Big Data's Biggest Problem

Extracting Useful Information From The Internet

Abraham Thomas
Founder and Chief Data Officer, Quandl Inc.
www.quandl.com

Part 1

The Knowledge Navigator

Apple's Vision of the Future

- http://www.youtube.com/watch?v=QRH8eimU_20
- Produced in 1987
- Meant to be a vision statement for the next 25 years
- How did we do?

Astonishingly Prescient

Astonishingly Prescient

- Hardware: form factor, multi-touch, voice recognition, UI

Astonishingly Prescient

- Hardware: form factor, multi-touch, voice recognition, UI
- Applications: communication, calendar, productivity

Astonishingly Prescient

- Hardware: form factor, multi-touch, voice recognition, UI
- Applications: communication, calendar, productivity
- The Internet: a vast repository of all human knowledge

Astonishingly Prescient

- Hardware: form factor, multi-touch, voice recognition, UI
- Applications: communication, calendar, productivity
- The Internet: a vast repository of all human knowledge

… But Not Quite Perfect

Where the Prediction Failed

Where the Prediction Failed

- Data Data Data

Where the Prediction Failed

- Data Data Data
- Search for “deforestation in the Amazon rainforest”. [link] The results are useless if you are a professional who wants to work with actual data.

Where the Prediction Failed

- Data Data Data
- Search for “deforestation in the Amazon rainforest”. [link] The results are useless if you are a professional who wants to work with actual data.
- The “information network” exists, but finding what you need takes time and effort. Getting it in useable format takes even more time and effort.

What Went Wrong?

What Went Wrong?

- It turns out that extracting relevant, useful, useable information from a vast network of data is a much harder problem than multi-touch, communications or even voice.

What Went Wrong?

- It turns out that extracting relevant, useful, useable information from a vast network of data is a much harder problem than multi-touch, communications or even voice.

- Why is this? The internet is designed for humans, not machines. It is inherently unstructured.

What Went Wrong?

- It turns out that extracting relevant, useful, useable information from a vast network of data is a much harder problem than multi-touch, communications or even voice.
- Why is this? The internet is designed for humans, not machines. It is inherently unstructured.
- Search engines attempt to extract structure and meaning from internet noise. They have a long way to go.

Part 2

The Evolution of Search

Generation 1:
Imposing Structure from the Outside

Generation 1: Imposing Structure from the Outside

- Lists of files and folders: Archie, Veronica, Gopher

Generation 1: Imposing Structure from the Outside

- Lists of files and folders: Archie, Veronica, Gopher
- Portals and closed gardens: AOL, Excite

Generation 1: Imposing Structure from the Outside

- Lists of files and folders: Archie, Veronica, Gopher
- Portals and closed gardens: AOL, Excite
- Hierarchical directories: DMOZ, Yahoo!

Generation 1:
Imposing Structure from the Outside

- Lists of files and folders: Archie, Veronica, Gopher
- Portals and closed gardens: AOL, Excite
- Hierarchical directories: DMOZ, Yahoo!
- Semi-manual indexes: Altavista, Infoseek, Lycos

Generation 1:
Imposing Structure from the Outside

- Lists of files and folders: Archie, Veronica, Gopher
- Portals and closed gardens: AOL, Excite
- Hierarchical directories: DMOZ, Yahoo!
- Semi-manual indexes: Altavista, Infoseek, Lycos

Non-scalable, because they limited their content or relied on in-house human input.

Generation 2:
One Algorithm to Rule Them All

Generation 2:
One Algorithm to Rule Them All

- Google's breakthrough: PageRank

Generation 2:
One Algorithm to Rule Them All

- Google's breakthrough: PageRank
- V1 did not even look at content; instead, relied on topological properties (metadata) to infer relevance and authority.

Generation 2:
One Algorithm to Rule Them All

- Google's breakthrough: PageRank
- V1 did not even look at content; instead, relied on topological properties (metadata) to infer relevance and authority
- Purely algorithmic approach eliminates human curation; leverages Moore's Law.

Smart, But Not Smart Enough

Smart, But Not Smart Enough

- At the end of the day, Google – like every other search engine – is merely a card catalog for the web.

Smart, But Not Smart Enough

- At the end of the day, Google – like every other search engine – is merely a card catalog for the web.
- The user still has to go to the indexed web page, locate the relevant information, and extract it for final use.

Smart, But Not Smart Enough

- At the end of the day, Google – like every other search engine – is merely a card catalog for the web.
- The user still has to go to the indexed web page, locate the relevant information, and extract it for final use.
- This is a far cry from the seamless “digital butler” experience promised in the Knowledge Navigator video.

And Google Knows This

And Google Knows This

- “I’m Feeling Lucky”

And Google Knows This

- “I’m Feeling Lucky”
- The one constant on their home page; the most valuable real estate on the whole web; for a button that nobody presses.

And Google Knows This

- “I’m Feeling Lucky”
- The one constant on their home page; the most valuable real estate on the whole web; for a button that nobody presses.
- Google knows that the future of the internet is this button. Google has known this since 1998. [link]

And Google Knows This

- “I’m Feeling Lucky”
- The one constant on their home page; the most valuable real estate on the whole web; for a button that nobody presses.
- Google knows that the future of the internet is this button. Google has known this since 1998. [link]
- We are finally seeing this future come true.

Generation 3: Vertical Search

Generation 3: Vertical Search

- Niche algorithms provide far richer, more structured results. Kayak, Quixey, Ark, Nexis, Thomas, Indeed, ...

Generation 3: Vertical Search

- Niche algorithms provide far richer, more structured results. Kayak, Quixey, Ark, Nexis, Thomas, Indeed, ...
- Google is now 10 or 12 different search engines, with a single common front-end. Images, Maps, Videos, News, Profiles, Travel, Weather, Finance, Calculator, ...

Generation 3: Vertical Search

- Niche algorithms provide far richer, more structured results. Kayak, Quixey, Ark, Nexis, Thomas, Indeed, ...
- Google is now 10 or 12 different search engines, with a single common front-end. Images, Maps, Videos, News, Profiles, Travel, Weather, Finance, Calculator, ...
- Cards provide “instant answers” without click-through.

Part 3

The Data Frontier

Everything But Data

Everything But Data

- The one area where niche search has not had success is for unstructured or semi-structured numerical data.

Everything But Data

- The one area where niche search has not had success is for unstructured or semi-structured numerical data.
- Because it's really hard. Identify, parse, organize, deliver.

Everything But Data

- The one area where niche search has not had success is for unstructured or semi-structured numerical data.
- Because it's really hard. Identify, parse, organize, deliver.
- At scale, this is close to impossible. It's a big data problem that dwarfs most internal big data challenges.

Big Data's Biggest Problem

Big Data's Biggest Problem

- Volume, velocity, variety. But not all are created equal.

Big Data's Biggest Problem

- Volume, velocity, variety. But not all are created equal.
- Most big data solutions have to deal with volume and velocity challenges. Moore's Law helps. But unanticipated variety is rarely an issue.

Big Data's Biggest Problem

- Volume, velocity, variety. But not all are created equal.
- Most big data solutions have to deal with volume and velocity challenges. Moore's Law helps. But unanticipated variety is rarely an issue.
- Data on the internet is completely the opposite. Individual datasets are small, but the variety is staggering.

Lucrative If You Can Solve It

Lucrative If You Can Solve It

- In aggregate, there is a huge amount of valuable data scattered across the internet.

Lucrative If You Can Solve It

- In aggregate, there is a huge amount of valuable data scattered across the internet.
- Consider “deforestation in the Amazon rainforest”. Useful for agriculture, logging, mining companies, climate scientists, economists, biologists, medical research and more.

Lucrative If You Can Solve It

- In aggregate, there is a huge amount of valuable data scattered across the internet.
- Consider “deforestation in the Amazon rainforest”. Useful for agriculture, logging, mining companies, climate scientists, economists, biologists, medical research and more.
- The data is out there; it’s just hard to find and hard to use.

But the Landscape is Changing

But the Landscape is Changing

- Technological advances + macro trends have enabled vertical search in other domains. E.g. “people search”.

But the Landscape is Changing

- Technological advances + macro trends have enabled vertical search in other domains. E.g. “people search”.
- Similar trends apply in the realm of raw numerical data.

But the Landscape is Changing

- Technological advances + macro trends have enabled vertical search in other domains. E.g. “people search”.
- Similar trends apply in the realm of raw numerical data.
- Vertical search for data is closer now than ever before.

Part 4

Building a Data Search Engine

How Google Does It

1. Crawl
2. Read
3. Rank
4. Direct

How Google Does It

1. Crawl = visit every webpage in the world
2. Read
3. Rank
4. Direct

How Google Does It

1. Crawl
2. Read = figure out what each page contains
3. Rank
4. Direct

How Google Does It

1. Crawl
2. Read
3. Rank = on query, order pages by usefulness
4. Direct

How Google Does It

1. Crawl
2. Read
3. Rank
4. Direct = send user to URL

How Google Does It

1. Crawl
2. Read
3. Rank
4. Direct

This works well for text.

How Google Does It

1. Crawl
2. Read
3. Rank
4. Direct

This works well for text.

It fails horribly for data.

Why This Fails For Data

1. Crawl: only a minority of pages host data
2. Read
3. Rank
4. Direct

Why This Fails For Data

1. Crawl
2. Read: data comes in too many formats
3. Rank
4. Direct

Why This Fails For Data

1. Crawl
2. Read
3. Rank: not enough link or semantic info
4. Direct

Why This Fails For Data

1. Crawl
2. Read
3. Rank
4. Direct: sending off to a URL is not enough

We Need a Deeper Process

1. ~~Crawl~~ Identify
2. ~~Read~~ Parse
3. ~~Rank~~ Organize
4. ~~Direct~~ Deliver

We Need a Deeper Process

1. ~~Crawl~~ Identify = locate numerical data
2. ~~Read~~ Parse
3. ~~Rank~~ Organize
4. ~~Direct~~ Deliver

We Need a Deeper Process

1. ~~Crawl~~ Identify
2. ~~Read~~ Parse = learn how data is formatted
3. ~~Rank~~ Organize
4. ~~Direct~~ Deliver

We Need a Deeper Process

1. ~~Crawl~~ Identify
2. ~~Read~~ Parse
3. ~~Rank~~ Organize = curate, classify, categorize the data
4. ~~Direct~~ Deliver

We Need a Deeper Process

1. ~~Crawl~~ Identify
2. ~~Read~~ Parse
3. ~~Rank~~ Organize
4. ~~Direct~~ Deliver = produce fresh usable data

A Successful Example

- Quandl: www.quandl.com
- Search engine for data: 8 million+ datasets, 1000s of sources
- Rich metadata, structure, polymorphism, API, ecosystem
- Focus on finance, economics, society data

But …
These Are HARD Problems!

1. Identify – computer vision, pattern-matching
2. Parse – ML, inference, heuristic methods
3. Organize – semantic analysis, data mining
4. Deliver – scale, schedule, integrate

Consider Parsing

- Format (CSV, JSON, HTML, PDF, DOC ...)
- Syntax (1,00 versus 1.00, YMD versus DMY versus MDY)
- Structure (rows, cols, splits, joins, nests, gaps)
- Insufficient AND extraneous information

Technology is Not Enough

Technology is Not Enough

- Try to advance all 4 stages in data pipeline.

Technology is Not Enough

- Try to advance all 4 stages in data pipeline.
- But recognize that we will not achieve perfection.

Technology is Not Enough

- Try to advance all 4 stages in data pipeline.
- But recognize that we will not achieve perfection.
- Technology does what it can; humans do the rest.

Build a Hybrid Process

Build a Hybrid Process

- Humans are very good at certain tasks: quick identification, classification, disambiguation.

Build a Hybrid Process

- Humans are very good at certain tasks: quick identification, classification, disambiguation.
- Technology is very good at certain tasks: repetition, polling, parallelization, rules-based transformation.

Build a Hybrid Process

- Humans are very good at certain tasks: quick identification, classification, disambiguation.
- Technology is very good at certain tasks: repetition, polling, parallelization, rules-based transformation.
- Design a process that will maximize joint efficiency.

Coming Full Circle

Coming Full Circle

- Return to the early days of search: use human beings to add or infer structure.

Coming Full Circle

- Return to the early days of search: use human beings to add or infer structure.
- But doesn’t this just run into the same scalability problem?

Coming Full Circle

- Return to the early days of search: use human beings to add or infer structure.
- But doesn't this just run into the same scalability problem?
- Not if these humans are "external". Crowdsourcing!

Build the Right Incentives

Build the Right Incentives

- What is the right incentive? Self-interest.

Build the Right Incentives

- What is the right incentive? Self-interest.
- “Data on Quandl is more valuable to the user than that same data elsewhere.”

Build the Right Incentives

- What is the right incentive? Self-interest.
- “Data on Quandl is more valuable to the user than that same data elsewhere.”
- Value can take many forms: convenience, affirmation, money.

Search Engine → Data Platform

Search Engine → Data Platform

- A 1-way search engine becomes a 2-way data platform: much more powerful and flexible.

Search Engine → Data Platform

- A 1-way search engine becomes a 2-way data platform: much more powerful and flexible.
- Behavioral and network effects begin to dominate.

Search Engine → Data Platform

- A 1-way search engine becomes a 2-way data platform: much more powerful and flexible.
- Behavioral and network effects begin to dominate.
- And now we are in Web 2.0 territory.

Part 5

The User-Driven World

Web 2.0

Web 2.0

- Value-add comes from User-Generated-Content (UGC).

Web 2.0

- Value-add comes from User-Generated-Content (UGC).
- Consumers are producers and producers are consumers.

Web 2.0

- Value-add comes from User-Generated-Content (UGC).
- Consumers are producers and producers are consumers.
- Web 2.0 is the flip side of the consumerization of enterprise.

Web 2.0

- Value-add comes from User-Generated-Content (UGC).
- Consumers are producers and producers are consumers.
- Web 2.0 is the flip side of the consumerization of enterprise.
- Network effects are critical.

Web 2.0 Examples

- Social networks: Twitter, Facebook, LinkedIn, Google+
- Publishing sites: Blogger, Tumblr, Medium
- Photo-sharing: Instagram, Snapchat, 500px
- Discovery: Reddit, Digg, StumbleUpon, Delicious
- Reviews: Yelp, TripAdvisor
- Expert Answers: Quora, Stack Exchange
- Location-based: Foursquare, NextDoor
- Implicit learning: Xero, Waze

Part 6

The Open Data Revolution

Limitless Raw Material

Limitless Raw Material

- None of this is possible without open data.

Limitless Raw Material

- None of this is possible without open data.
- First multinational agencies & governments, now businesses.

Limitless Raw Material

- None of this is possible without open data.
- First multinational agencies & governments, now businesses.
- Why do it?

Limitless Raw Material

- None of this is possible without open data.
- First multinational agencies & governments, now businesses.
- Why do it?
  - Governments: transparency, public good.

Limitless Raw Material

- None of this is possible without open data.
- First multinational agencies & governments, now businesses.
- Why do it?
  - Governments: transparency, public good.
  - Businesses: brand awareness, ecosystem creation.

Data Scarcity is Over

Data Scarcity is Over

- People will not pay for “mere” data.

Data Scarcity is Over

- People will not pay for “mere” data. \
- Data has no value in itself;
  Its only value lies in what can be done with it.

Data Scarcity is Over

- People will not pay for “mere” data.
- Data has no value in itself;
  Its only value lies in what can be done with it.
- People will pay for attributes of data:
  Uniqueness, Quality, Freshness, Convenience, Completeness

Data Ubiquity is Just Beginning

Data Ubiquity is Just Beginning

- It is more important than ever to find the right datasets, and extract the right insights from them.

Data Ubiquity is Just Beginning

- It is more important than ever to find the right datasets, and extract the right insights from them.
- Organizations have masses of data that they can monetize internally or externally (consumers = producers).

A Network Effect for Data

A Network Effect for Data

- 1 data point is rarely useful

A Network Effect for Data

- 1 data point is rarely useful
- 10 data points can tell a story

A Network Effect for Data

- 1 data point is rarely useful
- 10 data points can tell a story
- 100s of data points from 10s of datasets can generate insight

A Network Effect for Data

- 1 data point is rarely useful
- 10 data points can tell a story
- 100s of data points from 10s of datasets can generate insight

Data yields value when it is correlated and cross-referenced with other data. Context is everything.

A Confluence of Trends

- All these trends are coming together:
  - The open data movement
  - Crowd-sourcing and web 2.0
  - The consumerization of enterprise
  - Data ubiquity not scarcity
  - Network effects in data

- To create: the emerging universal data layer

Part 7

The Emerging Data Layer*

* credit Mark Suster of Upfront Ventures

Migrating to the Cloud

Migrating to the Cloud

- In the past you had to do everything yourself: memory, cycles, data, software

Migrating to the Cloud

- In the past you had to do everything yourself: memory, cycles, data, software
- Not any more. Everything is migrating to the cloud. “X as a Service”, for all values of X.

Migrating to the Cloud

- In the past you had to do everything yourself: memory, cycles, data, software
- Not any more. Everything is migrating to the cloud. "X as a Service", for all values of X.
- These services are increasingly organized like a traditional computational "stack".

The Cloud has Layers

Storage (S3)

The Cloud has Layers

Processing (EC2)

Storage (S3)

The Cloud has Layers

The Cloud has Layers

Business Logic (startup / existing company)

Management (RightScale, Heroku)

Processing (EC2)

Storage (S3)

The Cloud has Layers

Business Logic (startup / existing company)

+ custom software, data store

Management (RightScale, Heroku)

Processing (EC2)

Storage (S3)

The Cloud has Layers

Business Logic (startup / existing company)

App Layer (Marketo, Workday, Zendesk)

Management (RightScale, Heroku)

Processing (EC2)

Storage (S3)

The Cloud has Layers

Business Logic (startup / existing company)

App Layer (Marketo, Workday, Zendesk)

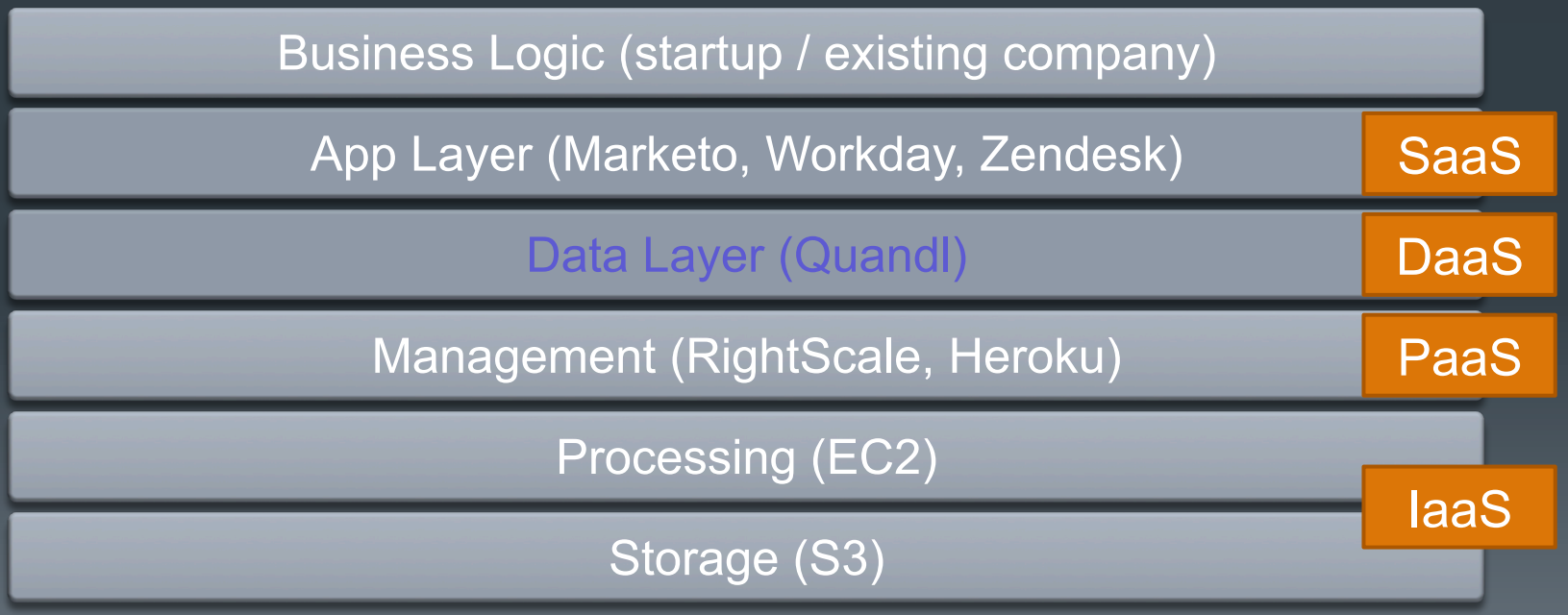
Data Layer (Quandl)

Management (RightScale, Heroku)

Processing (EC2)

Storage (S3)

The Cloud has Layers



Data as a Service

Data as a Service

- Everything is just an API call away.

Data as a Service

- Everything is just an API call away.
- Invoke memory, cycles, data and services, seamlessly.

Data as a Service

- Everything is just an API call away.
- Invoke memory, cycles, data and services, seamlessly.
- Get, put, publish, update, transform, merge:
  the data layer abstracts away the complexity.

An Explosion of Context

An Explosion of Context

- Value comes from context. Seamless merge of internal and external data massively boosts available context.

An Explosion of Context

- Value comes from context. Seamless merge of internal and external data massively boosts available context.
- Unlock massive new value in your data by correlating it with everything else in the data layer.

An Explosion of Context

- Value comes from context. Seamless merge of internal and external data massively boosts available context.
- Unlock massive new value in your data by correlating it with everything else in the data layer.
- And reduce costs at the same time.

Part 8

What Does This Mean For You?

End Users + Analysts

- Self-service, a la carte, granular data delivery
- Extracting insights from context
- Blurring roles: consumers = producers

Developers

- Data as a Service
- Abstraction and Simplification

Information Officers

- Merge internal and external data
- Unlock new sources / streams of value
- Monetize untapped data resources
- Cost effective

Publishers + Vendors

- Wider distribution into new markets
- More control of economics

Governments

- Transparency and governance
- Efficiency in providing services
- New enterprises and value creation

Thank You!

Abraham Thomas
Founder and Chief Data Officer, Quandl Inc.
thomas@quandl.com