



# PRG1/EPR: Lecture V18: Winter 2015/16

Prepared by:  
Prof. Dr. Visvanathan Ramesh

Department of Computer Science & Mathematics,  
University of Frankfurt &  
Frankfurt Institute for Advanced Studies

\*Source Credits: Adapted from Prof. John Canny's Data Science  
Course in UC Berkeley.



- **Overview of Data Analytics Trends – Examples of Intelligent Systems**
- **Recap of Data Structures, Random Surfer Model, Small Worlds**
- **Image Processing and Convolutional Neural Networks (Deep Learning)**
- **Introduction to Software Engineering**



- **Data Science and Analytics Overview**
  - Why Data Science?
  - Data Source Types and Quality
  - Examples of Data Analysis
    - Regression
    - Verifying Model Fit (Cross-Validation)

# Data Analysis Has Been Around for a While



FIAS Frankfurt Institute  
for Advanced Studies



GOETHE  
UNIVERSITÄT  
FRANKFURT AM MAIN

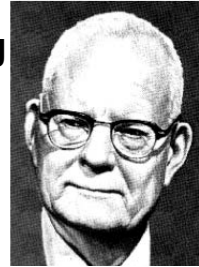
1935: “The Design of Experiments”

R.A. Fisher



1939: “Quality Control”

W.E. Demming

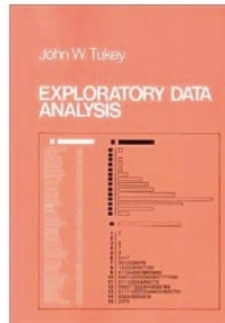


1958: “A Business Intelligence System”



Peter Luhn

1977: “Exploratory Data Analysis”



1989: “Business Intelligence”

Howard  
Dresner



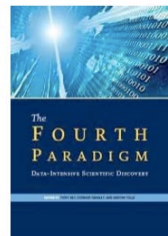
1997: “Machine Learning”



1996: Google



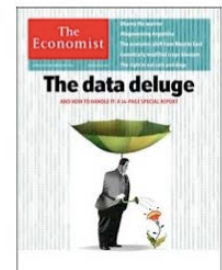
2007: “The Fourth Paradigm”



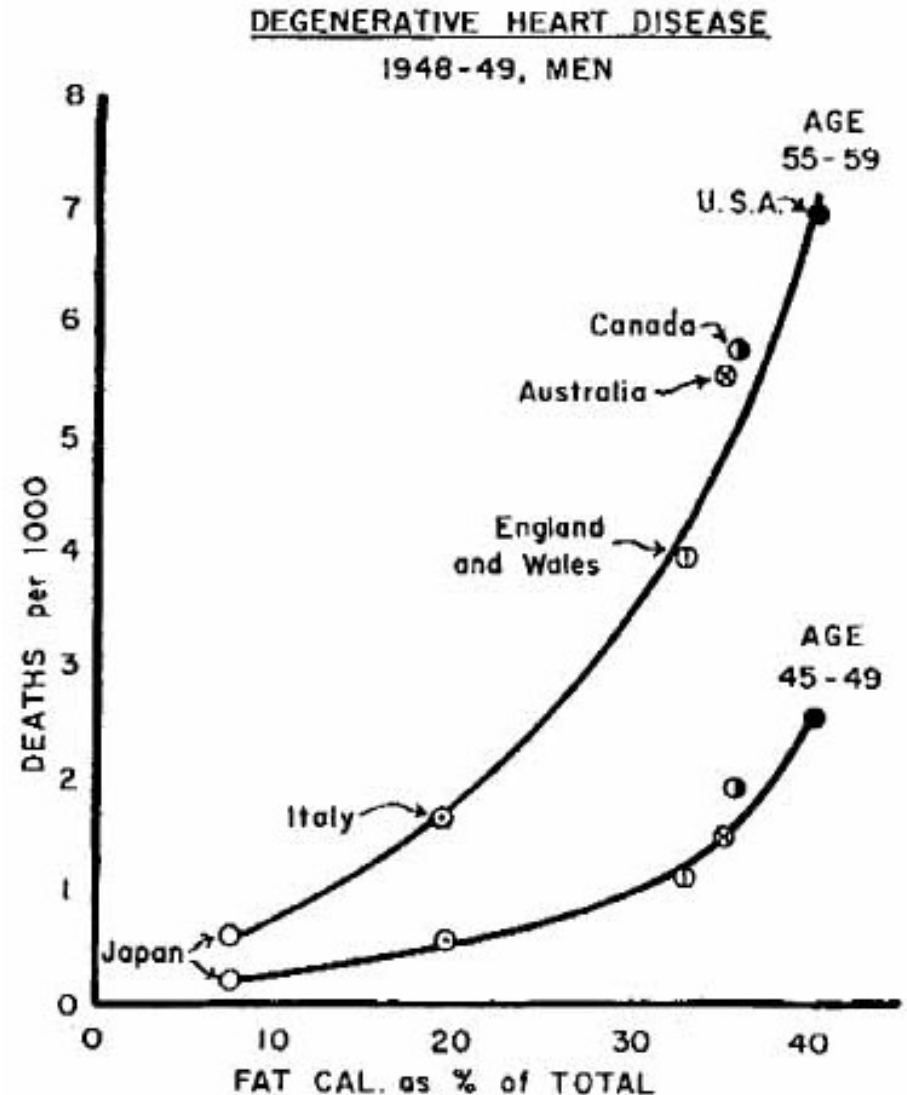
2009: “The Unreasonable Effectiveness of Data”



2010: “The Data Deluge”



**Seven Countries Study**  
**(Ancel Keys, UCB 1925,28)**  
**13,000 subjects total,**  
**5-40 years follow-up.**



# Data Science: Why all the Excitement?



**Example:  
Google Flu Trends:**

**Detecting outbreaks  
two weeks ahead  
of CDC data**

**New models are  
estimating which cities  
are most at risk for  
spread of the Ebola  
virus.**



# Why all the Excitement?



## elections2012

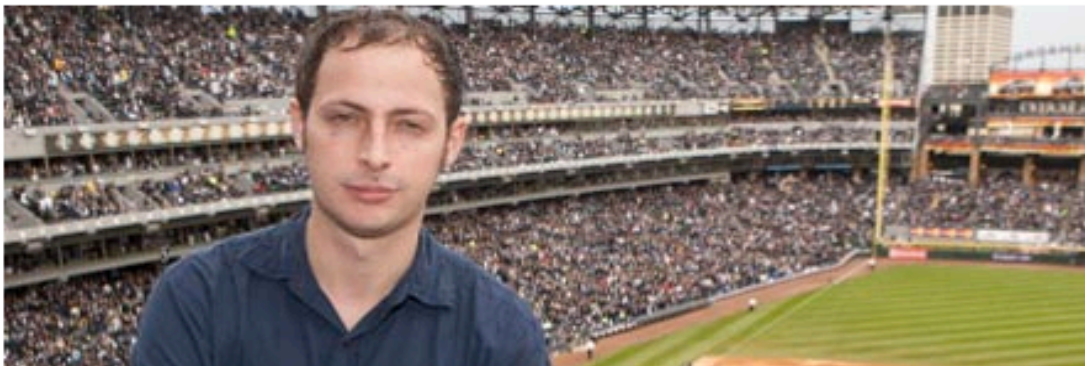
Live results | President | Senate | House | Governor |

### Numbers nerd Nate Silver's forecasts prove all right on election night

FiveThirtyEight blogger predicted the outcome in all 50 states, assuming Barack Obama's Florida victory is confirmed

**Luke Harding**

[guardian.co.uk](http://guardian.co.uk), Wednesday 7 November 2012 10.45 EST



*the signal and the  
and the noise and  
the noise and the  
noise and the no  
why most noise a  
predictions fail t  
but some don't n  
and the noise and  
the noise and the  
nate silver noise  
noise and the no*



...that was just one of several ways that Mr. Obama's campaign operations, some unnoticed by Mr. Romney's aides in Boston, **helped save the president's candidacy**. In Chicago, the campaign recruited a team of behavioral scientists to build an **extraordinarily sophisticated database**

...that allowed the Obama campaign not only to alter the very nature of the electorate, making it younger and less white, but also to create a portrait of shifting voter allegiances. **The power of this operation stunned Mr. Romney's aides on election night**, as they saw voters they never even knew existed turn out in places like Osceola County, Fla.

New York Times, Wed Nov 7, 2012



# Does Data Make Everything Clearer?



FIAS Frankfurt Institute  
for Advanced Studies



GOETHE  
UNIVERSITÄT  
FRANKFURT AM MAIN

## Epidemiological modeling of online social network dynamics

John Cannarella<sup>1</sup>, Joshua A. Spechler<sup>1,\*</sup>

<sup>1</sup> Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ, USA

\* E-mail: Corresponding [spechler@princeton.edu](mailto:spechler@princeton.edu)

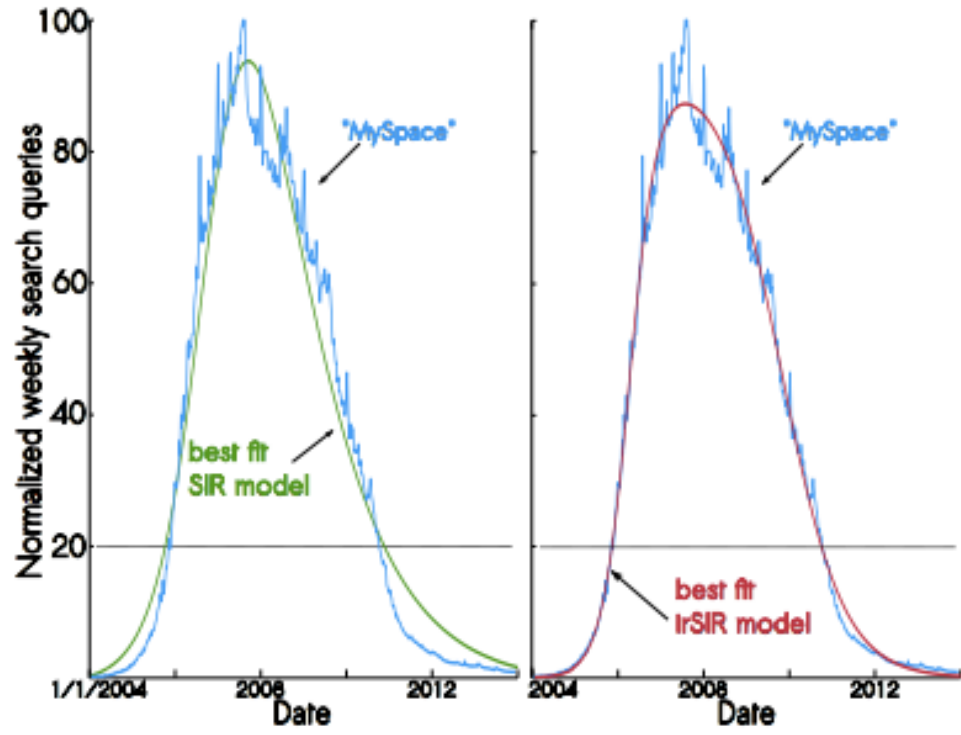
### Abstract

The last decade has seen the rise of immense online social networks (OSNs) such as MySpace and Facebook. In this paper we use epidemiological models to explain user adoption and abandonment of OSNs, where adoption is analogous to infection and abandonment is analogous to recovery. We modify the traditional SIR model of disease spread by incorporating infectious recovery dynamics such that contact between a recovered and infected member of the population is required for recovery. The proposed infectious recovery SIR model (irSIR model) is validated using publicly available Google search query data for “MySpace” as a case study of an OSN that has exhibited both adoption and abandonment phases. The irSIR model is then applied to search query data for “Facebook,” which is just beginning to show the onset of an abandonment phase. **Extrapolating the best fit model into the future predicts a rapid decline in Facebook activity in the next few years.**

# Data Makes Everything Clearer



## Searches for “MySpace”



## Searches for “Facebook”

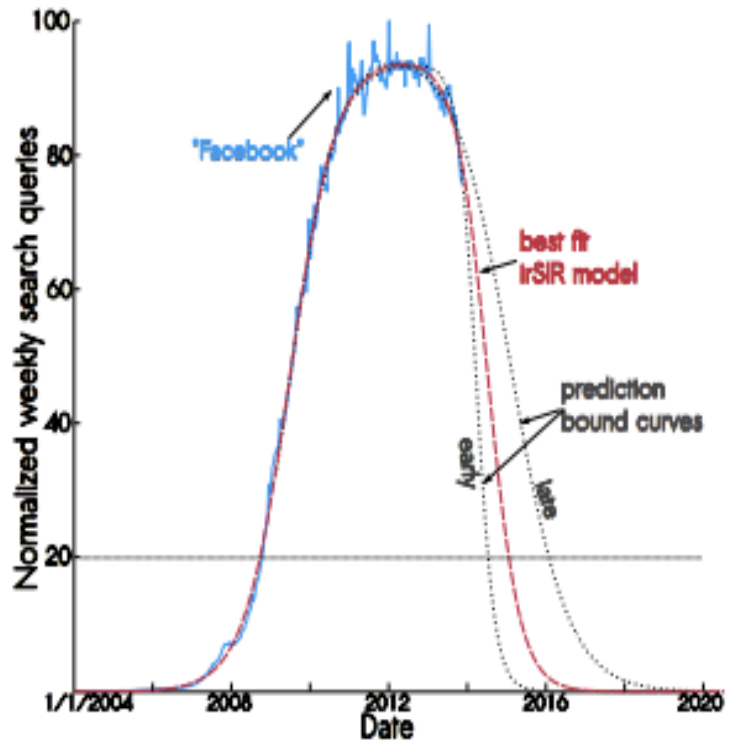
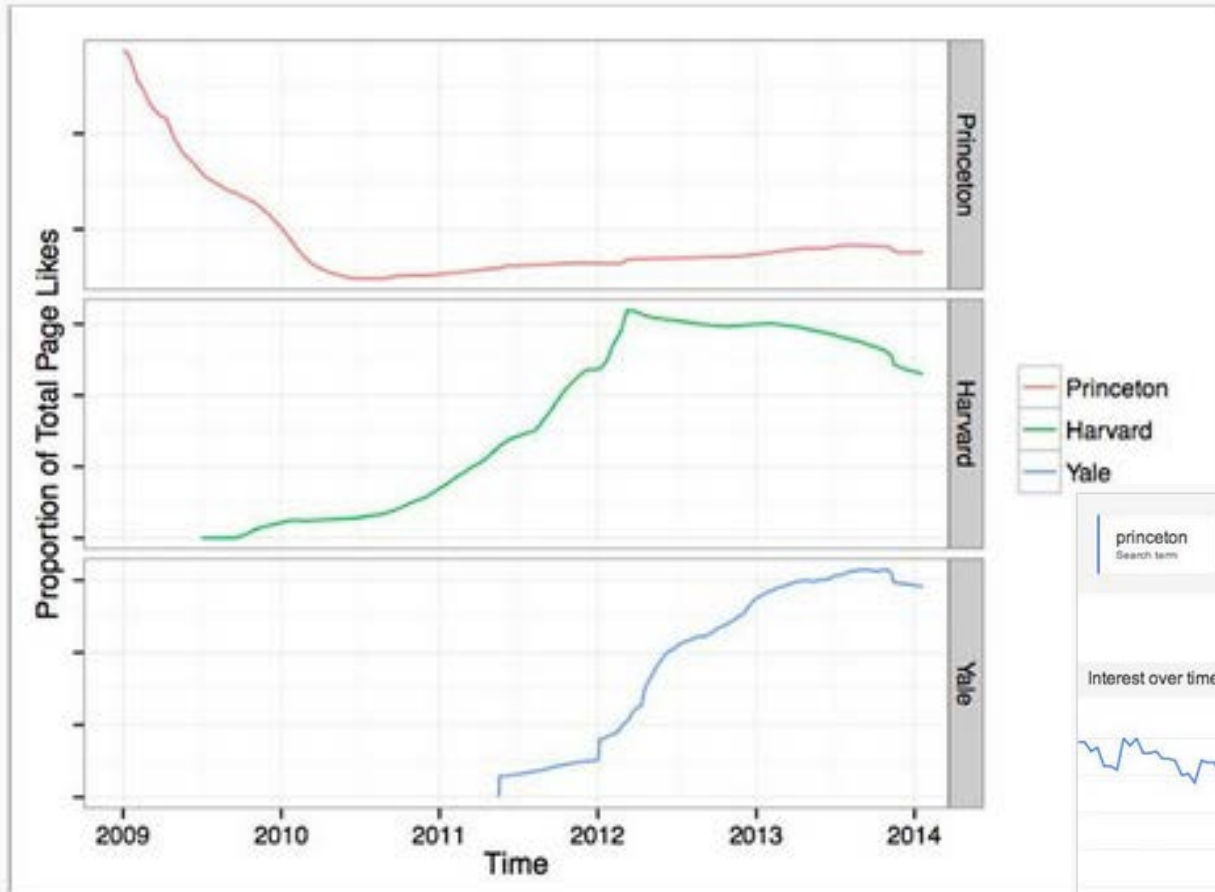


Figure 3: Data for search query “Myspace” with best fit (a) SIR and (b) IrSIR models overlaid. The search query data are normalized such that the maximum data point corresponds to a value of 100.

# Does Data Make Everything Clearer?

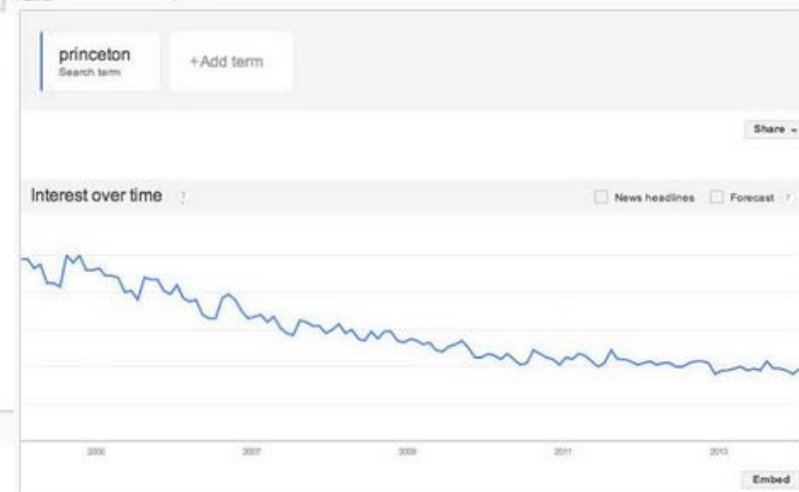


In keeping with the scientific principle “correlation equals causation,” our research unequivocally demonstrated that Princeton may be in danger of disappearing entirely. Looking at page likes on Facebook, we find the following alarming trend:



and based on Princeton search trends:

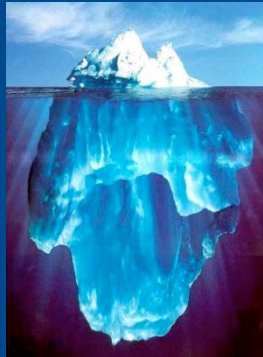
“This trend suggests that Princeton will have only half its current enrollment by 2018, and by 2021 it will have no students at all,...



# “Big Data” Sources



## It's All Happening On-line



Every:  
Click  
Ad impression  
Billing event  
Fast Forward, pause,...  
Server request  
Transaction  
Network message  
Fault

...

## User Generated (Web & Mobile)



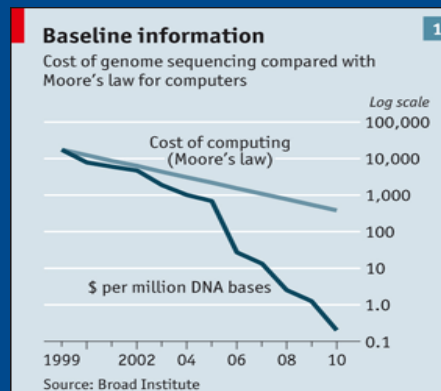
....



## Internet of Things / M2M



## Health/Scientific Computing



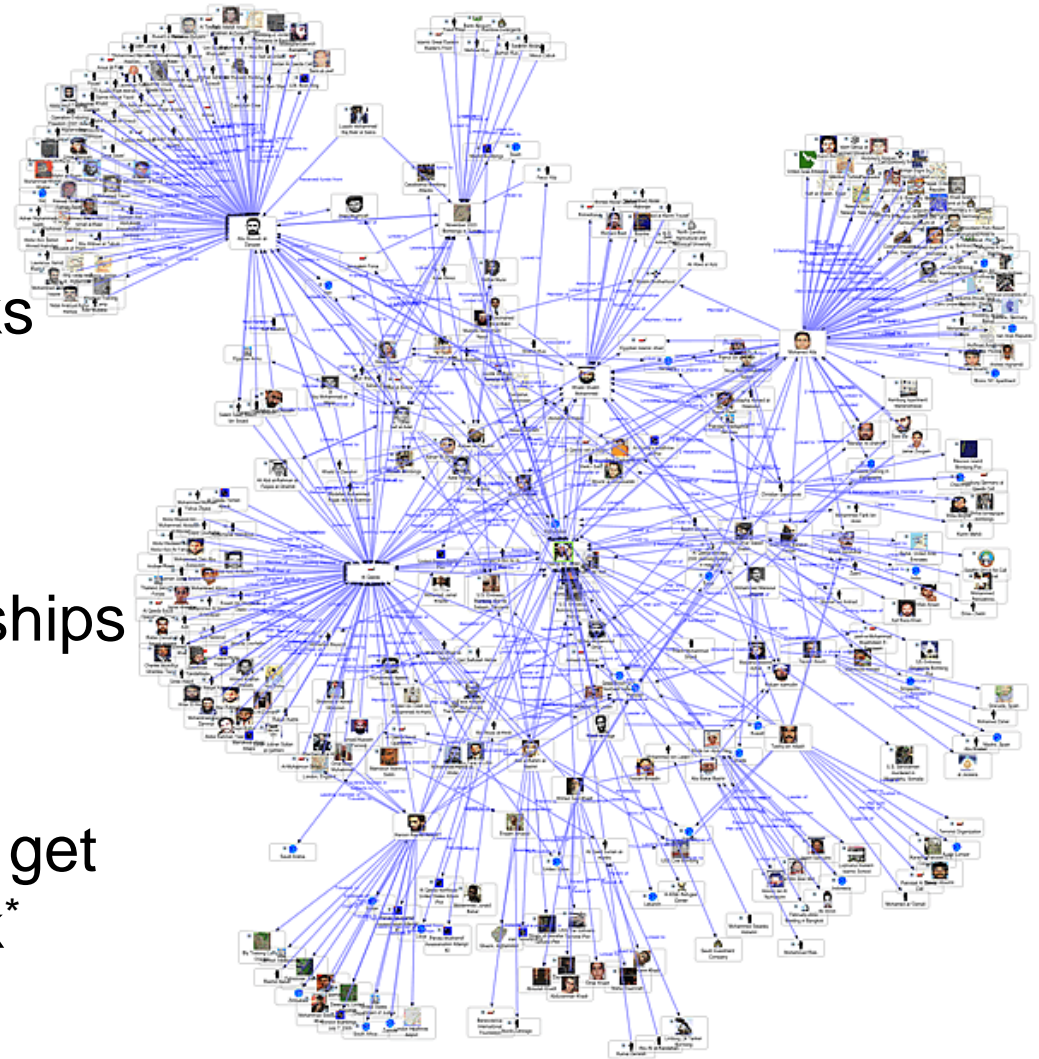
# Graph Data



Lots of interesting data  
has a graph structure:

- Social networks
- Communication networks
- Computer Networks
- Road networks
- Citations
- Collaborations/Relationships
- ...

Some of these graphs can get  
quite large (e.g., Facebook\*  
user graph)



# “Data Science” an Emerging Field



FIAS Frankfurt Institute  
for Advanced Studies



GOETHE  
UNIVERSITÄT  
FRANKFURT AM MAIN

## O'Reilly Radar report



# Example Machine Learning Competitions



FIAS Frankfurt Institute  
for Advanced Studies

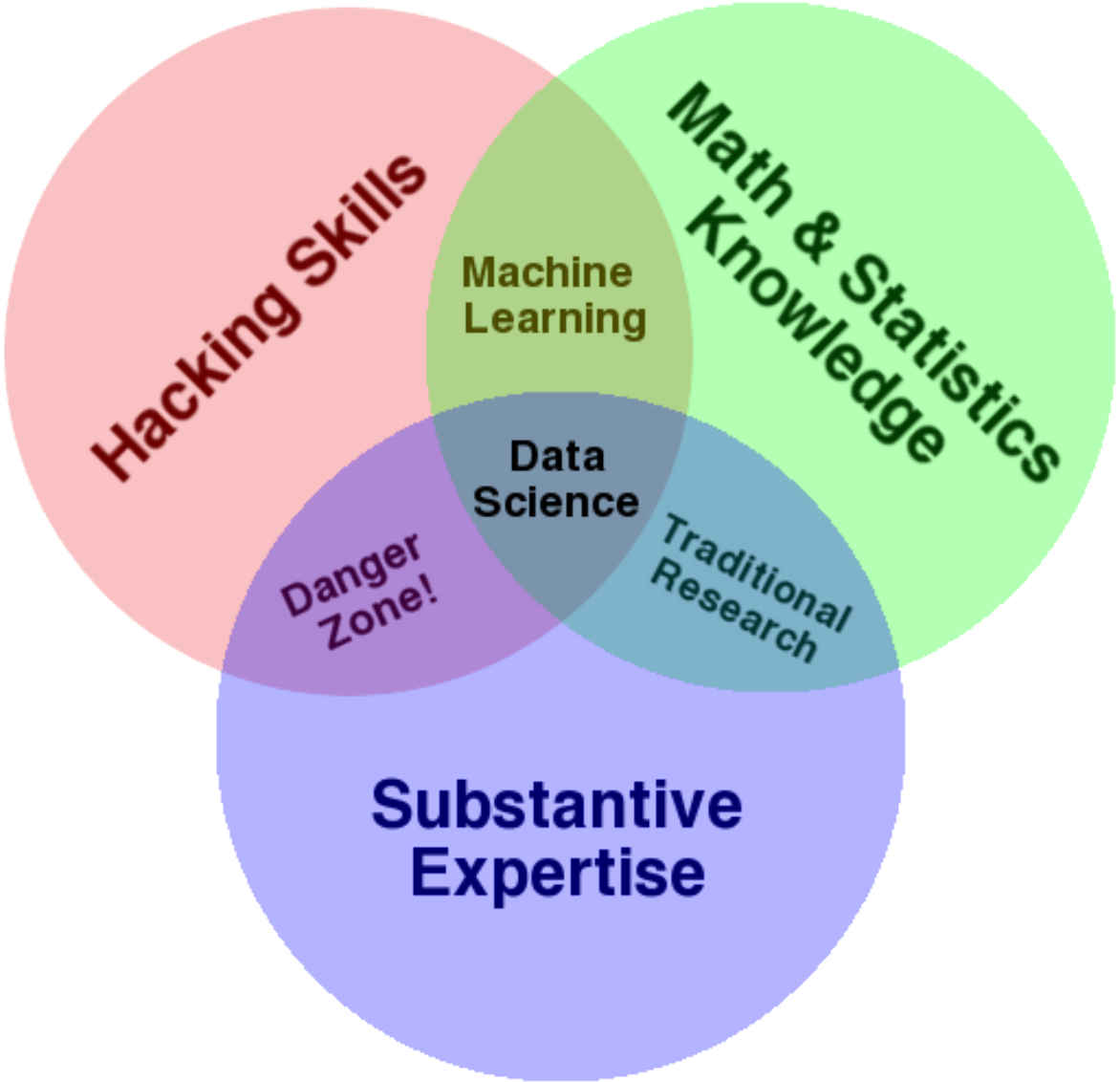


GOETHE  
UNIVERSITÄT  
FRANKFURT AM MAIN

kaggle

Active Competitions			
		<b>Flight Quest 2: Flight Optimization</b> Final Phase of Flight Quest 2	33 days Coming soon \$220,000
		<b>Packing Santa's Sleigh</b> He's making a list, checking it twice; to fill up his sleigh, he needs your advice	5.8 days 338 teams \$10,000
	Genentech	<b>Flu Forecasting</b>	41 days 37 teams
		<b>Galaxy Zoo - The Galaxy Challenge</b> Classify the morphologies of distant galaxies in our Universe	2 months 160 teams \$16,000
		<b>Loan Default Prediction - Imperial College Lon...</b> Constructing an optimal portfolio of loans	52 days 82 teams \$10,000
		<b>Dogs vs. Cats</b> Create an algorithm to distinguish dogs from cats	11 days 166 teams Swag

# Data Science – A Definition





# Contrast: Databases



FIAS Frankfurt Institute  
for Advanced Studies



GOETHE  
UNIVERSITÄT  
FRANKFURT AM MAIN

	Databases	Data Science
Data Value	“Precious”	“Cheap”
Data Volume	Modest	Massive
Examples	Bank records, Personnel records, Census, Medical records	Online clicks, GPS logs, Tweets, Building sensor readings
Priorities	Consistency, Error recovery, Auditability	Speed, Availability, Query richness
Structured	Strongly (Schema)	Weakly or none (Text)
Properties	Transactions, ACID*	CAP* theorem (2/3), eventual consistency
Realizations	SQL	NoSQL: Riak, Memcached, Apache River, MongoDB, CouchDB, Hbase, Cassandra,...

CAP = Consistency, Availability, Partition Tolerance

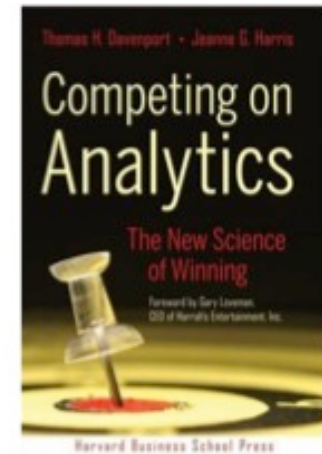
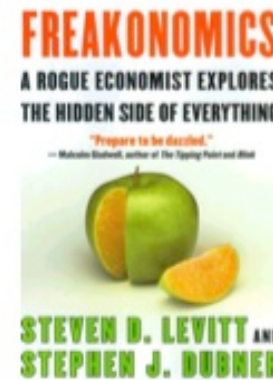
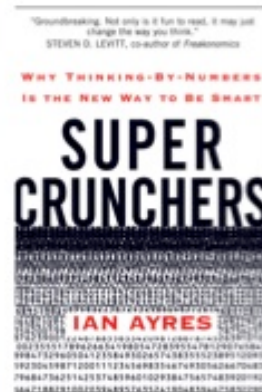
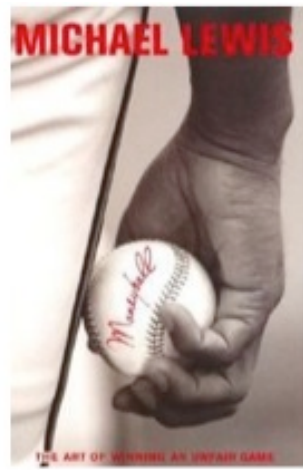
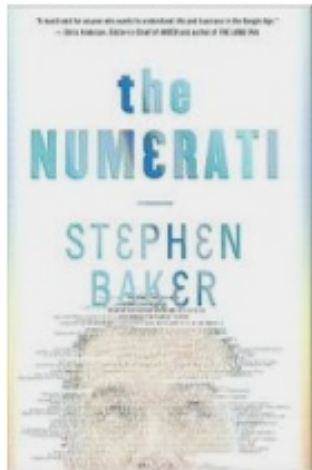
ACID = Atomicity, Consistency, Isolation and Durability

## Databases

Querying the past

## Data Science

Querying the future



**Business intelligence (BI)** is the transformation of raw data into meaningful and useful information for business analysis purposes. BI can handle enormous amounts of unstructured data to help identify, develop and otherwise create new strategic business opportunities - Wikipedia

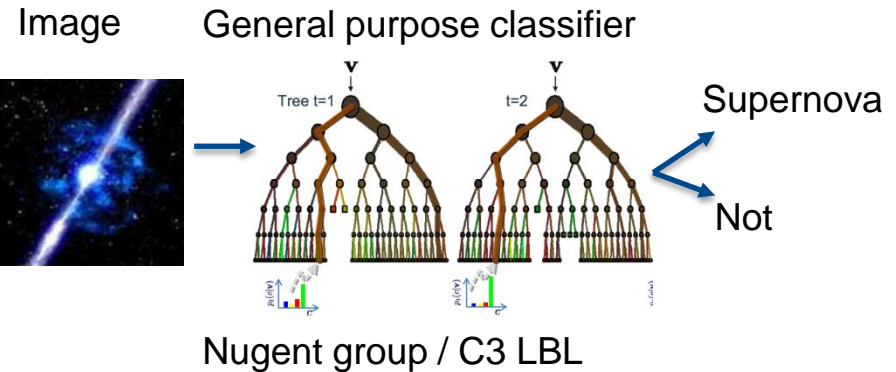
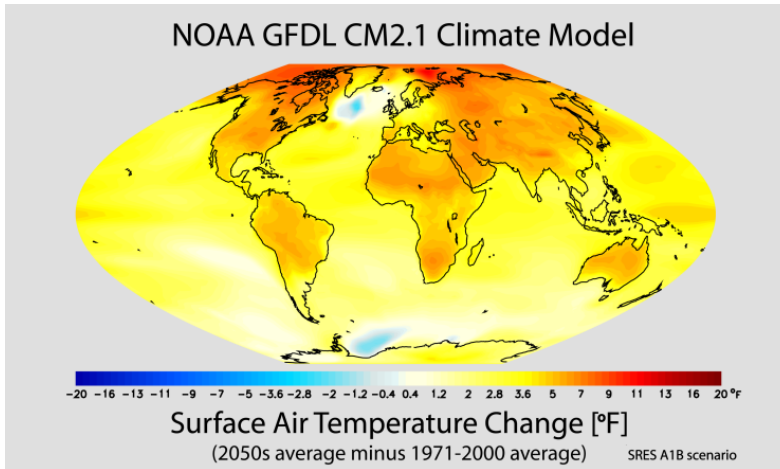
# Contrast: Scientific Computing



FIAS Frankfurt Institute  
for Advanced Studies



GOETHE  
UNIVERSITÄT  
FRANKFURT AM MAIN



## Scientific Modeling

Physics-based models

Problem-Structured

Mostly deterministic, precise

Run on Supercomputer or  
High-end Computing Cluster

## Data-Driven Approach

General inference engine replaces  
model

Structure not related to problem

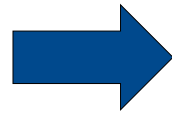
Statistical models handle true  
randomness, and **unmodeled  
complexity**.

Run on cheaper computer Clusters  
(EC2)

# Data Scientist's Practice

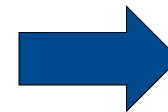


Clean,  
prep



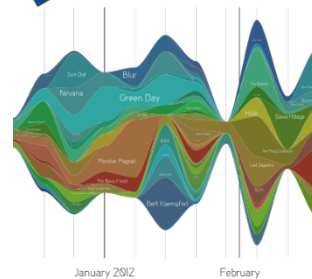
$$\begin{bmatrix} \cos 90^\circ & \sin 90^\circ \\ -\sin 90^\circ & \cos 90^\circ \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

Hypothesize  
Model



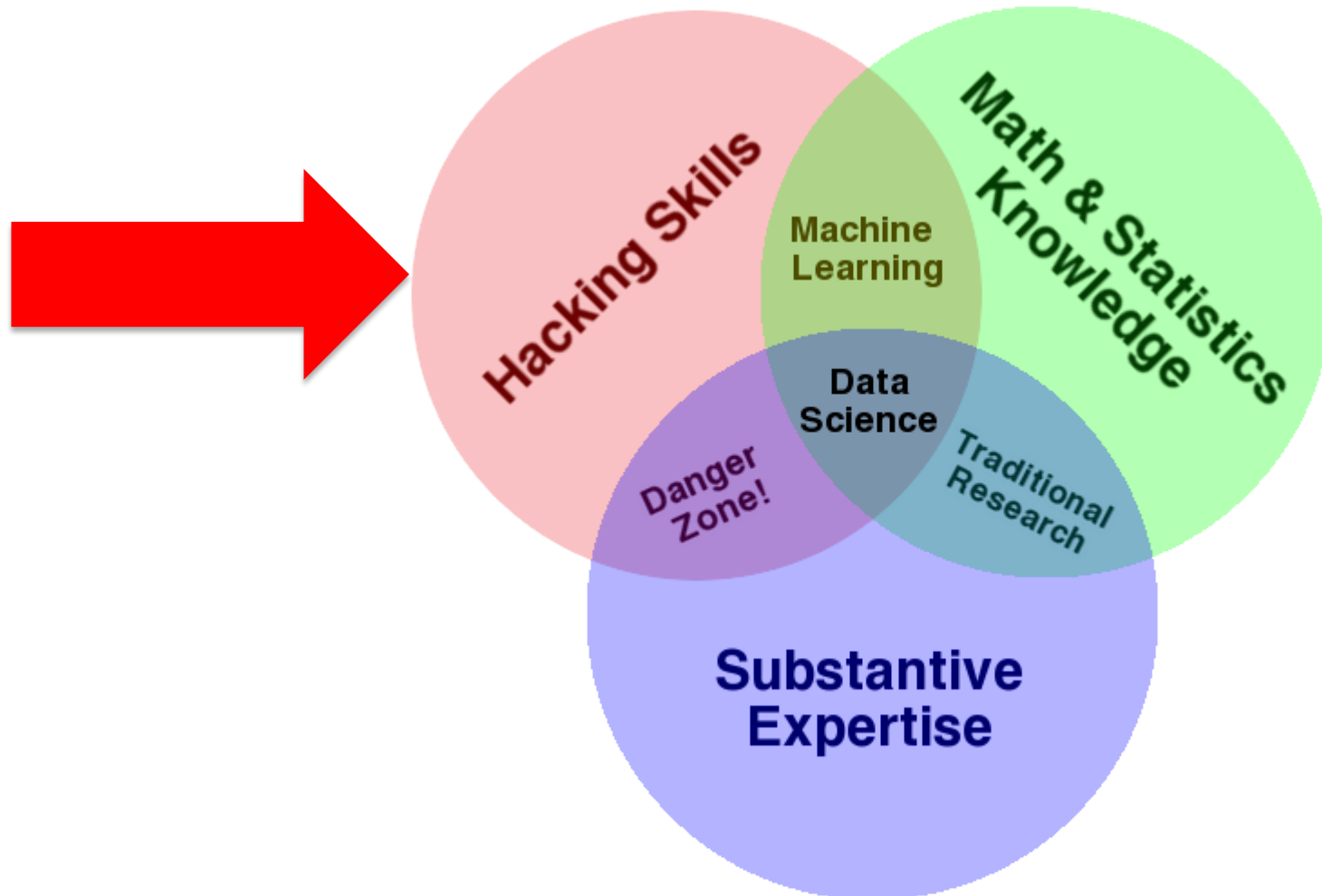
Large Scale  
Exploitation

Digging Around  
in Data

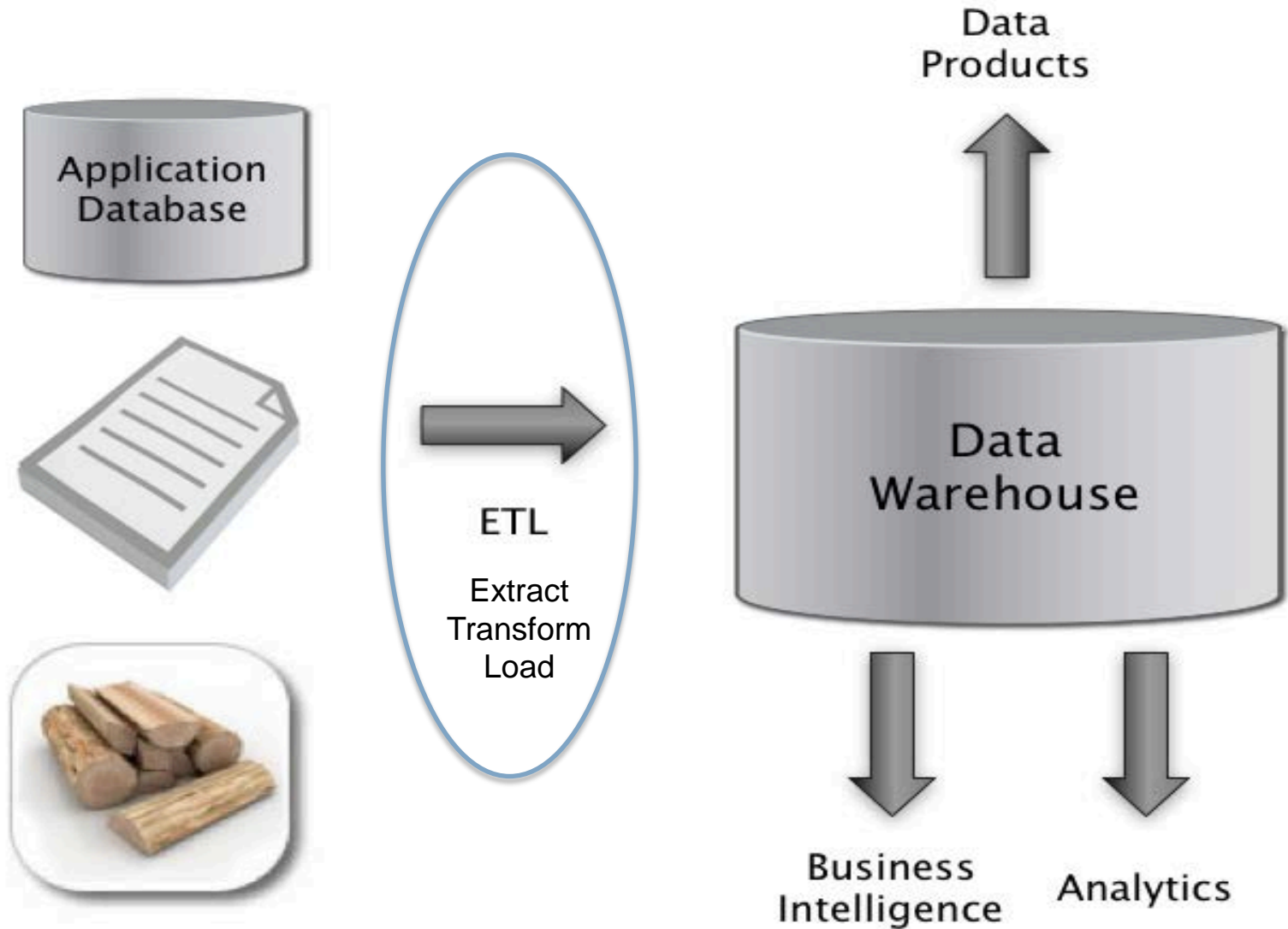


Evaluate  
Interpret

# Data Science – A Definition



# The Big Picture





## ETL

- We need to **extract** data from the **source(s)**
- We need to **load** data into the **sink**
- We need to **transform** data at the source, sink, or in a **staging area**
  
- Sources: file, database, event log, web site, HDFS...
- Sinks: Python, R, SQLite, RDBMS, NoSQL store, files, HDFS...



## Process model

- The construction of a new data preparation process is done in many phases
  - Data **characterization**
  - Data **cleaning**
  - Data **integration**
- We must efficiently move data around in space and time
  - Data **transfer**
  - Data **serialization** and **deserialization** (for files or network)





## Workflow

- The transformation **pipeline** or **workflow** often consists of many steps
  - For example: Unix pipes and filters
  - `$ cat data_science.txt | wc | mail -s "word count" myname@some.com`
- If the workflow is to be used more than once, it can be **scheduled**
  - Scheduling can be time-based or event-based
  - Use publish-subscribe to register interest (e.g. Twitter feeds)
- Recording the execution of a workflow is known as capturing **lineage** or **provenance**

# Data Sources at Web Companies



## Examples from Facebook

- Application databases
- Web server logs
- Event logs
- API server logs
- Ad server logs
- Search server logs
- Advertisement landing page content
- Wikipedia
- Images and video



# Data Source Types & Examples



## What is a table?

- A **table** is a collection of **rows** and **columns**
- Each row has an **index**
- Each column has a **name**
- A **cell** is specified by an (index, name) pair
- A cell may or may not have a **value**

# Tabular Data



## Fortune 500

	A	B	C	D	E	F	G	H	I
1	rank	company	cik	ticker	sic	state_location	state_of_incorporation	revenues	profits
2	1	Wal-Mart Stores	104169	WMT	5331	AR	DE	421849	16389
3	2	Exxon Mobil	34088	XOM	2911	TX	NJ	354674	30460
4	3	Chevron	93410	CVX	2911	CA	DE	196337	19024
5	4	ConocoPhillips	1163165	COP	2911	TX	DE	184966	11358
6	5	Fannie Mae	310522	FNM	6111	DC	DC	153825	-14014
7	6	General Electric	40545	GE	3600	CT	NY	151628	11644
8	7	Berkshire Hathaway	1067983	BRKA	6331	NE	DE	136185	12967
9	8	General Motors	1467858	GM	3711	MI	MI	135592	6172
10	9	Bank of America Corp.	70858	BAC	6021	NC	DE	134194	-2238
11	10	Ford Motor	37996	F	3711	MI	DE	128954	6561
12	11	Hewlett-Packard	47217	HPQ	3570	CA	DE	126033	8761
13	12	AT&T	732717	T	4813	TX	DE	124629	19864
14	13	J.P. Morgan Chase & Co.	19617	JPM	6021	NY	DE	115475	17370
15	14	Citigroup	831001	C	6021	NY	DE	111055	10602
16	15	McKesson	927653	MCK	5122	CA	DE	108702	1263
17	16	Verizon Communications	732712	VZ	4813	NY	DE	106565	2549
18	17	American International Group	5272	AIG	6331	NY	DE	104417	7786
19	18	International Business Machines	51143	IBM	3570	NY	NY	99870	14833
20	19	Cardinal Health	721371	CAH	5122	OH	OH	98601.9	642.2
21	20	Freddie Mac	37785	FMC	2800	PA	DE	98368	-14025

# Tabular Data (csv)



## Fortune 500

```
Fortune 500 with ticker and EDGAR - Plus Ticker and EDGAR.txt
rank,company,cik,ticker,sic,state_location,state_of_incorporation,revenues,profits
1,Wal-Mart Stores,104169,WMT,5331,AR,DE,421849,16389
2,Exxon Mobil,34088,XOM,2911,TX,NJ,354674,30460
3,Chevron,93410,CVX,2911,CA,DE,196337,19024
4,ConocoPhillips,1163165,COP,2911,TX,DE,184966,11358
5,Fannie Mae,310522,FNM,6111,DC,DC,153825,-14014
6,General Electric,40545,GE,3600,CT,NY,151628,11644
7,Berkshire Hathaway,1067983,BRKA,6331,NE,DE,136185,12967
8,General Motors,1467858,GM,3711,MI,MI,135592,6172
9,Bank of America Corp.,70858,BAC,6021,NC,DE,134194,-2238
10,Ford Motor,37996,F,3711,MI,DE,128954,6561
11,Hewlett-Packard,47217,HPQ,3570,CA,DE,126033,8761
12,AT&T,732717,T,4813,TX,DE,124629,19864
13,J.P. Morgan Chase & Co.,19617,JPM,6021,NY,DE,115475,17370
14,Citigroup,831001,C,6021,NY,DE,111055,10602
15,McKesson,927653,MCK,5122,CA,DE,108702,1263
16,Verizon Communications,732712,VZ,4813,NY,DE,106565,2549
17,American International Group,5272,AIG,6331,NY,DE,104417,7786
18,International Business Machines,51143,IBM,3570,NY,NY,99870,14833
19,Cardinal Health,721371,CAH,5122,OH,OH,98601.9,642.2
20,Freddie Mac,37785,FMC,2800,PA,DE,98368,-14025
21,CVS Caremark,64803,CVS,5912,RI,DE,96413,3427
22,UnitedHealth Group,731766,UNH,6324,MN,MN,94155,4634
23,Wells Fargo,72971,WFC,6021,CA,DE,93249,12362
24,Valero Energy,1035002,VLO,2911,TX,DE,86034,324
25,Kroger,56873,KR,5411,OH,OH,82189.4,1116.3
26,Procter & Gamble,80424,PG,2840,OH,OH,79689,12736
27,AmerisourceBergen,1140859,ABC,5122,PA,DE,77954,636.7
28,Costco Wholesale,909832,COST,5331,WA,WA,77946,1303
29,Marathon Oil,101778,MRO,2911,TX,DE,68413,2568
30,Home Depot,354950,HD,5211,GA,DE,67997,3338
```



```
HEADER  APOPTOSIS                05-OCT-10  3IZA
TITLE   STRUCTURE OF AN APOPTOSOME-PROCASPASE-9 CARD COMPLEX
COMPND  MOL_ID: 1;
COMPND  2 MOLECULE: APOPTOTIC PROTEASE-ACTIVATING FACTOR 1;
COMPND  3 CHAIN: A, B, C, D, E, F, G;
COMPND  4 SYNONYM: APAF-1;
COMPND  5 ENGINEERED: YES
SOURCE  MOL_ID: 1;
SOURCE  2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE  3 ORGANISM_COMMON: HUMAN;
SOURCE  4 ORGANISM_TAXID: 9606;
SOURCE  5 GENE: APAF1, KIAA0413;
SOURCE  6 EXPRESSION_SYSTEM: SPODOPTERA FRUGIPERDA;
SOURCE  7 EXPRESSION_SYSTEM_TAXID: 7108;
SOURCE  8 EXPRESSION_SYSTEM_STRAIN: SF21;
SOURCE  9 EXPRESSION_SYSTEM_VECTOR_TYPE: INSECT VIRUS;
SOURCE  10 EXPRESSION_SYSTEM_PLASMID: PFASTBAC1
KEYWDS  APOPTOSOME, APAF-1, PROCASPASE-9 CARD, APOPTOSIS
EXPDTA  ELECTRON MICROSCOPY
AUTHOR  S.YUAN,X.YU,M.TOPF,S.J.LUDTKE,X.WANG,C.W.AKEY
REVDAT  1  03-NOV-10 3IZA  0
SPRSDE  03-NOV-10 3IZA  3IYT
JRNL    AUTH  S.YUAN,X.YU,M.TOPF,S.J.LUDTKE,X.WANG,C.W.AKEY
JRNL    TITL  STRUCTURE OF AN APOPTOSOME-PROCASPASE-9 CARD COMPLEX
JRNL    REF   STRUCTURE                V. 18  571 2010
```



## Challenges:

- May be many missing fields (a particular sensor may not produce all types of output).
- Device may go offline for a while.
- Device may be damaged (permanently or intermittently).
- Timestamps usually critical but may not be accurate.
- Other meta-data (location, device ID) may have errors.



# Log Files – Example Apache Web Log



Processes, usually daemons, create logs  
e.g., httpd, mysqld, syslogd

```
66.249.65.107 - - [08/Oct/2007:04:54:20 -0400] "GET /support.html HTTP/1.1" 200  
11179 "-" "Mozilla/5.0 (compatible; Googlebot/2.1;  
+http://www.google.com/bot.html)"
```

```
111.111.111.111 - - [08/Oct/2007:11:17:55 -0400] "GET / HTTP/1.1" 200 10801  
"http://www.google.com/search?q=log+analyzer&ie=utf-8&oe=utf-8  
&aq=t&rls=org.mozilla:en-US:official&client=firefox-a" "Mozilla/5.0 (Windows; U;  
Windows NT 5.2; en-US; rv:1.8.1.7) Gecko/20070914 Firefox/2.0.0.7"
```

```
111.111.111.111 - - [08/Oct/2007:11:17:55 -0400] "GET /style.css HTTP/1.1" 200 3225  
"http://www.loganalyzer.net/" "Mozilla/5.0 (Windows; U; Windows NT 5.2; en-US;  
rv:1.8.1.7) Gecko/20070914 Firefox/2.0.0.7"
```

# “Splunking”



- Grab data from many machines
- Index it
- Check for unusual events:
  - Disk problems
  - Network congestion
  - Security attacks
- Monitor Resources
  - Network
  - Memory usage
  - Disk use, latency
  - Threads
- Dashboard for cloud administration.



## What is a file?

- A **file** is a named sequence of **bytes**
  - Typically stored as a collection of pages (or blocks)
- A **filesystem** is a collection of files organized within an hierarchical namespace
  - Responsible for laying out those bytes on physical media
  - Stores file metadata
  - Provides an API for interaction with files
- Standard operations
  - `open()/close()`
  - `seek()`
  - `read()/write()`



## Considerations for a file format

- Data model: tabular, hierarchical, array
- Physical layout
- Field units and validation
- Metadata: header, side file, specification, other?
- Plain text or binary
- Encoding: ASCII, UTF-8, other?
- Delimiters and escaping
- Compression, encryption, checksums?
- Schema evolution

## Read/Write time (180 MB tabular file)

	Read Time (Text)	Write Time (Text)	Read Time (Binary)	Write Time (Binary)
Pandas (Python)	88 secs	107 secs	**	**
Scala/Java	12 secs	30 secs	0.5-2* secs	0.5-2* secs

\*\* Pandas doesn't have a default binary file I/O library – you can use Python, but performance depends on what you pick.

\* 2 seconds is the time for sustainable read/write.  
May be faster due to caching

# File Performance – Compression



## Read/Write time (180 MB tabular file, Scala/Java)

Binary File	Read Time	Write Time	File Size
Gzip level 6 (Java default)	1-2 secs	11 secs	35 MB
Gzip level 3	1-2 secs	4 secs	37 MB
Gzip level 1	1-2 secs	2 secs	38 MB
LZ4 fast	1-2 secs	1-2 secs	63 MB

Text File	Read Time	Write Time	File Size
Gzip level 6 (default)	1-2 secs	22 secs	32 MB
Gzip level 3	1-2 secs	14 secs	35 MB
Gzip level 1	1-2 secs	12 secs	39 MB
LZ4 fast	1-2 secs	11 secs	63 MB



# “Table-Like” Data Model Example

---

**pandas: powerful Python data analysis  
toolkit**

*Release 0.17.0*

Wes McKinney & PyData Development Team



## Series: a named, ordered dictionary

- The keys of the dictionary are the **indexes**
- Built on NumPy's **ndarray**
- Values can be any Numpy data type object

## DataFrame: a table with named columns

- Represented as a Dict (col\_name -> series)
- Each Series object represents a column



# Operations



**map() functions**

**filter (apply predicate to rows)**

**sort/group by**

**aggregate: sum, count, average, max, min**

**Pivot or reshape**

**Relational:**

- union, intersection, difference, cartesian product (CROSS JOIN)
- select/filter, project
- join: natural join (INNER JOIN), theta join, semi-join, etc.
- rename



# Dirty Data: Errors in Data Sources

## The Statistics View:

- There is a process that produces data
- We want to model ideal samples of that process, but in practice we have non-ideal samples:
  - **Distortion** – some samples are corrupted by a process
  - **Selection Bias** - likelihood of a sample depends on its value
  - **Left and right censorship** - users come and go from our scrutiny
  - **Dependence** – samples are supposed to be independent, but are not (e.g. social networks)
- You can add new models for each type of imperfection, but you can't model everything.
- What's the best trade-off between accuracy and simplicity?



## The Database View:

- I got my hands on this data set
- Some of the values are missing, corrupted, wrong, duplicated
- Results are absolute (relational model)
- You get a better answer by improving the quality of the values in your dataset



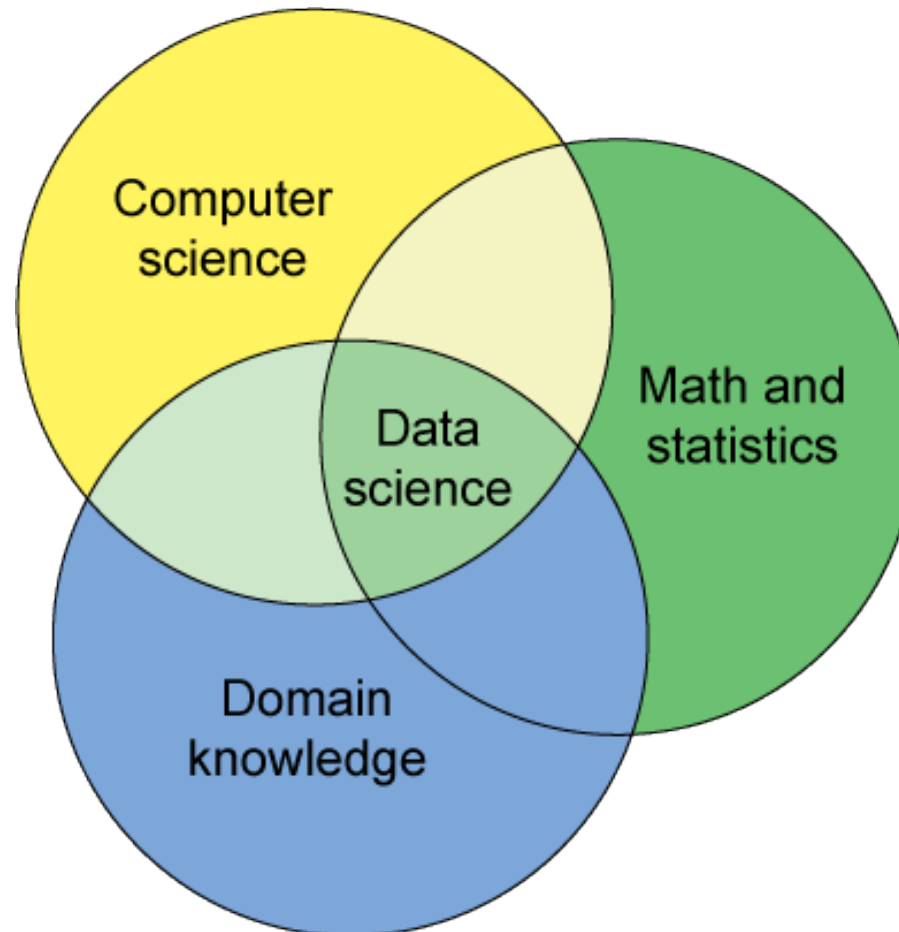
## The Domain Expert's View:

- This Data Doesn't look right
- This Answer Doesn't look right
- What happened?

**Domain experts have an implicit model of the data that they can test against...**

## The Data Scientist's View:

- Some Combination of all of the above





**(Source) Data is dirty on its own.**

**Transformations corrupt the data (complexity of software pipelines).**

**Data sets are clean but **integration** (i.e., combining them) screws them up.**

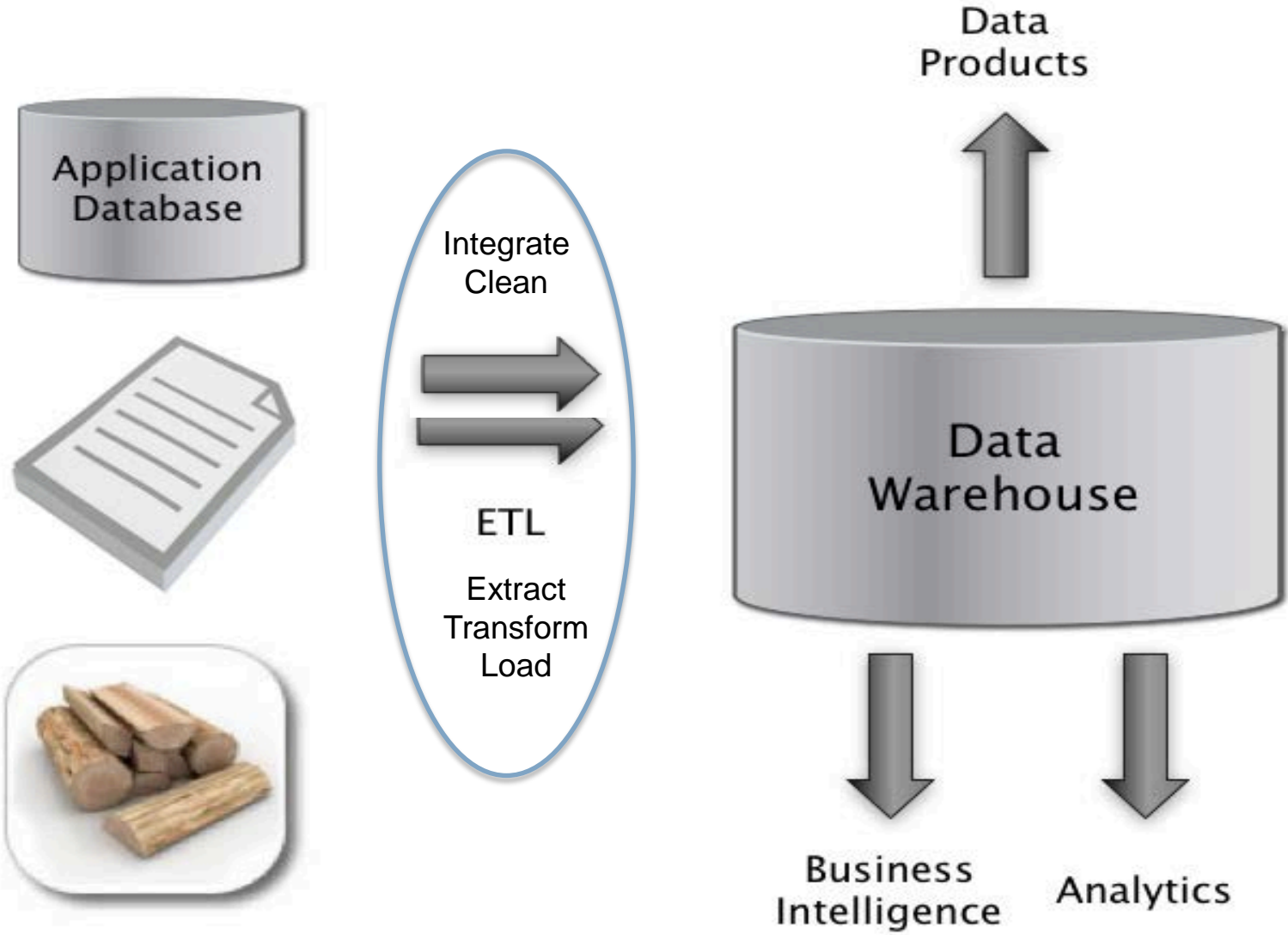
**“Rare” errors can become frequent after transformation or integration.**

**Data sets are clean but suffer “bit rot”**

- Old data loses its value/accuracy over time

**Any combination of the above**

# Where can Dirty Data Arise?



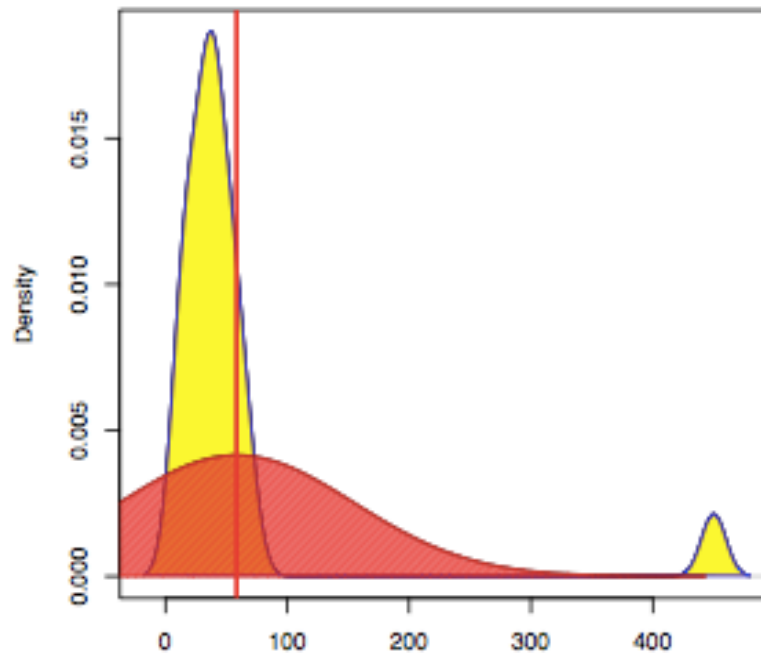


# Numeric Outliers



12	13	14	21	22	26	33	35	36	37	39	42	45	47	54	57	61	68	450
----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	-----

## ages of employees (US)



- ☉ median 37
- ☉ mean 58.52632
- ☉ variance 9252.041

Source: Joe Hellerstein's UCB CS 194 Guest Lecture

# Data Cleaning Makes Everything Okay?



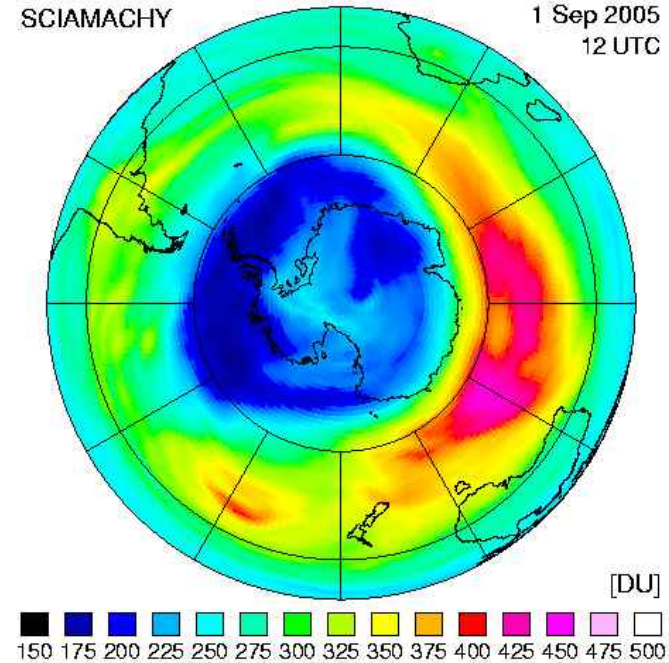
FIAS Frankfurt Institute  
for Advanced Studies



GOETHE  
UNIVERSITÄT  
FRANKFURT AM MAIN

**The appearance of a hole in the earth's ozone layer over Antarctica, first detected in 1976, was so unexpected that scientists didn't pay attention to what their instruments were telling them; they thought their instruments were malfunctioning.**

National Center for Atmospheric Research



**In fact, the data were rejected as unreasonable by data quality control algorithms**

## From Stanford Data Integration Course:

- 1) parsing text into fields (separator issues)
- 2) Naming conventions: ER: NYC vs New York
- 3) Missing required field (e.g. key field)
- 4) Different representations (2 vs Two)
- 5) Fields too long (get truncated)
- 6) Primary key violation (from un- to structured or during integration)
- 7) Redundant Records (exact match or other)
- 8) Formatting issues – especially dates
- 9) Licensing issues/Privacy/ keep you from using the data as you would like?

# Data Quality: Modern Definition?



## We need a definition of data quality which

- Reflects the **use** of the data
- Leads to **improvements in processes**
- Is **measurable** (we can define metrics)

## First, we need a better understanding of how and where data quality problems occur

- The **data quality continuum**

# Meaning of Data Quality (2)



## There are many types of data, which have different uses and typical quality problems

- Federated data
- High dimensional data
- Descriptive data
- Longitudinal data
- Streaming data
- Web (scraped) data
- Numeric vs. categorical vs. text data

# Meaning of Data Quality (2)



## There are many uses of data

- Operations
- Aggregate analysis
- Customer relations ...

**Data Interpretation : the data is useless if we don't know all of the **rules** behind the data.**

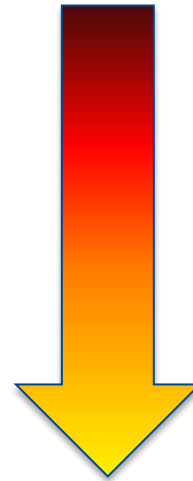
**Data Suitability : Can you get the answer from the available data**

- Use of proxy data
- Relevant data is missing



## Data and information is not static, it flows in a data collection and usage process

- Data gathering
- Data delivery
- Data storage
- Data integration
- Data retrieval
- Data mining/analysis





# Machine Learning





**Supervised:** We are given input samples ( $X$ ) and output samples ( $y$ ) of a function  $y = f(X)$ . We would like to “learn”  $f$ , and evaluate it on new data. Types:

- **Classification:**  $y$  is discrete (class labels).
- **Regression:**  $y$  is continuous, e.g. linear regression.

**Unsupervised:** Given only samples  $X$  of the data, we compute a function  $f$  such that  $y = f(X)$  is “simpler”.

- **Clustering:**  $y$  is discrete
- $Y$  is continuous: **Matrix factorization, Kalman filtering, unsupervised neural networks.**



## Supervised:

- Is this image a cat, dog, car, house?
- How would this user score that restaurant?
- Is this email spam?
- Is this blob a supernova?

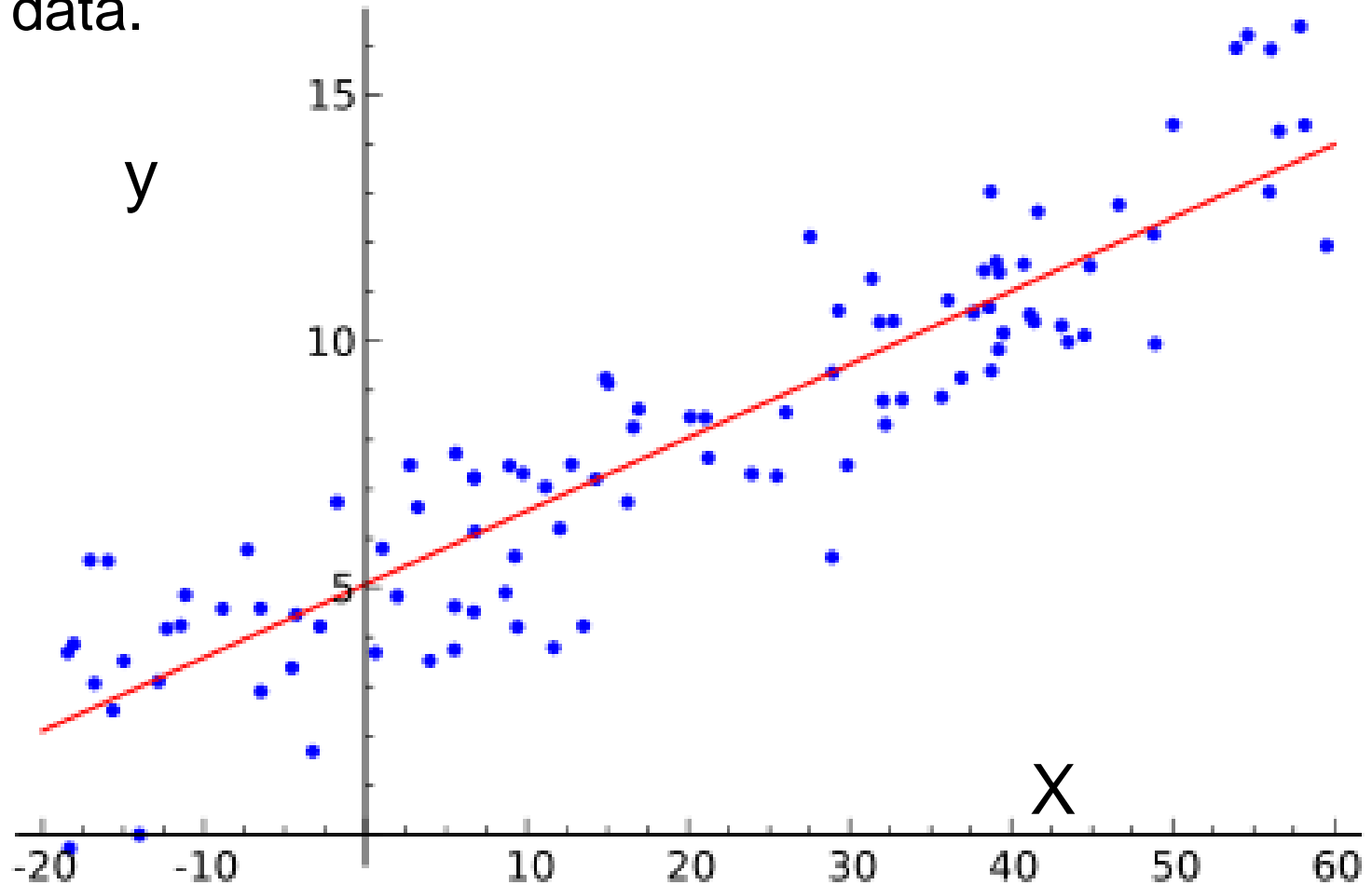
## Unsupervised

- Cluster some hand-written digit data into 10 classes.
- What are the top 20 topics in Twitter right now?
- Find and cluster distinct accents of people at Berkeley.

# Linear Regression



We want to find the best line (linear function  $y=f(X)$ )  
to explain the data.



The predicted value of  $y$  is given by:

$$\hat{y} = \hat{\beta}_0 + \sum_{j=1}^p X_j \hat{\beta}_j$$

The vector of coefficients  $\hat{\beta}$  is the regression model.

If  $X_0 = 1$ , the formula becomes a matrix product:

$$\hat{y} = X \hat{\beta}$$

We can write all of the input samples in a single matrix  $\mathbf{X}$ :

$$\text{i.e. rows of } \mathbf{X} = \begin{pmatrix} X_{11} & \cdots & X_{1n} \\ \vdots & \ddots & \vdots \\ X_{m1} & \cdots & X_{mn} \end{pmatrix}$$

are **distinct observations**, **columns of  $\mathbf{X}$**  are **input features**.

To determine the model parameters  $\hat{\beta}$  from some data, we can write down the Residual Sum of Squares:

$$\text{RSS}(\beta) = \sum_{i=1}^N (y_i - \beta x_i)^2$$

or symbolically  $\text{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$ . To minimize it, take the derivative wrt  $\beta$  which gives:

$$\mathbf{X}^T (\mathbf{y} - \mathbf{X}\beta) = 0$$

And if  $\mathbf{X}^T \mathbf{X}$  is non-singular, the unique solution is:

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

The exact method requires us to invert a matrix ( $\mathbf{X}^T \mathbf{X}$ ) whose size is `nfeatures x nfeatures`. This will often be **too big**.

There are many gradient-based methods which reduce the RSS error by taking the **derivative wrt  $\beta$**

$$\text{RSS}(\beta) = \sum_{i=1}^N (y_i - \beta x_i)^2$$

which was

$$\nabla = \mathbf{X}^T (\mathbf{y} - \mathbf{X}\beta)$$

# Stochastic Gradient



A very important set of iterative algorithms use **stochastic gradient** updates.

They use a **small subset or mini-batch**  $\mathbf{X}$  of the data, and use it to compute a gradient which is added to the model

$$\beta' = \beta + \alpha \nabla$$

Where  $\alpha$  is called the **learning rate**.

These updates happen **many times** in one pass over the dataset.

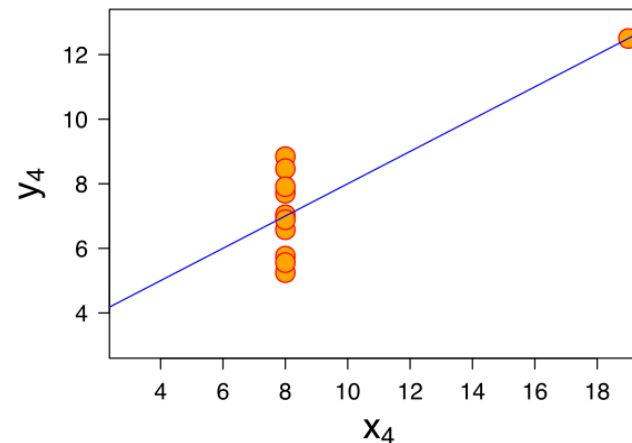
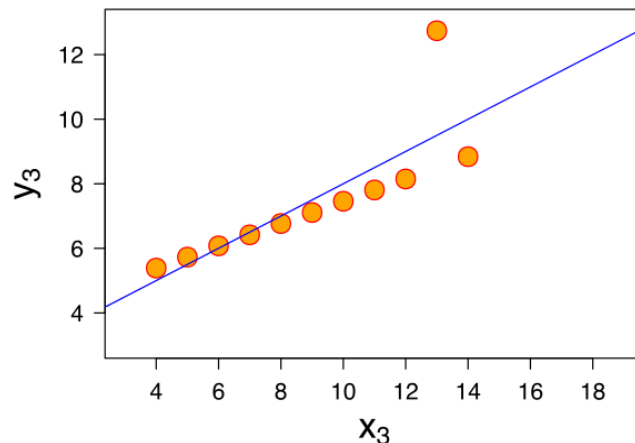
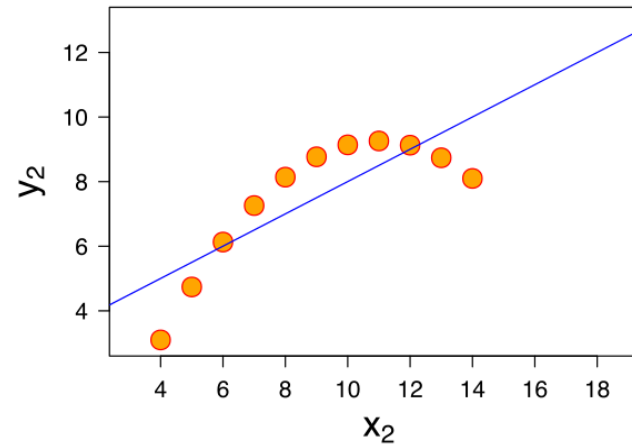
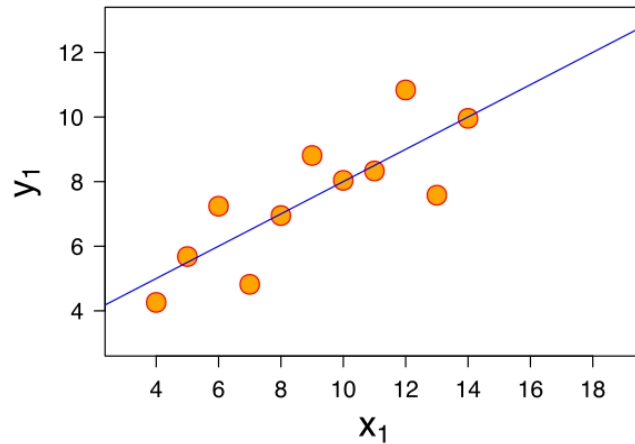
Its possible to compute high-quality models with very few passes, sometime with less than one pass over a large dataset.



# R<sup>2</sup>-values and P-values



We can **always** fit a linear model to any dataset, but how do we know if there is a **real linear relationship**?



# R<sup>2</sup>-values and P-values



**Approach:** Use a hypothesis test. The null hypothesis is that there is no linear relationship ( $\beta = 0$ ).

**Statistic:** Some value which should be small under the null hypothesis, and large if the alternate hypothesis is true.

**R-squared:** a suitable statistic. Let  $\hat{y} = X \hat{\beta}$  be a predicted value, and  $\bar{y}$  be the sample mean. Then the R-squared statistic is

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

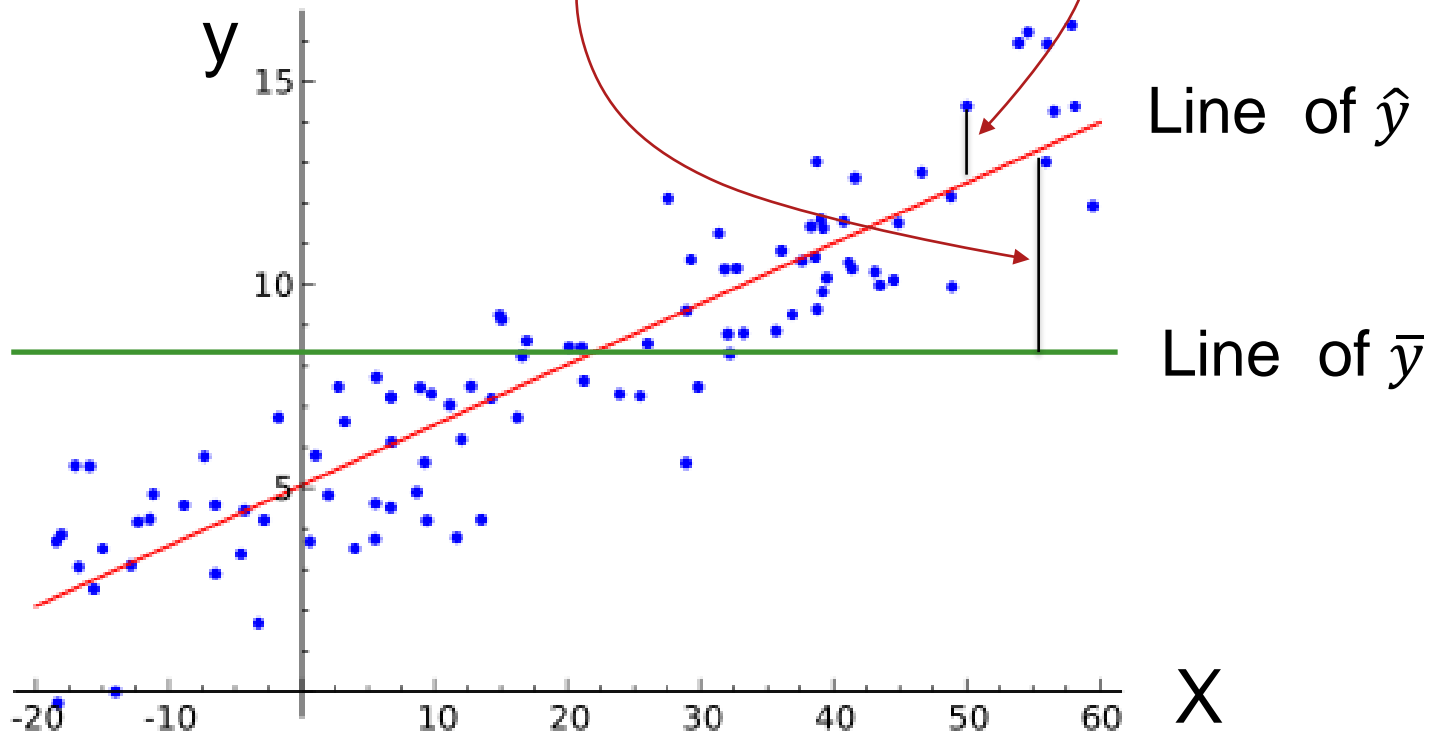
And can be described as the fraction of the total variance not explained by the model.

# R-squared



$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

Small if good fit



# R<sup>2</sup>-values and P-values



**Statistic:** From R-squared we can derive another statistic (using degrees of freedom) that has a standard distribution called an **F-distribution**.

From the CDF for the F-distribution, we can derive a **P-value** for the data.

The P-value is, as usual, the probability of observing the data under the null hypothesis of no linear relationship.

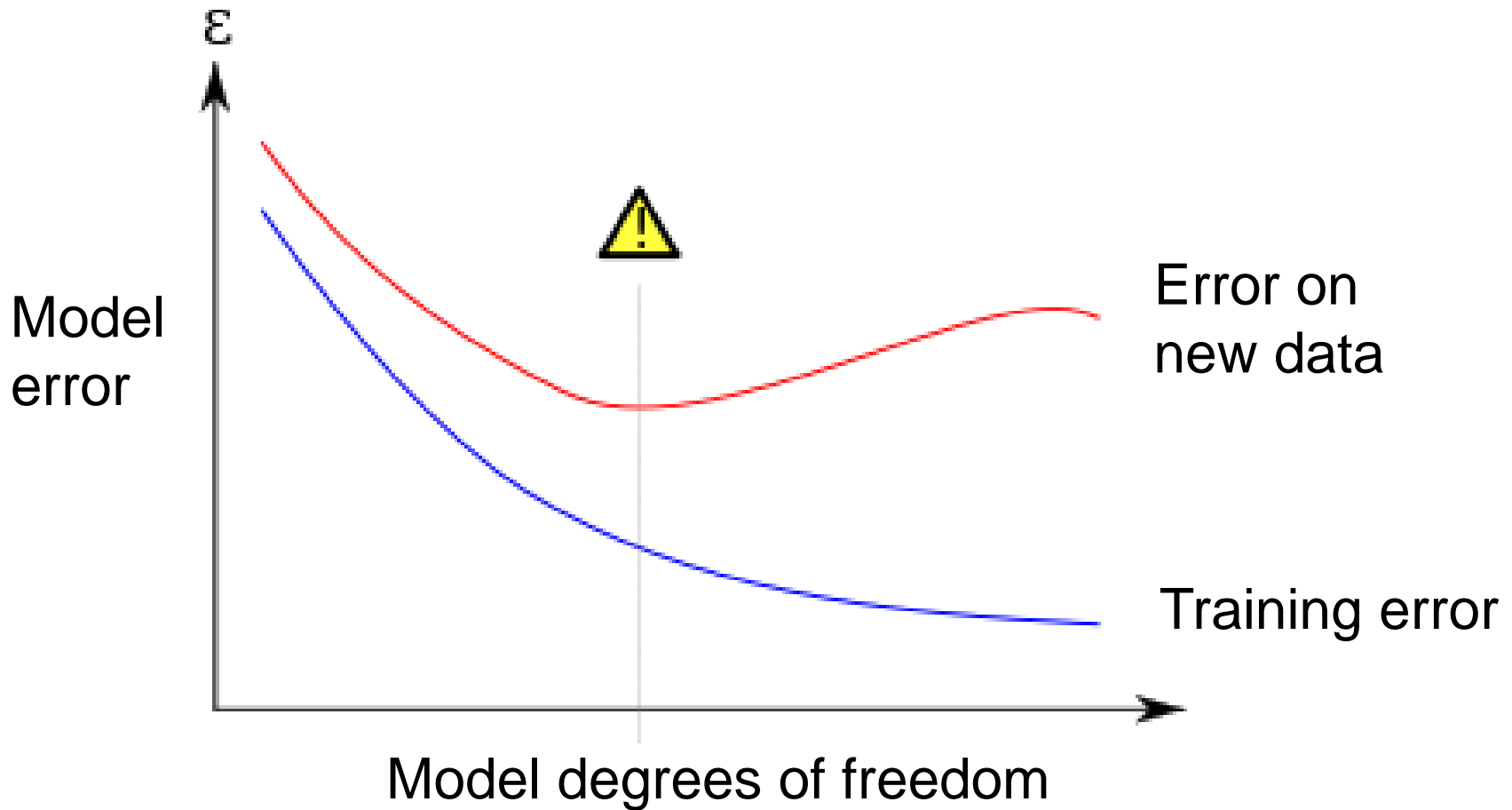
If **p is small**, say less than 0.05, we conclude that **there is a linear relationship**.

# Over-fitting



- Your model should ideally fit an **infinite sample** of the type of data you're interested in.
- In reality, you only have a **finite set** to train on. A good model for this subset is a good model for the infinite set, up to a point.
- Beyond that point, the model quality (measured on new data) starts to **decrease**.
- Beyond that point, the model is over-fitting the data.

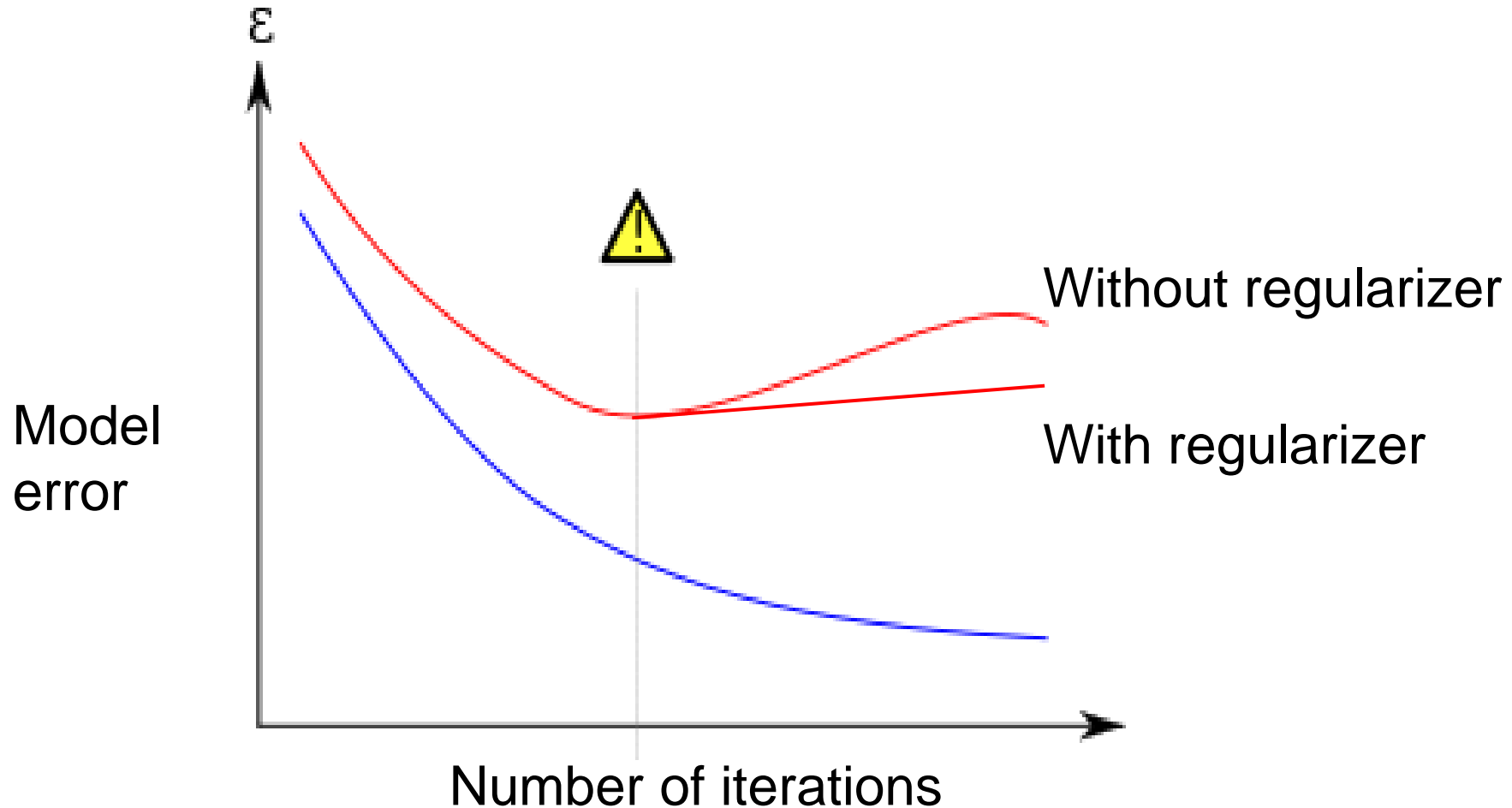
# Over-fitting



# Regularization and Over-fitting



Adding a regularizer:



# Cross-Validation



- Cross-validation involves **partitioning** your data into distinct **training** and **test** subsets.
- The test set **should never** be used to **train** the model.
- The test set is then used to **evaluate** the model after training.

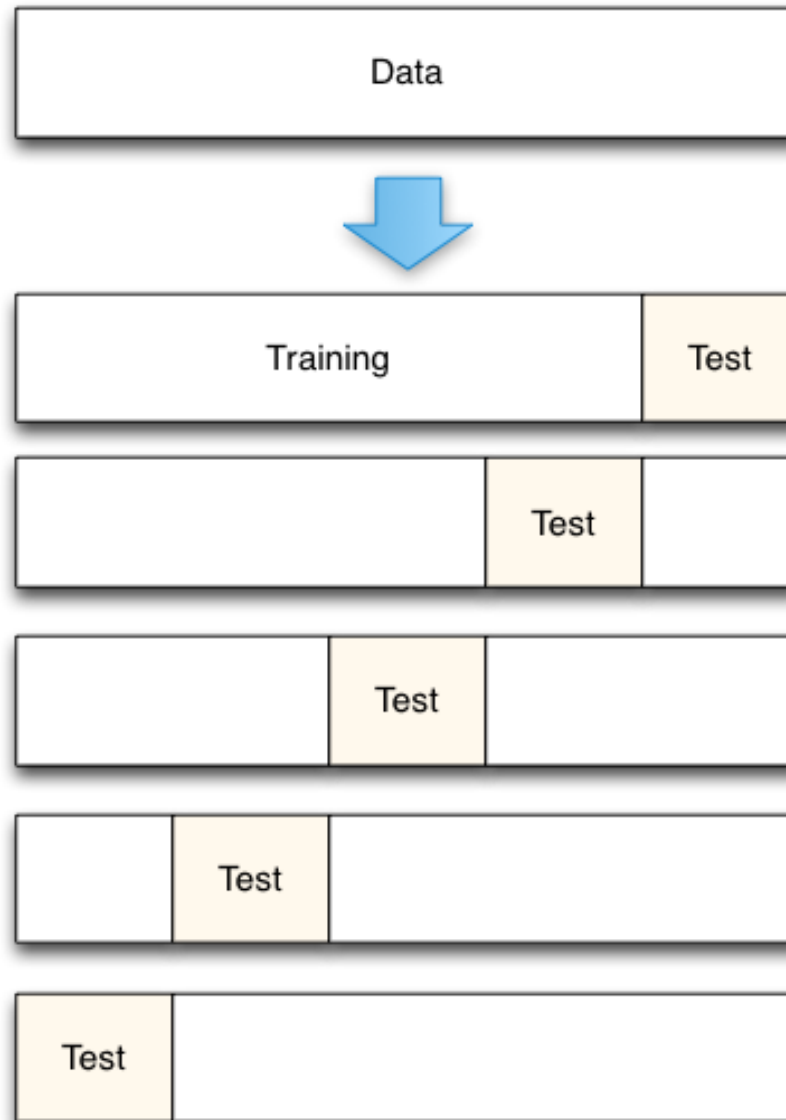


# K-fold Cross-Validation



- To get more accurate estimates of performance you can do this  $k$  times.
- Break the data into  $k$  equal-sized subsets  $A_i$
- For each  $i$  in  $1, \dots, k$  do:
  - Train a model on all the other folds  $A_1, \dots, A_{i-1}, A_{i+1}, \dots, A_k$
  - Test the model on  $A_i$
- Compute the **average performance** of the  $k$  runs

# 5-fold Cross-Validation





- **Exciting times to study Computer Science**
- **Advances in Sensing, Computing, Networking, Applied Mathematics, Statistics, Machine Learning, Artificial Intelligence enable Distributed, Networked, Human-like Intelligent Systems**
- **Engineering of Complex Systems require advanced platforms, tools and software/systems engineering practices.**

# 'Thank you' -- The End



FIAS Frankfurt Institute  
for Advanced Studies



GOETHE  
UNIVERSITÄT  
FRANKFURT AM MAIN



# Backup