UNIVERSITY of WASHINGTON

Data Science Curricula at the University of Washington

Bill Howe, PhD

Director of Research,
Scalable Data Analytics

University of Washington
eScience Institute

eScience Institute

http://escience.washington.edu

The University of Washington eScience Institute

- Rationale
  - The exponential increase in sensors is transitioning all fields of science and engineering from data-poor to data-rich
  - Techniques and technologies include
    - Sensors and sensor networks, databases, data mining, machine learning, visualization, cluster/cloud computing
  - If these techniques and technologies are not widely available and widely practiced, UW will cease to be competitive
- Mission
  - Help position the University of Washington at the forefront of research both in modern eScience techniques and technologies, and in the fields that depend upon them
- Strategy
  - Bootstrap a cadre of Research Scientists
  - Add faculty in key fields
  - Build out a "consultancy" of students and non-research staff

eScience Big Data Group

Bill Howe, Phd (databases, cloud, data-intensive scalable computing, visualization)
Director of Research, Scalable Data Analytics

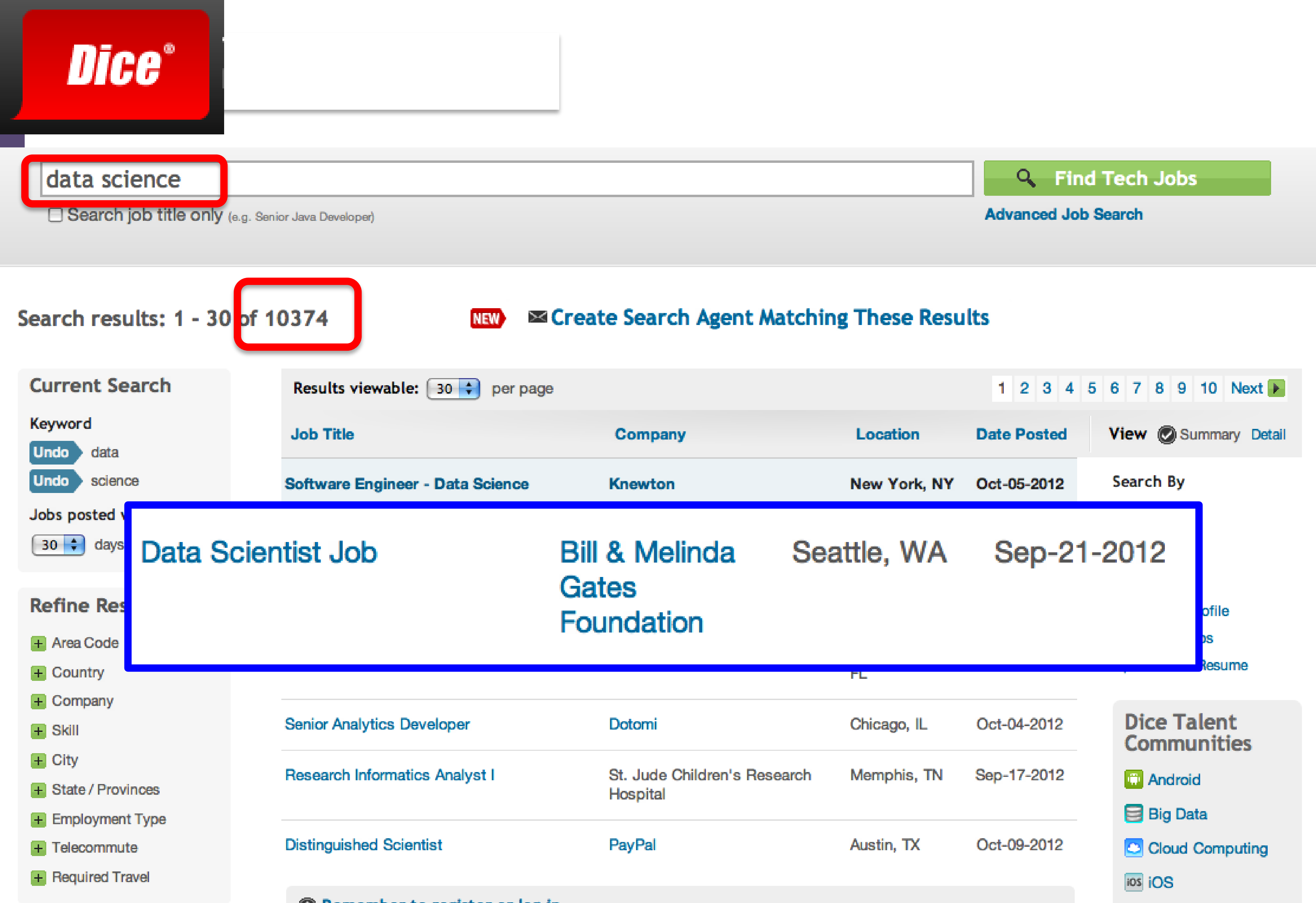
Staff
- Seung-Hee Bae, Phd (postdoc, scalable machine learning algorithms)
- Dan Halperin, Phd (postdoc; scalable systems)
- Sagar Chitnis, Research Engineer (Azure, databases, web services)
- (alumna) Marianne Shaw, Phd (hadoop, semantic graph databases)
- (alumna) Alicia Key, Research Engineer (visualization, web applications)

Students
- Scott Moe (2nd yr Phd, Applied Math)
- Daniel Perry (2nd yr Phd, HCDE)

Partners
- CSE DB Faculty: Magda Balazinska, Dan Suciu
- CSE students: Paris Koutris, Prasang Upadhyaya,
- UW-IT (web applications, QA/support)
- Cecilia Aragon, Phd, Associate Professor, HCDE (visualization, scientific applications)

eScience ~= Data Science

…modulo application area


Dice
data science
Find Tech Jobs
Search job title only (e.g. Senior Java Developer)
Advanced Job Search
Search results: 1 - 30 of 10374
NEW
Create Search Agent Matching These Results
Current Search
Keyword
Undo data
Undo science
Jobs posted
30 days
Refine Res
Area Code
Country
Company
Skill
City
State / Provinces
Employment Type
Telecommute
Required Travel
Results viewable: 30 per page
1 2 3 4 5 6 7 8 9 10 Next
Job Title
Company
Location
Date Posted
View Summary Detail
Search By
Software Engineer - Data Science
Knewton
New York, NY
Oct-05-2012
Data Scientist Job
Bill & Melinda Gates Foundation
Seattle, WA
Sep-21-2012
Senior Analytics Developer
Dotomi
Chicago, IL
Oct-04-2012
Research Informatics Analyst I
St. Jude Children's Research Hospital
Memphis, TN
Sep-17-2012
Distinguished Scientist
PayPal
Austin, TX
Oct-09-2012
Remember to register or log-in
Dice Talent Communities
Android
Big Data
Cloud Computing
iOS

UW Data Science Education Efforts

	Students				Non-Students	
	CS/Informatics		Non-Major		professionals	researchers
	undergrads	grads	undergrads	grads		
UWEO Data Science Certificate						
Graduate Certificate in Big Data						
CS Data Management Courses						
eScience workshops						
Intro to data programming						
eScience Masters (planned)						
Coursera Course: Intro to Data Science						

Previous courses:
Scientific Data Management, Graduate CS, Summer 2006, Portland State University
Scientific Data Management, Graduate CS, Spring 2010, University of Washington

UNIVERSITY of WASHING

Huge number of relevant courses, new and existing.

- Concepts in Computing with Data, Berkeley
- Practical Machine Learning, Berkeley
- Artificial Intelligence, Berkeley
- Visualization, Berkeley
- Data Mining and Analytics in Intelligent Business Services, Berkeley
- Data Science and Analytics: Thought Leaders, Berkeley
- Scalable Machine Learning, Berkeley
- Analyzing Big Data with Twitter, Berkeley
- Machine Learning, Stanford
- Paradigms for Computing with Data, Stanford
- Mining Massive Data Sets, Stanford
- Data Visualization, Stanford
- Algorithms for Massive Data Set Analysis, Stanford
- Research Topics in Interactive Data Analysis, Stanford
- Data Mining, Stanford
- Machine Learning, CMU
- Statistical Computing, CMU
- Machine Learning with Large Datasets, CMU
- Machine Learning, MIT
- Data Mining, MIT
- Statistical Learning Theory and Applications, MIT
- Data Literacy, MIT
- Introduction to Data Mining, UIUC
- Learning from Data, Caltech
- Introduction to Statistics, Harvard
- Data-Intensive Information Processing Applications, University of Maryland
- Statistical Inference, UPenn
- Introduction to Data Science, Columbia
- Dealing with Massive Data, Columbia
- Data-Driven Modeling, Columbia
- Introduction to Data Mining and Analysis, Georgia Tech
- Computational Data Analysis: Foundations of Machine Learning and Data Mining, Georgia Tech
- Applied Statistical Computing, Iowa State
- Data Visualization, Rice
- Data Warehousing and Data Mining, NYU
- Data Mining in Engineering, Toronto
- Machine Learning and Data Mining, UC Irvine
- Knowledge Discovery from Data, Cal Poly
- Large Scale Learning, University of Chicago
- Data Science: Large-scale Advanced Data Analysis, University of Florida
- Strategies for Statistical Data Analysis, Universität Leipzig
- Data Analysis, Johns Hopkins (via Coursera)
- Computing for Data Analysis, Johns Hopkins (via Coursera)

"I worry that the Data Scientist role is like the mythical "webmaster" of the 90s: master of all trades."

-- Aaron Kimball, CTO Wibidata

Breadth

tools	abstractions
Hadoop	MapReduce
PostgreSQL	Relational Algebra
glm(…) in R	Logistic Regression
Tableau	InfoVis

Depth

structures ⟷ statistics

structures	statistics
Management	Analysis
Relational Algebra	Linear Algebra
Standards	ad hoc files

Scale

desktop	cloud
main memory	distributed
R	Hadoop
local files	S3, Azure Storage

Target

hackers ⟷ analysts

Assume proficiency in Python, Java, R

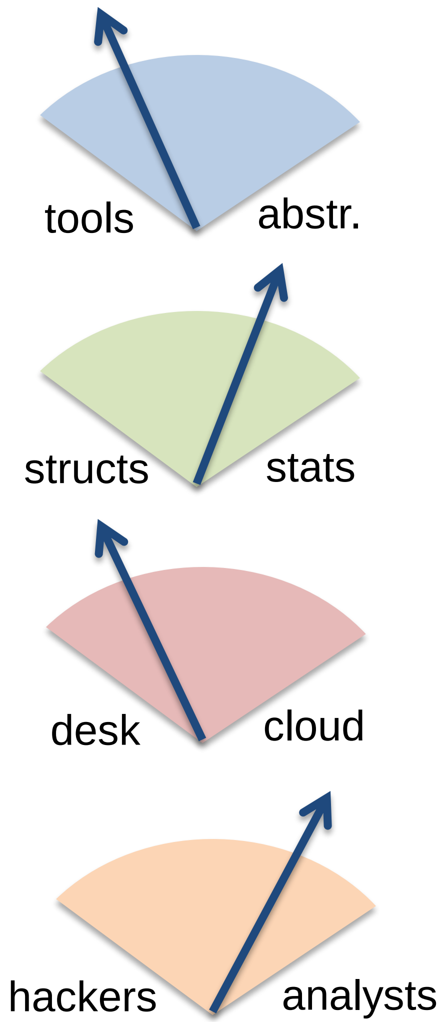
Assume little or no programming

Breadth

Depth

Scale

Target



Certificate in Statistical Analysis with R Programming

Approved by the UW Department of Statistics and UW Department of Applied Mathema

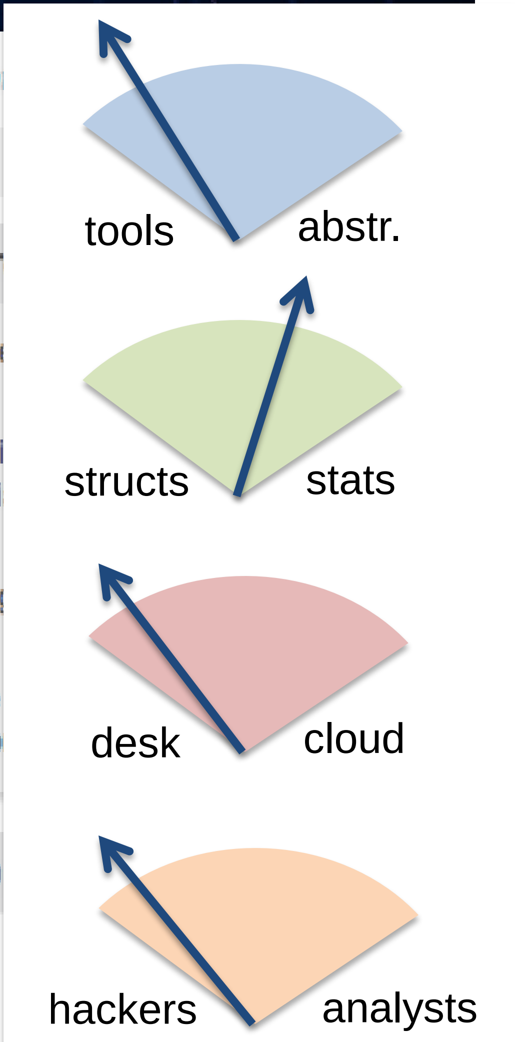
Certificates » **Statistical Analysis with R Programming**

Develop your statistical and analytical skills in the R programming environment. Master and apply a comprehensive range of statistical analyses and models, including linear regressions, multivariate analysis, machine learning algorithms and time series analysis. Learn and apply state of the art skills in data mining and big data management to derive meaning from raw data. Acquire a thorough understanding of the R programming source environment, and learn to maximize the visualization and graphical capabilities within R, including ggplot and lattice graphics. Use your skills in statistics and R to solve complex problems in such fields as finance, marketing, social media and genomics.

Program Featur

- Flexibility to take course both
- Virtual interaction with i instructor in real time vi (online program)
- Hands-on programming mining for analytics
- Instruction and real-life modeling technique fro

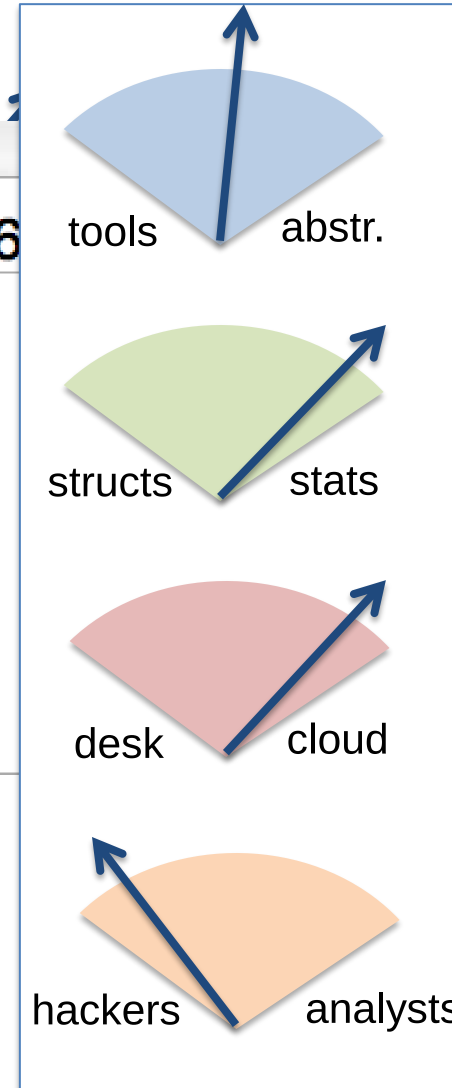
Who Should Ap



Carnegie Mellon University

of WASHINGTON

William W. Cohen



page | discussion | view source | history

Syllabus for Machine Learning with Large Datasets 10-6

This is the syllabus for Machine Learning with Large Datasets 10-605 in Spring 2012.

Contents [hide]

January

- Tues Jan 17. Overview of course, cost of various operations, asymptotic analysis.
- Thus Jan 19. Review of probabilities.
- Tues Jan 24. Streaming algorithms and Naive Bayes.
  - *New Assignment: streaming Naive Bayes 1 (with feature counts in memory).* PDF Handout
- Thus Jan 26. The stream-and-sort design pattern; Naive Bayes revisited.
- Tues Jan 31. Messages and records 1; Phrase finding.
  - **Assignment due: streaming Naive Bayes 1 (with feature counts in memory).**
  - *New Assignment: streaming Naive Bayes 2 (with feature counts on disk) with stream-and-sort.* PDF Handout

UNIVERSITY of WASHINGTON

Dan Suciu

Magda Balazinska

University of Washington
Computer Science & Engineering

CSE 344 Introduction to Data Management

CSE Home

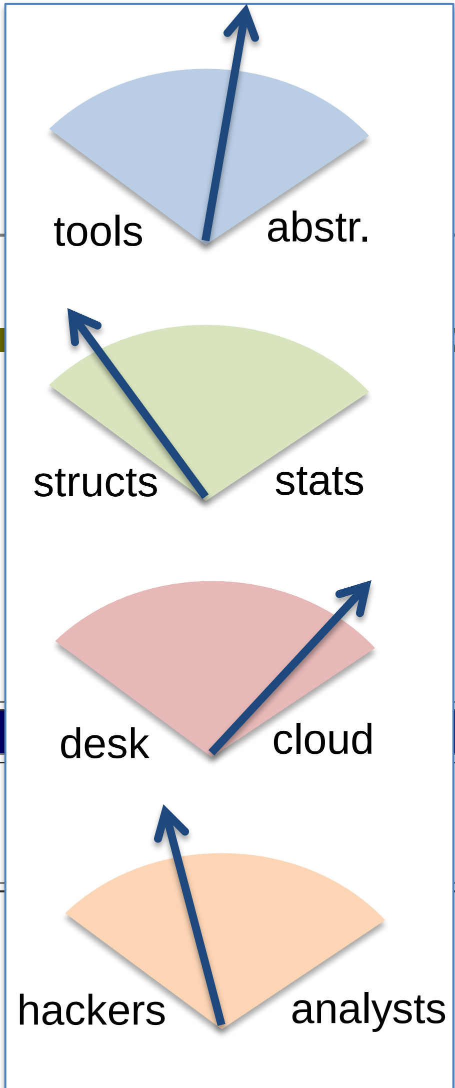
- Homepage
- Syllabus
- Calendar/Lecture notes
- Homeworks
- Webquiz
- Exams
- E-mail Archive
- Discussion Board
- Grades

CSE 344: Daily Schedule

Note that this schedule will be altered during the quarter. Please make sure to check it every week.

You can find the consolidated list of readings at the end of the page.

Week of	Monday	Wednesday		
Sep 24	Introduction lecture 1	Data Models lecture 2		
Oct 1	Aggregates in SQL lectures 4 and 5 (data) Homework 1 due Webquiz for lectures 1-3 due	Aggregates in SQL lectures 4 and 5 (data)		Upadhyaya



CS 194-16: Introduction to Data Science

Spring 2012

Course Information | Schedule | Resources | Spring 2011 Course

Jeff Hammerbacher

Mike Franklin

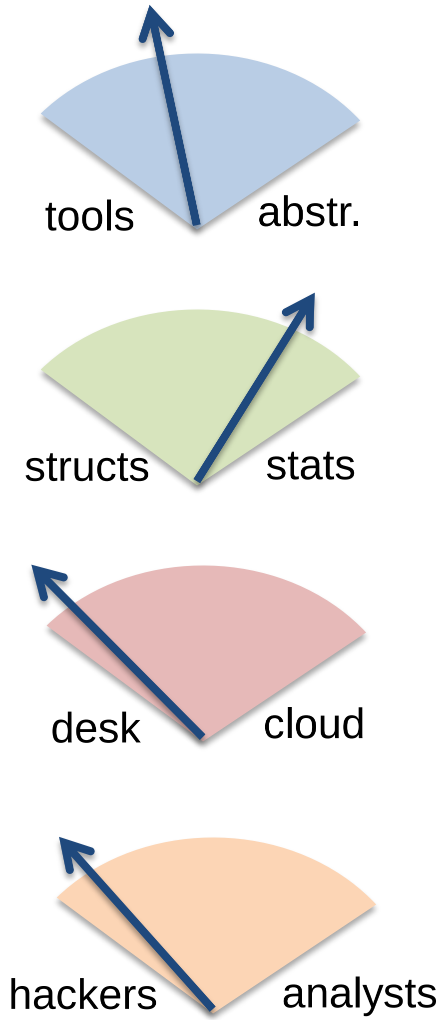
Schedule

The course will consist of five components: data preparation, data presentation, data products, observation, and experimentatio

There will be up to 6 guest speakers throughout the course; lecture dates may change based on guest speaker availability.

Regular Lectures

Component	Weeks	Class Dates	Lecture Materials	Homework
Introduction	1	1/17-1/19	Lecture 1 (slides, video) Lecture 2 (slides, video)	
Data Preparation	2 – 4	1/24-2/9	Lecture 3 (slides, video) Lecture 4 (slides, video) Lecture 5 (slides, video) Lecture 6 (slides) Lecture 7 (slides) Lecture 8 (slides)	Assignment 1
Data Presentation	5 – 6	2/14-2/23	Lecture 9 (slides) Lecture 10 (slides) Lecture 13 (slides)	Assignment 2
Data Products	7 – 10	2/28 – 3/22	Lecture 15 (IPython Notebook) Lecture 18 (slides) Lecture 19 (slides)	Assignment 3
Spring Break	11	3/27 – 3/29		
Observation	12 – 13	4/3 – 4/12	Lecture 20 (slides) Lecture 21 (slides)	
Experimentation	14	4/17-4/19		
Final Project	15	4/24 – 4/26		Final Project



COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Introduction to Data Science

Rachel Schutt

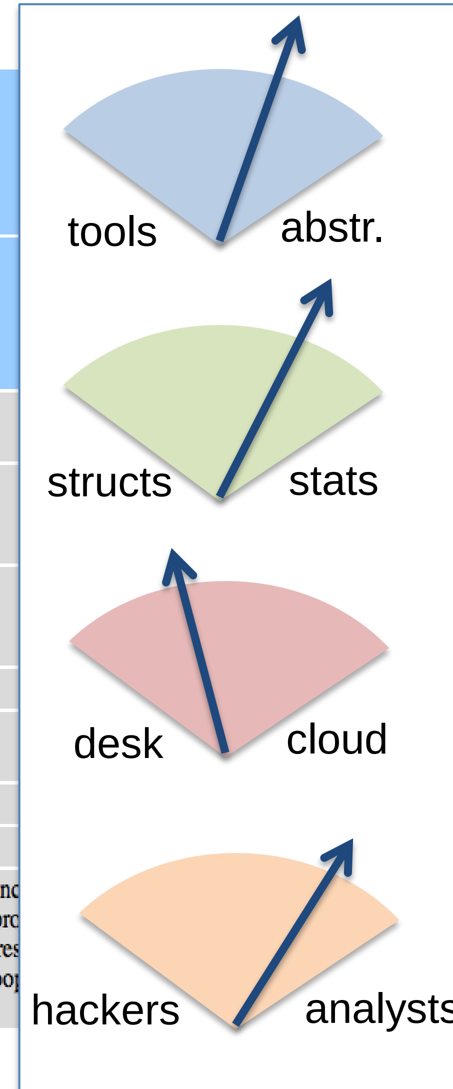
Columbia University Directory of Classes

NOTE: Course information changes frequently. Please re-visit these pages periodically for the most recent and up-to-date information.

Fall 2012 Statistics W4242 section 001

INTRODUCTION TO DATA SCIENCE

Call Number	61780
Day & Time Location	MW 6:10pm-7:25pm 313 Fayerweather
Day & Time Location	MW 7:40pm-8:55pm 313 Fayerweather
Points	3
Approvals Required	None
Instructor	Rachel R Schutt
Type	LECTURE
Course Description	This course is an introduction to the interdisciplinary and emerging field of data science, which lies at the intersection of statistics, computer scienc the social sciences. The course will be organized around three central threads: (1) statistical modeling and machine learning, (2) data pipelines, pro "big data" tools, and (3) real world topics and case studies. Correspondingly there will be (1) core lectures, 92) labs and (3) guest lectures from res who are experts in their fields. Topics and tools will include logistoc regression, predictive modeling, clustering algorithms, decision trees, Hadoop visualiziation, data journalism, R, python, javascript.



EMC²

DATA SCIENCE AND BIG DATA ANALYTICS

An ‘open’ course to unleash the power of Big Data

“We live in a data-driven world. Increasingly, the efficient operation of organizations across sectors relies on the effective use of vast amounts data. Making sense of big data is a combination of organizations havin the tools, skills and more importantly, the mindset to see data as the n “oil” fueling a company. Unfortunately, the technology has evolved fas than the workforce skills to make sense of it and organizations across sectors must adapt to this new reality or perish.”

- Andreas Weigend, Ph.D Stanford, Head of the Social Data Lab at Stanfo former Chief Scientist, Amazon.com

EMC² PROVEN PROFESSIONAL

ORDER YOUR VIDEO-ILT TODAY!

Become a certified Data Science Associate (EMCDSA)

- Immerse yourself in a near-classroom experience without the need for traveling
- Learn from top EMC subject-matter experts at your own pace

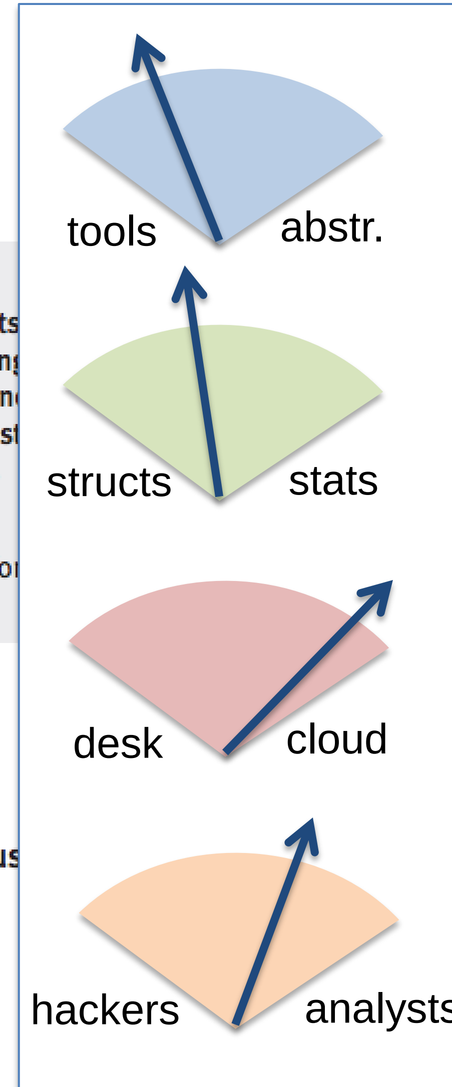
DATA SCIENCE AND BIG DATA ANALYTICS COURSE

An ‘open’ course and certification focused on concepts and principles applicable to any technology environment and indus

This course is intended for:

- Business and data analysts looking to add big data analytics skills
- Managers of business intelligence, analytics, or big data groups
- Database professionals looking to enrich their analytic skills
- College graduates considering data science as a career field

The course provides a hands-on practitioner’s approach to the techniques and tools required for analyzing Big Data.



100 YEARS W

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

| EV... | CONTACT US

Explore Programs | Online Learning | Student Resources

Bill Howe Richard Sharp Roger Barga

Approved by the UW Department of Computer Science & Engineering.

Certificates » Data Science » Winter 2013 Details » Introduction to Data Science

Winter 2013 Certificate

- Details
- Admissions
- Apply Now

Courses

- **Introduction to Data Science**
- Methods for Data Analysis
- Deriving Knowle...

Course Description

Introduction to Data Science

Bellevue, Classroom, Winter 2013

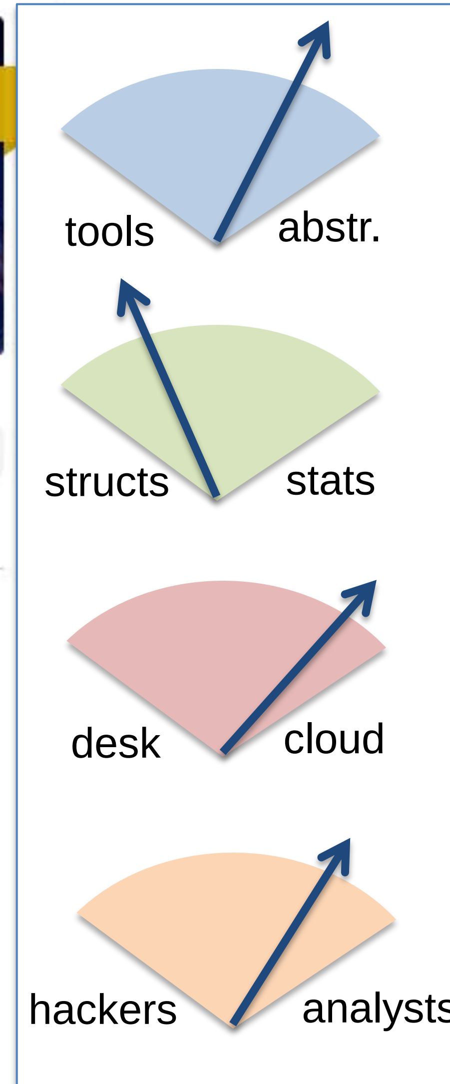
Instructor: Ernst Henle

Th, 1/10 - 3/14, 2013, 6-9 p.m.

Cost: $999 | 3 CEUs

This course is designed to introduce students to the data management, storage and manipulation tools common in data science and will apply those tools to real scenarios. An overview of different SQL and No-SQL database technologies is presented and the course finishes with a discussion of choosing the appropriate tool to get the job done.

Topics include:



UNIVERSITY of WASHINGTON

eScience Institute

Bill Howe

coursera

COURSES UNIVERSITIES ABOUT ▼

Course Dashboard

Users

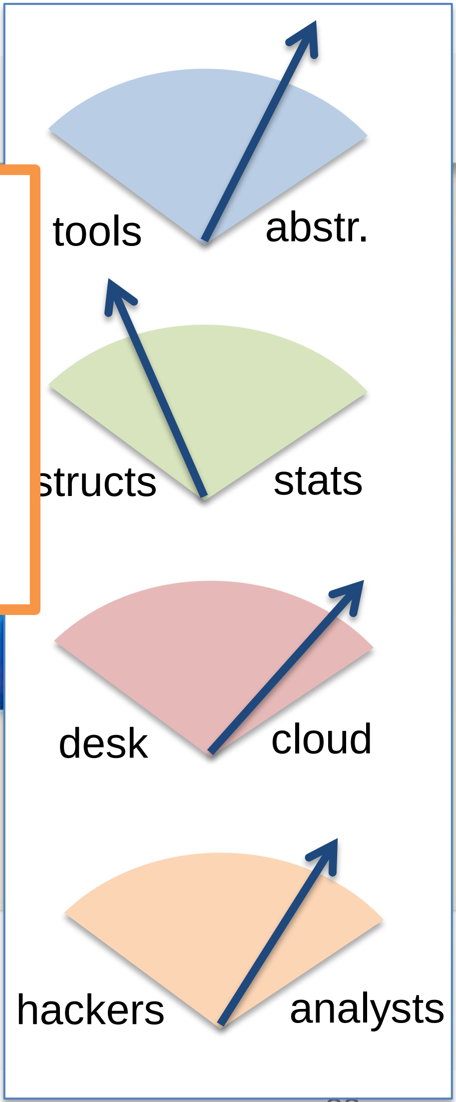
Total Registered Users	17834

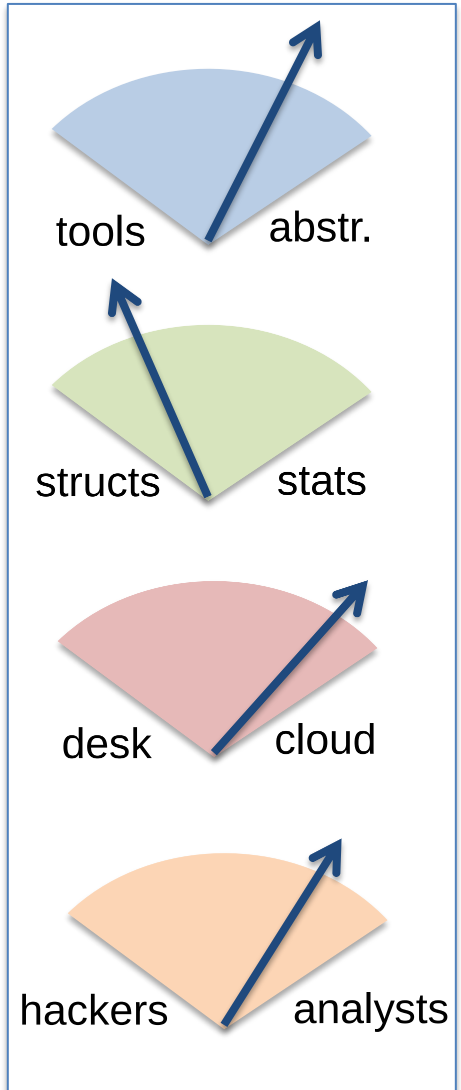
easy to obtain through conventional curricula. Introduce yourself to the basics of data science and leave armed with practical experience programming massive databases.

You are signed up

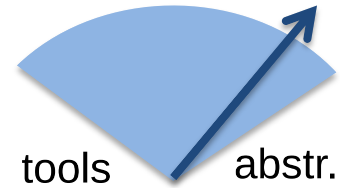
Next session: April 2013 (10 weeks long)

Statistics, Data Analysis, and Scientific Computing



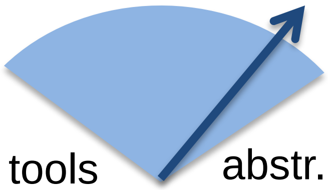

tools
abstr.
structs
stats
desk
cloud
hackers
analysts

What goes around comes around



- 2004 Dean et al. MapReduce
- 2008 Hadoop 0.17 release
- 2008 Olston et al. Pig: Relational Algebra on Hadoop
- 2008 DryadLINQ: Relational Algebra in a Hadoop-like system
- 2009 Thusoo et al. HIVE: SQL on Hadoop
- 2009 Hbase: Indexing for Hadoop
- 2010 Dietrich et al. Schemas and Indexing for Hadoop
- 2012 Transactions in HBase (plus VoltDB, other NewSQL systems)

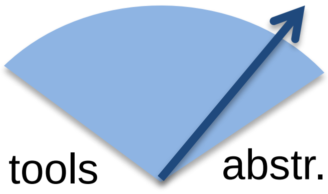
- But also some permanent contributions:
  - Fault tolerance
  - Schema-on-Read
  - User-defined functions that don't suck



What are the *abstractions* of data science?

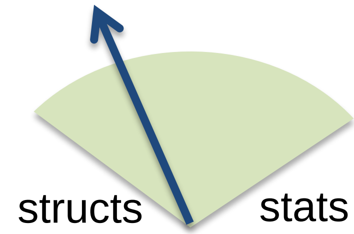
"Data Jujitsu"
"Data Wrangling"
"Data Munging"

Translation: "We have no idea what this is all about"



What are the *abstractions* of data science?

matrices and linear algebra?
relations and relational algebra?
objects and methods?
files and scripts?
data frames and functions?

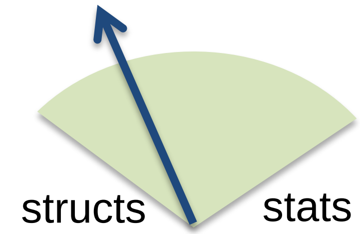


“80% of analytics is sums and averages”

-- Aaron Kimball, wibidata

God created the integers; all else is the work of man

Codd created relations; all else is the work of man



Three types of tasks:

1) Preparing to run a model

“80% of the work”

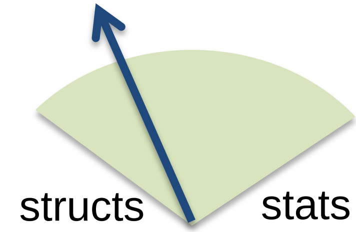
-- Aaron Kimball

Gathering, cleaning, integrating, restructuring, transforming, loading, filtering, deleting, combining, merging, verifying, extracting, shaping, massaging

2) Running the model

3) Interpreting the results

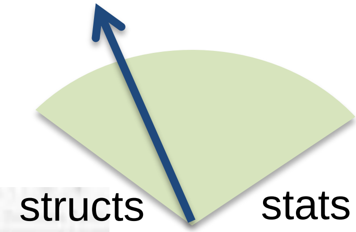
The other 80% of the work



Problem

How much time do you spend “handling data” as opposed to “doing science”?

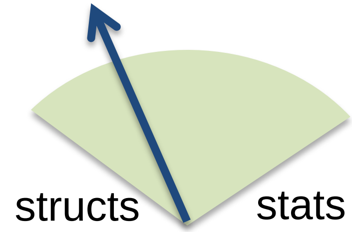
Mode answer: “90%”



- Databases and Statistical Packages
  - Many analysts download data to use in Excel/SAS/Matlab/R or their favorite programming language?
    FORTRAN??
  - Use matrix/vector operations
  - Most of these stat packages require data to fit in RAM
    - Taking samples from the full data to fit into ram results in loss of precision
  - External toolkits may also lack parallelism

src: Christian Grant, MADSkills

(Sparse) Matrix Multiply in SQL




```sql
SELECT A. row_number,  B.column_number, SUM(A.value * B.value)
FROM A, B
WHERE A.column_number = B.row_number
GROUP BY A.row_number,  B.column_number
```

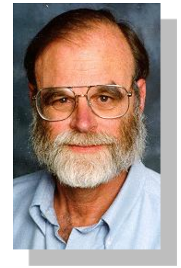
src: Christian Grant, MADSkills

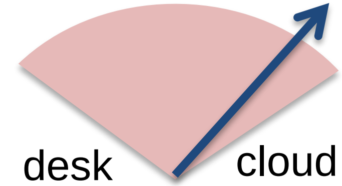
Aside: Schema-on-Write vs. Schema-on-Read

- A schema* is a shared consensus about some universe of discourse
- At the frontier of research, this shared consensus does not exist, *by definition*
- Any schema that does emerge will change frequently, *by definition*
- Data found "in the wild" will typically not conform to any schema, *by definition*
- But this doesn't mean we have to live with ad hoc scripts and files
- My answer: Schema-later, "lazy schemification"

* ontology/metadata standard/controlled vocabulary/etc.

Data Access Hitting a Wall





Current practice based on data download (FTP/GREP)
Will not scale to the datasets of tomorrow

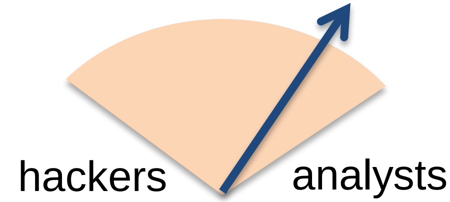
- You can GREP 1 MB in a second
- You can GREP 1 GB in a minute
- You can GREP 1 TB in 2 days
- You can GREP 1 PB in 3 years.

- Oh!, and 1PB ~5,000 disks

- At some point you need
  indices to limit search
  parallel data search and analysis
- This is where databases can help

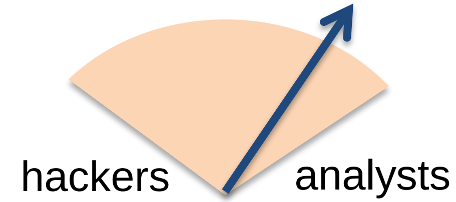
- You can FTP 1 MB in 1 sec
- You can FTP 1 GB / min (~1$)
- … 2 days and 1K$
- … 3 years and 1M$

[slide src: Jim Gray]



US faces 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



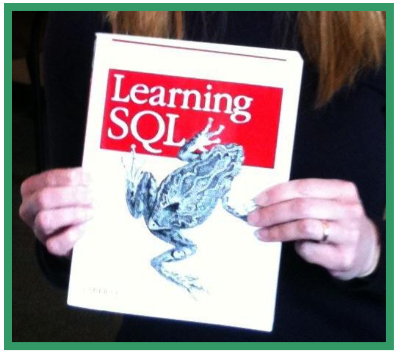
Biologists are beginning to write very complex queries (rather than relying on staff programmers)

Example: Computing the overlaps of two sets of blast results

```
SELECT x.strain, x.chr, x.region as snp_region, x.start_bp as snp_start_bp
    , x.end_bp as snp_end_bp, w.start_bp as nc_start_bp, w.end_bp as nc_end_bp
    , w.category as nc_category
    , CASE WHEN (x.start_bp >= w.start_bp AND x.end_bp <= w.end_bp)
   THEN x.end_bp - x.start_bp + 1
   WHEN (x.start_bp <= w.start_bp AND w.start_bp <= x.end_bp)
   THEN x.end_bp - w.start_bp + 1
   WHEN (x.start_bp <= w.end_bp AND w.end_bp <= x.end_bp)
   THEN w.end_bp - x.start_bp + 1
 END  AS len_overlap

 FROM [koesterj@washington.edu].[hotspots_deserts.tab] x
 INNER JOIN [koesterj@washington.edu].[table_noncoding_positions.tab] w
 ON x.chr = w.chr
 WHERE (x.start_bp >= w.start_bp AND x.end_bp <= w.end_bp)
 OR (x.start_bp <= w.start_bp AND w.start_bp <= x.end_bp)
 OR (x.start_bp <= w.end_bp AND w.end_bp <= x.end_bp)
 ORDER BY x.strain, x.chr ASC, x.start_bp ASC
```

We see thousands of queries written by non-programmers



UW Curricular Activities

	Students				Non-Students	
	CS/Informatics		Non-Major		professionals	researchers
	undergrads	grads	undergrads	grads		
UWEO Data Science Certificate						
Graduate Certificate in Big Data						
Database Courses						
eScience workshops						
Intro to data programming						
eScience Masters (planned)						
Coursera Course Intro to Data Science						

How do you deliver hands-on big data experience to 10k students?

Cloud vendors' free tiers?
- 1 micro instance is not "big data"

Cloud vendors' academic discounts? (e.g., Amazon's education grants)
- $100 / head = $1M
- invites abuse: free credits with no obligation to complete course

Out of pocket?
- Don't want to be the only non-free Coursera course
- Unclear that we can require it (perhaps analogous to a textbook?)

10k Students on 10k GB for $10k

- Requirements
  - Inexpensive: Need a fixed, small budget; O(10k) maximum
  - Fair: All students need to be able to complete the assignment
- Non-solutions
  - Fixed cluster
    - Fairness problems; no quality of service guarantees
  - Autoscaling cluster
    - No upper bound to cost
  - Budget cap (via, e.g., Amazon's IAM)
    - Fairness problems: Different students consume different levels of resources depending on background, etc.

10k Students on 10k GB for $10k

- Key idea: 10k students all working on the same assignment = lots of redundancy
- Students debug locally on scaled down datasets
- Then submit 10k jobs
- Prune the queue aggressively
  - Remove duplicates
  - Detect typical mistakes syntactically; return cached results
  - Global common subexpression elimination (feasible thanks to abstractions)

10k Students on 10k GB for $10k

- Another approach we are considering: Kaggle-Kaggle-style Prize assignments

student	date	runtime	output	quality	notes	votes
sarah123	4/23/12	5 min 44 sec	result.txt	45%	I removed the dirty data for this run	456
jane456	4/22/12	3 min 23 sec	result.txt	23%	I used gradient descent this time	97

:

:

- Pay-to-play: Students submit successful jobs to access leaderboards
- Cast votes for their preferred solution.
- Grade determined by
  f(votes(your_solution), grade(your_solution), grade(solutions_you_voted_for))
- We run the top-k highest ranked solutions on the full size dataset

- Problem: Easy to game the system and just coast *http://escience.washington.edu*

billhowe@cs.washington.edu

Coursera course:

https://www.coursera.org/course/datasci

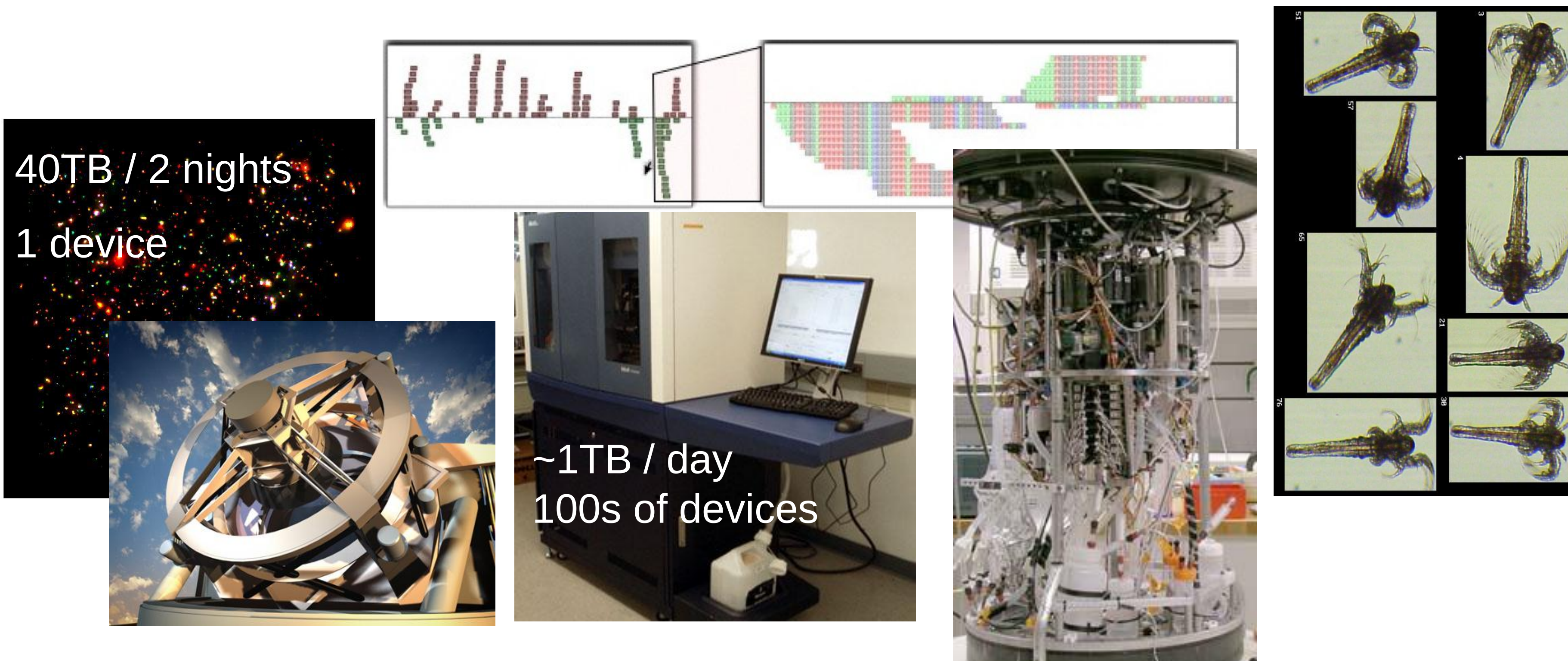
Certificate program:

http://www.pce.uw.edu/courses/data-science-intro

Science is becoming a database query problem

Old model: "Query the world" (Data acquisition coupled to a specific hypothesis)
New model: "Download the world" (Data acquisition supports many hypotheses)

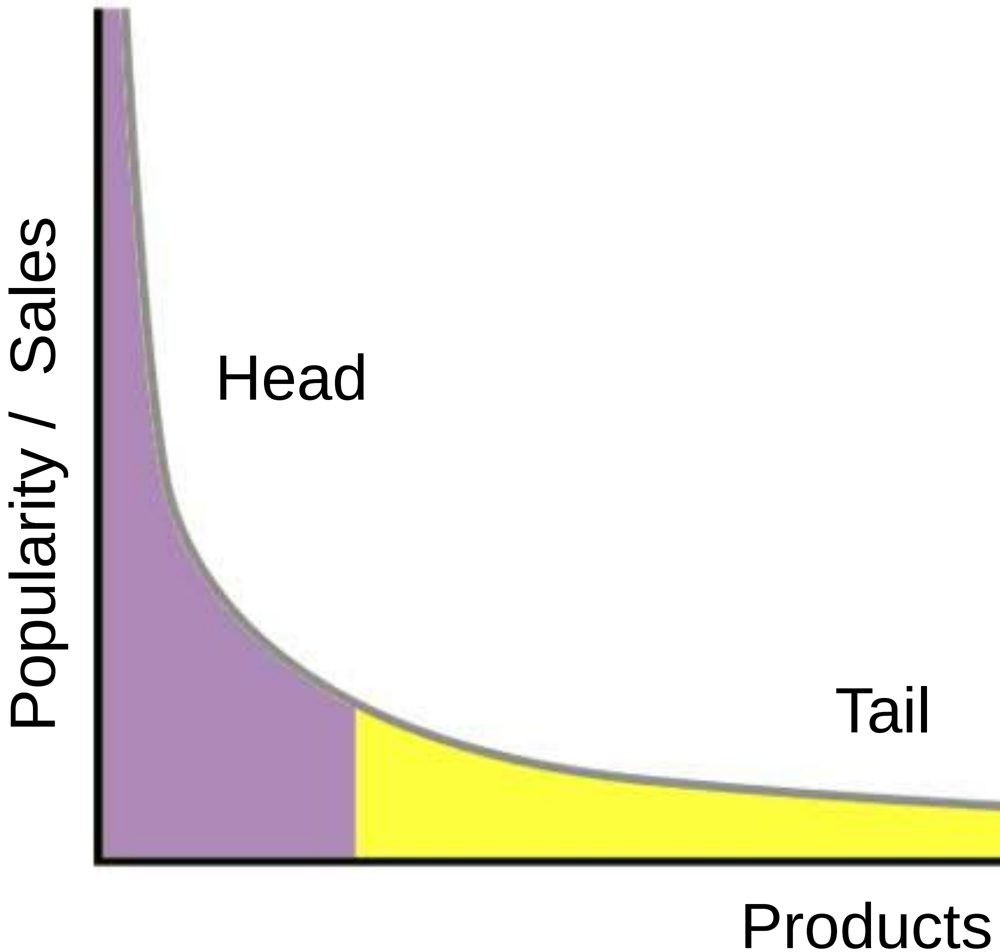
- **Astronomy: High-resolution, high-frequency sky surveys (SDSS, LSST, PanSTARRS)**
- **Biology: lab automation, high-throughput sequencing,**
- **Oceanography: high-resolution models, cheap sensors, satellites**

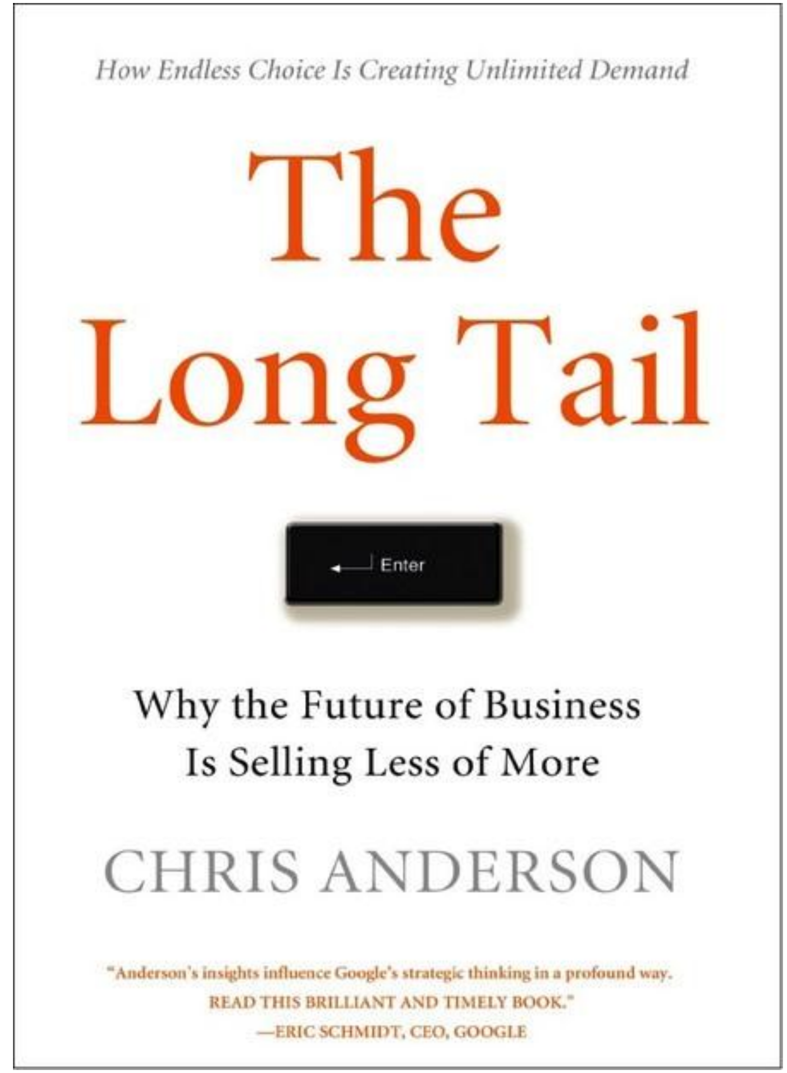


UNIVERSITY of WASHINGTON

[src: Carol Goble]

- Power distribution
- 80:20 rule





First published May 2007, Wired Magazine article 2004

A “Needs Hierarchy” of Science Data Management

“As each need is satisfied, the next higher level in the hierarchy dominates conscious functioning.”

-- Maslow 43