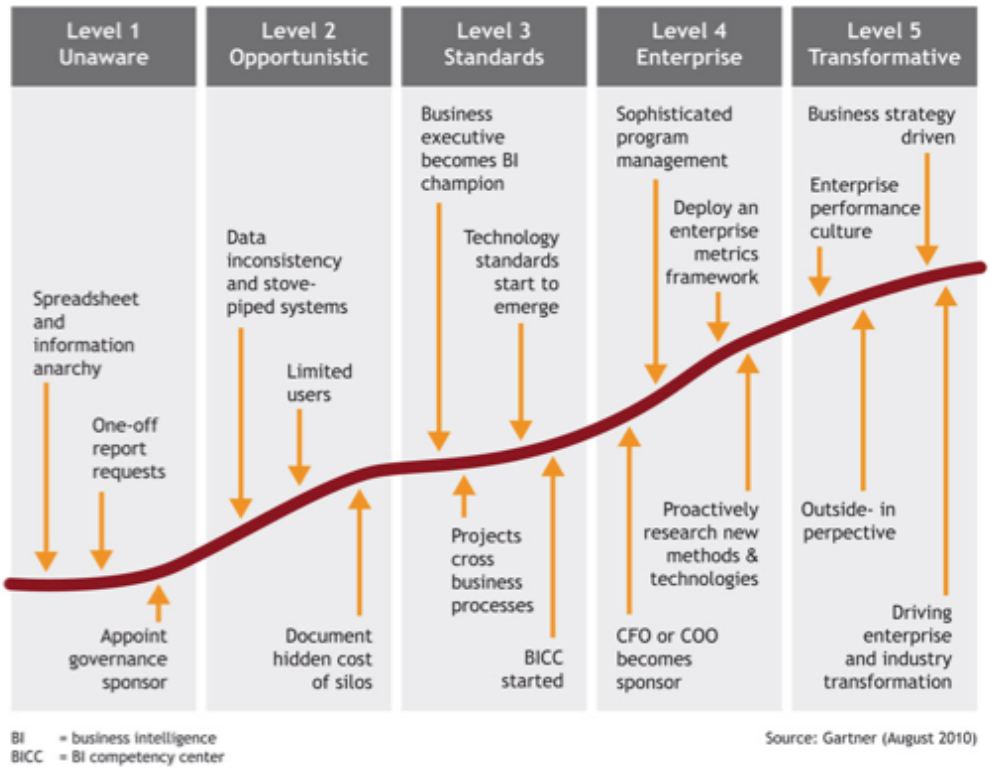
Innovation Fund Final Report: Exploration of Data Analytics Tools for Library Data

Principal Investigators: Jim Hahn, Jen-chien Yu and Megean Osuchowski
Project Timeline: October 2015 - March 2016
Budget: $10,413.00 (Spent: $8,325 as of March 25)
Final Report Submitted on: 03/31/16

Project Summary

The “Exploration of Data Analytics Tools for Library Data” project (the Project) was funded in Fall 2015. The Team included principal investigators and an Academic Hourly hired with funding by the Innovation grant. Though the Library is currently using pockets of business intelligence at many different levels across the organization, using Gartner’s business intelligence maturation roadmap as a reference, we see an opportunity to strengthen our business intelligence maturation levels from opportunistic to standardized, enterprise levels, and finally for business transformation.



In particular, through this initiative, we hope to remove hidden costs of data silos and open business intelligence up to more faculty and staff users to better support data informed decision making. Therefore, the technologies that we are advocating have been selected particularly for their abilities for open business intelligence across the Library organization.

Goals and Objectives:

From October 2015 to March 2016 we completed the following:

A. Reviewed literature and best practices for library data analytics

We reviewed literature on data warehousing and decision support systems in libraries and the corporate world; drew on industry reports from Gartner, and investigated what our peer institutions are doing in this area. The Team members also attended various webcasts and data user meetings related to library data dashboards and analytics.

B. Gathered requirements for data visualizations and use cases of data dashboards

We discussed data and analytics needs with members from Library IT, Scholarly Commons, Voyager users and the Library Assessment Committee. The Team used information collected during this project to decide on what data/data sources (e.g. reference transaction data from Desk Tracker, bibliographic data from Voyager, etc.) and data analytics tools (e.g. Statistica, Tableau, Zoomdata, etc.) to focus on.

C. Tested analytics and visualization tools; demonstrated tools for stakeholders and gathered feedback.

We purchased one Tableau Desktop license. The Team created test workbooks[1] which connected to Voyager data by way of the CARLI reports server and made them available via Tableau Server (access is available on campus through AITS). The Team also installed a trial version of Zoomdata (http://www.zoomdata.com/) and created sample visualizations. The trial period of Zoomdata software has since expired.

The Team held two open presentations (February 3rd and February 9th) and invited Library Executive Committee members, Library IT managers, Library Business

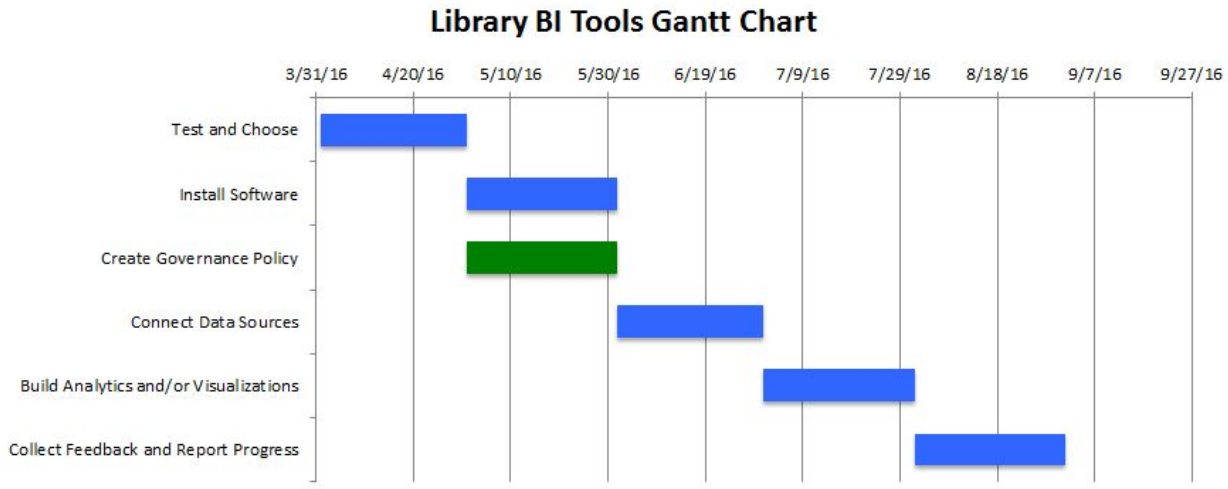
[1] A Tableau term that means a bundle of data with visualizations and analytics

Office and other library colleagues who have expressed interested in this project. The Team shared a PowerPoint presentation (attached) which summarized findings from our review of literature and best practices and demonstrated test visualizations from Tableau and Zoomdata. The presentations ended with open discussion and participants shared their comments on the literature review, test visualizations and suggestions for next steps; their comments are also attached with this report. We gathered additional data/data sources to collect for visualization and analytics (see Recommendations section C. of this report for sample holistic data points).

D. Created a roadmap for future work : phase 2

The team identified next steps, assuming the Library will support the development of library-wide data analytics solutions, and timeline of future work.

April- Library IT tests candidate recommendations (of a Tableau / Microsoft stack) or chooses alternative(s).
May- Library IT installs selected software. A data governance policy is authored by key stakeholders among Library Staff, Library Faculty and Library IT.
June- IT connects required data sources to newly installed software.
July- A group of Faculty and Staff stakeholders creates analytics and visualizations.
August- Get feedback from faculty and staff and provide updated phase 2 report.



Limitations and Challenges

In the Innovation Fund proposal, we presented several “overall objectives” as examples of what problem(s) library data analytics tools can solve. After diving into the Project it is clear to the Team that some objectives can’t be studied at the present time. For example, the objective of “Reduce staff time and resources so assessment and IT staff can utilize expertise in other areas” - while it is possible to achieve such objective in the future, the Library would have to have an advanced and robust data analytics solution in place first, and it will take dedicated time and resources to get a robust solution in place.

Recommendations

A. Further work is needed to identify data analytics solutions supported by the Library. More specifically, Tableau and Microsoft Server tools are both viable choices but offer different features and serve different needs. An implementation of both software tools is possible, with Tableau serving visualization needs and Microsoft tools (Microsoft SQL Server Integration Services, Microsoft SQL Server, Microsoft SQL Server Analysis Services, SQL Server Reporting Services and/or Microsoft Excel) for extracting data, transforming data, and loading data.

B. For extracting, transforming, and loading, we recommend loading all the operational data required for visualization to the centralized storage system of a campus or Library data warehouse (i.e. Online Analytical Processing or “OLAP” cubes). A business case could be made for utilizing the AITS campus data warehouse since Library resources represent a significant business investment for campus.

C. Future research uses of data analytics and visualization tools by library Faculty and Staff include:

    a. Gaining a deep understanding users’ information seeking behavior and information discovery patterns by combining data from multiple sources like Easy Search, IWonder, and VuFind searches, with functionality to annotate. Researchers will follow an IP, timestamp, or session ID among Easy Search, IWonder, and VuFind gaining profound insight into Internet based user services across our digital touch points.

b. Providing a holistic picture of the usage of the Library Gateway, unit web pages and other web-based services (iWonder, VuFind or Easy Search logs)

c. Supporting “Bean Counter-like” reports that are based on Voyager data

d. Other high value visualizations and analytics which pull together different data sources to help make decisions. As an example: if collection developers are looking at withdrawing books, which of the candidate withdrawals are held in less than 5% of OCLC member libraries and are digitized in Hathitrust? And, a study of connections among the inventory software program at Oak Street and Voyager collections data. Combining data silos in Business Intelligence tools would help to compare Oak Street inventory to what is in Voyager (to chart possible discrepancies) and also would chart how the collection from Oak Street has evolved over time.

Appendix -- Feedback Collected from open presentations

Use cases

- Reporting/visualization interface to replace Bean Counter
    - Same bean counter queries should be able to be imported into new system
- Some are interested in performing/viewing text analysis as well as data analysis
    - Example: IM chat data
- People are especially interested in the possibility of cross-referencing data from multiple sources or ability to look across multiple sources
    - Example: collection-level analysis that incorporates Voyager & e-resource data, or incorporating data on physical and virtual usage of library spaces
- It may be best to offer individuals who want to examine data more deeply read-only access to a database to query and examine it
    - Some are interested in the ability to code or tag data in the dashboard, look deeply into the data
    - Custom SQL queries
- It could be interesting to identify basic information needs for units and provide a dashboard for each unit
- Interested in predictive and diagnostic analysis, but not there yet

Recommendations

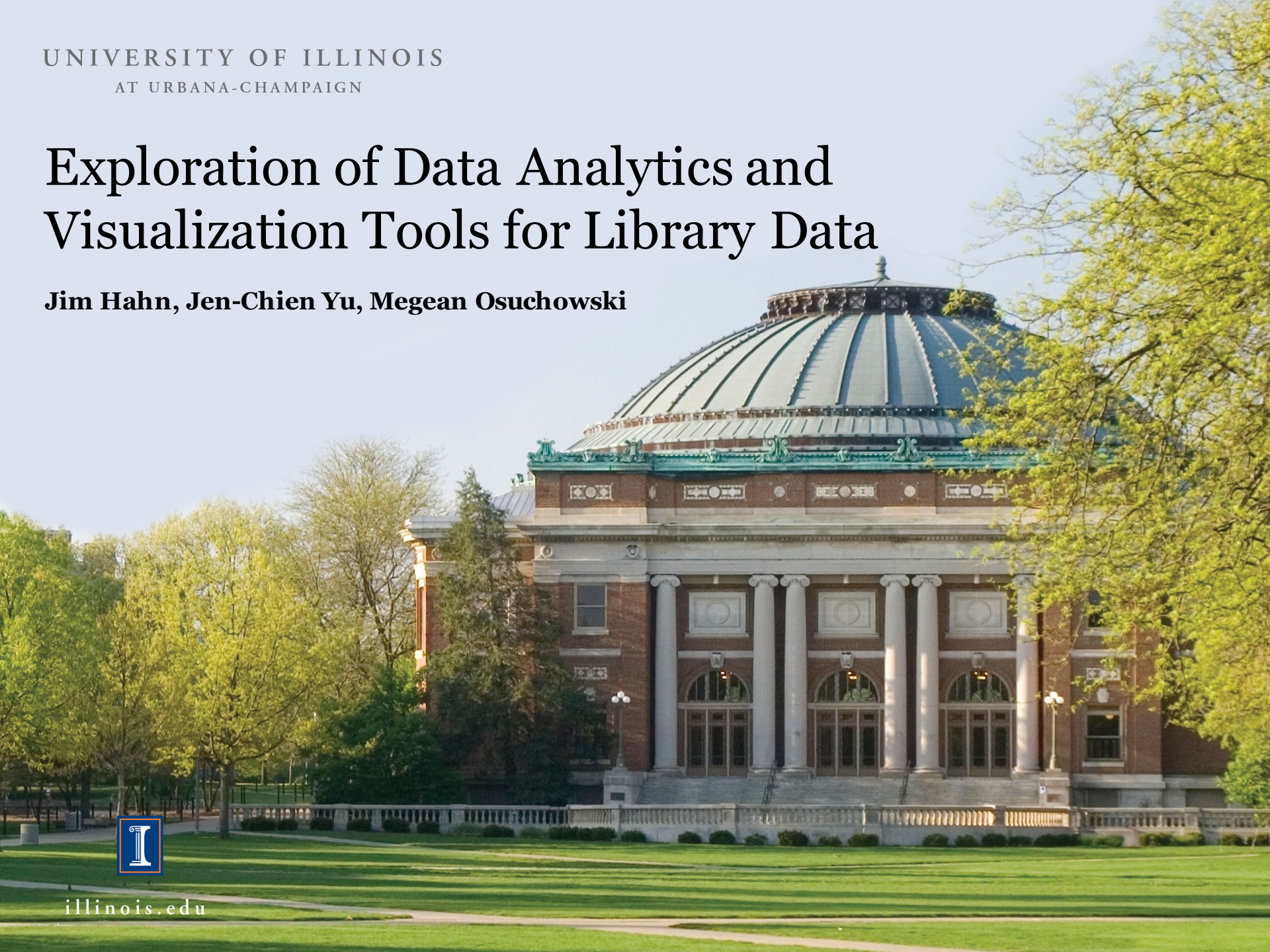
- IT has very limited developer resources and prefers a system that is easy to maintain
- Very positive support on the idea of the project
- Use a solution that's adopted and/or supported by other groups on campus
- Long-term planning & solutions: Think how tools will operate with future additions (ETL, data warehouse, etc). How well will these tools operate as our business intelligence needs mature?
- Data governance will be very important
    - The system will only be successful if we can ensure the accuracy of the data
        - Who will be responsible for data curation and accuracy?
        - Limit access to editing the data
    - What library data is public, secure, or institutional?
    - Must allow flexibility for data integration & analysis decisions
    - Data governance policy should be created with expertise and advice of data experts and those supporting the technology
- Follow-up investigation into data integration solutions. What can be integrated and how?
    - What are the best tools for data storage and analysis?
        - Numerical and textual data exists

    - This investigation and then consequential implementation will allow for more complex questions and queries and push the library towards greater BI maturity opportunities. Should not be skipped.
- Provide staff training on reading and understanding data to help ensure accurate translations of available data
- Consider tools that could be usable for not only internal library data, but can also provide or teach to patrons. It would be one less system to learn.
- Would something like a CIC library data warehouse be useful?

UNIVERSITY OF ILLINOIS
AT URBANA-CHAMPAIGN

Exploration of Data Analytics and Visualization Tools for Library Data

Jim Hahn, Jen-Chien Yu, Megean Osuchowski

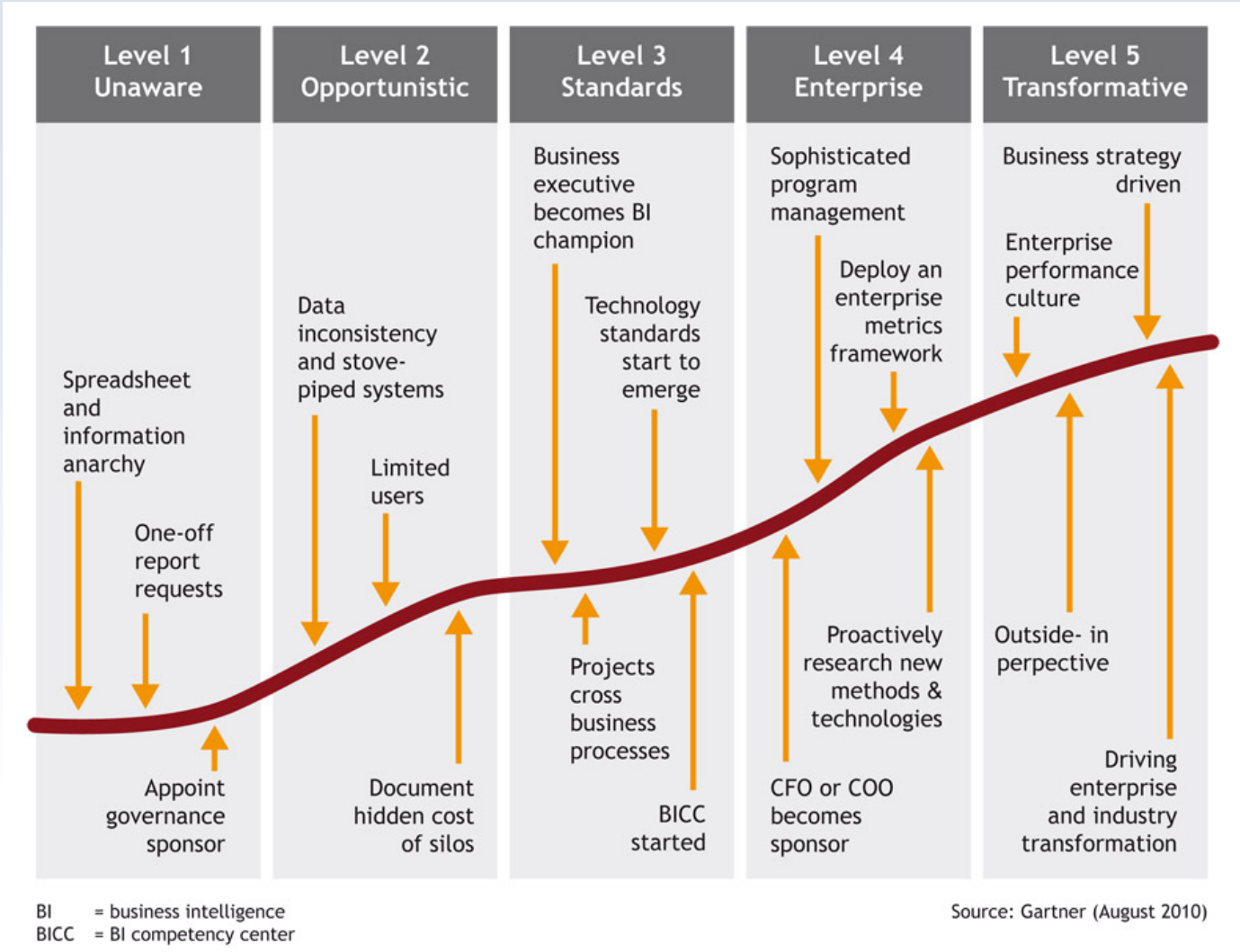


illinois.edu

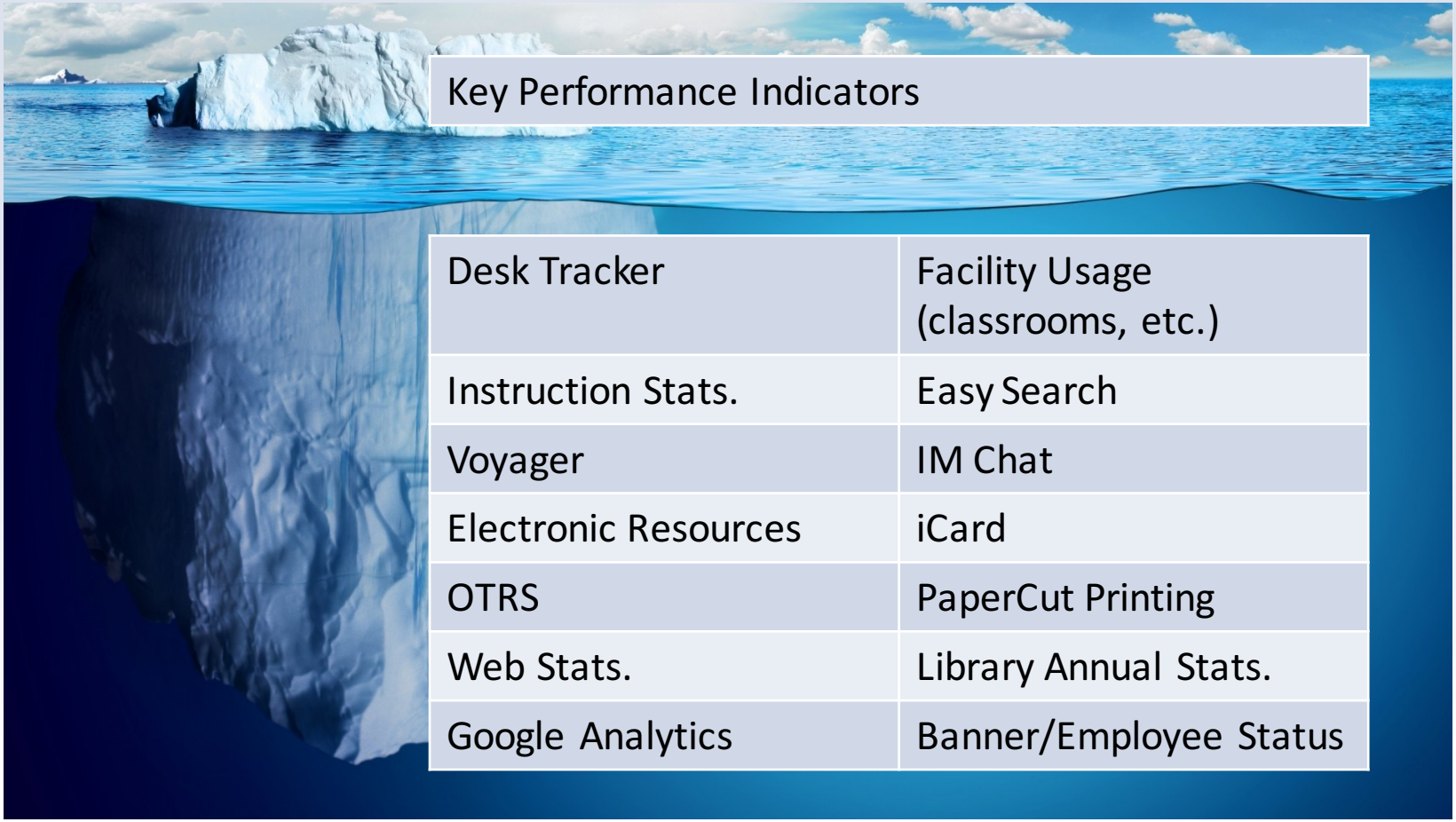
Overview of Analytics Exploration

1) Proposal
2) Technology Options
3) Top Candidates
4) Proposed Next Steps
5) Questions and Discussion

Proposal – BI Maturation

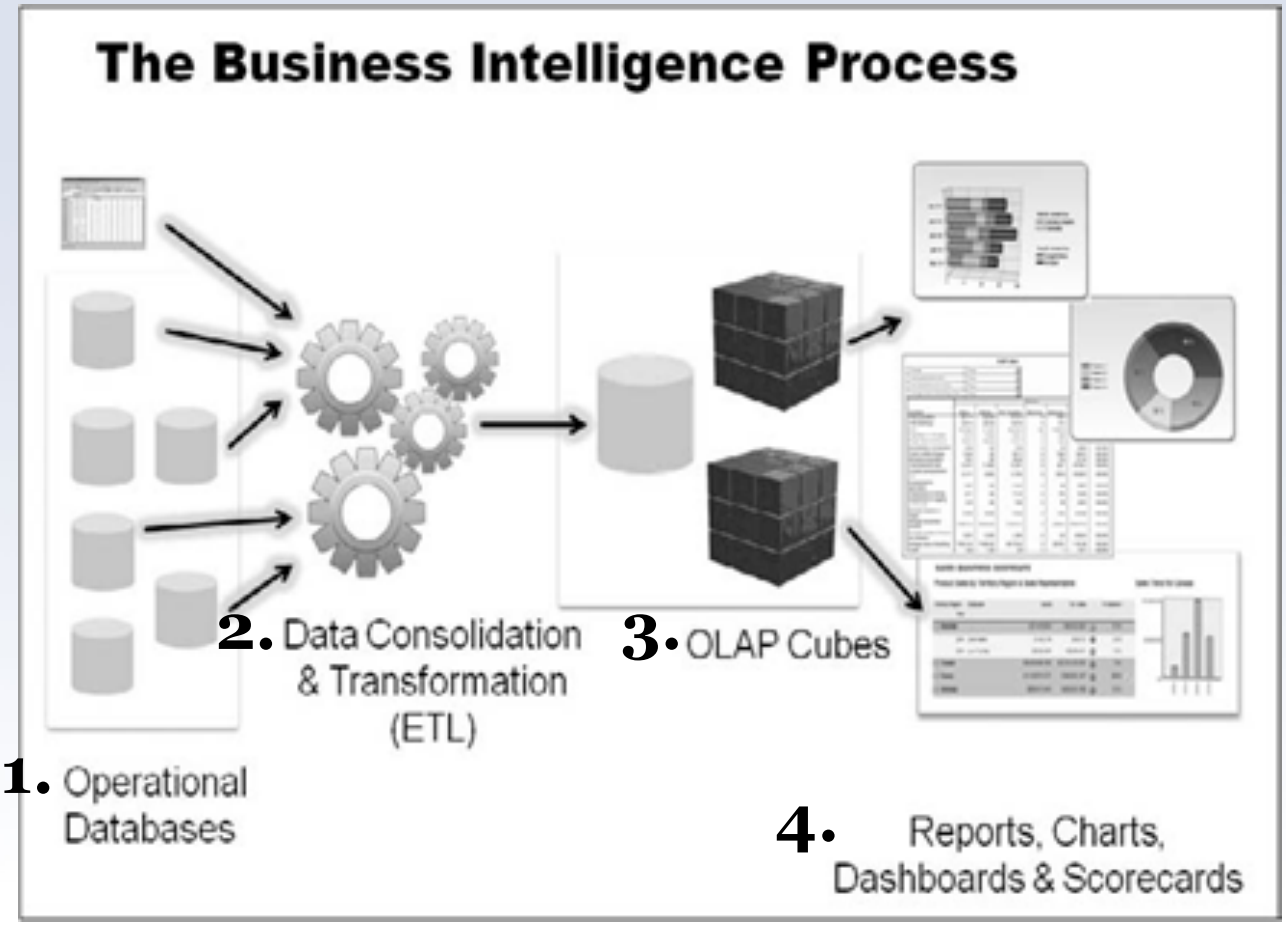


Proposal – Unlocking Data

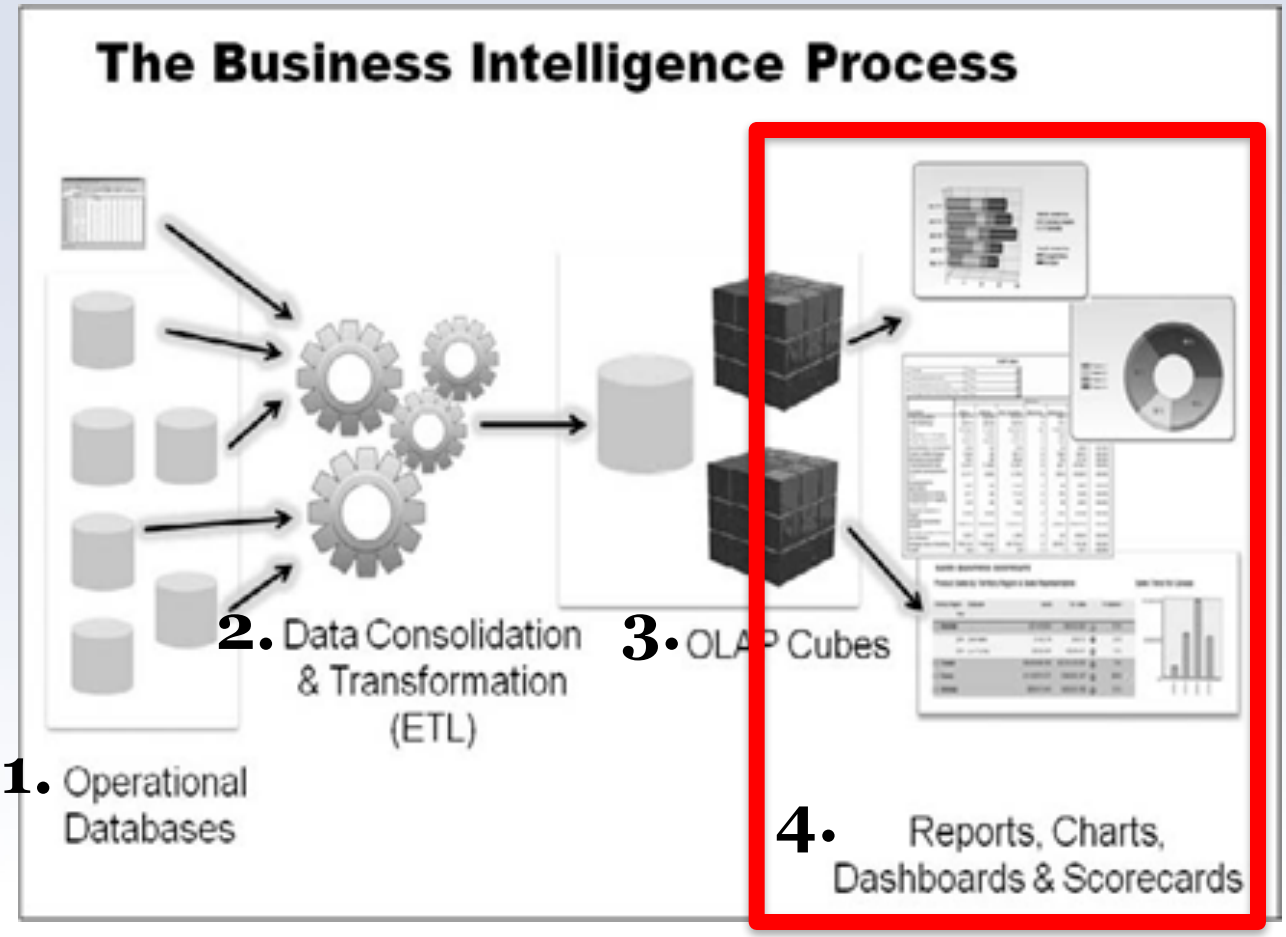


Key Performance Indicators	
Desk Tracker	Facility Usage (classrooms, etc.)
Instruction Stats.	Easy Search
Voyager	IM Chat
Electronic Resources	iCard
OTRS	PaperCut Printing
Web Stats.	Library Annual Stats.
Google Analytics	Banner/Employee Status

Tech. Options – Building Process



Tech. Options – Building Process

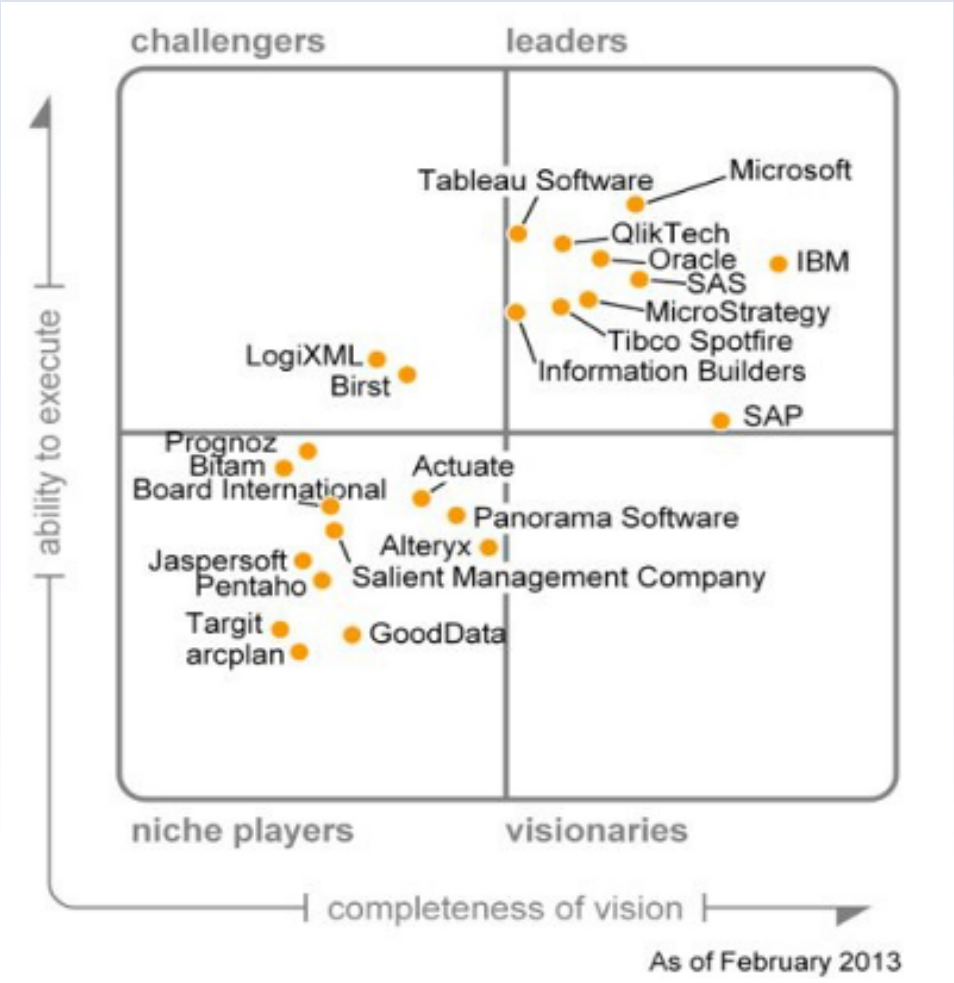


Tech. Options – Software List

Phase 2: Extract, Load, and Transform	**SSIS, Informatica, Python*, Perl*, Statistica ETL**, Websphere DataStage, SAP BusinessObjects, Cognos Data Manager, Oracle Data Integrator, SAS Data Integration Studio, Pentaho Data Integration
Phase 3: Data Warehouse and Analytics	**SSAS, Looker, SAS, R*, Python*, Perl*, Matlab*, Gephi**, **Weka**, Siebel Business Analytics Applications, SAP's BusinessObjects XI, Hyperion System 9 BI+, Pentaho's Open BI Suite, IBM's Cognos 8 BI, TIBCO Spotfire's Enterprise Analytics
Phase 4: Visualization and Reporting	**SSRS, Looker, Zoomdata, Tableau, Excel, Python*, R*, D3.js***, Dygraphs, Zingchart, NetCracker, Quirkos, Splunk, Datameer, Decibel Insight, GoodData, HappyMetrix, PlanZent

*Significant Software Developer Time Involved

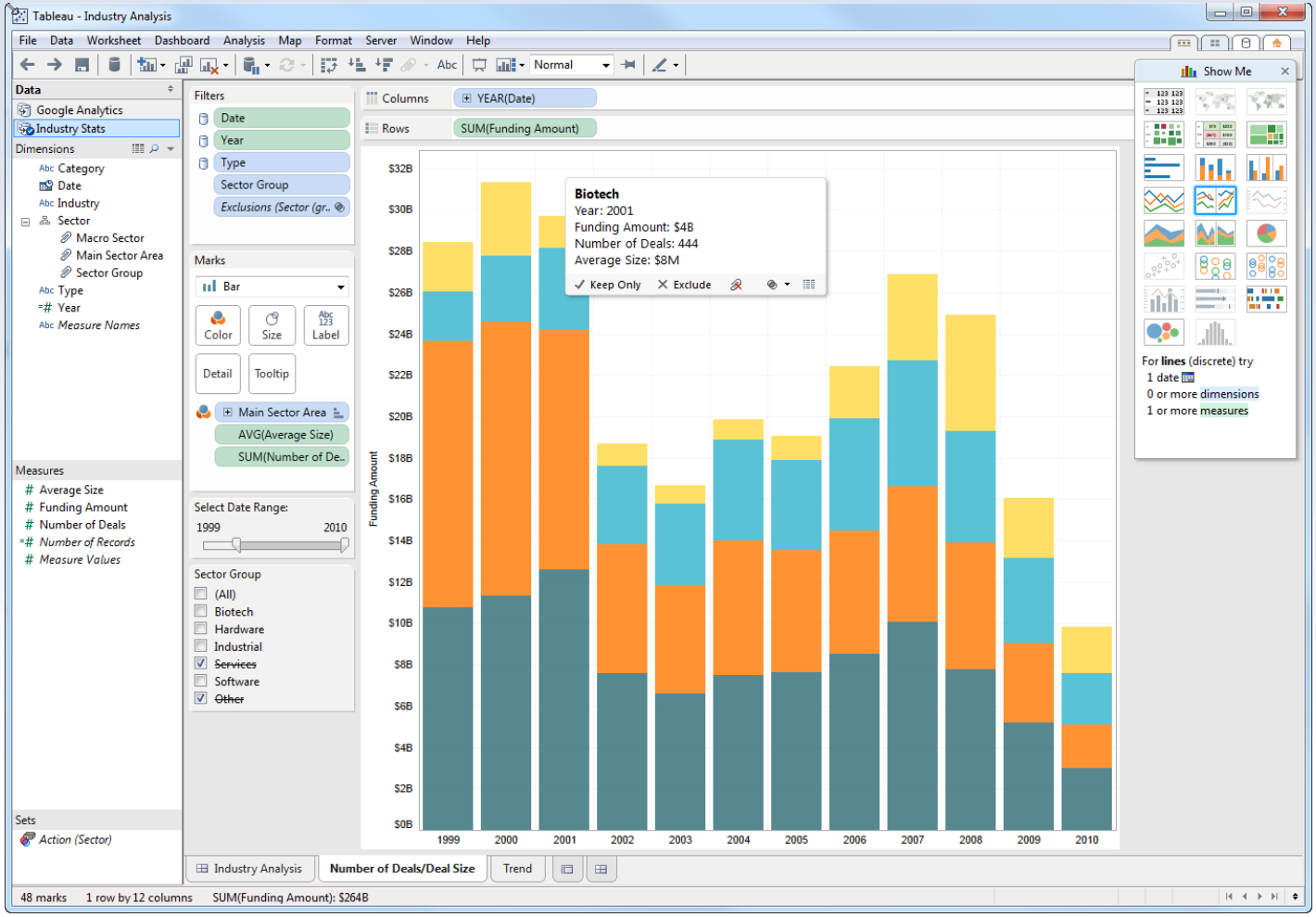
Top Candidates – Microsoft



Top Candidates – Microsoft

Pros	Cons
• Has integrated technology for all business intelligence phases • Integrates with Excel • Easier AITS integration • Very stable and supported enterprise software • Widely used among data scientists and data analysts • Campus licensing	• Requires a database administrator for all new report types • Strict requirements for data format, roles, etc. create limited accessibility for faculty • Poor web integration options • Cannot handle heterogeneous data sources very well

Top Candidates – Tableau



Top Candidates – Tableau

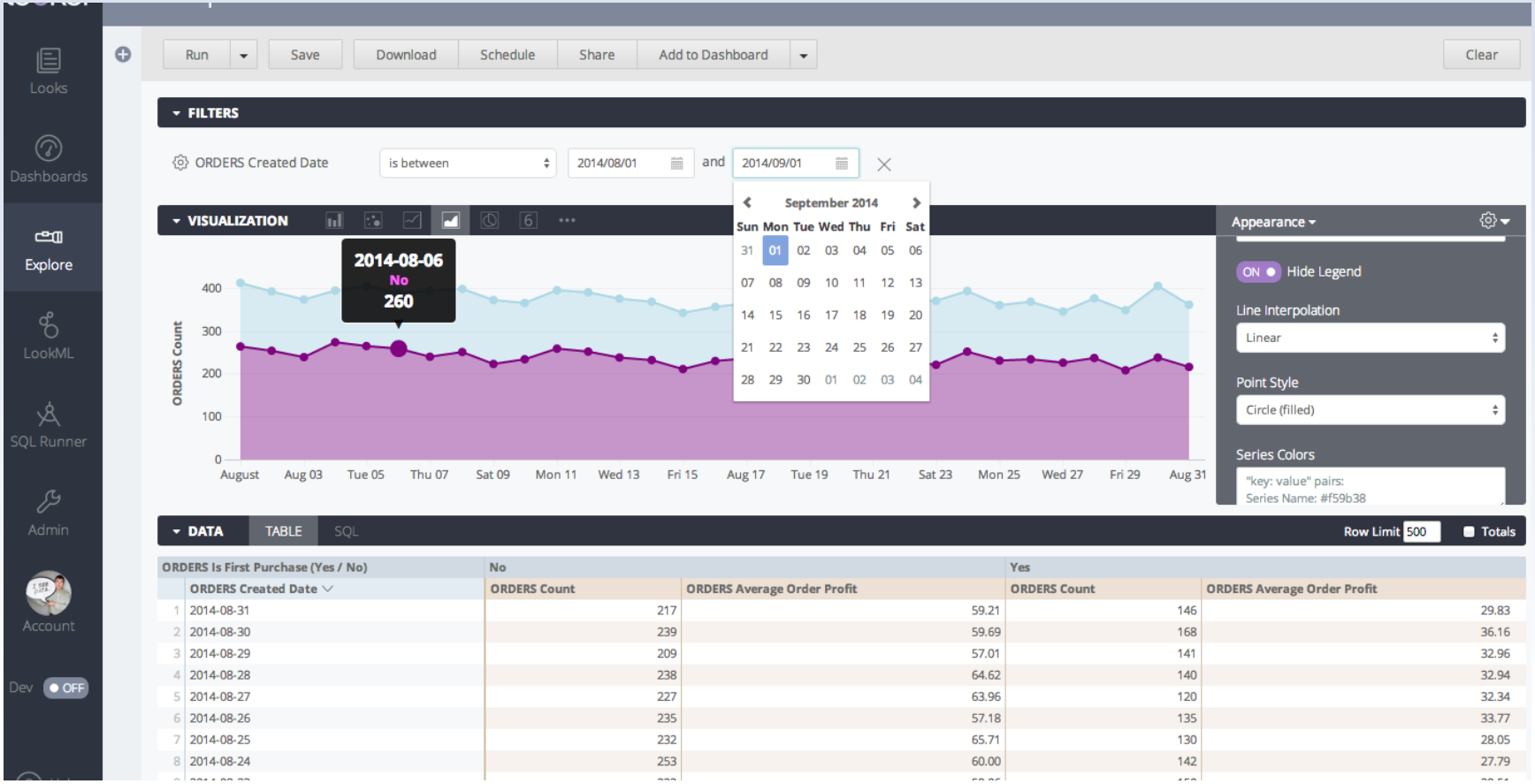
Pros	Cons
• Non-technical users can easily create dashboards • Many graph types available • Can create simple, computed attributes fairly easily (e.g. time * cost = money_spent) • Widely used among data scientists and a data analysts • Shared campus Tableau server available at no cost	• Mostly a visualization tool (computations are very limited) • Limited to built-in graph types (i.e. no ability to build new graph types) • Desktop Tableau instances are licensed separately

Top Candidates – ZoomData

Top Candidates – ZoomData

Pros	Cons
• Non-technical users can easily create dashboards • Very extendable; web developers can build reusable graphs with D3.js • Wordpress integration with iFrames • LDAP compatible • Web-based • Can connect to live data and distributed systems • Responsive representatives • Works on Redhat	• Smaller company (created in 2012 and recently raised $22.2 million) • Is only a visualization tool • Limited graph types available without further web development

Top Candidates – Looker

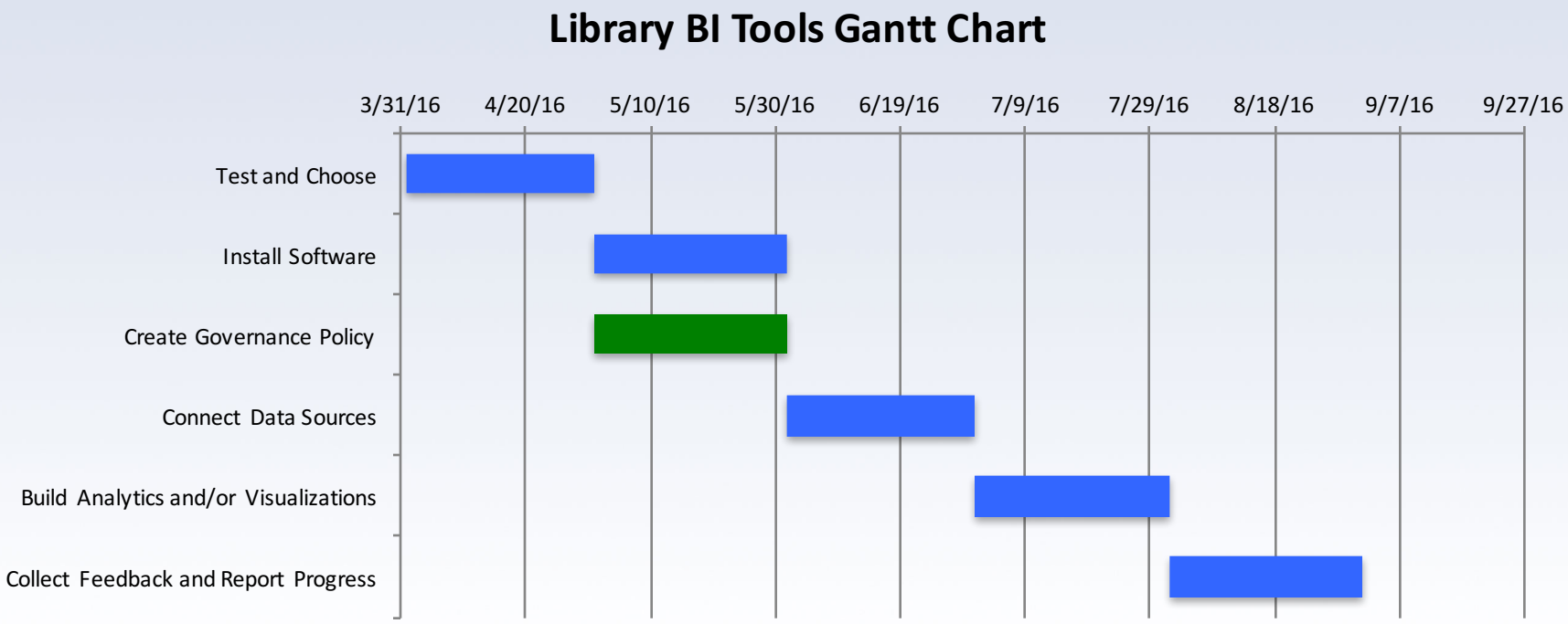


ORDERS Is First Purchase (Yes / No)	No		Yes	
ORDERS Created Date	ORDERS Count	ORDERS Average Order Profit	ORDERS Count	ORDERS Average Order Profit
1 2014-08-31	217	59.21	146	29.83
2 2014-08-30	239	59.69	168	36.16
3 2014-08-29	209	57.01	141	32.96
4 2014-08-28	238	64.62	140	32.94
5 2014-08-27	227	63.96	120	32.34
6 2014-08-26	235	57.18	135	33.77
7 2014-08-25	232	65.71	130	28.05
8 2014-08-24	253	60.00	142	27.79

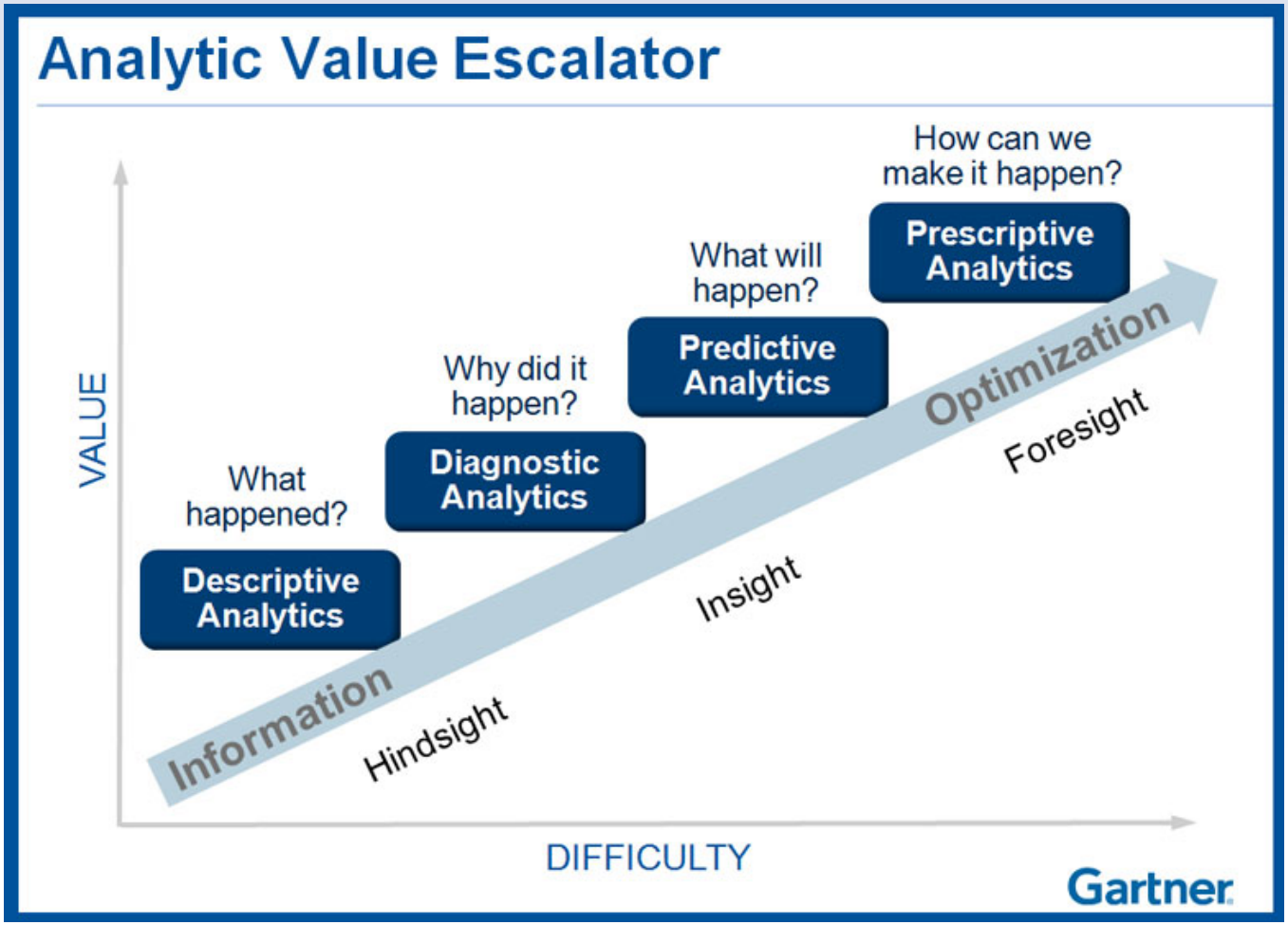
Top Candidates – Looker

Pros	Cons
• Non-technical users can explore SQL databases fairly easily • Non-technical users can create dashboards with the use of pre-created queries made by analysts • Used in modern small to midsized technology communities (e.g. Autodesk)	• Midsized company (created in 2011 and $96M for last funding) • Mostly a visualization tools, but SQL queries are supported • No graph extensibility • Requires an analyst to maximize utilization • Representatives want to talk to UIUC purchasing reps. before demoing

Proposed Next Steps



Proposed Next Steps



Questions?

Citations

- https://www.gartner.com/doc/2983817?ref=AnalystProfile&srcId=1-4554397745
- http://www.datascienceassn.org/content/data-warehouse-and-data-management-solutions-analytics-magic-quadrant-2015
- http://www.menlo-technologies.com/menlo-technologies-and-business-intelligence-why-looker-caught-our-eye/
- https://www.crunchbase.com/organization/zoomdata#/entity
- https://docs.treasuredata.com/articles/tableau-desktop-odbc