Big Data for Software Engineering Estimation: Pros and Cons

Srinivasa Gopal
Ramanujan Society for Academic Research and Promotion of Science

About the Author

Srinivasa Gopal

Co-Founder – Ramanujan Society for Academic Research and Promotion of Science

Education :

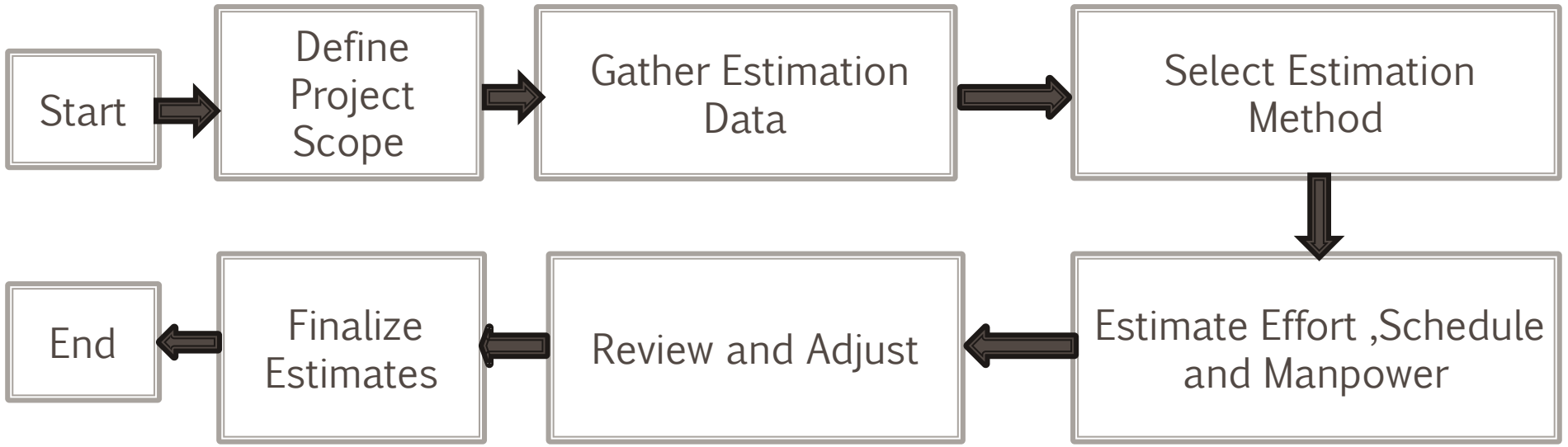
B.Tech in Mech Engg , Indian Institute of Technology, 1990
MS in Industrial Systems Engineering, University of Regina, Sask, Canada – 1992
MS in Information Technology, International Institute of Information Technology, Bengaluru, 2013

Awards : Governor General of Canada Gold Medal - 1992

Work Experience : Mainly in IT and Quality Assurance in multinational companies such as Infosys, Emirates Airlines, Unisys Corp, Land Mark IT, MasTech, GAVS Info Services

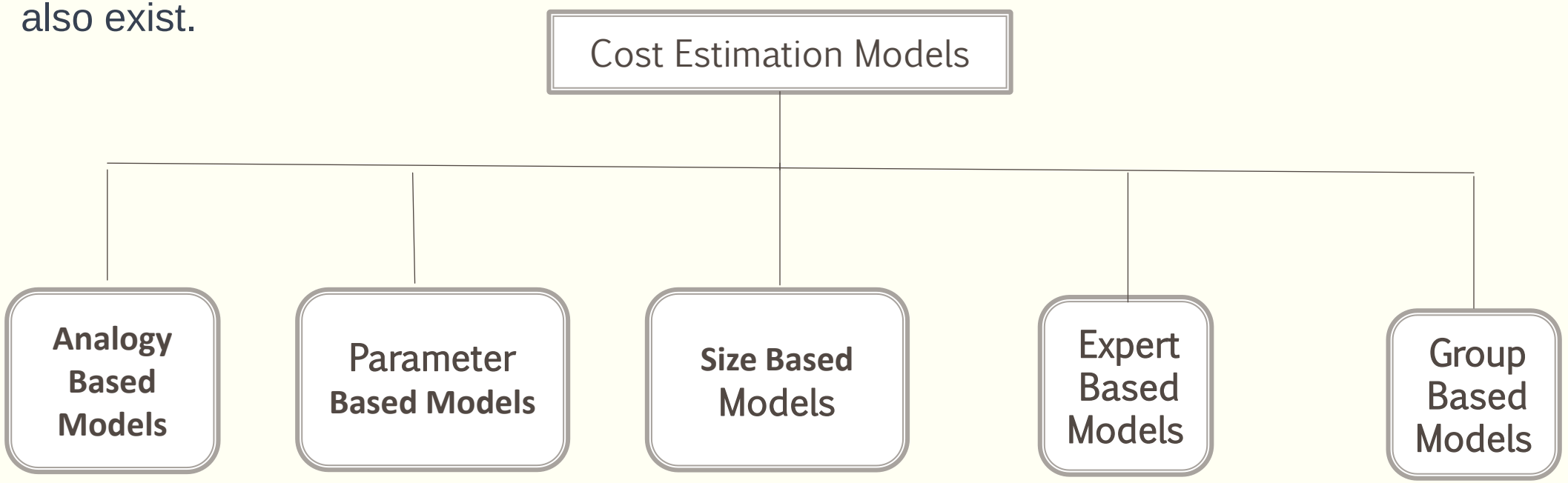
What is Software Engineering Estimation ?

Software engineering estimation involves predicting effort, schedule, and manpower needed for a successful software project. Crucial for development, maintenance, and support, estimates guide budgets and planning. Outputs include effort (measured in person-hours), schedule (time required), and manpower. Accurate estimation is vital for effective project management, guiding project scope and resource allocation.



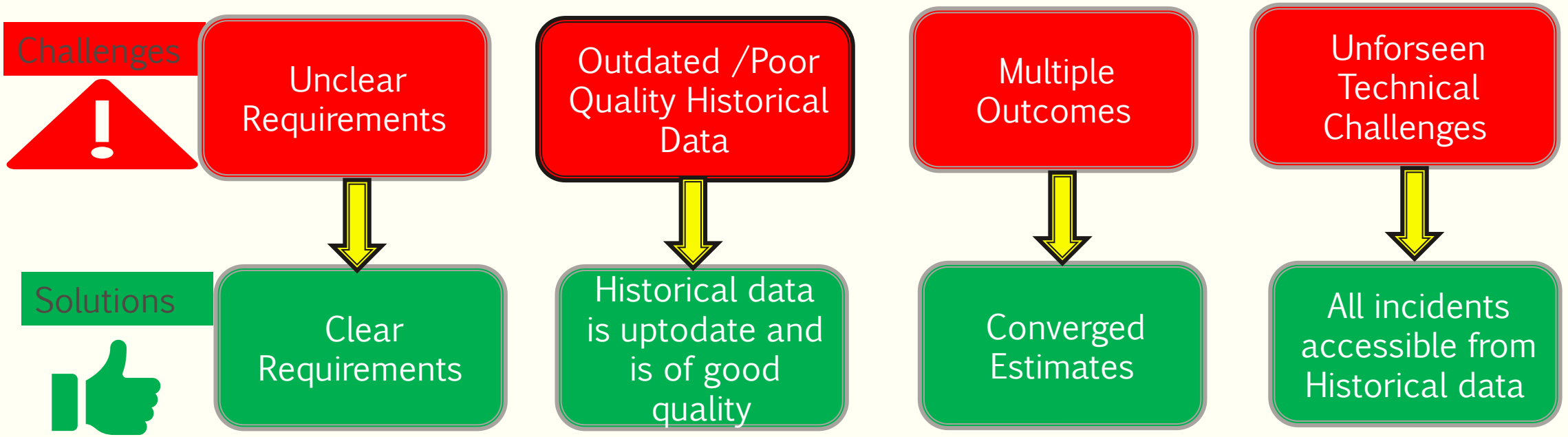
What are different types of estimation models ?

Estimation methods in software engineering include analogy-based models, comparing the current project to past similar ones using historical data such as the ISBSG project data set; parameter-based models, using statistical analysis based on software size, complexity, and team experience; size-based models, estimating effort from code size metrics; expert-based models, relying on experts' judgment; and group-based models, employing collective estimation techniques. Hybrid models combining these approaches also exist.

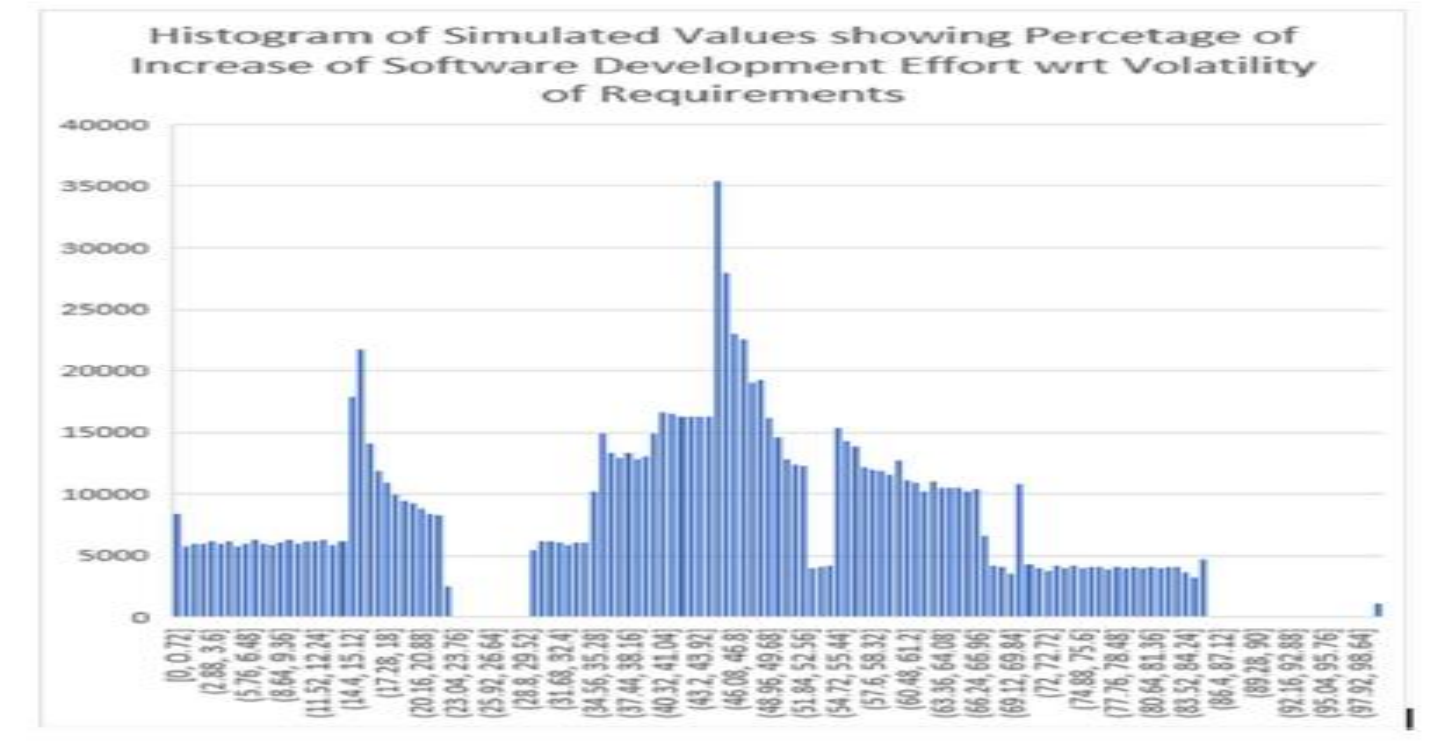


Challenges in Software Engineering Estimation

Software projects often encounter delays due to unclear or changing requirements. Estimation models rely on historical data like ISBSG's, draw on a vast dataset from diverse global projects. Outdated data hampers accuracy, especially if the current project varies significantly. Different estimation models yield varied results due to diverse algorithms and factors. Unanticipated technical challenges can cause delays and increased costs. To enhance accuracy, projects need clear requirements, updated data, appropriate techniques, effective scope management, open communication, and continuous estimate reviews.



Monte Carlo Simulation of wide variance in software effort due to volatility of requirements



Simulation shows volatile requirements impact effort disparity. Real-time strategies using historical data patterns or predictive corrective methods aid rapid estimation adjustments

Monte Carlo Simulation of the effect of Peer Review on Testing Effort

: MEAN DEFECT REMOVAL EFFORT UNDER DIFFERENT INSPECTION/TESTING RATIOS

Percent Defect removed via Inspection /Testing	Effort
100 percent removed via Inspection	4.30381
60 percent removed via Inspection	6.014812
40 percent removed via Inspection	6.879689
20 percent removed via Inspection	7.736111
0 percent Inspection, 100 percent Test	8.594696

Simulated output table demonstrates review impact on test effort reduction. Real-time feedback and historical data enable early defect identification, fostering efficient project management.

Monte Carlo Simulation of the effect of problem and solution complexity on Software Engineering Effort.

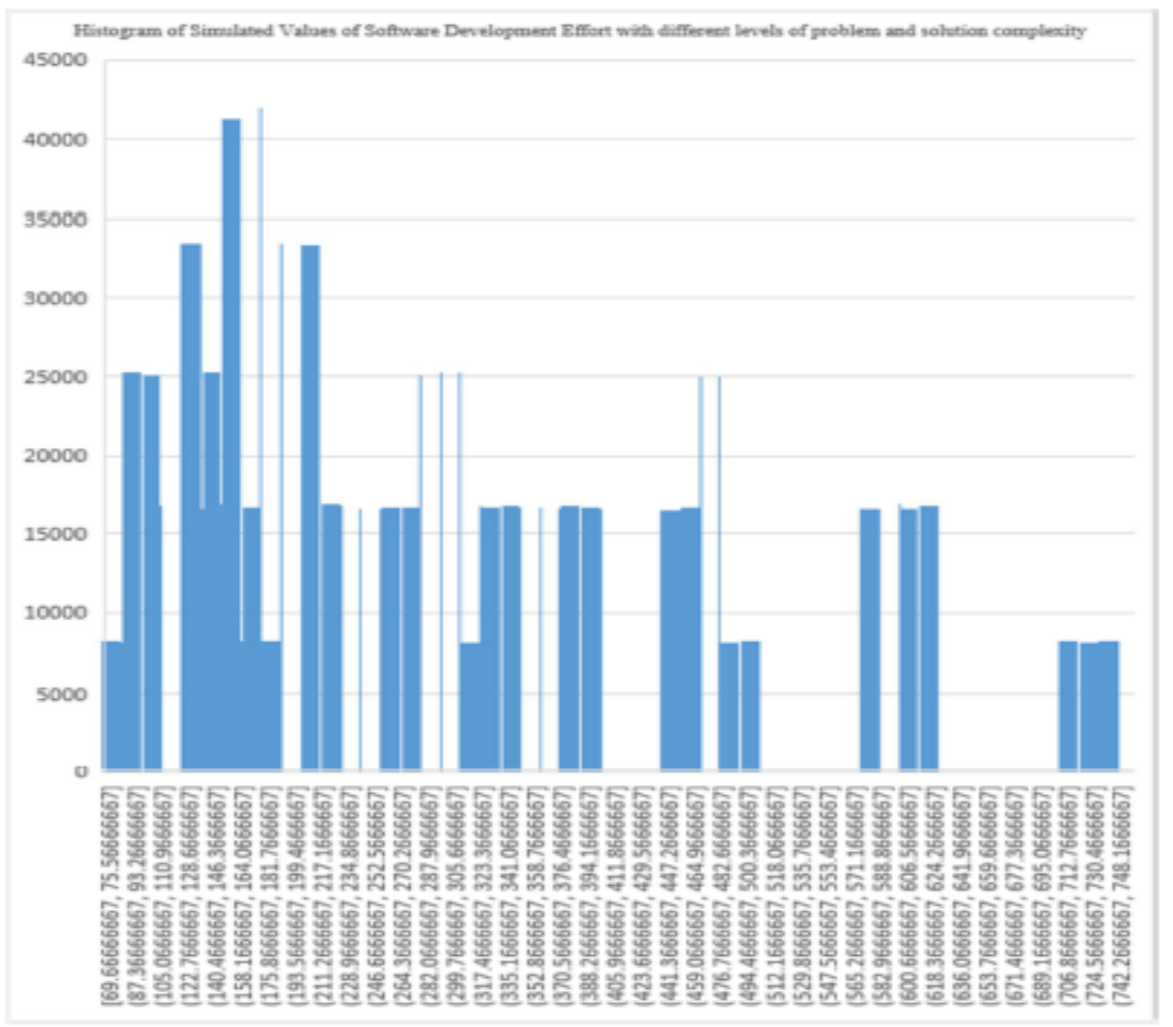
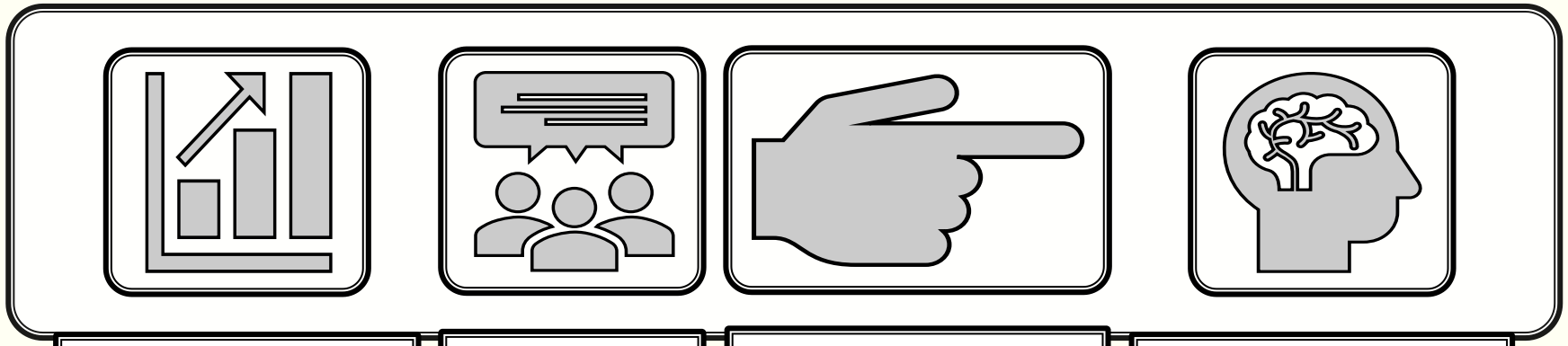


Figure highlights effort disparities, complicating accurate estimation amid varying problem complexities. Real-time feedback and historical data enable swift revisions, improving accuracy

Some unforeseen factors that influence Software Engineering Effort

Estimation challenges arise from factors like recent experience, peer code review, external dependencies, and innovative development. Parametric and analogy-based models lack nuanced parameters, ignoring peer review and external dependencies. Innovation-driven projects face unpredictable challenges. Traditional models overlook these aspects, making Agile and Scrum methodologies preferable for their adaptability.



Experience Impact:

Recent task experience improves accuracy .

Peer Code Review

Identifying improvements leads to rework.

External Dependencies

Market conditions, social and cultural influences and issues, legal restrictions, government or industry standards, financial considerations, and physical environmental elements.`

Innovative Development

Unpredictable challenges in cutting-edge projects

Unforeseen Factors that influence Software Engineering Effort(Cont..)

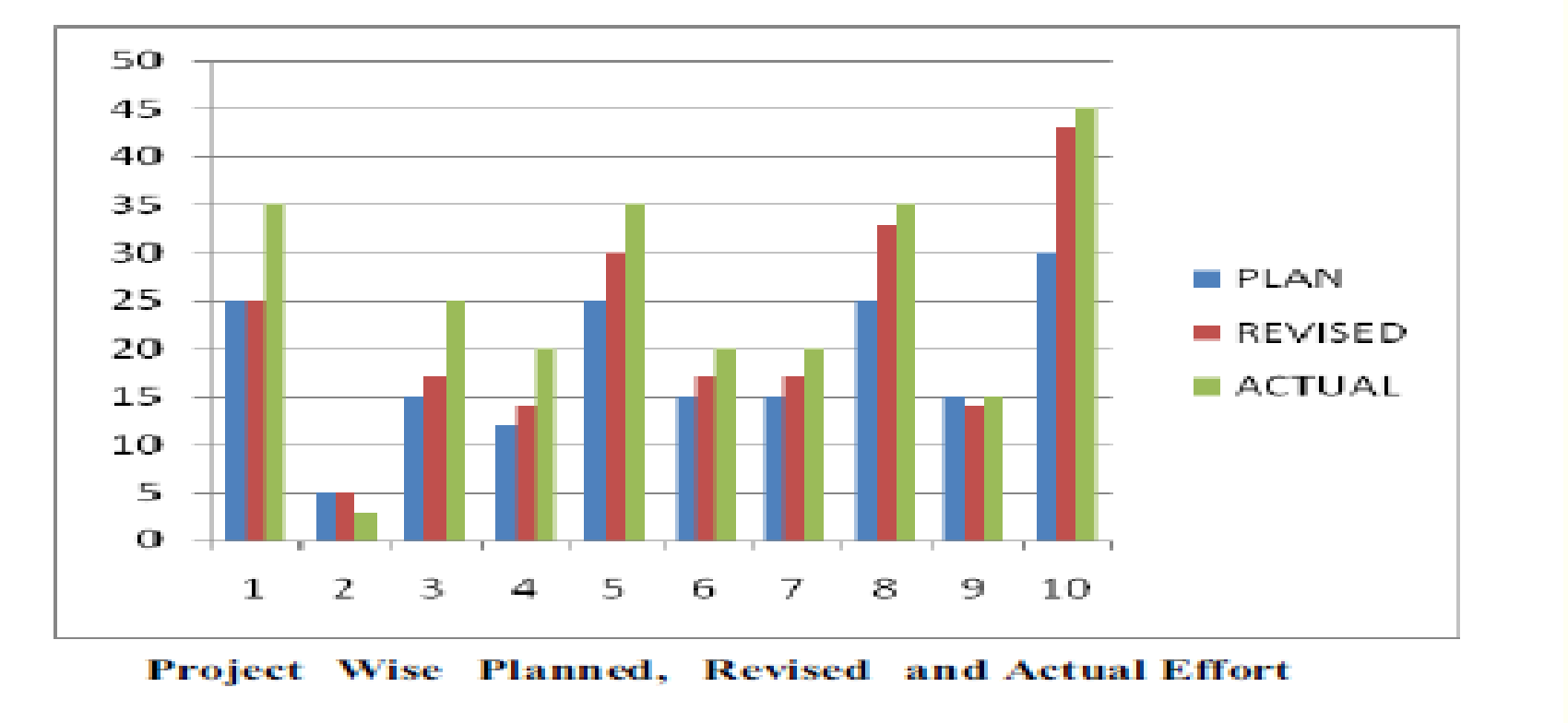
Adherence to processes impacts software development. While models like COCOMO don't consider it directly, it indirectly affects accurate estimation and risk mitigation techniques, enhancing project outcomes

Incorporating newly identified parameters into historical data collection process

Unforeseen parameters discovered during projects should be documented, incorporated into historical data repositories in addition to datasets available via ISBSG, collected regularly, and analyzed to support informed decision-making and continuous improvement efforts within organizations. By capturing hitherto unknown parameters or newly discovered parameters alongside explicit data, organizations can enhance their knowledge base, leading to more accurate estimations, better project planning, and improved overall project outcomes. For example ISBSG has historical data from 11128 Development & Enhancement projects ,1673 Maintenance & Support applications from around the world representing diverse industry and business types

Effect of adding parameters dynamically

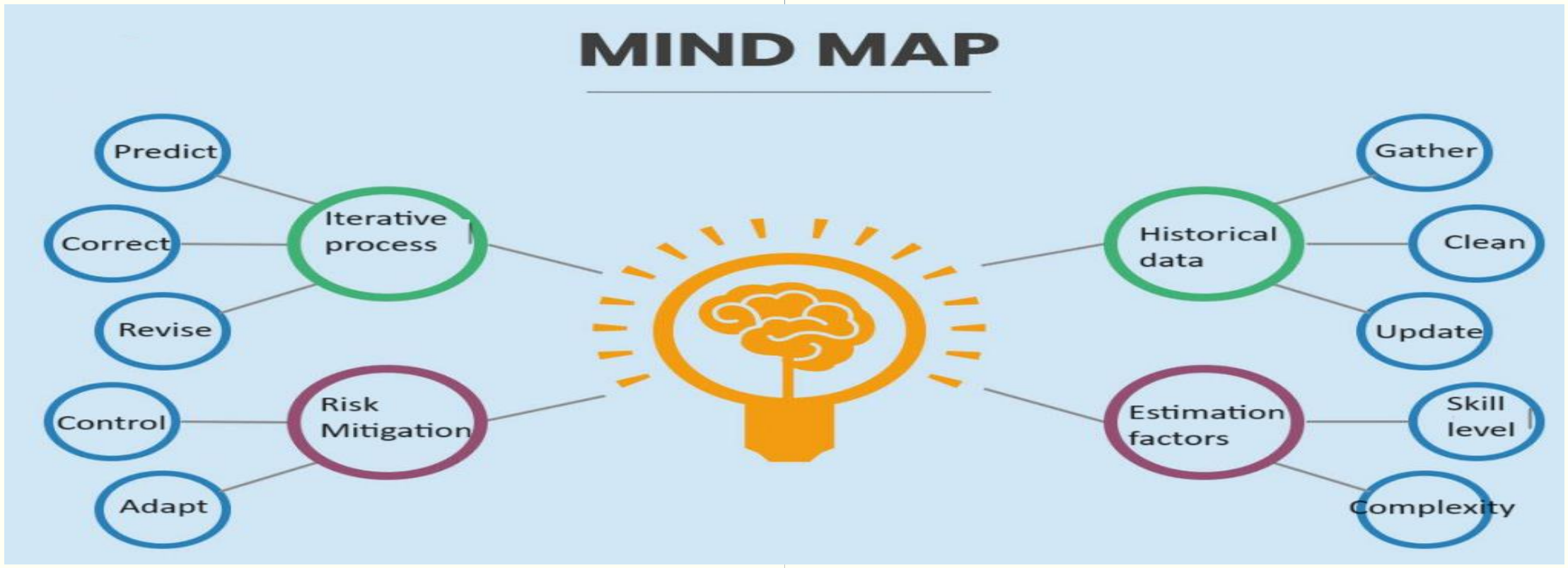
Figure below shows how estimation accuracy increases due to iteratively adding newly discovered parameters into the estimation process.



Source : *Srinivasa Gopal and Meenakshi D'Souza. 2012. Improving estimation accuracy by using case based reasoning and a combined estimation approach. In Proceedings of the 5th India Software Engineering Conference (ISEC '12). Association for Computing Machinery, New York, NY, USA, 75–78.* Click to read

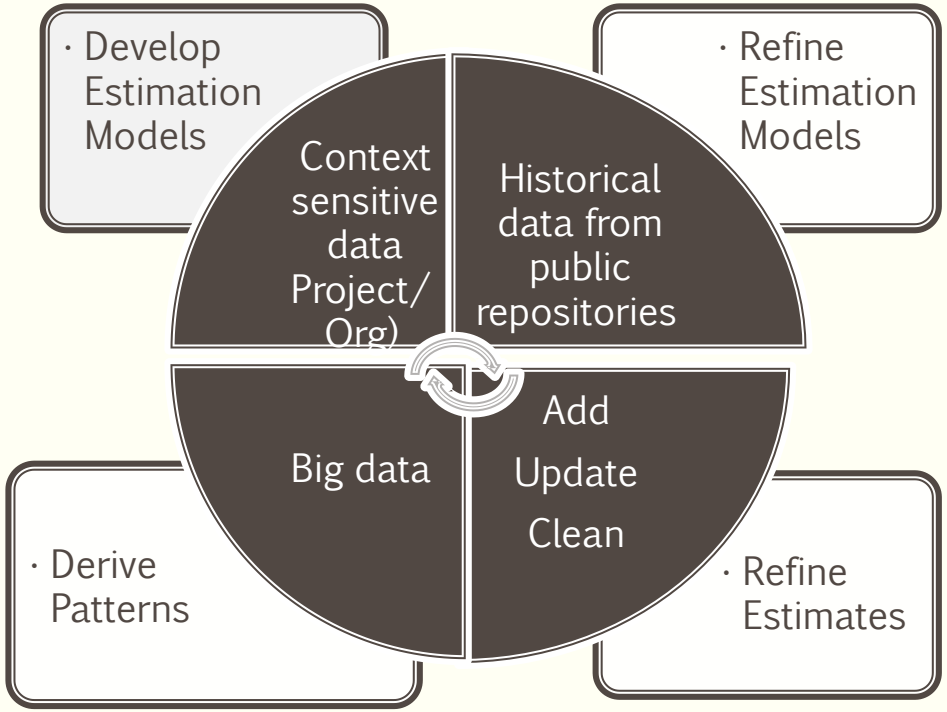
Harnessing Context-Sensitive Data ,Historical data , Dynamic parameter addition and Iterative Modeling

Accurate software engineering estimation relies on context-sensitive data, offering vital project background and specifics. Estimation models forecast project aspects like effort and cost. Project parameters vary based on the project's context. Gathering historical data on completed projects, including risks, failures, and team experiences, is essential. Incorporating new factors, like training recency, is crucial. Continuous updates to scenario modeling, simulations, and risk strategies enhance the estimation model's accuracy, adapting it to evolving project contexts.



Leveraging Big Data for Enhanced Software Estimation

Leveraging big data in software engineering estimation enhances accuracy by identifying patterns from past projects, predicting effort required. Big data tracks project development time, aiding precise estimates. It also helps identify risks by analyzing factors like changing requirements, guiding mitigation strategies. Predictive analytics algorithms and scalable, secure platforms enable accurate extrapolation from past patterns, improving estimation precision



Publicly Available Big Data Repositories for Software Estimation Research

ISBSG (International Software Benchmarking Standards Group): ISBSG (https://www.isbsg.org/) offers a comprehensive database of software projects and their related metrics. It provides valuable data for software estimation, benchmarking, and project management.

PROMISE Repository: The PROMISE repository (http://promise.site.uottawa.ca/SERepository/) provides a collection of datasets for software engineering research. It includes datasets related to software cost estimation, software defects, and other software engineering metrics.

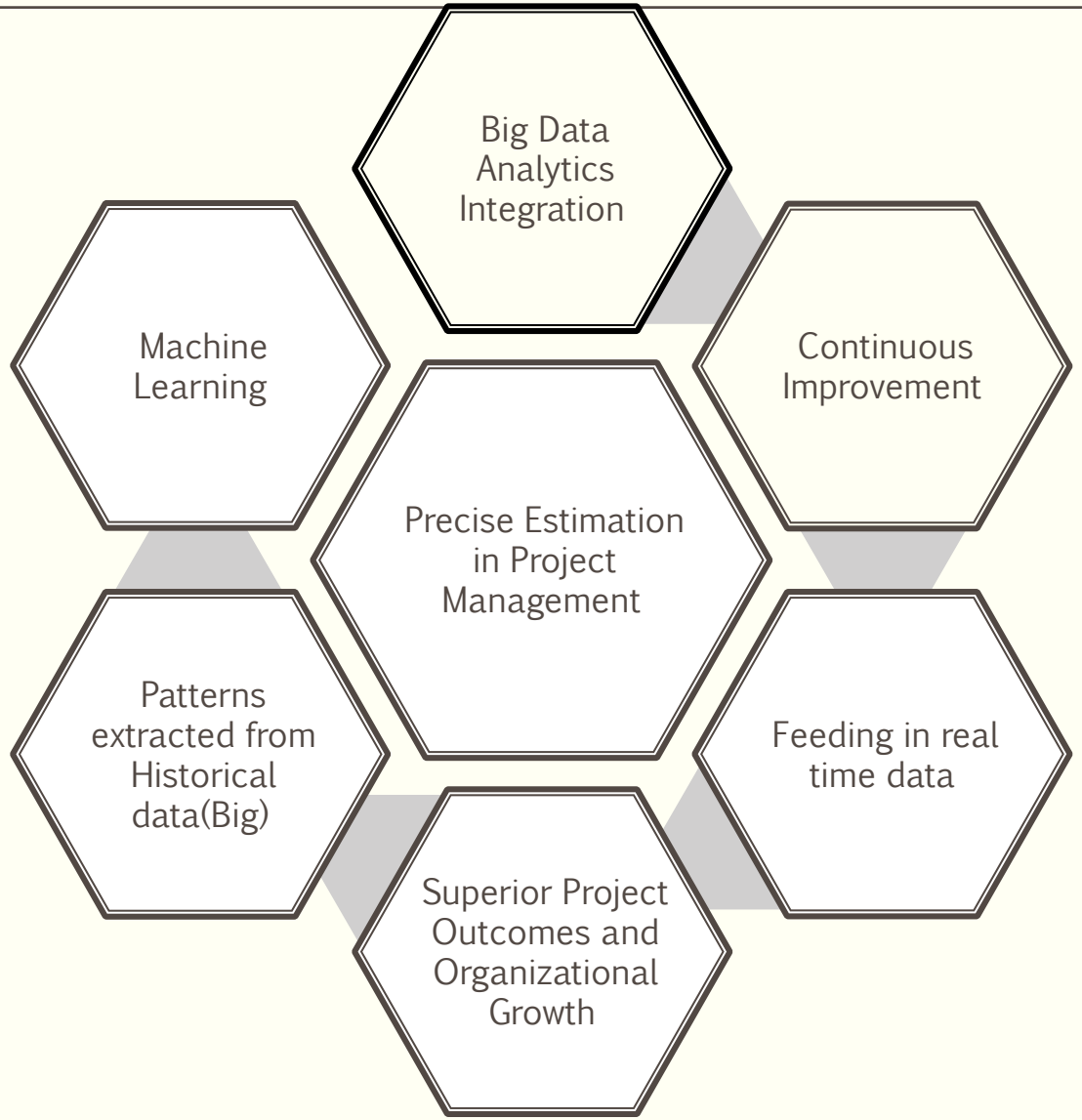
SEACRAFT Repository: The SEACRAFT (Software Engineering Artifacts) repository contains datasets related to software engineering artifacts, including software

NASA Software Engineering Laboratory (SEL) Dataset: NASA SEL provides datasets related to software projects conducted at NASA. These datasets include various metrics and can be used for software estimation research. Access to these datasets might require contacting NASA directly.

Apache Software Foundation Datasets: Apache provides a variety of software projects, and their development activities are publicly available. You can find repositories of their projects on GitHub (https://github.com/apache) and use this data for certain types of software estimation research.

GitHub Archive: While not specific to software estimation, GitHub Archive (https://www.gharchive.org/) provides a massive dataset of GitHub activity. Researchers can analyze this data to gain insights into software development trends, which might indirectly inform estimation models.

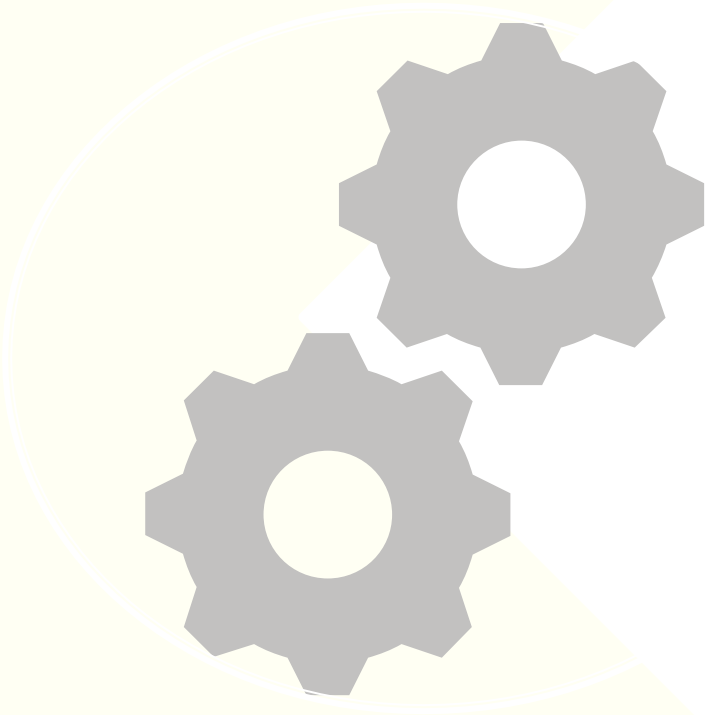
Enhancing Project Estimation Accuracy through Predictive Corrective Approach and Big Data Analytics



Big Data: Proactive Risk Management in Project Environments

Big data aids precise risk identification, analyzing vast data for patterns, enabling real-time monitoring, and predictive analytics foresee issues, ensuring project success

Big Data: Transforming Project Estimation with Efficiency and Precision



Big Data revolutionizes project estimation, enhancing efficiency and precision in decision-making and resource allocation for optimal outcomes.

Empowering Project Leadership: Informed Decision-Making with Big Data Insights

Big data revolutionizes project management decision-making by harnessing vast datasets such as ISBSG's dataset. In-depth analysis unveils intricate patterns, enabling leaders to anticipate challenges and make informed choices. Real-time feedback and machine learning-driven models foster agile responses. Data-driven strategies grounded in empirical evidence minimize uncertainties, empowering decision-makers. Continuous improvement, guided by big data, refines strategies in real-time, ensuring confident, precise decisions, and successful project outcomes

Improved Transparency

Leveraging big data in project estimation promotes transparency by providing real-time, accurate insights. Historical data analysis fosters realistic expectations, allowing open discussions about challenges and resources. Real-time monitoring enables instant progress tracking, aligning projects with initial estimates. Predictive analytics prepare stakeholders for future scenarios, fostering trust. This transparency builds mutual trust, grounding decisions in data, enhancing collaboration, and ensuring successful project outcome

Improved Project Communication

Big data transforms project communication through data-driven insights and real-time reporting. Decision-makers craft strategies grounded in concrete data, ensuring shared understanding. Visual representations simplify complexities, making progress accessible. Predictive analytics anticipate trends, enabling proactive discussions. Historical data offers valuable lessons, promoting knowledge sharing. Access to reliable data fosters open dialogue, enhancing communication channels. This transparency and proactive approach cultivate collaboration, ensuring effective project management and informed stakeholders

The Price of Insights: Understanding the High Costs of Big Data Collection and Storage

The high cost of managing big data stems from its sheer volume, velocity, and variety. Companies dealing with vast customer data invest heavily in storage infrastructure, high-speed processing systems, sophisticated software, and skilled professionals. Additionally, cybersecurity measures and redundant backups add to expenses. Despite challenges, businesses justify costs due to invaluable insights and competitive advantages gained from effective big data utilization

Decoding Complexity: Challenges in Analyzing Big Data

Analyzing big data presents formidable challenges due to its vast volume, rapid generation, diverse formats, and inherent complexity. The sheer scale, often reaching petabytes or exabytes, overwhelms traditional tools, necessitating robust processing. Real-time influx from sources like social media demands swift handling, a hurdle for conventional systems. Diverse data types and intricate algorithms deepen the complexity. Additionally, concerns about data quality, scalability, stringent security, and high costs add to the difficulty. Overcoming these challenges requires advanced technologies, skilled professionals, and meticulous planning to extract meaningful insights from big data.

Balancing Act: Legal, Security, and Privacy Challenges in Big Data Estimation

In big data estimation, businesses face legal, security, and privacy challenges. Sensitive data inclusion leads to constant threats of breaches and violations. To counter this, robust cybersecurity, encryption, and access controls are vital. Transparent communication, clear consent mechanisms, and legal adherence maintain trust. Striking a balance between innovation and legal compliance, businesses must ensure user privacy and data security while harnessing big data for estimation..

Thank You