

# Data Science

Alexander Schliep

CSE

Gothenburg University | Chalmers


**Data**



## Changing face of data...



## Every 60 seconds

-  **98,000+** tweets
-  **695,000** status updates
-  **11million** instant messages
-  **698,445** Google searches
-  **168 million+** emails sent
-  **1,820TB** of data created
-  **217** new mobile web users

**Yottabytes**

# Interesting sources of data

- Sensor networks
- Smart phones
- Quantified self
- Internet of things
- Personalized medicine
- Citizen Science



# Technological success stories



INTELLISAFE

AUTOPILOT

TRAVEL CALMER, SAFER, CLEANER

POA 708



# Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu,  
Zhifeng Chen, Nikhil Thorat  
melvinp,schuster,qvl,krikun,yonghui,zhifengc,nsthorat@google.com

Fernanda Viégas, Martin Wattenberg, Greg Corrado,  
Ludovic Hughes, Jeffrey Dean

... our models can also learn to perform implicit bridging between language pairs never seen explicitly during training...

**Abstract**  
We propose a simple, elegant solution to use a single Neural Machine Translation (NMT) model to translate between multiple languages. Our solution requires no change in the model architecture from our base system but instead introduces an artificial token at the beginning of the input sentence to specify the required target language. The rest of the model, which includes encoder, decoder and attention, remains unchanged and is shared across all languages. Using a shared wordpiece vocabulary, our approach enables Multilingual NMT using a single model without any increase in parameters, which is significantly smaller than previous proposals for Multilingual NMT. Our method often improves the translation quality of all involved language pairs, even while keeping the total number of model parameters constant. On the WMT'14 benchmarks, a single multilingual model achieves comparable performance for English→French and surpasses state-of-the-art results for English→German. Similarly, a single multilingual model surpasses state-of-the-art results for French→English and German→English on WMT'14 and WMT'15 benchmarks respectively. On production corpora, multilingual models of up to twelve language pairs allow for better translation of many individual pairs. In addition to improving the translation quality of language pairs that the model was trained with, our models can also learn to perform implicit bridging between language pairs never seen explicitly during training, showing that transfer learning and zero-shot translation is possible for neural translation. Finally, we show analyses that hints at a universal interlingua representation in our models and show some interesting examples when mixing languages.

11.04558v1 [cs.CL] 14 Nov 2016

# IBM Watson



"a technology platform that uses natural language processing and machine learning to reveal insights from large amounts of unstructured data"

<http://www.ibm.com/smarterplanet/us/en/ibmwatson/what-is-watson.html>



# **Application success stories**

*Case Study:*

# **Influences in English Literature**



# Large-scale literature analysis

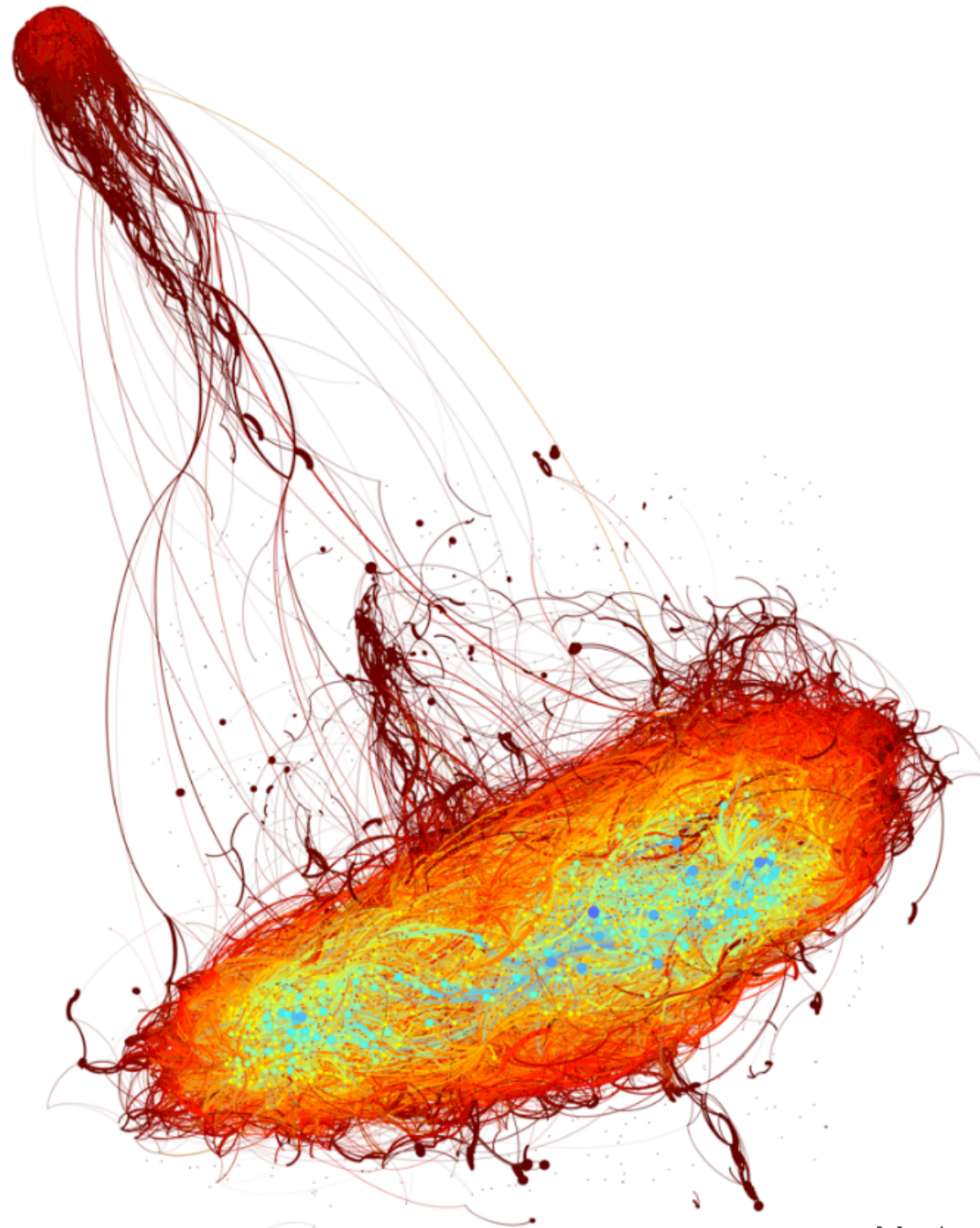
- 4357 novels
- 150 Years (average of 29 books per year)
- British (73%), Irish (5%), and American (22%)
- Male (55%), Female (36%), and Anonymous (9%)
- 1875 unique authors (2.32 books per author)

Author	Title	Distance
Austen, Jane	<i>Pride and Prejudice</i>	0.000000
Austen, Jane	<i>Emma</i>	1.260236
Austen, Jane	<i>Sense and Sensibility</i>	1.268725
Austen, Jane	<i>Mansfield Park</i>	1.421373
Austen, Jane	<i>Northanger Abbey</i>	1.600394
Austen, Jane	<i>Persuasion</i>	1.673071

Author	Title	Distance
Gaskell, Elizabeth	<i>Ruth</i>	
Craik, Dinah Maria	<i>Olive</i>	Dickens, Charles <i>A Tale of Two Cities</i> 0.000000
Church A. B. Mrs.	<i>Greymore a Story of Country Life</i>	Kirkland, Caroline Matilda <i>The Fountain and the Bottle</i> 1.361071
Grant, Louisa	<i>Charles Stanley</i>	Milman, Edward Augustus <i>Arthur Conway; or, Scenes in the Tropics</i> 1.385395
Tainsh, Edward Campbell	<i>One Maiden Only</i>	Liddell Charles Francis <i>Hidden Links; or, The Schoolfellows</i> 1.466322
		Dickens, Charles <i>The Old Curiosity Shop</i> 1.492650
		Armstrong, Francis Claudius <i>The Pirates of The Foam</i> 1.500570
		Spofford, Harriet Elizabeth Prescott <i>Sir Rohan's Ghost</i> 1.508781
		Fay, Theodore Sedgwick <i>Norman Leslie; A Tale of the Present Times</i> 1.509204
		Shillaber Benjamin Penhallow <i>Knitting Work; A Web of Many Textures</i> 1.534282
		Dickens, Charles <i>Barnaby Rudge</i> 1.544074
		Paulding, James Kirke <i>Chronicles of the City of Gotham</i> 1.548381

<http://www.matthewjockers.net/slides-etc/>





*Case Study:*

# **Society and policy**





# UNITED NATIONS GLOBAL PULSE

Harnessing big data for development and humanitarian action

Search  SEARCH



Home

- ABOUT
- PROJECTS
- LABS
- NEWS
- CHALLENGES
- PRIVACY
- PARTNERSHIPS
- RESOURCES
- CONTACT
- HOME

### SUBSCRIBE TO OUR NEWSLETTER

GO

## Projects

Welcome to the repository of Global Pulse's projects. Find out more about collaborative research, prototypes and experiments analyzing digital data to support global development and humanitarian action.



National Citizen Feedback Dashboard For Enhanced Local Government Decision-Making



Tracking The Impact Of Climate Anomalies



Publication: Integrating Big Data Into The Monitoring And Evaluation Of Development Programmes (2016)



Using Financial Data To Understand Macroeconomic Issues In Cambodia



Monitoring Social Response Before And After Natural Disasters With Data Analytics



Making Ugandan Community Radio Machine-Readable Using Speech Recognition Technology



Sex Disaggregation Of Social Media Posts

## BROWSE BY LAB

- Jakarta
- Kampala
- New York

## BROWSE BY PROGRAMME

- Climate & Resilience
- Data Privacy & Protection
- Economic Well-being
- Food & Agriculture
- Gender
- Humanitarian Action
- Public Health
- Real-time Evaluation
- The Sustainable Development Goals (SDGs)

## BROWSE BY REGION

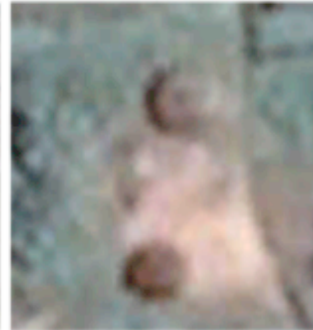
- Africa
- Asia
- Europe
- Global
- Latin America and the Caribbean
- Northern America
- Oceania

# Measuring Poverty

Photo



Satellite image





*Case Study:*

# **Ecology**

# eBird

- Quantified Bird Watching
- Bird watcher as "sensors"
- Citizen Science

From <http://ebird.org>

# eBird

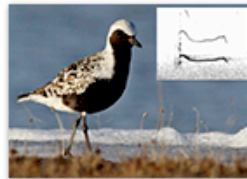


[Home](#) [About](#) [Submit Observations](#) [Explore Data](#) [My eBird](#) [Help](#)

[Sign In or Register](#)

[Language](#) ▾

## View and Explore Data



### [Search Photos and Sounds](#) <sup>NEW</sup>

Explore media through the Macaulay Library



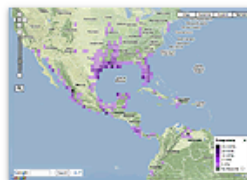
### [Explore a Region](#)

Recent sightings, checklists, birding activity, best hotspots, and top birders for a county, state, province, or country.



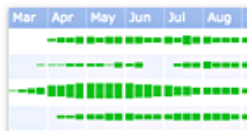
### [Explore Hotspots](#)

Discover the best places for birding nearby or around the world.



### [Species Maps](#)

Explore interactive range maps by species or subspecies — zoom in for details



### [Bar Charts](#)

Find out what birds to expect throughout the year in a region or location

### **Your Totals**

Track your totals and compare with other eBirders.

#### [Yard Totals](#)

How many species and checklists have you submitted for your yard?

#### [Patch Totals](#)

How many have you submitted for your favorite birding patches?

#### [Top 100](#)

Compare with the top eBirders in your region.

### **Species You Need**

Tools to find species you haven't seen yet.

#### [Target Species](#)

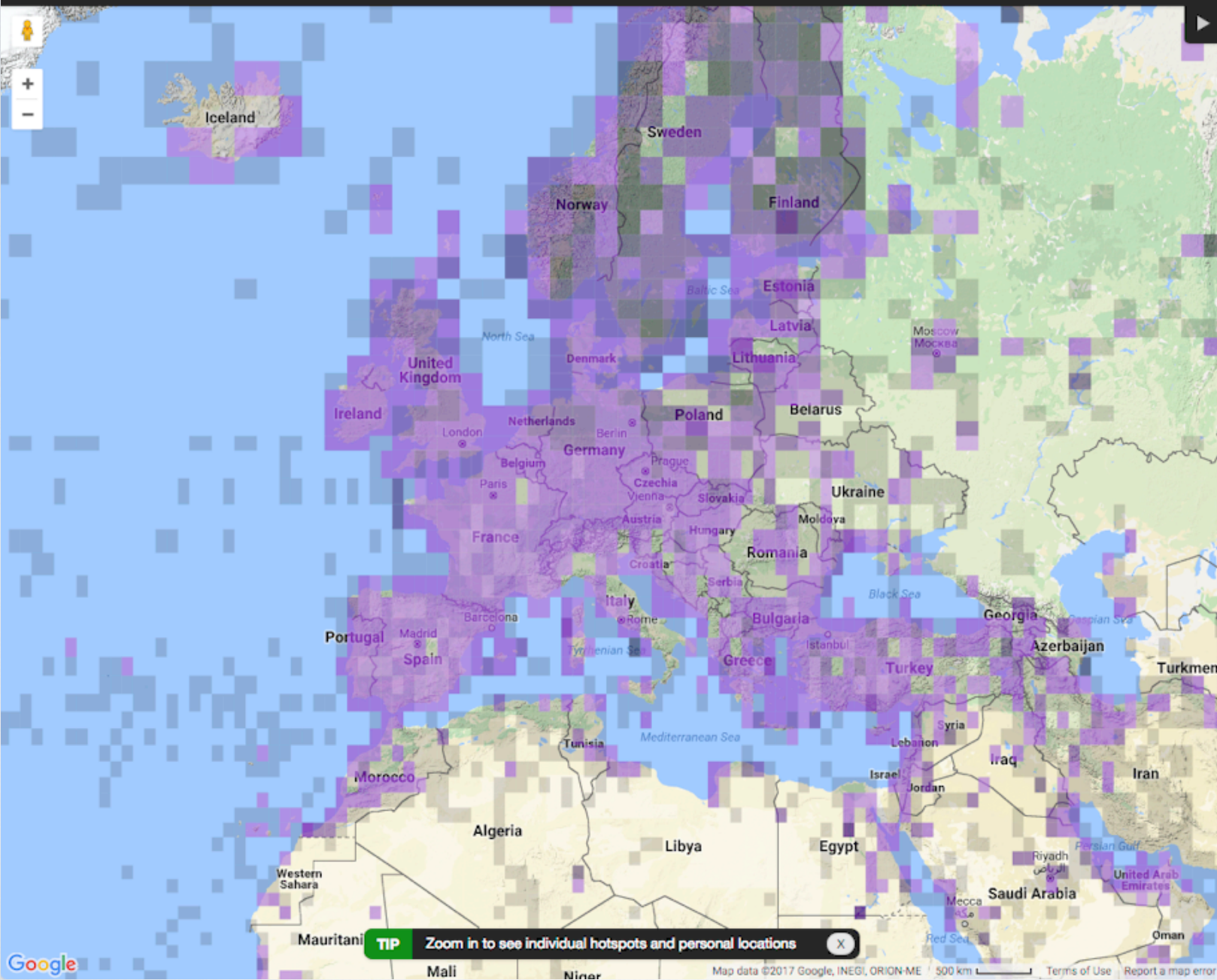
Prioritized list of county, state, or life birds that you can expect to find in a region

#### [Alerts](#)

Reports and email alerts for rarities and species you haven't seen



Species: Willow Warbler Date: Year-Round, All Years Location: Enter place name or address...



Zoom Tool  
Full Species Range

- Terrain
- Street
- Satellite
- Hybrid

Explore Rich Media  
Only show location photos, audio, or video

Show Points So  
Display points at both scales when possible (points max)

- 40-100%
- 25-40%
- 10-25%
- 2-10%
- 0-2%
- Not reported

TIP Zoom in to see individual hotspots and personal locations

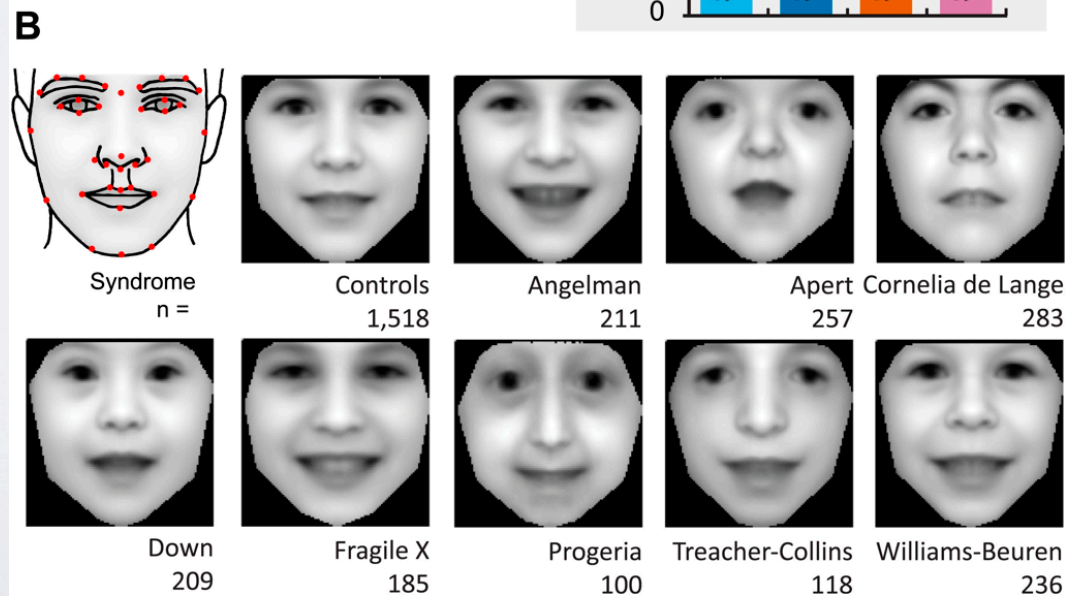
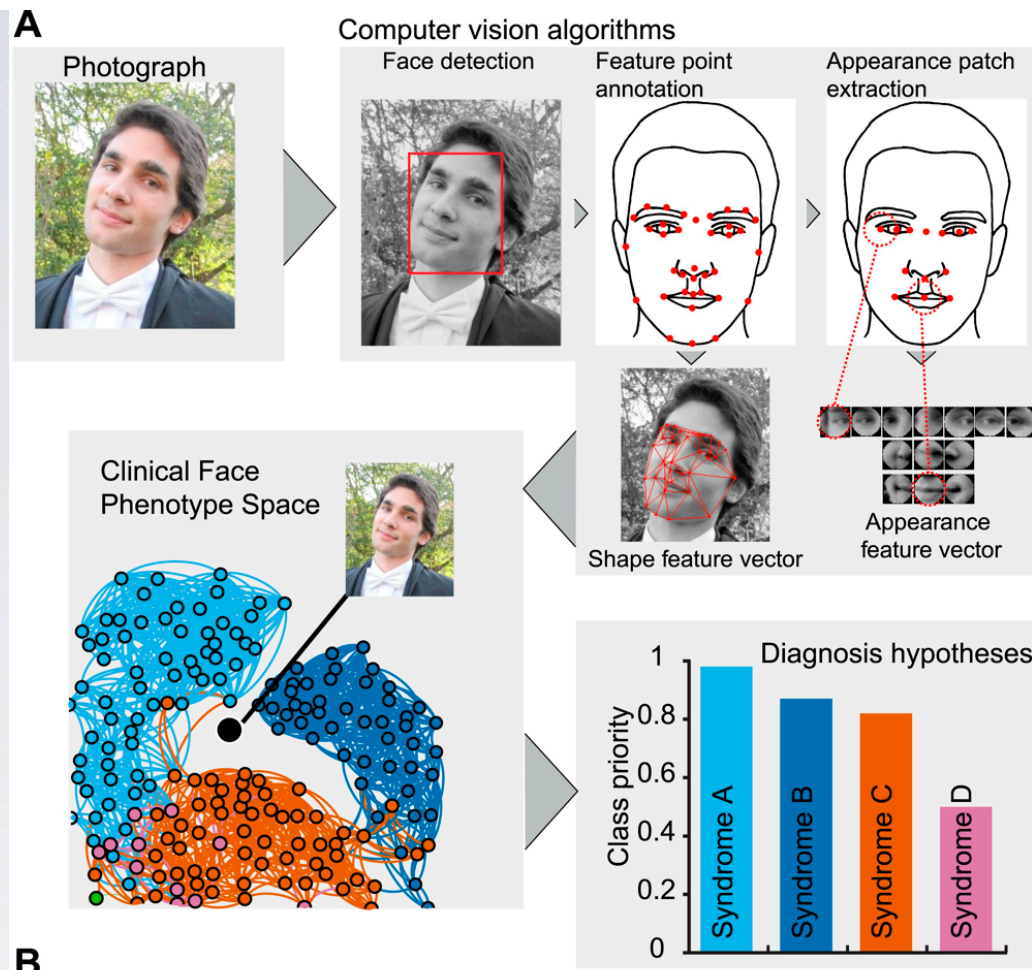
From <http://ebird.org>

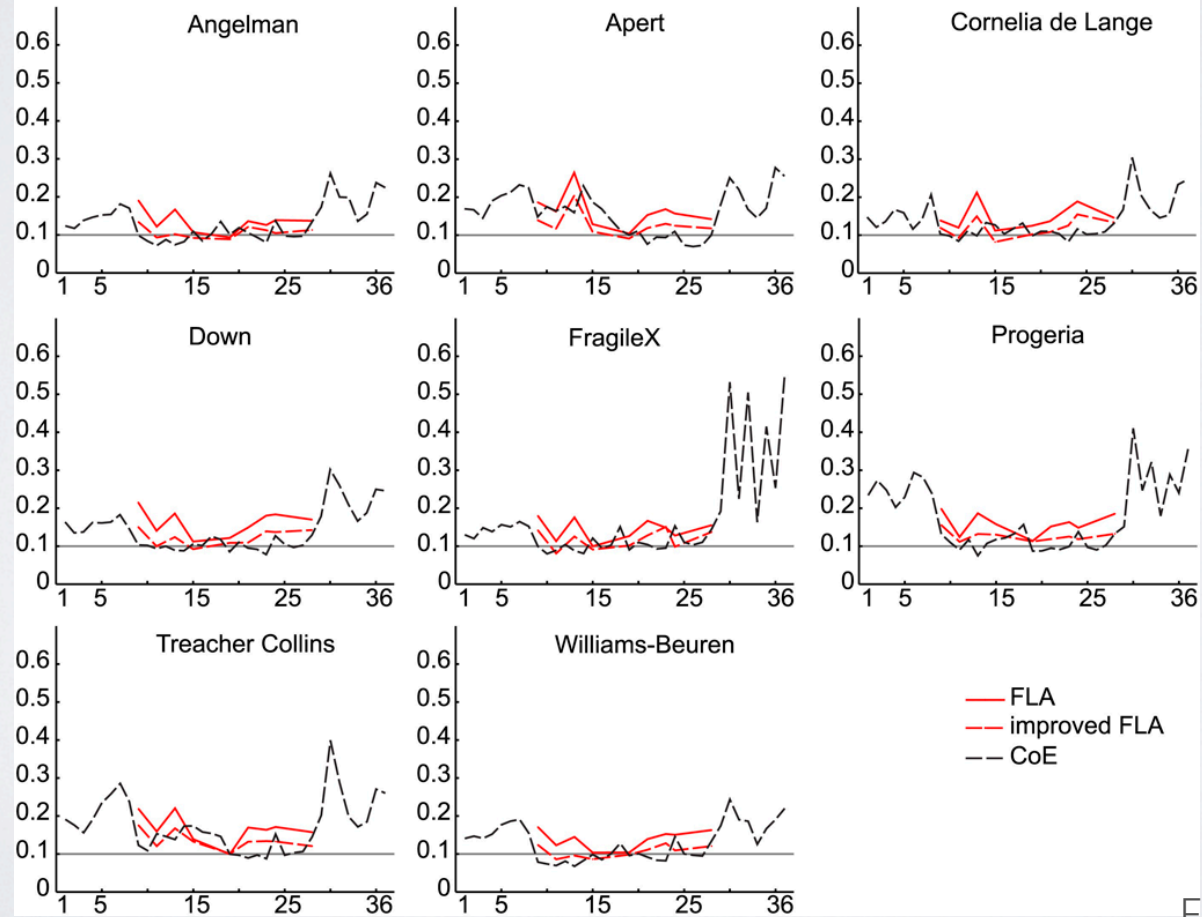
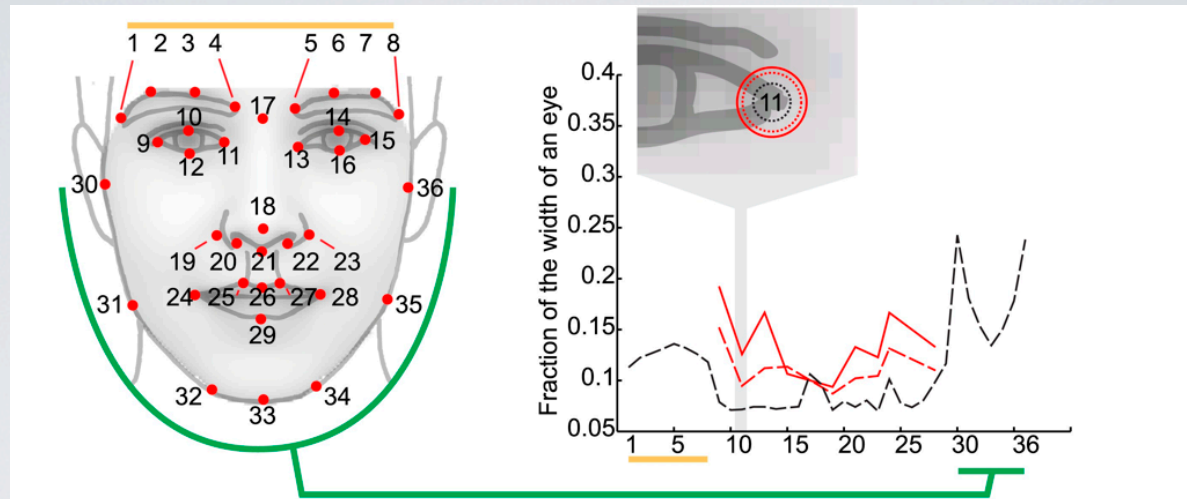
*Case Study:*

# **Diagnosing rare genetic diseases from photographs**



# Diagnosing rare genetic diseases from photographs







# Possible Definitions

# Introducing Data Science

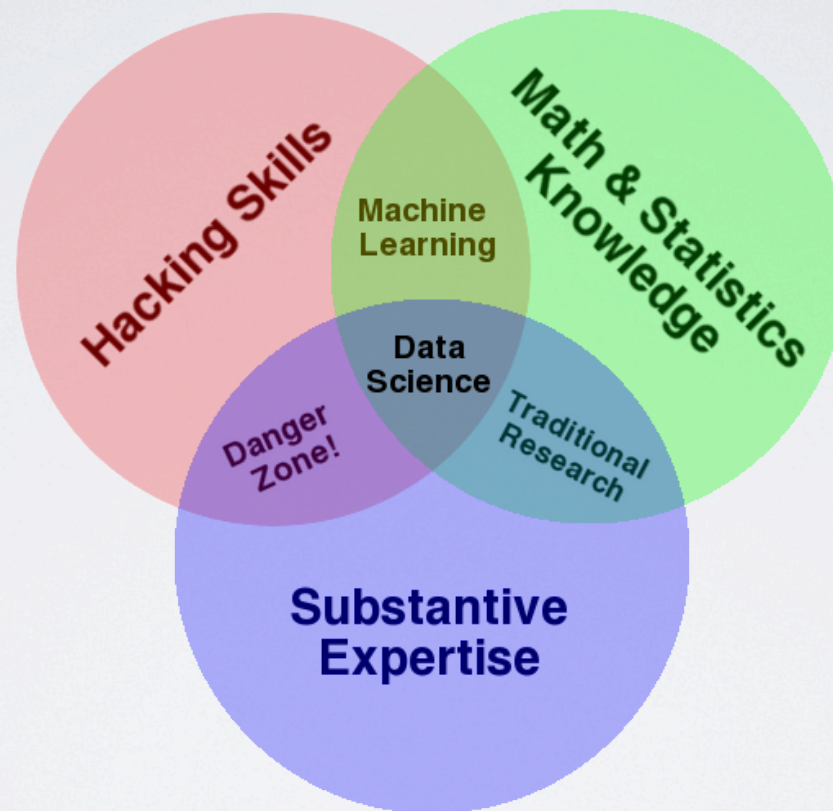
Data Science is concerned with **extracting meaning from big data**.

Central topics within Data Science include:

- **data mining**
- **machine learning**
- **databases**
- **the application of data science methods** in natural sciences, life sciences, humanities and social sciences, as well as in industry and society.



# The Data Science Venn Diagram



# Big Data techniques and

## Techniques

- A/B testing, Association rule learning, Classification, Cluster analysis, Crowdsourcing, Data fusion and data integration, Data mining, Ensemble learning, Genetic algorithms, Machine learning, Natural language processing, Neural network, Network analysis, Optimization, Pattern recognition, Predictive modelling, Regression, Sentiment analysis, Signal processing, Spatial analysis, Statistics, Supervised learning, Simulation, Time series analysis, Unsupervised learning, Visualization

## Technologies

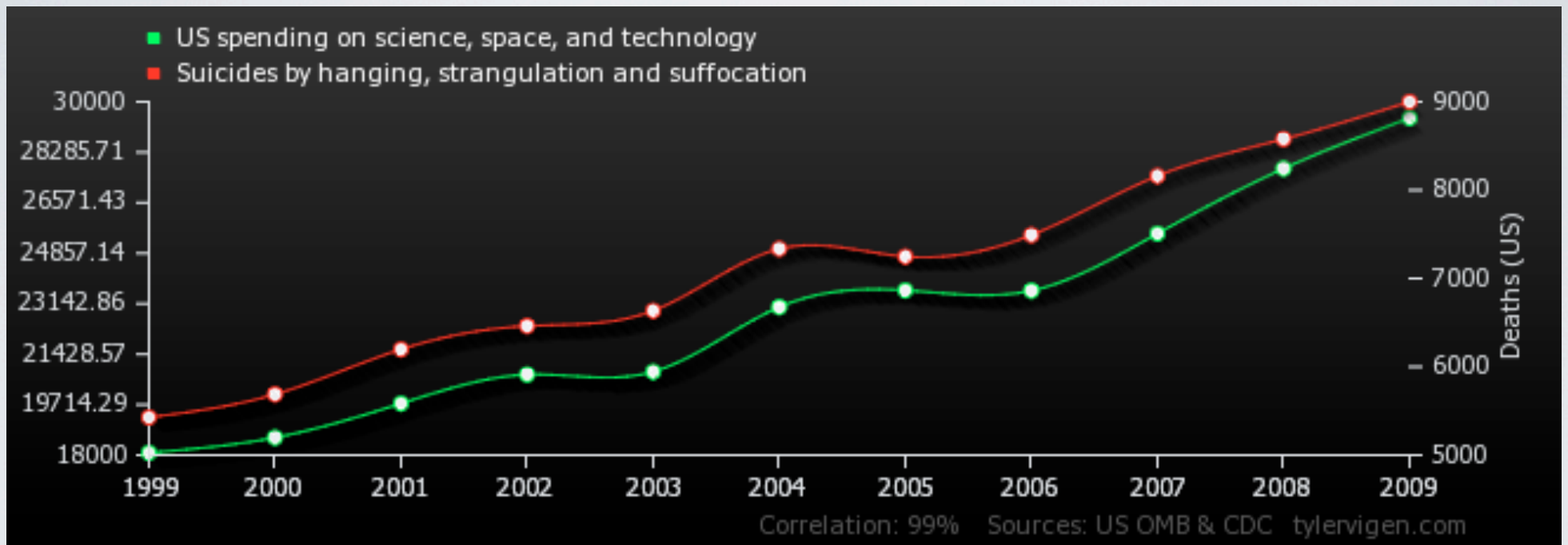
- Big Table, Business Intelligence (BI), Cassandra, Cloud computing, Data mart,

McKinsey Global Institute (2011) "Big data: The next frontier for innovation, competition, and productivity"



**Necessary skills**

# Statistics



<http://tylervigen.com/spurious-correlations>



# Algorithms: Tera → Peta Bytes

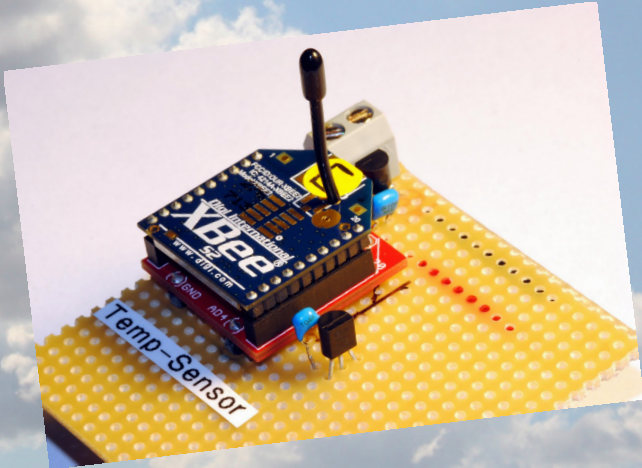
- RAM time to move
  - 15 minutes
- 1Gb WAN move time
  - 10 hours (\$1000)
- Disk Cost
  - 7 disks = \$5000 (SCSI)
- Disk Power
  - 100 Watts
- Disk Weight
  - 5.6 Kg
- Disk Footprint
  - Inside machine
- **RAM time to move**
  - **2 months**
- **1Gb WAN move time**
  - **14 months (\$1 million)**
- **Disk Cost**
  - **6800 Disks + 490 units + 32 racks = \$7 million**
- **Disk Power**
  - **100 Kilowatts**
- **Disk Weight**
  - **33 Tonnes**
- **Disk Footprint**
  - **60 m<sup>2</sup>**

May 2003 Approximately Correct

See also *Distributed Computing Economics* Jim Gray, Microsoft Research, MSR-TR-2003-24



# Systems





# Domain knowledge

- Science
- Humanities
- Industry
- Business
- Sports
- Art, ...

# Study program



# Big Data Seminars at Chalmers

Speakers from industry and academia

Abstracts and some presentation slides online:

<https://www.chalmers.se/en/areas-of-advance/ict/research/big-data/Pages/>

# Some relevant courses

CIU187 Information visualization

FFR105 Stochastic optimization  
algorithms

FFR135 Artificial neural networks

MVE186 Computer intensive statistical  
methods MSA100

MVE440 Statistical Learning for Big Data  
(MSA220)

RRY025 Image processing (ASM420)

TDA231 Algorithms for machine learning  
and inference (DIT 380)

TIN173 Artificial intelligence (DIT410)

TMS150 Stochastic data processing and  
simulation (MSG400)

DAT300 ICT support for adaptiveness and  
security in the smart grid (DIT 668)

SSY115 eHealth

VVT105 Geographical information systems

From the Applied Data Science MS program:

Applied Machine Learning

Techniques for Large-scale Data



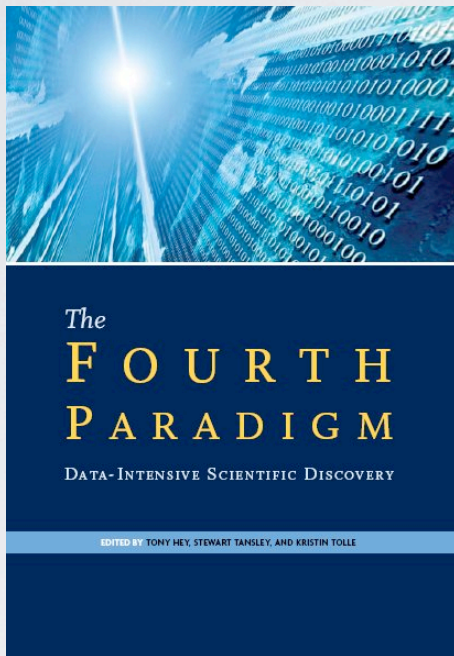
# Some Master's projects

- Constructing a Context-aware Recommender System with Web Sessions (3Bits Consulting AB)
- Machine Learning for On-line Advertising Using Contextual Information (Admeta)
- The Identification of Target Proteins from Patents – Mining of biological entities from a full-text patent database (AstraZeneca)
- Browser Fingerprinting (Burt)
- Learning to rank, a supervised approach for ranking of documents (Findwise)
- Entity Entity Disambiguation in Anonymized Graphs Using Graph Kernels (Recorded Future)
- Using Classification Algorithms for Smart Suggestions in Accounting Systems (SpeedLedger)
- Cluster User Music Sessions (Spotify)
- Extracting Data from NoSQL Databases – A Step towards Interactive Visual Analysis of NoSQL Data (TIBCO Software)

**Job market**



# The Fourth Paradigm



Increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore massive datasets.

<http://research.microsoft.com/en-us/collaboration/fourthparadigm/>

# Shortage of talent

"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."

McKinsey Global Institute (2011) "Big data: The next frontier for innovation, competition, and productivity"

[http://www.mckinsey.com/insights/business-technology/big data the next frontier for innovation](http://www.mckinsey.com/insights/business-technology/big-data-the-next-frontier-for-innovation)



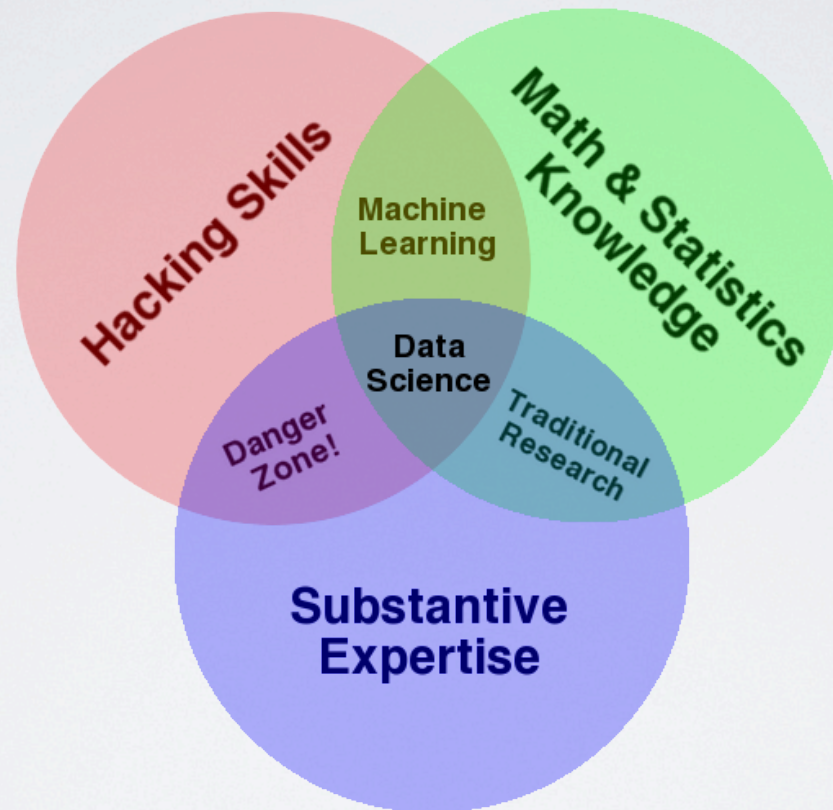
**"If you want a career in medicine these days you're better off studying mathematics or computing than biology."**

Sir Rory Collins, head of clinical trials at Oxford University  
BBC 10/14/2016

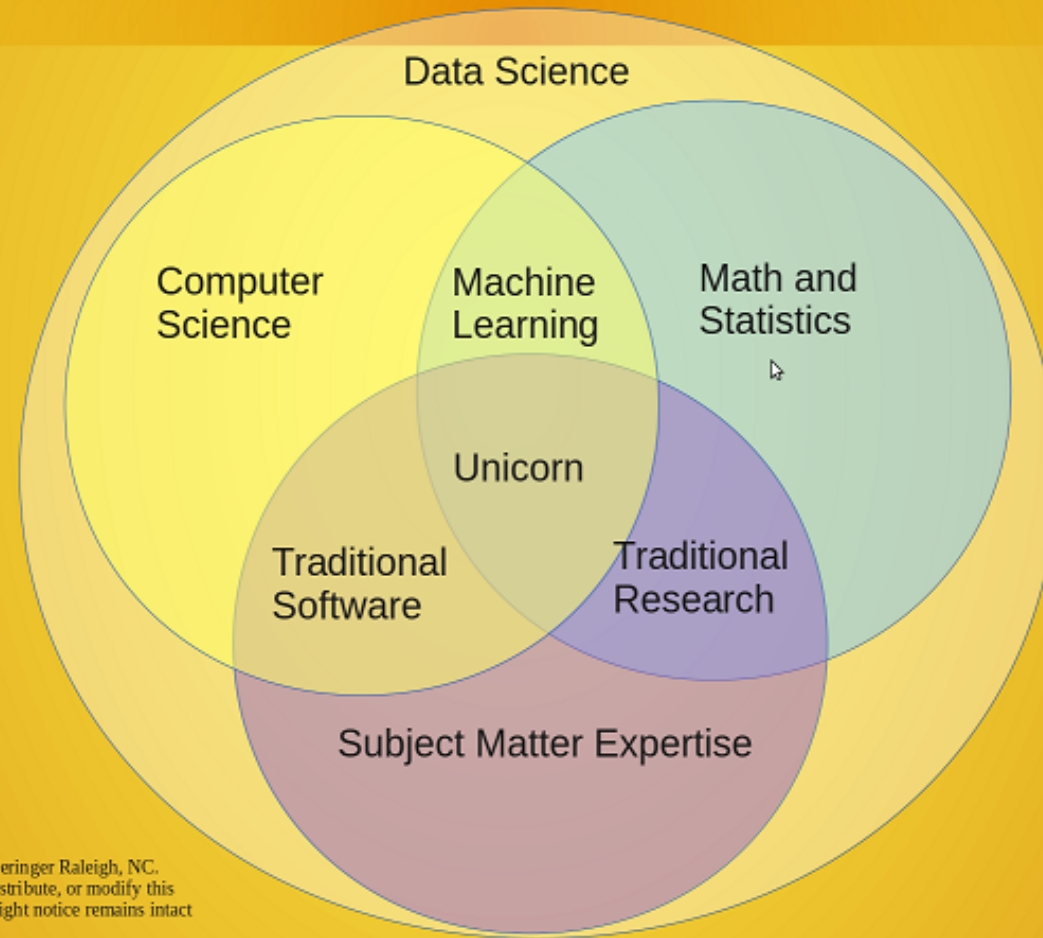
**A perspective ...**



# The Data Science Venn Diagram



# Data Science Venn Diagram v2.0



Copyright © 2014 by Steven Geringer Raleigh, NC.  
Permission is granted to use, distribute, or modify this  
image, provided that this copyright notice remains intact



**Thank you.**