Hi, I am Rekha Joshi!
Seven Habits Of
Highly Effective Big Data Programmers

Rekha Joshi (@rekhajoshm)
Principal Engineer, Intuit
Humanity Took 4.5 billion years, to know it is 4.5 billion years old!

**Human Timeline:**
5 Million Years

- 3.7 Million: First Humans
- 2 Million: First Tools
- 1 Million
- Present Day
- 360,000: Use of Fire
- 100,000: Migrations out of Africa
- 12,000: Agricultural Revolution
- 40,000: Modern Humans Complex Language

**Agricultural Revolution!**

**Industrial Revolution!**

**Scientific Revolution!**
Human Future Timeline:

500 Years


Future Scenarios:
1. Momentum: Future is just like today, only bigger
2. Breakdown: Present and unforeseen problems crash the system
3. Breakthrough: Present and unforeseen technology solve pressing problems
4. What the World Wants: Present technology, human ingenuity, values and persistence create the future people envision and desire

Transportation/Communication:
Roads linking villages, towns, nations
Sailing Ships

Letters

Telegraph
Telephone
Radio Television
Computer
Internet

Computer Revolution! Data Revolution!
Data Revolution - Human Urge To Connect Dots And Know!

Curious
Data Revolution - Sharing
What is Big Data?

“Does this count as big data?”
More Than Just Data!
Value Of Big Data == Value From Big Data

So you got big data?
Yeah. We have billions of records, hundreds of billions of fields of data, thousands of users.

What can you do with it?
Not much... we collect it.
Big Data Use Cases

- Agriculture
- Business
- Climate
- Consumer
- Ecosystems
- Education
- Energy
- Finance
- Health
- Local Government
- Manufacturing
- Ocean
- Public Safety
- Science & Research
Open Data

Stanford Large Network Dataset Collection

- Social networks: online social networks, edges represent interactions between people
- Networks with ground-truth communities: ground-truth network communities in social and information networks
- Communication networks: email communication networks with edges representing communication
- Citation networks: nodes represent papers, edges represent citations
- Collaboration networks: nodes represent scientists, edges represent collaborations (co-authoring a paper)
- Web graphs: nodes represent webpages and edges are hyperlinks
- Amazon networks: nodes represent products and edges link commonly co-purchased products
- Internet networks: nodes represent computers and edges communication
- Road networks: nodes represent intersections and edges roads connecting the intersections
- Autonomous systems: graphs of the internet
- Signed networks: networks with positive and negative edges (friend/foe, trust/distrust)
- Location-based online social networks: Social networks with geographic check-ins
- Wikipedia networks and metadata: Talk, editing and voting data from Wikipedia
- Twitter and Memetracker: Memetracker phrases, links and 487 million Tweets
- Online communities: Data from online communities such as Reddit and Flickr
- Online reviews: Data from online review systems such as BeerAdvocate and Amazon

SNAP networks are also available from UF Sparse Matrix collection. Visualizations of SNAP networks by Tim Davis.

Social networks

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Nodes</th>
<th>Edges</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ego-Facebook</td>
<td>Undirected</td>
<td>4,039</td>
<td>88,234</td>
<td>Social circles from Facebook (anonymized)</td>
</tr>
<tr>
<td>ego-Gplus</td>
<td>Directed</td>
<td>107,614</td>
<td>13,673,453</td>
<td>Social circles from Google+</td>
</tr>
<tr>
<td>ego-Twitter</td>
<td>Directed</td>
<td>81,306</td>
<td>1,768,149</td>
<td>Social circles from Twitter</td>
</tr>
<tr>
<td>soc-Epinions1</td>
<td>Directed</td>
<td>75,879</td>
<td>508,837</td>
<td>Who-whom network of Epinions.com</td>
</tr>
<tr>
<td>soc-LiveJournal1</td>
<td>Directed</td>
<td>4,847,571</td>
<td>68,993,773</td>
<td>LiveJournal online social network</td>
</tr>
<tr>
<td>soc-Pokec</td>
<td>Directed</td>
<td>1,632,803</td>
<td>30,622,564</td>
<td>Pokec online social network</td>
</tr>
<tr>
<td>soc-Slashdot0811</td>
<td>Directed</td>
<td>77,360</td>
<td>905,468</td>
<td>Slashdot social network from November 2008</td>
</tr>
<tr>
<td>soc-Slashdot0922</td>
<td>Directed</td>
<td>82,168</td>
<td>948,464</td>
<td>Slashdot social network from February 2009</td>
</tr>
<tr>
<td>wiki-Vote</td>
<td>Directed</td>
<td>7,115</td>
<td>103,689</td>
<td>Wikipedia who-votes-on-whom network</td>
</tr>
</tbody>
</table>

More info.
Impact of Big Data

Open government data is important because the more accessible, discoverable, and usable data is, the more impact it can have. These impacts include, but are not limited to: cost savings, efficiency, fuel for business, improved civic services, informed policy, performance planning, research and scientific discoveries, transparency and accountability, and increased public participation in the democratic dialogue. Below are just a few examples of citizens leveraging open data. While we don't endorse any particular use, we're always interested in new examples: Submit impact suggestions here.
Intuit Mission

To improve our customers’ financial lives so profoundly … they can’t imagine going back to the old way!
Who We Serve

Consumers

Small Businesses

Accounting Professionals

60+ products and services, covering 24% of US GDP alone!
Big Data Technologies At Intuit

Data-Driven Products
- TurboTax
- mint.com
- QuickBooks
- intuit Demandforce
- intuit Payroll

Data Products
- Search
- Business Lookup
- Unified Profile
- Personalization
- A/B Testing
- Intuit Business Graph

Data pipeline
- elasticsearch
- Apache Kafka
- Spark

Data Platform
- VERTICA
- others
Big Data Technologies At Intuit

Apache, Open Source Projects
- Hadoop, Hive, Spark, Kafka, Cassandra and ecosystem
- Docker, Mesos, Terraform, Ansible, aws-cli, Chef, Gradle
- ElasticSearch, Kibana, Grafana, Sensu, Prometheus
- Yhat, H2O, Spark MLlib

Enterprise
- Cloudera CDH
- Amazon Cloud, AWS services, AWS ML platform
- Vertica, Splunk, New Relic, PagerDuty

Proprietary
Challenges
This Too Shall Pass...
Too much, Too Big!
It’s Easy to Lose Your Way
Alienation?
Afraid Of Failures
Unsure?
I am not a great programmer:

I am a good programmer with great habits!
Explore - Shaping The Correct Future
Explore - Shaping The Correct Future

Data Scientist

Data Engineer

Statistician
Not Just Linearly - Being Creative
Not Just Linearly - Being Creative

The problems of tomorrow will NOT be just blown-up versions of problems of today!
Design And Redesign

As Fast As Possible, As Slow As Necessary

Steve Yegge
Someday my foot won't fit in my mouth.

Following

Staff Software Engineer at Google
Attended University of Washington
Lives in Seattle, WA
Design And Redesign
Design And Redesign
Newton SAW the Apple!
Newton SAW the Apple!
Have Evaluation Built In
Continuous AB Evaluations
Placing data at the heart of every important product decision
Get Your Mathematics Kung Fu On!
Get Your Mathematics Kung Fu On!
Get Your Mathematics Kung Fu On!
Networking is The King!
Networking is The King!
Networking is The King!
The 8th Habit: Contribute Back!
The Seven Habits Of Highly Effective Big Data Programmers

[Print Replica] Kindle Edition

by Rekha Joshi (Author)

Be the first to review this item

See all formats and editions

Kindle

$0.00 kindleunlimited

Subscribers read for free
$3.99 to buy

Big Data is becoming a world in its own right and being a Big Data Programmer a distinction in itself. A highly effective Big Data Programmer is as much a deadly combination of effective intelligence and mysterious elements as a James Bond. What is that combination, exactly?

This book unravels the Seven Habits of Highly Effective Big Data Programmers.

Format: Print Replica

Available on these devices

See the Best Books of the Month

Want to know our Editors’ picks for the best books of the month? Browse Best Books of the Month, featuring our favorite new books in more than a dozen categories.
Interested in Joining?

goo.gl/BLPfyR
Thank You!

https://www.linkedin.com/in/rekhajoshm

@rekhajoshm