DS501: Course Introduction
Welcome to
Introduction to Data Science!

Prof. Randy Paffenroth
rcpaffenroth@wpi.edu

Worcester Polytechnic Institute

WPI

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

WPI

But, does anyone really care?

- https://www.youtube.com/watch?v=3_1reLdh5xw

Ok, let the good time roll!

Ok, let the good time roll!
Oh, wait...

*Given a large mass of data, we can by judicious selection construct **perfectly plausible** unassailable theories—*
***all** of which, **some** of which, or **none** of which may be right.*
- Paul Arnold Srere

*"It's tough to make **predictions**, especially about the future."*

— Yogi Berra

WPI

Objectives for today

- Discuss the course mechanics (syllabus, grading, etc.)
- Start to get inspired about data science!
- Have a brief Python tutorial (depending on time...)

Basic course information

- Course number: DS501
- Course name: Statistical Methods for Data Science
- When/where:
  - Wednesdays from 6:00pm--8:50pm - HL154

WPI

Teaching Assistant/Grader

- TBD (depending on final class size)

WPI

Instructor information

- Randy Paffenroth (a.k.a. “Dr. Paffenroth”, “Prof. Paffenroth”, or “Randy”)
- Office location: AK124
- Office hours: 1-2pm on Tuesdays and 5-6pm on Wednesdays. **Other times are available by appointment, and walk-ins are always welcome if I am around and not otherwise indispos**
- Best ways to contact me:
  - WPI email: rcpaffenroth@wpi.edu
  - Gmail and Google hangouts: randy.paffenroth@gmail.com
  - Office phone: (508) 831-6562
- I should be able to turn around email questions relatively quickly 9am-5pm, Monday-Friday. My availability at night and on weekends is more limited and I certainly check my email far more infrequently, but you may feel free to try and contact me.

WPI

But, who am I really?

- I have two bachelor's degrees, one in Math and one in CS, from Boston University.
- I have a Ph.D. in applied mathematics from University of Maryland.
- Before coming to WPI I was a **Program Director** at a small company (50 people), and before that I worked at the California Institute of Technology and another small company (3 people!).

WPI

High level course goals and learning objectives

- This course provides an overview of Data Science, covering a broad selection of key challenges in and methodologies for working with big data.
- Topics to be covered include:
  - **Data Gathering**
  - **Data Storage**
  - **Business Intelligence**
  - **Basic Statistics**
  - **Machine Learning**
  - **Large Scale Data**
  - **Graph Data**
  - **Visualization**
  - **High Dimensional Data**
  - **Deep Learning**

WPI

High level course goals and learning objectives

- Professional skills, such as
  - communication,
  - presentation, and
  - storytelling with data, will be fostered.
- Students will acquire a working knowledge of data science through hands-on projects and **case studies** in a variety of business, engineering, social sciences, or life sciences domains.
- Issues of ethics, leadership, and teamwork are highlighted.

WPI

Computing background for the course

- You will need to be able get your hands dirty playing with, processing, and plotting data using the ***Python*** computer language!
  - and that will be the officially supported language for the course and all lecture examples will be in Python.
- Now, with that being said, this is not intended to be a programming course (i.e., your code will not be graded), but actually working with data will be extremely important (i.e., the results of the code will be graded)!

Python

- Python itself can be found at:
  - http://www.python..org
- We will also be making use of iPython/Jupyter notebooks. They makes developing Python code much easier. They can be found at:
  - http://jupyter.org
- Good place to start:
  - Learning Python By Mark Lutz O'Reilly Media, September 2013.
    - http://shop.oreilly.com/product/0636920028154.do
    - Available for free from the library.

WPI

Python

- There are also numerous Youtube videos
  - Though you have to be careful which ones you trust! :-)
  - https://www.youtube.com/watch?v=EUEHOYl0mRg

Why Python? a Data Science analysis...

- http://www.kdnuggets.com/2015/05/r-vs-python-data-science.html

Textbook

- **There is no official textbook for the course.**
  - Though, maybe we should write one someday :-).

Suggested texts

- Other texts that would be useful for the course are:
  - **Learning Python** by Mark Lutz O'Reilly Media, September 2013.
    - http://shop.oreilly.com/product/0636920028154.do
    - Available for free from the library.
  - **Python for Data Analysis** by Wes McKinney, October 2012.
    - http://shop.oreilly.com/product/0636920023784.do
    - Available for free from the library.
  - **Big Data: A Revolution That Will Transform How We Live, Work, and Think** by Viktor Mayer-Schönberger, et. al.
    - http://www.amazon.com/Big-Data-Revolution-Transform-Think/dp/0544227751/ref=sr_1_1?s=books&ie=UTF8&qid=1453149188&sr=1-1&keywords=big+data
  - T**he Fourth Paradigm: Data-Intensive Scientific Discovery**, by Tony Hey, et. al.
    - http://www.amazon.com/The-Fourth-Paradigm-Data-Intensive-Scientific/dp/0982544200

WPI

Course activities

- **Lectures**: The lectures and *in class discussions* are an important part of the course.
  - The base lecture notes will be posted on the class web page before after class along with any annotations made during class.
- **Case studies**: Doing the case-studies is how you get experience solving interesting problems.
  - The case studies are to be done in teams of 2-4, see the syllabus for details of the collaboration policy.
- **Exams**: There will be a midterm and final exam.
  - The final exam will be non-cumulative (i.e., the midterm exam will cover roughly the first half of the course and the final exam will cover roughly the second half of the course.

Suggestions for case studies

- Working on teams is hard… that is actually part of the point the exercise!
- There are various ways to organize yourselves…
  - Each case study will have multiple questions, so you could imagine each person working on one part (though some parts build on each other). However, that will likely not work best.
  - A better idea is to have teammates check each other's work. Alice does problems 1 and 3, Bob does problems 2 and 4, Alice double checks problems 2 and 4, and Bob double checks problems 1 and 3.
- I normally do not allow teams of size 2 and I think they are, in general, a bad idea.
  - However, I know that are people taking the class and working full time so I want to make allowances.

WPI

Course requirements and grading standards

Case Studies and Presentations (4 assignments)	50%
Midterm exam	20%
Final exam	30%

The midterm exam and final exam will be in class, noncumulative, and open note, but **no collaboration will be allowed** and the exams be graded based upon demonstrated understanding of key concepts. For each exam, you are allowed to bring in up to 4 8 ½ by 11 sheets of paper (either printed or handwritten) with whatever notes you want for the exam. The case studies will be performed in **groups of 2-4** and will be graded based upon the quality and completeness of presentations and the submitted reports.

I reserve the right to curve the final grades (either up or down) based upon the aggregate performance of the class.

WPI

Advice on doing the case studies

- ***You will be expected to hand in an iPython notebook with your code and a set of presentation slides for each case study.***
  - Make sure that the presentation slides are stand alone (i.e., I should not need to run your code to understand your results).
  - In addition, make sure that your code is well commented! If I can't figure out what you were doing then partial credit it hard to give.
- Depending on class size either all case study groups, or a subset, will be selected to present their work.
  - **You will not know if your team is going to be selected before the due date!**
- Make sure it is clear where **each part** of **each question** is answered.
  - Don't make the grader have to second guess which part of your write up answers which part of each question!
- The details are being finalized, but you will likely be grading each other on your project presentations.
- Remember, each case study is 12.5% of your final grade!

WPI

Important dates

	Assigned	Due
Case Study 1	January 27	February 10
Case Study 2	February 17	March 2
Case Study 3	March 23	April 6
Case Study 4	April 13	April 27

Midterm exam	March 2
Final exam	April 27

WPI

Collaboration and Academic Honesty Policy

Collaboration is prohibited on the exams. Collaboration is encouraged on case studies and you will be allowed to select your own teams of 2-4 for the the case studies. On case studies you **may** discuss problems across teams, but each team is responsible for generating solutions and writing up results on their own **from scratch**. All violations of the collaboration policy will be handled in accordance with the WPI Academic Honesty Policy.

WPI

Collaboration and Academic Honesty Policy

- As examples, each of the following would be a violation of the collaboration policy (this list is not exhaustive):
  - Two different homework teams share a solution to any assigned problem.
  - One homework or project team allows another homework or project team to copy any part of a solution to an assigned problem.
  - Any code or plots are shared between homework or project teams.
- As examples, each of the following would not be a violation of the collaboration policy:
  - Students within a team sharing solutions and code for a problem.
  - Students from different teams discussing an assignment at the level of goals, where ideas for solutions can be found in the book or notes, what parts are more challenging, or how one might approach the problem.
- Of course, you can ask Prof. Paffenroth or the TA any questions you like, show them code, etc.
- If there is any doubt as to what is allowed and what is not allowed, please just ask!

WPI

Student responsibilities and course policies

- Accommodation for Special Needs or Disabilities
  - If you need course adaptations or accommodations because of a disability, or if you have medical information to share with me, please make an appointment with me as soon as possible. If you have not already done so, students with disabilities who believe that they may need accommodations in this class are encouraged to contact the Office of Disability Services as soon as possible to ensure that such accommodations are implemented in a timely fashion. This office is located in the West St. House (157 West St), (508) 831-4908.
- Accommodation for Religious Observance
  - Students requiring accommodation for religious observance must make alternate arrangements with Prof. Paffenroth at least one week before the date in question.

WPI

Student responsibilities and course policies

- Personal Emergencies
  - In the event of a medical or family emergency, please contact Prof. Paffenroth to work out appropriate accommodations.
- Make-up Exam Policy
  - Make-up exams will only be allowed in the event of a documented emergency or religious observance. The exam dates are listed on the syllabus and you are responsible for avoiding conflicts with the exams.

WPI

Student responsibilities and course policies

- Late Assignment Policy
  - As the case studies are team based, and presentations will be expected on the day the case studies are due, **late case studies will not be accepted!**

WPI

Student responsibilities and course policies

Fairness!

http://www.whaleoil.co.nz/wp-content/uploads/2015/09/cat-begging.jpg

Dr. Paffenroth's keys to success!

- Do the case studies!
  - The case studies are perhaps the most important part of the learning experience in the class.
  - You need to get “your hands dirty”!
- Attend the lectures and ***ask questions***!
  - There are ***no*** dumb questions, and I can guarantee you that any question you want to ask will help your classmates as much as you.
  - I will also *post* the annotated slides after each class on myWPI.

WPI

Course syllabus available online, along with a lot of other stuff…

http://my.wpi.edu

WPI

What is Data Science?

- What do **you** think it is?
- What do you need to know to do it?
- How is it used?
- What got **you** interested in it?

What is Data Science? Some views...

What is Data Science? Some views...

...on any given day, a team member could author a multistage processing pipeline in Python, design a hypothesis test, perform a regression analysis over data samples with R, design and implement an algorithm for some data-intensive product or service in Hadoop, or communicate the results of our analyses to other members of the organization.

- Jeff Hammerbacher describing the data science group he put together at Facebook.
- What is Data Science?
An O'Reilly Radar Report – Mike Loukides

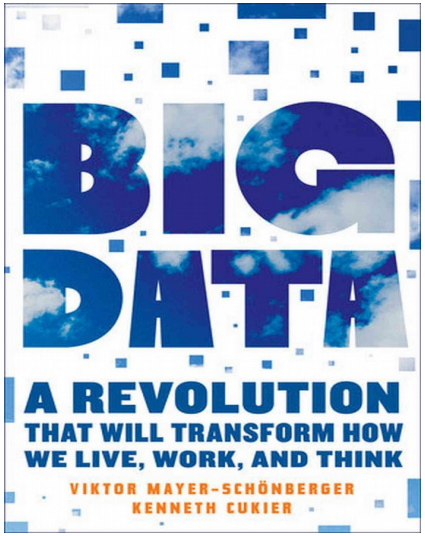
What is Data Science? Some views...

...on any given day, a team member could author a multistage processing pipeline in Python, design a hypothesis test, perform a regression analysis over data samples with R, design and implement an algorithm for some data-intensive product or service in Hadoop, or communicate the results of our analyses to other members of the organization.

- Jeff Hammerbacher describing the data science group he put together at Facebook.
- What is Data Science?
An O'Reilly Radar Report – Mike Loukides

"The "data scientist," which combines the skills of the statistician, software programmer, infographics designer, and storyteller."

- Big Data: A Revolution that Transform How We Live, Work, and Think, Viktor Mayer-Schönberger and Kenneth Cukier.

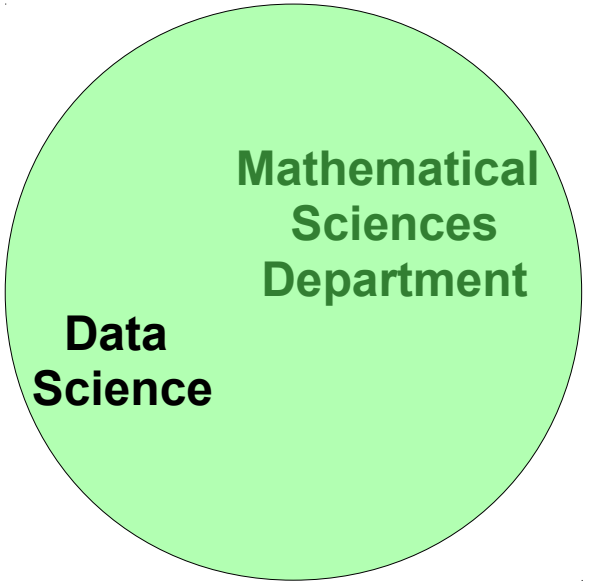


What is Data Science? Another view...

Data Science

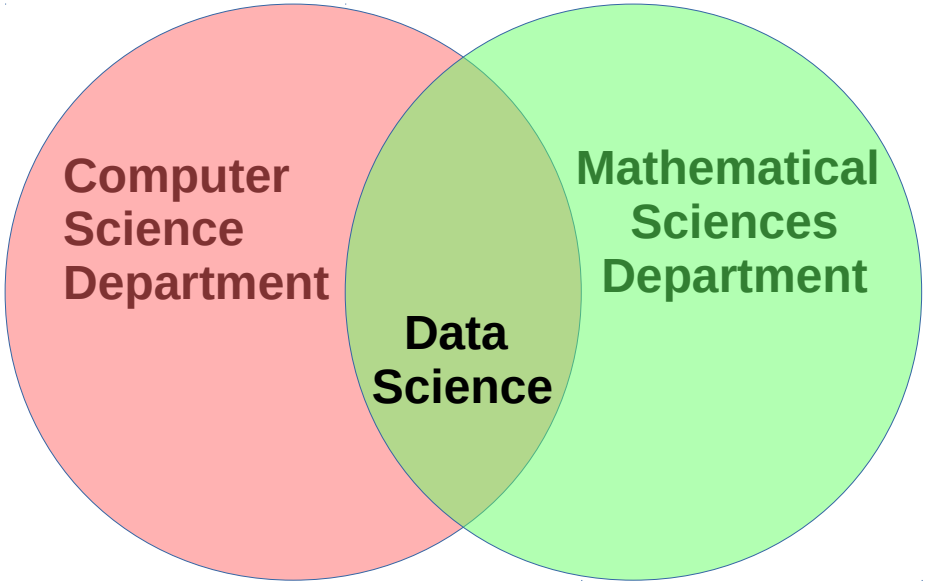
- Based upon Drew Conway's Data Science Venn Diagram
  - http://en.wikipedia.org/wiki/Data_science
  - http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

What is Data Science? Another view...



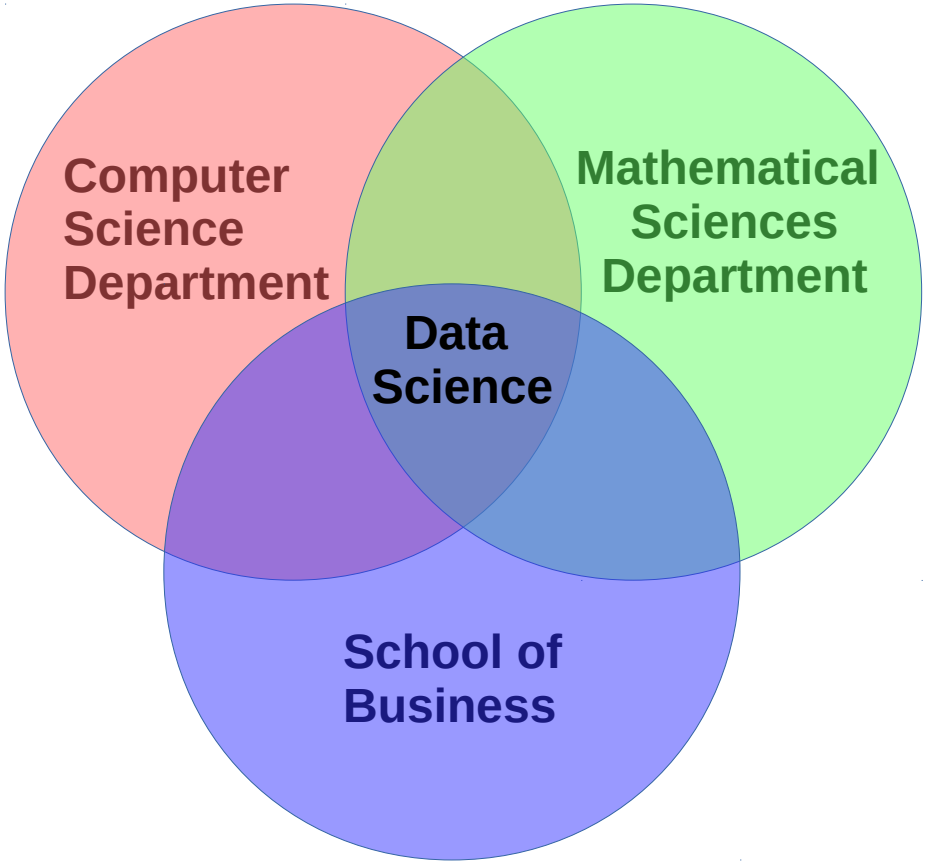
- Based upon Drew Conway's Data Science Venn Diagram
  - http://en.wikipedia.org/wiki/Data_science
  - http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

What is Data Science? Another view...



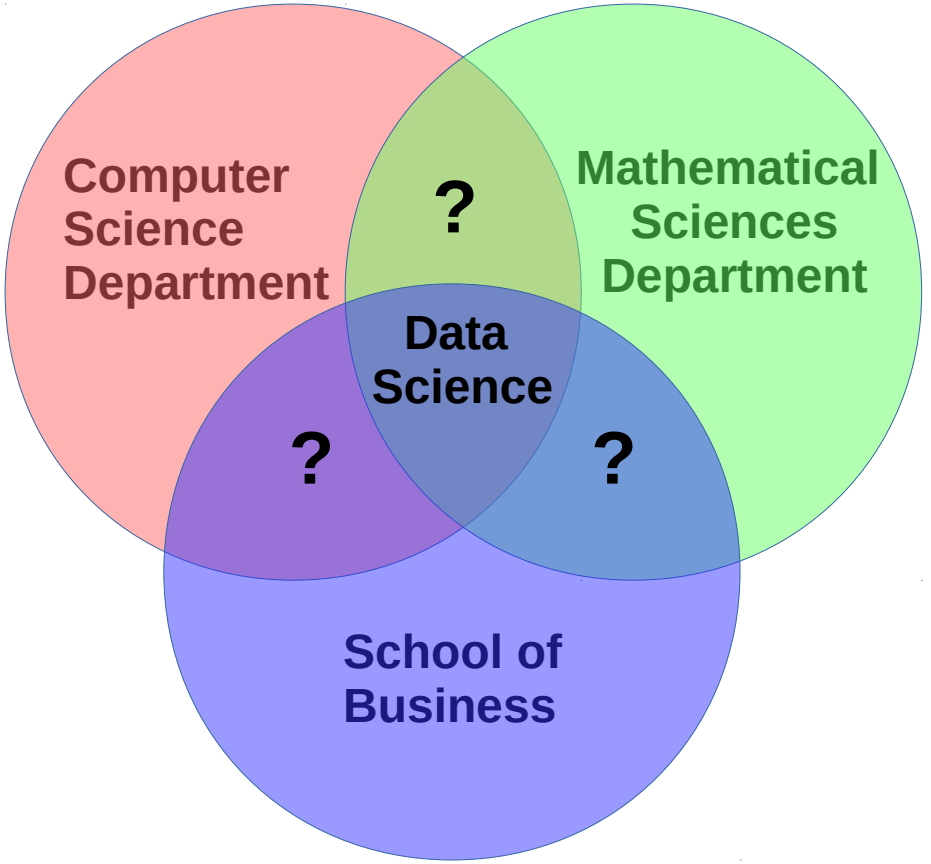
- Based upon Drew Conway's Data Science Venn Diagram
  - http://en.wikipedia.org/wiki/Data_science
  - http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

What is Data Science? Another view...



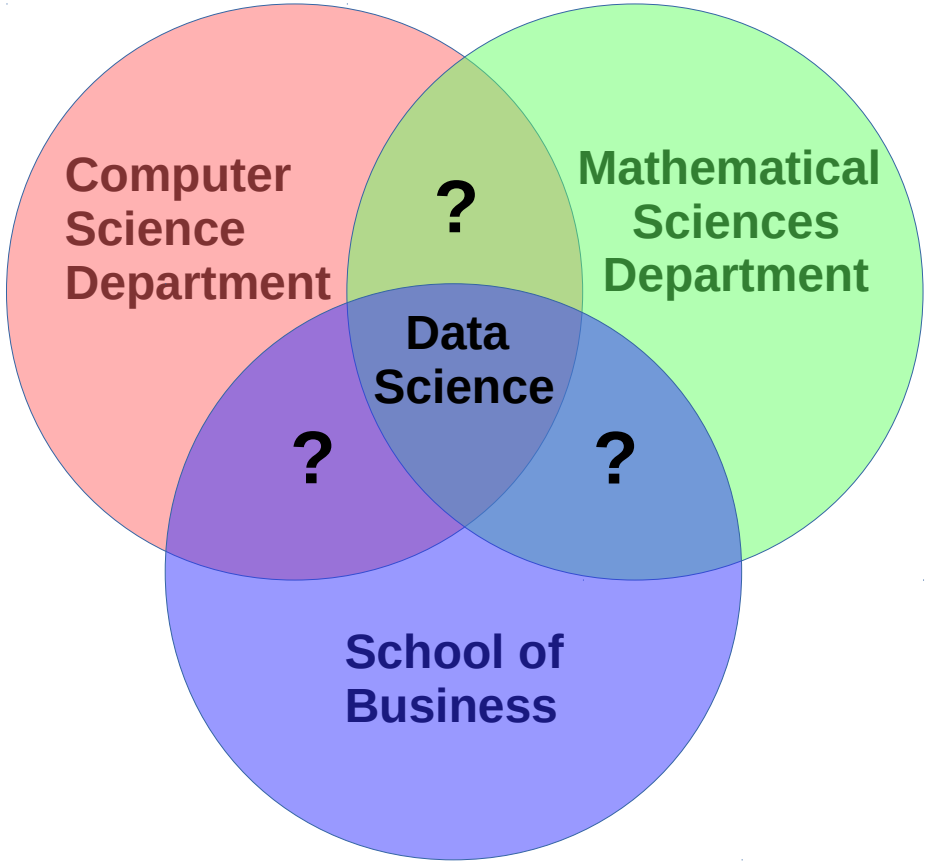
- Based upon Drew Conway's Data Science Venn Diagram
  - http://en.wikipedia.org/wiki/Data_science
  - http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

What is Data Science? Another view...



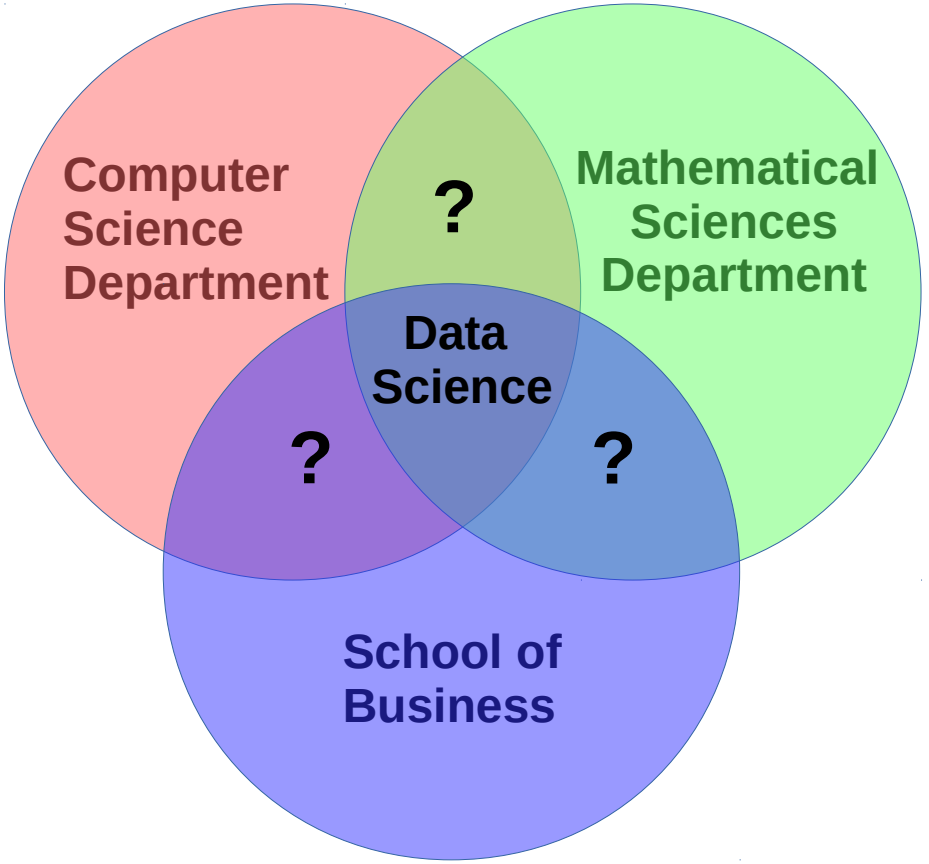
- Based upon Drew Conway's Data Science Venn Diagram
  - http://en.wikipedia.org/wiki/Data_science
  - http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Which is most important?



http://en.wikipedia.org/wiki/View_of_the_World_from_9th_Avenue

Which is most important?



http://en.wikipedia.org/wiki/View_of_the_World_from_9th_Avenue

How big is "data" these days?

- https://web-assets.domo.com/blog/wp-content/uploads/2015/08/15_domo_data-never-sleeps-3_full_v4.png

What is Big Data?

- The are many examples of "data", but what makes some of it “big”?

http://pl.wikipedia.org/wiki/Green_Giant#mediaviewer/Plik:Jolly_green_giant.jpg

What is Big Data?

- The are many examples of "data", but what makes some of it “big”? The classic definition revolves around the **three Vs**.
- **Volume, velocity, and variety.**

http://pl.wikipedia.org/wiki/Green_Giant#mediaviewer/Plik:Jolly_green_giant.jpg

What is Big Data?

- The are many examples of "data", but what makes some of it “big”? The classic definition revolves around the **three Vs**.
- **Volume, velocity, and variety.**
  - **Volume:** There is a just a lot of it being generated all the time. Things get interesting and “big”, when you can’t fit it all on one computer anymore. Why? There are many ideas here such as MapReduce, Hadoop, etc. that all revolve around being able to process data that goes from Terabytes, to Petabytes, to Exabytes.

http://pl.wikipedia.org/wiki/Green_Giant#mediaviewer/Plik:Jolly_green_giant.jpg

What is Big Data?

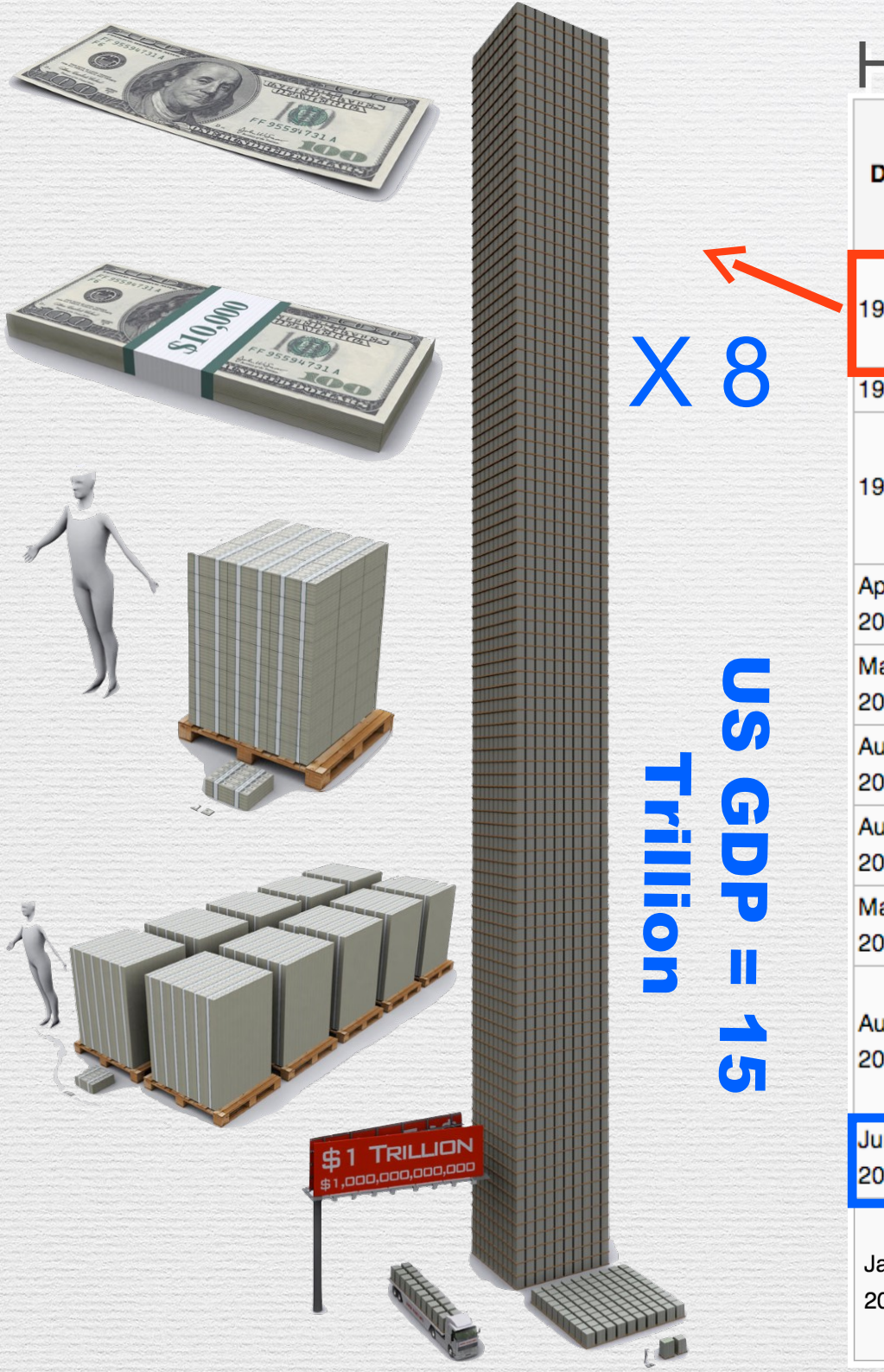
- The are many examples of "data", but what makes some of it “big”? The classic definition revolves around the **three Vs**.
- **Volume, velocity, and variety.**
  - **Volume:** There is a just a lot of it being generated all the time. Things get interesting and “big”, when you can’t fit it all on one computer anymore. Why? There are many ideas here such as MapReduce, Hadoop, etc. that all revolve around being able to process data that goes from Terabytes, to Petabytes, to Exabytes.
  - **Velocity:** Data is being generated very quickly. Can you even store it all? If not, then what do you get rid of and what do you keep?

http://pl.wikipedia.org/wiki/Green_Giant#mediaviewer/Plik:Jolly_green_giant.jpg

What is Big Data?

- The are many examples of "data", but what makes some of it “big”? The classic definition revolves around the **three Vs**.
- **Volume, velocity, and variety.**
  - **Volume:** There is a just a lot of it being generated all the time. Things get interesting and “big”, when you can’t fit it all on one computer anymore. Why? There are many ideas here such as MapReduce, Hadoop, etc. that all revolve around being able to process data that goes from Terabytes, to Petabytes, to Exabytes.
  - **Velocity:** Data is being generated very quickly. Can you even store it all? If not, then what do you get rid of and what do you keep?
  - **Variety:** The data types you mention all take different shapes. What does it mean to store them so that you can play with or compare them?

http://pl.wikipedia.org/wiki/Green_Giant#mediaviewer/Plik:Jolly_green_giant.jpg



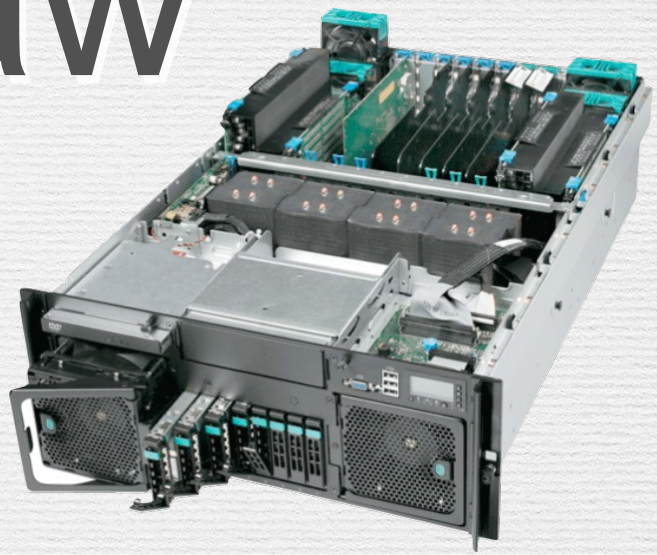
X 8

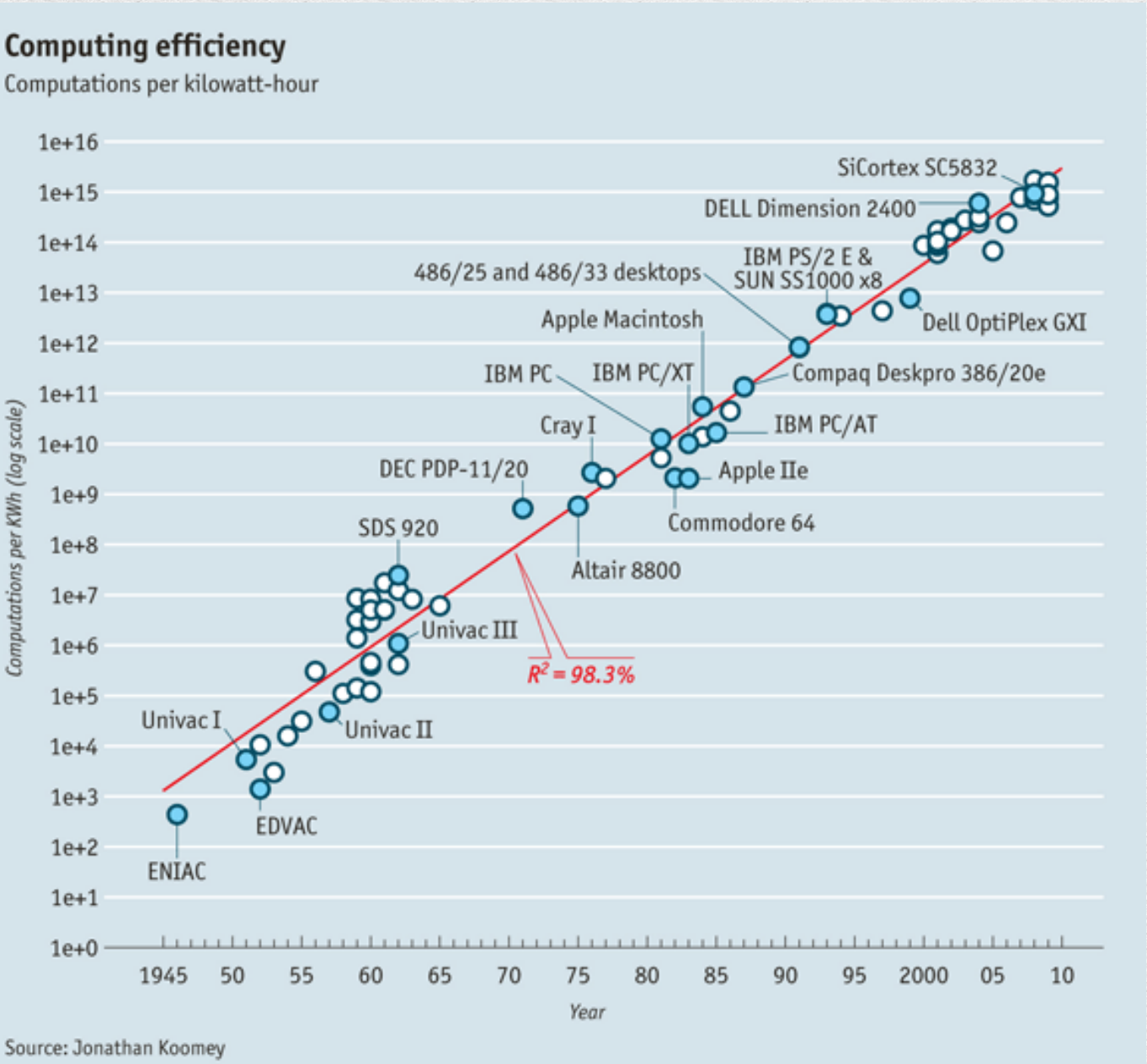
US GDP = 15 Trillion

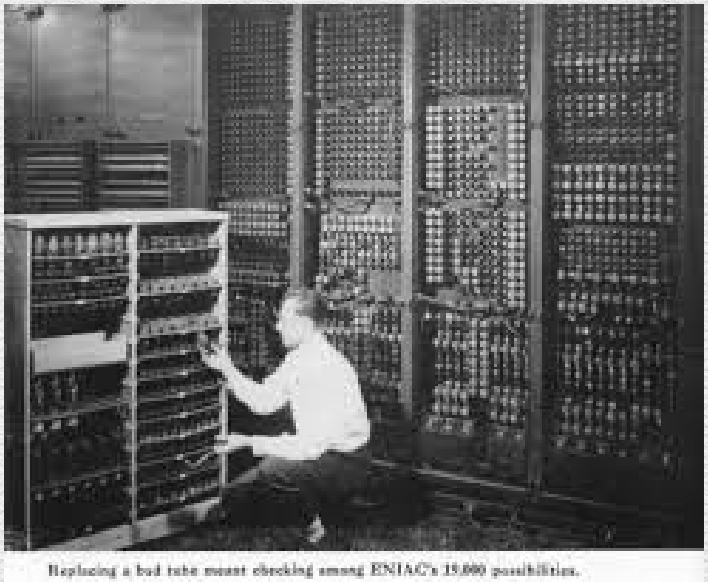
Hardware Cost for 1 GFlops

Date	Approximate cost per GFLOPS	Approximate cost per GFLOPS inflation adjusted to 2012 dollars[46]	Least expensive platform able to achieve 1 GFLOPS
1961	US $1,100,000,000,000 ($1.1 trillion)	US $8.3 trillion	About 17 million IBM 1620 units costing $64,000 each
1984	$15,000,000	$33,000,000	Cray X-MP
1997	$30,000	$42,000	Two 16-processor Beowulf clusters with Pentium Pro microprocessors[48]
April 2000	$1,000	$1,300	Bunyip Beowulf cluster
May 2000	$640	$836	KLAT2
August 2003	$82	$100	KASY0
August 2007	$48	$52	Microwulf
March 2011	$1.80	$1.80	HPU4Science
August 2012	$0.75	$0.73	Qua… Syste…
June 2013	$0.22	$0.22	Sony Playstation 4
January 2015	$0.08	$0.08	Celeron G1830 R9 295x2 System

Moore's Law







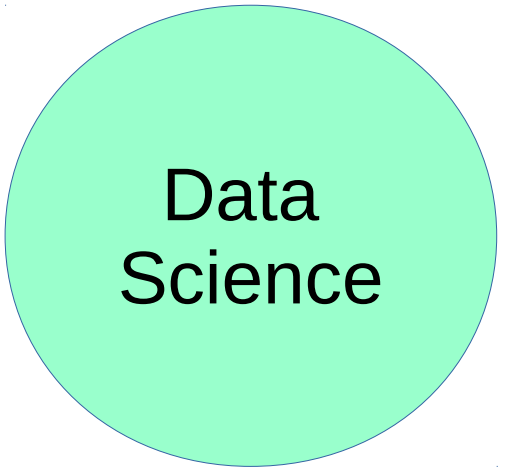
Replacing a bad tube meant checking among ENIAC's 19,000 possibilities.

What do they have to do with each other?

- Are Big Data and Data Science the same thing?

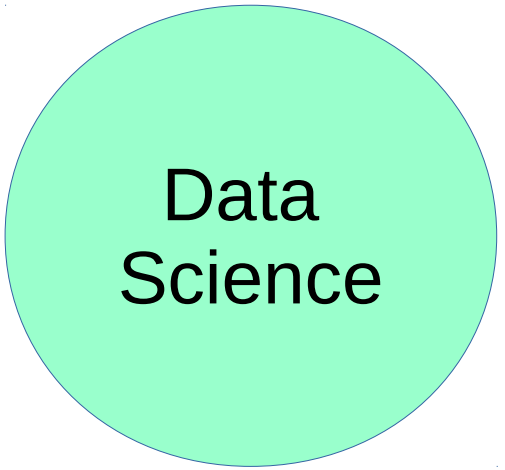
What do they have to do with each other?

- Are Big Data and Data Science the same thing?
  - I wouldn't say so...



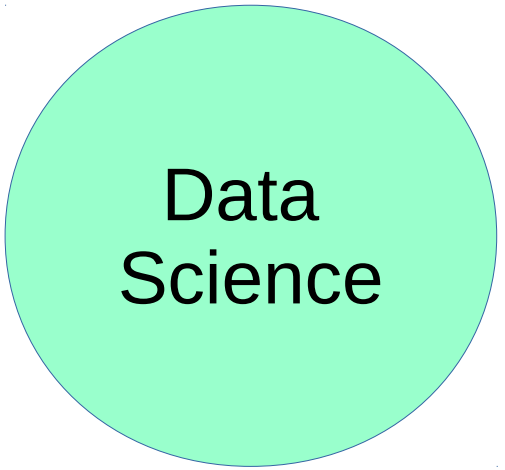
What do they have to do with each other?

- Are Big Data and Data Science the same thing?
  - I wouldn't say so...
  - Data Science can be done on small data sets.



What do they have to do with each other?

- Are Big Data and Data Science the same thing?
  - I wouldn't say so...
  - Data Science can be done on small data sets.
  - And not everything done using Big Data would necessarily be called Data Science.



What do they have to do with each other?

- Are Big Data and Data Science the same thing?
  - I wouldn't say so...
  - Data Science can be done on small data sets.
  - And not everything done using Big Data would necessarily be called Data Science.
  - But there certainly is a substantial overlap!

The good

Experiments, observations, and numerical simulations in many areas of science and business are currently generating terabytes of data, and in some cases are on the verge of generating petabytes and beyond. Analyses of the information contained in these data sets have already led to major breakthroughs in fields ranging from genomics to astronomy and high-energy physics and to the development of new information-based industries.

- Frontiers in Massive Data Analysis, National Research Council of the National Academies

The good

Experiments, observations, and numerical simulations in many areas of science and business are currently generating terabytes of data, and in some cases are on the verge of generating petabytes and beyond. Analyses of the information contained in these data sets have already led to major breakthroughs in fields ranging from genomics to astronomy and high-energy physics and to the development of new information-based industries.

- Frontiers in Massive Data Analysis, National Research Council of the National Academies

There will be a shortage of talent necessary for organizations to take advantage of big data. By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions.

- http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

The good

Experiments, observations, and numerical simulations in many areas of science and business are currently generating terabytes of data, and in some cases are on the verge of generating petabytes and beyond. Analyses of the information contained in these data sets have already led to major breakthroughs in fields ranging from genomics to astronomy and high-energy physics and to the development of new information-based industries.

- Frontiers in Massive Data Analysis, National Research Council of the National Academies

There will be a shortage of talent necessary for organizations to take advantage of big data. By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions.

- http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it.

- Hal Varian, Google's Chief Economist, http://www.mckinsey.com/insights/innovation/hal_varian_on_how_the_web_challenges_managers

The bad

Statistical rigor is necessary to justify the inferential leap from data to knowledge, and many difficulties arise in attempting to bring statistical principles to bear on massive data. Overlooking this foundation may yield results that are not useful at best, or harmful at worst. In any discussion of massive data and inference, it is essential to be aware that it is quite possible to turn data into something resembling knowledge when actually it is not. Moreover, it can be quite difficult to know that this has happened.

- Frontiers in Massive Data Analysis, National Research Council of the National Academies

The ugly

Would you sign the blueprint?

- Ivo Babuška, private communications

So, there is one thing that I really wany to stress.

- The three V's are all extremely important and make a Data Scientists job interesting and important, but I want to push for a 4th V!

So, there is one thing that I really wany to stress.

- The three V's are all extremely important and make a Data Scientists job interesting and important, but I want to push for a 4th V!
  - Veracity:

So, there is one thing that I really wany to stress.

- The three V's are all extremely important and make a Data Scientists job interesting and important, but I want to push for a 4th V!

  - Veracity:
    - Ok you have made a prediction, do you bet the farm (or your job on it)?

So, there is one thing that I really wany to stress.

- The three V's are all extremely important and make a Data Scientists job interesting and important, but I want to push for a 4th V!

  - Veracity:
    - Ok you have made a prediction, do you bet the farm (or your job on it)?
    - Or, maybe you do an *experiment* to see if the predictions you are making are correct?

So, there is one thing that I really wany to stress.

- The three V's are all extremely important and make a Data Scientists job interesting and important, but I want to push for a 4th V!

  - Veracity:
    - Ok you have made a prediction, do you bet the farm (or your job on it)?
    - Or, maybe you do an *experiment* to see if the predictions you are making are correct?
    - Is one experiment enough to bet the farm?

So, there is one thing that I really wany to stress.

- The three V's are all extremely important and make a Data Scientists job interesting and important, but I want to push for a 4th V!

  - Veracity:
    - Ok you have made a prediction, do you bet the farm (or your job on it)?
    - Or, maybe you do an *experiment* to see if the predictions you are making are correct?
    - Is one experiment enough to bet the farm?
    - Is a million?

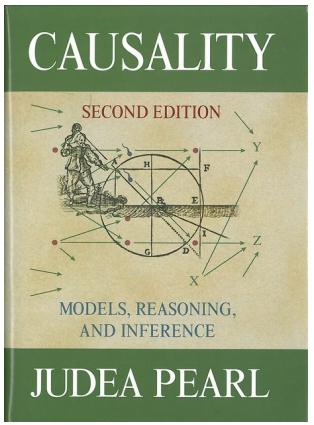
So, there is one thing that I really wany to stress.

- The three V's are all extremely important and make a Data Scientists job interesting and important, but I want to push for a 4th V!

  - Veracity:
    - Ok you have made a prediction, do you bet the farm (or your job on it)?
    - Or, maybe you do an *experiment* to see if the predictions you are making are correct?
    - Is one experiment enough to bet the farm?
    - Is a million?
    - How do you ***think critically*** about data, and all the things that go along with it?

But... what does getting "knowledge" from data really mean? Are we searching for causality?

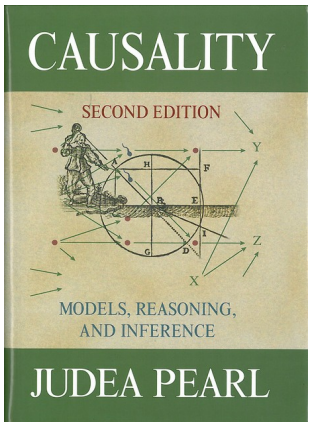
But... what does getting "knowledge" from data really mean? Are we searching for causality?



"Causation: The relation between mosquitoes and mosquito bites. Easily understood by both parties but never satisfactorily defined by philosophers and scientists."

- http://freshspectrum.com/causation/ Michael Scriven, Evaluation Thesaurus, 1991

But... what does getting "knowledge" from data really mean? Are we searching for causality?



"Causation: The relation between mosquitoes and mosquito bites. Easily understood by both parties but never satisfactorily defined by philosophers and scientists."

- http://freshspectrum.com/causation/ Michael Scriven, Evaluation Thesaurus, 1991

Most strikingly, society will need to shed some of its obsession for causality in exchange for simple correlations: not knowing why but only what.

- Big Data: A Revolution that Transform How We Live, Work, and Think, Viktor Mayer-Schönberger and Kenneth Cukier.

Can you even be *certain*?

Can you even be *certain*?

- For real world problems, I claim that you will never be certain of any inferences from data.

Can you even be *certain*?

- For real world problems, I claim that you will never be certain of any inferences from data.
    - I mean, what happens to your carefully thought out marketing plan for some rocking slacks when the Martians land.

Can you even be *certain*?

- For real world problems, I claim that you will never be certain of any inferences from data.
    - I mean, what happens to your carefully thought out marketing plan for some rocking slacks when the Martians land.
- What is *unacceptable* is when the data you actually have does not support the conclusion you report.

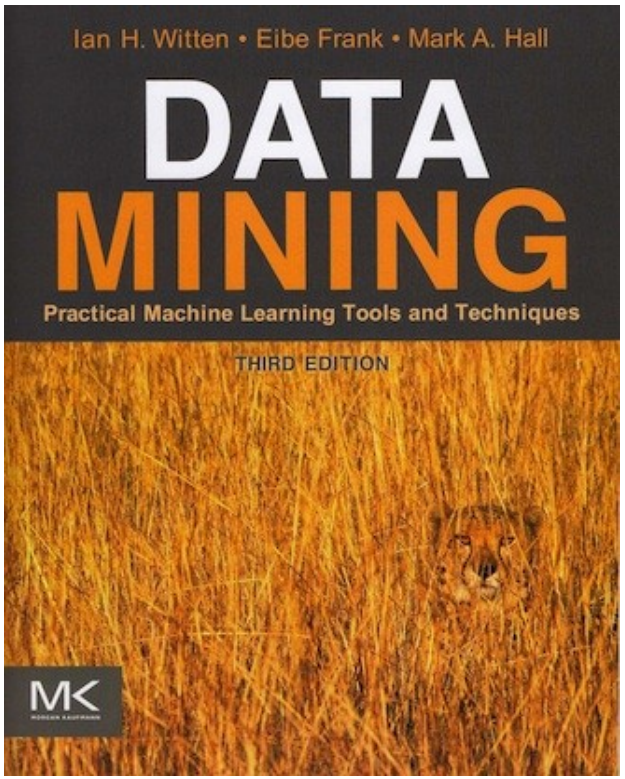
Can you even be *certain*?

You will never know (in time) whether you were right... but you could have known when you where wrong.

It can be easy to fool yourself!

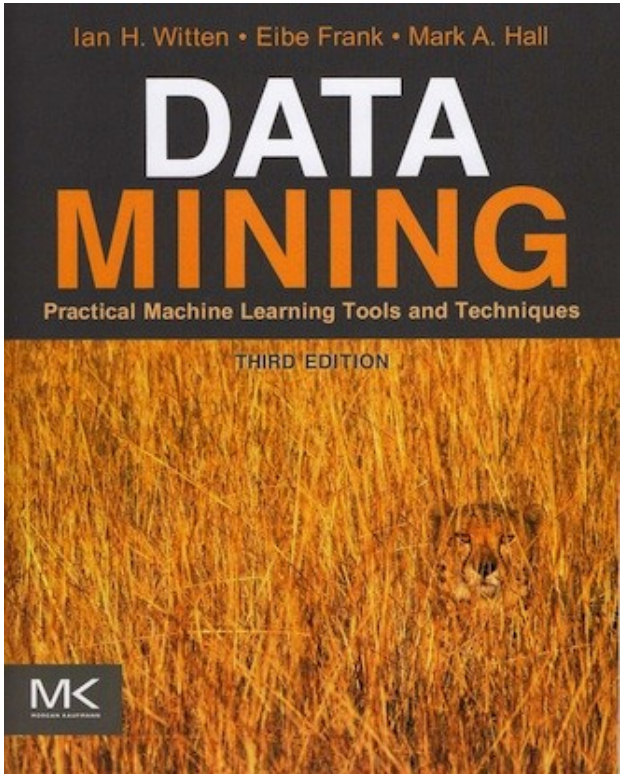
It can be easy to fool yourself!

Human beings are really good at pattern detection...



It can be easy to fool yourself!

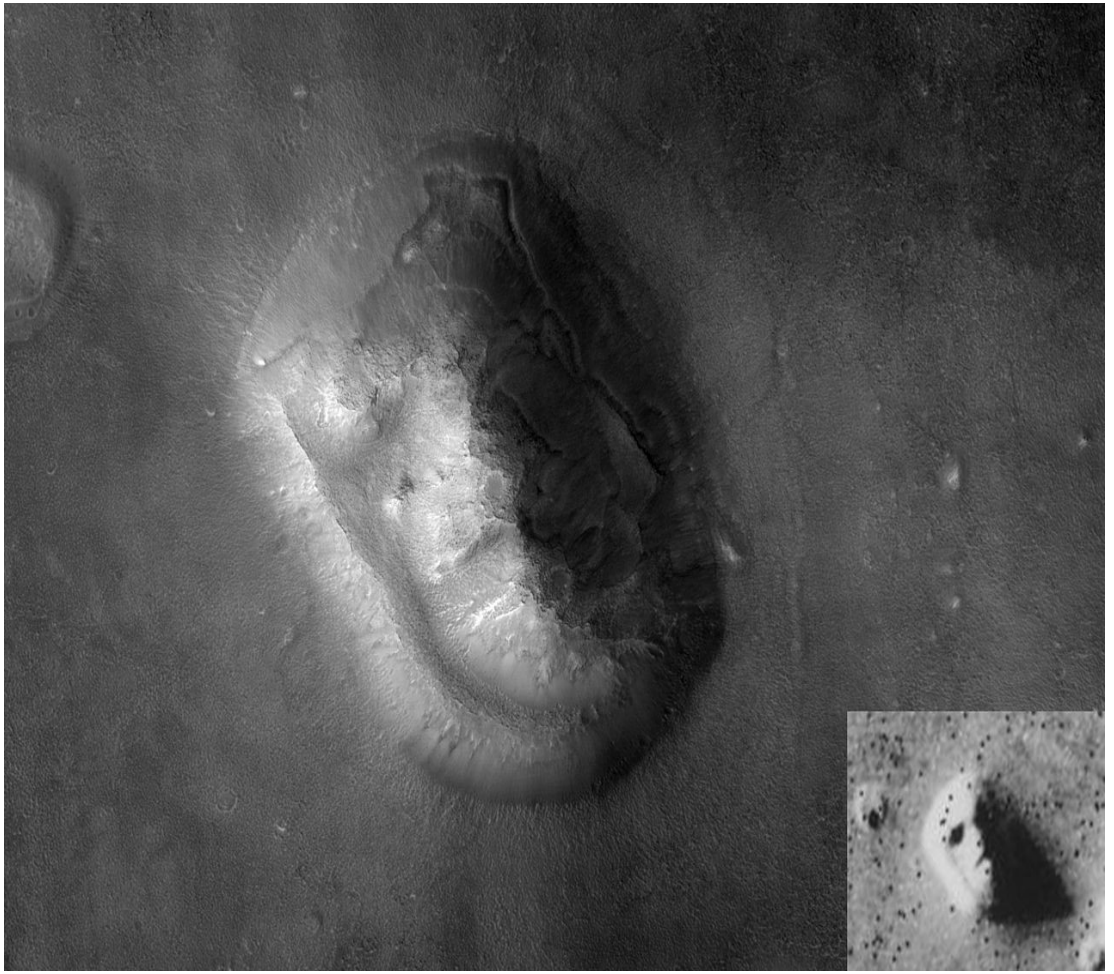
Human beings are really good at pattern detection...



Perhaps a bit too good!

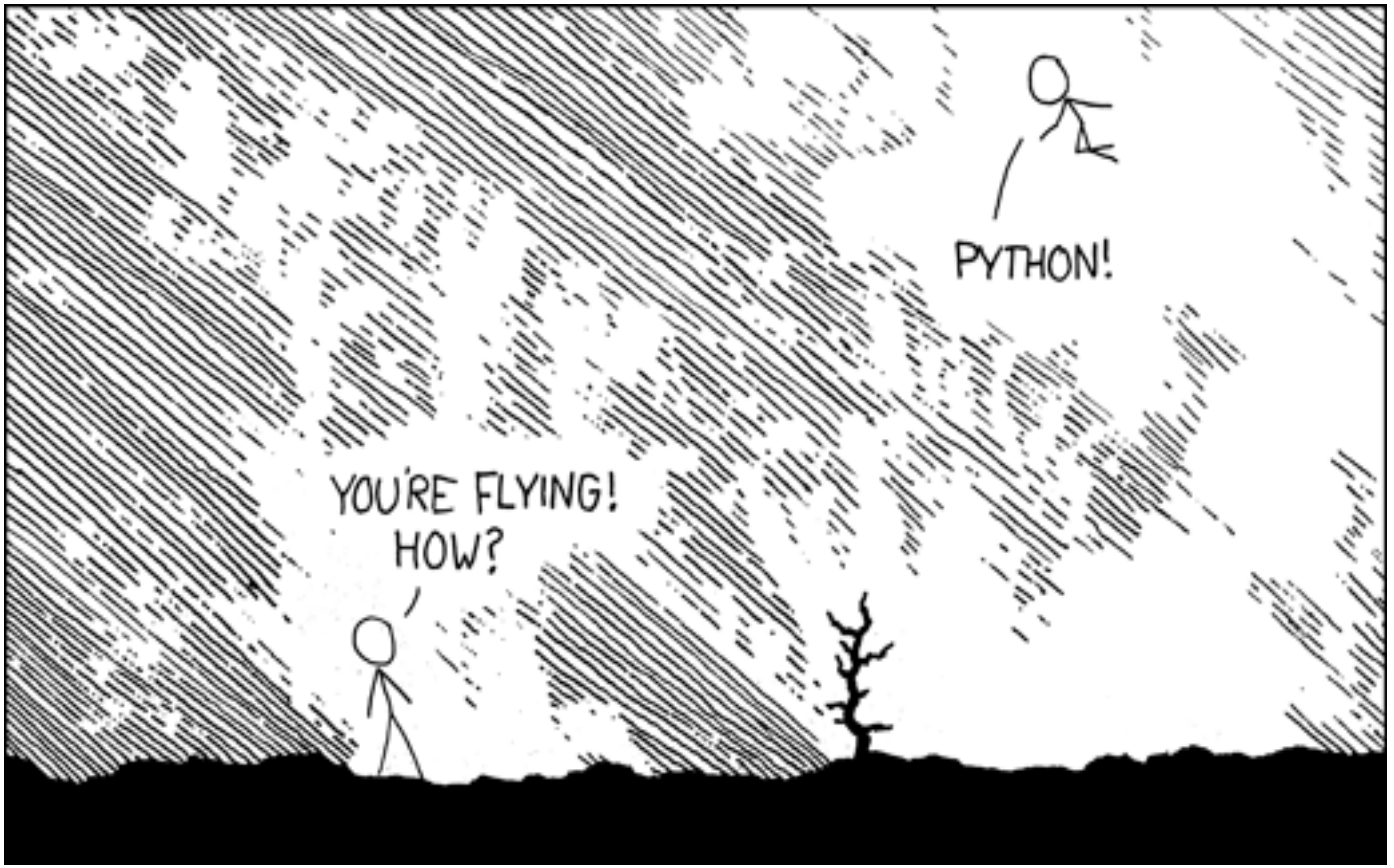
http://en.wikipedia.org/wiki/Cydonia_(region_of_Mars)

It can be easy to fool yourself!



http://en.wikipedia.org/wiki/Cydonia_(region_of_Mars)

Python tutorial



http://xkcd.com/353/

Important web pages

- Python documentation:
  - https://docs.python.org/2/
- IPython
  - http://ipython.org/
- NumPy
  - http://docs.scipy.org/doc/numpy/user/index.html
- Matplotlib
  - http://matplotlib.org/
- Scipy
  - http://www.scipy.org/

I prefer the 2.7.x versions

- 2.x is more stable.
- 2.x has more libraries for them.
- However, 3.x is also quite nice.

Good news, there is an easier way!

- Enthought Canopy!
  - https://www.enthought.com/products/canopy/
- Enthought provides a complete Python development environment for Windows and Mac (and Linux too, but if you use Linux you probably don't need it :-)
  - Which will be the focus today.

Problem one: A pattern!

Carl

- Likes Coca-cola? Yes
- Age 20-35? Yes
- Height > 6'0"? No
- Facebook friends > 30? Yes
- High school sports? Yes
- Math Ph.D.? No
- Married? No
- Monday Night Football fan? **???**

Problem one: A pattern!

Carl

- Likes Coca-cola? Yes
- Age 20-35? Yes
- Height > 6'0"? No
- Facebook friends > 30? Yes
- High school sports? Yes
- Math Ph.D.? No
- Married? No
- Monday Night Football fan? **???**

Joe

- Likes Coca-cola? Yes
- Age 20-35? Yes
- Height > 6'0"? No
- Facebook friends > 30? Yes
- High school sports? Yes
- Math Ph.D.? No
- Married? No
- Monday night football fan? **Yes**

Problem one: A pattern!

Carl

- Likes Coca-cola? Yes
- Age 20-35? Yes
- Height > 6'0"? No
- Facebook friends > 30? Yes
- High school sports? Yes
- Math Ph.D.? No
- Married? No
- Monday Night Football fan? **???**

Joe

- Likes Coca-cola? Yes
- Age 20-35? Yes
- Height > 6'0"? No
- Facebook friends > 30? Yes
- High school sports? Yes
- Math Ph.D.? No
- Married? No
- Monday night football fan? **Yes**

• Having found Joe, who is such a perfect match for Carl, do you feel justified in predicting that Carl likes Monday Night Football fan? Why or why not?

Problem one: A pattern!

Carl

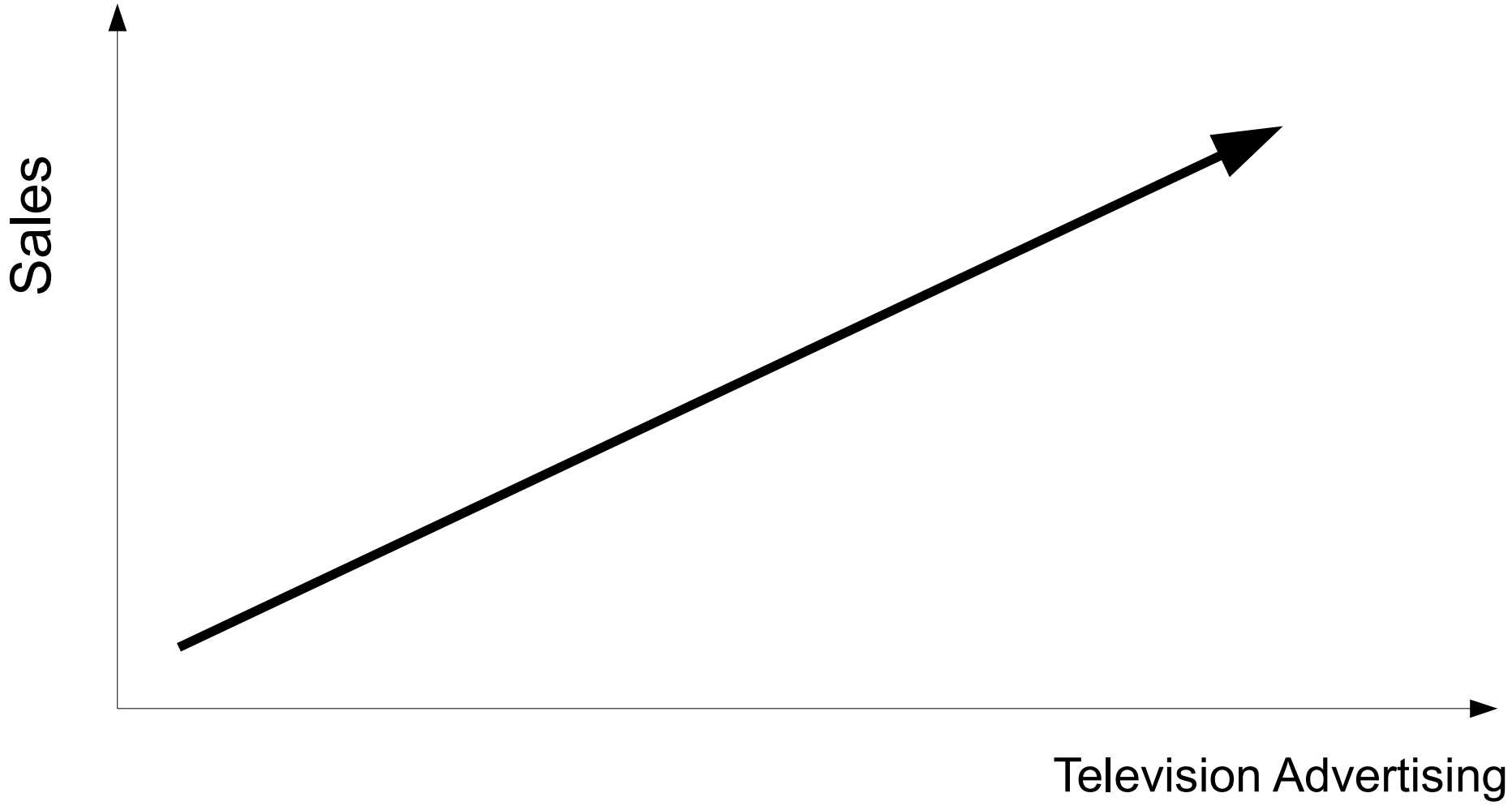
- Likes Coca-cola? Yes
- Age 20-35? Yes
- Height > 6'0"? No
- Facebook friends > 30? Yes
- High school sports? Yes
- Math Ph.D.? No
- Married? No
- Monday Night Football fan? **???**

Joe

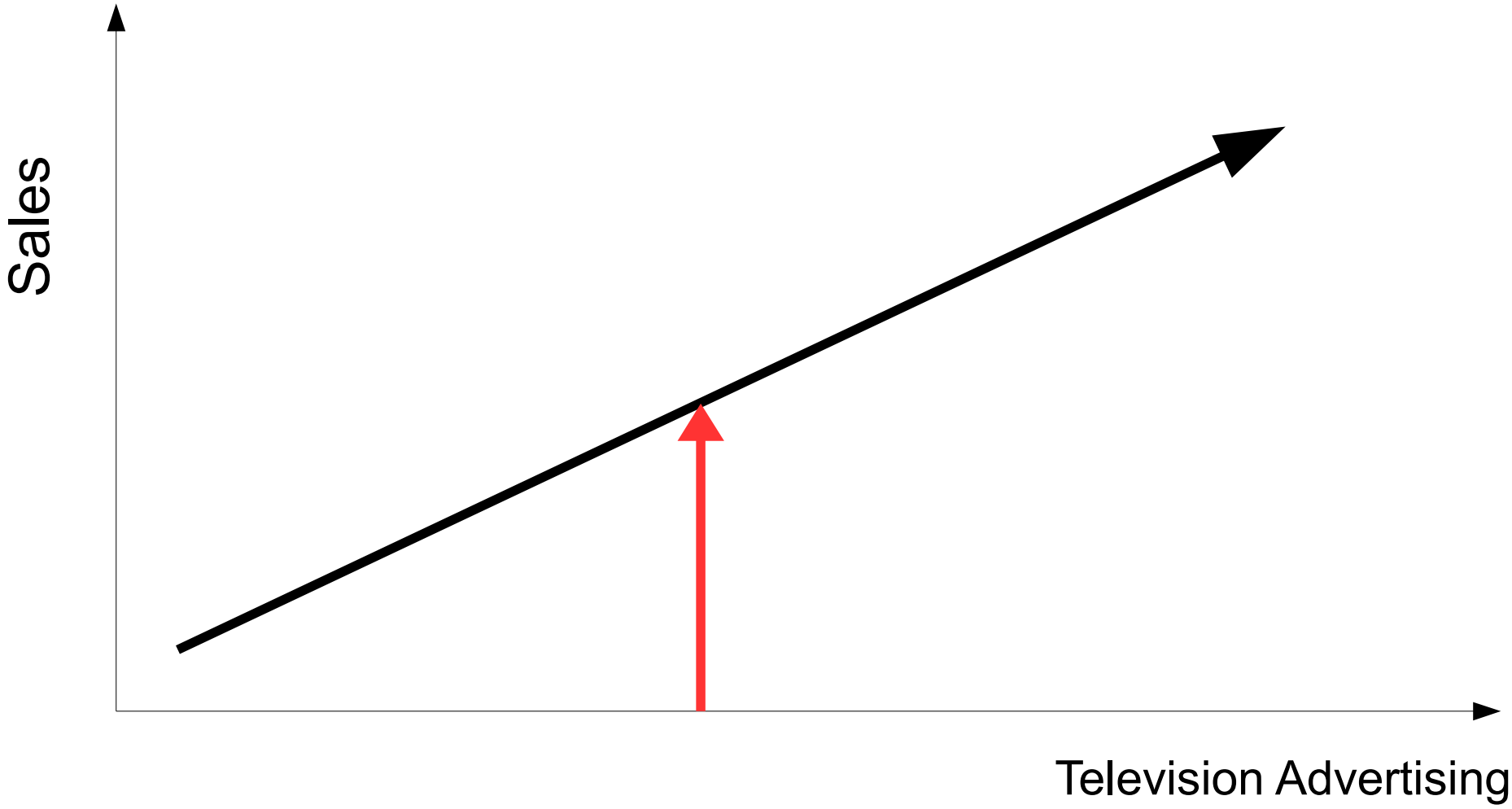
- Likes Coca-cola? Yes
- Age 20-35? Yes
- Height > 6'0"? No
- Facebook friends > 30? Yes
- High school sports? Yes
- Math Ph.D.? No
- Married? No
- Monday night football fan? **Yes**

• Having found Joe, who is such a perfect match for Carl, do you feel justified in predicting that Carl likes Monday Night Football fan? Why or why not?
• Suppose that your money is on the line for this prediction. What other questions might you ask?

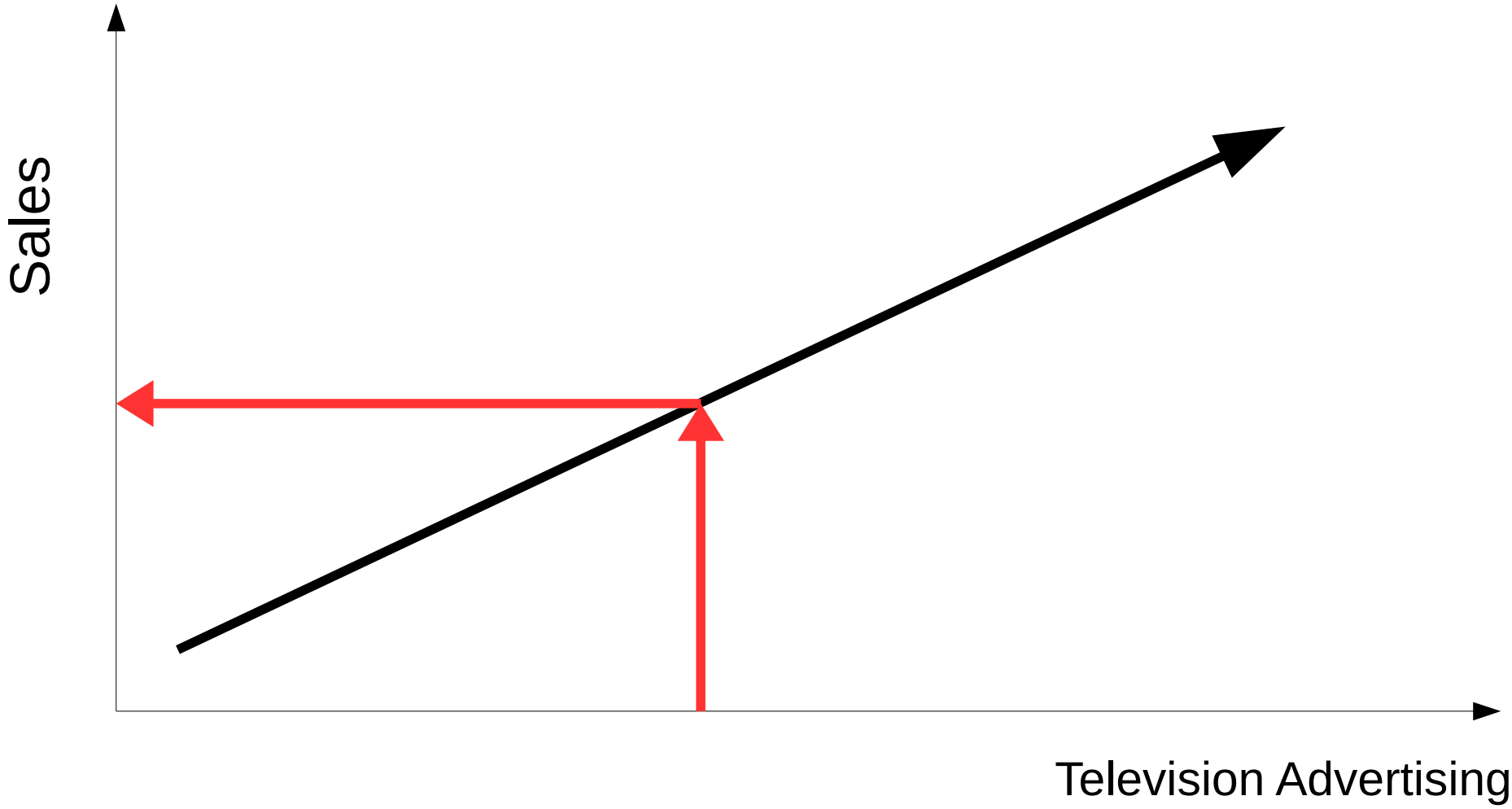
Problem two: We can save the company if we can just get our advertising right!



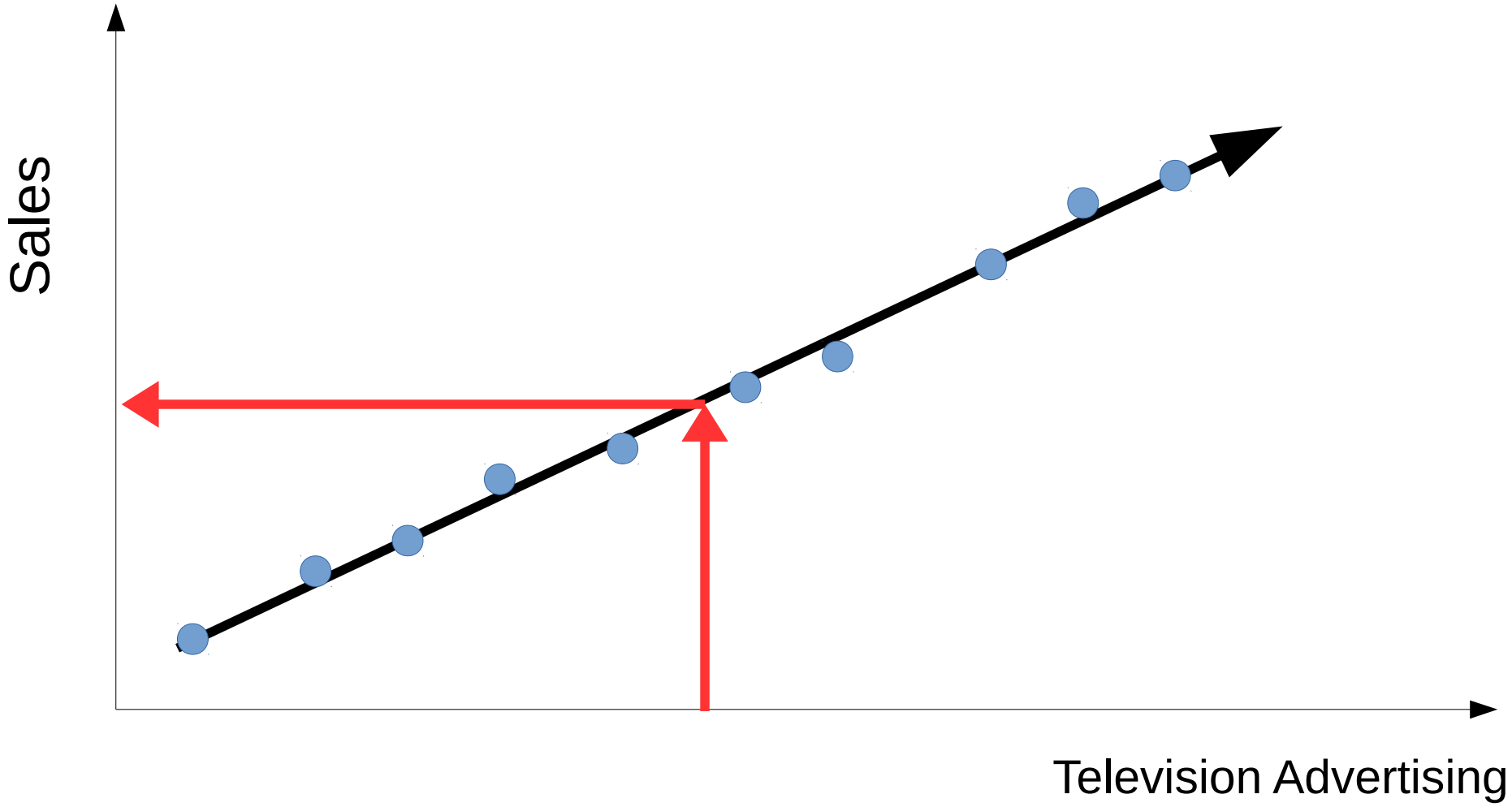
Problem two: We can save the company if we can just get our advertising right!



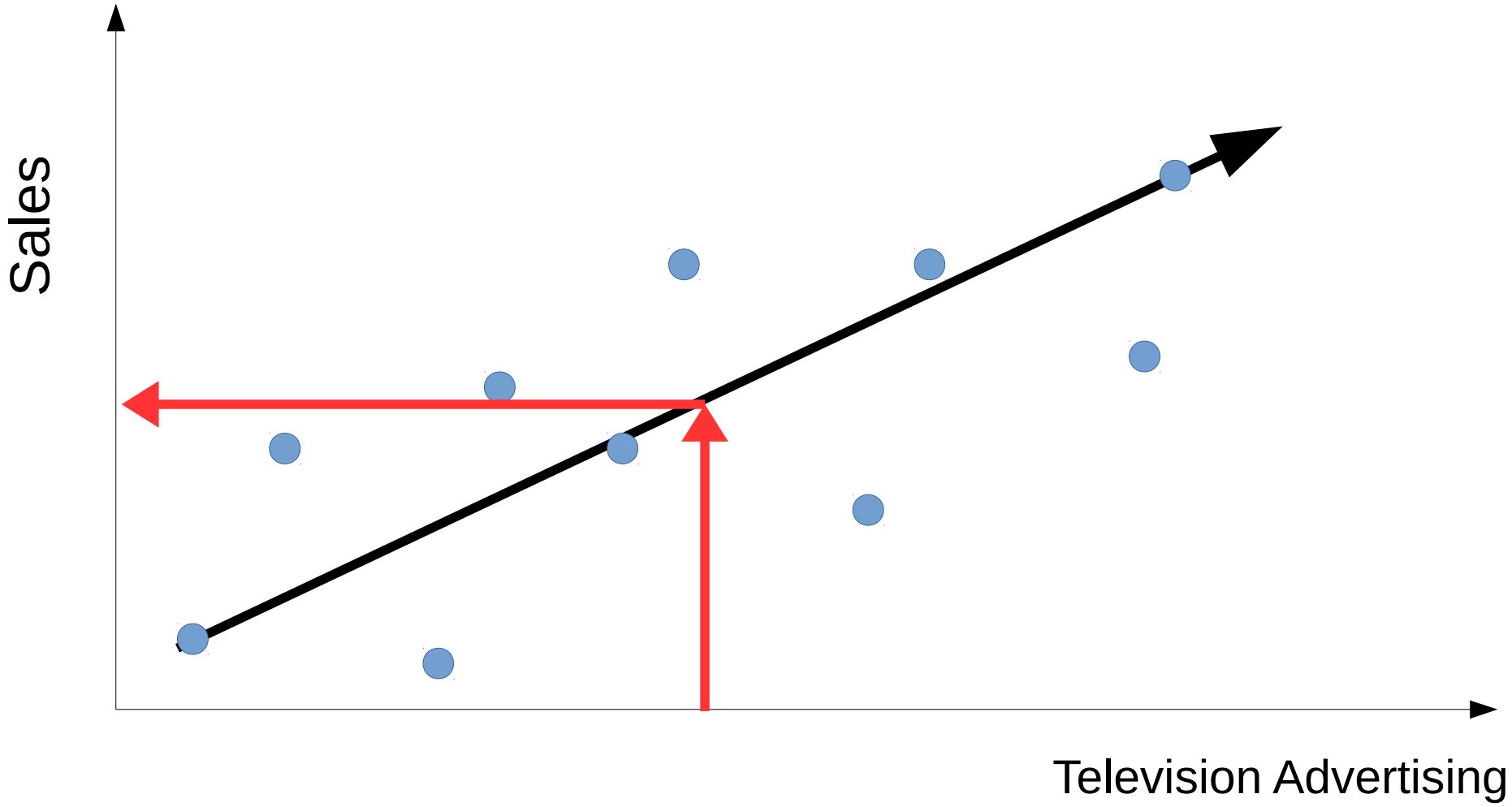
Problem two: We can save the company if we can just get our advertising right!



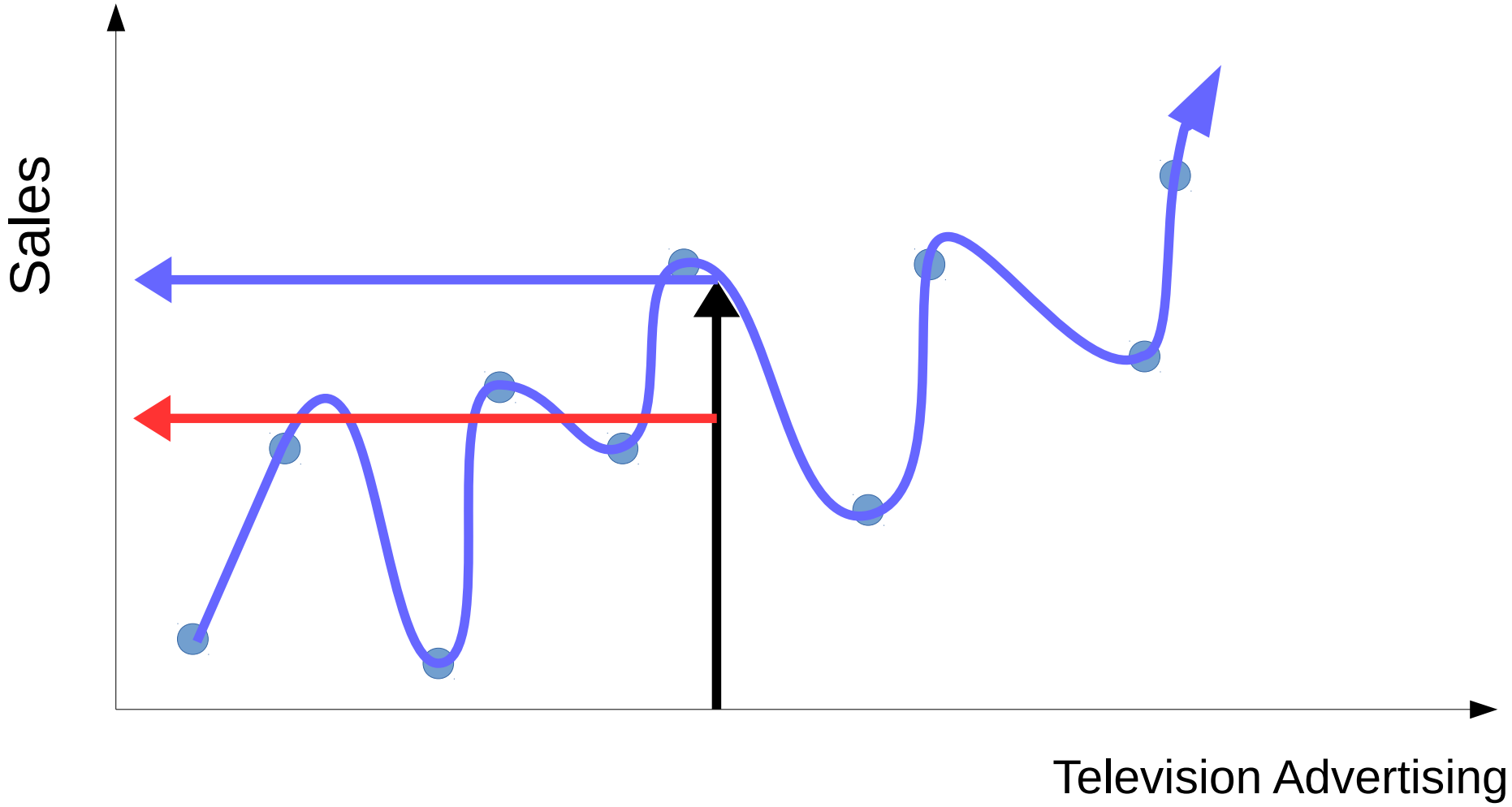
Problem two: We can save the company if we can just get our advertising right!



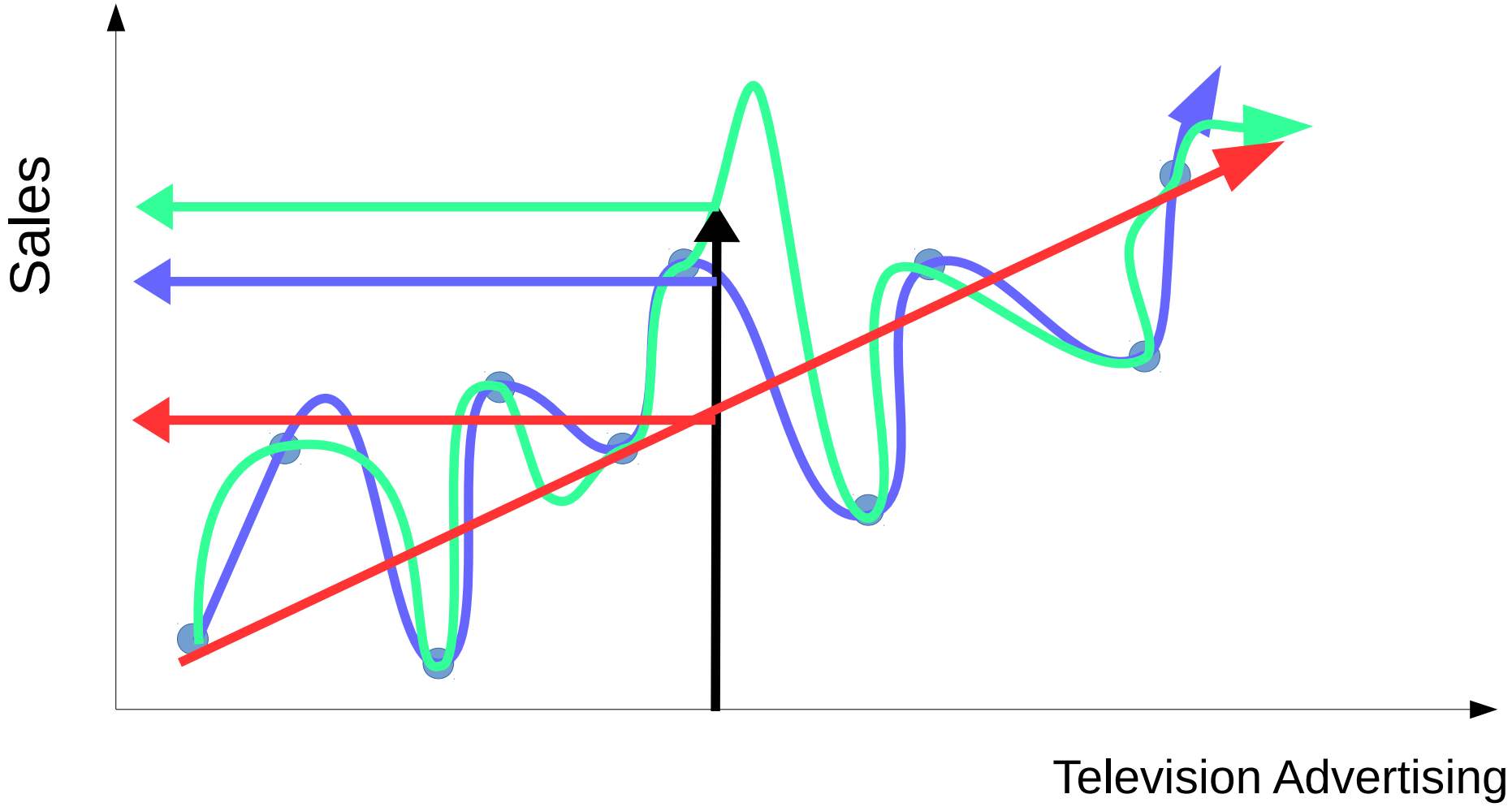
Problem two: We can save the company if we can just get our advertising right!



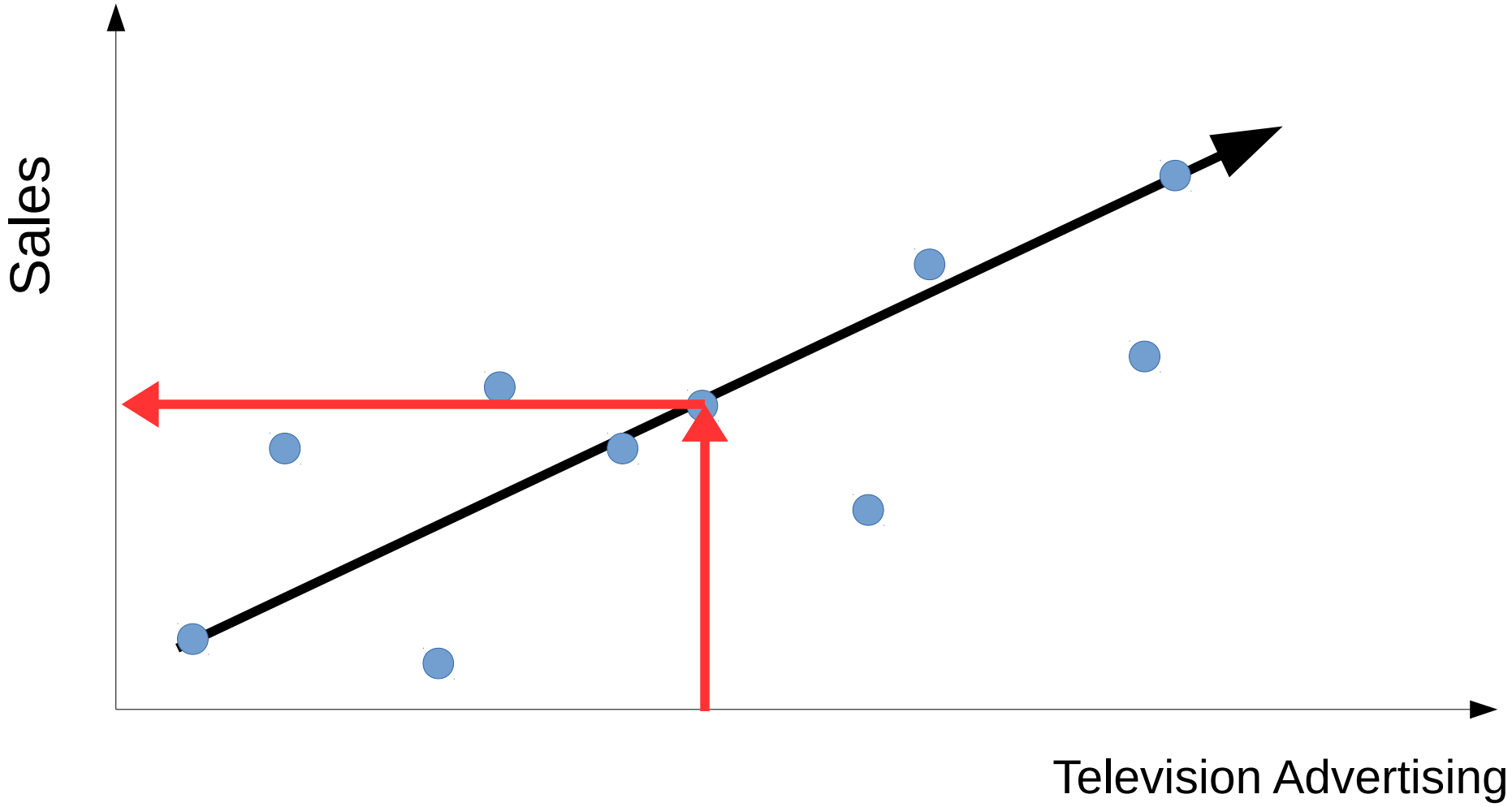
Problem two: We can save the company if we can just get our advertising right!



Problem two: We can save the company if we can just get our advertising right!



Problem two: We can save the company if we can just get our advertising right!



Problem two: We can save the company if we can just get our advertising right!