Intro to Data Science

Data Exploratation

Ed Podojil Data Engineer/Scientist, Animoto

Welcome!

My email: **ed@animoto.com**

Alice's email: **alice@yipit.com**

Agenda

What is Data Science?

The Data Mining Workflow

Working at the Unix Command Line

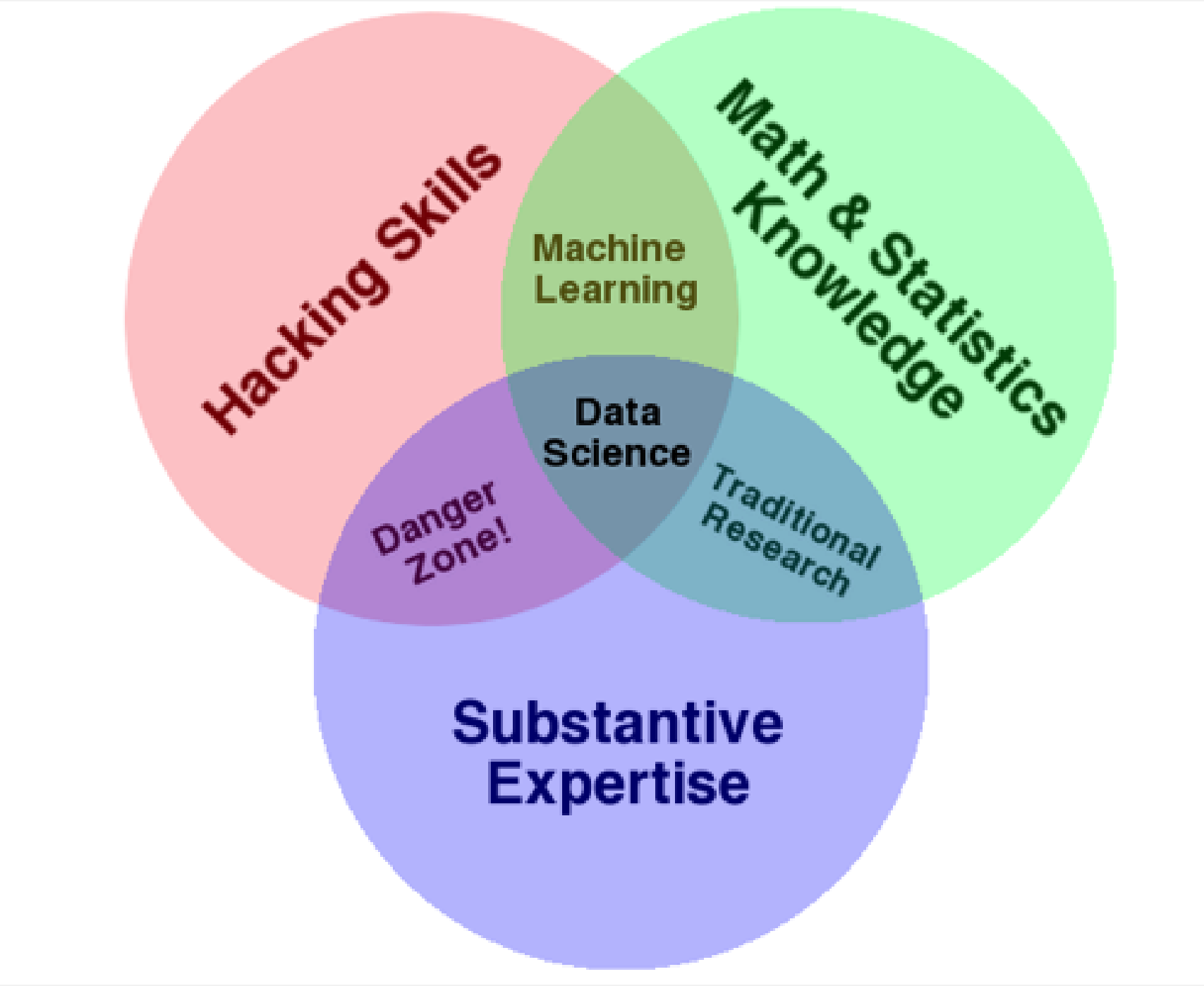
Visualizing Data with R & ggplot2

What is Data Science?

What is Data Science?

Tools and techniques for extracting information from data

Interdisciplinary, problem-oriented subject


Hacking Skills
Math & Statistics Knowledge
Substantive Expertise
Machine Learning
Data Science
Danger Zone!
Traditional Research

and...

Communication skills

What is Data Science?

Tools and techniques for extracting information from data

Interdisciplinary, problem-oriented subject

Application of scientific techniques to practical problems

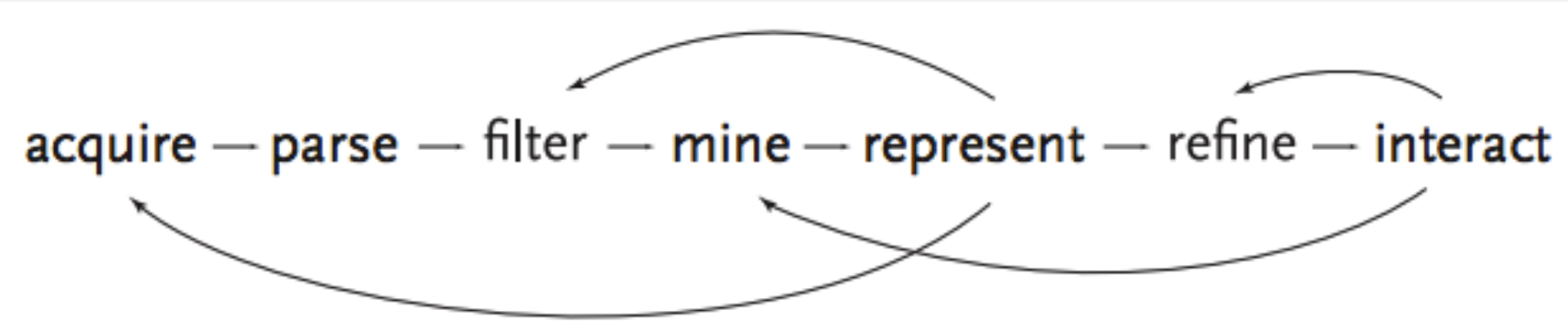
Rapidly growing!

Who Uses Data Science?


NETFLIX
amazon
facebook
Google
LinkedIn
FiveThirtyEight
Nate Silver's Political Calculus
foursquare
Etsy
2012
BARACKOBAMA.COM
Microsoft

The Data Science Workflow


COMPUTER SCIENCE
MATHEMATICS, STATISTICS, AND DATA MINING
GRAPHIC DESIGN
INFOVIS AND HCI
acquire
parse
filter
mine
represent
refine
interact


acquire — parse — filter — mine — represent — refine — interact


Business understanding
Data understanding
Data preparation
Modeling
Evaluation
Deployment
Data

III. Working at the Unix Command Line

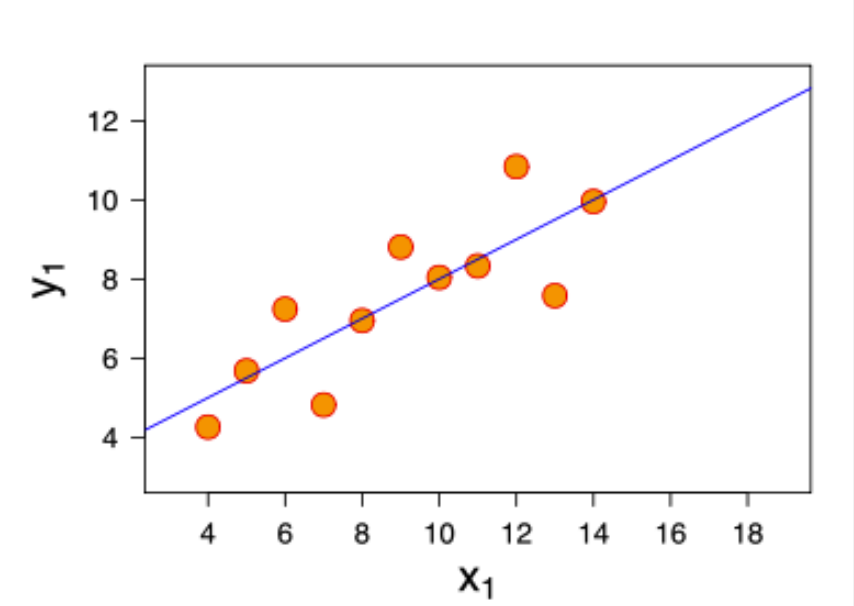
Bash tools we'll be using often:

- **Navigation:** ls, cd
- **Create and delete:** cat, touch, mv, cp, mkdir, rm, rmdir
- **View and search:** head, tail, less, cat, grep
- **Edit and Interact:** vim, awk, sed, tr, sort, uniq, wc
- **Combine steps:** Pipe (|)
- **Learn More:** man, apropos

IV. Visualizing data with R & ggplot2

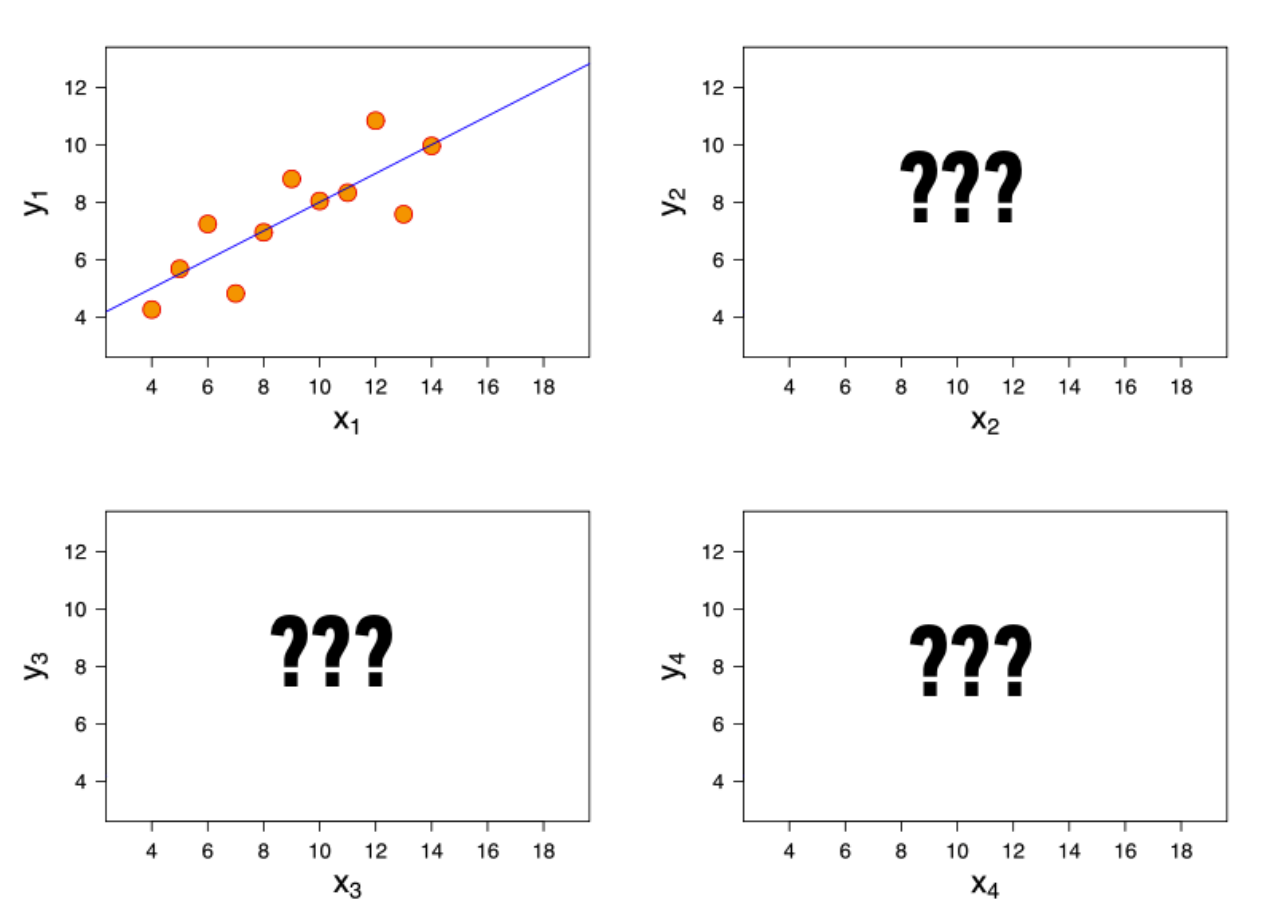
Why Visualize Data?

Consider this:



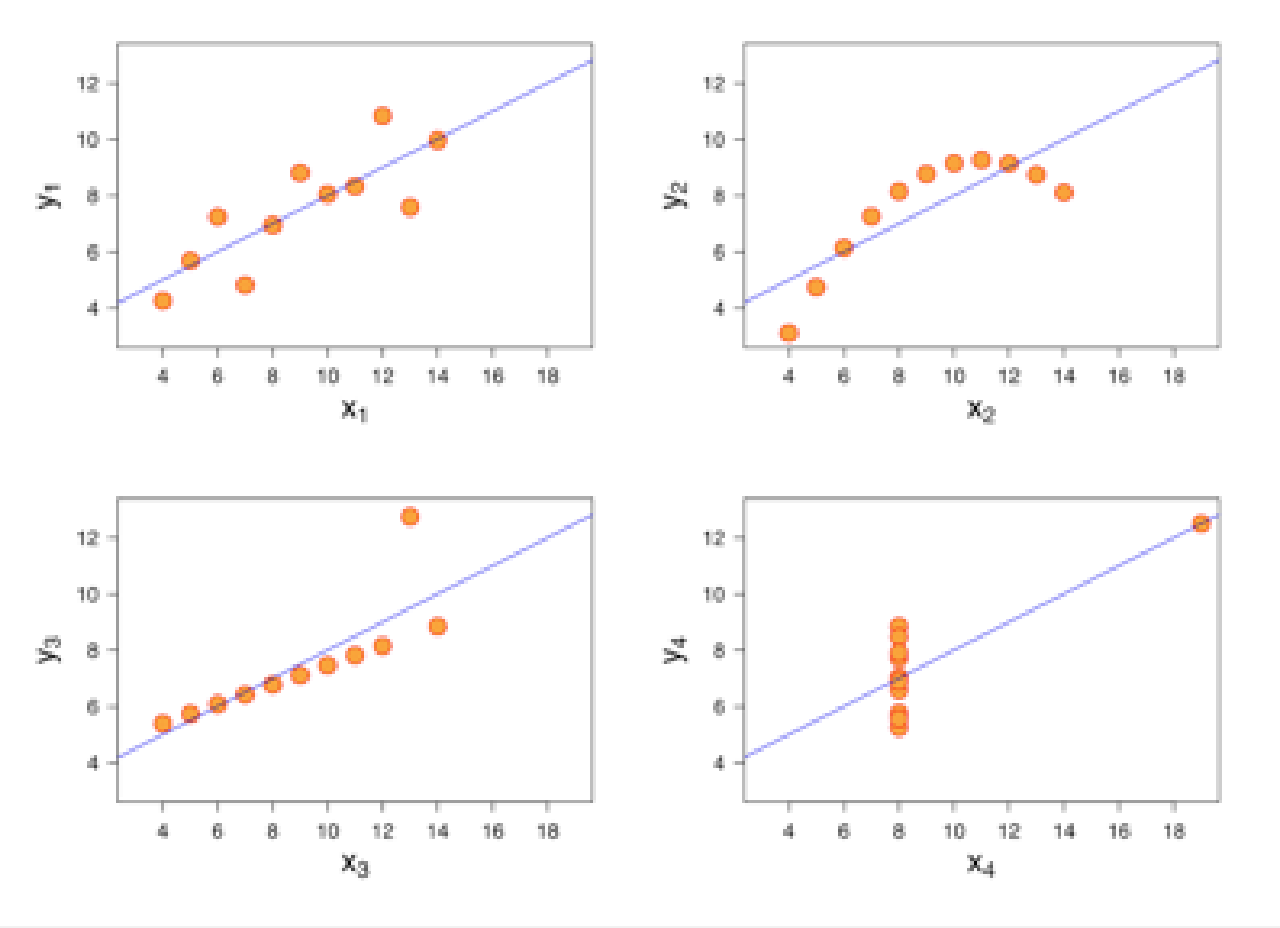
- Eleven (x, y) points
- mean(x) = 8, mean(y) = 7.5
- variance(x) = 11, variance(y) = 4.1
- correlation(x,y) = 0.8
- Best fit: y= 3.0 + .5x

Suppose we have three more datasets with the exact same characteristics.



How similar are these?

Not very!



Exercise

Visualize Data with R & ggplot2

Become familiar with the R environment

Explore Data in R

Visualize Data using ggplot2

Let's load some data

read.csv defaults to header=T, sep=",", and escapes quotes ("\"")

```
df <- read.csv('http://heypodo.com/public/etc/pace.csv')
```

we're using ggplot2, so let's load it in as well, and check out the data

```
library(ggplot2)
head(df)
summary(df)
```

Generate a plot

A 2d scatterplot of the data shows an obviously nonlinear relationship

```
ggplot(df, aes(y=speed, x=pop)) + geom_point()
```

We can confirm this with ggplot's smoother

```
ggplot(df, aes(y=speed, x=pop)) + geom_point() + geom_smooth(method="lm")
```

Generate the first model

```
linear.fit <- lm(pop ~ speed, data=df)
summary(linear.fit)
```

Fitting a linear model to this dataset produces significant coeffs
with an R-squared of ~43%, which is not bad,
but based on the shape of the data, we can probably do better

```
Call:
lm(formula = pop ~ speed, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.42046 -0.51769 -0.00158  0.65962  1.13502

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.6880     0.2778  13.275 6.17e-09 ***
speed         0.9672     0.3094   3.126  0.00803 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8205 on 13 degrees of freedom
Multiple R-squared: 0.4292, Adjusted R-squared: 0.3853
F-statistic: 9.774 on 1 and 13 DF,  p-value: 0.008029
```

Generate a new plot

This scatterplot shows the relationship after a log-log transformation based on this (and the previous) plot, we should expect the transformed data to produce a better linear fit

```
ggplot(df, aes(y=log(speed), x=log(pop))) + geom_point()
```

Why does this work?

The nonlinear relationship we saw before is an example of a **"power law"**

This is a linear fit on the transformed variables...
note that R-squared has nearly doubled

```
log.fit <- lm(log(speed) ~ log(pop), data=df)
summary(log.fit)
```

Let's verify with a new ggplot smoother

```
ggplot(df, aes(y=log(speed), x=log(pop))) + geom_point() + geom_smooth(method="lm")
```

Practice

Go through each of these data sets included in R and visualize their traits using ggplot2

Do any of them have data that can follow the log-log power law?

```
library(MASS) # Load data from S (R's non open source version)
data(crabs) # Morphological Measurements on Leptograpsus Crabs
data(mammals) # Brain and Body Weights for 62 Species of Land Mammals
data(wtloss) # Data of a weight loss patient
```

Discussion