERS

- Mixing and matching integers, long, double, float
  - E.g., comparing integer (say 50) with double (e.g., 48.39283948938291).
    - Risk Loss of Precision (the double turned into an integer)
      - Read more: https://docs.oracle.com/javase/tutorial/java/data/converting.html

6 PROBLEM AREA: MODELS

Man prefers to believe what he prefers to be true – Francis Bacon, 1561-1626

- Threats to Validity: the extracted data for visualization is not unlike a research project - so the model you select is subject tot he same questions of a research project:
- Are the conclusions correct?
- Are the changes in the independent variable indeed responsible for the variation in the dependent var?
- Are there hidden values (confounding values)?
- Is there strong evidence of causality in the data and in the visualization?
- What do you know about the sources of data? Are there "internal threats to validity"? (bad design, bias); "external threats"? (the data don't reflect the population being studied? generalizing to wrong population of interest?)

SUMMARY

- Actual data in files: values within range; no nulls
- Data Cleansing:
  - Convert data types to same type by field
  - Address missing values, case issues
  - Address data that cannot be compared to others
  - Statistical outliers vs. mistakes
  - Strings, characters and case

SUMMARY

- Quality data versus Good data
  - Are the data relevant to the phenomenon of interest?
    - [supposedly > 1/3 of the data collected are useless in application, costing $3.3 trillion/year (Vanson Bourne)]
    - There are many techniques for …
      - "cleaning the data"
      - checking the quality of the data [appropriateness for what's being studied]
      - reviewing the interpretation of the message from the visualization.
  - Do the data conform to research practices/validity?

http://www.cioinsight.com/it-strategy/big-data/why-separating-good-data-from-bad-is-a-challenge.html

In light of “good”/“quality” data - Wrap up with some examples and tutorials…

“What is good data” times from Tableau:

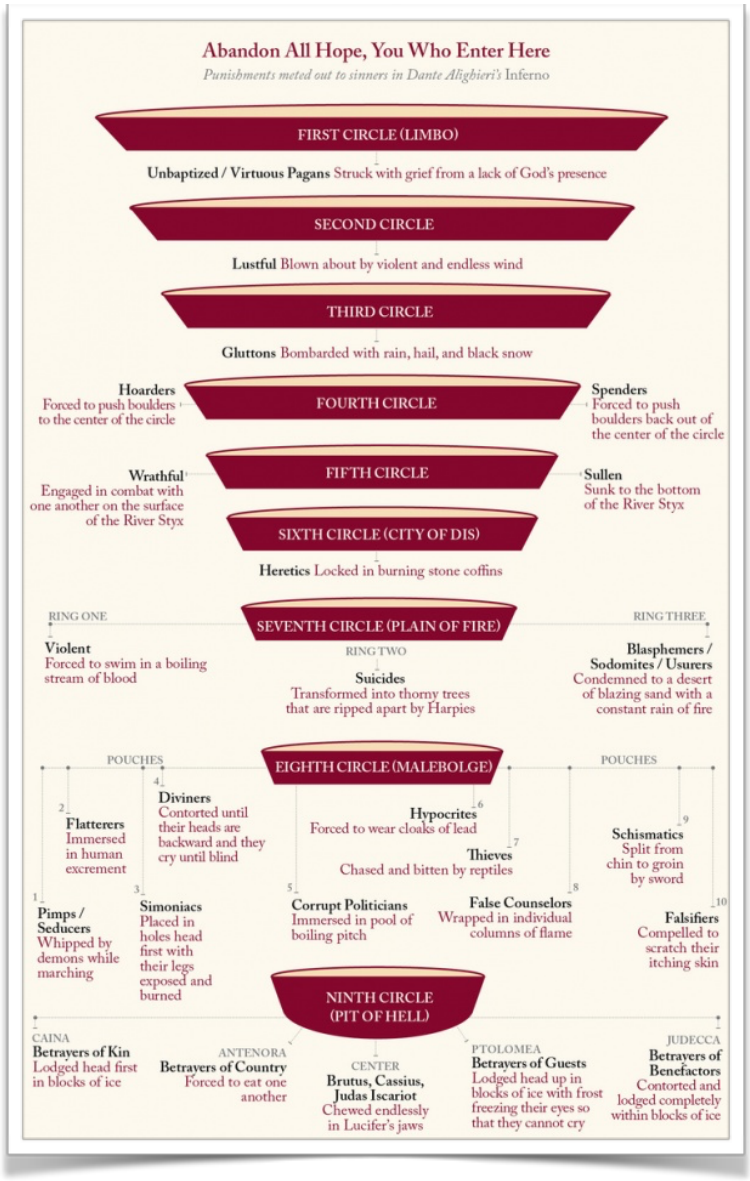
www.tableau.com/asset/10-tips-to-create-useful-beautiful-visualizations

Poor visualizations: http://viz.wtf Ask yourself (a) why the visualization is considered poor *and* (b) what kind of data do you think was used to create this visualization?

A collection of links http://cs.colby.edu/courses/S14/cs251/goodbad.php

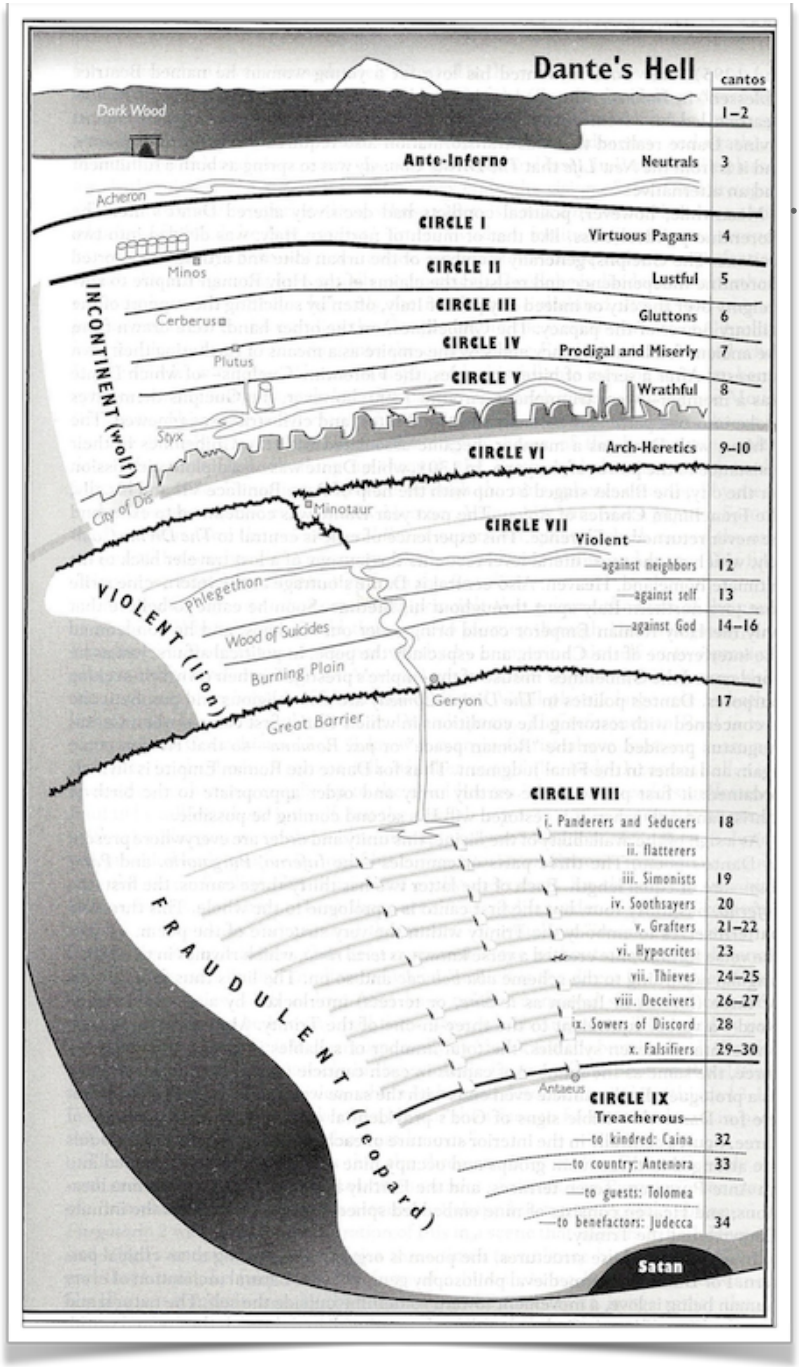
(kind of fun to review the myriad perspectives)

A bit of fun: visualizing Dante



This and the next slide demonstrate three very different ways of expressing the same concepts…

What are the data behind the graphics?


Dante's Hell
cantos
Dark Wood
1–2
Ante-Inferno
Neutrals
3
Acheron
CIRCLE I
Virtuous Pagans
4
Minos
CIRCLE II
Lustful
5
CIRCLE III
Gluttons
6
Cerberus
CIRCLE IV
Prodigal and Miserly
7
Plutus
CIRCLE V
Wrathful
8
INCONTINENT (wolf)
Styx
CIRCLE VI
Arch-Heretics
9–10
City of Dis
Minotaur
CIRCLE VII
Violent
—against neighbors
12
—against self
13
—against God
14–16
Phlegethon
Wood of Suicides
VIOLENT (lion)
Burning Plain
Great Barrier
Geryon
17
CIRCLE VIII
i. Panderers and Seducers
18
ii. Flatterers
iii. Simonists
19
iv. Soothsayers
20
v. Grafters
21–22
vi. Hypocrites
23
vii. Thieves
24–25
viii. Deceivers
26–27
ix. Sowers of Discord
28
x. Falsifiers
29–30
FRAUDULENT (leopard)
Antaeus
CIRCLE IX
Treacherous
—to kindred: Caina
32
—to country: Antenora
33
—to guests: Tolomea
—to benefactors: Judecca
34
Satan


INFERNO CANTO. XXIII.
Ei ne uerranno dietro piu crudeli,
Chel cane a quella leure, chegli acceſſa.
Gia mi ſentia tutti arricciar li peli
De la paura; e ſtaua indietro intento;
Quando diſſi; Maeſtro ſe non celi
Te e me toſtamente, i ho pauento
Di malebranche: noi glihauem gia dietro:
Io glimmagino ſi; che gia li ſento.
mamente poi come ſeron ad uſcirneſu ori.
Taciti, ſoli, e ſenza compagnia, Partiti da dieci demoni, nandaron taciti e ſoli e ſenza la compagnia di quelli, lun dinanzi e laltro dopo, come uanno per uia fra ti minori, Andaua adunque Virg. inanzi e Dante dopo lui, per la ragione che habbiamo gia piu uolte detto. VOlto era in ſu la fauola d'Iſopo, Iſopo fu poeta Greco, e ſcriſſe fauole, che hanno in ſe moralita, e fra quelle una dun topo, che giunto ad un feſſo pieno dacqua per paſſarlo, e temendo dannegarſi, una rana ſe gliofferſe di paſſarlo ſeluo con propoſito però di ſummergerlo, e perche meglio le riuſciſſe il fraudolente penſiero, ſe lo legò ſu la ſchiena, ma

REFERENCES (OTHER THAN IN THE SLIDE SET)

- Nelwamondo, F., et al., (2006). Missing data … https://arxiv.org/pdf/0704.3474.pdf
- Junninen, H., et al., (2004). Methods of imputation of missing values in air quality data sets. *Atmospheric environment, 38*(8), 2895-2907.
- Mussa, A. (2005). Use of genetic algorithms and neural networks to approximate missing data in databases. *Computational cybernetics*, DOI 10.1109/ICCCYB.2005.1511574
- Jenez, J, M., (2010, Oct.). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial intelligence in medicine, 50*(2), 105,115.

Other great sites to review as examples …

- http://www.designyourway.net/blog/inspiration/when-infographics-go-bad-or-how-not-to-design-data-visualization/
- http://www.dummies.com/programming/big-data/big-data-visualization/data-visualization-examples-of-the-good-and-the-bad/
- https://eagereyes.org/blog/2008/ny-times-the-best-and-worst-of-data-visualization
- http://www.scribblelive.com/blog/2014/05/12/data-visualization-charts-form-the-u-s-congress-floor-the-good-the-bad-and-the-ugly/
- https://curriculum.code.org/csp/unit2/10/
- http://www.kdnuggets.com/2014/07/spotting-bad-data-visualizations.html [In a ridiculously poorly designed website!]