INSEAD

The Business School for the World®

Prof. Anton Ovchinnikov

Prof. Spyros Zoumpoulis

DSB Classes 01-02, January 24, 2018

- "Data Science for Business" Course Intro
- UDJ regression recall, "Sarah Gets a Diamond" case
- From Excel to R

Why are we here?

INSEAD
The Business School for the World®

- "All-things-digital/data" (AI, Machine Learning, …) is at the top of every leader/CEO/manager/recruiter agenda
  - Forbes: "The Top 10 Business Trends That Will Drive Success In 2018" AI is at #1 https://www.forbes.com/sites/ianaltman/2017/12/05/the-top-business-trends-that-will-drive-success-in-2018/#66bebaf0701a
  - Fortune "Five Big Business Trends to Watch in 2018" – AI is #2 http://fortune.com/2018/01/02/five-big-business-trends-to-watch-in-2018/
  - Economist, WSJ, FT, … – "data" regularly on the cover/1st page
- Demand for data-savvy managers is strong in every industry, in every geography: GMAC Recruiters Report https://www.gmac.com/market-intelligence-and-research/research-library/employment-outlook/2017-corporate-recruiters-survey-report.aspx

ARTWORK Tamar Cohen, *Andrew J. Bubolz*, 2011, silk screen on a page from a high school yearbook, 8.5" x 11"

Data Scientist:

The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.

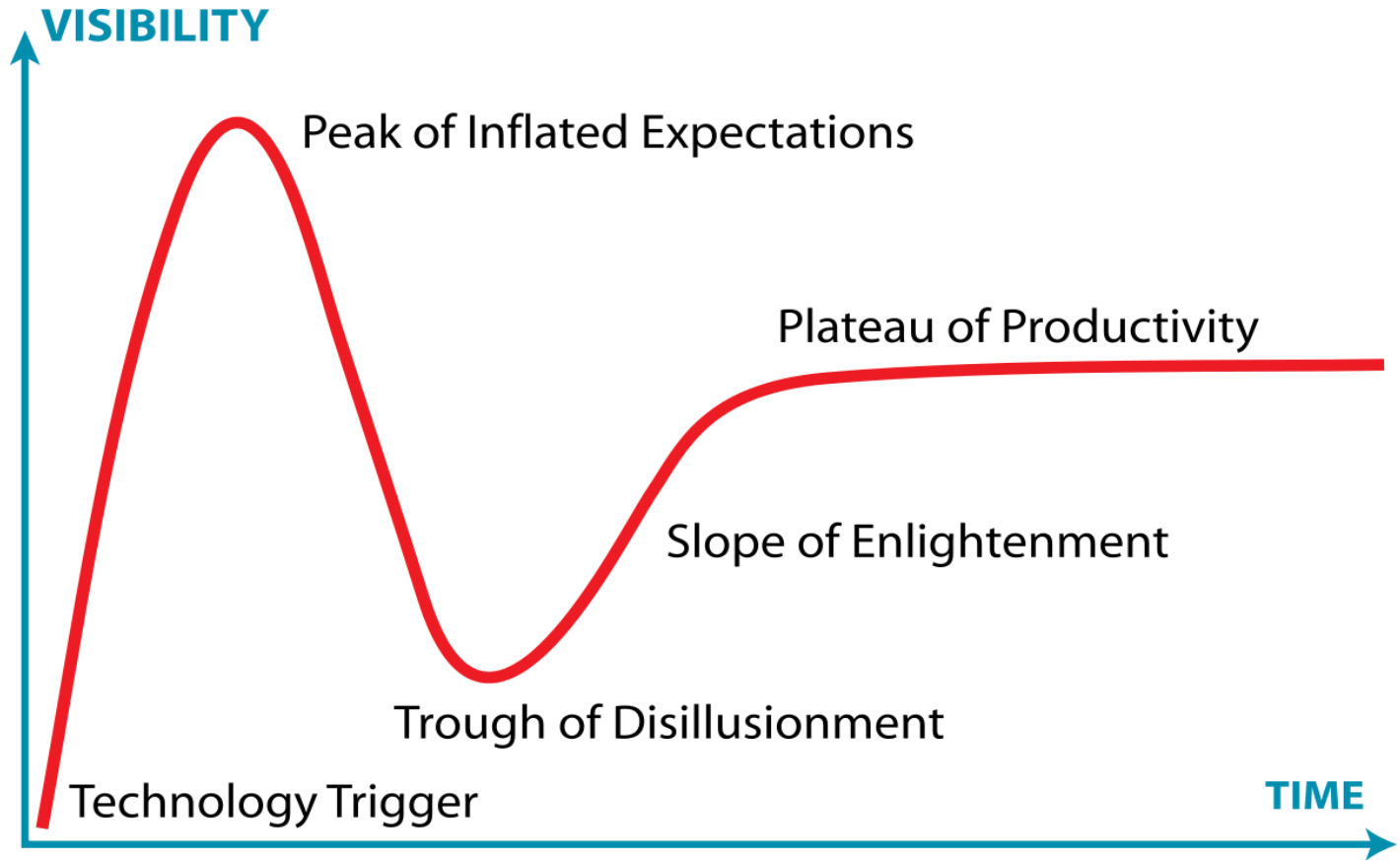
by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

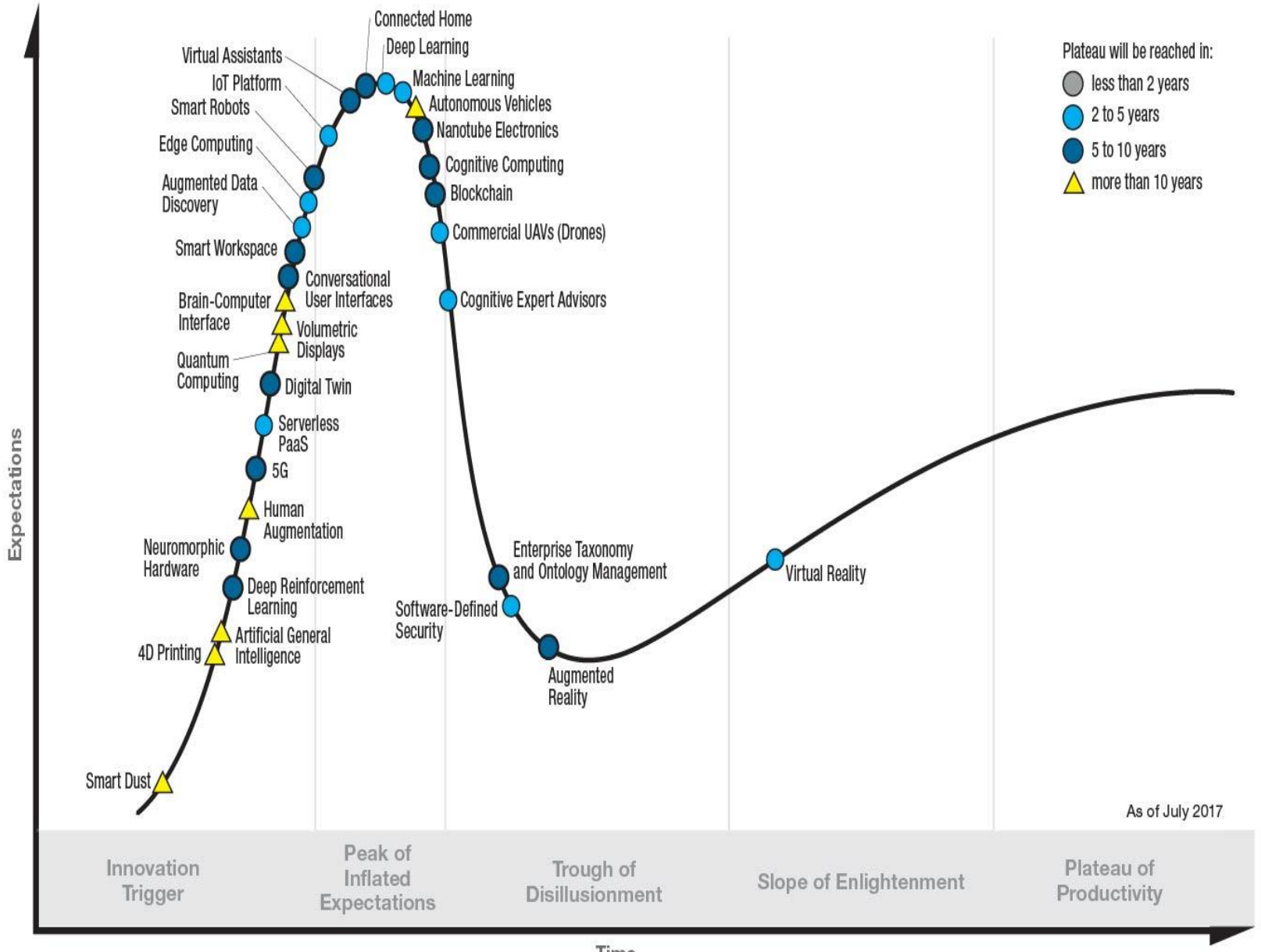
Is “all-things-digital” just a hype?

INSEAD
The Business School for the World®

The perception of highly publicized new technologies tends to follow a consistent pattern, Gartner hype cycle www.gartner.com

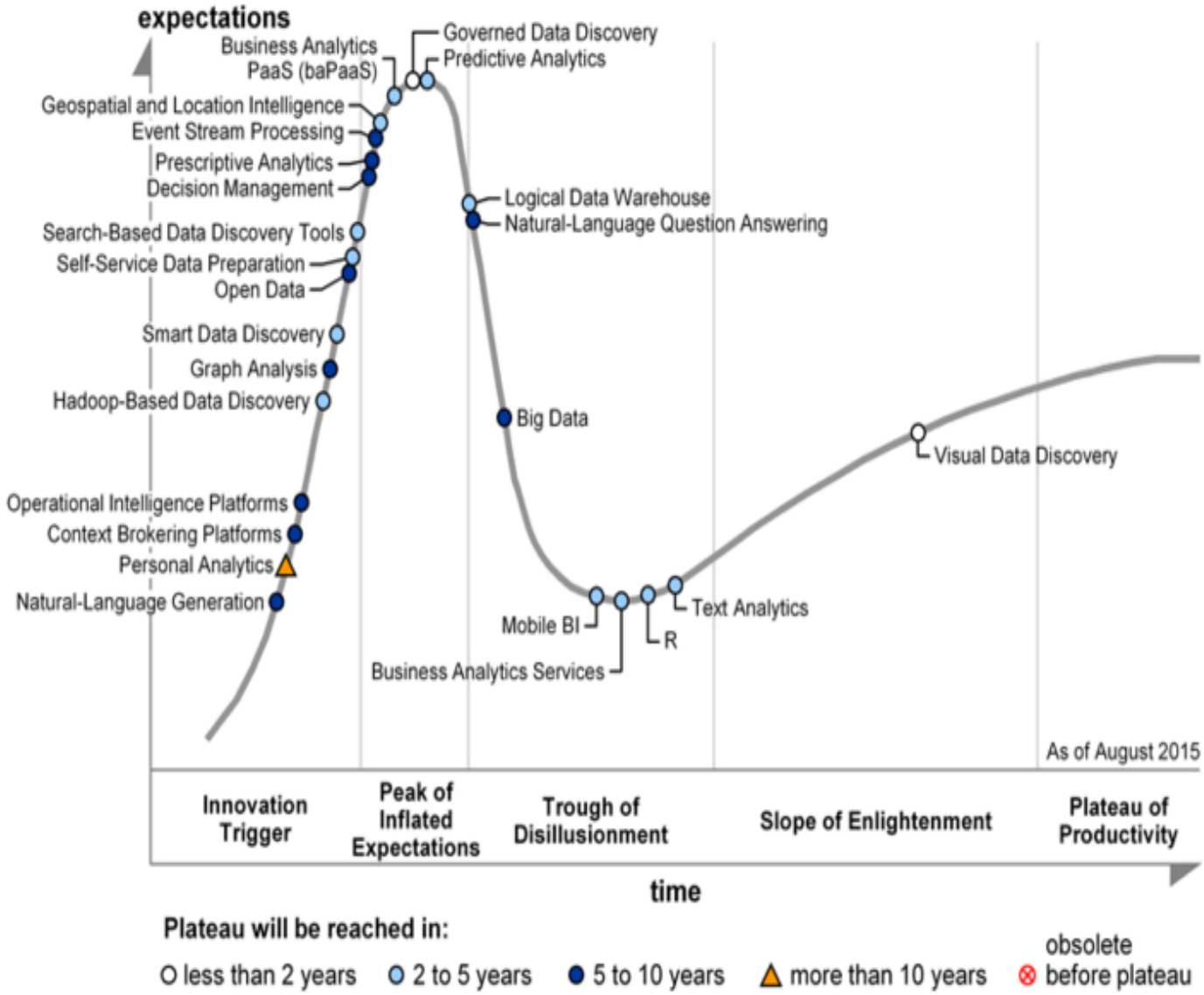


Gartner Hype Cycle for Emerging Technologies, 2017



Specific tech on the way to plateau

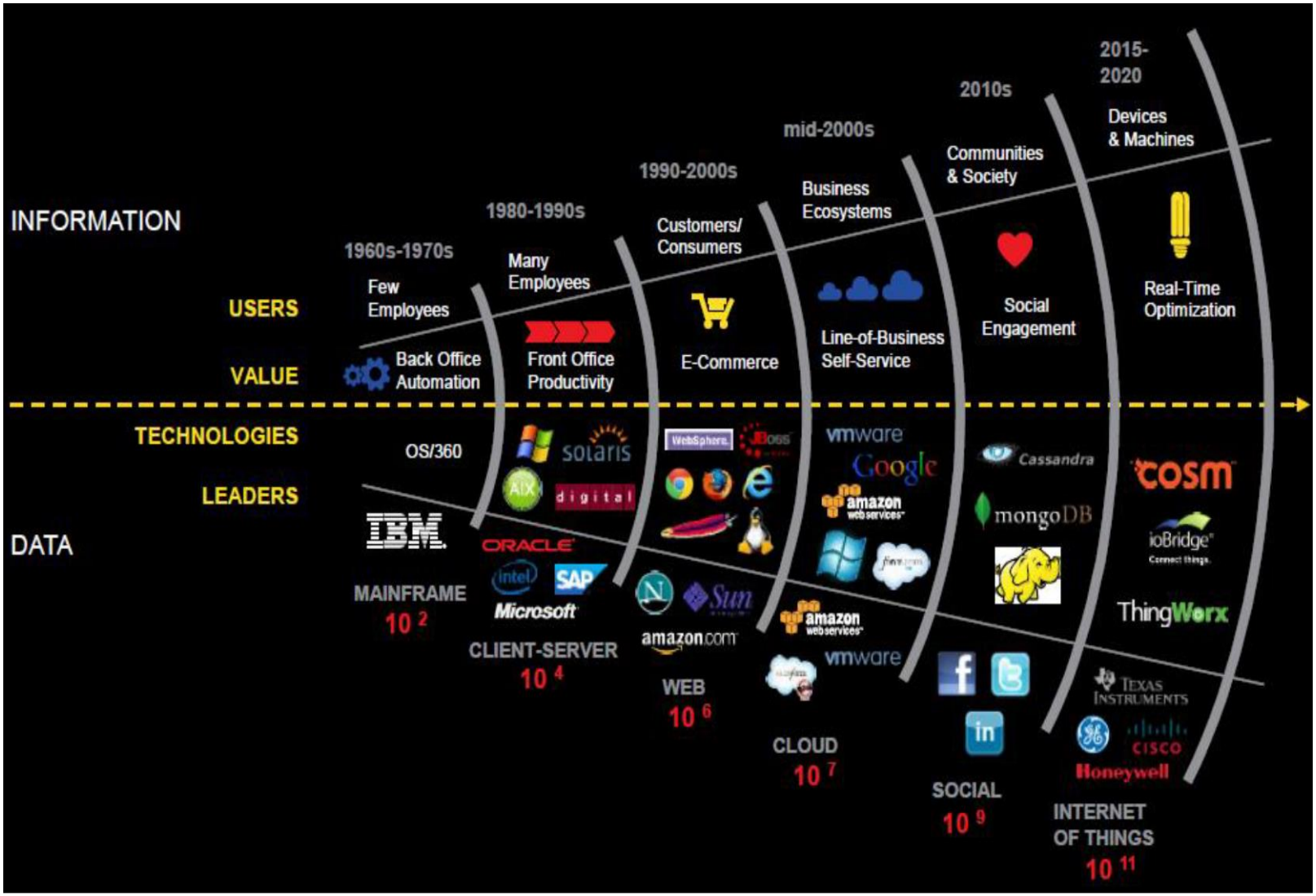
INSEAD
The Business School
r the World®



Source: Gartner (August 2015)

Big Data vs Smart Data: What makes data "Big"?

INSEAD
The Business School for the World®



3Vs: Volume, Veracity, Velocity

Big Data vs Smart Data: What makes data "Smart"?

INSEAD
The Business School for the World®

Smart data is:

- Data that is right for the decision
- Supports (and is supported by) analytics, expertise and machines
- Hits your key business drivers: customer acquisition, loyalty, growth, risk optimization, etc.

Big Data vs Smart Data: Examples

INSEAD
The Business School for the World®

"Big data"	"Smart data"
• Full-motion video feed from security cameras at a bank branch	• Customer arrival patterns by time of day; security alert
• Real-time website click-stream data	• Purchase behavior segmentation
• Raw twitter feed	• Sentiment analyses
• Your examples?	• Your examples?

Big Data vs Smart Data: What makes data "Smart"?

INSEAD
The Business School for the World®

Smart data is:

- Data that is right for the decision
- Supports (and is supported by) analytics, expertise and machines
- Hits your key business drivers: customer acquisition, loyalty, growth, risk optimization, etc.

Data Engineering

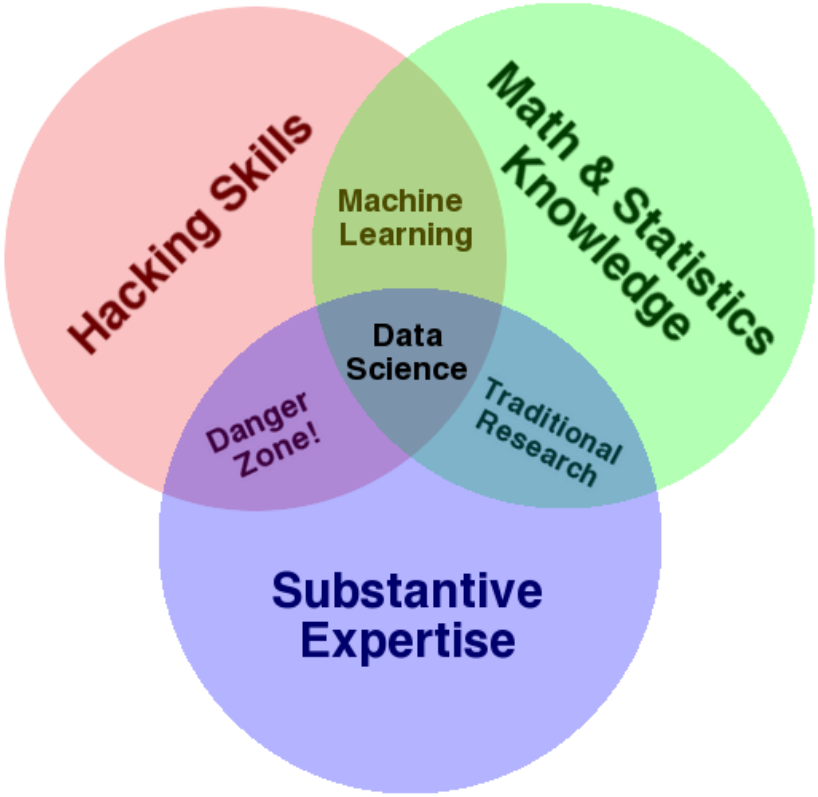
Data Analytics

Business Expertise

The nexus of "Smart Data": Data Science

INSEAD
The Business School for the World®

Data Engineering



Analytics

Business Expertise

Source: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Course objectives

INSEAD
The Business School for the World®

- "To build your capability in data science so that you can effectively add value through the intelligent management and use of data in your organizations."
- Three key elements: analytics techniques, business applications, basic coding/programming (in R).
- Main classes of techniques / "modules" of the course:
  - Regression
  - Time-series models

  } "Predicting quantities" ("stats", building on UDJ)

  - Classification
  - Clustering and Segmentation

  } "Predicting events" (machine learning)

Structure of the course

INSEAD
The Business School for the World®

- SESSIONS 1-2 (AO): Data analytics process; from Excel to R
  - Tutorial 1: Getting comfortable with R
- SESSIONS 3-4 (AO): Time Series Models
- SESSIONS 5-6 (AO): Intro to classification, logistic regression and machine learning — Due: A1: Yahoo/Tumblr case
  - Tutorial 2: Midterm R help / classification
- SESSIONS 7-8 (SZ): Classification cont.; Dimensionality reduction, principle component analyses — Due: A2: Credit defaults case
- SESSIONS 9-10 (SZ): Segmentation and clustering
  - Tutorial 3: Q&A on R for three main modules
- SESSIONS 11-12 (SZ): Catch-up and wrap-up; Guest speaker — Due: A3: Boats case
  - "Tutorials 4,5": Hands-on help on projects — Due: 1-pager project proposal, data
- SESSIONS 13-14 (AO+SZ): Project presentations

Grading

INSEAD
The Business School for the World®

- Case assignments [group]
  - 3 in total, 15% grade each
- Participation [individual], 15%
- Course project [group], 40%
  - Purpose: "develop a data analytics solution to a business problem"
  - Samples of topics on course site; proposals due on **Feb 14**
  - All groups will present in the last 2 classes (Feb 20)
  - The Profs and the TA will act as an advisors to all groups on all aspects of the projects [don't hesitate to ask for help!]

Course materials

INSEAD
The Business School for the World®

- Main resource: course website(s)
  - All "common" content (links, readings, etc.) – open-source website: http://inseaddataanalytics.github.io/INSEADAnalytics/SGP18J.html
    - Created with and hosted on GitHub (open-source collaboration platform, will learn about it as we go)
  - All section-specific content (models seen in class, assignment submissions, etc.) – INSEAD course portal https://mbacourse.insead.edu/courses/DSC/SGP_P3_Data_Science_for_Business_OvchinnikovZoumpoulis_1830_July/pages/default.aspx
- No mandatory textbook; optional textbooks:
  - Data Science for Business (DSB)
  - Forecasting Principles and Practice (FPP), https://www.otexts.org/book/fpp

About your professor(s)

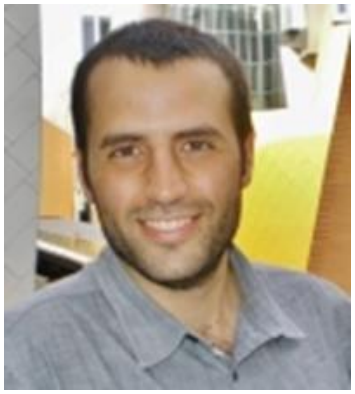
INSEAD
The Business School for the World®

Anton Ovchinnikov (AO)

- [off - chin[ny] - caugh]
- anton.ovchinnikov@insead.edu
  610D
- Visiting Professor of Technology, Operations and Decision Sciences
- Distinguished Professor of Management Analytics at the Queen's University in Canada
- Was born and grew up in Russia/Siberia, co-owned a design business before academia
- Love sailing/windsurfing, skiing and travel

Spyros Zoumpolis (SZ)

- spyros.zoumpoulis@insead.edu
  610D
- Professor of Decision Sciences and Technology Management



TA: Marat Salikhov

About you

INSEAD
The Business School for the World®

- Name, Background, Career Aspirations
- “I am taking this course because…”

Questions?

INSEAD
The Business School for the World®

- on Course Objectives?
- Structure?
- Grading?
- Deliverables?

INSEAD

The Business School for the World®

Linear regression "recall" from UDJ

"Sarah Gets a Diamond" case

Lets Recall: Linear Regression

INSEAD
The Business School for the World®

- You think that there may be a dependency between something you are interested in, Y, and something you know, X
  - X is called "independent variable"
  - Y is called "dependent variable"
  - The dependency itself is called "correlation"
- Knowing X does not tell you everything about Y, there can be different Ys even for the exact same X
- Thus the (linear) dependency is of the form of the following linear model:

  $$Y = a + b * X + error$$

  - a is called "intercept"
  - b is called "coefficient"
  - error has a zero mean ("unbiased") and distribution $\sim Normal(0, Standard\ Error)$

Linear Regression Continued…

INSEAD
The Business School for the World®

- You may often know many different Xs that may be correlated with Y -> multiple linear regression:

$$\boldsymbol{Y = a + b_1 * X_1 + b_2 * X_2 + b_3 * X_3 + \cdots + error}$$

- Significance (p-value of coefficients)
  - >0.1 - "insignificant" (data does not support the relationship)
  - <0.05 - "significant" (for situations where error is particularly undesirable, use <0.01)
  - Between 0.05 and 0.1 - "grey area" (no evidence of either significance or not)
- Types of variables:
  - Continuous (weight, income, height, etc…): Use "as is"
  - Categorical ("Red/Green/Blue", "Optimistic/Pessimistic/Neutral", etc…): Transform into binary/"dummy" variables (e.g., X_1=1 if "Red" and 0 otherwise, X_2=1 if "Green" and 0 otherwise). A category with n values requires $n-1$ dummy variables. Why?

"Sarah Gets a Diamond" case

INSEAD
The Business School for the World®

- The file "0102 Sarah Gets a Diamond data.xls" contains data about various attributes for >9000 diamonds. For the first 6000 you also have the price. The goal is to predict the prices of the remaining diamonds (i.e., those with IDs 6001 and above). From UDJ you know how to do it using regression…
- Next – **mini-competition**:
  - Form groups of 4-5 (I asked you to think about the groups before class; if unsure, see me and I will assign you)
  - Use tools you've learned in UDJ to predict prices, in your groups, in BORs (316-328)
  - Create a spreadsheet containing your group name in cell A1 (be creative), your names (last name first) in cells A3 to A7, and your forecasts of the prices of the diamonds with ID 60001 and above in cells A8 to A3149
- Email these spreadsheets to me (anton.ovchinnikov@insead.edu) and my assistant (alyssa.liao@insead.edu) by the beginning of the break (13:30 for AB and 17:15 for AA); we will resume with the competition results after the break

Results: TO BE UPDATED

INSEAD

The Business School for the World®

Now flip it over to you

INSEAD
The Business School for the World®

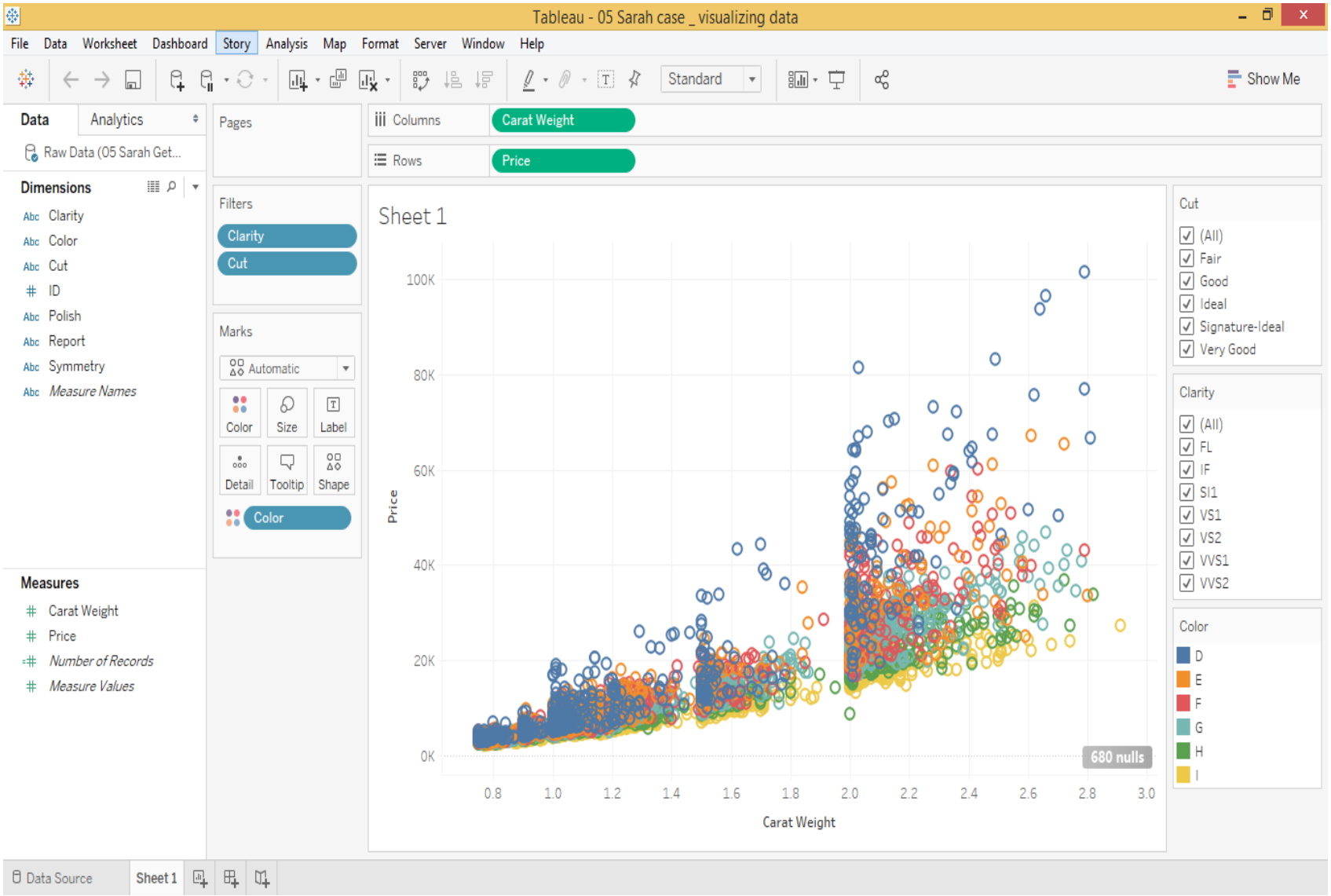
- Winning team:
  - Screen sharing
  - WebEx "meeting room" [running heading on portal]
  - https://insead.webex.com/meet/sgp.102insead.edu

- Approach?
- Why did you take that approach?

- Others – prepare your questions

Visualization: Tableau

INSEAD
The Business School for the World®



From Excel to R

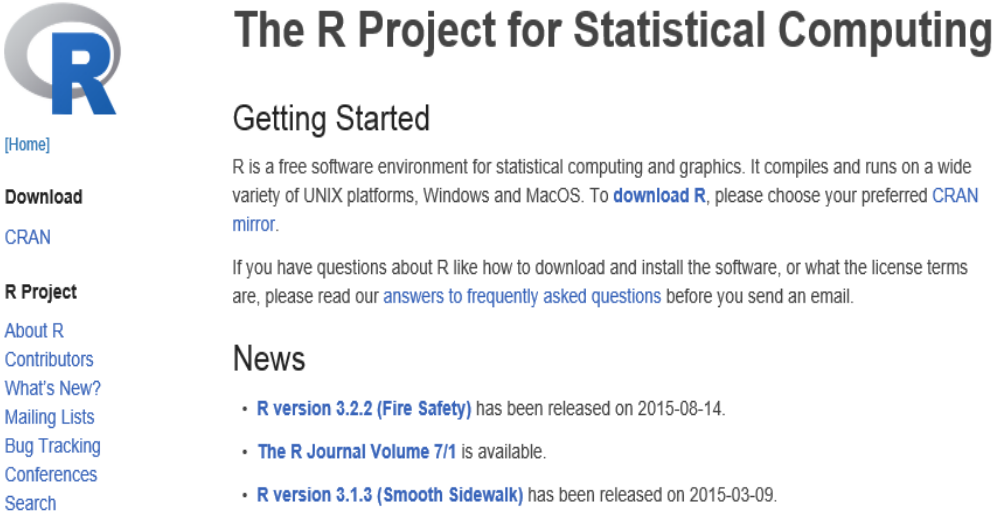
INSEAD
The Business School for the World®

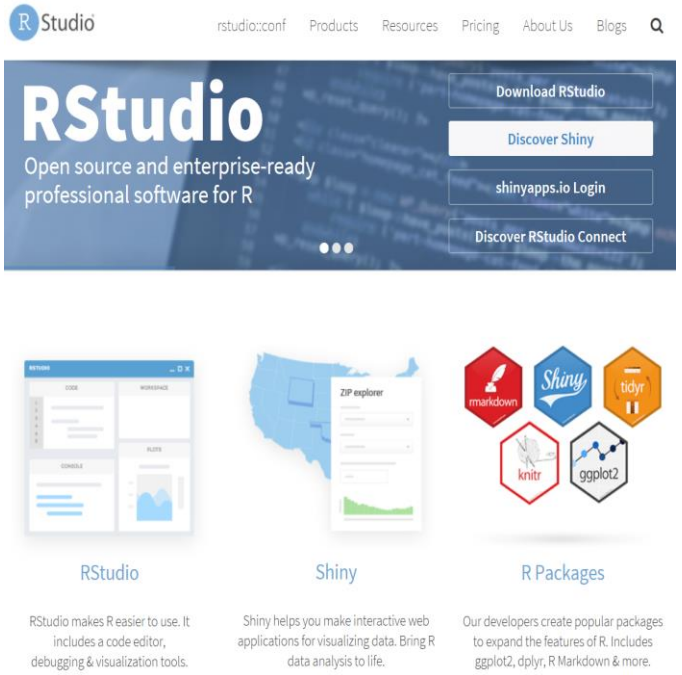
- Next: Intro to data analyses in R
  - What is R?
  - Why R?
  - "How to" R?: By following a simple code for the Sarah's case you just did in Excel
- Tonight: tutorial on R
  - Sarah's case simple code revisited + a simple problem to do on your own
- Sessions 3-4: time series models in R
  - How? Again, by following a simple code that I will provide
- Assignment 1, Yahoo/Tumblr case: combining Excel and R for startup valuation. How? By modifying the code from sessions 1-4 and the tutorial

What is R?

INSEAD
The Business School for the World®

- R is an open-source (free) programming language for statistical computing, [just one letter, "R"]
  - CRAN: R archive network that hosts "packages" (more next)
- Convenient user interface: R-Studio





- Main "competitor": language called "python"
  - R is more for data and learning, python is more for development and coding. The two are mostly identical in their analytical capabilities (can run each-others code via interpreters)
  - https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis

Why R?

INSEAD
The Business School for the World®

- Free, open source software
- Ultimate cross-platform operability
- Over>2,000 open-source packages
  - Forefront of developing new methods and tools (most stats/data science research papers come with R implementations of the proposed methods)
- Engaged community of data-scientists
- Numerous excellent learning and help resources, free online "books" and inexpensive courses (datacamp, udemy, coursera – see course website)
- Learning Data Science is effectively inseparable from learning R (or python)
- "Why R" vs "Why coding?"
  - Replicability: anyone who has your code can replicate exactly what you've done. Fundamentally different from Excel
  - Algorithmic thinking

"How to" R?

INSEAD
The Business School for the World®

- Some fundamentals:
  - CSV data format ("comma separated values") [note: regional settings]
  - variables, vectors/matrices and dataframes
    - data.frame is the "mother" of all datatypes in R
  - commenting out: # symbol
  - packages and libraries
    - The standard installation comes with only basic commands; most functionality is in packages/libraries
    - Installing packages (need to do that only once) + calling libraries (need to do that every time you relaunch R); will see later today
  - getting help:
    - Question mark and command opens help: e.g., "?plot"
    - or just search it online!

"How to" R? – Sarah's case

INSEAD
The Business School for the World®

Here is the code to create a model with MAPE ~ 7.6%

Yes, only six lines of code! Main function: `lm` ("linear model")

```
diamond.data<-read.csv(file.choose(), header=TRUE, sep=",") #load the data into the diamond.data dataframe
diamond.data.training<-subset(diamond.data, ID<=6000) #separate ID 1...6000 into "training"
diamond.data.prediction<-subset(diamond.data, ID>=6001) #separate ID 6001...9142 into "prediction"
fit<-lm(Price~Carat.Weight+Cut+Color+Clarity+Polish+Symmetry+Report, data=diamond.data.training) #run a multiple linear regression model (lm) on the training data, call it "fit"
predicted.prices<-predict(fit, diamond.data.prediction) #use the "fit" model to predict prices for the prediction data
write.csv2(predicted.prices, file = "Predicted Diamond Prices.csv") #export the predicted prices into a CSV file
```

Open the file "0102 Sarah Gets a Diamond -- Linear Regression.R" from INSEAD portal in RStudio, and lets proceed from there step-by-step

Additional R commands, "ideas"

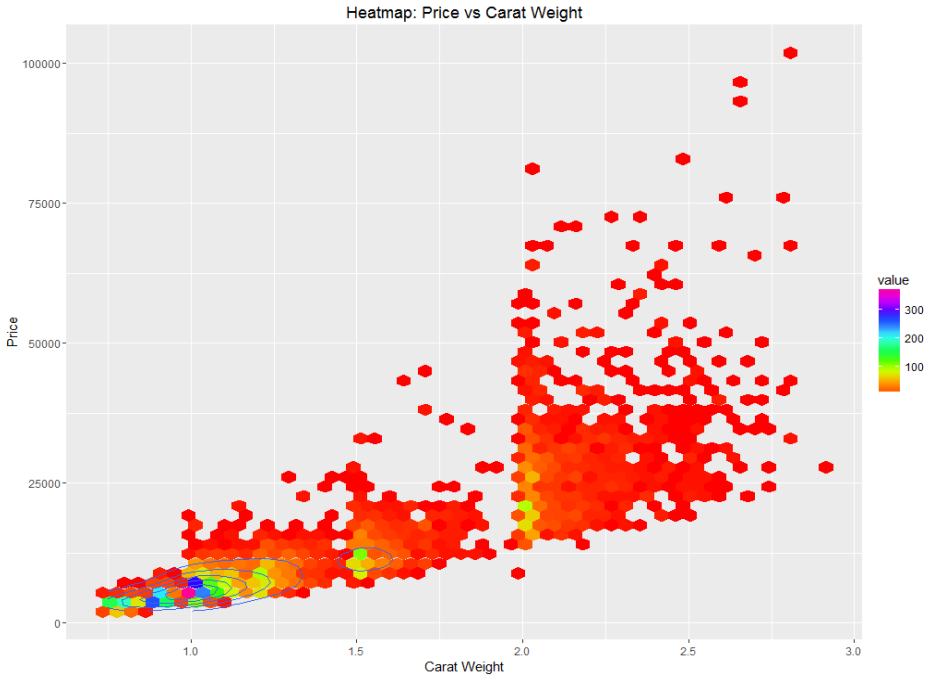
INSEAD
The Business School for the World®

- Continuous vs categorical variables – recognized automatically (nothing special you need to do!)
- Exploring the data: `str, head, tail, summary`
- Basic plotting: `plot, hist, abline, par`
- Log-transforms:
  - `fit.log<-lm(log(Price)~log(Carat.Weight)+…`
- Interactions:
  - `fit.log.i<-lm(log(Price)~log(Carat.Weight)*Color+…`
- More:
  - Advanced plotting
  - Cross-fold validation / holdout (pre-assessing the quality of the model)
  - Stepwise variable selection

Advanced "Tableau-like" plotting

INSEAD
The Business School for the World®

```
#install.packages("ggplot2") # installing a package -- do this only once, the first time you plan to use it

library(ggplot2) # calling a library from the package -- do this every time you plan on using it

ggplot(diamond.data.training, aes(x=Carat.Weight, y=Price, shape=Cut, color=Color)) +  geom_point(size=3)
#create a Tableau-like plot of price vs carat with color representing "color" and point shapes representing cut

heatmap<-ggplot(diamond.data.training, aes(Carat.Weight, Price))  + geom_hex(bins=50)
+scale_fill_gradientn(colours = rainbow(10) )+ggtitle("Heatmap: Price vs Carat Weight")+ theme(aspect.ratio =
0.8)+ labs(x = "Carat Weight", y = "Price")+geom_density2d()

print(heatmap)
```



Cross-fold validation / holdout

INSEAD
The Business School for the World®

- Recall: What was the task? To predict the prices of new diamonds, for which we know the features (Xs) but not the prices (Y). How can we recreate this task using the current data on 6000 diamonds?
- Main idea: set a portion of the data (e.g., 1000) as a "holdout" ("testing") sample. Build a model on the remaining 5000 ("training data") and use it to predict the prices of the holdout data
  - But we already know the prices of those 1000 diamonds(!)
  - We can thus measure the errors in predictions:
    - Estimate how well our model will perform, select the best model
    - Automatic resampling/"rotation" of holdout (`rms` package)

Cross-fold validation / holdout

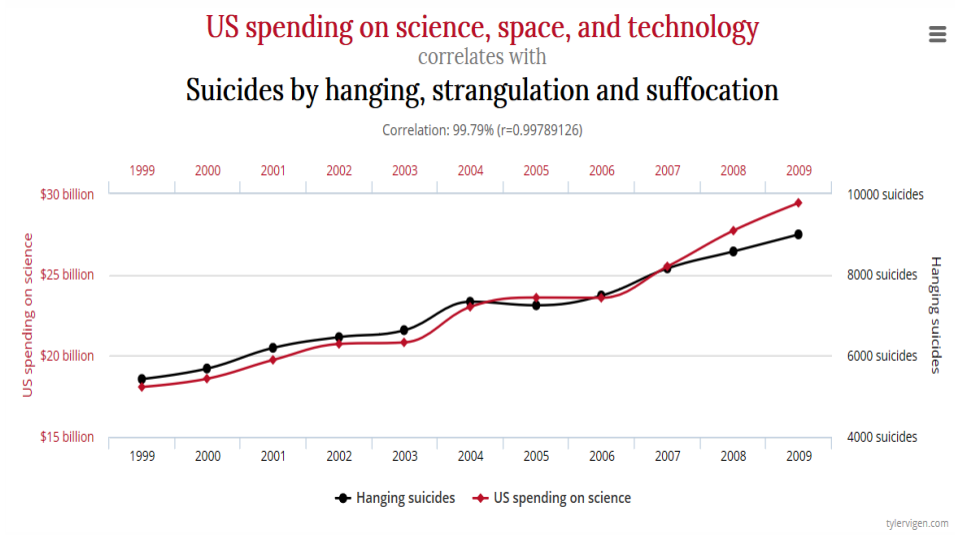
INSEAD
The Business School for the World®

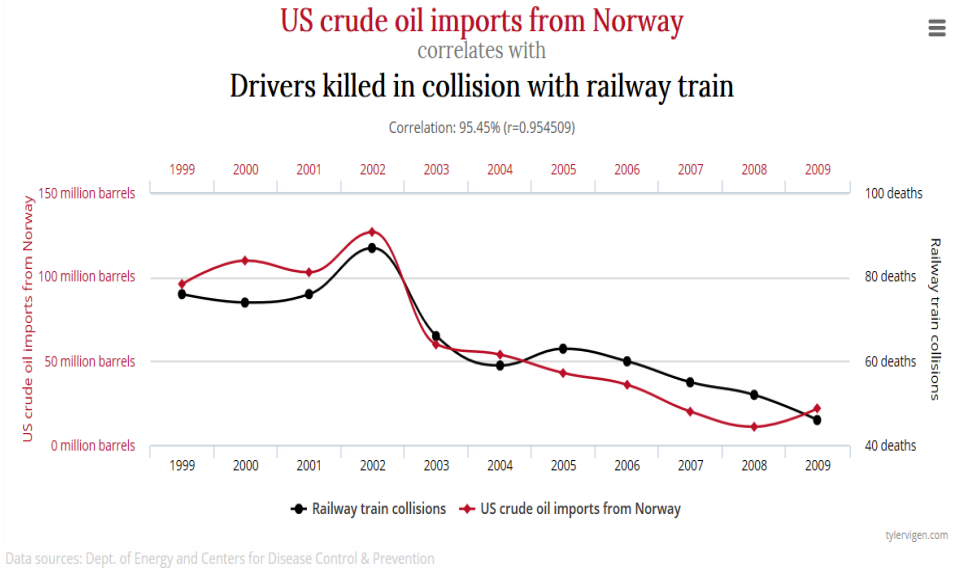
```
diamond.data.testing<-subset(diamond.data, (ID>=5001 & ID<=6000)) #withold 1000 datapoints into a "testing" data

diamond.data.training<-subset(diamond.data, ID<=5000) #redefine the training data


fit<-lm(Price~Carat.Weight+Cut+Color+Clarity+Polish+Symmetry+Report, data=diamond.data.training) #build a model on training data

predicted.prices.testing<-predict(fit, diamond.data.testing) #predict the prices of the 1000 diamonds left for testing the model

percent.errors <- abs((diamond.data.testing$Price-predicted.prices.testing)/diamond.data.testing$Price)*100 #calculate absolute percentage errors

mean(percent.errors) #display Mean Absolute Percentage Error (MAPE)


# repeat the same for the log model

fit.log<-lm(log(Price)~log(Carat.Weight)+Cut+Color+Clarity+Polish+Symmetry+Report, data=diamond.data.training)

predicted.prices.testing.log<-exp(predict(fit.log, diamond.data.testing))

percent.errors.log <- abs((diamond.data.testing$Price-predicted.prices.testing.log)/diamond.data.testing$Price)*100

mean(percent.errors.log)
```

Why holdout is so important? Spurious correlations/overfitting

INSEAD
The Business School for the World®

- In the world of "Big Data" it is not difficult to find variables that would correlate almost perfectly with any data ("overfitting"), without reflecting any underlying relationship - "spurious correlations"





Source: http://tylervigen.com/spurious-correlations

- The best model is thus **not** one that does best on training data (has high r^2 ~ overfitting), but one that does best on the holdout

Variable selection

INSEAD
The Business School for the World®

- In the world of "Big Data" 000s of variables can be easily mined/created (e.g., cut*color*clarity interactions result in 5*6*7=350 new variables).
- How to select which variables should be in the model?
- Forward/Backward/Stepwise regressions:

```
fit.log.step<-step(lm(log(Price)~log(Carat.Weight)*Color*Cut + … ,direction="backward")
summary(fit.log.step)
fit.log.step<-step(lm(log(Price)~log(Carat.Weight)*Color*Cut + …,direction="both")
summary(fit.log.step)
```

- Main idea (forward example): Find which single X explains most of Y (most correlated). Add X1 to the model. Given that X1 is already in the model, find which other variable adds most explanatory power. Add it and re-estimate the model. Repeat until no variable can be added
  - `step` function uses significance. More advanced: `stepAIC` ["Akaike information criterion"], LASSO/Ridge – penalties for many variables

Summary of Sessions 1-2

INSEAD
The Business School for the World®

- This course is about Data Science: most-demanded, highest-paid ("sexiest") skill in the MBA job market today
- You will learn the basics of several analytical (stats and machine learning) techniques, learn the basics of coding (in R, leading open-source soft), and most importantly, learn how to use this knowledge in business applications
- Key "Excel to R" learning: codified solutions to analytical problems are often easier than Excel-like equivalents and more powerful/advanced
  - More so, they can be automated to produce complete reports in minutes ("notebooks", *.Rmd files, TBD later in the course)
  - But coding is new to many of you; it will take practice, and we will move slowly. Tutorial 1 (**19:15 tomorrow, this room, 307**) will help with the basics of R
- … and if nothing else, you've learned my name and office ## ;)

Next…

INSEAD
The Business School for the World®

- Finish group formation (add/drop)

- Tutorial 1: (19:15 tomorrow) Getting comfortable with R
- Sessions 3-4: (Fri) Time-series analyses
  - Same as today, R code will be provided and we will work through it learning the underlying methods, how to code them, and how to use them
- Group Assignment 1: Yahoo's acquisition of Tumblr
  - Due 8:25am of ABSessions 5-6 morning (next Tue, Jan 30), upload to INSEAD portal


INSEAD
The Business School
for the World®
Europe | Asia | Middle East