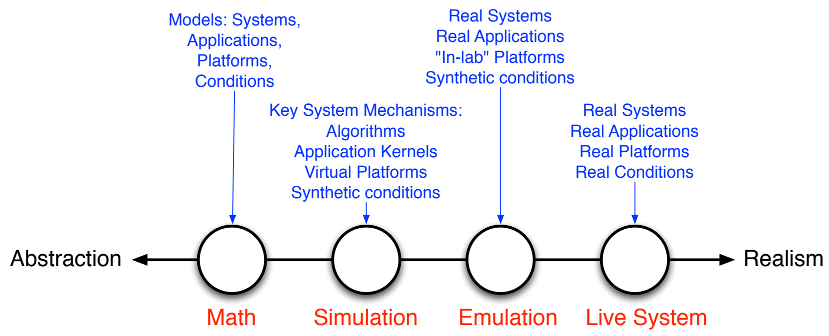
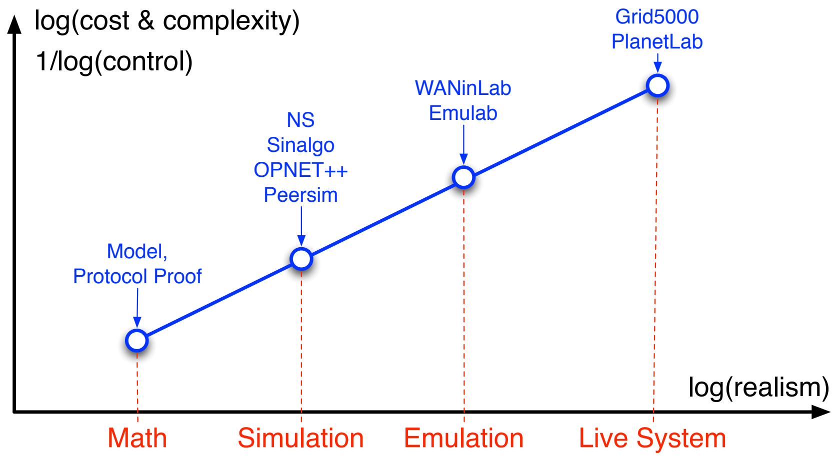
Making the Point

Sébastien Tixeuil
Univ. Pierre & Marie Curie - Paris 6
Sebastien.Tixeuil@lip6.fr

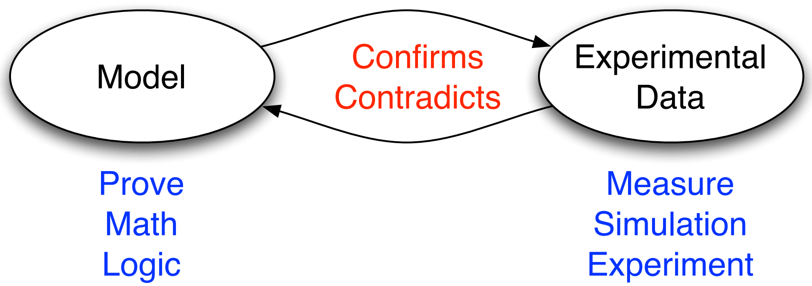
Studying Networks



Studying Networks



Agenda



Agenda

- **Writing Proofs**
- **Managing Experimental Data**
  - Classical *vs.* Exploratory
  - Practicalities

How to Write a Proof

How to write proofs: a quick guide. Eugenia Cheng.
http://www.math.uchcago.edu/~eugenia

Proof

- **Begining**: things we assume to be true, including the definitions of the things we talk about
- **Middle**: statements, each following logically from the things before it
- **End**: the thing we're trying to prove

How to write proofs: a quick guide. Eugenia Cheng.
http://www.math.uchcago.edu/~eugenia

Kinds of things to try and prove

$x = y$

$x \Rightarrow y$

$x \iff y$

x is purple

$\forall x, p(x)$ is true

$\exists x$ such that $p(x)$ is true

How to write proofs: a quick guide. Eugenia Cheng.
http://www.math.uchcago.edu/~eugenia

Example 1. *Using the field axioms, prove that $a(b-c) = ab - ac$ for any real numbers a, b, c. You may use the fact that $x.0 = 0$ for any real number x.*

BEGINNING field axioms
definition $x - y = x + (-y)$
given $x.0 = 0$

MIDDLE

$$\begin{aligned} a(b-c) &= a(b+(-c)) && \text{definition} \\ &= ab + a(-c) && \text{distributive law} \\ \\ ac + a(-c) &= a(c+(-c)) && \text{distributive law} \\ &= a.0 && \text{additive inverse} \\ &= 0 && \text{given} \\ \therefore\ a(-c) &= -(ac) && \text{definition of additive inverse} \\ \\ \therefore\ ab + a(-c) &= ab - ac \end{aligned}$$

END $\therefore$ by line 2, $a(b-c) = ab - ac$ as required $\square$

Example 2. *Let f and g be functions $A \xrightarrow{f} B \xrightarrow{g} C$. Show that if f and g are injective then $g \circ f$ is injective*

BEGINNING definition of injective
definition $(g \circ f)(a) = g(f(a))$
assumption that f and g are injective i.e.
$\forall a, a' \in A \quad f(a) = f(a') \Longrightarrow a = a'$
$\forall b, b' \in B \quad g(b) = g(b') \Longrightarrow b = b'$

MIDDLE

$$\begin{aligned} (g \circ f)(a) = (g \circ f)(a') &\Longrightarrow g(f(a)) = g(f(a')) && \text{by definition} \\ &\Longrightarrow f(a) = f(a') && \text{since g is injective} \\ &\Longrightarrow a = a' && \text{since f is injective} \\ \\ \therefore (g \circ f)(a) = (g \circ f)(a') &\Longrightarrow a = a' \end{aligned}$$

END i.e. $g \circ f$ is injective, as required $\square$

Example 3. *Prove by induction that $\forall n \in \mathbb{N}, 1 + \cdots + n = \dfrac{n(n+1)}{2}$*

BEGINNING Principle of Induction

MIDDLE

$$\begin{aligned} \text{for } n = 1, \text{LHS} &= 1 \\ \text{RHS} &= \frac{1(1+1)}{2} \\ &= 1 \end{aligned}$$

$\therefore$ result is true for $n = 1$

If result is true for $n = k$ then

$$\begin{aligned} 1 + \cdots + k + (k+1) &= \frac{k(k+1)}{2} + (k+1) \\ &= \frac{k(k+1) + 2(k+1)}{2} \\ &= \frac{(k+1)(k+2)}{2} && \text{i.e. result true for } n = k+1 \end{aligned}$$

$\therefore$ result true for $k \Longrightarrow$ result true for $k+1$

END $\therefore$ by the Principle of Induction, the result is true for all $n \in \mathbb{N}$ $\square$

Traps and Pitfalls

What is Wrong ?

$$\begin{aligned} a(b-c) &= ab-ac \\ ab+a(-c) &= ab-ac \\ a(-c) &= -ac \\ ac+a(-c) &= 0 \\ a(c+(-c)) &= 0 \\ a.0 &= 0 \\ 0 &= 0 \qquad \square \end{aligned}$$

What is Wrong ?

$$\begin{aligned} a(b-c) &= ab+a(-c) \\ &= ab-ac \qquad \square \end{aligned}$$

What is Wrong ?

$$\begin{aligned} a(b-c) &= ab+a(-c) \\ &= ab+a(-c)+a.1 \\ &= ab+a(1-c) \\ &= ab-ac \qquad \square \end{aligned}$$

What is Wrong ?

$$\begin{aligned} a(b-c) &= ab+a(-c) \\ a(-c) &= -ac \quad \text{because if you add ac to both sides then both sides vanish which means they're inverse} \\ \therefore\ ab+a(-c) &= ab-ac \qquad \square \end{aligned}$$

Beware Incorrect Logic

- Negating a statement incorrectly
- proving the converse of something instead of the thing itself

$$\forall \varepsilon > 0 \ \exists \delta > 0 \text{ s.t. } \forall x \text{ satisfying } 0 < |x - a| < \delta, \quad |f(x) - l| < \varepsilon$$

$$\exists \varepsilon > 0 \text{ s.t. } \forall \delta > 0 \ \ \exists x \text{ satisfying } 0 < |x - a| < \delta \ \text{ s.t. } \ |f(x) - l| \geq \varepsilon$$

Additional Pitfalls

- Incorrect assumptions
- Incorrect use of definitions, or use of incorrect definitions

$$\begin{aligned} \mathsf{f(a) = f(a')} &\implies \mathsf{a = a'} \\ \mathsf{g(a) = g(a')} &\implies \mathsf{a = a'} \end{aligned}$$

$$\begin{aligned} \mathsf{(g \circ f)(a) = (g \circ f)(a')} &\implies \mathsf{g(a) \circ f(a) = g(a') \circ f(a')} \\ &\implies \mathsf{a = a'} \end{aligned}$$

$\therefore$ g $\circ$ f is injective. $\square$

Assumptions

- You need to *justify* everything *enough* for your *peers* to understand it
- If in doubt, *justify* things *more* rather than less

Practicalities

Practicalities

- Write the **begining** very carefully
- Write the **end** very carefully
- Try and manipulate both ends to meet in the middle, from *big* leaps to *smaller* ones
- Pretend to be more *stupid* (or *sceptical*, or *untrusting*) that you are

$x = y$ or "something equals something else"

$$\begin{aligned} x &= a \\ &= b \\ &= c \\ &= d \\ &= y \end{aligned}$$

$$\begin{aligned} x &= a \\ &= b \\ &= c \\ \\ y &= e \\ &= d \\ &= c \\ \\ \therefore x &= y \end{aligned}$$

$x \Longrightarrow y$

$$\begin{aligned} x &\Longrightarrow a \\ &\Longrightarrow b \\ &\Longrightarrow c \\ &\Longrightarrow d \\ &\Longrightarrow y \end{aligned}$$

We know that a $\Longrightarrow$ b

Also a $\Longleftrightarrow$ x

and b $\Longleftrightarrow$ y

$\therefore$ x $\Longrightarrow$ y

$x \Longleftrightarrow y$

$$\begin{aligned} x &\Longrightarrow a \\ &\Longrightarrow b \\ &\Longrightarrow c \\ &\Longrightarrow d \\ &\Longrightarrow y \end{aligned}$$

Conversely
$$\begin{aligned} y &\Longrightarrow p \\ &\Longrightarrow q \\ &\Longrightarrow r \\ &\Longrightarrow x \end{aligned}$$

Hence x $\Longleftrightarrow$ y

$$\begin{aligned} x &\Longleftrightarrow a \\ &\Longleftrightarrow b \\ &\Longleftrightarrow c \\ &\Longleftrightarrow d \\ &\Longleftrightarrow y \end{aligned}$$

x is purple

"x is purple" means y

We know a and

$$\begin{aligned} \mathsf{a} &\Longrightarrow \mathsf{b} \\ &\Longrightarrow \mathsf{c} \\ &\Longrightarrow \mathsf{d} \\ &\Longrightarrow \mathsf{y} \end{aligned}$$

$\therefore$ x is purple as required

$\forall x, p(x)$ is true

Prove that any rational number can be expressed as $\frac{m}{n}$ where m and n are integers that are not both even.

Let x be a rational number. So x can be expressed as $\frac{p}{q}$ where p and q are integers and $q \neq 0$.

...

$\exists x$ s.t. $p(x)$ is true

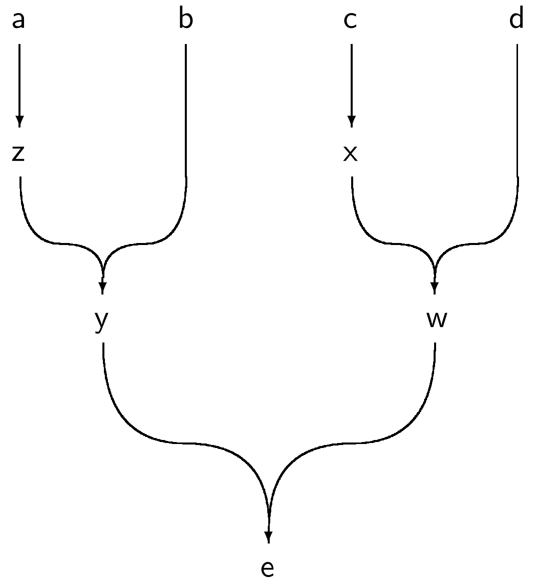
$$\exists\ \delta > 0\ \text{ s.t. }\ |x| < \delta \implies |x^2| < \frac{1}{100}$$

Put $\delta = \frac{1}{10}$. Now $|x^2| = |x|^2$ so we have

$$|x| < \frac{1}{10} \implies |x^2| < \frac{1}{100}$$

If a, b, c, d are true then e is true

$$\begin{array}{ccc} \mathsf{a} & \Longrightarrow & \mathsf{z} \\ \mathsf{b\ and\ z} & \Longrightarrow & \mathsf{y} \\ \mathsf{c} & \Longrightarrow & \mathsf{x} \\ \mathsf{x\ and\ d} & \Longrightarrow & \mathsf{w} \\ \mathsf{y\ and\ w} & \Longrightarrow & \mathsf{e} \end{array}$$



How to write proofs: a quick guide. Eugenia Cheng. http://www.math.uchicago.edu/~eugenia

Proof by Contradiction

- We are trying to prove that some statement *P* is true
- We say «suppose *P* were not true» and find a contradiction
- Since *P* being false gives a contradiction, we deduce that *P* must be true

How to write proofs: a quick guide. Eugenia Cheng. http://www.math.uchicago.edu/~eugenia

Exploratory Data Analysis

NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/

Approach

- **Exploratory Data Analysis** employs a variety of (*mostly graphical*) techniques to:
  - maximize *insight* into a data set
  - uncover underlying *structure*
  - extract *important* variables
  - detect *outliers* and *anomalies*
  - test underlying *assumptions*
  - develop parsimonious *models*
  - determine *optimal* factor settings

NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/

Graphical techniques

- *Plotting the raw data* (data traces, histograms, bihistograms, probability plots, lag plots, block plots, and Youden plots)
- *Plotting simple statistics* such as mean plots, standard deviation plots, box plots, and main effect plots of the raw data
- *Positioning* such plots so as to maximize

Classical *vs.* Exploratory

Classical Data Analysis

1. Problem
2. Data
3. Model
4. Analysis
5. Conclusion

Exploratory Data Analysis

1. Problem
2. Data
3. **Analysis**
4. **Model**
5. Conclusion

NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/

Classical vs. Exploratory

- **Models**
- **Focus**
- **Techniques**
- **Rigor**
- **Data Treatment**
- **Assumptions**

Model

- **Classical**
  - *imposes models* (both deterministic and probabilistic). *e.g.* regression models, analysis of variance. The most common probabilistic model assumes that the errors are normally distributed.

Model

- **Exploratory**
  - does not impose deterministic or probabilistic models on the data. In fact, EDA allows the data to suggest admissible models that best fit the data.

Focus

- **Classical**
  - *On the Model*. Estimate model parameters, generate predicted values from the model.
- **Exploratory**
  - *On the Data*. Structure, outliers, and models suggested by the data.

Techniques

- **Classical**
  - *Quantitative*. Mean, Variance, ANOVA, T-test, chi^2 tests, F-Test.
- **Exploratory**
  - *Graphical*. Scatter plots, Character plots, box plots, histograms, bihistograms, probability plots, residual plots, mean plots.

NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/

Rigor

- **Classical**
  - Probabilistic *foundation* of Science. Rigorous, formal, objective.
- **Exploratory**
  - Suggestive, indicative, insightful. Subjective, depend on interpretation.

NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/

Data Treatment

- **Classical**
  - Maps all data into *few* numbers. Loss of information.
- **Exploratory**
  - Shows *all* data. No loss of information.

NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/

Assumptions

- **Classical**
  - Tests based on classical techniques are very sensitive. Yet they depend on underlying assumptions. that could be *unkown* or *untested*.
- **Exploratory**
  - Makes no assumptions.

NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/

Quantitative Techniques

- Hypothesis testing
- Analysis of variance
- Point estimate and confidence intervals
- Least squares regression

NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/

Graphical Techniques

- Testing assumptions
- Model Validation
- Estimator Selection
- Relationship identification
- Factor Effect determination
- Outlier Detection

NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/

EDA Example

EDA Example

X	Y
10.00	8.04
8.00	6.95
13.00	7.58
9.00	8.81
11.00	8.33
14.00	9.96
6.00	7.24
4.00	4.26
12.00	10.84
7.00	4.82
5.00	5.68

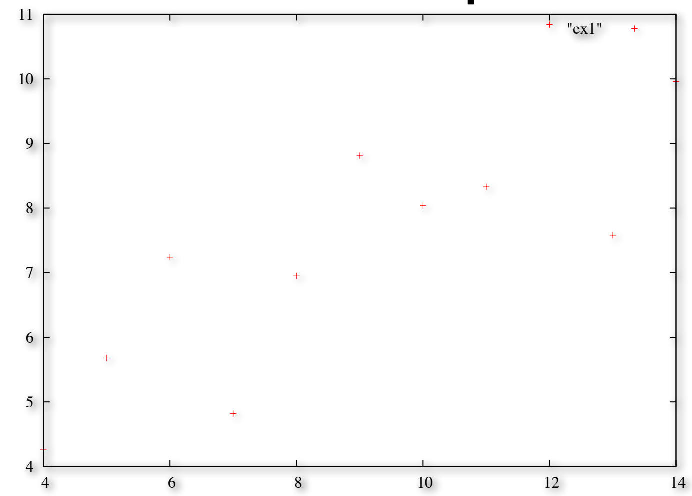
NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/

EDA Example (DS1)

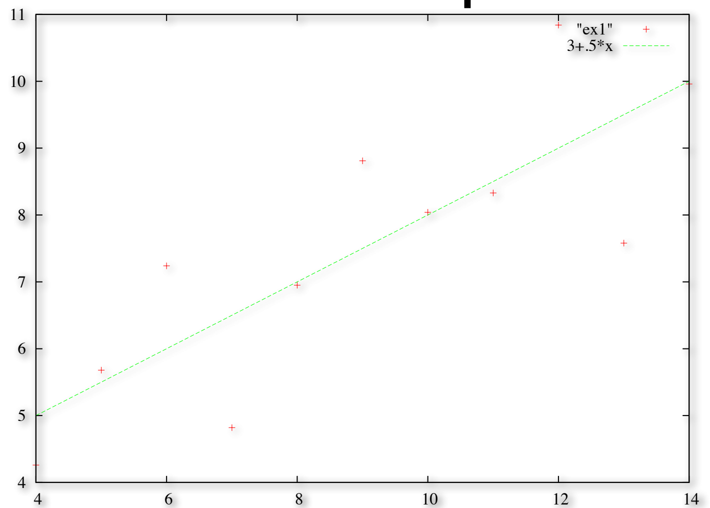
- N = 11
- Mean of X = 9.0
- Mean of Y = 7.5
- Intercept = 3
- Slope = 0.5
- Residual Standard Deviation = 1.237
- Correlation = 0.816

NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/

EDA Example



EDA Example



EDA Example

X2	Y2	X3	Y3	X4	Y4
10.00	9.14	10.00	7.46	8.00	6.58
8.00	8.14	8.00	6.77	8.00	5.76
13.00	8.74	13.00	12.74	8.00	7.71
9.00	8.77	9.00	7.11	8.00	8.84
11.00	9.26	11.00	7.81	8.00	8.47
14.00	8.10	14.00	8.84	8.00	7.04
6.00	6.13	6.00	6.08	8.00	5.25
4.00	3.10	4.00	5.39	19.00	12.50
12.00	9.13	12.00	8.15	8.00	5.56
7.00	7.26	7.00	6.42	8.00	7.91
5.00	4.74	5.00	5.73	8.00	6.89

NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/

EDA Example (DS2)

- N = 11
- Mean of X = 9.0
- Mean of Y = 7.5
- Intercept = 3
- Slope = 0.5
- Residual Standard Deviation = 1.237
- Correlation = 0.816

NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/

EDA Example (DS3)

- N = 11
- Mean of X = 9.0
- Mean of Y = 7.5
- Intercept = 3
- Slope = 0.5
- Residual Standard Deviation = 1.236
- Correlation = 0.816

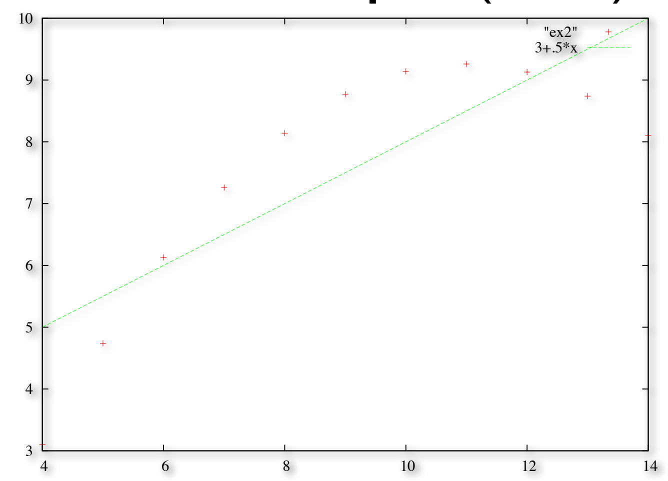
NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/

EDA Example (DS4)

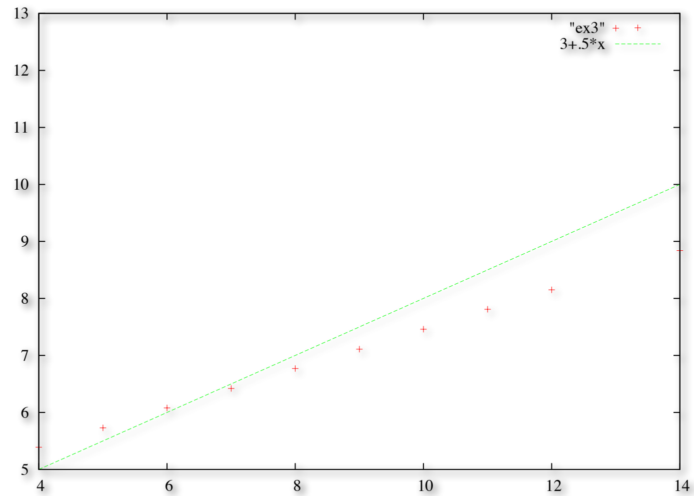
- N = 11
- Mean of X = 9.0
- Mean of Y = 7.5
- Intercept = 3
- Slope = 0.5
- Residual Standard Deviation = 1.236
- Correlation = 0.817

NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/

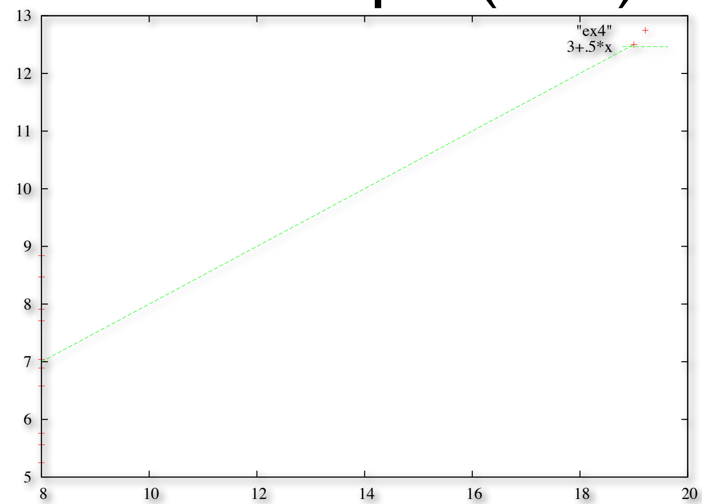
EDA Example (DS2)



EDA Example (DS3)



EDA Example (DS4)



Four Basic Tools

Univariate Data

- Most basic tools operate on *univariate* data, *i.e.* a list of *single* responses

Data Sets

- **Flow DS**: This data set was collected by Bob Zarr of NIST in January 1990 from a heat flow meter calibration and stability analysis. The response variable is a calibration factor.

Data Sets

- **Walk DS**: A random walk can be generated from a set of uniform random numbers by the formula :

$$R_i = \sum_{j=1}^{i} (U_j - 0.5)$$

- where U is a set of uniform random numbers

NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/

Data Sets

- **Beam DS**: This data set was collected by H.S. Lew of NIST in 1969 to measure steel-concrete deflections. The response variable is the deflection of a beam from center point.

NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/

Run-sequence Plot

- Considers *Univariate* Data
- **Vertical axis:** response variable Y(i)
- **Horizontal Axis:** Index i (i=1,2,3,...)

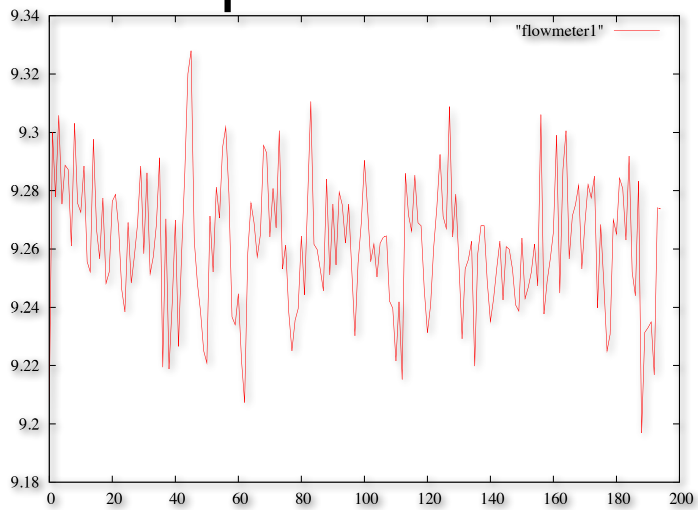
NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/

Run-sequence Plot

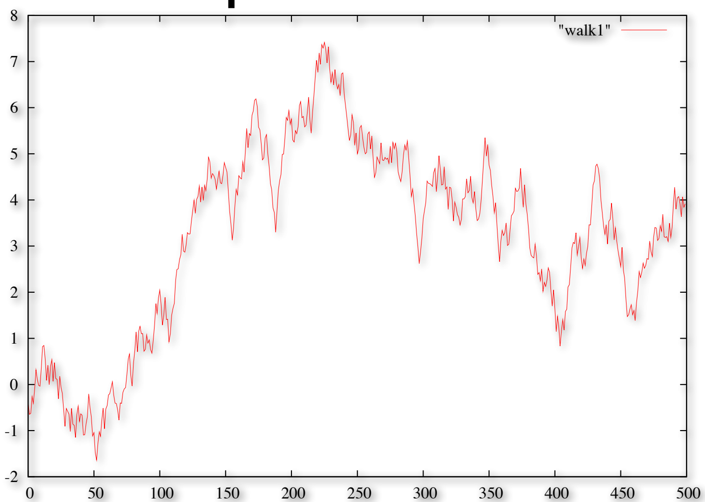
- Used to answer the questions
  - Are there any *shifts* in *location* ?
  - Are there any *shifts* in *variation* ?
  - Are there any *outliers* ?

NIST/SEMATECH e-Handbook of Statistical Methods,
http://www.itl.nist.gov/div898/handbook/

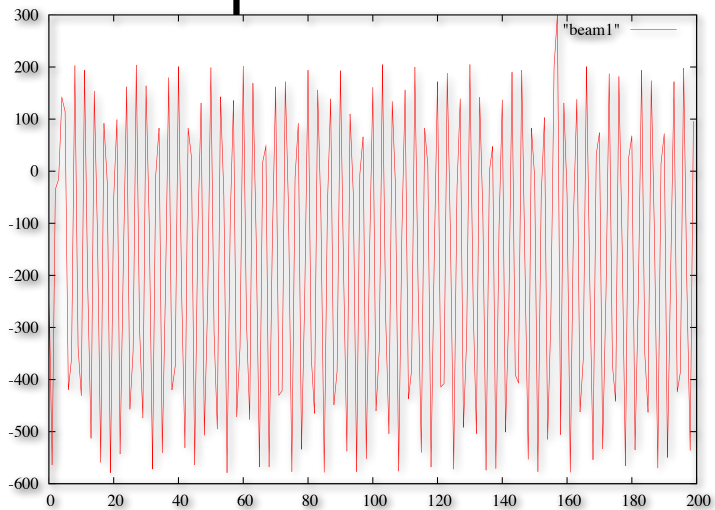
Run-sequence Flow DS



Run-sequence Walk DS



Run-sequence Beam DS



Lag Plot

- Considers *univariate* data
- **Vertical Axis**: *Y(i)* for all *i*
- **Horizontal Axis**: *Y(i-1)* for all *i*

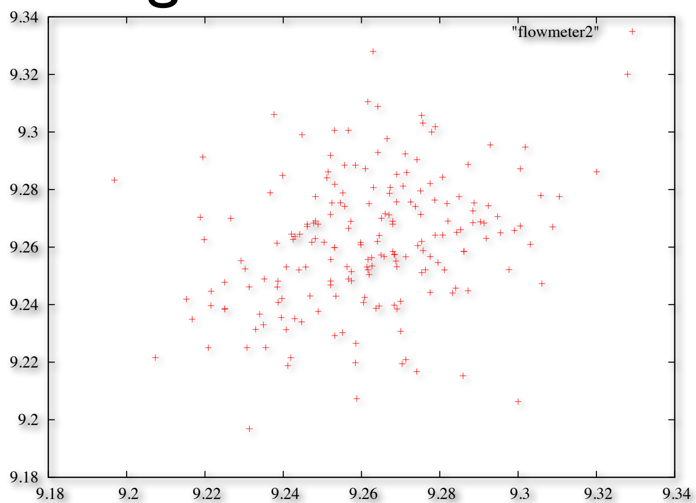
NIST/SEMATECH e-Handbook of Statistical Methods,
http://www.itl.nist.gov/div898/handbook/

Lag Plot

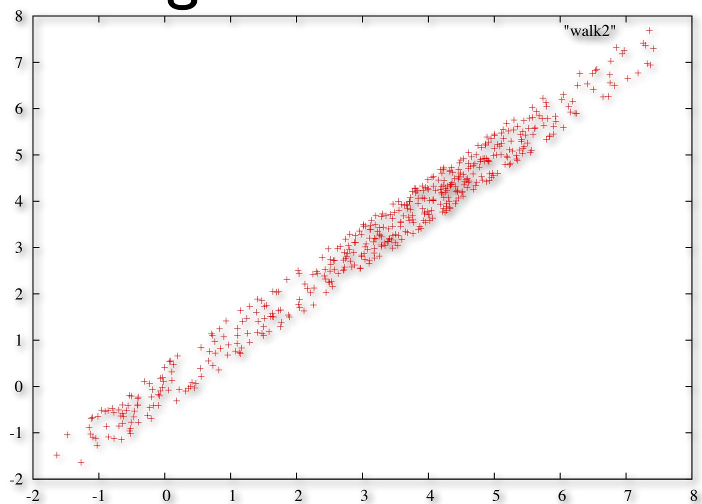
- Are the data *random* ?
- Is there *serial correlation* in the data ?
- What is a suitable *model* for the data ?
- Are there *outliers* in the data ?

NIST/SEMATECH e-Handbook of Statistical Methods,
http://www.itl.nist.gov/div898/handbook/

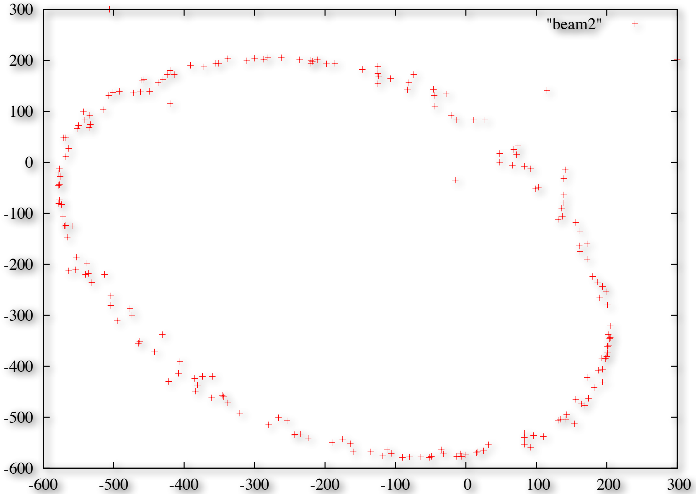
Lag Plot Flow DS



Lag Plot Walk DS



Lag plot Beam DS



Histogram

- Considers univariate data
- Split the range of the data into equal-sized bins, then for each bin the number of points from the data for each bin are counted
- **Vertical axis:** Frequency
- **Horizontal axis**: Response variable

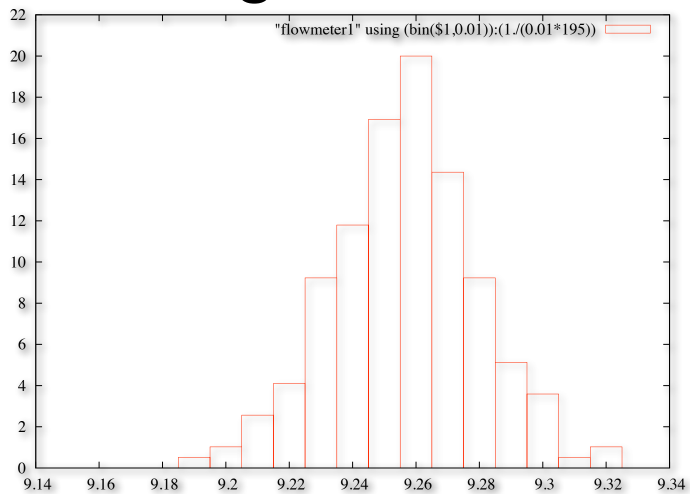
NIST/SEMATECH e-Handbook of Statistical Methods,
http://www.itl.nist.gov/div898/handbook/

Histogram

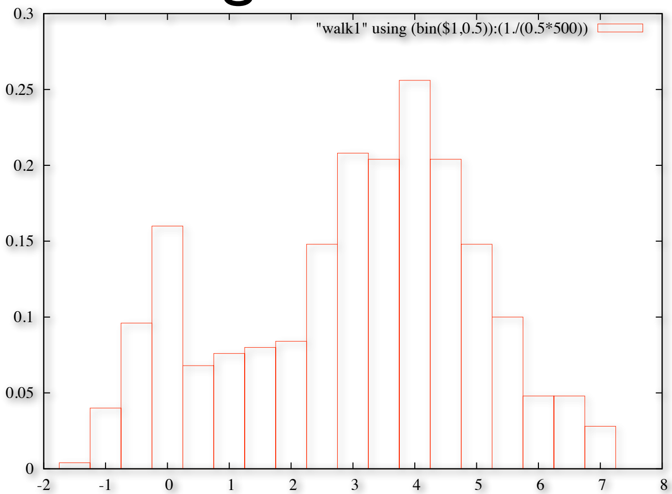
- Used to answer the following questions
  - What kind of population do the data come from ?
  - Where are the data located ?
  - How spread out are the data ?
  - Are the data symmetric or skewed ?
  - Are there outliers in the data ?

NIST/SEMATECH e-Handbook of Statistical Methods,
http://www.itl.nist.gov/div898/handbook/

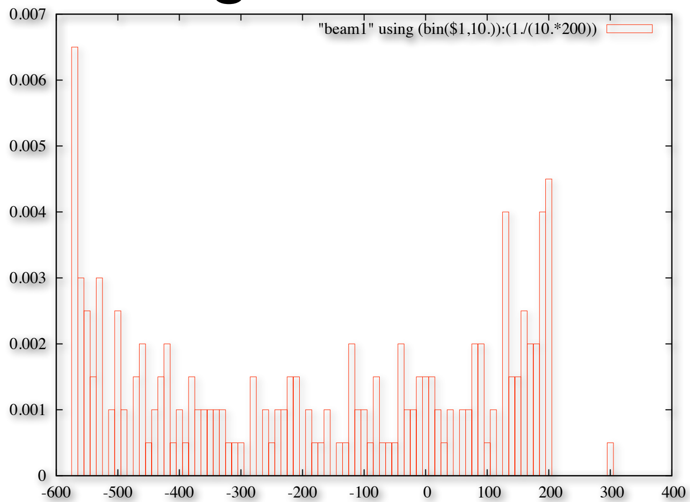
Histogram Flow DS



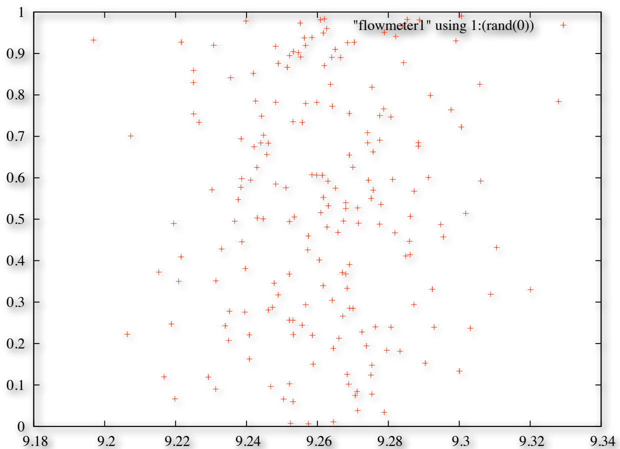
Histogram Walk DS



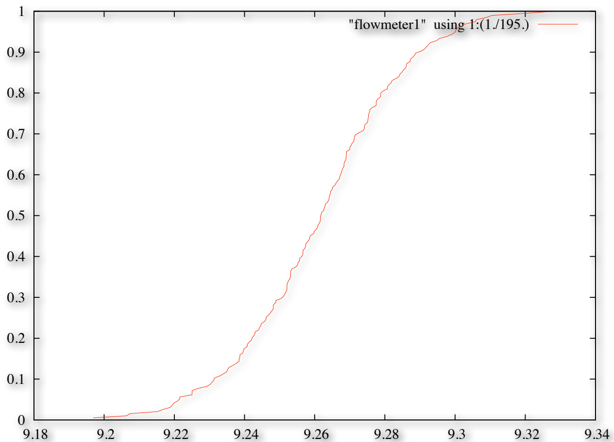
Histogram Beam DS



Beyond Histograms : Jitter Plots



Beyond Histograms : (Normal) Cumulative



(Normal) Probability Plot

- Considers univariate data
- **Vertical axis:** Ordered Response values
- **Horizontal axis:** Normal order statistics median

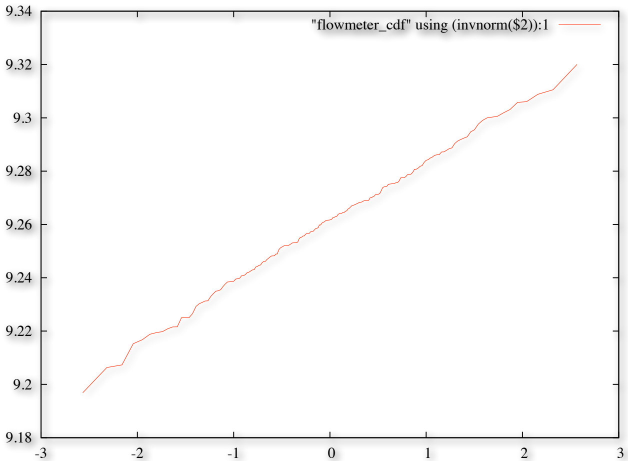
NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/

(Normal) Probability Plot

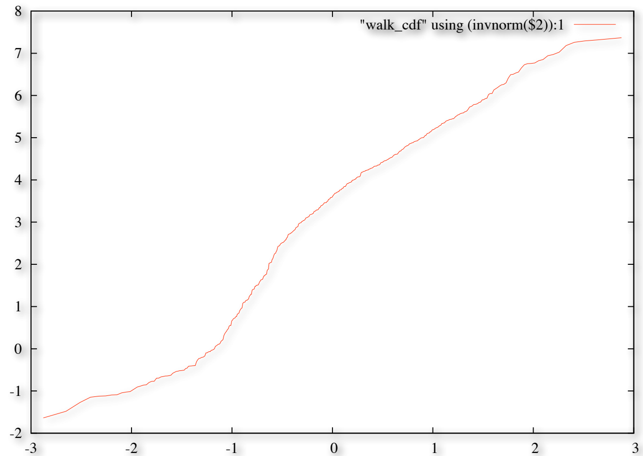
- Used to answer the following questions:
  - Are the data normally distributed ?
  - What is the nature of the departure from normality (data skewed, shorted than expected tail, longer than expected tails, etc.) ?

NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/

(Normal) Probability Plot Flow DS



(Normal) Probability Plot Walk DS



(Normal) Probability Plot Beam DS