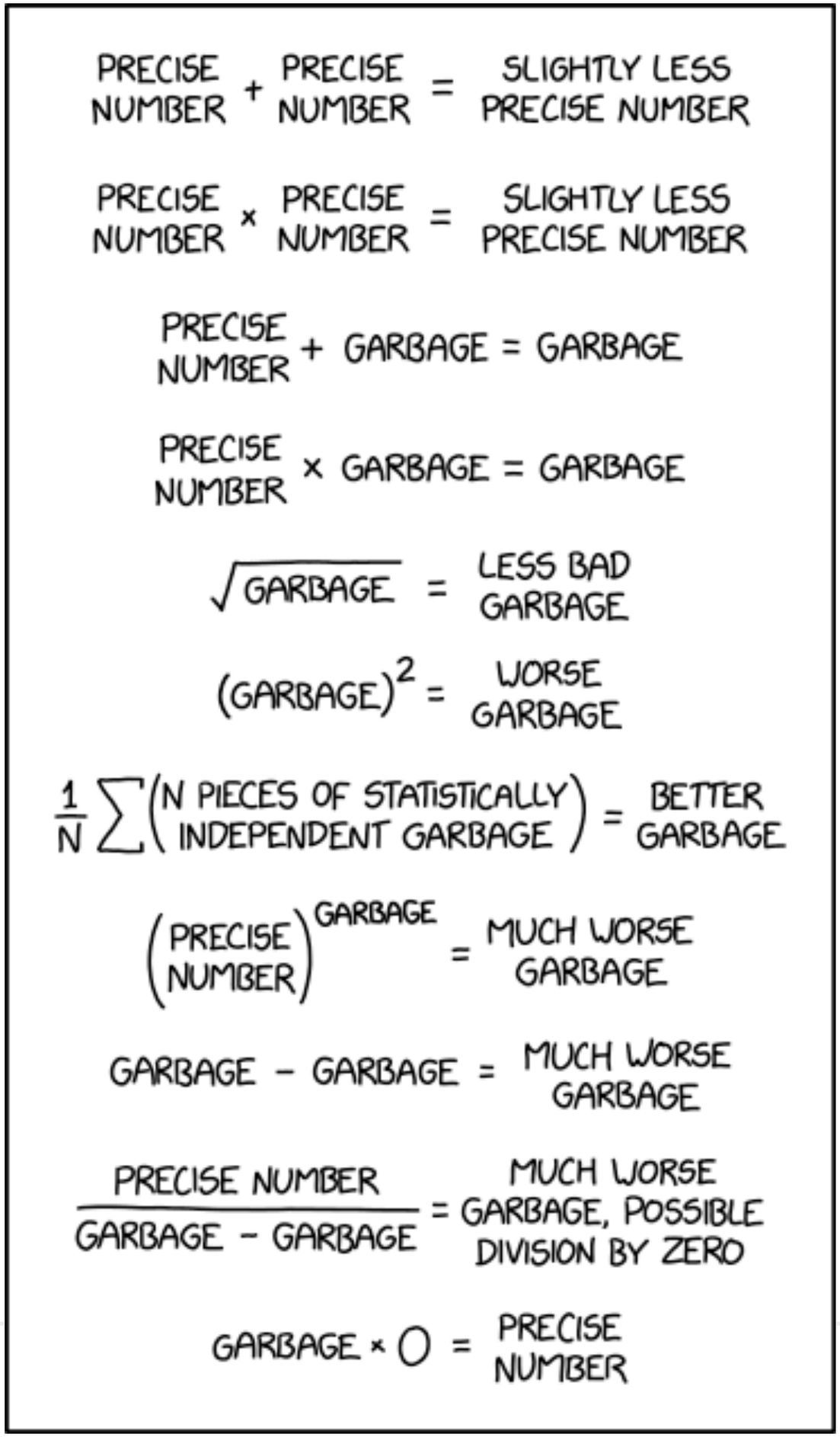
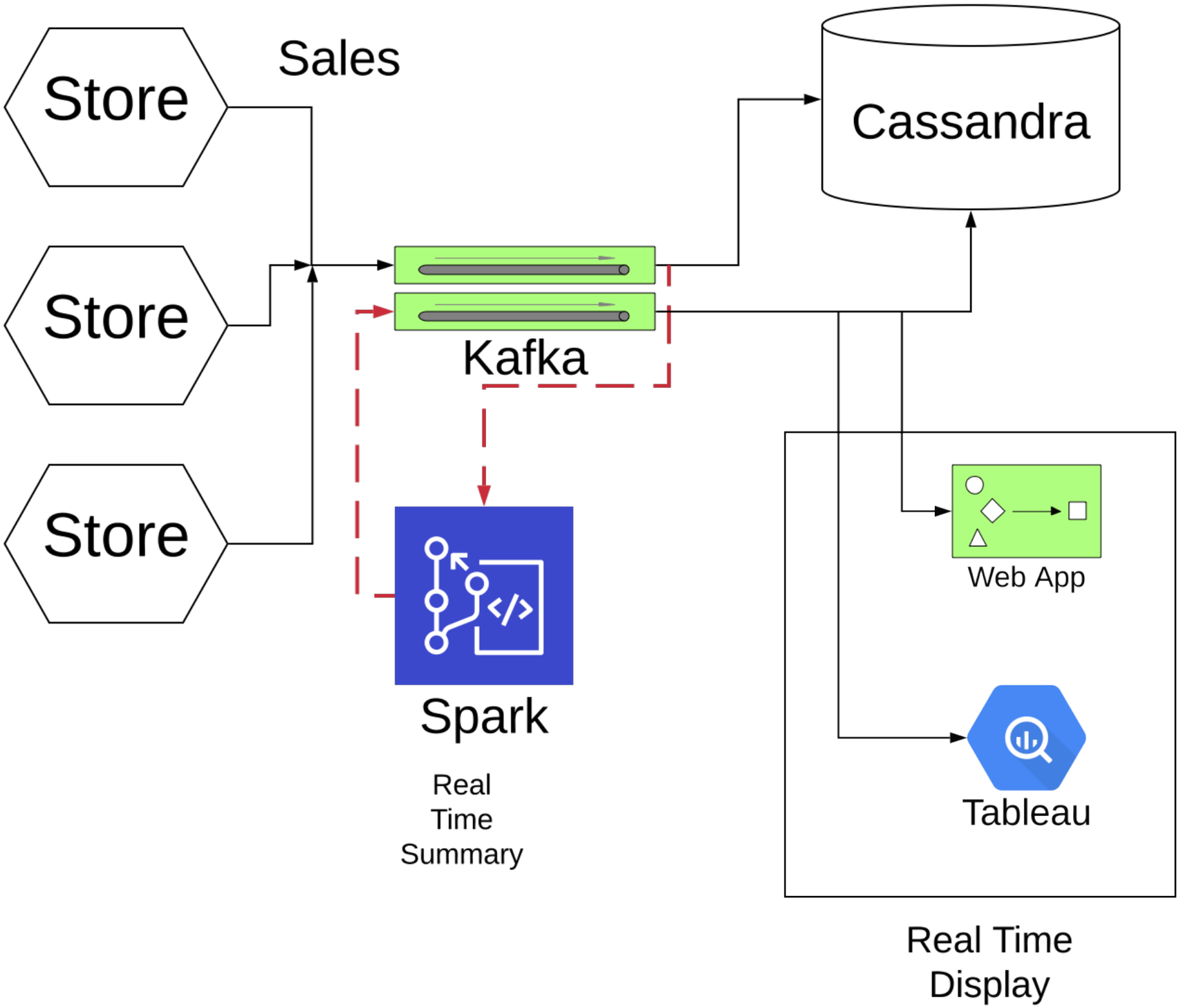
CS 696 Intro to Big Data: Tools and Methods
Fall Semester, 2020
Doc 24 Display End Remarks
Apr 23, 2020

https://xkcd.com/2295/



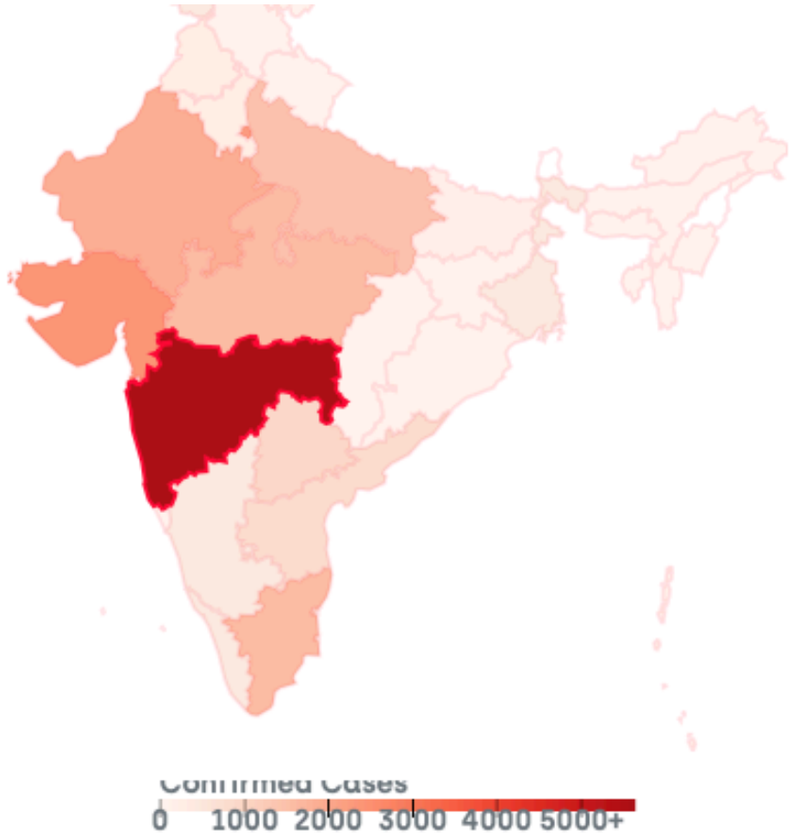
Displaying Results



https://www.covid19india.org

Compiled from State Govt. numbers, know more!

State/UT	Confirmed	Active	Recovered	Deceased
Maharashtra	↑431 5,649	4,591	↑67 789	↑18 269
Gujarat	↑229 2,407	2,125	↑40 179	↑13 103
Delhi	↑92 2,248	1,476	↑113 724	↑1 48
Rajasthan	↑153 1,888	1,517	↑70 344	↑1 27
Tamil Nadu	↑33 1,629	949	↑27 662	18
Madhya Pradesh	↑35 1,587	1,355	↑4 152	80
Uttar Pradesh	↑112 1,449	1,255	↑11 173	21
Telangana	↑15 943	725	194	↑1 24
Andhra Pradesh	↑56 813	669	↑24 120	↑2 24
Kerala	↑11 437	127	↑1 308	2
Karnataka	↑9 427	279	↑2 131	17
West Bengal	↑31 423	335	73	15
Jammu and Kashmir	↑27 407	310	↑11 92	5
Punjab	↑27 278	209	↑4 53	16
Haryana	↑9 264	103	↑11 158	3
Bihar	↑15 141	97	42	2
Odisha	↑4 83	50	↑2 32	1
Jharkhand	46	40	4	2
Uttarakhand	46	23	↑4 23	-
Himachal Pradesh	39	21	16	2
Chhattisgarh	36	8	↑3 28	-
Assam	35	15	19	1
Chandigarh	27	13	14	-
Andaman				

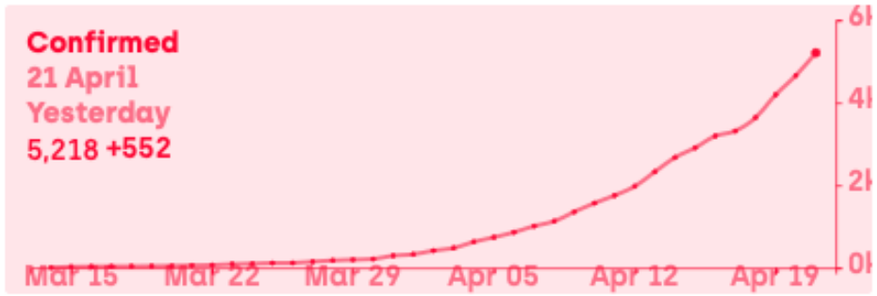


Spread Trends

Cumulative | Daily

Scale Modes Uniform Logarithmic

Maharashtra

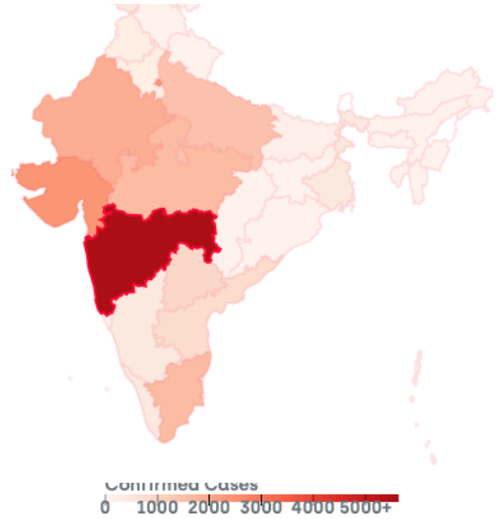


https://www.covid19india.org

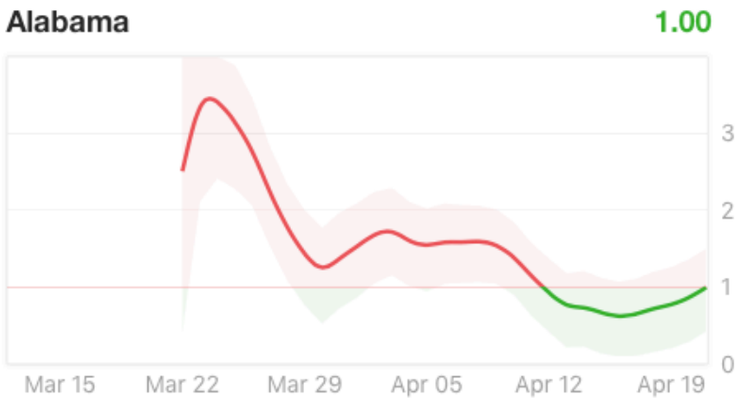
Use react components

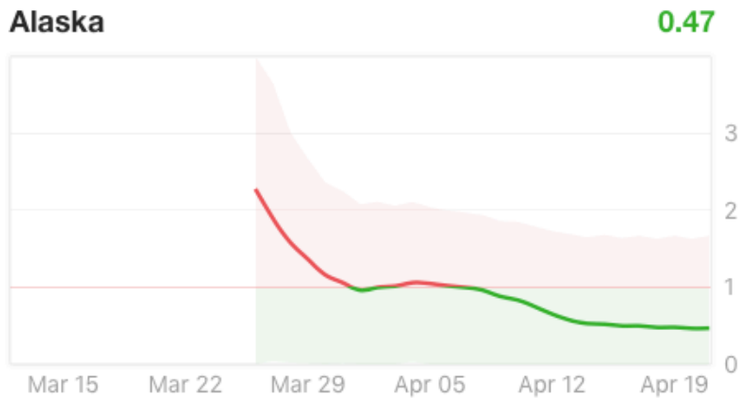
Compiled from State Govt. numbers, know more!

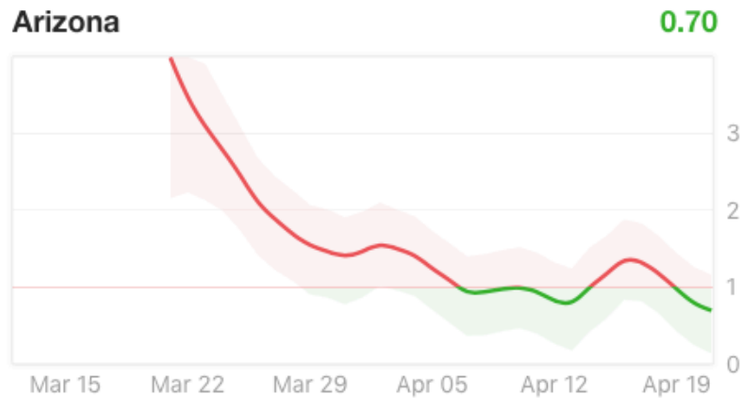
State/UT	Confirmed	Active	Recovered	Deceased
Maharashtra	↑431 5,649	4,591	↑67 789	↑18 269
Gujarat	↑229 2,407	2,125	↑40 179	↑13 103
Delhi	↑92 2,248	1,476	↑113 724	↑1 48
Rajasthan	↑153 1,888	1,517	↑70 344	↑1 27
Tamil Nadu	↑33 1,629	949	↑27 662	18
Madhya Pradesh	↑35 1,587	1,355	↑4 152	80
Uttar Pradesh	↑112 1,449	1,255	↑11 173	21
Telangana	↑15 943	725	194	↑1 24
Andhra Pradesh	↑56 813	669	↑24 120	↑2 24
Kerala	↑11 437	127	↑1 308	2
Karnataka	↑9 427	279	↑2 131	17
West Bengal	↑31 423	335	73	15
Jammu and Kashmir	↑27 407	310	↑11 92	5
Punjab	↑27 278	209	↑4 53	16
Haryana	↑9 264	103	↑11 158	3
Bihar	↑15 141	97	42	2
Odisha	↑4 83	50	↑2 32	1
Jharkhand	46	40	4	2
Uttarakhand	46	23	↑4 23	-
Himachal Pradesh	39	21	16	2
Chhattisgarh	36	8	↑3 28	-
Assam	35	15	19	1
Chandigarh	27	13	14	-
Andaman and Nicobar Islands	↑1 18	7	11	-
Ladakh	18	4	14	-
Meghalaya	12	11	-	1
Goa	7	-	7	-
Puducherry	7	3	4	-



https://rt.live

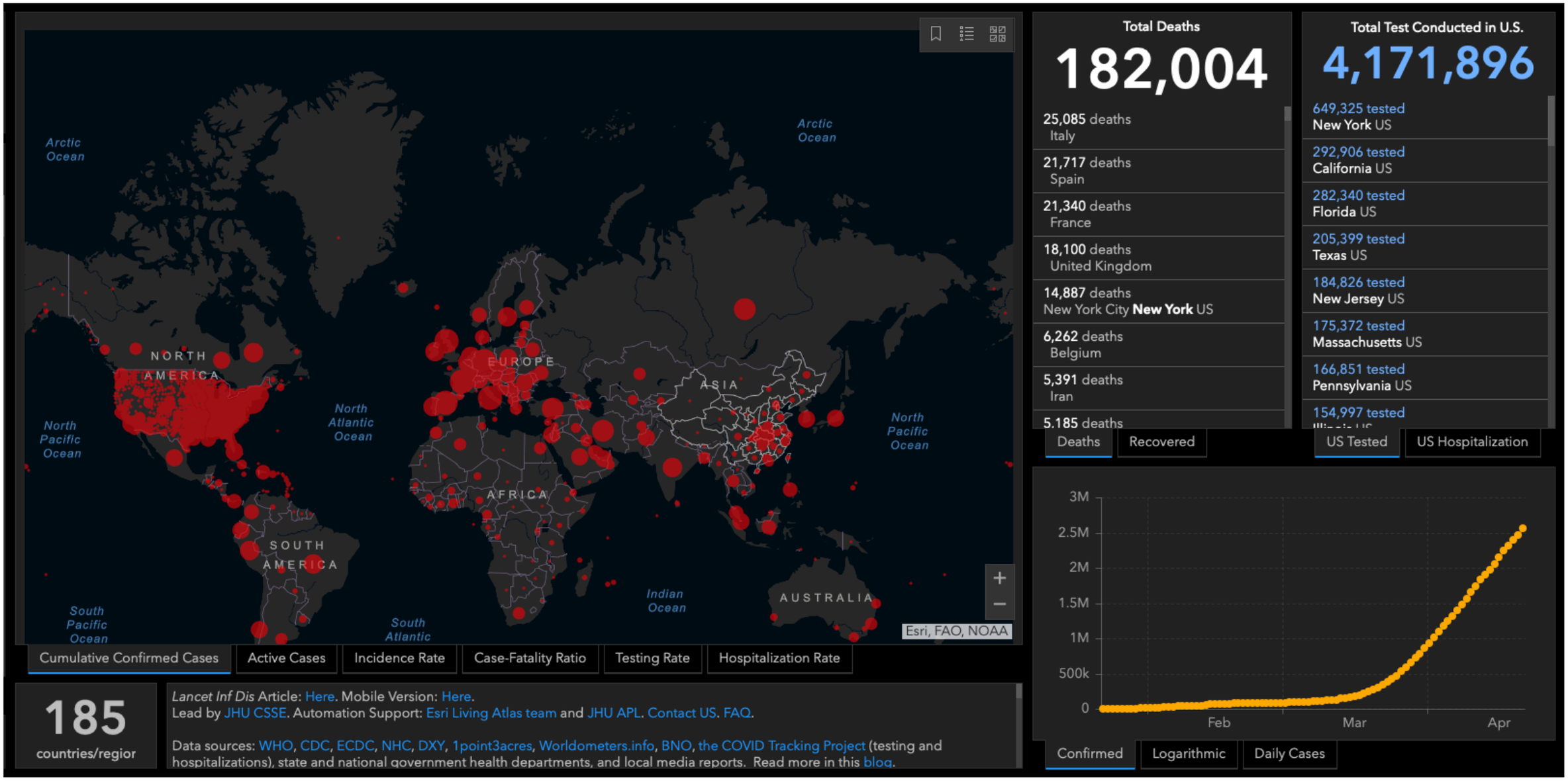






svg files + javascript

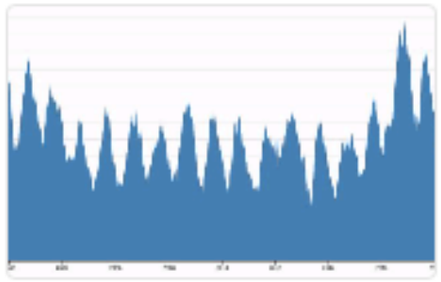
https://www.covidtracker.com



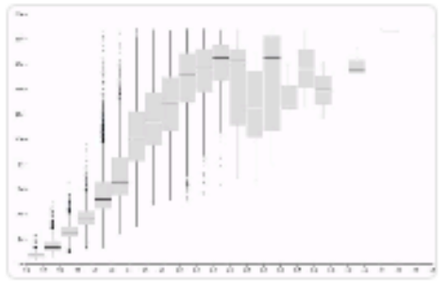
ArcGIS

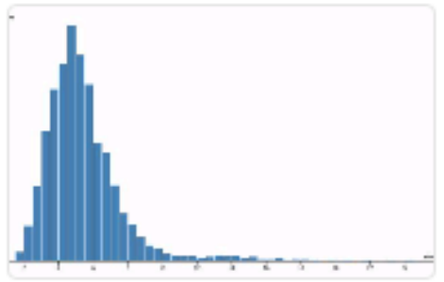
D3 - Data-Driven Documents

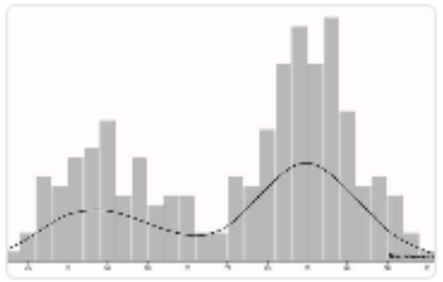
https://d3js.org

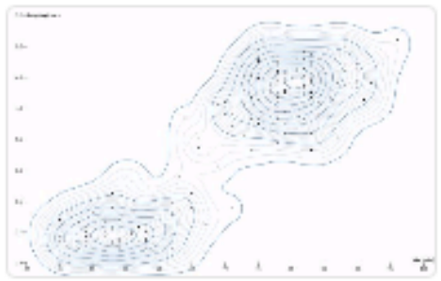

Moving average

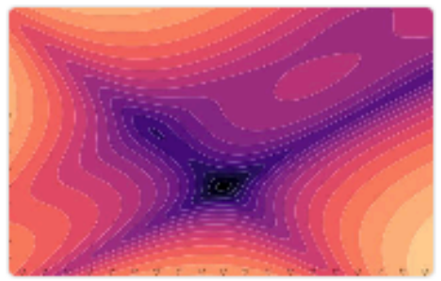

Bollinger bands

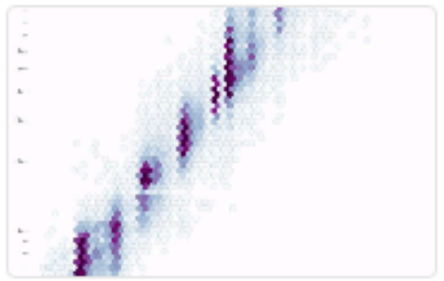

Box plot

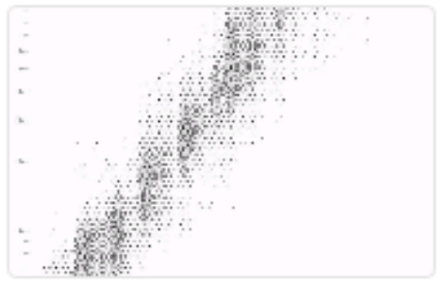

Histogram

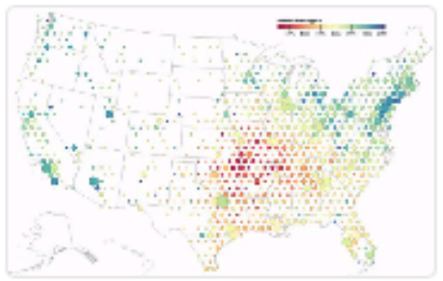

Kernel density estimation

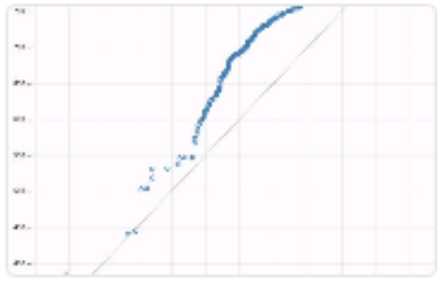

Density contours

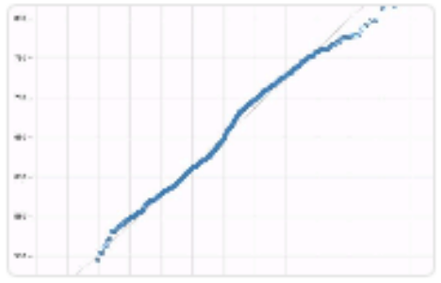

Contours


Hexbin


Hexbin (area)


Hexbin map


Q–Q plot


Normal quantile plot


Parallel sets

JavaScript library for manipulating documents based on data

Tableau

https://www.tableau.com

Tableau Software

American interactive data visualization software company

Based on visualization research from Stanford

Visualization techniques for exploring and analyzing

Relational databases

N-dimensional data

No programming

Tableau - Book1

trumptweets

Connection: Live / Extract

Filters: 0 | Add

Connections Add

trumptweets
Text file

Files

Use Data Interpreter

Data Interpreter might be able to clean your Text file workbook.

- snowtest.csv
- testtweets.csv
- trumpttest.csv
- trumptweets.csv
- US_WeatherE...16-2019.csv
- weathertest.csv

New Union

trumptweets.csv

Sort fields: Data source order

Show aliases | Show hidden fields | 1,000 rows

Id	Link	Content	Date	Retweets	Favorites	Mentions	Hashtags	Geo
1698308935	https://twitter.com/r...	Be sure to tune in an...	5/4/2009 8:54:25 PM	500	868	null	null	null
1701461182	https://twitter.com/r...	Donald Trump will be...	5/5/2009 3:00:10 AM	33	273	null	null	null
1737479987	https://twitter.com/r...	Donald Trump reads ...	5/8/2009 3:38:08 PM	12	18	null	null	null
1741160716	https://twitter.com/r...	New Blog Post: Celeb...	5/8/2009 10:40:15 PM	11	24	null	null	null
1773561338	https://twitter.com/r...	"My persona will nev...	5/12/2009 4:07:28 PM	1,399	1,965	null	null	null
1776419923	https://twitter.com/r...	Miss USA Tara Conne...	5/12/2009 9:21:55 PM	27	26	null	null	null
1786560616	https://twitter.com/r...	Listen to an intervie...	5/13/2009 7:38:28 PM	14	16	null	null	null
1796477499	https://twitter.com/r...	"Strive for wholenes...	5/14/2009 6:30:40 PM	18	25	null	null	null
1806258917	https://twitter.com/r...	Enter the "Think Like...	5/15/2009 4:13:13 PM	14	8	null	null	null
1820624395	https://twitter.com/r...	"When the achiever ...	5/17/2009 12:22:45 ...	20	48	null	null	null
1826225450	https://twitter.com/r...	"Don't be afraid of b...	5/17/2009 5:00:03 PM	38	67	null	null	null
1836131903	https://twitter.com/r...	"We win in our lives ...	5/18/2009 4:26:00 PM	64	102	null	null	null
1849558306	https://twitter.com/r...	"...these days...we c...	5/19/2009 7:43:39 PM	18	17	null	null	null
1859044981	https://twitter.com/r...	"Always know you co...	5/20/2009 3:25:39 PM	29	53	null	null	null
1864367186	https://twitter.com/r...	Read a great intervie...	5/21/2009 12:29:47 ...	8	10	null	null	null
1878373267	https://twitter.com/r...	"Keep it fast, short a...	5/22/2009 4:59:39 AM	52	70	null	null	null

Data Source | Sheet 1

Tableau - Book2

Standard

Show Me

Data | Analytics

US_WeatherEvents_201...

Dimensions

- Airport Code
- City
- County
- EndTime(UTC)
- Event Id
- Severity
- StartTime(UTC)
- State
- Time Zone
- Type
- Zip Code
- *Measure Names*

Measures

- Location Lat
- Location Lng
- *Latitude (generated)*
- *Longitude (generated)*
- *Number of Records*
- *Measure Values*

Pages

Filters

Marks

Automatic

Color | Size | Text | Detail | Tooltip

Columns

Rows

Sheet 1

Drop field here

Drop field here

Drop field here

Data Source | Sheet 1

http://highscalability.com

how to scale software - primarily web sites & backends

Hacker News

https://news.ycombinator.com

Martin Fowler Bliki

A website on building software effectively

https://martinfowler.com

Author

Works at ThoughtWorks

Software Architecture Guide

https://martinfowler.com/architecture/

What is architecture?

Why does architecture matter?

Application Architecture

Application Boundary

Microservices Guide

Serverless Architectures

Micro Frontends

GUI Architectures

Presentation Domain Data Layering

Martin Fowler - Recent Posts

Exploratory Testing

Waterfall Process

Continuous Delivery for Machine Learning

Don't get locked up into avoiding lock-in

Micro Frontends

ThoughtWorks Technology Radar

Techniques

Tools

Platforms

Languages & Frameworks

Adopt

Trial

- Worth pursing
- Try on projects that can handle risk

Assess

- Worth exploring
- How will it affect your enterprise

Hold

- Proceed with caution

TECHNIQUES - Adapt

1. Container security scanning
2. Data integrity at the origin
3. Micro frontends
4. Pipelines for infrastructure as code
5. Run cost as architecture fitness function
6. Testing using real devices

TECHNIQUES - Trial

7. Automated machine learning (AutoML)
8. Binary attestation
9. Continuous delivery for machine learning (CD4ML)
10. Data discoverability
11. Dependency drift fitness function
12. Design systems
13. Experiment tracking tools for machine learning
14. Explainability as a first-class model selection criterion
15. Security policy as code
16. Sidecars for endpoint security
17. Zhong Tai

Zhong Tai

An approach to delivering encapsulated business models

Deliver first- rate services without the costs of traditional enterprise infrastructure and enabling existing organizations to bring innovative services to market at breakneck speeds

Developed at Alibaba

Conway's Law

Organizations which design systems ...

are constrained to produce designs which are copies of the communication structures of these organizations

"If you have four groups working on a compiler, you'll get a 4-pass compiler."

Eric S. Raymond

"If the parts of an organization do not closely reflect the essential parts of the product then the project will be in trouble ...

Therefore: Make sure the organization is compatible with the product architecture."

James O. Coplien and Neil B. Harrison

TECHNIQUES - Assess

18. BERT
19. Data mesh
20. Ethical bias testing
21. Federated learning
22. JAMstack
23. Privacy-preserving record linkage (PPRL) using Bloom filter
24. Semi-supervised learning loops

Data Mesh

Architectural paradigm that unlocks analytical data at scale

Data mesh shifts to a paradigm that draws from modern distributed architecture

LANGUAGES & FRAMEWORKS

Trial

78. Arrow
79. Flutter
80. jest-when
81. Micronaut
82. React Hooks
83. React Testing Library
84. Styled components
85. Tensorflow

Assess

86. Fairseq
87. Flair
88. Gatsby.js
89. GraphQL
90. KotlinTest
91. NestJS
92. Paged.js
93. Quarkus
94. SwiftUI
95. Testcontainers

Platforms

TRIAL
Apache Flink
Apollo Auto
GCP Pub/Sub
Mongoose OS
ROS

ASSESS
AWS Cloud Development Kit
Azure DevOps
Azure Pipelines
Crowdin
Crux
Delta Lake
Fission
FoundationDB
GraalVM
Hydra
Kuma
MicroK8s
Oculus Quest
ONNX
Rootless containers 49. Snowflake
Teleport

Delta Lake

Open-source storage layer by Databricks

- Attempts to bring transactions to big data processing

What every computer science major should know

Dr. Matt Might
University of Utah

http://matt.might.net/articles/what-cs-majors-should-know/

What should every student know to get a good job?

What should every student know to maintain lifelong employment?

What should every student know to enter graduate school?

What should every student know to benefit society?

Portfolio verse Resume

A resume says nothing of a programmer's ability

Portfolio
- Personal blog
- Projects
- Github
- Open source projects

Technical Communication

Lone wolves in computer science are an endangered species

In smaller companies, whether or not a programmer can communicate her ideas to management may make the difference between the company's success and failure

Writing for Computer Science by Zobel.

Even a Geek Can Speak by Asher.

Unix Philosophy

linguistic abstraction and composition

Should be able to

Navigate and manipulate the filesystem;
Compose processes with pipes;
Comfortably edit a file with emacs and vim;
Create, modify and execute a Makefile for a software project;
Write simple shell scripts.

Unix Philosophy

Sample tasks

Find the five folders in a given directory consuming the most space

Report duplicate MP3s (by file contents, not file name) on a computer.

Take a list of names whose first and last names have been lower-cased, and properly recapitalize them.

Find all words in English that have x as their second letter, and n as their second-to-last.

Directly route your microphone input over the network to another computer's speaker.

Replace all spaces in a filename with underscore for a given directory.

Report the last ten errant accesses to the web server coming from a specific IP address.

Systems administration

Every modern computer scientist should be able to:

Install and administer a Linux distribution.

Configure and compile the Linux kernel.

Troubleshoot a connection with dig, ping and traceroute.

Compile and configure a web server like apache.

Compile and configure a DNS daemon like bind.

Maintain a web site with a text editor.

Cut and crimp a network cable.

Programming languages

Programming languages rise and fall with the solar cycle.

A programmer's career should not.

The best way to learn how to learn programming languages is to learn multiple programming languages and programming paradigms.

To truly understand programming languages, one must implement one.

Programming languages

Racket

C

JavaScript

Squeak

Java

Standard ML

Prolog

Scala

Haskell

C++

Assembly

Racket

Aggressively simple syntax

For a small fraction of students, this syntax is an impediment.

To be blunt, if these students have a fundamental mental barrier to accepting an alien syntactic regime even temporarily, they lack the mental dexterity to survive a career in computer science.

Racket's powerful macro system and facilities for higher-order programming thoroughly erase the line between data and code.

If taught correctly, Lisp liberates

How to Design Programs

https://htdp.org

Squeak

Squeak is a modern dialect of Smalltalk, purest of object-oriented languages

It imparts the essence of "object-oriented."

Introductions to Squeak

http://wiki.squeak.org/squeak/377

Architecture

There is no substitute for a solid understanding of computer architecture

- transistors
- gates
- adders
- muxes
- flip flops
- ALUs
- control units
- caches
- RAM
- GPU

Operating systems

Any sufficiently large program eventually becomes an operating system

To get a better understanding of the kernel, students could:

Print "hello world" during the boot process;

Design their own scheduler;

Modify the page-handling policy; and

Create their own filesystem.

Networking

Computer scientists should have a firm understanding of the network stack and routing protocols within a network

Every computer scientist should implement the following:

- an HTTP client and daemon;
- a DNS resolver and server; and
- a command-line SMTP mailer.

No student should ever pass an intro networking class without sniffing their instructor's Google query off wireshark.

Security

Computer scientists must be aware of the means by which a program can be compromised

At a minimum, every computer scientist needs to understand:

- social engineering
- buffer overflows
- integer overflow
- code injection vulnerabilities
- race conditions
- privilege confusion

Metasploit: The Penetration Tester's Guide

Security Engineering: A Guide to Building Dependable Distributed Systems

Software testing

Software testing must be distributed throughout the entire curriculum

He uses test cases turned in by students against all other students

Students don't seem to care much about developing defensive test cases, but they unleash hell when it comes to sandbagging their classmates

Visualization

The modern world is a sea of data

The Visual Display of Quantitative Information by Tufte

Graphics and simulation

There is no discipline more dominated by "clever" than graphics.

The field is driven toward, even defined by, the "good enough."

As such, there is no better way to teach clever programming or a solid appreciation of optimizing effort than graphics and simulation.

Over half of the coding hacks I've learned came from my study of graphics.

Topics I left out

Databases
Artificial intelligence
Machine learning
Robotics
Software engineering
Parallelism
User experience design

Disarmingly Forthright MSCS Advice

Nick Black

http://nick-black.com/dankwiki/images/8/85/Msadvice.pdf

Read it

If you'll only take away two things

Read the damn man pages

Check your damn return values

You’re a CS MS student. Act it

Join the ACM and IEEE

Don’t embarrass yourself

- Passwords
- Backups

If you don’t have at least 100 semi-frequent, provocative/informative RSS feeds you’re checking a few times daily, you’re not learning enough

Programming

Vast majority of code you'll read is laughably broken

> if you aren't, at any given time, scandalized by code you wrote five or even three years ago, you're not learning anywhere near enough

Seek out, study, and bookmark good code

Learn to program axiomatically

> take each element of the system, language, and toolchain, and learn it throughout

Keep all your projects in source control systems like git or svn